Semiconductors & Transistors

AN OVERVIEW OF THEORY, FABRICATION, AND APPLICATION

by Scott McEvoy

Introduction

The purpose of this paper is to provide an overview of semiconductors and transistors. More specifically, the purpose of this paper is to discuss the theory of how semiconductors and transistors operate, how they are fabricated, and aspects to consider in the design of various technologies.

The first section of this paper provides an overview of the physics principles that govern the most vital components of semiconductors: electrons and holes. It then discusses how electrons and holes move throughout larger structures, and how those movements can be modeled.

The second section looks at what happens when two different types of semiconductor materials are joined. These compounds – called PN junctions – are the major underlying technology in almost every semiconductor device responsible for the operation of larger components.

The third section of this paper looks at MOSFETs, which are electrically controlled switches that use two PN junctions to regulate current.

The fourth section looks at a BJTs, which are similar to MOSFETs, but provide unique performance benefits due to their varied structure.

Lastly, the fifth section of this paper discusses semiconductor device fabrication techniques and how each of these play a role in the creation of integrated circuits and other electronic devices.

Throughout the entirety of this paper, various materials such as figures, equations, and descriptions have been taken from Chenming Hu's text on these subjects. Due to the profuse use of Hu's material, his text has not been cited at each instance. Rather, it can be assumed that almost everything in this book has been adopted from Hu. Hu's text can be found here: https://people.eecs.berkeley.edu/~hu/Book-Chapters-and-Lecture-Slides-download.html.

Electrons and Holes in Semiconductors

Necessary Background

In order to understand how semiconductors truly work, one must first have a basic understanding of quantum mechanics. But since quantum mechanics isn't very basic, the task is rather formidable. Like Carl Sagan said, "If you wish to make an apple pie from scratch, you must first invent the universe." Fortunately for us, the universe of semiconductors is limited to the atom. Unfortunately, however, to get a clear picture of the atom, we first need to understand some rather unintuitive quantum mechanical rules. With this in mind, this section begins with a naïve model of the atom, and slowly adds the complexities of quantum mechanics along the way, ultimately giving us a slightly less-naïve model of the atom. Once we understand how atoms and electrons behave, we can then begin to understand how semiconductor devices work.

The Structure of the Atom

A Naïve Model

Every atom is made up of three key particles: protons, neutrons, and electrons.[1] Protons have a positive electric charge, electrons have a negative electric charge, and neutrons have no electric charge.[2] Under ordinary conditions, the protons and neutrons of an atom are held together by the nuclear force, creating a positively charged nucleus which sits at or near the center of the atom. The electrons move around the nucleus in various orbitals creating an "electron cloud". While this naïve model provides a good understanding for the basics of atomic structure and electron motion, the actual story is much more complex.

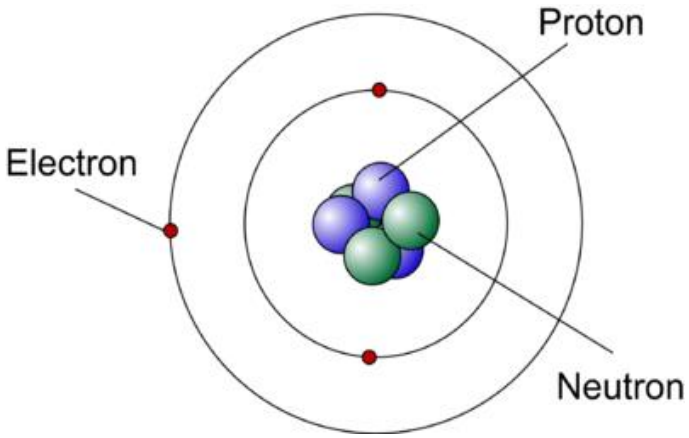

Figure 1: A naïve model of the atom

[1] Except for Hydrogen, which is only made up of one proton and one electron.

[2] $m_{proton} = 1.6726\times10^{-27}$ kg; $m_{neutron} = 1.6929\times10^{-27}$ kg; $m_{electron} = 9.11\times10^{-31}$ kg. The m_{proton} and the $m_{neutron}$ are approximately 1,836 and 1,839 times that of $m_{electron}$, respectively.

Coulomb's Law & The Uncertainty Principle

Electrons in an atom are governed by several forces and principles. First, electrons are attracted to the nucleus due to the electromagnetic force as described by Coulomb's Law, which states that particles of opposite charge are attracted to one another at a force inversely proportional to their distance.[3] However, if this were the only force governing electron motion, the electrons would fully collapse into the nucleus. The reason this does not occur is due to a quantum mechanical relationship called uncertainty, which states that the more precisely a particle's position (x) is known, the less precisely its momentum (p) can be known, and vice versa. This relation is written as

$$\Delta x \Delta p \geq \frac{1}{2}\hbar$$

Equation 1: The Uncertainty Relation

where $\hbar$ is the reduced Planck constant, Δx is the uncertainty of the particle's position, and Δp is the uncertainty of the particle's momentum. Based on this relationship, if we claim to know that a given electron is at the nucleus – i.e. Δx is reduced to the drastically smaller radius of the nucleus – then in order to keep the uncertainty relationship balanced, Δp must increase radically, resulting in a momentum that would eject the electron from the atom. Since this is not a common result in stable atoms, we can conclude that there is a minimum space that an electron can be restricted to.[4] When we combine this restriction with the Coulombic force between electrons and protons, we find that each electron should exist within an electrostatic potential well surrounding the nucleus.

The Photoelectric Effect

For an electron to move outside of this electrostatic potential well, it needs to gain an amount of energy greater than that of the attractive Coulombic force pulling it towards the nucleus. One way to gain such energy is by absorbing a photon (i.e. a packet of electromagnetic radiation, such as light or heat). However, as Einstein discovered in his photoelectric effect experiments, the energy (E) of a photon can only be transferred to an electron in discrete packets such that

$$E = hf$$

Equation 2: The Planck-Einstein Relation

where h is Planck's constant and f is the frequency of the photon. Thus, for an electron to move outside of its electrostatic potential well by means of photon absorption, the photon must have a sufficient frequency.

---

[3] Coulombs Law states that $F = k_e \frac{q_1 q_2}{r^2}$, where k_e is Coulomb's constant ($k_e = 8.99\times10^9 \frac{Nm^2}{C^2}$), q_1 and q_2 are the signed magnitudes of the charges, and the scalar r is the distance between the charges.

[4] This is not always the case in unstable atoms. For example, an atom with a super abundance of protons can provide such a strong positive charge that an electron is drawn in and "captured" by the nucleus, changing the proton and electron into a neutron and emitting an electron neutrino – a process known as "electron capture".

Wave-like Behavior & The Pauli Exclusion Principle

One unexpected consequence of this relationship came from De Broglie, who combined Einstein's relativistic energy-momentum relationship with the wavelength of light relationship to show that the wavelength of light was inversely proportional to its momentum.

$$p = \frac{E}{c}$$

Equation 3: The Einstein Relativistic Energy-Momentum Relationship

$$\lambda = \frac{c}{f}$$

Equation 4: The Wavelength of Light Relationship

$$\lambda = \frac{h}{p}$$

Equation 5: The De Broglie Wavelength Relationship

This deceptively simple result shows that all matter – including particles such as electrons – actually has both particle-like and wave-like behavior. The equation that describes this behavior is the particle's solution to the three-dimensional Schrödinger wave equation.[5]

For an electron, the solution to the three-dimensional Schrödinger wave equation is of the form

$$e^{\pm \boldsymbol{k} \bullet \boldsymbol{r}}$$

Equation 6: 3-D Schrödinger Wave Equation Solution Form for an Electron

where the wave vector, $\boldsymbol{k}$, is equal to $2\pi/\lambda_{electron}$. This solution represents an electron wave for each energy level E (i.e. a quantum state). It also describes the four quantum numbers associated with a given $\boldsymbol{k}$. These four numbers – n, the principle quantum number, ℓ, the angular momentum quantum number, m_ℓ, the magnetic quantum number, and m_s, the spin quantum number – define the quantum state of an electron. According to the Pauli exclusion principle, two or more identical electrons cannot occupy the same quantum state within a quantum system simultaneously. In other words, no two electrons within an electron system, such as an atom, molecule, or crystal, can have wave functions that perfectly overlap.

A Less Naïve Model

When all of these quantum mechanical properties are taken into consideration, we get a much clearer vision of the structure of an atom. At the center is a positively charged nucleus surrounded by orbiting electrons. The electrons orbit in electron clouds, where each forms a type

[5]The three-dimensional Schrödinger wave equation is $-\frac{\hbar}{2m_0}\nabla^2\psi + V(\boldsymbol{r})\psi = E\psi$, where m_0 is the free electron mass, $V(\boldsymbol{r})$ is the potential energy field that a crystal presents to the electron in the three-dimensional space, and E is the energy of the electron.

of three-dimensional standing wave – a wave form that does not move relative to the nucleus. This behavior is defined by an atomic orbital, a mathematical function that characterizes the probability that an electron appears to be at a particular location when its position is measured. Only a discrete – or "quantized" – set of these orbitals exist around the nucleus.

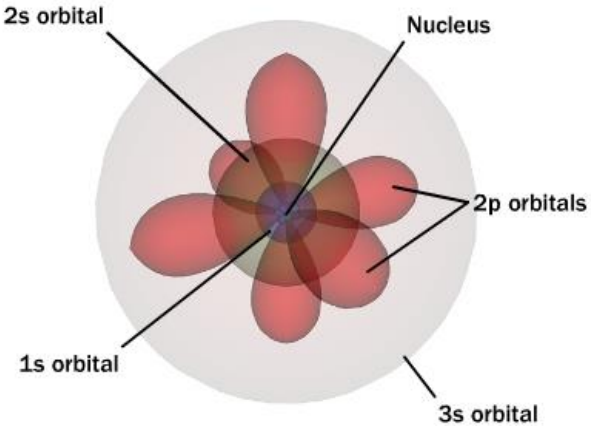

Figure 2: A less naïve model of the atom

Each atomic orbital corresponds to a particular energy level of the electron. The electron can change its state to a higher energy level by absorbing a photon with sufficient energy to boost it into the new quantum state. Or conversely, an electron in a higher energy state can drop to a lower energy state and radiate the excess energy as a photon – a process called spontaneous emission. If an incoming photon strikes an electron with enough energy, it can eject the electron out of the atomic structure, creating a positively charged ion.[6] Since electrons closer to the nucleus have both a stronger attraction towards the nucleus and a lower energy, the electrons in the highest orbitals – the valence band electrons – are more likely to be the ones ejected from structure. These are the same electrons which are most likely to be transferred from one atom to another, or shared between atoms via bonding, creating molecules, compounds, and crystals.

Semiconductors

Energy Band Model

Semiconductors are crystalline or amorphous solids with distinct electrical characteristics. When many atoms are brought into close proximity, as in a crystal, the discrete energy levels of the atoms are replaced with bands of energy states separated by gaps between the bands. The band containing valence electrons is called the "valence band". These are the electrons that bond to other atoms. Beyond the valence band is the "conduction band". Electrons in this band move most easily throughout a structure during electrical conduction and thus provide the electrical current across a semiconductor. Since it takes energy for an electron to move further away from an atom, an electron in the valence band requires an additional amount of energy to reach the

[6] Atoms with an equal number of protons and electrons are electrically neutral. If an atom has more or fewer electrons than its atomic number, then it becomes respectively negatively or positively charged as a whole. Such charged particles are called ions.

conduction band. This amount (E_g), called the "gap band" energy, is the difference between the conduction band energy (E_c) and the valence band energy (E_v).

$$E_g = E_c - E_v.$$

Equation 7: The Gap Band Energy

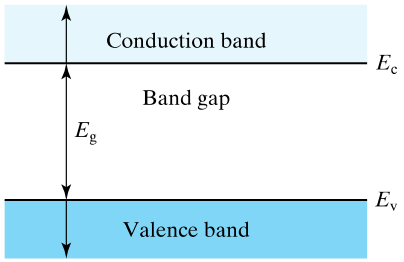

Figure 3: The energy band diagram of a semiconductor

Electrons & Holes

One distinguishing property of semiconductors is that they can be manipulated to either conduct the current of charge carriers or resist such current. In an intrinsic semiconductor, the two charge carriers – electrons and holes – are negatively and positively charged particles, respectively, and exist in pairs. This is because holes, which are simply the void of an electron in the valence band, are created when electrons move from the valence band to the conduction band. This usually occurs when electromagnetic radiation, such as light or heat, is absorbed by an electron. Naturally then, the equation that gives us the probability that an electron occupies the conduction band – or any given energy state for that matter – is a function of energy and temperature. This equation is called the Fermi Function:

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$$

Equation 8: The Fermi Function

where E_F is the Fermi Level, k is the Boltzmann constant, T is the temperature, and E is the inquired energy state.

But not all semiconductors are intrinsic. Rather, the number and ratio of charge carriers in an intrinsic semiconductor can be manipulated by introducing dopants to the material. Dopants either add electrons – producing an N-type semiconductor – or add holes – producing a P-type semiconductor. In either case, the abundant charge carrier is called the majority carrier while the less abundant charge carrier is called the minority carrier.

Electron & Hole Motion

In a Si crystal – the most common intrinsic semiconductor used in modern electronics – each Si atom uses each of its four valence electrons to form covalent bonds with its four neighbors, providing a relatively stable crystallography. Consequently, there are few mobile electrons and holes in Si at low temperatures. As the temperature increases, more electron-hole pairs are

formed, which in turn increases the total dopant concentration. These two factors – temperature and total dopant concentration – determine how frequently the holes and electrons collide with phonons and dopant ions, thus causing them to lose their momentum. These characteristics are called the hole mobility and the electron mobility.

In the presence of an electric field, charge carriers can "drift" in-line or against the direction of the field, gaining a "drift velocity". The drift velocity is a function of charge carrier mobility and the electric field. When the quantity of the charge carriers and their associated charges are taken into account, we can further determine how many of those charges are moving at the drift velocity; a value called "drift current density". When we consider particle diffusion from high-concentration locations toward low concentration locations – a common thermodynamic process proportional to the gradient of the carrier concentration – we can determine diffusion current density. Since both drift and diffusion are perturbations to the same thermal motion, and both are slowed down by the same collisions, we can relate the electron and hole diffusion constants to their respective mobilities. This relation is called the Einstein relationship.

$$D_n = \frac{kT}{q}\mu_n \; and \; D_p = \frac{kT}{q}\mu_p$$

Equation 9: Einstein Mobility-Diffusion Relationship

Furthermore, since these motions are so related, we can calculate the total current density (J) by adding the drift current density to the diffusion current density, such that

$$J_n = \mathrm{q}D_n\frac{dn}{dx} + qn\mu_n\xi \; and \; J_p = -\mathrm{q}D_p\frac{dp}{dx} + qp\mu_p\xi$$

Equation 10: Total Current Density

These equations are particularly helpful when dealing with semiconductors having both N-type and P-type regions, which we will discuss in the next section.

PN Junctions

Basic PN Junction Structure and Behavior

PN junctions occur where P-type semiconductors meet N-type semiconductors. When a P-type region comes into contact with an N-type region, holes from the P-type region will diffuse into the N-type region. Similarly, electrons from the N-type region will diffuse into the P-type region. When these holes and electrons come into contact with one another they recombine, resulting in the emission of energy, usually in the form of light or heat, and the elimination of both particles.

The Depletion Region

The recombination of electrons and holes creates an area depleted of charge carriers, called "the depletion region". The depletion region is a result of averaging the Fermi Levels between the two materials into an equilibrium Fermi Level. While the Fermi Level of the N-type region is closer to the conduction band, the Fermi Level of the P-type region is closer to the valence band. Since there is only one Fermi Level at equilibrium, the Fermi Levels between these two are averaged to only one value (E_F). This results in an E_F close to neither E_V nor E_C. Therefore, both electron and hole concentrations are quite small. This is where the depletion region gets its name; from the layer's depletion of electrons and holes.

The behavior of the depletion region can be described by dividing the PN junction into three regions – the neutral regions at $x > x_P$ and $x > -x_N$, and the depletion layer or depletion region, where $p = n = 0$.

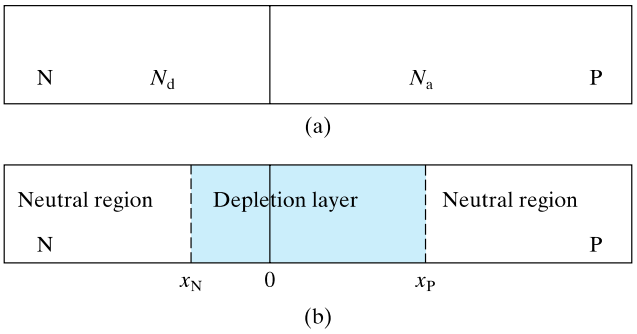

Figure 4: (a) a step PN junction, and (b) a depletion region approximation

On the P-side of the depletion layer ($0 \leq x \leq x_P$), the electric field is described by the following equation:

$$\xi(x) = \frac{qN_a}{\varepsilon_s}(x_P - x)$$

Equation 11: P-side Depletion Layer Electric Field

Similarly, the field on the N-side of the depletion layer is described by this equation:

$$\xi(x) = -\frac{qN_x}{\varepsilon_s}(x - x_N)$$

Equation 12: N-side Depletion Layer Electric Field

When these two equations are combined, we find that $|x_P|$ and $|x_N|$, the widths of the depletion layers on the two sides of the junction, are inversely proportional to the dopant concentration:

$$N_a|x_P| = N_d|x_N|$$

Equation 13: Depletion Layer Width to Dopant Concentration Relation

As this equation shows, the more heavily doped side holds a smaller portion of the depletion layer. A junction that is highly asymmetrical with regards to this equation is called a "one-sided junction". These can be described as either an N^+P junction or a P^+N junction, where N^+ and P^+ denote the heavily doped sides. In either case, the depletion layer penetrates primarily into the lighter doped side.

No matter how the depletion region is distributed, what is left behind are the ionized impurities from which the charge carriers came. Since the impurities come from the two sources of doped material, two regions appear in the PN junction: a region of positively charged ionized impurities and a region of negatively charged ionized impurities. This special distribution of charges creates an electric field with an electric potential that acts as a barrier, preventing the further displacement of electrons and holes. This voltage differential is called the built-in potential (ϕ_{bi}).

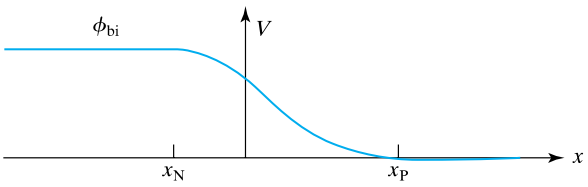

Figure 5: The electric potential across the depletion layer of a PN junction

A built-in potential is present at the interface of any two dissimilar materials, and is described by the following equation:

$$\phi_{bi} = \frac{kT}{q}\ln\frac{N_d N_a}{n_i^2}$$

Equation 14: Built-in Potential

From this equation, it is clear that the built-in potential is determined by N_a and N_d. The larger N_a or N_d is, the larger ϕ_{bi} is. Typically, ϕ_{bi} is about 0.9 V for a silicon PN junction.

If one is interested in determining the electric potential distribution, Poisson's equation is useful for when the charge density is known:

$$\frac{d^2V}{dx^2} = -\frac{d\xi}{dx} = -\frac{\rho}{\varepsilon_s}$$

Equation 15: Poisson's Equation

where ε_s is the semiconductor permittivity (for Si, that is twelve times the permittivity of free space), ρ is the charge density, and ξ is the electric field.

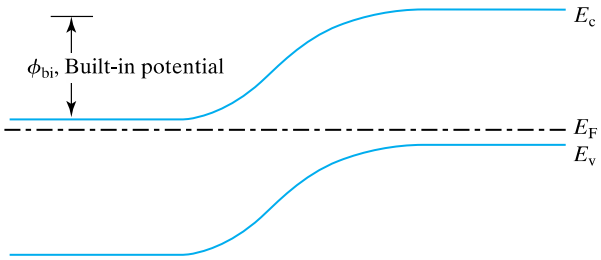

Figure 6: Energy band diagram of a PN junction

Another useful parameter is the total width of the depletion layer (W_{dep}). If we want to find this value, we simply take the sum of the widths of the depletion layers of the two sides of the junction:

$$x_N + x_P = W_{dep}$$

Equation 16: Total Width of the Depletion Layer

Since the depletion layer width is determined by the lighter doping concentration, the heavily doped side is often hardly depleted at all. When combined with Poisson's equation, W_{dep} can be determined by the following equation:

$$W_{dep} = \sqrt{\frac{2\varepsilon_s \phi_{bi}}{qN}}$$

Equation 17: Total Width of the Depletion Layer

Current Across the PN Junction

Once the depletion region reaches equilibrium – when the drift current equals the diffusion current – the potential barrier acts as an obstacle for the diffusion current in the device. However, it is possible to reduce the height of this potential barrier by the application of an external voltage.

By applying an external voltage, the height of the potential barrier is modified. If a positive voltage drop is applied between the P and N regions, the barrier height is reduced. A reduced barrier no longer prevents electrons and holes from diffusing across the structure, creating an electric current across the junction due to the diffusion mechanism. Under these conditions, the PN junction is said to operate under a "forward bias". A forward bias of V reduces the barrier height from ϕ_{bi} to $\phi_{bi} - V$. In this situation, electrons can now diffuse from the N-side into the P-side. This is called minority-carrier injection. Similarly, holes are injected from the P side into the N side. Moreover, according to the Shockley Boundary Condition, a PN junction with a forward bias, V, raises the minority carrier densities at the edges of the depletion layer by the factor $e^{\frac{qV}{kT}}$.

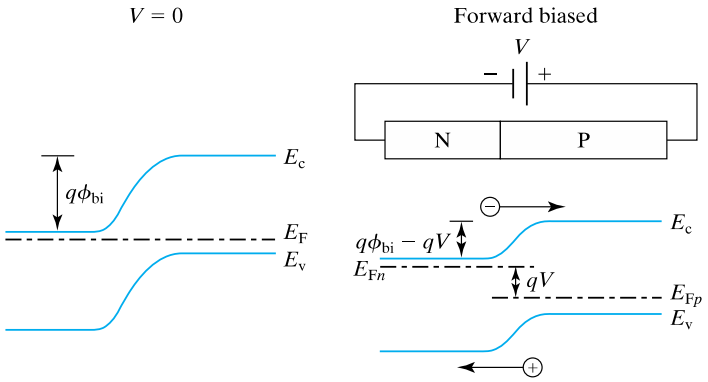

Figure 7: A forward bias reducing the junction barrier to $\phi_{bi} - V$

When a positive voltage is applied to the N region relative to the P region, the PN junction is said to have a "reverse-bias". Here, the barrier height increases, preventing the electron and hole diffusion and resulting in an electric current that is negligible. Under reverse-bias, the depletion region gets wider in order to dissipate the larger voltage drop across it. This bias-behavior in the PN junction is described as a rectifying current-voltage (I-V or IV) and is illustrated in Figure 5. As a device, it is called a "rectifier" or a "diode".

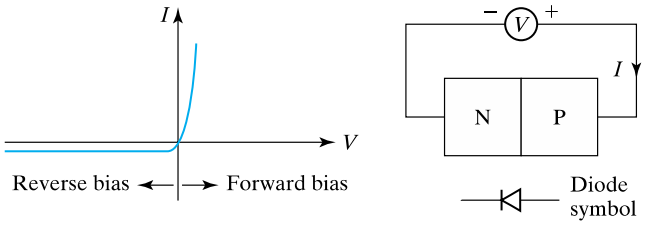

Figure 8: The rectifying characteristics of a PN junction

The total current across a PN diode can be written as

$$I = I_0(e^{\frac{qV}{kT}} - 1)$$

Equation 18: Total Current Across a PN Diode

$$I_0 = Aqn_i^2(\frac{D_p}{L_pN_d} + \frac{D_n}{L_nN_a})$$

Equation 19: Reverse Saturation Current

where A is the diode area, L_n and L_p are the diffusion lengths, and I_0 is the reverse saturation current. I_0 is known as the reverse saturation current because the diode current saturates at $-I_0$ at large reverse bias.

One factor that contributes to the forward current and the reverse current is net carrier recombination or generation in the depletion region. This contribution is called the "SCR current" for space-charge region current, which is just another name for the depletion region. The SCR current is described by the following equation:

$$Net\ recombination(generation) rate\ per\ unit\ volume = \frac{n_i}{\tau_{dep}}(e^{\frac{qV}{2kT}} - 1)$$

Equation 20: Depletion Region Recombination/Generation

where τ_{dep} is the recombination (or generation) lifetime in the depletion layer. To determine whether there is generation or recombination, look to the sign of the rate. When the rate is negative, there is net generation. When this occurs, the generated carriers are swept by the field into the N and P regions as an additional current component.

Under reverse bias, another source of current should also be taken into consideration: the reverse leakage current. Although described above as negligible, the reverse leakage current can become very important when the number of transistors increases, as in modern technology. This current is a function of the width of the depletion layer as well as the several other key design parameters.

$$I_{leakage} = I_0 + A\frac{qn_iW_{dep}}{\tau_{dep}}$$

Equation 21: Leakage Current

In DRAM and imager technology, junction leakage current is a very important issue that generates substantial noise.

To model factors like leakage current, the depletion layer and the neutral N and P regions may be modeled as a parallel plate capacitor with a capacitance:

$$C_{dep} = A\frac{\varepsilon_s}{W_{dep}}$$

Equation 22: PN Junction Capacitance Model

where C_{dep} is the depletion-layer capacitance and A is the area. Capacitance is an unwelcome capacitive load to devices and circuits. C_{dep} can be lowered by reducing the junction area and increasing W_{dep} by reducing the doping concentration(s) and/or applying a reverse bias.

PN Junction Breakdown

As noted above, a reverse-biased PN junction conducts a negligibly small current. This is still generally true. However, when a critical reverse bias is reached, a PN junction can begin to conduct a much higher current. This condition is called junction breakdown.

One form of junction breakdown is called “tunneling breakdown”. Tunneling breakdown occurs when a heavily doped junction is reversed biased and only a small distance separates the large number of electrons in the P-side valence band and the empty states in the N-side conduction band. In this situation, electrons can quantum mechanically tunnel through the junction. This is

known as tunneling breakdown, and it is the dominant breakdown mechanism when N is very high and V_B is quite low.

"Avalanche breakdown" is the mechanism of diode breakdown at a higher V_B. With increasing electric field, electrons traversing the depletion layer gain higher and higher kinetic energy. As they gain this energy, some of the electrons will strike the impurities ions that compose the depletion layer. Some of them will even have enough energy to raise an electron from the valence band into the conduction band. This phenomenon is call impact ionization. The electrons and holes created by impact ionization are themselves also accelerated by the electric field. Consequently, they and the original carrier can create even more carriers by impact ionization. This "avalanche" of electrons can ultimately generate a substantial current across the junction.

Metal-Semiconductor Junction

Overview

We just discussed the behaviors of semiconductors of different types forming junctions. But semiconductors can form junctions with a variety of materials. In integrated circuits, one of the most common junctions occurs between a metal and a semiconductor. Generally, there are two kinds of metal-semiconductor junctions: (1) Junctions between metal and lightly doped semiconductors, called Schottky diodes, and (2) Junctions between metal and heavily doped semiconductors, called low-resistance ohmic contacts (which are basically electrical shorts).

Schottky Diodes

The energy barrier at a metal-semiconductor interface is characterized by the Schottky barrier height, ϕ_B, which is the single most important parameter of a metal-semiconductor contact. Actually, there are two energy barriers within ϕ_B: ϕ_{Bn}, the barrier against electron flow between the metal and the N-type semiconductor, and ϕ_{Bp}, the barrier against hole flow between the metal and the P-type semiconductor. These two energy barriers are the primary sources of the band gap in a metal-semiconductor junction:

$$\phi_{Bn} + \phi_{Bp} \approx E_g$$

Equation 23: Metal-Semiconductor Band Gap Relation

In many metals, there is a clear trend that ϕ_{Bn} increases with an increasing metal work function. Since the work function is the minimum energy needed to remove an electron from a solid to a point in the vacuum immediately outside the solid surface, an equation exists that relates these values. This equation is known as the Schottky Mott Rule, which states that:

$$\phi_{Bn} = \psi_M - \chi_{Si}$$

Equation 24: Schottky-Mott Rule

where ψ_{M} is the metal work function and χ_{Si} is the silicon electron affinity. Although the Schottky Mott Rule is a naïve interpretation that doesn't completely predict Schottky barrier behavior, it does provide a good approximation for many materials.

One unusual consequence of creating a metal-semiconductor barrier is that the heights of the Schottky Barriers in metal semiconductor contacts often show little dependence on the value of the semiconductor or metal work functions. This is known as Fermi-Level pinning. Fermi-Level pinning is a result of placing a semiconductor up against a metal, which creates electron states within the bandgap, the nature of which (and their occupation by electrons) tends to pin the center of the band gap to the Fermi Level. Hence, "Fermi-Level pinning".

Although metal-Si contacts are functionally useful, silicide-Si contacts are much more prevalent in IC technologies. Silicide-Si contacts are created when metals react with silicon at a moderate temperature to form metal-like silicides. These silicide-Si interfaces are more stable than the metal-Si interfaces and free of native silicon dioxide. Nonetheless, the term metal-silicon contact is commonly understood to include silicide-silicon contacts.

The current of electrons flowing from Si over the energy barrier into metal is denoted as $J_{S\rightarrow M}$. This current can be predicted quite accurately by thermionic emission theory, which states that the thermally induced flow of charge carriers from a surface or over a potential-energy barrier occurs because the thermal energy given to the carrier overcomes the work function of the material. Mathematically, this means that $J_{S\rightarrow M}$ is only a function of $\phi_{\mathrm{B}} - V$. Thus, the metal work function and the bias voltage determine how many electron possess sufficient energy to surpass the peak of the energy barrier and enter the metal in a metal-semiconductor contact..

Although Schottky and PN diodes follow the same *IV* expression,

$$I = I_0(e^{\frac{qV}{kT}} - 1)$$

Equation 18: Total Current Across a PN Diode

I_0 of a silicon Schottky diode can be 10^3-10^8 times larger than a typical PN junction diode, depending on ϕ_{B} (i.e. the metal employed). Since a smaller ϕ_{B} leads to a larger I_0, a larger I_0 means that a smaller forward bias, *V*, is required to produce a given diode current. This property makes the Schottky diode the preferred rectifier in low-voltage and high-current applications.

For such application, a Schottky contact with a relatively small ϕ_{B} is used to obtain a large I_0 and a small forward voltage drop. However, ϕ_{B} cannot be too small, or else the large I_0 will increase the power loss when the diode is reverse biased and can cause excessive heat generation. The resultant rise in temperature will further raise I_0 and can lead to a destructive spiraling process called "thermal runaway". Thus, it is best to avoid very small metal work functions for such applications.

The second difference between a Schottky diode and a PN junction diode is that the basic Schottky diode operation involves only the majority carriers. Negligible injection of minority carriers also means negligible storage of excess minority carriers. Therefore, Schottky diodes can

operate at higher frequencies than PN junction diodes, making them particularly useful in circuits using oscillators and high-speed clocks.

Ohmic Contacts

Ohmic contacts are low-resistance junctions that occur where a metal meets a heavily-doped semiconductor. Since semiconductor devices are connected to each other in an IC through metal, the semiconductor-to-metal contacts should have sufficiently low resistance so they do not over degrade the device performance. Thus, low-resistance ohmic contacts play a vital role in ICs.

An important feature of all good ohmic contacts is that the semiconductor is very heavily doped. The depletion layer of the heavily doped Si is only tens of Å thin because of the high dopant concentration. When the potential barrier is very thin, the electrons can pass through the barrier by quantum tunneling, since the tunneling probability increases as a material becomes thinner. As in a Schottky barrier, if a small voltage is applied across the contact, the balance between $J_{S\rightarrow M}$ and $J_{M\rightarrow S}$ is broken and the barrier for $J_{M\rightarrow S}$ is reduced from ϕ_{Bn} to $(\phi_{Bn} - V)$. However, due to IC design desires, the voltage across an ideal ohmic contact is zero. Thus, IC design specifications should take great care with voltages across ohmic contacts.

Metal-Oxide Semiconductor Field Effect Transistors (MOSFETs)

Basic MOSFET Structure and Behavior

A MOSFET – or metal-oxide semiconductor field effect transistor[7] – is a semiconductor device that comprises four nodes: (1) a “source”, (2) a “drain”, (3) a “body”, and (4) a “gate”. For an N-channel MOSFET, the body is coupled to a P-type region of a semiconductor substrate, which contains within it two N-type regions, located a length *L* apart from one another. The source is coupled to one of the N-type regions and the drain is coupled to the other. The gate is coupled to the P-type region along the length *L*, with a dielectric layer disposed in between. This configuration is illustrated in Figure 9.

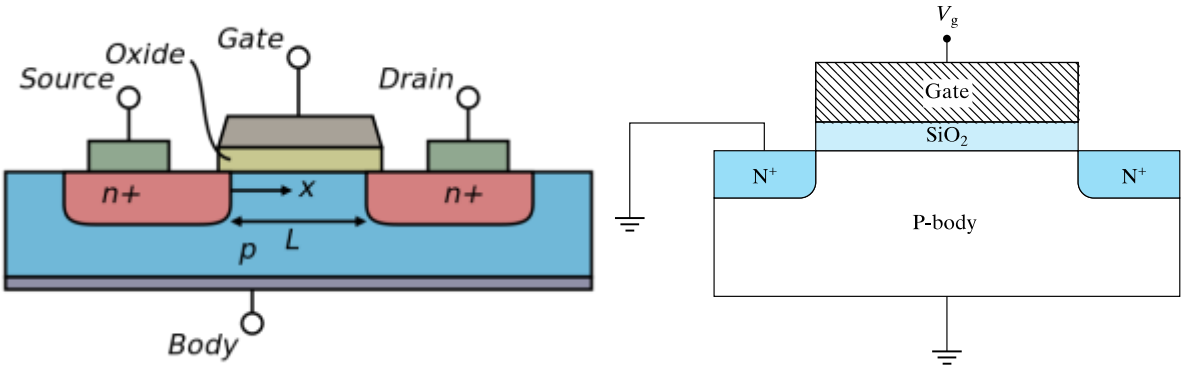

Figure 9: Two schematics of an N-MOSFET

When a voltage is high at the gate and low at the body, an electric field is generated that penetrates the dielectric material and pushes the holes in the p-type region away from the gate. This creates a negatively charged "inversion layer" or "channel" at the semiconductor-dielectric interface. When a voltage is high at the drain and low at the source (referred to as V_{ds}), electrons from the source can flow across the channel, creating an electrical current (I_{ds}). Varying the voltage between the gate and body modulates the conductivity of this layer and thereby controls the current flow between drain and source. This is known as “enhancement mode”.

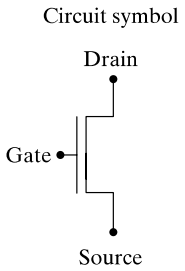

Figure 10: Circuit symbol for a MOSFET

[7] The name “field-effect transistor” or “FET” refers to the fact that the gate turns the transistor (inversion layer) on and off with an electric “field” through the oxide.

When no voltage is applied to the gate, the holes of the P-type body remain at the semiconductor-insulator interface in such abundance that the source and drain PN junctions have a reverse-bias. This reduces the conduction channel width so drastically that only a very small current is able to pass through the transistor. Since the voltage applied to the gate (V_g) controls whether a current flows or not, it is said that V_g acts as a "switch". When V_g is high, the switch is "on", and when V_g is low, the switch is "off". Thus, when the MOSFET is on, the current is called the "on-state current", I_{on}. Alternatively, when the MOSFET is off, the current is called the "off-state leakage current", I_{off}. This behavior can be seen in Figure 11, which demonstrates the IV characteristics of a basic MOSFET.

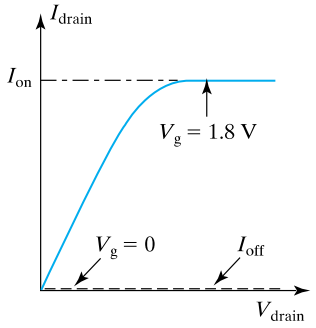

Figure 11: IV characteristics of a basic MOSFET

MOSFET C-V Characteristics

Surface Accumulation

What Figure 11 does not show is a condition known as "surface accumulation", which occurs when V_g is negative (i.e. when a negative voltage is applied to the gate).

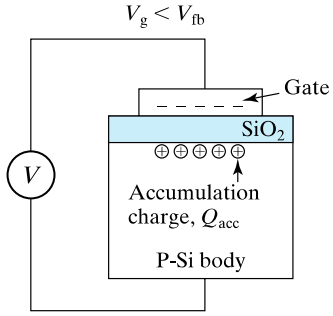

Figure 12: A MOS body biased into surface accumulation

Under surface accumulation, the negative V_g attracts a large number of holes that accumulate at or near the surface of the dielectric-body interface, forming an "accumulation layer". These "accumulation-layer holes", which have an "accumulation charge", Q_{acc}, create a surface voltage, ϕ_s, and a voltage across the oxide, V_{ox}, which is non-zero. This results in an electric field in the direction of the gate voltage which is even stronger than the gate-body potential difference alone.

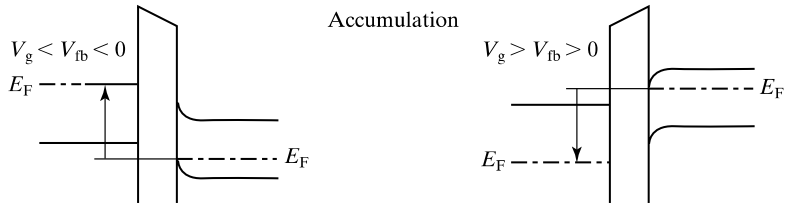

Figure 13: Energy band diagrams showing surface accumulation for an N-type Device (N^+ gate over P-substrate) on left and a P-type device (P^+ gate over N-substrate) on right

Flat-Band

Increasing V_g positively over time can result in a condition where, although a negative voltage is still applied to the gate, there is no surface electric field in the substrate. This condition is referred to as "flat band", and occurs where the energy band (E_C and E_V) of the substrate is in line, or "flat", with the gate's E_C and E_V across the Si-SiO_2 interface. Since both sides have equal energy bands, the electric field in the dielectric is also zero. The "flat-band voltage", is the gate voltage, V_{fb}, is the difference between the Fermi levels at the two terminals

$$V_{fb} = \psi_g - \psi_s$$

Equation 25: Flat-Band Voltage

where ψ_g and ψ_s are the gate work function and the semiconductor work function.

Since V_g is supposed to equal V_{fb} at flat-band, if $V_g \neq V_{fb}$, the difference must be picked up by ϕ_s, V_{ox}, and the poly-gate depletion layer, ϕ_{poly}. Thus, the gate voltage at flat-band can be described by the following equation:

$$V_g = V_{fb} + \phi_s + V_{ox} + \phi_{poly} = V_{fb} + \phi_s - \frac{Q_{sub}}{C_{ox}} + \phi_{poly}$$

Equation 26: Gate Voltage at Flat-Band

where Q_{sub} is all of the accumulation, inversion, and depletion-layer charge, and C_{ox} is the capacitance that may be present at the SiO_2-Si interface.

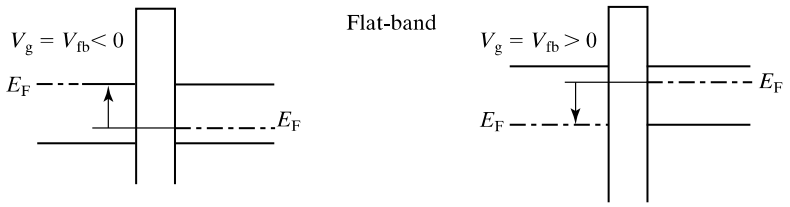

Figure 14: Energy band diagrams showing Flat-Band for an N-type Device (N^+ gate over P-substrate) on left and a P-type device (P^+ gate over N-substrate) on right

Surface Depletion

As V_g becomes more positive than V_{fb}, a third condition called "surface depletion" can begin to occur. In this situation, V_g begins to push the holes in the P-type layer away from the dielectric-body interface, creating a depletion region at the surface, where electron and hole densities are both small. In this scenario, the width of the depletion region can be determined by examining the gate voltage, such that:

$$V_g = V_{fb} + \phi_s + V_{ox} = V_{fb} + \frac{qN_aW_{dep}^2}{2\varepsilon_s} + \frac{qN_aW_{dep}}{C_{ox}}$$

Equation 27: Gate Voltage at Surface Depletion

Since this equation can be solved to yield W_{dep} as a function of V_g, V_{ox} and ϕ_s can be determined as well.

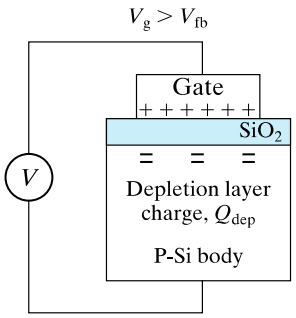

Figure 15: A MOS body biased into surface depletion

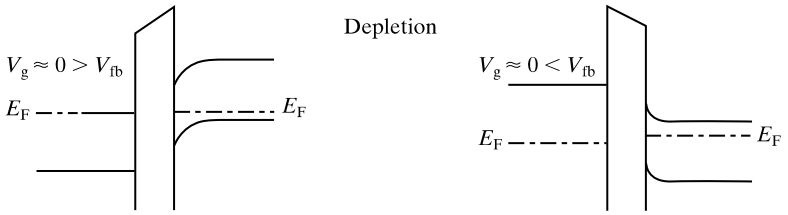

Figure 16: Energy band diagrams showing surface depletion for an N-type Device (N^+ gate over P-substrate) on left and a P-type device (P^+ gate over N-substrate) on right

Threshold Voltage

At some point, V_g can become so positive that the surface is no longer in depletion but at the threshold of inversion. The term inversion means that the surface is inverted from P-type to N-type, or electron rich. When this occurs, the surface electron concentration, n_s, becomes equal to the bulk doping concentration, N_a. In the IV curve in Figure 11, inversion occurs when V_g reaches the critical value of 1.8 V. This critical value is called the "threshold voltage", V_t. At V_t, the channel between the source and the drain flips its type. This turns the MOSFET to the on-state, and allows for a current of I_{on} to flow. The threshold voltage can be determined by the following three equations:

$$\phi_{st} = \pm 2\phi_{\mathrm{B}}$$
Equation 28: Surface Voltage at Threshold

$$\phi_{\mathrm{B}} = \frac{kT}{q} \ln \frac{N_{sub}}{n_i}$$
Equation 29: Band Gap Voltage

$$V_t = V_{fb} + \phi_{st} \pm \frac{\sqrt{qN_{sub}2\varepsilon_s|\phi_{st}|}}{C_{ox}}$$
Equation 30: Threshold Voltage

In these three equations, the positive signs are for a P-substrate and the negative signs are for an N-substrate.

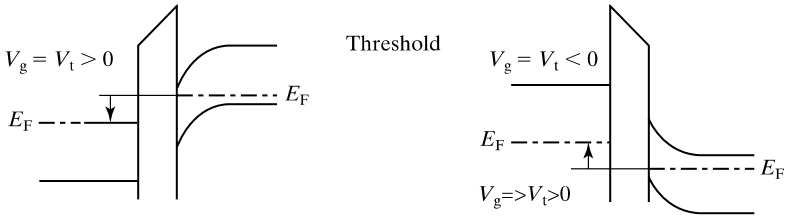

Figure 17: Energy band diagrams showing threshold voltage for an N-type Device (N^+ gate over P-substrate) on left and a P-type device (P^+ gate over N-substrate) on right

One particular disruption to the threshold voltage comes in the form of something called the "body effect". Since V_t increases with increasing body-to-source reverse bias, V_{sb}, this "body effect" is deleterious to circuit speed. To adjust for the body effect, the T_{oxe} can be reduced or W_{dmax} can be increased. Such technological developments can be found in various FinFET models.

$$V_{\mathrm{t}}(V_{\mathrm{sb}}) = V_{\mathrm{t0}} + \alpha V_{\mathrm{sb}}$$
Equation 31: Threshold Voltage Body Effect Equation

$$\alpha = \frac{3T_{\mathrm{oxe}}}{W_{\mathrm{dmax}}}$$
Equation 32: Alpha Body Effect Equation

$$V_{\mathrm{t}} = V_{\mathrm{t0}} + \frac{\sqrt{qN_{\mathrm{a}}2\varepsilon_{\mathrm{s}}}}{C_{\mathrm{oxe}}}\left(\sqrt{2\phi_{\mathrm{B}} + V_{\mathrm{sb}}} - \sqrt{2\phi_{\mathrm{B}}}\right)$$
Equation 33: Threshold Voltage Compensating for Body Effect Equation

In equations 31 and 33, V_{t0} is the threshold voltage in the absence of body bias.

Additional Considerations

Several other factors also affect charge accumulation and depletion in a MOS body. Failure to account for these factors can result in applying the incorrect V_t, which create a variety of logic failures throughout an IC.

One such factor to consider is the build-up of charge in the oxide layer. Presence of electric charge in the gate dielectric, such as "fixed oxide charge" (silicon ions at the SiO_2 interface), "mobile oxide charge" (mostly from sodium ion contaminants from the water in the fabrication process), and "interface traps", can modify both V_{fb} and V_t.

It should also be noted that V_t decreases with L, a fact known as "V_t roll-off", caused by drain induced barrier lowering (DIBL), described as such:

$$V_t = V_{t-long} - (V_{ds} + 0.4V) * e^{-\frac{L}{l_d}}$$

Equation 34: Roll-off Equation

$$I_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j}$$

Equation 35: DIBL Equation

Since V_t is a sensitive function of L, even the small (a few nm) manufacturing variations in L can cause problematic variations in V_t, I_{off}, and I_{on}. To allow L reduction, the above V_t equation states that l_d must be reduced, i.e., T_{oxe}, W_{dep}, and/or X_j must be reduced.

Lastly, one should consider the "poly-Si gate depletion effect". Because a depletion layer is present in the gate, one may say that a poly-silicon-gate capacitor is added in series with the oxide capacitor. The poly-Si gate depletion effect occurs when this gate capacitance drops as the capacitor is biased deeper into the inversion region due to increasing poly-depletion. The poly-depletion effect is undesirable because a reduced C means reduced Q_{inv} and reduced transistor current. The traditional solution is to dope the poly-Si heavily. Unfortunately, very heavy doping may cause "dopant penetration" from the gate through the oxide into the substrate. Poly-gate depletion is eliminated in advanced MOSFET technology by substitution of the poly-gate with a metal gate.

CMOS Technology

While the majority of the devices so far have dealt with NFETs, PFETs are not all that different. To create a P-channel MOSFET (or PFET), one can simply swap the P-type regions and N-type regions, as seen in Figure 18 below. If these regions are swapped however, to operate in the same state – "on" in both a PFET and an NFET – the voltage between the gate and the body (V_g and V_{dd}) must be swapped as well.

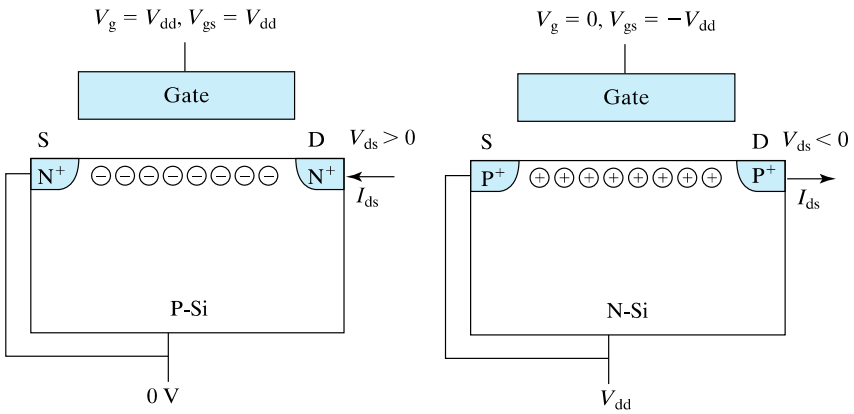

Figure 18: An NFET in the on state (left), and a PFET in the on state (right). When V_g is equal to V_{dd}, an inversion layer is present and the NFET is turned on. With its body and source connected to V_{dd}, a PFET responds to V_g in exactly the opposite manner.

Figure 18 also shows where the two devices got their names. The reason N-channel MOSFETs (or N-MOSFETs, or simply NFETs) are called N-channel is because the conduction channel (i.e. the inversion layer) is electron rich; the two N-type regions create a channel of flowing electrons that traverse the P-type substrate when the gate provides a positive voltage. P-channel MOSFETs (or P-MOSFETs or simply PFETs) on the other hand, allow for two P-type regions to create a current of holes that traverse the N-type substrate when the gate provides a negative voltage. This symmetric behavior between PFETs and NFETs can be used to make a very useful class of devices called CMOS circuits.

The complementary nature of NFETs and PFETs makes it possible to design low-power circuits called CMOS or complementary MOS circuits. One example of such a low power circuit is the CMOS inverter, shown in Figure 13.

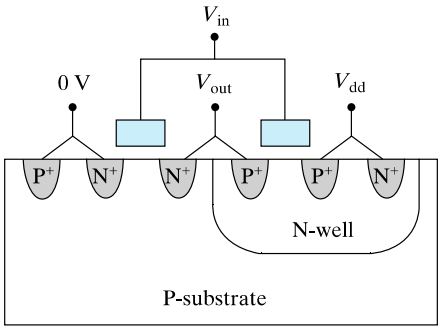

Figure 19: A CMOS inverter where the PFET "pulls up" and the NFET "pulls down"

In a CMOS inverter circuit, when $V_g = V_{dd}$ (i.e. when V_{in} is high), the NFET is on, the PFET is off, and the output node is pulled down to the ground ($V_{out} = 0$). When $V_g = 0$, the NFET is off, the PFET is on, and the output node is pulled up to V_{dd}. Thus, a CMOS inverter is said to consist of a PFET "pull-up device" and an NFET "pull-down device". In either case, one of the two transistors is off and there is no current that flows from V_{dd} through the two transistors directly to the ground. This allows CMOS circuits to consume much less power than other types of circuits.

The details of these characteristics can be seen in the voltage transfer curve (VTC) for the CMOS inverter.

VTCs show V_{in}/V_{out} pairs across a range of input voltages for a given circuit. Since the input voltage can be anywhere below the threshold voltage and still produce a perfect output voltage equal to driving voltage, the input voltage need not be a perfect "high" value to produce a "1". This process is similar for PFETs. Thus, perfect "0" and "1" outputs can be produced by somewhat corrupted inputs. This regenerative property can be thought of as a cushion for input voltage variation. If the voltage stays within the cushion (i.e. the margin), then the output will still result in the desired outcome. If the voltage goes outside of this margin, the output will produce a "bit error" – the production of a value opposite of what was intended. This information can be useful in determining digital circuit characteristics, such as noise margin. In the case of the CMOS inverter, the PFET pull-up device and the NFET pull-down device create a highly nonlinear VTC. This nonlinearity gives the inverter its ability to refresh digital signals and provides the much-needed noise margin in a noisy digital circuit.

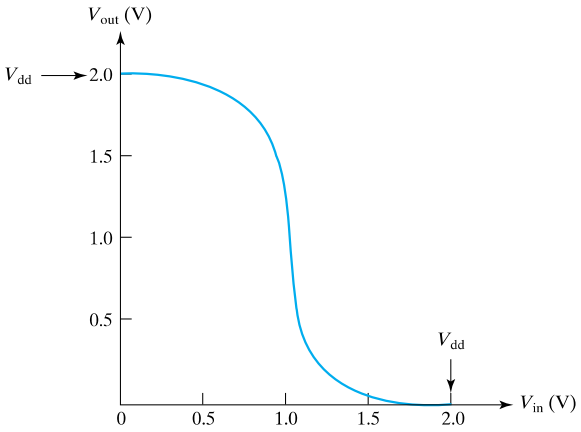

Figure 20: VTC of a CMOS inverter

MOSFET noise arises from the channel, gate, substrate thermal noise, and the flicker noise. While the thermal noise is a white noise, the flicker noise per bandwidth is proportional to $1/f$. The flicker ($1/f$) noise is reduced if the trap densities in the gate dielectric or the dielectric–semiconductor interface are reduced.

The inverter also provides information about propagation delay, where the delay time, τ_d, is proportional to the time it takes for the MOSFET to reach threshold voltage, such that

$$\tau_d = \frac{CV_{dd}}{4}\left(\frac{1}{I_{onN}} + \frac{1}{I_{onP}}\right)$$

Equation 36: MOSFET Propagation Delay

These values can help to determine fabrication process variations that result in threshold voltages that fall outside of the designed specifications. Consequently, this information can help reduce and control power consumption, which is described in the following equation:

$$P = kCV_{\mathrm{dd}}^2 f + V_{\mathrm{dd}} I_{\mathrm{off}}$$

Equation 37: MOSFET Power Consumption

where $k < 1$ accounts for the activity of the circuit. The first term is the "dynamic power" and the second, the "static power".

Naturally, it is highly desirable to have large I_{on}s without using a large power supply voltage, V_{dd}. It is also desirable to reduce the total load capacitance, C (including the junction capacitance of the driver devices, the gate capacitance of the driven devices, and the interconnect capacitance). Both capacitance and cost reductions provide strong motivations for reducing the size of the transistors and therefore the size of the chip. In addition, speed has benefited from the relentless push for smaller L, thinner T_{ox}, and lower V_{t} ; and power consumption has benefited greatly from the lowering of V_{dd}.

Very small MOSFETs are also prone to have excessive leakage current called I_{off}, which is a major power consumption concern in VLSI. The basic component of I_{off} is the "subthreshold current":

$$I_{\mathrm{off}}(\mathrm{nA}) = 100 * \frac{\mathrm{W}}{\mathrm{L}} * e^{-\frac{qV_{\mathrm{t}}}{\eta kT}} = 100 * \frac{W}{L} * 10^{-\frac{V_{\mathrm{t}}}{S}}$$

Equation 38: MOSFET Subthreshold Current

S is the "subthreshold swing". To keep I_{off} below a given level, there is a minimum acceptable V_{t}. Unfortunately, a larger V_{t} is deleterious to I_{on} and speed. Therefore, it is important to reduce S by reducing the ratio $T_{\mathrm{oxe}}/W_{\mathrm{dep}}$.

MOSFET Current Model & Alternative Designs

MOSFET Current & Carrier Mobility Model

When a small V_{ds} is applied to a MOSFET, the drain to source current, I_{ds}, is

$$I_{\mathrm{ds}} = WQ_{\mathrm{inv}} v = WQ_{\mathrm{inv}} \mu_{ns} \xi = WQ_{\mathrm{inv}} \mu_{ns} \frac{V_{\mathrm{ds}}}{L} = WC_{\mathrm{oxe}} \left(V_{\mathrm{gs}} - V_{\mathrm{t}}\right) \mu_{ns} \frac{V_{\mathrm{ds}}}{L}$$

Equation 39: MOSFET Drain to Source Current Model

where W is the "channel width", Q_{inv} is the inversion charge density, ξ is the channel electric field, and L is the channel length, μ_{ns} is the electron surface mobility, or the "effective mobility". As would be expected, surface mobility, μ_{ns}, is a function of the average of the electric fields at the bottom and the top of the inversion charge layer, ξ_{b} and ξ_{t}.

Electron and hole surface mobilities, μ_{ns} and μ_{ps}, are well-known functions of the average electric field in the inversion layer, which can be roughly expressed as $(V_{\mathrm{gs}} + V_{\mathrm{t}})/6T_{\mathrm{ox}}$. As this effective vertical field increases, the surface mobility decreases. At typical operating fields, surface mobilities are only fractions of the bulk mobilities.

GaAs MESFET

The most obvious way to improve speed is to use a semiconductor with higher carrier mobility than Silicon, such as Germanium (Ge) or strained Si. Such high carrier mobility allows the carriers to travel faster and the transistors to operate at higher speeds. GaAs and some other compound semiconductors have much higher electron mobilities than Si. Unfortunately, it is very difficult to produce high-quality MOS transistors in the materials. There are too many charge traps at the semiconductor/dielectric interface for MOSFET application. Fortunately, a Schottky junction can serve as the control gate of a GaAs FET in place of an MOS gate. This device is called a MESFET, for metal-semiconductor field-effect transistor. In selecting a metal material, a large Schottky barrier height is desirable for minimizing the input gate current (i.e., the Shottky diode current).

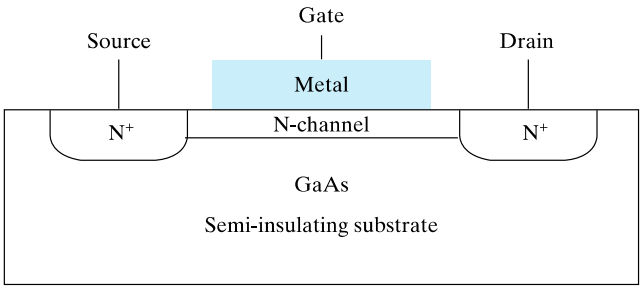

Figure 21: Schematic of a Schottky gate MESFET

HEMT

HEMTs, or high electron-mobility transistors, provide an epitaxial interface of two semiconductors that is smoother than the Si-SiO_2 interface. This device does not suffer from mobility degradation by surface scattering as the MOSFET does. Thus, the electron mobility is very high and the device speed is very fast. However, these devices are expensive to produce.

JFET

If the Schottky junction in the MESFET is replaced with a P^+N junction, the new structure is called a JFET or junction field-effect transistor. JFETs provide a low input current and capacitance device because its input is a reverse-biased diode. The major benefit of JFETs is that they can be fabricated with bipolar transistors and coexist in the same IC chip.

Bipolar Junction Transistors (BJTs)

Introduction to the BJT

A bipolar junction transistor (bipolar transistor or BJT) is a semiconductor devices that consists of three doped semiconductor regions: (1) the emitter region, (2) the base region, and (3) the collector region. In an NPN transistor, the emitter and the collector are N-type and the base is P-type. In a PNP transistor, these types are flipped. The base is physically located between the emitter region and the collector region, creating two PN junctions. Each semiconductor region is connected to a node of a matching name: emitter (E), base (B), and collector (C). The emitter node is coupled to a voltage source, V_{BE}. The collector node is coupled to a voltage source, V_{CB}. The base is then connected to the other terminal of each voltage source, V_{BE} and V_{CB}. A schematic of an NPN BJT is shown in Figure 22 below.

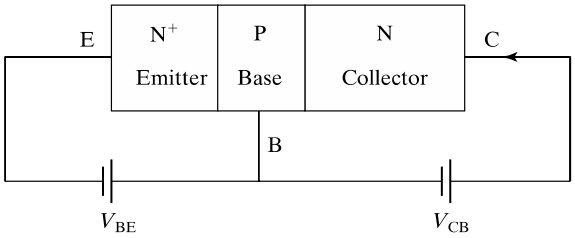

Figure 22: Schematic of an NPN BJT

The most common mode of operation for a BJT is called "forward-active" mode (or simply "active" mode). In active mode, the base-emitter PN junction is forward biased and the base-collector PN junction is reverse biased. This allows the negative terminal of V_{BE} to push the abundant electrons in the emitter past the PN junction's field – a process called "electron injection". Once these electrons reach the base, they are then attracted in two directions: across the base-collector PN junction towards V_{CB}^{+}, and back towards V_{BE}^{+}. The flow of electrons across the base-collector PN junction towards V_{CB}^{+} is called the "collector current", I_C. I_C is determined by the rate of electron injection from the emitter into the base, which is determined by V_{BE}. This rate of injection is proportional to $e^{qV_{BE}/kT}$. Thus, I_C is generally independent of V_{CB}, so long as the base-collector PN junction is reverse-biased (or lightly forward-biased).

To analogize BJTs to MOSFETs, the emitter is equivalent to the source, the collector is equivalent to the drain, and the base is the equivalent of the gate. This makes I_C equivalent to I_{ds}. This analogy also means that the emitter is often connected to ground. The distinguishing operational feature of BJTs, however, is their inherent ability to amplify current. Amplification, or "gain", is achieved by allowing V_{CB} to contribute to the collector current, I_C. MOSFETs lack this feature – there is no voltage source that directly contributes to I_{ds}. BJTs however, have this feature built directly in to their configuration, making them incredibly useful for high performance devices. Thus, BJTs are often used as amplifiers or switches, giving them wide applicability in electronic equipment, including computers, televisions, mobile phones, audio amplifiers, industrial control, and radio transmitters

Regions of Operation

As noted above, the most common mode of operation – or "region of operation" – for a BJT is active mode. However, three other modes exist: (1) saturation, (2) cut-off, and (3) reverse-active. These are determined by the way in which E, C, and B are connected to the two voltage sources, V_{BE} and V_{CB}. For an NPN, if E is connected to the negative terminal of V_{BE} (V_{BE}^-), B is connected to V_{BE}^+ and V_{CB}^-, and C is connected to V_{CB}^+, then the BJT is said to operate in "forward-active" mode. If E is connected to V_{BE}^-, B is connected to V_{BE}^+ and V_{CB}^+, and C is connected to V_{CB}^-, then the BJT is said to operate in "saturation" mode. If E is connected to V_{BE}^+, B is connected to V_{BE}^- and V_{CB}^-, and C is connected to V_{CB}^+, then the BJT is said to operate in "cut-off" mode. If E is connected to V_{BE}^+, B is connected to V_{BE}^- and V_{CB}^+, and C is connected to V_{CB}^-, then the BJT is said to operate in "reverse-active" mode.

Reverse-active mode is characterized by switching the roles of the emitter and collector regions as configured in active mode. Because most BJTs are designed to maximize current gain in active mode, the flow of electrons in reverse-active mode is several times smaller. This transistor mode is seldom used, usually being considered only for failsafe conditions and some types of bipolar logic. The reverse bias breakdown voltage to the base may be an order of magnitude lower in this region.

In saturation mode, both junctions are forward-biased, which facilitates a high flow of current from both the emitter and the collector to the base. If the load requiring current is placed on the node C, then saturation mode corresponds to a logical "on", or a closed switch.

Lastly, cut-off mode results in biasing conditions opposite of saturation, where both junctions are reverse-biased. In cut-off, there is very little current, which – dependent on the required load current – also corresponds to a logical "off", or an open switch.

For a PNP, the connections for each region of operation are all flipped with respect to the NPN configurations. However, since electron mobility is larger than hole mobility, NPN transistors – with their higher transconductances and speeds – are preferred over PNP transistors. For this reason, NPN BJTs are almost exclusively used since high performance is the BJT's competitive edge over MOSFETs. Thus, each of the modes listed in Figure 23 below is for the ubiquitous NPN BJT.

Mode (NPN)	Applied Voltages	B-E PN Junction Bias	B-C PN Junction Bias
Forward-active	E < B < C	Forward	Reverse
Saturation	E < B > C	Forward	Forward
Cut-off	E > B < C	Reverse	Reverse
Reverse-active	E > B > C	Reverse	Forward

Figure 23: Table of regions of operation for NPN and PNP BJTs

Current & Gain

In forward-active mode, the base–emitter junction is forward-biased while the base–collector junction is reverse-biased. V_{BE} determines the rate of electron injection from the emitter into the

base, and thus uniquely determines the collector current, I_C, regardless of the reverse bias, V_{CB}. This allows us to derive the equation for collector current as a function of V_{BE} as follows:

$$I_C = A_E \frac{qn_i^2}{G_B}(e^{qV_{BE}/kT} - 1)$$

Equation 40: BJT Collector Current

where A_E is the area of the emitter, and G_B is a parameter called the "base Gummel number". The Gummel number is a value that simplifies the complexities of the I_C model because it contains all of the subtleties of transistor design that affect I_C: changing base material through $n_{iB}(x)$, nonconstant minority carrier (i.e., electron) diffusion constant in the base (i.e., nonconstant D_B, nonuniform base dopant concentration through $p(x) = N_B(x)$, and even the high-level injection condition where $p > N_B$. The Gummel number is defined by Equation 41.

$$G_B \equiv \int_0^{W_B} \frac{n_i^2}{n_{iB}^2} \frac{p}{D_B} dx$$

Equation 41: BJT Base Gummel Number

As discussed above, one of the key features of a BJT is its inherent ability to amplify a signal by allowing V_{CB} to contribute a current, I_B, to the collector current, I_C. This amplification parameter, called "common-emitter current gain", β_F, is measured as the ratio of I_C to I_B:

$$\beta_F = \frac{I_C}{I_B} \approx \frac{G_E}{G_B}$$

Equation 42: BJT Common-Emitter Current Gain

where G_E is the emitter Gummel number. The common-emitter current gain, as shown above, is roughly equivalent to the ratio of the emitter Gummel number and the base Gummel number because the collector current is highly dependent upon how many electrons are injected into the base; a characteristic that is highly dependent upon the factors associated with the base Gummel number noted above. This is similar for the base current and the base Gummel number. Thus, the common-base current gain is similarly situated, as shown in Equation 43.

$$\alpha_F \equiv \frac{I_C}{I_E} = \frac{\beta_F}{1 + \beta_F}$$

Equation 43: BJT Common-Base Current Gain

To raise the common-emitter current gain, dopant concentration in the emitter is usually increased. Unfortunately, under such heavy doping, the band gap energy of the emitter can be substantially reduced, resulting in a narrow band gap. This effect is called the "heavy doping effect" or "band gap narrowing". Nonetheless, the common-emitter current gain can be increased without such effects by using a base material with a smaller band gap than the emitter material. Because the emitter-base junction is made of two different semiconductor materials, this device configuration is known as a "heterojunction bipolar transistor", or simply an "HBT".

Moreover, although many devices wish for a high gain, increasing V_{BE}, and thus I_B, can result in the undesirable, but unavoidable, side effect of a hole current flowing from the base into the emitter. Since they lack the energy to reach the conduction band of the emitter terminal contact (usual a metal with high Schottky barrier), they accumulate on the emitter's contact surface. This behavior is shown in Figure 24.

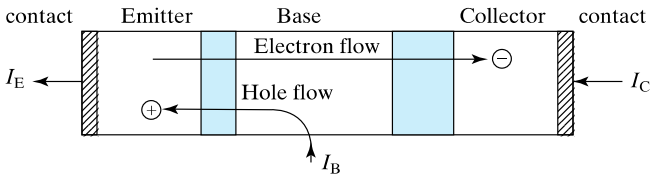

Figure 24: Schematic of electron and hole flow paths in BJT

Such accumulation can result in an increased capacitance, consequently lowering the collector current at high V_{BE} levels. This is called "β_F Fall-Off".

The Early Effect & The Kirk Effect

When V_{CE} increases, the base-collector depletion region widens and the neutral base width decreases, leading to an increase in I_C. Such base-width modulation by V_{CB} can result in a significant increase of output conductance in the active region. Since output conductance is defined as $\partial I_C / \partial V_{CE}$, an increasing output conductance can limit the voltage gain produced with a BJT. This rise of I_C due to base-width modulation is called the "Early effect". The Early effect can be suppressed with a lightly doped collector.

When the base-emitter PN junction is forward-biased, excess holes are stored in the emitter, the base, and even the depletion layers. The sum of these excess charges is called the "excess carrier charge", Q_F. Q_F is linearly proportional to I_C. The ratio of Q_F to I_C is called the "forward transit time", τ_F.

$$\tau_F = \frac{Q_F}{I_C}$$

Equation 44: Forward Transit Time

The forward transit time is often called the "storage time", as it sets a high-frequency limit to BJT operation.

If there were no excess carriers stored outside the base, the following equation would apply:

$$\tau_F = \tau_{FB} = \frac{W_B^2}{2D_B}$$

Equation 45: Forward Transit Time Without Excess Carrier Storage

where τ_{FB} is the base transit time. In general, however, τ_F is greater than τ_{FB} because excess carrier storage in the emitter and in the depletion layer are significant. In order to minimize τ_F,the emitter, depletion layers, and base width, W_B, should be reduced. Moreover, τ_{FB} should be

reduced by building a drift field into the base with a graded base doping. This significant increase of τ_{FB} at large I_C due to base widening is known as the "Kirk Effect".

BJT Circuit Modeling

A small-signal model is an AC equivalent circuit in which the nonlinear circuit elements are replaced by linear elements whose values are given by the first-order (i.e., linear) approximation of their characteristics curve near the bias point. Small-signal modeling is useful in analyzing and approximating electronic circuits containing nonlinear devices with linear equations. For a BJT, the simpler small-signal model consists of resistors, capacitors, and current sources. Some of the parameters the small-signal models looks at are:

$$g_m \equiv \frac{dI_{\mathrm{C}}}{dV_{\mathrm{BE}}} = \frac{I_{\mathrm{C}}}{kT/q}$$

Equation 46: Transconductance

$$C_\pi = \frac{dQ_{\mathrm{F}}}{dV_{\mathrm{BE}}} = \tau_{\mathrm{F}} g_m$$

Equation 47: Input Capacitance

$$r_\pi = \frac{dV_{\mathrm{BE}}}{dI_{\mathrm{B}}} = \frac{\beta_{\mathrm{F}}}{g_m}$$

Equation 48: Input Resistance

For computer simulation of circuits, however, the Gummel–Poon model is widely used for both DC and dynamic currents because it is far more comprehensive, including consequences such as the Early effect and the high-level injection effect.

Device Fabrication

Introduction to Device Fabrication

Typically, a semiconductor integrated circuit (IC) device fabrication process involves over one hundred steps. Generally, however, there are only seven core steps involved:

(1) Wafer creation
(2) Oxidization
(3) Lithography
(4) Etching
(5) Doping & Diffusion
(6) Thin-Film Deposition
(7) Interconnect

In the following sections, these steps are explored in greater depth to show just how and when different techniques are useful. But first, it is useful to review just how semiconductor crystals are structured.

A crystalline solid consists of atoms arranged in a repetitive structure. A "unit cell" is the smallest group of atoms of a substance that has the overall symmetry of a crystal of that substance, and from which the entire lattice can be built up by repetition in three dimensions. The length of a unit cell is call the "lattice constant". The lattice constant of Silicon, the most common semiconductor material, is 5.43 Å. The primitive cell is a minimum volume cell (a unit cell) corresponding to a single lattice point of a structure with discrete translational symmetry.

Miller indices indicate where the plane of the crystal intersects the xyz axis. Miller indices come in (abc) form, where a, b, and c are lattice constants. At 1/a, 1/b, and 1/c, the abc plane intersects the xyz axis. The form [abc] indicates the direction in the crystal normal to the (abc) place (i.e., the direction the electron travels with respect to the plane. Thus, the (100) plane intersects the x-axis at 1 lattice constant and the y-axis and z-axis at infinity. These indices will be used later in this section to discuss wafer orientation.

Wafer Creation

The first step in fabricating a semiconductor IC is the creation of a crystallized wafer. But to create a wafer, we first need a crystallized ingot from which to slice the wafer. One of the most popular bulk crystal growth techniques for ingot creation is called the Czochralski method.

Czochralski Method

Under the Czochralski method, high-purity, semiconductor-grade silicon is melted in a crucible at 1425°C, usually made of quartz. If desired, dopant impurity atoms such as boron or phosphorus can be added to the molten silicon in precise amounts to dope the silicon, thus changing it into p-type or n-type silicon. A precisely oriented rod-mounted seed crystal is then

dipped into the molten silicon. Once it reaches a desired depth, the rod is slowly pulled upwards and rotated simultaneously. By precisely controlling the temperature, rate of pulling, and speed of rotation, it is possible to extract a large, single-crystal, cylindrical ingot from the melt. This process is normally performed in an inert atmosphere, such as argon, within an inert chamber, such as quartz.

Once the ingot is collected, it can then be sliced. Silicon is usually cut along the (100) plane to obtain uniformity and good device performance. A flat or a notch is then cut along the (011) plane in order to precisely and consistently orient the wafer as desired during device fabrication.

Oxidation of Silicon

SiO_2 layers of precisely controlled thickness are produced during IC fabrication by reacting Si with either oxygen gas or water vapor at an elevated temperature. In either case the oxidizing species diffuses through the existing oxide and reacts at the Si–SiO_2 interface to form more SiO_2.

Growth of SiO_2 using oxygen and water vapor is referred to as "dry" and "wet" oxidation, respectively. Dry oxidation is used to form thin oxide films, while wet oxidation is preferred for thicker oxides due to the fact that water vapor diffuses through SiO_2 faster than oxygen.

This process occurs within an oxidation furnace system, as shown in Figure 25 below.

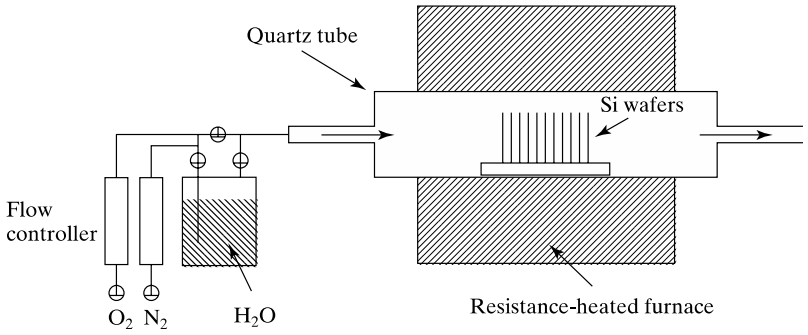

Figure 25: Schematic drawing of an oxidation system

Once the semiconductor is oxidized, it is then ready to have designed IC components put onto it.

Lithography

Lithography is the process of selecting which areas of the oxide layer are to be removed. Photolithography, also termed "optical lithography" or "UV lithography", uses light to transfer a pattern from a photomask to a light-sensitive chemical "photoresist" on the substrate.

The major steps in the lithography process involve applying a photoresist, aligning a photomask, and selectively exposing the photoresist to UV light. The photoresist itself is a UV light sensitive material that is uniformly applied to the top surface of the oxidized wafer. The photomask is a

quartz photoplate containing the patterns to be produced. Where the photomask is opaque, the UV light will be blocked, and where it is clear, the UV light will shine through. If the photoresist is a "positive resist", the areas where the light strikes will weakened and set for removal. If the photoresist is a "negative resist", the opposite will occur. This process is illustrated in Figure 26 below.

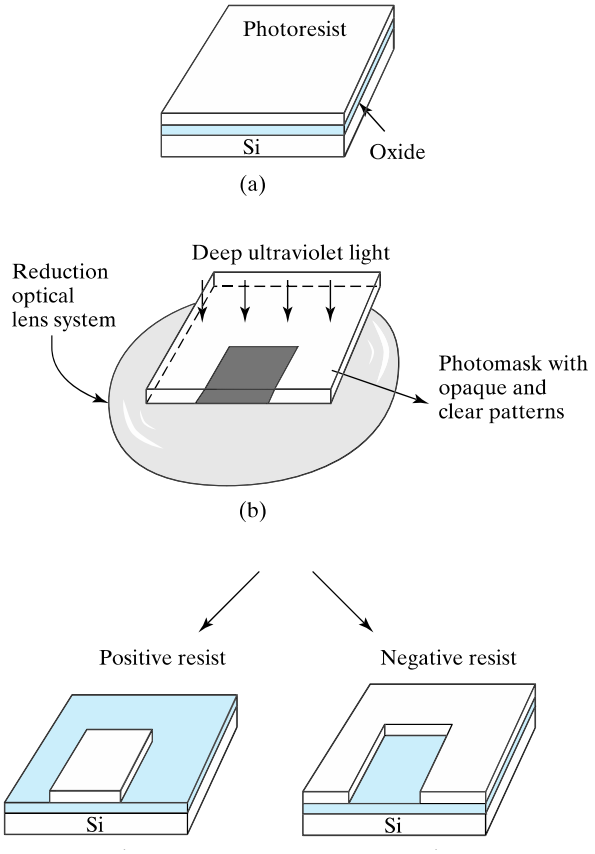

Figure 26: Major steps in the lithography process: (a) application of resist; (b) resist exposure through a mask and an optical reduction system; and (c) after development of exposed photoresist.

Due to optical diffraction limits, the resolution of the minimum feature size of the selected areas is limited to the lithography resolution,

$$Lithography\ Resolution = k\lambda$$

Equation 49: Lithography Resolution Limit

where λ is the wavelength of the light and k is the resolution limit, which is highly dependent upon the dielectric constant of the medium. As the lithography resolution limit relation show, the shorter the wavelength, the better the resolution. One way to reduce the wavelength is through a process called "wet lithography". Wet lithography sends the light through a material with a higher dielectric constant, such as water – hence "wet" – in order to reduce the wavelength.

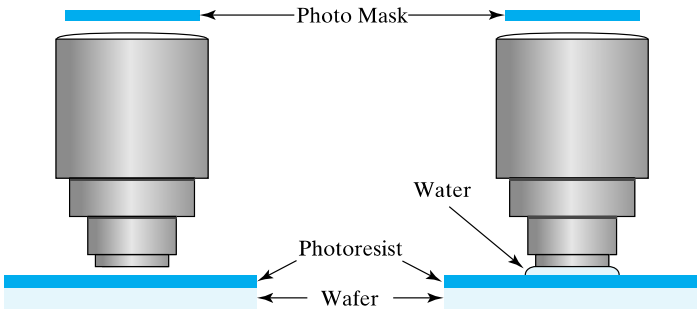

Figure 27: Schematics of dry lithography (left) and wet lithography (right)

Another way to reduce the wavelength is by using a "phase-shift photomask", which produces a 180° phase shift between the two clear regions on either side of a thin dark line. Since the two phases have electric fields of opposite sign, they partially cancel each other out, resulting in thinner lines being produced.

"Electron beam" lithography is another type of lithography process that helps to enhance resolution by using a focused stream of electrons to deliver energy to expose an electron resist. Although this process has slower exposure rates than optical lithography, its speed can be increased by using multiple electron beams.

Many other lithography processes exist to help enhance the resolution of the features. One reason for all of these processes is because lithography is the most difficult and expensive process among all the IC fabrication steps. So great care should be taken in each lithography step in order to reduce costs and increase efficiency.

Etching

Once the photoresist is exposed to UV light and the pattern is formed in the resist by lithography, the wafer is ready for etching. Etching is the process of removing patterns selected during lithography.

"Wet" etching often uses a chemical such as HF to remove selected features. Since this process is isotropic – i.e., without preference to direction – the etched features are generally larger than the dimensions of the resist patterns. Fortunately – although more expensively – this failure can be overcome through "dry" etching.

Dry etching is the process of removing a photoresist by exposing it to a bombardment of ions (usually a plasma of reactive gases such as fluorocarbons, oxygen, chlorine, boron trichloride; sometimes with addition of nitrogen, argon, helium and other gases) that dislodge portions of the material from the exposed surface. Since the ions move in a vertical direction (anisotropically), they tend to retain the features that lie underneath the oxide layer. This distinction is shown in Figure 28.

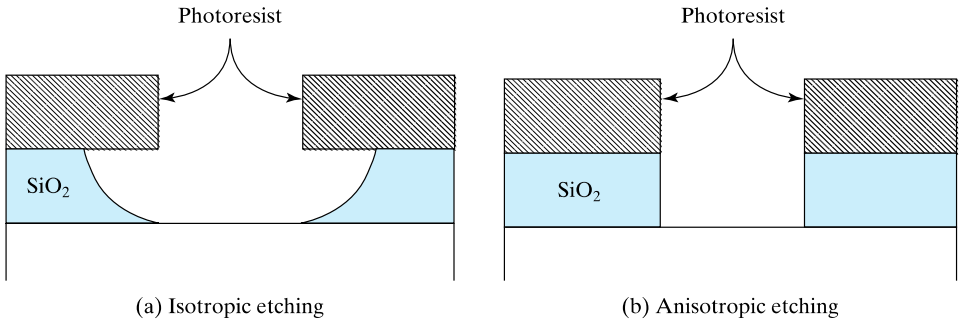

Figure 28: Comparison between wet etching results (left) and dry etching results (right)

Doping & Diffusion

Once the wafer has been patterned and etched, a common step is to then dope select regions of the wafer with various impurities in order to create N-type regions and P-type regions.

"Ion Implantation" is the most important doping method because of the precise control it provides. In ion implantation, ions of a dopant impurity are placed in an electric field and then accelerated into a wafer. These dopant ions, such as boron, phosphorous, or arsenic, are generally created from a gas source to retain their purity.

Once the dopants are introduced to the wafer, they are driven deeper into the semiconductor material through diffusion. Diffusion is commonly performed in an open tube system similar to the furnace used in the oxidation process. This application of heat – a process called "annealing" – results in the recrystallization of the doped semiconductor material, wherein the dopants are diffused almost evenly throughout. Faster annealing allows for shallower junctions. Thus, processes such as flash annealing or laser annealing are helpful in high performance devices.

Thin-Film Deposition

After dopants have been implanted throughout the semiconductor wafer, a common step is to deposit a thin layer of a material on top of the wafer, such as poly-Si, a common gate material. There are several ways to deposit these thin films.

"Sputtering" is a thin-film layer deposition technique that occurs in a low-pressure gas environment wherein a source material, called a sputtering target, and a wafer are juxtaposed. By applying a voltage between the target and the wafer, the gas is ionized and then accelerated towards the target. When the ions impact the target, atoms or molecules are ejected from the material, readily traveling back to the wafer, where they form the desired thin film. Since sputtering is anisotropic, it cannot deposit uniform films on the vertical portions of a wafer. This problem is solved by a process called "chemical vapor deposition".

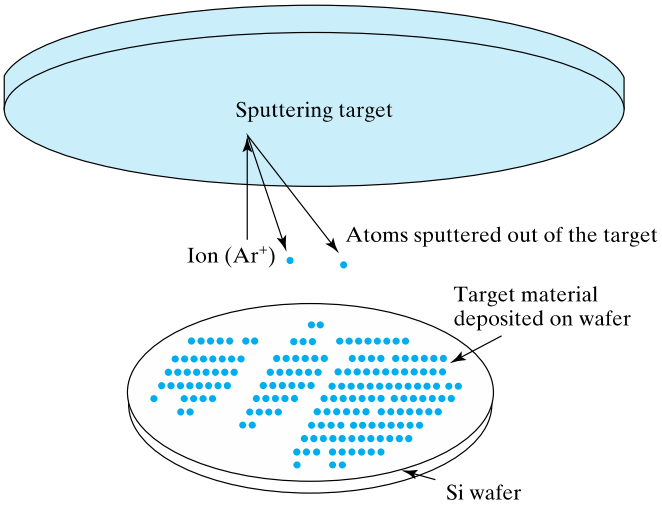

Figure 29: Sputtering process

Chemical vapor deposition, or "CVD", is a process that combines two different gases to create a compound that is then deposited on a substrate, producing a conformal film of uniform thickness. This process can be enhanced by using high temperatures to promote particle movement, thus increasing the uniformity of application and thickness. This process is shown in Figure 30.

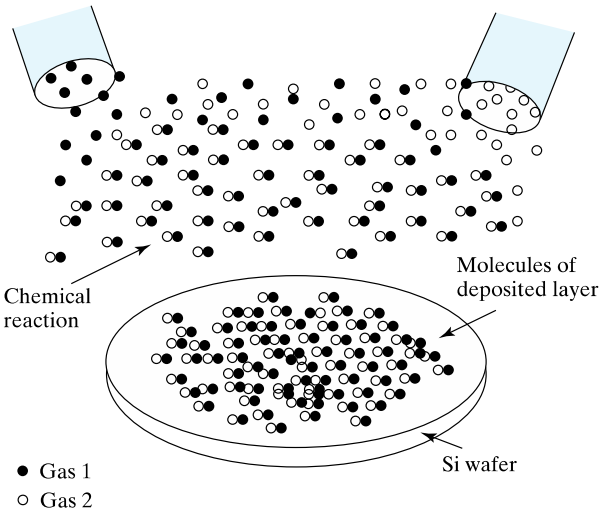

Figure 30: Chemical vapor deposition process

Epitaxy is another thin-film deposition technique that produces a crystalline overlayer on a crystalline substrate. During epitaxy, an arriving atom is moved across the surface of the substrate until it locks on to a seed crystal in the substrate. Once the atom attaches, it extends the lattice pattern of the substrate crystal. This process can be altered to a variation called "selective epitaxy deposition", wherein an etching gas is introduced to simultaneously etch away the material. In general, epitaxy is useful for fabricating lightly doped layers of crystal Si over a heavily doped substrate.

Interconnect – The Back-End Process

The fabrication techniques presented thus far can be interleaved many times over, creating complex processes that result in several hundred steps. Each of these processes are referred to as "front-end" processes, because they are involved in the creation of the semiconductor device itself. The "back-end" processes are the processes that involve connecting the semiconductor devices to one another (and external terminals) through metal lines.

Connecting semiconductor components in an IC – such as transistors – by metal lines is sometimes called "metallization". In older metallization processes using metals like aluminum, a layer of the metal was deposited on the substrate – typically through sputtering – and then selectively removed through lithography and dry etching processes. Today, however – due to electromigration problems with aluminum – the metallization process has been adapted to something called the "damascene process", which is better suited for metals without electromigration issues, such as copper.

The damascene process involves covering the wafer with a dielectric material, etching a trench in the dielectric, lining the trench with a film such as TiN, depositing the copper over the liner, and then planarization through chemical mechanical polishing (or "CMP"). The reason a liner is necessary is because copper diffuses rapidly when it touches a dielectric material. Once the damascene process is complete, the lines connecting the various terminals should be ready for operation.

One very important thing to note about metallization is that, due to the large number of metal layers and the number of processing steps involved, interconnection consumes a huge part of the IC fabrication budget. As with lithography processes, great care should be taken at each step to reduce costs and increase efficiency.