

12 Instrumentation and Detectors

Ian S. McLean · James Larkin · Michael Fitzgerald

Department of Physics and Astronomy, University of California,
Los Angeles, CA, USA

1	<i>Introduction</i>	508
2	<i>Classification of Instruments</i>	509
2.1	Camera Systems	509
2.2	Coronagraphs	511
2.3	Spectrometer Systems	512
2.4	Integral Field Spectrometers	517
2.5	Polarimeter Systems	520
2.6	Interferometers	521
3	<i>Detectors and Materials</i>	523
3.1	Classification of Detectors	523
3.2	Semiconductors	525
3.3	Photoconductors	526
3.4	Photodiodes	527
3.5	Applications to CCDs and IR Arrays	528
3.6	Detectors for High Energy	532
3.7	Thermal Detectors	533
3.8	Coherent Detectors	533
4	<i>Cryogenics and Vacuum Systems</i>	535
	<i>References</i>	538

Abstract: This chapter contains a broad introduction to astronomical instruments and detectors. The basic design principles for cameras, spectrometers, polarimeters, and interferometers are given, together with some practical material on instrument building techniques, including vacuum-cryogenic methods. Different detector technologies are introduced, such as CCDs and infrared arrays, together with basic information on semiconductors.

Keywords: Cameras, Coronagraphs, Detectors, Interferometers, Polarimeters, Spectrometers

List of Abbreviations: *AO*, Adaptive optics; *CCD*, Charge-coupled device; *IF*, Intermediate frequency; *IFS*, Integral Field Spectrometer; *IR*, Infrared; *LO*, Local oscillator; *MCP*, Microchannel plate; *MOS*, Metal oxide semiconductor; *NASA*, National Aeronautics and Space Administration; *PMT*, Photomultiplier tube; *PSF*, Point spread function; *SIS*, Superconductor Insulator Superconductor; *UV/O/IR*, Ultraviolet/optical/infrared; *VLT*, Very large telescope; *VPH*, Volume Phase Holographic

1 Introduction

When viewed across the electromagnetic spectrum, astronomical instrumentation can appear quite different, in part because of the wide range of technologies involved, but variety is fundamentally limited by the properties that can be measured. Following the scheme offered in McLean (2008), most instruments can be placed in one of four categories: a photometer or radiometer for measuring the brightness and direction of radiation; a spectrometer for measuring the distribution of brightness as a function of wavelength, frequency, or energy; a polarimeter to determine the degree of alignment or handedness of wave vibrations; or an interferometer which relies on coherent phase relationships to achieve interference effects prior to detection.

All instruments contain some kind of radiation detector and, once again, the range of detector technologies is broad. For most applications in astronomy, however, there are three main classifications of detectors as follows. Photon detectors in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma rays to the far-infrared. Thermal detectors in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical conductivity; thermal detectors have a broad spectral response but are often used for infrared and submillimeter detection. Coherent detectors in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

Breakthrough discoveries in astronomy often rely on the development of new technologies. Among the technologies that have made a difference, both on the ground and in space, in recent years are larger and more sensitive charge-coupled devices (CCDs) and infrared array detectors; more powerful cameras and spectrometers; improved methods for building very large optical/IR telescopes and for building efficient survey telescopes; advances in optics and detectors for X-ray astronomy; devices for the study of the cosmic microwave background; new digital signal processing techniques and new receiver/antenna designs for radio astronomy. Especially impressive has been the advent of adaptive optics techniques that enable large ground-based telescopes to operate at their ultimate diffraction limit.


When the silicon CCD was introduced into astronomy in 1974 (see Janesick 2001), it completely revolutionized astronomical imaging. From a modest 10,000 pixels in the early days, individual CCDs are now 4–16 million pixels, and many instruments employ large mosaics of CCDs to push the total number of pixels up to about one billion. A remarkable feature of the CCD is that it can also detect X-ray photons, and so CCDs are found in the X-ray cameras on the Chandra X-ray Observatory (Garmire et al. 2003). Modern CCDs can now be used for ultraviolet work, but competing devices such as the microchannel plate (MCP) have been used successfully on space missions such as GALEX (Morrissey et al. 2005). Silicon's band gap restricts the use of CCDs to wavelengths less than 1,100 nm, and thus a different kind of detector is needed for the infrared. Usually referred to as infrared arrays, these devices do not employ the charge-coupling principle, and semiconductor materials with a lower band gap must be used. In the near-infrared, these array detectors are now at the 4–16 megapixel size (McLean 2008). For even longer wavelengths, devices known as Transition Edge Sensors (TES) that rely on superconductivity have been developed into arrays for submillimeter astronomy. At the other end of the spectrum, high-energy astronomers have obtained images using large pixel arrays of cadmium zinc telluride (CZT) to detect gamma rays. One example is the Burst Alert Telescope (BAT) on NASA's *Swift* satellite which has 32,768 CZT detectors covering a focal plane of 1.2×0.6 m (Gehrels et al. 2004).

As telescopes have grown in size and instruments have become more costly, a strong trend toward general user facilities and large collaborations has occurred. Most astronomical instruments are therefore highly automated and capable of remote operation. Inevitably, these instruments must be well engineered for reliability. Many important factors constrain the design of new astronomical instrumentation. Engineering details are beyond the scope of this article, but the following sections will serve as a starting point. Clearly, the choice of instrument and the details of the design must depend on the science to be done. This chapter contains an introduction to astronomical detectors and provides basic design principles for cameras, spectrometers, polarimeters, and interferometers. Also included is some material on instrument building, including vacuum-cryogenic techniques. More details can be found in other chapters in this volume and in the references given.

2 Classification of Instruments

Astronomical instruments are often grouped into four classes: (1) photometers/cameras, (2) spectrometers, (3) polarimeters, and (4) interferometers. Although the methods of implementation differ considerably, variations of these instruments exist from X-ray wavelengths to radio wavelengths. The descriptions which follow are mainly applicable for UV, visible, and infrared wavelengths (UV/O/IR).

2.1 Camera Systems

In the simplest camera, the detector (CCD or other array device) is placed directly in the focal plane of the telescope behind a light-tight shutter. Filters to select different wavelength bands are therefore located in the converging beam from the telescope. An alternative approach, shown in  Fig. 12-1, is to collimate the beam by placing a lens after the focal plane at a distance s equal

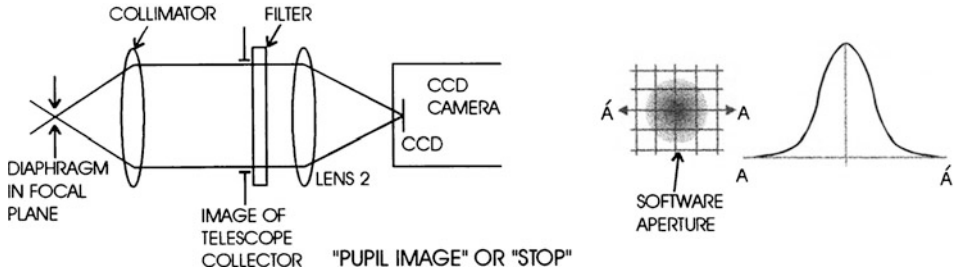


Fig. 12-1

The layout for a basic camera system in which optics are used to collimate the diverging beam from the telescope focal plane and reimage the field at a different magnification. Photometry is performed using software apertures on the digital image (From McLean 2008)

to its focal length ($s = f_{\text{coll}}$). The field is reimaged onto the detector with a lens (or mirror) with $s' = f_{\text{cam}}$. By selecting the focal lengths of the collimator and camera lenses, one can either magnify or reduce the scale; $m = f_{\text{cam}}/f_{\text{coll}}$. Filters of arbitrary thickness can be located in this collimated beam. Moreover, the filters can be placed near the image of the primary mirror created by the collimating optics; this is called the “pupil” image. In addition, a circular aperture or stop can be placed at the pupil image to reject stray light from outside the beam. A pupil stop is important, especially in infrared cameras where the pupil is at cryogenic temperatures and so it becomes a cold stop.

Matching the spatial or spectral resolution to the physical size of the detector pixels is important. There are two factors to consider: (1) maximizing observing efficiency and (2) obtaining accurate brightness measurements (photometry). In general, the image is either critically sampled, meaning that there will be about 2 pixels (the Nyquist limit) across the resolution element, or it will be oversampled, implying that there may be about 5 pixels across the resolution element. In a spectrometer, the width of the entrance slit is usually the determining factor.

The plate scale of the telescope is given in seconds of arc per mm ("/mm) by

$$(\text{ps})_{\text{tel}} = \frac{206,265}{f_{\text{tel}}} \quad (12.1)$$

Here, f_{tel} is the focal length of the telescope in millimeters ($f_{\text{tel}} = D_{\text{tel}} \times F$ where F is the focal ratio or f /number) and the numerical factor is the number of seconds of arc in 1 radian. For direct imaging, the angle on the sky subtended by the detector pixel is

$$\theta = (\text{ps})_{\text{tel}} d_{\text{pix}} \quad (12.2)$$

where d_{pix} is the physical pixel size in mm; pixels are usually square. Calculating the required magnification factor can be done as follows:

- choose a value for the diameter of the seeing in seconds of arc, θ_{see}
- decide on the sampling ($p = 2-5$ pixels)
- divide seeing diameter by sampling factor to get angular size of 1 pixel, $\theta_{\text{pix}} = \theta_{\text{see}}/p$
- derive the plate scale at the detector from $(\text{ps})_{\text{det}} = \theta_{\text{pix}}/d_{\text{pix}}$
- required magnification (m) is then

$$m = \frac{(\text{ps})_{\text{tel}}}{(\text{ps})_{\text{det}}} \quad (12.3)$$

where $m = f_{\text{cam}}/f_{\text{coll}}$ as before.

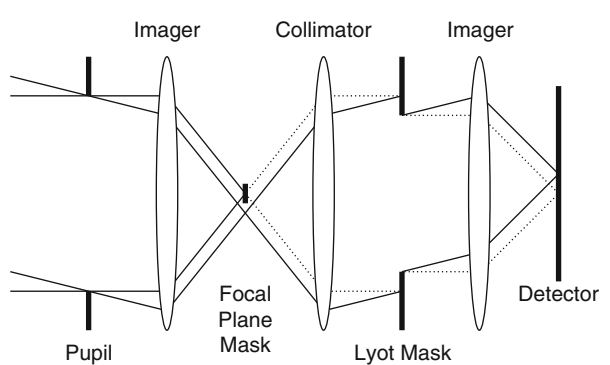
Note that m also defines an effective focal length ($\text{EFL} = mf_{\text{tel}}$) for the entire optical system. If $m > 1$, then the optical components are a magnifier, whereas if $m < 1$ (the usual case), then the optics are called a “focal reducer.” We can also relate the pixel size in seconds of arc to the f -number of the focal reducer optics (or simply, “the camera”) by

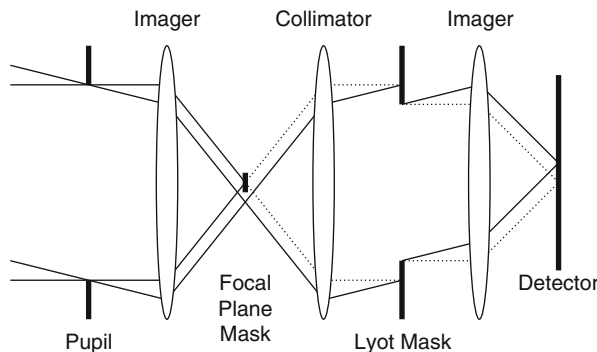
$$\theta_{\text{pix}} = 206,265 \frac{d_{\text{pix}}}{D_{\text{tel}}(f/\text{number})_{\text{cam}}} \quad (12.4)$$


where $(f/\text{number})_{\text{cam}} = f_{\text{cam}}/D_{\text{cam}} = F_{\text{cam}}$.

For example, if $d_{\text{pix}} = 18\mu\text{m}$ and $D_{\text{tel}} = 10\text{ m}$, then $\theta_{\text{pix}} = 0.37''/(f/\text{number})_{\text{cam}}$. Assuming seeing of $0.6''$ and 3-pixel sampling, this implies $\theta_{\text{pix}} = 0.2''$ which leads to $F_{\text{cam}} = 1.856$.

2.2 Coronagraphs

Astronomical investigations often seek to measure emission from material that lies at a small angular separation from a much brighter source. Distinguishing light from the much fainter source is the observational challenge. Coronagraphs attempt to suppress or steer the transmission of initially on-axis light while simultaneously allowing off-axis light to transmit relatively unimpeded. Reducing the unwanted light subsequently reduces the noise in measurements of the off-axis brightness. Many coronagraphic devices in astronomical instruments use a combination of a focal plane mask and a Lyot stop. As shown in  Fig. 12-2, a traditional Lyot coronagraph uses an opaque disk as the focal plane mask together with an undersized circular pupil as the Lyot stop or mask (Lyot 1939).



 Fig. 12-2

In the Lyot coronagraph, a focal plane mask is used to block the on-axis light, while the Lyot mask is a pupil stop that suppresses initially on-axis light that is diffracted by the pupil stop and focal plane mask. Initially off-axis light is transmitted through the system, although the intensity is reduced compared to a normal image by the action of the Lyot mask in particular

An image of a bright object, such as a star, is not an infinitesimal point in the telescope image plane. Even in the absence of wavefront errors, diffraction by the telescope pupil causes the image of a point-like source to be extended. For example, an image of a point source formed by a system with a circular pupil of diameter D will have an Airy pattern (see [Chap. 6](#)), characterized by a central core of width $\sim \lambda/D$ surrounded by concentric rings. An opaque circular focal plane mask can be used to block light from the Airy core and central bright Airy rings. By itself, however, this focal plane mask has done little to improve the ability to detect faint emission from around the star. Initially on-axis starlight is diffracted by the telescope pupil to positions beyond the extent of a finite focal plane mask, that is, to the outer Airy rings. In the pupil plane, located after the focal plane mask, this light is concentrated near the edge of the pupil image. By placing an undersized stop at this location (the Lyot mask), on-axis starlight that was initially diffracted to locations in the outer focal plane is also blocked.

Quantifying the performance of a coronagraphic imaging system is usually done in terms of the *contrast*, which measures the detectable flux relative to that of the on-axis point source. The precise definition can vary, but for point sources, this is commonly given in terms of the flux ratio between the on-axis point source *if it were not occulted* and the minimum detectable off-axis source flux. Contrast will be a function of position in the focal plane. The architecture of the coronagraph has implications for the contrast. For example, the size of the focal plane mask sets an *inner working angle*, inside of which off-axis sources are also occulted and thus undetectable. The diameter of the Lyot mask affects the amount of diffracted light from the on-axis source that is transmitted to the detector. Smaller Lyot masks result in more suppression of light from on-axis sources. However, this will not necessarily increase the achieved contrast, because decreasing the Lyot mask size also decreases the transmitted flux of any off-axis sources.

Variations of the Lyot coronagraph exist which use alternative focal plane masks. These masks manipulate the phase of the light rather than the amplitude, in order to steer it to regions in the Lyot plane that are masked. The Four Quadrant Phase Mask and the Optical Vortex Coronagraph are two examples of such architectures (Rouan et al. 2000; Palacios 2005). These devices have the advantage of reducing the inner working angle of the system. The effectiveness of diffraction control can be enhanced by adding an *apodizer* to the pupil prior to the focal plane mask. An apodizer smoothly tapers the transmission in the pupil, which has the effect of reducing the amplitude of rings in the focal plane compared to a hard-edged pupil stop (Soummer et al. 2003). Guyon (2006) reviews the performance limitations of interferometric and Lyot-style coronagraph designs. *Shaped-pupil coronagraphs* rely on novel shapes in the pupil plane to create regions in the focal plane that are devoid of diffracted on-axis light (Kasdin et al. 2003). Such systems often do not require smooth tapering of transmission profiles in the pupil mask, which can be difficult to manufacture precisely, and also do not require focal plane masking. Only a portion of the image plane achieves high contrast, which is a disadvantage.

2.3 Spectrometer Systems

[Figure 12-3](#) shows the essential features of a classical UV/O/IR spectrometer in which light enters through a narrow slit. The width of the slit must be matched to either the seeing conditions or the diffraction disk. As the beam diverges from the slit plane, it is collimated and directed to the dispersing system, after which the spectrally dispersed beam is collected by the camera optics and re-imaged onto the detector. Long-slit, multislit, and slitless spectrometers

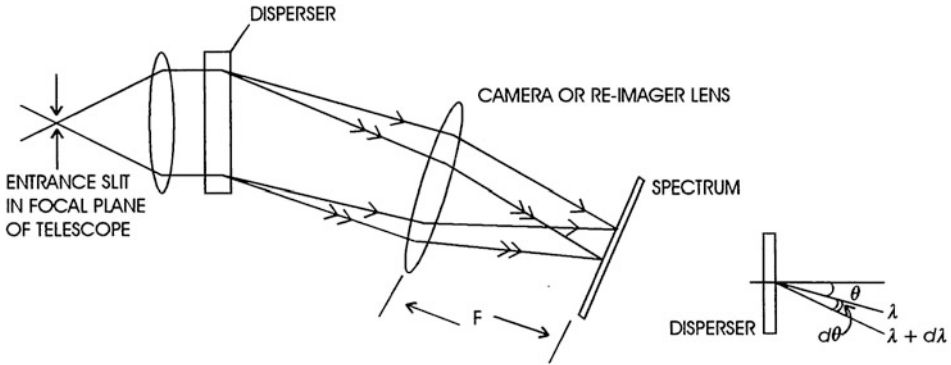


Fig. 12-3

Essential features in the optical layout of a spectrometer are illustrated. The beam is collimated before intersecting the dispersive element and then the spectrum is reimaged with camera optics onto the detector (From McLean 2008)

are in wide use throughout astronomy. In almost all cases, a two-dimensional detector records the spectrum. With the advent of optical fibers to collect light from many locations in the focal plane and feed it to the entrance slit, a huge multiplex advantage can be obtained which facilitates large-scale spectroscopic surveys. Following McLean (2008), for a given wavelength (λ) range, the important design quantities are (► 12.1) the resolving power ($R = \lambda/\Delta\lambda$), (► 12.2) the slit width, (► 12.3) the diameter of the collimated beam, (► 12.4) the sampling or matching of the slit width to the detector pixels, and (► 12.5) the resulting f /number of the camera system.

Linear dispersion (L.D.) relates an interval of length (dx in mm) along the spectrum to a wavelength interval ($d\lambda$ in Ångstroms or nanometers)

$$\text{L.D.} = \frac{dx}{d\lambda} = \frac{dx}{d\theta} \frac{d\theta}{d\lambda} = F \frac{d\theta}{d\lambda} \quad (12.5)$$

Here, f_{cam} is the focal length of the spectrograph camera and $d\theta/d\lambda$ is the angular dispersion of the prism or grating device. The units are usually expressed as $\text{mm}/\text{Å}$, but a more useful form is the Reciprocal Linear Dispersion which is simply the inverse of the above expression in $\text{Å}/\text{mm}$. Expressions for angular dispersion follow from the basic equations for prisms and gratings. For a diffraction grating, the equation is

$$m\lambda = d(\sin i + \sin \theta) \cos \gamma \quad (12.6)$$

where d is the spacing of adjacent grooves or slits, i is the angle of incidence of the collimated beam, θ is the angle of the emergent diffracted beam, γ is the angle out of the normal plane of incidence (usually 0° , hence $\cos = 1$), and m is an integer called the “order” of interference. For zero order ($m = 0$), $\sin \theta = -\sin i$ or $\theta = -i$. The negative sign comes from the fact that we have chosen to call i and θ positive when on the *same* side of the normal. Whenever the rays cross over the normal, the angle of diffraction is taken to be negative. With this sign convention, the equation applies when the grating is used in transmission and when the grating is used in reflection. There is an alternative form of the equation that uses a negative sign between the terms to describe a reflection grating. In that case, the angles are positive if they are on *opposite*

sides of the normal. If the medium on either side of the grating is not a vacuum, then a more general form would be $(n_1 \sin i + n_2 \sin \theta)$. From (12.6), the angular dispersion of a grating is given by

$$\frac{d\theta}{d\lambda} = \frac{m}{d \cos \theta \cos \gamma} \quad (12.7)$$

Substituting for m/d gives

$$\frac{d\theta}{d\lambda} = \frac{\sin i + \sin \theta}{\lambda \cos \theta \cos \gamma} \quad (12.8)$$

Usually $\cos \gamma \sim 1$ and therefore angular dispersion is determined entirely by i and θ for a given λ . Many combinations of m and d yield the same A.D. provided the grating angles remain unchanged. Typical “first-order gratings” ($m \sim 1$) have 300–2,400 grooves/mm; the number of lines per mm is given by $T = 1/d$. Coarse-ruled reflection gratings (large d) can achieve high angular dispersion by making i and θ very large, typically 60° . Such gratings are called “echelles” and have groove densities from 20 to 200 lines/mm with values of m in the range 10–100. This results in severe overlap of the orders unless a second disperser of lower resolving power is used at right angles to the first to “separate” the orders.

In practice, spectrometers are usually slit-width or seeing-limited. Taking $\theta_{\text{see}} = p \times \theta_{\text{pix}}$, where p is the number of pixels across the slit image, and converting to seconds of arc, gives a form which shows explicitly the trade-offs of size versus resolution:

$$R = \left(\frac{\sin i + \sin \theta}{\cos i} \right) \frac{D_{\text{coll}}}{D_{\text{tel}}} \frac{206,265}{p \theta_{\text{pix}}} \quad (12.9)$$

This formula makes it clear that as telescopes get larger, the spectrograph (defined by the beam size D_{coll}) gets larger too, all else being equal.

By tilting the facets of a reflection grating through an angle θ_B (known as the *blaze* angle) with respect to the plane of the grating surface, it is possible to maximize the grating efficiency in the direction in which light would have been reflected in the absence of diffraction. Grating efficiency is a maximum when the angle of incidence and angle of diffraction are related by $(i + \theta) = 2\theta_B$. The separation between the beams ($i - \theta$) is just the spectrograph angle ϕ . Thus,

$$m\lambda_B = 2d \sin \theta_B \cos(\phi/2) \quad (12.10)$$

A special case occurs when $\phi = 0$, for then the incident ray enters along the normal to the facet and the diffracted ray leaves along the same direction. This is the “Littrow” condition and the incident and diffracted angles measured relative to the grating normal are now equal to each other and to the blaze angle. The grating equation simplifies to $m\lambda_B = 2d \sin \theta_B$, and the resolving power is given by

$$R = \frac{2D_{\text{coll}} \tan \theta_B}{\phi D_{\text{tel}}} \quad (12.11)$$

The only way to work in the Littrow condition is with a central obscuration in the optics. Alternatively one can use the “near” Littrow condition by moving off by a $10\text{--}20^\circ$ or the “quasi” Littrow condition by going out of the plane ($\gamma > 0^\circ$).

Prisms find applications in spectrographs both in the role of primary disperser in (usually) low-resolution instruments and as a cross-disperser in high-resolution echelle spectrographs. The basic layout of a prism disperser is shown in Fig. 12-4. From the definition of angular dispersion:

$$\frac{d\theta}{d\lambda} = \frac{d\theta}{dn} \frac{dn}{d\lambda} = \frac{B}{D_{\text{cam}}} \frac{dn}{d\lambda} \quad (12.12)$$

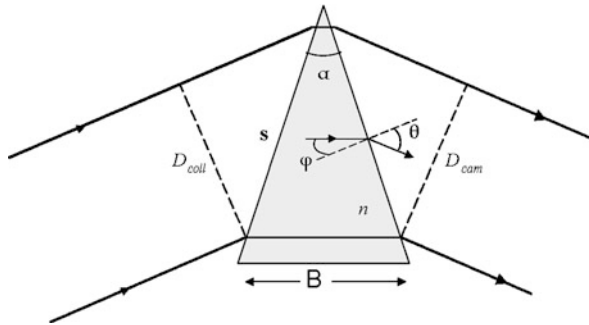


Fig. 12-4

The relationship of angles and lengths in a prism at minimum deviation are used to derive the resolving power (From McLean 2008)

In this expression, $dn/d\lambda$ describes the wavelength dependence of the refractive index n . The purely geometric term $d\theta/dn$ can be derived by differentiating Snell's law applied to the second surface and then doubling the rate to account for both surfaces. At minimum deviation, the angle $\phi = \alpha/2$ giving $d\theta/dn = [2s \sin(\alpha/2)/s \cos \theta]$. However, $2s \sin(\alpha/2) = B$, the base length of the prism, and $s \cos \theta = D_{cam}$ the emergent collimated beam width toward the camera. The resolving power of a prism is $R = B (dn/d\lambda)$, and for a slit-limited instrument, the resolving power is given by

$$R = \frac{\lambda}{\theta_{res} D_{tel}} B \frac{dn}{d\lambda} \quad (12.13)$$

A popular way to convert a camera into a spectrograph is to deposit a transmission grating on the hypotenuse face of a right-angled prism and use the deviation of the prism to bring the first order of diffraction on axis. Such a device is called a "grism" and the basic geometry (not to scale) is illustrated in Fig. 12-5. A grism can be placed in a filter wheel and treated like another filter. The basic relationships needed to design a grism are

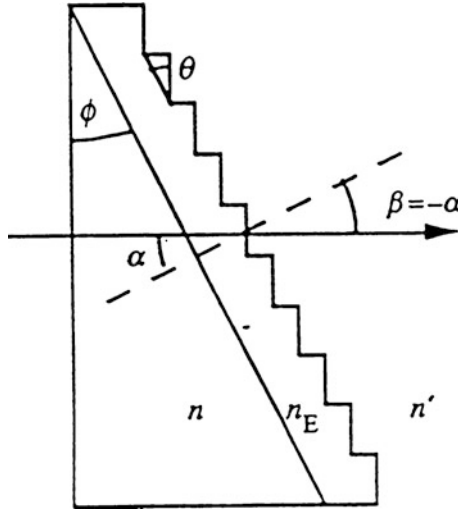
$$m\lambda_c T = (n - 1) \sin \phi \quad (12.14)$$

and

$$R = \frac{EFL}{2d_{pix}} (n - 1) \tan \phi \quad (12.15)$$

where λ_c is the central wavelength, $T (= 1/d)$ is the number of lines per mm of the grating, n is the refractive index of the prism material, and ϕ is the prism apex angle. EFL is the effective focal length of the camera system, and d_{pix} is the pixel size. The factor of 2 assumes that 2 pixels are matched to the slit width. In practice, the number of free parameters is constrained by available materials and grating rulings, and given conditions within the camera system. Resolving powers (2 pixels) of $R \sim 500$ – $2,000$ are practical.

Most astronomical gratings are of the surface relief kind in which the grooves are formed on the surface of the substrate (direct ruling) or as a replicated grating in a material bonded to the substrate. Reflection gratings can be coated with a reflective surface such as silver or gold, where the latter is particularly useful in the infrared. An alternative technology for grating fabrication is the Volume Phase Holographic or VPH grating. Not to be confused with a normal holographic grating which is another method for creating a surface relief grating, a VPH grating



■ Fig. 12-5

A simplified schematic giving the basic geometry of a *grism* – a transmission diffraction grating deposited on the hypotenuse face of a right-angled prism (From McLean 2008)

is an optical substrate in which the refractive index varies periodically throughout the body of the grating (Barden et al. 2000; Baldry et al. 2004). The grating body is made from a thin (3–30 μm) slab of dichromated gelatine (DCG) trapped between glass plates. Light passing through a VPH transmission grating obeys the following grating equation:

$$m\lambda = n_i \Lambda_g (\sin \alpha_i + \sin \beta_i) \quad (12.16)$$

where m is an integer representing the order, n_i is the refractive index of the medium, Λ_g is the grating period (equivalent to groove spacing) and is the projected separation between the fringes; $\Lambda_g = \Lambda / \cos \varphi$ where φ is the slant angle between the grating normal and the plane of the fringes. The angles of incidence (α_i) and diffraction (β_i) are relative to the grating normal with the convention that zero order (no diffraction) corresponds to $\beta_i = -\alpha_i$. The equation applies to each layer, where $i = 0$ is the air, $i = 1$ is the glass substrate, and $i = 2$ is the DCG layer. High diffraction efficiency can occur when light is effectively reflected from the plane of the fringes, that is, when $\beta_2 + \varphi = \alpha_2 - \varphi$ in the DCG layer. This behavior is the same as Bragg diffraction of X-rays from the atomic layers in a crystal lattice. In both cases, because the thickness of the medium is much greater than the wavelength, constructive interference occurs for radiation scattered in that direction. The Bragg condition implies

$$m\lambda = 2n_2 \Lambda \sin \alpha_{2b} \quad (12.17)$$

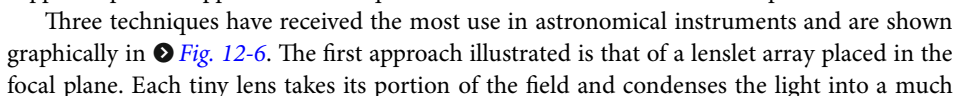
where n_2 is the refractive index of the DCG layer and α_{2b} is the “Bragg angle” or angle of incidence with respect to the plane of the fringes $\alpha_{2b} = \alpha_2 - \varphi$. At wavelengths sufficiently displaced from the Bragg condition, there is no diffraction. Diffraction efficiency also depends on the semiamplitude of the refractive index modulation (Δn_2) and the grating thickness (d). The DCG holds a fringe pattern generated by holography which provides planes of constant refractive index separated by a length $\Lambda = 1/\nu_g$. The index variations are the result of density variations

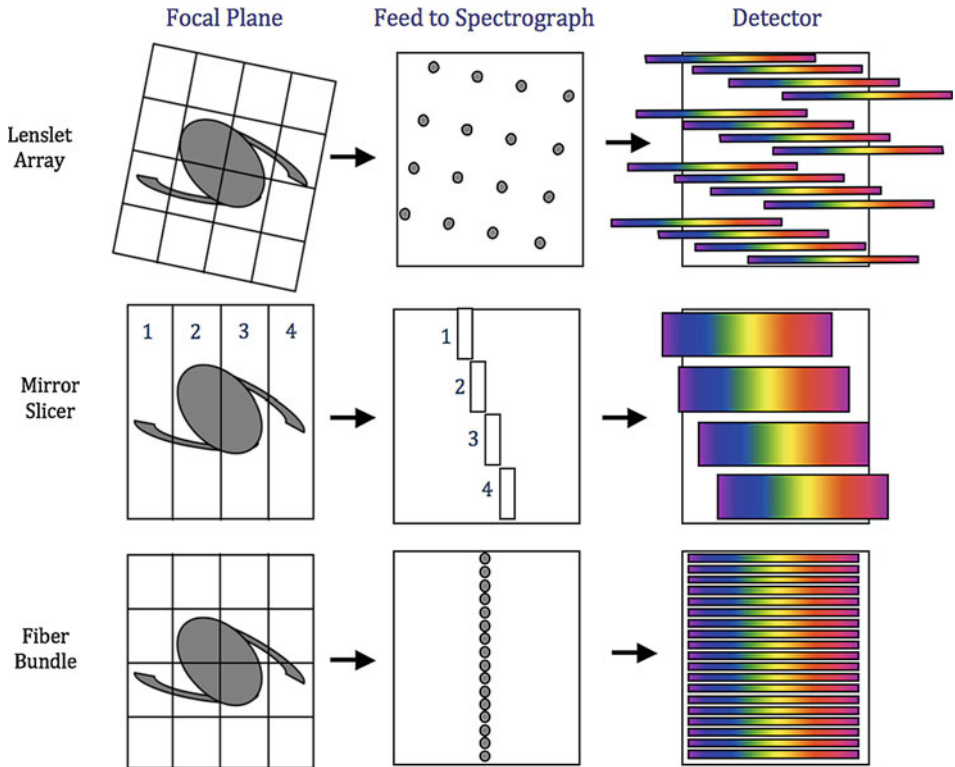
which are trapped into the material by exposure to light (the fringe pattern) because those regions collapse to a different density in the process of swelling the gelatin with water and then drying it rapidly. One form for the refractive index is $n_2(x, z) = n_2 + \Delta n_2 \cos[2\pi\nu_g(x \sin \gamma + z \cos \gamma)]$, which gives the variation in the x, z plane where z is the optical axis through the VPH, and γ is the angle between the normal to the planes and the z -axis. Line densities can range from 300 to 6,000 lines/mm, and index modulations of 0.02–0.1 are typical. Because of the Bragg condition, it is necessary to articulate the camera to a new angle to tune to a new wavelength.

2.4 Integral Field Spectrometers

An integral field spectrograph (IFS) samples a rectangular or other two-dimensional field of view and produces spectra for each spatial location. It does this by reformatting the focal plane in such a way that a traditional dispersing spectrograph can simultaneously disperse adjacent regions of the field without overlapping on the detector. Spectra from the various pieces of the field can then be reassembled into a cube of data covering two spatial dimensions and one of wavelength. Depending on the choices of field of view, spectral resolution, and spectral bandwidth, such cubes can contain tens, hundreds, or even thousands of spectral channels, all taken simultaneously and all covering a contiguous field. Almost all integral field spectrographs at telescopes use a standard CCD or infrared array as the detector. Because practical detectors are limited in terms of the total number of pixels, and since each spatial location uses hundreds or thousands of pixels for their respective spectra, the field of view of an IFS is often much smaller than in a traditional camera.

Nevertheless, having adjacent and simultaneous spectra can be a powerful scientific advantage in many cases. Examples include, measuring line ratios across such diverse objects as a high-redshift galaxy or a Jovian moon in an individual exposure in which slit losses from a classical spectrometer, changing seeing conditions and effects from adaptive optics on the point spread function are all minimized. In crowded fields like the Galactic Center, an integral field spectrometer can be used without fear of pointing errors since the entire field is reconstructed and synthetic apertures can be used on the cubes of data much like aperture photometry on individual images. In most cases, some portion of the field of view will also contain blank regions of sky, and thus atmospheric emission lines are recorded simultaneously with science photons. Depending on the design, a surprising benefit can be a very low wavefront error which is crucial for high-resolution applications. In particular, lenslet arrays and fiber bundles sample the focal plane prior to any of the spectroscopic optical elements like the collimators, cameras, and gratings. Consequently, image quality at the input to the reformatting optics remains the final image quality of the cubes. For example, in this way, the OSIRIS instrument at the Keck Observatory has a measured wavefront error below 25 nm across its entire field. In high-contrast applications, like the Gemini Planet Imager and Lyot Project, an integral field spectrograph with a lenslet array dramatically reduces noncommon path chromatic errors because the speckle pattern is recorded at the lenslet array where essentially no refractive elements, except for a window, have been in the beam path. Speckle coherence is maintained across a broad bandwidth which supports speckle suppression techniques to increase contrast with faint companions.

Three techniques have received the most use in astronomical instruments and are shown graphically in  Fig. 12-6. The first approach illustrated is that of a lenslet array placed in the focal plane. Each tiny lens takes its portion of the field and condenses the light into a much



■ Fig. 12-6

A summary of the three most common methods of creating an integral field spectrograph is illustrated. For each model, the image plane is shown on the *left*, the spectrograph input in the *middle*, and detector plane on the *right* (Based on a similar figure in Allington-Smith and Content (1998))

smaller pupil image. Thus, approximately one focal length behind the lenslet array, a grid of small pupil images serves as the entrance plane to a traditional spectrograph. The lenslet array is rotated compared to the dispersion axis of the spectrograph so that spectra are interleaved on the detector. Traditionally, lenslet-based spectrographs at optical wavelengths have been optimized for large spatial coverage with either low spectral resolution or small wavelength coverage. Lenslet designs were first proposed by Courtes (1982) and have since been implemented in a variety of instruments (e.g., Bacon et al. 1995). The first diffraction-limited use of a lenslet-based spectrograph was OSIRIS (Larkin et al. 2006) working in the near-IR with the Keck Adaptive Optics System (Wizinowich et al. 2000). This integral field spectrograph can produce spectra with as many as 1,800 spectral channels from 1,000 spatial locations in a 16×64 pattern. Similar lenslet based spectrographs with many more spatial elements and much shorter spectra ($R \sim 45$) have been chosen and built for the high-contrast Lyot Project (Hinkley et al. 2008) and Gemini Planet Imager projects. For the latter, a 200×200 lenslet yields $\sim 40,000$ spectra with each spectrum having only 18 spectral channels, but covering a full 20% bandpass.

Advantages of a lenslet-based IFS include excellent image quality because only optics in front of the lenslet array can affect the wavefront error. Lenslet arrays are also manufactured

in very large formats allowing nearly infinite expansion of the field of view. These commercial arrays also have no internal surfaces, approximately 98% fill factors and high transmission. It is generally easy to change the plate scale of a lenslet-based IFS because they are essentially like detector pixels in the sense that reimaging optics in front of them can be used to magnify the field, and their fast focal ratios make them relatively insensitive to the input focal ratio. Among the disadvantages of a lenslet-based IFS is the complexity of the data. Thousands of very long spectra are scattered over the detector in a complex pattern which must be mapped and individually calibrated. In principle, this is comparable to reductions of multislit data and has the advantage that unlike slits which can be repositioned, the lenslets have a fixed geometry. But with dense packing of spectra, sophisticated algorithms are often required to extract all of the spectra and reassemble the final cube. Lenslet array spectrographs must put at least some pixel gap between neighboring spectra. So the total number of elements in the final data cube will be at most half the number of original detector pixels. The other methods suffer similar pixel issues but for different reasons. A challenging problem with a lenslet IFS which must be considered carefully is crosstalk between neighboring spectra on the detector. Since they are staggered in wavelength space, a bright line from one spectrum could blend into a neighboring spectrum at a different wavelength creating in effect a ghost line. With good optics and knowledge of the packing geometry, this contamination can be eliminated or at least understood, but it does put pressure on the optical design, detector utilization, and data reduction algorithms.

A second technique, and perhaps the most commonly implemented one, is an image slicer. A parallel set of narrow mirrors is used to divert different parts of the focal plane into different directions. A second set of mirrors redirects these portions back into something close to a slit-like arrangement (a pseudo slit), which is then fed into a traditional spectrograph. The first cryogenic version of such an instrument for astronomy was the 3D Spectrograph (Krabbe et al. 1997). Other early successful instruments are PIFS at Palomar (Murphy et al. 1999) and SPIFFI (Tecza et al. 1998) for the VLT. The latter can be fed by an adaptive optics system where the combination is called SINFONI. For diffraction-limited spectrographs, there are some significant disadvantages to slicer-based spectrographs. The first is the relatively large size of the slicing mirrors which are usually ~ 1 mm wide. For small plate scales, this forces the design to long focal ratios which can be cumbersome. The second disadvantage is that diffraction off the individual slits causes the pupil on the grating to grow through the same process that causes pupil diffraction in lenslet-based designs. But perhaps the most important difficulty is with image quality. Particularly in the long axis of the slit, all optical elements, including the grating, affect the wavefront. However, if the wavefront error issue can be overcome, as some recent designs for SNAP and JWST indicate, then a slicer offers two very major advantages. First, it is more efficient in utilizing detector pixels (maybe by a factor of 2 or more). This translates directly into the number of spatial pixels or spectral bandwidth of the spectrograph. Secondly, each source is spread over fewer pixels which mitigates the effect of detector noise and thus increases sensitivity. In a simple comparison, the sensitivity gain could be as much as 40% at detector limited flux levels.

The third technique uses a bundle of fiber optic cables arranged to sample a region of the focal plane. This bundle is then reorganized into a linear arrangement in order to produce a slit-like pattern to feed the spectrograph. Optical fibers require a relative thick cladding and so the fill factor is usually quite low. Fibers also have fast output focal ratios making the collimator and camera optics more difficult. Fiber-based IFS systems have been used for relatively bright targets. Early examples of such spectrographs include 2D-FIS and HEXAFLEX (Garcia et al. 1994) and INTEGRAL at WHT (Arribas et al. 1998). Many of the inherent problems with fibers

can be overcome by using lenslets coupled to the input and sometimes the output of the bundles to give them higher fill factors in the focal plane. Focal ratio degradation within the fibers and coupling losses at the lenslet-fiber boundary still need to be considered.

2.5 Polarimeter Systems

To measure polarization properties, such as the fraction polarized, the direction of vibration, and the handedness of rotation, all polarimeters “convert” the polarization information into brightness modulations which are directly measurable with an electronic detector. Polarimeters can be created from camera and spectrometer designs by adding a polarization modulator. A typical approach is to construct the polarization modulator in two parts. A retardation device comes first to introduce a known and controllable phase shift into the beam, and this is followed by a fixed polarizer (also called an analyzer) that only allows one plane of polarization to pass unhindered and reduces others by the factor $\cos^2 \theta$, where θ is the angle between the polarizer’s axis and the plane of polarization in the beam. The intensity transmitted by the analyzer is therefore modulated by the action of the phase retardation device.

Linear polarization is described by three parameters: intensity (I), degree (or fraction) of linear polarization (p), and the direction of the (fixed) plane of vibration projected on the sky (θ). Circular polarization is similarly described by three parameters: intensity (I), degree of circular polarization (q), and handedness of the rotation of the electric vector (+ or -). A more convenient way to express polarization information is to use the four Stokes parameters (I, Q, U, V). These quantities are phenomenological, that is, they are more directly related to actual measurements. The Stokes parameters are easily related to the amplitudes (E_x, E_y) of the electric vector in two orthogonal directions and to the phase difference (δ) between the two components (e.g., Clarke and Grainger 1971; Tinbergen 1996). The degree of linear and circular polarization is given by

$$p = \frac{[Q^2 + U^2]^{\frac{1}{2}}}{I}, q = \pm \frac{V}{I} \quad (12.18)$$

and the direction of vibration of the linearly polarized part is given by

$$\tan 2\theta = \frac{U}{Q} \quad (12.19)$$

and it follows that

$$\begin{aligned} Q &= Ip \cos 2\theta \\ U &= Ip \sin 2\theta \\ V &= Iq \end{aligned} \quad (12.20)$$

The intensity of light transmitted by a retarder of retardance τ at angle ψ followed by a perfect polarizer with principal plane at $\phi = 0^\circ$ or $\phi = 90^\circ$ (upper/lower signs, respectively) is given by

$$I' = \frac{1}{2} [I \pm Q(G + H \cos 4\psi) + \pm UH \sin 4\psi \pm V \sin \tau \sin 2\psi] \quad (12.21)$$

where


$$G = \frac{1}{2}(1 + \cos \tau), \quad H = \frac{1}{2}(1 - \cos \tau), \quad \tau = \frac{2\pi}{\lambda} \delta \quad (12.22)$$

There are several special cases and multiple ways to solve these equations for the Stokes parameters. Although harder to measure than other properties, polarization from astronomical

sources can be detected from X-rays to radio waves. Because it may contain information about the formation of the early universe, one important area of interest is the polarization of the cosmic microwave background.

2.6 Interferometers

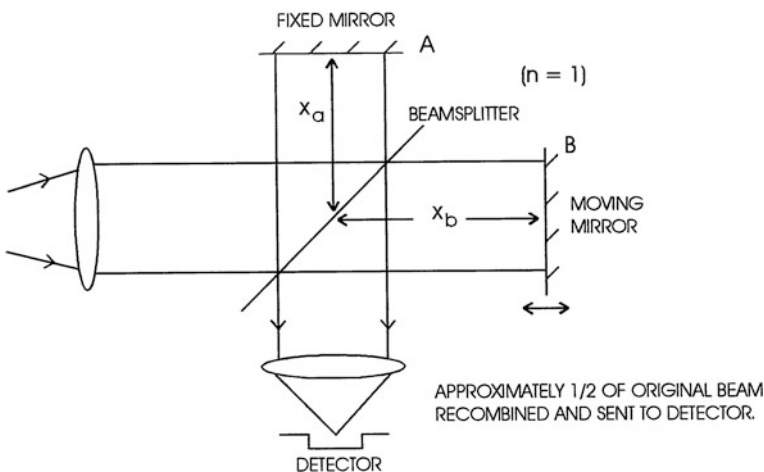
Interferometer techniques in astronomy are applied in two different ways, one as a collection method and the other as a detection method. Combining the light collected by many widely separated telescopes overcomes the diffraction limit of an individual telescope. Single-aperture telescopes can be equipped with interferometer equipment for specific detection purposes. Several types of detection interferometers have been used for spectroscopy, such as the Fourier transform spectrometer (FTS) which is a scanning Michelson interferometer and the Fabry-Perot interferometer which is an imaging spectrometer.


A typical FTS is shown in  Fig. 12-7. For a collimated monochromatic beam, the intensity at the detector is determined by the “path difference” $\Delta x = 2(x_b - x_a)$, where x_a refers to the arm containing the fixed mirror A and x_b is the distance to the scanning mirror B. The phase difference is given by $k\Delta x$ where $k = 2\pi/\lambda$. The fraction of the incident beam in the output is given by

$$T(k, \Delta x) = \frac{1}{2}[1 + \cos(2k\Delta x)] \quad (12.23)$$

from which it follows that $T = 1$ when the combining beams are in phase and $T = 0$ when they are 180° out of phase. Given an incident beam whose spectrum is $I(k)$, the signal F measured in the output is

$$F(\Delta x) = c \int I(k) T(k, \Delta x) dk = \text{constant} + \frac{c}{2} \int I(k) \cos(2k\Delta x) dk \quad (12.24)$$



 Fig. 12-7

The principle of the scanning Michelson interferometer is shown. As the mirror is scanned, the intensity recorded by the detector is modulated to produce an interferogram. The spectrum can be extracted by an inverse Fourier transform (From McLean 2008)

where c is a constant. The measured signal $F(\Delta x)$ is called the interferogram, and the last integral is the Fourier cosine transform of the spectrum. Therefore, the transform of the interferogram is $I(k)$.

The Fabry-Perot interferometer is an imaging spectrometer formed by placing a device called an “etalon” in the collimated beam of a typical camera system. One arrangement is shown in **Fig. 12-8**. The etalon consists of two plane parallel plates with thin, highly reflective coatings on their inner faces. The plates are in near contact but separated by a distance d . Assuming that the refractive index of the medium in the gap is n (usually $n = 1$) and θ is the angle of incidence of a ray on the etalon (usually very small), then multiple reflections and destructive interference within the gap occur and the wavelengths transmitted with maximum intensity obey the relation

$$m\lambda = 2nd \cos \theta \tag{12.25}$$

For monochromatic light, the image is a set of concentric rings. To ensure that a sufficiently narrow band of light passes through the system, it is necessary to “prefilter” the light. This can be done with a very narrow band interference filter. Usually, a circular aperture isolates the central order which has an angular diameter $\delta\beta = \sqrt{(8/R)}$ and the free spectral range is given by

$$\Delta\lambda_{\text{FSP}} = \frac{\lambda}{m} = \frac{\lambda^2}{2nd} \tag{12.26}$$

The resolving power ($R = \lambda/\delta\lambda$) is

$$R = \frac{2Fnd}{\lambda} \tag{12.27}$$

where $F (= \Delta\lambda_{\text{FSP}}/\delta\lambda)$ is called the “finesse” of the etalon, which is a measure of the plate quality and the reflectance (r) of the coatings; $F = \pi\sqrt{r}/(1 - r)$ and typical values are 30–50. Defining $\delta = (2\pi/\lambda) (2nd \cos \theta)$, the transmitted intensity is $I(\delta) = I(0)/[1 + (2F/\pi)^2 \sin^2 (\delta/2)]$. One application of Fabry-Perot etalons is the Taurus Tunable Filter (Bland-Hawthorn and Kedziora-Chudczek 2003) which allows wide-field narrow-band imaging with a CCD.

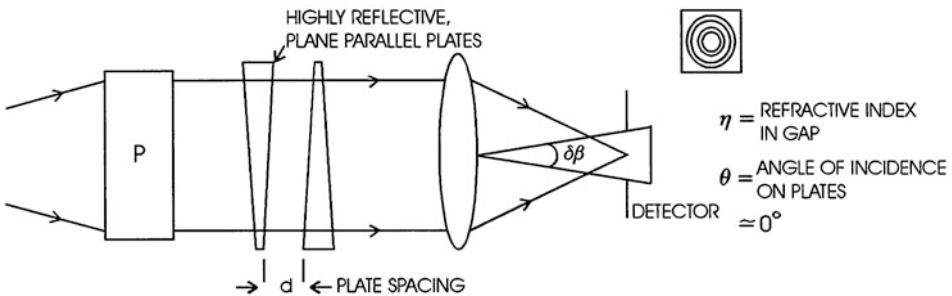


Fig. 12-8 One typical layout for a Fabry-Perot interferometer is shown. The device P is used to narrow the range of wavelengths fed to the etalon (From McLean 2008)

3 Detectors and Materials

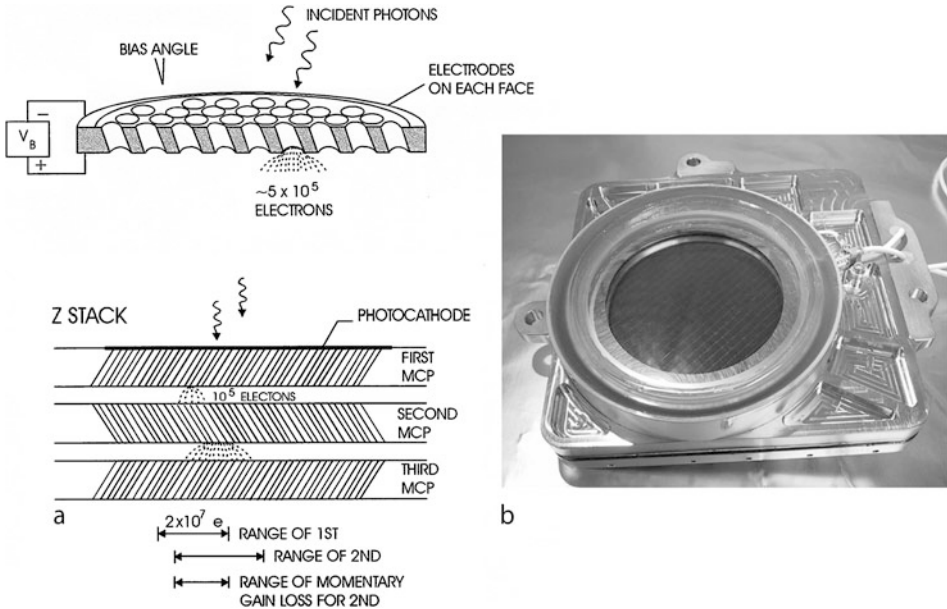
3.1 Classification of Detectors

High-energy photons are usually detected with particle detectors, but for lower energy photons, detectors are generally grouped into three broad classes:

1. *Photon detectors* in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma rays to the far-infrared.
2. *Thermal detectors* in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical conductivity; thermal detectors have a broad spectral response but are often used for infrared and submillimeter detection.
3. *Coherent detectors* in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

Photon detectors can be subdivided into (◆ 12.1) photoemission devices employing the external photoelectric effect in which the photon causes a charge carrier (electron) to be ejected from the material and (◆ 12.2) photoabsorption devices that use the internal photoelectric effect in a semiconductor to free a charge carrier within the material. The most well-known detector in the *photoemission* category is the photocathode of a photomultiplier tube (PMT) in which an electron is emitted from the photocathode surface and subsequently amplified by a cascade of impacts with secondary surfaces before being detected as a charge pulse.

Photoemissive materials can provide excellent detectors far into the ultraviolet. Most importantly, it is possible to create ultraviolet imaging devices based on this process. For example, long, narrow curved tubes or “microchannels” of lead oxide (PbO) can perform the same function as the secondary surfaces in a PMT resulting in a large pulse of electrons emerging from the end provided there is a potential gradient. As shown in ◆ Fig. 12-9a, such channels can be packaged very close together (like straws in a box) to make a two-dimensional array called a microchannel plate (MCP). MCPs are used across most of the UV and have been the detector of choice for almost all major UV missions. For example, the GALEX spacecraft employs two MCPs (◆ Fig. 12-9b) fabricated by the experimental astrophysics group at the University of California Berkeley, Space Sciences Laboratory. This group has been responsible for the development of most of the UV detectors in space. For example, there were seven on EUVE, four in two different instruments on SOHO, two in FUSE and a custom-designed MCP for COS (Hubble), to give just a partial list (Siegmond et al. 2007). The length of the microchannel is typically 50–100 times the diameter of the channel, which implies a large surface-to-volume ratio and the tendency to trap residual gas unless exceptional measures on cleanliness and plate conditioning are employed. Because MCPs are operated at potentials of a few thousand volts, residual gases can lead to destructive discharges. The channels have diameters ranging from 5 to 25 μm on 10 to 40 μm centers, and plates with active areas as large as $100 \times 100 \text{ mm}^2$ are available; the GALEX detectors are 75 mm in diameter with an active area of 68 mm. The response of the MCP is a strong function of the angle of incidence of the photons. Photocathodes can be placed on the top face or on a window in proximity focus immediately above the MCP. Materials with



■ Fig. 12-9

(a) The structure of a microchannel plate (MCP) device. (b) The GALEX MCP detector (Credit: Experimental Astrophysics Group, UC Berkeley (From McLean 2008))

large work functions such as CsI, CsTe, and KBr have good UV quantum efficiency but very low response to visible photons. More recently, gallium nitride (GaN), which has a band gap of 3.4 eV, has been added to the list of photocathodes available for UV astronomy. Microchannel plate detectors use a variety of anode structures. One of the simplest is a single resistive anode in which the location of the event is determined by the amount of charge or current “divided” between amplifiers attached to the corners. Other anode structures include the wedge and strip anode, the spiral anode, and the delay Line, each of which is described as “continuous” anodes. It is also possible to utilize “discrete” anode structures at the expense of many more amplifiers and encode the event location through direct detection. One such system is called the Multi-Anode Microchannel Array (MAMA). MAMA detectors with $1,000 \times 1,000$ pixels were constructed for Hubble’s STIS and ACS instruments.

Detectors employing *photoabsorption* make up the largest category. There are many possible outcomes, including chemical change as in photography, but absorption in semiconductor devices is the important one for astronomy. There are essentially two basic types of interactions, the photoconduction effect and the photovoltaic (or photodiode) effect. The photoconductor is composed of a single uniform semiconductor material in which the conductance is changed by the creation of free charge carriers in the material when photons are absorbed. There is usually always an external applied electric field. In the photodiode (or photojunction), internal electric fields and potential barriers are created by suitable junctions between different materials or deliberate variations in the electrical properties of the material so that photogenerated carriers in these regions respond to those fields.

3.2 Semiconductors

When individual atoms come close together to form a solid crystal, electrons in the outermost orbits or upper energy levels of adjacent atoms interact to bind the atoms together. Because of the very strong interaction between these outer or “valence” electrons, the upper energy levels are drastically altered, with the result that the outer electrons are shared between the different atomic nuclei. In fact, the energy levels are spread out into a “band.” The lowest band of energies, corresponding to all the innermost orbits of the electrons, is filled with electrons because there is one electron for each atom. This band of filled energy levels is called the “valence band.” Conversely, the upper energy band is empty of electrons because it is composed of the combined unoccupied higher energy levels of the individual atoms in the crystal. It is called the “conduction band” for reasons that will become apparent. Thus, the individual atoms have a gap in energy between the inner filled levels and the outer unoccupied levels. The energy region between the valence band and the conduction band in the crystal must be a “forbidden energy gap” (E_G). The crystal must be pure and contain atoms of only one kind; otherwise, additional energy levels corresponding to those atoms will be formed. More importantly, the periodic or repetitive crystalline structure must be unbroken to avoid distortions in the energy levels caused by abnormal sharing of electrons. In practice, both of these conditions are violated in real crystals, and such departures from the simplified model contribute to degraded performance.

In metals, the valence and conduction bands overlap, and so any of the many valence electrons are free to roam throughout the solid to conduct electricity and heat and to move in response to the force of an electric field. Insulating materials, on the other hand, have a highly ordered structure and a very wide forbidden energy gap. The conduction band is totally empty of electrons and so cannot contribute to an electrical current flow. Electrons in the completely filled valence band cannot move in response to an electric field because every nearby orbit is occupied. In a semiconductor, a few electrons can be elevated from the valence band to the conduction band across the forbidden gap merely by absorbing heat energy from the random, microscopic, jostling motions of the crystal structure at normal “room” temperature. Thermal energy is given approximately by

$$E_{\text{th}}(\text{eV}) = kT = 0.026(T/300)\text{eV} \quad (12.28)$$

where k is Boltzmann’s constant and T is the absolute temperature. At room temperature ($T = 300 \text{ K}$), the thermal energy is quite small at 0.026 electron volts. Electrons promoted to the conduction band can then conduct electricity, that is, they are free to move under the influence of an electric force field. Interestingly, the corresponding vacancies or “holes” left in the valence band allow it to contribute to electrical conductivity as well because there is now somewhere for electrons in adjacent atoms to go; descriptions of solid-state devices therefore refer to “electron-hole” pairs.

Most semiconductor crystals have band gap energies around 1 eV, but the range is from almost 0 to about 3.5 eV. Visible light photons have energies around 2.25 eV (for 550 nm). As the number of electrons which can be promoted to the conduction band by absorbing heat will vary with the temperature of the crystal, typically as $\exp(-E_G/2kT)$, those semiconductors with larger band gaps are preferred because transistors and other devices made from them will be less sensitive to environmental changes. If the semiconductor is cooled to a low enough temperature, random elevation of valence electrons to the conduction band can be virtually eliminated. The primary semiconductors, like silicon and germanium, belong to the “fourth

column” elements of the periodic table, which also includes carbon. Each of these elements has four valence electrons. Compounds of elements on either side of the fourth column can be formed, and these alloys will also have semiconductor properties. For example, gallium arsenide (GaAs) and indium antimonide (InSb) are III–V (or “three–five”) compounds, and mercury-cadmium telluride (HgCdTe) is one possible II–VI (or two–six) compound.


When a photon is absorbed in the crystalline structure of silicon, its energy is transferred to a negatively charged electron, the photoelectron, which is then displaced from its normal location in the valence band into the conduction band. When the electron reaches the conduction band, it can migrate through the crystal. Migration can be stimulated and controlled by applying an electric field to the silicon crystal by means of small metal plates called “electrodes” or “gates” connected to a voltage source. For each semiconductor, there is a wavelength of light beyond which the material is insensitive to light because the photons are not energetic enough to overcome the forbidden energy gap (E_G) in the crystal. The cutoff wavelength is given by

$$\lambda_c = \frac{hc}{E_G} \quad (12.29)$$

where h is Planck’s constant and c is the speed of light; $hc = 1.24$ for wavelengths in microns and energy in electron volts.

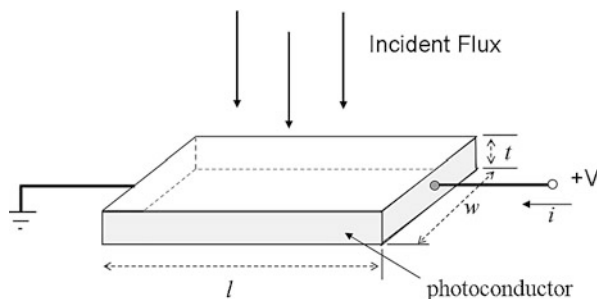
All of the materials mentioned so far are “intrinsic” semiconductors because each has a well-defined band gap intrinsic to the material. It is also possible to create an “extrinsic” semiconductor in which impurity atoms produce intermediate energy levels within the forbidden gap. For example, when silicon atoms in the crystal structure are deliberately replaced with other atoms, the semiconductor is said to be “doped.” If the impurity atom has more valence electrons than the semiconductor, then it will donate these *negative* charges to the conduction band; such a material is called *n-type*. Conversely, if the impurity atom has fewer valence electrons than the semiconductor, then a *positively* charged hole is left in the valence band ready to accept any available electrons; this material is called *p-type*. In *p-type* material, there is an excess of holes and so electrons are said to be the minority carriers of charge, whereas the opposite is true for *n-type* material. Because of the much lower transition energies, extrinsic semiconductors are used in far-infrared photon detection.

3.3 Photoconductors

This is the simplest application of a semiconductor for detection of photons. Photons are absorbed and create electron-hole pairs. If the material is extrinsic rather than intrinsic, then E_i must be substituted for E_G . Also, for extrinsic materials, there are limits on solubility of the dopants and high concentrations introduce unwanted conductivity modes. In practice, both electrons and holes contribute to the photocurrent, but it is usually the electrons that dominate. The main parameters in the construction and operation of a photoconductor are shown in  Fig. 12-10. For this discussion, it is assumed that the detector has been cooled to eliminate thermally generated charges.

The average photocurrent (I) between the terminals that is generated by an incident flux with power P (watts) is given by

$$I = (e\eta P/h\nu)(v\tau/l) \quad (12.30)$$



■ Fig. 12-10

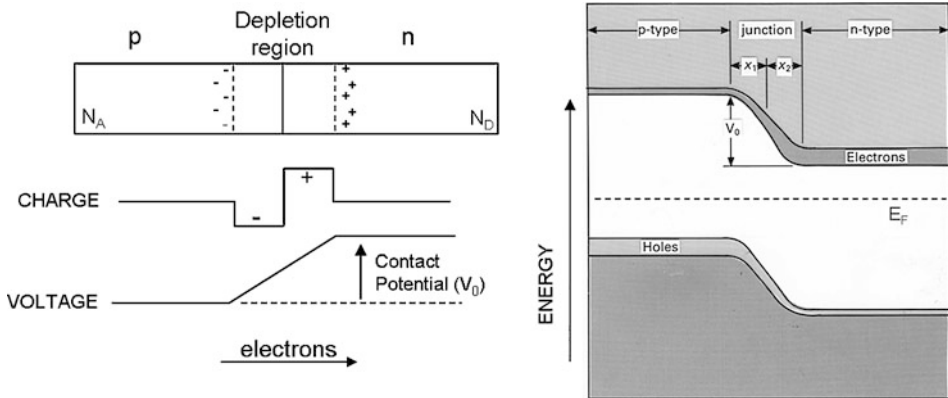
Construction and operation of a semiconductor used in photoconduction mode (From McLean 2008)

In this expression, η is the quantum efficiency and $P/h\nu$ is just the photon arrival rate. The quantity τ is called the mean carrier lifetime and measures how long the photogenerated charge exists before recombination. Values are usually less than to much less than a few milliseconds but depend on doping and temperature. The average charge carrier velocity is v , which is related to the applied electric field across the photoconductor $E = V/l$ by $v = \mu E$ where μ is called the mobility of the charge carrier. Thus, l/v is the transit time across the device from one terminal to the other, and the quantity $G = v\tau/l$ is just the ratio of mean carrier lifetime to transit time. It is known as the “photoconductive gain.” The response (S) of the detector (in amps per watt or volts per watt) is just I/P or V/RP where V is the bias voltage across the photoconductor and the resistance R due to the photocurrent is $l/\sigma A$ and the conductivity $\sigma = ne\mu$, where n is the average density of carriers. It follows that $S = (e\eta G/hc)\lambda$. Finally, the root mean square noise for a photoconductor is given by $\sqrt{(4eGIB)}$ where B is the electrical bandwidth of the measurement.

3.4 Photodiodes

Junctions between p- and n-type regions are used many times in semiconductor structures to produce different devices. One such device is the photodiode. When a pn junction is formed as shown in [Fig. 12-11](#), electrons from the n region tend to diffuse into the p region near the junction and fill up some of the positively ionized states or holes in the valence band, thus making that p-type region more negative than it was. Similarly, the diffusion of holes from the p to the n side leads to an increasingly more positive electrical potential.

A narrow region forms on either side of the junction in which the majority charge carriers are “depleted” relative to their concentrations well away from the junction. As the concentration of electrons in the n-type material is usually very much larger than in the p-type material, the flow of electrons would tend to be one way were it not for the fact that the diffusion process itself begins to build up an electrostatic potential barrier which restrains the flow of electrons from the n-type region; the buildup of electrons on the p side makes it negatively charged which starts to repel further diffusion. The magnitude of this potential barrier (V_0) depends on the impurity concentrations, that is, on the number of donor electrons at the junction that are available for transfer to nearby acceptor levels and is just equal to the required shift of the energy bands



■ Fig. 12-11

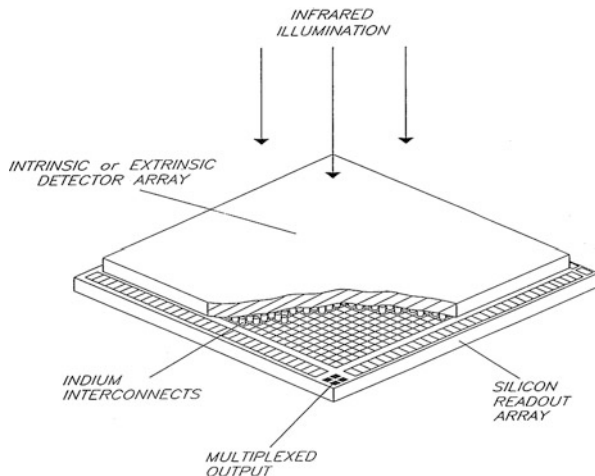
Formation of a pn junction between p-doped and n-doped materials results in a region depleted of carriers and the creation of a potential barrier (From McLean 2008)

needed to ensure that the Fermi level (E_F) remains constant throughout the crystal. The Fermi level is the energy at which there is a 50/50 chance of the corresponding electron energy state or orbit being occupied by an electron. For an intrinsic semiconductor, E_F lies halfway between the valence and conduction bands, whereas for an n-type doped semiconductor, the Fermi level moves up toward the conduction band and conversely p-type doping lowers the Fermi level.

When a positive voltage is applied to the p side of the junction, it will tend to counteract or reduce the built-in potential barrier and attract more electrons across the junction, whereas a negative voltage on the p side will enhance the internal barrier and increase the width of the depletion region; these conditions are called “forward” and “reversed” bias, respectively. Therefore, on one side of a pn junction, there is a region which is more negative, and on the other side, there is a region which is more positive than elsewhere in the crystal. When light of the correct wavelength is absorbed near the junction, an electron-hole pair is created and the potential difference across the junction sweeps the pair apart before they can recombine. Electrons are drawn toward the region of greatest positive potential buried in the n-type layer which therefore behaves like a charge storage capacitor. Of course, as more electrons accumulate, the positive potential is progressively weakened. In the photodiode, an electron-hole pair is created within the depletion region by the absorption of a photon, and the charge carriers are immediately separated by the electric field across the junction. The current due to an incident photon flux (signal and background) of power P is just $I = e\eta P/h\nu$, and the root mean square noise is given by $\sqrt{(2eIB)}$ where B is the electrical frequency bandwidth of the measurement. Comparing these results to the photoconductor shows that $G = 1$ for the photodiode and the noise is less by a factor of $\sqrt{2}$ because recombination does not occur in the depletion region.

3.5 Applications to CCDs and IR Arrays

Most infrared arrays are either photodiodes or photoconductors. Infrared arrays employ a hybrid construction. They are like a sandwich (► Fig. 12-12) in which the upper slab is the IR sensor (e.g., InSb, HgCdTe; Si:As, Ge:Ga), and the lower slab is a silicon readout device.



■ Fig. 12-12

The “hybrid” structure of infrared array devices: the two slabs are separated by a grid of tiny indium bumps that remain soft at cryogenic temperatures (From McLean 2008)

The infrared layer is a tightly packed grid of individual pixels with minimum dead space between them. Both slabs are provided with a grid of electrical connections in the form of tiny raised sections – referred to as “bumps” – of an electrical conductor called indium; indium remains soft at low temperatures. The two slabs are literally pressed together to enable the indium bumps to mate. A microscopic array of “switches” made from Metal–Oxide–Semiconductor Field-Effect Transistors (MOSFETs) is used to access the signal from each IR pixel (whether photodiode or photoconductor). Charge storage may occur on the junction capacitance of the IR sensor itself (in the case of a photodiode) or on a separate storage capacitor associated with the silicon circuitry.

The entire structure is often called a Focal Plane Array (FPA) or a Sensor Chip Assembly (SCA), and the silicon readout integrated circuit part by itself is called a ROIC or mux. In the infrared array, stored charge is read out *directly* from each pixel in turn using a source follower transistor which permits nondestructive sampling of the signal voltage. The charge does not “pass through” any other pixels in the array. This is quite different from the CCD discussed below.

Silicon CCDs (charge-coupled devices) are used widely in astronomy from X-ray to near-infrared wavelengths. The CCD is an array of individual pixels each one of which can absorb photons and utilize the energy to release an electron within the semiconductor. To confine the electron within a pixel requires a special electrostatic field to attract the charged electron to a specific spot. Creating a storage region capable of holding many charges is more complicated. It is achieved by applying metal electrodes to the semiconductor silicon together with a thin (100 nm = 0.1 μm) separation layer made from silicon dioxide, which is an electrical insulator. The resulting structure behaves like a parallel plate capacitor which can therefore store electrical charge. It is called an MOS (metal-oxide-semiconductor) structure.

An electric field is generated inside the silicon slab by the voltage applied to the metal electrode. If the material is p-type, then a positive voltage on the metal gate will repel the holes which

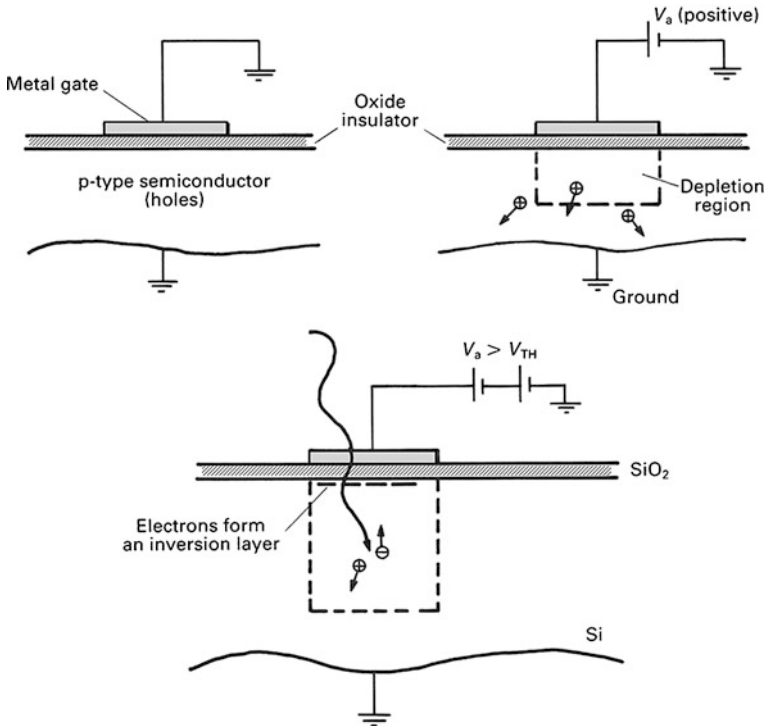


Fig. 12-13

Development of a single metal-oxide-semiconductor (MOS) storage well, the basic element in a CCD, is shown for different applied gate voltages (From McLean 2008)

are in the majority and sweep out a region depleted of charge carriers. These conditions are illustrated in Fig. 12-13. When a photon is absorbed in this region, it produces an electron-hole pair, but the hole is driven out of the depletion region and the electron is attracted toward the positively charged electrode. The MOS capacitor is the combination of two parallel plate capacitors, namely, the oxide capacitor and the silicon depletion region capacitor, and therefore, the capacitance is proportional to the area of the plates (electrodes) and inversely proportional to their separation. As the voltage on the plate can be controlled, then the depletion width can be increased or decreased, and so the capacity to store charge can also be controlled. The depletion region is an electrostatic “potential well” into which many photogenerated charges can be collected. Typically, the number of electrons stored is just $Q = CV/e$, where e is the charge on the electron (1.6×10^{-19} C), V is the effective voltage, and the capacitance C is given by the “parallel-plate” formula $C = A\kappa\epsilon_0/d$ in which A is the area of the pixel or gate electrode, d is the thickness of the region, κ is the dielectric constant of the SiO_2 insulator (~ 3.9), and ϵ_0 is the permittivity of free space (8.85×10^{-12} farad/m). As the voltage on the electrode increases, the “depth” of the well increases; other ways are needed to create sidewalls to the well. Eventually, at a certain “threshold” voltage, even the minority charge carriers due to impurities, electrons from a p-type semiconductor, will be drawn to the electrode.

The unique feature of the CCD that gives it its name is the way in which the photogenerated charge, and hence the image of the scene, is extracted from the MOS storage and detection site. It is called “charge coupling.” To transfer charge from under one electrode to the area below an adjacent electrode, raise the voltage on the adjacent electrode to the same value as the first one. This is like lowering the floor of the adjacent potential well. The charges will now flow, like water, and be shared between both regions. Transfer can be in either direction, and by connecting sets of electrodes together, the entire charge stored on the two-dimensional imaging area can be moved simultaneously in that direction. When the voltage on the original electrode is reduced to zero volts, transfer is complete because the collapse of the storage well pushes any remaining charges across to the new electrode. Because it takes three electrodes to define 1 pixel, three of the above transfers are required to move the two-dimensional charge pattern by one whole pixel step along the direction at right angles to the electrode strips. The process of raising and lowering the voltage can be repeated over and over. These drive or “clock” pulses can be described in a diagram called a timing waveform. Finally, there is another set of electrodes at right angles to the first to enable charges to be moved along that row to the output amplifier where the charge is converted to a voltage that can be further amplified and digitized by an analog-to-digital converter. Details and applications can be found in Janesick (2001) and McLean (2008).

The digital signals recorded by the detector system – usually called Data Numbers (or DN) or sometimes Analog-to-Digital Units (ADU) – must be turned back into microvolts and then into electrons and finally to photons in order to calibrate the system. The relation between DN and microvolts at the CCD or infrared array output depends on the “gain” of the amplifiers in the system, and conversion between microvolts at the output and an equivalent charge in electrons requires knowledge of the capacitance (C) of the output node of the on-chip amplifier. Counts or DN recorded in a given time by the camera system are linearly related to the numbers of electrons in the charge packets by

$$S = \frac{(N_e + N_D)}{g} + b \quad (12.31)$$

where S is the recorded output signal in Data Numbers (or counts), N_e is the number of electrons in the charge packet (ηN_p), and the system photon transfer gain factor is g electrons/DN; b is the electronic offset or bias level (in DN) for an empty charge packet and N_d is the residual dark current signal still present after cooling the device. Both the bias (b) and the dark current (N_d) can be determined from measurements without illumination and can therefore be subtracted. There are two ways to derive the transfer factor g in electrons/DN, either by calculation, knowing the overall amplifier gain and the capacitance of the CCD/IR array, or by a series of observations of a uniformly illuminated scene at different brightness levels.

Let V_{fs} be the full scale voltage swing allowed on the A/D unit, and n be the number of bits to which the A/D can digitize. The full scale range is therefore subdivided into 2^n parts, the smallest part – the least significant bit or LSB – is simply 1 DN. Thus, the voltage corresponding to 1 DN at the A/D unit is $V_{fs}/2^n$; as an example, suppose the full scale voltage is 10 V and the A/D is 16 bits then 2^n is 65,536 and so the ratio is 0.0001525 V, or 152.5 μ V at the A/D is equivalent to 1 DN. Similarly, for a 14 bit A/D, the range is 16,384 and 1 DN corresponds to 610 μ V. To get the number of microvolts corresponding to 1 DN at the CCD rather than at the A/D, divide the number derived above by the total gain product A_g of all the amplifiers in the system; usually this means the on-chip amplifier (A_{SF}), the preamplifier (A_{pre}), and a postamplifier (A_{post}). To convert this number of microvolts to an equivalent charge of electrons, multiply by the CCD

capacitance (C) and divide by the value of the charge on the electron (e). Therefore,

$$g = \frac{V_{fs}C}{2^n A_g e} \quad (12.32)$$

where $e = 1.6 \times 10^{-19}$ Coulombs.

Assuming that there are only two sources of noise when imaging a uniformly illuminated flat field, photon noise from the signal and readout noise from the detector, these two noise sources should be independent and random. Therefore, adding them together in quadrature gives the total noise

$$(\text{noise})^2 = p^2 + R^2 \quad (12.33)$$

This expression applies to photoelectrons and not to counts (DN). The measured quantities, the mean signal (S_M) and its variance (V_M), are in DN. To convert from electrons to DN in (12.33), divide each noise term by g (electrons/DN) to give

$$\left(\frac{\text{noise}}{g}\right)^2 = \left(\frac{p}{g}\right)^2 + \left(\frac{R}{g}\right)^2 \quad (12.34)$$

The left-hand side is now exactly V_M , the observed variance in DN. Also, the mean number of photoelectrons is $g(e^-/\text{DN})S_M(\text{DN})$ or gS_M , and the photoelectron noise (p) on this number is simply the $\sqrt{gS_M}$ for Poisson statistics, so $p^2 = gS_M$. Hence, (12.34) becomes

$$V_M = \frac{1}{g}S_M + \left(\frac{R}{g}\right)^2 \quad (12.35)$$

Equation 12.35 is just the equation of a straight line in a signal-variance plot of $y = V_M$ and $x = S_M$. Plotting these observed quantities (noise-squared and signal) as the illumination changes will yield a straight-line of gradient (slope) $m = 1/g$ with the value of the intercept on the V_M axis when $S_M = 0$ giving $(R/g)^2$ which yields R as g is known from the slope.

3.6 Detectors for High Energy

Gamma rays cannot be readily focused, even using the grazing incidence mirrors that operate successfully for X-rays. In addition, gamma rays are different from other photons in that they may not be fully absorbed. In fact, there are three regimes: total absorption by the photoelectric effect, Compton scattering, and pair production. Pair production dominates above 10 MeV. Pair conversion telescopes use devices like spark chambers and silicon strip detectors to track the particles produced. Sheets of silicon detectors can be stacked in towers to produce the solid-state equivalent of a spark chamber. An incident gamma-ray photon is forced to pair-produce by a plate of high atomic weight material, and the electrons and positrons from the conversion cause ionization in the silicon strip. Another useful device is the proportional counter. Cylinders of solid germanium with a central axial node and surrounded by a cylindrical cathode. The incoming gamma ray creates ion-electron pairs, and the electrons are attracted to the anode such that the number released is proportional to the gamma ray's energy. These kinds of units can form a "pixel" in a large array of germanium detectors. Another semiconductor material with good stopping power is cadmium zinc telluride, sometimes called simply CZT. When ionizing radiation interacts with the CZT crystal, electron-hole pairs are created in proportion to the energy of the incoming photon. CZT elements can be tightly packed to make an

array detector for gamma rays. A large array of CZT elements is used, for example, in the Burst Alert Telescope (BAT) on the *Swift* satellite (Gehrels et al. 2004). Finally, Earth's atmosphere forces high-energy gamma rays to pair-convert, and the resulting shower of particles produce Cherenkov light which can be detected by wide-field telescopes equipped with "cameras" using hundreds of PMTs as the pixel elements.

3.7 Thermal Detectors

In the class of thermal detectors, there is only one type that is used widely in astronomy and that is the "bolometer," which is described in more detail in [Chap. 14](#). Semiconductor bolometers, based on silicon or germanium, are well developed for far-infrared and submillimeter astronomy. Essentially, a bolometer consists of a sensitive thermometer and a high cross-section absorber that absorbs almost all of the incident radiation falling on it, that is, QE (η) \sim 100%. The absorber has a heat capacity of C joules per kelvin. The thermometer and absorber are connected by a weak thermal link to a heat sink at a low temperature, and the conductivity of this link is G watts per kelvin. If the detector element of the bolometer absorbs an amount of energy $E = \eta P \Delta t / h\nu$ in a time interval Δt from a source with power P, that energy is converted to heat which raises the temperature by an amount $\Delta T = T - T_0 = E/C$ above that of a heat sink at T_0 . The temperature rise decays exponentially as power in the absorber flows out to the heat sink via the weak link and the time constant is $\tau = C/G$. Temperature rise is proportional to the absorbed energy. In the classical circuit, a constant bias current, generated by the bias supply and a load resistor, flows through the bolometer. Provided that the bias power remains constant

$$T = T_0 + (P + P_{\text{bias}})/G \quad (12.36)$$

and the temperature rise causes a change in bolometer resistance, and consequently a change in the voltage across it which can be measured. Differential on/off source measurements are normally required to remove bias levels. A more detailed account is given in Rieke (2003). Arrays of bolometers now exist and will be described in the chapter on far-infrared detectors.

3.8 Coherent Detectors

A coherent detector or "receiver" is any device that directly responds to the electric field of the wave. The most important form of receiver is the heterodyne or superheterodyne which functions by mixing signals of different frequencies and detecting a signal at the difference or "beat" frequency between the original two frequencies. Depending on the frequency of the incoming wave, the electric field may be converted to an electrical signal which is then amplified before being mixed with a local oscillator. For frequencies below 1 GHz, cryogenic transistor preamplifiers are used. From 1 to 40 GHz, FET, parametric and maser amplifiers are employed, but above 40 GHz, the mixer must precede the preamplifier in order to reduce the frequency before amplification. The local oscillator (LO) produces a strong signal with a frequency that is close to but different from the signal frequency. The beat frequency, or intermediate frequency (IF), is $\nu_{\text{IF}} = \nu_{\text{S}} - \nu_{\text{LO}}$. Subsequent amplification and filtering of the intermediate frequency signal by a large factor then follows and the resultant signal is rectified by a diode and integrated. The key element in this entire process is the mixer. This device must be a nonlinear device that converts power from the original frequencies to the beat frequency, and is typically a diode because it

is a good approximation to a square-law device in terms of its current-voltage (I - V) behavior. If $I = V^2$, then the output current is proportional to the input power P , because V^2 is proportional to the electric field E^2 which is a measure of P . Within the mixer, the electric fields sum as vectors and the resultant power is the square of the amplitude of the electric fields. One of the best modern mixers is a junction made by separating two *superconductors* (not semiconductors) with a thin insulator; this is the SIS mixer.

The simplest radiometer (photometer) measures the average total power received over a well-defined radio frequency bandwidth $\Delta\nu$ and over a time interval τ . Just as for optical and infrared wavelengths, the weak astronomical source is measured against a background of many other radio signals such as the cosmic microwave background, the atmosphere, and the noise in the receiver itself. Power is usually expressed as a temperature, and the total system power is T_{sys} . The total noise in the measurement is given by the practical form of the radiometer equation:

$$\sigma_T = T_{\text{sys}} \left[(1/\Delta\nu_{\text{RF}}\tau) + (\Delta G/G)^2 \right]^{1/2} \quad (12.37)$$

where ΔG represents possible fluctuations in gain. If those fluctuations are negligible, then the equation simplifies to its ideal form ($T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$). In a manner similar to chopping in the infrared, one way to minimize fluctuations in receiver gain and atmospheric emission is to perform differential measurements by switching rapidly between two adjacent feeds as first suggested by Robert Dicke in the 1940s. The main drawback of Dicke switching is that the measured noise is doubled to $2 T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$ as a result of the difference measurement.

No quantum detector that uses the liberation of a (photo)electron by an incident photon can work at wavelengths longer than about 0.2 mm. Across the entire radio spectrum, the electromagnetic field, or the current which it induces in an antenna, is applied to a nonlinear element (diode) or mixer. The mixer either measures the total power or changes the signal frequency to one which is more easily measured. Frequently used devices are the Schottky diode and the superconducting junction (Zmuidzinas and Richards 2004).

When a metal and a semiconductor are brought into contact, the majority charge carriers of the semiconductor leave the contact zone until the Fermi levels of the metal and semiconductor are equalized. As a result, a “barrier” or “depletion region” empty of majority carriers appears in the semiconductor; typical barrier widths are $\sim 0.3 \mu\text{m}$. Even without any voltage across the junction, a current can flow through it. This is the Schottky diode. In practice, the semiconductor used is heavily doped gallium arsenide (GaAs) with a very thin lightly doped layer on the surface to reduce quantum mechanical tunneling and a contact layer of metal on top. Cooling the detector to a temperature of about 20 K yields a low-noise detector. Above 300 GHz (wavelengths shorter than 1 mm), the capacitance of the diode creates an RC filter which reduces its response.

When a thin insulating barrier is created between a normal metal and a superconducting metal (*not* a semiconductor) or between two superconductors, the structures are called SIN and SIS junctions. Nonlinear current can flow by quantum mechanical tunneling, and hence the device can be used as a mixer. When the voltage V is large enough that occupied states on one side are opposite vacant states on the other side, then a tunneling current can flow. Conversely, no current flows if $eV < 2\Delta$. Absorption of a photon can excite a charge carrier to the energy where the tunnel effect occurs. The high-frequency limit is about the same as the Schottky diode.

The output of the mixer is a signal with a frequency of $\nu_{\text{IF}} = \pm(\nu - \nu_{\text{LO}})$, where ν_{LO} is the frequency of the local oscillator, which is passed to the intermediate frequency (IF) amplifier and

then to a rectifying and smoothing circuit called a “detector.” This terminology seems strange compared to our previous usage where the detector is the device that receives the photons and converts them to an electrical signal. The detector part of the radio receiver is usually called the “backend” and can be quite complicated with a number of options, even for the same “front end” system. Rather than a single detector, a “backend” spectrometer often receives the output from the IF amplifier. The “spectrometer” usually consists of a number of electrical filters tuned to different frequencies with detectors on their outputs, a digital correlation computer.

4 Cryogenics and Vacuum Systems

Many detectors require cooling for optimum performance. There are several categories of cooling systems that might be required in astronomical instruments:

1. *Thermo-electric coolers and liquid circulation coolers.* These systems normally operate over the range -20°C to -50°C and are suitable for photomultiplier tubes, certain CCDs which have low dark currents, and high-speed applications such as telescope-guiding cameras.
2. *Liquid and solid cryogenics.* Dry ice (solid CO_2) is cheap and readily available. Coming in the form of a “snow,” it is most often used as the coolant for GaAs PMTs. Temperatures around -76°C (or 197 K) can be achieved with dry ice. Liquid nitrogen (LN_2) is also relatively cheap and can cool detectors (and other components) to -196°C (77 K), which is the normal boiling point of liquid nitrogen. Liquid helium (LHe) is more expensive, but it is needed to cool the low band gap semiconductor detectors and bolometers used in infrared instruments to -269°C (4 K). For liquid cryogenics, the cooling ability is expressed in terms of the product of the density (ρ) and the latent heat of vaporization (L_V):

$$\begin{aligned}(\rho L_V)_{\text{LHe}} &= 0.74 \text{ W hr l}^{-1} \\ (\rho L_V)_{\text{LN}_2} &= 44.7 \text{ W hr l}^{-1}\end{aligned}$$

For example, a 10 W heat load boils away 1 liter (l) of LHe in 0.07 h, whereas 1 l of LN_2 lasts 4.5 h. By attaching a vacuum pump to the LN_2 vent and reducing the pressure above the liquid, it is possible to solidify the nitrogen and achieve ~ 65 K.

3. *Electrical heat engines or closed-cycle refrigerators.* Because LHe is fairly expensive compared to liquid nitrogen and requires special care in handling, many infrared instruments and submillimeter/radio receivers employ multistage closed-cycle refrigerators (CCRs). Typical two-stage Gifford-McMahon systems using 99.999% pure helium gas at about 300 psi as the working fluid, such as the 350 model from CTI Inc. (now part of Brooks Automation Inc., USA), can provide two cold levels, usually 65 and 10 K, and extract heat at a rate of about 20 and 7 W from each stage. Both single-stage and triple-stage versions are available, and the larger units provide about 100 W of cooling power. If the mass to be cooled is very small, for example, a single CCD, then much simpler systems such as small Stirling cycle coolers can be used. Vibration damping is needed, especially for CCR-cooled instruments in sensitive AO systems, and counterbalance weights may also be required.
4. *^3He systems.* Even lower temperatures are obtained using ^3He systems in submillimeter and far-infrared bolometers. The basic principle of a ^3He cryostat is to condense helium-3 gas by bringing it in contact with a pumped helium-4 reservoir (yielding ~ 1.2 K). Low temperatures below 300 mK are then obtained (for small samples) by reducing the vapor pressure on top of the liquid helium-3 using an internal sorption pump (charcoal).

The heat H (J) removed from a mass m (kg) which is cooled from a temperature T_h to T_c is given by

$$H = mC(T_h - T_c) \quad (12.38)$$

where C is the specific heat of the material in joule/kg per K ($\text{J kg}^{-1}\text{K}^{-1}$). The specific heat of a substance usually changes with temperature, and it depends on the conditions under which the heat is applied (i.e., constant volume C_v or constant pressure C_p), but for solids the difference is generally small. For aluminum, C is about 900 J/kg·K, copper is 385 J/kg·K, steel is about 450 J/kg·K, and water is 4,190 J/kg·K. Specific heat is also tabulated in cal/g·K; using the conversion 4.19 joules per calorie gives 1 cal/g·K for water. As an example, to cool a mass of 1 kg (2.2 lb) of copper from 290 K to 80 K requires the removal of $1 \times 385 \times 210 = 80,850$ J and 2.3 times that amount for aluminum. If heat removal is to be accomplished in $t = 1$ h (3,600 s), then the average power is $H/t = 22.5$ W.

The rate of transfer of heat Q_H (in watts) by conduction along a rod of uniform cross-sectional area (A) and temperature gradient dT/dx is given by

$$Q_H = -kA \frac{dT}{dx} \quad (12.39)$$

where k is called the thermal conductivity ($\text{W m}^{-1}\text{K}^{-1}$) and is about 240 for Al, 400 for Cu, and only 0.9 for glass. In steady state conditions, we can write dT/dx as $\Delta T/L$, where L is the length of the conductor. If we think of heat flow as “current” and temperature difference as “voltage,” then by analogy with Ohm’s law ($V = IR$) for electrical circuits, we can define a “thermal resistance” such that

$$\Delta T = Q_H R, R = \frac{1}{k} \frac{L}{A} \quad (12.40)$$

Since k is a function of temperature, it is often more convenient to integrate over the required temperature range and give Q_H in the form

$$Q_H = \frac{A}{L} [I_{T_h} - I_{T_c}] \quad (12.41)$$

where L is the total length of the conductor and I is a tabulated property for many materials called the thermal conductivity integral which accounts for the variation of thermal conductivity (k) with temperature. In this expression, T_h and T_c represent the hot and cold temperatures between which the heat is flowing. If A and L are measured in cm^2 and cm, respectively, then I is in W/cm. Typical values are shown in [Table 12-1](#).

In large cryogenic infrared instruments, heat transfer by radiation is the dominant mechanism. The power radiated from a body of area A at an absolute temperature T is given by

$$Q_H = \varepsilon \sigma A T^4 \quad (12.42)$$

where $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ is the Stefan-Boltzmann constant and ε is the emissivity of the surface; $\varepsilon = 1$ for a perfectly black surface and is less than 0.1 for a shiny metallic surface. Polished aluminum foil yields an emissivity of about 0.05 and goldized Kapton gives ~ 0.04 ; anodizing increases the emissivity by a factor of 10 or more. The net rate of heat transfer by radiation from a body at temperature T_h onto a body at temperature T_c is given by

$$Q_H = \sigma A_c F_{hc} [T_h^4 - T_c^4] \quad (12.43)$$

Here, F_{hc} is an “effective” emissivity which also depends on the geometry of the cryostat (or dewar) configuration, such as concentric cylinders or plane-parallel plates, both of which are

■ Table 12-1

Values of the thermal conductivity integrals in W/cm for four materials

Temperature (K)	OFHC copper (W/cm)	6061 aluminum (W/cm)	Stainless steel (W/cm)	G-10 fiberglass (W/cm)
300	1,520	613	30.6	1.00
250	1,320	513	23.4	0.78
200	1,120	413	16.6	0.55
150	915	313	10.6	0.37
100	700	211	6.3	0.19
77	586	158	3.2	0.11
50	426	89.5	1.4	0.07
10	25	3.6	0.03	0.005

quite realistic. In both cases, when the emissivities of the two surfaces are small (<5%) and equal, then $F_{hc} \sim \varepsilon/2$.

For a given temperature differential, radiation load is minimized by reducing the surface area and achieving the lowest emissivity (shiniest) surfaces. It is also possible to add n “floating” shields which reduce the radiated heat load on the innermost body by a factor of $(n + 1)$, but careful application is critical. Various forms of floating shields are available. One form is called Multiple Layer Insulation (MLI). Typical emissivities range from 0.03 for polished aluminum and gold foil to 0.32 for polished anodized aluminum to 0.95 for a matt black surface. A useful rule of thumb is that if $\varepsilon = 0.05$ and the hot and cold temperatures are 290 and 80 K, respectively, then the radiation load is ~ 10 – 11 W/m². Note that radiation load is very sensitive to T_h .

Air at 20°C and 1 atmosphere of pressure (1 atmosphere = 760 torr and 1 torr = 132 Pa) contains about 2.7×10^{19} molecules per cubic centimeter, and the average distance traveled between collisions, called the mean free path, is about 7×10^{-6} cm. In general,

$$\lambda_{mfp} = \frac{1}{\sqrt{2}n\pi d^2} \quad (12.44)$$

where n is the number density of molecules and d is the diameter of the molecule. A “rough” vacuum is about 10^{-3} torr (mean free path = 5 cm) and a “high” vacuum would be 10^{-6} torr (mean free path 5×10^3 cm). The capacity of a vacuum pump is specified in terms of the pumping speed at the inlet, which is just the volume rate of flow $S = dV/dt$ l/s. The throughput of the flow is $Q_p = PS$ (torr-l/s) and the throughput of the pumping line is given by $Q_p = C\Delta P$, where ΔP is the pressure gradient and C is called the conductance which depends on gas pressure and viscosity. Since this equation is also analogous to Ohm’s law ($V = IR$) for electrical circuits, the net pumping speed of a pump and a system of pumping lines is given by

$$\frac{1}{S} = \frac{1}{S_{\text{pump}}} + \frac{1}{C_{\text{lines}}} \quad (12.45)$$

where the net conductance is found by adding the individual conductances like their electrical counterparts. Two equations are given for C (in l/s). The first corresponds to viscous flow when the mean free path is small and the other to molecular flow when the mean free path is large compared to tube dimensions and C is independent of pressure. Both apply to air at 20°C:

$$C = 180 \frac{D^4}{L} P_{av} \text{ or } C = 12 \frac{D^3}{L} \quad (12.46)$$

It is assumed that the tube is circular with diameter D and length L in cm and the pressure is in torr. Finally, the pump down time (in seconds) of a system with volume V from pressure P_0 to P assuming a constant net pumping speed S and no outgassing is

$$t = 2.3 \frac{V}{S} \ln \left(\frac{P_0}{P} \right) \quad (12.47)$$

Typically, the chamber is rough pumped to about 5×10^{-2} torr with a mechanical pump and then pumped to a lower pressure with a diffusion pump or turbomolecular pump. Typical pump speeds are 100 l/s at the inlet.

Developing modern astronomical instrumentation requires a broad knowledge of basic physics, engineering practice, and software. This chapter has touched on some of the background knowledge needed to appreciate what goes into the designing and building of such instruments. The design specification flows down from the scientific requirements. Initial optical, mechanical, and thermal calculations need to be followed by detailed ray tracing and finite element analysis. Many instruments require vacuum cryogenic systems. Analog and digital electronics are needed to operate the instrument and detector, and software to control the entire system must be robust and easy to use. Many other factors come into play when building astronomical instruments. See McLean (2008) for a more complete review of astronomical instruments and detectors.

References

- Allington-Smith, J., & Content, R. 1998, *PASP*, 110, 1216
- Arribas, S., et al. 1998, *Proc. SPIE*, 3355, 821
- Bacon, R., et al. 1995, *A&AS*, 113, 347
- Baldry, I. K., Bland-Hawthorn, J., & Robertson, J. G. 2004, Volume phase holographic gratings: polarization properties and diffraction efficiency. *Pub. Astron. Soc. Pac.*, 116, 403–414
- Barden, S. C., Arns, J. A., Colburn, W. S., & Williams, J. B. 2000, Volume-phase holographic gratings and the efficiency of three simple volume-phase holographic gratings. *Pub. Astron. Soc. Pac.*, 112, 809–820
- Bland-Hawthorn, J., & Kedziora-Chudczer, L. 2003, Taurus tunable filter: seven years of observing. *Pub. Astron. Soc. Australia*, 20, 242–251
- Clarke, D., & Grainger, J. 1971, *Polarized Light and Optical Measurement* (Oxford/New York: Pergamon Press)
- Courtes, G. 1982, *ASSL*, 92, 123
- Garcia, A. A., Rasilla, J. L., Arribas, S., & Mediavilla, E. 1994, *Proc. SPIE*, 2198, 75
- Garmire, G. P., Bautz, M. W., Ford, P. G., Nousek, J. A., & Ricker, G. R., Jr. 2003, *Proc. SPIE*, 4851, 28. X-ray and gamma-ray telescopes and instruments for astronomy, ed. J. E. Trümper, & H. D. Tananbaum
- Gehrels, N., Chincarini, G., Giommi, P., Mason, K. O., Nousek, J. A., Wells, A. A., White, N. E., Barthelmy, S. D., Burrows, D. N., Cominsky, L. R., et al. 2004, The *Swift* gamma-ray burst mission. *ApJ*, 611, 1005–1020
- Guyon, O. 2006, Theoretical limits on extrasolar planet detection with coronagraphs. *Astrophys. J. Suppl. Ser.*, 167, 81–99
- Hinkley, S., Oppenheimer, B. R., Brenner, D., Parry, I. R., Sivaramakrishnan, A., Soummer, R., & King, D. 2008, *Proc. SPIE*, 7015, 32
- Janesick, J. R. 2001, *Scientific Charge-Coupled Devices* (Bellingham: SPIE)
- Kasdin, N. J., et al. 2003, Extrasolar planet finding via optimal apodized-pupil and shaped-pupil coronagraphs. *Astrophys. J.*, 582, 1147–1161
- Krabbe, A., Thatte, N. A., Kroker, H., Tacconi-Garman, L. E., & Tecza, M. 1997, *Proc. SPIE*, 2871, 1179
- Larkin, J., et al. 2006, *Proc. SPIE*, 6269, 42
- Lyot, B. 1939, A study of the solar corona and prominences without eclipses. *Mon. Not. R. Astron. Soc.*, 99, 580–594
- McLean, I. S. 2008, *Electronic Imaging in Astronomy: Detectors and Instrumentation* (2nd ed.; Heidelberg: Springer)
- Morrissey, P., Schiminovich, D., Barlow, T. A., Martin, D. C., Blakkolb, B., Conrow, T.,

- Cooke, B., Erickson, K., Fanson, J., Friedman, P. G., et al. 2005, The on-orbit performance of the *Galaxy Evolution Explorer*. *ApJ*, 619, L7–L10
- Murphy, T. W., Matthews, K., & Soifer, B. T. 1999, *PASP*, 111, 1176
- Palacios, D. M. 2005, An optical vortex coronagraph. *Proc. SPIE*, 5905, 196
- Rieke, G.H. 2003, *The Measurement of Light from the UV to the Sub-millimeter* (Cambridge: Cambridge University Press)
- Rouan, D., et al. 2000, The four quadrant phase mask coronagraph. I. Principle. *PASP*, 112, 1479–1486
- Soummer, R., Aime, C., & Falloon, P. E. 2003, Stellar coronagraphy with prolate apodized circular apertures. *A&A*, 397, 1161–1172
- Tecza, M., Thatte, N. A., Krabbe, A., & Tacconi-Garman, L. E. 1998, *Proc. SPIE*, 3354, 394
- Tinbergen, J. 1996, *Astronomical Polarimetry* (Cambridge, UK: Cambridge University Press)
- Siegmund, O., Vallerger, J., Tremsin, A., and McPhate, J. 2007, Microchannel plates: recent advances in performance, *Proc. SPIE*, 6686, 66860W
- Wizinowich, P., Acton, D. S., Shelton, C., Stomski, P., Gathright, J., Ho, K., Lupton, W., Tsubota, K., Lai, O., Max, C., Brase, J., An, J., Avicola, K., Olivier, S., Gavel, D., Macintosh, B., Ghez, A., & Larkin, J., 2000, First Light Adaptive Optics Images from the Keck II Telescope: A New Era of High Angular Resolution Imagery, *PASP*, 112, 315
- Zmuidzinas, J., & Richards, P. L. 2004, Superconducting detectors and mixers for millimeter and submillimeter astrophysics. *Proc. IEEE*, 92, 1597–1616

