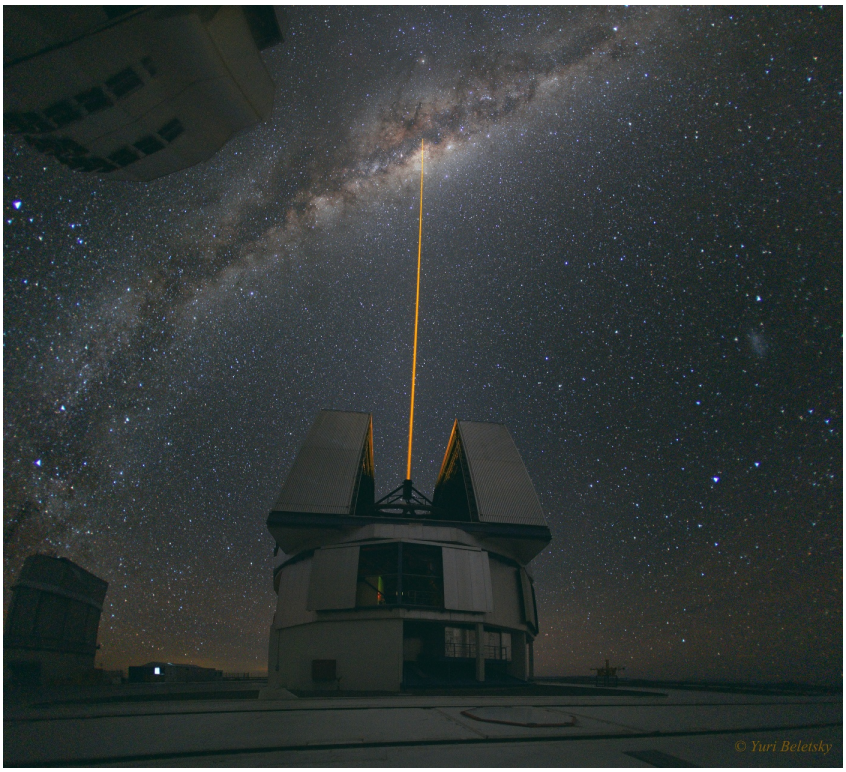
astronomical techniques

the course

astronomical techniques



course information

background

Astronomical Techniques (PHY217) is a compulsory first-semester course for second-year undergraduates doing dual-honours *Physics and Astrophysics* or *Physics and Astrophysics with Study Abroad* (BSc or MPhys). It is NOT available as an optional course for any other undergraduates.

PHY217 is a level 2 half-module, worth 10 credits, and is taught by Professor Vik Dhillon. The lecture notes and past exam papers for PHY217 can be found on the web at: http://www.vikdhillon.staff.shef.ac.uk/teaching.html. I will also be handing out a CD-ROM at the start of the course which contains a complete set of the lecture notes.

aims & objectives

This course aims to provide an understanding of the tools and techniques used by astronomers to study the Universe, with an emphasis on ground-based optical observations. Topics covered include: astronomical telescopes, astronomical instruments and electronic detectors. It builds on *Introduction to Astrophysics* (PHY104) and the topics covered in the first-year astronomy lab (PHY115 & PHY116: *Professional Skills in Physics and Astronomy I & II*). The module is designed to prepare students choosing to do observing projects in their third year (PHY319: *Astronomy Project*, including the La Palma field trip), fourth year (PHY480: *Research Project in Astronomy*), and those intending to spend a year abroad on La Palma (PHY473: *Working at the ING*; PHY474: *Extended Research Project in Astronomy*). As part of this module, all students must do an observing project using the University's 16-inch telescope. On successful completion of this course you should be able to:

- Assess the relative merits of different telescope and mounting designs, and different observing sites.
- Understand the effect of optical aberrations and the Earth's atmosphere on astronomical observations, and how they can be corrected.
- Understand the operating principles of imagers, photometers and spectrographs.
- Describe the operating principles of modern optical detectors.
- Follow the procedures required to reduce and calibrate astronomical data.
- Understand noise sources and predict the signal-to-noise ratio of an astronomical observation.

syllabus

This course is split into three main parts, *telescopes*, *instruments*, and *detectors*, which are presented in the same order that a photon from an astronomical source meets them.

telescopes

I. **introduction**

II. **telescope optics**
- i. **basic principles**
- ii. **refractors**
- iii. **reflectors**
- iv. **catadioptric telescopes**
- v. **visual use of telescopes**
- vi. **example problems**

III. **telescope mountings**
- i. **equatorial mountings**
- ii. **alt-azimuth mountings**
- iii. **coudé and nasmyth**
- iv. **tubes and trusses**

IV. **domes and sites**
- i. **the atmosphere**
- ii. **observatory sites**
- iii. **dome design**

V. **corrections**
- i. **autoguiding**
- ii. **active optics**
- iii. **adaptive optics**

VI. **removed from course: interferometry**

instruments

I. introduction

II. imagers
  i. simple imagers
  ii. focal reducers and extenders
  iii. re-imagers
  iv. sampling theory
  v. example problems

III. photometers
  i. single-pixel versus multi-pixel photometers
  ii. fluxes and magnitudes
  iii. photometric systems
  iv. extracting photometric data
  v. calibrating photometric data
  vi. example problems

IV. spectrographs
  i. the grating equation
  ii. basic spectrograph design
  iii. dispersion and spectral resolution
  iv. blazes and grisms
  v. free spectral range and order sorting
  vi. echelle spectrographs
  vii. removed from course: integral-field and multi-object spectrographs
  viii. removed from course: atmospheric dispersion
  ix. removed from course: reducing spectroscopic data
  x. removed from course: calibrating spectroscopic data
  xi. example problems

detectors

I. introduction

II. CCDs
  i. the physics of semi-conductors
  ii. the structure of CCDs
  iii. charge coupling
  iv. output electronics

---

text books

The on-line course notes provide all of the information you need for this course. If you wish to read around, however, I would recommend the following texts (roughly in decreasing order of importance):

- *To Measure the Sky* by Frederick R. Chromey (CUP), ISBN: 9780521747684. The book is currently in its first edition and costs around 37 pounds in paperback. There are 3 such texts in the Information Commons.

  This book, which came out quite recently, is probably the closest text you will find to the content of my course. It covers a great deal of the material I cover, and at a very similar level. Strongly recommended.

- *Electronic Imaging in Astronomy* by Ian S. McLean (Springer-Verlag), ISBN: 9783540765820. This book is currently in its second edition and costs around 63 pounds in both hardback and paperback. There are 3 such texts in the Information Commons.

  As its name implies, this book covers the detector aspects of my course, although it also has much useful information on instrumentation. The book tends to go into more detail than is required for PHY217, but it is written in a very readable style and I strongly recommend it.

- *Astronomy: Principles and Practice* by A. E. Roy and D. Clarke (Institute of Physics Publishing), ISBN: 9780750309172. This book is currently in its fourth edition and you should be able to find a new paperback copy for around 35 pounds. The Information Commons holds 6 copies of this book. Beware of older editions, as they don't cover modern detectors and instrumentation.

  This is an excellent book which will come in useful throughout your degree. It covers a much wider range of topics than my course, but the parts on telescopes are covered at just the right level of detail.

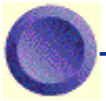
- *Astrophysical Techniques* by C. R. Kitchin (Taylor & Francis Inc), ISBN: 9781420082432. This is currently in its fifth edition, and is available in hardback for around 35 pounds. The Information Commons holds 3 copies of this book.

  This book covers a much wider range of material than my course, but unfortunately does not go into enough depth on some of the subjects that I require you to know. It may, however, be of use as a reference text.

---

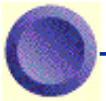
useful links

There are quite a few "Astronomical Techniques" courses with on-line resources. Here is a selection:



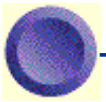
_ASTR 511: Astronomical Techniques

A graduate introductory course at the University of Virginia, by Robert W. O'Connell.



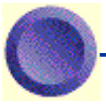
_Astronomical Techniques

An first-year undergraduate course at the University of Bristol, by Ben Maughan.



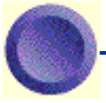
_Bill Keel's Lecture Notes: Astronomical Techniques

Notes for a senior/graduate student course at the University of Alabama, by Bill Keel.



_ASTR 306/406: Astronomical Techniques

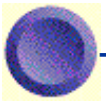
A final-year/graduate course at Case Western Reserve University, by Chris Mihos.



_big eyes

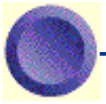
A league table of the world's largest optical telescopes,

both built and on the drawing board, with useful links to each observatory's web site.



_astronomy simulations and animations

A really useful selection of Flash animations and simulations for astronomy education, provided by the University of Nebraska-Lincoln. I'd strongly recommend you play around with these to get an intuitive feeling for some of the concepts involved in this course.



_QuCAM

Simon Tulloch's web site - a fantastic resource for all things CCD.

---

assessment

exam (80%): This will be sometime in weeks 13-15, exact date, time and place to be announced, and will be of 2 hours duration. You must answer question 1, which is worth 50% of the paper, and then two questions from a choice of four, each of which is worth 25%. Past exam papers can be found on the web at: http://www.vikdhillon.staff.shef.ac.uk/teaching.html

observing project (20%): Further details of the observing project, which is compulsory, can be found here.

observing project

As part of PHY217, you will be expected to complete a simple observing project using the 16-inch telescope on the roof of the Hicks Building. This project is designed to give you basic hands-on experience of astronomical observing and data reduction, and can be completed in a few hours of telescope time. You are encouraged to design your own project, but it is important to discuss the feasibility with me before starting your detailed planning. The observing must be done in groups of ***three*** students, so please find others interested in doing the same project as you. If you can't think of a project, or can't find a group to work with, I shall be happy to recommend some options. You must notify me by email of your final choice of project and partners by the deadline at the end of week 2: **Friday, 11 October 2013.** If you have not chosen a project and partners by then, you shall be assigned them by me! (For further information on choosing a project, please see below.)

There are 3 aspects to the observing project:

1. **Planning:** Well before your scheduled observing run, you must discuss with me or Paul Kerry (D16): which objects to observe, what time of night to observe, what filters you require, what sequence of exposures you require, etc. You will need to include a section on your planning in the final report.

2. **Observing:** Your observing sessions will be supervised by Paul Kerry and must be completed in a specified period: **Monday, 14 October - Friday, 13 December 2013 (weeks 3-11)**, although please note that observing is not usually possible over weekends and there may be short periods when Paul Kerry is unavailable.

   Sign-up sheets will be posted on the Astronomy Noticeboard outside the Astronomy Lab (E36), along with full instructions on how to contact Paul Kerry on the night. Although you should be able to complete all your observations in a single session, to allow for the vagaries of British weather we expect you to sign up for at least two

evenings per week until you have successfully completed your observing. If you cannot do this, you must discuss the problem with me or Paul Kerry before the start of the designated observing period, or as soon as the problem (e.g. illness) becomes apparent.

Attendance at the observing is compulsory - you will not receive any marks for the project if you fail to show up or, if the weather is bad for part of the specified observing period, you have not made every effort to sign up for other time slots. Note that, unless previously agreed with me or Paul Kerry, if you are unable to attend a successful observing session with the other members of your group, it will not be possible for you to observe at a later date on your own. Note also that no resit of the observing project is possible, so missing it will make it much more difficult to pass the module.

We strongly advise signing up for observing as soon as possible: students who fail this module tend to be those who leave signing up until the last minute and then suffer from poor weather at the end of the observing period. This is no excuse, as there are usually clear periods at the start of the observing period which no students sign up for. Only if the entire period is unusable, or if you have genuinely serious reasons as to why you could not do the observations (which in most cases must be supported by documentary evidence), will this component not count towards the final mark.

3. **Data reduction and report:** After you have obtained your observations, you will need to contact Paul Kerry to help you reduce and analyse your data using the computers and software available in the Astronomy Lab. Note that this element of the project, and the subsequent write-up, must be your own work - do not work in your observing groups.

   Your write-up must follow the same style as for a formal laboratory report. There must be sections describing the planning stage, the observations (a description of the equipment used, the observing conditions and the data that was taken), the data reduction and the data analysis. You will be penalised if you omit an analysis of the errors, and if you fail to compare your results with literature values. Please submit your reports to the departmental office by the deadline: **Thursday, 19 December 2013.** Note that this is the final week of term, and has been set to make the observing window as long as

possible. However, you will undoubtedly have other pieces of work to hand in around this time, so it is in your interests to complete your observing and hand in your report as early as possible in the semester.

Choosing a project

You are free to observe any object you wish. However, it is important to note the following limitations.

- You will only be able to use the imager, not the spectrograph. Hence, you will only be able to measure the brightness and colours of objects, and how they vary with time. The CCD camera on the 16-inch telescope has UBVRI filters and a field of view of approximately 18' x 12'.

- You will not be able to observe for more than a single session of about 4 hours. Hence, you will not be able to monitor the variability of an object with a period substantially longer than this, unless you are attempting to observe a specific event (e.g. the transit of an extrasolar planet).

- It is only a 16-inch telescope, mounted in the centre of a large city. Hence, you will not be able to observe particularly faint objects - it is recommended that you do not attempt to observe objects much fainter than about 15th magnitude.

- It is important that you can, in principle, make some kind of simple measurement from your data, i.e. taking pretty pictures of a galaxy just for the sake of it is not acceptable, but measuring the H-R diagram of an open cluster to determine the turn-off position is acceptable.

Some examples of the data obtained for previous projects are given here. Before you email me with your final choice of project, it is imperative that you come to see me to discuss your ideas so that I can confirm with you that the project is feasible. Past experience suggests that the best projects tend to be one of the following, although we are always keen for students to show initiative and come up with their own ideas for projects:

- **Hertzsprung-Russell (HR) diagram of an open cluster.** The aim here is to measure the distance and age of an open cluster. A list of open clusters is available here.

  Make sure that you pick a cluster that is visible from Sheffield at the start of the night. You can do this using *The Sky* in the astro lab, Stellarium on your own PC, or the on-line ING Object Visibility page (where you must enter the longitude and latitude of Sheffield in the following format: 358 30 51 53 22 50 185).

  The cluster you select must also be small enough so that the majority of the cluster fits within the 18' x 12' field of view of the CCD, i.e. don't select one much larger than 20' in diameter. However, don't pick one that is too compact either, as the individual stars in the cluster will be difficult to resolve. The cluster should also have a reasonably large number of stars (definitely greater than 50; greater than 100 would be best). Finally, the cluster should not be too distant or reddened and hence faint, and must be of sufficient age to show a relatively clear main-sequence turn off. The latter two items can be checked by inspecting existing HR diagrams of the cluster using WEBDA (simply enter the name of your chosen cluster in the *Display the Page of the Cluster* box).

  You can download isochrones computed by the University of Padova from here. To give you a good first guess at which isochrone is likely to fit best, go to the WEBDA page for your cluster and click on "*General menu for Isochrone plots (basic)*" - it is recommended you choose Padova isochrones with Solar metallicity (Z=0.019). You can plot an HR diagram for your cluster on this page, and then overplot the best-fit isochrone.

  When you construct an observed HR diagram, you must correct for both atmospheric extinction and interstellar extinction. The atmospheric extinction correction can be made by assuming standard values for the extinction coefficient, as given in table 2, and then transforming all of your measured magnitudes to above-atmosphere values. Be careful, however, as it is possible that you will have already corrected for atmospheric extinction if you used a photometric zero point determined from one of the cluster stars. To correct for interstellar extinction, you must use the formula $(B\text{-}V)_0 = (B\text{-}V) - E(B\text{-}V)$, where $(B\text{-}V)_0$ is the intrinsic colour index of the cluster (i.e.

corrected for interstellar extinction), *(B-V)* is your observed (i.e. uncorrected) colour index, and *E(B-V)* is the colour excess (or *reddening*) in magnitudes. Hence you will find that you will have to shift your data in the *x*-direction on the HR diagram in order to align it with the isochrone, and the value you shift it by is equal to the reddening, *E(B-V)*. Once you have determined *E(B-V)* in this way (and checked it against the value given by WEBDA), you will then have to correct your *V*-band apparent magnitudes using the equation: $V_0 = V - A_V$, where V_0 is the intrinsic *V*-band apparent magnitude of the cluster (i.e. corrected for interstellar extinction), *V* is your observed (i.e. uncorrected) *V*-band apparent magnitude, and A_V is the *visual extinction* in magnitudes. The ratio $A_V / E(B-V)$ is usually denoted by the symbol R_V and a generic value for our galaxy covering a large wavelength range is $R_V = 3.2 \pm 0.2$. Hence the formula to correct your *V*-band apparent magnitudes becomes: $V_0 = V - [R_V \times E(B-V)] = V - [(3.2 \pm 0.2) \times E(B-V)]$. Once you have corrected the *y*-axis of your data in this way, the difference between the isochrone and your data in the *y*-direction on the HR diagram will give you the distance modulus of the cluster.

To determine the age of your cluster, you will have to download a series of isochrones of different ages, using the WEBDA value for the age as a guide. The isochrone which best matches the main-sequence turn-off point gives the age of the cluster. It is likely you will find that the determination of the interstellar extinction, distance and age of the cluster will be an iterative process.

- **Light curve of a delta Scuti star.** The aim here is to measure the period and hence distance of a delta Scuti star. A list of delta Scuti variables is available here.

  Make sure that you pick a delta Scuti star that is visible for at least one orbital period around the start of the night - see the description of how to do this in the HR-diagram project above.

  The delta Scuti star you select must have a magnitude of less than V~15, and the brighter the better. The orbital period must be less than ~4 hours, i.e. ~0.17 days. The amplitude of the pulsation must be as great as possible. The lower limit is dependent on the brightness of the target you select, but I would avoid objects with

pulsation amplitudes of less than, say, 20%, i.e. ~0.2 magnitudes.

Once you have obtained your light curve, you must attempt to estimate the period of the pulsation in days. The simplest way of doing this would be, of course, to measure the separation between two repeated features in your light curve, such as two consecutive peaks or troughs. However, this would effectively ignore the rest of the data you have obtained, and so it is hoped that you will employ a more sophisticated approach in the period determination. You should then use your measured period to estimate the absolute *V*-band magnitude using the period-luminosity relations for delta Scuti stars given by Petersen and Hog (1998) and/or McNamara, Clementini and Marconi (2007). You can then use this absolute *V*-band magnitude in conjunction with your measured (mean) apparent *V*-band magnitude (so make sure that you observe with the *V* filter!) to derive the distance to the star in parsecs via the distance modulus equation.

- **Light curve of an eclipsing cataclysmic variable star.** The aim here is to use the eclipse width to measure the radius of the accretion disc. A list of cataclysmic variables (CVs) is available here.

  Make sure that you pick a CV that is visible for at least one orbital period around the start of the night - see the description of how to do this in the HR project above.

  The CV you select must be eclipsing, i.e. it must have an *EB* entry in the table of 1 or 2. The quiescent, out-of-eclipse brightness, indicated by the *Mag1* entry, must be less than V~15, and the brighter the better. The orbital period must be less than ~4 hours, i.e. an *Orb. Per.* value of less than ~0.17 days.

  Once you have obtained your light curve, you should measure the width of the eclipse in units of days. Dividing this by the orbital period in days will convert the eclipse width to phase units, i.e. the fraction of the orbit spent in eclipse, and dividing this by two will give the half-width of the eclipse. You should then use your eclipse half-width in phase units in conjunction with "*Method 1*" outlined by Harrop-Allin and Warner (1996) to plot the disc radius as a function of inclination and mass ratio. You can then use this plot to select the most likely range of values of the disc radius.

- **Light curve of an asteroid.** The aim here is to measure the rotation period of an asteroid and determine if the asteroid is likely to be a rubble pile or a solid body. A list of minor planet light curve parameters is given here, and a version of this list ordered by period is given here.

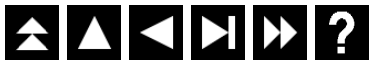
  Using the above list, you should select asteroids which have rotation periods of less than ~4 hours. The amplitude of the variability, given by the *Variation* column, must be as great as possible. The lower limit is dependent on the brightness of the target you select, but I would avoid objects with amplitudes of less than, say, 20%, i.e. ~0.2 magnitudes.

  To determine the magnitude, right ascension and declination of the targets, which are all time dependent, you need to enter the names of the asteroids you have selected in the large box at the centre of the Minor Planet and Comet Ephemeris Service, and then click on the *Get ephemerides/HTML page* button. The asteroid you select must have a magnitude of less than V~15, and the brighter the better. Make sure that you pick an asteroid that is visible for at least one orbital period around the start of the night - see the description of how to do this in the HR project above.

  Once you have obtained your light curve, you should attempt to measure its rotation rate. There are various ways in which this can be done: visual inspection of your light curve; folding your data on trial periods; Fourier analysis, etc. Once you have determined the rotation rate, you should compare this to the maximum rate assuming it is a rubble pile. A basic discussion of this topic and links to the literature can be found here. Note that the correct physical description of the situation is that gravity supplies the centripetal force required to keep the rubble in circular motion. As the rotation rate increases, the required centripetal force increases until a point comes when gravity is no longer able to provide it, and the object flies apart. You should not have to mention the term "centrifugal force" in your explanation, which is an imaginary force.

---

timetable



lectures: 20 lectures in total.

10 in weeks 1-5 and 8-12 on Tuesdays at 10:00-10:50 in Lecture Theatre 9 of the Hicks Building

AND

10 in weeks 1-5 and weeks 8-12 on Fridays at 09:00-09:50 in Lecture Theatre 4 of the Hicks Building.

Note that there are no lectures in week 6 as I shall be away commissioning an instrument in Thailand. Week 7 is a reading week.

labs: There are no day-time labs associated with PHY217, although all students must do an observing project, as detailed here.

problems classes: Weeks 2-6 and 8-11 on Thursdays at 11:00-11:50 in room E36 (the Astronomy Lab) of the Hicks Building

OR

Weeks 2-6 and 8-11 on Fridays at 11:00-11:50 in room E36 (the Astronomy Lab) of the Hicks Building.

Note that there are no problems classes in week 1 (October 3 and 4) or week 12 (December 19 and 20). Week 7 is a reading week.

The problems classes will be led by Dr Stuart Littlefair, who is Year Tutor for Second-Year Astronomy, assisted by two PhD students. You need only attend one problems class each week - you will be told which one you are assigned to (Thursday or

Friday) at the start of the semester.

---

contacting me

If you need help with any aspects of this course, please feel free to email me at this address: **vik.dhillon@sheffield.ac.uk**.

If you would prefer to see me in person, my office is E40 in the Hicks Building.

astronomical techniques

telescopes

telescopes: introduction

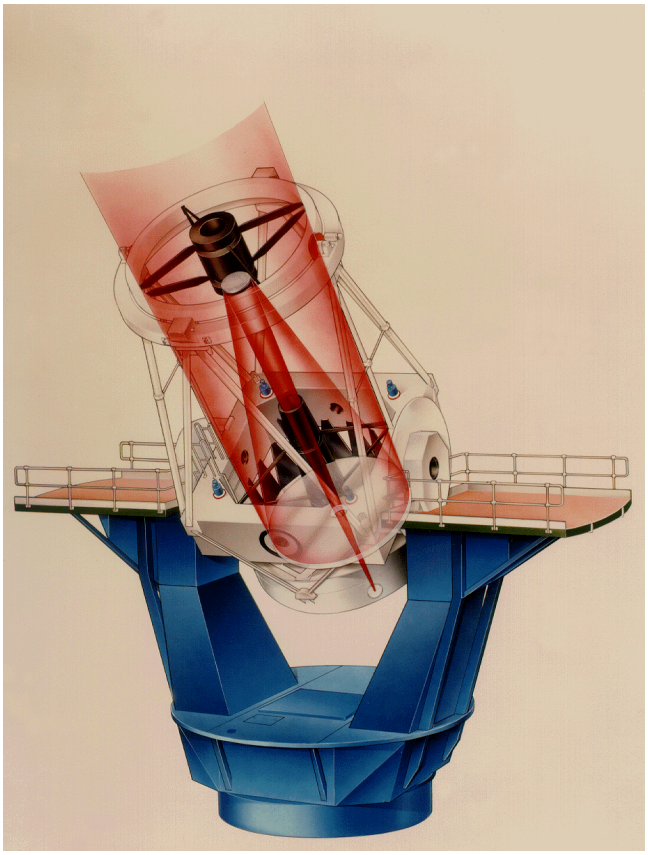
The question of who invented the telescope is a controversial one. It is usually credited to the Dutch optician Hans Lippershey in 1608, although recent research suggests that a form of telescope may have been discovered by the Englishman Leonard Digges around 1550. However, one thing is clear - the first person to systematically use a telescope for astronomy was the Italian scientist Galileo Galilei in 1610.

The purposes of a telescope are two-fold:

1. to allow the collection of energy (i.e. photons) over a larger area so that fainter objects can be studied;
2. to provide a higher angular resolution so that objects can be studied in greater spatial detail.

We shall begin this part of the course by studying the principles of telescope optics, before moving on to look at the various different types of astronomical telescope, how they are mounted, and where they are sited.

telescopes



II. telescope optics

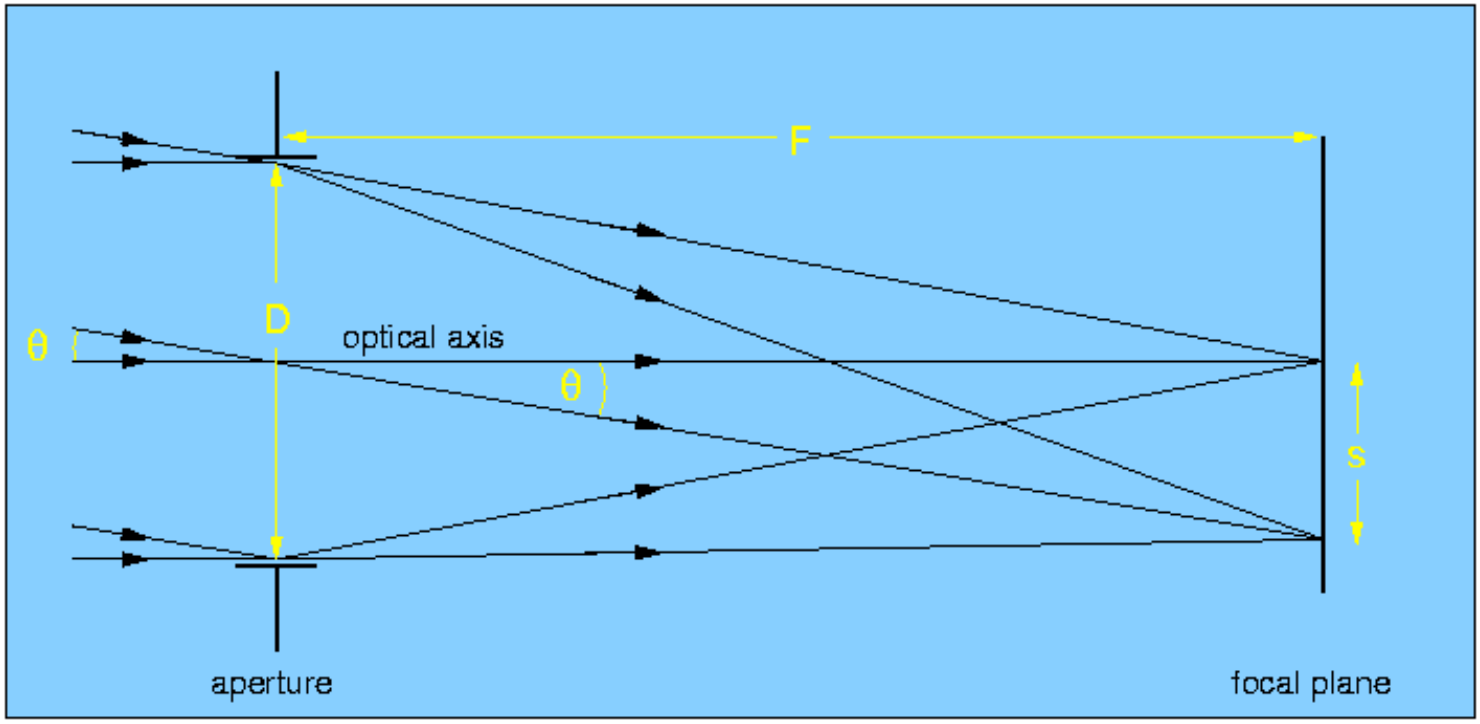
basic principles

aperture, focal length, plate scale and focal ratio

Figure 1 shows the basic function of a telescope. At the left, light from two distant sources arrives at the telescope. As all astronomical sources are effectively at an infinite distance compared to the dimensions of a telescope, the rays from each source can be assumed to be parallel.

The light is collected by an *aperture* of diameter *D* which forms an image of the two sources at the *focal plane*. The line through the centre of the aperture and the focal plane, and orthogonal to them, defines the *optical axis* of the system. Since the sources are at infinity, the *focal length, F,* of the telescope is defined by the distance between the focal plane and the aperture. In astronomical telescopes, the aperture is usually a lens or a mirror, and is circular, hence the amount of light collected is proportional to D^2. The aperture diameter is the single most important parameter affecting the performance of a telescope and hence most telescopes are usually referred to in terms of it, e.g. the ESO 3.6m telescope in Chile, where the 3.6m is the diameter of its primary mirror.

figure 1: A schematic showing the function of a telescope. The "aperture" in the figure represents an optical element, such as a mirror or lens, with the ability to bring light to a focus.



Two concepts which are extensively used by astronomers when describing the specifications of a telescope are the *plate scale* and the *focal ratio*:

- *Plate scale, p*: This relates the size of the image in the focal plane, *s*, to its angular size on the sky, θ. Given that θ is usually small, figure 1 shows that these two quantities can be related by the equation

  $s = F\theta$,

  where θ is measured in radians. This equation shows that if an astronomer wishes to study an object of angular size θ in high spatial detail, *s* needs to be large, and therefore a long focal-length telescope is required.

  Astronomers usually refer to the plate scale in units of arcseconds per mm. As shown in the example problems, the plate scale is then given by

  $p = 206265 / F$,

  where *F* is in mm.

- *Focal ratio, f*: This is defined as the ratio of the focal length of the telescope to its diameter, i.e.

$f = F / D$.

The term is often used in photography, where it refers to the *speed* of the camera. The larger the focal ratio, the "slower" the camera, as the amount of light falling on a given area of the focal plane is smaller. This can be understood in two ways: for larger F at fixed D, the plate scale is smaller and hence the light is spread out over a larger area; for smaller D at fixed F, less light reaches the given area on the focal plane. For a "faster" camera, the reverse is true.

resolving power

The *resolving power* (or *resolution*) of an astronomical telescope is a measure of its ability to distinguish fine detail in an image of a source. Aberrations due to the optical design or flaws in the manufacture and alignment of the optical components can degrade the resolving power, as does peering through the Earth's turbulent atmosphere. We shall come back to address these issues in later lectures. However, even if a telescope is optically perfect and is operated in a vacuum, there is still a fundamental lower limit to the resolving power it can achieve. This is known as the *theoretical resolving power*, and in this section we shall explore its origins and consequences.

Figure 2 shows a point source at infinity, i.e. a star radiating light into space. Considering the light as a wave, it is possible to define *wavefronts*, which are points of the wave at equal phase. As the wavefronts propagate from the star at the speed of light, their curvature decreases until they can be considered plane-parallel waves when they arrive at the telescope. A crude analogy would be to consider the waves created when a stone is thrown into the centre of a pond. Close to where the stone entered the water, the wavefronts would be clearly defined circles, whereas by the time they reach the edge of the pond, they would be more or less straight and parallel to the edge of the pond.

figure 2: Schematic showing wavefronts radiating from a point source at infinity and imaged by a telescope. The "aperture" in the figure represents an optical element, such as a mirror or lens, with the ability to bring light to a focus.

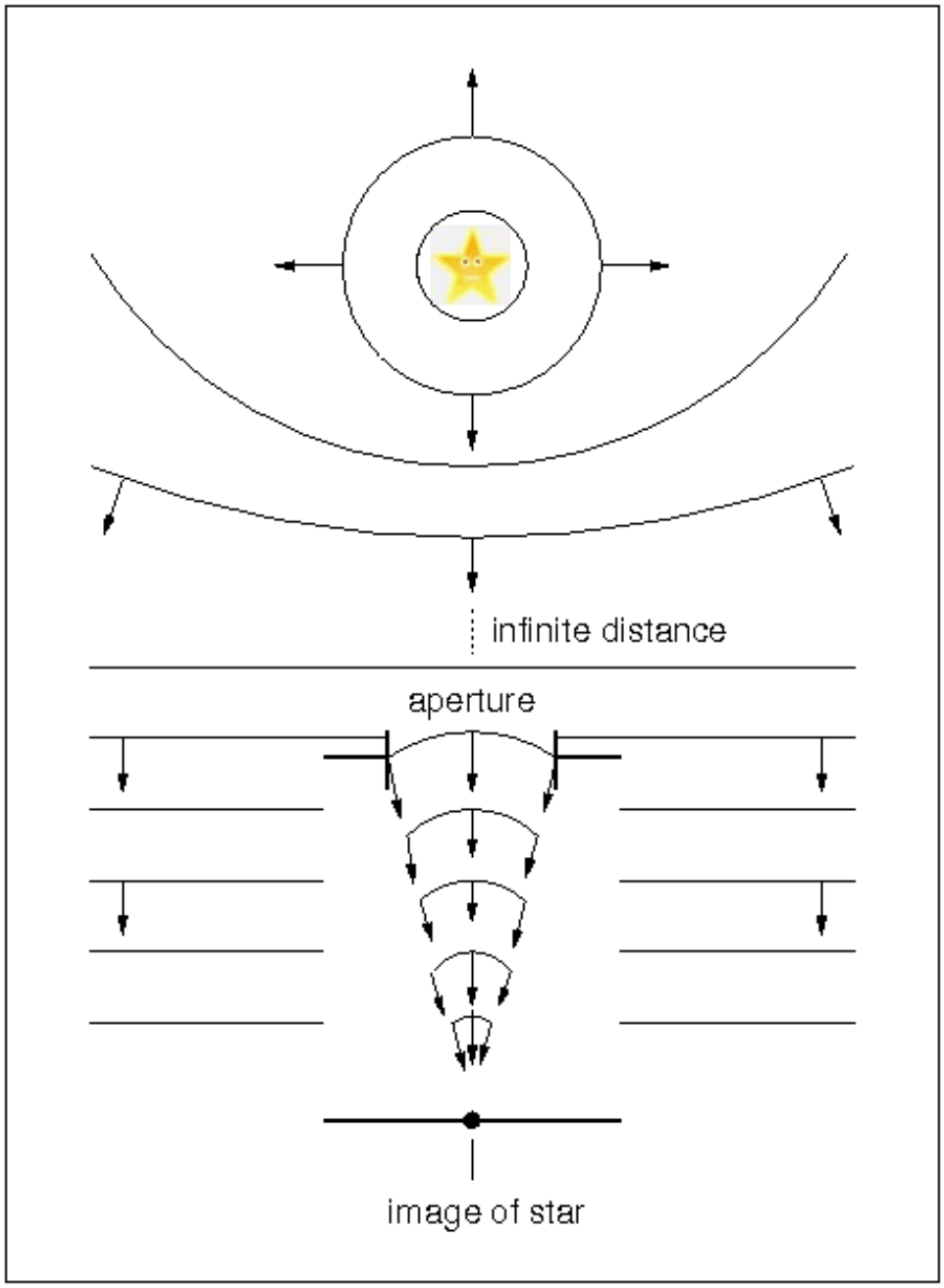
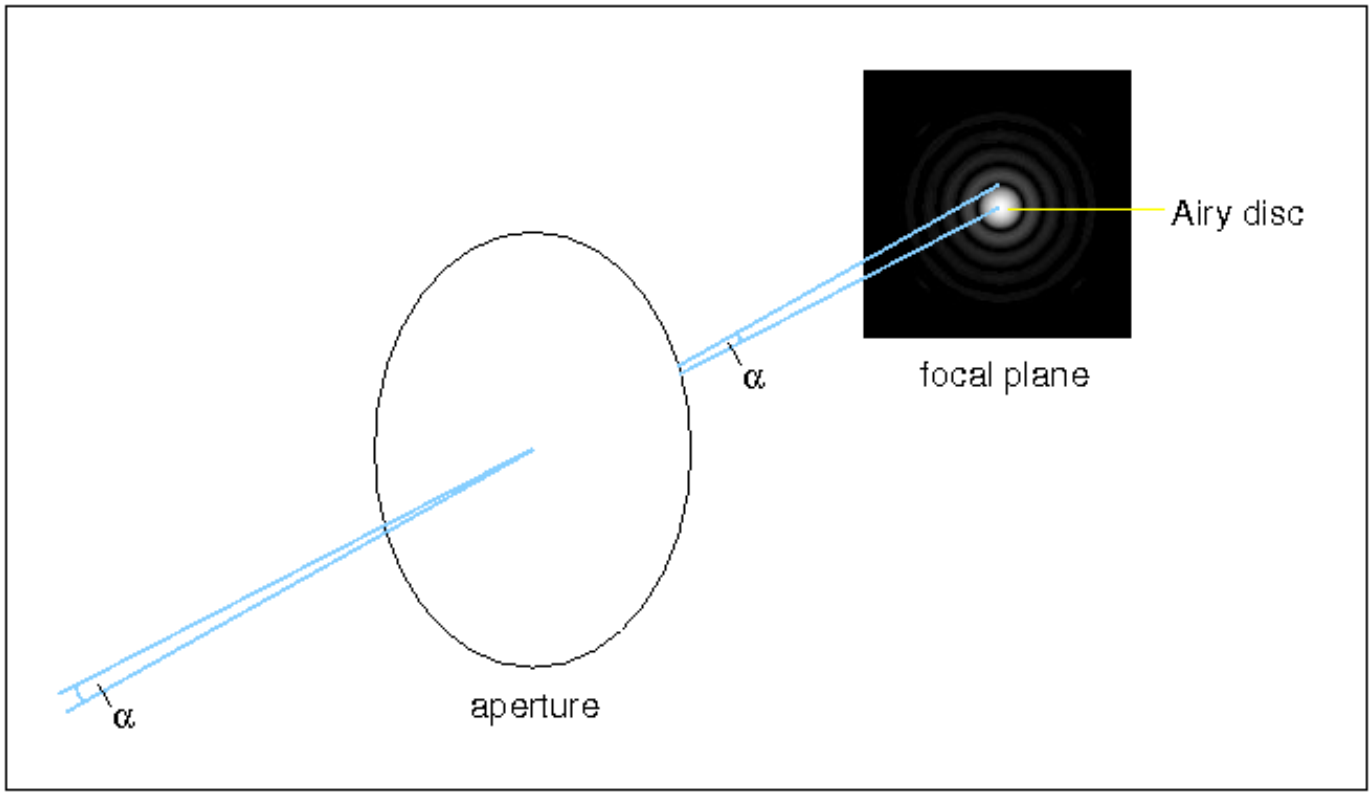


Figure 2 also shows that the telescope brings the plane-parallel waves incident upon the aperture into focus, forming an image of the star in the focal plane. It does this by inducing a phase change on the wavefront which varies across the telescope aperture. However, since the aperture does not cover the entire wavefront radiated by the star, but only a very small portion of it, a diffraction pattern is produced.

The diffraction pattern produced when imaging a point source with a lens-based telescope is shown in figure 3. The image appears as a spot surrounded by concentric rings which decrease in brightness with increasing distance from the centre. The bright central spot, known as the *Airy disc* after the British astronomer who first studied it, is theoretically predicted to

contain 84% of the light, and the first ring contains less than 2%.

figure 3: Schematic showing the diffraction pattern produced in the focal plane of a telescope when imaging a point source. α is the angle between the centre of the Airy disc and the first minimum and denotes the theoretical resolving power of the telescope.

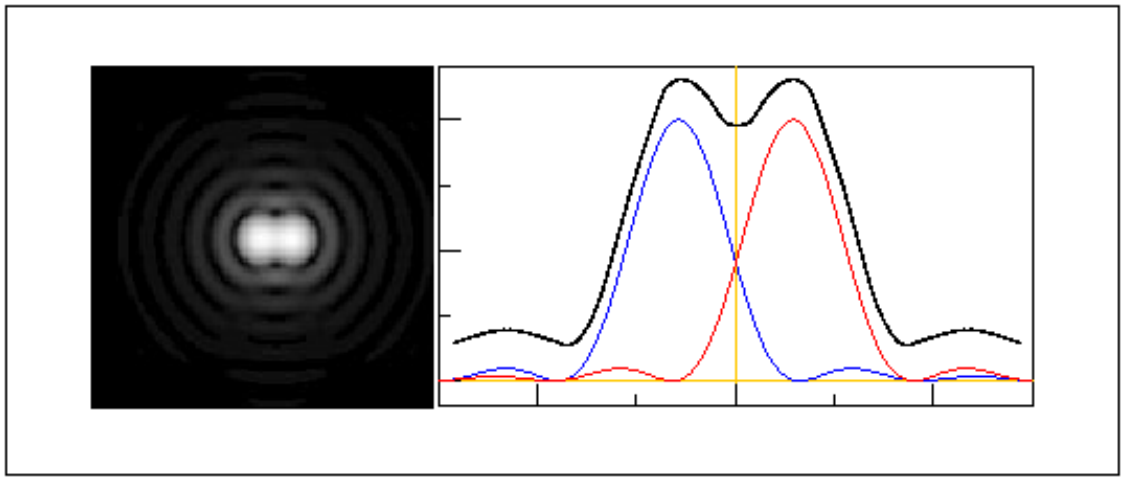


The size of the Airy disc puts a limit on the resolving power of a telescope. According to *Rayleigh's criterion* for resolution, two point sources are said to be just resolved when the centre of one Airy disc falls on the first minimum of the other diffraction pattern. This results in a 20% drop in intensity between the maxima, as illustrated in figure 4. An expression for Rayleigh's criterion can be obtained from theory by calculating the positions of minima of intensity in a point-source diffraction pattern. For the first minimum, this gives

$$\alpha = 1.22\lambda / D,$$

where λ is the wavelength of light and α is measured in radians and is the angle subtended at the aperture by the centre of the Airy disc and its first minimum in the focal plane (see figure 3). α is commonly referred to as the *theoretical resolving power* or *diffraction-limited resolution* of a telescope.

figure 4: Left: image of the overlapping diffraction patterns from two point sources separated by an angle α. Right: The black curve shows a cut through the image on the left. According to Rayleigh's criterion, the two sources are just resolved.



Rayleigh's criterion shows that to increase the resolving power of a telescope it is necessary to observe at shorter wavelengths and/or increase the diameter of the aperture. For small amateur telescopes, Rayleigh's criterion provides a realistic estimate of the resolution obtained, as calculated in the example problems. For large research telescopes, however, the resolution is dominated not by diffraction but by the *seeing*. We shall return to this topic later.

Having covered the basic principles of telescope optics, we turn now to look at the various different types of telescope design.

---

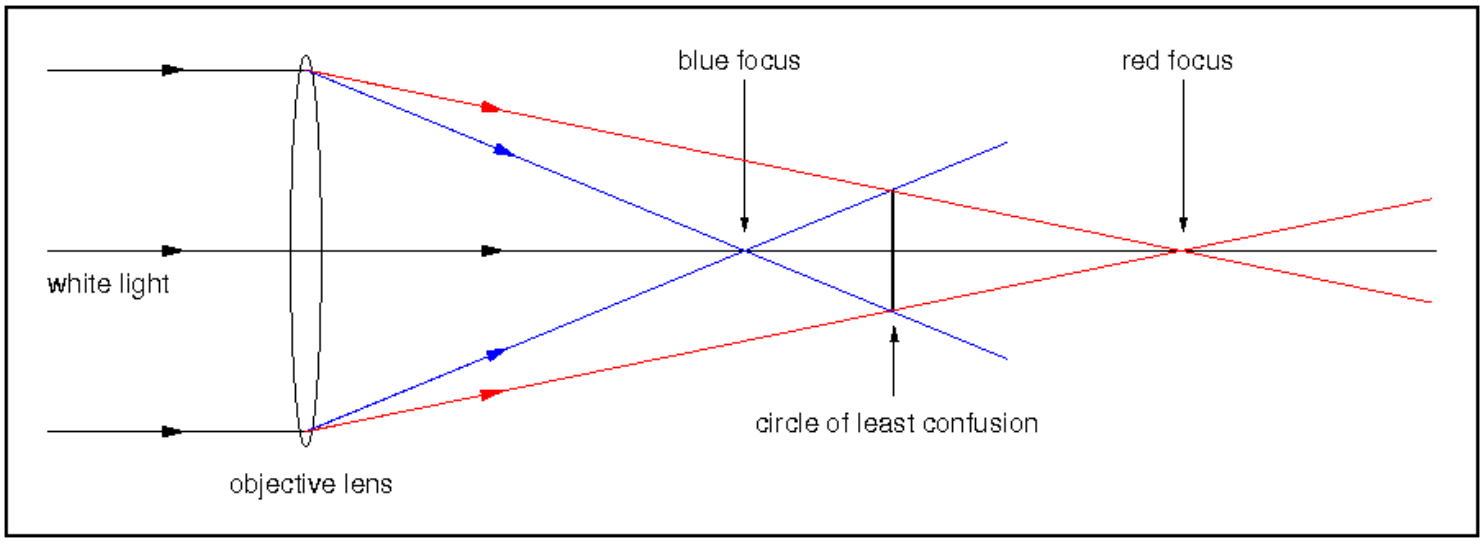
refractors

A *refractor* focuses the light from an astronomical source using a lens, known as the *objective*. Galileo's telescope, for example, was a refractor. A photograph of a replica of Galileo's refractor alongside a modern-day equivalent is shown in figure 5.

figure 5: Photograph showing a replica of Galileo's refractor alongside a modern-day equivalent.

The objective lens is usually made of glass and in its simplest form has a bi-convex shape where both surfaces are sections of a sphere. The lens collects the parallel light from an astronomical source and forms an image of it, as shown in figure 6.

figure 6: Schematic of a simple refracting telescope, illustrating the effect of chromatic aberration.

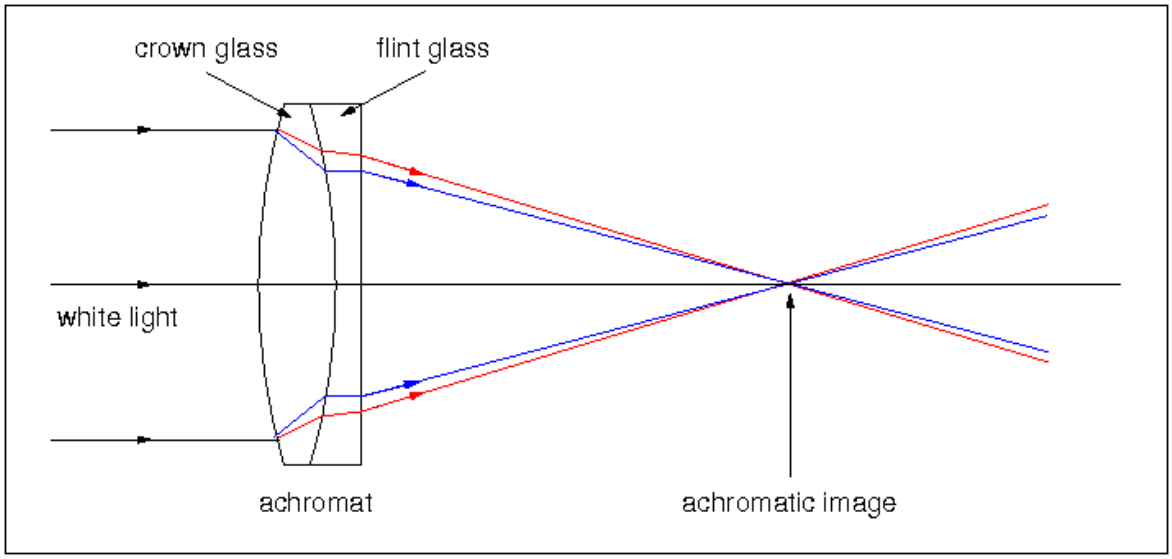


Unfortunately, the image produced by a simple refractor suffers from a major defect, known as *chromatic aberration*. This aberration is the largest to affect astronomical telescopes. Chromatic aberration arises because the glasses used to make lenses typically have refractive indices which decrease as the wavelength of light increases, resulting in the focal length of the objective increasing as the wavelength increases. Hence there is no single focal plane in which light of all wavelengths from a star is in focus, as shown in figure 6. Moving the focal plane towards the lens will provide a point-like blue image surrounded by a red halo. Similarly, moving the focal plane away from the lens will result in a point-like red image surrounded by a blue halo, as shown in figure 7.

figure 7: Photograph of the Lagoon Nebula, showing the blue halos due to chromatic aberration. The bright star at the top is probably quite red, and hence lacks a halo.

It is possible to minimise the effects of chromatic aberration in a given telescope by positioning the focal plane between these two extremes. The resulting image is known as the *circle of least confusion* and its position is shown in figure 6. It is the point of best focus, where the image of a star would appear as a filled circle. This is still a relatively poor image, however, and for a long time it was believed to be impossible to correct chromatic aberration any further. Then, in 1758, the English optician John Dolland patented the *achromatic doublet* (or *achromat*). The achromat is able to eliminate dispersion, i.e. the fact that each colour is deviated by a different amount, whilst remaining convergent. It does this by using two lenses made of different materials, typically crown and flint glass. The positive crown-glass lens strongly converges the light, but also disperses it. A less-powerful negative flint-glass lens is thus cemented to it which has a higher dispersion (i.e. exhibits a larger spread in refractive index with wavelength) and is able to undisperse the light without completely negating the overall convergence of the beam. The situation is shown schematically in figure 8.

figure 8: Schematic showing how an achromatic doublet can correct for chromatic aberration by bringing light of different colours to the same focus.



An achromatic doublet can only make the focus the same for two wavelengths and there is residual chromatic aberration at intermediate wavelengths. Nevertheless, achromats deliver order-of-magnitude reductions in chromatic aberration. Further improvement is possible by adding a third lens. These are called *apochromats*, and they can bring three wavelengths (e.g. red, green and blue) to the same focus. Even greater improvements can be obtained by adding more lenses, so-called *super-apochromats*, but these become economically unviable for the large apertures required for telescope objectives. An example of a top-of-the-range 4-inch (100mm) apochromatic refractor for the amateur market, costing about £3,000, is shown in figure 9.

figure 9: The 4-inch Takahashi FSQ-106 apochromatic refractor.

One implication of using achromats to minimise aberrations is that the focal length becomes quite long. Hence refracting telescopes are generally quite slow, with typical focal ratios of about 20. The plate scale is therefore quite small, which spreads the image of a point source over a large area in the focal plane. This makes refractors unsuitable for faint objects, but ideal for planetary observations and astrometry (i.e. measuring the positions of stars).

Refractors are relatively stable against changes in temperature during the course of a night, as changes in the optical properties of the front surface of the objective lens due to expansion/contraction tend to be cancelled out by the back surface. The sealed tube and use of a lens (rather than a mirror) also mean that refractors tend to require little maintenance and hence are optically relatively stable. Despite these advantages, modern professional telescopes are all reflectors. This is due to the cost of making large apochromatic lenses (and even then, there is still some residual chromatic aberration) and the length of the tube required to accommodate the long focal length of the objective (which has a knock-on effect on the size and cost of the dome). The most important reasons of all, however, are insurmountable: first, the lens in a refractor has to be held by its edge, and as the lens becomes larger it starts sagging under its own weight and hence distorts the image; second, as the diameter of the lens increases, so does its thickness and thus so does the amount of light absorbed in the glass. As a result, the largest refractor currently in existence is the 40-inch (1m) telescope at Yerkes Observatory near Chicago, built in 1897. Figure 10 shows the immense scale of the telescope tube and dome - it is unlikely that a bigger refractor will ever be built.

figure 10: A photograph of the 40-inch, f/19 Yerkes refractor.

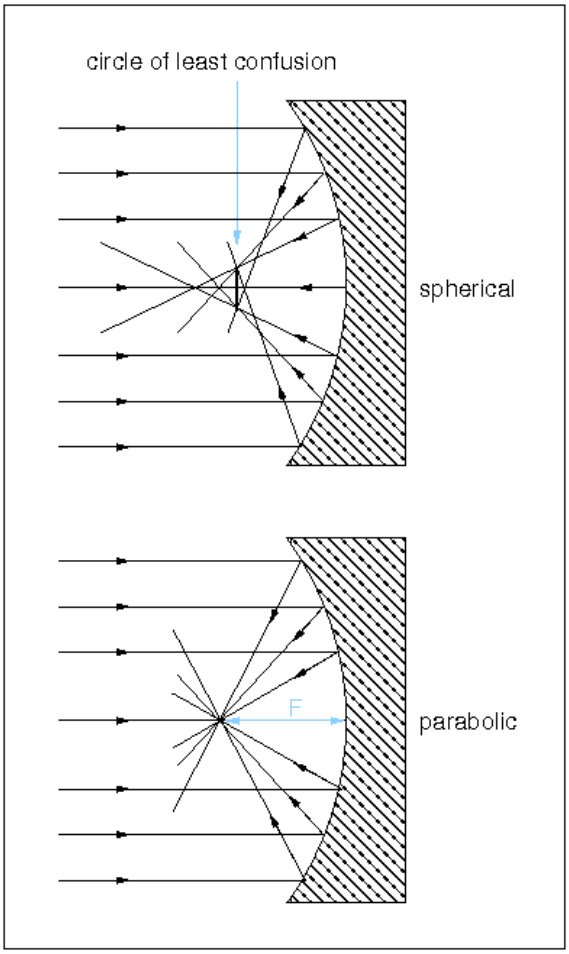
reflectors

A spherical concave mirror can be used to collect the light from an astronomical object and image it, as shown in figure 11. Since mirrors use reflection rather than refraction to form an image, they are inherently free of the most destructive aberration of all - chromatic aberration. Mirrors are also simpler than lenses in that they have only one optical surface. The optical properties of the substrate are unimportant: mirrors are usually made of a rigid, hard (i.e. polishable) material with a low thermal expansion coefficient (such as the glass Pyrex or the glass-ceramic Zerodur), and coated with a thin layer of aluminium, silver or gold to give high reflectivity. A telescope which uses a mirror to collect and focus light is known as a *reflector*.

spherical aberration

Unfortunately, mirrors are not aberration free. The top panel of figure 11 shows that the rays from the edge of a spherical mirror come to a focus nearer the mirror than do rays from the centre of the mirror, causing a point source to be imaged as a blurred disc. This defect is known as *spherical aberration*.

figure 11: The top panel is a schematic showing how a spherical concave mirror can collect light and bring it to a focus. The resulting image suffers from spherical aberration, but this can be removed using a parabolic mirror, as shown in the bottom panel.

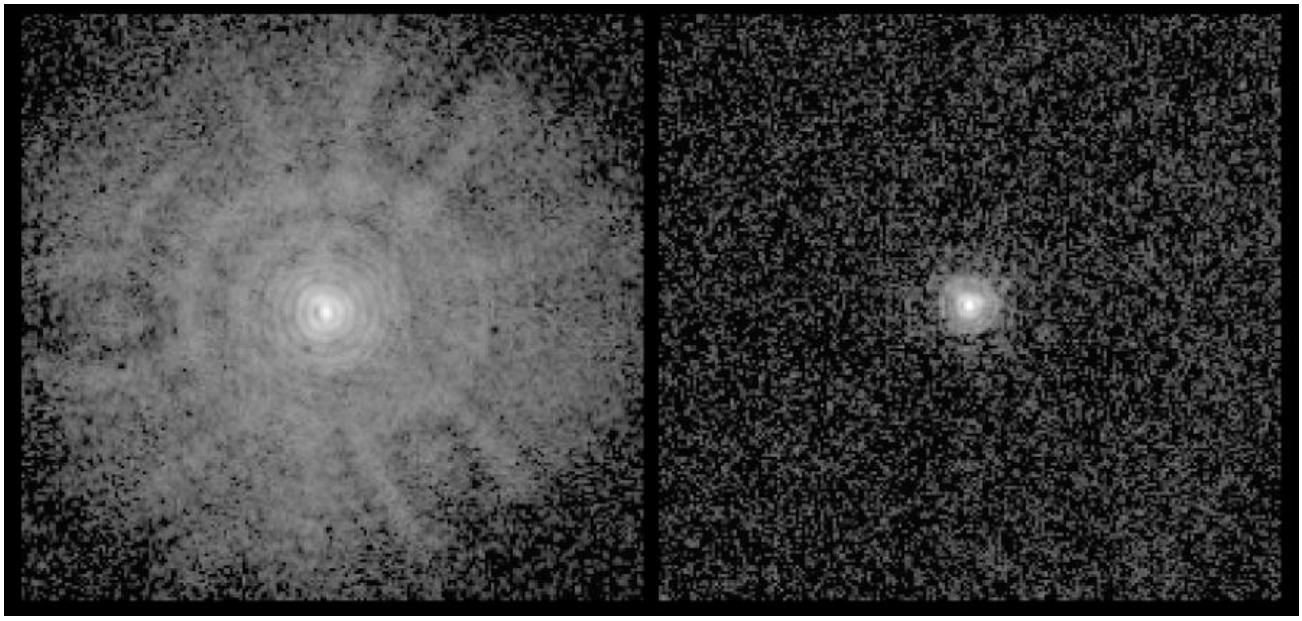


The effects of spherical aberration in mirrors can be reduced by positioning the focal plane coincident with the circle of least confusion, as shown in figure 11, but this still leaves an unsatisfactory blurring of the image.

However, the Scottish astronomer James Gregory realised in 1663 that spherical aberration can be eliminated if the curve of the surface of a mirror is a parabola. This is also shown in figure 11, where all rays parallel to the axis of a parabolic mirror are reflected to meet at the focus of the parabola.

Creating mirrors with parabolic surfaces is not straightforward, as special grinding and polishing techniques are required in order to create the deeper central depression of a parabola compared to a spherical surface. A famous example of spherical aberration is given by the Hubble Space Telescope (HST), which suffered from spherical aberration due to a mistake during the manufacture of its (hyperbolic) 2.4m mirror, as shown in figure 12. Corrective optics were later installed by astronauts on a space shuttle servicing mission and the telescope is now functioning perfectly.

figure 12: Left: Image of a star taken with the HST, showing the blurred, circularly-symmetric image characteristic of spherical aberration. Right: Image of a star taken with the corrected HST, showing just the Airy disc and diffraction rings, indicating that the spherical aberration has been successfully removed: the image of the star on the left covers approximately 2 arcseconds, whereas that on the right covers only 0.05 arcseconds.



It should be noted that lenses, and hence refractors, suffer from spherical aberration in exactly the same way as mirrors, but it is a much smaller effect than chromatic aberration. Spherical aberration can be eliminated from refractors using *aspheric* lenses, in which the curvature of the surfaces is not constant, but these are difficult and expensive to produce. It is more common to remove spherical aberration from refractors by designing achromats in such a way that the negative lens cancels out the spherical aberration introduced by the positive lens, whilst still correcting for chromatic aberration. [It is even possible to design achromats which also correct for coma (see below) - so called *aplanatic* lenses].

coma, astigmatism, curvature of field and distortion of field

Until now, we have been considering only *on-axis* aberrations, such as chromatic and spherical aberration. These are the aberrations suffered when imaging a point source in the centre of the field of view. There also exist *off-axis* aberrations, such as *coma*, *astigmatism*, *curvature of field* and *distortion of field*. These are caused when the light rays are not parallel to the optical axis, i.e. when imaging objects towards the edge of the field of view. We shall look at each of these aberrations in turn, before going on to describe the various different types of reflecting telescopes that have been designed to combat them.

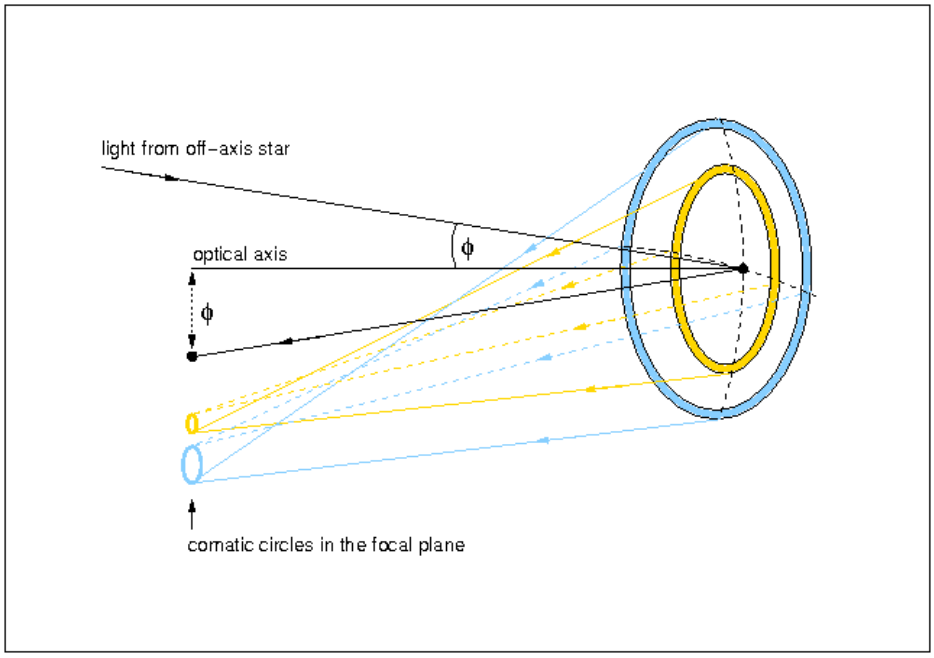
It is important to note that all four of these off-axis aberrations affect both mirrors and lenses, and hence both refracting and reflecting telescopes. It should also be noted that all the aberrations we look at in this course are inherent to the lens or mirror; there exists another whole class of aberrations which result due to poorly manufactured, badly mounted and mis-aligned optics which we shall not consider here.

The largest off-axis aberration is *coma*. It derives its name from the comet-like appearance of the resulting images, as shown in figure 13. At the centre of the focal plane, the images of the stars are point-like. Moving towards the edge of the image, however, the stellar images become increasing comet-like, with the faint, fanned "tail" always pointing away from the centre of the field.

figure 13: An example of coma. The lower image shows an enlarged view of the bottom left-hand corner of the upper image.

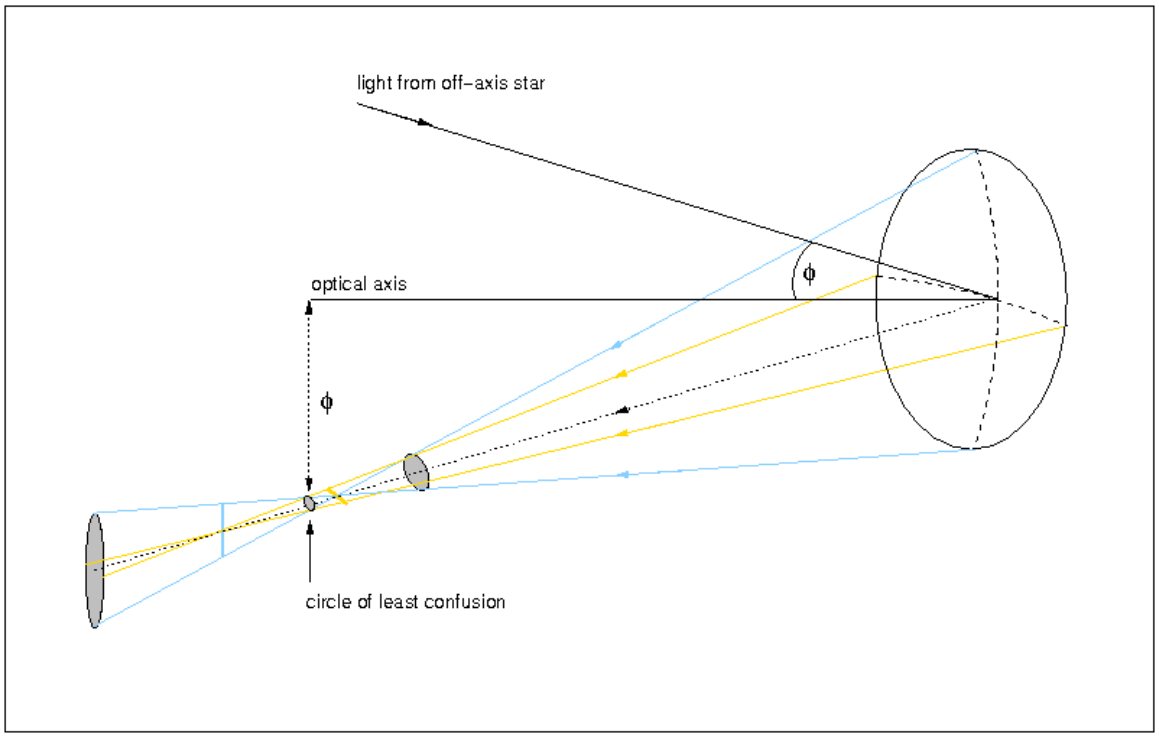
We can understand how coma arises by considering the mirror as a series of annuli, as shown in figure 14. If the mirror is used to image an off-axis star, each annulus will form a separate annular image, or *comatic circle*, of the star in the focal plane. Each point on the annular image corresponds to the meeting of two rays from diametrically opposite points of the corresponding annulus on the mirror, with smaller mirror annuli forming smaller comatic circles in the focal plane. The result of these overlapping comatic circles is a comet-like image like the ones shown in figure 13. The size of comatic aberration increases linearly with the angle ϕ between the incoming ray and the optical axis.

figure 14: A schematic showing how coma is produced by a parabolic mirror.



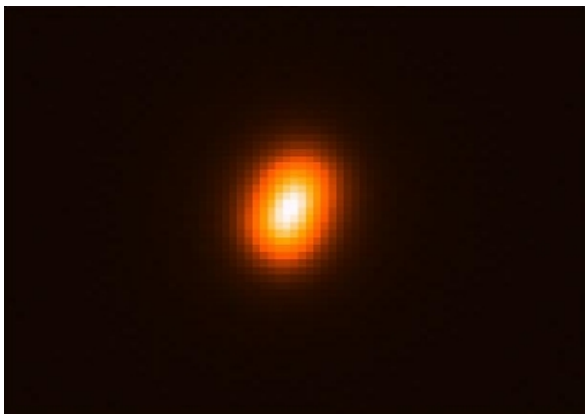
The next most important off-axis aberration is *astigmatism*. Considering rays from just the horizontal and vertical planes, figure 15 shows how astigmatism arises: when the rays in the vertical plane are in focus, the rays in the horizontal plane form a horizontal line. Similarly, at the point where the rays in the horizontal plane are in focus, the rays in the vertical plane form a vertical line. (Considering all other planes would result in an elliptical image, not a line). Elsewhere the image is elliptical, except at the circle of least confusion, where the image is circular and at its smallest extent. The distance between the two line images, and so the size of the circle of least confusion, is proportional to the square of the off-axis angle ϕ.

figure 15: A schematic showing how astigmatism is produced by a parabolic mirror.



Astigmatism can easily be recognised in a telescope if the images of off-axis point sources are elliptical, as shown in figure 16. The presence of astigmatism can be confirmed by adjusting the telescope focus, i.e. moving the focal plane, first one way and then the other. The ellipse should turn into a circle and then into an ellipse again on the other side of the best focus, but with its major axis rotated by 90^{o} (as shown in figure 15). It is important not to confuse this aberration, which affects only off-axis images, with the eye defect of the same name, which affects the entire field of vision and is due to a misshapen eye lens.

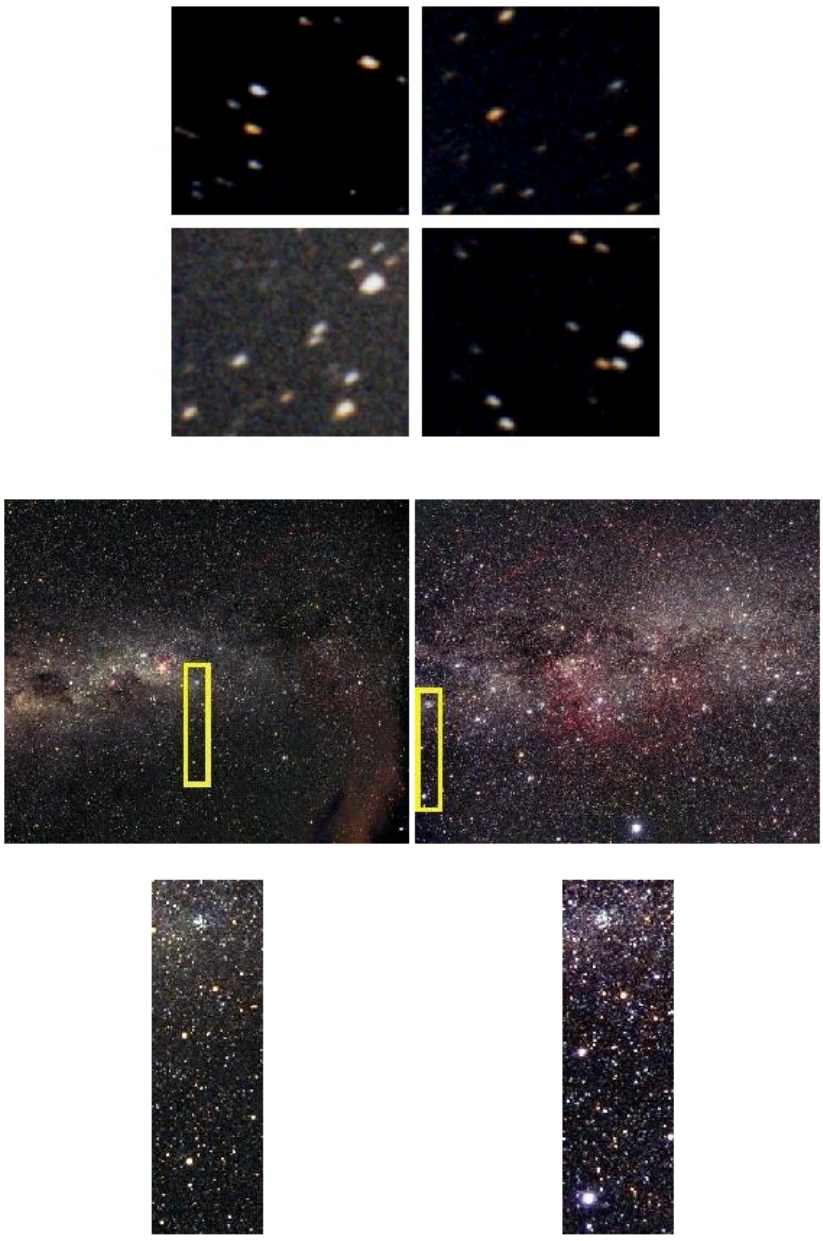
figure 16: An image of a star showing astigmatism.



The two remaining off-axis aberrations are *curvature of field* and *distortion of field* and we shall not deal with

them in detail here. Curvature of field is said to be present when the points of best focus lie on a curved rather than a flat focal plane. An example of the resulting image is shown in the upper panel of figure 17, showing how the blurring occurs in a radial direction with respect to the centre of the field. Curvature of field can be reduced using field flattening lenses.

figure 17: Top: An example of curvature of field. Each image shows a small region at the four corners of a much wider-field photograph. Bottom: An example of distortion of field - two images of the Milky Way taken with the same telescope. The two yellow boxes in the top panels contain the same patch of sky, and their contents are enlarged in the bottom panels.

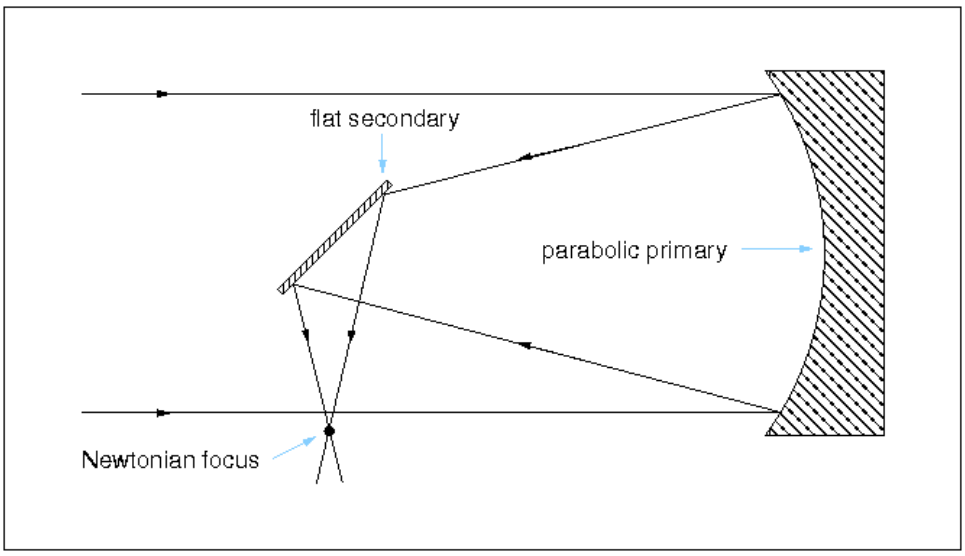


Distortion of field is said to be present when the plate scale varies over the focal plane, as shown in the lower panel of figure 17. The yellow box in the left-hand image is near the field centre. The positions of the two brightest stars in this yellow box should be compared with the positions of the same two stars in the right-hand image, which are now at the field edge. It can be seen that, although the individual stellar images appear aberration-free, their separation is much greater in the right-hand image, i.e. the plate scale is smaller.

Newtonian reflectors

James Gregory was never able to bring his telescope design, known as the *Gregorian*, into practical use, and it is Isaac Newton who is credited with making the first working reflecting telescope in 1668. His design, known as the *Newtonian*, is shown in figure 18. The Newtonian is a two-mirror telescope in which the first mirror in the light path, known as the *primary*, is a concave parabola. The *secondary* mirror has no curvature at all and is hence referred to as a *flat*. It simply folds the light through 90^{o}, placing the focal plane just outside the incoming beam. The focal ratio at the Newtonian focus is typically about 5. The secondary mirror is inclined at an angle of 45^{o} with respect to the primary. The base of the flat is actually elliptical in shape so as to minimise the size of the circular shadow it casts on the primary.

figure 18: A schematic of a Newtonian reflector.



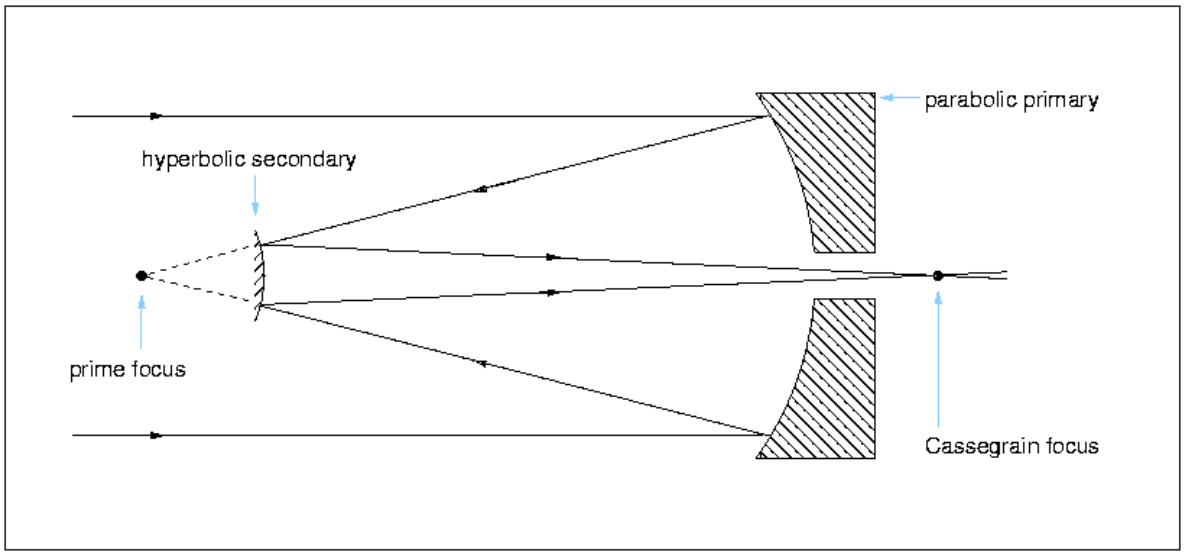
Although small amateur telescopes still adopt a Newtonian configuration, visual access to the focus becomes inconvenient as the telescope becomes larger, and mounting instrumentation there would unbalance the telescope, as demonstrated in the upper panel of figure 19. Hence Newtonians are rarely found in professional observatories.

figure 19: Top: A photograph of a modern Newtonian reflector. Bottom: A replica of Newton's original reflecting telescope.

Cassegrain reflectors

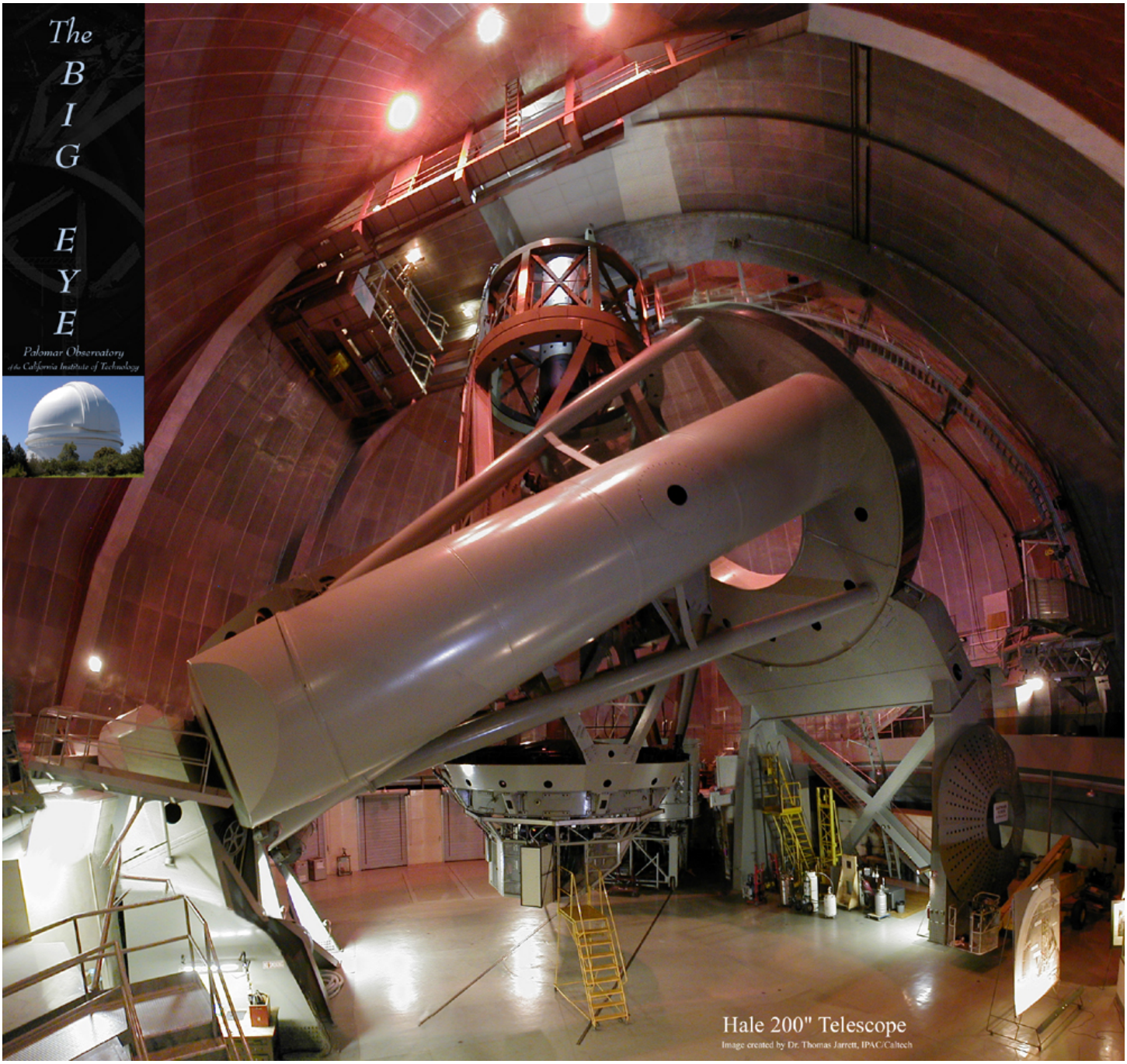
In 1672, the French priest Laurent Cassegrain developed another reflecting telescope design which is now named after him - the *Cassegrain*. This design, shown in figure 20, has been adopted by the majority of the world's largest telescopes due to the convenience of mounting instrumentation at the focus.

figure 20: A schematic of a Cassegrain reflector.



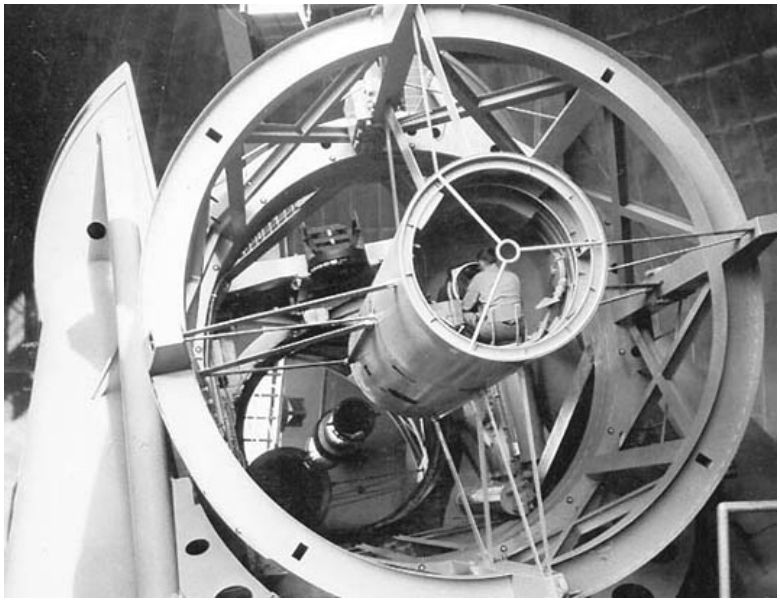
The Cassegrain telescope has a concave parabolic primary mirror like the Newtonian, but it employs a convex hyperboloidal secondary. This increases the focal length of the telescope and reflects the beam back towards the primary, where it passes through a hole bored in the centre of the mirror and comes to a focus just below it. This is a much more easily accessible focus than the Newtonian, and an ideal place to mount large and heavy instrumentation, as shown in figure 21. Compared to a Newtonian, nothing is lost by having a hole in the primary, as this region of the mirror lies under the shadow of the secondary. Moreover, because the beam is folded back on itself, it is possible to have a much longer focal length telescope without a correspondingly long tube: the focal ratio of a typical Cassegrain focus is 15.

figure 21: Photograph of the 5m Hale Telescope on Mount Palomar, California. This telescope is a Cassegrain reflector and was the largest telescope in the world between 1948 and 1993.



Another convenient feature of Cassegrain telescopes is that the focus can be adjusted by moving the secondary mirror towards or away from the primary, allowing the instrumentation to remain fixed in place. In addition, many large research Cassegrains have removable secondary mirrors, giving access to *prime focus* (see figures 20 and 22). This focus is optically equivalent to the Newtonian, and provides a much smaller focal ratio and hence larger field of view than the Cassegrain focus. The wider field makes prime-focus imaging more susceptible to off-axis aberrations than Cassegrain-focus imaging, hence lens-based correctors are usually required at prime focus.

figure 22: Photograph of an observer in the prime focus cage of the Hale telescope. Nowadays, remote operation of prime-focus instrumention means that it is no longer necessary for astronomers to spend the night in the cage!



Ritchey-Chretien reflectors

Both the Newtonian and Cassegrain telescopes suffer from significant off-axis aberrations, primarily coma. To remedy this, the American and French optical designers George Ritchey and Henri Chretien jointly developed the *Ritchey-Chretien* telescope around 1910. The Ritchey-Chretien is a modified form of the Cassegrain design, with a concave hyperbolic primary and a convex hyperbolic secondary. The advantage of this design is that both spherical aberration *and* coma are removed. Astigmatism and field curvature are also reduced, all at the expense of a larger secondary mirror. Hence the Ritchey-Chretien delivers significantly better imaging performance over a wider field of view than a Cassegrain, but with a slightly lower light grasp. It is much more complex to manufacture and test a hyperbolic mirror than a paraboloid, hence the Ritchey-Chretien design tends to be expensive and is found mainly in large research telescopes. The best known example of a Ritchey-Chretien telescope is the 2.4m Hubble Space Telescope, shown in figures 23.

figure 23: Photograph of the HST, a Ritchey-Chretien telescope.

catadioptric telescopes

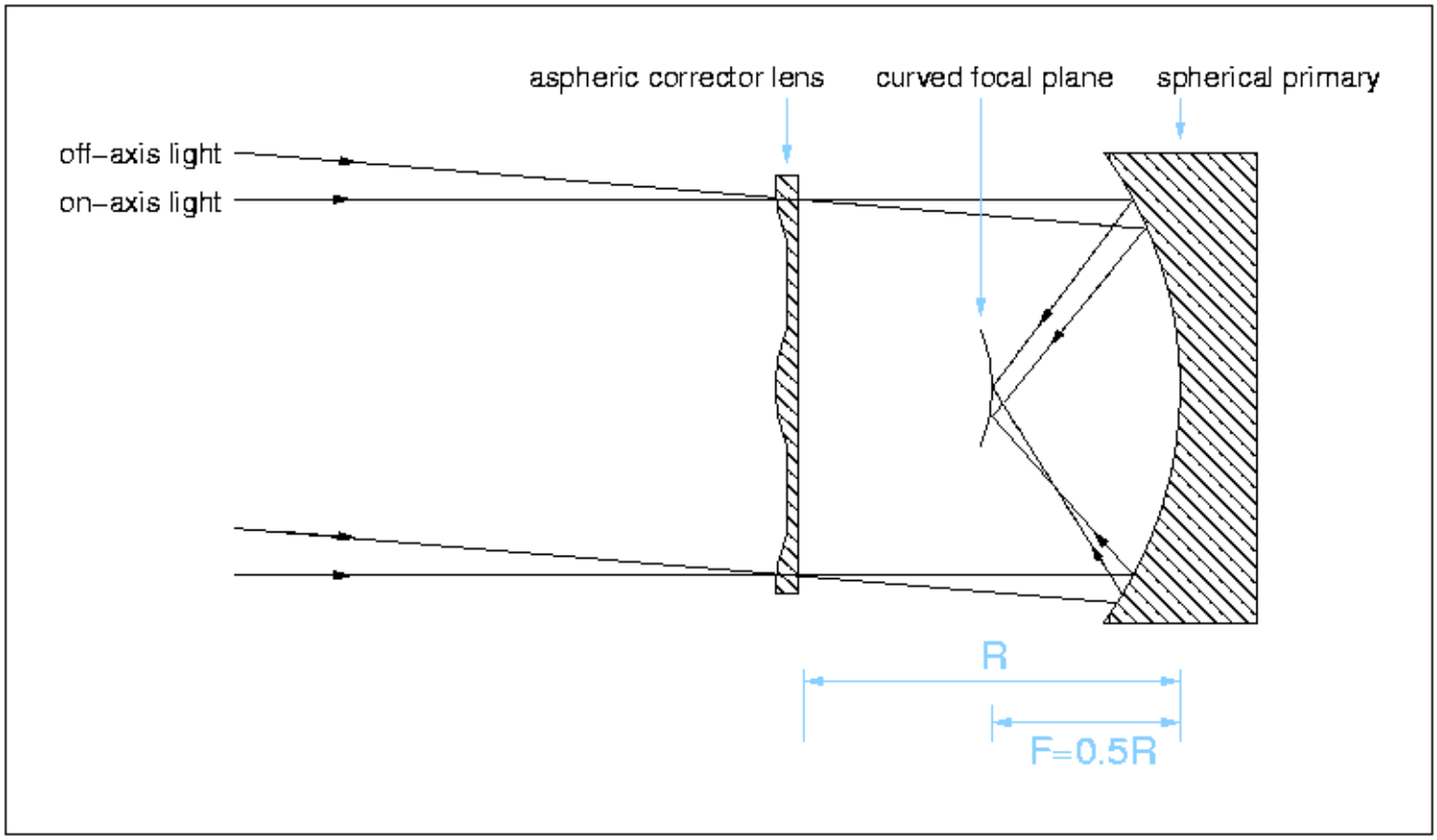
So far, we have been considering telescopes composed of only mirrors, or only lenses. It is possible, of course, to combine mirrors and lenses; such telescope designs are known as *catadioptric* and they offer certain advantages over refractors and reflectors for some applications. We shall look at three different types of catadioptric telescope here: the *Schmidt*, the *Schmidt-Cassegrain* and the *Maksutov*, noting that many other types exist (e.g. the *modified Dall-Kirkham*).

Schmidt telescopes

Due to the coma inherent in Cassegrain telescopes, the field of view tends to be limited to tens of arcminutes. This can be extended to a degree or so using a Ritchey-Chretien design, but fields of view significantly greater than this are almost impossible to achieve with these designs.

In 1930, the Estonian-born optical designer, Bernhard Schmidt, proposed a catadioptric telescope design capable of imaging fields of more than ten degrees across. He decided to use a concave spherical primary, but with an entrance aperture (or *aperture stop*) placed at the *centre of curvature* of the mirror (see figure 23). By symmetry, the spherical mirror then treats every point in the field the same, so all rays are effectively "on-axis" and are imaged onto a curved focal plane. (You can think about this in terms of the shape of the mirror seen by beams coming from different angles; at the centre of curvature, they all see the same spherical shape, whereas closer to the mirror the shape of the mirror depends on the off-axis angle.) Such a design therefore eliminates all off-axis aberrations such as coma and astigmatism. The drawback is that spherical mirrors, as we have already seen, suffer from spherical aberration. Schmidt's design corrects for this using a thin (so it absorbs little light) lens placed at the centre of curvature, thereby allowing wide-field, almost aberration-free views.

figure 23: A schematic of a Schmidt telescope. The corrector lens is placed at the centre of curvature of the primary mirror. The focal length, F, is given by half the radius of curvature, R.



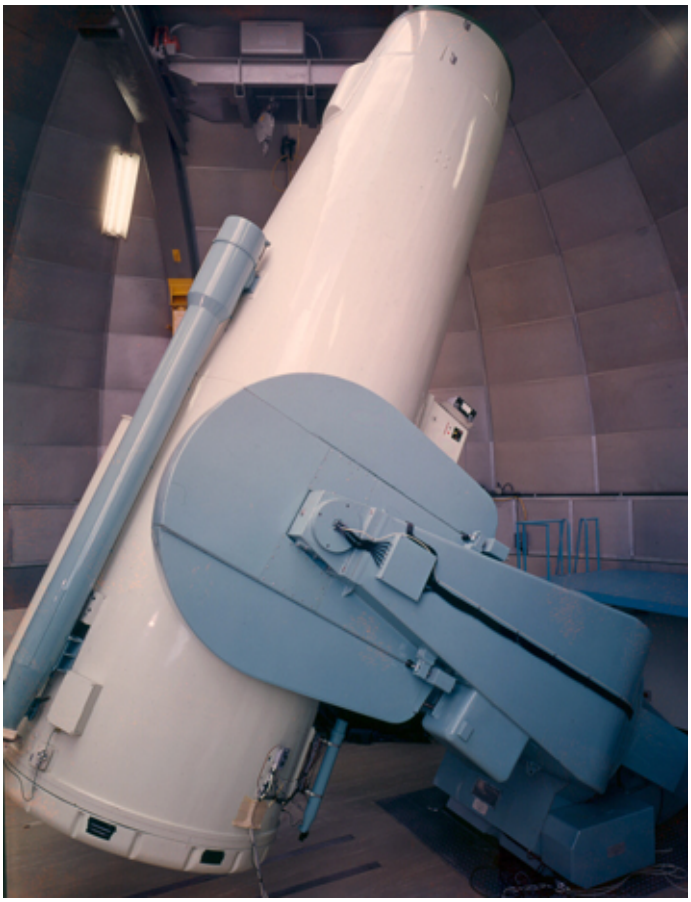
A schematic of a Schmidt telescope is shown in figure 23. In practice, unfortunately, Schmidt telescopes are not completely aberration free. The thin corrector lens, which is usually aspherical, introduces some chromatic aberration and residual spherical aberration. It is also impossible to achieve the idealised scenario of all rays effectively emanating from the centre of curvature by using a finite-sized aperture, hence some off-axis aberrations are still present. The curved focal plane also introduces difficulties, requiring either the use of a field-flattening lens or actually bending the detector (e.g. a photographic plate) to the correct shape. Despite these drawbacks, Schmidt telescopes provide excellent images over a very wide field of view and they were used throughout the 20th-century to perform a number of important sky surveys.

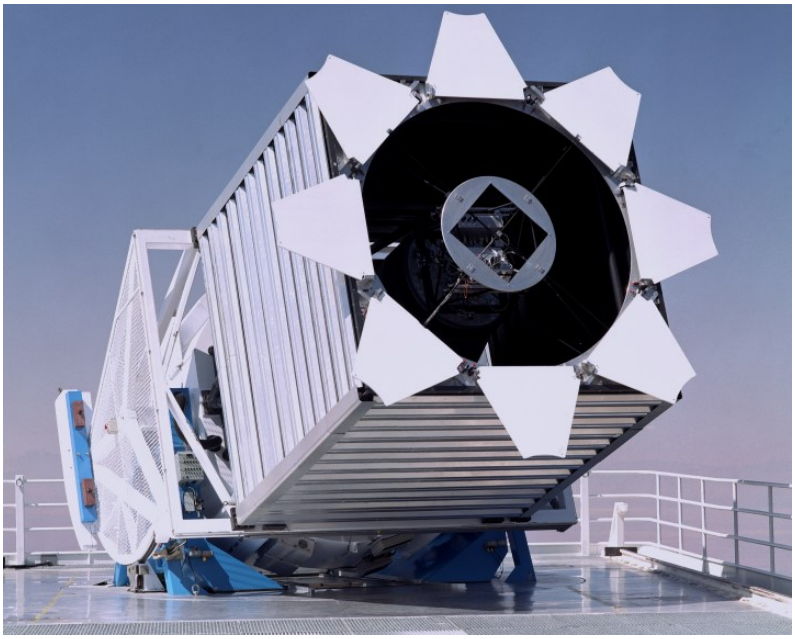
A well known example of a Schmidt design is shown on the left-hand side of figure 24 - the UK Schmidt Telescope (UKST) in New South Wales. This telescope has a corrector lens with a diameter of 1.24m. This defines the aperture of the telescope, but the primary mirror is actually far larger (1.83m) in order to collect light from a wide angle on the sky. The requirement for a wide field of view equates to a large platescale and hence a short focal length; the focal length of the UKST is only 3.07m, i.e. the telescope is a very fast f/2.5. With this design, 6.4^{o} x 6.4^{o} of the sky can be imaged in a single exposure.

The largest Schmidt telescope in the world has an aperture of 1.34m, only slightly larger than the UKST. It is unlikely that a bigger one will ever be built, for the same reason that the diameter of the biggest refractor in the world is only 1.0m: the

corrector is a lens and a much larger size would be difficult and expensive to make, and impossible to hold rigidly by its rim. Another problem is the length of the tube required, which the design dictates must always be approximately twice the focal length of the primary. The inaccessible position and awkward shape of the focal plane also makes the use of modern instrumentation difficult. For this reason, one of the latest sky surveys, the Sloan Digital Sky Survey (SDSS; right-hand side of figure 24), uses a 2.5m Ritchey-Chretien design with a lens-based corrector placed close to the focal plane, providing a field of view of 3^{o}. Although the SDSS telescope takes over 4 times as many pointings to cover the same patch of sky as the UKST, it is the only feasible way of using a larger aperture telescope to survey the sky for the faintest objects.

figure 24: Left: Photograph of the 1.24m UK Schmidt Telescope in New South Wales. Right: Photograph of the 2.5m Sloan Digital Sky Survey telescope in New Mexico.

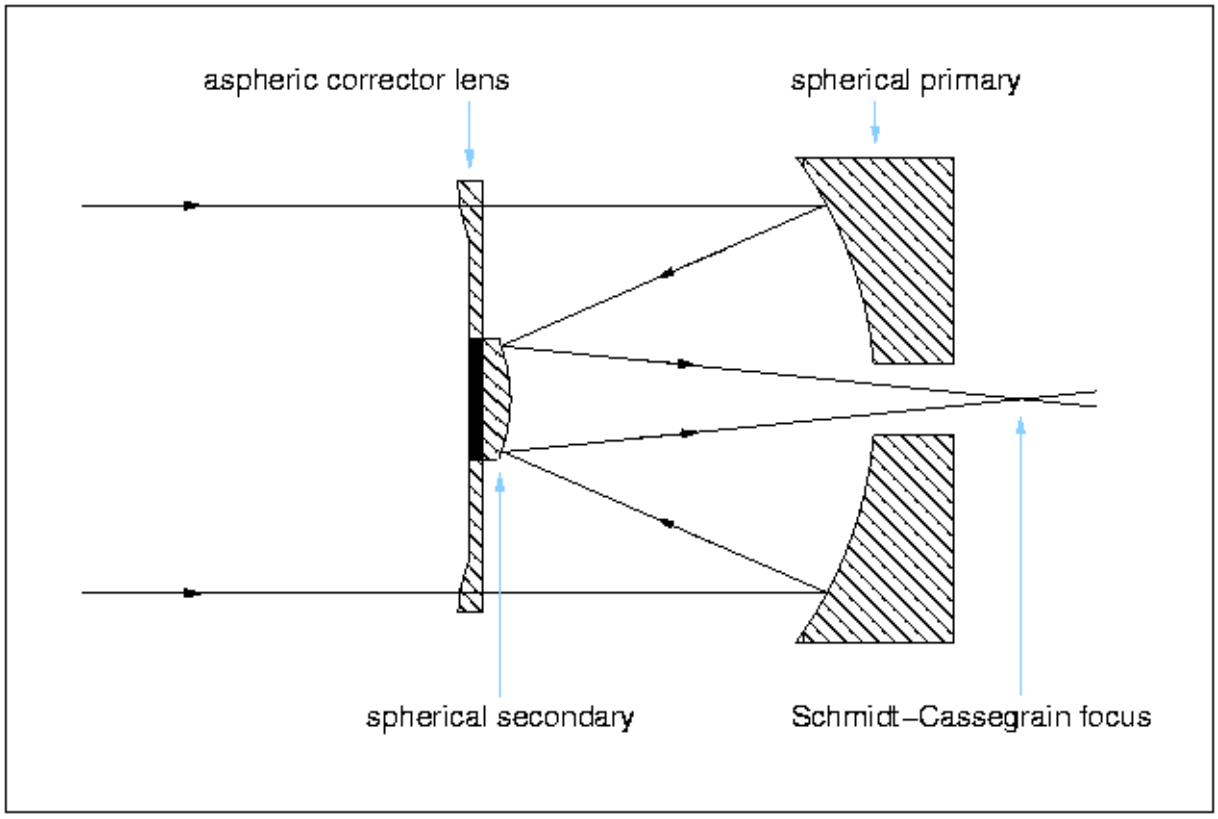




Schmidt-Cassegrain telescopes

The *Schmidt-Cassegrain* is not a telescope you will find in major research observatories, but it is arguably the most widespread design used in the amateur telescope market. As its name implies, it is a hybrid of the Schmidt and Cassegrain designs. The light passes through a corrector lens and reflects off a concave spherical primary, just like in a Schmidt telescope. The focal ratio of the primary is higher, however, and the corrector is not placed at the centre of curvature, so the reflected light hits the underside of the corrector, where a convex spherical secondary mirror reflects it back down through a hole in the primary and focuses it, as shown on the left-hand side of figure 25.

figure 25: Left: A schematic of a Schmidt-Cassegrain telescope. Right: A photograph of one of the market-leading 8-inch Schmidt-Cassegrain telescopes, which currently costs about £2500.



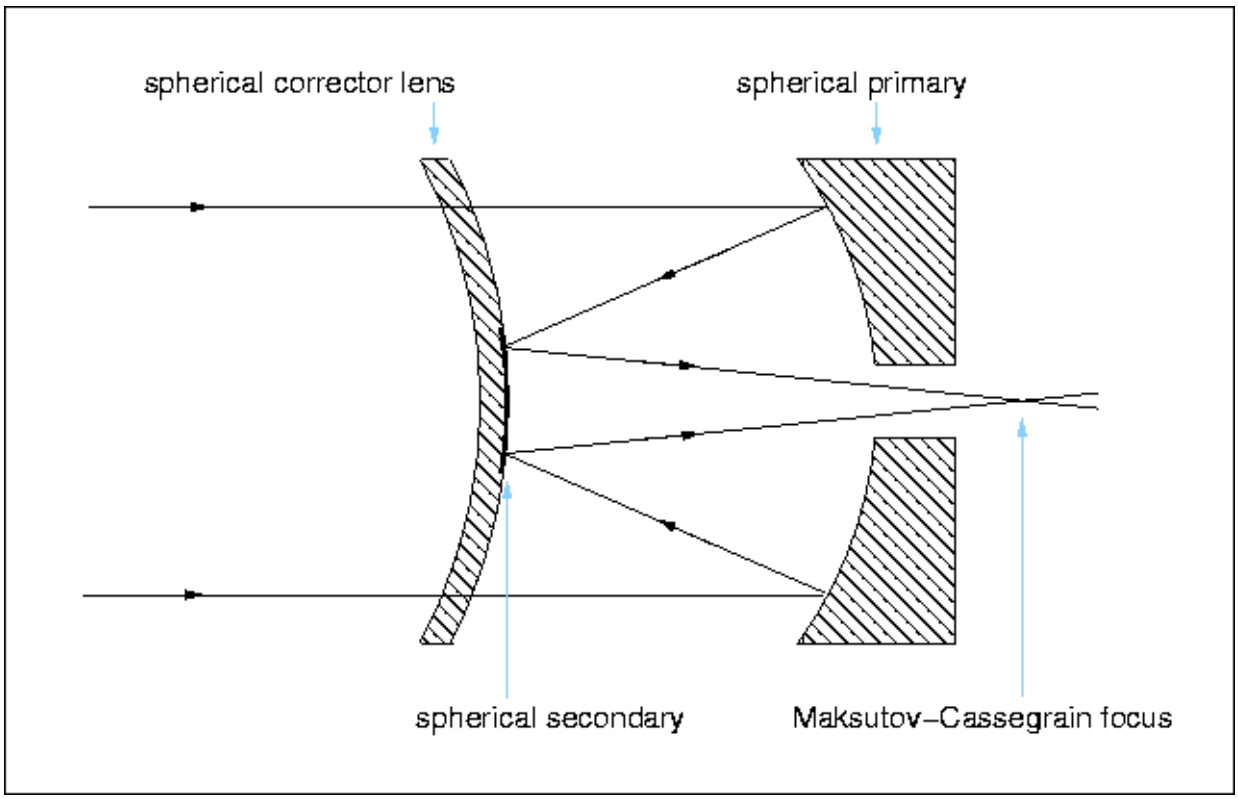
The great advantage of the Schmidt-Cassegrain design is that it is cheap to mass produce, thanks to the use of spherical mirrors. The resulting spherical aberration is dealt with by the corrector lens, a feature it borrows from the Schmidt telescope. Unlike the Schmidt telescope, however, the Schmidt-Cassegrain suffers from off-axis aberrations like coma and astigmatism because the corrector lens is not placed at the centre of curvature. The Schmidt-Cassegrain retains one of the main advantages of the Cassegrain - a long focal length in a short tube, giving a telescope which is compact and portable yet able to provide small enough plate scales for detailed planetary viewing. This versatility makes them ideal for the

amateur astronomy market. The typical focal ratio of a Schmidt-Cassegrain telescope is f/10 and they are available in apertures ranging from approximately 4-16 inches. Larger apertures become impractical due to the cost of manufacturing the corrector and mounting it without flexure (especially if the secondary mirror is attached to it). An example of a Schmidt-Cassegrain telescope is shown on the right-hand side of figure 25.

Maksutov telescopes

The corrector plate in a Schmidt-Cassegrain telescope is aspherical and hence difficult to shape and polish accurately. In 1941, the Russian optician Dmitri Maksutov announced a new catadioptric design, the *Maksutov telescope*, which does away with the aspheric corrector and replaces it with a meniscus lens with spherical surfaces. The spherical inner face of the corrector lens has a small spot aluminized on it which acts as the secondary mirror, redirecting the light through a hole in the spherical primary, as shown on the left-hand side of figure 26. Hence such telescopes are also known as *Maksutov-Cassegrains*.

figure 26: Left: A schematic of a Maksutov telescope. Right: A photograph of one of the world's best-selling telescopes - a 90mm Maksutov-Cassegrain, which currently costs about £500.



It is often said that Maksutov telescopes offer the ultimate in imaging performance. This is true to a point, as it is easier to shape and polish spherical surfaces to a high accuracy than the aspheric surfaces found in Schmidt-Cassegrain correctors. In order to minimise off-axis aberrations like coma and astigmatism, the focal lengths of Maksutovs tend to be higher than Schmidt-Cassegrain telescopes, with typical focal ratios of f/15. This results in a smaller secondary mirror and hence a reduced central obstruction, which improves image contrast. On the down-side, the slow focal ratio reduces the field of view of a Maksutov, and the deep curve and thickness of the meniscus lens becomes prohibitively expensive to manufacture and mount in large apertures. For this reason, the largest Maksutovs that can readily be purchased have only 8-inch primary mirrors.

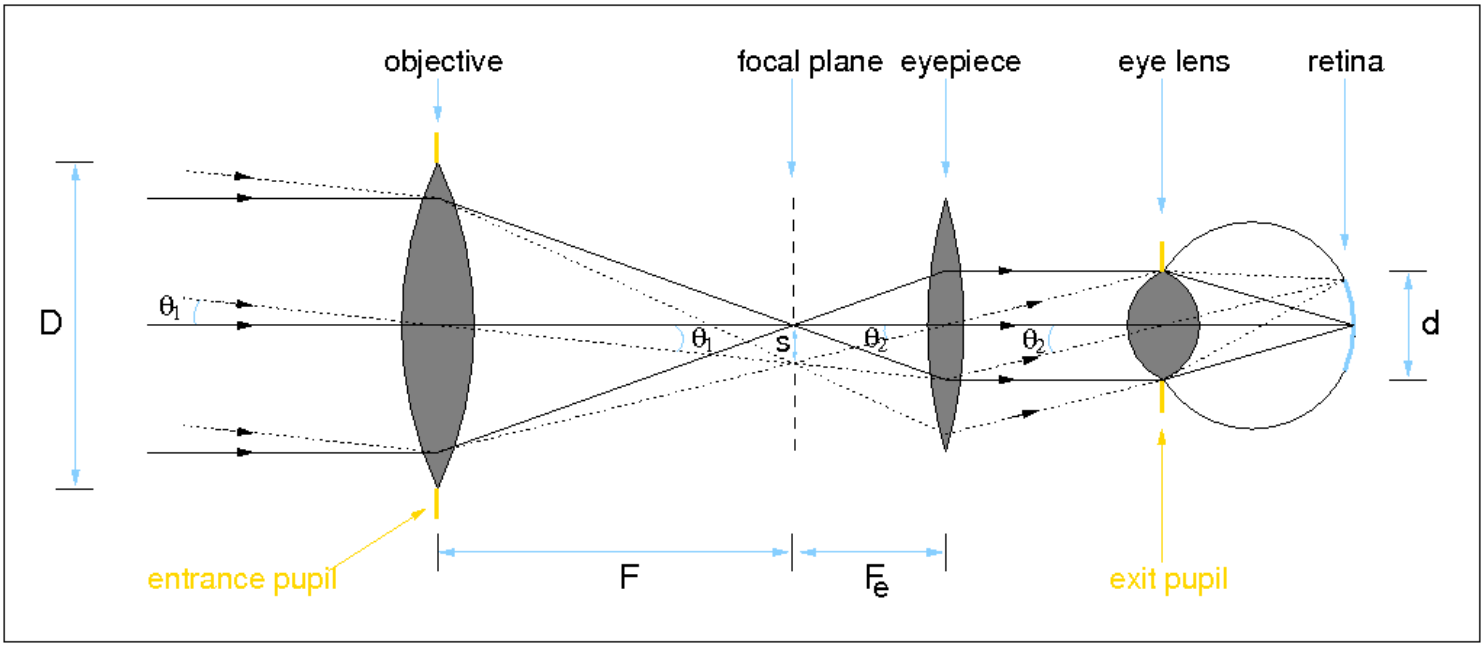
---

visual use of telescopes

The role of the eyepiece and the eye

It is rarely possible to look through modern research telescopes with the eye. This is because observing time on them is so precious and because the eye is such a poor detector compared to the electronic equivalents. On the other hand, amateur telescopes are often used visually, in which case an eyepiece is required, as shown in figure 27.

figure 27: Schematic of a refracting telescope, illustrating the role of the eyepiece and the eye. The same concepts would apply if considering a reflecting telescope.



The role of the eyepiece can be thought of as a magnifying glass which takes the diverging rays of light from each point in the focal plane and makes them parallel. The lens in the eye then focuses this parallel light, creating an image of the object on the retina which is the correct way up. However, since the brain inverts all images falling on the retina (if it didn't, the everyday world around us would look upside down!), the view in a telescope appears upside down. As can be seen from figure 27, the image recorded by an electronic detector placed at the focal plane of the objective would also appear upside down.

Looking in more detail at figure 27, it is useful to define the following terms: the *entrance pupil* is the area at the entrance of the telescope that can accept light. Clearly, in figure 27, this is defined by the objective lens. The *exit pupil* is the image of the entrance pupil formed by the eyepiece. Only rays which pass through the exit pupil can exit the system, and it therefore represents the optimum position for the eye when observing with a telescope. If the eye is placed closer or further away from the eyepiece than the exit pupil, not all the light gathered by the telescope will be collected by the eye. Light is also lost if the diameter of the exit pupil is greater than the size of the dark-adapted pupil of the eye, i.e. about 8mm. The distance of the last lens of the eyepiece to the exit pupil is known as the *eye relief*, and for comfortable viewing is usually designed to be around 10mm in an amateur telescope.

For simplicity, figure 27 assumes that the eyepiece (or *ocular*) is a single positive lens. We have already seen that such a lens would introduce chromatic aberration and hence combinations of lenses are often used. The simplest combinations, composed of two lenses, are the *Huygens* and *Ramsden* eyepieces. More complex, and expensive, eyepieces use three or more lenses and are able to provide better correction for chromatic

aberration, wider aberration-free fields-of-view, and comfortable eye relief. Common designs are the *Kellner*, *Orthoscopic*, *Plössl*, *Erfle* and *Nagler* eyepieces, some of which are shown in figure 28.

figure 28: Photograph of some typical telescope eyepieces. There is a huge variety of designs, with very different sizes and weights, and prices ranging from tens of pounds to many hundreds.

Magnifying power

From figure 27, it can be seen that the objective, of focal length F, focuses parallel light to a point image in its focal plane; the eyepiece, of focal length F_e, takes a point source in *its* focal plane and converts it back to parallel light. Two distant point sources separated by an angle θ_1 will produce two point images separated by a distance $s = F \theta_1$ in the focal plane of the objective and, since this is also the focal plane of the eyepiece, the resulting two beams of parallel light will diverge at an angle $\theta_2 = s / F_e$. Hence, an observer looking through the eyepiece will see parallel light from two sources separated by an angle θ_2 on the sky. Without the help of the telescope, the eye would receive parallel light from the two sources separated by an angle θ_1. Therefore, since $\theta_2 > \theta_1$, the image has been magnified. Given that $s = F \theta_1 = F_e \theta_2$, the *magnification* (or *magnifying power*) of the telescope/eyepiece combination is simply defined as:

$M = \theta_2 / \theta_1 = F / F_e$.

When the magnification of a telescope is defined in this manner, it clearly only has any meaning when the telescope is being used with an eyepiece. Hence magnification is not a term used in the research vocabulary of professional astronomers, whose telescopes are never used visually. The term is relevant to the amateur astronomy market, of course, but it is often misused as the primary indicator of telescope performance. As we have seen, it is the diameter, focal length and telescope design which affect the performance of a telescope; the magnification is something which can be altered simply by using different eyepieces and is not an inherent property of the telescope.

Magnification limits

A further inspection of the on-axis rays in figure 27 shows that the triangle formed by the diameter of the objective, D, and the image in the focal plane at a distance F from the objective is similar to the triangle formed by the diameter of the exit pupil at the eyepiece, d, and the image in the focal plane at a distance F_e from the eyepiece, i.e.

$$D / d = F / F_e = M.$$

Hence an alternative way to express magnification is as the ratio of the diameters of the objective and the exit pupil. This naturally leads to a lower limit to the useful magnification of a telescope; the magnification must be sufficiently high to make the exit pupil equal to or smaller than the diameter of the dark-adapted pupil of the eye, otherwise not all of the light collected by the telescope will be gathered by the eye. This condition can be written as

$$M \geq D / d,$$

where d is now the diameter of the eye's pupil.

Another lower limit to the useful magnification of a telescope can be obtained by comparing the resolving power of the telescope to that of the eye. It is impossible to separate two point sources by increasing the magnification if they are not resolved in the focal plane. However, if the two point sources *are* resolved in the focal plane, it is possible to use magnification to bring them above the resolution limit of the human eye. Hence a lower limit to the useful magnification of a telescope is the factor by which its resolution, α, needs to be multiplied by to make it equal to the resolution of the eye, α_e; any lower, and the full resolving power of the telescope is not being properly exploited. This limit can be expressed mathematically as

$$M \geq \alpha_e / \alpha.$$

Assuming that the resolution of the telescope is limited by diffraction, and that the resolution of the eye is approximately four times its diffraction limit (due to the relatively poor quality of the eye lens), this equation can be rewritten as

$$M \geq 4 \times (1.22\lambda / d) / (1.22\lambda / D) = 4 D / d.$$

Assuming again that the diameter of the dark-adapted eye pupil, d, is 8mm, we finally obtain

$$M \geq D / 2,$$

where D is expressed in mm. Hence, matching the resolution of the telescope to the resolution of the eye imposes a lower magnification limit which is four times higher than that imposed by the exit pupil diameter.

It is not possible to increase the magnification of a telescope indefinitely by changing the eyepiece. Ultimately, magnification is limited by the difficulty in making usable eyepieces of very short focal length and by the progressive impairment of vision as the beam that the eye accepts becomes smaller than about 1mm. Hence, using the expression derived above for the magnification of a telescope in terms of the diameters of the objective and exit pupil, we get

$$M \leq D,$$

where D is again expressed in mm. Hence an 8-inch (200mm) reflector will have a useful upper magnification limit of approximately 200.

An example problem shows how these magnification limits can be used in practice. It should be emphasized, however, that the magnification limits described here are only guides, and the actual limits of magnification depend on numerous factors, such as the characteristics of the astronomer's eye and telescope, the atmospheric conditions and the objects being observed.

Visual resolving power

The theoretical resolving power, α, of a telescope based on Rayleigh's criterion can be written as

$\alpha \approx 140 / D$,

where α is in arcseconds, D is in mm, and we have assumed an observing wavelength of 550nm.

A skilled observer, however, can actually resolve two stars which are separated by an even smaller angle than given by this expression. This is possible because an acute eye in good observing conditions can actually detect a drop in the intensity between two close point sources of only 5%, whereas the drop in intensity implied by Rayleigh's criterion is 20%. A number of alternatives to Rayleigh's theoretical criterion have therefore been proposed, including *Dawes empirical criterion*, which can be approximated by:

$\alpha \approx 115 / D$,

where α is in arcseconds, D is in mm and we have again assumed an observing wavelength of 550nm. Hence Dawes' limit gives resolving powers which are about 20% better than Rayleigh's limit. *Sparrow's limit* is even smaller, stating that it is possible to resolve objects even when they are separated by half of the angle given by Rayleigh's criterion, although to do this requires the use of electronic detectors and data analysis techniques to search for the flat-topped intensity profile of the two closely-separated point sources.

---

example problems

1. **Show that the plate scale, *p*, of an astronomical telescope with focal length, *F*, is given by the relation**

   ***p* = 206265 / *F*,**

   where *p* is measured in arcseconds per mm.

   We have seen that the relationship between the size of an image in the focal plane, *s*, and its angular size on the sky, θ, is given by

   $s = F\theta$.

   If *F* is measured in mm, the platescale in radians per mm is given by

   $p = \theta / s = 1 / F$.

   There are 180^{o} in π radians and hence (180x3600/π) arcseconds in 1 radian. Therefore, the platescale in arcseconds per mm is given by

   $p = 206265 / F$.

2. **An amateur astronomer has a 20 cm telescope with a focal ratio of *f*/8. What is the diffraction-limited resolution of this telescope when observing light of wavelength 550nm? Quote your answer both in arcseconds on the sky and mm in the focal plane.**

   Rayleigh's criterion tells us that the diffraction-limited resolution of a telescope, α, is given by

   $\alpha = 1.22\lambda / D$.

   In the question, the diameter of the telescope aperture, *D*, is 0.2m.

Remembering to use the same units for the wavelength, λ, this gives α = 3.355x10^{-6} radians. There are (180x3600/π) arcseconds in 1 radian, so the diffraction-limited resolution of the amateur astronomer's telescope is 0.69 arcseconds on the sky.

To convert resolution from angular units to a linear distance in the focal plane, it is necessary to know the plate scale, *p*, which in turn requires a knowledge of the telescope focal length, *F*. The focal length can be determined from the focal ratio, *f*, as follows: *F* = *Df* = 1600mm. The platescale can then be calculated: *p* = 206265 / *F* = 129 arcseconds per mm (or "/mm). The diffraction-limited resolution expressed as mm in the focal plane is then simply 0.69/129=0.005mm.

3. An amateur astronomer has a dark-adapted eye pupil of diameter 8mm and uses a 25 cm telescope with a focal ratio, *f*/10. What focal length eyepieces are required to operate the telescope at the minimum and maximum useful magnifications?

In the question, the diameter of the telescope aperture, *D*, is 250mm and its focal length, *F* = *Df* = 2500mm. A lower limit to the useful magnification of a telescope is given by the expression

$M \geq D / d,$

where *d* is the diameter of the eye's pupil. Hence *M* ≥ 31. The magnification is related to the focal length of the eyepiece, by the relation

$M = F / F_e.$

Hence an eyepiece of focal length F_e = 80mm would be required.

An upper limit to the useful magnification of a telescope is given by the expression

$M \leq D,$

where *D* is expressed in mm. Hence *M* ≤ 250, and an eyepiece of focal length $F_e = F / M$ = 10mm would be required.

telescopes

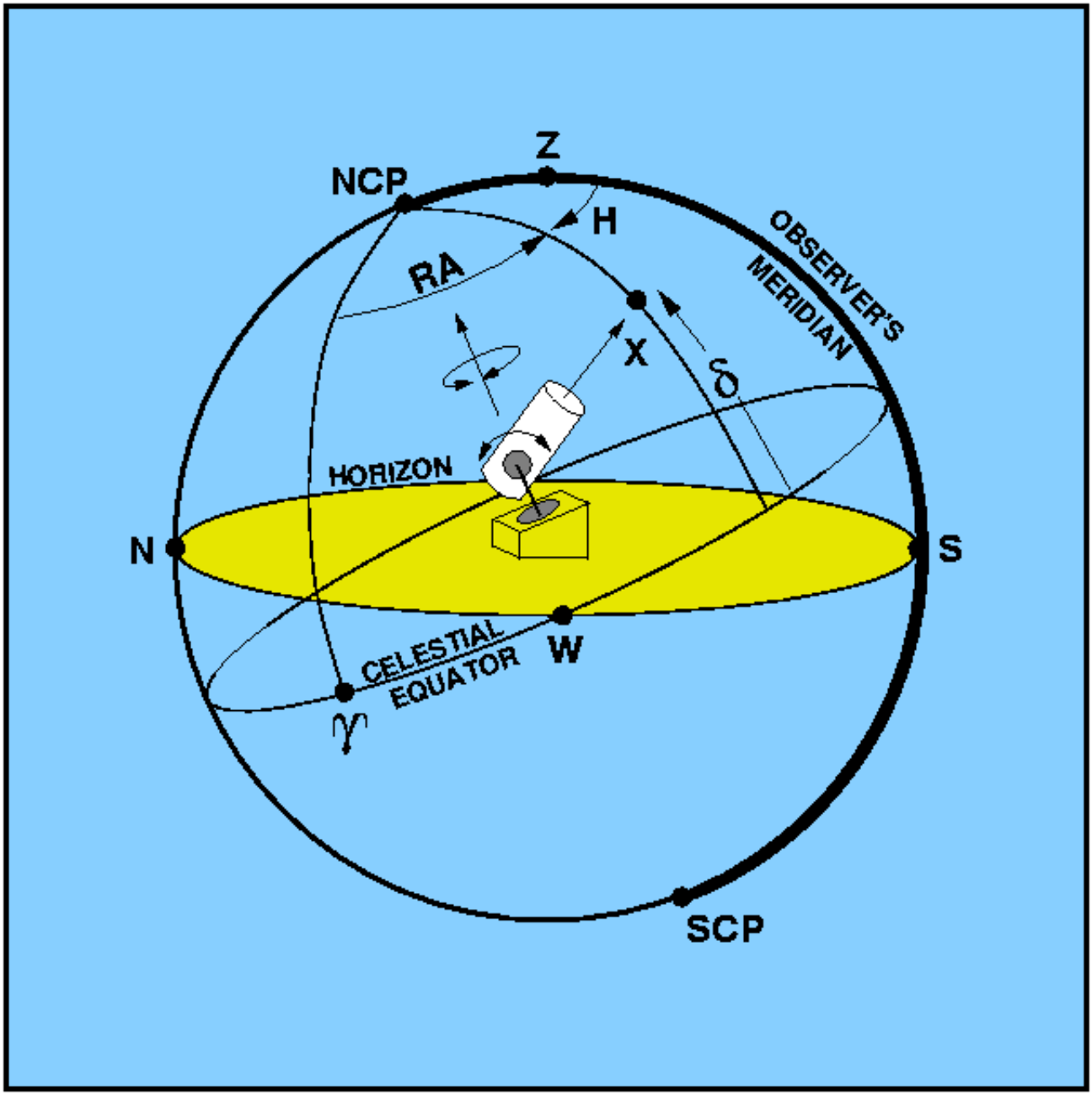
III. telescope mountings

equatorial mountings

A telescope mounting must allow the telescope to be pointed at any point in the sky and track the object as it moves due to the rotation of the Earth, maintaining the relative alignment of the optics as it does so.

Before the 1980's, all large telescopes were mounted on *equatorial mountings*. This system is illustrated in figure 29. One axis is parallel to the Earth's rotation axis and is known as the *polar axis*. The second axis is at right angles to this and is known as the *declination axis*. The beauty of this design is that there is a direct relation between the equatorial coordinates of an object and the axes of the telescope: rotation about the polar axis scans out a circle of constant declination and varying right ascension; rotation about the declination axis scans out a circle of constant right ascension and varying declination. As well as making it easy to point to an object, this also makes it simple to track the motion of an object due to the Earth's rotation by driving just the polar axis at an equal angular velocity, but in the opposite direction, to the angular velocity of the Earth on its rotation axis.

figure 29: Schematic of an equatorial mounting.



In practice, unfortunately, it isn't possible to keep an object perfectly in the centre of the field of view merely by rotating the polar axis at the sidereal rate. This is due to a number of factors, including inaccuracies in the telescope drive systems, flexure, and misalignment of the mount axes with respect to the celestial pole. To combat this image motion, which would cause smearing in long exposures, it is necessary to *guide*, where small adjustments to the position of the telescope are made to keep the object at the desired position in the focal plane. On all professional telescopes, this is done automatically using an *autoguider*, which continuously measures the position of a guide star somewhere in the field of view and nudges the telescope to keep the guide star locked onto a particular pixel on the autoguider's detector.

There are several different types of equatorial mounting; which one is adopted for a particular application depends on the size and type of the telescope being mounted. We shall look at the three main equatorial mount designs here: the *German mounting*, the *fork mounting* and the *English mounting*.

The German mounting

An example of a German equatorial mounting is shown in figure 30. In this design, the declination axis is in the form of a beam across the top of the polar axis. The telescope is attached to one end of the beam and a counterweight to the other. In large German mounts, the polar axis is often supported on a pier. The advantage of the German design is that it can handle both refractors and reflectors of all focal lengths, and it can access all areas of the sky. The disadvantage is that it suffers from the need for meridian reversal: to prevent the telescope from crashing into the pier whilst tracking objects crossing the observer's meridian, it is usually necessary to stop the observations, rotate the polar and declination axes through 180^{o} (so placing the telescope on the opposite side of the pier), and then restart the observations. This, and the large footprint of the German mount due to its long declination axis, can make it rather cumbersome to use, a problem exacerbated by the sometimes awkward positioning of the focal plane of the telescope.

figure 30: Examples of German equatorial mountings. Top left: A photograph of a 120mm refracting telescope. Top right: A photograph of the 50 cm Mons reflector on Tenerife. Bottom: pt5m, the Durham-Sheffield 0.5m robotic telescope on La Palma.

The German mount is a popular choice in the amateur astronomy market, but is rarely used in larger research telescopes. The reason for this is that the offset mounting of the telescope induces large torques which cause bending of the telescope and axes, and stresses on the bearings.

The fork mounting

The equatorial fork is probably the most popular mount found in the amateur astronomy market today, and is also the mount of choice for many large equatorially-mounted research telescopes. The design of the polar axis in a fork mount is very similar to that of the German mount. However, rather than supporting just one side of the declination axis passing through the telescope, as in the German mount, the fork mount supports the declination axis on both sides of the telescope, as shown in figure 31.

figure 31: Examples of equatorial fork mountings. Left: A photograph of an 8-inch Schmidt-Cassegrain telescope. Right: A photograph of the reflecting 2.5m Isaac Newton Telescope on La Palma.

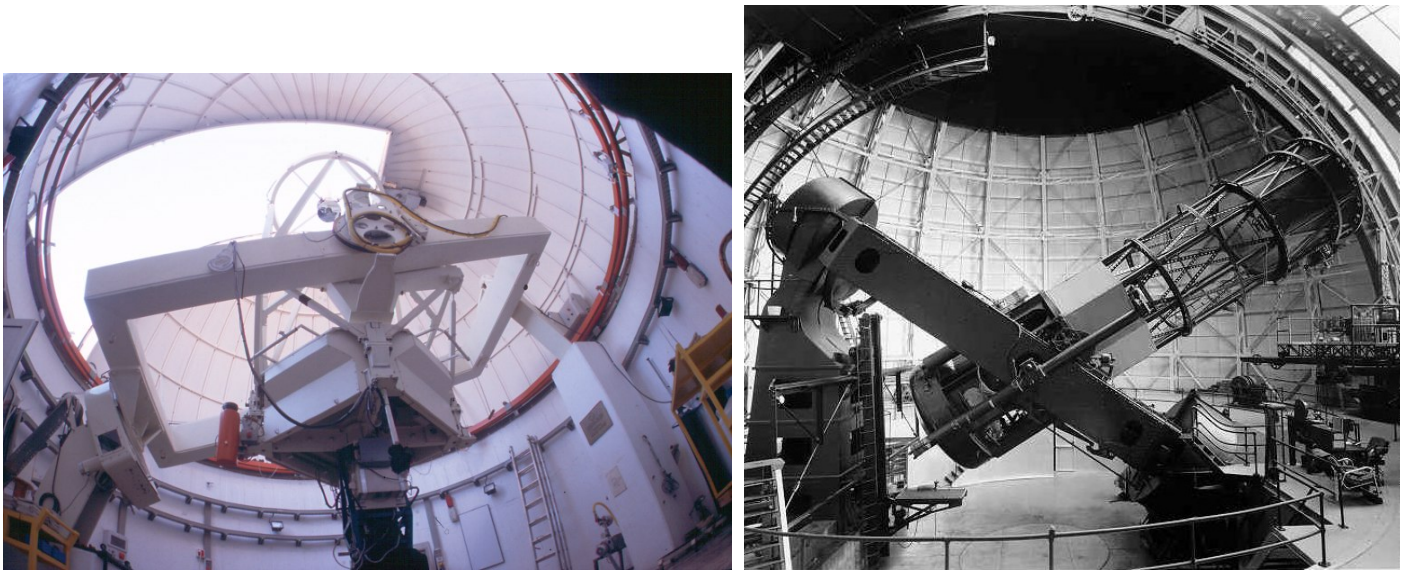
The fork mount has a number of advantages over the German mount. First, it does not require meridian reversals. Second, it is generally easier to access the focal plane. Third, no large counterweights are required and hence fork mounts are usually more compact than German mounts. There are disadvantages to the fork mount, however. The main one is that the fork arms (or *tines*) tend to be short in length to retain stiffness. This then makes it difficult to mount refractors, which generally have longer tubes than reflectors, without losing access to the sky around the celestial pole. Even if the telescope tube is short enough to rotate freely between the tines, the design still restricts the space for instrumentation mounted beneath the primary mirror.

For large telescopes, the fork mount tends to be stiffer than the German mount, thanks to the way in which both sides of the declination axis passing through the telescope are held. In both the German and fork designs, however, the telescope is mounted at the end of the polar axis, which stresses the bearings and induces flexure.

The English mounting

The English mounting, or *yoke mounting*, aims to reduce the flexure in the polar axis inherent in the German and fork mounts by supporting both sides of the polar axis of the telescope on two piers, as shown in figure 32. A cradle, or *yoke*, is used as the polar axis, and both sides of the declination axis of the telescope are supported by it.

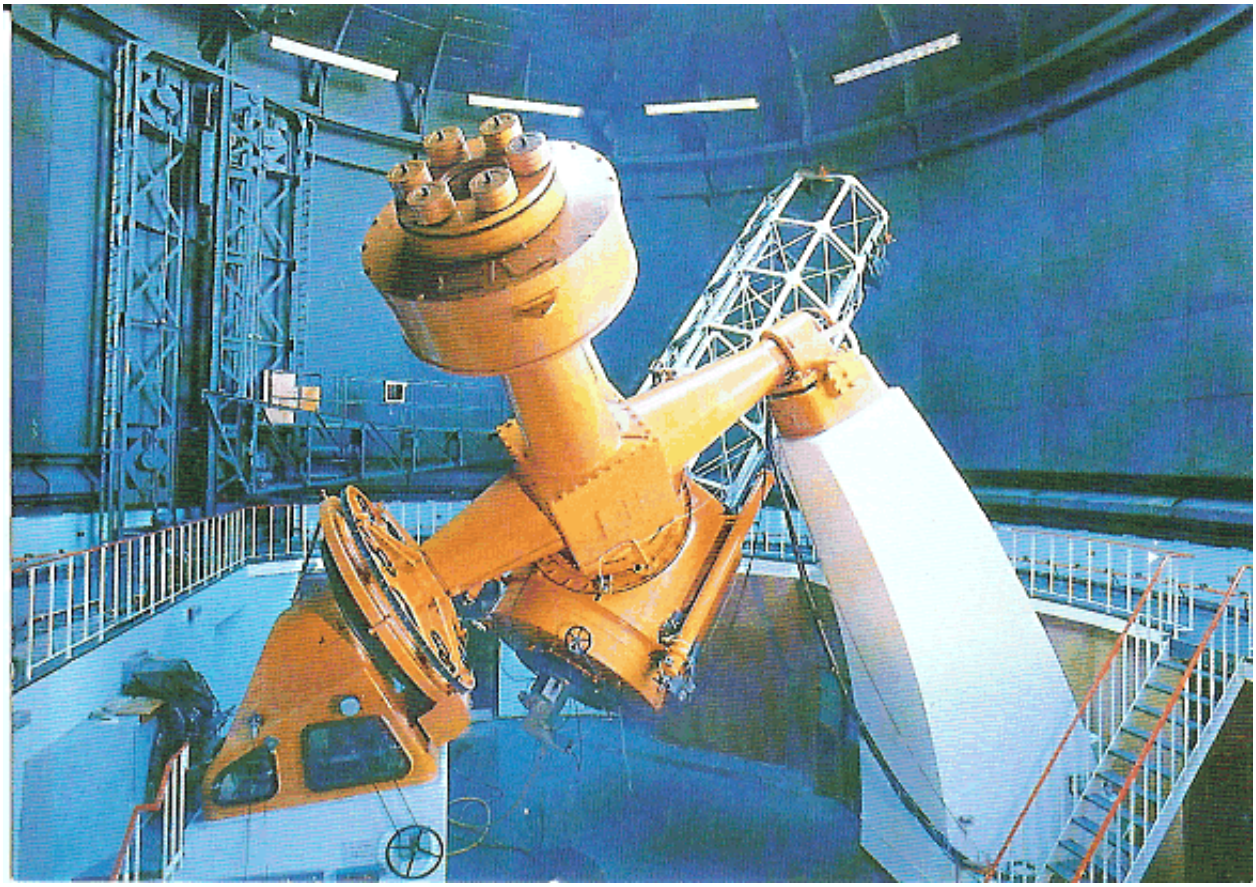
figure 32: Two examples of English mountings. Left: A photograph of the 1.5m Carlos Sanchez Telescope on Tenerife. Right: A photograph of the 100-inch Hooker Telescope in California, with which Edwin Hubble discovered the expansion of the Universe.



Although providing stable polar and declination axes, the disadvantage of the English mounting is that the sky around the celestial pole is unobservable, irrespective of the length of the telescope. One way round this is to use a *modified English mounting*, shown in the left-hand panel of figure 33, where the yoke is replaced by a single beam. The declination axis is mounted in a similar way to the German mount, with the telescope at one end and a counterweight at the other. Hence, although the modified English design provides access to the sky around the celestial pole for telescopes of all lengths, it does so at the expense of reintroducing the flexure in the declination axis inherent to the German design.

figure 33: Left: A photograph of the 1.9m Radcliffe Telescope in South Africa, which has a modified English mounting. Right: A

photograph of the 3.9m Anglo-Australian Telescope in New South Wales, which has a horseshoe mounting.



Another variant of the English mounting is shown in the right-hand panel of figure 33. This is known as the *horseshoe mounting*, or the *open yoke mounting*. It aims to combine the advantages of the German, English and fork mounts, whilst eliminating their disadvantages. It does this by adopting an English mount design with the end of the yoke open, thereby adopting a horseshoe shape, so that the telescope can still point towards

the celestial pole. The mount is relatively compact, no meridian reversals are required, the polar regions are accessible, and all masses lie between their points of support. It is no suprise, therefore, that the world's largest telescopes built before the 1980's were all horseshoe mounted, including the Mayall 3.8m Telescope on Kitt Peak, the 5m Hale Telescope, and the 3.9m Anglo-Australian Telescope. Note the subtle difference in the location of the declination axis in the latter two telescopes; in the Anglo-Australian Telescope it is located within the horseshoe itself, whereas in the Hale Telescope it is within a yoke.

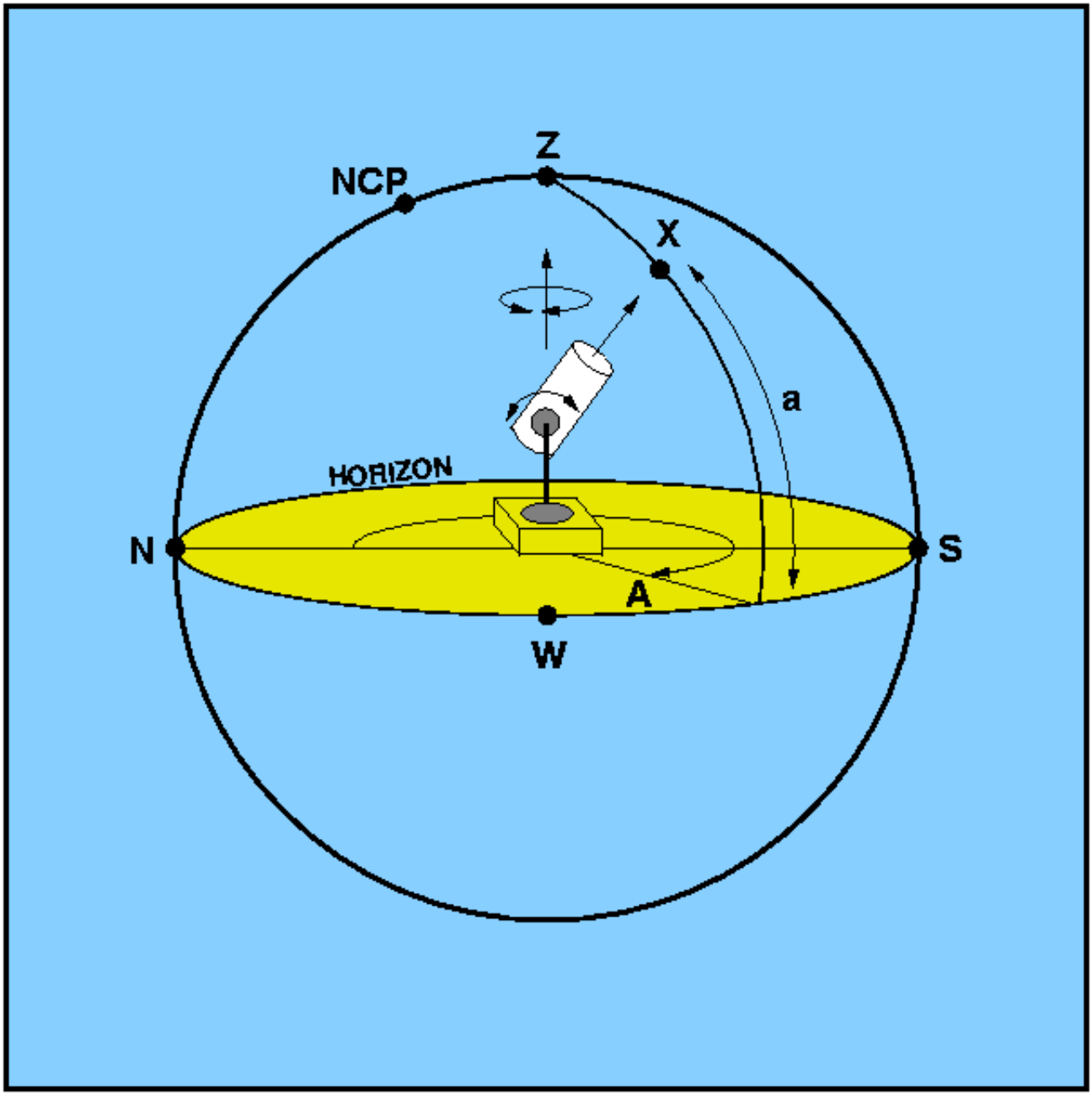
---

alt-azimuth mountings

All equatorial mountings have a fundamental flaw - the polar axis must be aligned with the Earth's rotation axis. This means that unless the telescope is situated at the Earth's poles, the polar axis will be inclined relative to the ground (or, strictly speaking, the local gravity vector), thereby creating stresses on the axes which vary as the telescope moves. The sheer size of the structure required to support the 3.9m Anglo-Australian Telescope, shown in figure 33, demonstrates the impractability and expense of mounting telescopes significantly larger than this on an equatorial mount; the largest equatorially-mounted telescope ever built - the 5m Hale Telescope - is only slightly larger than this.

One solution would be to reduce the length of the horseshoe so that it is stiffer, but this then reduces the space available beneath the primary mirror to mount instrumentation. A much better solution is to use a fork mount, but incline the tines so that they are parallel with the local gravity vector. This places the centre of mass of the telescope directly above the point where the telescope is mounted on the ground, which, for a given mass of mount, gives a much stiffer structure that is largely free from variable stresses. In this design, the length of the tines can also be increased to provide plenty of room for instrumentation without significantly degrading the structural stability.

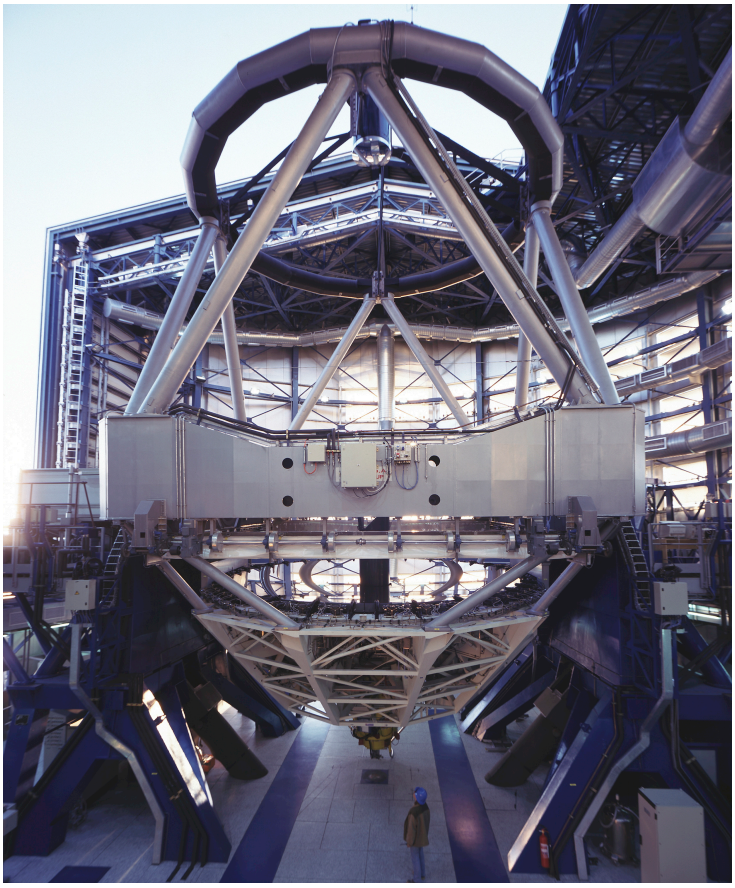
figure 34: Schematic of an alt-azimuth mounting.



A mounting of this type is called an *alt-az* or *alt-azimuth* mount, as its two axes are aligned with respect to the altitude and azimuth axes of the horizontal coordinate system, as shown in figure 34. The tines of the fork are mounted on a rotating pad which acts as the azimuth axis. The altitude axis is formed by mounting both sides of the telescope near the top of the tines. Two examples of alt-azimuth mounts are shown in figure 35.

figure 35: Two examples of alt-azimuth mountings. Left: A photograph of the 8.2m Very Large Telescope in Chile, which is a Ritchey-Chretien design. Right: A photograph of an 8-inch amateur telescope - a Newtonian on a Dobsonian mount. The

Dobsonian was popularised by the amateur astronomer John Dobson and is a particularly cheap and stable alt-azimuth design.



The primary disadvantage of the alt-azimuth design is that it is necessary to move both axes simultaneously in order to track the motion of an object as it moves across the sky due to the rotation of the Earth. Not only this, but the speed that the axes have to be moved varies depending on where the object is in the sky. The calculations required to move the two axes are quite complex (a coordinate conversion between the equatorial and horizontal systems) and need to be performed many times a second. This requires the use of a computer, but computers only became readily available in the 1970's. It is for this reason that every major research telescope made since the 1980's uses an alt-azimuth mount, but every telescope made before this time was mounted equatorially.

Even with computer control, the azimuth axis has a finite speed limit which means it is unable to track objects which pass within a few degrees of the zenith. This is referred to as the *zenith blind spot*, and is more of an annoyance rather than a major problem, as it means that observations have to be halted for a few minutes whilst the telescope catches up with

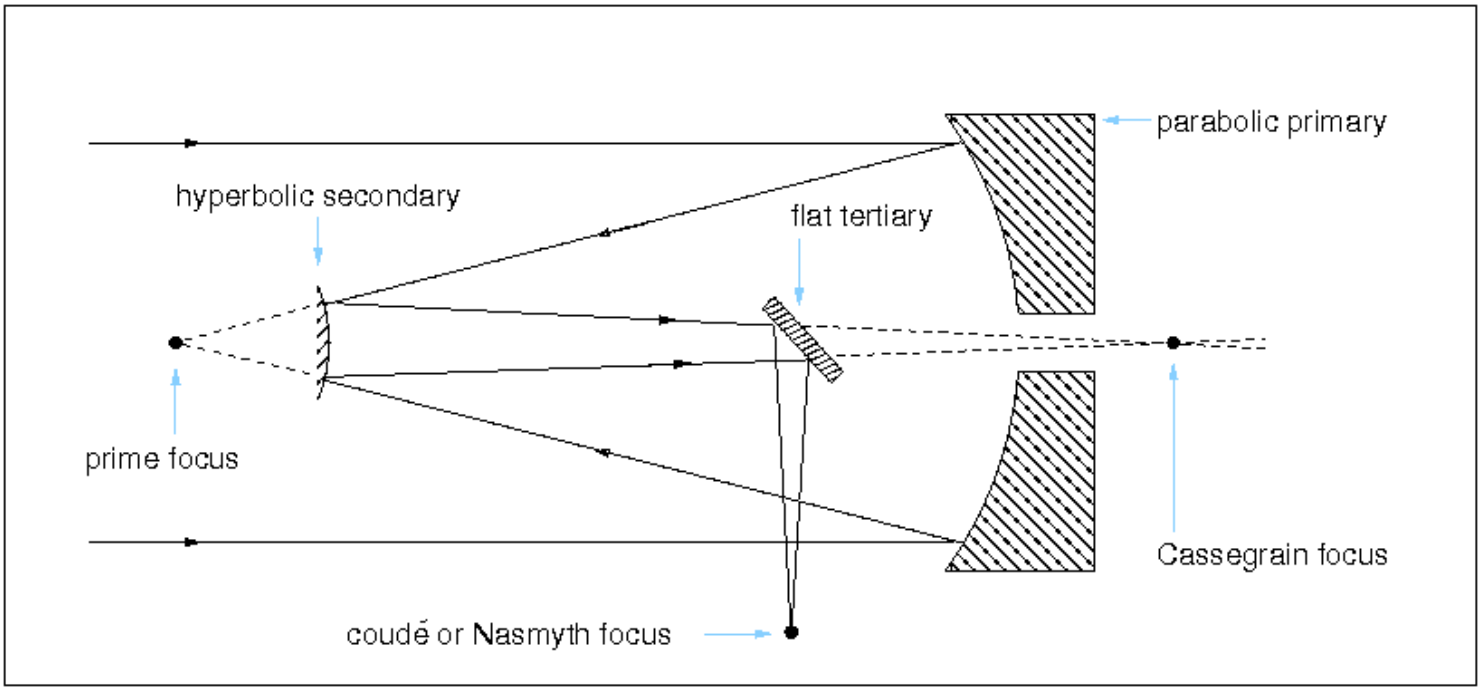
the object on the other side of the zenith. Another problem with alt-azimuth mounts is that the field of view rotates as an object is tracked across the sky, which if uncorrected would lead to the trailing of star-light in an image. To compensate for this, research telescopes are usually fitted with rotating mounts to which the instruments are attached, as shown in figure 38. These rotators rotate the instruments in the same direction and at the same rate as the field of view rotates. A similar effect can also be achieved by keeping the instrument fixed, but using derotation optics in front of the instrument.

---

coudé and nasmyth

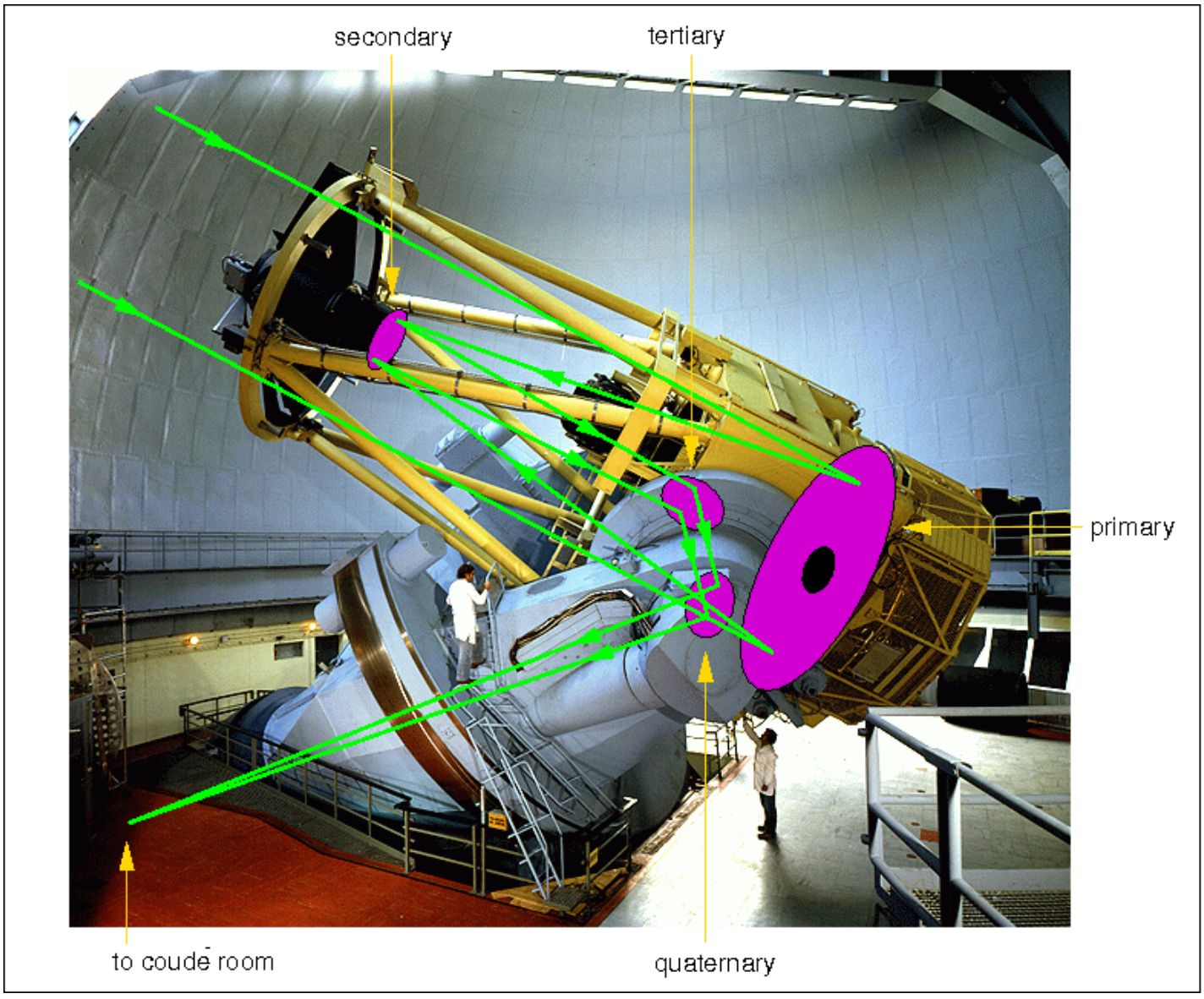
Some astronomical instruments, e.g. high-resolution spectrographs, are too large, too heavy and too sensitive to flexure to be mounted below the primary mirror of a Cassegrain or Ritchey-Chretien telescope. It is possible, however, to provide a stable platform for an instrument by redirecting the light off the telescope using one or more additional mirrors, as shown in figure 36.

figure 36: Schematic of a Cassegrain telescope showing the use of a tertiary mirror to direct light to a coudé or Nasmyth focus. The same diagram applies to a Ritchey-Chretien telescope, except that the primary would be hyperbolic rather than parabolic. In the case of the coudé focus, additional flat mirrors after the tertiary are sometimes employed to direct the light to a given location off the telescope.



The *coudé focus* (from the French word for elbow) is usually found on equatorially-mounted telescopes. A typical coudé design uses a flat mirror (the *tertiary*) to redirect the light along the declination axis of the telescope and then another flat mirror (the *quaternary*) to direct the light down the (fixed) polar axis into a room near the base of the telescope in which the instrument is mounted, as shown in figure 37. The instrument hence remains stationary whilst the telescope moves. In addition to the light lost at each reflection, the main drawback of the coudé design is that the field of view rotates as the telescope tracks an object, so derotation optics are usually required to correct for this (which causes yet more light loss).

figure 37: Left: A photograph of the ESO 3.6m Cassegrain reflector in Chile. A schematic of the four mirrors and light path of a typical coudé arrangement are overlaid. Right: A photograph of a typical coudé room. The light from the telescope comes through the pipe at the upper-left and falls on the instrumentation on the optical bench.

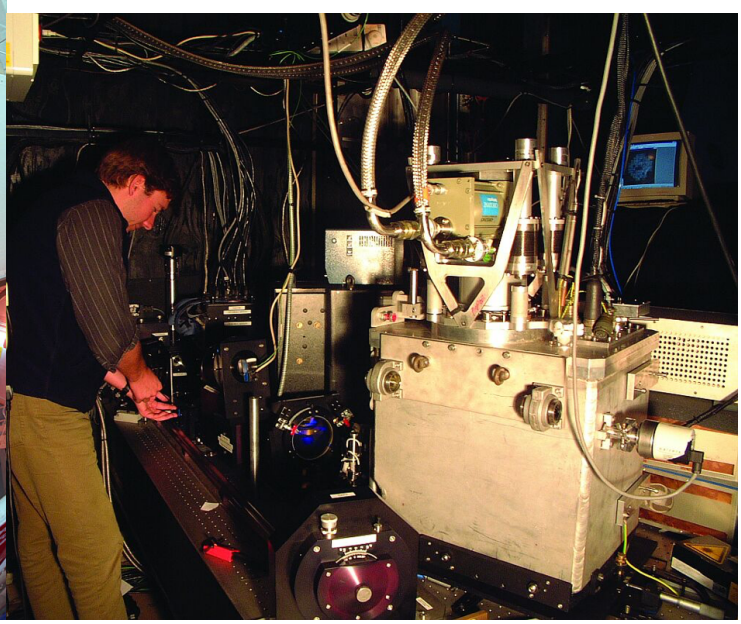


The *Nasmyth focus* (named after the 19th-century Scottish engineer James Nasmyth) is usually found on alt-azimuth mounted telescopes. A tertiary mirror redirects the light horizontally along the altitude axis to the side of the telescope. Hence the Nasmyth focus moves with the telescope azimuth axis, but the beam remains horizontal with respect to the ground. Like in the coudé design, the field of view rotates at the Nasmyth focus whilst tracking an object, but this can be compensated for by using derotation optics or by mounting the instrument on a rotating platform which rotates with the field of view, as shown in figure 38.

figure 38: A photograph of an instrument (ULTRACAM) mounted on one of the Nasmyth focii of the 8.2m Very Large Telescope in Chile.

Many modern research telescopes have two Nasmyth foci, one at each end of the altitude axis. The tertiary mirror can then be tilted to direct light to either Nasmyth focus, or retracted to allow light to pass straight through to the Cassegrain (or Ritchey-Chretien) focus. This configuration allows three different instruments to be permanently attached to the telescope, each accessible at the flick of a switch, giving astronomers significantly more flexibility. Some telescopes have horizontal platforms at the tops of the tines, such as the one the person is standing on in figure 38. It is possible to mount particularly large, heavy or sensitive instruments horizontally on these Nasmyth platforms using an optical bench and derotation optics. These instruments are often enclosed in light-tight, temperature-controlled laboratories, such as shown in figure 39.

figure 39: Left: A photograph of the 4.2m William Herschel Telescope on La Palma showing the two laboratories (GHRIL and GRACE) mounted at the two Nasmyth focii. It should be noted that both labs rotate with the azimuth axis of the telescope. The black turret at the centre of the image houses the tertiary mirror. Right: A photograph of the inside of one of the laboratories. Light from the telescope enters the lab just behind where the person is standing, passing through some derotation optics before entering the complex instrumentation (an adaptive optics imager) on the optical bench.



---

tubes and trusses

The tube or trusses of a telescope must perform the task of holding the optical components at the correct separation and at the correct relative alignment, regardless of the orientation of the telescope on the sky. Smaller telescopes tend to use tubes and larger telescopes trusses - we shall look at each in turn.

tubes

Most amateur telescopes use tubes to mount the optics because they are simple, cheap and can be made stiff enough to hold small-aperture mirrors and lenses without becoming too heavy to mount. A tube also has the advantage of providing a closed optical system, which helps to reduce stray light (and hence improves image contrast) and protects the inner surfaces of the lenses and/or mirrors from dirt and dew. As shown in figure 40, tubes can be made from a number of different materials, including steel, aluminium, carbon fibre, fibreglass and plastic, with the ideal material being stiff, light-weight, cheap and possessing a low coefficient of thermal expansion. The latter property ensures that the distance between the optical components of a telescope does not alter significantly as the temperature changes, although this can usually be compensated for by refocusing.

There are two main disadvantages to using tubes. The first, which applies primarily to open-ended tubes such as used with Cassegrain telescopes, is that they are susceptible to *tube currents*, where the trapped volume of warm air in the tube mixes turbulently with the colder outside air, degrading the seeing. This is not such a problem with closed-ended tubes, such as used with refractors and Schmidt-Cassegrain telescopes, but even closed-tube telescopes are often fitted with fans and vents to bring the optics into thermal equilibrium with the surroundings as quickly as possible. The second disadvantage of tubes is that they cannot be used to mount really large optics - to retain rigidity, they simply become too

heavy, placing too much strain on the telescope mounting.

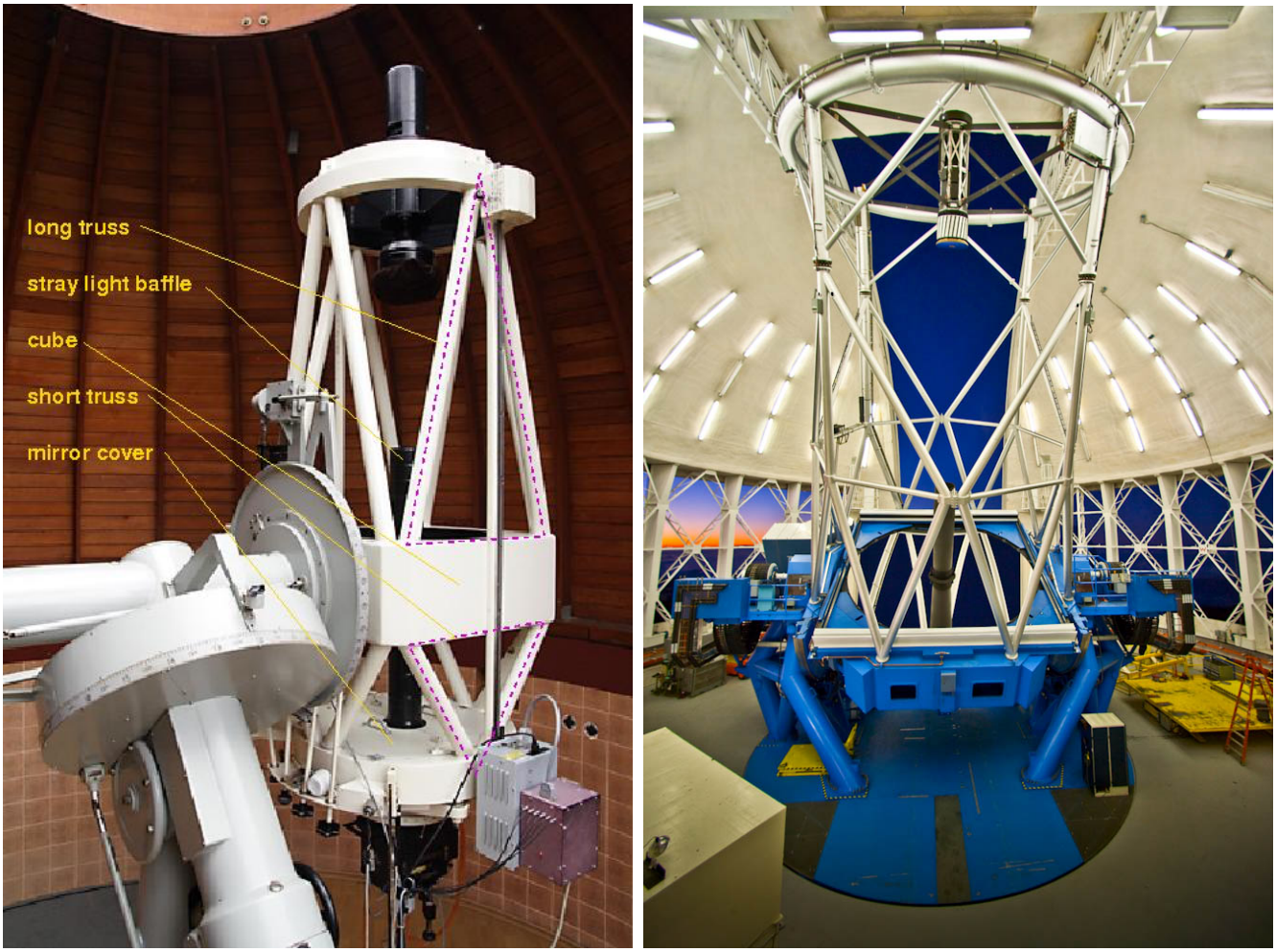
figure 40: Examples of different telescope tubes. Left: photograph of the 26-inch Thompson Refractor at Herstmonceux, which has a closed-ended steel tube. Right: photograph of a 10-inch Newtonian telescope with an open-ended carbon-fibre tube. Note the fans and ventillation slots at the base of the primary mirror.

trusses

A truss is a rigid structural element which is usually made up of triangular shaped units. Triangles are used because each of a triangle's three internal angles are fixed by the side opposite it. Hence it is impossible to change the angle between two sides of a triangle without changing the length of the side opposite it. This means that if a load is applied anywhere on a triangle, its angles, unlike those of other shapes (like squares), will not change.

The most commonly used truss design in telescopes is the *Serrurier truss*, named after its inventor in 1935, the US engineer Mark Serrurier. The Serrurier truss is composed of eight individual trusses, four holding the primary mirror and four holding the secondary mirror. The two sets of four trusses are joined at the declination/altitude axis, in a structure generally referred to as the *cube*, as shown in the left-hand panel of figure 41.

figure 41: Left: Photograph of the 0.6m Ostrowik telescope in Poland, which employs a Serrurier truss design. Right: Photograph of the 8.1m Gemini North telescope on Hawaii.



The lengths of the short and long sets of trusses are selected so that they both flex by exactly the same amount, keeping the primary and secondary mirrors on the same optical axis regardless of the telescope orientation. In addition, when tipped over horizontally, the vertical pair of trusses oppose bending by gravity, while the horizontal pair form a parellogram with the top ring and cube (or the cube and primary mirror cell), thereby ensuring that the two mirrors always remain parallel to each other, maintaining optical alignment and hence image quality.

Almost all telescopes larger than ~1 m in aperture use Serrurier trusses, as only this design can provide a stiff yet light-weight structure to keep the mirrors aligned. Telescopes with Serrurier trusses are open, and hence they do not suffer from tube currents and are able to come into thermal equilibrium with their surroundings much more quickly than telescopes

with tubes. The disadvantage of the open design is that the optics, particularly the upward-facing primary mirror, are exposed to the environment and hence must be covered when not in use and cleaned/re-aluminised regularly. The open design is also susceptible to stray light, although this can be alleviated to some extent through the use of baffles (see figure 41). Some large telescopes use a variant of the Serrurier truss, as shown in the right-hand panel of figure 41, where the cube is eliminated in order to minimise the thermal input into the light beam of the telescope (which is of importance for infrared observations).

---

telescopes

IV. domes and sites

the atmosphere

Astronomers studying the Universe from the surface of the Earth have to peer through the atmosphere. This has a number of undesirable consequences:

- *Sky background:* Dust and molecules in the atmosphere scatter light from the Moon and the surface of the Earth (e.g. street lamps), making the entire sky appear to glow. The molecules in the atmosphere also emit light. The sky therefore acts as a noisy background to the light received from astronomical objects, making the detection of faint targets particularly difficult.

- *Atmospheric extinction:* Dust and molecules in the atmosphere scatter and absorb the light from astronomical objects, dimming the images obtained by astronomical telescopes. The amount of dimming depends on the angle of the object above the horizon and the local conditions in the atmosphere at the time of observation. This type of extinction is prefixed with the word *atmospheric* in order to distinguish it from interstellar extinction.

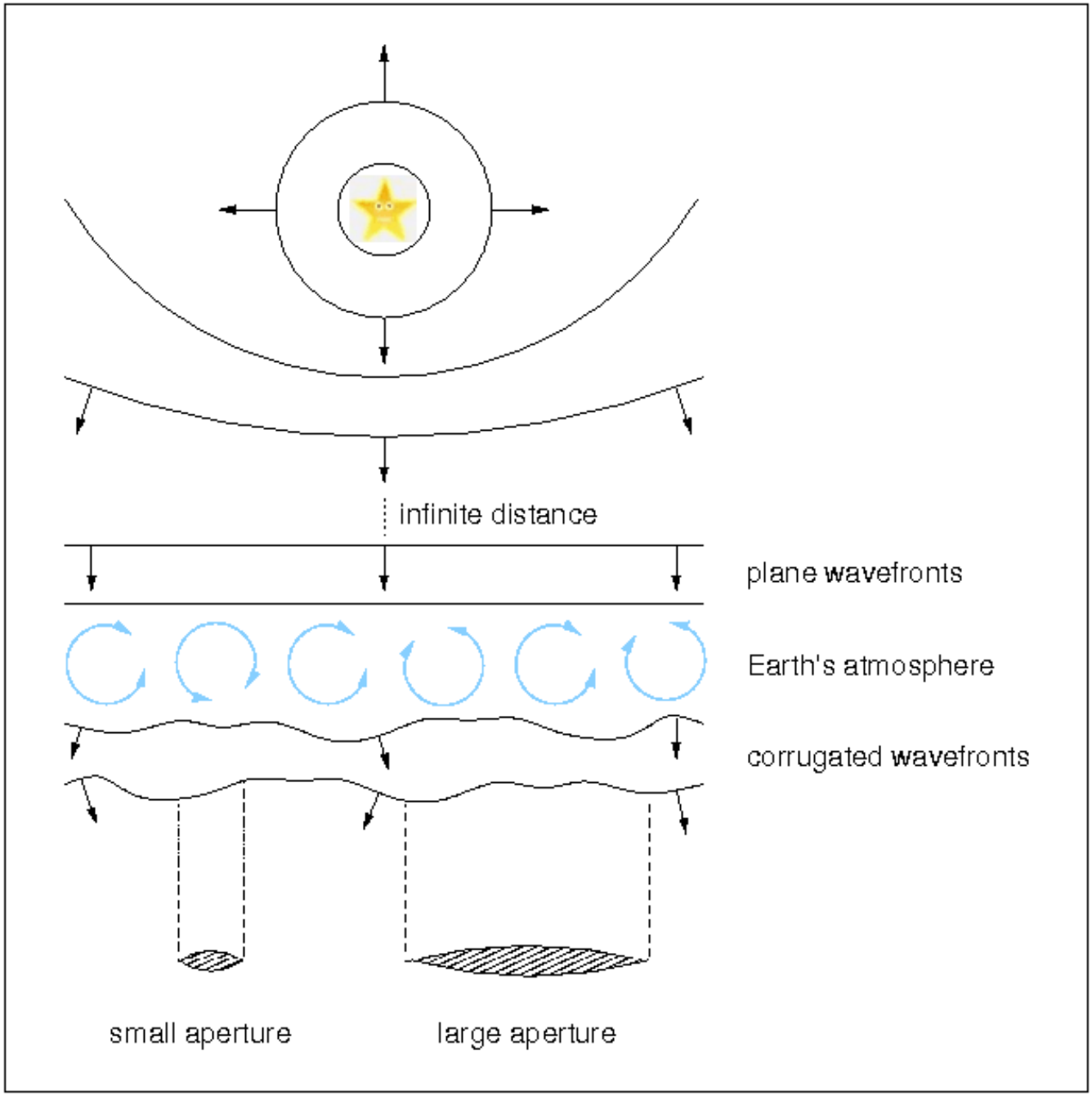
- *Transparency variations:* Related to atmospheric extinction, clouds (i.e. water droplets) can absorb and scatter the light from astronomical objects. The amount of absorption tends to be much more variable than that due to extinction, as the clouds are blown across the field of view of the telescope by the wind. The resulting variations in the amount of light received from an astronomical object range from partial attenuation due to thin cloud, to complete obscuration by thick cloud.

- *Seeing:* Completely unrelated to the above effects, turbulence in the atmosphere degrades the resolution of the image recorded by a telescope. We shall look at the causes and effects of seeing in more detail below. The turbulence also causes a variation in the brightness of an image recorded by a telescope, a phenomenon known as *scintillation* (or, less formally, *twinkling*), but this will not be considered further here.

It is very important not to confuse seeing and transparency. The two effects are largely unrelated: in poorer seeing, the image of a star will be more blurred, but its brightness will remain approximately constant; in poor transparency, the light from a star will be dimmed but its blurring will be largely unaffected.

Turbulence in the atmosphere occurs on all scales and results in adjacent pockets of air with slight temperature differences between them. Since the temperature of air affects its density, which in turn affects its refractive index, some rays of light from an astronomical source are bent by more than others. In terms of wavefronts, the plane wavefront from a star is corrugated by the atmosphere, as some parts of it are retarded in phase by more than others. This is shown schematically in figure 42.

figure 42: Schematic showing the effect of atmospheric turbulence on image formation in both a small and large-aperture telescope.

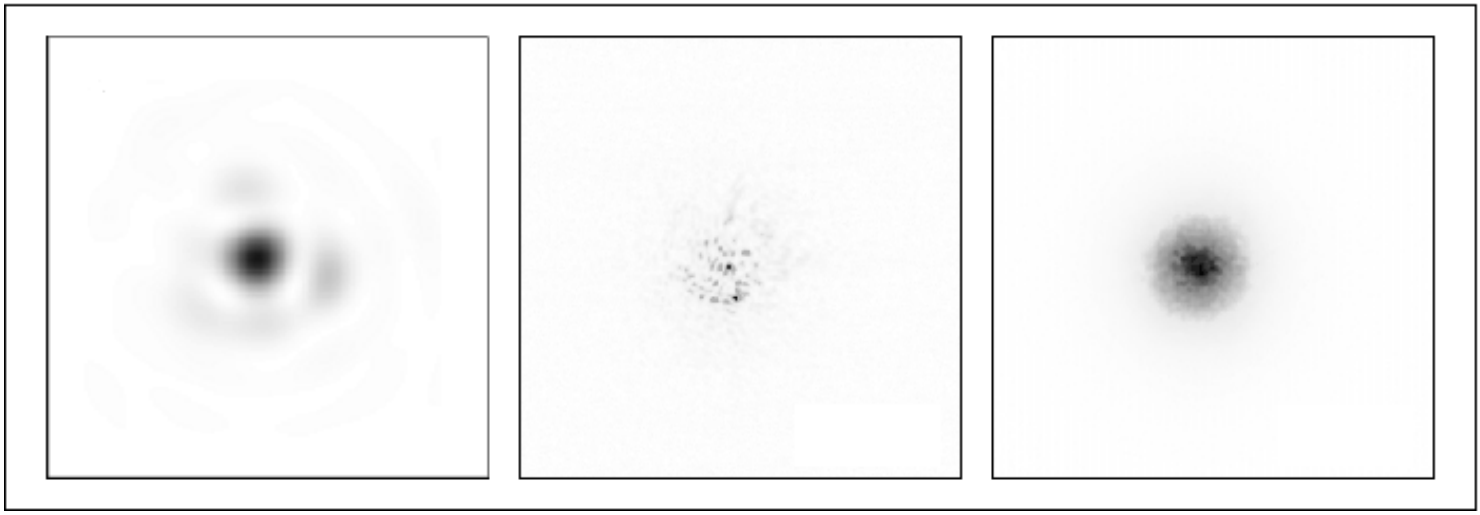


It can be seen from figure 42 that the image of a star will look very different through telescopes of different aperture. Small-aperture telescopes will collect portions of the wavefronts which do not have many corrugations. We have already seen that straight portions of a wavefront produce diffraction-limited images, hence the instantaneous image of a star in a small-aperture telescope (of order 10 cm, the typical size of the straight portion of a wavefront) will generally appear diffraction limited, as shown in the left-hand panel of figure 43, but the image will dance around on timescales of a fraction of a second. This is because the straight portions of successive wavefronts have different slopes with respect to the telescope aperture and hence will be focused onto different points in the focal plane.

The effect of exposing for many seconds or minutes then is to average out this image motion into a single blurred image of the star, similar to that shown in the right-hand panel of figure 43, This is known as the *seeing disc*.

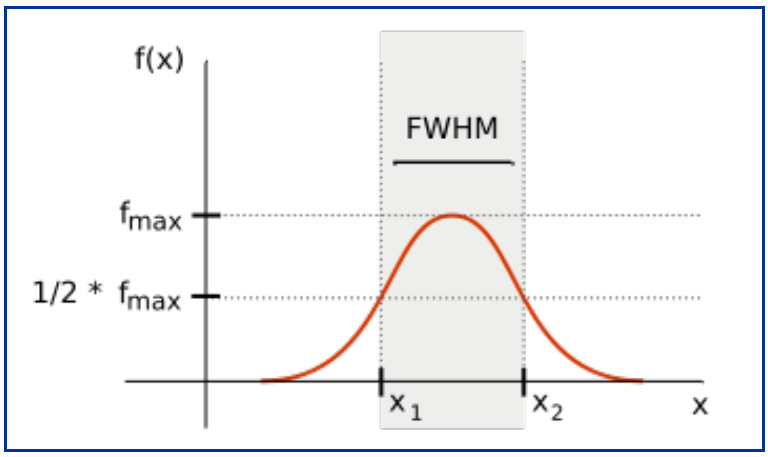
Large aperture telescopes (>> 10 cm) collect wavefronts with many corrugations. Each of the straight portions of the wavefront are simultaneously brought to a sharp focus, but at positions in the focal plane dependent on their tilts. Hence the instantaneous image appears as small bright spots, or *speckles*, superimposed on a faint, blurred disc, as shown in the central panel of figure 43. The effect of accumulating many such images over the course of seconds or minutes is to average them out into a seeing disc similar to that seen through a small telescope, as shown in the right-hand panel of figure 43. Note that, since each speckle in the instantaneous image is made up of all of the straight portions of the wavefront with the same tilt, and the maximum separation of these straight portions is given by the diameter of the aperture, the speckles will have a typical size given by the diffraction limit of the aperture. In other words, each speckle is a diffraction-limited image produced by the large-aperture telescope, explaining why the speckles appear so small in the central panel of figure 43.

figure 43: Left: Example of a short-exposure (of order milliseconds) image of a point source through a ~10 cm ground-based telescope. The fact that the first diffraction ring can just be seen surrounding the Airy disc implies that the instantaneous resolution is close to the diffraction limit, but significant image motion would cause blurring in a longer exposure, similar to that shown in the right-hand panel. Centre: Short-exposure image of a point source through a ~1 m ground-based telescope. The image is broken up into bright, dancing speckles, which are smeared-out in the longer-exposure image shown in the right-hand panel.



A measurement of the full-width at half-maximum (FWHM) of the seeing disc gives a numerical value for the seeing, as shown in figure 44. The value x_2-x_1 is typically measured in pixels from an astronomical image and then converted to arcseconds using the plate scale of the telescope. Seeing as low as 0.1 arcseconds has been recorded on the Earth's surface in Antarctica. The typical seeing at premier astronomical observatories, such as those found in Hawaii, Chile or the Canaries, is between 0.5 and 1.0 arcseconds. The best we've recorded from the roof of the Physics building in Sheffield is about 2 arcseconds! Given that the diffraction limit of a 10 cm telescope is approximately 1 arcsecond, it can be seen that in all but the smallest-aperture telescopes, the resolving power of a ground-based telescope is limited not by diffraction but by the seeing.

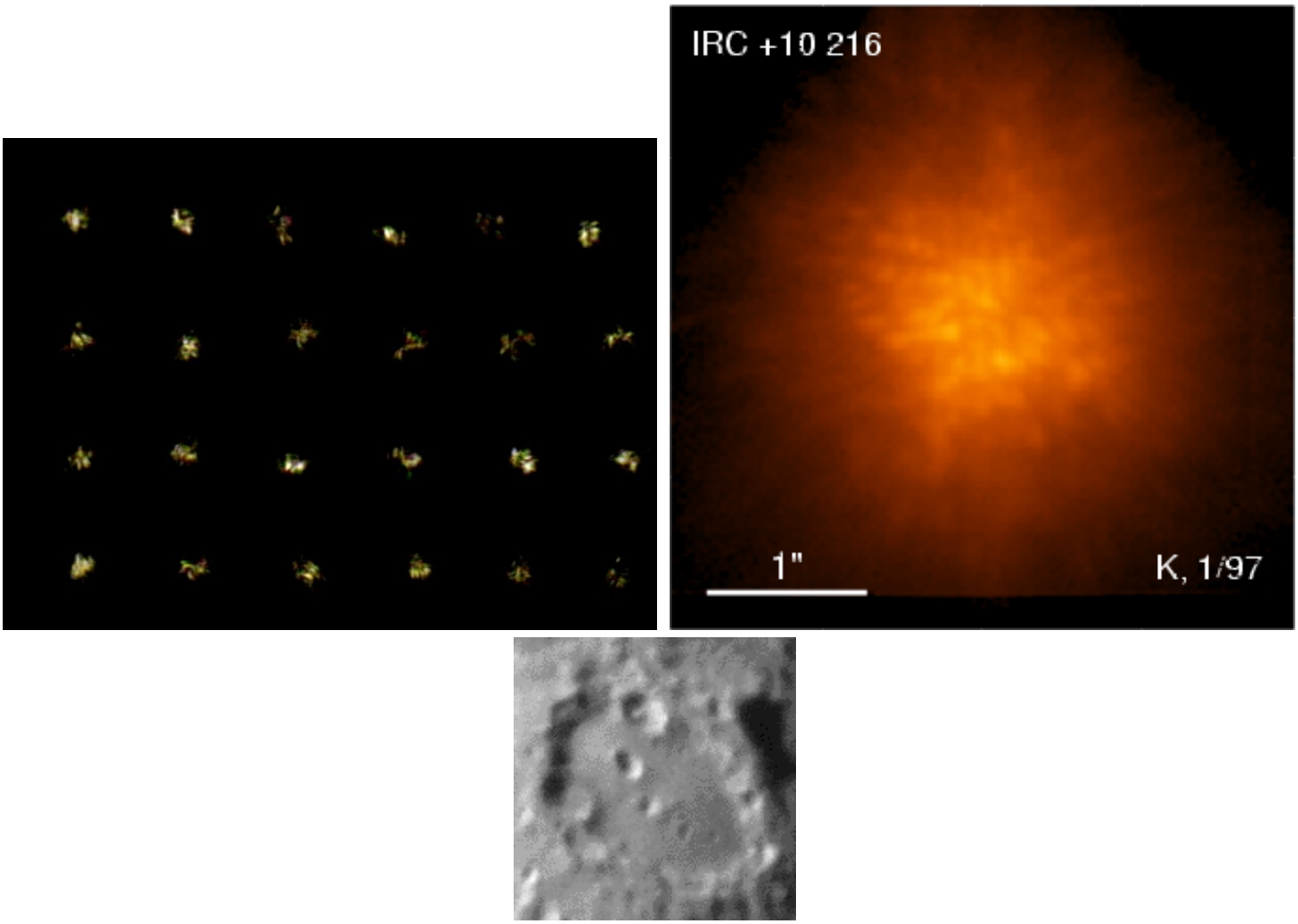
figure 44: Schematic showing the seeing profile of a star and how the FWHM is measured.



A good way to appreciate the disastrous consequences of astronomical

seeing on image quality is to look at movies of both point sources and extended sources, such as those shown in figure 45. To minimise the effects of the seeing shown in the movies below, astronomers continually seek new places to site their telescopes, a topic we shall turn to next.

figure 45: Examples of the effects of astronomical seeing on image quality. Left: Images of a star obtained on a large amateur telescope using a webcam with 0.01 s exposures. Each image is separated in time by about 1 second. Centre: A movie of a star observed with the Russian 6 m telescope, showing the speckle pattern. Note the image scale. Right: An amateur astronomer's movie of the lunar crater Clavius taken in bad seeing conditions.



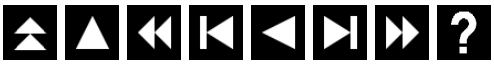
Scintillation (*For Advanced Readers*)

As well as blurring the images of stars, atmospheric turbulence also induces intensity variations, known as *scintillation*. A detailed description of this phenomenon can be found in the paper by Osborn et a. (2016), and

references therein. It should be noted that scintillation is actually an interference phenomenon, caused by diffraction of formerly plane-wave light distorted by atmospheric turbulence (see Little 1951, Chandrasekhar 1952 and this web page by Kuehne).

---

observatory sites



There are no optical telescopes in the UK which can now be regarded as world class. Arguably, the same applies to the whole of continental Europe. The reason is simple: the world's largest telescopes now cost hundreds of millions of pounds to build, and hence it makes sense to site them in the best possible places for astronomical observing, none of which are located in Europe. The factors which dictate whether or not an astronomical site is a good one are discussed below.

latitude

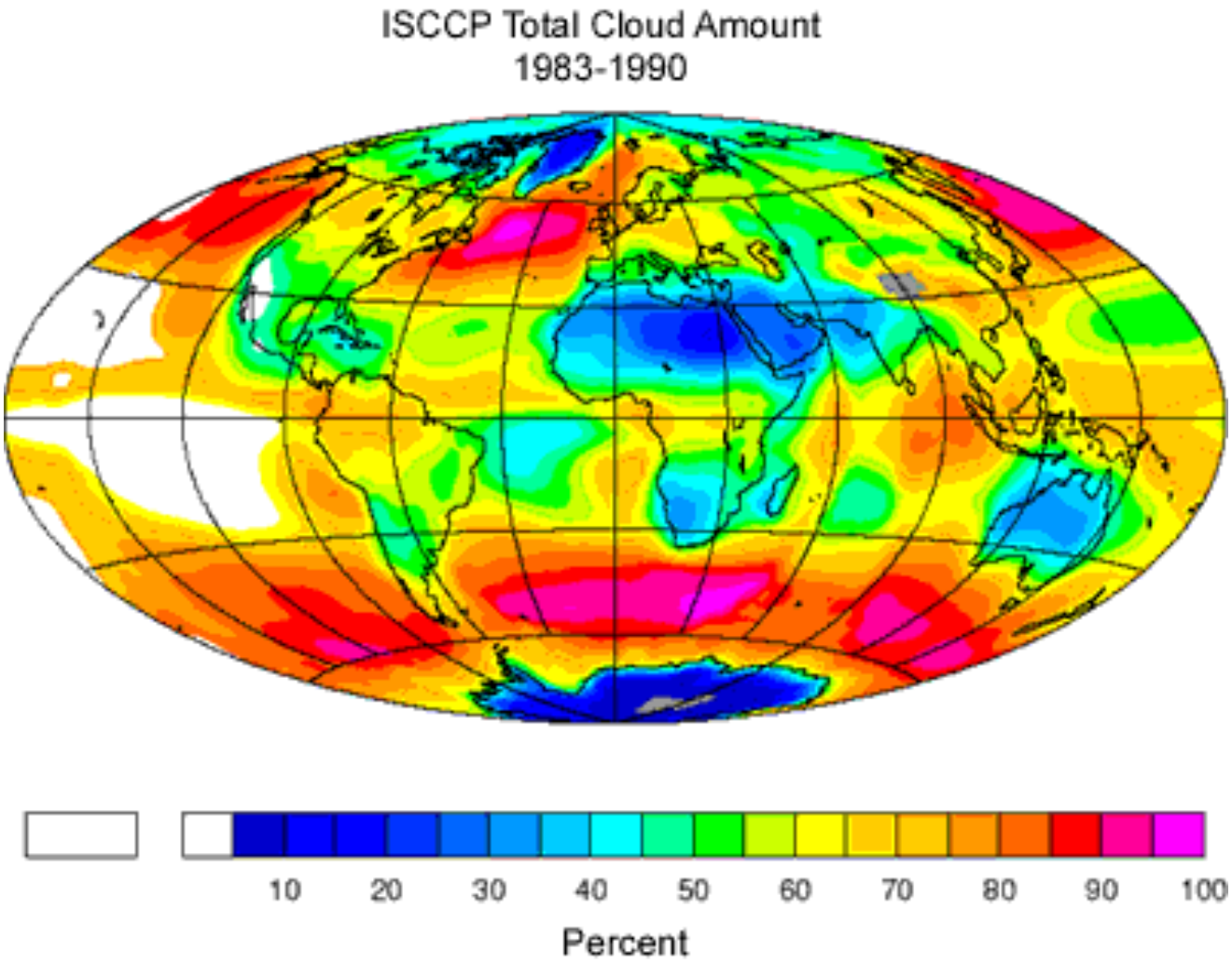
The first decision to be made is whether to site a telescope in the northern or southern hemisphere. The southern sky is arguably the more interesting, as it contains both the galactic centre and the two nearest galaxies to our own - the Small and Large Magellanic Clouds. The northern sky has an historic advantage in that it is arguably the better studied sky, hence there tend to be more members of each class of astronomical object known in the north (although this situation is rapidly changing). Neither argument is overwhelming, however, and the fact is that telescopes are required in both hemispheres in order to access the entire sky. Telescopes sited close to the equator can access much of both hemispheres, of course, so this is often a good compromise. Siting telescopes closer to the poles has the disadvantage that the Sun never gets sufficiently below the horizon in the summer months for the sky to become truly dark, although the longer winter nights can be advantageous for certain astronomical projects, e.g. monitoring stellar variability.

cloud cover

Although radio telescopes can peer through cloud, optical telescopes cannot. Hence, finding a site with relatively little cloud cover means that a

telescope can be used for a larger number of nights per year, maximising the investment made in the facility. This is one of the primary reasons there are no major research telescopes in the UK! Desert regions, including the Arctic, Antarctica, Australia, parts of Africa and the western coast of the Americas appear to provide some of the best cloud conditions in this respect, as shown in figure 46.

figure 46: A map of the world's percentage cloud cover, averaged over a 7-year period.

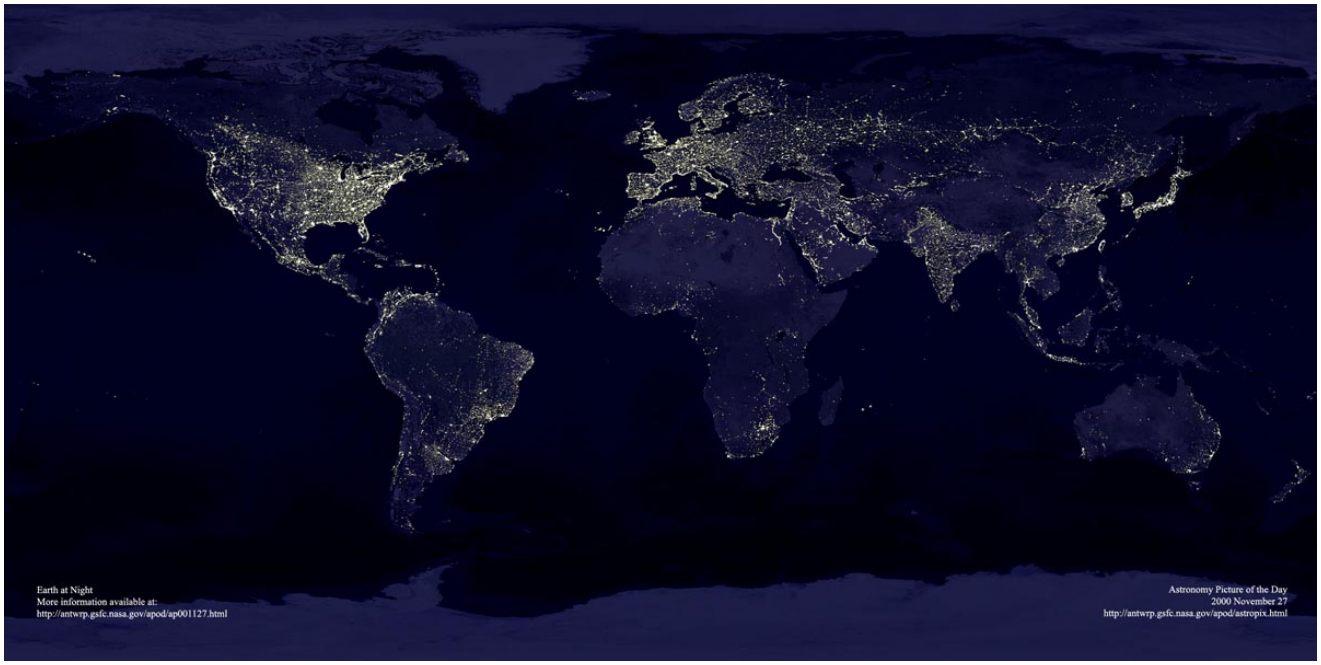


light pollution

Light from terrestrial sources, such as street lights and industry, increases the amount of background light detected by telescopes. This increases the background noise in astronomical images, making it more difficult to

detect faint astronomical sources. Since the faintest astronomical sources often turn out to be the most interesting ones, the world's largest telescopes are sited in regions with low light pollution. Figure 47 shows a map of the world's artificial night sky brightness. It can be seen that some of the regions which show low cloud cover in figure 46, such as the sparsely populated western coasts of South America and Africa, are also largely free of light pollution. Figure 47 also clearly shows why no major telescopes are now built in mainland Europe.

figure 47: A map of the artificial night sky brightness in the world.



seeing

Selecting a site with good seeing is of prime importance, as lower seeing improves both the spatial resolution and the signal-to-noise ratio of astronomical images, thereby exploiting the apertures of the world's largest telescopes to their maximum. The atmosphere over the sea tends to be much less turbulent than the atmosphere over land, as the sea exhibits an essentially smooth, constant-temperature surface compared to the land. Some of the best astronomical sites are therefore located on small islands in the middle of oceans, such as Hawaii and the Canaries, as these small land masses cause little additional turbulence. For the same reason, coastal regions that receive winds predominantly from the direction of the ocean, such as the western coasts of the Americas and

Africa, also exhibit excellent seeing.

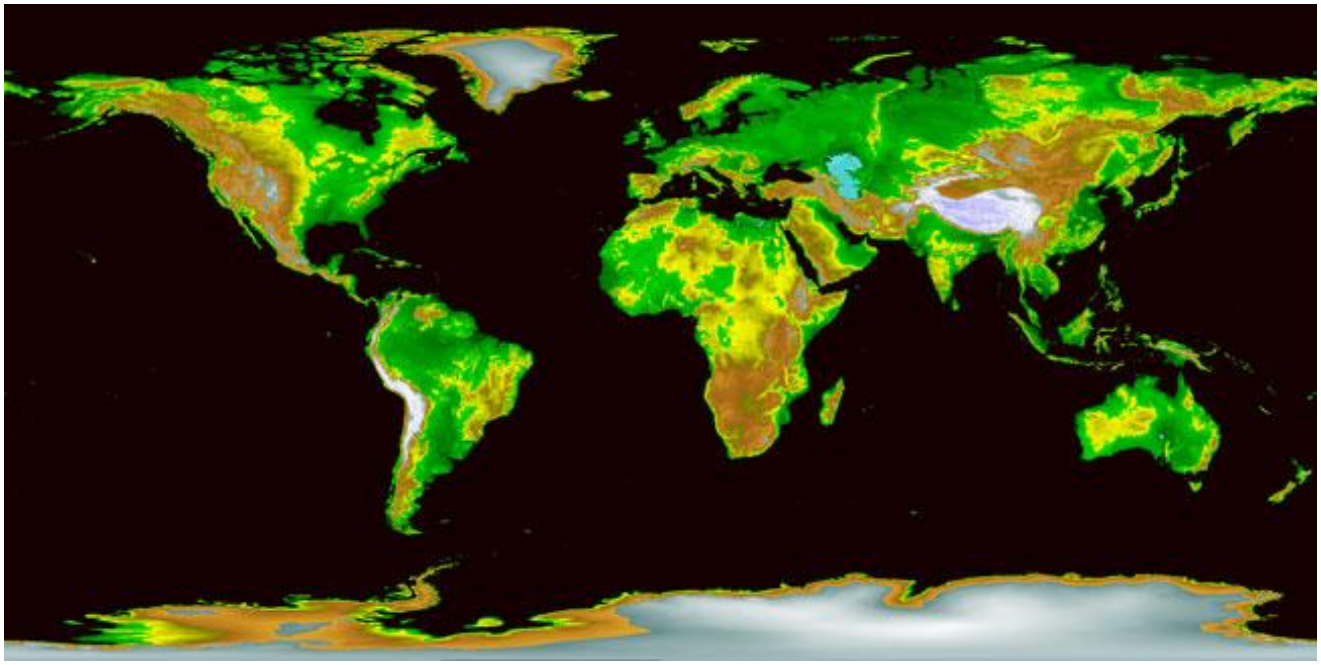
The best seeing on Earth has recently been shown to exist in Antarctica. Atmospheric turbulence is caused by heat from the ground rising through the atmosphere, and wind stirring the atmosphere up. At mid-latitudes, there are numerous layers of strong winds (e.g. the jet stream) which cause lots of turbulence; above Dome C on the high plateau in central Antarctica, however, the ground is cold and winds are low throughout the atmosphere, leading to low turbulence and hence excellent seeing.

height above sea level

Siting telescopes at altitudes of approximately 2000 m above sea level provides a number of advantages for astronomical observations. It places the telescope above a significant fraction of the Earth's atmosphere, reducing the thickness of air that sources are observed through. This reduces the atmospheric extinction and can also reduce the seeing if some of the turbulent layers lie below the altitude of the telescope. High-altitude telescopes are also often above the local inversion layer in the atmosphere, meaning that local cloud formation occurs below the telescope, significantly increasing the number of usable nights at the observatory compared to a telescope sited below the inversion level.

Figure 48 shows that the mountainous western coasts of the Americas and Africa provide favourable altitudes to site astronomical telescopes, as do the volcanic island-chains such as Hawaii and the Canaries. The Antarctican plateau, and parts of the Arctic, are also at a high altitude.

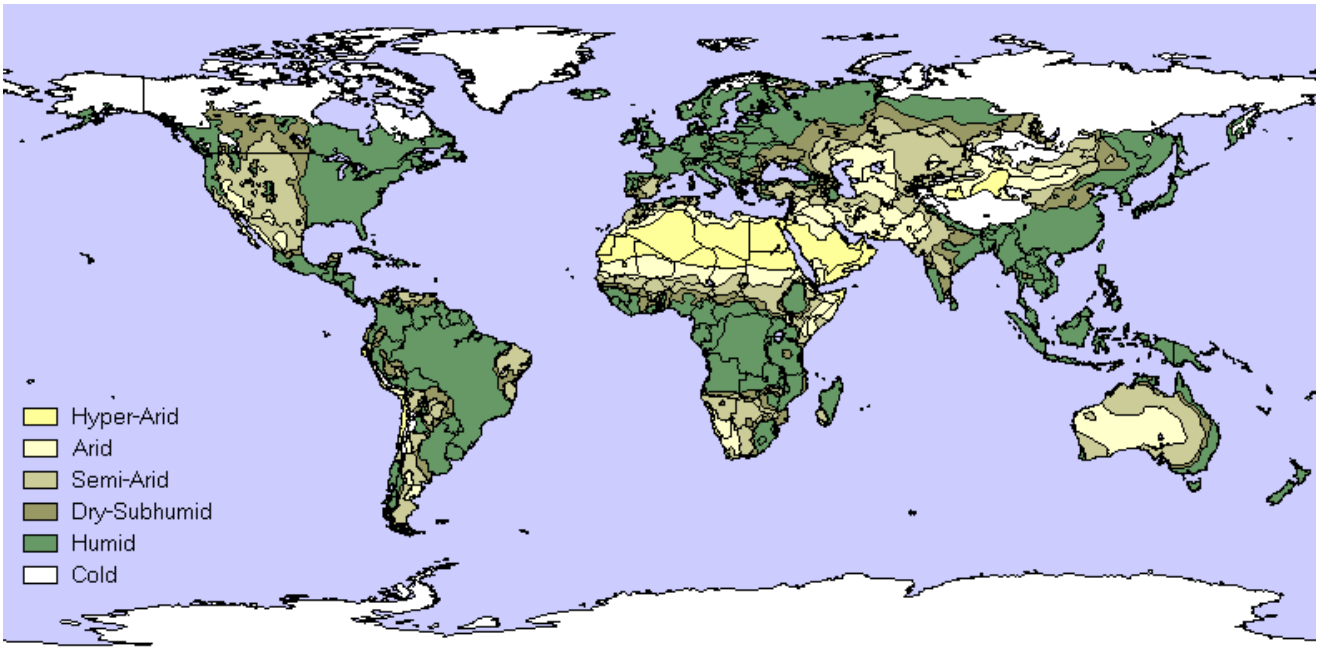
figure 48: A colour-coded map of the world where areas in white, brown, yellow and green indicate decreasing height above sea level.



humidity

Water vapour in the Earth's atmosphere has only a marginal impact upon optical observations, but it is a very significant source of absorption of the infrared light from astronomical sources. Since most large optical telescopes built today have also been designed to operate in the near infrared, finding sites with low water-vapour content is of importance. Figure 49 shows that the western coastal regions of the Americas and Africa are extremely arid (in fact, the Atacama desert in Chile is the driest place on Earth). Counter-intuitively, although not apparent from figure 49, parts of the Antarctican plateau and the Arctic also exhibit very low humidity.

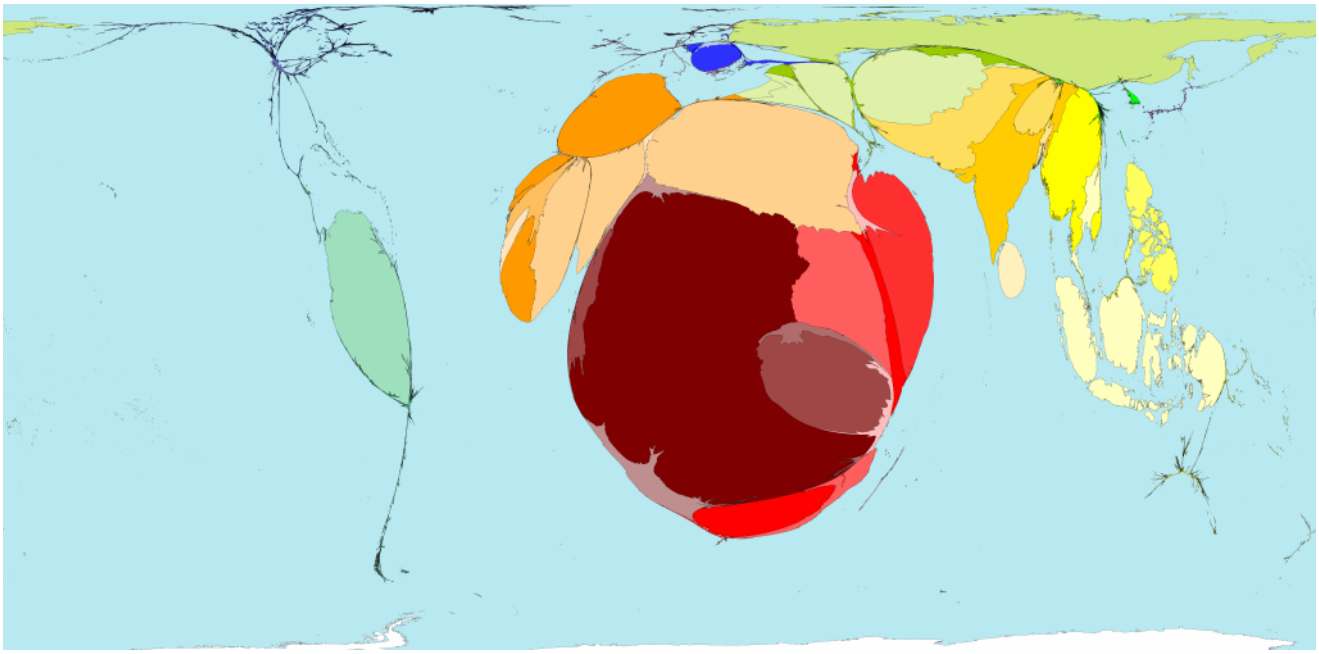
figure 49: A map showing how the average humidity varies with location in the world.



political stability

As well as geographic and atmospheric factors, political factors are also important when determining where to site a major telescope. For example, figure 50 shows the war zones in the world in 2002. Clearly, it would be extremely difficult to build and operate a telescope in a war zone. The same is true of any region in the world which is politically unstable - it is essential that the huge investment involved in building and operating a major telescope is not jeopardised by problems in recruiting staff, buying materials and services, accessing the site, etc.

figure 50: A map of the world in which territory size shows the proportion of deaths worldwide in 2002 directly attributed to war or conflict that happened there.



other factors

There are many other factors which come into play when selecting the optimum site for an observatory, some of which are related to those discussed above. These include:

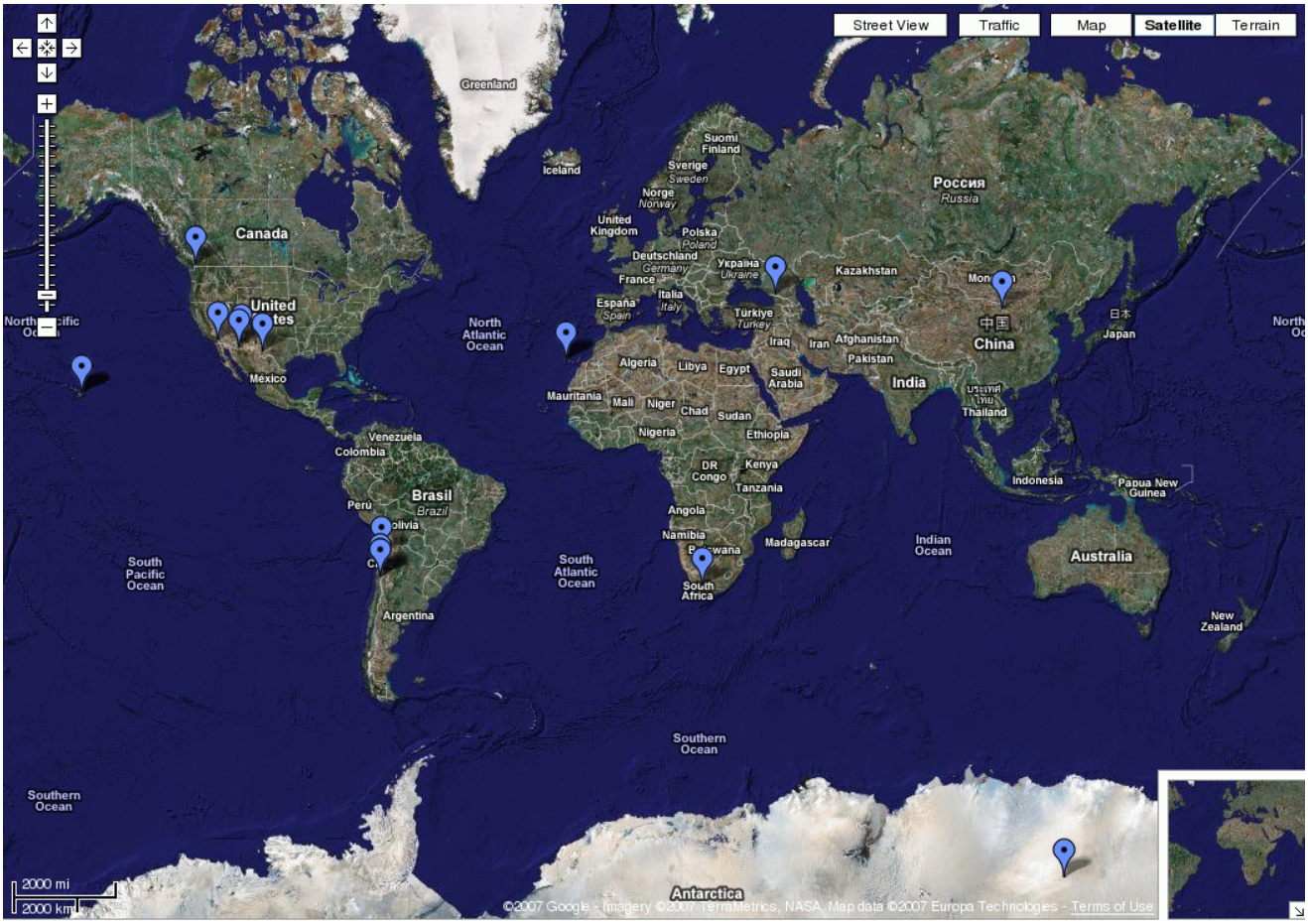
- seismic activity - telescopes built in seismically-active areas must be designed to survive strong earthquakes, which increases the complexity and cost of the facility. The Andes, for example, are very seismically active, whereas Antarctica is not.
- auroral activity - polar telescopes may be affected by light pollution from aurorae.
- rainfall - this is linked to cloud cover and humidity. Clearly, it is desirable to build telescopes in areas with low precipitation, as domes must be kept closed in the rain to prevent damage to the telescope.
- wind speed - strong winds can shake a telescope, blurring the images it produces. It is possible to design telescopes and domes in such a way that they are protected from wind shake and wind-blown debris (e.g. dust), but there are limits to how well these work when winds become very strong.
- dust/aerosols - widespread dust/aerosols in the atmosphere, such as produced by sandstorms in desert regions, or industrial pollution, causes additional extinction and hence attenuation of star-light. For example, due to their proximity to the Sahara, the Canary Islands are

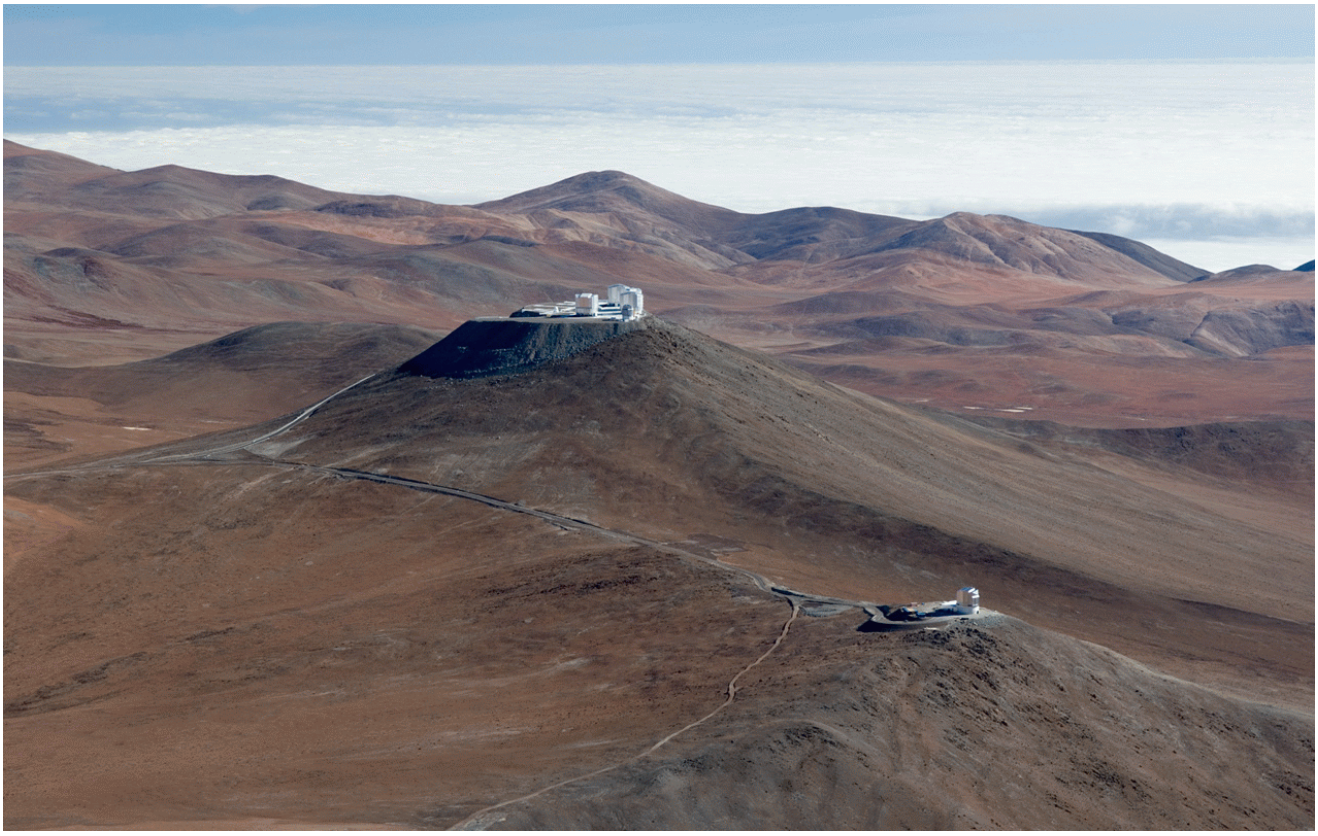
particularly susceptible to this.
- temperature - the sky glows in the infrared, significantly increasing the background noise in infrared images of astronomical sources. At cold sites, such as in Antarctica, this thermal emission from the sky is reduced, significantly improving the sensitivity of infrared observations.
- accessibility - sites must be accessible in order to build and operate telescopes there. Very remote telescopes without road access, for example, are extremely costly.

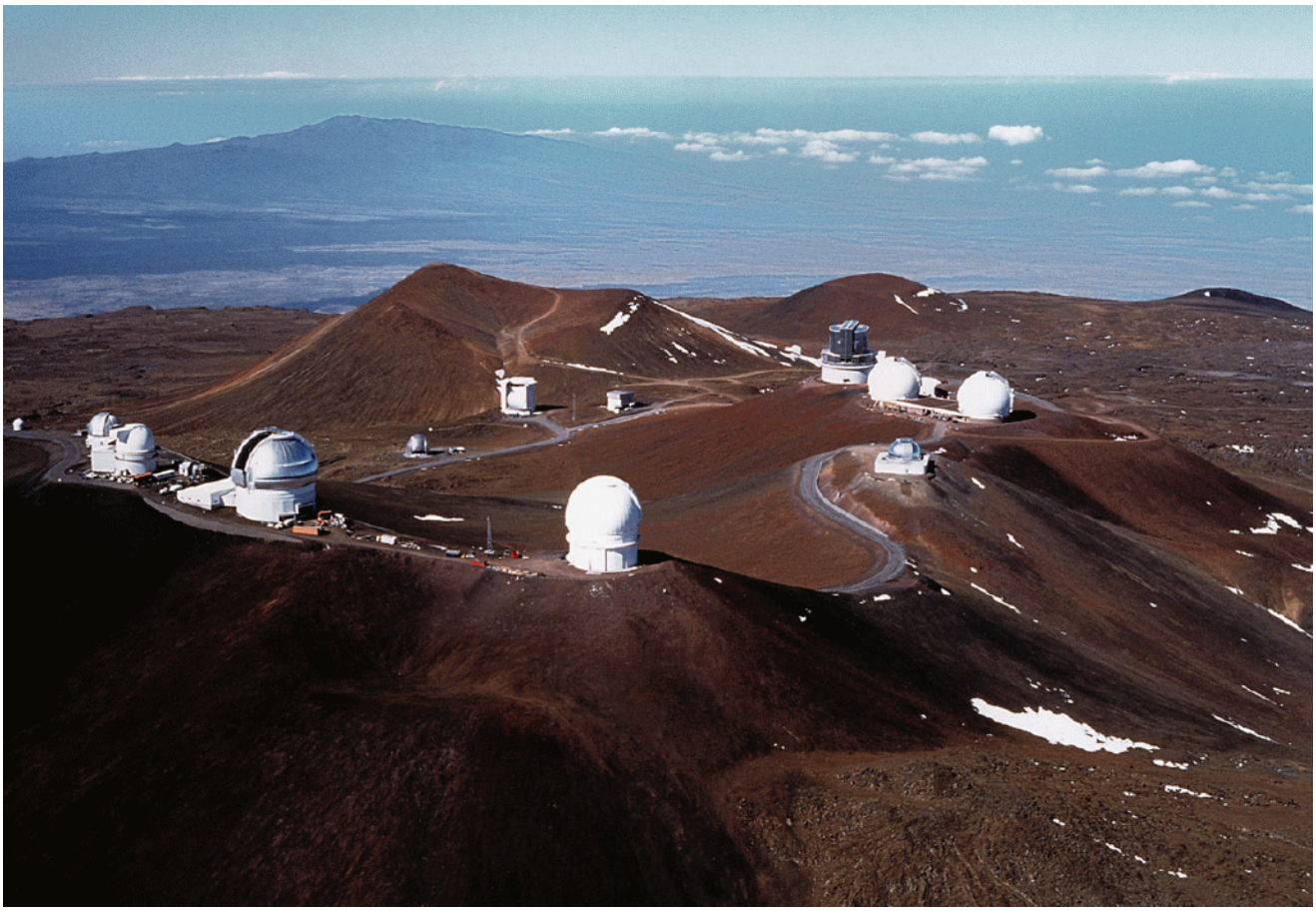
sites of the world's largest telescopes

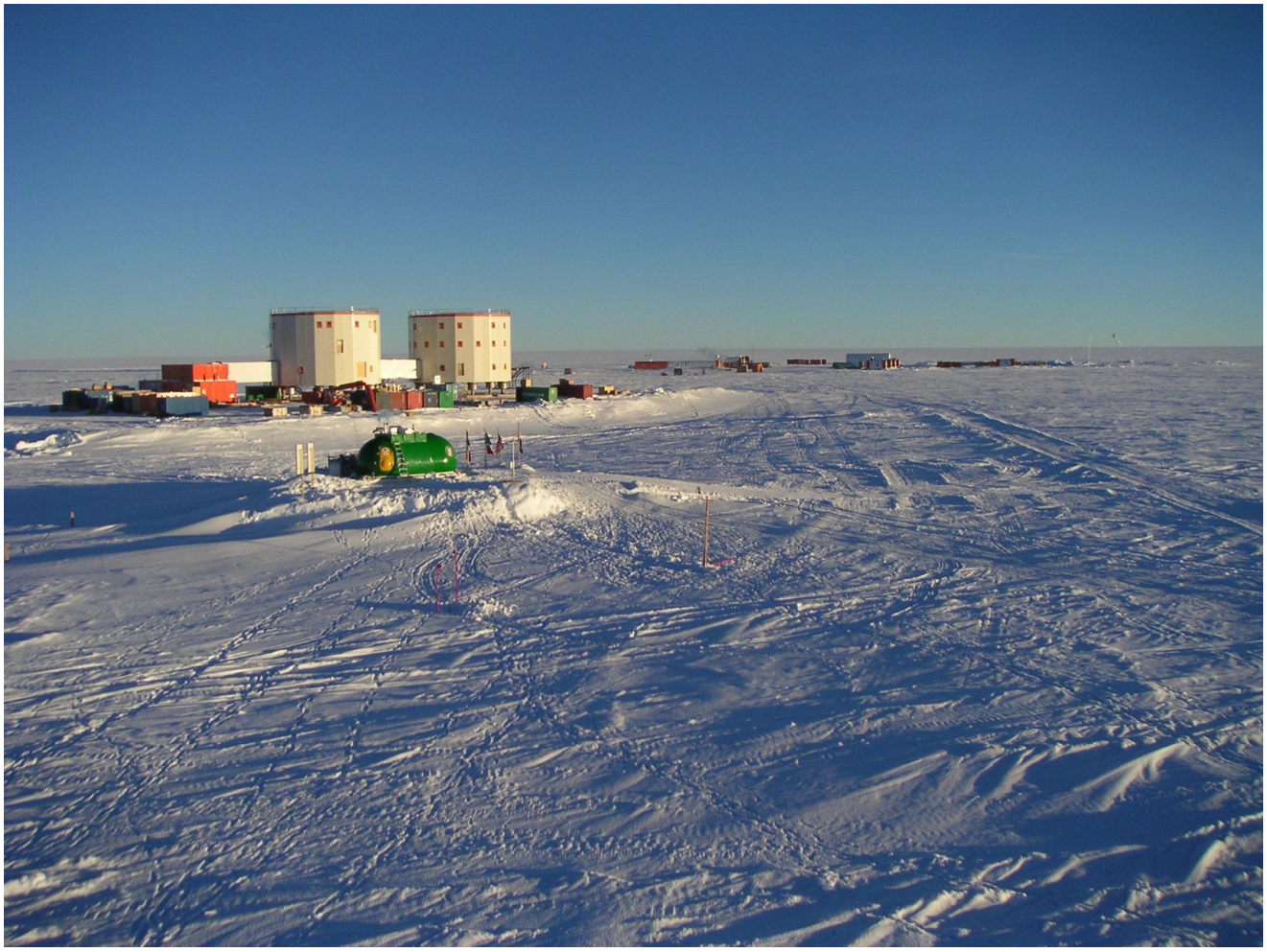
Figure 51 shows the sites of the world's largest telescopes. They are tightly clustered in only a few major sites, predominantly Chile, Hawaii and the Canary Islands. The reasons for this are clear from the preceding discussion: these sites have low cloud cover, low light pollution, good seeing and low humidity. They are all at high altitude and are politically stable.

figure 51: A map showing the location of the world's 23 largest telescopes (with apertures ranging from 4m to 10.4m), of which 9 are in Chile, 4 are in Hawaii and 2 are in the Canary Islands. Below that are photographs of, from top to bottom: Paranal Observatory in the Atacama Desert, Chile; Roque de los Muchachos Observatory on La Palma in the Canary Islands; Mauna Kea Observatory on Hawaii; Concordia Station at Dome C in Antarctica.


Street View
Traffic
Map
Satellite
Terrain
Greenland
Iceland
Canada
United States
México
North Atlantic Ocean
North Pacific Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
Suomi Finland
Sverige Sweden
Norge Norway
Россия Russia
United Kingdom
Polska Poland
Deutschland Germany
France
España Spain
Italia Italy
Україна Ukraine
Kazakhstan
Türkiye Turkey
Iraq
Iran
Afghanistan
Pakistan
China
日本 Japan
India
Algeria
Libya
Egypt
Saudi Arabia
Mauritania
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
DR Congo
Kenya
Tanzania
Angola
Namibia
Madagascar
South Africa
Venezuela
Colombia
Perú
Brasil Brazil
Argentina
Thailand
Indonesia
Papua New Guinea
Australia
New Zealand
Antarctica
2000 mi
2000 km
©2007 Google - Imagery ©2007 TerraMetrics, NASA, Map data ©2007 Europa Technologies - Terms of Use







astronomy from space and antarctica

The factors which degrade astronomical images, such as seeing, sky background, transparency variations and extinction are all atmospheric-induced phenomena. These effects can all be removed at a stroke by siting optical telescopes in space. The best known example of this is the 2.4 m Hubble Space Telescope (HST, see figure 23), which has helped to revolutionise astronomy since 1990 with its diffraction-limited imaging capability. The main drawback with siting telescopes in space is the cost: the HST cost many billions of dollars to build and operate, approximately ten times the sum required for the largest ground-based telescopes. The primary reason for this cost difference is accessibility - rockets rather than lorries have to be used to transport the telescope to the site. Another drawback with space is the risk involved in the launch, and the great difficulty of fixing problems, servicing the telescope and upgrading the instrumentation once the telescope has been deployed. NASA solved this to some extent by ensuring the HST possesses extremely

reliable/redundant systems and by using Space Shuttle servicing missions, but these were both costly solutions. Moreoever, the servicing missions, which cost nearly $1 billion each, were of limited number and were very risky for the astronauts involved.

Figure 51 also shows Dome C in Antarctica. Although not currently host to a major telescope, it is probably only a matter of time before the unique characteristics of this site are exploited. Antarctica offers imaging performance part way between that of the next best ground-based sites and space. Yet the difficulties of operating telescopes in Antarctican conditions, and of getting equipment and people there, means that the cost of building and operating a large telescope in Antarctica is also part way between that of a spaced-based mission and a ground-based telescope at a more temperate latitude. The same argument applies to the Arctic, which has large mountains bordering the Arctic Ocean that are very promising sites for telescopes.

---

dome design

Telescopes are packed with sensitive and expensive optical and electronic components. The purpose of a dome is to protect them from the environment, i.e. sunlight, humidity, rain, snow, dust and wind.

As well as protecting the telescope, a dome should also:

- Allow the telescope to access the entire sky. Hence they must have some form of opening and the ability to track with the telescope.
- Eliminate the temperature gradient between the interior and exterior, as the mixing of hot and cold air around the opening in the dome will produce turbulence, causing *dome seeing*.
- Provide space for amenities, such as cranes, control rooms, workshops, instrument storage and aluminising equipment.

There are many different varieties of domes, so this discussion shall be restricted to the three main types, which shall be referred to as: the *clamshell* design, the *classical* design, and the *box* (or *can*) design.

clamshell domes

The clamshell design is generally spherical in shape and uses two sets of interleaved shutters, as shown in figure 52. The great advantage of this design is that the entire sky is visible when the dome is open. This means that there is no need for dome rotation and tracking, and temperature differences between the inside and outside of the dome are almost instantaneously eliminated on opening. The disadvantages of this design are that it does not protect the telescope from wind-shake, which can blur astronomical images, or stray light, which can increase the background noise level. It is also impractical to house telescopes much larger than ~2 m in diameter in a clamshell dome as the shutters would become too heavy to maneouvre into position.

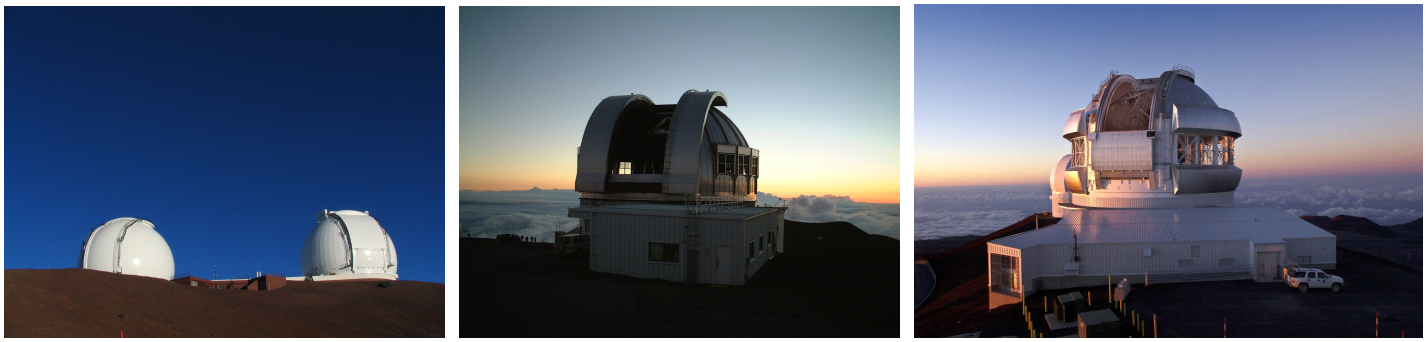
figure 52: Clamshell domes. Top left: the 2 m Liverpool Telescope (LT) on La Palma, with the dome closed. Top right: the LT partially opened, exposing the telescope within. Bottom left: a mosaic showing how the two sets of shutters open on a smaller clamshell dome. Bottom right: the open clamshell dome of pt5m, the Durham-Sheffield 0.5m robotic telescope on La Palma.

classical domes

The classical dome design dominates the skyline at most astronomical observatories. Like the clamshell, the classical dome is spherical in shape, which ensures minimal wind resistance and allows rain, snow and ice to run off, preventing it from collecting in large quantities on the outer surface. Unlike the clamshell, the classical design also allows a crane to be mounted on the inside of the roof, which is very useful for maintenance work (e.g. mirror re-aluminising) and instrument changes. A shutter is required in order to allow the telescope to view the sky: this can either be of the *wing* variety or the *over-the-top* variety - see figure 53. The latter is preferred due to its lower wind resistance when open, but this comes at an increased complexity and cost.

The great advantage of the classical dome design over the clamshell is the protection it affords from wind shake and stray light, and the fact that domes up to ~10 m in diameter can be housed for an affordable price. There are, however, two main disadvantages. First, it is more difficult to equalise the internal and external temperatures when the dome is opened, resulting in poor dome seeing. This problem can be alleviated to some extent with the installation of air-conditioning to keep the interior of the dome at the same temperature as the exterior, and the use of large fans and ventilation slots to rapidly flush the dome with exterior air when opening. Second, the entire dome must be able to rotate and track the motion of the telescope, significantly increasing the cost and complexity of the structure.

figure 53: Classical domes. Left: the twin 10 m Keck Telescopes on Hawaii, with closed over-the-top shutters. Centre: the 3.8 m UK Infrared Telescope (UKIRT) on Hawaii, showing open wing shutters and ventilation windows. Right: the 8.1 m Gemini Telescope in Hawaii, with open ventilation slots and over-the-top shutter. Notice the lower shutter, which can also be used as a wind shield.

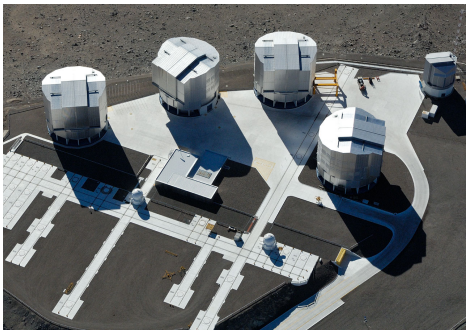
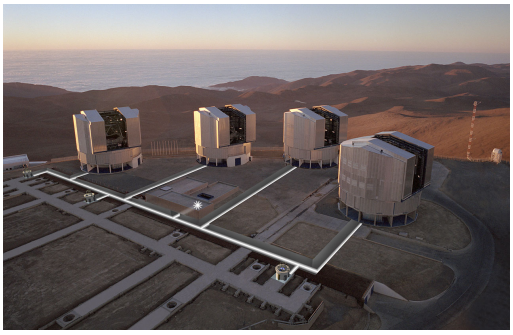


box domes

The curved components of the classical dome design are both complex and expensive to produce, which means that many (but not all) of the latest telescopes, including the next generation of extremely large (>10 m) telescopes, adopt a more box-like (or can-like) design. The use of straight sections of steel makes the design simpler and cheaper to manufacture. The disadvantage of this design, however, is its increased wind resistance

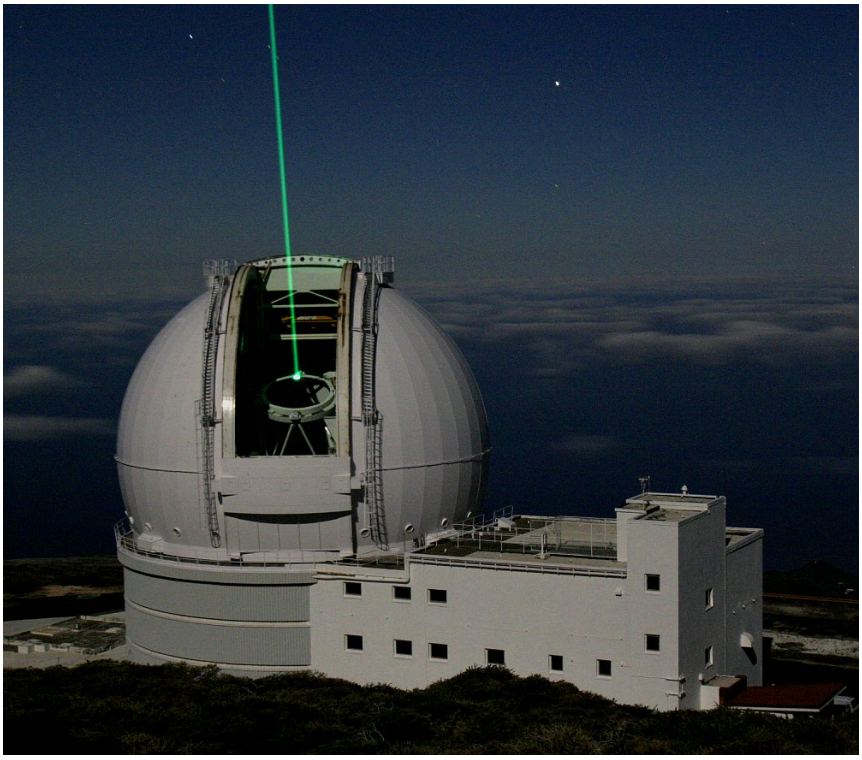
and the potential for snow and ice to collect on the flat roof. Hybrid systems intermediate between the box and classical designs are also in existence - see figure 54.

figure 54: Box domes. Left: the four 8.2 m Very Large Telescopes (VLT) in Chile, with the domes closed. Centre: the VLT with the domes opened - note the innovative wing shutter that does not protrude when opened. Right: the twin 6.5 m Magellan Telescopes in Chile, a hybrid design which uses straight girders/panels but maintains the overall classical dome shape.


---

telescopes



V. corrections

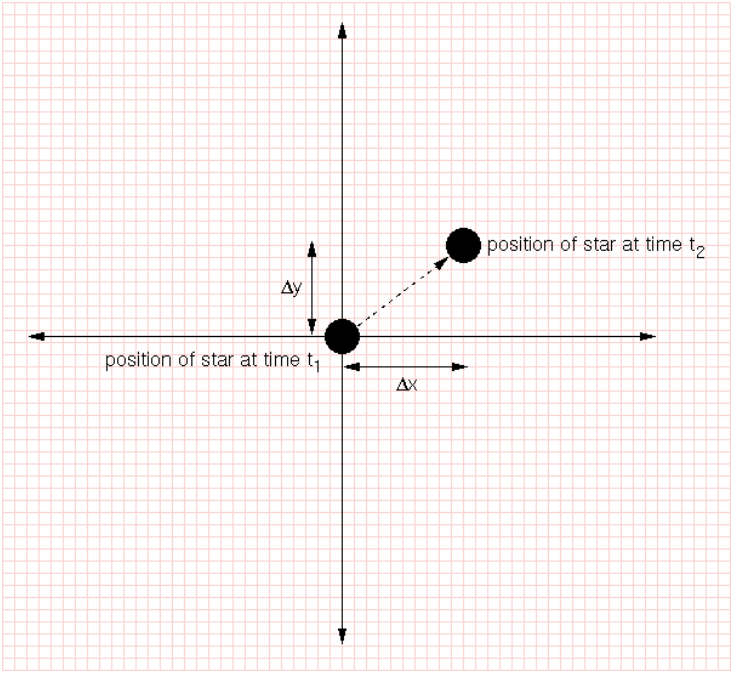
i. autoguiding
ii. active optics
iii. adaptive optics

autoguiding

We have already seen that equatorial telescopes can track the motion of an astronomical target across the sky by rotating the polar axis at the sidereal rate. In practice, it isn't possible to keep an astronomical object perfectly in the centre of the field of view by doing this. This is due to a number of factors, including imperfections in the telescope drive systems, flexure, and misalignment of the mount axes with respect to the celestial pole. The same problem also affects alt-az telescopes, of course, but in this case the misalignment is with respect to the horizon and zenith.

To combat this image motion, which would cause smearing in long exposures, it is necessary to *guide*, where small adjustments to the position of the telescope are made to keep the object at the desired position in the focal plane. Although this can be done manually using the telescope handset, it is usually done automatically using an *autoguider*, which continuously (typically on 0.1-1 s timescales) measures the position of a guide star somewhere in the field of view and nudges the telescope to keep the guide star locked onto a particular pixel on the autoguider's detector (see figure 55). To prevent over correction, i.e. "chasing" the guide star in response to fluctuations in its position due to seeing, and also to allow for the fact that the guide star will most probably have moved slightly in the time it takes to make the position measurement and correct for it, it is best to move the telescope by only a fraction of the measured offset. This parameter, called the *aggressiveness* or *gain* of the autoguider correction, typically has a value of ~0.3.

figure 55: The principle of autoguiding. Autoguiding begins at time t_1. At time t_2, the star has drifted by Δx, Δy pixels on the autoguider CCD. With a knowledge of the plate scale and orientation of the field of view, these pixel shifts are converted to arcseconds of motion of the two telescope axes. The shifts are multiplied by the aggressiveness (a factor between 0 and 1), and then applied to the telescope in order to shift the star back towards its initial position.

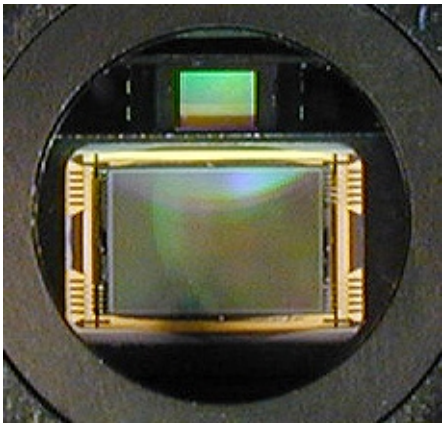


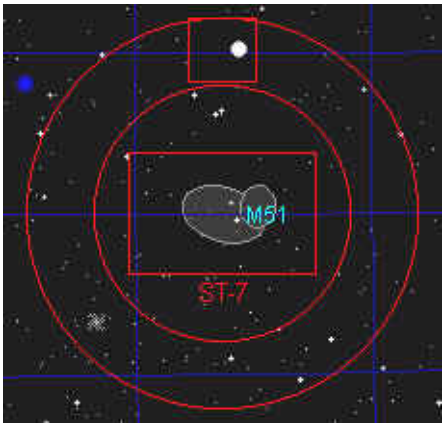
There are two main types of autoguider - *off-axis autoguiders* and autoguiders on separate *guidescopes* - each of which are described below.

off-axis autoguiders

Off-axis autoguiders use stars located in the periphery of the field of view, outside the area of scientific interest. On small telescopes, it is common to incorporate an off-axis autoguider CCD alongside the main science CCD, and house both in the same camera body. An example of such an arrangement is shown in figure 56. The advantage of this setup is that there are no moving parts and there is no flexure between the autoguider CCD and the science CCD, which means that any corrections measured from a star in the autoguider CCD will be applicable to the science target in the main CCD. The disadvantage of off-axis autoguiders of this type is the lack of flexibility in choosing a guide star - if no suitable star falls on the small autoguider CCD, then the whole camera body has to be rotated and/or the telescope has to be moved until one is found. Such autoguiders also have to peer through the same filter that is being used for the scientific observations, significantly reducing the amount of light from the guide star in the short exposure times and hence reducing the guiding accuracy.

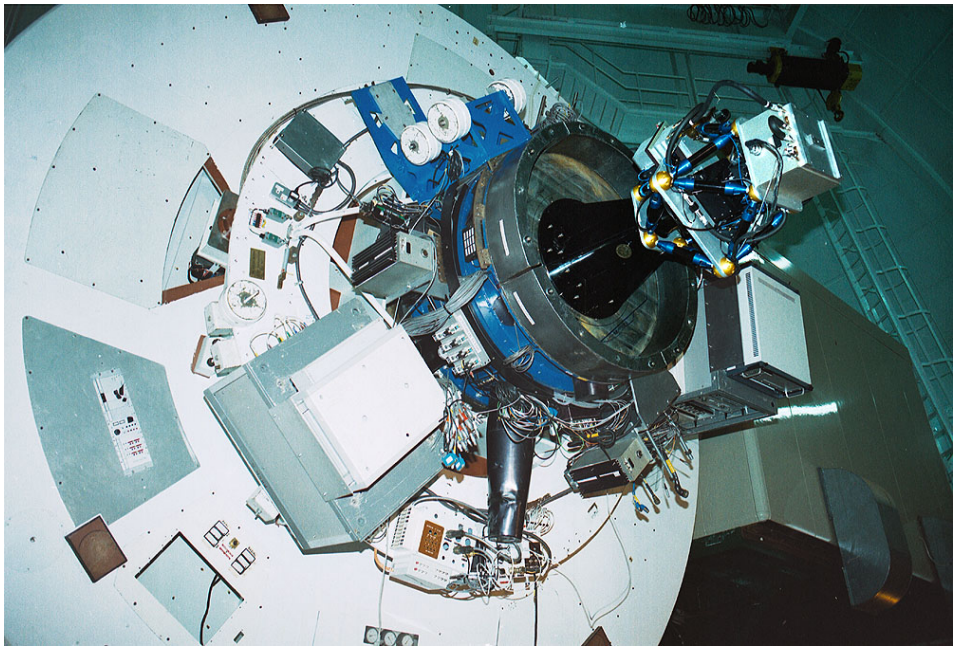
figure 56: Left: Photograph of the 657 x 495 pixel TC-237 off-axis autoguiding CCD sitting above the main KAF-1602E science CCD in an SBIG ST-8 camera. Right: Outline of the field of view of an SBIG ST-7 camera on an 8" Schmidt-Cassegrain telescope, showing the galaxy M51 on the science CCD and a bright guide star on the integral off-axis autoguiding CCD.





Large, professional telescopes tend to use a different form of off-axis autoguider, where a separate autoguider CCD is mounted on an adjustable stage which is able to select any guide star in an annulus around the central field of view (see figure 57). The great advantage of this setup is that it is almost always possible to find a suitable guide star by moving the autoguider CCD, and this process is often automated. Another advantage is that the autoguider CCD is a completely separate system, and hence can be used to autoguide any instrument which accesses the centre of the field of view. The disadvantage of this autoguider design is the complexity and cost of the system, and the fact that the autoguider CCD may experience different flexure to the science CCD, resulting in inappropriate guiding corrections being applied.

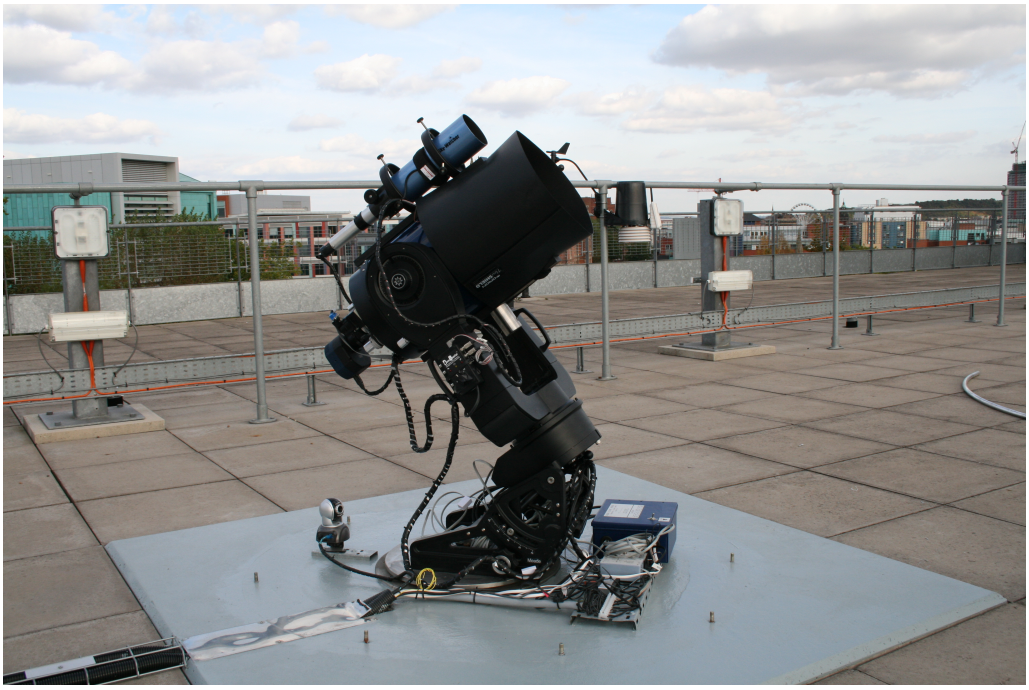
figure 57: Off-axis autoguiding using pick-off optics. Left: photograph of the Cassegrain focus of the 4.2 m William Herschel Telescope on La Palma. Autoguiding is provided by the blue-coloured Acquisition & Guidance (A&G) unit, which uses an internal pick-off mirror that can travel radially and azimuthally in an annulus around the centre of the field of view. Right: photograph of a Hutech Mitsuboshi off-axis guider for the amateur market. The pick-off prism directs off-axis light upwards for autoguiding. The prism can be manually adjusted radially and azimuthally to select different guide stars.



guidescopes

A guide scope is a separate, smaller-aperture telescope equipped with a CCD autoguider, which is mounted on the side of the main telescope, as shown in figure 58. The advantage of this setup, which is usually only found on amateur telescopes, is that the guide scope generally has a short focal length and hence wide field of view, resulting in a high probability of finding a bright guide star somewhere on the autoguider CCD. No moving parts are required and, assuming it is a bright point source, even the science target itself can be selected as the guide star. The disadvantage of a guide scope is that it can experience significantly different flexure to the main telescope as it tracks an object across the sky, resulting in a gradual drift of the field on the science CCD. Another disadvantage is that the short focal length of the guide scope provides a large plate scale, which means that small movements in the position of the guide star may be difficult to detect.

figure 58: A photograph of ROSA, the 10" robotic telescope on the roof of the Department of Physics and Astronomy at Sheffield, with the dome removed. The piggy-back mounted guide scope, a 3-inch refractor, can be seen on top of the main telescope. A CCD autoguider has been inserted in place of the eyepiece.

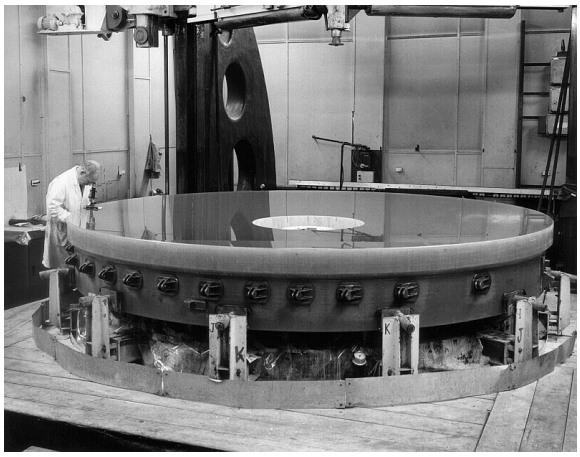


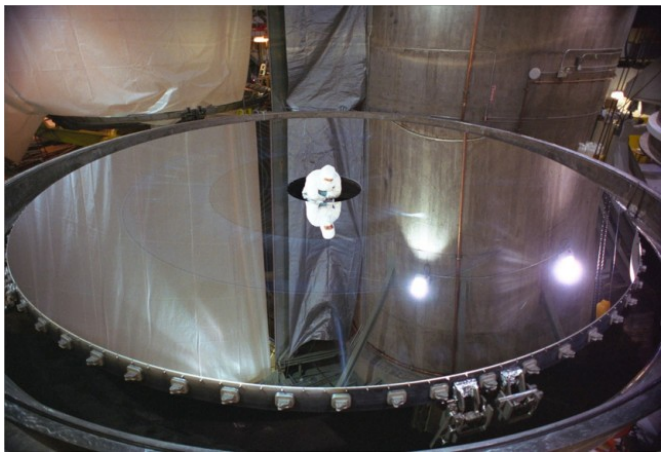
---

active optics

Prior to the 1990s, the primary mirrors of the world's largest telescopes had thicknesses of approximately 15% of their diameters. This ensured that the mirror was rigid and kept its shape regardless of the orientation on the sky. Unfortunately, this resulted in extremely heavy mirrors - the 4.2 m mirror of the William Herschel Telescope (WHT) on La Palma, for example, has a thickness of 56 cm and a weight of 16.5 tonnes (see figure 59). To make the next generation of 8 m class telescopes, thinner mirrors had to be used, as otherwise they, and their mounts, would have become impossibly heavy to move accurately around the sky. So, for example, if the 8.2 m primary mirror of the Very Large Telescope (VLT) in Chile had been constructed as a traditional thick mirror, it would have had a thickness of over 1 m and weighed approximately 100 tonnes. In contrast, the thin mirror that was actually built for the VLT has a thickness of only 0.18 m (i.e. 2% of the diameter) and weighs only 24 tonnes (see figure 59).

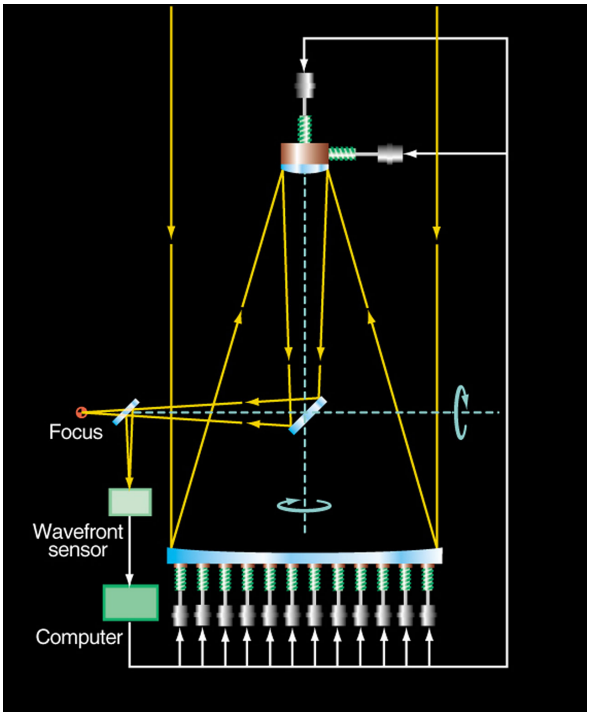
figure 59: Thick and thin mirrors. Left: photograph of the 4.2 m primary mirror of the WHT, which has a thickness/diameter ratio of 13%. Right: photograph of the 8.1 m primary mirror of the Gemini North telescope, which has a thickness/diameter ratio of only 2%.

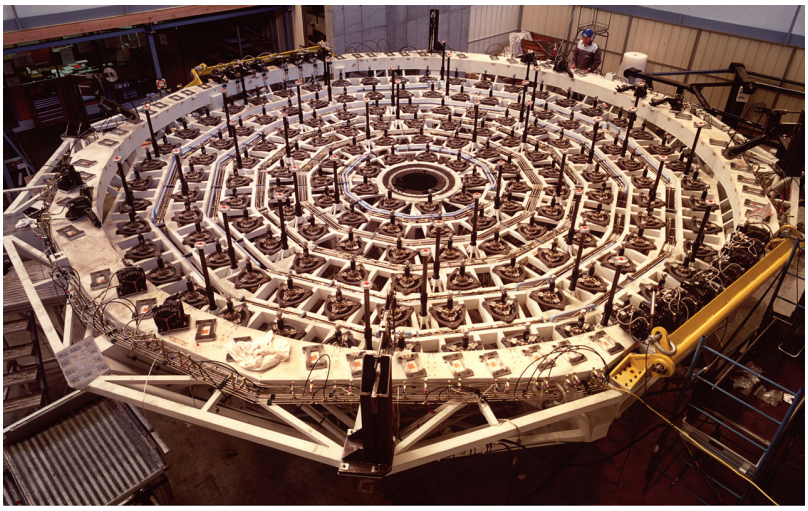




The trouble with using thinner mirrors is that they are distorted from their ideal shape by gravity as they are tilted to view different regions of the sky. To combat this, a technique known as *active optics* was developed, where the thin (and hence flexible) mirror is mounted on a set of actuators, which are devices that transform an electrical signal into linear motion. The system works as follows (see figure 60): whilst the telescope is observing a science target, the light from a bright reference star somewhere in the field of view around the science target is picked off and the quality of its image is measured. This measurement is usually performed by a *wavefront sensor*, which we shall discuss in more detail later. A computer-controlled set of actuators is then used to push and pull different parts of the primary mirror so that its shape changes in such a way as to improve the quality of the reference star image. Only about one measurement and correction cycle per minute is required, as the orientation of the telescope whilst tracking a target does not change significantly on timescales shorter than this. Note that, as well as correcting the shape of the primary mirror, many active optics systems also correct for the change in the position of the secondary mirror (see figure 60). This corrects for flexure and temperature-dependent changes in the length of the Serrurier truss, further improving the image quality.

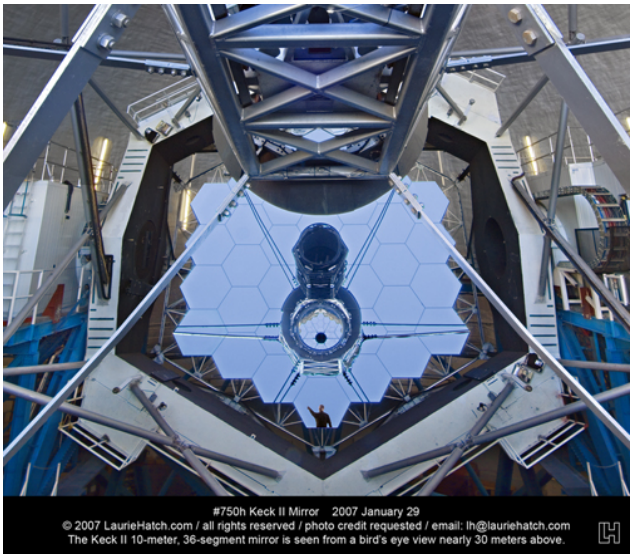
figure 60: Left: schematic showing the components of a typical active optics system. Note the actuators (shown as green springs) controlling the primary mirror shape and the vertical and horizontal position of the secondary mirror. Right: photograph of the VLT primary mirror support, showing 150 actuators arranged in six concentric rings.

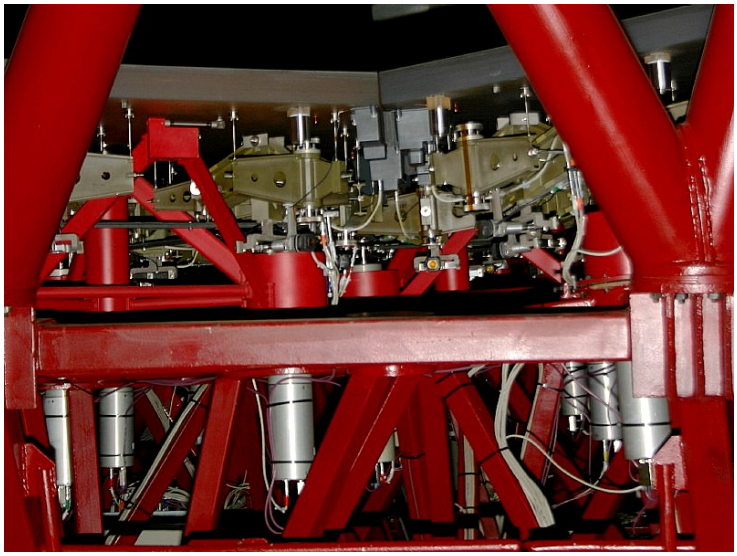




An alternative, cheaper solution to single (so-called monolothic) mirrors is to use segmented mirrors. This approach has been adopted, for example, on the twin 10 m Keck telescopes (figure 61). The segments are typically hexagonal in shape and usually have an asymmetric profile so that when all of the segments are combined they form, for example, a hyperbolic shape. Active optics systems are essential for segmented-mirror telescopes, as the position of each segment needs to be carefully controlled so that the overall shape of the primary mirror is retained as it is tilted to different sky positions. Each segment is therefore mounted on its own set of actuators, as shown in (figure 61).

figure 61: Left: photograph showing the segmented primary mirror of one of the twin 10 m Keck telescopes on Hawaii, composed of 36 hexagonal mirror segments, each 1.8 m wide, 8 cm thick and weighing only 400 kg. Note the reflection of the person for scale. Right: photograph showing the actuators beneath the hexagonal mirror segments of the 10.4 m Gran Telescopio Canarias on La Palma.





Segmented mirrors are the only feasible way of constructing telescopes with apertures significantly in excess of 8 m, as monolithic mirrors would become extremely expensive and ultimately impossible to manufacture, transport, install and maintain. The disadvantage of segmented mirrors is that they often require asymmetric profiles, making them difficult to manufacture. The active optics systems required to support them is also complex, and the gaps between the segments (typically a few mm) can cause low-level diffraction effects and increased infrared background in the final image.

---

adaptive optics

Active optics only counteracts the effect of gravity and other deformations on the telescope mirrors, and operates on timescales of tens of seconds. Correction for the effects of the Earth's atmosphere, on the other hand, is the realm of *adaptive optics*, which operates on timescales of hundredths of a second. Before we describe the basic principles of adaptive optics, it is necessary to define three important terms that characterize the effects of turbulence in the Earth's atmosphere on a wavefront propagating through it: the Fried (pronounced *free-d*) parameter, the coherence time and the isoplanatic angle.

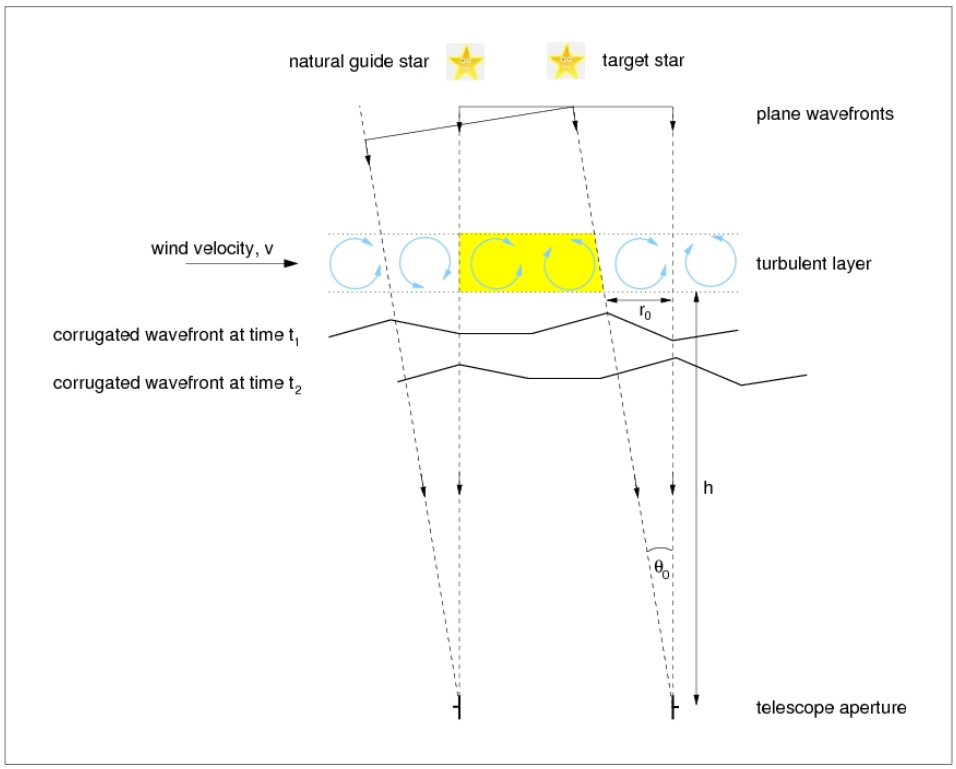
fried parameter

As shown earlier in figure 42, a distant point source like a star will produce a plane wavefront above the Earth's atmosphere. As this plane wavefront propagates through the atmosphere, it will be randomly distorted by moving cells of air with different indices of refraction. The size of the turbulent cells is characterized by the *Fried parameter*, r_0. The Fried parameter indicates the length over which a wavefront remains unperturbed, i.e. approximately planar, as shown in figure 62. (Strictly speaking, r_0 refers to the length of the wavefront over which the phase changes by 1 radian). The larger the Fried parameter, the better the atmospheric conditions. At a good observing site, the Fried parameter typically has a value of r_0 = 10 cm at an optical wavelength of λ = 500 nm.

Assuming a model for the atmosphere known as *Kolmogorov turbulence*, the Fried parameter is theoretically predicted to vary with wavelength as: $r_0 \alpha \lambda^{6/5}$. Hence one would predict a value of r_0 ~ 70 cm at a near-infrared wavelength of λ = 2.5 μm. This expectation is borne out by observations. As we shall see below, the Fried parameter determines the sizes of the individual lenslets and mirror segments in the wavefront

sensor and deformable mirror of an adaptive optics system, implying that adaptive optics correction in the infrared requires fewer elements (and is hence easier) than in the optical.

figure 62: Schematic showing how the Fried parameter, r_0, coherence time, $t_0 = t_2 - t_1$, and isoplanatic angle, θ_0, are defined.



We have already seen that the size of a diffraction-limited image is proportional to λ/D. We have also seen that large telescopes (defined here as those with $D >> r_0$) are not limited by diffraction, but by the seeing. In fact, the size of a seeing-limited image is proportional to λ/r_0. Hence the seeing in a large-aperture telescope is proportional to $\lambda^{-1/5}$. So, for example, if the seeing when observing at 500 nm is 1", it would be ~0.7" when observing in the near-infrared at 2.5 μm.

coherence time

The turbulent cells responsible for distorting the plane wavefronts from a star generally evolve on longer timescales than the time it takes the wind to move a cell by its own size. Hence it is the wind velocity, v, at the altitude of the turbulence that determines the temporal variation of the

wavefronts entering the telescope, as shown in figure 62. A turbulent cell would move its own size in a time $t_0 = t_2 - t_1$, given by

$t_0 = r_0 / v$.

At a good observing site on a typical night, v = 10 m/s and r_0 = 10 cm at λ_0 = 500 nm. Hence t_0 = 10 ms. Detailed arguments lead to a more accurate version of this equation: $t_0 \sim 0.314\ r_0 / v$. t_0 is known as the *coherence time* and indicates the timescale on which the wavefront sensor and deformable mirror of an adaptive optics system must operate. The dependence of t_0 on r_0 implies that adaptive optics correction in the infrared can be much slower (and hence is easier) than in the optical.

isoplanatic angle

Another parameter of importance in adaptive optics is the *isoplanatic angle*, θ_0. Suppose that there are two stars close together on the sky. What angle would these two stars have to be separated by in order for them to pass through approximately the same turbulent region of the atmosphere? Figure 62 shows that this can be estimated from the angle over which the turbulence pattern is shifted by a distance of only r_0, in which case the beams from the two stars would share a substantial fraction of the turbulent region (shaded in yellow). Assuming that the turbulent layer is at an altitude h above the telescope, the isoplanatic angle is hence given by

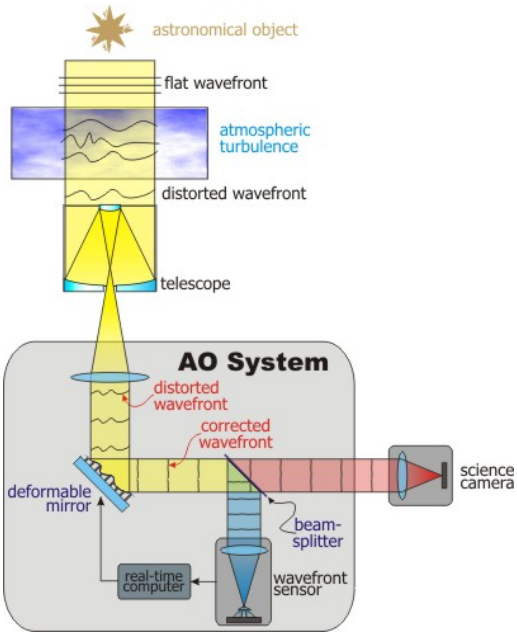
$\theta_0 = r_0 / h$.

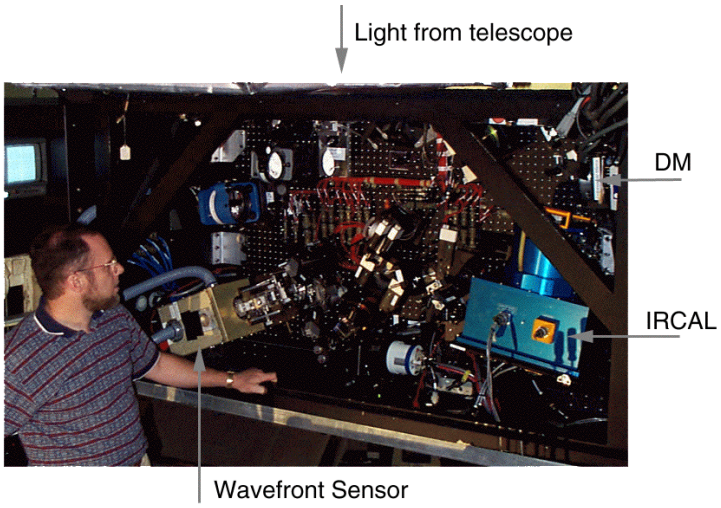
At a good observing site on a typical night, h = 10 km and r_0 = 10 cm at λ_0 = 500 nm. Hence $\theta_0 = 10^{-5}$ radians, which is ~2". Detailed arguments lead to a more accurate version of this equation: $\theta_0 \sim 0.314\ r_0 / h$. The isoplanatic angle determines the area on the sky over which adaptive optics correction is effective. The dependence of θ_0 on r_0 implies that much wider fields (and hence more extended objects) can be corrected with adaptive optics in the infrared than in the optical, making the technique much more attractive in the infrared. The increased isoplanatic angle in the infrared also means that many more natural guide stars are available, as discussed below.

basic principles of adaptive optics

The principles of adaptive optics correction are shown in figure 63. A plane wavefront from a star is corrugated by turbulence in the Earth's atmosphere. The diverging beam beyond the focal plane of the telescope is then made parallel using a collimator, and the collimated beam reflects off a *deformable mirror*, which is adjusted in shape to match that of the wavefront. As a result, the reflected wavefront becomes planar again, and the corrected beam is then focused and detected by a science camera. The shape of the deformable mirror is adjusted hundreds of times a second, using information provided by a *wavefront sensor*, which picks off the (unwanted) blue light in the beam using a *dichroic beamsplitter*. Note that the *shape* of the wavefront is independent of wavelength, which is why it is then possible to sense and correct at different wavelengths; you can think of an infrared wavefront as being a squashed (in phase) version of an optical wavefront.

figure 63: Left: Schematic showing the principal components of an adaptive optics system. Right: Photograph showing the main components of the adaptive optics system on the 3m Shane Telescope at Lick Observatory, California: the deformable mirror (DM), the science camera (IRCAL) and the wavefront sensor.

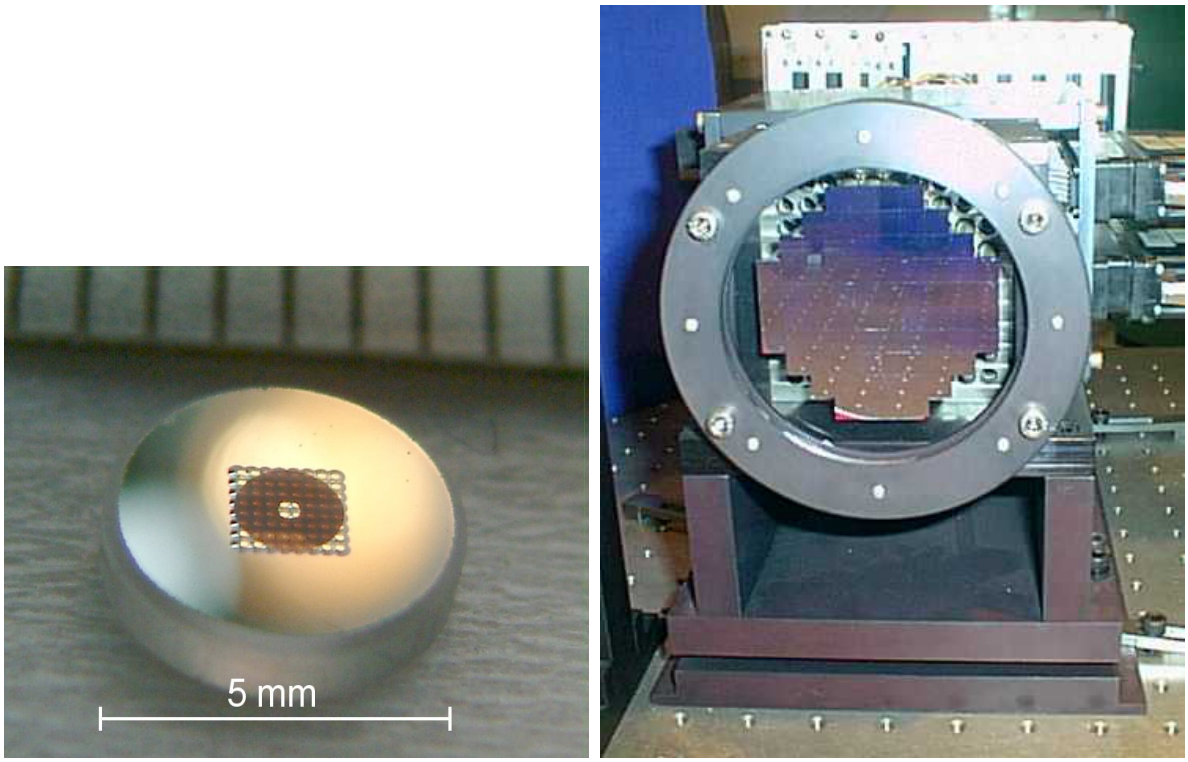




wavefront sensors and deformable mirrors

The two most critical elements of an adaptive optics system are the deformable mirror and the wavefront sensor. There are many different types of each, so we shall concentrate here on the conceptually simplest: the *Shack-Hartmann wavefront sensor* and the *segmented deformable mirror* - see figure 64.

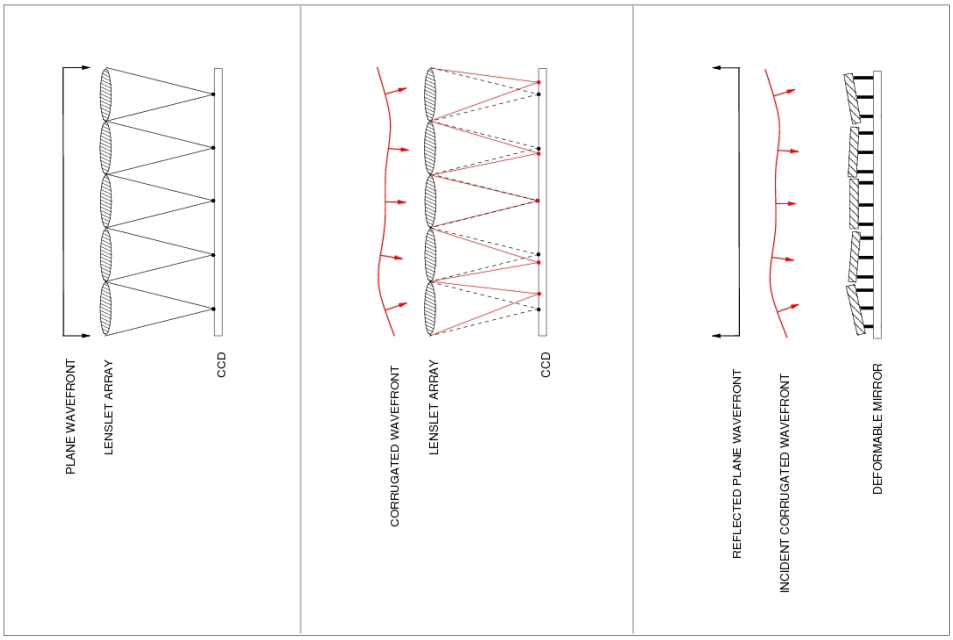
figure 64: Photograph of a typical lenslet array for use in a Shack-Hartmann wavefront sensor. Photograph of a 76-element segmented deformable mirror, used in the NAOMI adaptive optics system on the 4.2m William Herschel Telescope (WHT) on La Palma.

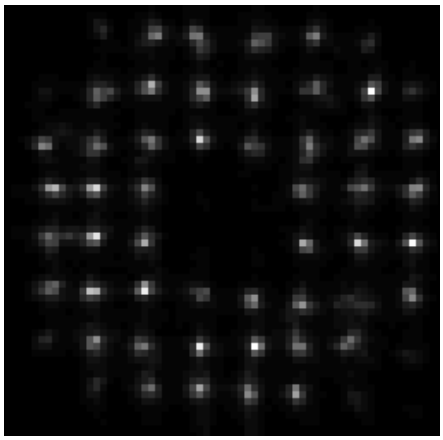


The Shack-Hartmann wavefront sensor consists of a lenslet array which the corrugated wavefront is incident upon. A plane wavefront incident on the lenslet array would produce a regular series of spots on a high-speed detector in the focal plane. A corrugated wavefront, on the other hand, would produce irregularly spaced spots, where the tilt of each section of the wavefront can be determined by measuring the displacement of the spot from the fiducial position (defined by illuminating the lenslet array

with a plane wavefront). This is shown in the left-hand and central panels of figure 65, together with a movie of data from a real Shack-Hartmann wavefront sensor.

figure 65: Left: Schematic showing the principle of Shack-Hartmann wavefront sensing. Right: Movie of images obtained by the JOSE Shack-Hartmann wavefront sensor on the WHT when illuminated by a bright star (reload this page to restart the animation). The blank region at the centre is the shadow of the secondary mirror.





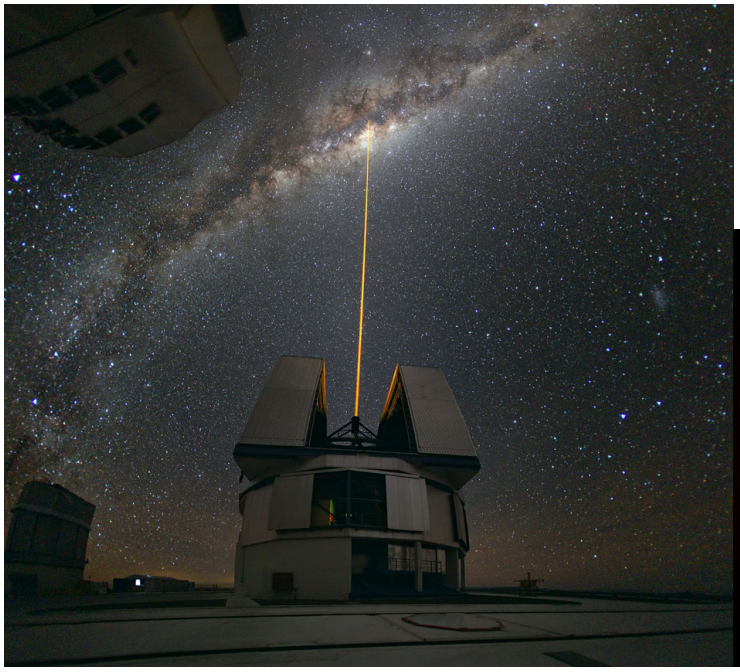
The tilt of the wavefront at each lenslet is then used to set the tilt of each corresponding element in the segmented deformable mirror (to a value equal to half of the tilt of the wavefront), so that the reflected wavefront becomes planar, as shown in the right-hand panel of figure 65. For accurate correction, it is essential that the delay between sensing the wavefront and adjusting the shape of the deformable mirror is no greater than the coherence time of the atmosphere, which is typically a few milliseconds in the optical part of the spectrum at a good astronomical site. Hence high-speed computer processors to measure the wavefront and move the mirror in a real-time correction loop are also an essential component of any adaptive optics system, as indicated in figure 63. It is also vital that the sizes of the lenslets, and hence the mirror segments, are well matched to the typical values of the Fried parameter at the observing site and wavelength of interest.

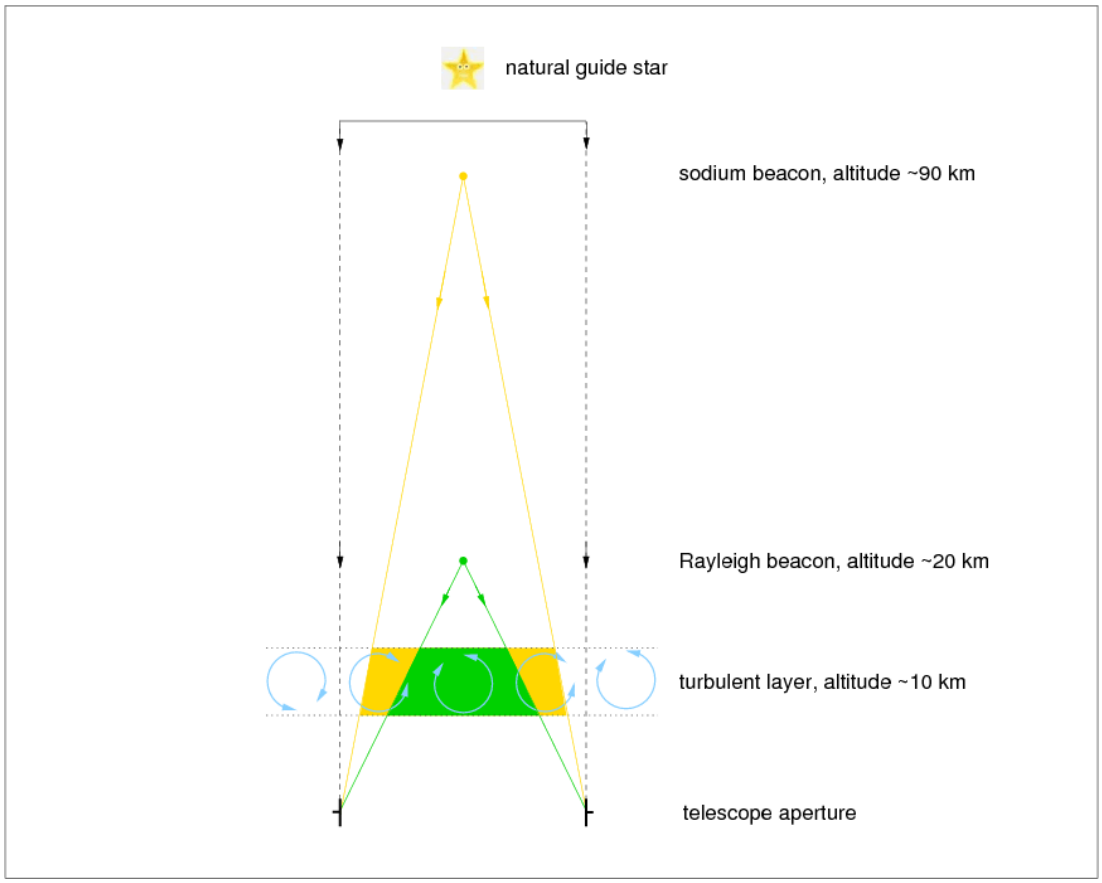
laser guide stars

In order for a Shack-Hartmann wavefront sensor to measure the tilts of a wavefront accurately, it is necessary to observe a bright point source to provide a sufficient signal-to-noise ratio in the short exposure times. It is possible that the science target itself can be used to sense the wavefront, e.g. if it has a bright, point-like central structure, such as an active galactic nucleus or young stellar object. Unfortunately, many astronomical targets are either faint, extended, or both. One way round this is to observe a bright star close to the target, but such a *natural guide star* would have to be within the isoplanatic angle of the target, otherwise the target and guide stars would be sampling different turbulence in the atmosphere (as shown in figure 62).

As calculated above, the isoplanatic angle in the optical part of the spectrum is only a few arcseconds, and this only increases to a few tens of arcseconds in the infrared. Hence only a very small fraction (or order 1-10 per cent) of the sky is actually correctable using adaptive optics with natural guide stars. The only way of significantly increasing the sky coverage is to generate an artificial guide star close to the target using a laser: a so-called *laser guide star* (figure 64).

figure 66: Top left: Photograph of a laser beam emanating from the 8.2 m Very Large Telescope in Chile. Top right: Photograph of the laser guide star produced by the ALFA adaptive optics system on the Calar Alto 3.5 m telescope in Spain. The sodium beacon is the point-like image at the centre; the plume to the right is the Rayleigh back-scattered light. Bottom: A schematic illustrating the cone effect.





There are two types of laser guide star. The first, known as a *Rayleigh beacon*, uses the Rayleigh back-scattering of light from molecules in the lower atmosphere to produce an artificial star at altitudes of approximately 20 km. The second type is the *sodium beacon*, which uses a laser tuned to the yellow sodium *D* lines around 589 nm. This excites sodium atoms (deposited by micrometeorites) in the mesosphere at an altitude of

approximately 90 km, which subsequently re-emit the light, producing an artificial star, as shown in figure 66.

Although much more more costly and complex than Rayleigh beacons, sodium beacons have one major advantage: they do not suffer as badly from the *cone effect*, resulting in superior adaptive optics correction. This is shown schematically in the bottom panel of figure 66, where it can be seen that the higher altitude of the sodium beacon means that it shares much more of the turbulence experienced by light from the star than the lower-altitude Rayleigh beacon. It is important to note, however, that a sodium laser focused to around 90 km altitude also produces an out-of-focus halo of light from Rayleigh back-scattering at lower altitudes. Fortunately, this halo can be ignored by using a pulsed laser and sensing the wavefront a few microseconds after the pulse has been launched; in this way, the scattered light from lower down in the atmosphere is rejected and only light which has travelled for several microseconds up to the sodium layer and back again is detected.

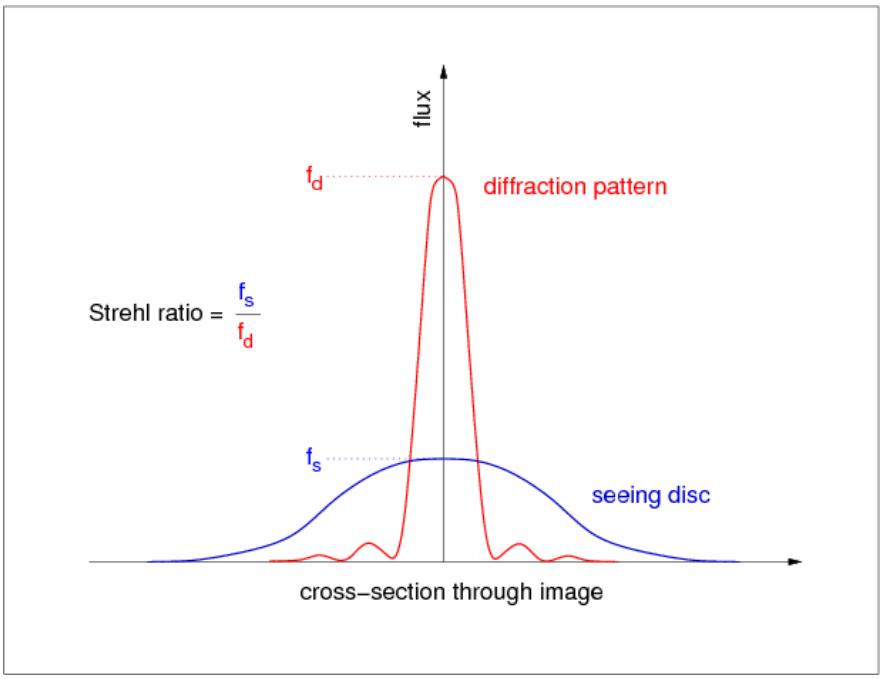
Since the laser passes twice through the turbulence, once on the way up and once on the way down, any overall tip and tilt of the wavefront, which manifests itself as overall image motion in the focal plane, is cancelled out. (The higher-order corrugations are not cancelled out as the laser is focused to a spot above the turbulent layer.) The stellar image, however, only passes once through the turbulence and hence will display this image motion which, if not corrected, will smear the image and thereby degrade the spatial resolution. Therefore, a natural guide star is still required for correction of this tip and tilt, but since only the image centroid needs to be measured, the guide star does not need to be anywhere near as bright, or lie as close to the target, as a natural guide star that is being used for higher-order corrections.

No matter which type of laser beacon is used, an artificial star is created that is (usually) above the typical altitudes at which turbulence occurs. The artificial star can be placed within the isoplanatic angle of the science target, resulting in sky coverage approaching 100% for adaptive optics (assuming a natural guide star is also available for the tip-tilt correction). Since the laser light is monochromatic, a simple notch filter can be used to direct all of the laser light to the wavefront sensor, leaving the rest of the light to be directed to the science detector.

adaptive optics in practice

We have already seen that the spatial resolution of a seeing-limited image is usually characterised by the full-width at half-maximum (FWHM) of a stellar profile, measured in arcseconds. This method becomes unreliable, however, as the spatial resolution approaches the diffraction limit, as measurement of the FWHM becomes complicated by the presence of diffraction rings. A more useful measure in this case is the *Strehl ratio*, which is the ratio of the intensity at the peak of the observed seeing disc divided by the intensity at the peak of the theoretical Airy disc, as shown in figure 67.

figure 67: The seeing disc of a star superposed on the theoretical diffraction pattern. The Strehl ratio is the ratio of the peak intensities of the two profiles.

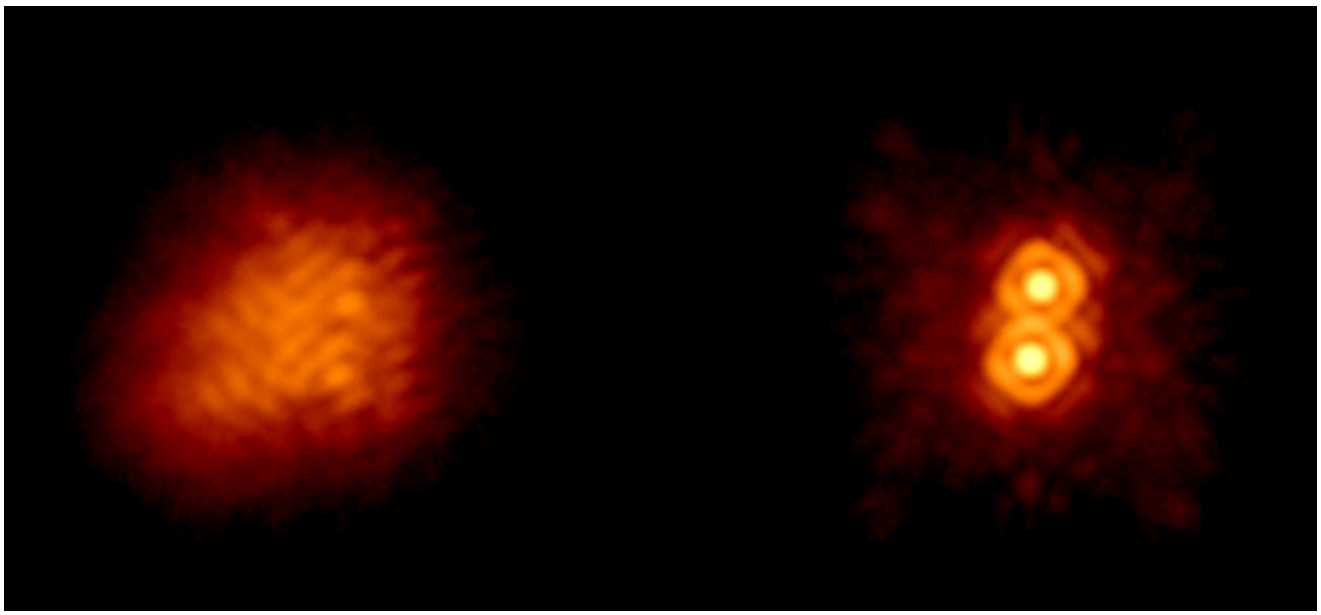


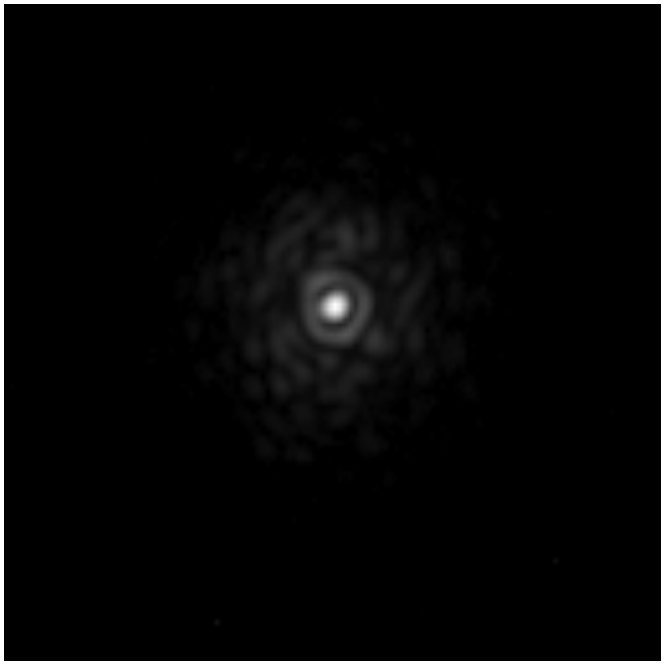
The Strehl ratio is the most commonly-used parameter to characterise the performance of an adaptive optics system. The Strehl ratio recorded by a telescope without adaptive optics is typically only a few per cent, but this can rise to over 50% if adaptive optics is used. The higher the Strehl ratio, the more the image is concentrated and hence the higher the spatial resolution. A higher Strehl ratio also means the image of a star is concentrated onto fewer detector pixels, minimizing the noise due to the sky background and the detector. In spectroscopy, a higher Strehl ratio implies that narrower slits can be used, which in turn means that the

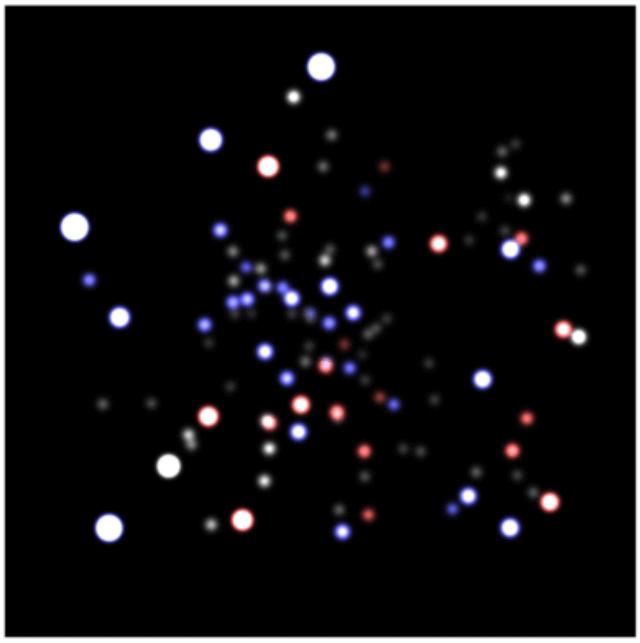
whole spectrograph can be made more compact.

Most large telescopes in the world are now equipped with adaptive optics systems. The most advanced such systems incorporate laser beacons and wavefront sensors/deformable mirrors with 1000+ elements, delivering diffraction-limited imaging in the infrared across most of the sky (see figure 68). However, diffraction-limited imaging on large-aperture telescopes is still not achievable in the optical. As discussed above, this is due to the lower values of the Fried parameter and coherence time, implying that an unfeasibly large number of adaptive elements and corrections per second would be required.

figure 68: Adaptive optics in action: Top: Images of the binary star IW Tau without (left) and with (right) adaptive optics on the 5.1 m Hale Telescope in California. The separation of the two stars is 0.3". Bottom left: A movie showing images of a star taken with AO correction turned off and then on. Bottom right: Arguably the most famous AO result to date - a movie of the orbits of stars around the Galactic centre, taken using the 8.2 m Very Large Telescope in Chile, which was used to infer the presence of a supermassive black hole. The image is only 3 arcseconds across.







---

telescopes

VI. interferometry

removed from course

instruments: introduction

The purpose of an astronomical instrument is to extract information from the photons collected by a telescope. In principle, all of the following photon properties are measurable:

1. the direction of the photon;
2. the time of arrival of the photon;
3. the energy of the photon;
4. the polarization of the photon;
5. the bunching properties of the stream of photons.

Imagers are used to study the spatial distribution of the photons from a source, i.e. they primarily measure property (i). *Photometers* are used to study the brightness of sources, and how they vary with time, i.e. they primarily measure property (ii). *Spectrographs* look at the wavelength distribution of light, i.e. they primarily measure property (iii). In this part of the course, we shall look at all three of these instruments. We shall not cover *polarimeters*, which exploit property (iv), as this is a relatively niche area of research. Instruments which exploit property (v), the realm of quantum optics, are even more niche and will not be discussed here either.

instruments

II. imagers

simple imagers

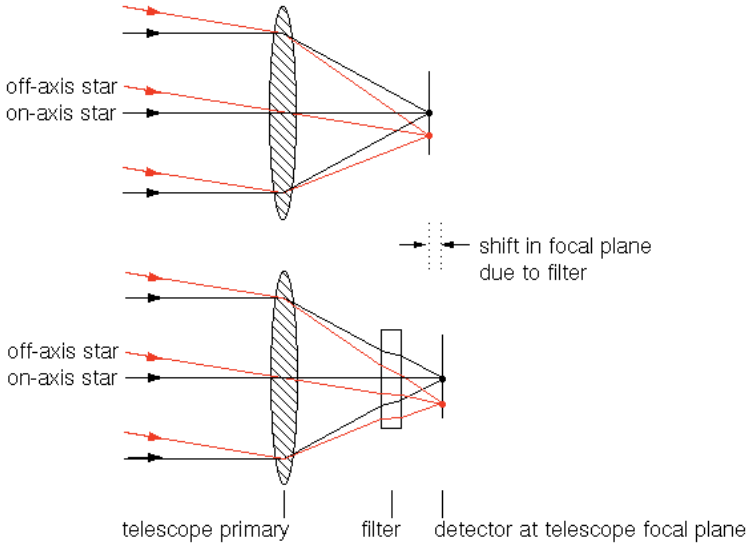
Imagers are astronomical instruments that have the capability of forming images of sources. The simplest form of imager is when a multi-pixel detector, such as a CCD, is placed at the focal plane of a telescope, in front of which is usually placed a filter in a filter wheel, as shown in figure 69.

figure 69: A photograph of the simple imager used by Sheffield astronomy undergraduates at the Cassegrain focus of the 0.5 m Mons telescope on Tenerife. From right-to-left are shown: the CCD detector, the filter wheel, the "flipper" (a diagonal mirror housed in the aluminium box which can be moved into the beam to direct the light upwards to the eyepiece) and the bottom of the primary mirror cell.

Since the filter in a simple imager is in the converging beam from the telescope, it causes the focal plane to move away from the primary, as shown in figure 70. To compensate for this when using a small telescope,

the detector is usually moved outwards. On large telescopes, however, which generally have multiple instruments that require focusing, the position of the detector and filter is usually fixed and the focal plane is moved back towards the primary by moving the secondary mirror away from the primary. Inserting filters of different thicknesses and/or refractive indices causes the focal plane to move by different amounts. This can be compensated for by adjusting the telescope focus each time, or eliminated entirely by ensuring that the filters all have the same *optical thickness*, i.e. the product of the thickness and refractive index of each filter in the wheel is the same.

figure 70: Left-top: A simple imager - parallel beams of light from an on- and off-axis star are brought to focus by a telescope onto an array detector placed at the focal plane. Left-bottom: the insertion of a filter forces the telescope focal plane outwards. Right: the QSI 532 CCD camera with a 5-position filter wheel, as used on Sheffield's 10" robotic telescope, ROSA.



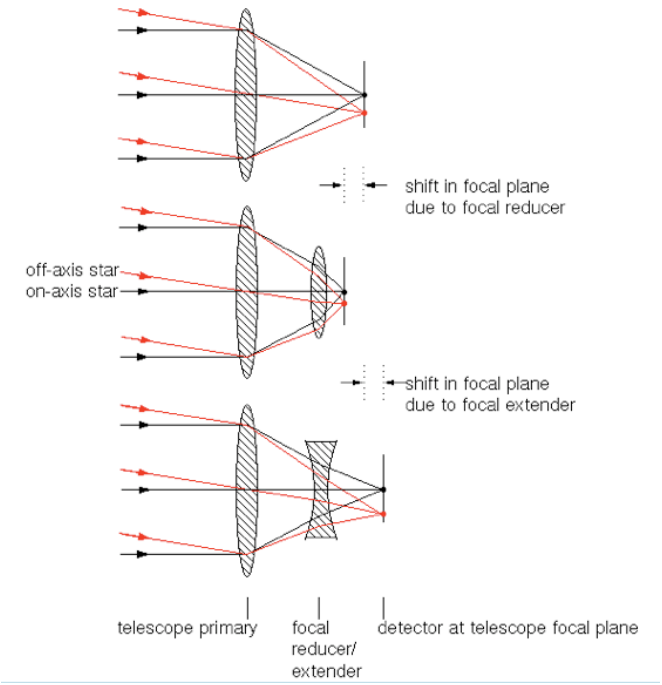
focal reducers and extenders

One problem with the simplest form of imager we have just seen is that the plate scale, i.e. the number of arcseconds per pixel on the detector, may be unsuitable for the size of the detector and its pixels, or vice versa. For example, a telescope with a focal length of 10 m will have a plate scale of approximately 20 "/mm. If used with a CCD of 1000 x 1000 pixels, each of 10 microns, the field of view would be only 3.3' x 3.3', and each pixel would correspond to only 0.2". If wide field imaging is required, such a setup would be of limited use. Moreover, if the seeing at the site is typically 2", the seeing disc would have a FWHM of 10 pixels, which is very oversampled (more on this later). Replacing the telescope for one of a shorter focal length, or the detector for one with more and/or larger pixels, is usually neither a practical nor economical solution, so what can be done?

One possibility is to use a *focal reducer*. This is a positive lens (or combination of lenses) which is usually placed in front of the focal plane, in the converging beam from the telescope, as shown in figure 71. The lens acts to shorten the focal length of the telescope, resulting in a larger plate scale and hence a larger field of view. Figure 71 also shows the opposite of a focal reducer - the *focal extender* or *Barlow lens* - which has negative optical power and acts to lengthen the focal length of the telescope, and hence decrease the plate scale and field of view.

Note that the focal length of a simple imager (such as shown in the top-left panel of figure 71) is simply the focal length of the telescope, but the focal length of a multi-element astronomical imager (like that shown in the centre- and bottom-left panels of figure 71) is a more complicated function of the focal lengths of the individual elements. In the limiting case of inserting a filter, which has no optical power (i.e. infinite focal length), the focal length remains unaltered, although the distance between the last lens and the focal plane changes (as shown in figure 70).

figure 71: Left-top: the light path of a simple imager. Left-middle: the light path of an imager with a focal reducer. Left-bottom: the light path of an imager with a focal extender. Note the shift in the position of the focal plane when using a focal reducer/extender, which can be compensated for by refocusing the telecope. This shift should not be confused with the change in the focal length induced by the focal reducer/extender. Right top: photograph of a 0.63x focal reducer. Right bottom: photograph of a 2.5x focal extender (or Barlow lens).



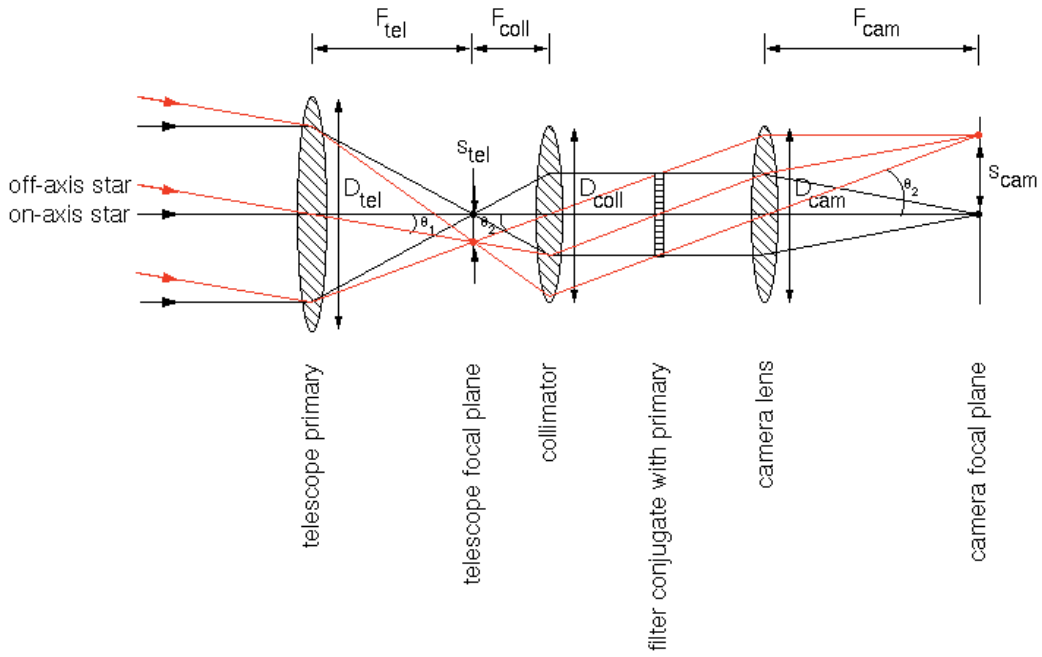
The top-right panel of figure 71 also shows a photograph of the 0.63x focal reducer used in the imager on Sheffield's *f*/10 robotic telescope, ROSA, which reduces the focal length of the telescope by a factor of approximately 0.63, i.e. the focal ratio changes from *f*/10 to *f*/6.3. (The precise amount of reduction depends on the characteristics of the telescope and the positioning of the focal reducer). Since the plate scale is inversely proportional to the focal length, this means that the plate scale (and hence the field of view) is increased by a factor of 1.6. Similarly, the bottom-right panel of figure 71 shows a photograph of a 2.5x Barlow lens, which decreases the field of view by this factor.

---

re-imagers

Focal reducers and extenders are commonly found on amateur telescopes. On professional telescopes, however, a more complex way of changing the plate scale and field of view is usually employed. This method, which shall be referred to as a *re-imager*, is still a form of focal reducer/extender, but uses a *collimator* placed after the focal plane of the telescope to make the diverging beam parallel. (To do this, the collimator must be placed at a distance equal to its focal length from the focal plane). The parallel beam then passes through a filter before it is focused onto the detector by a *camera* lens. A schematic of a re-imager is shown in figure 72.

figure 72: The light path of an on-axis and an off-axis star through a typical astronomical re-imager. The plate scale at the camera focal plane is different to the plate scale at the telescope focal plane due to the action of the collimator and camera lens. Note the position of the filter, which is placed conjugate with the telescope primary (see text for details).



The optical layout shown in figure 72 is similar to that of the telescope, eyepiece and eye system shown in figure 27. From Figure 72 it can be seen that the two stars are separated by a distance $s_{tel} = F_{tel} \theta_1 = F_{coll} \theta_2$ in the telescope focal plane, and by a distance $s_{cam} = F_{cam} \theta_2$ in the camera focal plane. We can therefore define a magnification, M, as follows:

$M = s_{cam} / s_{tel}$.

With this definition, if our re-imager has $M < 1$ we have a focal reducer (i.e. a demagnifier), and if $M > 1$ we have a focal extender (i.e. a magnifier). We can now derive a relation between M and the focal lengths of the collimator and camera lenses, as follows:

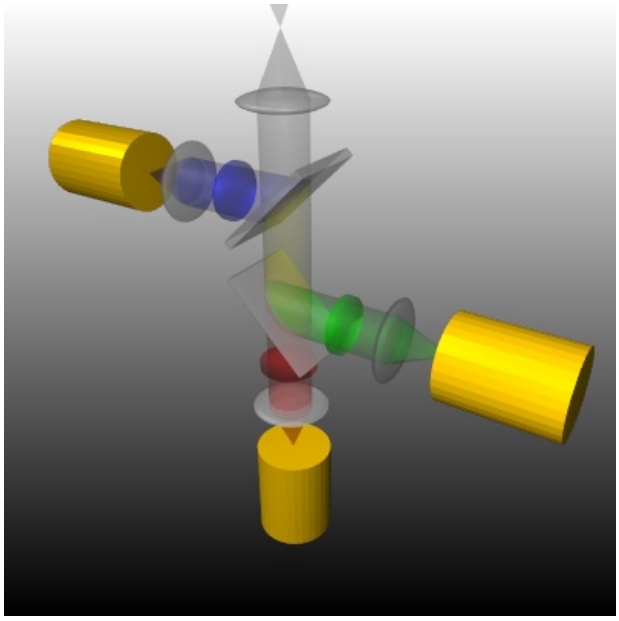
$M = F_{cam} \theta_2 / F_{coll} \theta_2 = F_{cam} / F_{coll}$.

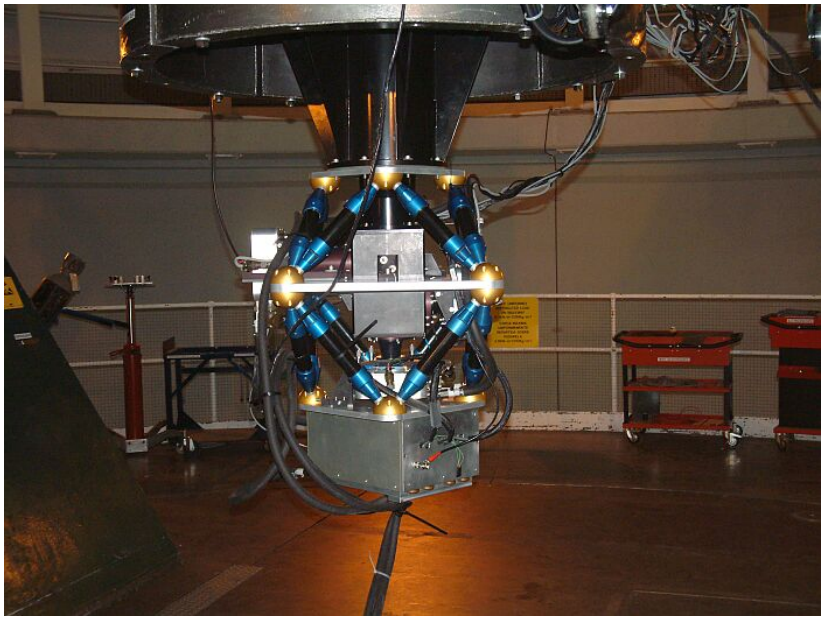
Hence, the magnification is given by the ratio of the camera to collimator focal lengths. This is an important equation, of use not only for re-imagers but also for spectrographs. A demonstration of when this equation is useful is given in the example problems.

Re-imagers have a number of advantages over the simple focal reducers/extenders shown in figure 71. First, since the filter is placed in a parallel beam, there is no change of focus if filters of different optical thickness are used, and no optical aberrations are introduced. Second, the filter can be positioned so that all of the rays incident on the primary mirror pass through the filter, i.e. the filter is said to be *conjugate* with the primary (and coincident with the so-called *pupil image*), as shown in figure 72. The advantage of this is that the minimum-sized (and hence cheapest) filter is then required to collect light from all angles. A circular aperture, or *stop*, is also often placed at this position in order to reject stray light from the system. Third, the collimated beam is optically easier to control than a diverging/converging beam, and hence it is more straightforward to provide space for additional optics, e.g. for adaptive optics or simultaneous multi-colour imaging (see figure 73).

figure 73: Left: A schematic ray-trace through ULTRACAM, the Sheffield/Warwick/UKATC-built three-colour imager. Light from the telescope focal plane at the top is collimated and then passes through two dichroic beamsplitters. The first reflects just the blue light, which then passes through a filter before

being re-imaged by a camera onto a CCD (the yellow cylinder). The second dichroic reflects just the green light, leaving the red light to fall onto the CCD at the bottom. Right: ULTRACAM mounted at the Cassegrain focus of the 4.2 m William Herschel Telescope on La Palma.





---

sampling theory

The issue of sampling was briefly mentioned in the context of the number of detector pixels in the seeing disc of a star when using a focal reducer. In this section we shall look at why sampling is an essential consideration when designing astronomical instruments, including both imagers and spectrographs.

Sampling is the process of converting a *continuous* signal, in this context an image or spectrum in the focal plane of an astronomical instrument, into a *discrete* signal, by selecting values at evenly-spaced points in the focal plane. This latter function is performed by the pixels in a detector. The question is, how finely spaced (i.e. how big) do the detector pixels have to be in order to faithfully recover the input signal? The solution is given by the *Nyquist sampling theorem*, which states that:

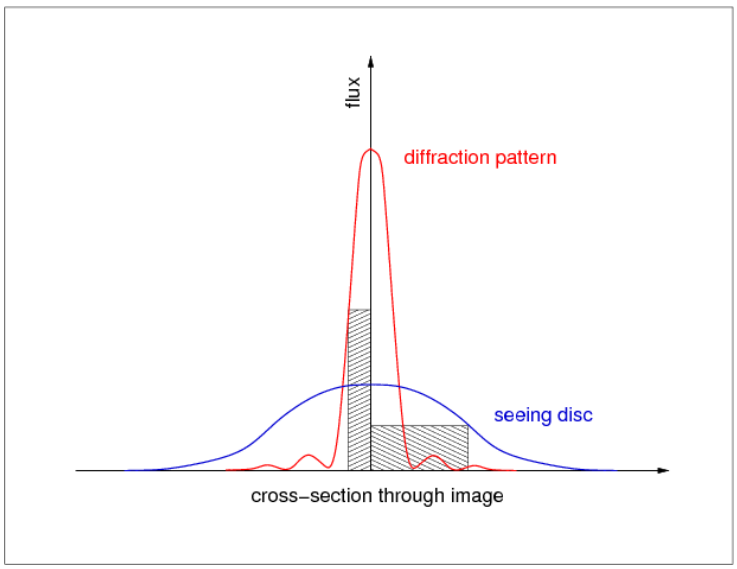
The sampling frequency should be greater than twice the highest frequency contained in the signal.

In other words, if the *smallest* feature in an image or spectrum has a dimension of x μm (i.e. the *highest* spatial frequency is $1/x$ cycles [or "features"] per μm), then the pixel size must be no larger than $x/2$ μm (i.e. the sampling frequency must be greater than $2/x$ cycles per μm) in order to detect all of the information in the image or spectrum.

The smallest real features present in an image will be those at the resolution limit, which is normally given by the seeing. If the seeing is 1", the Nyquist sampling theorem tells us that we would require 2 pixels across the seeing disc in order to sample the image optimally, i.e. the plate scale should be 0.5 "/pixel. This situation is shown schematically in figure 74. If we have more than 2 pixels across the seeing disc, the image is said to be *oversampled*, whereas if we have less than 2 pixels across the seeing disc, the image would be *undersampled*. A degree of oversampling is normally acceptable (for example to cope with the fact that the size of a pixel across the diagonal is greater than across a side), but too much oversampling results in a needlessly reduced field of view. Undersampling,

however, is rarely desirable, as it means that the image recorded by the detector is of a lower spatial resolution than the image delivered by the telescope, and undersampling can also cause problems during data reduction. The only occasions when undersampling might be desirable is if increasing the field of view or decreasing the contribution of detector (readout) noise is of paramount importance.

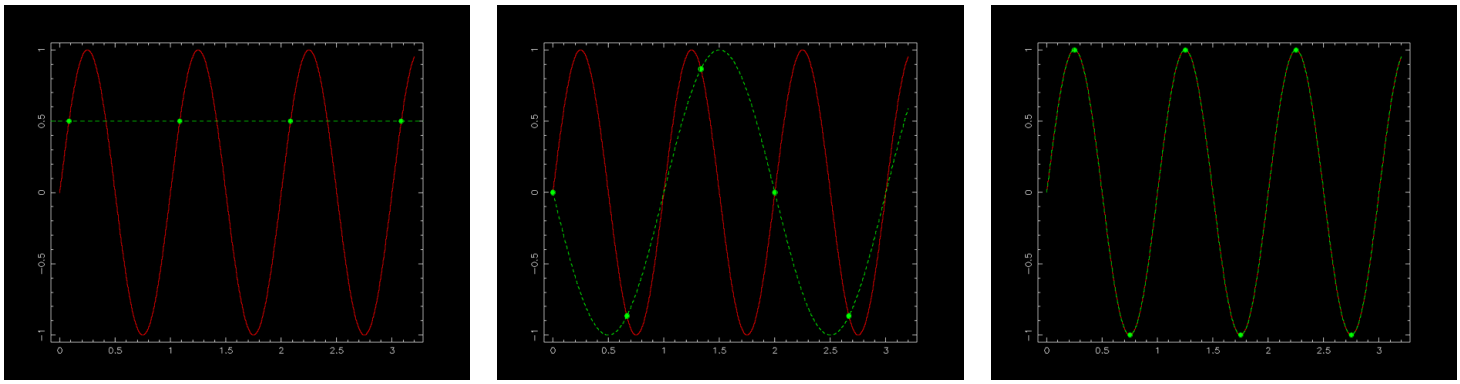
figure 74: Nyquist sampling of stellar profiles. Two profiles are shown - one at the theoretical diffraction limit of the telescope and another smeared into a seeing disc. The narrow hatched box, which is equal to half the FWHM of the Airy disc in the diffraction-limited profile, indicates the pixel size required to sample the profile at the Nyquist frequency. The wider hatched box, equal to half the FWHM of the seeing-limited profile, shows the pixel size required to sample the seeing disc at the Nyquist frequency.



To understand why there is a critical sampling frequency that determines whether or not an image or spectrum is optimally sampled, it is easiest to consider the case of a signal consisting of a sine curve, as shown in figure 75. Sampling the continuous sine curve only once per cycle would not reveal any variation at all (left-hand panel of figure 75). Sampling at 2/3 times the period of the sine curve (central panel of figure 75) would begin to reveal variation, but one would be fooled about its period. (This effect is known as *aliasing* and is most commonly seen in movies of the spokes in a rotating wheel, where an alias of the true period of the wheel is produced by the frame rate of the camera being below the Nyquist frequency).

Sampling the sine curve twice per cycle (right-hand panel of figure 75), as suggested by the Nyquist sampling theorem, is the minimum required to detect all of the peaks and troughs in the signal. Note, however, that one could be unlucky and still hit upon only the points at which the sine curve crosses the *x* axis - this is why the Nyquist sampling theorem states that the sampling frequency should be *greater than* twice the highest frequency contained in the signal, and why, strictly speaking, the theorem only applies to an infinitely long signal.

figure 75: A signal consisting of a continuous sine curve with unit period is shown in red. Left: Sampling only once per cycle (green dots) results in the signal being wongly interpreted as a constant (green dashed line). Centre: Sampling every 2/3 of a cycle results in the signal being interpreted as a sine curve, but of an incorrect period. This is aliaising. Right: Sampling twice per cycle, as suggested by the Nyquist theorem, allows the original signal to be recovered.



An example of how consideration of the Nyquist sampling theorem is important for the design of an imager is given in the example problems.

---

example problems

1. **An 8 m, *f/8* telescope is equipped with a re-imager and a detector of 2000 x 2000 pixels, where each pixel is 24 μm in size. What magnification must the re-imager have to give a field of view of 8' x 8'?**

   The focal length of the telescope is F = 8 x 8 = 64 m. Hence the plate scale is 206265 / 64000 = 3.2 "/mm. The field of view without the re-imager is therefore 2000 x 0.024 x 3.2 = 154". Hence the required re-imager magnification is 154"/480" = 0.32x.

2. **In the above example, what focal ratio must the camera have?**

   The focal length of the telescope plus re-imager, F_{sys}, is given by:

   $F_{sys} = F_{tel}\, M$.

   Since $M = F_{cam} / F_{coll}$, we can write:

   $F_{sys} = F_{tel}\, F_{cam} / F_{coll}$.

   Now, $F_{tel} = f_{tel}\, D_{tel}$, $F_{cam} = f_{cam}\, D_{cam}$ and $F_{coll} = f_{coll}\, D_{coll}$. Hence:

   $F_{sys} = f_{tel}\, D_{tel}\, f_{cam}\, D_{cam} / f_{coll}\, D_{coll}$.

   In order to collimate the beam from the telescope, the focal ratio of the collimator must be equal to the focal ratio of the telescope, i.e. $f_{coll} = f_{tel}$. Hence:

   $F_{sys} = D_{tel}\, f_{cam}\, D_{cam} / D_{coll}$.

   Also, since both are in the same collimated beam, the diameter of the camera must be at least as big as the diameter of the collimator, but

there is no point in making it any larger, i.e. $D_{cam} = D_{coll}$. So,

$F_{sys} = D_{tel}\, f_{cam}$.

Hence the focal ratio of the camera must be:

$f_{cam} = F_{sys} / D_{tel} = F_{tel}\, M / D_{tel} = f_{tel}\, M$ = 8 x 0.32 = f/2.6.

3. A 4 m, *f/10* telescope is equipped with a re-imager with a collimator of focal length 300 mm and a camera of focal length 100 mm. If the seeing is 0.8", what detector pixel size in μm is required to sample the image optimally?

The Nyquist sampling theorem tells us that we require at least 2 pixels across the seeing disc in order to sample the image optimally, i.e. the pixel size should be 0.4".

To convert this to μm, we need to know the plate scale of the telescope plus re-imager, p_{sys}:

$p_{sys} = 206265 / F_{sys}$,

where F_{sys} is the focal length of the telescope plus re-imager. In problem 2 above, we derived a relation between F_{sys} and the focal lengths of the telescope, collimator and camera:

$F_{sys} = F_{tel}\, F_{cam} / F_{coll}$.

Hence

$p_{sys} = 206265\, F_{coll} / F_{tel}\, F_{cam}$ = (206265 x 300) / (40000 x 100) = 15.5 "/mm.

Therefore, each pixel must be of size 0.4 x 1000 / 15.5 = 26 μm.

---

instruments

III. photometers

single-pixel versus multi-pixel photometers

Astronomical photometry is the measurement of the brightness of sources. Strictly speaking, it is not necessary to know the spatial distribution of the photons in order to make a photometric measurement - all that matters is the total number of photons received from the source. Hence it is possible to perform photometry with a single-pixel detector, such as a photomultiplier tube or an avalanche photodiode - see figure 76.

figure 76: Single pixel detectors. Left: avalanche photodiodes (APDs). Right: a photomultiplier tube.

There are a number of disadvantages, however, to single-pixel photometers:

- The pixel must be larger than the seeing disc and typical guiding errors of the source, otherwise it will not collect all of the photons. To allow for poor guiding and/or nights with poor seeing, the pixel is usually designed to be oversized, which means that extra background light from the sky is also detected, degrading the signal-to-noise ratio of the observation.
- Photometry of extended sources and crowded fields, e.g. galaxies and clusters of stars, is virtually impossible, due to the difficulty of isolating the light of individual sources/regions in a reliable manner.
- Single-pixel detectors can only record the brightness of one source at a time. This means that they can only be used in photometric conditions, i.e. when the transparency is good. If a source fades during conditions of poor transparency, it is impossible to tell if this is due to clouds or to an intrinsic change in the brightness of the source.
- Simultaneous measurement of the sky brightness is impossible with single-pixel detectors, resulting in the likelihood of subtracting an incorrect sky level from the source signal.

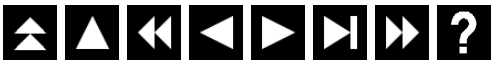
For the above reasons, single-pixel photometers are very rare noawadays, although they did play a key role in the development of the subject many decades ago. Instead, almost all photometry is now performed using imagers, which by definition possess multi-pixel detectors. *Imaging photometry* has a number of major advantages:

- The same instrument that is used to provide imaging can also provide photometry.
- The brightness of a source can be calculated by adding up all of the photons that fall within a software-defined circle (or *aperture*) centred on the source. The radius and position of this circle can be adjusted at the data reduction stage, i.e. after the observations have been performed, to cope with guiding errors and seeing variations, thereby maximising the signal-to-noise ratio.
- Photometry of crowded fields and extended sources is straightforward.
- It is possible to measure the brightness of the target and a (non-variable) comparison star at the same time, thereby allowing correction for transparency variations.
- The sky brightness can be measured simultaneously with, and at almost the same position on the sky as, the source.

We shall not discuss single pixel detectors again in this course, as shall assume from now on that all photometry is performed using imagers.

---

fluxes and magnitudes



The brightness of an astronomical source is usually given either in relative terms - the magnitude scale, or in absolute terms - the flux scale. The former is a logarithmic brightness scale, whereas the latter is a linear brightness scale. You will have already come across fluxes and magnitudes during the first-year astronomy course at Sheffield, but it is useful to revisit the basic concepts here.

fluxes

Before defining flux, it is important to define luminosity. The *luminosity, L,* of a source is defined as the total amount of radiant energy emitted over all wavelengths per unit time in all directions. The units of luminosity are joules per second (J s^{-1}) or watts (W), so you can think of luminosity as the power of the source.

Of course, it is impossible to intercept all of the energy emanating from an astronomical source and measure it. In practice, only a small fraction of the energy is ever detected, the fraction depending on the area of the collector and the distance of the collector from the source. This collected quantity is known as the *flux*, *F*, and has units of watts per square metre (W m^{-2}). The fraction of the luminosity collected by every square metre of detector located a distance *d* m from the source is simply the luminosity divided by the surface area of a sphere of radius *d* which is centred on the source, i.e.

$F = L / 4 \pi d^2$,

which shows that the flux obeys an inverse-square law with distance.

The electromagnetic radiation from most astronomical sources spans many orders of magnitude of wavelength and it is not possible to measure the flux at all wavelengths using the same equipment. Hence the flux is

often measured within a limited wavelength range and quoted in terms of unit wavelength interval, F_λ (in units of W m^{-2} nm^{-1}), or unit frequency interval, F_ν (in units of W m^{-2} Hz^{-1}). Both F_λ and F_ν are usually referred to as the *monochromatic flux* (or *flux density*) and, as the monochromatic fluxes of astronomical sources are small, the *jansky* (Jy) unit is often used, where 1 Jy = 10^{-26} W m^{-2} Hz^{-1}. F_ν and F_λ are related by the equation:

$$F = F_{bol} = \int_0^\infty F_\nu \, d\nu = \int_0^\infty F_\lambda \, d\lambda.$$

The flux, F, in the above equation is also sometimes referred to as the *bolometric flux,* F_{bol} (also in units of W m^{-2}), as it represents the total flux emitted over all wavelengths or frequencies.

Conversion between F_λ and F_ν can be achieved using the relations $F_\nu \, d\nu = F_\lambda \, d\lambda$ (which follows from differentiating the above integral) and $c = \nu \lambda$, leading to

$F_\nu = F_\lambda \lambda^2 / c$, and

$F_\lambda = F_\nu c / \lambda^2$.

You need to be careful with units when using the above flux conversion equations. For example, when converting from F_λ in units of W m^{-2} nm^{-1} into F_ν in units of W m^{-2} Hz^{-1}, you need to enter λ in units of nm and c in units of nm/s. Alternatively, you can convert F_λ to units of W m^{-2} m^{-1} by multiplying by 10^9 and then enter λ and c in units of m and m/s, respectively.

Conversion of F_λ or F_ν to the photon flux, N_λ or N_ν, can be achieved using the relation $E = h\nu = hc / \lambda$, leading to

$N_\nu = F_\nu / E = F_\nu \lambda / h c$, in units of photons s^{-1} m^{-2} Hz^{-1}, and

$N_\lambda = F_\lambda / E = F_\lambda \lambda / h c$, in units of photons s^{-1} m^{-2} nm^{-1}.

magnitudes: definition

The Greek astronomer Hipparchos is usually credited with the origin of the magnitude scale. He assigned the brightest stars he could see with his eye a magnitude of 1 and the faintest a magnitude of 6. However, in terms of the amount of energy received, a sixth magnitude star is not 6 times fainter than a first magnitude star, but more like 100 times fainter, due to the eye's non-linear response to light. This led the English astronomer Norman Pogson to formalize the magnitude system in 1856. He proposed that a sixth magnitude star should be precisely 100 times fainter than a first magnitude star, so that each magnitude corresponds to a change in brightness of $100^{1/5}$ = 2.512. For example, a star of magnitude 2 is 2.512^1=2.512 times fainter than a star of magnitude 1, a star of magnitude 6 is 2.512^2=6.3 times fainter than a star of magnitude 4, and a star of magnitude 25 is 2.512^5=100 times fainter than a star of magnitude 20. Note how it is only the magnitude difference that determines the brightness ratio of two stars, not the absolute values of their magnitudes.

Hence, *Pogson's ratio* of 2.512 leads us to *Pogson's equation*:

$$F_1/F_2 = 2.512^{-(m_1-m_2)},$$

where F_1 and F_2 are the fluxes of two stars, m_1 and m_2 are their magnitudes, and the minus sign in front of the exponent accounts for the fact that numerically larger magnitudes refer to fainter stars. It is important to note that the flux, F, in the above equation, and all the equations given below, refers to any *linear* measurement of the brightness of a star, e.g. the number of counts, joules, photons, etc, received.

Taking logarithms of Pogson's equation, we obtain:

$$log_{10}(F_1/F_2) = -(m_1-m_2) \; . \; log_{10}(2.512) = -0.4 \; (m_1-m_2).$$

More conveniently, we can write:

$$F_1/F_2 = 10^{-0.4(m_1-m_2)}, \text{ and}$$

$$m_1-m_2 = -2.5 \; log_{10}(F_1/F_2).$$

Note that the factor of 2.5 in the latter equation is not equal to 2.512 rounded down. It is an exact value, due to the fact that $\log_{10}(2.512) = \log_{10}(100^{1/5}) = 0.2 \log_{10}(100) = 0.2 \times 2 = 0.4 = 1/2.5$, precisely.

magnitudes: apparent and absolute

The magnitude of a source, m, defined above is known as the *apparent magnitude*, as it is the value measured from the Earth and does not take into account the distance of the source: a star may be intrinsically brighter than another star and yet have a higher apparent magnitude because it is further away from the Earth. To represent the intrinsic brightness of a star, the *absolute magnitude*, M, is used, which is defined as the apparent magnitude a star would have if it is 10 parsecs (pc) from the Earth, in the absence of any interstellar extinction. Hence, due to the fact that the flux of a source drops as the inverse square of the distance, d, we can write

$$F_d/F_{10pc} = [L / 4 \pi d^2] / [L / 4 \pi 10^2] = 1/(d/10)^2,$$

and

$$m_d - m_{10pc} = -2.5 \log_{10}(F_d/F_{10pc}).$$

Hence,

$$\log_{10}(F_d/F_{10pc}) = \log_{10}(1) - 2 \log_{10}(d/10),$$

and we can then write:

$$m_d - m_{10pc} = -2.5 [0 - 2 \log_{10}(d/10)],$$

or,

$$m - M = 5 \log_{10}(d/10).$$

This equation relates the apparent and absolute magnitude of a source with its distance, where d is in parsecs. For example, the brightest star in the sky, Sirius, has an apparent magnitude of m = -1.5 and distance of d

= 2.6 pc, so it has an absolute magnitude of M = 1.4. The quantity m - M is known as the *distance modulus*. If m - M = 0, 5, 10, 15, 20, 25 magnitudes, then d = 10 pc, 100 pc, 1 kpc, 10 kpc, 100 kpc and 1 Mpc, where 1 pc = 3.26 light years.

magnitudes: zero points

It can be seen that the magnitude scale is a relative one that depends on the ratio of the fluxes of two stars. It makes sense, therefore, to define a zero point, i.e. to choose a star that represents a magnitude of 0. We can then measure the magnitudes of all other stars with respect to this one. The A0V star Vega was chosen as this so-called *primary standard* because it indeed does have a magnitude close to zero as determined by Hipparchos' crude system, it is easily observable in the northern hemisphere for more than 6 months of the year, it is non-variable, relatively nearby (and hence unreddened by interstellar dust), and it has a reasonably flat and smooth optical spectrum. However, because Vega is too bright to observe with modern telescopes and instruments without saturating their detectors, and because it is not always observable, an all-sky network of fainter *secondary standards* has also been defined, where the magnitude of each star relative to Vega has been carefully calibrated. Over the years, refinements in the definition, number and measurement accuracy of the primary and secondary standards has resulted in the apparent magnitude of Vega now being 0.03 in the V-band, and it is also thought that Vega may be slightly variable, but for the purposes of this course we can ignore this few per cent offset and assume it is 0 in all bands.

If we have measured the flux of a standard star, F_{std}, and we know its catalogue magnitude, m_{std}, we can determine the magnitude of another star, m_1, whose flux, F_1, we have also measured, as follows:

$$m_1 - m_{std} = -2.5 \log_{10}(F_1/F_{std}).$$

Rearranging this equation, we obtain:

$$m_1 = m_{std} - 2.5 (\log_{10}F_1 - \log_{10}F_{std}) = m_{std} + 2.5 \log_{10}F_{std} - 2.5 \log_{10}F_1.$$

The first two terms on the right-hand side of the above equation can be

collected together into what is known as the *zero point*, $m_{zp} = m_{std} + 2.5 \log_{10} F_{std}$, giving

$m_1 = m_{zp} - 2.5 \log_{10} F_1$.

A useful way of thinking about zero points is to note that the above equation can be rewritten as:

$m_1 - m_{zp} = - 2.5 \log_{10} (F_1/1)$.

This means that if F_1 is the number of counts, photons or joules received per second from a star of magnitude m_1, then the zero point is the magnitude of a star that would give **one** count, photon or joule per second when measured with the same equipment.

magnitudes: rule of thumb

A useful rule of thumb when thinking about magnitudes is that:

- 1 magnitude corresponds very approximately to a 100% change in brightness (strictly speaking, it is a factor of 2.512, i.e. 151.2%);
- 0.1 magnitude corresponds approximately to a 10% change in brightness (strictly speaking, it is a factor of 1.096, i.e. 9.6%);
- 0.01 magnitude corresponds to a 1% change in brightness (strictly speaking, it is a factor of 1.009, i.e. 0.9%);
- 0.001 magnitude (i.e. 1 millimagnitude) corresponds to a 0.1% change in brightness (strictly speaking, it is a factor of 1.0009, i.e. 0.09%);

It is also helpful to have a feeling for the magnitudes of various well-known astronomical objects:

-26.7 = Apparent *V*-band magnitude of the Sun
-12.9 = Apparent *V*-band magnitude of the Full Moon
-4.7 = Apparent *V*-band magnitude of Venus (brightest planet)
-1.5 = Apparent *V*-band magnitude of Sirius (brightest star)
0 ~ Apparent *V*-band magnitude of Vega (α Lyrae)
6 ~ Apparent *V*-band magnitude of the faintest stars visible to the

naked eye under good conditions

16 ~ Apparent *V*-band magnitude of the faintest stars visible to the eye through a 16'' telescope

30 ~ Apparent *V*-band magnitude of the faintest objects detected by the Hubble Space Telescope

---

photometric systems

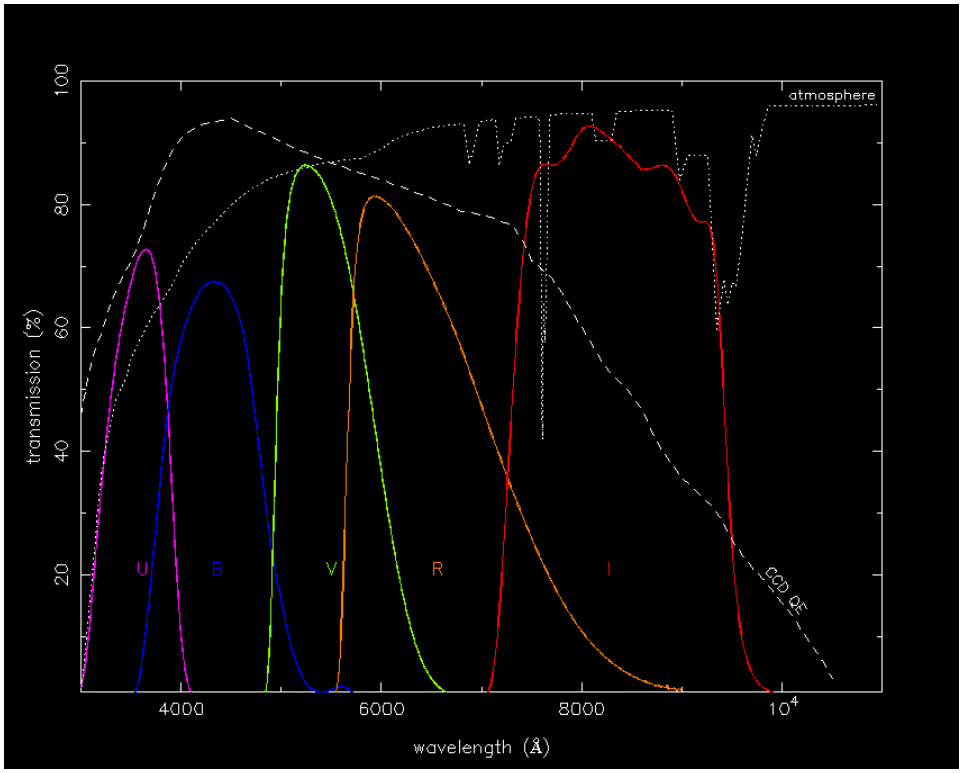
In a perfect world, astronomers would not bother with photometry. It would be far better to have spectra of all astronomical sources. That way, the flux of the source would be known at every wavelength, and detailed information on the physical conditions, chemical composition and kinematics of the source could be determined. In practice, however, not enough photons can always be collected to split the light into its component wavelengths with a usable signal-to-noise ratio. Also, for reasons that will become apparent when we discuss spectrographs, it can be difficult to measure accurate absolute fluxes from spectroscopy.

The next best option, then, is to use filters to isolate chunks of the spectrum and take images of sources through these filters. The amount of the spectrum that a filter allows through is known as the *bandpass*. Filters are usually categorized into *narrow-band* filters, which have a bandpass of order 10 nm and typically isolate a spectral line, and *broad-band* filters, which have a bandpass of order 100 nm. The central wavelength of the filter bandpass is known as the *effective wavelength*. Most modern filters are constructed of different coloured glasses, often in conjunction with thin-film coatings to help define the bandpasses and minimise reflection at the surfaces.

A *photometric system* (also known as a *filter system*) consists of a set of filters that provides coarse spectral information about a source. There is a compromise to be made when defining a photometric system - how many filters should there be, and what bandpasses and effective wavelengths should they have? A photometric system with too many filters, each with a very narrow bandpass, would make it difficult to detect sufficient photons from a source, and strong absorption/emission features in the spectrum might adversely affect some of the bandpasses. Conversely, a photometric system with too few filters, each with a very wide bandpass, would provide insufficient spectral information. The best compromise, and most widely used photometric system, is the broad-band *UBVRI* system shown in the top panel of figure 77. The original three-filter *UBV* system (which stands for "Ultraviolet", "Blue" and "Visual") was defined by Johnson and Morgan in the early 1950's. In the mid 1960's, Johnson added *R* and *I* ("Red" and "Infrared") filters to this system, but these were superseded by the shorter effective-wavelength *R* and *I* filters introduced by Cousins in the mid 1970's. Hence the *UBVRI* filters in use today are commonly referred to as the *Johnson-Morgan-Cousins* photometric system. The only change to this system since then has been the prescription of the filters: the advent of CCDs, which have completely different spectral sensitivities to the photomultiplier tubes that were used to define the original *UBRVI* system, led Bessell in 1990 to come up with a new recipe for making *UBVRI* filters out of common coloured glasses that more closely reproduces the original Johnson-Morgan-Cousins filter profiles when used with CCDs. It is Bessell's *UBVRI* filters that are found in most of the world's observatories today.

It is standard practice to refer to apparent magnitudes in a particular photometric system by the name of the filter. Hence if the apparent magnitude of a star in the *B* band is $m_B = 15.5$, this is often referred to simply as $B = 15.5$. It is also standard practice to refer to the magnitude difference of a star observed through two different filters, the *colour index*, by the names of the filters. Hence if the apparent magnitudes of a star are $B = 15.5$ and $V = 15.0$, then the colour index is $B-V = 0.5$. The most commonly quoted colour indices in the *UBVRI* system are *U-B*, *B-V*, *V-R* and *R-I*.

figure 77: Left: Filter profiles of the Johnson-Morgan-Cousins *UBVRI* system, the most widely-used broad-band photometric system in the world. Also plotted is the transmission of the atmosphere (dotted line) and the quantum efficiency of a typical CCD (dashed line). Right: photograph of *UBVRI* filters.



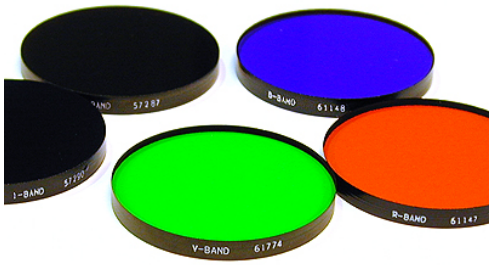


Table 1 lists the key characteristics of the Johnson-Morgan-Cousins *UBVRI* photometric system, with values obtained from Chris Benn's ING signal program. From left-to-right are tabulated the filter name, the effective wavelength (λ_{eff}), the bandpass ($\Delta\lambda$; FWHM), the approximate apparent magnitude of Vega (m_{Vega}), the band-averaged monochromatic fluxes in both frequency (F_ν) and wavelength (F_λ) units of a V=0 A0V star, and the photon flux (N_λ) in units of photons s^{-1} cm^{-2} $Å^{-1}$ (the latter units are used for convenience as they result in more easily remembered values). Note than 1 Å = 10^{-10} m = 0.1 nm. Examples of how to use the data in table 1 to convert between fluxes and magnitudes are given in the example problems.

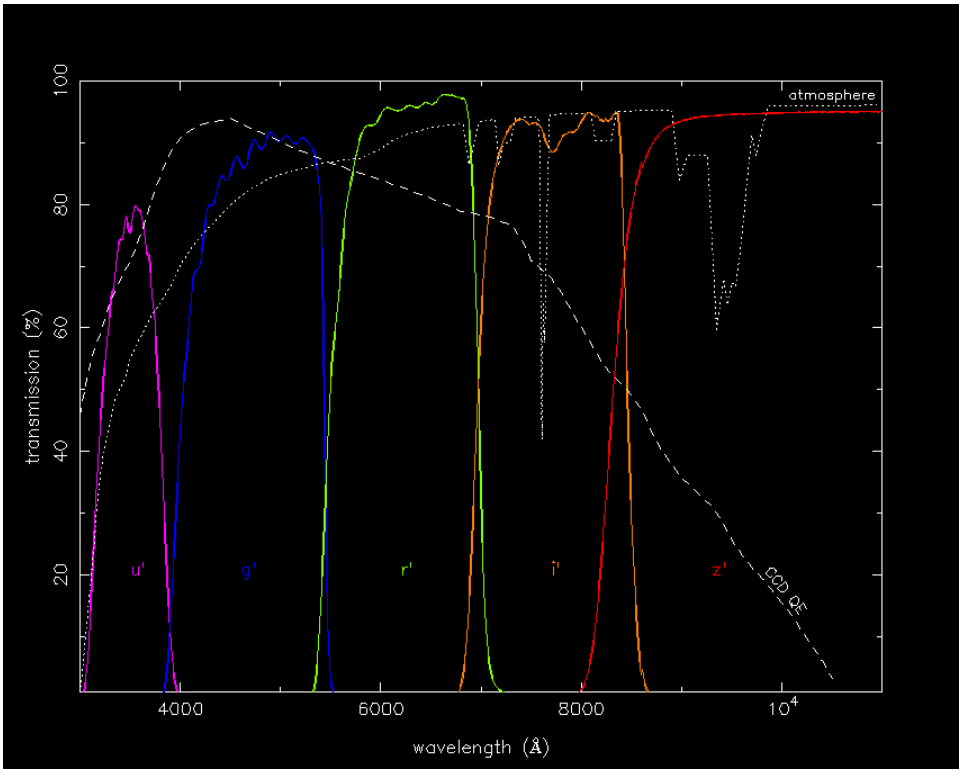
table 1: characteristics of the Johnson-Morgan-Cousins *UBVRI* photometric system. m_{Vega} refers to the magnitude of the star Vega in the filter. Strictly speaking, the magnitude of Vega is V=0.03 and all colours, e.g. *U-B, B-V*, etc, are zero. Magnitudes defined in this way are referred to as being in the *Vega magnitude system*. Alternative magnitude systems do exist, most prominently the *AB magnitude system*, in which the *V*-band magnitude of Vega is still 0.03 but only a star with a flat spectrum, i.e. F_ν = constant, has the same magnitude in all filters (and hence zero colour). Note that any photometric (filter) system can be used with any magnitude system, and it is important not to confuse the two.

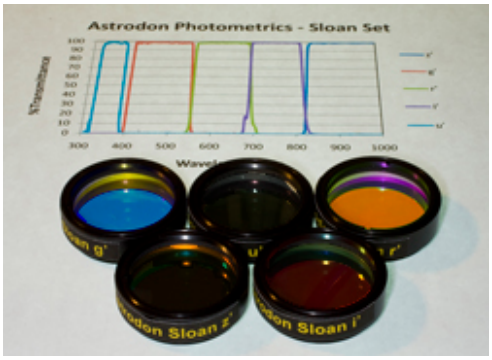
filter	λ_{eff} (nm)	$\Delta\lambda$ (nm)	m_{Vega}	F_ν (W m^{-2} Hz^{-1})	F_λ (W m^{-2} nm^{-1})	N_λ (photons s^{-1} cm^{-2} $Å^{-1}$)
U	360	50	0	1.81×10^{-23}	4.19×10^{-11}	759
B	430	72	0	4.26×10^{-23}	6.91×10^{-11}	1496
V	550	86	0	3.64×10^{-23}	3.61×10^{-11}	1000
R	650	133	0	3.08×10^{-23}	2.19×10^{-11}	717
I	820	140	0	2.55×10^{-23}	1.14×10^{-11}	471

Although widely used, the *UBVRI* photometric system was not chosen by the Sloan Digital Sky Survey (SDSS), which is one of the most important projects in astronomy today. Instead, they defined a new system: the SDSS photometric system is composed of five filters, *u'g'r'i'z'*, which have broader bandpasses than *UBVRI*, as shown in figure 78. Their broader bandpasses, in conjunction with their higher transmissions, make *u'g'r'i'z'* filters superior to *UBVRI* for photometry of faint sources. Other advantages of the SDSS filters are that there is

minimal overlap between the bandpasses, they are not truncated at the red end of the spectrum and hence cover the entire optical range, and each bandpass (apart from the red end of *z'*) is defined by a combination of coloured glass and thin-film coating, making their bandpasses very stable. With 25% of the sky now surveyed, and a catalogue containing *u'g'r'i'z'* magnitudes for 230 million celestial objects, it is likely that this photometric system will become dominant in the future.

figure 78: Left: Filter profiles of the Sloan Digital Sky Survey (SDSS) *u'g'r'i'z'* system. Also plotted is the transmission of the atmosphere (dotted line) and the quantum efficiency of a typical CCD (dashed line). Right: photograph of *u'g'r'i'z'* filters.





extracting photometric data

Determining the uncalibrated signal from a source in an image is usually a three-step process. The first step is to measure the centre of the source, which for the rest of this discussion we shall assume is a star. The second step is to estimate the sky background at the position of the star. The third step is to calculate the total amount of light received from the star.

centroiding

The first step - accurately determining the centre of the star in the image (or *centroiding*) - is usually achieved by adding up the light from the star along the rows and then the columns of the CCD, giving a one-dimensional stellar profile in the *x* direction and another in the *y* direction. The resulting profiles, which are known as *Point Spread Functions* or *PSFs* are then fit with a one-dimensional Gaussian (or similar) function, akin to those shown in figure 80. The position of the centre of the Gaussian fit is then used as an estimate of the centre of the star, which can change as a function of time due to guiding and seeing variations. This technique can fail in crowded fields, or if the stellar PSF is very faint or highly non-Gaussian; a number of more complex centroiding techniques exist for such cases.

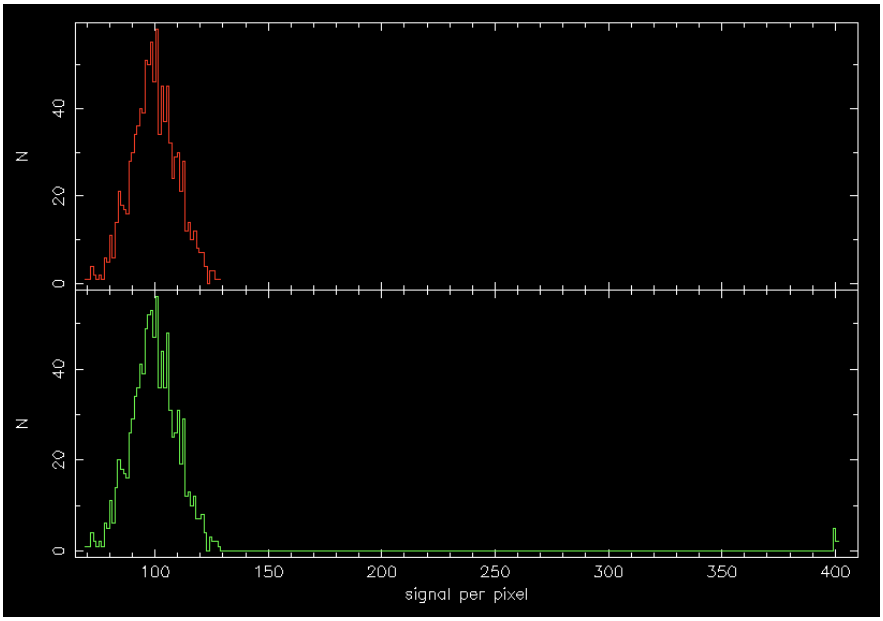
sky background

Every pixel containing light from the star also contains light from the night sky. The great advantage of imaging photometry is that the level of this sky background can be estimated simultaneously with, and at almost the same position on the sky as, the star. To determine the sky background, the usual procedure is to measure the signal in an annulus centred on the star, as shown in figure 81. Such a measurement ensures that, to first order, any gradient present in the sky background is cancelled out. Clearly, the inner radius of the annulus must be large enough to avoid contamination with light from the star at the centre, and the outer radius must be large enough to ensure that the annulus contains sufficient pixels for a robust estimate of the sky signal.

Unfortunately, the annulus is unlikely to contain signal from the sky alone. There

will also be contributions from cosmic rays, hot pixels, faint stars, and the wings of the PSF of the central star. All of these will add a positive skew to the histogram of pixel values in the annulus. The mean of these pixel values will then not be an accurate representation of the sky background. Instead, the sky level is usually determined using a more robust estimator, such as the median (i.e. the central value) or a clipped mean, as shown in figure 79.

figure 79: The top histogram shows the distribution of the signal in a sky annulus containing 1000 pixels. The mean and median of this distribution have the same value, 99.5, which represents the background sky level. The bottom histogram shows the distribution of the signal in the same sky annulus, but with the addition of 7 cosmic rays, each with a signal of 400, resulting in the small peak at the right-hand side. This skews the mean of the distribution to a value of 101.8, but the median effectively rejects these outliers and remains at the true sky level of 99.5. Therefore, adopting the mean rather than the median would result in a 2% error in the sky signal.

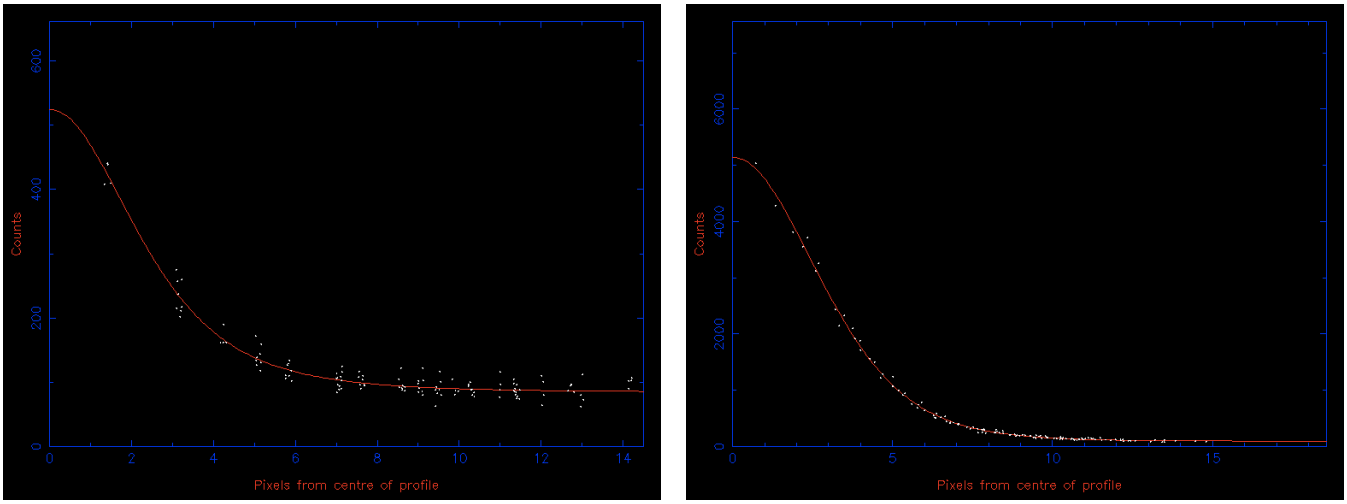


extraction

There are two ways in which the total signal from a star can be *extracted* from an image. The first is *profile fitting*, where the star is fitted with a function that closely matches its shape (e.g. a Gaussian, modified Lorentzian, or Moffat profile), as shown in figure 80, and the function is then integrated to calculate the area underneath it. It is also possible to use the bright stars in an image to define an empirical profile, which is then fitted to the star of interest. Profile-fitting techniques are mainly used for faint stars and/or crowded fields, and they often assume that all the stars in an image have the same PSF. The latter

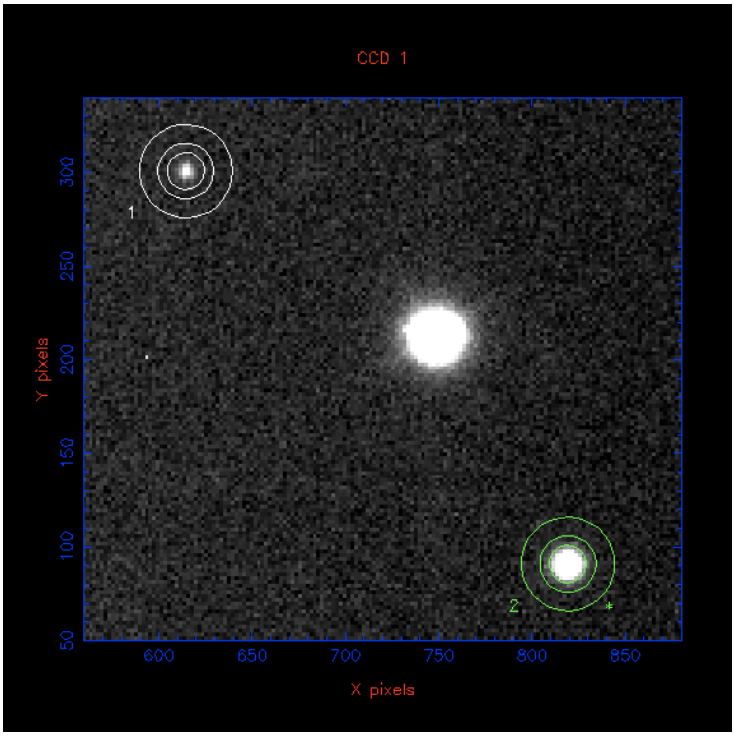
assumption can be erroneous in the presence of off-axis aberrations or seeing-induced spatial variations in the PSF.

figure 80: Left: Moffat fit to the faint star (aperture 1) in figure 81. Right: Moffat fit to the bright star (aperture 2) in figure 81. Note how their FWHM are almost identical, even though the bright star appears to cover more pixels than the faint star in figure 81. The background sky level is given by the vertical offset of the fit from the *x* axis.



The second, more commonly-used approach to extracting photometric data is known as *aperture photometry*. In this case, a software *aperture*, usually a circle or ellipse, is centred on the star, as shown in figure 81. The total signal from the star can then be calculated by summing the signal from each pixel that falls inside the aperture, and then subtracting the previously-determined sky background from each pixel.

figure 81: Aperture photometry from a CCD image. The target, an eclipsing cataclysmic variable star, is labelled as aperture 1. The non-variable comparison star is labelled as aperture 2 (in green). The unmarked bright star is saturated and hence unusable. The inner circle defines the total signal from the target. The annulus defined by the two outer circles is used to calculate the signal from the sky.

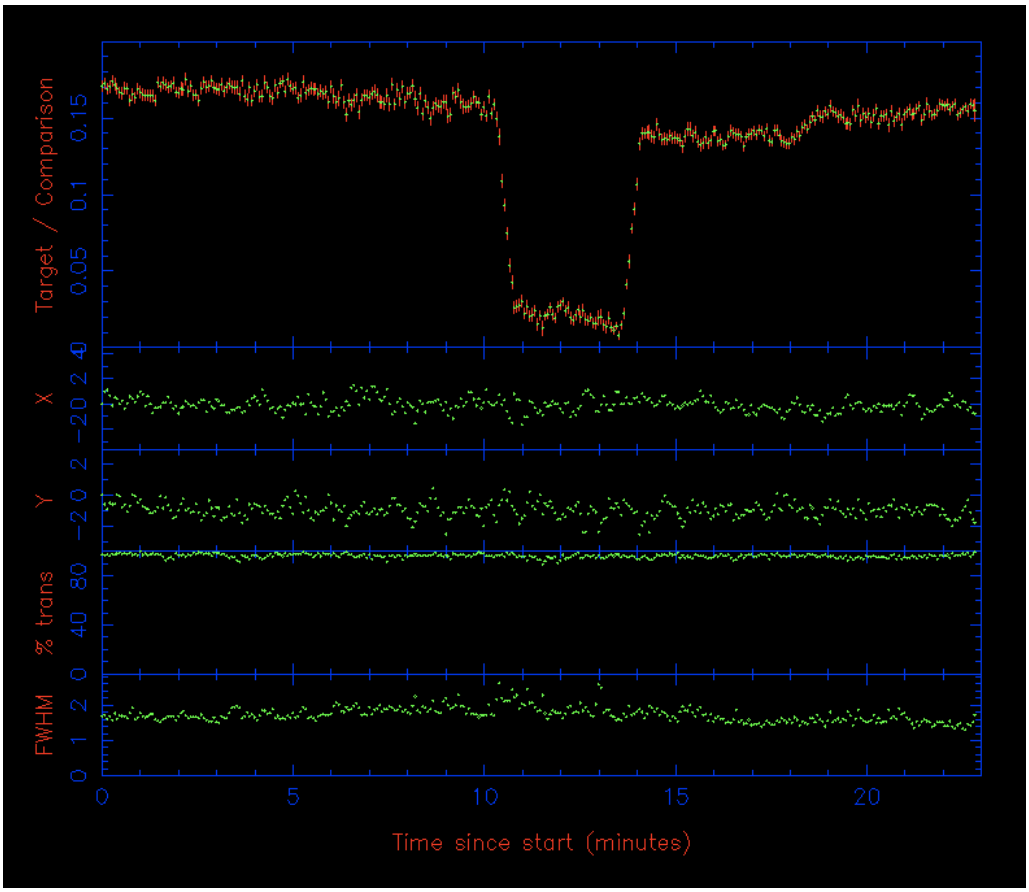


The optimum aperture size to use depends on the brightness of the star. As the aperture size is increased, an increasing fraction of the star light will obviously be included in the sum. However, the increasing aperture size also allows in extra light from the sky, as well as extra detector noise, both of which will degrade the signal-to-noise ratio. Hence, as the aperture size is increased, a point will come when the combination of the extra sky and detector noise will dominate over the extra light included from the star. If one is interested in the total signal from the star, a common procedure then is to plot a *curve of growth*, which is a plot of the signal from the star as a function of aperture radius. This will asymptotically reach a maximum value that can be estimated from the graph.

If, instead, one is interested in the relative signal from the star, e.g. when studying variability, then measuring the total signal is not important. In this case, the signal from the variable star (aperture number 1 in figure 81) is divided by the signal from a nearby, non-variable *comparison star* on the same CCD image (aperture number 2 in figure 81). (Equivalently, from m_1-m_2 = -2.5 $log_{10}(F_1/F_2)$, the *magnitude* of the comparison star is *subtracted* from the magnitude of the variable star.) This technique is known as *differential photometry* and is used to correct for transparency and seeing variations, which might otherwise mimic the intrinsic variations of the variable star. The resulting plot of the relative signal from the variable star versus time is known as a *light curve*, an example of which is shown in figure 82. The optimum-sized aperture to use in this case is one which minimises the scatter in the light curve, which can be deduced by trial and error by changing the aperture size so as to minimise the standard deviation in a flat portion of the light curve. It is

important to note here that the same sized aperture must be used for both the variable star and the comparison star, as the fraction of the total signal in an aperture of a given size is the same regardless of the brightness of a star (as demonstrated in figure 80). Hence, if the seeing increases, the fraction of light lost from the apertures of both stars is the same and dividing one by the other will then correct for the resulting signal variation.

figure 82: Differential photometry obtained with ULTRACAM. The top panel shows the light curve of an eclipsing cataclysmic variable: the signal from aperture 1 (the variable) in figure 81 is divided by the signal from aperture 2 (the comparison star) and plotted as a function of time. The two panels below this show the centroid of aperture 2 in both the *x* and *y* directions. The units are pixels, where 1 pixel = 0.3" and zero corresponds to the position of the star in the first CCD image. The variability in the centroid is due to the autoguider correcting the telescope tracking errors. The panel below this shows the sky transparency, as determined from the signal of the comparison star, where 100% represents the maximum signal observed. The bottom panel shows the seeing (i.e. the FWHM of the PSF) in arcseconds, as measured from aperture 2.



calibrating photometric data

Once the sky-subtracted signal of a star has been extracted from a CCD image, it is usually desirable to calibrate the signal by converting it to a magnitude tied to a photometric system. Unless very accurate photometry is required, this involves only 5 steps:

1. Convert the signal from the target star, which is usually in units of *counts*, to a signal per unit time interval. This can be achieved by dividing the signal by the exposure time in seconds, giving units of counts per second.
2. Calculate the instrumental magnitude from the number of counts per second.
3. Determine the extinction coefficient, and then correct the instrumental magnitude to the above-atmosphere value.
4. Repeat the above steps for a standard star and use the resulting above-atmosphere instrumental magnitude of the standard star to calculate the zero point.
5. Use the zero point to transform the above-atmosphere instrumental magnitude of the target star to the required photometric system.

The above steps are described in greater detail below.

instrumental magnitudes

Ths sky-subtracted signal from the target star, F, can be converted to a so-called *instrumental magnitude* using the formula:

$m_{inst} = -2.5 \log_{10} (F/t_{exp})$,

where t_{exp} is the exposure time in seconds and F is most likely to be in counts. The above formula follows from the more general equation $m_1 - m_2 = -2.5 \log_{10}(F_1/F_2)$ by setting $m_2 = 0$ and $F_2 = 1$, i.e. assuming a zero magnitude star gives a flux of unity.

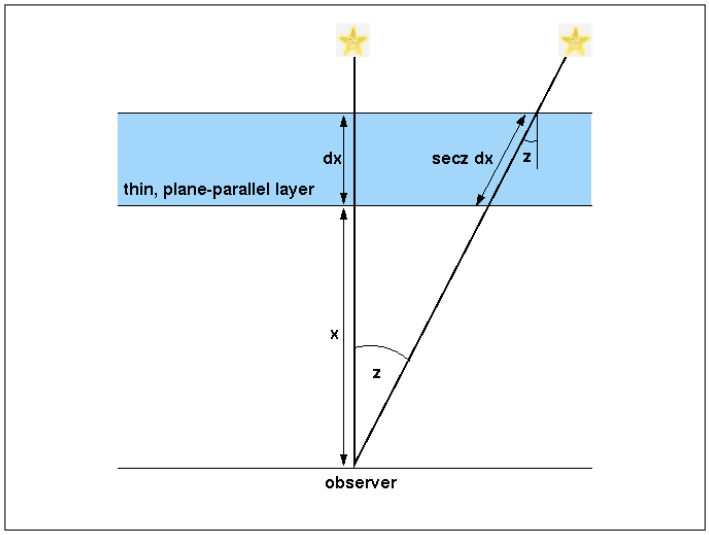
The instrumental magnitude depends on the characteristics of the telescope, instrument, filter and detector used to obtain the data. For example, if 150,000 counts are observed from a star in 1 second, the instrumental magnitude of the star is -12.9. At first glance, this suggests that the star is as bright as the Full Moon, which has an apparent V-band magnitude of -12.9. However, the latter is in the *UBVRI* photometric system, whereas the former is in an arbitrary photometric system defined by the observer's equipment; it is hence meaningless to compare the two values.

extinction

The next step is to convert the instrumental magnitude, which is measured on the surface of the Earth, to the instrumental magnitude that would be observed above the atmosphere. This is necessary because the Earth's atmosphere absorbs light from the star, and the amount of light absorbed depends on the angle of the star above the horizon. Hence if we are to compare the magnitudes of two stars reliably, we must ensure that this atmospheric effect, known as *extinction*, is removed.

We can derive a simple equation for the extinction correction by assuming that the atmosphere is a series of thin plane-parallel layers. Figure 83 shows one such layer, of thickness dx at an altitude x. The path length through the layer for light from a star at a zenith distance z is equal to dx/cos z = dx sec z. The term sec z is known as the *airmass*, and is sometimes given the symbol X. At the zenith, sec z = 1, and this increases to a value of 2 at a zenith distance of 60^o. For zenith distances greater than $\sim 60^o$, the plane-parallel approximation breaks down and a relation that takes the curvature of the atmosphere into account should be used. A number of such relations exist, e.g. $X = \sec z\ [1 - 0.0012\ (\sec^2 z - 1)]$, but it is inadvisable to observe objects at the large zenith distances where use of this equation becomes important.

figure 83: A thin, plane-parallel layer in the Earth's atmosphere. As the zenith distance of the star increases, the path length through the atmosphere increases, and hence the absorption increases.



If the monochromatic flux from a star incident on the layer is F_λ, then the decrease in flux, dF_λ, on passing through the layer will be proportional to F_λ and the path length through the layer. We can thus write:

$dF_\lambda = -\alpha_\lambda F_\lambda \sec z \, dx$,

where the constant of proportionality, α_λ, is known as the *absorption coefficient*, with units of m^{-1}. The absorption coefficient is a function of the composition and density of the atmosphere, and hence the altitude of the layer, x. Rearranging this equation, dropping the λ subscripts for clarity, and then integrating from the top of the atmosphere t to the bottom b, we obtain:

$$\int_t^b dF / F = - \sec z \int_t^b \alpha \, dx.$$

Hence,

$$F_b / F_t = F / F_0 = \exp (- \sec z \int_t^b \alpha \, dx),$$

where for clarity we have renamed the above-atmosphere flux $F_t = F_0$ and the flux measured at the ground by the observer $F_b = F$. Given the general relation between fluxes and magnitudes, $m_1 - m_2 = -2.5 \log_{10}(F_1/F_2)$, we can then write:

$$m - m_0 = -2.5 \log_{10}(F / F_0) = -2.5 \log_{10} [\exp (- \sec z \int_t^b \alpha \, dx)] = 2.5 \sec z \log_{10} e \int_t^b \alpha \, dx.$$

Defining the *extinction coefficient*, k, as:

$$k = 2.5 \log_{10} e \int_t^b \alpha \, dx,$$

we finally obtain:

$$m = m_0 + k \sec z,$$

where m_0 is the magnitude of a star observed above the atmosphere and m is the magnitude of a star observed at the Earth's surface at zenith distance z. Hence if the extinction coefficient in the V band is $k = 0.15$ magnitudes/airmass, then a star would appear 0.15 magnitudes fainter at the zenith than it would appear above the atmosphere, and 0.3 magnitudes fainter than above the atmosphere when at a zenith distance of 60^o.

The dominant source of extinction in the atmosphere is Rayleigh scattering by air molecules. This mechanism is proportional to λ^{-4}, which means that extinction is much higher in the blue than the red. The extinction can also vary from night to night depending on the conditions in the atmosphere, e.g. dust blown over from the Sahara can increase the extinction on La Palma during the summer by up to 1 magnitude. Table 2 lists the extinction coefficients on a typical (undusty) night on La Palma in *UBVRI*. For reference, the night sky brightness on La Palma when the Moon is 0% (Dark), 50% (Grey) and 100% (Bright) illuminated is also listed. All values in table 2 have been taken from Chris Benn's ING signal program.

table 2: typical extinction and sky brightness values in the Johnson-Morgan-Cousins *UBVRI* photometric system at a high-quality astronomical site.

filter	λ_{eff} (Å)	k (magnitudes/airmass)	m_{sky} (magnitudes/arcsecond^2)		
			Dark	Grey	Bright
U	3600	0.55	22.0	20.0	17.7
B	4300	0.25	22.7	20.7	18.4
V	5500	0.15	21.9	19.9	17.6
R	6500	0.09	21.0	19.7	17.5
I	8200	0.06	20.0	18.9	16.7

To measure the extinction on a particular night, it is necessary to measure the signal from a non-variable star at a number of different zenith distances. Inspecting the equation $m = m_0 + k \sec z$, it can be seen that the extinction would be given by the gradient of a plot of the instrumental magnitude of the star versus sec z and the y-intercept would give the above-atmosphere instrumental magnitude. Although such a plot would give the most accurate answer, it is also possible to obtain an estimate of k from just two measurements of the instrumental magnitude of a star at two different zenith distances: subtracting $m_{z_1} = m_0 + k \sec z_1$ from $m_{z_2} = m_0 + k \sec z_2$ eliminates m_0, allowing k to be derived (see the example problems).

Note that no explicit extinction correction is required when performing differential photometry. This is because the target and comparison stars are always observed at the same airmass and hence suffer the same extinction. Hence, when the target signal is divided by the comparison star signal to correct for transparency variations, the variation due to extinction present in the comparison star is removed from the target star.

For very accurate photometry, the wide bandpass of broad-band filters has to be taken into account when correcting for extinction. Having a single, average extinction coefficient for each filter means that a blue star would actually suffer from more extinction than corrected for. Conversely, a red star would be over-corrected for extinction. The solution is to introduce a colour-dependent *secondary extinction coefficient,* k_2, which modifies the above extinction correction equation to:

$m = m_0 + k \sec z + k_2 C \sec z$,

where C is the *colour index*, e.g. when correcting a V-band magnitude for extinction, $C = B - V$. The secondary extinction coefficient is usually of order a hundredth of a magnitude, so it will be ignored for the remainder of this course.

zero points

The final step in photometric calibration is to convert the above-atmosphere instrumental magnitude of a star, m_{inst0}, to a magnitude tied to a photometric system. To do this, it is necessary to observe a primary or secondary photometric standard star to derive the zero point. Using the steps outlined above, the above-atmosphere instrumental magnitude of the standard star, $m_{inst0std}$, should first be obtained. The zero point magnitude, m_{zp}, can then be calculated from the difference between the catalogue magnitude of the standard star, m_{std}, and the above-atmosphere instrumental magnitude of the standard star:

$m_{zp} = m_{std} - m_{inst0std}$.

The calibrated magnitude of the target star, m, can then be determined by simply adding its above-atmosphere instrumental magnitude to the zero point:

$m = m_{zp} + m_{inst0}$.

A demonstration of how the above steps are performed in practice is given in the example problems.

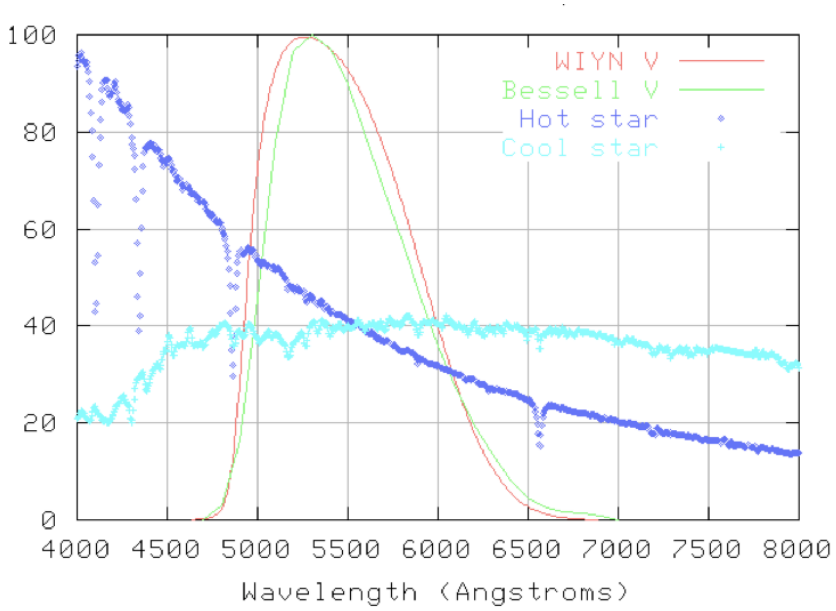
Each filter in a photometric system will have a different zero point. Once the zero point has been measured for a particular telescope, instrument, filter and detector combination, it should remain unchanged, although dirt and the degradation of the coatings on the optics will cause minor changes to the zero point on long timescales. To determine the zero points for the *UBVRI* system, the photometric standards measured by Landolt can be used:

- UBVRI photometric standard stars around the celestial equator, with a searchable catalogue.
- UBVRI photometric standard stars in the magnitude range 11.5-16.0 around the celestial equator, with a searchable catalogue.

For very accurate photometry, the wide bandpass of broad-band filters also has to be taken into account when converting instrumental magnitudes to standard values. This is because the telescope, instrument, filter and detector used by an observer will always have a slightly different response to light as a function of wavelength than those used by the astronomers who originally defined the magnitudes of the

photometric standard stars. The biggest discrepancy is often in the profile of the filter. Figure 84 shows two *V* filters used at the Kitt Peak National Observatory in Arizona. The WIYN *V* filter has a sharper increase in transmission on the blue side of the profile than the Bessell *V* filter, allowing more light to enter from hot, blue stars than from cool, red stars. This creates a systematic error that makes blue stars seem a bit brighter than red stars when observed through the WIYN *V* filter.

figure 84: A plot by Michael Richmond showing two different *V* filters and the spectra of hot and cold stars, demonstrating why correcting for colour terms is necessary when performing high-accuracy photometry (see text below for details).



Fortunately, it is straightforward to correct for this systematic error by observing a field with many standard stars possessing a range of colours. The advantage of observing a single field is that all of the stars will then be at the same airmass and hence extinction effects are cancelled out. We have already seen that the zero point, m_{zp}, is equal to the difference between the catalogue magnitude of the standard star, m_{std}, and its above-atmosphere instrumental magnitude, $m_{inst0std}$:

$$m_{zp} = m_{std} - m_{inst0std}.$$

Hence, if the telescope, instrument, filter and detector combination being used matches that of the photometric system perfectly, then we can write for stars of all colours that:

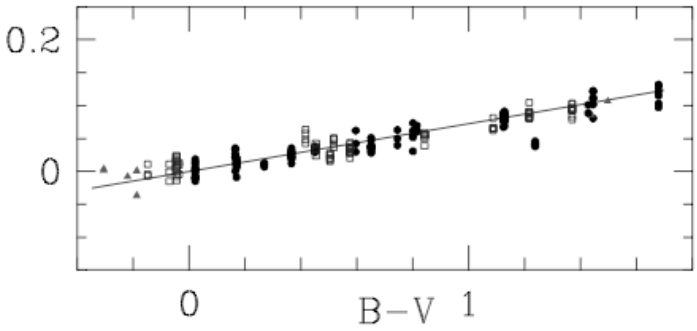
$$m_{std} - m_{inst0std} - m_{zp} = 0.$$

In practice, however, the observer's equipment is never identical to that used to define the photometric system, resulting in the above equation being modified to:

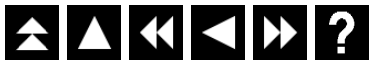
m_{std} - $m_{inst0std}$ - m_{zp} = c C.

where c is the *colour term*, C is the *colour index*, e.g. $C = B - V$. Hence, the colour term is equal to the gradient of the line in a plot of m_{std} - $m_{inst0std}$ - m_{zp} versus C, i.e. a plot of the difference between the catalogue magnitudes of the standard stars and their calibrated magnitudes as a function of colour; the y intercept is set to zero by using a zero point calculated from a star of $B - V = 0$. An example of such a plot is shown in figure 85, which has a gradient of 0.072. Colour terms are only significant if stars of extreme colours are being observed, and we shall ignore them for the remainder of this course.

figure 85: A plot of the difference between the catalogue magnitudes of a set of standard stars and their calibrated magnitudes (y axis) as a function of colour (x axis). The gradient of the line is equal to the colour term.



example problems



1. A star has a measured *V*-band magnitude of 20.0. How many photons per second are detected from this star by a 4.2 m telescope with an overall telescope/instrument/filter/detector efficiency of 30%?

The relation between fluxes and magnitudes is given by the equation:

m_1-m_2 = -2.5 $\log_{10}(F_1/F_2)$.

We can see from table 1 that a V = 0 star has a monochromatic flux of F_λ = 3.61 x 10^{-11} W m^{-2} nm^{-1}. Substituting these values into the equation above, we obtain:

20-0 = -2.5 $\log_{10}(F_1/3.61 \times 10^{-11})$,

which gives F_1 = 3.61 x 10^{-19} W m^{-2} nm^{-1} at the effective wavelength of the V band, 550 nm.

The number of photons s^{-1} m^{-2} nm^{-1} is then given by dividing this flux by the energy of a single 550 nm photon, i.e.

$N_1 = F_1 \lambda / h c$ = 1.0 photons s^{-1} m^{-2} nm^{-1}.

The total number of photons detected by a 4.2 m telescope in the V band can be obtained by multiplying this number by the collecting area of the telescope, the bandpass of the filter and the efficiency of the telescope/instrument/filter/detector:

N_1 = 1.0 x π x 2.1^2 x 86 x 0.3 = 357 photons s^{-1}.

2. A non-variable star is tracked across the sky on a photometric

night. At a zenith distance of 30^o, 100,000 photons are detected from the star in a 1 minute integration in the *V* band, which drops to 88,000 photons at a zenith distance of 60^o. What is the extinction coefficient in the *V* band on the night in question? You may assume that the signal from the sky has been subtracted from these measurements.

The instrumental magnitude of the star at each zenith distance is given by

$m_{inst} = -2.5 \log_{10} (F/t_{exp})$,

and these can be corrected for extinction using the equation:

$m_{inst} = m_{inst0} + k \sec z$.

Hence, we can write:

$m_{inst1} = -2.5 \log_{10} (100{,}000/60) = m_{inst0} + k \sec 30$, and

$m_{inst2} = -2.5 \log_{10} (88{,}000/60) = m_{inst0} + k \sec 60$.

Subtracting these two equations to eliminate m_{inst0} and rearranging for k, we obtain:

k = 0.16 magnitudes per airmass in the *V* band.

3. The star Vega is observed on the same night with the same equipment at a zenith distance of 45^o. If 1 x 10^7 photons are detected from Vega in a 10 s exposure, what is the *V*-band magnitude of the star in question 2?

The instrumental magnitude of Vega at a zenith distance of 45^o is:

$m_{instst d} = -2.5 \log_{10} (F/t_{exp}) = -2.5 \log_{10} (1 \times 10^7/10) = -15$.

The above-atmosphere instrumental magnitude of Vega is then given by:

$m_{inst0std} = m_{inst} - k \sec z = -15 - 0.16 \sec 45 = -15.2$.

In comparison, the above-atmosphere instrumental magnitude of the star in question 2 is:

$m_{inst0} = m_{inst1} - k \sec z = -2.5 \log_{10} (100{,}000/60) - 0.16 \sec 30 = -8.2$.

Table 1 tells us that Vega has a magnitude of $V = 0$. Hence the zero point of the telescope/instrument/filter/detector combination is

$m_{zp} = m_{std} - m_{inst0std} = 0 - (-15.2) = 15.2$

The calibrated V-band magnitude of the star in question 2 is then given by:

$m = m_{zp} + m_{inst0} = 15.2 + (-8.2) = 7$.

4. Using the same equipment as described in question 1, if the sky has a brightness of $V = 19.9$ magnitudes per square arcsecond, and the extraction aperture has a radius of 2.5", how many photons per second from the sky are recorded in the aperture?

The solution to this question follows the method used in question 1, but remembering that the sky is not a point source like a star. Hence, unlike a star, the larger the aperture used to measure the sky, the higher the total number of photons that will be detected from the sky. For this reason, sky magnitudes are quoted in per square arcsecond units and the area of the aperture used to extract the signal from the CCD image must be taken into account.

The relation between fluxes and magnitudes is given by the equation:

$m_1 - m_2 = -2.5 \log_{10}(F_1/F_2)$.

We can see from table 1 that a $V = 0$ star has a monochromatic flux of $F_\lambda = 3.61 \times 10^{-11}$ W m^{-2} nm^{-1}. Substituting these values into the equation above, we obtain:

$19.9\text{-}0 = -2.5 \log_{10}(F_1/3.61 \times 10^{-11})$,

which gives $F_1 = 3.96 \times 10^{-19}$ W m^{-2} nm^{-1} arcsecond^{-2} from the sky at the effective wavelength of the *V* band, 550 nm.

The number of photons s^{-1} m^{-2} nm^{-1} arcsecond^{-2} is then given by dividing this flux by the energy of a single 550 nm photon, i.e.

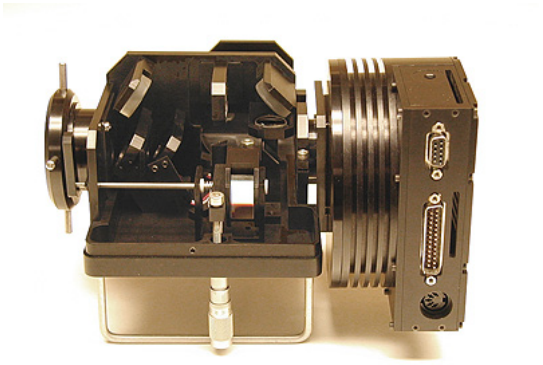
$N_1 = F_1 \lambda / h c = 1.1$ photons s^{-1} m^{-2} nm^{-1} arcsecond^{-2}.

The total number of photons per second from the sky detected in the extraction aperture by a 4.2 m telescope in the *V* band can be obtained by multiplying the number above by the collecting area of the telescope, the bandpass of the filter, the efficiency of the telescope/instrument/filter/detector, and the area of the aperture:

$N_1 = 1.1 \times (\pi \times 2.1^2) \times 86 \times 0.3 \times (\pi \times 2.5^2) = 7720$ photons s^{-1}.

---

instruments



IV. spectrographs

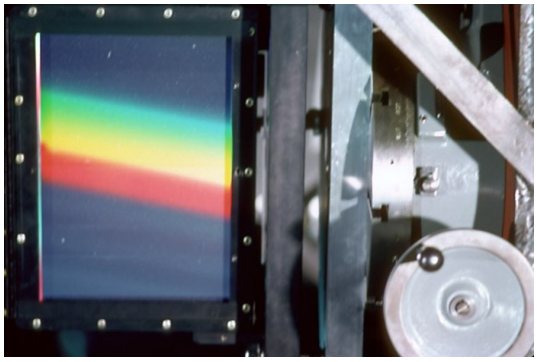
the grating equation

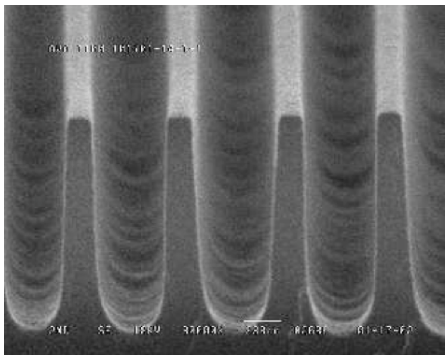
An astronomical spectrograph splits, or *disperses*, the light from a source into its component wavelengths. Some means of dispersing the light is therefore required. This function used to be performed by a prism, which exploits the fact that light of different wavelengths are *refracted* by different amounts, with blue light being refracted more than red light, as shown in figure 86. Prism spectrographs are only rarely found in astronomical spectrographs nowadays. There are a number of reasons for this, including: the dispersion is non-linear, with light in the blue part of the spectrum being dispersed more than the red, making it more difficult to analyse; the dispersion is not very high, and the only way of significantly increasing it is to use two or more prisms in tandem, which starts to become inefficient due to the loss of light at each air-glass surface and absorption within the glass itself.

figure 86: A photograph showing how white light entering a prism from the left is dispersed into a spectrum on exiting at the right. Note how the blue light is refracted by the greatest angle and also dispersed the most.

The almost universal choice for the dispersing element in modern astronomical spectrographs is the *diffraction grating*. A diffraction grating consists of a large number of fine, equidistant, parallel lines ruled onto a transparent glass plate so that light can pass between the lines, but not through them. Strictly speaking, this would be a *transmission grating*; a *reflection grating* would have the lines ruled onto a reflective glass plate so that only the light falling between the lines is reflected. An example of each is shown in figure 87. Gratings with up to 2400 lines per mm are available, produced either with a diamond tool or photolithographic (holographic) etching.

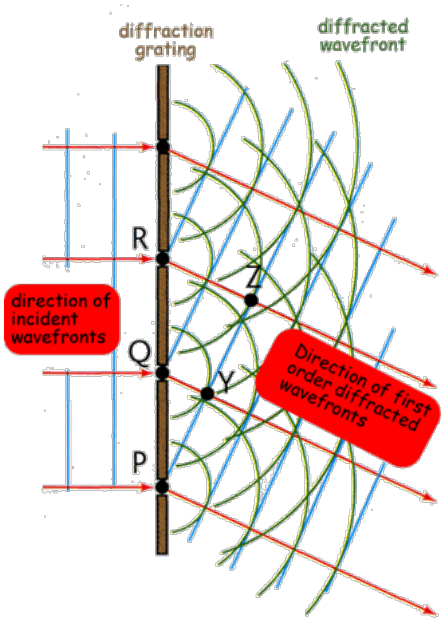
figure 87: Left: photograph of a transmission grating for the amateur astronomy market with 100 grooves per mm. Centre: photograph of a reflection grating mounted in a professional spectrograph. Right: a microscopic view of the grooves in a typical diffraction grating.

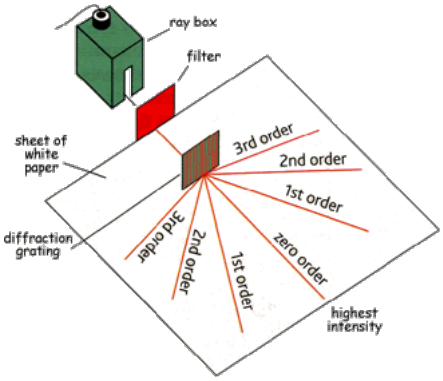


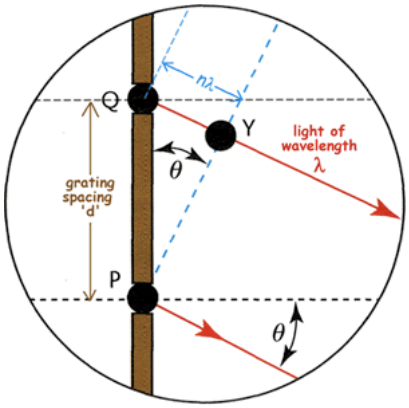


Light incident upon a grating will be diffracted by the lines, producing a series of secondary wavelets emanating from the gaps between the lines, as shown in the left-hand panel of figure 88. As each diffracted wavefront emerges from a gap, it reinforces wavefronts from each of the other gaps, i.e. there is constructive *interference*, but only at certain angles. For example, in the left-hand panel of figure 88, the *monochromatic* wavefront emerging at P reinforces the wavefront emitted from Q one cycle earlier, which reinforces the wavefront emitted from R one cycle earlier, etc. The effect is to form a new wavefront PYZ which travels in a certain direction and contributes to the first-order diffracted beam. A similar diagram could be drawn for a second-order diffracted beam, but in this case the wave emerging at P reinforces the wavefront emitted from Q *two* cycles earlier, etc. The end result is shown in the centre-left panel of figure 88, which shows that multiple orders emerge from a diffraction grating.

figure 88: Schematic showing how parallel, monochromatic wavefronts are diffracted by the gaps in the grating, forming secondary wavefronts which constructively interfere (left), transmitting light in certain directions only (centre-left). Constructive interference only occurs when the path difference between the wavefronts is equal to a whole number of wavelengths (centre right - for incident waves perpendicular to the grating; right - for arbitrary angle of incidence).







The condition for constructive interference can be derived by inspecting the centre-right panel of figure 88. The wavefront emerging from slit P reinforces a wavefront emitted *n* cycles earlier by the adjacent slit Q. This earlier wavefront therefore must have travelled a distance of *n* wavelengths from the slit. Therefore the perpendicular distance QY from the slit to the wavefront is equal to $n\lambda$, where λ is the wavelength of the light. Since the angle of diffraction of the beam, θ, is equal to the angle between the wavefront and the plane of the slits, it follows that $\sin\theta$ = QY/QP, where QP is the grating spacing (i.e. the centre-to-centre distance *d* between adjacent slits). Substituting *d* for QP and $n\lambda$ for QY and rearranging gives the *grating equation*:

$$n\lambda = d \sin\theta.$$

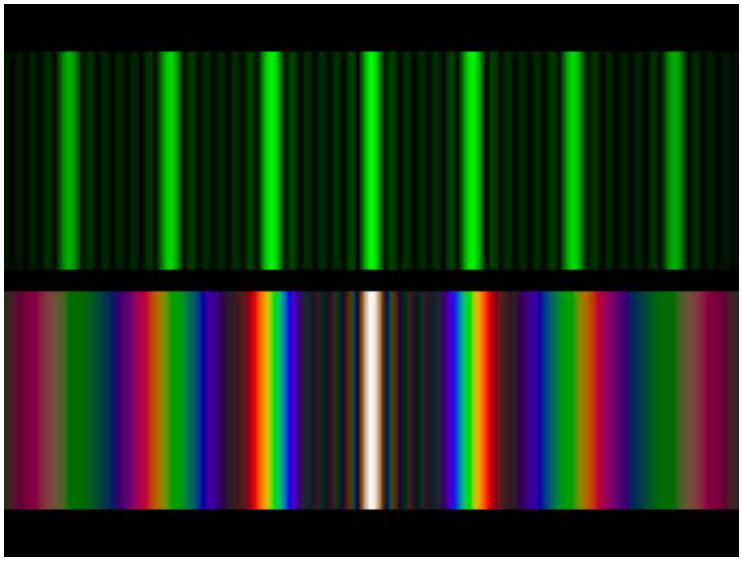
In deriving the above form of the grating equation, we assumed an incident beam perpendicular to the grating. If, instead, the incident beam makes an angle ϕ with respect to the grating, then the right-hand panel of figure 88 shows that the total path difference becomes $x + y = n\lambda$. Given this, we arrive at the more generalized form of the grating

equation:

$n \lambda = d (\sin \theta + \sin \phi)$.

The discussion above has considered only a monochromatic incident wavefront. If more than one wavelength is present, each wavelength will be diffracted through different sets of angles as defined by the grating equation. The diffraction grating will thus disperse the light incident upon it into its component wavelengths, as shown in figure 89. It can be seen that the zeroth order is undispersed, which follows from the grating equation - when $n = 0$, then $\theta = 0$ and hence light of all wavelengths is undeviated. The first orders (n = ±1) fall either side of the zeroth order, with the blue light deviated the least and red the most, opposite to that observed with a prism (figure 86). This follows from the grating equation - for a fixed n and d, θ must *increase* as λ increases. Outside the first order can be seen the second (n = ±2) and third (n = ±3) orders, which appear successively wider and fainter. As we shall see, the higher dispersion in higher orders is exploited in high-resolution spectrographs, in conjunction with blazed gratings to overcome the efficiency loss.

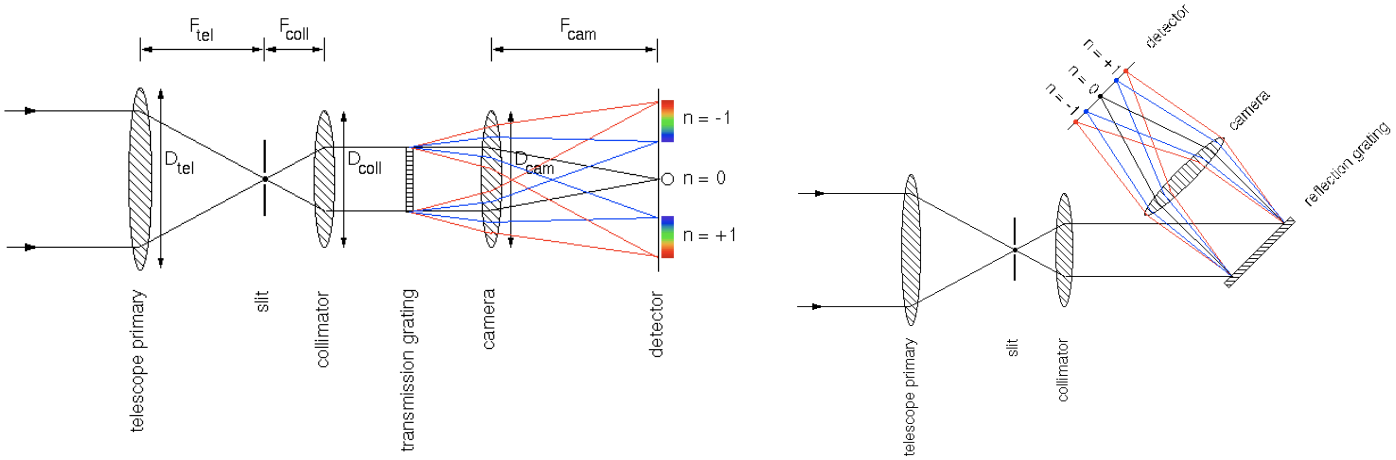
figure 89: Photograph showing the diffraction pattern produced by a monochromatic light source (top) and a white light source (bottom) incident on a diffraction grating. The central bright line is the zeroth order, with orders ±1, ±2 and ±3 shown either side of this. The fainter lines between the bright lines are due to the diffraction pattern produced by the finite width of the grooves in the grating.

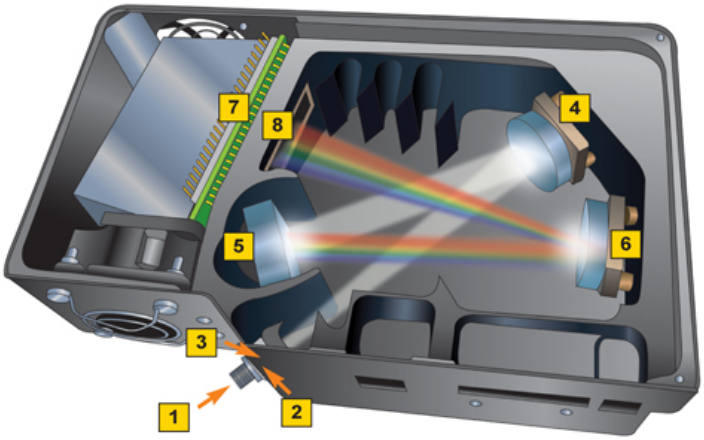


basic spectrograph design

Most astronomical spectrographs have the same basic design, regardless of whether they are to be used for low-, medium- or high-resolution spectroscopy. A schematic is shown in figure 90.

figure 90: Top left: schematic of an astronomical spectrograph incorporating a transmission grating. The slit is in the focal plane of the telescope. The grating is conjugate with the primary mirror. The detector is in the focal plane of the camera. Top right: as above, but incorporating a reflection grating. Bottom: the light path through a commercial spectrograph, showing the principal components: the slit (2), collimator (4), reflection grating (5), camera (6) and detector (7).



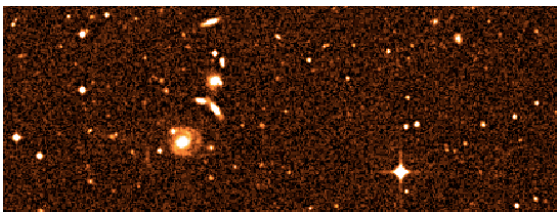


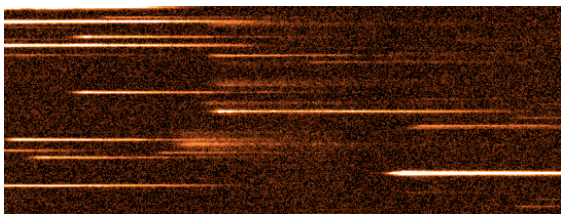
The five basic components of an astronomical spectrograph are the *slit*, *collimator*, *grating*, *camera* and *detector*, each of which are described in more detail below. Note that, with the exception of the slit and grating, spectrographs have the same basic layout as re-imagers, and hence much of the material we covered on the latter topic will be applicable here.

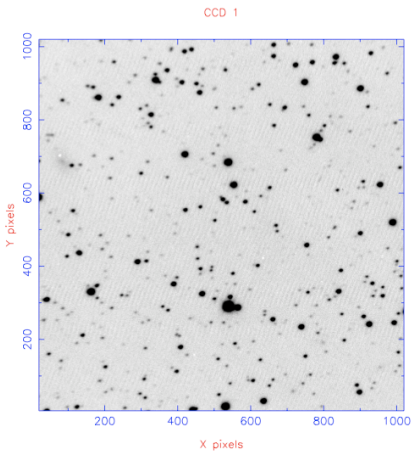
slit

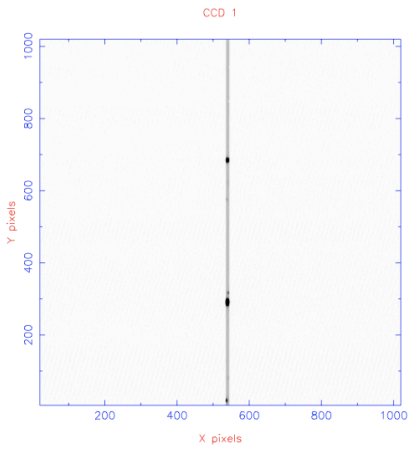
The slit is a mask with a narrow rectangular aperture that is placed in the focal plane of the telescope. The slit has two main functions. First, the slit acts as a means of isolating the region of interest on the sky; only light falling on the slit may enter the spectrograph, as shown in figure 91. Without a slit present, the spectra from sources either side of the target would overlap, contaminating the target spectrum. Additional sky background from either side of the target would also be recorded, degrading the signal-to-noise ratio of the spectrum.

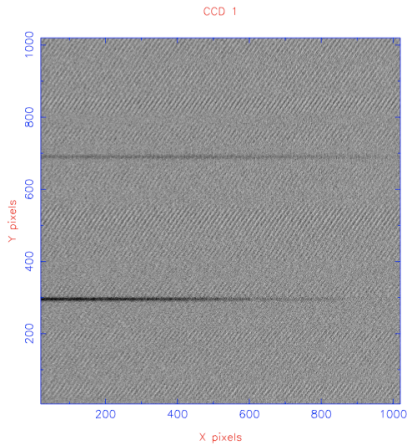
figure 91: Top left: Photograph of an adjustable slit. Top middle: image of a star field. Top right: the same star field, but imaged by a slitless spectrograph. Note how the spectra of objects at the same *y* position in the image overlap. Bottom: ULTRASPEC data showing the image of a star field recorded with the slit and grating removed from the beam (left), the image of the star field when just the slit is inserted into the beam (centre), and the spectra of the stars on the slit that result when the grating is also inserted into the beam (right). The *y* axis is the spatial direction and the *x* axis is the dispersion direction.











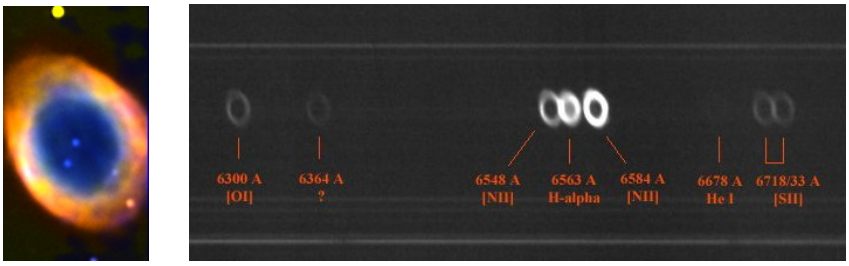
The second function of the slit is to provide stable spectral resolution. This can be understood by noting that a spectrum is essentially an infinite number of images of the telescope focal plane, each shifted slightly in wavelength. Hence, without a slit present, the spectral resolution of a star would be defined by the width of the star, i.e. the seeing. However, since seeing varies with time, the spectral resolution would then vary with time. The situation is even worse if a source is extended, such as the planetary nebula shown in the top panel of figure 92. In this case, individual spectral lines would appear as multiple images of the extended source, each shifted in wavelength, making detailed analysis of the lines almost impossible. The solution is to illuminate a narrow slit with the source, so that the spectrum becomes an infinite number of images of the slit, not the source. The minimum width of the spectral lines would then be defined by the slit width, as shown in the bottom panel of figure 92, and would hence be stable.

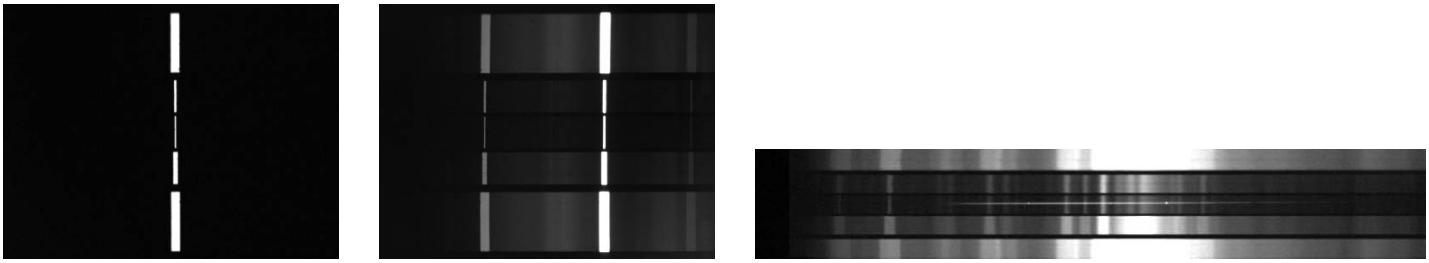
Spectrographs re-image, or *project*, the slit in the focal plane of the telescope onto the detector, which lies in the focal plane of the camera. As we discovered when looking at re-imagers, the ratio of the camera to collimator focal lengths determines the magnification of the image of the slit on the detector: $M = F_{cam} / F_{coll}$. Hence a slit of width 50 μm in a

spectrograph with a collimator of focal length 500 mm and camera of focal length 250 mm would be projected to a width of 25 μm on the detector.

We shall discuss the issue of the optimum slit width to use when we look at spectral resolution.

figure 92: Top: Image (left) and slitless spectrum (right) of the Ring Nebula (M57), showing that individual spectral lines take on the appearance of the nebula. Bottom left: Photograph of light illuminating five different slits of width, from top-to-bottom, 400 μm, 100 μm, 50 μm, 200 μm and 400 μm. Bottom middle: spectra of a hydrogen lamp, produced by illuminating the slits in the left-hand panel. Note how the spectrum is made up of images of the slit shifted in wavelength. Bottom right: spectra of the night sky emission lines, produced by illuminating the slits in the left-hand panel; the central spectrum also shows the spectrum of a star superimposed on the sky spectrum. Note how the width of the sky lines is set by the width of the slit.





collimator

The collimator in an astronomical spectrograph takes the diverging light from the slit, makes it parallel and directs the collimated beam towards the grating. Without a collimator, the diverging light from the slit would hit the grating, resulting in variable angles of incidence as a function of

position on the grating. Inspecting the grating equation, it can be seen that varying the angle of incidence implies that the angle of diffraction must also vary for a fixed wavelength and spectral order. Hence the same wavelength of light in the same order would be imaged onto different positions on the detector, blurring the resulting spectrum. With a collimator, however, all angles of incidence on the grating are equal and no such blurring would result.

Collimators can be either lenses or mirrors and, as discussed in the section on re-imagers, the collimator must have the same focal ratio as the telescope and be positioned at a distance equal to its focal length from the telescope focal plane.

grating

The diffraction grating splits the light into its component wavelengths and can be of either the transmission or reflection variety, as shown in figure 90. The grating is usually positioned in the collimated beam so that it is conjugate with the primary mirror, which is advantageous as the minimum-sized (and hence cheapest) grating is then required to collect light from all angles incident on the primary mirror.

The grating equation shows that changing the angle of the grating with respect to the incident beam, ϕ, changes the angle of the diffracted beam, θ. Hence most spectrographs have gratings that can be tilted in order to adjust the start and end wavelengths of the spectrum.

camera

The role of the camera in an astronomical spectrograph is to collect the spectrally-dispersed beams from the grating, which are still collimated, and make them converge so that the spectrum is imaged onto the detector. The camera can be either a lens, mirror, or catadioptric system.

detector

The dispersed light is ultimately imaged by the spectrograph onto the

detector, forming a *spectrum*, as shown in the bottom-right panel of figure 91. We can define the direction along the slit (i.e. the vertical direction in the bottom-right panel of figure 91) as the *spatial* axis, and the direction along the spectrum (i.e. the horizontal direction in the bottom-right panel of figure 91) as the *dispersion* axis.

Clearly, single pixel detectors are inappropriate for use in spectrographs, as the amount of light at each wavelength must be recorded. In principle, a detector composed of a one-dimensional array of pixels could be used to record a spectrum. However, such a detector would suffer from the same disadvantages that were listed for single-pixel detectors used for imaging, i.e. no spatially extended sources could be studied, no simultaneous sky background could be measured, and oversized pixels in the spatial direction would have to be used to collect all of the stellar flux, degrading the signal-to-noise ratio due to the additional sky background detected. For this reason, the detector of choice on virtually all of the world's astronomical spectrographs is two-dimensional - the CCD detector.

---

dispersion and spectral resolution

Two of the most important properties of a spectrograph are the *dispersion*, which sets the wavelength range of the spectrum, and the *spectral resolution*, which sets the size of the smallest spectral features that can be studied in the spectrum.

dispersion

Recalling that the grating equation is given by

$n \lambda = d (\sin \theta + \sin \phi)$,

the *angular dispersion* is defined as the rate of change of the angle of the dispersed light, θ, with wavelength, λ. We can obtain an expression for the angular dispersion by differentiating the grating equation with respect to wavelength, noting that ϕ is a constant:

$d(\sin \theta) / d\lambda = \cos \theta \, (d\theta / d\lambda) = n / d$.

Hence,

$d\theta / d\lambda = n / d \cos \theta$.

Thus the angular dispersion of a spectrum, in units of radians per unit wavelength, is greater for higher orders of the spectrum (larger n) and smaller values of the grating spacing d.

It is generally more convenient to express dispersion in terms of a linear scale at the detector rather than an angle. Hence, *linear dispersion* is defined as the rate of change of the linear distance, x, along the spectrum with wavelength:

$dx / d\lambda = (dx / d\theta) (d\theta / d\lambda) = F_{cam} (d\theta / d\lambda)$,

where F_{cam} is the focal length of the spectrograph camera, which is equal to the inverse of the plate scale, $p = d\theta / dx$. This equation is often inverted to give the *reciprocal linear dispersion*, i.e. the wavelength range for a particular length at the detector:

$d\lambda / dx = d \cos \theta / n F_{cam}$.

Note that the reciprocal linear dispersion is a length divided by a length, and this can lead to confusion with units. Generally, the reciprocal linear dispersion is expressed in

units of nm/mm or Å/mm. For example, if a spectrograph has a linear dispersion of 10 Å/mm and the CCD detector being used has a size of 20 mm in the dispersion direction, then the *wavelength range* of the resulting spectrum will be 200 Å.

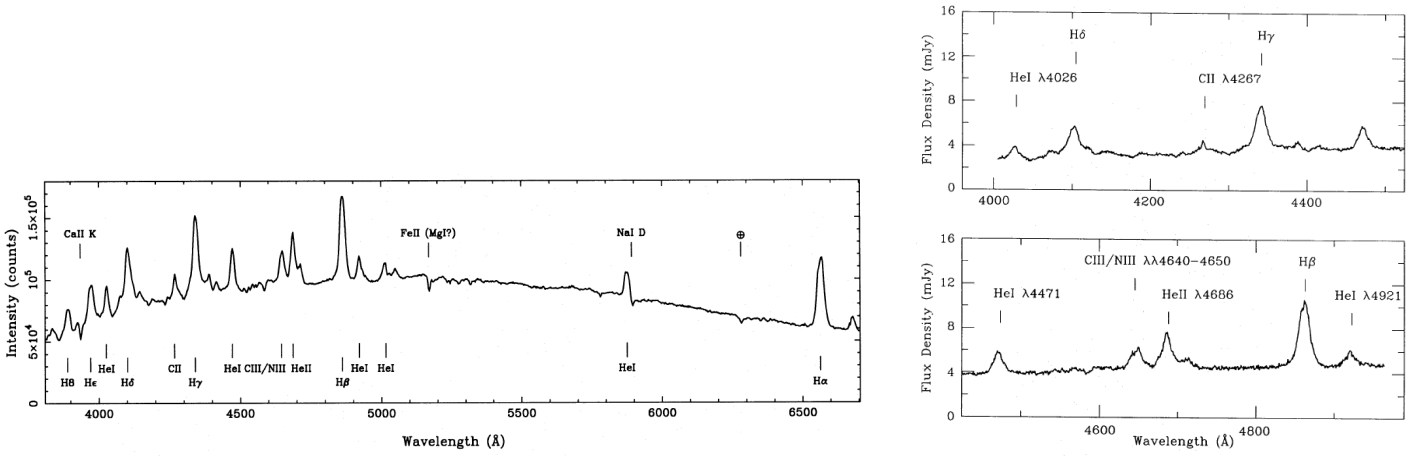
spectral resolution

The *spectral resolution* or *spectral resolving power*, R, of a spectrograph is defined as the ability to distinguish between two wavelengths separated by a small amount $\Delta\lambda$. Spectral resolution is usually quoted either in terms of $\Delta\lambda$ (usually in units of nm or Å) or in terms of the dimensionless quantity:

$R = \lambda / \Delta\lambda$.

As a rough guide, spectrographs with $R < 1000$ are regarded as *low resolution* and they generally do not allow the spectral lines from astronomical sources to be resolved. Spectrographs with $1000 < R < 10{,}000$ are regarded as *intermediate resolution*, and these do enable the study of the broadest spectral lines. However, only *high-resolution* spectrographs, with $R > 10{,}000$, enable the narrow spectral lines emitted by most stars to be studied in detail. Note that, in comparison, broad-band photometry has an effective spectral resolution of $R \sim 5$. Examples of low- and intermediate-resolution spectra of the same star are shown in figure 93.

figure 93: Low resolution (left) and intermediate resolution (right) spectra of the cataclysmic variable star V1315 Aql.



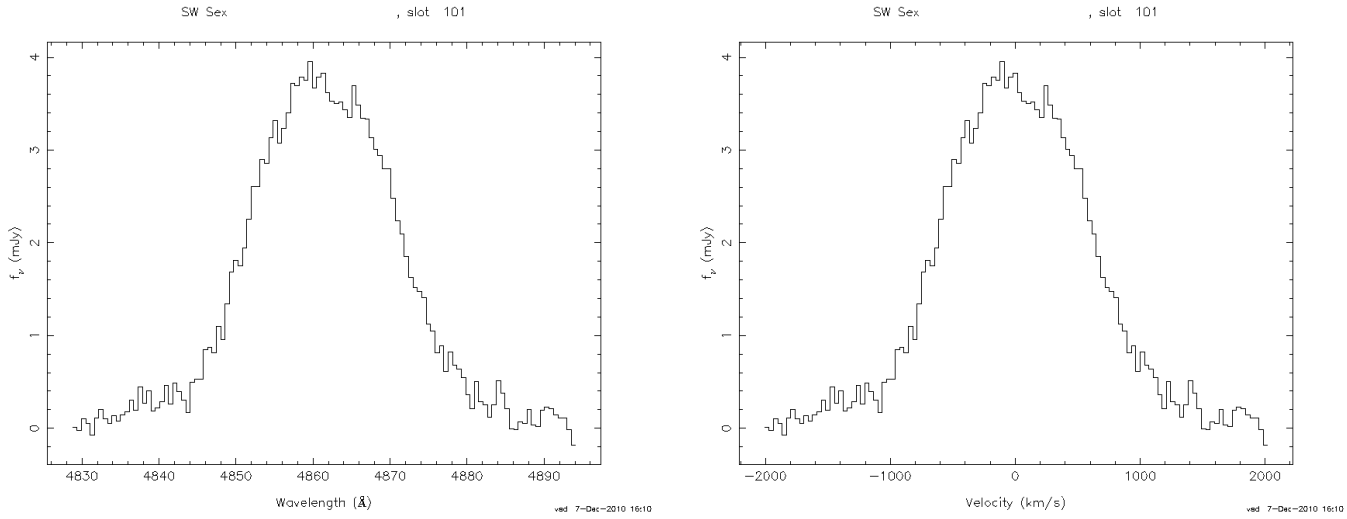
The inverse of the spectral resolution is equal to the expression describing the non-relativistic Doppler shift:

$1 / R = \Delta\lambda / \lambda = v / c$,

where v is the radial velocity of the source and c is the speed of light. Hence a spectral resolution of $R = 10{,}000$ would enable a wavelength shift of $\Delta\lambda = 0.0001 \times \lambda$ to be measured, which implies that only Doppler shifts larger than $v = 0.0001c = 30$ km/s

could be measured. The above formula also implies that it is possible to plot spectra on both wavelength and velocity scales, as shown in figure 94.

figure 94: Spectra of the Hβ emission line in the cataclysmic variable star SW Sex, plotted on a wavelength scale (left) and a velocity scale (right).



In a similar way that there is a limit to the *spatial* resolving power of a telescope, there is a limit to the *spectral* resolving power of a spectrograph. Diffraction by the grooves in the grating form a diffraction pattern, as shown in the upper panel of figure 89, which in cross section would look similar to that shown in the right-hand panel of figure 4. Adopting Rayleigh's criterion, two spectral lines would be said to be just resolved when the maximum of the diffraction pattern of one line falls on the first minimum of the diffraction pattern of the other. The *diffraction-limited spectral resolution* (or simply *limiting resolution*) of a spectrograph is then given by,

$R = Nn$,

where N is the total number of lines used across the grating. It can be seen that a higher spectral resolution can be obtained by increasing the order of the spectrum or increasing the total number of rulings in a grating, which for a fixed-sized grating implies increasing the ruling frequency. For example, a grating may have 300 lines/mm, in which case a 20 mm diameter grating used in first order would have a diffraction-limited spectral resolution of $R = Nn$ = 300 x 20 x 1 = 6,000, but working in the second order with a 600 lines/mm grating would give $R = Nn$ = 600 x 20 x 2 = 24,000, i.e. four times the spectral resolution.

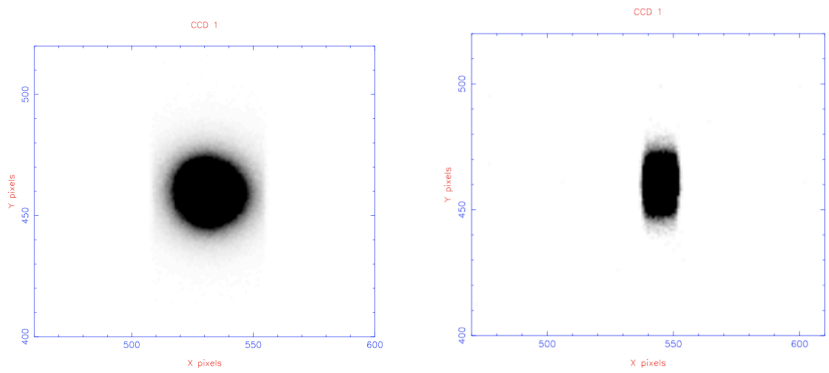
In reality, however, in the same way that the *spatial* resolution of a telescope is limited by the seeing, pixel size and/or quality of the telescope optics, not by diffraction, the *spectral* resolution of a spectrograph is limited by the slit width, detector sampling and/or spectrograph optics, not by diffraction, as discussed below:

- **slit width:** We have seen that it is useful to think of a spectrum as essentially an infinite number of images of the slit, each shifted slightly in wavelength. If the

projected slit width at the detector is much less than the limiting resolution of the spectrograph, then no loss of resolution will result. However, if the slit is widened so that the width of its image in the spectrum rises above the limiting resolution, then it is the width of the slit that will define the resolution of the spectrum, not diffraction. This point is illustrated in figures 89 and 92 (bottom panel), showing spectral resolution limited by diffraction and the slit width, respectively.

To maximise spectral resolution, therefore, it seems obvious that the slit width should always be kept smaller than the limiting resolution. Unfortunately, this is rarely possible, as the slit would be so narrow compared to the seeing disc of the star that very little light would pass into the spectrograph. This trade-off between spectral resolution and throughput is illustrated in figure 95, which shows images of the same star passing through a wide slit and a narrow slit. In the wide slit case, all of the light is able to pass into the spectrograph, but the resolution would be defined by the seeing (and image motion due to guiding errors), not the slit. In the narrow slit case, the resolution of the spectrograph would be defined by the width of the slit and would more closely approach the limiting resolution, but only a fraction of the light from the star would pass through the slit.

figure 95: Image of a star observed through a 5" wide slit (left) and a 1.5" slit (right) with ULTRASPEC.



- **detector sampling:** The size of the pixels in the CCD detector also has an influence on the effective resolution of a spectrograph. The Nyquist sampling theorem tells us that the sampling frequency should be greater than twice the highest frequency contained in the signal. In the context of spectrographs, this means that there must be at least two CCD pixels per spectral resolution element, as indicated by figure 74, otherwise it is the size of the CCD pixels that will define the resolution of the spectrograph, not diffraction or the slit width.

- **spectrograph optics:** The quality of the spectrograph optics, i.e. the collimator, grating and camera, can also degrade the resolution of a spectrograph from the limiting case by introducing optical aberrations. High quality optics are therefore essential to maximize the spectral resolution.

Ideally, then, with perfect optics, the projected slit width would be smaller than the limiting resolution of the spectrograph, and the detector pixel size would be less than

half the limiting resolution of the spectrograph. In practice, however, the limiting resolution is rarely achieved for astronomical spectrographs, and the spectral resolution is defined by the slit, which is usually matched in width to the seeing and projects to two detector pixels.

It is important not to confuse dispersion and spectral resolution. If the reciprocal linear dispersion is 1 Å/pixel, this does not mean that the spectral resolution is $\Delta\lambda$ = 1 Å. As discussed above, most well-designed spectrographs will have a slit width in the telescope focal plane that is approximately equal to the seeing, and this slit width will project to two pixels on the CCD detector. In this case, the spectral resolution would be 2 Å, i.e. the spectral resolution is generally twice the reciprocal linear dispersion. Some example calculations involving the concepts of dispersion and resolution are given in the example problems.

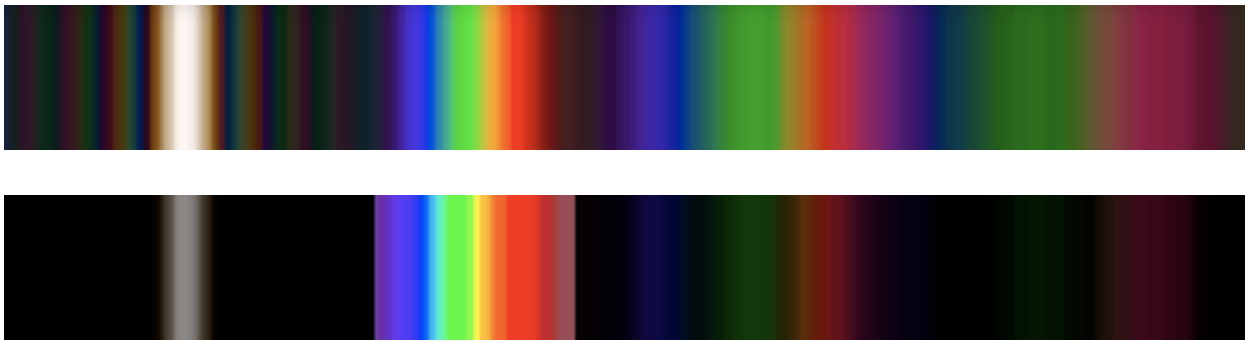
---

blazes and grisms

blazes

The fact that diffraction gratings produce multiple spectra, or orders, each with a different dispersion, makes them very versatile but also very inefficient. Most of the light is directed to the zeroth order, which is undispersed and hence useless for spectroscopy. Only ~10% of the light is directed to the first order, which is the most commonly used order in astronomical spectrographs. Some means of directing all of the light into the order of interest would therefore be highly desirable, as illustrated in figure 96.

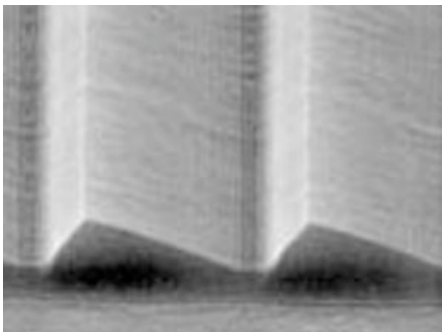
figure 96: Top: the diffraction pattern produced by a white light source incident on an unblazed (top) and blazed (bottom) diffraction grating, where only the orders n = 0, +1, +2, +3 are shown. Note how the spectra get fainter as the order increases in the unblazed case, and how most of the light is directed into the order of interest (n = +1) in the blazed case.

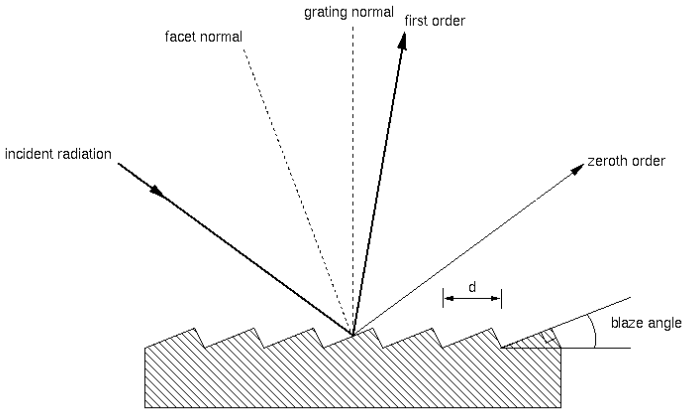


The solution is to employ the fact that gratings are most efficient when the rays emerge from the grating as if by direct reflection. By creating tilted facets, or a *blaze*, in the grating, as shown in figure 97, it is possible to direct the majority of the light (~70%) into the order of interest.

figure 97: Left: a microscopic view of the tilted facets in a typical blazed diffraction grating. Right: an illustration of the saw-tooth profile

of a blazed reflection grating. The tilt of the facet ensures that most of the light is diffracted to the first order rather than the zeroth order.





Having covered gratings, dispersion, resolution and blaze, you should now be able to understand a typical table from a spectrograph user manual, such as that shown in table 3. The first column gives the grating name, where "R" refers to a diamond ruled grating, "H" to a holographic-etched grating, "1200" (for example) to the number of grooves per mm and "B" to the fact that the grating is optimised for blue light. The second column gives the blaze wavelength of the grating in Å, i.e. the wavelength of peak grating efficiency. The third and fourth columns give the dispersion of the grating in units of Å/mm and Å/pixel, where each detector pixel is 13.5 μm in size. The fifth column gives the wavelength range of the resulting spectrum, obtained by multiplying the dispersion in Å/pixel by the number of CCD pixels in the dispersion direction (4096). Only 3500 of these 4096 pixels are usable, so the sixth column lists the usable wavelength range. The final two columns list the slit widths in arcseconds that must be used to obtain a projected slit width at the detector of 54 μm (i.e. 4 pixels) and 27 μm (i.e. 2 pixels). The latter case corresponds to sampling at the Nyquist critical frequency, and the spectral resolution is then given by twice the dispersion. Using a slit which projects to less than 2 pixels will not improve the spectral resolution, as the slit will then be undersampled by the CCD pixels. The 54 μm projected slit width is well matched to the typical seeing on La Palma, resulting in good throughput but oversampling by the CCD pixels; the spectral resolution in this case will then be defined by the slit, not by the pixels.

table 3: Details of the gratings available in the blue arm of the ISIS spectrograph on the 4.2 m William Herschel Telescope on La Palma.

ISIS wavelength coverage and resolution with EEV12							
Grating	**Blaze**	**Dispersion (Å/mm)**	**Dispersion (Å/pix)**	**Total Spectral range (Å)**	**Unvignetted range (3500 pixels)**	**Slit-width for 54 mu at detector (in arcsecs)**	**Slit-width for 27 mu at detector (in arcsecs)**
R158B	3600	120	1.62	6635	5670	0.8	0.4
R300B	4000	64	0.86	3539	3024	0.8	0.4
R600B	3900	33	0.45	1825	1560	0.9	0.45
R1200B	4000	17	0.23	940	803	1.1	0.55
H2400B	Holo	8	0.11	442	378	1.2	0.6

grisms

removed from course

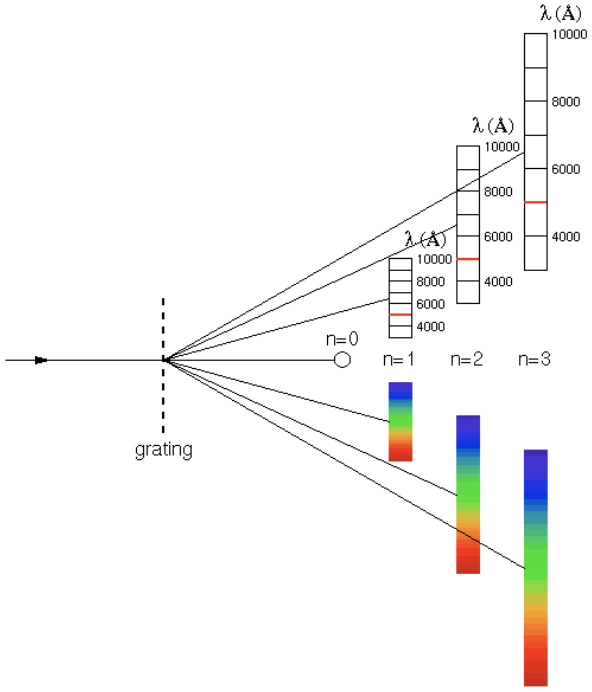
---

free spectral range and order sorting

free spectral range

The grating equation, $n\,\lambda = d \sin\theta$, tells us that light of wavelength λ in the first order is diffracted by exactly the same angle θ as light of wavelength λ/2 in the second order, as well as λ/3 in the third order, etc. In practice, this means that different spectral orders will overlap, as illustrated in figure 98.

figure 98: Schematic showing the overlap between the first, second and third spectral orders produced by a diffraction grating. The wavelength range in each spectrum is limited by the CCD detector, which is only sensitive to light in the range ~3000-10,000 Å. The horizontal red lines mark the cut-point of the commonly-used *GG495* order-sorting filter.



For example, figure 98 shows that light of wavelength 10,000 Å in the first order will fall on the same location on the detector as light of wavelength 5000 Å in the second order and 3333 Å in the third order, etc. The amount of overlap between the orders gets worse as the order number increases. Assuming that each order contains the same, fixed range of wavelengths, as depicted in figure 98, then the amount of spectrum in a given order that does not overlap with the next order up is known as the *free spectral range*. We can derive an expression for the free spectral range in a given order by noting that two wavelengths in adjacent orders, λ_1 and λ_2, that fall on top of each other must satisfy the relation $n\,\lambda_1 = (n + 1)\,\lambda_2$. Setting λ_2 as the minimum wavelength present in each order (e.g. 3000 Å in figure 98), then the free

spectral range, *FSR*, is given by:

$FSR = \lambda_1 - \lambda_2 = \lambda_2 / n$.

Hence, for order $n = 1$ in figure 98, $FSR = 3000 / 1 = 3000$ Å, and for order $n = 2$, $FSR = 3000 / 2 = 1500$ Å.

order sorting

The first-order spectrum of an astronomical source will be contaminated by second-order light. If the grating being used is blazed to the first order, the contaminating second-order spectrum is likely to be very weak, but it can still be problematic if observing a blue object in the red end of the first order. For example, if one is interested in the spectral range 6000-10,000 Å in the first order, this region will be contaminated by light from 3000-5000 Å in the second order.

One way of eliminating the second-order contamination is to use an *order-sorting filter*. This is a filter which obscures all light below a certain wavelength and transmits everything above it. For example, a common order-sorting filter is *GG495*, which transmits all light above 4950 Å and blocks everything below it. This filter would be inserted in the collimated beam of the spectrograph, thus leaving the spectrograph focus unchanged, and will prevent all light below 4950 Å from hitting the grating (see figure 98). Since, in the above example, one is only interested in the range 6000-10,000 Å, this is not a problem, and it has the advantage that the contaminating second-order light between 3000-5000 Å is almost entirely eliminated.

The free spectral range is largest in the first order, making order-sorting relatively straightforward. This is the reason why spectrographs designed to work in the first order are so popular. However, if higher spectral resolution is required, it is much better to work in higher orders. In this case, the use of order-sorting filters is impractical, as the reduced free spectral range means that too much of the spectrum would have to be blocked. An alternative approach to order sorting is therefore required, a subject we shall turn to now.

---

echelle spectrographs

We have seen that the diffraction-limited resolution of a spectrograph is given by $R = Nn = Wn / d$, where N is the total number of grooves in the grating, n the order of the spectrum, W the width of the grating and d the groove spacing. Hence, to increase R we can do one of three things:

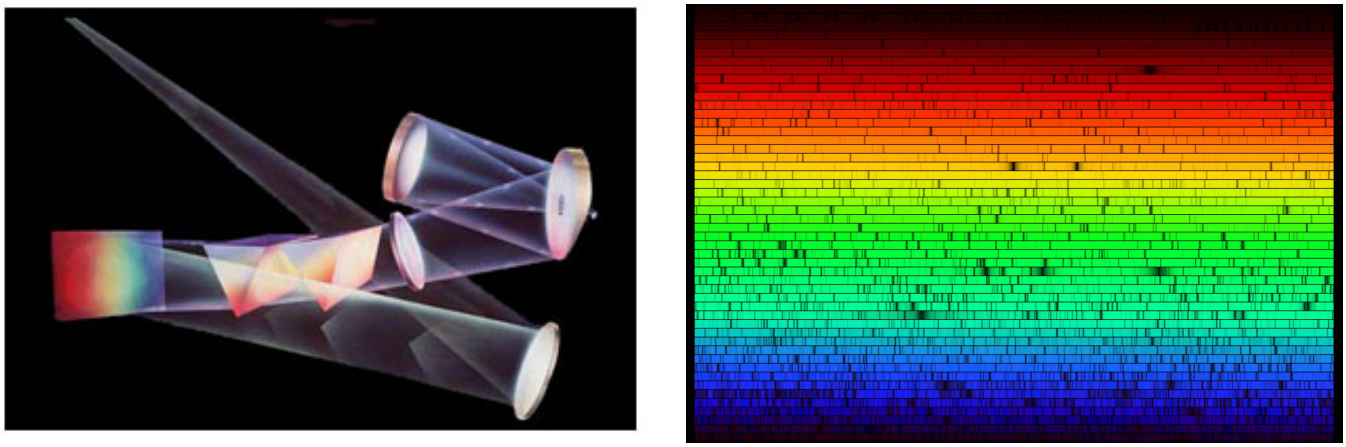
1. Increase W. However, making large gratings is difficult and expensive, and the camera and collimator must then be at least as large as the grating, significantly increasing the size and cost of the whole spectrograph. The largest diffraction gratings in the world are of widths ~ 1 m.
2. Decrease d. However, there is a limit to the maximum ruling frequency that can be manufactured - approximately 2400 lines/mm.
3. Increase n. It is relatively straightforward to design a spectrograph to work with a high-order spectrum, but the small free spectral range then results in severely overlapping orders.

If high-resolution spectroscopy is required, the easiest approach is to increase n which, from the grating equation, implies the use of increased angles of incidence and diffraction. A grating optimised for high-order work is known as an *echelle*, which comes from the French word for stairs (which describes the blazed profile of the grating). Echelle gratings do not need particularly high ruling frequencies as they operate in high orders. Typical echelles have 20-200 lines/mm and are blazed so that the maximum grating efficiency is in the range n = 10-100.

Simply inserting an echelle in place of a standard grating in a spectrograph, however, would result in multiply overlapping orders. The free spectral range is very small at high orders, hence order sorting filters are of limited use. The solution is to use an additional dispersing element, known as a *cross disperser*. The cross disperser, which can be either a low dispersion grating or prism, is usually mounted after the echelle grating with its dispersion axis perpendicular to it, as shown in the left-hand panel figure 99. This means that the overlapping orders are separated in the spatial direction on the detector, as shown in the right-hand panel of figure

99. Although each order has only a very narrow wavelength range, there are so many of them, each covering slightly different wavelengths, that very wide spectral coverage is achieved. To prevent the orders overlapping in the spatial direction, only a very short slit must be used. *Echelle spectrographs* (or *high-resolution spectrographs*) can deliver resolutions in excess of R = 100,000, and are most famously employed in the detection of extrasolar planets by measuring the tiny Doppler wobble of the host stars.

figure 99: Left: schematic of the light path through the Hamilton Echelle Spectrograph on the 3 m Shane Telescope at Lick Observatory, California. Right: cross-dispersed Solar spectrum covering almost the entire optical range, obtained using the HIRES echelle spectrograph on the 10 m Keck telescope, Hawaii. These orders are relatively straight - some echelle spectrographs produce tilted and curved orders, making data reduction quite challenging.



example problems

1. **A grating is used in first order, with the spectrum observed at an angle of 15^{o} with respect to the grating face. If the spectrograph has a camera of focal length 300 mm, how many lines per mm are required to give a reciprocal linear dispersion of 20 Å/mm?**

   The reciprocal linear dispersion is given by the formula:

   $d\lambda / dx = d \cos\theta / n F_{cam}$.

   Rearranging for d gives

   $d = (d\lambda / dx) . (n F_{cam} / \cos\theta) = 20 \times 1 \times 300 / \cos(15^{o}) = 6212$ Å $= 6.212 \times 10^{-7}$ m, i.e. 1610 lines/mm.

   Note how the units of d are given in Å, as the reciprocal linear dispersion is given in Å/mm and the camera focal length in mm.

2. **A grating of width 50 mm and 300 lines/mm is used to produce a second-order spectrum. What is the limiting spectral resolution in Å at a wavelength of 550 nm?**

   The limiting spectral resolution is given by the formula:

   $R = Nn = \lambda / \Delta\lambda$.

   We require $\Delta\lambda$:

   $\Delta\lambda = \lambda / Nn = (550 \times 10^{-9}) / (50 \times 300 \times 2) = 1.8 \times 10^{-11}$ m $= 0.18$ Å.

3. **A spectrograph has a reciprocal linear dispersion of 66 Å/mm**

and a diffraction-limited spectral resolution of 3.3 Å. What are the maximum projected slit width and detector pixel size required to exploit this resolution?

The projected slit width that matches the diffraction-limited spectral resolution = 3.3 / 66 = 0.05 mm = 50 μm, which is the maximum permissable size.

Nyquist sampling theory tells us that at least two detector pixels are required across one resolution element to optimally sample the spectrum, so the pixel size must be less than 25 μm.

4. **A spectrograph is being designed for a 2 m *f*/10 telescope, which is located at a site where the typical seeing is 1". What ratio of camera to collimator focal lengths would you recommend using if you have a detector with a pixel size of 25 μm?**

   If the typical seeing is 1", then it would be best to use a slit width of approximately this size so as to obtain reasonable throughput whilst still ensuring that the resolution of the spectrograph is defined by the slit.

   For optimal sampling, the 1" slit should project to at least 2 pixels on the detector. The platescale of the telescope is 206265 / 20000 = 10.3 "/mm, so the physical width of the slit is approximately 100 μm, which must project to 2 x 25 = 50 μm. Hence the magnification of the spectrograph must be 0.5, implying that the ratio of camera to collimator focal lengths must be 0.5.

5. **What changes would you have to make to the spectrograph in question 4 above if you wanted to use it on an 8 m *f*/10 telescope?**

   The collimator and camera in a spectrograph work as a re-imager, changing the focal length (and hence platescale) delivered by the telescope to a different value at the detector. The focal length of the telescope plus collimator/camera system, F_{sys}, is given by:

$F_{sys} = F_{tel} M$,

where M is the magnification, given by $M = F_{cam} / F_{coll}$. Using this equation, we derived the following relation in example problem 2 in imagers:

$F_{sys} = D_{tel} f_{cam}$.

F_{sys} determines the platescale at the detector, i.e. the number of arcseconds per pixel. A well designed spectrograph will project a slit width equal to the seeing to two pixels on the detector. Assuming the 8 m and 2 m telescopes we are considering experience the same seeing, and both use CCDs with the same pixel size, this means that F_{sys} for the two telescopes must be equal, i.e.:

$F_{sys8m} = F_{sys2m} = D_{tel8m} f_{cam8m} = D_{tel2m} f_{cam2m} = 8 f_{cam8m} = 2 f_{cam2m}$.

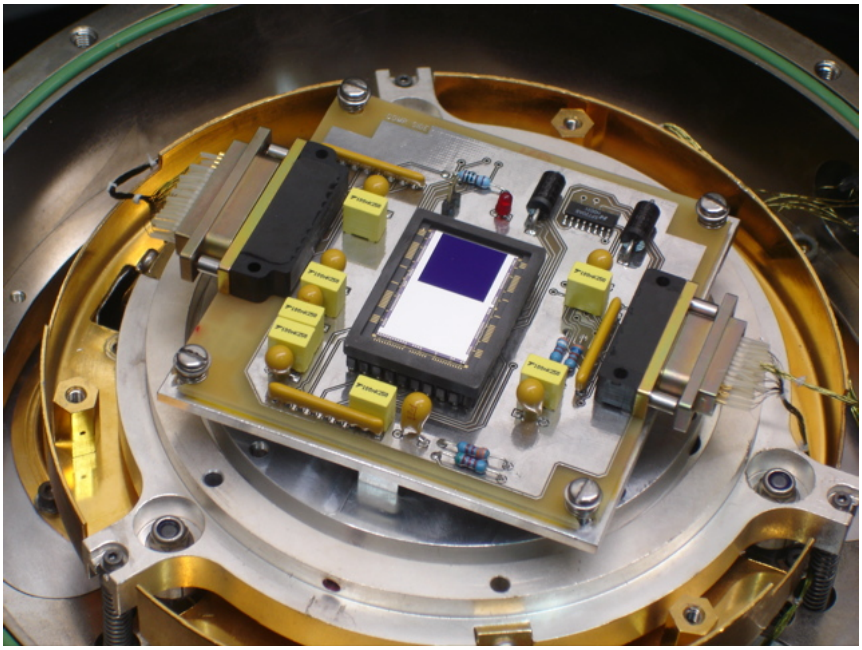
Hence:

$f_{cam8m} = f_{cam2m} / 4$.

This means that the focal ratio of the spectrograph camera must be reduced by a factor of 4 when moving the spectrograph from a 2 m telescope to an 8 m telescope. Since $f_{cam8m} = F_{cam8m} / D_{cam8m}$, this can be achieved either by decreasing the focal length of the camera by a factor of 4 or increasing the diameter of the camera by a factor of 4. The former is not an option as changing the focal length of the camera would change the magnification and hence the size of the projected slit on the detector. Hence the only option is to increase the diameter of the camera or, strictly speaking, the light beam entering the camera, by a factor of 4. This means that the diameter of the grating and collimator must also increase by a factor of 4, since all lie in the same collimated beam. Hence spectrographs on the world's largest telescopes are enormous and very expensive instruments.

---

astronomical techniques



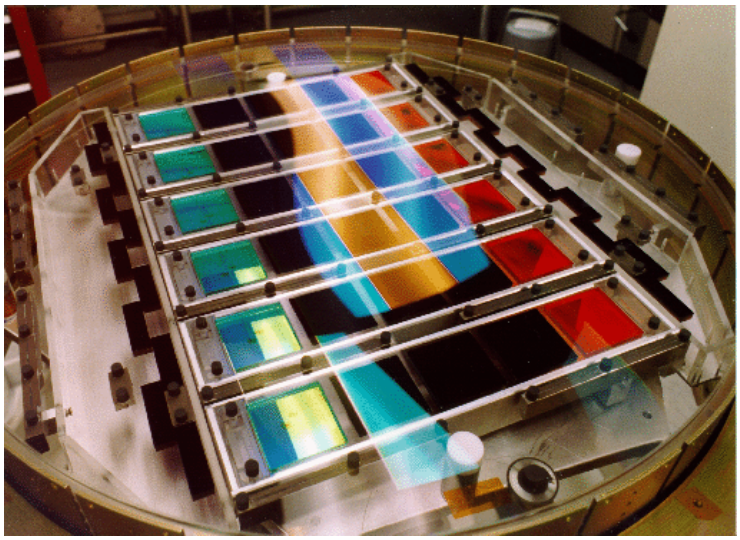
detectors

detectors: introduction

An astronomical detector is a device, typically located in the focal plane of a telescope or instrument, that has the ability to record the photons incident upon it. We have seen that for imaging, photometry and spectroscopy, detectors composed of a two-dimensional array of picture elements, or *pixels*, are essential for efficient operation. Furthermore, detectors with the ability to record as many of the incident photons as possible are highly desirable - wasting photons is a cardinal sin in astronomy, given how faint astronomical sources are and how much money and effort goes into building bigger telescopes.

Towards the end of 1969, Willard Boyle and George Smith, whilst trying to develop a video phone at Bell Labs in the USA, invented the *charge-coupled device* or *CCD*, a discovery for which they were awarded a share of the 2009 Nobel Prize in Physics. CCDs are almost perfect astronomical detectors. They are multi-element, small, linear, stable, low-power, low-cost devices with excellent sensitivity over a wide wavelength range. Astronomers pioneered the use of CCDs in the 1970s and nowadays you would find it extremely difficult to find any other type of detector in use at a major telescope. For this reason, we shall concentrate exclusively on CCDs in this part of the course.

detectors



II. CCDs

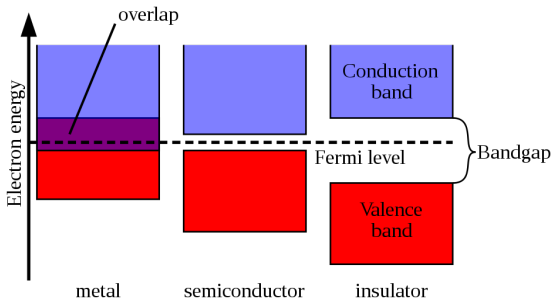
the physics of semi-conductors

One way of detecting light is to harness the photo-electric effect, where electrons are emitted from matter as a result of the absorption of a photon. To make a useful detector, however, some means of storing the electrons at the location where they were emitted, and then counting how many have been emitted in a given time, is required. The best way of doing this is to use a semi-conductor, so in this section we shall review the physics of this material and how it can be used to store and control the motion of electrons.

metals, semi-conductors and insulators

The energy levels available to bound electrons in an isolated atom are confined to certain discrete values. When isolated atoms are brought together to form a solid, they interact and their energy levels are split into a large number of closely spaced levels, creating a series of *bands*, as shown in figure 100.

figure 100: Schematic showing the valence and conduction bands in a metal, semi-conductor and insulator.

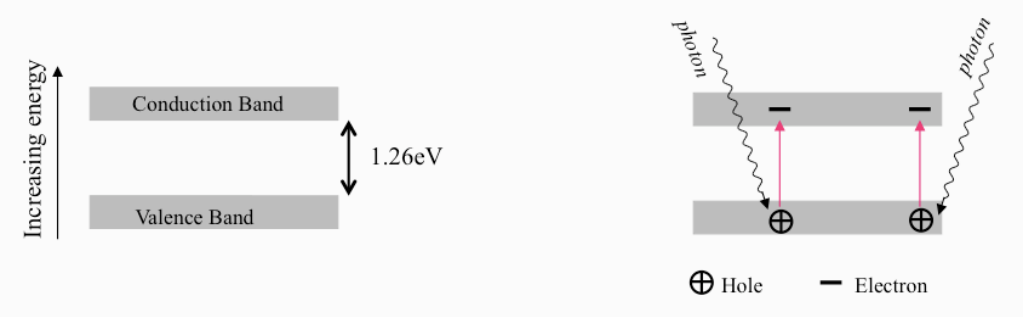


The highest energy levels interact the most and hence their bands are the widest. The highest energy band that is filled with electrons is known as the *valence band*. The band above this, composed of even higher energy, unoccupied levels, is called the *conduction band*. In an *insulator*, there is a wide gap between the valence and conduction bands, the *band gap*, which electrons are forbidden to occupy. The electrons in the completely filled valence band cannot move in response to an electric field because every nearby orbit in the valence band is occupied, and they do not have sufficient energy to move up into the conduction band. In a *metal*, the

valence and conduction bands overlap, i.e. there is no band gap. In this case, the valence-band electrons are free to move throughout the solid in response to the force of an applied electric field, e.g. as produced when both ends of the metal are attached to a battery.

In a *semi-conductor*, the band gap is small. This means that it is sometimes possible for an electron to be promoted from the valence band to the conduction band by absorbing energy, either from the thermal motion of the atoms in the solid, or from an incoming photon, as shown in figure 101. Once an electron is in the conduction band, it can move and hence conduct electricity. The promoted electron leaves behind a *hole* in the valence band, and this too can conduct electricity, as it offers a place for neighbouring electrons to occupy. Since there are fewer charge carriers (electrons in the conduction band, holes in the valence band) in a semi-conductor than in a metal, semi-conductors are poorer conductors than metals, but better than insulators.

figure 101: Schematic showing how a photon of energy greater than 1.26 eV (the band gap of silicon) is able to promote an electron to the conduction band, leaving behind a hole in the valence band.



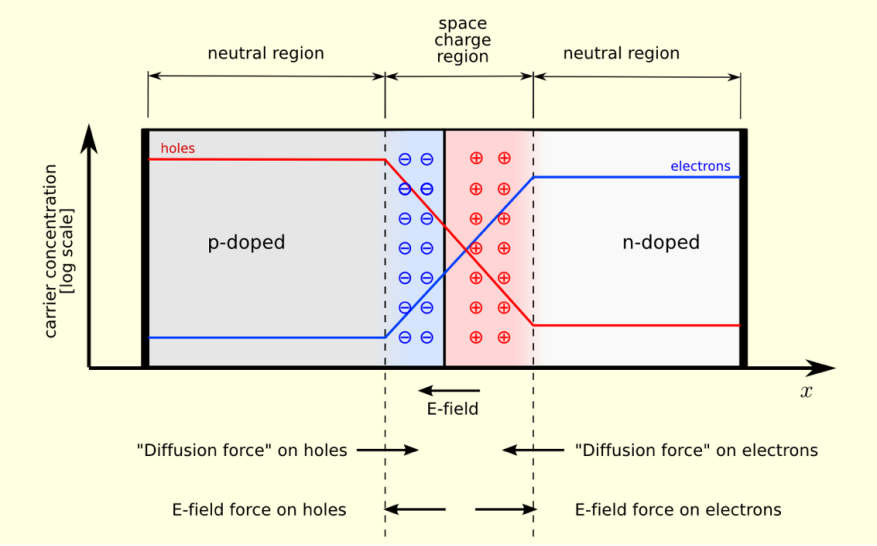
The semi-conductor of choice for CCDs is silicon, as it has a relatively large band gap, and hence is less likely to suffer from random elevation of electrons to the conduction band via thermal excitation, although this effect can be alleviated by cooling. Electrons that are elevated to the conduction band via the absorption of a photon are known as *photo-electrons*. The likelihood of absorbing a photon in silicon is a function of the photon energy, with red (low-energy) photons passing deeper into the silicon before being absorbed. A point comes when the reddest light does not have enough energy to overcome the band gap and promote a valence electron into the conduction band. For silicon, this occurs at a wavelength of approximately 1100 nm.

Unfortunately, once a photo-electron has been created in a silicon crystal, it is free to migrate through the crystalline structure, away from the point of origin, and recombine with a hole. Hence some means of separating the photo-electrons from the holes and storing them close to the point of origin is required. The best way of doing this is to use a p-n junction, as described below.

p-n junctions

The properties of a semi-conductor can be altered by deliberately adding small amounts of impurities, a process known as *doping*. Doped silicon in which the impurity has more valence electrons than undoped silicon will donate these *n*egatively-charged electrons to the conduction band and is hence known as *n type*. Conversely, doped silicon with a deficit of valence electrons compared to undoped silicon will leave a *p*ositively charged hole in the valence band ready to accept any electrons, and is hence known as *p type*. If an n-type and p-type semi-conductor are brought together, a *p-n junction* is formed. Electrons from the *n* region will diffuse to the *p* region to fill up some of the holes in the valence band, thus making the *p* region more negative than it was. Similarly, the diffusion of holes from the *p* region to the *n* region will make the *n* region more positively charged than it was. Therefore, a narrow zone, the *depletion region* or *space charge region*, either side of the junction will form where the majority charge carriers are depleted relative to their concentrations well away from the junction, as shown in figure 102. Eventually, a potential barrier is created by the build up of electrons and holes on the *p* and *n* sides, respectively, repelling further diffusion.

figure 102: Schematic showing a p-n junction. The concentration of electrons and holes are shown by the blue and red lines, respectively. The grey regions are neutral, the red zone is positively charged and the blue zone is negatively charged. The directions of the electric field, the electrostatic force on the electrons and holes, and the direction in which the diffusion tends to move electrons and holes, are shown at the bottom.



The concentration of holes buried in the n-type region will have a more positive

electrical potential than elsewhere in the crystal, whereas the surplus of electrons in the p-type region will make it more negative. This means that any electron-hole pairs created near the junction by incoming photons will be swept apart and the electrons will be stored in the region of greatest positive potential in the n-type layer, which therefore acts as a charge storage capacitor.

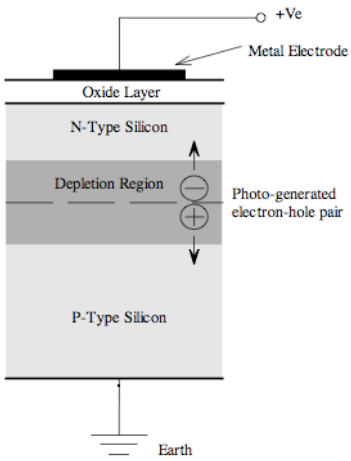
It possible to control the height of the potential barrier and the width of the depletion region by applying a voltage across the junction, which in turn controls the capacity of the p-n junction to store charge. If a positive voltage is applied to the *p* side of the p-n junction, it will tend to attract more electrons across the junction and decrease the potential barrier (the *forward bias* condition). More usefully for CCDs, a positive voltage applied to the *n* side would increase the potential barrier and the width of the depletion region (the *reverse bias* condition).

---

the structure of CCDs

Having looked at the physics of semi-conductors, we are now in a position to understand the structure of CCDs. There are two types, *buried channel* and *surface channel*. The surface channel device is the simplest CCD structure, but suffers from poor charge-handling performance as the electrons are stored and transfered close to the surface of the crystal, where the many irregularities and defects present cause charge traps. Surface-channel CCDs will not be considered further here. Instead, almost all scientific CCDs are buried-channel devices, which employ a p-n junction to bury the depletion region well below the silicon surface. A schematic of a buried-channel CCD pixel is shown in figure 103.

figure 103: Schematic cross-section through a CCD pixel.



The bottom layer is a semi-conductor substrate, composed of a p-type silicon layer and an n-type silicon layer, i.e. a p-n junction. Over this is grown a thin layer of silicon dioxide, which is an insulator. On top of this is placed a metal *electrode* or *gate*, made of a transparent material known as polysilicon. The entire structure is known as a *MOS* capacitor, as it is composed of layers of, from top-to-bottom, *M*etal, *O*xide and *S*emi-

conductor.

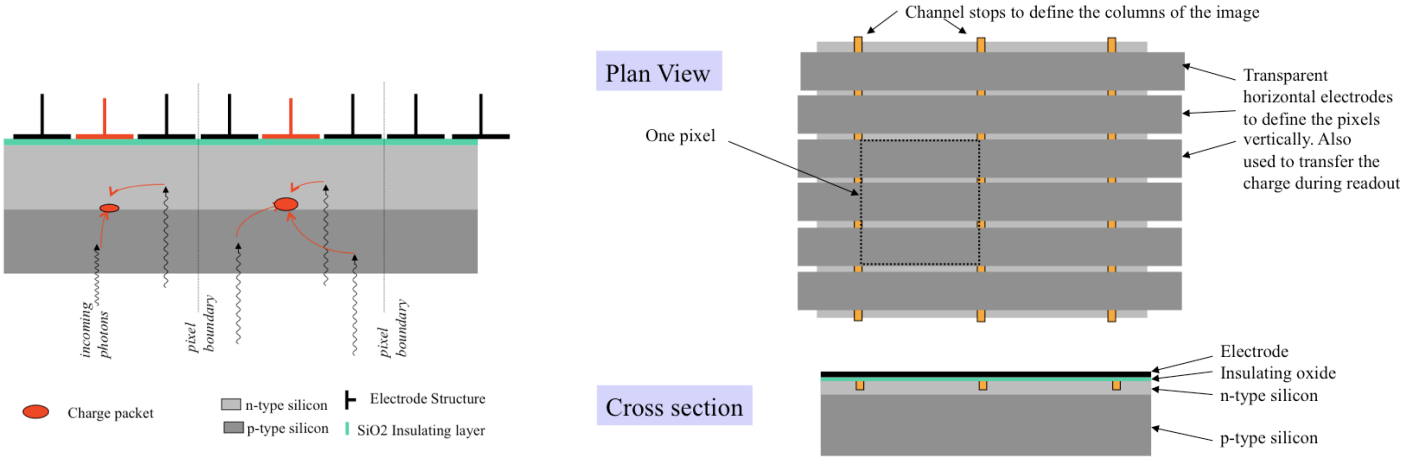
By grounding the p-type silicon substrate and applying a positive voltage to the n-type layer, the p-n junction is reverse biased, widening the depletion region and creating a potential well in the n-type silicon that confines the electrons in the vertical direction in figure 103 (for more detail, see figures 6.12a,b of McLean, where it should be noted that in actual fact a separate positive voltage is applied to the electrode and the n-type silicon). Incident light generates electron-hole pairs in the depletion region, and the electrons migrate upwards into the n-type silicon layer where they are trapped in the vertical direction by the potential well. The electrons are therefore able to build up, with the amount of negative charge directly proportional to the level of incident light. The depth of the electron-collection layer (or buried channel) in the silicon can be controlled by altering the positive voltage on the electrode at the surface.

Note that it is important for the depletion region to permeate throughout as much of the thickness of the silicon as possible. This is because the depletion region is devoid of electrons, and hence any electrons that then appear in the potential well in the n-type silicon can safely be assumed to have been produced by incident photons. A thicker depletion region also ensures that there are no field-free regions in the silicon, as any photons striking such regions would be able to diffuse in a random manner away from the point of generation and either recombine with a hole and so not be counted, or be collected by the wrong pixel.

To create an array of CCD pixels, some means of isolating the charge packets held under each pixel is required. To prevent the collected charge from moving left and right in figure 103, a strip of insulator is implanted on either side, in structures known as *channel stops*. In the other dimension, the charge packet is held in place by splitting the electrode into a triplet, with the central electrode held at a more positive potential than the two on either side. Together, the channel stops and the electrode triplet define the pixel, as shown in figure 104.

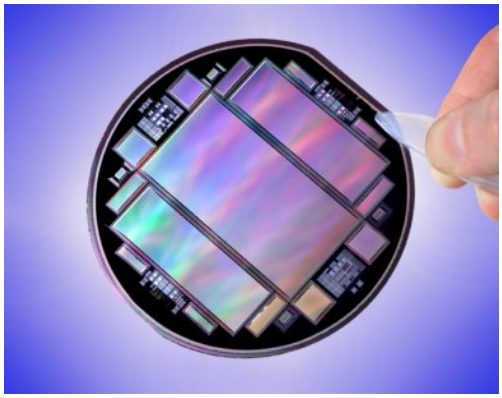
figure 104: Left: schematic cross-section through three pixels of a CCD, showing how the charge packets in each pixel are confined in the left-right direction by setting the central electrode (red) of a triplet to a higher positive voltage. Right: schematic of a small portion of a CCD. The combination of channel stops and triplet electrode structure is used to confine the charge

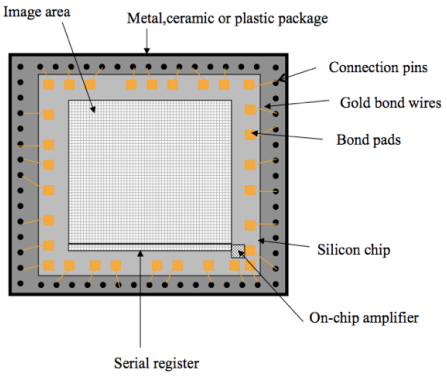
packets in each pixel in both the *x* and *y* directions.

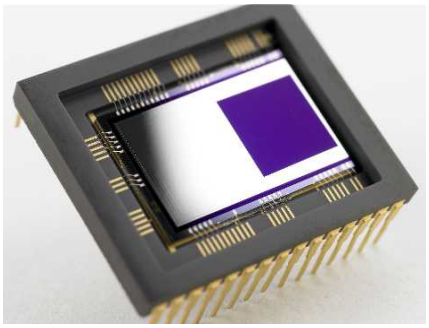


CCDs are manufactured on silicon wafers using the same photo-lithographic techniques used to manufacture computer chips. Scientific CCDs are very big - only a few can be fitted onto a wafer, as shown in the left-hand panel of figure 105. This is the main reason why CCDs are so costly - large-format, science-grade CCDs cost many tens of thousands of pounds. Electrical connections are made to the outside world via a series of bond pads and thin gold wires positioned around the chip periphery, as shown in the central and right-hand panel of figure 105.

figure 105: Left: photograph of a 6" silicon wafer with three large CCDs and assorted smaller devices. Centre: schematic showing how CCDs are packaged. Right: photograph showing an E2V CCD97 package.

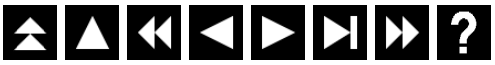






---

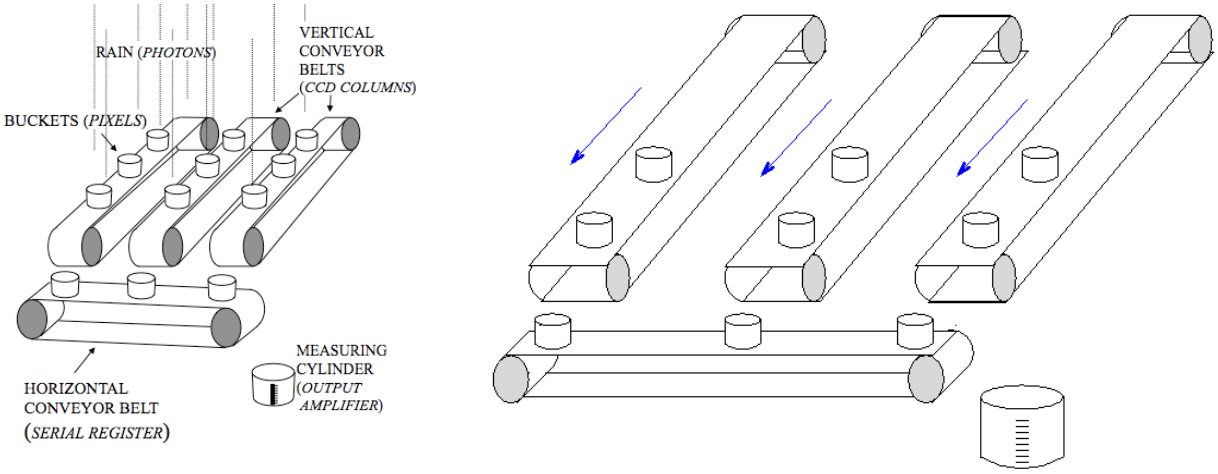
charge coupling



The unique feature of a CCD, which gives it its name, is the way in which the photo-electrons are extracted from the storage site, a process known as *charge coupling*.

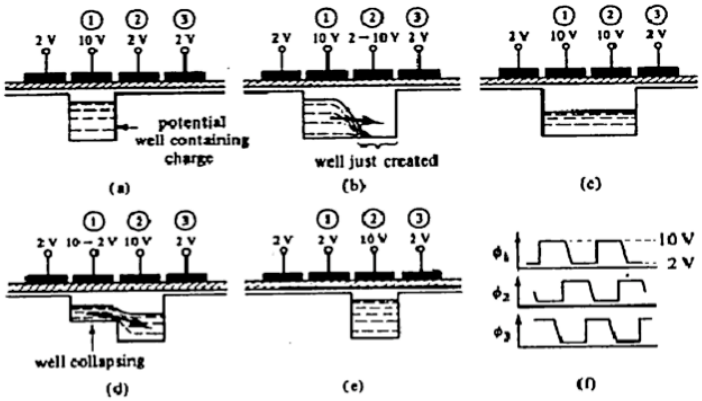
Once the CCD has been exposed to light for the required amount of time, each pixel will contain a charge packet of size proportional to the number of incident photons. To measure this charge, it is necessary to move these charge packets, one by one, off the chip. The easiest way of visualizing this process is by thinking of taking a CCD image as analogous to measuring the rain falling on a field, as shown in figure 106. The first step is to distribute a large number of buckets in a rectangular pattern of rows and columns over the field - these are the pixels. After it has stopped raining, you can measure the amount of water collected in each bucket, i.e. the amount of charge in each pixel, by shifting the entire array of buckets, one row at a time, on parallel conveyor belts towards a perpendicular conveyor belt at one end of the field. When a single row has been transferred onto the conveyor belt at the end of the field, the row is shifted, one bucket at a time, towards a measuring point at the end of this conveyor belt, where the amount of water in each bucket is recorded. Once the whole row has been measured, the next row is shifted onto the conveyor belt at the end of the field, and the process is repeated until every bucket in the field has been measured. By plotting the amount of rain in each bucket as a two-dimensional grey-scale image, where white represents a full bucket and black an empty one, it is possible to visualise the pattern of rainfall over the field. Replacing the rain in this analogy by photons, this is how an image of the sky can be recorded.

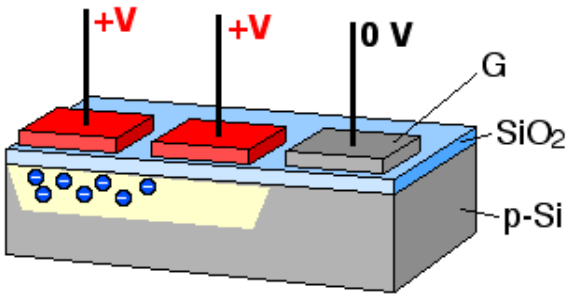
figure 106: schematic (left) and animation (right; reload page to restart) showing the analogy between constructing an astronomical image with a CCD and measuring the rainfall over a field - see text for details.



So how are the conveyor belts implemented electronically? We saw in figure 104 that the pixels in most CCDs are defined by three electrodes. This so-called *three-phase* structure can be exploited to move charge. The voltage in an electrode adjacent to the electrode holding the charge packet is raised to the same level. This allows the charge to flow, like water, and be shared between the two electrodes. Decreasing the voltage of the original electrode then completes the transfer, pushing all of the charge across to the adjacent electrode, which is held at the higher voltage level. Since there are three electrodes in each pixel, three of the above transfers are required to move the charge packet by one pixel. The process of raising and lowering the voltages to move charge is known as *clocking*, and is illustrated in figure 107. The clock pulses can be described by a *timing waveform*, also shown in figure 107, and are straightforward to implement electronically.

figure 107: schematic (left) and animation (right) showing how charge is clocked from one pixel to the next by modulating the voltage on adjacent electrodes.





The process of transferring charge from one electrode to the next is not perfect. The *charge transfer efficiency*, CTE, is defined by the equation:

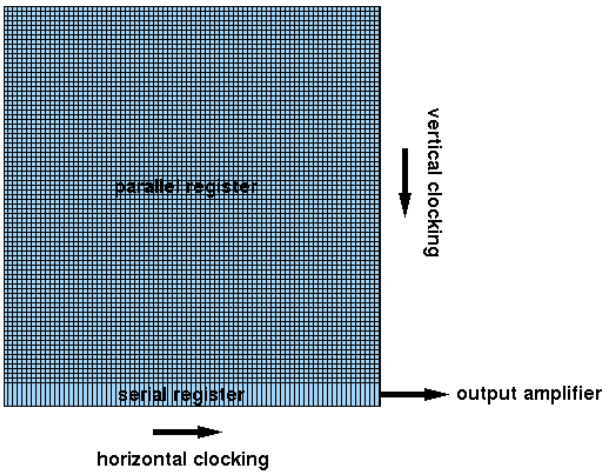
$$CTE = 1 - [(N_0 - N_t) / N_0],$$

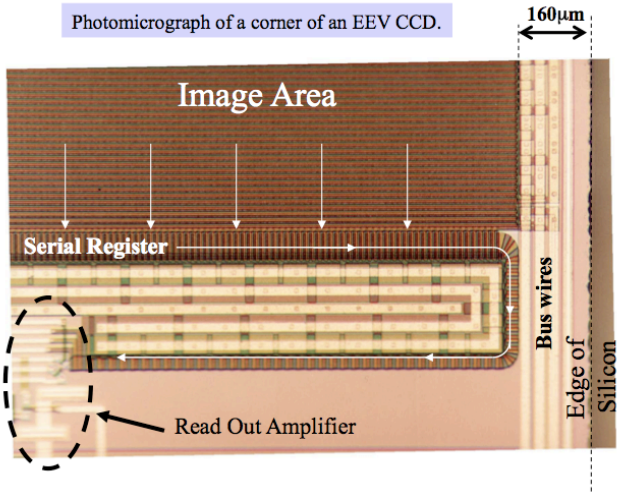
where N_0 is the number of charges under an electrode and N_t is the number transferred to the next electrode. It is essential to have extremely high values of CTE in an astronomical CCD. For example, if there are 1000 e^- under an electrode and the CTE is 99%, then after 100 transfers only $1000 \times (0.99)^{100} = 366$ e^- = 36.6% of the original charge will have been transferred! The net effect is to trail the image of a CCD in the direction of the charge transfer. In practice, CCDs with CTEs greater than 99.999% are required, as demonstrated in the example problems.

Figure 108 shows the structure of a typical CCD. The electrodes in the main image area of the CCD, known as the *parallel register*, are arranged so that they move the charge vertically downwards (a process known as *vertical clocking*). At the bottom of the parallel register is a single row, the *serial register*, in which the electrodes are arranged perpendicularly to those in the parallel register, so that they move the charge horizontally (a process known as *horizontal clocking*). At the end of the serial register is an *output amplifier*.

figure 108: Left: the structure of a CCD, indicating the parallel register, serial register and output amplifier. Right: photomicrograph of the corner of a CCD, with arrows showing the direction in which the charge is shifted. In this particular CCD, the serial register is bent double to move the output amplifier away

from the edge of the chip, which is useful if the CCD is to be butted against another in a mosaic.





A complete clocking sequence, known as a *read out*, therefore consists of the following steps:

1. A vertical shift of the entire parallel register by one pixel. This delivers a row of charge to the serial register.
2. A horizontal shift of all the pixels in the serial register. This delivers each charge in that row to the output amplifier, one pixel at a time.
3. Another vertical transfer. This moves the next row in the image into the serial register.
4. Another horizontal transfer, etc.

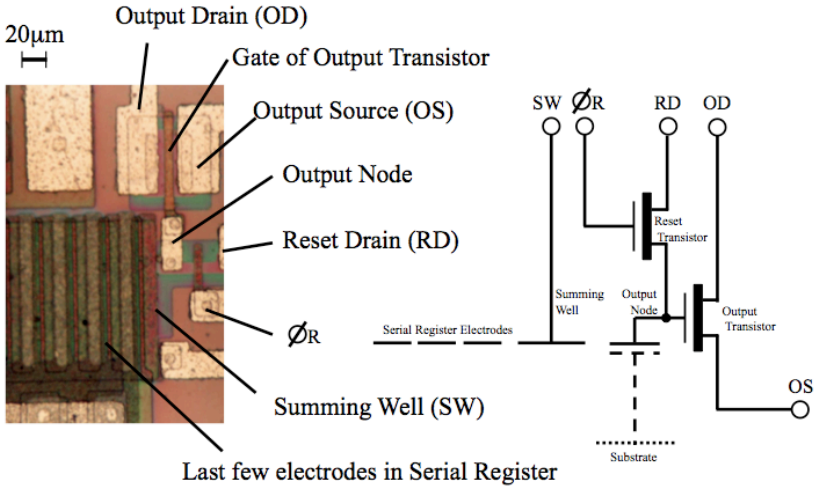
The above process is repeated until all of the pixels in the image area have been delivered to the output amplifier, where they are detected, measured and recorded. We shall turn now to look at how this is done.

---

output electronics

A detailed view of the output amplifier of a CCD is shown in figure 109. As each packet of electrons of total charge *Q* leaves the last pixel in the serial register, it is passed into a capacitor of capacitance *C*. This causes a small change in the voltage, *V*, across the capacitor, given by $V = Q / C$. This voltage change is first amplified and then measured. To measure the voltage, it is necessary define a *reset* level, i.e. the voltage across the capacitor in the absence of the charge contained in the CCD pixel. A technique known as *correlated double sampling (CDS)* is then used to measure the difference between the reset level voltage and the final voltage across the capacitor containing the charge packet. The faster the CDS is performed, the less accurate the measurement of the voltage. Finally, the voltage produced by each charge packet is digitized using an *analogue-to-digital* converter, producing the number of *analogue-to-digital units (ADUs)*, or simply *counts*, in each pixel, which are then written to a computer disk.

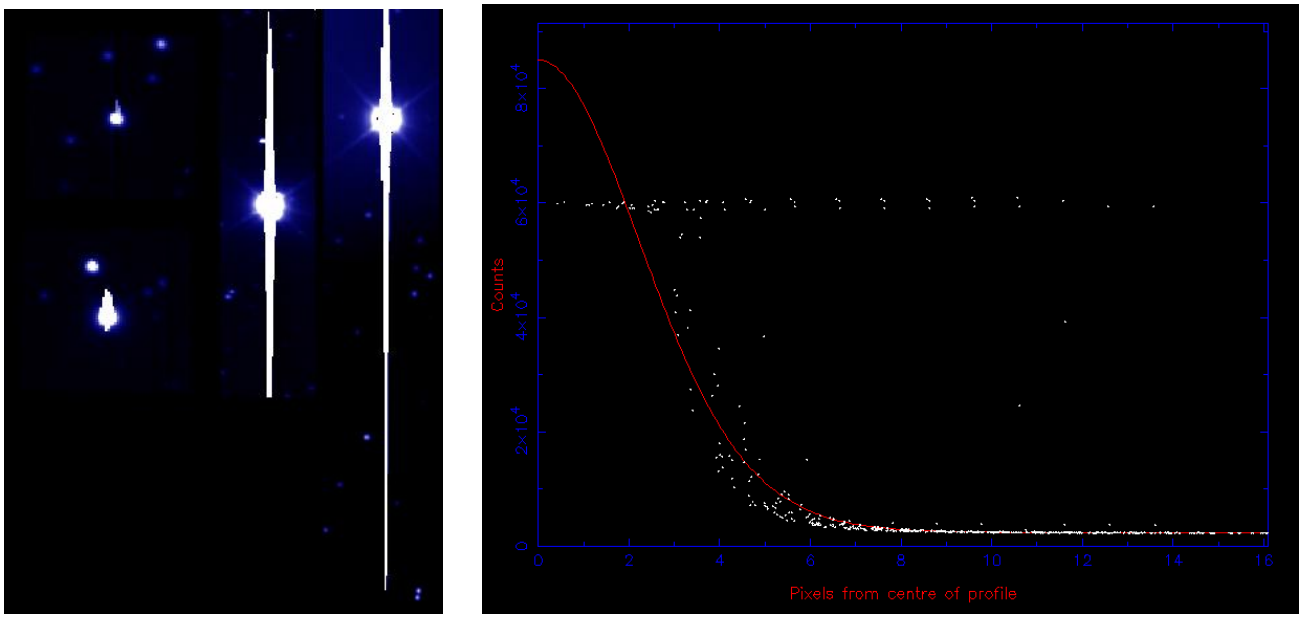
figure 109: Photomicrograph of the output amplifier of a CCD and its circuit diagram.



There are a number of noise sources in the output electronics and the charge-measurement process that result in an uncertainty in the measurement of the charge contained in each pixel. This is known as *readout noise*, which can be reduced to a level of ~3 e^- per pixel in a well-designed and well-optimized CCD.

Each CCD pixel has a maximum charge carrying ability, known as the *full-well capacity*. Typically, the full-well capacity of a pixel is hundreds of thousands of electrons. If this is exceeded, no further electrons can be detected by that pixel and it is said to be *saturated*. The channel stops prevent any excess charge from spilling horizontally into adjacent pixels, so the electrons spill vertically into adjacent pixels, creating streaks in saturated images, known as *blooming*, as shown in figure 110. At the other end of the scale, the readout noise defines the lowest number of electrons that can be recorded. The ratio of the full-well capacity to the readout noise is known as the *dynamic range* of the CCD, and is typically of order 100,000:1.

figure 110: Left: three CCD images showing the effects of saturation on a star. The level of saturation increases from left to right, evident as an increased level of blooming. Right: the flat-topped point-spread-function of a saturated star.



Most high-quality CCDs have 16-bit analogue-to-digital converters (ADCs). This means that the ADC can divide a specified voltage into 2^{16}-1 = 65535

parts. For example, if the amplified voltage across the capacitor is 10 volts, voltage intervals as small as 152.6 μV can be measured, and each such interval would represent 1 count. Matching the amplified voltage to the resolution of the ADC is crucial. This is controlled by a parameter known as the *gain*, often given in units of e^-/ADU. For the low signal levels typical in astronomy, it is important that the gain is set to a value such that the readout noise is optimally sampled. For example, if the readout noise of a CCD is 2 e^-, and the gain is set to 4 e^-/ADU, then it will be impossible to distinguish between pixels containing 1 or 2 units of readout noise, effectively increasing the readout noise. This so-called *quantization noise* is like a rounding error, but it can be rendered negligible by choosing an appropriate gain value: in the above example, a gain of approximately 1 e^-/ADU should be used.

One consequence of using low gains to reduce quantization noise, however, is that the dynamic range of the CCD is defined not by the full-well capacity of a pixel but by the resolution of the ADC. For example, a gain of 1.5 e^-/ADU would mean that a 16-bit ADC is capable of counting 1.5 x 65535 = 98302 e^- before saturating. If the full-well capacity of the pixel is 200,000 e^-, this means that the ADC would saturate well before the full-well capacity of the pixel is reached, i.e. the maximum number of electrons that can be counted is 98302, not 200,000. Another example is given in the example problems.

Note that the error due to quantization noise is given by the expression N_{fwc} / (2^n . $12^{0.5}$), where N_{fwc} is the *effective* full-well capacity of the pixel in electrons (i.e. the maximum number of electrons that can be counted by the ADC without saturating) and n is the number of bits in the ADC. For a well optimised CCD, N_{fwc} / $2^n \cong g$, where g is the gain in e^-/ADU, and hence the quantization noise is given by g / $12^{0.5}$. Quantization noise must be added in quadrature with other CCD noise sources to give the total noise. For example, in a CCD with g = 1 e^-/ADU and readout noise of 4 e^-, the total noise becomes 4.01 e^-, so quantization noise is unimportant. But in a CCD with g = 8 e^-/ADU and readout noise of 4 e^-, the total noise becomes 4.6 e^-.

---

improving performance

Great strides have been made in recent years in enhancing the performance of CCDs. In this section, we shall briefly look at how four of the most important characteristics of CCDs can be improved: dark current, quantum efficiency, readout speed and readout noise.

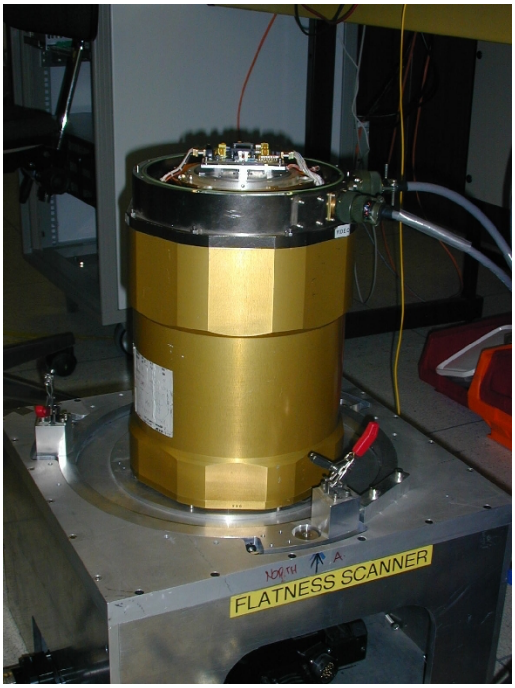
dark current

Valence electrons in a CCD can be promoted into the conduction band by absorbing energy, either from the random thermal (i.e. heat-generated) motion of the atoms in the silicon lattice, or from an incoming photon. The former mechanism is known as *dark current* and the electrons produced by it are indistinguishable from photo-electrons.

Dark current can be very substantial. At room temperature, the dark current of a standard CCD is typically 100,000 e^-/pixel/s, which is sufficient to saturate most CCDs in only a few seconds. One way round this is to take very short exposures, but astronomical sources are usually too faint for this to be possible without paying a heavy penalty in readout noise.

The solution is straightforward - cool the CCD. Dark current decreases rapidly with decreasing temperature. The typical operating temperatures of CCDs are in the range 150 to 263 K (i.e. -123 to -10^{o}C). At major observatories, most CCDs are cooled to the bottom end of this range, generally using liquid nitrogen. The resulting dark current can be as low as a few electrons per pixel per hour. To prevent the liquid nitrogen from having to be continuously replenished as it boils off, CCDs are usually mounted in *cryostats*, as shown in figure 111. Cryostats employ a vacuum and radiation shield to minimize conductive and radiative heating by the surroundings.

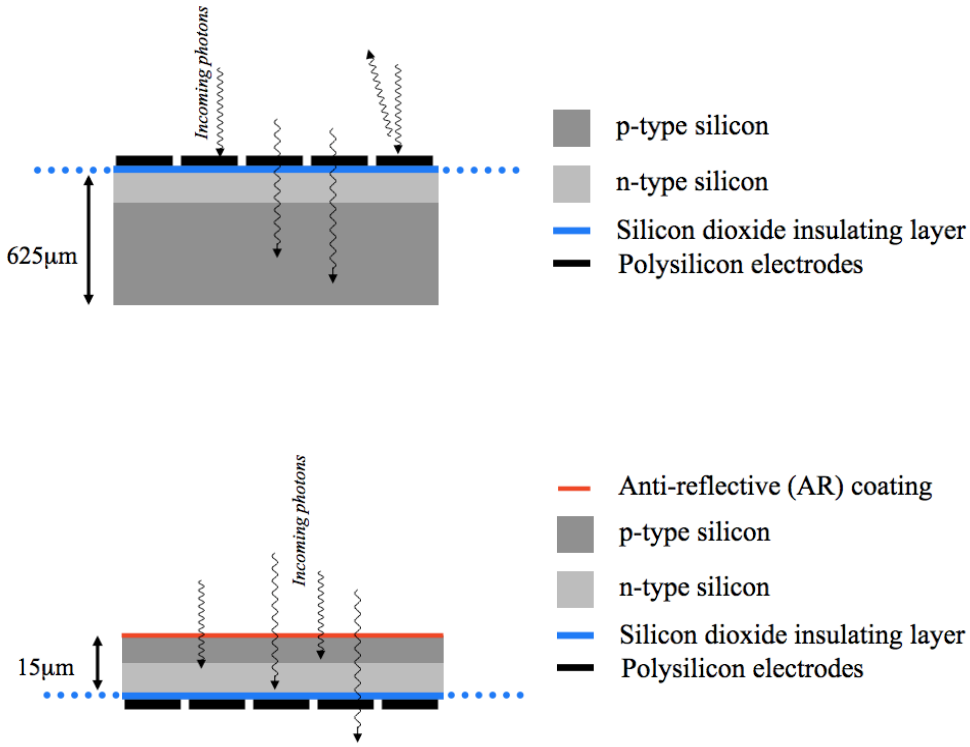
figure 111: Photograph of the gold-coloured ULTRASPEC cryostat in the lab, with the lid open. The CCD mounted on its circuit board can be seen at the top.



quantum efficiency

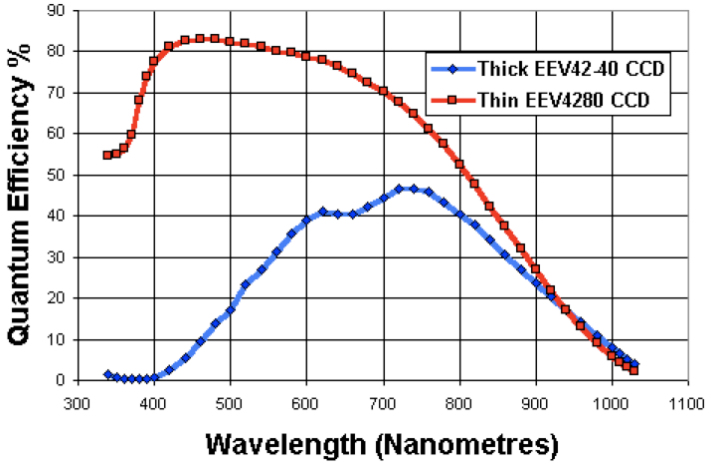
Thus far, we have been considering CCD structures which involve the photon passing through the electrode structure to get to the depletion region in the silicon substrate beneath. Unfortunately, the electrode structure absorbs and reflects many of the incident photons, particularly at blue wavelengths, preventing them from producing electron-hole pairs in the depletion region. This arrangement is known as a *front-side illuminated* CCD, and is illustrated in the top panel of figure 112.

figure 112: Top: schematic of a thick, front-side illuminated CCD. Bottom: schematic of a thinned, back-side illuminated CCD.



The percentage of photons incident on a CCD which successfully produce electrons that are measured at the output of the CCD is known as the *quantum efficiency (QE)* of the device. The QE varies with wavelength and, for a front-side illuminated device peaks around 50%, as shown in figure 113. As well as absorbing photons, the complicated nature of the electrode structure precludes the use of an anti-reflection (AR) coating, which would otherwise boost the QE of a front-side illuminated CCD.

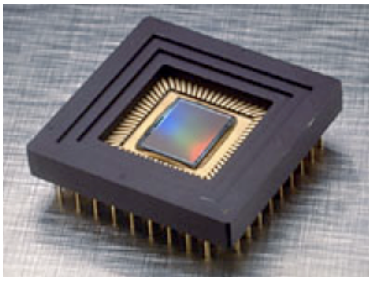
figure 113: Graph comparing the QE of a thick front-side illuminated CCD with that of a thinned, back-side illuminated CCD.

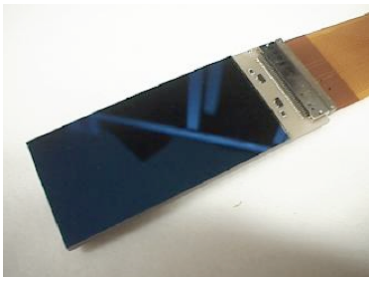


One way to significantly improve the QE of a CCD is to turn it over and illuminate it from the back side. The photons in such a *back-side illuminated* device do not have to pass through the electrodes to get to the depletion layer, eliminating the QE loss from absorption. Furthermore, the back-side of the chip can be readily AR coated, significantly reducing the QE loss by reflection, as shown in figure 114. However, to avoid absorption of photons before they get to the depletion region, the thick silicon substrate must be *thinned*, mechanically and/or chemically, to only ~15 μm, as shown in the bottom panel of figure 112. Such thinned, back-side illuminated CCDs, or *back-thinned* CCDs, have QEs of over 90% and are particularly good in the blue part of the spectrum, as shown in figure 113.

Back-thinned CCDs do have a number of disadvantages. Thinning can reduce the red response because red photons need more absorption length and if this is not there they will pass right through the silicon. Thinned CCDs are also mechanically fragile, prone to warping and expensive to manufacture compared to thick CCDs. Another problem is that the thin silicon layer produces interference fringing in the red part of the spectrum, as shown in figure 115. This is correctable to some extent at the data reduction stage, but fringing can limit the accuracy of measurements, particularly when doing spectroscopy.

figure 114: Left: photograph of a thick front-side illuminated CCD, with the electrode structure on the top. Right: photograph of a thinned, back-side illuminated CCD, which appears almost black thanks to the anti-reflection coating.

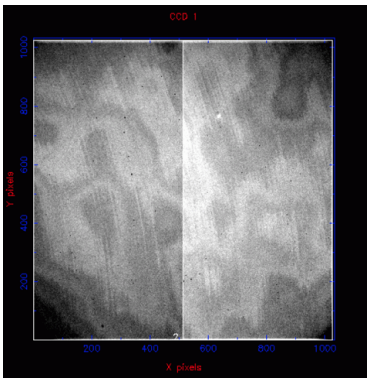


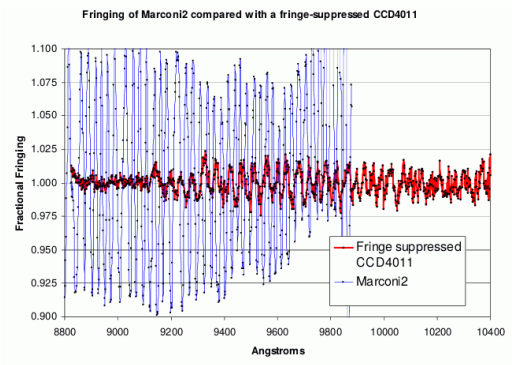


Fortunately, there is a way round the poor red-response of back-thinned CCDs, and that is by using thicker silicon (typically ~40 μm). Simply making the silicon thicker however, would mean that red photons would tend to generate charge below the depletion region. In this field-free

region, the charge would then be able to diffuse in a random manner away from the point of generation before being attracted to the depletion region of a pixel for storage. The probability of being collected in the correct pixel is therefore reduced, effectively degrading the spatial resolution of the image. This problem can be alleviated by increasing the thickness of the depletion region, which is achieved by reducing the doping concentration of the silicon, i.e. using purer, higher-resistivity silicon. Such devices are known as *deep-depletion CCDs*. As well as significantly higher QE in the red, these devices exhibit much lower levels of fringing, due to the thicker silicon altering the condition for interference and also allowing more red light to be absorbed by the silicon rather than being reflected off the top and bottom surfaces and interfering. It is possible to obtain even lower levels of fringing in deep-depletion devices by altering the thicknesses of adjacent electrodes, a process sometimes referred to as *anti-etaloning* or *fringe suppression*, thereby further breaking the interference condition.

figure 115: Left: an ULTRACAM *z'* CCD image showing fringing. Right: fringe amplitude as a function of wavelength for a deep-depletion CCD (blue) and a fringe-suppressed deep-depletion CCD (red).





Readout speed (*For Advanced Readers*)

CCDs are slow to read out due to the serial nature of the clocking and charge measurement processes. Typical dead times between CCD exposures are of order tens of seconds to minutes. Increasing the clocking and charge measurement speeds can decrease the dead time, but only at the expense of poor charge-transfer efficiency and readout noise, respectively. One way round this is to use a different CCD architecture,

known as a *frame-transfer CCD*. Further details of frame-transfer CCDs can be found in the paper on the high-speed camera ULTRACAM by Dhillon et al. (2007).

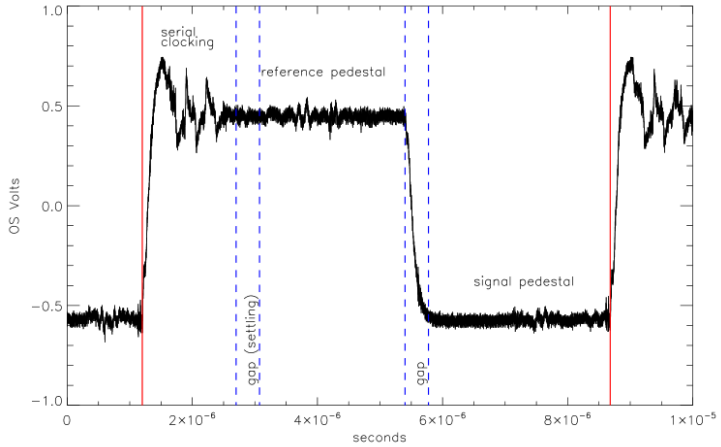
Readout noise (*For Advanced Readers*)

There are a number of ways in which readout noise can be reduced in CCDs. One way is ensure that the CCD is manufactured using low-noise amplifiers at the outputs. A second way is to use a different CCD architecture, known as an *electron-multiplying CCD* (or EMCCD). Further details of EMCCDs can be found in the paper by Tulloch and Dhillon (2011).

A third way of reducing CCD readout noise is to optimize the way in which the charge content of each pixel is measured. This measurement is performed by the *video processor* in the CCD controller. The video processor measures the size of the voltage step produced by the dumping of the charge in a pixel onto the output node of the CCD, as described earlier. Two measurements have to be made, a *reference* voltage before the charge is dumped (i.e. after the capacitor has been "reset"), and a *signal* voltage after the charge in each pixel has been dumped into it, as shown in figure 116. However, due to random thermal agitation of electrons in the reset electronics, there is uncertainty in the reset level, given by the *reset noise* = $(kTC/e)^{0.5}$ e^-, where k is Boltzmann's constant, e is the charge on the electron, T is the temperature of the capacitor and C is its capacitance. This noise is also sometimes referred to as *kTC* noise and is of order 100 e^-. Fortunately, with suitable electronic design of the CCD output, the effect of reset noise can be eliminated as the (unknown) reference voltage is "frozen" once set, which means that taking the difference between the voltage measured after reset and after the pixel charge dump eliminates the reference voltage regardless of its value. This technique is known as *correlated double sampling* (or CDS), because two samples of the CCD output voltage are taken per pixel and the offset due to the reset noise in each sample is the same, i.e. it is correlated. Two common implementations of CDS are the *dual slope integrator* (or differential averager) and *clamp and sample*, which are described in more detail by Hegyi and Burrows (1980), Hopkinson and Lumb (1982), Tulloch (2013a) and Tulloch (2013b). The dual slope integrator takes two samples of equal duration, one before and one after the pixel dump. In clamp and sample, only a single sample after the pixel dump is taken (prior to this,

the signal processing chain is clamped to ground for a brief period to establish a known reset level and then the clamp is open to let the signal go through the chain). The dual slope integrator is hence slower than clamp and sample, but results in better noise performance.

figure 116: The change in voltage at the output of a CCD as the charge in a single pixel is measured. The reference level is at 0.5 V and the signal level is at -0.5 V. The difference between the two gives the charge in the pixel. Note that this waveform represents a large signal that is close to the full-well capacity of the pixel. Typical astronomical images will have much smaller waveform amplitudes - the sensitivity of a modern CCD is approximately 8 μV/e^-. From Tulloch (2013a).



In practice, it isn't possible to measure the reference and signal voltages perfectly as both suffer from noise induced by the on-chip electronics, such as the MOSFETs in the output amplifier. The result is noise in the final output signal of the CCD, as shown in the top panel of figure 117. The power spectrum of this noise is shown in the left-hand panel of figure 118, where the *x*-axis is the frequency at which the pixels are read out by the CCD. At high frequencies, or fast pixel rates, the output signal from the CCD is dominated by *white noise* (also known as Johnson noise) from the amplifier, which is present at all frequencies and is constant with frequency. At lower frequencies, or slow pixel rates, the output of the CCD is dominated by *1/f noise* (also known as flicker noise), where the noise increases with decreasing frequency. The effect of the high-frequency white noise can be minimised by integrating the reset level and the signal level, i.e. taking an average of each level using a dual-slope integrator

instead of a single "snap-shot" measurement. Taking the difference of the average reset and signal levels using CDS also corrects for the lowest-frequency *1/f* variations that occur on timescales longer than the signal-processing time for each pixel, as given by the separation of the vertical solid lines in figure 117. However, intermediate-frequency *1/f* variations on timescales shorter than the pixel time are not corrected by the averaging, or by taking the difference between the reset and signal levels. Hence, as the integration time is increased, the readout noise decreases due to the reduction of the white noise component until a noise floor is reached of approximately 2 e^-/pixel set by the *1/f* noise, as shown in the right-hand panel of figure 118. Note that building a CCD controller that is capable of reaching the intrinsic noise floor of a modern CCD is extremely demanding.

figure 117: Simulated output of a CCD (in Volts) as a function of time, for 11 pixels. The solid, vertical red lines show the pixel boundaries. The vertical, dashed red lines show the charge dump event for each pixel. The top panel shows just the noise from the output amplifier: two main noise components are present - the short-timescale (white) noise and the longer timescale flicker (*1/f*) noise. The power spectrum of these data are shown in figure 118. The middle panel shows the addition of the reset (or *kTC*) noise. The bottom panel shows the addition of the signal. From Tulloch (2013a).

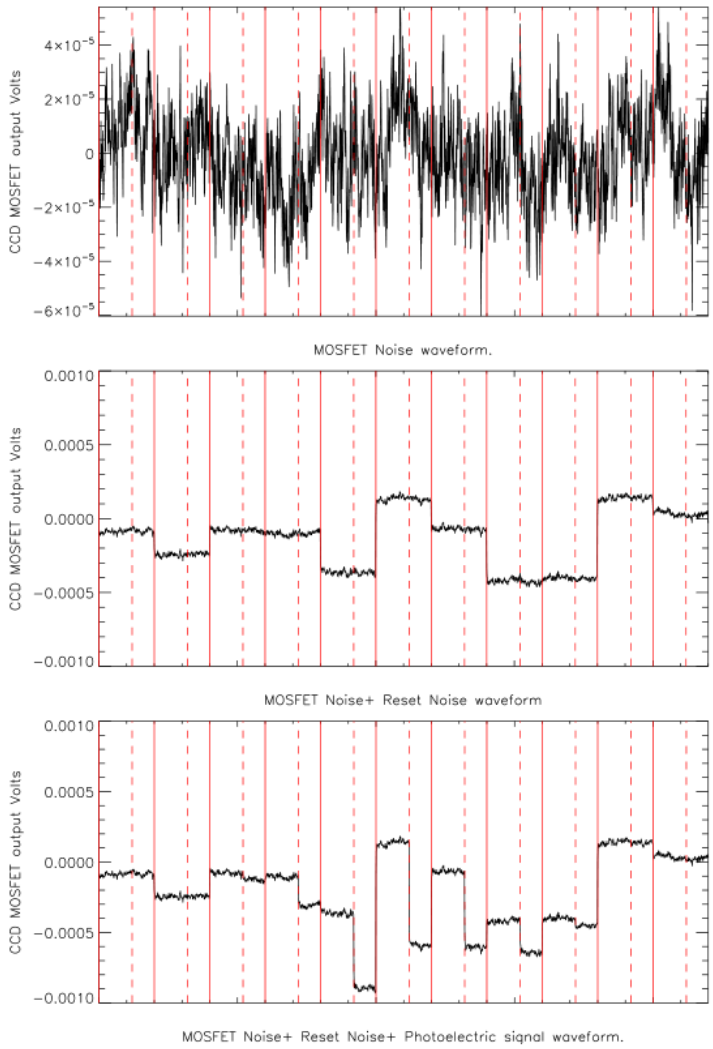
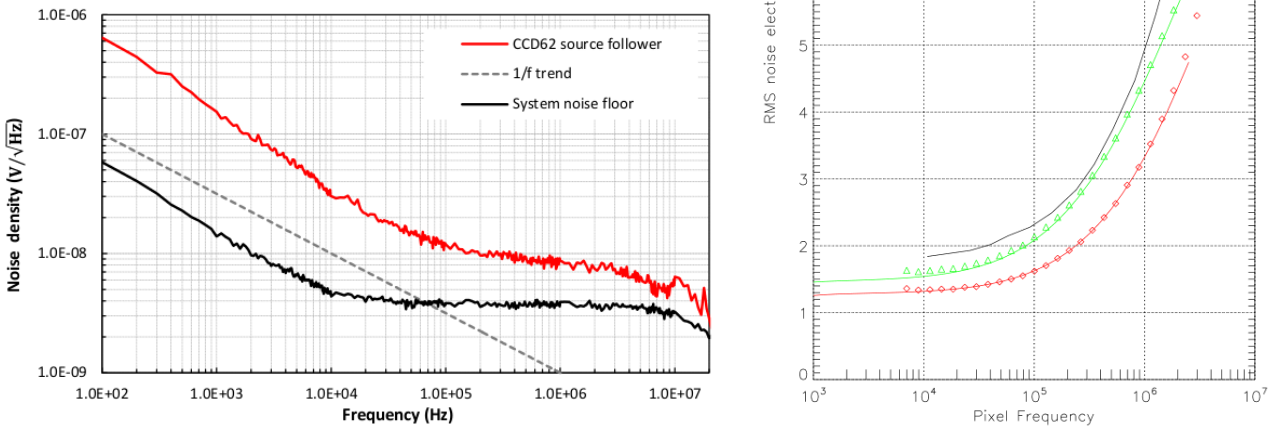


figure 118: Left: power spectrum of the output of a CCD (red curve). *1/f* noise is dominant at lower frequencies, and the curve flattens off to white noise at higher frequencies. The point at which it flattens, called the *corner frequency* (or knee), lies at around 150 kHz for this CCD, and is where the white noise is approximately equal to the *1/f* noise. Reading the CCD slower than the corner frequency will not decrease the readout noise any further. Right: the relationship between CCD readout noise and the pixel frequency. The latter is proportional to the time taken to integrate the reset and signal levels in the CDS. It can be seen that a noise floor is reached around 20 kHz for this particular CCD, due to *1/f* noise. It can also be seen that the dual-slope integrator performs better than the clamp and sample technique at all but the highest frequencies. Both techniques gives better readout noise than predicted by the CCD manufacturer (solid curve), most probably because they include the effect of

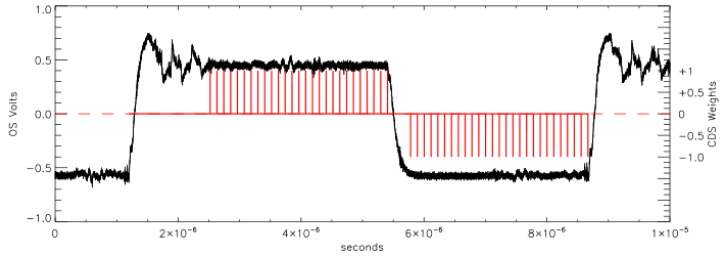
noise added by the CCD controller. From Tulloch (2013a).



Recent developments in CCD controller technology, in particular faster ADCs, now allow the use of *digital CDS* (DCDS). In this technique, rather than integrating the reset and signal levels using analogue electronics, a series of rapid "snap-shot" measurements of each level are made and digitised, as shown in figure 119. With multiple samples of each level now available, a range of averaging and weighting schemes can be adopted to maximise the signal-to-noise ratio of the CCD output. There have been claims of significantly reduced readout noise using DCDS (e.g. Cancelo et al. 2011), but Tulloch (2013a, 2013b, 2015) has shown that DCDS does not make all that much difference to the readout noise. According to Tulloch (priv. comm.), *assuming that one has a well-designed analogue CCD controller with which to compare*, the only advantages of DCDS are: 1. Cleaner/simpler CCD output electronics, e.g. no noisy switches, and; 2. Improved dynamic range due to the use of floating point arithmetic, effectively giving 1-2 more bits of ADC resolution - this means that quantization noise is not as important for high gains (e^-/ADU). See also the papers on DCDS by Alessandri et al. (2015) and Alessandri et al. (2016), which support Tulloch's conclusions.

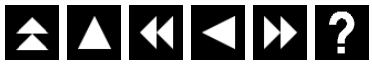
figure 119: The change in voltage at the output of a CCD as the charge in a single pixel is measured. The reference level is at 0.5 V

and the signal level is at -0.5 V. The red lines show the times that the waveform is sampled for DCDS. In this scheme, the samples all have equal weights but opposite signs (+1 in the first pulse group, -1 in the second group), so this is the digital equivalent to the analogue dual-slope integrator (differential averager). From Tulloch (2013a).



---

example problems



1. **A pixel at the centre of the imaging area of a CCD contains 1000 e^-. The CCD has 2048 x 2048 pixels and a CTE of 99.999%. How many electrons will the pixel contain at the output of the CCD?**

   Assuming that the CCD is a three-phase device, moving the charge by one pixel will require 3 transfers.

   To transfer the charge down the parallel register will therefore require 3 x 1024 = 3072 transfers.

   To transfer the charge along the serial register will require another 3 x 1024 = 3072 transfers.

   Assuming the CTE is the same in both the parallel and serial registers, the number of electrons in the pixel at the output is given by 1000 x $(0.99999)^{(3072 \text{ x } 2)}$ = 940 e^-.

   This is only 94% of the original charge. To increase this to the more acceptable value of 99% would require a CTE of 99.9999%.

2. **A CCD pixel contains 80,000 e^-. If the full-well depth of the pixel is 100,000 e^-, the gain is 1.2 e^-/ADU and a 16-bit ADC is used, will the pixel be saturated?**

   The pixel contains only 80,000 e^-, which will therefore *not* saturate the 100,000 e^- full-well depth of the pixel.

   However, on readout, the 80,000 e^- will be converted to 80,000/1.2 = 66,667 counts, which is above the 2^{16}-1 = 65535 saturation limit of the ADC. Hence the pixel will be saturated.

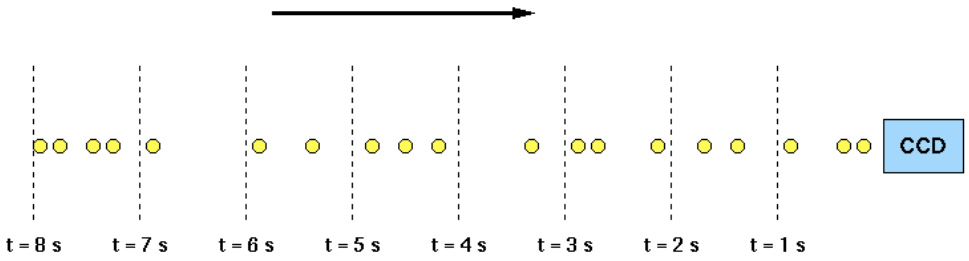
detectors

III. signal-to-noise

photon statistics

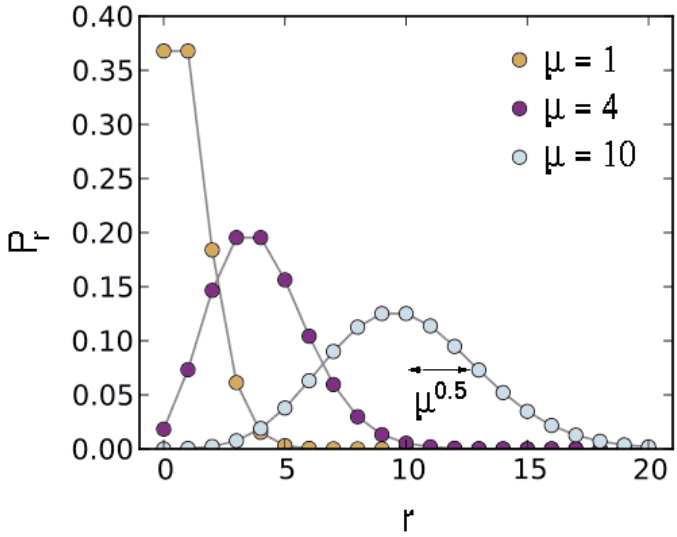
The vast majority of astronomical sources produce photons via random processes distributed over vast scales. Therefore, when these photons arrive at the Earth, they are randomly spaced, as illustrated in figure 115. The number of photons counted in a short time interval will vary, even if the long-term mean number of photons is constant. This variation is known as *shot noise*. It represents the irreducible minimum level of noise present in an astronomical observation.

figure 115: Photons from a faint, non-variable astronomical source incident on a CCD detector. Because the signal is low in this example, it is easy to see that there is a substantial variation in the number of photons detected in each 1 s time interval.



If one plots a histogram of the number of photons arriving in each time interval, the resulting distribution is known as a *Poisson distribution*, as shown in figure 116. Poisson statistics are applicable when counting independent, random events which occur, when measured over a long period of time, at a constant rate. The Poisson distribution is therefore applicable to the counting of photons from astronomical sources, the counting of photons from the sky, or the production of thermally-generated electrons in a semi-conductor (i.e. dark current).

figure 116: Poisson distributions for mean values of μ = 1, 4 and 10. P_r is the probability of observing *r* events. For each value of μ, the mean of the distribution is at μ and the standard deviation is $\mu^{0.5}$. As μ increases, the Poisson distribution approaches the Normal distribution.



An important property of the Poisson distribution is that the standard deviation is equal to the square root of the mean, as indicated in figure 116. Hence if the mean is μ then the error on the mean is $\mu^{0.5}$. For example, if μ = 10, then the fractional error is $\mu^{0.5}/\mu$ = 31.6%. If μ = 100, then the fractional error is $\mu^{0.5}/\mu$ = 10%. This shows that as more photons are counted, the noise becomes a smaller fraction of the signal, i.e. the signal-to-noise ratio improves. For more examples, see the example problems.

removed from course: proof that the variance of the Poisson distribution is equal to the mean.

---

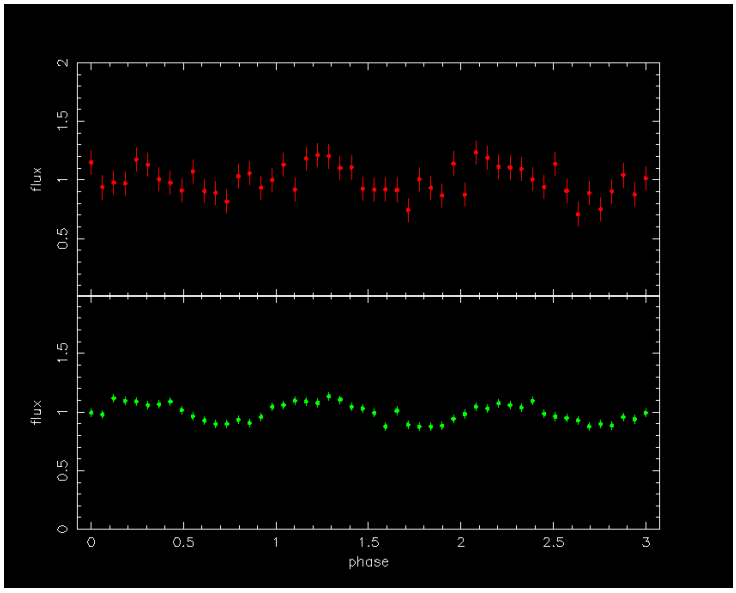
signal-to-noise ratio

Measuring the signal from an astronomical source, *S*, is not sufficient information on its own to determine whether or not the source is visible in an image or spectrum. A source might be bright but indistinguishable from the noise, *N*, if the noise is high. Conversely, a faint source might be visible if the noise is low. The correct statistic to use is therefore the *signal-to-noise ratio* (SNR):

SNR = signal / noise = *S* / *N*.

A signal-to-noise ratio of 10, for example, implies that the noise is one tenth of the signal. Hence the size of the error bar is 10%. The SNR required for an astronomical observation depends on the scale of the feature being studied. For example, imagine that you wish to determine the amplitude of the brightness variation of a variable star. If the amplitude of variation is, say, 0.1 magnitudes then, recalling the rule of thumb for magnitudes, this is a brightness variation of approximately 10%. It would be difficult to make out a 10% brightness variation if the typical scatter in the data (denoted by the size of the error bar) is also 10%, as demonstrated in figure 117. Hence a SNR of 10 would not be appropriate in this case.

figure 117: Two simulated light curves of a variable star, where the amplitude of the variability is 10% of the mean signal. The top light curve has a SNR of 10, i.e. the noise, and hence the error bars, are the same size as the amplitude. The bottom light curve has a SNR of 33 and one can be much more confident that the variability has been detected.

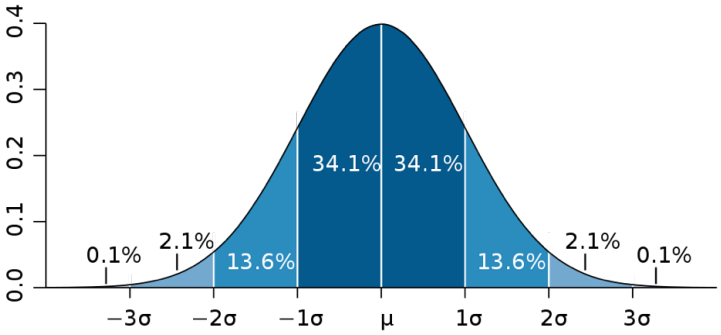


So what would be an appropriate SNR? Most astronomers would not believe a result unless it is at least *"three sigma"*, by which they mean that the quantity being measured is at least three times bigger than its error bar (denoted by the standard deviation, σ). In the above example of a variable star, the quantity we wish to measure is the amplitude of the variability, which is 10%. Hence an astronomer would not believe that a variability has been detected unless the error bar on the measured amplitude is less than 3.3%, i.e. a signal-to-noise ratio of 33 (see figure 117). Note, however, that this is a simplified and overly-conservative analysis, as the fact that there are multiple points in the light curve means that the SNR of the individual points could be worse than this and the variability would still be detectable.

Many astronomers would query even a 3σ result. Assuming that the errors on a measurement are normally distributed, one would expect a measurement to deviate by chance by 3σ from the mean once in every 370 measurements, i.e. the result is 99.7% significant. This is illustrated in figure 118. However, this assumes that there are no systematic errors in the data (e.g. in the flux calibration), which can move the position of the mean from its true value, effectively making it more likely that a 3σ result can occur by chance. To provide increased protection against systematic errors, therefore, astronomers often adopt 5σ as a measure of the believability of a result; a 5σ deviation occurs by chance once in every 1,744,278 measurements, i.e. it is 99.99994% significant.

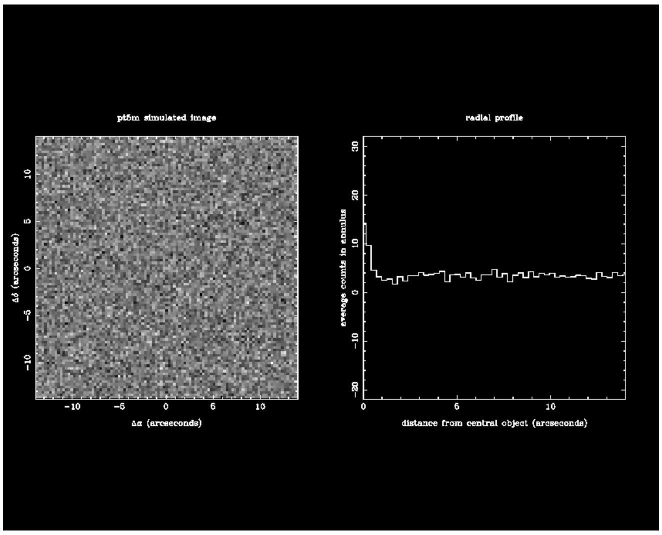
figure 118: Normal (or Gaussian) distribution. Regions which are less

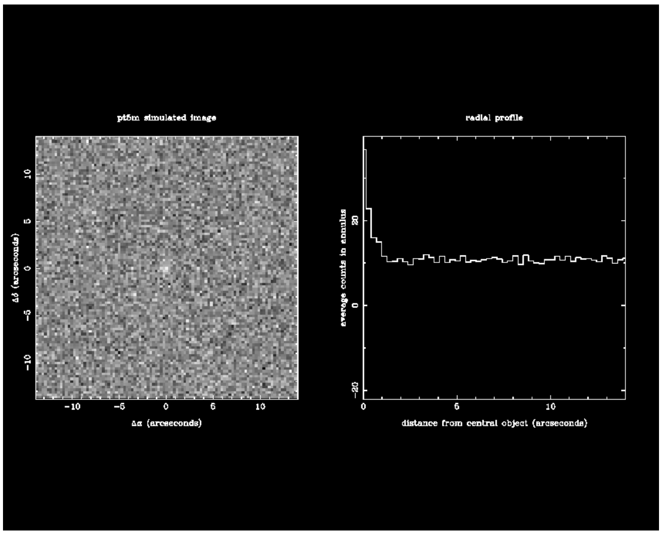
than one, two and three standard deviations (1σ, 2σ and 3σ) from the mean are coloured in different shades of blue. These regions account for 68.3%, 95.4% and 99.7% of the area under the curve, respectively.

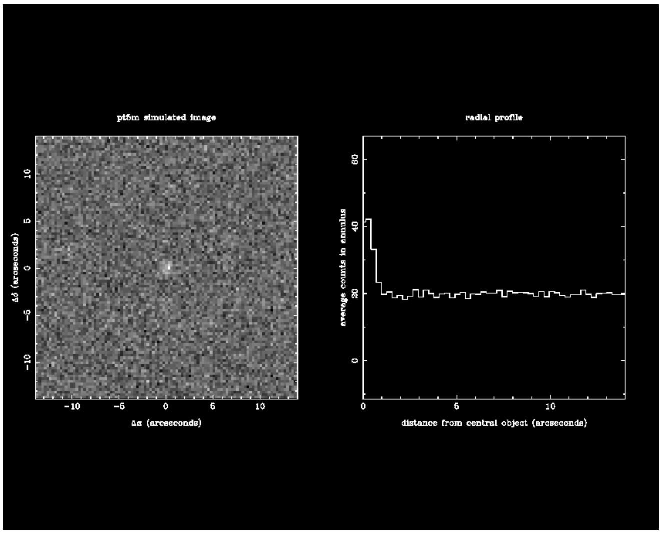


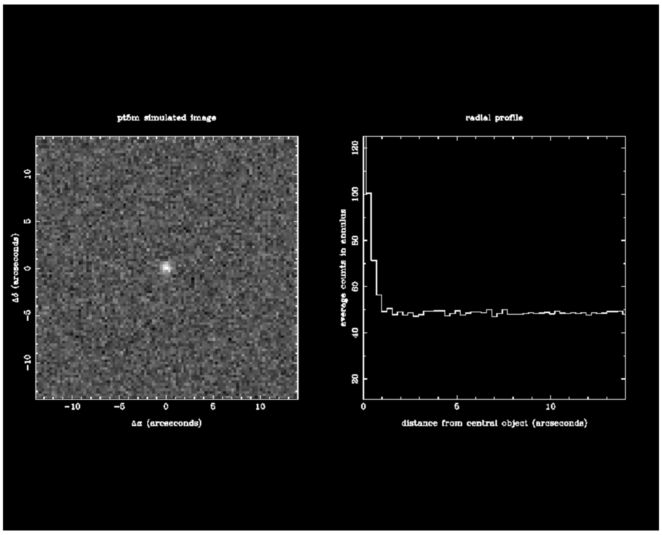
It is useful to have a feel for what a stellar image of a particular SNR looks like. Simulated images of a star detected at a SNR of 1, 3, 5, 10, 100 and 1000 are shown in figure 119, along with their corresponding radial profiles (a form of cross-section through the image). As expected, the star is invisible in the SNR = 1 image, and only just visible as a faint smudge in the SNR = 3 and 5 images. The star becomes readily visible to the eye at SNR = 10, and detecting the star at SNR = 100 or higher gives good-quality data.

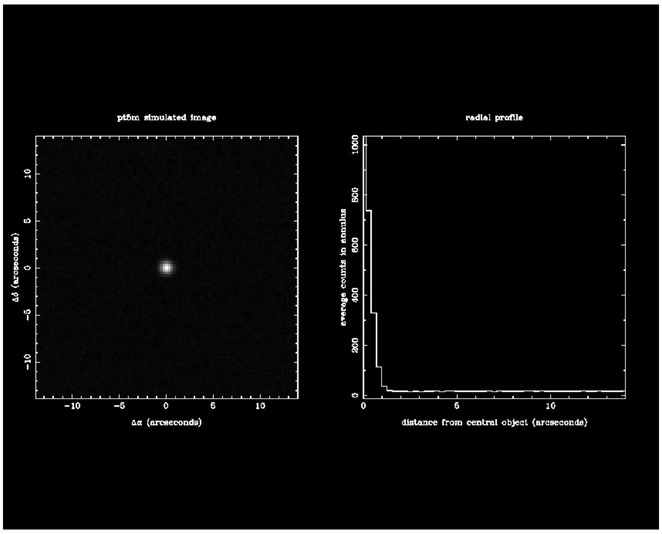
figure 119: Simulated images and radial profiles of a star detected at a SNR of 1 and 3 (top panel), 5 and 10 (central panel), and 100 and 1000 (bottom panel). In each image, the maximum count level is set to white and the minimum count level to black.

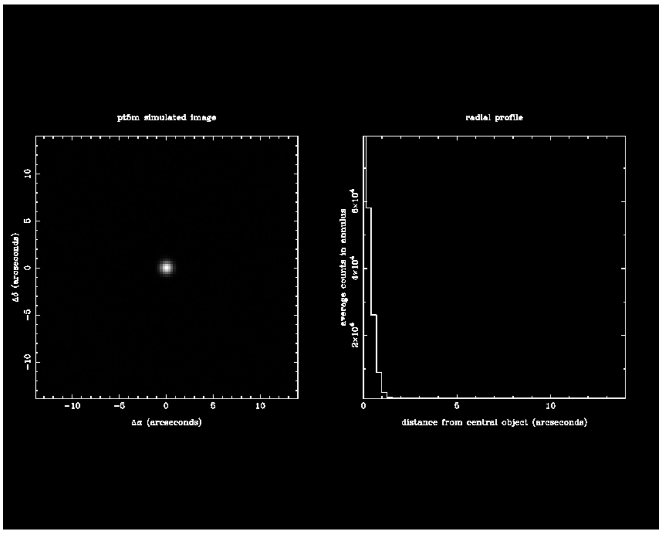




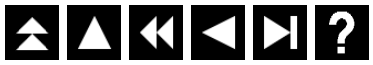








the ccd equation



To write an equation for SNR we need to know the various noise sources that contribute to an astronomical measurement with a CCD. These are:

- Random fluctuations in the detected photons from the source, i.e. the shot noise on the source signal, $(S_{obj})^{0.5}$.
- Random fluctuations in the detected photons from the sky background, i.e. the shot noise on the sky signal, $(S_{sky})^{0.5}$.
- Random fluctuations in the thermally-generated electrons produced in the CCD, ie. the shot noise on the dark current, $(S_{dark})^{0.5}$.
- Time-independent detector noise, i.e. readout noise, R. Note that no square root is required here as there is no signal associated with this noise source and hence Poissonian statistics do not apply.

Assuming that all of the above noise sources are independent, the total noise, N, is given by the square root of the sum of the squares of the individual errors:

$$N = (S_{obj} + S_{sky} + S_{dark} + R^2)^{0.5}.$$

Hence the SNR is given by the equation:

$$\text{SNR} = S / N = S_{obj} / (S_{obj} + S_{sky} + S_{dark} + R^2)^{0.5}.$$

It is important to realise that the above equation applies even *after* subtracting the mean sky background and dark current levels from each pixel, since the shot noise from these sources will still be present.

To use the above equation correctly, one must be careful with the units used for each of the terms. Typically, when predicting the SNR of an observation, we have:

- S_{obj} in units of photons per second.

- S_{sky} in units of photons per second per pixel.
- S_{dark} in units of electrons per second per pixel.
- R in units of electrons per pixel.

A few things are worthy of note in the above list. First, S_{obj} is the *total* number of photons from the object, which will probably be spread out over a number of pixels, whereas S_{sky} is the number of sky photons *per pixel*. Second, S_{obj} and S_{sky} are in photon units not electrons. Third, S_{obj}, S_{sky} and S_{dark} will increase with exposure time, but R will not.

So, to clarify the SNR equation given above, we need to account for the exposure time of the CCD image, t, the number of pixels that the object is spread over, n_{pix}, and the conversion efficiency of photons to electrons, which is given by the quantum efficiency, QE, of the CCD expressed as a number between 0 and 1. The latter conversion from photons to electrons is essential, as otherwise one would predict a higher signal-to-noise than measured, i.e. the signal in the SNR equation must be the *detected* signal, not the signal emitted by the source. The resulting equation is sometimes referred to as the *CCD equation*:

$$\mathrm{SNR} = (S_{obj} \,.\, t \,.\, QE) \,/\, [\, (S_{obj} \,.\, t \,.\, QE) + (S_{sky} \,.\, t \,.\, QE \,.\, n_{pix}) + (S_{dark} \,.\, t \,.\, n_{pix}) + (R^2 \,.\, n_{pix}) \,]^{0.5},$$

which can be simplified to:

$$\mathrm{SNR} = [S_{obj} \,.\, (t \,.\, QE)^{0.5}] \,/\, [\, S_{obj} + n_{pix}\, (S_{sky} + (S_{dark} \,/\, QE\,) + (R^2 \,/\, QE \,.\, t))\,]^{0.5}.$$

Sometimes, S_{obj} and S_{sky} are given in counts. In this case, we must also convert them into electron units. This is because Poisson statistics are only applicable when counting independent, random events, which means that the noise is only equal to the square root of the signal if the signal is in units of the detected quantity, i.e. electrons. The conversion from counts to electrons can be performed by replacing QE in the above equation by the CCD gain, g, in units of e^-/ADU:

$$\mathrm{SNR} = (S_{obj} \,.\, t \,.\, g) \,/\, [\, (S_{obj} \,.\, t \,.\, g) + (S_{sky} \,.\, t \,.\, g \,.\, n_{pix}) + (S_{dark} \,.\, t \,.\, n_{pix}) + (R^2 \,.\, n_{pix})\,]^{0.5},$$

which can be simplified to:

$$SNR = [S_{obj} \,.\, (t \,.\, g)^{0.5}] \,/\, [\, S_{obj} + n_{pix} \,(S_{sky} + (S_{dark} \,/\, g\,) + (R^2 \,/\, g \,.\, t))\,]^{0.5}.$$

Similarly, S_{obj} and S_{sky} are sometimes given in flux units, and these must be converted into electrons before the CCD equation can be used. In this case, one must first divide the flux by the energy of a single photon to give the number of photons, and then multiply by the *QE* to give the number of electrons.

We can use the CCD equation to define three limiting cases: the object-limited case, the background-limited case and the readout-noise limited case.

- **Object limited:** In this case, the object signal per pixel is much larger than the sky signal, dark current or readout noise per pixel. Hence,

  $$SNR = S_{obj} \,/\, (S_{obj} + S_{sky} + S_{dark} + R^2)^{0.5} \sim (S_{obj})^{0.5}.$$

  Hence the SNR increases as the square root of the object signal. Since the object signal is proportional to the exposure time, this means that the SNR is proportional to the square root of the exposure time. If the exposure time is doubled, the SNR will increase by $2^{0.5} \sim 1.4$. The object signal is also proportional to the area of the telescope aperture, which is proportional to the square of the diameter. Hence the SNR is proportional to the diameter of the telescope aperture - if the diameter is doubled, the SNR will double.

- **Background limited:** In this case, the sky signal per pixel is much larger than the object signal, dark current or readout noise per pixel. Hence,

  $$SNR = S_{obj} \,/\, (S_{obj} + S_{sky} + S_{dark} + R^2)^{0.5} \sim S_{obj} \,/\, (S_{sky})^{0.5}.$$

  This is similar to the object-limited case. Both the object and sky signal increase linearly with exposure time, hence the SNR is proportional to the square root of the exposure time. Both the object and sky signal also increase linearly with telescope area, hence the

SNR is proportional to the diameter of the telescope aperture.

For a given sky signal, the SNR will increase linearly with the object signal. For a given object signal, however, the SNR decreases as the square root of the increasing background level. This is why it is so important to minimize light pollution and observe faint objects when the Moon is new.

- **Readout-noise limited:** In this case, the readout noise per pixel is much larger than the object signal, sky signal or dark current per pixel. Hence,

  SNR = $S_{obj} / (S_{obj} + S_{sky} + S_{dark} + R^2)^{0.5} \sim S_{obj} / R$.

  Since the readout noise is independent of integration time or telescope aperture, the SNR will now increase linearly with exposure time and as the square of the telescope aperture diameter.

To maximise SNR, one should always try to expose for long enough to obtain object- or background-limited data, as otherwise one pays a significant penalty for reading out the CCD. However, it isn't always possible to avoid the readout-noise limited regime, particularly when exposure times must be kept short in order to sample short time-scale variability.

Some calculations illustrating how to use the CCD equation are given in the example problems.

---

example problems

1. **Suppose that we detect 1000 photo-electrons from an astronomical source in a certain time interval. In the absence of other noise sources, what is the likely error in the measurement and what signal-to-noise ratio (SNR) is obtained?**

   From Poisson statistics, an estimate of the error is simply the square root of the number of detected photo-electrons,

   $\sigma = (1000)^{0.5} = 31.6$.

   Thus we can write that the number of photo-electrons = 1000 ± 32.

   The SNR is simply the signal divided by the noise, which in this case is again the square root of the number of detected photo-electrons:

   $SNR = S / N = S / S^{0.5} = S^{0.5} = (1000)^{0.5} = 31.6$.

2. **A CCD detects a total of 600 counts/s from an astronomical source, and 1200 counts/s/pixel from the sky. If the exposure time is 1000 s, the readout noise of the CCD is 5 e^-/pixel and the dark current is 2 e^-/pixel/s, what is the error in the measurement of the source and what is the SNR? You may assume that the gain of the CCD is 1.2 e^-/ADU and that the light from the source is spread over 50 pixels.**

   The error in the measurement, i.e. the noise N, is given by the denominator of the CCD equation:

   $$N = [(S_{obj} . t . g) + (S_{sky} . t . g . n_{pix}) + (S_{dark} . t . n_{pix}) + (R^2 . n_{pix})]^{0.5}.$$

Note that we have chosen the form here that includes the CCD gain, *g*, as this is required to convert the object and sky signals, which are in units of counts, into units of photo-electrons. Hence

$N = [(600 \times 1000 \times 1.2) + (1200 \times 1000 \times 1.2 \times 50) + (2 \times 1000 \times 50) + (5^2 \times 50)]^{0.5} = 8534\ e^-$.

The SNR is then

SNR = $(S_{obj} \cdot t \cdot g) / N = (600 \times 1000 \times 1.2) / 8534 = 84.$

3. A detector has incident upon it a total of 1 photon/s from an astronomical source and a total of 2 photons/s from the sky underneath the source. If the dark current and readout noise are negligible, how long an exposure is required to achieve a SNR of 30?

In this question, the *QE* of the detector is not given and so we are unable to convert the incident photons into detected photo-electrons. So, we must assume that it is a perfect detector, i.e. 100% *QE*.

Since the dark current and readout noise are negligible, we can write for the SNR that:

SNR = $(S_{obj} \cdot t) / [(S_{obj} \cdot t) + (S_{sky} \cdot t)]^{0.5}$.

Rearranging for *t*, we obtain:

$t = \text{SNR}^2 \cdot [(S_{obj} + S_{sky}) / (S_{obj})^2]$.

Hence

$t = (30^2) \cdot 3 / 1 = 2700$ seconds.
