ASTROPHYSICS AND SPACE SCIENCE LIBRARY 360

Jingquan Cheng

The Principles of Astronomical Telescope Design

ASSL

Springer

The Principles of Astronomical Telescope Design

Astrophysics and Space Science Library

Recently Published in the ASSL series

Volume 360: *The Principles of Astronomical Telescope Design*, by Jingquan Cheng. Hardbound ISBN: 978-0-387-88790-6, February 2009

Volume 354: *Sirius Matters*, by Noah Brosch. 978-1-4020-8318-1, March 2008

Volume 353: *Hydromagnetic Waves in the Magnetosphere and the Ionosphere*, by Leonid S. Alperovich, Evgeny N. Fedorov. Hardbound 978-1-4020-6636-8

Volume 352: *Short-Period Binary Stars: Observations, Analyses, and Results*, edited by Eugene F. Milone, Denis A. Leahy, David W. Hobill. Hardbound ISBN: 978-1-4020-6543-9, September 2007

Volume 351: *High Time Resolution Astrophysics*, edited by Don Phelan, Oliver Ryan, Andrew Shearer. Hardbound ISBN: 978-1-4020-6517-0, September 2007

Volume 350: *Hipparcos, the New Reduction of the Raw Data*, by Floor van Leeuwen. Hardbound ISBN: 978-1-4020-6341-1, August 2007

Volume 349: *Lasers, Clocks and Drag-Free Control: Exploration of Relativistic Gravity in Space*, edited by Hansjorg Dittus, Claus Lammerzahl, Salva Turyshev. Hardbound ISBN: 978-3-540-34376-9, September 2007

Volume 348: *The Paraboloidal Reflector Antenna in Radio Astronomy and Communication Theory and Practice*, by Jacob W. M. Baars. Hardbound 978-0-387-69733-8, July 2007

Volume 347: *The Sun and Space Weather*, by Arnold Hanslmeier. Hardbound 978-1-4020-5603-1, June 2007

Volume 346: *Exploring the Secrets of the Aurora*, by Syun-Ichi Akasofu. Hardbound 978-0-387-45094-0, July 2007

Volume 345: *Canonical Perturbation Theories Degenerate Systems and Resonance*, by Sylvio Ferraz-Mello. Hardbound 978-0-387-38900-4, January 2007

Volume 344: *Space Weather: Research Toward Applications in Europe*, edited by Jean Lilensten. Hardbound 1-4020-5445-9, January 2007

Volume 343: *Organizations and Strategies in Astronomy: Volume 7*, edited by A. Heck. Hardbound 1-4020-5300-2, December 2006

Volume 342: *The Astrophysics of Emission Line Stars*, by Tomokazu Kogure, Kam-Ching Leung. Hardbound ISBN: 0-387-34500-0, June 2007

Volume 341: *Plasma Astrophysics, Part II: Reconnection and Flares*, by Boris V. Somov. Hardbound ISBN: 0-387-34948-0, November 2006

Volume 340: *Plasma Astrophysics, Part I: Fundamentals and Practice*, by Boris V. Somov. Hardbound ISBN 0-387-34916-9, September 2006

Volume 339: *Cosmic Ray Interactions, Propagation, and Acceleration in Space Plasmas*, by Lev Dorman. Hardbound ISBN 1-4020-5100-X, August 2006

Volume 338: *Solar Journey: The Significance of Our Galactic Environment for the Heliosphere and the Earth*, edited by Priscilla C. Frisch. Hardbound ISBN 1-4020-4397-0, September 2006

Volume 337: *Astrophysical Disks*, edited by A. M. Fridman, M. Y. Marov, I. G. Kovalenko. Hardbound ISBN 1-4020-4347-3, June 2006

For other titles see www.springer.com/astronomy

The Principles of Astronomical Telescope Design

Jingquan Cheng

National Radio Astronomy Observatory Charlottesville, Virginia, USA

Springer

Jingquan Cheng
National Radio Astronomy Observatory
520 Edgemont Rd.
Charlottesville, Virginia
USA

ISSN 0067-0057
ISBN 978-0-387-88790-6 e-ISBN 978-0-387-88791-3
DOI 10.1007/b105475-3

Library of Congress Control Number: 2008942086

Printed on acid-free paper

springer.com

This book is dedicated to those who have provided help and encouragement in my thirty-year pursuit of astronomical telescope knowledge

Contents

Preface of English Edition

Progress in astronomy has been fueled by the construction of many large classical and modern telescopes. Today, astronomical telescopes image celestial sources not only across the wide electromagnetic spectrum from 10 m radio waves to 100 zm (10^{-19} m) gamma rays, but also through other spectra in gravitational waves, cosmic rays, and dark matter. Electromagnetic and other waves or particles cover a very wide energy density range. Very high energy cosmic rays have energy a billion times greater than that accelerated at Fermilab and some light dark matter particles have tiny energies beyond the detection limit of the finest existing quantum devices. Now astronomical telescopes are very large, very expensive, and very sophisticated. They are colossal in size, extremely demanding in technology, and terribly high in cost. Because of the technology, scale of construction, and the desire of scientists to plumb the depths of the Universe, astronomy today epitomizes the oft-used expression "Big Science."

Over the past 400 years, the size, the wave or particle types, and the spectral coverage of astronomical telescopes have increased substantially. Currently, large optical telescopes have apertures as large as 10 m (78 m^2). It is important to note that the total optical collecting area around the world in the past 20 years has more than tripled. At radio wavelengths, the largest collecting area of a single telescope is still dominated by the 300-m Arecibo telescope (roughly 70,000 m^2) although a 500-hundred-meter-diameter Aperture Spherical radio Telescope (FAST) is under construction in China. For interferometers, the Very Large Array (VLA, roughly 13,000 m^2) located in New Mexico (USA) is currently dominant. By comparison, the Atacama Large Millimeter Array (ALMA), presently being constructed in northern Chile, will have a collecting area of roughly 6,000 m^2. In gravitational wave detection, the Laser Interferometer Gravitational wave Observatory (LIGO) has two very long laser interferometer arms, each 4 km long (much longer if multi-reflection is taken into account). The sensitivity acquired by this instrument is as high as 10^{-21}. For cosmic ray detection, one site of the Pierre Auger Observatory has 30 fluorescence detectors and 1,600 water Cherenkov detecting stations over a surface area of 6,000 km^2 on earth. In the search for

dark matter particles, thousands of detectors are located inside ice layer between 1,400 and 2,400 m underground at the South Pole. Detectors are also located at other underground or underwater locations all over the world. Some of these detectors are working at extremely low temperatures of 20–40 mK.

At the current time, plans are underway to construct optical telescopes with apertures up to 42 m, radio telescope arrays up to a square kilometer aperture area, and space telescopes of diameters up to 6.5 m. Extremely sensitive gravitational wave detectors, large cosmic ray telescopes, and the most sensitive dark matter telescopes are also under construction. Larger aperture area, lower detector temperature, and sophisticated technology greatly improve the sensitivity of telescopes. This means more detecting power for fainter and far away objects and increasing clarity of star images. However, it is not just the size and accuracy of a telescope that matters; the gain in efficiency that results from performing many functions simultaneously and the ability to measure spectra and to monitor rapid variation are also important figures of merit.

Interferometry was pioneered by radio interferometers. A resolution of 50 milliarcsecs was routinely obtained by the VLA. Long baseline interferometry at millimeter wavelengths, using the Very Long Baseline Array (VLBA), can achieve a thousand times better angular resolution than that of the VLA. In the optical field, an important breakthrough has been achieved in optical interferometers. Another important achievement is the development of active and adaptive optics (AO). Active and adaptive optics holds promise to transform a whole new generation of optical telescopes which have large aperture size as well as diffraction limited image capability, improving the angular resolution of ground-based telescopes. In nonelectromagnetic wave detections, extremely low temperature, vibration isolation, adaptive compensation for interference, superconductor transition edge sensors, and SQUID quantum detectors are widely used for improving instrument sensitivity and accuracy. All of these are pushing technologies in many fields to their limiting boundaries. In general, modern telescope projects are very different from any other comparable commercial projects as they heavily involve extensive scientific research and state of the art innovative technical development.

To write a book on these exciting and multi-field telescope techniques is a real challenge. The author's intention is to introduce the basic principles, essential theories, and fundamental techniques related to different astronomical telescopes in a step-by-step manner. From the book, the reader can immediately get into the frontier of these exciting fields. The book pays particular attention to relevant technologies such as: active and adaptive optics; artificial guide star; speckle, Michelson, Fizeau, intensity, and amplitude interferometers; aperture synthesis; holographic surface measurement; infrared signal modulation; optical truss; broadband planar antenna; stealth surface design; laser interferometer; Cherenkov fluorescence detector; wide field of view retro-reflector; wavefront, curvature, and phasing sensors; X-ray and gamma ray imaging; actuators; metrology systems; and

many more. The principles behind these technologies are also presented in a manner tempered by practical applications. Telescope component design is also discussed in relevant chapters. Because many component design principles can be applied to a particular telescope design, readers should reference all relevant chapters and sections when a telescope design project is undertaken.

The early version of this book started as lecture notes for postgraduate students in 1986 in Nanjing, China. The notes had a wide circulation among the postgraduate students. In 2003, the Chinese version of this book was published. The book was well received by the Chinese astronomical community, especially by postgraduate students. With a wide circulation of the Chinese version, requests were received from English speaking students for an English language version. The translation of this book started in 2005. The basic arrangement of the book remains unchanged. The book is intended to target postgraduate students, engineers, and scientists in astronomy, optics, particle physics, instrumentation, space science, and other related fields. The book provides explanations of instruments, how they are designed, and what the restrictions are. This book is intended to form a bridge between the telescope practical engineering and the most advanced physics theories. During the translation of this English version, many experts and friends provided great help both with technical contents and the English language. Among these reviewers, Dr. Albert Greve of IRAM reviewed all chapters of this book. In the language aspect, Ms. Penelope Ward patiently reviewed the entire book. Without this help, the book translation project would not have succeeded.

Charlottesville, Virginia *Jingquan Cheng*
December 2008

Preface of Chinese Edition

Astronomical telescopes as important tools in the exploration of the universe are based on scientific theory and technology developments. This book provides a systematic discussion on these design principles behind various astronomical telescopes. The development of astronomical telescopes usually reflects the highest technology achievements of the times. Therefore, this book pays more attention to these telescope-related technologies. Some of these relevant technologies are applied not only in astronomy, but also in other fields, such as telecommunication, aerospace, remote sensing, military, high energy physics, atmospheric sciences, and so on. Special technologies used in astronomy include active and adaptive optics, artificial laser guide star, speckle, Michelson, Fizeau, intensity and amplitude interferometers, holographic surface measurement, infrared signal modulation, optical truss, broadband planar antenna, stealth surface design, and many more. Widely used techniques include optical mirror manufacture, mirror supporting, air and hydrostatic bearings, Stewart platform, encoders and actuators, system simulation, vibration control, homologous structural design, laser ranger, laser lateral positioning, wide field retro-reflectors, carbon fiber reinforced composites, tilt-meters, accelerometers, precision surface manufacturing, X-ray and gamma ray imaging, lightning protection, three-dimensional surface measurement, etc. This book also provides discussions on wind, temperature, and earthquake-induced effects on telescope performance. The telescope foundation design is also discussed.

The writing of this monograph has taken 16 years of time. The author had prepared multiple versions of manuscripts in order to best explain these complicated telescope-related theories and principles. The notes soon expanded as fresh information was gained through design and research practice. This book reflects the author's experiences and knowledge in the astronomical telescope field. This book is intended to be used by scientists, engineers, and students in astronomy, optics, communications, aerospace, remote sensing, structure, military, high energy physics, atmospheric science, mechanics, metrology, and other related fields.

During preparation of this book, many scientists, engineers, and friends provided advice and help. These include Yang Ji, Jiang Shi-yang, Ye Bian-xie, Wang She-guan, Ai Gong-xiang, Cui Xiang-qun, Zhao Gong, Cai Xian-de, Cheng Jing-yun, Fan Zhang-yun, Huang Ke-liang, Liang Ming. Among these, Prof. Jiang She-yang had reviewed all chapters of this book. Many publishers and science organizations also kindly granted permission to the author for use of the many figures in this book. The publication of this book was supported partly by Beijing Astronomy Observatory, China.

Beijing *Jingquan Cheng*
December 2002

Chinese Edition:
ISBN 7-5046-3392-5
The Principles of Astronomical Telescope Design, Jingquan Cheng
Chinese Science and Technology Press, 2003, Beijing

Author's other books:

Contributor of the book
ISBN 0-387-95512-7
The Design and Construction of Large Optical Telescopes,
Editor: Pierre Y. Bely
Springer, 2003, New York

Authored book
Chinese Edition:
ISBN 7-5046-4274-6
The Principles and Applications of Magnetism, Jingquan Cheng
Chinese Science and Technology Press, 2006, Beijing

Translators:
Zhang Yong, Miao Xinli, and Zheng Yi, Nanjing Institute of Astronomical Optics and Technology, China.

Expert Reviewers:
Yang Ji, Purple Mountain Observatory, Jiang Sheyang, Ye Binxun, Wang Sheguan, Ai Gongxiang, Cui Xiangqun, Zhao Gong, Cai Xiande, Beijing Astronomy Observatory, Cheng Jingyun, Shanghai Maritime University, Huang Keliang, Nanjing Normal University, Chen Lei, Nanjing Institute of Science and Technology, Fang Zhangyun, University of Arizona (in the above names, the first is the family name in normal Chinese order)

Albert Greve, Institut de Radioastronomie Millimetrique, France, Franz Koch, European Southern Observatory, Torben Anderson, Lund Observatory, Sweden, Bryan Colyer, Rutherford Appleton Laboratory, U.K., Larry Stepp, Thirty Meter Telescope project, Fang Shi, Jet Propulsion Laboratory, Ming Liang, National Optical Astronomy Observatory, John H. Bieging, Steward Observatory, James Lamb, Owens Valley Radio Observatory, Antony Stark, Harvard-Smithsonian Center for Astrophysics, Darrel Emerson, Fred Schwab, A. Richard Thompson, Dale A. Frail, Ken Kellermann, Bill Shillue, Jeffery Mangum, Clint Janes, David Hogg, Geoff Ediss, Paul Vanden Bout, Antonio Hales, Scott Ransom, Bill Cotton, Jim Braatz, Tim Bastian, Anthony J. Remijan, Darrell Schiebel, Rob Reid, and Richard Bradley, National Radio Astronomy Observatory

Language Editors:
Penelope Ward, Marsha Bishop, Ellen Bouton, Jennifer Neighbours, Tony Rodriguez, and Nicholas Emerson, National Radio Astronomy Observatory.

Cover Concept designers:
Patricia Smiley and Jingquan Cheng

Cover insert picture:
Aerial view of the Paranal summit with the four VLT Telescopes, in the front, YEPUN, and from left to right in the background, ANTU, KUEYEN and MELIPA (©ESO).

Cover background picture:
Star-Forming Region LH 95 in the Large Magellanic Cloud (NASA, ESA, Hubble Heritage team, and G. Gouliermis) (©STScI).

Chapter 1
Fundamentals of Optical Telescopes

This chapter provides a general overview of optical telescope history, astronomical requirements, optical aberrations, optical telescope system design, and modern optical theory. In this chapter, important optical concepts such as angular resolution, light collecting power, field of view, telescope efficiency, atmospheric seeing, geometrical aberrations, wavefront error, ray tracing, merit function, optical and modulation transfer function, point spread function, Strehl ratio, and imaging spatial frequency are introduced. The concept discussions are arranged in a systematic way so that readers can learn step-by-step. The chapter provides many important formulas of optical system design and evaluation. Emphases are placed on both the traditional geometric aberrational theory and the modern optical theory. At the end of the chapter, image properties of a segmented mirror system are also discussed in detail.

1.1 A Brief History of Optical Telescopes

Visible light is the only part of the electromagnetic radiation that can produce a response in human eyes. The wavelengths of visible light range between 390 and 750 nm. This region is also known as the optical or visible (VIS) region. The human eye is a complex organ composed of a light collector and detector. The eye is very sensitive. If one injects seven photons of 500 nm wavelength into a human eye within a time interval of 100 ms, the eye will produce a response (Schnapf and Baylor, 1987). The rod cells in the human eye play an important role for the sensitivity under a dark environment. A single photon may produce a response from a rod cell. Cone cells are less sensitive. However, as a light collector, the collecting area of the eye is very limited. The maximum iris opening (pupil) is only about 6 mm in diameter. It can only collect a very small part of light within a cone angle from a radiation source. The separation between light-sensitive cells in the eye is 2.5 μm. The angular resolution of the eye is about

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_1,

1 arcmin. To detect very faint sources or to separate two closely located celestial sources, one has to use optical telescopes.

The invention of the optical telescope is surrounded by controversy. One story puts it in a shop of a Dutch lens maker, Hans Lippershey, in October of 1608. As the story goes, two children were playing with his lenses, put two together, peered through them at a distant church tower and saw it wonderfully magnified. This was the first known optical telescope. In July of 1609, Galileo Galilei developed the first astronomical optical telescope. With his simple telescope, he made important early discoveries in astronomy. His telescope is formed by the combination of a concave and a convex lens, today known as the Galileo telescope system [Figure 1.1(a)]. The front convex lens is called an objective because it is close to the object being viewed and the rear concave lens is called an eyepiece because it is closest to the observer's eye. Binoculars based on Galileo telescope design, known as opera glasses, were also invented in the same period.

In 1611, Johannes Kepler invented another type of longer telescope comprising two convex lenses [Figure 1.1(b)], known as the Kepler telescope system. The Kepler telescope forms an upside down image, but with a slightly larger field of view.

Telescopes for personal use with the eye, including binoculars, are called afocal optical systems because rays of light from a distant object that are parallel when they enter the objective are also parallel when they exit the eyepiece. For afocal telescopes, an important parameter is angular magnification. The angular

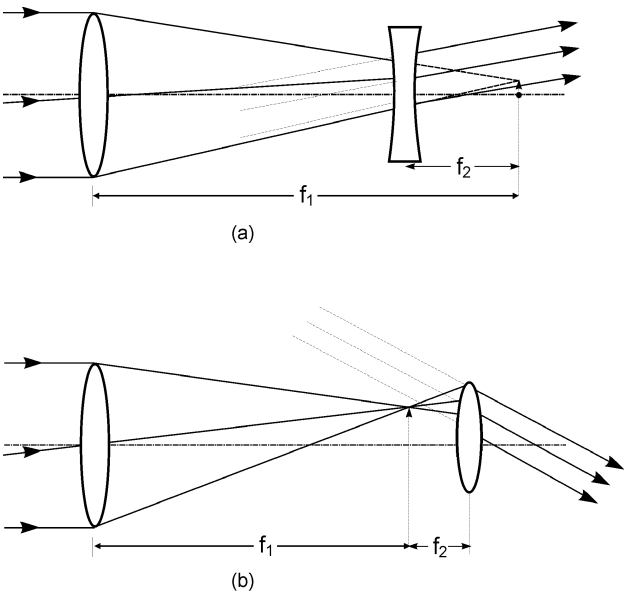


Fig. 1.1. (**a**) Galileo telescope and (**b**) Kepler telescope.

magnification or magnification factor is the ratio between the angle subtended by the output image and the angle subtended by the input object. Kepler telescopes usually have larger magnification than Galileo ones.

Large telescopes are often used to form a direct image, for example, an image on a detector. For these large telescopes, another parameter, resolution, is used instead of the magnification. Resolution is the ability to separate two closely located objects. Angular resolution is discussed in Section 1.2.1. However, the magnification concept is still used for the secondary mirror of modern telescopes.

Earlier telescopes were made of simple glass lenses, and the chromatic aberration caused by the change in refractive index with wavelength was a serious problem. To reduce chromatic aberrations, Christiaan Huygens proposed using lenses with smaller curvature. This, however, resulted in long tube lengths for early refractive telescopes. A telescope built in 1655 had an objective lens of 5 cm diameter and a tube length of 3.6 m. Johannes Hevelius built an even longer telescope with a tube length of 46 m.

In 1664, James Gregory proposed a spherical error-free reflecting optical telescope with conic section mirrors, known as the Gregorian telescope. However, the elliptical concave secondary mirror used in this system was difficult to make at that time. The system was never built by Gregory. Isaac Newton was the first to construct a usable reflecting telescope. In January of 1670, he produced a much simpler optical system with a parabolic mirror and an inclined small flat mirror, today known as a Newtonian telescope. Meanwhile, a French optician N. Cassegrain came up with a third configuration for a reflecting telescope, called the Cassegrain telescope. Instead of the concave secondary mirror of a Gregorian telescope, the Cassegrain system uses a convex hyperboloid. Reflecting telescopes are free from chromatic aberration and can have a short tube length. The fabrication of a reflecting telescope was still difficult in the early days. Mirror material stiffness and tighter surface requirements were the main problems.

In 1672, Newton claimed that there was no way to eliminate the chromatic aberration of a refracting optical telescope. However, achromatic lenses were invented 86 years later by John Dollond in 1758. The first use of achromatic lenses in astronomical telescopes was in 1761. Achromatic lenses bring two colors (e.g., red and blue) into focus in the same plane. The telescope performance improved and the tube length of refracting telescopes was reduced about ten times. This accelerated the development of large refracting telescopes. If three colors (e.g., red, green, and blue) are brought into focus in the same plane, the lenses are called apochromatic ones.

Binoculars of the Kepler system have an inverted image. In 1854, a double prism Z-shaped configuration to erect the image was invented by Ignazio Porro. Binoculars using roof prisms appeared as early as the 1880s. These are the two optical designs most commonly used in binoculars today.

In 1888, a 91.44 cm refracting telescope was built east of San Jose, California at Lick Observatory. In 1895, the Yerkes Astronomical Observatory produced a refracting telescope of 1 m diameter, the largest refracting telescope in the world.

At the same time, progress was also being made in constructing large reflecting telescopes. In the 1840s, Lord Ross produced a 1.8 m reflecting optical telescope. In 1856, the first optical telescope with a coated silver surface mirror was produced. In 1917, the 2.54 m Hooker reflecting telescope was built at Mount Wilson Observatory. In 1934, a new method of vacuum aluminum coating was invented. The reflectivity was improved significantly by this aluminum coating. In 1948, the giant 5 m Hale optical telescope was built at Palomar Observatory, 130 km southeast of the Hooker telescope.

At the beginning of the 20th century, two special types of astronomical telescopes were developed; astrometric and solar telescopes. Astrometric telescopes require high positional accuracy, while solar ones require reduction of the Sun's heat. Today, astrometric telescopes are merged with normal optical telescopes, while solar telescopes still remain a special type of telescope. At the same time, efforts were made to increase the field of view of telescopes. In 1931, a wide field so-called catadioptric telescope involving reflecting and refracting components known as a Schmidt telescope was invented. The field of view of a Schmidt telescope can be 6×6 square degrees.

Beginning in the 1960s, there were continuous developments in optical telescope technology. In 1969, the USSR built the 6 m Bolshoi Teleskop Azimultalnyi also known as the Big Telescope Alt-azimuthal (BTA). It used, for the first time in a large optical telescope, an altitude-azimuth mount. Compared to equatorial mounts, altitude-azimuth mounts have a more direct load path for transferring the weight of telescope to the ground. One area where this is particularly important is in the difference between declination axis bearings on equatorial telescopes, which must support the weight of the telescope over a range of orientations, and elevation axis bearings on altitude-azimuth mounts, which have a fixed orientation. The new mounting makes optical telescopes with even larger diameters possible. In 1979, the 4.5 m Multiple Mirror Telescope (MMT) was built in the U.S. This telescope combined six 1.8 m sub-telescopes in one structure. Although the MMT was modified in 1998 into a single mirror 6.5 m telescope, the original version nevertheless had a great impact on modern telescope design.

In 1990, the Hubble Space Telescope (HST) was launched into space. This was the first major space telescope in astronomy. In the HST, 24 actuators were on the mirror back, intended for deforming the mirror to improve imaging performance. This was an attempt to control a telescope actively even in space. The first working active optics system was built in 1989 on the ESO New Technology Telescope (NTT) in Chile. In 1992, the 10 m Keck I segmented mirror telescope (SMT) was built. The primary mirror of this telescope is composed of 36 hexagonal segments of 1.8 m size. This represents a milestone in modern telescope design.

In 1997, the 10 m Hobby-Eberly Telescope (HET) was built with a segmented spherical mirror and a fixed altitude mounting. In 1998, the second Keck II telescope was completed.

In 1999, the Subaru 8.2 m and the northern Gemini 8 m telescopes were built. The Very Large Telescope (VLT) with four 8.2 m telescopes was built in 2000, and the southern Gemini telescope was completed in 2002. These seven telescopes are large single mirror ones.

The 10 m Gran Telescope Canarias (GTC) in Spain and the 10 m South African Large Telescope (SALT) were built in 2003 and 2005, respectively. The Chinese 4 m Large sky Area Multi-Object fiber Spectroscopic Telescope (LAMOST), a reflecting Schmidt telescope with a field of view of 5×5 square degrees, was built in 2008. It involves both segmented mirror and active optics technologies. The giant Large Binocular Telescope (LBT) had its first light for both mirrors in 2008. This is a MMT-style Fizeau interferometer instrument with two large 8.4 m sub-telescopes.

Now the 22 m Giant Magellan Telescope (GMT), the 30 m Thirty Meter Telescope (TMT), and the 42 m European Extremely Large Telescope (E-ELT) are under serious design study. These are a new generation of modern astronomical optical telescopes.

With the development of modern computer and control systems, active optics control soon expanded into the higher frequency region, which became adaptive optics. Adaptive optics, using natural guide stars, was developed in the 1970s and 1980s by the military, but was first used in astronomy in 1989 at the European Southern Observatory. However, natural guide stars have limited sky coverage. The use of laser guide stars was proposed by Foy and Labeyrie (1985) in 1985 and was realized by the military first before 1990. It expands the sky coverage of the adaptive optics. Even so, the field of view with laser guide stars is still limited. At the turn of the century, multi-laser guide stars, atmospheric tomography, and multi-conjugated adaptive optics brought a new revolution in ground-based optical telescope design.

The development of optical interferometers is parallel to the development of telescopes. In 1868 Hippolyte Fizeau proposed for the first time an interferometric method with two separated apertures for measuring the diameter of a star. In 1891, Albert Michelson realized this technique which is now called a Michelson interferometer. A Michelson interferometer with a real and a mirror imaging telescope from the sea surface was first realized in radio wavelength in 1945 by Pawsey et al. (1946). Aperture synthesis in radio wavelengths was realized afterwards.

In 1956, Hanbury-Brown and Twiss realized an optical intensity interferometer. In 1970, Antoine Labeyrie produced a high-resolution optical speckle interferometer. An optical Michelson interferometer was also realized by Labeyrie in 1976. In 1995, a Fizeau interference image from separated optical telescopes was obtained by the Cambridge Optical Aperture Synthesis Telescope (COAST).

The development of optical telescopes provides astronomers with larger and larger light collecting area and higher and higher angular resolution. At the same time, it pushes the design of optical systems, mounting structures, sensors, actuators, control systems, and receiver systems to their present limits, producing design revolutions in many related fields.

1.2 General Astronomical Requirements

In optical telescope design, three basic requirements are high angular resolution, large light collecting power, and large field of view. When an image is formed by an optical telescope, the first question one would ask is: Is the source a single star or made up of closely spaced stars? To resolve closely spaced stars, high angular resolution is necessary. The second question one would ask is: Is a telescope able to detect very faint stars? The photons collected by a telescope from a star are proportional in number to the area of the telescope aperture, while the photon number from the star per unit area on the earth is inversely proportional to the square of the distance to the star. To detect a very faint or a very distant star, a large aperture, high telescope efficiency, and good clear site are required. Together these define the light collecting power of the telescope. The third question one would ask is: How many stars can one record inside the image field? For this question, a large field of view is required. This section will discuss all these related topics.

1.2.1 Angular Resolution

The angular resolution of an optical telescope is its ability to resolve two closely located point objects. Three factors influence the angular resolution of a telescope. These are diffraction of the aperture, aberrations of the optical system, and atmospheric turbulence.

In Gaussian optics, light from a single point object will arrive at a single point in the image space. The form of Gaussian optics known as paraxial or first-order optics is the first approximation of a practical optical system. In Gaussian optics, the sine of an angle is replaced by the angle itself. There is no aberration in a Gaussian system.

The term of aberration in optics describes the geometrical difference between a practical image and the corresponding Gaussian image. If the third order aberrations (first few terms of the aberration) are included, the second approximation of a practical optical system is referred to as classical optics where only third power terms of the aberrations are considered.

Geometrical optics is the third approximation for a practical optical system, where all the aberration terms are included. Geometrical optics does not fully describe a practical optical system which is influenced by the wave features of light.

One important property of light is interference. By including the wave effect, the fourth approximation of a practical optical system is referred to as physical optics. Physical optics includes diffraction, interference, and aberrations in a practical optical system. For optical telescope design, this stage of approximation is generally adequate.

The next stage in optics will be quantum optics where light is studied as quantized photons.

The aberrations and wavefront errors of an optical system are discussed in latter sections. In this section the aperture diffraction and atmospheric turbulence are discussed.

In physical optics, light also acts like a wave. Therefore, the actual image of a point source even without aberrations is not a sharp Gaussian image but a diffraction pattern. The size of this pattern determines the angular resolution of the optical system. The pattern is a function of the size and shape of the main mirror (aperture), due to the effects of Fraunhofer diffraction. It has many names depending on the conditions and the parameters being used, such as Fraunhofer pattern, diffraction pattern, far field pattern, radiation pattern, point spread function, intensity pattern, and, in some case, Airy disk. Some of these refer to the complex amplitude and phase, others to the amplitude squared (intensity).

If the Fraunhofer diffraction of an aperture field, S, is considered, then any small area ds on the aperture will have a contribution of light radiation in a direction of P (Figure 1.2). $F(x,y)$ is a complex field function in the aperture. The radiation contribution of an element $ds = dxdy$ in a direction of P will be $F(x,y)\,dxdy$, but with an added phase of $(2\pi/\lambda)(lx+my)$, where (l,m,n) are the direction cosines of the direction P and λ the wavelength of light. An aperture, as entrance pupil, is an opening through which light or radiation is admitted. It is usually a projection of the primary mirror which determines the cone angle of a bundle of rays that come to a focus in the image plane.

The radiation amplitude in the direction (l,m) is the integral of the contributions of each small area (Graham Smith and Thompson, 1988):

$$A(l,m) = C \iint_{Aperture} F(x,y) \exp\left[-\frac{2\pi \cdot j}{\lambda}(lx+my)\right] \cdot dxdy \qquad (1.1)$$

where $j^2 = -1$ and C a constant with the unit of $[\text{length}]^{-2}$. The phase term in the equation for each small area includes two parts. One is the phase of the aperture

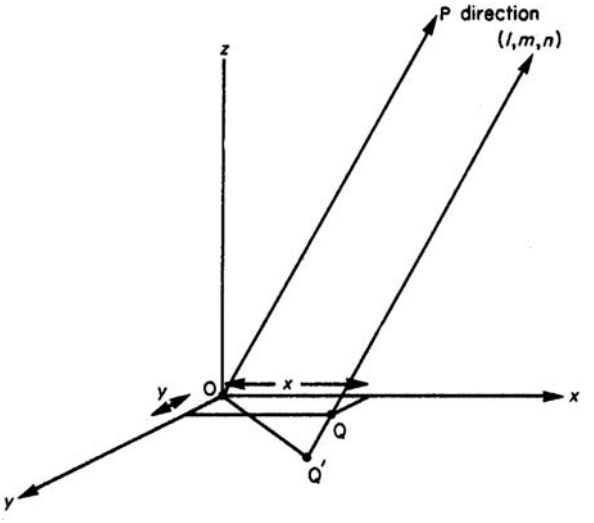


Fig. 1.2. Radiation contribution to a direction by a small area on an aperture field (Graham Smith and Thompson, 1988).

function itself of $F(x,y)$ and the other is the phase due to a path length of $QQ' = lx+my$. Mathematically, this equation represents approximately the Fourier transform of the aperture function $F(x,y)$, where x/λ and y/λ are dimensional variables. Therefore, the Fraunhofer diffraction pattern is the Fourier transform of the aperture field.

For a rectangular aperture with sides a and b, placed symmetrically about the origin and being uniformly illuminated, we have $F(x,y) = F_0$. By separating the variables:

$$A(l,m) = C \cdot F_0 ab \frac{\sin(\pi \cdot la/\lambda)}{\pi \cdot la/\lambda} \frac{\sin(\pi \cdot mb/\lambda)}{\pi \cdot mb/\lambda} \tag{1.2}$$

This is the diffraction pattern for a rectangular aperture without aberration.

For a uniformly illuminated circular aperture of a radius a, the above equation becomes:

$$A(w,\phi) = C \cdot F_0 \int_0^a \int_0^{2\pi} \exp\left[-\frac{2\pi j}{\lambda} rw\cos(\theta - \phi)\right] \cdot rdrd\theta \tag{1.3}$$

where (r,θ) is the polar coordinates in the aperture plane and (w,ϕ) in the image plane. It is equal to:

$$A(w,\phi) = C \cdot F_0 \frac{2J_1(2\pi aw/\lambda)}{2\pi aw/\lambda} = A(0,0)\frac{2J_1(2\pi aw/\lambda)}{2\pi aw/\lambda} \tag{1.4}$$

where $A(0,0) = C \cdot F_0$ is the amplitude of the radiation field at the origin, $J_1(x)$ the first-order Bessel function, $r\cos\theta = x$, $r\sin\theta = y$, $w\cos\phi = l$, $w\sin\phi = m$, and $w = (l^2 + m^2)^{1/2}$ the sine of the angle of the radiation direction measured from the z-axis. The intensity (power, energy) distribution of the diffraction pattern is:

$$I(w,\phi) = A^2(0,0)\left[\frac{2J_1(2\pi aw/\lambda)}{2\pi aw/\lambda}\right]^2 \tag{1.5}$$

This equation shows that the diffraction pattern of a circular aperture is axially symmetric with bright and dark rings. This pattern is called the Airy disk. The cross sections of amplitude and intensity patterns of an Airy disk are shown in Figure 1.3. The radius of the first zero (dark ring) of an Airy disk is at $1.22\lambda/d$, where d is the aperture diameter.

Generally, the intensity distribution of the diffraction pattern is called the point spread function (PSF) of an aperture field. The diffraction pattern of a

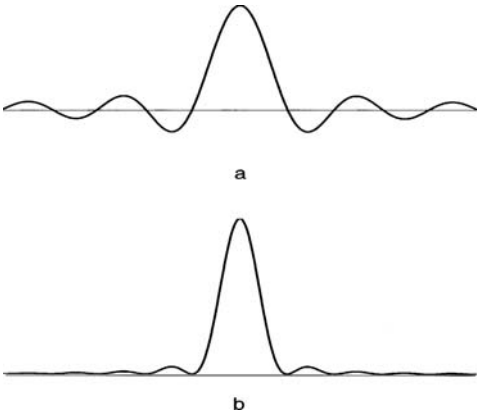


Fig. 1.3. The cross sections of the field pattern (a) and the intensity diffraction pattern (b) of a circular aperture.

dual-reflector telescope is slightly different due to the presence of a central blockage. For circular aperture and circular blockage, the amplitude pattern is:

$$A(w,\phi) = A(0,0)\left[\frac{2J_1(2\pi aw/\lambda)}{2\pi aw/\lambda} - \beta^2\frac{2J_1(2\pi\beta aw/\lambda)}{2\pi\beta aw/\lambda}\right] \tag{1.6}$$

where β is the blockage ratio. The blockage ratio is the relative radius of the secondary mirror blockage on the aperture plane. The intensity pattern is:

$$I(w,\phi) = A^2(0,0)\left[\frac{2J_1(2\pi aw/\lambda)}{2\pi aw/\lambda} - \beta^2\frac{2J_1(2\pi\beta aw/\lambda)}{2\pi a\beta w/\lambda}\right]^2 \tag{1.7}$$

This equation shows that the radius of first dark ring becomes smaller due to central blockage. Tables 1.1 and 1.2 give the radius of the first dark ring and the energy distribution in different rings of the diffraction pattern for the wavelength of 550 nm when central blockage is considered.

Table 1.1. The radius of the first dark ring in the intensity diffraction pattern of a circular aperture with different central blockage and at a wavelength of 550 nm

Central blockage ratio	0.0	0.1	0.2	0.3	0.4	0.5
Diameter of first dark ring (in arcsec)	13.8/d	13.6/d	13.2/d	12.6/d	11.9/d	11.3/d

Note: The unit of the aperture diameter d is in centimeter.

Table 1.2. The energy distribution in the intensity diffraction pattern of a circular aperture with different central blockage and at a wavelength of 550 nm

Central blockage ratio	0.0	0.1	0.2	0.3	0.4	0.5
Central spot	83.78%	81.84%	76.38%	68.24%	58.43%	47.86%
First ring	7.21%	8.74%	13.65%	21.71%	30.08%	34.99%
Second ring	2.77%	1.90%	0.72%	0.47%	1.75%	7.29%
Third ring	1.47%	2.41%	3.98%	2.52%	0.39%	0.17%

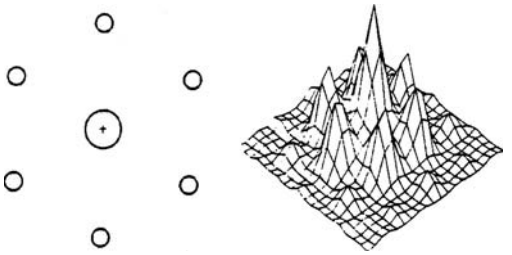


Fig. 1.4. An array of apertures in an interferometer and the resulting diffraction pattern.

Using the same method, the diffraction patterns for an unfilled aperture field can be derived. Figure 1.4 shows an array of apertures in an interferometer and its radiation pattern or its point spread function.

It is the intensity diffraction pattern that determines the angular resolution of an optical telescope. The angular resolution is defined as the minimum angle between two distinguishable objects in image space. In modern optics, the resolution is expressed as a cut-off spatial frequency of the aperture field (Section 1.4). The cut-off spatial frequency of an aperture is represented by the wavelength to diameter ratio.

Without the concept of spatial frequency, one requires empirical criteria for the separation of two stellar images with the same brightness. Three criteria were used in classical optics: the Rayleigh, the Sparrow, and the Dawes criteria.

The Rayleigh criterion is widely used. In 1879, Rayleigh suggested that if two star images of the same brightness are so close that the first dark ring of one pattern is at the center of another, then two images are resolved [Figure 1.5(a)]. In this case, the brightness in the middle of two images is about 0.735 of the maximum for the combined image.

The Sparrow criterion is less strict. From this criterion, if two images get so close that the faint middle point just disappears, then two stars are resolved. If two stars become even closer, there will be only one peak in the image [Figure 1.5(c)]. The Dawes criterion was derived after a long period of studies. It is in between the Rayleigh and Sparrow criteria. In Dawes criterion, there is

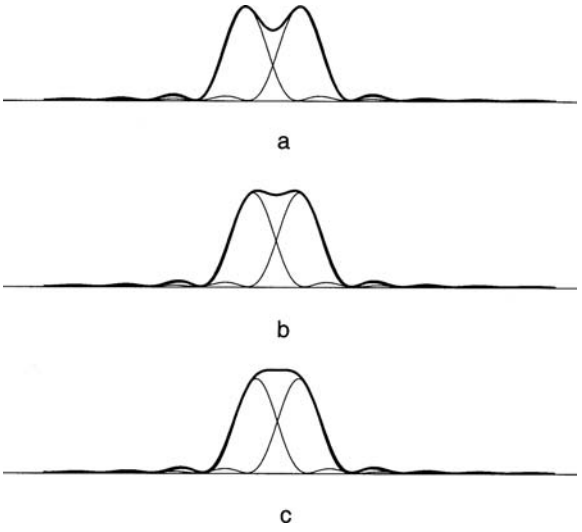


Fig. 1.5. Three resolution criteria: (**a**) Rayleigh, (**b**) Dawes, and (**c**) Sparrow.

just a tiny faint spot between two bright peaks which a human eye can just detect [Figure 1.5(b)]. These criteria are close to the cutoff spatial frequency in modern optical theory.

These three empirical criteria can also be applied to stars of different brightness. Figure 1.6 is the intensity diffraction pattern of two stars that meet the Rayleigh criterion, for which the magnitude ratio is 1.5. In this case, a faint spot

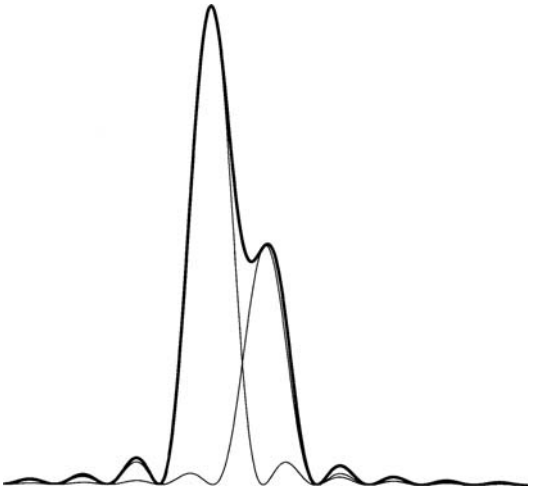


Fig. 1.6. Brightness distribution in case of the Rayleigh criterion when the stellar magnitude ratio is 1.5.

exists between two bright peaks. For a circular aperture, these three criteria are expressed as:

$$q_r = 1.22\frac{\lambda}{D}\,(\text{Rayleigh}) \tag{1.8}$$

$$q_d = 1.02\frac{\lambda}{D} \quad (\text{Dawes}) \tag{1.9}$$

$$q_s = 0.95\frac{\lambda}{D} \quad (\text{Sparrow}) \tag{1.10}$$

The angular resolution for apertures with blockage is better since the radius of the first dark ring is reduced.

The cut-off spatial frequency and Rayleigh criterion represent the diffraction limited resolution. The diffraction limit can be achieved by space telescopes, small ground-based telescopes, and ground-based telescopes with adaptive optics or interferometers. Another way to achieve diffraction-limited resolution when observing through the atmosphere is to use super-resolution (SR) techniques, most of which employ computer processing. One of these is to process many shifted short exposure frames of the same image. One type of image shift is caused by atmospheric turbulence, which is called the seeing effect. The seeing moves the image slightly and randomly in image plane. Another super-resolution technique moves the image in the longitude direction, called longitudinal super-resolution. In this technique, both in- and out- of focus images are collected as the out of focus image provides more spatial frequency information. Using template images of two point sources of different separations is also a super-resolution technique. Extrapolating in frequency domain by an assumed analytic function is another super-resolution technique.

However, super-resolution is largely a matter of data processing techniques. The ultimate image intensity distribution of a telescope is still set by the corresponding spatial frequency distribution of the aperture field.

Using the Rayleigh criterion, angular resolutions for different aperture sizes can be calculated. The 5 m Hale telescope of the Palomar Observatory should have an angular resolution of $\omega = 0.028$ arcsec at $\lambda = 550$ nm. In fact, this telescope has an angular resolution of about $\omega = 1$ arcsec. This resolution is not diffraction limited but is seeing limited because of the atmospheric turbulence.

The main origin of the atmospheric turbulence is wind. Wind has a wide frequency band and different wind frequencies will excite different scales of the turbulence. Within a single scale of the turbulence, the temperature is the same but the temperature is different when the scales are different. Temperature variations cause changes in the refractive index of the atmosphere, resulting in gradients of refractive index. The spatial scale of the wind frequency variation ranges from a few millimeters to a few hundred meters. Therefore, random differential or anomalous refraction occurs when stellar light passes through a turbulent atmosphere.

The effect of atmospheric turbulence can be described by two physical parameters, i.e., atmospheric seeing and atmospheric scintillation. Atmospheric seeing describes the random motion and spread of a stellar image caused by a telescope observing through the atmosphere. Atmospheric scintillation describes a sudden brightness change of the stellar image due to atmospheric turbulence.

Figure 1.7 shows the distortion of a plane wavefront of stellar light passing through the atmosphere. On a large scale, the wavefront distortion is significant but the orientation of the reference plane remains unchanged, while on a small scale, the wavefront distortion is small but the wavefront tilt is serious. That is to say, the star image observed by a large aperture telescope has a large image spread but the image position is stable. This is the atmospheric seeing effect.

For the image observed by a small aperture telescope, the turbulence will move in and out of the space above the telescope aperture rapidly. The slope of the wavefront and the path length (or focus/defocus) changes swiftly. The situation is opposite to the large aperture case. For a small telescope, the image is sharp but it jumps from one position to the other or produces rapid variation in apparent brightness or color. This is the atmospheric scintillation or twinkling effect.

If the turbulence is far away from the telescope aperture, the wavefront change will produce path length changes. Small scale, high atmosphere turbulence will have a large effect on scintillation, while large scale, low atmospheric turbulence, which includes near-ground turbulence and turbulence inside the dome, will have a large effect on seeing.

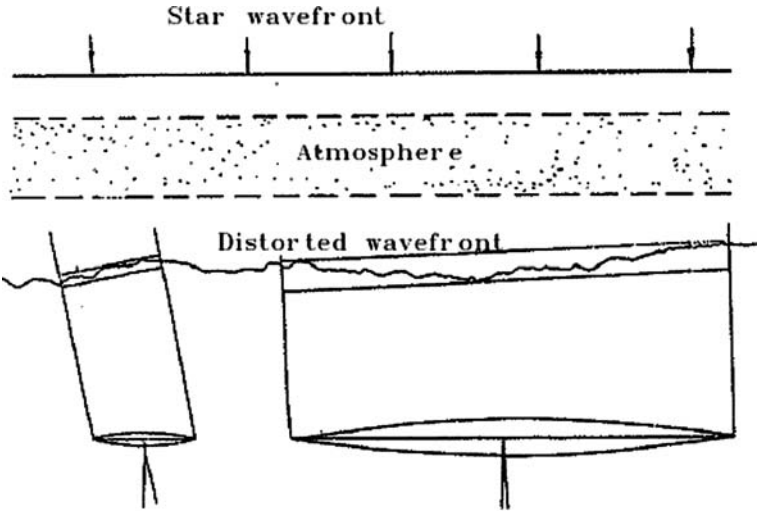


Fig. 1.7. Distortion of a stellar plane wavefront when it passes through the atmosphere.

Fig. 1.8. Change of the atmospheric seeing with the elevation angle.

The seeing is measured by the diameter of the image spread. Different locations on the earth have different seeing values. The seeing is a very important parameter for the observatory site selection. For a low altitude site, the atmospheric layer through which the stellar light passes is thick and the seeing is poor. Therefore, the site of an optical observatory is usually high above sea level.

The seeing of a good site is generally less than 0.5 arcsec while that of a poor site can be up to 3–5 arcsec (Figure 1.8). The atmospheric turbulence also depends on the direction of the star. When the incident angle of light is close to the horizon, the light path through the atmosphere is longer and the seeing will be poorer.

For a large diameter optical telescope, the seeing is the most serious limit to the angular resolution. The seeing is caused by the characteristics of the observing site and its dome. To improve the telescope angular resolution, a good site and improved dome and the mirror seeing (Sections 1.2.4 and 2.4.1) are important. However, to achieve diffraction limited angular resolution, adaptive optics should be used (Section 4.1).

1.2.2 Light Collecting Power and Limiting Star Magnitude

1.2.2.1 Light Collecting Power

The light collecting power, also known as the sensitivity, is a measure of the power of an optical telescope to detect faint objects. Telescopes with a large light-collecting capacity can detect very faint star light. In astronomy, the brightness of a celestial object is measured in magnitude. Early in the second century BC, Hipparchus divided the stars a human eye can detect into six

magnitudes. In the nineteenth century, astronomers re-examined the star magnitude using modern optical methods. The relationship between the magnitude m and the radiation intensity E is a logarithmic function:

$$m_1 - m_2 = 2.5 \log \frac{E_2}{E_1} \tag{1.11}$$

From the formula, large star magnitude means small radiation power density received on the earth. Only optical telescopes with large light collecting power can detect the faint radiation from a star of a larger magnitude.

In astronomical observations, the light collected by an optical telescope observing a celestial object can be expressed as:

$$N(t) = Q \cdot A \cdot t \cdot \Delta\lambda \cdot n_p \tag{1.12}$$

where A is the aperture area of the telescope, t the integration time, $\Delta\lambda$ the bandwidth, and n_p the number of photons per unit time, unit area and unit bandwidth from the celestial object which arrive on the earth. The function Q represents the combined quantum efficiency of the telescope and detector.

For stars with zero magnitude in the standard visible V band, the photon number which arrives above the atmosphere per square centimeter per second time is $\Delta\lambda \cdot n_p = 1007$. From the equation, it follows that the light-collecting power of a telescope is a linear function of the aperture area and the integration time. The integration time of an observation is usually limited. An essential way to improve the light-collecting power is to increase the aperture area of the telescope. Therefore, astronomers always want larger telescopes.

The function Q in Equation (1.12) includes the quantum efficiency of both the telescope and the detector. Compared with the human eye and photographic plate, modern Charge-Coupled Device (CCD) detectors have very high quantum efficiency. In this section, only factors which influence the quantum efficiency of the telescope are discussed.

For reflecting telescopes, the reflection loss is one of the main factors that influence the quantum efficiency of a telescope. Different mirror surface coatings have different reflecting loss. The reflecting loss of a fresh copper alloy mirror is about 45%. The loss increases as the surface becomes tarnished. The reflecting loss was a main reason for the low quantum efficiency of the early optical telescopes.

When a chemical silver coating is used on a glass mirror, the reflecting loss is only about 5% for a fresh coating (Figure 1.9). However, sulfur dioxide in the atmosphere will tarnish the silver surface very quickly and the reflecting efficiency is reduced.

Mirrors with vacuum aluminum coating are used today for almost all reflecting telescopes. The coating is made in a vacuum chamber with an air pressure of 0.007 pascal. The thickness of the coating is controlled to be half wavelength of yellow light. For this condition, the reflection loss is slightly lower than 10% in the visible band and is about 12% in the ultraviolet region of 250 nm. The

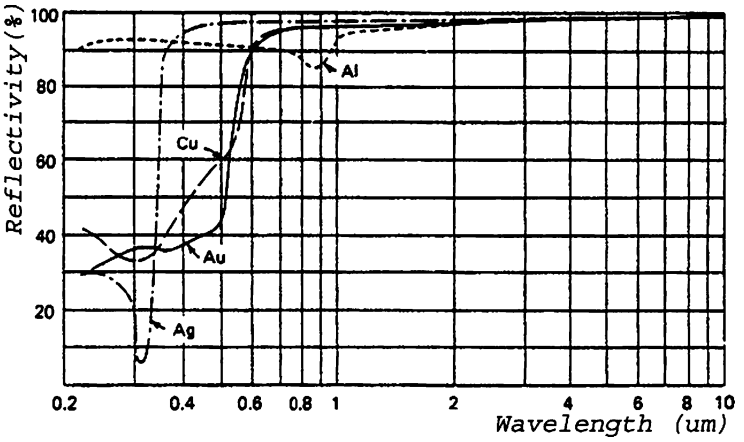


Fig. 1.9. Reflectivity curves of gold, silver, copper, and aluminum coatings.

reflection improves gradually in the infrared region; the reflection loss is only 9% at the wavelength of 1 μm. The loss is only 1% at the wavelength of 50 μm.

The aluminum coating on a mirror can oxidize when it is in contact with air. This will slightly increase the reflection loss. In the long term, dust and chemical reaction with air will increase the reflection loss to about 15%. For maintaining a low reflection loss, aluminum-coated mirrors should be washed every three to six months and recoated every one to two years.

For dual-reflector telescopes, the reflecting efficiency is the square of that for a single mirror telescope. The overall reflection loss increases. At the Coude focus, the light is reflected by three or more mirrors, the reflection loss may reach 60% or more for aluminum mirror coating. In this case, high-efficiency coating should be used. In ultraviolet and far ultraviolet regions, the reflectivity of most coatings decreases rapidly. Only gold, platinum, and indium coatings remain highly reflective. The wavelength used for these coatings can be extended to between 2.3 and 19 nm.

Generally, if a coating material absorbs visible light, the complex refractive index of the material is:

$$\tilde{n} = n - ik \tag{1.13}$$

where k is the absorption index and n the refraction index. From this formula, the reflection efficiency of a thick coating is:

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \tag{1.14}$$

For improving the reflection efficiency, a thin multi-layer dielectric interference coating can be used. If a periodic multi-layer dielectric coating consists of

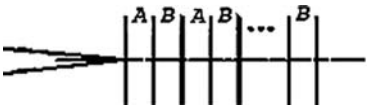


Fig. 1.10. A multi-layer dielectric interference coating.

N layers of material A and B, then the highest reflectivity is obtained when the optical thickness of the coating equals an odd number of a quarter of wavelength (Figure 1.10). In this case, if the refractive index of the surface is n, the reflectivity is:

$$R = \left[\frac{1/n - (n_A - n_B)^{2N}}{1/n + (n_A - n_B)^{2N}}\right]^2 \tag{1.15}$$

Figure 1.11 shows the reflectivity curves of some multi-layer dielectric interference coatings. The reflectivity of these coatings is more than 95% over wide bandwidths. In the figure, some multi-layer dielectric interference coatings with very low reflectivity are also shown. These coatings can be used on lens surfaces to improve the transmission of the visible light. When lenses are introduced in a telescope system, refracting loss occurs on the surface between the lens and air. The transmission efficiency between two transparent media is:

$$R = [(n' - n)/(n' + n)]^2 \tag{1.16}$$

where n and n' are the refractive indexes of the first and second material. For glass with $n = 1.74$, the loss due to reflection is about 7.3%. When $n = 1.51$, the loss is about 4.1%.

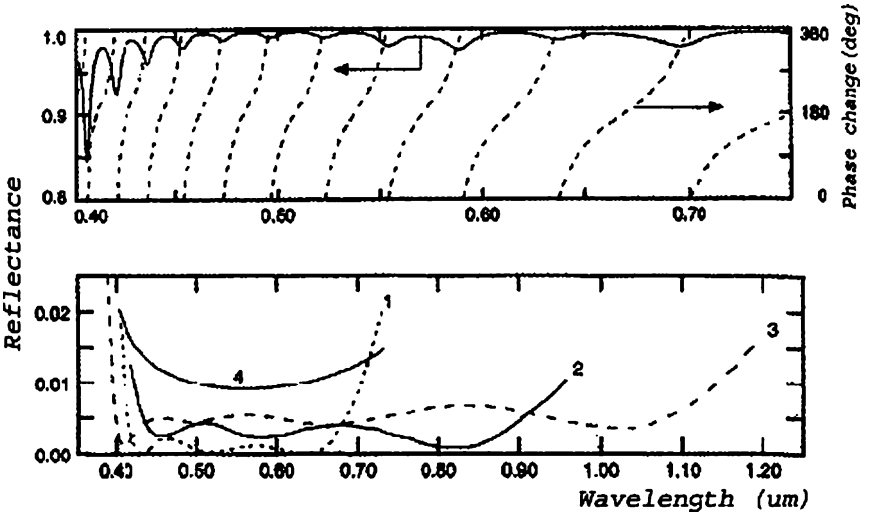


Fig. 1.11. Reflectivity curves of some multi-layer dielectric interference coatings.

Since the refraction index of a glass is different for different wavelengths, the transmission loss is slightly higher for blue light than for red light. If the lens surface is coated with MgF ($n = 1.38$) of a quarter wave optical path length, the transmission loss can be reduced to 1.3–1.9%. The transmission efficiency of a multi-layer dielectric interference coating can reach 99.7%. For one or two narrow bands, the transmission efficiency of a multi-layer dielectric coating can reach nearly 100%.

Besides the reflection and refraction losses, the quantum efficiency can also be affected by absorption of the lens material. If the absorption coefficient of the material is τ, the transmissivity of the medium is:

$$T = e^{-\tau \cdot l} \tag{1.17}$$

where l is the length of the medium. The reciprocal of the transmissivity is the degree of transparency. The logarithm of the degree of transparency is known as the optical density.

1.2.2.2 Limiting Star Magnitude

If the aperture size and quantum efficiency of a telescope-detector system are known, its light-collecting power is then determined. Astronomers are usually more interested in another parameter: the penetrating power of a telescope. The penetrating power, or the limiting star magnitude the telescope can reach, is closely related to the observing mode. Different observing modes have different penetrating power.

The following discussion on penetrating power is based on past discussions of Baum (1962), Bowen (1964), Bahner (1968), and Disney (1972, 1978). Parameters used in the discussion are all listed in Table 1.3. For simplicity, the star

Table 1.3. Parameters used in the discussion of the limiting star magnitude

D	aperture diameter of the telescope	f	telescope focal length
t	exposure time	g	line number per millimeter
n_p	photon number arrived on the earth per unit time, unit area, and unit bandwidth from a star	f_c	camera focal length of the spectroscope
		F_c	camera focal ratio of the spectroscope
S	photon number arrived on the earth per unit time, unit angular area from sky background	d	projected diameter of the grating
		W	grating width
$\Delta\lambda$	spectrum resolution or the spectrum bandwidth	m	optimal photon number per unit area of the photographic plate
Q	combined telescope-detector efficiency	O	order of the grating used
p	pixel dimension	ω	width of the entrance slit
β	pixel field angle	B	relative signal to noise ratio
F	focal ratio of the telescope	e	the efficiency of entrance slit

energy distribution in the discussions is assumed as a square instead of as a circle, so that a factor of $\pi/4$ is omitted. The whole discussion is divided into three related parts: (a) photometry, (b) photography or CCD observation, and (c) spectroscopy.

(a) Photometry

When a telescope is used in photometric mode, the photon number detected from a star by a telescope is a function of the telescope quantum efficiency, aperture size, integration time, spectrum bandwidth, and the photon number per unit area, unit time, and unit bandwidth arrived from a star:

$$N_0(t) = QD^2t \cdot \Delta\lambda \cdot n_p \tag{1.18}$$

Since the signals detected from any astronomical object are random variables with a Poisson distribution, the detecting error or noise is:

$$\delta N_0(t) = \sqrt{N_0(t)} = \sqrt{QD^2t \cdot \Delta\lambda \cdot n_p} \tag{1.19}$$

In the same time, the photon number received from the sky background is:

$$N_s(t) = \beta^2 S \cdot QD^2t \cdot \Delta\lambda \tag{1.20}$$

where β is the pixel field angle and S the photon number per unit time, unit angle, and unit area of the sky background. The sky background noise is:

$$\delta N_s(t) = \sqrt{N_s(t)} = \sqrt{\beta^2 S \cdot QD^2t \cdot \Delta\lambda} \tag{1.21}$$

If B is the tolerable signal to noise ratio of the observation, then:

$$B = \delta N/N_0 = (\delta N_0 + \delta N_s)/N_0 \tag{1.22}$$

There exist two situations in the observation. The first is the star brightness limited where the star brightness is much greater than that of the sky background ($n_p >> \beta^2 S$). In this case, the limiting star magnitude (or the limiting star brightness) detected is:

$$1/n_p = B^2 D^2 t \cdot Q \cdot \Delta\lambda \rightarrow D^2 t \tag{1.23}$$

The second is the sky background limited where the sky background brightness is far greater than that of the star. Then we have:

$$1/n_p = B \cdot D \cdot t^{1/2} \cdot Q^{1/2} \cdot \Delta\lambda^{1/2}/(\beta \cdot S^{1/2}) \rightarrow D \cdot t^{1/2} \tag{1.24}$$

Equations (1.23) and (1.24) show that the limiting star magnitude in photometric mode is proportional to the square of the aperture diameter when the star brightness is dominant and, for faint star observations, is proportional to the diameter only when the sky background is dominant. In the latter case, the increase of the penetrating power is slower than the increase of the light-collecting power for faint star observation in photometric mode.

Affected by the atmospheric seeing, the diameter of the entrance pupil is usually set to be larger, such as $\beta = 10$ arcsec, during the photometric observation. Under this assumed condition, the transition of penetrating power from D^2 to D occurs at the magnitude of 16.5. For fainter stellar photometric work, the penetrating power of the telescope is proportional to the aperture diameter only. Photometry can be done using CCD chips. However, the basic formulation remains unchanged when the CCD detectors are used.

(b) Photography and CCD observation

Photography and the CCD are used in imaging observations. The detecting principle is, in fact, identical as in the photometry so that Equations (1.23) and (1.24) listed above are also valid in the photography or CCD imaging observations. Since the images in this case are much sharper than those in photometric mode, the angular size used will be different. If the observation is seeing limited, then β may be chosen as about 1 arcsec (note, some of the best sites have a medium FWHM around 0.6 arcsec). Then the transition of the penetrating power from D^2 to D is at about the magnitude of 20 for photography or CCD imaging depending on the telescope-detector quantum efficiency, where the atmospheric seeing matches the image size of the CCD and the comparison of the transition of the penetrating power is made for the same integration time. In practice, the exposure time of the CCD or photographic plate is limited due to the dynamic range of detector response. Therefore, an ideal exposure time exists; this is especially true for the photographic work. For a photographic plate, if the exposure is too long or too short, the plate efficiency decreases rapidly. In both cases, the contrast of the faint stellar images will decrease. If the photon number per unit area of a photographic plate to reach its ideal exposure is m, then:

$$mf^2 = QD^2 t_0 \Delta\lambda \cdot S \tag{1.25}$$

where t_0 is the ideal exposure time. It is:

$$t_0 = mf^2/(QD^2\Delta\lambda \cdot S) \tag{1.26}$$

where f is the telescope focal length. Inserting t_0 into Equation (1.24), the limiting star magnitude reached during an ideal exposure condition, $1/n_p$, is:

$$1/n_p = B \cdot m^{1/2} f/(\beta \cdot S) \to f \tag{1.27}$$

This equation shows that the limiting star magnitude of a photographic plate depends on focal length, but not on the aperture diameter of the telescope. However, if the aperture is too small, the required exposure time may be too long to be realized as $t_0 \propto D^{-2}$.

For small focal length of the pixel size limited case, the star image diameter is $\beta = p/f$. Then:

$$1/n_p = B \cdot m^{1/2} f^2/(p \cdot S) \to f^2 \tag{1.28}$$

The limiting star magnitude depends on focal length squared, but not on the aperture size. Following the above two equations, if $\beta = 1.25$ arcsec and $p = 0.18$ mm, the relationship between the limiting star magnitude and the focal length are as shown in Figure 1.12. The change of slope is at about $f = 3$ m. The slope is about 5 for smaller focal length and is about 2.5 for larger focal length.

(c) Spectroscopy

The general arrangement of a spectrograph (spectroscope) is shown in Figure 1.13. In the figure, the collimator (a lens or mirror with a light source

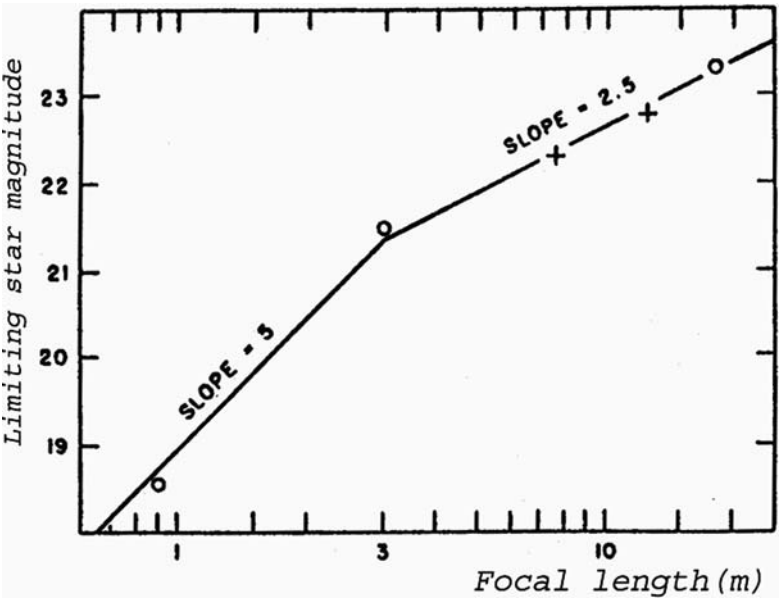


Fig. 1.12. The relationship between the limiting star magnitude and the focal length for photography and CCD observations when the seeing is about 1 arcsec.

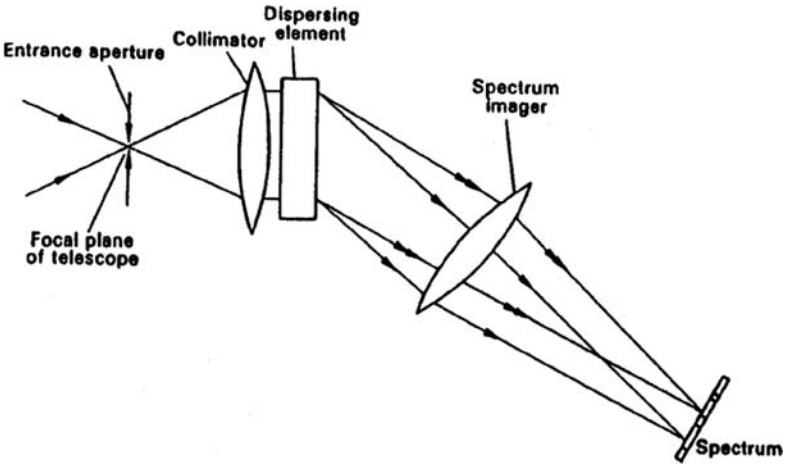


Fig. 1.13. General arrangement of a spectrograph.

at its focus) and camera (an image-forming device) are lenses and the diffraction component is a grating. A diffraction grating is an optical component with a surface covered by a regular pattern of parallel lines, typically separated by a distance comparable to the wavelength of light. Light rays that pass through such a surface are bent as a result of diffraction and this diffraction angle depends on the wavelength of the light. For a celestial object with an angular diameter β, its linear dimension on the detector is $DF_c\beta$, where F_c is the camera focal ratio. To ensure the spectral resolution (the power to resolve features in electromagnetic spectrum) of the spectroscopy, it should have (Figure 1.14):

$$f_c\Delta\theta \geq D \cdot F_c\beta \tag{1.29}$$

When $f_c\Delta\theta < DF_c\beta$, the images will overlap each other (Figure 1.14). In this situation, it is necessary to adjust the entrance width to $W' = f_c\Delta\theta$, so that both sides of the images are cut out. The efficiency of the spectrograph is reduced. To satisfy the condition set by Equation (1.29), a larger grating and collimator

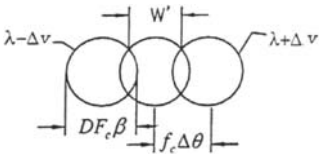


Fig. 1.14. The image positions in the spectrum after diffraction.

are required; however, these are usually difficult. Therefore, two cases exist: (a) nongrating-limited case and (b) grating-limited case.

In the grating limited case, the entrance width is reduced and the spectrograph efficiency decreases by a factor of e:

$$e = f_c \Delta\theta / (D \cdot F_c \beta) = (W/D)(\Delta v/\beta) O_g \tag{1.30}$$

where W and O_g are grating parameters defined in Table 1.3 and Δv is the spectral resolution, not the waveband. Under this situation, the limiting star magnitude for spectrograph work is almost the same as that in photometric mode. However, Equation (1.18) should be modified into:

$$N_0(t) = eQ \cdot D^2 t \cdot \Delta v \cdot n_p \tag{1.31}$$

Equation (1.20) becomes:

$$N_s(t) = e\beta^2 SQ \cdot D^2 t \cdot \Delta v \tag{1.32}$$

For the star brightness limited case, the limiting star magnitude is:

$$1/n_p = \left[B^2 QDt \cdot (\Delta v)^2\right](W \cdot O_g) \sim Dt \tag{1.33}$$

For the sky background limited case, the limiting star magnitude is:

$$1/n_p = B \cdot m^{1/2} f/(\beta \cdot S) \to f \tag{1.34}$$

If the grating size is limited, an image slicer is often used for avoiding the star light loss caused by the width of the entrance aperture. The limiting star magnitude is unchanged. An image slicer is a device to split the star image into small dimensional patches so that these patches can be rearranged along the slit of the spectroscope without loss of the star light. A group of optical fibers can be used as an image slicer.

From the above analysis, the grating parameter WO_g has a great influence on the spectrograph efficiency. Therefore, it is important to improve the grating parameters in order to increase the spectrograph efficiency and to increase the limiting star magnitude. To improve WO_g, one has to increase the grating dimension W. Since $W \sim d$, it is also necessary to increase the collimator diameter, which is difficult in practice.

Figure 1.15 shows the relationship between the spectrum resolution $\Delta\lambda$ and β for different telescope diameters when the grating parameter is $WO_g = 100$. The lines of seeing limit and the grating limit divide the $(\beta, \Delta\lambda)$ plane. It produces a special area (small triangular shaded area) of high efficiency spectroscopy. This area starts from an aperture of 3 m and ends at an aperture of 5 m. In order to achieve an even higher spectral resolution, the optimal aperture size is between

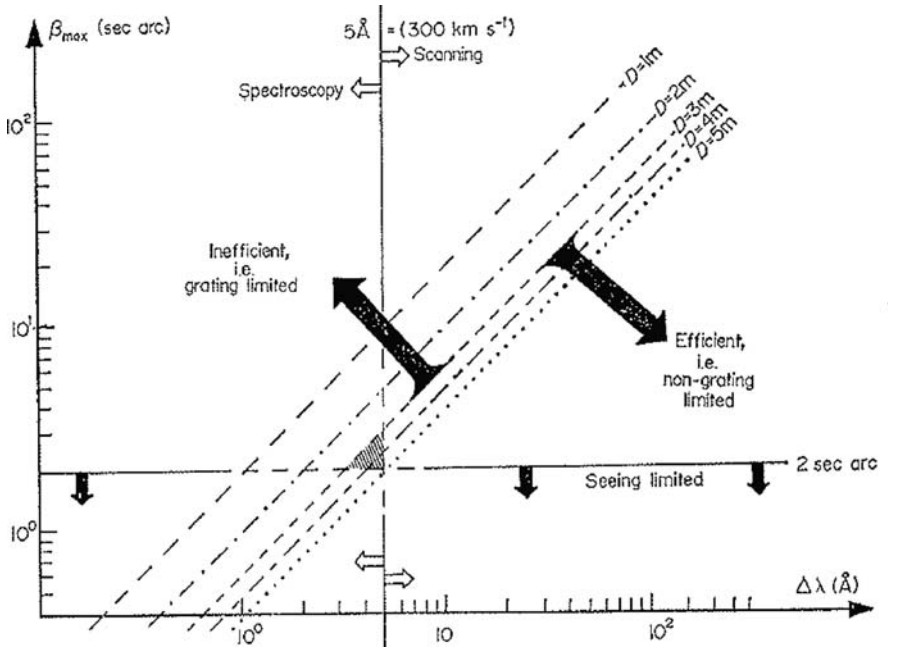


Fig. 1.15. The relationship between limiting star magnitude, telescope diameter, and spectrum resolution (Disney, 1972).

2 and 3 m. This shows that the limiting star magnitude and the spectral resolution increases are far less than the increase of the light-collecting power.

The discussion so far does not include the case of both diffraction- and detector-noise-limited observations. Table 1.4 shows the relationship between the penetrating power, the telescope aperture diameter, and the integration time for all nine cases of the astronomical observations. In the diffraction-limited case, the penetrating power is proportional to high power of the aperture area.

Table 1.4. The relationship between limiting star magnitude, aperture diameter, and the integration time

Noise source	Effective diameter of the source star		
	Seeing limited	Grating limited	Diffraction limited
Star limited	$\propto D^2t$	$\propto Dt$	$\propto D^4t$ or $\propto D^2t$ (space or adaptive telescope)
Sky limited	$\propto Dt^{1/2}$	$\propto D^{1/2}t^{1/2}$	$\propto D^3t^{1/2}$ (space telescope)
Detector limited	$\propto D^2t^{1/2}$	$\propto Dt^{1/2}$	$\propto D^2t^{1/2}$ (radio telescope)

This applies to all space telescopes, ground telescopes with adaptive optics, and very small aperture (smaller than 10 cm) optical telescopes. In optical region, the detector noise is not an issue, so only four cases exist. These are: (I) star brightness limited case: $\propto D^4 t$ for point sources and $\propto D^2 t$ for extended targets; (II) sky background limited case: $\propto Dt^{1/2}$; (III) grating limited case: $\propto Dt$; and (IV) grating and sky background limited case: $\propto D^{1/2} t^{1/2}$.

In the above four cases, case I includes photometric observation of stars brighter than magnitude 16.5, photography or CCD observations of stars brighter than magnitude 20, and low-resolution spectroscopy and spectral surveys of stars brighter than magnitude 20. Case II includes photography or CCD observations of stars fainter than magnitude 20, low-resolution spectroscopic work, photometry and low-resolution spectroscopic surveys of stars fainter than magnitude 16.5. Case III includes medium or high-resolution spectroscopy of stars brighter than magnitude 20 and low-resolution spectroscopy of bright emission lines of faint galaxies. Case IV includes spectroscopic work of faint celestial objects.

Table 1.4 shows that the increase of penetrating power of ground optical telescopes is not always proportional to the increase of the light-collecting power. The penetrating power increases very fast for space or adaptive optics telescopes on point sources. For a given light-collecting power telescope, there is no clear cutoff limiting star magnitude for all observations. The star magnitude reached can be increased by increasing the exposure time. This is why so many Hubble Space Telescope pictures are produced by hundreds of hours of multi-exposure measurements with modern star guiding techniques. For some types of spectroscopic, photographic or CCD work, medium aperture telescopes are still efficient, while large aperture telescopes show many restrictions requiring large collimator, large grating, and adaptive optics.

1.2.3 Field of View and Combined Efficiency

For collecting more photons, a telescope needs a large light-collecting power, or larger aperture size. However, optical telescopes are also imaging instruments. For an imaging device, the field of view is important when the information gathering ability is concerned. With a large field of view, a telescope can gather more information in a single exposure and, therefore, the telescope has higher information delivery efficiency.

Generally, the image blur size is the main factor which limits the field of view of a telescope. However, the observing mode, the detector type, the field vignetting, the atmospheric seeing, and the differential atmospheric refraction also have an influence on the field of view.

The geometrical aberrations of an optical system determine the image blur size besides the atmospheric seeing effect. For a paraboloid prime focus system, coma limits the field of view. The coma size equals $3\Phi/(16F^2)$, where Φ is the half field angle and F the focal ratio. Figure 1.16 is the relationship between coma,

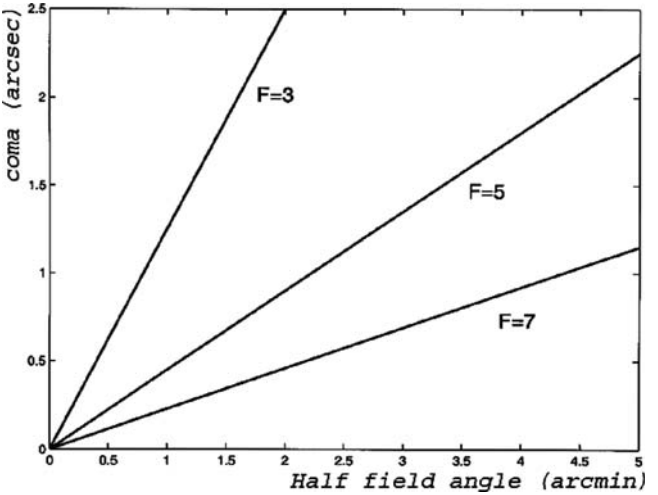


Fig. 1.16. The relationship between coma, half field angle, and focal ratio for a single paraboloidal reflector.

the half field angle, and the focal ratio for a prime focus system. Without adaptive optics, the allowable image size is about 1 arcsec. Therefore, the effective field of view of this system is only a few arcmin.

A classical Cassegrain system has the same coma as the prime focus one. The field of view is also limited. An R-C optical system (refer to Section 1.3.1) has no coma and therefore can have a slightly larger field of view, up to 30 arcmin for typical focal ratios used. Field correctors increase the field of view. The field of view can reach 1° to 2° with a field corrector. A three-mirror system can provide a field of view of about 5°. A Schmidt telescope has the largest usable field of view of about 6°.

Different observation modes require different fields of view. Photometry with single pixel detectors uses a very small field of view since only one star is observed. Photometry with a CCD has a larger field of view. The photographic or CCD observations have a larger usable field of view determined by the size of the plate or CCD. In spectroscopic work, the usable field of view is small except when an objective prism or multi-object fibers are used. An objective prism is a thin prism placed on the entrance pupil of a telescope. The multi-object fibers involve movable fiber heads on the star image plane. The fibers capture the light from the objects in the field and then are rearranged on a slit opening of a spectrograph. Because the length of the entrance slit is limited by d/D, a ratio between grating size and telescope diameter, the number of stars observed is limited. Nevertheless, this improves the spectroscope efficiency.

The field of view for normal spectroscopic work is limited by the size of the diffraction component. When the diffraction component is a grating, the

relationship between the grating dimension d and maximum usable field of view Φ is:

$$\Phi = (d - d_0)/(FD) \tag{1.35}$$

where $d_0 = L/F$ is the dimension of the diffraction component needed for a zero field of view, L the distance between the focal plane and the diffraction component, and F and D the focal ratio and diameter of the telescope.

Field vignetting is another problem in wide field instruments. If the half field angle is θ, the ratio between the projected aperture area at this angle and that at the image center is a measure of the vignetting effect. The star magnitude reduction caused by this factor can be expressed as:

$$\Delta m = -2.5 \log[A(\theta)/A(0)] \tag{1.36}$$

where $A(\theta)$ is the aperture area at the field angle θ and $A(0)$ is the on-axis aperture area. With this formula, the star magnitude can be calibrated. Figure 1.17 shows the relationship between the field angle and the star magnitude reduction rate for the UK 1.2 m Schmidt telescope. In Schmidt telescopes, the primary mirror diameter is always larger than the aperture diameter; however, vignetting still exists at large field angles.

The differential atmospheric refraction is another factor limiting the usable field of view. The differential atmospheric refraction is derived from the elevation angle dependent atmospheric refractive index. Because of this, the scale within a large field of view is not the same, so that the image blur size on the edge increases. This effect is largest near the horizon. For a long

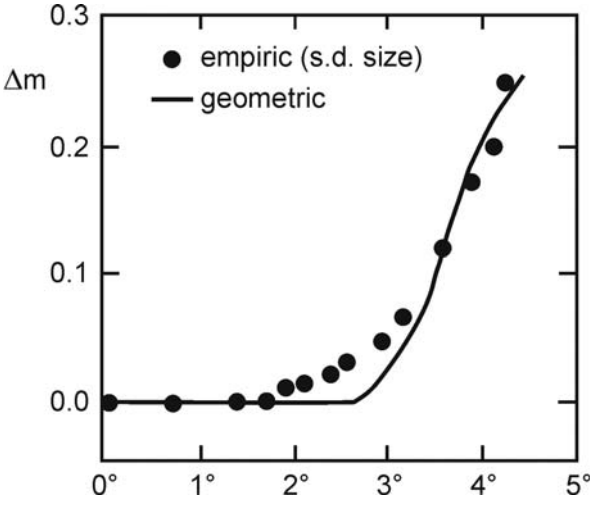


Fig. 1.17. The reduction of star magnitude as a function of the half field angle for the UK 1.2 m Schmidt telescope (Dawe, 1984).

exposure, a special field corrector (lensm) for the differential atmospheric refraction is necessary (Section 1.3.4).

The overall telescope efficiency is a complex parameter. Different observing modes and different objects observed require different evaluation methods for this parameter. However, a widely used parameter in optics is etendue. The etendue specifies the throughput or capacity to transmit radiation for an optical instrument. It is the product of the aperture area and the field of view expressed as $E \sim D^2\Phi^2$.

The overall efficiency should include the penetrating power, which is expressed as $E_f \sim D^{\alpha}\Phi^{\beta}$ (α and β are determined by the observation mode and other conditions as discussed in early sections), the angular resolution, the spectral range, and the field of view. All these are heavily influenced by atmospheric absorption and seeing. In the next section, the site selection and atmospheric seeing are discussed briefly. More discussion on seeing is in Section 4.1.6. Over all the factors which influence telescope efficiency, the astronomers in charge are the most important one for the output of a telescope. With sufficient knowledge and full understanding of the celestial objects observed, even medium or small size telescopes could produce the most important scientific results.

1.2.4 Atmospheric Windows and Site Selection

The radiation from celestial objects covers nearly the whole spectrum of electromagnetic waves. However, there are only two transparent windows due to the absorption and scattering from various particles of the earth's atmosphere. These windows are the optical one at visible wavelengths and the radio one at radio wavelengths (Figure 1.18). The optical window is from 300 to 700 nm wavelengths. Within this band, scattering from the atmosphere is small and the transmission efficiency is very high. When the wavelength is smaller than 300 nm, the radiation is seriously absorbed by oxygen atoms, oxygen molecules, and ozone. The long wavelength part of the radio window is stopped by the ionosphere, between 100 and 50 km above sea level. The radio window extends into millimeter and submillimeter regions. At wavelengths around 1 μm, the

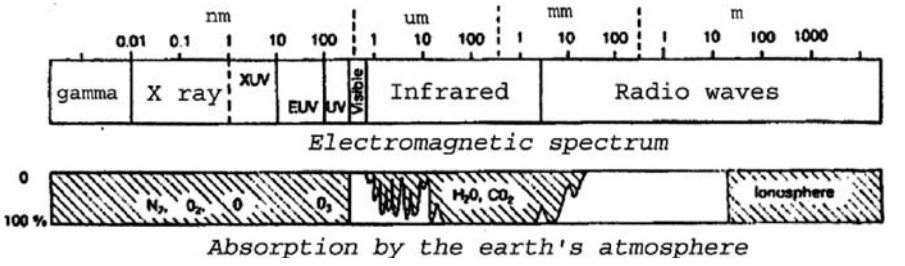


Fig. 1.18. The spectrum of electromagnetic waves and the atmospheric absorption.

major infrared absorption is by water molecules. In the infrared region, water vapor, carbon dioxide, and ozone produce a series of absorption bands, leaving a number of very narrow infrared windows. These nearly transparent windows are in the region of 8–13 μm, 17–22 μm, and 24.5–42 μm. When the observatory altitude increases to about 3,500 m, some windows in the far infrared region can be used as the water vapor content in the atmosphere is reduced. These narrow windows are connected to the submillimeter windows in the radio region. The details of the infrared and radio windows will be discussed in latter chapters.

The transparency of the optical window provides an important observing condition for optical astronomy. However, atmospheric turbulence as well as human activity makes different observational conditions for different sites on the earth. In the beginning of modern astronomy development, the telescope sites were near large cities. Gradually, telescopes were moved to far away high mountain sites. Serious site selection activities were started in the early 1950s in order to get the best results from optical telescopes.

Many factors influence the site selection. Generally, these factors fall within the following categories: (a) factors related to the site atmospheric conditions. These include the number of clear nights per year with no or few clouds, seeing, scintillation, rain, snow, wind, and atmospheric attenuation; (b) factors related to site natural conditions. These include site altitude, latitude, topology, temperature variation, sand storm, dust condition, and earthquake activity; (c) factors related to human activities. These include sky brightness, city light, and atmospheric pollution; and (d) logistic factors. These include water and power supply, road condition, and living facilities.

In general, site selection for large telescopes requires a long-term data collection and analysis. Among these data, the information of seeing measurement and investigation is the most important. Usually, the telescope diameter to seeing ratio is an important indication of the telescope efficiency. Roddier (1984) pointed out that the signal-to-noise ratio achieved in the high-resolution interferometer observations is also proportional to the diameter to seeing ratio. A site with good seeing is essential for astronomical observations.

The quantitative description of seeing is the Fried expression. When the star light passes through the atmosphere with variable refraction index, the maximum telescope dimension not being affected by seeing, or the correlation length of the refraction index in the atmosphere, r_0, is expressed as the Fried length or Fried parameter:

$$r_0 = \left[1.67\lambda^{-2} \cdot \cos^{-1}\gamma \int_0^L C_N^2(h)dh \right]^{-3/5} \tag{1.37}$$

where λ is the wavelength, γ the zenith distance, $\boldsymbol{h}$ the altitude, L the optical path length, and C_N the refraction index structure constant. The index structure

constant C_N is determined by the refractive index structure function which is used to describe the property of a nonstationary random variable:

$$D_N(\vec{\rho}) = <|n(\vec{r}) - n(\vec{r}+\vec{\rho})|^2> = C_N^2|\vec{\rho}|^{2/3} \tag{1.38}$$

where $n(\vec{r})$ is the refractive index spatial distribution and $\vec{\rho}$ a directional vector between two nearby points. In the optical region, the index structure constant C_N^2 is closely related to the temperature structure constant C_T^2:

$$C_N = 80 \times 10^{-6} P C_T / T^2 \tag{1.39}$$

where P is the atmospheric pressure in mbar and T the absolute temperature. The temperature structure constant is also related to the temperature structure function:

$$D_T(\vec{\rho}) = <|T(\vec{r}) - T(\vec{r}+\vec{\rho})|^2> = C_T^2|\vec{\rho}|^{2/3} \tag{1.40}$$

From these relationships, the seeing can be derived through the measurement of the temperature structure function while the temperature structure function can be measured by thermal sensors or echo sound radar.

The measured site index structure constant of the Canada-French-Hawaii observatory is $C_N = 2 \cdot 10^{-15}$ ($h = 0$ m) and $C_N = 1 \cdot 10^{-17}$ ($h = 100 \sim 150$ m). The Fried length of this site is $r_0 = 0.4''$. The measured temperature structure constant of the site is ${C_T}^2 = 0.1^{\circ}\mathrm{C}^{-2/3}$ ($h = 10$ m) and $C_T = 0.02^{\circ}\mathrm{C}^{-1/3}$ ($h = 50$ m). The calculated medium Fried parameter is $r_0 = 0.3''$.

It follows from these numbers that the contribution to the seeing of this site by the lower atmospheric layers between 10 and 150 m is about 0.3–0.4 arcsec. The seeing including the upper layer contribution is about 0.35–0.45 arcsec. If the dome seeing of 0.25 arcsec is added, the total medium seeing is about 0.7 arcsec.

Figure 1.19 shows the measured seeing statistic of this telescope site. The expectation of the seeing effect in observation is 0.75 arcsec, a number which is very close to that derived from the seeing measurement.

Further study shows that the index constant is related to the wind velocity at the upper level where the pressure is 200 mbar. Therefore, this wind velocity number is also related to atmospheric seeing number. Figure 1.20 shows this relationship. Using this relationship, one can also use the meteorological data for the site selection purpose.

Experiences show that the best astronomical sites are located at high mountains of coastal regions or isolated islands where a cold sea current from the west is dominant. At these sites, the air flow is smooth and the correlation length is long. The mountain sites are high above clouds. The number of clear nights without clouds per year is high. The water vapor content is also lower, so that the absorption and attenuation due to the atmosphere is small.

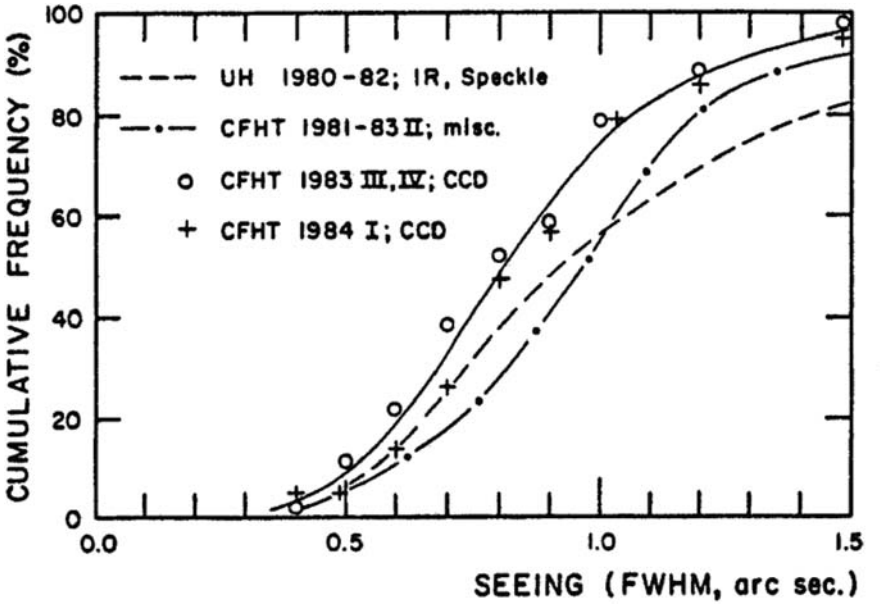


Fig. 1.19. The seeing statistic distribution measured at Hawaii French-Canada-Hawaii telescope site (Racine, 1984).

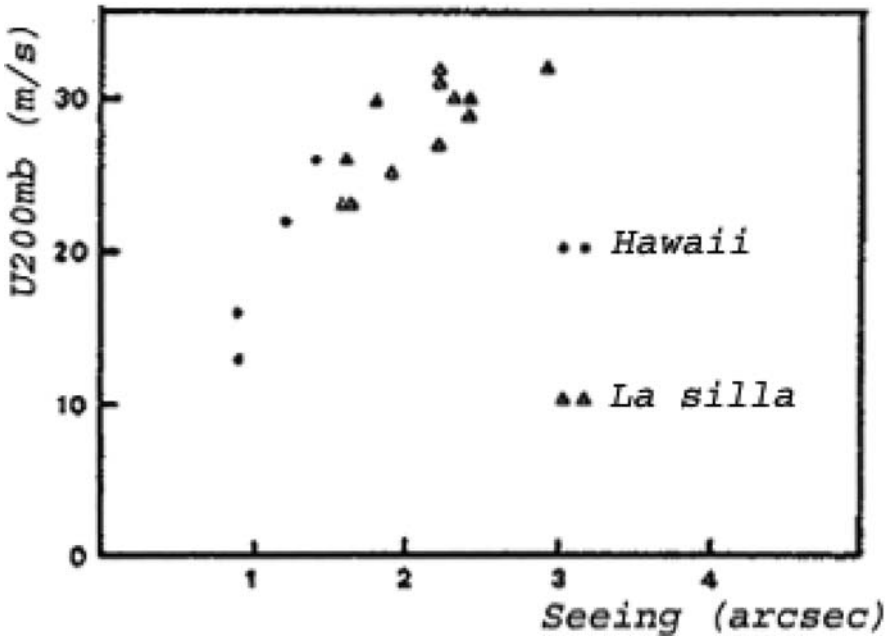


Fig. 1.20. The relationship between the seeing and the wind velocity at the 200 mbar level in Hawaii and in La Silla (Vernin, 1986).

A good observatory site has to be away from human activity to avoid bright sky background polluted by the city illumination. Today the best optical sites are Hawaii, northern Chile, and the Canary Islands. However, getting approval to build new telescopes in Hawaii has become difficult because of environmental or other issues, so that northern Chile will become a major site in the near future for optical, infrared, and millimeter wavelength telescopes. However, the South Pole may be one of the best observation sites on the earth as it has the lowest water vapor content and lowest wind velocity (Section 6.3). The only problems of the South Pole are severe weather conditions, strong ground-layer seeing, and the difficulties in access. Along with the development of extremely large optical telescopes, even higher and more advantageous sites are under investigation.

1.3 Fundamentals of Astronomical Optics

1.3.1 Optical Systems for Astronomical Telescopes

Limited by the availability of large transparent homogeneous glass material, the absorption and scattering, and the shape deformation under edge support, modern astronomical optical telescopes are mainly made of mirrors, not lenses. A few medium-size Schmidt telescopes use catadioptric systems, containing both mirrors and lenses. The largest aperture of these telescopes is only about 1.2 m.

The optical systems used for optical telescopes include the prime focus, Newtonian focus, Cassegrain focus, Nasmyth focus, and Coude focus systems. These (focus) systems have different focal ratio, focal position, and aberrations. In telescope design, other optical systems, such as Schmidt system, three-mirror system, and folding optics, are also used. This section provides an overview of these systems.

1.3.1.1 Prime Focus and Newtonian Focus Systems

The prime focus system is a system with only one mirror [Figure 1.21(a)]. When a paraboloidal primary mirror is used, light beams parallel to the axis will be reflected to a single focal point. In geometrical optics, an on-axis point star will form a sharp image. Traditionally the focal ratio of a prime focus system is in the range of $F/1$ to $F/5$. Small size telescopes use a larger focal ratio. Large size telescopes use a smaller focal ratio because of practical constraints of structural flexure and cost. Modern large telescopes use even smaller focal ratio from $F/1$ to $F/0.5$.

The prime focus of a paraboloidal mirror has no spherical aberration. Coma is the main off-axial aberration of the system. The astigmatism and field curvature, being proportional to the square of the half field angle, are less serious for a small field of view. The usable field of view of a prime focus system is limited, so correctors are needed for wide field observations. The prime focus is located within the light path, so that only simple small detectors are used. Since the prime focus

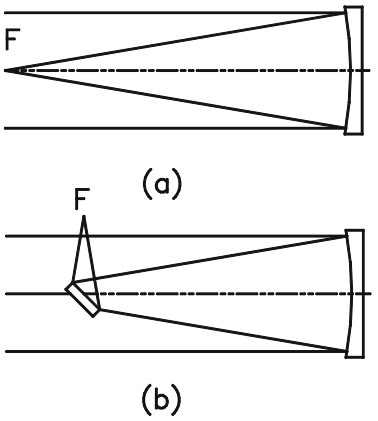


Fig. 1.21. (**a**) Prime focus and (**b**) Newtonian focus.

involves a single reflecting surface, the light loss due to absorption is small. The prime focus of a hyperboloid mirror is also used for some Ritchey–Chretien (R-C) telescopes. There are also a few prime focus telescopes using a spherical primary. Specially designed correctors are required for this type of prime focus system.

A Newtonian focus system is the same as the prime focus, except an inclined mirror reflects the beam to one side of the tube [Figure 1.21(b)]. The Newtonian system has the same properties as the prime focus system, except an added light loss from the inclined mirror. The Newtonian focus is easier to access. The focal stations can be changed by rotating the inclined mirror. However, this focus is in front of the tube, so that the system is not popular in large astronomical telescopes.

1.3.1.2 Cassegrain and Nasmyth Focus Systems

If a secondary mirror is inserted before the prime focus, the optical system becomes a dual reflector or Cassegrain one. The new focus formed is the Cassegrain focus. The Cassegrain focus is usually located behind the primary mirror [Figure 1.22(a)].

In a Cassegrain system, the secondary mirror directs light away from the incident beam, so that it is suitable for large size instruments. The access of the focus is also easy. The Cassegrain focus is very important for astronomical observations. The focal ratio of this system typically ranges from $F/7$ to $F/15$ due to the magnification of the secondary mirror.

Depending on the mirror surface shape, dual-reflector telescopes include classical Cassegrain, Gregorian, R-C, and quasi-R-C systems. A classical Cassegrain system has a paraboloidal primary mirror and a hyperboloid secondary mirror. The two foci of the hyperboloid mirror coincides with the prime focus and the Cassegrain focus. The path lengths of all on-axis rays to the Cassegrain

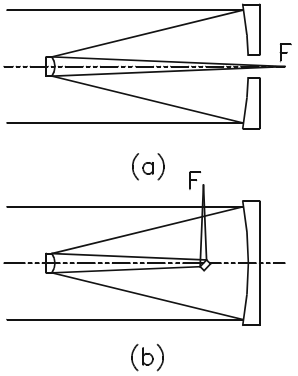


Fig. 1.22. (**a**) Cassegrain focus and (**b**) Nasmyth focus.

focus are identical. Therefore, the classical Cassegrain focus is also free from spherical aberration. The Cassegrain focus has coma, astigmatism, and field curvature. The focal ratio of a Cassegrain system is normally larger than a prime focus system, so that the usable field of view is larger. The tube length of a Cassegrain system is short compared to its focal length.

If a concave elliptical secondary mirror is added to a prime focus system, the new Cassegrain system becomes a Gregorian one. A Gregorian system has similar properties as a classical Cassegrain except for a longer tube length and larger secondary mirror for a similar field diameter. For this reason, Gregorian systems are not very often used in astronomy. However, if a folding mirror is added, the tube length becomes very short. The new system can be used for extremely large telescopes. Another important feature of a Gregorian system is that it has a real (not virtual) exit pupil that is close to the prime focus. If a chopping mirror is used in this location, the primary mirror illumination will not change. If a deformable mirror is used in the location, the primary mirror surface error can be compensated in an area by area pattern.

If a Cassegrain system satisfies both equal path length and the Abbe sine conditions, the system will form sharp images for both on-axis and off-axis objects. The Abbe sine condition states the sine of output angle in image space being proportional to the sine of the input angle in the object space. This Cassegrain system will have no spherical aberration or coma. This system is the Ritchey–Chretien system. For an R-C system, both primary and secondary mirrors are hyperboloids. The useful field of view of an R-C system is larger than that of a classical Cassegrain one. When the field of view is large, the ideal focal surface is curved due to field curvature. The primary mirror of an R-C system has spherical aberration so that a field corrector is required at the prime focus. Dual reflector systems also include other primary and secondary shapes. When the primary is a spherical one, the cost of the system is low. This special dual reflector system is often used.

In a dual reflector system, the focal position can be moved out of the light path by inserting an inclined mirror at the elevation axis. This new focus is called the Nasmyth focus [Figure 1.22(b)]. At the Nasmyth focus of an alt-azimuth telescope, the platform is fixed. The alt-azimuth system is discussed in Chapter 3. The Nasmyth focus is suitable for very large and heavy detectors. The properties of the Nasmyth focus are the same as a Cassegrain focus.

1.3.1.3 Coude Focus

For installation of giant instruments, the focal position can be moved away from the telescope through a series of mirrors. The light beam is usually through both declination and polar axes of an equatorial telescope, or the elevation and azimuth axes of an alt-azimuth telescope. The new focus is stable and fixed. This focus is the Coude focus and the system is the Coude system (Figure 1.23). The focal ratio of a Coude system is large. The instrument on this focus is located in a large laboratory where the telescope provides the star light for further analysis. The mirrors before the Coude focus can be flat or axially symmetrical. When using curved mirrors, the system focal ratio can be changed to satisfy the observation requirements. In the Coude focus, the field will rotate as the telescope moves. For an equatorial telescope, the field rotation is the same as the sidereal motion of stars.

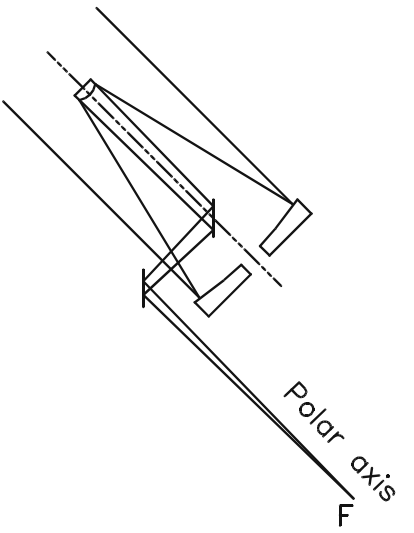


Fig. 1.23. Coude focus.

1.3.1.4 Schmidt and Three-Mirror Optical System

The Schmidt system was invented by Schmidt in 1931. It is a catadioptric or a reflecting-refracting optical system (Figure 1.24). The primary of the system is a spherical mirror. However, the entrance pupil is located at the center of curvature of the primary. The spherical mirror suffers from spherical aberration but is symmetrical for all field angles. For beams incident from different angles, the imaging condition is identical except the projection of the pupil is not the same. An aspherical lens corrector is used at the entrance pupil for the compensation of the spherical aberration. The corrector has two sides; on side is a plane and the other a special curved surface.

If a ray is parallel to and at a distance y_0 from the axis, it will be reflected by the primary to its focus point S_0. If the ray is parallel to and at a distance y from the axis, it will be reflected to another on-axis point S by the spherical mirror. The distance between these two points is:

$$SS_0 = \frac{y^2 - y_0^2}{4R} \tag{1.41}$$

where R is the radius of the spherical primary mirror. For compensating this spherical aberration, a small angle θ has to be introduced for the incoming ray:

$$\theta = \frac{y^3 - y_0^2 y}{R^3} \tag{1.42}$$

If the corrector is at the center of curvature of the spherical mirror and the refractive index is n, the slope of the incoming ray should be:

$$\frac{dx}{dy} = \alpha = \frac{\theta}{n-1} \tag{1.43}$$

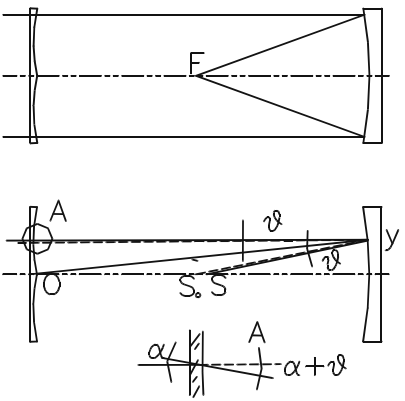


Fig. 1.24. Schmidt optical system.

Therefore, the curved shape of the corrector surface can be expressed as:

$$x = \frac{y^4 - 2y_0^2 y^2}{4(n-1)R^3} \tag{1.44}$$

To minimize chromatic aberration, it is necessary to set $y_0 = 0.433$ D, where D is the diameter of the corrector. To assure a large field without vignetting, the diameter of the spherical primary mirror should be larger than that of the corrector. One way to define a Schmidt telescope is to use a group of three numbers: $a/b/c$, where a is the diameter of the corrector, b that of the spherical primary mirror, and c the focal length. The corrector is very thin and a diameter up to 1.2 m is possible. For correcting chromatic aberrations, an achromatic corrector can be made of two glass plates with different refractive index. The disadvantages of a Schmidt telescope are: (a) the aspherical corrector is not easy to make and its size is limited; (b) the focal surface is curved; (c) the focal position is inside the light beam; and (d) the tube length is twice the focal length. Extremely wide angle, low image quality Schmidt systems are also used as fluorescence detectors for gamma ray or cosmic ray observation (Section 10.2.3).

To challenge the aperture limit of the Schmidt telescope, a type of reflecting Schmidt telescope can be used. The corrector of this design is replaced by a shaped siderostat reflector. This profiled reflector produces a path length change of the beam to compensate spherical aberrations of the primary mirror. However, as the reflecting corrector moves to a different pointing, the surface shape of the corrector has to change by a very small amount, requiring active mirror control. The Large sky Area Multi-Object Spectroscopic Telescope (LAMOST) in China is such a 4 m reflecting Schmidt telescope.

Another catadioptric system is the Maksutov one. The corrector used in this system is a thick crescent achromatic lens. The Maksutov telescope is free from spherical aberration, coma, and chromatic aberrations. The astigmatism is small but field curvature is large. Since the thickness of the corrector is about one tenth of the diameter, the aperture size of the system is very limited.

A wide field of view can be achieved by a three-mirror system. In an early three-mirror system design, the primary and secondary mirrors form an afocal Mersenne beam compressor. In this afocal system, both mirrors, one concave and one convex, are parabolic with their foci at the same position. The tertiary mirror is spherical with its center of curvature at the vertex of the secondary mirror. The aberration of this system is similar to a Schmidt system without a corrector.

In an improved Paul–Baker design, the spherical aberration is corrected by making the secondary mirror with a central radius the same as the tertiary mirror (Wilson, 2004, p227). Using the aspheric plate theory discussed in Section 1.3.4, the Paul–Baker system is finally derived by making all spherical aberration, coma, and astigmatism zero. The field is curved. This system has a large obstruction ratio, but has a useful field of view over one degree. Willstrop

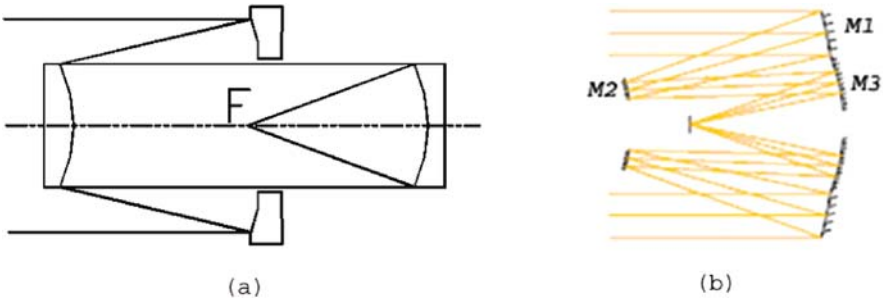


Fig. 1.25. Willstrop's three-mirror system (**a**) and short tube three-mirror system (**b**) (Liang et al., 2005).

improved this Paul–Baker system by placing the tertiary mirror sufficiently behind the primary and the final image is formed on the primary mirror plane. The primary mirror is also modified to keep some of the field curvature. The new system has a wider field of view, known as the Willstrop (1984) system [Figure 1.25(a)]. In the system, the primary mirror is a quasi-paraboloid, the secondary and tertiary are quasi-spherical (Cheng and Liang, 1990). Since there are many parameters in the system for optimization, the field of view can reach 4°. The disadvantages of the Willstrop system are its long tube length and a focal position inside the beam.

A short tube version of the three-mirror system was developed by Liang et al. (2005). In this system, the tertiary mirror is moved to the primary mirror plane. The image is formed behind the secondary mirror [Figure 1.25(b)]. The new design is compact and the tube length is half of the old system. This short tube three-mirror design is now used in the 8.4 m Large Synoptic Survey Telescope (LSST). The design has a usable field of view of 3.1°. The etendue (defined in Section 1.2.3) of the LSST will be 319 $\mathrm{m}^2\ \mathrm{deg}^2$, the largest instrument ever built.

1.3.1.5 Folding and Other Optical Systems

When the aperture of a telescope becomes very large, the tube will be extremely long. The weight and deformation of the tube increase as a cubic power of the length, the moment of inertia increases at an even higher power of the length. Therefore, a short and compact optical system is desirable. In this situation, a folding optical design can be used. A folding system is obtained by inserting a flat mirror in the middle of an existing optical system. The reflection of the secondary mirror can bring the primary and secondary in the same plane. The tube length becomes half as the original system.

Figure 1.26 is a folding Gregorian system. In this system, the primary and secondary mirrors form a continuous surface. If the system is used in the infrared regime, the joined primary and secondary mirrors can be made by diamond

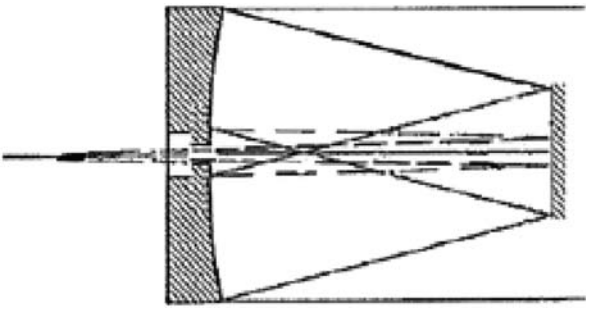


Fig. 1.26. A folding Gregorian optical system.

turning of aluminum material. This folding optical system is widely used in the military as very sensitive infrared sensors.

A major folding optical system is proposed for a planned future 100 m OverWhelmingly Large (OWL) telescope design (Figure 1.27). A flat mirror is added to the beam path of a spherical prime focus system. In this way, the tube length of this giant telescope becomes two-thirds of the original system. The focal position is located between the primary and the flat folding mirrors. The focal position access is easier. The spherical aberration of the system is corrected by a smaller four-mirror field corrector. However, this project has now been replaced by a smaller, less ambitious 42 m European Extremely Large Telescope (E-ELT) project. Apart from folding optical system, a special grazing incident system, also called the Wolter optical system, is used in X-ray imaging (Section 9.2.2).

In space telescope design, optical systems used include Fresnel lens and photon sieves, both diffractive systems. A Fresnel lens forms a real image at

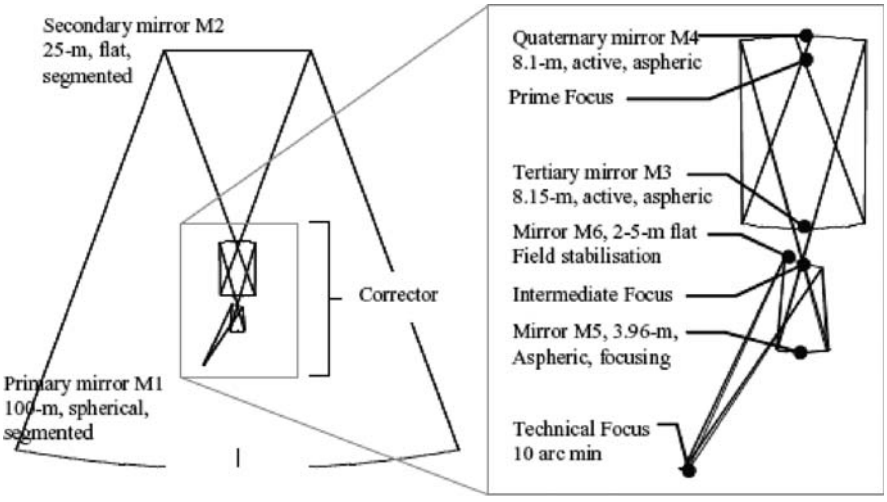


Fig. 1.27. Optical layout of the OWL telescope (Dierickx et al., 2004).

the focus by alternating zones of a half wave phase delay. A Fresnel lens reduces the amount of material required compared to a conventional spherical surface lens by breaking the lens into a set of concentric annular sections. The total number of zones is $N = D^2/(8\lambda F)$, where D is the lens diameter and F the focal length. The Fresnel lens has advantages of low weight, low cost, and low tolerance required. The Fresnel lens is tuned to a particular wavelength. Radiation from other wavelengths is bent to different focal positions. This chromatic aberration is its main drawback. However, with information from all wavelengths, post-processing of the data can retrieve real images of all wavelengths within the spectral range (Lo and Arenberg, 2006).

Photon sieves are another diffractive optical system. Its principle is similar to that of the Fresnel plate with alternative opaque and transparent circular zones. However, instead of continuous rings, many holes are located in the transparent portions of an opaque plate. These holes are unevenly distributed and not connected, so that the structural integrity is improved. The sieves can be made by membranes pulled flat in space. The diameter of the holes reduces as the radius increases. The radius of the hole center is $r_n^2 = 2F\lambda + \lambda^2$, where F is the focal length and the hole diameter of the ring n is defined by $w = \lambda F/(2r_n)$. The holes are distributed randomly to avoid diffraction spikes in the far field. So far, a photon sieve with over 10 million holes has been constructed for astronomy (Anderson and Tullson, 2006).

Special noncontinuous mirror systems are also discussed in this book. These include Davies–Cotton optics of spherical sub-mirrors on paraboloidal surfaces used in air Cherenkov telescopes (Section 9.3.3) and spherical (Section 5.3.3) and flat (Section 9.2.4) mirror interferometers.

1.3.2 Aberrations and Their Calculations

Gaussian optics, or the first-order optics, assumes the light beam passing lenses is infinitesimally small and near the axis of the system. It is also known as paraxial optics. A practical optical system is not a Gaussian one. If P' is the image of an object P, the light rays from point P will not exactly pass the point P' in a practical optical system; they will cover a small area around the point P'. This means that the image of a point object is not a sharp point even if the diffraction effect is not considered. The geometrical difference in the image plane between a Gaussian image point and real image spread is called geometrical aberration. A Gaussian image corresponds to a Gaussian wavefront. The wavefront deviation from a perfect (Gaussian) wavefront is called wavefront aberration, or wave aberration, or wavefront error. As light rays are perpendicular to the wavefront, the geometrical aberration is proportional to the slope of the wavefront aberration.

For an axial symmetrical image system, if $O\eta\xi$ is a coordinate plane in the object space and $O'yz$ a coordinate plane in image space, we can derive all possible terms of its wave aberration. For rays from point $P(\eta, \xi)$ to point

$P'(y,z)$, the corresponding wave aberration W will be a function of all the related coordinates as:

$$W = W(\eta, \xi, y, z) \tag{1.45}$$

Because the system is axially symmetric, only two polar distances and one angle difference are involved. Use the following three variables:

$$\begin{aligned} R &= \eta^2 + \xi^2 = r^2 \\ R' &= y^2 + z^2 = r'^2 \\ u &= y\eta + z\xi = r \cdot r' \cos(\phi - \psi) \end{aligned} \tag{1.46}$$

The wave aberration can be expressed as a power series of these three variables. In the series, the constant and the first-order terms represent Gaussian optics and it produces a perfect Gaussian image. Taking the second power terms from the rest of the series:

$$W_1 = \frac{A}{4}R^2 + \frac{B}{4}R'^2 + C \cdot u^2 + \frac{D}{2}R \cdot R' + ER \cdot u + FR' \cdot u \tag{1.47}$$

This is the first-order wave aberration. The geometrical aberrations T_{Ay} and T_{Az} are related to the wavefront slope, so:

$$k \cdot T_{Ay} = \frac{\partial W}{\partial y} \tag{1.48}$$

$$k \cdot T_{Az} = \frac{\partial W}{\partial z} \tag{1.49}$$

where k is a constant. If the object is assumed in the meridian plane, $\xi = 0$, then the first-order geometrical aberrations can be expressed as:

$$\begin{aligned} k \cdot T_{Ay} &= By(y^2 + z^2) + F\eta(3y^2 + z^2) + (2C + D)\eta^2 y + E\eta^3 \\ k \cdot T_{Az} &= Bz(y^2 + z^2) + 2F\eta yz + D\eta^2 z \end{aligned} \tag{1.50}$$

In the formulas, the coefficient A disappears and all the terms are the third power of the coordinates. Therefore, the first-order aberration is also called the third-order aberration. The coefficients B, C, D, E, and F in the expressions are aberration coefficients. These coefficients are used by scientists and engineers in the UK, the US, and Australia. If the above formulas are changed into (Slyusarrev, 1984):

$$\begin{aligned} k \cdot T_{Ay} &= S_I y(y^2 + z^2) + S_{II}\eta(3y^2 + z^2) + (3S_{III} + S_{IV})\eta^2 y + S_V \eta^3 \\ k \cdot T_{Az} &= S_I z(y^2 + z^2) + 2S_{II}\eta yz + (S_{III} + S_{IV})\eta^2 z \end{aligned} \tag{1.51}$$

then the coefficients S_I to S_{IV} are respectively called spherical aberration, coma, astigmatism, distortion, and field curvature ones. These coefficients are widely used by scientists and engineers in China, Russia, and Europe.

For a given optical system, the refraction indexes are different for radiation of different wavelength and, therefore, there are chromatic aberrations in the wave aberration expression:

$$W_{0c} = \frac{1}{2} C_I (y^2 + z^2) + C_{II} \eta y \tag{1.52}$$

where C_I is the axial (longitudinal) chromatic aberration coefficient and C_{II} the lateral (transverse) chromatic aberration coefficient. The first-order aberrations and chromatic aberrations are major image errors in an optical system.

1.3.2.1 Spherical Aberration

The first term in the geometrical aberration formulas is the spherical aberration. In this term, there are no coordinates of the object space and, therefore, the spherical aberration is a constant over the field of view. Spherical aberration is an image error of an on-axis object. Using a polar coordinate system, the terms are:

$$\begin{aligned} k \cdot T_{Ay} &= S_I r'^3 \sin\psi \\ k \cdot T_{Az} &= S_I r'^3 \cos\psi \end{aligned} \tag{1.53}$$

Combine both equations:

$$(kT_{Ay})^2 + (kT_{az})^2 = S_I^2 r'^6 \tag{1.54}$$

Equation (1.54) shows that the spherical aberration is a set of circles around a common center. The radii of these circles are proportional to the spherical aberration coefficient S_I and increase as cubic power of the radius at the entrance pupil.

The calculation shows that if a star image is divided into five rings with equal radial separation, then 34% of the light is concentrated in the first ring of an area of 4%. The light percentages of the other rings are respectively 20, 17, 15, and 14%, ignoring the diffraction effect. If the image plane moves along the axis, the image size will change.

In Figure 1.28, point F represents the Gaussian image point. The axial spherical aberration δ equals the distance between the intersection of the edge beam with the axis and the Gaussian image point. When the image plane is moved to point H_1, the image size is the smallest. The distance between this point and the Gaussian image point is $FH_1 = 3\delta/4$. Moving an image plane is equivalent to adding a linear term in the spherical aberration. In Figure 1.29, the abscissa is the spherical aberration measured from Gaussian position and the

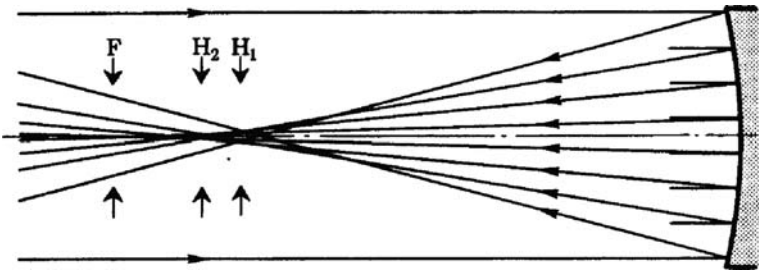


Fig. 1.28. The image shape of spherical aberration along the axis.

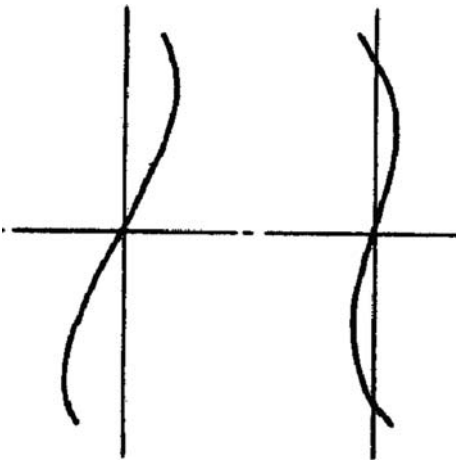


Fig. 1.29. The change of spherical aberration by moving the image plane.

ordinate is the ray distribution over the entrance pupil. The left side of Figure 1.29 is the distribution of spherical aberration on the focal plane. The right side of the figure represents the spherical aberration distribution after adding a linear term. By adding a correct linear term, the image size is greatly reduced (by ∼4 times). However, the brightest image position is not the position of the smallest image size; it is at the point H_2, where more light rays intersect with the axis.

1.3.2.2 Coma

The coma is related to the second terms of the polynomials; these are:

$$\begin{aligned} k \cdot T_{Ay} &= S_{II}\eta(3y^2 + z^2) \\ k \cdot T_{Az} &= 2S_{II}\eta \cdot yz \end{aligned} \tag{1.55}$$

Combine both equations:

$$(kT_{Ay} - 2S_{II}\eta \cdot r'^2)^2 + (kT_{Az})^2 = (S_{II}\eta \cdot r'^2)^2 \tag{1.56}$$

This formula shows that the coma pattern is formed by circles with shifted centers. The rays with the same r' still form a circle and the radius of the circle is proportional to r' squared. The center of the circle has a distance of $2S_{II}\eta r'^2/k$ from the Gaussian image point. Therefore, the coma aberration is formed by circles within a 60° angle, starting from the Gaussian image point.

The coma does not possess axial symmetry. The size of coma can not be reduced by moving the image plane. Figure 1.30 is the change of coma along the axis. In the meridian plane, the distance between the Gaussian image and the coma spread is called meridian (or tangential) coma. The meridian (in radial direction) coma is about three times the sagittal (in direction perpendicular to the radius) coma.

Another important term is the coma constant which represents the size of meridian coma. It is defined as:

$$OSC' = \frac{3S_{II}r'^2}{k} \tag{1.57}$$

1.3.2.3 Astigmatism and Field Curvature

If the coefficients S_I, S_{II}, and S_V are zero, then the geometrical aberrations are linear functions of the coordinate y or z. Equation (1.51) can be simplified as:

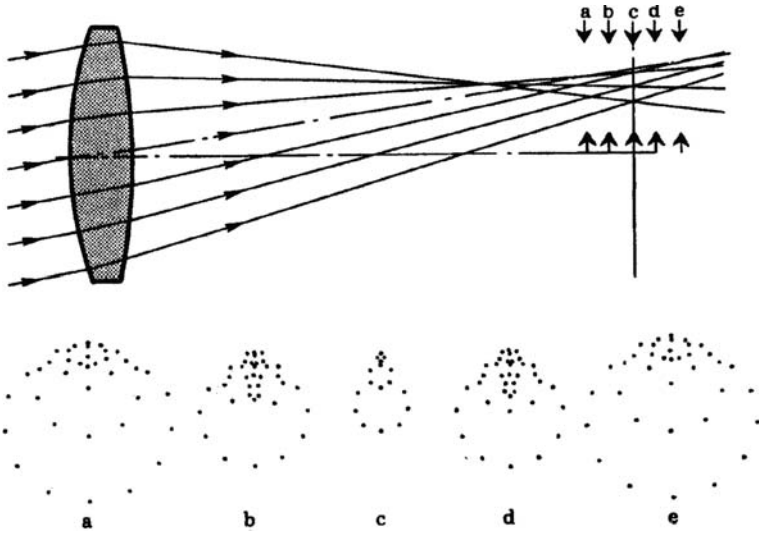


Fig. 1.30. The image of coma along the axis.

$$\left[\frac{kT_{Ay}}{(3S_{III}+S_{IV})\eta^2 r'}\right]^2+\left[\frac{kT_{Az}}{(S_{III}+S_{IV})\eta^2 r'}\right]^2=1 \tag{1.58}$$

In this case, the image is an ellipse with uniform illumination. The radii of two axes are all proportional to $\eta^2 r'$. If the image plane is moving along the optical axis, a linear term is added to this expression (Equation 1.58). When the distance is adjusted so that the aberration along one of the axes is zero, then the ellipse becomes a straight line at two axial positions. One of the lines is along the major axis of the ellipse and the other is along the minor axis. The minimum image spread is between these two axial positions. The coefficient S_{III} is called astigmatism.

If there is no astigmatism $S_{III}=0$, then the system forms a perfect image on a curved surface. The coefficient S_{IV} is called field curvature which produces a curved image surface.

1.3.2.4 Distortion

If $S_V \neq 0$ and all other aberrations are zero, only one expression of the geometrical aberration exists. This expression is:

$$kT_{Ay}=S_V\eta^3 \tag{1.59}$$

This means that the image position is moved nonlinearly in radial direction. The displacement of the movement is a function of η^3. Therefore, it is not a simple scale change. This aberration is called distortion. Figure 1.31 shows two types of the distortion. When the object is a square shape, the distorted images are either (a) barrel shape or (b) cushion shape as in Figure 1.31.

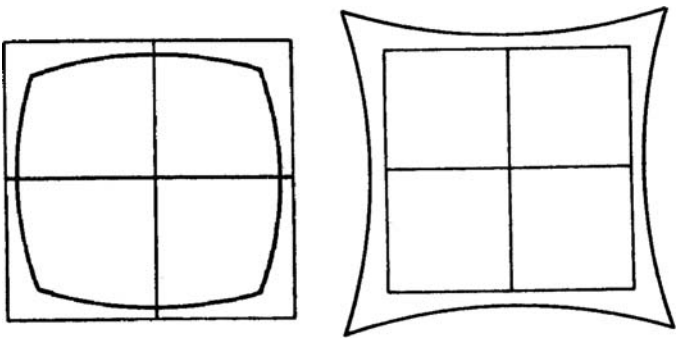


Fig. 1.31. Barrel and cushion distortions.

1.3.3 Formulas of Telescope Aberrations

The same as other optical systems, the aberration of an astronomical telescope is the sum of the aberrations of each optical component. A telescope system involves many reflectors and the object is at infinite distance. The aberration of a telescope system is, therefore different from other lens optical systems.

A simple spherical surface with a radius of r can be expressed as:

$$x = \frac{1}{2r}(y^2 + z^2) + \frac{1}{8r^3}(y^2 + z^2)^2 + \frac{1}{16r^5}(y^2 + z^2)^3 + \cdots\cdots \tag{1.60}$$

For an axially symmetrical surface, the general expression is:

$$y^2 + z^2 = 2rx - kx^2 + \alpha x^3 + \beta x^4 + \gamma x^5 \tag{1.61}$$

This expression can be written in the form of Equation (1.60) as:

$$\begin{aligned} x =& A(y^2 + z^2) + B(y^2 + z^2)^2 + C(y^2 + z^2)^3 + D(y^2 + z^2)^4 \\ &+ E(y^2 + z^2)^5 + \cdots\cdots \end{aligned} \tag{1.62}$$

The relationship between coefficients of Equations (1.61) and (1.62) is:

$$\begin{aligned} A =& \frac{1}{2r} \\ B =& \frac{k}{8r^3} \\ C =& \frac{k^2 - \alpha r}{16r^5} \\ D =& \frac{5k^3 - 10\alpha rk - 4\beta r^2}{128r^7} \\ E =& \frac{7k^4 - 21\alpha rk^2 - 12\beta r^2 k + 6\alpha^2 r^2 - 4\gamma r^3}{256r^9} \\ r =& \frac{1}{2A} \\ k =& \frac{B}{A^3} \\ \alpha =& \frac{2B^2 - CA}{A^5} \\ \beta =& \frac{5BCA - 5B^3 - A^2 D}{A^7} \\ \gamma =& \frac{14B^4 - 21B^2 CA + 6A^2 BD + 3A^2 C^2 - A^3 E}{A^9} \end{aligned} \tag{1.63}$$

Referring to Equation (1.60), the expression of an axially symmetrical surface can be expressed as (Wilson, 2004):

$$x = \frac{1}{2r}(y^2 + z^2) + \frac{1}{8r^3}(1 + b_s)(y^2 + z^2)^2 + \cdots\cdots \tag{1.64}$$

where b_s is the conic constant. Another parameter used for a nonspherical axially symmetrical surface is the eccentricity ε. The relationship between the conic constant and the eccentricity is $b_s = -\varepsilon^2$. The eccentricity is defined by the ellipse radii. If the two radii of an ellipse are a_1 and a_2 and $a_1 > a_2$, then $(a_1/a_2)^2 = 1/(1-\varepsilon^2)$. For a spherical surface, $\varepsilon = 0$, for a paraboloid $\varepsilon = 1$, for an ellipsoid $1 > \varepsilon > 0$, and for a hyperboloid $\infty > \varepsilon > 1$. Another expression of the conic constant is the aspherical coefficient G, $G = b_s/(8r^3)$. If a surface has only the fourth power term r^4, the aspherical coefficient is the coefficient of this fourth power term. One has to be careful when the aspherical coefficient is used. Because the aberration is usually normalized, one may produce errors when the aberration is expressed by the aspherical coefficient.

In the above equations, higher order terms have no influence on third-order aberrations; therefore, one may drop these terms. If the focal length f is introduced, the general surface expression can be written as:

$$x = \frac{1}{4f}(y^2 + z^2) + \frac{1}{64f^3}(1 + b_s)(y^2 + z^2)^2 + \cdots\cdots \tag{1.65}$$

Many optical scientists made detailed calculations to derive the aberration formulas for an aspherical surface. The basic approach of these calculations is to derive the wave aberration as in Equation (1.47) first. Then, each term is examined so that the aberration formulas of the surface are derived.

The wave aberration of a system is the optical path length difference between all rays and the principle ray (chief ray, first auxiliary ray), i.e. the ray which passes through the center of the entrance pupil. The second auxiliary ray (marginal ray) is the ray from object to the edge of the entrance pupil. With the aid of a computer, the calculation of the wave aberration becomes easier. From the wave aberrations, the geometrical aberrations may be derived.

For a single reflector surface, the formulas of the aberration coefficients are:

$$\begin{aligned}
S_I &= -\left(\frac{y_1}{f_1}\right)^4 \frac{f_1}{4}(1 + b_1) \\
S_{II} &= -\left(\frac{y_1}{f_1}\right)^3 \frac{1}{4}[2f_1 - s_{p1}(1 + b_1)]u_{p1} \\
S_{III} &= -\left(\frac{y_1}{f_1}\right) \frac{1}{4f_1}[4f_1(f_1 - d) + s_{p1}^2(1 + b_1)]u_{p1}^2 \\
S_{IV} &= +\frac{H^2}{f_1}
\end{aligned} \tag{1.66}$$

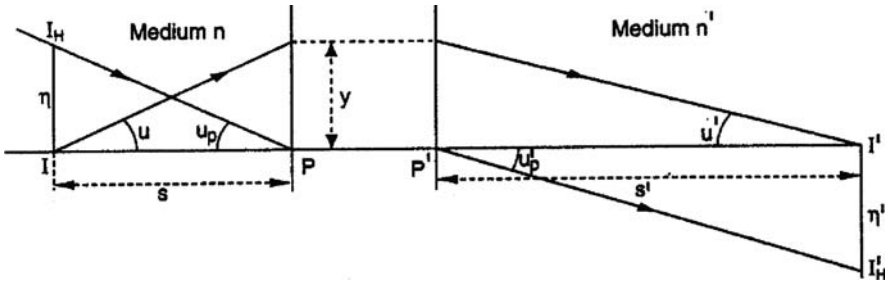


Fig. 1.32. The parameters used in Lagrange invariants calculation (Wilson, 1968).

where b_1 is the conic constant of the surface, f_1 the focal length, y_1 the height of incidence of the second auxiliary ray, s_{p1} the distance between the entrance pupil and the reflecting surface (note: this distance will be positive if it is in the same direction as the incident ray and will be negative if it is not), and $H = nu\eta$ is the Lagrange invariant (Figure 1.32). In a telescope system, the Lagrange invariant is $H = nyu_p$, where n is the refraction index. In a reflecting system, $n = -1$, u_p the incident ray angle and y the radius of the ray incidence on the aperture plane.

For an optical system, parameter normalization can be used so that we have:

$$\begin{aligned} y_1 &= +1 \\ u_{p1} &= +1 \\ f &= \pm 1 \\ H &= -1 \end{aligned} \tag{1.67}$$

If the reflecting surface is also the entrance pupil of the system, the aberration coefficients are:

$$\begin{aligned} S_I &= 1/4(1 + b_1) \\ S_{II} &= 1/2 \\ S_{III} &= 1 \\ S_{IV} &= -1 \end{aligned} \tag{1.68}$$

From these expressions, only the spherical aberration coefficient is related to the conic constant of the reflector. All other aberration terms are constants. To eliminate spherical aberration, the conic constant of the reflector has to be -1, corresponding to a paraboloidal reflector.

The main aberration of a paraboloid reflector is coma. When the field of view increases, astigmatism and field curvature become serious. In Equation (1.68), the effect from moving the entrance pupil is not included. When the entrance pupil is moved away from the reflector, the change of the aberration coefficients will be:

$$\begin{aligned}\delta \cdot S_{II} &= s_{p1} \cdot S_I \\ \delta \cdot S_{III} &= 2s_{p1} \cdot S_{II} + s_{p1}^2 S_I\end{aligned} \tag{1.69}$$

where s_{p1} is the distance of the entrance pupil away from the primary. The coma of a reflector system can therefore be eliminated by adjusting the entrance pupil position. If the entrance pupil of a spherical primary is moved to the center of curvature, the new system is coma free. This is the principle of a Schmidt telescope.

A dual reflector system or a Cassegrain system is shown in Figure 1.33. In the following formulas, the Subscripts 1 and 2 represent the parameters of the primary and secondary mirrors respectively. If no Subscript, the parameters are those for the system. The formula for the system focal length is:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 \cdot f_2} \tag{1.70}$$

where d is the distance between the primary and secondary. If m is the magnification of the secondary mirror, then $f = mf_1$. The aberration formulas for a general dual reflector optical system are (Wilson, 1968):

$$\begin{aligned}
S_I &= \left(\frac{y_1}{4}\right)^4 (-f\varsigma + L\xi) \\
S_{II} &= \left(\frac{y_1}{4}\right)^3 \left[-d\xi - \frac{f}{2} - \frac{s_{p1}}{f}(-f\varsigma + L\xi)\right] u_{p1} \\
S_{III} &= \left(\frac{y_1}{4}\right)^2 \left[\frac{f}{L}(f+d) + \frac{d^2}{L}\xi + s_{p1}\left(1 + \frac{2d}{f}\xi\right) + \left(\frac{s_{p1}}{f}\right)^2 (-f\varsigma + L\xi)\right] u_{p1}^2 \\
S_{IV} &= H^2 \left[\left(\frac{m}{f}\right) - \left(\frac{m+1}{f - md}\right)\right] \\
\varsigma &= \frac{m^3}{4}(1 + b_1) \\
\xi &= \frac{(m+1)^3}{4}\left[\left(\frac{m-1}{m+1}\right)^2 + b_2\right] \\
L &= f - md
\end{aligned} \tag{1.71}$$

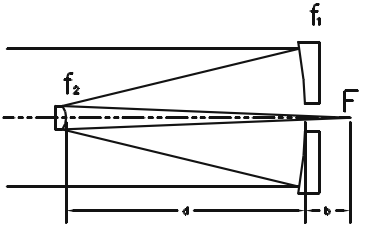


Fig. 1.33. The basic parameters of a dual reflector system.

where b_1 and b_2 are conic constants of the primary and secondary mirror. In the following formulas, b represents the back distance between the primary mirror and the system focus (Figure 1.33).

For a dual reflector system where the entrance pupil is at the primary mirror plane, $s_{p1} = 0$ and the primary mirror is parabolic, $b_1 = -1$. To eliminate spherical aberration, the conic constant of the secondary mirror has to be:

$$b_2 = -\left(\frac{m-1}{m+1}\right)^2 = -\frac{(f-f_1)^2}{(f+f_1)^2} \tag{1.72}$$

This is a classical dual reflector system. If $m<0$, the secondary mirror is a hyperboloid; the system is a classical Cassegrain system. If $m>0$, the secondary mirror is an ellipsoid, the system is a Gregory (or Gregorian) system. The coma coefficient of these systems is $1/2$, the same as for a primary mirror system. The astigmatism is $-m/(1-mb/f) \approx -m$, larger than that of a primary mirror system with the same focal ratio.

If both spherical aberration and coma are eliminated in a dual reflector system, then the system is an R-C system. From Equation (1.71), the solutions of $S_I = S_{II} = 0$ are:

$$\begin{aligned} b_1 &= \frac{2(1-mb/f)}{m^3(1-b/f)} - 1 \\ b_2 &= -\frac{(m-1)[m^2(1-b/f)+1+b/f]}{(m+1)^3(1-b/f)} \end{aligned} \tag{1.73}$$

These formulas show that both the primary and the secondary mirrors are hyperboloids. The main aberration of this system is astigmatism as:

$$S_{III} = \frac{2(1+b/f)-4m}{4(1-mb/f)} \tag{1.74}$$

This is not very large. Therefore, the usable field of view of an R-C system is about 40 arcmin, far larger than that of a classical Cassegrain system.

Using the same method by assuming $S_I = 0$, and assuming either $b_1 = 0$ or $b_2 = 0$, it is possible to find the conic constants of the other mirror when one mirror of the system is spherical. The solutions are:

$$\begin{aligned} b_1 &= 0 \\ b_2 &= -\frac{(m-1)[(m^2-1)(1-mb/f)+m^3]}{(m+1)^3(1-mb/f)} \end{aligned} \tag{1.75}$$

or

$$\begin{aligned} b_1 &= -\left(1+\frac{(m^2-1)(1-mb/f)}{m^3}\right) \\ b_2 &= 0 \end{aligned} \tag{1.76}$$

From the calculations of the system aberrations, it is possible to find the magnitude of the third-order aberrations for a particular system.

Apart from Equation (1.70), there exist the following additional relationships in a dual reflector telescope system. These are:

$$\begin{aligned} (f+f_1)d &= f_1(f-b) \\ -(f-f_1)f_2 &= f_1(d+b) \\ -(f-f_1)f_2 &= f(f_1-d) \end{aligned} \tag{1.77}$$

If there is no vignetting in a field angle of 2ϕ and the diameter of the primary mirror is D, then the diameter of the secondary mirror is:

$$(d+b)D/f + 2d\phi = (f_1+b)D/(f_1+f) + 2d\phi \tag{1.78}$$

1.3.4 Field Corrector Design

1.3.4.1 Corrector Design for Prime Focus System

Equation (1.68) shows that the prime focus suffers from serious coma for paraboloidal or hyperboloidal reflectors. The coma is inversely proportional to the cubic power of focal ratio and it limits the field of view. The prime focus of a hyperboloid reflector also suffers from spherical aberration. When the focal ratio is smaller, direct photographic work at the prime focus is difficult and a corrector is necessary.

Early correctors made of lenses were used to correct coma for paraboloidal primary mirrors. The purpose was to make the system free from spherical aberration, coma, and field curvature altogether. However, the spherical aberration free condition was difficult to meet for paraboloidal reflectors. At the same time, chromatic aberration from lens corrector is added into the system.

Lenses with optical power suffer from chromatic aberration. The term of optical power means the ability for lenses to converge or diverge light. Any lens except a plane one has optical power. For this, in 1935, Ross stated that the correction of both first-order chromatic aberrations, longitudinal and transverse, was impossible with separated lenses, whether or not the corrector is afocal. If the first-order longitudinal and transverse chromatic coefficients for each lens are C_1 and C_2, then for a two lens system (Wilson, 1968):

$$\begin{aligned} \sum C_1 &= (C_1)_1 + (C_1)_2 \\ \sum C_2 &= [E_1(C_1)_1 + E_2(C_1)_2]H \\ E_i &= \frac{1}{H^2}\frac{y_p}{y} \end{aligned} \tag{1.79}$$

where y_p and y are the height of the first and second auxiliary rays on the lenses and H the Lagrange invariant being unity for a normalized system. From the equations, if the chromatism is corrected, the necessary condition is either $E_1 = E_2$, (system without space) or $C_1 = 0$, (zero power for lenses).

Therefore, Ross started with a two-lens afocal corrector with zero spacing and zero thickness. An afocal set of lenses has zero optical power. The focal length of the telescope is not affected and chromatic aberrations are not added. If a corrector is placed near the focal point, the size is small.

For a paraboloidal reflector, if both spherical aberration and coma are corrected, astigmatism becomes larger. Therefore, Ross eliminated coma and astigmatism, $S_{\mathrm{II}} = S_{\mathrm{III}} = 0$, and retained some spherical aberration. The system aberrations are:

$$\begin{aligned} (S_I)_{cor} &= f\left(\frac{1+E}{E^2}\right) = g\left(\frac{f}{f-g}\right)^2 \approx g\left(\frac{f}{f-g}\right) \\ (S_V)_{cor} &= k\left(\frac{f-g}{g}\right)^2 \end{aligned} \tag{1.80}$$

where g is the distance between the lenses and the focus, k a constant, f the focal length, and $E = (f-g)/g$ a parameter describing how far the lenses are away from the focus. The first equation assumes that the astigmatism of the primary is negligible and it requires a corrector free from astigmatism. Under this condition, the system's spherical aberration depends only on the parameter g. To reduce spherical aberration, the corrector has to be near the focus. However, this produces large distortion limiting the field of view and requires lenses with larger curvature which produce higher order aberrations. Therefore, Ross proposed a corrector with three lenses: a doublet and a field lens [Figure 1.34(a)]. In this system, a meniscus field lens with large curvature is added towards the primary mirror for correcting the spherical aberration arising from the doublet set to achieve a relatively large field of view.

The field of view of this corrector is still small if a small focal ratio is used. The performance improves when a thick meniscus lens of high refractive index is used. This limits the work at short wavelengths. Therefore, Wynne (1967) proposed a four-lens corrector. This corrector consists of two afocal pairs of lenses providing more freedom to make spherical aberration, coma, and astigmatism free for the system.

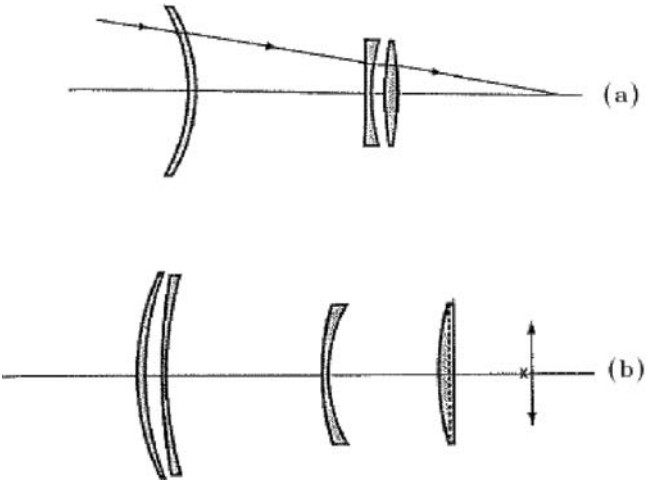


Fig. 1.34. (**a**) Ross three-lens corrector and (**b**) Wynne's four-lens correctors for prime focus (Wynne, 1967).

If the left Subscripts A and B represent the coefficients of lens pair A and B and the right superscript represents the aberration coefficients at the entrance pupil, then from the spherical aberration, coma, and astigmatism free condition, it follows:

$$\begin{aligned}
&{}_{B}S_{I} = -{}_{A}S_{I} \\
&{}_{A}S'_{II} = {}_{A}S_{II} + E_{A} \cdot {}_{A}S_{I} = 1/4 \\
&{}_{B}S'_{II} = {}_{B}S_{II} + E_{B} \cdot {}_{B}S_{I} = {}_{B}S_{II} - E_{B} \cdot {}_{A}S_{I} = 1/4 \\
&{}_{A}S'_{III} + {}_{B}S'_{III} = 2E_{A} \cdot {}_{A}S_{II} + 2E_{B} \cdot {}_{B}S_{II} + {}_{A}S_{I}(E_{A}^{2} - E_{B}^{2}) = 0
\end{aligned} \tag{1.81}$$

where E_A and E_B are parameters describing how far the two pairs of lenses are away from the focus, $E = d/h_M h_L$, with d the distance to the primary mirror from the corrector pair, h_M and h_L, the height of the secondary auxiliary ray on the primary mirror and on the pair. In the formulas, an equal amount of coma is corrected by each pair and no original astigmatism is assumed. The solutions of the above equations are [Figure 1.34(b)]:

$$\begin{aligned}
{}_{A}S_{I} &= -{}_{B}S_{I} = \frac{1}{2(E_{A} - E_{B})} \\
{}_{A}S_{II} &= -{}_{B}S_{II} = -\frac{E_{A} + E_{B}}{4(E_{A} - E_{B})}
\end{aligned} \tag{1.82}$$

This corrector has separations between each doublet avoiding larger aberrations. To reduce higher order chromatism, the lens glass used should have low dispersion. This type of corrector was used on the 5 m Palomar telescope and it had a field of view of ± 12.5 arcmin. The design is also scaleable for other

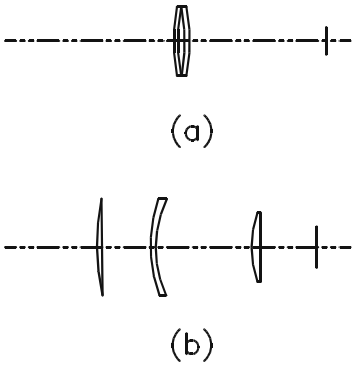


Fig. 1.35. Prime focus correctors for an R-C system.

telescopes. Generally, the larger the corrector is relative to the primary mirror, the better performance it will have.

The field corrector design for nonparaboloidal prime focus is easier as spherical aberration exists in the uncorrected system. Figure 1.35 shows two types of prime focus correctors for an R-C primary. Figure 1.35(a) has two lenses and (b) three lenses. These correctors are compact and efficient. The field of view of the three-lens system is ± 30 arcmin.

All lenses discussed so far are spherical ones. Aspherical lenses (plates) have also been used in astronomy over the past thirty years. Aspherical plates used in field correctors were first proposed by Paul in 1935. Meinel, Gascogne (1968, 1973), Schulte, Kohler, and Su (1963, 1967) produced various aspherical plate corrector designs as shown in Figure 1.36. Figure 1.36(a) is a one-aspherical-plate corrector used for an R-C prime focus. It corrects both spherical aberration and coma. However, the system astigmatism is far larger than that of a hyperboloid mirror. The field of view is limited. Correctors with wide field of view have three or four aspherical plates.

In the corrector design, aspherical plates similar to a Schmidt corrector with spherical aberration only are also used. Following Equations (1.69) and (1.71), a dual reflector system plus one aspherical plate has aberration coefficients as (Wilson, 1968):

$$\begin{aligned} S_I &= \left(\frac{y_1}{4}\right)^4 (-f\varsigma + L\xi + \delta S_I^*) \\ S_{II} &= \left(\frac{y_1}{4}\right)^3 \left[-d\xi - \frac{f}{2} + \frac{s_{pl}}{f}\delta S_I^*\right] u_{p1} \\ S_{III} &= \left(\frac{y_1}{4}\right)^2 \left[\frac{f}{L}(f+d) + \frac{d^2}{L}\xi + \left(\frac{s_{pl}}{f}\right)^2 \delta S_I^*\right] u_{p1}^2 \\ L &= f - md \end{aligned} \tag{1.83}$$

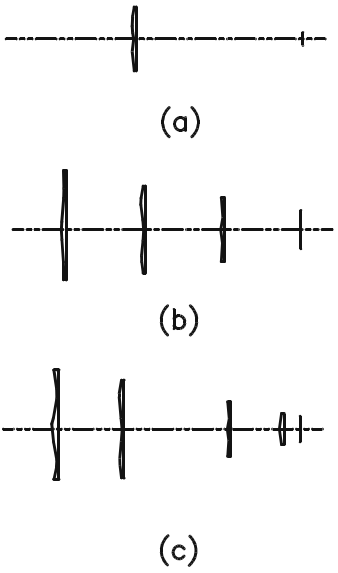


Fig. 1.36. Various aspheric plate correctors.

where δS_I^* is the spherical aberration of the aspherical plate, s_{pl} the distance between the plate and the primary mirror, and d the distance between the primary and the secondary mirror. The distance is positive if the ray hits the primary before it hits the aspherical plate. For a corrected system, all three equations are zero.

If the thickness of a simple aspherical plate varies as the fourth power of the radius, $t = Gr^4$, where G is the aspherical coefficient, and the refractive index of the material is n, then its spherical aberration is:

$$\delta S_I^* = 8(n-1)Gy^4 \tag{1.84}$$

An aberration reduction factor of l/f is needed for its contribution at the focal plane as the aspherical plate is not located at the pupil plane, where l is the distance between the plate and the focus and f the system focal length. For a primary mirror system, the parameters in Equation (1.83) become:

$$m = -1, L = f, d = 0, \xi = 0, u_{p1} = 1 \tag{1.85}$$

Given:

$$S_{pl}/f = E \tag{1.86}$$

and after normalization, $f = y_1 = u_{pl} = 1$. For a prime focus corrector of three aspherical plates, the system aberration coefficients are (Shao and Su, 1983):

$$\begin{aligned} S_1 + S_2 + S_3 &= f\varsigma \\ E_1 S_1 + E_2 S_2 + E_3 S_3 &= f/2 \\ E_1^2 S_1 + E_2^2 S_2 + E_3^2 S_3 &= -f \end{aligned} \tag{1.87}$$

From the equations, aberrations of three aspherical plates required can be obtained. The surface shape of these plates can be determined if the material is determined. For correcting the field curvature, the last aspherical plate has a meniscus shape. This plate is known as a field lens, or a field plate. An aspherical plate corrector can have a field of view of up to one square degree over a wide wavelength range. However, the design of these correctors is purely theoretical and their application is seriously limited by the difficulties of the aspherical plate manufacture.

1.3.4.2 Correctors for Cassegrain System

The corrector design for a Cassegrain system is the same as for the primary system. Simple Cassegrain correctors are made of two or three spherical lenses (Figure 1.37). When the required field of view increases, the number of lenses required increases. A classical Cassegrain system is free from spherical aberration, but has serious coma. With a simple corrector with only two lenses it is difficult to obtain a large field of view. An R-C system, which is free from spherical aberration and coma, has both hyperboloid primary and secondary. The corrector design is easier. If a Cassegrain corrector has three lenses, the distortion can also be corrected. The correction improves as the eccentricity of

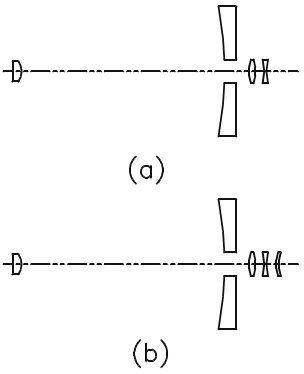


Fig. 1.37. Lens type Cassegrain focus correctors.

the primary e_1^2 increases. The correction for a R-C system is better than for a classical Cassegrain system.

If the surface shape of the primary and secondary is not fixed and the original system has spherical aberration and coma, the best corrected system can be obtained for certain eccentricities of the primary. This special Cassegrain system is named a quasi R-C system. The usable field of view of a corrected quasi R-C system can reach one square degree.

If the corrector lenses are made of glasses with different refractive indexes, the monochromatic performance of the system can be well adjusted.

The same as for the primary system, aspherical plates are also used for correctors of quasi R-C, R-C, classical Cassegrain, and spherical primary Cassegrain systems. These correctors have one, two, or three aspherical plates together with a field plate. When an aspherical plate corrector is used for a spherical primary Cassegrain system, the field of view is large and the field lens is not required. This corrector system has three aspherical plates and the field of view after correction is about 1.5°. The field of view of a spherical prime focus system can reach 40 arcmin with an aspherical plate corrector.

Aspherical plates are similar to reflecting mirrors in optical design. Therefore, reflecting mirror correctors are also used, especially for large aperture optical telescopes. A four-mirror prime focus corrector has been proposed for a future very large telescope (Figure 1.27).

The field corrector design can be combined with the design of filters or differential atmospheric refraction correctors. One type of lens corrector made of two split prisms is used for both the differential atmospheric refraction and the field corrections. It has a special name of "lensm." This is one lens made of two pieces of glasses of different refractive indexes glued together at an inclined plane. It serves as a field corrector as well as a differential atmospheric refraction corrector. As the lensm rotates around its axis, the refracting angle from the prism elements varies according to the elevation angle of the telescope. The lensm compensates the differential refraction of the atmosphere.

1.3.5 Ray Tracing, Spot Diagram, and Merit Function

Classical aberration theory deals with low order aberrations of the system. The analytic calculation of higher order aberrations is very difficult and time consuming. For overcoming the difficulties, ray tracing is used. Ray tracing can provide accurate image spread as geometrical optics expected. Ray tracing also provides optical path length error or wavefront error. The wave aberration can also be used to confirm the geometrical aberration of the system.

The earliest ray tracing formulas were derived by L. Seidel in 1856. The formulas are for a lens optical system with spherical refracting surfaces. The invention of modern computers makes ray tracing even more important. Ray tracing is now used not only in optics, but also in computer graphics and many other important fields.

In general, ray tracing involves the following steps: (a) from the initial starting position and the ray's direction, the intersection point of the ray on a (mirror or lens) surface is determined; (b) at the intersection point, the surface normal is determined; (c) the reflecting or refracting directions of the ray are determined; and (d) starting from the intersection point and the new ray's direction, a new round of ray tracing is performed.

The first important task of ray tracing is to provide the expression of a ray starting at an initial point $X_0 = (x_0, y_0, z_0)^T$and with a ray direction of $X_d = (x_d, y_d, z_d)^T$, where the superscript T means the transpose of a vector and the directional vector used is a normalized one. The simple ray vector expression is:

$$\begin{aligned} x &= x_d t + x_0 \\ y &= y_d t + y_0 \\ z &= z_d t + z_0 \end{aligned} \tag{1.88}$$

In a matrix form, these equations become $X = X_d t + X_0$. For a given lens surface, the equation is:

$$F(x, y, z) = 0 \tag{1.89}$$

Therefore, the equation which is used for solving the intersection point is:

$$F(x_d t + x_0, y_d t + y_0, z_d t + z_0) = 0 \tag{1.90}$$

This equation may have n solutions. The best thing in ray tracing is that only the smallest positive solution along the ray direction is useful for us. This solution is the first intersection point between the ray and the surface.

Different algebraic methods are used to solve the above equation. For high order equations, iterative methods can be used.

If the surface is a simple plane, the surface equation is:

$$ax + by + cz + d = 0 \tag{1.91}$$

The solution of the intersection point is:

$$t = \frac{ax_0 + by_0 + cz_0 + d}{ax_d + by_d + cz_d + d} \tag{1.92}$$

For a plane surface, the matrix expression is $\vec{N} \cdot \vec{p} = c$, where $\vec{N}$ is the normal vector of the plane, $\vec{p}$ a vector from the origin to a point on the

plane, and c a constant. Then, the intersection point between the ray and the plane is:

$$S = \frac{c - \vec{N} \cdot \vec{X}_0}{\vec{N} \cdot \vec{X}_d} \tag{1.93}$$

For a second order surface, a general equation is:

$$ax^2 + 2bxy + 2cxz + 2dx + ey^2 + 2fyz + 2gy + hz^2 + 2iz + j = 0 \tag{1.94}$$

The above equation can be rewritten as:

$$X^T Q X = 0 \tag{1.95}$$

with

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad Q = \begin{bmatrix} a & b & c & d \\ b & e & f & g \\ c & f & h & i \\ d & g & i & j \end{bmatrix} \tag{1.96}$$

The equation for the intersection point is:

$$At^2 + Bt + C = 0 \tag{1.97}$$

The three coefficients in the equation are:

$$\begin{aligned} A &= X_d' Q X_d \\ B &= 2X_d' Q X_o \\ C &= X_o' Q X_o \end{aligned} \tag{1.98}$$

For some typical second-order surfaces, the solutions have very simple forms. If the surface is a sphere with a center of $X_s(x_s, y_s, z_s)$and a radius of S_r, the three coefficients for the solution are:

$$\begin{aligned} A &= x_d^2 + y_d^2 + z_d^2 = 1 \\ B &= 2(x_d(x_0 - x_s) + y_d(y_0 - y_s) + z_d(z_0 - z_s)) \\ C &= (x_0 - x_s)^2 + (y_0 - y_s)^2 + (z_0 - z_s)^2 - S_r^2 \end{aligned} \tag{1.99}$$

If a secondary order surface has its equation as:

$$F = x^2 + y^2 + z^2 - 2rx - e^2x = 0 \tag{1.100}$$

then:

$$\begin{aligned} A &= x_d^2 + y_d^2 + z_d^2 \\ B &= 2(x_0x_d + y_0y_d + z_0z_d) - 2rx_d - e^2x_d \\ C &= x_0^2 + y_0^2 + z_0^2 - 2rx_0 - e^2x_0 \end{aligned} \tag{1.101}$$

Algebraic method used in ray tracing is generally less effective than geometrical method. Using geometrical method to find the intersection point between a ray and a spherical surface, two triangles are formed. One is the triangle EOP, where E is the ray starting point, O is the center of the spherical surface, and P is the intersection point between the ray and the surface, and the other is a triangle OPD, where D is a point in the ray direction and the line OD is perpendicular to the line PD. The distance ED is derived as a scalar product of the vector EO and a normalized ray direction vector. From these two triangles, the distance PD is calculated using the length OD and OP, which is the radius of the spherical surface. OD and OP are two edge lines of a rectangle triangle. If this distance PD is larger than or equal to zero, the ray will intercept the spherical surface and the intersection point P is found. If this distance is negative, there is no intersection point for the ray and the surface. For a ray which intersects with an aspherical surface, a similar, but iterative method can be used in the ray tracing.

By finding the intersecting point between the ray and the surface, the surface normal is determined from the surface equation. The directions of reflecting or refracting rays are then determined. If $\vec{I}$ is the normalized ray vector and $\vec{N}$ the normal vector of the surface, then $\vec{R}$, the reflecting ray vector, and $\vec{T}$, the refracting ray vector, are (Glassner, 1989, Figure 1.38):

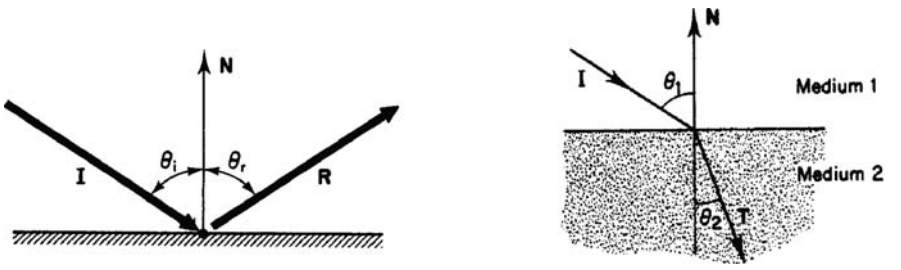


Fig. 1.38. Reflecting and refracting ray tracing.

$$\begin{aligned}
\vec{R} &= \vec{I} - 2(\vec{N}\cdot\vec{I})\vec{N} \\
\vec{T} &= n_{12}\vec{I} + (n_{12}C_1 - C_2)\vec{N} \\
n_{12} &= n_2/n_1,\ C_1 = \cos\theta_1 = \vec{N}\cdot(-\vec{I}), \\
C_2 &= \cos\theta_2 = \sqrt{1 - n_{12}^2(C_1^2 - 1)}
\end{aligned} \tag{1.102}$$

where n_2 and n_2 are the refractive indexes of the medium 1 and the medium 2.

From ray tracing, the intersection points between the ray and the image plane can be found. The diagram of these intersection positions in the image plane is known as the spot diagram. To derive a useful spot diagram, the starting rays are usually uniformly distributed over the entire entrance aperture. They are parallel to each other. The spot diagram represents the energy distribution of the image of a point source in the ray direction.

In optical system design, spot diagrams of different wavelengths and different field angles are derived. They are used for optical system evaluation. Today, ray tracing is also used for wavefront and phase error calculation and stray light control. From the phase error distribution on the aperture plane, a diffraction pattern of a practical optical system can be derived through Fourier transform of the complex aperture field. Both wavefront error and stray light estimations are performed by tracing the ray backwards from the detector surface. For stray light ray tracing, Lambertian scattering as well as specular reflection of the light are used. These are discussed in Section 2.4.2.

For evaluation of an optical system, a mathematical indicator is required to describe overall quality of the system. This indicator is known as the merit function. There are different merit functions. Basically, two aspects are included: the spreading of the spot diagram, which is measured by the distance squared between the weighted spots and their center of gravity, and the scale distortion of the image's center of gravity at a certain field angle. Combining these factors, the merit function has a general form as (Su et al., 1983):

$$\begin{aligned}
\Phi &= \frac{1}{1+\eta}(\Phi_1 + \eta\Phi_2) \\
\Phi_1 &= \frac{1}{n\sum_{j=1}^{e} a_j \sum_{\lambda=1}^{S} I_\lambda} \sum_{j=1}^{e} a_j \sum_{\lambda=1}^{S} I_\lambda \sum_{k=1}^{n} \left[(y - y_0)^2 + z^2\right] \\
\Phi_2 &= \frac{1}{\sum_{j=1}^{e} b_j} \sum_{j=1}^{e} \left(\frac{\sum_{\lambda=1}^{S} I_\lambda \sum_{k=1}^{n} y}{n\sum_{\lambda=1}^{S} I_\lambda} - \frac{\sum_{j=1}^{e} y_c b_j \tan w_j}{\sum_{j=1}^{e} b_j \tan^2 w_j} \tan w_j \right)^2 b_j
\end{aligned} \tag{1.103}$$

where e is the number of field angles, s the number of wavelengths, n the number of rays at each field angle of w_j, a_j, b_j, η, and I_λ the weight functions.

The merit function will reach the minimum at the optimum image plane. Using ray tracing, spot diagram, Fourier transform, and merit function, modern optical design plays a very important role in the development of modern astronomical telescopes.

1.4 Modern Optical Theory

1.4.1 Optical Transfer Function

One important concept in modern optical theory is the optical transfer function. A transfer function is usually used to refer to linear, time-invariant (LTI) systems. The transfer function of a LTI system is a linear mapping of the Laplace transform from the input to output. Linearity here means that the input and output satisfy the superposition law. For an input as a sum of two input components, the output will be the sum of two outputs caused by the two input components individually. Time invariance means that the output will not change except for a given time delay if the input has a similar time delay. If the input is made of a pulse (function) very short in time while maintaining its area, the output is an impulse response function of the system. The transfer function is the Laplace transform of the impulse response function. The output of this system is a convolution of the impulse response function and the input signal both in time domain. In Laplace space, the output (Laplace) is simply a product of the input (Laplace) and the transfer function.

For a stable LTI system, if both the input and output are represented as phasors, complex harmonic vectors with amplitude, phase, and angular frequency, then the output will retain the same angular frequency as the input. In this case, the system transfer function can be replaced by a mapping of the Fourier transform from the input to output. The Fourier transform is a special case of the Laplace transform where the real part of the variable vanishes. This type of transfer function is often referred to as the system frequency response. The frequency response includes two components: amplitude gain and phase shift both in frequency domain.

An optical system is treated as a linear, space-invariant, and incoherent system. Space-invariant is the same as time-invariant except in the space domain. Incoherence means that the lights from two different points of an object (input signal of an optical system) do not have a stable relative phase and are unable to produce interference. Only the lights from the same point possess the quality of interference. The optical transfer function (OTF) of this system is a linear mapping of the Fourier transform of the input (object/source) to output (image). The transform is in the spatial frequency domain instead of the

temporal frequency domain. The modulation transfer function (MTF) is the amplitude part of the OTF and the phase transfer function (PTF) is the phase part. If the phase transfer function is zero, the modulation transfer function is the same as the optical transfer function.

As in temporal frequency, the spatial frequency is a measure of how often a structure repeats per unit of distance. The spatial frequency can be used to describe a pattern, a source, an image, or an intensity distribution. All these can be represented as a sum of infinitely repeated brightness samples (or gratings) with different spatial frequencies. If p_0 is the period of the brightness variation, the spatial frequency is defined as v_0 =unit of length/p_0 expressed in a unit of cycles per unit length (Figure 1.39). In astronomy, a length is sometimes equal to the sine of an angle. For small field angle, the sine can be replaced by the angle itself so that the spatial frequency may also be in cycles per unit angle.

The modulation transfer function is related to the contrast change (or modulation). The contrast at a particular frequency v_0 is defined as $C(v_0) = (l_{\max} - l_{\min})/(l_{\max} + l_{\min})$, where l is the illumination of the pattern

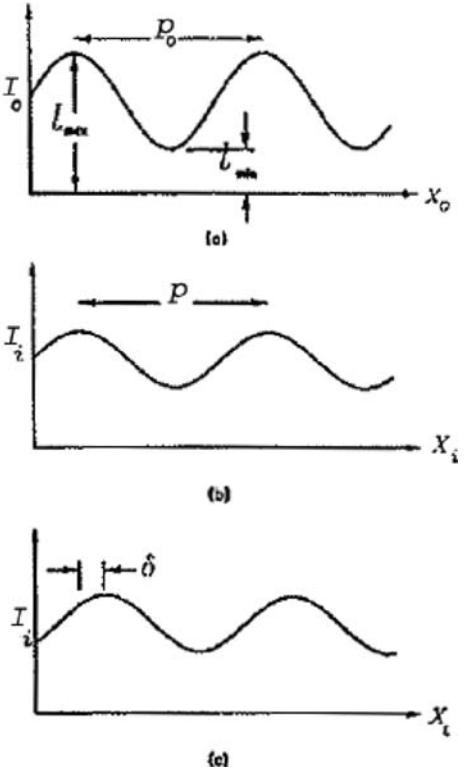


Fig. 1.39. Period, contrast, and phase shift at one spatial frequency.

[Figure 1.39(a)]. The modulation transfer function at this frequency is the contrast ratio of the output (image) to input (object):

$$MTF(v_0) = \frac{C_i(v_0)}{C_o(v_0)} \tag{1.104}$$

In general, any passive optical system does not provide amplitude gain so that $MTF(v) \leq 1$. In a practical optical system, only when $v \to 0$, then $MTF(v) \to 1$. When the spatial frequency, v, approaches the limiting resolution of the system, $MTF(v) \to 0$. The spatial frequency corresponding to the limiting resolution is called the cutoff frequency v_c. Information of the input, where the spatial frequencies are higher than the cutoff frequency, vanishes in the output. Usually, the spatial frequencies are normalized using the cutoff frequency, so that the frequency becomes $v_n = v/v_c$.

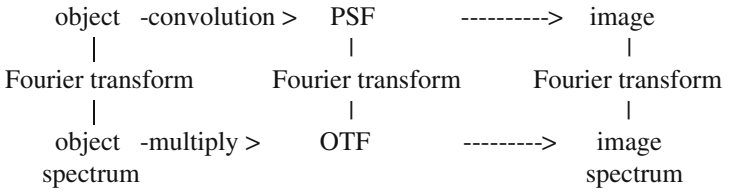
When a transverse displacement δ of the brightness variation appears in the image (output) compared with that in the object (input) [Figure 1.39(c)], then the phase transfer function of this frequency is not zero, it is equal to:

$$PTF(v_0) = \frac{2\pi\delta}{p_0} \tag{1.105}$$

The optical transfer function is a function of both the modulation and phase transfer functions:

$$OTF(v) = MTF(v)\exp[iPTF(v_0)] \tag{1.106}$$

As the impulse response and transfer functions are a Laplace pair in a LTI system, the point spread and optical transfer functions are a Fourier pair in an optical system. A point source is a "pulse" in the space domain. Generally, the following important relationships exist in all optical systems:



One related parameter in optics is the encircled energy (EE) of a point image. The encircled energy is a function of radius in the image plane. It is calculated by first determining total energy of the PSF. Circles of increasing

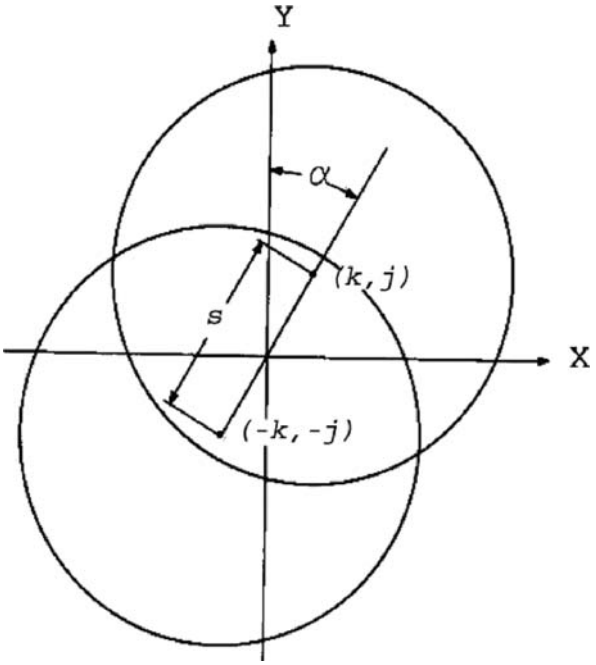


Fig. 1.40. Coordinate system for the aperture field auto-correlation.

radius are then created and the EE is the energy within each circle divided by the total energy.

The optical transfer function can also be calculated from the aperture field function through auto-correlation (Figure 1.40):

$$OTF(k,j) = \frac{\int_{-1}^{1}\int_{-1}^{1} P^*(x+k, y+j)P(x-k, y-j)dxdy}{\int_{-1}^{1}\int_{-1}^{1}[P(x,y)]^2 dxdy} \tag{1.107}$$

where $k = (s/2)\sin\alpha, j = (s/2)\cos\alpha$, and s is the distance between centers of the aperture field P and its conjugated field P^*. The area integral is taken within the overlapped area and the cut-off frequency is $v_n = 2s/\lambda$ when $s = 1$.

To judge whether the phase transfer function exists or not, the following rules can be used: (a) If there is no phase term for an aperture field, then the phase transfer function vanishes; (b) if an aperture field function is a Hermitian one, where its complex conjugate is equal to the original function with the variable changed in sign, then the phase transfer function vanishes; and (c) if both phase and amplitude of the aperture field are even functions, then the

phase transfer function exists. In modern optics, there also exist the following important relationships:

Complex aperture function	-auto-correlation >	Optical transfer function
|		|
Fourier transformation		Fourier transformation
|		|
Field pattern	-squared >	Point spread function

Of the above four functions, the point spread function is a real one and the optical transfer function is either a real or Hermitian one since the function is real-valued if and only if the Fourier transform of it is either real or Hermitian. The field (amplitude in some references) pattern (as used in radio science) includes both amplitude and phase. The optical transfer function and aperture field function are defined in a closed domain, while the field pattern and point spread function are defined in an open domain. When a function in a closed domain is derived from a function in an open one, error exists if the integral is carried out over a limited range.

Using the point spread function to derive the optical transfer function, truncation is necessary. To reduce the error, it is better to normalize the total energy of the point spread function over the truncated area (Figure 1.41). If the

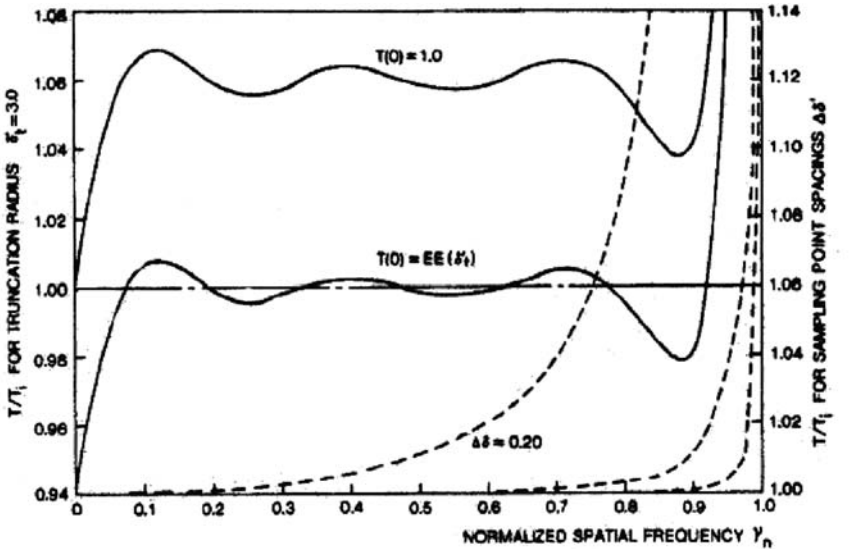


Fig. 1.41. The errors produced when different normalizations are used to derive the modulation transfer function from the point spread function. The *dotted lines* are the errors caused by different sample point spacing (Wetherell, 1980).

integration area used is three Airy diameters, then the energy within this range is 0.943 of the total energy. This number is used for the energy normalization. In this way, a more accurate result is derived. The sample rate is also important as it has direct impact on errors at high spatial frequencies. In general, it requires 10–20 samples in an Airy radius for assuring the accuracy of the optical transfer function.

The aperture field and field pattern are a Fourier pair. With the knowledge of the field pattern, the aperture field function can be calculated. This is the base of holographic surface measurement in radio antennas (Section 8.4.1). For deriving higher spatial resolution in the aperture field, more sampling points of the field pattern are required.

To avoid a complex number problem, the optical transfer function can be replaced by the modulation transfer function in some cases. For a circular aperture, the modulation transfer function has the form (Wetherell, 1974, 1980):

$$\begin{aligned} MTF(v_n) &= \frac{2}{\pi}[\arccos v_n - v_n \sin(\arccos v_n)] \\ &= \frac{2}{\pi}\left[\arccos v_n - v_n\sqrt{1 - v_n^2}\right] \end{aligned} \tag{1.108}$$

For a circular aperture with a blockage ratio of ε, the following relationship applies (Figure 1.42):

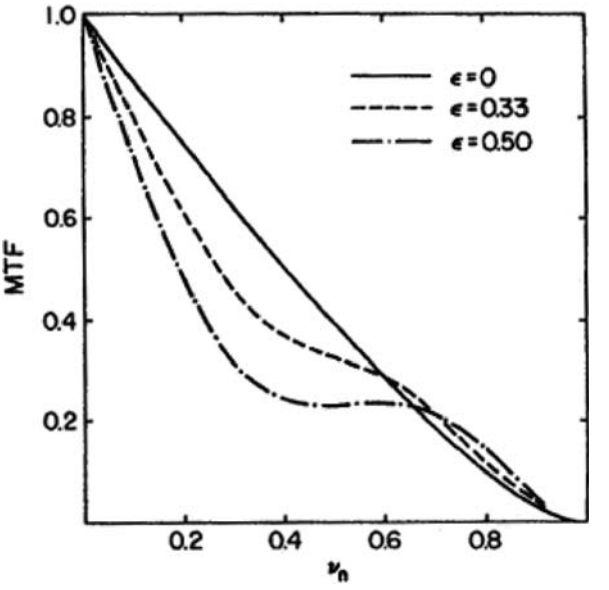


Fig. 1.42. The modulation transfer functions of a circular aperture with different blockage ratios (Wetherell, 1980).

$$
\begin{aligned}
MTF(v_n) &= \frac{2}{\pi}\frac{A+B+C}{1-\varepsilon^2} \\
&\quad 0 \le v_n \le \varepsilon \\
&B = \varepsilon^2[\arccos(v_n/\varepsilon) - (v_n/\varepsilon)\sqrt{1-(v_n/\varepsilon)^2}] \quad 0 \le v_n \le \varepsilon \\
&B = 0 \quad v_n > \varepsilon \\
&C = -\pi\varepsilon^2 \quad 0 \le v_n \le (1-\varepsilon)/2 \\
&C = -\pi\varepsilon^2 + \left\{\varepsilon \sin X + \frac{X}{2}(1+\varepsilon^2) - (1-\varepsilon^2)\tan^{-1}\left[\frac{1+\varepsilon}{1-\varepsilon}\tan\frac{X}{2}\right]\right\} \\
&\quad 1-\varepsilon \le 2v_n \le 1+\varepsilon \\
&C = 0 \quad 2v_n > 1+\varepsilon \\
&X = \arccos\frac{1+\varepsilon^2-4v_n^2}{2\varepsilon}
\end{aligned}
\tag{1.109}
$$

1.4.2 Wave Aberrations and Modulation Transfer Function

The advantage of using the modulation transfer function is that the system modulation transfer function is simply a product of individual component modulation transfer functions:

$$MTF = MTF_1 \cdot MTF_2 \cdot MTF_3 \cdot MTF_4 \cdots \tag{1.110}$$

If there are a number of independent factors in an optical system, which influence the image performance, then the system modulation transfer function is also a product of the individual modulation transfer functions of these factors:

$$MTF = MTF_d \cdot MTF_w \cdot MTF_r \cdot MTF_p \cdots \tag{1.111}$$

where subscript d is the aperture, w the geometrical aberration, r the randomly distributed wavefront error, and p the pointing error. The first term MTF_d is the modulation transfer function of an ideal optical system and the final modulation transfer function of the system is the product of this ideal function with a number of degradation functions.

Geometrical aberrations can be expressed as wavefront phase errors in the aperture field. Therefore, the modulation transfer functions of these aberrations can be calculated through aperture field auto-correlation. Figure 1.43 illustrates the modulation transfer functions of some geometrical aberrations. Figure 1.44 shows the modulation transfer functions of a circular aperture plus defocusing errors. In this figure, negative values appear for some parts of the modulation transfer function. This negative value is called spurious resolution. In this case, a

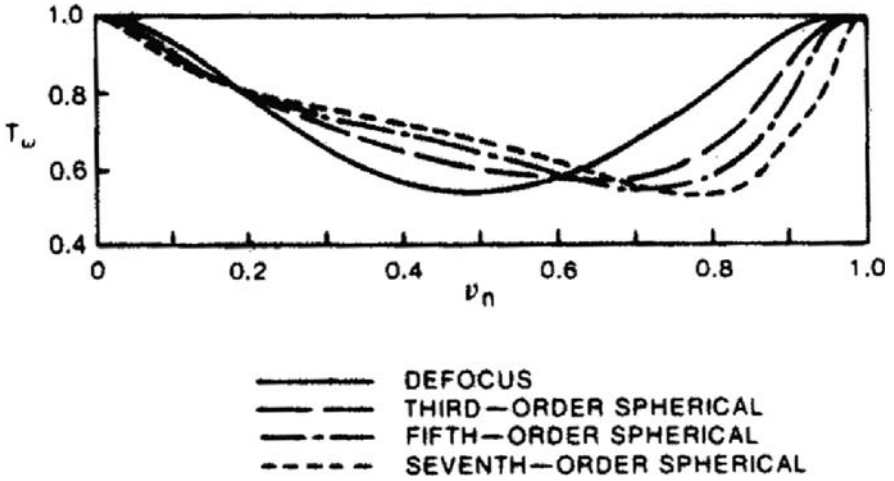


Fig. 1.43. The modulation transfer function of some geometrical aberrations (Wetherell, 1980).

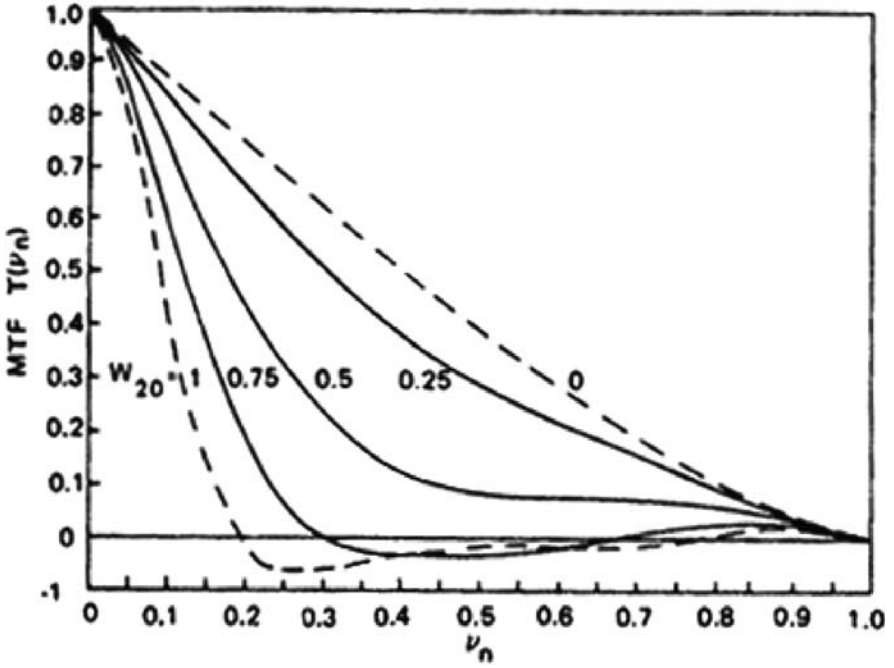


Fig. 1.44. The modulation transfer function of circular aperture plus defocusing errors (Wetherell, 1980).

180° phase delay exists, so that the sign of the amplitude is reversed. For example, if an object has four dark lines between five bright lines, the image will have four bright lines between five dark ones.

All aberrations in Figure 1.43 are axially symmetrical. The modulation transfer function is expressed by only one curve for each one. However, some aberrations are axially asymmetrical so that more curves are needed for the modulation transfer functions. Figure 1.45 shows the modulation transfer function curves of coma and astigmatism.

If lower order aberrations are removed, the wavefront error left is a high spatial frequency, randomly distributed ripple error. These spatial frequencies are usually larger than five cycles per aperture radius. As an example, the primary mirror of the Hubble Space Telescope has high frequency surface errors with 1 mm periodicity, a frequency of 1,200 cycles on the 1.2 m radius. O'Neill (Wetherell, 1980) made statistical analysis of the ripple error effect on the modulation transfer function as:

$$MTF_m = \exp[-k^2\omega_m^2(1 - C(v))] \tag{1.112}$$

where $k = 2\pi/\lambda$, ω_m the rms wavefront error at the central spatial frequency, and $C(v)$ the auto-correlation function of the residual wavefront error over a normalized aperture field. In general, the auto-correlation function of random wavefront error is $C(v) = \exp[-4v_n^2/l^2]$, a Gaussian function, where v_n is the normalized spatial frequency, l the correlation length normalized to the mirror diameter, and the reciprocal of l the spatial frequency of these ripples. The modulation transfer function or the degradation function of this ripple error is shown in Figure 1.46.

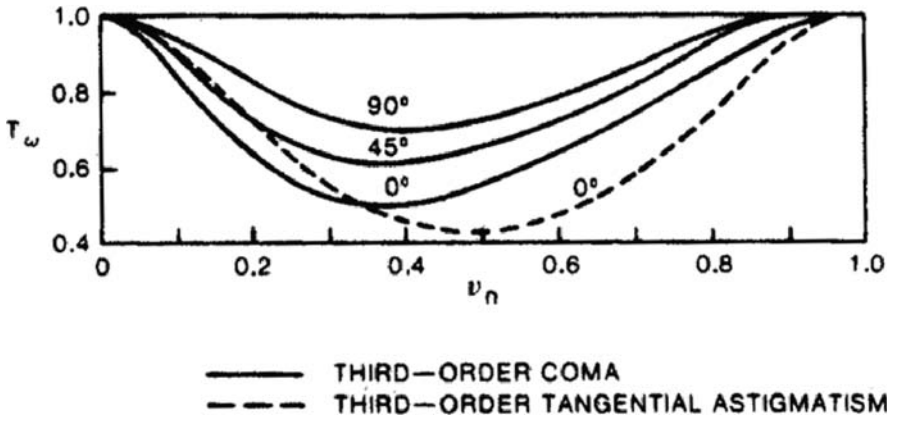


Fig. 1.45. The modulation transfer function of coma and astigmatism (*dotted line*) (Wetherell, 1980).

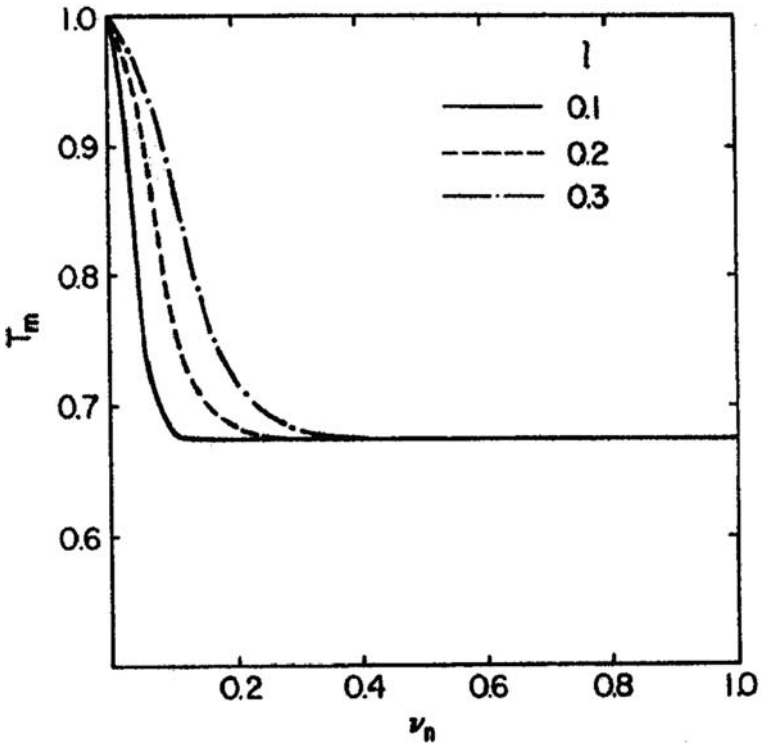


Fig. 1.46. The degradation functions of ripple errors with different correlation length l (Schroeder, 2000).

The errors with even higher spatial frequencies are referred to as micro ripple. In this case, the normalized correlation length $l \to 0$ and the modulus transfer function is:

$$MTF_h = \exp[-k^2\omega_h^2] \tag{1.113}$$

The product of MTF_m and MTF_h is the modulus transfer function of all random ripple errors.

Another type of wavefront error is the angular deviation of wavefront which produces random pointing changes. This type of error can be expressed in an axially symmetrical format:

$$P_r = \exp[-\alpha^2/2\sigma'^2] \tag{1.114}$$

where σ' is the standard deviation of pointing error and α the mean pointing direction both in angular units. To normalize σ', use $\sigma = \sigma' D/\lambda$. Then the transfer function of a random pointing error is:

$$MTF_p(\nu_n) = \exp[-2\pi^2\sigma^2\nu_n^2] \tag{1.115}$$

From this expression, it follows that the modulation transfer function of a random pointing error will be reduced as the spatial frequency increases. Therefore, the random pointing error makes the energy distribution more uniformly spread over the image blur. It eliminates all high frequency image details.

The modulation transfer function for the aperture blockage is complex (Figure 1.47). If $v_n \geq (1-\varepsilon)/2$, then

$$MTF_{\varepsilon}(v_n) = (1-\varepsilon^2)^{-1} \tag{1.116}$$

This shows that the high frequency value of the transfer function may be larger than unity. In a mathematical sense, the resulting image is a struggle between two Bessel functions from the aperture and blockage. The energy in the image is transferred from one ring to another so that high frequency details are amplified instead of decreased.

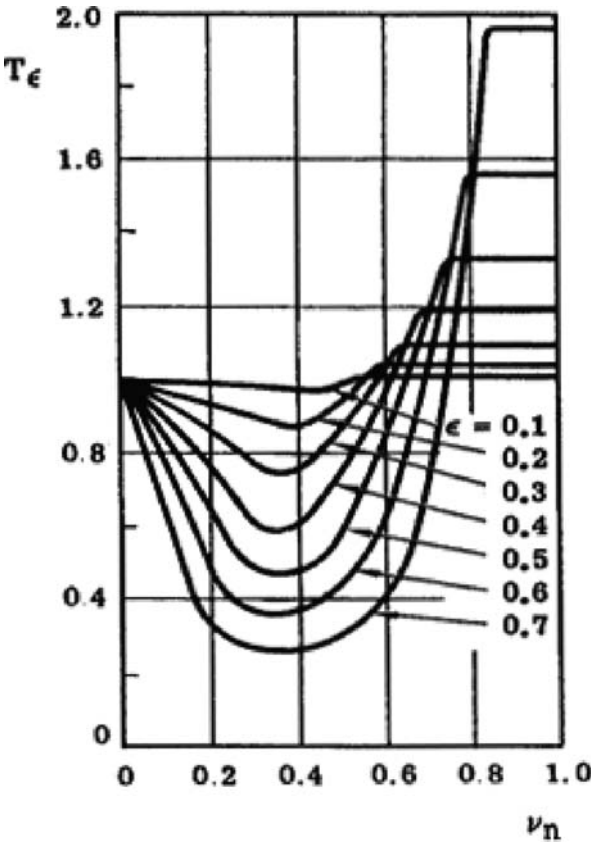


Fig. 1.47. The modulation transfer function of central blockage (Wetherell, 1980).

1.4.3 Wavefront Error and the Strehl Ratio

The Strehl ratio is defined as the energy ratio between the maxima of observed image of a system in the presence of aberration and an ideal diffraction limited image from a point source. The Strehl ratio can be calculated from the MTFs of the systems with and without aberration. It is the area ratio under the MTF curves of systems with and without aberrations:

$$St = \frac{\int MTF(v)dv}{\int MTF_i(v)dv} \tag{1.117}$$

where MTF_i is the MTF of an ideal system and MTF is that of the system with aberrations. Since the field pattern, which is (approximately) the square root of the point spread function, is a Fourier transform of the aperture field, the Strehl ratio can also be represented as a square of Fourier transform of the aperture field at zero spatial frequency point (i.e. $\exp(-2\pi i(v_x x + v_y y)) = 1$):

$$St = \frac{1}{\pi^2}\left|\int_0^{2\pi}\int_0^1 \exp(2\pi \cdot iW(\rho,\theta))\rho d\rho d\theta\right|^2 \tag{1.118}$$

where $W(\rho,\theta)$ is the phase on the aperture plane. The Strehl ratio can be derived using Taylor expansion of the exponential function in the above equation as:

$$St \cong \exp\left[-\left(\frac{2\pi}{\lambda}\sigma\right)^2\right] \approx 1 - \left(\frac{2\pi}{\lambda}\sigma\right)^2 \tag{1.119}$$

where σ is the rms path length error.

A wavefront rms error of $\lambda/14$ will produce a Strehl ratio of 0.8. If the wavefront rms error is small ($\sigma < \lambda/\pi$), one is in a weak aberration regime and the central core of the point spread function remains essentially diffraction limited. With the reduction of the image maximum in the center, the image blur angle is enlarged (Dalrymple, 2002):

$$\theta = \frac{\theta_D}{\sqrt{St}} \cong 2.4\frac{\lambda}{D}\sqrt{e^{\phi^2} - 1} \tag{1.120}$$

where θ_D is the diffraction limited image angle, ϕ the phase error, $\theta_D \cong 2.4\lambda/D$. Using this enlarged image angle, the Strehl ratio can be represented as:

$$St = \left(\frac{\theta_D}{\theta}\right)^2 \tag{1.121}$$

When the wavefront error is large ($\sigma \geq \lambda/\pi$) due to atmospheric turbulence, one is in a strong aberration regime and the central core of point spread function is obscured. The energy redistribution will fill up the null regime of the Airy disk. The signal is made up entirely of scattered energy and system noise. The system modulation transfer function is:

$$MTF = MTF_i \exp\left[-\left(\frac{2\pi\sigma \cdot R}{\lambda l_z}\right)^2\right] \tag{1.122}$$

where MTF_i is the modulation transfer function of the system without wavefront error, l_z the correlation length along the optical path (refer to Section 2.4.1), and R a dimension related to the aperture. Its Fourier transform, the point spread function, becomes a Gaussian:

$$\begin{aligned} PSF(r) &= \frac{1}{2\pi\xi^2}\exp\left[-\left(\frac{r}{\sqrt{2}\xi}\right)^2\right] \\ \xi &= \frac{\sqrt{2}\sigma F}{l_z} \end{aligned} \tag{1.123}$$

where r is the radius in the image plane and F the system focal length. The blur angle which contains p% of the encircled energy is:

$$\theta_{p\%} = \frac{4\sigma}{l_z}\sqrt{-\ln(1-p)} \tag{1.124}$$

For example, $\theta_{50\%} = 3.33\sigma/l_z$. The Strehl ratio in this case is $St \approx (l_z/\sigma)^2$. The wavefront correlation length plays an important role in the expression. As a limiting case, if the correlation length is the aperture dimension, a tilt of the image is produced.

1.4.4 Image Spatial Frequency

In the section, most formulas and discussions are presented, for simplicity, in a one-dimensional space. In the real situation, both the aperture and image are two-dimensional. The derivation of all the formulas is exactly analogous in the two-dimensional case. The far field pattern $A(\sin\phi)$ and aperture field function $P(x_\lambda)$ are a Fourier pair as (Kraus, 1986):

$$A(\sin\phi) = \int_{-\infty}^{\infty} P(x_\lambda) \cdot e^{i2\pi \cdot x_\lambda \sin\phi} dx_\lambda$$
$$P(x_\lambda) = \int_{-\infty}^{\infty} A(\sin\phi) \cdot e^{-i2\pi \cdot x_\lambda \sin\phi} d\sin\phi \tag{1.125}$$

where $x_\lambda = x/\lambda$ is the one-dimensional spatial frequency. The sine function of a small angle can be replaced by the angle ϕ as:

$$A(\phi) = \int_{-a_\lambda/2}^{+a_\lambda/2} P(x_\lambda) \cdot e^{i2\pi \cdot x_\lambda \sin\phi} dx_\lambda \tag{1.126}$$

where $a_\lambda = a/\lambda$ is the aperture cutoff frequency (Figure 1.48).

The optical transfer function is an auto-correlation of the aperture field:

$$OTF(x_{\lambda 0}) = \int_{-a\ \lambda/2}^{+a_\lambda/2} P(x_\lambda - x_{\lambda 0}) P^*(x_\lambda) dx_\lambda \tag{1.127}$$

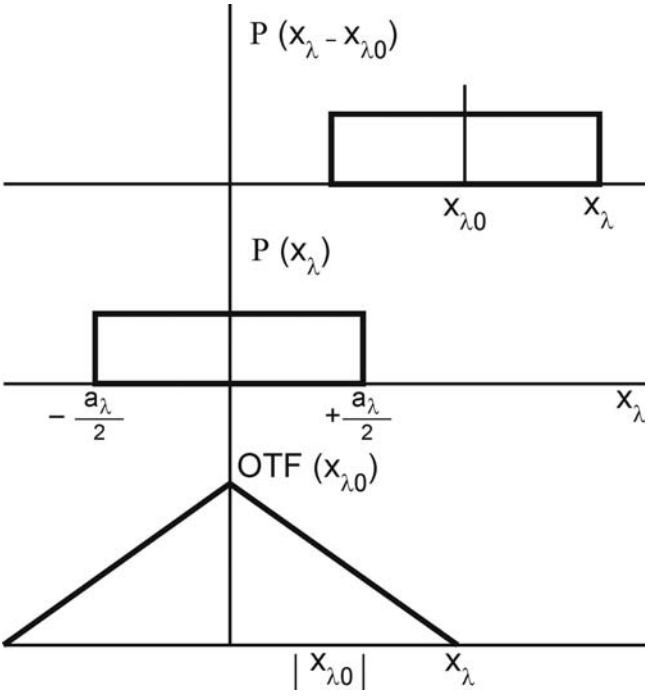


Fig. 1.48. The auto-correlation of the complex aperture function (Kraus, 1986).

where $OTF(x_{\lambda 0})$ is the optical transfer function [Figure 1.48(c)] and $P(x_\lambda)$ the aperture field function [Figure 1.48(b)]. For spatial frequency $x_{\lambda 0}$ larger than a_λ, the auto-correlation is zero. If the source brightness distribution is $B(\phi_0)$, then the image in the spatial frequency domain is the product of Fourier transform of the source brightness and the optical transfer function:

$$\bar{S}(x_\lambda) = \bar{B}(x_\lambda) OTF(x_{\lambda 0}) \tag{1.128}$$

This formula shows that a telescope as an optical system is a spatial frequency filter. Higher spatial frequency information is filtered out and only the lower frequency information is retained in its image.

To expand the spatial frequency beyond the cutoff one limited by the aperture size, interferometer is necessary. If two apertures are used with a separation of s_λ (Figure 1.49), the cutoff spatial frequency becomes:

$$x_{\lambda c} = a_\lambda + s_\lambda \tag{1.129}$$

Larger cutoff spatial frequency allows more information retained in the image. The normalized far field pattern of an interferometer is (Figure 1.50):

$$A(\phi) = A_n(\phi) \cos(\pi s_\lambda \sin\phi) \tag{1.130}$$

where $A_n(\phi)$ is the normalized far field pattern of the individual aperture element. The point spread function or the power pattern of the interferometer is:

$$|A(\phi)|^2 = |A_n(\phi)|^2 \cos^2(\pi s_\lambda \sin\phi) = |A_n(\phi)|^2 [1 + \cos(2\pi s_\lambda \sin\phi)]/2 \tag{1.131}$$

This formula shows that the image formed by a one-dimensional interferometer [Figure 1.51(c)] involves fringes with bright and dark lines [Figure 1.51(b)]

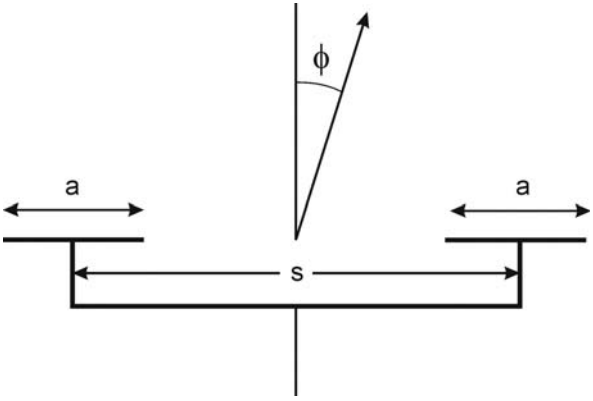


Fig. 1.49. The interferometer with two apertures.

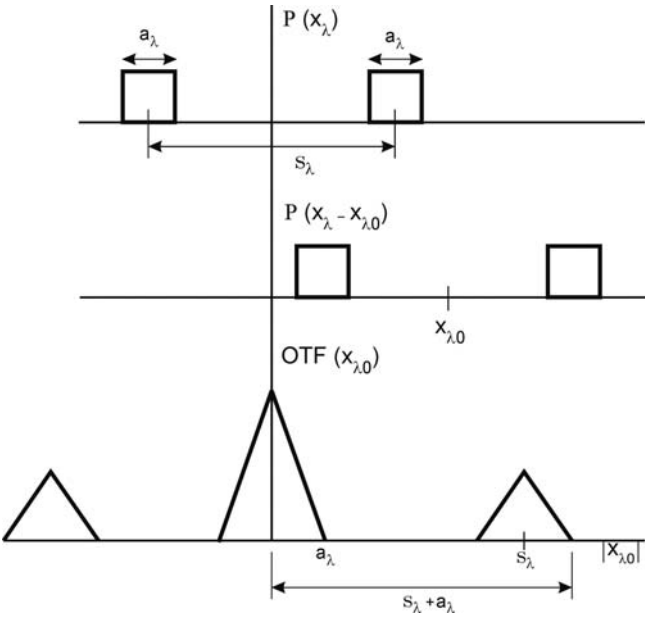


Fig. 1.50. The auto-correlation of the field of two separated apertures (Kraus, 1986).

modulated by the point spread function of each aperture [Figure 1.51(a)]. The fringe separation is the reciprocal of s_λ.

If a celestial source is observed by an interferometer, the image intensity is the convolution of the point spread function and the source brightness distribution (Kraus, 1986):

$$\begin{aligned} S(\phi_0, s_\lambda) &= |A_n(\phi)|^2 \int_{-\alpha/2}^{\alpha/2} B(\phi)[1 + \cos(2\pi s_\lambda \sin(\phi_0 - \phi))]d\phi \\ &= |A_n(\phi)|^2 \left\{ S_0 + \int_{-\alpha/2}^{\alpha/2} B(\phi) \cos(2\pi s_\lambda \sin(\phi_0 - \phi))d\phi \right\} \end{aligned} \tag{1.132}$$

where α is the source angular dimension, S_0 the source flux density, and ϕ_0 the angle in the image plane. For a uniform brightness source, if the source size is far smaller than the reciprocal of s_λ, then the image pattern is the same as the point spread function of the interferometer [Figure 1.52(a)]. If the source size is slightly smaller than the reciprocal of s_λ, then the contrast of the image fringes is reduced [Figure 1.52(b)]. If the source size equals the reciprocal of s_λ, then the

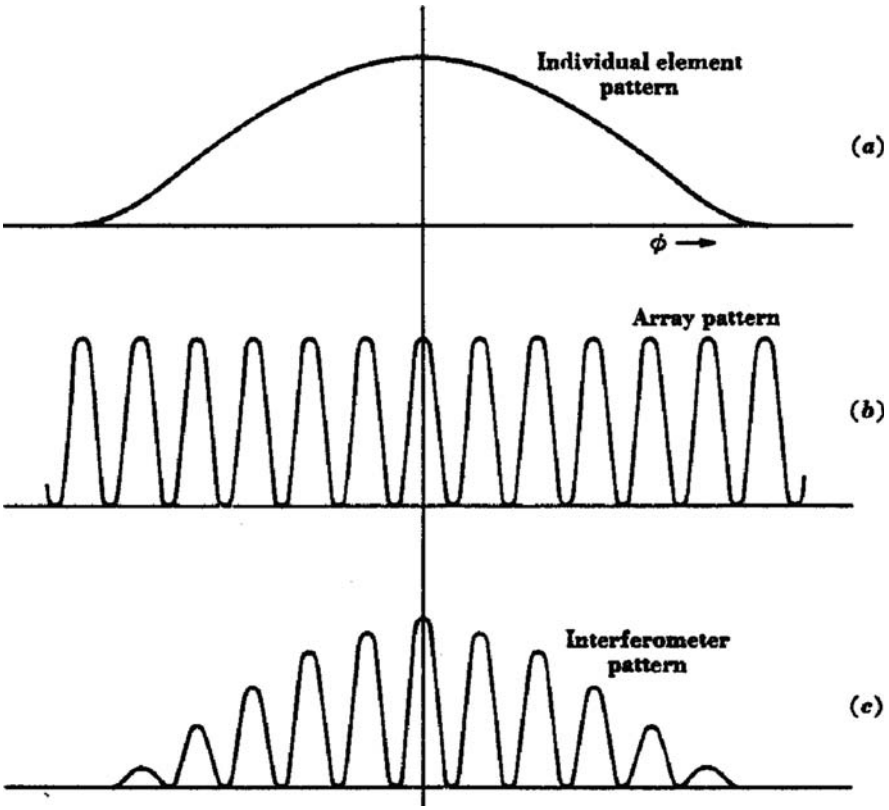


Fig. 1.51. The point spread function of a single aperture (**a**), of a double slit interferometer (**b**), and of a two element interferometer (**c**) (Kraus, 1986).

fringes of the image disappear [Figure 1.52(c)]. In the above formula, the first term is a constant and the second term is:

$$V(\phi_0, s_\lambda) = \frac{1}{S_0} \int_{-\alpha/2}^{\alpha/2} B(\phi) \cos(2\pi s_\lambda \sin(\phi_0 - \phi)) d\phi$$

$$= \frac{1}{S_0} \left[\cos 2\pi s_\lambda \phi_0 \int_{-\alpha/2}^{\alpha/2} B(\phi) \cos 2\pi s_\lambda \sin \phi d\phi + \right. \qquad (1.133)$$

$$\left. \sin 2\pi s_\lambda \phi_0 \int_{-\alpha/2}^{\alpha/2} B(\phi) \sin 2\pi s_\lambda \sin \phi d\phi \right]$$

This variable term can also be expressed as a cosine function with a displacement $\Delta\phi_0$:

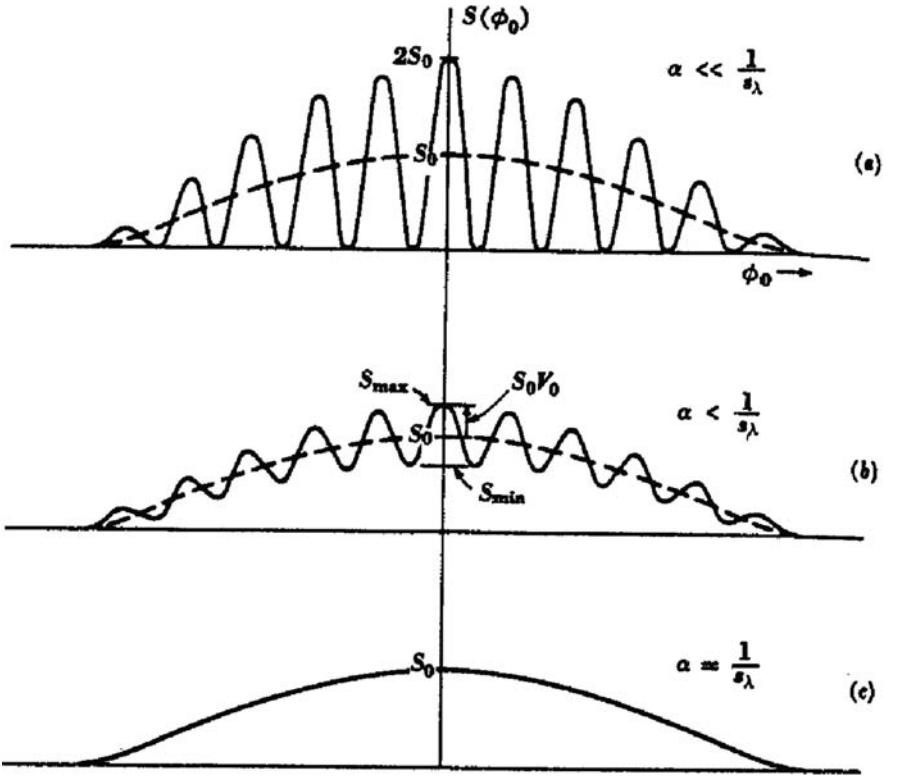


Fig. 1.52. The interferometer pattern (**a**) of a point source; (**b**) of a uniform extended source of angle $\alpha < 1/s_\lambda$; and (**c**) of a uniform extended source of angle $\alpha = 1/s_\lambda$ (Kraus, 1986).

$$\begin{aligned} V(\phi_0, s_\lambda) &= V_0(s_\lambda)\cos(2\pi s_\lambda(\phi_0 - \Delta\phi_0)) \\ &= V_0(s_\lambda)[\cos 2\pi s_\lambda\phi_0 \cos 2\pi s_\lambda\Delta\phi_0 + \sin 2\pi s_\lambda\phi_0 \sin 2\pi s_\lambda\Delta\phi_0] \end{aligned} \tag{1.134}$$

where the term $V_0(s_\lambda)$, which represents the amplitude of the fringe pattern, is named as the visibility function, a function of the baseline between two apertures of an interferometer measured in wavelength. The angle represents the fringe displacement from the position with a point source. From above equations, we have:

$$\begin{aligned} V_0(s_\lambda)\cos 2\pi s_\lambda\Delta\phi_0 &= \frac{1}{S_0}\int_{-\alpha/2}^{\alpha/2} B(\phi)\cos 2\pi s_\lambda\phi d\phi \\ V_0(s_\lambda)\sin 2\pi s_\lambda\Delta\phi_0 &= \frac{1}{S_0}\int_{-\alpha/2}^{\alpha/2} B(\phi)\sin 2\pi s_\lambda\phi d\phi \end{aligned} \tag{1.135}$$

Therefore:

$$V_0(s_\lambda)e^{j2\pi s_\lambda \Delta\phi_0} = \frac{1}{S_0}\int_{-\alpha/2}^{\alpha/2} B(\phi)e^{j2\pi s_\lambda \phi}d\phi \tag{1.136}$$

where $V_0(s_\lambda)\exp(j2\pi s_\lambda \Delta\phi_0)$ is called the complex visibility function. This formula is also valid for extended sources. Therefore, in a general form:

$$V_0(s_\lambda)e^{j2\pi s_\lambda \Delta\phi_0} = \frac{1}{S_0}\int_{-\infty}^{\infty} B(\phi)e^{j2\pi s_\lambda \phi}d\phi \tag{1.137}$$

This formula shows that the source bright distribution is the Fourier transform of the complex visibility function. If visibility is available for different spatial frequencies, the inverse Fourier transform of it provides the source brightness distribution:

$$B(\phi_0) = S_0\int_{-\infty}^{\infty} V_0(s_\lambda)e^{-j2\pi s_\lambda(\phi_0-\Delta\phi_0)}ds_\lambda \tag{1.138}$$

For symmetrical source distributions, the fringe displacement is zero or at the one-half fringe ($\Delta\phi_0 = s_\lambda/2$) position.

In correlation interferometers, the complex visibility function related to each baseline is collected and the source brightness is derived through Fourier transform. In the spatial frequency domain, the response at zero is unity for a single aperture telescope and the response reduces as the frequency increases (Figure 1.53). For the correlation interferometer, the response is band limited. The peaks are related to different baselines. However, there is no response at the zero spatial frequency. Therefore, information on large scale of the object is lost. The concept of correlation interferometer is discussed in Section 7.3.1.

In a two-dimensional case, the visibility is a function of two variables. So the source intensity is:

$$B(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{-j2\pi(ux+vy)}dudv \tag{1.139}$$

where $u = s_{\lambda x}$ and $v = s_{\lambda y}$ are the separations in x and y directions and $x = \cos\alpha$ and $y = \cos\beta$ are directional cosines. The $u-v$ plane, which is perpendicular to the source direction, is an important concept in the aperture synthesis telescopes. For proper Fourier transformation, a good $u-v$ coverage is essential. The theory of aperture synthesis telescopes is discussed in Section 7.3.2.

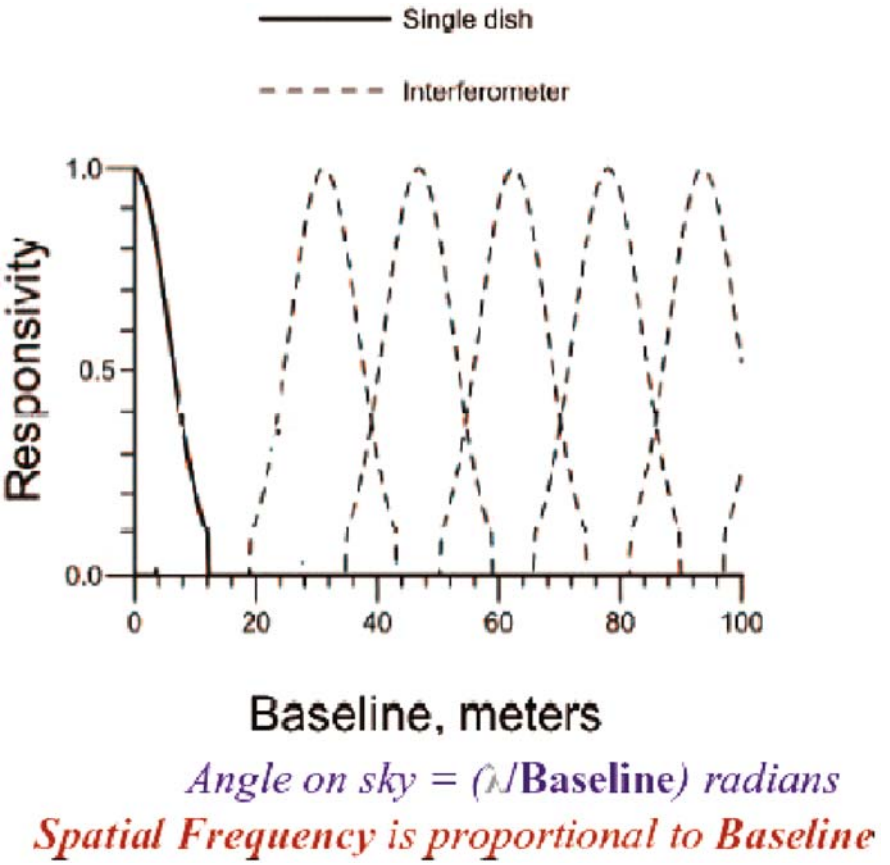


Fig. 1.53. The responsivity of a single dish and a correlation interferometer with different baselines (Emerson, 2005).

The above discussion is for monochromatic light. If a wide waveband is used, the visibility function will be modulated by another function of $\sin(\pi\tau\Delta(c/\lambda))/(\pi\tau\Delta(c/\lambda))$ and the visibility reduces rapidly, where τ is the time delay of the beams and $\Delta(c/\lambda)$ is the source bandwidth. The waveband effect and the spatial and temporal coherence theories are discussed in Section 4.2.4. The important Weiner–Khinchin and Van Cittert–Zernike theorems are discussed in Section 7.3.3.

1.4.5 Image Property of a Segmented Mirror System

Segmented and other primary mirror arrangements are discussed in Chapter 2. The point spread function of a perfectly segmented primary mirror is (Yaitskova et al., 2003):

$$\begin{aligned} PSF(w) &= \left(\frac{AN}{\lambda F}\right)^2 \left|\frac{1}{N}\sum_{j=1-N} \exp\left(i\frac{2\pi}{\lambda F}\vec{w}\cdot\vec{r}\right)\right|^2 \\ &\times \left|\frac{1}{A}\int \theta(\xi)\exp\left(i\frac{2\pi}{\lambda F}\vec{w}\cdot\vec{\xi}\right)d^2\vec{\xi}\right|^2 \\ &= \left(\frac{AN}{\lambda F}\right)^2 GF(w)\cdot PSF_{seg}(w) \end{aligned} \tag{1.140}$$

where $\vec{w}$ is a position vector in the image plane, $\vec{\xi} = \vec{x} - \vec{r}_j$ a local vector of each segment, for particular hexagonal segments, $\theta_j(\vec{x} - \vec{r}_j) = 1$ means inside the segment aperture and $\theta_j(\vec{x} - \vec{r}_j) = 0$ outside the segment aperture, $A = \sqrt{3}d^2/2$ the segment mirror area, d the separation between two segments, N the number of segments, λ the wavelength, F the focal length, PSF_{seg} the point spread function of a segment, and GF a grid factor, i.e. a Fourier transformation of the segmentation grid, usually a periodic array of sharp peaks. The functions, GF and PSF_{seg}, are always equal to unity for the position of $w = 0$ on the image plane. Usually, other zeros of the segment point spread function coincide with the peaks of the GF term, so that only the central peak is observed.

If hexagonal segments are arranged in M concentric bands around the central segment, the total number of segments is $N = 3\,M(M+1)+1$. Then,

$$\begin{aligned} GF(w) = &\Big\{\sin[(3\,M+1)\beta + (M+1)\sqrt{3}\alpha] \\ &\times \frac{\sin[M(\beta-\sqrt{3}\alpha)]}{N\sin(2\beta)\sin(\beta-\sqrt{3}\alpha)} + \sin[(3\,M+2)\beta - M\sqrt{3}\alpha] \\ &\times \frac{\sin[(M+1)(\beta+\sqrt{3}\alpha)]}{N\sin(2\beta)\sin(\beta+\sqrt{3}\alpha)}\Big\}^2 \end{aligned} \tag{1.141}$$

where $\alpha = (\pi d/2\lambda F)w_x$ and $\beta = (\pi d/2\lambda F)w_y$ are normalized coordinates in the image plane, d the distance between centers of segments, F the focal length, and λ the wavelength. The function shows a double $\pi/3$ symmetry in the image plane.

The segment field pattern function is (Yaitskova et al., 2003):

$$\begin{aligned} FPF(w) = \frac{1}{2\sqrt{3}\alpha}\Big[&\sin(\sqrt{3}\alpha-\beta)sinc(\alpha/\sqrt{3}+\beta) + \\ &+\sin(\sqrt{3}\alpha+\beta)sinc(\alpha/\sqrt{3}-\beta)\Big] \end{aligned} \tag{1.142}$$

Along two orthogonal directions in the image plane, the segment point spread function is:

$$\begin{aligned} PSF_{seg}(\alpha) &= \frac{\sin(\alpha/\sqrt{3})\sin(\alpha\sqrt{3})}{\alpha^2} \\ PSF_{seg}(\beta) &= \frac{1-\cos(2\beta)+2\beta\sin(2\beta)}{6\beta^2} \end{aligned} \tag{1.143}$$

If random piston errors with a zero mean exist between segments, the segment point spread function remains unaffected. However, the function GF is changed by introducing a noisy speckle background. The angular size of the main speckle field is defined as the full width at half maximum (FWHM) of the segment point spread function PSF_{seg} and is independent of the piston error. It is (Yaitskova et al., 2003):

$$\varepsilon_{speckle} \approx 2.9\lambda/\pi d \tag{1.144}$$

For $d = 1.5$ m and $\lambda = 0.5$ μm, this angle is 0.07". The appearance of speckle brings a reduction of the Strehl ratio. The expression of the Strehl ratio for an instant realization of random piston phase errors δ_j is given by:

$$St = \frac{1}{N}\left[1 + \frac{2}{N}\sum_{j>1}^{N}\cos(\delta_j - \delta_1)\right] \tag{1.145}$$

If all piston errors are independent and have a Gaussian distribution with a zero mean, then it becomes:

$$St = \frac{1}{N}\left[1 + (N-1)e^{-\phi^2}\right] \tag{1.146}$$

where ϕ is the rms wavefront phase error on the aperture field. The ratio of the average intensity to the central image peak intensity is (Yaitskova et al., 2003):

$$Ra = \frac{1-e^{-\phi^2}}{1+(N-1)e^{-\phi^2}} \tag{1.147}$$

The situation of a segmented aperture with random segment tip-tilt errors is more complicated. The point spread function can also be expressed as a product of a few terms, i.e. an incoherent combination of the segment point spread functions and interferences between segments. For weakly coherent segmented mirrors with large and different tip-tilt errors, the interference terms between segments can be neglected. In this case, the observed point spread function is simply an addition of N spots of light reflected by each segment at a tilted angle.

However, as the number of segments increases, the second term contributes more and more to the point spread function. This case can also be regarded as a product of a grid function and a modified point spread function for the segment. This modified point spread function PSF'_{seg} of a segment is a convolution of the segment field pattern function (Equation 1.142) with a Q function which depends on the statistics of the tip-tilt error distribution.

$$PSF'_{seg}(\vec{w},\phi) = \left| \int FPF(\vec{w}')Q(\phi,\vec{w}-\vec{w}')d^2\vec{w}' \right|^2 \tag{1.148}$$

For a normalized Gaussian distribution of the tip-tilt errors, the Q function is:

$$\begin{aligned} Q(\phi,\vec{w}-\vec{w}') = &\left(\frac{2\pi}{\lambda F}\right)^2 \frac{d^2}{2\pi(2.7\phi)^2} \\ &\times \exp\left[-\left(\frac{2\pi}{\lambda F}\right)^2 \frac{(\vec{w}-\vec{w}')^2 d^2}{2(2.7\phi)^2}\right] \end{aligned} \tag{1.149}$$

and the Strehl ratio in this case, which is not strongly dependent on the number of segments, is (Yaitskova et al., 2003):

$$St(\phi) \approx 1 - \phi^2 + \frac{\phi^2}{4}\left(2.34 + \frac{2}{N}\right) \tag{1.150}$$

References

Anderson, G. and Tullson, D., 2006, Photon sieve telescope, SPIE Proc., 6265, 626523.

Bahner, K., 1968, Large and very large telescope projects and consideration, ESO Bulletin, No. 5.

Barlow, B. V., 1975, The astronomical telescope, Wykeham Publications (London) Ltd, London.

Baum, W. A., 1962, The detection and measurement of faint astronomical sources, in Astronomical techniques, ed. Hiltner, WA,, Astronomical Techniques, Chicago.

Born, M. and Wolf, E., 1980, Principles of optics, 6th ed. Pergamon Press, Oxford.

Bowen, I. S., 1964, Telescopes, AJ, 69, 816.

Cao, C., 1986, Optical system for large field telescopes, Conference on large field telescope design, Nanjing Astronomical Instrument Institute, Nanjing.

Cheng, J., 1988, Field of view, star guiding and general design of large Schmidt telescope, Proceedings of ESO conference on VLT and their instruments, Munich, Germany.

Cheng, J. and Liang, M., 1990, High image quality Mersenne-Schmidt telescope, SPIE Proc. Adv. Technol. Telescope (IV), 1236, p243–249.

Dalrymple, N. E., 2002, Mirror seeing, ATST project CDR report #0003, NOAO.

Dawe, J. A., 1984, The determination of the vignetting function of a Schmidt telescope, in Astronomy with Schmidt telescopes, ed. Capaccioli, M, E. Reidel Pub. Co., Dordrecht.

Dierickx, P., et al., 2004, OWL phase A, status report, Proc. SPIE, 5489, 391.
Disney, M. J., 1972, Optical arrays, Mon. Not. RAS., 160, 213–232.
Disney, M. J., 1978, Optical telescope of the future, ESO Conf. Proc. 23, 145–163.
Emerson, D., 2005, Lecture notes of NRAO summer school on radio interferometry, National radio astronomy observatory.
Foy, R. and Labeyrie, A., 1985, Feasibility of adaptive telescope with laser probe, Astron. Astrophys., 152, L29.
Gascoigne, C. S. R., 1968, Some recent advances in the optics of large telescopes, Quart. J. RAS., 9, 18.
Gascoigne, C. S. R., 1973, Recent advances in astronomical optics, Appl. Opt., 12, 1419.
Glassner, A. S., 1989, An introduction to ray tracing, Academic Press, London.
Gramham Smith, F., and Thompson, J. H., 1988, Optics, 2nd edition, John Wiley & Sons Ltd., New York.
Hecht, H. and Zajac, A., 1974, Optics, Addison-Wesley Pub. Co, London.
Jiang, S. 1986, Review of multi-object spectroscope, Conference on large field telescope, Nanjing Astronomical Instrument Institute, Nanjing.
Kraus, J. D., 1986, Radio astronomy, Cygnus-Quasar Books, Powell, Ohio.
Learner, R., 1980, Astronomy through the telescope, Evans Brothers, London.
Liang, M., et al., 2005, The LSST optical system, Bull. Am. Astron. Soc., 37, 2005.
Lo, A. S. and Arenberg, J., 2006, New architectures for space astronomical telescopes using Fresnel optics, SPIE Proc., 6265, 626522.
Pawsey, J. L., Payne-Scott, R. and McCready, L. L., 1946, Radio frequency energy from the sun, Nature, 157, 158.
Racine, R., 1984, Astronomical seeing at Mauna Kea and in particular at the CFHT, IAU Colloq. No. 79, 235.
Reynolds, G. O., et al., 1989, The new physical optics notebook: tutorials in Fourier optics, SPIE Press,
Roddier, F., 1979, Effect of atmosphere turbulence on the formation of infrared and visible images, J. of optics, 10, 299–303.
Roddier, F., 1984, Measuring atmospheric seeing, in IAU Colloq. No. 79, eds. Ulrich MH and Kjar K, Garching bei Munchen, Germany.
Schnapf, J. L. and Baylor, D. A., 1987, How photoreceptor cells respond to light, Sci. Am., 256, 40–47.
Schroeder, D. J., 2000, Astronomical optics, Academic Press, San Diego.
Shao, L.-Z. and Su, D.-Q., 1983, Improvement of chromatic aberration of an aspherical plate corrector for prime focus, Opt. Acta, 30, 1267–1272.
Slyusarev, G. G., 1984, Aberration and optical design theory, 2nd ed. Adam Hilger Ltd., Bristol.
Steward, E. G., 1983, Fourier optics: an introduction, Ellis Horwood Limited, Chichester.
Stoltzmann, D. E., 1983, Resolution criteria for diffraction-limited telescopes, Sky Telescope, 65, 176–181.
Su, D.-Q., 1963, Discussion on corrector design for reflecting telescope system, Acta Astron, 11.
Su, D.-Q., et al., 1967, Automatic design of corrector system for Cassegrain telescopes, Acta Astron, 17.
Su, D.-Q. and Wang, Y.-L., 1974, Optimization of aberrations for astronomical optical system, Acta Astron, 15.
Su, D.-Q. and Wang, L.-J., 1982, A flat-field reflecting focal reducer, Opt. Acta, 29, 391–394.

Su, D.-Q., et al., 1983, Spot diagram and lest square optimization, Nanjing Astronomical Instrument Institute, Nanjing.

Vernin, J., 1986, Astronomical site selection, a new meteorological approach, SPIE Proc., 628, 142.

Wetherell, W. B., 1974, Image quality criteria for the Large Space Telescope, in Space optics, eds. Thompson B. J. and Shannon R. R., National Academy of Science, Washington.

Wetherell, W. B., 1980, The calculation of image quality, in Applied optics and optical engineering, Vol. 8, Academic Press, New York.

Willstroop, R. V., 1984, The Mersenne-Schmidt telescope, in IAU Colloq. No. 79, eds. Ulrich M. H. and Kjar K., Garching bei Munchen, Germany.

Wilson, R. N, 1968, Corrector systems for Cassegrain telescopes, Appl. Opt., 7, 253–263.

Wilson, R. N, 2004, Reflecting telescope optics I, 2nd ed. Springer, Berlin.

Wynne, C. G., 1967, Afocal correctors for Paraboloidal mirrors, Appl. Opt., 6, 1227–1231.

Yaitskova, N., et al., 2003, Analytical study of diffraction effects in extremely large segmented telescopes, J. Opt. Soc. Am. A, 20, 1563–1575.

Yi, M., 1982, Design of aspherical correctors for Cassegrain system, Acta Astron., 23, 398.

Chapter 2
Mirror Design For Optical Telescopes

The reflector mirror is the most important component of an astronomical optical telescope. This chapter provides discussions on the requirements for astronomical optical mirrors; the ways to reduce mirror weight, mirror cost, and mirror materials; the methods of mirror figuring, polishing, and surface coating; the design of mirror support mechanism; the concept of mirror seeing; and the stray light control. Emphasis is placed on various mirror designs for modern large optical telescopes. These include the thin mirror, honeycomb mirror, segmented mirror, and multi-mirror telescope concepts. When discussing all these concepts, important formulas and their restrictions are provided for the reader's reference so that they may use them in their mirror design practice. The discussion on the mirror support system is thorough and comprehensive, including both the positional and flotation support systems. A new mirror support system using a hexapod platform is also introduced. In the stray light control section, a new scattering theory based on the bidirectional reflectance distribution function is also introduced.

2.1 Specifications for Optical Mirror Design

2.1.1 Fundamental Requirements for Optical Mirrors

An optical astronomical telescope, as a very sensitive light collector, comprises a number of important components. Among these, the reflecting primary mirror is the most important. The telescope efficiency is directly related to its area, its reflectivity, and its surface accuracy. The mirror area and reflectivity have been discussed in Section 1.2.2. The mirror surface accuracy is related to wavefront errors which affect the image Strehl ratio. The image Strehl ratio and the wavefront error were briefly introduced in Section 1.4.3.

To obtain sharp star images, a rigorous tolerance is used for the mirror surface precision. The ideal primary mirror shape is determined through optical design, ray tracing, and system optimization. In geometrical optics, this ideal

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_2,

surface shape ensures a small acute star image spot in the focal plane. This corresponds to a perfect planar Gaussian wavefront on the aperture plane. However, mirror surface shape imperfection always exists due to the mirror manufacture, mirror support, thermal variation, and other reasons. The wavefront error is twice the mirror surface error due to the double reflection.

Generally, the characteristic mirror surface or wavefront error is expressed by the root mean square (rms) of the distance errors to an ideal mirror or wavefront surface. Statistically, the average value of the errors can be made equal to zero by choosing a best fit reference surface, and the rms then is the standard deviation of the error. The square of the rms error is the variance. The ratio between the rms and the peak error depends on the error distributions. For a uniform error distribution, the peak error is twice the rms value. For a triangular error distribution, the peak error is 3.46 times the rms. For a sine error distribution, the peak error is 2.83 times the rms. The peak of a finite sample from a Gaussian distribution is not fixed; being typically 6 to 8 times the rms. When more than one independent factor (in mathematics, independent error terms are orthogonal to each other) exists, the combined rms error is the root sum square (rss) of the rms errors of individual factors.

According to electromagnetic wave theory, if the wavefront deviates from an ideal one, the radiation energy of the image will be redistributed resulting in: (a) a decrease in image sharpness; (b) an increase in image size; and (c) a decrease in image central energy, and the Strehl ratio of the image decreases.

For an axial symmetrical aperture, the diffraction radiation energy distribution at a position P is:

$$I(P) = \left(\frac{Aa^2}{\lambda R^2}\right)^2 \left| \int_0^1 \int_0^{2\pi} e^{i(2\pi\phi/\lambda - v\rho\cos(\theta-\psi) - 1/2u\rho^2)} \rho d\rho d\theta \right|^2 \tag{2.1}$$

where A is the radiation amplitude on the aperture plane, ϕ the wavefront phase error, a the aperture radius, ρ and θ polar angles in the aperture plane, r and ψ polar angles in the image plane, z the axial distance between the aperture and the image, R the distance between image position P and the point to be integrated on the aperture plane, $u = (2\pi/\lambda)(a/R)^2 z$, and $v = (2\pi/\lambda)(a/R)r$.

Without wavefront aberrations, the maximum on-axis intensity is:

$$I_{\phi=0}(P_{r=0}) = \pi^2 \left(\frac{Aa^2}{\lambda R^2}\right)^2 = \left(\frac{\pi^2 A^4}{\lambda^2 R^2}\right) I_{z=0} \tag{2.2}$$

This image intensity is $[a^2/(\lambda R)]^2$ times stronger than the radiation intensity on the aperture plane. This amplification is named the Fresnel coefficient. The Fresnel coefficient indicates that larger aperture, shorter wavelength, and small, fast focal ratio produce a higher intensity image. The Strehl ratio of a practical system is given by the ratio of Equations (2.1) and (2.2).

After the first (piston) and second (tilt) aberration terms in a Taylor expression have been removed, the wavefront error becomes the difference between the practical wavefront and its best fit Gaussian one. The Strehl ratio is:

$$S = \frac{I(P)}{I_{\phi=0}(P_{r=0})} = \frac{1}{\pi^2}\left|\int_0^1 \int_0^{2\pi} e^{ik\phi_P}\rho d\rho d\theta\right|^2 \tag{2.3}$$

where ϕ_P is wavefront deviation away from its best fit Gaussian wavefront. If Φ^n represents the ensemble average of the n-th power of wavefront error ϕ:

$$\Phi^n = \frac{\int_0^1 \int_0^{2\pi} \phi^n \rho d\rho d\theta}{\int_0^1 \int_0^{2\pi} \rho d\rho d\theta} = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \phi^n \rho d\rho d\theta \tag{2.4}$$

The wavefront error variance $(\Delta\phi)^2$is:

$$(\Delta\phi)^2 = \frac{\int_0^1 \int_0^{2\pi} (\phi - \Phi)^2 \rho d\rho d\theta}{\int_0^1 \int_0^{2\pi} \rho d\rho d\theta} = \Phi^2 - (\Phi)^2 \tag{2.5}$$

If the rms wavefront error is $\Delta\phi < \lambda/2\pi$ and there is no correlation between errors in Nsub-apertures ($N$is a large number), the corresponding Strehl ratio is:

$$S = 1 - \left(\frac{2\pi}{\lambda}\right)^2 (\Delta\phi)^2 \cong \exp\left[-\left(\frac{2\pi}{\lambda}\right)^2 (\phi_P)^2\right] \tag{2.6}$$

This equation is the same as Equation (1.119). The wavefront error discussed is a small and randomly distributed one with very small correlation lengths, it has no repeatable pattern, and has a continuous first derivative (slope). In this case, the image intensity loss is independent upon the wavefront error details. However, if the wavefront error is large, the formula has error. The reader may reference Section 7.1.2 for a better understanding of this formula.

Since the mirror surface error is half of the wavefront error, Equation (2.6) provides an important criterion for the mirror surface requirement of an optical telescope system. Table 2.1 lists relative image intensities for different wavefront errors. Usually, a relative image intensity of 67% is acceptable; the corresponding rms mirror surface error allowed is, therefore, 1/20th of the wavelength.

The image size of a ground-based astronomical optical telescope without adaptive optics is limited by the site atmospheric seeing. To achieve maximum

Table 2.1. The relationship between wavefront rms error and relative image intensity

$\Delta\phi$	$\lambda/10$	$\lambda/12$	$\lambda/14$	$\lambda/16$	$\lambda/18$	$\lambda/20$	$\lambda/22$	$\lambda/24$
S	0.674	0.760	0.817	0.857	0.885	0.906	0.921	0.933

telescope efficiency, 90% of image energy should be within the best seeing disk and 80% of the received energy should fall within a diameter of $0.15''$ to $0.30''$; for telescopes without adaptive optics. For telescopes in space or with adaptive optics, details of the Airy disk can be resolved. The mirror surface rms error should be smaller than 1/40th of the wavelength. A target image accuracy of $0.02''$ may be required. These tolerances are very stringent, so that the mirror manufacture and support are demanding for space optical telescopes. In some publications, the Fried parameter, which is related to FWHM of image size, is used as the error tolerance specification. A Fried number of 60 cm is equivalent to a FWHM of 0.17 arcsec (Hill, 1995).

2.1.2 Mirror Surface Error and Mirror Support Systems

Mirror surface error comes from three major sources: mirror manufacture, mirror support, and other influences. Mirror manufacture produces a fixed surface error from polishing and testing, the mirror support system produces an elevation dependent surface error, and other influences include actuator error and wind and thermal induced errors. The elevation dependent errors are from the gravity force which varies with the telescope pointing. Two typical gravity directions relative to the mirror are the axial and radial ones. To balance these gravity components, mirror supports on both directions are necessary. The design of these mirror supports are discussed in this section.

2.1.2.1 Axial Support for Optical Mirrors

Mirror diameter-to-thickness ratio (d/t), also known as aspect ratio, is very important in mirror support design. The smaller the aspect ratio is, the heavier the mirror and the higher the costs are. Classical mirrors have their aspect ratios between six and eight. These thick mirrors are easy to support. However, their thermal and gravitational inertias bring trouble to telescope designers. The first large thin mirror used is in the UK Infrared Telescope (UKIRT) built in 1973, with an aspect ratio of 16. Afterwards, thin mirrors with larger and larger aspect ratios were used in astronomy. As the mirror aspect ratio becomes large, the mirror support system design becomes critical. The surface deformation is also more sensitive to the support conditions.

Surface deformation of thin mirrors under an axial support system can be predicted by using classical thin plate theory. Under the thin plate assumption, the deformation of a plate is approximately a function of plate diameter and

thickness, being called the scaling law in the telescope mirror support design. The scaling law states that the rms surface deformation of a mirror is proportional to the fourth power of the diameter and is inversely proportional to the square of the thickness when the support conditions are not changed. With this scaling law, if the deformation of one mirror is known, then the deformations of other mirrors under a similar support condition can be accurately predicted.

From the scaling law, wavefront rms error curves are drawn in Figure 2.1 for mirrors of different diameter and aspect ratio under different support ring conditions (Cheng and Humphries, 1982). Four sets of wavefront rms error curves represent one-ring, two-ring, three-ring, and four-ring axial support systems, respectively. The figure shows that mirrors with an aspect ratio of 15 can be reasonably well supported by a one-ring axial support system up to a diameter of 1.25 m; by a two-ring support system up to 2.25 m; by a three-ring support system up to 4.5 m; and by a four-ring support system up to ∼6 m. The proportionality constants in these curves are derived from the data produced by the ESO CAT telescope, the 4 m KPNO telescope, the 3.8 m UKIRT, and from

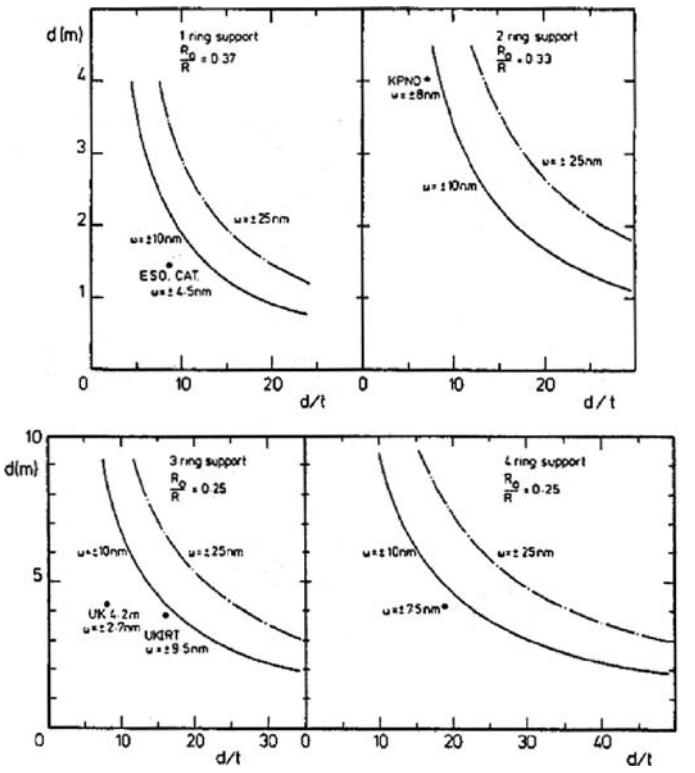


Fig. 2. 1. RMS wavefront errors for mirrors with different diameter and different aspect ratio on different rings of support system.

the finite element analysis (FEA). These curves can be used to estimate the surface rms errors for mirrors of different size and different aspect ratio. If the specification of a mirror is provided, the number of support rings required can be roughly predicted.

More accurately, the mechanical surface rms error, not wavefront rms error, of a thin plate under an axial support system is (Nelson et al., 1982):

$$\delta_{rms} = \xi \frac{q}{D} A^2 \tag{2.7}$$

where A is the plate area, q the areal density, and D the bending stiffness of the plate. The value of ξ reflects the support condition and is called the support efficiency. If there are N support points on a thin plate, then the average support efficiency of each point γ_N can be used in the expression of the rms surface error. The average support efficiency is:

$$\gamma_N = \xi N^2 \tag{2.8}$$

For a practical mirror support system, each support point may have its own support efficiency, resulting in a larger edge deformation as the edge support points may have lower support efficiency. The support efficiency of any point may be close to, but never reach an ideal value. This ideal value is the support efficiency when $A \to \infty$ and $N \to \infty$.

When $A \to \infty$ and $N \to \infty$, the plate deformation is determined by only two factors: the arrangement of the support points and the plate area related to each support point (A/N). Under this assumption, the ideal support efficiency of each support point can be found for three basic support point arrangements, i.e., triangular grid, square grid, and hexagonal grid (Figure 2.2). The deformation of these three grid supports can be found analytically through linear superposition. Therefore, the average support efficiency of each point γ_∞ for these three cases can be derived. In terms of the rms surface error, these average support efficiencies are (Nelson et al., 1982):

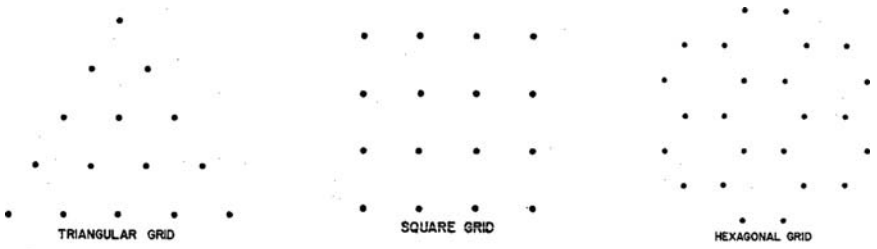


Fig. 2.2. Three basic grid arrangements: triangular, square, and hexagonal ones.

$$\begin{aligned} \gamma_{triangular} &= 1.19 \times 10^{-3} \\ \gamma_{square} &= 1.33 \times 10^{-3} \\ \gamma_{hexagonal} &= 2.36 \times 10^{-3} \\ \delta_{rms} &= \gamma_i \frac{q}{D} \left(\frac{A}{N}\right)^2 \end{aligned} \tag{2.9}$$

The above efficiencies are independent of the Poisson ratio of the plate material. If the maximum, or the peak surface deformation, is considered, the average support efficiencies of these three cases are respectively $4.95 \cdot 10^{-3}$, $5.80 \cdot 10^{-3}$, and $9.70 \cdot 10^{-3}$. The triangular grid has the highest support efficiency (or the lowest support efficiency number). This efficiency number can be set as a standard in the discussion of a thin mirror support system.

For circular thin plates on a multi-ring support system, the support points within a ring are at the same radius. The deformation of the plate is a superposition of deformations introduced by each support point. If the number of the support rings is $n\,(i = 1, 2, \cdots, n)$, the number of the support points of each ring is k_i, the weighting factor of each support ring is ε_i, and the angle between adjacent points in a ring is ϕ_i, then the surface rms deformation can be expressed as:

$$\delta_{rms}(r, \theta) = \sum_{i}^{n} \varepsilon_i \delta_i(k_i, \beta_i, r, \theta - \phi_i) \tag{2.10}$$

where β_i is relative support radius of ring i. By providing δ_i in Equation (2.10), all the terms are added after the weighting factor ε_i being considered. In the calculation, δ_i can be expressed in Zernike polynomial forms. If n is large, the calculation of δ_{rms} and the optimization of the support radius β_i are generally difficult.

The simplest case involves one support ring with only two variables: the relative radius of the support and the number of support points. Figure 2.3 shows the relationship between the relative support radius and the rms surface error for a one-ring support system.

If one support point is used, optimization of the radius is not necessary. The support efficiency is $\xi = \gamma_1 = 2.62 \cdot 10^{-3}$. If two support points are used, optimization of the radius is necessary. At the optimum radius of about 0.35, the support efficiency is $\xi = 2.16 \cdot 10^{-3}$ and the average support efficiency of each point is $\gamma_2 = 2^2 \cdot \xi$. The rms surface error decreases slightly. However, the support efficiency of each point decreases greatly. Figure 2.3 also shows the efficiencies for three, six, and more support points within one ring. When a continuous support ring is used, the optimum support radius is 0.683. The rms surface error is only 4% of that using an outer edge support. This efficiency increase demonstrates the importance of the mirror support system optimization.

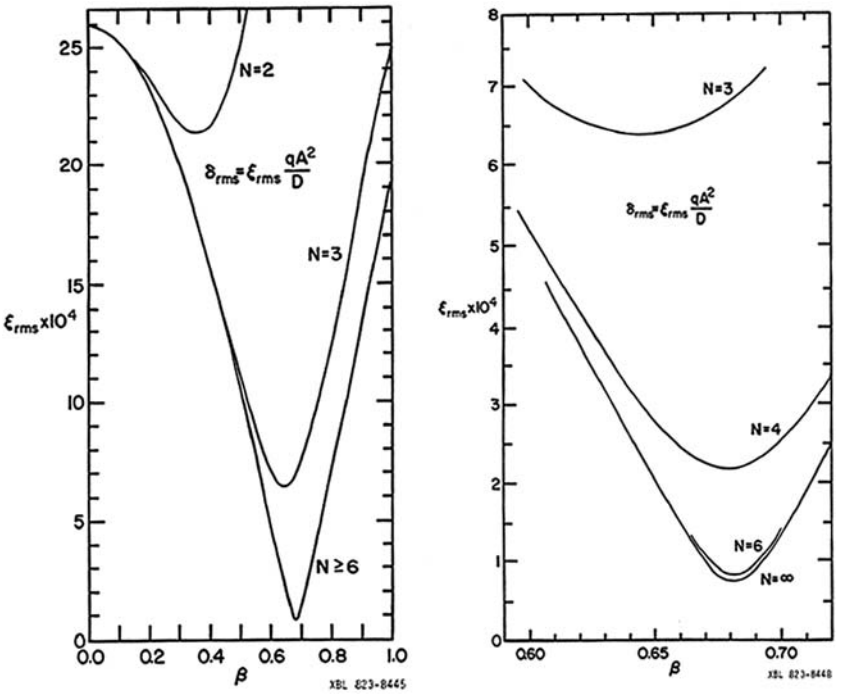


Fig. 2.3. RMS errors of the mirror surface as a function of relative support ring radius for different one ring support conditions and the figure on the right shows enlarged details (Nelson et al., 1982).

For two- or more-ring support systems, the plate surface deformation is also a function of Poisson ratio of the mirror material. Assuming the Poisson ratio is 0.25, by adding an additional point on the plate center in a six-point one-ring system, a seven-point two-ring support system is formed. After the radius optimization, the optimum arrangement has a support efficiency of $\xi = 0.045 \cdot 10^{-3}$ and the average support efficiency of each point of $\gamma_7 = 2.40 \cdot 10^{-3}$. An eight-point two-ring system does not produce a satisfactory result and, thus, has never been used in practice. A nine-point two-ring system is the best among two-ring support systems. However, nine support points still can not form an integrated triangular arrangement. Under this support condition, the reduction of the rms surface error is still limited even compared with a seven-point two-ring support system. The average support efficiency of each point decreases greatly (Figure 2.4). A 12-point two-ring support system forms a real integrated triangular grid. Under this condition, both the rms surface error and the average support efficiency of each point are improved. The support efficiency is $\xi = 0.013 \cdot 10^{-3}$ and the average support efficiency of each point is $\gamma_{12} = 1.88 \cdot 10^{-3}$. It is worth noting that γ_{12} is merely 1.6 times worse than $\gamma_{triangular}$.

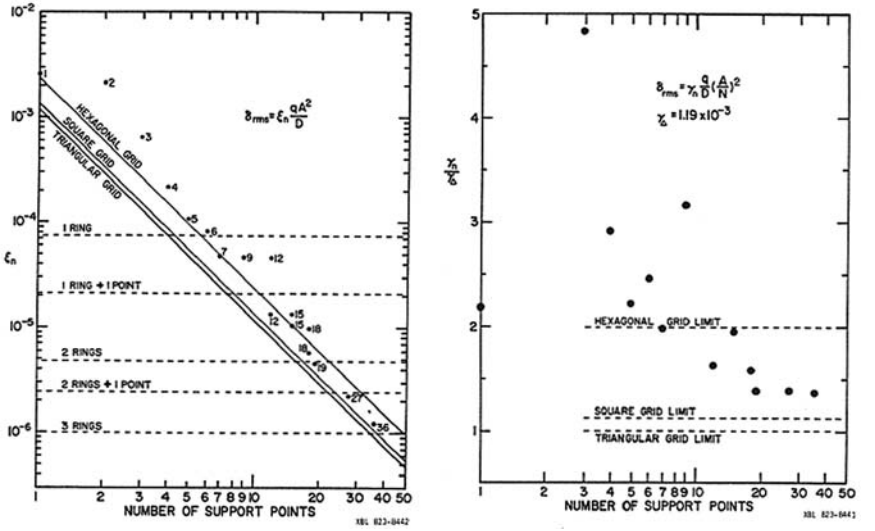


Fig. 2.4. The support efficiency as a function of the number of support points for different support systems (*left*) and the normalized efficiency of each point as a function of the number of support points (*right*) (Nelson et al., 1982).

By adding more points, a 15-point two-ring system has an even better average support efficiency, γ_{15}. When an 18-point two-ring system is used, the support efficiency increases again. If an additional support point, the 19th, is added at the center of this system, the support efficiency ξ and the average support efficiency of each point γ_{19} decreases rapidly. As the support point number further increases, the support efficiency ξ goes down gradually and the efficiency of each point γ_N approaches $\gamma_{triangular}$. For example, the support efficiency of each point for an optimized 36-point system is $\gamma_{36} = 1.4\gamma_{triangular}$. However, too many support points produce many more variables in the system optimization. Therefore, careful calculation is necessary as a minute change of parameters can result in large variation of the surface rms error.

Figure 2.4 gives the ratio between the support efficiency of each point γ_N and the triangular grid efficiency $\gamma_{triangular}$. The overall support efficiency ξ is also included. All the data points in the figure are for optimum support systems.

For a large thin mirror, the support point number N is usually inversely proportional to the average support area. If the mirror thickness is t, then the following relationship exists:

$$\delta_{rms} \sim \frac{1}{(tN)^2} \tag{2.11}$$

For improving the surface accuracy, adding more support points or using a thick mirror are necessary. For the same surface accuracy, a thinner mirror

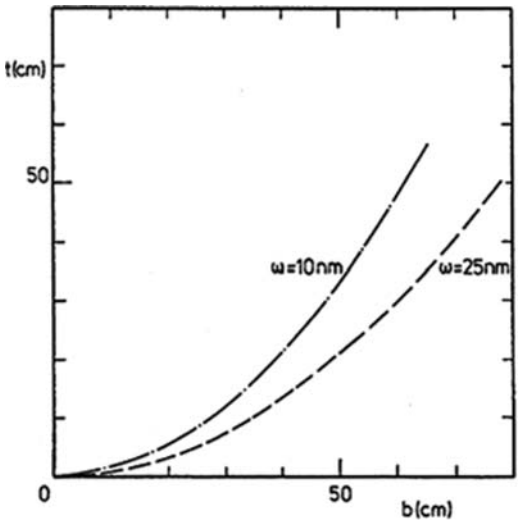


Fig. 2.5. The relationship between mirror thickness, support distance and the surface rms error.

requires more support points than a thicker one. More support points reduces the distance between each support point. If the most efficient triangle grid support is used and the mirror is of Cer-Vit material with $E = 9.2 \times 10^{10}\,\mathrm{Nm}^{-2}$, $v = 0.25$, and $\rho = 2{,}500\ \mathrm{kg\ m}^{-3}$, the relationship between the mirror thickness, the distance between support points, and the surface rms error is shown in Figure 2.5.

2.1.2.2 Radial Support for Optical Mirrors

The maximum mirror deformation caused by a radial (lateral) mirror support system occurs when the telescope points to horizon. At this position, the gravity and support forces are both perpendicular to the mirror surface (Figure 2.6). The major mirror deformation at this position is still along the axial direction z. The strain ε_z produced in the z direction is due to the Poisson effect of the support forces:

$$\varepsilon_z = -\frac{v}{E}(\sigma_x + \sigma_y) \tag{2.12}$$

where E is the Young modulus, v the Poisson ratio, and σ_x and σ_y the stresses in x and y directions, respectively. If the x direction is along the vertical line, the stress in this direction is caused by gravity and the radial supporting forces. The stress in the y direction is not related to gravity and is determined by the mirror support conditions.

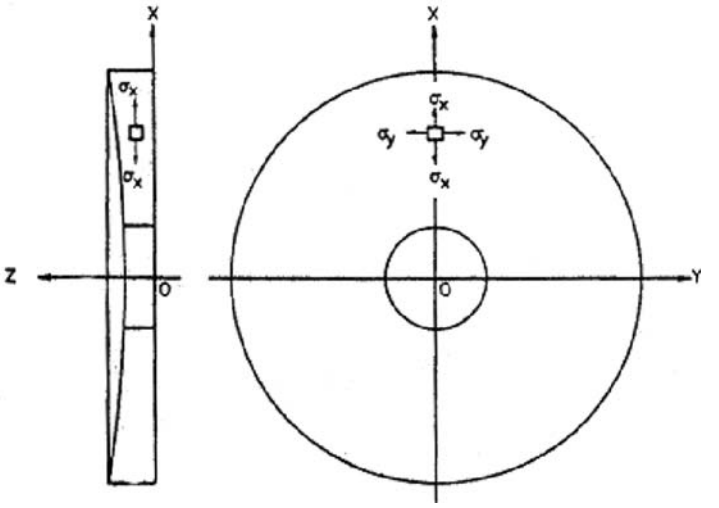


Fig. 2.6. Mirror stress distribution under the radial mirror support system (Cheng and Humphries, 1982).

Three radial mirror support systems exist and their force conditions are illustrated in Figure 2.7. Figure 2.7(a) shows the force condition when a mercury belt is used (a rubber torus filled with mercury and with a fixed inner contact area). Figure 2.7(b) shows the force condition for a cosine radial support system (counterweight and cantilever system in the radial direction). And Figure 2.7(c) shows the force condition of a vertical push-pull support system of which all supporting forces are parallel to the vertical axis (counterweight and cantilever

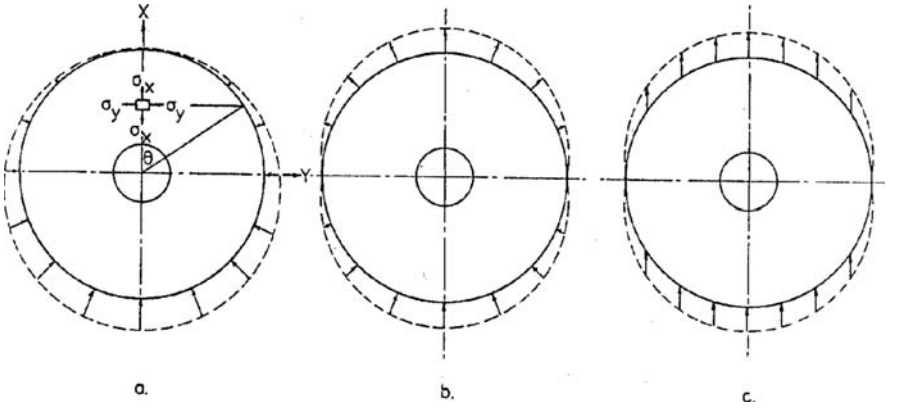


Fig. 2.7. Force conditions of (a) mercury belt radial support, (b) the cosine lateral force radial support, and (c) the vertical push-pull radial support (Cheng and Humphries, 1982).

system in the vertical direction). The stresses along the y direction of these three radial support systems are:

$$\begin{aligned} \sigma_{ya} &= -k_a(1-\cos\theta)\sin\theta \\ \sigma_{yb} &= k_b\cos\theta\sin\theta \\ \sigma_{yc} &= 0 \end{aligned} \tag{2.13}$$

where θ is the polar angle in the mirror plane and k_a and k_b are positive constants. From the formulas, it is found that the stresses along the y direction in System (a) and in the bottom part of System (b) are of the same sign as stresses in the vertical direction. The contributions from these stresses to the surface error in the z direction are added to that from stresses in the vertical direction. Therefore, the minimum mirror surface deformation along the z axis happens only in System (c) where stresses in the y direction vanish.

If a paraboloidal mirror has a flat-back surface and is supported as in configuration (c), a small section of the mirror on the vertical symmetrical plane is shown as in Figure 2.8. Since the mirror thickness in the z direction is expressed as $z = (x^2/(4F)) + t_0$, where t_0 is the thickness at the vertex and F the focal length, then the deformation caused by the Poisson effect is $w = z \cdot \varepsilon_z$, where ε_z is the strain in the z direction. The deformation of the mirror surface is:

$$w = \frac{v\rho g}{E}\left(\frac{R_0}{12fd} + R_0 t_0\right) - \frac{v\rho g t_0}{E}\cdot x - \frac{v\rho g}{12Efd}\cdot x^3 \tag{2.14}$$

where ρ is the material density, g the gravitational acceleration, R_0 the radius of the central hole, t_0 the thickness at the central hole, f the focal ratio, and d the mirror diameter. The first and second terms in this expression are constant and

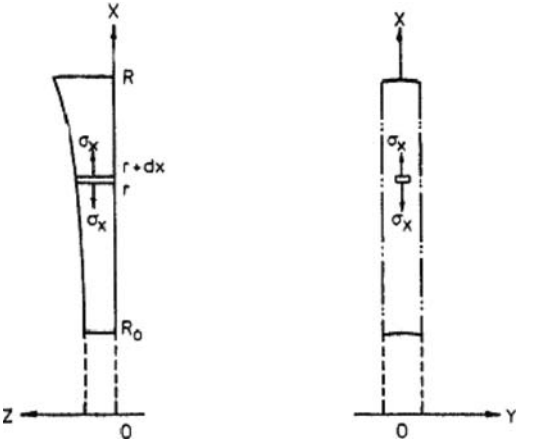


Fig. 2.8. Mirror stress under a vertical push-pull lateral support system.

linear terms, respectively. Neither of these have an influence on the surface rms error. Only the third term will cause astigmatism of the wavefront. This undesirable deformation has its maximum at a position of $x = d/2$ and its value is:

$$w_{\max} = \frac{v\rho g}{96Ef} \cdot d^2 \tag{2.15}$$

If Cer-Vit material is used with the Young modulus of $E = 9.2 \times 10^{10} N \cdot m^{-2}$, the Poisson ratio of $v = 0.25$, and the density of $\rho = 2{,}500$ kg m^{-3}, then the relationship between the maximum surface error, the mirror diameter, and the f-ratio for a flat-back mirror under radial support system is shown in Figure 2.9. The maximum of the undesirable deformation caused by the Poisson effect is proportional to the diameter squared and is inversely proportional to the focal ratio. In general, this deformation is small in comparison with errors of the axial support case and will not produce serious effects on the telescope image. This is why the radial mirror support is less important than the axial one.

The depth of a curve is called sagitta. For a large parabolic mirror, the sagitta is $S = d/(16f)$. A large sagitta value of a flat-back mirror produces different thermal inertia along the radius, resulting in thermal-induced surface error. A meniscus mirror with uniform thickness avoids this thermal problem. The lateral support system for a meniscus mirror is slightly more complicated than that of a flat-back one. The main concern is that the lateral support forces have to pass through the mirror section center of gravity they support. The distance between combined support force and center of gravity produces a

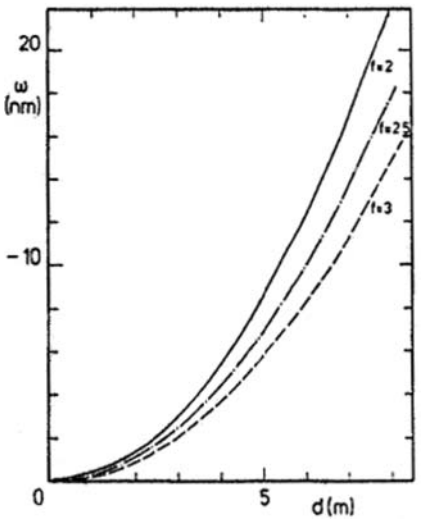


Fig. 2.9. Maximum flat-back mirror deformation under a push-pull lateral support system (Cheng and Humphries, 1982).

harmful bending moment which may produce large mirror surface deformations. The deformation caused by the Poisson effect is generally small and most of it varies linearly with the distance to the support point. However, deformations caused by bending moments are large and vary in a high power nonlinear fashion with the distance to the support point.

To reduce the deformation caused by the bending moment, the lateral support forces can be distributed inside small holes over the mirror back surface. In this way, the distance between the lateral support point and the local center of gravity of the mirror section is reduced. The combined lateral support force is on the same plane as the center of gravity of the mirror.

The bending moment caused by the lateral support of a meniscus mirror is proportional to both the diameter and thickness but inversely to the focal ratio. The deformation caused by this bending moment is proportional to the mirror area but inversely to the focal ratio and the square of the mirror thickness. Therefore, when the mirror aspect ratio increases, the mirror lateral support design becomes more important to the mirror surface deformation.

2.1.3 Surface Error Fitting and Slope Error Expression

An ideal Gaussian wavefront on an aperture plane is flat in shape. For any deformed wavefront, there exist many Gaussian reference wavefronts. However, only one among these has the minimum deviation from the deformed wavefront. This particular reference wavefront is the best fit wavefront. Relative to the original coordinate system, the best fit wavefront may have coordinate rotation, coordinate shift, and focal length change. The difference between the deformed and the best fit wavefronts is the wavefront error or the path length error. The wavefront error produced by a mirror is twice the mirror surface error. Wavefront error relative to the wavelength is called wavefront phase error. A wavefront error of half wavelength is equal to a wavefront phase error of 180°.

When an ideal telescope mirror, either paraboloid or hyperboloid in shape, is under gravity loading, the surface shape will change. For a deformed mirror surface, there is a best-fit reference surface. Detailed formulation of the best fit surface is provided in Section 7.1.4.

The best fit process for optical mirrors, where the f-ratio is large (comparing with a radio dish), is much easier. This is especially true when a multi-ring axial mirror support system is used. One convenient solution in the mirror support optimization is to consider only the axial coordinate shift. The principle is named equal softness. The ideal best fit surface of a mirror is a surface with an axial displacement from the original one. This simplification will reduce the workload in the mirror support optimization. The wavefront rms errors mentioned in the previous section were also obtained by using this simplified best fit method.

Apart from wavefront error, slope error is also used to describe the mirror deformation. The slope error is a measure of surface long-range modulations or

zones with ripple wavelengths (not optical wavelength) typically in the centimeter to tens of centimeter range. Because of the mirror slope error, the wavefront distortion is produced. The wavefront slope error is twice the mirror slope error. The slope error produces image blur size increase and resolution decrease. The resulting image blur size and resolution can be computed using geometrical optics. The slope angle is proportional to the image blur angle. The wavefront slope can be directly detected by a number of wavefront sensors, such as the Hartmann one. Maximum image diameter is about four-times the maximum mirror slope error, or twice the maximum wavefront slope error. However, limitation exists when the slope error instead of wavefront error is used for optical systems. If the wavefront or mirror surface ripple amplitude becomes so small relative to the wavelength (as may occur, for example, in going from visible to infrared) that the geometrical optics will be no longer valid then the effects of slope error may be greatly reduced.

Usually, the mirror slope error (S) is proportional to the rms surface error and is inversely proportional to the effective mirror support distance (u). The effective support distance is defined by the formula $N\pi u^2 = A$, where A is the mirror area and N the number of support points. Therefore, the mirror slope error can be expressed as (Nelson et al., 1982):

$$S = g_N \frac{q}{D} \left(\frac{A}{N}\right)^{3/2} \sim \frac{1}{t^2} \left(\frac{A}{N}\right)^{3/2} \tag{2.16}$$

where q is the areal density, D the diameter, and t the thickness. In this equation, the constant g_N can be obtained from the calculation and, in most cases, it can be expressed as a function of mirror support efficiency γ_N:

$$g_N = 9\gamma_N \tag{2.17}$$

2.2 Lightweight Primary Mirror Design

2.2.1 Significance of Lightweight Mirrors for Telescopes

The primary mirror is the most important component of an optical telescope. Its surface should maintain high accuracy under the telescope operating conditions. The weight and cost of the mirror are determining factors for the telescope total weight and total cost.

The mirror cell supports the primary mirror through a support system. Most support systems involve floating counterweight cantilever devices or air pads so that small changes in support position produce little effect on the mirror surface shape. However, any support system has a limited dynamic range. Therefore, the mirror cell has to be stiff enough so that its deformation does not exceed this dynamic range. The dimension and material density of traditional mirror cells

are usually larger than that of the mirror. Therefore, the weight of the cell is in the same magnitude of the mirror.

The telescope tube supports the primary mirror at one side and the secondary assembly at the other side. Therefore, the tube weight, including the center section, is related to the weight of both the mirror and cell. The weight of the mount structure is also related to the mirror weight directly or indirectly. Table 2.2 lists relative weights of all telescope components in a classical telescope relative to the primary mirror weight. From the table, one would find how important the mirror weight reduction is to the telescope weight.

The cost of any engineering project is always proportional to the structural weight. The cost–weight ratio is an indicator of the structure precision and complexity. As the mirror weight is a deciding factor on the overall telescope weight, therefore, the reduction of the mirror weight is very important in telescope design. To build an extremely large telescope with a nonstop increase of the aperture, the mirror weight reduction is a necessary first step.

In the past few decades, mirror weight reduction had been a major research topic for telescope scientists and engineers. A number of techniques developed in this aspect include: (a) using a thin mirror; (b) using a honeycomb mirror; (c) building a multiple-mirror telescope; (d) building a segmented mirror telescope (SMT); and (e) using mirrors made of metal, or carbon fiber reinforced plastic (CFRP) composite, or other special materials. These techniques are discussed in the following sections.

2.2.2 Thin Mirror Design

Cheng and Humphries (1982) and Nelson et al. (1982) pointed out that the surface error of any thin mirror may be reduced by an increase of the mirror support points and, in theory, a mirror can have a very large aspect ratio. Traditional telescope mirrors had their aspect ratios smaller than 10. Newly designed monolithic mirror telescopes have their mirror aspect ratios much larger than 20 and newly built segmented mirror telescopes have aspect ratios as large as 110. In an extreme case, the thin adaptive secondary mirror has an aspect ratio of 320.

Table 2.2. Weight ratios of major components for an optical telescope.

Name of component	Relative weight
Primary mirror	1.0
Primary mirror cell	1.5–3.3
Tube	3.5–10.0
Yoke	6.0–16.5
Mounting structure	6.0–20.0
Total	18.0–50.0

A key issue in the use of thin mirrors is the mirror support system. The factors which limit the increase of mirror aspect ratio are assembly and disassembly methods of the mirror, the maximum stresses during installation, the wind disturbances, and the resonant vibration.

The stresses of a primary mirror, except an extremely thin one, in normal working conditions are negligible. However, high stresses are induced during assembly and disassembly. The maximum stress $\sigma_{\max}$ of a circular mirror during assembly is:

$$\sigma_{\max} = Kq\frac{d^2}{t} \tag{2.18}$$

where K is a constant determined by the mirror lifting condition, t the thickness, d the diameter, and q the density of the material. The condition of using this simplified formula is that the lifting force applied is on the middle plane of the mirror. When the lifting force is on the bottom of the mirror, this formula is still correct for the maximum stress estimation.

Traditional thick mirrors were lifted on the central holes. The lifting forces are applied on the bottom surface around the hole. If the mirror material has a density of $q = 2{,}500$ kg/m^3 and a Poisson ratio of $\nu = 0.3$, the relationship between the maximum stress, the diameter, and the aspect ratio during a central hole lifting is shown in Figure 2.10. In the figure, the maximum permissible stress for Cer-Vit material of about 3×10^6 N m^{-2} is also plotted. From this figure,

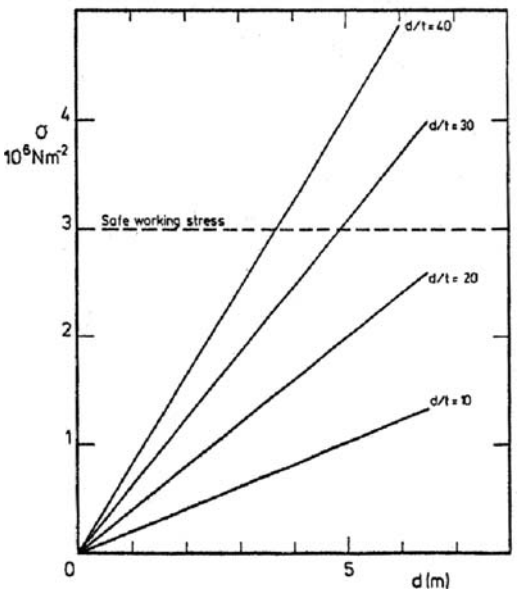


Fig. 2.10. The relationship between maximum stress, diameter, and aspect ratios.

mirrors of large aspect ratios using a central hole lifting may have unacceptable maximum stresses. The maximum stress in this case is in the tangential direction.

To reduce the stresses, the lifting position may move to the outer edge of a mirror. When the mirror is lifted on its outer edge, the maximum stress is still in the tangential direction. However, the maximum stress reduces to half of the numbers shown in Figure 2.11.

To further reduce the maximum stress during assembly, the lifting position should move to the middle radius of a mirror. If the lifting force is applied on a continuous circle with a radius of 0.67 R (R is the radius of the mirror), the maximum stress is only one tenth of that when the lifting force is applied on the central hole. The maximum stress in this case changes from the tangential to the radial direction at the lifting radius. In Figure 2.11, stress distributions for different radius lifting are listed, where g represents the gravitational acceleration, ρ the density, σ_r the radial stress, and σ_t the tangential stress.

Normally, telescope mirrors use a single ring lifting. However, difficulties arise when the lifting radius is in the middle of sophisticated mirror support mechanisms. To solve the problem, a combination of the mirror cell and lifting

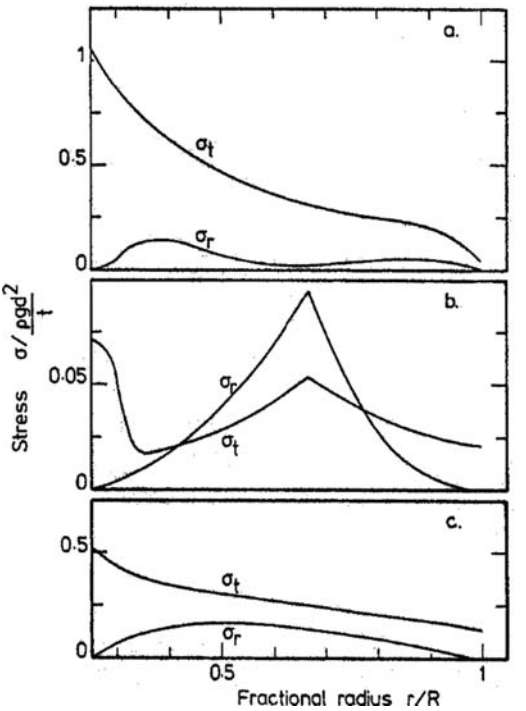


Fig. 2.11. Stress distribution under (a) a central ring support, (b) a 0.67-radius ring support, and (c) outer ring support (Cheng and Humphries, 1982).

mechanism is made during the assembly and disassembly processes. The mirror is lifted with its mirror cell, so the safety of the mirror is insured.

Another lifting method involves multiple vacuum lifting points on top of a mirror, which is often used in the mirror polishing process. Extremely thin mirrors can only be lifted with distributed multiple support point floating devices.

The wind disturbance on a mirror becomes serious as the mirror becomes very thin and is exposed to outside air flow. Mirror supports include floating ones, which do not take any additional loads, and positioning ones (hard points), which do take additional loads. Usually at least three positioning supports are used. If these positioning supports are evenly located on the outer radius of the mirror, the maximum deformation caused by a pressure load P is:

$$w_{\max} = 1.9 \times 10^{-3} \frac{\pi P d^4}{t^3} \tag{2.19}$$

This formula shows that the maximum deformation caused by wind is proportional to the cubic power of the aspect ratio. To increase the wind resistance of a mirror, the positioning support radius should be optimized. An optimized radius of the positioning supports has a normalized radius of 0.67. The maximum deformation in this case will reduce to a quarter of the above value. Further improvement in the wind resistance can be achieved by increasing the number of load bearing support (positioning) points or by using an adaptive optics mirror support system.

The natural frequency of a thin mirror can be expressed as:

$$v_R = \frac{2\phi \cdot t}{\pi \cdot d^2} \sqrt{\frac{E}{12(1-v)\rho}} \tag{2.20}$$

where ϕ is a constant determined by the mode shape. If the hard points are arranged on the outer radius with a free edge, the value of ϕ is 9.1. When Cer-Vit material is used, the natural frequency of a mirror is:

$$v_R = 1.14 \times 10^4 \frac{t}{d^2} (Hz) \tag{2.21}$$

where the unit of d and t is in meters. This formula shows that a 5 m mirror with an aspect ratio of $d/t = 20$ has a natural frequency of about v_R =100 Hz. If the aspect ratio increases to 50, then the natural frequency would be reduced to 27 Hz. If three hard points are moved to 0.7 radius, the natural frequency will reduce by another factor of 4.

If the stiffness of the hard points is considered, the relationship with the frequency of the piston mode of a rigid mirror is approximately (Hill, 1995):

$$K_i = (2\pi v)^2 m/3000 \tag{2.22}$$

where m is the mirror mass in kg, v the frequency, and K_i the stiffness of one hard point in N/mm. A small frequency number of the piston mode is undesirable. High stiffness of the hard points will minimize the wind and actuator error induced displacements and vibration amplitudes. To increase the natural frequency of the mirror system, it may be necessary to add more hard support points.

The friction of a mirror support system is another consideration when an aspect ratio is selected. The friction produces support force errors. A classical counterweight cantilever system usually has a friction coefficient of 0.1∼0.3%. To ensure its optical performance, the aspect ratio of a mirror with this system should satisfy the following relationship:

$$d^2/t \leq 2500(cm) \tag{2.23}$$

For an air bag support system, the friction coefficient is about 0.01%. The corresponding number in the right hand side is 25,000 cm. Other factors, which restrict the use of very thin mirrors, include the mirror casting, mirror polishing, and mirror transportation. If we want to use an even larger aspect ratio, technology improvements in these fields are required.

2.2.3 Honeycomb Mirror Design

A honeycomb mirror is a sandwiched structure including face plate, honeycomb core, and base plate. The base plate may have holes for mirror supporting and ventilation. Honeycomb mirrors are light in weight, high in stiffness, and rigid in bending. The earlier applications of this type of mirror are the 4.5 m old Multi-Mirror Telescope (MMT) which was made of six 1.8 m honeycomb mirrors (note: The old MMT telescope was converted to a single mirror 6.5 m telescope in 1998) and the 2.4 m Hubble Space Telescope (HST). The largest honeycomb mirror has a diameter of 8.4 m.

Honeycomb mirrors are made by removing materials in the honeycomb holes or by fusing glass plates and core together at high temperatures. A rotational honeycomb mirror casting method was developed by the mirror laboratory of the University of Arizona for large honeycomb mirrors with a paraboloidal surface shape. The principle to form a paraboloidal shape when the glass is in a liquid form is the same as that of a rotational mercury mirror as discussed in Section 2.2.6. During the mirror casting process, the furnace is heated to 1,178 C and is rotating at a constant speed, few revolutions per minute, to shape the front mirror surface.

The rigidity of a honeycomb mirror is nearly equivalent to a solid one of a similar thickness, but with only a small fraction of the weight. The bending stiffness of a honeycomb mirror is approximately:

$$D = \frac{E(h+t)t^2}{2(1-v^2)} \tag{2.24}$$

where t is the thickness of the upper or bottom plates and h the thickness of the honeycomb core in the middle. The weight of a honeycomb mirror is $(2t + \alpha h)/(2t + h)$-times that of a solid one, where $\alpha(\alpha \ll 1)$ is the relative density of the core compared with the face plate.

A honeycomb mirror also has lower thermal inertia than that of a solid one. The low thermal inertia reduces the temperature gradient within the mirror. If air ventilation is applied to the honeycomb cells, then normal borosilicate glass with a relatively larger coefficient of thermal expansion (CTE) can be used for large optical telescopes. The thermal time constant τ of a plate is:

$$\tau = \frac{\rho \cdot c \cdot t^2}{\lambda} \tag{2.25}$$

For borosilicate glass, the density is $\rho = 2{,}230$ kg/m^3, the specific heat $c = 1{,}047$ J/Kg °C, the thermal conductivity $\lambda = 1.13$ W/m °C, and t the thickness of the wall. The thermal time constant of a honeycomb mirror can be derived from its wall thickness. To further reduce its thermal time constant, ventilation may be added. Under air ventilation, the energy is conserved and the following formula exists (Hill, 1995):

$$m_g \dot{T}_g c_g = \dot{m}_a c_a (T_{exit} - T_{input}) \tag{2.26}$$

where subscripts g and a are for glass and air, T is the temperature, T_{exit} and T_{input} the exit and input air temperature. The specific heat of the air c_a is 711 J/Kg °C and the thermal time constant is related to air flow rate as:

$$\dot{m}_a = \frac{m_g c_g}{\tau \cdot c_a \eta} \tag{2.27}$$

where $\eta = 0.7$ is the heat transfer coefficient (heat coupling coefficient) of the forced convection and τ the thermal time constant.

The size of the honeycomb hexagonal cells is determined from the maximum deformation of the mirror face plate during polishing. For a given pressure loading, the deformation of the center point of a honeycomb structure is:

$$w = 0.00111 \frac{qb^4}{D} \tag{2.28}$$

where q is pressure loading, b the distance between two opposite sides of the cell, and D the bending stiffness of the face plate. If the average pressure during the polishing is $q = 0.084\, N/cm^2$, the top plate thickness is 2 cm, and the maximum

deformation allowed is 1/20th of the visible wavelength, then the size b of the cell can be calculated. The thickness of honeycomb side wall is determined by the mirror's global surface deformation and is usually a quarter of the top plate thickness. With ventilation holes on the mirror back, the installation of the mirror support system is easy. A honeycomb mirror requires no special designed support devices. Figure 2.12 shows the axial and lateral support systems of a honeycomb mirror.

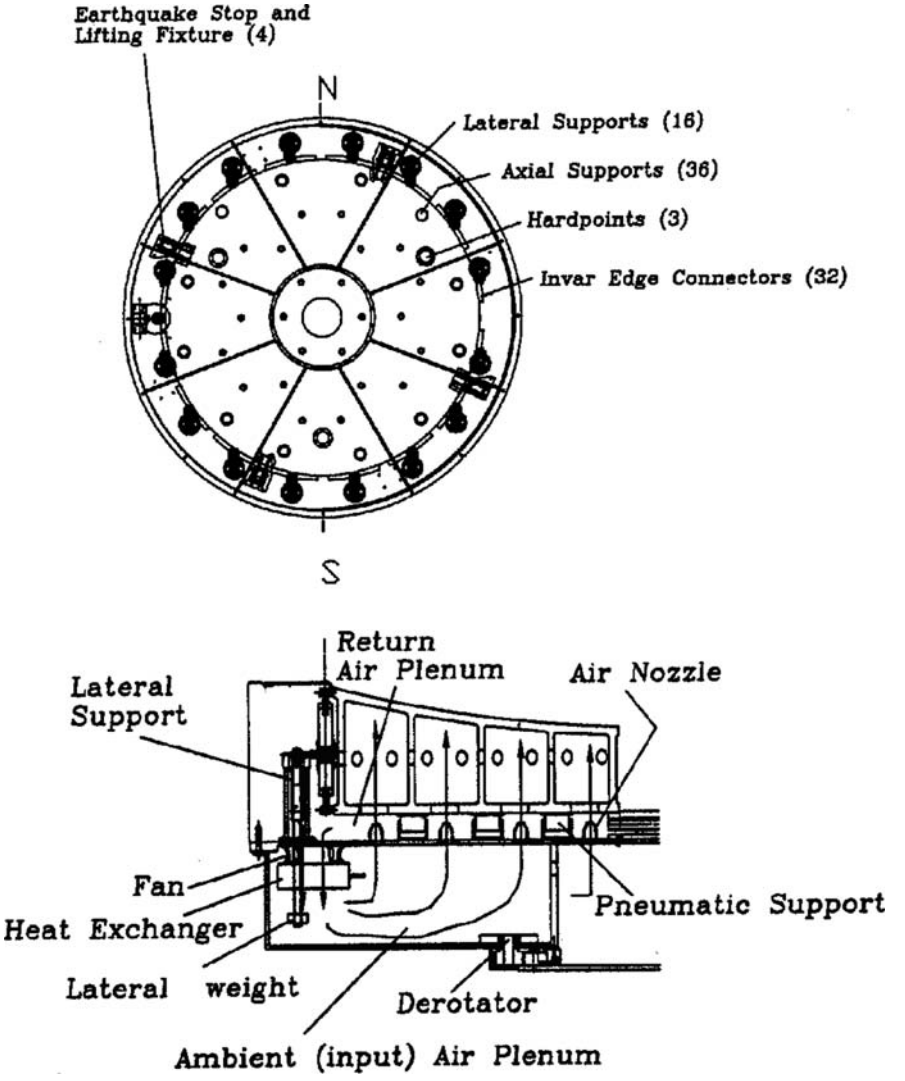


Fig. 2.12. An 8 m honeycomb mirror and its support system (West et al., 1997).

2.2.4 Multi-Mirror Telescopes

The name multi-mirror telescope comes from the old Multi-Mirror Telescope (MMT) which was built in 1979 with an equivalent aperture size of 4.5 m. This telescope had six independent tubes (or sub-telescopes) each with an aperture size of 1.8 m but firmly connected together by a large elevation (or tube) structure as shown in Figure 2.13. This design provided a new way to reduce the weight of a primary mirror.

Since the aperture area of an individual telescope tube accounts for only $1/n$ of that of the whole telescope, where n is the number of sub-tubes used, so the aspect ratio of the mirror is equivalent to $n^{1/2}$ of that of a monolithic primary mirror. Other advantages for a MMT telescope design are short tube length and large lateral tube dimension, so that the tube is stiff and the required dome is small.

For spectroscopic observation, when the co-focusing condition is met, the six sub-telescope foci are so arranged that they form a straight line on the entrance slit of the spectroscope. This avoids energy loss which happens in single aperture telescopes used in spectroscopic mode without an image slicer. Therefore, in this mode, the MMT works as one telescope.

Another intentional operation of this old MMT telescope was that the radiation collected by each sub-telescope was directed to a common focus in co-phasing condition, thus forming a Fizeau interferometer with a much larger baseline. However, the tube structure made of steel to support both the primary and secondary mirrors had serious uncorrectable thermal distortions and the telescope was lacking in optical path length equalization devices or optical delay lines to compensate these random phase differences. The wavefront co-phasing was nearly impossible. The field of view on its common focus was also limited due to a small angle between beams from all sub-telescopes in the common image plane. All these are reasons leading to the failure in this interferometer mode although some fringes were obtained occasionally. After 19 years of continuous struggle with the spectroscopic and independent small telescope observations, a conversion of the old MMT into a new 6.5 m single mirror telescope was finally made in 1998.

The old MMT was gone, but the idea to build a MMT-type Fizeau interferometer (Section 4.2.3) remains. For a coherent diffraction limited common focus image, all the sub-telescopes are required both co-phasing each other through sophisticated optical delay lines and free from atmospheric turbulence through adaptive optics. In the past, the technologies required to fulfill these two tasks were not ready, but now they are within reach.

A newly built MMT-type telescope is the Large Binocular Telescope (LBT) with two 8.4 m mirrors completed in 2008. With a separation between two mirror centers of 14.4 m, its common massive elevation structure is on hydrostatic pads to reduce the structural deformation. Now both sub-telescopes are in perfect working condition and used as an independent telescope, but the final target of this telescope as a Fizeau interferometer instrument has still not been realized.

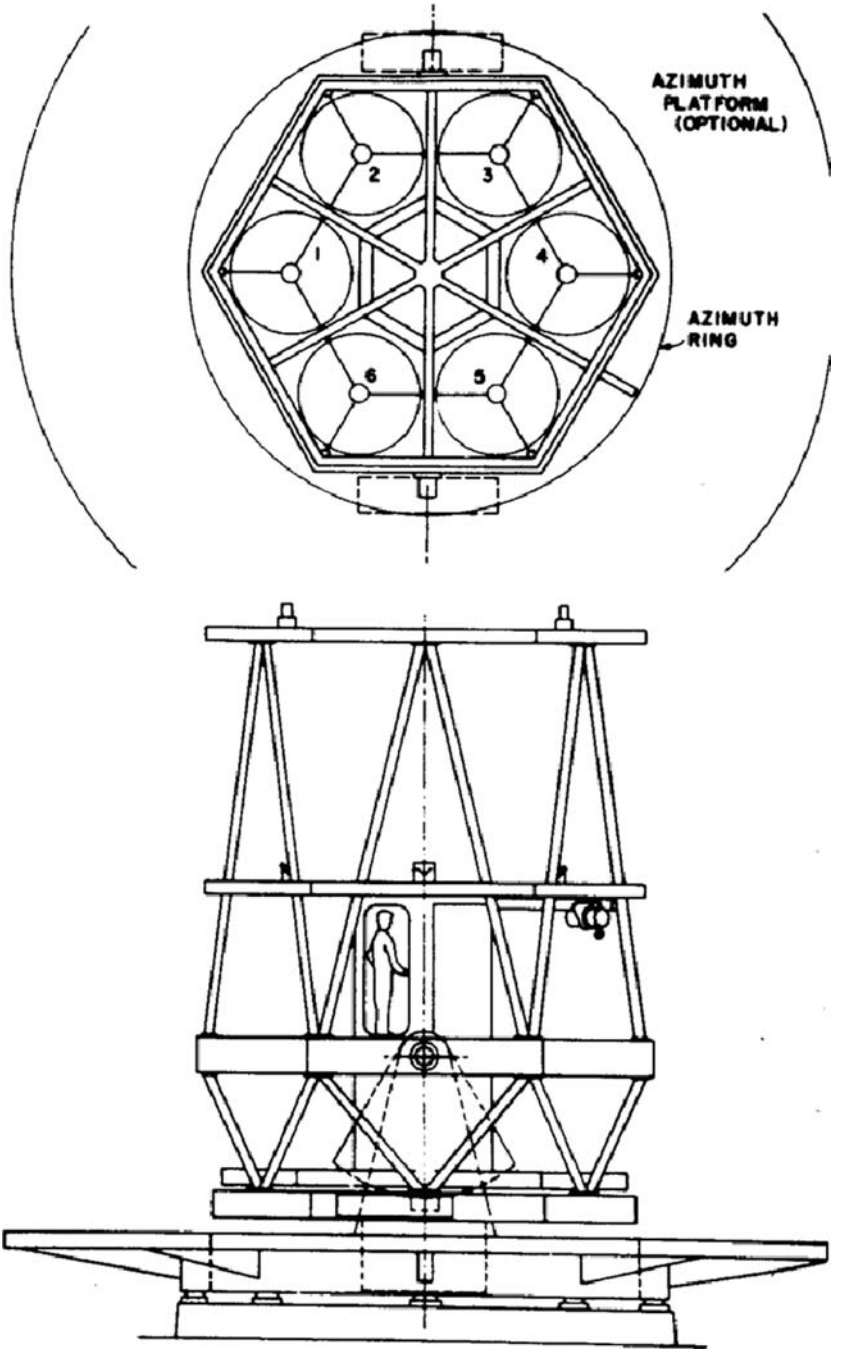


Fig. 2.13. The arrangement of the old MMT telescope.

2.2.5 Segmented Mirror Telescopes

The Segmented Mirror Telescope (SMT) represents a new approach to obtain a light weight primary mirror for extremely large optical telescopes. The extremely large optical telescopes are telescopes with their diameter far beyond 10 meters. Compared with the MMT design, the SMT, with all segments of the primary mirror reflecting light to a common secondary mirror can have both a large field of view and co-phase interferences between individual segments. The advantages of the SMT design include great mirror weight reduction, large cost savings, easy mirror handing and transportation, and small dome size.

A SMT telescope consists of many mirror segments, making up a larger light collecting area. Each of the 10 m Keck I and II telescopes has 36 1.8 m hexagonal mirror segments with a thickness of only 8.7 cm (Figure 2.14). Since the deformation of a thin mirror is proportional to the fourth power of diameter, the mirror support systems for smaller segments are much simpler in comparison with that for a monolithic larger primary mirror. The smaller segment diameter and repeatable segment patterns also lower the mirror manufacture, mirror polishing, and mirror transportation cost.

Two surface shapes are used for the SMT telescopes: a spherical one as used in the HHT and SALT telescopes and paraboloidal one as used in the Keck and Gran Telescopio Canarias (GTC). The GTC was built by Spain, Mexico, and University of Florida. The proposed TMT, GMT, and E-ELT will have a paraboloidal surface shape and the OWL a spherical surface shape.

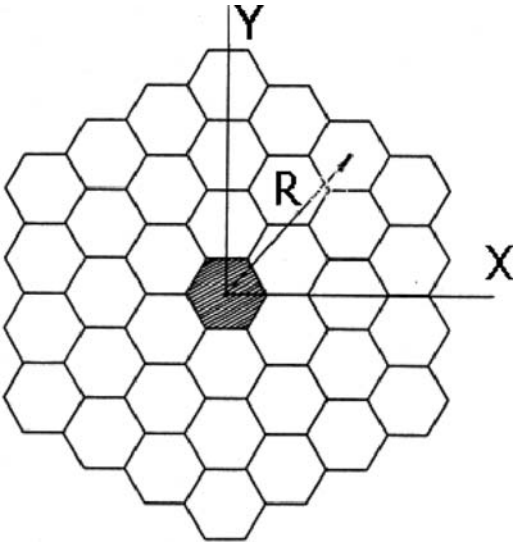


Fig. 2.14. The primary mirror of the 10 m Keck telescope.

With a spherical surface shape, the mirror segments are all identical, so that the mirror manufacture is easy and the cost is low. However, a complicated field corrector is needed to correct the spherical aberration although the field of view is wider. In general, the spherical primary mirror shape limits the usage of the telescope, so most telescopes use a paraboloidal surface shape.

With a paraboloidal surface shape, the system optical design is easier, but with a relatively smaller field of view. However, the mirror segments are different between rings. All these segments have off-axis paraboloidal surface shape, bringing difficulties in segment manufacture and polishing. The SMT telescopes require an accurate position control of each mirror segment to achieve a smooth coherent mirror surface. The strategy of the mirror segment position control is discussed in Section 4.1.4. In this section, the off-axis paraboloidal mirror segment manufacture is discussed.

The formulae for an axial symmetrical, conic surface are (Nelson et al., 1985):

$$
\begin{aligned}
Z(X,Y) &= \frac{1}{K+1}\left[k - [k^2 - (K+1)(X^2+Y^2)]^{1/2}\right] \\
Z(X,Y) &= \frac{1}{2k}(X^2+Y^2) + \frac{1+K}{8k^3}(X^2+Y^2)^2 \\
&\quad + \frac{(1+K)^2}{16k^5}(X^2+Y^2)^3 + \frac{5(1+K)^3}{128k^7}(X^2+Y^2)^4 \cdots\cdots
\end{aligned}
\tag{2.29}
$$

where k is the radius of curvature at the vertex, K the conic constant, Z the coordinate along the axis, and O the vertex of the surface. When the global coordinate system is replaced by a local one of $p(x,y,z)$ (Figure 2.15), the conic surface can be expressed as a trigonometric series:

$$
z = \sum_{ij} \alpha_{ij} \rho^i \cos j\theta \quad (i \geq j \geq 0, i-j = even) \tag{2.30}
$$

In this expression, the first few coefficients are listed as:

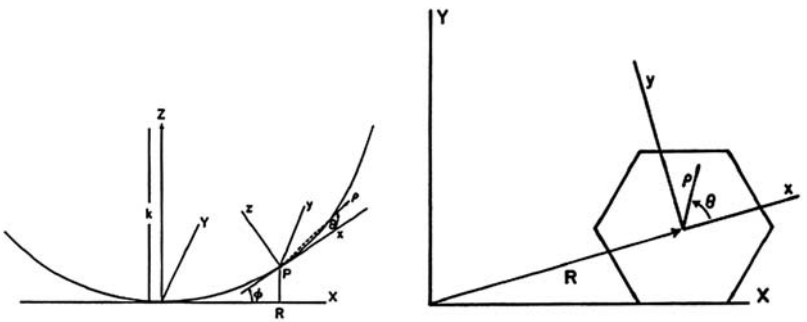


Fig. 2.15. The global and local coordinate systems for an off-axis, conic segment surface (Nelson et al., 1985).

$$\alpha_{20} = \frac{a^2}{k}\left[\frac{2 - K\varepsilon^2}{4(1 - K\varepsilon^2)^{3/2}}\right] \text{(focus)}$$

$$\alpha_{22} = \frac{a^2}{k}\left[\frac{K\varepsilon^2}{4(1 - K\varepsilon^2)^{3/2}}\right] \text{(astigmatism)}$$

$$\alpha_{31} = \frac{a^3}{k^2}\left[\frac{K\varepsilon[1 - (K+1)\varepsilon^2]^{1/2}(4 - K\varepsilon^2)}{8(1 - K\varepsilon^2)^3}\right] \text{(coma)}$$

$$\alpha_{33} = \frac{a^3}{k^2}\left[\frac{K^2\varepsilon^3[1 - (K+1)\varepsilon^2]^{1/2}}{8(1 - K\varepsilon^2)^3}\right] \tag{2.31}$$

$$\alpha_{40} = \frac{a^4}{k^3}\left[\frac{8(1+K) - 24K\varepsilon^2 + 3K^2\varepsilon^4(1 - 3K) - K^3\varepsilon^6(2 - K)}{64(1 - K\varepsilon^2)^{9/2}}\right]$$

(spherical aberration)

$$\alpha_{42} = \frac{a^4}{k^3}\left[\frac{K\varepsilon^2[2(1 + 3K) - (9 + 7K)K\varepsilon^2 + (2 + K)K^2\varepsilon^4]}{64(1 - K\varepsilon^2)^{9/2}}\right]$$

$$\alpha_{44} = \frac{a^4}{k^3}\left[\frac{K\varepsilon^2[1 + 5K - K\varepsilon^2(6 + 5K)]}{64(1 - K\varepsilon^2)^{9/2}}\right]$$

where a is the projected radius of a hexagonal mirror segment, $\rho = (x^2 + y^2)^{1/2}/a$, $\theta = \tan^{-1}(y/x)$, and $\varepsilon = R/k$. The other coefficients are expressed in a generalized form as:

$$\alpha_{ij} \approx a^i\varepsilon^j/k^{i-1} \tag{2.32}$$

When $i > 4$, these coefficients are very small for a hexagonal mirror segment and can be neglected. The off-axis conic surface described in the above expression can be transformed into a symmetrical spherical surface by applying elastic deformations.

For a spherical surface, K equals zero and it leaves only two coefficients of α_{20} and α_{40} in the segment surface expression. The values of these two coefficients are $a^2/(2k)$ and $a^4/(8k^3)$, respectively. By comparing the above off-axis conic shape with a best-fit spherical shape, the required elastic deformations for the surface shape transformation can be derived. The required deformations, which equal the differences between these two surface expressions, are (Lubliner and Nelson 1980):

$$\begin{gathered} w = \sum_{ij} a_{ij}\rho^i \cos j\theta \\ w \cong \alpha_{20}\rho^2 + \alpha_{22}\rho^2 \cos 2\theta + \alpha_{31}\rho^3 \cos\theta + \\ \alpha_{33}\rho^3 \cos 3\theta + \alpha_{40}\rho^4 + \alpha_{42}\rho^4 \cos 2\theta \end{gathered} \tag{2.33}$$

With classical thin plate theory, the required deformations are obtained by applying forces, moments, and distributed surface loads on the segment. These forces, moments, and distributed loads are also expressed in sine and cosine series as:

$$\begin{aligned} M(\theta) &= M_0 + \sum_n (M_n \cos n\theta + \bar{M}_n \cos n\theta) \\ V(\theta) &= V_0 + \sum_n (V_n \cos n\theta + \bar{V}_n \cos n\theta) \\ q(r,\theta) &= q_0 + q_1 r \cos n\theta + q_2 r \sin \theta \\ V_0 &= -q_0/2 \\ M_1 + aV_1 &= -q_1 a^3/4 \\ \overline{M}_1 + a\overline{V}_1 &= -q_2 a^3/4 \end{aligned} \tag{2.34}$$

The coefficients used in the above expressions derived from the plate deformation formulae are:

$$\begin{aligned} M_0 &= \frac{D}{a^2}[(2+v)\alpha_{20} + 4(3+v)\alpha_{40}] \\ V_0 &= -\frac{D}{a^3}(32\alpha_{40}) \\ M_1 &= \frac{D}{a^2}[2(3+v)\alpha_{31} + 4(5+v)\alpha_{51}] \\ V_1 &= -\frac{D}{a^3}[2(3+v)\alpha_{31} + 4(17+v)\alpha_{51}] \\ M_n &= \frac{D}{a^2}[(1-v)n(n-1)\alpha_{nn} + (n+1)[n+2-v(n-2)]\alpha_{n+2,n}] \\ V_n &= \frac{D}{a^3}[(1-v)n^2(n-1)\alpha_{nn} + (n+1)(n-4-vn)\alpha_{n+2,n}] \\ q_0 &= 64D\alpha_{40}/a^4 \\ q_1 &= 192D\alpha_{51}/a^5 \\ q_2 &= 192D\beta_{51}/a^5 \end{aligned} \tag{2.35}$$

where D is the plate bending stiffness and v the Poisson ratio. If a sine component exists in the deformation formula, then a sine term in shearing force, or moment, or distributed load expression is necessary. After applying these required loads, the problem of off-axis paraboloidal surface manufacture turns into simple spherical surface manufacture. The only difference is that shearing

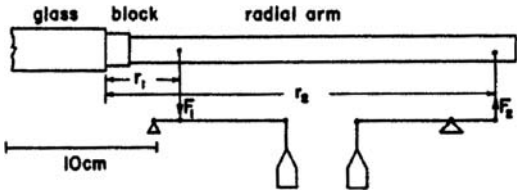


Fig. 2.16. The methods for applying shear force and bending moment during the off-axis mirror segment manufacture.

forces, or moments, or distributed loads have to be applied on the mirror segment during the segment polishing process.

Figure 2.16 shows a method of applying shearing forces and moments around a segment edge. The distributed loads can be provided in the axial mirror support system. Since the mirror segment is polished under internal stresses, this mirror manufacturing method is called "stressed polishing." If a mirror segment under stress has been fabricated into a spherical shape, an off-axis paraboloidal shape can be obtained by releasing all the applied loads. Using this stressed polishing method, an ideal off-axis paraboloidal shape can be achieved step-by-step.

2.2.6 Metal and Lightweight Mirrors

Traditional mirror materials include glass ceramic materials, fused quartz, and other glasses. Nontraditional mirror materials include metals, metal alloys, SiC, and CFRP composites. The main motivation of using nontraditional materials is to reduce the weight and cost. Mirror material properties are discussed in next section. Liquid mirrors can also be formed by rotating liquid mercury inside flat dishes.

Metals or their alloys were used as optical mirror materials in the early days of mirror manufacture. They were replaced by glasses at the beginning of the 20th century because glasses have high surface smoothness and lower thermal expansion coefficients. Recent attempts at using metal mirrors include two Italian test optical telescopes: one is a 1.5 m one and the other is a 1.4 m one. The thermal sensitivity of a metal mirror is lower than that of a borosilicate (BSC) glass one as metals have high thermal conductivities. High thermal conductivity reduces temperature gradient inside a mirror. However, large metal thermal expansion produces large surface deformation.

Suitable metal mirror materials are aluminum, steel, titanium, beryllium and their alloys. The hardness of aluminum is low, so a coating of phosphor nickel alloy is used. The coated surface can be polished to required smoothness as used in a number of test optical telescopes. Steel and stainless steel are good metal mirror materials. Stainless steel with a hard alloy surface coating can be polished to optical surface quality. Beryllium and titanium mirrors have both

been used in space infrared telescopes. Beryllium is the mirror material of the James Webb Space Telescope (JWST) primary mirror as discussed in Section 5.3.2.

In general, metal mirrors are built by casting. However, a promising technique for large diameter mirrors is through micro-welding. The main obstacle of using metal mirrors in optical telescopes is the long-term shape warping. Many reasons produce warping of a metal mirror. Among these, the mirror's shape and thermal treatment are major ones. Asymmetrical mirror shape can produce larger warping, so that a feasible metal mirror shape is a meniscus of uniform thickness. The Italian 1.4 m aluminum mirror had a warping of one wavelength per ten years. Metal mirrors are easiest to build and the lowest in cost. If active optics is applied, metal mirrors may be used as candidates for future extremely large telescope mirrors.

Carbon Fiber Reinforced Plastics (CFRP) composite is another material for optical mirrors. The CFRP replica technique is a new achievement in the mirror manufacturing field. After 20-years practice, the CFRP replicated mirrors have been used in millimeter wavelength, infrared, optical, and X-ray telescopes. The great advantages of a CFRP mirror are light weight, high surface accuracy, high thermal stability, high surface smoothness, and low manufacture cost. The areal density of a CFRP mirror can be only a few kilograms per square meter.

To achieve a highly accurate mirror surface, a high precision mold is essential. A major problem in CFRP replication is the volume contraction of the resin material during the solidification process. The resin contraction is large and it produces mirror surface deformations. Therefore, it is necessary to reduce the resin contents in the CFRP mirror body. Other problems of the replication are air bubbles and the print-through of the ribs or fibers which overlap each other in the mirror surface. Air and water exist inside the resin in liquid form. During the solidification, water and air can turn into bubbles as temperature increases. Surface print-through is caused by the residual stress during the solidification process. The stress caused by overlapping of fibers in the mirror body may release and the fine print-through will appear on the mirror surface. All these reduce the smoothness and accuracy of the mirror. Without internal stresses, the surface smoothness of a CFRP replication mirror can be better than that of the mold surface used in the replication.

If a CFRP mirror's diameter is small, the mirror could be made by several symmetric layers of uni-directional fibers to form a meniscus shape. If the diameter mirror is large, a sandwiched structure should be used to guarantee the shape's stability. The sandwiched structure includes a top and bottom layer, both are curved in shape and a middle spacing part which can be formed from a number of CFRP short tubes with strictly the same length. The curing of CFRP parts should be done at a relatively low temperature. High temperature produces higher residual stresses. After the replication, the sandwich mirror is removed from the mold. The symmetry of fiber layers of each CFRP part should also be maintained to assure the mirror's long term stability.

Different carbon fibers have different thermal expansion coefficients (Section 8.3.1). For optical mirrors, low thermal expansion carbon fibers are preferred. Replication technology for a CFRP mirror depends largely on the mold's precision, workmanship, and proficiency of the technique. There are no unconquerable difficulties in the process. The replication technology can also be used in the manufacturing of deformable mirrors used in active or adaptive optics (Section 4.1.3).

Another CFRP replication method is to build a carbon fiber mirror blank and metal film on a mold surface first. The metal film is produced through electrical forming. The next step is to glue the metal film to the CFRP mirror blank using a thin layer of epoxy resin. This method is mainly used for mirrors of smaller aperture size.

Recently, silicon carbide (SiC) has been used as one optical or infrared mirror material (Section 9.1.3). A silicon carbide molecule is like diamond with half of the carbon atoms replaced by silicon atoms. SiC is an abrasive material. However, sintering (hot pressing), chemical vapor deposition (CVD), and reaction bonding lead to silicon carbide mirror blanks. One of these approaches is to obtain a soft blank through iso-static pressing of pure silicone carbide powder. The soft blank called "green-body" is workable to produce shape change. After milling the blank into its final geometry, the substrate is sintered at 2,000°C. The hardened segment is finally grounded and polished.

In CVD approach, gaseous chemicals react on a heated surface (often graphite) to form solid crystalline material. The process is slow, but it will produce a 100% dense, pure compound. This method can also produce a mirror surface with an integrated rib structure. However, the hardness of the compound makes mirror figuring and polishing time-consuming and difficult. In silicone carbide mirror polishing, diamond powder is the only abrasive used.

The reaction bonding is a cast and chemical process. First, high-grade silicone carbide is manufactured by chemical leaching to purify the base SiC abrasive. Leaching is a process of extracting a substance from a solid by dissolving it in a liquid. Then the powder is molded as the SiC is suspended in a silica-based gel. The substrate is heated to 950°C to remove the inert materials through evaporation. Small mirror blanks can be assembled to form a large mirror blank in this stage. Next, the substrate is heated again to 1,550°C in the presence of methane gas in vacuum. The carbonized substrate is immersed in molten silicon which fills the voids. This process produces a substrate of 83% SiC and it can be polished to a smoothness of 10 Å.

In this section, it is worth mentioning the rotational mercury mirror experiment. One project is the 6 m diameter Large Zenith Telescope (LZT) east of Vancouver, Canada. The project was developed from a 2.7 m one. This telescope mirror is a large plate filled with mercury, rotating at a constant rate over a precision air bearing. The plate container has a roughly parabolic shape to reduce the mass of mercury. The thickness of the mercury is only 1 mm.

The mercury has a relatively high reflectivity (∼80%). The rotating speed of the plate ω is directly related to the focal length $F = g/(2\omega^2)$, where g is the

Table 2.3. Thermal and structural properties of some mirror materials

	Al	Steel	Invar	BSC glass	Fused silica	Cer-Vit	CFRP	SiC	Titanium
$\alpha[K^{-1}]$ CTE	23×10^{-6}	11×	1×	3.2×	0.05×	0.05×	0.2×	2×	12×
$\lambda[W/mK]$ conductivity	227	251	10	1.13	1.31	1.61	10	150	21.9
$c[J/kgK]$ capacity	879	502	500	1047	770	821	712	670	523
$\rho[kg/m^3]$ density	2700	7750	8130	2230	2200	2530	1800	3140	4650
$\delta = \lambda/c\rho$	95×10^{-6}	6.5×	2.5×	.48×	.77×	.79×	7.8×	.86×	9.0×
δ/α	4.16	0.59	2.46	0.15	15.4	15.8	39.0	0.43	0.75
$E[N/m^2]$ modulus	72×10^9	210×	145×	68×	66×	91×	105×	430×	100×
v Poisson	0.34	0.28	0.30	0.20	0.17	0.24	0.32	0.15	0.36
$E/\rho g(1-v^2)$	307×10^4	299×	199×	322×	315×	389×	662×	1726×	252×

Note: $\delta = \lambda/c\rho$ is thermal sensitivity and $E/\rho g(1-v^2)$is bending stiffness. BSC glass is borosilicate glass.

acceleration of gravity (for a 6 m mirror, the speed of rotation is 8.5 s/revolution). However, this type of mirror also has drawbacks: (a) its location has to be far away from any vibration source. (b) The rotating speed has to be smooth and uniform. (c) The mirror can only be used in zenith position. If star tracking is required, the optical device at the focus is complicated. (d) Because of the low viscosity of mercury, the aperture size is limited. And (e) there is vaporization of mercury, contamination from sulfur and phosphate in air, and surface ripples caused by a gentle breeze. The last issue may be solved by clamping a thin stretched film of Malar over the top of the mirror surface.

At present, a Large Aperture Mirror Array (LAMA) with 66 individual 6.15 m mercury mirror telescopes is planned and another large mercury mirror on the pole area of the moon is proposed for astronomy.

Table 2.3 lists thermal and mechanical properties of some common mirror materials. Some special details of the mirror materials will be discussed in the next section.

2.3 Mirror Polishing and Mirror Supporting

2.3.1 Material Properties of Optical Mirrors

Mirror materials should have special properties in order to maintain a stable, high-precision surface shape. Ceramic materials, such as Cer-Vit, Zerodur, and fused quartz are major optical mirror materials. New mirror materials include CFRP, SiC, metals, and alloys.

What are the basic requirements of optical mirror materials? First, the material should have excellent shape stability so that the mirror can maintain its high precision shape over a very long period. Second, the CTE of the material should be close to zero so that the shape of the mirror will not change when temperature changes. Third, the material should have enough rigidity and hardness to sustain stresses induced during fabrication and transportation. And fourth, the material surface should be smooth after polishing and capable of being coated with a thin reflective metal film in a vacuum condition. Some soft materials, which cannot be polished, can also be used by coating a layer of hard material on their surfaces.

Mirror material selection is a tradeoff process. Many factors influence the material selection. These include mechanical and thermal properties, material availability, cost, weight, transportation, fabrication, and others. Space optical telescopes require light weight materials.

Surface smoothness, or roughness, is one important mechanical property for mirror material. Surface roughness is defined as surface height rms error in an extremely high spatial frequency (small scale) range. The roughness measurement of the optical surface requires an extremely high spatial frequency of 100–200 μm^{-1}. The surface roughness directly influences light scattering on the

surface. In the optical region, the total integrated scattering (TIS) is a function of surface roughness:

$$TIS = \left(\frac{4\pi\sigma}{\lambda}\right)^2 \tag{2.36}$$

where σ is the roughness and λ the wavelength. Since the TIS is inversely proportional to the square of the wavelength, mirrors used in optical and ultraviolet regimes require a very small surface roughness number (Figure 2.17).

Surface roughness is related to mirror material and mirror fabrication. Glass materials of fused silica or borosilicate glass have a surface roughness of 8 Å (Angstroms = 10^{-10} m) after fine polishing. After normal polishing the roughness achieved is of 25 Å. Stainless steel can reach a roughness of 40 Å, Invar of 47 Å, and aluminum of 53 Å. The roughness of a silicon carbide mirror can be 8–12 Å. These values are after fine polishing. Research indicates that the resulting surface roughness in fine polishing is related to the lubricant used. When using special lubricants, aluminum material can also be polished for optical telescopes. Table 9.1 of Section 9.2.1 lists the surface roughness number of some mirror materials. The requirement for the TIS of an optical mirror surface is around 10^{-3}.

Before the invention of glass ceramics and fused silica, the only material for optical telescope mirrors was borosilicate (BSC) glass. BSC has a relatively high thermal expansion coefficient and is still used for building large honeycomb mirrors. However, borosilicate glass has a very low thermal expansion coefficient

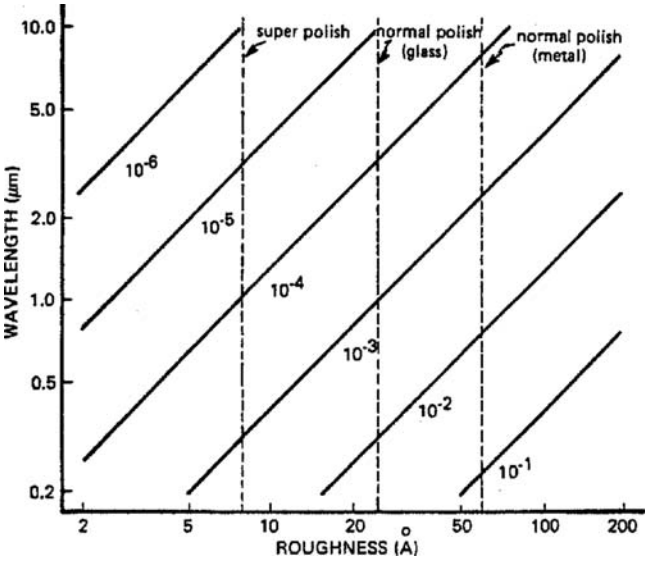


Fig. 2.17. Surface roughness and total integrated scattering.

(about $0.8\times10^{-6}\,\mathrm{k}^{-1}$) at a low temperature of 40 Kelvins (Section 5.3.2). Therefore, it is a candidate mirror material for modern space infrared telescopes.

Glass ceramic material is made by adding chemical additives (crystal seeds) into liquid glass to germinate fine crystals through thermal treatment. The crystals form a polycrystalline structure with an ultra-low thermal expansion coefficient. However, it is difficult to produce very large, thin, or special shaped glass ceramic mirror blanks due to residual crystallization stresses. It is also not possible to form a honeycomb mirror shape through this casting process.

Another mirror material is fused silica (or fused quartz made form quartz crystals). Fused silica is made by melting naturally high purity silica sand at around 2,000°C using either an electrically heated furnace (electrically fused) or a gas/oxygen-fuelled furnace (flame fused). Fused silica is translucent or opaque. A large mirror blank can be made by fusing small pieces of blanks together at about 1,500°C. It is the material for the honeycomb primary mirror of the Hubble Space Telescope. This primary mirror was made of five small pieces: top plate, bottom plate, inner annulus ring, outer annulus ring, and egg-crate core (Figure 2.18).

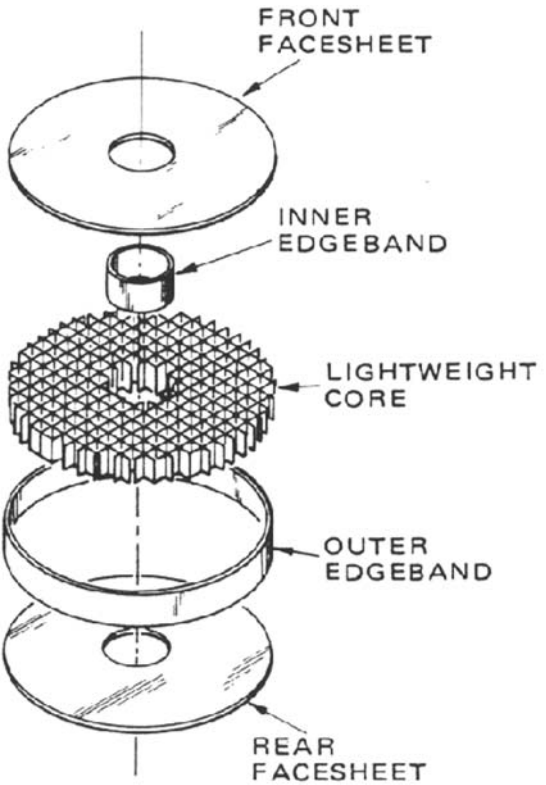


Fig. 2.18. Five components of the HST primary mirror.

Large and very thin primary mirrors of the SUBARU and GEMINI telescopes are also made by fusing hexagonal-shaped fused silica segments. Two steps are used in their manufacture: first to fuse all segments into a flat blank and, second, to soften the blanks into an ideal meniscus shape on a convex mold. Since the mirror is made of several segments, special attention has to be paid in the optimization of thermal distortion. There are small CTE differences between all segments. Different segment arrangement produces different surface rms errors for the same thermal loading. The loading includes absolute temperature change (range about 25°C) and axial temperature gradient (typically 3°C) both in production and in telescope operation. The optimization is through the finite element analysis.

2.3.2 Optical Mirror Polishing

Single point diamond turning (SPDT) is an efficient method for the manufacture of metal mirrors. The achievable surface accuracy and roughness using this technique are about 3 and 1 μm, respectively. These mirrors can only be used in the infrared region. They usually do not meet the requirements for large optical mirrors.

To produce glass-type optical mirrors, grinding and polishing are necessary. Four variables affect the removal of mirror surface material in the grinding and polishing process. These are the pressure, the relative speed, the contacting area between the mirror surface and lapping tool, and the abrasive used in the process. Improvement in any of these variables leads to an improvement in the grinding and polishing efficiency. A simplified model assumes linear relationships between the efficiency and any of the first three variables.

Polishing a parabolic surface is much more difficult than polishing a spherical surface because a good surface contact between the tool and the mirror blank is difficult to maintain for a paraboloid shape. The maximum deviation of a paraboloidal surface from a spherical one can be expressed as $0.00032D^4/F^3$ (where D is the mirror diameter in meters and F is the focal length in meters). This expression indicates that the larger the mirror diameter or the smaller the focal length is, the more difficult the mirror polishing will be.

At present, three methods exist in aspherical mirror grinding and polishing. The first one uses traditional grinding tools, the second one uses deformable grinding tools, and the third one involves a pre-stressed mirror blank.

According to the size of the lapping tools used, the first method can be further divided into three sub-classes: one using a full size tool, one using a medium size tool, and one using a small size tool. Using a full size tool to polish a spherical surface is easy. Large-size grinding tools have a large contact area with the blank resulting in high polishing efficiency. However, when the mirror shape departs from a sphere, the contact area required between the tool and the blank at each radius should be different as different amounts of mirror material need to be removed. This requires special contacting patterns on the lapping tool. The

tool pattern is related to the material removed from the mirror and the moving range of the lapping tool.

In general, a full size tool is difficult in polishing an aspherical surface of a small focal ratio. Therefore, a medium size tool is necessary. The medium size tool can remove material of a specific mirror radius. However, one major problem is that the surface under a medium size tool is usually asymmetrical. To correct this, the abrasives can be added only from desirable directions so that the mirror grinding is done in a particular part under the tool, but not in the other parts. This, unfortunately, provides only limited improvement. A small size tool can be easily used to modify the surface shape within a small radius range. Small tools are often used by experienced opticians when polishing large aspherical mirrors. However, care has to be taken as a small size tool introduces high spatial frequency ripples on the mirror surface. These ripples are difficult to remove and to be corrected. Now a small tool with computer control plays a very important role in the modern aspherical mirror fabrication.

A deformable tool can keep a good contact between the tool and the mirror blank when an aspherical surface shape is involved. Two types of deformable tools are used in optical manufacture, passive and active ones. The deformation of a passive tool is from the tool design. There are no external forces or moments applied on the tool except the gravity. When the 4.2 m William Hershel Telescope (WHT) mirror was polished, Brown designed a large full-size polishing tool with a number of deep ring ribs on the tool back. There was no radial rib connection between these rings. The bottom plate of the tool was very thin. With this structural arrangement, the tool was "soft" in the radial direction but "stiff" in the circumferential direction. During the mirror polishing, sandbags are placed on top of the tool to insure a good surface contact between the tool and the blank. Therefore, the contacting area increased and the aspherical surface shape was manufactured.

The active deformable tools involve force and/or moment actuators. The deformation is controlled in real-time through some positional and orientational encoders. R. Angel used an active deformable tool with force actuators for the manufacture of the Vatican f/1.0 primary mirror.

The third polishing method is called stressed polishing where a pre-stressed mirror blank is used instead of an unstressed mirror blank. The mirror shape required under a pre-stressed condition is only a simple spherical shape. However, after the surface has been polished, the desired complex mirror or lens shape can be obtained by releasing the preloaded stresses. The simple surface shape can be a plane or a sphere which is easy to make. The pre-stresses are from either vacuum or force actuators. The method has been used in the manufacture of the off-axis paraboloidal mirror segments of a SMT and the Schmidt corrector plates. Applying this method, iterations and additional corrections using plasma or ion polishing may be needed.

The formulation of pre-stressed polishing of an off-axis conic surface is in Section 2.2.5. For extremely large aperture segmented mirror telescopes, a study shows that the astigmatism is the only important term on the off-axis mirror

segments (TMT design report). Other terms are very small. The astigmatism can be easily eliminated by applying a moment across one mirror diameter. This approach simplifies the stressed mirror blank design for the stressed polishing. After a bending moment has been applied to the mirror blank, the polishing of mirror segments can be done on a planet-type polishing machine. This increases the efficiency and lowers the cost for the mirror segment polishing of an extremely large telescope. A planet-type polishing machine involves a large rotating flat lapping tool and a number of mirror blanks which are floating on top of the tool. Retaining rings (frames) are used to limit the mirror blank's motion so that the blank rotates about both the machine axis and the mirror axis. These two rotational movements produce a uniform material polishing of the mirror surface. If the flat surface tool is replaced by a spherical one, this planet polishing machine can be used for spherical mirror polishing. By putting a stressed mirror blank on top of the spherical tool, the machine can be used for the mass production of off-axis mirror segments for large segmented mirror telescopes. Stressed polishing is a trial and error method for achieving an accurate mirror surface. To avoid iterations, final ion beam polishing or plasma figuring are required.

One problem in mirror polishing is caused by the mirror deformation under the weight of the polishing tool and the mirror itself. To solve this problem, air cushions are used as the mirror support system. In this way, the mirror is floating on the top of the cushions so that the weight of the tool will not produce any local mirror deformation. An air cushion support usually has three axial symmetrical groups of pads arranged in rings. Air valves are used between groups to control the damping of the system. The mirror supported is in the same condition as if it were floating inside a fluid of the same density. This mirror support is called an astatic support.

By using an air-cushion support system the local surface deformation will be very small and it will not influence the surface precision. During the polishing of the UK 1.2 m Schmidt objective prism, a special viscous syrup bag was used for the very thin corrector support. The back side of this Schmidt corrector was glued to the syrup bag and its radial edge was constrained by roller bearings to avoid radial movement. This support arrangement produced a high quality, very-thin Schmidt object prism of 1.2 m size.

Cell print-through is a problem when a honeycomb mirror is under polishing. Honeycomb mirrors have a thin top surface and elastic deformation occurs when the polishing force is applied. This degrades the surface accuracy. To overcome the cell print-through, a special vacuum polishing tool can be used (Figure 2.19). The tool draws air out from the contacting surface between the tool and the mirror to eliminate the force applied on the mirror surface while the removal of mirror material is not affected. Polishing non-spherical surface with magneto-rheological fluid or ferro-fluid is a new technique. The viscosity of these fluids can be changed when magnetic field is applied. Therefore, the rate of material removal can be easily controlled.

The ion beam and plasma figurings are also important manufacture methods for astronomical optics. These two methods are mainly used in the final finishing stage of a mirror to achieve precise surface shape modification. The ion beam

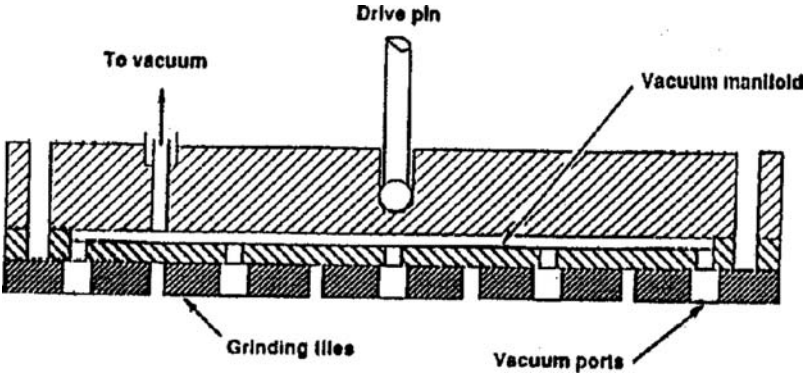


Fig. 2.19. A vacuum polishing tool used for honeycomb mirror manufacture.

figuring is a physical process of bombarding the mirror surface with high speed ions. The mirror should have a low surface roughness before the ion figuring is applied. The whole ion beam process is carried out in a vacuum chamber. The removal of mirror material by ions can be expressed by a beam removal function (BMF). One characteristic of this method is its noncontacting property. So the material removal speed has no relationship with the mirror surface shape. Usually the optical surface is facing down during the ion figuring. This method can achieve a surface precision of about 0.02 visible wavelengths. The main restriction of this method is the dimension of the vacuum chamber.

Plasma polishing is different from ion beam figuring because it is a process of chemical erosion using plasma gas. Some special gas in a plasma state is added in the polishing process, it reacts with the material on the mirror surface. Then the reaction produces active compounds, which detach from the mirror surface. For a fused silica mirror, the reaction is:

$$CF_4 + SiO_2 \rightarrow SiF_4 + CO_2 \tag{2.37}$$

The plasma itself has moment in the polishing which will further accelerate the chemical reaction. Plasma polishing can be carried out in a low vacuum condition and it can also produce a high surface precision. It is also a noncontacting polishing method. The efficiency of plasma polishing is higher than that of ion beam figuring.

The discussion in this section was mainly focused on the polishing of astronomical optical surfaces. Other manufacturing methods such as optical surface replications are discussed in Section 2.2.6.

2.3.3 Vacuum Coating

Vacuum coating is used to increase the reflectivity of a mirror surface. A metal material is evaporated onto the surface and becomes a thin layer of deposition.

Before the coating, the old film should be removed from the mirror surface. Then the surface is cleaned. The cleaning of the surface is a process of chemical and mechanical reaction. Usually a detergent or a mixture of mild sulfuric and chromic acids is used. After cleaning with acid, the mirror surface should be washed with water. Then the mirror dries in air. Some observatories also use dry ice for the mirror surface cleaning though this process is usually used between coatings to remove dust from the mirror surface.

The coating is done inside a vacuum chamber. The chamber is a large barrel-like container. If the mirror to be coated has a larger dimension, it is usually placed vertically in the chamber. In this position, no metal fuses or other objects will fall on the mirror surface. Metal fuses are arranged at equal distances around the mirror. Then the chamber is evacuated. When the air pressure reaches 10^{-4} to 10^{-5} mmHg (1 mmHg $= 1.33\times10^{-2}$ Pa) and can be maintained at this level, the coating process can be started. If a single layer of aluminum film is needed, the coating material, usually aluminum filaments, is placed above a few tungsten heating coils. When the temperature of the filaments is over 600°C, the aluminum melts and attaches to the heating coils. When the temperature reaches 1,200°C, the aluminum evaporates. The evaporating aluminum molecules radiate to the mirror and deposit on the surface. In the visible wavelength range, aluminum coating is widely used. For infrared wavelengths, gold or silver has a higher reflectivity. The obvious drawbacks of the silver coating are a low adhesive force and the tendency to oxidize. These can be solved by an additional coating of Al_2O_3 or SiO_x. Gold, or silver, or platinum are also used for mirrors in X-ray imaging systems.

2.3.4 Mirror Supporting Mechanisms

The basic goal of a mirror support is to hold the mirror in the telescope so that the forces of gravity, wind, and telescope acceleration do not significantly change the surface and the position of the mirror. The mirror support includes positioning ones and floating ones. The position of a mirror is defined by a few positioning support points (hard points). The positioning support and its related displacement actuator carry a very small portion of the mirror weight. Most of the mirror weight is carried by "floating" supports to avoid the mirror surface deformation. The floating support is known as astatic flotation which mimics the buoyant force felt if the mirror were floating in a liquid of its own density. The direction of the gravity load of a mirror changes as the elevation changes, so that both axial and radial positioning and floating support systems are used in the mirror support system.

2.3.4.1 Positioning Support Systems for Optical Mirrors

Any rigid body has six degrees of freedom. Therefore, the best mirror support is the so-called "kinematic" mounting, which fixes just six rigid body degrees of

freedom of a mirror. These six degrees of freedom can be applied on a single point, but stresses will be produced around it, or on three or more points. In some cases, the axial and radial positioning points are separately grouped, each with three support points. In some cases, there is no constraint in the mirror axial rotational direction, resulting only five constraints of the mirror positioning system.

Positioning support points can be chosen near the outer edge, or the middle radius, or the inner radius (at the central hole of a primary mirror) of a mirror. In general, the mirror and its cell are made of different materials so that differential thermal expansion may happen when temperature changes. This thermal effect is small when the mirror uses the central hole as its positioning location. The effect will be serious when support radius increases. However, there are two cases where the thermal effect is not a problem even for outer edge positioning. One is when both the mirror and its cell have low CTEs and another is where the positioning constraint degrees of freedom are not affected by the differential thermal expansion.

The HST primary mirror has its positioning device at its outer edge. Above the mirror, the constraint is from a zero-expansion CFRP tube truss. Below the mirror, the constraint is from a low-expansion titanium alloy mirror cell so that the relative movement between the mirror and its positioning device is very small as temperature changes. The advantage of placing the positioning support points on the outer edge is that the mirror will have a higher resonant frequency. For space telescopes, since there is no gravity, the weight of the tube truss would not produce deformation of the mirror when the telescope is in orbit.

For many ground-based telescopes, the mirror cells are made of steel. Radial shear forces may be produced due to differential thermal expansion between the mirror and cell. Therefore, most mirrors use the central hole for positioning location. The position defining points bear little of the mirror's weight. To avoid mirror surface deformation caused by small friction force, the axial and radial contact areas in the central hole positioning system are very small. In the radial direction, the mirror positioning is through a thin tube extended from the mirror bottom support plane. The contacting part is a spherical surface inside the inner mirror hole. To further reduce the contact stress, several vertical slots are made on the sphere surface to absorb any possible stresses.

For mirrors with a small aperture size, the mirror positioning points may be located at the middle radius. These points are on the back of the mirror. This arrangement can be found in a number of secondary mirror support systems. However, if the mirror diameter is not so small, then the force caused by differential thermal expansion remains a problem. An improvement can be made by adding radial flexible springs at the positioning support points. These springs absorb thermal stress between the mirror and its cell.

The three-point mirror support can evolve into a six-point, or nine-point, or more point mirror support through a whiffle-tree design. A whiffle-tree is a beam, or plate, structure, which distributes the support force from one point to two, or three, ends of a beam, or a plate. This force redistribution can be cascaded as a tree structure. However, the degrees of freedom involved are kept the same as a single point support. Differential thermal expansion also exists in a

whiffle-tree support system. To overcome this, the beams which transfer the mirror load to the positioning points should have the same thermal expansion coefficient as the mirror material. Invar is a favorable material used for whiffle-tree beams.

The mirror middle radius positioning is mostly used for mirror axial positioning. The radial mirror positioning is usually at the outer edge of the mirror. Generally three clockwise or anti-clockwise linkage bars in the tangential direction can be used. One end of these linkage bars attaches to the mirror and the other end attaches to the cell. These three tangential linkage bars will fix the mirror in the radial direction. It allows dimensional variation between the mirror and the cell. If the mirror is in zenith position, these linkage bars are free from any loading. When the mirror tilts, the linkage bars will generate a lifting force to counteract the component of the mirror weight along the radial direction. Temperature change and differential thermal expansion have no influence on this type of design. The link bar positioning system allow rotation in axial direction. It constrains only two degrees of freedom.

For very thin mirrors, more positioning support points are required. These support points can also take additional loads, increasing the mirror stiffness, but they will not produce deformation of the mirror surface. These support points are usually equipped with sensors for active or adaptive mirror support force or position control.

A new style of mirror positioning has been developed from the Stewart platform (Parks and Honeycutt, 1998). The basic principle of the six-beam Stewart platform will be discussed in Section 3.1.3. In this hexapod platform,

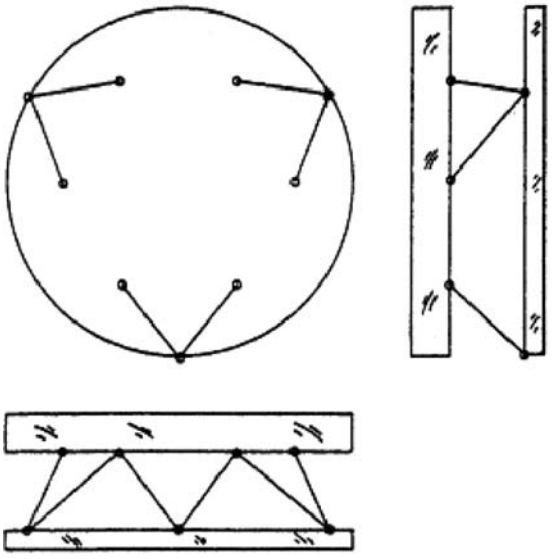


Fig. 2.20. Hexapod axial support system for optical mirrors (Parks and Honeycutt, 1998).

each rod has universal joints on both ends. The six rods provide six degrees of freedom for a stable mirror positioning support. Figure 2.20 illustrates a hexapod mirror supporting structure. In this support, the length of each supporting rod is relatively long so that an axial movement of the mirror will be produced when temperature changes. To reduce this temperature effect, the supporting rod can be bent into an 'L' shape so that the distance between the mirror and the cell reduces. If there is a radial force component, a hexapod platform may produce astigmatism of the mirror. Therefore, the hexapod platform support is not an ideal solution for the mirror's radial support.

Parks and Shao extended this hexapod support system to a more complicated 18-point mirror support system. The 18 support points are arranged in two rings. These rings have radii of 0.408 and 0.817 of the mirror radius. Three groups of six points are formed with 12 points at the outer ring and six at the inner ring. One hexapod support device is used for each support group so that the weight of the mirror is evenly distributed to all 18 points. To avoid over-constraint of the mirror, every hexapod platform is connected to a Y-shaped cell with a wire rope in tension. These wire ropes and the connected

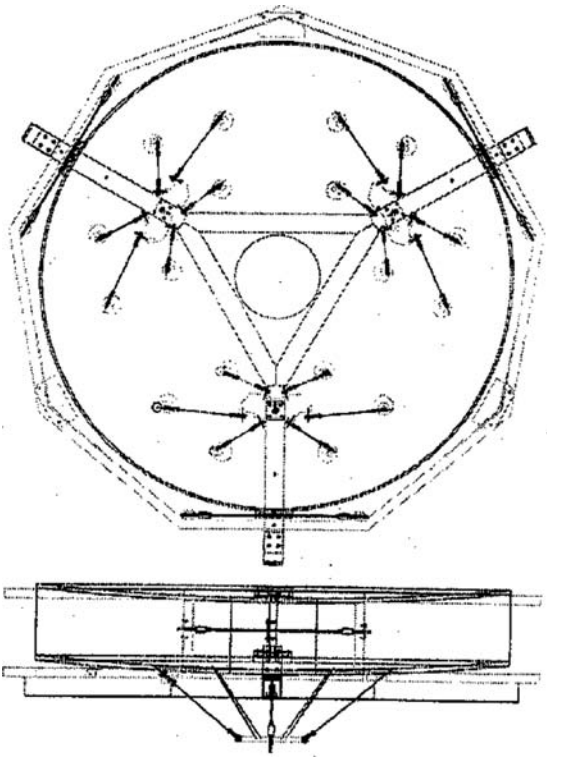


Fig. 2.21. An 18-point supporting device and its six-beam linkage subsystem (Parks and Honeycutt, 1998) (Note: Pre-stressed steel wires are used for radial supporting).

hexapods provide a positioning of just six degrees of freedom for the whole mirror and it turns out a very stable mirror support system.

To reduce the support system weight, the lower platforms are made of a triangle truss. In this design, any two of all six rods should not align through one common point for maintaining its stability. Figure 2.21 shows the arrangement of this 18-point hexapod support device. In the radial direction, pre-stressed steel wire ropes are used.

2.3.4.2 Flotation Support Systems for Optical Mirrors

In general, most of the mirror weight is taken up by floating support mechanisms. There are two types of flotation support systems: mechanical and pneumatic ones. A mechanical mirror support system usually involves a counter-weight and cantilever mechanism. The support force generated by this counter-weight and cantilever system follows a sine law of the mirror's elevation angle. This is the same law governing the axial force component change of the mirror weight. Therefore, no force adjustment is necessary in a normal passive support system. The cantilever length ratio produces a magnification of the load applied to the mirror, therefore, the counter-weight required is smaller. The mechanical flotation support system can be used on both axial and radial support. For radial support of a thin mirror, a thin membrane can be used to transfer the support forces from the cantilever system to the support point in radial direction as shown in Figure 2.22. This avoids the effect on the mirror from the bending moment of the system. The main problem of a mechanical flotation support system is the friction involved. It affects thin mirrors.

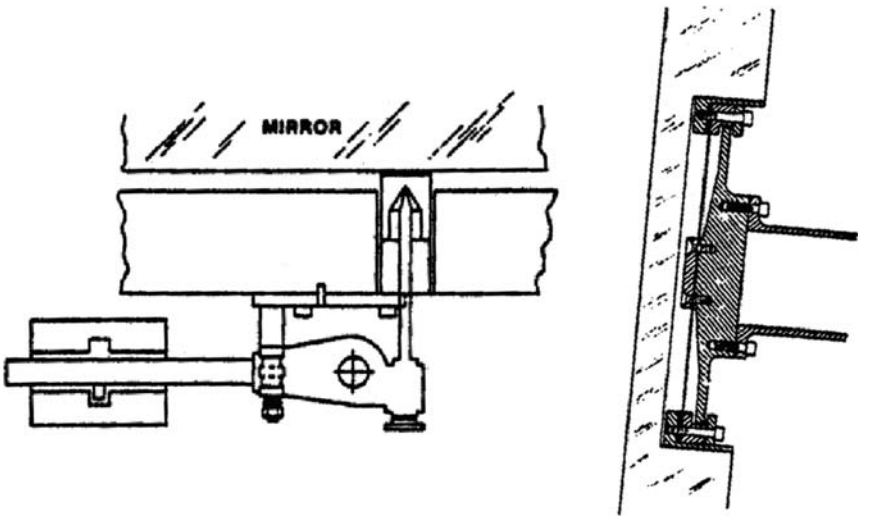


Fig. 2.22. Axial and radial counter-weight and cantilever support systems (Keck).

Air cushions or air cylinders are pneumatic flotation mirror support systems. They are mainly used in axial direction. The support force of these systems is proportional to the mirror contacting area, therefore, the contacting area should remain constant during the telescope operation. The air pressure is adjusted by a pressure regulator so that the required supporting force can be obtained. The regulator follows the motion of the tube to produce sine-law governed air pressure. Force sensors used in the support points may also provide signals to control the air pressure. The air cushions are soft allowing height and tilt adjustments, producing smaller friction forces. When the pneumatic system is not pressurized, the mirror rests on a set of spring-loaded rest pads.

For radial mirror support, a mercury bag is often used. In this system, the mirror is surrounded by a ring-shaped bag filled with mercury. The bag is held by the mirror cell while the mirror floats inside the bag. The force applied on the mirror is proportional to a constant contacting width of the mercury bag.

Another flotation support is a vacuum secondary mirror support system. The principle of a vacuum support is the same as an air cushion system but with a negative air pressure.

In the mirror support system, force sensors can be used to measure the supporting forces. One type of force sensor is the strain gauge. The usage of force sensors is essential for active mirror surface control as discussed in Section 4.1.3.

2.4 Mirror Seeing and Stray Light Control

2.4.1 Mirror Seeing Effect

Generally, seeing effect is produced by the density inhomogeneities in air along the optical path. Thermal nonuniformities are the main reason behind the air density and air refractive index variation. When a mirror surface has a different temperature from the surrounding air, convection will dissipate these heat nonuniformities. Two types of convection occur over a horizontally placed surface, a natural one and a forced one. Natural convection produces large-scale air bubbles, while forced convection has a thin boundary layer, small scale eddies, and fast time scales (Figure 2.23). The type of air convection can be described by Froude number, which is the ratio of Reynolds number squared to Grashof number, both are introduced in Section 8.1.3 (Dalrymple, 2002):

$$Fr = \frac{\mathrm{Re}^2}{Gr} = \left(\frac{VL}{v}\right)^2 \frac{\rho \cdot v^2}{\Delta\rho g L^3} = \frac{\rho V^2}{\Delta\rho g L} \tag{2.38}$$

where V is the wind velocity, L a length scale, v the kinematic viscosity, ρ the air density, g the gravity, and $\Delta\rho$ the magnitude of the density fluctuation of the air.

For $Fr >> 1$, the forced convection dominates; for $Fr << 1$ the natural convection dominates; and for $Fr = 1$ the convection is mixed. For a heated

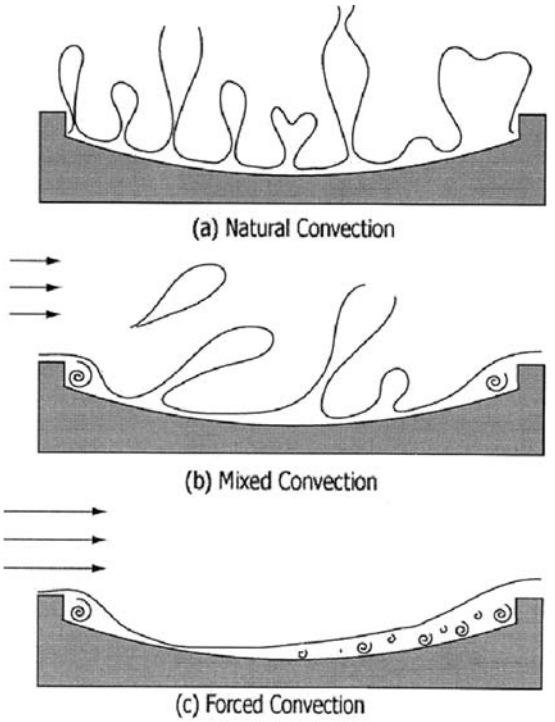


Fig. 2.23. Natural and forced convection air flow over the mirror surface (Dalrymple, 2002).

mirror, the length scale is the same as the diameter of the mirror, $L = D$. If the air pressure remains constant, then $\Delta\rho/\rho = \Delta T/T$, where T is the temperature and ΔT the temperature difference. One can map the convection regimes for particular mirror length scale in wind velocity and temperature difference space base on the Froude number. The Froude numbers between 0.1 and 10 correspond roughly to mixed convection, and higher and lower Froude numbers correspond to forced and natural convection respectively. It is suggested the natural convection produces the most aberration and the forced convection the least.

Air density fluctuation affects optical beams in different ways. For small-scale turbulence, image energy scatters widely and Strehl ratio reduces; for intermediate-scale turbulence, it will produce beam spread and image blurring, resulting in loss of both resolution and contrast; for large-scale turbulence, it will produce tilt-induced image shift, as jitter. A general pattern is a composition of all these three effects. Fast tip/tilt correction can remove jitter (Section 4.1.5). The mirror convection is better kept in the forced convention region, so that the boundaries are smooth and flat.

For a precise expression, the wavefront variance due to the air density change is:

$$\sigma^2 = 2G^2 \int_0^{L_{opt}} <\rho'^2> l_z dz \tag{2.39}$$

where G is the Gladstone–Dale parameter ($G = 0.22$ cm^3/g over the optical wavelength), ρ' the fluctuating density, which is roughly 10% of the total density variation $\Delta\rho$ in the air flow, $\rho' = 0.1\rho\Delta T/T$, l_z the correlation length along the optical axis, and L_{opt} the total path length through the disturbance. In many cases, $l_z = 0.1 \sim 0.2L_{opt}$. The total path length through the disturbance is related to the disturbance layer thickness above the mirror; for natural convection, it is of the scale of the mirror's diameter or larger. The formula of the disturbance layer thickness is:

$$L_{opt} \cong 0.184 \frac{L^{1.5}\Delta T_C^{0.5}}{V} + 0.0392 \frac{L^{0.8}}{V^{0.2}} \tag{2.40}$$

where L is the upstream heated length (m), ΔT_C the average temperature difference over the length ($^\circ C$), and V the wind velocity (m/s). For a 4 m diameter mirror in the natural convection case and the wavelength of $\lambda =$ 550 nm, the phase error is about:

$$\phi = \frac{2\pi\sigma}{\lambda} \approx 0.2\pi\rho \cdot G \frac{\Delta T}{T} \frac{\sqrt{2l_z L_{opt}}}{\lambda} \approx 0.48\pi\Delta T \tag{2.41}$$

In the forced convection case, the turbulent flow thickness is much smaller than that of the natural convection. The boundary layer thickness over a flat plate is:

$$\delta = 0.37 \mathrm{Re}_x^{-0.2} x; \quad \mathrm{Re}_x > 10^5 \tag{2.42}$$

and for a velocity of 1 m/s at x = 4 m, this is 12 cm. Higher wind velocity reduces this even further. In general, $L_{opt} = \delta$, and $l_z = 0.1\delta$. We see that the wavefront error is down from a natural convection case by one to two orders of magnitude.

When the temperature variation is $\Delta T = 1K$ for a natural convection case, the wavefront error is small ($\sigma < \lambda/\pi$) and the mirror seeing or the blur angle is:

$$\theta_M = \frac{\theta_D}{\sqrt{S}} \tag{2.43}$$

where θ_D is the diffraction limit image angle, $\theta_D \cong 2.4\lambda/D$, and S the Strehl ratio. The Strehl ratio for this weak aberration case is:

$$S \cong \exp\left[-\left(\frac{2\pi}{\lambda}\sigma\right)^2\right] = e^{-\phi^2} \tag{2.44}$$

Therefore, the mirror seeing is $\theta_M \cong 0.2 arc\,\sec$.

In a strong aberration regime, the central core of the point spread function is obscured and the signal is made up of scattered energy and system noise. The blur full angle containing p percent of the encircled energy is:

$$\theta_{M,p\%} = \frac{4\sigma}{l_z}\sqrt{-\ln(1-p)} \tag{2.45}$$

For $p = 50\%$, $\theta_{M,50\%} = 3.33\sigma/l_z$. If the temperature variation is $\Delta T = 2\,K$, the wavefront error is strong ($\sigma \geq \lambda/\pi$) and the mirror seeing is:

$$\theta_{50\%} \cong 0.45 arc\,\sec \tag{2.46}$$

Racine's experimental formula of the mirror seeing is: $\theta = 0.4(T_M - T_e)^{1.2}$, where T_M is the mirror temperature and T_e is the air temperature (Dalrymple, 2002). Figure 2.24 shows the mirror seeing as a function of temperature difference between the mirror and surrounding air.

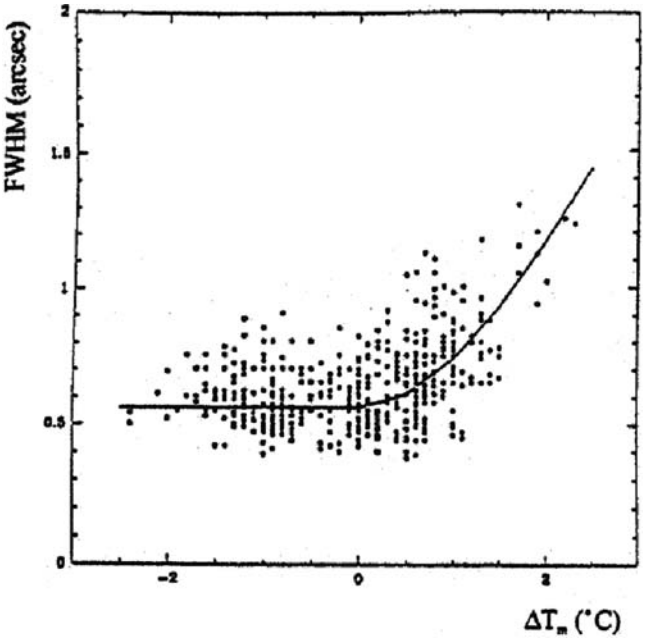


Fig. 2.24. Relation between mirror seeing and difference between mirror and air temperatures (Mountain et al., 1994).

2.4.2 Stray Light Control

Stray light is any light which does not come from the celestial target sources and yet illuminates the detector. Stray light creates an unwanted background and lowers the sensitivity. For optical telescopes, the source of stray light is the light from celestial objects outside the field of view and the light inside the field of view which does not go to the right position on the focal plane. Both lights are called "off-axis" sources. To overcome stray light, proper design of baffles and stops is necessary. Ray tracing is a way to predict the unwanted stray light. However, for infrared and millimeter wavelength telescopes, thermal emission of the telescope and the surrounding surfaces, including baffles and stops, is a major source of stray light. To overcome these thermal emissions, infrared telescopes may require a design with no baffles in their optical system.

Ray tracing starting from the focal detector is the most effective way of finding and eliminating the stray light in telescopes. In optical system design, ray tracing usually starts from the object space. However, this is not effective in finding stray light in a system. Ray tracing from the detector is like positioning oneself at the detector and looking outward. The first step is to determine the sources which are out of the field of view and still can be seen directly. To block these sources, baffles and stops are required. The next step is to find any object, optical or structural, visible to the detector directly or by the reflection of the optical surfaces. These objects are called "critical objects." The last step is to find any object, which is seen by the detector and illuminated by stray light sources. These are called "illuminated objects." If an object is on both the critical and illuminated list, it is on a first-order stray light path. For these objects it is necessary to move them away or to block them. In this way stray light can be reduced by factors of 100 or more. For objects not on the first-order path, the paths with most power must be blocked or removed. However, second-order stray light paths are much more numerous and further ray tracing is necessary.

Stray light ray tracing programs usually use a Monte Carlo approach. A random number generator is used over a selected area, to select only a few random rays to represent all the possible rays in the area, both in position and in direction. In the ray tracing process, each time a ray intersects an object; additional reflected, refracted, and scattered rays are generated. If the secondary rays are shot towards a light source, then the brightness of the surface where the primary ray intersects should be calculated. The power of the primary rays is weighted by the surface scattering rate. This stray light ray tracing is almost the same as the ray tracing in the computer graphical render program. The process is time-consuming because of the intersection calculations. Several approaches can be used to speed up the computations. These are: (a) Use faster computers; (b) Use specialized hardware, especially parallel processors; (c) Speed up computations by using more efficient algorithms; and (d) Reduce the

number of ray-object computations. The ray-object computation reduction includes adaptive depth control, bounding volume, and first-hit speedup.

2.4.2.1 Baffle and Stop Design

There are different stops in an optical system. The aperture stop, or the entrance pupil, limits the size of the incoming beam. Objects in the space outside the desired beam are not seen by the detector. The aperture stop is usually the edge of a primary mirror. However, in infrared telescopes, the aperture may be located at the secondary mirror. The field stop limits the field of view. The field stop is located at the focal plane.

Baffles are usually conic or cylindrical tubes designed to block unwanted light paths. To further suppress scattered light, the baffle sides facing the detector may have a series of concentric rings, called "vanes." For a Cassegrain system, two sets of baffles are required. One is around the secondary mirror, and the other is above the primary mirror as shown in Figure 2.25. The dimensions of the baffles can be found from the following formulae (Bely, 2003):

$$\begin{aligned} x_u &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} \\ r_u &= x_u(\theta - \theta_0) + \theta_0 f_1 \\ x_l &= \frac{-c_1 b_2 + b_1 c_2}{a_1 b_2 - a_2 b_1} \\ r_l &= \frac{-c_1 a_2 + c_2 a_1}{b_1 a_2 - b_2 a_1} \end{aligned} \tag{2.47}$$

where f_1 is the primary mirror focal length and θ_0 is the semi-angle one wishes to protect. The other parameters are:

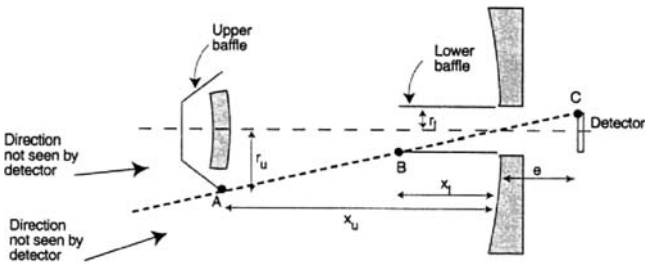


Fig. 2.25. Baffle design for Cassegrain systems (Bely, 2003).

$$
\begin{aligned}
\theta_0 &= D/2f_1 \\
a &= \theta_0^2(f_1+e)^2(m+1) + \theta_0\theta(f_1+e)(mf_1(m-1) - e(m+1)) \\
b &= -(f_1+e)^2\theta_0^2((2m+1)f_1 - e) - \theta_0\theta(f_1+e)((mf_1)^2 + e^2) \\
&\quad + f_1\theta^2(f_1^2(m^3 - 3m^2) - 2f_1e(m^2 - m) + e^2(m+1)) \\
c &= \theta_0^2(f_1+e)^2 f_1(mf_1 - e) \\
&\quad - f_1\theta^2((mf_1)^3 + 2f_1^2e(m^2 - m) - f_1e^2(m^2 - m) - e^3) \\
a_1 &= \theta_0(f_1+e) - \theta(m^2f_1 + e) \\
b_1 &= -(f_1+e)m \\
c_1 &= \theta_0(f_1+e)e + \theta(m^2f_1^2 - e^2) \\
a_2 &= \theta_0 x_u - \theta_0 f_1 - f_1\theta \\
b_2 &= -f_1 \\
c_2 &= -\theta_0 f_1 x_u + \theta_0 f_1^2
\end{aligned}
\tag{2.48}
$$

where D is the diameter of the primary mirror and m is the magnification of the secondary mirror. Baffle surfaces usually have a diffuse black coating to absorb the incoming light. However, none of these coatings will absorb all of the light. At normal incidence, the absorption is a constant. At other angles, surface scattering occurs. When the incident angle is near 90°, the scattering increases to values larger than unity at the specular direction. This scattering can be controlled by placing zigzag vanes to make all the light strike the baffle at a normal incidence.

2.4.2.2 Stray Light Analysis

Specular reflection and scattering are two different but related surface properties in optics. Specular reflection occurs on an ideal reflecting surface or mirror (Figure 2.26). It follows the law of reflection. The optical design is based on specular reflection. The scattering of a surface is described by a bidirectional reflective distribution function (BRDF). BRDF is a ratio between radiation scattered of a unit angular area and surface irradiation weighted with the cosine of the projected

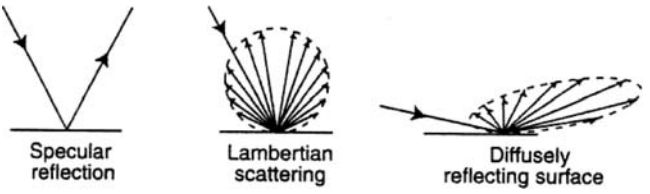


Fig. 2.26. Reflecting and scattering from surfaces (Bely, 2003).

solid angle. For an observer viewing from a different polar angle, the projected solid angle of a surface irradiance area is the solid angle of the area multiplied by a cosine of the polar angle. The expression of BRDF is (Bennett and Mattsson, 1999):

$$BRDF = \frac{dE_s/(Ad\Omega_s \cos\theta_s)}{E_i/A} \approx \frac{E_s/\Omega_s}{E_i \cos\theta_s} \tag{2.49}$$

where E_s is the radiation over an angular area Ω_s with a reflecting angle of θ_s, A the illuminated area on the surface, and E_i the surface total irradiation at a point.

If the surface is a perfectly diffuse reflector, light is scattered uniformly, the intensity of the scattered beam varies as the cosine of the angle from the normal of a surface. This is called "Lambertian scattering." The intensity (the photons per second) is the same for Lambertian scattering which has a constant of BRDF $= (1/\pi)\ sr^{-1}$.

An important property of the BRDF is that the half sphere surface integral of the product of BRDF and cosine of the polar angle must be less than or equal to unity. The integration is the reflectance ratio or total integrated scattering of a surface:

$$\int_{\Omega} BRDF \cos\theta \cdot d\Omega = \iint DRDF \cos\theta \sin\theta \cdot d\theta \cdot d\phi \leq 1 \tag{2.50}$$

Another less well-defined parameter is bidirectional scattering distribution function (*BSDF*) which is the scattered power per unit solid angle divided by the incident power:

$$BSDF = \frac{dE_s/d\Omega_s}{E_i} \approx \frac{E_s/\Omega_s}{E_i} \tag{2.51}$$

The BSDF simply uses the cosine-corrected scattered radiance rather than solely the surface irradiance (which has the effect of removing the factor of $\cos\theta_s$ from the projected solid angle) to yield scatter per unit illuminated surface area per unit solid angle.

All surfaces used in telescopes are in between these two types of scattering. The scattered light is concentrated in the specular direction, but a significant portion of it is around this direction. The flux transferred from a small scattering surface of area dA into an elementary solid angle $d\Omega$ can be expressed as:

$$d\Phi = BRDF \cdot E_i dA \cos\theta_i \cos\theta_s d\Omega \tag{2.52}$$

where θ_i is the incident angle and E_i the incident flux density. The BRDF depends on polarization and wavelength. A perfect surface produces specular reflection and has a BRDF infinite in the reflection direction. For lenses and windows, a bidirectional transmission distribution function, BTDF, is used.

When the BRDF of a surface is known, one can calculate the amount of power that is scattered from one surface to another. It is:

$$\begin{aligned} P_c &= \pi \cdot P_s(BRDF)(GCF) \\ GCF &= A_c \frac{\cos\theta_s \cos\theta_c}{\pi \cdot R_{sc}} \end{aligned} \tag{2.53}$$

where P_s is the incident power on the scattering surface area, R_{sc} the distance between the scattering and scattered surface area, $\theta_s$the scattering angle, and θ_c the scattered angle. GCF is the geometry configuration factor.

The BRDF of a mirror surface is related to the roughness of the surface. However, at infrared wavelengths, dust becomes dominant in scattering. The dust percentage is related to the cleanliness level. It is not practical to have a cleanliness level higher than 500 for large optics. The dust coverage of this level is 1%.

References

Bely, P., 2003, The design and construction of large optical telescopes, Springer, New York.

Bennett, J. M. and Mattsson L., 1999, Introduction to surface roughness and scattering, 2nd edn, Optical Society of America, Washington D. C.

Cheng, J. and Humphries, C. M., 1982, Thin mirrors for large optical telescope, Vistas in astronomy, 26, 15–35.

Classen, J. and Sperling, N., 1981, Telescopes for the record, Sky and Telescope, Vol. 61, Apr. 1981, 303–307.

Dalrymple, N. E., 2002, Mirror seeing, ATST project report #0003, NOAO.

ESO, 1986, Very Large Telescope Project, ESO's proposal for the 16 meters very large telescope, Venice workshop, 29, Sep. 2, Oct.

Hill, J. M., 1995, Mirror support system for large honeycomb mirrors, UA-95-02, Large Binocular Telescope tech memo, steward observatory, University of Arizona.

Lubliner, J. and Nelson, J., 1980, Stressed mirror polishing. 1 a technique for producing nonaxisymmetric mirrors, Applied Optics, 19, 2332.

Mountain, M. et al., 1994, The Gemini 8 m telescopes project, SPIE 2199, 41–55.

Nelson, J. E., Lubliner, J. and Mast, T. S., 1982, Telescope mirror supports: plate deflection on point supports, UC TMT Report No. 74, The University of California.

Nelson, J. E., Mast, T. S. and Faber, S. M., 1985, The Design of the Keck observatory and telescope, Keck Observatory Report No. 90, the University of California and California Institute of Technology.

Parks, R. E. and Honeycutt, K., 1998, Novel kinematic equatorial primary mirror mount, SPIE 3352, 537–543.

Swings, J. P. and Kjar, K. eds., 1983, ESO's Very Large Telescope, Cargese, May.

West, S. C. et al., 1997, Progress at Vatican Advanced technology telescope, SPIE Proc. 2871, 74–83.

Chapter 3
Telescope Structures and Control System

This chapter provides a comprehensive discussion on telescope structural design and analysis. Different telescope mounting designs are discussed in this chapter. Emphasis is placed on the altitude-azimuth mounting system. Formulas for star coordinator transformation and the zenith blend spot determination are provided. Formulas of a Steward platform are also introduced. Detailed information on telescope tube structure, secondary mirror vane structure, bearings, encoders, drive system, and control system design are also discussed. In the encoder part, a number of methods for increasing the encoder resolution are provided. The telescope pointing, tracking, star guiding, and the pointing correction formulas are discussed. In the final part of this chapter, static and dynamic structural analysis, wind and earthquake influence on structure, structure vibration control, and foundation design are all introduced. These contents are also useful for radio or other wavelength telescopes.

3.1 Telescope Mounting

3.1.1 Equatorial Mounting

Many existing small or medium size optical telescopes use equatorial mountings (mounts), including the largest 5 m Hale optical telescope at Mt. Palomar. If radio telescopes are included, the Green Bank 47 m, is the largest equatorial mounting telescope. The most important feature of an equatorial mounting is that one of its axes is parallel to the earth rotation axis, known as the right ascension or the polar axis. The other axis, the declination, is perpendicular to the polar one. With this axial arrangement, relative motion of celestial bodies is compensated by a constant polar axial movement. The field of view of an equatorial mounting is stationary and there is no zenith blind spot, which exists on an alt-azimuth mounting telescope. The zenith position always has the best observing condition in comparison with other sky positions.

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_3,

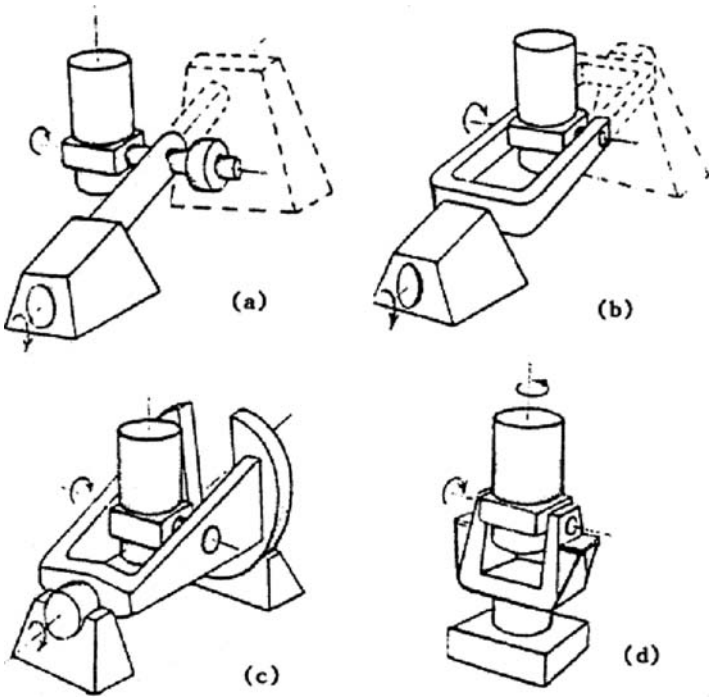


Fig. 3.1. (**a**) Asymmetrical equatorial mounting, (**b**) yoke and fork mounting, (**c**) a horseshoe mounting, and (**d**) alt-azimuth mounting.

Generally, there are two types of equatorial mountings: symmetrical and asymmetrical ones. Asymmetrical equatorial mountings, including the German and the English design Figure 3.1(a), are suitable for small aperture telescopes. As shown in the figure, the German mounting has only one polar bearing in solid lines and the English mounting has two polar bearings, shown in added dotted lines. The tube of an asymmetrical mounting is away from the polar axis, and a counter-weight is added on the other side of the axis for structural balance. The German mounting is a cantilever design suitable only for small telescopes. The English mounting can be used for medium size (2–4 m) telescopes. However, part of the polar region is blocked by one polar bearing, the English mounting is only suitable for low latitude observatory sites.

The tube bending in an asymmetrical equatorial mounting is complex, therefore, large telescopes use symmetrical mountings. There are three symmetrical equatorial mountings: the yoke design, the fork design, and the horseshoe design. A yoke mounting has two polar bearings and the tube is inside a yoke frame [dotted line in Figure 3.1(b)]. A fork mounting has only one polar bearing and two fork arms hold the tube as cantilevers [solid line in Figure 3.1(b)]. The horseshoe mounting has one polar bearing in the back and one giant horseshoe

hydrostatic bearing in its front [Figure 3.1(c)]. Please note that a fork structure used in an alt-azimuth mounting is commonly referred to as a yoke structure.

A fork mounting was developed from the German design and a yoke mounting from the English design. With a yoke mounting, the polar region is blocked by one polar bearing. To access the polar region with this design, tube offset is necessary as used in the 3.8 m UKIRT in Hawaii. The fork mounting is popular and is further divided into simple fork one and polar-disk fork one. The latter design has a stronger disk block in front of the polar bearing. The yoke mounting is used in low latitude sites and the fork one is used in high latitude sites.

The horseshoe mounting for very heavy equatorial telescopes include a simple horseshoe and a yoke-style horseshoe one. The structure of horseshoe is always massive and heavy, so that hydrostatic bearings are used. Hydrostatic bearings can support very heavy loads and have very little friction (Section 3.2.4).

3.1.2 Altitude-Azimuth Mounting

3.1.2.1 Mechanical Advantages of an Alt-Azimuth Mounting

Among equatorial mountings, symmetrical fork design is widely used, while asymmetrical ones are only used for very small telescopes. However, the cantilever fork arm of a symmetrical equatorial mounting has its own problem when it is at different hour angles (Figure 3.2). At zero hour angle, the bending moments of both cantilever arms of the fork appear in the meridian plane and a symmetrical deformation is produced. At an hour angle of 6^{h}, the bending moments are in the fork arm plane and an asymmetrical deformation is produced. Nonconsistent

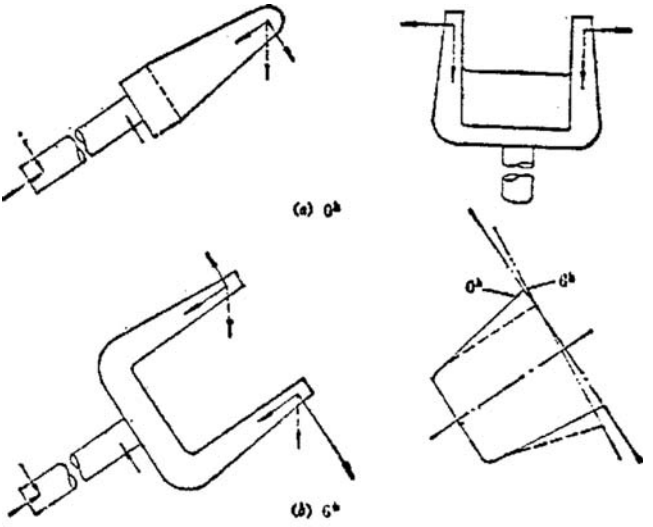


Fig. 3.2. The force conditions of a fork mounting at different hour angles.

deformations of the fork appear introduce a relative position change between the tube and polar axes, resulting in tracking and pointing errors. For building larger fork mounting telescopes, it is necessary to increase the length of the fork arm to accommodate longer tube length. If the length and cross-section area of a fork arm are L and $d \times d$, then the angular deflection of the arm end is proportional to L^3/d^2. If the length doubles, then the dimension of the cross-section should increase by a ratio of $2^{3/2}$ to keep the angular deflection constant. Therefore, the rate of the fork arm weight increase is $2 \times 2^{3/2} \times 2^{3/2} = 2^4$, leading to a very heavy structure for larger telescopes (Cheng and Xu, 1986).

An altitude-azimuth (or alt-azimuth) mounting is a special equatorial fork mounting where the polar axis is in the vertical direction. With this change, there is no bending of the yoke arm and the load on the yoke is pure compression and shear (from now on, we shall refer to the fork arm as the yoke arm). The telescope tube weight is directly transferred to the azimuth axis. The bending of tube structure happens only in the vertical plane. Neither of the two axes, which support the telescope weight, change direction with respective to the gravity. The mounting is sturdiest and simplest. The reduction of weight and cost is significant.

For telescope design, bending and torsion are the worst among deformations, while compression and shear introduce no pointing errors. The compression stiffness is also much larger than the bending stiffness, so that the compression and shear effects are generally small. For this reason, heavy and large radio telescopes first adopted the alt-azimuth mounting in the 1950s. The 6 m USSR Bolshoi Teleskop Azimultalnyi (Big Telescope Alt-azimuthal, BTA) built in 1976 is the first modern optical telescope which uses the alt-azimuth mounting. Now the alt-azimuth mounting has become the standard for all large optical telescopes.

The alt-azimuth mounting has a smaller sweeping volume than that of an equatorial mounting, so that alt-azimuth mount telescopes can be housed inside smaller and more compact astro-domes (a hemisphere or other shaped building for an astronomical telescope, which has an opening to allow star light to come in. The opening is computer controlled to follow the telescope's motion). The last advantage of the alt-azimuth mounting is that its design is independent of the geographical latitude, while the design of an equatorial mounting is always latitude dependent.

3.1.2.2 Coordinate Transformation and the Zenith Blind Spot

In astronomy, an equatorial coordinate system is the most commonly used system for indicating the positions of stars and other celestial objects. It is basically a projection of the earth's latitude and longitude coordinates onto the celestial sphere. By direct analogy, lines of latitude become lines of declination, measured in degrees, and lines of longitude become lines of right ascension, measured in degrees or in hours. The right ascension of a star is measured eastward from the vernal equinox, a point at which the sun crosses the celestial

equator in March. For equatorial mounting telescopes, a star's local hour angle is just the difference between the local sidereal time and the star's right ascension. With the polar axis rotates at a constant sidereal rate, the star image will be fixed inside the telescope field of view.

An alt-azimuth mounting telescope uses an altitude-azimuth, or a horizontal, coordinate system with the local horizon as the fundamental plane. This plane divides the sky into upper and lower hemispheres. The pole of the upper hemisphere is the zenith. The alt-azimuth coordinates are, respectively, elevation, or altitude, measured from horizon, and azimuth, measured eastward from north. If the star elevation position is measure from the zenith point, it is named as the zenith angle. The transformation from equatorial coordinates into alt-azimuth ones is (Figure 3.3):

$$\begin{aligned} \tan A &= \frac{\sin t}{-\sin\phi\cos t + \cos\phi\tan\delta} \\ \cos Z &= \sin\phi\sin\delta + \cos\phi\cos\delta\cos t \end{aligned} \tag{3.1}$$

where A is the azimuth angle, Z the zenith distance, φ the geographical latitude of the observatory site, δ the declination, and t the local hour angle of the celestial object. During the star tracking, both the azimuth angle and the zenith distance

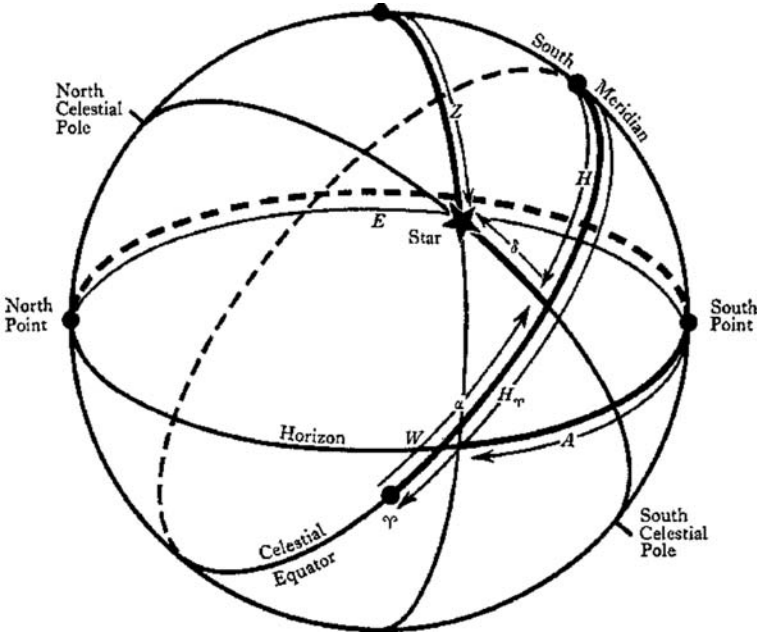


Fig. 3.3. The relationship between alt-azimuth and equatorial coordinate systems.

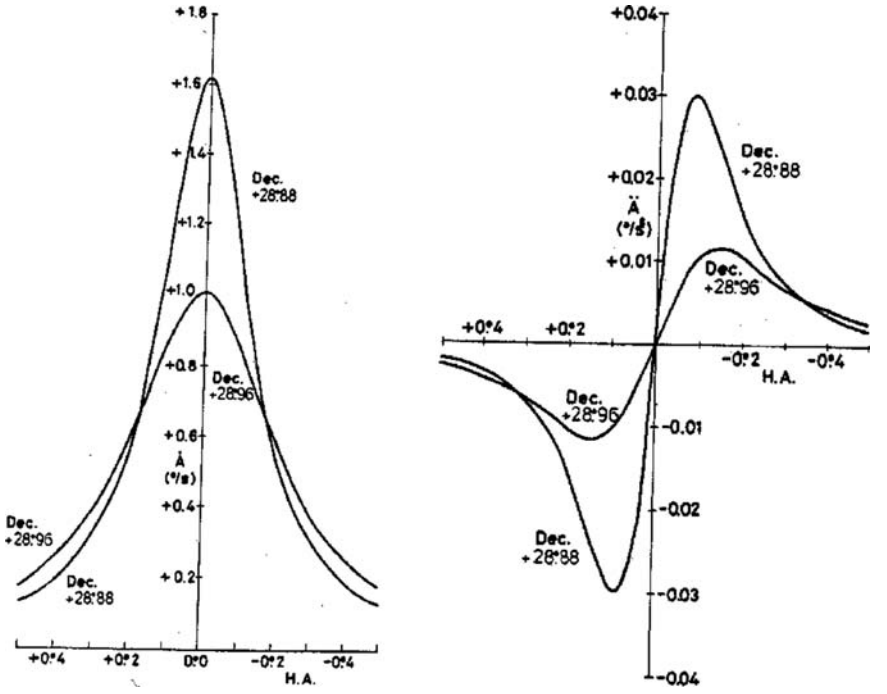


Fig. 3.4. The azimuth velocity (**a**) and the azimuth acceleration (**b**) for celestial objects with 0.1° and 0.2° zenith distances at a latitude of 28.75° (Watson, 1978).

vary as defined in Equation (3.1). The corresponding velocities of the two alti-azimuth coordinates during the star tracking are:

$$\begin{aligned} \frac{dZ}{d\tau} &= \cos\phi \sin A \\ \frac{dA}{d\tau} &= \frac{\sin\phi \sin Z + \cos Z \cos A \cos\phi}{\sin Z} \end{aligned} \tag{3.2}$$

Equation (3.2) indicates that the absolute elevation velocity is never faster than that of the local hour angle, while the azimuth velocity may reach arbitrarily high values as the celestial object close to the zenith point [Figure 3.4(a)]. For stars, the transitional azimuth velocity over the local meridian plane equals to $\mathrm{dA}/\mathrm{d}\tau = \cos\delta/\sin(\phi - \delta)$. The corresponding accelerations of the two alt-azimuth coordinates during star tracking are:

$$\begin{aligned} \frac{d^2Z}{d\tau^2} &= \cos\phi \cos A \left[\sin\phi + \frac{\cos\phi \cos A}{\tan Z}\right] \\ \frac{d^2A}{d\tau^2} &= -\frac{\cos\phi \sin A}{\sin^2 Z}[\sin Z \cos Z \sin\phi + \cos\phi \cos A(1 + \cos^2 Z)] \end{aligned} \tag{3.3}$$

From the formulas, the azimuth acceleration has a sign change at the transit and it reaches a high value when the star is near the zenith point [Figure 3.4(b)]. Alt-azimuth mounting telescopes with a computer control system can point and track celestial objects in most of the sky area. However, the tracking of a star is seriously limited by the maximum azimuth velocity and acceleration. A small, inaccessible blind spot exists around the zenith point as the azimuth angle will suddenly change sign when a celestial body crosses the meridian.

An alt-azimuth mounting telescope tracking a star will keep pace with it until a point, east of the meridian, where $\mathrm{d}A/\mathrm{d}\tau$ of Equation (3.2) reaches its maximum allowable azimuth velocity. Using the actual star azimuth velocity at the meridian, the tracking limitation happens for stars whose declination lies in the range (Borkowski, 1987):

$$\phi - \arctan\frac{\cos\phi}{|V| - \sin\phi} < \delta < \phi + \arctan\frac{\cos\phi}{|V| + \sin\phi} \tag{3.4}$$

where V is the maximum azimuth tracking velocity. This formula determines the north-south limits of the blind spot. The blind spot at the zenith is also determined by the maximum azimuth acceleration and the slewing velocity. To get an exact shape and size of the blind spot, it is necessary to draw all three groups of curves of the tracking velocity, tracking acceleration, and slewing velocity. The contours of maximum azimuth tracking velocity can be calculated from (Watson, 1978):

$$\tan Z = \frac{\cos\phi}{\dfrac{\dot{A}}{\omega} + \sin\phi}\cos A \tag{3.5}$$

where ω is the angular velocity of a star in the polar coordinate system. Figure 3.5(a) shows the contours for various maximum azimuth tracking

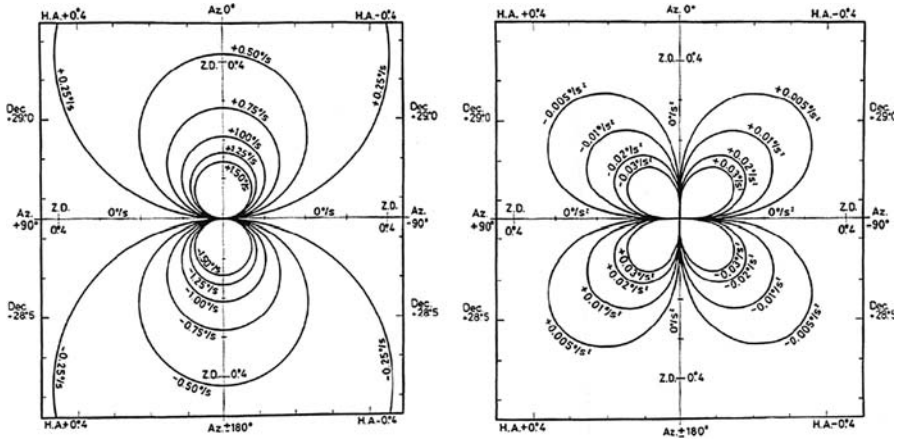


Fig. 3.5. Contours of various maximum azimuth tracking velocities (*left*) and tracking accelerations (*right*) at a latitude of 28.75° (b, *right*) (Watson, 1978).

velocities. These contours contain east-west and nearly south-north symmetry. When the maximum azimuth tracking velocity is a constant, the zenith angles of the maximum contour are different for different site latitude φ. For $\phi = 0$, $Z = \tan^{-1}(\cos A \cdot \omega / \dot{A})$.

The contour for the maximum azimuth tracking acceleration is given by:

$$\sin(2A)\cos^2\phi\cos^2 Z - \frac{1}{2}\sin A \sin(2\phi)\cos Z + \frac{1}{2}\sin(2A)\cos^2\phi + \frac{\ddot{A}}{\omega^2} = 0 \tag{3.6}$$

Figure 3.5(b) shows the contours for various maximum azimuth accelerations. However, to derive a precise shape of the blind spot, the slewing limitation effect (maximum azimuth velocity) has to be considered.

The blind spot due to the slewing limitation is formed in the following way: as a telescope is trying to follow a celestial object through the blind spot at the zenith, it accelerates from the tracking velocity to the slewing limit with a maximum acceleration and it slews over the blind spot. Then in the last stage, it decelerates to the tracking velocity and meets the celestial object on the other side of the meridian plane (Figure 3.6). Assuming the change of the hour angle inside the blind spot is H, the formula for the maximum slewing velocity is:

$$\frac{H}{\omega} = \frac{A}{\dot{A}_{\max}} + 2\frac{\dot{A}_{\max} - \dot{A}}{\ddot{A}_{\max}} - \frac{\dot{A}^2_{\max} - \dot{A}^2}{\ddot{A}_{\max}\dot{A}_{\max}} \tag{3.7}$$

Its approximate form is:

$$\cos Z = \cos A \tan\phi + \sin A \sec\phi \cot\left(\frac{A\omega}{\dot{A}_{\max}}\right) \tag{3.8}$$

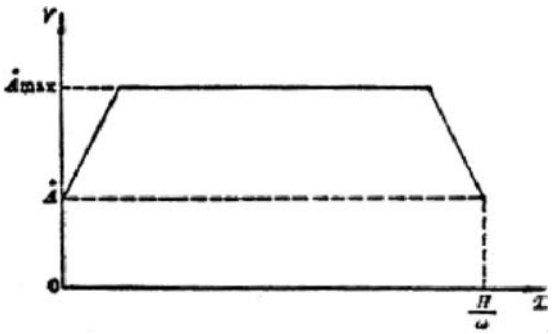


Fig. 3.6. Azimuth velocity curve while a telescope crosses the blind zenith spot.

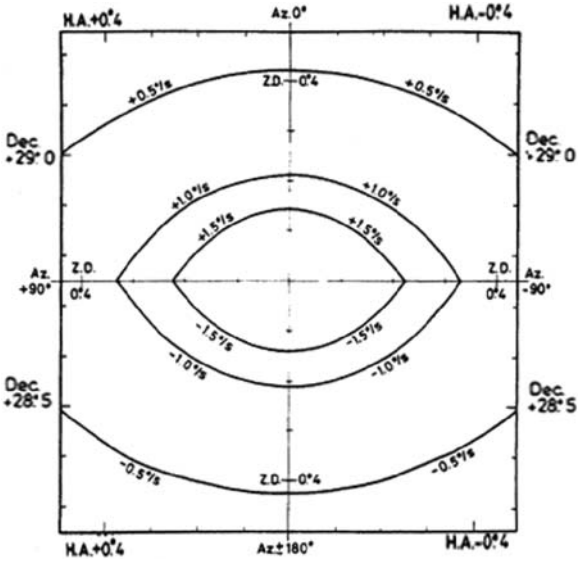


Fig. 3.7. Contours produced by various maximum azimuth slewing velocity for a latitude of 28.75° (Watson, 1978).

Figure 3.7 shows the contours produced by various maximum azimuth slewing velocities.

The size and shape of the real blind spot can be derived by adding all the relevant contours together. Of course, the elevation drive and field rotation also contribute to the blind spot size. In telescope design, the optimum tracking velocity and tracking acceleration can also be determined through the analyses of these contour maps. In general, the size of the blind spot is not very big. However, the velocity and acceleration required to observe very near to the zenith will be quite high.

3.1.2.3 Field Rotation and Its Compensation

When an alt-azimuth mounting telescope is tracking a celestial object, the orientation of its field will change with time. The field angle, or parallactic angle, and its rate of change are given by:

$$\tan p = \frac{\sin t}{\tan\phi\cos\delta - \sin\delta\cos t}$$
$$\frac{dp}{d\tau} = -\frac{\cos\phi\cos A}{\sin Z} \tag{3.9}$$

To achieve a stationary and high quality image of a sky area from an alt-azimuth mounting telescope, field rotation compensation is required. Using this equation, it is easy to follow the field rotation in Cassegrain or Nasmyth focal positions through a mechanical device. However, if a heavy and cumbersome detector is used, the device would be too difficult to build. In this case, special field de-rotation devices are used.

There are two different field de-rotation systems: refractive and reflective ones. The refractive one relies on internal reflection within a Dove prism [Figure 3.8(a)] and the reflective one uses a three-mirror system, named the K-mirror [Figure 3.8(b)].

In a field de-rotation device, if the rotation angle of the incident beam $\bar{A}$ is θ and the rotation angle of the de-rotation device is $\theta/2$, then the system transfer equation of a Dove prism will provide the rotation angle of the output beam $\bar{A}'$. The system transfer equation is formed from the Characteristic matrix. For the Dove prism, the characteristic matrix of the prism of single reflection is R^1, then we have:

$$\begin{aligned} &\bar{A}' = S^{-1} R^n S \cdot \bar{A} \qquad \bar{A} = \begin{bmatrix} 0 \\ \sin\theta \\ -\cos\theta \end{bmatrix} \\ &R^n = R^1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta/2 & \sin\theta/2 \\ 0 & -\sin\theta/2 & \cos\theta/2 \end{bmatrix} \end{aligned} \tag{3.10}$$

where S and S^{-1} are coordinate conversion matrixes of the prism. The result is:

$$\bar{A}' = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{3.11}$$

This solution means that the output beam remains stationary as the de-rotating device follows the half angle of the field rotation. For the K-mirror

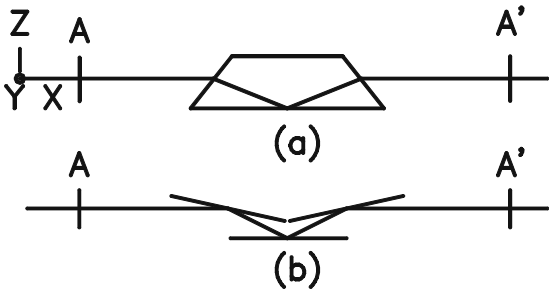


Fig. 3.8. (**a**) The refractive and (**b**) the reflective field de-rotation devices.

reflective field de-rotation device, the matrix R^3, describing triple reflections, is equal to R^1 for a single reflection. The solution is the same. One problem with these field de-rotation devices is that light polarization may be introduced. Therefore, in some cases, field de-rotation may be realized through software after the CCD observation, not by a special hardware.

In adaptive optics, when multi-laser guide stars are launched from an alt-azimuth telescope, the guide star positions in the sky may rotate around the telescope axis. To achieve stationary laser guider stars, a field de-rotation device is needed as discussed in Section 4.1.7.

3.1.3 Stewart Platform Mounting

The Stewart or hexapod platform was first reported by V. E. Gough in 1956 and fully discussed by D. Stewart in 1965. It is formed of six support struts, with variable lengths, and two platforms, one on top and the other fixed on the bottom. The major advantage of this platform is that all platform movements can be controlled by adjusting the lengths of six support struts. The platform movements include three linear translations and three angular rotations. When it is used in telescopes, the azimuth and elevation axes rotations are omitted. In addition, the Stewart platform is extremely stiff and stable as six struts form three stable triangles. The only unstable situation is when some support struts are in the same plane of the top platform.

The Stewart platforms were first used in flight simulators and now are used in many areas. The first optical telescope employing a Stewart platform mounting is the 1.5 m German Ruhr-University optical telescope in Chile built in 1998 (Figure 3.9 is a sketch). The telescope uses six long struts and the same top and bottom platform size so that the sky coverage is relatively larger. The primary mirror cell as its top platform and six struts are all of low expansion CFRP material. To ensure high pointing accuracy, the telescope employs gyroscopes as well as displacement sensors. The actuators inside struts are precision ball screws.

The Array for Microwave Background Anisotropy (AMiBA) is another telescope where a Stewart platform mounting is used. The Stewart platform mounting has been used in many secondary mirror systems. In this case, as the loading applied is in tension direction, backlash or play may be produced if no pretension is used. This backlash can be avoided by the use of tension springs between the mirror cell and the platform base. In the HST, a modified Stewart platform was used where three eccentric wheels, instead of linear screws, are used for strut length adjustment.

The operating principle of a Stewart platform is based on Grodzinski formula on the degrees of freedom (DOF) of a link mechanism:

$$F = 6(n-1) - \sum_{1}^{g} (6-f) \tag{3.12}$$

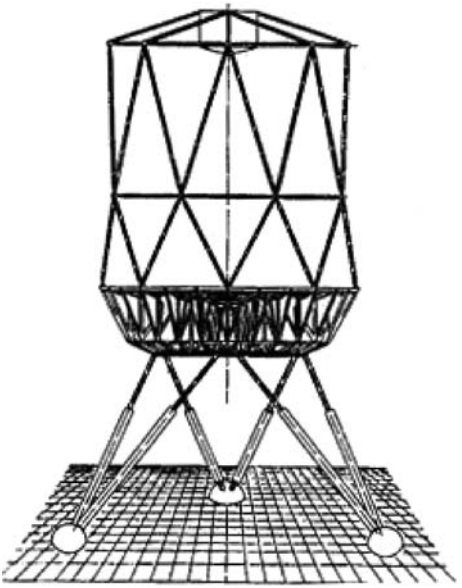


Fig. 3.9. Optical telescope using a hexapod platform mounting structure.

where F is the total DOF of the system, f the DOF of each joint, n the number of the components, and g the number of the joints. A Stewart platform consists of a total of 14 components: the top and bottom platforms and the top and bottom parts of six struts.

If the platform is not locked, each joint of the bottom struts attached to the bottom platform has two DOF in two rotational directions. Each joint inside a strut formed by two parts has only one displacement DOF, only its length changing. Each joint of the top struts attached to the top platform has three rotational DOF.

When the platform is locked, each joint of the struts attached to the bottom platform has only one DOF. It allows the rotation around only one axis. The joint between two parts of one strut has zero DOF and each joint of the struts attached to the top platform has three DOF.

By using Grodzinski's formula, total DOF of a Stewart platform, when it is not locked, is:

$$F = 6(14 - 1) - (6 \times 18 - 36) = 6 \tag{3.13}$$

And total DOF, when it is locked, is:

$$F = 6(14 - 1) - (6 \times 18 - 30) = 0 \tag{3.14}$$

These results confirm that a hexapod platform can produce motions in all six directions: three in translation and three in rotation. So, it can replace

translation and rotational devices and simplify telescope structure design. The platform can be used as the telescope, the secondary mirror, or the mirror mount.

The top platform motion of a hexapod can be expressed by the motions of three vertexes of the equiangular triangle on the platform. If X, Y, and Z are translations of three vertexes, these translations can be decomposed into two parts: x, y, z, caused by the platform translation and fx, fy, fz caused by the platform rotation:

$$\begin{aligned} X &= x + fx \\ Y &= y + fy \\ Z &= z + fz \end{aligned} \tag{3.15}$$

The translational parts of the three vertexes are synchronized with the translation of the platform. The relationship between the rotational part of $fx_i, fy_i, fz_i$and the top platform rotation can be expressed as:

$$\begin{aligned} fx_1 &= +S \sin\psi \cos\theta \\ fy_1 &= +S(1 - \cos\psi \cos\theta) \\ fz_1 &= -S \sin\theta \\ fx_2 &= +0.577S(1 - \cos\psi \cos\phi + \sin\phi \sin\theta \cos\psi) \\ fy_2 &= -0.577S(\sin\psi \cos\phi + \sin\phi \sin\theta \cos\psi) \\ fz_2 &= +0.577S \sin\phi \cos\theta \\ fx_3 &= -0.577S(1 - \cos\psi \cos\phi + \sin\phi \sin\psi \cos\theta \\ fy_3 &= +0.577S(\sin\psi \cos\phi + \sin\phi \sin\theta \cos\psi) \\ fz_3 &= -0.577S \sin\phi \cos\theta \end{aligned} \tag{3.16}$$

where the subscript i means the contribution from the vertex L_i, S is the distance between a vertex and its facing edge of the top platform, and the rotational angles, θ, ϕ, and ψ are shown in Figure 3.10. From these formulas, the amount of the strut length change of the mechanism can be derived.

If the motion of the top platform is known, the calculation of the rod length changes of all six struts is an inverse coordinate transformation. The above formulas are for this coordinator transformation. In comparison with the inverse coordinate transformation, the forward coordinate transformation is more difficult. In this case, the active velocity vector of active joints in joint space is given and the generalized movement of the top platform in task space is required. This forward formulation is provided by Shi and Fenton (1992).

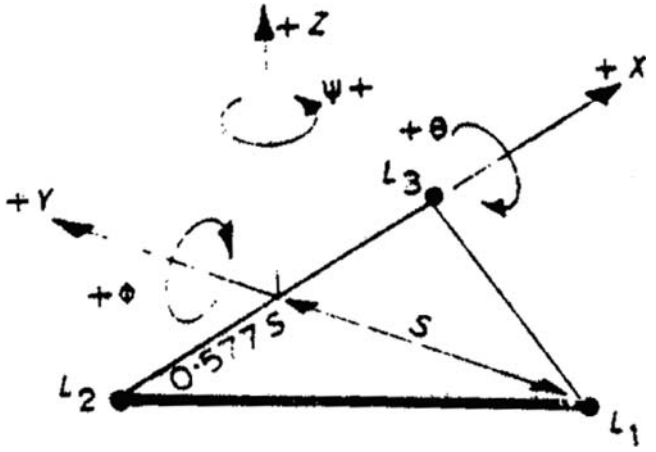


Fig. 3.10. The relationship between the motion of the platform and the motions of three vertexes.

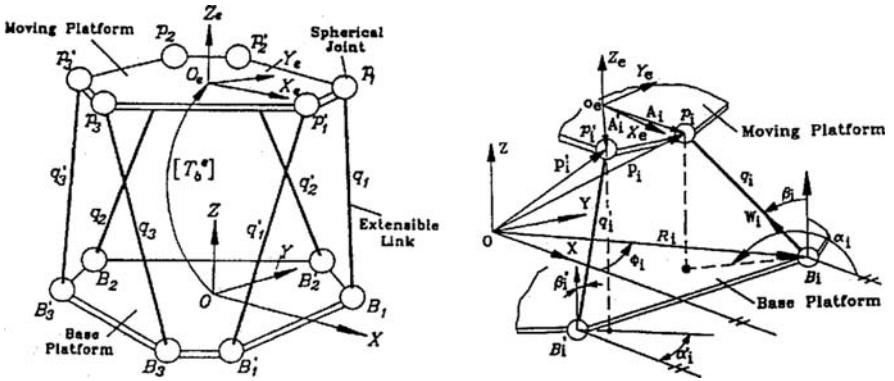


Fig. 3.11. The positions of vertexes in a hexapod platform (Shi and Fenton, 1992).

As shown in Figure 3.11, the transformation from a globe coordinate system, $O(XYZ)$, which is fixed to the bottom platform, to a local one, $O_e(X_eY_eZ_e)$, which is fixed to the top platform, can be expressed by a transformation matrix $\left[T_b^e\right]$:

$$\left[T_b^e\right] = \begin{bmatrix} \left[R_b^e\right] & P_b^e \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.17}$$

where, $\left[R_b^e\right]$ and P_b^e are rotational and translational matrixes. As shown in Figure 3.11, the positions of three nodes p_i can be expressed as:

$$p_i = B_i + w_i q_i \tag{3.18}$$

where, $B_i = \overline{OB}_i$ is a constant vector fixed to the bottom platform, q_i is the length of each rod, and w_i the directional vector of each rod.

$$B_i = \begin{bmatrix} R_i \cos\phi_i \\ R_i \sin\phi_i \\ 0 \end{bmatrix} \tag{3.19}$$

The directional vector of each rod is:

$$w_i = \begin{bmatrix} \sin\beta_i \cos\alpha_i \\ \sin\beta_i \sin\alpha_i \\ \cos\beta_i \end{bmatrix} \tag{3.20}$$

Inserting the above relationship into Equation (3.18) gives:

$$p_i = \begin{bmatrix} R_i \cos\phi_i + q_i \sin\beta_i \cos\alpha_i \\ R_i \cos\phi_i + q_i \sin\beta_i \sin\alpha_i \\ q_i \cos\beta_i \end{bmatrix} \tag{3.21}$$

Differentiating the above formula produces the velocity of these three vertexes:

$$v_{p_i} = \frac{dp_i}{dt} = \begin{bmatrix} \dot{q}_i \sin\beta_i \cos\alpha_i + q_i\dot{\beta}\cos\beta_i \cos\alpha_i - q_i\dot{\alpha}\sin\beta_i \sin\alpha_i \\ \dot{q}_i \sin\beta_i \sin\alpha_i + q_i\dot{\beta}\cos\beta_i \sin\alpha_i - q_i\dot{\alpha}\sin\beta_i \cos\alpha_i \\ \dot{q}_i \cos\beta_i - q_i\dot{\beta}_i \sin\beta \end{bmatrix} \tag{3.22}$$

The linear velocity of the top platform is related to the velocity of these three vertexes. Assuming that the origin of the local coordinate system is located at the center of gravity of the triangle with vertexes of p_i, then:

$$v_0 = \frac{dp_0}{dt} = \frac{1}{3}\frac{d}{dt}\sum_{i=1}^{3} p_i = \frac{1}{3}\sum_{i=1}^{3} v_{p_i} \tag{3.23}$$

On the other hand, the three vertex velocities can also be expressed as a sum of the linear velocity, v_0, and the angular velocity, ω, of the top frame:

$$v_{p_i} = \left[R_b^e\right][\omega]A_i + v_0 \tag{3.24}$$

where $A_i = [A_1 A_2 A_3]$ is the positional vectors of the three vertexes in the top platform coordinate system, $A_i = O_e p_i$ and the angular velocity ω is:

$$[\omega] = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{3.25}$$

If the term $[v_{p_1}, v_{p_2}, v_{p_3}] - [v_0, v_0, v_0]$ is expressed as a matrix $[C]$, then the Equation (3.24) may be transformed as $\left[R_b^e\right]^T [C] = [\omega][A]$. The top platform angular velocity can be expressed by the following three formulas:

$$\begin{aligned} a_x C_{12} + a_y C_{22} + a_z C_{33} &= \omega_y A_{2x} - \omega_x A_{2y} \\ a_x C_{13} + a_y C_{23} + a_z C_{33} &= \omega_y A_{3x} - \omega_x A_{3y} \\ n_x C_{11} + n_y C_{21} + n_z C_{31} &= \omega_z A_{1y} \end{aligned} \tag{3.26}$$

In the above equations, there are six unknowns as the rates of the angular change $\dot{\alpha}_i$ and $\dot{\beta}_i$. Therefore, another six equations are necessary to derive the required solutions. Three of these six equations can be found from the velocity constraint equations of a rigid body (Figure 3.12), which are:

$$\begin{aligned} v_{p1} \cdot (p_1 - p_2) &= v_{p2} \cdot (p_1 - p_2) \\ v_{p2} \cdot (p_2 - p_3) &= v_{p3} \cdot (p_2 - p_3) \\ v_{p3} \cdot (p_3 - p_1) &= v_{p1} \cdot (p_3 - p_1) \end{aligned} \tag{3.27}$$

The remaining three equations can be found from the velocity of the other three vertexes:

$$\dot{q}_i' = v_{p_i'} \cdot w_i' \tag{3.28}$$

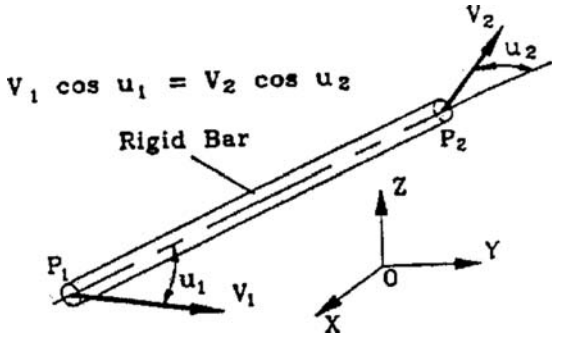


Fig. 3.12. The velocity constraint equations of a rigid body (Shi and Fenton, 1992).

where $w_i' = [\sin\beta_i' \cos\alpha_i', \sin\beta_i' \sin\alpha_i', \cos\beta_i']$ is a unit vector representing the orientation of the extensive link bar, and

$$v_{p_i'} = [R_b^e][\omega]A_i' + v_0 \tag{3.29}$$

is the velocity of point p_i' expressed in terms of the three point velocities, v_{pi} and $A_i' = O_e p_i'$ is a constant vector expressed with respect to the top frame. Equations (3.27) and (3.28) represent six equations with six unknowns. They can be solved for $\dot{\alpha}_i$ and $\dot{\beta}_i$. The linear and angular velocities of the top platform, v_0 and ω, can be found from Equations (3.23) and (3.26). The forward transformation problem can be solved.

For a hexapod platform as shown in Figure 3.13, where there are only three vertexes on the top platform, the equations derived are simpler. They are:

$$p_i = \begin{bmatrix} R\cos\Phi_i + q_i F_{1i}\cos\gamma_i - q_i F_{2i}\sin\gamma_i\cos\theta_i \\ R\sin\Phi_i + q_i F_{1i}\sin\gamma_i - q_i F_{1i}\cos\gamma_i\cos\theta_i \\ q_i F_{2i}\sin\theta_i \end{bmatrix} \tag{3.30}$$

$$\begin{aligned} F_{1i} &= \cos\phi_i = \frac{q_i^2 + b_i^2 - q_i'^2}{2b_i q_i} \\ F_{2i} &= \sin\phi_i = \left[1 - \left(\frac{q_i^2 + b_i^2 - q_i'^2}{2b_i q_i}\right)^2\right]^{1/2} \end{aligned} \tag{3.31}$$

where θ_i are three intermediate variables. Differentiating the above equations:

$$v_{p_i} = \begin{bmatrix} -\cos\gamma_i(\dot{q}_i F_{1i} + q_i \dot{F}_{1i}) - \sin\gamma_i(q_i F_{2i}\cos\theta_i + q_i \dot{F}_{2i}\cos\theta_i - q_i F_{2i}\cos\theta_i\dot{\theta}) \\ -\sin\gamma_i(\dot{q}_i F_{1i} + q_i \dot{F}_{1i}) + \cos\gamma_i(q_i F_{2i}\cos\theta_i + q_i \dot{F}_{2i}\cos\theta_i - q_i F_{2i}\cos\theta_i\dot{\theta}) \\ \dot{q}_i F_{2i}\sin\theta_i + q_i F_{2i}\sin\theta_i - q_i F_{2i}\cos\theta_i\dot{\theta} \end{bmatrix} \tag{3.32}$$

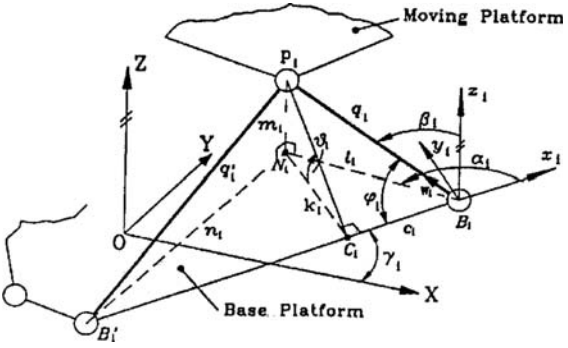


Fig. 3.13. Special hexapod platform with only three vertexes on the top frame (Shi and Fenton, 1992).

$$\dot{F}_{1i} = \frac{2q_i(q_i\dot{q}_i - \dot{q}'_i\dot{q}''_i) - \dot{q}_i(q_i^2 + b_i^2 - q'^2_i)}{2b_iq_i^2}$$
$$\dot{F}_{2i} = -\frac{2q_i(q_i\dot{q}_i - \dot{q}'_i\dot{q}''_i) - \dot{q}_i(q_i^2 + b_i^2 - q'^2_i)}{2q_i[4b_i^2q_i^2 - (q_i^2 + b_i^2 - q'^2_i)^2]^{1/2}} \tag{3.33}$$

The above equations and the velocity constraint equations of rigid body determine all the motion of the top platform. Substituting these equations into Equation (3.26), three equations with three unknown variables can be solved. The solutions, then, are substituted into Equation (3.32) for linear velocities of three top vertexes. The forward problem can be thus solved.

The hexapod platform is also used for supporting very accurate optical benches. Some fine tuning benches use double layers of hexapod platforms: one for coarse tuning and another for fine tuning control. In space telescopes, the platform is also used for vibration control purposes. Parks and Honeycutt (1998) had applied this hexapod platform to optical mirror support which is discussed in Section 2.3.4.

3.1.4 Fixed Mirror or Fixed Altitude Mountings

For cost reduction purposes, some telescopes use fixed mirror or fixed altitude mountings. An early photographic zenith tube, which recorded the time of a star transiting over the meridian, was a telescope which used a fixed mirror mounting at zenith direction. The liquid mercury mirror telescope is also a fixed-mirror mounting design. In radio astronomy, the Arecibo 300 m radio antenna is the largest telescope which employs a fixed mirror mounting. The star tracking of this antenna is through the movement of its receiver on top of the telescope. The fixed mirror design is also used in some solar telescopes (solar towers), where the tracking of the sun is through a pair of heliostat mirrors. A fixed mirror mounting is in fact a special case of the fixed altitude mounting.

The fixed altitude mounting design was developed from the early fixed position sextant. The Hobby–Eberly Telescope (HET) is the first modern telescope using this type of mounting. In the HET design, the primary mirror is fixed at a particular altitude and it could only rotate in steps in azimuth direction. In both fixed mirror and fixed altitude mounting designs, the primary mirror does not change direction with respect to gravity, therefore, the mirror support and control system are simplified.

Three large optical telescopes have been built by using fixed mirror or fixed altitude mounting designs. These are the 4 m LAMOST (reflecting Schmidt telescope) and the 10 m HET and SALT (segmented mirror telescope). The LAMOST has a fixed segmented 6.67 m primary spherical mirror and a 4.4 m deformable reflector corrector which is also a segmented mirror and acts as a siderostat mirror. The HET and SALT are of fixed altitude design with the tube axis 35° away from the zenith. The star tracking of these telescopes is through the motion of the detectors inside the 12° field of view. For extending the sky

coverage, both telescopes can rotate in steps about the azimuth axis through a number of hydrostatic bearings. Even so, the sky coverage is very limited. However, the cost of a fixed altitude telescope is only about 20% of that of its fully steerable conventional counterpart.

3.2 Telescope Tube and Other Structure Design

3.2.1 Specifications for Telescope Tube Design

A tube structure of an optical telescope connects its primary and secondary mirrors. A general requirement for a tube structure is to ensure that the relative position is maintained between these two optical mirrors. Tubes for small telescopes are truly cylindrical in shape. As the telescope size increases, the deflections of two ends of such a cylinder structure at low elevation angle involve both lateral and tilt components. The primary and secondary mirrors, which are fixed at both ends of the tube, will deflect in opposite directions, resulting in a pointing error and coma and affecting the telescope performance.

The pointing error due to relative positional changes between the primary and secondary mirrors is given by the formula:

$$\delta = u_1/f - (m-1)(u_2 + r_2\theta)/f \tag{3.34}$$

where u_1 and u_2 are the radial displacements of the primary and secondary mirrors with respect to the telescope focus, θ the relative angle change between the primary and secondary mirrors, f the system focal length, m the magnification factor of the secondary mirror, and r_2 the radius of curvature at the vertex of the secondary mirror. The gravity induced pointing error is repeatable and its calibration and correction are possible in the control system.

However, the position changes between two mirrors also produce coma as:

$$l_c = \frac{3(m-1)^2}{32F^2 f}\left[(e_2^2(m-1) + m + 1)u + (m^2+1)r_2\theta\right] \tag{3.35}$$

where u is the relative radial position change between the primary and secondary mirrors, F the focal ratio, and e_2^2 the eccentricity of the secondary mirror. This coma does not include coma of the telescope optical system which is a function of field angle. The coma caused by the relative position changes is evenly distributed all over field of view.

The above equation has two terms. An increase of one term may be compensated by a reduction of another term. If the sum of these terms is zero, then the tube structure is coma free. One special tube design is named the coma-free tube design, where the displacement induced coma is compensated by the tilt induced one. This design principle is also used in the optimization of the rotational center of a chopping secondary mirror.

The axial position change between two mirrors δl introduces an axial displacement of the focus position δf:

$$\delta f = (m^2 + 1)\delta l \tag{3.36}$$

The positional tolerance between two mirrors can also be derived from the wavefront error as discussed in Section 7.1.5.

3.2.2 Telescope Tube Design

Most small and medium size optical telescope tubes use simple "A"-shape trusses first proposed by Mark U. Serrurier in 1935 for the 5 m Hale telescope. The "A"-shape truss is called the "Serrurier truss" [Figure 3.14(a)]. Using this truss design, four very long "A" trusses are above the central block and four very short ones are below the block. These trusses intersect each other on one plane within the central block to avoid any unwanted moments applied on the tube central block. The other ends of these trusses support the primary mirror cell and secondary mirror top ring.

With this design, when the tube is in vertical direction, the primary and secondary mirrors shift along the axis. There is no pointing error introduced. When the tube is in horizon direction, the top and bottom truss members do not

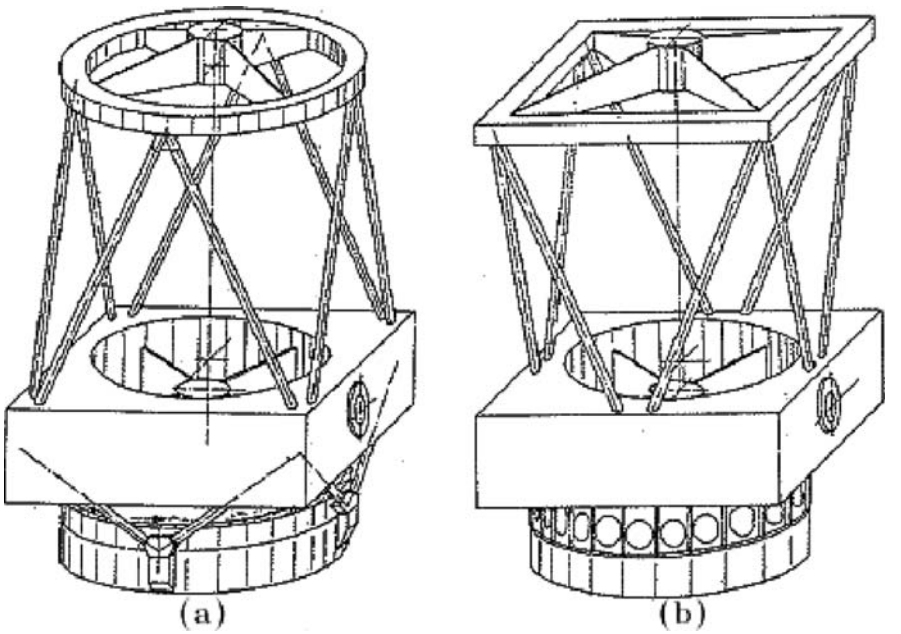


Fig. 3.14. (**a**) Conventional and (**b**) inverted "A" truss tube designs.

support the mirror weight and their lengths remain unchanged. The side trusses deform downwards, so that the primary and secondary mirrors, which are defined by four "A" trusses, deform laterally in gravitational direction. No mirror tilt is produced. The lateral displacements of both mirrors can be adjusted to be the same by altering the cross-section area of their members. In this way, neither pointing error nor coma is produced by the tube.

If the weight of a mirror is W, the cross-sectional area of a truss member is A, the distance between the center block and the mirror is L, and the width of the central block is D, then the lateral displacement of the mirror when the tube is horizon pointing is:

$$\delta = \frac{\left[(D/2)^2 + L^2\right] W/4}{AED/2} = \frac{\left[(D/2)^2 + L^2\right] W}{2AED} \tag{3.37}$$

where E is the Young modulus of the truss material.

The deformation of such a truss is roughly proportional to the square of its length, so that the beam cross section for the secondary mirror support is generally large while that for the primary is very small. A large cross sectional area results in weight and thermal time constant increase.

With modern computer control, the repeatable pointing error and coma can be compensated by actively adjusting the secondary mirror position. The equal deformation criterion becomes less demanding. Some telescopes use only the top trusses and its bottom trusses are replaced by short parallel cantilever beams for larger primary mirror displacement. To improve the mirror seeing, the primary mirror surface can even be lifted above the central block to allow the natural wind ventilation over the surface. The top "A" trusses can even be inverted so that a square top vane support frame is formed. This top frame has high rigidity as discussed in the next section [Figure 3.14(b)]. For small optical telescopes, especially, for space telescopes, a central tower for the secondary mirror from the primary mirror inner hole is used and the tube trusses are eliminated. For large telescopes, more "A" trusses on the top part of the tube are also used to reduce the lateral displacement of the secondary mirror.

To improve thermal and mechanical performances of the tube, a two-layer or multi-layer (multibay) truss structure is used for very large optical telescopes with one truss atop of the other (Figure 3.15), so that the truss cross sectional area and weight are reduced and thermal performance is improved.

The natural frequency of a cantilever beam with distributed and concentric mass at the free end is estimated by the formula (Schneermann, 1986):

$$f = \frac{1}{L^2}\sqrt{\frac{EA}{\rho I}\frac{1}{(1 + 0.23m_c/m)}} \tag{3.38}$$

where E is the Young modulus, I the moment of inertia of the beam, A the cross section, L the length, ρ the density of the material, m the distributed mass, and

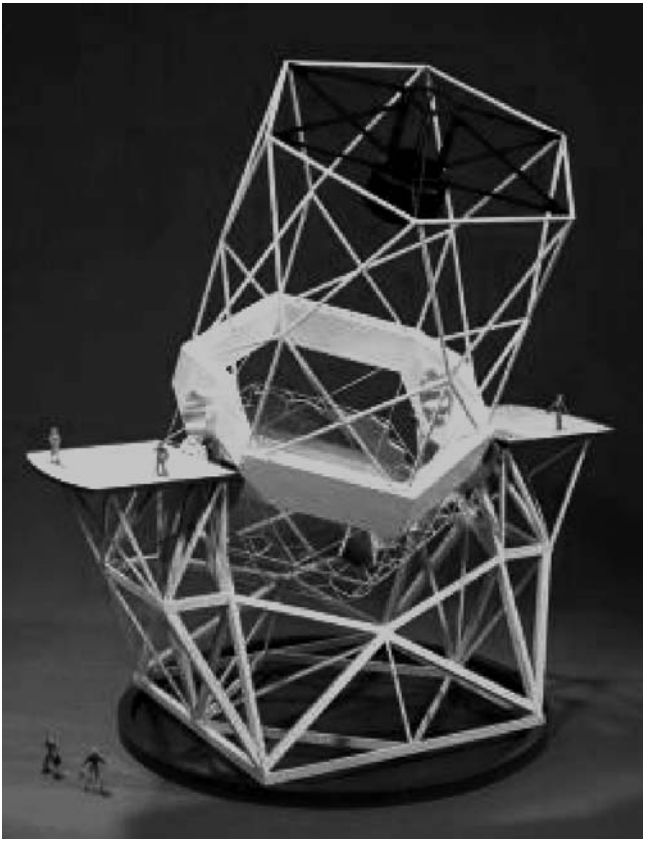


Fig. 3.15. The Keck 10 m telescope structure (Keck).

m_c the concentric mass at the beam end. With this formula, the resonance frequency of the tube can be estimated.

A related topic in tube design is the location of the elevation bearings. In many cases, the bearings are outside the central block and then the central block bending is inevitable. If the bearings are inside the central block, there will be no bending of the central block at the bearing location.

The primary mirror cell is on the lower side of a tube. The mirror cell normally has a circular bottom plate and a strong cylindrical side ring. The mirror cell supports the primary mirror in both axial and radial directions. In most telescopes, the mirror cell also provides support for the Cassegrain instruments.

For large segmented mirror telescopes, the primary mirror cell is a truss structure for achieving maximum stiffness-to-weight ratio. The mirror segments are supported on top of the truss with three nodes for each mirror segment. A large steel truss structure has temperature gradient induced error either in open air or inside a dome. These errors produce segment deformations of the primary

mirror. The thermal error of an open air truss is discussed in Section 8.1.2 The error for trusses inside the dome is usually less.

For space-based or extremely large telescopes, the tube and mirror cell may have a shape similar to a backup structure and feedleg of a radio telescope. In this design, the truss structure supporting the secondary mirror becomes "feedlegs." The support beams extend directly from the primary mirror cell to the secondary mirror position. This "feedleg" is referred to as a metering structure. Figure 3.16 is the primary mirror cell and the metering structure of the proposed 100 m Over-Whelmingly Large (OWL) telescope.

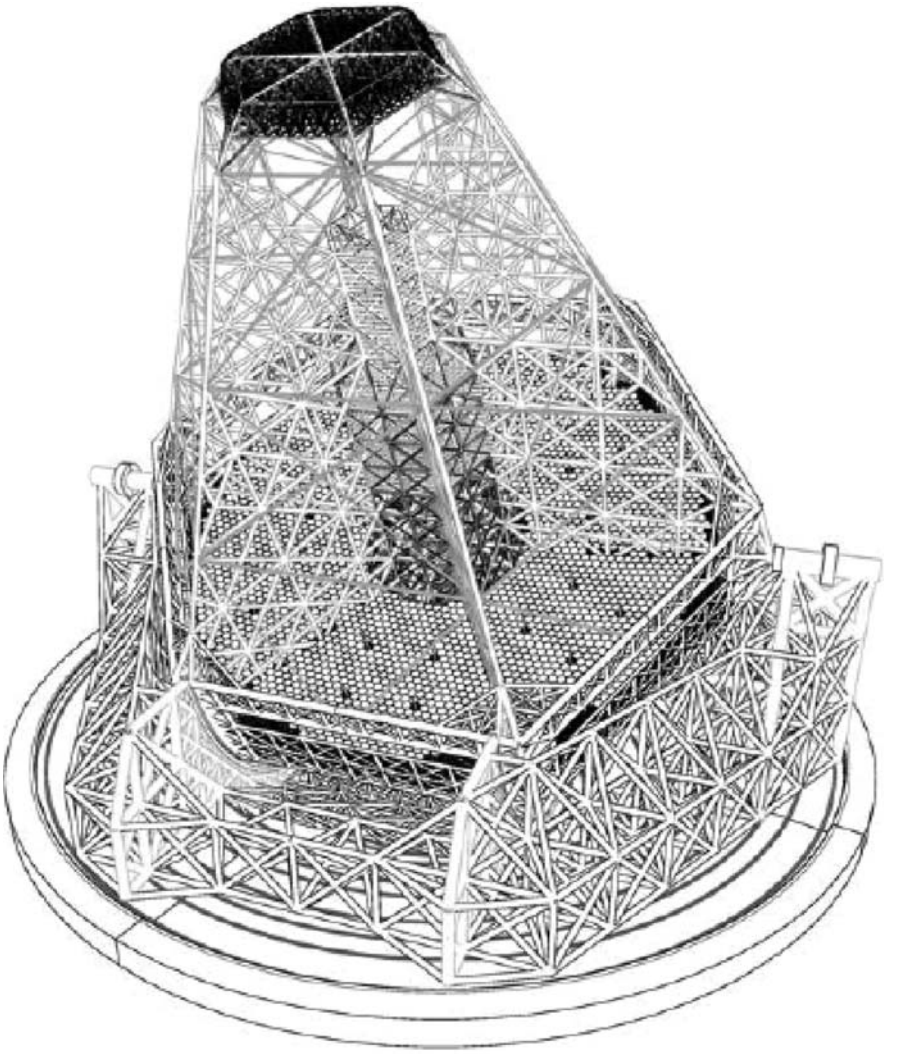


Fig. 3.16. Proposed metering structure of the OWL telescope (Brunetto et al., 2004).

3.2.3 Support Vane Design for Secondary Mirror

To support the secondary mirror on top of a tube, a cross vane structure as shown in Figure 3.14(a) is used. The vane support structure has three advantages: (a) stiff and stable; (b) smaller aperture blockage; and (c) easy to manufacture and assemble.

The classic vane structure has a symmetric cross shape without offset, where the vanes are arranged in two perpendicular directions. The vanes are wider in the axial plane (height) and narrower in the aperture plane (width). The radial stiffness is high, but the lateral stiffness is small. With four vane beams, the natural frequency is (Cheng, 1988):

$$f = \frac{1}{\pi}\sqrt{[(4EI/L) + (12EIr^2/L^3)]/J} \tag{3.39}$$

where E is the Young modulus, I the moment of inertia of the vane beams, L the length of the beams, J the moment of inertia of the secondary unit, and r the radius of the secondary unit. The vane stiffness reduces as the aperture size increases. To maintain the same natural frequency, the vane width has to increase as the 5/3 power of the aperture. This brings more aperture blockage. An alternative way is to apply pre-stress to the vane structure. The resonance frequency of a pre-stressed vane structure is (Bely, 2003):

$$\begin{aligned} f &= f_0\sqrt{(1+P)/P_{Euler}} \\ P_{Euler} &= \pi^2 EI/L^2 \end{aligned} \tag{3.40}$$

where f_0 is the frequency without preload, P the preload, and P_{Euler} the Euler critical load of the vane beam, I the moment of inertia of the beam, and L the length of the vane. However, the increase of resonance frequency is limited.

An offset vane structure as shown in Figure 3.17(a) can be used for solving the problem. In this design, two opposite vanes are no longer in a straight line

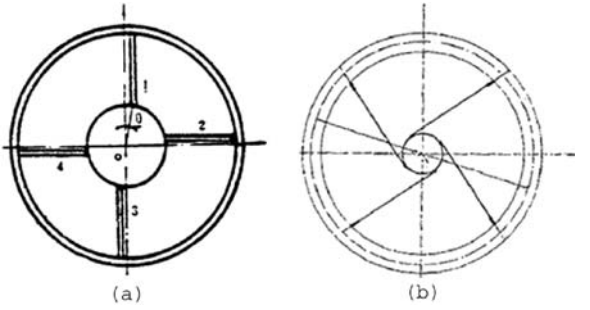


Fig. 3.17. The supporting vane structure of the secondary mirror.

and the moments produced by two pairs of vanes cancel each other exactly. The torsional stiffness improves. Calculation shows that a small offset angle of 1.13° can double the structural resonance frequency. In an extreme case of an offset angle of 45°, a very stiff vane structure is formed [Figure 3.17(b)].

In the vane design, each vane beam is made of one or two thin members. If two members are used, the two vane beams and the central unit form a stable triangle. The tilt of the secondary mirror under gravity can be adjusted through the member cross-section optimization. When the center of gravity of the central unit is on the support ring plane, the cross-sectional areas of the top and bottom members should be the same if they are symmetrical about the ring plane.

However, if the center of gravity of the secondary assembly is not on the symmetric plane, the cross-sectional areas of the top and the bottom members have to be optimized to guarantee that the axis of the secondary assembly does not tilt when the tube rotates. The cross-sectional area of the member near the center of gravity should be larger than that far away from the center of gravity.

In addition to the vane design improvement, a square vane frame can be used as in Figure 3.14(b). In this new design, four stiff and stable triangles are formed on top of the tube. This increases the tube stability and lowers the weight of the ring.

Some very large telescopes use octagonal-shape vane support rings with more than four supporting vanes. These are variations of a circular vane support structure. For extremely large telescopes, high stiffness CFRP composite and structure damping may be necessary in the secondary mirror vane support structure.

3.2.4 Telescope Bearing Design

Ball bearings are used at the rotation axes of small and medium optical telescopes. The friction coefficient of ball bearings ranges between 0.001 and 0.003. Even for large telescopes, if the light beam does not pass through the elevation axis, the elevation bearing diameter is small. The friction moment is small so that ball bearings are suitable for the elevation axis of most optical telescopes. As discussed in the previous section, the elevation bearings are usually located in both sides of the tube central block. The center line of the two elevation bearings divides equally a square formed by four support points of the upper and lower tube trusses. Self-aligning roller or ball bearings are often used for allowing misalignment of the bearing sockets and the bending of elevation axes.

The elevation bearing has a direct influence on pointing and tracking accuracy. The bearing shafts used could be light weight tubes. On one end of the shaft the fork arm is connected through a screwed flange and on the other end the elevation bearing is attached. During bearing assembly, the shaft's deformation under gravity usually makes the alignment difficult when a long shaft is used. To overcome this, a balance weight can be used to correct the end tilt of the shaft on the bearing side.

The radial runout of a bearing produces pointing instability. The runout is the displacement of one bearing surface relative to a fixed point when one raceway rotates with respect to another raceway. Generally, this radial run-out is caused by small irregular gap or irregular contact between bearing rings and rollers (or balls). When the bearing is under loading, the small gap or runout will increase. To reduce this gap or runout, preloading the bearing is necessary. The preloading eliminates the gap between bearing ring and rollers. The preload force is applied by squeezing the inner bearing ring through a tapered insert between the bearing and shaft. The required preload force can be calculated from the bearing's fact sheets. When applying the preload, lubricating is necessary for reducing friction in the metal-to-metal contacting area. Without preloading, the very top roller can be hand turned and, after preloading, the turning becomes impossible.

A three-row roller bearing with two rows in the horizon plane and one in the vertical plane is suitable for the azimuth axis (Figure 3.18). This bearing has high resistance against turning moments. However, a large roller bearing is expensive and its friction moment is also significant. A new azimuth bearing design involves a small diameter thrust bearing in the bottom and a number of radial rollers holding a fine machined cylindrical surface on the top. It is like a reversed conic cone. The thrust bearing takes all the axial loading and the radial rollers ensure the alignment of the bearing axis. Some of these radial rollers also act as drive pinions to rotate the big cylinder surface through friction force. Radial preload of these rollers is necessary for the stability of this bearing system. The roller preload forces applied on the machined cylinder surface should be constant through an elastic deformed large truss frame (Figure 3.19). This type of azimuth bearing design is low in cost and has been adopted by the Wisconsin, Indiana, Yale, and NOAO (WIYN) 4 m telescope, the Astrophysical Research Consortium (ARC) 3.5 m telescope, and a number of millimeter wavelength telescopes.

For extremely large optical telescopes, hydrostatic bearings are preferred as their azimuth bearings. Hydrostatic bearings have enormous loading capability and very low friction forces. Hydrostatic bearings are made of bearing pads, which have smooth edges and few recesses, filled with a high pressure viscous fluid, which supports the load and lubricates the main bearing surfaces. The main bearing surfaces are on the moving part of a structure and are flat or cylindrical in shape. The main advantages of hydrostatic bearings are: (a) very small friction coefficient; (b) extremely high stiffness; (c) great load bearing

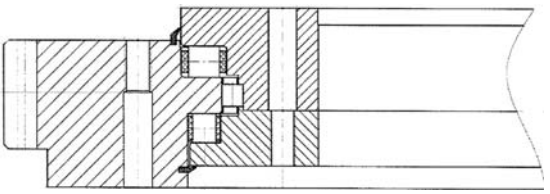


Fig. 3.18. A typical three-row roller azimuth bearing with gear rim.

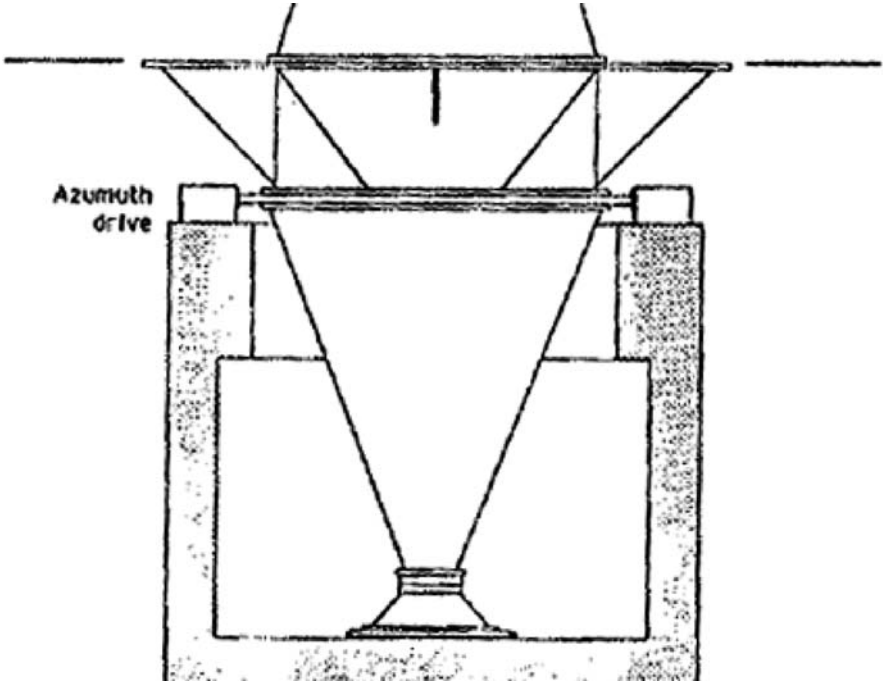


Fig. 3.19. The azimuth bearing system using a thrust bearing, a cylinder surface, and radial rollers on an elastic frame.

capability; (d) easy in bearing surface manufacture; and (e) permits a larger dimension tolerance than ball or roller bearings. However, hydrostatic bearings also have disadvantages. First, it requires high rigidity in the bearing surface. Because there are just three or four bearing pads which support the bearing surface, the support forces are more concentrated than in ball or roller bearings where the load is more uniformly distributed. Second, a hydrostatic bearing requires a complex compressor unit to provide fluid with sufficient pressure. Third, hydrostatic bearings can generate heat, which will lead to temperature increase of the telescope structure. To avoid this effect, pre-cooling of the fluid is necessary. The heat generated in a unit time within a volume of fluid, dQ/dt, can be expressed as:

$$\frac{dQ}{dt} = \frac{dV}{dt} \cdot \Delta P \tag{3.41}$$

where dV/dt is the fluid volume which passes through the bearing surface in a unit time and ΔP the pressure drop in the hydrostatic bearing system. If the oil volume through the bearing system is $0.819{*}10^{-5}$ m^3/s and the bearing pressure drop is 24.5 kg/m^2, then the heat generated is about 20 W. And fourth, the fluid

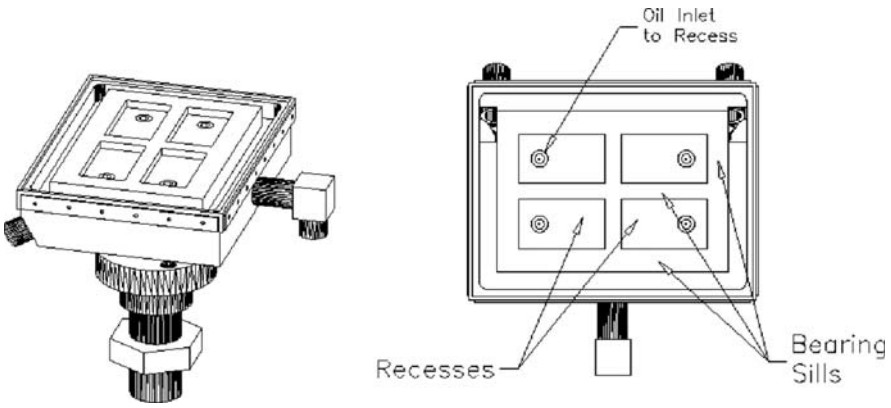


Fig. 3.20. The design of hydrostatic bearing oil pads (Eaton, 2000).

viscosity used in hydrostatic bearings is a function of temperature. If heat is generated inside hydrostatic bearings, the flux volume as well as the pressure will change in the system.

Hydrostatic bearing oil pads are shown in Figure 3.20. The oil pads should be self-aligned with the main bearing surface so that a smooth bearing surface contact is maintained for a long service life. The self-aligned property can be achieved by using a spherical pad back or by using double-layer oil pads. Generally, the number of oil pads is determined by the number of constrained degrees of freedom. In order to increase the static pressure, the oil pad number can be more than the degrees of freedom. The oil pad is made of bronze material to avoid main bearing surface damage should the hydraulic system fail.

In hydrostatic bearing design, the concept of viscosity is important. Viscosity is known as resistance to flow. Viscosity of a liquid is equivalent to the shear modulus of a solid. Shearing deformation of a solid, δ, can be expressed as $Fh/(AG)$,where F is the shearing force, h the height, A the cross-sectional area of shear, and G the shear modulus. In a hydrodynamic system, the relationship between the shear force and the relative velocity can be derived by considering two plates closely spaced apart at a distance h, and separated by a fluid or gas. Assuming that the plates are very large, with a large area A, such that the edge effects may be ignored, and that the lower plate is fixed, let a force F be applied to the upper plate, then the relative velocity between the two plate surfaces, U, can be expressed as $Fh/(A\mu)$, where μ is the absolute viscosity of a liquid or gas, in a unit as $\mathrm{Ns/m^2}$, or Pascal-second. The Pascal-second is also named Poise, which equals 100 centipoises, or *cp*. The ratio of the absolute viscosity to density is the dynamic viscosity or the kinematic viscosity. The product of the dynamic viscosity and the velocity gradient is the shear stress between layers. The unit of the absolute viscosity is $\mathrm{cm^2/s}$ or stoke ($1\ \text{stoke} = 1\ \mathrm{cm^2/s}$).

The viscosity of oil is between 100 and 1,000 cp and that of air is about $170 \cdot 10^{-4}$ cp, one ten thousandth of that of oil. The rate of viscosity change

against temperature is called the viscosity index. The higher the viscosity index is, the smaller the variation of the viscosity against the temperature. According to the definition of the viscosity, the friction force of a hydrostatic bearing F can be calculated from (Eaton, 2000):

$$F = \frac{UA\mu}{h} \tag{3.42}$$

where U is the linear fluid speed of the bearing surface layer, A the effective area of the hydrostatic bearing, and h the oil film thickness. The oil film thickness is not affected by the speed of rotation. It is given by (Bely, 2003):

$$h = \sqrt[3]{12\frac{Q\mu \cdot l}{b\Delta p}} \tag{3.43}$$

where Q is the oil flow, Δp the drop in pressure over the gap, l the length of the gap, and b the total width of the gap. The stiffness of the bearing is approximately:

$$k = 3\frac{W}{h}(1-\beta) \tag{3.44}$$

where W is the load, and β the pad pressure ratio, which is the pressure in the recess with the load lifted to the pressure required to lift the load. Typically it is 0.7 for a film thickness of 50 μm. Another formula of oil pressure change is:

$$\frac{dP}{dl} = \frac{12\mu}{wh^3}\frac{dV}{dt} \tag{3.45}$$

where w is the width of the fluid cross section and dV/dt the flowing fluid volume per unit time. To regulate the oil flow and to respond to the changes in loading, a hydrostatic bearing should have regulators that change the recess pressure in response to the flow rate. This is done through narrow tube (capillary), gaps between two cylinder side walls, or through an orifice. The formulas for these regulators with a pressure drop, δP, is respectively:

$$\begin{aligned} \delta P &= \frac{8L\mu}{\pi R^4}\frac{dV}{dt} \\ \delta P &= 6\frac{L\mu}{\pi R\delta R^3}\frac{dV}{dt} \\ \delta P &= \frac{\gamma}{169d^4}\left(\frac{dV}{dt}\right)^2 \end{aligned} \tag{3.46}$$

where L is the length, R the radius or mean radius, δR is the difference between them of the tube, γ is the fluid weight density, and d the diameter of the orifice.

The oil film thickness of the VLT telescope hydrostatic bearing is 50 μm. The bearing has a stiffness of 5 kN/μm and a friction moment of only 100–200 Nm. Since hydrostatic bearings have small friction force, the damping is also small. To increase damping of the 5 m Hale telescope, a small friction wheel at the northern side of the polar axis is added. Modern telescopes may use electromagnetic dampers in the system. If a static magnetic field exists around a moving metal component, damping forces will be produced due to the induced surface eddy current.

Hydrostatic bearing pads include few recesses. The recess area should be so designed that the static pressure from the area is sufficient to support the telescope load. The edge area of the pad contacts the bearing surface when the bearing is not in operation. Therefore, it should be precise, smooth and a little bit soft. To reduce the thermal effect on the telescope, pre-cooled oil may be used for the bearings.

The theory of air bearings is the same as the hydrostatic bearings. The working pressure of an air bearing is many times lower than that of a hydrostatic one since air has a much lower viscosity. The gap of an air bearing is about 15% of the hydrostatic bearing film thickness, just 10 μm. An air bearing has a higher stiffness, but the bearing surface shape has to be very accurate.

3.2.5 Structural Static Analysis

3.2.5.1 A Brief Introduction to Finite Element Analysis

Structural analysis is of vital importance for modern telescope design. With the development of computer and software technology, finite element analysis (FEA), instead of pure analytic calculations, becomes an important tool for engineers in telescope structural design. The FEA is based on structure elastic theory where a practical structure is modeled as a group of discrete elements. The structural deformations and stresses can be derived by solving linear or nonlinear equations with applied loads and constraints as the boundary conditions. The structure performance can be predicted before the construction.

The force and displacement equation of the FEA analysis is:

$$[K]\{u\} = \{F\} \tag{3.47}$$

where $[K]$ is the stiffness matrix, $\{\mu\}$ the displacement matrix, and $\{F\}$ the external force matrix. As the equation states, the internal reactions balance the external forces. Figure 3.21 shows a simple bar element under two external forces. The balance condition requires that the sum of these forces is zero, $F_2 = -F_1$. Because the element is under a pair of balanced forces F_1 and F_2, the relative elongation or the strain ε_x is:

$$\varepsilon_x = \frac{\Delta L}{L} = \frac{u_2 - u_1}{L} \tag{3.48}$$

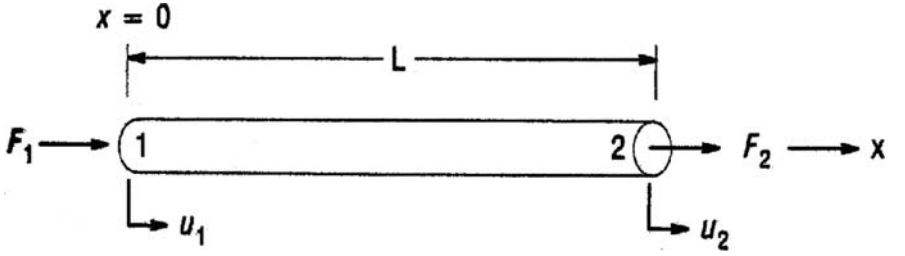


Fig. 3.21. The balance of external forces on a rod element.

where u_1 and u_2 are displacements at both ends of the bar. For an elastic material, the relationship between the stress and strain is:

$$\sigma_x = E\varepsilon_x \tag{3.49}$$

where E is the Young modulus of the material. According to the definition, the stresses at both ends are:

$$\begin{aligned} \sigma_x &= \frac{-F_1}{A} \\ \sigma_x &= \frac{-F_2}{A} \end{aligned} \tag{3.50}$$

Combining above equations, the derived force and displacement equations are:

$$\begin{aligned} -F_1 &= \frac{EA}{L}u_2 - \frac{EA}{L}u_1 \\ -F_2 &= \frac{EA}{L}u_2 - \frac{EA}{L}u_1 \end{aligned} \tag{3.51}$$

And using matrix expression, the equation is:

$$\frac{EA}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix} \tag{3.52}$$

Namely:

$$[K_i]\{u\} = \{F\} \tag{3.53}$$

In the above equation, $[K_i]$ is the stiffness matrix of the ith element. The global stiffness matrix is an assembly of individual element stiffness matrices.

For a single truss system, if one of the displacements u_i is constrained, then the other displacement can be determined by solving the displacement equation.

In the FEA, a complete structure is converted into many discrete small elements. These elements form a set of linear displacement equations. When the external forces and boundary conditions are applied, the linear equations can be solved using Gaussian elimination routines. However, a few general assumptions in the linear static FEA analysis exist: (a) elastic materials are used; (b) small deformations are assumed; and (c) external forces are applied slowly to the structure. At present, many FEA codes are available and the FEA has become a standard approach in modern telescope design.

With the development of the FEA codes, nonlinear static analysis can also be performed now. There are three types of nonlinear analysis: (a) Changing status, e.g. tension-only cable is either slack or taut, or a roller support is either in contact or not; (b) geometric nonlinearities, such as large deformations or large rotations; and (c) material nonlinearities, e.g. nonlinear stress-strain relationship or temperature-dependent material properties.

3.2.5.2 Purposes of Static Structural Analysis for Telescopes

Static structural analysis provides surface deformations of the primary and secondary mirrors, relative positional changes between optical components, and stresses within the structure under gravity, temperature change, and static wind loading. The mirror surface rms error and the line of sight pointing error are derived from special programs after the FEA static analysis. These errors are important to telescope designers.

Reliable modeling is crucial in the FEA analysis. Shell elements have only five degrees of freedom from which one rotational degree of freedom in the element plane is missing. Three node triangular plate elements are generally more rigid than they should be since the stress change inside the element is not considered. All nodes of the solid elements have only three translational degrees of freedom while those of the beam elements have six degrees of freedom. If a beam element is connected to two-dimensional or three-dimensional elements, the mismatching of degrees of freedom will bring errors in the calculation. The absent rotational degrees of freedom should be constrained by using other translational degrees of freedom to make the analysis more precise.

In the FEA analysis, correct bearing modeling is important. Bearings, especially large ones, have not only axial and radial stiffness but also the overturning stiffness. The element parameters used for the bearing are determined by these stiffness numbers. One method in bearing modeling is to arrange both bearing rings in the same circle and to connect corresponding nodes of both rings by $2N$ spring elements, where N is the node number in each bearing ring. These spring elements include N elements in an axial direction and N elements in a radial direction. Nodes of both rings are connected to the same central node for both circles with rigid elements. The axial stiffness of the bearing is a sum of

axial spring stiffness. If radial bearing stiffness is K_R, the spring constant of each radial spring, K_{radial} is derived from the following equation:

$$K_R = 2 \sum_{n=-N/4}^{+N/4} K_{radial} \cos \frac{2\pi \cdot n}{N} \tag{3.54}$$

If 48 spring elements are used in a bearing model, the cosine summation in the formula is 15.252. For over-turning stiffness K_M, the stiffness of each axial spring, K_{axial} should satisfy the following equation:

$$K_m = \left(\frac{D}{2}\right)^2 \sum_{n=-N/4}^{+N/4} K_{axial} \cos \frac{2\pi \cdot n}{N} \tag{3.55}$$

where D is the diameter of the bearing. The analysis of hydrostatic bearing is also important for stress prediction of each component under loading.

When a structure involves materials with different coefficients of thermal expansion, thermal analysis is essential. Thermal loadings include temperature gradients and absolute temperature change. A bi-metal effect should be avoided in any precision structures. If different coefficients of thermal expansion materials are used, soft connection is recommended to absorb thermal induced stresses. Shaped sandwiched structures made of materials with different coefficients of thermal expansion have shape changes when temperature changes. The discussion of this is in Section 8.3.2.

Integrated modeling is now an essential part of the structure analysis. Structure-optical analysis can predict the optical performance under structural loadings. In the analysis, the deformed mirror surfaces are expressed in a Zernike polynomial form. The surface shape change is added to the optical system for detailed ray tracing. From ray tracing, image spot size and line of sight pointing error are produced. The deformation of a very large SMT-type primary mirror under gravity usually retains a repeatable pattern with noticeable correlation between segments. A simplified (approximate) image spread function through a Fourier transform of the aperture phase function may not provide accurate image intensity distribution as small but repeatable surface error produces serious scattering and interference in the image plane. In this case, a simple relationship between surface rms error and Strehl ratio does not exist (Section 7.1.2) and ray tracing is more accurate. In the combined structure-optical analysis, if random distributed errors, such as positioning or manufacture ones, are involved, the Monte Carlo method should be used in performance evaluation. For this analysis, random number generation is required. One group of random number is for one design variable. The product of random number and the error tolerances will provide a simulation of the system under random distributed error. The procedure should be repeated a few more times to determine the possible performance range of the system. The structural dynamic analysis is discussed in Section 3.4.2.

3.3 Telescope Drive and Control

3.3.1 Specifications of a Telescope Drive System

High accuracy of star pointing and tracking is a basic requirement for the optical telescope drive system. To achieve this, a frequent pointing model update is necessary. This requires a quick pointing check for hundreds of stars all over the sky in the beginning of each night observation. Therefore, there are the following different movement modes for an optical telescope.

3.3.1.1 Slewing

Slewing is a rapid motion used to move a telescope rapidly from one location to another in the sky. Slewing is also used to change a telescope position in order to change its configuration allowing it to replace instruments or to stow the telescope in a predetermined position. The maximum velocity of slewing is about 1–3°/s. The maximum slewing acceleration is about 0.1–0.3°/s^2. Large telescopes usually have smaller slewing velocity and acceleration. During the emergency braking, the telescope's motion due to the inertia from slewing should be less than 2°. Special purpose telescopes, such as near earth object searching ones, require very high slewing rate.

3.3.1.2 Star Acquisition

Star acquisition is required when a telescope is only a few tens arcsec away from a target star. The velocity of the telescope in this mode is less than 2 arcmin/s and the pointing error produced is called blind pointing error. This blind pointing error should be smaller than about 1 arcsec after calibration and pointing model updating. The star acquisition should bring the star right on the desired fiducial position. A telescope with poor pointing accuracy brings difficulties for astronomers to verify the required celestial targets. If the target is too faint or is an extended (nonpoint) source, a bright reference star nearby of known position is needed. If the target is pointed with an offsetting (or blind offsetting) from a reference star through an open loop movement of the telescope, the offset pointing accuracy is about 0.1 arcsec. The above pointing requirements are for ground-based telescopes. For space telescopes with an order smaller star image, the corresponding pointing requirement may be an order higher than that of the ground-based counterparts.

3.3.1.3 Star Tracking

During star tacking, the telescope is synchronous with the motion of a celestial target. Since the star guiding device can realize even higher performance than star acquisition, tracking pointing accuracy of a telescope is generally high, at

below 0.1–0.02 arcsec. Space telescopes have very high tracking accuracy, allowing repeated exposure of the same sky area. The tracking velocity of an equatorial telescope is 15 arcs/sec. For an alt-azimuth telescope, it is given by the following formula:

$$\begin{aligned} T^2 &= \left(\frac{dA}{dt}\right)^2 \cos^2 Z + \left(\frac{dZ}{dt}\right)^2 \\ &= \left(\frac{dh}{dt}\right)^2 \cos^2 \delta + \left(\frac{d\delta}{dt}\right)^2 \approx \omega^2 \cos^2 \delta \end{aligned} \tag{3.56}$$

where ω is the rate of rotation of the celestial sphere. Generally, the maximum azimuth tracking velocity of an optical telescope is between 0.5 and 1°/s and the maximum elevation tracking velocity is about 15 arcsec/s. The maximum azimuth acceleration is about $\pm 0.02^\circ/\mathrm{s}^2$. The tracking error is expressed as:

$$\varepsilon^2 = (\Delta A)^2 \cos^2 Z + (\Delta Z)^2 \tag{3.57}$$

In the expression, the term $\cos Z$ is small near the zenith.

In addition to the above three basic modes of telescope motion, telescopes may have the following special modes of motion for meeting special observational requirements.

3.3.1.4 Scanning

A telescope may be required to scan a sky area back and forth, line by line, or follow a spiral curve from the center outwards. The rate of scanning depends on the integration time of the detector. A typical scanning speed is 20 arcsec/s.

3.3.1.5 Chopping and Fast Switching

In some cases, telescopes or their secondary mirrors are required to switch back and forth between two sky positions. The switching angle is between a few arcsec and a few arcmin. The switching speed equals the slew speed of the telescope. In radio telescopes, the switching angle used is larger, up to 1.5°.

3.3.1.6 Whole Sky Survey

It is often required that the telescope points to all sky positions. For convenience, the azimuth angle is usually required to move within the range of 360° or beyond, and the elevation angle is required to move within the full range of 0–90°.

In all these modes of telescope motion, the basic requirements are smooth, accurate, and highly repeatable positioning. Therefore, a high quality drive and control system is required.

3.3.2 Trends in Drive System Design

Earlier equatorial mounting telescopes used an accurate worm gear system to achieve a constant velocity around the polar axis. The advantages of a worm gear system are high accuracy, high smoothness, larger reduction ratio, high stiffness, and high tolerance for structural imbalance. The disadvantages are low efficiency (14–15%), high cost, limited size, high alignment requirement, and the irreversibility (self locking) in motion. Because of the irreversibility, the worm gear drive requires a buffer protection mechanism to protect the telescope when a sudden change of speed happens. This nonlinear property of a worm gear makes it incompatible with modern servo control systems.

From the 1970s, the worm gear had been abandoned for large optical and radio telescopes. Spur gears (or helical gears) were used in many 4 meter class optical telescopes before 1980. Spur gears have higher moment transmission efficiency (85%) and have no irreversibility problem. The backlash of the gear is eliminated through a torque biased pinion pair system. In this system, an identical pinion drive is added and it has a constant torque difference with the original one as the axis is in driving. The torques on both pinions are equal and opposite in direction when the telescope has a zero velocity. These two pinion systems act together to drive the telescope.

A spur gear drive is a linear system which is suitable for the telescope control system. In the control system, a highly accurate angular encoder provides positional information so that the gear accuracy required is reduced. A major requirement is the smoothness of motion. The cost of the spur gears is still high, so that the roller (friction) or direct (torque and linear motors) drive system is becoming more popular in telescope design.

Roller drives are inherently smoother, less expensive, and more accurate than a gear drive. They can achieve a larger reduction ratio, so that the drive train stiffness is improved. In a roller drive system, a small roller is simply pushing against a cylindrical journal surface on the telescope axis. The moment transferred by the rollers is determined by the friction coefficient.

Adequate pressure between the roller and journal surface has to be kept to prevent radial slippage. The radial runout of a drive roller changes the distance between two axes which may cause variation of the contact stress. The maximal contact stress between two cylindrical surfaces for a Poisson ratio of 0.3 is (Bely, 2003):

$$\tau_{\max} = 0.591 \left[\frac{P(1/r_1 + 1/r_2)}{L(1/E_1 + 1/E_2)} \right] \tag{3.58}$$

where P is the pressure between two wheels, E Young modulus, r_1 and r_2 the radii of the roller and journal, and L the length of contact. Typically the pressure is provided by a loading roller on top of the drive roller for minimizing moments on the roller surface. Excessive contact stress between the roller and journal

surface is harmful, but inadequate contact stress may cause slippage and error. An optimal contact stress is derived through experiments.

Axial slippage, caused by misalignment between the axes, is more harmful as it introduces jitter in the drive. The drive will jump after enough error has accumulated and the roller or journal surfaces may degrade quickly. The roller should be made of slightly softer material (brass) than the journal. The roller design has to ensure parallelism between the roller and journal axes.

Another consideration in a roller drive is the resistance against any external turbulence, such as wind. The required drive moment has to be higher than the turbulence moment. A roller drive requires a braking device, especially when it is used on the elevation axis. If a self-aligned journal is driven by several rollers, care has to be taken to eliminate axial eccentricities and positional error caused by temperature or other factors. A self cleaning device is also recommended to minimize surface degradation from any foreign objects (i.e., dirt) on the drive surface.

A direct motor drive, which eliminates all the gear or roller trains between the motor and telescope axis, is another choice for the telescope drive. For small size telescopes, commercial torque motors can be used for their direct drive systems. However, for large telescopes, their direct drive motors are specially ordered. These noncontact linear motors have a number of 1 m long curved racetrack magnetic and winding pads facing the race. The air gap in between is of the order of a few millimeters.

Direct drives with no moving parts are friction free, stick-slip free, low tolerance, and low maintenance. Direct drives provide the highest stiffness, which increases the structural locked rotor frequency. Using a direct motor drive, the drive force is not concentrated on one point, but distributed over an area, so the local deformation is minimized. The disadvantages of the direct drive are high cost, torque ripple, and electromagnetic cogging. The cogging is a condition at low speed where motor rotation is jerky as two magnetic fields interfere with each other. Generally, the ripple and cogging is much less than 1%.

3.3.3 Encoder Systems for Telescopes

An angular encoder provides an accurate position indication of a shaft. Various angular encoders, such as optical encoders, Inductosyns (a trademark for a resolver/synchro assembly), grating tapes, and gyroscopes, have been used in astronomical telescopes.

3.3.3.1 Optical Encoder

An optical encoder is a binary position transducer. The fundamental principle may be described as a glass disk with radial gratings of different periods illuminated by light so that a binary signal can be obtained at the other side of the disk using photoelectrical sensors. These binary signals provide either an absolute or incremental angular position of a shaft. An absolute encoder provides a unique angular

position relative to a reference position while an incremental one provides the change of the angular position.

The coded disk is a key component of an optical encoder. For absolute encoders, each bit of resolution requires an additional code ring on the disk. More bits involve more code rings. The size and cost of the encoder increase. Figure 3.22 shows an 8-bit absolute classical coded disk. It is coded with gratings of gradually reduced periods. However, this disk has an ambiguous problem as more than one bit in the code rings may change when the angle changes. If dirt appears on the disk, an error appears.

To avoid the ambiguity of the classical disk, a Gray code disk is often used. In a Gray code, only one bit changes between any successive code words, producing the least uncertainty. A 4-bit Gray code disk is shown in Figure 3.23(a). The rule of a Gray code grating arrangement is: (a) a 1-bit Gray code has two words 0 and 1; (b) the first 2^{n-1} code words of an n-bit Gray code equals the code words of a (n-1)-bit Gray code written in order with a leading number of 0 appended; and (c) the last 2^{n-1} code words of a n-bit Gray code equals the code words of a (n-1)-bit Gray code, written in a reverse order, with a leading number of 1 appended. An 8-bit Gray code is arranged as: ring 0, 000 000; ring 1, 001 001; ring 2, 010 011; ring 3, 011 010; ring 4, 100 110; ring 5, 101 111; ring 6, 110 101; and ring 7, 111 100. The disk code is easily changed to binary numbers by a simple logic circuit.

An incremental encoder has only one grating ring. Generally, with the same angular resolution, the cost of an incremental encoder is much lower than that of an absolute one. An incremental (or tachometer-type) code disk is shown in Figure 3.23(b). The output of the encoder consists of a sine or a square wave signal whose resolution is determined by the line number on the disk.

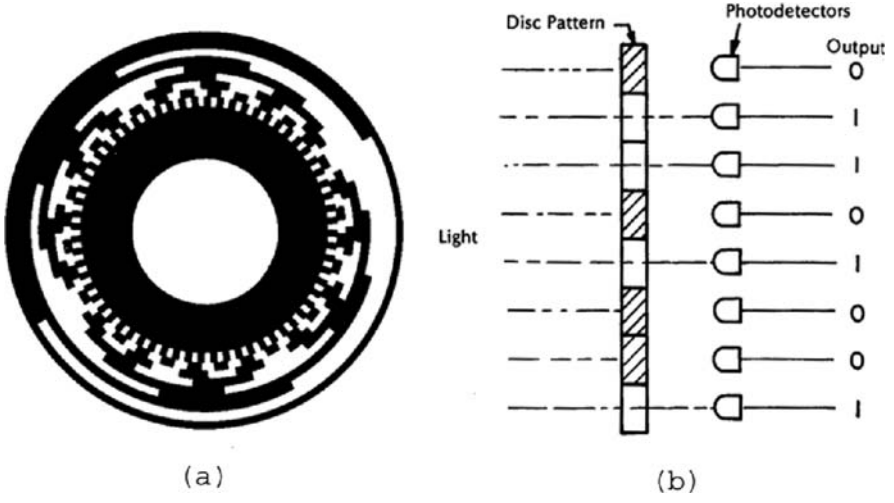


Fig. 3.22. (**a**) An 8-bit code disk and (**b**) schematic diagram of an encoder.

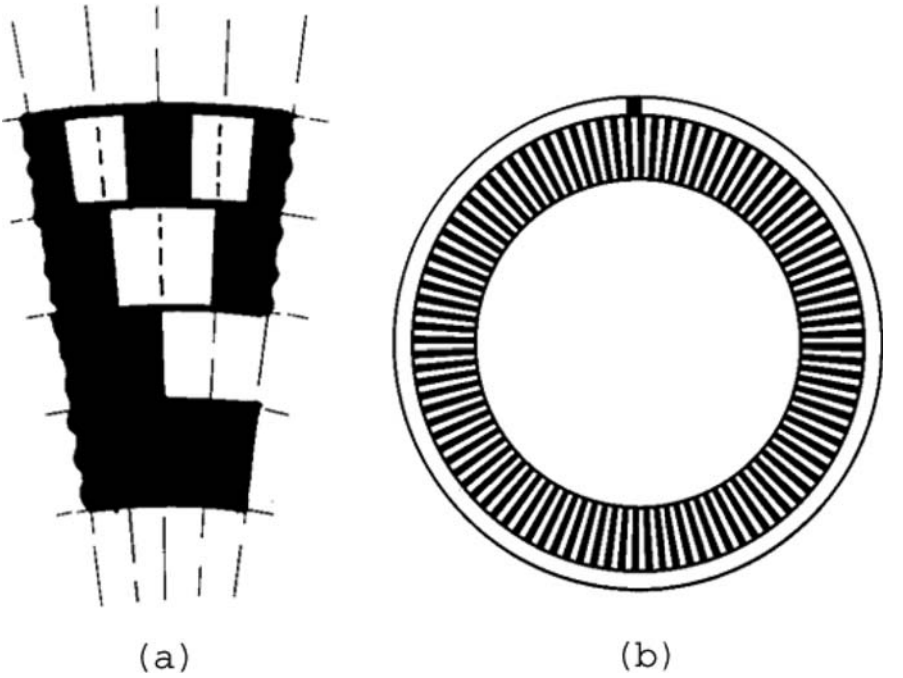


Fig. 3.23. (**a**) A 4-bit Gray code disk and (**b**) an incremental code disk.

If only a single output is received, the direction of the motion is difficult to obtain and the resolution is also fixed. The velocity can be measured. Therefore, it can be used as a velocity transducer in a feedback loop. Both the direction and finer resolution can be achieved if more tracks are used on the disk. This type of encoder is a quadrature one. A quadrature encoder has at least two output signals: Channel A and B as shown in Figure 3.24. The Channel B code ring has a 90° phase offset from the Channel A code ring, resulting in a phase shift in the output signal. Alternatively, this code ring offset can be replaced by an optical receiver position offset.

The first benefit of the quadrature encoder is its ability to detect the direction of rotation. A convention of clockwise (CW) rotation represents Channel A leading Channel B and that of counterclockwise (CCW) represents Channel B leading Channel A. Using one count per encoder cycle (P1) as an example, if Channel A rises before Channel B, a CW count is generated on the rising edge of Channel A. If Channel A falls behind Channel B, a CCW count is generated on the falling edge of Channel A. As shown in output P1, the opposite edges of the Channel A output must be used to generate CW and CCW count pulses. This directional ability is critical.

This encoder has a higher resolution. In P2 output option, both rising and falling edges of Channel A are used to generate counts, resulting in a doubling of

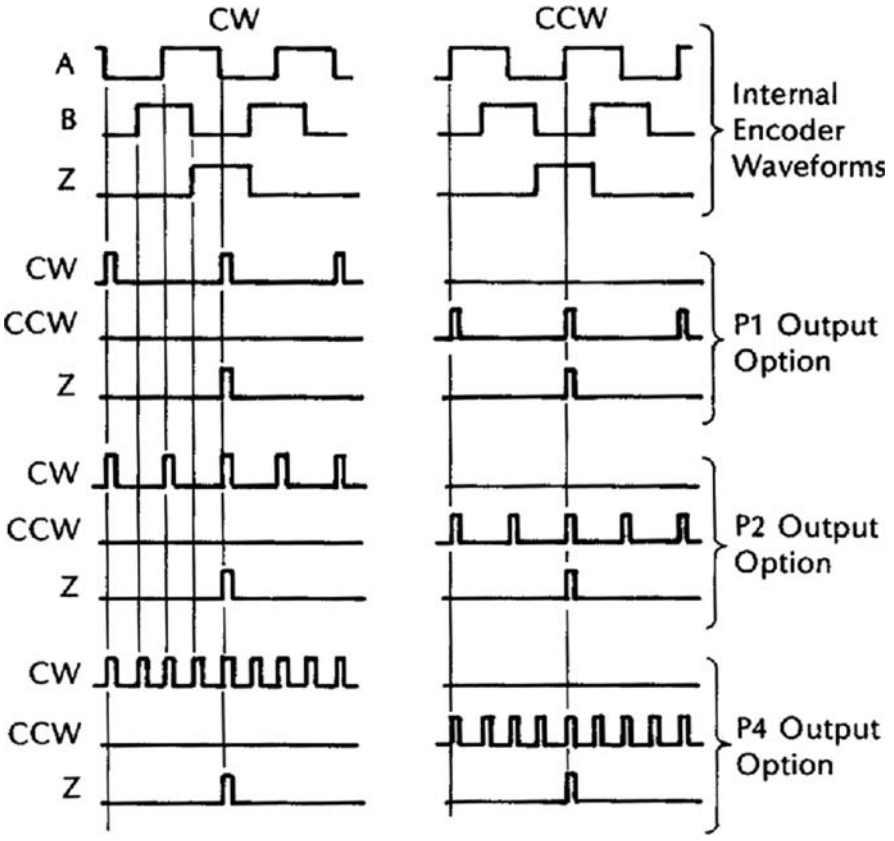


Fig. 3.24. Resolution enhancement for an incremental encoder.

resolution while leaving both edges of Channel B unused. In P4 output, both rising and falling edges of Channels A and B are used to generate counts, effectively increasing resolution in the P1 output by four.

Another feature of this type of encoder is the index pulse, occurring only once per encoder revolution. The index pulse can be used to preset and reset the position of the encoder.

To further improve the encoder resolution for both absolute and incremental encoders, one or more secondary scanning gratings have to be used as shown in Figure 3.25. For a coherent light beam passing though the scanning grating, the light intensity is (Ieki et al., 1999):

$$E_1(\xi) = \frac{i}{\lambda Z}\exp\left(\frac{i2\pi Z}{\lambda}\right)t_1(x)\int \exp\left(i\frac{2\pi}{\lambda}\frac{(x-\xi)^2}{2Z}\right)dx \qquad (3.59)$$

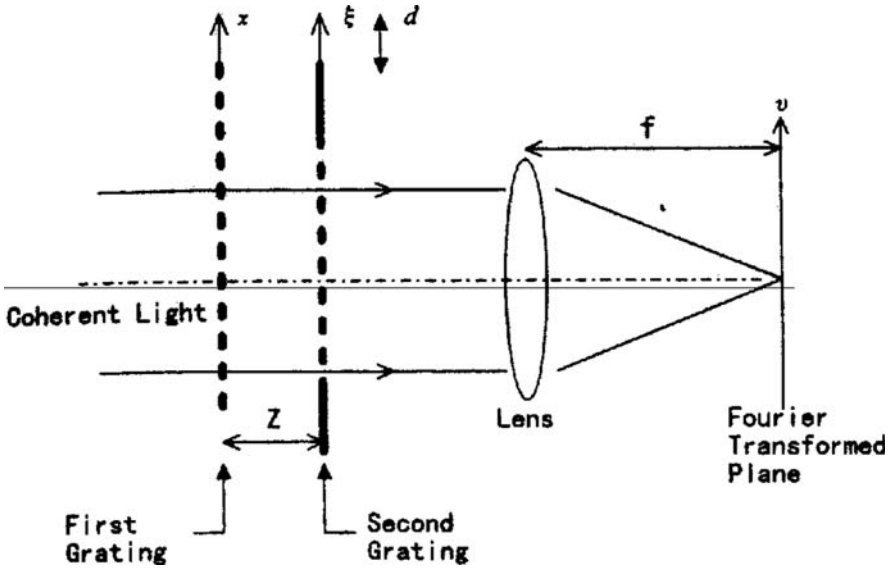


Fig. 3.25. The principle of using a scanning grating in incremental encoders (Ieki et al., 1999).

where t_1 is transmissivity of the grating and λ the wavelength of the light. If the period of the scanning and encoder grating is the same as P, and the spatial frequency in the focal plane is $\omega = v/\lambda f$, where v and f are defined in Figure 3.25, then the light intensity on the focal plane is (Ieki et al., 1999):

$$|E_2(\xi)|^2 = \left| \frac{1}{Z} \iint t_1(x) t_2(\xi) \exp\left[i2\pi \left(\frac{(x-\xi)^2}{2Z} - \omega\xi \right) \right] dx \right|^2 \tag{3.60}$$

To express $t_2(\xi)$ as a Fourier series, then:

$$t_2(\xi) = \sum_{k=-\infty}^{\infty} {}_2C_k \exp\left(\frac{i2\pi k\xi}{P} \right) \tag{3.61}$$

In the above equations, ${}_2C_k$ is a Fourier coefficient and its expression is:

$${}_2C_k = \frac{1}{P} \int_{-P/2}^{P/2} t_2(t) \exp\left(-\frac{i2\pi kt}{P} \right) dt \tag{3.62}$$

If L is the length of the scanning grating, $M = Z/\lambda P^2$, where Z is defined in Figure 3.25, the variation in light intensity caused by both gratings is:

$$\left|E_2\left(\frac{n}{P}\right)\right|^2 = \frac{PL}{Z^{1/2}}\left|\sum_{k=-\infty}^{\infty} {}_1C_{n-k2}\, C_k \exp\left[-i\pi(n-k)\left(\frac{2d}{P} + M(n-k)\right)\right]\right|^2 \tag{3.63}$$

where d is the relative displacement between two gratings in the x direction. So:

$$\begin{aligned}
|E_2|^2 \approx & \frac{1}{12} + {}_1C_{12}\, C_1 \cos\left(\frac{2\pi d}{P}\right)\cos(\pi M) \\
& + \sum_{k=3,5,7,9\ldots}^{\infty} {}_1C_{k2}\, C_k \cos\left(k\frac{2\pi d}{P}\right)\cos(\pi k^2 M) \\
& + 2\sum_{k=1,3,5,7,9\ldots}^{\infty} {}_1C_{k2}^2 C_k^2 \cos\left(k\frac{2\pi d}{P}\right) \\
& + \sum_{m=1}^{\infty}\sum_{k=1,3,5,7\ldots}^{\infty} {}_1C_{k1}\, C_{m2}\, C_{k2} C_m \\
& \times \left[\cos\left((k+m)\frac{2\pi d}{P}\right) + \cos\left((k-m)\frac{2\pi d}{P}\right)\right]\cos\left[\pi M(k^2 - m^2)\right] \\
m = & \ 1, 3, 5, 7, \ldots; k(= 1, 3, 5, 7\ldots) > m
\end{aligned} \tag{3.64}$$

When $M = 0$, the light intensity can be expressed as a series. Taking the first two terms of this expression, the intensity is an exact cosine function of the relative displacement d. In practice, four groups of the scanning gratings are used instead of one for increasing the resolution.

As shown in Figure 3.26, the light intensity on the focal plane for a normal scanning grating is not a true cosine curve. To get a cosine intensity curve, two methods can be used. The first is to use scanning gratings with uneven grating spacing as shown in Figure 3.27. The other is to use two light sources with different widths instead of one wide width. If the light intensity is a cosine function, fine calibration can increase greatly the resolution.

Combining incremental gratings of different periods on the same disk is another way to achieve a very high resolution absolute encoder. If cosines and sines with different frequencies are available, the higher resolution absolute position can be defined through calibration. In this way, very high accuracy (>24-bits) absolute encoders can be made from grating rings of much lower orders such as 3-bit to 15-bit ones.

The absolute angular positioning is a necessary requirement for a telescope. An absolute encoder may be coupled directly with the drive axis of the telescope so that there are no other error sources in the system. However, when the encoder used has a lower resolution, the encoder can be used through a gear or

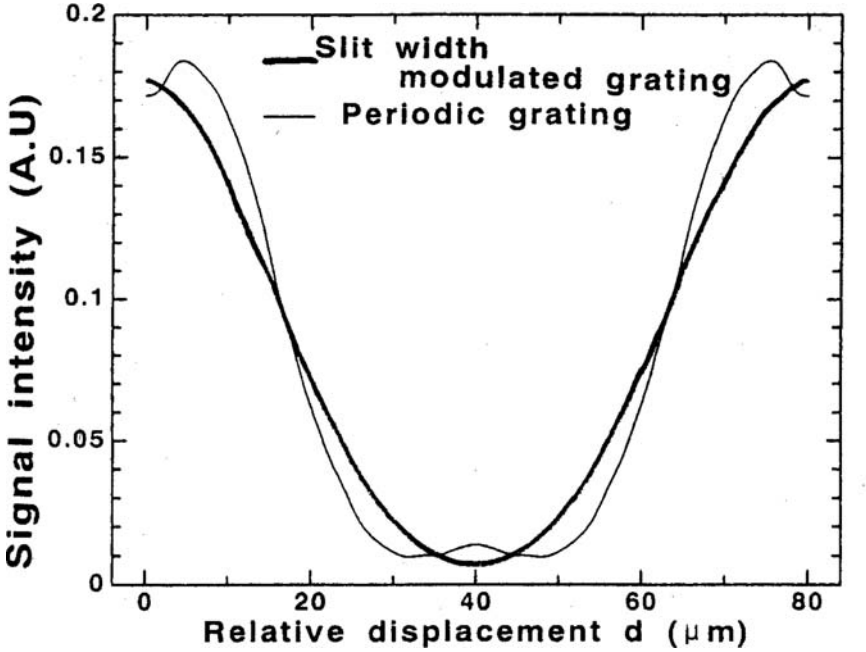


Fig. 3.26. The relationship between light intensity and displacement of an incremental encoder with a scanning grating, A.U means arbitrary unit. (Ieki et al., 1999).

roller train with magnification. However, the profile error of the gear or roller influences the accuracy of the angular measurement.

Combining a modern optical grating with a CCD camera can also achieve very high linear or angular accuracy. If a 16-bit incremental disk with an additional 16-bit absolute position pattern is used together with a CCD camera, an absolute resolution of 24-bit disk is achievable. The optical encoder in analog control systems requires a digital-analog converter. Today magnetic data storage is widely used. Magnetic encoders may soon play roles in the encoder field.

3.3.3.2 Inductosyns

Another angular transducer used for telescopes is the Inductosyn. Unlike an optical encoder, an Inductosyn is an analog device. Its output signal is not discrete but continuous.

An Inductosyn is a multiple pole synchro paired with a resolver. Just as an LVDT (linear variable differential transformer) measuring device, a synchro is a rotational variable differential transformer. A synchro consists of a stator and a rotor as shown in Figure 3.28. The rotor has one coil and the stator has n coils composing $2n$ poles. The interval of angle between poles is $360^\circ/n$. The stator of

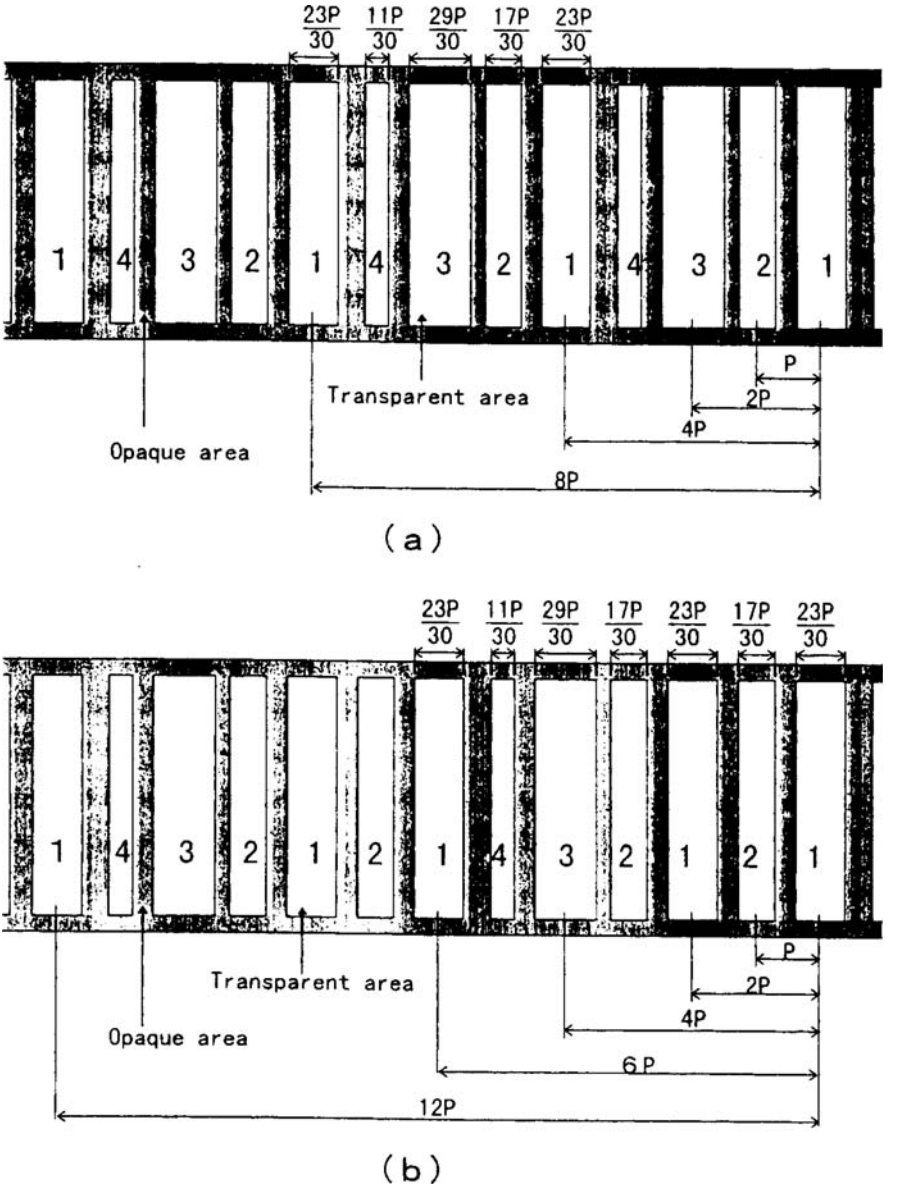


Fig. 3.27. Detailed dimensions of two special scanning gratings for an incremental encoder which provide a cosine light intensity change (Ieki et al., 1999).

a basic synchro has three coils. For this type of synchro, when a current $e_r(t)=E_r \sin\omega_e t$ passes through the rotor and the shaft of the rotor has an angle of θ from its zero position of the stator, the stator coils will produce signal waves phase shifted by 120° from each other and so called sine/cosine

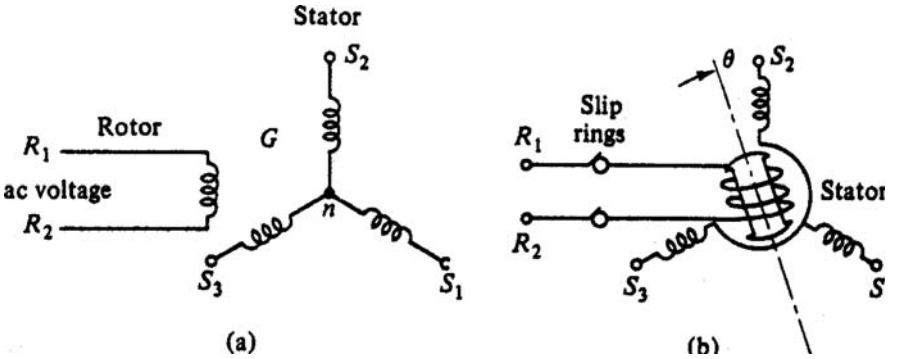


Fig. 3.28. Operational principle of a synchro.

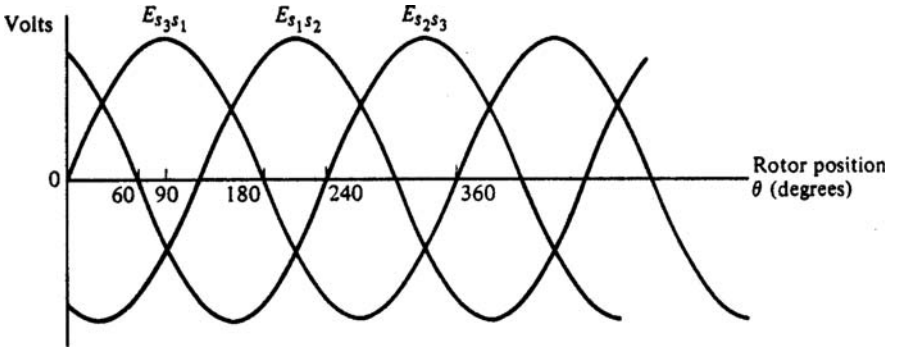


Fig. 3.29. The output voltage produced by the stator coils of a synchro.

signals as shown in Figure 3.29. A synchro-to-digital converter can be used to produce a binary absolute code from the sine/cosine outputs.

$$\begin{aligned} E_{S1} &= KE_r \cos(\theta - 240^\circ) \\ E_{S2} &= KE_r \cos\theta \\ E_{S3} &= KE_r \cos(\theta - 120^\circ) \end{aligned} \tag{3.65}$$

In the above equations, K is a constant. If the stator coils are connected as in Figure 3.28, the voltages produced will be:

$$\begin{aligned} E_{S1,S2} &= \sqrt{3}KE_r \cos(\theta + 240^\circ) \\ E_{S2,S3} &= \sqrt{3}KE_r \cos(\theta + 120^\circ) \\ E_{S3,S1} &= \sqrt{3}KE_r \cos\theta \end{aligned} \tag{3.66}$$

By calibrating the produced voltages, a high angular resolution of the shaft position can be determined unambiguously.

Just as a synchro, if the stator has only one coil and the rotor has two perpendicular coils, the rotational transformer is now a resolver. The advantage of a resolver is that the output voltages of the rotor are cosine and sine components of the voltage $U_0 = U \sin \omega \cdot t$ applied on the stator.

The inductosyn is a multi-pole synchro/resolver as shown in Figure 3.30. The stator of an inductosyn is named a scale and the rotor a slider. Both the scale and slider are circular plates with coils. The gap between them is very small. The coils in scale and slider have the same period. The distance between two coils in the slider is a multiple of a quarter period. So, the voltages of these two coils are cosine and sine components:

$$\begin{aligned} U_{12} &= KU_1 \sin(2\pi \cdot x/p) \\ U_{22} &= KU_1 \cos(2\pi \cdot x/p) \end{aligned} \tag{3.67}$$

where x is the linear distance and p the period. An inductosyn is an incremental encoder. However, using coils with different periods on both plates, a modern inductosyn can be an absolute encoder. Compared to optical encoders, an inductosyn costs less and is more durable in a harsh environment. Therefore, it can be used when temperature change or system vibration are concerns. The negative side is a periodic error introduced by a very small difference in the maximum amplitude of the sine and cosine outputs. This difference in the maximum amplitude is interpreted as a phase difference that results in a positional error. The MMT observatory and NRAO have developed circuits that reduce the periodic effect to a usable 24-bit level.

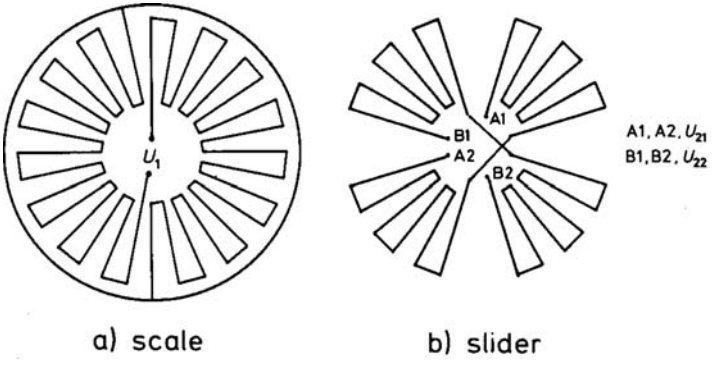


Fig. 3.30. Details of Inductosyn scale and slider.

3.3.3.3 Other Angular Encoders

An optical grating tape plus Moire fringe readers are also used as angular encoders for the 4.2 m William Hershel Telescope (WHT). The Moire fringe reader has a scanning grating which has its lines at a small angle with the grating lines. The accuracy of the grating tape used is about 1 μm and the resolution is a function of the diameter of the surface where the grating tape is attached. The problem in using a very long grating tape is the inconsistency of the period. Therefore, a computer lookup-table is used for its calibration. The grating tape provides absolute angular positioning.

A high-bit absolute encoder can also be formed by a high-bit incremental encoder plus some absolute positioning devices. These positioning devices include a tilt meter, proximity sensors, resolver, or grating marks. The ARC 3.5 m telescope uses this electromagnetic proximity sensor with a repeatability of 1 μm as a positioning indicator. The sensor is placed every 10–15°. Between these angles, an incremental encoder is used. The encoder resolution of this telescope is also amplified through a roller drive device.

The tachometer is another type of sensor. It is usually connected on the motor shaft. The voltage produced from the tachometer is proportional to the shaft velocity. The tachometer's information is used in the velocity control loop. However, a tachometer does not provide angular position of the axis.

Precision encoders should be mounted correctly to assure their accuracy and resolution. When an encoder is coupled with a drive shaft, it is necessary to avoid any force or moment which may apply to the encoder shaft. The coupling should have a high torsional stiffness, while keeping a minimum radial and axial stiffness.

For new Stewart platform mountings or space-based telescopes, gyroscopes are usually used for angular positioning purposes. The discussion of gyroscopes is in Section 5.2.1.

3.3.4 Pointing Error Corrections

No telescope structure and control system are perfect so the pointing error, which is a difference between the commanded and real position of a telescope, always exists. Factors causing pointing errors include the error in a drive system and encoders, the axial nonperpendicularity, the tube deformation, the atmospheric differential refraction, the atmospheric seeing, the wind and thermal induced error, and even the power voltage variation. Pointing errors include repeatable one and nonrepeatable one. Most of the pointing errors are repeatable. However, errors from the backlash, clearance, friction, and hysteresis are nonrepeatable and are not correctable.

Two methods, a physical one and a mathematical one, are used for the pointing error correction in astronomical telescopes. Using a physical method, physical laws of pointing error induced are studied, especially for those which produce significant pointing errors. The pointing error functions related to the

telescope positions, $\Phi A_i(A,Z)$ or $\Phi Z_i(A,Z)$, are established for all the error sources, most of them from structural deformations or encoder offset. The atmospheric refraction and other sources, such as temperature or humidity, are also included in the formulation. All these functions are linearly independent, so that total pointing error is the sum of individual terms.

The coefficients of all error functions are determined using least square fitting of the known stars coordinates from an all sky pointing check. The derived coefficients are used for subsequent pointing error correction in observation. The physical method involves few terms and the pointing improvement is significant and straightforward. The functions involved are mostly triangular ones as the gravity components follow sine and cosine laws. For an alt-azimuth mounting telescope with an azimuth bearing being supported on n points, a set of sin nA and cos nA terms is also required in the error correcting formulas.

Using mathematical method, the terms used do not link directly to any physical error sources. The function terms are arbitrarily selected, but they are linearly independent. The coefficients are also derived through the least square fitting after observational pointing check. Using this method, more terms in pointing error correction formulas are usually required. The optimization is a multi-variable one and has a solution in theory. However, the convergence may be slower.

In practice, the correction formula used by astronomers is a mixture of both methods. For an English style equatorial telescope, widely used pointing correction formulas are (Cheng, 1987):

$$\begin{aligned}\Delta_\delta &= a_{10} + a_{11}\sin t + a_{12}\cos t + a_{13}\sin\delta + a_{14}\cos\delta \\ &\quad + a_{15}\cos t\sin\delta + a_{16}\cos t\sin^2 t + a_{17}\cos^3 t \\ \Delta_t &= a_{20} + a_{21}\sin t + a_{22}\sin\delta + a_{23}\cos\delta + a_{24}\sin t\cos t \\ &\quad + a_{25}\sin t\tan\delta + a_{26}\cos t\tan\delta + a_{27}\sin^3 t\tan\delta \\ &\quad + a_{28}\cos^2 t\sin t\tan\delta\end{aligned} \tag{3.68}$$

where t is the hour angle, δ the declination, a_{10}, a_{11}, a_{12}, a_{20}, a_{25}, a_{26} the error coefficients related to hour angle, a_{16}, a_{17}, a_{24}, a_{27}, a_{28} the error coefficients related to the fork arm, a_{22}, a_{23} the nonorthogonal error coefficients of the two axes, and a_{13}, a_{15} the error coefficients of tube deformation. In addition, if the encoder has a n-bit subdivision, it may cause a pointing error with a period of $\pi/2^{n-1}$.

Widely used pointing formulas for an alt-azimuth telescope from Wallace are (Mangum, 2005):

$$\begin{aligned}\Delta_A &= -IA - CA\sec E - NPAE\tan E - AN\sin A\tan E \\ &\quad - AW\cos A\tan E - ACEC\cos A - ACES\sin A + \Delta A_{obs}\sec E \\ \Delta_E &= IE - AN\cos A + AW\sin A + HECE\cos E \\ &\quad + HESE\sin E + \Delta E_{obs} + R\end{aligned} \tag{3.69}$$

where A is the azimuth angle, E the elevation angle, IA and IE the zero point offsets of azimuth and elevation encoders, CA the nonorthogonality between optical (the tube) and elevation axes, $NPAE$ the nonorthogonality between axes, AN and AW the north-south and east-west tilts of the azimuth axis (north and west are positive), $ACEC$ and $ACES$ the cosine and sine components of the azimuth centering error, $HECE$ and $HESE$ the cosine and sine components of the tube vertical flexure, $\Delta A_{obs}, \Delta E_{obs}$ the azimuth and elevation correction applied by the observer, and R the atmospheric refraction coefficient. Sometimes, tangent terms instead of sine terms are used in the pointing error correction formulas.

If the pointing errors are known, the real azimuth and elevation encoder angles required to point to the object are given by:

$$\begin{aligned} A &= A_{demand} + \Delta_A \\ E &= E_{demand} + \Delta_E \end{aligned} \tag{3.70}$$

where A_{demand} and E_{demand} are the catalog azimuth and elevation angles of the object.

The real angular distance in a sphere is a function of cosine elevation, so the total pointing error is:

$$\Delta = \left[(\Delta A \cos E)^2 + \Delta_E^2 \right]^{1/2} \tag{3.71}$$

If a cross elevation component is included, the total pointing error is:

$$\Delta = \left[(\Delta_A \cos E)^2 + \Delta_E^2 + (\Delta_{EL} \sin E)^2 - \Delta_A \Delta_{EL} \sin 2E \right]^{1/2} \tag{3.72}$$

where Δ_{EL} is the cross elevation error. The effect of atmospheric refraction is usually:

$$\Delta Z = 60 \frac{P}{760} \frac{273}{273 + T} \tan Z \tag{3.73}$$

where Z is the zenith distance, P the atmospheric pressure with its unit mmHg, and T the absolute temperature. A more accurate expression of atmospheric refraction can be found in the work of Yan (1996).

3.3.5 Servo Control and Distributed Intelligence

Modern telescopes require high pointing and tracking accuracy which is difficult for an open-loop control system. Earlier classical telescopes use accurate worm

or spur gear drives without angular encoders. The accuracy reached is much worse than the arcsec and subarcsec level required for modern optical telescopes. To improve the pointing and tracking accuracy, angular encoders and closed-loop control systems are used in modern telescopes.

The transfer function of a modern control system is usually expressed as a Laplace transform. The Laplace transform of a function $f(t)$ is:

$$F(s) = \int_0^\infty f(t)e^{-st}dt \tag{3.74}$$

The Laplace transform can be expressed as $F(s) = L[f(t)]$, where s is a complex variable. For a discrete data system used in digital control, the equivalent one is z transform defined as $z = \exp(Ts)$, where T is the sample period and s the same variable used in a Laplace transform. Some Laplace and z transforms of time functions are listed in Table 3.1.

For the simple spring-mass-damper system shown in Figure 3.31, the corresponding dynamic equation is:

$$mx'' + bx' + kx = u(t) \tag{3.75}$$

where x is the displacement, u the force applied, m the mass, b the damping, and k the stiffness. When a sensor is attached to the mass block, the equation for the displacement measurement is:

$$y(t) = px(t) \tag{3.76}$$

where p is the gain in measurement. In Laplace space, the above two equations become:

$$\begin{aligned} (ms^2 + bs + k)X(s) &= U(s) \\ Y(s) &= pX(s) \end{aligned} \tag{3.77}$$

Table 3.1. Table of Laplace and z transforms

Time function	Laplace transform	z-transform
Unit step	$1/s$	$z/(z–1)$
e^{-at}	$1/(s+a)$	$z/(z–e^{-aT})$
$d^n f(t)/dt$	$s^n F(s)–s^{n-1}f(0)–\ldots f^{(n-1)}(0)$	

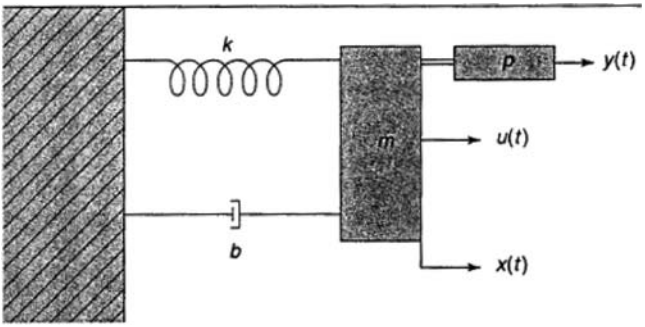


Fig. 3.31. A simple mass-spring-damper system.

where $s = \sigma + j\omega$. The transfer functions or system gains which are defined as ratios between the outputs and inputs in Laplace space of these two equations are respectively:

$$\begin{aligned} G(s) &= \frac{X(s)}{U(s)} = \frac{1}{ms^2 + bs + k} \\ H(s) &= \frac{Y(s)}{X(s)} = p \end{aligned} \tag{3.78}$$

In telescope control, the sensor measurement provides a feedback for the control system and the control loop is closed as shown in Figure 3.32, where $R(s)$ is the input signal, $C(s)$ the output, $G(s)$ the gain of the drive loop, $H(s)$ the gain of the feedback loop, and $\delta(s)$ the input and feedback signal difference. For a telescope, the input is a required angular position, the output is the real telescope pointing, and the feedback is from the angular encoder measurement. The system transfer function of this feedback control is:

$$M(s) = \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} \tag{3.79}$$

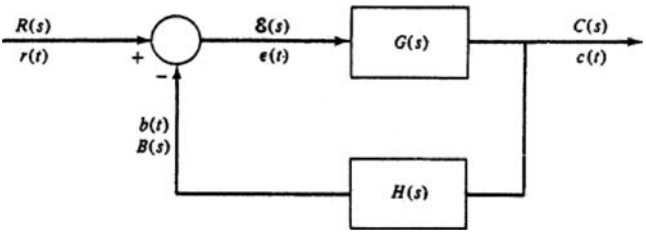


Fig. 3.32. A typical feedback close loop control.

The differential equation of a mechanical, hydraulic, or other system is only a part of the drive system or the controller. Fortunately, this part can be perfectly replaced by an electric network analogy and be combined with control electronics in a mathematical form. The controller's behavior can be adjusted by change of electronics of the system according to system requirements. Generally, three different controllers, proportional, proportional integral (PI), and proportional integral derivative (PID) ones, are used in classical analog feedback control.

The proportional controller makes changes to output proportional to the feedback error. The PI controller makes changes to output proportional to both the magnitude and duration of the error. The PID controller makes changes to output proportional to not only the magnitude and duration, but also the rate of change of the error. The PID control law is:

$$u = K_p e + k_i \int e dt + K_d \frac{de}{dt} \tag{3.80}$$

where u is the corrective command signal, e the error signal, and K_p, K_i, and K_d the proportional, integral, and derivative gains, respectively. The traditional PI control has limited bandwidth since it is slow in response when the gain is small and is unstable when the gain is large. The gain is also tied to the lagging (a constant servo error) in response. Smaller gain produces a larger lagging. It corrects low frequency error, but not the high frequency one. The PID controller avoids saturation of the integral and overshooting when the error is larger. It produces a faster response and wider control bandwidth.

The use of computers brings two major changes in feedback control. First, the control data can be in discrete form instead of continuous. Second, the function of some electronic hardware can be replaced by computer calculations. Digital control can perform exactly the same as or better than an analog one. Modern telescope feedback control is now more complex and more accurate. It is a digital cascade system, generally involving three level loops: acceleration (current), velocity (rate), and position ones.

Generally, the telescope current loop is integrated with a servo amplifier. The velocity one is a PID one as a pair of motors is involved for dampening the unwanted anti-resonance mode in the drive system. It acts as a low pass filter. The velocity controller should also include an acceleration limit in its step response. The position controller has a direct impact on telescope pointing accuracy. It is usually a PID one or some type of combination controller with signal feedforward. In a combined controller, when the position error is small, a proportional control is used as it provides rejection on turbulence. A motion profiler is added to feedforward the future position or/and velocity commands for avoiding resonance in lower frequency ranges.

In the control system, the position signals are supplied by encoders, the velocity signals by tachometers, and the torque signals through the voltage measurement on a resistor cascaded in motor circuits. The continuous analog

signals are converted to digital ones and they are compared with the instruction to produce an error signal. The error signal is put into a reversible counter for error calibration through a digital-to-analog converter. The challenge of this system is the balance between the transient response and the steady-state behavior.

To further improve the accuracy of the drive system, modern telescopes use the following methods: (a) increasing the sampling frequency; (b) using a maximum or minimum value control, such as used in star guiding discussed in the next section; (c) using a state space controller to allow more inputs into the control system, and (d) using adaptive or dynamic control such as the Kalman filter which continuously updates system gain to achieve the best telescope performance. A high sampling frequency can reduce the error caused by time delay. By using maximum or minimum value control, the disadvantage of classical control where merit function does not change quickly is removed.

The state space controller is based on mathematical modeling of all subsystems, including frictional force, temperature, and motor stiffness. The advantage is that all the relevant information measured is utilized in the state and output equations:

$$\begin{aligned} \dot{x}(t) =& Ax(t) + Bu(t) + Ew(t) \\ y(t) =& Cx(t) + Du(t) + Fw(t) \end{aligned} \tag{3.81}$$

where $x(t)$ is the state variable, $u(t)$ the input, $w(t)$ the noise, and $y(t)$ the measurement output. By using the state space method, the previous spring mass damper system can be expressed as:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} x =& \begin{bmatrix} 0 & 1 \\ -\dfrac{k}{m} & -\dfrac{b}{m} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{m} \end{bmatrix} u \\ y =& \begin{bmatrix} p & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned} \tag{3.82}$$

where $x_1 = x(t)$, $x_2 = \dot{x}(t)$, m is the mass, k the stiffness, b the damping, and p is a constant for sensor. The three coefficient matrixes on the right hand sides of the equations are called the ABC matrixes (Section 3.4.3). However, if the close-loop information is from the axial encoders, not from a real star, this is still an open-loop pointing system. To close the loop, a guiding star is required. With star guiding, the pointing error reduces significantly.

The adaptive or dynamic control using a linear Kalman filter (Crassidis and Junkins, 2004) has been used in a number of telescopes. The Kalman filter is an efficient recursive genetic filter that estimates the state of a dynamic system from a series of incomplete and noisy measurements and evolves to improve its performance over iterations. The Kalman filter used in a state space equation is most commonly referred to as a linear quadratic estimation (LQE). The

feedback controller using a Kalman filter is referred to as a linear quadratic regulator (LQR). The cost function of LQE is the square of the residual error. This least square method was originally proposed by Carl Gauss. Therefore, the combination of LQE and LQR is now an optimum liner quadratic Gaussian (LQG) control system. In this LQG system, the control loop gain is carefully updated according to a mathematical model of the system and the sequential state estimation from the measurements. The deviation from the required profile will be very small and the system time constant is also small. The Kalman filter control method is a very powerful one in adaptive control. Wodek Gawronski (2007) carried out an extensive study of a LQG system applied to very large outdoor antennas. Using the LQG controller in velocity and position loops, the pointing performance under wind turbulence is hundreds of times better than a system using classical PI controllers in both velocity and position loops (Gawronski and Souccar, 2003).

In modern telescopes, more and more functions are fulfilled by computers. Therefore, a distributed control system is usually used for avoiding frequent failure caused by one component of the system. A distributed control system utilizes many microprocessors or computers. They are connected with the main control computer through interfaces. The microprocessors used are near to the components they control, such as encoders, motors, tachometers, and other devices. The circuits are short and the system is simple. This distributed control system realizes a mechanical-electrical integration. When some parts disassemble, they can be tested independently. One microprocessor needs only two sets of cables: one for the power supply; and the other for transmitting data. In this system, microprocessors communicate with each other through the main control computer. Because microprocessors accomplish all local control, the main control computer will be used for sending commands, collecting information, and processing data.

Now remote observation is possible for many astronomical telescopes. Astronomers can operate the telescopes through the internet, despite being several thousands of kilometers away from the site. The remote control reduces travel, improves efficiency, and cuts the cost of telescope operation.

3.3.6 Star Guiding

The control systems discussed in the previous section can only achieve a pointing accuracy inherently limited by the encoder resolution if a guiding star is not used. Other errors from optical (distortion), mechanical (optical, boresight, encoder alignment), thermal (alignment change), and other origins (atmospheric diffraction) also produce blind pointing errors before pointing correction can be made. Almost all repeatable errors are compensated after the pointing correction so that the residual blind pointing error is the encoder error plus nonrepeatable pointing errors. Compared with real star position, the control system is still open loop and its accuracy is limited.

The residual pointing error which produces a slight drift of the target star within a period is tracking error. If the drift is smaller than FWHP of the point spread function, the image degradation is insignificant for a short period. However, the required tracking accuracy for long or multi- exposures is very high. Fortunately, the stars in the field of view provide a final check of the telescope pointing, so that the control loop can be closed during tracking. The star(s) used for this pointing check is called the guide star. The tracking with a guide star is called star guiding and the automatic image centroiding device is called the guider.

During star tracking, information from the guider is fed back to either a main control loop or an additional loop for fine steering a small tip-tilt mirror to achieve a much higher pointing accuracy. Using a fine steering tip-tilt mirror has advantages of fast response and wider bandwidth so that high frequency atmospheric and wind errors are corrected. Tip-tilt mirrors are discussed in Section 4.1.5.

The pointing accuracy achieved without star guiding is about 0.8–0.5 arcsec. It will be 0.1–0.02 arcsec when star guiding is used for ground-based telescopes and about 0.004 arcsec (in 1,000 s duration time) for space telescopes. In the future, space telescopes may require a 0.1-marcsec pointing accuracy. This can only be achieved through a special star guider as the gyroscope would drift at a rate of 1–1.5 μarcsec/s.

Star guiding is not new. Early star guiding was performed visually through a small guiding telescope mounted on the side of a telescope where the photographic work was performed. The pointing correction was adjusted manually by hand. This type of star guiding has a differential line of sight error. Guiding using a star in the field of view became possible after the invention of electronic detectors. After that, the tracking accuracy was greatly improved. Guiding using a single star may have field rotation around the guided star. To avoid this error, one additional star sufficiently far away from the first one is used. This device is also called a derotator which produces a stable image within the field of view. A derotator can be kept off if the guide star is the same as the science target or the derotation of the image is fulfilled by software after the observation.

A star guider is a continuous balance device. In the past, a pyramid reflector with four photoelectric detectors placed symmetrically around (Figure 3.33) was

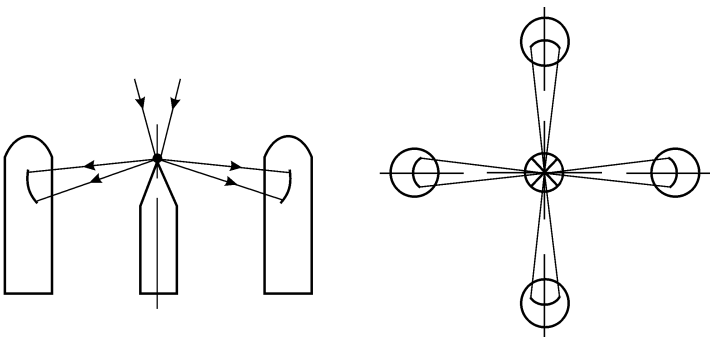


Fig. 3.33. Early continuously balanced star guiding with a pyramid reflector.

used. If starlight illuminates the center of the pyramid reflector, the light intensity received by each detector is identical. Once the star deviates from the center, the light intensities received by detectors will vary. The centroid position error is used for the pointing correction. The problem of this device is the response difference between four photoelectric detectors.

Other early star guiders include the semicircle disk flux modulation device and quadrant detector. The semicircle disk device comprises a Fabry lens and a semicircle disk rotating uniformly around the optical axis. A Fabry lens is a lens which forms an image of the telescope primary mirror. When starlight is on axis, the light energy modulated by the semicircle disk remains constant without alternative components. When the star is away from the axis, alternative components appear and the amplitude and phase provide the information of the offset. The quadrant detector is still in use today and its guiding principle is the same as the pyramid reflector shown in Figure 3.34. When starlight passes through the center of the sensor, the output current waveform is shown in the upper line. When the star drifts away from the center, the waveform is shown in the bottom line. The x and y centroid positions of starlight are:

$$\begin{aligned} I_x &= I_1 + I_2 - I_3 - I_4 \\ I_y &= I_1 - I_2 - I_3 + I_4 \end{aligned} \tag{3.83}$$

Today, star guiding is usually performed by CCD detectors. If four pixels are involved, the calculation of the centroid is the same as the quadrant detector. If more pixels are involved, the center of gravity of the image is calculated mathematically. The accuracy of these guiding devices can reach one order smaller than the pixel size. Therefore multiple- and extremely long exposures become possible for modern optical telescopes. The software tools for the image analysis used include the Image Reduction and Analysis Facility (IRAF) developed by the National Optical Astronomy Observatory (NOAO).

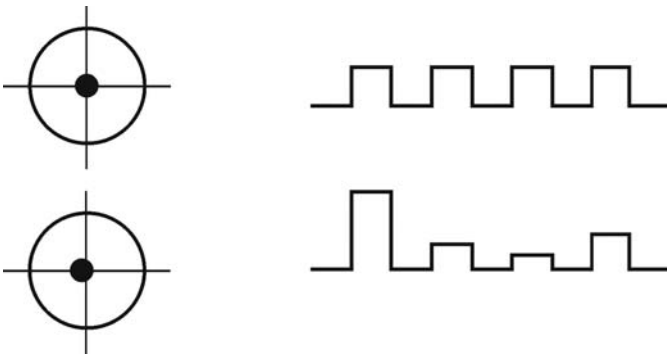


Fig. 3.34. The current output waveforms of a quadrant detector.

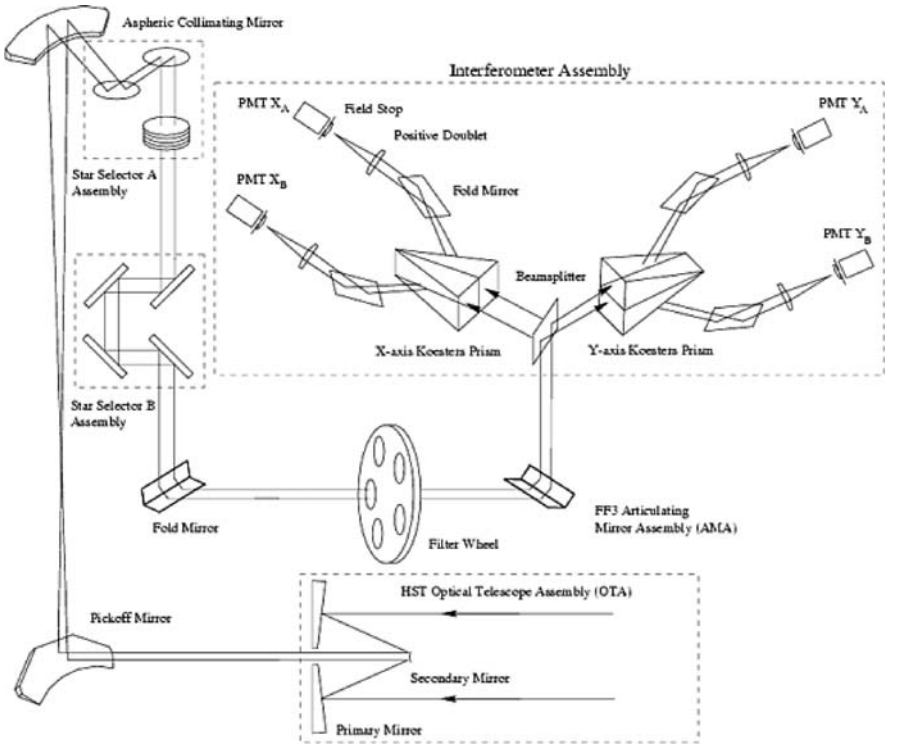


Fig. 3.35. The layout of the fine guidance sensor of the HST (STSCI).

The most accurate star guiding instrument is a type of shearing interferometer wavefront slope sensor, which has a *mas* accuracy. One of these is the HST fine guidance sensor (Figure 3.35). The sensor consists of a polarizing beam splitter followed by two Koesters prisms. The polarizing beam splitter divides the incoming collimated star light into two plane polarized beams. The splitter then directs each beam to a Koesters prism and its associated optics. The Koesters prisms include two halves of fused silica joined along a coated dielectric beam splitter. The dielectric layer divides a beam into two equal intensity parts, imparting a 90° phase lag in the transmitted beam. This division and phase shift gives the Koesters prism its interferometric properties: the beam reflected from one side of the prism interferes constructively or destructively with the beam transmitted from the other side. The degree of interference between the two beams is directly related to the wavefront tilt of the incoming wavefront relative to the dielectric surface (Hu, 2007). If the two beams have a quarter wavelength difference, one output has a maximum intensity and the other zero intensity.

Because the celestial target is faint, a conventional TV cannot meet the requirement for star guiding and a digital integration TV has to be used. The operation is as follows: after analog-to-digital conversion, a digital signal is

stored in memory. The signal is then integrated for improving its signal-to-noise ratio. After digital-to-analog conversion, the stored signal is mixed with a synchronous signal and input back to the TV monitor for display. Because the integration time is longer than the frame duration, a faint star image can be displayed brightly in the screen. In a digital TV, the integration time can be adjusted to meet the requirement for display. Inside the memory, the new and old image signals provide the correction of the pointing error.

3.4 Structural Dynamic Analysis

3.4.1 Wind and Earthquake Spectrums

3.4.1.1 Random Property of Wind

Wind as a natural random phenomenon caused by the flow of air includes two parameters: its direction and velocity. Generally, wind velocity is a sum of two components: a constant mean term, V_m, and a variable random term, $v(t)$:

$$V(t) = V_m + v(t) \tag{3.84}$$

The mean wind velocity is a function of the height above the ground level z and the ground roughness z_0:

$$V_m(z) = V(z_{ref}) \ln(z/z_0) / \ln(z_{ref}/z_0) \tag{3.85}$$

where $V(z_{ref})$ is the wind velocity at a reference height of 10 m above the ground level. In Table 3.2, different ground roughness number is listed.

As a random variable, the wind velocity can be expressed by its power spectrum. The power spectrum of wind, $S(f)$, gives the distribution of the wind energy along the frequency axis. Integration of the power spectrum gives the variance of the wind velocity, σ^2. The square root of the variance is the standard deviation or the rms deviation. If a random variable follows a Gaussian distribution, its probability between two velocities V_i and V_m is:

$$P(V_i - V_m) = \frac{1}{\sigma_v \sqrt{2\pi}} \exp\left[-\frac{(V_i - V_m)^2}{2\sigma_v^2} \right] \tag{3.86}$$

Table 3.2. Ground roughness of different areas

	Open field	Agricultural area	Village
Roughness	0.01	0.05	0.3

The probability between $-\sigma$ and σ is 68.3%, the probability between $-2\ \sigma$ and $2\ \sigma$ is 95.4%, and that between $-3\ \sigma$ and $3\ \sigma$ is 99.7%.

There are different spectra which may fit to the natural wind. Commonly used is Davenport's power spectrum which is low in the high frequency regime. Simiu (1974) corrected the high frequency discrepancy and gave a Simiu power spectrum (Figure 3.36):

$$S(f) = (V_*^2/f)(200n/(1+50n)^{5/3}) \tag{3.87}$$

where $V_* = V(z)/2.5\ln(z/z_0)$ is the wind shear velocity, $n = fz/V(z)$, and f the frequency. Note that the power spectrum is proportional to the square of the wind velocity. Another expression of this power spectrum is:

$$\frac{fS(f)}{V_*^2} = \frac{200\,n}{(1+50n)^{5/3}} \tag{3.88}$$

Shear velocity of the wind is related to the variance:

$$6V_*^2 = \sigma_v^2 = \int_0^\infty S(f)df \tag{3.89}$$

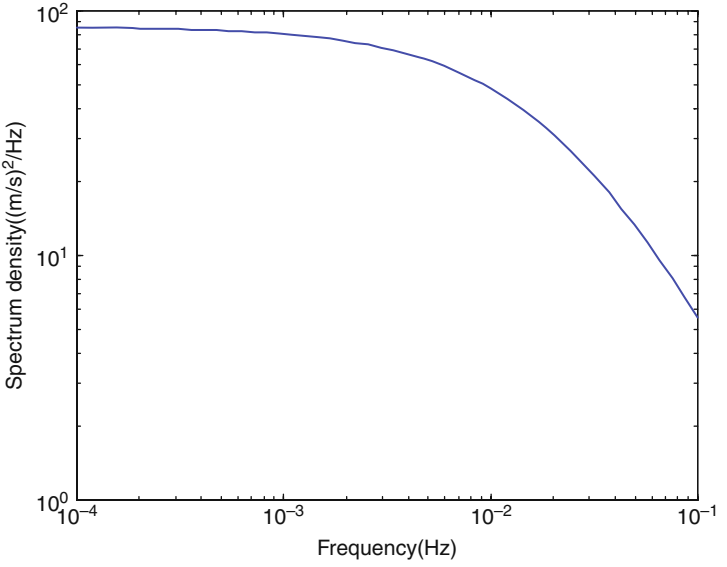


Fig. 3.36. The Simiu spectrum of wind.

The wind head pressure can be expressed as:

$$P = \frac{1}{2}\rho V^2 \tag{3.90}$$

where ρ is the air density. Note that the air density is a function of the height above sea level. For example, the 10 m/s wind head pressure is 61 N/m^2 at sea level and is 38 N/m^2 at a 5,000 m altitude.

3.4.1.2 Wind Loading on Structures

Wind loading on an object is a function of the wind head pressure P, the cross sectional area A, and the shape of the object. The expression is:

$$F = C_D PA = \frac{1}{2} C_D A \rho V^2 \tag{3.91}$$

where C_D is the drag coefficient which describes the shape or the dynamic performance of an object. When both constant and random wind velocities are considered in structural analysis, two methods can be used: (a) using an equivalent wind velocity; and (b) using the power spectrum to derive the random wind effect and then add it to the effect from the constant wind part.

Since the wind follows a Gaussian distribution, the random part of the wind velocity can be expressed as:

$$F(v) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(v - V_m)^2}{2\sigma_v^2}] \tag{3.92}$$

where V_m and σ_v are the mean and the standard deviation of the wind velocity. They are expressed as:

$$\begin{aligned} V_m &= \int_{-\infty}^{\infty} v f(v) dv \\ \sigma_v^2 &= \int_{-\infty}^{\infty} (v - V_m)^2 f(v) dv \end{aligned} \tag{3.93}$$

The wind loading is proportional to the wind head pressure, so it has a form of Kv^2. The mean and the standard deviation of the wind loadings are:

$$\begin{aligned} P_m &= \int_{-\infty}^{\infty} Kv^2 f(v) dv = K[V_m^2 + \sigma_v^2] \\ \sigma_p^2 &= \int_{-\infty}^{\infty} (Kv^2 - P_m)^2 f(v) dv = K^2 \sigma_v^2 [4V_m^2 + 2\sigma_v^2] \end{aligned} \tag{3.94}$$

Generally, the wind loading including the static and the variable parts is described as an rms value which is the sum of the mean wind loading and the standard deviation of the random part. This is:

$$P_{rms} = K[V_m^4 + 6V_m^2\sigma_v^2 + 3\sigma_v^4]^{1/2} \tag{3.95}$$

Therefore, the equivalent wind velocity is:

$$V_{equ} = [V_m^4 + 6V_m^2\sigma_v^2 + 3\sigma_v^4]^{1/4} \tag{3.96}$$

The equivalent wind velocity is only a little larger than the mean wind velocity. Generally, the mean square deviation of a sum of two random variables is smaller than the sum of the mean square deviations of two random variables. For a 9 m/s mean wind velocity, if the wind velocity within 90% of the time is below 11 m/s, the standard deviation is about 1.28 m/s from the Gaussian distribution table. The equivalent wind velocity is about 9.385 m/s. If the standard deviation is calculated from friction velocity of the wind, it will be about 1.632 m/s. The equivalent wind velocity is 9.422 m/s.

Another way in finding the random part of the wind loading is (Figure 3.37):

$$\begin{aligned} F(t) &= \frac{1}{2}\rho A C_D (V_m + v(t))^2 \\ &= \frac{1}{2}\rho A C_D V_m^2 + \rho A C_D V_m v(t) + \frac{1}{2}\rho A C_D v(t)^2 \end{aligned} \tag{3.97}$$

The first term of the equation is the static wind loading and the rest are the dynamic loading part. The third term is small and may be neglected. For calculating the second term, a concept of loading spectral function is necessary.

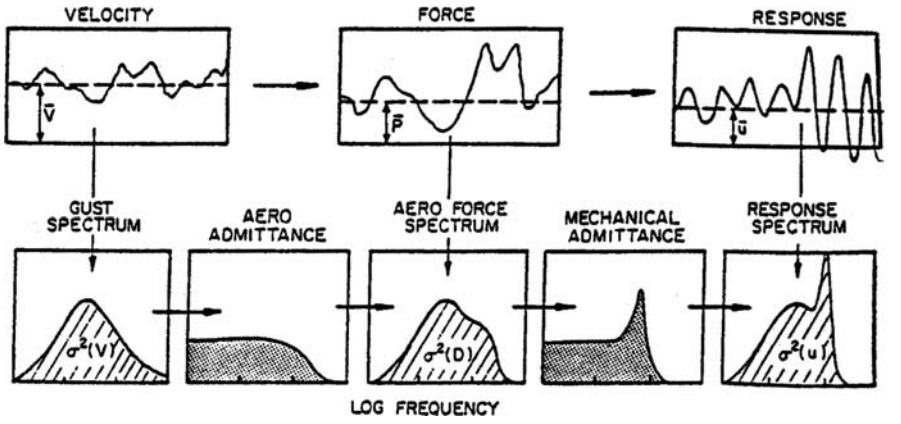


Fig. 3.37. Procedure to calculate the dynamic reaction of wind on a structure.

Using the second term, the loading spectrum is represented by a velocity power spectrum:

$$S_D(f) = (\rho A C_D V_m)^2 S(f) \tag{3.98}$$

where $S(f)$ is the wind velocity power spectrum. For large structures with its dimension approximately the same as the wavelength of wind, $fA^{1/2}/V_m \approx 1$, an aerodynamic admittance term $|X_{aero}(f)|^2$ should be added in the above expression:

$$S_D(f) = |X_{aero}(f)|^2 (\rho A C_D V_m)^2 S(f) \tag{3.99}$$

In most cases, the admittance term is approximately unity. Using the FEA analysis, the structural response spectrum $S_R(f)$ can be directly derived from the wind loading spectrum $S_D(f)$ as:

$$S_R(f) = [H]^2 S_D(f) \tag{3.100}$$

where $[H]$ is the transfer matrix. After the response spectrum is derived, the variable wind loading on the structure can be derived by the integration of the spectrum over the frequency range. The dynamic effect together with the static wind effect is then the total wind effect on the structure.

Using this process, since the response is frequency related, the residual structural response can be derived if part of the response in some frequency range is compensated by the control system. The topic on how to derive the response spectrum $S_R(f)$ from the force spectrum $S_D(f)$ will be discussed in the next section.

3.4.1.3 Vortex Shedding Resonance

Vortex-shedding is another wind effect on structures. Vortex-shedding is caused by alternating low pressure zones generated on the downwind side of a long slender beam. The vortex shedding frequency excited is proportional to the wind velocity and the Strouhal number, but is inversely proportional to the beam diameter. Over a wide range of the Reynolds number, the Strouhal number measured is about 0.12 (Sarioglu and Yavuz, 2000). The resonance of a slender beam occurs when the following formula exists:

$$V_{cr} \approx 0.447 \Lambda (d/L)^2 \tag{3.101}$$

where V_{cr} is the wind velocity in m/s, d the diameter, L the length of the bar in m, and Λ is a constant determined by the constraint conditions shown in Table 3.3.

Table 3.3. The support conditions and the constant used in Equation (3.100)

Support condition	One end fixed	Both ends simple supported	One end fixed and other end simple supported	Both ends fixed
Λ	10,600	29,800	46,500	67,400

3.4.1.4 Wind Pressure Distribution on a Mirror Surface

A normalized wind pressure distribution on a mirror surface is shown in Figure 3.38, where the elevation angle is 90° and 45°, respectively. When the wind pressure on the mirror surface is used, attenuation of the dome should also be considered. The ratio of wind pressure reduction on the mirror from a dome is about 1/8 on average. The dome can cause resonance at certain wind frequencies. The dome can also shift the wind power from lower to higher frequencies. All these should be considered in the wind analysis.

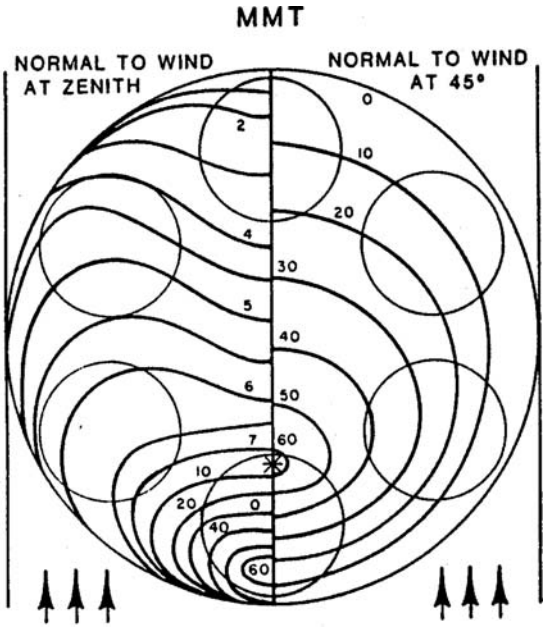


Fig. 3.38. Wind pressure distribution over the mirror surface at the altitude angles of 90° and 45° (Forbes and Gaber, 1982).

3.4.1.5 Earthquake Response Spectrum

An earthquake is a random motion of the ground. It is typically represented by its acceleration response spectrum. This spectrum includes two in the horizontal directions and one in the vertical direction. The earthquake acceleration response spectrum is different from the ground motion spectrum. The acceleration response spectrum is generated by a set of one degree of freedom mass spring systems with different resonant frequencies and certain damping ratio under the earthquake condition. The response spectrum can be acceleration, velocity, or displacement ones. Generally, the ground acceleration is amplified in the response spectrum by a factor of three within 2–10 Hz range when the damping ratio is small (Figure 3.39). A common value of damping ratio of a response spectrum is 1%. The ground acceleration in Figure 3.39 is about 0.3 g.

A magnitude is used to represent the intensity of the earthquake. The formula of a Richter scale magnitude M is (Richter, 1958):

$$\log_{10} E_f = 11.4 + 1.5\,M \tag{3.102}$$

where E_f is the energy in Joule released from the earthquake. Generally, an earthquake of magnitude 5 on a Richter scale will cause serious damage (Table 3.4). The earthquake magnitude is also related to the ground acceleration. The higher the earthquake magnitude is, the higher the ground acceleration will be. The time interval is related to response frequency.

The effect of the earthquake on a structure is represented by its response spectrum. When the frequency response analysis is made, a square root of the square sum of the response at each resonant mode is used as the structural response under the earthquake condition.

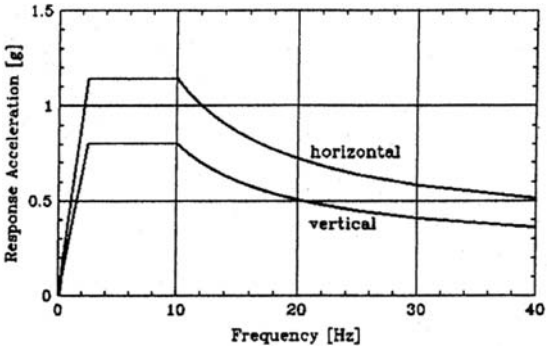


Fig. 3.39. A typical earthquake response spectrum.

Table 3.4. Relationship between maximum ground acceleration and earthquake magnitude

Earthquake magnitude	Maximum ground acceleration	Time interval (seconds)
5.0	0.09	2
5.5	0.15	6
6.0	0.22	12
6.5	0.29	18
7.0	0.37	24
7.5	0.45	30
8.0	0.50	34
8.5	0.50	37

3.4.2 Dynamic Simulation of Telescope Structures

A dynamic simulation, which provides the telescope response to the control system and to the turbulences, is important in the design study.

3.4.2.1 Modal Analysis

A modal analysis, which provides the resonance frequencies of a structure and their vibrational shape, is used for the estimation of the controllability of a structure. It is a necessary step before the structural transient simulation and frequency response analysis.

Considering a typical dynamic equation of a structure:

$$[M]\{\ddot{u}\} + [K]\{u\} = 0 \tag{3.103}$$

where $[M]$ is the mass matrix, $[K]$ the stiffness matrix, and $\{u\}$ the displacement matrix. The solution of the equation is:

$$\{u\} = \{\phi\}e^{i\omega t} \tag{3.104}$$

By replacing the solution into the dynamic equation:

$$([K] - \omega^2[M])\{\varphi\} = 0 \tag{3.105}$$

it becomes an eigenvalue equation. The values ω and $\{\varphi\}$ corresponding to its nontrivial solutions are the circular resonance frequency and its related modal shape. The circular resonance frequency is 2π-times the resonance frequency in Hz. A structure with n mass attached degrees of freedom will have n resonance frequencies. When a structure vibrates, the structural shape is a linear combination of all the modal shapes.

3.4.2.2 Transient Analysis

A transient response analysis can be done either iteratively or in the modal space. The iterative method solves the following equation:

$$[M]\{\ddot{u}(t)\} + [B]\{\dot{u}(t)\} + [K]\{u(t)\} = \{P(t)\} \tag{3.106}$$

where $[B]$ and $\{P(t)\}$ are the damping and load matrixes. Using the initial conditions, the equation can be solved by replacing the acceleration and velocity with displacements at different time intervals as:

$$\begin{aligned} \{\dot{u}_n\} &= \frac{1}{2\Delta t}\{u_{n+1} - u_{n-1}\} \\ \{\ddot{u}_n\} &= \frac{1}{\Delta t^2}\{u_{n+1} - 2u_n + u_{n-1}\} \end{aligned} \tag{3.107}$$

Substituting the above equations into the differential equations, one obtains:

$$\begin{aligned} [A_1]\{u_{n+1}\} &= [A_2] + [A_3]\{u_n\} + [A_4]\{u_{n-1}\} \\ [A_1] &= \left[M/\Delta t^2 + B/2\Delta t + K/3\right] \\ [A_2] &= 1/3\{P_{n+1} + P_n + P_{n-1}\} \\ [A_3] &= \left[2\,M/\Delta t^2 - K/3\right] \\ [A_4] &= \left[-M/\Delta t^2 + B/\Delta t - K/3\right] \end{aligned} \tag{3.108}$$

The problem can be solved. By using the modal space approach, the physical coordinate u is transformed into a modal coordinate ξ as:

$$\{u\} = [\phi]\{\xi\} \tag{3.109}$$

where $[\phi]$ is the modal shape matrix. Substituting the modal coordinate into the dynamic equation, then:

$$[M][\phi]\left\{\ddot{\xi}\right\} + [K][\phi]\{\xi\} = \{P(t)\} \tag{3.110}$$

Multiply each term by $[\varphi^T]$:

$$\left[\phi^T\right][M][\phi]\left\{\ddot{\xi}\right\} + \left[\phi^T\right][K][\phi]\{\xi\} = \left[\phi^T\right]\{P(t)\} \tag{3.111}$$

The resulting matrixes of the equation are all diagonal ones. The new stiffness matrix equals the square of the circular frequency. The above equations are now a set of independent second order equations:

$$\left\{\ddot{\xi}\right\} + [\Omega]\{\xi\} = [\phi]^T\{p_i(t)\} \tag{3.112}$$

A new damping matrix is a product of the physical damping and model shape matrix $[\phi]$:

$$[\phi]^T[B][\phi] \tag{3.113}$$

The dynamic equation including the damping is:

$$\ddot{\xi} + (b/m)\dot{\xi} + \omega^2\xi = p(t)/m \tag{3.114}$$

The solution of this equation is:

$$\begin{aligned}\xi(t) = e^{-bt/2m}&\left[\xi_0\cos\omega t + \frac{\dot{\xi}_0 + (b/2m)\xi_0}{\omega}\right] \\ &+ e^{-bt/2m}\frac{1}{m\omega}\int_0^t e^{-b\tau/2m}p(\tau)\sin\omega(t-\tau)d\tau\end{aligned} \tag{3.115}$$

Usually, the first term equals zero without the initial displacement condition. The solution can be derived by transforming the displacements in mode space into the physical displacements.

3.4.2.3 Frequency Response Analysis

The frequency response also named as harmonic response is performed in the frequency domain. Any continuous cyclic load will cause a sustained cyclic response in a structural system. Harmonic response analysis allows one to predict the sustained dynamic behavior of a structure. One can verify whether or not the structure can overcome resonance, fatigue, or other effects of forced vibrations (Koch, 2008).

In the frequency response analysis, the input loading is over all the frequency range. The response calculated can be the displacements of some nodes or the stresses of some elements. The response includes both the amplitude and phase.

The frequency response can be derived in physical or modal space. The basic equations are:

$$[-\omega^2 M + i\omega B + K]\{u(\omega)\} = \{P(\omega)\} \tag{3.116}$$

In the physical space, the calculation is the same as the transient analysis. Using a modal space transformation, the displacement in modal space, ξ, is independent of each other and the solution is:

$$\xi = \frac{p}{-\omega^2 m + i\omega b + k} \tag{3.117}$$

Figure 3.40 shows a flat plate excited by a frequency-varying loading. The plate response is picked up by a sensor. As the loading frequency increases, the response approaches the maximum values in four distinct frequencies. These are resonance frequencies of the plate. To get exact response amplitude, smaller frequency interval should be used near the resonant frequencies (Figure 3.41). In general, at least five samplings should be used within the FWHM of the resonant frequency peak point.

3.4.2.4 Forced Vibration Analysis

The earthquake induced vibration is a typical case of the forced vibration. To simulate this, the most effective method is to add a giant mass to the foundation of the structure. This mass can be a million times greater than that of the structure. Then, a dynamic force is applied on this mass point. The force applied

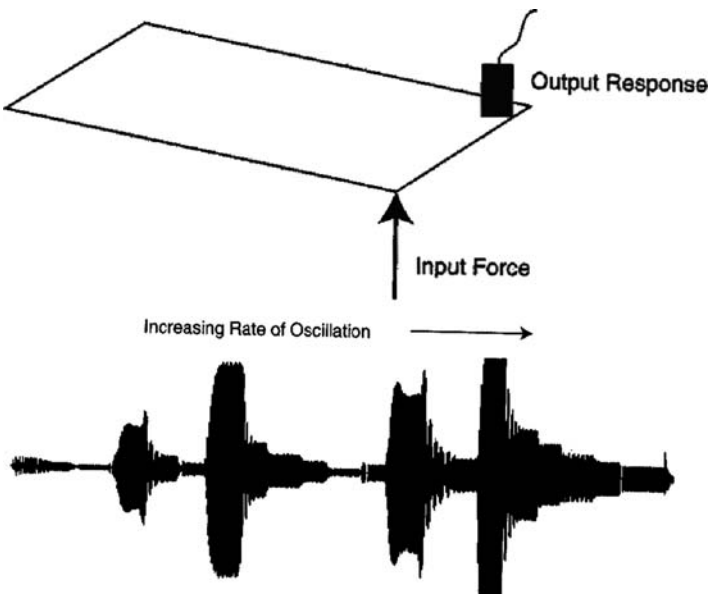


Fig. 3.40. A test device for frequency response analysis (Avitabile, 2001).

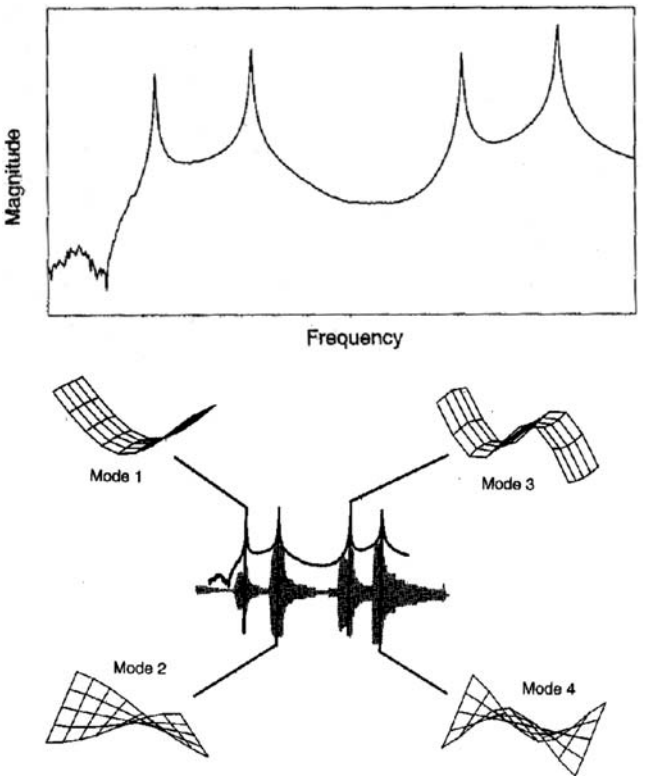


Fig. 3.41. Relationship between response frequency and modal shape (Avitabile, 2001).

can be in the form of acceleration, or velocity, or displacement. If acceleration is applied, the nodal acceleration should be:

$$\ddot{u}_b = \frac{1}{M_L} P \tag{3.118}$$

where M_L is the giant mass and P is the acceleration applied. When a velocity function F is applied, then P equals:

$$\begin{aligned} P &= (F_N - F_{N-1})/\Delta t \quad t_N \neq 0 \\ P &= 0 \qquad\qquad\qquad\quad t_N = 0 \end{aligned} \tag{3.119}$$

When a displacement function F is applied, then P equals:

$$\begin{aligned} P &= \frac{2}{\Delta t_1 + \Delta t_2}\left(\frac{F_N - F_{N-1}}{\Delta t_2} - \frac{F_{N-1} - F_{N-2}}{\Delta t_1}\right) \quad t_N \neq 0 \\ P &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad t_N = 0 \end{aligned} \tag{3.120}$$

3.4.2.5 Spectrum Response Analysis

Spectrum response analysis is used to analyze the effect of a random loading on a structure. The loading can be force, acceleration, or displacement. To express a random function, the auto-correlation function or the power spectrum can be used. They are a Fourier pair. The auto-correlation function is defined as:

$$R(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T u(t)u(t-\tau)dt \tag{3.121}$$

The power spectrum is:

$$S(\omega) = \lim_{T\to\infty} \frac{2}{T}\left|\int_0^T u(t)e^{-i\omega t}dt\right|^2 \tag{3.122}$$

The rms value of a random function is:

$$\bar{u}^2 = R(0) = \frac{1}{2\pi}\int_0^\infty S(\omega)d\omega \tag{3.123}$$

From frequency response analysis, an input loading $F(\omega)$ on a structure will produce a response $u(\omega)$ as:

$$u(\omega) = H(\omega)F(\omega) \tag{3.124}$$

where $H(\omega)$ is the system transfer function. If there are several input loadings, then:

$$u(\omega) = H_a(\omega)F_a(\omega) + H_b(\omega)F_b(\omega) + \cdots \tag{3.125}$$

Using matrix expression, it is:

$$u(\omega) = [H_a(\omega)H_b(\omega)\cdots]\begin{bmatrix} F_a(\omega) \\ F_b(\omega) \\ \cdots \end{bmatrix} \tag{3.126}$$

In many situations, the response can be represented by Fourier transform of its auto-correlation or the power spectrum:

$$S_{uu} = [H_a(\omega)H_b(\omega)\cdots]\begin{bmatrix} F_a(\omega) \\ F_b(\omega) \\ \cdots \end{bmatrix}[F_a^*(\omega)F_b^*(\omega)\cdots]\begin{bmatrix} H_a^*(\omega) \\ H_b^*(\omega) \\ \cdots \end{bmatrix} \tag{3.127}$$

If input loadings are represented by their power spectra,

$$\begin{aligned} S_{aa} &= [F_a(\omega)][F_a^*(\omega)] \\ S_{ab} &= [F_a(\omega)][F_b^*(\omega)] \\ S_{bb} &= [F_b(\omega)][F_b^*(\omega)] \end{aligned} \tag{3.128}$$

then the response power spectrum is:

$$S_{uu} = [H_a(\omega)H_b(\omega)\cdots] \begin{bmatrix} S_{aa} & S_{ab} & \cdot \\ S_{ba} & S_{bb} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} H_a^*(\omega) \\ H_b^*(\omega) \\ \cdots \end{bmatrix} \tag{3.129}$$

In the above formula:

$$\begin{aligned} &S_{ab} = S_{ba}^* \\ &S_{aa}, S_{bb} = real \geq 0 \end{aligned} \tag{3.130}$$

3.4.3 Combined Structural and Control Simulation

Combined structural and control simulation is based on the state space equations. The structural state space equations come from structure dynamic equations with the external load, Bu, applied. The output equations give sensor readings. They are:

$$\begin{aligned} &M\ddot{q} + Kq = Bu \\ &y = C_{oq}q + C_{ov}\dot{q} \end{aligned} \tag{3.131}$$

Transforming the physical coordinates into the modal coordinate and adding the damping term, the equations become:

$$\begin{aligned} &\ddot{q}_m + 2Z\Omega\dot{q}_m + \Omega^2 q_m = M_m^{-1}\Phi^T Bu \\ &y = C_{oq}\Phi q_m + C_{ov}\Phi\dot{q}_m \end{aligned} \tag{3.132}$$

where Ω is the resonance frequency matrix, Φ the modal shape matrix, M_m the modal mass matrix, and $2Z\Omega/M_m{}^{-1}$ the modal damping matrix. If state variables defined are $x_1 = q_m$ and $x_2 = \dot{q}_m$, the state equations become:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\Omega^2 x_1 - 2Z\Omega x_2 + M_m^{-1}\Phi Bu \\ y &= C_{oq}\Phi x_1 + C_{ov}\Phi x_2 \end{aligned} \tag{3.133}$$

Normally, the state space equations can be expressed by using the ABC matrix as:

$$A = \begin{bmatrix} 0 & I \\ -\Omega^2 & -2Z\Omega \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ M_m^{-1}\Phi^T B \end{bmatrix}$$
$$C = [C_{oq}\Phi, C_{ov}\Phi] \tag{3.134}$$

To form the state space equations, the relationship between the input and the output is very important. The control system itself is represented by a number of state space equations between the input and output of the system (Section 3.3.5). The telescope performance, therefore, can be predicted by solving these two sets of the state space equations. During the simulation, the wind or friction forces can be included in the calculation.

3.4.4 Structure Vibration Control

Structure vibration control is extremely important for large telescopes. In general, a tuned mass damper, or a viscoelastic layer, or a motion profiler can be used in this aspect.

3.4.4.1 Tuned Mass Dampers

If a small mass-spring system is added to another base mass-spring system with the same resonance frequency, the combined system will have two resonance frequencies which are located on both sides of the original one (Figure 3.42). In

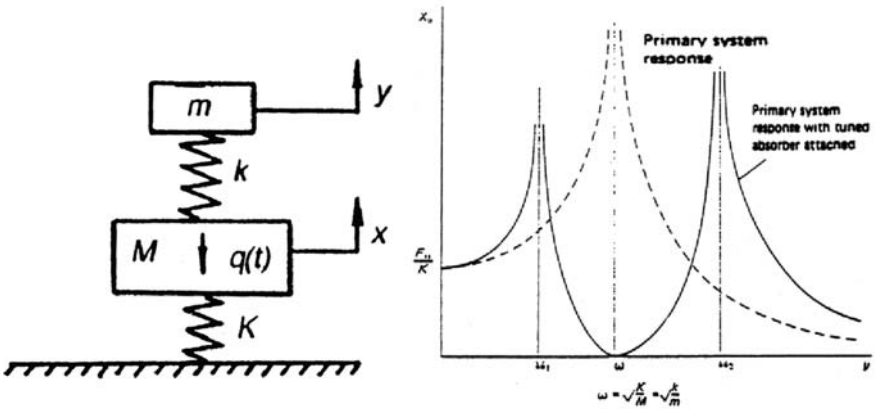


Fig. 3.42. A two-mass and two-spring system and its frequency response.

this case, the original base mass will keep its position undisturbed when vibration is excited. The system dynamic equations are:

$$\begin{aligned} M\ddot{x} + Kx + k(x - y) &= Qe^{ip_0 t} \\ m\ddot{y} + k(y - x) &= 0 \end{aligned} \tag{3.135}$$

where $q(t) = Q\exp(ip_0 t)$ is the loading, M the mass of the base system, K the spring constant of the base system, m the mass of the added system, and k the spring constant of the added system. When $\Delta = (1-p^2)(f^2-p^2)-vp^2f^2 \neq 0$, the equations have solutions as:

$$\begin{aligned} x &= \frac{Q}{K}\frac{f^2 - p^2}{\Delta}e^{ip_0 t} \\ y &= \frac{Q}{K}\frac{f^2}{\Delta}e^{ip_0 t} \end{aligned} \tag{3.136}$$

where $p = p_0/\omega_0$, $\omega_0 = (K/M)^{1/2}$ the natural frequency of the base system, $f_0 = (k/m)^{1/2}$ the natural frequency of the tuned mass damper, $v = m/M$ the mass ratio, and $f^2 = {f_0}^2/{\omega_0}^2$ the modulation factor. When $\Delta = 0$, Equation (3.135) has no solution. The new resonant frequencies of the system are:

$$\omega_{1,2} = \sqrt{\left[\left[1 + f^2(1+v) \pm \sqrt{[1 + f^2(1+v)]^2 - 4f^2}\right]/2\right]} \tag{3.137}$$

A tuned mass damper is suitable for damping out a particular frequency of a structure. It requires exactly the same resonance frequency for both the base and the added systems. However, if the frequency of outer loading is the same as the resonance frequency of the combined system, the resonance will be excited.

Adding damping into the above system improves the tuned mass performance. The new damper is named as a dynamic vibration absorber. The damping force can be friction or from electromagnetic induction. The equations for frequency response of the combined system with damping are:

$$\begin{aligned} M\ddot{x} + Kx + k(x - y) + \mu_0(\dot{x} - \dot{y}) &= Qp^{\alpha}e^{ip_0 t} \\ m\ddot{y} + k(y - x) + \mu_0(\dot{y} - \dot{x}) &= 0 \end{aligned} \tag{3.138}$$

In this set of equations, an amplitude change of the loading p^{α} is introduced and μ_0 is the damping coefficient of the dynamic vibration absorbers. The solutions of this system are:

$$\begin{aligned} x &= \frac{Q}{K}P^{\alpha}\frac{f^2 - p^2 - i\mu p}{b_1 + i\mu p b_2}e^{ip_0 t} \\ y &= \frac{Q}{K}p^{\alpha}\frac{f^2 + i\mu p}{b_1 + i\mu p b_2}e^{ip_0 t} \end{aligned} \tag{3.139}$$

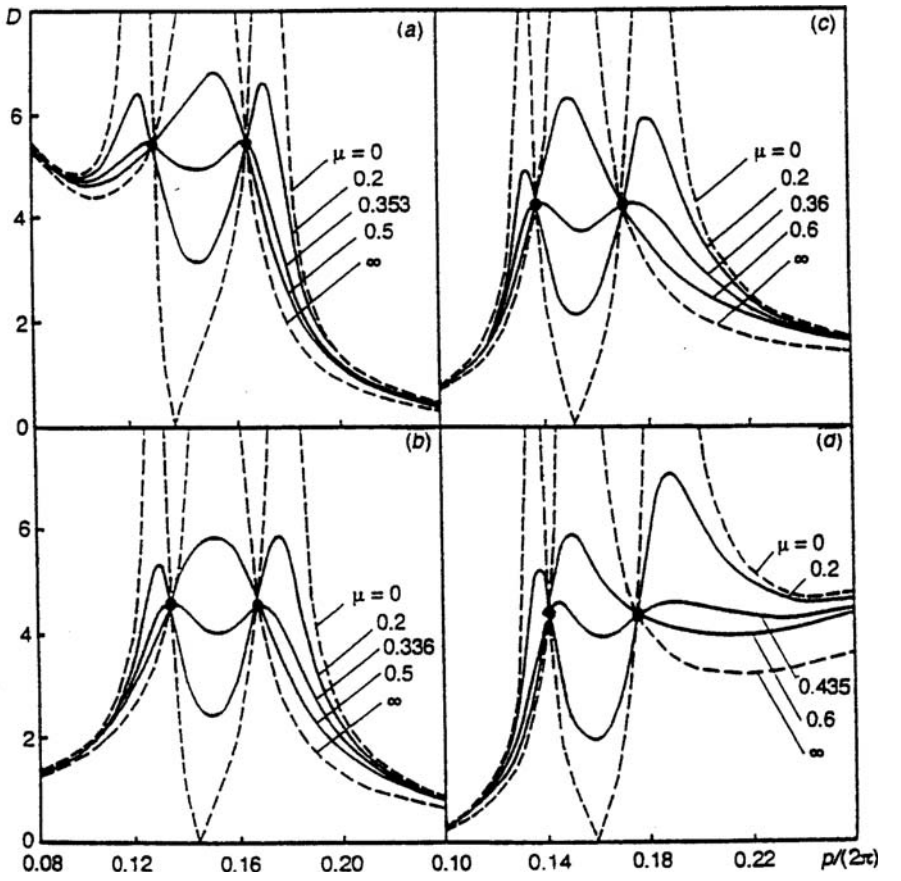


Fig. 3.43. Frequency response curves of different damping μ, different amplitude α, and different modulation (**a**) $\alpha = -2$, $f^2 = 0.735$, (**b**) $\alpha = 0$, $f^2 = 0.827$, (**c**) $\alpha = 2$, $f^2 = 0.909$, and (**d**) $\alpha = 4$, $f^2 = 1$ when $\nu = 0.1$ (Korenev and Reznikov, 1993).

where $b_1 = (1-p^2)(f^2-p^2)-\nu p^2 f^2$, $b_2 = 1-p^2(1-\nu)$, $\mu = \mu_0/(m\omega_0)$. In the expressions the amplitude coefficients are complex numbers.

The calculated frequency response curves are shown in Figure 3.43. By adding damping, the two infinite resonance peaks disappear, resulting in smooth hump-shaped response curves. One characteristic of the curves is that all the curves pass through two common points and most responses are lower than one particular amplitude value after the system optimization. Under the optimal conditions, the frequency of a tuned mass absorber is near to, but not the same as the resonance frequency of the base system. It requires a small damping coefficient. Over damping has exactly the same effect as without damping.

To obtain the best results, the dynamic absorbers should be arranged at those nodal points where the structural vibrations are the largest.

3.4.4.2 Viscoelastic Layer Damping

Polymer materials with both elastic and plastic characteristics, such as silicone rubber, can be used together with small mass as dynamic absorbers. Another polymer material application is the viscoelastic layer damper. The polymers consist of many chain-like organic molecules. When the polymer deforms, some of its mechanical energy converts into heat resulting in damping of the system.

The stress-strain relationship of the viscoelastic materials can be expressed as:

$$\sigma = (E' + iE'')\varepsilon = E'(1 + i\eta)\varepsilon \tag{3.140}$$

where σ is the stress, ε the strain, E' the storage modulus, E'' the loss modulus, and η the material loss ratio. The viscoelastic material usually has a higher material damping and a smaller storage modulus, resulting very little load-bearing capability.

An unconstrained viscoelastic damping layer can be formed by adding this material directly on top of a structure surface as shown in Figure 3.44(a). In the figure, the lower part is the structural member with a modulus of E. If an alternative force is applied to this structure, the stress-strain relationship is:

$$P = b\varepsilon[(Eh + E't)^2 + (E''t)^2]^{1/2} \tag{3.141}$$

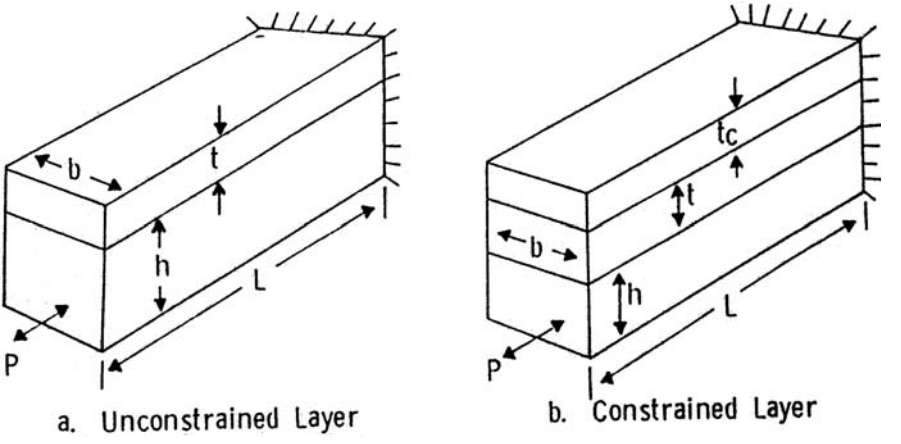


Fig. 3.44. Unconstrained and constrained viscoelastic damping layers.

where b and t are the width and thickness of the damping layer, $E'+iE''$ the modulus of the damping layer, and h the thickness of the structural layer. The total energy stored in the structure is:

$$U_s = bL\varepsilon^2(Eh + E't)/2 \tag{3.142}$$

where L is the length of the viscoelastic damping layer. Within one cycle, the energy loss of the system is:

$$D_s = \pi bLt\varepsilon^2 E'' \tag{3.143}$$

Since the elastic modulus of the damping layer is low, the energy loss ratio within one cycle is:

$$\eta_s = D_s/2\pi U_s = tE''/(Eh + E't) \approx tE''/Eh \tag{3.144}$$

The damping of this arrangement is less effective.

To further improve the damping, a constrained viscoelastic layer is developed as shown in Figure 3.44(b). In this arrangement, an additional structural layer is on top of the damping layer. This additional structure layer is fixed at one end and free at the other. When an alternative force is applied to the system, the damping layer bears a shear force. Its strain increases gradually from zero at the fixed end to the free end. The strain is:

$$\varepsilon = x\delta/tL \qquad 0 \le x \le L \tag{3.145}$$

where δ is the structure displacement at the free end. Assuming the shear modulus of the viscoelastic material is $G=G'+iG''$, the total energy loss within one cycle is:

$$D_s = \pi G''\delta^2 Lb/3t \tag{3.146}$$

The energy stored is:

$$U_s = (Eh/2 + G'L^2/6t)b\delta^2/L \tag{3.147}$$

Since the elastic modulus of the damping layer is lower, the energy loss ratio within one cycle is:

$$\eta_s = \pi EG''L^2/3Eth \tag{3.148}$$

Because the length is longer than the thickness and the height, the constrained viscoelastic layer damping produces a higher damping efficiency. The main disadvantage of this is that the viscoelastic layer may yield as the length

increases. In order to overcome this problem, segmented constrained viscoelastic layer damping or multi-layer viscoelastic layer damping can be used. This damping design has been used on the secondary vane structure of some large infrared telescopes.

3.4.4.3 Optimization of the Motion Profile

The telescope drive is one source which causes the structural vibration. An impulse of current or voltage includes excitations of wide frequency band. To restrict the frequency band, a filter can be used in the control circuit. However, limited bandwidth in the frequency response will make the control system less effective in suppressing external disturbances. Therefore, an effective method in vibration control is to optimize the motion profile (or to shape the input signal) in order to reduce high frequency excitations in the drive system.

One method of motion profile optimization is to replace a simple large step function with two smaller ones. The amplitude of the smaller step is half of the required large step, so that the vibration induced from the first step is compensated by that from the second step. The time interval between two steps should be exactly a half of the vibration period which equals the lowest resonant frequency of the structure. To perform this profile optimization, it is necessary to know this resonant frequency.

Another method of motion profile optimization is to shape the displacement and velocity curves so that they are smoother and include no high frequency components as shown in Figure 3.45.

The last profile optimization method involves a profile generator in the control system. The generator produces curves for acceleration, velocity, and displacement. These generated curves are fed forward into the control loops. Since these curves include no high frequency components, the structure resonant frequency is not excited by the drive system. One set of the ideal motion curves suggested by D. Woody includes a Gaussian velocity curve and error function curves for both acceleration and displacement. The formulas for these velocity, displacement, and acceleration curves are:

$$\begin{aligned} V(t) &= \frac{1}{t_0\sqrt{\pi}}\exp\left(-\frac{t^2}{t_0^2}\right) \\ S(t) &= \frac{1}{2}(1 - erf(t/t_0)) \\ A(t) &= \frac{-2t}{t_0^3\sqrt{\pi}}\exp\left(-\frac{t^2}{t_0^2}\right) \end{aligned} \tag{3.149}$$

where t_0 is a quarter period of these functions.

Fourier transforms of these curve functions represent the energy distribution in the frequency domain. The transforms show that the energy attenuates exponentially with increase of frequency. The Fourier coefficients of the

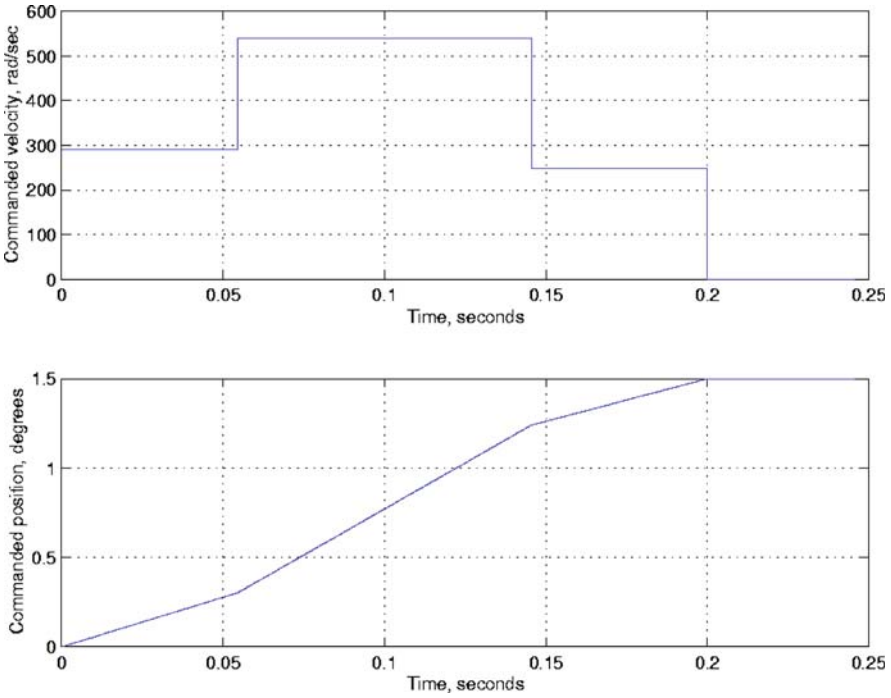


Fig. 3.45. Optimized velocity (*up*) and displacement curves (*bottom*) (Anderson, 1998).

displacement and velocity functions reach the maximum values at a zero frequency and that of the acceleration reaches the maximum value when the frequency equals $1/(4t_0)$. When the frequency equals $1/t_0$, the energy reduces to less than 0.1% of the maximum value. If inverse Fourier transforms are made with a finite frequency band to represent the original motion functions, the residual errors are $<10^{-5}$ for a cut-off frequency of $0.8/t_0$ and are $<10^{-9}$ for a cut-off frequency of $1/t_0$. This shows that by using this motion curve profiling, all high frequency components in the drive system are removed, so that structural vibration is controlled. However, the disadvantage of this method is its incapability in overcoming external disturbances.

3.4.5 Telescope Foundation Design

A foundation of a telescope has a direct effect on the structural stability. Foundation design is a special topic outside this book, therefore, only basic formulas and essential knowledge of the foundation design are provided in this section for reference. Generally, a foundation is a concrete solid platform supported by a number of circular or rectangular reinforced concrete columns. For

high precision structures, the columns should reach the bottom of the soil layer and be supported on the surface of the base rock.

The behavior of an unburied foundation part is the same as other structures. However, the property of a buried foundation part is entirely different. The stiffness of a buried foundation is dependent on the soil properties, such as the shear modulus G and the Poisson ratio v. The stiffness is also related to the soil contacting area. A large soil contact area and a high soil modulus produce a high stability to the foundation.

For telescope foundations, the most important characteristic is not its load bearing capacity but its dynamic stiffness, or its spring constant. The spring constant affects the telescope vibration and pointing. For a simple circular footing with a radius r and a height h, the dynamic spring constants in the vertical and horizontal direction are respectively (Arya et al., 1984):

$$\begin{aligned} k_v &= \frac{4Gr}{1-v}\left[1+0.6(1-v)\left(\frac{h}{r}\right)\right] \\ k_z &= \frac{32(1-v)Gr}{7-8v}\left[1+0.55(2-v)\left(\frac{h}{r}\right)\right] \end{aligned} \tag{3.150}$$

The dynamic spring constants in the rocking and torsion directions are respectively:

$$\begin{aligned} k_\phi &= \frac{8Gr^3}{3(1-v)}\left[1+1.2(1-v)\left(\frac{h}{r}\right)+0.2(2-v)\left(\frac{h}{r}\right)^3\right] \\ k_\theta &= \frac{16Gr^3}{3} \end{aligned} \tag{3.151}$$

If L is the horizontal length perpendicular to the rocking direction and B the length in another direction, the dynamic stiffness of a rectangular foundation can also be calculated from the above equations by using an equivalent radii. These equivalent radii in vertical and horizontal, rocking, and torsion directions are respectively:

$$\begin{aligned} r_{v,z} &= \sqrt{BL/\pi} \\ r_\phi &= \sqrt{[4]BL^3/\pi} \\ r_\theta &= \sqrt{[4]BL(B^2+L^2)/6\pi} \end{aligned} \tag{3.152}$$

The dynamic spring constant of a complex foundation can be calculated by adding the related spring constants of the foundation components together. For instance, a circular platform supported by few columns entirely under the ground has its rocking stiffness as a sum of two parts: one is the rocking stiffness

Table 3.5. Soil's dynamic shear modulus and Poisson ratios (Arya et al., 1984)

Soil type	Shear modulus	Soil condition	Poisson ratio
Soft clay	17–28 $\times 10^6$ N/m^2	Saturated clay	0.45∼0.5
Stiff clay	56∼112	Partially saturated	0.35∼0.45
Very stiff to hard clay	>112	Dense sand/gravel	0.4∼0.5
Medium dense sand	28∼84	Medium dense sand/ gravel	0.3∼0.4
Dense sand	56∼112	Silt	0.3∼0.4
Medium dense gravel	84∼140		
Dense gravel	112∼224		

of the circular platform and the other is a sum of the product of the column dynamic stiffness in the vertical direction and the relative radius squared of the column to the symmetrical plane of the foundation. More columns results in a higher stiffness. However, excessively dense columns can reduce the soil shearing modulus. The real telescope dynamic behavior can be predicted by adding the foundation stiffness to the structural FEA model.

Table 3.5 lists some soil dynamic shear modulus and Poisson ratios. The dynamic shear modulus is generally larger than the static shear modulus. When the dynamic shear modulus is large, the static shear modulus is about half of the dynamic shear modulus. When the dynamic shear modulus is small, the static shear modulus is about 1/20th of the dynamic shear modulus. The soil conditions in New Mexico of the US are: from the surface down to about 2 m, the soil dynamic shear modulus is $7 \cdot 10^6$ N/m^2 and the Poisson ratio is 0.15; for a depth up to 7 m, the soil dynamic shear modulus is $2 \cdot 10^7$ N/m^2 and the Poisson ratio is 0.32; as the depth increases further, the soil dynamic shear modulus becomes $3.5 \cdot 10^7$ N/m^2 and Poisson ratio becomes 0.06.

References

Anderson, T, 1998, A first study of MMA antenna offset performance, ALMA memo, 231, National Radio Astronomy Observatory, US.

Arya, S., O'Neill, M. and Pincus, G., 1984, Design of structures and foundations for vibrating machines, Gulf Publishing Co., Houston, Texas.

Avitabile, A., 2001, Experimental modal analysis, Sound Vibration, 35 (1), 20–31.

Bely, P. Y., 2003, The design and construction of large optical telescopes, Springer, New York.

Borkowski, K. M., 1987, Near zenith tracking limits for altitude-azimuth telescopes, Vol. 37. Acta Astronomica, Poland, 79–88.

Brunetto, E., et al., 2004, OWL, opto-mechanics, phase A, Proc. SPIE 5489, 571–582.

Chatfield, C., 1996, The analysis of time series, an introduction, 5th edn. Chapman & Hall/CRC, London.
Cheng, J., 1987, Pointing error correction for 3.8 m United Kingdom Infrared Telescope, Vol. 28, No. 3. Acta Astronomia Sinica, Nanjing, China.
Cheng, J., 1994, Damping and vibration control, ALMA memo 125, National Radio Astronomy Observatory, Charlottesville.
Cheng, J., 2006, The principles and applications of magnetism, Chinese Science and Technology Press, Beijing, China, in Chinese.
Cheng, J. and Li, G., 1988, Mechanical properties of crossed-vane type supporting structure, Astronomical Instrument and Technology, No. 1, p 5–10, Nanjing, China.
Cheng, J. and Xu, X., 1986, Some problems of alt-azimuth mounting telescopes, Progress in Astronomy, Vol. 4, No. 4, 322–337, Shanghai, China.
Crassidis, J. L. and Junkins, J. L., 2004, Optimal estimation of dynamic system, Chapman & Hall/CRC, London.
Davenport, A. G., 1961, The spectrum of horizontal gustiness near the ground in high winds, Quart. J. R. Meteorol. Soc., 87, 194.
Dyrbye, D. and Hansen, S. O., 1996, Wind loads on structure, John Wiley & Sons, New York.
Eaton, J. A., 2000, Report on application of hydrostatic bearings to the azimuth axis of the TSU 2 m telescope, Tennesse State University.
Forbes, F. and Gaber, G., 1982, Wind loading of large astronomical telescopes, SPIE, 332, 198–205.
Gawronski, W., 2007, Control and pointing challenges of large antennas and telescopes. IEEE Trans. Control Syst. Technol., 15, 276.
Gawronski, W. and Souccar, K., 2003, Control systems of the Large Millimeter Telescope, IPN Progress Report 42-154, JPL, NASA.
Haojian, Y., 1996, A new expression for astronomical refraction, Astro. J., 112, 1312.
Hu, Q., 2007, General design of astronomical telescopes, Nanjing Institute of Astronomical Optical Technology.
Ieki, A., et al., 1999, Optical encoder using a slit-width-modulated grating, J. Modern Opt., 46 (1), 1–14.
Juvinall, R. C. and Marchek, K. M., 1991, Fundamentals of machine component design, John Wiley & Sons, New York.
Koch, F., 1997, Analysis concepts for large telescope structures under earthquake load, SPIE, 2871, 117.
Koch, F., 2008, private communication.
Korenev, B. G. and Reznikov, L. M., 1993, Dynamic vibration absorbers, John Wiley & Sons, New York.
Mangum, J. G., 2001, A telescope pointing algorithm for ALMA, ALMA memo 366, National Radio Astronomy Observatory, Charlottesville.
Mangum, J. G., 2005, ALMA notes, NRAO, Charlottesville.
Parks, R. E. and Honeycutt, K., 1998, Novel Kinematic equatorial primary mirror mount, SPIE 3352, 537–543.
Richter, C. F., 1958, Elementary seismology, Freeman, San Francisco.
Sarioglu, M. and Yavuz, T., 2000, Vortex shedding from circular and rectangular cylinders placed horizontally in a turbulent flow, Tur. J. Eng. Environ. Sci., 24, 217–228.
Schneermann, M. W., 1986, Structural design concepts for the 8 meter unit telescopes of the ESO-VLT, SPIE Proc., 628, 412.

Shi, X. and Fenton, R. G., 1992, Solution to the forward instantaneous kinematics for a general 6-dof Stewart platform, Mech. Mach. Theory, 27 .(3), 251–259.
Simiu, E. 1974, Wind spectra and dynamic alongwind response, J. Struct. Div., ASCE, 1897, ST_9.
Simiu, E. and Scanlan H., 1986, Wind effects on structure, John Wiley & Sons, New York.
Stewart, D., 1966, A platform with six degrees of freedom, Proc. Inst. Mech. Eng., 180 (15), Part 1, 371.
Tedesco, J. W., et al., 1999, Structural dynamics, theory and applications, Addison-Wesley, Montlo Park, California.
Wallace, P. T., 2000, Manual of TPOINT software, TPOINT Software, Abingdon.
Watson, F. G., 1978, The zenithal blind spot of a computer-controlled alt-azimuth telescope, MNRAS, 183, 277–284.

Chapter 4
Advanced Techniques for Optical Telescopes

This is the most important and key chapter of the whole book. It discusses active and adaptive optics and various interferometers used in astronomy. It covers nearly all aspects of active and adaptive optics, including wavefront sensors, curvature sensors, phasing sensors, actuators, deformable mirrors, tip-tilt correction, adaptive secondary mirror, phase corrector, metrology system, laser guide stars, atmosphere tomography, and multi-conjugate adaptive optics. It also discusses all astronomical interferometers, including speckle, Michelson, Fizeau, intensity, and amplitude interferometers. Theories, principles, formulas, and application of these devices and interferometers are provided. Important temporal and spatial coherence theories are also discussed in depth. This chapter includes many important formulas and figures. It is the first time that so much related information have been packed within a short chapter of a book.

4.1 Active and Adaptive Optics

4.1.1 Basic Principles of Active and Adaptive Optics

Active and adaptive optics are the most noteworthy achievements in astronomical optics today. In the past, all astronomical telescopes were passive systems where no built-in optical correcting devices were used for improving the system performance. The image quality depended solely on the precision and stability of the optics, in order to reduce gravity and thermal influences. This requires a rigid structure and low thermal expansion mirrors. The weight and the cost of this type of design increase as the third power of the telescope diameter.

For less weight and dramatic cost reduction, light weight optics with some built-in adjusting devices was introduced to cope with gravity and thermal influences. These mirror shape and distance adjusting devices which worked at low frequencies are termed active optics. The first active optics system was in the European Southern Observatory (ESO) New Technology Telescope (NTT)

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_4,

completed in 1989. In 1992, the active concept was applied to the 10 m Keck I segmented mirror telescope. Active optics can be used for correcting defects within a telescope at frequencies up to a few Hertz (Bely, 2003).

Expanding the active concept to a higher frequency range, wind and atmospheric influences can also be corrected. These devices are known as adaptive optics which can achieve diffraction limited image quality. Adaptive optics began with the analysis of a star image or its wavefront. Early systems relied on natural guide stars. By using natural guide stars, limited sky coverage as well as a small field of view is expected. Laser guide stars, introduced in the 1990s, expands the telescope sky coverage. The field of view is defined by the isoplanatic condition in adaptive optics. The isoplanatic patch is a small field area in which the atmospheric disturbance remains approximately constant. Within this area, all light collected can be assumed to have passed through the same atmospheric volume. Therefore, if the disturbance is compensated for one direction, the disturbance will also be compensated within this isoplanatic area. To expand the field of view used in adaptive optics, multi-laser guide stars, atmospheric tomography, and multi-conjugated adaptive optics have been introduced. With these techniques, diffraction limited imaging over all the sky and within full field of view becomes possible for existing and future optical telescopes.

Within an optical telescope, there are many factors which influence the image quality. These are (Wilson, 1982):

(a) optical design (residual aberrations);
(b) optical manufacture errors;
(c) errors from the mirror support system, from position changes between optical components, and during pointing and tracking;
(d) long-term variations of the errors in (c);
(e) thermal behavior of optical components;
(f) long-term variations in material properties of optical components (e.g. mirror warping);
(g) effects from local air and atmosphere, namely mirror, dome, and atmospheric seeing;
(h) wind induced errors of optical components; and
(i) errors due to structural (mirror) resonance.

In the frequency domain, the above errors have different time scales as:

(1) a steady DC part as errors in (a) and (b);
(2) a very low frequency part, such as in (f);
(3) a low frequency part, such as in (d), where the time scale is months or tens of days, and in (e), where the time scale is hours:
(4) an intermediate frequency part, such as in (c), at a frequency of about 10^{-3} Hz for tracking and of 10^{-2} Hz for pointing, and in (h), where the range is of 0.1 to a few Hertz; and
(5) a high frequency part, such as errors in (i), where the range is 5–100 Hz, and in (g), where a wide frequency range exists from 0.02 to 1,000 Hz.

Within a telescope system, the low or intermediate frequency source-induced wavefront variations can be corrected either by changing the primary mirror shape or by adjusting the separation between mirrors. If the primary mirror is a thin monolithic one, the change of the mirror shape can be accomplished by applying forces (or moments) on the back of the mirror. For a segmented primary mirror, segment positional adjustment is necessary. These corrections are slow and the system belongs to the category of active optics.

Without damping, the response of a large primary mirror with low resonant frequency to a control command is slow. Therefore, small-size deformable or tip-tilt mirrors instead of the primary mirror are often used in adaptive optics. For the adaptive secondary mirror, damping is added through a tiny air gap between a very thin deformable mirror and a reference plate (mirror). The thin deformable mirror also has a fast response to the control command. Deformable, tip-tilt, and adaptive secondary mirrors are adaptive optics devices used for compensating wavefront errors due to atmospheric disturbance.

In mathematics, if $T(v)$ is the optical (or modulus) transfer function of a telescope with aberrations, $<A(v)>$ is the optical transfer function of the atmosphere for a long exposure, and v is the spatial frequency, then the optical transfer function of the system for a long exposure is given by:

$$G(v) = T(v) \cdot <A(v)> \tag{4.1}$$

The system is aberration and seeing limited. A long exposure means that the integration time is longer than the atmospheric coherence time (Section 4.1.6).

For an active optics system, the goal of its control is to make $T(v)$ the same as the optical transfer function of an ideal one without aberrations, namely $D(v)$. The optical transfer function of an active optics telescope is limited only by the atmosphere. The system is now seeing limited. The filter added by the active optics is $F(v)$:

$$F(v) = D(v)[T(v)]^{-1} \tag{4.2}$$

Adaptive optics is intended to correct all the defects, including atmospheric turbulence so that the system approaches diffraction limited performance. The optical transfer function of an ideal adaptive optics system is equal to the optical transfer function of an ideal telescope, namely:

$$G_0(v) = D(v) \tag{4.3}$$

where $G_0(v)$ corresponds to a diffraction limited transfer function of the telescope with a Strehl ratio of unity.

Open loop control is used in some active optics systems where the effects, such as from gravity or temperature, are predictable. This open loop control is a type of "lookup table correction." Closed loop control is used for most active and

adaptive optics systems. The closed loop system includes local loop and system loop ones. Active optics uses mostly a local closed loop, while adaptive optics uses more the system loop.

The feedback signal of a system closed loop is the wavefront error of a guide star detected by a wavefront sensor. This signal is processed by a computer and the commands are generated for individual actuators. For active optics, the sampling rate is usually low, while for adaptive optics, the rate is high. The sampling rate of some wavefront sensors is limited. Therefore, many systems rely only on local feedback loops. These local loops are referred to as the metrology system. A metrology system may include distance, displacement, angle, individual image position, or optical path length sensors.

Actuators in active optics can change the surface shape of a primary mirror. Generally, low spatial frequency errors can be corrected, while high spatial frequency errors are difficult to remove as the number of actuators used is limited. Mirrors required by active optics should be thin, smooth, and free from high spatial frequency ripples.

For adaptive optics, since the Fried parameter varies as the 6/5th power of the wavelength, the wavefront slope change remains the same for different wavelengths. The images are better for longer wavelength. A small guide star limiting magnitude (numerically smaller than 13) is present in adaptive optics, as sufficient photons are required for fast wavefront sensing. This limits the sky coverage. The field of view with traditional adaptive optics is also limited by the atmospheric isoplanatic angle.

To extend the sky coverage limited by the guide star magnitude, a laser guide star can be used. A laser guider star can be generated in any sky area. However, the laser guide star has a cone effect as the disturbance volume of the guide star is not the same as that of the object observed. To overcome this, more laser guide stars are used. Multiple laser guide stars produce a detailed 3-D map of the atmospheric disturbance, and the technique is referred to as atmospheric tomography. Atmospheric tomography with deformable mirrors located on different conjugated planes of the turbulence layers are the basis for powerful multiconjugate adaptive optics (MCAO) systems. The resulting compensation of turbulence is now beyond the isoplanatic patch and covers the whole field of view.

4.1.2 Wavefront Sensors

Generally, there are three different groups of sensors in active and adaptive optics: (a) wavefront sensor, (b) curvature sensor, and (c) phasing sensor. The wavefront sensor provides wavefront error, the curvature sensor provides wavefront curvature and edge slope, and the phasing sensor provides piston error between two adjacent mirror segments. Wavefront sensors are discussed in this section, phasing and curvature ones are discussed in Sections 4.1.4 and 4.1.5.

Real-time wavefront sensing is a major requirement in active and adaptive optics. Wavefront sensors can gain information either in the pupil plane or in the

image plane. The Shack–Hartmann sensor is a pupil plane one which is known as a direct method. The curvature sensor is in the focal plane and is an indirect method. Other indirect methods include phase retrieval, "phase diversity," and multi-dither techniques. The iterative phase retrieval is through an inverse Fourier transform of the image pattern. The phase diversity technique uses two images, one in focus and another off focus by a known distance. The multi-dither technique works through the modulation of phase at a much higher frequency. Some of these indirect methods are similar to the curvature sensors and out-off-focus holographic method discussed in Sections 4.1.5 and 8.4.1.

The basic requirement of wavefront sensors is to detect wavefront error with enough sensitivity and spatial resolution within a finite time range. The required sensitivity and accuracy is usually between 0.1 and 0.03 wavelengths or smaller. The spatial resolution should match the number of actuators used in the system. The sampling rate is determined by the frequency of error correction of a system, less than $0.01s$ for adaptive optics and much longer for active optics. In general, wavefront sensors for adaptive optics are more difficult. For adaptive optics using natural guide stars, the sensor should have high quantum efficiency and low noise and it should work with white light of incoherent sources to reach a higher stellar magnitude.

Major wavefront sensors are based on slope measurement in the concept of geometric optics or interference. Geometric optics assumes the light ray is orthogonal to the local wavefront and the direction of a ray can be found by focusing a sub-pupil beam. The geometric optics wavefront sensors include Shack–Hartmann, pyramid prism, and others. Using an interference concept, the amplitude of an original wavefront is split or transformed. Interference between the original and newly produced wavefronts forms fringes which provide slope information. The interference sensors include transverse grating shearing, phase contrast, and others.

Sensors of both concepts provide information of wavefront local slopes. From the local slopes, the wavefront shape is derived by orderly connecting of the slope lines as in Figure 4.1. Modern wavefront sensors also include other types, such as those from the curvature measurement. These sensors are introduced in later sections.

4.1.2.1 Shack–Hartmann Wavefront Sensor

In a Shack–Hartmann wavefront sensor, a lenslet array divides the pupil into sub-apertures. For a plane wavefront, images of sub-apertures are formed right on the foci of the lenslet. If the wavefront is disturbed, images of sub-apertures shift away from the focii. The displacements are proportional to the local slopes of the wavefront (Figure 4.2). A Shack–Hartmann sensor has its own reference wavefront generated by a reference light source in the instrument. This plane wavefront provides precise foci positions of the lenslet array.

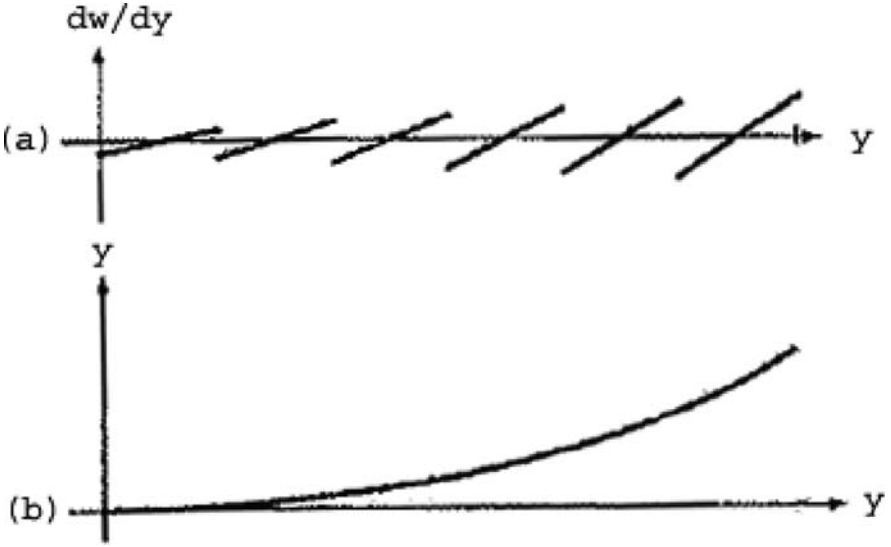


Fig. 4.1. Reconstruction of wavefront from local slope measurements.

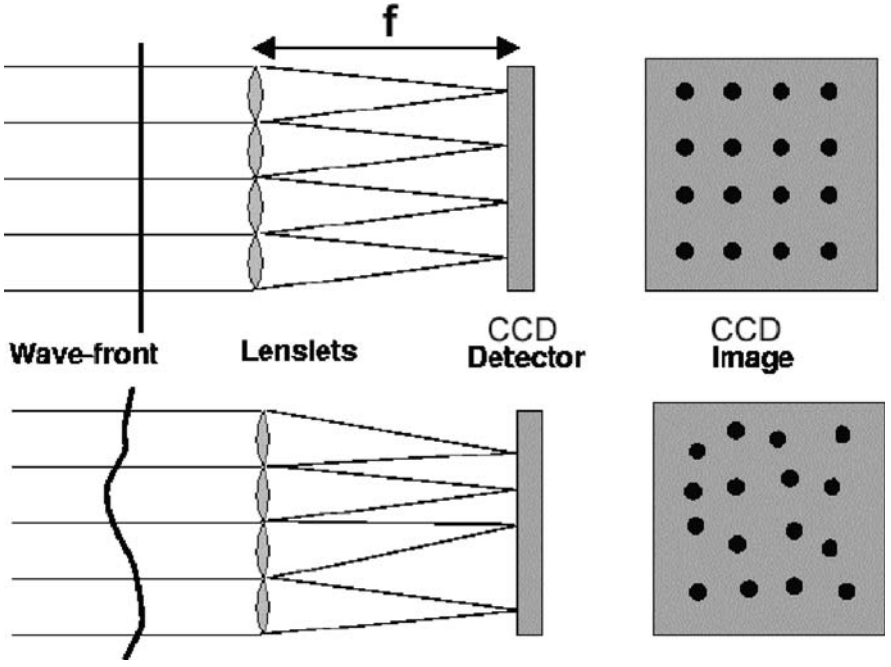


Fig. 4.2. Principle of the Shack–Hartmann wavefront sensing method (Max, 2003).

To derive wavefront error, the images of the entire pupil should be sampled and the estimation of sub-aperture image positions is made either by quad-cell or center of gravity approaches. The quad-cell approach is discussed in the star guiding part of Section 3.3.6. The center of gravity approach in estimating the focal position error x, when the wavefront error exists, is:

$$\begin{aligned} x &= \frac{\lambda}{2\pi \cdot S} \int\limits_{subaperture} \frac{\partial \phi(\vec{r})}{\partial r_x} d\vec{r} \\ x &= \frac{\sum\limits_{i,j} x_{ij} I_{ij}}{\sum\limits_{i,j} I_{ij}} \end{aligned} \tag{4.4}$$

where $\phi(\vec{r})$ is the wavefront error function, S the exit sub-aperture pupil area, and I_{ij} the intensities of light on the detector pixel. The exit pupil of the sub-aperture has a smoothing effect on the incident wavefront.

The slope variance from the atmospheric turbulence over a circular sub-aperture, d, is given (Tatarskii, 1971):

$$< S_j^2 > = 0.98 \frac{6.88}{4\pi^2} \lambda^2 d^{-1/3} r_0^{-5/3} \tag{4.5}$$

This variance does not depend on the wavelength λ as the Fried parameter, r_0, is proportional to $\lambda^{5/6}$. This manifests that the Shack–Hartmann sensor is an achromatic device for all white light. As the lenslet has a focusing effect, the sensor with a wide bandwidth is fairly sensitive to faint stars.

For estimating the error of the measurement arising from photon noise, let β be the radius of the image formed by a sub-aperture. For a point source, $\beta = \lambda / d$ when the sub-apertures are smaller than r_0, and $\beta = \lambda / r_0$ when the sub-apertures are larger. The image intensity is determined by a probability density distribution. Each arriving photon permits one to determine image position with an error of β. When n photons are detected during exposure time, the photon error of the centroid position (i.e. slope) becomes $\beta / n^{1/2}$. This is the same after repeating the same measurement n times. By multiplying the slope error by $2\pi d / \lambda$ to obtain the variance of the phase difference between the edges of the sub-aperture in square radians:

$$< \varepsilon_{phot}^2 > = \frac{4\pi^2}{n} \left(\frac{\beta d}{\lambda} \right)^2 \tag{4.6}$$

The photon flux is proportional to the square of sub-aperture diameter d. It means that for a given β, the photon noise caused error of a Shack–Hartmann sensor is independent of the size of its sub-apertures. This conclusion applies

only to the ideal detector, however, in real systems with CCDs, larger sub-apertures are selected for fainter guide stars.

The Shack–Hartmann wavefront sensor is the most popular sensor in adaptive optics. The accuracy reached is about $\lambda/40$ peak to peak. As mentioned earlier, the Shack–Hartman sensors measure only the wavefront slope, not the phase. Therefore, one has to be careful when this device is used for segmented mirror telescopes as the wavefront in a segmented mirror aperture may not be a continuous one.

The Hartmann–Shack wavefront sensor can also be applied to solar telescopes where the source is extended instead of a sharp point one. In this case, instead of simply finding the image positions from the lenslet array, the cross-correlation functions between the lenslet images are used for the determination of the wavefront slope changes. The displacements between the cross-correlations represent the wavefront slope change between the sub-apertures. In this way, the wavefront shape of the beam can be reconstructed.

4.1.2.2 Pyramid Prism Sensor

A pyramid prism sensor (Figure 4.3) is based on a novel concept proposed by Carlo S. Ragazzoni (1996). This sensor includes a small pyramid prism at the focal point. The wavefront slope information is derived in the exit pupils of a refocusing lens. Since the vertex angles of this prism are close to 180°, the four

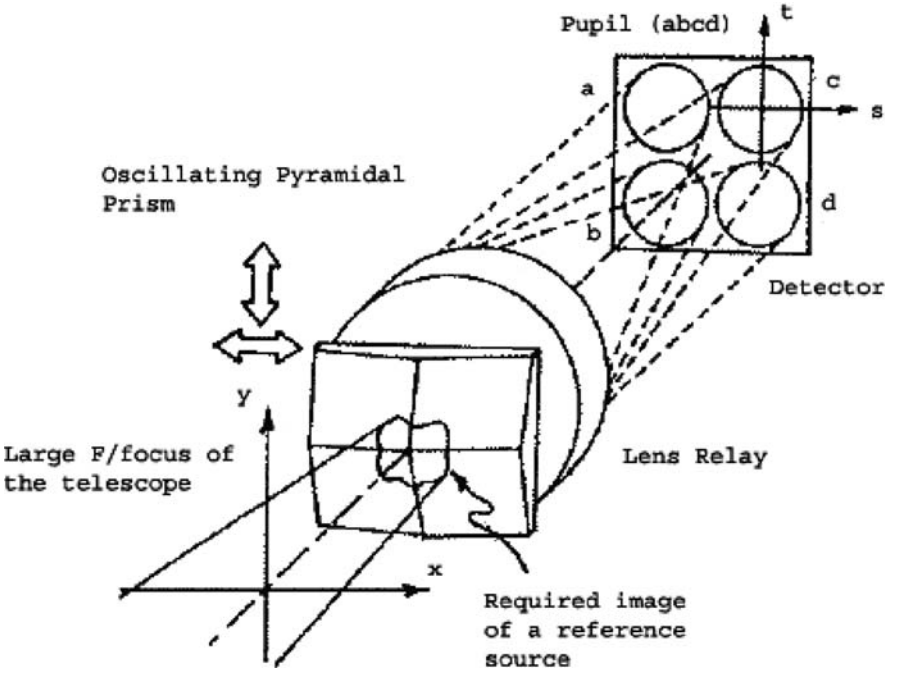


Fig. 4.3. Pyramid prism wavefront sensor (Ragazzoni, 1996).

surfaces of the prism will deflect a star light in slightly different directions. In the refocused pupil plane, four images of entrance pupil are obtained.

In geometric optics, the prism edges act as knife-edges of the traditional Foucault test. Any deviation of the light ray from their undisturbed path results in a fluctuation of the light intensity outside the focal point. In the refocused pupil plane, the intensity distribution of each image position is related to wavefront local slope. If the wavefront function is W, then each side of the prism will be illuminated by rays with a given range of ∇W. If the refocused pupil plane coordinate is (s, t) and the image plane coordinate is(x, y), then:

$$x = \frac{\partial W}{\partial s} F \quad y = \frac{\partial W}{\partial t} F \tag{4.7}$$

where F is the distance between the exit pupils (or the image position) and the nominal focal plane. When the prism is modulated in both x and y directions or the star image is steered relative to the prism by a tip-tilt mirror located on the exit pupil, the relative intensity in one cycle of each of the four faces can be obtained. The modulation can be a small square or a small circle around the apex of the prism. If the modulation amplitude R is larger than the maximum image blur size, the corresponding relative intensity can be derived from:

$$I_{ab}(s, t) = I_0(s, t) T\left(\frac{\partial W(s, t)}{\partial s} F\right) \tag{4.8}$$

where I_{ab} is light collected in the pupils a and b co-added together after a certain integration time and T is relative transparency due to modulation. The sum of light from other two pupils is the same except the transparency is now 1-T. The same is also true for relative intensity along the t direction. In the calculation, the effect from scintillation can be removed by using the wavefront derivative which is the difference between opposite pairs of pupils. The sum of the imaging intensity among an opposite pair of pupils should be constant without scintillation. If the intensity functions a, b, c, and d are linear, then the derivative of the wavefront in one direction is:

$$\frac{\partial W(s, t)}{\partial s} = \sin\left[\frac{\pi}{2}\frac{(a + b) - (c + d)}{a + b + c + d}\right]\frac{R}{F} = \frac{R}{F}\sin\left(\frac{\pi}{2} S_s\right) \tag{4.9}$$

This derivative is a function of the modulation amplitude. The other derivative has a similar form. One advantage of this sensor is that the sensitivity of the measurement can be adjusted by varying the modulation amplitude. As the correction of the wavefront error progresses, a higher sensitivity is required. Therefore, this sensor performs better than a Shack–Hartmann one in a closed control loop. This prism modulation in geometric optics concept is necessary as some value of the signal function S_s without modulation can be infinite – only the sign of the wavefront slope is derived.

However, if diffraction is considered, the prism acts as four spatial corner filters. Each filter has a quadrant corner to allow the light to go through and the other three quadrants are blocked. Therefore, the intensity distribution of the four pupil images is:

$$I_i(x,y) \propto FT^{-1}[H_i(f_x,f_y)FT[P(x,y)\exp(i\phi(x,y)]] \tag{4.10}$$

where H_i is the corresponding spatial corner filter, $P(x,y)$ and $\phi(x,y)$ the wavefront amplitude and phase functions on the aperture plane. Therefore, the signal derived from the pyramid prism sensor is (Esposito et al., 2000; Costa et al., 2003):

$$S_x(x_p,y_p) \propto \int_{-B(y_b)}^{B(y_p)} \frac{\sin[\phi(x,y_p)-\phi(x_p,y_p)]}{2\pi(x-x_p)} \frac{\sin[a_{tt}(x-x_p)]}{x-x_p} dx \tag{4.11}$$

where a_{tt} is the applied tilt modulation and B the boundary of the chord which is perpendicular to the axis. The signals we obtain from the pyramid sensor are proportional to the integral of a product of two terms. One term is the sine of the phase difference between each point on the chord, $y = y_p$, weighted by the distance between these two points. And the other is the modulation term. In the first term, more distant points with a higher probability of large phase difference will be weighted less. When the phase difference is small, the sine can be replaced by the angular value as:

$$\sin(\phi(x,y_p)-\phi(x_p,y_p)) \cong \phi(x,y_p)-\phi(x_p,y_p) \tag{4.12}$$

Therefore the sensor can be modeled as a linear system. From another point of view, if the phase over the aperture plane is the sum of a phase term of high spatial frequency and another of the low spatial frequency, then the sine term can be expressed as:

$$\begin{aligned}\sin(\phi(x)-\phi(x_p)) = &\sin(\phi_L(x)-\phi_L(x_p))\cos(\phi_H(x)-\phi_H(x_p)) \\ &+ \sin(\phi_H(x)-\phi_H(x_p))\cos(\phi_L(x)-\phi_L(x_p))\end{aligned} \tag{4.13}$$

If the phase varies rapidly, then in the first term of this expression the cosine of the phase difference will erase the contribution from the sine term unless the point x is really near the point x_p. In this case, since the term of the sine over the distance between point x and x_p inside the integral will be very large, its contribution to the integral will be decreased by the effect of the cosine, but it will not be zero. We see that the high spatial frequency terms themselves have an effect of "modulation." They act as delta functions inside the integral which linearize the system. Therefore, the sensor may not need any physical modulation when it is inside a closed control loop.

As wavefront sensors, a single pyramid prism sensor is very small in size and it can be moved freely inside the focal plane for finding the star image. This is especially suitable for the multi-guiding star sensing for observation beyond the isoplanatic patch. The pyramid prism used as a phasing sensor for segmented mirror telescopes will be discussed in Section 4.1.4.

Pyramid prisms with such a large vertex angle are difficult to make using a traditional prism fabrication technique. One method to avoid this problem is to form the required prism from two pyramid prisms with a small vertex angle difference. These two prisms are placed back to back, so that the net effect of these two prisms is the same as a prism with a nearly 180° vertex angle. A design using a front prism of 30° base angle and a back prism of a 28.338° base angle was reported by Esposito et al. (2003). The resulting half separation angle is $\delta = \alpha_1(n_1 - 1) - \alpha_2(n_2 - 1)$, where α_i and n_i are the base angles and the refractive indexes. Another way is to use two reflective pyramid prisms to form the wavefront sensor as shown in Figure 4.4 (De Man et al., 2003).

A new method of producing the required pyramid prism is called deep X-ray lithography (Ghigo et al., 2003). The material used for the pyramid prism is a polymer material called PMMA (Polymethyl methacrylate). It has a very high molecular weight and high X-ray resistance. During the prism manufacture, the PMMA sheet is bonded to a metal base. When the PMMA sheet (for example 500 μm in thickness) is placed on a turntable, a polished knife-edge is placed on top of the sheet as an X-ray mask. A plane X-ray beam from a synchrotron is directed onto the PMMA with an inclination angle necessary to obtain the desired pyramid vertex angle over the knife-edge. After one side is exposed, the turntable will rotate 90° for the second side of the prism. After all four sides are exposed to the X-ray, the irradiated PMMA part is removed by means of a suitable chemical developer. The X-ray exposure time is about 1 h for each side and the development time is about 24 h. The micro-roughness of the knife-edge required is about 10–20 nm. The refractive index of PMMA is 1.491 at 589 nm wavelength.

4.1.2.3 Interferometer Wavefront Sensor

A shearing interferometer wavefront sensor is based on the interference concept and is similar to an amplitude interferometer discussed in Section 4.2.4. In a

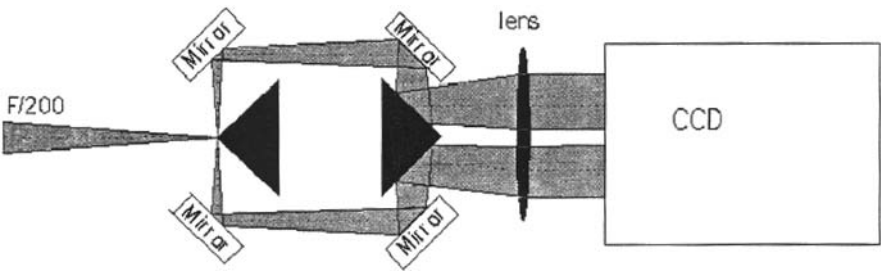


Fig. 4.4. Reflective pyramid prism wavefront sensor (De Man et al., 2003).

shearing interferometer, the beam is amplitude-split into two and these two beams produce interference fringes from which the wavefront phase is derived.

A typical interferometer wavefront sensor is the rotating radial grating one where the star image is focused on a rotational grating but away from its rotating center (Armitage and Lohmman, 1965). When the grating rotates with an angular velocity, a wavefront transverse shearing is realized and fringes are formed in the exit pupil. If the wavefront phase and amplitude are $\phi(x, y)$ and $A(x, y)$, the radiation is:

$$u(x, y) = A(x, y) \exp(-ik\phi(x, y)) \tag{4.14}$$

The lens $L1$ in Figure 4.5 generates the Fourier transform of $U(x, y)$on the pupil plane. If the Fourier transform of the transmission function of the grating is$M(x_0, t)$, then in the Fourier plane, the amplitude distribution of light after the grating is:

$$U(x_0, y_0, t) = \tilde{U}(x_0, y_0)M(x_0, t) \tag{4.15}$$

where $\tilde{U}(x_0, y_0)$ is the Fourier transform of $U(x, y)$. After the lens $L2$, an image is formed on a photon detector. So the intensity on the detector is:

$$I(x, y, t) = |U(x_0, y_0, t)|^2 \tag{4.16}$$

For a rotating grating, with a sinusoidal transmission ratio, the light intensity on the detector is given by:

$$I(x, y, t) = \frac{1}{2} + \gamma(\frac{1}{2}) \cos\{k[\phi(x - s, y) - \phi(x + s, y)] + 2\omega t\} \tag{4.17}$$

where $\omega = 2\pi v/g$, v is the radial speed of the grating, g is the period of the grating, and γ the degree of coherence of the light. In Figure 4.5, the interference is formed between images of the ± 1 order of the diffraction grating with a shearing length of $s = \lambda Z/g$. The phase differences on the detector are the optical path difference between two shearing wavefronts. If the shearing length

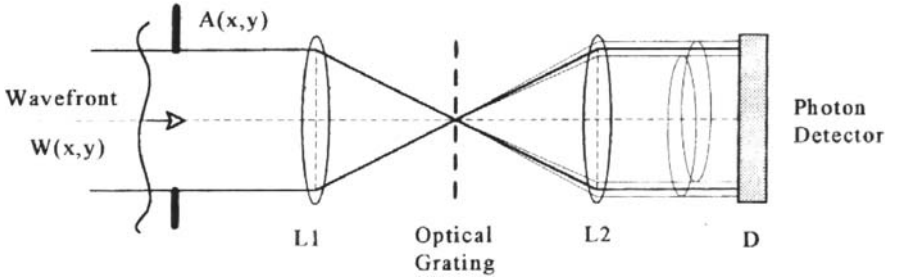


Fig. 4.5. Transverse grating shearing interferometer (Hardy et al., 1977).

is small, the path difference is proportional to the wavefront slope. However, one shearing device provides information only in one direction and another identical device is required in the perpendicular direction in order to get the full information of the wavefront shape.

An improved radial rotating grating interferometer is the nutating grating interferometer. In this sensor a steering mirror shifts the star image in the grating plane, so that a single grating can produce information in both directions. Other similar wavefront sensors include rotating soft knife-edge ones. The "soft" means that the transparency of this knife-edge is a continuous function instead of a step one. The principle of this sensor remains the same.

4.1.2.4 Phase Contrast Wavefront Sensor

The phase contrast wavefront sensor, or the quadrant phase mask sensor, is based on the phase contrast technique proposed by Frits Zernike for observing transparent objects with a microscope in 1930 (Bloemhof and Wallace, 2004). The technique allows the visualization of phase disturbances by introducing a phase shift filter in the Fourier plane. Normally, when a pupil is reimaged, only the intensity, the square of the amplitude, is recorded, the phase is lost. In this sensor, a phase shift of $\pi/2$ is produced in the central small area of $\sim\lambda/D$ in the focal plane, where D is the telescope diameter. The remaining high order beams outside the area are unaltered (Figure 4.6). After this transformation, the high order beams and the retarded zero order beam interfere in the pupil plane to form a phase re-distribution pattern.

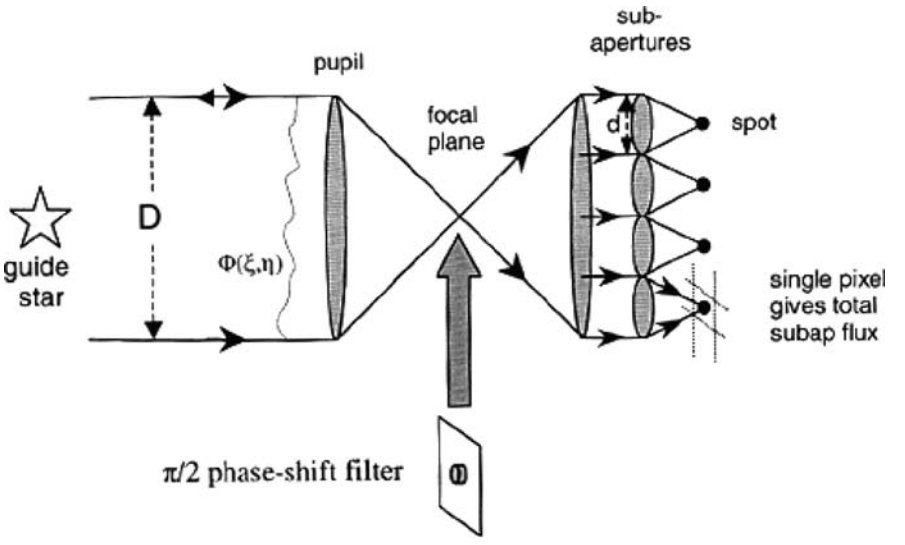


Fig. 4.6. Phase contrast sensor for adaptive optics (Bloemhof and Wallace, 2004).

For a telescope, the light beam can be expressed as $A\exp(ik\phi)$, where A is the amplitude and $k\phi$ is the phase of the wavefront. If the phase is small($k\phi < \pi/3$), the exponential function can be expanded as polynomials: $\exp(ik\phi) \sim 1 + ik\phi$. This expression shows that the phase of a wavefront can be regarded as a sum of two parts. One is a larger undiffracted component of the on-focus light and the other is a phase diffracted component which is 90° out of phase with the undiffracted part.

In the focal plane, the undiffracted rays are spatially localized within the central area of (asymptotically =)λ/D in the focal plane, while the diffracted rays would impinge at off-axis positions in the focal plane. If a 90° phase shift filter having approximately the diffraction-limited size of λ/D is placed at the focal point, then the undiffracted component would shift by 90° in phase and would be in phase (or phase opposition) with the diffracted component.

This can be realized by inserting a small dielectric plate with a 90° phase delay of roughly a diffraction-limited size in the focal point, so that the central rays are retarded by a quarter wave path. When these transformed rays together with diffracted rays are re-imaged in the pupil plane, the phase signal will become observable as a small variation in intensity across the re-imaged pupil. The intensity variation is proportional to $1 \pm 2k\phi$ (plus or minus depending on whether the focal plane spot advances or retards the phase).

The measurement device is similar to a Shack–Hartmann one. However, only one pixel is needed for each sub-aperture as intensity is recorded while at least four pixels are needed by the Shack–Hartmann wavefront sensor as the slope of the wavefront is recorded. This reduces the read out noise.

Dielectric material has a limited transmission bandwidth. A new beam splitter approach can achieve broadband operation. Because the phase difference between beams transmitted and reflected by a thin beam splitter is exactly $\pi/2$, the broadband phase contrast device can be realized by a 50/50 beam splitter in front of the focus (Figure 4.7). The two beams from the beam splitter are directed into the front and back sides of a thin mirror with a small hole at the focal point. The small pinhole size is λ/D. The output beams from both sides of the mirror will have a central ray 90° shifted against the off-axis rays. The re-imaged intensity of the output beams is derived. The beams of two sides can also be combined electronically to increase the signal-to-noise ratio.

4.1.3 Actuators, Deformable Mirrors, Phase Correctors, and Metrology Systems

After wavefront error is measured, the next task is to compensate this for achieving a better star image. Both active and adaptive optics require some types of wavefront error compensation devices. In active optics, the devices used are mechanical ones including displacement, moment, or force actuators. These actuators produce positional or surface shape changes of the mirror components. In adaptive optics, the compensation devices used include phase correctors and

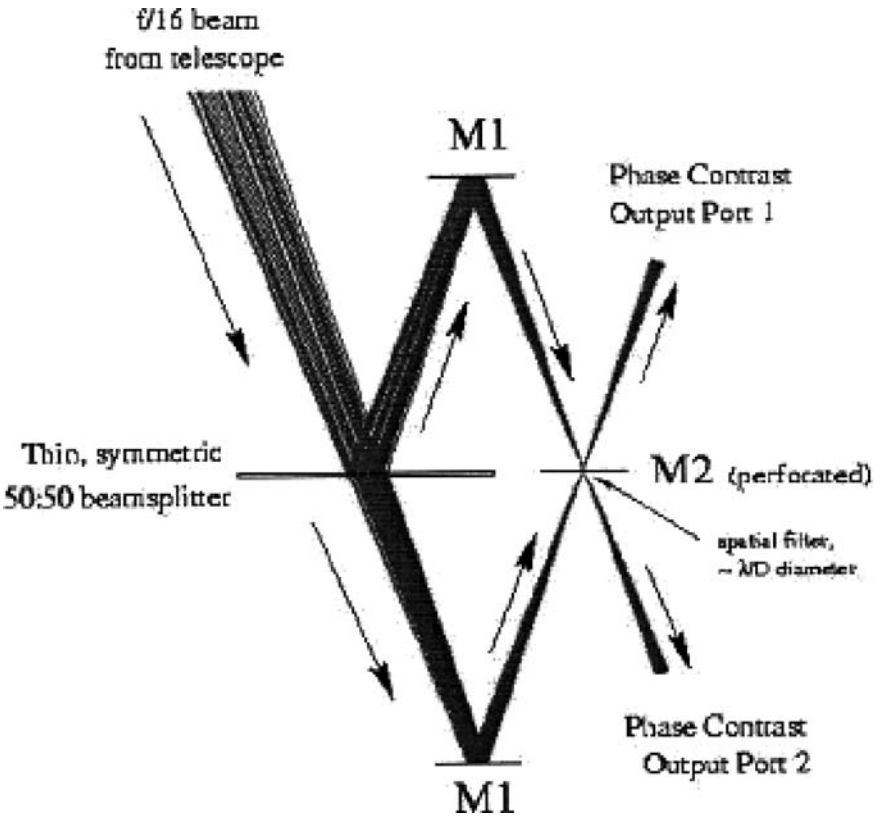


Fig. 4.7. Broadband phase contrast sensor arrangement (Bloemhof and Wallace, 2004).

deformable, tip-tilt, and adaptive secondary mirrors. The tip-tilt mirrors, including membrane and bimorph ones, correct only the first few terms of the wavefront aberrations and the adaptive secondary mirrors, at present, are only used in a few telescopes. These tip-tilt and adaptive secondary mirrors are discussed in later Sections 4.1.5 and 4.1.9. In this section some of the metrology systems used are also discussed.

4.1.3.1 Actuators

Mechanical mirror actuators are developed from the mirror support mechanisms. A passive mirror support system provides a fixed support force or a fixed positioning to a mirror system. This compensation of gravity loading is through either a counterweight mechanism or an air pressure regulator. Actuators are special mirror support systems which can change the mirror loading conditions. Because of this change, the mirror surface shape or position changes, resulting in the wavefront error compensation.

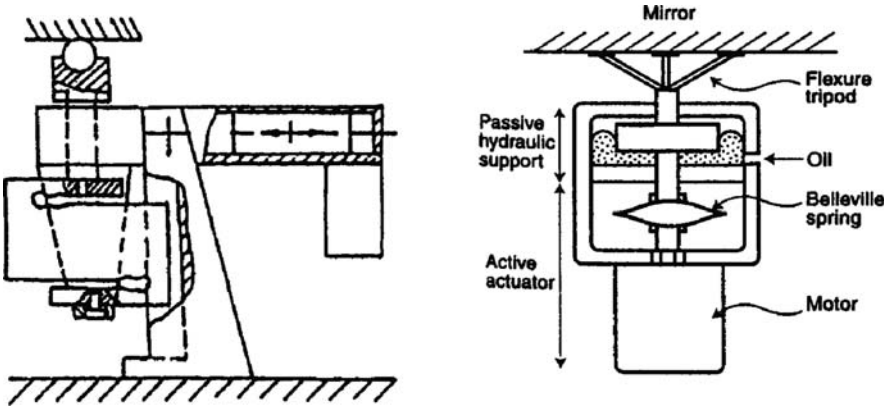


Fig. 4.8. Typical force actuators (ESO).

Force actuators are mainly used in telescopes with a thin monolithic primary mirror. Figure 4.8 shows two types of force actuators. One has a motor on a cantilever beam. It moves the counterweight along the beam so that the supporting force applied to the mirror back changes. The other force actuator is a two-layer system. A passive hydraulic part carries a fixed axial support load and a motor-driven spring part is for the active optics correction. Force actuators are relatively easy to design. They can achieve great sensitivity, high accuracy, and large dynamic range. Force actuators are often designed with load cells. The load cell on the mirror supporting point can measure the applied force and feedback the required force information.

To overcome frictional force, the mirror contact point often uses steel ball bearings. Within the elastic range, the deformation of a thin mirror and the force applied has a linear relationship. If the mirror is thin, the cross-talk between actuators is small. The total mirror surface change equals the sum of surface changes produced by each force actuator.

Moment actuators are variations of the force actuators. If there are two blocks glued on the mirror back, a variable moment will be produced by varying the push or pull force applied to these two blocks. Moment actuators are also used in the mass production of off-axis paraboloidal mirror segments in the prestressed polishing.

Displacement actuators are more difficult compared with force or moment actuators. They require a precise positional control even when the support force has been changed. Displacement actuators are essential in the segmented mirror telescope. The mirror segments have to be adjusted to a fraction of the operating wavelength which is about 10–50 nm. At the same time, it should also be stable and accurate as a passive mirror support system.

A typical displacement actuator used in the Keck telescope is shown in Figure 4.9. In this actuator, a ball screw will produce an accurate position change. The load variation can be absorbed by a preloaded spring and a

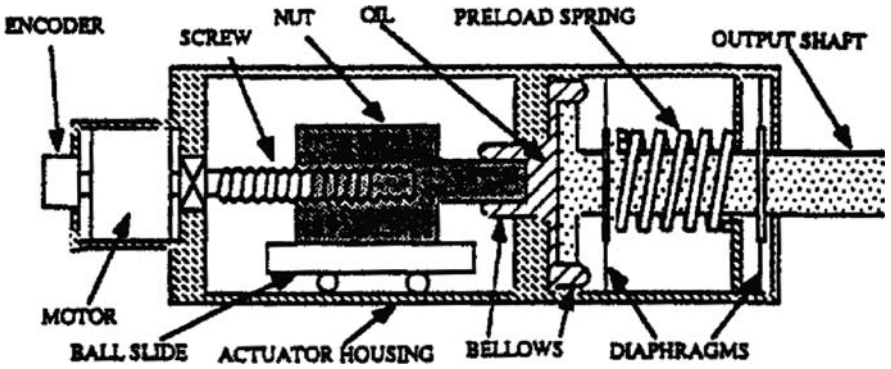


Fig. 4.9. Typical displacement actuator (Keck).

hydraulic system. Therefore, the position accuracy will not be affected. An encoder on the back provides information of the position change. The Keck actuator has a resolution of 4 nm over a full range of 1 mm.

Displacement actuators developed for the James Webb Space Telescope (WJST) have to operate in a very low temperature of 40 K. They consist of a pair of stepper-motor and lead-screw assemblies, arranged one after another through a differential spring coupling. The two assemblies act as a fine and a coarse stage for the positioning.

4.1.3.2 Deformable Mirrors

In adaptive optics, small deformable mirrors, made of very thin glass plates driven by piezoelectric or magnetic actuators, play a pivotal role. There are two types of deformable mirrors: one with discontinuous surfaces and the other with a continuous surface. The mirror segment with discontinuous surface may have no tilt degrees of freedom if the surface plate is supported by only one (or one group) piezoelectric actuator. To improve the mirror performance, one segment of surface plate can be supported by three (or three groups) actuators. Some deformable mirrors have continuous thin mirror surfaces. The continuous deformable mirror has a lower ability (4–8 times less) in the wavefront compensation than a segmented deformable mirror with the same number of piezoelectric cells.

There are generally two effects in the piezoelectric actuators: piezoelectric and electrostrictive ones. Piezoelectricity is a property exhibited by only a small number of poled dielectric materials called piezoelectric materials, such as lead zirconate titanate (PZT). Within these materials, the application of displacement (or strain) creates an electric field in the materials, and vice versa. The piezoelectric effect is a linear one, but with high hysteresis.

Different from piezoelectricity, electrostrictivity is a property of the ferroelectric materials, such as lead magnesium niobate (PMN). Electrostrictivity induces displacement through applying an electric field, but not vice versa. The induced displacement by this effect is repeatable and approximately proportional to the square of the applied field. Piezoelectricity is a subclass of ferroelectricity. Therefore, both piezoelectric and electrostrictive effects may exist for piezoelectric materials. Another ferroelectric material is liquid crystal used in phase correctors for adaptive optics.

PZT ($Pb(Zr, Ti)O_3$) exhibits the strongest piezoelectric effect. Its effect is direction oriented, forming a tensor. The effect of a longitudinal electric field E_3 will change the relative thickness to:

$$\frac{\Delta h}{h} = d_{33} E_3 \tag{4.18}$$

where h is the cell height, d_{33} the longitudinal piezoelectric coefficient, and the number 3 refers to the z axis. If voltage $V_3 = E_3 h$ is used, then

$$\Delta h = d_{33} V_3 \tag{4.19}$$

Values of d_{33} are typically 0.3–0.8 μm/kV. If the electric field applied is transversally in the x direction, then:

$$\frac{\Delta h}{h} = d_{31} E_1 \quad \Delta h = \frac{h}{w} d_{31} V_1 \tag{4.20}$$

where w is the x dimension of the cell and d_{31} roughly 3/8 of d_{33}, but with an opposite sign. From these equations, the most efficient cell actuators are formed by a stack of individual cells loaded in the longitudinal direction.

PMN ($Pb(Mg_{1/3}Nb_{2/3})O_3$) exhibits an electrostrictive effect. Compared with PZT actuators, PMN ones require lower voltage for the same amount of displacement stroke. The relative deformation is proportional to the square of the applied field:

$$\frac{\Delta h}{h} = aE^2 = a\left(\frac{V}{h}\right)^2 \tag{4.21}$$

The deformation induced by an electric field of a piezoelectric cell may be a superposition of both piezoelectric and electrostrictive effects. The piezoelectric effect requires a permanently polarized piezoelectric ceramic, while the electrostrictive effect does not require a permanent polarization. The hysteresis of the electrostrictive effect is less (<3%). Therefore the electrostrictive effect is more stable, with less aging. The major drawback of using the electrostrictive effect is its temperature dependence.

Deformable mirrors can also be made by gluing small magnets onto the back of a face sheet. The magnets are driven by printed micro voicecoils on a back plate. Another type of deformable mirror is called the micromirror array. These arrays have a high fill factor (>95%), large stroke (5–10 μm), and small pixel size (<200 μm). The response of the micromirror array is very fast (100 μs).

Bonding two opposite polarized piezoelectric ceramic wafers together with an array of electrodes inside forms a bimorph mirror. The properties of bimorph and membrane mirrors will be discussed in Section 4.1.5. They are mostly used for wavefront curvature compensation.

4.1.3.3 Liquid Crystal Phase Correctors

The liquid crystal phase corrector, or phase modulator, is a newly developed adaptive optics device. It can replace a deformable mirror in an adaptive optics system. Liquid crystal is a state between solid and liquid. In solid state, molecules have both positional and orientational order. In liquid state, both orders vanish. However, liquid crystal material retains an orientational order, but no positional one. With cylindrical molecular shape and ferroelectric property, the material has birefringence. Linearly polarized light parallel to the major axis has one refractive index (ordinary ray), while perpendicular to the axis it has another refractive index (extraordinary ray). The effective refractive index for a polarized light between these two directions is:

$$n(z) = \frac{n_e^* n_o}{(n_e^2 \sin^2\theta + n_o^2 \cos^2\theta)^{1/2}} \tag{4.22}$$

where n_0 and n_e are ordinary and extraordinary ray refractive indexes and θ the angle between the incoming polarized light and the molecular major axis. The refractive index is a complex number, and the sign "*" represents its conjugative.

The liquid crystal phase corrector is similar to the liquid crystal display. The liquid crystal material is sandwiched between two glass plates and the separation is maintained by spacers. On each glass plate, a thin film of material acts as a transparent electrode, usually Indium Tin Oxide (ITO). The inner layer of the plate is an alignment layer which is used to anchor the liquid crystal molecules. In the liquid crystal display, the two face plates have alignment films perpendicular to each other, however, in phase correctors, the two face plates have alignment films parallel to each other. When an electric field is applied between the face plates, the molecular alignment will change as shown in Figure 4.10. The tilt angle is a function of the field intensity, so that a phase modulation for a polarized beam going through the device will be (Restaino, 2003):

$$\Delta\phi = \frac{2\pi}{\lambda}\int_{-d/2}^{d/2}[n(z) - n]dz + <\Delta\phi>_{thermal} \tag{4.23}$$

where d is the thickness of the cell, usually a few microns, and n the refractive index in absence of the electric field. The last thermal fluctuation term is usually

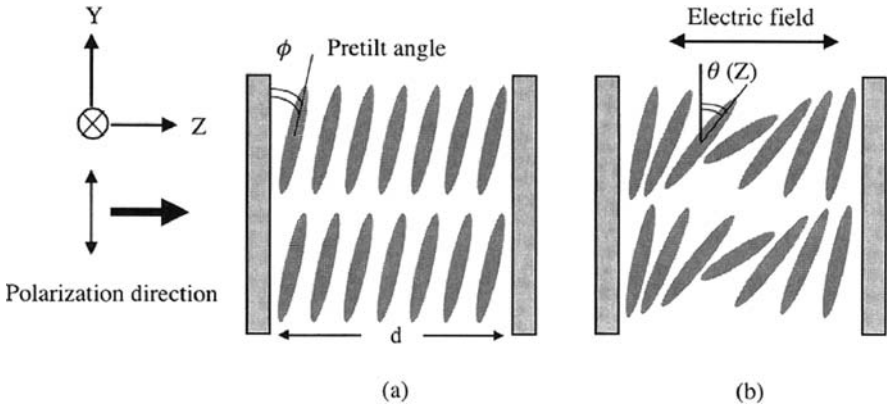


Fig. 4.10. Liquid crystal cell without (**a**) and with electric field (**b**).

negligible, of the order of 1.7×10^{-7} radians. This phase modulator is for polarized beams. If an unpolarized beam is used, two devices are needed. Another way is to put in optical contact a quarter-wave plate between a liquid crystal phase corrector and a mirror. When the light goes through the corrector, one polarized light component is retarded. After being reflected from the mirror, the other polarized light component will be retarded.

4.1.3.4 Metrology Systems

An active and adaptive optics system relies on wavefront sensors to form the system control loop. Without wavefront sensors, the local feedback signal is from metrology systems. A metrology system is formed by displacement, tilt, acceleration, path length, or temperature sensors. For radio telescopes, an optical guiding telescope may be a part of the metrology system. The laser ranger and quadrant detector metrology systems are discussed in Section 7.2.4. Thermal sensors, accelerometers, and tilt meter are discussed in Section 8.2.4.

In a segmented mirror telescope, displacement sensors are used between mirror segments. The Keck telescopes use a special capacitance displacement sensor between segments (Figure 4.11). The sensor is made of low expansion ceramic block and coated with metal film. The displacement is measured from the capacitance of the sensor. Similar capacitance sensors are also used in an adaptive secondary mirror system (ref. Section 4.1.9).

An improved capacitance sensor involves metal films on both sides of segments, one with a larger square area and the other with two vertically placed small rectangles. The displacement is measured from differential capacitance between two capacitors. It provides tilt information as well.

A similar inductive sensor also involves printed coils on both sides of the segments, one being a big coil and the other two coils placed vertically. When

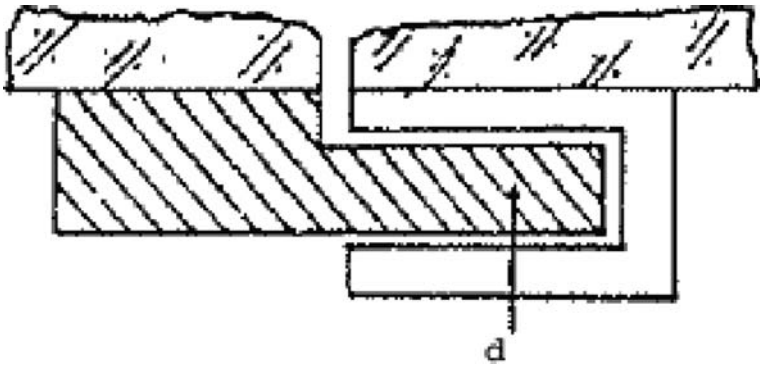


Fig. 4.11. A capacitance displacement sensor for a segmented mirror system.

both mirror segments are aligned, no voltage is generated from the big coil as alternative currents flow through two small coils in opposite direction. This sensor can tolerate a humid environment.

For a multi-mirror telescope, not only the position, but also the optical path of each mirror influences the interference. Therefore, specially designed phasing sensors are used. One sensor used in the old MMT has a laser light source located at the common focus and an inverted paraboloidal small mirror reflects light to the mirror focus. The sensor can generate interference fringes at the focus of the inverted mirror. From the fringe map, the phase difference between mirrors can be determined (Figure 4.12).

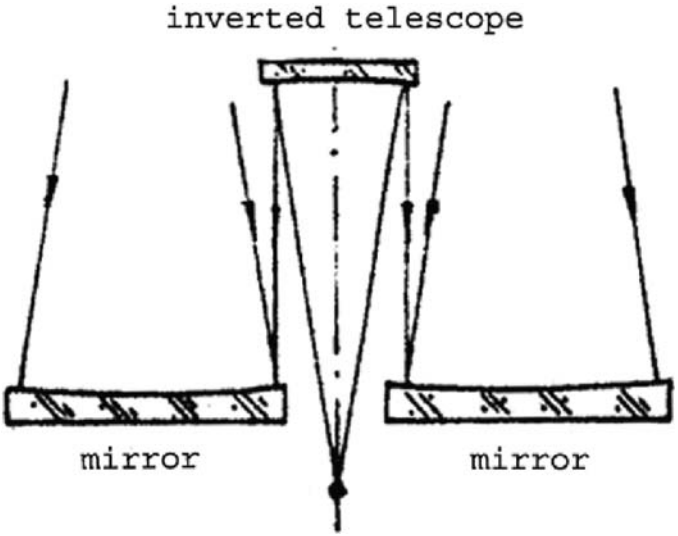


Fig. 4.12. An inverted mirror used for phase measurement between mirrors in the old MMT telescope.

Some sensors, such as force, displacement, strain, or angular ones, may be imbedded within actuators. These actuators form feedback control loops themselves.

4.1.4 Active Optics System and Phasing Sensors

4.1.4.1 Monolithic Mirror Active Optics

There are two major active optics systems for optical telescopes: one for a monolithic mirror with force and moment actuators and the other for a segmented mirror with displacement actuators.

The closed loop of a monolithic mirror active optics includes wavefront sensor, control computer, and actuators on the primary and secondary mirrors (Figure 4.13). The wavefront sensor provides phase error information which is usually expressed as Zernike polynomials where all low order terms with physical meanings (Table 4.1) are considered while high order terms are usually ignored as the number of actuators involved is limited. From the error expression, the computer produces the required mirror adjustment from the classic thin plate theory and the actuators produce the required mirror surface or mirror separation changes to compensate the wavefront error. The loop is therefore closed.

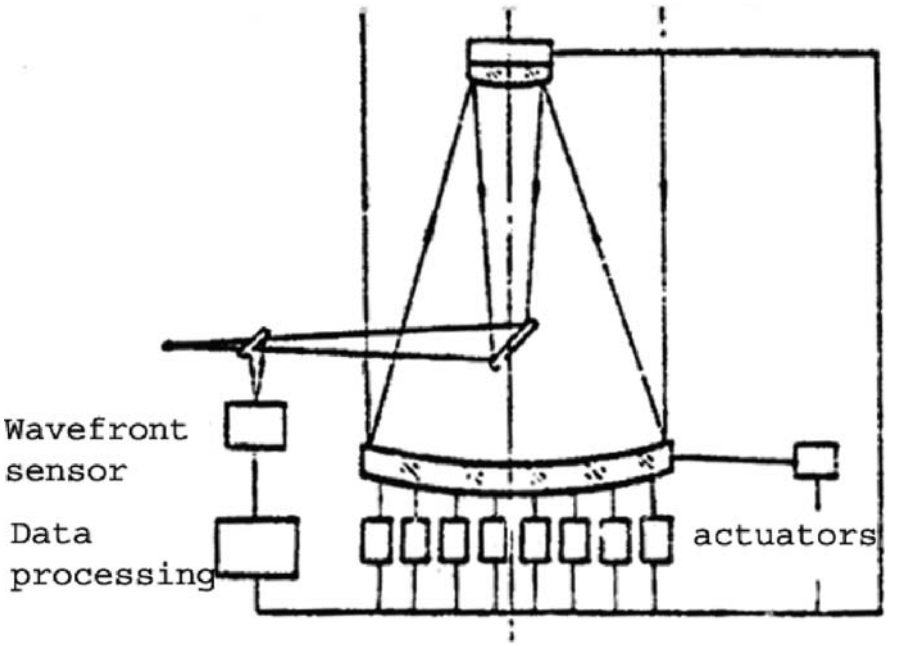


Fig. 4.13. Active control system for monolithic mirror telescopes.

Table 4.1. Physical meanings of the low mode terms of Zernike polynomials

Term	Meanings	Causes
$Z_1 = 1$	Constant	
$Z_2 = 2r\cos\theta$	Tilts	Lateral focusing
$Z_3 = 2r\sin\theta$	Tilts	Lateral focusing
$Z_4 = \sqrt{3}(2r^2 - 1)$	Defocus	Axial focusing
$Z_5 = \sqrt{6}r^2\sin 2\theta$	Astigmatism	Axial support
$Z_5 = \sqrt{6}r^2\cos 2\theta$	Astigmatism	Axial support
$Z_7\sqrt{8}(3r^3 - 2r)\sin\theta$	Coma	Lateral support
$Z_7\sqrt{8}(3r^3 - 2r)\cos\theta$	Coma	Lateral support
$Z_9 = \sqrt{8}r^3\sin 3\theta$	Coma	Axial defining point
$Z_{10} = \sqrt{8}r^3\sin 3\theta$	Coma	Axial defining point
$Z_{11} = \sqrt{5}(6r^4 - 6r^2 + 1)$	Spherical	Axial support

Table 4.1 lists low order terms and their physical meanings of the Zernike polynomials. The normalized Zernike polynomials defined by Noll (1976) are:

$$\begin{aligned} Z_{even,j}(m \neq 0) &= \sqrt{n+1}\cdot R_n^m(r)\sqrt{2}\cos m\theta \\ Z_{odd,j}(m \neq 0) &= \sqrt{n+1}\cdot R_n^m(r)\sqrt{2}\sin m\theta \\ Z_j(m = 0) &= \sqrt{n+1}\cdot R_n^0(r) \\ R_n^m(r) &= \sum_{s=0}^{(n-m)/2} \frac{(-1)^s(n-s)!}{s![(n+m)/2-s]![(n-m)/2-s]!} r^{n-2s} \end{aligned} \tag{4.24}$$

where n is the radial degree and m the azimuth frequency. The values of n and m are always integral and satisfy $m \leq n$, $n - |m| = even$. The coefficients of radial polynomials R_n^m are zero when $n - |m| = odd$. The index j is a mode ordering number and is a function of n and m. All the modes of Zernike polynomials are orthogonal each other and each of them represents a particular surface shape within a circle.

4.1.4.2 Segmented Mirror Active Optics

Active optics control for a segmented mirror telescope is different. The main purpose is to ensure the correct position of each segment in the system. For a telescope with N segments, there are in total $6N$ degrees of freedom. However, within each segment, the effects from two lateral translations and the in-plane rotation are much smaller than the other three degrees of freedom (i.e., tip, tilt, and piston). The number of all these three important degrees of freedom of segments is $3N$. However, three are global tip, tilt, and piston which determine the mirror's position in space. Therefore, only $3N - 3$ degrees of freedom are left for the segment positional adjustment.

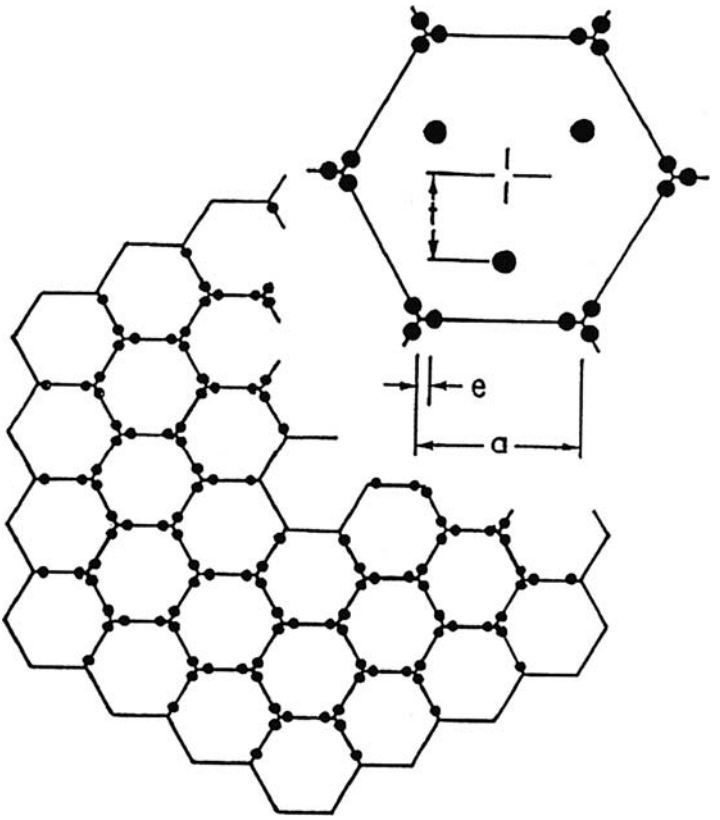


Fig. 4.14. Sensor (between segments) and actuator (within segment) distribution of the 10 m Keck telescope (Nelson et al., 1985).

The Keck telescope mirror has a total of 105 degrees of freedom and it has 168 edge displacement sensors (Figure 4.14). As the number of displacement sensors is larger than the number of degrees of freedom involved, a least-square reconstruction process can be used to derive the required actuators' position.

If all segments were at their neutral positions, the reading of all displacement sensors would be zero. If all of the segments are pointed in different directions, there would be N image spots in the focal plane. The root mean square of the image spot size can be used as a rough estimation of the quality of the mirror segment control.

Assuming the entire mirror is a flat plane in the x-y plane, the reading S_j of each displacement sensor is the z coordinate difference between neighboring segments:

$$S_j = \Delta Z \tag{4.25}$$

When a segment is in its neutral position, the z coordinate of the three actuators relative to a neutral position will all be zero, $P_{ij} = 0$, where the

subscript j is the segment number and $i = 1, 2, 3$ represents three actuator positions of a segment. When the three actuators have position errors, the height of each point on the segment will be:

$$Z_j = \alpha_j x + \beta_j y + \gamma_j t \tag{4.26}$$

where,

$$\begin{aligned} \alpha_j &= \tfrac{1}{3t}(2P_{1j} - P_{2j} - P_{3j}) \\ \beta_j &= \tfrac{1}{\sqrt{3}t}(P_{2j} - P_{3j}) \\ \gamma_j &= \tfrac{1}{3t}(P_{1j} + P_{2j} + P_{3j}) \end{aligned}$$

where t is the distance between actuator position and the segment center (Figure 4.14). The relationship between the reading of each displacement sensor and the actuator positions is:

$$S_j = \sum_n A_{jn} P_n \tag{4.27}$$

where A_{jn} is a constant matrix determined by the support geometry, j the number of the displacement sensor, and n the number of the displacement actuator. A_{jn} is a sparse matrix in which most of its elements are zero. By solving Equation (4.27) or calculating the inverse of coefficient matrix, the displacements required for the mirror position correction are obtained.

In the actual control process, the inverse of the sum matrix A_{jn} is stored inside the computer. The inverse of the coefficient matrix can be done through a singular value decomposition process. If S_j is the measurement, P_n will be calculated in real time. The calculated actuator position together with information from the wavefront sensor is used for segmented mirror surface control.

Usually there are different requirements in different stages of active optics for a segmented mirror telescope. In the early stage, co-focus and co-alignment of all segments is required. Image spot centroiding or spot size can be used to guide the segment tilt correction and the segment co-focusing. Single segment wavefront sensor can also be used to control the tip-tilt of the segments. In this stage, interference is not possible between mirror segments. In the later stage, a demanding co-phasing of the segments is required. In this stage, the piston error between segments is of vital importance. The phasing sensors discussed in the following sub-sections are used for this purpose. After the piston errors are detected, the segment will move using the displacement actuators. This iteration may continue to achieve the best surface rms error.

Without atmospheric turbulence, the Airy disk of a co-focus segmented mirror telescope is determined by the mirror segment diameter, while the Airy disk of a co-phase telescope is determined by the whole mirror diameter. The stellar image in the latter case is much sharper than in the first case.

4.1.4.3 Dispersed Fringe Phasing Sensor

To achieve co-phase of a segmented primary mirror, newly developed phasing sensors, which measure the piston phase error between two aligned mirror segments, are of vital importance. Phasing sensors are generally used after segment tip-tilt adjustment. All the mirror segments are parallel each other and only piston errors exist.

One phasing sensor, the dispersed fringe sensor (DFS) based on a transmissive grism, is used for the coarse phasing stage (Shi, 2003). The grism is a combination of a prism and a grating with the grating on one side of the prism.

The dispersed fringe sensor can measure a piston error more than 2π radians between two aligned mirror segments. This is not possible for any other wavefront sensors as the information more than 2π radians will overlap the error portion within the 2π radians.

The dispersed fringe sensor is realized by inserting a grism in the imaging optical path after a collimator (Figure 4.15). For a transmissive grism, the wavelength dispersion along the dispersion direction x is:

$$\lambda(x) = \lambda_0 + \frac{\partial \lambda}{\partial x} = \lambda_0 + C_0 x \tag{4.28}$$

where λ_0 is the central wavelength and C_0 the linear dispersion coefficient. If two mirror segments have only a piston wavefront error, coherent addition of the two beams will result in an intensity modulation within the spectrum. The intensity $E(x)$ at any point along the dispersion direction is a sum of the fields from the dephased segments:

$$\begin{aligned} E(x) =& E_1 e^{i[(2\pi/\lambda(x))\cdot L]} + E_2 e^{i[(2\pi/\lambda(x))\cdot(L+\delta L)]} \\ =& E e^{i[(2\pi/\lambda(x))\cdot L]} \left(1 + e^{i[(2\pi/\lambda(x))\delta L]}\right) \end{aligned} \tag{4.29}$$

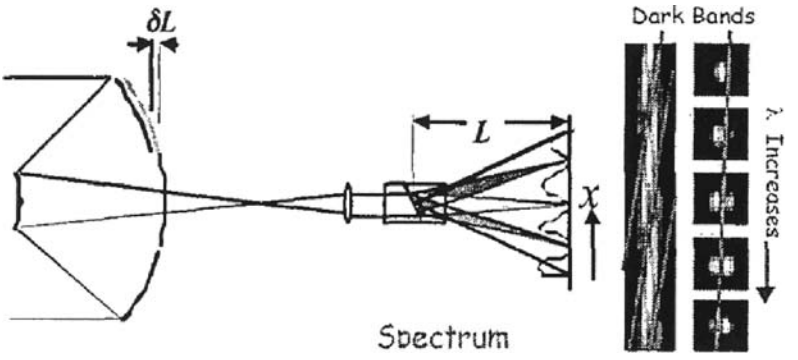


Fig. 4.15. Arrangement of a dispersed fringe sensor (Redding et al., 2000).

where E_1 and E_2 are the amplitude of the light reflected from the two segments, $E_1 = E_2 = E$ when they come from two equal areas of mirror segments. L is diffraction length and δL is the optical path difference between the two mirror segments.

Along the dispersed spectrum, the wavefronts may add constructively or destructively, depending on the local wavelength and the piston error. Periodic dark fringes are formed within the point spread function (PSF) when the destructive condition is met. Along the dispersion direction, the fringe intensity pattern has the form:

$$I(x, y) = I_0 \left[1 + \gamma \cos(2\pi \frac{\delta L}{\lambda(x)} + \phi_0(y)) \right] \tag{4.30}$$

where $I(x, y)$ is the dispersed fringe intensity along the dispersion direction x, γ the fringe visibility, and ϕ_0 a phase constant depending on where the dispersed fringe sensor is extracted along the y direction. The fringe orientation depends on segment gap direction and the best fringe contrast happens when the dispersion direction is parallel to the inter-segment edge. The pattern shows that a larger piston error will cause more fringes. A very small piston error may produce an incomplete fringe pattern period. Usually a number of rows of the pattern are averaged to determine the parameters of the right-hand side terms of the equation. When the segments are co-phased, the wavefront is coherently added for the entire wavelength and the spectrum will not show any modulation. However, a very small piston error can still be detected by adding a predetermined piston phase difference, so that the total phase error is large enough to be detected.

The coarse co-phasing requires that the maximum detectable piston is at least one depth of focus, which is $\pm 2\lambda F^2$, where F is the focal number of the mirror.

In practice, however, the resulting spectral pattern from the dispersed fringe sensor is usually not an ideal cosine curve. The signal is affected by the grism spectrum efficiency, the wavelength dependent detector efficiency, and the spectrum uniformity of the light source. All these are unwanted spectral features. Therefore, it is necessary to find a calibration spectrum which is formed by summing up all of the intensities across the measured spectrum. By removing spectrum defects, a clean dispersed fringe sensor spectrum will be produced. The dispersed fringe technique is the baseline for coarse phasing in JWST wavefront sensing and control.

4.1.4.4 Template Phasing Sensor

The template phasing sensor, also named the phasing camera, is an earlier sensor used for estimating the piston error between segments in the Keck telescopes (Chanan et al., 1998, 2000). The sensor is basically a Shack–Hartmann instrument with its lenslet array preceded by a mask at the exit pupil. This mask defines small circular sub-apertures at the centers of the inter-segment edges by

cross hairs. The alignment of these sub-apertures is crucial. The diameter of these sub-apertures is about 12 cm, smaller than the atmospheric coherence diameter of about 20 cm at a wavelength of 500 nm. A cross hair of 30 mm wide is across each sub-aperture providing a dead band between segments. This ensures that the results will be insensitive to atmospheric turbulence in all conditions. The resulting interference pattern is compared with a set of calculated theoretical diffraction patterns. The lenslet array can also be replaced by a prism array or a combination of an objective lens and a prism array. The latter combination has a better image quality and large focal ratio.

Let ρ be the position vector in the sub-aperture plane and w the position vector in the image plane, in the sub-aperture plane, (η, ξ) forms a rectangular coordinate. If one half of the sub-aperture has a piston error of $\delta/2$ and the other half has a piston error of $-\delta/2$, the complex aperture field within the sub-aperture is:

$$P(\rho, k\delta) = \begin{cases} \exp(ik\delta) & \eta \geq 0 \\ \exp(-ik\delta) & \eta < 0 \end{cases} \tag{4.31}$$

where $k = 2\pi/\lambda$. The image complex function is simply the Fourier transform of the aperture field:

$$\begin{aligned} A(w, k\delta) = & \frac{1}{\pi a^2} \int_0^{\pi} \int_0^{a} \exp(ik\delta) \exp(ik\rho \cdot w) \rho d\rho d\theta \\ & + \frac{1}{\pi a^2} \int_{-\pi}^{0} \int_0^{a} \exp(-ik\delta) \exp(ik\rho \cdot w) \rho d\rho d\theta \end{aligned} \tag{4.32}$$

where a is the radius and (ρ, θ) circular coordinates on the sub-aperture. If small effects due to the aperture cross hairs are ignored, then the imaginary part in the image field vanishes and we have:

$$A(w, k\delta) = \frac{1}{\pi a^2} \int_0^{\pi} \int_0^{a} \cos(k\delta + k\rho \cdot w) \rho d\rho d\theta \tag{4.33}$$

The intensity in the image plane is:

$$I(w, k\delta) = A^2(w, k\delta) \tag{4.34}$$

When both segments are in phase, the intensity in the image plane is:

$$I(w, 0) = \left[\frac{2J_1(kaw)}{kaw} \right]^2 \tag{4.35}$$

If the two segments are out of phase by $\delta = \lambda/4$, in which case $k\delta = \pi/2$, and:

$$A\left(w, \frac{\pi}{4}\right) = \frac{2}{\pi} \int_0^{\pi} \frac{u \cos u - \sin u}{u^2} d\theta$$
$$u = kaw \cos(\theta - \psi) \tag{4.36}$$

where (w, ψ) is the circular coordinates in the image plane. For an arbitrary δ, the image can be expressed from the above two expressions:

$$I(w, k\delta) = \left[(\cos k\delta) A(w, 0) + (\sin k\delta) A(w, \frac{\pi}{2})\right]^2 \tag{4.37}$$

Figure 4.16 shows the theoretical diffraction patterns for a sequence of 11 equally spaced values of δ. With these patterns stored in the computer, the

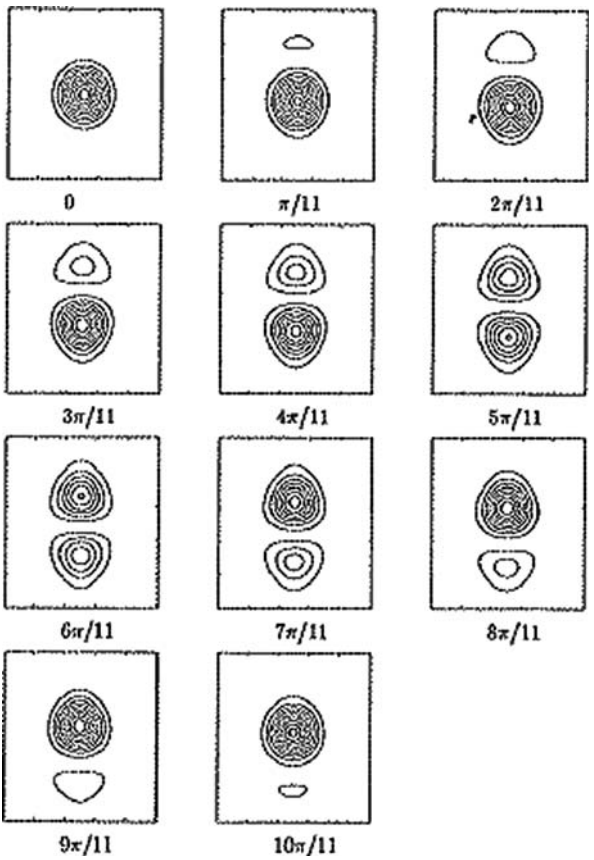


Fig. 4.16. Theoretical narrowband diffraction patterns for a split circular aperture with a physical step (Chanan et al., 1998).

sensor could determine the piston error by comparing with this template. In the above discussion, narrowband monochromatic light is assumed.

The diffraction pattern of the above equation is a periodic function with a period of $\lambda/2$. Therefore, there is an ambiguity of half of a wavelength.

If the light used is spanned over a finite wavelength interval $\Delta\lambda \approx 2\pi\Delta k/k$ and the condition of $\Delta\lambda << \lambda^2/2\delta$ is violated, the Equation (4.36) has to be integrated over k. The simple case is for a relatively small wavelength spanning, $\Delta\lambda << \lambda$. In this case, the $A(w, k\delta)$ functions can be evaluated at the midpoint of the k interval. The triangular functions have to be explicitly averaged over k. If the bandwidth in k is a Gaussian one:

$$g(k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(k-k_0)^2}{2\sigma_k^2}\right) \tag{4.38}$$

In this distribution, the full width of half maximum (FWHM) is $\Delta k = [8\ln(2)]^{1/2}\sigma_k$. To perform the k average, we obtain:

$$\begin{aligned} <I(w, k\delta)> &= \alpha_1 A^2(w, 0) + \alpha_2 A(w, 0) A\left(w, \frac{\pi}{2}\right) + \alpha_3 A^2\left(w, \frac{\pi}{2}\right) \\ \alpha_1 &= \frac{1}{2}[1 + \exp(-2\sigma_k^2\delta^2)\cos 2k_0\delta] \\ \alpha_2 &= \exp(-2\sigma_k^2\delta^2)\sin 2k_0\delta \\ \alpha_3 &= \frac{1}{2}[1 - \exp(-2\sigma_k^2\delta^2)\cos 2k_0\delta] \end{aligned} \tag{4.39}$$

If $\sigma_k\delta \to 0$, this leads to the narrowband situation. If $\sigma_k\delta \to \infty$, the intensities from two semicircular sub-apertures simply add incoherently. We obtain:

$$I(w, \infty) = \frac{1}{2}\left[I(w, 0) + I(w, \frac{\pi}{2})\right] \tag{4.40}$$

In this broadband equation, the diffraction pattern washes out when $\sigma_k\delta$ becomes significant. The length scale for the piston error in this case, named coherence length, is:

$$l = \lambda^2/2\Delta\lambda = 1.334/\sigma_k \tag{4.41}$$

For a bandwidth of 10 nm and a central wavelength of 891 nm, the coherence length is $l = 40\ \mu$m. Figure 4.17 shows a typical broadband sequence of the image patterns. The boxes are 4 arcsec on each side. Using a narrowband or a broadband pattern template, the segment piston error can be detected.

Another way to increase the detecting range is to use two narrowband measurements of slightly different wavelengths. This is similar to the encoder technique discussed in Section 3.3.3. The disadvantage of this technique is that the mask of sub-apertures has to align accurately with the segment edge. Otherwise the pattern detected will not match the template.

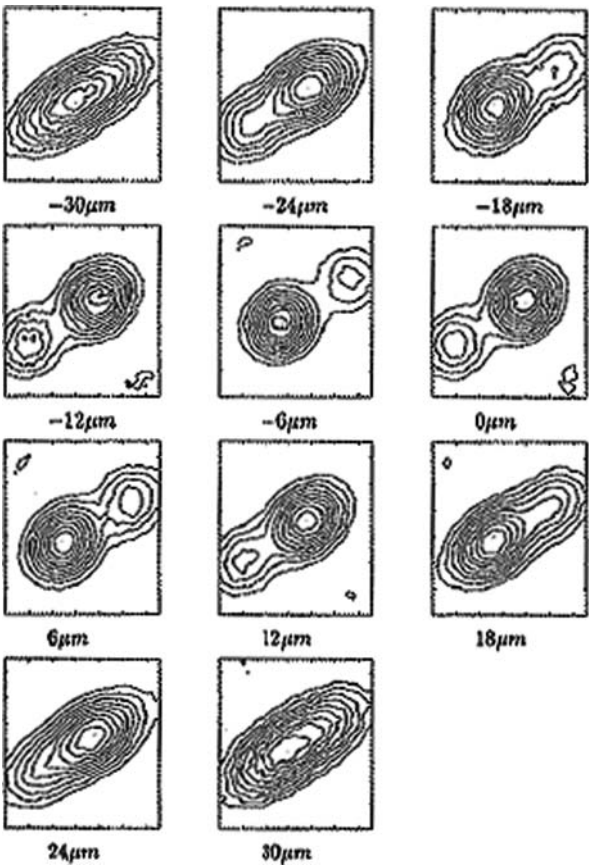


Fig. 4.17. Theoretical broadband diffraction patterns for a split circular aperture with a physical step (Chanan et al., 1998).

4.1.4.5 Young–Shack–Hartmann Phasing Sensor

The Young–Shack–Hartmann phasing sensor was developed from the Young–Hartmann sensor [Figure 4.18(a)] which is based on a Hartmann array. In this Young-Hartmann sensor, an aperture plate is placed in the pupil plane followed by diffractive optics, such as a grating. The central beam (zero order), which is used to get the wavefront slope information, is filtered out and the first order beams from two adjacent segments overlap to create a Young's interference pattern as:

$$I(x) = 1 + \cos(2\pi\delta/\lambda) = 1 + \cos\phi \tag{4.42}$$

If two closely spaced light wavelengths are used, the piston error between segments can be calculated from the two patterns. The advantages of using a

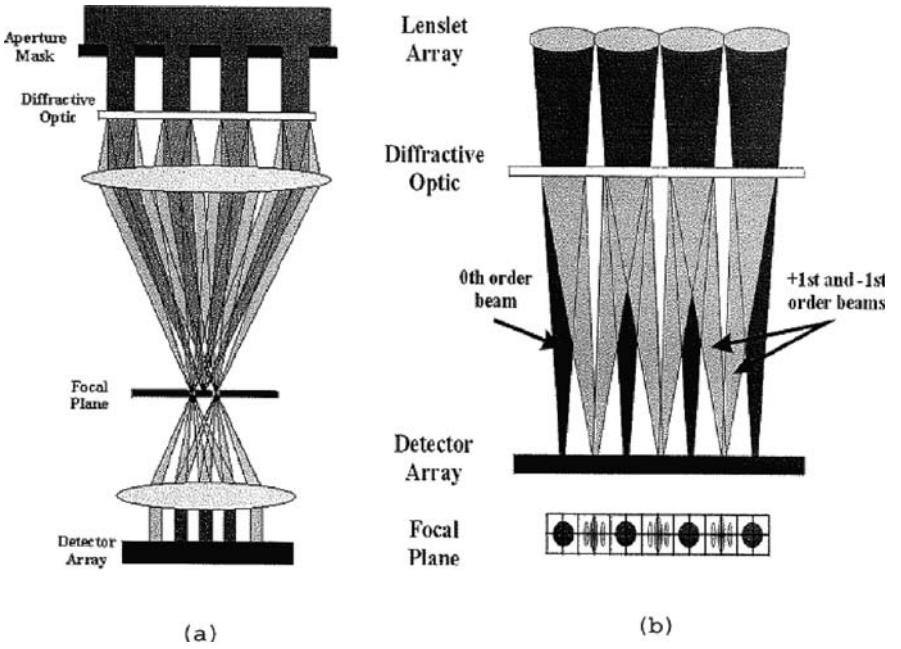


Fig. 4.18. (**a**) Young–Hartmann sensor and (**b**) Young–Shack–Hartmann sensor (Walker et al., 2001).

Young–Hartmann sensor are simplicity in calculation, larger segment areas sampled, and relaxed alignment tolerance. However, in the co-phase process, it is necessary to use different pairs of wavelengths for course and fine phase alignment. There are also other error sources which will alter the diffraction patterns and make the calibration difficult.

A Young–Shack–Hartmann sensor is similar to the above concept (Walker et al., 2001) where diffractive optics follows the lenslet array in the pupil plane [Figure 4.18(b)]. Therefore, the zero order beams are used for tip-tilt error measurement and the first order beams from adjacent segments overlap to create an interference pattern for the piston error measurement.

The same as the Young–Hartmann sensor, the piston error detection can be achieved by using a pair of different wavelengths of light.

4.1.4.6 Mach–Zehnder Phasing Sensor

A Mach–Zehnder phasing sensor as shown in Figure 4.19 (Yaitskova et al., 2005) has its beam split into two and one pinhole is placed in the focal plane of one arm acting as a spatial filter. The beam after the pinhole acts as a reference wave which is coherent with the other beam. The size of the pinhole is about that of the seeing disk so that the beam becomes a lower pass filtered version of the original beam.

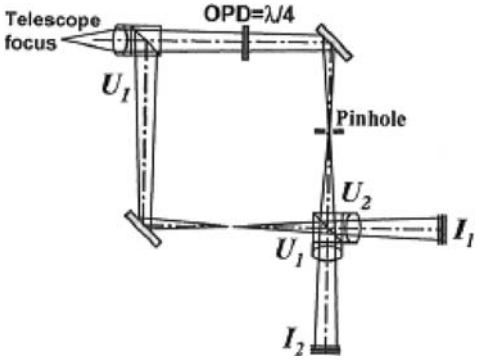


Fig. 4.19. Schematic representation of a Mach–Zehnder phasing sensor (Yaitskova et al., 2005).

If $U_1(\xi)$ and $U_2(\xi)$ are complex amplitudes of the two beams in the pupil plane, then the complex amplitudes in the focal plane are the Fourier transforms $u_1(w)$ and $u_2(w)$. In the reference beam, a filter is used so that:

$$U_2(\xi) = \frac{1}{\lambda}\int u_1(w)t(w)\exp\left(-i\frac{2\pi}{\lambda}w\xi\right)d^2w \tag{4.43}$$

where λ is the wavelength and $t(w)$ the filter function in the focal plane. The Fourier transform of $t(w)$ is $T(\xi)$. The above equation is equivalent to:

$$U_2(\xi) = \frac{1}{\lambda}\int U_1(\xi')T(\xi-\xi')d^2\xi' \tag{4.44}$$

The two output beams of the sensor will be:

$$\begin{aligned} I_1(\xi) &= \frac{1}{2}\{|U_1(\xi)|^2+|U_2(\xi)|^2+2\mathrm{Re}[U_1^*(\xi)U_2(\xi)\exp(i\theta)]\} \\ I_2(\xi) &= \frac{1}{2}\{|U_1(\xi)|^2+|U_2(\xi)|^2-2\mathrm{Re}[U_1^*(\xi)U_2(\xi)\exp(i\theta)]\} \end{aligned} \tag{4.45}$$

where θ is a phase shift between two arms, and Re is the real part of the complex number. The sensitivity of the instrument can be improved by using the difference between two output beams which is the signal of the instrument. If the filter function (pinhole) is a Gaussian one with the FWHM being a, the shape of a hole aperture field is:

$$t(w) = \exp\left[-(2\sqrt{\ln 2}w/a)^2\right] \approx \exp\left[-(w/0.6a)^2\right] \tag{4.46}$$

In a classical implementation of the Mach–Zehnder interferometer, the pinhole is smaller than the Airy disk ($a << \lambda/D$), where D is the pupil diameter. Then the diffracted wave is a real function:

$$U_2(x) \approx \; u_1(0)\frac{1}{l\sqrt{\pi}}\exp\left[-(x/l)^2\right] = D\sqrt{St}\frac{1}{l\sqrt{\pi}}\exp\left[-(x/l)^2\right] \qquad l = \lambda/0.6\pi a \tag{4.47}$$

where St is the Strehl ratio. If the incoming beam has a phase distribution as $\phi(x)$, then:

$$U_1^*(x)U_2(x) \approx D\sqrt{St}\frac{1}{l\sqrt{\pi}}\exp[-i\phi(x)] \tag{4.48}$$

The signal is:

$$S(x) \approx D\sqrt{St}\frac{1}{l\sqrt{\pi}}\cos[\theta - \phi(x)] \tag{4.49}$$

For a fixed delay between two arms, $\theta = \pi/2$, and a small phase value $\phi(x)$,

$$S(x) \approx D\sqrt{St}\frac{1}{l\sqrt{\pi}}\phi(x) \tag{4.50}$$

In the case of a very large pinhole and $a >> \lambda/D$, the situation is different. If $U_1(x)$ can be expressed as a Taylor series:

$$U_1(x') = U_1(x) + \sum \frac{1}{n!}\frac{d^n U_1}{dx^n}\bigg|_x (x' - x)^n \tag{4.51}$$

Therefore, $U_2(x)$ is:

$$U_2(x) = U_1(x) + \frac{1}{l\sqrt{\pi}}\sum \frac{1}{n!}\frac{d^n U_1}{dx^n}\bigg|_x \int_{-\infty}^{\infty} \exp[-(x'/l)^2]\cdot(x')^n dx' \tag{4.52}$$

In the last sum of this expression, all the odd terms are zero. The integral of all the even terms with $n = 2m$ is proportional to l^{2m}. If the $U_1(x)$ function does not change significantly within the domain l, we can retain only two terms in the sum:

$$U_2(x) = U_1(x)\left\{1 - \frac{l^2}{4}\left(\frac{d\phi}{dx}\right)^2 + i\frac{l^2}{4}\frac{d^2\phi}{dx^2}\right\} \tag{4.53}$$

The signal becomes:

$$S(x) = \left[2 - \frac{l^2}{2}(\frac{d\phi}{dx})^2\right]\cos\theta - \left(\frac{l^2}{2}\frac{d^2\phi}{dx^2}\right)\sin\theta \tag{4.54}$$

This expression shows that when the phase delay is zero between two arms, the instrument measures the square of wavefront slope; when the phase delay is $\pi/2$, it measures the wavefront curvature. The condition for applying the above formula to segmented mirror systems is $a >> \lambda/0.6\pi d$, where d is the segment size.

When a piston phase error exists between segments, the signal is:

$$\begin{aligned} S(x) = &\left\{\sin(\Delta\phi)sign(x)\left[1 - \Phi\left(\frac{0.6\pi a|x|}{\lambda}\right)\right]\right\}\sin\theta \\ &- \left\{2 - [1 - \cos(\Delta\phi)]\left[1 - \Phi\left(\frac{0.6\pi a|x|}{\lambda}\right)\right]\right\}\cos\theta \end{aligned} \tag{4.55}$$

where $\Delta\phi$ is the phase jump between segments and $\Phi(x)$ is an error function. For a $\pi/2$ phase delay between two arms, the expression is:

$$\begin{aligned} S(x) &= \sin(\Delta\phi)sign(x)\left[1 - \Phi\left(\frac{0.6\pi a|x|}{\lambda}\right)\right]; \\ \Phi\left(\frac{0.6\pi a|x|}{\lambda}\right) &= \frac{1.2\pi a}{\lambda\sqrt{\pi}}\int_0^x \exp\left[-\left(\frac{0.6\pi a}{\lambda}\right)^2 x'^2\right]dx' \end{aligned} \tag{4.56}$$

If the segment edge is at $x = 0$, the signal will have the shape as in Figure 4.20 for two different pinhole functions. However, limitations exist in this sensor arrangement, including pinhole position relative to the inter-segment boundary, presence of gap, piston, tip, and tilt errors between segments, rolled-off segment edges, and atmospheric distortion.

4.1.4.7 Pyramid Phasing Sensor

As a wavefront sensor, a pyramid prism can also be used as phasing sensors with a nonambiguous range of piston error between segments being $\pm\lambda/4$. If a subaperture corresponds to a physical mirror step δ, the wavefront sensor signal is (Pinna et al., 2006):

$$S = C + A\sin[2\pi\frac{2\delta}{\lambda}] \tag{4.57}$$

where λ is the wavelength, C and A are constants. If another wavelength of $\lambda + \Delta\lambda$ is used for the signal difference measurement, the equivalent wavelength is extended to $\lambda(\lambda+\Delta\lambda)/\Delta\lambda$. This is usually many times the wavelengths used. This is the base for the pyramid phasing sensor. The pyramid phasing sensor

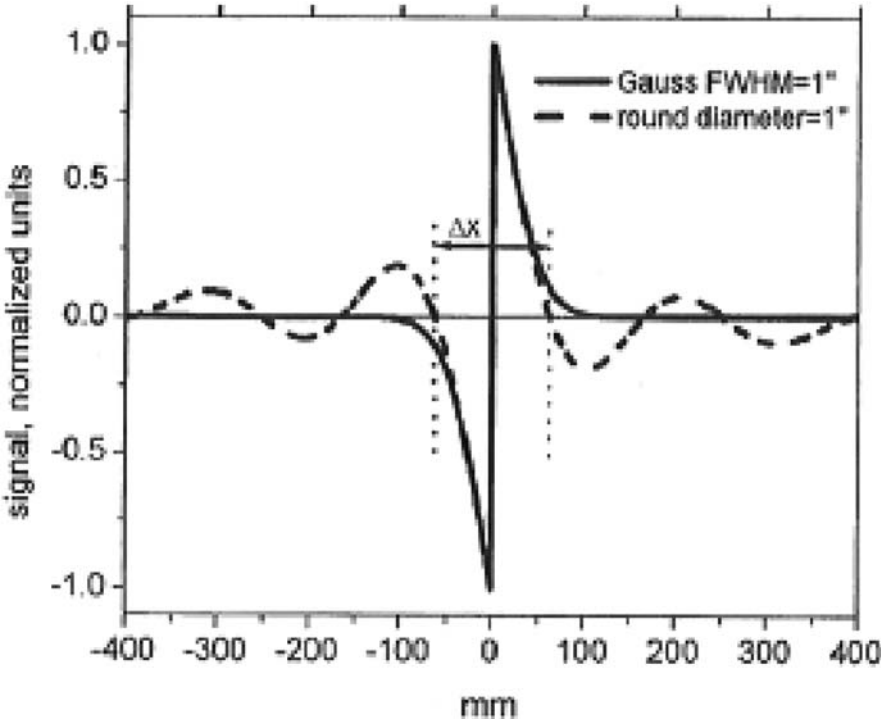


Fig. 4.20. Mach–Zehnder signal at the edge between two segments due to 1/4 wavelength piston steps. The *solid line* is for a Gaussian pinhole shape and the *dotted line* is for a top-hat pinhole shape (Yaitskova et al., 2005).

can also be realized by wavelength or segment sweeping. The advantages of a pyramid phasing sensor are the tip-tilt terms of segments are also detected at the same time.

4.1.5 Curvature Sensors and Tip-Tilt Devices

Wavefront correcting devices for active optics are usually acted on the primary mirror of a lower resonant frequency. Compared with adaptive optics which is diffraction limited, a poor Delivered Image Quality (DIQ) is expected for the active optics system. The DIQ equals the FWHM of the image energy spread in a unit of arcsec.

To improve the telescope DIQ, it is necessary to compensate wavefront errors caused by the atmospheric turbulences. Without a complex and expensive deformable mirror, the wavefront compensation is difficult. However, if only the first few terms of the wavefront error induced by the atmospheric turbulence are considered, the phase compensation task is a lot easier. A fast responding small tip-tilt mirror, or membrane mirror, or steerable secondary mirror will be much

Table 4.2. Zernike–Kolmogorov residual errors $\Delta_J = x_j(D/r_0)^{5/3}$

$x_1 = 1.0299$	$x_2 = 0.582$	$x_3 = 0.134$
$x_4 = 0.111$	$x_5 = 0.0880$	$x_6 = 0.0648$
$x_7 = 0.0587$	$x_8 = 0.0525$	$x_9 = 0.0463$
$x_{10} = 0.0401$	$x_{11} = 0.0377$	$x_{12} = 0.0352$
$x_{13} = 0.0328$	$x_{14} = 0.0304$	$x_{15} = 0.0279$
$x_{16} = 0.0267$	$x_{17} = 0.0255$	$x_{18} = 0.0243$
$x_j \approx 0.2944\ \mathrm{J}^{-\sqrt{3}/2}$ (For large J)		

easier and the reduction of residual rms atmospheric wavefront error will be significant (Table 4.2 and Section 4.1.6).

To apply tip-tilt compensation, full knowledge of the wavefront information is not required. A special type of image plane wavefront sensor is used. It provides wavefront Laplacian together with the slope on the aperture edge and is named the curvature sensor because the Laplacian of a surface represents its curvature. Although the wavefront function can be reconstructed through solving a Poisson equation, the wavefront error correction task can be easily done by special groups of deformable mirrors: tip-tilt and membrane ones when curvature information alone is available. A system with curvature sensors and tip-tilt or membrane mirrors is a tip-tilt device, which is a system between active optics and adaptive optics. It improves greatly the DIQ of a large infrared telescope ($\geq$4 m) and can have a large influence on the DIQ of small optical telescopes. It can compensate wind-induced vibrations, thermal effects, and significant effects of atmosphere induced wavefront disturbances.

4.1.5.1 Dual-Image Curvature Sensors

Roddier (1988) and Roddier and Roddier (1993) proposed the wavefront curvature measurement through intra- and extra-focal (before and after focus) images (Figure 4.21). If I_1 and I_2 are the irradiance distributions in intra- and extra-focal planes with their distance to focus being l, then the irradiance distributions represents the irradiances in two conjugated off pupil planes of object space, one before and the other after the pupil plane. The distance from two conjugated planes to the pupil is $\Delta z = f(f - l)/l$, where f is the telescope focal length.

From the irradiance transmission theory, when parallel light propagates along axis z, its luminosity variation follows the equation:

$$\frac{\partial I}{\partial z} = -(\nabla I \cdot \nabla W + I\nabla^2 W) \tag{4.58}$$

where $I(x, y, z)$ is the luminosity distribution and $W(x, y)$ the wavefront phase function. The luminosity is nearly uniform with a value of I_0 inside the pupil and

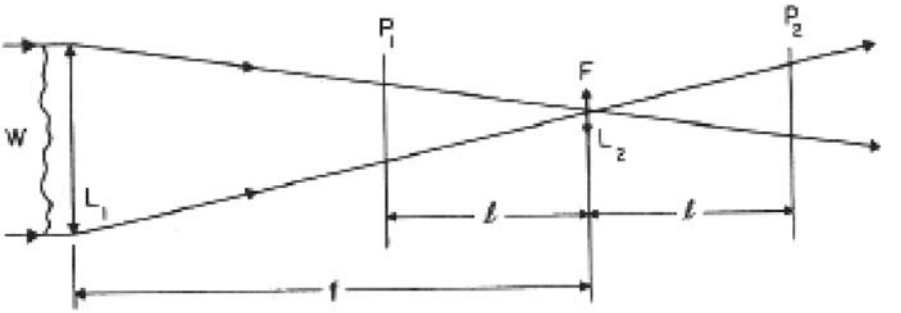


Fig. 4.21. The illumination difference between planes P1 and P2 is a measure of the local curvature distribution of the incoming wavefront (Roddier, 1988).

zero outside. Luminosity variations are mostly zero except at the edge of the aperture where $\nabla I = -I_0 \vec{n} \delta_C$, where δ_C is a Dirac function of unity at the aperture edge and zero elsewhere and $\vec{n}$ a unit vector perpendicular to the aperture edge and pointing outwards. By applying these conditions to Equation (4.58), it yields:

$$\frac{\partial I}{\partial z} = I_0 \frac{\partial W}{\partial n} \delta_C - I_0 P \nabla^2 W \tag{4.59}$$

where $P(x, y)$ is a special function with unity inside the aperture and zero outside. The expression $\partial W/\partial n = \vec{n} \cdot \nabla W$ is the radial wavefront slope at the pupil edge. For geometrical optics approximation, the luminosity distributions on these two planes are:

$$I_1 = I_0 - \frac{\partial I}{\partial z} \Delta z, \quad I_2 = I_0 + \frac{\partial I}{\partial z} \Delta z \tag{4.60}$$

The normalized difference between illuminations of these two planes is:

$$S = \frac{I_1 - I_2}{I_1 + I_2} = \frac{f(f-l)}{l} \left(\frac{\partial W}{\partial n} \delta_C - P \nabla^2 W \right) \tag{4.61}$$

If R is the mirror radius, the condition for the above equation is:

$$\frac{\lambda f(f-l)}{l} \ll r_0^2 \ll R^2$$

where r_0 is the correlation scale of the intensity fluctuations. The above equation shows that the normalized illumination difference has two parts: one is proportional to the wavefront radial tilt at the edge and the other is proportional to the curvature or the Laplacian of the wavefront. These two parts do not

overlap. From information of both terms, the wavefront can be reconstructed by solving the equation with Neumann boundary conditions. Since the wavefront curvature is derived directly, therefore, the wavefront sensor based on this theory is called a curvature sensor.

Advantages for a curvature measurement are: (a) the curvature is a scalar with only one measurement being necessary at each position instead of two for the slope measurement; (b) the spectrum of the atmospheric distorted wavefront decreases on an order of $k^{-11/3}$, where $k = 2\pi/\lambda$ is the wave number. The slope measurements with a power spectrum of $k^{-5/3}$ have a relatively large correlation length over the pupil, while the spectrum of the wavefront curvature decreases only on an order of $k^{1/3}$(Roddier, 1988). The curvature is smoother. The correlation between curvature fluctuations is small. (c) in the temporal behavior, the slope measurement presents a large correlation time. The curvature measurement has a much smaller correlation time.

There are several ways to get these intra- and extra-focus images. A beam-splitting prism and relay optics provide both images on one detector. A vibrating membrane mirror in front of the focus can have two images alternatively. This type of sensor has high sampling frequency up to several KHz. For wide field work, offset CCDs can be arranged before and after the focus, the illumination can be calculated through the analysis of many star images.

4.1.5.2 Single-Image Wavefront and Curvature Sensor

The dual-image method can be replaced by a simple single-image one. The principle remains unchanged. When light passes a point $\mathbf{r}$ on the entrance pupil and arrives at a new off focus position with a distance l to the focus, the light intensity variation relative to an average intensity at the point has an exact same expression as Equation 4.61 (Hickson and Burley, 1994). In this equation, both the intensity and wavefront are functions of entrance pupil coordinates.

The intensity fluctuation is mainly due to the wavefront curvature. However, scintillation may also have an effect. Scintillation is caused by the spatial variations in the amplitude of the incident radiation introduced by the turbulent air layer distance change. In the two-image curvature sensor, the intensities are measured and combined at two locations with same distance before and after the focus, so that the scintillation effect is cancelled, while the curvature effects are enhanced as they are equal in magnitude but with opposite signs.

The scintillation effect increases with the turbulent layer distance. These distances are of the order of several kilometers, a value much larger than the distance $f - l$. Therefore, the scintillation effect is practically constant for images both in the pupil and near the focus. Based on this assumption, the curvature information of the wavefront can be subtracted from a single out-off-focus image. The technique is to achieve the variation or fluctuation of the local intensity $\Delta I(r)$ through subtracting the local intensity $I(r)$ a constant average signal I_0, which is obtained through integrating the image over a period of time.

In practice, the average intensity is stored in computer memory. Variations in the wavefront tilt, represented by the Dirac function and the defocus would appear as difference signals at the edge of the pupil.

Compared with two image sensors, the single image curvature sensor is simpler and has improved signal-to-noise ratio since the former ones require separate detectors and beam splitting devices. In the new method, the exact distance between the image and focal plane is usually not critical.

The outputs of a curvature sensor are curvature and edge wavefront tilt. In most cases, wavefront phase is not sought when the deformable mirrors are not, but tip-tilt or membrane mirrors are used in the system. However, the curvature sensors can also be worked as wavefront sensors as the wavefront phase information can be extracted from a map of its gradient (Laplacian) plus its edge radial tilt as a Neumann-type boundary condition.

This direct wavefront measurement is different from another technique, named as phase retrieval. Phase retrieval method works in the diffraction regime and requires taking monochromatic image of point source either in focus or near focus. Like any interferometer technique, phase retrieval is sensitive to vibrations and to turbulence, limiting its application in ground-based telescopes at long wavelengths. The technique of extracting wavefront phase from curvature information works with wideband relatively long exposures in visible wave with ground-based telescopes (Roddier and Roddier, 1993). It works in geometrical optics regime like the Hartmann test method. Best results can be obtained with large amount of defocus, well outside the so-called caustic zone, which is volume near the focus when aberrations are presented in the image. The observation of such off-focus images is sensitive to the mirror alignment or mirror figure errors. In the past, it was referred as the eye-piece test or the inside-and-outside test.

Reconstructing the wavefront from defocused images can be done simply solving the differential equation by an integration-like algorithm or Fast Fourier Transform algorithm (Roddier, 1991). The basic idea behind the FFT algorithm is that the Laplacian operator $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is equivalent to a multiplication by $u^2 + v^2$ in the Fourier space, where u and v are variables in Fourier space. Therefore, one can divide the Fourier transform of the Laplacian, namely $FT[\nabla^2 W]$ by $u^2 + v^2$. After that, the wavefront function is an inverse Fourier transform of this function. It is $FT^{-1}\{FT[\nabla^2 W]/(u^2 + v^2)\}$. However, this method is for a wavefront function without boundary. Therefore, the wavefront Laplacian has to multiply an aperture function (a pupil transmission function). In Fourier plane, this is equivalent to a convolution with Fourier transform of the aperture function. So, iteration is necessary to derive the boundary conditions outside the image domain.

A new improved method is through iteratively compensating the effect of the estimated aberrations on the defocused images exactly as in an active optics control loop. Residual aberrations are again estimated and compensated until the noise level is reached.

An aberration of wavefront will produce a deviation of the optical ray based on its slop direction. This produces a displacement of the ray on the image plane:

$$\begin{cases} x' = x + C\partial W(x,y)/\partial x \\ y' = y + C\partial W(x,y)/\partial y \end{cases}$$

If $I(x,y)$ is the intensity at point N and $I'(x,y)$ is the intensity at another related point N'. The flux conservation requires that:

$$I(x,y)d^2N = I'(x',y')d^2N' = I'(x',y')Jd^2N$$
$$J = \begin{vmatrix} \partial x'/\partial x & \partial x'/\partial y \\ \partial y'/\partial x & \partial y'/\partial y \end{vmatrix} = \begin{vmatrix} 1 + C\partial^2 W/\partial x^2 & C\partial^2 W/\partial xy \\ C\partial^2 W/\partial xy & 1 + C\partial^2 W/\partial^2 y \end{vmatrix} \tag{4.62}$$

Then, the image compensation requires changing the intensity $I'(x,y)$ into $I(x,y)$. As the wavefront is expressed in terms of Zernike polynomials, the Jacobians J for first 15 or so terms are computed for the image intensity improvement. The reconstruction of wavefront is obtained through addition of the compensated Zernike terms to the residuals.

In this way, the sensor is not only a curvature, but also a wavefront sensor. In radio telescopes, a similar technique, out-off-focus holography, is used for antenna surface error measurement (Nikolic et al., 2008). The antenna surface error is the half phase error of the wavefront (Section 8.4.1).

4.1.5.3 Tip-Tilt and Curvature Compensation Devices

From curvature information, it is possible to derive the slope distribution and the wavefront tilt terms. If only the tilt and focusing errors are corrected, the devices are simple and a small steerable mirror is used. This small mirror is called a tip-tilt mirror, a mirror with two axial tilting. Small size secondary mirrors can also be treated as tip-tilt mirrors. If the wavefront curvature term is also compensated, a small bimorph or membrane mirror is necessary.

A bimorph mirror (Figure 4.22) consists of two piezoelectric ceramic wafers which are bonded together and oppositely polarized, parallel to the axes. An array of electrodes is deposited on the inner side of the wafers. Both front and back surfaces are grounded. By applying voltages to the electrodes, one layer shrinks

Fig. 4.22. A bimorph mirror structure (Roddier and Roddier, 1999).

while the other expands, resulting in an induced curvature. The deformation of the mirror can be expressed by a Poisson equation:

$$\frac{\partial^2 z'}{\partial t^2} = A\nabla^4 z' + B\nabla^2 V \tag{4.63}$$

where A and B are constants, z' the surface deformation, and $V(x, y, t)$ the applied voltage. In the curvature compensation, signals from the curvature sensor can be amplified and directly applied to the bimorphs for the required compensation.

A membrane mirror (Figure 4.23) consists of a stretched reflective membrane inside a partial vacuum chamber. The mirror is deformed by means of electrostatic forces between electrodes on the mirror back and the chamber frame. For a linear response, a fixed bias voltage is added to the signal voltages. The thin air left inside the chamber can damp excessive vibration of the membrane and improve its response. This mirror has similar characteristics to a bimorph. Its deformation equation under electrostatic force is:

$$\frac{\partial^2 z'}{\partial t^2} = A\nabla^2 z' + BP \tag{4.64}$$

where A and B are constants, z' the surface deformation, $P(x, y, t)$ the applied electrostatic voltage. This equation has the same form as Equation (4.61). The compensation is straight forward.

4.1.6 Atmospheric Disturbance and Adaptive Optics Compensation

With atmospheric disturbance, the telescope system becomes seeing limited. The atmosphere optical transfer function is a spatial coherence function of the

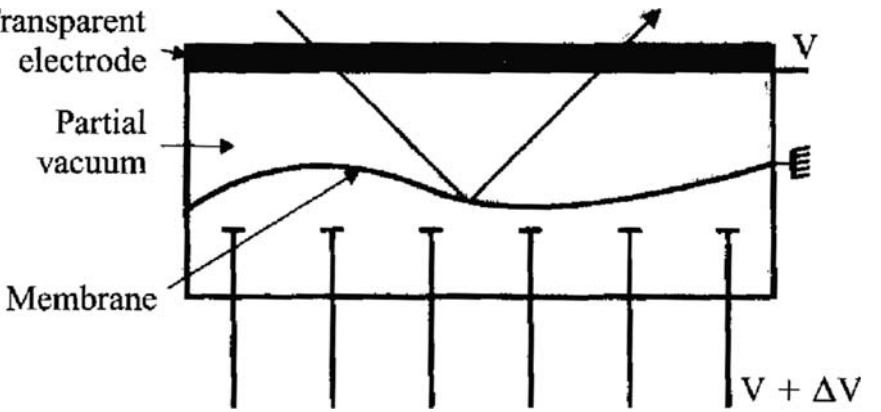


Fig. 4.23. A membrane mirror structure (Roddier and Roddier, 1999).

optical field, which is the covariance of complex light wave functions (Max, 2003):

$$A(v) = C_\Psi(v) = <\Psi(x)\Psi^*(x+r)> \tag{4.65}$$

where $\Psi(x) = \exp[i\phi(x)]$ is the light wave function with the phase $\phi(x) = kz - \omega t$. The bracket in the formula means taking an average value. The covariance is also called an auto-correlation function. In this formula, the spatial frequency has been replaced by a length of separation r, $v = 2\pi/r$. Sometimes, the spatial coherence function can be expressed by using the phase structure function $D_\phi(r)$:

$$\begin{aligned} C_\Psi(v) &= <\exp i[\phi(x) - \phi(x+r)]> \\ &= \exp[-<|\phi(x) - \phi(x+r)|^2>/2] = \exp[-D_\phi(r)/2] \end{aligned} \tag{4.66}$$

A structure function is a property of a stably varying nonstationary random variable, such as the atmosphere temperature or refractive index distribution. It represents the periodicity of the variable's change. It equals a mean or an average of the squared variable difference:

$$D_f(r) = <(f(x) - f(x+r))^2> \tag{4.67}$$

The definition of a covariance, or correlation, function is:

$$B_f(r) = <f(x+r)f(r)> = \int_{-\infty}^{\infty} dx f(x+r)f(x) \tag{4.68}$$

The structure function is related to the covariance function as:

$$\begin{aligned} D_f(r) &= <(f(x+r) - f(x))^2> \\ &= 2<(f(x))^2> - 2<f(x+r)f(x)> \\ &= 2[B_f(0) - B_f(r)] \end{aligned} \tag{4.69}$$

The atmosphere as a turbulent flow, where the local spatial scale l and the velocity V are associated, has its energy per unit mass $\propto V^2/2$. The energy dissipation rate per unit mass per unit time will be proportional to $\propto \varepsilon = V^2/\tau = V^2/(l/V) = V^3/l$, where τ is a time scale. If this expression is rearranged, one gets $V \propto (\varepsilon \cdot l)^{1/3}$. This means that the velocity associated with eddies of a particular size is proportional to the cubic root of length. This is Kolmogorov's scaling law which is true for an inertial range between an inner scale of 1–0.1 mm

and an outer scale of 10–10,000 m. Beyond the range, the assumed relationship between the scale and velocity will not hold. The energy (power) density is, therefore, proportional to $V^2 \propto \varepsilon^{2/3} l^{2/3}$. The power spectrum density of the velocity fluctuation over a small spatial frequency interval will be:

$$S(v)dv \propto V^2 \propto \varepsilon^{2/3} l^{2/3} \tag{4.70}$$

Since $v = 2\pi/l$, in two dimensions the power spectrum of the velocity fluctuation in the atmosphere is:

$$S(v) \propto v^{-5/3} \tag{4.71}$$

This dependence of $S(v)$ on v gives the Kolmogorov –5/3 power law. The relationship between 2D and 3D power spectra is:

$$2\pi v^2 S_{3d}(v) = S(v) \tag{4.72}$$

So the 3D power spectrum is:

$$S_{3d}(v) \propto v^{-11/3} \tag{4.73}$$

The velocity fluctuation power spectrum is in Fourier space. In physical space, the structure function is more important. If a function has its power spectrum of $S(v)$, then the structure function can be expressed as:

$$D_f(r) = 2 \int S(v)(1 - e^{ivr}) d^3 v \tag{4.74}$$

Since the integral of a power law remains a power law, the power law indices can be derived. The calculation is difficult. Using the 2D power spectrum, the velocity structure function of the atmosphere is:

$$D_V(r) = <[V(x+r) - V(x)]^2> = C_V^2 |r|^{2/3} \tag{4.75}$$

The temperature fluctuations are passively derived from the velocity field and it influences the refractive index changes. Therefore, the temperature and index structure functions are:

$$\begin{aligned} D_T(r) &= <[T(x+r) - T(x)]^2> = C_T^2 |r|^{2/3} \\ D_N(r) &= <[N(x+r) - N(x)]^2> = C_N^2 |r|^{2/3} = C_N^2 (l^2 + z^2)^{1/3} \end{aligned} \tag{4.76}$$

Since the refractive index of air is:

$$n - 1 = \frac{77.6 \times 10^{-6}}{T} (1 + 7.52 \times 10^{-3} \lambda^{-2}) \left(P + 4810 \frac{e}{T} \right) \tag{4.77}$$

where P is the pressure in mbar, T the temperature in K, and e the water vapor pressure also in mbar. In general, the coefficient is:

$$C_N^2 = \left(77.6 \times 10^{-6} \frac{P}{T^2}\right) C_T^2 \text{ at } \lambda = 0.5\ \mu\text{m} \tag{4.78}$$

During night time, this is $C_N^2 \approx 10^{-13} - 10^{-15}\ \text{m}^{-2/3}$. The phase is calculated using the refractive index multiplied by the optical path dz:

$$\phi(x) = k \int_h^{h+\delta h} dz \times n(x, z) \tag{4.79}$$

From this expression, the phase covariance function is:

$$\begin{aligned} B_\phi(r) &= k^2 \int_h^{h+\delta h} dz' \int_h^{h+\delta h} dz'' < n(x, z') n(x + r, z'') > \\ &= k^2 \int_h^{h+\delta h} dz' \int_{h-z'}^{h+\delta h-z'} dz B_N(r, z) \approx k^2 \delta h \int_{-\infty}^{\infty} dz B_N(r, z) \end{aligned} \tag{4.80}$$

where $B_N(r) = B_N(r, z)$ is the refractive index covariance function, $k = 2\pi/\lambda$ the wave number, and h the height of the atmospheric layer. From these formulas, the relationship between the phase structure function and the refractive index structure function is:

$$D_\phi(r) = k^2 \delta \text{h} \int_{-\infty}^{\infty} dz [D_N(r, z) - D_N(0, z)] \tag{4.81}$$

where $D_N(r)$ is the refractive index structure function. Therefore, the phase structure function can be derived from the refractive index structure function:

$$\begin{aligned} D_\phi(r) &= k^2 \delta h C_N^2 \int_{-\infty}^{\infty} dz [(r^2 + z^2)^{1/3} - z^{2/3}] \\ &= 2.914 \left(\frac{2\pi}{\lambda}\right)^2 r^{5/3} \int_0^L C_N^2(h) dh = A^* r^{5/3} \end{aligned} \tag{4.82}$$

where L is the height from the earth's surface. Introducing the Fried parameter as $r_0 = (6.88/A^*)^{3/5}$ where A^* is defined in the above equation, the phase structure function expressed in scalar distance is (Fried, 1965):

$$D_\phi(r) = 6.88 (r/r_0)^{5/3} \propto r^{5/3} \tag{4.83}$$

The Fried parameter is also named as the atmospheric correlation length. It is the distance over which the optical phase distortion has a mean square value of 1 rad^2. The seeing, or the FWHM of the point spread function, is $0.98\lambda/r_0$. From this correlation length, we can derive the atmospheric correlation time using average wind velocity:

$$\begin{aligned} &\tau_0 \approx 0.3(r_0/\bar{V}); \\ &\bar{V} = \left[\int dh \cdot C_N^2(h) |V(h)|^{5/3} / \int dh \cdot C_N^2(h) \right]^{3/5} \end{aligned} \tag{4.84}$$

The reciprocal of the coherence time is called the Greenwood frequency, which is the frequency required to compensate the atmospheric turbulence. At the Hawaii Haleakala site, the Greenwood frequency is about 20 Hz (Tyson, 2000). From the Fried's parameter, the atmospheric isoplanatic angle is:

$$\begin{aligned} &\theta_0 \approx 0.314 \cos\xi \left(\frac{r_0}{\bar{h}}\right); \\ &\bar{h} = \left(\int dz \cdot z^{5/3} C_N^2(z) / \int dz C_N^2(z) \right)^{3/5} \end{aligned} \tag{4.85}$$

where ξ is the zenith angle (Hardy, 1998). In the same reference, this zenith angle dependent term also appears in the equation of the phase structure function. The isoplanatic angle θ_0 corresponds to a Strehl ratio reduction of 0.38. This is a very important parameter in adaptive optics. If the adaptive correction has a time delay τ, then the residual wavefront variance is $\sigma^2 = 28.4(\tau/\tau_0)^{5/3}$. If the guide star used for the adaptive correction is at an angle θ away from the source, then the residual wavefront variance will be $\sigma^2 = (\theta/\theta_0)^{5/3}$.

Traditionally, the wavefront phase error is expressed as a function of the pupil coordinates, known as zonal expression. The error can also be expressed as Zernike polynomials, known as modal expression. Each term in the Zernike expression is called a mode of wavefront phase error. In modal expression, the wavefront phase error is:

$$\phi(r,\theta) = \sum_j a_j Z_j(r,\theta) \tag{4.86}$$

For wavefront phase error caused only by atmospheric disturbance and expressed in a modal form, when the first J modes are corrected, the corrected part of the phase error can be written as:

$$\phi_c = \sum_{j=1}^{J} a_j Z_j \tag{4.87}$$

The remaining mean square residual error is:

$$\Delta_J = \int dr W(r) < [\phi(r) - \phi_c(r)]^2 > = < \phi^2 > - \sum_{j=1}^{J} < |a_j|^2 > \qquad (4.88)$$

By applying the atmosphere phase structure function, the above expression can be solved for the remaining mean square residual errors when the first J modes of atmosphere turbulence are corrected. The residual phase errors for the first J modes corrected are listed in Table 4.2 (Section 4.1.5). From this table, one finds that the contributions from the first few terms, i.e., the tip-tilt and curvature terms, are significant for atmospheric disturbances. This is why a tip-tilt correction is very important in telescope imaging.

The optical transfer functions of the atmospheric disturbance, the telescope, and the diffraction limited image are shown in Figure 4.24. These are seeing limit, aberration limit, and diffraction limit. For a telescope with a 0.5 m

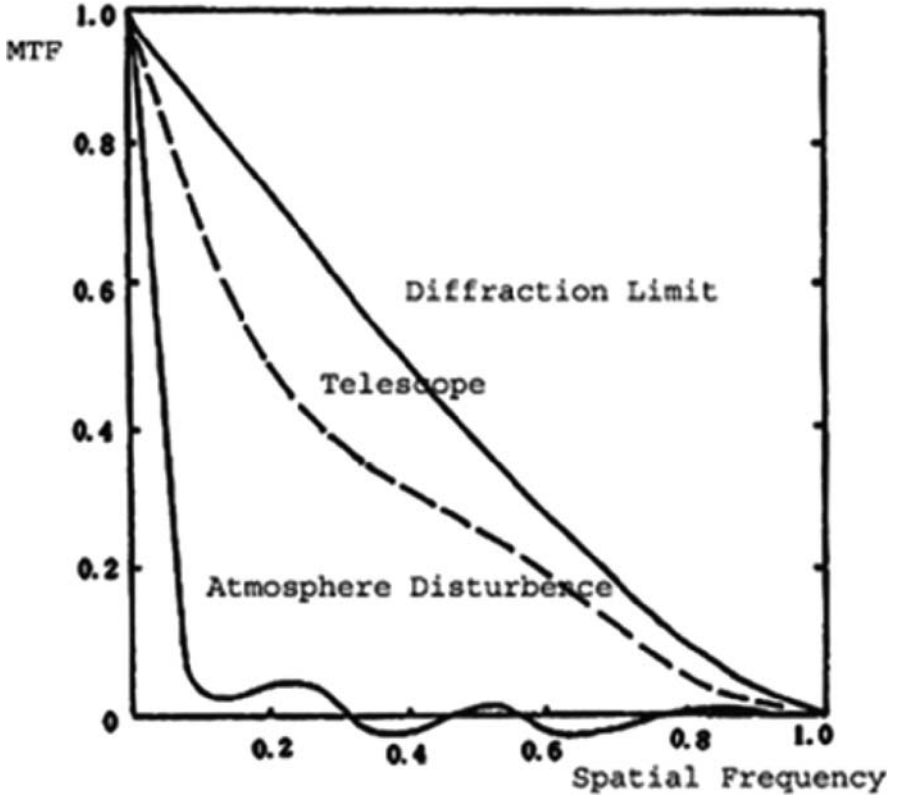


Fig. 4.24. Optical transfer function for atmosphere, telescope, and the diffraction limited aperture.

aperture size, the diffraction limit is about 24 cycle/arcsec and the seeing limit about 1 cycle/arcsec (Section 4.1.5). The aberration effect is in between. To improve the seeing and aberration effects, the wavefront phase error caused by the atmosphere and the aberrations should be compensated.

Different from active optics, the correction of seeing in adaptive optics has to be performed at a higher frequency, at about 100–200 Hz or higher. The deformable mirror used as actuators should have changeable sub-areas determined by Fried parameter r_0. For a good observatory site, r_0 in the optical regime usually equals 10 cm. Therefore, a telescope has a primary mirror of 2.5 m, about 500 piezoelectric ceramic actuators are required for its deformable mirror. Adaptive optics also requires a much higher temporal sampling frequency; so, a guide star of magnitude about 10, depending on wavefront sensor used, is needed. For active optics, the magnitude can be as low as magnitude 17 (0.01 Hz) to 20 (0.1 Hz).

Figure 4.25 shows a typical adaptive optics system. The star light (1) passes through the atmosphere (2) and is collected by the telescope (3). It arrives at a deformable mirror (4) at the telescope pupil. After reflection from the deformable mirror, most of the light reaches the focus, and some of the light reaches a wavefront sensor (6). Measured wavefront signals are processed and amplified by the computer (8) to control the deformable mirror.

4.1.7 Artificial Laser Guide Star and Adaptive Optics

One limitation in adaptive optics is the small size of the isoplanatic angle in the optical region. Within this small angle (<5 arcsec in the optical regime for best seeing), finding a natural guide star (NGS) bright enough (m<10) near the target is very difficult. The average number of stars brighter than m magnitude in the optical regime is only about $3\exp(0.9m)$ stars/rad^2. At infrared

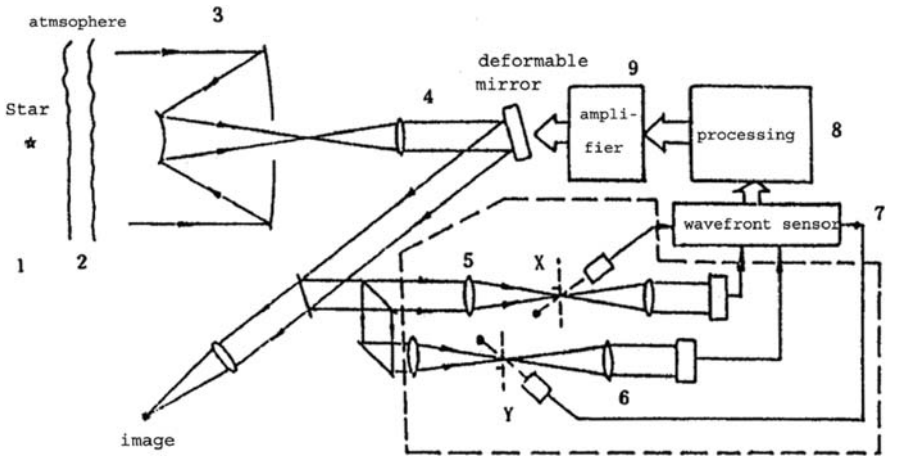


Fig. 4.25. A typical adaptive optics system (Hardy et al., 1977).

wavelength (λ=1 μm), the isoplanatic patch is slightly larger (∼10 arcsec), however, the probability in finding a bright NGS within the patch is still low. This limits the sky coverage. To fill the gap between the available natural guide stars, artificial Laser Guide Stars (or laser beacons) are needed.

Two types of laser guide stars (LGSs) exist, the Rayleigh and sodium ones.

4.1.7.1 Sodium Beacon and LGS Cone Effect

Sodium beacon guide stars are created by using a laser specially tuned to 589.2 nm to energize a layer of sodium atoms which is naturally present in the mesosphere at an altitude of around 90 km. The sodium atoms then re-emit the laser light, producing a glowing artificial star. The sodium laser guide stars are less bright, but at a much higher altitude.

The biggest problem of using a laser guide star is the cone effect. The height of a LGS is limited, so that the atmospheric turbulence detected is within a cone volume from the artificial star. Any star observed around a LGS will pass through the atmosphere outside the cone volume. Those outside turbulences are not detected and are not compensated. The cone effect is also called focal anisoplanatism (Tallon and Foy, 1990). The wavefront variance due to focal anisoplanatism can be expressed as (Tyson, 1997):

$$\sigma_{cone}^2 = (D/d_0)^{5/3} \tag{4.89}$$

where D is the telescope aperture and d_0 the focal anisoplanatism parameter (unit is meter). d_0 is defined as:

$$d_0 = \lambda^{6/5} \cos^{3/5\beta} \left[19.77 \int \left(\frac{z}{z_{LGS}} \right)^{5/3} C_n^2(z) dz \right]^{-3/5} \tag{4.90}$$

where z_{LGS} is the altitude of the artificial guide star (in km), β the zenith angle, and C_n the atmospheric refractive index structural constant. This formula is difficult to use; however, the focal anisoplanatism parameter can be expressed in a much simpler way as a linear function of the altitude of the LGS. In different models, d_0 (in meter) has a different expression as follows (Tyson, 1997):

$$\begin{aligned} d_0[HV5/7\} &= 0.018 z_{LGS} + 0.39 \\ d_0[SLC - Day] &= 0.041 z_{LGS} + 0.299 \\ d_0[SLC - Night] &= 0.046 z_{LGS} + 0.42 \end{aligned} \tag{4.91}$$

where HV means Hufnagel–Valley turbulence model and SLCs are other turbulence models, one of the SLCs is for daytime and the other for night time.

If the wavefront error is one tenth of the wavelength, then d_0 for a 4 m aperture telescope will be 7 m from the SLC-Night model of Equation (4.91). It follows that the laser guide star should be located at an altitude of about 143 km.

This is not realistic. In this case, multiple laser guide stars may be used for overcoming the cone effect from a single LGS. For multiple laser guide stars, d_0 will be calculated from the following formula:

$$d_0[multiple - star] = 0.23N_{LGS} + 0.95 \tag{4.92}$$

where N_{LGS} is the number of artificial laser stars.

4.1.7.2 Rayleigh Beacon

The Rayleigh backscattering guide stars rely on the λ^{-4} Rayleigh scattering of aerosols and dust in atmosphere at an altitude of 10–20 km. They can be made very bright ($m < 5$) with a copper vapor laser without fine tuning of wavelength. Astronomers also plan to use ultraviolet lasers as guide stars. However, the height of a Rayleigh guide star of this type limits its usage in astronomy.

When a pulsed laser beam is focused at altitudes between 10 and 20 km above ground, Rayleigh back-scattering will take place due to the air density fluctuations. According to LIDAR theory (LIDAR means LIght Detection And Ranging, which is similar to a radar), the brightness of this kind of laser guide star is directly proportional to the atmospheric density where the laser back-scattering takes place. The number of photons received from this type of LGS on the ground is given by:

$$F_{Rayleigh} = \eta T_A^2 \frac{\sigma_R n_R}{4\pi z_0^2} \frac{\Delta z \lambda_{LGS} E}{hc} \tag{4.93}$$

where η is the efficiency of the telescope and receiver, T_A the atmospheric transmission between the telescope and laser guide star, σ_R the area of Rayleigh dispersion section, n_R the atmospheric density, Δz the altitude range where Rayleigh scattering has taken place, λ_{LGS} the wavelength of the laser guide star, z_0 the distance between the telescope and the laser guide star, E the laser energy, h the Plank constant, and c the speed of light. For Rayleigh scattering, the product of the laser beacon area and atmospheric density equals approximately:

$$\sigma_R n_R \approx 2.0 \times 10^{-4} \exp[-(z_0 + z_t)/6 \text{ km}] \tag{4.94}$$

where z_t is the altitude where the laser beam sends out. Suppose that the wavelength of the laser guide star is 351 nm, for a laser beam generator with an aperture of D_{proj}, the altitude range where the Rayleigh backscatter takes place in the atmosphere is:

$$\Delta z = 4.88\lambda z_0^2/(D_{proj} r_0) \tag{4.95}$$

If a laser beam generator with a 1 m aperture size and the altitude of the observatory site is 3 km above sea level, the altitude of the atmospheric layer, in which the Rayleigh laser guide star is formed is about 20 km. The total height of

the backscatter layer is about 33 m. With all these assumptions, η equals 0.075 and T_A equals 0.85, the photon energy collected by the telescope is $6.2 \times 10^5 E$ [J/pulse]. For a sub-aperture of a diameter of the Fried parameter r_0, the number of photons collected is $N_s = (\pi/4) \cdot r_0^2 F = 1.1 \times 10^{+4} E$. If N_S equals 150 photons, the energy of the laser generator needed is 14 mJ/pulse.

4.1.7.3 Other Limitations

Because Rayleigh scattering can take place at any altitude in the atmosphere, it is very important to get rid of those unwanted back-scattering photons through a range gating technique. The technique is also important for sodium LGS. Since the photons scattered from the bottom of the atmosphere reach the telescope at a shorter time than the photons back-scattered from a laser guide star, a switch can be used to turn off the receiver before the photons were scattered from the LGS. For a Rayleigh star at an altitude of 20 km, the total time for photons arriving at the telescope is 132 μs. The receiver will not open if the time interval is less than 132 μs. Such range gating ensures that the wavefront sensor is opened when the laser beam scattered from the required height.

The resonance effect of sodium atoms takes place at a wavelength of 589.1583 nm at a much higher altitude. The number of photons from a sodium laser guide star collected by the telescope can be expressed as:

$$F_{Sodium} = \eta T_A^2 \frac{\sigma_{Na} \rho_{col}}{4\pi z_0^2} \frac{\Delta z \lambda_{LGS} E}{hc} \tag{4.96}$$

where σ_{Na} is area of resonance radiation, ρ_{col} the abundance of sodium atoms inside a unit cylinder of the atmosphere (between 3×10^9 atom/cm^2 and 1×10^{10} atom/cm^2), and z_0 the altitude of the laser guide star which is about 92 km. The product of the radiation area of this kind of laser guide star and the abundance in the unit cylinder equals approximately 0.02. Because the number of sodium atoms at such an altitude is limited, the intensity of the sodium laser star can only reach 1.9×10^8 photons. For a sub-aperture of a diameter of Fried parameter r_0, the number of photons collected is $N_s = (\pi/4) r_0^2 F = 550E$. If the N_S required equals 150 photons, the required laser energy should be 272 mJ/pulse.

Another problem with a laser guide star is the pointing error. If a wavefront slope occurs in the atmospheric layer, the actual position of the laser guide star cannot be determined (Figure 4.26) as the paths of the light rays are the same on the way up as on the way down. In this situation, multiple artificial laser stars, multiple waveband observations, or a faint closeby natural guide star can be used for the pointing (tip-tilt) control. The multicolor laser probe is only applicable to sodium resonant scattering at 90 km, it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength.

Multi-laser guide stars are required inside the field of view when atmosphere tomography and multi-conjugate adaptive optics are performed. If these

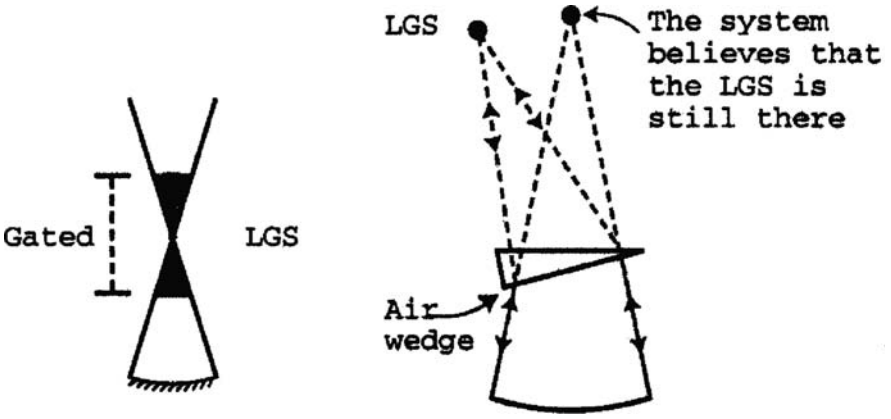


Fig. 4.26. Range gating and pointing error from the laser guide star.

multi-laser guide stars are launched from the same alt-azimuth telescope doing the observation, a field de-rotation device may be necessary to fix the star positions in the sky. A reflecting field de-rotation device, named a K-mirror, can be used. Figure 4.27 shows a K-mirror field de-rotation device used for the proposed Thirty Meter Telescope multi-LGS launching system.

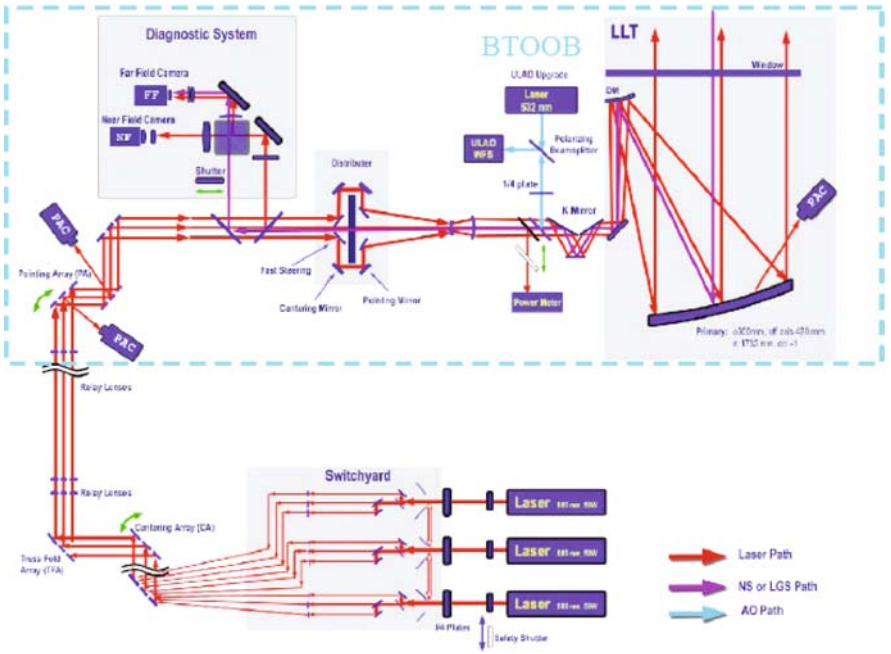


Fig. 4.27. A design of a multi-laser guide star projection system for the Thirty Meter Telescope (Liang, 2004).

4.1.8 Atmosphere Tomography and Multi-Conjugate Adaptive Optics

4.1.8.1 Atmosphere Tomography

By using a wavefront sensor, the atmospheric turbulence sensed is within a cylinder or a cone volume above the telescope aperture. Stars from other directions will pass through different atmospheric volumes. The resulting anisoplanatism poses a problem not only for laser guide stars, but also for natural guide stars. To overcome this anisoplanatism problem, it is necessary to obtain 3D details of the wavefront deformation from the atmospheric turbulence. Generally, using more guide stars within the field of view, one can restore more 3D details of the atmospheric turbulence. The wavefront error gained by a wavefront sensor is the sum or the projection of the atmospheric turbulence in the guide star direction. The problem to find 3D details from more than one projection of an image is a reverse problem, named tomography. Atmospheric tomography, or turbulence tomography, is a special term for the technique to determine 3D details of atmosphere turbulence through a number guide stars sensing.

To retrieve 3D details from the projections of a function, different methods can be used. Among those, the Radon transformation (Shepp, 1982; Liang et al., 2000), which is a special Fourier transformation, is the most successful one. However, with this approach, projections of about 180 degrees range are required to reconstruct a complete 3D image. This is difficult for the atmosphere tomography, where the wavefront information or the projection gained is from a very limited angular range.

Early atmosphere tomography (turbulence tomography or layer oriented tomography) was based on a traditional ray tracing technique. In this approach, a wavefront zonal approach is used and the atmosphere volume under the guide stars is divided into several layers. At any point on each layer, there are two sets of numbers, representing the slope changes of wavefront in two perpendicular directions. These slope numbers are unknowns. Then rays are traced from each guide star to the telescope aperture plane. The sum of wavefront changes at each ray passing zonal area of each atmosphere layer provides the wavefront slopes in the telescope pupil plane. The sum of errors is detected through the wavefront sensor. If the number of traced rays from each guide star is N, then one guide star provide $2N$ equations. These N rays intersect one atmospheric layer at N points. The unknown slope number required on one layer is $2N$. If the number of the atmospheric turbulence layers is more than one, then more than one guide star is required to solve this inverse problem. In a zonal approach, the ray intersecting points of each layer are usually different from the given grid points in each atmospheric layer, producing errors in calculation. For improving this situation, Ragazzoni et al. (1999, 2000) proposed an alternative modal approach based on Zernike polynomial expression of wavefront error.

A rectangular coordinate is used in zonal expression of wavefront error. Polar coordinate is used in the modal expression. In this modal expression,

a sum of Zernike polynomials up to a given order Q is used for wavefront error:

$$W(\rho,\theta)=\sum_{n,m=0}^{Q}\rho^{n}[A_{nm}\cos(m\theta)+B_{nm}\sin(m\theta)] \tag{4.97}$$

where $n \geq m$ and $n-m$ is an even number (when $m=0$, B_{n0} has no meaning). The total number of independent coefficients in this expression is:

$$\frac{(Q+1)^2+(Q+1)}{2}=\frac{Q^2+3Q+2}{2} \tag{4.98}$$

In order to transfer the polar expression into a zonal expression, it is necessary to use special formulas of cosine and sine:

$$\begin{gathered}\cos(m\theta)=\cos^{m}\theta-\frac{m(m-1)}{1\cdot 2}\cos^{m-2}\theta\sin^{2}\theta+\\ \frac{m(m-1)(m-2)(m-3)}{1\cdot 2\cdot 3\cdot 4}\cos^{m-4}\theta\sin^{4}\theta-\ldots\\ \sin(m\theta)=m\cos^{m-1}\theta\sin\theta-\\ \frac{m(m-1)(m-2)}{1\cdot 2\cdot 3}\cos^{m-3}\theta\sin^{3}\theta+\ldots\end{gathered} \tag{4.99}$$

In the right-hand sides of these formulas only power terms of the cosine and the sine of a base angle are used. These cosine and sine terms can be replaced by corresponding x and y values of a rectangular coordinate system. Therefore, the cosine and sine of any multiples of angle can be expressed as rectangular coordinates x and y as:

$$\begin{aligned}\cos(m\theta)=&(x^2+y^2)^{-m/2}[a_{0m}y^{m}-a_{1m}y^{m-2}x^2+a_{2m}y^{m-4}x^4+\ldots]\\ \sin(m\theta)=&(x^2+y^2)^{-m/2}[b_{0m}y^{m-1}x-b_{1m}y^{m-3}x^3+b_{2m}y^{m-5}x^5-\ldots]\end{aligned} \tag{4.100}$$

The Zernike polynomials of wavefront error, therefore, can be written in rectangular coordinates by replacing the cosines and sines:

$$\begin{aligned}W(x,y)=\sum_{n,m=0}^{Q}\Big\{&(x^2+y^2)^{(n-m)/2}\Big[A_{nm}(a_{0m}y^{m}-a_{1m}y^{m-2}x^2+\ldots)\\ &+B_{nm}(b_{0m}y^{m-1}x-b_{1m}y^{m-3}x^3+\ldots)\Big]\Big\}\end{aligned} \tag{4.101}$$

This expression still represents a circular aperture. The coefficients used in this expression remain the same as in the Zernike polynomials. With coordinate transformation, this formula represents all circles of different size and different

location. If the aperture radius is smaller than unity, a constant k can be used. If the aperture is offset, a shift of the origin by an amount of $(\Delta x, \Delta y)$ can be used. The derived expression for any circular aperture will be:

$$W'(x', y') = W(\Delta x + kx', \Delta y + ky') \tag{4.102}$$

With the above transformations, it is now possible to solve the tomography inverse problem. In this process, the atmosphere is divided into M layers and N guide stars are needed for wavefront sensing (Figure 4.28). The atmospheric layers start from the aperture plane upwards. On each layer,

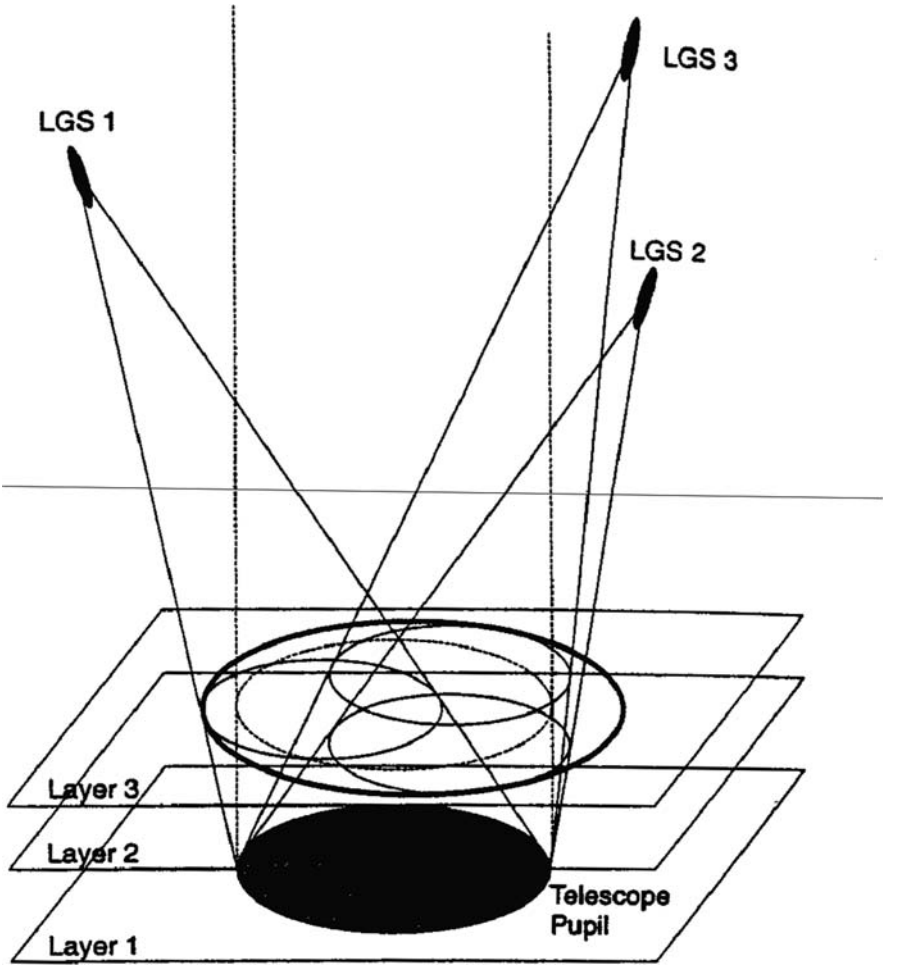


Fig. 4.28. Layered atmosphere turbulence and *sub-circles* of each laser guide star. The large meta-pupil includes all the sub-circles of the same layer (Ragazzoni et al., 1999).

there is a small intersecting circle from the i-th guide star and also a large circle which includes all the smaller circles. This large circle is called the meta-pupil of this layer.

For the i-th guide star, the wavefront phase expression at the aperture plane has P Zernike coefficients:

$$L_i = \left[a_4, a_5, \cdots a_{p+3} \right]^T \tag{4.103}$$

In the Zernike wavefront expression, the first three coefficients are respectively piston and tilt. These coefficients can be neglected when an appropriate reference surface is used. This expression of the wavefront phase is a summation of corresponding coefficients of each atmospheric layer:

$$L_i = \sum_{j=1}^{M} L_{ij} \tag{4.104}$$

where j is the number of atmospheric layer. In a similar way, we can define the Zernike wavefront expansion on the meta-pupil on each layer, $W_j, j = 1.2 \cdots M$. The contribution of i-th star to the j-th layer is:

$$L_{ij} = A_{ij} W_j \tag{4.105}$$

where the matrix A_{ij} is of a size $P \times P$. It is worth noting that this is an exact relationship. W_j is defined on a large circle which includes any sub-circle regions L_{ij}. From the above formula, we have:

$$L_i = \sum_{j=1}^{M} A_{ij} W_j \tag{4.106}$$

If all the contributions are added together, we can find:

$$\begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_N \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1M} \\ A_{21} & A_{22} & \cdots & A_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1} & A_{N2} & \cdots & A_{NM} \end{bmatrix} \cdot \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix} \tag{4.107}$$

or in a compact form:

$$L = AW \tag{4.108}$$

In the same way, the wavefront phase expressions of the j-th meta-pupil can be projected onto a meta-region of the telescope pupil plane:

$$W_{Tj} = T_j W_j \tag{4.109}$$

where T_j is also a matrix of $P \times P$. The wavefront turbulence at the meta-region of the telescope pupil plane, free from focal anisoplanatism, is the sum of these contributions from all atmospheric layers:

$$W_T = \sum_{j=1}^{M} W_{Tj} \tag{4.110}$$

or in a compact form:

$$W_T = TW \tag{4.111}$$

If the number of guide stars used is larger than the number of the layers, the reverse problem can be solved although, in reality, an infinite number of turbulent layers in the atmosphere exist and the wavefront sensor data are noisy. The distribution of atmospheric turbulence can be estimated in 3D tomography form after an optimum filtering technique is applied. The compensation of the atmospheric turbulence for any target star around the region can be made by summing up the contributions in the target direction.

4.1.8.2 Multi-Conjugate Adaptive Optics

A single deformable mirror cannot correct the wavefront deformation from the atmosphere beyond the isoplanatic patch inside the field of view. In order to extend the isoplanatic patch to the entire field of view, more deformable mirrors are required. This multi-mirror correction technique is called Multi-Conjugate Adaptive Optics (MCAO). If a mirror or lens with the same focal ratio is used after the telescope focus, the combined new optical system is an afocal system. The mirror or lens becomes a collimator. For an afocal system, all the planes in object space above the aperture have their conjugated planes in the image space after the collimator. The relationship of these conjugated planes is (Figure 4.29):

$$\frac{h'}{f'} = \left(1 + \frac{d}{D}\right) - \left(\frac{d}{D}\right)\frac{h}{f} \tag{4.112}$$

where D and d are aperture diameters of the telescope and the collimator, f and f' their focal lengths, and h and h' the distances of the conjugated planes before the telescope aperture and after the collimator.

If there are a number of atmospheric layers which cause phase turbulence, then we can place the same number of deformable mirrors on their conjugated

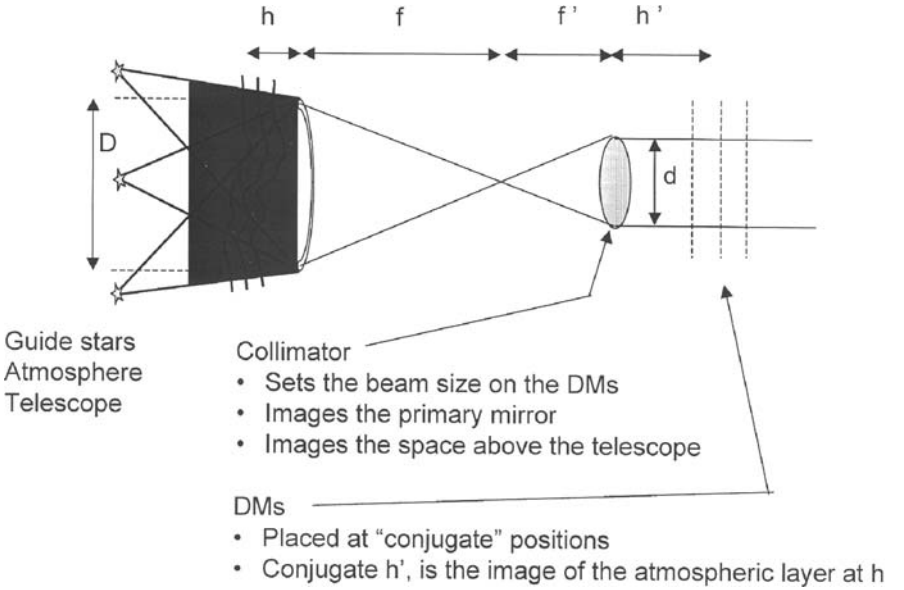


Fig. 4.29. Conjugated planes of a telescope used in multi-conjugate adaptive optics.

positions. These deformable mirrors will compensate the turbulences of the conjugated layers in the atmosphere. In this case, the adaptive optics correction will not be limited to the isoplanatic patch, but will be extended to the entire field of view of the telescope. In this system, multiple guide stars have to be used for gaining 3D details of the atmospheric turbulence through atmosphere tomography. By using this MCAO system, wide field diffraction limited images can be obtained for any ground-based optical telescopes.

Multi-conjugate adaptive optics requires the use of the atmosphere tomography technique; however, the tomography technique may be used without a multi-conjugated adaptive optics.

4.1.9 Adaptive Secondary Mirror Design

For correcting wavefront errors at high temporal frequency, most adaptive optics systems use small deformable mirrors. These small mirrors have a fast response to computer commands. However, small deformable mirrors and related optical components after the telescope focus can significantly reduce the telescope throughput, introduce polarization as well as thermal emissions. Therefore, it is desirable to apply adaptive optics directly to the secondary mirror. Experiences have shown that the telescope throughput with an adaptive secondary mirror can be more than 95% at optical and infrared wavelengths, while the throughput with small deformable mirrors is only 80% at optical wavelengths and only 93% at infrared wavelengths (Lee et al., 2000).

To produce a quiet, light weight, deformable secondary mirror with a correlation length matching the atmosphere Fried's parameter, the mirror has to be extremely thin. The "quiet" here means the deformation of the mirror will not disturb the operation of the telescope. An extremely thin mirror has low resonant frequencies. To overcome this problem, thin air gap damping has to be used. The first successful example of this type of adaptive secondary mirror is the mirror used on the 6.5-m MMT telescope.

In this mirror system, the secondary mirror has a diameter of 642 mm, a magnification of 12, and a thickness of only 2 mm. For grinding and polishing this very thin mirror, two pieces of low thermal expansion Zerodur blanks were used. They were ground to a matching concave and convex spherical shape. A kind of pitch, a liquid that is very viscous at room temperature, is used to glue these two pieces together. The thickness of the glue layer is about 0.1 mm. After gluing together, the top piece can be ground and polished as a single thick piece of glass. When the thickness of the top piece is about 2 mm, the difference between the required hyperboloid and the nearest sphere is only about 80 μm. After the polishing is done, the top piece was separated by using a hot oil bath at 120°C. The mirror will deform by a few wavelengths after the separation.

A big problem of this mirror is the resonance under any excitation. The radial support of the mirror is through the central hole, the axial support is done by 336 magnetic voice coil actuators. Different from piezoelectric actuators, the voice coil actuators act through magnetic force with no stiffness in them. The resonance problem is solved by placing on the back another thick reference mirror. The thickness of the reference mirror is 50 mm. The width of the air gap between two mirrors is 40 μm. At this gap width, there is enough viscosity in the air film between the mirror and reference plate. This very thin air film produces enough damping to extend the control bandwidth up to nearly 1,000 Hz. An aluminum coating is applied on the inner sides of both mirrors, on the reference plate side the coating is around each actuator, on the mirror back it is on the whole surface. The air gap and aluminum coating form very sensitive capacitance displacement sensors. These sensors can be read to <10 nm at a 40 kHz rate, serving as a local metrology system. The capacitance of each capacitor sensor is about 65 μF. The measurement of the air gap through these capacitors can be accurate to 3 nm.

The shape change of the adaptive secondary mirror is driven by electromagnetic force. All 336 magnets are glued to the back of the mirror and 336 voice coils are placed on holes of the matching reference mirror. The distance between coils and magnets is 0.2 mm. To reduce the thermal effect, all the coils are on 10-cm long aluminum fingers for heat conduction to a large aluminum back plate behind the reference plate. Inside the back plate there is circulating cooling fluid. The fluid is a 50/50 mixture of distilled water and methanol.

The control of each coil current is a proportional (P) or proportional-derivative (PD) filter. It does not contain an integrative part in order to avoid any extra phase lag and to increase the bandwidth. However in this system, when the stiffness of some modes is comparable to or larger than

the proportional gain of the loop, a static error is produced. Therefore, a feed forward (FF) force is added. This FF force is derived by multiplying the measured stiffness matrix with the command variation coming from the wavefront computer (Riccardi et al., 2002). In a static condition, the root mean square error of the secondary mirror achieved is about 88 nm. The adaptive secondary mirror of the MMT telescope obtains an overall Strehl ratio of 98% at 10 μm wavelength.

The conjugated plane of a Cassegrain secondary mirror is located below the aperture plane, far away from any atmospheric turbulence layer. Therefore, the correction of wavefront is in a limited isoplanatic patch. By using an adaptive Gregorian secondary mirror, the conjugated plane is at a height of a few hundred meters above the aperture. Therefore, the turbulence at this altitude can be compensated successfully. The correction may influence most of the field of view if the dominant atmospheric turbulence layer is within that height. Unfortunately, in practice, the dominant atmospheric turbulence layer is not at such a height. Therefore, it is difficult to achieve wide angle wavefront correction using an adaptive secondary mirror system.

4.2 Optical Interferometers

Adaptive optics provides a new way of improving the angular resolution of a ground-based telescope beyond the atmospheric seeing limit. To achieve the same or even higher angular resolutions, interferometry techniques should be used. These include the speckle, Michelson, Fizeau, intensity, and amplitude interferometers. In this section, these interferometry techniques are discussed. Related techniques, such as the correlation interferometer and the aperture synthesis technique will be introduced in Chapter 7.

Existing optical and infrared interferometer projects include VLTI, CHARA, LBTI, MRO, COAST, GI2T, SUSI, ISI, PTI, NPOI, SUSI, IOTA, and Keck interferometers.

4.2.1 Speckle Interferometer Technique

An early form of speckle imaging is called the “lucky image” which is a super-resolution technique (Section 1.2.1) involving a high-speed camera with exposure times short enough (100 ms or less) so that the changes of atmosphere during the exposure are minimal. With the lucky imaging, those exposures least affected by the atmosphere (typically around 10%) are chosen and combined into a single image by shifting and adding, so that a much higher resolution than would be possible with a single, longer exposure, which includes all the frames, is derived. In 1970, Antoine Labeyrie (1970) showed that high-resolution

information of an object could be obtained from the speckle patterns through Fourier analysis, leading to the speckle interferometry technique.

In this technique, the object function (stellar source intensity distribution) is retrieved from the measurement of its auto-correlation by using a number of very short exposures. If I_0 is an ideal source distribution in the sky and $\tilde{I}_0$ its spatial frequency spectrum or its Fourier transform, then:

$$I_0(x) = \int \tilde{I}_0(v) \exp(2\pi i v x) dv \tag{4.113}$$

where v is the spatial frequency. Because of atmospheric disturbances, distorted images will be produced. If n short exposures are taken, then a series of distorted images $I_n (n = 1, 2, \cdots, m)$ can be obtained:

$$I_n(x) = \int \tilde{I}_0(v) \tilde{F}_n(v) \exp(2\pi i v x) dv \tag{4.114}$$

where $\tilde{F}_n$ is the instantaneous optical transfer function of the atmosphere during each short exposure. Remember, $\tilde{F}_n$ changes randomly from one exposure to another. In Figure 4.30, the real parts of the two transfer function examples of $\tilde{F}_n$ are shown. In the figure v_A is the spatial cut-off frequency for the telescope aperture and v_L is the atmospheric cut-off frequency for a long exposure. For a

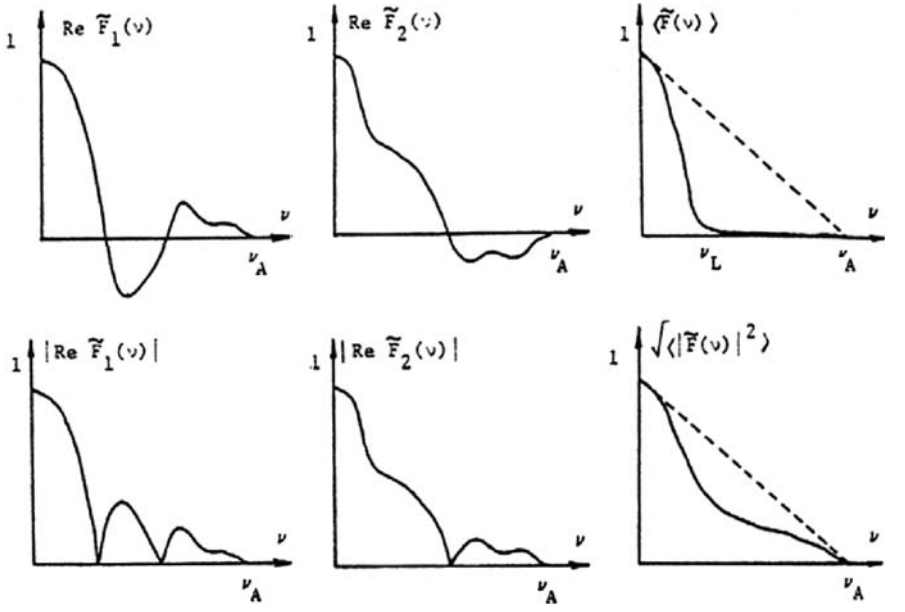


Fig. 4.30. Real part of atmospheric transfer function during short exposures, and their ensemble effect (*above*); the absolute values of them and their ensemble (*below*) (Liu and Lohmann, 1973).

long exposure, the final image is the average of many short exposure images. The image produced is:

$$\begin{aligned} I_L(x) &= \int \tilde{I}_0(v) < \tilde{F}_n(v) > \exp(2\pi i v x) dv \\ I_L(x) &= \frac{1}{N} \sum I_n(x) \\ < \tilde{F}_n(v) > &= \frac{1}{N} \sum \tilde{F}_n(v) \end{aligned} \tag{4.115}$$

The optical transfer function of the atmosphere for a long exposure has a cut-off spatial frequency much lower than the diffraction limit of the telescope. During a short exposure, the transfer function of the atmosphere has contributions beyond this cutoff frequency limit (for a long exposure). However, values in this region may swing in both positive and negative directions. The effects of these contributions cancel out when a long exposure is made.

To overcome this cut-off frequency limit, Labeyrie recorded a series of short exposure images to derive their power spectrum through Fourier transform:

$$\left|\tilde{I}_n(v)\right|^2 = \left|\tilde{I}_0(v)\right|^2 \left|\tilde{F}_n(v)\right|^2 \tag{4.116}$$

The above formula contains no negative terms of the atmospheric modulus transfer function. The average power spectrum of these images will be:

$$\frac{1}{N} \sum \left|\tilde{I}_n(v)\right|^2 = \left|\tilde{I}_0(v)\right|^2 < \left|\tilde{F}_n(v)\right|^2 > \tag{4.117}$$

If the mean square of the atmospheric transfer function $< \left|\tilde{F}\right|^2 >$ is measured from a standard star, then it is possible to obtain the spatial power spectrum term of the target star $\left|\tilde{I}_0\right|^2$. The Fourier transform of the power spectrum is the auto-correlation of the object function (Section 7.3.3). That is:

$$\int \left|\tilde{I}(v)\right|^2 \exp(2\pi i v x) dv = \int I_0(x' + x) I_0(x') dx' \tag{4.118}$$

From this formula, the ideal star image of certain types of objects can be retrieved. This retrieved image is not seeing limited, but diffraction limited.

The speckle interferometry technique provides the modulus in spatial frequency of the observed star, which is equivalent to the visibility fringe amplitude in a Michelson interferometer. However, there is, in fact, no light interference before the image is taken in the speckle interferometer. The fringes are derived from cross-correlation between similar but shifted images.

For simple symmetry cases, such as measuring the diameter of a star or the separation between binary stars, precise results can be obtained. The technique can also be used for stars with known spatial distribution. However, for stars

with an unknown pattern, it is difficult to reconstruct the ideal image without information of the phase of the object transformation. When an asymmetric object is involved, the result will be a mixture of the object with its mirror symmetrical images.

An estimate of the phase of the object transformation can also be made from the same short exposure photographs, but by a different averaging procedure. The procedure uses the auto-correlation of $\tilde{I}_n$. This average is (Knox, 1976):

$$<\tilde{I}(v)\tilde{I}^*(v+\Delta v)> = \tilde{I}_0^*(v)\tilde{I}_0(v+\Delta v)<\tilde{F}(v)\tilde{F}^*(v+\Delta v)> \qquad (4.119)$$

The shift, Δv, is made small compared to the correlation width of $\tilde{F}(v)$. This ensures that the auto-correlation of $\tilde{F}(v)$, evaluated at Δv, has a measurable value with respect to the diffraction limit of the telescope. The equation shows that the auto-correlation contains phase information of $\tilde{I}(v)$, in the form of a phase difference. If $\theta(v)$ represents the phase of the object transformation, then the phase of the auto-correlation of $\tilde{I}(v)$ is given by:

$$\begin{aligned} &phase[\tilde{I}(v)\tilde{I}^*(v+\Delta v)] \\ &= \theta(v+\Delta v) - \theta(v) + phase[\tilde{F}(v)\tilde{F}^*(v+\Delta v)] \end{aligned} \qquad (4.120)$$

From theoretical grounds, Knox (1976) demonstrates that the last term, the phase of the auto-correlation of $\tilde{F}(v)$, should be negligible and it has little effect on the reconstruction. To recover the phase function of the object itself, these phase differences are summed up together outwards from the origin. For small values of phase errors, the phase differences can be approximated by a derivative:

$$\theta(v+\Delta v) - \theta(v) \approx \frac{d\theta(v)}{dv}\Delta v \qquad (4.121)$$

Therefore, the phase is given by an integral operation:

$$\theta(v) = \int_0^v \frac{d\theta(u)}{du} du \qquad (4.122)$$

With the knowledge of the phase, the correct spatial distribution of the object can be recovered.

One type of speckle interferometer is shown in Figure 4.31. After the focal plane, there is a microscope (2). The size of individual speckles in the focal plane is about λf, where f is the focal ratio. For the telescope primary focus, the speckle size is only a few microns, which is too small for most detectors. For matching the detector's pixel size, magnification is necessary. Item (3) is a shutter with a filter. The shutter provides micro-second exposures to freeze

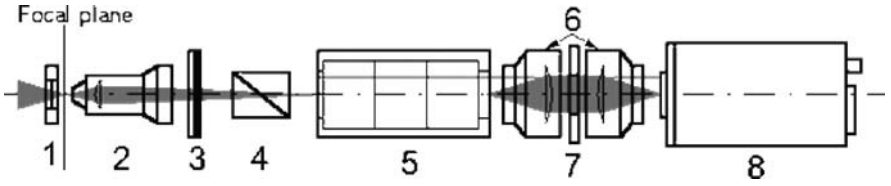


Fig. 4.31. A system layout of a speckle interferometer.

the speckle movement. The filter ensures a quasi-monochromatic light condition. Item (4) is an atmosphere dispersion compensating prism. Items (5)–(8) are detectors. The detector is a 3-stage image intensifier coupled to a full frame CCD camera. At present, the application of the speckle interferometry technique can distinguish binary star structure of 1–5 μarcsec with a stellar magnitude of about 14∼16 mag.

4.2.2 Michelson Interferometer

The basic principle of Michelson stellar interferometry was first proposed by Hippolyte Fizeau in 1862. In 1891, Albert A. Michelson successfully measured the diameter of Jupiter's moons by using this technique. In 1907, Michelson won the Nobel Prize in physics with the interferometer work one of his many achievements. Michelson interferometers with separated telescopes were first realized in radio wavelength in 1945 by Martin Ryle and in optics in 1976 by Antoine Labeyrie. Michelson interferometers used in astronomy is different from Michelson interferometers in optics, which have two arms perpendicular to each other.

Michelson interferometry is an important method to achieve high angular resolution in both optical and radio astronomy. Its principle is shown in Figure 4.32 where two slits are placed at a distance D on the telescope aperture plane, so that monochromatic light from each point of a celestial body with an angular diameter of ϕ_0 will generate cosine interference fringes at the focal plane. Between the maxima there is zero intensity in these fringes. However, fringes from different points of the source will overlap, reducing the fringe's visibility. The fringe's visibility is related to both the slit distance and the angular diameter of the star. Michelson interferometer is a pupil plane interferometer.

The "adding interferometer" is a basic Michelson interferometry technique. From Equation (1.135) for a symmetrical source, the visibility is:

$$V_0(s_\lambda) = \pm \frac{1}{S_0} \int_{-\alpha/2}^{\alpha/2} B(\phi) \cos 2\pi s_\lambda \phi d\phi \tag{4.123}$$

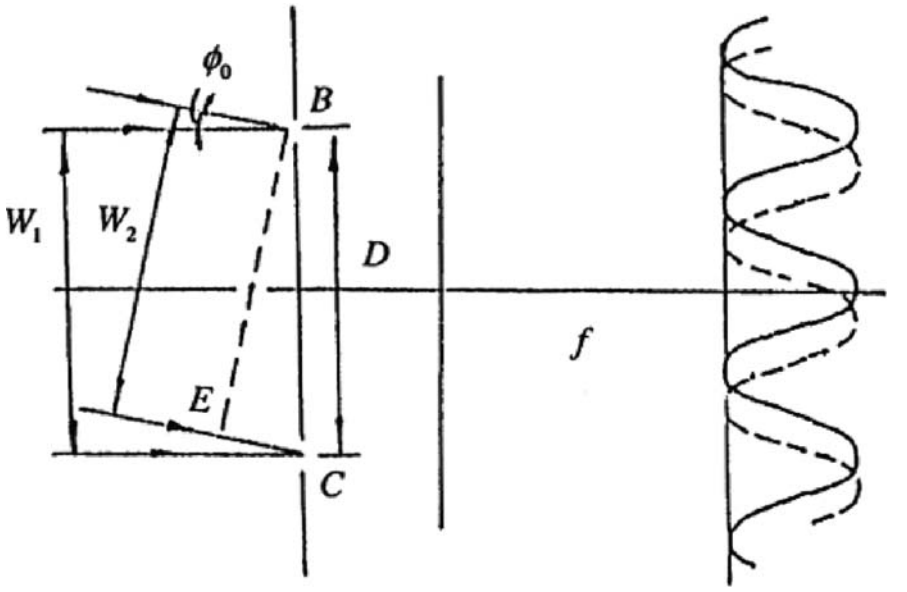


Fig. 4.32. Principle of a Michelson interferometer.

where S_0 is the flux density of the source, $B(\phi)$ the source brightness distribution, $s_\lambda = D/\lambda$, and α the source extent. If the source is a uniform illuminated one, $\alpha B(\phi) = S_0$, the formula reduces to:

$$V_0(s_\lambda) = \pm \frac{\sin 2\pi s_\lambda(\alpha/2)}{2\pi s_\lambda(\alpha/2)} \tag{4.124}$$

As the source extent becomes small compared to the fringe spacing, the visibility approaches unity, but as the fringe spacing becomes very small compared to the source extent, the visibility tends to zero. The visibility is also zero when the source extent is equal to the fringe spacing or integral multiples. By measuring the baseline change ΔD at which the fringe visibility drops to a minimum, the angular diameter of the source is derived.

The Michelson interferometer may have a baseline extending beyond the diameter of the telescope known as the stellar interferometer as shown in Figure 4.33. For this arrangement, ΔD in the above formula should be replaced by ΔL, the change of the baseline L. Using this arrangement, Michelson achieved a resolution of 0.0047 arcsec.

In the 1980s, the Centre de Recherche en Geodynamique et Astrometrie (CERGA) interferometer obtained fringes from two 25 cm telescopes with a baseline of 35 m. These two telescopes are able to move along a precise rail in a south-north direction (Figure 4.34). In between, there is an optical path length equalizer (OPLE) on a movable platform to compensate the optical path

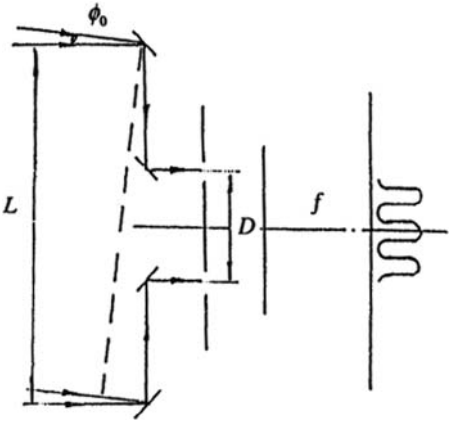


Fig. 4.33. Michelson stellar interferometer.

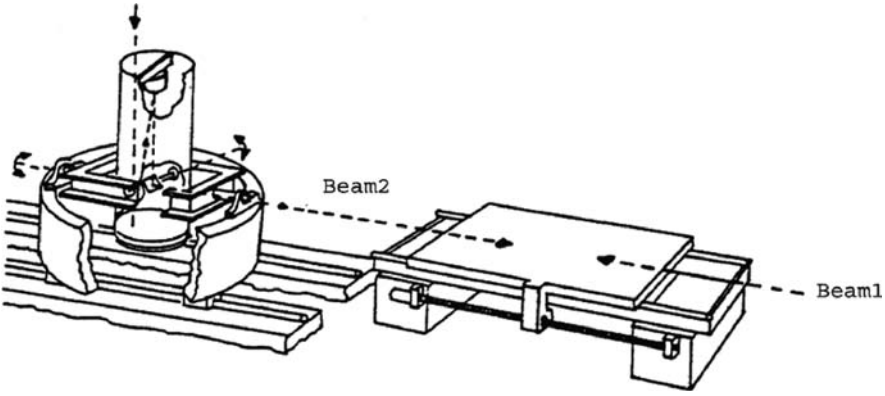


Fig. 4.34. Precision platform of the CERGA interferometer (Marriotti and Di Benedetto, 1984).

difference between two beams. The beams were combined at a prism combiner (Figure 4.35). The Michelson interferometer requires high precision in the path length compensation. In Section 4.2.4, the coherence length, which is the path length maximum difference to obtain interference, is defined. In optical range, it is approximately:

$$L = \frac{\lambda^2}{n\Delta\lambda} \tag{4.125}$$

where λ is the central wavelength and n the refraction index. If the spectral width $\Delta\lambda$ is of the order of angstroms, the coherence length is about 1 mm. This requires a very accurate control of the beam's phase to a fraction of wavelength.

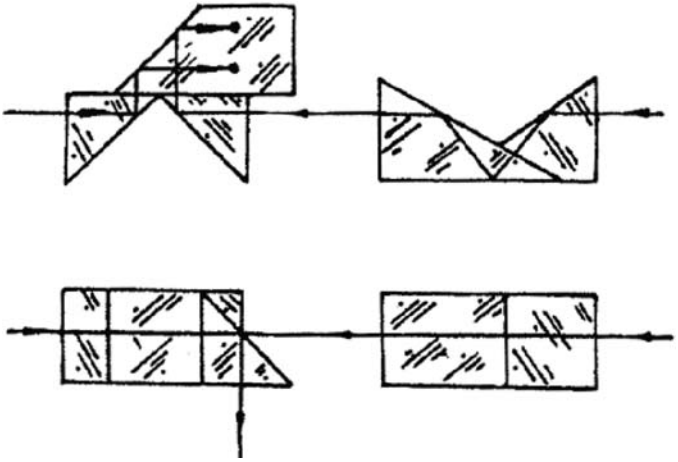


Fig. 4.35. A beam combiner for the Michelson interferometer.

The optical path length error variation between two beams through the atmosphere is (Davis, 1997):

$$\sigma_{OPL} = 0.42\lambda_0 \left(\frac{L}{r_0}\right)^{5/6} \tag{4.126}$$

where L is the baseline length, r_0 the Fried parameter, and λ_0 the wavelength. Since r_0 varies as $\lambda^{6/5}$, the atmosphere induced path length error is mostly independent of wavelength. Therefore, interference in the long wavelength radio region is much easier than in the optical regime. The formula is valid for a baseline within the outer scale of atmospheric turbulence. Beyond that, the path length error increases much more slowly and will reach a limiting value depending on the site and meteorological conditions.

The atmospheric seeing also affects the wavefront shape within each aperture. This also reduces the visibility unless special efforts are made for its compensation.

The path difference from the source to the apertures of an interferometer ranges from zero to 0.9 times baseline depending on star position relative to the baseline. Therefore, the most important component of an OPLE system is the optical delay line. Two types of delay line are shown in Figure 4.36, where the combiners are beam-splitters. Since a phase change exists due to reflection, the two beams are complementary in intensity. Assuming no loss in reflection and refraction, the intensities of the two beams for a point source at delay tracking position are:

$$\begin{aligned} I_1 &= I_0(1 + |\gamma_{12}(0)| \cos \phi(t)) \\ I_2 &= I_0(1 - |\gamma_{12}(0)| \cos \phi(t)) \end{aligned} \tag{4.127}$$

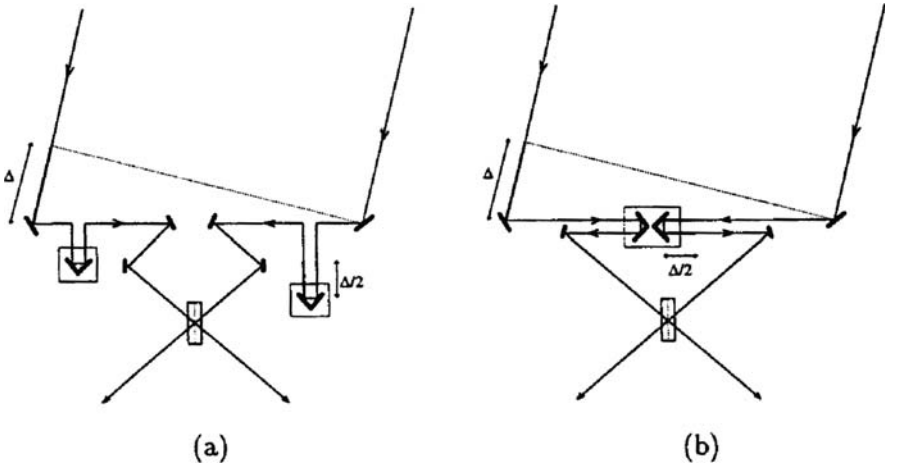


Fig. 4.36. Two configurations of optical delay line: (**a**) A dual delay line system and (**b**) a differential delay line.

where I_0 is the incident beam intensity, $\gamma_{12}(0)$ the degrees of coherence, and $\phi(t)$ the varying random phase fluctuation dominated by atmospheric turbulence. Using the average of the squared intensity difference for a large number of samples, the correlation between two beams which is the square of the degree of coherence can be obtained:

$$< (I_1 - I_2)^2 > = 4I_0^2 |\gamma_{12}(0)|^2 < \cos^2 \phi(t) > = 2I_0^2 |\gamma_{12}(0)|^2 \qquad (4.128)$$

where the angular brackets means a time average. The time average of the squared cosine of a random number is 0.5. This formula shows an optical technique to obtain the correlation between two beams. The details of correlation interferometer and aperture synthesis telescope are discussed in Sections 7.3.1 and 7.3.2.

Another beam correlation technique involving path modulation was used by Shao (Davis, 1997). In this method, the path length of one beam is modulated in a triangular waveform with the amplitude of one wavelength. During each cycle of the sweep, the output intensity is put into four time bins, denoted as A, B, C, and D. Then the phase and amplitude of the fringe are:

$$\begin{aligned} Phase &= \tan^{-1}\left(\frac{D-B}{C-A}\right) \\ Amplitude &= \sqrt{(C-A)^2 + (D-B)^2} \end{aligned} \qquad (4.129)$$

A special type of Michelson interferometer is the nulling interferometer. The nulling interferometer is an important tool for studying faint planets in close

proximity to a much brighter star. The technique was first proposed by Bracewell and MacPhie (1979). In a two element interferometer, the null is achieved by shifting the phase of one element by half a wavelength to cancel out the light at the pointing center. At the same time, the light of a nearby planet, which is at an angle θ away from the center, is added together. The light has a phase difference of $\pi + \phi = \pi + \theta \cdot D/\lambda$, where D is the separation of the two beams and $\phi = \theta \cdot D/\lambda$ the phase (Figure 4.37). By adjusting the baseline, light from the planet interferes constructively.

A coronagraph is an instrument similar to a nulling interferometer. The coronagraph provides details of faint coronal features of the sun. However, the coronagraph uses a masking technique in the focal plane. It does not rely on the interferometer technique.

In a multi-aperture telescope array, a very important correlation interferometer is named the aperture synthesis telescope, which forms an imaging of celestial sources through the Fourier transform of the visibility as a function of baseline projection. This imaging technique was developed first in the radio region.

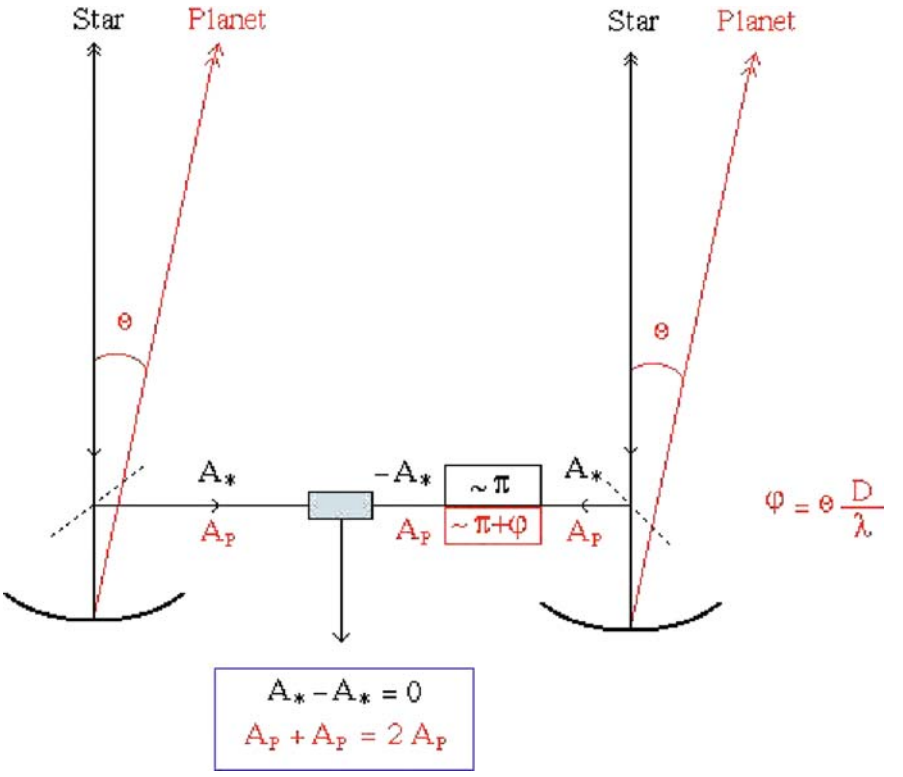


Fig. 4.37. Principle of a nulling interferometer.

4.2.3 Fizeau Interferometry

Optical interferometry can be made either by division of the amplitude or division of the wavefront. By division of the amplitude, usually through reflection and refraction in a semi-transparent beam-splitter, amplitude interferometers are constructed. By division of the wavefront, usually with two apertures, wavefront interferometers are constructed. The Michelson interferometer is a wavefront interferometer.

The beam combination of a Michelson interferometer is done in the pupil plane. Therefore, the traditional Michelson interferometer is a pupil plane interferometer. If the beam combination is in the focal plane, another type of wavefront interferometer, a focal plane one, is constructed. The Fizeau interferometer, which has a similar arrangement as the Michelson interferometer, produces an interference pattern in the focal plane so that it is a focal plane interferometer or a Fizeau mode of the Michelson interferometer.

A Fizeau interferometer can involve two or more divisions of wavefront. If two wavefronts of circular apertures are involved, the image pattern is an Airy pattern of the aperture, modulated by Young's two slit fringes with the spacing set by the baseline. The central fringe of the image corresponds to zero path length difference between the two wavefronts. At this point, the image is achromatic as beams interfere constructively for any wavelengths.

The Michelson interferometer is also known as the Michelson mode or pupil plane mode of the Michelson interferometer. In Michelson mode, the beams from two apertures are combined in pairs so that the visibility of each baseline is measured. This visibility corresponds to a particular spatial frequency of the source in a direction conjugated to the baseline direction. To gather information on other spatial frequencies, more baselines are required. Two major limitations exist in Michelson interferometry: (a) limited field of view and (b) difficult in the beam pair reconfiguration. In the Michelson mode, there are in fact no fringe patterns, but the brightness, on the detector. The brightness is modulated by adjusting the path length difference. Zero path length difference corresponds to a single achromatic fringe point. The field of view acquired is related to the bandwidth. For a relative bandwidth of 0.1, the maximum fringe number observed is about 10. For studying other spatial frequencies of the source, the visibility functions of different baselines are needed. Beam pair reconfiguration is necessary. An array of 30 telescopes has 435 baselines. A continuous beam pair reconfiguration during observation involves adjusting very delicate and accurate beam combiners and optical delay lines. Any design for a quick reconfiguration of a Michelson interferometer array is extremely difficult.

The field of view of the Fizeau interferometer, which is determined solely by the off-axis optical performance of the system, is wider. The Fizeau interferometer also moves the beam combination from the pupil plane to the focal plane. Therefore, the beam reconfiguration is done by simply blocking or unblocking beams from array telescopes. The beam blocking and unblocking only takes seconds instead of tens of minutes. The Fizeau interferometer also can use more

array telescopes in the image formation as long as they are all co-phased. More sub-apertures (or telescopes) are involved; sharper interference patterns (or fringes) are formed.

The main difficulty of the Fizeau interferometer is the co-phase requirement of each sub-aperture beam. In the Michelson interferometer, the knowledge of the baselines and optical paths to a fraction of wavelength is sufficient as the path length modulation can provide further information on path length difference. However, in the Fizeau interferometer, fine control of the phase of all combining beams to a fraction of a wavelength is necessary. Otherwise, the interference pattern will not be formed. For a ground-based Fizeau interferometer array, many factors influence the path lengths of sub-telescopes. The atmospheric phase fluctuation is the most important one. Even so, a stable phase lock had been achieved between two sub-telescopes of the old MMT telescope decades ago using less advanced sensors (Angel et al., 2002). Fringes for on- and off-axis objects had also been recorded.

The Fizeau interferometer can also be used as a nulling interferometer if the beam phase difference in the center is intentionally set as half wavelength. In 1995, a Fizeau interference image was obtained from the Cambridge Optical Aperture Synthesis Telescope (COAST) in optical wavelength.

There are a number of on-going Fizeau interferometer projects including the Center for High Angular Resolution Astronomy (CHARA) array and the Large Binocular Telescope Interferometer (LBTI). In the LBTI Fizeau interferometer, two individual beams will be theoretically locked in a co-phase condition and the interference pattern will appear on their common focus. The US Stellar Imager (SI) project is a space Fizeau interferometer of 10–30 mirrors of 1 m size with the maximum baseline of 500 m in space. The shape of the mirrors is either spherical or flat. All mirrors are co-focused as well as co-phased through optical measuring devices to form a wide field interferometer image. The interference pattern formed by flat mirrors is discussed in Section 9.2.4.

4.2.4 Intensity Interferometer

As the baseline increases, the phase fluctuation due to the atmospheric disturbance increases. Therefore, there is a practical limit for a ground-based passive optical interferometer. This limit is about several hundred meters to a few kilometers equivalent to a resolution less than 1 marcsec or smaller. To obtain an even higher angular resolution, Hanbury Brown and Twiss (1954) developed another technique called intensity interferometry. In intensity interferometry, the atmosphere and instrument induced phase fluctuations are ignored and the interference between beams from stellar objects is not required. Instead, the power spectrum in the spatial frequency domain of the stellar objects is measured. The power spectrum density of the source in the spatial frequency domain equals approximately the cross-correlation of intensities of two beams received by a pair of apertures over a baseline.

The intensity interferometry is based on the optical coherence theory. In physical optics, white light is a type of Gaussian random variable. Its electric vector is a superposition of Fourier components with the amplitudes and phases statically independent and randomly distributed. The real part of the electric vector can be represented as a Fourier integral (Brown, 1974):

$$V^{(r)}(t) = \int_0^\infty a(v)\cos[\phi(v) - 2\pi vt]dv \tag{4.130}$$

where $a(v)$ is the amplitude, v is frequency, and $\phi(v)$ the phase. If the imaginary part is included, then the complex optical vector can be represented as:

$$V(t) = V^{(r)}(t) + iV^{(i)}(t) = \int_0^\infty a(v)\exp[i(\phi(v) - 2\pi vt)]dv \tag{4.131}$$

where $V^{(i)}(t)$ is the conjugate function of $V^{(r)}(t)$. Sometimes the electric vector can be expressed as an integral over an even wider frequency range:

$$V(t) = \int_{-\infty}^\infty v(v)\exp[-2\pi ivt]dv \tag{4.132}$$

where

$$v(v) = \frac{1}{2}a(v)\exp[i\phi(v)] \tag{4.133}$$

Therefore, $V(t)$ is a random variable defined over all values of t. Astronomical observations can only be carried out over a finite time interval: $-T < t < T$. So a truncated optical vector $V_T(t)$ over the above time interval is defined. The time average of the light intensity is:

$$\begin{gathered}\tfrac{1}{2}\langle V^*(t)V(t)\rangle = \left\langle V^{(r)2}(t)\right\rangle = 2\int_0^\infty G(v)dv \\ G(v) = \lim_{T\to\infty}\frac{|v_T(v)|^2}{2T}\end{gathered} \tag{4.134}$$

where $G(v)dv$ includes contributions of light intensity over the frequency (Note: When temporal frequency is used, we drop the word "temporal.") range from v to $v+dv$. The above formula represents a fact of energy conservation that the light intensity in the time domain is the same as that in the frequency domain.

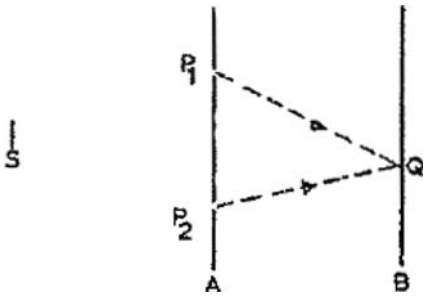


Fig. 4.38. Two pin holes split the source wavefront into two parts.

If two pinholes on a screen are illuminated by a light source, the electrical field at a point Q on the second screen is (Figure 4.38):

$$V_Q(t) = k_1 V_1(t) + k_2 V_2(t+\tau) \tag{4.135}$$

where $V_1(t)$ and $V_2(t)$ are optical field vectors at two pinholes, k_1 and k_2 the complex amplitude transmission factors of the two beams from the pinholes to point Q, and τ the time difference for the transmission between the pinholes and point Q. By ignoring a factor of $1/2$, the intensity on the screen is:

$$I_Q = < V_Q^*(t) V_Q(t) > \tag{4.136}$$

This is:

$$I_Q = |k_1|^2 I_1 + |k_2|^2 I_2 + 2\mathrm{Re}[k_1 * k_2 \Gamma_{12}(\tau)] \tag{4.137}$$

where $\Gamma_{12}(\tau)$ is the mutual coherence function between the two beams:

$$\Gamma_{12}(\tau) = < V_1^*(t) V_2(t+\tau) > \tag{4.138}$$

From the definition of the autocoherence function, there exists:

$$\Gamma_{ii}(0) = < V_i^*(t) V_i(t) > \tag{4.139}$$

By introducing a dimensionless term, the complex degree of coherence, $\gamma_{12}(\tau) = (I_1 I_2)^{-1/2} \Gamma_{12}(\tau)$, the formula becomes:

$$I_Q = |k_1|^2 I_1 + |k_2|^2 I_2 + 2\{I_{1Q} I_{2Q}\}^{1/2} \mathrm{Re}[k_1 k_2 \gamma_{12}(\tau)] \tag{4.140}$$

where I_{1Q} and I_{2Q} are the intensities produced at Q by the radiation of the two pinholes separately.

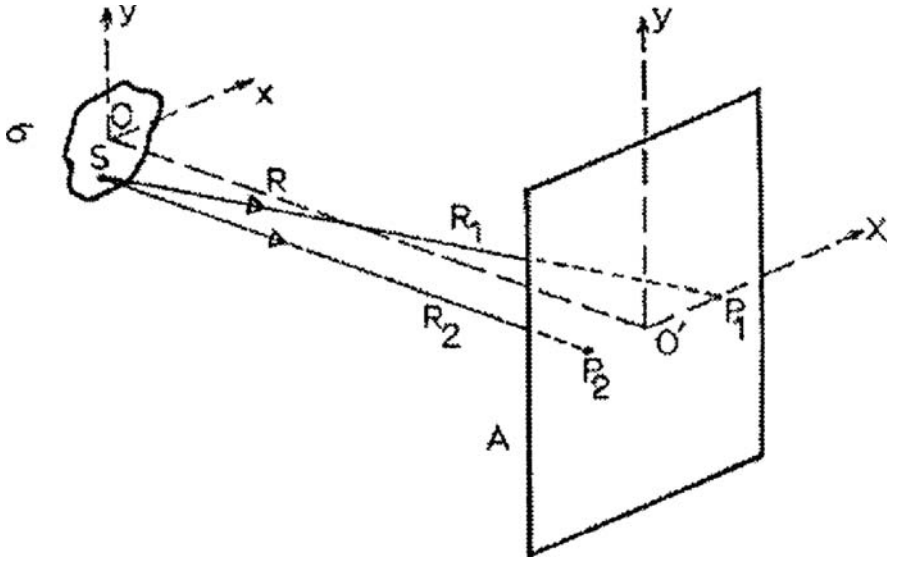


Fig. 4.39. Spatial coherence and the angular size of the source.

Assuming there is no time delay between two pinholes and point Q (Figures 4.38 and 4.39), then the complex degree of coherence in the above formula can be replaced by its value at the origin $\gamma_{12}(0)$. However, if the separation between the pinholes, or the angular size of the light source, or the wavelength of the light is changed, the degree of coherence will vary, this is referred to as spatial coherence.

In case the light source lies in a plane parallel to and far away from the screen, it remains equidistant from the two holes and the optical path difference between them is far less than the light coherence length $(c/\Delta v)$. For a quasi-monochromatic light source, namely $\Delta v/v_0 << 1$, dividing the source into a large number of small independent patches and letting $V_{m1}(t)$ and $V_{m2}(t)$ be the complex wave amplitudes at two pinholes due to a sub-area source $d\sigma_m$, then the mutual coherence function is expressed as:

$$\Gamma_{12}(0) = <V_1^*(t)V_2(t)> = \sum_m V_{m1}(t)V_{m2}(t) + \sum_{m\neq n}\sum V_{m1}(t)V_{n2}(t) \qquad (4.141)$$

The second term in this expression is zero as $V_{m1}(t)$ and $V_{n2}(t)$ come from independent sub-areas of the light source. The above formula becomes:

$$\Gamma_{12}(0) = \sum_m V_{m1}(t)V_{m2}(t) \qquad (4.142)$$

If the intensity over the unit area of the light source is I and the distances from two holes to the light source are R_1 and R_2, respectively, then

$$\Gamma_{12}(0) = \int_{\sigma} \left(\frac{I(S)}{R_1 R_2} \right) \exp[2\pi i (R_1 - R_2)/\lambda] dS \tag{4.143}$$

The complex degree of coherence is

$$\gamma_{12}(0) = (I_1 I_2)^{-1/2} \int_{\sigma} \left(\frac{I(S)}{R_1 R_2} \right) \exp[2\pi i (R_1 - R_2)/\lambda] dS \tag{4.144}$$

where I_1 and I_2 are:

$$I_i = \int_{\sigma} (I(S)/R_i^2) dS \tag{4.145}$$

If x and y are the coordinates on the source plane and X_1 and X_2 the coordinates of two pinholes on the screen lying on the X axis, then:

$$R_1 - R_2 \approx (X_1^2 - X_2^2)/2R - (X_1 - X_2)x/R \tag{4.146}$$

with $R_1 \approx R_2 \approx R$, then:

$$\gamma_{12}(0) = \frac{\exp(i\psi) \int_{\sigma} \int I(x,y) \exp[-2\pi i (X_1 - X_2)x/\lambda R] dx dy}{\int_{\sigma} \int I(x,y) dx dy}$$
$$\psi = (2\pi/\lambda)[(X_1^2 - X_2^2)/2R] \tag{4.147}$$

This is the important Van Cittert–Zernike theorem on spatial coherence (Section 7.3.3). It states that the complex degree of coherence of light beams at two separate locations is approximately a normalized Fourier transform of the source intensity distribution. The coefficient $\exp(i\psi)=(2\pi/\lambda_0)(OP_1–OP_2)$ in the formula represents a relative phase-shift and is unity in our case, where O is the origin of the source and P_1 and P_2 the positions of two pinholes (Figure 4.39). The complex degree of coherence is also known as the visibility function in interferometers.

If a plane wave from a distant light source passes a beam splitter M and arrives at point P_1 and P_2 with $MP_1 = MP_2$, then the degree of coherence, if we can measure it, would be $\gamma_{12}(0) = 1$ (Figure 4.40). If we move one point by a distance of $l = \tau \cdot c$ to P_1', then the spatial coherence is not changed. However, there is a time delay between the fields at these two points. Now, the two fields

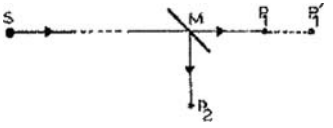


Fig. 4.40. Temporal coherence and a time delay of one beam.

are $V_1(t)$ and $V_2(t+\tau)$, respectively. The mutual coherence function for temporal coherence is [ref. Equation (4.142)]:

$$\Gamma_{12}(\tau) = <V_1^*(t)V_2(t+\tau)> = 4\int_0^\infty G_{12}(v)\exp(-2\pi iv\tau)dv \qquad (4.148)$$

Therefore, the complex degree of coherence is:

$$\gamma_{12}(\tau) = \frac{\int_0^\infty G_{12}(v)\exp(-2\pi iv\tau)dv}{\int_0^\infty G_{12}(v)dv} \qquad (4.149)$$

$$G_{12}(v) = \lim_{T\to 0}\frac{v_{T1}{}^*(v)v_{T2}(v)}{2T}$$

where $G_{12}(v)$ is the mutual spectrum density in the frequency domain of the two beams at the two points. The mutual coherence varies as a Fourier transformation of the mutual frequency spectrum density of the two light beams. If the spectra of two beams are identical as in our case, then the mutual coherence function is the auto-correlation function of the light. This formula can also be derived in mathematics as the Weiner–Khinchin theorem (Section 7.3.3).

If the mutual spectrum density is uniformly distributed over a narrow bandwidth Δv around a mean frequency v_0 and $\Delta v << v_0$, then we have another very important relationship for the complex degree of coherence:

$$\gamma_{12} = \left[\frac{\sin \pi v\tau}{\pi v\tau}\right]\exp(-2\pi iv_0\tau) \qquad (4.150)$$

This function reaches its first zero at a time delay of $\tau_0 = 1/\Delta v$. This time interval is referred to as the coherence time of the light and the corresponding distance in space is the coherence length $l_0 = c/\Delta v$. A narrow spectrum width corresponds to a long coherence time and coherence length. Beyond this time or length, the interference between two beams is not possible.

From the above discussion, the basic principle of the intensity interferometer can be found. In an intensity interferometer, the spatial intensity distribution of the source is obtained through the cross-correlation of two beam intensities collected at different locations. For a light source which forms images simultaneously at two locations, the image intensity at each location can be expressed as:

$$I_{i=1,2} = V_i^*(t)V_i(t) \qquad (4.151)$$

The cross-correlation between these intensities is:

$$\begin{aligned} <I_1(t)I_2(t+\tau)> &= <V_1{}^*(t)V_1(t)V_2{}^*(t+\tau)V_2(t+\tau)> \\ &= <V_1^{(r)2}(t)V_2^{(r)2}(t+\tau)> + <V_1^{(r)2}(t)V_2^{(i)2}(t+\tau)> \\ &\quad + <V_1^{(i)2}(t)V_2^{(r)2}(t+\tau)> + <V_1^{(i)2}(t)V_2^{(i)2}(t+\tau)> \end{aligned} \tag{4.152}$$

Since $V_i^{(r)}(t)$ and $V_i^{(i)}(t)$ are Gaussian random variables, then

$$\begin{aligned} <V_1^{(r)2}(t)V_2^{(r)2}(t+\tau)> &= \frac{1}{4}\bar{I}_1\bar{I}_2 + 2\Big[<V_1^{(r)}V_2^{(r)}(t+\tau)>\Big]^2 \\ &= \frac{1}{4}\bar{I}_1\bar{I}_2 + \frac{1}{2}\{\mathrm{Re}[\Gamma_{12}(\tau)]\}^2 \end{aligned} \tag{4.153}$$

Similarly we have expressions for other terms:

$$\begin{aligned} <V_1^{(i)2}(t)V_2^{(r)2}(t+\tau)> &= \tfrac{1}{4}\bar{I}_1\bar{I}_2 + \tfrac{1}{2}\{\mathrm{Im}[\Gamma_{12}(\tau)]\}^2 \\ <V_1^{(r)2}(t)V_2^{(i)2}(t+\tau)> &= \tfrac{1}{4}\bar{I}_1\bar{I}_2 + \tfrac{1}{2}\{\mathrm{Im}[\Gamma_{12}(\tau)]\}^2 \\ <V_1^{(i)2}(t)V_2^{(i)2}(t+\tau)> &= \tfrac{1}{4}\bar{I}_1\bar{I}_2 + \tfrac{1}{2}\{\mathrm{Re}[\Gamma_{12}(\tau)]\}^2 \end{aligned} \tag{4.154}$$

Substituting these into Equation (4.152), then,

$$<I_1(t)I_2(t+\tau)> = \bar{I}_1\bar{I}_2 + \Gamma_{12}^2(\tau) = \bar{I}_1\bar{I}_2[1 + |\gamma_{12}(\tau)|^2] \tag{4.155}$$

In the observation, only the intensity fluctuations are our main concern. The above formula can be expressed as:

$$<I_1(t)I_2(t+\tau)> = \bar{I}_1\bar{I}_2 + <\Delta I_1(t)\Delta I_2(t)> \tag{4.156}$$

Therefore, we have:

$$<\Delta I_1(t)\Delta I_2(t+\tau)> = |\Gamma_{12}(\tau)|^2 = \bar{I}_1\bar{I}_2|\gamma_{12}(\tau)|^2 \tag{4.157}$$

This is the base of the intensity interferometer. If the beams at two separated locations are partially coherent, then the fluctuations of the intensity are correlated. Their cross-correlation is proportional to the square of the degree of coherence $\gamma_{12}(\tau)$, while the degree of coherence is the Fourier transform of the source spatial intensity distribution. In the above analysis, a linear polarized radiation is assumed. If unpolarized light is discussed, then the equation is:

$$<\Delta I_1(t)\Delta I_2(t+\tau)> = |\Gamma_{12}(\tau)|^2 = \frac{1}{2}\bar{I}_1\bar{I}_2|\gamma_{12}(\tau)|^2 \tag{4.158}$$

The advantage of an intensity interferometer is that the interference of the light is not required so that the telescopes used are simply light collectors. For obtaining the spatial intensity distribution of a source, cross-correlation between intensities collected by two telescopes located at a distance is performed. Two photoelectric receivers, such as photomultiplier tubes, with an accurate clock circuit are used at the focii of both telescopes. The cross-correlation coefficient calculated from the output signals is used to determine the spatial distribution of the source.

In 1956, Hanbury Brown and Twiss built the Narrabri Intensity Interferometer with two telescopes of 6.5 m diameter in Australia. The telescopes as light collectors had 270 small hexagonal segments and had an image size of 2.5 cm. Both telescopes can move on a circular track with a diameter of 188 m. The smallest angular diameter measured by this interferometer was about 0.00047 arcsec, a very small angle even by today's standard. The bandwidth of the instrument was very narrow (about 100 MHz) so that it could only be used for very few bright stars (magnitude less than 2.5). The interferometer was closed after all bright stars were observed.

The major disadvantage of an intensity interferometer is that the exact source image is difficult to form as the visibility phase information is not preserved in the cross-correlation, the same as in the speckle interferometer case. And the number of stars observed is limited due to a narrow bandwidth used. The bandwidth in the Michelson interferometer is about 1,000 GHz, while that in the intensity interferometer is about 1,000 times smaller. In principle, the bandwidth can be expanded to increase the light signal received, but the signal to noise ratio also increases. Besides these, the cross-correlation between two intensity signals is usually difficult to establish.

An intensity interferometer can work over a long baseline so that it is possible to realize the intensity interference by using existing telescopes, such as the gamma-ray Cherenkov telescopes (Chapter 9), which are separated at a distance of several hundred meters to a few kilometers. The possible resolution would be better than 10^{-5} arcsec.

4.2.5 Amplitude Interferometer

The same as the speckle interferometer, amplitude interferometry is a single aperture interference technique. However, the speckles obtained in the speckle interferometer are in the focal plane while the fringes obtained in the amplitude interferometer are in the pupil plane. The principle of the amplitude interferometer is to obtain interference fringes by self-shearing the wavefront in the pupil plane. For a coherent wavefront of a point source, the fringes produced are coherent holographs. If the wavefront is from a source of finite angular size, the light from different sub-areas of the source is incoherent. The interference fringes are called incoherent holographs. An amplitude interferometer is also named the

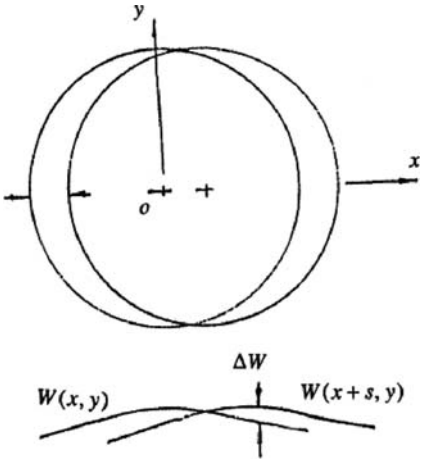


Fig. 4.41. Principle of shearing interferometer.

incoherent holograph, or amplitude splitting interferometer, or pupil plane interferometer.

The principle of a shearing interferometer has been discussed in Section 4.1.2.3. Figure 4.41 shows an original and a lateral sheared wavefront. If the wavefront phase is $W(x, y)$ and its amplitude $A(x, y)$, then the radiation can be expressed as:

$$U(x, y) = A(x, y)e^{jkW(x,y)} \tag{4.159}$$

When the beam is laterally sheared with a distance of s, the phase of the new wavefront will be $W(x + s, y)$. The wavefront phase difference after shearing is $\Delta W = W(x, y) - W(x + s, y)$. When the amount of lateral shearing s is small, ΔW can be expressed as $(\partial W/\partial x)s$. Because phase differences exist between two wavefronts as:

$$\Delta W = W(x, y) - W(x + s, y) \tag{4.160}$$

the intensity of the fringes become:

$$I(x, y) = 2A(x, y)(1 + \cos(\frac{2\pi}{\lambda}\Delta W)) \tag{4.161}$$

The wavefront shear includes lateral, folding, rotational, and mirror-imaged ones. The mirror-imaged and rotational shears are more often used in astronomy.

A mirror-imaged shear uses both the original off-axial wavefront and its mirror image. If an aberration free wavefront formed by a point source with an off-axis angle of α is used, the wavefront is a plane tilted by the same angle α with respect

to the pupil plane. When this wavefront and its mirror image interfere, the fringes produced are parallel lines with a separation of $\lambda/2\alpha$. Assuming the beam amplitude is unity, the observed illumination is (Roddier, 1989):

$$I(\vec{r}) = \left[1 + \cos\left(\frac{2\pi}{\lambda} 2\alpha \cdot \vec{r}\right)\right] \tag{4.162}$$

The wavefront difference between these two wavefronts is $\Delta W = 2\vec{r}\alpha$. When the instrument is illuminated by an incoherent source with a brightness distribution of $O(\alpha)$, then the illumination is a sum of the illuminations produced by each object point. It is given:

$$I(\vec{r}) = \int O(\alpha)[1 + \cos(4\pi\alpha \cdot \vec{r}/\lambda)]d\alpha \tag{4.163}$$

or

$$I(\vec{r}) = \widehat{O}(0) + \mathrm{Re}\widehat{O}(2\alpha/\lambda) \tag{4.164}$$

where $\widehat{O}(\alpha)$ is the Fourier transform of the object brightness distribution $O(\alpha)$. The interferogram obtained is a DC term (or constant term) and the real part of the object Fourier transform. Taking the Fourier transform of this interferogram, three patterns will be derived: a Dirac impulse function at the origin, the source, $O(\alpha)$, and its mirror image, $O(-\alpha)$, in both sides. The source intensity distribution can, therefore, be resolved as long as these images do not overlap.

The problem is that as the number of points of the source increases the signal to noise ratio reduces rapidly. Therefore, it is necessary to restrict the angular size of the source being smaller than the telescope aberration or the seeing. Atmospheric disturbance distorts the fringes, but not the fringe's visibility, so that the technique is diffraction limited.

An alternative way to solve the problem is through two interferograms: a real part and an imaginary part of the source Fourier transformation. This can be realized by inserting a quarter wave plate in the optical path immediately after one fringe pattern is taken. Of course, the switching time between these two interferograms must be shorter than the atmospheric coherence time so that the wavefront distortions remain the same for both exposures. This is difficult in optics, but is feasible in infrared wavelengths.

Double Fourier transform interferometry is another related technique. It records interferograms as a function of phase delay. The Fourier transform taking the phase delay as a variable provides spatial spectral information of the sources.

In rotational shearing interference, polar coordinates are used in the expression. The wavefront difference is then:

$$\Delta W = W(\rho, \theta) - W(\rho, \theta + \phi) \tag{4.165}$$

where ϕ is the shearing angle. The expression of the incoming wavefront is:

$$W(\rho,\theta) = \sum_{n}^{k} \sum_{l}^{q} \rho^n (a_{nl} \cos l\theta + b_{nl} \sin l\theta) \tag{4.166}$$

After the shearing, the terms for $l = 0$ disappear. And all terms with a small l become negligible. Therefore, this technique is not sensitive to the telescope aberrations and atmospheric disturbance.

A classical design of rotational shearing devices is shown in Figure 4.42. After the light beam hits a beam splitter, two output beams arrive at roof prisms A and B. Between the two prisms, there is a small rotational angle of φ relative to the incoming beam direction. Interference will take place after these two beams are recombined. The disadvantage of this design is that the polarization states of the two output beams do not always match. This affects the fringe visibility. Therefore, unless only a small shearing angle is used, filters and polarizers may be necessary both at the input and output ends of the instrument.

In the amplitude interferometry, a number of methods are developed for estimating the phase of the Fourier transform of the source distribution. One is a triple shearing interferometry, where three overlapping pupil images with different shearing angles are recorded. The phase difference of one pair is the sum of the phase differences of the remaining two pairs. Another way is to record two interferograms at the same time, one with a shearing angle α, and the other with an angle 2α. From Figure 4.43 one can find that the measured frequency vectors are linked to the phase closure map. This reduces the noise in the image reconstruction.

During the observation of binary stars, each star will form a set of interference fringes. Because two sets of fringes come from different light sources, their intensities are superimposed and the Fourier transform of the intensity pattern will produce the spatial distribution of the binary stars. A high spatial resolution will

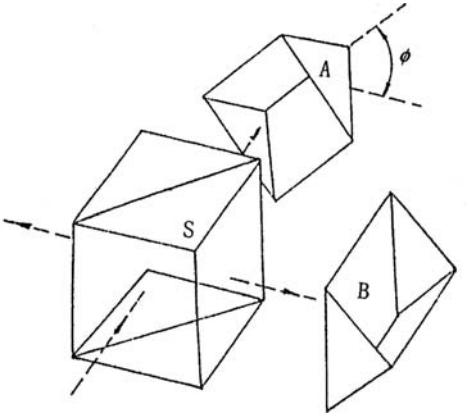


Fig. 4.42. Rotational shearing interferometer. (**A**) Rotating roof prism, (**B**) fixed roof prism, and (**S**) beam splitter.

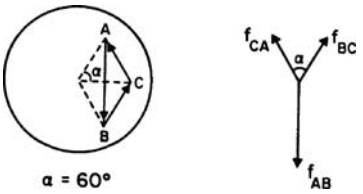


Fig. 4.43. Phase closure loop using two interferograms.

be obtained with a small rotational shearing angle. By increasing the rotational angle, the visibility of interference fringes decreases. To achieve the maximum contrast of the fringe pattern, a proper polarizer should be used on the roof prism to ensure that the coherence light beams have the same polarization direction. The biggest problem in applying an amplitude interferometry is its comparatively low signal to noise ratio, which decreases as the square root increase of the pixel number. For systems with large aberrations, the signal to noise ratio at the image plane is much worse than that at the Fourier plane used in the amplitude interferometry. Figure 4.44 is a direct comparison between optical modulation transfer functions by using a direct imaging method and by using a holographic method, where v_C is the system cutoff frequency. In the direct imaging method, only a small part of the information at the low frequency region is preserved. Reconstructing the light source spatial distribution is difficult. However, in the holographic method, all information within the cutoff frequency is kept despite aberrations and atmospheric turbulence.

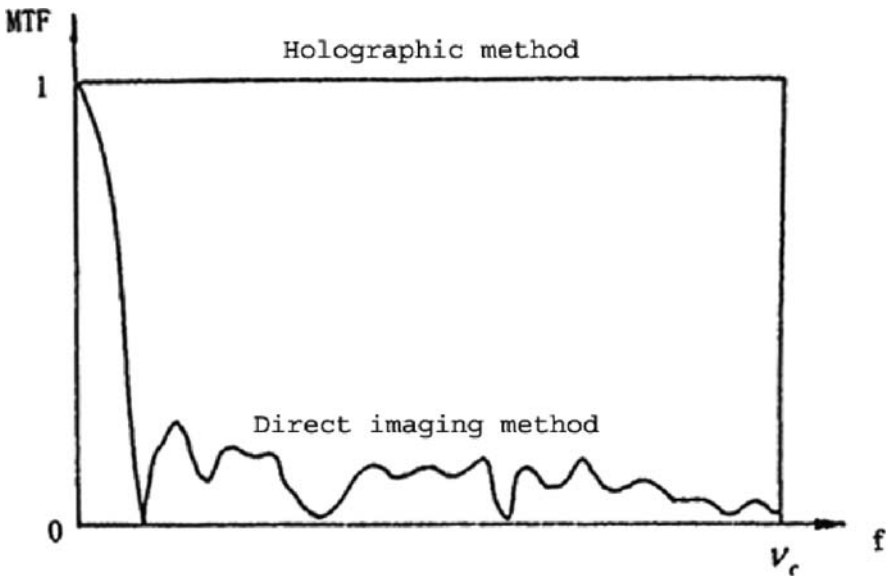


Fig. 4.44. Optical transfer functions using holographic and direct image methods.

References

Angel, R., et al., 2002, The 20/20 telescope: MCAO imaging at the individual and combined foci, in Beyond conventional adaptive optics, eds. Vernet, E., et al., ESO conference proceeding, No. 58.

Armitage, J. D. Jr. and Lohmann, A., 1965, Rotary shearing interferometry, Opt. Acta, 2, 185–192.

Barr, L. D., ed., 1986, Advanced technology optical telescopes III, Proc. SPIE, 628, 466–470, Tuscon.

Beddoes, D. R., et al., 1976, Speckle interferometry on the 2.5 m Isaac Newton telescope, J. Opt. Soc. Am., 66, 1247.

Bely, P. Y., ed., 2003, The design and construction of large optical telescopes, Springer, New York.

Bloemhof, E. E. and Wallace, J. K., 2004, Simple broadband implementation of a phase contrast wavefront sensor for adaptive optics, Opt. Express, 12 (25), 6240–6245.

Bracewell, R. N. and MacPhie, R. H., 1979, Search for nonsolar planets, Icarus, 38, 136–147.

Brown, H. R., 1974, The intensity interferometer, Taylor and Francis Halsted Press, London, 184.

Brown, H. R. and Twiss, R.Q., 1954, A new type of interferometer for use in radio astronomy, Phil. Mag., 45, 663–682.

Chanan, G., et al., 1998, Phasing the mirror segments of the Keck telescopes, the broadband phasing algorithm, Appl. Opt., 37, 140–155.

Chanan, G., et al., 2000, Phasing the mirror segments of the Keck telescopes II, the narrowband phasing algorism, Appl. Opt., 39, 4706–4714.

Cheng, J., 1987, Active optics and adaptive optics, Opt. Instrum. Technol, 4, 1–8.

Classen, J. and Sperling, N., 1981, Telescopes for the record, Sky Telescope, 61, 303.

Costa, J., et al., 2003, Is there need of any modulation in the pyramid wavefront sensor? Proc. SPIE, 4839, p288–298, Hawaii.

Darling, D., 2005, http://www.daviddarling.info/encyclopedia/N/nulling.html

Davis, J., 1997, Observing with optical/infrared long baseline interferometers, in High angular resolution in astrophysics, eds. Largrange, A.-M., et al., Kluwer Academic Publishers, The Netherlands.

De Man, H., Doelman, N. and Krutzen, M., 2003, First results with an adaptive test bench, SPIE Proc., 4839, 121.

Dyck, H. M. and Howell, R. R, 1983, Seeing measurements and Mauna Kea from infrared speckle interferometry, Pub. Asir. Soc. Pac., 95, 786–791.

Esposito, S., et al., 2000, Closed-loop performance of pyramid wavefront sensor, in Laser Weapons Technology, eds. Steiner, T., et al., SPIE Vol. 4034, 434.

Esposito, S., et al., 2003, First light adaptive optics system for large binocular telescope, SPIE Vol. 4839, Adaptive Optics System Technologies II, 164.

Fried, D. L., 1965, Statistics of a geometric representation of wavefront distortion, JOSA, 55, 1427–1435.

Ghigo, M., et al., 2001, Construction of a pyramidal wavefront sensor for adaptive optics compensation, in Beyond conventional adaptive optics, eds. Vernet, E., et al., ESO Conference Proceeding No. 58.

Ghigo, M., et al., 2003, Manufacturing by deep x-ray lithography of pyramid wavefront sensors for astronomical adaptive optics, SPIE Vol. 4839, Adaptive Optics System Technologies II, 259.

Hardy, J. W., et al., 1977, Real-time atmosphere compensation, JOSA, 67, 360–369.
Hardy, J. W., 1998, Adaptive optics for astronomical telescopes, Oxford University Press, Oxford.
Hickson, P. and Burley, G., 1994, Single-image wavefront curvature sensing, SPIE, 2201, 549–554.
Knox, K. T., 1976, Image retrieval from astronomical speckle patterns, JOSA, 66, 1236–1239.
Labeyrie, A., 1970, Attainment of diffraction limited resolution in large telescopes by Fourier analysing speckle patterns in star images, A & A, 6, 85.
Lee, J. H., et al., 2000, Why adaptive secondaries? Publ. Astron. Soc. Pac., 112, 97–107.
Liang, M., 2004, Design note of laser guide star system for TMT, Natioanl Optical Astronomy Observatory.
Liang, Z-P., et al., 2000, Principles of magnetic resonance imaging, a signal processing perspective, IEEE Press, New York.
Liu, C. Y. C. and Lohmann, A. W., 1973, High resolution image formation through the turbulent atmosphere, Opt. Commun, 8 (4), 372.
Lloyd-Hart, M., 2003, Taking the twinkle out of starlight, Spectrum, 40, 22–29.
Malacara, D., 1978, Optical shop testing, John Wiley and Sons, New York.
Marriotti, J. M. and Di Benedetto, G. P., 1984, Pathlength stability of synthetic aperture telescopes in the case of the 25 cm CERGA interferometer, in IAU Colloq. No. 79, eds. Ulrich M.-H. and Kjar K., 247, European Southern Observatory.
Max, C., 2003, Lecture notes on adaptive optics, http://www.ucolick.org/∼max/289C/Lectures/
Nelson, J. E., Mast, T. S., and Faber, S. M., 1985, The design of Keck observatory and Keck telescope, Keck Observatory Report, No 90, the University of California and California Institute of Technology.
Nikolic, B., et al., 2007, Measurement of antenna surface from in- and out-of-focus beam maps using astronomical sources, A&AP, 465, 679.
Noll, R. J., 1976, Zernike polynomials and atmospheric turbulence, JOSA, 66, 207–211.
Ohara, C. M., et al., 2003, PSF monitoring and in-focus wavefront control for NGST, Proc. SPIE, 4850, 416–427.
Pinna, E., et al., 2006, Phase ambiguity solution with the pyramid phasing sensor, SPIE Proc., 6267, 62672Y.
Ragazzoni, R, 1996, Pupil plane wavefront sensing with an oscillating prism, J. Modern Opt., 43, 289–293.
Ragazzoni, R., et al., 1999, Modal tomography for adaptive optics, Astro. Astrophys. 342, L53–L56.
Ragazzoni, R., et al., 2000, Adaptive corrections available for the whole sky, Nature, 403, 54–56.
Redding, D., et al., 2000, Wavefront control for a segmented deployable space telescope, Proc. SPIE, 4013, 546–558.
Restaino, S. R., 2003, On the use of liquid crystals for adaptive optics, in Optical applications of liquid crystals, ed. Vicari L., IOP Publishing Ltd., Bristol and Philadelphia.
Riccardi, A., et al., 2002, The adaptive secondary mirror for the 6.5 m conversion of the Multiple Mirror Telescope, in Beyond conventional adaptive optics, ed. Ragazzoni, R., ESO Conference Proceedings, Venice.
Roddier F., 1988, Curvature sensing and compensation: a new concept in adaptive optics, Appl. Opt., 27, 1223–1225.

Roddier, C. and Roddier, F., 1979, Image with a coherence interferometer in optical astronomy, in Image formation form coherence functions in astronomy, ed. van Schooneveld C., Proceedings of Vol. 76, IAU Colloq. No. 49, D. Reidel Pub. Co., Dordecht.

Roddier, C. and Roddier, F., 1989, Pupil-plane interferometry, in Diffraction-limited imaging with very large telescopes, ed. by Allion, D. M. and Mariotti, J. -M., p211–236.

Roddier F. and Roddier C., 1991, Wavefront reconstruction using iterative Fourier transforms, Appl. Opt., 30, 1325–1327.

Roddier, C. and Roddier, F., 1993, Wavefront reconstruction from defocused images and the test of ground-based optical telescopes, JOSA, A, 10, 2277.

Shepp, L. A., ed., 1982, Computer tomography, in Proceedings of symposia in applied mathematics, Vol. 27, American Mathematical Society, Providence.

Shi, Fang, et al., 2003, Segmented mirror coarse phasing with a dispersed fringe sensor: experiment on NGST's wavefront control testbed, SPIE Proc., 4850, 318–328.

Tallon, M. and Foy, R., 1990, Adaptive telescope with laser probe: isoplanatism and cone effect, Astro. Astrophys., 235, 549–557.

Tatarskii, V. I., 1971, The effect of the turbulent atmosphere on wave propagation, National Technical Information Service, Springfield.

Tyson, R. K., 1997, Principles of adaptive optics, Academic Press, San Diego.

Tyson, R. K., 2000, Introduction to adaptive optics, SPIE Press, Washington.

Ulrich, M. H. and Kjar, K., eds., 1981, Proceedings of ESC conference on: scientific important of high angular resolution at infrared and optical wavelengths, Garching, March.

Walker, C. B., Stahl, H. P. and Lloyd-Hart, M., 2001, Optical phasing sensors, http://optics.nasa.gov/tech_days/techdays_2001/38_MSFC_Optical_Phasing_Sensors.ppt#421,2,Outline

Wilson, R. N., 1982, Image quality consideration in ESC telescope projects, Opt. Acta, 29, 985–992.

Wyngaard, J. C., et al., 1971, Behavior of the refractive-index-structure parameter near the ground, JOSA, 61, 1646.

Yaitskova, N., et al., 2005, Mach-Zehnder interferometer for piston and tip-tilt sensing in segmented telescope: theory and analytical treatment, J. Opt. Soc. Am. A, 22, 1094–1105.

Zhang, Y. (chief editor), 1982, Astronomy in Chinese encyclopedia, Chinese Encyclopedia Publishing Company, Peking.

Chapter 5
Space Telescope Projects and their Development

Space orbit is a subject new to most professionals in astronomy, optics, radio, and other sciences. The orbits definition and the orbit environmental conditions, including upper atmosphere, heat sources, heat transfer, plasma, charged particles, torque produced by gravitational field, and torque produced by the aerodynamic drag, are introduced in this chapter. The altitude sensors and actuators for space telescope position control are also discussed. In the space telescope project part, two important space telescopes, the Hubble Space Telescope and James Webb Space Telescope, are discussed in depth. Emphasis is placed on their mirror design and the mirror material selection. The Space Interferometer Mission and other space projects are also introduced.

5.1 Orbit Environmental Conditions

Earlier and most present astronomical observations are made from the earth's surface. However, the spectrum coverage, angular resolution, and image quality of these observations are seriously limited by the earth's atmosphere. To eliminate the limitations imposed by the atmosphere, it is necessary to send astronomical telescopes high above the ground and into earth's orbit. Telescopes on orbits are referred to as space telescopes. Free from the atmospheric turbulence, space telescopes can achieve truly diffraction limited resolution and the observations can be performed over the entire electromagnetic spectrum. Another possible location for space telescopes is on the moon's surface where the atmosphere and wind turbulence disappear. Some moon-based optical telescopes are planned on the pole of the moon. The far side of the moon is an ideal location for radio telescopes, as it is free from the noisy torrent of radio signals emanating from the earth. Radio telescopes at this location will pick up the extremely faint signals left over from the early universe. The moon-based or planet based telescopes are still at an early stage and they are briefly mentioned in Section 11.5.

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_5,

In this chapter, major space telescope projects, mostly in optical and infrared regions are discussed. Space telescopes at other wavelengths will be discussed in later chapters together with their ground-based counterparts.

5.1.1 Orbit Definition

5.1.1.1 Low Earth Orbit

A low earth orbit (LEO) is that below the medium earth orbit (MEO) or the intermediate circular orbit (ICO). The medium earth orbit is in the region of space around the earth with an altitude between 2,000 and 35,786 km. Most typical LEO satellites are around 350 to 1,400 km above the earth. From Kepler's third law, the period T of a satellite is related to the semi-major axis a of its orbit as:

$$T^2 = \left(\frac{4\pi^2}{\mu}\right) a^3 \tag{5.1}$$

where $\mu = GM = 398{,}600.5\ \mathrm{km^3\ s^{-2}}$ is the earth's gravitational constant, $G = 6.673 \cdot 10^{-11}\ \mathrm{m^3\ kg^{-1}\ s^{-2}}$ the universal gravitational constant, and $M = 5.9742 \cdot 10^{24}\ \mathrm{kg}$ the mass of the earth. Since the LEO is near the earth, the energy required to place satellites in such an orbit is less and the data communication is easy. The period of an LEO is short and the velocity is high. The first major space astronomical telescope, the Hubble Space Telescope (HST), is on this orbit which is reachable by space shuttles so that astronaut-operated maintenance of the telescope is possible. In the LEO, the earth's magnetosphere shields the telescopes from harmful cosmic rays. However, telescopes on this orbit may have limited sky coverage due to the solar radiation and the direct communication to the ground is not continuous.

5.1.1.2 Geosynchronous Orbit

A geosynchronous orbit is a geocentric one that has the same orbital period as the sidereal rotation period of the earth, 23 h, 56 min and 4.090530833 s. It has a semi-major axis of 42,164 km (or an attitude at 35,786 km above mean sea level).

5.1.1.3 Geostationary Orbit

A geostationary orbit is a circular and equatorial geosynchronous one. Objects located on this orbit will maintain the same position relative to the earth's surface. They will cover about 70 degrees in both southern and northern hemisphere of the earth. Their communication with the ground is easy.

5.1.1.4 Polar Orbit

A polar orbit is an orbit which passes above, or nearly above, both poles of the earth. It, therefore, has an inclination angle of, or very close to, 90 degrees to the earth's equator. Since a satellite on this orbit will have a fixed orbital plane perpendicular to the earth's axis of rotation, it will pass over a region of different longitude on each of its revolutions. If the orbital period is an integer multiple of the sidereal day, the satellite will pass the same area of the earth periodically. A polar orbit cannot take advantage of the "free ride" provided by the Earth's rotation, and thus the launch vehicle must provide all of the energy for attaining orbital speed.

5.1.1.5 Sun-Synchronous Orbit

A sun synchronous orbit is a special polar orbit which crosses the equator at the same longitude on the same local solar time. The orbit remains approximately fixed with respect to the sun. Because the earth has more mass around the equator than at the poles, thus a small attractive force is produced towards the earth's equator if a satellite is on this orbit. This pulling force does not change the orbit's angle, but its intersection point with respect to the earth's equator. The force produced is determined by the satellite altitude and inclination angle relative to the earth's equator. By adjusting these two parameters, the change of the intersection point with the equator can be about 1° a day. That is to say, if the satellite orbit plane is perpendicular to the sun direction in the beginning, then it will remain so perpetually. This special polar orbit is a sun-synchronous orbit. Satellites in this orbit are especially good for solar panel operation. In astronomy, some space solar telescopes are placed on this type of orbit.s

5.1.1.6 Lagrangian Point

For astronomers, there are some special space locations, the Lagrangian points (also called libration points). Objects in these positions are stationary relative to a two-body system, such as the sun and earth. The Lagrangian points are not orbits in a normal sense, but stable or metastable space positions named after the 18th century mathematician and astronomer, Joseph Lagrange. Five Lagrangian points exist in any two-body system as shown in Figure 5.1. These five points are named $L1$ to $L5$. Points $L1$, $L2$ and $L3$ were firstly predicted by Leonhard Euler and points $L4$ and $L5$ were predicted by Lagrange.

An object at point $L1$, $L2$, or $L3$ is in a metastable condition. In the plane perpendicular to the line between the sun and earth, they are stable as an object displaced away from the central line in this plane would feel a force pulling it back towards the equilibrium point. Point $L1$ is located between the sun and the earth. The distance to the earth is 1.5 million km, about 0.01 of the distance between the sun and the earth. Point $L2$ is located on the opposite side from the earth with a distance to the earth the same as that of point $L1$. Point $L3$ is located in the opposite side of the sun. Its distance to the sun is about 1.05 of the distance between the sun and the earth. These three points are not stable in the

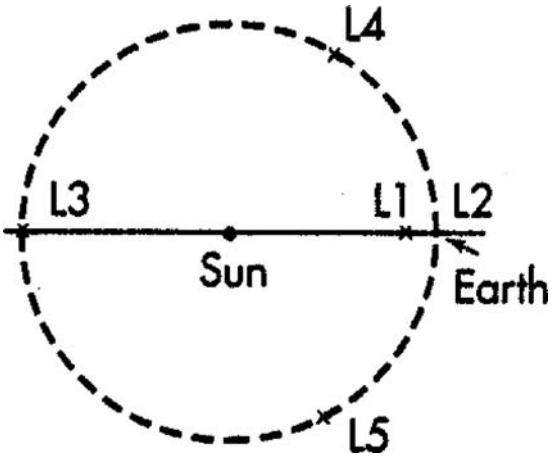


Fig. 5.1 Positions of Lagrangian points in sun-earth system.

central line direction between the sun and the earth. If an object drift away from these points in this direction, the gravitational attraction from the two mass would pull it away even further. In contrast to these three points, $L4$ and $L5$, which are located on the earth's orbit around the sun, are stable, more like a ball at the bottom of a bowl. Small perturbations will not move objects away from these points. These points form equilateral triangles with the earth and the sun. A few astronomical solar telescopes, including the solar heliospheric observatory (SOHO) launched by ESA in 1995 and advanced composition explorer (ACE) by NASA in 1997, were around the $L1$ point. The spacecraft were not exactly at this point, otherwise it would be difficult to track them from the earth as the sun is a source of interference. The James Webb Space Telescope (JWST) under construction will be placed at the $L2$ point.

5.1.2 Orbit Thermal Conditions

In orbital regime, the atmosphere becomes very thin. The main heat exchange within satellite and between the satellite and other celestial objects is through radiation and conduction. There is no convection in this regime, while convection is extremely effective in heat transfer on the earth's surface. This situation together with rapid position change of heat source (the sun) produces a very undesirable space thermal environment. When a satellite is exposed to the solar radiation, thermal energy arrives on the satellite exterior. However, the earth may block part of the solar radiation, when the satellite is inside the penumbra (half shadow) of the earth, or block all of the solar radiation, when the satellite is inside the umbra (full shadow). In the latter situation, the only radiation received is that coming from the earth's surface which is very weak. Rapid thermal change is a primary concern for the on-orbit thermal design.

Generally, there are two types of heat sources for a satellite in space: internal and external heat sources. The internal sources are from electronics and other components. The outer heat sources are solar radiation, the earth's radiation, and solar albedo (the reflected solar radiation from the earth). In the earth's orbit, the solar radiation flux is about 1350 W/m^2. The solar radiation received by a surface in space is (Thornton, 1996):

$$q_s = 1,350 a_s \cos\psi \tag{5.2}$$

where a_s is the absorption coefficient of the surface and ψ the angle between the solar flux vector and the surface normal. The radiation received from the earth (Figure 5.2) is:

$$q_e = \sigma T_e^4 a_e F \tag{5.3}$$

where σ is the Boltzmann constant, $T_e = 289\ K$ the black body temperature of the earth, a_s the absorption coefficient of the surface, and F a view factor of the surface relative to the earth. When only one side of the surface receives radiation from the earth, the view factor F is (Figure 5.3):

$$F = \cos\lambda / H^2 \tag{5.4}$$

where λ is the angle between the line, which connects the surface center and earth's center, and the surface normal, $H = r\ /\ R$, r the distance between the surface and the earth's center, and R the radius of the earth. The condition of this equation is $\lambda + \Phi_m \le \pi/2$, where Φ_m is the half angle extended from the surface center to the earth sphere. If both sides of the surface receive radiation, then the view factor will be:

$$F = \frac{2}{\pi}\left[\frac{\pi}{4} - \frac{\sin^{-1}}{2}\left[\frac{(H^2-1)^{1/2}}{H\sin\lambda}\right] + \frac{1}{2H^2}\{\cos\lambda\cos^{-1}\left[-(H^2-1)^{1/2}\cos\lambda\right] - (H^2-1)^{1/2}\left[1 - H^2\cos^2\lambda\right]^{1/2}\}\right] \tag{5.5}$$

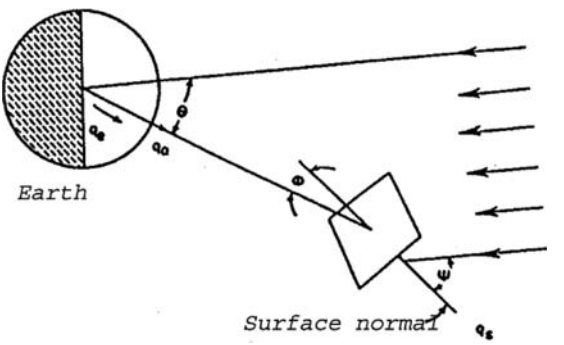


Fig. 5.2 Radiation received from the sun and the earth in the space (Thornton, 1996).

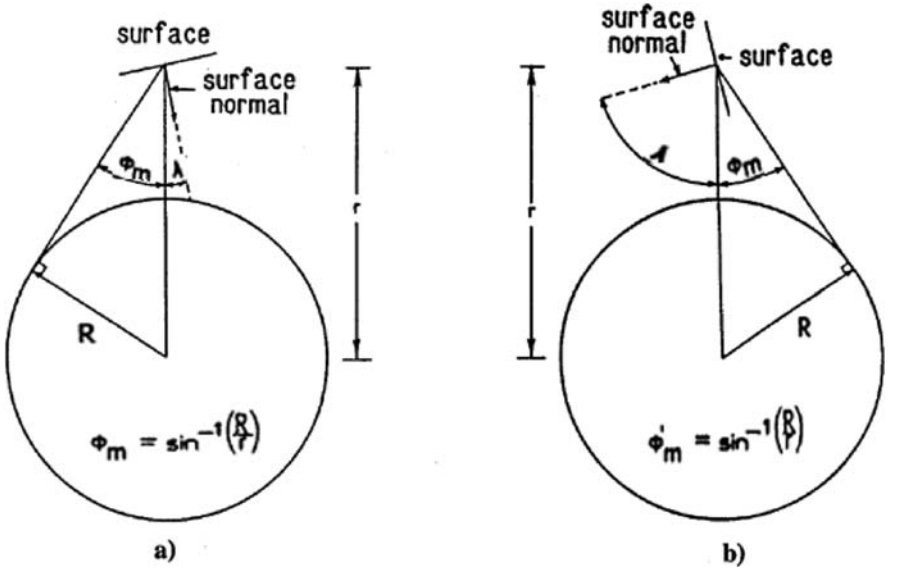


Fig. 5.3 View factor calculation for radiation received from the earth (Thornton, 1996)

Solar albedo is the earth's reflection of solar radiation. The albedo is proportional to the emissivity of the earth, AF, which is influenced by the season and geography. The approximate value is $AF = 0.36$. Solar albedo applies only on the day side of the earth. The heat received from solar albedo is:

$$q_a = 1,350(AF)Fa_s \cos\theta \qquad (W/m^2) \tag{5.6}$$

where a_s is the surface absorption coefficient and θ the angle between the line, connecting the surface center and the earth's center, and the surface normal. The total heat received by the surface is then:

$$q = q_s + q_e + q_a \tag{5.7}$$

In Table 5.1, typical radiation intensities in the low earth and geosynchronous orbits are listed. In Table 5.2, the orbit periods and the time for a satellite to pass through the earth's shadow are listed for these two orbits. Since the transit time through the penumbra is negligible compared with through the umbra, the penumbra can be disregarded. For polar orbit, the orbit surface has an angle τ with the earth's shadow and the time to pass the earth's shadow is:

$$t_s = \frac{\pi \cdot r}{V_s}\left\{1 - \frac{2}{\pi}\left[\sin^{-1}\left(\frac{\sin(\cos^{-1} R/r)}{\sin\tau}\right)\right]\right\} \tag{5.8}$$

Table 5.1 Typical radiation intensity for two types of orbits (Thornton, 1996)

Orbit type	Radiation intensity (W/m^2)		
	Solar radiation	Earth's radiation	Solar albedo
Low earth	1,350	310	380
Geosyn.	1,350	8	10

Table 5.2 Orbit periods and transit time passing the earth's shadow (Thornton, 1996)

Orbit type	Radius (km)	Period (s)	Transit time (s)	
			Umbra	Penumbra
Low earth	6,878*	5,675	21,443	8
Geosyn.	42,253*	86,044	4,167	128

*Data quoted as it was in the reference book.

where V_s is the velocity of the satellite, $V_s = (gR^2 \,/\, r)^{1/2}$, R the radius of the earth, and r the orbit radius.

Figure 5.4 shows the radiation distribution on a large 43 m space truss structure. Figure 5.4(a) is for the structure on geosynchronous orbit and Figure 5.4 (b) is on low earth orbit. From the figure, it can be found that the

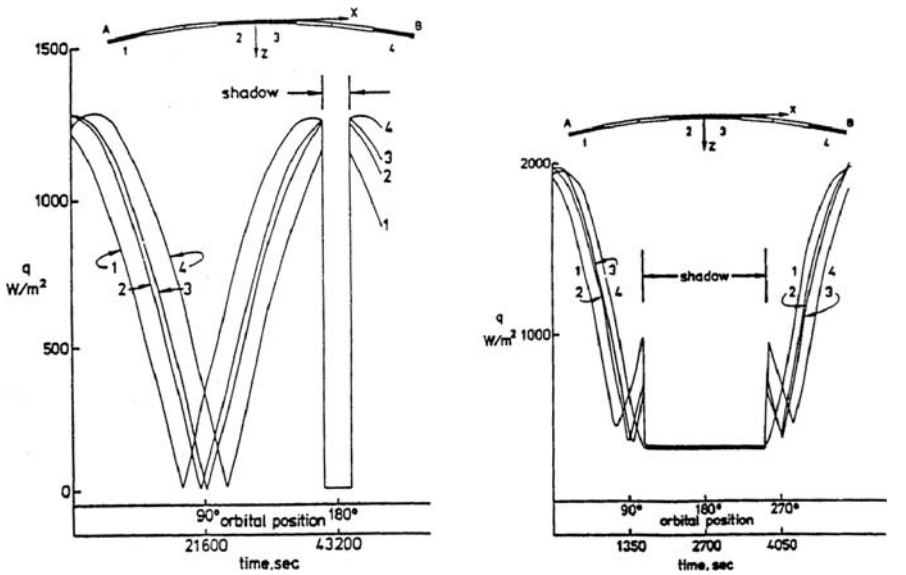


Fig. 5.4. Radiation intensity change of a 43 m space truss structure (AB) (a) in geosynchronous orbit and (b) in a low earth orbit (Thornton, 1996)

radiation intensity variation received on orbit is extremely high. The rapid variation can produce structural deformation, stress, and vibration. Therefore, insulation and the use of low thermal expansion materials are very important for space telescopes.

5.1.3 Other Orbit Conditions

5.1.3.1 Effects of the Upper Atmosphere

The upper atmosphere, although very thin, affects satellites by producing aerodynamic drag, lift, and heat. Highly reactive elements of the upper atmosphere, such as atomic oxygen, will also cause corrosion of materials. The acceleration caused by the atmospheric drag is:

$$a_D = -\frac{1}{2}\rho(c_D A/m)V^2 \tag{5.9}$$

where ρ is the density of the atmosphere, c_D the drag coefficient (for most satellites, c_D is about 2.2), A the cross section area, m the mass, and V the relative velocity between the satellite and the atmosphere. The denser the air is, the larger the drag. Therefore, satellites on LEO with their radius at perigee (the closest distance to the earth's center) being smaller than 120 km will have a very short lifetime, while the lifetime of a satellite on orbits higher than 600 km from the earth center will be more than 10 years.

5.1.3.2 Plasmas and Spacecraft Charging

The earth's magnetic field is similar to a bar magnet, roughly dipolar in shape. Local magnetic field intensity of the earth is:

$$B(R,\lambda) = (1+\sin^2\lambda)^{1/2}B_0/R^3 \tag{5.10}$$

where λ is the magnetic latitude, R the radial distance measured in earth radii, and $B_0 = 0.32 gauss$ the field flux intensity at the earth equator surface. The interaction between the solar wind and the earth's magnetic field causes the magnetic field on the night side of the earth to stretch into a very elongated structure, the magnetotail (Figure 5.5). In the middle of the magnetotail, there is a thin plasma sheet extending over 1,000 earth radii parallel to the solar wind flow. Some of the solar wind kinetic energy is converted to magnetic energy stored in the magnetotail after the interaction with the earth's magnetic field. However, some of the energy will be dissipated as magnetic substorms. These substorms produce energized hot plasmas (5 to 20 KeV) which extend into geosynchronous orbits, charging the surface of any spacecraft within it to high negative voltages.

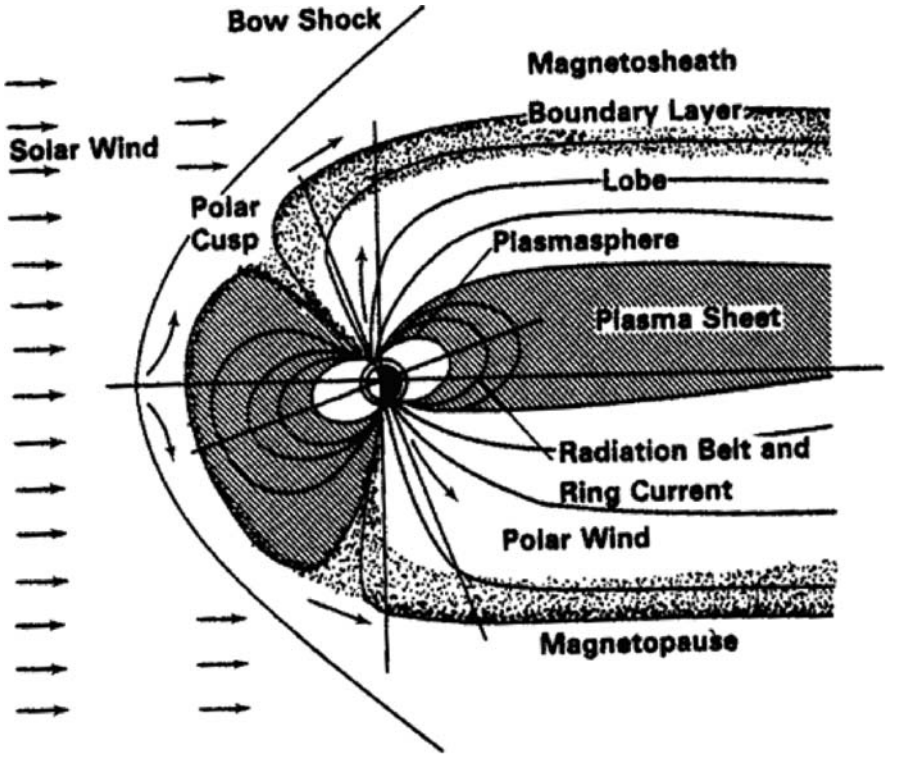


Fig. 5.5. A cross section of the earth magnetosphere (Wertz and Larson, 1991)

5.1.3.3 Trapped High Energy Particles on Space

The Van Allen belt is a small but very intense torus region of energetic charged particles, mainly protons and electrons (>30 KeV), captured by the magnetic field of the earth. These particles are trapped within 6,000 km or so of the earth's surface. As illuminated in Figure 5.6, the energetic electrons populate a pair of regions centered on $R_E \sim 1.3$ and $R_E \sim 5$, where R_E is the distance measured in the Earth's radius. It also provides flux of protons around the Earth. In low Earth orbit, energetic protons in the inner radiation belt contribute most to the total radiation dose. Solar cells, integrated circuits, and sensors can be damaged by this radiation. Miniaturization and digitization of electronics and logic circuits have made satellites more vulnerable to the radiation as incoming ions may be as large as the circuit's charge. Electronics on satellites must be hardened against radiation to operate reliably. Radiation hardening is a method of designing and testing electronic components and systems to make them resistant to damage or malfunctions caused by high-energy subatomic particles. Two hardening methods are: physical and logical ones. Using physical methods, hardened chips are often manufactured on insulating substrates instead of the usual

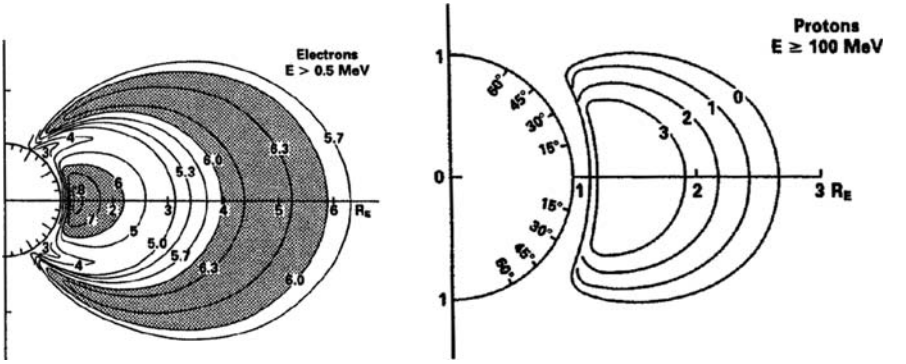


Fig. 5.6. (a) High energy electrons and (b) protons around the earth (Wertz and Larson, 1991)

semiconductor wafers. Silicon oxide (SOI) and sapphire (SOS) space-grade chips can survive doses many orders of magnitude greater than those of the usual semiconductor ones. Shielding is another physical way to reduce exposure of the bare device. Figure 5.7 shows how various shielding thickness (g/cm^2) of aluminum will affect the radiation doses in low altitude polar orbits. The unit 1 rad is the amount of radiation which deposits 100 *ergs* (6.25×10^7 MeV) per gram of target material (100 mils of aluminum sheet is equivalent to 0.686 g/cm^2). The total radiation dose consists of three components: proton, electron, and X-ray bremsstrahlung doses. The radiation dose strongly depends on the altitude. Below the altitude of 1,000 km, the radiation dose is of mainly energetic protons. The dose increases as approximately the 5th power of the altitude. At the

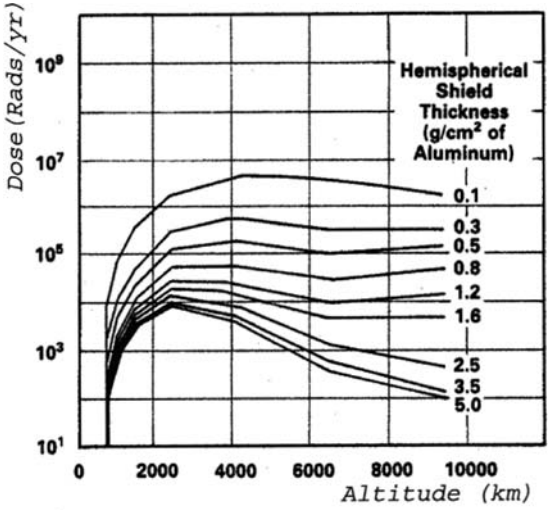


Fig. 5.7 Radiation dose as a function of altitude for low altitude polar orbits (Wertz and Larson, 1991)

synchronous orbit altitude the proton dose greater than 5 MeV is negligible and the bremsstrahlung dose is dominant for a shielding thickness greater than 1 cm. Low energy ions can affect space systems differently than the penetrating radiation. They deposit energy in the spacecraft skin and cause a temperature rise, significantly increasing the infrared background. They also degrade the effectiveness of painting and the protective glasses.

5.1.3.4 Solar Particle Events and Cosmic Rays

Solar particle events occur in association with solar flares, which are violent explosions in the solar atmosphere. Solar flares release a rare energetic particle flux which lasts from several hours to several days. However, they degrade the solar panel arrays and increase the background noise of many electro-optical sensors of a satellite.

Galactic cosmic rays are particles from outside the solar system. Cosmic rays pose a serious hazard because a single particle can cause malfunction of an electronic component, such as a microprocessor. This is named the single-event phenomenon (SEP). The SEP cannot be reliably predicted. Galactic cosmic rays can also generate background noise in sensors, detectors, and other electric components.

5.1.3.5 Gravity Gradient and Aerodynamic Torques

Torque applied on space telescopes will produce a rotation of the spacecraft. The same as how a magnet in a magnetic field is subject to a torque if the magnet is not aligned with the field line, a nonsymmetrical mass inside a gravity field will be subject to a torque if the mass main axis is not aligned with the field line. This torque is the gravity gradient torque. The gravity field lines of the Earth are straight and are usually perpendicular to orbits. The gravity gradient torque generated in the x direction is:

$$\tau_g = \frac{3\mu}{R^3} |I_z - I_y| \sin\theta \tag{5.11}$$

where μ is the earth's gravitational constant, R the radius from the spacecraft to the earth center, I_z the smallest moment of inertia of the spacecraft about the longest dimension axis, I_y the moment of inertia of the spacecraft about the other axis, and θ the angle between the z axis and the nadir vector in the x-z plane. The torque in the y direction can also be calculated.

Aerodynamic torque in a spacecraft is caused by the air drag in the low earth orbit and an offset between the mass center and the aerodynamic pressure center. The torque is equal to:

$$\tau_a = -\frac{1}{2} C_D \rho V^2 \int \vec{r} \times (\vec{N} \cdot \vec{V}) \vec{V} dA \tag{5.12}$$

where C_D is the drag coefficient, ρ the atmospheric density, V the velocity, $\vec{r}$ the vector from the mass center to element area dA, $\vec{N}$ the outwards normal of element area, and $\vec{V}$ the velocity vector.

5.1.3.6 Launch Conditions

Spacecraft launch places limitations not only on payload size and weight, but also in terms of thermal shock, pressure change, acceleration, and vibration. During launch, the temperature inside the rocket fairing (the cover of the payload) can reach 200°C for some launch vehicles. At the same time, pressure differences inside and outside the fairing are developed as the ambient atmospheric pressure continuously drops with altitude. Inside the fairing, higher pressure air is retained. The largest challenge during launch is the acceleration and random vibration. There are two acceleration components: axial and lateral. The lateral acceleration and random vibration come from the wind shear and orbit correction and the axial one is from acceleration. These random variables are expressed in the form of power spectral density as shown in Figure 5.8 for some US launch vehicles.

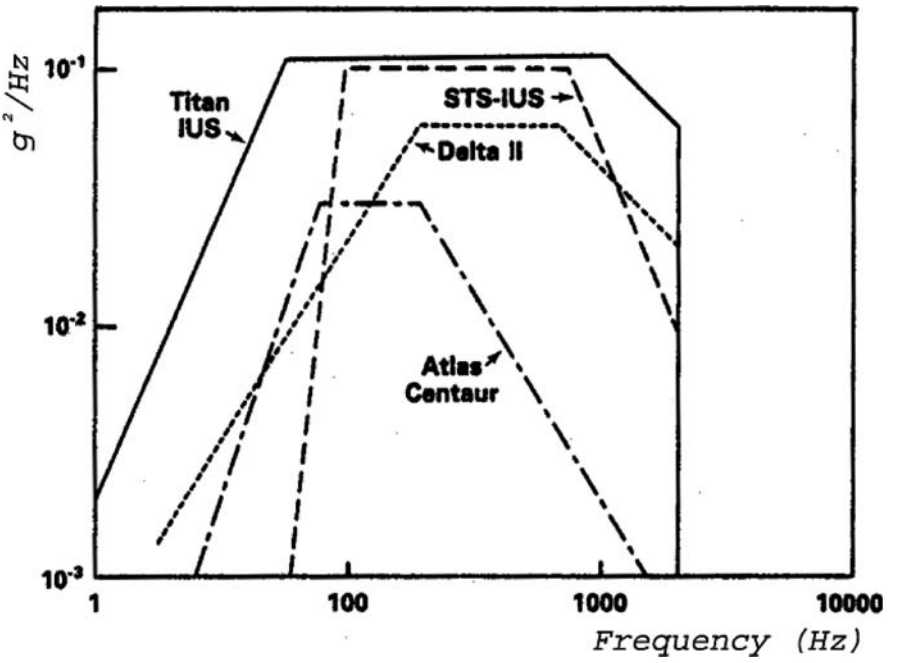


Fig. 5.8 Vibration environment for different launching systems (Wertz and Larson, 1991)

5.2 Attitude Control of Space Telescopes

The orientation of a spacecraft with a reference system is named the attitude. The space telescope attitude is measured by attitude sensors such as: gyroscopes, horizon indicators, sun sensors, and star trackers. Information from these sensors is fed back to the altitude actuators for pointing the spacecraft in the required orientation.

5.2.1 Attitude Sensors

5.2.1.1 Gyroscopes

Gyroscopes, including positional, rate, optical fiber, gas-bearing, and electrostatically supported ones, are very accurate attitude measuring devices. The principle behind a positional gyroscope is the conservation of angular moment within a fast rotating system. If there is no torque applied on a fast rotating system, the orientation of the system is maintained regardless of the motion of the platform to which it is attached. The principle for a rate gyroscope is Newton's second law. A change in moment is compensated by an addition of another equal moment in an opposite direction. The HST telescope used to have rate gyroscopes for its altitude control. When a torque is applied on the system, a moment change takes place in the system. This torque is measured and compensated by a balanced spring system. The measured torque is the input of a feedback system for the pointing correction. The HST used to have four gyroscopes. However, it can now be operated with only two gyroscopes together with star trackers.

An optical fiber gyroscope can be either of interference or of resonance types. Both types of gyroscope are based on the Sagnac effect. According to Sagnac theory, if two pulses of light are sent in opposite directions around a stationary circular loop, they will travel the same distance at the same speed. However, when the loop rotates about its axis, the pulse traveling in the same direction as the loop travels a slightly longer distance than the pulse travels in the opposite direction. As a result, the counter-rotating pulse arrives at the "end" point slightly earlier than the co-rotating pulse. If these two light beams interfere, then the rate of loop rotation can be calculated. This is an interference gyroscope. If these two beams remain inside the loop as in a Fabry–Perot interferometer, resonance is produced. Then, it is a resonant gyroscope. The resonant optical fiber gyroscope has a higher rate of resolution and is often used in space telescope altitude control.

A typical resonance optical fiber gyroscope consists of three fiber loops. These loops are coupled to each other with the second loop being a closed one. Laser light from a diode is sent into the first loop and enters the second loop through an optical coupler. The reflected light from the optical coupler together with the light transmitted from the second loop reaches a photo-sensor. The

second loop is a major one. Both couplers for this loop are asymmetrical with 90 to 99% of the light remaining inside the second loop. The transmitted and reflected light at the first coupler will produce resonance of a given wavelength inside the second loop. The transmitted light of this wavelength at the other coupler of the second loop enters a transmission port of the third loop. When the rate of the second loop rotation changes, the optical path lengths in both directions change and resonant wavelength and frequency vary. The rate of the loop rotation is related to the resonant frequency.

A gas bearing or electrostatically supported gyroscope is a very precise instrument (Merhav, 1996). Normally, mechanical gyroscopes have two sets of bearings: rotor spin and gimbal suspension. These bearings are noisy and easy to wear out, producing rate drifting in measurement. In a gas-bearing gyroscope, a fixed, high precision ball in the center supports a spinning rotor through a 4 μm gap hydrogen gas bearing. This one bearing functions as two bearings in other gyroscopes. The stator coils around the rotor generate a rotational magnetic field, so that the rotor rotates as an induction motor with a speed up to 9,000 rpm. The rotor can rotate about two other axes which are perpendicular to the spinning one. The gyroscope house has sensors to sense these two tilt components of the rotor for the positioning measurement.

An electrostatically supported gyroscope is the most accurate instrument. Its spinning rotor is a ball isolated from its surroundings, like a free spinning star encapsulated in an evacuated chamber and running under its own angular momentum. This hollow beryllium ball has a 0.02 mm gap from six levitation electrodes evenly distributed around the ball. These electrodes have a voltage of 150 V. In the ball rotational plane, four stator coils produce a rotating magnetic field. The induced eddy current on the ball surface allows it to spin up to a speed of 150,000 rpm. Once the speed is reached, the power is disconnected. The angular shifts in the directions perpendicular to the spinning plane are picked up by quantum measuring devices from magnetic fields produced by a thin superconductor layer on the ball surface in these directions.

5.2.1.2 Star Tracker, Horizon Indicator and Sun Sensors

A star tracker is another attitude sensor for spacecraft. The star tracker provides highly accurate, very stable pointing relative to a celestial sphere. The star tracker generally has a large field of view, of about 8 degrees, so that many stars are inside the field of view. The location of stars is compared with star maps stored in the computer to determine the star tracker's direction.

A horizon indicator is an infrared device which uses the contrast between the cold of deep space and the heat of the earth's horizon. They provide the direction of the earth's horizon. Two different horizon indicators are used: a scanning one with a rotating head and static one with a fixed head.

A sun sensor is a pinhole camera to form an image of the sun on a position sensitive detector or CCD surface. Therefore, the location of the sun is detected.

5.2.2 Attitude Actuators

The attitude actuators include thruster (hydrazine), reaction wheel, and magnetic torquer. A thruster is a mass rejection device mainly used for the orbit correction. It releases mass and, therefore, has a limited lifetime. It is not often used for space telescope pointing corrections.

A reaction wheel is a type of flywheel used by spacecraft to change their angular momentum. Reaction wheels are usually located at a corner of the rotational plane. They are equipped with magnets or coils to act as a brushless DC motor. The spacecraft rotation is realized through the acceleration or deceleration of these flywheels. A typical rotational speed of these wheels is 3,600 rpm. When the flywheel is fixed at this speed and the torque is balanced, the spacecraft has no angular change. If it is accelerated, a moment in the opposite direction will be produced on the spacecraft. The motion stops when the flywheel returns to its original speed. Three sets of the reaction wheels in perpendicular directions are needed for a spacecraft to rotate in all three axes.

Magnetic torquers are generally two sets of coils of uniform wire along perpendicular directions. When a voltage is applied on a coil winding, a magnetic dipole is produced. This dipole interacts with the earth's magnetic field and a torque is produced. The direction of the torque produced is determined by the cross product of the magnetic field and the magnetic dipole of the coil.

5.3 Space Telescope Projects

5.3.1 Hubble Space Telescope

The first major space telescope project was the Hubble Space Telescope (HST). The project, also named the Large Space Telescope (LST) and Space Telescope (ST), was proposed by the National Aeronautics and Space Administration (NASA) in the 1960s. In the beginning, the telescope size was not fixed. An early test tube structure made by Boeing Inc. had a 3 m aperture size. The 2.4 m aperture size was finally selected in 1975 as a result of cost analysis. At that time, the budget for a 3 m telescope was $334 M, for a 2.4 m one $273 M, and for a 1.8 m one $259 M. The space telescope plan was formally approved by the US Congress in 1977 with a total budget of $425 M to $475 M. The European Space Agency (ESA) also joined the project and shared 15% of its cost.

Along with the project development, the cost went up rapidly. By 1985, the HST budget reached $1.175B. The cost reached $1.6B when the telescope was ready for launch. The HST launch was delayed because of the space shuttle

disaster in 1986. The HST was put into a low earth orbit in 1990 with a total cost of $2.35B. The HST operational cost is about $230 M per year (Petersen and Brandt, 1995).

The HST telescope optics was made by the Perkin-Elmer Company. The company won the contract with a very low bid of $95 M. However, when the optical components finished in 1984, the company submitted a claim of $300 M, a number far above the bid price. The company also claimed that the components reached a much better accuracy than the specification required.

However, it was just this optics which gave the telescope a very poor image quality after launch. The HST optics suffered from serious spherical aberrations. During the manufacture of the HST primary mirror, a combined primary and secondary mirror test was not performed. The primary mirror was tested simply using nulling lens optics. Unfortunately, the primary null-corrector was misplaced by a very small distance of 1.3 mm so that the outer edge of the primary mirror was over-polished. The solar panels of the HST also had thermal caused vibration problems.

Immediately after launching of the HST, a special panel was formed to investigate this spherical aberration problem. In 1993 special correction optics was added to the HST during the first shuttle service mission for compensating spherical aberrations. In this mission, new solar panels were used in the telescope. To date, three additional service missions in 1997, in 1999, and in 2002 were performed for the HST. During these missions, three guiding gyros, new focal instruments, and a cooler system were installed.

The HST is a 2.4 m space telescope with a honeycomb fused silica primary mirror of 829 kg. The mirror was formed by fusing together five mirror components, the upper and bottom plates, the inner and outer rings, and the honeycomb grid (Figure 2.18). It is light in weight and high in stiffness. During mirror polishing, the mirror was supported on an air bag support system. The air bags provided uniform floating support forces so that the force condition was similar to that in orbit without gravity. The mirror was coated with aluminum and magnesium fluoride layers. Magnesium fluoride is inert to atomic oxygen which exists in the upper atmosphere.

Since there is no gravity loading in the orbit, the positioning of the primary mirror is through three outer edge points. These points also acted as the mirror's axial support. On the back of the HST primary mirror, 24 piezoelectric actuators were attached on a spring-loaded mirror cell. These actuators were intended for correcting the primary mirror surface shape. However, their forces of about 10 lbs each are too small to be of any practical use for this stiff mirror. Nevertheless, it was the first attempt at applying active optics to a major optical telescope.

The HST primary mirror cell is made of titanium, a material with low density, high strength, and low thermal expansion. The mirror cell design considered the mirror support under gravity on earth, under acceleration during launch, and without gravity in orbit. The mirror cell has a double deck structure. The upper layer was soft to insure the mirror support under gravity. The lower layer was strong to insure the mirror support during launch.

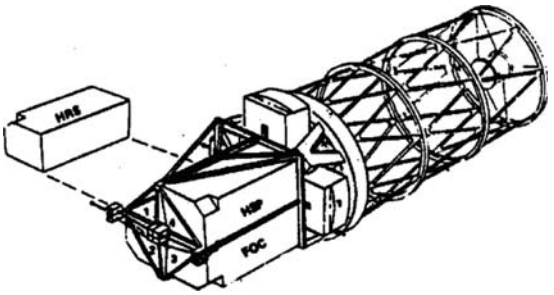


Fig. 5.9 The structure of the Hubble Space Telescope (NASA)

The HST tube was made of carbon fiber reinforced plastic (CFRP) as shown in Figure 5.9. The CFRP materials are discussed in Section 8.3. The main characteristics of the CFRP are very low coefficient of thermal expansion (CTE) and very high strength in the fiber direction. The CTE in the fiber direction is near zero or negative. However, in directions perpendicular to the fiber, the CTE is large and the stiffness is low. For this reason, ring-shaped CFRP parts may have large CTEs in the axial direction. In the HST tube design, negative CTE CFRP beam members were selected so that the whole structure had a zero axial CTE. The tube is a three-layer structure. The CTE for the first layer of trusses is $-0.006 \sim 0.04$ ppm, for the second layer $-0.01 \sim 0.056$ ppm, and for the third layer $-0.02 \sim 0.14$ ppm. The CTEs of the ring structures were large and positive so that the axial CTE of the whole tube is zero. The resonant frequency of the HST tube is 18 Hz. All the optical benches of the HST instruments are also made of CFRP thin plates.

The pointing of the HST used six rate gyroscopes and two precision star trackers. With its very accurate interference star guiding equipment (Section 3.3.4), the pointing accuracy of the HST is 0.007 arcsec matching its high diffraction limited angular resolution. The HST has a long cylindrical protective cover outside its tube for protecting its high-sensitive detectors away from the bright sun. Ring baffles inside the tube are also used. The whole telescope is insulated by highly efficient multi-layer thin aluminum films. The telescope has two solar panels of 2,400 W to provide the power needed. It has a number of focal instruments including the Wide Field/Planetary Camera (WF/PC), Corrective Optics Space Telescope Axial Replacement (COSTAR), Faint Object Camera (FOC), Faint Object Spectrograph (FOS), and High Resolution Spectrograph (HRS). The COSTAR is the optical device specially added for correcting its spherical aberrations of the primary mirror. The most recent HST shuttle service mission brought more focal instruments in infrared and ultra-violet regions.

The HST telescope weighs 11,000 kg with a length of 13.1 m and has an outer diameter of 4.27 m. It fits very well inside a space shuttle, which has a loading limit of 29,250 kg and a load dimension limit of 4.5 m$\times$ 18 m. The HST costs

much more than a ground-based telescope. However, its angular resolution (∼0.03 arcsec at 400 nm wavelength through dither technique) and sensitivity reached are about 100 times that of a ground-based telescope without adaptive optics. The HST provides very sharp images and has led to many new discoveries in astronomy. Because of this success, an even larger space telescope, the James Webb Space Telescope (JWST), is now under construction.

Another important space optical telescope is the Hipparcos Astrometry Satellite which is a small astrometric optical telescope. The satellite was launched on August 1989 and ended its observations on August 1993. The telescope had a split primary mirror so that it could observe two different sky areas at the same exposure. The real distance between two stars in the sky is the sum of the distance in the image plane and the angular distance between two split parts of the primary mirror. The star positions are found through a special metrology method, which requires a large number of overlapped star images. These star images form a huge net and the star positions can be accurately determined through mathematical iteration.

The Hipparcos was planned on a geo-synchronous orbit. However, it was accidentally launched to a highly elliptical orbit with a perigee of only 500 km. This made the telescope less efficient since it was affected by atmosphere drag, thermal radiation from the earth, and radiation from the Van Allen belt. It also suffered from reduced solar power supply and discontinuous communication with the ground base. Even so, the Hipparcos made very accurate measurement of about 120,000 stars. The position accuracy of these measurements is about 1 milliarcsec, the accuracy of star brightness is about 0.0015 magnitudes.

The most recently developed space telescope by NASA is a 0.95 m Schmidt one named the Kepler Mission to be launched in 2009. Its spherical main mirror is a 1.4 m diameter honeycomb one. One main objective of the mission is detection of terrestrial planets.

5.3.2 James Webb Space Telescope

The James Webb Space Telescope is a 6.5 m segmented mirror space mission operated by Goddard Space Flight Center (Figure 5.10). The planned launch date is no earlier than June 2013. The budget for the JWST was $1.3B before the year 2000 and is now about $5B. As the next major space telescope after the HST, it was named the Next Generation Space Telescope (NGST).

The JWST will work in the infrared region of 0.6∼0.27 μm and thus is an infrared instrument. It will be located at the *L2* Lagrangian point of the Earth-Sun system. The distance from this point to the earth is 1.5×10^6 km. The orbit provides sufficient solar power and is easy for communication between ground and the telescope. The telescope will be shielded from the strong sun, earth, and moon radiations by a number of insulation panels. The working temperature is very low at about 40∼60 K. Since this orbit has a constant distance to the sun, its thermal environment is stable year around.

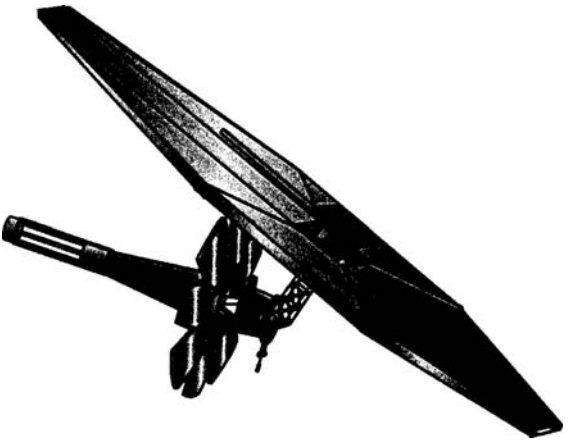


Fig. 5.10. The design of the James Webb Space Telescope (Stockman, 1997)

The HST telescope uses a traditional passive stiff mirror with a weight of 829 kg. The unit area mirror weight is 183.25 kg/m^2. However, the JWST telescope will use a new generation, thin, adaptive, segmented mirror with an expected unit area weight of about 10 to 15 kg/m^2. This mirror unit area weight includes the weight of the actuators and supporting structures attached to the mirror segments. That is to say that the total weight of this 6.5 m mirror will be 10 times less than that of the 2.4 m HST primary mirror.

Limitation for the JWST mirror is also imposed by the shroud size of the launch vehicle. Study shows that at least for this decade a monolithic mirror of 4 m size will fit within a 5 m payload fairing without difficulties. However, any mirrors larger have to be deployed through mirror segments. Even a 6 m fairing size ($\sim$5 m mirror diameter) requires a substantial investment in rocket technology. Fortunately, study of the JWST system shows a very small effect on the point spread functions between using a monolithic mirror and a segmented mirror as the JWST is not a coronagraph instrument where a monolithic primary mirror has a definite scientific advantage.

Both CFRP and beryllium are candidate materials for the JWST primary mirror. The CFRP material has the lowest contraction ratio when the temperature drops from room temperature to the telescope's operational temperature of 40–60 K. Figure 5.11 shows the percentage contraction of different materials in the range from 300 to 0 K. However, it is still difficult for the CFRP mirror segments to reach the tight surface requirements of the JWST after more than two decades of development.

In 1987, a 4.5 m seven-segment space-based antenna was produced in the US. The accuracy of the mirror surface limited its operation wavelength to only 5 mm. The positional accuracy between segments achieved is about 16 μm. In recent years, the replica technique for the CFRP mirrors has been improved and some CFRP mirrors are used at optical wavelengths. However, the largest

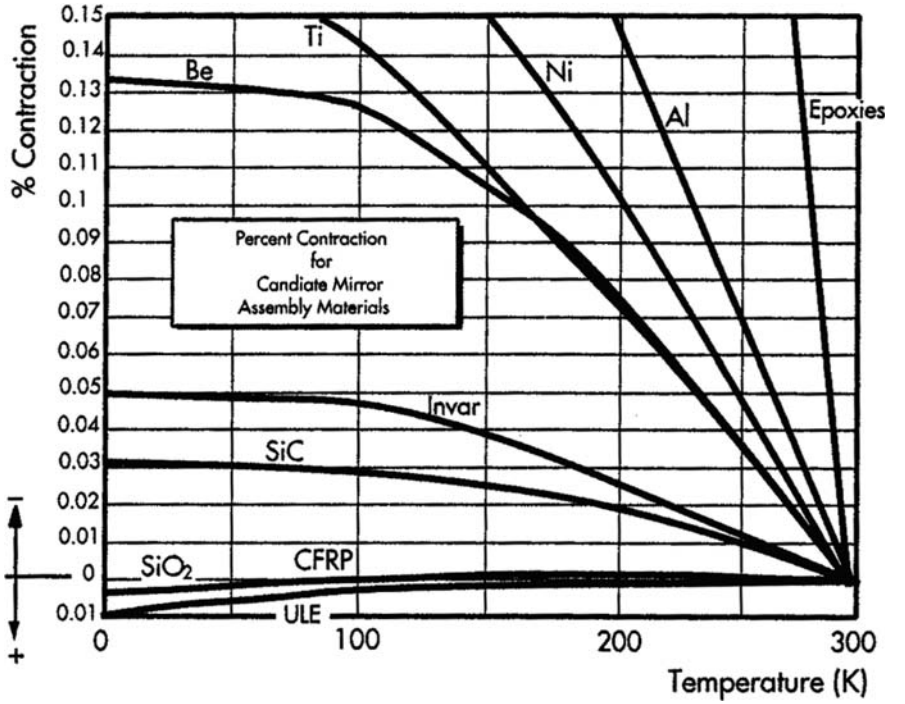


Fig. 5.11. Percentage contractions for different materials from 300 to 0 K (Stockman, 1997)

CFRP optical mirror produced to date is only 1.5 m in diameter. Even so, the long-term stability of the CFRP mirrors is still untested.

An alternative mirror concept is the combination of a thin glass surface with a CFRP backup structure. This is the University of Arizona's approach. In a design study, a 2 m NGST Mirror System Demonstrator (NMSD) test mirror was produced. Details of this test mirror are shown in Figure 5.12. It consists of a very thin glass surface supported by a strong CFRP box structure. The surface is controlled by a number of actuators. The very thin glass mirror has a thickness of only 2 mm. Under the gravity, the mirror surface shows deformations where the actuators push the mirror back. However, by subtracting the selfweight deformation of the glass mirror, the mirror surface becomes much smoother. Therefore, the mirror can be used for space application. But the study also shows that the complexity increases with larger aperture size and lower areal mirror density. In a 2 m test mirror, 50 actuators are involved and for a 6.5 m mirror, 10 times more actuators will be required.

Beryllium (Be) is another mirror candidate material. As a hazardous material, beryllium has ceramic-like property that makes it stiffer than glass even though it is lighter than aluminum. Its density is 1,850 kg/m^3. This material has

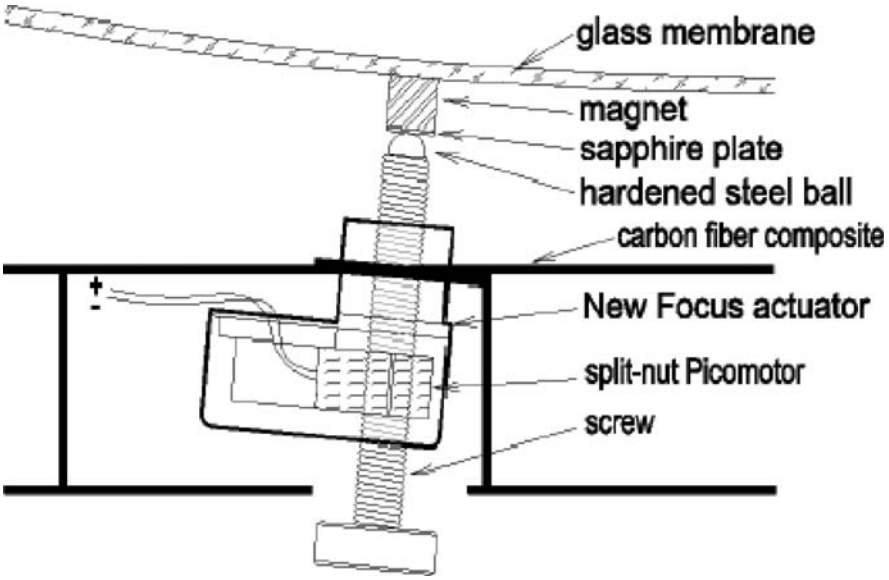


Fig. 5.12. One of the primary mirror designs for JWST (Burge et al., 1998).

been used successfully for the primary mirror of the Space Infrared Telescope Facility (SIRTF). The SIRTF mirror is a 0.85 m diameter very thin meniscus.

The SIRTF beryllium mirror had also been thoroughly tested at both liquid nitrogen and helium temperatures. The test shows that the surface pattern change of the mirror from ambient to cryogenic temperature is repeatable. Therefore, a cryogenic hit map of the mirror surface deformation can be obtained through a cooling test. The correction of the surface shape under cryogenic conditions can be made at room temperature. The test also shows that the beryllium mirror under the temperature of 100 K performs pretty constantly as the CTE of beryllium changes very little over this temperature range (Figure 5.13). This allows most of the surface information of beryllium mirrors to be obtained in a lower-cost nitrogen temperature test ($\sim$90 K), instead of a costly helium temperature test ($\sim$4 K). At present, beryllium has been selected as the mirror material for the JWST project.

Beryllium mirrors cannot be cast since it loses its strength during the melting process. However, it produces grain crystal structure during solidification from powder beryllium. To achieve the highest strength, beryllium must have a fine-grained structure through a powder metallurgy process with inert gas. The fine hexagonal crystal of beryllium is anisotropic in property. To form the mirror blank, the powder is heated to about 900°C while compressing by vacuum or pressure at 1,000 atmospheres. These processes are called vacuum hot pressing or hot isostatic pressing. Through variations in particle size, distribution, BeO content, and temperature, it is possible to produce a variety of beryllium grades

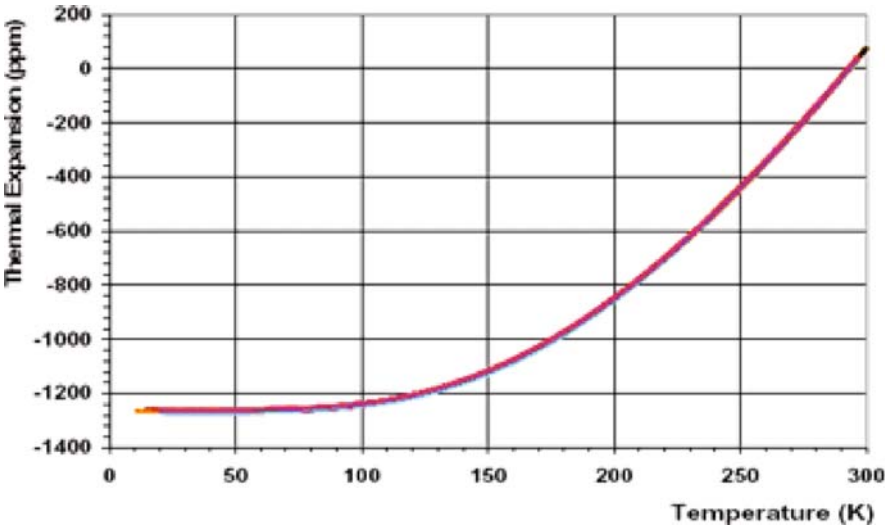


Fig. 5.13. Thermal expansion of beryllium (Parsonage, 2004).

with different properties. Generally a lower oxide content beryllium mirror has better rms error and low light scattering. The particle size should be about 40 to 110 μm with 99% less than 70 μm. The powder size can be tested from the light scattering.

Beryllium can be light-weighted by machining with conventional mills and lathes. Precautions must be taken when small particles (less than 10 μm) are produced. These particles can become airborne. Inhalation of beryllium dust can be fatal. However, there is no such danger during the wet polishing process. Beryllium mirrors can be bare polished to a roughness of 25 angstroms or be plated with electroless nickel, which has a matching CTE over a wide range of temperature. The latter one has a better surface roughness.

Further improvement for the JWST mirror is to combine beryllium material with weight reduction technology. Ball Aerospace & Technologies Corporation produced a test mirror segment using this approach (Figure 5.14). The beryllium mirror blank was made of O–30H spherical powder beryllium which is formed in an atomization process. The powder is not homogeneous; however, after the powder is blended and screened during the consolidation process (hot isostatic pressing), the resulting billet is homogeneous in bulk properties. At the back of the mirror segment, a number of machined triangular pockets reduce 92% of the mirror weight (Kendrick et al., 2003). This 1.4 m mirror segment requires very few actuators for its surface shape control.

The same as the University of Arizona's approach, the weight-reduced beryllium mirror segment is supported by a strong CFRP box reaction structure made of thin plates. The reaction structure is very light in weight and its areal density is only 2.6 kg/m^2. The connection between the mirror segment and the

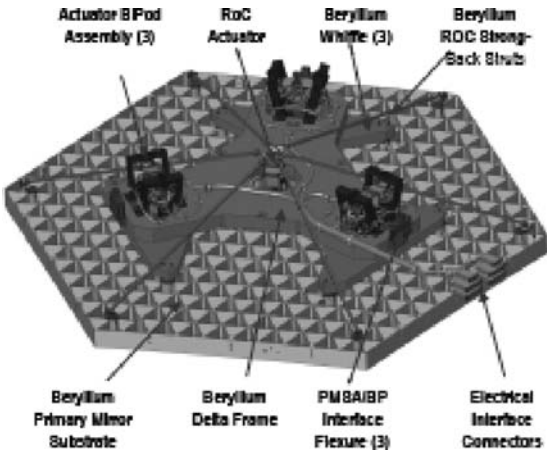


Fig. 5.14 Primary mirror segment design of the JWST (Atkinson et al., 2006)

reaction structure is through six actuators. These actuators are used to control both the curvature and position of the mirror segment.

In the mirror segment control, the co-phase between segments is very important. This is done in two steps: (a) a tip-tilt correction made for each mirror segment so that all the segments are parallel to each other; and (b) a dispersed fringe phasing sensor is used to find the piston errors between any two mirror segments. In both stages, a closed loop correction is performed under the orbit condition.

Mirror segment deployment is another technical challenge for the JWST project. There were two segment deployment plans. One is like flower petals which open up from both front and back sides of the primary dish. The other concept involves stacked mirror segments inside the fairing. The segments would deploy through rotation as well as translation. These two plans are illustrated in Figure 5.15. The JWST now uses the first deployment plan.

Besides the mirror manufacture and segment deployment, the JWST also has difficulties in its dynamic simulation. The telescope structure will suffer high acceleration and vibrations during the launch stage. At the same time, the temperature will change from the earth ambient environment to a very cold low temperature one. The thermal stresses induced may produce structural deformations and vibration. When the telescope is deployed, the accuracy and stability of each component is also important. The JWST requires no hysteresis for all its moving components.

5.3.3 The Space Interferometry Mission and Other Space Programs

While the JWST project is under development, another space program, the Space Interferometry Mission (SIM) or the SIM PlanetQuest, is also being

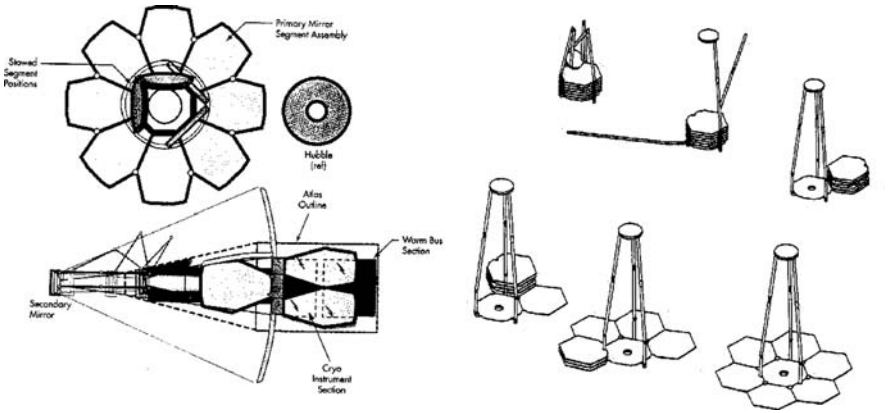


Fig. 5.15 Two concepts for deploying JWST mirror segments (Stockman, 1997)

planned at the Jet Propulsion Laboratory. The project is an optical interferometer operated from an Earth-trailing solar orbit. It was originally scheduled for launch in 2012, but recent news from NASA put the project in an indefinitely delay. Therefore, a small-scale SIM-Lite mission is now planned to save this project. The following text was written before this change.

The purpose of the SIM mission is to determine the positions of and the distances between stars several hundred times more accurately than any previous programs. The accuracy planned is one milliarcsecond, which is the thickness of a nickel, viewed at a distance of the moon. This accuracy will allow the SIM to probe nearby stars for earth-sized planets and to produce a new generation of star catalog. In its final orbit, the spacecraft will slowly drift away from the earth at a rate of approximately 17 million kilometers per year, reaching a maximum communication distance of about 95 million kilometers after 5.5 years. In this orbit the spacecraft will receive continuous solar illumination, avoiding occultation which occurs in a near earth orbit.

In its earlier design, all telescopes of the SIM system were movable along a rail on an optical bench. Recently, no movable telescopes on the platform are planned. The project includes three pairs of 0.3 m diameter optical telescopes. Each telescope has a siderostat device formed by two flat mirrors, which reflect star light through a particular path. Among three pairs of telescopes, two pairs form fringes of a bright star. The information obtained from the fringes is used for the system tracking and guiding, so that the system can reach the required stability. The last pair of telescopes is used for astrometry observation.

In the earlier revised design, although all telescopes are fixed on the bench, they still can form pairs out of any two telescopes so that the baseline can be adjusted in steps. The largest distance between any two telescopes is 10 m. As a space optical interferometer, there are too many difficulties in the design, manufacture, launching, and operation. Now, the system design has been changed again. The new design involves only fixed baselines. The length of the

observing baseline is now 9 m and those of two reference baselines are 7 m. The diameter of the telescopes is also reduced to 0.02 m and beam-reducers are used after siderostat mirrors in the system.

The absolute pointing accuracy of the SIM system is at the microarcsecond level. With the fringes from a bright star, the baseline orientation needs to be as good as a few tens of milliarcseconds. The optical path length delay of a stellar source between two telescopes of an interferometer can be expressed as a sum of two terms: a constant term and a dot product term of the baseline vector and unit star direction vector. To achieve milliarcsecond accuracy, the baseline length and the constant terms are solved by using an astrometric "grid" of stars over the entire sky. The limiting magnitude of the stars observed is about 20, which is also far beyond the magnitude of standard star catalog. Figure 5.16 lists the magnitude and accuracy of FK5 catalog, the Hipparcos catalogs, and the planned SIM catalog.

A schematic of the recent SIM design is shown in Figure 5.17. The positional accuracy of the optical system required is less than 1 nm. To achieve this accuracy, a precision metrology system is important. A major metrology system is located at six node positions of the optical bench. Two of these nodes form a guiding baseline. The other two nodes form a science baseline. On these nodes, there are special retroreflectors, or corner cubes, for position determination. These special corner cubes have two or three retroreflectors combined in one unit as shown in Figure 5.18. In the figure, three retroreflectors exist, however, one of them is not used. The axes of all retroreflectors in a corner cube pass through precisely a single vertex point.

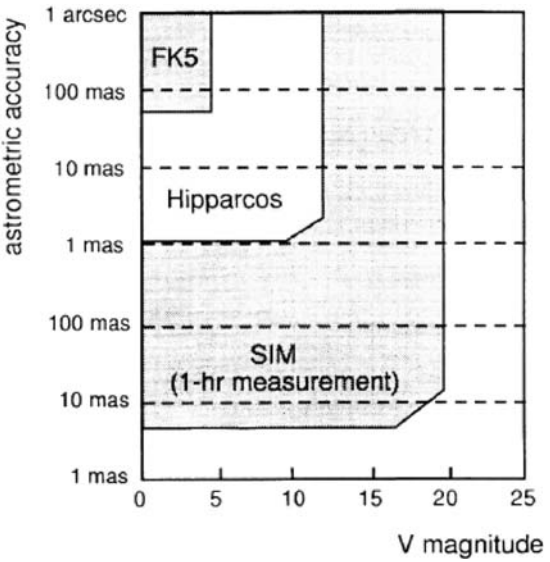


Fig. 5.16. The planned accuracy and magnitude of the SIM project (Shao, 1998)

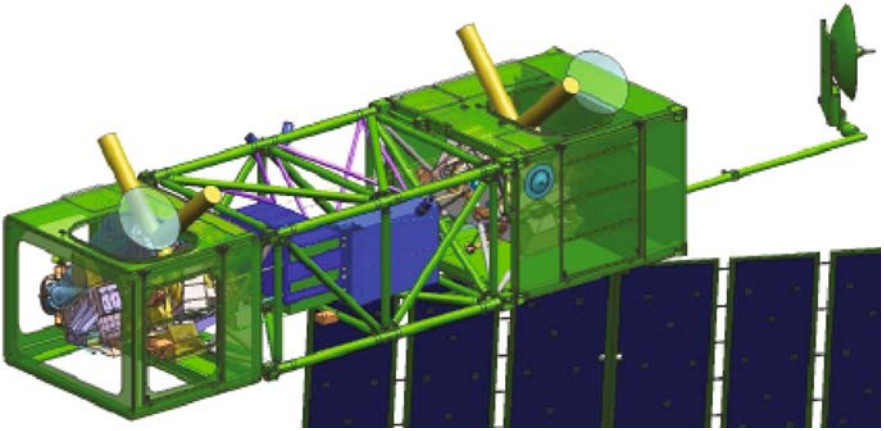


Fig. 5.17. A recent layout of the SIM project structure (Nemati, 2006)

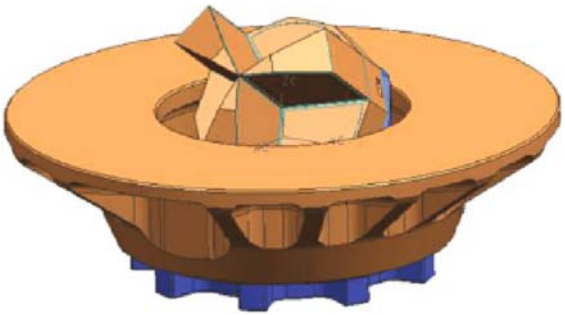


Fig. 5.18. Special multi-corner retroreflector used in the SIM project (Nemati, 2006)

The metrology system also includes several heterodyne laser ranger systems. The principle of the laser range system is discussed in Section 7.2.4. However, the SIM project requires higher accuracy than other projects, so that the frequency used in these laser rangers is also higher. This high laser frequency will be mixed with another laser frequency with a very small frequency difference. In this way, a small distance of picometer change is converted into a detectable phase change of the returned lower frequency wave signal. This wave frequency is the difference between the frequencies of two input laser signals. In space orbit the laser range is free from any error caused by the earth's atmosphere.

By triangulation, two laser beam lengths of a triangle can be accurately determined. If there are other laser rangers which form triangles over the same baseline, the length change of the baseline can be accurately determined. These laser measuring systems are similar to a truss system in structural design so they are named "optical trusses."

The structural vibration and thermal effect are extremely important in the SIM project, which result in a tiny length change of its interferometer baselines.

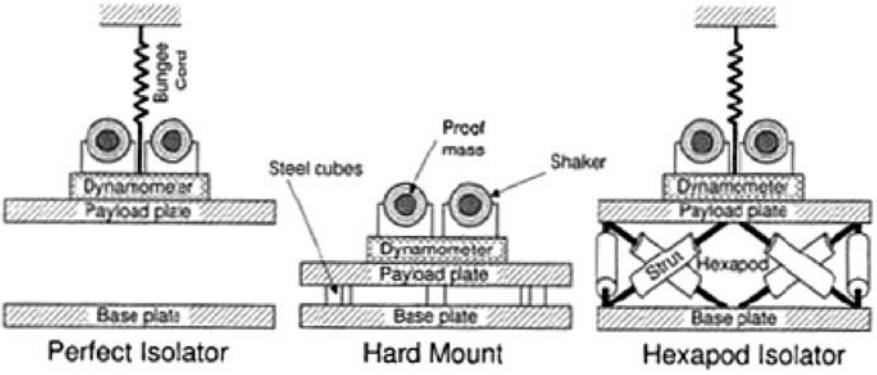


Fig. 5.19. Three types of vibration isolation systems (Goullioud et al., 2000)

These vibrations may be caused by the position control system (e.g. reaction wheels) of the spacecraft or by unavoidable thermal shock. A study of three vibration isolation systems has been carried out. These are: complete isolation, no isolation, and isolation through the Stewart platform (Figure 5.19). The study shows that complete isolation is unrealistic. If there is no isolation, the optical bench will not be stable even when very sophisticated active control devices are used. Only the Stewart platform isolation can remain stable by using an active control. The study of vibration isolation is still going on in the SIM project. Another detail of the SIM project is the optical delay line. The SIM project requires several stages, coarse and fine ones, in the delay line control.

The SIM project is a Michelson interferometer instrument; beams from the source are combined in the aperture plane after beam-reducers, producing fringes for celestial sources.

Other planned space missions include two Terrestrial Planet Finder (TPF) missions: a coronagraph named TPF-C and an interferometer named TPF-I. The TPF-C is an offset Cassegrain system with a primary mirror of 8 m size. A 12 m long deployed secondary tower supports the secondary mirror. At the image plane, an occulting mask is inserted to block the starlight so that the nearby planet image is formed at the camera's focus. The TPF-C is planned to be launched in 2016.

The TPF-I is a nulling interferometer with separated spacecraft. The TPF-I involves four collector and one combiner spacecraft in the system. The collector spacecraft carry telescopes and collect light from a target planet-and-star system. The light is relayed to a combiner spacecraft. The phases of these four collector systems are so adjusted that a $\pi/2$ phase difference exists between all the four beams. The phase difference between beams 1 & 3 is π and that between beams 2 & 4 is also π. The system first nulls the signals from pairs 1 & 3 and 2 & 4. Then the nulled outputs are cross-combined again for fringe final output. The nulling is at the star direction so that the planet information is recorded in the combined fringes. The TPF-I is planned to be launched in 2020.

Other on-going space test programs include the Fizeau Interferometer Testbed (FIT), the Formation Flying TestBed (FFTB), and the Synthetic Imaging Formation Flying Testbed (SIFFT). The ultimate goal of these programs is a sparse synthetic aperture optical telescope or a Fizeau interferometer with a large array of space-borne mirrors. The interferometer concept will involve 20 or more spherical mirrors over a large unfilled spherical surface. The aperture of this system will reach 0.5 km. Beams reflected from the mirrors are combined in the focal plane to form an interference image. The target angular resolution is between 60 and 120 microarcseconds. The interferometer patterns from sparse flat reflector mirrors are discussed in Section 9.2.4.

The FIT project is a small-scale laboratory test to ensure a close-loop control of separated articulated apertures at the nm-level accuracy for evaluating image synthesis algorithms and image reconstruction. The FFTB is a simulation of a stellar image interferometer with 30 mirror satellites at micrometer-level accuracy. The focal length of the mirrors is 5 km. The first stage of the project is to model the system through omni-directional radio frequency ranging sensors. The second stage is to have a close-loop control with laser ranging measurements to a micrometer level. The transverse position accuracy at this stage is at 10-cm level. The third stage is to steer target starlight into a baffle opening of the detector. The fourth stage is to control the target starlight to the center of the detector. The SIFFT is also a ground-based testbed to ensure cm-level positional precision of a multi-element flying array. The investigation includes multi-element formation (configuration) capture, formation maintenance, formation reconfiguration, and synthetic imaging maneuvers. Some of these test programs have reported very encouraging results for future further testing.

References

Atkinson, C. et al., 2006, Status of the JWST optical telescope element, SPIE Proc., 6265, 62650T.

Burge, J. H., et al., 1998, Lightweight mirror technology: using a thin face sheet with active rigid support, SPIE Proc. 3356, 690–701.

Cruise, A. M. et al., 1998, Principles of space instrument design, Cambridge university press.

Goullioud, R. et al., 2000, Micro-precision interferometer: scorecard on technology readiness for the space interferometry mission, SPIE Proc. 4006, 847.

Kendrick, S. E. et al., 2003, Lightweighted beryllium cryogenic mirror for both monolithic and segmented space telescopes, SPIE 4850, 241–253.

Merhav, S., 1996, Aerospace sensor systems and applications, Springer, New York.

Nemati, B., 2006, SIM planetquest: status and recent progress, SPIE Proc. 6268, 62680Q.

Parsonage, T., 2004, JWST Beryllium telescope material and substrate fabrication, SPIE Proc. 5494–5494, Glasgow, UK.

Petersen, C. C. and Brandt, J. C., 1995, Hubble vision, Cambridge University Press, Cambridge.

Shao, M., 1998, SIM the space interferometry mission, SPIE 3350, 536.

Stockman, H. S., 1997, The next generation space telescope, The association of universities for research in astronomy, Inc., Washington, D. C., NASA

Thornton, E. A., 1996, Thermal structure for aerospace applications, AIAA series, AIAA Press, Reston, Virginia.

Unwin, S. C. and Shao, M., 2000, The space interferometry Mission, SPIE Proc. 4006. 754.

Wertz, J. R. and Larson, W. J., 1991, Space mission analysis and design, Kluwer Academic Publishers, Boston.

Chapter 6
Fundamentals of Radio Telescopes

In this chapter, a brief review of radio astronomical telescopes is provided. The fundamental concepts of radio antennas, including radiation pattern, antenna gain, antenna temperature, antenna efficiency, and polarization, are introduced. These concepts are important for readers outside the radio antenna field. The emphasis of this chapter is placed on the parameter design of reflector radio telescope antennas. These parameter selection criteria are for both the primary focus and the Cassegrain focus reflector antennas. At the end of the chapter, characteristics of the offset antennas, where the polarization is usually a problem, are discussed. The receiver of radio telescope antenna is also introduced.

6.1 Brief History of Radio Telescopes

Electromagnetic (EM) radiation in space has both electric and magnetic components. These components are perpendicular to the direction of propagation. They oscillate 180° out of phase at right angles to each other. The EM radiation covers a very wide spectral range (refer to Figure 1.18 and Table 11.1) of which the visible light is only a very narrow portion. The EM radiation from roughly 10^{-3} to 10^{5} m, or a frequency band from 10^{12} to 10^{3} Hz, is called the radio spectrum. Radio telescopes are astronomical tools for the detection and collection of the radio waves coming from the universe. The radio spectrum is further divided into eight frequency bands; very low frequency (VLF) from 10 to 30 KHz, low frequency (LF) from 30 to 300 KHz, medium frequency (MF) from 300 KHz to 3 MHz, high frequency (HF) from 3 to 30 MHz, very high frequency (VHF) from 30 to 328.6 MHz, ultra high frequency (UHF), from 328.6 MHz to 2.9 GHz, super high frequency (SHF) from 2.9 to 30 GHz, and extremely high frequency (EHF) from 30 GHz and above. Some often used radio bands are L (1–2 GHz), S (2–4 GHz), C (4–8 GHz), X (8–12 GHz), Ku (12–18 GHz), K (18–26 GHz), Ka (26–40 GHz), V (40–75 GHz), and W (75–111 GHz) bands.

The development of optical telescopes has a history of four hundred years, while the development of radio telescopes has a history of only 80 years. Radio

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_6,

telescopes, or radio antennas for astronomy, are high-gain radio wave collectors and detectors. They are suited for the detection of faint radio signals from far away celestial bodies. In 1928, Karl Jansky developed a relatively directional antenna that turned out to be the first radio telescope. In 1932, he recorded radio signals coming from the center of the Milky Way galaxy. This marked the birth of radio astronomy. After Jansky's work, Grote Reber developed the first paraboloidal reflector radio telescope. Paraboloidal reflector antennas remain the most common type of radio telescopes today.

The wavelength of radio waves is much longer than that of its visible counterpart; therefore, a single radio antenna has a rather poor angular resolution. In 1945, Joseph Lade Pawsey et al. (1946) used an antenna overlooking the sea and created a radio interferometer by combining the reflected beam from the sea surface and the direct beam from the sky. In 1946, Ryle and Vonberg were the first to realize an interferometer made of two separate telescopes. An interferometer has a much higher angular resolution than a single antenna dish.

In the 1950s, two important large radio telescopes were completed. These were the Jodrell Bank 72.6 m antenna in England and the Mills Cross antenna in Australia. The Jodrell Bank paraboloidal antenna was the first astronomical telescope which employed a newly invented vacuum tube computer for its structural analysis. The structural calculation took almost a year of time of one early computer. The Mills Cross telescope consisted of two orthogonal long narrow aperture antennas, one in the south-north and another in the east-west directions. The length of each narrow aperture antenna was 450 m and the telescope formed a narrow pencil beam which had a relatively high angular resolution.

In the 1960s, more and larger radio telescopes were built for astronomy. These included a 91 m (300 ft) radio dish of the National Radio Astronomy Observatory (NRAO) and the 300 m fixed Arecibo antenna in Puerto Rico. In this period, John Ruze established an important antenna tolerance theory and Sebastian von Hoerner developed a new antenna homology design method. These new innovations made larger aperture, steerable radio antennas possible at short wavelengths. In 1972, the Max-Planck-Institute fur Radioastronomie (MPIfR) built a large, fully steerable homologous 100 m antenna in Effelsberg, Germany. In 1988, the aging 91 m NRAO telescope collapsed and, in 2000, it was replaced by the 100 m off-axis Green Bank Telescope (GBT). At the beginning of the 21st century, a Five-hundred-meter Aperture Spherical radio Telescope (FAST) with an active controlled surface began construction in China.

In 1972, the $78 M Very Large Array (VLA) project in New Mexico was approved in the US. This large radio interferometer has 27 movable antennas each having a diameter of 25 m. Three array arms form a 'Y' shape with each arm having a maximum length of 21 km. The project construction was completed in 1980. After the VLA, the Very Long Baseline Array (VLBA) composed of ten 25 m dishes was built, which extends across the whole US continent as well as the islands of St. Croix and Hawaii. In the long wavelength range, India and China had also built large meter wavelength aperture synthesis telescopes. The

Australia Telescope (AT) array with six 22 m dishes was built in the 1980s. The construction of these very large radio interferometer arrays makes the angular resolution in radio bands even better than those in the visible wavebands in spite of a 10^5 difference in the wavelength. An even more ambitious international project, the Square Kilometer Array (SKA) which includes low, middle, and high frequency arrays, is being planned.

Along with the construction of large radio telescopes, the usable radio band has been pushed into the shorter millimeter and submillimeter wavelength range. Since the 1960s, a number of small or medium size millimeter wavelength telescopes have been built around the world. However, as the wavelength becomes shorter, the telescope design becomes more difficult both in the surface and pointing requirements, making the design of the millimeter wavelength telescopes very different from that of other radio telescopes. Therefore, in this book, the design of millimeter wavelength telescopes is discussed separately in Chapter 8. Other astronomical telescopes are also discussed separately according to their wavelengths. However, it should be pointed out that all astronomical telescopes have many characteristics in common. Therefore, to fully understand all the design and structural problems of any type of telescope, readers should refer to all relevant chapters and sections.

6.2 Scientific Requirements for Radio Telescopes

The basic requirements of radio telescopes are similar to those of optical telescopes. These requirements are high sensitivity, high angular resolution, high dynamic range, and wide spectrum coverage. High sensitivity, which requires a large effective collecting area, is a common characteristic of all radio telescopes used in forefront research. The radiation collected on the earth from a celestial object in the radio band is very weak. If the received radiation is expressed as power per unit area orthogonal to the radiation direction, namely power flux density, then a flux density in the radio band for the strongest radio source, i.e. a quiet sun, is only about 10^{-20} Wm^{-2} Hz^{-1}. In radio astronomy, flux density is usually described with another smaller unit, *jansky* (Jy), specifically one jansky is equal to 10^{-26} $^{-}\mathrm{Wm}^{-2}$ Hz^{-1}. Some flux spectra of important radio sources are presented in Figure 6.1(a) (Kraus, 1986). In the radio range, a typical power flux density of celestial objects being studied is only about 1 mJy to 1 μJy. The flux density of celestial objects can be also expressed as a power of the frequency $S \sim v^a$, where a is the spectral index of the celestial objects. In Figure 6.1(b), selected typical spectral indexes of radio sources are marked on the relative flux density curves (Kause, 1986).

The sensitivity of a radio telescope is related to a number of factors, including antenna surface area, transmission loss, antenna surface accuracy, antenna radiation pattern, receiver noise performance, instantaneous bandwidth, and atmospheric attenuation. Different from observations using optical telescopes, radio telescopes are normally receiver noise dominated, so that the size of

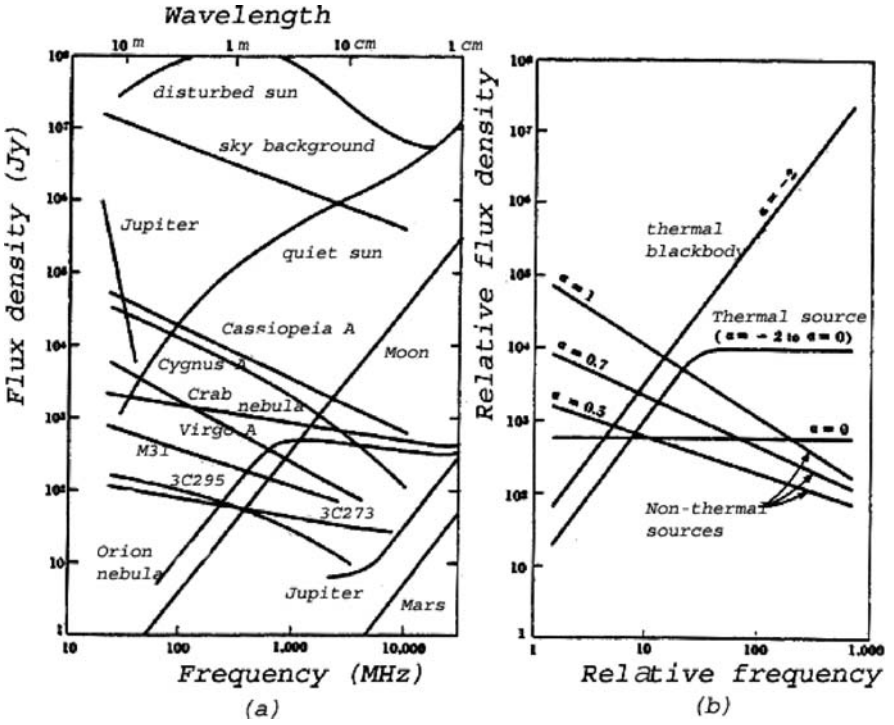


Fig. 6.1. (**a**) Spectra and (**b**) spectral index curves of selected radio sources (Kraus, 1986).

random fluctuations is usually independent of the strength of the radio source. They will always fluctuate randomly around a certain mean value irrespective of the existence of radio source or its radiation mechanism. The root mean square variation of the observed radiometer output is inversely proportional to the square root of the total integration time. That is to say, the accuracy of the measured mean value for N repeated measurements is $N^{1/2}$ times higher than that of a single measurement with the same integration time.

Within a narrow receiver band, the power spectrum of a radio source is usually assumed to be uniform. In this way, details of the spectrum of the radiation within the band are lost. Therefore, the root mean square error is also inversely proportional to the square root of the receiver frequency bandwidth. For receiver noise limited observation, the root mean square error ΔT_{rms} observed from a radio telescope is expressed by:

$$\Delta T_{rms} = M \frac{T}{\sqrt{\Delta v \cdot t}} \tag{6.1}$$

where M is a constant on the order of unity depending on the type of the receiver, T the system noise temperature given as the equivalent Rayleigh–Jeans power,

Δv the frequency bandwidth observed, and t the total integration time. The detection limit of a radio receiver is usually assumed to be between three and five times the root mean square of the error measured. If the receiver's fluctuation follows a Gaussian distribution, then the probability for generating a random measurement five times the rms is about 6×10^{-6}. In a radio receiver system, if the transmission efficiency (Section 7.1) is η_t, then the minimum measurable antenna temperature change is:

$$(\Delta T_a)_{\min} = 5\,M\frac{T}{\eta_t \cdot \sqrt{\Delta v \cdot t}} \tag{6.2}$$

If the source is extended, a contribution of the directional lobe efficiency η_B should be added to the above formula. The new formula becomes:

$$(\Delta T_B)_{\min} = 5\,M\frac{T}{\eta_t\eta_B \cdot \sqrt{\Delta v \cdot t}} \tag{6.3}$$

where the directional lobe efficiency, or the main beam efficiency, η_B, is:

$$\eta_B = \frac{1}{\lambda^2}\int\limits_{mainlobe} A(l,m)d\Omega \tag{6.4}$$

where $A(l,m)$ is the effective antenna area relative to the direction of (l,m). The concept of the main beam will be discussed in Section 6.4.1.

Among the factors which influence the sensitivity of radio astronomical observations at frequencies less than 5 GHz is the confusion noise within the solid angle of the antenna main beam (Condon, 2002). The confusion noise is the integral of signals from very faint radio sources within the solid angle of the main beam. As the main beam passes through these confusion sources while tracking the primary source of interest, the antenna temperature fluctuates. The amplitude difference (or amplitude of deflection) is expressed as the flux density of an assumed point source passing through the center of the beam. If the total number of radio sources, which have a flux density greater than a certain value S within the mainlobe is $N(0)$, then the following approximation exists (von Hoerner, 1961):

$$N(x) = aS^x \tag{6.5}$$

where the power spectral index x is determined by the cosmic model applied. The constant a can be determined from the actual observations after the cosmic model is determined. If within the main-beam solid angle there are n radio sources which have flux density greater than a given value $S_{\lim}$, then the number of radio sources whose flux density is greater than a certain value S within the main-beam solid angle, will be:

$$N = n\left(\frac{S}{S_{\lim}}\right)^{x} \tag{6.6}$$

The number of sources, whose flux density is between S^x and S^{x-1} within the main-lobe solid angle, will be:

$$dN = nx\left(\frac{S^{x-1}}{S_{\lim}^{x}}\right)dS \tag{6.7}$$

Assuming the distributions of all radio sources are random, then their total number should obey a Poisson distribution, which has a standard deviation of $(dN)^{1/2}$. The background radiation caused by these radio sources should have a similar standard deviation of $\sigma(dN)^{1/2}$. Unfortunately, the integration of the above approach may not converge.

Condon (2002) pointed out that the confusion is directly proportional to the telescope beam solid angle and follows a –0.76 power law of the frequency at the centimeter wavelengths. He provided a simple approximation of the rms confusion in a Gaussian beam as:

$$\frac{\sigma_c}{mJy \cdot beam^{-1}} \approx 0.2\left(\frac{v}{GHz}\right)^{-0.76}\left(\frac{\theta_M\theta_m}{arc\,\sec^2}\right) \tag{6.8}$$

where θ_M and θ_m are FWHM major and minor diameters of the beam. In the detection, only sources stronger than the above $5\sigma_c$ can be detected reliably. There are about 25 beam areas per source stronger than $5\sigma_c$, so extracting more than one source per 25 beam areas from a confusion-limited image is dangerous. Condon (1974) provided more detailed formulas of confusion standard deviation from the cutoff amplitude difference for the exponent value between $-3 < x < -2$.

The resolution of radio telescopes depends on the size of the aperture and the observing wavelength. A larger aperture and a shorter wavelength improve the telescope resolution. In radio observations, the amplitude beam pattern (or field pattern if the phase term is included) and power beam pattern are used instead of the point or intensity spread function. For a circular aperture with a uniform illumination, the half-power beam width (HPBW) of the power beam pattern is $1.02\lambda/D$. The angular distance from the primary beam maximum to the first zero of the power beam pattern, is $1.22\lambda/D$.

To further improve the resolution of a radio telescope, the technique of interferometry is necessary. Types of interferometers include adding interferometers, correlation interferometers, and aperture synthesis telescopes. The angular resolution or the ultimate spatial frequency resolved for an interferometer is determined by the largest baseline length perpendicular to the observation direction and the observing wavelength. The basic theory of an interferometer is discussed in Sections 1.4.4, 4.2, and 7.3. Unlike optical telescopes, the atmospheric disturbance in the radio bands for an interferometer is nearly equal to

that for a single antenna at the same wavelength. Therefore, phase adjustment used in radio "active optics" can be performed after the observation has been taken through editing and calibration. The very long baseline radio interferometers can achieve a much higher spatial resolution than currently obtained in the optical wave bands (Refer to Preface of English Edition).

Radio interferometers provide the most useful research tool on fine structures of radio sources. An interferometer with a baseline separation D can provide information on an angular scale of λ/D. If there are interferometer pairs of different baselines, the visibility information at different spatial frequencies can be determined. When the visibility information is sufficient, aperture synthesis techniques allow mapping all of a source area. In a one-dimensional case, the visibility function $V(s_\lambda)$ is (Equation 1.136):

$$V(s_\lambda) = \int B(\phi) \exp[2\pi i s_\lambda \phi] d\phi \tag{6.9}$$

where $B(\phi)$ is the source brightness distribution and s_λ the spatial frequency expressed as the reciprocal of baseline length. If we obtain values of the function $V(s_\lambda)$ for a set of different baselines, the source image may be obtained by an inverse Fourier transform.

The dynamic range is to describe the ratio between the smallest and largest possible values in the observation. Higher dynamic range includes more information in the observation

The other requirement for radio telescopes is that the observation can be carried out over a fairly wide frequency range. Paraboloidal reflector antennas have this property and are popular in radio astronomy.

6.3 Atmospheric Radio Windows and Site Selection

The atmosphere is transparent or partly transparent in radio wave range from about 15 m (or 200 m) to 1 mm in wavelength. The cutoff in long wavelength end is caused by ionosphere condition which changes from time to time. This cutoff frequency is the ionosphere critical frequency. The ionosphere critical frequency is proportional to the square root of electron density as $f_{cri} = 9 \cdot 10^{-3} \cdot N^{-1/2}$, where N is the electron density per cm^3 and f_{cri} is in MHz. Generally, the ionosphere critical frequency is between 9 and 15 MHz. When the frequency is higher than the ionosphere critical frequency, radiation from space can penetrate the atmosphere.

In general, the atmospheric influences include three major groups: molecular absorption, atmospheric refraction, and the extinction or scattering by dust and aerosol particles. In the long meter to centimeter wavelength bands, the influence on radiation from the atmosphere is small. The effect and the absorption from water vapor or rain are not serious. However, when the frequency is higher

than 10 GHz, the absorption and scattering from the atmosphere, from rain, or from water vapor becomes serious. The higher the frequencies, the more serious the effects are expected in a few narrow windows. The atmosphere also affects the phase of the radio beams transmission just as in the optical region even when the attenuation is negligible. These phase fluctuations are particularly important in interferometry work.

Very low frequency radio waves between 3 and 30 kHz penetrate sea water. Low frequency radio waves between 30 and 300 kHz travel easily through brick and stone. As the frequency rises, the absorption effect from the atmosphere becomes important. At microwave or higher frequencies, molecular resonance absorption from the atmosphere (mostly water, H_2O and oxygen, O_2) becomes dominant. Molecular absorption takes place when the radiation energy equals the difference between two rotational or vibrational energy levels, namely

$$E_{high} - E_{low} = hv \tag{6.10}$$

where h is the Planck constant and v the radiation frequency. In the lower atmosphere, the absorption is mainly from the water molecules. The molecular collision also widens the absorption lines. Figure 6.2 shows the relationship between the density of major molecules of the atmosphere and the altitude above sea level. The low atmosphere has a high concentration of water molecules and the absorption is serious. Figures 6.3 and 6.4 show typical absorption

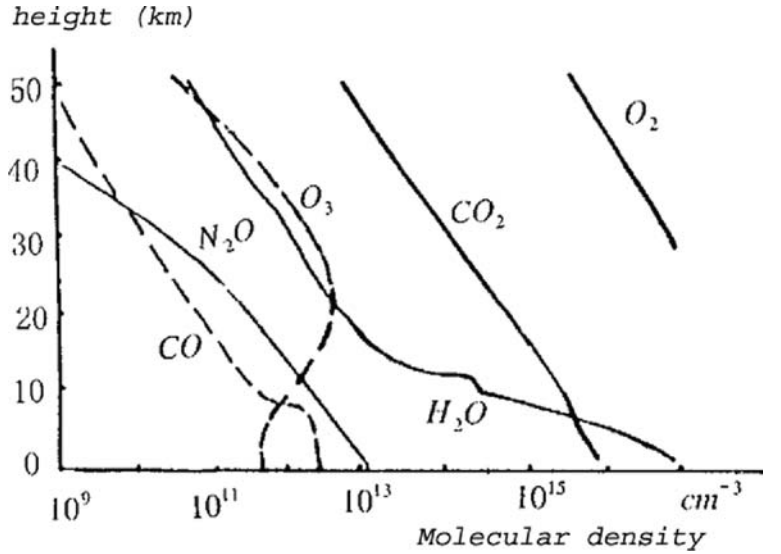


Fig. 6.2. Density distribution of molecules which influence the atmospheric absorption in summer.

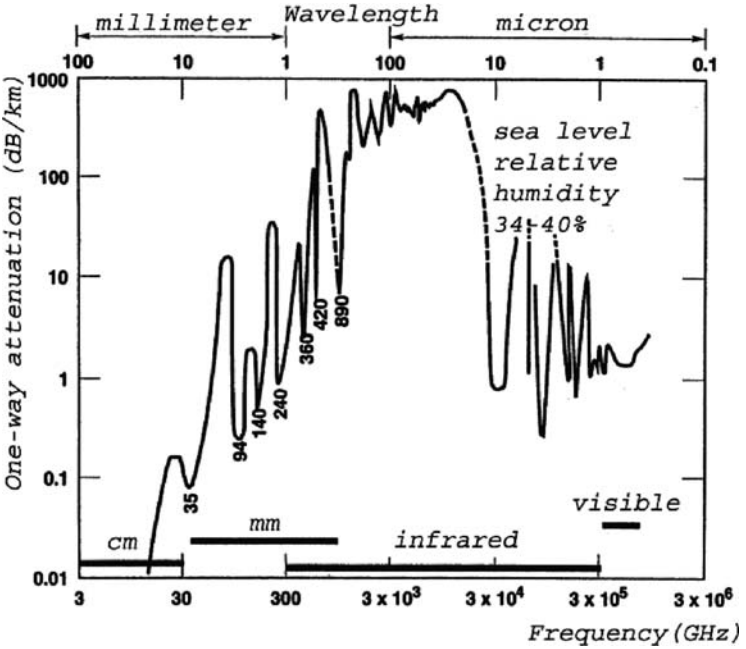


Fig. 6.3. Typical atmosphere absorption spectrum at sea level (NASA).

spectra of the atmosphere at sea level. A number of absorption lines from water and oxygen are indicated in Figure 6.4. Bi-molecular water structure $(H_2O)_2$ and tri-molecular structure $(H_2O)_3$ can be formed through hydrogen bonds. These compound molecules have their own absorption spectra depending on the temperature and the strength of the hydrogen bonds. The effects from these compound molecules have been considered in Figure 6.3. For different altitude above sea level, the absorption spectra vary widely because of water vapor content.

Atmospheric refraction is a phenomenon caused by regular or irregular refractive index changes of the atmosphere. Refraction is very small for wavelengths longer than several centimeters. Regular or repeatable atmospheric refraction influences only the apparent positions of radio sources and it can be corrected by using pointing formulas. The refraction index of the atmosphere can be expressed as (Bean, 1962):

$$N = \frac{77.6}{T}\left(P + \frac{4810e}{T}\right) \tag{6.11}$$

where T is the temperature in K, P the air pressure in *millibar*, and e the water vapor pressure in *millibar*. This formula is used for frequencies up to 100 GHz. However, when the frequency is above 134 GHz, the radio refractive index becomes very close to the optical refractive index. Discussion of atmosphere

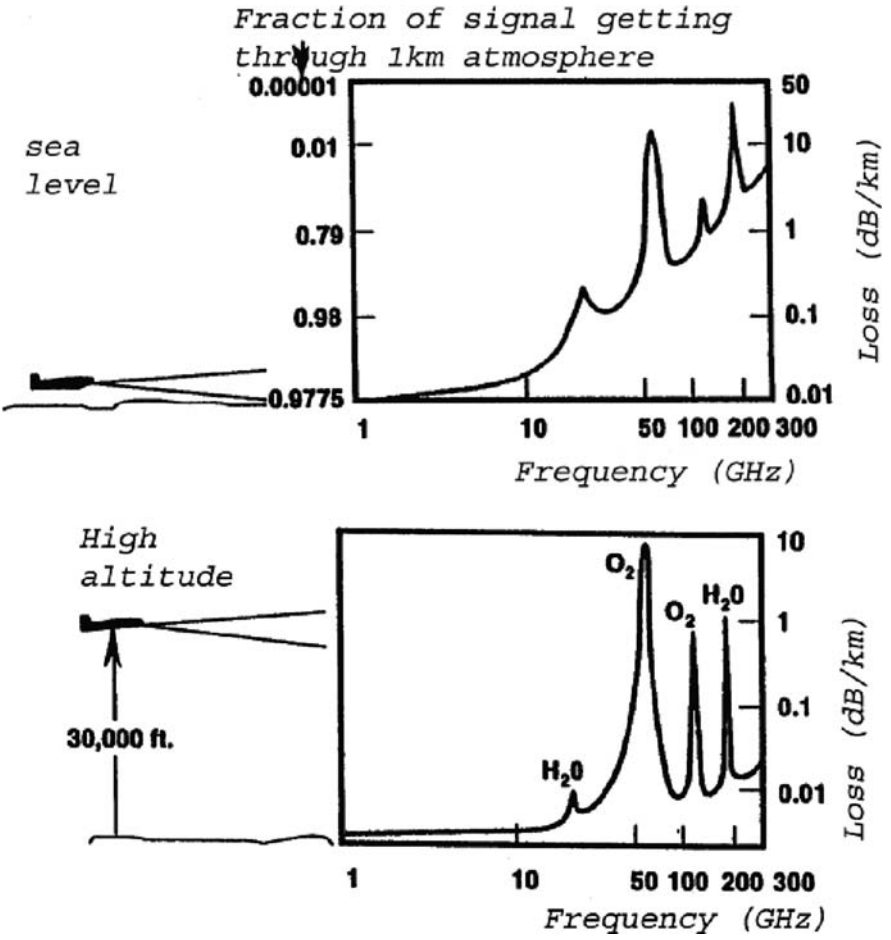


Fig. 6.4. Atmospheric absorption spectra at different altitudes (NASA).

refraction in the optical region can be found in Section 3.3.4. In radio bands, irregular refractive index change is small; the absolute value is similar to that in the optical region. The maximum root mean square value of the path length difference is just about one centimeter per km distance, which corresponds to 2 arcsec of the position change. Furthermore, this path length change follows a power law discussed in Section 4.1.6 so that it does not affect observations with large aperture radio telescopes. However, at millimeter wavelengths, the effect is more serious.

Scattering and attenuation by dust and aerosol particles can also produce more problems for high frequency observation. Scattering takes place when the refractive index of particles is different from the atmosphere. The attenuation is caused by the imaginary part of the refractive index. The radiation energy

absorbed by aerosol particles turns into heat. The aerosols in the atmosphere are mainly water in liquid form, such as rain and fog. Ice attenuation is about 10% of that of liquid water because of a smaller imaginary part of the refractive index. Water particles in fog and clouds are far smaller than the wavelengths in millimeter and centimeter wave bands. The attenuation of this type does not depend on particle size but on the density.

Figure 6.5 shows the attenuation by rain and fog in different conditions; the abscissa is frequency in GHz and the ordinate is the attenuation in dB/km. The curves from top to bottom are for rainfall of 100, 50, 25, 12.5, 2.5, 1.25, and

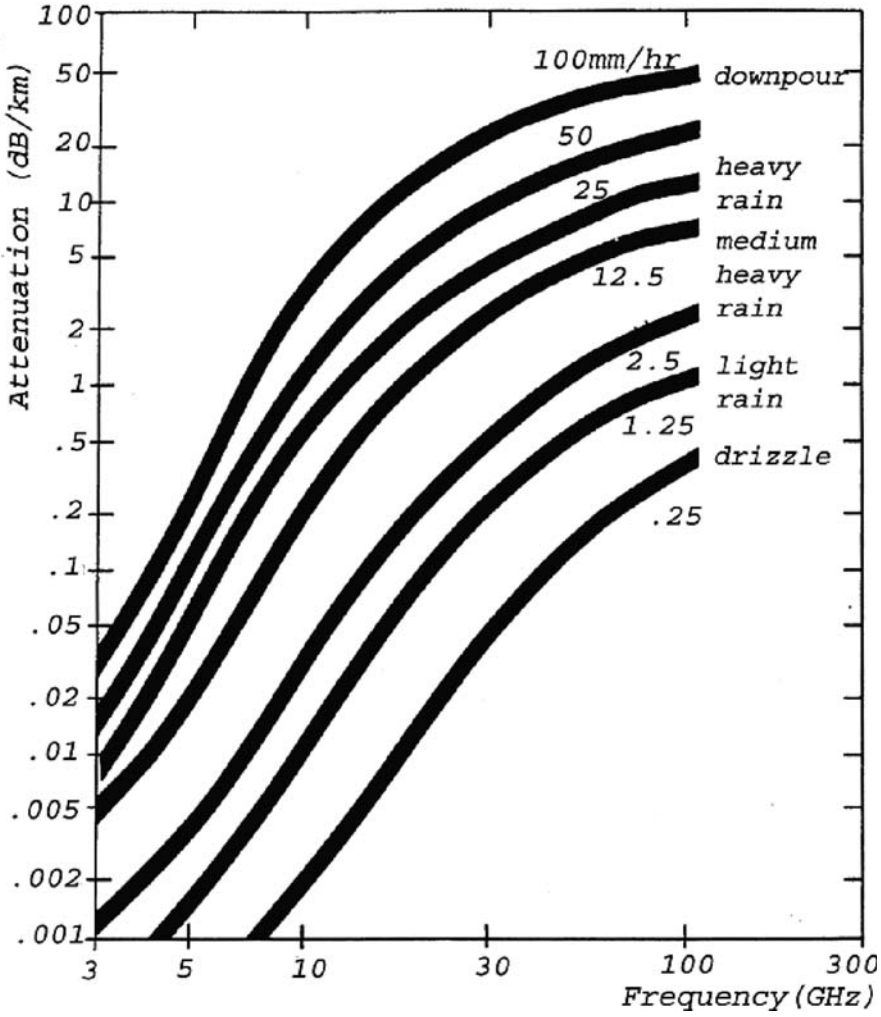


Fig. 6.5. Extinction spectra of rain and mist in different conditions (NASA).

0.25 mm/hr, respectively. For clouds and fog when the water vapor density is 1 g/m^3, the visibility in the optical region is 20 m, which corresponds to 800 dB/km, and the attenuation coefficient is far less in radio wavebands. The visibility is a measure of the distance at which an object can be clearly discerned. However, with the increase of frequency, cloud and fog affect the astronomical observations. The effects from dust and aerosol particles in the atmosphere also include the random thermal noise produced by molecular radiation. These noises fluctuate continuously.

Major atmospheric effects on radio observation occur at high frequency. Therefore, site selection becomes necessary only for observations in these frequency regions. In the low-frequency regions, the observatory site should be away from man-made radio signal sources. In the high-frequency regions, the site should have lower water vapor contents and a lower dust density. The water vapor content is the height of the total water, liquid, solid or vapor condensed in liquid form in the atmosphere. High-frequency radio observatories are usually located in high mountain areas far away from cities. For observations in the millimeter wavelength, the water vapor contents of the observatory site should be below 4 mm most of the year, and for observations in the submillimeter wavelength, the water vapor content should be less than 1 mm during the observations. The 13.7 m Purple Mountain Observatory millimeter wavelength telescope is located at an altitude of 3,204 m in Delingha, Qinghai plateau. The 15 m JCMT millimeter wave telescope is located at an altitude of 4,200 m on Mauna Kea, Hawaii Island. The Atacama Large Millimeter Array (ALMA), built jointly by the US, Europe, Japan, Canada, and Chile will be sited in a high mountain area of the Atacama Desert in northern Chile with an altitude of 5,000 m. The South Pole Submillimeter wave Telescope is located at a South Pole Station with an altitude of 2,835 m. The water vapor contents of these last three sites are documented in Table 6.1 by Lane (1998). From the table, the best site on earth for the submillimeter and infrared wavelength observations is the South Pole. The South Pole also has a lower wind velocity and no rain all year around. The highest wind velocity from 1953 to 1987 was only 24 m/s. However, the South Pole has an annual low average temperature of -49°C. Access to the South Pole is also difficult.

Table 6.1. Water vapor contents of three telescope sites (mm of water) (Lane, 1998)

Percentage of time	South pole		Mauna Kea		Atacama	
	Winter	Summer	Winter	Summer	Winter	Summer
25%	0.19	0.34	1.05	1.73	0.68	1.1
50%	0.25	0.47	1.65	2.98	1.0	2.0
75%	0.32	0.67	3.15	5.88	1.6	3.7

6.4 Parameters of Radio Antennas

6.4.1 Radiation Pattern

The radiation pattern, or antenna pattern, or far field pattern, is the most important parameter of a radio antenna. It represents the directional, or angular, dependence of radiation from an antenna or the response of an antenna to a far away point source. It is similar to, but not exactly the same as, the point spread function used in optical telescopes. For antennas, it follows an important reciprocity law. Therefore, the antenna pattern is identical either for transmitting or for receiving radiation.

Depending on physical quantities used, the antenna pattern can be a field pattern, an amplitude pattern, a power pattern, or a polarization pattern. The antenna pattern is usually expressed in polar coordinates. Various parts of the radiation pattern are referred to as lobes or beams. In Figure 6.6(a), the lobe with the maximum response is the main (or major) lobe. The lobes closest to the main lobe are the first sidelobes. The ratio between the maximum of the first sidelobe and the main lobe is the sidelobe level. The other sidelobes are called the second sidelobe, the third sidelobe, etc. The lobes in the reverse direction of the main lobe are back lobes. The sidelobes are an important part of the spillover, while the spillover also includes radiations directly picked from sources outside the primary reflector of an antenna.

The pattern is usually normalized using the response at the center of the main lobe. The pattern can also be expressed in Cartesian coordinates [Figure 6.6(b)]. The directivity of a radio telescope can be described by half power beam width (HPBW), which is the angle θ where the power declines to a half of the maximum in the power pattern. For a field pattern, this angle is at a level of 0.707 of the maximum field position. The power pattern is often calibrated in decibel (dB) scale. The dB value was originally invented for the power transmission line attenuation. A 1 dB loss is equal to that of the fractional power lost along a one-mile telephone line. In mathematical notation, $P(\mathrm{dB}) = 10 \log_{10}(P_2/P_1)$, where P_2 is the output power, and P_1 is the input power. A 20 dB gain

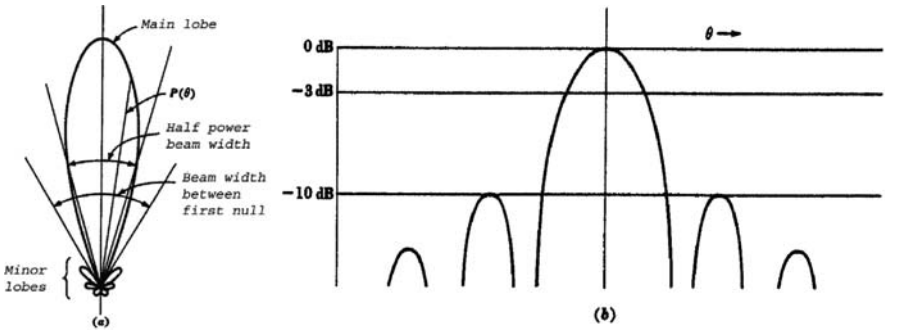


Fig. 6.6. Antenna pattern expressed in (**a**) polar and (**b**) Cartesian coordinates.

means that the output power is 100 times the input power. When using the dB scale, the HPBW is at the −3 dB position.

6.4.2 Antenna Gain

Antenna gain is defined as the ratio of the radiation intensity of an antenna in a given direction to the intensity that would be produced by a hypothetical ideal antenna that radiates equally in all directions (isotropically) and has no losses. An ideal hypothetical isotropic antenna transmits or receives radiation in all directions equally, whereas a high-gain antenna will preferentially radiate or receive in a particular direction. If we define the effective collecting area of an antenna in one direction as $A(v, \theta, \phi)$ square meters, where v is the frequency and θ and φ are the directional coordinates, then the power received when the antenna is pointed at a source with a brightness of $I(v, \theta, \phi)$ is:

$$P_1(\theta_0, \phi_0) = \int A(v, \theta - \theta_0, \phi - \phi_0) I(v, \theta, \phi) dv d\Omega \tag{6.12}$$

The normalized power pattern is $P(v, \theta, \phi) = A(v, \theta, \phi)/A_0$, where A_0 is the effective area of the antenna or the response at the center of the main lobe. In antenna field, the beam solid angle is a hypothetical angle defined as:

$$\Omega_A(v) = \iint P(v, \theta, \phi) d\Omega \tag{6.13}$$

where $P(\theta, \phi)$ is the normalized power pattern and the integral is within the main beam area. The beam solid angle is the angle through which all the power from a transmitting antenna would go if the power per unit solid angle were constant and were equal to the maximum value over this angle. If a radiation pattern is equal to unity everywhere, the maximum possible beam solid angle of this hypothetical isotropic antenna is 4π.

Therefore, the antenna directive gain is also represented as:

$$G = \frac{4\pi}{\Omega_A} \tag{6.14}$$

Antenna gain can be expressed in dBd or dBi units. dBd is referenced to a half-wavelength dipole and dBi is referenced to an ideal isotropic antenna. A half-wavelength dipole has a gain of 2.15 dBi. A single 20-wavelength-long Yagi antenna might have a gain of 20 dBd (ref. Section 7.2.1). Stacking two Yagis ideally adds 3 dB although the precise gain, higher or lower than 3 dB, is usually affected by mutual coupling between antennas.

An important relationship in antenna theory is that the product of the effective area and beam solid angle is equal to square of wavelength (Kraus, 1986). That is to say, the sensitivity of an antenna is inversely proportional to

the field of view for a given wavelength. An isotropic antenna can see the whole sky with equal sensitivity, but it has a very small effective collecting area. When the collecting area increases, the antenna field of view decreases. Some large scale details of the source are lost. This important relationship also governs the design of modern aperture synthesis telescopes.

This relationship also explains why wire antennas can be used with enough sensitivity for wavelengths longer than about 1 m in astronomy. The effective area is proportional to the square of the wavelength. These wire antennas include dipoles, Yagis, spirals, and helices. Arrays of these antennas may be used with an increased total collecting area. For wavelengths shorter than about 1 m, reflecting antennas are common. For a wide illumination angle, horn antennas can be used at the reflector focus.

For a radio source with a given flux density, the effective area determines the maximum energy received by a radio telescope. The ratio between the effective area and the actual geometric area of a reflector antenna is the aperture efficiency, although a simple wire antenna does not have a geometric area.

In radio astronomical observations, there is a very important relationship between the intensity observed and the pointing accuracy of the telescope. If the pointing error is equal to one tenth of the full width at half maximum (FWHM), the peak intensity observed will decrease to ∼0.97 of the maximum value.

6.4.3 Antenna Temperature and Noise Temperature

When an antenna receives radiation from an unpolarized radio source in the sky, the received power flux (W/Hz) can be expressed as:

$$P = \frac{1}{2}\iint A(\theta_0 - \theta, \phi_0 - \phi)B(\theta, \phi)d\theta d\phi \tag{6.15}$$

where A is the antenna effective area expressed as a function of the pointing direction, B the brightness distribution of a radio source, and θ_0 and ϕ_0 a reference direction. The factor $1/2$ takes into account that the antenna can only receive one polarized component of the radiation, while the source emits in both polarizations. The power received by an antenna is usually represented as the "antenna temperature" T_a (K), which is the temperature of a fictitious resistor:

$$T_a = \frac{1}{2k}\iint A(\theta_0 - \theta, \phi_0 - \phi)B(\theta, \phi)d\theta d\phi = \frac{P}{k} \tag{6.16}$$

where $k = 1.38 \times 10^{-23} WHz^{-1}K^{-1}$ is the Boltzmann constant. The temperature T_a is just an expression of power flux received and, often, it has nothing to do with any actual temperature.

In radio astronomy, the brightness of a radio source, in units of flux density per solid angle, can be expressed as a temperature of an equivalent blackbody

whether or not the source really is a blackbody. This equivalent temperature is the brightness temperature of the source. Usually a Rayleigh–Jeans law exists between radio brightness and its brightness temperature. When $h\nu \ll kT_B$, the radio brightness B is:

$$B \approx 2kT_B/\lambda^2 (W \cdot m^{-2} Hz^{-1} sterad^{-1}) \tag{6.17}$$

Using the brightness temperature T_B to express the signal received from the source, the antenna temperature T_a has the formula:

$$T_a = \lambda^{-2} \iint A(\theta_0 - \theta, \phi_0 - \phi) T_B(\theta, \phi) d\theta d\phi \tag{6.18}$$

The signal received in radio waves is extremely weak and is "noise-like" (white noise, containing all frequencies in the band). For convenience, it is often considered as an equivalent noise temperature corresponding to the power level received. Therefore, the output power of the receiver includes two parts: the antenna temperature, or the signal, and the added noise of the system also named the system temperature:

$$P_{tot} = P_a + P_{sys} \Rightarrow T_{tot} = T_a + T_{sys} \tag{6.19}$$

The system temperature or the system noise temperature in a radio telescope should be kept as low as possible. The system noise temperature of a radio telescope can be divided into several contributions:

$$T_{sys} = T_{bg} + T_{sky} + T_{spill} + T_{loss} + T_{cal} + T_{rx} \tag{6.20}$$

where T_{bg} is the noise from the microwave and galactic background, T_{sky} the noise from the atmospheric emission, T_{spill} the noise due to spillover and ground scattering, T_{loss} the noise due to losses in feed, T_{cal} the injected nosie, and T_{rx} the reciever noise temperature. The first three contributions are sky position related. The noises due to spillover, ground scattering, and the feed are from the antenna. The system temperature varies from tens to hundreds of Ks depending upon the observing wavelengths, while the signal or the "noise" from the source is generally only hundredths of Ks for a 10 m size antenna.

The system noise factor F is defined as the input signal to noise ratio S_i/N_i divided by the output signal to noise ratio S_o/N_o:

$$F = \frac{S_i/N_i}{S_o/N_o} \tag{6.21}$$

For a linear cascaded system, the system noise factor and noise temperature can also be calculated from Friis' formula:

$$\begin{aligned} F_{1n} &= F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \cdots + \frac{F_n - 1}{G_1 G_2 \cdots G_{n-1}} \\ T_{Nn} &= T_{N1} + \frac{T_{N2}}{G_1} + \frac{T_{N3}}{G_1 G_2} + \cdots + \frac{T_{Nn}}{G_1 G_2 \cdots G_{n-1}} \end{aligned} \tag{6.22}$$

where F_i, G_i, and T_{Ni} are noise factor, gain, and noise temperature of each subsystem, respectively.

6.4.4 Antenna Efficiency

Antenna efficiency is another important parameter in antenna design. It is expressed as:

$$\eta = \eta_1 \eta_2 \eta_3 \eta_4 \cdot \cdots \cdot \eta_n \tag{6.23}$$

where $\eta_1, \eta_2, \ldots \ldots \eta_n$ are respectively antenna aperture field efficiency, blockage efficiency, surface error efficiency, component positioning error efficiency, etc. Another expression of the antenna efficiency for reflector antennas is a ratio between the effective and geometric areas:

$$\eta = \frac{A_e}{A_g} \tag{6.24}$$

where A_e and A_g are the effective and the geometric aperture areas. The geometric area is the projected antenna aperture on the plane perpendicular to the electronic axis. The antenna efficiency, antenna gain, and antenna effective area are related to each other.

Among factors of antenna efficiency, the aperture field and surface error ones are very important. The antenna surface error is discussed in Section 7.1.2. The antenna aperture efficiency is due to the illumination by the feed. The highest aperture field efficiency corresponds to a uniform aperture illumination. Optical telescopes always have a uniform aperture illumination. However, the aperture field illumination for radio antennas is always nonuniform and illumination by the feed outside the reflector (spillover) is unwanted. First, no illumination pattern (radiation pattern) of a feed has a sharp cutoff at the aperture edge so edge taper, or grading, is inevitable. Second, solid angles extended from a feed to a same aperture sub-area δA in the dish center and on the edge are not the same (Figure 6.7).

The illumination grading of a radio antenna aperture field can be a Gaussian function:

$$g_{Gaussian}(\rho) = \exp\left[-\alpha\left(\frac{r}{r_0}\right)^2\right] \tag{6.25}$$

where $\alpha = (T_e/20)\ln 10$ and T_e is the edge taper in dB, or in the following expressions (Christiansen & Hogbom, 1985):

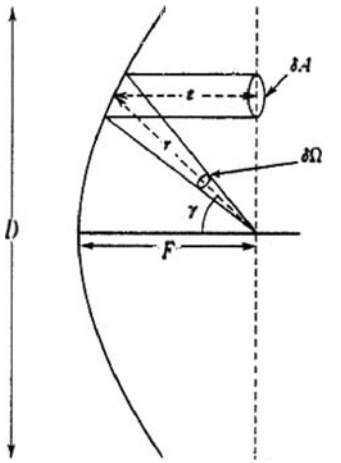


Fig. 6.7. Sub-area δA at the aperture plane and its corresponding illumination angles $\delta\Omega$.

$$g(\rho) = K + [1 - (\rho/a)^2]^p \tag{6.26}$$

where K and p are constants and ρ/a the normalized aperture radius. Even if there are no other losses, the antenna radiation pattern as well as the aperture field efficiency will be affected by the values of K and p. Figure 6.8 shows the illumination distributions at aperture plane for different K and p. Table 6.2 lists antenna parameters for different aperture illuminations. The antenna aperture field efficiency can be determined by the following formula:

$$\eta = \frac{4\pi}{\lambda^2} \frac{\left| \int g(\rho) \rho d\rho \right|^2}{\int |g(\rho)|^2 \rho d\rho} \tag{6.27}$$

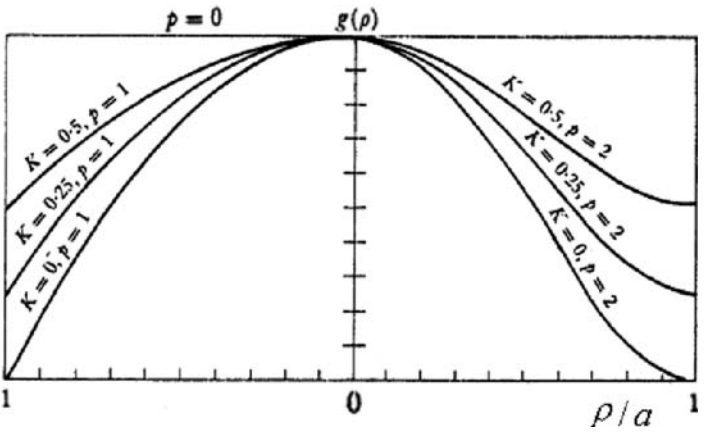


Fig. 6.8. Different aperture illumination functions (Christiansen and Hogbom, 1985).

Table 6.2. The aperture illumination properties when $g(\rho) = K - [1 - (\rho/a)^2]^p$

K	p	Main lobe half power width λ/D	Distance to first zero λ/D	Sidelobe level	Aperture efficiency
0	0	1.02	1.22	17.6	100
1	0	1.27	1.62	24.7	75
2	0	1.47	1.03	30.7	55
1	0.25	1.17	1.49	23.7	87
2	0.25	1.23	1.68	32.3	81
1	0.5	1.13	1.33	22.0	92
2	0.5	1.16	1.51	26.5	88

6.4.5 Polarization Properties

The polarization of an electromagnetic wave is defined from the trace of (or) the orientation of the electric field vector, specifically, its time-varying direction and relative magnitude. According to the trace, the polarization may be linear, circular, or elliptical. The polarization of an antenna in a given direction is its ability to respond to electromagnetic radiation of the same polarization. If the antenna polarization matches the polarized direction of the electromagnetic radiation, the antenna will have a maximum gain.

The cross-polarization means an electric vector response with a direction perpendicular to a certain reference plane (E plane). The co-polarization means a response of the polarized component parallel to the intended polarization direction. The cross-polarization (field) represents polarization orthogonal to a specified polarization, usually the co-polarization. Figure 6.8 shows the responses of co-polarization and cross-polarization for a paraboloidal surface. For this reflector antenna, because of the surface curvature, a different direction has different effective polarization. From the boundary conditions of a paraboloidal surface for the electric field:

$$\vec{n} \times (\vec{e}_0 + \vec{e}_1) = 0 \tag{6.28}$$

where $\vec{n}$ is the unit normal vector perpendicular to the reflector and $\vec{e}_0$ and $\vec{e}_1$ directional vectors of co-polarization and cross-polarization. For co-polarization direction, the components of the electric field E have the same direction in the aperture. However for cross-polarization, the electric field has opposite directions in different quadrants (Figure 6.9). They are 180° out of phase.

Because of the symmetry of the reflector, the cross-polarization disappears on co-polarization planes [Figure 6.10(a)]. In the figure, a small difference between co-polarizations in 0° and 90° directions is caused by the asymmetry of feed illumination on the aperture plane. The cross-polarization reaches a maximum value at the 45° and 135° directions. In these two directions, the sidelobe of cross-polarization is at the same position of the first minimum of

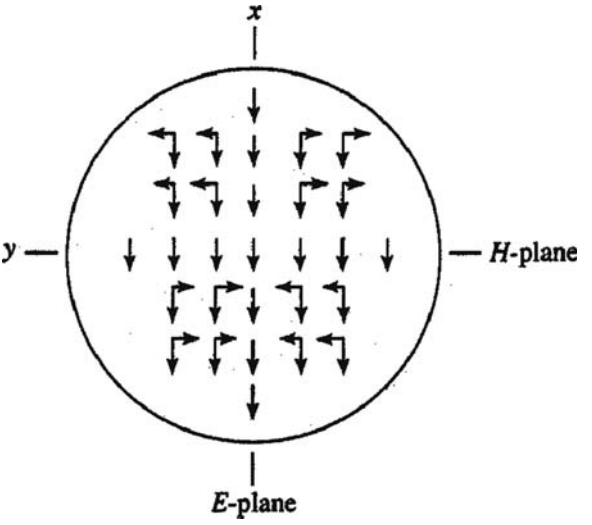


Fig. 6.9. Co-polarization (E-plane) and cross polarization (H-plane) of a paraboloidal reflector antenna.

co-polarization [Figure 6.10(b)]. The cross-polarization sidelobe level of a paraboloidal reflector is related to its focal ratio. It is more serious for smaller focal ratios. If the focal ratio is 0.25, the cross-polarization is at −16 dB. If it is 0.6, the cross-polarization becomes −28 dB (Figure 6.11). If the feed has a positional error in the radial direction, the symmetry of aperture illumination distorts and the cross-polarization becomes more serious. This coma sidelobe level due to the feed radial displacement increases very slowly for a large focal ratio.

The polarization efficiency is the output to input energy ratio in the co-polarization plane. A radio source usually has random polarization. However, any feed can only respond to radiation of one polarization, so that the highest

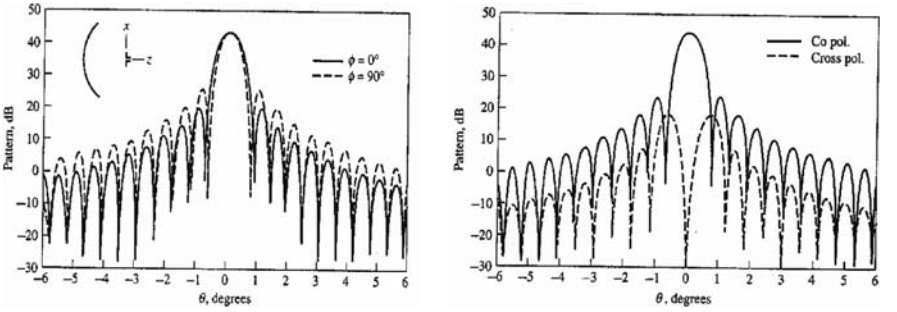


Fig. 6.10. (**a**) Radiation pattern of co-polarization in 0 and 90 degrees planes for a parabolic antenna and (**b**) patterns of co-polarization and cross-polarization in the directions of 45 and 135 degrees planes (Stutzman and Thiele, 1998).

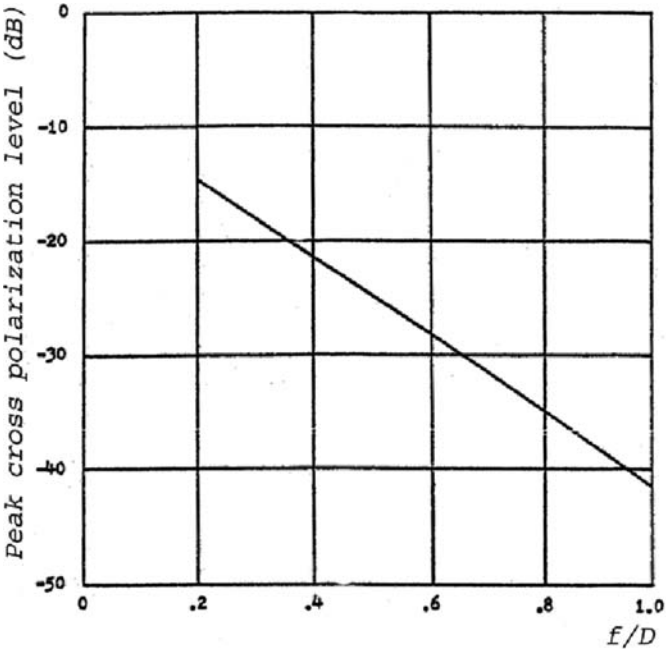


Fig. 6.11. Relationship between cross-polarization level and focal ratio of paraboloidal antennas.

energy efficiency is only 0.5. In most cases, cross-polarization in radio telescopes is harmful. It increases the sidelobe level, decreases the signal to noise ratio, and makes polarimetry more difficult.

6.4.6 Optical Arrangement of Radio Antennas

For radio telescope design, important issues are the selections of the optical system and its basic parameters. Reflector radio telescopes can be a single reflector, a dual reflector system, or a multi-mirror system. A single reflector antenna is a prime focus system and a dual reflector one is either a Gregorian or Cassegrain system. At very high frequency (10^2 GHz), the feed size is small, a Ritchey–Chretien optical system (Section 1.3) with a wider field of view can also be used for multi-feed applications. In the telecommunication field, antennas are often designed with a shaped surface profile to maximize the gain at a specific frequency. However, such antennas have a complex beam shape, higher sidelobe level, and a narrow frequency band. They are not widely used in radio astronomy.

6.4.6.1 Parameter Selection for Parabolic Reflector Antenna

The prime focus system plays a very important role in radio astronomy. In such a system, a feed is located at the prime focus (also called apex) and the mirror is a

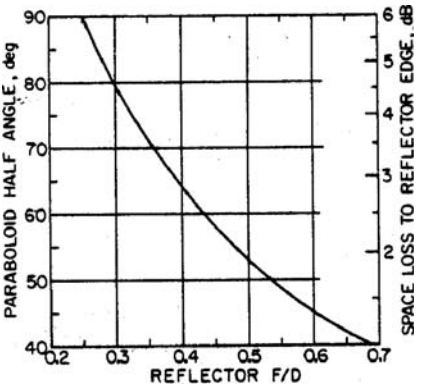


Fig. 6.12. Relationships between half extending angle and reflector f-ratio and between space loss to reflector edge in dB and f-ratio (Meeks, 1976).

paraboloidal reflector. Figure 6.12 shows the half extending angle from the focus to the reflector edge and space loss to reflector edge as functions of the reflector focal ratio. The extending angle decreases while the focal ratio increases. For radio telescopes, the primary focal ratio is between $f/\,0.25$ and $f/\,0.8$. When the focal ratio is $f/\,0.43$, the paraboloidal half angle is $\theta = 60$ degrees and the total feed angle towards the primary mirror is 120 degrees. Because the distance to the focus from the reflector edge is longer than that to the reflector center, there will be space attenuations for radio waves. This attenuation is proportional to $(a\,/\,\rho)^2$, where a is the radius and $\rho\,/\,a$ the normalized radius of the aperture (refer to Figure 6.7). Such attenuation produces edge illumination tapering on the aperture even if the feed pattern is perfectly uniform. This figure shows the space attenuation at the parabolic edge in dB for different focal ratios. If the focal ratio is $f/\,0.43$, the edge space attenuation is -2.5 dB. The attenuation will decrease as the focal ratio increases. The aperture space attenuation plus the nonuniform feed power pattern produces the real aperture illumination of an antenna.

Spillover is another important consideration in the antenna design. Spillover is the unwanted radiation received from side lobes and back lobes. For a single reflector system, the feed can "see" the warm earth surface outside the reflector edge, at a temperature of about 300 K. For a Cassegrain system, the spillover comes mainly from cold sky with a lower temperature. In a single reflector system, the amount of spillover is determined by the focal ratio and the feed design. In general, if the focal ratio is $f\,/\,0.3$, the spillover is about 4%. The spillover increases as the focal ratio increases. If the focal ratio is $f\,/\,1$, the spillover is about 15%.

A flat plate reflects beam at equal angle on the other side of the axis for a feed with an angular displacement, but curved reflector modifies the result slightly. The ratio of the beam maximum angle to the feed rotational angle is named as beam deviation factor(BDF). The BDF for a paraboloidal reflector is (Ruze, 1969):

$$BDF = \frac{\int_0^1 \frac{f(r)r^3 dr}{1+(r/2F)^2}}{\int_0^1 f(r)r^3 dr} \tag{6.29}$$

where $f(r)$ is the aperture illumination function, r the normalized radius, and F the focal length of the paraboloid. Figure 6.13 shows the BDF value for different focal ratios when the illumination edge tapering is -10 dB. As the focal ratio increases, the curvature of the reflector reduces, and the BDF will gradually approach unity. As the focal ratio is 0.43, the BDF = 0.84; and when the focal ratio is 0.8, the BDF = 0.94.

If a feed has a radial displacement, the incoming beam will shift in an opposite direction, changing the beam pattern, reducing the antenna gain, and increasing the sidelobe level. This type of sidelobe near to the electrical axis is named the coma lobe. Figure 6.14(a) gives the relationship between the radial displacement of the feed in HPBW (half power beam width) and the focal ratio when the gain loss is -1 dB. From this figure, it follows that the linear field of view is approximately proportional to the square of the focal ratio and the area field of view is proportional to the 4th power of the focal ratio. Therefore, a large focal ratio antenna has a large field of view and is more suitable for multiple feed applications. The survey range of a f / 0.43 single reflector antenna is approximately ± 4 main lobe widths, while the range of a f / 0.8 one is ± 15 main lobe widths.

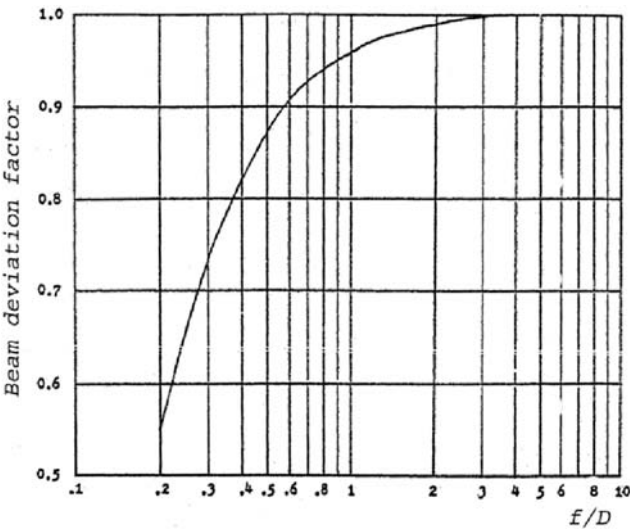


Fig. 6.13. The BDF as a function of focal ratio when illumination tapering at edge is -10 dB.

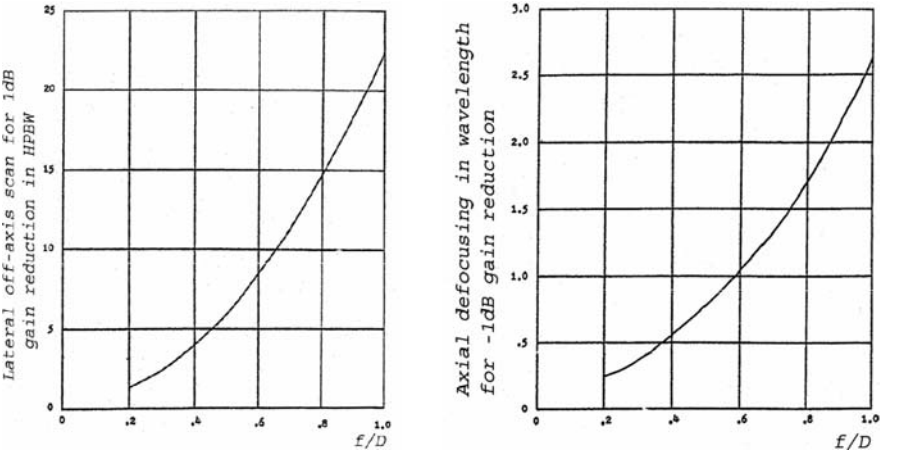


Fig. 6.14. (**a**) Relationship between feed radial displacement and focal ratio, and (**b**) relationship between feed axial displacement and focal ratio while the gain loss is −1 dB (Ulich, 1981).

An axial displacement of the feed also produces gain loss and increases the HPBW and sidelobe level. Figure 6.14(b) gives the relationship between the axial displacement of a feed in wavelength and the focal ratio when the gain loss is −1 dB. In this figure, we assume that the aperture illumination is a quadratic parabolic function. From this figure, one can find that a large focal ratio requires a lower feed positional accuracy.

The deflection of the reflector edge under gravity is also related to the focal ratio as the feed leg weight and moment change for the same antenna aperture size. By increasing the focal ratio, the reflector edge deformation increases. This becomes a problem for medium or large size antenna design. Figure 6.15 shows the relationship between the reflector edge deformations in wavelength and the antenna focal ratio when the gain loss is −1 dB. In homologous design, it is arranged that the paraboloid deformation will mainly lead to a change in focal length and to produce an antenna gain loss unless the feed is moved to refocus the system. For structures with small focal ratios, the gravity induced deformation does not affect the antenna gain. However, when the focal ratio is far more than $f/$ 0.4, then the effect is a constant as large focal ratios are not sensitive to the feed axial shift.

The blockage by a feed is small, only about 0.2–0.3% of the aperture area depending upon wavelength. In short wavelengths, the feed leg (tripod or quadrapod) blockage of about 2–4% is smaller than that of the secondary mirror blockage for a Cassegrain system. The feed leg blockage in longer wavelengths can be serious. Using a simple waveguide type horn, a single reflector antenna can achieve excellent electrical performance in aperture illumination, spillover loss, and others for focal ratios between 0.2 and 1. The polarization of a single

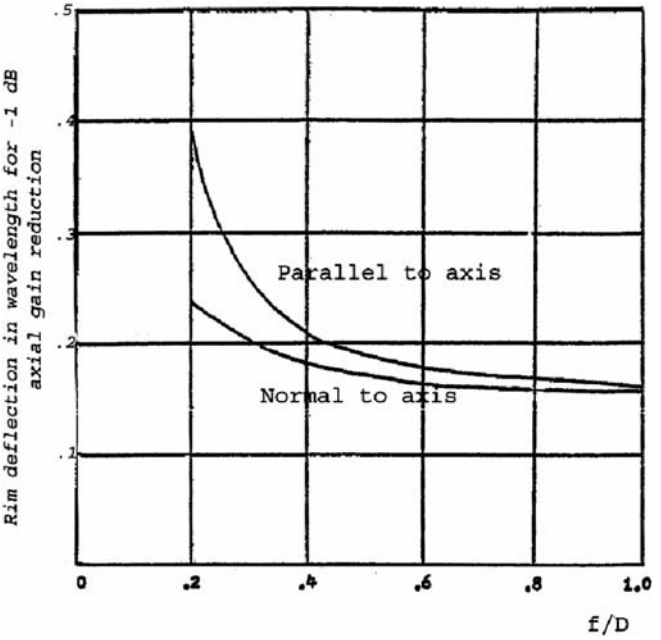


Fig. 6.15. Relation between edge deformation of primary reflector and focal ratio while the gain loss is −1 dB (Ulich, 1981).

reflector system has been discussed earlier. In general, a single paraboloidal reflector with a focal ratio of 0.6–0.8 is a good choice for radio astronomy. This is used for small aperture telescopes. For large aperture telescopes, a smaller focal ratio is preferred as a long feed leg increases the structural cost and deflection. This is also true for antennas within a radome. The radome is a spherical structure with a frame and membranes that allows radio waves in or out. The radome protects the antenna from wind and the Sun, but brings noise and attenuation.

6.4.6.2 Parameters of Cassegrain Antennas

A Cassegrain antenna includes a paraboloidal primary reflector and a hyperbolic subreflector (Figure 6.16). A Cassegrain system can be replaced by an equivalent single paraboloidal reflector system. The diameter of the equivalent system is the same. The focal length of the equivalent system F_e equals a product of the prime focal length F_m and the magnification factor of the secondary mirror m. Generally, there are the following relationships among its parameters:

The tangent of half angle of the primary mirror is:

$$\tan(\phi_v/2) = D_m/(4F_m) \tag{6.30}$$

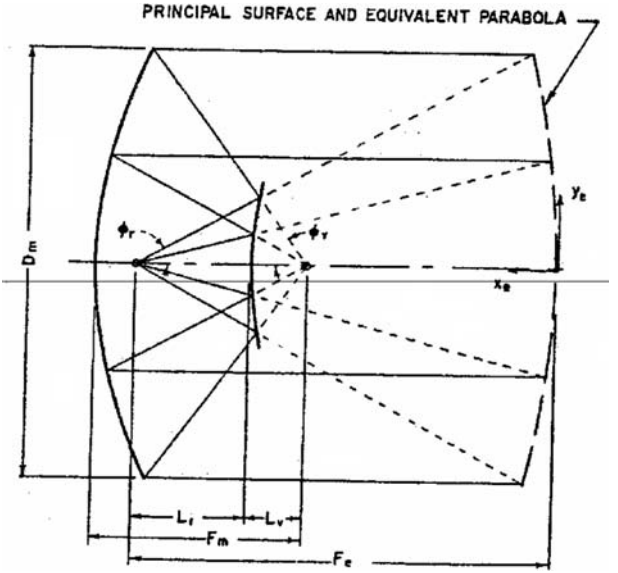


Fig. 6.16. Cassegrain optical system and its equivalent paraboloid.

The tangent of half angle of the equivalent paraboloid is:

$$\tan(\phi_r/2) = D_m/(4F_e) \tag{6.31}$$

The half angle of the secondary mirror is:

$$\frac{1}{\tan\phi_v} + \frac{1}{\tan\phi_r} = 2F_e/D_s \tag{6.32}$$

The eccentricity of the secondary mirror is:

$$e = \frac{\sin[(\phi_v + \phi_r)/2]}{\sin[(\phi_v - \phi_r)/2]} \tag{6.33}$$

The magnification factor of the secondary mirror is:

$$m = \frac{F_e}{F_m} = \frac{\tan(\phi_v/2)}{\tan(\phi_r/2)} = \frac{e+1}{e-1} \tag{6.34}$$

In the above formulas, D_s is the effective diameter of the secondary mirror. In a Cassegrain antenna, there are three basic parameters: the prime focal ratio $f = F_m/D_m$, the diameter of the secondary reflector D_s, and the magnification factor m.

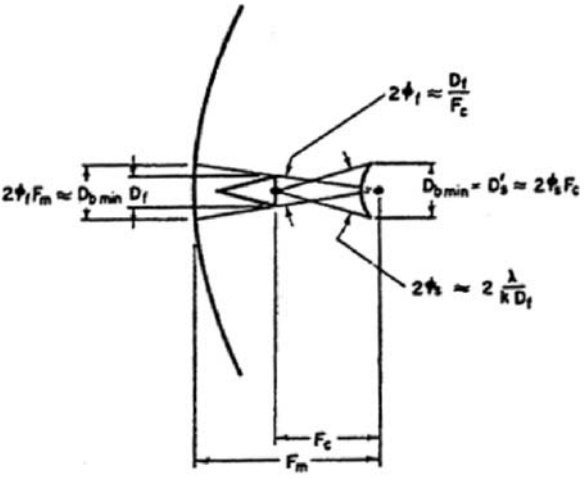


Fig. 6.17. Minimum blockage condition of a Cassegrain antenna (Hannan, 1961).

For radio telescope design, it is desirable to meet the minimum blockage condition. Under such a condition, the blockage of the secondary mirror equals the shadow of the feed on the primary reflector (Figure 6.17):

$$F_c/F_m \approx kD_f^2/(2F_c\lambda) \approx D_f/D_s' \tag{6.35}$$

where D_s' is the actual diameter of the secondary mirror, D_f the actual diameter of the feed, and k the ratio of the effective to the actual diameters of the feed which is usually slightly less than one. From this formula, when the feed is located at the vertex of the paraboloid, the feed size can be equal to the secondary mirror diameter.

For a Cassegrain system, the secondary mirror diameter may be determined by considering the diffraction and the spillover at the mirror edge. The secondary mirror diameter should be at least 10 times as large as the wavelength used. In this way, the diffraction effect is small and geometrical optics can be used in the system analysis. When the secondary mirror diameter is small and the magnification factor is large, a very long feed horn or a feed very close to the secondary is required. This usually is not favorable either for the antenna configuration or for the horn manufacturer. This is also the reason why some existing Cassegrain telescopes do not satisfy the minimum blockage condition. Their secondary mirrors are usually larger and the feed horn is comparatively shorter.

Because of the secondary mirror blockage, the gain reduces and the sidelobe level increases. Figure 6.18 gives the relationship between the gain loss and the central aperture blockage. In the same figure, the relationship between the sidelobe level and the blockage is also provided. If the blockage is 1%, the gain loss is 3% and the sidelobe level increases by 1.5 dB. A large aperture blockage will greatly reduce the telescope performance. In addition to the secondary mirror blockage, the feed leg will also generate blockage. Such feed leg blockage

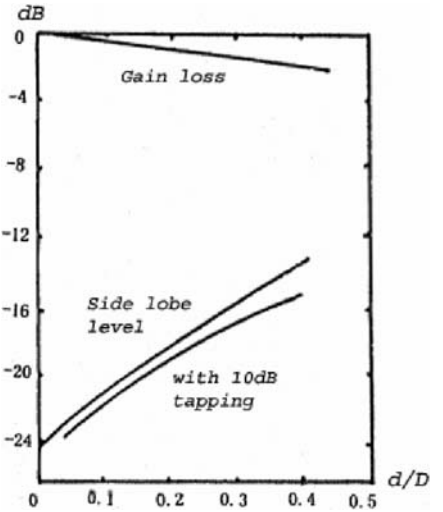


Fig. 6.18. Increase of gain loss and sidelobe level caused by central aperture blockage.

includes two parts: a plane wave one and a spherical wave one. The discussion of the feed leg blockage will be in Section 7.1.6.

For a Cassegrain system, the reflection from the secondary mirror may produce a sine wave stationary ripple in the receiver. This ripple noise will affect the spectrum line observation. Figure 6.19(a) shows the relationship between the antenna temperature fluctuation caused by this stationary ripple effect and

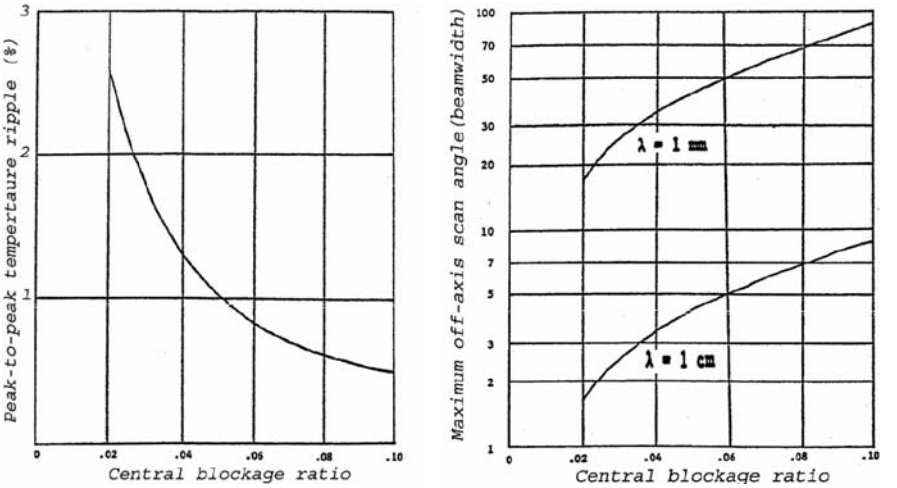


Fig. 6.19. The stationary ripple effect (Peak value percentage) caused by the obstruction of the secondary mirror (**a**) and the maximal scanning angle ($f/0.43$, $d = 0.4\,\mathrm{m}$) determined by the obstruction of the secondary mirror, with a unit of beam width (**b**) (Ulich, 1981).

the secondary mirror central aperture blockage. The curve is for a primary focal ratio of f / 0.43 and a secondary mirror diameter of 0.4 m. The ripple effect will reduce if the secondary mirror diameter increases. Such a ripple effect can be eliminated by a small cone located at the center of the secondary mirror. However, the cone diameter depends on the wavelength involved. A special size cone can only be used for a very narrow frequency range.

The usable field of view of a Cassegrain system is usually restricted by the sidelobe level and spillover power, not by the gain loss. Figure 6.19(b) gives the relationship between the maximal feed scanning angle in the main beam width and the system blockage for a focal ratio of f / 0.43 and a secondary mirror diameter of 0.4 m. The two curves in the figure represent respectively the wavelengths of $\lambda = 1$ mm and $\lambda = 1$ cm. The physical diameter of a multi-feed cluster in a Cassegrain focus is, in fact, independent of the wavelength since the beam width is proportional to the wavelength.

If the wavelength is fixed, the feed scanning angle is nearly proportional to the secondary mirror diameter. Figure 6.20 provides the relationship between the coma sidelobe level in dB at the maximum scanning angle and the central blockage for the wavelength of 1 mm and 1 cm. In this figure, the focal ratio and the secondary mirror diameter used are the same as those used in Figure 6.18. For a large secondary mirror blockage, the diameter of the multi-feed cluster is limited by the coma lobe level.

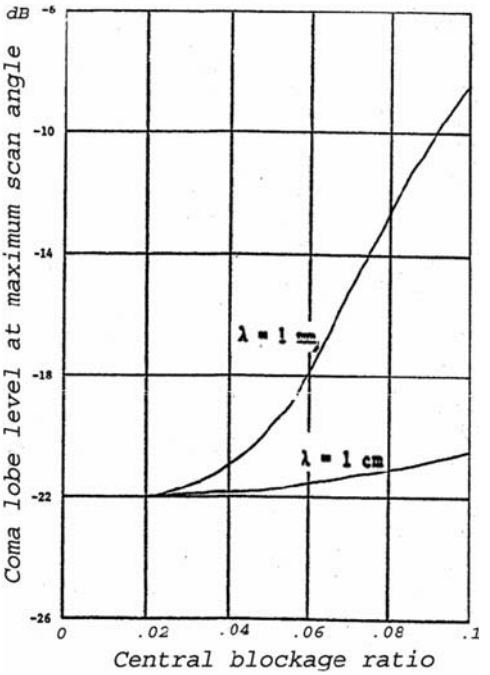


Fig. 6.20. Relation between sidelobe level (dB) at maximal scanning angle and central obstruction for wavelength of 1 mm and 1 cm (Ulich, 1981).

For high-frequency radio wave observations, the sky background is strong. The background noise can be eliminated by switching the secondary mirror back and forth. This will produce a signal of the source without sky background. For the pointing and phase error correction in an array telescope, in order to get large scale information, fast switching of the secondary mirror of some antennas is also required. Because of this, the secondary mirror size and its moment of inertia should be small.

The cross polarization of a Cassegrain system is relatively small. The focus position of the system is easy to access. The structure is compact. Its spillover loss is low and the feed design is easy. Either a high efficiency, very big horn feed or a short small feed equipped with a phase correction lens can be designed for the system. In the design of a Cassegrain radio antenna, attention has to be paid in the selection of the focal ratio, the focal position, and the diameter of the secondary mirror to achieve an optimal performance. The parameter selection is a compromise based on the observational method, the structural design, and the project budget. Table 6.3 is parameter comparison between the prime focus and Cassegrain focus systems.

6.4.7 Characteristics of Offset Antennas

An offset, or off-axis, or asymmetrical, antenna is a special type of antenna where the feed is offset to the side of the primary reflector. The advantage of an offset

Table 6.3. Characteristics comparison between the prime focus and Cassegrain focus antennas

	Prime focus system	Cassegrain system
Prime focal ratio with maximal gain	0.35∼0.55	0.05∼0.40
Major parameters in the system	Primary focal length	Primary focal ratio F secondary diameter d, magnifying ratio m
Difficulties for installing auxiliary instruments	Easy	Difficult
Feed length	Short	m times longer
Aperture blockage	Small	Comparatively large
Noise temperature at zenith	30°K	10°K
Sidelobe level	−25 dB	−20 dB ∼ −17 dB
Gain change caused by axial defocus		feed: m times smaller Secondary mirror: more serious
Coma aberration effect caused by feed radial displacement		m times smaller while at a small angle
Main beam offset caused by the feed radial displacement		m times larger

configuration is that the feed or subreflector does not cast a shadow on the dish. Offset antennas are often referred to as asymmetrical or clean aperture designs. One important issue with the offset antenna is its complex polarization feature.

A prime focus offset antenna is shown in Figure 6.21. The angle between the feed and the reflector axes is θ_0 and the half angle of the feed to the primary reflector is θ_c. In such a system, three coordinate systems, a spherical coordinate $(\hat{\rho}', \hat{\theta}', \hat{\phi}')$ with the origin at the feed center and aligned with the feed axis, a rectangular coordinate $(\hat{x}', \hat{y}', \hat{z}')$ with the origin at the feed center also aligned with the feed axis, and a rectangular coordinate $(\hat{x}, \hat{y}, \hat{z})$ with the origin at the feed center and its axes parallel to the primary reflector axes. The transformations from the unit vectors $(\hat{\rho}', \hat{\theta}', \hat{\phi}') \rightarrow (\hat{x}', \hat{y}', \hat{z}') \rightarrow (\hat{x}, \hat{y}, \hat{z})$ are (Chu & Turrin, 1973):

$$\begin{aligned}
\hat{\rho}' &= \sin\theta' \cos\phi' \cdot \hat{x}' + \sin\theta' \sin\phi' \cdot \hat{y}' + \cos\theta' \cdot \hat{z}' \\
\hat{\theta}' &= \cos\theta' \cos\phi' \cdot \hat{x}' + \cos\theta' \sin\phi' \cdot \hat{y}' - \sin\theta' \cdot \hat{z}' \\
\hat{\phi}' &= -\sin\phi' \cdot \hat{x}' + \cos\phi' \cdot \hat{y}'
\end{aligned} \tag{6.36}$$

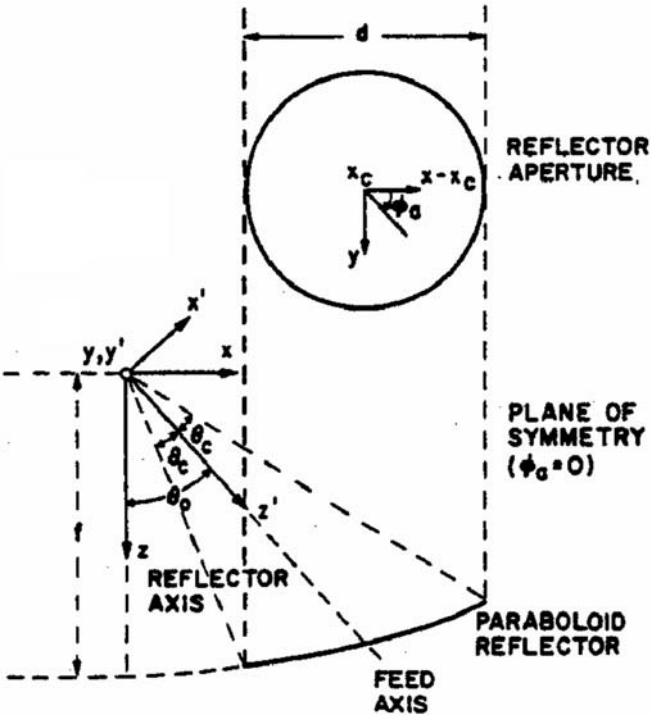


Fig. 6.21. Coordinate systems of the prime focus offset antenna (Chu and Turrin, 1973).

$$\begin{aligned}\hat{x}' &= \sin\theta_0 \cdot \hat{x} - \cos\theta_0 \cdot \hat{z} \\ \hat{y}' &= \hat{y} \\ \hat{z}' &= \sin\theta_0 \cdot \hat{x} + \cos\theta_0 \cdot \hat{z}\end{aligned} \tag{6.37}$$

The position vectors in $(\hat{x}, \hat{y}, \hat{z})$ and $(\hat{x}', \hat{y}', \hat{z}')$ are identical to each other; i.e., $\rho \equiv \rho'$. Representing both sides of the last equation in terms of $(\hat{x}, \hat{y}, \hat{z})$ yields:

$$\begin{aligned}\sin\theta\cos\phi &= \sin\theta'\sin\phi'\cos\theta_0 + \cos\theta'\sin\theta_0 \\ \sin\theta\cos\phi &= \sin\theta'\sin\phi' \\ \cos\theta &= -\sin\theta'\cos\phi'\sin\theta_0 + \cos\theta'\cos\theta_0\end{aligned} \tag{6.38}$$

The radiation far field from the feed can be expressed by:

$$E_f = E'_\theta\hat{\theta}' + E'_\phi\hat{\phi}' \tag{6.39}$$

The reflected field from the primary reflector is:

$$E_r = -E_f + 2\hat{n}(E_f \cdot \hat{n}) \tag{6.40}$$

where $\hat{n}$ is a unit vector normal of the reflector surface. Considering the coordinate transformations of a paraboloidal reflector, there will be:

$$\begin{aligned}E_r = &\frac{\hat{x}}{t}\{[\sin\theta'\sin\theta_0 - \cos\phi'(1+\cos\theta'\cos\theta_0)]E'_\theta + \sin\phi'(\cos\theta' + \\ &+ \cos\theta_0)E'_\phi\} + \frac{\hat{y}}{t}\{-\sin\phi'(\cos\theta' + \cos\theta_0)E'_\theta + [\sin\theta'\sin\theta_0 - \\ &- \cos\phi'(1+\cos\theta'\cos\theta_0)]E'_\phi\}\end{aligned} \tag{6.41}$$

where $\hat{n} = -(\hat{\rho} + \hat{z})/(2t)^{1/2}$ and $t = 1 + \cos\theta'\cos\theta_0 - \sin\theta'\sin\theta_0\cos\phi'$ for a paraboloidal surface. The $\hat{z}$ component is absent in the aperture because of the unique focusing property of a paraboloid. However, its currents flow on the reflector surface. We consider a balanced feed radiation in the following form:

$$\begin{aligned}E_{fx} &= \frac{F(\theta', \phi')}{\rho}(\cos\phi' \cdot \hat{\theta}' - \sin\phi' \cdot \hat{\phi}')\exp(-jk\rho) \\ E_{fy} &= \frac{F(\theta', \phi')}{\rho}(\sin\phi' \cdot \hat{\theta}' + \cos\phi' \cdot \hat{\phi}')\exp(-jk\rho)\end{aligned} \tag{6.42}$$

The formulae correspond to two principal linear polarizations. An important special case of the above formulae is for a circularly symmetric feed, where

$F(\theta', \phi')$ is independent of ϕ'. Then, the principle polarization component of the reflected field becomes:

$$M = E_r \cdot \frac{\hat{x}}{\hat{y}} = \frac{F(\theta', \phi')}{t\rho} [\sin\theta' \sin\theta_0 \cos\phi' - \\ - \sin^2\phi'(\cos\theta_0 + \cos\theta') - \cos^2\phi'(1 + \cos\theta_0 \cos\theta')] \tag{6.43}$$

while the cross polarization component is:

$$N = E_{rx} \cdot \hat{y} = -E_{ry} \cdot \hat{x} = -\frac{F(\theta', \phi')}{t\rho} [\sin\theta' \sin\theta_0 \sin\phi' - \\ - \sin\phi' \cos\phi'(1 - \cos\theta_0)(1 - \cos\theta')] \tag{6.44}$$

where $M^2 + N^2 = F^2/\rho^2$ and if N vanishes when $\theta_0 = 0$, then the system is not an offset system.

From the above formulae, the rotation of the polarization vector due to offset in a paraboloidal aperture has the same magnitude and is in the same sense for any orientation of the incident linear polarization. The reflected field of a circular polarized wave will remain circular polarized but in an opposite rotating sense and with a phase shift of $\tan^{-1}(N/M)$. In the feed, radiation is circularly polarized everywhere. No cross polarization will appear in the reflected radiation, but a small beam displacement will occur because of variation in phase shift across the reflector. Figures 6.22(a) and (b) give the detailed values of this kind of cross-polarization for linear excitation and beam displacements of circularly polarized excitation.

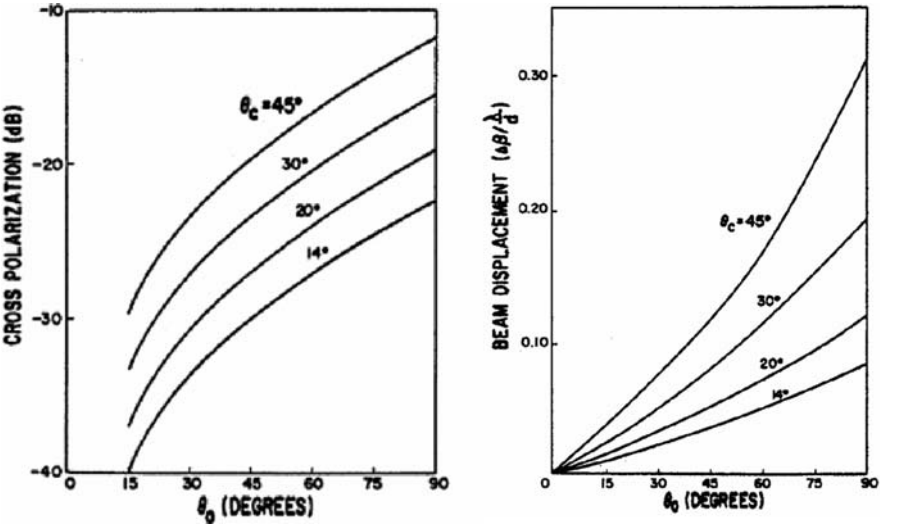


Fig. 6.22. (**a**) Maximum cross polarization (dB) of linearly polarized excitation and (**b**) beam displacement of circularly polarized excitation (Chu and Turrin, 1973).

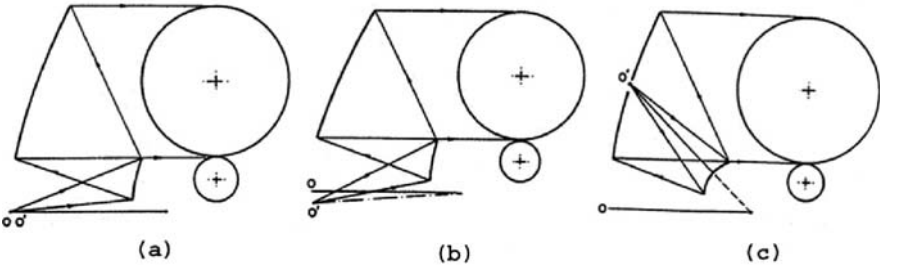


Fig. 6.23. Three configurations of Cassegrain off-set antennas.

Three offset configurations for a dual-reflector antennas are shown in Figure 6.23. The first one [Figure 6.23(a)] has its Cassegrain focus located on the primary reflector axis. The polarization characteristic of this configuration is exactly the same as a primary focus offset one, except that the focal ratio is larger and the influence of cross polarization is smaller.

The second configuration [Figure 6.23(b)] is an optimized Cassegrain system. Its main feature is that the secondary reflector rotates around the prime focus in the direction opposite to the primary reflector with a small angle of β and the feed itself also rotates toward the primary reflector with a small angle of θ_β. The focal length of the equivalent paraboloid system is (Figure 6.24):

$$F_{eq} = F\frac{\left|e^2 - 1\right|}{(e^2+1) - 2e\cos\beta} \tag{6.45}$$

where e is the eccentricity of the subreflector and F the prime focal length. If $\beta = 0$, the above formula is the same as that of a Cassegrain system. However, there exists the following relationship between the tilted angles α and β (Figure 6.24):

$$\tan\frac{\alpha}{2} = \frac{e+1}{e-1}\tan\frac{\beta}{2} \tag{6.46}$$

where the tilted angle of the feed relative to the axis of the equivalent paraboloid is $\theta_\beta - \alpha$. If this angle equals zero, namely $\theta_\beta = \alpha$, there is no tilt between the feed axis and the equivalent paraboloid axis. The equivalent paraboloid system is an axially symmetric one. This system has no serious cross polarization problem. Such an optimized offset system has been used both in astronomy and in communication, such as in the GBT telescope.

The third system [Figure 6.23(c)] is a special offset one, called an open Cassegrain system. In the system, the secondary mirror tilts toward the center of the primary mirror. The tilt angle is opposite in the direction from an

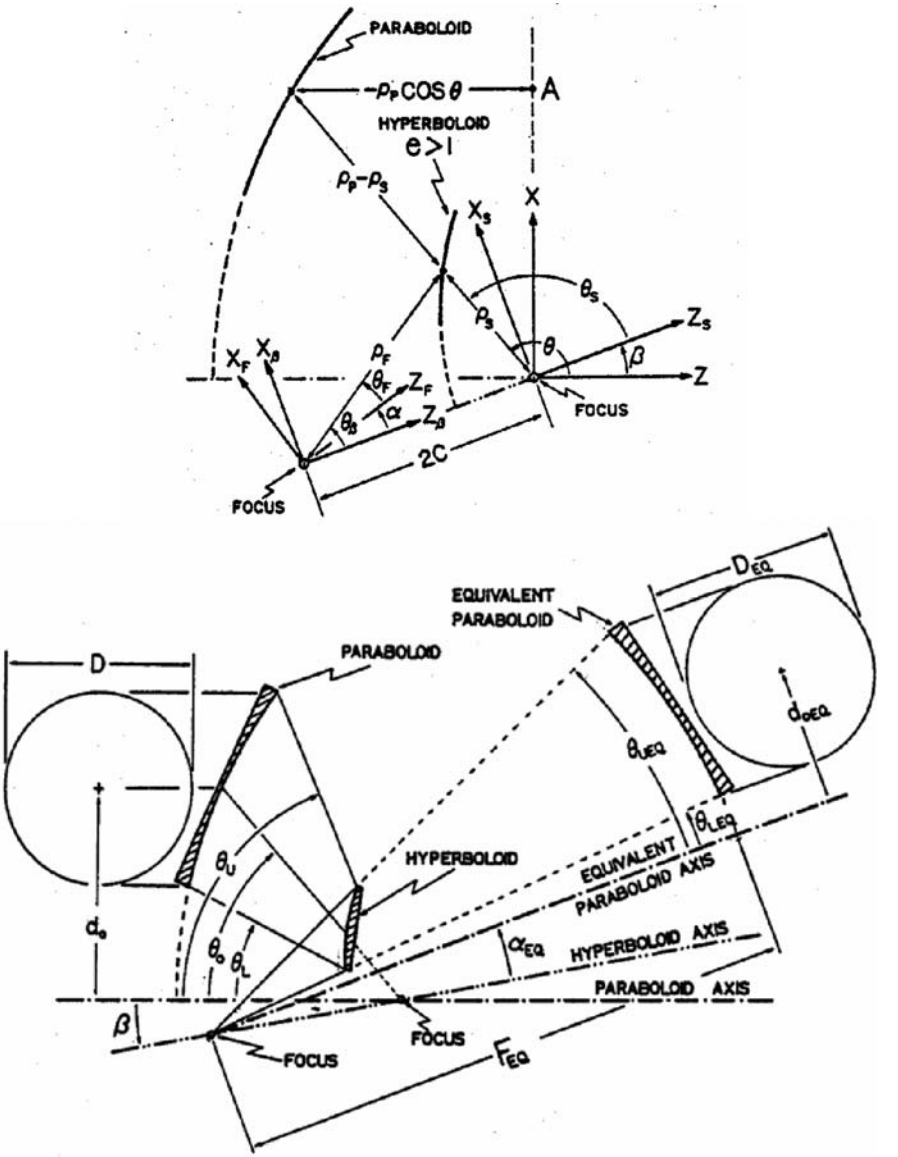


Fig. 6.24. Principle of optimized off-axis radio telescope (*upper*) and the analysis method of equivalent paraboloid (*lower*) (Rusch et al., 1990).

optimized offset system. The cross-polarization of the system is very serious. The asymmetric geometry gives a higher beam separation (squint). Additional reflectors have to be added at the focus to correct the cross polarization of the system.

6.5 Radio Telescope Receivers

A receiver of a radio telescope has two functions: to filter and to detect radio emission from astronomical sources. The first stage of a receiver is named the front end. It is often a low-noise amplifier (LNA) connected to a feed antenna, usually a feed horn. Since the signals received by radio telescopes are very weak, the amplification ratio is large. Therefore, the noise performance of this amplifier is crucial. This leads to extraordinary efforts, such as using very special transistors and cryogenic cooling, to reduce the noise in the LNA.

Generally a radio telescope receiver employs a heterodyne scheme. It mixes the radio frequency (RF) signals from the LNA with a phase-locked local oscillator (LO) reference signal inside a mixer, which is a nonlinear component. The mixer is a frequency down-converter. The local oscillator signal frequency is close to the frequency of the RF signal. The output from the mixer is a lower frequency one, called the intermediate frequency (IF) signal. There exist two reasons for using a mixer in a receiver system. First, it is very difficult to build high quality amplifiers, filters, and other components for higher frequencies. Second, if all the amplification of the signals is done at high frequencies, some of the signal may escape back into the antenna and produce unwanted feedback.

For a nonlinear mixer component, the input and output relation is:

$$F(x) = c_0 + c_1 x + c_2 x^2 + ... \tag{6.47}$$

If two radio signals with different frequencies are input into a nonlinear component, then the output will be:

$$\begin{aligned} F(A\cos(\omega_1 t + \phi_1) + \cos\omega_2 t) &= c_0 + c_1[A\cos(\omega_1 t + \phi_1) + \cos\omega_2 t] + \\ &+ c_2[A\cos(\omega_1 t + \phi_1) + \cos\omega_2 t]^2 + ... \\ &= c_0 + c_1 A\cos(\omega_1 t + \phi_1) + c_1\cos\omega_2 t + c_2 A^2\cos^2(\omega_1 t + \phi_1) + \\ &+ c_2\cos^2\omega_2 t + 2c_2\cos(\omega_1 t + \phi_1)\cos\omega_2 t + ... \end{aligned} \tag{6.48}$$

By ignoring the constant term and the multiples of the input frequencies, the remaining terms are sum and difference terms of the input frequencies:

$$\begin{aligned} 2c_2\cos(\omega_1 t + \phi_1)\sin\omega_2 = c_2\cos[(\omega_1 + \omega_2)t + \phi_1] + \\ + c_2\cos[(\omega_1 - \omega_2)t + \phi_1] \end{aligned} \tag{6.49}$$

If higher frequencies are filtered out, the output from the mixer is the difference between the input frequencies. This is a much lower frequency IF signal, but it retains the same phase information of the original RF signal. For very high frequency observation, two or more stages of LOs and mixers may be necessary and the output from the final stage is the low frequency (LF) one.

After the mixers, signals with low frequencies will further be amplified in the back end of the receiver. Inside the receivers, the signal is usually represented by a voltage proportional to the electric field (as collected by the antenna). For many observations, power or power density, which is the square of voltage, is measured. A device which produces an output proportional to the square of the input voltage is required. This device is called a square-law detector. The square-law detector is formed by a diode circuit loaded with an input signal on a resistor. The diode rectifies the signal and the output is proportional to the square of the input. Some averaging, with either analog or digital electronics, is incorporated into the detector.

The last stage of a receiver is an analog-to-digital converter (ADC). This ADC provides integrated power with as many bits as needed to represent the count over the integration interval. In many cases, the IF signal is digitized without the analog square law detection and the total power is derived from digital hardware or software.

Radio telescopes are, sometimes, also used as radars by receiving the reflected radio waves from planets or other celestial objects. These radio waves are emitted by the telescopes themselves.

References

Barrs, J. W. M., 2007, Paraboloidal reflector antennas in radio astronomy and communication, theory and practice, Astrophysics and space science library, Springer, London, 348.

Bean, B. R., 1962, Proc. IRE, 50, 260–273.

Christiansen, W. N. and Hogbom, J. A., 1985, Radiotelescopes, Cambridge Press, Cambridge.

Chu, T. and Turrin, R. H., 1973, Depolarization properties of offset reflector antennas, IEEE AP-21, 339.

Condon, J. J., 1974, Confusion and flux-density error distributions, Ap J, 188, 279–286.

Condon, J. J., 2002, Continuum 1: general aspects, in Single-dish radio astronomy: techniques and applications, edited by Stanimirovic, S. et al., ASP 278, 155–171.

Cuneo, W. J. Jr.,(ed), 1980, Active optical devices and applications, Proc. SPIE, 228.

Emery, R. J. and Zavody, A. M., 1979, Atmospheric propagation in the frequency range 100–1000 GHz, Radio Electron Eng 49, 370–380.

Findlay, J. W., 1971, Filled-aperture antennas for radio astronomy, Ann. Rev. Astro. Astroph., 9, 272–292.

Hannan, P. W, 1961, Microwave antennas derived from Cassegrain telescope, IRE Trans, AP-9, 140.

Kraus, J. D., 1986, Radio astronomy, Cygnus-Quasar Books, Powell, Ohio.

Lane, A. P., 1998, Submillimeter transmission at South Pole, in Astrophysics from Antarctica, ASP Conf. Proc., eds. G. Novak and R. H. Landsberg, 141, 289.

Meeks, M. L. ed., 1976, Astrophysics, Part C: radio observations, Academic Press, New York.

Pawsey, J. L., Payne-Scott, and McCready, L. L., 1946, Radio frequency energy from the sun, Nature, 157, 158.

Ren, S. Q., 1975, Collection of microwave noise papers, Science Press, Beijing.

Roy, A. E. and Clarde, D., 1982, Astronomy: Principle and practice, 2nd ed. Adam Hilger Ltd, Bristol.
Rudge, A. W. et al., 1982, The handbook of antenna design, Peter Peregrinus Ltd, London.
Rusch, W. V. T. et al., 1990, Derivation and application of the equivalent paraboloid for classical offset Cassegrain and Gregorian antennas, IEEE AP-38, 1141–1149.
Ruze J., 1968, Feed support blockage loss in parabolic antennas, Microwave J, 12, 76–80.
Ruze, J., 1969, Small displacements in parabolic reflectors, Internal report, Lincoln Lab, MIT Press.
Ruze, J., 1965, Lateral-feed displacement in a paraboloid, IEEE Trans, AP-13, 660–665.
Ruze, J., 1966, Antenna tolerance theory – a review, Proc of IEEE, 54, 633.
Stutzman, W. L. and Thiele, G. A., 1998, Antenna theory and design, John Wiley & Sons, Inc., New York.
Ulich, B. L, 1981, Millimeter wave radio telescopes, gain and pointing characteristic, Int. J. Infra Millimeter Waves, 8, 293.
Von Hoerner, S., 1961, Very large antennas for the cosmological problem; I. Basic considerations, Publ. NRAO Greenbank, 1, 19.
Williams, W. F., 1965, High efficiency antenna reflector, Microwave J, 8 79–82.
Zarghamee, M. S., 1967, On antenna tolerance theory, IEEE Trans., AP-15, 777.

Chapter 7
Radio Telescope Design

This is an important chapter concerning radio telescope design. In the first part of this chapter, all the major design issues of radio telescopes are discussed, which include the reflector surface transmission loss, the antenna tolerance theory, the antenna homology design, the antenna surface best fitting, the antenna component positional tolerance, the antenna aperture blockage, the ground radiation pick-up, and antenna best fitting through ray tracing. Theories and formulas are provided so that the readers will have a full understanding of antenna design requirements and they can apply these theories and formulas to their design project. In the second part, different types of radio antenna structure, wind loading on antennas, and active radio telescope structure control are discussed. An overview of typical steerable paraboloidal antenna backup structure design is provided. The active control of a radio telescope involves secondary mirror position control and primary mirror surface control. These control systems are all used only in a few antennas at present. In the last part of the chapter, theory and realization of radio interferometers is discussed together with a number of existing large interferometers. The data calibration in radio interferometers is also discussed so that readers can easily get into this most difficult radio observation field.

7.1 Antenna Tolerance and Homologous Design

7.1.1 Transmission Loss of Electromagnetic Waves

The surface of a filled aperture radio telescope may be continuous, such as the case of a metal or metal-coated surface, or discontinuous, such as a parallel wire, wire grid, or wire mesh surface. The discontinuous reflector surface is generally lighter in weight, lower in cost, and lower in wind resistance. Wind resistance is a major consideration in radio telescope design. The wind resistance of a mesh surface is proportional to the cross sectional area, but not to the surface area. Because of this, reflectors used solely at long wavelengths (>5 cm) are all made

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_7,

of wire mesh surfaces. Wire mesh surfaces are not used for short wavelength operation because the allowable gaps in the reflecting surface are proportional to wavelength. For antennas used in shorter wavelengths, continuous metal surfaces are always used.

To calculate the reflectivity of a metal surface or a mesh surface, a concept of the "intrinsic impedance of a medium" is necessary. The intrinsic impedance is a ratio of the electric to magnetic field amplitudes. In fact, impedance and resistance regard the same concept: the opposition of the flow of electrical field. The resistance is a constant part regardless the field frequency, while the impedance includes the reactance which is purely a frequency dependent part. For a metal surface, it is defined as (Christiansen and Hogbom, 1985):

$$Z_0 = \sqrt{\frac{j\omega\mu}{\sigma + j\omega\varepsilon}} \tag{7.1}$$

where σ is the material conductivity, μ the relative permeability, ε the relative permittivity, $j^2 = -1$, $\omega = 2\pi v$, and v the frequency of the radiation. For a metal surface, $Z_0 = [\omega\mu/2\sigma]^{1/2}\cdot(1+j)$. If the metal is copper, $Z_0 = 2.6\times10^{-7}v^{1/2}\cdot(1+j)$. When the frequency is $v = 10^8$ Hz, then $|Z| = 3.7\times10^{-3}\Omega$. For other metals, the absolute value of the impedance or the resistivity is no more than 100 times that of copper. Therefore, the intrinsic impedance of all metals is less than 0.1Ω at $v = 10^8$ Hz and less than 1Ω at $v = 10^{10}$ Hz.

The intrinsic impedance of free space (vacuum) is $120\pi = 377\Omega$. If Z_{01} and Z_{02} are intrinsic impedances of free space and a metal surface and V_1, V_1', and V_2 are amplitudes of the incident, reflected, and transmitted electromagnetic waves, then:

$$\frac{V_1}{Z_{02} + Z_{01}} = \frac{V_1'}{Z_{02} - Z_{01}} = \frac{V_2}{2Z_{02}} \tag{7.2}$$

where $Z_{01} = 120\pi$ and $Z_{02} = R_{02}(1+j)$. For a copper surface at $v = 10^8$ *Hz*, $Z_{02} = 2.6\times10^{-3}(1+j)$. The transmission loss between the reflected and the incident waves of this surface is very small, only $|V_2|/|V_1| \sim 2\times10^{-5}$ or about 94 dB. The same is true for other metal surfaces. Therefore, the reflector surface of an antenna can be made of any type of metal plates or coated with any metal films. The thickness of the metal coating need be only skin-deep.

Compared with metal surfaces, a wire mesh has a larger transmission loss. Since there is free space after the wire mesh, $Z_{02} = Z_{01}\cdot Z_s/(Z_{01}+Z_s)$, where Z_s is the intrinsic impedance of the wire mesh. The transmission loss of this reflector surface is $V_2/V_1 = 2Z_{02}/(Z_{02}+Z_{01}) = 1/(1+Z_{01}/2Z_s)$. The impedance of a wire mesh includes both reactive (conductive) and resistive parts. The resistive part is small so that the impedance equals approximately the reactive part:

$$X_S \cong \frac{377d\log[d/2\pi r]}{\lambda} \tag{7.3}$$

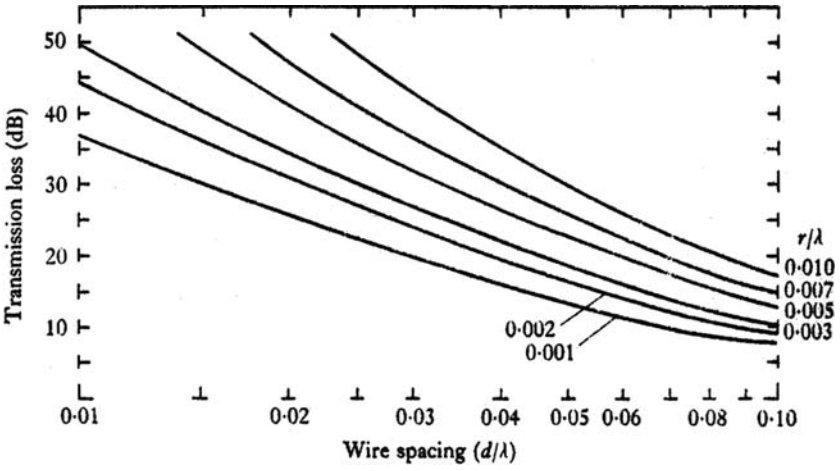


Fig. 7.1. Transmission loss of a wire mesh net (Christiansen and Hogbom, 1985).

where r and d are the radius of and the spacing between the wires. By using this formula, the transmission loss of various wire meshes can be calculated. The calculated results are shown in Figure 7.1.

For a copper wire mesh with a wire radius of 1 mm and a spacing of 10 mm, the reactive part of the impedance is 6Ω while the resistive part is only 0.013Ω for a wavelength of 30 cm. The transmission loss is $V_2/V_1 \sim j/30$ or about 30 dB. Even for a stainless steel wire mesh which has a resistive impedance 1,000 times that of copper, the transmission loss is still quite small. However, the transmission loss of wire meshes at high frequencies is significant.

For a paraboloid reflector, if the incident beam has an angle with a mesh surface, then the wire spacing used in the above transmission loss formula should be multiplied by a factor of the cosine of the incident angle for deriving the correct surface transmission loss.

7.1.2 Antenna Tolerance Theory

Any deviation of a reflector surface shape away from its ideal shape introduces path length or phase errors in the pupil plane. Because of reflection, the path length errors introduced are twice the effective surface errors or half path length errors. Path length errors in units of wavelengths represent phase errors. The wavefront phase errors affect the radiation beam pattern in two major aspects: they produce a gain loss and increase the sidelobe level.

From a Kirchhoff surface integral, the antenna gain on a particular direction is expressed as (Zarghamee, 1967):

$$G(\theta,\phi) = \frac{4\pi}{\lambda^2} \frac{\left| \int_A f(\vec{r}) e^{j\vec{k}\cdot\vec{r}} \cdot e^{j\delta(\vec{r})} ds \right|^2}{\int_A f^2(\vec{r}) ds} \tag{7.4}$$

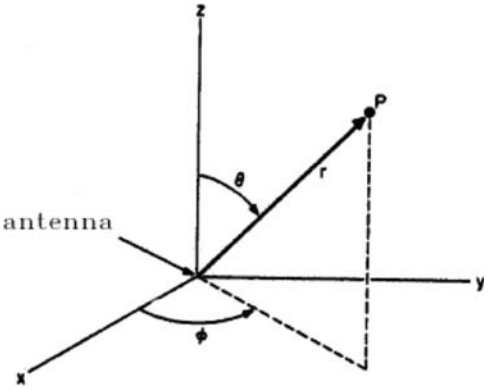


Fig. 7.2. Aperture position and observing direction.

where $f(\vec{r})$ is the aperture illumination, $\vec{k} = 2\pi \cdot \vec{p}/\lambda$, $\vec{p}$ a unit vector in the direction of observation, $\vec{r}$ the aperture positional vector, $\delta(\vec{r})$ the phase error on the aperture plane, A the area of the aperture, and (θ, ϕ) the angles defining the direction of observation (Figure 7.2). If the aperture is divided into N sub-apertures and there is no correlation between sub-apertures, the axial field will be a vector sum of individual contributions. It is a simple addition if no phase error exists [Figure 7.3(a)]. If phase error does exist, the amplitude of the field sum is smaller [Figure 7.3(b)].

To evaluate the field sum, the phase of the individual vector is needed. In the case of a Gaussian distributed phase error $\delta(\vec{r})$ which has a zero mean and a standard deviation of $\sigma(\vec{r})$ and the phase errors are so correlated that the phase errors between two points $\vec{r}_1$ and $\vec{r}_2$ are given by:

$$\sigma^2(\vec{r}_1 - \vec{r}_2) = [\sigma^2(\vec{r}_1) + \sigma^2(\vec{r}_2)](1 - e^{-\tau^2/c^2}) \tag{7.5}$$

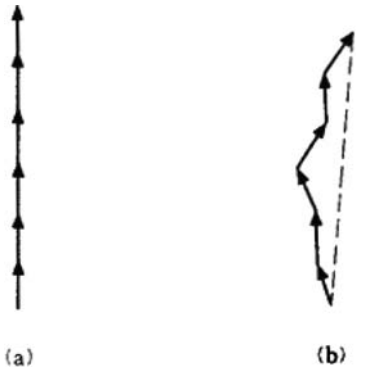


Fig. 7.3. Addition of radiation vectors without (**a**) and with (**b**) phase errors.

where $\tau = |\vec{r}_1 - \vec{r}_2|$ is the distance between two positions and c the correlation radius which is much larger than the wavelength and sufficiently smaller than the aperture dimension, then the expected antenna gain is:

$$G(\theta,\phi) = \frac{4\pi}{\lambda^2}\frac{\left|\int\limits_A f(\vec{r})e^{-\sigma^2}\cdot\sigma^2\cdot e^{j\vec{k}\cdot\vec{r}}ds\right|^2}{\int\limits_A f^2(\vec{r})ds} + \left(\frac{2\pi c}{\lambda}\right)^2\sum_{n=1}^{\infty}\frac{1}{nn!}e^{-(\pi\cdot c\cdot u/\lambda)^2/n}\frac{\int\limits_A f^2(\vec{r})e^{-\sigma^2}(\sigma^2)^n ds}{\int\limits_A f^2(\vec{r})ds} \tag{7.6}$$

where $u = \sin\theta$. When $\sigma(\vec{r}) = \sigma$ is a constant over the aperture, the above equation becomes (Ruze, 1966):

$$G(\theta,\phi) = G_0(\theta,\phi)e^{-\sigma^2} + \left(\frac{2\pi c}{\lambda}\right)^2 e^{-\sigma^2}\sum_{n-1}^{\infty}\frac{\sigma^{2n}}{n\cdot n!}e^{-\sigma^2/n} \tag{7.7}$$

This is the Ruze formula which defines the antenna tolerance theory. If the phase error is small, one may drop the last term:

$$G(\theta,\phi) \approx G_0(\theta,\phi)e^{-\sigma^2} \tag{7.8}$$

In antenna design, this simplified formula is used for assessing the gain loss caused by the reflector surface error. From this formula, the antenna efficiency due to the aperture phase error is:

$$\eta \approx e^{-\sigma^2} \tag{7.9}$$

The phase rms error is related to the path length rms error which is twice the effective surface rms error ε:

$$\left(\frac{4\pi\varepsilon}{\lambda}\right)^2 = \sigma^2 \tag{7.10}$$

and the Ruze formula becomes:

$$G = G_0\exp\left[-\left(\frac{4\pi\varepsilon}{\lambda}\right)^2\right] \tag{7.11}$$

Figure 7.4(a) shows the relationship between the effective surface rms error and antenna gain loss. The peak surface error is roughly 3 times the effective surface rms error depending on the error distribution (Section 2.1.1).

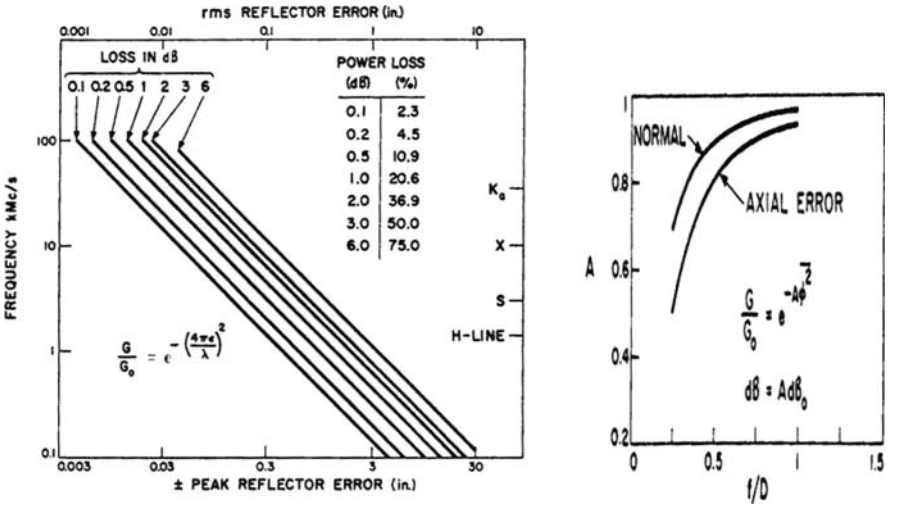


Fig. 7.4. (**a**) Gain loss and (**b**) correction factors caused by the surface deviation of the telescope (Ruze, 1966).

The surface rms error may not be the same as the effective surface rms error. The effective surface rms error is calculated by minimizing the path length error weighted with the illumination edge tapering. If the surface error is that perpendicular to the surface (normal error) or parallel to the electrical axis (axial error), correction factors are necessary in the gain loss formulae. The relationship between these correction factors and the focal ratio is shown in Figure 7.4(b). The formula linking the effective surface error with the axial one Δz and the normal one Δn is:

$$\varepsilon = \frac{\Delta z}{1 + (r/2f)^2} = \frac{\Delta n}{\sqrt{1 + (r/2f)^2}} \tag{7.12}$$

From Equation (7.11), the maximum gain of an antenna is achieved when $\lambda = 4\pi\varepsilon$. For this case, the gain loss is about 4.3 dB. The gain is:

$$G_{\max} \approx \frac{1}{43}\left(\frac{D}{\varepsilon}\right)^2 \tag{7.13}$$

The antenna gain is also related with radius of the correlation regions. Figure 7.5 shows the relationship between the antenna gain loss η_A/η_0, the effective surface error ε/λ, and the radius of the correlation regions r_c.

In the Ruze formula, the assumption of a uniform error distribution, in general, underestimates the gain. However, the error is small for the effective surface

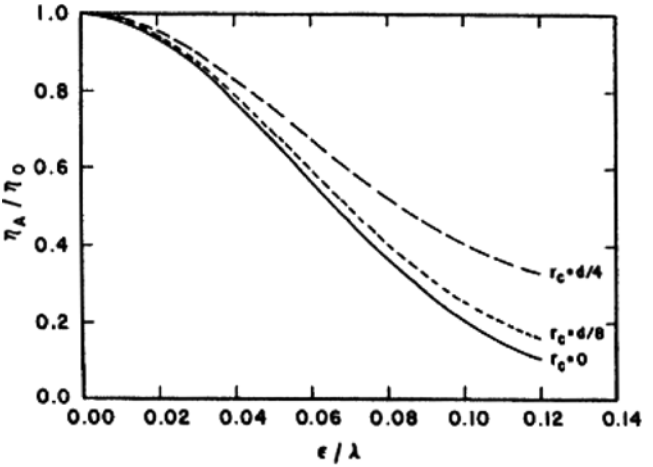


Fig. 7.5. Relationship between gain, surface error, and correlation radius (Ulich, 1976).

deviation of less than about one twentieth of the wavelength, but becomes increasable larger as the surface deviation increases.

When the effective surface rms error over the aperture is not a constant, the gain loss can be calculated only for very few particular cases. If the first and second order variances of the effective surface deviations are:

$$\begin{aligned} \varepsilon_0^2 &= \frac{\int_A \varepsilon^2 f(\vec{r})ds}{\int_A f(\vec{r})ds} \\ \eta_0^4 &= \frac{\int_A (\varepsilon^2 - \varepsilon_0^2)^2 f(\vec{r})ds}{\int_A f(\vec{r})ds} \end{aligned} \tag{7.14}$$

then the gain expression becomes (Zarghamee, 1967):

$$G \approx G_0 \exp[-(4\pi\varepsilon_0/\lambda)^2] \cdot \exp[0.25(4\pi\eta_0/\lambda)^4] \tag{7.15}$$

The above equations indicate that if the surface error is not uniformly distributed, the tolerance loss will be a little larger. However, when the surface error is smaller in the aperture center and larger at the edge, the increase of the gain loss is very small. When the surface error is larger in the center and smaller at the edge, the increase of the gain is larger particularly for larger tolerance cases. The uneven distribution of the surface error also affects the main lobe width. If the surface error is small in the center and large at the edge, the main lobe width decreases.

7.1.3 Antenna Homology

For reflector antennas, the gain loss due to the surface shape deviation is an important design consideration. The surface shape deviation limits the operational frequency range for either prime focus or Cassegrain focus optical systems. Many factors contribute to the reflector surface shape deviation. These include manufacturing error in the backup structure (the backup structure is the structure which supports the reflector surface panels), manufacturing error in the panels, panel adjustment error, and errors caused by gravity, wind, and thermal loadings. Most of these errors are random in nature. However, the surface errors under gravity, static wind, and typical temperature patterns are repeatable and they can be predicted by using finite element analysis.

According to the antenna tolerance theory, Von Hoerner (1967) defined three important natural limits, stress, gravitational, and thermal limits, in antenna structural design. The stress limit is set when the weight of a structure produces a stress at its base which equals the maximum allowed strength of the material. If the maximum allowable stress of a material is $S_{\max}$ and the density is ρ, then the maximum height of a structure made of this material is h_{max}:

$$h_{\max} = K_1(S/\rho) \tag{7.16}$$

The constant K_1 is unity in the case of a straight post. The maximum height for a tall steel structure is about 1,800 m. This is the first stress limit in antenna design.

The second stress limit is set by the maximum deformation caused by the structure's self weight. If the modulus of a material is E and the height is h, then the maximum deformation Δh caused by its self weight is:

$$\Delta h = K_2(h^2\rho/E) \tag{7.17}$$

The deformation caused by self weight is proportional to the square of the height. If the top of a structure is smaller and the base is larger, the deformation reduces. However, this type of structure is not steerable. The second stress limit is much smaller than the first. In fact, the first stress limit is never reached.

For an antenna structure which can be held at two points, if the dimension is D, then the deformation caused by its self weight may determine its operating frequency. Using 1/16th wavelength as the maximum allowable deformation, then the gravitational limit is:

$$\lambda = 5.3K_3(D/10,000)^2(\text{cm}) \tag{7.18}$$

where D is the antenna diameter in cm.

The third natural limit is set by the temperature difference within an antenna structure. If the temperature at one end of a structure is ΔT degrees higher than that at the other end, the dimension of this end will increase by an amount of $\alpha\Delta T$,

where α is the coefficient of thermal expansion of the material. Then the reflector surface error is about $0.03\alpha(\Delta TD/10{,}000)$(cm). For a steel structure with an assumed temperature difference of 5°C, the thermal limit is:

$$\lambda = 2.4(D/10,000)(\text{cm}) \tag{7.19}$$

This is the thermal limit for an open-air antenna made of steel. Figure 7.6 plots these three natural limits in antenna design. The temperature difference within an antenna can be reduced through radome, insulation, ventilation, or using lower thermal expansion materials. Therefore, the thermal limit may change accordingly.

In antenna structural design, another natural limit is from the wind. The wind produces both stress and deformation limits. For the maximum stress case, a survival wind speed is normally set as 44–60 m/s (100–136 mph) depending on site conditions. At the same time, ice, snow, and sandstorm loadings should also be considered. For the wind deformation case, the width of a structure is assumed as $l = D/\sqrt{2}$, where D is the diameter, then the deformation under a wind loading F_{oh} is $\Delta l = (S/E)(F_{oh}/F_{sv})l$, where S is the maximum allowable stress, F_{sv}, the maximum wind loading, and E the Young modulus of the material. For a steel structure, $S/E = 6.7\times10^{-4}$, if we want $\Delta l = \lambda/16$, and using 11 m/s wind as the operating wind limit, then the wind-induced wavelength limit or wind limit is:

$$\lambda = 7.5(D/10,000)(\text{cm}) \tag{7.20}$$

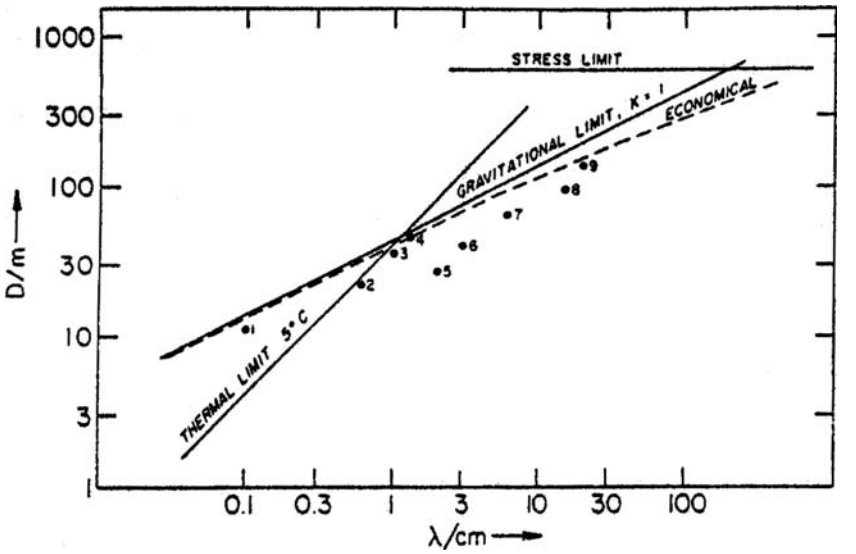


Fig. 7.6. Natural limits of radio antenna design (von Hoerner, 1967).

Comparing this with the gravity limit, the wind limit is more serious even for small size antennas. The wind limit can also be more serious than the thermal one.

Among all natural limits, the gravitational one exists for all antenna structures. It seriously limits antenna aperture size. How to pass this gravitational limit? Generally, there are four possibilities: (a) avoiding the deformation by fixing the antenna in elevation axis (use a fixed mirror or fixed altitude mount (Section 3.1.4); (b) using active surface control; (c) using lever and counterweight mechanism to support the surface panel as in an optical telescope; and (d) using homologous structural design. Using option (a), many fixed mirror or fixed altitude telescopes have been built. Option (b) is used for a few antennas. There is no antenna which uses option (c). However, option (d) is the most attractive way to surpass the gravity limit. The homologous concept is to design a shape of a given type (i.e., a paraboloid) without gravity, but it deforms into another shape of the same type when it is under gravity.

Most large antennas use a truss-type backup structure. How can an antenna truss structure be homologous? First, if a surface is a paraboloid without gravity and it deforms to surfaces of other paraboloids under gravity at just two different elevation angles, then from the superposition law between force and deformation, the surface will remain a paraboloid at any other elevation angle. Second, six points are required to define a paraboloid surface. For an antenna truss structure with S top nodes, a set of $2(S-6)$ conditions has to be satisfied for fitting the nodes on two different paraboloids. Third, topology theory requires that a truss structure with p nodes needs a minimum of $3(p-2)$ members just for its stability. For S top nodes of a truss, there are at least $3(S-2)$ members to connect them. Each member has one degree of freedom for truss adjusting. Since

$$3(S-2)-2(S-6)=S+6>0 \tag{7.21}$$

so more degrees of freedom than the condition required exist in the truss structure. This means the problem is solvable in a mathematical sense and homologous antenna truss structures do exist. This is the homologous theory in radio antenna design.

For achieving a homologous design, a key issue is to determine the deviations between a gravity-deformed antenna surface and the best fit paraboloid. If this deviation is known, the structural optimization can be performed.

The deformation of any structure under an external loading can be expressed in a matrix FEA equation as:

$$[K]\{X\}=\{F\} \tag{7.22}$$

where $[K]$ is the stiffness matrix, $\{X\}$ the displacement matrix, and $\{F\}$ the loading matrix. In antenna structural design, parameters in the stiffness matrix, such as member areas and positions of joints, are optimized so that the surface

node displacements satisfy the homologous condition. The process is a constrained nonlinear optimization problem. The problem is solvable in a mathematical sense.

7.1.4 Antenna Surface Best Fitting

Surface best fitting and surface rms error calculation are important steps in homologous antenna design. For an undeformed surface with a paraboloidal shape, the surface node coordinates should satisfy the following equation (Cheng, 1984):

$$X^2 + Y^2 = 4f(Z + c) \tag{7.23}$$

where f is the focal length and c the vertex z coordinate. Under gravity, the surface nodes will move away from their initial positions. For the new surface node positions, a best fit paraboloid exists. This new paraboloid in a new coordinate system is:

$$X_1^2 + Y_1^2 = 4f_1(Z_1 + c) \tag{7.24}$$

Assuming the new vertex of this best fit paraboloid is (u_a, v_a, w_a) and the focal length change is h. Relative to the initial paraboloid, there are also coordinate rotations of ϕ_x and ϕ_y in x and y directions. The relationship between the initial and the new coordinate systems are (Figure 7.7):

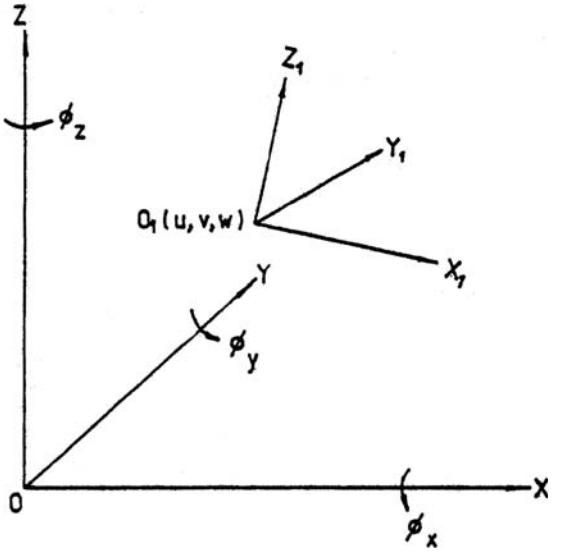


Fig. 7.7. Initial and new coordinator system in surface best fitting.

$$\begin{aligned} X_1 &= (X - u_a) - (Z + c)\phi_y \\ Y_1 &= (Y - v_a) + (Z + c)\phi_x \\ Z_1 &= (Z - w_a) + X\phi_y - Y\phi_x \\ f_1 &= f + h \end{aligned} \tag{7.25}$$

By ignoring higher order terms, the new best fit paraboloid in the initial coordinate system is:

$$\begin{aligned} X^2 + Y^2 + 2(Z + c)Y\phi_x - 2(Z + c)X\phi_y - 2X(u_a + 2f\phi_y) \\ - 2Y(v_a + 2f\phi_x) - 4(Z + c)(f + h) + 4fw_a = 0 \end{aligned} \tag{7.26}$$

At any point on a paraboloid surface, the direction cosines of the surface normal are:

$$\begin{aligned} 2\cos\alpha_1 &= -\frac{X_i}{\sqrt{f(f + Z_i + c)}} \\ 2\cos\alpha_2 &= -\frac{Y_i}{\sqrt{f(f + Z_i + c)}} \\ 2\cos\alpha_3 &= \frac{2f}{\sqrt{f(f + Z_i + c)}} \end{aligned} \tag{7.27}$$

If the displacements of the surface nodes i under gravity is (u_i, v_i, w_i) and the distance to this best fit paraboloid is Δ_i then the following equations exist:

$$\begin{aligned} X - (X_i + u_i) &= \pm\Delta_i \cos\alpha_1 \\ Y - (Y_i + v_i) &= \pm\Delta_i \cos\alpha_2 \\ Z - (Z_i + w_i) &= \pm\Delta_i \cos\alpha_3 \end{aligned} \tag{7.28}$$

In the above expressions, the node is actually on the best fit paraboloid, so that the coordinates also satisfy Equation (7.26). By ignoring the higher order terms and using these two equations, the expression of the distances between the deformed surface points to the best fit surface becomes:

$$\begin{aligned} X_i(u_i - u_a) + Y_i(v_i - v_a) + 2f(w_i - w_a) - 2(Z_i + c)h \\ + Y_i(2f + Z_i + c)\phi_x - X_i(2f + Z_i + c)\phi_y \\ = \pm\frac{\Delta_i}{2\sqrt{f(f + Z_i + c)}}(X_i^2 + Y_i^2 + 4f^2) \\ \Delta_i = \pm\frac{1}{2\sqrt{f(f + Z_i + c)}}[X_i(u_i - u_a) + Y_i(v_i - v_a) \\ + 2f(w_i - w_a) - 2(Z_i + c)h + Y_i(2f + Z_i + c)\phi_x \\ - X_i(2f + Z_i + c)\phi_y] \end{aligned} \tag{7.29}$$

The sum of all distances squared to the best fit paraboloid is:

$$G = \sum_{i=1}^{N} \Delta_i^2 \tag{7.30}$$

By using the least square optimization, the minimum distance solution to the best fit paraboloid is derived from the equations as:

$$\frac{\partial G}{\partial u_a} = \frac{\partial G}{\partial v_a} = \frac{\partial G}{\partial w_a} = \frac{\partial G}{\partial \phi_x} = \frac{\partial G}{\partial \phi_y} = \frac{\partial G}{\partial h} = 0 \tag{7.31}$$

This provides the following matrix equations:

$$\left[\sum_{i}^{N} \frac{1}{2f(f+Z_i+c)}[A]\right]\{x\} = \left\{\sum_{i}^{N} \frac{X_i u_i + X_i v_i - 2f w_i}{2f(f+Z_i+c)}\{B\}\right\} \tag{7.32}$$

$$[A] = \begin{bmatrix} X_i^2 & X_i Y_i & -2fX_i \\ X_i Y_i & Y_i^2 & -2fY_i \\ X_i(Z_i+c) & fY & -2f^2 \\ X_i(Z_i+c) & Y_i(Z_i+c) & -2f(Z_i+c) \\ X_i Y_i(2f+Z_i+c) & Y_i^2(2f+Z_i+c) & -2fY_i(2f+Z_i+c) \\ X_i^2(2f+Z_i+c) & X_i Y_i(2f+Z_i+c) & -2fX_i(2f+Z_i+c) \end{bmatrix}$$

$$\begin{bmatrix} 2X_i(Z_i+c) & -X_i Y_i(2f+Z_i+c) & X_i^2(2f+Z_i+c) \\ 2Y_i(Z_i+c) & -Y_i^2(2f+Z_i+c) & X_i Y_i(2f+Z_i+c) \\ 2f(Z_i+c) & -fY_i(2f+Z_i+c) & fX_i(2f+Z_i+c) \\ 2(Z_i+c)^2 & -Y_i(Z_i+c)(2f+Z_i+c) & X_i(Z_i+c)(2f+Z_i+c) \\ 2Y_i(Z_i+c)(2f+Z_i+c) & -Y_i^2(2f+Z_i+c)^2 & X_i Y_i(2f+Z_i+c)^2 \\ 2X_i(Z_i+c)(2f+Z_i+c) & -X_i Y_i(2f+Z_i+c)^2 & X_i^2(2f+Z_i+c)^2 \end{bmatrix}$$

$$\{x\}^{-1} = \{\, u_a \quad v_a \quad w_a \quad h \quad \phi_x \quad \phi_y \,\}$$

$$\{B\}^{-1} = \{\, X_i \quad Y_i \quad f \quad Z_i + c \quad Y_i(2f+Z_i+c) \quad X_i(2f+Z_i+c) \,\} \tag{7.33}$$

The geometrical rms distance surface error is:

$$\sqrt{\frac{G}{N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \Delta_i^2} \tag{7.34}$$

The geometrical rms distance surface error is not the effective rms surface error. The effective surface error should be the half path length error. The

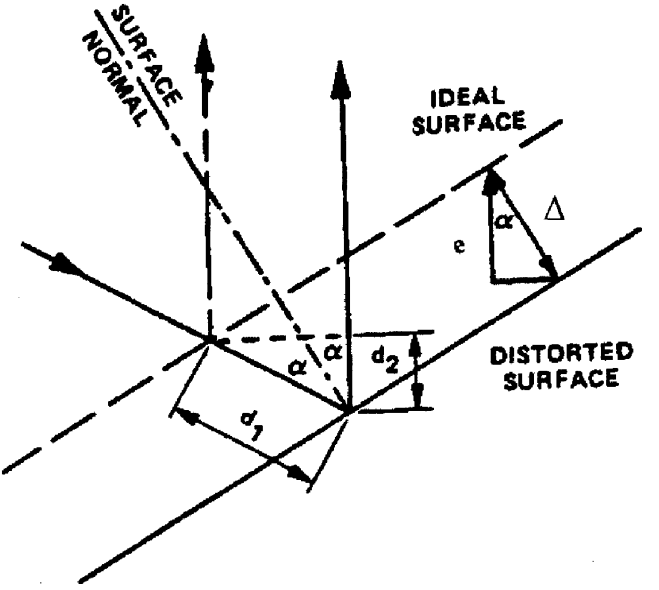


Fig. 7.8. Half path length error and minimum distance error.

relationship between the minimum distance error and the minimum half-path length error is (Figure 7.8):

$$e = \Delta \cos\alpha = d_1 \cos^2\alpha = \frac{1}{2}(d_1 + d_2) \tag{7.35}$$

where α is the angle between the paraboloidal axis and the surface normal. Since the illumination taper exists on the aperture, a weighting factor of illumination should be added when the rms half path length error is calculated.

7.1.5 Positional Tolerances of Antenna Reflector and Feed

The radiation pattern of an antenna will change when the position of its reflectors or feed changes. This positional change also affects the antenna gain and its pointing. On the other hand, the gain loss and antenna pointing error also set the positional specifications for the reflectors and feed of an antenna.

If the wavefront phase error is $\delta(r, \varphi)$, the gain of a circular aperture antenna is (Equation 7.4):

$$G = \frac{4\pi}{\lambda} \frac{\left| \int_0^{2\pi} \int_0^1 f(r,\varphi) e^{j\delta(r,\varphi)} r dr d\varphi \right|^2}{\int_0^{2\pi} \int_0^1 f^2(r,\varphi) r dr d\varphi} \tag{7.36}$$

where r is the normalized radius and $f(r, \varphi)$ the aperture illumination. When the phase error is small, the relative gain is:

$$\begin{aligned} \frac{G}{G_0} &\cong 1 - \overline{\delta^2} + \bar{\delta}^2 \\ \overline{\delta^2} &= \frac{\int_0^{2\pi}\int_0^1 f(r,\varphi)\delta^2(r,\varphi)rdrd\varphi}{\int_0^{2\pi}\int_0^1 f(r,\varphi)rdrd\varphi} \\ \bar{\delta} &= \frac{\int_0^{2\pi}\int_0^1 f(r,\varphi)\delta(r,\varphi)rdrd\varphi}{\int_0^{2\pi}\int_0^1 f(r,\varphi)rdrd\varphi} \end{aligned} \tag{7.37}$$

For calculating the antenna gain loss, it is necessary to know the wavefront phase distribution over the aperture plane. The path length error caused by a small positional change of the reflectors or the feed can be calculated. These path length errors and their normalized values are listed in Table 7.1 (Lamb, 2001):

In the table, F is the primary focal length, r the normalized radius, θ_p the half extending angle between the point on the primary reflector and the prime focus, θ_f the half extending angle between the reflected beam position on the secondary mirror and the feed, m the magnification factor, c - a the distance between the secondary mirror and the prime focus, and c + a the distance between the secondary mirror and the feed, φ the polar angle of a point in the aperture plane, φ_0 the positional angle of a given point in the aperture plane, and $\sin\theta_p = (r/F)/[1+(r/2F)^2]$ (Figure 7.9). If F_{sys} is the system focal length, $F_{sys} = mF$, then $\sin\theta_f = (r/F_{sys})/\,[1+(r/2F_{sys})^2]$. The signs of all parameters are important. A positive sign means an optical path length increase. The positive sign of Δz is in the direction away from the primary reflector. The sign of the rotational angle $\Delta\alpha$ is the same as the vector product of $\vec{z} \times \vec{r}$.

Table 7.1. Optical path length error caused by small positional changes of reflectors or feed (Lamb, 2001)

	Optical path length error	Path length error relative to the aperture center
Feed axial displacement Δz	$-\Delta z \cos\theta_p$	$\Delta z(1-\cos\theta_f)$
Feed lateral displacement Δr	$-\Delta r \sin\theta_f \cos(\varphi-\varphi_0)$	
The rotation of primary reflector $\Delta\alpha$	$F\Delta\alpha[(r/F)+\sin\theta_p]\cos\varphi$	
The axial displacement of the secondary reflector Δz	$\Delta z(\cos\theta_p+\cos\theta_f)$	$\Delta z[(1-\cos\theta_p)+(1-\cos\theta_f)]$
The lateral displacement of the secondary reflector Δx	$-\Delta r[\sin\theta_p-\sin\theta_f]\cos(\varphi-\varphi_0)$	
The rotation of the secondary reflector $\Delta\alpha$	$-(c-a)\Delta\alpha[\sin\theta_p+m\sin\theta_f]\cos(\varphi-\varphi_0)$	

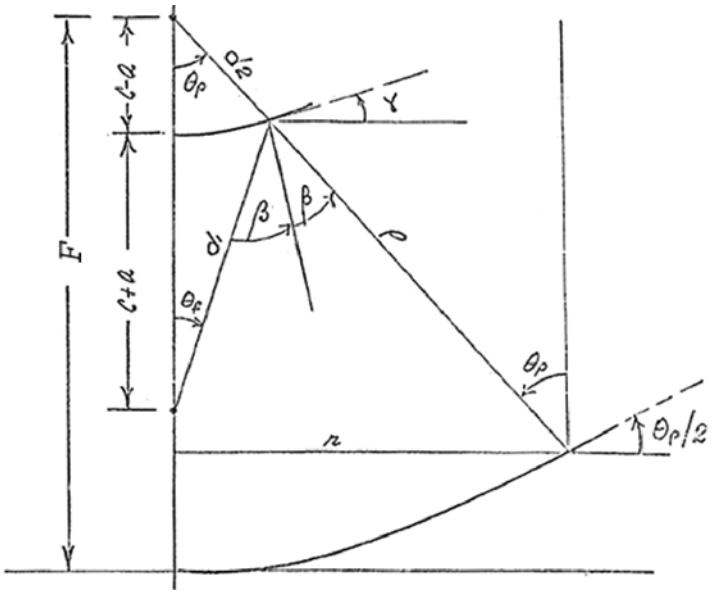


Fig. 7.9. Notations used in path length error calculation.

The calculation of this table is time consuming, but not very difficult. As an example, the optical path length error l for an axial displacement of the secondary reflector Δz is:

$$l = 2\Delta z \cos\left(\frac{\theta_p - \theta_f}{2}\right) \cos\left(\frac{\theta_p + \theta_f}{2}\right) = \Delta z[\cos\theta_p + \cos\theta_f] \qquad (7.38)$$

If the optical path length is normalized by using the path length of the central ray, then the other expression in the table can be derived.

In the path length calculation, it is very useful to draw a simple sign diagram. For a lateral displacement of the feed in x-direction in a prime focus system, the path length on the positive x side becomes smaller, while the path length on the negative x side becomes longer. This sign rule should be kept throughout all the calculations. When the feed has an axial displacement Δz, the path length error is:

$$\varepsilon = \Delta z(1 - \cos\theta_f) = \Delta z \frac{2(r/2F_{sys})^2}{1 + (r/2F_{sys})^2} \qquad (7.39)$$

The radius in the right-hand side of this formula is not normalized. The path length or phase errors are axially symmetric and have high order terms in this case. If the illumination function is $I(r) = 1 - kr^2$, then the gain loss has a simple form:

Table 7.2. The *ALAD* value for different illumination and focal ratio

F_{sys}/D	0.2	0.3	0.5	0.8	1.0	2.0	3.0	5.0
$k=0$	0.14	0.33	0.64	0.84	0.88	0.97	0.99	1.00
$k=0.7$	0.13	0.32	0.58	0.75	0.79	0.86	0.87	0.88
$k=0.9$	0.13	0.30	0.53	0.70	0.75	0.77	0.77	0.77

$$\frac{G}{G_0} = 1 - \frac{\left(\frac{2\pi\Delta z}{\lambda}\right)^2}{3\left(\frac{4F_{sys}}{D}\right)^4} ALAD \tag{7.40}$$

where $ALAD$ is a factor related to the illumination and the system focal ratio as listed in Table 7.2. The gain loss is proportional to the square of the axial displacement and fourth power of the system focal length. If the feed has a radial displacement, its phase error is:

$$\begin{aligned} \delta &= -\frac{2\pi\Delta r}{\lambda}\sin\theta_f \cos(\varphi - \varphi_0) \\ &= -\frac{2\pi\Delta r}{\lambda}\frac{r/F_{sys}}{1+(r/2F_{sys})^2}\cos(\varphi - \varphi_0) \end{aligned} \tag{7.41}$$

This expression is no longer axially symmetrical and it results in wavefront tilts. Using the same illumination function, the gain loss is:

$$\frac{G}{G_0} = 1 - \frac{2\left(\frac{2\pi\Delta x}{\lambda}\right)^2}{\left(\frac{4F_{sys}}{D}\right)^2} ALLD \tag{7.42}$$

where the correction factor $ALLD$ is given in Table 7.3. For more complex subreflector lateral displacements or reflector rotations, the axial loss must be determined from Equation (7.37) by numerical integration.

Table 7.3. The *ALLD* value for different illumination and focal ratio

F_{sys}/D	0.2	0.3	0.5	0.8	1.0	2.0	3.0	5.0
$k=0$	0.27	0.49	0.74	0.88	0.93	0.98	0.99	1.00
$k=0.7$	0.26	0.43	0.62	0.72	0.75	0.78	0.79	0.88
$k=0.9$	0.25	0.40	0.57	0.65	0.67	0.71	0.73	0.77

For calculating the pointing error caused by the positional movement of the reflectors or feed, we can express the path length or phase error terms with reference to a plane which is inclined at an arbitrary angle θ' to the aperture. Inserting the error term into Equation (7.37) and differentiating with respect to θ' one can find the boresight angle θ'_m for the maximum antenna gain. This angle represents the real pointing direction with the reflectors or feed position movements.

Applying this method and ignoring the axial symmetrical terms which do not produce pointing errors, a general path length error expression with respect to an inclined plane is (Ruze, 1969):

$$\varepsilon = [r \sin\theta' \cos\varphi - \sum_n \varepsilon_n \sin\theta_n \cos\varphi] \tag{7.43}$$

where the summation includes several lateral displacement terms of ε_n. As $\bar{\varepsilon} = 0$ and letting $u = \sin\theta'$, we have:

$$\frac{\partial}{\partial u}\int_0^{2\pi}\int_0^1 f(r)[u - \sum_n \frac{\varepsilon_n}{r}\sin\theta_n]^2 \cos^2\varphi \cdot r^3 dr d\varphi = 0 \tag{7.44}$$

with the result:

$$u_m = \sum_n \frac{\varepsilon_n}{f_n} \frac{\int_0^1 \frac{f(r)\cdot r^3 dr}{1+(r/2F_n)^2}}{\int_0^1 f(r)\cdot r^3 dr} = \sum_n \frac{\varepsilon_n}{F_n} BDF_n \tag{7.45}$$

where $u_m = \sin\theta'_m$, F_n is the focal ratio, and BDF_n the beam deviation factor. This formula shows that the pointing error by a number of factors is a sum of contributions from each factor. This is the basis of the pointing error calculation. The beam deviation factor involved is:

$$BDF_n = \frac{\int_0^1 \frac{f(r)\cdot r^3 dr}{1+(r/2F_n)^2}}{\int_0^1 f(r)\cdot r^3 dr} \tag{7.46}$$

In Table 7.4, the beam deviation factor values for the illumination function of $f(r) = 1 - kr^2$ are listed for different f-ratios.

Table 7.4. The beam deviation factor for different illumination and focal ratio

F_{sys}/D	0.2	0.3	0.5	0.8	1.0	2.0	3.0	5.0
$k = 0$	0.51	0.69	0.86	0.94	0.96	0.99	0.99	1.00
$k = 0.9$	0.57	0.74	0.88	0.95	0.97	1.00	1.00	1.00

By using this principle, one set of pointing error formulas is listed in the following equations:

$$\begin{aligned}
&\text{Secondary rotation} && \theta_{hr} = k_1 \tan^{-1}\left(\frac{4cb}{F(m+1)}\right)\\
&\text{Secondary shift} && \theta_{ht} = k_2 \tan^{-1}\left(\frac{h(m-1)}{Fm}\right)\\
&\text{Feed shift} && \theta_{ft} = k_3 \tan^{-1}\left(\frac{ds}{Fm}\right)\\
&\text{Primary rotation} && \theta_{pr} = -(1+k)\gamma\\
&\text{Primary shift} && \theta_{pt} = k \tan^{-1}(\frac{dv}{F})
\end{aligned} \tag{7.47}$$

where F is the primary focal length, $2c$ the distance between the system and the prime focii, m the magnification factor of the secondary mirror, k_i the beam deviation factor of the primary, $k_3 = 1$, b the rotational angle of the secondary mirror, h the displacement of the secondary mirror, ds the displacement of the feed, γ the rotation angle of the primary, and dv the displacement of the primary mirror. In these formulas, the unit of the displacement is the same as the focal length and the unit of the angle is in radians. Figure 7.10 shows the sign convention. In this set of formulas, the optical path length increases in $+x$ direction, and decreases in $-x$ direction. The up-left pointing direction is positive and the up-right direction is negative.

In radio antenna design, the wind-induced pointing errors for some individual terms may be large, but the sum can be small as some component errors have different signs.

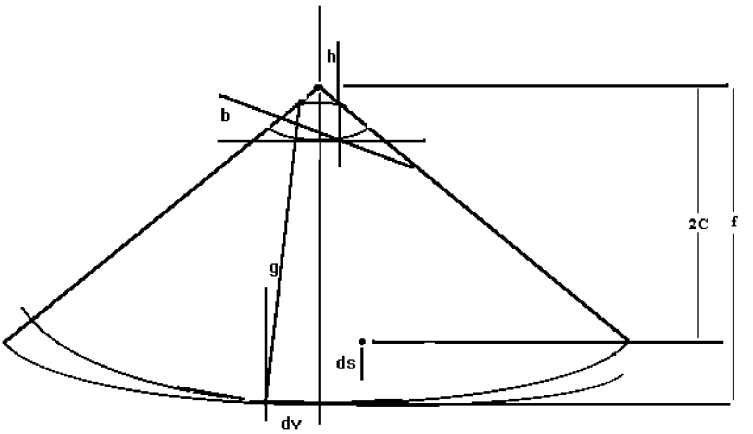


Fig. 7.10. The sign convention in the pointing error formulas.

7.1.6 Aperture Blockage and Ground Radiation Pickup

All axial symmetrical radio antennas have some aperture blockage. In a prime focus system, the feed and feed legs (or feedlegs) produce an aperture blockage and in a Cassegrain system, the blockage is caused by the secondary mirror and feed legs.

Unlike the secondary mirror vane support structure in optical telescopes, the feed legs used in radio telescopes are at an angle with the incoming radiation. Both the incoming and reflected beams will hit the feed legs. Therefore, two blockage effects exist: a plane wave one for the incoming beam and a spherical wave one for the reflected beam from the primary mirror (Figure 7.11). The plane wave blockage is the projected shadow of the feed leg and secondary mirror. The spherical wave blockage is located between the bottom feed leg

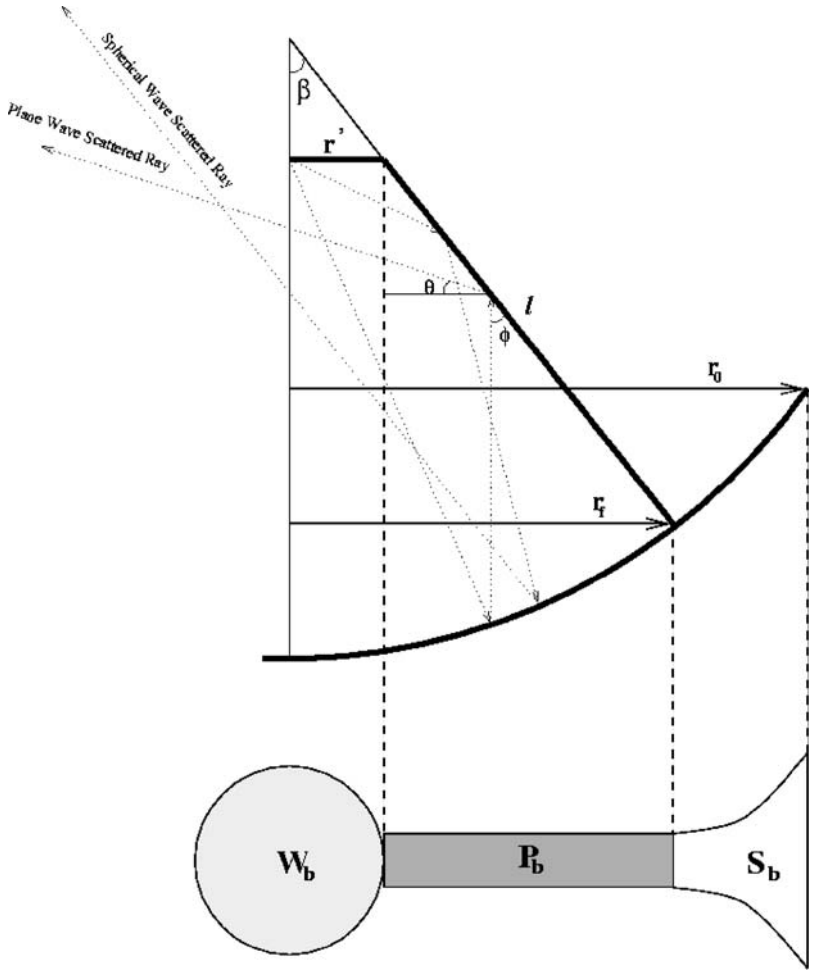


Fig. 7.11. Plane and spherical wave secondary mirror and feed leg aperture blockage (Cheng and Mangum, 1998).

support and the aperture edge. The shadow area is determined by the lines coming from the prime focus to the edges of the dish. Both types of blockage are affected by the aperture illumination (Lamb and Olver, 1986).

By considering two typical illuminations: a uniform one and a tapered Gaussian one, the voltage distribution functions on the aperture are given by (Cheng and Mangum, 1998):

$$\begin{aligned} E_{uniform}(r) &= 1.0 \\ E_{taperedgaussian}(r) &= \exp[-\alpha(r/r_0)^2] \end{aligned} \tag{7.48}$$

where

$$\alpha = (T_e/20)\ln 10 \tag{7.49}$$

with T_e the Gaussian illumination edge taper in dB and r/r_0 the normalized radius. The plane wave blockage from the secondary mirror is given by:

$$\begin{aligned} W_{uniform} &= \pi r_{sub}^2 \\ W_{taperedgaussian} &= \pi r_{sub}^2 \{\frac{1}{\alpha}(1 - \exp[-\alpha(r_{sub}/r_0)^2]\} \end{aligned} \tag{7.50}$$

where r_{sub} is the radius of the secondary mirror and assuming r_t the effective radius for the tapered reflector, and $r_t = r_0/\alpha^{1/2}$.

The plane wave blockage from the feed legs is:

$$\begin{aligned} P_{uniform} &= n_{leg} w(r_f - r_{sub}) \\ P_{taperedgaussian} &= \frac{n_{leg}\sqrt{\pi} w r_t}{2}[erf(r_f/r_t) - erf(r_{sub}/r_t)] \end{aligned} \tag{7.51}$$

where n_{leg} is the number of the feed legs, r_f the bottom radius of the feed leg at the aperture surface, w the width of the feed leg, and the error function used is:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \tag{7.52}$$

The spherical wave blockage from the feed leg is:

$$\begin{aligned} S_{uniform} &= \frac{n_{leg} w}{r'}[r_0^2/2 - r_f^2/2 - fD\tan\beta(r_0 - r_f) + \frac{\tan\beta}{12fD}(r_0^3 - r_f^3)] \\ S_{taperedgaussian} &= \frac{n_{leg} w}{2r'}[r_t^2\{\exp(-r_f^2/r_t^2) - \exp(-r_0^2/r_t^2) \\ &\quad + \frac{\tan\beta}{4fD}[r_f \exp(-r_f^2/r_t^2) - -r_0 \exp(-r_0^2/r_t^2)]\} \\ &\quad + \sqrt{\pi} r_t \tan\beta(fD - \frac{r_t^2}{8fD})[erf(r_f/r_t) - erf(r_0/r_t)]] \end{aligned} \tag{7.53}$$

where D is the primary reflector diameter, f the focal ratio, β the angle between the feed leg and reflector axis, and r' the distance between the primary focus and the feed leg top position in the plane perpendicular to the axis. If there is an illumination taper, the effective area of the antenna is $A_{collect} = \pi r_t^2[1 - \exp[-(r_0/r_t)^2]]$.

In radio antenna design, it is necessary to consider not only the blockage, but also the noise pickup (spillover). In a Cassegrain system, some of the noise reflected by the secondary mirror and feed leg are coming from the warm ground. If the bottom surface of the feed leg is a plane, the radiation picked up by the feed from this surface has an angle $\phi = 2\beta$ with the optical axis. When the angle θ is used to represent the radiation source elevation direction, then $\theta = 90^\circ - 2\beta$. Figure 7.12 shows the relationship between the angle β, angle θ, and the relative radius of the feed leg bottom support R_f, where $R_f = 1$ means a dish edge feed leg support and $R_f = 0$ means dish center feed leg support. As the feed leg support radius increases, the angle β increases from 0° to 65° and the angle θ reduces greatly from 90° to minus numbers. For elevation angles smaller than 60°, warm ground noise pickup is possible. In this analysis, only a single reflection from the feed leg bottom surface is considered.

For understanding the noise pickup from multiple feed leg reflections, a ray tracing program is used. Figure 7.13 shows the results from ray tracing of the feed leg multi-reflections. In the figure, the x axis is a ratio of the ray angle from

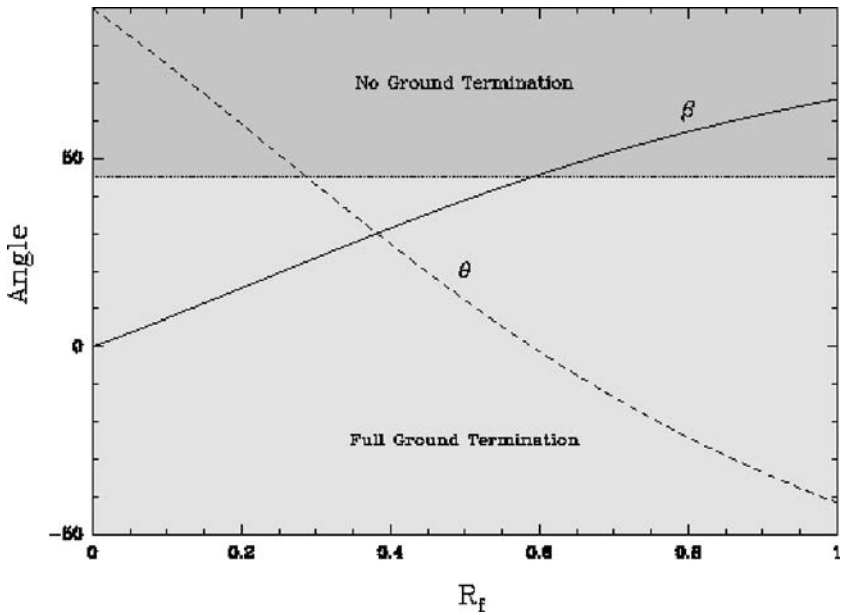


Fig. 7.12. The spillover pickup direction for flat bottom feed leg reflection (Cheng and Mangum, 1998).

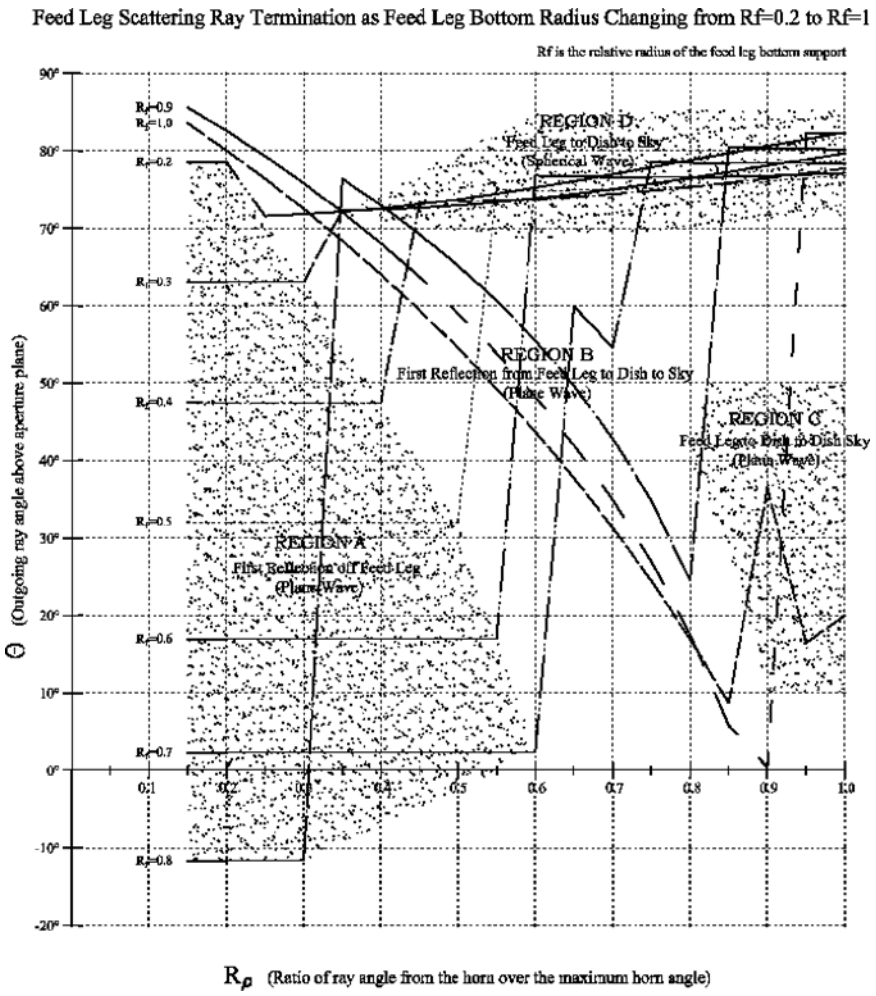


Fig. 7.13. The spillover pickup directions for different feed leg support radii (Cheng and Mangum, 1998).

the axis relative to the maximum horn angle and the y axis is the outgoing ray angle above the aperture plane, representing a spillover pickup direction.

In ray tracing, four distinct regions of the outgoing ray directions (noise pickup directions) are located. In Region A, the ray from the feed is reflected from the secondary and primary mirrors, and hits the feed leg. This is within the plane wave blockage area. The results are a group of parallel lines, which are located at the bottom left side. As the feed leg support radius increases, the angle θ of the outgoing ray becomes smaller and the ray may terminate to a warm ground.

In Region B, the curves run from top left to bottom right. These are also rays of the plane wave blockage but they hit the primary mirror twice. The four

curves of this group start from a feed leg relative radius of $R_f = 0.7$ to $R_f = 1.0$. These rays terminate at high elevation angles with no ground noise pickup.

A small group of curves on the right side is the Region *C*. These are also rays of the plane wave blockage but they hit the primary mirror three times. The rays also terminate to the warm ground.

Curves in Region *D* are all at the top right corner. These are rays of the spherical wave blockage and the noise picked up comes from the cold sky. The ray hits feed leg first after reflected from the secondary mirror. Then it is reflected from the primary reflector.

Generally, the noise picked up by rays in the spherical wave blockage area is very small and that of the plane wave blockage area is serious for a larger feed leg support radius.

The above discussion is based on feed leg beams with a flat bottom surface. If the feed leg bottom surface is not a plane, but a sharp angle of 2α, then the direction of the noise source is:

$$\theta = 90 - 2\sin^{-1}(\sin\beta \cdot \sin\alpha) \tag{7.54}$$

Figure 7.14 shows the noise pickup angles derived from the above formula for different feed leg bottom half angles. When the bottom half angle is small, the noise picked up all comes from the cold sky. This is the same principle as used in stealth airplanes. In this way, a low noise pickup feed leg is produced for any feed leg support radius.

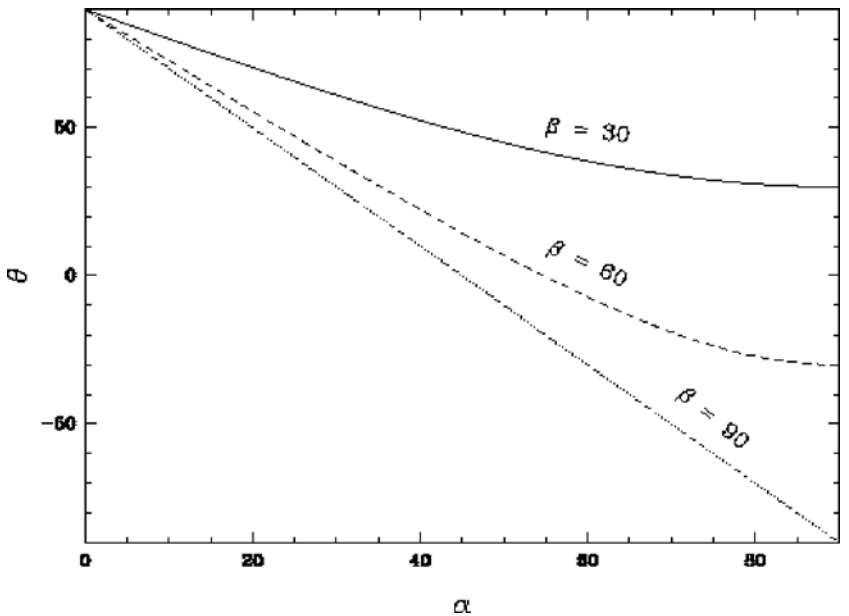


Fig. 7.14. The spillover pickup for different feed leg bottom angles (Cheng and Mangum, 1998).

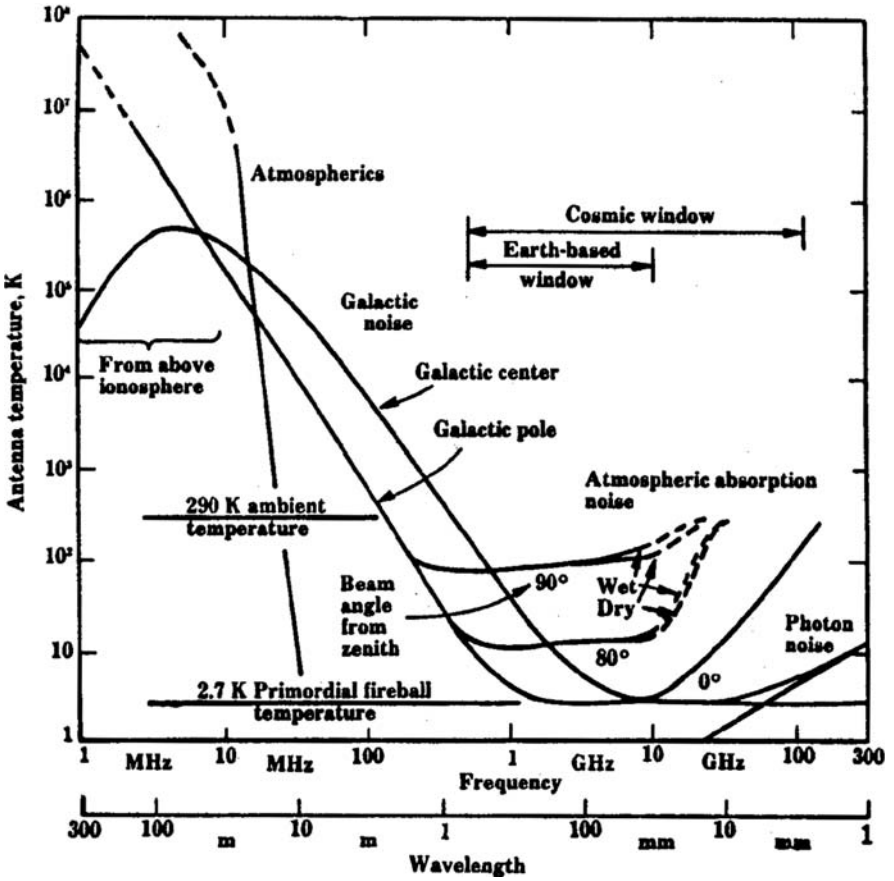


Fig. 7.15. The noise sources and noise temperature (Kraus, 1986).

The sky noise temperature is both frequency and position related. Figure 7.15 shows spectra of some noise sources in the sky. At low frequencies, the noise is mainly from the galaxy center and polar region. When the wavelength is 1 m, the sky noise temperature is a few hundreds of K; when it is 1 cm, the noise is only about 3 K. However, in the millimeter wavelength region, the noise coming from the atmosphere is about a few tens of K. The noise temperature of the ground is nearly a constant, at about 290 K.

7.1.7 Antenna Surface Fitting Through Ray Tracing

Recently, a new antenna surface fitting method had been developed by James Lamb (1998). This surface fitting is based on the ray path length calculation. The basic path length formula for a deformed paraboloid is:

$$\begin{aligned} &Path(x, y, z, \Delta x, \Delta y, \Delta z, \Delta\theta_x, \Delta\theta_y) \\ &= \sqrt{(x-\Delta x)^2 + (y-\Delta y)^2 + (z-f-\Delta z)^2} \\ &\quad + [(-z-f) - \sin(\Delta\theta_x)\cdot y - \sin(\Delta\theta_y)\cdot x] \\ &\quad + \frac{1}{4f}[(2-\cos(\Delta\theta_x) - \cos(\Delta\theta_y))\cdot(x^2+y^2)] \end{aligned} \tag{7.55}$$

where x, y, and z are deformed surface coordinates, Δx, Δy, and Δz are the required focus shift for optimum gain, and $\Delta\theta_x$ and $\Delta\theta_y$ the required pointing change. In the surface fitting, the deformed surface coordinates are input data and the focus shift and pointing change are unknowns before the fitting is done.

In the surface fitting, the ray path lengths are calculated with no focus shift and no pointing change. Therefore, the relative path length function is derived from the difference between the ray path length function and its main value given as:

$$P0 = \text{Path}(x, y, z, 0, 0, 0, 0, 0) - \text{mean}[\text{Path}(x, y, z, 0, 0, 0, 0, 0)] \tag{7.56}$$

The rms wavefront error without best fitting is given by:

$$\text{rmserror0} = \text{stdev}(P0) \tag{7.57}$$

It is easy to find the wavefront tilt in both axes. These tilts are:

$$\begin{aligned} \delta\theta_{x0} &= \text{slope}(y, P0) \\ \delta\theta_{y0} &= \text{slope}(x, P0) \end{aligned} \tag{7.58}$$

If the wavefront tilt is corrected, new ray path length function is derived from:

$$P1 = P0 - \delta\theta_{x0} y - \delta\theta_{y0} x \tag{7.59}$$

where $\delta\theta_{x0}$ and $\delta\theta_{y0}$ are preliminary pointing correction. After the pointing correction, next step is to derive the required focus shift. For solving focus shift required, special sine and cosine functions are required for the angle between the ray towards the primary focus and the dish axis (Figure 7.9). These special functions are as:

$$\begin{aligned} \sin\theta_p(r_p) &= \frac{r_p/f}{1+(r_p/2f)^2} \\ \cos\theta_p(r_p) &= \sqrt{1-\sin\theta_p{}^2(r_p)} \end{aligned} \tag{7.60}$$

To find the optimum focus position of the antenna, we need to move the prime focus in order to maximize the antenna gain. Since this lateral motion removes coma aberrations, we can multiply the wavefront path length by this coma to project out the coefficient from the coma formula defined by Ruze. We also know the pointing changes from the beam deviation factor (BDF) so that we simultaneously remove this pointing term. The coefficients for the lateral shifts of the focus are therefore found from:

$$\Delta x = \frac{\sum_0^K P1_k \cdot \left[\sin\theta_p\left(\sqrt{x_k^2+y_k^2}\right)\cdot\cos[\arctan(y_k/x_k)] - \frac{\mathrm{BDF}\cdot x_k}{f}\right]}{\sum_0^K \left[\sin\theta_p\left(\sqrt{x_k^2+y_k^2}\right)\cdot\cos[\arctan(y_k/x_k)] - \frac{\mathrm{BDF}\cdot x_k}{f}\right]^2} p$$

$$\Delta y = \frac{\sum_0^K P1_k \cdot \left[\sin\theta_p\left(\sqrt{x_k^2+y_k^2}\right)\cdot\sin[\arctan(y_k/x_k)] - \frac{\mathrm{BDF}\cdot y_k}{f}\right]}{\sum_0^K \left[\sin\theta_p\left(\sqrt{x_k^2+y_k^2}\right)\cdot\sin[\arctan(y_k/x_k)] - \frac{\mathrm{BDF}\cdot y_k}{f}\right]^2} \tag{7.61}$$

$$\mathrm{BDF} = \frac{f\sum_0^K \left[\left(\sqrt{x_k^2+y_k^2}\right)\cdot\sin\theta_p\left(\sqrt{x_k^2+y_k^2}\right)\right]}{\sum_0^K (x_k^2+y_k^2)}$$

After the focus shifted, the pointing offset is readjusted by remove the BDF term as:

$$\begin{aligned}\Delta\theta_x &= \delta\theta_{x0} - \frac{\mathrm{BDF}\cdot\Delta y}{f}\\ \Delta\theta_y &= \delta\theta_{y0} - \frac{\mathrm{BDF}\cdot\Delta x}{f}\end{aligned} \tag{7.62}$$

The relative path length function after the pointing offset is given by:

$$P2 = \mathrm{Path}(x,y,z,0,0,0,\Delta\theta_x,\Delta\theta_y) - \mathrm{mean}[\mathrm{Path}(x,y,z,0,0,0,\Delta\theta_x,\Delta\theta_y)] \tag{7.63}$$

The rms path length error after pointing correction is the standard deviation of the relative path lengths:

$$\mathrm{rms\,error2} = \mathrm{stdev}(P2) \tag{7.64}$$

The axial focus shift is derived from mean cosine term minus individual cosine terms as:

$$\Delta z = \frac{\sum_0^K P1_k \cdot \left[z_{avg} - \cos\theta_p \left(\sqrt{x_k^2 + y_k^2} \right) \right]}{\sum_0^K \left[z_{avg} - \cos\theta_p \left(\sqrt{x_k^2 + y_k^2} \right) \right]^2} \tag{7.65}$$

$$z_{\text{avg}} = \text{mean}\left[\cos\theta_p \left(\sqrt{x_k^2 + y_k^2} \right) \right]$$

Now we can use these derived values for the focus and pointing offsets and calculate the rms path length error. This is as:

$$\begin{aligned} P3 &= \text{Path}(x, y, z, \Delta x, \Delta y, \Delta z, \Delta\theta_x, \Delta\theta_y) \\ &- \text{mean}[\text{Path}(x, y, z, \Delta x, \Delta y, \Delta z, \Delta\theta_x, \Delta\theta_y) \\ \text{rms error3} &= \text{stdev}(P3) \end{aligned} \tag{7.66}$$

The rms path length error is the standard deviation of this vector. The difference of it with its mean value is the individual path length error. The antenna surface error is half of this rms path length error divided by the cosine function of the surface normal angle with the antenna axis. In principle, it is necessary to adjust Δ_x, Δ_y, Δ_z, θ_x, and θ_y through iteration to further minimize the residual path length error. However, in practice we have found that this gives no improvement on the above estimates.

7.2 Radio Telescope Structure Design

7.2.1 General Types of Radio Antennas

7.2.1.1 Radio Antennas

In the high frequency region, major radio telescopes are high gain filled (or continuous) or unfilled aperture reflector antennas which collect and focus the radio wave into receiver feeds. Nonreflector antennas usually have low gain and have no physical geometrical aperture area. They are often used at very long wavelengths or as feeds for reflector antennas. Low gain antennas can be added together to form high-gain phased arrays. These phased arrays or interferometers are also used in radio astronomy.

A dipole antenna is made of two line conductors with equal length. Generally, the radiation pattern of a dipole is axially symmetric and lacks a directional mainlobe. The lobe shape is related to the dipole length. The dipole length is twice the conductor length. In the dipole plane, when the length is very small relative to the wavelength, the angle of the HPBW is 90°. As the length increases, the lobe angle decreases. The angle of the HPBW is 78° when the length is half the wavelength (Figure 7.16(a)). The HPBW lobe angle becomes

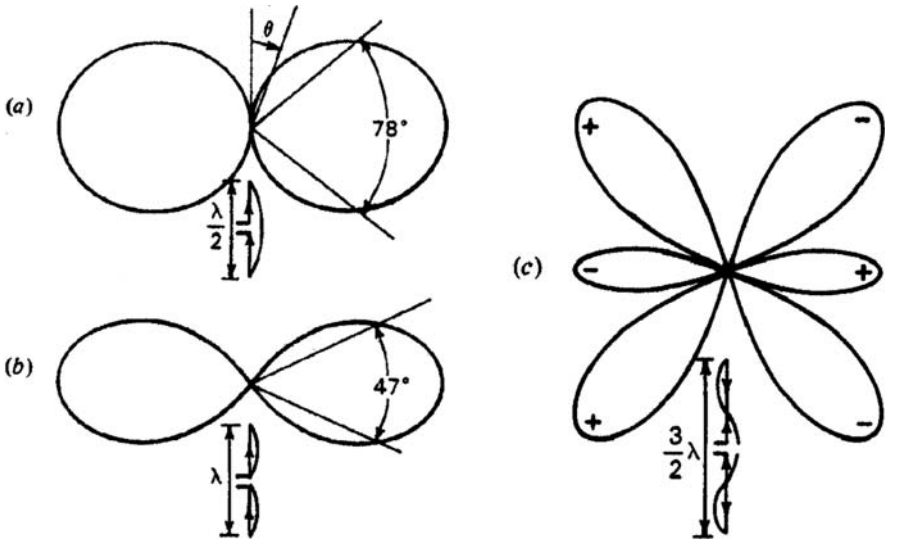


Fig. 7.16. Radiation patterns of dipole antennas with different lengths.

47° if the length is one wavelength (Figure 7.16(b)). However, further increase of the dipole length complicates the radiation pattern and makes the major lobe in the dipole plane disappear. In Figure 7.16(c), the plus and minus signs mean that the phases are 180° apart. This type of pattern is unfavorable for most antenna work. The dipole antenna has a very low gain ($\leq 4.4\,\mathrm{dB}$) and a poor directivity. It can be used as feeds for reflector antennas. To achieve high gain and high directivity, dipole arrays are used. A dipole array is a basic radio telescope type in the low frequency range. If all dipoles are arranged in a plane, a surface array is formed. A surface array has a great directional property (Figure 7.17).

If active dipoles and parasitic (nonactive) elements are used together, a Yagi antenna is formed. The parasitic elements used in Yagi antennas are slightly longer and are served as reflectors. Yagi antennas can also be used together as an array telescope or used as feeds in aperture reflector antennas. Broadband planar antennas are discussed in Section 8.4.4. Other types of nonreflector antennas are introduced in the feed part of this section as they usually are used as feed for reflector antennas.

Many different types of reflecting antennas are used in radio astronomy. From the reflector number involved, antennas can be divided into single reflector, dual reflector, and multi-reflector ones. From the shape of reflector used, antennas include circular, cylindrical, or offset ones. From the sky coverage, antennas can be fully steerable, partly steerable, meridian, or fixed ones. Circular single and dual reflector antennas are the most commonly used in radio astronomy.

Reflector antennas include two major parts: the reflector(s) for collecting and reflecting the radiation and the receiver for detecting the radiation received. The receiver is also known as the feed or feed horn. The size of the primary

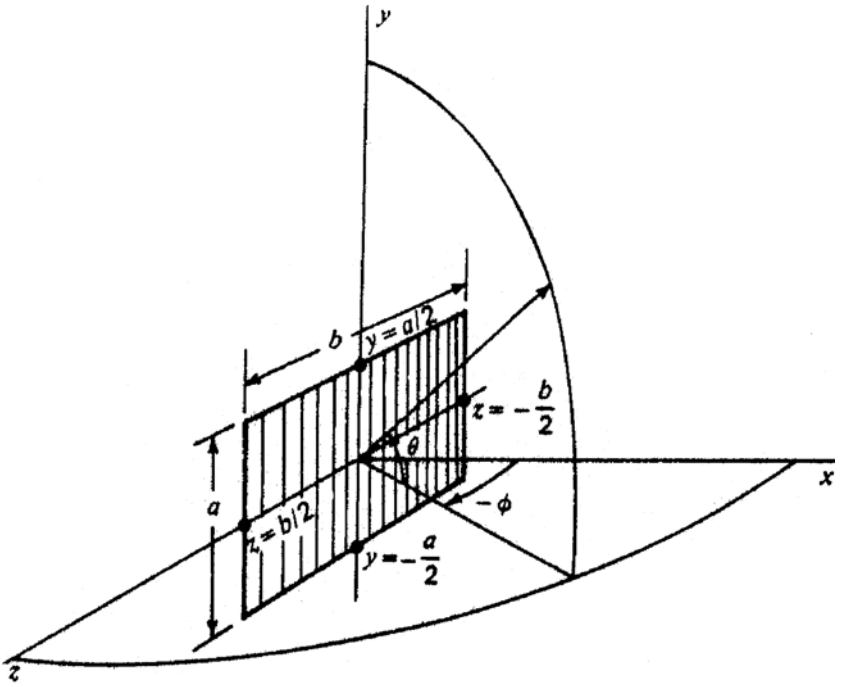


Fig. 7.17. The dipole array radio telescope.

reflector determines the radiation collecting power. When only one reflector surface is involved, the antenna is a prime focus system. When two reflectors are involved, the antenna is a Cassegrain or Gregorian system.

A fixed antenna can also track the star through the motion of its receiver. The tracking range is limited. The largest existing fixed reflector antenna is the 300 m Arecibo one in Puerto Rico. With a spherical primary reflector, it has a field of view of 22° (Figure 7.18). In China, the largest 500 m fixed reflector antenna (FAST) is now under construction. It will have the highest gain and the highest resolution among single dish antennas for the same wavelengths.

Important meridian radio telescopes include the early Kraus telescope in Ohio, US and the RATAN-600 antenna in Russian. The Kraus telescope has a main reflector B and a steerable plane reflector C. The main reflector is a part of a large paraboloid (Figure 7.19). This telescope was built in the 1950s and was demolished in 1998 after a long service time of 40 years. The RATAN-600 telescope has a large ring reflector formed by 995 surface panels with a diameter of 756 m. The angle of these plates can be adjusted to receive radiation from slightly different elevation directions. The ring reflector reflects the radiation through a cylindrical reflector to a feed as in Figure 7.20(a). The RATAN-600 telescope also has other configurations as in Figure 7.20(b) and (c). Since the

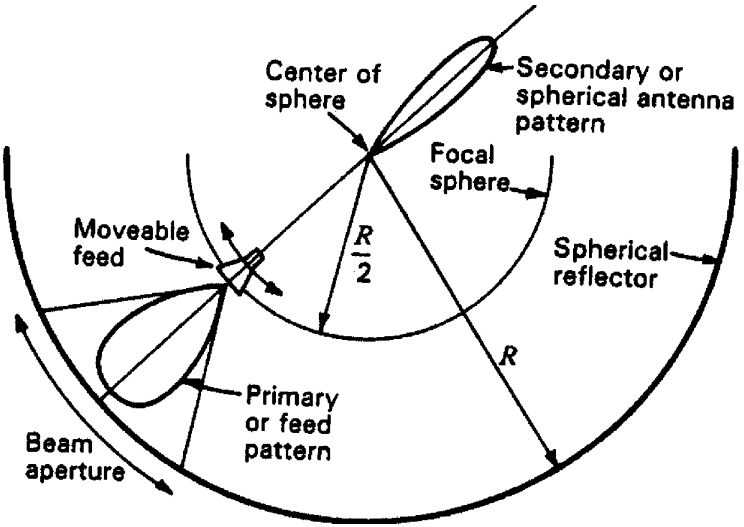


Fig. 7.18. Fixed radio antenna which tracks the sky through the movement of feed.

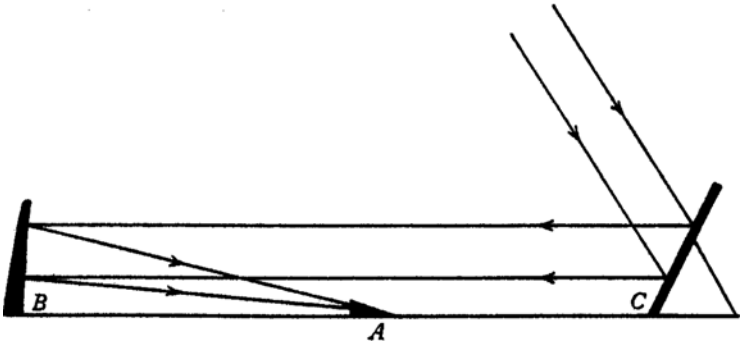


Fig. 7.19. The Kraus radio telescope.

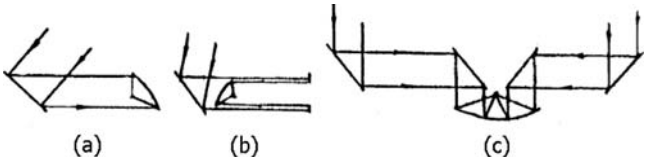


Fig. 7.20. Main configurations of the RATAN-600 radio telescope.

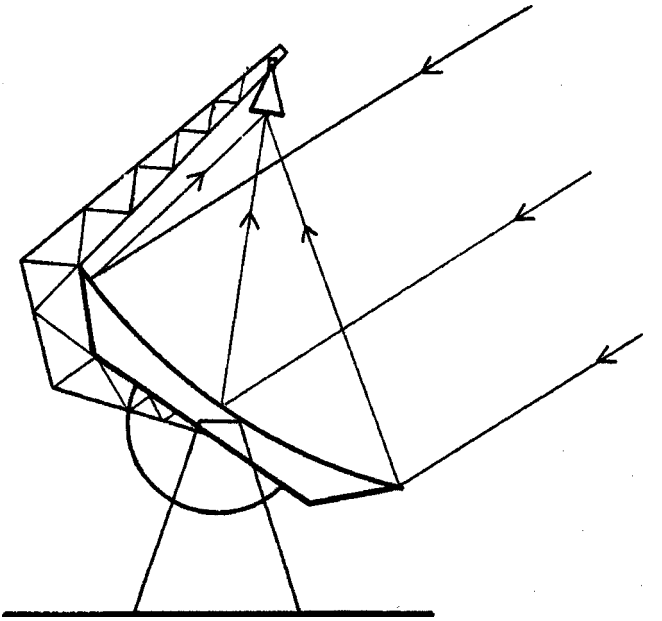


Fig. 7.21. Cross section of a cylinder offset reflector telescope.

dimensions of these telescopes are very large, they have high sensitivity and high resolution in special directions.

Figure 7.21 shows a cross section of a cylindrical surface reflector antenna. If a cylindrical reflector has its axis in north-south direction, then this antenna can track the movement of celestial objects through rotation of the reflector. The observation of a different declination can be achieved by a phase control technique on the receiving dipoles. Such a cylindrical antenna is low in cost. However, the site topology for this telescope has to match the north-south directional requirement.

The optical system of an offset antenna is shown in Figure 7.22. The reflector is an off-axis paraboloid and the feed is away from the incoming beam. An offset design reduces or eliminates the blockage so that the antenna gain increases. However, the polarization is usually complex which limits its usage. The offset prime focus system is rarely used. Dual reflector offset systems with the receiver located on or near the axis of the primary reflector have a smaller polarization effect. The largest offset antenna is the 100 m Green Bank Telescope (GBT) of the National Radio Astronomy Observatory (NRAO).

7.2.1.2 Feed and Feed Horns

The feed element is an actual "antenna" in a reflecting antenna system. The feed interfaces a transmission line or waveguide containing the radio-frequency

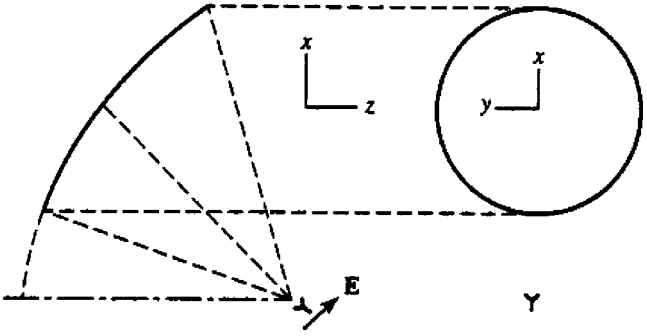


Fig. 7.22. The optical system of an offset telescope.

energy to free space while a reflector is purely a passive focusing surface. The feed can be any one of a multitude of antenna types: a dipole, a dipole array, a helical antenna, a slot antenna, or various horn antennas. At higher frequencies a horn-type feed, or feed horn, becomes more feasible and efficient.

A helical antenna as shown in Figure 7.23 has a wide bandwidth and is circularly polarized. Its radiation pattern is determined by the ratio between its diameter and the wavelength used. The axial radiation occurs at the following frequency range:

$$0.75 < C/\lambda < 1.25 \tag{7.67}$$

where $C = \pi D$. A type of slot antenna shown in Figure 7.24 has two openings in a metal surface which form a magnetic dipole of the horn. The distance between these two openings must be smaller than $\lambda/2$.

Horn antennas include circular, conical, multi-mode, and hybrid mode horns. They have good frequency response and are simple in their structural design. A hybrid mode horn possesses axially symmetric amplitude and phase

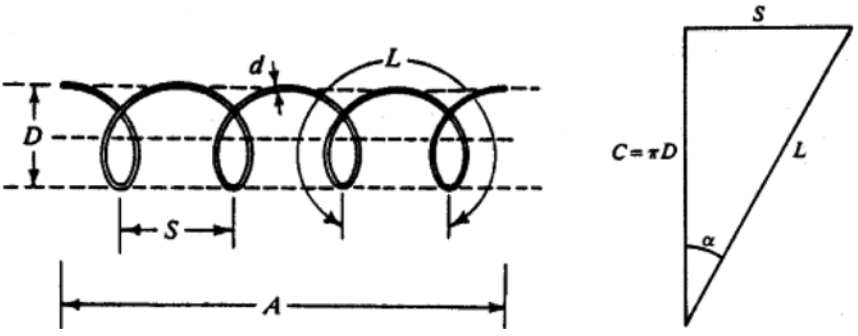


Fig. 7.23. Helical wire antenna and its design parameters.

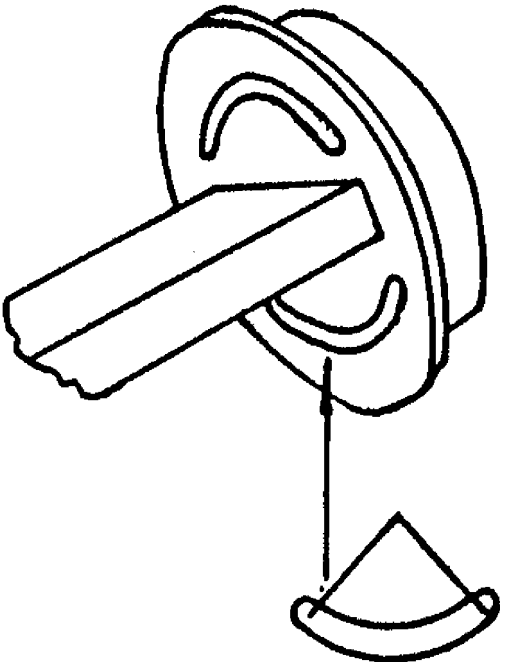


Fig. 7.24. A type of slot antenna used as a feed in a radio telescope.

property. Conical horns have good circular polarization under excitation from the base mode TE_{11}. The *TE* means a transverse electric mode whose electric field vector is normal to the direction of propagation and the Subscripts are mode numbers. A *TM* mode is a transverse magnetic one.

The radiation pattern of these horns is related to the horn aperture diameter. When the horn diameter is larger than the wavelength used, the axially symmetrical radiation pattern may not be retained. The cross-polarization of a horn is also related to its edge shape. These disadvantages can be all eliminated by using a multi-mode horn. In a multi-mode horn, the change of the cross-section shape and the conical angle can excite more transmission modes. The composition of these modes makes the field axially symmetric, so that the gain and the pattern symmetry are improved (Figure 7.25).

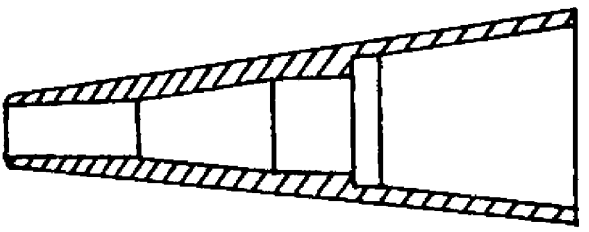


Fig. 7.25. A dual-mode horn antenna.

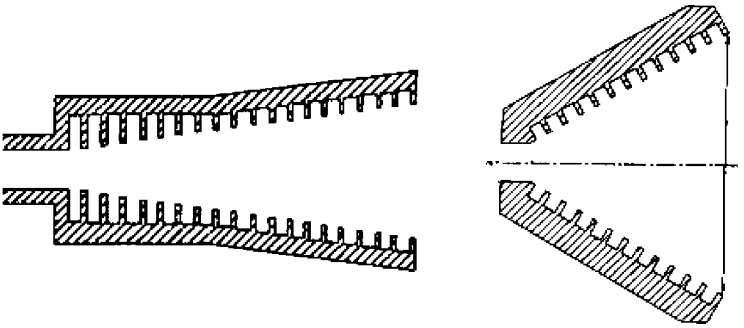


Fig. 7.26. Corrugated horn antennas (**a**) with small angle and (**b**) with wide angle.

The hybrid conical horn or corrugated horn was developed in the 1960s. It is a high efficient feed horn. The performance of the horn is improved by applying a corrugated inner surface to the horn. The boundary condition of this type of horn is different; both *TE* and *TM* modes have the same boundary condition, so that a symmetrical hybrid mode pattern is achieved. The sidelobe level of this horn is very low, therefore, it has been used in many radio telescopes.

There are two types of corrugated horns: small and large conical angle ones as shown in Figure 7.26. The radiation patterns of both types are similar. The horn with a large conical angle has a spherical wavefront at its aperture, which is more suitable for a prime focus system, while that with a small conical angle has a plane wavefront at its aperture, which is more suitable for a large focal ratio system. The horn with a large conical angle is called a scalar horn.

7.2.1.3 Radio Antenna Mountings

Radio telescopes can be in the open air, inside a radome, or inside an astro-dome. A radome protects a telescope from solar and wind effects. However, absorption and membrane radiation produce gain loss and noise. These effects are very detrimental for high frequencies (e.g. millimeter or submillimeter waves). An astro-dome with an observing opening seems to have all the advantages required. However, astro-domes are expansive. Therefore, most radio telescopes are open air ones.

Similar to optical telescopes, radio telescopes can have either an equatorial or alt-azimuth mounting. Radio telescopes involving a large reflector have since the early days more often used an alt-azimuth mounting. Two different alt-azimuth mounting types are used for radio telescopes: king post and wheel-on-track ones. The King post design is used for small and medium-size radio telescopes. The wheel-on-track design is used for large aperture radio telescopes.

In a king post design, the antenna weight is supported by a vertical post or tower structure with an azimuth drive system built inside. A yoke-style structure supports the antenna dish and the elevation axis. The elevation drive gear is

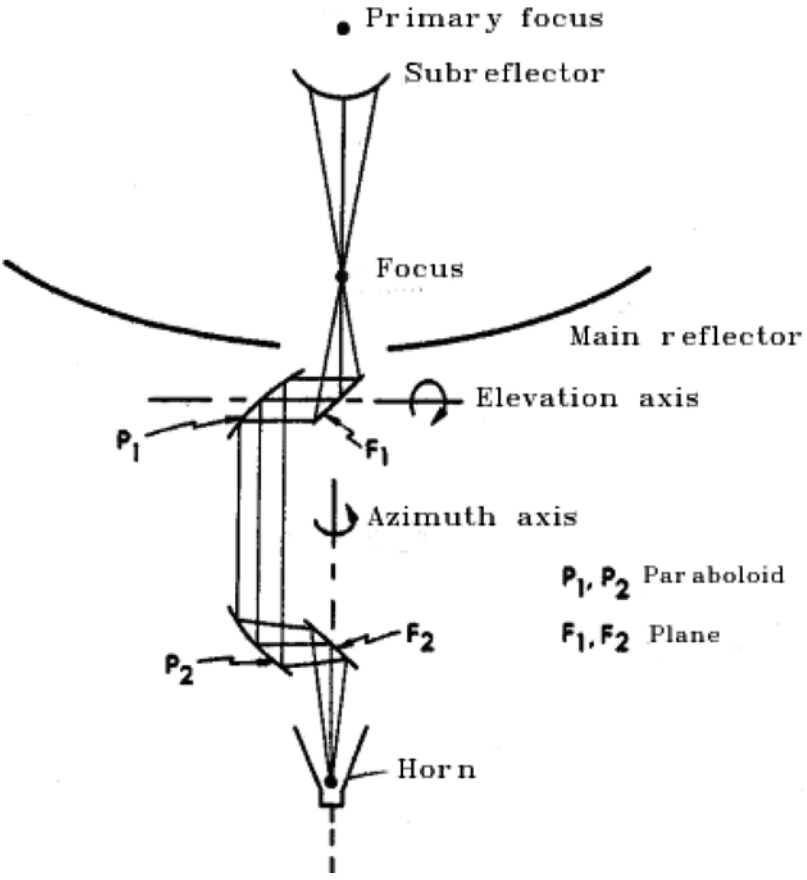


Fig. 7.27. Wave-guide system used in communication antennas.

in the middle of the yoke arms. In a wheel-on-track antenna, a wave guide system can insure a stationary feed position. This wave guide system is widely used in communication antennas. Figure 7.27 shows this type of wave guide system where mirrors transfer the radiation from the Cassegrain focus to a fixed feed center. The wave guide system has high gain loss and adds more noise to the system, so it is not widely used in radio astronomy. Astronomical telescopes use mostly prime or Cassegrain foci as their feed positions.

7.2.2 Steerable Parabolic Antenna Design

The structural design of a steerable parabolic antenna is similar to that of an optical telescope, so that readers should reference related sections in early chapters. A small diameter radio telescope can have a monolithic reflector plate. The dish is made of either a hydroformed aluminum plate or a sandwiched composite surface coated with aluminum film. The dish with a feed at its apex is then

connected with two rotational axes. Hydroforming is a process of forming aluminum plate to a rigid and precise mold by using fluid or gas under pressure. The advantages of the process are high rigidity of the dish shell which requires only simple back support, high accuracy largely determined by the mold, and low cost for both material and labor when large numbers of dishes are required. However, the FEA shows that the deformation of this type of dish under wind loading is proportional to the fourth power of the diameter, so that the diameter of this type of antenna is limited. Another problem with hydroforming is the mold cost which can be as high as $6 M for a 12 m diameter mold (in 2005).

The sandwiched composite design is discussed in Section 8.3.2. Sandwiches of glass fiber plates with a form or rib core are used in many fields. The technique can be used to form a large diameter (15 to 18 m) monolithic dish surface. The mold for it can be made from identical adjustable sections so that the cost is lower. A new surface metallization method by thermal spraying aluminum coating on the top of the mold surface improves manufacturing efficiency. The metal surface is then transferred to the composite surface for reflecting radio waves. The method can be used for both mass production and test experiments.

In general, a medium or large radio antenna includes a backup structure, surface panels, a subreflector, feed legs, a feed, elevation axis and drive, and azimuth axis and drive. The backup structure is usually a truss structure. The primary reflector surface is welded to or supported by the backup structure. For high precision antennas, the surface panels are supported by adjusters, which are spring loaded differential screws. With the adjusters, the panel position can be accurately controlled.

The design of the backup structure is the most important task in radio telescope design. The purpose of the backup structure is to accurately support and maintain the shape of the reflector surface. A typical backup structure is a two-layer space truss structure. The top layer nodes support rings of surface panels. The space between truss layers provides stiffness of the backup structure against gravity. In general, any antenna backup structure retains some homologous properties. For small antennas, the backup structure design is simple. However, when antenna size becomes large and the operation frequency is high, optimization of the backup structure is necessary.

The primary concern in optical telescope design is the deformation, not the stress, of the structure components. This is different for radio telescope design. Radio telescopes involve larger structures; therefore, a primary concern is the maximum stress induced inside the structure. Stress distribution of a circular plate supported at different radii is shown in Figure 7.28. When a circular plate is supported at 0.67 of the plate radius, the maximum stress is only one tenth of that of an inner ring support case. The smaller stress level of the backup structure can make it lighter in weight and simpler in design. A relatively larger support radius requires a stronger supporting structure. This can be accomplished by increasing the supporting truss height and width. A typical design is to combine the elevation axis, the drive gear, and the support structure together. This strong square structure also provides a separate support of the

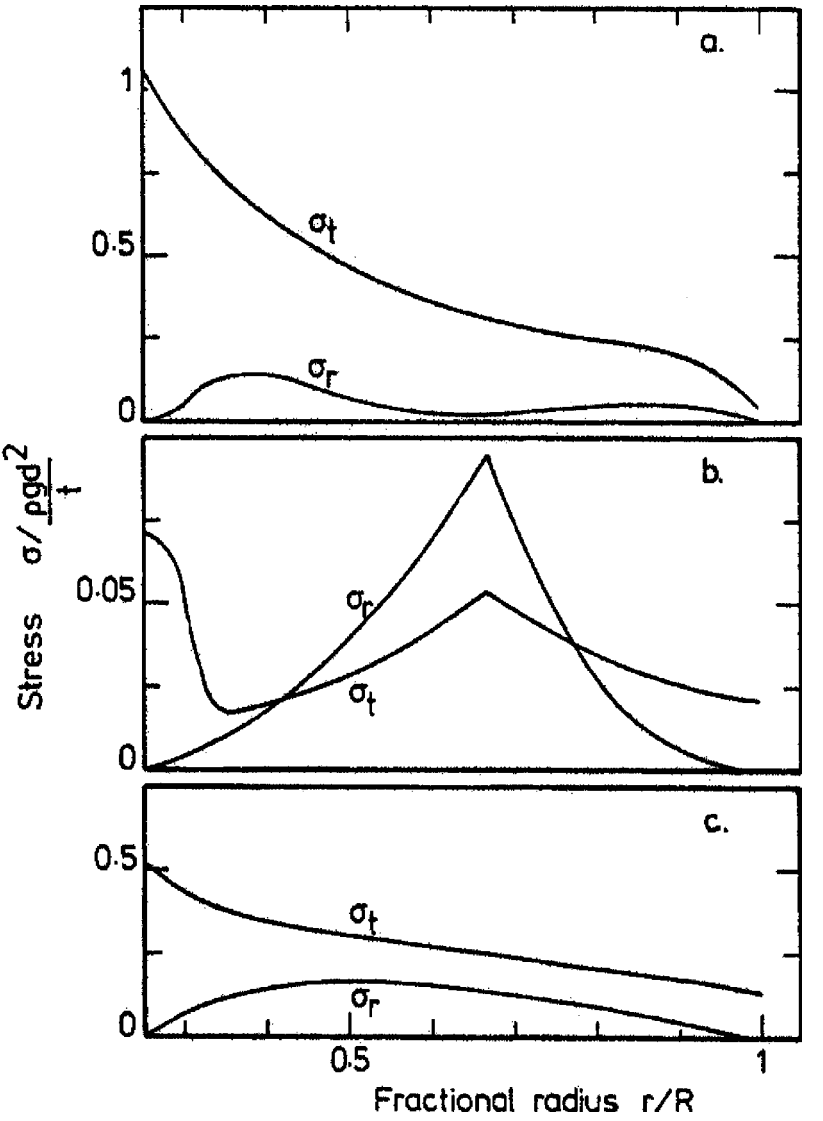


Fig. 7.28. The stress distribution of a circular plate under (**a**) an inner ring, (**b**) 0.67 radius, and (**c**) outer ring support (Cheng and Humphries, 1982).

subreflector and feed leg, so that the backup structure will have no added unsymmetrical loading. The homology is realized.

Different antennas usually use different backup structure support systems. A conservative approach is to build a strong box structure from the elevation support axis. This box structure forms the central part of the backup structure. Then the dish truss is supported through the box structure at its inner radius. In this design, the deformation of the dish central part is very small, but the

deformation increases as the radius increases. This deformation pattern agrees well with the homology design principle. The disadvantage of this design is that the weight of the backup structure is heavy and the cost is high. For an axial symmetrical truss structure, there are more truss members in the central part per unit area than those at the edge of the dish, while the area in the central part is smaller than that on the edge. Therefore, the central part is over-designed. This type of dish also has a lower resonant frequency.

Large or very large backup structures prefer a medium radius support in reducing their stress level. A commonly used support has four or six symmetrical hard points. These hard points are connected to a large structure formed by elevation bull gear and elevation axis. The other part of the backup structure is not connected to the support structure to avoid over constrained condition (Figure 7.29). The feed legs are also supported by these few points. Therefore, the dish homology is assured.

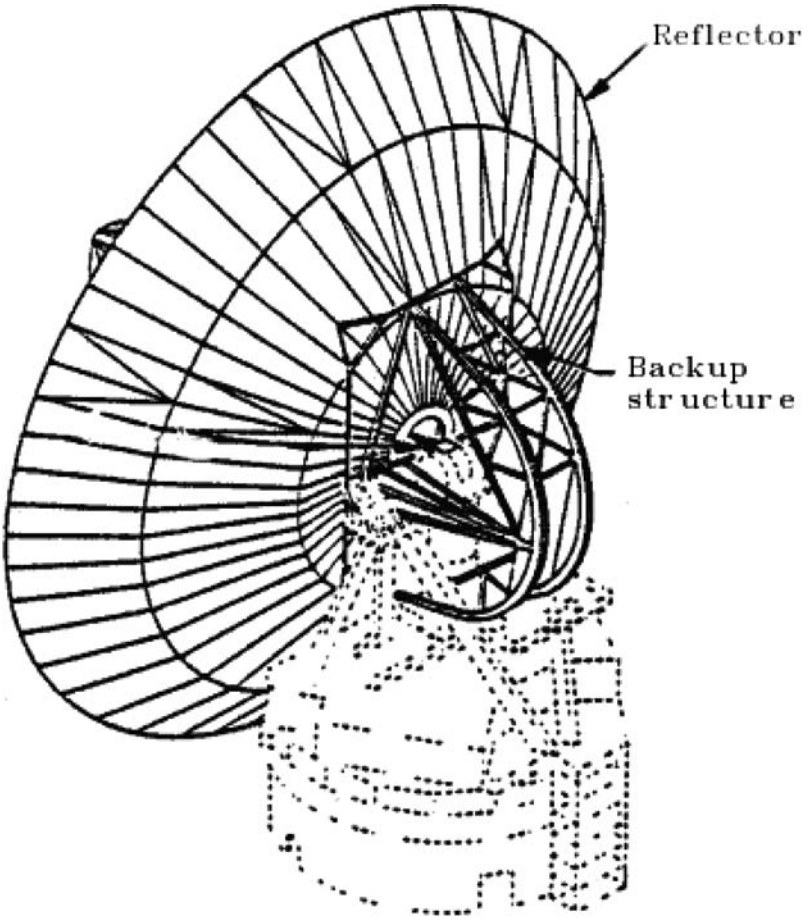


Fig. 7.29. A typical backup structure system.

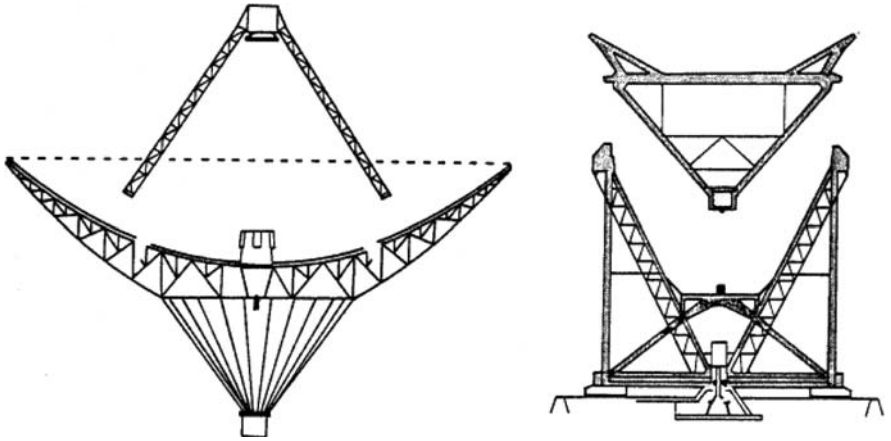


Fig. 7.30. The structural design of the German Effelsberg 100 m radio telescope.

Another backup structure design involves an inverse pyramid truss. In this design, the vertex of the pyramid is located on the axis behind the dish. From the vertex point, radial trusses support the backup structure at a middle radius. The entire support structure is axially symmetrical and so is the deflection pattern. This vertex point is then connected to a separate frame structure which connects the elevation bearings, the bull gear, and the feed legs. There are only four symmetric contact points for the dish and the frame. The German 100 m Effelsberg radio telescope uses this design concept (Figure 7.30).

The optimization of backup structure is usually time-consuming. If the backup structure is formed by connecting many two-dimensional radial ribs, the optimization can be performed in a two-dimensional plane instead of a three-dimensional space. In this case, half of the weight of all members between ribs should be added to the attached joint in the analysis.

If the reflector surface is a wire mesh one, the central part of the backup structure can have only one layer of truss as used in the Westerbork radio telescopes. If a radial prestressed structure is used, the rigidity of the structure will increase and the weight of dish decreases. This type of structure can only be used for low frequency or small size radio telescopes as the thermal stress may change the stress when temperature changes.

The distance between elevation axis and reflector surface is an important consideration in the backup structure design. The smaller the distance is, the less the counter weight required. This distance is also related to the operational elevation range. The smaller the distance is, the smaller the operational elevation range as it may be difficult for the dish to reach very low elevation angles. To overcome this, an asymmetrical elevation structure can be used. The dish center is offset from the azimuth axis. However, this asymmetrical design may require a counter weight in the azimuth structure. Using offset design, a path length error will be produced when the dish tracks celestial objects.

A separation between elevation axis and the dish surface is necessary. Therefore, a counter weight is always needed. To reduce the counter weight and the cost, it is better to place it far away from the elevation axis. However, this lowers the structure resonant frequency and increases the moment of inertia. It may not be desirable if the antenna requires fast motion in operation.

In a king-post design, an azimuth bearing supports a vertical yoke which holds the elevation bearings. The over-turning stiffness of the azimuth bearing is directly related to the pointing accuracy of the telescope. To increase the structural stability, a larger diameter azimuth bearing is preferred. The king post does not always provide room for a larger azimuth bearing. Therefore, very large radio antennas use wheel-on-track design.

In wheel-on-track design, there is no azimuth bearing, but azimuth wheels on a ring track. Usually, three or four groups of wheels are rotating on a ring track producing the required azimuth movement of the telescope. The structure on top of these wheels connects to a small central bearing at the azimuth rotation center. All the wheels are not perfect cylinders, but have a cone shape with a slightly inclined surface. Atop of these wheel groups, an 'A'-shaped truss structure is used as an elevation axis support (Figure 7.31). In this structure, two triangles on both sides are formed base to ensure the stability of the structure. Since the structure is tall, solar heating may produce noticeable pointing errors. Therefore, sun shielding of long truss members is necessary. Thermal effects on antennas will be discussed in Chapter 8. Wind also has influence on large antenna structures. Wind effect is discussed in Chapter 3 and wind loading on antenna structures is discussed in the next section.

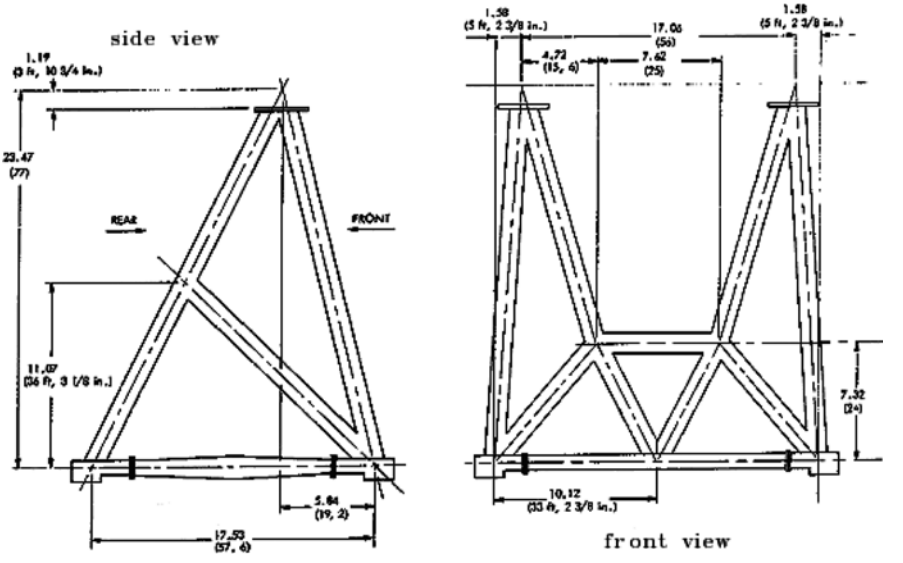


Fig. 7.31. The elevation axis support structure of a wheel-on-track design.

7.2.3 Wind Effect on Antenna Structures

In the past 40 years, a number of wind tunnel measurements have been performed in an effort to understand wind effects on antenna structures. Based on these measurements, wind pressure distributions at different elevation angles for a paraboloidal antenna surface are shown in Figures 7.32 and 7.33. In these figures, the contour lines are the pressure difference between the front and back of a paraboloidal dish surface. By using these pressure distributions, the antenna surface and pointing performance under wind loading can be predicted through finite element analysis.

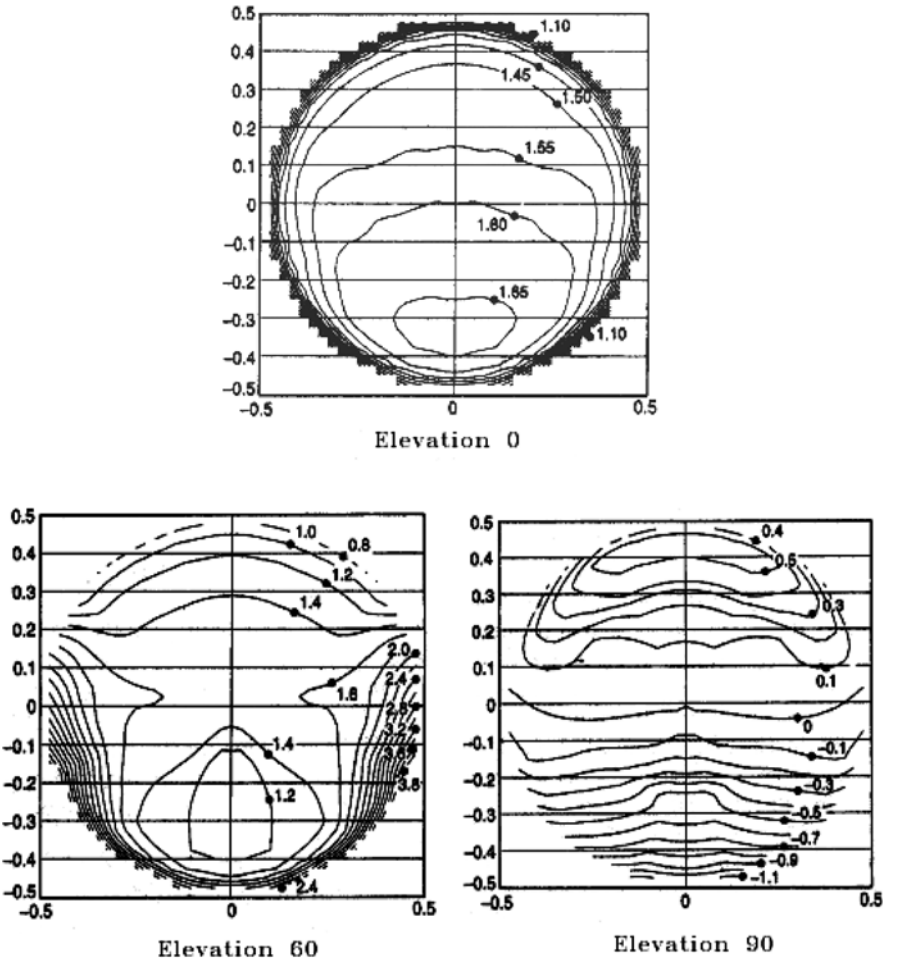


Fig. 7.32. Wind pressure over the aperture when the elevation angle is 0, or 60, or 90 degrees (Levy, 1996).

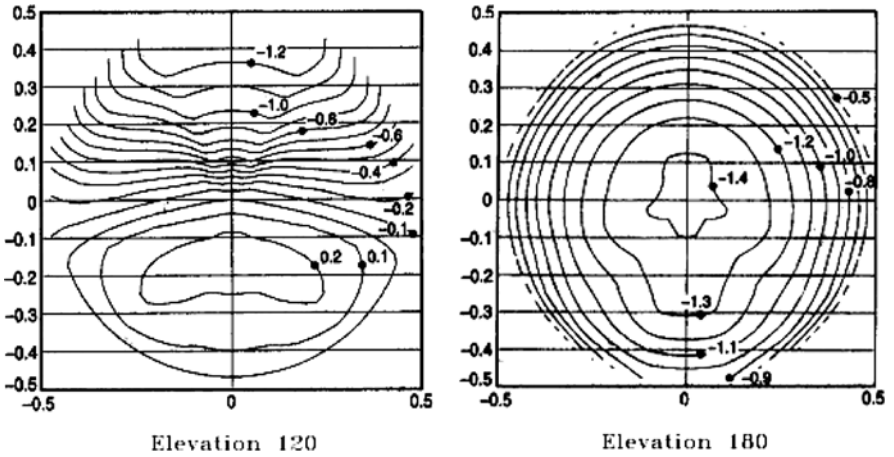


Fig. 7.33. Wind pressure over the aperture when the elevation angle is 120 or 180 degrees (Levy, 1996).

Besides wind pressure on the antenna surface, the wind effects on antennas also include the axial and tangential forces at its vertex and the moment on the elevation axis. The wind-induced axial force has its highest value when the elevation is about 60°. The wind-induced tangential force is generally small. The positive direction of this force is downwards when the dish points to the horizon. At a low elevation angle, the bottom part of dish has more wind pressure than the upper part of the dish so that a net negative tangential force is produced. The wind-induced moment has its highest value when the elevation is about 120°. The wind forces and moments on antennas with the elevation angle are shown in Figures 7.34, 7.35, and 7.36. Among these loadings, the maximum axial force and the maximum moment from the wind are very important in antenna structural design. These two maximum loadings are approximately:

$$\begin{aligned} F_{\max} &= 1.5A\left(\frac{1}{2}\rho V^2\right) \\ M_{\max} &= 0.15DA\left(\frac{1}{2}\rho V^2\right) \end{aligned} \tag{7.68}$$

where A is the antenna aperture area, D the diameter, ρ the density of the air, and V the wind velocity. Wind as a turbulence has a direct effect on antenna pointing performance. A new control strategy using Linear Quadratic Gaussian (LQG) optimal control theory is very effective for very large antennas in suppressing the wind turbulences (Gawronski, 2007).

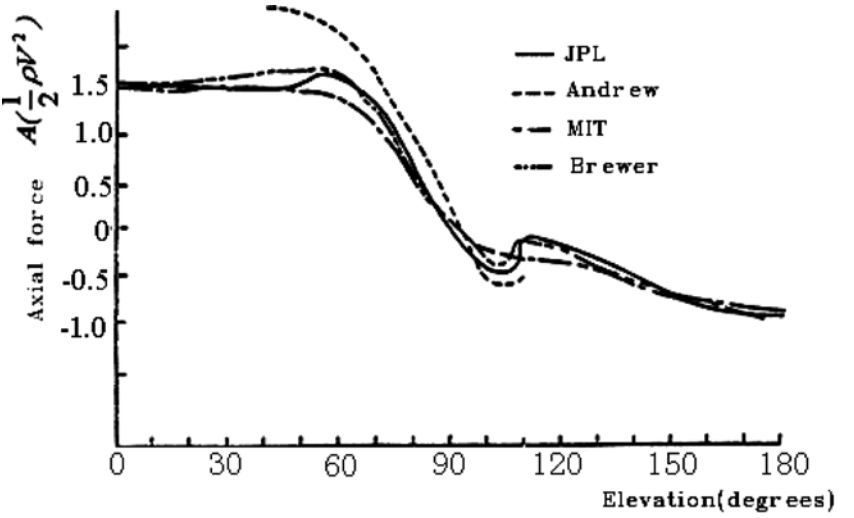


Fig. 7.34. Axial force of front wind as a function of elevation angle (Hirst and McKee, 1965).

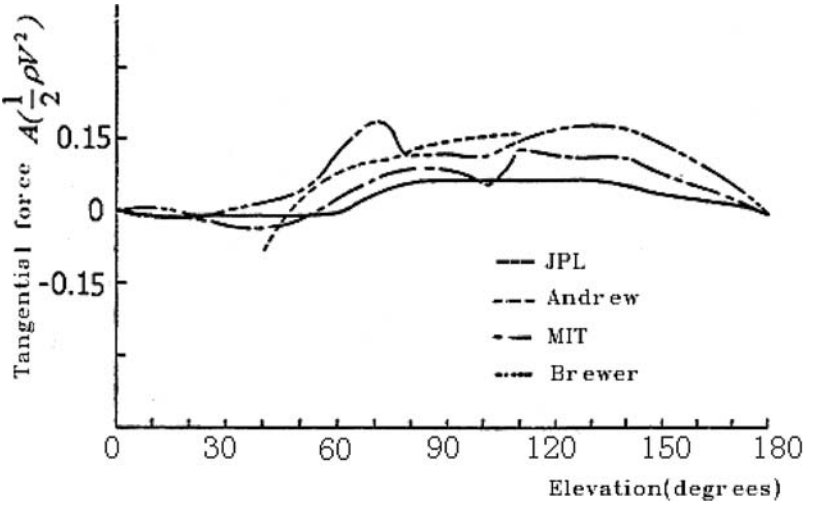


Fig. 7.35. Tangential force of front wind as a function of elevation angle (Hirst and McKee, 1965).

7.2.4 Active Control of Radio Telescopes

Active optics control of radio telescopes is still limited. In the past, a few attempts were made in relation to the subreflector's shape control. The shape of this type of subreflector is controlled by actuators to cancel aberrations, mainly the

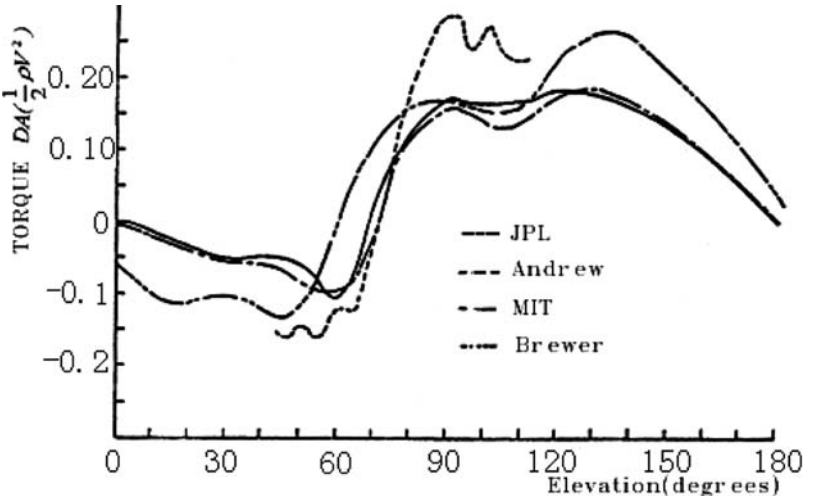


Fig. 7.36. Turning moment of front wind as a function of elevation angle (Hirst and McKee, 1965).

astigmatism, arising from the primary reflector deformation. However, there is no wide application for this technique in astronomy. In this section, only two types of active optics control experiments are discussed. These are the subreflector lateral position control and the laser ranger main reflector surface shape control. In Section 8.2.5, some active optics compensation methods used in the millimeter wavelength antennas are discussed.

For a precision radio telescope, a position change of the subreflector will produce both pointing error and gain loss. In general, to monitor the subreflector tilt is easy, but to determine the lateral displacement is fairly difficult. A laser device together with a mirror fixed on the subreflector provides tilt information for the subreflector. As the subreflector tilts, the reflected laser beam has an angle change twice as large as the tilt angle of the subreflector. However, the lateral displacement of the subreflector is hard to measure. In radio astronomy, a laser quadrant displacement detector is used for the lateral displacement measurement of the subreflector.

The monitoring and control of the primary reflector surface shape of radio telescopes is still in an experimental stage. There are two approaches: using edge displacement sensors as used in segmented mirror optical telescopes (Section 4.1.4) and using a laser range system which was tried in the GBT telescope. An edge-displacement sensor method has been planned in some large aperture submillimeter wavelength telescopes. The test of the laser range system on the GBT telescope was not fully successful. Presently, GBT surface measurements still rely on the holographic method (Section 8.4.1). The surface active control of the telescope is mostly done through a lookup table according to the telescope elevation.

7.2.4.1 Laser Quadrant Displacement Detector

The application of a quadrant detector for star guiding has been discussed in Section 3.3.6. The laser quadrant displacement detector was developed based on the same principle. Figure 7.37 is a schematic of a laser quadrant displacement detector system. The system includes a laser generator, a beam expander, a beam reducer, a quadrant detector, x and y rails, displacement actuators, and a control circuit.

The propagation of a laser beam in the atmosphere follows the Gaussian beam theory (Section 8.4.3). A major laser beam parameter is the width of the beam waist, which is the radius at which the field amplitude drops to $1/e$ of the peak amplitude or the field intensity drops to $1/e^2$ of its peak intensity. The formula of the beam waist width in quasi-optics is:

$$W(z) = W_0\left[1 + \left(\frac{z\lambda}{\pi W_0^2}\right)^2\right]^{1/2} \tag{7.69}$$

where W_0 is the inertial beam waist width, z the propagation distance, λ the wavelength, and $W(z)$ the beam waist width at the distance of z. By differentiating the above formula $dW(z)/dW_0 = 0$, an optimum inertial beam waist width is derived:

$$W_0 = \left(\frac{z\lambda}{\pi}\right)^{1/2} \tag{7.70}$$

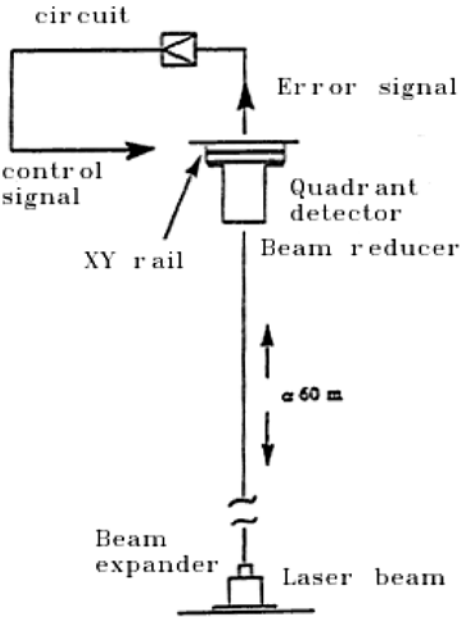


Fig. 7.37. Laser quadrant displacement detector for radio telescope secondary mirror control.

If the wavelength is 750 nm and the distance is 60 m, the optimum beam waist width required is 3.45 mm. The beam waist width of a laser generator is normally smaller than this so that a beam expander is needed in front of the laser generator. A beam expander is an afocal optical system formed by two lenses with different focal lengths. For a quadrant detector, there is a minimum image size where the optimum resolution occurs just a little bit above it. The beam radius for an optimum resolution of a typical quadrant detector is about 0.5–0.6 mm. Therefore, a beam reducer, another afocal system, is needed in front of the detector. The calculation of the beam reducer is the same as that for the beam expander. The detector should be located at the exit pupil of this beam reducer, so that the beam reducer tilt will not produce position errors of the laser beam on the quadrant detector.

The lateral movement range of this detector is about twice the image size. For increasing the movement range, a larger image size on the detector is required. In this device, a filter centered at the laser wavelength with a bandwidth of 20 nm is used. The atmosphere has a small effect on laser light scattering. The energy losses of the laser beam include the absorptions from the atmosphere, glasses, and filter, and the losses at lenses, glasses, and detector surfaces. The atmospheric absorption is about 2% for a distance of 10 m.

During the displacement measurement, laser light modulation and synchronous detection are necessary, so that the noise of measurement is reduced. The device can be locked with the subreflector, so that the relative displacement between the subreflector and the primary reflector is detected.

7.2.4.2 Laser Ranger System

Laser ranger systems are widely used in many important fields. One large-scale experiment was carried out on the largest steerable offset GBT (Figure 7.38). The telescope has a shortest working wavelength of 3 mm. The whole testing system included two important components: the laser ranger and the retroreflectors system (Payne and Parker, 1990).

There were 18 laser rangers for the telescope in the test. Twelve of these were located on the ground around the telescope and six of these were on the feed leg structure. There were many retroreflectors both on the ground and on the telescope structure. At some antenna structure locations, two retroreflectors were connected back-to-back as coordinate transformation tools. The positions of these retroreflectors could be measured accurately using multiple laser rangers on the ground, while the positions of the laser rangers on the moving antenna structure could be determined through these back-to-back retroreflectors.

When the laser ranger positions on the antenna feed leg were determined, the surface shape and relative distance from the reflector to the feed could be determined by laser rangers and retroreflectors at corners of each surface panel. The data was used by the actuators to adjust the reflector surface shape and pointing direction.

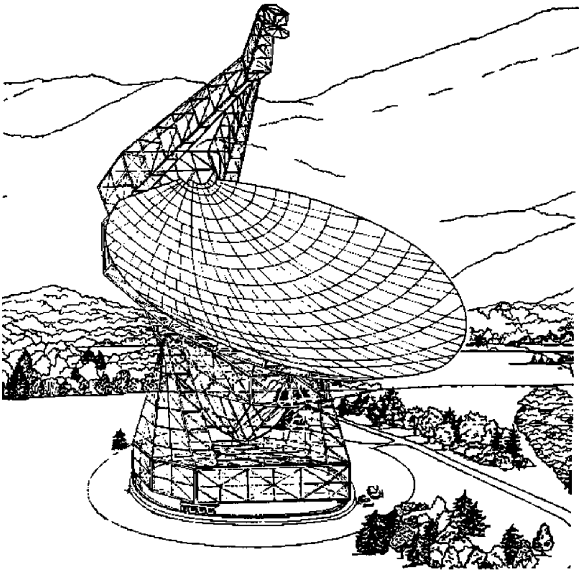


Fig. 7.38. Green Bank 100 m telescope (NRAO).

The laser ranger block diagram is shown in Figure 7.39. In the system, a laser diode of 780 nm wavelength was amplitude modulated at a frequency of 1.5 GHz. The laser beam was directed to a retroreflector. After reflection, the beam was mixed with a local oscillator (LO) signal with a frequency of 1,500.001 MHz, producing an output frequency of only 1 kHz. This low frequency output had the same phase as the reflected high frequency laser beam. This phase represented a small remainder path length of $2d/\lambda$, where d was the distance between the laser ranger and the retroreflector and λ the wavelength of the modulation frequency of the laser signal. These phase details were calibrated by a standard clock with a frequency of 20 MHz and an accuracy of 100 μm in path length was achieved.

Since $\lambda << 2d$, there was an uncertainty in distance which was multiples of the wavelength. This uncertain distance was 10 cm for a frequency of 1.5 GHz. For determining the absolute distance, a method of tuning the frequency or using a lookup table could be used. In this way, the distance uncertainties could be eliminated completely.

Since the beam was transmitted through the atmosphere, there were two effects from the atmosphere refractive index change: a large-scale one and a small-scale one. The large-scale change of the index follows the following law:

$$n_g = 1 + \frac{n_{g0} - 1}{1 + T/273}\frac{P}{100} - \frac{5.5 \times 10^{-8} e}{1 + T/273} \tag{7.71}$$

where T is the temperature in degrees C, P the pressure in mmHg, e the partial pressure of the water vapor in mmHg, n_{g0} the refraction index at 20°C and

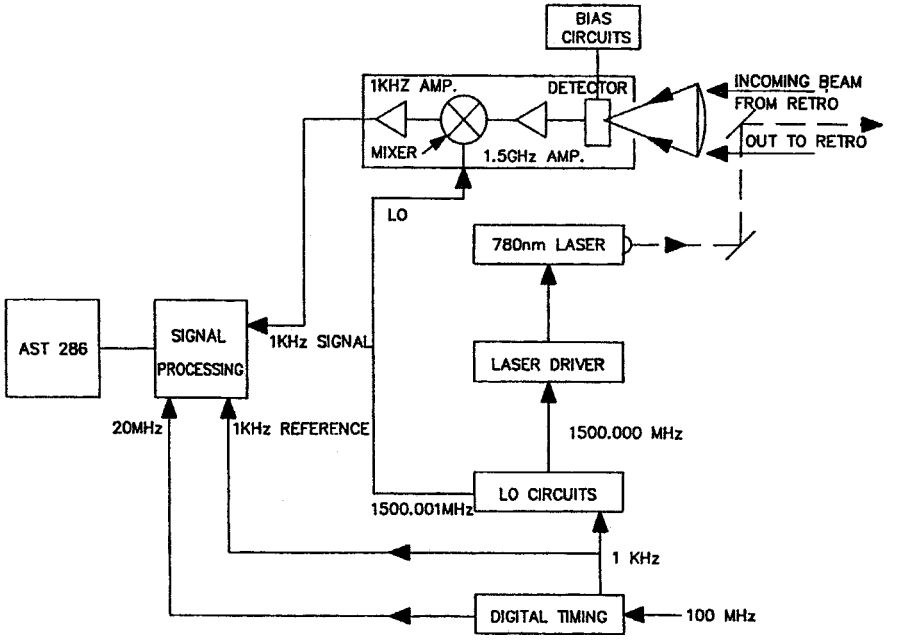


Fig. 7.39. The 100 m GBT telescope laser range system (Payne and Parker, 1990).

760 mmHg, and $dn_g/dt = 10^{-6}$ when the temperature is 20°C. The relationship of $dn_g/de = 0.5 \times 10^{-7}$ is not related to the temperature and the pressure. The small-scale change of the index had a larger influence on the measurement accuracy. In a distance of 60 m, this small-scale effect was about 60 μm/°C.

The absolute and relative accuracies of the GBT laser range system were 50 μm+1 ppm and 5 μm+0.2 ppm, where ppm represents in one millionth of the measured distance. The error came from three sources: (a) the drift in the electronic system; (b) the error of average atmospheric refractive index; and (c) the local air refractive index change. The sum of these errors was about 1.5∼2 mm. Among these errors, the most important one came from the drift of the electronic system. For correcting this phase drift, a device inside the laser ranger is used. It could direct the emitting signal back to the detector for calibration. Therefore, the distance measurement was more accurate.

The atmospheric index change could be calibrated through the distance change between two fixed retroreflectors. The measurement accuracy improved by increasing the sampling. One factor which affected the stability was the ground height change with the time. To monitor this, a hydrostatic leveler with an accuracy of $0.01''$ could be used.

In the laser ranger system, the return beam to the emitter should be avoided. The circuit should be shielded from outer electromagnetic interference. Since the atmospheric refractive index was related to the temperature, a high accurate temperature calibration was required.

Standard retroreflectors are formed by three perpendicular planes which pass through a common point. These retroreflectors can reflect a beam back to its incoming direction. Figure 7.40 shows these retroreflectors used on each panel's corner. When the position of the retroreflector was determined, the required adjustment for the actuator was known. The surface shape could be improved.

The standard retroreflector has a limited field of view. For enlarging the field of view, a wide field retroreflector was also used in the system as shown in Figure 7.41. The retroreflector was formed by two half spheres of different radii. These two half spheres are connected center-to-center. An entrance pupil is in between two half spheres. For ensuring the accuracy of the center, the retroreflector was actually made from one sphere and a thick lens using glue with the same refractive index. The field of view of this type of retroreflector was up to $\pm 65^\circ$. If the main beam is the light which passes through the center, then the condition for an edge beam to reflect back is:

$$R = \frac{\sqrt{1-h^2} + \sqrt{n^2-h^2}}{n^2-1} \tag{7.72}$$

where $0 \leq h < 1$ is the height of the beam relative to the radius of the small half sphere, n the index ratio between the glass and air, and R the ratio between the large and small radii of both half spheres. For any R, there were only two groups of beams that were reflected back to the incoming direction. One was the main beam and the others were beams of which the height satisfied the above equation. If the retroreflector was made of BK7 glass, the index ratio is $n = 1.511186/1.000290 = 1.51075$. When the laser wavelength is $\lambda = 0.78\,\text{nm}$, the required radius ratios for each beam height to reflect back are listed in Table 7.5.

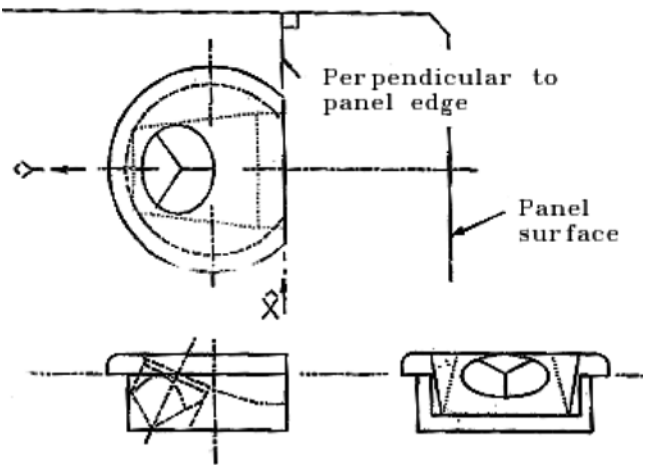


Fig. 7.40. The retroreflector used in the 100 m GBT panel surface (NRAO).

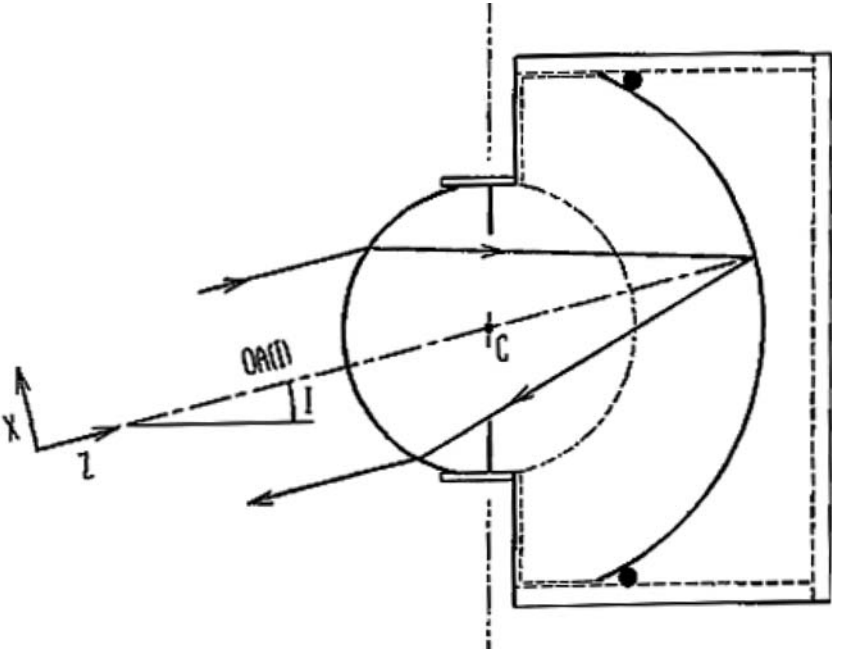


Fig. 7.41. The wide angle retroreflector used in the 100 m GBT (NRAO).

Table 7.5. The relationship between the spherical lens radius and the return ray height

h	0	0.2	0.4	0.6	0.8	0.9	0.96	0.98
R	1.9579	1.9318	1.8508	1.7051	1.4673	1.2861	1.1280	1.0518

In this wide field retroreflector, the radius of the small half sphere was 50 mm and large radius 96.5 mm. The beam height, h, to reflect the beam back was 0.20882. Using computer ray tracing, all rays with their height smaller than h would deviate by small angles and all rays with their height greater than h would spread out quickly. The spot diagram had a dense central area and a light outer ring. The spot diagram remained the same for all the incoming beam directions within the field of view. The field of view was related to the diameter of the entrance pupil.

Since the light reflected inside the glass sphere, the equivalent vertex in the air was far away from its center. This extra distance should be removed in the path length calculation. This distance was:

$$L = n(R_1 + R_2) - R_1 \tag{7.73}$$

7.3 Radio Interferometers

7.3.1 Fundamentals of Radio Interferometers

Resolution of a filled aperture telescope is related to the diameter and wavelength. The wavelengths in the radio band are 10^5–10^9 times those of visible light. Therefore, with any filled aperture radio antenna it is impossible to achieve the same resolution of optical telescopes. For increasing the resolution in the radio band, interferometers are used. The theories relating to interferometers have been discussed in Sections 1.4 and 4.2. In this section, theories of radio interferometers are briefly discussed.

For a two-element interferometer shown in Figure 7.42, if $\vec{D}$ is the antenna baseline vector and $\vec{i}$ the unit vector towards the radio source, one of the antennas will have a time delay for the signal to arrive relative to the other (Rogers, in Meeks, 1976):

$$\tau = -\frac{(\vec{D} \cdot \vec{i})}{c} \tag{7.74}$$

where c is the speed of light in a vacuum. This formula can be expressed in an equatorial coordinate system as:

$$\tau = -\frac{D}{c}[\sin\delta_B \sin\delta_S + \cos\delta_B \cos\delta_S \cos(L_S - L_B)] \tag{7.75}$$

where δ and L are declination and hour angle of the source and the baseline, and the Subscripts B and S represent the baseline and the source. This small time delay can be expressed as a phase delay of the wavefront:

$$\phi = \omega\tau \tag{7.76}$$

where ω is the angular frequency of the radio wave. The angular resolution of an interferometer is:

$$\frac{d\phi}{d\theta} = -\left(\frac{\omega}{c}\right) D_T \tag{7.77}$$

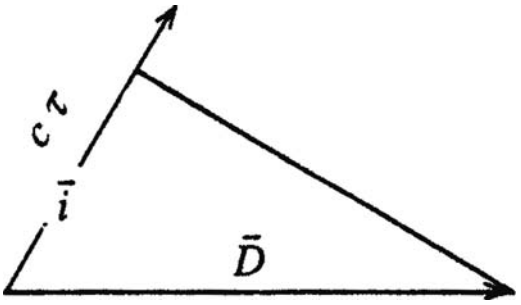


Fig. 7.42. Principle of a radio interferometer.

where D_T is the projection of the baseline vector $\vec{D}$ on the plane perpendicular to the source direction.

Radio interferometers are divided into two groups: adding interferometers and correlation interferometers. Since any two antennas of an interferometer have a phase difference for any particular source, by compensating the delay of one antenna, the signals from both antennas can add together to extract the visibility function, which is the Fourier transform of the source distribution. An adding interferometer is the same as a phased array. One property of a phased antenna array is that the pointing of an array can be adjusted electronically instead of physically moving the antenna elements. The adding interferometer works exactly the same as a single unfilled aperture telescope. The information produced by an adding interferometer contains a constant term and a baseline related term (Equation 1.132). The baseline related term is the visibility function. The visibility function is a Fourier transform of the source intensity distribution (Equation 1.138). From the complex visibility, the source image can be formed through a Fourier transform.

If signals from two antennas are multiplied and accumulated (cross-correlated) instead of being added, then the interferometer is a correlation interferometer. The information provided by this interferometer is the visibility function. A correlation interferometer which forms an image of the sky source is called an aperture synthesis telescope. An aperture synthesis telescope involves many baselines to fill the u-v plane.

In a phased array, signals are simply added together and so are the noises. In radio astronomy, the noise is much larger than the signal received. In a correlation interferometer, the noises are suppressed as there is no correlation between two random noise terms of two beams. If the signals and noises received by two antennas are V_{is} and V_{in}, $(i = 1, 2)$, then the outputs from an adding interferometer and a correlation interferometer can be expressed respectively as (Emerson, 2005):

$$\begin{aligned} <(V_{1s} + V_{2s} + V_{1n} + V_{2n})^2> &= V_{1s}^2 + V_{2s}^2 + V_{1n}^2 + V_{2n}^2 + 2V_{1s}V_{2s} \\ <(V_{1s} + V_{1n})(V_{2s} + V_{2n})> &= V_{1s}V_{2s} \end{aligned} \tag{7.78}$$

The above formula shows that a correlation interferometer is superior over an adding interferometer if the noise is significant. The correlation interferometer is free from noise terms if the statistical fluctuation of signals is ignored, while the adding interferometer has square terms of noises.

The visibility is a function of the baseline vector on the u-v plane. If the visibilities over the u-v planes are all known, the source distribution can be derived through an inverse Fourier transform. Figure 7.43 is an example of this type of one-dimensional inverse transform. It shows that the sun does not have a brightness sharp increase at its edge in the radio wavelength.

The signal correlation between two antennas can be achieved either by a digital technique or through a phase-switching technique. If a digital technique

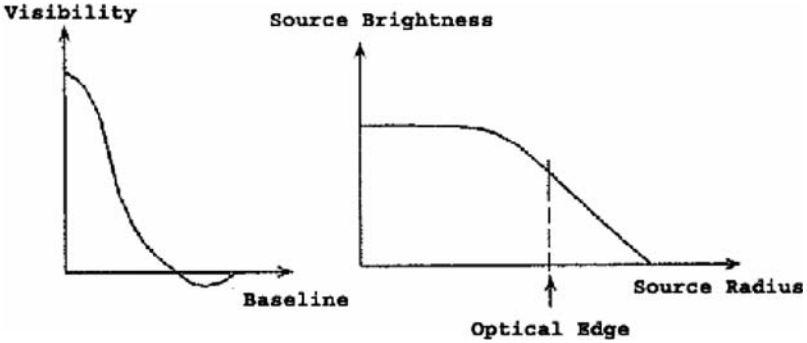


Fig. 7.43. The visibility function and the illumination distribution at the edge of the sun in radio wave.

is used, the output from each antenna has to be converted into a digital form. The sampling rate in an analog-to-digital converter should be at least twice as much as the bandwidth used. This sampling rate may limit the bandwidth used in the observation.

To avoid this, a continuous phase-switching technique may be used. Using this technique, signals of two elements add and subtract before averaging. Then the correlation of two beams is derived from the difference between the square terms of the sum and the square terms of the difference. This process retains information over a wide bandwidth. In mathematics, it is:

$$<(V_1 + V_2)^2> - <(V_1 - V_2)^2> = 4<V_1 \cdot V_2> \tag{7.79}$$

An earlier phase switch technique is through the measurement of the sine $S(\vec{D})$ and cosine $C(\vec{D})$ terms of the visibility function. The measurement is done from a special circuit shown in Figure 7.44. The formula used is:

$$C(\vec{D}) + iS(\vec{D}) = \int_{\phi} B(\theta) e^{2\pi i \vec{D}/\lambda} d\theta \tag{7.80}$$

The left hand side of the equation is the visibility function and the source brightness distribution is $B(\theta)$ on the right hand side.

7.3.2 Aperture Synthesis Telescopes

The basic principle of an aperture synthesis telescope is based on the correlation interferometer. As stated earlier, the cross-correlation of two beams is a single value of the visibility function relative to one baseline. If all the values or samples of the visibility functions in the (u, v) plane are known, then the source intensity distribution can be derived by an inverse Fourier transform. An aperture synthesis telescope is just such a mapping instrument with an array of antennas. The sampling of visibility function in the u-v plane is also through the rotation of the earth.

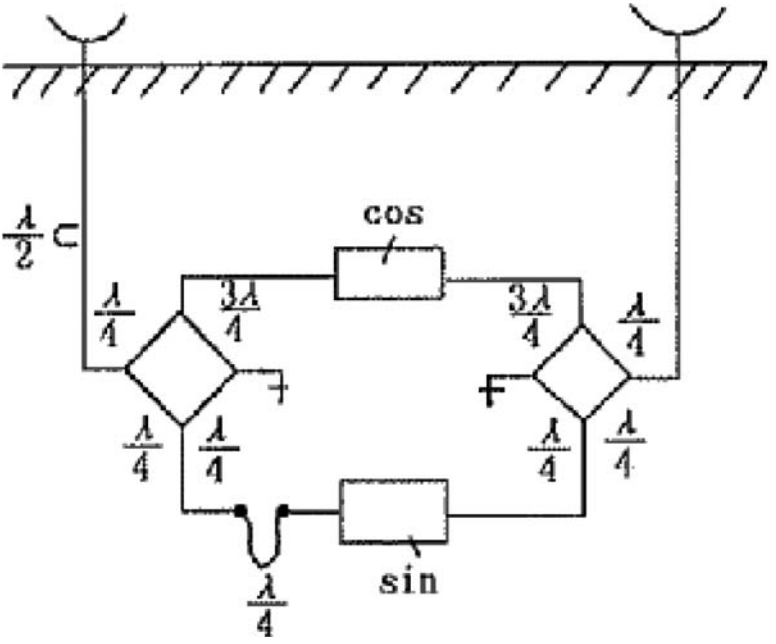


Fig. 7.44. The circuit for measuring sine and cosine terms of correlation function.

When we discuss the aperture synthesis, two assumptions are necessary. (a) the sources of interest are a long way away. All we have measured is a source surface brightness without the third dimension, the depth and (b) all the radiations received come from a small portion of the celestial sphere centered at a particular direction. This direction is called the phase tracking center. The synthesis image is in a plane of (l, m). The baseline vector has components (u, v, w) where w points in the direction of interest (Figure 7.45).

If L_x, L_y, and L_z are the Cartesian coordinate differences of an antenna pair, the baseline components (u, v, w) are given by:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin H_0 & \cos H_0 & 0 \\ -\sin\delta_0 \cos H_0 & \sin\delta_0 \sin H_0 & \cos\delta_0 \\ \cos\delta_0 \cos H_0 & -\cos\delta_0 \sin H_0 & \sin\delta_0 \end{pmatrix} \cdot \begin{pmatrix} L_x \\ L_y \\ L_z \end{pmatrix} \tag{7.81}$$

where H_0 and δ_0 are the hour angle and declination of the phase tracking center and λ is the center frequency of the receiver system. By eliminating H_0 from the expressions for u and v, we obtain the equation of an ellipse in the plane (Thompson et al., 2001):

$$u^2 + \left(\frac{v - (L_z/\lambda)\cos\delta_0}{\sin\delta_0} \right)^2 = \frac{L_x^2 + L_y^2}{\lambda^2} \tag{7.82}$$

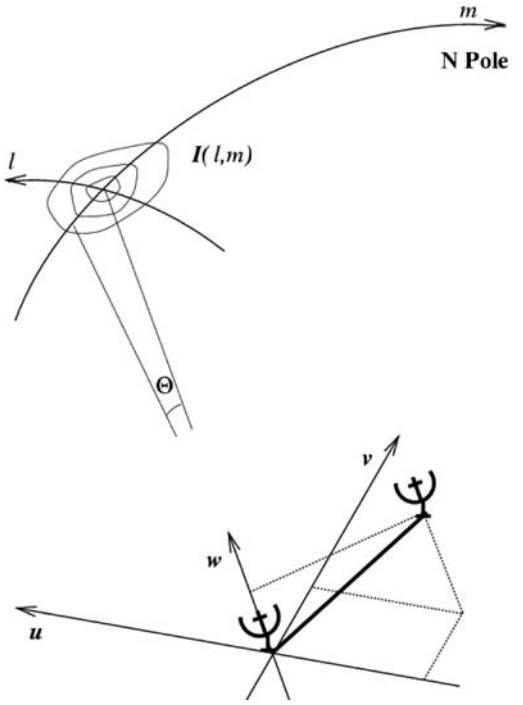


Fig. 7.45. Relationship between baseline and image coordinates.

This equation shows that as an interferometer observes a point source on the celestial sphere, the rotation of the earth causes the baseline to trace out an ellipse locus in the (u, v) plane (Figure 7.46).

In an aperture synthesis telescope, the baselines of array antennas should be optimized to form baselines that sample the u-v plane as well as possible. Some arrangements have multiple pairs of antennas with the same separations, called "redundant baselines." For a fixed number of antennas, redundant baselines reduce the number of sampled positions in the u-v plane, but some arrays (e.g. Westerbork) include them to assist calibration. In the Stanford five-antenna array, the separation length ratio between antennas is 1:1:4:3, achieving the least redundancy for a one-dimension five-antenna array with integral spacings. However, with modern computers there is no reason to stick to integer ratios, as integer ratios produce much worse sidelobes (grating rings) than noninteger ones. For minimizing the sidelobes of an array, Kogan (1997) provides an approach for (u, v) plane optimization. In his approach, if vector $\vec{r}_i$ determines the position of array elements at the array aperture measured in wavelength, then the beam pattern of the array along a direction is determined by:

$$P(\vec{e}) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{n=1}^{N} e^{-i2\pi(\vec{r}_k-\vec{r}_n)\vec{e}} = \frac{1}{N}\sum_{k=1}^{N} e^{-2\pi\cdot\vec{r}_k\vec{e}}\,\frac{1}{N}\sum_{n=1}^{N} e^{-2\pi\cdot\vec{r}_n\vec{e}} = |V(\vec{e})|^2 \qquad (7.83)$$

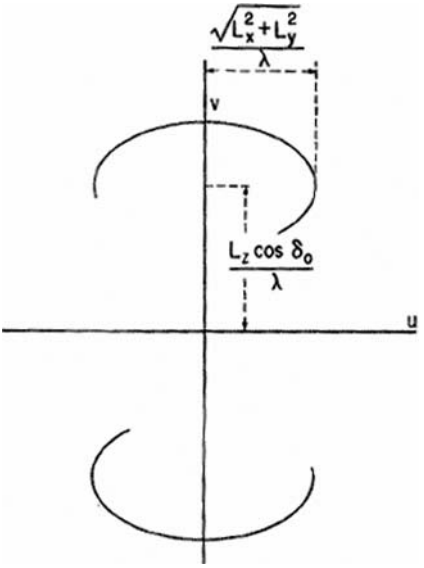


Fig. 7.46. Aperture synthesis by using the rotation of the earth (Thompson, 1999).

where $\vec{e}$ is the vector of the direction on the sky, $V(\vec{e}) = (1/N)\sum_{n=1}^{N} e^{-2\pi\cdot\vec{r}_n\vec{e}}$ a voltage beam pattern, and N the number of elements. Therefore, the relative change of the beam pattern on a shift of a given element along the given direction is:

$$\frac{dP_{rn}(\vec{e})}{P} = 4\pi(\vec{e}\cdot\Delta\vec{r}_n)\frac{\sum_{k=1,k\neq n}^{N}\sin\ 2\pi(\vec{r}_k-\vec{r}_n)\cdot\vec{e}}{|\sum_{n=1}^{N} e^{i\,2\pi\cdot\vec{r}_n\cdot\vec{e}}|} \tag{7.84}$$

7.3.3 Weiner–Khinchin and Van Cittert–Zernike Theorems

In mathematics, an important relationship, named the Weiner–Khinchin (or Khintchin) theorem, exists between the auto-correlation function and the power spectral density of a stationary random process. This theory states the power spectral density is the Fourier transform of the corresponding auto-correlation function:

$$S_{xx}(f) = \int_{-\infty}^{\infty} r_{xx}(\tau)\exp(-j2\pi f\tau)d\tau \tag{7.85}$$

where $r_{xx}(\tau)$ is the auto-correlation function. It is defined as:

$$r_{xx}(\tau) = \int_{-\infty}^{\infty} x(\tau)x(t-\tau)d\tau = <x(\tau)x(t-\tau)> \tag{7.86}$$

When the above theorem is used in electromagnetic waves, it is referred to the temporal coherence as discussed in Section 4.2.4. Therefore, the auto-correlation of the electromagnetic signal in time will provide the source spectral distribution in the frequency domain.

In Section 4.2.4, another very important theorem on the spatial coherence of electromagnetic waves is presented, the Van Cittert–Zernike theorem. The theorem states that the spatial correlation function (or cross-correlation function) between two beams with baseline difference in space is the Fourier transform of the source brightness distribution. The spatial cross-correlation function is referred to the visibility function (Equations 1.137 and 4.147). However, the Weiner–Khinchin relationship is a mathematically rigorous one, while the Van Cittert–Zernike one is only an approximation. There are two basic assumptions for the Van Cittert–Zernike theorem: (a) the observation performed is confined in a *u*-*v* plane; the source distance is infinity to the observers and (b) the sky source brightness distribution is limited to a very small region (usually $<1^\circ$) in the sky as the response of a radio antenna is limited to sources which lie within its primary beam.

The above two theorems have an important bandwidth limitation. The time difference between two beams should be less than the coherence time, which is the reciprocal of the bandwidth. The corresponding path length is the coherence length (Section 4.2.4).

The advantage of using a cross-correlation operation is that not only the amplitude but also the phase is recorded. In Section 8.4.1 the holographic measurement of the antenna surface is discussed, which is through the cross-correlation between responses of an aperture antenna and another reference horn antenna. The reference antenna is usually fixed or has a much wider, flat main beam shape both in amplitude and in phase so that the phase difference recorded between the measured aperture antenna and the reference one represents the phase distribution of the measured aperture antenna. Through Fourier transform of the recorded radiation pattern, the phase on the aperture or the antenna surface deviation from an ideal paraboloid shape is derived with high accuracy.

7.3.4 Calibration: Active Optics After Observation

The relationship between the visibility function $V(u,v)$ and the source brightness function $B(l,m)$ is:

$$V(u,v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} B(l,m)\, e^{-j2\pi(ul+vm)}\, dl dm \tag{7.87}$$

In fact, the antenna primary beam has a direct effect on the integral of the equation. Including this effect, the formula is:

$$V(u,v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(l,m)B(l,m)\, e^{-j2\pi(ul+vm)}\, dldm \tag{7.88}$$

where $A(l,m)$ is the normalized reception pattern of the interferometer pair. The term $(ul+vm)$ in the exponential is the geometric phase difference $\Delta\phi_g$ produced by the differential path length between the radiations from the source located at (l,m) to each antenna, compared with a fictitious source at the phase tracking center (Figure 1.1). Following Equation (7.75), the total geometric phase difference ϕ_g, or geometric group delay τ_g, for the interferometer with baseline components (L_x, L_y, L_z) is (Fomalont and Perley, 1999):

$$\phi_g = 2\pi \frac{c}{\lambda}\tau_g = \frac{2\pi}{\lambda}(L_x \cos H \cos\delta - L_y \sin H \cos\delta + L_z \sin\delta) \tag{7.89}$$

Therefore, the first-order differential geometric delay between radiations from a source direction and a reference direction, when the baselines differ from those assumed, is:

$$\begin{aligned}\Delta\phi_g = 2\pi \frac{c}{\lambda}\Delta\tau_g = \frac{2\pi}{\lambda}&(\Delta L_x \cos H \cos\delta - \Delta L_y \sin H \cos\delta + \Delta L_z \sin\delta \\ &+ \Delta\alpha \cos\delta(L_x \sin H + L_y \cos H) + \Delta\delta(-L_x \cos H \sin\delta \\ &+ L_y \sin H \sin\delta + L_z \cos\delta))\end{aligned} \tag{7.90}$$

where α and δ are true source position. The observed visibilities $\tilde{V}_{ij}$ are not the true visibilities V_{ij}, where i and j are the antenna numbers in the pair. Generally, a linear relationship can be assumed between these two visibilities:

$$\tilde{V}_{ij}(t) = g_i(t) g_j^*(t) G_{ij}(t) V_{ij}(t) + \varepsilon_{ij}(t) + \eta_{ij}(t) \tag{7.91}$$

where t is the time, g_i and g_j the variable baseline-based complex gain related to antenna i and j, G_{ij} the relative stationary baseline-based complex gain, ε_{ij} the offset, and η_{ij} the complex random noise.

In optical observations, the wavefront phase error in the aperture plane can be corrected by using a wavefront sensor and actuators before the final image is formed (see Chapter 4). The amplitude in optics is assumed to be unaffected by the atmosphere or the primary mirror. However, in a radio interferometer, the active optics correction is performed not before but after the observation. The correction or calibration includes both amplitude and phase of the visibility function. The visibility is one representation of the aperture field. The amplitude here is referred to as the gain. For each pair of antennas, the gain is not expected to change much.

In the correction, the first task is editing. It checks the gains and other obvious errors. If the gains cannot be determined for some or all antenna pairs over a

period, these data should be removed. Other suspect data may be removed as well. This task was handled by astronomers in the past. Now it is done through display by software.

Real field calibrations include direct calibration, calibration through radio sources, and self-calibration. Direct calibration is necessary for all aperture synthesis arrays. It is the first part of active optics. For an aperture synthesis array, all antennas have to follow a given specification to ensure their uniform reception pattern and pointing performance. The amplitude should be stable to a few percent. If a pair is formed from antennas of different design, gain amplitude calibration is necessary. If one antenna is shaded by another as the projected baseline is smaller than the diameter, the best procedure is to remove the data, although calibration is possible if all the information is available.

Antenna pointing or tracking accuracy should be better than one tenth of FWHP of the primary beam. Large errors can reduce sensitivity of the array. Factors which affect pointing error are discussed in Chapters 3, 7, and 8. In Equation (7.75), the differential geometric delay is included. However, this does not include the time differences in signal propagation through different electronic paths especially when signal frequencies are different. The calibration of this small delay can be done by observing a strong and isolated source. The required delay is adjusted in steps for maximizing the coherence. Another method is to use the phase slope respecting with the frequency which is caused by a delay error. When the delay is optimized, the phase slope will be zero.

Calibration through radio sources is the most important part of the active optics in the aperture synthesis. The task is to check the phase difference between antenna pairs. Since there is no absolute phase reference, an antenna phase offset can be determined by observing a sky calibrator. The calibrator works like the laser guide star in adaptive optics. If the array is not completely phase- or gain-stable, periodic observations of calibrators can be used to monitor these changes. Since the calibrator observations are not concurrent or co-located in the sky, some arrays require antennas to perform fast switching movement.

Self-calibration is mainly for phase correction although amplitude can also be checked through amplitude closure in some cases (Cornwell and Fomalont, 1999). The function of self-calibration works exactly the same as a deformable mirror in optical telescopes. Self-calibration includes redundant calibration, calibration through phase and amplitude closure, and phase data optimization after the observations. If an array includes a few pairs which measure the same spacing, or the same (u, v) sample, then these data can be used as redundant calibration. For arrays with more than three antennas, the sum of the phases of three closed baselines should be zero. These are good constraints in the solution finding. For amplitude closure, a loop of four elements is needed. The calibration through optimization is mainly a mathematical approach in optimizing the problem for finding the best fitting visibility functions.

7.3.5 Very Large Array, Expanded Very Large Array, and Square Kilometer Array

Among aperture synthesis methods, there is a special technique called super synthesis. By using this technique, both the antenna regrouping as well as the earth rotation is used in improving the *u-v* plane coverage. The simplest super synthesis telescope is formed by two elements. In one direction, the baseline changes through the movement of one antenna; in the other direction, the baseline change is through the earth's rotation. The largest existing aperture synthesis telescope is the Very Large Array (VLA) of NRAO. This telescope consists of 27 antennas of 25 m diameter. The total collecting area is equivalent to a single dish of 130 m diameter. The area covered by the VLA is 14,000 m^2. The cost of the project was $78 M in the 1980s. The VLA antennas are arranged on Y-shaped rails. Two of three arms are 21 km long and the third one is 19 km. Altogether, 27 antennas form 351 baselines. Being a Y, the VLA's *u-v* sampling is maximized by an observation of 8 h (1/3 of a day). Now an Expanded Very Large Array (EVLA) project is under construction to improve its sensitivity and bandwidth.

Another ambitious multi-national project proposed is the Square Kilometer Array (SKA). It has a total collecting area of one square kilometer, a hundred times more than that of the VLA array. The number of antennas is in the thousands. One basic design of this huge array involves planar aperture arrays for low frequency bands and a small or medium size steerable dish array jointed with an aperture core array in the center for intermediate and high frequency bands. The three frequency ranges make three sub-programs of the SKA project, named SKA-low, SKA-mid, and SKA-high. For achieving a large field of view, a focal plane array is also proposed. One design of the planar aperture arrays has 64 wide-band antenna elements stationary over a one square meter area. The steering of the beam is achieved by adjusting the phase delay of each antenna element. This technique has been demonstrated in the Netherlands Thousand Element Array (THEA) test project.

One type of parabolic dish design used for the SKA involves a newly developed hydroforming technique to produce a thin aluminum stretched parabolic surface stiffened with a simple truss structure. Another technique is to produce a larger fiber glass sandwiched composite dish structure using an on-site molding. The cost of each antenna unit is a driving factor in the SKA antenna design. As the antenna cost goes up, the sensitivity, the resolution and the frequency range all go down if the budget is fixed. The proposed completion date for the SKA-low and SKA-mid is 2020. When they complete, the SKA instruments will provide the highest sensitivity and greatest resolution for attacking many important questions relating to our universe.

Another smaller project is the Frequency Agile Solar Radiotelescope (FASR), which is a solar radio array. It also has three sub-arrays called FASR-high, FASR-mid, and FASR-low. The FASR-high and mid are reflector arrays and FASR-low is a dipole array. The project will be finished in a few years.

7.3.6 Very Long Baseline Interferometer

The Very Long Baseline Interferometer (VLBI) is another very high-resolution aperture synthesis concept. It is also called the independent local oscillator interferometer. The antenna units of the VLBI are separated far from each other. Therefore, independent and highly stable oscillating signals are used to replace the normally used connected local oscillator signals. After the oscillating signal is mixed with the signal from the radio source, the information is preserved in a medium frequency band. This medium frequency information is recorded on DVD disks. The disks are used for final data processing through a computer. In the VLBI observation, there are several factors which will influence the relative time delay. These include the following terms:

$$\tau_t = \tau_g + \tau_c + \tau_i + \tau_p \tag{7.92}$$

where τ_g is the delay caused by the baseline, τ_c that caused by the synthesis time difference between two stations, τ_i that caused by instruments (amplifier, waveguide, cable, mixer, etc.), and τ_p that caused by atmosphere, ionosphere, and plasma regions in the intergalaxy area. Among these four factors, only τ_g is predictable. All other factors are variables of time. Therefore, in the VLBI observation, a very accurate clock should be used to reduce the influence of τ_c and τ_i. The influence from τ_p can be compensated by a phase closure method. Any three antennas will provide one phase closure check. The VLBI also requires wide band recording so that it can be scanned quickly during the correlation operation.

The baseline of some VLBIs reaches the diameter of the earth, providing resolutions up to 0.0001 arcsec depending on wavelength. Most VLBI networks use existing antennas to form interferometer arrays. However, the *u-v* coverage of existing antennas is usually poor. To avoid this, NRAO built a large-scale Very Large Baseline Array (VLBA) which includes ten antennas of 25 m diameter and covers a wide area from the border with Canada to the border with Mexico, from the Virgin Islands in the east to the Hawaii islands in the west. The effective baseline length is over 3,000 km.

In Europe, a European VLBI Network (EVN) was formed by joining antennas from Italy, Germany, the Netherlands, the UK, France, Sweden, Russia, and Poland. The diameters of these antennas range from 15 to 100 m. It forms a powerful imaging instrument. In Australia, a smaller project is the Australia Telescope (AT). It is a combination of aperture synthesis telescope and the VLBI. It includes a 6 km baseline aperture synthesis telescope with six antennas of 22 m diameter and two separated antennas of 64 m diameter. In China, four antennas form a VLBI. The Chinese one is geographically in between the three major VLBI arrays of the US, Europe, and Australia. It can join with any one of these VLBIs. The millimeter wavelength VLBIs have also been realized in recent years. They have the highest angular resolution of about 5×10^{-5} arcsec.

7.3.7 Space Radio Interferometers

The longest baseline of the ground-based VLBIs is limited by the earth's diameter. For achieving even higher angular resolution, it is necessary to send radio antennas into space orbit. The resolution of the space VLBIs can be 10–1,000 times those of the ground-based ones. Early in August 1986, a space VLBI experiment was performed using two 4.9 m antennas of the Tracking and Data Relay Satellite System (TDRSS). One of the antennas was pointed to a radio source and the other antenna was pointed to a satellite which performed observation of the same source and sent signals back to the earth. The purpose of the latter TDRSS antenna is to define the relative position between the earth and the satellite. In this experiment, the baseline achieved is 1.4 times the earth's diameter. The frequency used in this experiment is 2.3 GHz. In 1987, the frequency of the space radio interferometer reached 15 GHz. In 1997, a Japanese 8 m deployable antenna for the VLBI Space Observatory Program (VSOP) was sent into earth orbit. The VSOP worked only at 1.6 and 5.0 GHz since the planned high frequency receiver was damaged during launch. An improved 9 m VSOP-II antenna is now under production and is planned to be launched in 2010. This antenna will work at the higher frequencies of 8, 22, and 43 GHz.

NASA and ESA had planned a space radio telescope project, UASAT, an umbrella-type antenna of 15 m. This antenna will perform space VLBI work with existing ground-based antennas. Unfortunately, the project was rejected and it became a new project, the International VLBI Satellite (IVS). The IVS has wide spectral coverage from 4.5 to 120 GHz and has a 20 m class space antenna. A Russian planned space VLBI program is Radioastron. This is a 10 m space deployable radio antenna working at 0.3–25 GHz. This program was delayed but has been rescheduled for launch in 2009.

For space radio antenna design, the structure weight is a primary consideration. For reducing the structure weight, a number of methods have been tried. These include using CFRP material, using a prestressed umbrella structure, or using an air-inflated thin film structure. The umbrella structure has a low prestress level so that the stiffness is low. If radial prestress is introduced on the edge towards the center, the stiffness improves and the structure becomes very stable.

Using an air-inflated thin film structure to form a parabolic surface, the thickness of the thin film should satisfy a formula of $t = t_0(1 + 0.42r/R)$. However, it is not so easy to meet this thickness requirement. Air-inflated thin film structures also have problems in their deployment. A decade ago, a US air-inflated thin film space antenna failed to open after reaching orbit. The project terminated. Advanced Radio Interferometry between Space and Earth (ARISE) is a NASA inflatable telescope with a diameter of 25 m. Its launch date was in 2008, but it had delayed. For very large space radio telescopes, deployable or robot-assembled truss structures are necessary.

References

Cheng, J.. and Chiew, S. P., 1994, Structural aspects of steerable parabolic antenna design, J Inst Eng, Singapore, 34 (5), 47.
Cheng, J. and Humphries, C. M., 1982, Thin mirrors for large optical telescopes, Vistas in Astronomy, Vol. 26, pp 15–35.
Cheng, J. and Mangum, J., 1998, Feed leg blockage and ground pickup for Cassegrain antennas, ALMA memo 197, NRAO.
Cheng, J., 1984, Steerable parabolic antenna design, Ph D thesis, University of Wales.
Christiansen, W. N. and Hogbom, J. A., 1985, Radiotelescopes, Cambridge University Press, Cambridge, UK.
Cornwell, T. J. and Fomalont, E. B., 1999, Self-calibration, In: Synthesis imaging in radio astronomy II, ed Taylor, G. B. et al., ASP Conference, 180.
Emerson, D., 2005, Lecture notes of NRAO summer school on radio interferometry.
Fomalont, E. B. and Perley, R. A., 1999, Calibration and editing, In Synthesis imaging in radio astronomy II, ed. Taylor, G. B. et al., ASP Conference, 180.
Gawronski, W., 2007, Control and pointing challenges of large antennas and telescopes, IEEE Trans. Control Syst. Technol., 15, 276–289.
Goldman, M. A., 1996, Ball retro-reflector optics, GBT memo 148, NRAO.
Hirst, H. and McKee, K. E., 1965, Wind forces on parabolic antennas, Microw J, Nov, 43–47.
Kelleher, K. S. and Hyde, G., 1984, Reflector antennas, in Antenna Engineering handbook, eds. Johnson, R. C. and Jasik, H., McGraw Hill Inc, New York.
Kogan, L., 1997, Optimization of an array configuration minimizing side lobes, MMA memo 171, NRAO.
Korolkov, D. V. and Pariiskii, Yu. N., 1979, The Soviet Ratan – 600 Radio telescope, sky and telescope, Apr. 324–329.
Kraus, J. D., 1986, Radio Astronomy, Cygnus-Quasar Books, Powell, Ohio.
Lamb, J. W. and Olver, A. D., 1986, IEE Proc., 133, 43.
Lamb, J. W., 1998, Best fit focus for distorted paraboloid, internal report, Ovens Valley Radio Observatory.
Lamb, J. W., 2001, Formulas on antenna path length change caused by small positional changes of reflector or feed, Ovens Valley Radio Observatory.
Lee, K. F., 1984, Principles of antenna theory, John Wiley & Sons, New York.
Levy, R., 1996, Structural engineering of microwave antennas, IEEE Press, New York.
Meeks, M. L.(ed.), 1976, Astrophysics, Part C: radio Observations, Academic Press, New York.
Norrod, R. D., 1996, On possible locations for a GBT quadrant detector, GBT memo 143, NRAO.
Payne, J. and Parker, B., 1990, The laser ranging system for the GBT, GBT memo 57, NRAO.
Rohlfs, K., 1986, Tools of radio astronomy, Springer-Verlag, New York.
Ruze, J., 1966, Antenna tolerance theory – a review, Proc. IEEE, 54, 633.
Ruze, J., 1969, Small displacement in parabolic reflectors, MIT Lincoln Lab Report.
Setti, G. and Wielebinski, R., 1988, European consortium for very long baseline interferometer, an advance science and technology network, Radiosterrenwacht, The Netherlands.
Smithers, T., 1981, The design of homologically deforming cyclically symmetric structures, Ph. D. thesis, Darwin College, Cambridge.

Thompson, A. R., 1999, Fundamentals of radio interferometry, in Synthesis imaging in radio astronomy II(edited by Taylor, G. B. et al.), ASP Conference Series V180.
Thompson, A. R., Moran, J. M. and Swenson, G. W. Jr., 2001, Interferometry and synthesis in radio astronomy, 2nd edn, John Wiley & Sons Inc., New York.
Ulich, B. L., 1976, Optimum radio telescope geometry, NRAO internal report No. 2.
Von Hoerner, S., 1967, Design of large steerable antennas, Astro. J., 72, 35.
Von Hoerner, S., 1967, Homologous deformations of steerable telescope, Proc. ASCE ST5, 461–485.
Von Schooneveld, C., ed., 1979, Image formation from coherence functions in astronomy, (Astrophysics and space science library) D. Reidel Pub. Company, Dordrecht, Holland.
Whiteoak, J. B., 1987, The Australia telescope project: going along nicely, thank you, Aust. J. Astron., Vol. 2, No. 2, p. 54–55.
Whitney, A. R., 1977, Very long baseline interferometer for geology and astronomy, Shanghai Observatory.
Xiang, D.-L., 1986, Introduction of radio astronomical method, Purple Mountain Observation.
Zarghamee, M. S., 1967, On antenna tolerance theory, IEEE Trans AP, AP-15, 777.
Zhang, D.-Q., 1985, Fundamentals of Microwave antennas, Press of Beijing Institute of Technology, Beijing.

Chapter 8
Millimeter and Submillimeter Wavelength Telescopes

The design of open-air millimeter and submillimeter wavelength telescopes brings a new challenge to telescope designers. In this chapter, the thermal environment of outdoor antennas and its effect on the antenna surface and pointing errors are discussed in detail. The antenna panel and backup structure designs are discussed herein. In order to achieve superior performance of antenna backup structure, new CFRP material, structure insulation, and forced ventilation are all used in millimeter wavelength telescopes. In this chapter, the reactionless chopping secondary mirror, sensors, and the metrology system are also discussed. Special sections are added for the discussion of CFRP material properties, the shape changes of CFRP aluminum honeycomb sandwiched structures, and joint performance for CFRP-metal components. The readers can acquire basic knowledge of these specialized fields. In the last part of the chapter, antenna surface holographic measurement, surface panel adjusting strategy, and quasi-optics and broadband planar antennas are also discussed. Lightning protection and active optics for the millimeter wavelength telescopes are also mentioned. The contents of this chapter are also important for extremely large optical telescope design.

8.1 Thermal Effects on Millimeter Wavelength Telescopes

Since the 1980s, significant developments in receiver technology and in antenna manufacture have been achieved. These developments brought a new wave of millimeter and submillimeter wavelength telescope projects. These include the UK and Netherlands 15 m James Clerk Maxwell Telescope (JCMT), the IRAM 6×15 m millimeter wavelength interferometer, the IRAM 30 m millimeter wavelength telescope, the Japanese 45 m millimeter wavelength telescope, the German and the US 10 m Heinrich Hertz Telescope (HHT), the Swedish and ESO 15 m Submillimeter Telescope (SEST), the 10.4 m Caltech Submillimeter

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_8,

Observatory (CSO) telescopes, the 12 m Atacama Pathfinder Experiment (APEX) telescope, the 10 m South-Pole Submillimeter Telescope (SPST), and the 50 m Large Millimeter Telescope (LMT) of Mexico and the US. In addition, the GBT works at the longer millimeter wavelengths. At the same time, the Berkeley Illinois Maryland Association (BIMA) array, the Millimeter Array at Owens Valley, and the SubMillimeter Array (SMA) have been built. The two Californian arrays are now joined together as a Combined Array for Research in Millimeter-wave Astronomy (CARMA). The Atacama Large Millimeter Array (ALMA) of the US, Europe, and Japan, and the Caltech Cornell Atacama Telescope (CCAT) are now under construction.

8.1.1 Characteristics of Millimeter Wavelength Telescopes

From antenna tolerance theory, the millimeter and submillimeter wavelength telescopes demand a very tight surface rms error, between 15 and 25 μm (Figure 8.1). This tight requirement makes these telescopes different in design from the radio antennas discussed in the past two chapters. These millimeter and submillimeter wavelength telescopes work at very much higher frequencies and they usually are in an open-air environment.

The beamwidth, even for a medium size antenna (10–20 m) in this wavelength regime, is only a few tens to a few arcsec. The main lobe of an antenna is approximately a Gaussian function $f(x) = exp[-4\ln(2)x^2]$, where x is the Half Power Beam Width (HPBW). From this, a 1/5th HPBW pointing error will produce a 10% gain loss, so that the pointing requirement for

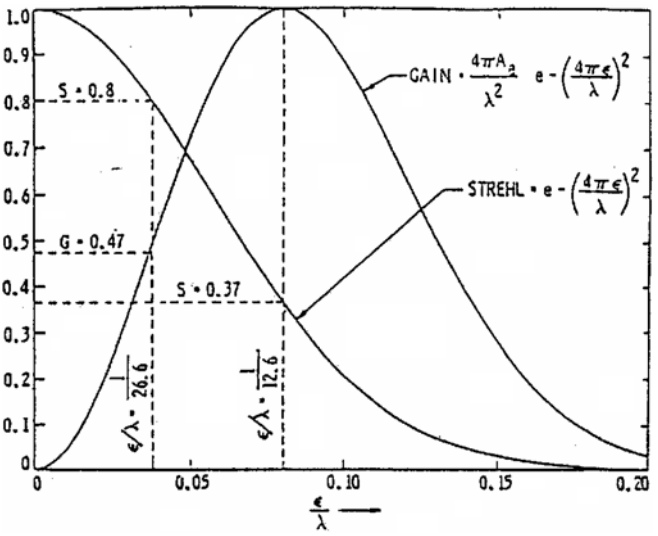


Fig. 8.1. Antenna gain and Strehl ratio as functions of surface error to wavelength ratio.

millimeter and submillimeter wavelength telescopes is also critical, near to or below 1 arcsec. The pointing performance of an optical telescope can be improved using a star guiding mechanism and the telescope is usually inside a dome. This is not the case for millimeter or submillimeter wavelength telescopes. The radio point sources on the sky are sparse, and antennas are constantly exposed to solar heating as well as wind loading. The lack of two-dimensional feed arrays also makes radio star guiding difficult.

Tight surface and pointing tolerances and serious solar and wind loadings make millimeter and submillimeter wavelength telescope design more difficult than that of other long-wavelength radio telescopes. The telescopes generally require a thermally stable and insensitive backup structure, precision reflector panels, accurate pointing and tracking control, a sophisticated metrology system, and a reactionless chopping subreflector. In this chapter, these requirements are discussed in detail.

Figure 7.6 shows natural limits defined by Von Hoerner (1967) for antenna design. Large and accurate millimeter or submillimeter wavelength telescopes have surpassed the thermal limit of steel material by applying following techniques: (a) using low coefficient of thermal expansion (CTE) materials, such as CFRP (carbon fiber reinforced polymer) or Invar; (b) using an astro-dome or radome to shield from the sun and wind; (c) using forced ventilation, controlled heating, or low time constant structural design; and (d) using active surface control.

For millimeter and submillimeter wavelength antennas, lower CTE CFRP and Invar are widely used. The CFRP material has been used in the HHT, the SMA, and the ALMA backup structures. The JCMT and HHT are inside astro-domes. Some 13.7 m antennas are inside radomes. The BIMA, the IRAM, and the ALMA-J antennas use forced ventilation in their backup structure. The Owens Valley antennas use a low thermal time constant, thin tube open backup structure where the temperature gradient is reduced through natural air ventilation. The GBT is the only one where an active surface control was tried.

Radomes are spherical weatherproof structures with membranes to protect antennas from solar heating and wind loading. Radomes have been used for a number of millimeter wavelength telescopes. However, the radomes have disadvantages such as: (a) blockage if metal frames are involved. The dome membrane absorbs the radiation in millimeter and submillimeter wavelengths. The solar light on the membrane also produces thermal noise; (b) the membrane absorption increases with the increase of frequency; and (c) reflection from the membrane is a periodic function of the wavelength. All these produce a gain loss of 10–20%. They also add significant background noise. Figure 8.2 shows the noise level of a radome for a 13.7 m diameter antenna. The gain loss plus the radome noise results in a lower overall antenna efficiency. Therefore, there is no radome for any newly proposed millimeter and submillimeter wavelength antennas. The CCAT uses an astro-dome which has a circular opening avoiding these disadvantages.

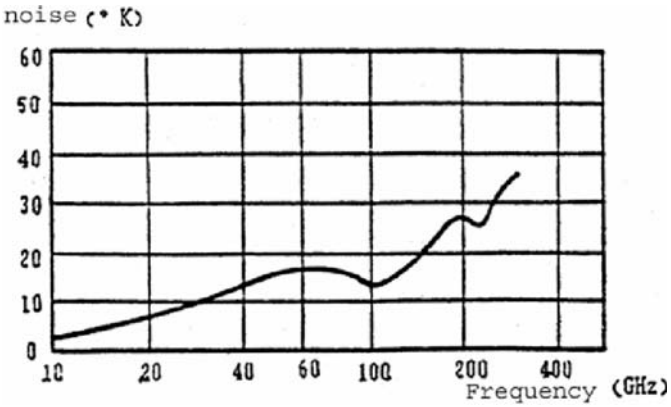


Fig. 8.2. Noise level produced by a 13.7 m antenna radome.

8.1.2 Thermal Conditions of Open Air Antennas

The air around an antenna can be assumed to be an infinite heat reservoir. The air temperature is an important factor which influences the antenna temperature. The air temperature can be represented as (Greve et al., 1992):

$$T_A(t) = T_{A0} - \delta T_A \cos[\omega(t - t_0)] \tag{8.1}$$

where $T_{A0} = <T_A(t)>$ is the daily average temperature, $\omega = 2\pi/24\,h$, δT_A the amplitude of temperature variation, and t_0, about 1–2 h, a time delay of the variation with respect to the middle day time. A typical range of the air temperature is between -20 and 30°C. The heat transfer between the air and antenna is mainly through convection.

Radiation exists between the sky and antenna. The sky temperature is represented by Swinbank's equation:

$$T_S(t) = 0.0553 T_A^{1.5}(t) \tag{8.2}$$

The visible sky temperature is also represented approximately by:

$$T_S(t) = T_A(t) - \delta T_S \tag{8.3}$$

where $\delta T_S \approx 15 - 20$°C. The visible sky temperature depends on water vapor and cloud condition. Heat transfer from the antenna to the sky peaks at the infrared region of wavelengths between 8 and 13 μm. The infrared sky temperature is lower, at about 50 K.

The ground is another heat reservoir. Its temperature is represented approximately by:

$$T_G(t) = T_{G0} - \delta T_G \cos[\omega(t - t_g)] \tag{8.4}$$

where $T_{G0} = <T_G(t)>$ is the daily average ground temperature and t_g, about 1–2 h, a time delay of the temperature variation with respect to the middle day time. In the formula, an ideal cosine form is used. The actual ground temperature expression is more complex. Using the air temperature to represent the ground temperature requires an offset:

$$T_G(t) = T_A(t) + \delta T_G \tag{8.5}$$

The ground radiation property varies as the composition changes of the ground. Typical absorption and emission coefficient of the ground is about 0.6–0.7.

For out-door antennas, the sun is an important heat source. It produces structural thermal deformation. Solar radiation measured from low earth orbit is roughly 1,366 watts per square meter, though it fluctuates by about 6.9% during a period of a year (1,412 W/m^2 in early January and 1,321 W/m^2 in early July) due to the distance variation between the earth and the sun. By considering the atmospheric absorption, the solar radiation at a high site is about $S_0 = 1{,}290$ W/m^2. The relationship between the solar radiation and the elevation angle β is:

$$S(t) = S_0(1 + d)e^{-B/\sin\beta(t)} \tag{8.6}$$

where $d \approx 0.05$ is the diffusivity and $B = 0.1$ the transparency factor. The elevation angle $\sin\beta = \cos H\cos\delta\cos\varphi + \sin\delta\sin\varphi$ can be calculated from the hour angle H, the declination δ of the sun, and the site latitude φ.

8.1.3 Heat Transfer Formulae

Heat transfer occurs through conduction, convection, radiation, and any combination of these. Conduction is the transfer of thermal energy from a region of higher temperature to a region of lower temperature through direct molecular communication within a medium or between mediums, but without a flow of the material. The governing law of the heat conduction is:

$$q_c = -kAdT(x)/dx \tag{8.7}$$

where k is the conductivity, A the cross sectional area, T the temperature, and x the direction in which the heat flows from hot to cold.

Radiation is the transfer of heat through electromagnetic radiation. Hot or cold, all objects radiate energy at a rate equal to their emissivity times the rate at which energy would radiate if they were a black body. No medium is needed for radiation to occur; radiation works even in a perfect vacuum. The heat radiated by a blackbody is:

$$q_r = \sigma \cdot AT^4 \tag{8.8}$$

where A is the radiation area and σ the Stefan–Boltzmann constant. Other surface parameters include e, the emissivity, and a, the absorptivity. If the radiation is between two objects with surface areas being F_1 and F_2, then:

$$\begin{aligned} q_{r1-2} &= \sigma \cdot e_1 a_2 (T_1^4 - T_2^4) F_1 \varphi_{12} \\ \varphi_{12} &= \frac{1}{\pi F_1} \int_{F_1} \int_{F_2} \frac{\cos\beta_1 \cos\beta_2}{s^2} \, df_1 df_2 \end{aligned} \tag{8.9}$$

where φ_{12} is a view factor. For radiation between two infinite long parallel plane surfaces, the above formula becomes:

$$q_{r1-2} = \sigma (T_1^4 - T_2^4) F[(1/e_1) + (1/e_2) - 1] \tag{8.10}$$

Convection is a combination of conduction and the thermal energy transfer by fluid or gas circulation where hot particles move to cooler areas. Unlike the case of pure conduction, currents in fluid or gas are involved either into a fluid or gas or within a fluid or gas. In most antenna cases, the flow currents occur in air.

In convection, a heat transfer coefficient or convection coefficient, h, is used. Unlike thermal conductivity, the heat transfer coefficient is not a material property. The heat transfer coefficient depends upon the geometry, fluid, temperature, velocity, and other characteristics of the system in which convection occurs. The heat transfer coefficient must be found experimentally for every system analyzed. Using the heat transfer coefficient, the general formula for convection is:

$$q_{cv} = Ah(T_s - T_a) \tag{8.11}$$

where T_s is the surface temperature, T_a the nearby fluid or gas temperature, and A the area. In heat convection, fluid dynamics, and wind resistance calculations, some dimensionless parameters are important. These are Reynolds, Prandtl, Grashof, and Nusselt numbers defined in air as:

$$\begin{aligned} \mathrm{Re} &= \frac{VL}{v} = \frac{\rho VL}{\mu} \\ \mathrm{Pr} &= \frac{C_p \mu}{\kappa} = \frac{v}{\alpha} \\ Gr &= g\beta \frac{(T - T_\infty)L^3}{v^2} \\ Nu &= \frac{VL}{v} = \frac{hL}{\kappa} \end{aligned} \tag{8.12}$$

where V is the air flow velocity, L the length of the surface, ρ the air density, v the kinetic viscosity of the air, $\mu = v\rho$ the absolute viscosity, C_p the specific heat, α the thermal diffusivity, g the gravitational acceleration, $(T - T_\infty)$ the temperature difference between the surface and the nearby air, β the expansion coefficient of the air, and κ the conductivity of the air. The expansion coefficient of the air is the reciprocal of absolute temperature $\beta = 1/T(air)$. Among these numbers, the Nusselt number is directly related to the convection coefficient, h. Table 8.1 lists some basic parameters for air at different temperatures and Figure 8.3 shows the air density change with altitude.

Convection includes natural and forced ones. Forced convection involves both laminated and turbulence flow. In natural or free convection, air surrounding a

Table 8.1 Some air parameters at different temperatures

Temperature T(°C)	ρ [kg/m^3]	C_p [J/kg K]	$\alpha\ 10^{-6}$ [m^2/s]	κ W/m [K]	$v\ 10^{-6}$ [m^2/s]	$\mu\ 10^{-6}$ [Ns/m^2]	Pr
0	1.252	1011	19.2	0.0237	13.9	17.403	0.72
20	1.164	1012	22.0	0.0251	15.7	18.275	0.71
40	1.092	1014	24.8	0.0265	17.6	19.219	0.71
60	1.025	1017	27.6	0.0279	19.4	19.885	0.70

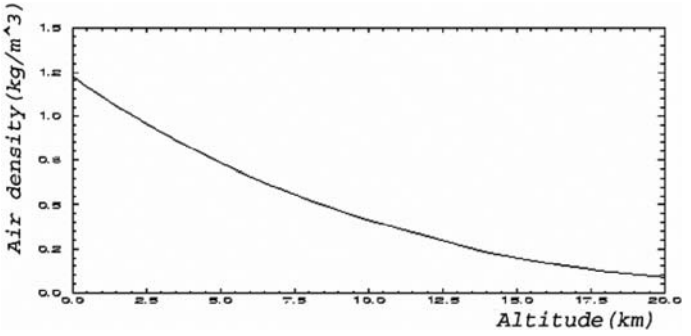


Fig. 8.3. Relationship between air density and altitude.

heat source receives heat, becomes less dense and rises. The surrounding, cooler air then moves to replace it. This cooler air is then heated and the process continues, forming a convection current. Forced convection, by contrast, occurs when wind, fans or other means are used to move the air and create an artificially induced convection current. From the average Nusselt number over a surface, an average convection coefficient including both free and forced convection can be derived:

$$\overline{\mathrm{Nu}} = \frac{\bar{h}_c L}{\kappa} \tag{8.13}$$

For natural convection without wind or fans, if the surface of a plate faces up horizontally and $10^4 \leq Gr\,\mathrm{Pr} \leq 10^7$, then (Incropera and De Witt, 1990):

$$\overline{\mathrm{Nu}} = 0.54(Gr\,\mathrm{Pr})^{1/4} \tag{8.14}$$

When $10^7 \leq Gr\,\mathrm{Pr} \leq 10^{11}$, then:

$$\overline{\mathrm{Nu}} = 0.15(Gr\,\mathrm{Pr})^{1/3} \tag{8.15}$$

If the surface of a plate faces down horizontally and $10^5 \leq Gr\,\mathrm{Pr} \leq 10^{10}$, then the Nusselt number for natural convection is:

$$\overline{\mathrm{Nu}} = 0.27(Gr\,\mathrm{Pr})^{1/4} \tag{8.16}$$

When a horizontally placed plate is under a stable air flow and $\mathrm{Re} \leq 5 \times 10^5$, then the Nusselt number for a forced convection under a laminar air flow is:

$$\overline{\mathrm{Nu}} = 0.664\ \mathrm{Re}^{0.5}\ \mathrm{Pr}^{1/3} \tag{8.17}$$

The transition from a laminar flow to a turbulent one occurs at a certain point from the edge for a strong wind or air current. The distance between the edge and the transit point is named the critical distance. At this distance, the convection is a mixture of both laminar and turbulent flow and the Nusselt number is:

$$\overline{\mathrm{Nu}} = 0.036(\mathrm{Re}^{0.8} - 23200)\ \mathrm{Pr}^{1/3} \tag{8.18}$$

However, in reality, due to variation of wind or air vibration, the distance between the edge and transit point becomes so short that the whole plate surface is almost under turbulent air flow. In this case and $5 \times 10^5 < \mathrm{Re} < 5 \times 10^7$, the Nusselt number is:

$$\overline{\mathrm{Nu}} = 0.036\ \mathrm{Re}^{0.8}\ \mathrm{Pr}^{1/3} \tag{8.19}$$

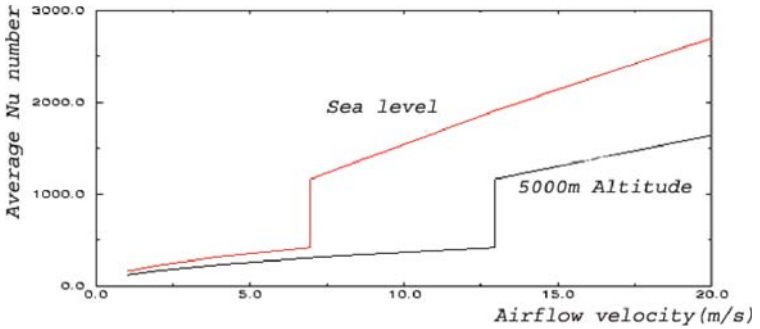


Fig. 8.4. The average Nusselt number changes with air velocity on the mountain top (5,000 m) and at sea level (Cheng, 1998).

The Nusselt number of air is also a function of air density which varies with altitude. This affects the transition wind speed between laminar and turbulent flows both at sea level and on the mountain top. At sea level, the transition occurs at about 7 m/s, while on the 5,000 m mountain top, it is about 13 m/s for a 1 m long plate (Figure 8.4).

The sudden transition from laminar to turbulent flow has an important impact on the cooling efficiency at the mountain top (Cheng, 1998). The cooling efficiency of a laminar flow is lower than that of a turbulent one. Because of the transition wind speed change, the cooling efficiency at the mountain top is much lower than that at sea level. Therefore, all heat producing equipment has to be de-rated at the mountain top. Some equipment can only achieve an efficiency of 22% compared with when it is at sea level. Table 8.2 lists the estimated efficiency loss of some electrical equipment or components at high altitude when forced air cooling is involved.

The thermal time constant is a parameter widely used in the heat transfer process. The thermal time constant is the time interval needed for a system to change from one state to a fraction of another state. For a first-order linear system, the step response can be expressed by the difference between unity and

Table 8.2. Percentage of efficiency loss for electrical equipment used in high mountainous areas

	2,000 m	3,000 m	4,000 m	5,000 m
Generator	25%	40%	55%	70%
Compressor	35%	55%	75%	95%
Vacuum pump	20%	30%	40%	50%
Transmission line	10%	20%	30%	40%
Transformer	5%	15%	25%	35%
Motor	5%	15%	25%	35%

an exponential function with a minus power. The system time constant is the time interval where the system achieves 63% of the final state. Four times the system time constant achieves a fraction of 98.2% of the final state. In heat transfer, this four times the system time constant is named the thermal time constant.

Heat conduction is a first-order linear dynamic system. The governing equation in one dimension is $\partial T/\partial t = (k/\rho C_p)\partial^2 T/\partial x^2$; this equation could be solved by separation of variables. By applying certain boundary conditions, the general solution of the time needed to reach the same temperature on both sides of a wall with a thickness W is:

$$T(x,t) = \sum_{n=1}^{\infty} C_n \sin\left(\frac{n\pi \cdot x}{W}\right) e^{-(k/\rho C_p)(n\pi/W)^2 t} \tag{8.20}$$

Then, the thermal time constant for conduction is:

$$\tau = 4\rho C_p W^2/\kappa\pi^2 \tag{8.21}$$

where ρ is the density, C_p the heat capacity or specific heat, and k the conductivity. In some references, the thermal time constant is also expressed as the square of the thickness divided by the thermal diffusivity of a material. Thermal diffusivity is a function of the density, heat capacity, and conductivity. For convection or radiation, the thermal time constant of a rod is $\tau = \rho C_p A/hC$, where h is the convection or radiation coefficient, A the cross sectional area, and C the circumferential length.

8.1.4 Panel Thermal Design

Reflector panels are important components for radio telescopes. In short wavelengths, the required surface rms accuracy should be about 10 to 30 μm. Errors of an antenna surface come from two sources: the panels and the backup structure. The backup structure design is discussed in Sections 8.1.5 and 8.2.2. In this section and Section 8.2.1, panel design is discussed.

Factors that influence panel surface accuracy include panel manufacture, gravity, wind, thermal-induced deformation, and panel adjusters. Among these factors, thermal-induced deformation is the most important one. Two thermal conditions exist for antenna panels: absolute temperature change and temperature gradient. When the overall temperature changes, we have an absolute temperature case and when the temperature of one part is different from the other, we have a temperature gradient case.

8.1.4.1 Absolute Temperature Error

For an antenna dish made of only one material, an overall temperature change will only produce a focal length change which can be compensated by adjusting

Table 8.3. Thermal properties of some materials

Material	CTE α ($\times 10^{-6} C^{-1}$)	Conductivity κ (mC^{-1})	α/κ ($\times 10^{-1}$ m^{-1})
Aluminum	23	156	1.5
CFRP	0∼5	4.2	0∼12
Steel	12	52	2.3
Invar	0.9	16	0.56

the secondary mirror position. However, if the panel and the backup structure are fixed together and are made of different materials, an absolute temperature change will produce a panel surface error. If the CTEs of the panel and backup structure are α_p and α_b, respectively, then the panel surface rms error at a temperature different from the initial one, T_0, will be (Lamb, 1992):

$$\varepsilon = \frac{|(T_p - T_0)\alpha_p - (T_b - T_0)\alpha_b|}{8\sqrt{3}R_0} {d_p}^2 \tag{8.22}$$

where T_p and T_b are the panel and backup structure temperatures after the temperature change, R_0 the radius of curvature of the panel, and d_p the dimension of the panel. Thermal properties of some structural materials are listed in Table 8.3. For an aluminum panel of 1 m dimension with a radius of curvature of 7 m fixed on a steel backup structure, the absolute temperature-induced surface error will be about $\varepsilon \approx 0.11$ μm/°C. Considering the range of temperature change for antennas, this absolute temperature-induced error is serious. This is the reason why flexible panel adjusters are used between the panels and the backup structure for all millimeter and submillimeter wavelength antennas.

Even using flexible panel adjusters, the expansion or contraction of the aluminum panel produces bending moments on the adjusters. These bending moments will also cause panel surface error. The softer the adjuster is, the smaller the bending moment and the induced surface error will be.

8.1.4.2 Temperature Gradient Error

Under solar radiation, a temperature gradient will arise between the panel top and bottom surfaces. This temperature gradient also produces a curvature change of the panel. Assuming the original panel is a flat one, the curvature radius induced by a temperature gradient will be:

$$c = \frac{\Delta T \alpha_p}{t} \tag{8.23}$$

where t is the thickness of the panel, ΔT the temperature difference between the top and bottom surfaces, and α_p the CTE of the panel material. The panel front

and back temperature difference can be calculated from the heat flux density, which passes through the panel, and the thermal conductivity:

$$\Delta T = \frac{It}{\kappa} \tag{8.24}$$

where I is the heat flux density and κ the thermal conductivity. From the above two formulas, the temperature gradient-induced curvature radius is irrelevant to the panel thickness as well as the front and back temperature gradient. It is only a function of the heat flux passing through the panel (Lamb, 1992):

$$c = \frac{\alpha_p I}{\kappa} \tag{8.25}$$

To derive the heat flux passing through a panel, it is necessary to produce a thermal model. Under a stable thermal condition, the total input heat flux equals the total output heat flux. The input heat flux, W_s, comes from the sun and the outputs include heat radiated to the sky, W_r, the heat through air convection, W_c, and the heat passing through the panel, W_i:

$$W_s = W_r + W_c + W_i \tag{8.26}$$

For an unpolished aluminum surface, the absorptivity is the same as its emissivity, ~0.25. If the air temperature is 300 K, the sky temperature in the infrared region is 50 K, and the temperature of the panel top surface is 10 K above the air temperature, 300 K, then:

$$\begin{aligned} W_s &= 0.25 \times 1200 = 300\,\mathrm{W/m^2} \\ W_r &= 0.25 \times 5.67 \times 10^{-8} \times (310^4 - 50^4) = 130\,\mathrm{W/m^2} \end{aligned} \tag{8.27}$$

For deriving the convection heat transfer coefficient, it is necessary to calculate the following dimensionless parameters of the air:

$$\begin{aligned} \mathrm{Re} &= \frac{VL}{v} = \frac{\rho VL}{\mu} = \frac{1.164 \times 6 \times 8}{18.24 \times 10^{-6}} = 3.063 \times 10^6 \\ Gr &= g\beta \frac{(T - T_\infty)L^3}{v^2} = 9.8 \times \frac{1}{300} \times \frac{10 \times 1}{(15.7 \times 10^{-4})^2} \\ &= 13.25 \times 10^4 \\ \mathrm{Pr} &= \frac{C_p \mu}{\kappa} = \frac{v}{\alpha} = 0.71 \end{aligned} \tag{8.28}$$

For natural convection, if the panel surface is in a horizontal position and $10^4 \leq Gr\,\mathrm{Pr} \leq 10^7$, then (Incropera and De Witt, 1990):

$$\begin{aligned} \overline{\mathrm{N}}\mathrm{u} &= 0.54(Gr\,\mathrm{Pr})^{1/4} = 9.457 \\ h_{c1} &= \overline{\mathrm{N}}\mathrm{u}\,\kappa/L = 9.457 \times 0.0251/1 = 0.237\,\mathrm{W/m^2K} \end{aligned} \tag{8.29}$$

For forced convection, if $5 \times 10^5 < \mathrm{Re} < 5 \times 10^7$ under turbulent flow, then:

$$\begin{aligned} \overline{\mathrm{N}}\mathrm{u} &= 0.036\mathrm{Re}^{0.8}\ \mathrm{Pr}^{1/3} = 4961 \\ h_{c1} &= \overline{\mathrm{N}}\mathrm{u}\,\kappa/L = 4961 \times 0.0251/8 = 15.5\,\mathrm{W/m^2K} \end{aligned} \tag{8.30}$$

Using an average of the above two convection coefficients, 8 W/m^2 K, the heat flux through the panel is between 100 and 130 W/m^2. For a 1 m size aluminum flat panel, the induced surface rms error is about 1.5 μm.

If back rib stiffeners exist on the panel, the surface error induced would be twice larger. If the aluminum panel plate is glued to its ribs as used in early antennas, then the temperature gradient-induced surface rms error will increase significantly due to poor glue conductivity. If the panel is an aluminum sandwiched structure with a core made of aluminum honeycomb and the core density is 1:10–1:80, then the surface rms error would be 10–80 times the above value. This results in a surface rms error of 15–120 μm.

For the CFRP sandwiched panel with an aluminum honeycomb core, analysis becomes difficult. The CTE of this type of panel is about $4.5 \cdot 10^{-6}$ K^{-1}. Using the conductivity of aluminum, its $\alpha/\kappa = C \cdot 2.9 \cdot 10^{-8}$ m/W, where C is the cross sectional density of aluminum honeycomb, the surface rms error is about $0.3C$ μm. If the honeycomb is made of CFRP material and its CTE is zero, then the surface rms error is very small. However, if the CTE of the CFRP is not zero, but $2 \cdot 10^{-6}$ K^{-1}, then the surface rms error is $2C$ μm, a relatively large number for millimeter wavelength antennas.

Generally, machined aluminum and nickel sandwiched panels have better thermal stability.

In millimeter or submillimeter wavelengths, paint applied on top of panels is not an option due to the absorption from the paint. However, in long wavelength radio regions, white paint is used on top of panels to lower the absolute temperature and temperature gradient of panels. In the visible region, white paint has a lower absorption coefficient while, in the infrared region, it serves as a blackbody with a higher emissivity. Therefore, the heat flux through the reflector surface is greatly reduced.

8.1.5 Backup Structure Thermal Design

A structure of height h, width w, and CTE α will produce a small tilt β when two sides of it have a temperature difference Δt:

$$\beta = \frac{\alpha h \Delta t}{w} \tag{8.31}$$

For a steel structure, if the height is the same as the width, the tilt angle produced is $\beta = 2.5\Delta t(\text{arcsec})$. Solar radiation is the reason for the temperature difference. Heat conduction may equalize the temperatures on both sides of a structure; however, a time interval is required for the heat to flow from one side to the other. This thermal time constant of the structure is:

$$\tau = 4\rho C_p w^2/\kappa\pi^2 \tag{8.32}$$

where C_p is the heat capacity, ρ the density, and k the conductivity. If the width of a steel structure is 4 m, the time for a thermal balance is about 3.3 days. This shows that passively to achieve a temperature balance for a practical antenna structure is impossible. A temperature gradient always exists inside antenna structures.

Inside a backup structure, three typical temperature gradients are: (a) axial, (b) side-by-side, and (c) radial ones (Figure 8.5). The axial temperature gradient only produces a simple focus change. This focus change is:

$$\frac{\Delta F}{F} = \frac{2F\alpha\Delta t}{d} \tag{8.33}$$

where d is the height of the backup structure and F the focal length. The focus change will not affect the surface error. However, if the refocusing rate is slower than the temperature change, the defocusing produces an equivalent surface rms error as:

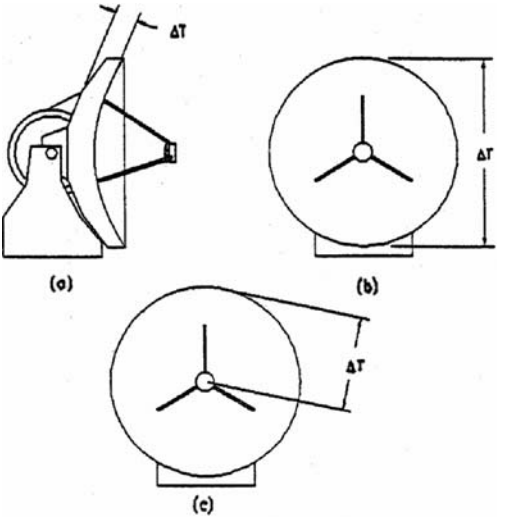


Fig. 8.5. Three ideal temperature gradients on the antenna backup structure (Lamb, 1992).

$$\varepsilon_{eff} = \frac{0.02}{(F/D)^2}\Delta F \tag{8.34}$$

If the focal length is $F = 3.2$ m, the diameter $D = 8$ m, the backup structure height $d = 1$ m, and $\alpha = 12 \cdot 10^{-6}\ \mathrm{K}^{-1}$, then $\Delta F = 250$ μm/K. When the antenna does not refocus, the equivalent surface rms error is $\varepsilon_{\mathrm{eff}} / \Delta t = 30$ μm/K.

The side-by-side temperature gradient produces a simple pointing error:

$$\Delta\theta = \frac{\alpha d \Delta t}{2D} \tag{8.35}$$

Using the above backup structure data, the pointing error produced is about $\Delta\theta/\Delta t = 0.75$ arcsec/K.

The radial temperature gradient produces both the focus change and surface error. The focus change is:

$$\frac{\Delta F}{F} = \frac{10 F d \alpha \Delta t}{D^2} \tag{8.36}$$

And the surface error is:

$$\varepsilon = \frac{\alpha d \Delta t}{40\sqrt{2}} \tag{8.37}$$

For the same backup structure data, the focus change is $\Delta F / \Delta t = 18$ μm/K and the surface error is $\varepsilon/\Delta t = 0.2$ μm/K. If the antenna does not refocus in time, the equivalent surface error is $\varepsilon_{\mathrm{eff}}/\Delta t = 18$ μm/K.

All the above gradients are typical (ideal) ones for antenna structures. The real temperature distribution on an antenna is usually complicated. Very often, a small local temperature gradient produces larger antenna surface rms errors.

During the past 20 years, efforts have been made to measure the real temperature distribution on antenna structures. The measurement data shows that the rms temperature difference of the Owens Valley 10 m open-air, thin steel tube backup structures is about 1 K in breezy conditions, that of the IRAM 15 m antennas with a closed CFRP backup structure and some ventilation outlets is between 0.8 and 1 K at night time, and that of the BIMA steel backup structure with forced ventilation and sun shielding is between 0.6 and 1 K.

Analytical calculation shows that the temperature gradient-induced surface rms error for an 8 m steel backup structure antenna is about 23 μm. The pointing error is about 1.4 arcsec. The rates of the error changes are 14 μm/h and 2.5 arcsec/h. From these numbers, millimeter and submillimeter wavelength antennas may require the use of lower expansion CFRP materials for their backup structures.

Two different thermal effects also occur on the yoke structures. First, the temperature difference between two elevation bearing arms will produce a tilt of the elevation axis. Using Equation (8.31), if the height of the yoke is the same as

the width, then a pointing error of $2.5\Delta t$ arcsec is produced. If a steel yoke is covered by a polished aluminum surface with an air gap and foam in between, and taking the following parameters into consideration:

Solar radiation: 1,260 W/m^2
Total absorption coefficient of aluminum surface: 0.04
Thickness of steel structure: 0.025 m
Structural area factor: 2.4

The structural area factor is the area ratio between the structural area under the sun and the structure area where temperature increases. With a steel density of 7,800 kg/m^3 and a specific heat of 418 J/kg·K, the time required for a 1-K temperature increase of one side of the structure is only 2.2 h. This means that the elevation axis will have 0.57 arcsec pointing error every half hour. Direct antenna measurement shows the thermal-induced axis tilt is about 3–5 arcsec/day for this type of yoke structure.

Second, if the temperature difference is between two sides of one yoke arm, the effect is more serious than the first one. In this case, the height of the support arm is much larger than the width. Therefore, it will produce an even larger pointing error. Calculation shows that the pointing error from this can be tens of arcsec within a half hour interval. Therefore, a yoke without thermal insulation will not satisfy the millimeter and submillimeter wavelength requirement. For achieving an even higher pointing performance, an independently supported, thermally stable, low CTE reference structure may be necessary for very accurate millimeter and submillimeter wavelength antennas. The reference structure is inside the yoke and free from wind and other loadings. The reference structure provides reference for the encoder to correct any thermal or wind-induced pointing errors. The reference structures are in fact used in ALMA antennas.

Solar observation using precision antennas will have a large heat flux on the subreflector, feed leg, and the feed areas. In the millimeter and submillimeter wavelength region, one method to avoid the heat from the sun is through the use of special panels where fine triangular or circular grooves are machined or cast on the panel surface (Lamb, 1999b, 2000). Since most of the solar radiation falls in a wavelength range shorter than 4 μm, those relatively wide and shallow grooves can scatter away radiation energy, while the surface is still within Ruze tolerance criterion for the millimeter and submillimeter wavelength antennas. For triangular grooves, the depth should be less than or equal to $\varepsilon/0.289$ in order to meet the Ruze criterion, where ε is the rms surface error produced from these fine grooves. If the blaze angle is $\alpha = 5^\circ$, the width of the grooves is $w \leq 230$ μm. This is close to the diffraction limit of a submillimeter wave of $\lambda = 300$ μm. When the sun is on the dish axis, these grooves will scatter the solar radiation in an annulus of a width of subreflector diameter $d_s$and a radius of $2\alpha f$, where f is the focal length of the primary reflector. The reduction factor of the solar radiation is $d_s/(32\alpha f)$. Circular grooves can also reduce the solar radiation but with a smaller reduction factor.

However, panels with grooves will produce strong scattering. Sand blasting, chemical etching, or other treatment of the panel surface can also produce an

ideal diffusing pattern with a FWHP angle of about 45 degrees. It may reduce subtantially the heat flux on the subreflector and focus areas. The surface diffusivity and the bidirectional reflectance distribution function (BRDF) were discussed in Section 2.4.2. With the BRDF measurement from panel samples, the flux density on the subreflector surface and focal plane can be calculated (Schwab and Cheng, 2008). For a 12 m antenna, the flux intensity reflected from chemical etching panels on the subreflector is only about 4,000 W/m^2.

8.2 Structural Design of Millimeter Wavelength Antennas

8.2.1 Panel Requirements and Manufacture

Figure 8.6 shows the relationship between the rms surface error and the antenna efficiency over the operating wavelength axis. The figure requires an rms surface error smaller than 60 μm for antennas operating in the millimeter wavelength and smaller than 20 μm for antennas in the submillimeter wavelength. The atmospheric submillimeter wavelength window has a low wavelength limit between 0.45 and 0.35 mm. Of these reflector surface error requirements, a major part is from the panel surface errors.

Reflector panels for long-wavelength radio telescopes are usually fixed (through welding or through screws and nuts) to the top of the backup structure. The panel also stiffens the backup structure. However, the panels may deform during the panel assembly. The surface accuracy of this type of panels is limited. If materials of the panel and backup structure are different, the panel will have a larger deformation when temperature changes. This panel design is not suitable for millimeter and submillimeter wavelength antennas.

For millimeter and submillimeter wavelength telescopes, the panels are supported on the backup structure through adjusters. These adjusters are flexible in some directions so that thermal differential expansion will not deform

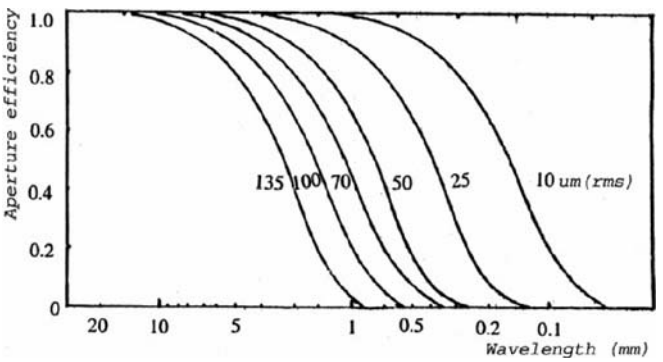


Fig. 8.6. Relationship between surface rms error and aperture efficiency.

the panel surface shape. At the same time, the position of panels can be accurately adjusted through its spring-loaded differential screws. The panel adjusting can be performed either from the back or the front of the reflector.

The major part of panel surface error comes from thermal deformation and panel manufacture. The panel thermal deformation is related to the panel design, while the manufacturing error is related to the production process.

In the millimeter and submillimeter wavelength region, there are a number of panel designs and panel production processes. One type of panel is made by stretching a thin aluminum plate on top of either a stiffening rib frame or shaped honeycomb stiffener. Other types of panels use a replicated CFRP surface or electroformed metal surface with machined or crush-shaped honeycomb stiffeners. If the panel surface is a very thin aluminum sheet, stretching on a mold is always necessary. The mold used is made of aluminum, steel, or composite material. Normally, one mold is needed for panels of one ring. A unique type of mold involves many adjustable screws on a rigid bed. On top of each screw, a spherical bearing supports a small flat surface. The mold top surface shape can vary by adjusting the height of the screws. Therefore, a single mold can be used for all panels of any reflector. If a high precision surface shape is required, a mold made of CFRP or Pyrex glass materials is necessary.

Panel surface errors also depend on the production process. However, the same panel accuracy can be achieved by different production processes. The determining factor of surface accuracy in the panel manufacture is the effort put into the process, not the manufacturing process itself. With more effort applied and more experiences gained, a better accuracy can be achieved. In most cases, the panel accuracy depends on the mold accuracy. Molds made of metal materials have limited accuracy; while molds made of Pyrex glass have a better accuracy. Variables that affect the panel's accuracy include the restoring stress and the surface contact between the mold and panel sheet. A vacuum is usually used to ensure a perfect surface contact between the mold and panel sheet.

Microwave antenna panels are typically made by stretching a thin aluminum sheet on top of a mold. The stiffening aluminum z-shaped rib formed frame is then glued on the back of the surface sheet. These aluminum ribs have many parallel slots near the stretched sheet so that they can bend easily to fit a curved panel shape. The length of these slots is about half of the rib's thickness. Since the glue used has poor conductivity, thermal deformations of this type of panel are larger. They are not accurate enough for millimeter and submillimeter wavelength telescopes.

Stretching a thin aluminum sheet on top of a mold stiffened with a crush-shaped aluminum honeycomb produces high accuracy millimeter wavelength panels. The HHT and JCMT used this type of surface panel.

Machined aluminum panels were used for the recent BIMA, ALMA-US, ALMA-J, and other antennas. Two types of the machined panels exist. One is machined from a cast aluminum blank and the other from a thick aluminum plate. The machined panels from cast aluminum are heavier in weight, but the machine time required is less. By using vacuum casting and an embedded metal

surface inside the cast mold, a fine panel surface can be cast. The area density of this type of panel is about 25 kg/m^2.

Panels directly machined from aluminum plates are costly. Computer-controlled machining is required. The machining time is longer. However, the rib wall (2 mm) and panel surface thickness (2.5 mm) can be very thin. Therefore, the panel area density is a lot smaller (15 kg/m^2). Some panels use a specially made machine for their front surface fine. The cost of these precision light-weight aluminum panels is higher, at \$3,000/m^2 in the year 2000. The accuracy of these panels are within $3 \sim 6$ μm (Figure 8.7).

The accuracy of some panels can be further improved by fine adjustment of their supporters. Most antenna panels use four adjusters. Some panels use five adjusters. These over-constrained panel supporters can bend the panel to fit a better surface shape. However, when the panel is in the cutting process, only three supports are used for defining the position of the panel. To avoid over-constraint, spherical washers are used. Damping is also necessary in the panel cutting process. Damping foam or thin lead plates can reduce vibration of the panels under machining. The cutting depth should be well controlled for the best surface accuracy. Temperature control during final cutting is important. Human appearance should be avoided during the cutting process. Some panel cutting uses a high speed cutter supported on an air bearing. Using a single point cutter, a special groove pattern can be made for defusing the solar heat (Lamb, 1999b). Chemical etching is another way to diffuse radiation. However, the itching and stress release after cutting may deform the surface shape. These deformations are usually complex and trial and error may be needed to improve the final panel accuracy.

Panels of replicated CFRP or electroformed nickel surfaces are also used in millimeter and submillimeter wavelength antennas. These surfaces are stiffened by honeycomb cores. The panels are very stiff, very accurate, and very stable under thermal loadings.

Besides the panel surface shape, the panel support and adjustment are also important. A good panel support system ensures accurate panel position, while it produces no thermal or other deformations on the panel. Spring-loaded differential screws are used in all the panel adjusters. Adjusters can be either

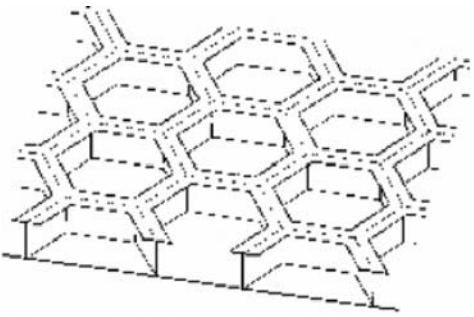


Fig. 8.7. The back side of a precision panel made from a thick aluminum plate.

manual- or motor-driven. Within one panel, only one rigid adjuster is allowed; other adjusters should be flexible in some directions, so that thermal stresses of the panel can be absorbed. Modern panel adjusting can use a digital coding system for its positioning recognition.

The measurement of the panel surface relies on computer-controlled three-dimensional measuring machines. The data from the measurement are fit to a best-fit paraboloid. If tape targets are glued on the panel surface, the surface shape can also be measured by using a photogrammetry technique. Photogrammetry was developed from aeronautic topological surveys. It uses projected images from different view positions to calculate the target coordinates through reverse transformation. The accuracy of measurement is about a quarter millionth to one millionth based on the number of photos used. Viewing angle of the camera also influences the measurement accuracy. Several types of targets are used in photogrammetry, including an encoded one for the orientation determination and scaling one for dimension calibration. All of the targets are made from high reflectivity fine glass ball based white paint on black plastic film. The black background has very low reflectivity, while the targets have the highest reflectivity.

Photogrammetry can also be used for antenna surface panel coarse adjustment. It is used prior to the more accurate radio holographic method which is discussed in Section 8.4.1. Another type of industry surface measurement uses an optical holographic principle. Some birefringence crystals will produce different reflective indexes when different voltages are applied. If voltages of F and $F \pm \Delta F$ are applied, the crystal will produce three different indexes of n and $n \pm \Delta n$. When a collimated laser light is incident, three refracting beams come out of the crystal. With a lens, three foci are formed. By blocking the middle image on the focal plane, two coherent point sources produce an interference pattern on any curved surface. Since the separation of the two light sources can be adjusted electrically, the fringes can be tuned to have different resolutions. Therefore, the surface shape is determined. However, this method has not been used in the radio telescope field yet.

In the reflector assembly, the main tools for panel alignment are the theodolite, tape, and gauges. Laser range systems are also used in the antenna assembly process. The preliminary panel adjustment could achieve an rms surface error of about 100 μm. The photogrammetry method is usually used in between. The final panel adjustment is through the radio holographic method with an accuracy of about 10 μm.

In panel design, two formulae are important. One is the paraboloidal curve length along a radial direction and the other is the ring area of the paraboloidal surface. The paraboloidal curve length along a radial direction from the vertex is:

$$L = \frac{1}{4F}\left[x\sqrt{4F^2 + x^2} + 4F^2 \ln\left(\frac{x + \sqrt{x^2 + 4F^2}}{2F}\right)\right] \tag{8.38}$$

where F is the focal length and x is the radial coordinate. The paraboloidal ring area between two radii is:

$$S = \frac{8\pi\sqrt{F}}{3}\left[\left(\sqrt{\frac{x_2^2}{4F}+F}\right)^3 - \left(\sqrt{\frac{x_1^2}{4F}+F}\right)^3\right] \tag{8.39}$$

where x_2 and x_1 are two radial coordinates of the ring area.

8.2.2 Backup Structure Design

To maintain an accurate reflector surface shape, millimeter and submillimeter wavelength telescopes require high accuracy backup structures. Generally, the backup structures used for millimeter or submillimeter wavelength telescopes are made of, or partly made of, low CTE CFRP materials. Different from steel truss structures, CFRP material members have unique directional properties. In the fiber direction, the strength and modulus are higher and, in the directions perpendicular to the fibers, the strength and modulus are lower. The torsional stiffness of CFRP structures is also a problem. For a detailed discussion of CFRP see Section 8.3. In this section, the CFRP truss joint design is also discussed.

In a steel truss structure, beam members are welded together. However, the beam members for the CFRP truss can not be welded together and the joints are mostly made from other materials. The materials for CFRP joints include steel, stainless steel, and Invar. CFRP materials are light in weight, so that the weight of the joints becomes dominant in CFRP trusses. Another problem with CFRP truss structure is the joint size. The size of a spherical joint in a three-dimensional truss is determined by the beam diameter and the angles between jointed beams. When the angle between beams is small, the size of the joint will be large to ensure that all the beams are aligned to the center of the joint [Figure 8.8(a)]. This increases the joint weight. To overcome this problem, a complex sub-joint design is used [Figure 8.8(b)]. The sub-joint connects several nearby beams within a small angular range. This increases the complexity and the cost. Another problem with steel joints is the effective CTE of the connected beams may increase as the length ratio between the CFRP and steel parts reduces. This is especially true for relatively short beams. The CTE difference produces truss deformation.

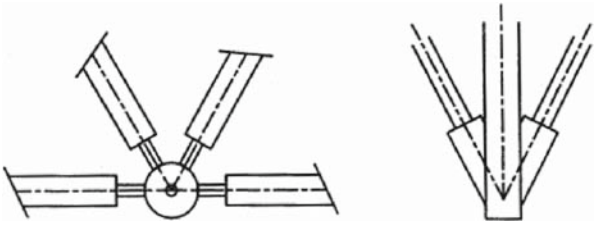


Fig. 8.8. Joint and angle size between truss members ((a), *left*) and complex joint for small truss angle between truss members ((b), *right*).

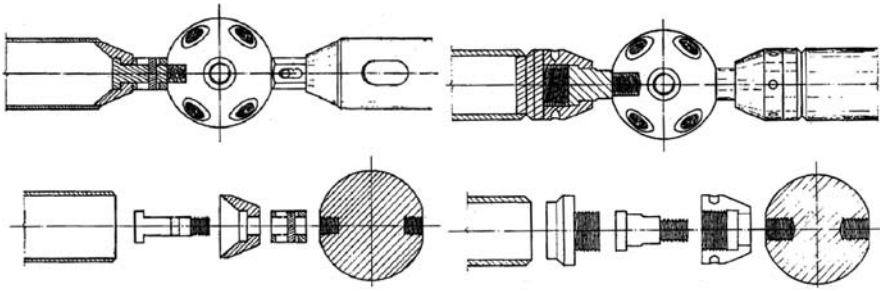


Fig. 8.9. Two types of truss connections (Chiew, 1993).

For precision submillimeter wavelength antennas, even if the joint is made of steel or other metals, joint welding is usually avoided to prevent truss deformation. Therefore, specially designed joints have to be used. In the architecture field, a number of screwed joints have been developed; two designs are shown in Figure 8.9. These joints are for metal tubes. For CFRP tubes, a larger connecting area between the inside surface of the CFRP tube and the outer surface of the joint is necessary.

If the CFRP tube is a unidirectional protruded one, a special joint design has to be used as shown in Figure 8.10. This design produces a staged connection between the metal part and CFRP part for avoiding the end stress concentration in a simple adhesive-bonded joint. The end stress level in a simple joint can be many times higher than the average stress value of the joint. A simply increase of the bonded length may not solve the stress concentration problem as the center part of the bonded line may have no stress at all. Unidirectional CFRP tubes have serious disadvantages as the radial stiffness is very low and the radial CTE is very large. The metal cover in the design restrains the expansion of the tube. In this design a screw is used for securing the connection between the cover and the header parts. If the CFRP tube is simply glued to the outer diameter of a metal header, differential thermal expansion may separate the CFRP tube from the connected header part.

In some antenna backup structures, CFRP beams are only used in the axial direction as this direction has more influence on surface error than other directions. Using only CFRP plates, the backup structure can be a shell

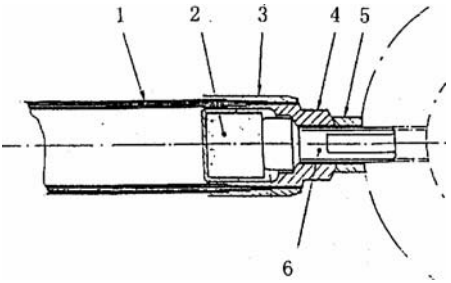


Fig. 8.10. Axial aligned CFRP tube head design: (1) CFRP tube; (2) foam; (3) metal cover; (4) metal head; (5) nut and (6) screw (IRAM).

structure. The backup structure can also be built by using a CFRP aluminum honeycomb sandwiched structure. In this design, the backup structure is made of many fan-shaped sections. All sections are joined side-by-side and the top surfaces are used for panel support. The structure is radially weak in torsion. The thermal problem of a channel sectional sandwich is discussed in Section 8.3.2. Another type of backup structure is a combination of CFRP turss and CFRP aluminum honeycomb sandwich sub-panels. The sub-panels are on top of a sparse CFRP truss and the surface panels are supported on top of the sub-panels.

When low CTE material is used in the backup structure, the connection design between the backup structure and its support is important. Any thermally induced stresses should be avoided. A ring of separated thin plates for the connection can be used. This thin ring plate support allows dimensional difference in the radial direction. The structural axial symmetry is also maintained and the thermal stresses are low. Other stress-free connections include usage of flexible hinges.

In some millimeter wavelength array telescopes, accurate phase or path-length calibration is important. Astronomers have concluded that the best approach in phase calibration is to switch the whole telescope back and forth from a target to a calibrator. In general, the distance between the target and a nearby calibrator is about 3°. The switch period is about 1.5 s. The telescope, therefore, will move at a high acceleration of about $24^\circ/s^2$. Special on-the-flight observation also requires high acceleration movement of the telescope. All these demand the telescope structure have a higher natural resonant frequency.

Millimeter and submillimeter wavelength telescopes also require few mirrors of room temperature in its optical system for reducing the thermal noise level. Therefore, a large receiver cabin in the Cassegrain focus is desirable. The cabin acts as a balance weight of the dish structure, but it increases the moment of inertia on the elevation axis. For avoiding serious thermal effects, the outer surfaces of the backup structure, yoke, and pedestal of a millimeter wavelength telescope should be well insulated from direct solar heating.

8.2.3 Design of Chopping Secondary Mirror

Strong atmospheric emission in the millimeter or submillimeter wavelength region produces high background level in the observation. To eliminate this background noise, a chopping subreflector, or nutator, is often used. The principle of this method in observation is the same as in infrared observation which will be discussed in Section 9.1.2.

For a chopping secondary, the relationship between the tilting angle of the secondary (subreflector) $\Delta\theta$ and the tilting angle of the beam $\Delta\phi$ is (Radford, 1990):

$$\Delta\phi = \left[R\left(\frac{BDf}{f} - \frac{BDF}{F}\right) - \frac{f}{l}(BDf + BDF)\right]\Delta\theta \tag{8.40}$$

where f and F are the primary and Cassegrain focal lengths, BDf and BDF the corresponding beam deviation factors, and l and R the distances from the subreflector's vertex to the primary focus and to the rotation axis of the subreflector. When the subreflector has a tilt angle, pointing changes and coma are produced. Optimization of the pivot axis can eliminate the effect of coma aberration (Lamb, 1999a). However, for reducing the torque required in the mirror chopping process, the pivot axis may be better not at the optimal position. The pivot axis is usually near the vertex of the subreflector, so that the moment of inertia of the mirror is reduced. However, the system retains its coma effect.

There are several designs for the chopping subreflector assembly. One design places a counterweight, which has the same moment of inertia, at the opposite side of the secondary so that the reaction force caused by the motion of the secondary is compensated by the motion of the counterweight. The net force applied on the antenna structure is zero. Another clever chopping mirror design is called the reactionless subreflector as shown in Figure 8.11.

In this reactionless chopping mirror design, there is no fixed frame to push the subreflector and the forces applied to the subreflector are from voice coil motors. The two voice coil motors are located on a frame, named the motor frame. The motor frame and the subreflector are supported on sets of torque bearings (Figure 8.12). These bearing axes of the motor frame and the subreflector are parallel to each other. The motors generate forces through voice coils in the magnetic field. When a current runs inside the coils, forces are produced to push or pull the subreflector. The coil's position is controlled by a plate with spring and damping properties. The reactions of these forces are absorbed or

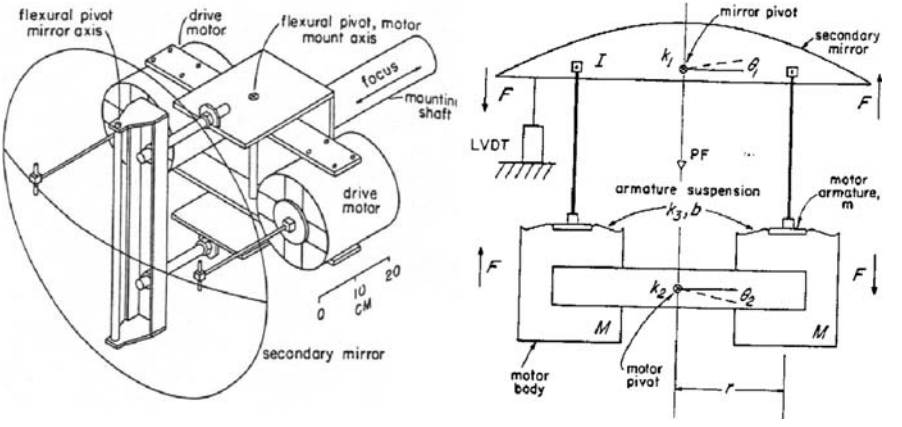


Fig. 8.11. A schematic design of a chopping subreflector (Radford, 1990).

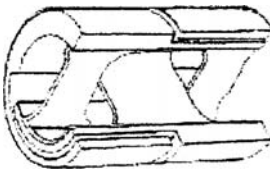


Fig. 8.12. Structure of a torque bearing.

compensated within the system so that there is no reaction in the antenna structure.

If I is the moment of inertia of the subreflector, M the mass of the motor, m the mass of the coil, and r the radial distance of the motor, then the moment of inertia of the subreflector plus the coil is $I + 2\ mr^2$ and the moment of inertia of the motor is $2\ Mr^2$. By considering the spring constant k_3 and the damping ratio b between the motor and coil, the whole system is a coupled harmonic oscillator with two subsystems. Under the thrust forces from the motors, the main mode of motion is an anti-symmetric one. However, the frequencies of two subsystems are not the same, that is $2k_1/(I + 2mr^2) \neq 2k_2/2Mr^2$, where k_1 and k_2 are spring constants of two torque bearings. This produces small coupling to the undesirable symmetric mode. The moment of inertia of the whole system is:

$$I' = \left[(I + 2mr^2)2Mr^2\right] / \left[(I + 2mr^2) + 2Mr^2)\right] \tag{8.41}$$

Assuming the thrust force of each motor is F, and a constant acceleration and deceleration is assumed for the system, the 10 to 90% amplitude response time will be $t \approx 1.1(\Delta\theta I'/2rF)^{1/2}$. This time period is about 10 ms. The chopping angle of the subreflector is about 3 to 4° with a chopping frequency of about 10 Hz. The anti-symmetric mode resonant frequency is $\omega^2 = 2k_3\ r^2\ /\ I'$ and the damping ratio is $\xi^2 = b^2r^2\ /2k_3\ I'$. In Figure 8.11, the LVDT is a displacement sensor which provides displacement data for the feedback control during the chopping process. The block diagram of the chopping mirror control system is in Figure 8.13. The purpose of the control is to suppress the symmetric mode between the subreflector and the motor support. To suppress the wind effect on the system, an increase of the system damping is necessary.

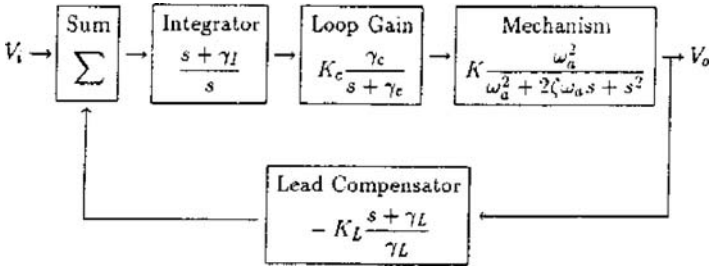


Fig. 8.13. Control system of the chopping subreflector (Radford, 1990).

8.2.4 Sensors, Metrology, and Optical Pointing Telescopes

The environment conditions of an open-air millimeter and submillimeter wavelength telescope are much more severe than those of an optical telescope inside an astro-dome. The turbulence loadings, including wind and thermal ones, produce deformations of the antenna structures. For measuring these tiny changes of the antenna structure, thermal sensors, tilt meters, and accelerometers are widely used. These and related reference structures form a local metrology system of the antenna. In some antennas, optical telescopes are also used for the pointing calibration.

Thermal sensors are temperature measuring devices. Three types of temperature sensors exist. These are thermal couple, resistance temperature detector (RTD), and thermistor. Among these sensors, RTD and thermistor are used in antennas. The principle of temperature measurement is simple. Most thermal sensors are based on the resistance change when temperature changes. The RTDs are made of metal materials, such as platinum, nickel, copper, or iridium. The variation of resistance with temperature for the RTDs can be represented as $R = R_0(1 + a_1T + a_2T^2 + ...)$. The number of terms used depends on the material and the required accuracy. All the coefficients in the formula are usually constants. For platinum, $a_1 = 0.00385\Omega/\Omega/^\circ\text{C}$ and for nickel, $a_1 = 0.00617\Omega/\Omega/^\circ\text{C}$.

Thermistors are made of ceramic materials by combining two or more metal oxides. Most of them are made of oxides of cobalt, copper, magnesium, zinc, or nickel. There are two groups of thermistors: negative temperature coefficient (NTC) and positive temperature coefficient (PTC) ones. Most sensors used in the telescope structure are the NTC thermistors. They are small in size, high in accuracy, and low in cost. The relationship between the temperature and the resistance of this type of sensors is given by:

$$1/T = A + B(ln\ R) + C(ln\ R)^3 \tag{8.42}$$

where T is temperature in K. By using this formula to fit a curve, the accuracy can reach $\pm 0.005^\circ$C within the range of 50°C. These sensors have a rate of resistance change with temperature in the range between $-3\%\ /\ ^\circ$C and $-6\%\ /\ ^\circ$C. Figure 8.14 shows the curves of resistance change for platinum RTD and some NTC thermistors. Before using thermal sensors, hot bath resistance calibration is necessary.

If thermal deformations, such as surface error, pointing error, or focal length change, are calculated using the FEA for some typical temperature distributions, then the real time temperature effects can be estimated by fitting the measured temperature data with the typical temperature distributions. After the fitting, the surface deformations or pointing errors are derived through a linear combination of the stored deformation data sets. Therefore, some of these errors, especially the pointing and focusing errors, can be compensated in real time. However, the real time error compensation is never perfect since the time

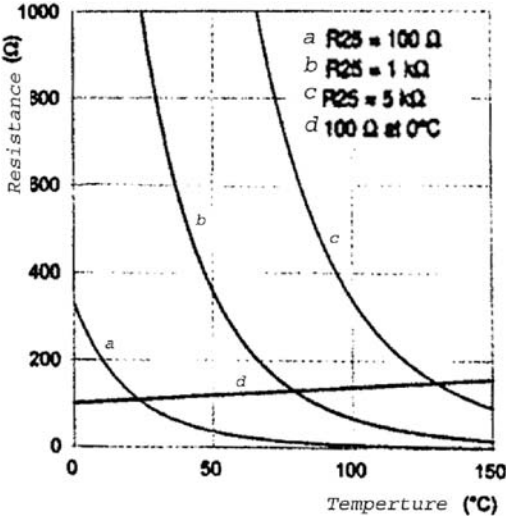


Fig. 8.14. Resistance curves of Pt RTD (d) and NTC thermistors (Lavenuta, 1997).

constant of the structure is usually different and the measured temperature data are not dense enough.

Tilt meters measure tiny angle changes relative to a vertical direction. Tilt meters are also used in millimeter or submillimeter wavelength antennas. Tilt meters include both capacitance and inductance types. Figure 8.15 shows a diagram of an inductance-type tilt meter. In this figure, A is a pendulous mass, B a position sensor, and C a torque motor. The pendulous mass attaches to the torsional suspended armature of a torque motor. When the pendulous

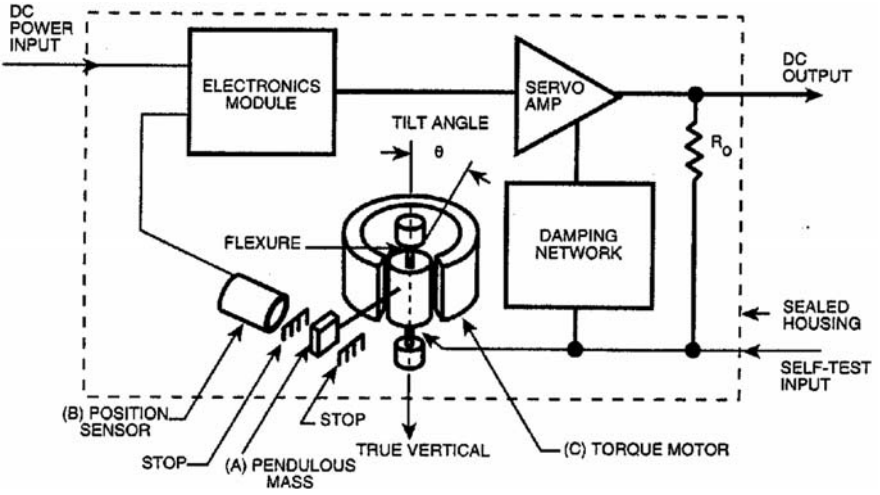


Fig. 8.15. Structure of a tilt meter.

mass is away from the neutral position, the sensor produces a small signal to feed a direct voltage to the torque motor. The pendulous mass will swing back to its neutral position. The sine of the tilt angle can be detected from the voltage applied on its motor.

Tilt meters can be installed directly on the foundation or on top of the azimuth axis. When it is on the foundation, the tilt of the foundation can be monitored. When it is on top of the azimuth bearing, the tilt of the bearing can be monitored. If a tilt meter is placed away from the azimuth axis, the angular velocity due to the antenna movement will produce an eccentric tilt. The reading may not be the same as the structure is not in motion. The tilt meters only measure the tilt angles of the positions they are attached. Therefore, unwanted local deformation and prestress of the mounting should be avoided.

Accelerometers are also used in telescope structures, especially in structures which require fast motion. Using accelerometers, the telescope's dynamic parameters can be accurately determined. They can be used to check the vibration of a structure when high accurate accelerometers are used. Accelerometers also include both capacitance and inductance ones. Figure 8.16 is a diagram of a capacitance-type accelerometer. The main mass of the system is supported by a number of cantilever beams. A group of capacitors are formed between a main mass plate and many fixed plates. Since a high frequency square wave voltage with 180° phase difference is applied on the fixed plates of these capacitors, there is no voltage output from the moving part (main mass) of the capacitors when there is no acceleration. If the meter is under acceleration, there will be voltage output from the moving part of the capacitors. Acceleration applied is proportional to the output voltage of the moving part. The resistance of position change will move the main mass back to its neutral position.

Independent precision reference structures were first used in optical telescopes when a heavy counterweight in traditional equatorial optical telescopes was used. These structures usually are inside and separated from the structures which carry the counterweight. By using the same principle, a separated reference structure could be used as an encoder support for precision millimeter and submillimeter wavelength telescopes when the yoke is under thermal and wind loadings. However, these structures have to be thermally stable and free from resonant vibration. In

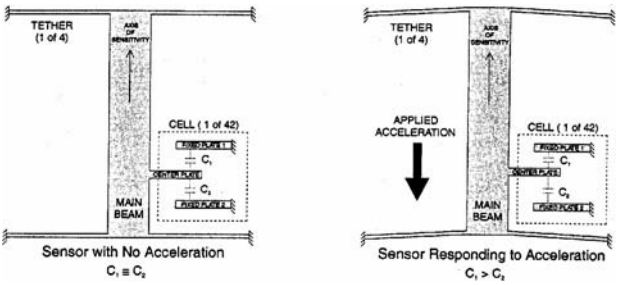


Fig. 8.16. Structure of a capacitance accelerometer.

some cases, optical trusses made of laser range systems can be used (Section 5.3.3). Sensors together with these reference structures are called the metrology system. For large precision millimeter and submillimeter wavelength telescopes, the metrology system may be very important to assure the telescope pointing and stability.

Adding small pointing optical telescopes to millimeter wavelength telescopes is always helpful. However, the optical axis of the pointing telescope may not be the same as the radio axis of the antenna. Therefore, it only provides a reference in the pointing calibration. Most optical pointing telescopes used are located near the center of the antenna. Using a small polished surface on both the main dish and subreflector may provide a more accurate optical pointing reference. It may lead to optical star guiding during millimeter or submillimeter wavelength observations.

8.2.5 Active Optics Used in Millimeter Antennas

Homologous antenna design insures that the reflector surface remains a paraboloid shape when the elevation changes. This practice greatly improves large antenna performance in a passive way. Further performance improvements usually require complex active surface control which is up to today not widely used. Active surface control was tried with the GBT. The LMT and CCAT, under development, involve some type of active surface control. All of these are in the testing stage.

Besides active surface control on the primary and secondary mirrors, Greve et al. (1996a,b, 1994) proposed another wavefront compensation method on a small reflector near the focus, where "small" and "near" are relative concepts compared with focal length. The mirror used should be located in a far field from the focus. This distance is about $z = 2\ d^2\ /\ \lambda \approx 500\ \lambda$ for a horn aperture size of $d \approx 10 - 15\lambda$. At this distance, the field distribution is nearly the same as that in the aperture plane. Therefore, compensation of surface errors is easier. Differing from optical telescopes, real time measurement of wavefront shape in radio antennas is difficult. So the compensation is still limited to some constant terms of coma and/or astigmatism.

Astigmatism is an aberration easily produced if the expansions of a dish along two perpendicular directions are different. The measurement of the astigmatism can be done by measuring the radiation pattern of a point source before and after the focus. Generally, a small defocusing distance of less than two wavelengths is desirable. When the defocus distance is large, the radiation pattern will involve more defocusing effect instead of astigmatism. The defocusing can be obtained by axial adjustment of the subreflector. When the antenna is in focus, astigmatism is difficult to detect. However, if it is out of focus, the pattern will no longer be symmetric about the azimuth or altitude axes. Astigmatism is the ratio of HPBWs in both azimuth and altitude directions when a defocusing exists. It is:

$$A(z/\lambda) = \theta_a(z)/\theta_e(z) \tag{8.43}$$

where $\theta_a(z)$ and $\theta_e(z)$ are HPBWs at both azimuth and elevation directions. From the formula of astigmatism, the corresponding phase error is:

$$\Phi_\alpha(\rho, \varphi) = \alpha'\rho^2 \cos[2(\varphi - \varphi_0)] = 2\alpha'\rho^2 \cos^2(\varphi - \varphi_0) - \alpha'\rho^2 \tag{8.44}$$

where $\alpha' = 2\pi\alpha/\lambda$ is a dimensionless astigmatism coefficient, α the amplitude of the wavefront deformation, ρ a normalized radius, and φ a polar angle in the aperture plane. The wavefront error caused by defocusing can be expressed by using the square term of the radius. If the wavefront error is known, the radiation pattern can be derived by Fourier transform.

In the case of an extended radio source, the radiation pattern produced is a convolution of the radiation pattern and the source distribution. If the measured and an ideal pattern are known, the compensation can be done by adding small pieces of polymer materials on the metal surface of the small mirror near the focus. The polymer material should have a different refractive index from that of the air. The small mirror inserted can be a Nasmyth one.

Another active compensation applied in millimeter wavelength telescopes is for coma induced by a chopping secondary. During the chopping process, the subreflector is away from its neutral position and coma is produced. Using a similar small mirror with polymer pads, the coma can be compensated. However, during coma correction, the mirror has to switch in and out of the optical path as the secondary is in and out of its neutral position. The compensation is basically the same as the astigmatism correction.

8.2.6 Antenna Lightning Protection

Lightning is caused by build-up of electrostatic charge in clouds. One region of a cloud builds up a positive charge and the other a negative one. The build-up process is not completely understood yet, but the bottom of the cloud usually ends up being negatively charged and the top positively charged. If the build-up (separation) of charge becomes large enough, such as a few MV/m, the negative charges may leap to the positive side of another cloud or to the ground. The former is lightning between clouds and the latter is lightning between cloud and ground. The discharge between cloud and ground can be made more than once. The average lapse time of each discharge is about 0.2 s. The longest lapse time is 1∼1.5 s. Lightning can damage bearings, control circuits, and even a steel-made antenna structure. For antennas made of CFRP material, lightning will produce serious damage as the resistance of the CFRP structure is large.

Lightning protection devices for an open-air telescope include a lightning rod on top of the structure, the conducting lines in the middle part, a grounding system beneath the structure, and an automatic high voltage switching system. For protecting the telescope, it is necessary to keep the same electrical potential

over the whole structure body. Electromagnetic shielding is also needed for its electrical circuit equipment.

In the lightning process, there is no conductor line between cloud and ground. The charge is through a plasma channel. The completion of the channel takes a relatively long time, about 0.01 s. The air temperature inside this channel may reach about 5,000–6,000 K. After the channel is completed, there is a first discharge. The average current of the first discharge is about 30 KA, and may reach a maximum of 200 KA. The time for the first discharge is about 10–100 μs. After the first discharge, there is a second discharge. The time of the second discharge is very short, only about 0.25 μs. The current in the second discharge is also small; it is about one quarter of the first discharge. Since the time of the second discharge is very short, the change of current is very large, and can reach 10^{11} A/s. The magnetic field during the lightning is between 25 kHz and 1 MHz.

There are two aspects to lightning protection. One is to lead the lightning charge safely to the ground. If the structure itself is used as conductor lines, it is necessary to reduce the voltage difference between the structure and the ground. If lightning rods and conductor lines are used outside the structure, it is necessary to insure the structure is inside the protection radius. The protection radius for a lightning rod with a height of h is (Bazelyan & Raizer, 2000):

$$\begin{aligned} R &= (r_s h - h^2)^{1/2} && r_s \geq h \\ R &= r_s && r_s < h \end{aligned} \tag{8.45}$$

where r_s is the lightning striking radius which is related to the lightning current (Table 8.4). One new concept is called the rolling sphere principle of protection. The method states that when a sphere of diameter 20–60 m rolls over the structure, the untouched parts of the structure are all within the protected area. A smaller diameter of the sphere means higher protection level.

The copper or steel grounding should connect reliably to the structure parts or the steel rebars of the foundation. The steel grounding plate should have an anti-corrosion treatment. To evaluate the resistance of the grounding, soil resistivity is needed. The resistivity of soil can be derived from the following formula (Vijayaraghavan et al., 2004):

Table 8.4. Relationship between lightning current and striking radius (Bazelyan and Raizer, 2000)

Current	25 KA	50 KA	75 KA	100 KA	125 KA	150 KA	175 KA	200 KA
Mini radius	50 m	80	100	110	120	130	140	150
Maxi radius	75 m	150	230	300	370	450	520	600

$$\rho = 2\pi SR \tag{8.46}$$

where ρ is the resistivity in ohm-m, S the distance between electrodes in m, and R the resistance measured between two electrodes on earth in ohm. The represented value is derived when the electrode depth is 0.85 S. The resistance of a single vertically grounded rod electrode is:

$$R_e = \rho/2\pi L[\log(8L/d) - 1] \tag{8.47}$$

where R_e is the resistance of the rod electrode, L the length buried in the soil, and d the diameter of the rod in m. If the electrode is 16 mm in diameter and buried 3 m into the ground, the resistance is about 1/3 of the soil resistivity in a unit of ohm. The soil resistivity varies from place to place. The value ranges from 10–1,000 ohm-m. For protecting the signal lines buried underground, a grounding line should be on top of these lines. A copper grounding line can protect the lines with a 90° angle below the grounding line.

The second aspect in lightning protection is the induction effect caused by the lightning current. If the area of a coil is A and its distance to a shielding surface is d, the inducting voltage when a current is going through the shielding surface is:

$$u = C_S A\mu_0 \frac{dH}{dt} = C_S A\mu_0 \frac{i}{2\pi d}\frac{1}{T} \tag{8.48}$$

where $\mu_0 = 1.2566 \cdot 10^{-6}$ Vs/Am is permissivity of the vacuum, H the magnetic intensity, i the electric current, T the time interval of the current, and C_S the shielding efficiency. In general, a sealed metal box has a very high shielding efficiency; it is between 300 and 1,000 dB . However, the shielding efficiency SE reduces if there are n openings in the shielding box:

$$SE = 20log\left(\frac{\lambda}{2L}\right) - 20\,log\ n \tag{8.49}$$

where L is the largest dimension of the opening, λ the wavelength, and n the number of openings which have their dimensions smaller than half of the wavelength. The shielding efficiency is independent of the shape of the opening. For protecting some important circuit devices, double layer shielding may be necessary. The shielding layer should be thicker than 8 mm for important parts.

8.3 Carbon Fiber Composite Materials

8.3.1 Properties of Carbon Fiber Composites

Carbon as a light element with density of 2,268 kg/m^3 can exist in various forms. Graphitic structure inside carbon fibers is carbon atoms in a form of hexagonal

layers with very strong bonds between atoms. Anisotropic graphitic carbon has a maximum Young modulus in the hexagonal layer plane of 1,000 GPa, five times the modulus of steel. The modulus perpendicular to the layer is only about 35 GPa. The degree of order of the graphitic form and the porosity in carbon fiber determines the fiber axial modulus (Fourdeux et al., 1971).

Carbon fibers have superior structure properties with tensile modulus and strength many times greater than other materials, such as steel, titanium, or aluminum (Figure 8.17). The density of carbon fibers is between $1{,}760 \sim 2{,}170\ \text{kg/m}^3$, four-times smaller than that for steel. Thermal conductivity of carbon fibers varies from 8.5 to 640 W/mK. The pitch-based fibers have larger conductivities, while pan-based fibers have very small conductivity. The conductivity of fiber is related to the modulus; the higher the modulus, the higher the conductivities. All carbon fibers have negative axial CTEs. The CTEs of carbon fibers are related to the modulus as in Figure 8.18. The CTEs in transverse direction of carbon fibers are larger, about $5.5 \sim 8.4 \times 10^{-6}/^{\circ}\text{C}$.

Carbon fibers have very good structural and thermal properties. However, they require materials to hold and to transfer loads to fibers. These resin materials are called matrix. Polymers, as matrix materials, are cheap, but they have inferior strengths, modulus, and CTEs (Table 8.5).

The properties of CFRP composites depend on the fiber arrangement and matrix to bond them together. Unidirectional lamina (UDL) in a resin matrix has a Young modulus in the fiber direction:

$$E_1 = V_f E_f + (1 - V_f) E_m \tag{8.50}$$

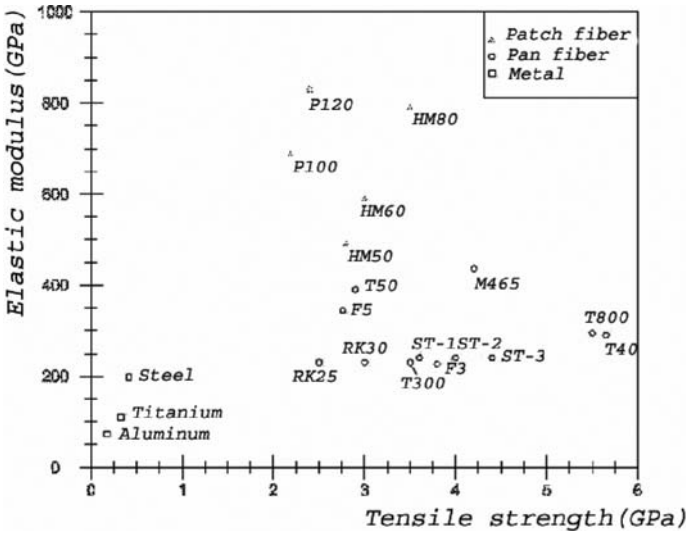


Fig. 8.17. Modulus and strength of different carbon fibers (Cheng, 2000).

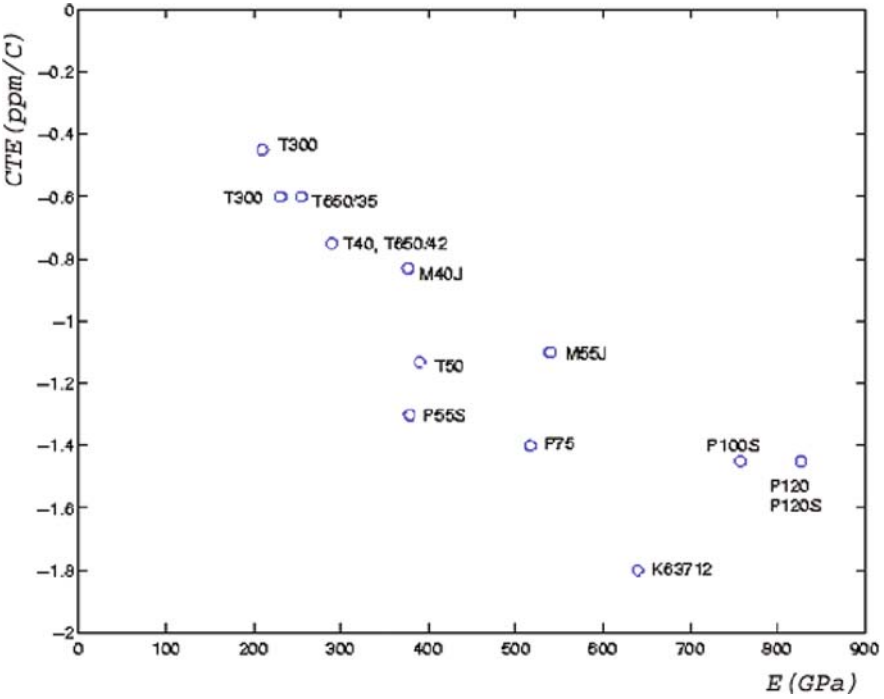


Fig. 8.18. Modulus and CTE of some carbon fibers (Cheng, 2000).

Table 8.5. Structural and thermal properties of a typical polymer

	Modulus	Strength	Density	CTE
Epoxy	15∼35 GPa	35∼85 MPa	1,380 kg/m^3	70×10^{-6}/°C

where V_f is the volume fraction of fibers, E_f the modulus of fibers in the axial direction, and E_m the modulus of the matrix. The strength of the composite has exactly the same formula as the modulus. In practice, the fiber volume is limited to about 70%. The modulus in the transverse direction is:

$$E_2 = \frac{1}{V_f/E_f + (1 - V_f)/E_m} \tag{8.51}$$

The formula for CTEs of a unidirectional lamina or tube is complex. The CTE calculation has to consider the fact that carbon fiber is only a transverse isotropic material. Rosen and Hashin gave formulas for CTE of a unidirectional CFRP composite which involve bulk modulus of fiber and matrix, and effective properties of composite assemblage. Approximations of these formulas are:

$$\alpha_1 = \frac{\alpha_f E_f V_f + \alpha_m E_m (1 - V_f)}{V_f E_f + (1 - V_f) E_m}$$
$$\alpha_2 = \left[v_f V_f + v_m (1 - V_f) \right] \frac{\alpha_f E_f V_f + \alpha_m E_m (1 - V_f)}{V_f E_f + (1 - V_f) E_m} \tag{8.52}$$

where v_f and v_m are the Poisson ratio of fiber and matrix and α_f and α_m the CTEs of fiber and matrix. Compared with complex formulas, the above formulas give a slightly smaller CTE value along the fiber direction.

Off-axis modulus of a CFRP plate decrease as a 4th power of the cosine angle. Figure 8.19 shows off-axis modulus for CFRP plates. The transverse CTE has the maximum at 0° angle.

The CFRP laminate involving two layers of UDLs has one very important property in their CTEs. If the fibers are arranged at a shallow and symmetrical angle, the longitudinal CTE can be negative even for T300 carbon fibers. As shown in Figure 8.20, the laminate CTEs (T300 fibers) can be as low as −2 ppm/°C.

Figure 8.21 shows the CTEs and the modulus of a number of laminas and quasi-isotropic plates. High-modulus unidirectional laminates (UDLs) have negative CTEs but spread in a wider range. For composite laminas, the transverse CTEs range from 15 to 30 ppm/°C. Detailed composite CTE data varies according to manufacturer. Figure 8.22 shows modulus and yield strength of different metals and CFRP plates.

For stability reasons, carbon fibers are usually symmetrically arranged around the middle layer surface so that there is no truly bending-isotropic CFRP plate unless the fiber arrangement is asymmetric about its central plane.

8.3.2 Thermal Deformation of Shaped Sandwiched Structures

A sandwich structure, with top and bottom surfaces made of high modulus materials and middle cores made of light weight materials such as aluminum

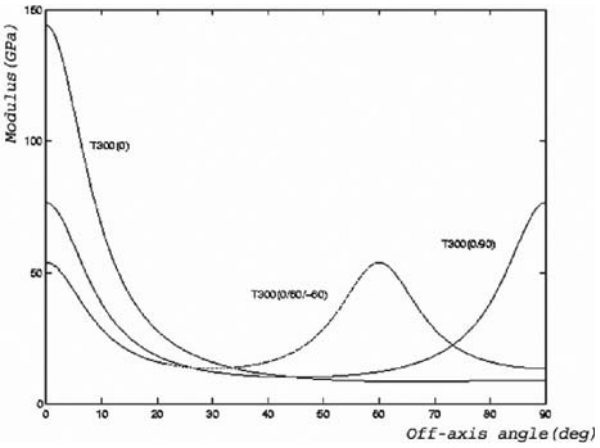


Fig. 8.19. Off axial modulus of different types of CFRP (Cheng, 2000).

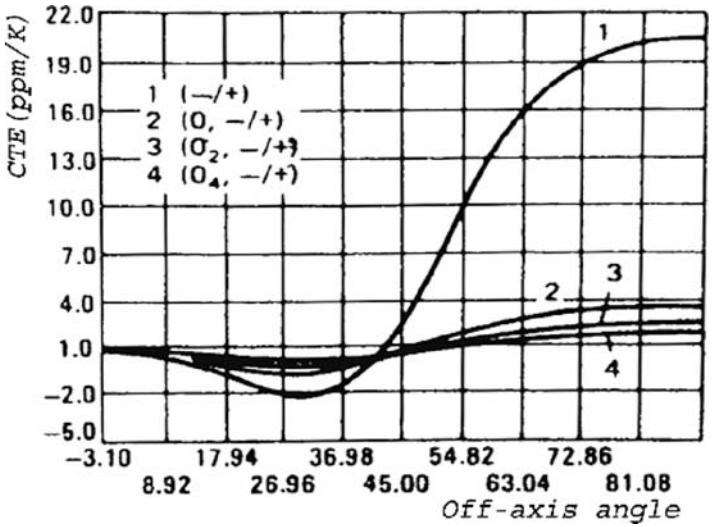


Fig. 8.20. The CTEs of CFRP with small half angle between fibers.

honeycomb or plastic foam, is highly efficient. Under outside load, the top and bottom surfaces undergo tension and compression, while the middle core only produces shear. The CFRP-aluminum honeycomb sandwiches combine high stiffness of CFRP plates and light weight of aluminum honeycomb. The bending stiffness of such a sandwiched structure is:

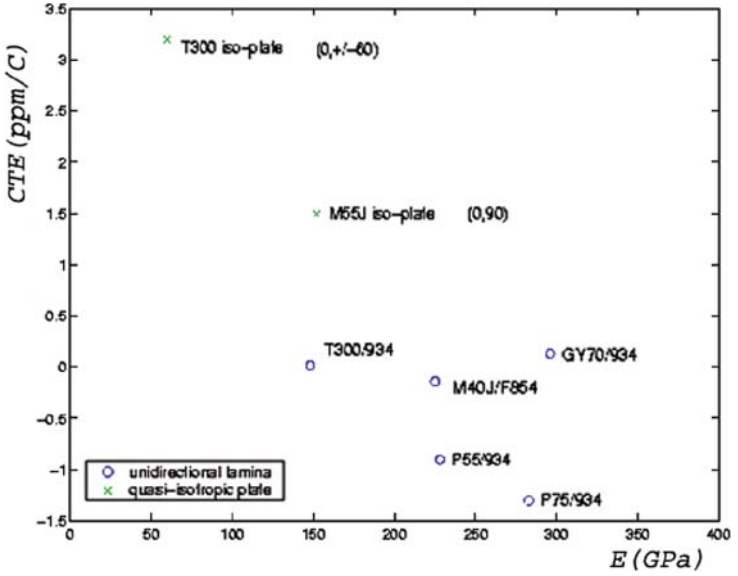


Fig. 8.21. Modulus and CTEs of different CFRPs (Cheng, 2000).

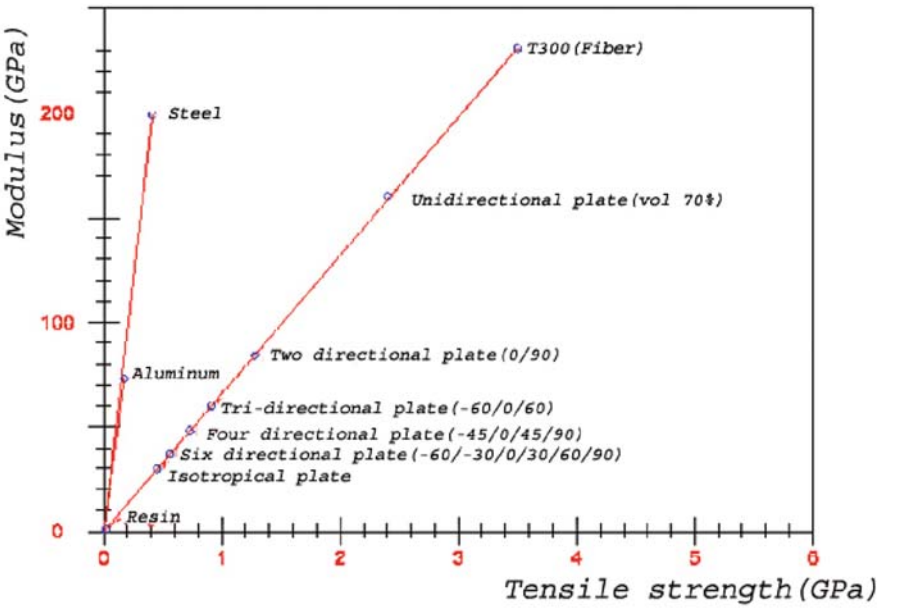


Fig. 8.22. Modulus and strength of different CFRP plates (Cheng, 2000).

$$D = \frac{1}{3}\left[\frac{2E_f}{1-v_f^2}\left(\frac{3}{4}h_c^2 t_f + \frac{3}{2}h_c t_f^2 + t_f^3\right) + \frac{E_c}{1-v_c^2}\frac{h_c^3}{4}\right] \tag{8.53}$$

where E_f and E_c are Young modulus of the surface plate and core, v_f and v_c their Poisson ratios, and t_f and h_c the thickness of the surface layer and the core in between. The last term is the bending stiffness of the core plate. Since $E_c << E_f$ and $t_f << h_c$, therefore Equation (8.53) can be reduced to:

$$D = \frac{1}{2}\frac{E_f {h_c}^2 t_f}{1-{v_f}^2} \tag{8.54}$$

The bending stiffness of a sandwiched plate is much higher than that of a solid plate made of a same weight CFRP. For a thickness ratio of 20 between core and face plates, the bending stiffness is about 300 times that of a monocoque construction.

A CFRP-aluminum honeycomb sandwiched structure is thermally anisotropic. Thermal deformation of shaped sandwiched structures was studied by Cheng (2003). Figure 8.23(a) shows a T-shaped sandwiched structure with outer surfaces of CFRP and the core of aluminum honeycomb. When temperature changes, a shape change occurs. Taking two parallel top and bottom lines of the horizontal sandwiched plate, the top one is made of CFRP and the bottom one includes

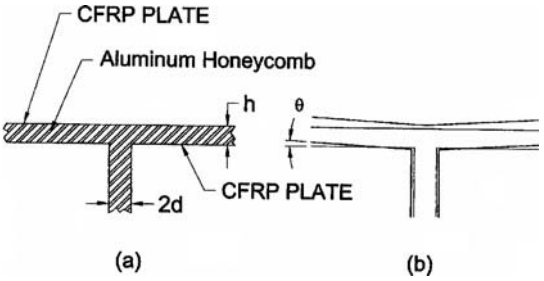


Fig. 8.23. Thermal deformation of T-shaped sandwich (Cheng, 2003).

a small portion of honeycomb. CFRP has a lower CTE than that of the honeycomb. If the lengths of these two lines are the same at one temperature, the lengths of two lines will be different when temperature changes. To accommodate this length change, the horizontal surface will bend. When temperature increases, the surface bends upwards [Figure 8.23(b)]. By assuming the CTE of CFRP is very small, the angle of the bending is approximately:

$$\theta \approx d\alpha_{AL}\Delta T/h \tag{8.55}$$

where α_{AL} is the CTE of aluminum, h the height of the top sandwich, and ΔT the temperature change. For an L-shaped sandwich structure, the shape change occurs again as temperature changes. The stress distribution and deformation around the corner are complex. There is also a small displacement between the vertical CFRP edge and the starting point of the top plate (Figure 8.24):

$$\Delta h = h\alpha_{AL}\Delta T \tag{8.56}$$

If $d<h$, the first-order approximation gives a good estimation of the angle change:

$$\theta \approx d\alpha_{AL}\Delta T/h \tag{8.57}$$

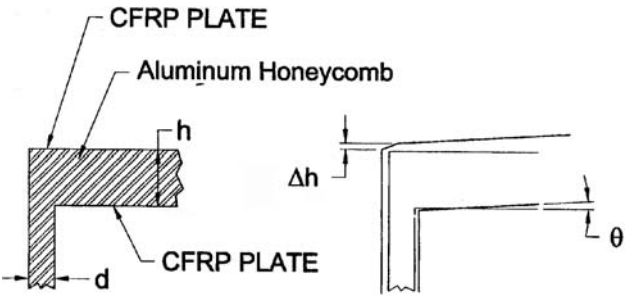


Fig. 8.24. Thermal deformation of L-shaped sandwich (Cheng, 2003).

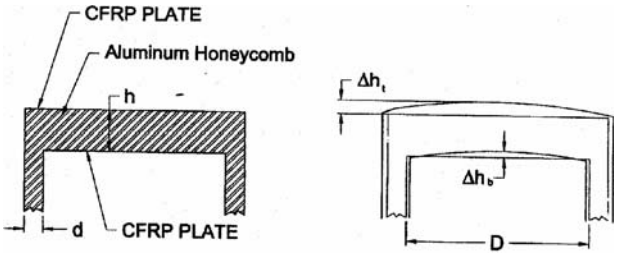


Fig. 8.25. Thermal deformation of C-shaped sandwich (Cheng, 2003).

Two T-shaped structures form a channel-shaped sandwiched structure as in Figure 8.25(a). Shape change also occurs for this structure when temperature changes. If both vertical parts are constrained, the top surface will bend as in Figure 8.25(b). If $d<h$, the radius of curvature of the upper surface is:

$$R = \sqrt{1+\theta^2} * D/(2+\theta) \tag{8.58}$$

where θ is the angle change of one corner as expressed above. The maximum height change of the lower side of the top plate is:

$$\Delta h_b = R - R * \cos\theta \tag{8.59}$$

And the maximum height change of the upper side is:

$$\Delta h_t = \Delta h_b + h\alpha_{AL}\Delta T \tag{8.60}$$

If $d = 0.04$ m and $h = 0.14$ m, we can get a linear relationship between the maximum height change and temperature as shown in Figure 8.26. Again, the stresses around the corners are complex. For avoiding such a thermal shape change, modifying the corner is necessary. The Cheng theory on thermal shape change (Cheng, 2006) plays a significant role in the backup structure design of the ALMA-US antennas, the South Pole Telescope (SPT), and the Atacama Pathfinder Experiment (APEX) projects.

CFRP honeycomb sandwiches are also widely used in space antenna structures. To build a curved sandwich, separated top and bottom CFRP sheets through laying epoxy wetted carbon fibers on top of a mold are necessary. In this case, overlap of fibers is unavoidable. Fiber overlap will produce surface print-through and residual stress. To avoid these, smaller hexagonal shaped CFRP UDL thin plates are used to match the curved shape of the mold. The fiber in one layer should remain in the same direction. On top of the layer, the hexagonal CFRP thin plates in another direction are used. After all layers are placed, vacuum curing is performed under atmospheric pressure. The curing is at a relatively lower temperature lasting 15 to 25 hours. The CFRP face sheets are placed on both sides of the aluminum honeycomb core to form a sandwiched structure.

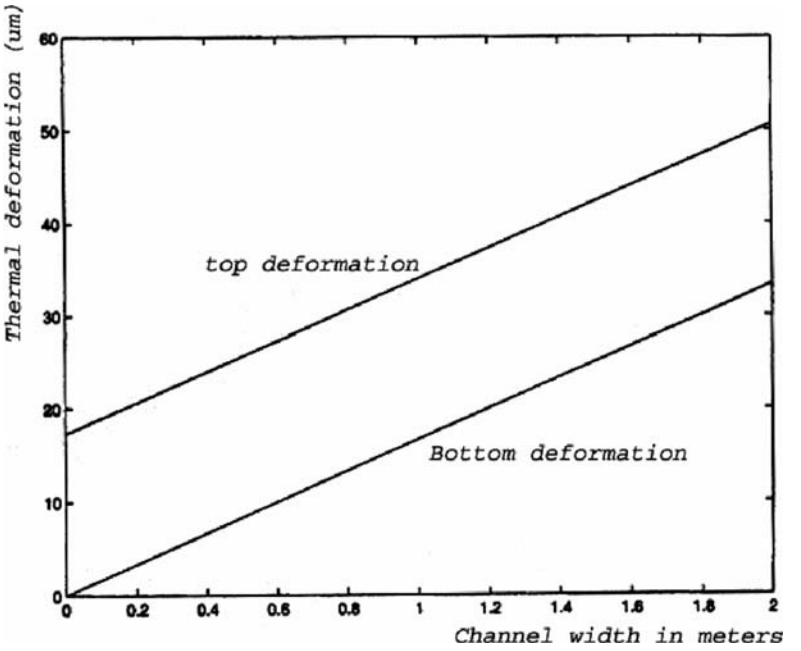


Fig. 8.26. Thermal deformation of a C-shaped sandwich with different widths (Cheng, 2003).

8.3.3 CFRP-Metal Joint Design

CFRP parts, depending on the fibers used and the fiber volume involved in one direction, are very strong in their modulus and strength. The modulus and strength of steel are 207 and 2 GPa. However, a unidirectional T300 CFRP tube can have a modulus and strength as high as 160 and 270 GPa. High-quality CFRPs have even higher modulus and strength. However, the adhesives used to glue CFRP with metal are extremely weak with typical modulus and strength of 2 GPa and 70 MPa. These strength numbers of adhesives given by manufacturers or by textbooks are mean values of the failed samples which only ensures 50% failure rate in all the test samples (Cheng, 2008). To account for the statistical uncertainty, the measured ultimate strength should not be used directly; instead, the measured value minus three times the measured standard deviation is typically used as the 100% failure criterion. Further complications of joint design come from the uneven stress distribution, the elastic-plastic law governing the adhesive materials, the fatigue under cyclic loading, mismatching of thermal expansion, chemical additives in adhesive, resin pot life, surface preparation, and curing processes.

8.3.3.1 Stress Distribution of a Simple Lap Joint

Any structural joint without real mechanical fasteners (note, some fasters may not transfer force between adherends) can be modeled as variations of a simple plate lap-joint. In a simple plate lap joint, adhesive, mainly epoxy resin, is used between CFRP and/or metal adherends.

The adhesive is poor in tension, so the load should transfer between adherends by shearing instead of tension. The shear stress estimated from dividing the longitudinal force with the bonding area does not represent the real stress inside the adhesive. This estimation also ignores the normal peer stress concentration produced by the longitudinal force.

Shear stress edge concentration has been known for years through Elastic foundation (Elsf) theory. However, Elsf theory ignores the normal peer stress edge concentration. Frostig et al. (1999) studied the adhesive-bonded joint using Closed Form High-Order (CFHO) theory and the stresses inside the bonding area are fully understood.

Generally, the shear stresses reach a maximum value at a distance of about one adhesive layer from the edge of overlap where the shear stress is zero. The shear stress goes down quickly to its minimum in the middle of the joint. The magnitude of the maximum is about 4 to 10 times the average shear stress. The transverse normal stresses at the upper and lower interfaces display different signs: one in tension and the other in compression. At a distance of one adhesive layer thickness from the edge, the upper and lower interface stresses become identical and, at about four-times the adhesive layer thickness, they vanish almost completely. Normal stress concentration exists at the edge of bonding. By using a spew-fillets edge or reversed tapered ones, the stress level at the edge could be reduced, but stress concentration still exists. Simply increasing the joint length does not help to lower the maximum stress as the middle part of the joint may have zero stress (Cheng, 2000).

8.3.3.2 Creep Model of Epoxy Material

The proper model for epoxy material is spring and viscous dashpot in series and parallel. The simplest one involves two springs with modulus of E_0 and E_1 and one dashpot with damping of η_1. One spring and one dashpot form a Voigt element. The strain response to a constant stress σ_0 is (Dean and Mera, 2004):

$$\varepsilon(t) = \frac{\sigma_0}{E_0} + \frac{\sigma_0}{E_1}[1 - \exp(-t/t_0)] \tag{8.61}$$

The loss factor is related to the relaxation time t_0 by:

$$t_0 = \eta_1 / E_1 \tag{8.62}$$

A more accurate model involves more Voigt elements. Therefore, polymer material has a very broad band distribution of relaxation times. Relaxation or creep is slow moving or deformation of a material to permanently relieve stresses. It occurs as a result of long-term exposure to stresses that are below the yield of the material. The creep strain is:

$$\varepsilon(t) = \frac{\sigma_0}{E_1}\exp(t/t_0)^m \tag{8.63}$$

The exponent m represents the broad band distribution of relaxation times and t_0 the mean or effective value of relaxation time which is dependent on temperature, humidity, stress level, and physical aging of the adhesive (Dean and Mera, 2004). The creep strain per unit stress is named the creep compliance function which is a function of time. An empirical relationship between relaxation time and some factors is as:

$$t_0 = A \cdot (RH^{-n} + B)\exp\left(-\alpha\sigma_0^2 - \beta T\right) \tag{8.64}$$

where RH is relative humidity, T absolute temperature, σ_0 stress, α, β, A, and B constants. A small relaxation time increases the rate of degradation.

8.3.3.3 Fatigue Model of the Epoxy Joint

Fatigue is a progressive and localized structural damage that occurs when material is subjected to cyclic loading. The maximum stress values are less than the ultimate tensile stress limit and below the yield stress limit of the material. Fatigue starts with dislocation movements, eventually forming persistent slip bands that nucleate short cracks. Fatigue is a stochastic process, often showing considerable scattering even in controlled environments. Damage is cumulative. Materials do not recover when it rests. Fatigue life is influenced by temperature, surface finish, presence of oxidizing or inert chemicals, residual stresses, contact (fretting), etc.

Fatigue initiates at multi-material interface corners, where singularity of stress and elevated stress is produced, under cyclic loading. The crack then propagates and produces delamination. The stress state at the interface corner drives the fatigue crack initiation. Experiments show correlation exists between cyclic stress intensity, the number of cycles, and the duration time under loading (Veer et al., 2004 and Wu, 2000).

Under small-scale yield (SSY) theory, there is a small annual elastic region near the interface corner (K-annulus) where the stress intensity characterizes the stress field at the interface corner. Near an interface corner, the asymptotical stress field can be expressed as (Wu, 2000):

$$\sigma_{ij}^m = K_1 r^{\lambda_1 - 1} f^{1m}(\theta) + K_2 r^{\lambda_2 - 1} f^{2m}(\theta) \tag{8.65}$$

where K_1 and K_2 are stress intensities, λ_1–1 the stress singularities (it is -0.032 for a 90 degree opening in the metal-epoxy interface), r the radius around the singularity point, m the material numbers, and $f(\theta)$ describes angular variations of stresses. On the basis of experiments, the following relationship between differential stress and cyclic number and duration time may exist:

$$\Delta K = C_1 - C_2 \log(N) - C_3 \log(\Delta t) \tag{8.66}$$

where C_i are constants, N the number of cycles, and Δt the loading duration time at each cycle. One experiment exhibits that $C_1 = 70.5$ ($\mathrm{MPa}^{-1}\ \mu\mathrm{m}^{-0.332}$) and $C_2 = 11.3(\mathrm{MPa}^{-1}\ \mu\mathrm{m}^{-0.332})$ (Wu, 2000). Veer et al. (2004) provided test results between cyclic number and duration time under loading of each cycle. From the test, the coefficient for the load duration time is at a similar order to that for cyclic number.

8.3.3.4 Failure Due to Differential Thermal Stresses

When materials of two adherends are different, thermal stresses of adhesive may be developed during temperature variation. Carbon fibers have a low coefficient of thermal expansion in the fiber direction which is near to zero. However, epoxy has a coefficient of thermal expansion between 30 to 70 ppm/°C. Unidirectional CFRP has very high thermal expansion in perpendicular directions. If Invar and a unidirectional CFRP tube are glued together, the stress produced is serious. The stress produced is:

$$\sigma = \Delta\alpha \cdot \Delta T \cdot r \cdot E/d \tag{8.67}$$

where $\Delta\alpha$ is the coefficient difference of thermal expansion, ΔT the temperature change, r the mean radius of the tube joint, E the modulus of the epoxy material, and d the thickness of the joint layer.

Temperature-induced failure also exists after temperature reaches a post-curing one, including relaxation of the epoxy resin, resin-filler separation, or softening of the resin due to high humidity. Low temperature or excessive desiccation may cause the adhesive to be hard or brittle.

8.3.3.5 Chemical Reasons for Failure

Adhesives may hold materials together by invading tiny pores and undercuts in the adherends thereby locking them together mechanically, or by molecular attraction of the adhesive and adherend that generates cohesion. Metal is a nonporous solid where the molecular attraction is strongest for the same material. For a CFRP-metal joint, liquid adhesive is used. The adhesive and the adherends are firmly stuck together after heat set, chemical reaction, or solvent loss. For better contact between epoxy and adherend, diluents are necessary for reducing the viscosity of

the liquid adhesive. The viscosity of old "solid" epoxy resin and polyamide curing agent is 400,000 cP. New liquid epoxy and curing agent has a viscosity of 15,000 to 50,000 cP. The diluents will evaporate during storage resulting in high viscosity and low pot life. The diluents also reduce solid contents of the epoxy.

Epoxies formed by polar members, such as epoxide, hydroxyl, amine and others, have high adhesion to metal, glass or other materials. The wetting depends on the surface energy of the material. Metals have high surface energy (500 mj/m^2). Water has a low surface energy of 72 mj/m^2. Wax has a surface energy of 22 mj/m^2. Low energy liquids will be attracted to a high energy surface. Liquid epoxy has a surface energy of 31 mj/m^2 and it will wet metal easily. Teflon has a surface energy of 18 mj/m^2 and it will be adhered by a few materials. A surface of high surface energy contaminated by low surface energy materials, such as oxide and oil, will behave as a low surface energy. The bonding will be poor.

Surface preparation is to keep the surface chemically active and free from contamination and laitance through grit blasting. Sanding does not remove contaminants resulting in low bonding strength. The CFRP surface should be sanded and cleaned with a solvent, such as methanol.

Besides diluents, fillers and modifiers are also used in epoxy. The filler slows the curing reaction and reduces the exotherm (heating) and the shrinkage during curing. Some filler will reduce the modulus and strength of the epoxy. The modifiers may improve surface properties, but impact on strength, flexibility, and thermal shock resistance. In epoxy bonding, small E-glass beads are mixed into the adhesive to insure the thickness of bonding. The volume of these beads may take little loading in the joint.

Primer or promoter is also used. The primer is for improving the wetting of the metal surface. The silane adhesion promoter increases the bonding between steel and epoxy. However, some primers with 10 to 30% talc filler cannot take much shear loading. Therefore, the thickness of primer layer is restricted to be 5–8 μm. A primer layer thicker than this will produce early separation of joint bonding.

Curing includes two reactions: conversion and crosslinking. The reaction is exceedingly complex. The conversion is formed at a slightly lower temperature. More thorough crosslinking and few unreacted groups will be produced at a postcuring temperature much above the ambient. Degrees of cure are an indication of the epoxy strength and modulus.

Galvanized corrosion is another cause of decrease of effective bonding area. An electrolyte, such as salt water, can form a circuit to pass current between two different materials. The epoxy also has free ions. Therefore, top sealing of the CFRP surface is important to avoid moisture invading into the joint area as epoxy is hydrophilic in nature.

8.3.3.6 Failure Due to Other Reasons

Joint design has a direct impact on joint reliability. Generally, a few principles should be followed: (a) the joint should avoid tension stress within adhesives;

(b) provide maximum bonding area, (c) maintain a thin and uniform adhesive layer; and (d) avoid stress concentrations.

The thickness of the glue layer affects the strength of the joint. An excessively thick layer tends to creep under prolonged loading. Thin glue lines are generally desirable but even this rule has limits as a starved joint tends to fail. One guideline lists a desirable adhesive thickness of 0.33 mm. The strength reduces to 78% at 1 mm, to 48% at 2 mm, and to 36% at 3 mm (Rodriguez, 2007).

If clamping pressure is applied to obtain a thin glue layer, high spots are stressed continuously. This strain may result in a change of shape and failure due to plastic flow or fracture. There is an ideal thickness of glue line, but it may be difficult to realize.

The adhesive should apply to both metal and CFRP surfaces. Pressure using a roller should be applied on the freshly bonded joint to ensure 100% air removal and surface contact. Care must be taken to ensure that too much adhesive is not squeezed out of the edge of the joint. It is critical that the bond line should have consistent thickness as the rest of the joint.

Surface treatment, chemical mixing and degradation, and curing temperature are other concerns. The chemicals should have proper viscosity and mix thoroughly within their pot life.

8.4 Holographic Measurements and Quasi-Optics

8.4.1 Holographic Measurements of Antenna Surfaces

Microwave holography as a method in evaluating radio antenna dish surface shape is not new (Scott and Ryle, 1977). It is based on the well-known relationship between the far-field radiation pattern and the aperture field distribution of an antenna. This Fourier pair involves both amplitude and phase. The phase on the aperture field represents reflector surface deviation.

The antenna far-field pattern is measured over a limited range of $n\theta$ around the main beam, i.e., ~1 to 2 degrees, with a sampling interval θ smaller than λ/D, which is half of the maximum frequency period in the signal as requested by Nyquist criterion, and, from the information, the aperture field details with a spatial resolution $\sim D/n$, where D is the diameter of the measured antenna, are derived.

To perform microwave holography, a stable transmitter is necessary. The transmitter can be a celestial radio point source, a beacon signal from an artificial communication satellite, or a special earth-bound transmitter. Modern earth-bound transmitters make use of two phase-locked lasers driving a photomixer, rather than a millimeter-wavelength oscillator. These transmitters allow observations at two well-separated frequencies for improving the phase accuracy. However, earth-bound transmitters may be located in near field range (Fresnel region with a distance smaller than D^2/λ and far field region with a distance larger than 2 D^2/λ). The reflections from the ground and other metal objects may also be a problem. In the near field range, field correction is necessary (Baars, 2007).

The phase in radiation field is derived through correlation between two signals: one from a reference antenna and the other from the testing antenna. The reference antenna is usually fixed. In this way, the scanning of the testing antenna will produce true phase difference compared with a fixed signal. However, the rotation of the testing dish about its axis produces a relative phase change between the two received signals. This phase change as a function of elevation angle (Figure 8.27) should be removed before the correlation. Usually, a small-size reference horn antenna is used where the beam pattern is low and flat over a wide angular range. In this case, the reference antenna can follow the motion of the testing dish. A holography instrument package with two horns is usually used at the subreflector location, where one horn is facing the reflector and another facing the transmitter.

In the holographic measurement, the dynamic range, the noise, and the spillover of the receivers play important roles. If the signals from the testing and reference antennas are $S_1(t)$ and $S_2(t)$, the correlation receiver will produce a multiplied term of two signals as shown in Figure 8.28(a). The signals received include noises as (D'Addario, 1982):

$$\begin{aligned} S_1(t) &= V_1 \cos(2\pi f_o t) + n_1(t) \\ S_2(t) &= V_2 \cos(2\pi f_0 t + \phi) + n_2(t) \end{aligned} \tag{8.68}$$

In the circuit, a low-pass filter immediately follows the correlator. Since the noise terms are Gaussian random variables, so that:

$$\begin{aligned} &< n_1 > = < n_2 > = 0 \\ &< n_1{}^2 > = kT_{S_1}B \\ &< n_2{}^2 > = kT_{S_2}B \end{aligned} \tag{8.69}$$

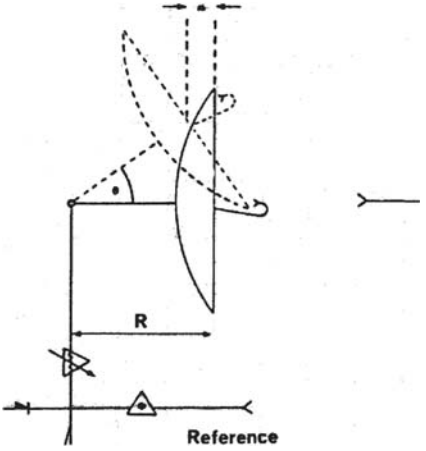


Fig. 8.27. Phase difference between a measured antenna and the reference antenna (Kitsuregawa, 1990).

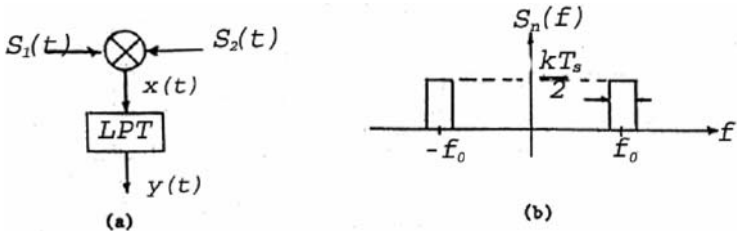


Fig. 8.28. (a) Schematic of a correlation receiver and (b) the power spectrum of receiver noise after the low-pass filter (D'Addario, 1982).

where B is the bandwidth used, normally $B < f_0$ as shown in Figure 8.28(b), k the Planck constant, and T_i the noise temperature. The output of the multiplier is:

$$\begin{aligned} x(t) &= S_1(t)S_2(t) \\ &= V_1V_2\cos(2\pi\cdot f_0t)\cos(2\pi\cdot f_0t+\phi) + n_1(t)n_2(t) \\ &\quad + V_1n_2(t)\cos(2\pi f_0t) + V_2n_1(t)\cos(2\pi\cdot f_0t+\phi) \end{aligned} \tag{8.70}$$

Let $X = <y(t)>$ be the "signal" and let $\sigma = (<y^2(t)> - X^2)^{1/2}$ be the "rms noise," where $y(t)$ is the output of the low-pass filter.

$$X = <x> = <y> = \frac{1}{2}V_1V_2\cos\phi \tag{8.71}$$

To obtain the rms noise σ, the power spectrum of $y(t)$ is needed which is derived from the power spectrum or Fourier transform of auto-correlation of $x(t)$:

$$\begin{aligned} S_x(f) &= F.T.\{<x(t)x(t+\tau)>\} \\ &= F.T.\Bigg\{<\Bigg[\frac{1}{2}V_1V_2cos\phi + \frac{1}{2}V_1V_2cos(4\pi f_0t+\phi) + n_1(t)n_2(t) \\ &\quad + V_1n_2(t)cos(2\pi f_0t) + V_2n_1(t)\cos(2\pi f_0t+\phi)\Bigg] \times \Bigg[\frac{1}{2}V_1V_2cos\phi \\ &\quad + \frac{1}{2}V_1V_2cos(4\pi f_0t + 4\pi f_0\tau + \phi) + n_1(t+\tau)n_2(t+\tau) \\ &\quad + V_1n_2(t+\tau)cos(2\pi f_0t + 2\pi f_0\tau) + V_2n_1(t+\tau)cos(2\pi f_0t+\pi)\Bigg] >\Bigg\} \\ &= F.T.\Bigg\{\frac{1}{4}V_1^2V_2^2cos^2\phi + \frac{1}{8}V_2^1V_2^2cos4\pi f_0\tau + \rho_1(\tau)\rho_2(\tau) \\ &\quad + V_1^2\rho_2(\tau)cos(2\pi f_0\tau) + V_2^2\rho_1(\tau)cos(2\pi f_0\tau)\Bigg\} \end{aligned} \tag{8.72}$$

where $F.T.$ denotes the Fourier transform, $\rho_i(t) = < n_i(t),\ n_i(t+\tau) >$, $i = 1,2$. The Fourier transform can be worked out term by term. Since the power spectrum of noises is known and a transform of the product of two terms is a convolution of two transformed terms. The resulting power spectrum as shown in Figure 8.29 is:

$$\begin{aligned} S_x(f) = {} & \frac{1}{4} V_1^2 V_2^2 \cos^2\phi \cdot \delta(f) + \frac{1}{16} V_1^2 V_2^2 [\delta(f - 2f_0) \\ & + \delta(f + 2f_0)] + KT_{S_1} KT_{S_2} B \cdot tri(f/B) [\frac{1}{2} tri(f/B) \\ & + \frac{1}{4} tri((f - 2f_0)/B)] + \frac{1}{2} V_1^2 KT_{S_2} rect(f/B) \\ & + \frac{1}{2} V_2^2 KT_{S_1} rect(f/B) \end{aligned} \tag{8.73}$$

where *tri* and *rect* are triangular and rectangular functions. If the low-pass filter has an ideal rectangular bandwidth W, then:

$$\begin{aligned} < y^2 > = {} & \int_{-\infty}^{\infty} S_y(f) df = \int_{-W}^{W} S_x(f) df = \\ = {} & \frac{1}{4} V_1^2 V_2^2 \cos^2\phi + W(KT_{S_1} KT_{S_2} B + V_1^2 KT_{S_2} + V_2^2 KT_{S_1}) \end{aligned} \tag{8.74}$$

The first term is the square of the "signal." So the "rms noise" is:

$$\sigma = \sqrt{< y^2 > - X^2} = \sqrt{W(KT_{S1} KT_{S2} B + V_1^2 KT_{S2} + V_2^2 KT_{S1})} \tag{8.75}$$

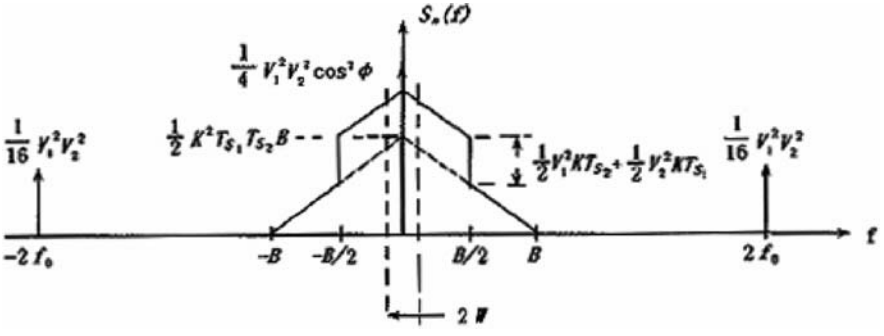


Fig. 8.29. The power spectrum of the signal after the low pass filter (D'Addario, 1982).

If both receivers are the same, so that $T_{S1} = T_{S2} = N_0 / K$, then:

$$\begin{aligned} \sigma_e &= \sqrt{W\left(N_0^2 B + 2N_0\left(\frac{1}{2}V_1^2 + \frac{1}{2}V_2^2\right)\right)} = \sqrt{\sigma_n^2 + \sigma_1^2 + \sigma_2^2} \\ SNR_e &= \frac{X}{\sigma_e} = \sqrt{\frac{(1/4)V_1^2 V_2^2}{W[N_0^2 B + 2N_0(1/2)(V_1^2 + V_2^2)]}}\cos\phi \end{aligned} \tag{8.76}$$

Because the signal received from the testing antenna V_1 is always strong, then:

$$SNR_e \approx \sqrt{\frac{V_2^2}{4WN_0}}\cos\phi \tag{8.77}$$

This shows that the signal-to-noise ratio of the reference antenna is a dominant factor in the overall signal-to-noise ratio.

The beam pattern of the reference antenna can be expressed as:

$$E_r(\theta,\phi) = \frac{j}{\lambda r}e^{-jkr}g_r(k_x,k_y) = E_0 g_r(k_x,k_y) \tag{8.78}$$

where g_r (k_x, k_y) is the Fourier transform of the aperture field. The cross-correlation of the responses from the testing and reference antennas is:

$$EE_r{}^* = |E_0|^2 g(k_x,k_y)g_r^*(k_x,k_y) \tag{8.79}$$

If M_k is a measurement of the cross-correlation and ε_k is noise with zero mean, then:

$$M_k = |E_0|^2 g(k_x,k_y)g_r^*(k_x,k_y) + \varepsilon_k \tag{8.80}$$

If there are K measurements at various directions (k_x, k_y), preferably on a rectangular grid, the best linear estimation of the testing aperture field must be:

$$\hat{F}(x,y) = \sum_{k=1}^{K} \frac{M_k}{|E_0|^2 g_r^*(k_x,k_y)} e^{j(k_x x + k_y y)} W_k \tag{8.81}$$

where W_k is a weighting which may be adjusted to control the error and the resolution. It is easily shown that:

$$\hat{F}(x,y) = F(x,y)^{**}b(x,y) + \delta(x,y)$$

$$b(x,y) = \sum_{k=1}^{K} \exp(j(k_x x + k_y y)W_k) \tag{8.82}$$

$$<\delta(x,y)> = 0$$

The last formula follows from $< \varepsilon_k > = 0$. If for all $k \neq h$, there exists $< \varepsilon_k \ \varepsilon_h > = 0$. We find:

$$<\delta^2> = \sum_{k=1}^{K} \left| \frac{W_k}{|E_0|^2 g_r^*(k_x,k_y)} \right|^2 <\varepsilon_k^2> \tag{8.83}$$

The above formula shows that the variance of the error in the estimated aperture field is independent of the aperture position (x, y). When the measurement is discrete on a rectangular grid and each increment is $\Delta l = \Delta\ k_x\ /\ k$ and $\Delta m = \Delta\ k_y / k$, using $l = k_x / k$ and $m = k_y / k$, the weighting becomes $W_k = \Delta l \Delta m / (1 - l^2 - m^2)^{1/2}$. The variance of the error in the aperture field is:

$$<\delta^2> = \left(\frac{\Delta l \Delta m}{|E_0|^2} \right)^2 \sum_{k=1}^{K} \frac{<\varepsilon_k^2>}{\left|g_r(k_x,k_y)\right|^2} \tag{8.84}$$

From the formula, if we take K measurements, the estimated surface errors are:

$$\Delta z = \frac{1}{16\sqrt{2}} \frac{\Delta l \Delta m D^2\ K^{1/2} \sigma_{AV}}{\lambda <M_0>} \tag{8.85}$$

where σ_{AV} is the rms of the measurement errors, $<\mathrm{M_0}>$ is the expected value of the on-axis measurement, D is the aperture diameter, and we have taken the reference antenna pattern as a constant. If the desirable resolution in the aperture plane is Δ, then $K \geq (D/\Delta)^2$ points must be sampled. The sampling theorem requires Δl, $\Delta m \leq \lambda / D$, then:

$$\Delta z = \frac{1}{16\sqrt{2}} \frac{\lambda D \sigma_{AV}}{\Delta <M_0>} \tag{8.86}$$

where $<\mathrm{M_0}> / \sigma_{\mathrm{AV}}$ is the signal-to-noise ratio.

The other holographic technique is based on phase retrieval from an out-off-focus (OOF) power pattern (Nikolic et al., 2007). This is a method similar to the single image curvature sensor discussed in Section 4.1.5. In this technique, only the power pattern is measured. The aperture phase information is recovered by numerical post-processing. The iteration is started from an assumed wavefront

Zernike function. The power pattern of the antenna is measured at two or more different positions at focus and out-off-focus. The phase change across the aperture is known for a fixed defocusing amount. This small phase change has a great effect on defocused images. Generally, the path length change due to defocus of the telescope by a distance is given by:

$$\delta(x, y; dZ) = dZ\left(\frac{1-a^2}{1+a^2} + \frac{1-b^2}{1+b^2}\right) \tag{8.87}$$

where dZ is the defocus distance, $a = r/(2f)$, $b = r/(2F)$, $r = (x^2+y^2)^{1/2}$, f and F the focal lengths of the primary reflector and the telescope. The positive of the defocusing distance corresponds to moving the secondary mirror away from the primary. The aperture function is then:

$$A(x, y) = \theta(R^2 - x^2 - y^2)I(x, y)\exp[\varphi(x, y) + \delta(x, y; dZ)] \tag{8.88}$$

where θ is the Heaviside step function which describes the truncation of the aperture function at the edge of the primary reflector, φ the phase of the aperture field, R the radius of the primary, and I the power pattern at the defocused position.

8.4.2 Surface Panel Adjusting

Holographic measurement provides real surface shape of an antenna at a particular elevation angle. If the surface measurement is done at two different elevation angles, the surface shape can be predicted for all elevation angles. The task of surface panel adjustment is to achieve the smallest error within the working elevation range of the antenna.

The antenna surface rms error under gravity follows the superposition law. The surface rms error at one elevation angle can be represented as (Woody et al., 1998):

$$S_g = D_y\hat{g} \cdot y + D_z\hat{g} \cdot z \tag{8.89}$$

where $\hat{g} = [g_x, g_y, g_z]$ is a gravitational acceleration vector. If the elevation angle is α, then $g_x = 0$, $g_y = g\sin\alpha$, and $g_z = g\cos\alpha$, where $y = [0, 1, 0]$ and $z = [0, 0, 1]$ are unit vectors in radial (horizon pointing) and axial (zenith pointing) directions of the dish and x is along the elevation axis, and S_g, D_y, and D_z are functions of surface coordinates (x, y).

The panel adjustment is equal to adding a constant term in the above surface error equation. That is:

$$S_g = D_y\hat{g} \cdot y + D_z\hat{g} \cdot z + T \tag{8.90}$$

If the surface after adjustment is a perfect paraboloid at a particular elevation angle α_0, then the constant term T is equal to:

$$T = -D_y \hat{g}_{\alpha 0} y - D_z \hat{g}_{\alpha 0} z \tag{8.91}$$

where $\hat{g}_{\alpha 0}$ is the gravitational acceleration vector at this angle. Using this type of adjustment, the surface will be worse when the elevation angle is away from an elevation of α_0. In the past, it was common to adjust the surface perfectly at about 45° in elevation. In this case, the maximum surface rms error from the horizon and zenith pointing is $[2(H_1{}^2+H_2{}^2)]^{1/2}/2$, where H_1 and H_2 are surface rms errors at the horizon and zenith pointing before the panel adjustment. However, a new surface adjusting method is to optimize the surface error within the whole working elevation range, thus:

$$T = -\frac{1}{2}\left[(\cos\theta_1 + \cos\theta_2)D_y + (\sin\theta_1 + \sin\theta_2)D_z\right] \tag{8.92}$$

where θ_1 and θ_2 are elevation angles at two extreme locations. If these two angles are 0° and 90°, the resulting maximum rms error in between is $(H_1+H_2)^{1/2}/2$. This is $2^{1/2}$ smaller than adjusting the surface perfectly at about 45° elevation. However, in this case, there is no perfect surface at any elevation angle. The above solution is based on the antenna structure being axial symmetrical. For a symmetrical structure, the cross-correlation between D_y (x,y) and D_z (x,y) is zero. The rms error after adjustment can be represented as:

$$\sigma_g^2 = \sum_i^N w_i (D_{yi}\hat{g}\cdot y + D_{zi}\hat{g}\cdot z + T_i)^2 \tag{8.93}$$

where w_i is the weighting function. The adjustment value is determined through optimization after integrating.

8.4.3 Quasi-Optics

In radio wave range, the wavelengths are either greatly in excess of or comparable with the transverse dimensions of a conventional receiver apparatus. In the case that the wavelength is much larger than the transverse dimension of an apparatus, the guided-current method in electronics is used. For the case that the wavelength is in the range of the transverse dimensions, the guided-wave method is used. In the case where the wavelength is shorter than the transverse dimensions of the apparatus, the directed-wave method of classical optics can be used. The transverse direction is perpendicular to the wave vector direction.

In millimeter or submillimeter wavelength range, the wavelengths are in between the guided-wave and the directed-wave ranges. Within this range, the

wave propagation is treated by a new quasi-optics theory, Quasi-optics or long-wave optics, which is a powerful tool in the millimeter and submillimeter wave field.

The transverse field in quasi-optics is not uniform and follows a Gaussian distribution in amplitude (Figure 8.30). The beam is therefore called a Gaussian beam. The field amplitude is:

$$E(r,z) = E(0,z)\exp\left[-\left(\frac{r}{\omega_0(z)}\right)^2\right] \tag{8.94}$$

where z is the coordinate in the beam direction, r the transverse coordinate, and $\omega_0(z)$ the beam waist. The beam waist is the beam transverse width where the field amplitude reduces to $1/e$ of its maximum value. During the beam propagation, the beam waist will change continuously as defined in the formula:

$$\omega(z) = \omega_0(0)\left[1 + \left(\frac{\lambda z}{\pi\omega_0^2(0)}\right)^2\right]^{0.5} \tag{8.95}$$

The radius of curvature of a Gaussian beam wavefront follows:

$$R = z + \frac{1}{z}\left(\frac{\pi\omega_0^2(0)}{\lambda}\right)^2 \tag{8.96}$$

From these formulas, a simple lens equation in quasi-optics becomes:

$$\frac{1}{f} = \frac{1}{R_2} - \frac{1}{R_1} \tag{8.97}$$

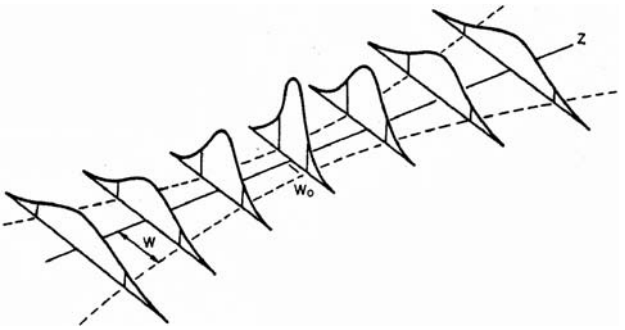


Fig. 8.30. The profile of Gaussian beam amplitude distribution in quasi-optics.

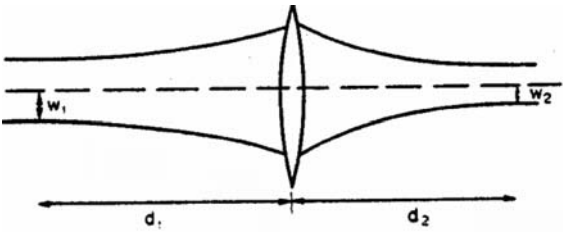


Fig. 8.31. The image formation of a Gaussian beam through a lens.

where R_1 and R_2 are the radii of curvature of the wavefronts at the lens position and f the focal length of the lens (Figure 8.31). Combining the above two formulas, then:

$$R_2 = d_2 + \frac{1}{d_2}\left(\frac{\pi\omega_2^2}{\lambda}\right)^2$$
$$R_1 = d_1 + \frac{1}{d_1}\left(\frac{\pi\omega_1^2}{\lambda}\right)^2 \tag{8.98}$$

By using quasi-optics, a number of antenna characteristics can be easily explained. Quasi-optical components are also widely used in millimeter and submillimeter wavelength range. These include quasi-optical retroreflectors, polarizers, beam splitters, and wave interferometers. These components are much simpler than complex wave guiders. Some of these components, such as the retroreflector formed by two perpendicular surfaces, are frequency independent. For more information on the quasi-optical system, Gaussian beam, and their applications, please refer to Goldsmith (1982, 1998).

8.4.4 Broadband Planar Antennas

Broadband or frequency-independent planar antennas first developed in the 1960s are based on the angular concept, instead of the scaling concept used for other antennas. They are in a plane with no characteristic length and their patterns are self-complementary. The metal part has the same shape as the nonmetal part.

An early broadband antenna is a Bow-tie one. This planar antenna has two triangles formed by two perpendicular lines. The Bow-tie antenna has no main lobe in the direction perpendicular to the antenna plane. The response includes two side lobes away from the normal direction. Another is a spiral antenna consisting of two self-complementary Archimedean spiral arms suspended in free space. This antenna has a main lobe in the normal direction. These two types of antenna have scale limitations. The spiral antennas also have circular polarization and are not suitable for linear polarization purposes.

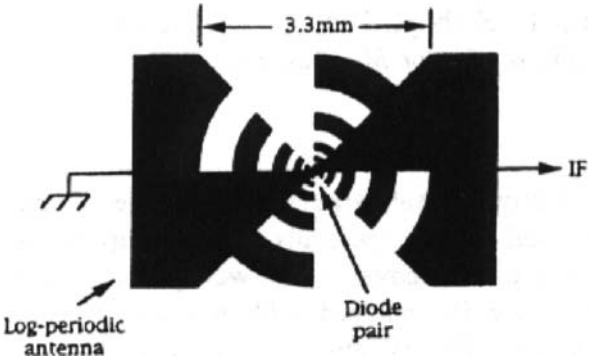


Fig. 8.32. Log-periodic frequency independent antenna used for 26–260 GHz (Kormanyos et al., 1993).

Duhamel and Isbell (Kormanyos et al., 1993) developed a new frequency-independent log-periodic antenna (Figure 8.32). This antenna has a main lobe in the normal direction and retains a linear polarization response. Similar to dipoles, the characteristic length of each tooth is its radius and the radii ratio of two teeth is a constant. If the antenna has a response at a frequency of f_1, then it has responses at frequencies of $f_1\tau$, $f_1\tau^2$, and $f_1\tau^3$, where τ is the radii ratio of two nearby teeth.

Because this type of antenna has a main lobe, it can couple to a Gaussian beam. However, the HPBW of this antenna is quite large, at about 40°, so that the usage is limited. Its directional property can be improved by an extended hemispherical silicon lens as shown in Figure 8.33.

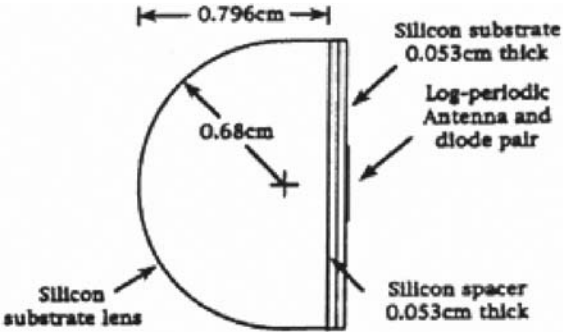


Fig. 8.33. Frequency independent antenna and the lens used for improving its directivity (Kormanyos et al., 1993).

References

Baars, J. W. M., 1983, Technology of large radio telescopes for millimeter and submillimeter wavelength, in Infrared and millimeter waves, 9, 241–091.

Baars, J. W. M., and Mezger, P. G., et al., 1983, Design features of a 10 m telescope for sub-millimeter astronomy, Advanced technology optical telescope II, Proc. SPIE, 444.

Baars, J. W. M. et al., 2007, Near field radio holography of large reflector antennas, IEEE, AP, 49, No 5.

Bazelyan, E. M. and Raizer, Y. P., 2000, Lightning physics and lightning protection, Institute of Physics, London.

Cheng, J., et al., 1998, 12 m antenna design for a joint US-European array, SPIE 3357, 671–685.

Cheng, J., 1998, Forced air cooling at high altitude, ALMA memo 203, NRAO.

Cheng, J., 2000, Design of carbon fiber composite antenna dishes, SPIE Proc. 4015, 597–604.

Cheng, J., 2003, Thermal shape change of some CFRP-aluminum honeycomb sandwiched structures, SPIE Proc. 4837, p 331–347, Hawaii.

Cheng, J., 2006, Thermal deformation of shaped carbon fiber-aluminum core sandwiched structures (I), (II), and (III), ALMA memos 557–559, NRAO.

Cheng, J., et al. 2008, Study on simple CFRP-metal joint failure, SPIE Proc. 7018, 70183F

Chernenkoff, R. A., 1992, Cyclic creep effects in single-overlap bonded joints under constant-amplitude testing, in Cyclic deformation, fracture, and nondestructive evaluation of advanced materials, ASTM STP 1157, eds. Mitchell, M. R. and Buck, O., American Society for testing and materials, Philadelphia, 190–204.

Chiew, S. P., 1993, Features of local proprietary space-frame systems, Steel News Notes, 8 (2), Singapore Structural Steel Society.

D'Addario, L., 1982, Holographic antenna measurements: further technical considerations, 12 m memo 202, NRAO.

Dean, G. D. and Mera, R. D., 2004, Modelling creep in toughened epoxy adhesives, DEPC-MPR 003, June, ISSN: 1744-0270, NPL report.

Fourdeux, A. et al., 1971, In Carbon fibers: their composites and applications, Plastics Institute, London, p 57.

Frostig, Y., Thomsen, O. T., and Mortensen, F., 1999, Analysis of adhesive-bonded joints, square-ended, and spew-fillethigh-order theory approach, J. Eng. Mech., 1298–1307, Nov.

Goldsmith, P. F., 1982, Quasi-optical techniques, In: Infrared and millimeter waves ed. K. J. Button, 7, Academic Press, New York.

Goldsmith, P., 1998, Quasi-optical system, Gaussian beams, quasi-optical propagation, and applications, IEEE Press, New York.

Greve, A et al., 1992, Thermal behavior of millimeter wavelength radio telescopes, IEEE Trans, AP 40, 1375.

Greve, A. et al, 1994, Astigmatism in reflector antennas: measurement and correction, IEEE Trans AP 42, 1345.

Greve, A. et al., 1996a, Coma correction of a wobbling subreflector, IEEE Trans. AP 44, 1642.

Greve, A. et al., 1996b, Near-focus active optics: An inexpensive method to improve millimeter-wavelength radio telescope, Radio Sci, 31, 1053.

Incropera, F. P. and De Witt, D. P., 1990, Introduction to heat transfer, John Wiley & Sons, New York.

Kitsuregawa, K., 1990, Advanced technology in satellite communication antennas, Artech House, Boston.
Kormanyos, B. K. et al., 1993, A planar wideband 80–200 GHz subharmonic receiver, IEEE Trans. Microw Theory Tech. 41, 1730.
Lamb, J. W., 1992, Thermal considerations for mmA antennas, ALMA memo. 83, NRAO.
Lamb, J. W., 1999a, Optimized optical layout for MMA 12 m antenna, ALMA memo. 246, NRAO.
Lamb, J. W., 1999b, Scattering of Solar Flux by Panel Grooves, ALMA memo. 256, NRAO.
Lamb, J. W., 2000, Scattering of Solar Flux by Panel Grooves: update, ALMA memo. 329, NRAO.
Lavenuta, G., 1997, Negative temperature coefficient thermistors, Sensors, May, 46.
Nikolic, B., et al., 2007, Measurement of antenna surface from In- and out-of-focus beam maps using astronomical sources, A&AP, 465, 679.
Radford, S. J. E., et al., 1990, Nutating subreflector for millimeter wave telescope, Rev. Sci. Instrum., 61, 953–959.
Rodriguez, T., 2007, Internal report, NRAO.
Scott, P. F. and Ryle, M., 1977, A rapid method for measuring the figure of a radio telescope reflector, Mon. Not. R. Astro. Soc., 178, 539–545.
Schwab, Fred and Cheng, Jingquan, 2008, Flux concentration during solar observation for ALMA antennas, ALMA memo. 575, NRAO.
Silver, S., 1949, Microwave antenna theory and design, McGraw-Hill, New York, 174–175.
Veer, F. A. et al, 2004, Failure criteria for transparent acrylic adhesive joints under static, fatigue and creep loading, Proc. of the 15th European conf of advanced fracture mechanics for life and safety assessments (ECF 15), Stockholm, 1–8.
Vijayaraghavan, G. et al., 2004, Practical grounding, bonding, shielding and surge protection, Butterworth-Heinemann, London.
Von Hoerner, S., 1967, Design of large steerable antennas, the Astro. J., 72, 35.
Woody, D., et al, 1998, Measurement, Modeling and Adjustment of the 10.4 m Diameter Leighton Telescopes, SPIE Proc. 3357, 474.
Wu, Derick, et al., 2000, Prediction of fatigue crack initiation between underfill epoxy and substrate, Electronic components and technology Conference.

Chapter 9
Infrared, Ultraviolet, X-Ray, and Gamma Ray Telescopes

This chapter provides an overview of the design and principles of infrared, ultraviolet, X-ray, and gamma ray telescopes. The infrared astronomical ground observation is seriously limited by atmospheric thermal emissions. To reduce the sky background noise, a chopping technique and special structural design are usually used. To avoid the atmospheric emission and absorption altogether, balloon, aircraft, and satellite-borne infrared telescopes are used. The emphasis of this chapter is placed on X-ray and gamma ray telescopes. In the X-ray region, grid collimator, focusing collimator, and grazing optics telescopes are used. Detailed discussion is provided on the grazing imaging X-ray optics and its manufacture. In the gamma ray regime, the coded mask collimators, Compton scattering telescope, pair production telescope, air Cherenkov telescopes, and extensive air shower arrays are used. The principles of all these telescopes, including the Davies–Cotton optics used in the air Cherenkov telescope, are also introduced. Important infrared, ultraviolet, X-ray, and gamma ray telescopes are also introduced in this chapter.

9.1 Infrared Telescopes

9.1.1 Requirements of Infrared Telescopes

Infrared (IR) radiation is electromagnetic radiation whose wavelength is longer than that of visible light, but shorter than that of submillimeter waves. Its wavelength is in the range between 0.75 and 350 μm. IR radiation extends well into the submillimeter wave range so that, in astronomy, infrared radiation is often discussed together with submillimeter radiation. The infrared spectrum is further divided into three sub-divisions; near-infrared (NIR) with wavelengths between 0.75 or 1 and 5 μm, middle-wavelength infrared (MWIR) between 5 and 25 or 40 μm, and far-infrared (FIR) between 25 or 40 μm and 200 or 350 μm. Sometimes, near- and middle-infrared regimes are simply called near-infrared regime. Short-wavelength infrared (SWIR) and long wavelength infrared

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_9,

(LWIR) are also used in some references to denote wavelengths between 1.4 and 3 μm and between 8 and 15 μm, respectively. In the infrared region, wave numbers instead of wavelength or frequency are often used, especially in the analysis of atomic spectra. The definition of the wave number $\tilde{\nu}$ is:

$$\tilde{\nu} = 1/\lambda = \nu/c \tag{9.1}$$

where λ is the wavelength, ν the frequency, and c the velocity of light in vacuum. The wave number has the dimensional unit of m^{-1}.

Infrared is a thermal form of radiation and its detection is evidence of a thermal process. In the detection of very weak celestial infrared radiation, three important factors are; the sensitivity of the detector, the atmospheric influences, and the radiation of the telescope structure itself. Detector sensitivity is not the subject of this book. In this section, only the latter two aspects are discussed.

The atmospheric influences in the infrared region include atmospheric attenuation and atmospheric emission. Attenuation is the reduction in amplitude and intensity of a signal radiation as it propagates through a medium. The attenuation includes scattering and absorption. Atmospheric attenuation in the infrared is caused mainly by molecules of water vapor, carbon dioxide, ozone, methane, nitrous oxide, and carbon monoxide occurring below the tropopause where the troposphere ends and the stratosphere begins. The altitude of the tropopause is between 7 and 20 km depending on the location on the earth. These atmospheric molecules can selectively absorb infrared radiation. Only in a few narrow wavelength ranges can infrared light make it through to ground level.

Figure 9.1 shows the absorption spectrum caused by major molecules in the atmosphere, in the regime from 1 to 16 μm. In the figure, the solar spectrum at sea level is a sum of all individual spectra. Because of atmospheric absorption in

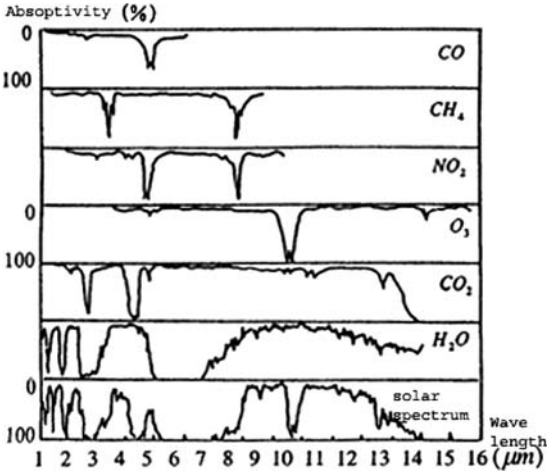


Fig. 9.1. Infrared absorption caused by atmospheric molecules.

the infrared regime, there are only a few windows in near-infrared and middle-infrared regimes. These windows are expressed alphabetically as I (0.75–0.92 μm), J (1.1–1.4 μm), H (1.45–1.8 μm), K (1.9–2.5 μm), L (3.05–4.1 μm), M (4.5–5.5 μm) and Q (17–28 μm) bands. The atmospheric absorption is a function of altitude. High mountain altitude suffers less from this absorption. For wavelengths longer than 25 μm, the atmosphere is nearly opaque even at high mountain altitudes, except for some very narrow windows around 34 and 350 μm. Figure 9.2(a) shows the transmissivity at four different altitudes above the ground. For an altitude of 28 km above ground, the atmosphere is almost transparent. Therefore, infrared observations have to be performed from high mountains, balloons, airplanes, rockets, or space.

The atmospheric emission is very strong in infrared regime in comparison with radiations from celestial objects. This atmospheric emission sets the lower limit for observational background noise. The atmospheric infrared emission peaks at a wavelength of about 10 μm. The atmosphere is not a black body. Its flux density is:

$$B = \varepsilon \cdot P \tag{9.2}$$

where P is the Planck function, which can be calculated using an effective temperature of the atmosphere, and ε the atmospheric emissivity. The emissivity of a black body is unity. The relationship between the emissivity and the transmissivity of an object is:

$$\varepsilon = 1 - T_r \tag{9.3}$$

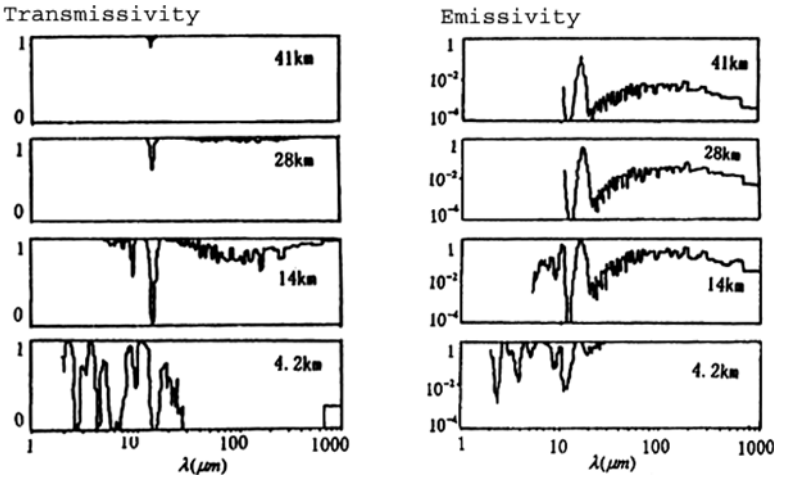


Fig. 9.2. Infrared transmissivity (**a**) and emissivity (**b**) of the atmosphere at four altitudes.

where T_r is the transmissivity. Figure 9.2(b) shows the emissivity of the atmosphere at four different altitudes above the ground. In addition to the atmospheric emission, the variations in emissivity and path length on timescales of the order of one second also introduces "sky noise" which are fluctuations in total power or phase on detectors. Sky noise causes systematic errors in the measurement of astronomical sources.

In infrared observations, radiation from the telescope structure is also a background noise source. For a black body, at an absolute temperature T, the intensity of radiation emitted is:

$$P = \frac{2hc^2/\lambda^5}{e^{hc/\lambda kT} - 1} \tag{9.4}$$

where $h = 6.626 \times 10^{-34} Js$ is a Planck constant and $k = 1.38 \times 10^{-23} J/K$ a Boltzmann constant. Figure 9.3 shows relative radiation spectra of a black body with different absolute temperatures. For objects at a temperature of 300 K, the maximum of the emitted radiation is at 10 μm of the infrared regime.

Total background noise in an infrared observation can be expressed as:

$$S = \varepsilon \cdot \Omega \cdot P(\lambda, T)\Delta\lambda \tag{9.5}$$

where Ω is the combined throughput of the telescope and detector and $\Delta\lambda$ the bandwidth. For a diffraction-limited field of view with a circular aperture, the background noise is:

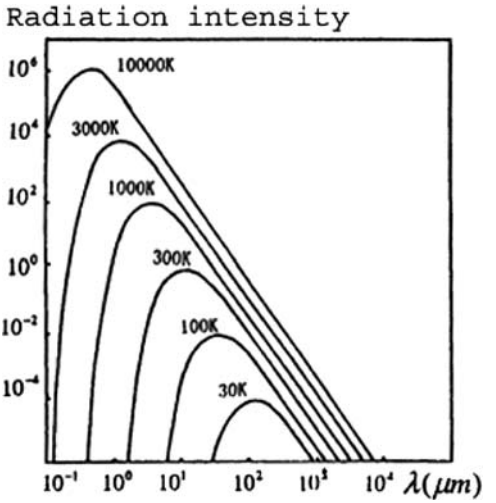


Fig. 9.3. Relative radiation spectra of a black body at different temperatures.

$$S = \frac{4.45\varepsilon \cdot C_1}{\pi\lambda^2(e^{C_2/\lambda T} - 1)} \frac{\Delta\lambda}{\lambda} (\text{unit : Watt}) \tag{9.6}$$

where $C_1 = 3.74 \times 10^{-4}\,\text{W}\mu\text{m}^2$ and $C_2 = 14,400\,\mu\text{mK}$. If $\lambda = 30\,\mu\text{m}$, $T = 270\,\text{K}$, $\varepsilon = 0.01$, and $\Delta\lambda/\lambda = 0.1$, then $S = 1.19 \times 10^{-10}\text{W}$. This background noise is very often 10^7 times higher than the signal detected.

Strong noise level in infrared observation places different requirements on infrared telescope design. These are: (a) it is necessary to reduce blockage with high emissivity within the detector's field of view; (b) it is necessary to maintain a stable signal (or noise) level anywhere within the field; (c) it is favorable to reduce both structural temperature and temperature gradient in an infrared telescope; and (d) it is necessary to use a chopping or other mechanisms to remove the sky background noise.

With these requirements, infrared telescope design is different from optical telescope design. Infrared observation by simply using an existing optical telescope is usually not an ideal solution. For airborne, balloon-borne, or space infrared telescopes, there are additional design requirements.

9.1.2 Structural Properties of Infrared Telescopes

By the end of the last century, great progress had been made in infrared detector technology. Old single-cell detectors have gradually been replaced by new two-dimensional ones. This technology change also has a great impact on infrared telescope design. Modern infrared telescopes loaded with infrared CCDs may compensate sky background noise without using a mechanical chopping device. The field of view used is also enlarged. However, the major task of reducing background radiations in the field of view for infrared telescopes remains unchanged.

Infrared telescopes usually have a relatively large focal ratio and a small field of view. With a large focal ratio, the size of the secondary mirror is small. The central aperture blockage and the thermal background noise from the blockage are reduced. A smaller secondary mirror is also favorable for a mechanical chopping device, which is necessary when the single-cell detector is used. However, the field of view should not be smaller than the size of the first ring of the diffraction pattern; otherwise some radiation power from the source is lost.

With the usage of two-dimensional detectors, the focal ratio and the field of view can be slightly larger. The focal ratio can reach $f/7$. The size of the secondary mirror in infrared telescopes is always smaller than the size required without vignetting. The secondary mirror forms the entrance pupil of the telescope. In this way, the detector only sees through the primary mirror which has a high reflectivity and thermal radiation from many room temperature parts, such as the primary mirror cell and the tube frames, will not be in the field of view.

In infrared telescopes, there are no baffles located inside the optical path. The baffles, especially the black ones as used in optical telescopes, are main

sources of infrared background noise. The secondary mirror cell and support mechanism should also be located behind the mirror surface so that direct radiation from them is blocked from the detector.

The secondary mirror support vanes and upper part of the tube structure should be polished and coated with high reflectivity coatings. Rough and low reflectivity surfaces have higher emissivity. Figure 9.4 shows the relationship between surface roughness and emissivity for aluminum. The secondary mirror support vanes should have a 'T'-shaped cross section so that the projection area of vanes will not change as the field angle changes. The background radiation from the vanes will remain constant all over the field. In the center of the secondary mirror, a reflecting spherical surface or a tilted mirror can be used for reducing or eliminating the central hole radiation of the primary mirror.

The primary and secondary mirrors should have a very-low emissivity coating layer. Gold or silver are usually used. The emissivity of silver coating of 0.05 is better than that of aluminum which is about 0.1. However, silver coating alone does not last long so that a protection layer on top is necessary. Gold has an even lower emissitivity of 0.03.

To reduce thermal radiation, some modern infrared telescopes use a heat pipe cooling system for their primary and secondary mirrors. Smaller mirrors near the detector area are cooled inside a cryogenic container.

Infrared telescopes normally use the chopping technique for compensating sky background noise. The main reasons are: (a) It is much easier to amplify an alternating current than a direct current; (b) some detectors are only sensitive to changes of temperature. If the incident flux does not change, the response may drift; and (c) chopping provides necessary discrimination against a bright sky

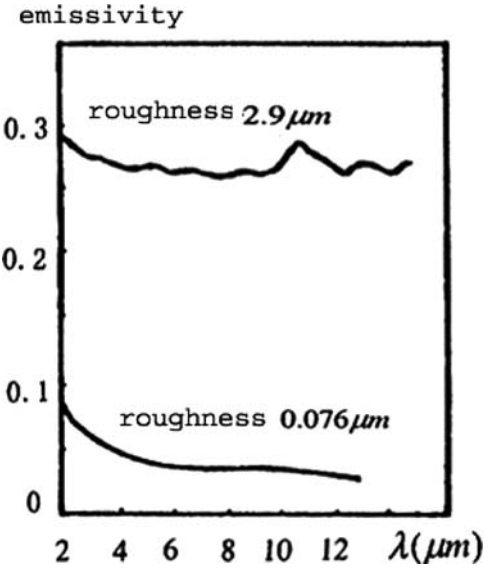


Fig. 9.4. Emissivity of aluminum surfaces with different roughness in normal direction.

background. In some cases, chopping can be performed through data processing when an infrared CCD is used.

Using the chopping technique, two neighboring patches of sky (beams) are presented alternately to the detector and their difference is taken electronically by phase sensitive rectification. If the sky background is identical within two beams, it will cancel out and only the signal from a source located in one beam is recorded. The frequency of chopping between two beams is between 1 and 1,000 Hz depending on the response time of the detector. Mechanical chopping may produce structural vibration. To suppress vibration in large infrared telescopes, reactionless chopping mirror design (Section 8.2.3) and vibration damping may be used. Using a chopping secondary instead of other chopping mirrors has the advantage of less noise as no additional mirror is required. Passive damping is usually used in the thin, very long secondary mirror vane structure of large infrared telescopes. The damping is through constrained viscoelastic damping layers (Section 3.4.4).

Chopping can also be realized by using other mechanisms including a rotating segmented mirror, rotating chopper, modulation disk, or fast switching of the whole telescope (Figure 9.5). A rotating chopper is used together with a fixed mirror. When the observation is in the far infrared regime, these two mirrors may have different temperature and the edges of the segments may emit strongly and may reflect warm objects, producing thermal noise. The chopping angle is also limited as the two beams may include different portions of the telescope optics. This can be avoided when the chopping mirror is located at the telescope exit pupil. A chopping Gregorian secondary mirror is near the exit pupil and is used in some millimeter telescopes. During chopping of this Gregorian mirror, the illumination on the primary surface remains unchanged. Therefore, a large angle of chopping is possible. The background noise during chopping remains constant. Fast switching of the whole telescope is never used in infrared telescopes.

In infrared observation; the response to radiation is a thermal process. Therefore, the chopping frequency of an infrared telescope also depends on the time constant of the detector:

$$2\pi f \tau << 1 \tag{9.7}$$

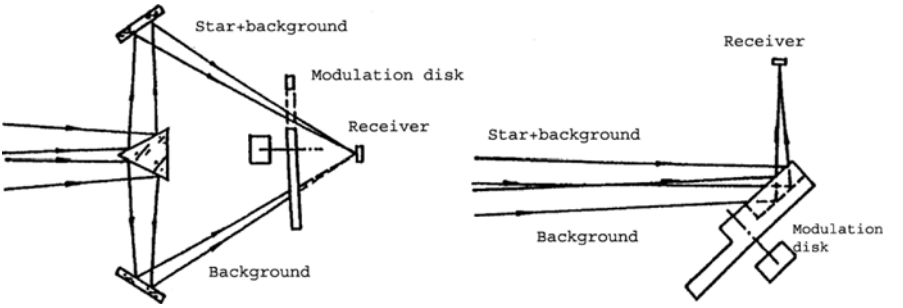


Fig. 9.5. Two chopping methods used in infrared observation.

where f is the chopping frequency, $\tau = C / G$ the time constant of the detector which is a function of the heat capacitance C of the detector pixel and the conductivity G of the detector material. During the chopping, it is necessary to avoid multiples of 50 or 60 Hz preserved for the power line. It is also undesirable to use a too low frequency as $1/f$ noise may become significant. With fast development of infrared CCDs, the mechanical chopping technique is gradually replaced by computer post-processing.

Another important technique in using existing optical telescopes for infrared observation is to have a cold Lyot stop at the exit pupil of the optical path so that all unwanted radiations are blocked and absorbed at this plane. The background noise is reduced. This cold stop is also called a pupil plane baffle. The alignment of the stop is usually critical. The Lyot stop has another application in segmented mirror telescopes. It can eliminate the diffraction effect from the gaps between mirror segments.

In infrared telescopes with a chopping secondary, the temperature gradient on the primary mirror surface may produce a repeated pattern of thermal background (usually happens in a thick primary mirror). This pattern becomes a "false signal," which is different from any other random background noise. It has two effects: (a) the signal-to-noise ratio is difficult to increase by simply increasing the integration time; and (b) the false signal is noncoherent. The only way to reduce or eliminate it is through careful calibration. The most difficult problem is the variation of this temperature gradient during the observation. The signal offset caused by a temperature gradient ΔT is given by:

$$\Delta\phi = 4.67\varepsilon\lambda^3 \frac{\Delta\lambda}{\lambda}\frac{dP}{dT}\frac{dS}{S}\Delta T \tag{9.8}$$

where P is the Planck function, S the mirror surface area, and T the temperature. When the temperature gradient is small, one can take only the linear part of the Planck function so that:

$$\frac{dP}{dT} = \frac{C_1 C_2}{\pi\lambda^6 T^2}\frac{e^{C_2/\lambda T}}{\left(e^{C_2/\lambda T}-1\right)^2}\left(W\cdot micron^3\cdot steradian\right) \tag{9.9}$$

where $C_1 = 3.74\cdot 10^{-4}\ \mathrm{W}\cdot\mu\mathrm{m}^2$, $C_2 = 14{,}400\ \mathrm{K}\ \mu\mathrm{m}$.

The mirror surface accuracy required for infrared telescopes is lower than that for an optical telescope. Therefore, metal or other materials are often used in infrared telescopes. Beryllium, aluminum, and silicon carbide mirrors are also used in space infrared telescopes. CFRP aluminum honeycomb sandwich mirrors are used. The weight and cost of these mirrors are lower than those of optical ones.

In the infrared regime, sky background radiation during day and night is similar. Therefore, infrared observation can be carried out in daytime or

moon-lit nights. In daytime, the use of optical guiding stars is difficult so that some infrared telescopes may require higher pointing and tracking ability without the help from the guiding star in comparison with optical telescopes.

The largest ground-based infrared telescope is the 10 m Keck telescope, the largest airborne infrared telescope is the 2.5 m SOFIA telescope, the largest balloon-borne infrared telescope has a diameter of about 1 m, and the largest rocket infrared telescope has a diameter of about 0.15 m. Generally, airborne and ground-based infrared telescopes have higher pointing accuracy. Balloon-borne telescopes have lower background noise and relatively high loading capacity.

9.1.3 Balloon-Borne and Space-Based Infrared Telescopes

For effective infrared observation, ground-based infrared telescopes have to be at an altitude above 3 km. The height for airborne telescopes is above 25 km, that for balloon-borne telescopes is above 50 km, and that for rocket telescopes is above 100 km. However, all these telescopes still suffer from some atmospheric influences. To remove all atmospheric influences, space infrared telescopes are needed.

Balloon-borne infrared telescopes played a very important role in early infrared astronomy. Large balloons have high loading capacity. Usually, a large balloon has a weight of about 5,000 kg, a volume of about 60,000 m^3, and a loading capability of about 500 kg. Below the balloon, a gondola (basket or payload) is attached with a cable ladder. The infrared telescope is located inside the gondola.

The balloon obscures the telescope from the zenith sky at an angle of about 20°. If extremely long suspension is used, the zenith angle obscuration can be smaller, down to 2°. In the balloon-borne telescope design, gondola stabilization is a main design consideration. The dominant motion of a balloon system is the induced motion from the stratospheric winds at an altitude between 10 and 50 km. The wind speed increases very rapidly at an altitude of about 30 km. A typical wind speed at that altitude is about 45 km/h. The motion that a gondola must be isolated from is a slow rotation of the entire balloon system. This rotation is not constant and a reverse rotation is possible occasionally. The largest speed of this motion is about one revolution every eight minutes. Another gondola motion is a pendulum one due to the suspension from the balloon. The period of this motion is about 15 s. Rocking or double pendulum modes of the system may also occur at the system mass center with a frequency of about 1 to 2 Hz. Usually, the azimuth stability of the gondola is maintained through a magnetometer. Gyroscopes or star trackers can also be used on azimuth and elevation axes for short-term pointing correction (Section 5.2.1).

Balloon-borne telescope structure is different from other telescopes. The simple alt-azimuth system is usually not satisfactory. The preferred mount system is a three-axis one with an azimuth, an elevation, and a cross-elevation

axis. The elevation axis is above the azimuth one and the elevation axis supports a yoke where the cross-elevation axis is located. The elevation and cross-elevation axes provide the tube motion in elevation.

Balloon-borne telescopes are usually small in size and are less in cost so that they can be used at the starting stage for high-altitude infrared observation. Airborne telescopes are relatively larger and more expensive than the balloon-borne ones. They usually serve as major tools for high-altitude infrared observation.

Two important airborne infrared telescopes are the Kuiper Airborne Observatory (KAO) and the Stratospheric Observatory For Infrared Astronomy (SOFIA). The KAO was a 0.91 m infrared telescope sat on a modified C-141 cargo airplane. It flew in a high stratosphere layer with an altitude of 13 km above the lower troposphere. At this altitude, 95% water vapor absorption of infrared waves from the atmosphere is removed.

The telescope of the KAO is set on top of four vibration absorbers above a spherical air bearing. The telescope used a star tracker and gyroscopes for its pointing and tracking control. The pointing accuracy achieved was about 2 arcsec. The KAO observations were started in 1971 and were terminated in 1995.

The SOFIA project is a new 2.5 m airborne infrared telescope, the largest among all airborne telescopes. It was built jointly by NASA and the German astronomical community. The telescope with an oversized 2.7 m primary mirror and a small 0.4 m secondary mirror is set on a Boeing 747-SP aircraft. It works over a wide wavelength range from 0.3 μm to1.6 mm. The first detection of star light was done when the airplane was on the ground in 2004 and its full operation will begin in 2009.

The telescope is located in an open cavity in the aft section of the aircraft. It has a limited sky coverage of elevations between $+20^\circ$ and $+60^\circ$ restricted by its enclosure. The telescope is expected to have a pointing accuracy better than 1 arcsec and a tracking accuracy better than 0.5 arcsec. The typical seeing at high altitude is about diffraction limited for all wavelengths longward to 15 microns. This telescope has a similar vibration control as the KAO one.

Space infrared telescopes began in the 1980s. Space telescopes remove all the influences from the earth's atmosphere. In addition, space infrared telescopes usually have their mirrors cooled to a very low temperature so that the background noise is further reduced.

In the early days, space infrared telescopes had a very small mirror diameter. The first-ever space infrared survey telescope was a 0.57 m telescope named the Infrared Astronomical Satellite (IRAS) launched on January 25, 1983. This was a joint US, Netherlands, and UK project. IRAS was an $f/9.6$ Ritchey–Chretien telescope mounted in a liquid helium cooled cryostat (Figure 9.6). It had a field of view of 30 arcmin. The telescope mirrors were made of beryllium and cooled to approximately 3 K. The telescope carried 72 kg of liquid helium.

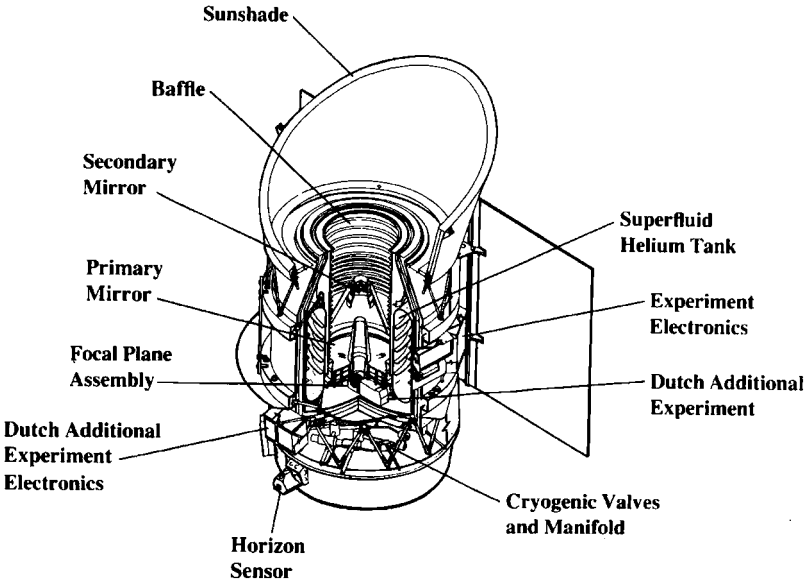


Fig. 9.6. Infrared Astronomical Satellite (IRAS) (NASA).

The focal plane assembly of this telescope contained survey detectors, visible star sensors for position reconstruction, a low-resolution spectrometer (LRS), and a chopped photometric channel (CPC). The focal plane assembly was located at the Cassegrain focus. After nine months of operation, the liquid helium was exhausted and observation was terminated. IRAS detected about 350,000 infrared sources, increasing the number of cataloged astronomical sources by about 70%.

After IRAS, there came several small infrared space telescopes. The Infrared Space Observatory (ISO) of the European Space Agency (ESA), a 60-cm diameter Richey–Chretien system with an overall focal ratio of $f/15$ was one among these. With 2,200 liters of liquid helium for cooling, ISO was launched in November 1995. The telescope was accidentally pointed to the warm earth in May 1996. The temperature of the instruments increased to 10 K. However, this telescope's working life was still much longer than the expected 18 months.

The Spitzer Space Telescope [formerly Space Infrared Telescope Facility (SIRTF)] which had a diameter of 0.85 m was launched into orbit by a Delta rocket in August 2003 (Figure 9.7). During its mission, the Spitzer space telescope obtained images and spectra between wavelengths of 6.5 and 180 μm. The telescope together with three science instruments was cryogenically cooled to a low temperature of 5.5 K. Spitzer space telescope had an expected mission life of 5 years, but it is now still working in orbit.

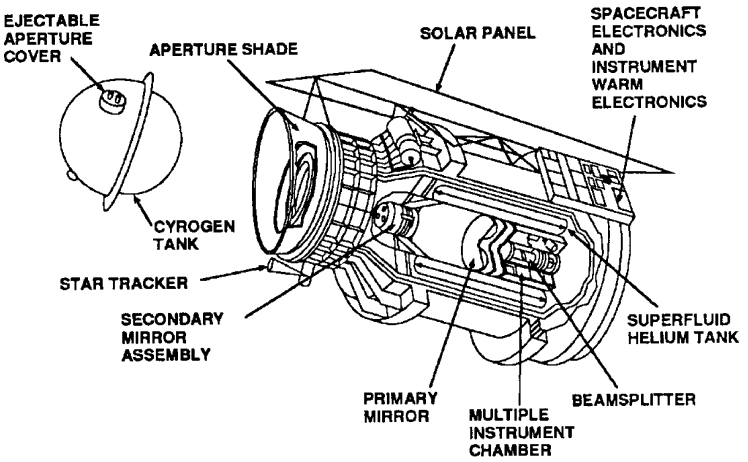


Fig. 9.7. Early design of Spitzer Space Telescope (NASA).

Astro-F or IRIS (Infrared Imaging Surveyor) were the old names of a Japanese space infrared telescope used to operate on a sun-synchronous polar orbit. After its launch in 2005, the telescope was renamed as AKARI. AKARI was a 70-cm infrared telescope cooled to 6 K by liquid helium. However, it ran out of helium by August 26, 2007. During the mission period, AKARI completed the far-infrared all-sky survey covering about 94% of the sky and the mid-infrared survey on more than 5,000 individual pointed observations.

In the far infrared and submillimeter wave regime, the Herschel Space Observatory is scheduled to be launched to an earth-sun L2 point April, 2009. Its former name was the Far Infrared and Submillimeter Telescope (FIRST). The telescope has a 3.5 m primary mirror made by brazing together 12 silicon carbide (SiC) mirror segments. The mirror is the largest silicon carbide mirror ever made. To provide the mechanical stability during the mirror manufacture, temporary stiffening ribs were positioned on the internal surface of the segments in addition to the external surface stiffeners that are part of the final mirror configuration. After the assembly process was completed, the temporary stiffening ribs were removed from the mirror blank. The thickness of the final mirror is only 3 mm. The mirror reaches the correct profile through polishing. The reduction of the mass from the mirror blank to the final mirror is 480 kg, with a final mirror weight of 240 kg. The post-machining surface has a rms error of 170 μm. The surface final rms error is only 1.5 μm and the surface roughness is less than 30 nm. The complete telescope has a total wavefront rms error less than 6 μm. The mass of the complete telescope is 315 kg.

Of course, the most important space infrared telescope project of all is the James Webb Space Telescope (JWST) as discussed in Chapter 5.

9.2 X-Ray and Ultraviolet Telescopes

9.2.1 Properties of X-Ray Radiation

Visible light is divided into the different colors of red, orange, yellow, green, blue, and violet in wavelength order. Farther away from the violet radiation, electromagnetic waves have shorter and shorter wavelength, or higher and higher photon energy. Adjacent to violet radiation is the ultraviolet (UV) one with wavelengths from 10 to 390 or 400 nm. In the near ultraviolet (NUV) regime (200–400 nm), astronomical observations are similar to those in the optical regime. However, the absorption of the mirror coating material is similar to that in other UV and X-ray bands. Higher energy UV radiation is named extreme ultraviolet (EUV), or vacuum ultraviolet (VUV). The wavelength is in the range 10 to 200 nm. The observations in this higher energy UV regime are similar to those in the X-ray regime. When a reflecting telescope is used for ultraviolet observation, it is necessary to coat the mirror with a lower absorption gold or iridium layer. The sun is a source of ultraviolet radiation; however, the atmosphere absorbs almost all UV radiation except that with wavelengths between 310 and 400 nm. The wavelength of this part is near to the visible wavelength. The ultraviolet absorption is mostly due to the ozone layer (Figure 9.8). Since ultraviolet telescopes are basically similar to optical or X-ray telescopes, ultraviolet telescope structure and design are omitted from this book. However, space ultraviolet telescope projects are covered in the later Section 9.2.5.

Further away from UV radiation in the electromagnetic wave spectrum is the X-ray regime. The wavelength of X-rays is from 0.01 to 10 nm, which is about the size of an atom. X-ray radiation has even higher energy. The energy of

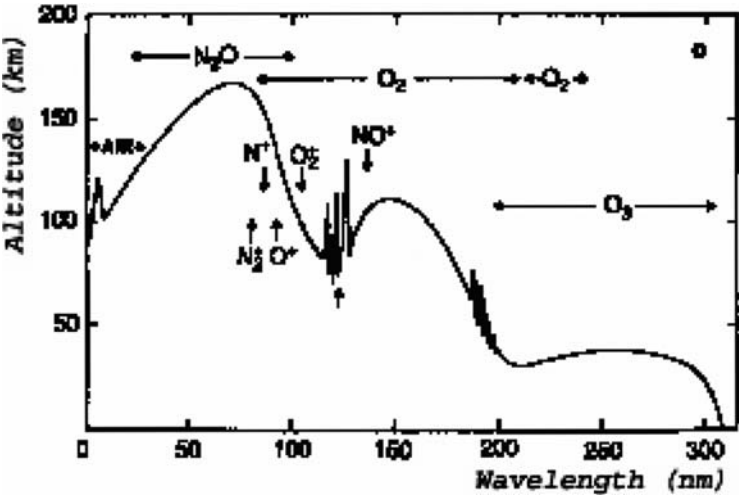


Fig. 9.8. The altitude reached by ultraviolet radiation from the sky.

electromagnetic waves is described in a unit of electron volt (eV). An electron volt is the amount of energy acquired by a single electron when it is accelerated through an electrostatic potential difference of one volt in a vacuum, which is approximately 1.602×10^{-19}joules. The relationship between the electron volt and the wavelength is:

$$\lambda(\text{nm}) \cdot E(\text{eV}) = 1240 \tag{9.10}$$

where E (eV) is the energy of the photon in eV and λ the wavelength in nm. According to the energy of the radiation, the X-ray is further divided into soft and hard X-ray bands. Soft X-rays have energies from 120 to 1,200 eV or have wavelengths from 10 to 1 nm. Hard X-rays have energies from 1.2 to 120 keV or have wavelengths from 1 to 0.01 nm. However, the division of soft and hard X-rays is not strict. In some references, X-rays may include a small portion of soft gamma rays with the wavelengths from 0.01 to 0.001 nm.

There are a number of mechanisms that have been discovered in astronomy for converting energy into high frequency X-rays. These are super-hot black body (tens of millions degrees) radiation, electron synchrotron radiation, inverse Compton scattering, and thermal bremsstrahlung. The word "bremsen" means braking and "strahlung" radiation. The most common situation for bremsstrahlung is the emission from hot gas as the electrons collide with nuclei due to their random thermal motions. The X-ray sources in astronomy include the sun, supernova remnants, pulsars, and active galactic nuclei.

When X-rays or gamma rays interact with materials, there will be photoelectric, Compton, or electron pair effects. The photoelectric effect is a phenomenon in which electrons are emitted from material after the absorption of energy from electromagnetic radiation such as X-rays or visible light. The emitted electrons are referred to as photoelectrons. When a photon with energy and momentum collides with a stationary electron, some of the energy and momentum is transferred to the electron, but both energy and momentum are conserved in this elastic collision. This phenomenon is the Compton effect. If the energy of the photon is more than twice that of the electron stationary mass, the photon that collides with the material will disappear and it produces an electron/positron pair. This process is the electron pair effect.

In the X-ray regime, the photoelectric and Compton effects are important and, in the high-energy gamma ray regime, the electron pair effect is dominant. The absorptivity of material is a combination of these three effects. Figure 9.9 is the total absorptivity τ of some materials caused by the photoelectric, Compton, and electron pair effects. The contributions from these effects for the material Pb are also shown in this figure, where v is radiation frequency, $\hbar = h/(2\pi)$ the Dirac constant, h the Planck constant, and $mc^2 = 0.511\,\text{MeV}$.

The reduction in radiation intensity after interaction between an X-ray or gamma ray and a material is:

$$I = I_0 \exp(-ux) \tag{9.11}$$

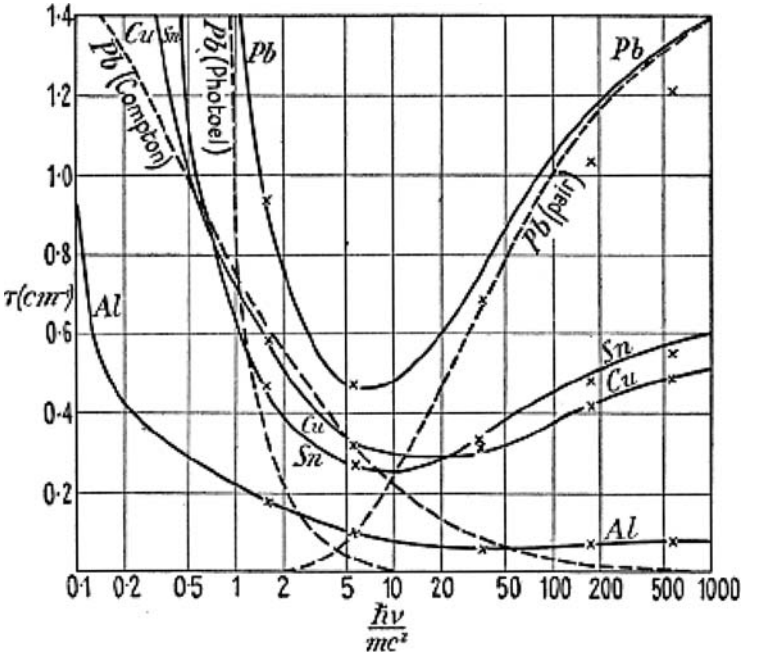


Fig. 9.9. Total absorptivities of materials caused by photoelectric, Compton, and pair production effects. The dashed lines are individual contributions from the three effects for Pb and the solid lines are combined effects for Ph, Al, Sn, and Cu (Heitler, 1954).

where I_0 is the intensity of the incident radiation, x the thickness of the material, and u a linear absorption coefficient. In the high frequency regime the absorption coefficient is roughly proportional to the density of the material. The mass absorption coefficient is defined as $u_m = u/\rho$, where ρ is the density of the material.

X-ray absorption in the atmosphere depends on the energy of radiation. Hard X-rays can penetrate to an altitude of 20–40 km while soft X-rays only reach higher altitude (Figure 9.10). Therefore, all X-ray observations have to be carried out from rockets or from space. Not only the earth's atmosphere, but also interstellar mediums absorb strongly soft X-ray and UV radiations. Therefore, most diffuse soft X-rays detected are from nearby space of about tens of parsec ($3.0847 \cdot \times 10^{16}$ m) range.

In the X-ray regime, the indices of refraction of all materials are slightly less than unity so that the lens optical system for X-ray observation is impossible. On the other hand, strong absorption of materials in the X-ray regime also makes reflector optical systems for X-ray observation totally different from those in other wavelengths. The complex index of refraction $n - ik$ of a metal material at frequencies which are higher compared with the relaxation time of conduction electrons ($\sim 10^{-9}$ s) is given by (Bennet, 1979):

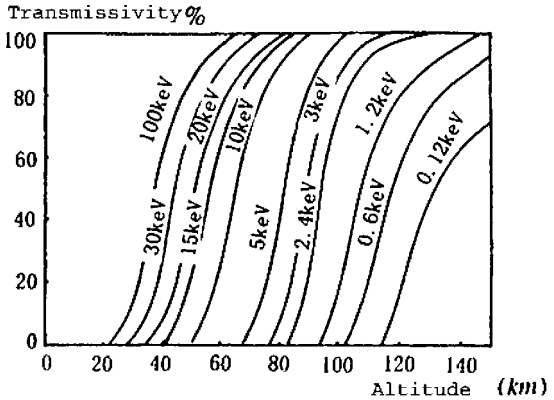


Fig. 9.10. Transmissivity of the atmosphere in the X-ray regime.

$$n^2 - k^2 = 1 - \frac{4\pi N e^2}{m\omega^2} \tag{9.12}$$

where n is the real part of the complex index (n is much greater than k), k the extinction coefficient (the imaginary part of the index) of the material, N the number of participating electrons per unit volume, m the effective mass of the electron, e the electron charge, and ω the circular frequency of the incident radiation. The inner electrons of the material are too tightly bound to be affected by the visible photons. However, all electrons having their binding energy less than hv in the X-ray regime participate in the refraction index, where v is the frequency of the radiation and h the Planck constant. Therefore, the number N in Equation (9.12) is close to:

$$N = \eta Z \tag{9.13}$$

where η is the number of atoms per unit volume and Z the atomic number of the material. Since $k << n < 1$, $n + 1 \approx 2$, then:

$$(1 - n) \rightarrow \frac{2\pi\eta Z e^2}{m\omega^2} \quad (if: N \rightarrow \eta Z) \tag{9.14}$$

To maximize $(1 - n)$, the atomic number of the material selected for the reflector should be as large as possible. In reflecting optics, a critical angle is the angle that provides total internal reflection without any flux loss. From Snell's law, the critical angle ϕ measured from the reflecting surface, as is usual in the case of an X-ray system, is given by:

$$\phi = \sqrt{2(1 - n)} \rightarrow \frac{2e}{\omega}\sqrt{\frac{\pi\eta Z}{m}} \tag{9.15}$$

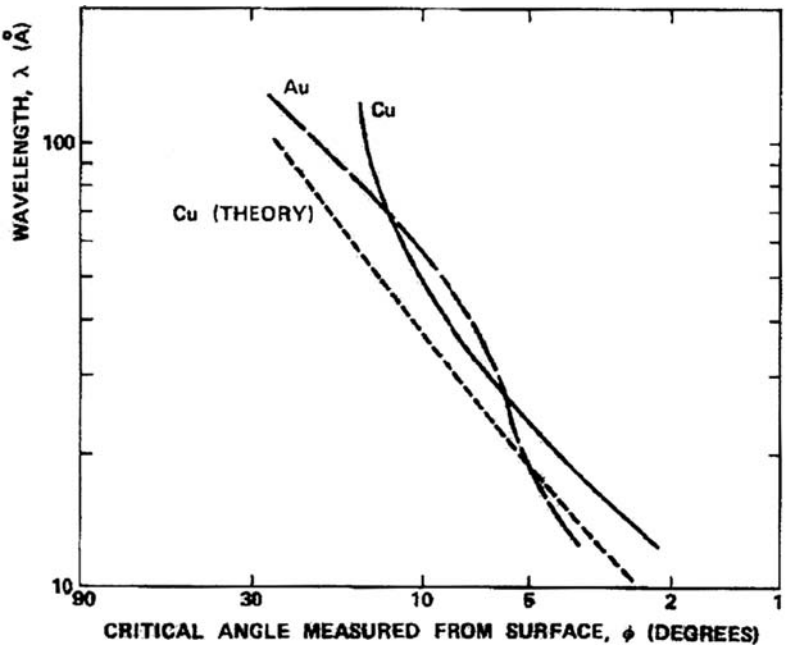


Fig. 9.11. The critical angles measured from a surface for copper and gold in the soft X-ray regime [theoretical values are calculated from Equation (9.17)].

This formula gives the theoretical critical angle of a material in terms of fundamental material constants. In the soft X-ray regime, the agreement between theory and measurement for metals, such as copper, is fairly good (Figure 9.11). In general, the critical angles for metals in the X-ray regime are quite small, only about 1 to 2° from the reflector plane. Another simplified formula of the critical angles in the X-ray regime is $\phi(\text{deg}) \approx \rho^{1/2}/\text{E}$, where E is the X-ray photon energy in keV and ρ the density in g/cm^3.

The Total Integrated Scattering (TIS) of a surface used in the visible regime is represented by a scalar scattering theory. The TIS is the ratio between diffuse reflectance and the total reflectance (Bennett and Mattsson, 1999). The total reflectance includes both specular and diffuse reflectances. The TIS expressed in surface rms micro-roughness (similar to but slightly different from the surface rms error which includes both micro-roughness error and large-scale surface errors) and angle of incidence is:

$$TIS = 1 - \exp\{-[4\pi\delta(\sin\psi)]^2\} \cong [4\pi\delta(\sin\psi)/\lambda]^2 \tag{9.16}$$

where δ is the surface rms micro-roughness and ψ the incident angle measured from the reflector surface. Using this equation in the X-ray regime gives the relationship between the wavelength, the surface micro-roughness, the total

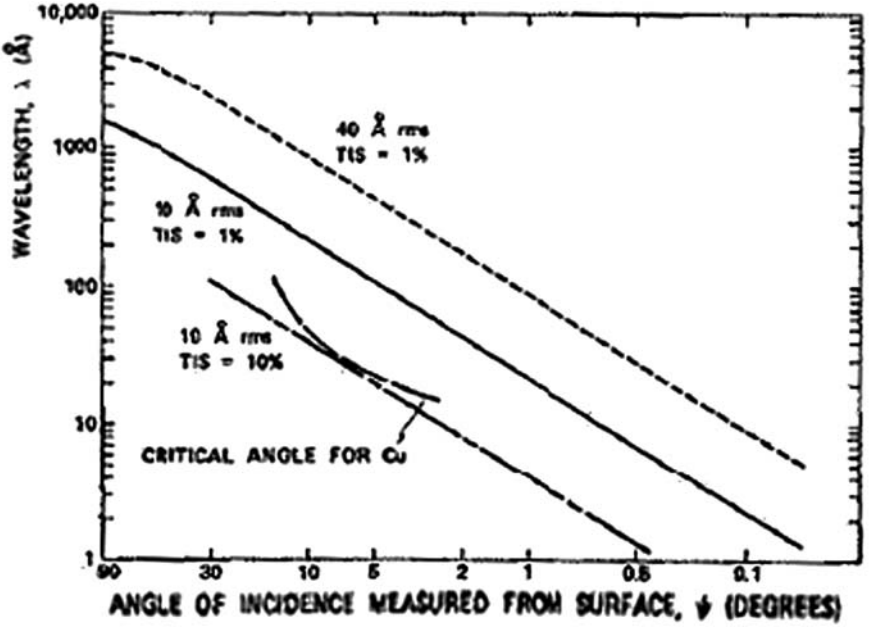


Fig. 9.12. Relationship between surface roughness, total integrated scattering, and incident angle predicted by Equation (9.16). The curved line is the measured value.

integrated scattering, and the incident angle. This relationship is valid as long as the polarization effect is negligible. Figure 9.12 shows the curves derived from this formula. From this figure, it follows that the best surface which can be used at near normal incidence in the visible regime is the same as that at a grazing incidence in the soft X-ray regime. Two of the three lines are for a 1-nm rms surface micro-roughness with 1 and 10% TIS value. The other line is for a 4-nm rms surface micro-roughness with 10% TIS value. The curved line is the measured critical angle of a copper surface of 1 nm micro-roughness in 1 to 10 nm wavelength range. For a high-quality optical surface used either in the visible or X-ray regime, the rms micro-roughness should be 1 nm or less.

Absorption and scattering in the X-ray regime makes the reflection only possible at small grazing angles. This leads to the X-ray grazing telescopes. Besides the grazing telescopes, narrow bandwidth X-ray normal reflecting optics can also be realized through multi-layer dielectric coating. The application of the latter is photon frequency limited in small bandwidth.

For understanding the required surface micro-roughness of X-ray grazing telescopes, the surface micro-roughness can be expressed as a two-dimensional Fourier series. The bases of the series have a grating shape. The surface

scattering is the sum of all these grating effects. If the grating separation is d, the spatial frequency is $1/d$. At normal incidence, the grating equation is:

$$d = \lambda / \sin\theta \tag{9.17}$$

where θ is the scattering angle measured from the surface normal. The range of d can be determined for which the scattering angle is approximately between 1 and 90. When the wavelength is 500 nm of the visible light, this range of d is between 0.5 and 28 μm. At a grazing angle, the above equation is replaced by:

$$d = \lambda / \varepsilon \tag{9.18}$$

where ε is a small scattering angle away from the specular reflecting direction. In the X-ray regime, λ is small and the concerned scattering angle is also small. They are within a range between arcs and arcmin. If $\lambda = 1$ nm, the grating effect produced scattering angle is 7 arcsec for $d = 28$ μm and the produced scattering angle is 7 arcmin for $d = 0.5$ μm. Generally, a polished surface has a smooth and continuously changed micro-roughness power spectral density. A polished surface of 1-nm rms micro-roughness or less in the optical regime should also ensure a good X-ray scattering performance at grazing angles from 1 arcsec to 1 degree away from its specular reflecting direction. This relaxes the surface smoothness requirement for the X-ray grazing telescopes. The micro-roughness requirement for multi-layer coating X-ray mirror surfaces is much higher. Table 9.1 gives the range of surface micro-roughness for various optically finished surfaces.

9.2.2 X-Ray Imaging Telescopes

Any detector has a fundamental directional property defined by its mechanical frame. However, this frame-defined directional accuracy is usually poor. To

Table 9.1. Variance of surface roughness of several mirror materials (nm)

Name	Upper limit of surface roughness	Lower limit of surface roughness	Average roughness
Fused quartz	3.0	0.4	1.3
SiC	1.9	0.4	1.0
Cu (diamond turned)	34.0	1.0	4.9
Electroless nickel	4.4	1.1	1.8
Titanium	3.9	1.3	2.7
Copper	6.3	1.3	3.0
Alumium	8.1	1.9	5.3
Invar	7.0	2.0	4.7
Stainless steel	4.7	3.3	4.0

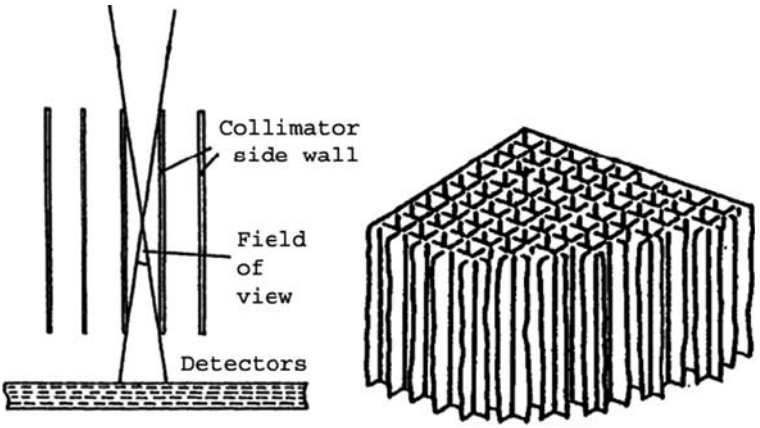


Fig. 9.13. A grid-style collimator.

improve its pointing accuracy, mechanical collimators can be used in front of a detector. A grid style collimator is simple (Figure 9.13). However, this collimator does not retain a focusing effect. A "lobster eye" is a focusing collimator (Figure 9.14). Both are used in X-ray telescopes.

Fixed grid collimators have limited pointing accuracy. To improve their performance, a modulation collimator, which involves several groups of parallel wire grids, can be used [Figure 9.15(a)]. These grids form repeatable patterns in the same pointing direction. The angular resolution and period of these grids are determined by the separations between wires of one grid and the distances between grids. When a new grid of parallel wires is inserted, or removed, the resolution and period in the pointing pattern will change. If observations are made with different grid groupings on the same sky area, the X-ray source positions can be determined through analysis of the periods and resolutions of

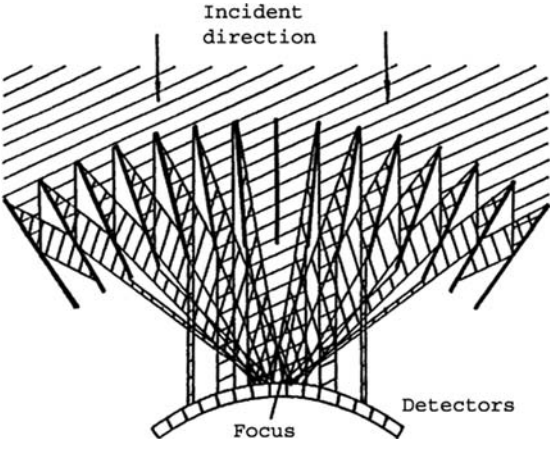


Fig. 9.14. The "lobster eye" focusing collimator.

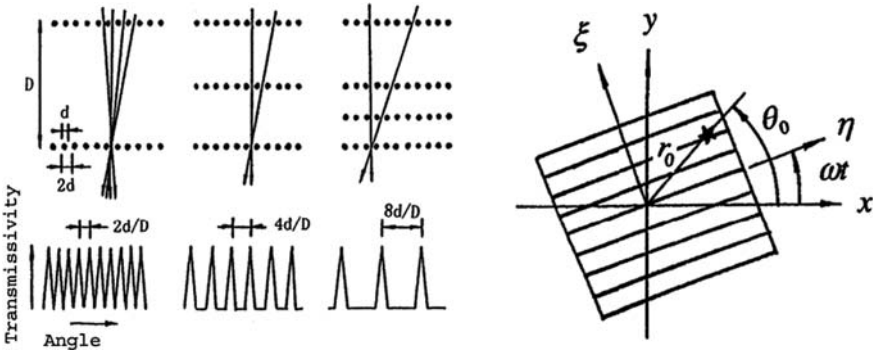


Fig. 9.15. (**a**) Parallel and (**b**) rotational modulation wire gird collimators.

images from different combinations of the grids. The pointing accuracy of this type of collimator is fairly high, reaching 1 arcmin.

An alternative to this type of collimator is called a rotational modulation collimator [Figure 9.15(b)] which is also formed by groups of parallel wire grids. However, these grids can rotate about an axis which is perpendicular to a grid plane, so that the X-ray source image will also be modulated on the detector plane. Demodulating the images can determine the source directions. For a perfect sub-collimator with a pitch of d, the cross section through the modulation pattern is a triangular waveform, which can be represented by a sum of sinusoids:

$$M = a_0 + a_1 \cos(\mathbf{k} \cdot \mathbf{x} - \alpha) + \text{high order terms} \tag{9.19}$$

where $\mathbf{x}$ is the angular distance in the sky from the spin axis of the rotating sub-collimator, and $\mathbf{k}$ a vector whose direction is perpendicular to the slits and whose magnitude is 2π divided by the pitch. The function is always non-negative, since it represents a probability.

During the rotation of a sub-collimator, the X-ray emitted by a point source produce a modulation profile, which may be computed from Equation (9.19) by choosing a source position $\mathbf{x} = (\mathbf{x_0}, \mathbf{y_0})$, and letting the wave vector $\mathbf{k}$ rotate through a range of orientation angles φ. The collimator phase α, which represents the offset of the modulation pattern from the spin axis, is a function of the orientation angle which is known in advance. The flux of X-rays will be approximately proportional to:

$$F = a_0 + a_1 \cos[(2\pi/d)(x_0 \cos\phi + y_0 \sin\phi) + \alpha] \tag{9.20}$$

A number of different collimators have been used in X-ray and gamma ray observations. However, very high resolution and sensitivity are difficult to obtain by using these simple collimators. At the same time, the bombing from

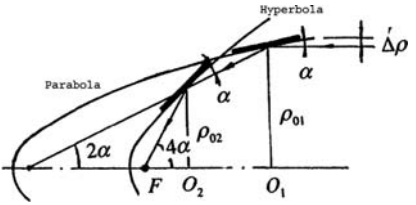


Fig. 9.16. A basic grazing incidence telescope.

cosmic rays also adds serious background noise on the detector plane, increasing the difficulties in the data processing.

Real two-dimensional X-ray images are obtained through the use of grazing incidence telescopes, which were first proposed by Hans Wolter, known as the Wolter-type telescope or Wolter optics (Cheng, 1988). The basic Wolter-type optics (Wolter I type) is shown in Figure 9.16. It consists of two axially symmetric ring reflecting surfaces. The X-ray photons strike the surfaces at a very small grazing angle to avoid material absorption and scattering. The first surface is a portion of a paraboloid and the second a portion of a hyperboloid. As in a Cassegrain system, the focus of the paraboloid overlaps the virtual focus of the hyperboloid and the final image is formed at the real focus of the hyperboloid mirror.

Four basic parameters are needed to determine a grazing incidence telescope. These are: (a) the grazing angle at the center of the entrance ring, α ; (b) the radius at the center of the entrance ring, ρ_{01}; (c) the distance between the centers of two surfaces, d; and (d) the half width of the entrance ring, $\Delta\rho$. The equation for the reflecting surface of this telescope is:

$$\rho^2 - \rho_0^2 = 2kz - (1-\delta)z^2 \quad (\rho^2 = x^2 + y^2) \tag{9.21}$$

Other parameters of the telescope, where b is the distance between the secondary mirror and the system focus, are defined by the following equations:

$$\rho_{02} = \rho_{01} - d\tan 2\alpha$$

$$k_1 = -\rho_{01}\tan\alpha \qquad k_2 = -\rho_{02}\tan 3\alpha$$

$$\delta_1 = -1 \qquad \delta_2 = -\left[\frac{\sin 2\alpha}{\sin 4\alpha - \sin 2\alpha}\right]^2$$

$$b = \rho_{02}/\tan 4\alpha \qquad f = \rho_{01}\sin 4\alpha$$

There are a number of other types of grazing incidence telescopes. As shown in Figure 9.17, in one system the secondary mirror is a ring of convex hyperboloid with reflection at its outer surface. The other system has two orthogonally placed cylindrical reflecting surfaces. The efficiency of this latter system is high but the resolution is limited. In some grazing telescopes, three or more mirrors are involved. Two three-mirror grazing telescopes are shown in Figure 9.18. In the first system, the first two reflectors form the basic Wolter I optics and

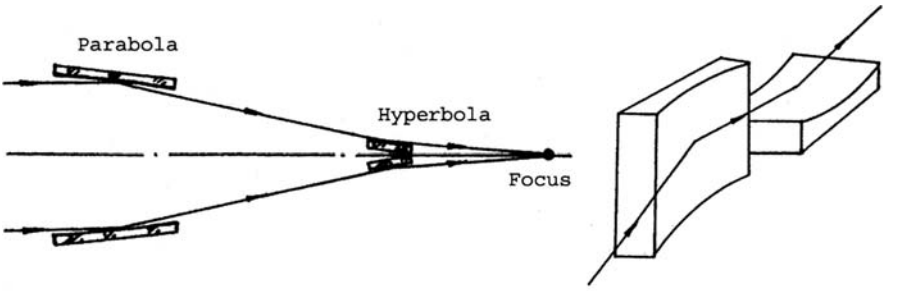


Fig. 9.17. Other types of grazing telescopes.

the secondary and third form another Wolter optics. This three-mirror grazing telescope has larger off-axis aberrations; however, the image quality can be improved by adjusting the first two reflector shapes. The second system is a special grazing telescope with its first subsystem being a grazing telescope and the second a grazing microscope. This telescope has a higher resolution but a smaller field of view.

A single grazing telescope has limited light collecting area. Therefore, a nested array of grazing telescopes is often used (Figure 9.19). The total light collecting area of this system is the sum of each individual telescope.

The mirror of a grazing telescope can be made either by diamond turning of metal material or through optical polishing. However, more and more grazing mirrors are made through a replication process using electroformed surfaces. These surfaces are bonded to CFRP ring-type bodies. Using replication, the surface micro-roughness is usually improved. A new design of the nested grazing telescope is multi-nested thin foil optics, which involves many rings of densely packed small conical reflectors separated by radial comb-like alignment bars.

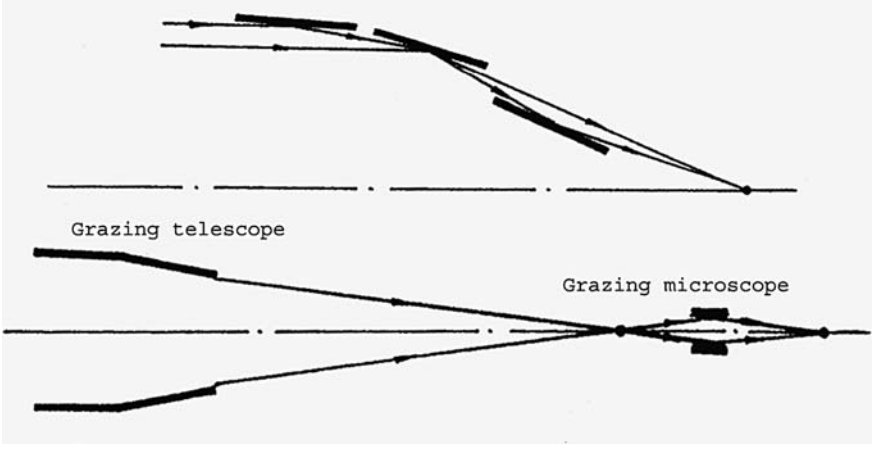


Fig. 9.18. Two multi-mirror grazing telescopes.

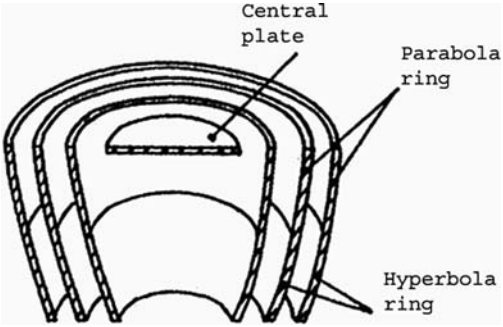


Fig. 9.19. Nested mirror grazing telescope.

These thin glass shells or aluminum alloy foils are made by slumping thin glass of about 0.4 mm in thickness on a precise mandrel at ∼600°C or pushing a stack of thin aluminum alloy small foils against a conical mandrel at ∼200°C for a period of time. Once the glass mirror cools, a layer of iridium is sputtered on its inner surface. These small reflector sections are then radial supported by comb-like alignment bars to form multi-ring densely nested grazing telescopes as used in the Constellation X SXT and the New Exploration X-ray Telescope (NEXT) projects.

In X-ray optics, the wavelength is small. The required optical path length error is small. However, the path length error is the product of the surface error and the sine of the grazing angle. Therefore, the surface tolerance is not too tight.

Grazing telescopes are not used for photon energy above ∼100 keV in gamma ray and hard X-ray regimes. In these regimes, absorption becomes so serious that only coded mask detectors are used. The coded mask detectors are discussed in the gamma-ray telescope sections.

9.2.3 Space X-ray Telescopes

The first X-ray astronomical observation was performed in 1949. Afterwards, the number of X-ray observations increased very rapidly. Early X-ray observations were made from rockets. The detectors were pieces of photographic films shielded from visible and ultraviolet light by beryllium or aluminum foils. By varying the thickness of the foils it was possible to detect X-rays of different energies.

In 1970, NASA sent from Kenya two X-ray detectors into orbit. This was the Small Astronomical Satellite-A (SAS-A, Uhuru, or Explorer 42). The instruments used were two argon-filled proportional counters sensitive to X-rays in the range from 2 to 20 keV. Mechanical collimators were also used in this instrument for improving the angular resolution. Three years of SAS-A operation produced a large amount observation data.

After this satellite, a number of other X-ray satellites were sent into space: including ANS of the Netherlands, Ariel-5 of the UK/USA, the Aryabhata satellite of India/USSR, SAS-3 satellite, and Orbit Solar Observatory (OSO-8). In 1977, the High Energy Astronomical Observatory (HEAO-1) was sent into orbit. This was a large satellite with a weight of 175 kg. It used a collimator for improving the angular resolution, but it was not an imaging telescope yet.

Grazing telescopes started from 1963. In 1970, images of the sun were obtained with a small grazing telescope. In 1978, the HEAO-2 satellite, also named the Einstein Observatory, was sent into orbit. The Einstein Observatory consisted of a nested grazing X-ray telescope with four groups of mirrors. The diameter of the mirrors was about 0.5 m and the telescope was used for X-rays with energy lower than 3.5 keV. The effective area was 300 cm^2 for X-rays with energy lower than 0.75 keV, 200 cm^2 for energy lower than 2.5 keV, and 50 cm^2 for energy lower than 3.5 keV. The mirrors were made of glass material, coated with chromium and nickel to obtain the required reflectivity. This satellite ceased operation in March 1983.

After the Einstein Observatory, another X-ray telescope, the European X-ray Observatory SATellite (EXOSAT), was launched by the ESA in 1983. EXOSAT consisted of two high-resolution grazing telescopes with a diameter of 52 cm. Its effective mirror area was 90 cm^2. Each telescope had three mirror elements with a focal length of 1,090 mm and a shortest working wavelength of 0.83 nm. The HPBW of the point image blur was 10 arcsec. The scattering loss was less than 6%. The mirror surface was made of beryllium material using an epoxy replica technique and was coated with gold. The weight of each telescope was 7 kg. EXOSAT developed a high-precision surface replicating technique and a number of testing devices for grazing optics. EXOSAT ceased operating in May 1986.

After EXOSAT, X-ray projects include the X-Ray Telescope (XRT) of the UK, the Roentgen X-ray telescope (ROSAT) on the Kvant-1 module of Mir space station of USSR, and Astro-C of Japan. An important one among these is the Rontgenstrahlen Satellit (ROSAT) of Germany, the UK, and US. This was an 80-cm grazing telescope with four nested mirror groups and a focal length of 2.25 m. The telescope was sent into orbit in 1990 and worked until 1999. ROSAT was a very successful large telescope with a weight of 2.5 tons and a dimension of 2.4 m × 4.7 m × 8.9 m. ROSAT was able to sweep X-ray sources a hundred times fainter than those located by the Einstein observatory. It obtained information of more than 100,000 X-ray sources.

After ROSAT, the X-ray projects include Astro-D of Japan, the Rossi X-ray Time Explorer (RXTE), and the X-ray Astronomical Satellite (SAX) of Italy and the Netherlands. In 1999, the US sent into orbit the Chandra X-ray Observatory (Figure 9.20). Its other name was Advanced X-ray Astrophysical Facility (AXAF). Chandra had a 1.2 m grazing telescope of six nested mirror groups with an effective collecting area of 1,100 cm^2, a focal length of 10 m, and a field of view of 30 arcmin. Its X-ray observing range was from 0.1 to 10 KeV and the resolution is 0.5 arcsec – as good or better than a modern large ground-based

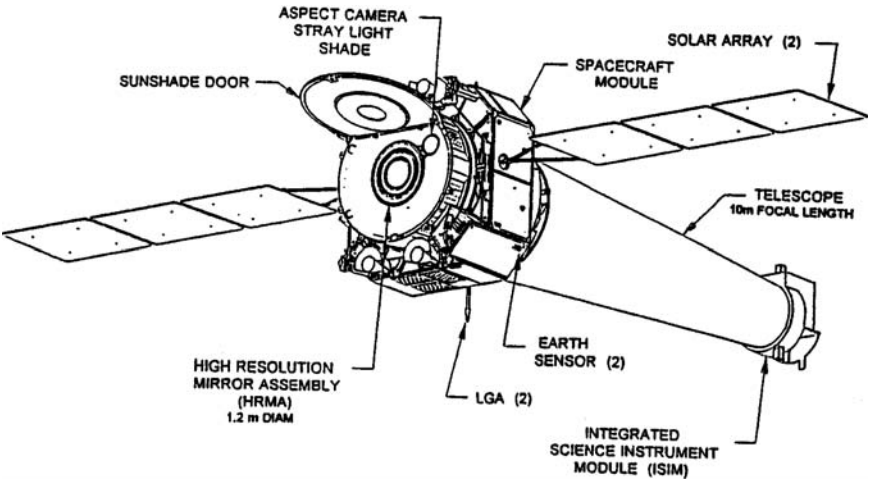


Fig. 9.20. Chandra X-ray Observatory.

optical telescope. The mirror was coated with iridium. It could observe continuously on a highly eccentric ellipse orbit with a period of 64 h.

In 1999, an X-ray Multi-mirror Mission (XMM-Newton) telescope was also sent into orbit by ESA. XMM consists of three co-aligned grazing telescopes. Each of these telescopes had a collecting area of 1,500 cm^2. Reflecting gratings were installed on two of these telescopes. The spectral resolution could be adjusted to $E/dE \sim 20$–50 or $E/dE \sim 200$–800.

In 2000, Astro-E featuring a cryogenically cooled "X-Ray Spectrometer" built by Japan and the US was launched. The XRS provided about five-times greater spectroscopic resolution than the Chandra, though only over the high end of Chandra's spectral range. The XRS also had a wider field of view than spectroscopes on other satellites. The XRS ran out of helium coolant after two and a half years, but the payload also included a low-energy "X Ray Imaging Spectrometer (XRIS)" and a "Hard X-Ray Detector (HXD)" that could continue to operate.

The planed space X-ray telescopes include NEXT of Japan and International X-ray Observatory (IXO) of ESA, NASA, and JAXA (Japan Aerospace Exploration Agency). The IXO had an early name of Constellation X. These new projects will use multi-nested thin foil grazing telescopes and will be placed into orbit after 2010.

9.2.4 Microarcsecond X-ray Image Mission

The MicroArcsecond X-ray Image Mission (MAXIM) will be the next step of the X-ray space mission. As an interferometer, the resolution can be 100 μas

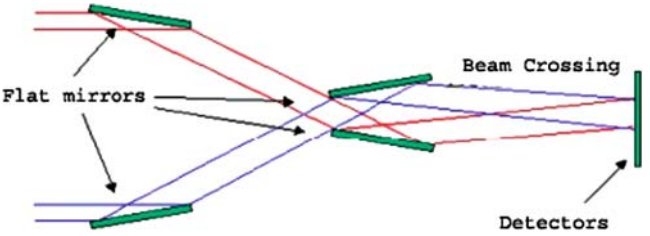


Fig. 9.21. Schematic of the MAXIM interferometer (NASA).

(microarcseconds), a factor of 1,000 higher than that of the HST and 10^6 times higher than that of the Chandra X-ray Observatory, even for a short baseline of 1 m.

To avoid the difficulties of building diffraction limited X-ray optics, a special flat mirror 'x' configuration interferometer is being studied for this project where the basic mirror arrangement is shown in Figure 9.21. This is a quasi-Fizeau interferometer (Section 4.2.3) where the beam from each mirror is combined in the detector plane and fringes are formed. The formed fringe spacing s is proportional to the wavelength λ and the focal length L, and it is also inversely proportional to the beam separation d (Figure 9.22). For a plane mirror surface, a small angle θ exists between wavefronts of two incoming beams so that two flat mirrors will form fringes of parallel straight lines with a sine wave form as in the Newton ring interferometer (Figure 9.23). In this interferometer, all flat mirrors are arranged in a ring around the axis. Four mirrors along two perpendicular lines will produce a checkerboard-type dot pattern. As the number of mirrors increases, the pattern will become complex first, then it starts to clear out the regime around the center point. If there are 32 mirrors in the ring, the interference pattern will be similar to the diffraction pattern

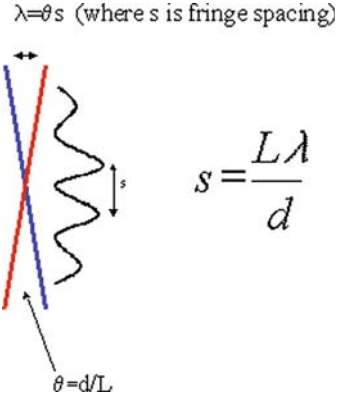


Fig. 9.22. Two beams are crossing at a small angle and form a wide fringe (NASA).

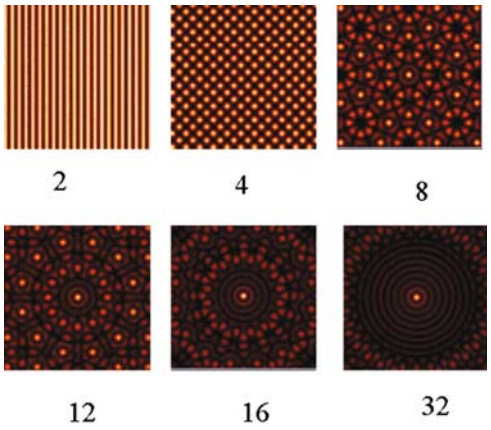


Fig. 9.23. Interferometer patterns formed by a ring of flat mirrors (NASA).

resulting from a circular aperture, being a diffraction limited one. However, this telescope has its resolution near to $\lambda/2d$, not the number of λ/d for a simple filled aperture telescope, where d is the ring diameter formed by all the flat mirrors. For this special X-ray interferometer, if two sources are $\lambda/2d$ apart, their images can be well resolved.

A pathfinder mission for this project will include two important spacecraft: one for the optical system and the other for the detectors. The optics spacecraft has two rings of 64 flat mirrors, with 32 mirrors in each ring, forming the 'x'-type arrangement. All mirrors are finely adjusted to have exact zero null on the axis of the image plane while they are 450 km away from the detector spacecraft. Each mirror is about 3 cm in width and 90 cm in length. The beam size formed is about 3 cm. The diameter of the first ring of mirrors is 1.4 m and that for the second ring is 0.3 m. The separation between the two mirror rings is 10 m and the crossing angle is about 2 arcsec. The fringe formed is only 100 μm for the 1-nm wavelength X-ray.

A laser metrology system will be used for mirror fine position adjustment of this pathfinder mission. The required accuracy is between 1 and 10 nm for an exceptionally stable system pointing performance. The achieved pointing stability is about 300 μas in the laboratory, with expected reconstruction in orbit to 30 μas. The target acquisition of the system takes place in X-ray band, but the maintenance of the pointing of the main spacecraft cannot be done in X-ray band. The pointing information is generated by two visible interferometers based on the design used on SIM demonstrators (Section 5.3.3). Two visible interferometers lie approximately perpendicular to each other, one for the pitch and one for the yaw alignments. The SIM has a target resolution of 4 μas so that the Pathfinder requirements are directly achievable using the same approach. To stabilize the spacecraft, laser propulsion may have to be used. The laser thrust can reach thrust forces of a few micro-Newton.

In the detector spacecraft, a quantum calorimeter array will provide the necessary spectrum resolution of $E/\Delta E = 100$–1,000. The detector will be about 30 mm^2 with 300 μm or smaller pixel size. Energy resolution of 2 eV at 1 keV will allow the fringe coherence to be maintained. The interferometer has a wide field of view, so that an array of 3 cm CCDs will surround the quantum calorimeter to increase the field of view for the observation of extended objects and to center on poorly known target positions.

However, the MAXIM is a very challenging project and its final design concept is still under development.

9.2.5 Space Ultraviolet Telescopes

The ultraviolet regime ranges from 10–91 to 390–400 nm in wavelength. The range between 310 and 390–400 nm is reachable by ground ultraviolet telescopes. However, observations in the rest wavelength range have to be carried out at high altitude, or in space.

The earliest space ultraviolet observation was with a simple photo-electronic detector in 1964. The first space ultraviolet telescope was the Orbit Astronomy Observatory (OAO-1) launched in 1966 for the analysis of stellar energy distribution in the ultraviolet regime. However, the OAO-1 did not send back any observational data.

The OAO-2 launched in 1968 consisted of a 20-cm survey telescope in the ultraviolet regime. The OAO-3 (or the Copernicus) launched in 1972 was an 80-cm Cassegrain telescope which lasted for three years. The USSR also had a few space ultraviolet satellites. The European and France ultraviolet satellites included the TD-1A and D2B-AURA.

An important US/Europe ultraviolet project, the International Ultraviolet Explorer (IUE), which worked at the wavelength of 115∼350 nm, was launched in 1979 (Figure 9.24). Its length was 4.2 m and its weight was 700 kg. Other ultraviolet satellites included the USSR/France 0.8 m ASTRON space telescope for 150 to 350 nm wavelengths in 1983 and the 0.5 m Astro-1 in 1990 and the 0.38 m Astro-2 in 1995. Both Astro-1 and -2 are space shuttle based ultraviolet telescopes.

In EUV wavebands, grazing telescopes are necessary. The successful Apollo-Soyuz Test Project (ASTP) in 1979 included a 37-cm EUV grazing telescope made with four nested mirrors (Figure 9.25). Two similar grazing telescopes were the wide field camera of the UK ROSAT satellite in 1990 and the US Extreme Ultraviolet Explorer (EUVE) in 1992. The EUVE was a 40-cm diameter gold-coated grazing telescope. The Orbiting Retrievable Far and Extreme Ultraviolet Spectrometer (ORFEUS) was a conventional 1 m telescope which was launched twice by the space shuttle in 1993 and in 1996.

The Far Ultraviolet Spectroscopic Explorer (FUSE) launched in June 1999 worked until July 2007 when its reaction wheel seized. The FUSE telescope also

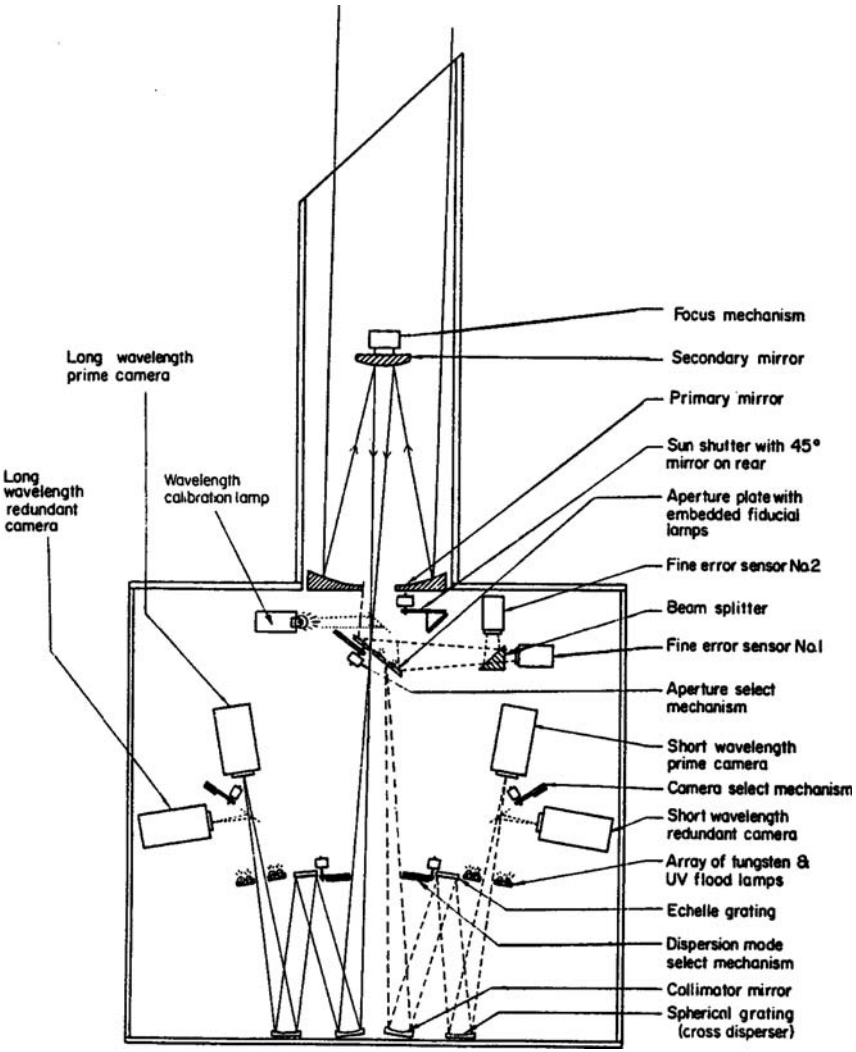


Fig. 9.24. The telescope and instruments of IUE (PPARC).

included four groups of mirrors. These mirrors of a size of 39 cm by 35 cm in an off-axis parabola shape were coated both with silicon carbide for reflectivity at the shortest ultraviolet wavelengths and with lithium fluoride over aluminum for reflectivity at longer wavelengths.

The NASA, South Korea, and France Galaxy Evolution Explorer (GALEX) survey telescope was launched in 2003. It was a 50-cm Richey–Chretien ultraviolet telescope with a field of view of 1.2 degrees working at wavelengths between 135 and 280 nm. The detectors were near-ultraviolet and far-ultraviolet micro-channel image intensifier ones. The mission lasted 29 months.

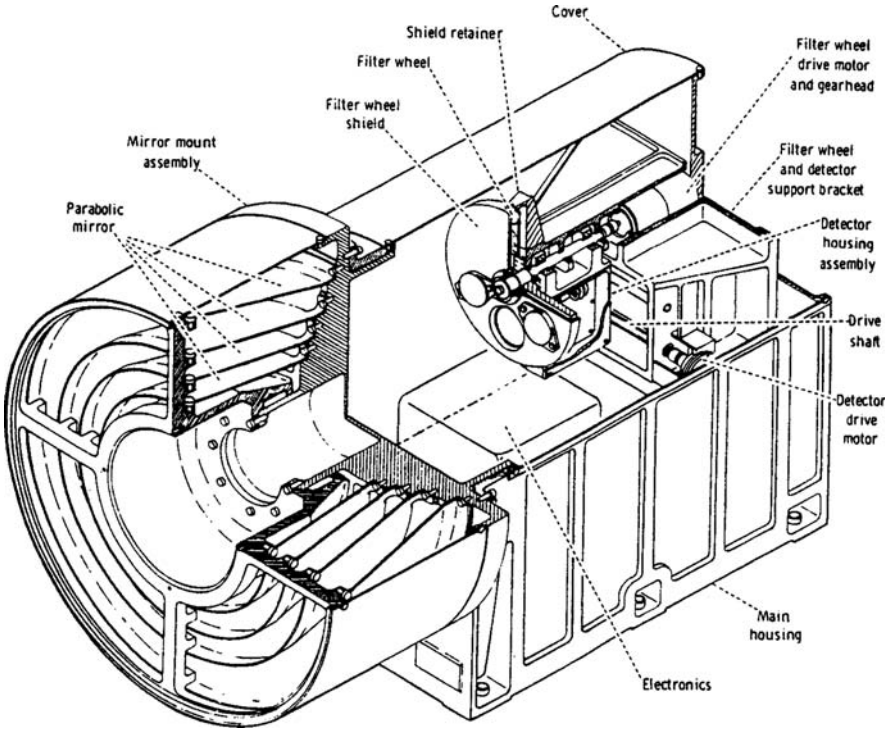


Fig. 9.25. The ASTP EUV telescope (NASA).

9.3 Gamma Ray Telescopes

9.3.1 Gamma Ray Fundamentals

Gamma rays are high-energy electromagnetic radiations with extremely short wavelengths. Normally, the wavelength is smaller than 0.001 nm, so that the photon energy is higher than 1.2 MeV. However, there is no clear division between X-rays and gamma rays. In some references, radiation with wavelengths between 0.01 and 0.001 nm is also called soft gamma rays. Therefore, the energy of soft gamma rays may be smaller than 1.2 MeV.

The generation of gamma rays is directly related to interaction of high energy particles, including cosmic rays and high-energy gamma rays, with interstellar particles or atmospheric molecules. If the collision occurs in the atmosphere, the gamma rays and particles generated from this collision are usually called secondary gamma rays and secondary particles. The gamma ray generation is also related to the matter–anti-matter annihilation process, radioactive decay, and the acceleration of particles. The collision of electrons and positrons produces a pair of gamma rays. Likewise, colliding gamma ray photons can produce electron and positron pairs. Therefore, gamma rays hold the key to

unlocking the mystery between matter and energy and between matter and antimatter.

The most distinct feature of gamma rays is their high energy and lower absorption in the universe. For gamma rays with energies between 10 MeV and 10^9 MeV, the universe is almost transparent when the high energy gamma rays pass through the whole disk of our galaxy. Only 1% of the photons interact with the galaxy materials.

In the radiation generation process, the number of photons generated decreases greatly as the photon energy increases; therefore the flux density of gamma rays is less than that of X-rays. When the photon energy is larger than 50 MeV, the flux density of gamma rays is only $4 \times 10^{-14}/\text{cm}^2\,\text{s/rad}^2$. This small flux density makes rocket observation of gamma rays quite difficult. The secondary gamma ray and cosmic ray effects from the atmosphere and the dark current of detectors make balloon-borne observations also difficult. Gamma rays with energy below 10 GeV generally cannot penetrate the atmosphere. Therefore, direct gamma ray observations have to be performed on mountain tops, in high-flying balloons, or in orbit. Indirect gamma ray observations, which detect the secondary particles (effects) from their interaction with the atmosphere, water, ice, or other molecules, can be performed on ground or under water. One of these effects is the Cherenkov (also known as Cerenkov) effect which is the cone-shaped bluish flashes generated from high-speed secondary particles. The Cherenkov effect and Cherenkov telescopes will be discussed in Section 9.3.3.

9.3.2 Gamma Ray Coded Mask Telescopes

In the gamma ray regime even grazing telescopes are not an option because of the high photon energy level. Up to now, the only method for improving the pointing accuracy in gamma ray observations is through coded mask type collimators. In an extreme case, the collimator mask is just a coded pinhole pattern. If a telescope has only one pinhole, it is the same as a pin-hole camera as used in the optical regime. The collecting area of one pin hole is too small so that a group of coded pinholes is used. The telescopes of this type are named coded mask imagers.

When gamma rays come from one direction, a pattern is produced in the detector surface through the coded mask. If the direction changes, then the pattern changes. All coded mask imagers have their own point spread functions. The image generated from an imager is a convolution or correlation of the mask's function and the object distribution (Fenimore and Cannon 1978):

$$P = (O * A) + N \tag{9.22}$$

where P is the recorded image, A the mask's function, N the noise, O the object, and the star $*$ means the correlation. After deconvolution, the object is solved for by:

$$\widehat{O} = R\bar{F}^{-1}[\bar{F}(P)/\bar{F}(A)] = O + R\bar{F}^{-1}[\bar{F}(N)/\bar{F}(A)] \tag{9.23}$$

where $\bar{F}$, $\bar{F}^{-1}$, and R are the Fourier transform, the inverse Fourier transform, and the reflection operator, respectively. A is usually defined as an array of ones and zeros where the ones have the same pattern in the array as do the pinholes in the aperture surface. Therefore, the Fourier transform of A may have small terms. These small terms can cause the noise to dominate the reconstruction process.

To simplify the situation, a correlation method may be used. The reconstruction of the object is defined as:

$$\hat{O} = P * G = O * (A * G) + N * G \tag{9.24}$$

where G is called the post-processing array which is chosen such that A *G approximates a delta function. The delta function represents an infinitely sharp peak: a function that has the value zero everywhere except at x = 0 where its value is a finite number. G usually has ones in the same location as the A array and has minus ones when zeros appears in the A array. In this case, the reconstructed object becomes:

$$\hat{O} = O + N * G \tag{9.25}$$

Another method is to make the auto-correlation of a coded mask being unity. In this case, the post-processing array is not necessary. The object function, therefore, is a convolution of the mask function and the image function. To achieve an auto-correlation of the mask function being a delta function, ideal patterns can be found based on cyclic difference sets (Gunson and Polychronopulos, 1976). For an integer number n based set, the cyclic difference set (CDS) is a unique sub-set of numbers. Mathematically, a cyclic difference set of modulo n is a set of s positive numbers $\{a_1 = 0, a_2, a_3, \ldots, a_s\}$ less than n, with the property that all differences a_i-a_j of modulo n for i not equal to j are different. The number n is called the modulus of CDS. By assigning transparent elements to these cyclic difference set numbers on the mask surface, the required coded mask function can be formed (In't Zand, 1992).

To apply cyclic difference sets of base number N for constructing a coded mask, one first visualizes a regular pattern of N equally spaced locations in the mask surface and then puts the number for these locations from 1 to N moving diagonally down through the grid from the upper left corner to bottom right and wrapping back up to the top of the next column when the bottom is reached. For example, given a base 15 set, we get a three row pattern of numbers as follows:

[1]	[7]	[13]	[4]	[10]
[11]	[2]	[8]	[14]	[5]
[6]	[12]	[3]	[9]	[15]

If one cyclic difference set for a base of 15 has the numbers: 1, 2, 3, 5, 6, 9, and 11, then a coded mask pattern is to assign transparent elements on these numbers' position. Higher pinhole area percentage is a consideration for the instrument sensitivity. In practice, the coded mask pattern is now generated by computers.

In the sky, a gamma ray burst is a rare phenomenon. To catch this type of rare event with a low signal to background ratio, a wide angle camera is necessary. One wide angle camera has a dome-like shape. There are three groups of patterned pinholes on the dome surface and three groups of detectors in three different orientations are located behind this mask. The camera has a field of view of $3\pi \cdot \text{rad}^2$. The pointing accuracy of this type of imager can reach 1 arcmin.

A coded mask type telescope has three parameters: resolving angle, beam width, and period of image pattern. To increase the resolution of these gamma ray telescopes, one can increase the distance between the mask and the detector. The coded mask can have circular, square, or rectangular holes. Square holes have a pencil-shaped beam pattern. Rectangular holes have a fan-shaped beam pattern. The lower energy limit for using this type of imager is determined by the diffraction of the gamma ray photons and the upper energy limit is determined by the transparency of the mask frame. In general, these imagers are only used in the relatively low energy (LE) X-ray/gamma ray regime (0.1–0.5 MeV) or in the so-called "hard X-ray" regime (10–150 keV).

The detectors of the gamma ray telescopes are important. The detector is located behind the collimator mask. In the high energy regime, the collimator mask becomes transparent, so the detectors are placed on or behind a type of tracker. The tracker is a type of positioning device which mixes with the detector for tracking and reconstruction of single gamma rays (multiplicity = 1) in the detection. The requirements for trackers include positional accuracy, multiple event identification, and energy resolutions. The tracker, usually as a passive device, is also related to detection threshold.

Existing gamma ray detectors include scintillators, Compton scattering detectors, and pair production detectors. The Compton scattering and pair production telescopes are discussed in the next section. The scintillators are made of scintillate crystals of an alkali halide salt, such as sodium iodide (NaI) or cesium iodide (CsI) doped with an activator, or semi-conductor materials such as germanium or cadmium-zinc-telluride (CdZnTe, or CZT). The material emits low-energy (usually in the visible range) photons when struck by a high-energy charged particle. When used as a gamma ray detector, the scintillator does not directly detect the gamma rays. Instead, the gamma rays produce charged particles in the scintillator which interact with the material and emit photons. These lower energy photons are subsequently collected by photomultiplier tubes (PMTs).

The combination of the coded mask collimator and the CZT detector forms the state-of-the-art gamma ray instruments. They have been used in the SWIFT

gamma ray space project, serving as a Burst Alert Telescope (BAT) for rarely occurring gamma ray bursts.

9.3.3 Compton Scattering and Pair Telescopes

In the medium energy and high energy regimes, Compton scattering and pair production telescopes are used for direct gamma ray observation. The Compton scattering effect mostly occurs in the gamma-ray medium energy (ME) regime (0.5–30 MeV) and the pair production effect mostly in the high energy (HE) regime (30 MeV–10 GeV) and beyond (40 GeV). By using these effects together with well-designed trackers, both the direction and energy of incoming gamma ray photons can be determined.

Compton scattering or Compton effect is the decrease in energy of an X-ray or gamma ray photon when it interacts with materials. The interaction of a gamma ray with an electron transfers part of its energy to the charged particle, making it recoil. The photon with remaining energy emits in a different direction from the original one so that the overall momentum of the system is conserved. Compton scattering telescopes are based on this effect as shown in Figure 9.26.

A Compton scattering telescope is typically a two-level instrument. The first level is a converter and the second an absorber. Both levels are formed by scintillate materials (S_1 and S_2). Scintillators emit light when traversed by energetic particles. Many materials radiate, but most also absorb that radiation so that the light may never get out in some materials. Scintillators are widely used in other particle detectors.

On top of the converter and absorber layers, there are shielding layers A_1 and A_2 for blocking low energy photons. When a gamma ray photon γ_0 with energy $E_{\gamma 0}$ hits the top level S_1, it scatters off an electron e_1 in the scintillator. At same time, the gamma ray itself will change direction and becomes a new photon γ_1 with reduced photon energy. The angle between the incoming and the new photon direction is θ. The energy of the new photon is $E_{\gamma 1}$. The new photon

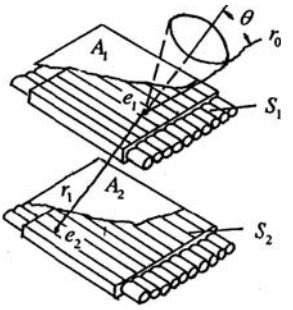


Fig. 9.26. Compton scattering telescope.

travels down into the second level of the scintillator where it produces another electron e_2 and another photon γ_2. The process can be expressed as:

$$\begin{aligned} E_{\gamma 0} &= E_{\gamma 1} + E_{e1} \\ E_{\gamma 1} &= E_{\gamma 2} + E_{e2} \\ \cos\theta &= 1 + mc^2\left(\frac{1}{E_\gamma} - \frac{1}{E_{\gamma 1}}\right) \end{aligned} \tag{9.26}$$

where $mc^2 = 0.51$ MeV with m the mass of the electron. If the second level material completely absorbs the scattered photon, $E_{\gamma 2} = 0$, $E_{\gamma 1} = E_{e2}$, then:

$$\cos\theta = 1 + mc^2\left(\frac{1}{E_{e1} + E_{e2}} - \frac{1}{E_{e2}}\right) \tag{9.27}$$

During the scattering process, phototubes or other devices which view the two levels can determine approximately the interaction points in the two layers and the amount of energy deposited in each layer. Therefore, the angle θ can be estimated. Unfortunately, while the angle can be estimated, the azimuth direction from where the photon came cannot be determined. The gamma ray may have come from anywhere in a ring direction of the sky, which makes analyzing Compton scattering telescope data particularly challenging. Compton scattering telescopes have relatively small effective areas since only a small number of incident gamma rays actually produce the Compton effect on the top converter level. The effective area of a Compton scattering telescope is the geometrical area of the detectors weighed by the probability of interaction.

A pair production or pair gamma ray telescope is shown in Figure 9.27. The name is from the pair production effect in the HE gamma ray regime (>30 MeV), where pair production is dominant for most materials. Pair production is a process in which a gamma ray of sufficient energy is converted into an electron and a positron. In the pair production process, a third body is required for system momentum conservation. The standard design of this type of telescope is a layered structure with converter layers interwoven with tracking materials (tracker). The converter layer is typically a high atomic number material (e.g. heavy metal such as lead). The converter layer provides the target for creating the initial electron/positron pair, while the tracking material detects the pair.

One type of tracking system is a spark chamber which is a gas-filled regime crisscrossed with wires. Once the pair has been created in one of the converter layers, they traverse the chamber and ionize the gas. The charged particle drifts from tracker to wire. Triggering the detector causes the wires to be electrified, attracting free electrons which provide a detectable signal. The trail of the signal

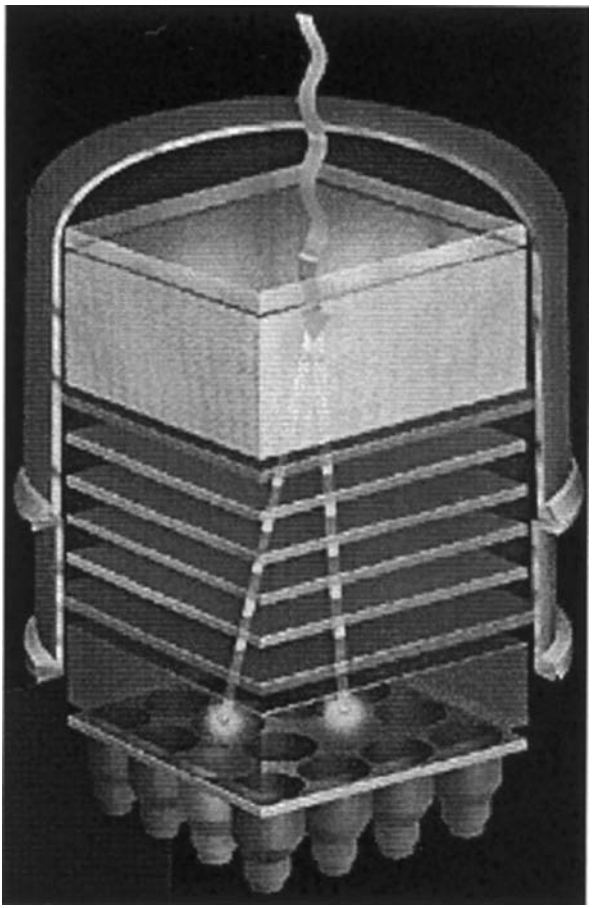


Fig. 9.27. Pair production telescope on the Compton gamma ray telescope.

provides a three-dimensional picture of the pair's path. Another type of tracking material is silicon strips. In one plane, the silicon strips are in the x-direction, and, in another, they are in the y-direction. The position of a particle can be determined more precisely than in a spark wire gas chamber. The incoming gamma ray direction can be calculated from the tracks of the charged pair. In addition, through absorption of the pair by a scintillator detector or a calorimeter at the bottom of the telescope, the total energy of the gamma ray detected can be determined. Calorimeters are also used in other high energy particle and dark matter detection. They measure very small temperature changes by estimating the energy absorbed in a system. The calorimeter can be a scintillator block detector or a liquid (such as argon) detector which stops the particle so that the change of energy detected provides a measure of total energy of the particle.

9.3.4 Space Gamma Ray Telescopes

Except very few located on high mountains, most direct gamma ray telescopes are located in space orbit. The earliest space gamma ray telescope was the third Orbit Solar Observatory (OSO-3) which detected 50-MeV gamma rays from the galaxy disk direction in March 1967. Afterwards, major gamma ray satellites included the fifth Orbit Geophysical Observatory (OGO-5) in April 1968, the seventh Orbit Solar Observatory (OSO-7) in September 1971, the second Small Astronomy Satellite (SAS-2) in November 1972, Thor-Delta-1A (TD-1A) in March 1972, Gamma and Granat in December 1989 of the USSR, High Energy Astronomy Observatories (HEAO-1) in August 1977, COS-B in August 1978, HEAO-C in September 1979, and the Compton Gamma Ray Observatory (CGRO) in April 1991. The HEAO-C mainly worked in the 0.6–10 MeV gamma ray regime. The COS-B of Europe was a sensitive sky survey instrument. The CGRO was a large gamma ray facility for study of spectrum line and gamma ray bursts. It consists of four instruments: BATSE works at energies between 0.03 and 1.2 MeV; OSSE at energies between 0.06 and 10 MeV; COMPTEL at energies between 1 and 30 MeV; and EGRET at energies between 20 MeV and 30 GeV.

The High Energy Transient Explorer 2 (HETE-2) of NASA, Japan and France was put into space in 2000 following the failed launch of HETE-1 in 1996. HETE 2's main instrument was a gamma-ray spectrometer array and a wide-field hard X-ray coded mask telescope. The International Gamma-Ray Astrophysics Laboratory (Integral) of ESA was launched by a Russian Proton booster in 2002. The Integral which weighed 4.1 tons was designed both for the X-ray and low-energy gamma ray regimes. It carries a gamma ray spectrometer and a gamma ray and a X-ray imager both being coded mask telescopes. The Integral is still in operation. The Swift Gamma Ray Burst Explorer is a three telescope space observatory including BAT in gamma ray, XRT in X-ray, and UVOT in ultraviolet bands. The Swift was launched in November 2004 and is still in operation. The BAT is a coded mask detector made of 5 mm lead plate. The Gamma-Ray Large Area Space Telescope (GLAST), also named the Fermi gamma-ray space telescope, is an international and multi-agency mission launched in June 2008. It has detectors for GeV gamma rays.

Generally, space gamma ray telescopes have a smaller photon collecting area and a poor resolution. When the telescope is used for very high-energy gamma ray detections, the chance of detection with such a small area is very limited. Space gamma ray telescopes usually have an energy detection upper limit of 10 GeV. Therefore, ground- or space-based indirect gamma ray telescopes are important for high-energy gamma ray detection. Major ground-based gamma ray telescopes include very few direct gamma ray telescopes on mountain tops and mostly indirect air Cherenkov telescopes and extensive air shower arrays. Some very large area extensive air shower arrays and ground- and space-based fluorescence telescopes are discussed with cosmic ray telescopes in the next chapter.

9.3.5 Air Cherenkov Telescopes

As mentioned early, the number of high energy photons generated in the universe decreases rapidly as the energy increases. To capture very high-energy (VHE) or ultra high-energy (UHE) gamma ray photons, an extremely large detecting area is required. The VHE regime covers the energies between 10 GeV and 100 TeV (1 TeV = 10^{12} eV), also known as TeV gamma ray, and the UHE gamma rays are more energetic than the VHE gamma rays (above 100 TeV). To acquire a very large detecting area, ground-based indirect observation of the gamma ray is necessary. In space, an area larger than 1 m^2 will be very expensive; however, one can achieve a collecting area of more than 10,000 m^2 by making an array telescope on the ground. Major ground gamma ray telescopes include the Air Cherenkov Telescopes (ACTs) and Extensive Air Shower (EAS) arrays. These telescopes are all based on the Cherenkov effect. The telescopes do not detect gamma rays directly, but Cherenkov air showers caused by gamma ray interaction. The ACTs are used in the VHE regime and the EAS arrays in the UHE regime.

When gamma ray photons hit atmospheric molecules, they produce electron/positron pairs. These particles then interact, through bremsstrahlung and Compton scattering, losing some of their energy to create secondary gamma ray photons which, in turn, produce more electron/positron pairs as long as the photon energy remains higher than 1.022 MeV. Then they go again through bremsstrahlung, and so on. The result is a cascade of electrons/positrons and photons which travel down through the atmosphere until the particles run out of energy. This phenomenon is called the Cherenkov air shower or Cherenkov effect named after the Russian physicist who made comprehensive studies of this phenomenon.

The created charged particles travel through the atmosphere at a speed faster than the speed of light in the medium, c/n, where n is the index of refraction of the medium. This results in an electromagnetic shock wave along their path. The coherent wavefront of radiation has a similar effect as a sonic boom from a supersonic aircraft. The radiation produces a conic shape of faint, bluish light, known as Cherenkov radiation or fluorescence. The Cherenkov effect occurs not only in the atmosphere, but also inside ice and water. The induced radiation not only occurs in the optical regime, but also occurs in the radio regime. In the radio regime, the effect is called the Askaryan effect. Askaryan is another Russian-born scientist. The Askaryan effect is discussed in Section 10.2.3.

Not only gamma rays but also cosmic rays or neutrinos produce the Cherenkov effect although these rays or particles have distinct origin and characteristics. The effect happens for very high energy (VHE), ultra high energy (UHE) (100 TeV–100 PeV), and extreme high energy (EHE) (100 PeV–100 EeV) (1 EeV=10^{18} eV) cosmic rays and neutrinos. A very high-energy cosmic ray Cherenkov effect produces even more particles including electrons/positrons, gamma rays, pions, nuons, and neutrinos. For this reason, gamma-ray Cherenkov

telescopes have a direct application for cosmic ray detection. Sometimes, gamma rays, cosmic rays, and neutrinos share the same telescope facility for their detection. The difference between gamma ray and cosmic ray detections will be discussed in this section and later in Chapter 10.

The conic angle of the bluish faint light generated in an air shower is named the characteristic angle. It is given by the Mach relation:

$$\cos\theta = \frac{c}{vn} \tag{9.28}$$

where v/c is the particle speed in the medium in a unit of the speed of light in vacuum c, and n the refraction index. The condition, that the Cherenkov effect occurs, is $v/c > 1/n$. The attenuation length of the faint visible light in the atmosphere is about 12 km, in water, is between 2 and 3 m, and inside pure ice, is about 24 m. The attenuation length is the distance where the light intensity falls to $1/e$ of the initial value.

The ACTs are ground-based optical telescopes for observing the faint Cherenkov bluish radiation working only on moonless and cloudless nights. However, only 1 in 1,000 air showers in the atmosphere is initiated by gamma rays and all the others are likely derived from cosmic ray effects. Air showers produced by cosmic rays are called hadronic air showers which include hadrons such as electrons, protons, and pions.

Both gamma ray and cosmic ray produced air showers are presented in Figure 9.28. From the shape, a gamma-ray air shower has less momentum in the lateral directions and it appears more symmetrical and more compact, while a hadronic air shower is asymmetrical and wide spreading in lateral directions. If the direction of an air shower is the same as the axis of an optical reflector, a gamma-ray air shower will produce a small circular image at the center of the field. When the direction of an air shower is parallel to and at a distance (e.g. 120 m) away from the reflector axis, a gamma-ray air shower will produce a small elliptical image. The major axis of the ellipse will point to the center of the field, while the minor axis will be very short. Figure 9.29 shows an image of a gamma-ray produced Cherenkov air shower formed by an optical reflector. For

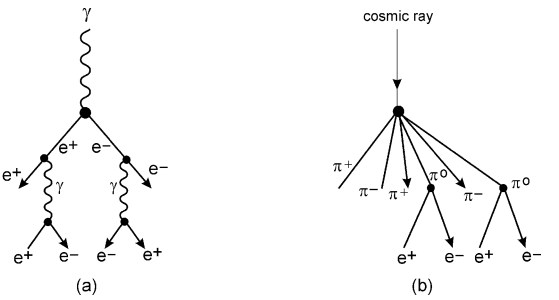


Fig. 9.28. Air showers produced by (**a**) gamma ray and (**b**) cosmic ray.

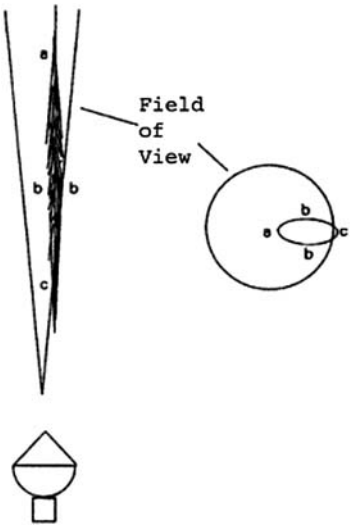


Fig. 9.29. The image technique for the Cherenkov air shower telescope (VERITAS, 1999).

hadronic air showers, the image will be more spreading and the minor axis of the image will be not so short. From the particles generated, air showers initiated by gamma rays contain only electron/positron pairs and gamma rays, while that from cosmic rays contains gamma rays, electron/positron pairs, pi-mesons (pions), nuons, and neutrinos.

In ground-based gamma ray observations, a major task is to distinguish gamma-ray air showers from those with the neutral pions π^0 produced by a cosmic ray. This gamma/hadron separation technique is very important in high energy astrophysics.

The ACT telescope is also known as an imaging ACT (IACT). The air Cherenkov photon number produced by gamma rays or cosmic rays is a function of their energy level (Figure 9.30). Therefore, to detect a relatively low-energy air shower requires larger photon collecting area as well as high detector efficiency. The smaller ACT telescope works for higher particle energy ranges.

The Major Atmospheric Gamma Imaging Cherenkov (MAGIC) telescope array formed from 17 m diameter dishes has an overall area of 236 m^2. The sensitivity can be as high as 0.6–1.1 photons/m^2; the lowest energy detected is about 14 GeV (Magnussen, 1997). The Very Energetic Radiation Image Telescope Array System (VERITAS), a US, UK and Ireland project formed from 10 m diameter dishes, has a sensitivity of 16 photons/m^2. The High Energy Stereoscopic System (HESS), a Germany, France, Namibia, South Africa, and UK project, also formed from 10 m diameter dishes, has the same sensitivity as the VERITAS dishes. Their lowest energy detected is about 100 GeV. Very small dishes detect very high-energy gamma ray events. The collaboration

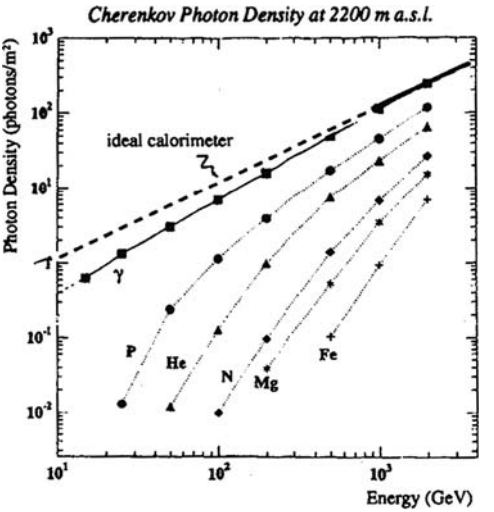


Fig. 9.30. Air Cherenkov photon density at 2,200 m at sea level for gamma ray and cosmic rays of different energy levels (Chantel et al., 1997).

between Australia and Nippon for a Gamma Ray Observatory in the Outback, called the CANGAROO project, involves six telescopes of 3.8 to 10 m diameter, and has the lowest detecting sensitivity.

An ACT unit is an optical reflector and an ACT array involves more reflectors. For improving its detecting sensitivity, a major consideration in the ACT unit design is the aperture size. The requirement for its image size is low. Therefore, its design is in between a light bucket, which is simple, and an optical telescope, which is expensive. For this reason, it uses a special Davies–Cotton optics as shown in Figure 9.31. In this system, the primary reflector formed by

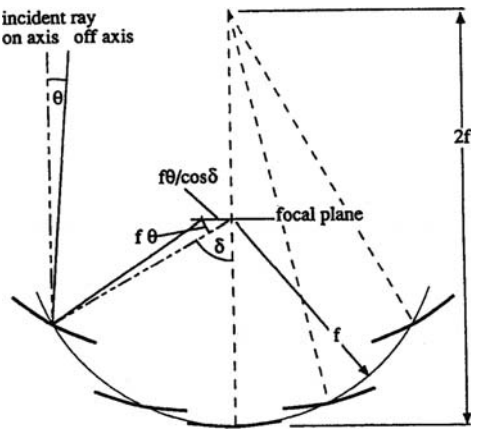


Fig. 9.31. Davies–Cotton optics used for ACT design (VERITAS, 1999).

many small segments is not a continuous surface. As a whole, it has a paraboloidal shape, however, all its segments are spherical ones. The radii of curvature of these segments are twice the paraboloid focal length. The on-axis image is sharp while the off-axis image is poor.

From Figure 9.31, one can find that an angle δ exists between off-axis beams and the focal plane. Therefore, the image size is magnified by a factor of $1/\cos\delta$. The image quality is reduced (Figures 9.32 and 9.33). During mirror adjustment, a retroreflector at the on-axis center of curvature can make the mirror alignment easier. Because the image is poor the requirement for its segment is also low. The general requirement is to achieve 80% of energy within an area of 1 mrad^2, equaling an rms image spread of 0.28 mrad for the telescope.

For this reason, a new segment production technique is to slump a flat circular glass blank on top of a concave spherical mold at high temperature. After the segment is roughly in shape, slight polishing of the surface will be sufficient for the ACT telescopes. Light-weight aluminum honeycomb sandwiched

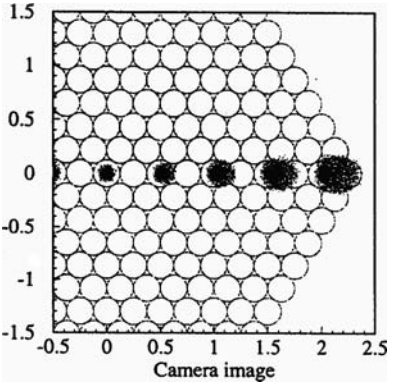


Fig. 9.32. Image spread of Davies–Cotton optics (VERTAS, 1999).

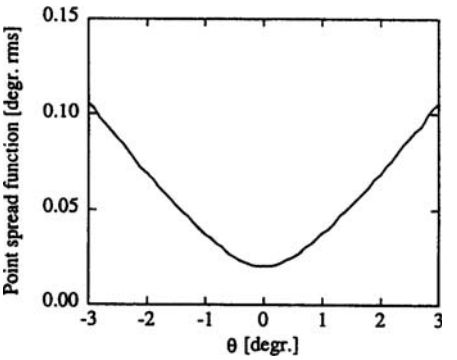


Fig. 9.33. Point spread function of Davies–Cotton optics (VERITAS, 1999).

mirror segments are also used for some ACT arrays. These aluminum surfaces are diamond machined or polished. To prevent dew or ice deposits, heating circuits can be embedded under the top aluminum layer.

The ACT telescope backup structure is similar to the backup structure of a large radio telescope. For reducing the structure weight, CFRP frames can be used in the telescopes. Some modern gamma-ray ACT telescopes have used active mirror adjustment devices.

If a number of images are formed at different locations for a given gamma-ray air shower, then the air shower's precise direction can be determined. Two methods are used for determining the direction of an air shower. One is to combine all individual images by extending the lines of all major axes of them. These major axes of the images define the source direction in space at which the air shower core is located. Usually more than one intersection point exists for all images. Therefore, a best fit method can be used in the data process. Figure 9.34(a) shows the reconstruction of the core location by using this method. Another method is called shower axis projection. Using this method, a shower axis in space is projected into the focal plane of each telescope. The shower direction and the impact point at each telescope are estimated by minimizing the global width of the shower. In this method, all the images are used regardless of their location relative to the shower core [Figure 9.34(b)].

The Davies–Cotton optics has also been used in the solar power industry as a solar power collector. Therefore, an alternative of the ACT telescope is named the solar tower. The telescope formed is called the solar tower telescope. In this design, an individual heliostat mirror serves as a reflector to direct light beams to a secondary mirror on top of a tower. The photomultiplier detectors are at the focal position of the tower. For increasing the efficiency of this type of telescope,

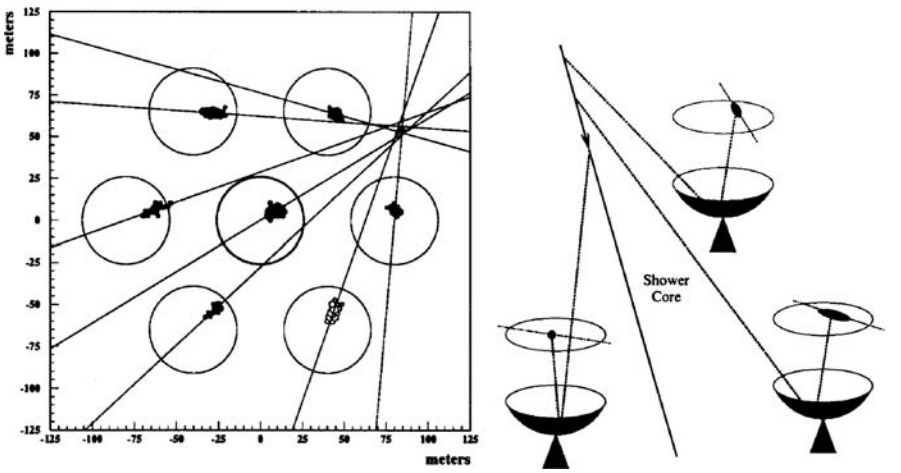


Fig. 9.34. Reconstruction of core location by using (**a**) combination of images and (**b**) using shower axis projection.

the facility can be used as a solar power station in the daytime, while it serves as a gamma ray telescope on cloudless moonless nights. Solar Tower telescopes usually have an even larger photon collecting area and can be used for even lower energy level gamma ray observation.

Major Solar Tower gamma ray telescopes are the Solar Tower Atmospheric Cherenkov Effect Experiment (STACEE) in Albuquerque, NM and Keck Solar Two at Barstow, CA.

9.3.6 Extensive Air Shower Array

In the UHE gamma ray regime (100 TeV–100 PeV), an air shower technique (AST) involving the Extensive Air Shower (EAS) array is used for gamma ray detection. In this technique, a giant array on the ground collects secondary photons and particles. Some of the large-area EAS arrays also collect air Cherenkov radiations.

There are two reasons for applying the EAS arrays for UHE gamma ray detection. First, the ACTs operate only on cloudless moonless nights while the EAS array operates continuously. This is very important for searching transient sources, such as gamma ray bursts, or for studying time variation of "steady" sources. Second, the air shower spread area is directly related to the energy of the gamma ray photon. The higher the photon's energy is, the larger the area the air shower spreads. The collecting area of an ACT array is still limited and the energy of the photon detected is still below 200 GeV. For higher photon energy of 100 TeV, a very large-area EAS array is necessary.

In an EAS array, a large number of detectors, which are generally shielded from low energy photons, are sparsely located over a much larger area. They detect the air shower particles that have survived at the detection altitude. The detectors of an EAS array generally include shielded plastic scintillators, water tank, or other particle detectors with photomultiplier tubes. The photomultiplier tubes record the light from the Cherenkov effect inside scintillators, water, or liquid produced by the secondary gamma-ray air shower particles. ESA array facilities can also be used for cosmic ray detection. Therefore, they may be called cosmic ray telescopes instead of gamma ray ones, especially for very large area arrays where the chance to detect high-energy cosmic ray particles is much higher than gamma ray photons. Since the sparse detectors of the array sample only a small part of the shower, it is more difficult to discriminate gamma ray from cosmic ray by using this type of instrument.

There are a number of differences between the ACT and the EAS arrays. These differences are: (a) The EAS array can be used day and night, while the ACT array can only be used on moonless cloudless nights; (b) the ACT array works at the pointed direction (+/−2 degrees), while the EAS array has a much wider field of view ($>45^\circ$); (c) the energy threshold for ACT arrays is about 200 GeV, while that for the EAS arrays is >50 TeV; and (d) the background rejection of ACT arrays is excellent (about 99%), while that for EAS arrays is very limited (about 50%).

Cosmic rays, which are charged particles and arrive from all directions due to the effect of the interstellar magnetic field, are the major source of background noise in gamma ray detection. To search for photon signals in the background, there are a few basic methods: (a) Improve the detectors' angular resolution. The smaller the regime the detector covers, the smaller the background level. (b) Use good gamma/hadron separation in the data reduction, since the gamma ray flux is about 3–4 orders lower than that of cosmic rays. Gamma ray observation has a terrible background. Air showers induced by cosmic rays are more likely asymmetrically spread. They are also more likely to include hadrons and muons. Although it is impossible to identify the primary source on an event by event basis, one can find parameters on which to select in order to decrease the number of hadrons relative to the number of photons. If the expected background is B, the significance of M measured events is approximately given by $(M - B)/B^{1/2}$. If imposing some cutoff, then the quality factor may be equal to $Q = [(M_c - B_c)/B_c^{1/2}]/[(M - B)/B^{1/2}]$. In this way, 99% of the hadrons are reduced while over 50% photons are retained. The quality factor can reach about nine by this method (Mincer, 2001). And the last, (c) compare observation results with other instruments for redundancy.

Multi Institution Los Alamos Gamma Ray Observatory (MILAGRO) is an EAS array instrument. It is a water-filled pond of 60 m × 80 m × 8 m at an altitude of 1.63 km. There are two layers of photo-multiplier tubes (PMTs): 450 at a depth of 1.4 m below the water surface and 273 at a depth of 6 m. When an air shower hits the detector, the secondary charged particles at a speed faster than light in water emit Cherenkov radiation along a 41 degrees cone. Photons in these showers are typically five-times more numerous as the charged particles. By detecting these photons, a gamma ray or cosmic ray effect is detected.

Very large-area EAS arrays work in the UHE regime (100 TeV–100 PeV) and the fluorescence detectors in the EHE (Extreme High energy) regime (100 PeV–100 EeV). These high energy detectors are more often used for cosmic ray detection instead of for gamma ray detection. For this reason, very large-area EAS arrays and the ground- or space-based fluorescence detectors are discussed with cosmic ray telescopes in Chapter 10.

9.3.7 Major Ground-Based Gamma Ray Projects

The Whipple VERITAS 10 m telescope array in Arizona, US is a US, UK, and Ireland facility with an altitude of 1.3 km. It consists of seven 10 m telescopes. The MAGIC array has two telescopes of 17 m size with active mirror surface in La Palma, Spain (Figure 9.35). The Major Atmospheric Cerenkov telescope Experiment (MACE) is another ACT array of two 17 m telescopes located at an altitude of 4.2 km at Hanle in the Ladakh regime of Northern India. Its energy threshold is about 20 GeV. The MACE and Multi-element Ultra-Sensitive Telescope for Quanta of Ultra-high Energy (MYSTIQUE) are parts of the Gamma-ray Astrophysics through Coordinated Experiments (GRACE)

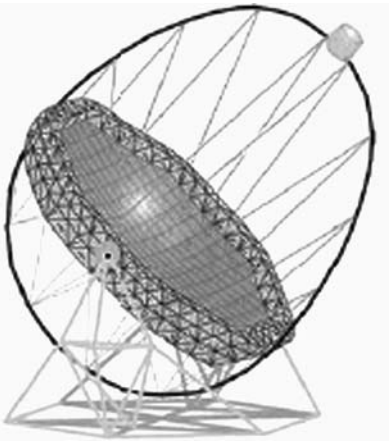


Fig. 9.35. Sketch of the MAGIC telescope (Moralejo et al., 2004).

project. Other ACT arrays include the HESS four 12 m telescope array of Germany, France, Namibia, South Africa and the UK in Namibia, the Cherenkov Array at Themis (CAT) in France, and CLUE (Cherenkov Light Ultraviolet Experiment) at La Palma, Spain.

The solar tower arrays include SOLAR TWO, also named Keck Solar Two, Solar Tower Atmospheric Cherenkov Effect Experiment (STACEE) of the US, and CELESTE of France. The SOLAR TWO has 32–64 heliostats; each has an area of 40 m^2. The reflectors are distributed over a circle of about 200 m in diameter. The secondary mirror on the central tower is a spherical one with a curvature of 6 m. STACEE includes 48 to 64 heliostats with an area of 37 m^2 and it is located in Albuquerque, NM. It is part of the National Solar Thermal Test Facility (NSTTF).

The EAS gamma ray facilities include Multi Institution Los Alamos Gamma Ray Observatory (MILAGRO) in the US and ARGO-YBJ in Tibet. The ARGO-YBJ array detectors are resistive plate chambers (RPCs) with scintillator material on mountain top. Other large-area EAS arrays forming cosmic ray telescope facilities will be discussed in Chapter 10.

References

Bennett, H. E., 1979, Techniques for evaluating the surface finish of X-ray optics, in Proceedings of the SPIE, Vol 184, ed. M. Weisskopf.

Bennett, J. M. and Mattsson, L., 1999, Introduction to surface roughness and scattering, 2nd ed., Optical Society of America, Washington D. C.

Chantel, M. et al., 1997, Nucl. Instrum. Meth (Magnussen, N.,)

Chen, H., 1985, Infrared physics, Defense industrial press, Beijing.

Chen, P., 1978, Photometry and instruments of ground based infrared astronomy, J. Yunnan Observatory, (2).

Cheng, J., 1988, X-ray imaging optical system, Tech. Opt. Inst., (3) 28–34.
Fagio, G. G., 1979, A review of infrared and sub millimeter astronomy with balloon-borne telescopes, Infrared Phys, 19, 341–351.
Fazic, G. G. ed., 1977, Infrared and submillimeter astronomy, Proceedings Vol 63, D. Reidel Pub. Co., Dordrecht.
Fenimore, E. E. and Cannon, T. M., 1978, Coded aperture imaging with uniformly redundant arrays, Appl. Opt., 17, 337.
Giacconi, R, Harmon, N. F, Lacey, R. F. and Szilagyi, Z., 1965, Aplanatic telescope for soft X-ray, JOSA, 55, 345.
Gunson, J. and Polychronopulos, B., 1976, Optimum design of a coded mask x-ray telescope for rocket applications, MNRAS, 177, 485.
Heitler, W., 1954, The quantum theory of radiation, 3rd ed., Oxford University Press, UK.
Huang, T., et al., 1986, Astronomy beyond visible light, Science Press, Beijing.
In't Zand, J. J. M., 1992, PhD Thiese, University of Utrecht.
Jet Propulsion Laboratory, 1974, The NASA/JPL 64-meter-diameter antenna at Goldstone, California: project report, Tech memo 33–671, NASA.
Kaplan, D. et al., 1976, A large Infrared telescope for spacelab, Final report, ESA-CR(P)-833- 1.
Kitchin, C. R., 1984, Astrophysical techniques, Adam Hilger Ltd, Bristol.
Korsch, D., Wyman, C. and Perry, L. M., 1979, Influence of alignment and surface defects on the performance of X-ray telescopes, in Proceedings of the SPIE, 181 ed. M. Weisskopf.
Meinel, A B. and Meinel. M. P, 1986, Very large optics of the future, Optics News, 12(3).
Mincer, A. I., 2001, Gamma ray astronomy with air shower arrays, CP587, GAMMA2001: Gamma ray astrophysics 2001, ed. S. Ritz, AIP, New York.
Moralejo, A. et al, 2004, The MAGIC telescope for gamma-ray astronomy above 30ReV, Mem. S. A. It., 75, 232.
Qiang, Z., 1984, Astronomical infrared detectors, Progress in Astronomy, 12, 167–177.
Swanson, P. N. et al., 1986, System concept for a moderate cost large deployable reflector (LDR), Opt. Eng., 25, 1045–1054.
Traub, W. A., and Stier M. T., 1976, Theoretical atmospheric transmission in the mid- and far-infrared at four altitudes, Appl. Opt., 15, 364–377.
Van der Hucht, K. A, and Vaiana, G., eds. 1978, New instrumentation for space astronomy, Pergamon Press, Oxford.
VERITAS, 1999, VERITAS proposal, submitted to the Department of Energy by Iowa state University. Purdue University, the Smithsonian Astrophysics Observatory, and Washington University.
Weekes, T. C., 2001, The next generation of ground-based gamma ray telescopes, CP587, GAMMA2001: Gamma ray astrophysics 2001, ed. S. Ritz, AIP, New York.
Weisskopf, M., 1979, Space optics: imaging X-ray optics workshop, Proc. SPIE, 184, p301.
Ye, Z., 1986, Detection techniques for cosmic ray radiations, Science press, Beijing.

Chapter 10
Gravitational Wave, Cosmic Ray and Dark Matter Telescopes

This chapter provides an overview of all nonelectromagnetic wave telescopes for astronomy, which include gravitational wave, cosmic ray, and dark matter telescopes. The theory, design principles, structural arrangement, and the limitations of these special telescopes are introduced step-by-step. In the gravitational wave telescope part, resonant bar and laser interferometer telescopes are discussed in detail. The vibration isolation, the low detecting limit, and the detection of gravitational waves through pulsar observation and microwave background observation are also discussed. Other space gravity probes are also introduced. In the cosmic ray telescope part, extensive air shower arrays, optical fluorescence telescopes, radio fluorescence telescopes, and space magnetic spectrometers are discussed together with their operating principles. In the dark matter part, various neutrino telescopes are discussed. These include scintillator, heavy water, liquid argon, high Z neutron, and radio Askaryan effect detectors. The cold dark matter telescopes include cryogenic calorimeters (or thermistors), scintillation detectors, and resonant cavity detectors. The chapter also provides reviews of existing or planned telescopes for nonelectromagnetic waves or particles.

10.1 Gravitational Wave Telescopes

10.1.1 Gravitational Wave Fundamentals

So far, we have examined telescopes designed exclusively for electromagnetic (EM) radiation. Electromagnetic radiation delivers important information about our universe because this radiation is directly related to physical changes in temperatures and in electrical charges. However, there are other carriers of information that are equally important in our understanding of the universe. Among of these, gravitational wave is one which is directly related to the mass, energy, and acceleration of celestial objects.

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_10,

The existence of gravitational waves predicted by Einstein's theory of general relativity. In Einstein's view, space-time is analogous to a flat piece of fabric. When a concentration of mass or energy appears, there will be a curvature on the space-time field. Gravity is directly related to the space-time curvature. When the curvature is weak, it produces the classical Newtonian gravity that governs motion within our solar system. When the curvature is strong, it will behave nonlinearly and can even amplify itself to produce a space-time singularity. The space-time curvature produced by a concentrated mass is extremely slight and, consequently, very difficult to detect. However, when the space-time curvature varies rapidly, it should produce curvature ripples that propagate through the universe at the speed of light. This is the case where mass or energy accelerates rapidly. The spreading of the space-time curvature is called gravitational (G) waves.

In our universe, the EM waves are direct results of charged particle acceleration. Similarly, the G waves are results of the mass or energy acceleration. The EM and G waves are both fields which propagate at the speed of light in space. The characteristics of the EM waves, including frequency, wavelength, polarization, and energy, are determined by the charge and acceleration of the moving entities. The same is true for the G waves. The G waves may have different frequency, magnitude, and polarization as the mass which generates the G wave has different shape and different acceleration. Photon particles can be used to represent the EM waves, while the G waves can also be associated with a special type of particle: the graviton. Of course, questions exist in the theory of gravitons. One of these questions is: does the graviton have energy or mass in a vacuum?

Using space-time to describe the G waves, the space-time without curvature will be in an isotropic coordinate system. This system can be assumed to be a perfect spherical surface. However, when a G wave exists, the space-time will not be isotropic any more. In one direction of the space-time field, the dimension will be expanding and contracting alternately. In another perpendicular direction, the dimension will be contracting and expanding alternately. Therefore, the G wave will propagate outwards in a direction perpendicular to both the above mentioned directions. The G wave is also a transverse wave. If a G wave propagates in the z direction, the wave will have two types of polarizations: one in the x and y direction or in a '+' (plus) direction and the other in the 45° angle direction from x and y axes or in the '×' (cross) direction. If the G wave is a '+' polarized wave, the space dimensions along the x and y coordinators will change periodically. If the G wave is '×' polarized, the dimensions along the 45° angles of the x or y planes will change periodically. The magnitudes of these two components of a G wave can be represented as $h_+(t)$ and $h_\times(t)$, which are relative dimension changes (or strains) of the space-time field. Any G wave is the sum of these two polarized components. However, theories of the G waves are almost entirely theoretical and untested at present.

Although there are similarities between G waves and EM waves, they are actually two completely different wave forms. Linear equations govern the EM

waves while those for G waves are nonlinear. Nonlinear equations will not have the property of superposition and solving nonlinear equations is also more difficult. When the G wave is weak, linear approximations may be used, therefore, the solution may be easier to find.

The energy of the G waves is much weaker than that of the EM waves. The estimated energy of a G wave is only 10^{-38} that of a corresponding EM wave, so that only acceleration of those masses with an astronomical magnitude will produce an observable G wave. The EM wave will interact with many materials, but the EM waves only interact occasionally with each other through interference. The G wave will not interact with most materials, so that the G wave can penetrate the earth without being noticed. However, the G waves themselves can interact with each other and can interact with the space-time. Because the EM waves can interact with many materials, telescopes of large surface area can be used for the collection and detection. This is especially important for the study of very weak EM waves. However, the G wave has no interaction with materials, therefore, it is impossible to collect and detect these very weak G wave signals directly. The only way to determine the existence of the G wave is through the measurement of the space-time deformation induced variations of a mechanical structure that has a very low resonant frequency and cannot respond quickly. The wavelength of the EM wave is smaller than the dimension of the source, while the wavelength of the G wave is the same or larger than the dimension of the source. The G wave frequency possibly detected is lower, generally below 10 kHz. The flux of the EM wave is inversely proportional to the distance squared to the source, but the strain of the G wave is only inversely proportional to the distance to the source.

The existence of a mass will produce a curvature in space-time. From this theory, the trajectory of light passing through the edge of the sun will bend which has been confirmed by astronomical observation. The acceleration of a mass will produce a G wave in space. However, from Newton's third law, acceleration in one direction of a mass will accompany acceleration in another direction of another mass. The momentum of two masses, that is the product of their velocity and mass, will be equal but in opposite directions. Following this, the G waves produced by both masses will cancel each other. However, if the motion of the two masses is not on a line, a G wave will be created. A type of motion to produce G waves is through rotation of a mass or masses. The G wave generated is related to the arrangement of masses, measured by what is called "quadropole (or higher order) moment" of the mass (means four mass concentrations of an object). The contribution of the dipole term of a mass (means two mass concentrations of an object) is zero. The larger the quadropole moment, the stronger the G wave generated. The quadropole moment of a sphere is zero. However, any shape away from a sphere will possess a quadropole moment. From this theory, binary star systems, especially those in a tight orbit, will be important sources of the G waves. Inside or around black holes, the movement of masses should have a spinning feature and a very high acceleration so that observable strong G waves may be produced. Other objects, such as supernova,

black hole binary, cosmic strings and domain wall, also produce strong G waves. The oscillatory characteristics of the G waves carry detailed information about their sources. Therefore, the observation of G waves is very important for astronomers as it could discover dark mass and other mysteries in our universe.

Russell Hulse and Joseph Taylor (Hulse, 1975), and Taylor and J. M. Weisberg (1989) published papers which showed the existence of G waves through the observation of a pulsar binary. The papers show that the tiny difference in the period change of the motion is equivalent to the energy loss through the G wave produced. In 1993, Hulse and Taylor won the Nobel Prize for this discovery.

The observation of G waves is to estimate the gravitational wave strain that acts on the detector. The strain, or relative length change, is the characteristic amplitude of the G wave which is very small. For example, the G wave strain generated by rotating a steel rod of 500 tons at the highest speed achievable on earth will be about 10^{-40} which is not detectable. The G wave generated by a neutron binary coalescence in Virgo will produce a strain of 10^{-21} on the earth. This is like measuring the size of an atom in a distance between the sun and the earth. This magnitude, however, can be measured by using modern detecting techniques. By estimation, the strain of the G waves on earth from various astronomical sources can be of the order of about 10^{-20}. The strains of G waves from black holes or asymmetrical supernova have a relatively large strain. Since the strain received from the earth is inversely proportional to the distance between the sources and the earth, for a number of G wave sources with their distance to earth smaller than 100 Mpc (1 pc equals 3.2 light years), the strains are larger, ranging from 10^{-20} to 10^{-22}. The frequency of binary star produced G waves is directly related to their periods, or their separations. If the separation of binary objects is smaller than 100 km, the period is a few seconds and the frequency of the G wave generated is about 100 Hz. If the separation is smaller (about 20 km), the frequency of the G wave produced will be higher (about 1,000 Hz).

10.1.2 Resonant Gravitational Wave Telescopes

Attempts to detect the mystery G waves started in the 1960s with Joseph Weber who used a large solid metal bar for this purpose. This bar detector is a resonant gravitational wave telescope or resonant detector. It is like a xylophone struck by a mallet, but the bar is struck by the gravitational wave. The material used for the bar should have a high quality factor such as aluminum or niobium. The quality factor, or Q factor, is the reciprocal of the damping coefficient in a dynamic system. To capture as much energy as possible from the G wave, the mass of the bar needs to be as large as possible and the bar should have a higher resonant frequency larger than 1 kHZ. The detector is influenced by many environmental factors, even sophisticated isolation and damping are applied. These factors include turbulence from seismic waves, vibrations due to sound

waves, strains induced by temperature variations, and pressure fluctuations caused by the Brownian motion of the air molecules. False detection can be avoided by using two carefully suspended metal bars at very low temperature and in nearly vacuum containers, each in a different far away underground location. The details concerning vibration isolation will be discussed in the next section.

When a bar is hit by a G wave burst with very short duration (∼1 ms) and with an impinging wavefront perpendicular to its axis, the bar will start to vibrate at those resonant modes which are coupled to the G wave for a very long time (∼1 h). To record the vibration, a number of very accurate strain sensors are attached. These sensors use very low noise DC SQUID (superconducting quantum interference device) amplifiers which operate near to the quantum mechanical energy resolution limit and have very low thermal noise contribution. Weber's early device included two aluminum bars 1,000 km apart. In this way, a number of turbulent factors are excluded. The sensitivity of the measurement was claimed as 10^{-18} and, in the following year, Weber published results of his G wave detection. However, later study showed that the responses he recorded are far larger than the real response caused by G waves. For many reasons, the noise level is much larger than the signal.

To understand the detecting power, we can use S_h to represent the power spectral density of the input noise at the bar detector (Pizzella, 1997):

$$S_h = \frac{\pi}{8} \frac{kT_e}{MQL^2} \frac{1}{f_0^3} \tag{10.1}$$

where k is the Boltzmann constant, M the mass, Q the quality factor, L the length of the bar, T_e the thermodynamic temperature plus a term due to the back action (feedback) from the sensor (for SQUID amplifier, the term is negligible), and f_0 the resonance mode coupled to the G wave. If the detecting bandwidth is Δf, then the minimum value of detecting the dimensionless strain amplitude of the G wave with SNR=1 is:

$$h \approx \frac{1}{\tau_g} \sqrt{\frac{S_h}{2\pi\Delta f}} \tag{10.2}$$

where τ_g is the recorded duration. The bandwidth of the resonant detector is limited by the sensors and by the noise of the electronic amplifier, because the resonant bar responds in the same way to the excitation due to the G wave and to the Brownian noise. The bandwidth is:

$$\Delta f = \frac{f_0}{Q} \frac{4T_e}{T_{eff}} \tag{10.3}$$

where T_{eff} is the noise temperature for the G wave detector. Summing these formulae, we have the minimum value of detecting amplitude:

$$h \approx \frac{L}{\tau_g v^2}\sqrt{\frac{kT_{eff}}{M}} \tag{10.4}$$

where v is the sound velocity in the bar material (for aluminum, $v = 5,400\,\mathrm{m/s}$). From the formula, it can be found that with $M = 2{,}300$ kg, $L = 3$ m, and $\tau_g = 1$ ms, in order to observe a G wave of the order of $2\cdot10^{-18}$ from the galactic center, the T_{eff} needed is only 0.2 K. If the same G wave is from the Virgo cluster, T_{eff} needed is only $1.4\cdot10^{-7}$ K. From this calculation, early room temperature resonant detectors will not be able to detect any real G wave signals. The low temperature and vacuum environment makes the detector more sensitive to the G wave signals. That is why modern resonant detectors are all operated at extremely low reachable temperature (0.02–6 K).

It is possible to compute the bandwidth from Equation (10.3). If f_0 is 1 kHz, $T_{\mathrm{eff}} = 1.4\cdot10^{-7}$ K, $T_{\mathrm{e}} = 0.1$ K, and $Q = 1\cdot10^{7}$ ($Q = 4\cdot10^{6}$ at 100 mK for Al5056), then $\Delta f = 300$ Hz.

Using a resonant detector, it is also possible to detect the G wave stochastic background. The relationship between the G wave density Ω and the power spectrum S_h of h is (Pizzella, 1997):

$$\Omega = \frac{4\pi^2}{3}\frac{f^3}{H^2}S_h(f) \tag{10.5}$$

where H is the Hubble constant. In the design of resonant detectors, the sensitivity will be larger when a larger mass is used (Equation 10.2). The bar material and dimensions are chosen as a good compromise between the cost and frequency constraints (the expected signals are at about 1 kHz). Moreover, the mass of the detector, where the cross-section is signal dependant, cannot be too large as the cooling power of actual refrigerators is still limited. A number of detectors have the mass of 2,300 kg and the dimension of about 3 m in length and 60 cm in diameter. The Young modulus of the material is a function of temperature. At room temperature, the resonance of such a bar is 874 Hz and, at 0.1 K operational temperature, the resonance is 920 Hz.

The crucial development of resonant detectors is to enlarge the bandwidth. High-Q of the material seems to limit the bandwidth, whereas, in fact, this is not true. Both strain and thermal Brownian motion noise exhibit the same resonant response near the mode frequency. Thus, the signal-to-noise ratio is not bandwidth limited but limited by the detector thermal noise. Significant limitation comes from the transducer and amplifier. This is why SQUID-type near quantum-limited amplifiers and microwave, inductive or optical transducers are now used in this type of detector.

In addition to bar resonant detectors, there is a new spherical shaped resonant detector named the resonant antenna. The term antenna here means

a G wave detector. The spherical detector can be heavier; a 3 m diameter sphere of aluminum will weigh 38 tons instead of 2.3 tons for a 3 m long cylinder. At the same time, spherical detectors are sensitive to G waves from all directions and of any polarization. A spherical detector is equivalent to five bar detectors, each one with the mass of the sphere. A 3 m sphere at 1 kHz resonance for a 1 ms burst has a sensitivity of 8×10^{-22}. In general, resonant detectors are narrow-band devices, which limit the windows of frequencies in the observation. In this aspect, their sensitivity is high while the cost is low.

10.1.3 Laser Interferometer Gravitational Wave Detectors

Resonant detectors detect G waves through strain change of a continuum body mass. Using finite element analysis, a large mass can be modeled as small mass elements connected by springs. If two masses are separated in space, a G wave burst will change the distance between them alternatively, even the distance change is tiny. If there are two groups of masses that are perpendicular to each other on a plane which is, in turn, perpendicular to the G wave propagation direction, the distance between masses of one group will expand and contract alternatively relative to the other group. If a G wave travels on the same plane of the masses, then the distance between masses in the G wave propagation direction will not change, while that in the other direction will change periodically. This is the operational basis of a laser interferometer gravitational wave telescope.

A laser interferometer gravitational wave telescope is a Fabry–Perot style Michelson interferometer which measures a very small distance change between two groups of objects separated in space. The basic arrangement of the detector is an L-shaped instrument as shown in Figure 10.1. In the interferometer, there are two perpendicular arms of the same length ($L_1 = L_2 = L$). On each arm, there are two test masses made of fused silica suspended by fine fused silica fibers. The masses are polished and coated as mirrors to reflect the laser beam back and forth. These mirrors have high reflectivity, low transmissivity, very low scattering and absorption. A laser source sends a beam through a beam

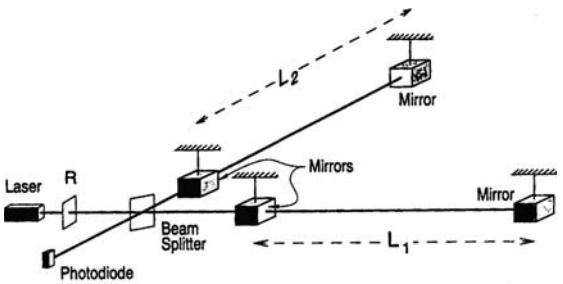


Fig. 10.1. A schematic view of the laser interferometer gravitational wave detector (Abramovici et al., 1992).

splitter to reach the two arms of the instrument. After multiple reflections between two mirrors of each arm, two laser beams return to a photodiode through the beam splitter. The suspended mass mirrors have a low resonant frequency of about 1 Hz. If the frequency of a G wave is larger than this frequency, the mirrors will freely move in a horizontal direction. If the propagation direction of a G wave is perpendicular to the instrument plane, all the mirrors will move periodically in the instrument plane, changing the lengths of the two arms. A length change ($dL = L_1 - L_2$) is produced as one arm length increases while the other decreases. The relative length change represents the characteristic amplitude of the G wave, $h = dL/L$.

A laser interferometer gravitational wave detector measures a tiny distance change between two groups of test mirrors through interference of two laser beams. In this way the existence of a G wave is recorded. The tiny distance change represents the change of optical path lengths of the two arms, so that fringes are recorded in the photodiode receiver. For improving the sensitivity of the instrument, the first mirrors near the beam splitter are partially transparent so light beams can bounce back and forth between the two mirrors of each arm before they reach the photodiode. The same as in a Fabry–Perot interferometer, the relative optical path lengths are multiplied by the number of multi-paths as mdL. The real phase error between two beams is:

$$\Delta\Phi \approx m\Delta L/\lambda = mhL/\lambda \tag{10.6}$$

This phase error can be detected. If all the possible measurement errors are excluded, the limiting accuracy is determined by photon shot noise,

$$\Delta\Phi \approx 1/\sqrt{N} \tag{10.7}$$

where N is the number of photons incident on the beam splitter during the integration time, which is about a period of the G wave. So the magnitude of the G wave detected is:

$$h_{\min} \approx \lambda/(mL\sqrt{N}) \tag{10.8}$$

Actually, $\Delta\Phi$ is proportional to m and $h_{\min}$ is proportional to $1/m$ when mL/c is less than half of the G wave period, where c is speed of light. If m is larger, there is no further improvement of $h_{\min}$. As an example, if $L = 4\,km, m = 400$, the frequency of the G wave is 100 Hz and the scatter and absorption loss is 1%, the total power available in the interferometer for the measurement is about 100 times the laser's output power. If the laser power is 60 W, this means that $N = 2 \times 10^{20}$ photons during a 10 ms integration time. The detecting wave amplitude is:

$$h_{\min} \approx 0.5\,\mu\text{m}/(400 \cdot 4\,\text{km}\sqrt{2 \times 10^{20}}) \approx 10^{-23} \tag{10.9}$$

This is the limiting number which the Laser Interferometer Gravitational wave Observatory (LIGO) telescope can achieve. From this formula, it can be seen that the longer the optical path, the more sensitive the G wave detection. However, the length of the optical path cannot be too long as it will affect the detection of high frequency G waves as the strain change may smooth out within the detecting period. The other approach in increasing the sensitivity is to limit the frequency bandwidth.

Another interferometer configuration that can be used as a G wave detector is a polarization Sagnac interferometer shown in Figure 10.2 (Beyerdorf et al., 1999). The principle of its detection is the same as in the earlier Michelson interferometer. The Sagnac interferometer most often used is in the fiber gyroscope configuration. In this interferometer, each portion of a separated beam travels the same path, but in counter-rotational directions. This makes the system less sensitive to the mirror surface errors and other defects. Unfortunately, these benefits are almost lost when multi-path techniques are introduced for the fringe enhancement.

Similar to resonant G wave detectors, errors of laser interferometer detectors also come from seismic turbulence, thermal effects, Brownian motion of air molecules, and other interference to the test masses. Laser frequency stability and bandwidth also have an important impact on the distance measuring accuracy. Since air molecules buffet the test mirrors and affect the phase of the laser beam, so that all of the important parts have to be isolated from air molecules. It is also necessary to control the instrument temperature and put all important components, including the light path, in cold and vacuum conditions. Even so, some nonGaussian bursts and other noises are unavoidable. Therefore, two identical instruments at two widely separated sites are necessary for signal discrimination. In observation, only signals detected by both instruments are recorded for analysis. To know the exact direction of the incoming G wave, three identical instruments may be necessary.

To measure test mass displacement, we measure the phase shift between two beams in the two different arms. The output power from the interferometer will

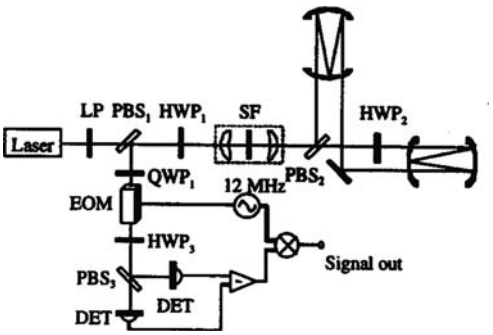


Fig. 10.2. The layout of the Sagnac interferometer (Beyersdorf, 1999).

be the input power modulated by the cosine squared of the phase shift. Therefore, if we have noise in the laser, we have noise in the displacement measurement. The readout noise of the laser is the shot noise plus the pressure noise. The shot noise on the length of the arm is (de Michele et al., 2001):

$$h_{sh}(f) = \frac{\delta L}{L} = \frac{1}{L}\sqrt{\frac{kc\lambda}{2\pi T(f) P_{in}}} \tag{10.10}$$

where L is the length of the arm, k the Plank constant, c the speed of light, λ the wavelength, $T(f)$ the transfer function of the system, and P_{in} the input laser power. In this formula, the noise decreases as the input power increases. The noise also decreases when the wavelength decreases. Therefore, it is essential to use a high power laser for minimizing the shot noise. Nd:YAG lasers, with a wavelength of 1,064 nm and power of 200 W in cw (continuous wave) are the lasers of choice for the new generation gravitational wave interferometers (Willke et al., 2006).

The pressure noise is due to the force of the light on the mirror:

$$h_{pr}(f) = \frac{\delta L}{L} = \frac{2}{mf^2}\sqrt{\frac{kT(f)P_{in}}{8\pi^3 c\lambda}} \tag{10.11}$$

where m is the test mass. The readout noise is the root sum square of the two terms. Other laser noises come from frequency noise (random change of frequency), intensity noise, and pointing and angle fluctuations. The laser system for a laser interferometer has two output beams, one of high power and one of low power. The frequency of the low power beam is compared with the frequency of a resonant cavity. The error is sent to a frequency stabilization servo-system to minimize frequency noise. The high power beam enters a pre-mode cleaner cavity and an intensity servo-system before sending to the interferometer. The beam retains a near-perfect TEM_{00} Gaussian mode and the intensity noise is reduced.

In the design of laser interferometer G wave detectors, one of the most important tasks is to isolate the test masses from any possible vibration. This is similar to the resonant detector design. The vibration isolation structures of the test mass typically include isolation stacks, vertical springs, filters, and double pendulums. The stack itself is a series of mechanical vibration filters that consist of masses and elastomer (rubber-like material) springs. One typical design consists of three legs, each consisting of layers of stainless steel masses and graphite-loaded silicone rubber. Sometimes, one stack is also supported by three mass-spring systems. They are encapsulated in stainless steel bellows. The stack should have low resonant frequency and reasonably high Q factor, so that the transmissibility above resonances falls off as f^{-2} per layer over the frequencies of interest. The damping of silicon rubber is smaller than required, so that synthetic graphite has to be added before the rubber is cured. The addition of synthetic graphite of 6% by weight can bring the Q factor down from $\sim$20 to $\sim$12.

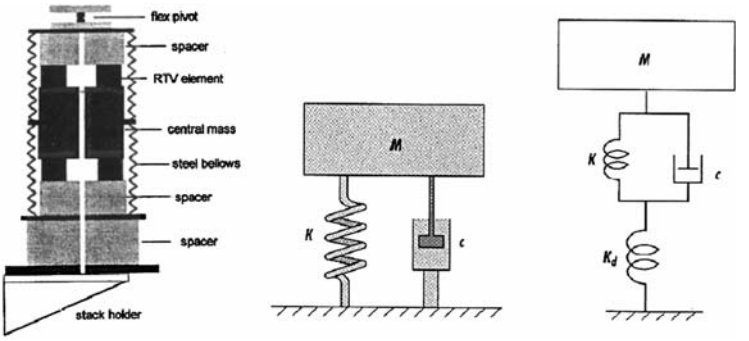


Fig. 10.3. Detailed stack design and vibration isolation principle (Plissi et al., 1998).

Single mass, spring, and dashpot system (middle of Figure 10.3) has its transfer function as:

$$T = \frac{As + B}{Cs^2 + As + B} \tag{10.12}$$

where $A = 1/k$, $B = 1/d$, $C = M/(dk)$, M is the mass, k the spring constant, and d the damping coefficient. For this system, the transmissibility above resonance falls off as f^{-1}. However, as another layer(s) of spring structure is added on the bottom of the spring and dashpot part as shown on the right side of Figure 10.3, the transfer function of the new system becomes:

$$T = \frac{As + B}{Ds^3 + Es^2 + As + B} \tag{10.13}$$

where $D = M/(kk_d)$, $E = M(k + k_d)/(dkk_d)$, and k_d the additional spring's constant. The transmissibility of the new system is inversely proportional to the square of the frequency. The new system is more effective in vibration isolation. Outside the stack devices are the stainless steel bellows. These bellows have large rotational stiffness, which can transmit vibration to the test mass. A rotational flexure has to be incorporated on top of each stack structure.

For further improving the vibration isolation, some G wave detectors also use cantilever-type springs to support the mass in a vacuum chamber. Usually a ring is on the top of three stack legs, which is known as a stack stabilizer. Above the stabilizer, another support ring is added to align the test mass' direction. From the support ring, cantilever-type springs are attached. These springs are made with a blade geometry and constructed from special maraging (precipitation hardened) steel. Suspended, the upper mass is connected from the tapered

end of the springs. The frequency of the bending mode for these cantilever blade springs is:

$$f = \frac{1}{2\pi}\sqrt{\frac{Eah^3}{4ml^3}} \tag{10.14}$$

where m is the mass suspended, l the length of the blade, a the width of the blade base, h the blade thickness, and E the Young modulus. The choice of frequency for this cantilever stage is for a good vertical isolation. At the same time, it is necessary to avoid long-term creeping of the blade. The frequency should be about 2.5 Hz and the maximum stress should be 50% of the elastic limit. The mirror is a double pendulum structure. The double pendulum is further suspended by two sets of cantilever springs from an upper mass. At that stage the vertical mode frequency will be about 2.8 Hz.

The double pendulum provides another seismic isolation from turbulence. The test mass, as a reflecting mirror, is the lower stage of the double pendulum. To minimize the influence of temperature change, four fused silica wires suspend a pair of double pendulums. One double pendulum is the reflecting mirror, another is a reaction mass. The test mass is made of fused silica. However, the best test mass material is sapphire, which has an extremely high quality factor (10^8, while the quality factor of fused silica is 10^6). The loss factor is the reciprocal of the quality factor. The loss factor of a pendulum system $\varphi_{pend}(\omega)$ is directly related with the loss factor of the material $\varphi_{mat}(\omega)$as:

$$\varphi_{pend}(\omega) = \varphi_{mat}(\omega)\frac{4\sqrt{TEI}}{Mgl} \tag{10.15}$$

where T is the tension in each suspended fiber, E the Young modulus of fibers, I the bending moment of each fiber ($I = \pi \cdot r^4/4$ for cylinder fiber with radius of r), M the mass of the pendulum, and l the length of the pendulum.

For active and adaptive vibration control of the test mass, an intermediate reaction mass and a final reaction mass are also suspended with the intermediate test mass and the test mass (Figure 10.4). On the intermediate reaction mass, actuator coils are used to sense any unwanted vibration. Unwanted seismic noises, even a small motion of the intermediate mass, have to be suppressed. Actuators also appear in the final reaction mass. Initial lock of the optical system is difficult, but electronic feedback loops damp out the pendulum modes. Of course, all the sensors, actuators, and servo electronics have to be functional. The best situation is that the sensing of resonant modes is done using the very high sensitivity G wave interferometer itself. Another test mass type support is an active Steward platform seismic isolation system as used in the JWST space telescopes. This part as a whole is also called the Seismic Attenuation System (SAS) in gravitational wave telescopes.

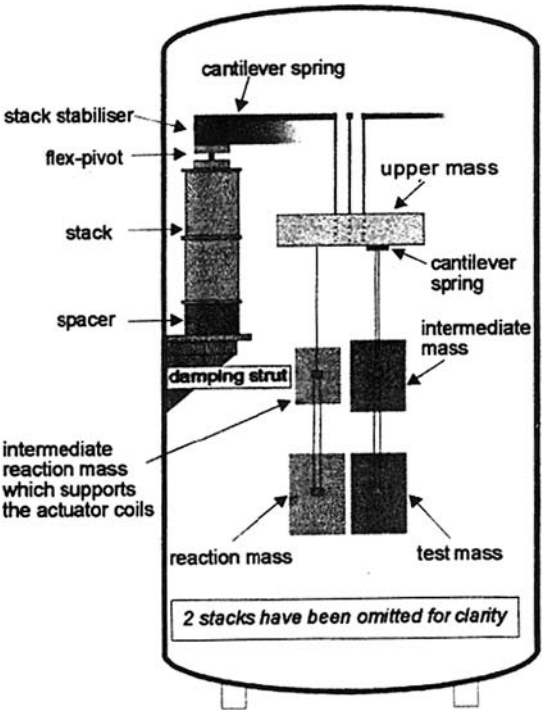


Fig. 10.4. Test mass suspension with reaction mass adjacent to it (Plissi et al., 1998).

In general, the resonant frequency of the test mass has to be as high as possible to reduce the thermal noise affect. The rms thermal displacement for an extended body at the mode frequency ω_0 is (Bernardini et al., 1999):

$$\sqrt{<z^2>} = \sqrt{\frac{kT}{m_{eq}\omega_0^2}} \tag{10.16}$$

where k is the Bolztmann constant, T the temperature, and m_{eq} the equivalent mode mass associated with the vibration mode, which is defined as:

$$\int_V \sigma \cdot u dV = \frac{1}{2} m_{eq}\omega_0^2 z^2 \tag{10.17}$$

where σ and u are the stresses and strains of the given mode and V the volume of the elastic body. A typical thermal noise number is about 10^{-19} m. After the attenuation of the suspending wires, this number may be as low as about 10^{-25} m.

The laser interferometer G wave detector requires vacuum tunnels for the laser light path. It also requires a clean mirror surface, high reflective coating, and negligible cross talk of the applied forces. Monitoring the tilt of the test mass and suspending the test mass with magnetic levitation can improve the stability

of the system. The laser used must have a very stable frequency, as even the pressure from the laser beam on the test mass is a concern in system design. For achieving high vacuum, baking of the vacuum tubes is necessary to outgas the system. For a stable laser output, a device called a mode cleaner, which itself is a complex electronic optical system, is used immediately after the laser source.

10.1.4 Important Gravitational Wave Telescope Projects

High-temperature resonant detectors do not provide the sensitivity required for G wave detection. In the 1990s, a number of lower temperature resonant detectors were installed. These included ALLEGRO (bar – US), ALTAIR, and AURIGA (bar – Italy), all at 4.2 K; EXPLORER (bar – Switzerland) at 2 K; NAUTILUS (bar – Italy) at 0.1 K, and NIOBE at 6 K, while TOKYO Crab operates at 4.2 K. The theoretical pulse sensitivity of these detectors range from 10^{-18} to 10^{-20}. However, no evidence shows that these detectors recorded any confirmed gravitational waves.

Beginning in the early 1990s, more attention was paid to the development of laser interferometer G wave detectors. In 1991, the LIGO project was funded by the US government. This $365 M project includes two instruments, each with an arm length of 4 km located respectively in Louisiana and in Washington states with a distance of 3,000 km. For increasing the path length difference, the multiple reflection number used is 50. The optical paths are inside stainless steel tubes with a diameter of 1.2 m and a vacuum of 10^{-12} the atmospheric pressure. Even so, the detection of G waves by the LIGO is still questionable.

In 2005, modification was begun on this detector. The new upgraded observatory is called Advanced LIGO. The major improvements include the reduction of seismic turbulence, using a higher power laser, completing the closed-loop control system, and using sapphire as the mirror material. After the modification, the sensitivity will increase by a factor of 15. Advanced LIGO(A-LIGO) seems likely to detect real gravitational waves for the first time. It is planned to finish in 2010.

Other existing laser interferometer detectors include VIRGO (France and Italy), GEO600 (the UK and Germany), and TAMA300 (Japan). VIRGO has an arm length of 3 km and a vacuum tube diameter of 1.2 m. GEO600 and TAMA300 have arm lengths of 600 and 300 m, respectively. In 2001, Japan started a Large-scale Cryogenic Gravitational wave Telescope, called LCGT. This will be a 3-km arm length instrument. At the same time, a project called the Australia Interferometer Gravitational wave Observatory (AIGO) was proposed. However, the possibility for "discovery" is still small. It is expected that the detection probability could reach one event a year in 2009–2010 and a few tens of events a year in 2013–2014.

It is impossible to avoid seismic turbulence with ground-based G wave detectors. They also have limited distances between test masses which produce limitations on the G wave detection. The detecting frequency range of ground-based G wave instruments is between 10 and 1,000 Hz. These correspond to astronomical

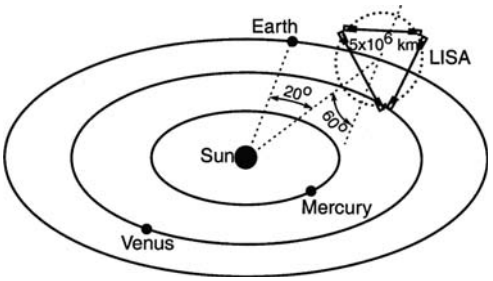


Fig. 10.5. The LISA project and its position in the solar system (Hughes et al., 2001).

events that involve very large mass/energy but a very short time scale. The probability of these events is relatively rare. To detect relatively long time scale, lower frequency G waves from the most interesting compact binary and merging black hole sources, space-based gravitational wave telescopes are necessary. Figure 10.5 shows a new space-based gravitational wave detector project: the Laser Interferometer Space Antenna (LISA). This space antenna is a joint NASA/ESA project to develop an equilateral triangular laser interferometer with three detecting stations/spacecraft (Figure 10.6). The side length of this huge triangle is about 5 Mkm. The mirrors used are proof masses suspended inside the spacecraft. The spacecraft shield the test masses from light pressure and the magnetic field of the Sun as well as their variations. The instrument will be placed in an orbit 20 degrees behind the Earth's orbit of the Sun to minimize the effects of the earth's gravity. It will be operated at a very low temperature and have a higher sensitivity and a lower frequency range (0.0001–0.1 Hz) than its earth-bound counterparts. It will also have a number of devices to suppress structural vibration, temperature

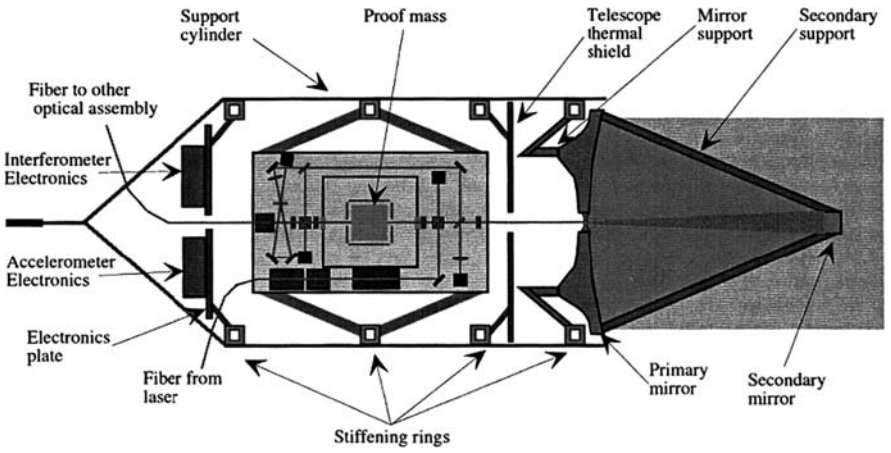


Fig. 10.6. The optics of the LISA system (Hughes et al., 2001).

variation, and other noises. The real challenge is to accurately measure and actively maintain the separations between test masses up to 10 nm accuracy. This will be performed through sets of gravitational reference sensors and micro-thrusters on the spacecraft. The optical assemblies include telescopes and optical benches for laser beam injection, transmission, reception, detection, and beam shaping. The optical benches will be solid glass blocks for attaching all components. The LISA project involves many other technical difficulties. With the aim of achieving success, a pathfinder of the project is scheduled to be launched in 2009, an unlikely goal.

10.1.5 Other Gravitational Wave and Gravity Telescopes

The ground-based laser interferometer telescopes can detect relatively high frequency gravitational waves. Space-based laser interferometer telescopes can detect lower frequency gravitational waves. For even lower frequency, such as nano-Hertz gravitational waves, a pulsar timing array can be used. The idea to use pulsars as natural detectors of gravitational waves was first explored independently by both M. V. Sazhin and S. Detweiler in the late 1970s. The basic concept is to treat the solar system and a distant pulsar as opposite ends of an imaginary arm in space. The pulsar acts as the reference clock at one end of the arm sending out regular signals which are monitored by an observer on the earth over some time-scale T. The effect of a passing gravitational wave at the pulsar or earth locations would cause a change in the observed rotational frequency by an amount proportional to the amplitude of the passing gravitational wave. If the uncertainties of the pulse time of arriving is e, then the detection sensitivity is larger than or equal to e/T and the frequencies are as low as about $1/T$ (Bertotti, et al., 1983). Since the solar system is also the subject of the impact of gravitational waves, the amplitudes detected are the sum of the gravitational wave effects at both the distant pulsar and the solar system. To determine the gravitational wave in a distant pulsar, a number of detecting results have to be used for subtracting the gravitational wave effect in the solar system. For the extremely low frequency end, a cosmic microwave background survey may be used as a gravitational wave background detector. The tiny temperature difference in the microwave background indicates small mass variation in space.

In summary, the cryogenic resonant detectors can detect gravitational waves of about 10^3 Hz in frequency (about 300 km in wavelength), the ground-based laser interferometers from 10 to 10^3 Hz, the space-based laser interferometers from 10^{-4} to 10^{-1} Hz, the pulsar observation detection at about 10^{-9} Hz, and the microwave background observation detection at about 10^{-16} Hz.

Einstein's general relativity theory includes not only the gravitational wave, but also the equivalence of space and time and the curvature of space. In the past, two gravity telescopes, Gravity Probe A and Gravity Probe B, were launched for observing these phenomena. Gravity Probe A launched in 1976 was a satellite based hydrogen maser clock. The clock rate was measured from the ground by

comparing the microwave signal from the clock to a maser on the ground and subtracting a signal of the Doppler shift from the spacecraft. The clock rate was measured for most of the duration of the flight which lasted 1 h and 55 min and compared to theoretical predictions. The stability of the maser permitted measurement of changes in a rate of one in 10^{14}. Gravity Probe A confirmed Einstein's prediction that gravity slows the flow of time. The observed effects matched the predicted effects to an accuracy of about 70 parts per million.

The purpose of the Gravity Probe B, which was launched in 2004, was to detect small wrapping of the spacetime around the earth. The probe was on an orbit 600 km over the poles of the earth. Einstein's theory predicts that large mass will wrap the local spacetime and large rotating masses would drag local spacetime around them. These two effects on the spacetime wrapping are "geodetic twisting" and "frame dragging". The main instruments on board the Gravity Probe B were four superconductor gyroscopes and a precision pointing telescope. The four gyroscopes provide a measurement redundancy. These precise instruments were well shielded from the magnetic field by superconductor ring structures at a very low temperature of 1.7 K. The gyroscopes and pointing telescope in Probe B were perfectly aligned to a fixed distant star position. The key parts of these gyroscopes were perfect spherical fused quartz and silicon balls coated with the superconductor material niobium. The ball diameter was 1.5 inches and the deviation from a true sphere was only 40 atomic layers. The ball was actively controlled to have its position unchanged inside the gyroscope chamber walls. The accuracy of these gyroscopes was about 0.1 milliarcsec, 30 million times more accurate than any other existing gyroscopes. However, minute spacetime wrapping would produce a tiny tilt of the gyroscope axes. These tilts would be detected by superconductor quantum interference devices (SQUID). The detecting accuracy was about 0.0001 arcsec. The predicted axis shift is about 0.0018°. After 18 months of data analysis for Probe B, the measured axial shift is within 1% of Einstein's prediction (Figure 10.7).

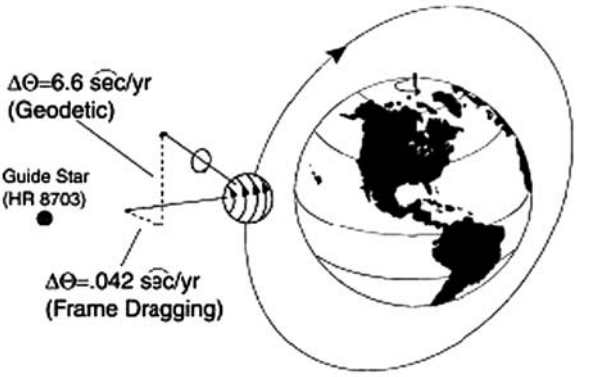


Fig. 10.7. The detection principle of Gravity Probe B (NASA).

10.2 Cosmic Ray Telescopes

10.2.1 Cosmic Ray Spectrum

Both the EM and G waves are fields where some type of physical quantities are assigned to every point in space, although they could also be considered as particles. Cosmic rays are indeed particles and most of them even have mass at zero velocity. Cosmic rays are high-speed charged particles, including electrons, positrons, protons, ions, anti-particles, and anti-matter. Major cosmic rays are composed of nuclei, from the simplest hydrogen nucleus (a proton) to the iron nucleus and beyond. The primary ones are mostly nuclei with atomic weights less than 56. The cosmic rays having the highest energy are moving at speeds close to that of light and have the highest energies of any naturally occurring particles. Since they are charged particles, the magnetic field around celestial bodies will bend the path of cosmic rays. Therefore, the incoming directions of cosmic rays appear random and they do not provide the direction of ray sources in most cases. The distribution of cosmic rays is also latitude dependent on the earth surface. In some references (Giovannelli, 2004), cosmic rays also include gamma rays and neutrinos which are not charged particles. In this book, we assume all the cosmic rays are charged particles; therefore, they include electrons and positrons, but not gamma rays and neutrinos.

The energy spectrum of cosmic rays has been measured from 10^6 up to 10^{21} eV. The highest energy of the cosmic ray is the same as the energy of a well hit tennis ball at 160 km/hr, but here it is all packed into a single atomic nucleus. The energy spectrum of cosmic rays at the top of the atmosphere spans many decades. Above 10 GeV, the cosmic ray spectrum can be described as (Bergstrom, 2004):

$$\frac{dN}{dE} \propto E^{-\alpha} \tag{10.18}$$

where α is 2.7 for energy smaller than $10^{16}\,\mathrm{eV} = 10\,\mathrm{PeV}$ and is 3.0 for energy between 10^{16} and 10^{18} eV (Figure 10.8). For energy above 10^{19} eV, the flux is even smaller. When the energy level is above 10^{18} eV, the cosmic ray flux is only 1 particle per square kilometer per year. Even so, the flux of cosmic rays is still 1,000 times that of the gamma rays.

Cosmic ray detection is similar to gamma ray detection. The detection can be achieved directly through their interaction with material or indirectly through secondary particles which include gamma rays, mesons, and neutrinos, from an air shower or other interactions.

Space cosmic ray detectors with limited collecting areas are mostly working at relatively lower energy bands. Almost all telescopes used for gamma ray detection can also be used for cosmic ray detection. However, in the UHE and EHE regions, cosmic ray flux itself becomes so small that the detectors require an extremely large collecting area for the detection of the secondary particles. Therefore, these UHE detectors are called cosmic ray telescopes, and are not

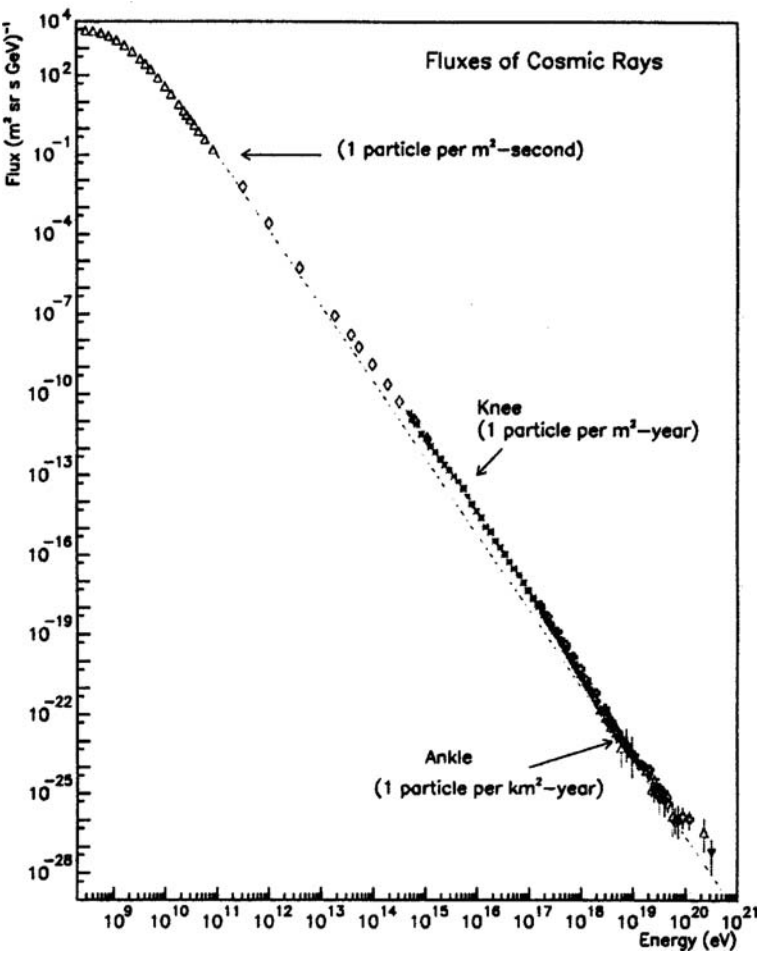


Fig. 10.8. The spectrum of high energy cosmic rays (Springer, 2000).

named gamma ray facilities. In some cases, cosmic ray telescopes are also used to observe neutrinos. In this case, the information recorded should come from the earth center direction, since only neutrinos can penetrate the earth without interaction with materials while cosmic rays will be blocked by the earth's shell. Indirect cosmic ray detectors are mostly EAS arrays and ground or space fluorescence telescopes.

Major observations of cosmic rays are through the Cherenkov air shower effect. When a high-energy cosmic ray particle enters the atmosphere it loses its energy via interactions with the nuclei of the air. At high energy level, these interactions create secondary particles, mostly pi-mesons (pions). These new particles go on to create more particles. This multiplication process is known as

particle cascade. Neutral pions produced will decay quickly into two gamma rays. The gamma rays from the neutral pions may also create new particles, electron/positron, by the pair-creation process. Electrons/positrons in turn may produce more gamma rays by the bremsstrahlung mechanism (Figure 10.9). Charged pions also decay but after a longer time. Therefore, some of the pions may collide with yet another nucleus of the air before decaying, which will react into a muon and a neutrino. The particle cascade looks like a pancake of particles traveling through the atmosphere at a speed near the speed of light in a vacuum but faster than the speed of light in air. This process continues until the average energy per particle drops below about 80 MeV. At this point, the interactions lead to the absorption of particles and the cascade begins to die. The altitude absorption beginning is known as the shower maximum. Cosmic rays with more energy will have their shower maximum at a lower attitude. At this time, though the number of particles in the pancake may be decreasing, the size of the pancake always grows as the interactions cause the particles to diffuse away from each other. When the pancake reaches the ground, it is roughly a few hundred meters or larger across and 1 to 2 meters thick.

The main difference between air showers from gamma rays and those from cosmic rays is the components of the air shower pancake. A gamma-ray air shower pancake will contain electrons/positrons and gamma rays, while the cosmic ray air shower pancake will also contain muons, neutrinos, and hadrons (protons, neutrons, and pions). The muons are positively charged and the neutrinos are neutral without charge. The muons can be detected by a high-voltage Geiger-Muller detector, silicon detector, or Cherenkov fluorescence detector. The neutrinos can be detected only by Cherenkov fluorescence detectors. The number of particles left in the pancake depends upon the energy of the primary cosmic ray, the observing altitude, and fluctuations in the shower development. The cosmic rays will hardly ever hit the ground but will collide (interact) with a nucleus of the air, usually at tens of kilometers altitude.

Since the flux of cosmic rays is about 1,000 times that of gamma rays in the same energy level, relatively lower energy cosmic rays can be observed through gamma ray facilities. On the ground, the ACTs are important tools for HE cosmic ray detection. However, as the energy level increases, the flux becomes extremely small. The cosmic ray telescopes in these regions should have a much larger area coverage and wider field of view. This requires dedicated cosmic ray telescopes instead of using gamma ray ones at UHE (100 TeV–100 PeV) or EHE (100 PeV–100 EeV) energy levels. These facilities are EAS arrays and fluorescence telescopes. Sometimes, they are used together as one telescope facility to provide redundance, coincidence, and timing check. In some references, names such as extremely high energy cosmic ray (EHECR) and extreme energy cosmic ray (EECR) are used. They are those cosmic rays with energy above 1 EeV or above 50 EeV, respectively. At these energy levels, the flux can be only 1 event per square kilometer per century. The detection will require special wide field of view space-based fluorescence detectors.

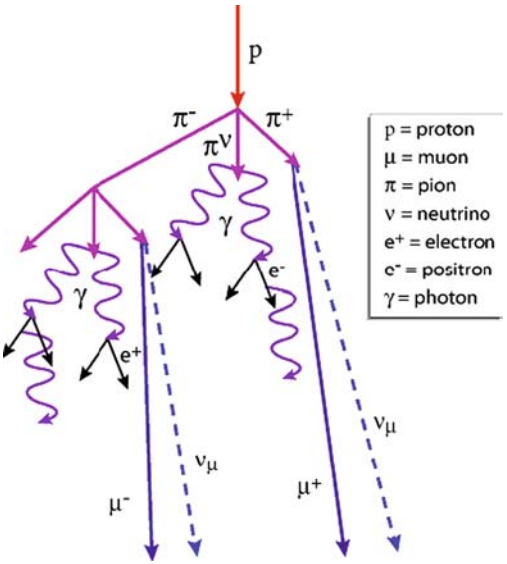


Fig. 10.9. Cosmic ray air shower composition.

10.2.2 Cosmic Ray EAS Array Telescopes

Cosmic ray extended air shower array (EAS array) facilities are mostly used in the UHE (100 TeV–100 PeV) regions. They are either at a high altitude or cover a large area on the ground. A few detectors are even located underground. At a high altitude, the required collecting area is smaller. The detectors or trackers can be emulsion chamber, resistive plate chamber, drift chamber, streamer tube detector, Geiger tube detector, or a variety of position-sensitive devices as used in particle physics. Some of these detectors allow the measurement of the particle direction. Some scintillator detectors are installed in pairs so that when both sensors indicate a response, the signal is recorded. One type of scintillator fiber hodoscope is formed by two layers of scintillator-doped polystyrene fibers: one in the x-direction and the other in the y-direction. The fibers are coated with black ink to optically isolate individual fibers. This provides accurate location of cosmic ray particle impact. High-altitude cosmic ray telescopes include ARGO, YangBaJing (4,300 m above sea level, in Tibet, China), INCA (Investigation on Cosmic Anomalies), and Chacaltaya Laboratory (5,220 m above sea level, near La Paz, Bolivia). The YangBaJing detectors are resistive plate chambers and the INCA uses emulsion chambers, which are multi-layered sandwiches of nuclear emulsion plates and lead plates.

The ground EAS array facilities usually involve larger collecting area and low cost water, ice, or other Cherenkov scintillation counters with PMTs. They allow the measurement of time of arrival with high accuracy. The Akeno Giant

Air Shower Array (AGASA) is a very large area surface array. It covers 100 km^2 and consists of 111 surface detectors (2.2 m^2, 5 cm thick scintillator) and 27 muon detectors (2.8–10 m^2, Fe/concrete absorber).

During a cosmic ray air shower, hundreds of particles arrive at shower maximum altitude per second but extensive air showers in extreme energy range are less common. Therefore, signal coincidence of several particle detectors is required. When looking for small showers with perhaps a few thousand particles, tens or hundreds of detectors separated by 10–30 m are usually used. For the much less frequent very large showers with billions of particles, the detectors can be placed at separations of about one kilometer.

The Pierre Auger Observatory is the largest ground-based cosmic ray telescope. It is located both in Pampa Amarilla, Mendoza, Argentina and in southeastern Colorado, US. This is an Extended Air Shower array and each site covers an area of 300 km^2. The determination of the energy of the cosmic ray is performed with both a surface EAS array and an atmosphere fluorescence detector array. The design of an atmosphere fluorescence detector will be discussed in the next section. The surface detectors consist of a triangular array of 1,600 water tanks spaced at 1.5 km. These water Cherenkov detectors have a cross sectional area of 10 m^2. They are 1.2 m deep and contain 12 tons of pure water. Within each water tank, three groups of PMTs are installed to detect the faint fluorescent light path. The attenuation length of the light in water is usually 2–3 m, which limits the size of the water tank with a minimum number of PMTs. A larger size tank requires more PMTs inside. In these detectors, there are three 9-inch photomultipliers facing downward on the upper surface of each tank. To avoid false images, these tanks have their inner walls of diffusively reflective material, called Tyvek, which is made up of overlapping polyethylene microfibers randomly oriented. The surface area is fully efficient for cosmic rays of energy $\geq 10^{19}$ eV with a zenith angle up to 60°.

MARCO (Monopole Astrophysics and Cosmic Ray Observatory) is another large area, underground experiment for rare components of cosmic rays (monopole, nuclearites as examples) as well as high-energy neutrinos. MACRO is located in Hall B of Gran Sasso National Laboratory (LNGS), located between the towns of L'Aquila and Teramo, about 120 km from Rome, Italy.

10.2.3 Cosmic Ray Fluorescence Detectors

Water Cherenkov scintillation counters with PMTs are seriously limited by attenuation of faint fluorescent light in water. However, the attenuation length of faint fluorescent light in the atmosphere is much longer, about 12 km. Therefore, the bluish fluorescent light (wavelength from 200 to 450 nm) showers can be detected at even greater distances. The light is emitted by nitrogen molecules when charged particles are passing near-by. Imaging devices (wide field of view

optical telescopes with photomultipliers as cameras) can be used for determining the track of air showers through the atmosphere. If an extremely wide field of view optical telescope is used, it can cover a much larger equivalent area. This special wide field optical telescope is called a cosmic ray fluorescence telescope or detector. Cosmic ray fluorescence detectors are mostly used in EHE (100 PeV–100 EeV) regions.

There are a number of designs for the cosmic ray fluorescence detectors. The simplest is a spherical mirror with PMTs at its focal plane. For a one-degree image spot size, this detector has a field of view of $15\times15^{\circ}$. This optical system is used for the HiRes (High Resolution Fly's Eye) project in Utah. A new version of this design is to put a stop at the curvature center of the mirror resulting in a field of view of $30\times 30^{\circ}$. The Pierre Auger Observatory's fluorescence detectors use this new design. To avoid the vignetting effect, the detector has a stop diameter of 1.7 m, while the mirror size is 3.5×3.5 m^2.

At the Pierre Auger Observatory, 30 units of fluorescence detectors are used alongside the EAS array. Each unit is a Schmidt camera without a corrector; it covers from 1.7 to 30.3° in elevation and 30° in azimuth. The camera contains 440 40 mm hexagonal photomultipliers in its focus. Figure 10.10 shows the design of this new version of fluorescence detector. This detector will have an image spot size of 1.5° when the stop size is 2.2 m. However, an improved new version of this Schmidt-type fluorescence detector involves one or two spherical ring correctors as shown in Figure 10.11. The improved new design with a stop size of 2.2 m can reach an image spot size of 0.55°.

In the HiRes project, two fluorescence detector arrays are used. They are separated at a distance of 12.6 km. If an image from one detector is used, the shower's plane can be constructed. If two images from two detectors are used from the array, the direction of the shower's axis can be determined, which is a

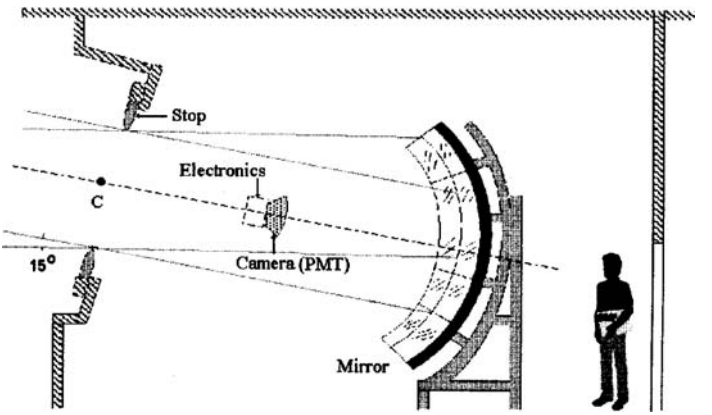


Fig. 10.10. Wide-field Schmidt camera without corrector (Cordero, 2000).

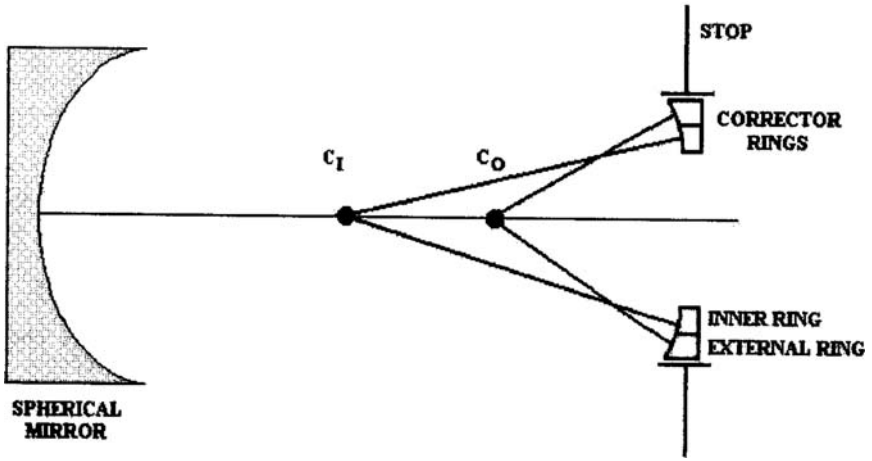


Fig. 10.11. Spherical mirror with ring corrector (Cordero et al, 2000).

line formed by the two planes. This technique is referred to as the stereo reconstruction technique.

A future Telescope Array (TA) project will have ten fluorescence detector stations. In each of these stations, two layers of 40 fluorescence detectors will cover 3–34° in elevation and all angles in azimuth. The area covered will be 30 times larger than AGASA.

For an even greater area of coverage, a cosmic ray fluorescence detector can be sent into orbit. The planned Extreme Universe Space Observatory (EUSO) project of ESO involves a detector with a field of view of 60° at an orbit of 380 km above the earth. The detector with a diameter of 3.5–4 m may be made of a double-sided, double Fresnel plate design. The pixel size of the image is about 1 km on ground. This great telescope will make it possible to catch very rare EECR (Extremely Energetic Cosmic Rays) particles. Another similar project proposed by Russia is called KLYPVE and is a 10 m diameter segmented Fresnel lens with a field of view of 15° on an earth orbit.

Cosmic ray telescopes can be used in neutrino detection if the detected fluorescence cones come from the earth direction instead of the sky direction. This removes all effects from gamma rays and cosmic rays. Therefore, fluorescence detectors discussed in this section can also serve as neutrino detectors.

In the radio regime, an Askaryan effect (Gorham, et al., 2007) can also be produced by particles traveling at a speed higher than the speed of light of the medium. The Askaryan effect describes a phenomenon, similar to the Cherenkov effect, whereby a particle traveling faster than the speed of light in a dense radio-transparent medium such as salt, ice, or the lunar regolith produces a shower of secondary charged particles which contain a charge anisotropy and thus emits a cone of coherent radiation in the radio or microwave part of the electromagnetic spectrum (0.2–1 GHz). So far the effect has been observed in silica sand, rock salt, and ice and is of primary interest in using bulk matter to detect ultra-high energy

particles (including neutrinos) between PeV (10^{15} eV) to ZeV (10^{21} eV). The effect is named after its postulator, Russian physicist Gurgen Askaryan. In ice, the cone angle of the effect is 53° and in rock salt it is 66°. The RF (radio frequency) Cherenkov cone can propagate through a solid and refracts into air space. The polarization of the radio signal is consistent with the Cherenkov effect, so that the tracking of the particle source is possible. The radio signal produced can be detected by earth- or space-based antennas, also named radio fluorescence antennas. These antennas can be underground, on the ground surface, or in orbit.

10.2.4 Magnetic Spectrometer Detectors

Traditional particle detection techniques in space have an energy limit of a few hundreds of MeV. Beyond this, the total energy measurement in space becomes impractical. Space-based magnetic spectrometers as an alternative can provide particle detection between several hundreds MeV to several TeV. Just as visible light of different colors will bend differently in a prism, the paths of charged particles of different momentum (or energy) will deflect differently in a magnetic field. In electromagnetism, the definition of momentum is different from classical Newtonian mechanics. It includes two terms: a mass velocity term and charge electromagnetic vector potential term.

$$\mathbf{p} = m\mathbf{V} + q\mathbf{A} \tag{10.19}$$

where m is the mass, V the velocity, q the charge, and A the electromagnetic vector potential. In a magnetic field, $\mathbf{B} = \nabla \times \mathbf{A}$ represents the field vector. For high-speed charged particles, the first term is very small. The rate of change of this momentum is the force vector of the particle in a magnetic field:

$$d\mathbf{p}/dt = q\mathbf{V} \times \mathbf{B} \tag{10.20}$$

The absolute value of the rate of change is related to its angular velocity $d\theta/dt$ or the linear velocity ds/dt, and the curvature of the trace ρ as:

$$\frac{d|\mathbf{p}|}{dt} = |\mathbf{p}|\frac{d\theta}{dt} = \frac{|\mathbf{p}|}{\rho}\frac{ds}{dt} \tag{10.21}$$

If the magnetic field is perpendicular to the particle velocity, the force is:

$$d\mathbf{p}/dt = q|\mathbf{B}|\frac{ds}{dt} \tag{10.22}$$

Then the momentum of the particle is related to the curvature of the motion:

$$\mathbf{p} = q\mathbf{B}\rho \tag{10.23}$$

In a magnetic spectrometer, there is a silicon tracker detector which consists of a number of detector layers. Some of these layers are inside the magnetic field for measuring charged particle bending. A time-of-flight system consisting of two double planes of scintillator counters, one in front of the magnetic field and the other behind, is used to record the time of the event to a time precision of 120 ps. By measuring the charge and the curvature of motion in a magnetic field, the particle momentum and the sign of the charge can be determined. This is the principle of the modern magnetic spectrometer for cosmic ray observation.

Since these spectrometers can resolve particle energy and charge, they can discover anti-matter from outer space. Anti-particles like positrons, antiprotons, and antideuterons are produced in the annihilation of neutralinos, which is one type of dark matter. Therefore, magnetic spectrometers can also serve as anti-matter detectors and indirectly as dark matter detectors.

The most important magnetic spectrometer is the Alpha Magnetic Spectrometer (AMS), which is an international space project. The experiment includes two instruments: AMS-1 and AMS-2. AMS-1 was a short period observation carried on a space shuttle. To avoid the influence of the unwanted orbit magnetic field, the magnetic field of AMS-1 was thoroughly blocked using neodymium-ferrous-boron magnetic materials. The observation lasted about 10 days in 1998. AMS-2 is a three-year experiment on the International Space Station, beginning in 2013 or 2014 as the instrument is already completed final assembly and experimental calibration. The magnetic field component of this magnetic spectrometer is made of a superconductor material. Besides a silicon tracker detector and the time-of-flight system, AMS-2 also has a transition radiation detector for detecting the transition radiation light by ultra-relativistic particles (for definition of relativistic particles, see Section 10.3.1). The transition radiation detector is a set of polypropylene fiber radiators with 5,248 straw tubes operated at a high voltage with a mixture of Xe and CO_2. AMS-2 also has a ring imaging Cherenkov counter and an electromagnetic calorimeter. The Cherenkov counter measures the velocity of relativistic particles moving away from its Cherenkov cone opening. The Cherenkov radiator consists of a set of aerogel and NaF tiles. In the detection plane, there are 680 multi-anode photomultipliers. The calorimeter is composed of layers of lead foils with glued scintillating fibers.

10.3 Dark Matter Detectors

10.3.1 Cold and Hot Dark Matter

Early in the 1970s, astronomers measured the rotation curves of some galaxies through redshifts of the H-alpha line along their radii (Rubin and Ford, 1970). These measured curves are quite flat from the galaxy center towards the edge. If most mass were concentrated in the galaxy center as we see from the visible band, the velocity should change inversely with the square root of the radius. This unusual phenomenon can only be explained by the existence of dark matter

in the universe. In astrophysics, dark matter is that matter that does not emit or reflect enough electromagnetic radiation to be detected directly, but whose presence may be inferred from its gravitational effects on the visible matter. Nowadays, astronomers, especially cosmologists, are convinced from indirect observation, the detection of major cosmology constants, the Big Bang theory, and the general relativity theory that only about 4% of the content of the Universe is in common luminous matter, about 24% belongs to dark matter, and the rest is dark energy. Dark energy is an even stranger component of the universe which keeps the universe from expanding.

The standard model of particles includes two families of elementary particles, i.e. leptons and hadrons, where "leptos" means thin or light and "hadros" means strong. Particle leptons are so-called "nonstrongly interacting" particles. The particle hadrons are again divided into mesons ("mesos": middle) and baryons ("baryos": heavy). The visible universe is only made up of baryons, including electrons, neutrons, and protons. Therefore, dark matter includes both nondetectable baryonic matter and other nonbaryonic matter.

The nonbaryonic matter can be further divided into relativistic and nonrelativistic particles. Particles with their velocity close to the speed of light are relativistic or "hot," and those with lower velocities are nonrelativistic or "cold." In some articles, warm dark matter refers to particles in between. Warm dark matter is one part of the cold dark matter. The search for dark matter includes the search for hard-to-detect baryonic matter, hot dark matter, and cold dark matter. Unfortunately, to date, the only known hot dark matter particles are neutrinos. All other dark matter particles are deduced only from theoretical modeling and still a mystery to astronomers.

Baryonic dark matter includes the unseen gas clouds, matter contained in planets, "failed" stars, brown dwarfs, or primordial black holes. This hidden galactic baryonic dark matter is called MAssive Compact Halo Objects (MACHOs), occurring mostly in the galactic halos. Gravitational wave telescopes are tools for detecting some of the hard-to-detect baryonic dark matter. The rest of the dark matter is nonbaryonic dark matter. The ordinary neutrino with minuscule mass is the only known nonbaryonic hot dark matter. Neutrino detection is discussed in the following two sections. However, ordinary neutrinos make only a small contribution to the density of nonbaryonic dark matter. Therefore, the detection of cold nonbaryonic dark matter is also necessary.

Other hypothetical cold dark matter candidates are axions and the Lightest Supersymmetric Particle (LSP). The LSPs include neutralinos, gravitinos, and sneutrinos. But there is plenty of room for other species, which are generally called Weakly Interacting Massive Particles (WIMPs). None of these are part of the standard model of particle physics, but they can arise in the extensions of the standard model. Many supersymmetric models naturally give rise to stable WIMPs in the form of neutralinos. The efforts to detect these cold dark particles are discussed in Section 10.3.4.

10.3.2 Detection of Neutrinos

The neutrino is one of the fundamental particles which make up the universe. It is one type of missing dark mass. It is also one of the least understood particles in the universe.

Neutrinos were predicted by Wolfgang Pauli as a missing part in beta decay of a neutron. Neutrinos carry some energy. They are similar to the more familiar electron, with one crucial difference: neutrinos do not carry an electrical charge. Because neutrinos are electrically neutral, they are not affected by the electromagnetic forces which act on electrons or cosmic rays. Neutrinos are affected only by a "weak" sub-atomic force of a much shorter range than electromagnetism, and are therefore able to pass through great distances in matter without being affected by it. The detection of neutrinos is, therefore, difficult, requiring large detection volumes or high intensity of neutrino beams. Neutrinos may carry information concerning the nuclear reaction that takes place at the core of a star. For example, neutrinos generated near the center of the sun could be detected on the earth in about eight minutes, while it takes around a million years for photons generated at the center of the sun to reach the earth. The photons inside the sun escape the sun via a very slow diffusion process.

Each type or "flavor" of neutrino is related to a charged particle (which gives the corresponding neutrino its name). Three types of neutrinos are known; the "electron neutrino" associated with the electron, and two other neutrinos associated with heavier versions of the electron called muon and tau. There is strong evidence that no additional neutrinos exist, unless their properties are unexpectedly very different from the known types.

Neutrinos are linked with dark matter and anti-matter in the universe. Lighter neutrinos were discovered in high energy accelerators. However, did any super-heavy counterparts to the lightweight neutrinos exist in the early universe? It is impossible to generate these hefty particles in accelerators which have limited energy levels. The existence of the missing mass relies on careful observation of neutrinos. Neutrinos may have their own anti-particles. If this is the case, the anti-neutrino emitted in a "double beta decay" may be immediately absorbed by the other neutrino. Verification of this interaction will be important for matter–antimatter asymmetry theory.

In particle physics, cross section or cross section area is a term for the effective area of collision. The numerical value of it is chosen so that, if the bombarding particle hits a circular area of this size perpendicular to its path and centered at the target nucleus or particle, the given reaction occurs; and, if it misses the area; the reaction does not occur. The cross section is usually different from the geometrical area, it depends on the energy. Boron (a chemical element with atomic number 5), when bombarded by neutrons with a speed of 1000 m/s, has a cross section of 1.2×10^{-22} cm^2. However, the cross section area when bombarded by solar neutrinos is very small, about 1.06×10^{-42} cm^2. The atomic radius is usually in the range of femtometers.

There are various ways to detect neutrinos. One is through a radio chemical material, such as ^{37}Cl, ^{127}I, or ^{71}Ga. In the past, a tank filled with a cleaning fluid, C_2Cl_4, was used for neutrino detection. The reaction between neutrinos and radio chemical material such as ^{37}Cl is;

$$v_e + {}^{37}Cl \rightarrow {}^{37}Ar + e^- \tag{10.24}$$

where v_e is an electron neutrino. By the detection of rare isotope ^{37}Ar, the neutrinos are detected. The first neutrino detection using this method was done in the Homestead mine, South Dakota. Since the neutrinos have small cross section, a large volume of fluid is required for detection. The high threshold for this reaction is at the energy of 0.814 MeV, which permits the observation of only a part of the solar neutrinos.

Another very efficient solar neutrino absorption process uses a gallium detector. The reaction is:

$$v_e + {}^{71}Ga \rightarrow {}^{71}Ge + e^- \tag{10.25}$$

This reaction has a much lower energy threshold of 0.233 MeV, so it is sensitive to a much larger fraction of the total spectrum of solar neutrinos (between 0.1 and 10 MeV). The GALLEX experiment in Gran Sasso National Laboratory, located between the towns of L'Aquila and Teramo, about 120 km from Rome, Italy, used 30 tons of gallium in an aqueous solution of gallium chloride and hydrochloric acid. The detection was discontinued in 1991. Another Russian SAGE experiment uses 55 tons of metallic gallium in 3,000 m^3 liquid at the Baksan Neutrino Observatory, situated in Prielbrusye, the Caucasus.

One type of neutrino or anti-neutrino detector involves a mixture of water and cadmium chloride. The anti-neutrino interacts with a proton of the target matter, giving a positron and a neutron. The positron annihilates an electron of the surrounding material, giving two simultaneous photons and the neutron that slows down until it is eventually captured by a cadmium nucleus, implying the emission of photons some 15 microseconds after those of the positron annihilation. These photons can be detected and the 15 microseconds identify the neutrino's interaction.

The simplest way to detect neutrinos is through a water tank fitted with PMTs. The principle of this type of detection is called elastic scattering. In the process, neutrinos scatter with matter and gain or lose energy from the collision partner without any additional matter being created or destroyed. The process is:

$$v + e^- \rightarrow v + e^- \tag{10.26}$$

The recoiling electrons emit Cherenkov light, detectable with PMTs that can be recorded. Pure water or even pure ice makes this method more attractive as the cost of water or ice is very low. One disadvantage of this method is a relatively high threshold for the neutrino detection. The threshold is between 6.5 and 9 MeV. Therefore, only the high-energy portion of neutrinos can be detected by using the water or ice Cherenkov method. This reduces the detectable neutrino flux from the sun by a factor of 10^{-4}. The other disadvantage of this method is the difficulty in distinguishing gamma ray signals generated by impurities in the water (or ice) or induced by cosmic rays. To remove the cosmic ray effect, the optimum method is to record the Cherenkov light coming from the earth's center direction instead of from the sky direction.

The number of interactions with neutrinos within a volume of water in a time interval is given by:

$$N_{v\cdot e} = \phi_v \Delta t \sigma_{v\cdot e} N_{t\arg et} = \phi_v \Delta t \sigma_{v\cdot e} \frac{10 M N_A}{A} \tag{10.27}$$

where ϕ_v is the neutrino flux, Δt the time interval, $\sigma_{v\cdot e}$ the cross section for interaction, $N_{t\arg et}$ the number of target electrons, M the water weight in kg, $N_A = 6.022 \times 10^{26}\,\text{kmol}^{-1}$ the Avogadro constant (or number), $A = 18\,\text{kg/kmol}$ atomic weight of water, and the factor of 10 means there are ten electrons for each H_2O molecule. The cross section for interaction is proportional to the neutrino's energy level. The cross section for neutrino interaction with ordinary material is approximately $5 \times 10^{-32}(E_v/\text{MeV})^2\text{cm}^2$, which is still very small. If the neutrino flux is $6.5 \times 10^6\text{cm}^{-2}\text{s}^{-1}$ and the time interval is one day, then the water weight required for the detection of one event is about $0.5 \times 10^6\text{kg}$. The cross section for the neutrino interaction rises linearly with neutrino energy. For neutrinos with energy above 5 MeV, the cross-section for the reaction is given by (Bergstrom, 2004):

$$\sigma_{v\cdot e} = C_x 9.5 \times 10^{-45} \left(\frac{E_v}{1\,\text{MeV}} \right) \text{cm}^2 \tag{10.28}$$

where C_x is a constant related with neutrino types (for electron neutrinos, $C_x = 1$, for other types of neutrinos $C_x = 1/6.2$) and E_v the energy of neutrinos.

On the basis of the above theory, water or ice Cherenkov neutrino telescopes are widely used. Water and ice have a suitable index of refraction and low absorption in UV and optical regions. The detector typically consists of an array of PMTs distributed in the medium. The PMTs have a good time resolution (~1 ns). The pattern of hit and the arrival time are used to fit the direction of the particle. To avoid the bombing produced by muons from a cosmic ray air shower, neutrino telescopes are generally built deep underground or under the sea. Muons are charged, and while they do not penetrate matter nearly as effectively as neutrinos, they do sometimes have sufficient energy to reach considerable depths. Even at about 1 km underground, approximately

4×10^{-8} such particles pass through rock every square centimeter, every second (Robinson, 2003). However, by going deep underground, we ensure that 99.9% of these muons are filtered out by the rock above. Further improvement is given by recording neutrinos coming from the other side of the earth. This will remove all the effects from cosmic rays. Both optical Cherenkov and radio Arskaryan effects can be used for neutrino detection in caves deep underground.

As mentioned in the cosmic ray telescope sections, neutrinos can also be detected by air fluorescence detectors or radio antenna detectors on the ground or in orbit used for cosmic ray observations. The air fluorescence detectors can be ground based or space based as in the cosmic ray detection. The radio antenna can be existing antennas or space-based antennas. If a neutrino particle hits the rock at the moon's edge, the radio cone signal produced could be received by a group of existing radio antennas. If a neutrino particle hits the rock from the earth's center direction, the signal can be picked up by space-based antennas. If a target of an ice cube underground is hit by the particle, the radio signal can be received by an array of horns located within the cone angle after it refracts into the air (Gorham, 2007). In ice, the cone angle of this effect is 53° and in rock salt it is 66°. If radio detectors are arranged inside a salt dome, the signal received can also track the source particle direction as well. The design of the detector is similar to Cherenkov telescopes. In neutrino detection, false alarms are unavoidable for many reasons including the detector itself. Therefore, coincidences between independent detectors will add great confidence to the detection. As a summary, neutrino detector types are listed in Table 10.1.

10.3.3 Status of Neutrino Telescopes

The first experiment to detect electron neutrinos was carried out in 1969 by using a liquid Chlorine target in the Homestead mine. It contained 615 tons of industrial solvent based on chlorine. Afterwards, the GALLEX project used 30 tons of gallium in an aqueous solution of gallium chloride and hydrochloric acid. GALLEX was discontinued in 1991. Another Russian SAGE experiment used 55 tons of metallic gallium at the Baksan Neutrino Observatory. In 2002, the

Table 10.1. Neutrino detector types

Detector type	Material
Scintillator	C, H
Water Cherenkov	H_2O
Heavy water	D_2O
Liquid argon	Ar
High Z/neutron	NaCl, Pb, Fe
Radio chemical	^{37}Cl, ^{127}I, ^{71}Ga
Radio Askaryan effect.	Ice, rock, or salt

Sudbury Neutrino Observatory Laboratory (SNOLAB) near Sudbury Ontario, Canada, built a 12 m diameter acrylic vessel located 2,200 m underground using 1,000 tons of heavy water, which was on loan from Atomic Energy of Canada Limited (AECL).

However, most new high-energy neutrino detectors use water or ice as the detecting medium. The major water-type detector facilities include DUMAND (Deep Underwater Muon And Neutrino Detector) in Hawaii, NESTOR (NEutrinos from Supernova and TeV sources Ocean Range) in the Mediterranean near Greece, NT-36 (in the future, NT-200) with 36 PMT groups in Lake Baikal as a Russian, German, and Hungarian project, KAMIOKANDE (Kamioka Nucleon Decay Experiment) near the city of Hida in Japan, and PAN (Particle Astrophysics in Norrland) in northern Sweden. DUMAND2 will extend the detecting area of DUMAND to an area of 20,000 m^2. NESTOR is also an underwater 100,000 m^2 large area neutrino telescope with many detector groups. In each group, there are ten phototubes (15 inch photocathodes), six of which are located in the radius of 7 m in the first layer and four in the second layer of 3.5 m lower down. A tower-type structure is made for 12 groups of detectors in the vertical direction. There are many tower structures to cover the targeted detecting area. All the groups are located down to a depth of 4,100 m underwater. Inside sea water, the attenuation length of the Cherenkov light is about 2 m. KAMIOKANDE is another major project with a tank of 2,140 tons of water underground. A very important ice detector facility is AMANDA (Antarctic Muon And Neutrino Detector Array), which is 1,400 m under the ice and has a total of 4,800 PMTs. This project is also known as IceCube. The Antarctic ice only becomes transparent to Cherenkov light at depths greater than 1,400 m. The solid pure ice has a maximum attenuation length of 24 m. The neutrino produced Cherenkov light inside the ice has a conic angle of 45°.

There are still other types of neutrino detectors. Among these, the MUNU (means magnetic moment and neutrino from the Greek) detector uses compressed CF_4 gas with 5 bars pressure as a medium. A neutrino interacts with an electron from the gas and the electron recoils and ionizes the gas, leaving a track behind it.

The experiments relying on the radio Arskaryan effect include Radio Ice Cherenkov Experiment (RICE), Fast On-orbit Recording of Transient Events (FORTE), Goldstone Lunar Ultra-high energy neutrino Experiment (GLUE), Saltdome Shower Array (SalSA), and Antarctic Impulsive Transient Antenna (ANITA).

Besides underground facilities for neutrino detection, ground- or space-based cosmic ray telescopes can also be used for neutrino detection. Using these telescopes for neutrino detection, the conic fluorescence light detected should not come from the sky, but from the earth center, which removes the effects from both cosmic rays and gamma rays. Therefore, neutrino telescopes may include some cosmic ray facilities.

10.3.4 Detection of Cold Dark Matter

Direct detection of cold dark matter looks for the recoil energy deposited by elastic scattering of particles off the nucleus of underground low-background detectors. Indirect detection looks for the products of dark matter annihilations. One of the indirect detections of the dark particle neutralinos is to detect the neutrinos decayed from the annihilation products after neutralinos were trapped inside the core of bodies, such as the sun or the galaxy center.

10.3.4.1 Cryogenic Dark Matter Detector

When a cold dark particle directly hits a nucleus of an absorber (Ge, Si, sapphire, LiF, AlO_2, CaW, etc), the momentum is conserved. Therefore, the material has a very small increase in energy (possibly below 1 keV) or temperature. The energy increase depends on the relative speed of the dark particle v, the hitting angle θ, and the masses of the dark particle and the nucleus m_x and m_N:

$$E = m_N v^2 (1 - \cos\theta) \left[\frac{m_x}{m_N + m_x}\right]^2 \tag{10.29}$$

This tiny energy increase (∼100 eV) can be measured by very sensitive cryogenic detectors such as a Transition Edge Sensor (TES) or a thermistor, both responding to a change in temperature by a change in resistance (Cheng, 2006). One semiconducting thermistor is the Neutron Transmuted Doped (NTD) germanium thermistor, which is a heavily doped semiconductor slightly below the metal insulator transition. One type of TES used in the dark matter search is the Z-sensitive Ionization and Phononbased detector (ZIP). The absorber size, its property, and the particle's cross-section area determine the probability of hitting the detector.

The transition edge sensor is a superconductor instrument. For any superconductor, there exists a critical or transition temperature. Below this temperature, the material is a superconductor with a zero resistance and above the temperature; the material is a normal one with a nonzero resistance. In the transition temperature, the change of resistance is very large (Figure 10.12). Transition widths are typically in the order of millikelvin. The transition edge sensor is a very sensitive microcalorimeter using the transition property from a normal material to a superconductor in very low temperature. The relative sensitivity in the transition edge sensor is defined as:

$$\alpha = \frac{T}{R}\frac{dR}{dT} \tag{10.30}$$

The energy resolution of the sensor is:

$$\Delta E \approx 2.35\sqrt{4kT^2\frac{C}{\alpha}} \tag{10.31}$$

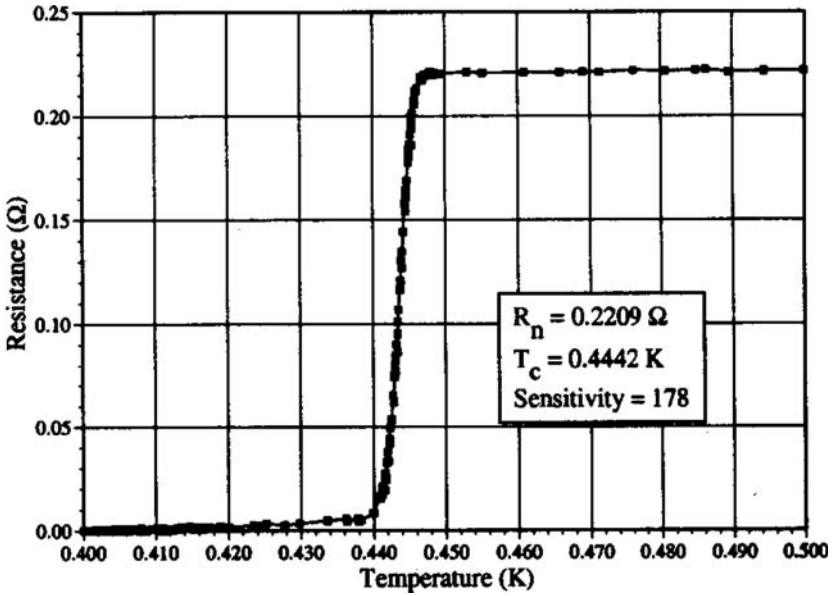


Fig. 10.12. The resistance transition curve of a molybdenum/gold bilayer.

where k is the Boltzmann's constant, T the temperature, C the heat capacity, and the factor of 2.35 is used to convert standard deviation into full-width at half maximum (FWHM) (just an easier way to measure the resolution of the devices). So for a good energy resolution, we want to operate the devices as cold as we can, and to have a low heat capacity and a high sensitivity.

The general arrangement of a transition edge sensor is shown in Figure 10.13. For dark matter detection, the absorber is some type of crystal material. The small current change of the sensor is picked out by a superconducting quantum interference device (SQUID), another extremely sensitive superconductor measuring device operating in very low temperature. The measuring system is a closed loop one.

In a normal operation, many factors can produce a small temperature increase in the sensor. Among these, cosmic ray and residual radioactivity from some materials or nearby rocks are important energy sources. The vibration of the device is also an energy source. Therefore, it is necessary to place the dark matter cryogenic detector deep inside mine caves and to shield it by heavy metals, such as copper and lead. The materials used for the device have to be pure and radiopoor enough to avoid any residual radioactivity. Rinsing with pure water and long time storing inside an underground cave can help in this aspect. The detector also needs a good vibration control as with the gravitational wave telescopes. Even so, redundant detection or other background discrimination techniques should be used for distinguishing electron recoils (Compton scattering) and nucleus recoils. In quantum mechanics a type of

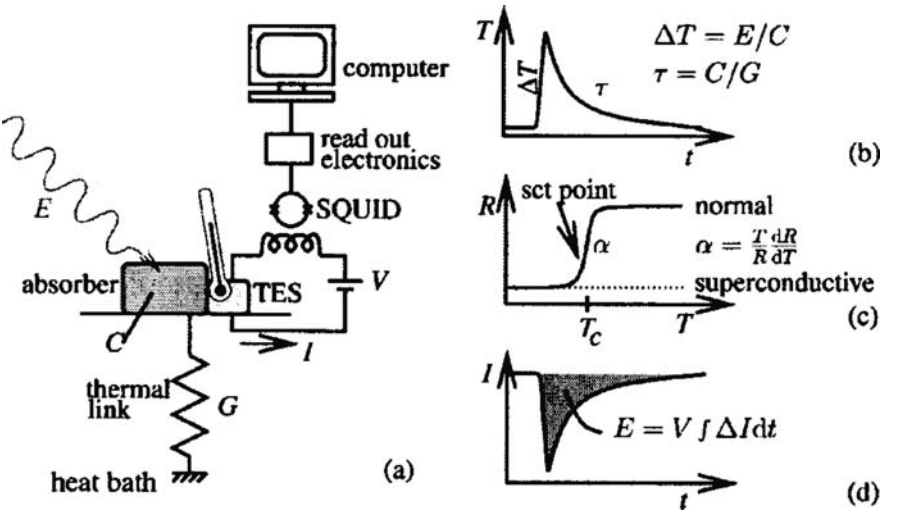


Fig. 10.13. Schematic of a typical transition edge sensor: (**a**) its diagram, (**b**) its temperature curve, (**c**) its resistance curve, and (**d**) its current curve (Bruijin, 2004).

vibrational motion, known as normal modes in classical mechanics, in which each part of a lattice oscillates with the same frequency is described as a particle of phonons. Therefore, it is necessary to decouple electrons from phonons in the measurement.

Another technique of background discrimination is to use the seasonal variation of the signal to find the cold dark matter trapped in our galaxy near the solar system. Since the earth is inclined to the direction of the sun's movement in our galaxy, a seasonal variation exists for relative speed between the heavy dark matter particles and the detector (Figure 10.14). This produces a seasonal variation of the detected signal.

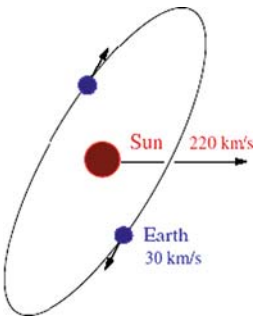


Fig. 10.14. Seasonal variation of relative velocity of cold dark matter trapped in the solar system.

This type of cryogenic dark matter thermister detector is also called a phononbased detector or phonon sensor. The typical temperature of the absorber kept in these detectors is about 20 to 40 millikelvins.

10.3.4.2 Scintillation Dark Matter and Resonant Cavity Detectors

A scintillation dark matter detector is similar to those used in gamma ray and neutrino detections. Scintillation dark matter detectors make use of another effect of the recoiling nucleus: ionization. The nucleus knocks some electrons off of surrounding atoms, resulting in excited ions known as excimers. These ions eventually recapture an electron and return to normal. In some materials, the process triggers the emission of light, called scintillation light. The light is captured by PMTs. The materials used in this type of detector include NaI, CsF_2, $CaWO_4$, BGO, liquid or gas xenon, and others. In some detectors, a layer of gas threaded by an electric field is used to amplify the scintillation light produced. The field accelerates the electrons that get kicked off by recoiling nuclei, thereby turning a handful of particles into an avalanche. In some facilities, the phonon detector and scintillation detector are combined in one for a better background discrimination.

In the search for axions with a mass range from 10^{-6} to 10^{-2} eV, a special method is to convert axions into microwave photons. This produces a type of resonant cavity detector. Under a strong magnetic field, if the resonant frequency of a cavity is equal to the axion energy, the axions can convert to microwave photons. Therefore, the search for axions becomes the detection of microwave photons. The detecting system includes two parts: a conversion cavity, which converts axions into photons under a strong magnetic field ($\sim$8 T), and a detection cavity to catch the photons. The photons are transferred to the detection cavity via a coupling hole. The detection cavity is set to be free from the magnetic field to avoid Zeeman splitting. The cavity is also kept at a low temperature of 1.3 K for background removal.

The search for cold dark matter is a new event in astronomy, which started in the 1990s. Major projects in this search include the UK Dark Matter Collaboration (UKDMC) formed by the University of Sheffield, Rutherford Appleton Laboratory, Imperial College of London, and the University of Edinburgh (liquid Xenon, Sodium Iodide (NaI) crystal scintillator), Dark Matter (DAMA) of Italy and China, Rare Object Search with Bolometers UndergrounD (ROSEBUD) of Spain, Project In CAnada to Search for Supersymmetric Objects (PICASSO) of Canada and the Czech Republic, Directional Recoil Identification From Tracks-I (DRIFT) of the UK and the US, Edelweiss of France and Russian, Cryogenic Dark Matter Search (CDMS) of the US, Cryogenic Rare Event Search with Superconducting Thermometers (CRESST) in Gran Sasso of Italy, and OTO Cosmo Observatory of Japan (Cline, 2000). Table 10.2 lists the details of these leading projects. So far, results of all dark matter experiments have been negative or inconclusive.

Table 10.2. Leading cold dark matter search projects (Cline, 2005)

Project	Location	Start year	Device	Material	Mass
UKDMC	UK	1997	Scintillation	NaI	5 kg
DAMA	Italy	1998	Scintillation	NaI	100 kg
ROSEBUD	Spain	1999	Cryogenic	AlO2	0.05 kg
PICASSO	Canada	2000	Liquid droplets	Freon	0.001
DRIFT	UK	2001	Ionization	Germanium	0.16 kg
Edelweiss	France	2001	Cryogenic	Liquid xenon	1.3 kg
ZEPLIN I	UK	2001	Scintillation	Liquid xenon	4 kg
CDMS II	US	2003	Cryogenic	Silicon, Ge	7 kg
ZEPLIN II	UK	2003	Scintillation	Liquid xenon	30 kg

References

Abramovici, A. et al., 1992, LIGO: The laser interferometer gravitational-wave observatory, Science, 256, 325.

Bergstrom, L. and Goobar, A., 2004, Cosmology and particle astrophysics, 2nd ed., Springer, Praxis, Berlin.

Bernardini, A. et al., 1999, Suspension of last stages for the mirrors of the Virgo interferometric gravitational wave antenna, Rev. Sci. Instrum., 70, 3463.

Bertotti, B., Carr, B. J., and Ree, M. J., 1983, Limits from timing of pulsars on cosmic gravitational wave background, Mon. Not. R. Astro. Soc., 203, 845–954.

Beyerdorf, P. T., Byer, R. L., and Fejer, M. M., 1999, The polarization Sagnac interferometer as a candidate configuration for an advanced detector, in Gravitational waves, ed. Meshkov, S., Third Edoardo Amaldi conference, Pasadena, AIP Proceeding. 523.

Bruijn, M. P. et al., 2004, Development of arrays of transition edge sensors for application in X-ray astronomy, Proceedings of 10th International Workshop on Low Temperature Detectors, eds. Flavio G., Nuclear Instruments and Methods in Physics Research, A 520, 443.

Cheng, J., 2006, The principles and applications of magnetism, Chinese Science and Technology Press, Beijing, in Chinese.

Cline, D. B.(editor), 2000, Sources and detection of dark matter and dark energy in the universe, Springer, New York.

Cordero-Davila, A., 2000, Optical design of the fluorescence detector telescope, in Observing ultrahigh energy cosmic rays from space and earth, eds. Salazar, H., Villasenor, L., and Zepeda, A., AIP Conf. Proc. 566.

Cordero-Davila, A. et al., 2000, Segmented spherical corrector ring 1: computer simulation, in Observing ultrahigh energy cosmic rays from space and earth, eds. Salazar, H., Villasenor, L., and Zepeda, A., AIP Conf. Proc. 566.

Cristinziani, M., 2002, Search for antimatter with the AMS cosmic ray detector, International Meeting on Fundamental Physics, IMFO, Spain.

De Michele, A., Weinstein, A., and Ugolini, D., 2001, The pre-stabilized laser for the LIGO Caltech 40 m interferometer: stability controls and characterization, LIGO document LIGO-T010159-00-R.

Giovannelli, F. and Sabau-Graziati, L., 2004, The impact of space experiments on our knowledge of the physics of the universe, Space Sci Rev, 12, Issue 1, p1–443.

Gorham, P. W. et al, 2007, Observations of the Askaryan effect in ice, Phys. Rev. Lett., 99, 171101.

Hughes, S. A. et al, 2001, New physics and astronomy with the new gravitational-wave observatories, Proceeding of Snowmass.

Hulse, R. A., 1975, A high sensitive pulsar search, Ph. D. thesis, Massachusetta University, Amherst.

Kuroda, K. et al, 2004, Large-scale Cryogenic Gravitational wave Telescope.

Maggiore, M., 1999, Gravitational wave experiments and early universe cosmology, Report IFUP-TH20/99, Universita di Pisa, Italy.

Meshkov, S., ed., 1999, Gravitational waves, Third Edoardo Amaldi conference, Pasadena, AIP Proceeding Vol 523.

National Research Council, USA, 2001, Astronomy & Astrophysics in the New Millennium.

Pizzella, G., 1997, Gravitational waves: the state of the art, in Conference of Proceedings 57, Frontier objects in astrophysics and particle physics, eds. Giovannelli, F. and Mannocchi, G., SIF, Bologna.

Plissi, M. V. et al., 1998, Aspects of the suspension system for GEO600, Rev. Sci. Instrum., 69 (8), 3055.

Pretzl, K., 2000, Cryogenic calorimeter in astro and particle physics, Nucl. Instrum. Meth. A 454, 114.

Robinson, M. et al., 2003, Measurement of muon flux at 1070 meters vertical depth in Boulby underground laboratory.

Rubin, V. and Ford, W. K., 1970, Rotation of the Andromeda nebula from spectroscopic survey of emission regions, Astrophy. J. 159, 379.

Salazar, H., Villasenor, L., and Zepeda, A. eds., 2000, Observing ultrahigh energy cosmic rays from space and earth, AIP Conference Proceedings, 566, New York.

Springer, R. W., 2000, Observing ultra high energy cosmic rays with the high resolution fly's eye detector, in Observing ultrahigh energy cosmic rays from space and earth, ed. Salazar, H., Villasenor, L., and Zepeda, A., AIP Conf. Proc. 566.

Taylor, J. H. and Weisberg, J. M., 1989, Further experimental tests of relativistic gravity using binary pulsars PSR 1913+16, Astrophys. J., 345, 434.

Willke, B. et al., 2006, Stabilized high power laser for advanced gravitational wave detectors, J. Phys.: Conf. Ser.. 32, 270–275.

Chapter 11
Review of Astronomical Telescopes

In this chapter, a general review for all electromagnetic and nonelectromagnetic wave or particle telescopes is provided. Detailed tables of various major ground- and space-based astronomical telescopes are presented. The atmosphere effect on electromagnetic radiations and the advantages of space electromagnetic wave observations are also discussed. The chapter also gives a brief summary of telescope-related space missions as well as the reconnaissance telescopes. These two sections are brief, but include almost all of the important space missions, the airborne reconnaissance telescopes, reconnaissance satellites, and ground telescopes for monitoring man-made near earth objects. From this chapter, readers will have a general picture of all the astronomical telescopes and their related applications.

11.1 Introduction

Two common characteristics that are shared among various astronomical telescopes are extremely high sensitivity, for detecting weak radiation and lower energy level, and very high precision, for quantifying the direction, size, and other properties of the emitting regions. The universe outside the earth was not physically accessible before the invention of spacecraft. Telescopes were the only tools for astronomers to study our universe. Through the captured waves or particles, astronomers acquired direct information of the universe. Now man-made spacecraft have landed on or flown by the moon and a few planets. Small samples and few details can be examined more closely. However, vast parts of the universe remain inaccessible to human beings. Thus, astronomical telescopes today remain extremely important in astronomy.

Even with the most advanced astronomical telescopes, we only reach radiation levels down to a low limit set by the instruments. Our knowledge of the universe is totally based on such limited information. To reach beyond these limits, astronomers demand ever increased sensitivity and increased precision. Therefore, new telescopes are colossal in size and extremely demanding in

J. Cheng, *The Principles of Astronomical Telescope Design*,
Astrophysics and Space Science Library 360, DOI 10.1007/b105475_11,

technology, so that the observation limit can be further extended. Because of this, astronomy has become a "big" science among scientific disciplines. After 400 years development, telescope project size in almost all wave/particle forms and all wavelengths has now passed the several billion US dollar threshold in cost. The telescope itself, with its close connection with physics, astronomy, optics, radio science, space science, technology, and many other fields, becomes a new established branch in astronomy. This branch is of vital importance for anyone who is devoting himself in these related fields. It also fills the gap between important physical theories and many practical applications.

Most astronomical telescopes today are those used at electromagnetic wavebands. A general review of electromagnetic waves and their interaction with the earth's atmosphere are provided in Section 11.2. Section 11.3 gives an overview of nonelectromagnetic wave or particle telescopes. Sections 11.4 and 11.5 are summaries of important ground-based and space-based astronomical telescopes. Sections 11.6 and 11.7 provide a brief history of space missions and reconnaissance telescopes.

11.2 Electromagnetic Wave and Atmosphere Transmission

Human astronomical observations began thousands of years ago in the narrow visible band of the electromagnetic spectrum, from which the optical telescope was invented in 1609. Now, astronomical telescopes cover all the spectral bands of electromagnetic waves. Early astronomical observation was started at sea level, but serious absorption, turbulence, and emission of the earth's atmosphere had forced observers in many wave bands to move gradually to higher and better terrain and eventually to realize observations on mountain tops, rockets, airplanes, balloons, and on spacecraft covering the entire electromagnetic spectrum.

Table 11.1 presents the electromagnetic spectrum from high to low frequencies. In the high frequency, the electromagnetic radiations are treated as particles with high energy and in low frequency, they are treated as waves. When frequency changes, the energy level ($E = hc/\lambda$) of the radiation also changes. The temperature of related astronomical processes which generate electromagnetic radiation is governed by Planck's blackbody radiation theory. Therefore, electromagnetic waves provide us important thermal information for the whole universe.

The earth's atmosphere imposes wavelength-dependent effects on the electromagnetic wave propagation; direct observations of different bands in the spectrum have to be performed at different altitudes above sea level. Extremely high energy gamma rays can reach an altitude of a few kilometers. High and low energy gamma rays, including part of hard X-rays, can reach a height of 40 km restricted from the electron/positron pair production and Compton effects. X-rays can reach a height between 70 and 100 km constrained by the photoelectric effect where the high energy photons collide with atmospheric molecules or atoms so that the energy of the photon is transferred to the electrons, ionizing the molecules or atoms. The ultraviolet photons reach a height between 50 and

Table 11.1. The electromagnetic spectrum (expanded from Davies, 1997)

Wavelengths (m) shorter than	Other units	Photon energy or frequency greater than	Usual name	Produced by temperatures in regions of (K)	Objects of interest
10^{-25}		80.6 EeV			
10^{-22}		80.6 PeV	EHE		Electron and positron annihilation
10^{-19}		80.6 TeV	UHE		
10^{-16}		80.6 GeV	Gamma ray		
10^{-13}		80.6 MeV	HE		Cosmic ray interactions with inter-stellar gas
10^{-12}		8.06 MeV		10^{8}	
10^{-11}		0.8 MeV	ME		
10^{-10}	0.1 nm	80.6 keV			Accretion disk in binary, hot gas in galaxy
			Hard X-ray	10^{7}	
10^{-9}	1 nm	8.06 keV			
10^{-8}	10 nm	0.8 keV	Soft X-ray	10^{6}	
10^{-7}	100 nm	80 eV	XUV/ EUV		White dwarf stars
			Far UV		
3×10^{-7}	200 nm		Ultraviolet	10^{5}	Flare stars O stars
4×10^{-7}	400 nm		Violet		
			Optical	10^{4}	B stars
7×10^{-7}	700 nm		Red		K, M stars
8×10^{-7}	0.8 μm			10^{3}	Circum-stellar dust

Table 11.1. (continued)

Wavelengths (m) shorter than	Other units	Photon energy or frequency greater than	Usual name	Produced by temperatures in regions of (K)	Objects of interest
10^{-5}	10 μm		Near infrared		Comets and asteroids
10^{-4}	100 μm	3 THz	Far infrared		
10^{-3}	1 mm	300 GHz	Millimeter	100	
10^{-2}	1 cm	30 GHz		10	Microwave background
10^{-1}	10 cm	3 GHz	Microwave	1	
1	1 m	300 MHz			
10	10 m	30 MHz	Radio wave		
10^2	100 m	3 MHz			Electrons Spiraling in magnet fields
10^3	1 km	300 kHz	Long wave		
10^4	10 km and greater	30 kHz	Very long wave		

100 km where the atmospheric molecules of oxygen, nitrogen, and ozone can be dissociated and ionized by the photons. The visible photon can reach the ground level, not being absorbed, but with scattering and phase changes. This forms the first atmospheric window in visible region.

Infrared radiation can reach a height of 5–10 km due to absorption from carbon dioxide, water vapor, and other molecules. Narrow atmospheric windows exist in the infrared, submillimeter, and millimeter regions at high and dry sites. It follows another important atmospheric window in the millimeter and radio wave regions. The long wavelength end of this radio window depends on the ionosphere which in turn is affected by solar activities. Long and very long radio waves can only reach a height of 90–500 km set by the ionosphere reflection.

The restrictions caused by the earth's atmosphere have forced astronomical observation to move to balloons, airplanes, rockets, and space orbits in many wavebands.

The induction effect of the electromagnetic waves makes the reflector system the most important one in electromagnetic wave telescopes except for very short and very long wavelengths. By using a large reflector system, very weak radiation signals are collected over a large aperture area and are focused to detectors. However, any telescope has its own sensitivity limitation which is set by its reflector size. To overpass the existing detecting threshold, larger and larger aperture telescopes are required and have been built. This trend is not the same for nonelectromagnetic wave or particle telescopes.

Figure 11.1 shows the opening up of the electromagnetic wave observations beyond the visible region as a function of time throughout the past 100 years mainly driven by science and technology improvement. It can be found that the observations of the highest energy regions in electromagnetic waves began only very recently.

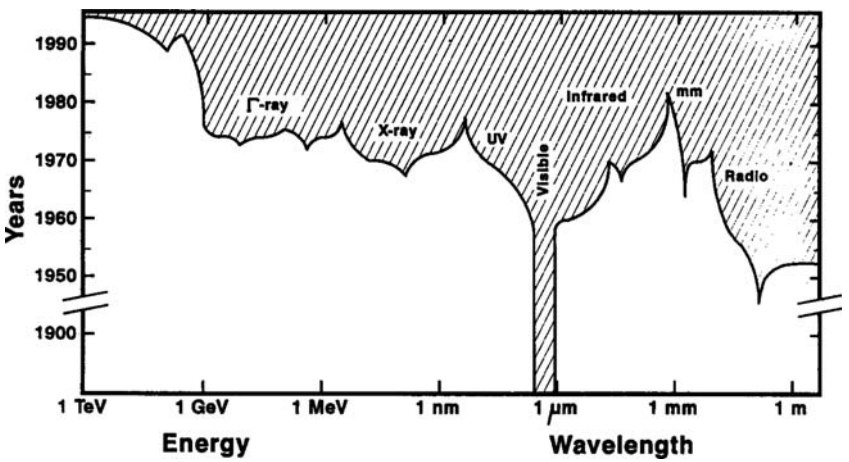


Fig. 11.1. A time chart showing the opening up of the electromagnetic wave observations in the last century (Giovannelli and Sabau-graziati, 1996, updated from Lena, 1988).

11.3 Nonelectromagnetic Telescopes

For the development of astronomical telescopes, the 20th century was the most important 100 years where observations of the spectrum opened up to include not only electromagnetic waves, but also nonelectromagnetic waves or particles, including gravitational waves, cosmic rays, and dark matter.

The foundation of a gravitational wave telescope is based on Einstein's prediction made in the early 20th century. The key to his theory is that the presence of mass will produce curvature in the space-time. This prediction was confirmed from an astronomical observation where light rays were indeed bent when they passed by the edge of the sun. From this theory, the motion of mass or energy would introduce ripples in the space-time and produce gravitational waves. The presence of gravitational waves in the universe was again confirmed from an indirect binary system observation where the system energy lost is exactly equal to the energy of the gravitational waves it produced.

Gravitational waves as a field is directly linked to dark matter, black holes, neutron stars, and others. What is the form of dark matter? How is a black hole formed and how does it change the space-time? And what does a neutron star of enormous mass with a dimension of only 20 km look like? All these questions are difficult to answer using direct observations in the electromagnetic spectrum. However, observations of the gravitational waves can provide important details for these interesting and mysterious objects. These include their forms, directions, energy levels, and the frequencies of their movements.

Attempts to detect the gravitational waves directly began in the 1950s. Both resonance bar and laser interferometer devices were used. However, to date, limited by high noise level and weak signal, no reliable gravitational wave observation results have been obtained from these already very sensitive instruments. The resonance bar detectors still require a lower temperature and the laser interferometers still require higher quality factor mirror material and more vibration suppressions. These improved bar devices and laser interferometers are still struggling in the search for gravitational waves. With these instruments, it is hoped that real observational results of the gravitational waves will come soon.

Cosmic ray observations started in 1912 when Victor Hess observed the ionized particles at high altitude during a balloon flight. The cosmic rays, mostly protons, including heavier atomic nuclei, have a wide energy spectrum up to about 10^{20} eV ($= 16$ J), which is over fifty million times more energetic than the particles produced by the Fermilab Tevatron accelerator. Cosmic rays with positive or negative charges and moving at very high speed will change direction when they pass through a stellar or interstellar magnetic field. Therefore, the observations of cosmic rays usually provide no real directions of their sources except in some special cases.

One of the most important questions or outstanding puzzles in astrophysics concerning cosmic rays is: How the energetic cosmic ray gains its energy in the universe? At present, we can only speculate about the conditions at their sources

because we know of no standard object, i.e., supernova, pulsar, or even a black hole, that could easily accelerate particles to such enormous energies at present. The sources of these cosmic rays must either have very strong magnetic fields or be enormous in size. To judge these speculations, direct or indirect cosmic ray observations are necessary. The frontier now in astronomy and particle physics is to study these extremely high energy cosmic ray particles.

As most cosmic rays cannot penetrate down to the lower level of the atmosphere, direct cosmic ray observations have to be performed at high altitude or in orbit. Indirect cosmic ray observations relying on the Cherenkov effect can be performed on the ground, underground, or underwater. One difficulty in high energy cosmic ray observations is that the ray flux is inversely proportional to an exponential function of the ray's energy level. As the cosmic ray energy increases, the chance to catch them becomes very difficult. At ground level, a large collecting area can be made for catching these rarely occurring but extremely high energy cosmic ray events. Now large ground facilities cover an extremely large area (6,000 km^2) and with an extremely wide field of view (30° in elevation and 360° in azimuth). It is hopeful that astronomers will soon have observational evidence concerning the mystery surrounding the extremely high energy issue.

The first dark matter particle observed was the neutrino which is a hot particle with no charge that travels near the speed of light. The neutrino, first predicted by Pauli in 1930, has a very small cross-sectional area when it hits other nuclei or particles. They are like chargeless electrons and always maintain the direction from where they were generated in the universe. Different from the electron which is affected by both electromagnetic and weak forces, the neutrino only has weak interactions exerting over an extremely small distance.

Neutrinos can penetrate the whole universe with little notice, even though they do interact with atomic nuclei in rare cases as its distance to them becomes extremely small. With a huge volume of detecting nuclei in liquid form, the interactions of solar and cosmic ray-induced neutrinos with the detector is possible as the solar neutrino fluxes and the number of detecting atoms are huge. However, modern cosmology theory shows that the neutrinos form only a small part of the missing dark matter. The search for the cold dark matter, i.e., particles with very low speed, started in the last 15 years. Cold dark matter detectors are extremely sensitive, underground, cryogenically cooled, quantum measuring devices with superconductor absorbers. However, the noise level remains a problem in the search for dark matter and so far, no reliable results have been obtained from any of these instruments. Real dark cold particle detections will not occur for a few years.

11.4 Ground Astronomical Telescopes

Ground and high mountain electromagnetic wave facilities include radio, millimeter, infrared, and optical telescopes as well as indirect gamma ray telescopes. Most of these telescopes have their collecting surfaces in a paraboloidal shape.

This reflector shape has wider waveband coverage. Therefore, some optical telescopes can be used in the infrared regime and some precise radio telescopes can be used in the millimeter wavelength regions. Generally, optical, infrared, and millimeter wavelength telescopes are sited at high and dry mountain areas. Longer wavelength radio telescopes are sited at ground level, but with less man-made radio interferences.

Optical telescopes have a long development history of 400 years. Table 11.2 lists major existing and planned optical and infrared telescopes. The existing largest single mirror optical telescope has a diameter of 8.4 m and the largest segmented mirror optical/infrared telescope has a diameter of 10.4 m. Dedicated infrared telescopes include the 3.8 m UKIRT and the 3.6 m Canada-France-Hawaii telescope (CFHT).

In the past 10 years, a new generation of large optical and infrared telescopes has been built. These include the Keck I and II (10 m), the Gemini South and North (8.3 m), Subaru (8.4 m), the LBT (two 8.4 m), the VLT (four 8.2 m), the HET (10 m), the SALT (10 m), and the GTC (10.4 m). Among these, some are segmented mirror telescopes.

Table 11.2. Major optical telescopes (expanded from Bely et al., 2003)

Name	Aperture (m)	Site	Observatory	Year
E-ELT	1×42	Chile	ESO	2017
TMT	1×30	Chile	UC, NOAO	2016
LAMA	66×6	Chile	Canada	undecided
GMT	1×22	Chile	UofA	2016
ATST	1×4	Hawaii	NSO	2014
LSST	1×8.4	Chile	NOAO	2014
Lamost	1×4	Beijing	NAO	2008
GTC	1×10.4	Canary Islands	Spain	2005
SALT	1×10	South Africa	S. Africa	2003
Magellan	2×6.5	Chile	US	2002
LBT	2×8.4	Arizona	UofA	2004
VLT	4×8.2	Chile	ESO	2001
Gemini	2×8.3	Hawaii, Chile	US, UK	2001
Subaru	1×8.4	Hawaii	JNO	1999
HET	1×9.2	Taxes	MAO	1999
MMT	1×6.5	Arizona	SMO	1999
Keck	2×10	Hawaii	Keck	1998
Herschel	1×4.2	Canary Islands	UK	1986
Bolshoi	1×6	Russian	Russian	1976
NOAO	2×4	US, Chile	NOAO	1974
Hale	1×5	US	Palomar	1949

Wide field of view telescopes are used for sky survey. Existing wide field telescopes are mostly Schmidt ones. However, the Schmidt telescope involves a refracting lens, so that the largest diameter is only 1.34 m. In 2000, a wide field dual reflector, the 2.5 m Sloan telescope, was built with a smaller field of view of 9 square degrees. A 4 m reflecting Schmidt telescope with a segmented mirror, LAMOST, was built in 2008 in China. An 8.4 m diameter three-mirror wide field reflecting telescope, Large-aperture Synoptic Survey Telescope (LSST), is also under construction in the US.

The new US large optical telescopes are the Thirty Meter Telescope (TMT) and the 22 m Giant Magellan Telescope (GMT), both as Giant Segmented Mirror Telescope (GSMT). ESO had also planned a 42 m European Extremely Large Telescope (E-ELT). All these projects are under development.

Solar telescopes are special ones. The largest solar telescope will be the 4 m Advance Technology Solar Telescope (ATST), now under construction. Mirror seeing of a solar telescope is a major concern. The energy input to the mirror surface can be more than 1,000 W/m^2. To reduce the heat, a flat mirror directs the beam down through a very long cold or vacuum tunnel and most of the heat is reflected by a cold Lyot stop to reduce the mirror seeing. This long tunnel formed tall structure is usually named the solar tower.

The radio telescope was invented in 1928 and the development was started from the long wavelength end. Meter and centimeter wavelength radio telescopes flourished after World War II. Early important radio telescopes included the Mills Cross, Kraus antenna, and the Jodrell Bank telescope. The 300 m Arecibo, 100 m Effesberg, and 100 m Green Bank ones are the largest fixed and steerable radio telescopes. A new 500 m Chinese fixed telescope is under construction. The millimeter wavelength telescopes started in the 1970s and were limited by the receiver technology. Radio interferometers started in the 1940s and were developed from the 1970s. The most important radio interferometers are the Vary Large Array (VLA) and Very Long Baseline Array (VLBA). Table 11.3 lists the major centimeter wavelength radio telescopes and Table 11.4 lists all the millimeter and submillimeter wavelength telescopes.

Ground-based electromagnetic wave telescopes also include gamma-ray Cherenkov telescopes. However, the accuracy of these Cherenkov telescopes is low, resulting in little impact on telescope design. Some of the gamma-ray Cherenkov telescopes are also used for cosmic ray observations. Some even use existing solar power facilities.

Major gravitational wave telescopes are listed in Table 11.5. The only space one listed among them is a planned space gravitational wave interferometer, LISA.

Neutrino detectors are similar to cosmic ray detectors. However, as the cosmic ray effects are dominant in the downwards direction, all the neutrino telescopes are located deep underground or underwater. The signal recorded by these detectors is coming from the other side of the earth. Table 11.6 lists major gamma ray, cosmic ray, and neutrino telescopes. Cold dark matter detectors are all deep underground. They are sensitive, cryogenically cooled, quantum detecting devices.

Table 11.3. Major existing and planned radio telescopes

Name	Aperture (m)	Site	Wavelength (mm)	Resolution (arcsec)
Fast	1×500	China	60–5,000	179
SKA	1 km^2	undecided	10–5,000	0.02/f(GHz)
Arecibo	1×300	Puerto Rico	60–900	60
ATCA	6×22	Australia	3.0–200	0.1
GBT	1×100	US	5.0–300	10
Effelsberg	1×100	Germany	3.0–730	10
Jodrell bank	1×76	UK	12.5–2,000	
Parks	1×64	Australia	1.3–900	50
VLBA	10×25	US	3–900	0.0001
EVLA	37×25	New Mexico	7–300	0.2/f(GHz)
Westerbork	14×25	The Netherlands	60–1,500	4
MERLIN	6×25–76	UK	13–2,000	0.01
Nancay	1×35–300	France	90–210	100
1hT(ATA)	500×6	US	30–200	3
GMRT	30×45	India	>210	2
LOFAR		Chile	>1,000	

Table 11.4. Major millimeter and submillimeter wavelength telescopes

Name	Aperture (m)	Site	Wavelength (mm)	Resolution (arcsec)	Year
ALMA	64×12	Chile	0.3	0.003	2012
CCAT	25×1	Chile	0.1	1.0	2013
LMT	50×1	Mexico	0.85	1.0	2008
APEX	12×1	Chile	0.3	1.4	2004
CARMA	10×6+6×10.4	US	1	0.2	2005
IRAM	6×15	France	1.5	0.6	2005
Nobeyama	1×64+6×10	Japan	3	1.5	2000
SMA	8×6	US	0.3	0.1	2003
CSO	1×10.4	US	0.3	7	1988
HHT	1×10	US	0.3	7	1995
IRAM	1×30	France	1.8	25	1999
JCMT	1×15	UK	0.3	5	1988
Nobeyama	1×45	Japan	3	16	1990
Delingha	1×13.4	China	3	50	1986
Datian	1×13.4	Korea	3	50	1988

Table 11.5. Major gravitational wave telescopes

Name	Mass or arm length	Sensitivity	Year
ALLEGRO	2,296 kg	1×10^{-21}	1996
AURIGA	2,230 kg	2×10^{-22}	1998
EXPLORER	2,270 kg	6×10^{-22}	1991
NAUTILUS	2,260 kg	2×10^{-22}	1997
NIOBE	1,500 kg	8×10^{-22}	1994
LIGO	4 km	10^{-20}	2002
VIRGO	3 km		2003
GEO600	600 m		2003
TAMA300	300 m		2005
AIGO	80 m		2015
LISA	5 Mkm		2020

11.5 Space Astronomical Telescopes

Space electromagnetic wave telescopes, developed from the balloon-borne and airborne ones, have definite advantages over the ground-based telescopes. These advantages are: (a) absence of restrictions owing to atmosphere absorption; (b) absence of atmosphere seeing and other phenomenon; (c) absence of restriction owing to meteorological conditions; (d) absence of the day–night cycle; and (e) access to the entire electromagnetic spectrum with only a few limitations, such as the absorption of the interstellar medium (ISM) in the extreme ultra-violet (EUV) range (91.2–10.1 nm) and the absorption of the ISM and interplanetary medium in the radio range. Space electromagnetic wave telescopes play an important role in astronomy. However, nonelectromagnetic wave or particle space telescopes play only a supplemental role to the observations carried out on the ground or underground, or underwater.

Today, balloon-borne and airborne telescopes are mainly used for infrared observation. Space telescopes in orbit are also named astronomical satellites. These telescopes have evolved from earlier smaller sizes to larger ones. Some of them have specific purposes and others have a number of instruments working at different spectrum bands.

In the EM wave spectrum, space telescopes are divided into three groups: short wavelength X-ray and gamma ray ones; medium wavelength optical, infrared, and UV ones; and long wavelength radio and millimeter wavelength ones. In the short wavelength group, the Einstein satellite is an important one with a nested X-ray grazing telescope. AXAF and XMM are missions that followed the Einstein. Table 11.7 lists the major space X-ray and gamma ray telescopes.

In the medium wavelength group, the HST is a very successful space telescope with very high pointing accuracy and angular resolution. The JWST is the next major space telescope under construction with a scheduled launch time of

Table 11.6. Major ground-based gamma ray, cosmic ray, and neutrino telescopes

Name	Location	Type	Area (m)
GRANITE	Arizona, US	ACT	1×10
CAT	France	ACT	17.7 m^2
MILARGRO	Los Alamos	Water Cherenkov	723 PTMs
CLUE	La Plama	UV Cherenkov	
CELESTE	France	Solar tower	
STACEE	Albuquerque	Solar tower	37 m^2
CAO (Crimean)	Ukraine	ACT	48 × 1.2
HEGRA	La Plama	EAS	200×200 m^2
MACE	India	ACT	1×21
TACTIC	India	ACT	4×3
HESS	Namibia	ACT	4×12
VERITAS	US	ACT	7×10
CANGAROO	Australia	ACT	4×10, 2×3.8
ARGO	China	EAS	
INCA	Bolivia	Scintillator	
Pierre Auger	US/Argentina	EAS/FD	300 km^2
HiRes	US	FD	
AGASA	Japan	EAS	100 km^2
EUSO	ESO	Space FD	
SNOLAB	Canada	Neutrino	1,000 t heavy water
DEMAND	Hawaii	Neutrino	Under water
NESTOR	Greece	Neutrino	Under water
KAMIOKANDE	Japan	Neutrino	50,000 t pure water
ICEcube	South pole	Neutrino	4,200 detectors

2013. Table 11.8 lists major space optical, infrared, and millimeter wavelength telescopes and, in Table 11.9, major long wavelength space telescopes are listed. Space observations of nonEM waves or particles have just started. There are only a few such telescopes to date.

Table 11.10 is a short list of all important future large space telescope projects planned in 2006. Some of them are under construction, some are under design study and some have already been cancelled.

11.6 Man's Space Missions

This and the next section provide astronomical telescope-related information for the reader's reference. In this section, man's space missions are reviewed and all missions are described in moon and planet order (Goebel, 2008). The next section is a brief summary of reconnaissance telescopes.

Table 11.7. Major space X-ray and gamma ray telescopes

Name	Year.month	Perigee	Apogee	Angle	Purpose	Note
SAS-1	1970.12	324 km	350 km	3.0	X ray	Uhuru
TD-1A	1972.3	534	539	97.5	Gamma	
Ariel 5	1974.10	513	557	2.9	X-ray	
COS-B	1975.8	342	99,837	90.1	Gamma	
HEAO-1	1977.8	424	444	22.7	X ray	
HEAO-2	1978.11	355	364	23.5	X ray	Einstein
HEAO-3	1979.9	424	457	43.6	Gamma	
EXOSAT	1983.5	356	1,91,581	72.5	X ray	
Astro-C	1987.2	510	673	31.1	X ray	
ROSAT	1990.6	560	578	53	X ray	
CGRO	1991.4	445	458	28.5	Gamma	
XTE	1995.12	565	583	23	X ray	
SAX	1996.4	555	605	4	X ray	
AXAF	1999.7	∼10,000	∼1,28,000		X ray	Chandra
XMM	1999.12	825.6	1,13,946		X ray	Newton
HXMT	>2010				X ray	China

11.6.1 Moon Missions

Immediately after the first satellite launching by the USSR in October 1957, a competitive space race started between the US and the USSR. The target of this race was the moon.

The US Air Force had five failed missions Pioneer 0 to 4 in 1958 and 1959. Then the Jet Proportion Laboratory (JPL) launched Ranger 1 to 9 between

Table 11.8. Major optical and infrared space telescopes

Name	Year.month	Perigee	Apogee	Angle	Purpose	Note
KAO	1970–1995				Infrared	Airborne
Salyut	1977.9	380	390	51.6	Sub-mm	USSR
IRAS	1983.1	896	913	99.0	Infrared	
Hopparcos	1989.8	500	36,000		Astrometry	
COBE	1989.11	900	900	99.0	Infrared	
Hubble	1990.4	563	568	28.5	Optical	
ISO	1995.11	1,036	70,578	5.2	Infrared	
SOFIA	2002				Infrared	Airborne
Spitzer	2003.8	Helio orbit			Infrared	
JWST	2013	L2 orbit			Infrared	
WSO	2010		35,800	51.8	UV	Russian

Table 11.9. Major radio space telescopes

Name	Year.month	Perigee	Apogee	Angle	Size	Note
REA-1	1968.7	5,829 km	5,864 km	120.9		US
Explor43	1971.3	146	1,22,146	28.7		US
REA-2	1973.6	1,100	1,100	59.0		US
Salyut-6	1983.1	896	913	99.0	10 m	USSR
VSOP	1997.2	570	21,527	31.2	8 m	Japan
SWAS	1998.12	LEO	600	70	0.55×0.7	US
WMAP	2001.6	L2 orbit			1.4×1.6 m	US
ARISE	2008	1,000	76,800	51.5	25 m	US

1961 and 1965. All failed except Ranger 7 to 9. In 1966 and 1967, the Langley center of NASA sent five successful Lunar Orbiter missions. At the same time, the JPL launched seven successful Surveyor missions with only two crashing into the moon's surface.

The USSR had two failed Luna E-1 and E-1A missions in 1959 followed by a Luna 2 which hit the moon, but did not return any data. In October 1959, Luna 3 brought pictures from the far side of the moon. This was the first time that humans saw the other side of the moon. From 1963 to 1965, Luna E-3 and E-4, Luna 7 and 8, and eight Luna E-6 all failed, but these were followed by a successful Zond 3 (Zond means to probe) mission which sent back images of the far side of the moon. In 1966, Luna 9 to 13 landed on the moon successfully.

Table 11.10. Future space major projects (Stahl, 2006)*

Project name	Proposed date
James Webb Space Telescope (JWST)	2011
Space Interferometer Mission (SIM)	2012
Laser Interferometer Space Antenna (LISA)	2014
Terrestrial Planet Finder Coronagraph (TPF-C)	2016
Constellation X (ConX)	2017
Terrestrial Planet Finder Interferometer (TPF-I)	2019
Large microwave	2019
Single Aperture Far-Infrared (SAFIR)	2022
Large Ultra-Violet Observatory (LUVO)	2024
Life Finder (LF)	2026
Black Hole Imager (BHI)	2028
Big Bang Observer (BBO)	2028
Stellar Imager (SI)	2030
Far-Infrared Sub-MM Interferometer (FIRSI)	2032
Planet Imager (PI)	2034

* Many projects in this list have been significantly delayed or cancelled.

The manned moon missions run parallel with the unmanned ones. In 1961, the USSR cosmonaut Yuri Gagarin became the first human in space with a 108 minute orbital flight. However, the manned moon mission program in the USSR was without luck and the attempt was finally stopped in 1975.

The US had successful manned moon missions with the Apollo spacecraft. The spacecraft included three modules: a command, service, and lunar module. Unmanned earth and lunar orbit flights were tested in 1968 and 1969 with Apollo 7 to 10. In July 1969, Apollo 11 brought the first man, Neil Armstrong, to the moon. Afterwards, 11 men followed on Apollo 11 to 17. The program ended in 1973.

Lunar exploration was revived in the 1990s. The US Clementine was sent out in 1994 with CCD cameras and a laser ranging system. The Lunar Prospector was launched in 1998. An ESA Small Missions for Advanced Research in Technology 1 (SMART-1) was launched in 2003 and it impacted the moon in 2006. The Chinese Chang'e-1 and Japanese Selenological & Engineering Explorer (SELENE), also called Kaguya, were launched in 2007. The Indian Chandrayaan-1 was launched in 2008. These are all lunar orbiting satellites. Chang'e-1, launched on October 24, 2007, was intentionally crashed on the moon surface on March 1, 2009 after running out its fuel.

With no atmosphere and ionosphere, the moon's surface is an ideal location for astronomical telescopes. The far side of the moon is free from the noisy radio signals emanating from the earth. Radio telescopes on the far side of the moon will pick up the extremely faint signals left over from the early universe. One concern is moon dust contamination. However, the Apollo data suggests that moon dust levitates electrostatically, so that magnetic shielding might protect a telescope. A few optical and radio telescopes have been proposed on the Lunar surface, including a 16 m Large Lunar Telescope (LLT), the Lunar Array for Radio Cosmology (LARC), and the Dark Ages Lunar Interferometer (DALI).

11.6.2 Mercury Missions

Mercury is the closest planet to the sun. NASA launched Mariner 10 in 1973. It had three flybys in 1974 and 1975 with a closest distance of 327 km from Mercury.

The Mercury Surface, Space Environment, Geochemistry, & Ranging (Messenger) probe was launched in 2004. The probe had an earth flyby in August 2005, Venus flybys in October 2006 and June 2007, and three Mercury flybys in January 2008, October 2008, and September 2009.

The ESA BepiColombo will be launched in 2013 and will arrive at Mercury in 2019.

11.6.3 Venus Missions

The US Mariner 1 failed in July 1962, but Mariner 2 arrived on Venus in December 1962. Mariner 5 and 10 had Venus flybys in October 1967 and February 1974.

The Pioneer Venus Orbiter was launched in 1978 and fulfilled 17 experiments in Venus orbit before 1992. Magellan arrived in Venus orbit on 10 August 1990 with a five-year observation period.

The USSR carried out six failed missions from 1961 to 1965. However, Venera 11, 13, 14, 15, and 16 all arrived on Venus between 1978 and 1983. Vega 1 and 2 also touched Venus on 11 and 16 June 1985.

ESA launched the Venus Express (VEX), which includes six instruments, in 2005 and the observations started in April 2006. Another planned Japanese Planet C mission will be launched in 2010.

11.6.4 Mars Missions

The US Mariner 3 launch failed in 1964, but Mariner 4 flew past Mars with returned data on 14 July 1965. Mariner 6 and 7 launched in February and March 1969 had flybys on 31 July and 4 August 1969. Mariner 8 failed in 1971, but Mariner 9 launched at the same time returned thousands of images over a period of nearly a year.

Viking 1 and 2 arrived on Mars on 20 July and 3 September 1975. Viking 2 failed in 1978 and Viking 1 operated until August 1980. The Mars Observer (MOB) failed in 1992. Both Mars Global Surveyor (MGS) and Mars Pathfinder succeeded in November and December 1996. The rover of Pathfinder, Sojourner, became the first vehicle on any planet.

The Mars Climate Orbiter (MCO) and Mars Polar Lander (MPL) launched in December 1998 and January 1999 and worked until December 2004 and March 2000, respectively. The Mars Surveyor 2001 (Mars Odyssey) launched on 7 April 2001 and arrived on Mars on 24 October, 2001. It remains in Mars orbit.

Mars Exploration Rover A (MER-A or Spirit) and MER-B (Opportunity) launched in June and July 2003 set down on the 3 and 24 January 2004, respectively. Both rovers are still in operation. The Mars Reconnaissance Orbiter was launched in August 2005 and arrived at Mars orbit in March 2006. It will operate until 2014. Another probe, Phoenix, launched on 4 August 2007 arrived at the Mars north pole on 25 May 2008. The Mars Science Lab will be launched in 2011.

The USSR had many failed Mars missions between 1960 and 1973. In 1988, Phobo 2 arrived on Mars and worked until 27 March 1989.

In July 1998, Japan sent a Mars orbiter, Planet B, also known as Nozomi (hope). The orbiter arrived at Mars in October 1999, but remained in a solar orbit. It was finally damaged by solar flare in 2003.

In June 2003, an ESA Mars Express including an orbiter and the lander Beagle 2 was launched, but the Beagle 2 was lost after it entered the Martian atmosphere in late 2003.

11.6.5 Jupiter Missions

The Jupiter probe, Pioneer 10 (Pioneer F) launched by NASA on 2 March 1972 crossed the Mars orbit on 25 May 1972, and entered the asteroid belt between

15 July 1972 and 15 February 1973. Pioneer 10 began its Jupiter observations between 3 November and 3 December 1973. After crossing Neptune's orbit on 13 June 1988, it sailed out of the solar system. Pioneer 11 launched 11 months later. The same pattern followed.

Voyager 1 and 2 launched in 1977 and reached Jupiter in January and July 1979. After sending back a great deal of data, the Voyagers went on to Saturn. In October 1989, the probe Galileo was launched by the space shuttle. It passed Venus on 19 February, passed asteroid Gaspra on 29 October 1991, and passed the earth on 8 December 1992. It released an atmospheric probe on 13th July 1994. In the later part of 1994, while still 18 months away from Jupiter, Galileo was able to photograph the impact of a comet on Jupiter's surface. The spacecraft entered the Jovian system in late 1995 and the mission was ended on 21 September 2003.

11.6.6 Saturn, Uranus, Neptune, and Pluto Missions

Pioneer 11, Voyager 1, and Voyager 2 all went to Saturn on 1st September 1979, 12th November 1980, and 25th August 1981 after a Jupiter flyby.

The only dedicated Saturn spacecraft is Cassini which carried a lander, Huygens, built by ESA. Cassini launched in October 1997 and entered Saturn orbit on 30th June 2004. Huygens was released on 25th December 2004 and reached Titan, a moon of Saturn, on 14th January 2005.

After Jupiter and Saturn flybys, Voyager 2 passed Uranus and Neptune on 24th January 1986 and 24th January 1989. It discovered ten small moons of Uranus. As of September 2007, Voyager 1 was 104 astronomical units (AU is the mean distance between the sun and earth) from the sun while Voyager 2 was 84 AU away.

The only Pluto mission, New Horizon, was launched on 19th January 2006. This spacecraft will have a Pluto flyby on 14th July 2015.

11.6.7 Asteroids and Comet Missions

Most asteroids reside between Mars and Jupiter. The asteroid belt filled a gap in the distance series between the sun and its planets defined by Bode–Titius Law.

The first asteroid to be photographed close-up was the asteroid 951 Gaspra in 1991 by Galileo while on its way to Jupiter. In 1993, Galileo imaged the asteroid 243 Ida.

The first asteroid mission was the JPL Near Earth Asteroid Rendezvous (NEAR) launched on 17th February 1996. In 1997, NEAR, also called NEAR-Shoemaker, had an asteroid 253 Mathilde flyby. On 14th February 2000, NEAR reached its objective, Eros, and went into orbit around it.

In early 2000, the Cassini probe imaged the asteroid 2685 Masursky. The Stardust comet probe launched in 1999 imaged asteroid Annefrank on 2nd

November 2002 and the ESA Rosetta comet probe launched in 2004 will perform observations of two asteroids in 2008 and 2010.

The Japanese ISAS MUSES-C probe launched on 9 May 2003 is a mission to the asteroid 1998 SF36 in an orbit between Earth and Mars. Another asteroid probe, Dawn, launched from Cape Canaveral on 27 September 2007, is to orbit asteroid Vesta and Ceres and perform detailed observations.

In 1986, comet Halley came back after 76 years. The ESA Giotto, launched on 2 July 1985, had a comet flyby at a distance of less than 600 kilometers and returned high-quality images. The Vega 1 and 2 also performed this comet flyby. The Japanese Sakigake and Suisei performed limited observations on the comet.

The JPL "Deep Space 1 (DS1)" probe launched on 24 October 1998 had a closest approach to the comet Borrelly at 22,000 km on 22nd September 2001.

The Stardust launched on 7th February 1999 had a rendezvous with comet Wild 2 (pronounced "Vilt 2") and returned samples from its coma and tail in a reentry capsule in the deserts of Utah on 15 January 2006. It was the first planetary sample-return mission to any place but the Moon. The Comet Nucleus Tour (CONTOUR) launched from Cape Canaveral on 3 July 2002 visited several comets including comet Encke in November 2003. On 15 August 2002, the probe performed its last engine burn to leave Earth orbit. The Deep Impact launched on 12 January 2005 had a flyby of comet Tempel 1 on 4 July 2005. The next target is comet Hartley 2 with the flyby scheduled for 11 October 2010.

The ESA "Rosetta" comet probe launched on 2nd March 2004 was designed to rendezvous with a comet and drop a lander named "Philae" to the comet's surface. The spacecraft's path was to take it by asteroid Steins on 5th September 2008, and then by asteroid Lutetia on 10th July 2010.

11.7 Reconnaissance Telescopes

With their superior angular resolution and very powerful sensitivity, astronomical telescopes have direct applications in reconnaissance.

After the invention of the optical telescope, the inventor, Hans Lippershey, sold his telescopes to the military. Optical telescopes can be used to monitor an approaching army or navy. The curvature of the earth limits the sight of telescopes so that observation towers were built in early days. The hot-air balloon, invented in 1783, could be used by the military but the use was limited by the unpredictable wind and the delayed delivery of information. The photograph and airplane invented in 1827 and 1903, respectively, made aerial surveillance possible. England was the first to perform an aerial survey of the German military early in 1936. During World War II, both the Allied and Axis powers used airborne cameras for military purposes. Radio radars developed rapidly in the war time, which are special radio telescopes.

The high-altitude U2 and SR-71 Blackbird airplanes were used for US intelligence in the 1950s. The USSR also developed a high-altitude Myasishchev M-55 airplane. The telephoto cameras used on these airplanes have a lens

diameter of about 1 m. The photos of the Cuban missile site gave most people a shock as the photos were so clear that the missile trailers could all be identified. Now airborne multi-wavelength telescopes are used for intelligence, such as those used in EP-3 airplanes. In 2001, a collision occurred between an EP-3 plane and a Chinese fighter. The radar early warning airplanes, such as the E-3 Sentry and Grumman E-2C Hawkeye, are other special airborne radio telescopes Now, telescopes are equipped inside Unmanned Aerial Vehicles (UAVs).

Satellites are also equipped with telescopes and their operation requires help from ground-based telescopes. In total, there have been about 4,500 satellite launches within the past 50 years. Among them, 850 are active satellites and the US owns more than half. The reconnaissance satellites make up about 40%. The satellites' access to the earth is easy and fast as the relative speed between satellites and the ground is about 8 km/s.

The American National Reconnaissance Office used the Keyhole (KH) series satellites (Anderson, 2007). KH-1 to KH-4 were Corona space cameras with diameter and 7-8 m resolution. The films were retrieved by C-119 airplanes in mid air. KH-5 to KH-6 were Argon and Lanyard and KH-7 to KH-8 were Gambit. KH-9, named Hexagen or Big Bird, was the last film-based space camera satellite. There were 20 launches of KH-9 from 1971 to1986 with only one failure.

The KH-11, named 1010, Crystal, or Kennan, included ten launches from 1986 to 1990 with one failure. These cameras were believed to be a similar size to the HST. The atmospheric seeing when looking down is much less serious than looking up and no adaptive optics was required for these telescopes. Their resolution was about 0.15 m.

From 1992 to 1999, four launches of the KH-12, named Advanced Keyhole, Ikon, or improved Crystal, were made by powerful Titan-IV rockets with a fairing length of about 5.1–5.9 m. The estimated diameter of them was about 2.4–3.1 m. The resolution might reach a few centimeters. The most recent launches were between 2001 and 2005 to sun-synchronous orbits. Segmented mirrors might be used on these space cameras.

Early in 2007, one satellite in the series was reported crashing in Peru and some radioactive isotope Pu-238 for power generation contaminated the site, causing a mysterious illness in residents. In February 2008, a malfunctioning bus-sized space camera US 139 was shot into pieces by an S-3 missile of the US Navy to avoid earth contamination. The missile used was also equipped with a small infrared telescope for target guiding.

Besides optical cameras, Lacrosse 1 to 5 radar imaging reconnaissance satellites, the Defense Meteorological Satellite Program (DMSP) and Automatic Nanosatellite Guardian for Evaluating Local Space (Angels) were telescopes in other wavelengths. It is believed that deployable radio dishes larger than 50 m in size have been in orbit for some years already.

From 1961 to 1994, about a hundred of the USSR Zenit (Kosmos) reconnaissance satellites with short average on-orbit time of a few days each were

launched. The early films and reusable cameras were retrieved by reentry capsules. In general, the electronics of early Russian space cameras were worse than the US ones.

With an increasing number of reconnaissance satellites, special groups were formed in many countries to study man-made near earth objects. Some early monitoring work was done by astronomical observatories but now this is usually done by the military. The laser guide star technique used in adaptive optics was first used in the US military. One major near earth object telescope is the 3.6 m Advanced Electro-Optical System (AEOS) in Maui, Hawaii, housed inside a collapsible cylinder dome for quick access to the sky. Its slew rate is about $18^\circ/s$ in azimuth and $5^\circ/s$ in elevation and it has a number of Coude foci equipped with unknown instruments. Another telescope array with four 1.8 m telescopes is used for redundant checking of sky images to identify any small moving object in orbit.

The most recent application of adaptive optics is the ground-based and airborne laser gun projects. The Air Borne Laser (ABL) gun with a 1.5 m diameter gold-plated beam expander and deformable mirror is on a Boeing 747-400 airplane for shooting down hostile ballistic missiles in early stages of flight. Three laser systems are involved; one for pointing stabilization, one for atmosphere phase error detection, and the other for shooting down the missile. Since the laser used is in the mega-watt-class, its effectiveness is still a question. The astronomical radio multi-beam technique is also the basis of another important innovation, the widely used global positioning system (GPS).

References

Anderson, G., 2007, The telescope, its history, technology, and future, Princeton University Press, Princeton.

Bely, P., et al., 2003, The design and construction of large astronomical telescopes, Springer, New York.

Davies, J. K., 1997, Astronomy from space, John Wiley & Sons and Praxis, New York.

Giovannelli, F. and Sabau-graziati, L., 1996, High energy multifrequency astrophysics today, in Italian physical society conference proceedings Vol. 57, eds. Giovannelli, F. and Mannocchi, G., Vulcano Workshop 1996 Frontier Objects in Astrophysics and Particle Physics, Vulcano.

Goebel, G., 2008, Missions to the planets, on public domain web site.

Hu, Q., 2007, General and structural design of astronomical telescopes, Nanjing Institute of Astronomical Optics and Technology, Nanjing, China.

Huang, Y., 1987, Observational astrophysics, Science Press, Beijing.

Lena, P., 1988, Observational astrophysics, Springer-Verlag, Berlin.

National Research Council, 2001, Astronomy and astrophysics in the new millennium, National Academic Press, Washington.

Stahl, H. P., 2006, Mirror technology road map for optical/IR/FIR space telescopes, SPIE Proc., 6265, 626504.

Appendix A
Abbreviations of Telescope Names

1hA, One Hectare Telescope
21CMA, 21 CM Array
AAT, Anglo-Australia Telescope
ACE, Advanced Composition Explorer
AGASA, Akeno Giant Air Shower Array
AIGO, Australia Interferometer Gravitational wave Observatory
ALMA, Atacama Large Millimeter Array
AMANDA, Antarctic Muon And Neutrino Detector Array
AMiBA telescope, Array for Microwave Background Anisotropy telescope
AMS, Alpha Magnetic Spectrometer
ANITA, Antarctic Impulsive Transient Anwtenna
APEX, Atacama Pathfinder EXperiment
ARC, Astrophysical Research Consortium
ARISE, Advanced Radio Interferometry between Space and Earth
ASTP, Apollo-Soyuz Test Project
ATA, Allen Telescope Array
ATCA, Australia Telescope Compact Array
ATST, Advance Technology Solar Telescope
AXAF, Advanced X-ray Astronomy Facility (Chandra X-ray Observatory)
BAT, Burst Alert Telescope (in Swift)
BIMA, Berkeley Illinois Millimeter Array
BTA, Bolshoi Teleskop Azimultalnyi (Big Telescope Alt-azimuthal)
BATSE, Burst And Transient Source Experiment
CACTUS, Converted Atmospheric Cerenkov Telescope Using Solar 2
CANGAROO, Collaboration between Australia and Nippon foe a Gamma Ray Observatory
CARMA, Combined Array for Research in Millimeter-wave Astronomy
CAST, CERN Axion Solar Telescope
CAT, Cherenkov Array at Themis
CDMS, Cryogenic Dark Matter Search

CERGA interferometer, Centre d'Etudes et de Recherches Geodynemiques et Astronomiques interferometer
CFHT, Canada-France-Hawaii Telescope
CGRO, Compton Gamma Ray Observatory
CHARA array, Center for High Angular Resolution Astronomy array (6×1 m)
CLUE, Cherenkov Light Ultraviolet Experiment
COAST, Cambridge Optical Aperture Synthesis Telescope
COMPTEL, imaging Compton Telescope
ConX, Constellation X
CRTNT, Cosmic Ray Tau Neutrino Telescopes
CXO, Chandra X-ray Observatory
DALI, Dark Ages Lunar Interferometer
DAMA, Dark Matter
DRIFT, Directional Recoil Identification From Tracks-I
DUMAND, Deep Underwater Muon And Neutrino Detector
E-ELT, European Extremely Large Telescope
EGRET, Energetic Gamma Ray Experiment Telescope
EUVE, Extreme Ultraviolet Explorer
EUSO, Extreme Universe Space Observatory
EVLA, Expanded Very Large Array
FASR, Frequency Agile Solar Radiotelescope
FAST, 500-hundred-meter-diameter Aperture Spherical Radio Telescope
FIRST, Far Infrared and Sub-millimeter Telescope (Herschel Space Observatory)
FORTE, Fast On-orbit Recording of Transient Events
FUSE, Far-Ultraviolet Spectroscopic Explorer
GAIA, Global Astrometric Interferometer for Astrophysics
GALEX, Galaxy Evolution Explorer
GBT, Green Bank Telescope
GI2T, Grand Interferometre a 2 Telescopes
GLAST, Gamma-ray Large Area Space Telescope (Fermi Gamma ray Space Telescope)
GLUE, Glodstone Lunar Ultra-high energy neutrino Experiment
GMRT, Giant Metre-wavelength Radio Telescope
GMT, Giant Magellan Telescope
GRACE, Gamma-ray Astrophysics through Coordinated Experiment
GSMT, Giant Segmented Mirror Telescope
GTC, Gran Telescope Canaries
HDMS, Heidelberg Dark Matter Search
HEAO, High Energy Astronomy Observatories
HEGRA array, High Energy Gamma Ray Astronomy array
HESS, High Energy Stereoscopic System
HET, Hobby-Eberly Telescope
HETE, High Energy Transient Explorer
HHT, Heinrich Hertz Telescope

HiRes, High Resolution Fly's Eye
HST, Hubble Space Telescope
HUT, Hopkins Ultraviolet Telescope
IACT, imaging air Cherenkov telescope
INCA, Investigation on Cosmic Anomalies
Integral, International Gamma Ray Astrophysics Laboratory
IOTA, Infrared Optical Telescope Array
IRAS, InfraRed Astronomical Satellite
ISI, Infrared Spatial Interferometer (3×1.65 m)
ISO, Infrared Space Observatory
IUE International Ultraviolet Explorer
IXO International X-ray Observatory
JCMT, James Clerk Maxwell Telescope
JWST, James Webb Space Telescope (NGST)
KAMIOKANDE, Kamioka Nucleon Decay Experiment
LAMOST, Large sky Area Multi-Object Spectroscopic Telescope
KAO, Kuiper Airborne Observatory
LARC Lunar Array for Radio Cosmology
LBA, Long Baseline Array (Australia)
LBT, Large Binocular Telescope (Columbus Telescope)
LCGT, Large-scale Cryogenic Gravitational wave Telescope
LIGO, Laser Interferometer Gravitational wave Observatory
LISA, Laser Interferometer Space Antenna
LLT, Large Lunar Telescope
LMT, Large Millimeter Telescope
LOFAR, LOw Frequency ARray
LSST, Large-aperture Synoptic Survey Telescope
MACE, Major Atmospheric Cerenkov telescope Experiment
MAGIC, Major Atmospheric Gamma ray Imaging Cherenkov telescope
MARCO, Monopole Astrophysics and Cosmic Ray Observatory
MERLIN, Multi-Element Radio-Linked Interferometer Network
MILAGRO, Multi Institution Los Alamos Gamma Ray Observatory
MMT, Multiple Mirror Telescope
MROI, Magdalina Ridge Observatory Interferometer (10 × 1.4 m)
MWA, Murchison Widefield Array
MYSTIQUE, Multi-element-Ultra-Sensitive Telescope for Quanta of Ultra-high Energy
NESTOR, NEutrinos from Supernova and TeV sources Ocean Range
NTT, New Technology Telescope
NPOI, Navy Prototype Optical Interferometer
OAO, Orbit Astronomy Observatory
OGO, Orbit Geophysics Observatory
OHANA, Optical Hawaiian Array for Nano-radian Astronomy
ORFEUS, Orbiting Retrievable Far and Extreme Ultraviolet Spectrometer
OSO, Orbit Solar Observatory

OSSE, Oriented Scintillation Spectrometer Experiment
OVLBI, Orbiting VLBI
OWL, Over-Whelmingly Large Telescope
PICASSO, Project In CAnada to Search for Supersymmetric Objects
PTO, Palomar Testbed Interferometer
RICE, Radio Ice Cherenkov Experiment
ROSEBUD, Rear Object Search with Bolometers Underground
SalSA, Saltdome Shower Array
SALT, South African Large Telescope
SAS, Small Astronomy Satellite
SEST, Swedish-ESO Sub-millimeter Telescope
SKA, Square Kilometer Array
SIM, Space Interferometry Mission
SIRTF, Space IR Telescope Facility (Spitzer Space Telescope)
SMA, Sub-Millimeter Array
SMT, segmented mirror telescope
SOAR telescope, Southern Astrophysical Research telescope
SOFIA Stratospheric Observatory For Infrared Astronomy
SOHO, Solar Heliospheric Observatory
Space FD, Space fluorescent detector
SPICA, Space Infrared Telescope for Cosmology and Astrophysics
SPT, South Pole submillimeter Telescope
SST, Spitzer Space Telescope
STACEE, Solar Tower Atmospheric Cherenkov Effect Experiment
SUSI, Sydney University Stellar Interferometer
SWAS, Submillimeter Wave Astronomy Satellite
Swift, Swift Gamma Ray Burst Explorer
TA, Telescope Array
TIM, Telescopio Infrarrojo Mexicano
TMT, Thirty Meter Telescope
TNG, Telescopio Nazionale Galileo
TPF-C, Terrestrial Planet Finder-coronagraph
TPF-I, Terrestrial Planet Finder-Interferometer
UKFMC, UK Dark Matter Colleboration
UKIRT, United Kingdom InfraRed Telescope
UIT, Ultraviolet Imaging Telescope
UVOT, Ultraviolet/Optical Telescope (in Swift)
VERITAS, Very Energetic Radiation Imaging Telescope Array System
VLA, Very Large Array
VLBA, Very Long Baseline Array
VLT(I), Very Large Telescope (Interferometer)
VSOP, VLBI Space Observatory Program (HALCA, Astro-G)
WHT, William Herschel Telescope
WIYN, Wisconsin, Indiana, Yale, and NOAO telescope
WMAP, Wilkinson Microwave Anisotropy Probe

WSO, World Space Observatory
XEUS, X-ray Evolving Universe Spectroscopy
XMM, X-ray Multi-mirror Mission (XMM-Newton)
XRS, X-Ray Spectrometer
XRT, X-Ray Telescope (in Swift)
XTE, X-ray Time Explorer
ZEPLIN, ZonEd Proportional scintillation in LIquid Noble gases

Appendix B
Prefixes for Standard Units

Symbol	Value	Name	Symbol	Value	Name
d	10^{-1}	deci	da	10^{1}	deca
c	10^{-2}	centi	h	10^{2}	hecto
m	10^{-3}	milli	k	10^{3}	kilo
μ	10^{-6}	micro	M	10^{6}	mega
n	10^{-9}	nano	G	10^{9}	giga
p	10^{-12}	pico	T	10^{12}	tera
f	10^{-15}	femto	P	10^{15}	peta
a	10^{-18}	atto	E	10^{18}	exa
z	10^{-21}	zepto	Z	10^{21}	zetta
y	10^{-24}	yocto	Y	10^{24}	yotta

Index

N