

# The Age of Silicon

Dr. André van Schaik

&

Dr. Richard Coggins

Computer Engineering Laboratory,  
School of Electrical and Information Engineering,  
The University of Sydney

andre@sedal.usyd.edu.au, richardc@sedal.usyd.edu.au

## Introduction

During this session participants will be given an overview of the history and future of silicon integrated circuits, the characteristics of analogue and digital signals, an introduction to silicon device functionality and physics, how analogue and digital integrated circuits can be designed for information processing and how these circuits can interact with their environment.

## History and Future of Silicon Integrated Circuits

### Overview

In 1965, Gordon Moore, at Fairchild Semiconductor, quantified the growth of semiconductors, saying that they had been doubling at regular intervals and would continue to do so. Dubbed 'Moore's Law', it quickly became a reliable predictor of future trends and is even used as a baseline strategy for the industry for the next 15 years. Extrapolating out to 2050 the numbers still look possible. Applications (smart everything) can certainly consume this production, on top of PC demand which consumes 60% of IC production. This techno-mantra has perhaps become a self-fulfilling prophecy.

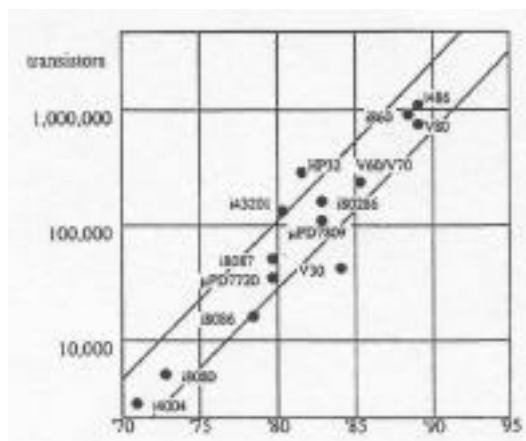
After 1947 and Shockleys, (Ohls) invention of the semiconductor transistor, the minaturisation, which was not possible for thermionic valves, became the hallmark of the semiconductor industry and the basis for Moore's Law. In 1950 William Shockley, John Bardeen, and Walter Brattain founded Texas Instruments and Fairchild and Moore (a member of Schockleys team) cofounded both Fairchild Semiconductor and Intel. During the 1950s Jack Kilby at Texas Instruments invents the Integrated Circuit (IC) while Fairchild invents the planar IC and this becomes nearly as significant as the semiconductor transistor itself and is the basis for the doubling law of Moore.

### Science to Production

Science produced solid-state electronics but more important for industry were several unprecedented production technologies. Two of the most important were the diffusion process and the oxide-masking process which departed from convention by reversing the usual science to technology development path. Diffusion allowed transformation of handcrafted ICs to mass production (the layering of conducting and insulating material was bypassed). The planar process, developed by Jean Hoerni, was the logical outgrowth of the diffusion/oxide masking technique. Planar or flat transistors are easier to manufacture than 3D 'mesa' transistors. Contact lithography and the planar technique co-developed for ever greater rates of production at higher yield. Robert Noyce at Fairchild realised that in addition the planar process allows integration of circuits on a single substrate. Patent attorneys asked engineers to consider the possibilities of the planar technique and they realised that metal could be deposited to interconnect circuits. Exponential advances in the planar process, lithography and photographic techniques set the industry on its exponential growth curve.

### Birth of Moore's law

On the 19<sup>th</sup> April 1965 Moore's (R&D Lab Director) article in Electronics magazine 'Cramming more components onto integrated circuits', predicted what would happen in the next 10 years in the industry. Based on a log-linear plot, a prediction of 65,000 components per IC by 1975 was achieved (see Figure 1). "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain constant for at least 10 years."



**Figure 1.** Chip transistor counts

Moore's prediction was based on three Fairchild data points:

- 1<sup>st</sup> planar Xtor 1959.
- 2<sup>nd</sup> is the 1964 IC with 32 components.
- 3<sup>rd</sup> late 1965 IC with 64 components.

In 1975 at the IEEE International Electron Devices Meeting Moore's prediction was on the mark, (a memory device with 65,000 devices was in production by Intel). His paper attributed this 2/3 growth to:

- Contact lithography replaced by photolithography which allowed higher yields.
- Finer rendering of images and lines
- 1/3 attributed to "circuit and device cleverness", allowing more efficient use of area (ended with CCD which used light beams to control doping instead of chemicals).

"There is no room left to squeeze anything out by being clever. Going forward from here we have to depend on the two size factors - bigger die sizes and finer dimensions". Moore then redrew his plot with a gentler slope doubling every 18 months. In 1995 Moore tested the relationship with industry output resulting in close correlation.

The Semiconductor industry marks evolution by Medium Scale Integration (MSI) in the 60s, Large Scale Integration (LSI) in the 70s, Very Large Scale Integration (VLSI) in the 80s and Ultra Large Scale Integration (ULSI) in the 90s. We are now seeing chips like the Intel Pentium III with ~8.2 million transistors with features sizes of 0.1um being clocked at speeds approaching 1GHz and memory chips (dynamic random access memory) with 4 billion binary digits (bits). In 2010 we may see terachips (with one trillion bits of memory operating at one trillion instructions per second with feature sizes of 10nm resolution (DNA coil width)). It is quite possible, that during the 21<sup>st</sup> century we will see computers with the same computing power as the human brain and beyond. We may also see a fusion of biology and electronics. However, it is not so clear as to whether we will understand how to program such computers to perform the feats that human brain achieves routinely.

### Perpetual Innovation Machine

Moore's law has become a universal law of the entire IC industry. Smaller feature sizes makes everything better. Speed goes up, power goes down, manufacturing yield goes up, reliability goes up. Electronics becomes cheaper and so is applied more widely. There is perhaps also a psychological component. "More than anything, once something like this gets established, it becomes more or less a self-fulfilling prophecy. The Semiconductor industry Association puts out a technological roadmap, which continues this generation [turnover] every three years. Everyone in the industry recognises that if you don't stay on essentially that curve they will fall behind. So it sort of drives itself."

### User Expectations Matter

Moore's Law is supplemented by the pull of software developments. Early software placed a premium on tight code, the full force of the law relaxed this constraint with programs proliferating to thousands, tens of thousands and now millions of lines of code. In 1995 Nathan Myhrvold at Microsoft examined the Basic Language, 1975: 4000 lines, 1995: 500 000 lines. Microsoft Word was 27000 lines and is now 2 million. Software contributes to the law as increased performance is consumed almost faster than IC improvements. Economists call this dynamically increasing returns. "We like what we get, we want more which spurs people to create more." As the marginal cost of additional processing power and memory goes to zero software has expanded to take on a larger influence in product development.

### Is the end in view?

Moore's law has been validated over time but has been constantly predicted to fail. A 1996 poll by Forbes - of 11 industry chiefs gave the law 14 years. A total of an amazing 45 years from its first pronouncement. At some point exponential growth may falter because of physical or economical limits. Moore's 2<sup>nd</sup> Law is "economics may constrain growth". Capital requirements for IC fabrication have grown from \$14M in 1966 to \$1500M in 1995 and in 1998 the first \$3billion dollar plant was built. By 2005 a fabrication plant will be \$10billion = half of Intel's net worth in 1995. Manufacturers will need to team up to afford these huge costs. Government organised consortia are starting to appear. Another economic threat to the law is lack of profit in making devices cheaper, Dan Hutcheson (VLSI research) predicts that the price per transistor will bottom out in 2005. The rate of growth may be limited by the size of the economy. The world economy grows about 3% annually but the size of the semiconductor industry is growing by 25%. It may grow to 50% of the world economy in 30 years. Dan Lynch (CyberCash) says "We'll be dead when Moore's Law is played out" because "Moore's law is about human ingenuity not physics"

## Enter the Equipment Maker

Early IC manufacturers also produced the equipment needed to manufacture them. Innovations in silicon semiconductor manufacturing lead to new processes or equipment. In-house developments spun off into independent suppliers.

Fairchild Semiconductor spawned 150 companies including Intel. Manufacturing equipment literally defines the limits of the technology. There are unusual relationships between capital equipment makers and chip makers given the free enterprise model. Collaboration occurs at all levels of interaction.

## Is Moore's Law unique?

An alternative view of the law can shed light on it. Moore joked that similar progress for air travel means a jet would cost \$500, travel around the world in 20mins and use 20 litres of fuel and be only the size of a shoe box. A better example, however, is the first 50 years of aircraft speed and performance. Personal Computer (PC) harddrives have gone from a megabyte to gigabyte in one decade and DNA based technologies have also seen similar advances. Railways in 1830 in the US started at 37km in length and doubled every decade for 60 years. By 1995 they could have been millions of kilometers in length, but is only 400,000km. It turned out to be uneconomical to connect small towns –roads are used instead. The analogy is flawed, however, as it deals with the implementation or diffusion of technology, whereas Moore's Law is about the pace of innovation. The law has become a benchmark of progress for the entire semiconductor industry. Actual underlying causes of the trend have not been proven. Carver Mead stated "Moore's Law is really about human activity, it is about vision, it is about what you are allowed to believe"

Year	Generation	Platform	User interface	Network
1951	Direct and batch use	Computer, vacuum tube, transistor, core, drum, magnetic tape	Card, paper tape, direct control evolving to batch op. system	None (originally stand-alone computers)
1965	Interactive timesharing via commands; minicomputers	Integrated circuit, disk, mini-computer; multiprogramming	Glass teletype and keypunch, control by command language	Telephone using modem, and proprietary wide-area networks
1981	Distributed PCs and workstations	Microprocessor PCs, workstations, floppy, small disk, distributed operating system	WIMP (windows, icons, mouse, pull-down menus)	Wide- and local-area networks
1994	World Wide Web access through PCs and workstations	Evolutionary PCs and workstations, servers everywhere, Web op. system	Browser	Optical-fiber backbone, World Wide Web, hypertext transfer protocol
1998	Web computers; network, telephone, TV computers	Client software from server using JAVA, ActiveX and so forth	Telephone, simple videophone, television access to the web	Subscriber digital lines for telephone or cable access for high-speed data
1998	SNAP: scalable network and platforms	PC uni- or multiprocessor commodity platform	Multimedia Web clients	System area network for clusters
2001	"Do what I say" speech controlled computers	Embedded in PCs, hand-held devices, phones, digital assistants	Speech	Infrared and radio LANs for network access.
2010	One info dial tone: phone, videophone, TV and data	Video-capable devices of all types	Video as primary data type	Single high-speed network access; home net
2020	Anticipatory by "observing" user behaviour	Room monitoring, gesture	Vision, gesture control	Home Net
2025	Body Net: vision, hearing, monitoring, control, communication, location	Artificial retina, cochlea, glasses for display, monitoring and recording of everything we see, hear, and say	Implanted sensors and actuators for virtually every part of a body	Body Network, gateway to local IR or radio nets everywhere, Humans are part of cyberspace
2048	Robots for home, office, and factory	General-purpose robot; appliances become robotic	Radar, sonar, vision, mobility, arms, and hands	IR and Radio LAN for home and local areas

**Table 1.** The table shows the evolution of computer classes in the context of the foregoing analysis, assuming that Moore's Law continues to hold.

## Beyond Moore's Law

Moore's Law predicts PCs by 2050 at  $10^{18}$  operations per second. Breakthrough technology currently being researched may allow  $10^{15}$  bytes of memory on a  $1\text{cm}^2$  chip. If this happens in the next decade the law could be accelerated by 30 years. On the other hand, the end of law by tomorrow would not stop improvements in systems software and networking.

One could assign three stages for computer evolution characterised by ever increasing performance at reduced price. The first is pre 70s when computers were rare. The second is post 70s. \$1M mainframes became \$100K minicomputers then \$20K workstations then \$2K PCs and now \$200 personal digital assistants (see Table 1). We are now entering the third evolutionary path which will require new sensors and transducers. New computing paradigms such as network computing are appearing. Single chip computers are likely to appear soon. The market for single chip computers may be 100 times greater than the current PC industry.

## Silicon Device Functionality and Physics

All the technology advancement predicted by Moore's law is not only driven by improved manufacturing processes, but also by a solid understanding of the physics behind the semiconductor technology. In this section we will give an introduction into the physics behind the semiconductors and the most common silicon devices. The focus is on passing on an intuitive understanding of the operation of these devices, without presenting too many formulas. The text is therefore not entirely rigorous and only ideal devices are presented, without their limitations.

### Conduction and Band Theory

Conduction requires the motion of charge carriers. The more easily the carriers move in response to an external field the more conductive the substance is. We typically divide substances into conductors and non conductors or insulators. Semiconductors fit into the broad region between these two extremes of the continuum.

In order to understand the underlying mechanism for semiconduction we will look at band theory and energy levels. For a single isolated atom, electrons are restricted to discrete energy levels. The exclusion principle permits only 2 electrons (with opposite spin) in any energy level. When 2 atoms are brought together, electrons occupying the same energy levels in each atom change energy levels slightly to produce 2 adjacent levels of energy. As more and more atoms are brought together the splitting continues (see Figure 2). For a large number of atoms, the original single energy levels split to become almost continuous bands of permitted energy. Note that the structure of these bands varies with atomic type, spacing, temperature etc.

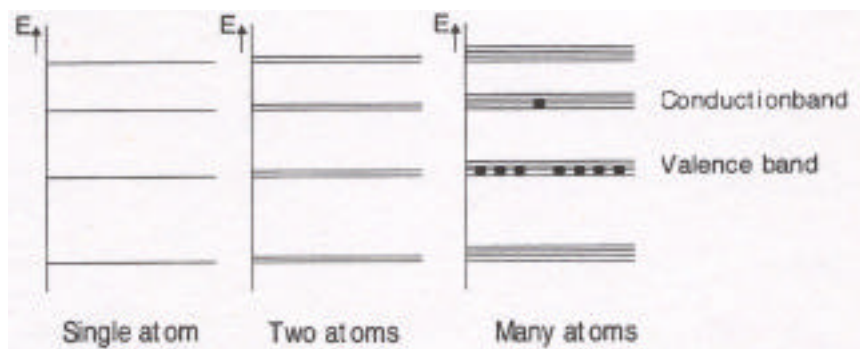


Figure 2. Formation of energy bands in semiconductor material.

Energy bands are ranges of energy where electrons *may* exist. We are interested principally in the bands that occur at the outer levels of an atom. Specifically the *valence* band, which contains the electrons which form chemical bonds, and the *conduction* band in which electrons are substantially free from direct atomic forces. It is the ease of movement of electrons between these two bands which determines whether substances are conductors, semiconductors or insulators.

The semiconductor material Si is tetravalent and it forms a covalent *diamond* crystal lattice, which has 4 adjacent tetrahedrally located atoms. The covalent bonding means that in general electrons are fairly tightly held in the valence band and thus not available for conduction. This explains the relatively low conductivity of pure semiconductors. Thermal or other energy is required to excite electrons into the conduction band.

### Holes

If an electron is excited into the conduction band it leaves a positively charged *hole* behind. A different electron, from another atom, may fall into the hole. In this fashion a long chain of "electron falling into hole" events can more easily be considered as the progression of a positive hole through the lattice.

It is important to remember that "holes" "conduct" in the valence band not the conduction band.

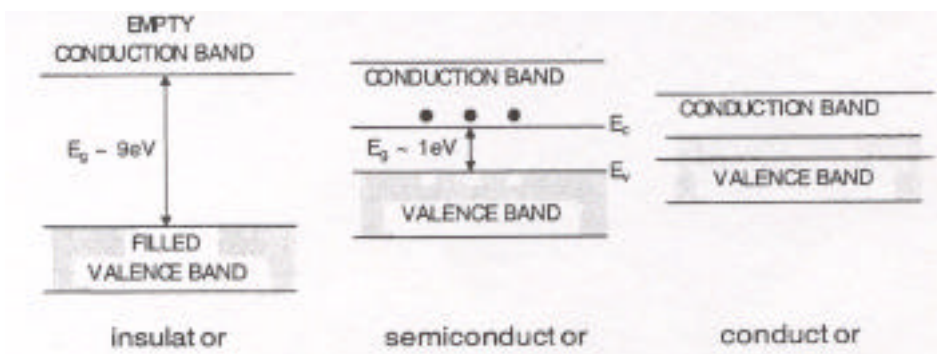


Figure 3. The band gap in insulators, semiconductors and conductors.

**Kinetic and Potential Energy**

The lowest energy level in the conduction band is labelled  $E_C$  and represents the minimum energy required for an electron to be in the conduction band. An electron with energy  $E_C$  is essentially at rest, i.e.,  $E_C$  is the potential energy of the electron. Electrons with energy higher than this are moving and have kinetic energy  $E_{ke}=E-E_C$ .

Similarly the highest energy in the valence band is labelled  $E_V$  and represents the minimum energy for a hole at rest. More energetic holes are lower in the valence band with kinetic energy  $E_{kh}=E_V-E$ . The band gap energy  $E_g=E_C-E_V$  is in general the minimum energy required to move from the valence to the conduction band. The larger the band gap, the less likely the material is to conduct electric currents (Figure 3).

$E_g$  decreases as temperature increases, so semiconductors become more like conductors at high temperatures. A table of room temperature bandgaps is shown below.

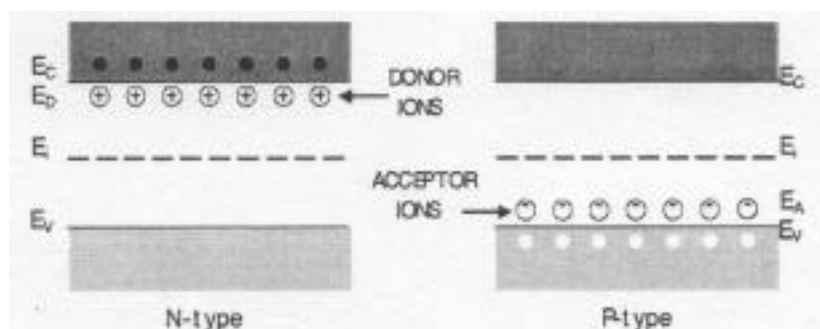


Figure 4. Doping of Silicon.

**Donors and Acceptors**

We can modify the parameters of semiconductors, by replacing some of the normal atoms with atoms of a different valence (Figure 4). This is known as *doping*. For Si substituting an atom of valence 5 in the lattice leaves an electron which is not held strongly in the covalent bonds, and is readily ionised at room temperature. Such atoms are known as *donor* atoms. The number density of donor atoms is written as  $N_D$ . Semiconductors with an excess of donor atoms are known as *n-type*. Similarly substituting an atom of valence 3 creates a hole in the covalent lattice which is also likely to be free to conduct at room temperature. These atoms are *acceptor* atoms, and have number density  $N_A$ . Semiconductors with an excess of acceptor atoms are known as *p-type*. These donors and acceptors create new positions in the energy band diagram as shown below. The thermal energy required to excite them into conduction is so low that we commonly assume they are completely ionised at room temperature.

Undoped semiconductors are called *intrinsic* semiconductors, doped semiconductors are known as *extrinsic* because the carriers are principally from non intrinsic dopant atoms.

## Electron and Hole Densities

The total number of electrons and holes in a semiconductor are labelled  $n$  and  $p$  respectively. Sometimes we subscript them to indicate the type of semiconductor they are in, e.g.,  $p_n$  is the number of free holes in an n-type semiconductor.

In an intrinsic semiconductor, there are “no” donors or acceptors and the hole and electron counts are due to ionisation of intrinsic atoms. Thus we define an intrinsic carrier density  $n_i = p = n$ . Typical values of densities are shown adjacent for Si (at 300K).

Particle	density (cm <sup>-3</sup> )
Si atoms	$5 \cdot 10^{22}$
$n_i$	$1.45 \cdot 10^{10}$
$(N_D, N_A, p, n)$	$10^4$ to $10^{18}$

## Thermal Equilibrium

New electrons and holes are produced by thermal energy. Thus, if there are no other inputs, we may assume that the rate of generation of electrons and holes is a function of thermal energy, say  $G = f_1(T)$ . Electrons and holes also recombine at some rate. Clearly for an electron hole pair to recombine, electrons and holes must exist, and indeed we would expect the number of holes and electrons to determine the rate. We may define a recombination rate  $R = pf_2(T)$ .

If there are no changing inputs there will finally be a thermal equilibrium level where the rate of generation equals the rate of recombination. At this point we have  $G = f_1(T) = pf_2(T) = R$ . We can rewrite this as  $pn = f_1(T)/f_2(T) = f_3(T)$ . This is true for all values of  $p$  and  $n$ , including intrinsic semiconductor, i.e.,  $n_i^2 = f_3(T)$ . Thus we get the *mass action law*:

$$pn = n_i^2$$

Additionally we have charge neutrality in the semiconductor so that  $p + N_D = n + N_A$ .

For an n-type semiconductor  $N_D \gg N_A$  and  $n \gg p$  and therefore  $n \approx N_D$ . From the mass action law we get  $p \approx n_i^2/n = n_i^2/N_D$ . As a result we see that n-type doping depresses the number of free holes. In n-type material we thus have many more electrons than holes. In other words, electrons are *majority* carriers, and holes are *minority* carriers.

Similar logic shows that for a p-type semiconductor  $n \approx n_i^2/N_A$ . For p-type we have holes as majority carriers, and electrons as minority carriers.

## Fermi Level

As noted before, energy levels are levels where electrons *may* exist. The probability that an electron fills a possible energy level is given by the Fermi-Dirac distribution function, or Fermi function  $f_D(E)$  given below.

$$F_D(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

where  $k$  is the Boltzmann constant ( $8.62 \times 10^{-5}$  eV/K) and  $E_F$  is the Fermi level, the point at which there is a 50% probability of occupation of that energy level by an electron. For  $E < E_F$  the level is more likely to be occupied, for  $E > E_F$  the level is less likely to be occupied. For typical doping levels  $E_V < E_F < E_C$ .

## Doping

By adding *donor* atoms to the intrinsic Si, we create an n-type semiconductor. Adding these atoms with a density  $N_D$  raises the Fermi level from its intrinsic level by:

$$E_F - E_i = kT \ln(N_D/n_i)$$

Similarly, by adding *acceptor* atoms with a density  $N_A$  to create p-type semiconductor we lower the Fermi level, as expressed by:

$$E_F - E_i = -kT \ln(N_A/n_i)$$

## PN Junctions

A plot of conduction, valence and Fermi levels is shown for adjacent, but non contacting p-type and n-type semiconductors in Figure 5. As explained previously, the Fermi levels differ.

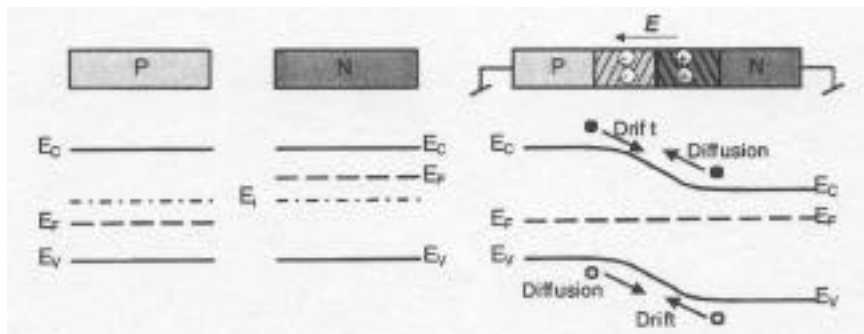


Figure 5. Energy band bending in the formation of a PN junction.

If the two sides are brought together there will be a diffusion current of holes from p-type to n-type, and electrons from n-type to p-type due to the different concentrations of holes and electrons in both types of material. This leaves a charged region in the centre of the junction where carriers have diffused away from the donor and acceptor atoms. This is called the *depletion region* as there are few charge carriers, or the *space charge region* as the ionised dopant atoms are not neutralised by carriers. This charged region results in an electric field which induces a drift current in the opposite direction. At equilibrium the currents cancel.

**Ideal Diode Equation**

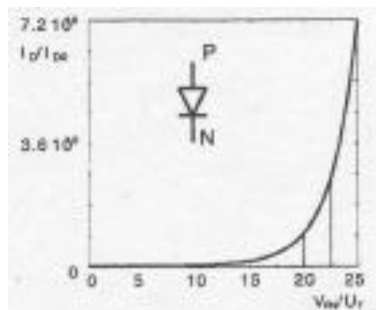


Figure 6. The current through an ideal diode as a function of the voltage across it.

When we forward bias a PN junction we add an additional voltage across the depletion region. This reduces the potential barrier and increases the diffusion current. It also decreases the depletion region width. The reduced potential and reduced width result in an approximately constant electric field and hence the drift current is substantially unchanged. The diffusion and drift currents thus do not cancel out anymore and a net current results. This is the diode current expressed by:

$$I = I_s (e^{V_w/U_T} - 1)$$

where  $I_s$  is the saturation current, which is a device constant, and  $U_T = kT/q$  is the thermal voltage, which is about 25mV at room temperature. This is the *ideal diode equation* and it is plotted in Figure 6.

**Bipolar Junction Transistors**

So if we apply a positive bias to the PN junction, we increase the majority carrier density at the edge of the depletion region, and therefore increase the diffusion current across the junction. The increased diffusion current creates an injected minority carrier density on the far side of the depletion region and a net current flows across the junction. A reverse bias across the junction depletes majority carriers near the depletion region edge. Note that any charge carriers available can move across the junction as in fact they will have the drift field in their favour. The limit is the carrier depletion at the edge of the depletion region. If we look at the last case in more detail, and imagine an unspecified source of electrons on the p side of the junction, the electrons could diffuse in the p-type region. Many would recombine with the holes in the region but some would eventually diffuse to the depletion region, get caught in the drift field and be swept across to the n-type region.

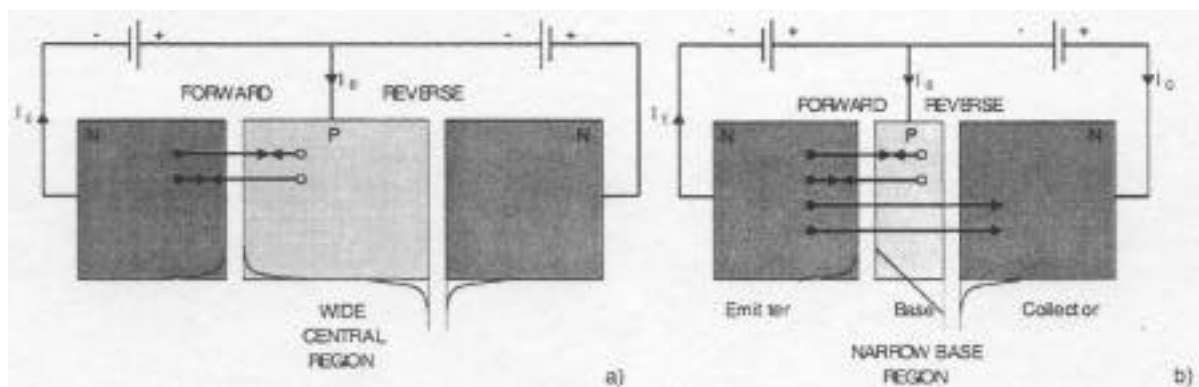


Figure 7. a) The BJT structure but with a wide central region; b) The BJT with a narrow base region.

Let us now look at the 3 layer BJT structure created by two diodes head to head and forward bias one diode while reverse biasing the second one. A diffusion injected minority carrier charged region will exist on either side of the forward biased NP junction (see Figure 7a). Similarly a diffusion/drift extracted minority carrier depleted region will exist on both sides of the reversed biased PN junction. The distance between the two junctions will result in recombination of minority carriers, before they can diffuse from the NP junction to the PN junction. If we decrease the width of the P region enough, however, a substantial number of minority carriers (electrons) will diffuse to the edge of the PN depletion region (see Figure 7b). They will then be carried by the drift field to the right hand N region where they will be majority carriers again. Thus the presence of injected minority carriers, in close proximity to the reverse biased junction has set up a current through the junction despite the reverse bias. This current is expressed as:

$$I_C = I_S e^{V_{BE}/U_T} (1 - e^{-V_{CE}/U_T})$$

where  $I_S$  is the specific current of the BJT and is a device constant.

The left hand N region in Figure 7b is known as the *emitter* as it injects carriers into the central region. The right hand N region is known as the *collector* as it collects some of the emitted carriers. The central P region is known as the *base* for historical reasons, as in the first point-contact transistor it was a block of Ge which formed the mechanical base of the device. A plot of the collector current is shown in Figure 8. The current depends exponentially on the base-emitter voltage. As the collector-emitter voltage becomes larger than a few  $U_T$  its influence on the current through the device can be ignored.

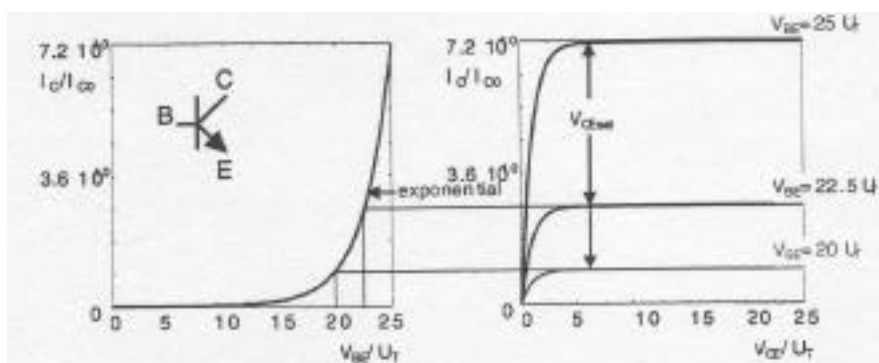


Figure 8. The collector current  $I_C$  through a BJT as a function of  $V_{BE}$  and  $V_{CE}$ .

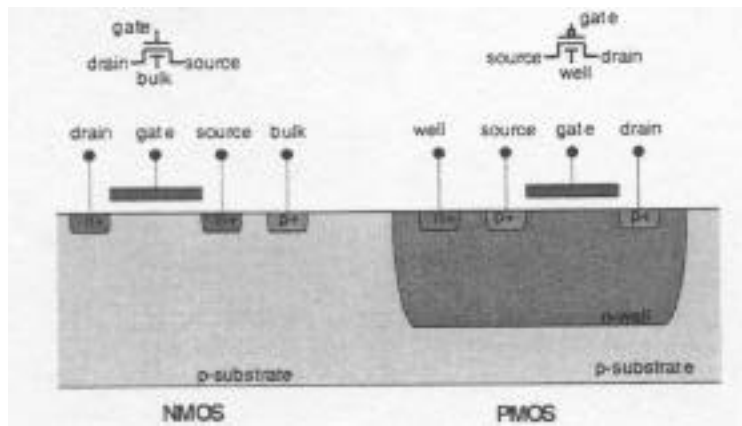
For charge neutrality in the base region  $N_{AB} + n_{PB} = p_{PB}$ . The current from base to collector is limited by  $n_{PB}$ ,  $p_{PB}$  comes from holes supplied by the base contact, and  $N_{AB}$  is a constant. Thus the base hole current permits injected electrons to enter the base from the emitter, which may diffuse to the collector. In this way the base/emitter current controls the flow of electrons from E to C. Note that the base hole current controls the number of electrons in the base region, but the electron current depends on other physical and mechanical properties, and can be much greater. In this way a small BE current may control a larger CE current.

### MOS Field Effect Transistors

MOSFETs are field effect transistors, i.e., they use electric fields to control the movement of charges, and thereby control current flow. In MOSFETs the gate is isolated from the channel by an insulating oxide in a metal oxide semiconductor (MOS) structure. Nowadays a layer of doped polycrystalline silicon is used instead of the metal layer as the gate terminal, but the name MOSFET is still used for the device (see Figure 9). MOS transistors come in two types, depending on the type of diffusion used for their drain and source terminals. Before studying the device operation we shall look at the mechanisms of channel formation while ignoring the influence of the drain and source electrodes.

### MOSFET inversion channel

The MOS structure consists of polysilicon gate electrode separated from a semiconductor substrate by an insulating layer of oxide. Initially we assume that the Fermi level in the polysilicon and the semiconductor are the same in isolation, so that there is no change to the energy bands in the composite device at equilibrium. (This is not an absolute requirement as we shall see - it simply eases understanding.) The MOS structure thus has a uniform charge throughout the semiconductor, with no surface charge on the metal.

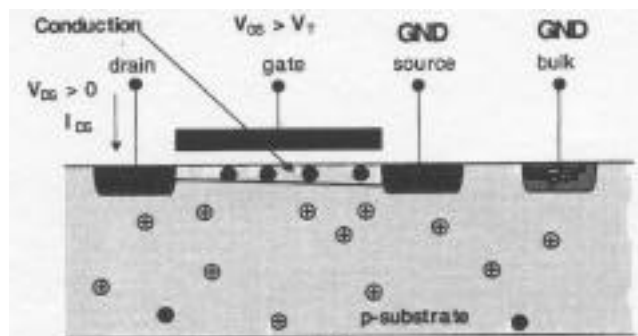


**Figure 9.** Simplified MOS transistor structures in a modern integrated circuit process and the transistor symbols.

For our example we will assume a p-type semiconductor as our substrate and n-type drain and source diffusion, i.e., an NMOS transistor. A similar result holds for the PMOS transistor. If an external negative voltage is applied to the gate electrode, a field is induced across the oxide and into the semiconductor. The field attracts the positively charged holes (the majority carriers) to the surface just under the oxide and the negatively charged gate. This is known as accumulation, where the majority carriers are attracted to the oxide surface.

If a small external positive voltage is applied, The positively charged holes move away from the positively charged gate and thus from the oxide. This creates a depletion region under the gate oxide. The space charge of the depletion region is balanced by the surface charge on the gate side. As the voltage increases, the depletion region widens. At the same time that the majority carriers (holes) are pushed away from the positively charged gate, the minority carriers (the electrons) are attracted to it.

If the external voltage is further increased, we will actually get a situation where we will have more electrons than holes at the silicon surface under the gate and the minority carriers have locally become the majority carriers. This is called inversion. Further increases in voltage tend to increase the electron count rather than widen the depletion layer, so the width reaches an effective maximum  $W_M$  under strong inversion.



**Figure 10.** Channel current through an NMOS transistor.

## MOSFET channel current

The drain and source terminals consists of n-type contact region on each side of the MOS structure, and we have a region of high resistivity between the 2 terminals, due to the reversed biased diode junction at one or other end. Thus no current flows. If we bias the MOS diode into accumulation, i.e., we will have lots of holes available for conduction, but one of the contacts will still be reverse biased and still no current can flow.

On the other hand, under positive bias the region under the gate will first deplete and then invert, forming a channel of free electrons between the contacts. This device in which we induce a channel under gate bias is called an *enhancement* MOSFET. A similar device can be generated with a lightly doped n-type channel between the source and drain, which has conductivity under zero bias. The conductivity can be increased by field induced channel enhancement as above or reduced, by field induced channel depletion. This is a *depletion* MOSFET. Modern day integrated circuit technology nearly exclusively uses enhancement MOSFETs. Both types of MOSFET can be formed with either p-type or n-type semiconductors. The type of the device is determined by the type of the channel under conduction, not the type of doping of the substrate. Thus n-channel MOSFETs are made on p-type substrates.

MOSFETs only conduct in the inversion mode. Thus for any type of MOSFET there is a threshold voltage on one side of which the channel is cutoff, and on the other the channel is conductive. The threshold voltage is commonly denoted  $V_T$ . In MOSFET IC technology the threshold voltage is a critical parameter and often special processing steps are added to set its value.

When the drain-source voltage ( $V_{DS}$ ) is zero, we may induce an inversion channel but we will still have no current through the devices as there is no lateral electric field that makes the carriers move right or left. If we increase  $V_{DS}$ , the current through the device will increase proportionally to this voltage, with a slight complication. As the drain voltage is increased, the electric field between the gate and the inversion channel at the drain side decreases and we will have less electrons in the channel on the drain side. The situation at the source end of the channel is unchanged, so the total number of electrons in the channel decreases, which in turn reduces the conductivity of the channel. The current through a MOS transistor as a function of  $V_{GS}$  and  $V_{DS}$  is given by:

$$I_D = ((V_{GS} - V_T) - (V_{DS}/2)) V_{DS}$$

Where  $\beta$  is a device constant which depends on the technology and the geometry of the device. The first term in the equation ( $V_{GS} - V_T$ ) is proportional to the amount of electrons induced in the channel by the gate-source voltage. The second term ( $V_{DS}/2$ ) is proportional to the amount of reduction of electrons in the channel by the increase in drain voltage. When the drain voltage increases to a voltage  $V_P = V_{GS} - V_T$ , there are effectively no electrons left at the drain end of the channel and the channel is said to be pinched off.  $V_P$  is called the pinch-off voltage. When the drain voltage is larger than the pinch-off voltage, it ceases to influence the current through the transistor, and the drain current can be simply described by:

$$I_D = \frac{\beta}{2n} (V_{GS} - V_T)^2$$

The drain current in both modes of operation is shown in Figure 11.

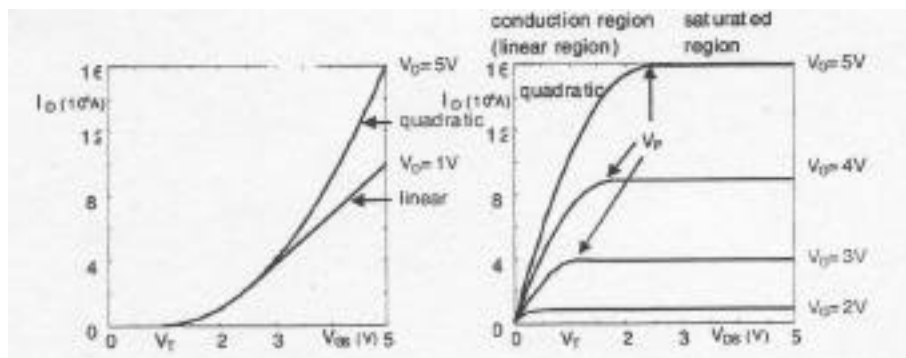


Figure 11. Drain current through an NMOS transistor as a function of  $v_{GS}$  and  $V_{DS}$ .

## Silicon Sensors and Actuators

The semiconductor devices discussed in the previous section can be connected on a chip into electronic circuits that process information (which we will discuss in the next section). A chip, however does not exist in isolation and ultimately interacts with the world around it through sensors and actuators. Most sensors and actuators are external to the silicon IC and we will not discuss these here. Some sensors can be built directly on the silicon that holds the electronic circuits and these days even some actuators are built on chip. Specifically, photonic sensors are for instance integrated with signal processing circuits in current digital cameras, or, as another example, in the Logitech Marble trackball, which optically measures the rotation of a ball to move the cursor on a computer screen. A recent development is to combine mechanical sensors and actuators on the micrometer scale with electronics directly on silicon is MicroElectroMechanical Systems (MEMS).

### Photonic Sensors

In our discussion of the semiconductor devices, we have considered only thermal energy as the source of new charge carriers. The ionisation energy can however also be provided by photon absorption. If a photon has energy  $h\nu$  comparable to the band gap energy  $E_g$ , energy transfers may occur between photons and carriers in the semiconductor. For instance, a photon with  $h\nu > E_g$  may be absorbed in the lattice, ionising an atom and creating an electron/hole pair.

## Absorption

If  $h\nu = E_g$  a simple transition may occur. For  $h\nu > E_g$  the additional energy is lost to the lattice as thermal energy. These band to band transitions are called intrinsic transitions. For  $h\nu < E_g$  absorption can only occur if energy levels exist in the bandgap, due to impurities or lattice defects. Such transitions are known as extrinsic transitions.

If we have a photon flux  $\phi_0$  photons/cm<sup>2</sup>s, and  $h\nu > E_g$  on a semiconductor, a fixed fraction of the photons will be absorbed per unit distance travelled. The photon flux decreases as  $\phi(x) = \phi_0 e^{-\alpha x}$ . We define  $\alpha$  as the *absorption coefficient*. We could regard it as the inverse of the light penetration depth.

Note that high energy photons (short wavelengths) are absorbed quickly whereas lower energy ones may pass straight through. Absorption can only occur for wavelength shorter than the critical wavelength which is given by  $\lambda_c = 1.24/E_g$   $\mu\text{m}$ . So for  $\lambda < \lambda_c$  absorption may occur, and as  $\lambda$  decreases  $\alpha$  increases and the penetration depth  $1/\alpha$  decreases. Note that as the energy increases (with decreasing wavelength) beyond  $E_g$  more energy is converted to heat in the lattice.

## Photodiodes

If we operate a diode under reverse bias with an optical input to the junction, electron-hole pairs will be generated and will separate and drift towards the opposite electrodes, and into the external circuit. To maximise high frequency response we want a short transit time through the depletion region, i.e. a narrow depletion region. To maximise efficiency we want the longest possible depletion region to absorb photons. Thus we have conflicting requirements for efficiency and response time.

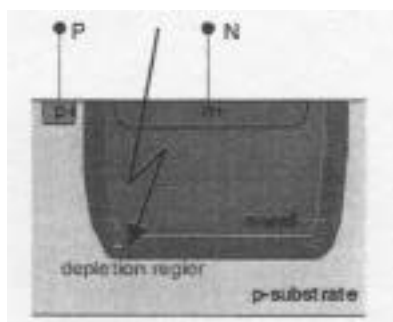


Figure 12. A photodiode structure.

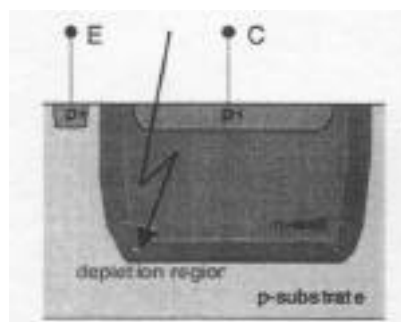


Figure 13. The photodiode structure adapted to form a photoBJT.

## Efficiency

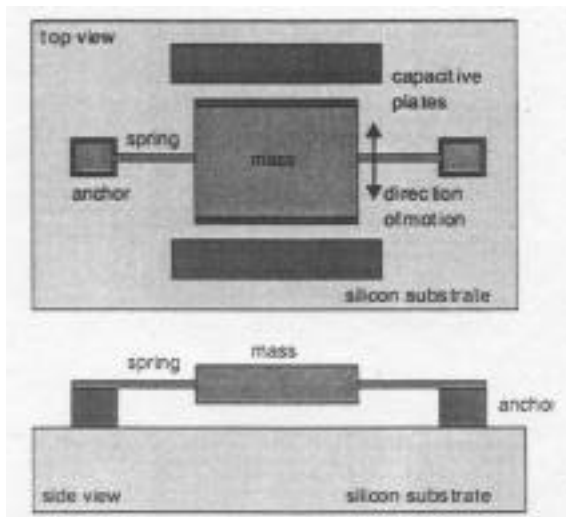
The quantum efficiency  $\eta$  is defined as the number of electron hole pairs generated over the number of absorbed photons. As might be expected the absorption coeff  $\alpha$  has a strong effect on  $\eta$ , and hence  $\lambda_c$  has a critical effect on  $\eta$ . For  $\lambda > \lambda_c$ , absorption is negligible and hence  $\eta$  is low. For  $\lambda < \lambda_c$ ,  $\alpha$  is large and all the absorption occurs near the surface. The surface of a semiconductor lattice has a large number of loose bonds, and these act like recombination centres, thus the recombination is much higher and  $\eta$  drops again.

Thus high quantum efficiency is usually only possible over a fairly narrow band. Si has good absorption .8-.9  $\mu\text{m}$ , i.e. for infra-red light.

## PhotoBJTs

The light sensitivity of the diode structure can also be used to forward bias the base-emitter junction of a BJT structure created with the same well as the photodiode if the base terminal is floating. The advantage of this procedure is that the current generated by the photons is now effectively the base current of the BJT and will be multiplied by the strong current gain of the BJT at the collector. We thus get a lot more current out of this device for the same photon flux. A disadvantage of this structure is that before the current gain kicks in, the base emitter junction needs to be forward biased. In darkness, the voltage over the base-emitter junction will be 0V and after the onset of light, the small photon induced current is needed to charge the well first and this is a slow process. The photoBJT is therefore a much slower device than the photodiode.

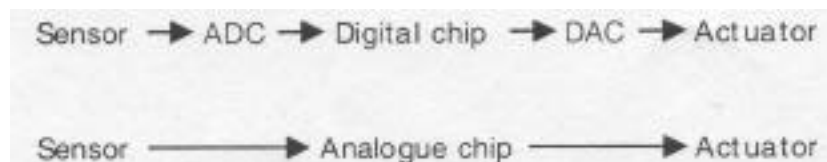
**MicroElectroMechanical Systems**



Micoelectromechanical Systems (MEMS) are enabled by recent advances in chip fabrication technology. The MEMS technology not only allows us to create electronic devices on the silicon, but also mechanical structures, such as springs, beams, masses, etc. We can use these MEMS as sensors and actuators at a micrometer scale, or by using them in and array, we can even generate effects on larger scales. The mechanical structures are created by depositing the structure on a supporting layer and subsequently etching away the supporting layer, so that the structure becomes free standing.

Capacitive sensing, piezo-electric effects, alternating electric and magnetic fields and temperature dependent mechanical deformation some examples of the mechanisms used to create MEMS sensors and actuators. Current applications in MEMS include micro-motors, accelerometers, pressure sensors, micro-optics and fluid pumps.

**Figure 14.** A simplified MEMS structure.



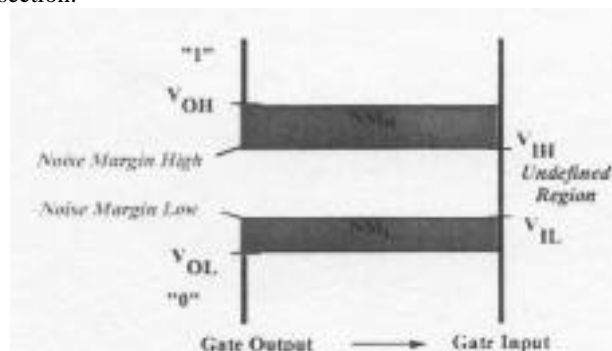
**Figure 15.** Sensors and actuators form the interface between chips and the analogue world.

**Digital and Analogue Information Processing**

The world around us is inherently analogue, i.e., signals, such as light or sound intensity, speed, temperature, etc., are continuous valued signals that vary over time in a continuous manner. Sensors and actuators therefore mostly capture or create analogue signals, such as voltages or currents. Most on chip processing these days, however, is digital, using only 1s and 0s to process the information over time in a step-wise manner. In order to connect the digital processing unit with the analogue parts, we need to convert the analogue signals to digital signals and vice versa, using Analogue-Digital Converters (ADC) and Digital-Analogue Converters (DAC), as shown at the top in Figure 15. Alternatively, we can directly process the voltages and currents from the sensors using analogue information processing circuits. In the following sections we will look at both means of on chip information processing.

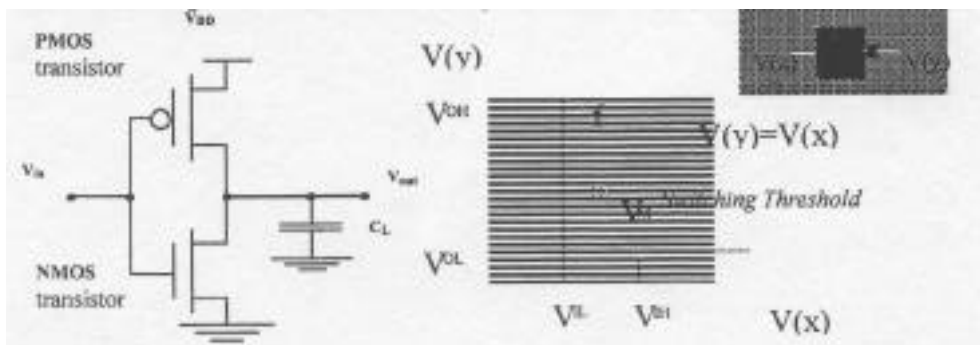
**Digital Information Processing**

Information is processed digitally usually by adopting a binary representation. A binary digit (bit) is directly related to the state of a digital circuit. Usually, a binary digit corresponds to a node in a circuit having a voltage close to ground or close to the positive power supply. Figure 16 shows the precise definition of logic '1' and logic '0' in a digital circuit. Since it is physically impossible to measure a voltage with infinite precision it is necessary to consider an undefined region or range of voltages for which it is not known whether the signal is logic '1' or logic '0'. Further, to guarantee correct operation of transistor switches it is necessary to define noise margins at their inputs and outputs. We will discuss this further in a later section.



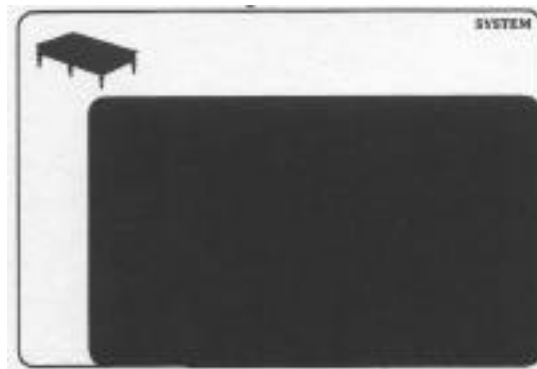
**Figure 16.** Mapping of binary digits onto voltages in digital circuits.

Groups of binary digits may be used to represent all types of information including integer or rational numbers, characters and special codes or instructions which computers can interpret in order to execute computer programs. Information processing can then be viewed as the mapping of a sequence of binary digits to a different sequence of binary digits. The information revolution has been based on the fact that silicon integrated circuits can perform such mappings at very high speed using very little space and power. The processing of binary digits in a silicon integrated circuit relies on transistors being used as switches. An example of this is shown in Figure 17.



**Figure 17.** A CMOS inverter and its voltage transfer characteristic.

The operation of the circuit is as follows. Initially the voltage at the input is low corresponding to a logic '0'. This means that the PMOS transistor is conducting and the capacitor  $C_L$  is charged up. As the voltage at the input is increased and reaches a value  $V_{IL}$  the PMOS transistor begins to reduce its conductivity while the heretofore high impedance NMOS transistor starts to conduct. The voltage range  $0V$  to  $V_{IL}$  is referred to as the input voltage low noise margin (see Figure 16) and represents the input voltage range which is interpreted as a logic '0' by the inverter circuit. As the input voltage continues to rise the PMOS transistor ceases to conduct while the NMOS transistor is conducting strongly, thus discharging  $C_L$  to ground. Hence we see that the circuit performs a logical inversion between its input and output. The design of a digital information processing system consists of the combining of such circuits in a hierarchical manner in order to implement the desired function. The designer deals with the complexity of the circuits by viewing the design at different levels of abstraction as shown in Figure 18.



**Figure 18.** The digital design hierarchy.

In order to illustrate how arithmetic processing can be performed in silicon integrated circuits are few more circuits are needed. Figure 19 shows a NAND gate circuit. This circuit can be used a building block for any arithmetic operation.

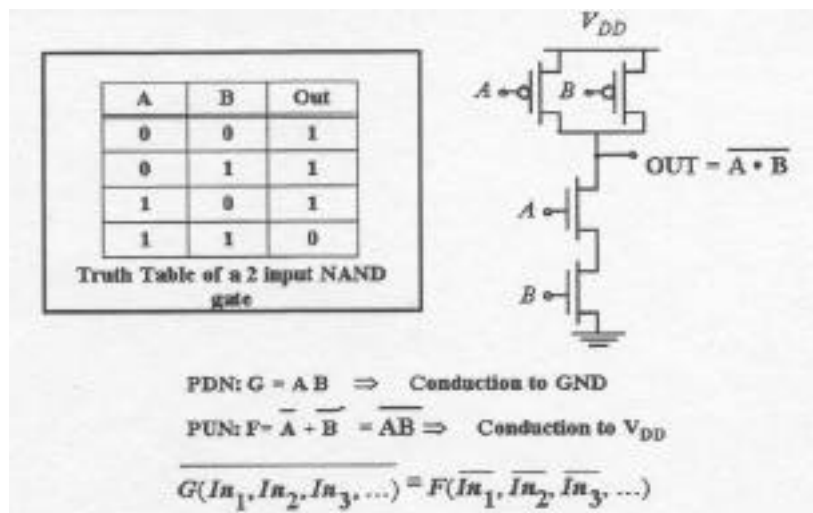


Figure 19. A nand gate circuit and its functional description.

The other important ingredient for complex information processing is memory. Figure 20 shows the two common types of digital memory circuits. The first circuit is the basis for a type of memory known as Static Random Access Memory (SRAM). The second circuit is known as Dynamic random access memory (DRAM). Each of these circuits can store one binary digit. SRAM uses the two back to back inverters to create a positive feedback loop to maintain the value of the binary digit stored in the loop when it is first closed. Additional circuitry is needed to read and write the binary digits. The DRAM circuit uses the principle of charge storage to store the binary digit. It has the advantage of being a very small simple circuit, and hence very high densities are achievable. It has the disadvantage of being slower to read and write and needing special circuitry to counteract the effects of charge leaking away from the capacitor.

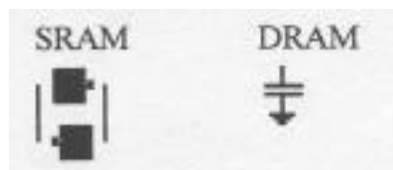


Figure 20. Digital memory circuits.

Given such circuits we are then in a position to design any arithmetic function we wish and store any intermediate results in memory. An example of such a function is binary addition shown in Figure 21. The function of the half adder can be described by the following logic equations:

- $sum = (a \text{ and not } (b)) \text{ or } (\text{not } (a) \text{ and } b)$
- $C_{out} = a \text{ and } b$

where the sum output is the sum of two binary digits a and b, and  $C_{OUT}$  is a binary digit representing the carry. The full adder then allows a for a carry input  $C_{IN}$  by adding the sum of the two operands to the carry in. Such circuits can then be cascaded to produce adders for the desired number of bits in the operands.

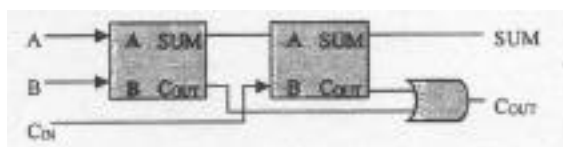


Figure 21. Full adder constructed from two half adders and an or-gate.

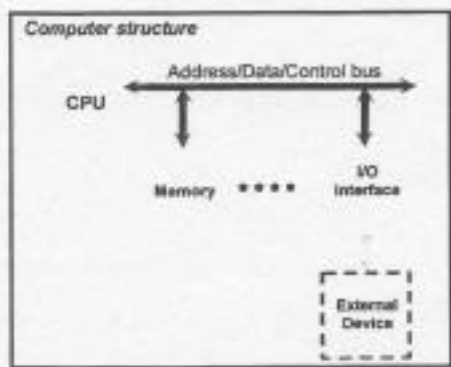


Figure 22. Structure of a desktop computer.

Finally, we observe that such circuits can be composed to build whole computer systems. The structure of a typical desktop computer is shown in Figure 22.

The CPU is the central processing unit which performs calculations based on the lists of instructions which comprise the operating system and user applications. Data and programs are stored in memory which is accessed by digital signals emanating from the CPU, travelling across the motherboard to the memory devices. In addition, circuitry is provided to connect the computer to external devices such as keyboards and modems etc. These are referred to as Input/Output (I/O) devices.

### Analogue Information Processing

So far we have discussed information processing in terms of transformations of binary digits. This is not the only way silicon integrated circuits can process information. Instead we can represent information by the actual values of voltages and currents themselves in the circuits. A combination of the transistor physics and the topologies of circuits then allow the transformation of the voltages and currents at the input side of a circuit to a desired mapping on the output side. An example of such a circuit is shown in Figure 23. Such circuits can perform information processing using very little area and power at the cost of reduced arithmetic precision due to noise on the currents and voltages and small variations in the fabricated components. Applications of such processing have included specialised circuits for implantable pacemakers and defibrillators and models of the human cochlear. This particular structure computes the vector length of an n-dimensional vector with only  $3n+5$  transistors.

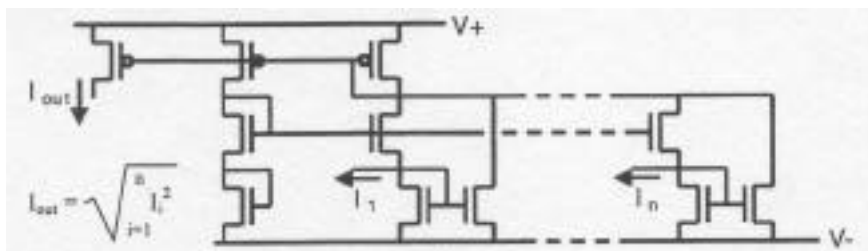


Figure 23. An analogue CMOS circuit to compute the vector length of the input currents  $I_1, \dots, I_n$ .

### References

[1] Robert R. Schaller, Moore's Law: past, present, and future, IEEE Spectrum June 1997, 53-59.  
 [2] A.S. Sedra and K.C. Smith, Microelectronic Circuits, 4<sup>th</sup> Edition, Oxford University Press, 1998.  
 [3] R.J. Coggins, M.A. Jabri, B.F. Flower and S.J. Pickard, "A Hybrid Analog and Digital VLSI Neural Network for Intracardiac Morphology Classification", IEEE Journal of Solid State Circuits, Vol. 30, No. 5, May 1995, 542-550.  
 [4] X. Arreguit, F.A. van Schaik, F. Bauduin, M. Bidiville, and E. Rieber, "A CMOS Motion Detector System for Pointing Devices," IEEE Journal of Solid State Circuits, Vol. 31, No. 12, December 1996, 1916-1921.