12 Optical Materials and Devices

In this chapter we will see the light associated with the intersection between the electronic and optical properties of materials, called *optoelectronics*, or, with a bit less electronics and a bit more religion, *photonics*.

We'll start by looking at how electronic energy can be converted into illumination (both visual and intellectual), then how light can be converted back to electronic energy and information, and close with methods for modifying it.

12.1 GENERATION

12.1.1 Incandescence

The simplest way to generate light is by heating something to produce *incandescence*. To find the states available to a thermal photon in a box of side length L, following the derivation of the electron density of states (from equation 11.36) we'll impose periodic boundary conditions on the radiation field so that the wave vector $\vec{k}$ is indexed by integers n_x, n_y, n_z

$$\begin{aligned} \vec{k} &= \frac{2\pi}{L}\left(n_x\hat{x} + n_y\hat{y} + n_z\hat{z}\right) \\ k^2 &= \left(\frac{2\pi}{L}\right)^2 r^2 \quad , \end{aligned} \tag{12.1}$$

which in the limit of large box can be taken to define a continuous variable r. In terms of the frequency,

$$\frac{2\pi}{c}\nu = k = \frac{2\pi}{L}r \tag{12.2}$$

or

$$r = \frac{L}{c}\nu \quad . \tag{12.3}$$

The total number of states in a volume $V = L^3$ that have indices between r and $r + dr$ is then given in terms of the density of states per volume N by a spherical shell

$$\begin{aligned} VN(r)\,dr &= 2 \cdot 4\pi r^2\,dr \\ &= 8\pi r^2\,dr \end{aligned}$$

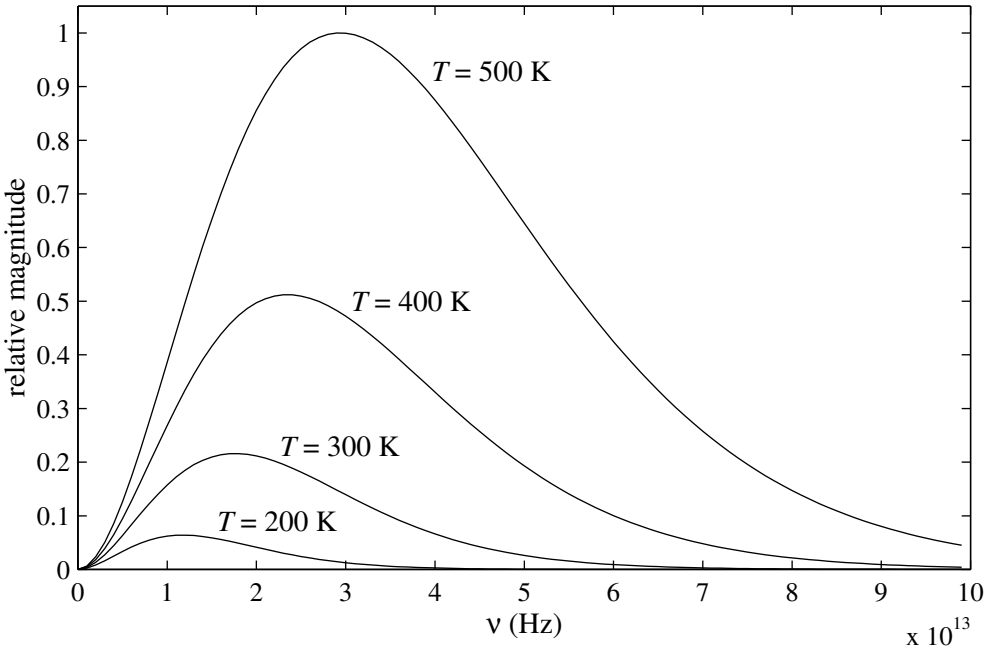

Figure 12.1. Planck's Law for thermal radiation.

$$= 8\pi \left(\frac{L}{c}\right)^2 \nu^2 \frac{L}{c}\, d\nu$$

$$= 8\pi \frac{V}{c^3} \nu^2\, d\nu$$

$$N(\nu)\, d\nu = \frac{8\pi}{c^3} \nu^2\, d\nu \quad . \tag{12.4}$$

The first factor of 2 comes from the two possible transverse photon polarizations (which will be discussed further in Section 12.3). The thermal photon energy per volume U in the box is then found by multiplying this density of states by the photon energy $h\nu$ and the Bose–Einstein distribution for the photon occupancy of the available sates,

$$U = h\nu \frac{8\pi}{c^3} \nu^2 \frac{1}{e^{h\nu/kT} - 1}$$

$$= \frac{8\pi h \nu^3}{c^3 \left(e^{h\nu/kT} - 1\right)} \quad . \tag{12.5}$$

This is *Planck's Law*, plotted in Figure 12.1.

The total energy per volume is found by integrating over the spectrum,

$$\int_0^\infty U(\nu)\, d\nu = \frac{8\pi h}{c^3} \int_0^\infty \frac{\nu^3}{e^{h\nu/kT} - 1}\, d\nu$$

$$= \frac{8\pi h}{c^3} \frac{1}{15} \left(\frac{kT\pi}{h}\right)^4 \quad . \tag{12.6}$$

This is done analytically in terms of the *Riemann zeta function*, arguably the most interesting integral in all of mathematics [Hardy & Wright, 1998].

Planck's Law applies to photons in thermal equilibrium, for example in a closed cavity with walls at a temperature T. If we open a hole in the cavity we'll disturb the distribution, but as long as the hole is not too large it can be used to sample the radiation. This idealization is called a *black-body* radiator, because any light entering the hole will have little chance of scattering out so it is an almost-ideal absorber and emitter. The total power per area R radiated from the hole is found by multiplying equation (12.6) by the speed of light c (to convert energy per volume to energy per time per area), dividing by 2 (because half the photons are headed towards the hole, and half away), and dividing by another factor of 2 (because the effective area of the opening must be scaled by the dot product of the surface normal with the uniformly-distributed photon orientations $\int_{-\pi/2}^{\pi/2} \cos\theta \; d\theta = 2$), giving

$$\begin{aligned} R &= \frac{c}{4} \int_0^\infty U(\nu) \, d\nu \\ &= \frac{\pi^2 k^4}{60 \hbar^3 c^2} T^4 \\ &\equiv \sigma T^4 \\ &= 5.67 \times 10^{-8} \; T^4 \; \frac{\mathrm{W}}{\mathrm{m}^2} \quad . \end{aligned} \tag{12.7}$$

This is the *Stefan–Boltzmann* Law. For real surfaces it must be corrected for the *emissivity* deviating from the black-body idealization of unit efficiency, but for a wide range of materials it is a good approximation. The presence of Planck's constant h in the formula indicates its quantum origin. The inability to derive the correct form for thermal radiation during the latter part of the 19th century, at a time when physics was widely viewed as having been completed as a theory, was an irritation that lead to a revolution with the development of quantum mechanics by Einstein and others.

12.1.2 Luminescence: LEDs, Lasers, and Flat Panels

Light produced by quantum transitions rather than thermal means is called *luminescence*. If the excitation mechanism is an electrical current or voltage it is called *electroluminescence*, if the photons are produced by electron bombardment it is called *cathodoluminescence*, and if the excitation is by photons it is called *photoluminescence*. When the decay time is fast (on the order of the nanosecond time scales for direct electron–hole recombination) the radiation is called *fluorescence*, and if the decay is slow (seconds, minutes, even hours) it is called *phosphorescence* and the material is called a *phosphor* [McKittrick *et al*., 1999].

We'll primarily be concerned here with the electroluminesence of semiconductor devices. The most important example of cathodoluminescence is the familiar *Cathode Ray Tube* (*CRT*), in which electrons emitted from a heated cathode are accelerated by an anode to strike a phosphor. A typical phosphor is ZnS doped with Cu as an *activator*; excited conduction electrons make a 530 nm (green) transition at a Cu ion. An increasingly significant application of photoluminescence is in *optical repeaters* for long-haul fiber links. These use silica fibers doped with erbium ions, which can be pumped by 980

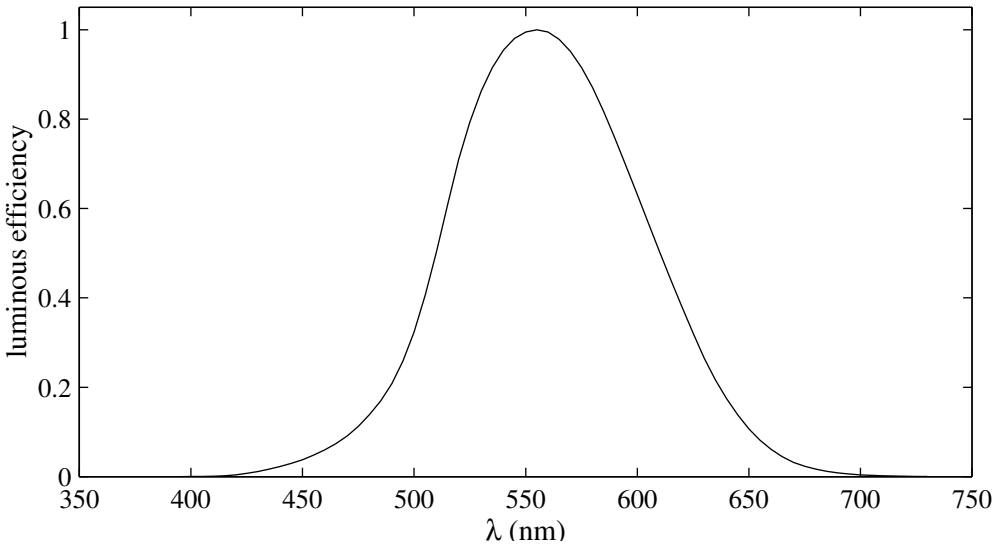

Figure 12.2. CIE luminosity function.

or 1480 nm light from a local laser into a metastable state with a transition that matches the 1.5 μm absorption minimum used for communications. When a signal photon arrives it stimulates emission from this excited state, providing gain in an all-optical system. This eliminates the losses, noise, and complexity of detecting the light, amplifying it electronically, and then generating it again [Delavaux & Nagel, 1995].

The apparent brightness of any light source is measured in *lumens*. According to the definition given in Chapter 2, one watt of 555 nm (green) light corrsponds to 683 lumens (lm). Unfortunately, for any other kind of light the meaning is less straightforward. In 1924 the *Commission Internationale de l'Eclairage* (*CIE*) used early perceptual studies to define a standard *luminosity function*, shown in Figure 12.2, that gives the relative sensitivity of the eye to wavelengths away from the 555 nm peak. The spectrum of a broadband light source must be weighted by this curve to determine its value in lumens. Even worse, later studies have shown that this curve underestimates the eye's sensitivity to short wavelengths, so lumen measurements are sometimes reported with more modern weightings. This is analogous to the ambiguity possible in specifying the reference level used for a decibel measurement, but now a function must be given.

Because of the eye's non-ideal response, an ideal source of white light over the visible range would produce about 200 lm per watt of power. A typical 75 W light bulb produces 1200 lm, giving 10–20 lm/W for incandescent sources. Fluorescent bulbs raise this to 50–100 lm/W by replacing electronic heating with electronic excitation of a mercury plamsa which releases ultraviolet photons that pump a phosphor coating. The most efficient lamps eliminate the down-conversion loss by directly using atomic transitions in a sodium vapor, approaching 200 lm/W [Hollister, 1987].

Semiconductors can display bulk electroluminescent effects through a range of mechanisms including carrier impact scattering, field emission around defects, and nanoscale quantum confinement in *porous silicon* [Fauchet, 1998]. More efficient, predictable, and

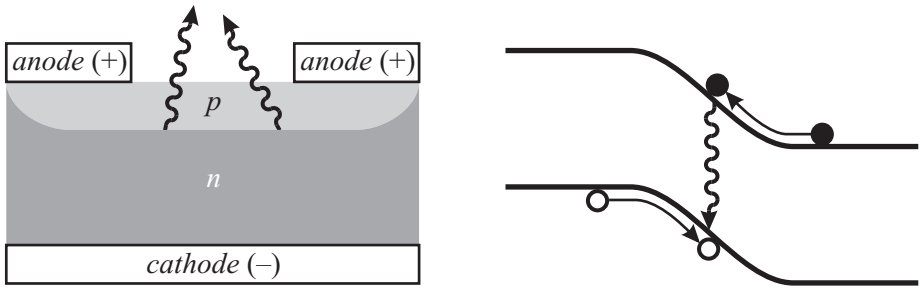

Figure 12.3. A Light–Emitting Diode.

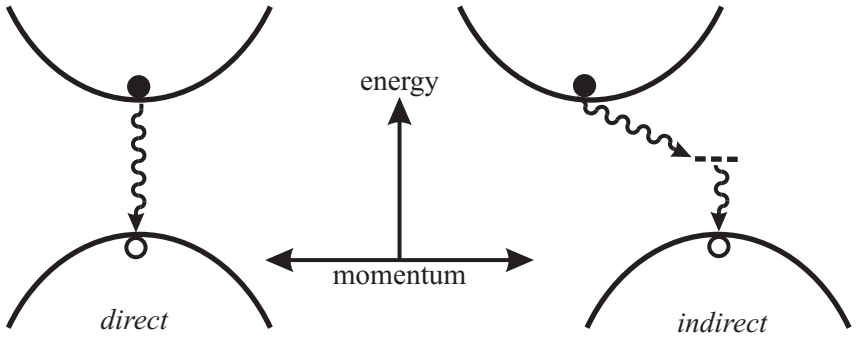

Figure 12.4. Semiconductor band gaps.

controllable is the *injection electroluminescence* associated with electrons and holes recombining at a p–n junction. In a *Light-Emitting Diode* (*LED*), shown in Figure 12.3, a photon is produced when a conduction electron falls into a valence hole. The junction is forward biased to drive excess carriers into the junction region, which unlike a conventional diode is wide and shallow to enhance the production and emission of light.

Some semiconductors, such as GaAs, have *direct band gaps* in which the conduction band minimum has the same crystal momentum as the valence band maximum. In others such as Si these are displaced, as shown in Figure 12.4. These *indirect band gap* materials make very inefficient emitters of light, because momentum conservation requires a phonon for electron–hole recombination. This reduces the recombination probability as well as releases energy to the lattice and broadens the radiation linewidth. For this reason, optoelectronics almost exclusively uses direct band gap semiconductors.

In $GaAs_{1-x}P_x$, as x is varied from 0 to 0.45, the band gap changes from 1.4 eV (IR) to 2 eV (red). This lets the color of an LED be selected by the composition. At higher concentrations the gap becomes indirect, but nitrogen impurity doping is used to introduce gap states that let concentrations up to $x = 1$ be used (2.2 eV, green). GaN has a larger direct band gap of 3.4 eV, making blue and even UV LEDs possible [Mukai *et al*., 1999]. Through improvements in band-structure engineering and light collection, LEDs have become competitive beyond information displays as direct sources of illumination; AlInGaP LEDs have been produced with outputs over 10 lm at efficiencies over 20 lm/W [Fletcher *et al*., 1993]. Less efficient but more versatile are *Organic Light–Emitting*

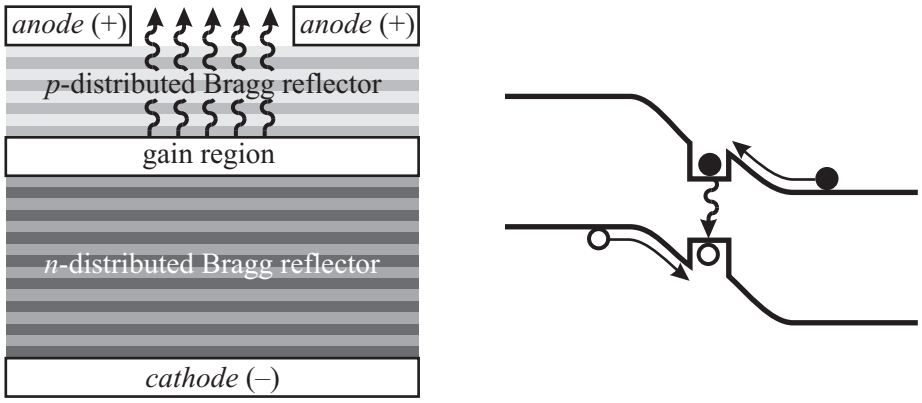

Figure 12.5. A Vertical-Cavity Surface-Emitting Laser.

Diodes (*OLEDs*), which can be produced as thin, flexible devices [Sheats *et al*., 1996; Friend *et al*., 1999].

A typical linewidth for an LED is 100 Å, which is too broad for fiber-optic links that rely on low dispersion and on wavelength-division multiplexing, and the lack of control over where the photons are emitted prevents their use in the diffraction-limited optics needed for storage applications. These are among the many reasons for the growing importance of *diode lasers*, which can provide linewidths below 1 Å with fundamental beam mode-shapes.

Lasers rely on *stimulated emission* [Corney, 1978]. If a system such as an atom or a electron–hole pair that can make a radiative transition is pumped into its excited state, then an incident photon at that frequency can stimulate a transition to the ground state with the emission of a photon, in the inverse process to the absorption of a photon driving a transition to the excited state. And because this is a resonant effect, the emitted photon matches the phase of the incident one. The result is two phase-coherent photons instead of one; if they can be kept around long enough there will be net optical gain with the mode shape determined by the mirrors defining the optical cavity.

Lasing requires maintaining a *population inversion* of excited states, and needs high-reflectivity mirrors that let photons pass through the gain medium many times. An elegant semiconductor solution is the *Vertical-Cavity Surface-Emitting Laser* (*VCSEL*), shown in Figure 12.5 [Lott *et al*., 1993; Someya *et al*., 1999].

This is still a p–n diode just like an LED, but the junction is now sandwiched between two mirrors. These are *Distributed Bragg Reflectors* (*DBRs*), periodic quarter-wavelength dielectric layers that scatter coherently at their interfaces, the opposite of an antireflection coating (Problem 9.6). The index of refraction can be controlled by varying the composition of $Al_xGa_{1-x}As$. Not only does this let the mirror also serve as a part of the semiconductor junction that can still be doped, it avoids the losses due to the conductivity of a metal mirror. The lower mirror reflectivity can be better than 99%; the upper one is intentionally slightly lower to couple some light out. The heart of the junction itself consists of undoped GaAs layers, called *quantum wells*. Because these have a smaller band gap, the carriers being injected across the junction by the forward bias

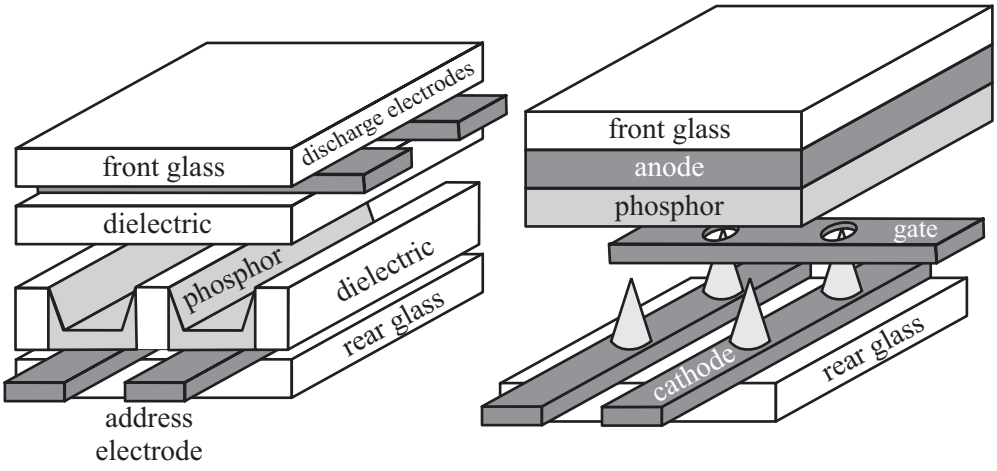

Figure 12.6. A Plasma Display Panel (left) and a Field Emission Display (right).

drop into these thin high-mobility layers where they can easily recombine. Since there is more than one interface, this is called a *heterojunction* device or a *heterostructure*.

The thickness of the gain region is chosen to support a single longitudinal mode, and the transverse mode structure is determined by the lateral shape. Because of the laser's gain, all of the light that emerges comes from this same cavity mode, providing the desired narrow linewidth and beam shape. VCSELs makes such good use of the carriers that, unlike early semiconductor lasers, just a few volts and mA are adequate for lasing at room temperature. The conversion efficiency from electricity to light can be over 50%. Because the light emerges from the top of the laser, these are easily integrated with other devices on a chip, and by coupling them into arrays it is possible to generate watts of output power.

Multiple light-emitting elements can be combined to form an *emissive display* (in Section 12.3 we'll look at alternatives that modulate rather than generate light). While many diodes can be fabricated on a single substrate, making a display this way would entail covering large areas with expensive silicon wafers. The challenge in developing displays is to find scalable technologies that can match the performance of the eye, which can resolve a spatial frequency of 60 cycles per degree [Campbell & Green, 1966], and compete with ambient light ranging from 100 cd/m^2 indoors to 10 000 cd/m^2 in sunlight. Two approaches that build on familiar light sources have received particular attention (Figure 12.6): *Plasma Display Panels* (*PDPs*) and *Field Emission Displays* (*FEDs*).

A PDP comprises an enormous number of tiny fluorescent lights [Bitzer, 1999]. Dielectric channels are coated with a phosphor and filled with a combination of inert gases such as He, Ne, and Xe. In columns above the channels are pairs of clear electrodes, encapsulated in a protective dielectric layer. An AC voltage is applied to them that is just below the breakdown voltage needed to ionize the gas. Below the channels are row electrodes that are used to turn the discharge on and off, creating plasmas that excite the phosphor with ultraviolet light. Because it's not possible to exercise fine control over the UV intensity, PDPs cycle the plasma many times in microsecond pulses to vary the brightness. Even though the physical structure is (relatively) straightforward,

this requires complex control strategies to prime the discharge because of the time scales associated with the surface charge in the channel and the ionization state of the gas [Rauf & Kushner, 1999]. PDPs are attractive because they can be fabricated over large areas, and can match the brightness of fluorescent lighting because they *are* fluorescent lighting, although their efficiency drops to lumens per watt because of the losses associated with the small channel size.

A more efficient alternative is an FED, which can be thought of as a huge number of tiny cathode ray tubes [Ghrayeb *et al*., 1997]. Although CRTs are a mature technology, they are inefficient because of the thermionc electron emission, and bulky because of the electron deflection optics. An FED replaces the single cathode with a huge number of tiny sharp metal tips. An electric field is applied by column gate electrodes above the rows of tips. Because the field gradient is significantly increased around the tip, the local potential difference produced by just a few volts can exceed the work function of the tip, causing it to emit electrons by *field emission* [Phillips *et al*., 1998]. These are then accelerated towards a phosphor by an anode, much like a CRT but now each pixel has its own emitters. Although FEDs are more difficult to manufacture than PDPs, the electronic control is straightforward, and the efficiency is increased over 10 lm/W because there are no thermal or down-conversion losses.

12.2 DETECTION

The fundamental processes that generate photons from electrons can be run in reverse, to convert photons to electrical signals. Simplest of all is a *photoconductor*. This mechanism was discovered in the early days of studying semiconductors, when a curious oscillation developed in the conductivity of a sample that was finally explained by the shadow of a ceiling fan rotating above it. Photons can excite carriers across the conduction band, reducing the resistance of the material. The energy of the photons of interest must be larger than the gap energy; for visible light common photoconductors include CdS (2.4 eV, 0.52 μm) and CdSe (1.8 eV, 0.69 μm). For longer wavelengths lower-energy excitations must used; HgCdTe with a gap around 0.12 eV is used for detection of infrared light below 10 μm. Beyond that, it's possible to introduce gap states with dopants that provide lower-energy excitations. Hg in Ge is an acceptor that sites sit 0.09 eV (14 μm) above the valence band, and Cu acceptors in Ge are 0.04 eV (32 μm) above it. Valence band electrons that are excited into these acceptors leave holes behind that increase the conductivity. At these low energies, the number of carriers produced by thermal excitation becomes significant compared to those excited by weak optical signals, requiring that the detectors be cooled. This can be done with a Peltier cooler (Chapter 15), liquid nitrogen (77 K), or liquid helium (4.2 K). .

The dominant noise mechanism in uncooled photoconductors is their $4kTR$ Johnson noise, which is intrinsic to this kind of detector because it works by measuring a resistance. A quieter alternative is to use an LED in reverse. An incoming photon can excite an electron–hole pair in the depletion region, which will then be swept apart by the junction field and measured as a current. Because carrier diffusion is a slow process, a faster variant is the *p–i–n photodiode*, shown in Figure 12.7. Thin p and n layers are sandwiched around a thicker insulating layer, expanding the depletion region to fill most of the diode. Now

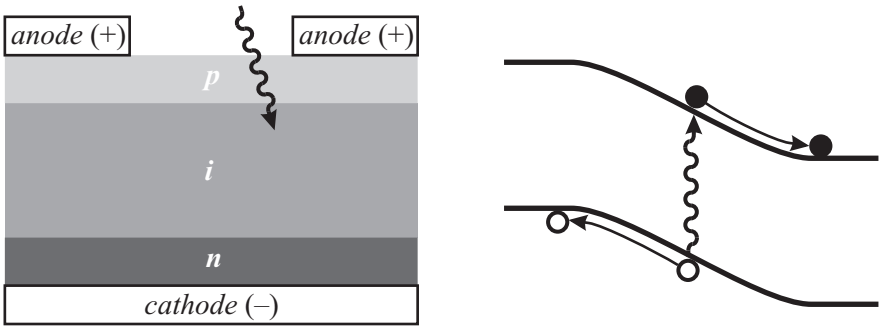

Figure 12.7. A p–i–n photodiode.

photocarriers are accelerated out by the junction field, and just have to diffuse through the thin doped layers. These devices can have response times approaching picoseconds.

If the power of the illumination of a photodiode is P, and it produces a current I, then the *quantum efficiency* η is defined to be

$$\eta = \frac{I/e}{P/h\nu} \quad . \tag{12.8}$$

The numerator divides the current by the charge to find the rate of electron–hole pair production, and the denominator divides the power by the photon energy to give the rate of photon arrival. The ratio is the probability that a photon will contribute to the current. The quantum efficiency falls off at long wavelengths because the photon energy is smaller than the gap energy, and at short wavelengths because the photon is absorbed before reaching the depletion region, producing an electron–hole pair that can recombine. At its peak, the quantum efficiency can approach unity; for Si this happens around 0.8 μm.

Photodiodes are ultimately limited by the shot noise of the photocurrent due to fluctuations in the rate of photon arrival. On top of this, there is noise from the *dark current* due to thermal carrier excitation, and Johnson noise associated with the load. While the shot noise limit is fundamental, the other two are instrumental and can be eliminated by *heterodyne* detection. This clever trick is the optical analog of electrical measurement techniques that will be covered in Chapter 14. The quantum transition probability to excite a transition is proportional to the intensity of the radiation, and hence the square of the electric field strength [Corney, 1978]. Heterodyne detection works by exploiting this nonlinearity. The idea is to add to the optical signal of interest, taken to have an electric field at the detector of $E_S e^{i\omega t}$, a much stronger local optical field $E_L e^{i[(\omega+\delta)t+\varphi]}$. The frequency shift δ is an intentional detuning, and φ is their relative phase. The total intensity is then

$$\begin{aligned} |E|^2 &= |EE^*| \\ &= E_S^2 + E_L^2 + E_S E_L \left(e^{i[\omega t - (\omega+\delta)t - \varphi]} + e^{i[-\omega t + (\omega+\delta)t + \varphi]} \right) \\ &= E_S^2 + E_L^2 + 2E_S E_L \cos(\delta t + \varphi) \\ &\approx E_L^2 \left(1 + 2\sqrt{\frac{E_S}{E_L}} \cos(\delta t + \varphi) \right) \end{aligned} \tag{12.9}$$

in the limit $E_L \gg E_S$. The magic happens in the product, where the fluctuations in the optical signal are scaled up by the much stronger local oscillator field, bringing them above the level of the photodiode's dark current and Johnson noise. The detected current will then just be proportional to equation (12.9), with the coefficient found from equation (12.8), so that in terms of the signal and local-oscillator powers

$$I = \frac{P_L \eta e}{h\nu}\left[1 + 2\sqrt{\frac{P_S}{P_L}}\cos(\delta t + \varphi)\right] \quad . \tag{12.10}$$

The magnitude of the current signal is then

$$\begin{aligned} S &= \left\langle \left(I - \langle I \rangle\right)^2 \right\rangle \\ &= \left(\frac{P_L \eta e}{h\nu}\right)^2 2\frac{P_S}{P_L} \quad . \end{aligned} \tag{12.11}$$

Because the detected current is shot-noise-limited (equation 3.33), the current noise magnitude is

$$\begin{aligned} N &= 2e\langle I \rangle \Delta f \\ &= 2e\frac{P_L \eta e}{h\nu}\Delta f \quad , \end{aligned} \tag{12.12}$$

where Δf is the measurement bandwidth. This gives a quantum-limited SNR of

$$\frac{S}{N} = \frac{P_S \eta}{h\nu \Delta f} \quad . \tag{12.13}$$

For an SNR of 1, the photon arrival frequency equals the bandwidth of the detector.

Heterodyne detection does require a local light source matched to the signal. An alternative mechanism to improve sensitivity is used in an *Avalanche PhotoDiode* (*APD*). When a p–n junction is illuminated, its I–V curve is shifted down by the photocurrent. Photodiodes are usually operated reverse-biased, where this current depends on the light intensity but is independent of the bias voltage. As the reverse bias is increased, avalanche breakdown is reached. Just short of that, a photocarrier can get enough energy from the junction field to excite another carrier by *impact ionization*, leading to a cascade that produces many electrons from one photon. This can lead to a gain of 100 or more in the current, although this does come at the expense of a slower response (because of the collisions) and more noise (because the thermal dark current also gets amplified).

An interesting thing happens as the bias becomes positive in Figure 12.8: the IV product changes sign. This means that the diode becomes a net exporter rather than importer of energy, generating power in a *photovoltaic* or *solar cell* [Chapin *et al*., 1954]. For best efficiency, the load must be chosen to maximize the IV product in that quadrant. Then the efficiency of the solar cell is limited by the energy lost from missing those photons with an energy below the gap, and from thermalizing carriers produced by photons with an energy above the gap. For a single junction under solar illumination that results in a maximum efficiency of about 30%; strategies for raising the efficiency include stacking multiple junctions ranging from highest to lowest band gap, and using concentrators to collect photons from a larger area.

The converse to creating a display out of an array of LEDs is to use an array of photodetectors to record an image. The challenge is to integrate as many detectors as possible,

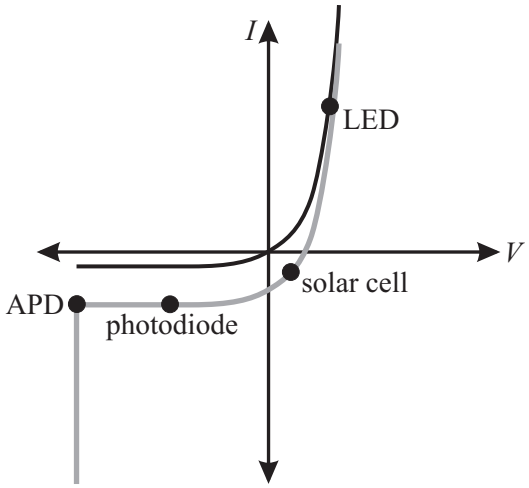

Figure 12.8. Biasing for a p–n junction.

while still managing to extract the signals from each device with acceptable fidelity. We've already seen one solution to this problem: a DRAM memory (Figure 11.16). A *CMOS imager* [Denyer *et al*., 1995] adds a photodetector to each cell, most simply just a MOS capacitor (Figure 11.9). The charge produced by the photodetector is read out just as the charge on a DRAM cell capacitor is. But now, rather than trying to maximize the capacitance, it's important to maximize the collection area of the photodetector. While this cell design, called a *Passive Pixel Sensor* (*PPS*), does do that by having just one transistor per cell, it suffers from noise and delay associated with charging a long read line with a small capacitor. The analog to SRAM is an *Active Pixel Sensor* (*APS*), which adds one or more transistors for *transimpedance* conversion of current to voltage with gain (Chapter 14). This comes at the expense of losing collection area for the photodetector, but that can be ameliorated by fabricating microlenses above the pixels. The remaining bane of CMOS imagers is *Fixed Pattern Noise* (*FPN*), the systematic image errors that come from pixel-to-pixel sensitivity variations and cross-talk in the readout lines. This is dealt with by schemes for differential readout and background subtraction.

Charge-Coupled Devices (*CCDs*) take advantage of the ability to manipulate the surface band structure to move charge out directly [Boyle & Smith, 1971]. The interfacial band-bending in an MOS capacitor is used in a MOSFET to introduce carriers into the conduction band, but it also forms a potential well that can store carriers that arrive by other means (Figure 11.9). A CCD pixel accumulates photo-induced charge in this well. But instead of reading it out through a wire, the pixels are connected as shown in Figure 12.9. The depth of the well is a function of the gate voltage, which at the beginning of a cycle is set to retain charge below every third electrode. Then, the potential on that line is dropped while it is raised on the neighboring cells, creating a single larger well that fills up with the charge. At the end of the cycle the second well is lowered while the first one is rasied, shifting the charge over by one pixel. By repeating this cyclic pattern, the charge on each pixel is sequentially shifted out of the end of the row. The long scan lines are now actively driven from the periphery of the chip, with each pixel needing only to

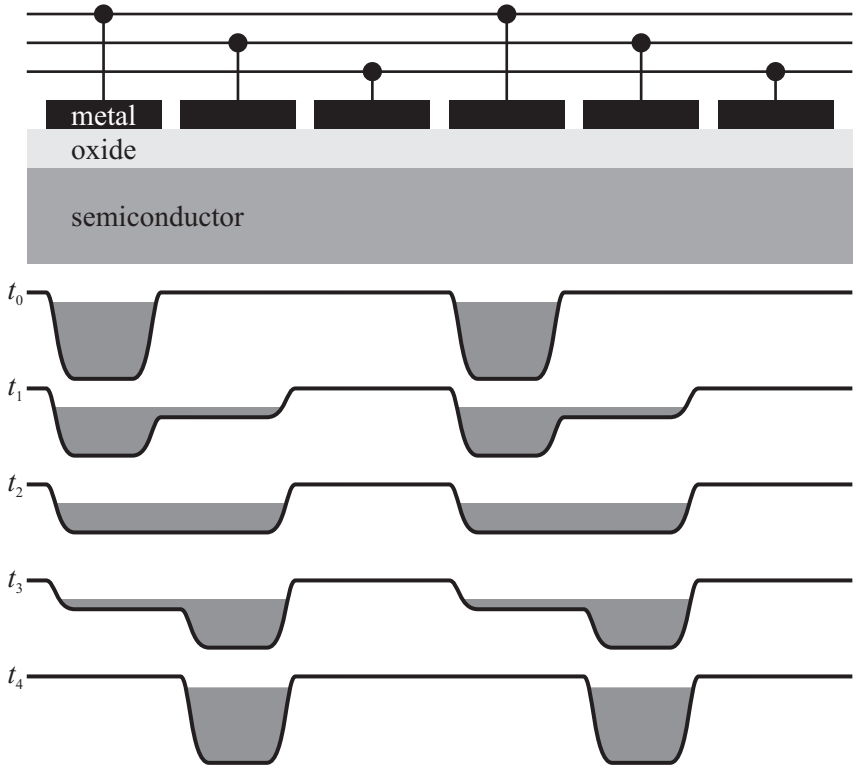

Figure 12.9. The operation of a CCD.

transfer charge to its neighbor. In practical CCDs there is an extra n-type layer on top of the p-type substrate, which forms a *buried channel* that moves the potential well down from the surface and its higher density of defect states. This CCD structure also finds application in analog memories and delay lines.

For low-light applications, CCDs are cooled to reduce the dark-current accumulation of thermally-induced charge, and they can be thinned to bring light in from the rear to completely avoid the losses associated with the wires and electrodes. With these enhancements, the quantum efficiency can approach 100% at matched wavelengths, and the noise introduced by the readout can be on the order of a single electron. The sensitivity is then determined solely by the collection time, permitting dimmer images to be recorded by accumulating charge for a longer time before scanning it out. Room-temperature CCDs read out at video rates cannot reach this sensitivity, but can still have readout noise of a few tens of electrons per pixel. A dominant contribution to this is the *reset noise* associated with resetting the readout circuit, which can be reduced by reading it out twice to perform a background subtraction.

Compared to CMOS imagers, CCDs offer good pixel density and noise performance, but they require higher power because of the charging currents associated with driving the readout cycle, and the device optimizations are not compatible with conventional CMOS design, requiring specialized fabs to make them and supporting chips to interface to them. CMOS imagers offer random pixel access and on-chip integration of related

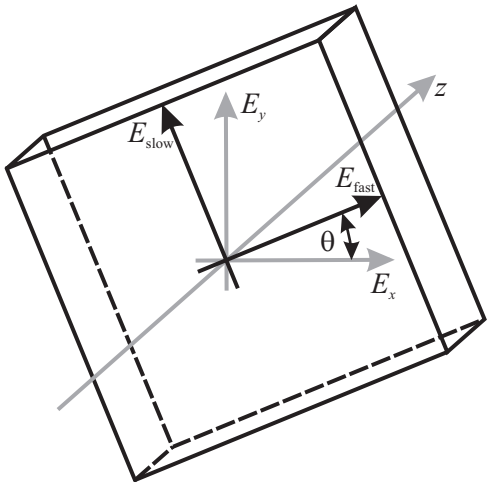

Figure 12.10. Axes in a birefringent crystal.

functions, historically with lower power, cost, and performance. But as this technology matures the performance difference is disappearing.

12.3 MODULATION

We've examined how to generate and detect light; this final section will look at passive and active mechanisms to modulate it.

12.3.1 Polarization

The eye cannot see polarization directly, but in many applications it does see the consequences of manipulating polarization states. This can be done with a *birefringent* material, one with an anisotropic ordering that causes its polarizibility ϵ and hence index of refraction n to depend on the orientation. The simplest of these are *uniaxial* materials that have a single optical axis with orthogonal fast and slow directions. For *calcite* ($CaCO_3$), these two refractive indices are $n_{\text{slow}} = 1.658$ and $n_{\text{fast}} = 1.486$, giving a birefringence difference of 0.172. *Quartz* (SiO_2) is less birefringent, with a difference of 0.009.

Just as ray matrices simplified the description of a series of optical elements in Section 9.2.1, the *Jones calculus* does the same for polarizing materials. If the transverse electric field components of a wave are $E_x e^{i\omega t}\hat{x} + E_y e^{i\omega t}\hat{y}$, we'll write the complex coefficients as a two-component vector (Figure 12.10)

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad . \tag{12.14}$$

If E_y/E_x is pure real the wave is said to be *linearly polarized* because the components move back and forth in phase; if $E_y/E_x = i$ or a multiple of it, the wave is *circularly polarized* because the vector rotates around a circle; and between these cases it is *elliptically polarized*.

The field components relative to the axes of a birefringent material are found with a rotation matrix

$$\begin{pmatrix} E_{\text{slow}} \\ E_{\text{fast}} \end{pmatrix} = \underbrace{\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}}_{\equiv \mathbf{R}(\theta)} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad . \tag{12.15}$$

Since the wave vector is $k = 2\pi/\lambda = n\omega/c$, the field components after propagating through a thickness of d will pick up a phase shift of e^{ikd} along each axis

$$\begin{pmatrix} E_{\text{slow}} \\ E_{\text{fast}} \end{pmatrix}' = \begin{pmatrix} e^{-in_{\text{slow}}\omega d/c} & 0 \\ 0 & e^{-in_{\text{fast}}\omega d/c} \end{pmatrix} \begin{pmatrix} E_{\text{slow}} \\ E_{\text{fast}} \end{pmatrix} \quad . \tag{12.16}$$

This can be written more symmetrically in terms of the sum

$$\sigma = (n_{\text{slow}} + n_{\text{fast}})\frac{\omega d}{2c} \tag{12.17}$$

and the difference

$$\delta = (n_{\text{slow}} - n_{\text{fast}})\frac{\omega d}{2c} \tag{12.18}$$

as

$$\begin{pmatrix} E_{\text{slow}} \\ E_{\text{fast}} \end{pmatrix}' = \underbrace{e^{-i\sigma}\begin{pmatrix} e^{-i\delta} & 0 \\ 0 & e^{i\delta} \end{pmatrix}}_{\equiv \mathbf{B}(d)} \begin{pmatrix} E_{\text{slow}} \\ E_{\text{fast}} \end{pmatrix} \quad . \tag{12.19}$$

The phase prefactor $e^{-i\sigma}$ can be left out unless the light will later be recombined with a reference beam. For a wave polarized along the laboratory axes, the change after passing through a birefringent material is found by rotating to the optical axes, applying the birefringence matrix, and then rotating back:

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix}' = \mathbf{R}(-\theta)\mathbf{B}(d)\mathbf{R}(\theta) \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad . \tag{12.20}$$

A *dichroic* material has absorption coefficients that depend on polarization; there are both linearly- and circularly-polarized dichroics. If a linear dichroic material completely absorbs one component while passing the other, it is a *linear polarizer* with a Jones matrix

$$\mathbf{L} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{12.21}$$

(ignoring the phase prefactor). Edwin Land developed synthetic polarizers using *herapathite*, which forms dichroic crystals that were discovered through the rather unusual laboratory accident of dropping iodine into the urine of dogs fed quinine [Land, 1951]. More stable polarizers are made from stretched sheets of *PolyVinyl Alcohol* (*PVA*) reacted with iodine.

In some magnetic materials, left- and right-circularly-polarized waves travel at different

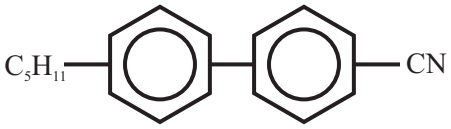

Figure 12.11. The liquid crystal 5CB.

speeds. A circularly-polarized basis is related to the linear one by

$$\begin{pmatrix} E_+ \\ E_- \end{pmatrix} = \underbrace{\frac{1}{2}\begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}}_{\mathbf{C}} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad . \tag{12.22}$$

Analogous to birefringence, after passing through a magnetic material these components become

$$\begin{pmatrix} E_+ \\ E_- \end{pmatrix}' = \begin{pmatrix} e^{i\theta_F} & 0 \\ 0 & e^{-i\theta_F} \end{pmatrix} \begin{pmatrix} E_+ \\ E_- \end{pmatrix} \quad , \tag{12.23}$$

where $\theta_F = (n_- - n_+)\omega d/2c$ is the *Faraday rotation* angle. In ferrite materials, $\theta_F = VBd$, where d is the thickness, B is an applied DC magnetic field, and V is the *Verdet constant*. For *Yttrium Iron Garnet* (*YIG*), $Y_3Fe_5O_{12}$, $V = 0.1\ ^\circ/\text{G}\cdot\text{cm}$

In the linear basis, Faraday rotation is

$$\begin{aligned} \begin{pmatrix} E_x \\ E_y \end{pmatrix}' &= \mathbf{C}^{-1} \begin{pmatrix} e^{i\theta_F} & 0 \\ 0 & e^{-i\theta_F} \end{pmatrix} \mathbf{C} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta_F & \sin\theta_F \\ -\sin\theta_F & \cos\theta_F \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \\ &= \mathbf{R}(\theta_F) \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad . \end{aligned} \tag{12.24}$$

This is simply a rotation by the Faraday angle (hence the name). A magnetic material that rotates polarization by 45° and that is between linear polarizers rotated relative to each other by that angle is called a *Faraday isolator*: linearly polarized light can pass in one direction but not the other. This violation of reversibility is possible because magnetic interactions change sign under time reversal. Faraday isolators are used for preventing light from coupling back into lasers, which is important for their mode structure and stability, and for selecting a lasing direction in a symmetrical ring laser.

12.3.2 Liquid Crystals

The most visible application of polarization is in *Liquid Crystal Displays* displays (*LCDs*). Liquid crystals are fluids with order intermediate between the long-range periodicity of crystals and the short-range correlations of ordinary liquids.

An example of a liquid crystal molecule is *5CB* (4-pentyl-4′-cyanobiphenyl), shown in Figure 12.11. The hexagons are *benzene rings*, with the circles showing the resonance between the two equivalent ways to alternate single and double bonds between the carbon atoms on the vertices. The long axis of this anisotropic molecule is called the *director*.

5CB forms a *nematic* liquid crystal, in which the positions of the molecules are random,

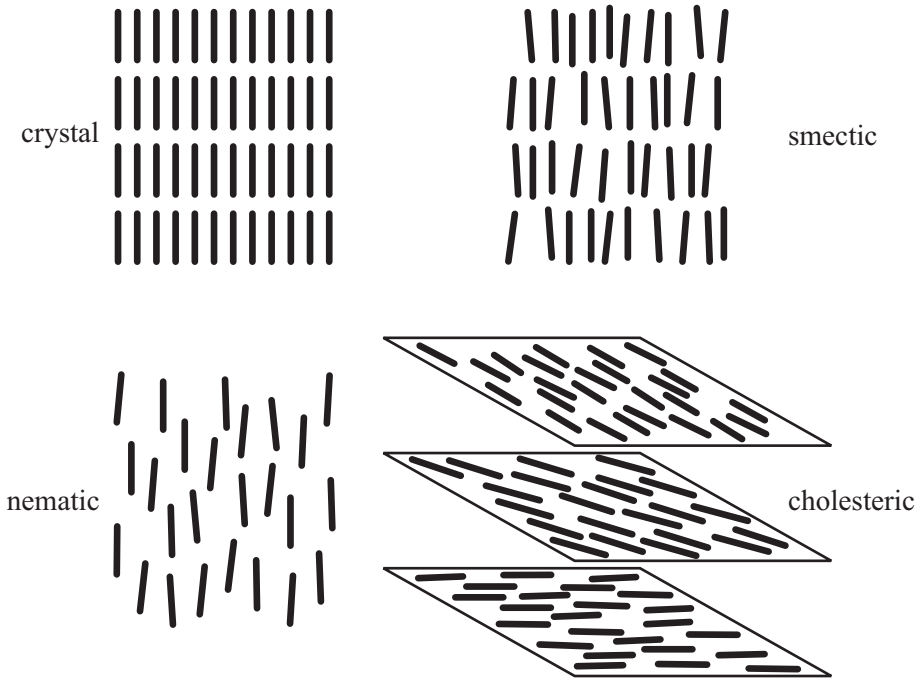

Figure 12.12. Liquid crystal ordering.

but they all point in the same direction. A *smectic* liquid crystal shares this long-range orientational order, and in addition the molecules are layered in planes. And in a *Cholesteric* the positions are random, but the directors are aligned and twist in a helix (Figure 12.12).

The dipole moment of an anisotropic liquid crystal can be used to electrically switch its orientation, in the scheme shown in Figure 12.13 [Schadt & Helfrich, 1971]. The liquid crystal is contained between two glass plates with a spacing on the order of 10 μm. The inner surfaces have a thin coating of the polymer *polyimide*, which has been rubbed with a cloth in one direction to produce nanometer-scale grooves. This rubbing is one of a number of "black magic" procedures in LCD production, which are poorly understood but essential operations that are more likely to be considered proprietary trade secrets than research topics.

The director will line up with the grooves because there would be a bending energy associated with crossing them. In a *Twisted Nematic* (*TN*) display the plates are rotated by 90°. This induces a net rotation, like a cholesteric, but it is the result of boundary conditions applied to a nematic. Because there are two possible rotation directions, a small amount of cholesteric is in fact added to break that symmetry. The glass plates have clear electrodes deposited on them, usually *Indium Tin Oxide* (*ITO*). Because the plates are so close together, when a few volts are applied across them the electrostatic dipole energy becomes more significant than the liquid crystal orientational energy, and the molecules rotate to align with the field.

Because of the anisotropy, liquid crystals are also birefringent. When the cell is in the twisted state, it can be considered to be a stack of infinitesimally-thick rotated birefringent plates. If the state of the light coming into the cell is $\vec{E}_0$, then after passing through the

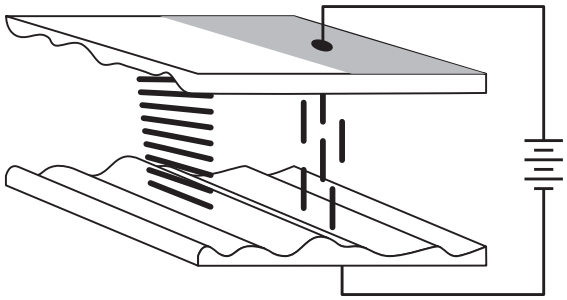

Figure 12.13. A twisted-nematic liquid crystal display.

first of these layers it is

$$\vec{E}_1 = \mathbf{R}^{-1}(\theta)\mathbf{B}(d)\mathbf{R}(\theta)\ \vec{E}_0 \quad , \tag{12.25}$$

where d is the layer thickness and θ is the angle change over the layer. After two layers,

$$\begin{aligned}\vec{E}_2 &= \mathbf{R}^{-2}(\theta)\mathbf{B}(d)\mathbf{R}^2(\theta)\ \mathbf{R}^{-1}(\theta)\mathbf{B}(d)\mathbf{R}(\theta)\ \vec{E}_0 \\ &= \mathbf{R}^{-2}(\theta)[\mathbf{B}(d)\mathbf{R}(\theta)]^2\vec{E}_0 \quad ,\end{aligned} \tag{12.26}$$

and after N layers

$$\vec{E}_N = \mathbf{R}^{-N}(\theta)[\mathbf{B}(d)\mathbf{R}(\theta)]^N\vec{E}_0 \quad . \tag{12.27}$$

If both θ and d are small, this reduces to [Chandrasekhar, 1992]

$$\mathbf{E}_N = \mathbf{R}(N\theta)\mathbf{B}(Nd)\ \vec{E}_0 \quad . \tag{12.28}$$

In this *adiabatic* limit the light rotates with the pitch of the liquid crystal, also picking up the phase shift of the unrotated cell's thickness. If crossed polarizing filters are put before and after the cell, aligned with the direction of the polyimide texture, then when no voltage is applied the polarized light exiting the first filter will be rotated to pass through the second. But when the molecules align with an applied voltage, they no longer rotate the light and the second filter blocks the transmission. This provides a switchable light valve based on moving molecules rather than macroscopic materials.

TN displays are addressed with row and column electrodes that rely on each pixel's nonlinear response to the field to isolate the part of the drive waveform intended for it. This limits the size of the display because the on–off voltage ratio becomes too small as the number of pixels is increased [Alt & Pleshko, 1974], reducing the contrast and increasing the switching time. For this reason, twisted nematics are used in LCDs for applications such as watches and control panels, but not in larger computer screens. One way to increase the resolution is by decreasing the voltage range over which the cell switches, which is done in a *Super-Twisted Nematic* (*STN*) by using a twist angle of 270° instead of 90°. The larger index of refraction change also leads to a chromatic aberattion, giving an objectionable color difference between the off and on states. This is eliminated in a *Double Super-Twisted Nematic* (*DSTN*) display by adding a second index-compensating film or LCD layer. Note that the same acronym is used in a *Dual Scan Twised Nematic*, which splits the display into subpanels that are addressed separately.

DSTN displays can reach hundreds but not thousands of pixels. For that, it's necessary

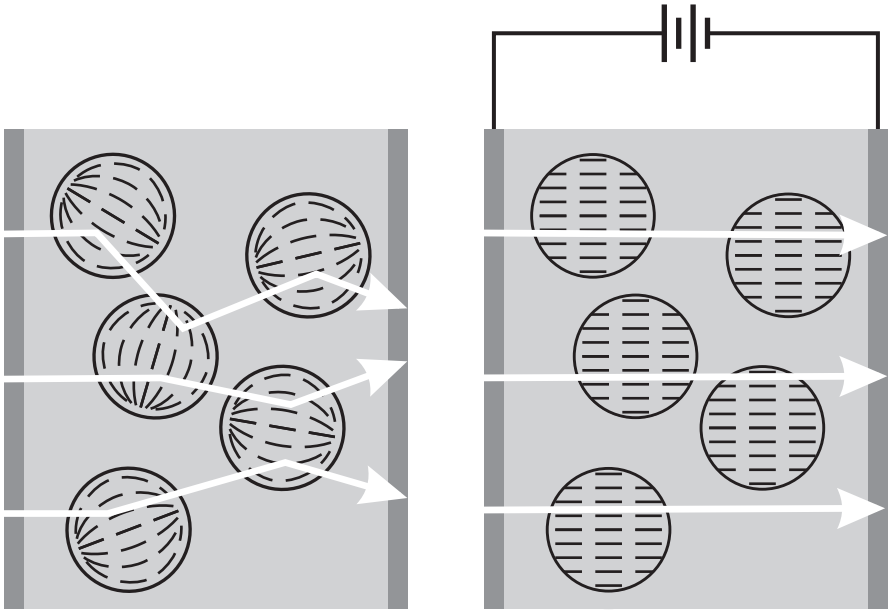

Figure 12.14. A polymer-dispersed liquid crystal panel.

to use some kind of active switch. A *Thin Film Transistor* (*TFT*) LCD does this with the same addressing scheme as a DRAM memory (Figure 11.15), where the capacitor becomes the pixel electrodes [Fischer *et al*., 1972; Brody, 1996]. This brings the contrast up from around 10:1 to 100:1, and the switching time down from about 100 ms to 10 ms. The transistors have been made using *amorphous silicon* (*a-Si*) with a *silicon nitride* (Si_3N_4) gate deposited on the glass, which has a mobility on the order of 1 $cm^2/(V \cdot s)$, and increasingly with *Polycrystalline silicon* (*p-Si*) because its mobility of $\sim$ 100 $cm^2/(V \cdot s)$ is close enough to that of single-crystal silicon (over 1000 $cm^2/(V \cdot s)$) for some of the supporting electronics to be integrated in the same process.

Manufacturing TFT panels requires lithographic fabrication over large areas, bringing down the yield (and increasing the cost) of acceptable panels because defects are so easy to see. Another limition of TFT panels is their power consumption: after passing through the polarizing filters, the liquid crystal, the electrodes and drive transistors, and the color filters, less than 10% of the light makes it out.

One approach to reducing the cost is to take advantage of existing CMOS processes to make small displays that are used with external optics. This is done in a *Liquid Crystal On Silicon* (*LCOS*) display by putting the liquid crystal on top of a CMOS wafer and using it in a reflection mode. A benefit of this approach is that the pixel spacing can become comparable to the wavelength of light, letting the display control color and optical elements by using diffractive structures [Alvelda & Lewis, 1998].

In the other direction, *Polymer-Dispersed Liquid Crystals* (*PDLCs*) are used to cover large areas, such as electronically-controllable windows [Fergason, 1985]. The idea is shown in Figure 12.14. The liquid crystal is contained in small voids in a polymer matrix. With no field applied, the directors line up based on the local asymmetry in their environment, giving an random distribution of orientations. This causes light to scatter

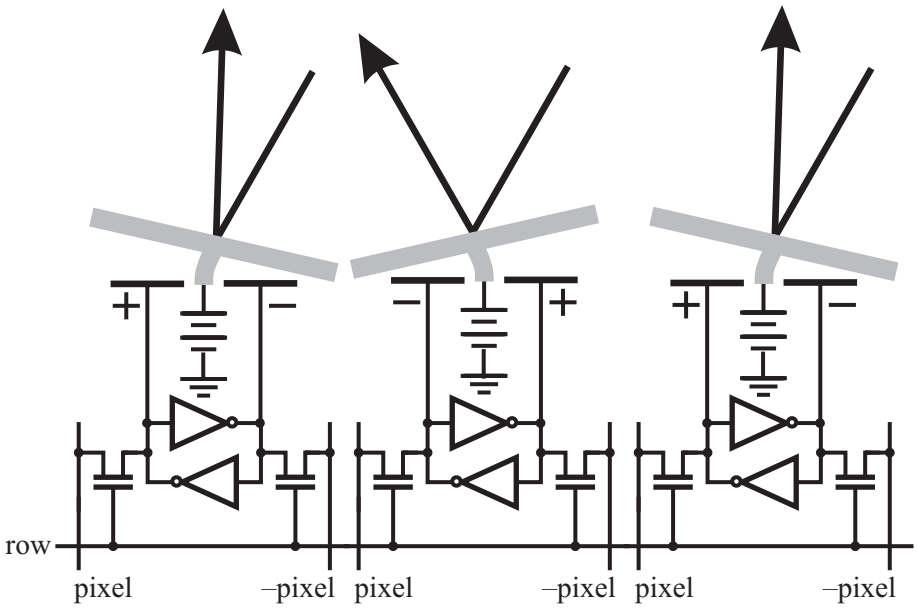

Figure 12.15. A Digital Micromirror Device.

many times, allowing it to be separated out in an optical system, or obscuring what is behind a window. When a voltage is applied to the electrodes, the dipole orientational energy once again dominates and the directors align. If the polymer is index-matched to the liquid crystal then light can pass straight through, switching the panel from opaque to clear.

12.3.3 Smoke and Mirrors

Small particles and shiny surfaces were among the first means used to modulate light; this final section will look at some of the new ways they are being reinvented to address serious limitations in more (recently) conventional displays.

Video projection is growing in importance, initially for presentations to groups, increasingly as a replacement to film in theaters, and ultimately as a way to illuminate smart spaces [Underkoffler *et al*., 1999]. In the last section we saw that about 10% of the light incident on a liquid crystal panel makes it through; all of the rest is dissipated internally. This represents a significant heat load in a display that is required to produce thousands of lumens, which is a particularly serious issue for the long-term stability of optical materials. Another problem with liquid crystals for projection applications is the display area lost to addressing and TFTs, which can be apparent when the pixels are magnified many times. And for video applications at 60 frames per second, the 17 ms switching time per frame is close to the time scales required to establish the molecular alignment, leading to blurring artifacts.

Figure 12.15 shows an alternative that is easy to understand but hard to implement, a *Digital Micromirror Device* (*DMD*). This starts with the layout of an SRAM memory, but then fabricates above it electrodes on either side of the inverters, and a mirror on a

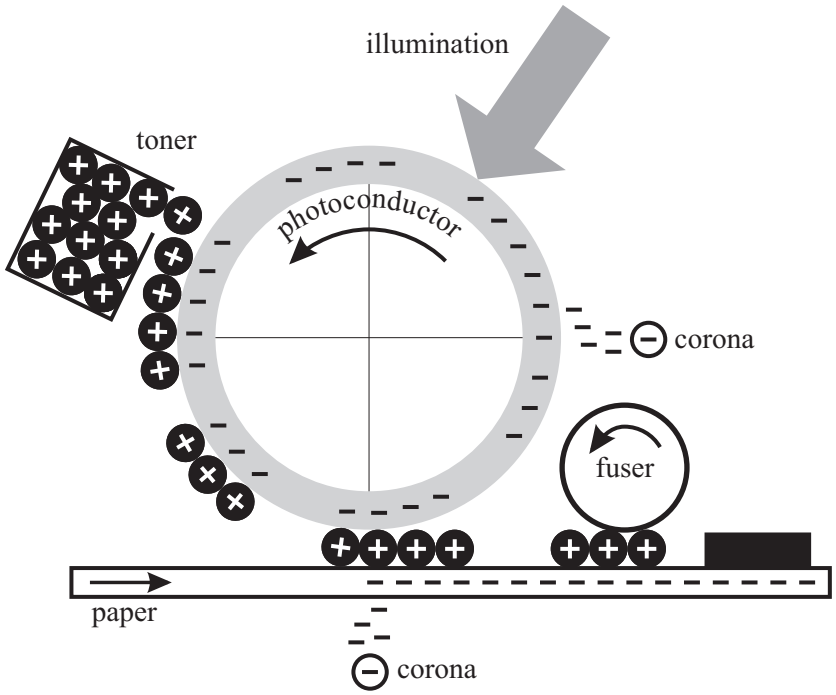

Figure 12.16. Electrophotography.

deformable support that can be electrically biased [van Kessel *et al*., 1998]. Depending on the bit stored in the cell below it, the mirror is tilted to one side or the other. This can be used to deflect incident illumination into or out of the exit optical path. But unlike an LCD, very little energy is absorbed by the mirrors, the mirrors can fill the surface area of the chip, and they switch in microseconds rather than milliseconds. Because of the difficulty in controlling the magnitude of the bending force, the mirrors are driven between stops in either direction, with grayscale variation coming from modulation of the switching waveform. Such a structure is an example of a *Micro-Electro-Mechanical System* (*MEMS*), extending CMOS fabrication techniques to selectively etch supporting layers to yield free-standing mechanical structures that bridge between the mechanical and electronic worlds [Rodgers *et al*., 1997]. Beyond the sophistication of the extra lithographic steps required to build them, MEMS encounter a host of forces that are not issues in larger machines. For DMDs, one of the biggest problems was simply preventing the mirrors from sticking to the substrate because of weak inter-atomic forces and capillary adhesion from moisture [Hornbeck, 1998].

The paper you're holding is one of the most interesting alternatives to a mirror for deflecting light. Its constituent fibers are translucent; the white color comes from photons bouncing many times and then diffusing back out. This lets it convert incident light from almost any direction into uniform background illumination, with contrast coming from absorption in the ink. The same mechanism occurs with the emulsion of fat globules in milk or water droplets in a cloud; it is related to the phenomena of *weak localization* in

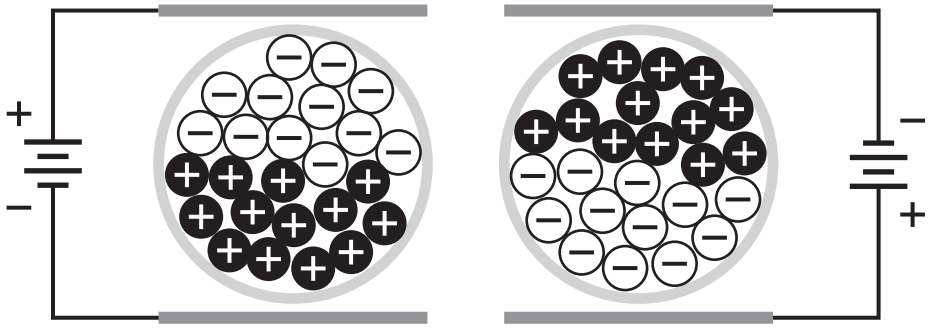

Figure 12.17. Microencapsulated electrophoretic electronic ink.

which coherent scatterers become trapped in random media [Yoo *et al*., 1989; Hastings *et al*., 1994].

The ubiquitous connection between a computer and a piece of paper is based on Chester Carlson's invention of *electrophotography* in 1938. The essential elements he used then appear today in laser printers and copiers. The printing cycles starts with a charge source, typically a *corona* discharge from a wire held at many kilovolts. This ionizes the air around it, attracting the positive ions and repelling the negative eletrons. These electrons accumulate on the surface of an insulating or semiconducting photoconductor. Materials used include selenium, amorphous silicon, and increasingly organic photoconductors because of their chemical and mechanical flexibility.

After it is charged, the photoconductor is illuminated with the desired image. This is done by focusing light reflected from a scanned document, or a linear array of light-emitting diodes, or by switching on and off a rastered laser beam [Starkweather, 1980]. The result is photo-induced carriers, with the positive ones being attracted to the surface electrons to neutralize their charge, leaving negative charge in the complement to the illumination. Then charged *toner* is applied, adhering to the photoconductor in these charged areas. These are pigmented thermoplastic particles, with sizes on the order of 10 μm. Their charge, opposite to that on the photoconductor, is developed through *triboelectricity*, the charge transfer that occurs between two objects rubbed together because of differences in their electron affinity. Finally, a piece of paper is brought in contact, itself charged to pull the toner off of the photoconductor. The final step is to use heat and pressure to fuse the toner to the paper, and to reset the photoconductor for the next pass. This can all happen very quickly, at speeds approaching 1000 pages per minute.

A piece of paper is an ideal display medium: it is thin, flexible, and non-volatile, and it offers high resolution, and high contrast. Its one liability is that what is printed cannot be changed. This is being remedied with the development of *electronic inks* that retain the contrast mechanism of printing, but also provide electronic addressability. This can be done with *microencapsulated electrophoresis*, shown in Figure 12.17 [Comiskey *et al*., 1998].

The synthesis starts with a solution of toner particles, with a color difference matched by a difference in their surface charge. These are then dispersed in a second liquid to form an emulsion of toner-containing droplets, on the order of 100 μm. Finally an *interfacial*

polymerization step mixes in a binary system that grows a clear shell at the droplet–solution interface. This is the process used to encapsulate ink into shells that burst under pressure in carbonless copy paper, but the introduction of surface charge on the particles permits them to be moved relative to each other because of their differential motion in an electric field.

The resulting contrast, resolution, and packaging are competitive with conventional printing because the mechanism is so similar, but now the image can be changed after it is put down. This could be done in a simple printer that needs just an electrode array to reuse a sheet of paper, or by integrating the drive electronics on the substrate as part of the printing process [Ridley *et al*., 1999]. While this technology is just beginning the technological scaling that more mature display technologies have been through, it promises to merge information display with the inks and paints used in everyday objects.

12.4 SELECTED REFERENCES

[Sze, 1998] Sze, S.M. (ed). (1998). *Modern Semiconductor Device Physics*. New York: Wiley-Interscience.

Advances in optoelectronic (and many other kinds of semiconductor) devices.

[O'Mara, 1993] O'Mara, William C. (1993). *Liquid Crystal Flat Panel Displays: Manufacturing Science & Technology*. New York: Van Nostrand Reinhold.

Everything you need to know to start your own LCD production facility.

[Pai & Springett, 1993] Pai, D.M., & Springett, B.E. (1993). Physics of Electrophotography. *Reviews of Modern Physics*, **65**, 163–211.

[Williams, 1993] Williams, Edgar M. (1993). *The Physics and Technology of Xerographic Processes*. Malabar, FL: Krieger.

The remarkable sophistication of, and insight into, the familiar copier.

12.5 PROBLEMS

(12.1) (*a*) How many watts of power are contained in the light from a 1000 lumen video projector?

(*b*) What spatial resolution is needed for the printing of a page in a book to match the eye's limit?

(12.2) (*a*) What is the peak wavelength for black-body radiation from a person? From the cosmic background radiation at 2.74 K?

(*b*) Approximately how hot is a material if it is "red-hot"?

(*c*) Estimate the total power thermally radiated by a person.

(12.3) (*a*) Find a thickness and an orientation for a birefringent material that rotates a linearly polarized wave by 90°. What is that thickness for calcite with visible light ($\lambda \sim 600$ nm)?

(*b*) Find a thickness and an orientation that converts linearly polarized light to circularly polarized light, and evaluate the thickness for calcite.

(12.4) Consider two linear polarizers oriented along the same direction, and a birefringent material placed between them. What is the transmitted intensity as a function of the orientation of the birefringent material relative to the axis of the polarizers?