Optical and infrared detectors: fundamentals
By B. Menard (supplemented and edited by N.Zakamska)

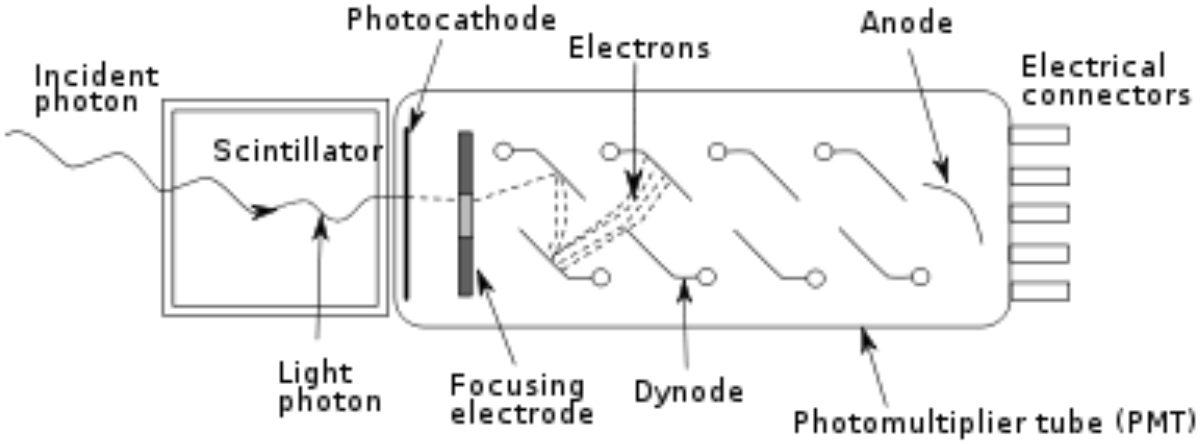
What would be the ideal detector?
- noiseless
- have 100 percent "quantum efficiency" (QE) (i.e., the probability that we detect a photon)
- be available with any desired pixel size
- measure x, y, lambda (or nu), t and polarization for each photon incident on it, through the visible and into the mid-IR
- linearity
The performance of the current detectors is definitely not there yet!

HISTORY

Through the late part of the 19th and first two-thirds of the 20th, the effective QE of photographic plates improved by a factor of order 100, culminating in Kodak IIIa-J, at ~0.5%. Plates were never satisfactory for photometry except in the very most careful hands-- which existed, but were not very numerous. They are nonlinear, nonuniform, and cannot be calibrated after the fact.

Nobel Prize in Physics 1921 was award to Albert Einstein "for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect".
https://en.wikipedia.org/wiki/Photoelectric_effect

Photomultipliers were developed during the war. After the Second World War, the 931A, the 1P21, and later the red-sensitive RCA7102, came to be used for astronomy. They were single-channel devices, suitable for aperture photometry and, again in capable hands, spectrophotometry. They had quantum efficiencies approaching 20% in the blue and a few percent in the red.



In the 1960s, the first photoelectric image tubes became available, both as a result of military research and research sponsored by astronomy, particularly through the Carnegie institution of Washington. This led to magnetically-focussed devices with good photocathodes and ~20 micron resolution. They were light amplifiers, and the recording medium was again photographic plates.

Electronic detectors in the form of low-light-level vidicon camera tubes, all employing the charge multiplication of energetic (typically a few keV) electrons bombarding some kind of `sticky' target, were developed in the 1960s and early 1970s.

Analog-to-digital converters were used from the very beginning with these devices, and the resulting pixellated images stored on magnetic tape and/or disk. Various schemes were also implemented for 1 and 2-d photon counting using electron multipliers of very high gain; some of these are still in use today in the UV.

The charge-coupled device was invented in 1969 at AT&T Bell Labs by Willard Boyle and George E. Smith (2009 Nobel Prize in Physics). The first CCDs used in astronomy came in the middle 1970s; devices from RCA, Fairchild, and TI were used, but by far the best devices were the TI ones developed with NASA support for Voyager and later the mission which became Hubble. These CCDs were developed for low light-level scientific applications and were the first high-sensitivity, low-noise, very high performance CCD devices available.
https://en.wikipedia.org/wiki/Charge-coupled_device

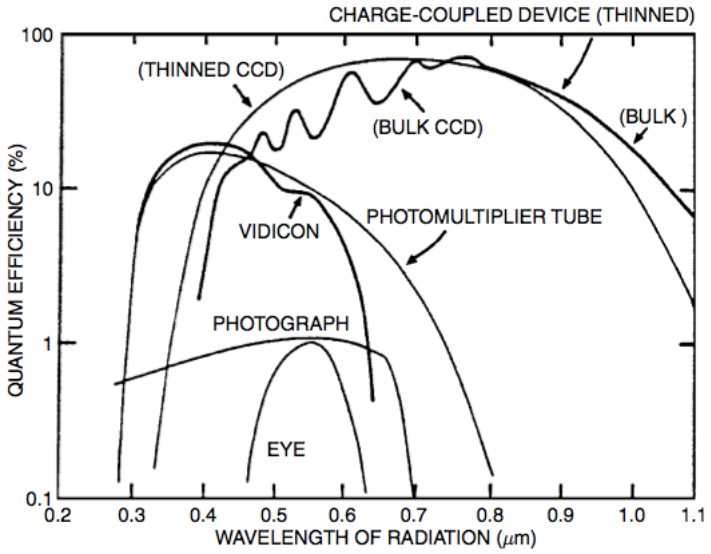


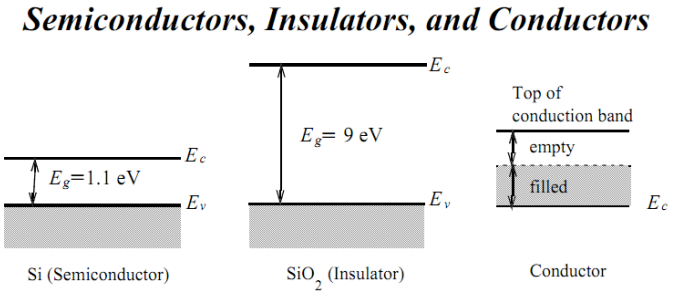
Fig. 3.2. QE curves for various devices, indicating why CCDs are a quantum leap above all previous imaging devices. The failure of CCDs at optical wavelengths shorter than about 3500 Å has been essentially eliminated via thinning or coating of the devices (see Figure 3.3).

PRESENT-DAY: solid-state photodetector imagers -- CCDs, CMOS imagers, IR imagers using InSb and HgCdTe photodiode arrays.

Essentially all of the current crop of detectors are MOS (metal-oxide-semiconductor) devices.

What is a semiconductor, and what semiconductor properties do imagers use? Semiconductors are loosely defined as solids
a) which have well-defined crystalline structures
b) in which the energy bands containing the valence electrons of the constituent atoms are filled. (Metals have partially filled bands, allowing electrons to carry momentum and current with very little energy input).
c) in which there exist continuum bands at most about three volts above the top of the filled valence band. This energy difference is called the BAND GAP. This is equivalent to the energy required to free an outer shell electron from its orbit about the nucleus to become a mobile charge carrier, able to move freely within the solid material.

The distinction between dielectrics (insulators) and semiconductors is fuzzy, but generally insulators have larger band gaps.
- Insulator: a material with a large band gap is an insulator.
- A semiconductor is a material with a small but nonzero band gap
- In conductors, the valence and conduction bands may overlap, so there may not have a band gap.



Many detectors are made of silicon. The property of silicon which allows the devices in principle to be made is simply that it is a semiconductor; in practice the availability of these devices is a direct consequence of the enormous commercial development of silicon semiconductor technology for innumerable other uses. The choice of silicon over other semiconductors is motivated by several factors, not the least of which is that silicon can be oxidized to form a robust dielectric layer, allowing devices to be built of several vertically differentiated and electrically isolated layers. MOS devices make use of M(etal) gates isolated by O(xide) from the S(emiconductor) substrate.

The band gap in silicon is about 1.12eV at room temperature. This is a fundamental limit! 1.12 eV corresponds to a photon at about 11000 A; this sets a firm upper limit on the wavelength of light which can be detected with silicon photodetectors.

The potential gradient perpendicular to the surface can be manipulated as well as the layer is grown, and can incorporate an electric field which attracts electrons toward the surface. Thus a photon above the band gap energy which interacts with the silicon can knock a valence electron into the conduction band, and physically the electron will be pulled toward the surface of the epitaxial layer, where it can be moved about by manipulating the voltages on gates built on the surface.

List of band gaps

Material	Symbol	Band gap (eV) @ 302K	Reference
Indium antimonide	InSb	0.17	[6]
Lead(II) selenide	PbSe	0.27	[6]
Lead(II) telluride	PbTe	0.29	[6]
Indium(III) arsenide	InAs	0.36	[6]
Lead(II) sulfide	PbS	0.37	[6]
Germanium	Ge	0.67	[6]
Gallium antimonide	GaSb	0.7	[6]
Indium(III) nitride	InN	0.7	[7]
Silicon	Si	1.11	[6]
Copper(II) oxide	CuO	1.2	[9]
Indium(III) phosphide	InP	1.35	[6]
Gallium(III) arsenide	GaAs	1.43	[6]
Cadmium telluride	CdTe	1.49	[8]
Aluminium antimonide	AlSb	1.6	[6]
Cadmium selenide	CdSe	1.73	[6]
Selenium	Se	1.74	
Copper(I) oxide	Cu_2O	2.1	[10]
Aluminium arsenide	AlAs	2.16	[6]
Zinc telluride	ZnTe	2.25	[6]
Gallium(III) phosphide	GaP	2.26	[6]
Cadmium sulfide	CdS	2.42	[6]

Left to themselves, these conduction band electrons would recombine back into the valence level within approximately 100 microseconds. Silicon has a useful photoelectric effect range of 1.1 to about 10 eV, which covers the near-IR to soft X-ray region (Rieke, 1994). Above and below these limits, the CCD material appears transparent to the incoming photons.

Gallium arsenide (GaAs) has six times higher electron mobility than silicon, which allows faster operation; wider band gap, which allows operation of power devices at higher temperatures and gives lower thermal noise to low power devices at room temperature; its direct band gap gives it more favorable optoelectronic properties than the indirect band gap of silicon; it can be alloyed to ternary and quaternary compositions, with adjustable band gap width, allowing light emission at chosen wavelengths, and allowing e.g. matching to wavelengths with lowest losses in optical fibers. GaAs can be also grown in a semi-insulating form, which is suitable as a lattice-matching insulating substrate for GaAs devices. Conversely, silicon is robust, cheap, and easy to process, whereas GaAs is brittle and expensive, and insulation layers can not be created by just growing an oxide layer; GaAs is therefore used only where silicon is not sufficient.

The summary of the differences between CMOS vs CCD detectors:
https://electronics.howstuffworks.com/cameras-photography/digital/question362.htm

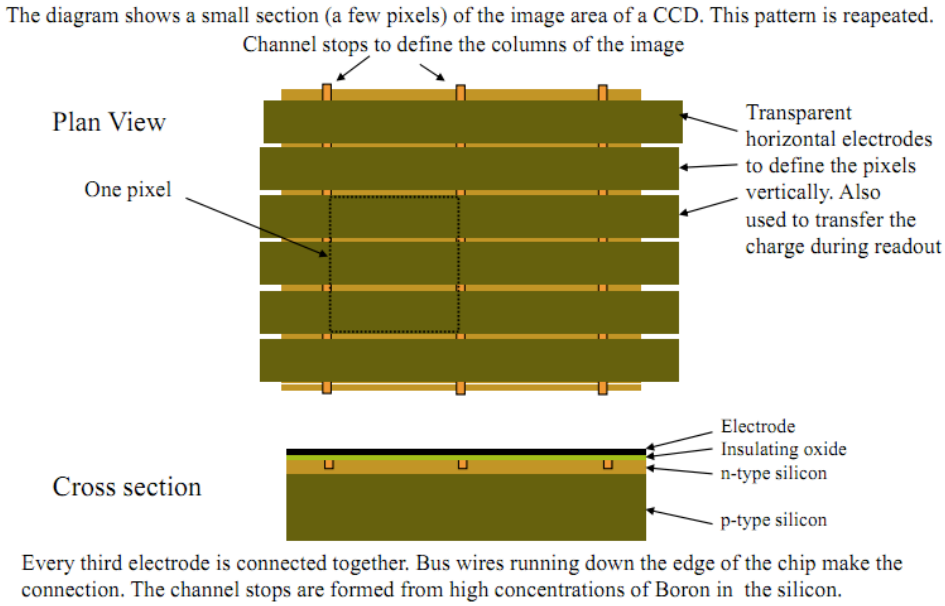
From here on, we will mostly discuss CCD detectors.

Advantages of CCD detectors for astronomy:

1. Very high quantum efficiency, often limited entirely by the reflection losses at their surfaces. CCDs in the visible and red can have QE as high as 90%.
2. Very low noise. CCDs have on-chip amplifiers with very low capacitance, achieving sensitivities of the order of one or even a few microvolts per electron, and the readout noise using proper circuitry is as small as a few electrons. The noise in a signal of N electrons is sqrt(N), so if the read noise is n electrons, detected photon statistics will dominate for all signals greater than n^2 electrons. If n is 3, which is achievable, then read noise is negligible for all observations for which the S/N > 3, N > 10 or so for measurements including the background!
3. Very large dynamic range. For a device with a read noise of 4 electrons and a capacity of 100,000 electrons, which is not unusual, the dynamic range is 25,000. The useful dynamic range of a photographic plate is less than 10, and for photoelectric imagers at most a few hundred.
4. Very good linearity, often better than a few percent over the full very large dynamic range of the device. The nonlinearity which exists is stable and can be calibrated relatively easily.
5. Good devices from the past two decades or so are quite uniform in response over their surfaces, and again the non-uniformities are stable. The big problem with calibrating the devices over their surface is not the device itself but scattered light.

6. Well-defined and stable intra-pixel response, facilitating PSF photometry with barely adequately sampled images and accurate astrometry.
7. Large sizes. LSST camera is 3.2 gigapixels with 10-micron pixels. At the relevant focal position can produce 0.2 arcsec sampling of the PSF, so that the full field of view is more than 3 degrees on the side!

Reading out the CCD



Reading out the CCD is the process of converting the charge in each pixel into a signal for conversion into a digital unit. One clock cycle moves each row of pixels up one column, with the top row being shifted off the array into what is called the output shift register or horizontal shift register. This register is another row of pixels hidden from view (i.e., not exposed to incident light) and serves as the transition between active rows on the array and the output of the device. Animation of the charge transfer process: https://upload.wikimedia.org/wikipedia/commons/thumb/6/66/CCD_charge_transfer_animation.gif/250px-CCD_charge_transfer_animation.gif

The transfer of the total charge from location to location within the array is not without losses. Each charge transfer (one of which occurs for each voltage change or clock cycle) has an associated efficiency. This efficiency value is the percent of charge transferred compared with that which was actually collected. Modern values for the charge transfer efficiency (CTE) are approaching 0.999 999 (i.e., 99.9999% efficient) for each transfer. Example: CTE has been declining on Hubble due to the damage from the space environment, http://www.stsci.edu/hst/instrumentation/acs/performance/cte-information.

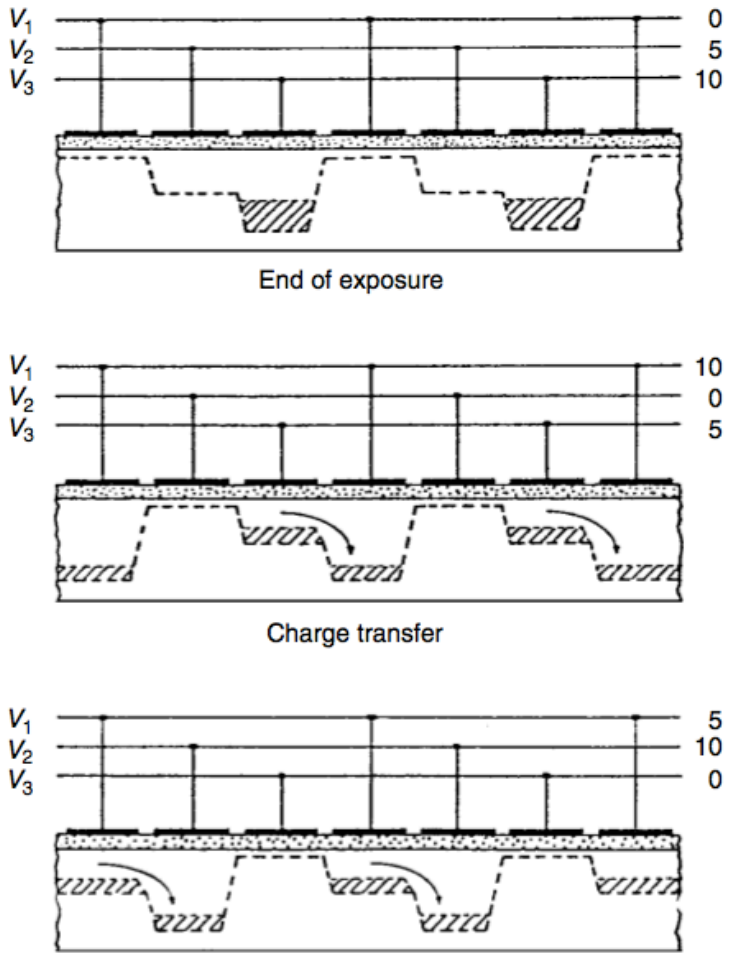


Fig. 2.2. Schematic voltage operation of a typical three-phase CCD. The clock voltages are shown at three times during the readout process, indicating their clock cycle of 0, 10, and 5 volts. One clock cycle causes the stored charge within a pixel to be transferred to its neighboring pixel. CCD readout continues until all the pixels have had their charge transferred completely out of the array and through the A/D converter. From Walker (1987).

Well-designed CCDs made of good material achieve levels of transfer of charge in a single pixel (three-phase) transfer of CTE = .99998 or better; at this level it is clearly better to use the charge transfer inefficiency, CTI = 1-CTE <= 2e-5. This number means that a fraction of the charge CTI is left behind in a transfer. After N transfers, a charge packet of size Q has a tail which is roughly Q*exp(-N*CTI*n), where n is the index of the pixel following the packet (n=0 is the packet).

For a 1024x1024 CCD, the charge collected in the last pixel to be read out has to be transferred over two thousand times! The loss in charge from a CCD pixel containing N electrons that is shifted 1024 times vertically and 1024 times horizontally is given by L(e) = 2048xNxCTI. CCDs with poor CTE generally show charge tails in the direction opposite readout for bright stars. These tails are the charge left behind as the image is shifted out. Take a look at the example ACS HST image: http://www.stsci.edu/itt/review/dhb_2014/ACS/acs_Ch47.html.

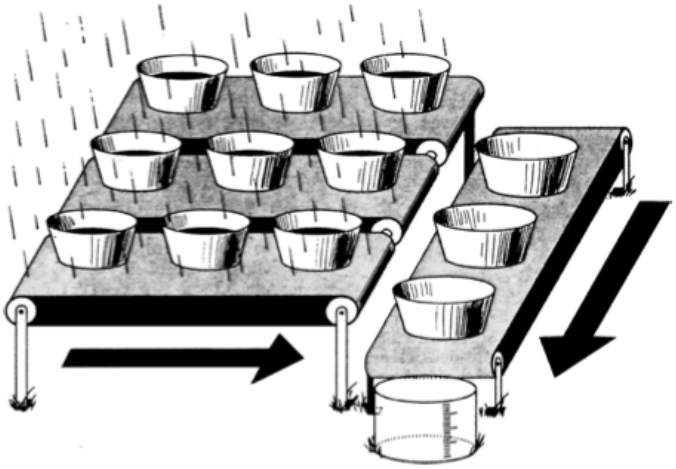


Fig. 2.1. CCDs can be likened to an array of buckets that are placed in a field and collect water during a rainstorm. After the storm, each bucket is moved along conveyor belts until it reaches a metering station. The water collected in each field bucket is then emptied into the metering bucket within which it can be measured. From Janesick & Blouke (1987).

Single monolithic CCDs usually have 2 or 4 output amplifiers available (one in each corner) and given the proper electronic setup, these large chips are often read out from 2 or 4 corners simultaneously, thus decreasing the total readout time by 2-4.

The charge collected within each pixel is measured as a voltage and converted into an output digital number. Each pixel's collected charge is sensed and amplified by an output amplifier. CCD output amplifiers are designed to have low noise and are built directly into the silicon circuitry; thus they are often referred to as on-chip amplifiers. These amplifiers must work with extremely small voltages and are rated, as to their sensitivity, in volts per electron. Typical values are in the range of 0.5 to 4 microvolts per electron

The output voltage from a given pixel is converted to a digital number (DN) and is typically discussed from then on as either counts or ADUs (analog-to-digital units). The amount of voltage needed (i.e., the number of collected electrons or received photons) to produce 1 ADU is termed the **gain** of the device.

Gain refers to the magnitude of amplification a given system will produce. Gain is reported in terms of electrons/ADU (analog-to-digital unit). A gain of 10 means that the camera digitizes the CCD signal so that each ADU corresponds to 10 photoelectrons.
https://www.photometrics.com/learn/imaging-topics/gain
For example, a photomultiplier by design has a very large gain, ~1e8.

With gain=10, if a pixel collects 1000 electrons (photons), the output pixel value stored in the computer would be 100 ADUs. For 17 234 electrons, the output pixel value would be 1723 ADUs -- not 1723.4. Digital output values can only be integer numbers and it is clear already that the discrimination between different pixel values can only be as good as the resolution of the gain and digital conversion of the device.

CCD readout time can reach ~ 1 min up to several minutes, so it is a major source of overhead in optical and especially infrared observations, where long exposure times are undesirable due to the strong thermal and airglow background. So deep observations in

the IR have to be split into multiple shallower observations, with the result that a lot of time is spent reading out the CCD. To reduce the read-out time, some instruments offer "partial" readout, where only part of the CCD is read out, or "fast" readout (at the expense of a higher read-out noise, but appropriate for bright objects), or detector-based binning of the data.

CCD data reduction and sources of noise

*** Dark currents

ref: http://inst.eecs.berkeley.edu/~ee130/fa07/lectures/Semiconductor_fundamentals_lec3.pdf

Every material at a temperature much above absolute zero will be subject to thermal noise within. For silicon in a CCD, this means that when the thermal agitation is high enough, electrons will be freed from the valence band and become collected within the potential well of a pixel. When the device is readout, these dark current electrons become part of the signal, indistinguishable from astronomical photons.

At room temperature, the dark current of a typical CCD is near 2.5 x 10^4 electrons/pixel/second. Typical values for properly cooled devices range from 2 electrons per second per pixel down to very low levels of approximately 0.04 electrons per second for each pixel. Although 2 electrons of thermal noise generated within a pixel every second sounds very low, a typical 15 minute exposure of a faint astronomical source would include 1800 additional (thermal) electrons within each CCD pixel upon readout. Bottom line: The CCD needs to be cooled!

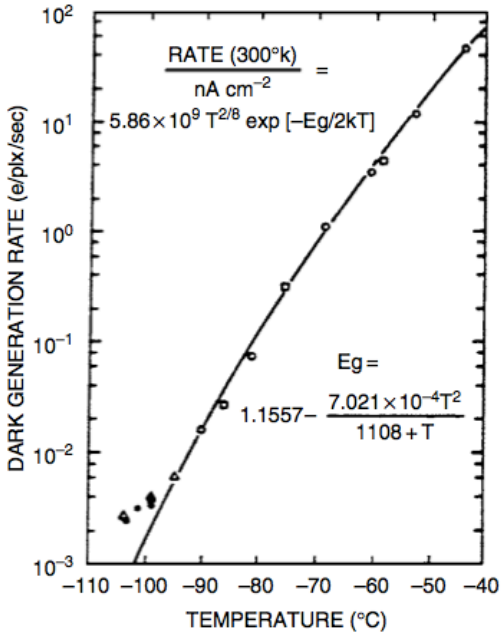


Fig. 3.6. Experimental (symbols) and theoretical (line) results for the dark current generated in a typical three-phase CCD. The rate of dark current, in electrons generated within each pixel every second, is shown as a function of the CCD operating temperature. E_g is the band gap energy for silicon. From Robinson (1988a).

*** Read noise

Read noise consists of two inseparable components. First is the conversion from an analog signal to a digital number, which is not perfectly repeatable.
Second, the electronics themselves will introduce spurious electrons into the entire process, yielding unwanted random fluctuations in the output. Good read noise values in today's CCDs are in the range of 10 electrons per pixel per read or less.

Read noise can be isolated via the subtraction of a bias frame. A bias frame is an image of zero exposure time where the CCD is read out without having been exposed to light. In this manner, thermal noise produced by the heat generated by the device's electronics and contributions of light to the exposure are at a minimum, thereby isolating the effect of read noise.

*** Bias

Bias or zero images allow one to measure the zero noise level of a CCD. For an unexposed pixel, the value for zero collected photoelectrons will translate, upon readout and A/D conversion, into a mean value with a small distribution about zero. To avoid negative numbers in the output image, CCD electronics are set up to provide a positive offset value for each accumulated image. This offset value, the mean `zero' level, is called the bias level. A typical bias level might be a value of 400 ADU (per pixel), which, for a gain of 10 e-/ADU = 4,000 e-. Bias frames amount to taking observations without exposure to light (shutter closed), for a total integration time of 0.000 seconds.

Variations in the mean zero level of a CCD are known to occur over time and are usually slow drifts over many months or longer, not noticeable changes from night to night or image to image.

*** Processing

reduced image = (raw image - dark) / (flat field - bias)

Here the bias is included in the dark frame.

There are a couple of simple techniques used to compensate for the effects outlined above.

-- A bias frame is an image exposed for zero seconds (or as short as the camera allows) with a closed shutter. It contains information about pixel-to-pixel variations in the read noise, often visible as a gradient in the bias frame, as well as any defects in the chip. Subtracting a bias frame from a light frame (i.e., a normal image) removes these effects.

-- A dark frame is an image exposed for a finite amount of time with a closed shutter. It contains the level of the dark current and any pixel-to-pixel variations in it. Subtracting a dark frame from a light frame corrects for the dark current. Dark frames are often taken with the same exposure times as the science images, but during the telescope "down time", for example, during the "day-time calibrations", so as not to waste valuable night-time hours. The bias information is also in the dark frame, so bias-subtraction is subsumed by dark subtraction.

-- A flat frame is an image exposed to a uniform light source. It contains information about pixel-to-pixel variations in quantum efficiency (i.e., response to light) and light variations due to the instrument configuration. Dividing a dark-subtracted light frame by a dark-subtracted flat frame corrects for this effect. Flats can be obtained either using special lamps in the dome (don't forget to turn it off after the exposure!) or by taking exposures of the twilight sky.

*** The CCD equation

$$\frac{\mathbf{S}}{\mathbf{N}} = \frac{N_*}{\sqrt{N_* + n_{\rm pix}(N_S + N_D + N_R^2)}}$$

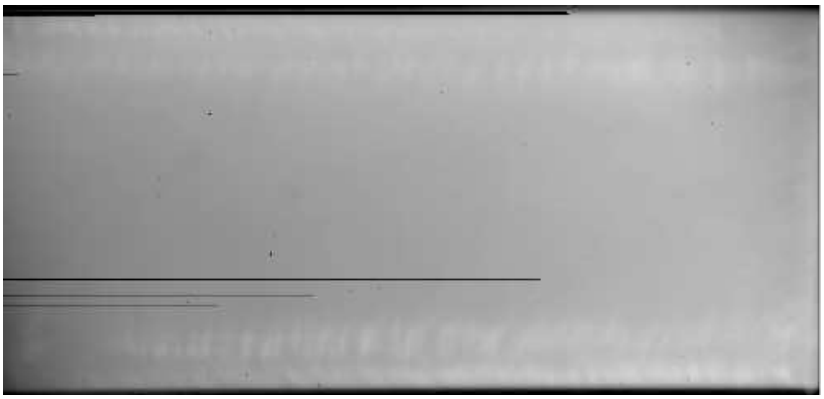
N_*: number of photons collected form the object of interest
The different noise contributions can first be introduced individually with a \sigma for each. Each sigma is due to Poisson statistics or simple noise term.
N_S: sky background per pixel
N_D: dark current per pixel
N_R: number of electron per pixel due to the read noise. This is not a Poisson statistic.

Alternative view of the CCD equation is as a function of exposure time, should be self-explanatory:

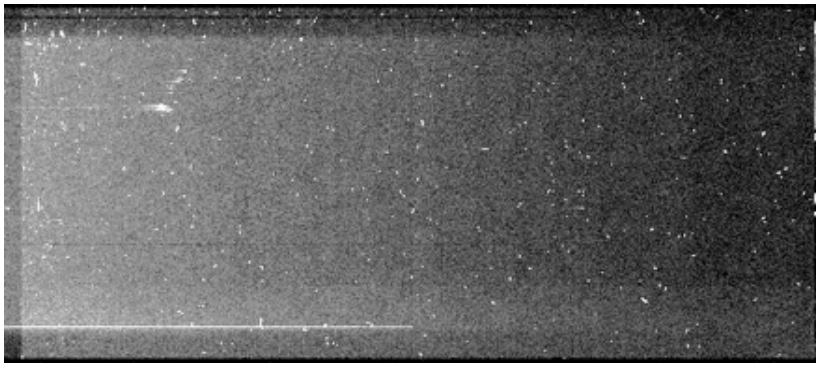
$$\frac{S}{N} = \frac{\dot{N}_* t}{\sqrt{\dot{N}_* t + n_{\rm pix}(\dot{N}_S t + \dot{N}_D t + N_R^2)}}$$

Image defects

Unless one pays a huge amount it is generally difficult to obtain a CCD free of image defects. The first kind of defect is a ‘dark column’. Their locations are identified from flat field exposures. Dark columns are caused by ‘traps’ that block the vertical transfer of charge during image readout. The CCD shown at left has at least 7 dark columns, some grouped together in adjacent clusters. Traps can be caused by crystal boundaries in the silicon of the CCD or by manufacturing defects. Although they spoil the chip cosmetically, dark columns are not a big problem for astronomers. This chip has 2048 image columns so 7 bad columns represents a tiny loss of data. Image shows Flat field exposure of an EEV42-80 CCD.



There are three other common image defect types: Cosmic rays, Bright columns and Hot Spots. Bright columns are also caused by traps. Electrons contained in such traps can leak out during readout causing a vertical streak. Hot Spots are pixels with higher than normal dark current. Their brightness increases linearly with exposure times. Cosmic rays are unavoidable. Charged particles from space or from radioactive traces in the material of the camera can cause ionization in the silicon. The electrons produced are indistinguishable from photo-generated electrons. Approximately 2 cosmic rays per cm^2 per minute will be seen. A typical event will be spread over a few adjacent pixels and contain several thousand electrons. Somewhat rarer are light-emitting defects which are hot spots that act as tiny LEDS and cause a halo of light on the chip. Their locations are shown in the image below which is a lengthy exposure taken in the dark (a ‘Dark Frame’).



CCDs in blue and UV observations

Blue cutoff of CCDs: The composition of a CCD is essentially pure silicon (by volume). The quantum mechanics of this element is thus ultimately responsible for the response of the detector to various wavelengths of light. Absorption properties of silicon are such that for light outside the range of about 3500 to over 8000A, the photons (1) pass right

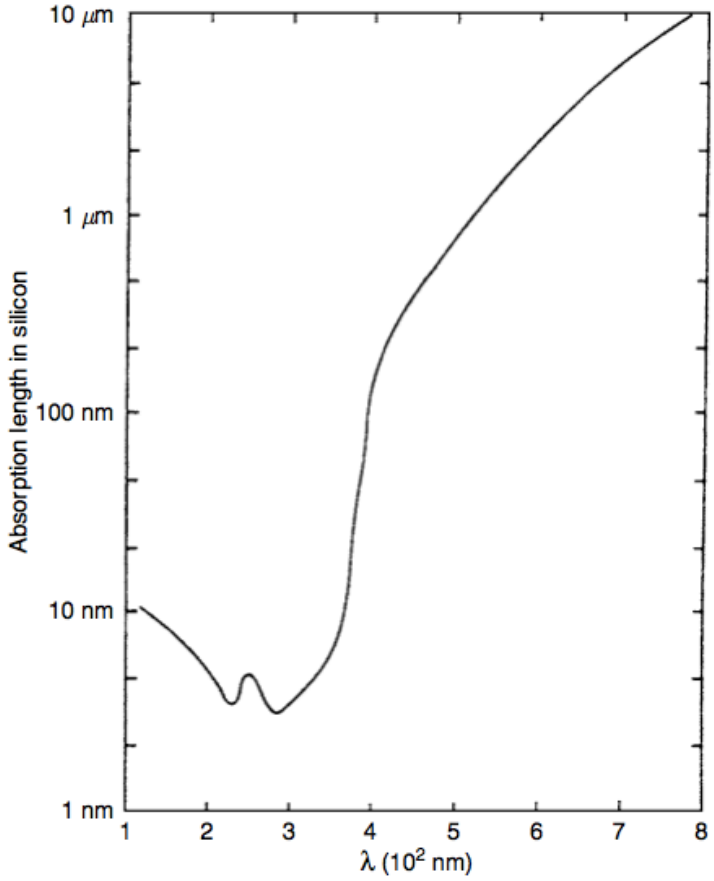


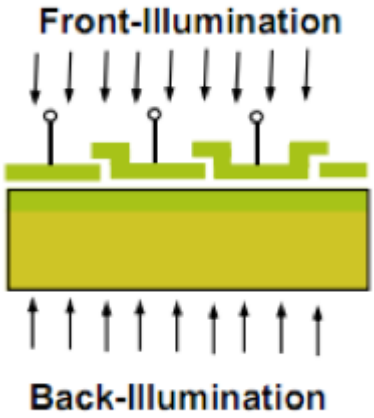
Fig. 3.1. The photon absorption length in silicon is shown as a function of wavelength in nanometers. From Reicke (1994).

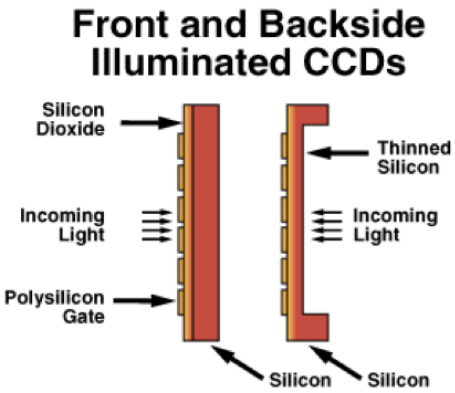
through the silicon, (2) get absorbed within the thin surface layers or gate structures, or (3) simply reflect off the CCD surface. Thus the quantum efficiency curve (as a function of lambda) of a typical CCD device will approximately mirror the photon absorption curve for silicon.

(Photons of energy 1.1 eV to near 4 or so eV generate single electrons. Above 4 eV, a photon can generate multiple electrons -- this is why there is a feature at 2000A.) From the absorption curve, it's clear that blue photons are absorbed while crossing a few

atomic layers of silicon. Then the circuitry (made out of silicon) blocks the blue light which is then not recorded by the detector.
The circuitry (the gate structures) covers the CCD.

For **front-illuminated** CCDs, illumination occurs on the front of the CCD with the photons being absorbed by the silicon after passing directly through the surface gate structures. These are thick, inexpensive CCDs used in consumer applications. Because of their thickness (300 microns) they are susceptible to cosmic rays. But for astronomical CCDs the circuitry blocks too much blue light, so they must be **back-illuminated**. Nearly the entire substrate must be mechanically and chemically removed so that photon-generated charges can reach the potential wells, leaving only a 20-micron thin layer beneath the circuitry. So many of these gossamer chips are damaged during fabrication that the survivors are worth tens of thousands of dollars apiece. Production failure rate: 99% (for the SDSS camera blue chip). Back-side illuminated devices, also known as thinned devices, are physically thinned to >=15 microns.





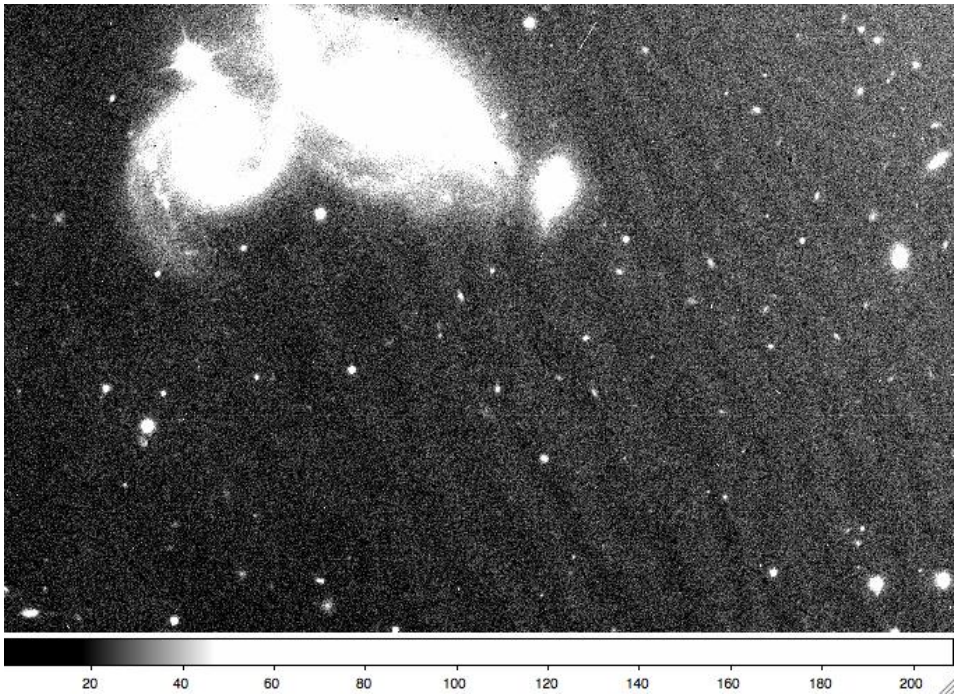
The incoming photons are now able to be absorbed directly into the bulk silicon pixels without the interference of the gate structures. The advantages in this type of CCD are that the relative quantum efficiency greatly exceeds that of a front-side device and the response of the detector to shorter wavelength light is improved since the photons no longer need to pass through the pixel gates.

Disadvantages of back-illuminated CCDs:
- shallower pixel well depths (due to the much smaller amount of material present per pixel);
- possible nonuniform thinning leading to surface and flat-field nonuniformities;
- increased expense incurred by the thinning and mounting process.

The z-band of the SDSS is front-illuminated. The red sensitivity of a thick frontside CCD on the other hand is satisfactory. It allows one to reach the theoretical limit in the

infrared. The thicker the substrate the better the sensitivity in the red region of the spectrum.

Another consequence of thinning a CCD is that it will be less sensitive to the red region of the spectrum but they are perfect in the blue region of the spectrum. The thinned layer is the cause of **interference fringes** of equal thickness which strongly modulate the sensitivity across the detector's surface. (from http://pages.astronomy.ua.edu/keel/techniques/).



**** Coating

Some coating materials allow CCDs to become sensitive to photons normally too blue to allow absorption by the silicon. They generally consist of organic phosphors that down-convert incident UV light into longer wavelength photons, easily detected by the CCD. One common phosphor coating, lumogen, eliminates the low-QE notch in the opacity curve, as it is responsive to wavelengths between 500 and 4200A. An interesting side note is that lumogen is the commercial phosphorescent dye used in yellow highlighting pens

Silicon has a very high refractive index. The fraction of photons reflected at the interface between two media n_i (from where the light is incident) and n_t (into which the light is transmitted) is [(nt-ni)/(nt_ni)]^2. For example for the air-water boundary, ni=1, nt=1.33, so the reflectivity of the water in air is 4%. In contrast, for the vacuum-silicon boundary, ni=1, nt=3.6, so R is 32%. Unless we take steps to eliminate this reflected portion, then

a silicon CCD will at best only detect 2 out of every 3 photons. The solution is to deposit a thin layer of a transparent dielectric material on the surface of the CCD (similar to impedance matching in electronics).