FUTURE DEVELOPMENTS IN MICROELECTRONICS AND THEIR IMPLICATIONS FOR COMPUTING

by

Richard Turton
Computer Officer
University of Newcastle Upon Tyne
Newcastle Upon Tyne, ENGLAND

The Twenty-first International Conference on the Unity of the Sciences
Washington, D.C. November 24-30, 1997

1 Introduction

There are several major technological advances that have helped to bring about the rapid growth in information technology, often referred to as the information revolution. The invention of the transistor, improvements in telecommunications and the introduction of 'user-friendly' software and the Windows environment have all played a significant role. But perhaps the most important step has been the development of the integrated circuit, commonly referred to as the silicon chip, in which many devices are placed side-by-side on the same slice of semiconductor.

The first integrated circuit, consisting of just two bipolar transistors, was developed by Jack Kilby and Robert Noyce in 1959. Over the next few years, developments proceeded at such a startling rate that in 1964 Gordon Moore made his now famous prediction that the number of devices in a integrated circuit would continue to double each year. Amazingly, over thirty years later this statement continues to apply. The rate of increase has dropped slightly - the figure is now approximately a factor of 1.6 per annum, equivalent to an increase by a factor of four every three years - but the growth rate is still exponential.

This remarkable achievement has resulted in a year-by-year decrease in the cost of computer hardware, leading to an exponential increase in the ownership of computers, and consequently in the number of people able to participate in the information revolution. The decrease in the physical size of the components also gives rise to a corresponding improvement in performance, allowing more powerful supercomputers to be constructed. Although this has a direct effect on a much smaller percentage of the population, the increase in computer power is of enormous importance in virtually all areas of research and development which employ mathematical modelling.

The future development of the integrated circuit is therefore significant from two

points of view; to continue to fuel the information revolution and to increase the type and scope of problems that can be satisfactorily tackled by numerical methods. Making predictions about the future is always a risky business, particularly when discussing a technology which is developing at such a frantic rate, but as a rough guide we can use past history as a guideline. It is well-known that the development of a successful new technology tends to follow an 'S'-shaped curve[1]. Typically there is a gestation period as the first expensive prototypes are introduced to the market. This is followed by a period of commercial investment. The product goes into mass-production, unit costs fall and rapid growth occurs until finally a saturation level is reached. This model suggests that the information revolution will show no signs of slowing down whilst the development of the integrated circuit continues at the present pace (i.e. along the steep section of the 'S'-shaped curve). The crucial question is, for how much longer can this trend continue?

We begin by considering what I will refer to as conventional technology, as typified by Si CMOS (Complementary Metal-Oxide-Semiconductor) integrated circuits which account for over 75% of the world semiconductor market. We consider how far the process of scaling can be expected to further reduce the size of the devices and look at other ways in which the number of devices per circuit could be increased. The main part of the paper considers other technologies which in the future could offer a viable alternative to the semiconductor transistor. We will concentrate principally on the behaviour of semiconductor-based devices as the dimensions are reduced into the quantum regime, but in Section 5 we mention briefly several more radical alternatives. Finally, I will draw some conclusions about the effects that these developments are likely to have on the future of the computer.

2 The CMOS Integrated Circuit

The objectives and technical requirements necessary to maintain the current rate of progress in integrated circuits are set out in the *National Technology Roadmap for Semiconductors* by a consortium of US industry, university and government research centres[2]. The most recent report, which dates from 1994, deals with the period up to 2010. It predicts that by this date circuits containing up to 64 billion transistors will be commercially available. The minimum feature size will be 0.07 μm, and the maximum clock frequency will be just over 1 GHz. (This should be compared with 1995 figures of 64 million devices per circuit, a minimum feature size of 0.35 μm and a maximum clock frequency of 0.3 GHz.)

How will these objectives be achieved? Historically, the maximum number of devices on a chip has been increased by varying three parameters - decreasing the physical size of the transistors and other components, increasing the die area (i.e. the size of the chip) and increasing the packing efficiency. In order to obtain some idea of how integrated circuits are likely to develop in the future, we will consider each of these three factors in turn.

The size of the transistors in an integrated circuit is altered from one generation to the next by a process known as scaling. In the idealized case[3], scaling requires that all of the dimensions, both lateral and vertical, are scaled by a factor $1/\kappa$, the voltages are also decreased by $1/\kappa$ and the doping concentrations are increased by κ. The main consequences are that the number of devices per unit area increases in proportion to κ^2, the switching delay decreases by $1/\kappa$, the power delay product (i.e. the amount of energy required per switching operation) decreases as $1/\kappa^3$ and the power dissipated per unit area remains constant. In practice there are deviations from this model. In particular, the voltage is generally not scaled as described above which

leads to the undesirable result that the amount of power per unit area increases with each generation of integrated circuit. (This is a potentially serious problem because the amount of heat generated per unit area of a chip is already comparable to that produced by the heating element in an electric cooker!)

Although scaling has proved to be very successful in the past, resulting in an average decrease in the minimum feature size of 13% per annum[4], we cannot expect this approach to be valid indefinitely. How much further can the process of scaling continue before the transistors cease to function correctly? Unfortunately, there is no simple answer to this question. The satisfactory operation of a transistor depends on a large number of parameters, many of which are dependent on each other. (For a more detailed discussion of these parameters see Reference 5.) In order to determine the minimum possible size for a working transistor we therefore need to optimize all of these parameters simultaneously. There are too many uncertainties to be able to solve this problem exactly. In the late 1980s it was predicted that the minimum possible feature size would be about 0.1 μm[1]. Smaller devices have since been fabricated in research laboratories, but it awaits to be seen whether the target of 0.07 μm can be achieved commercially by 2010.

The effects of scaling are not so beneficial for the interconnects, the thin strips of conductor that carry electrical signals from one device to the next. The main problem is the time it takes for a signal to propagate along the interconnect. This is referred to as the RC delay since it is proportional to the product of the resistance and the capacitance of the conducting strip. If the interconnects are scaled in the same way as the devices, the RC delay time at best remains constant, and for long interconnects it increases quite alarmingly. This suggests that the performance of the smallest scale integrated circuits will ultimately be limited by the interconnects and not by the switching speed of the devices. To avoid such problems the width and thickness

of the interconnects are often kept constant or even increased. This tactic is also unfavorable because the interconnects then take up a larger proportion of the surface area of the chip. The solution is to use multiple layers of interconnects. Current state-of-the-art circuits employ up to 5 layers of interconnect, and it is estimated that by 2010 up to 8 layers will be required[2].

The die area has increased historically by about 9% per annum[6] and there is every reason to believe that it will continue to increase at the same rate for the foreseeable future. The main restriction on the die area is that as the size of the circuit increases, the yield of working circuits decreases. Increases in the die area that are achieved without reducing the yield therefore reflect improvements in the quality of the wafers, in the clean room conditions, and in the fabrication processes generally. An alternative approach which could radically increase the die area is to incorporate a certain amount of redundancy into the circuit design. This is referred to as wafer scale integration (WSI). It should be noted, however, that although an increase in the die area has a beneficial effect on the cost per function, it does not lead to any increase in performance.

The third parameter, the packing efficiency, is generally used as a means of taking into account all of the other factors that determine the number of devices on a chip. It incorporates, for example, changes in the spacing between the devices and in the number of resolvable elements needed to construct a device. The packing efficiency can be calculated empirically from one generation of integrated circuit to the next by dividing the increase in the number of devices per chip by the increase expected due to the changes in die area and minimum feature size. Since the design of integrated circuits has been more or less optimized over the preceding years, the packing efficiency is now close to unity and so is not expected to play a significant role in future developments. However, one way to dramatically increase the packing density

would be to extend the circuit upwards from the plane so that there are several layers of devices stacked one on top of the other. The main practical difficulty with this arrangement is the problem of dissipating the heat produced by the devices in the middle layers. The complexity of the interconnects would also no doubt cause some headaches!

3 Low-Dimensional Semiconductor Heterostructures

A heterostructure is a device which consists of two types of semiconductor, the principle being to exploit one or more of the differences between these two materials in order to realize characteristics which cannot be achieved with a bulk material. In particular, the conduction electrons (and holes) in the two materials generally have quite different energies, which gives rise to a step-like change in the conduction (and valence) band edge at the heterojunction. This is referred to as a band offset (see Fig. 1).

The band offsets typically have magnitudes between 0.1 eV and 1.0 eV, which is comparable to the change in potential across a p-n junction. This suggests that the band offsets could be used to control the flow of carriers in the direction perpendicular to the heterojunction. This is similar to the function of a p-n junction in a diode or transistor. However, there is an important difference between the dimensions of a p-n junction and a heterojunction. A heterojunction typically has a width of about 0.5 nm since a crystal can be changed from one material to another over a distance of little more than one atomic spacing. In contrast, the minimum feasible size for the depletion region of a p-n junction in a device designed for room temperature operation is a few tens of nanometres. This suggests that heterostructures could be used to make devices with dimensions which are of the order of a hundred times

smaller than those of devices which rely on doping to control the flow of carriers.

However, making a transistor with characteristic dimensions of a few nanometres cannot be achieved simply by proportional scaling and replacing p-n junctions with heterojunctions. This is because on these length scales the effects of quantum theory become increasingly important. The behaviour of a transistor relies on treating the electrons as small solid particles which behave according to the Boltzmann transport equation. But as the dimensions of a device become comparable to the de Broglie wavelength of the electrons - which in a semiconductor is typically between about 10 nm and 50 nm - this analysis is no longer valid.

How do the electrons behave in this quantum regime? Let us consider an electron confined between two heterostructures placed back-to-back, for example, a thin layer of gallium arsenide sandwiched between two thicker regions of aluminium gallium arsenide. If the thickness of the gallium arsenide layer is comparable with the wavelength of the electrons, then the electrons form a standing wave in the cavity (Fig. 2). Such a structure is known as a quantum well. How does this structure affect the movement of the electrons? For electrons travelling in the plane of the gallium arsenide layer, we can picture these electrons as behaving like water waves propagating on the surface of a water tank. They are free to move in any direction across the surface, but cannot move out of this plane. We therefore describe the quantum well as a two-dimensional system. In the extreme case we could reduce all three dimensions so that they are comparable with the electron wavelength. This is a zero-dimensional structure, otherwise known as a quantum dot.

Another aspect of quantum theory which affects the transport behaviour of the carriers is the phenomenon of tunneling. According to quantum theory it is possible for a particle on one side of a classically forbidden region to spontaneously appear on

the other side by apparently tunneling through the forbidden region. This process would be catastrophic to the operation of a conventional transistor, but has provided inspiration for a new breed of devices which exploit this phenomenon. In particular, considerable effort has been invested in the concept of resonant tunneling[7]. Devices based on this principle are not only very fast but also exhibit multi-state behaviour which makes it possible to design functional devices (e.g. a single device can operate as a logic gate or as an N-state memory element). An integrated circuit for the future based on this technology could consist of an array of quantum dots, each functioning as a separate resonant tunneling transistor. If we assume that each quantum dot has lateral dimensions of 10 nm with a similar spacing between adjacent dots, this suggests that we could achieve densities of 250 billion devices per cm^2, with each device performing a function which would require several conventional transistors.

Although this sounds very promising, there are several problems with the above scenario. Many of these difficulties are not specific to the types of devices considered here, but are universal to any technology in which these levels of packing densities are anticipated. We will discuss this further in the following section.

4 Quantum Cellular Automata

In a conventional integrated circuit the devices are placed on the same piece of semiconductor for the sole purpose of reducing the cost per device. As far as the operation of the devices is concerned, great efforts are taken to ensure that each transistor behaves as though it is isolated from all of the others. In other words, the state of device A in Fig. 3 is determined only by the input signals received from the connecting wires and is not influenced by the behaviour of the neighbouring devices (unless they happen to be connected directly to device A by a wire).

However, as the individual components are reduced in size, it will no longer be viable to design a circuit based on isolated devices connected individually by electrical wiring. There are several reasons. We have already mentioned the time delays associated with scaling of the interconnects. Other problems arise from the sheer complexity and topological constraints imposed by the necessity of making physical connections between each device. Another consideration is that as the density of devices increases it becomes increasingly difficult to isolate one device from another.

One way around this problem is to design a circuit which exploits the interactions between the neighbouring devices. This not only eliminates the isolation problem, but also removes the necessity for making electrical contacts to each device. Instead we could envision an array of devices communicating possibly by tunneling, or some other mechanism, with each other. Electrical connections would be necessary only at the sides of the array to provide the input and to read the output from the array. This suggests that some kind of cellular automata is required in which the state of cell A in Fig. 3 depends explicitly on the states of the surrounding cells.

How would such a system work in practice? A system based on quantum dot technology which could meet these criteria has been suggested by Craig Lent at the University of Notre Dame, Indiana[8]. A cell in the proposed system consists of five quantum dots arranged in a face centered square. When two electrons are placed in the cell, the cell has two stable states, as shown in Fig. 4. We can see how the cells interact by considering a one-dimensional chain of cells - if a signal is applied to the cell at one end of the chain, induced polarization effects will cause the signal to propagate along the chain. (A Java simulation of this process can be found at the URL given in Reference 8.) Extending these ideas, it may be possible to build two, or even three, dimensional arrays of cells which function in a similar manner. Several experimental techniques already exist for making regular two-dimensional arrays of

quantum dots, and by using self-organized growth it is possible to achieve ordering in the third dimension as well.

It should be pointed out that this approach is not limited to the use of semiconductor microstructures. Similar cellular automata structures can also be achieved, for example, using single electron effects in metal-insulator junctions[9] and using polymer chains[10].

5 Other Technologies

Having focussed on one possible direction for the future of microelectronics, let us now take a brief look at some of the other 'alternative' technologies which could form the basis for computer processing in the 21st century.

The possibility of using superconducting Josephson junctions to perform a transistor-like function has been explored since the mid 1960s. The main attraction is the potential speed of these devices. This ability has been demonstrated in working circuits, but despite major investments by IBM and MITI (in Japan) this work has fallen short of producing a superconducting computer. The efforts in this area are currently focussed on a rather different approach based on the use of Rapid Single Flux Quantum (RSFQ) circuits[11]. These devices are potentially about 100 times faster than conventional CMOS circuits - simple circuits have been demonstrated to operate at up to 370 GHz and large scale circuits are expected to reach speeds in excess of 100 GHz. Power dissipation is also several orders of magnitude lower than in CMOS circuits. However, it is unlikely that this technology will compete with the levels of integration achievable with CMOS circuits - currently the linear dimensions are about a factor of ten times larger than conventional transistors.

Another alternative technology which has been around for a considerable length of time is optical computing. The fabrication of all-optical switching devices in the early 1980s gave the impression that the optical computer would become a reality in the near future[12,13]. However, subsequent progress in this field has proved disappointing. More promising results have been achieved with a hybrid device, the self electro-optic effect device (SEED). A simple digital optical processor based on the SEED was demonstrated by AT&T Bell Labs in 1990[14], but a general-purpose optical computer still remains a long way off. Nevertheless, the potential for massively parallel processing (which can be achieved due to lack of interaction between photons) may lead to to an important market for certain specialized applications.

The idea of using molecules to perform electronic functions was first proposed in 1974 by Aviram and Ratner[15], but molecular electronics has only recently emerged as a viable approach to computing. Even so, there are still no working demonstrations of processing elements based on molecular electronics. Some of the most promising work has been carried out on donor-bridge-acceptor polymer segments[10] and on proteins such as bacteriorhodopsin (bR)[16].

More radical alternatives include DNA computing and quantum computing. DNA computing is a very new field which began in 1994 following a seminal paper by Leonard Adleman[17]. The original paper described how strands of DNA could be used to solve combinatorial problems such as the travelling salesman problem. Since then it has been shown that similar techniques could be applied to crack sophisticated data encryption algorithms and solve other related problems that are beyond the abilities of any supercomputer. This raises an interesting question about how we should define the term 'computer' given that so much can be achieved with a simple test tube of DNA solution.

What is the ultimate limit in size for a microelectronic device? In theory, the spin of an electron could be used to store one quantum bit (or qubit) of information. If we have, say, eight electrons, then the system exists in a superposition of 2^8, or 256, quantum states. Consequently, if we have a quantum computer that is capable of working with these quantum bits, it would be possible to perform calculations on all 256 states at the same time. However, the construction of a quantum computer is a long way off. At present there are many fundamental issues to be resolved, but a $ 5 million project to investigate quantum computing and quantum information has recently been instigated.

6 Conclusions - The Implications for Computing

There seems little doubt that silicon-based CMOS circuits will continue to evolve over the next ten years at a rate sufficient to maintain the current exponential trend in the number of devices per circuit. However, looking further into the future we have seen that this technology will begin to run out of steam as the problems associated with further down-scaling become insurmountable. This suggests that we may move off the steep section of the 'S'-shaped curve, signaling an end to the information revolution, unless a replacement technology can be found. Fortunately, there are several possibilities, some of which we have discussed in this paper, which are potentially capable of sustaining the exponential growth rate well into the next millennium. The slightly worrying fact is that at present there is no major investment dedicated to converting one or more of these working prototypes into commercial reality. The problem may be that there are simply too many alternatives. If you are going to invest a few billion dollars in a new technology, you need to be 100% sure that it is the right one!

Whichever technology comes to the forefront, it seems certain that there will be

some radical changes. In the past, new technologies have tried to emulate the silicon integrated circuit formula of having isolated devices performing a simple switching function, but over the past decade there has been a growing realization that these alternative technologies will also require a completely different approach to computer architecture. In my view, this will be just the beginning of a fundamental change in the philosophy of computer processor design. The current approach is to build general-purpose computers which can tackle a wide variety of problems by breaking them down into combinations of a few primitive operations. This requires a vast number of individual devices all performing a similar function. In the future, we are likely to see a greater number of functional circuit elements of vastly increased complexity for performing specialized tasks. The logical conclusion of this strategy is that the general purpose computer will be superseded by a range of computers - maybe utilizing a variety of different technologies such as DNA and optical processing - dedicated to performing specific tasks.

References

1. James D. Meindl, *Chips for advanced computing*, Scientific American, p.54, Oct 1987. In this article an interesting analogy is made between the production of steel during the industrial revolution and that of silicon during the information revolution.

2. *National Technology Roadmap for Semiconductors* (1994 Edition). An on-line version exists at *http://www.sematech.org/public/roadmap/index.htm*

3. R. H. Dennard, F. M. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous and A. R. LeBlanc, *Design of ion-implanted MOSFETs with very small physical dimensions*, IEEE Journal of Solid State Circuits, **SC-9**, p.256 (1974)

4. S. M. Sze in *VLSI Technology* (2nd Edition), Ed. S. M. Sze (McGraw Hill, 1988)

5. R. J. Turton, *The Quantum Dot* (Spektrum Academic/Oxford, 1995)

6. J. D. Meindl, *Ultra-large scale integration*, IEEE Trans. Electron Devices, **ED-31**, p.1555 (1984)

7. F. Capasso, S. Sen, F. Beltram, L. M. Lunardi, A. S. Vengurlekar, P. R. Smith, N. J. Shah, R. J. Malik and A. Y. Cho, *Quantum functional devices: resonant-tunneling transistors, circuits with reduced complexity and multiple-valued logic*, IEEE Trans. Electron Devices, **ED-36** p.2065 (1989)

8. C. S. Lent, P. D. Tougaw, W. Porod and G. H. Bernstein, *Quantum cellular automata*, Nanotechnology, **4** p.49 (1993). Java simulations of quantum cellular automata are at *http://www.nd.edu/ lent/QCAhome.html*

9. see *http://www.nd.edu/ micro/*

10. H. Knoll and M. Mehring, *Controlled electronic transfer in molecular chains and segments* in Molecular Electronics - Properties, Dynamics and Applications, Eds G. Mahler, V. May and M. Schreiber (Marcel Dekker, 1996)

11. K. Likharev, *Superconductors speed up computation*, Physics World, p.39, May 1997. See also *http://pavel.physics.sunysb.edu/RSFQ/RSFQ.html*

12. S. D. Smith and D. A. B. Miller, *Computing at the speed of light*, New Scientist, p.554, 21 Feb 1980

13. E. Abraham, C. T. Seaton and S. D. Smith, *The optical computer*, Scientific American, p.63, Feb 1983

14. D. Brady, *Switching arrays make light work in a simple processor*, Nature, **344**, p.486 (1990)

15. A. Aviram and M. A. Ratner, *Molecular rectifiers*, Chem. Phys. Lett., **29**, p.277 (1974)

16. R. R. Birge, *Protein-based optical computing and memories*, IEEE Computer, **25**, p.56 (1992)

17. L. M. Adleman, *Molecular computation of solutions to combinatorial problems*, Science, **266**, p.1021 (1994)

Figure Captions

Fig. 1 At a heterojunction the abrupt change in the energies of the conduction electrons and holes gives rise to discontinuities in the conduction and valence band edges, ΔE_c and ΔE_v. In the GaAs–$Al_{0.3}Ga_{0.7}As$ system (which is the most studied case) the conduction and valence band offsets are 0.23 eV and 0.14 eV, respectively. (Note - the diagram is not to scale.)

Fig. 2 Schematic diagram showing the confinement of an electron state in a quantum well.

Fig. 3 In a conventional integrated circuit the devices are isolated from one another so that the state of device A is unaffected by the behaviour of the neighbouring devices. In contrast, in a cellular automata the behaviour of device A depends explicitly on the behaviour of the neighbouring devices.

Fig. 4 A system of five identical quantum dots arranged in a face-centered square and containing two electrons has two stable states.

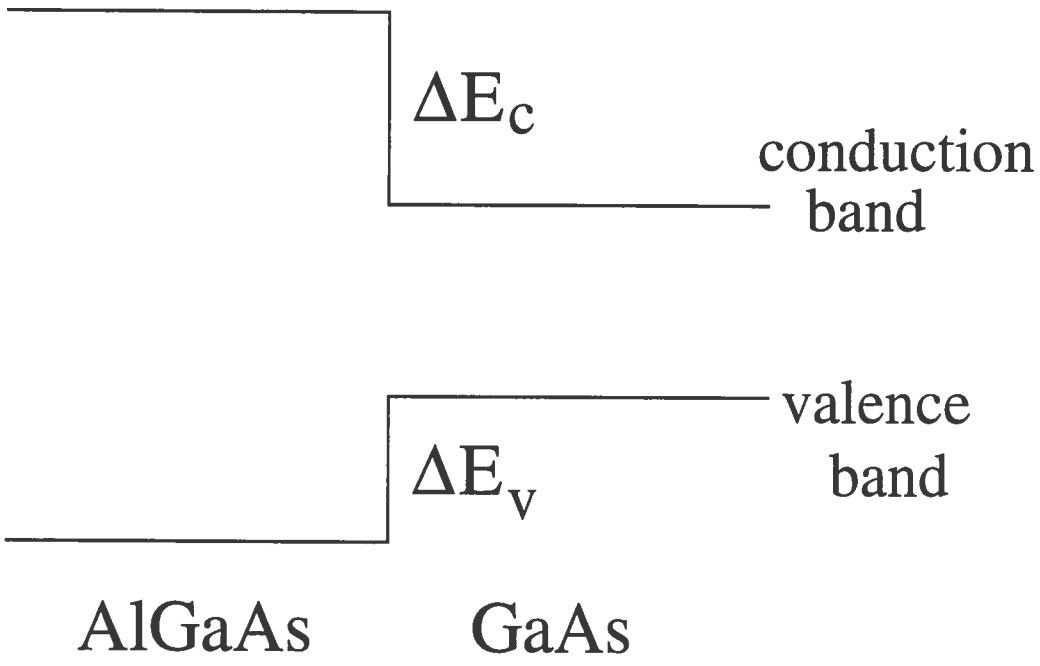


Fig. 1

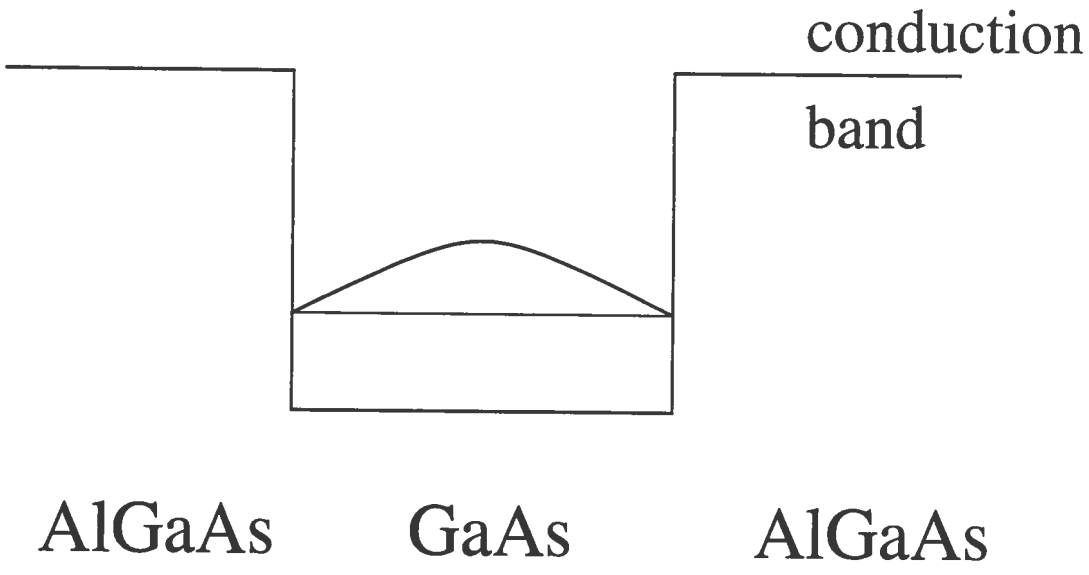


Fig. 2

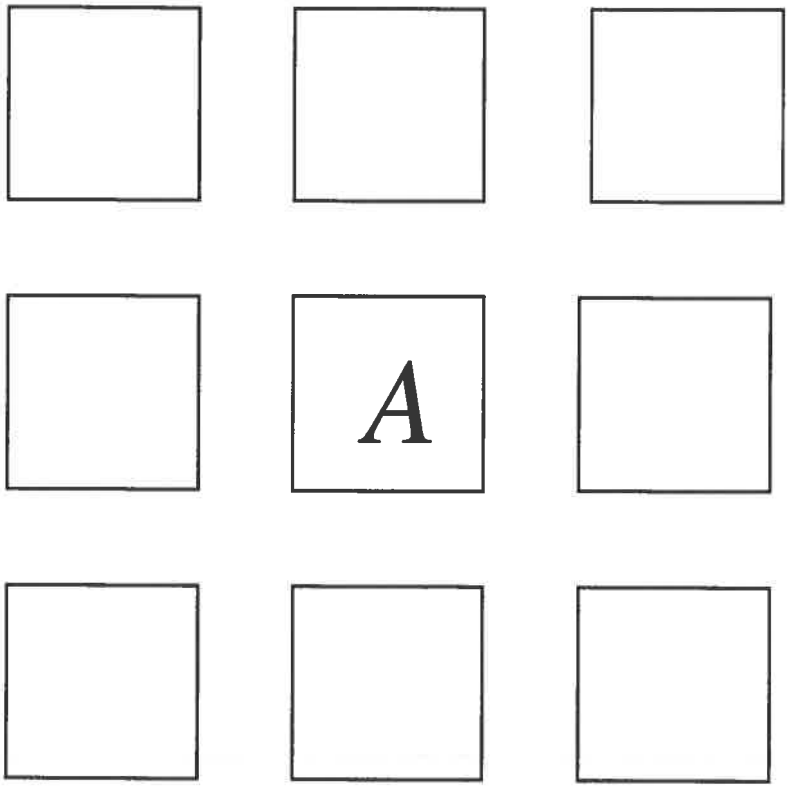
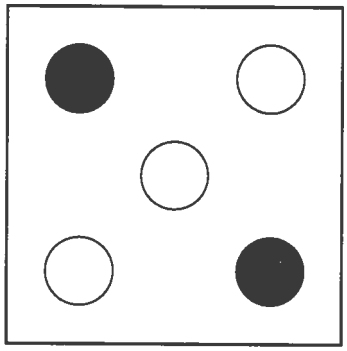
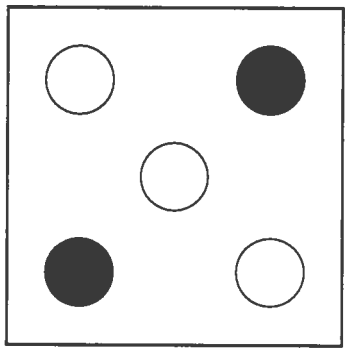


Fig. 3



0



1

Fig 4