MOORE'S LAW AND THE FUTURE
by
Maurice.V.Wilkes

The continual reduction of size has made microcomputers not only faster, but also cheaper. There has been no occasion to make a compromise between performance and cost. This is a reversal of the normal state of affairs. Gordon Moore, speaking at a meeting held at the Cavendish Laboratory in March 1997 to celebrate the centennial of the discovery of the electron, described it as a violation of Murphy's law.

The same has happened with disks. It seems that cost depends more on size than on complexity. I remember noting when I was quite young that a cheap (mechanical) watch could be brought for 5 shillings in English money, whereas the cheapest lawn mower cost several pounds. This was despite the fact that the watch was far more complex than the lawn mower. Cheap mechanical watches can no longer be bought, but that is beside the point.

In the computer world, the effect has been that the market for high performance processors has focused on the latest chip designs. There has been no point in companies continuing to offer older designs, made in less advanced foundries because they would have cost more to make. This has led to the situation in which the commodity processor chip, made and sold in large quantities and used in a whole range of workstations and PC's, is also the chip on which high performance workstations are based. This is a strange outcome indeed, and possible only in a looking-glass world in which Murphy's law is suspended.

It is to be noted that shrinkage alone would have produced faster chips, but not cheaper ones. However, the shrinkage was accompanied by an aggressive move to larger wafers and it was the two together that produced the low cost. Wafer diameter started at one or two inches in the early days of integrated circuits. By 2000 it had reached 12 inches, which is its present value. At first I was puzzled by the move to larger wafers. It increased the cost of foundries and it also increased the risk. It seemed to me that the industry was making things unnecessarily hard for itself. I now see that to reduce unit costs of chips was an essential feature of product strategy, just as important as the maintaining of Moore's law.

The statement about the latest processes producing the fastest and the cheapest chips is a statement about foundries, not about chips. To achieve low cost, the foundry must be fully loaded, but it need not be on one line of chips only; it is aggregate loading that counts.

The economics of foundry use are complex. New products go on to the latest foundries; older products may not be worth the cost of transferring and continue on older foundries. For some products, such as embedded processors, performance may not be a primary consideration, provided it reaches a certain minimum. The fact that older foundries have written off their initial investment also comes into the equation.

Interesting light is thrown on foundry capacity and how it is used by some data to be found on the website of the Taiwan Semiconductor Manufacturing Company (TSMC), said to be the largest of the foundry companies. This company operates eight foundries of varying ages, including one owned jointly, and has several foundries of the latest specification coming along.

THE SEMICONDUCTOR ROAD MAP

The extensive research and development effort making these advances possible is the result of a remarkable cooperative effort by the international semiconductor industry. At one time it would probably have been illegal under US monopoly laws for US companies to participate in such a scheme. Around 1980, significant and far-reaching changes in the laws were made. These made possible pre-competitive research in which the companies can collaborate with themselves and other organisations; the same companies can later develop products in the regular competitive manner. The instrument by which the pre-competitive research in the semiconductor industry is managed is known as the Semiconductor Industry Association (SIA). Membership is open to any organisation that can contribute to the joint effort. SIA has been active as a US organization since 1992 and it became international in 1998.

SIA is responsible for the International Technology Roadmap for Semiconductors (ITRS). Roadmaps are issued every three years, the current one being dated 2001. Each roadmap presents targets that must be attained over a number of years in the future if Moore's law and falling unit cost are to be maintained. Moore's law is explicitly defined as doubling the number of components on a chip every 18 months.

The roadmap is a published document available for all to read. The route that it describes is presented as the "best available industrial consensus" on the way the industry should move forward. It has been accepted as such by the various companies who have competed to achieve the targets set out and hopefully to be a bit ahead of the competition. The targets are aggressive, but they have proved realistic and the progress of the industry as a whole has followed the roadmaps closely. The SIA may be said to have achieved a remarkable success and to have made possible the rapid and orderly development of the industry. The merits of cooperation and competition have been admirably combined. It is to be noted that the major strategic decisions affecting the progress of the industry have in effect been taken at the pre-competitive level, in relative openness rather than behind closed doors. The decisions include the one to proceed to larger and larger wafers.

The roadmap indicates, in a series of tables, the developments that are necessary to achieve the various steps and indicates the research and development that will be necessary to achieve them. In the 2001 roadmap some of the targets present problems calling for serious research. If "manufacturable solutions"? as they are called? are not found, Moore's law will come to an end; this is called the Red Brick wall.

THE ULTIMATE FEATURE SIZE

What the ultimate feature size will be, and how close we will get to it, remains to be seen. Roadmap 2001 projects that by 2010 microprocessors will be 10 times faster than in 2001 and that by 2016 they will be 15 times faster. However the roadmap identifies major technical problems that will have to be overcome if the factor of 10 is to be achieved by 2010, and states that after that progress will stall unless research breakthroughs are achieved in most technical areas. This is the most specific statement about the CMOS end-point that has so far come from the SIA. I interpret it as meaning that there are good reasons for expecting that progress will be made according to Moore's Law up to 2005, but that there is a distinct chance that it will fail beyond that point. I have taken the information given above from the summary of Roadmap 2001 issued by the SIA and dated 6 February 2002. The figures given are rounded to the nearest factor of 5; more precise estimates are given in the Roadmap itself.

At each step the cost of producing chips will rise steeply, and it is cost that will ultimately be seen as the reason for calling a halt. The exact point reached will depend on the economic climate at the relevant time and on the financial strength of the semiconductor industry itself. These factors will determine how far the stopping point is short of where it would be if unlimited financial sources were available.

Although cost will be seen by the world in general as the reason for stopping, it will be only a symptom of the real cause, namely, fundamental limitations imposed by physics.

BEYOND THE ROAD MAP

Everyone is asking what will happen after the development of CMOS as we know it has come to an effective end. I claim no particular insight, but certain things can be said and I will attempt to say some of them.

The technical problems that have to be faced as shrinkage takes place are rooted in semiconductor physics and there are a variety of them. A particularly fundamental limitation is that, as the thickness of the insulating layers is reduced, quantum mechanical tunnelling becomes of importance. This leads to a very sharp cut off since, once the insulating layer is thin enough for tunnelling to occur, any further shrinkage of more than a mere fraction leads to virtual disappearance of insulating properties. We are already down to layers only 1.3 nm thick, that is, less than five atoms. Other physical problems, such as a shortage of electrons to represent a one, do not lead to such a sharp cut off.

The reason why tunnelling is fundamental is that it makes itself felt whatever the system of logic and whatever the semiconductor used. The current Roadmap mentions a number of variants on CMOS as we know it. All these variants need very thin insulating layers in the gates, and there is no future in hoping that one of them might restart the merry round of doubling the density of transistors every eighteen months. This consideration suggests that we should look to logical systems that exploit tunnelling as a feature, rather than continuing to treat it as a problem.

Tunnel diodes were last in the news in the 1960s and I have a personal interest dating from that time. These were the days of discrete semiconductor devices, with each transistor or diode on its own piece of silicon and packaged accordingly.

Over a certain range of currents, a tunnel diode presents a negative slope resistance. If such a diode is connected in series with a suitable resistor, the current passing through it has two stable values and the device thus constitutes a flip-flop. Tunnel diodes were very fast in operation and we set up, in the Computer Laboratory at Cambridge, a major project to investigate their applications and in particular the possibility of building a complete computer based on them. Unfortunately this project had to be abandoned because tunnel diodes became very difficult to acquire. This was partly because they presented serious manufacturing difficulties and partly because transistors were showing such great promise that the semiconductor industry naturally concentrated its resources on them.

One outcome of the project was the building of a tunnel diode memory for a pioneering instruction cache. Tunnel diodes still seem to me to have possibilities as the basis of a high-speed memory, although that does not appear to be a generally held view.

AFTER MOORE'S LAW

The most promising line of development that might lead to a breakthrough in very high-speed logic lies in the direction of single electron effects, that is, effects in which one electron more or less makes the difference between a zero and a one. However, to say this is to reveal how bare the cupboard is. Single electron effects have been demonstrated in a number of places, but little progress has been made towards the developing of a practical device that could form the basis of a computer. Effects have been demonstrated at temperatures well above that of liquid helium, but not yet at room temperatures. The subject is still in the hands of experimental physicists, and I fear that, even with exceptionally good luck, many tens of years must elapse before the construction of a working computer can be contemplated.

It is worth remarking that the Millikan oil drop experiment demonstrated a single electron effect. I am not suggesting that this could be the basis for a computer, although it might be better than the wet chemistry which some people are talking about.

PARALLELISM NO GENERAL SOLUTION

In the mid 1990s, it was almost universally assumed that when hardware development came to an end, further substantial advances in speed would be possible through the use of parallelism. Hardly a conference went by without some speaker asserting that the success of optimizing compilers would be repeated by the writing of parallelising compilers, that is, compilers which would take a conventional program written in C, or some other high level language, and optimize it to run on parallel hardware. This did not happen and indeed it was never likely to happen. There is all the difference in the world between applying optimizing algorithms for performing what are basically elementary operations and reworking the program in a way which challenges the human intellect.

This was an embarrassing period for me personally, since I thought that the expectations that were being so confidently based on parallelism were entirely misplaced. I said so, and was in consequence accused of being negative. It is now generally appreciated that it is by no means easy to achieve a speeding up of a whole program by a factor of two, let alone three or four.

It has, moreover, become abundantly clear that the development of a parallel program, which will give even a modest gain in speed over a simple serial one, is a long and arduous task. For this reason parallelism is for teams not for individuals. This is perhaps the most important lesson that has been learnt about parallelism in recent years. The lone user of a workstation working on his own application has insufficient time, and certainly insufficient temper, to engage in it. If he has a multiprocessor workstation, he may be able to profit by proprietary packages developed by industrial teams, for example a package for performing computer algebra, but that is all. On the other hand, research organizations devoted to particular application areas, such as atmospheric physics, with teams of programmers at their disposal can develop programs to run on large multiprocessor systems.

OPTICAL COMPUTING

If I had been asked thirty years ago what was the way to very high speed computing I would probably have said: “stop using heavy things like electrons and use light photons instead”. The difficulty is that, whereas electrons interact very easily with matter, photons interact only with great difficulty. A transistor has a length compatible with the wavelength of light and in that distance provides the non-linearity required for logic. By contrast, two light beams must travel together for several metres along a fibre, even one doped with erbium, for appreciable interaction to occur and to travel that distance takes many

nanoseconds. The resulting latency would be fatal in a general-purpose computer. It must be accepted as a fact that CMOS computers have already advanced far beyond what could be hoped for from any optical computer that might one day be developed.

SPECIAL-PURPOSE COMPUTERS

Since it is now apparent that continued doubling of speed is not part of man's birthright, and there are no major new developments in sight, one may expect that progress will slow down, at least temporarily, to a pedestrian rate. As a result some deeply ingrained expectations will come under attack. One of these is our attitude to special-purpose computing devices.

Up until now experience has been that, by the time a special device has been developed to solve a problem beyond the scope of existing general-purpose computers, the power of such computers has increased and there is no longer any need for a special-purpose device. Even in areas where special-purpose devices are well established, they have been overtaken by the development of more powerful general-purpose computers. Where, for example, is the Wang word processor today?

There are other reasons why special hardware devices may be unattractive, even if at first sight they appear to offer great speed advantages. Commonly, there is a need for the problem to be set up before the special device can be used and this requires the use of a general-purpose computer. For example, a specialized device for computing the eigenvalues of a matrix would require to have the elements computed by a general-purpose computer before it could be invoked. Since computing the elements may be a time consuming process, this detracts seriously from the speed advantage offered by the special device. Again, special-purpose devices usually have fixed limits to the size of problem that they can handle; for example in the above case, the rank of the matrix may be limited. Such limits, if built into the hardware of the special-purpose computer, make upgrading difficult, if not impossible.

Some of the objections to special-purpose computing devices that I have just mentioned will lose their weight if the development of general-purpose computers slows down to a low rate. More than that, the design of special-purpose devices will in some areas be the only way forward. These may take the form of special-purpose coprocessors designed to run under the control of a regular general-purpose computer.

KILLER PROBLEMS

It sometimes happens that a specialized computing technique is suggested which would be incapable of forming a basis for the development of a general-purpose computer, but which might enable a special-purpose computer for solving certain specific problems to be constructed. I know of two examples, both of startling novelty. One is DNA computing and the other is quantum computing. In both cases, it has been suggested that applications might be identified of such importance as to justify the manufacturer and marketing of a special-purpose computer. Such applications are called *killer* applications.

The DNA people have so far failed to identify a killer application, although they have been looking for a number of years. At one time they thought that the traveling salesman problem might fit the bill, since they were told by mathematicians that it was NP complete. However, it turned out that algorithms exist which yield approx-imate solutions of sufficient accuracy for all practical purposes. Non-mathematicians should be careful how they interpret what mathematicians tell them. Work continues on DNA computing as one

of a number of non-biological uses of DNA, but it is no longer to be taken seriously by the computer industry.

Quantum computing is different in that a quantum computer, if it existed, would have certain properties of universality and so could be claimed to be capable of supporting a general work load. However, universality is not the same as speed. I will not discuss quantum computing any further, since I regard its potential as being well beyond the range of twenty years or so that I have in mind. I may add that I have found little enthusiasm for quantum computing among the computer people I have talked to recently.

There is nothing inherently unsound in the idea of a killer application, but it is, in my view, extremely unlikely that the circumstances that would make it viable will ever be encountered.

DRAMs AND THE MEMORY GAP

Road map 2001 states that up until recently DRAMs have led the industry in the movement towards smaller and smaller feature sizes. Now high performance processor chips are beginning to share this honour.

I remarked earlier that in the past there has been no conflict between DRAM requirements for high performance workstations and for PCs, since in each case the need was for high storage capacity. However, the situation is now changing, since the speed of DRAM based memory is increasing at a rate of only 10% per annum, whereas the speed of processors is increasing at the rate of 60% per annum. The result is a growing mismatch between the speed of the memory and the speed of the processor. This is referred to as the *memory gap*. It only affects programs that do not fit into the cache, or rather those whose working sets do not fit into the cache. The reason why DRAM latency does not scale with the speed of logic circuits is that DRAMs are analogue devices in which information is stored on capacitors.

A measure of the severity of the memory gap is given by the statement that in 2001 a cache miss cost 128 clock cycles on an Alpha 21264 workstation running at 500 MHz. This is not perhaps a very serious matter, but the memory gap will become serious if and when we have processors running at 10 times the speed of present processors, while the speed of DRAMs has only increased by a much smaller factor. As a result, designers of high performance workstations would like to see more attention paid to the reduction of latency.

To its credit, the DRAM industry has made some attempt to meet the requirements of the designers of high performance workstations, but it is unwilling to spend very much money in doing so. Moreover those who control the DRAM industry do not appear to understand the importance of memory latency. They are under the delusion that typical access to data is by reading successive words from a block, and that it is a sufficient solution to the latency problem to provide a streaming mode. Or perhaps they think that, if this is not what users actually want, it is what they ought to want. The result is that for the last five years there has been little improvement in available DRAMs from a latency point of view.

I find myself regretfully forced to the view that, for a variety of interconnected reasons, there is unlikely to be any change in this unsatisfactory state of affairs. I know that I am not alone in this.

Underlying the problem is the different attitude taken to the market by DRAM manufacturers compared with that taken by the manufacturers of processor chips. Those

who make DRAMs do not consider that the market for a low latency DRAM would be large enough to make its development worthwhile. Yet more DRAMs than processor chips go in to high performance workstations. Thus we have the paradox that the processor industry is able to put into processor chips elaborate features like branch prediction and speculative execution, even though the market for those chips is several times smaller than the market which the DRAM manufacturers turn their noses up at. It is true that the number of DRAM chips that go into high performance workstations is a very small proportion of the total number sold, but the paradox remains.

The overall result of the situation is that workstation designers have effective control over the design of the processor, but not over that of the memory system. As far as memory is concerned they have to take what they are given.

One way to solve the problem would be for workstation designers to use SRAMs instead of DRAMs. These would be faster and since they are composed of logic circuits they would scale with the processors. Effectively the change would be a change from analogue methods to digital methods. It is noticeable that when, in some area, a change from analogue to digital is proposed, everyone without exception dismisses the idea as being quite impossible. This state of mind continues until suddenly people wake up to the fact that the change has actually taken place.

I am convinced that if we could look forwards to Moore's law continuing to hold over an indefinite number of doublings of processor speed, the change from DRAMs to SRAMs would take place sooner or later. An immediate change from DRAMs to SRAMs would result in 4–8 times as many chips being required; these figures come from the third edition of *Computer Architecture* by Hennessy and Patterson. As time went on, shrinkage would progressively reduce this number. A consensus would eventually be reached to the effect that, instead of a large memory with a long latency, it would be preferable to have a smaller memory with a short latency, and the change to SRAMs would take place.

WORKSTATIONS BECOMING MORE APPLICATION SENSITIVE

The growth of the memory gap has made the performance of workstations more sensitive to the application than was formerly the case. In the 1990s it was, broadly speaking, possible by using a suite of benchmark programs to assign a performance figure to a workstation which would apply over the whole range of programs encountered in a normal work load. This statement is less true, today partly because of the increasing memory gap and partly, I think? although I would not like to have to give chapter and verse? because there are now many users who run very large programs which together stretch workstations in all possible directions. None of these large programs fit into the cash. Some have working sets with a high degree of locality and some have working sets with a low degree of locality. Some are well served by the algorithms used for various forms of speculation now commonly implemented, while others are not so well served. The various methods available for hiding latency especially multi-threading are more effective for some workloads than for others. These factors all affect performance.

It has been clear for sometime that, in the absence of a revolutionary development in memory technology, this change was bound to occur sooner or later, and that it would affect the way computer vendors conduct their business. They would become more ready to develop variants on their systems that would enhance their performance for particular groups of users; their sales staffs could be relied on to follow this up with vigour.

In a recent issue of *Microprocessor Report*, a writer claims to have observed that the above is now happening. However, he is inclined to blame it on a fickle market which has lost its belief in general-purpose computers. If the trend has really begun, I would see it as a necessary consequence of the technical factors I have just mentioned, and I would expect it to become increasing more marked as time goes on.

The developments that I have been talking about will mean that users who require high performance will have to pay more. However, there is no need to become mesmerised about costs. Users who require the highest performance have always been able to find the means to pay for it. They will do so in the future, rather than give up.

One final comment: on the whole, I am inclined to think that there is a chance that we will see a real breakthrough in memory.

October 2002