Nanoelectronics

D.J. Paul

Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge, CB3 0HE, U.K.

Outline

1. Introduction
2. Silicon Nanoelectronics and Ultimate CMOS
3. Single Electron Devices
4. QuantumMechnanical Tunnel Devices
5. Spin Nanoelectronics or Spintronics
6. Fabrication Technology
7. Self-Assembly
8. Molecular Nanoelectronics
9. Alternative Architectures
10. Summary and Future Prospects

Glossary

Nanoelectronics: The field of devices controlled using the electronic properties of materials with dimensions below 100 nm in size.

Quantum mechanical tunnelling: The ability to use the wave properties of an electron to allow transmission (electron tunnelling) through a thin potential barrier.

Single electron devices: Devices which have switching properties controlled by the addition or substraction of one electron or through which only one electron may be transported at any one time.

Spin nanoelectronics: The use of the electron spin for computation or memory storage.

Transistor: A three terminal device used as a switch or amplifer where one terminal controls the electron current flowing between the other two terminals.

Abstract

Nanoelectronics is the field of computation and control in the nanometer scale regime (below one thousandth the diameter of a human hair) using the electronic properties of materials. Useful logical computation or storage of information may be achieved by a number of different concepts and devices. Logical circuits may be used for computation, communication, control systems or storage of information. This article reviews the major nanoelectronic devices, fabrication techiques and architectures. The devices include silicon transistors, single electron transistors, resonant tunneling diodes, magnetic spin devices and molecular devices.

1. Introduction

Many new technologies appeared during the 20th century. If one had to decide on which new technology had the largest impact on mankind, the microelectronics industry would certainly be one of the main contenders. Microelectronic components in the form of microprocessors and memory are used in computers, audio-visual components from hi-fis and videos to televisions, cars (the smallest Daimler-Benz car has over 60 microprocessors), communications systems including telephones and mobile phones, banking, credit cards, cookers, heating controllers, toasters, food processors..... the list is almost endless. It is almost impossible to live in the western world today and not use a microelectronic component in every day life or eat food which has not had a microchip involved in the management, harvesting, processing, distribution or sale of the product.

The microelectronics industry has only been around for the last 50 years or so. The first bipolar transistor was demonstrated at Bell Laboratories in 1948 and the first field effect transistor (FET) appeared in 1960. Since that date, the microelectronics industry has been growing at an exponential rate which was first analysed by Gordon Moore, founder of Intel, in 1965. Moore showed that the size of a transistor gate is halved every 18 months, a feature which is now referred to as Moore's Law. This halving in size has been driven by economics. The smaller the gate, the faster the transistor can switch, the less power it consumes and the larger the number of transistors which can be integrated onto the one silicon chip. The increase in number of transistors and the associated higher manufacturing yields reduces the cost per transistor and therefore it is economical to scale the transistor to smaller and smaller sizes.

At the start of 2001, the gate length on the silicon metal oxide semiconductor field effect transistors (MOSFETs) on the latest Intel Pentium IV microprocessor chip will be 100 nm with a gate oxide thickness of 2 nm and a total of 28 million transistors per processor. The microelectronics industry has therefore become nanoelectronics named after the Greek for a dwarf "nanos". This article will review the silicon nanoelectronic field and discuss how far the silicon MOSFET can be scaled down. At some point this scaling must end and therefore for the continued growth of the electronics industry, an alternative technology must be found. The remainder of the article will review other nanoelectronic devices and circuits presently in the research field which may eventually be able to take over the transistor mantle or open up new electronic applications.

2. Silicon Nanoelectronics and Ultimate CMOS

All semiconductors such as silicon are insulators until impurities are added to the system. Silicon has four outer shell electrons which form tetrahdral bonds to four other silicon atoms in a crystal and therefore all the electrons are bound and the material is insulating. If a group V atom such as phosphorus is added with five electrons in the outer shell then there is an additional electron after four electrons are used to bond to the sur-

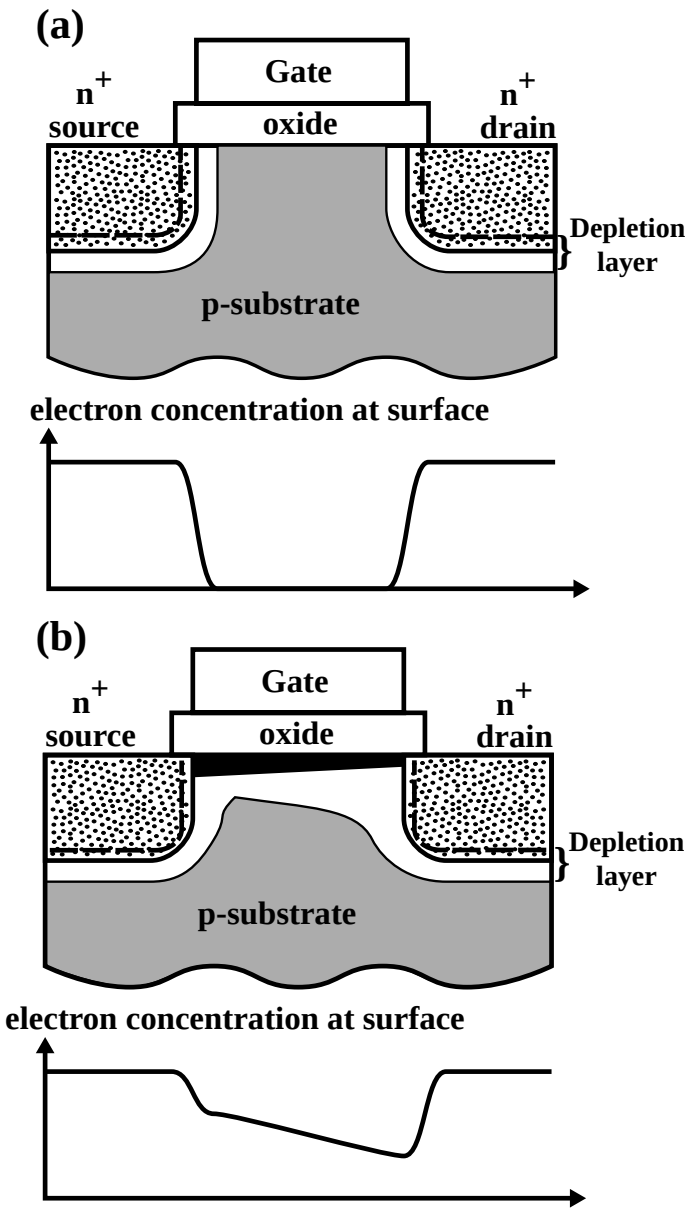


Figure 1:- The operation of a FET. (a) The transistor switched off with no voltage applied to the gate. The p region prevents electrons from travelling from the source to the drain (b) the transistor switched on with a voltage applied to the gate which attracts electrons to the gate and therefore the electrons may travel from the source to the drain.

rounding silicon atoms which allows electronic conduction when a voltage is applied. The silicon is said to be n-type as the conduction is through electrons. If a group III atom such as boron replaces a silicon atom then the boron can bond to three silicon atoms in the lattice but there is a hole left where the fourth bond used to be. Therefore bonding electrons from the outer shell of other silicon atoms can move to fill this hole and they leave a hole behind from where they came from in the crystal. Therefore the silicon can conduct as electrons move into holes in the system. This is p-type silicon. If p-type silicon is placed beside an n-type piece of silicon then a depletion layer forms where electrons and holes annihalate each other and electrons cannot flow between the two regions unless a large voltage is applied between them. Therefore a p-n junction stops electrons from being transported through the system.

The basic operation of a silicon MOSFET is shown in Figure 1. It is a p-silicon substrate with an insulator (silicon dioxide) and a metal gate on top. It is basically two back-to-back p-n junctions so that with no gate voltage applied no current flows from the source n-type contact to the drain n-type contact. As a voltage is applied to the gate, electrons are attracted to the gate and form a thin layer near the surface of the p-type silicon. Electrons can now be transported from the source to the drain and a current flows between the source and drain. The gate therefore switches the current between the source and drain off and on.

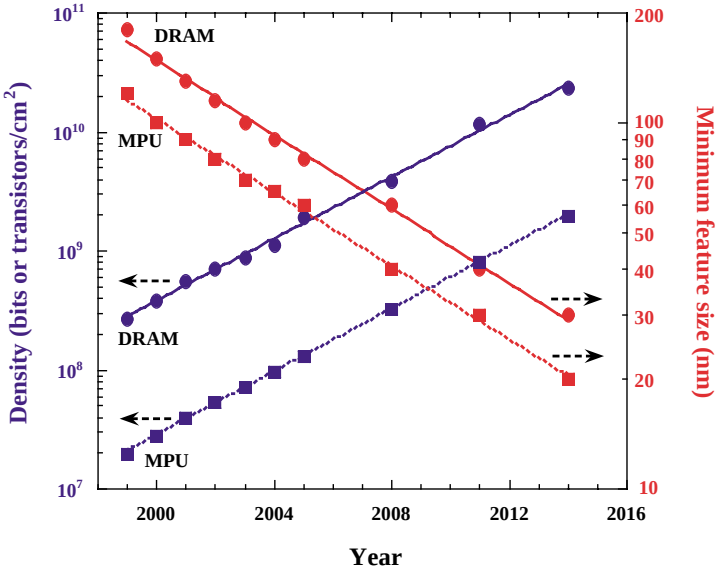


Figure 2:- Moore's law for the future as predicted from the SIA 2000 Roadmap. DRAM is dynamic random access memory used in a computer and MPU is microprocessing unit - the central processor of a computer.

Figure 2 shows the gate length and density on memory and logic chips found inside present computers along with predictions from the Semiconductor Industry Association (SIA) Roadmap for the next thirteen years. Already in 2001 dynamic random access memory chips (DRAM) have 270 million bits per cm^2. Each DRAM bit consists of a capacitor and a transistor with a gate length of 180 nm. The microprocessing units (MPU) such as Pentium and Power PC chips have 19.7 million transistors per cm^2 with gate lengths of 100 nm. Most of these transistors are in fact memory in the form of static random access memory (SRAM). A single SRAM cell consists of six transistors. Attempts to integrate the more compact DRAM memory cells onto MPU chips resulted in substantially poorer logic transistors while the six SRAM transistors are identical to those used in logic and therefore are easy to integrate. The MPU chips use a complementary metal oxide semiconductor (CMOS) architecture which incorporates two MOSFET transistors one which uses electrons as the conducting layer and a second which uses holes. The advantage of this architecture is that power is only dissipated (apart from that resulting from any leakage currents in the transistor) when the circuit is switched. This substantially reduces the power consumption of the chip. As the integration density of a chip is limited by the power dissipation, the ability to scale to smaller transistors aids the integration density from both the size and power requirements.

While the SIA Roadmap of Figure 2 suggests that CMOS MPUs may be scaled to 20 nm transistor gatelengths in 2014, the SIA Roadmap also contains a list of requirements which have to be met to allow the scaling of silicon to continue to this

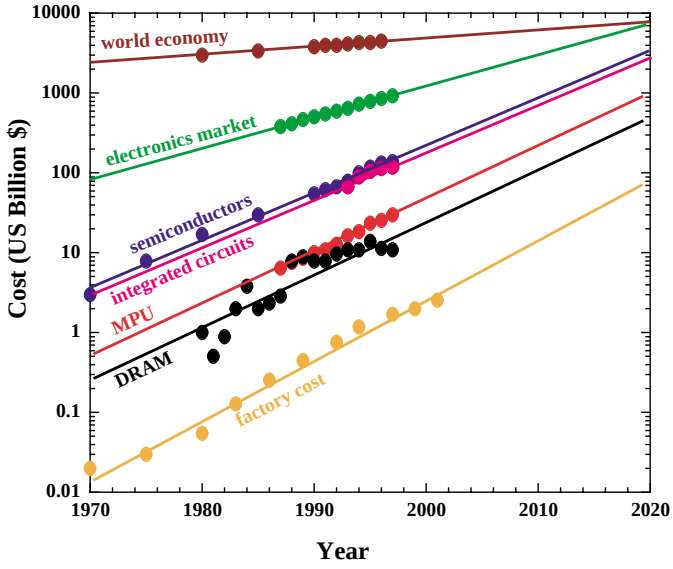


Figure 3:- The economics of the semiconductor industry with data taken from Semiconductor International.

level. Many of these requirements have no present or known solutions and therefore the ability to scale CMOS to 20 nm is not guaranteed. In the scaling process, all the parts of the transistor are reduced to try and keep all the electric fields in the device constant. In reality this cannot be achieved in short channel devices and the electric field is increased resulting in a number of problems, the most significant being drain induced barrier lowering (DIBL). The p-region in Figure 1 produces a barrier between the source and drain n-type contacts. If the gatelength is reduced then the source and drain n-regions deplete out carriers in the p-type region and the barrier between the source and drain is reduced i.e. drain induced barrier lowering. The way around this is to increase the p-type dopant density under the gate but this can only be achieved up to a certain level. Eventually electrons can quantum mechanically tunnel through the barrier between the source and drain when they are close enough together. Therefore the transistor can never be switched off and ceases to be a switch. NEC have, however, demonstrated a transistor which has a gatelength of only 8 nm. This demonstrates that transistors can be made to operate down to this gatelength although due to the DIBL, the gain or amplification of the transistor is much lower than that predicted by direct extrapolation from larger gatelength transistors.

A second problem is that electrons can quantum mechanic ally tunnel through the gate oxide when it becomes too thin. Therefore the oxide can only be scaled down to about 0.8 nm thick. Since transistors in 2001 already have a 2 nm thick oxide, this value is very close. One potential way around this problem is to find a new material for the gate oxide with a much higher dielectric constant. A number of materials including Ta_2O_5 have are being researched but to date the results are not adequate for manufacture.

A third major problem is that the fabrication process uses optical lithography which is a parallel process. The Rayleigh resolution criteria for optics states that the linewidth that can be achieved is proportional to the wavelength of the light and inversely proportional to the numerical aperture of the lens in the optical system. For the 100 nm gatelength transistors, a wavelength of light of 248 nm is used along with photosensitive resists which become more soluable when exposed to photons and can be selectively disolved away. By shining the light through a mask with transparent and opaque regions, the pattern on the mask can be replicated in the resist and this can be etched into the underlying material. There are only a limited number of known radiation sources with wavelengths below 248 nm and the appropriate lens and resists either do not exist or require substantial development before they can be used. While electrons can be used, almost all electron beam lithography systems are serial in nature and therefore much slower that the parallel lithographic processes. The serial processes therefore cannot be used on the present 300 mm wafers for maufacture. The radiaiton at these wavelengths is close to X-rays and therefore the photons are highly energetic. The lenses in the systems cost about $10 million (US dollars) and must be replaced annually as the energetic radiation damages the lens materials. The reduction in wavelength is not guranteed. One possible method is to use phase shifting technology. By incorporating sections on the mask which rotate the phase of the photons, the light interferes and it is possible to beat the Rayleigh resolution criteria. Intel have recently demonstrated 30 nm gatelength transistors and CMOS SRAM cells using such phase shifting technology with 248 nm wavelength resists and optical lithography systems. This at least demonstrates that optical lithography with tricks can produce the required linewidths to at least 2011 although other problems such as layer overlay and mask testing (metrology) have yet to be resolved.

A corollary to Moore's law is that to increase the yield and therefore reduce the cost per transistor on a chip by scaling down the transistor size, the cost of the semiconductor fabrication equipment doubles every four years. While the cost of a DRAM bit is now 7.6 microcent / bit and of a transistor on an MPU is now 123 microcent / transistor, the cost of semiconductor fabrication plants announced at the start of 2001 is about $2.5 US billion (Figure 3). At some point these plants may become too expensive and no return for the investment can be achieved by sales. Many in the the semiconductor industry therefore predict that the economics will end Moore's law before any physical limitation does (Figure 3).

3. Single Electron Devices

While CMOS transistors now have 100 nm gate-lengths, the number of electrons which are used in a switching operation are still tens of thousands. If this could be reduced to a situation where only one electron is used then the energy required to switch should be much lower. This is the basic philosophy of single electron transistors (SETs). There are a number of different types of SETs. The original type rely on Coulomb blockade while a second type are literally miniature flash memory where the addition of a single electron to the gate-memory node results in a large change to the current in the measuring transistor.

3.1 Coulomb blockade

Coulomb blockade involves an small island of charge situation between two electrodes. If the island is small enough and has N electrons then an energy gap opens up between the energy of the last (N^{th}) electron and the first empty electron state (N+1). This gap equals the square of the electron charge, e^2, divided by the capacitance of the island, C. Therefore if the is-

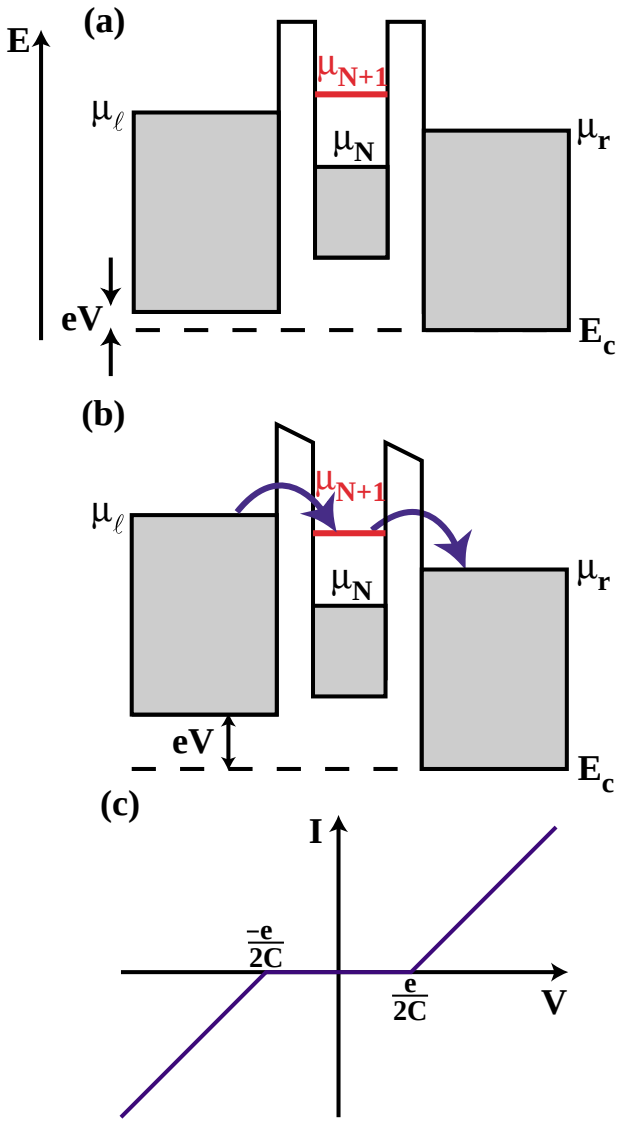


Figure 4:-(a) The Coulomb blockade regime. An energy gap forms between the chemical potential, E of the N^{th} and the $N+1^{th}$ electrons. (b) Single electron tunnelling (c) The current voltage characteristics.

land is small enough so that this energy gap is larger than the thermal energy in the system (k_BT where k_B is Boltzmann's constant and T is the temperature) then electrons cannot quantum mechanically tunnel through the system since the only free states that electrons may tunnel onto the island are above the energy of the electrons in the electrodes (Figure 4). If, however, a gate is used to electrostatically move the islands energy states with respect to the electrodes then the $N+1^{th}$ electron free state can be moved below that of the left electrode and electrons can quantum mechanically tunnel through the island one at a time (Figure 4(b)), hence the name single electron transistor. Therefore the current-voltage characteristics have zero current until the applied gate voltage corresponds to ±e/2C (Figure 4(c)).

A large number of demonstrators of such transitors have been fabricated in a number of different material systems. For room temperature operation of such devices, the island must be below about 10 nm in diameter. Numerous SETs using silicon and silicon dioxide which operate at room temperature have been demonstrate. SETs are more likely to be used for memory applications because they have no gain (i.e. amplification). It is therefore difficult to create logic circuits where the gain in a transistor or logic device overcomes the losses in the circuit and interconnects (i.e. resistance). While a number of logic architectures have been proposed and a few have been demonstrated, it is questionable about the scalability of the circuits to the levels of present CMOS MPUs (section 1).

3.2 Miniature Flash Memory

The second type of SET memory is really just a miniature version of the conventional CMOS flash memory which is found in mobile phones and MPEG music stick players, for example. The structure is that shown in Table I where the addition of a single electron to the memory node results in a substantial change to the electron current through the transistor channel. One potential problem of this approach is the robustness of the memory node to stray charge and fluctuations since one electron is enough to switch the device between memory states.

A second approach to this concept is to use a number of Si nanocrystals as nodes in the oxide rather than one. This approach has the advantage that it is more robust to single electron fluctuations in the system.

3.3 Yano Type Memory

The next type of SET memory is that demonstrated by Yano at the Hitachi Central Research Laboratories. It involves the fabrication using standard CMOS fabrication lines of two crossed poly-Si wires. The poly-Si consists of small grains of single-crystal silicon with grain boundaries between. This type of memory device uses the grains as memory nodes. One major problem is that conduction in the channel is the result of a percolation path through a large number of poly-Si grains. This is a random process and is difficult to control. Therefore the ability to mass produce such memory devices with the required control of properties may be the major problem with this apporach. Hitachi has demonstrated a 128 Mbyte memory chip with the technology although only half the devices operated.

3.4 Comparisons of Single Electron Devices

Table 1 summaries the experimental results in SETs with the present production memory of DRAM and flash produced using CMOS processing lines. Some of the performance is comparable to present CMOS DRAM and flash although noise because of the single electron nature is still a major barrier to manufacturable devices.

4. Quantum Mechanical Tunnel Devices

Quantum mechnanical tunnel devices rely on the ability of electrons to tunnel through thin barriers when the electron wave function can penetrate through the barrier. The transmission coefficient, T, for an electron of energy, E, impinging on a single barrier of height H and thickness b is given by

$$T \approx \frac{16E}{H} e^{-2\alpha b}$$

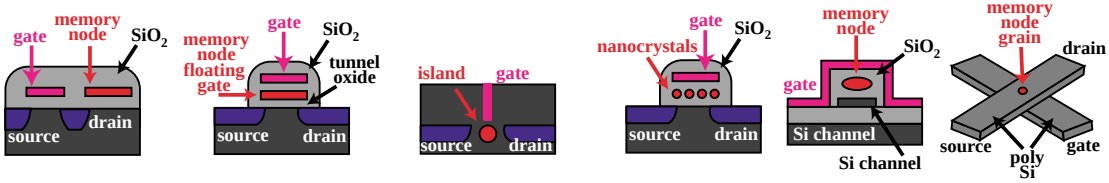
	Conventional Memory		Quantum Dot Memory			
	DRAM	Flash	SET	Nano-flash		Yano-type
				Multidot	Single dot	
device structure						
read time	~6 ns	~6 ns	1 ns	~10 ns	~10 ns	~20 µs
write time	~6 ns	1 ms	1 ns	~100 ns	<1 µs	~10 µs
erase time	< 1ns	~ 1 ms	< 1 ns	~1 ms	<1 ms	~10 µs
retention time	250 ms	~10 years	~ 1s	~1 week	~5 s	~1 day
endurance cycles	infinite	10^6	infinite	10^9	10^9	10^7
operating voltage	1.5 V	10 V	1 V	5 V	10 V	15 V
voltage for state inversion	0.2 V	3.3 V	< 0.1 V	0.65 V	0.1 V	0.5 V
electron number to write bit	10^4	250	1 (excluding no to change gate potential)	10^3	1 (excluding no to change gate potential)	2 (excluding no to change gate potential)
cell size	8.5 F^2/bit	~9F^2/bit	9-12 F^2/bit	9F^2/bit	9F^2/bit	2F^2/bit

Table 1:- A comparison of the performance of different SET and conventional memory technologies

The barrier thckness must therefore be below about 10 nm to allow an adequate amount of tunnelling current. There are two main types of devices which use quantum mechanical tunnelling to produce a negative differential resistance (NDR) current-voltage characteristic which may be used in a number of circuit applications.

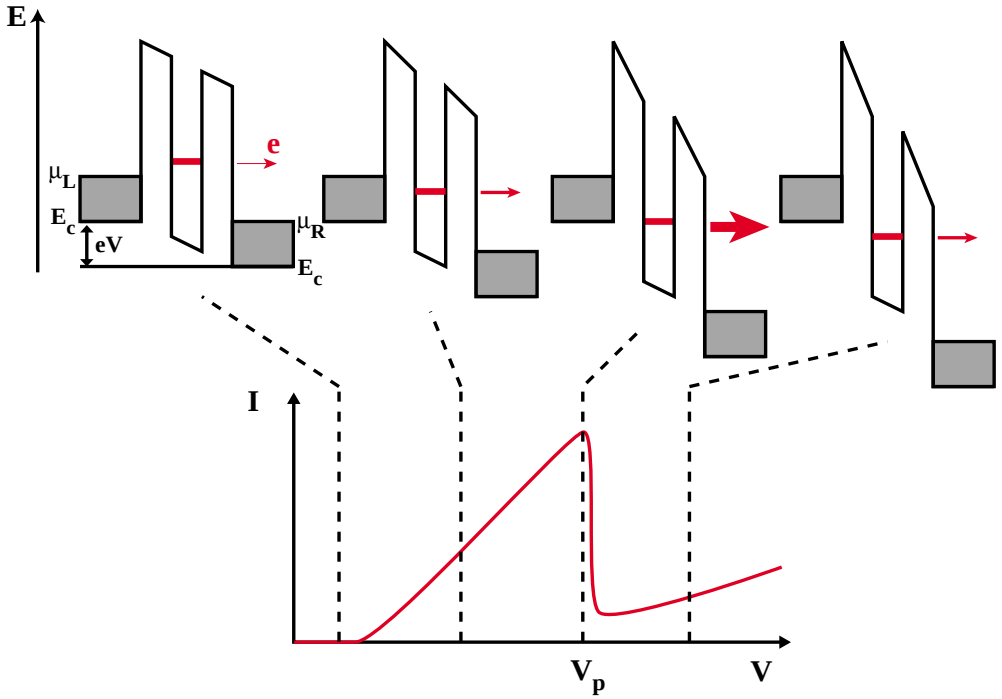


Figure 5:- A schematic diagram of the operation of a resonant tunneling diode. μ_r is the chemical potential for the right electrode and μ_l for the left - electron fills states from the top of the conduction band, E_c, to the chemical potential. The grey areas are filled electrons states in the heavily doped electrodes which provide reservoirs of electrons for tunnelling. Only when the subband in the central quantum well has the same energy as an electron in one of the electrodes can electrons tunnel through the system to the other electrode.

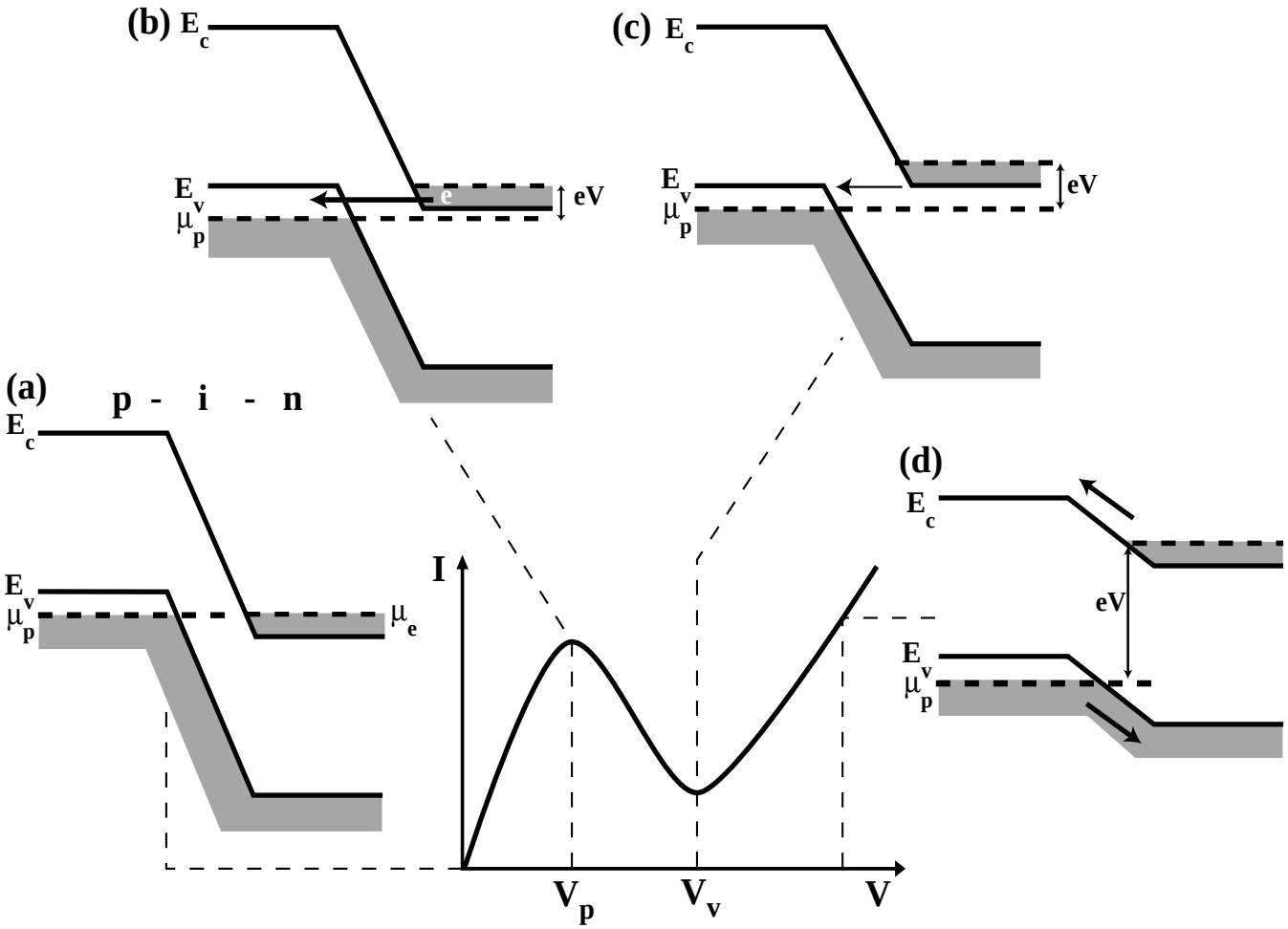


Figure 6:- The operation of the Esaki interband diode which has p-i-n layers. μ_p is the chemical potential in the p-region and μ_n is the chemical potential in the n-region. Grey areas correspond to states filled with electrons (in the valence band this corresponds to states without holes). (a) At zero bias electrons from the conduction band of the n-region cannot tunnel to the valence band because there are no hole states at the same energy. (b) A peak in the current voltage occurs when the reservoir of electrons in the conduction band of the n-region has the maximum overlap in energy with the reservoir of holes in the valence band. (c) The valley in the I-V occurs when the electron reservoir in the n-region becomes higher in energy than all the hole states in the valence band of the p-region. (d) Once the n- and p- states are misaligned, current can only flow due to second order processes by scattering and tunneling through defect states and deep levels in the bandgap and thermally excited processes.

4.1 Resonant Tunneling Diodes (RTDs)

The resonant tunnelling diodes involves a device with two electrodes with two tunnel barriers between the electrodes (Figure 5). The quantum well which is created by the confinement of the electron wave function between the two barriers produces a discrete set of allow electron energy states in the quantum well. Only when an electron from the electrode has an energy which corresponds to the allow state in the quantum well can it quantum mehanically tunnel through the two barriers and quantum well and reach the right electrode (Figure 5). Therefore since each electrde has a reservoir of electron states above the conduction band but below the chemical potential in the electrode (μ_L for the left electrode and μ_r for the right) a peak forms in the current when the two are aligned after the application of a voltage between the electrodes, V which corresponds to lifting one electrode by eV with respect to the other electrode. Once a further voltage is applied, the electrons do not have allowed states in the quantum well to tunnel through and the current drops. If there is a second subband then a second current peak will form when the reservoir of electrons has the same energy as the second subband.

The best results from RTD devices are in the III-V system and in particular the InP / InGaAs / InAlAs heterosystem which include a large number of circuit demonstrators. Silicon is, however, the dominant material in electronics and is substantially cheaper to manufacture. There has therefore been interest in reproducing the III-V based results in a silicon system. Silicon based diodes have had much poorer peak current densities and peak-to-valley current ratios (the two most important parameters of a RTD for circuit applications) although recent Si/SiGe have improved to the stage of having good enough performance to be used in circuits. Typical RTD performance is summarised in table 2. The RTD is the fastest transit time device and an InAs / AlGaAs oscillator operating at 712 GHz with a 0.3 μW output power has been demonstrated. This power is at present too low for many applications but there are very few other devices which can operate at this speed.

4.2 Esaki Diodes

While the RTD uses the one type of charge carrier (i.e.

Parameter	High speed InP RTD logic	Predicted high speed logic	Low power InP RTD memory	Predicted low power memory	Nanometer scaled RTDs	SiGe RTD	SiGe interband tunnel diode
Peak to valley current ratio	4	3	2	3	3	3.3	5.45
Peak current density (A/cm^2)	40k	10k	0.16	0.1	10k	25k	8k
Minimum feature size	2 μm	200 nm	500 nm	200 nm	50 nm	14 μm	14 μm
Peak Voltage	0.35	0.16	0.20	0.20	0.20	1.5	0.28
Maximum clocking frequency	12.5 GHz	6.25 GHz	592 kHz	56.8 MHz	6.25 GHz	no circuit	no circuit
RTD time constant	0.02 ns	0.04 ns	422 ns	4.4 ns	0.04 ns	~0.02 ns	~0.31 ns
Reference	Dortmund 1998	Dortmund 1998	Raytheon 1998	Dortmund 1998	Dortmund 1998	Cambridge 2001	Stuttgart 2000

Table 2:- Comparisons of different RTD parameters and predictions from circuit simulations of required device parameters for optimised performance.

either electrons or holes), the Esaki diode is an interband device where electrons tunnel from the conduction band to hole states in the valence band. The operation is described in detail in Figure 6. Esaki diodes have been around since Leo Esaki discovered and analysed the effect in 1958 using Ge diodes. All the devices produced in the early 1960s were fabricated by diffusing one species of doping to create the p-i-n diode. This results in poor uniformity and poor control of the performance. Recent epitaxial growth and the inclusion of SiGe intrinsic layers has allowed much better performance results (table 2) and good uniformity.

4.3 Tunnel Diode Circuits

To date, tunnel diodes are one of the few quantum devices which has successfully demonstrated room temperature circuit operation. InP based tunnelling static RAM memory cells using two RTDs and one FET have demonstrated 50 nW power dissipation with a 150 μm^2 cell footprint. This power dissipation is a factor of two hundred lower than an equivalent six transistor GaAs static RAM which has a footprint three times larger. The RTD can therefore be used to reduce power consumption, circuit complexity and the real estate useage on a chip.

In logic circuits there have also been a substantial number of demonstrators. Generic logic which can change functionality between NAND, NOR or NOT just by changing the relative sizes of the transistors in the circuit has demonstrated 12 GHz circuit operation using 20 μm minimum feature sizes. As a comparison, CMOS requires 180 nm minimum features to achieve 1 GHz operation. Therefore the scaling of this technology to nanometer sizes will result in substantially faster performance that CMOS. Double current peak RTDs have also been used to demonstrate multi-valued logic which allows a substantial reduction in device count for logic chips. Monolithic 4-bit analogue to digital converters have shown 2 gigabit per second sampling and static binary frequency dividers operating at 40 GHz have all been demonstrated. There are also circuit demonstrators of photodetectors, clock circuits and shift registers which all operate at far faster speeds than equivalent CMOS circuits and with reduced component counts.

RTD devices have therefore be used to demonstrate a large number of circuits which either provide higher speed or lower power dissipation than conventional CMOS or III-V FET technologies. At present all these demonstrators are in III-V technology and no silicon RTD or Esaki diode circuits have been demonstrated. One major problem of the III-V system is that since the tunneling current depends exponentially on the barrier thickness, only two atoms difference across a wafer can have a substantial change in the tunneling current. This limits the yield at present of the circuits and must be improved if RTD circuits are to compete with other technologies in the market. In addition for high frequency oscillations, the power must be increased to able to be used. One method that is being used is to build arrays of RTDs and this may allow oscillators to operate above 600 GHz with mWs of power.

5. Spin Nanoelectronics or Spintronics

While most nanoelectronic devices rely on charges being transferred around a device or circuit, the electron spin may also be used for memory storage or logical computing. The hard discs in computers are already a good example of how spins may be used to store information. Most new spintronic devices rely on spin injection rather than the measurement of a ferromagnetically

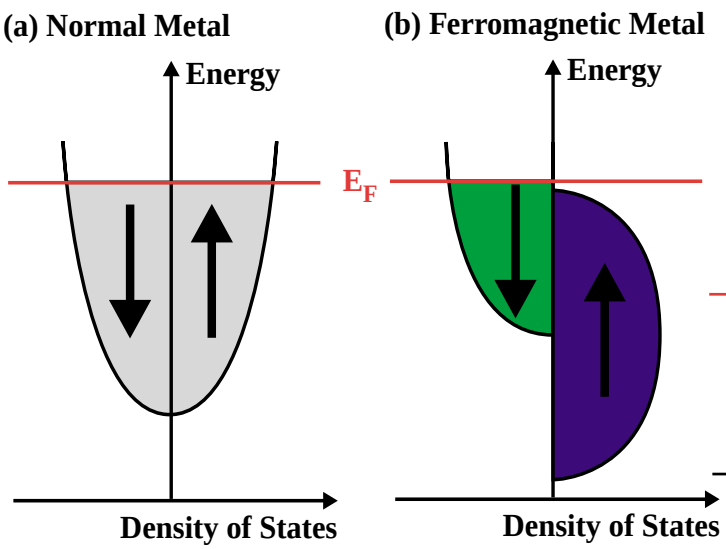


Figure 7:- The difference between a metal and a ferromagnet. (a) In a metal, the number of up spins equals the number of down spins around the Fermi surface. (b) In a ferromagnet, there is a larger number of one type of spin around the Fermi surface.

polarised particle.

The basic operation of a spin device relys on the ferromagentic material having a larger number of one spin close to the Fermi energy or surface as shown in Figure 7. If the ferromagnetic is put into a circuit then only (e.g. from Figure 7) the down spins can be transported through the system. Therefore switches can be built up by designing structures with different ferromagnetic and normal metallic layers.

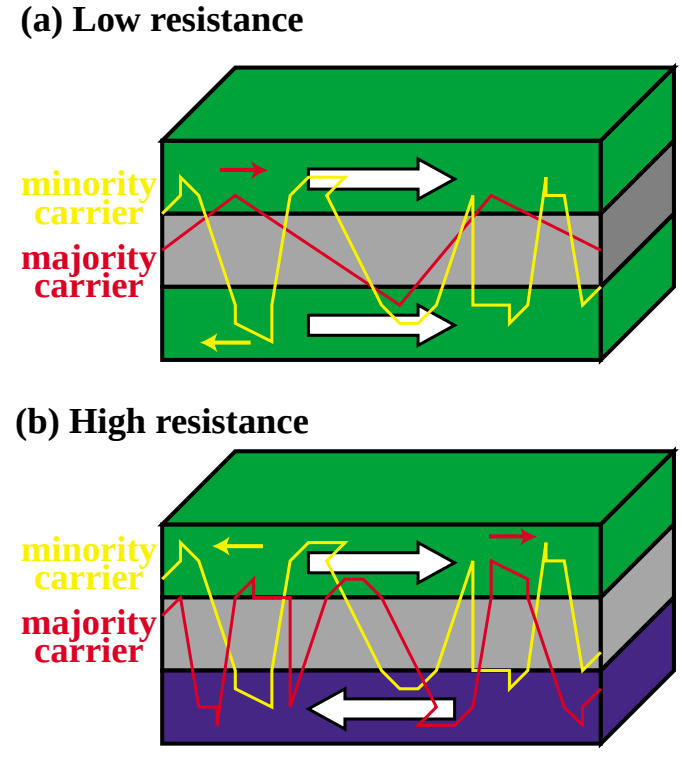


Figure 8:- The concept of the spin valve switch. (a) If the two ferromagnetic layers have the spins aligned then the majority-spin carrier has a low resistance path through the device. (b) If the ferromagnetic layers have opposite spins then both spin carriers see a higher resistance path through the device.

Figure 8 shows the basic operation of a spin valve. By having two ferromagnetic layers placed on either side of a metal, if both ferromagnets are polarised in the same direction then the majority spins in the system will have a low resistance path through the device. If, however, the ferromagnets have opposite polarisation then both the majority and minority spins find a high resistance path. Therefore the resistance of the system corresponds to the on and off states required for switching. The application of a magnetic field provides giant magnetoresistance (GMR) which is now the basis of all the devices for reading hard discs.

A similar effect can be used to produce a spin tunnel device. If two ferromagentic contacts are placed either side of a metal (or semiconductor) then electrons of one particular spin will be injected into the metallic layer if a bias is applied to the system (Figure 9(a)). If the second contact to the metal also has the same spin polarisation as the first then the electrons can pass into the second contact and the circuit has a low resistance. If,

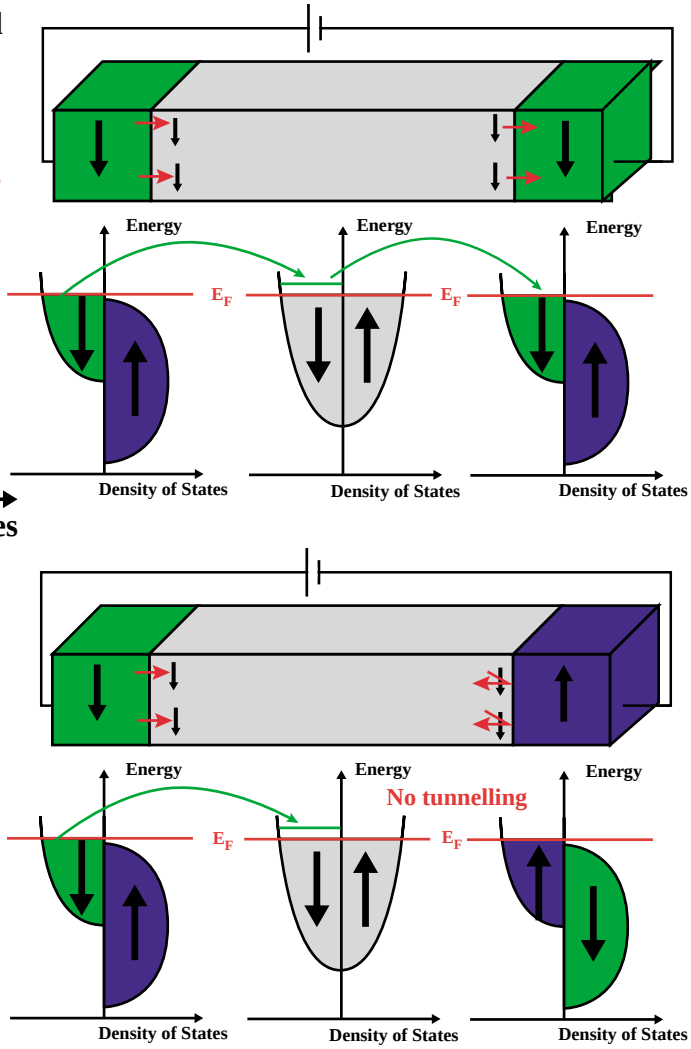


Figure 9:- The basic priciple behind a spin tunnel junction device. (a) If the two contacts to a metal or semiconductor have the same spin polarisation then electrons can transfer around the circuit and a low resistance is measured. (b) If, however, the contacts have opposite spin polarisation then the electrons cannot enter the second contact as there are no free states near the Fermi energy.

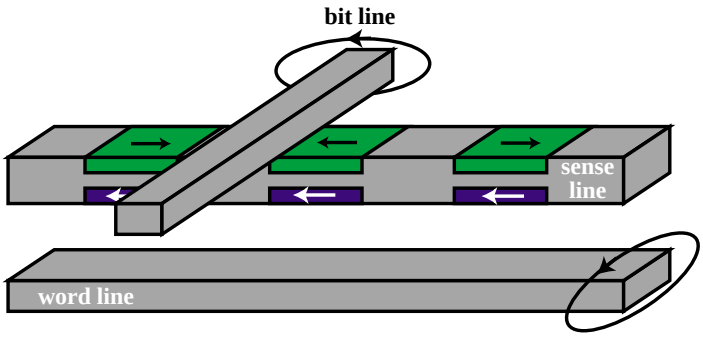


Figure 10:- A schematic representation of a MRAM device where a GMR element is magnetised by the magnetic fields created by currents in the word and bit lines.

however, the second contact is polarised opposite to that of the first contact (Figure 9(b)), no free states exists for the electrons to tunnel into close to the Fermi surface and no current can flow in the system. If the metal is replaced by a semiconductor with a large Rashba term then the application of an electric field to the semiconductor will rotate the spin of the electrons as they are transported across the semiconductor layer and the device can be switched on and off using the electric field from a gate. These effects can be used to form a number of different spintronic devices including switches and spin polarised filters.

The most successful magnet nanoelectronic device to date is the magentic random access memory (MRAM). This is shown in Figure 10 and corresponds to a GMR element which is written and read using the magnetic fields generated by currents in the word and bit lines. It is therefore a very compact memory along with being non-volatile and radiation hard. A substantial number of demonstrators have been produced with sizes above 1 MB and access times of about 50 ns using 0.25 μm dimensions. Such devices can be easily scaled to smaller sizes where the performance is predicted to increase substantially.

To date there have only been a few demonstrations of spin injection into semiconductors, in particular a spin polarised light emitting diode (LED) has been demonstrated where the output photons are polarised depending on the spin of the electron and holes in the system. Significant spin injection can only be achieved from magnetic semiconductors into normal semiconductors. Results from metallic ferromagnets into semiconductors has to date been very poor.

Spintronics is therefore at a relatively mature stage of research and potentially may allow both memory and logic functions in much smaller and robust systems than Si CMOS. In paricular, the magnetic systems may be able to be designed so that the electrical resistance of small interconnects does not impede the ultimate performance.

6. Fabrication Technology

At present all CMOS wafers are fabricated using a top-down approach where deep ultraviolet photons are shone through a patterned mask made of glass (or quartz) and chrome. Where the light is transmitted through the mask, it reacts with an optical resist on the wafer and this area may be disolved in a solvent to leave the mask pattern in the resist. This pattern is then transfered to the underlying substrate by etching. By blanket deposition of insulators or metallic layers along with lithography and etching, complex circuits are patterned onto chips.

The smallest feature which may be produced using this lithographic technique is given by the Rayleigh resolution criteria,

$$\frac{k\,\lambda}{NA}$$

where k is a constant, λ is the wavelength of light and NA is the numerical aperture of the lens in the system. At present 248 nm KrF excimer laser sources are predominantly used and 197 nm ArF lasers are available. To produce smaller features, the wavelength must be reduced and the numerical aperture of the lens in the optical system must be increased. By using phase shifting technology (i.e. parts of the mask also rotate the polarisation of the light) interference effects can be achieved which produce smaller features than the Rayleigh criteria. The major problem is that such phase shifting technology is extremely complicated to design and difficult to fabricate the appropriate masks.

Other lithographic systems at the research stages include extreme ultraviolet sources, X-ray and electron beams. In all cases these systems use smaller wavelengths to increase resolution. It must be stated that sources, masks, optics and resists are all required for a lithographic fabrication process. These are not always avaliable, for instance X-ray optics are extremely poor at present. Another problem is that the speed of exposure must be high enough to allow mass production (Figure 11). All early electron beam lithography systems used a single focused beam with a spot of a few nanometres. The pattern is then built up by rastering the beam over the chip area but for large complex chips this is very slow and inpracticable. Vector beam systems have been developed where a much larger standard feature (e.g. a rectangle or a repeated feature) can be exposed at one time substantially increasing the throughput.

Imprint lithography is a recent development where a master mold is imprinting into a polymer resist through pressure and sometimes additional heating. When the mold is removed the polymer typically has thick and thin parts in the approriate

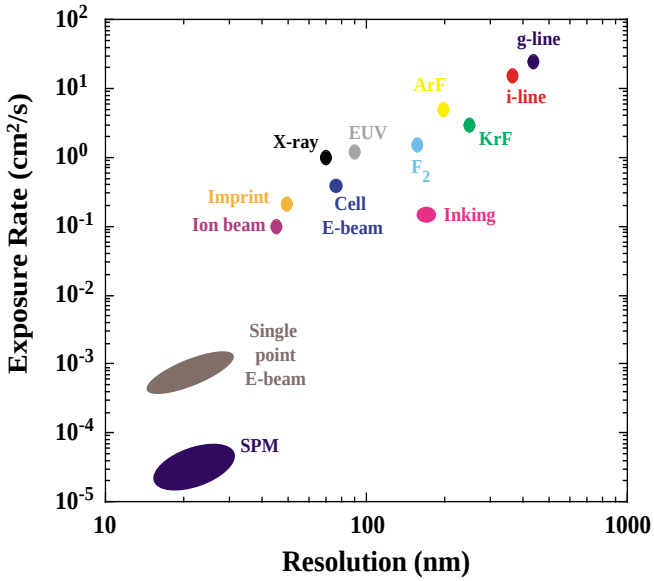


Figure 11:- A comparison of the exposure rate and hence throughput of different lithographic techniques compared to the resoluation of the techniques. g-line, i-line, KrF, ArF and F_2 are all sources for optical lithography. EUV is extreme ultraviolet.

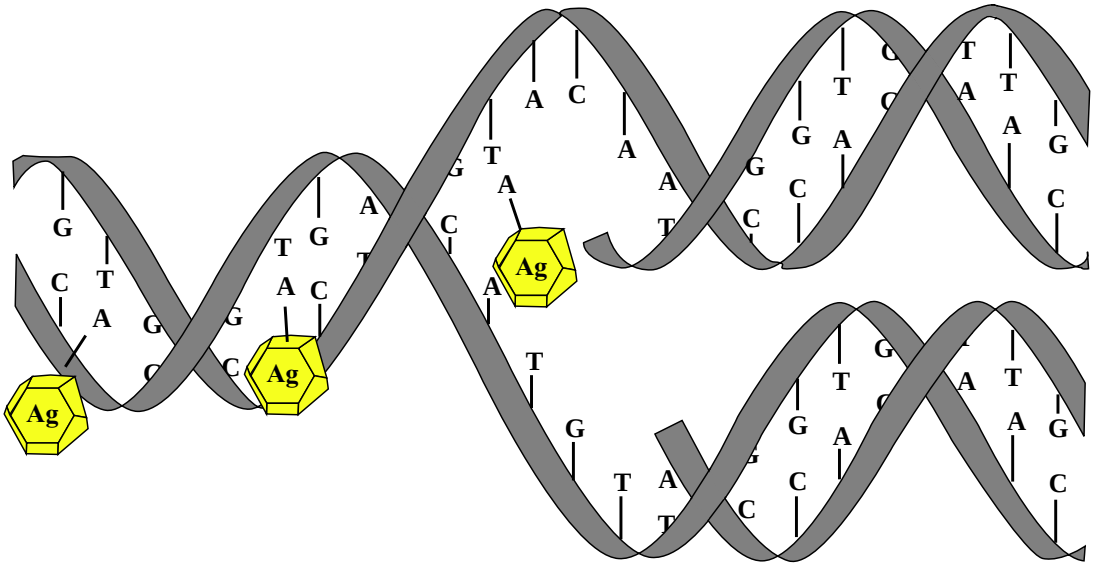


Figure 12:- A schematic diagram of DNA templating where silver particles with an attached adenine (A) selectively bonds to thymine (T). Guanine (G) bonds to cytosine (C).

pattern which after a descum to remove the thin parts of the polymer, may be transferred to the underlying substrate using etching. This technique can potentially be applied on a wafer scale and has demonstrate small features, the present limit of which appears to be more related to the ability to fabricate small features on the master mold rather than through the imprint technique itself. Features as small as 10 nm have been demonstrated although for high throughput, features are typically 50 nm or above (Figure 11). The major problem with such a technique is that it is difficult to align and calibrate the mold to different layers. As a first stage lithographic process this is adequate but for large scale nanofabrication, many different layers are required. The optical techiques easily allow scaling of the pattern to match to other layers through focusing but a mold is fixed. An additional potential problem is that the mechanical contact is prone to dirt and defects and may be damaged over time. A derivative of the imprint technique is inking where the polymer or ink is applied to the mold so that only raised features have any ink on them. The pattern is then transferred to the substrate through contact. These techniques show good promise to scaling to features of 10s of nm but may ultimately be limited by their inability to be flexible to other layers in real structures.

One of the major problems of these top-down lithographic approaches is that they require very clean environments otherwise dust and particulates can mask part of the exposed area. It is the lithography that drives the cleanliness aspects in semiconductor production. For high yields and hence low costs, clean rooms must have particulate densities extremely low and at sizes much smaller than the lithographic minimum feature size. The cost of lithography is also increasing at an exponential rate as the feature size decreases and may eventually stop the progress of microelectronics.

7. Self-Assembly

The top down approach of lithography described in the previous section is one of the expensive stages in fabrication of CMOS circuits and requires high levels of cleanliness to produce high yields. An alternative concept is to use a bottom up technique where the devices self-assemble into circuits. This is at a relatively early stage of research but if successful should be substantially cheaper and more robust to “dirty” environments than lithographical techniques.

Much of the early self-assembly was in the area of metallic and semiconductor dots (small particles) which have improved optoelectronic properties for detectors, LEDs and lasers. Most of these systems are not appropriate for electronic systems as no interconnects are available. In the last ten years a number of biological, organic and inorganic systems have been shown to be able to either self-assemble or to be used as a template for other functional groups.

One of the best examples was the use of DNA to selectively attach conducting particles (in the first instance silver particles). Since the DNA helix is held together by base pairs made up of 4 different bases each of which can only bond to a reciprocal base through the Watson-Crick base-pairing interactions, single stranded or double helix DNA can be designed to either selectively attach to other DNA molecules or selectively attach functional units (Figure 12). An adenine (A) on one strand pairs preferentially with a thymine (T) on another strand while

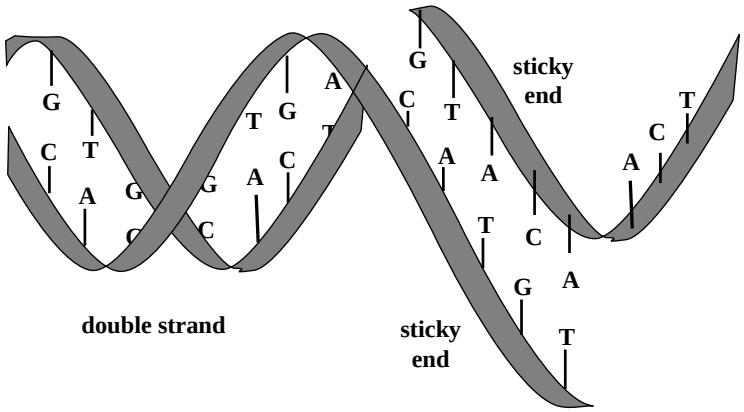


Figure 13:- The ideal behind sticky ends which may be used to selectively attach functional DNA double or single strands to other DNA or substrates or electrodes with appropriate sticky ends.

a guanine (G) pairs with cystosine (C). Single strands of DNA at the end of a double helix with a specific set of bases called a sticky end allows this end to only attach to another sticky end with the appropriate bases thereby potentially allowing small, functional, hybrid DNA molecules to be self-assembled into much larger circuits (Figure 13).

Numerous other examples of self-assembly exists including Langmuir-Blogett films, S-bacterial layers, self-assembled monolayers (SAMs) and anti-body-antigen recognition. All these techniques are still at a very early stage and substantial research is required to get any of the techniques to a manufacturing level.

8. Molecular Nanoelectronics

The first discussion of molecular nanoelectronics was the proposal by Aviram and Ratner in 1974 to produce a rectifier from organic molecules. The first example of a single molecular electronic device did not appear until 1990, the major problem relating to the difficulty of making individual electrical contacts to molecules which may only be a few nm in size. The development of the scanning tunnelling microscope (STM) basically enabled the first measurements in this field to be started and has remained one of the major tools in electrically characterising single molecules.

Some of the first demonstrations of electronic properties of single molecules by Purdue University included Coulomb blockade and Coulomb staircase when a STM tip measured the conduction through gold nanoparticles self assembled with SAMs. A second set of experiments at IBM, Zurich demonstarted a STM tip deforming a C_{60} bucky ball. The resulting mechanical defomation modifies the resonance tunnelling bands of the molecule and produces electromechanical amplification. Hitachi demonstrated a molecular abicus where a STM tip was used to move 0.25 nm high C_{60} molecules along monoatomic steps and then counted by imaging with the STM tip. While all these devices demonstrate functionality which may be used in circuits, the scaling to the level of present CMOS circuits would be impossible.

Another approach to molecular electronics is the use of organic molecules. A collaboration between Yale and South Carolina Universities demonstrated the conduction through a benzene molecule attached to two gold electrodes using thiol groups to bind the molecule to the gold. Benezene rings have delocalised π-electrons out of the plane of the molecule through which electrons can be transported when an appropriate bias is applied across the molecule. Carbon-carbon double and triple bonds provide similar orbitals out of the plane and therefore combinations of these polyphenylene molecules create conducting wires now known as Tour wires. Molecules could then be designed with conducting sections along with barriers created by methyl groups which do not have delocalised electron orbitals. Therefore organic chemists can design molecules where the high occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) through which electronic transport is controlled can be manipulated in a similar fashion to the band structure engineering in semiconductors. Molecular RTDs operating at room temperature have been demonstrated. The low temperature properties of some of the RTD designs demonstrate PVCRs of over 1000 at low temperatures while the room temperature properties require improvement before they can be used in circuits.

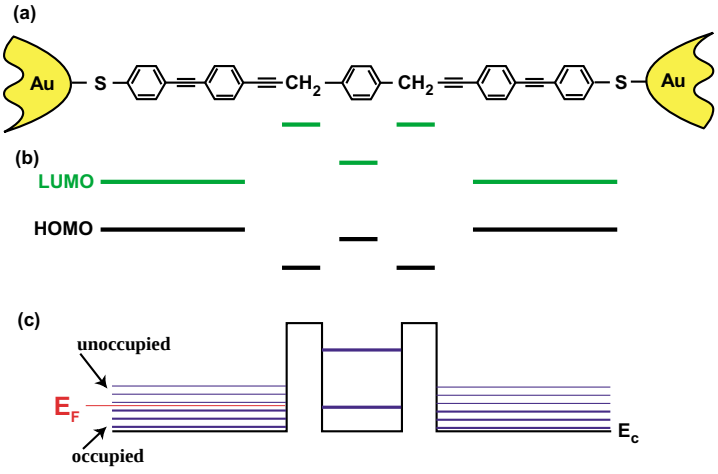


Figure 14:- (a) The chemical structrue of an organic RTD. (b) The HOMO and LUMO levels responsible for the electronic transport through the molecule. (c) The equivalent band structure for a semiconductor RTD.

Work at the Mitre Corporation has investigated possible architectures using organic molecules. A number of designs have been proposed based on diode logic using the Tour wires and diodes. AND, OR and XOR gate designs are given along with an adder. The major problem with such organic systems is that the conductivity is relatively poor through the interconnecting Tour wires. The RC time constants of most of the devices is likely to be relatively large and some rough "back-of-the-envelope" calculations suggest that typical speeds will be up to 10s of MHz rather than GHz as presently available with CMOS. The major limitation is the conductivity and unless better conductors or architectures for which the performance does not depend on resistance can be found, the organic systems will always be much slower than silicon.

Carbon nanotubes have also been used to conduct electrons and demonstrate reasonable conductivities for their size. A large number of groups have published results in the field. Both single and multiple walled tubes can be created which may potentially reduce resistance. Conductivities as high as 2000 Sm^{-1} which corresponds to a resistance per μm of 200 Ω have been measured, substantially better than organic polyphenylene molecules. A transistor was demonstrated although the gate was a silicon substrate and the carbon nanotube had been placed across two metal electrodes fabricated on top of the thermally oxidised silicon substrate. Metal-nanotube rectifiers have also been demonstrated. To date, no switch or three terminal device has been produced which is a basic requirement for most logical architectures.

Numerous groups have demonstrated measurements on nanocrystals, many have used cadmium selenide. CdSe nanocrystals can be prepared with sizes down to about 2 nm and with attached linker molecules which may bind to numerous surfaces including gold. A single electron transistor has been demonstrated where in a similar experiment to the carbon nanotube transistor, an oxidised silicon substrate is used as a gate to a CdSe nanocrystal placed between two gold electrodes on the SiO_2 surface.

DNA self-assembly was discussed in the precious section.

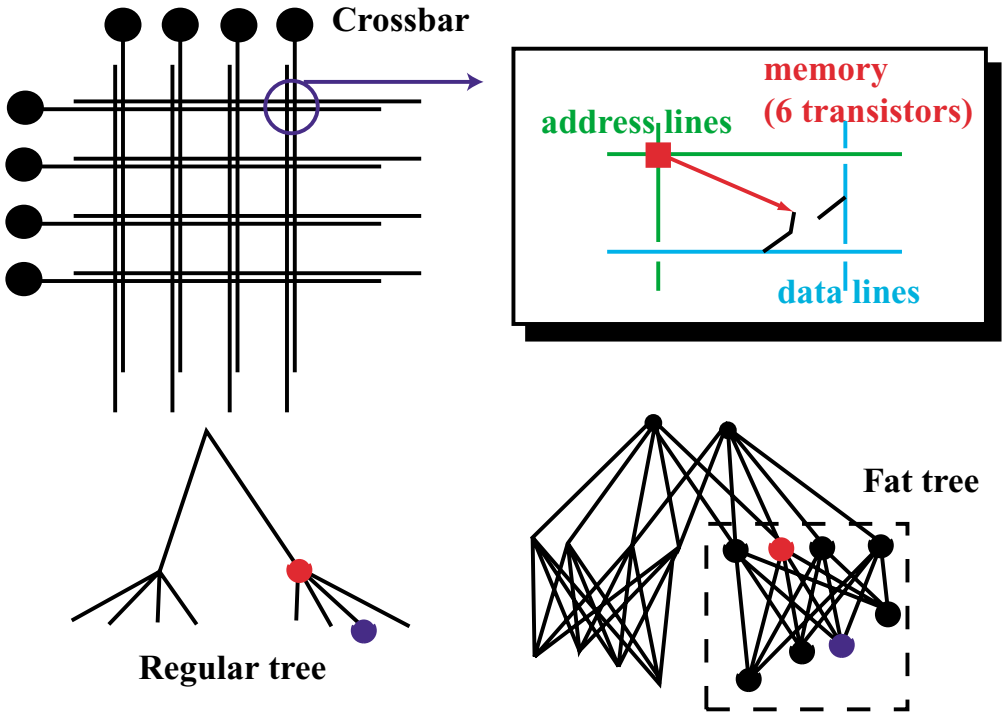


Figure 15:- Schematic diagrams of the Teramac concept. Normal architectures use a regular tree structure where there is only one or a few paths between any points in the circuit. The Teramac architecture supplies much larger levels of connectivity through a fat tree structure so that if one interconnect fails, another can be used. The overall structure employs a crossbar addressing scheme with numerous connections to each memory node.

There have also been a substantial number of paper measuring the conductivity of different typres of DNA. Results range from highly insulating behaviour to semiconducting and even metallic behaviour. The results come from many different structures and types of DNA and as yet no systematic investigation of varying specific parameters and the resulting changes in conductivity has been present. It would appear that the only consistent results are from DNA with self-assembled metallic nanoparticles. The best results to date correspond to Pd nanoparticles attached to a DNA strand which has demonstrated 100 Sm^{-1} conductivity.

There are a large number of additional ideas presently around in the literature. To date, however, there has not been any demonstration of an all molecular transistor or any structure with gain. As will be discussed in the next section, gain is crucial to the distribution of information over large circuits. For a switch or transistor, a voltage corresponding to the difference in HOMO and LUMO energies will be required to switch conduction on or off. In most of the proposed devices this corresponds to about 1 eV and hence 1 V must be applied to switch a device on or off. If the resistance of the molecule is about 200 Ω and you have 10^{10} molecules then the circuit will consume 10^{8} W! Molecular nanoelectronics is at a very early stage and potentially offers very cheap self-assembly fabrication routes but a substantial break through is required if the technology is ever to come to fruition.

9. Alternative Architectures

9.1 Fault Tolerant

The main reason for the high cost of CMOS fabrication plants is that the top-down architecture required from lithographic fabrication techniques is not defect tolerant. Therefore any transistor or interconnect which fails on the circuit potentially destroys the whole chip. Some redundancy can be built in but this is expensive and not ideal. Almost all nanoelectronic devices still rely on architectures which have no fault tolerance. This problem is substantially worse for applications specific integrated circuits (ASICS) where each application has a different circuit design unlike the mass repetition of memory or processor manufacture. Testing has become a substantial cost (> 60% of the total chip cost) as the new circuit design must be tested to check that no mistakes in circuit design or mask fabrication has taken place. With up to 20 mask levels in some chips and with up to 40 million transistors in designs, the potential for errors at some point in the design and fabrication process is substantial.

If an architecture could be found for which not all the transistors or interconnects are required to correctly function for the operation of a complete chip then this would substantially reduce costs. Most molecular ideas may require a fault tolerant architecture if they are ever going to be successful.

One of the few defect tolerant architectures which has appeared is the Teramac architecture. This concept is shown in

Figure 15 and relys on the ability to have a large number of interconnects in a system so that some path may always be found around defects and non-functional parts of a circuit. Before the system is run, a map of all the defects is found and the system is configured so that the defects can be circumvented. Therefore the Teramac paradigm is to build a cheap and defective computer, find the defects, configure the resources with software, compile the programme and then run the computer.

The major disadvantage with the Teramac concept is that the biggest problem with both CMOS and molecules is the interconnects. With CMOS as interconnects are reduced in size, the conductivity is constant with the resistance increasing faster than the decrease in capacitance from reduced size. Therefore the RC time constants increase for reduced interconnect size and the speed of the circuit correspondingly is reduced. High quality interconnects are a serious problem in many of the molecular schemes. Therefore using larger interconnect bandwidth to provide defect tolerance is not an option with most of the known nanoelectronic schemes.

9.2 Gainless Architectures

To date, no molecular transistor with gain has been demonstrated. In single electron transistors, for every electron put into the transistor you get almost one electron out of the transistor. These devices therefore have no gain. Gain serves two major purposes in architectures. Firstly it provides isolation between the input and output. Therefore if you have some voltage at the output then that is the answer from the inputed data. If there is no isolation then the logic gate could potentially run backwards so that the output could be defining the input and therefore the whole logic gate oscillates between different values and any output cannot be believed.

The second major reason for gain in a circuit architecture is related to transmiting information around a large circuit. All electrical interconnects have a resistance which corresponds to a loss in the system. Therefore if you send a small current across a large circuit, it will only reach its destination if the losses are small. The way around this is to amplify the signal which requires gain. Many circuits also connect one output from a logic gate to a number of new inputs to many logic gates (fan-out as it is known). The problem with a gainless system is obvious if one considers the case of one electron exiting the first logic gate output. It can only be transmitted through one of the new logic gates. Therefore for fan-out in any circuit, a large signal is required which will be divided between a number of paths in the circuit. Since this is repeated many times in large logic circuits, there must be gain if any signal is to pass through the large number of gates.

For all the known architectures which are used in electronics, gain is therefore an essential requirement.

9.3 Quantum Cellular Automata

The quantum cellular automata (QCA) provides an interconnectless method of information exchange and computation. It therefore has gained a substantial amount of interest. The basic concept relys on four (or five) quantum dots being produced with only two electrons in the system (Figure 16). The Coulomb repulsion of the electrons forces them to occupy diagonally opposite sites and therefore the arrangement dictates whether the cell is in a “1” or “0” state. As long as the electrons are forced to remain in this cell then other cells can be placed beside the first cell. If a field is applied to change the arrangement of the electrons in the first cell then this information is passed to neighbour cells and therefore is transport through the system.

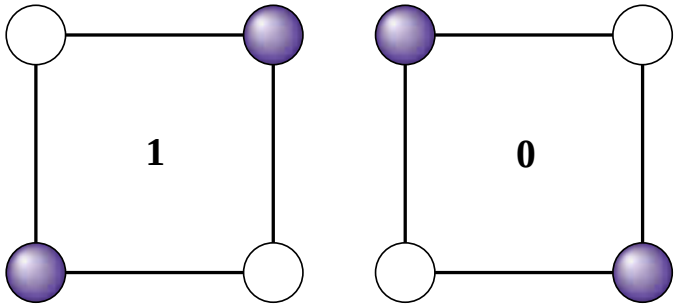


Figure 16:- The two stable states of a quantum cellular atomata cell providing the “1” and “0” states.

By appropriate arrangements, logic calculations can also be achieved. Figure 17 shows a majority voting gate where the output of the gate equals the value of at least two of the inputs which have the same value. This sort of gate will only function with an odd number of inputs so that either “1” or “0” can dominate. The major problem with the QCA architecture can be easily demonstarted when a similar logic gate is constructed with an even number of inputs, half of which are in a “1” state and the remainder in the opposite “0” state. The output is therefore not predictable and the architecture breaks down. The second problem with QCA is that it is also a gainless system and therefore if another cell applies a force to the output then the system may run backwards and the whole circuit becomes unstable and can oscillate between values.

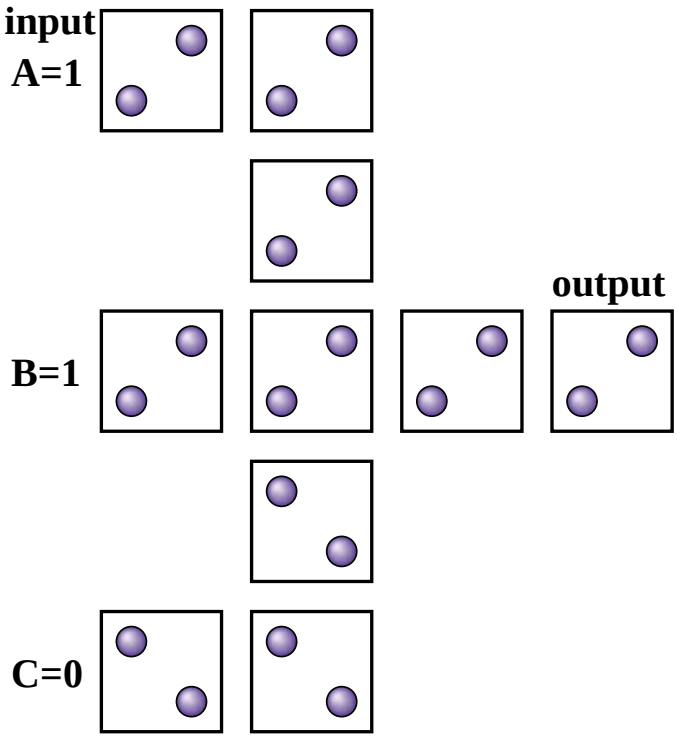


Figure 17:- An example of a QCA logic gate. This is a majority voting gate where the output will equal the majority of the inputs, in the case above the output = 1.

So far only a single cell has been demonstrated experimentally and no logic gates have been demonstrated. Calculations suggest that Si quantum dots of less than 2 nm size are required to produce cells which may operate at room temperature. There is some belief that molecular self-assembled quantum dot systems may be the ideal system to produce QCAs but much work is required. The QCA therefore is an interesting concept as it is an interconnectless architecture but substantial problems suggest that it may not be realisable. Calculations at University College London have also shown that it may be much slower than conventioanl CMOS circuits at relatively small integration densities.

9.4 Quantum Computing

Quantum computing is a radically different architecture which uses the interference properties of entangle quantum mechanical particles to allow each bit of quantum information (called a qubit) in the computer to be intimitely linked to every other qubit in the system. Quantum computing is therefore a massively parallel architecture at a level which is impossible in any classical architecture.

The important pieces of quantum mechanics for quantum computing is the Einstein-Podolsky-Rosen (EPR) paradox and the Bell inequality. If two particles are entangled and taken to the opposite ends of the universe then quantum theory determines that if you measure the properties of one, you automatically then know the properties of the other since they are intimately linked. An example is the spin of a photon where if one has a right-hand polarisation then the other after entanglement must have a left-hand polarisation through quantum theory. More over, the state of one after a measurement automatically defines the state of the other. Before the measurement it is impossible to predict exactly which state each particle may be in. This is as if information has travelled instantaneously across the enormous distance of the universe which is impossible from Einstein's special relatively theory.

A bit is a two state system from a physical point of view. In quantum mechanics, if a bit has two distinguishable states then it may also exists in a superposition of these two states called a qubit. If each of these superpositions are entangled so that there is a further superposition of each qubit, then the state of each qubit depends on the states of every other qubit. Gate operations on these qubits provide a time evolution of the system where every entangle particle is affected and so a massive parallel computation results. This enormous parallelism allows classically numerically intensive problems such as the travelling sales-man problem (known as NP problems in mathematics) to be solved in real time which cannot be solved with 100 years of present supercomputing time. Potential applications are factorisation for cryotography, fast database searching and modelling of quantum systems.

To produce a quantum computer, there are five criteria called the DiVincenzo checklist which must be adheard to:-

1. *A scalable physical system with well characterised qubits.*
2. *Initialisation of simple fiducial states.* This basically is setting the qubits into similar quantum states before any calculation can proceed. An example is setting all the spins in one system to be identical.
3. *Long relevant decoherence times, much longer than the gate operation.* As the decoherence time dictates the length of time the qubits can be entangle without loss of any information, any computation must be finished before the qubits loose information.
4. *A universal set of quantum gates.* For a quantum computer, two types of gate operation are required to produce a universal computer which may be designed and programmed to complete any computation task. These are the single- and two-qubit operations and from these a computer for factorisation, database searching or quantum system simulation may be built and operated. These two gate operations can be achieved by different techniques which depend on the specific two state system used for each qubit.
5. *A qubit specific measurement capability*. Once a calculation is complete, the information must be outputed from the qubits.

At present only a few systems with up to 5 or so qubits have been demonstrated. For most applications such as factorisation of large numbers for cryptography, 30,000 qubits or so are required although even 50 qubits will solve problems quickly which are impossibly slow on classical computers. Ion traps and nuclear magnetic resonance systems have so far demonstrated the largest number of qubit entanglements. For large integration, however, solid state quantum computers potential provide the best platforms. Examples for the two state systems required for qubits include semiconductor quantum dots, superconducting device, the nuclear spins of donors in silicon, the electron spins of donors in SiGe heterostructures and the use of surface accoustic waves with low dimensional structures in GaAs heterostructures.

Quantum computing is presently at an immature experimental stage but if many of the problems can be overcome, the potential power of the technique will allow many numerically intensive calculations to be completed which are presently impossible with classical computational machines.

10. Summary and Future Prospects

The major conclusion of any review of nanoelectronics must be that the silicon transistor through CMOS is the dominant technology and will remain so for the foreseeable future. Only a break in the relentless pursuit of Moore's law in silicon may provide a chance for other technologies to try to compete. Such a break is more likely to be the result of economics rather than technological problems.

RTDs and single electron memories have serious possibilities of niche markets and at least the RTDs are already being sold for specialist high speed digital to analogue converter applications. The spin and magnetic devices, in particular MRAM, have very clear markets and are already close to production. Molecules and self-assembly are much further from the market place and with present results, substantial breakthroughs are required if they are ever going to reach the market place.

A second significant conclusion of any nanoelectronics review will always be that the architectures for nanoelectronics are all the same ones used for CMOS and new architectures are required for most new nanoelectronic concepts and devices if they are to succeed. Quantum computing is such a new architecture but it only provides niche applications and is not applicable to most nanoelectronic systems. It is also a field with much

theory but little experimental demonstration to date. Much of the technology for quantum computing is way beyond present fabrication technology but this may change as the nanoelectronics research field progresses. The realisation of a successful fault-tolerant architecture would bring enormous benefits to all nanoelectronic fields including CMOS and molecular electronics. The research field is therefore ripe for major advances in the future.

11. References

Bennett, C.H. and DiVincenzo, D.P. (2000). Quantum information and computation, *Nature* **404**, 247-255

Compañó, R., Molenkamp, L. and Paul, D.J. (Eds.) (1999). Technology Roadmap for Nanoelectronics, European Commission, Brussels. (http://www.cordis.lu/esprit/ist/fethome.htm)

de Boeck, J. and Gorghs, G. (April 1999). Magnetoelectronics, *Physics World* **12**(4) 27-32

Ellenbogen, J.C. and Love, J.C. (1999). Architectures for molecular electronic computers: 1. Logic structures and an adder built from molecular electronic diodes. *Proceedings of the IEEE* **87**, 386-426

Heath, J.R., Kuekes, P.J., Snider, G.S. and Williams, R.S. (1998). A defect-tolerant computer architecture: opertunities for nanotechnology. *Science* **280**, 1716-1721

Lent, C.S. and Tougaw, P.D. (1997). A device architecture for computing with quantum dots. *Proceedings of the IEEE* **85**, 541-557

Reed, M.A. (1999). Molecular-Scale Electronics. *Proceedings of the IEEE* **87**, 652-658

Semiconductor Industry Association (2000). International Technology Roadmap for Semiconductors 2000 Update, SEMATECH, California.

van der Wagt, J.P.A. (1999). Tunneling-Based SRAM. *Proceedings of the IEEE* **87**, 571-595

Yano, K., Ishii, T., Sano, T., Mine, T., Murai, F., Hashimoto, T., Kobayashi, T., Kure, T. and Seki, K. (1999). Single-Electron Memory for Giga-to-Tera Bit Storage. *Proceedings of the IEEE* **87**, 633-651