THE HISTORY OF COMPUTERS TO THE YEAR 2000

R.W. Hamming

Rapporteur: Mr. F.R. Parker

Abstract:

In spite of the difficulties of making long-term predictions, the development of computers is so important to our society that it is necessary to try. Estimates of hardware developments are based on: the velocity of light; the size of molecules; and the problem of heat removal.

1. Introduction

It is well known that predicting the future is a dangerous sport, but the importance of the future of computing is too great to ignore it. Of course any sufficiently detailed prediction is bound to have a low probability of being right, nevertheless it is worth trying for the following reasons:

1. Money looms very large, because we have to choose machines and we have to choose systems. We have to plan for the future.
2. Scientific potential is even more important. What are the kinds of things we hope from computing to aid science in various areas, not only from computing but from various data processing forms?

2. Past Trends

Twenty five years ago, in 1950, at the very beginnings of electronic machines, the major computers available to most people were still basically relay machines, such as the IBM 601, 602A the Bell relay machines, and the Harvard Mk. 1. These could achieve operation speeds of about 1 operation per second. If at present we are close to 10^8 operations per second, where will we be 25 years from now? It is somewhat uncertain what an operation is when referring to one operation per second, but I will take it to mean a "fixed point multiply". In independent figures produced by extrapolation at Los Alamos a limiting speed of approximately 3.08×10^9 operations per second was arrived at. Quite recent trends have given a number which is not a multiply time, but a pulse rate, which increases by a factor of four every five years. The reason for using pulse rate is that it is more basic and we have managed to speed up machines, and multiplies, by shrewder organisation.

The reason for thinking there is an upper limit to the speed of computing is that there is supposed to be a finite velocity of light, and we use electromagnetic waves. Reviewing past trends we see that: from 1945 to 1960 we were using essentially valve machines; from 1960 to 1965 we were using encapsulated transistors; from 1965 to 1970, solid state devices; and from 1970 to 1975, we were putting a few thousand components on a chip, which is generally called the fourth generation of machines. We are now beginning to put some tens of thousands of components on a single chip. It is not known how long this is going to last, but the next probable stage will occur when the chip yields become very high and a chip can be made, a layer of insulation placed on top with a few holes in it, another layer laid on top, and another on top, and so on. This will make essentially a solid - solid state device. The reason for this, as will be explained, is the necessity for keeping all components of a high speed computer close together.

3. Future Limitations

In 1974 the general component packing density for integrated circuits was a few thousand components per one tenth square cm. of surface area used. It was achieved by chemical etching of photographically exposed surfaces. We are now using electron beam etching and a linear reduction of around 20 to 1 has been achieved, giving a 400 fold increase in the component density. In chemical etching the minimum wire widths are typically 10^{-3} cm (10^{4} nanometers), while the electron beam results in wire widths of less than 500 nanometers, which is the typical wave length of visible light. As 0.1 nanometer is about the typical spacing of atoms we have a definite limitation on decreasing the size of the components (which we must make small if we want to increase the speed). Thus our predictions from past exponential growth of computer speeds face a finite limitation in the future. What is a reasonable guess at this ultimate (soft) limitation? Suppose we immerse the component in liquid nitrogen (boiling temperature 77.4^{0} K) both to reduce the background thermal noise and increase our ability to remove the generated heat. Then wire widths of 100 atoms, meaning cross sections of possibly 5000 atoms or a bit more, seem to be a reasonable soft limit <u>if</u> we are to depend on the bulk properties of matter and avoid the uncertainty principle of quantum mechanics.

When we ask for the lengths of the longer wires in terms of wire width on currently produced integrated circuit boards we find that some lengths are almost 100 times the wire width. Therefore, provided geometry does not alter significantly, we can expect our wires of 100 atoms width to have some wire lengths of 1000 nanometers between elementary components. Similarly, when we look at the typical diffusion zone of current solid state devices we find a typical zone width of around 200 atoms. Impurities in solid state devices tend to occur at about one part per million, suggesting that such devices (excluding, as we are, basically new discoveries and taking about 1000 impure atoms as acceptable) will have about 10^{9}

atoms or more, which leads to a minimum linear dimension of a component of around 10^3 atoms.

The velocity of light enters our calculations, and it has a very subtle effect because it is an upper limit to the speed of signal. There is no use in having a very fast machine if it has to reach very far to get at the memory. In one nanosecond light goes one foot and in a picosecond (10^{-12} seconds) it goes 0.01 inch. Thus if you are going to have pulse rates that are very high, (10^{12} and so on) everything you want to access had better be within about 0.1 inch otherwise you will find your speed is being lost waiting to reach some component. Thus we have a limitation on the size of a single processor. Machines such as the Illiac with 256 parallel processors do exist, but it is impossible to believe that they would be used on other than very special problems. One need only consider the problems of writing an efficient Fortran compiler for 10 processors to understand the point. Similarly pipeline computers (which, like the name suggests, consist of a 'pipe' down which information flows and at each point there is a processor to do an allotted task) where the data stream has only known paths of influence on future data, have limited uses (such as, for example, processing speech in a telephone network). We are restricting ourselves to general purpose computers where there is essentially one or maybe two processors which you would expect to be used fairly efficiently.

The third limitation is heat dissipation, and it is a serious one, even now for planar computers. At Los Alamos the Cray 1 machine, which is built out of standard circuits, is driven hard to obtain its high speed, and generates a few hundred kilowatts of heat to dissipate. Now the smaller the components are made, the more components and the denser the components, therefore the more heat per square centimeter to be removed. Also, as speed increases the heat problem also rises. It is surprising what modern research into heat pipes has achieved for heat removal, nevertheless it is to be expected that well over 90% of the volume of the future fast computer will consist of heat removal components, with the computing components spread around the heat removal sections. In the solid-solid state devices a great deal of room has to be left for copper slugs and heat pipes. We can also use convection as well as conduction for heat removal in the form of some circulating liquid. Liquid hydrogen has a boiling temperature of 20.4° K and will tend to prevent electron migration in the atoms, while liquid helium is even lower at 4.2° K, but it does not seem likely that it will be worth the effort of holding these low temperatures, except in very special cases. There is also the phenomenon of heat super-conductivity, but it is unlikely that we will be using it by the year 2000.

How far can we limit the heat generated? Probably not much. It takes a certain minimum amount of heat to change the state of a device. Also the faster you want to change the state of a device, apparently the more energy you must use. The tighter we pack the components, the more changes of state per given volume, and this, of course, is multiplied by the higher rate of changing states. Together these two problems make a difficult, though not completely unsolvable problem.

A common suggestion is superconducting computers, but superconducting computers have to become non-superconducting occasionally. At present it is believed that the Josephson effect can switch currents in 0.01 nanoseconds and, for theoretical reasons it is unlikely to get much faster. This switching speed must be comparable to or less than the transmission speed if the computer is to operate fast. The main difficulty with the superconducting approach is cost, and we will probably not see a superconducting computer until some other technology uses super-conductivity enough to get the hardware developed and into mass production. Experience shows that, for a device as complex as superconducting integrated circuits, routine factory production is necessary to reduce the costs of fabrication. If you consider the history of magnetic cores and computing storage technologies, magnetic cores were in mass production and the constant work at low level engineering detail was improving them. There were many computing technologies which might have done better if only investment had been made in them, but cores were able to stay ahead until very recently. Thus there will not be superconducting machines unless someone else first uses superconductivity enough to pay the large initial developing costs, to produce the low cost, great reliability, and safety that is necessary for a large superconducting machine to come into general use.

In principle it is possible to beat the heat problem by building a completely thermodynamically reversible computer, which would do the computation, record the answer, and quickly reverse everything to re-absorb most of the heat. Although a theoretical possibility, it is unlikely before the year 2000.

In this way we are led to imagine a central computer, possibly smaller than a walnut in size, which is either immersed in liquid nitrogen, or suspended just above the boiling liquid surface by, for example, a quartz fibre. Similar to a space satellite immersed in space, communication between the computer and the large outside environment of peripheral equipment would be via electromagnetic signalling (probably laser beams) with electromagnetic power supplied in a similar way to the solar batteries of space vehicles. Physical repair of such devices will be a problem since one cannot hope to raise and lower the temperature many times before mechanical failures occur. To compensate for this there will probably be redundant parts, with the ability to switch the defective parts out where necessary.

One may ask, 'Why spend all this effort on making fast machines?'. There have always been problems bigger than we can compute, in terms of required time; the prime one to consider is weather prediction. At present poor models are used, but, on the other hand, it is no use predicting the weather one day in advance if it takes three days to make the prediction. It is evident that the prediction must be made 'faster than real time' and the more computing capacity available, the more detailed the model can be and thus the more accurate the predictions can be made. This is one of several problems which will justify a high-speed machine; in this case, because of the advance warning, knowledge of potentially catastrophic situations can save both lives and money.

Thus we have a viable image of the future machine, which can be developed in more detail if necessary.

4. Storage Devices

The storage devices have been basically magnetic. Although we began with vacuum tubes and Williams Tubes, magnetic devices have dominated the information storage mechanisms for many years, in the form of cores and backup stores: drums, disks and tapes. Recently solid state memories have started to replace cores as the main high speed storage unit, and there seems little doubt that this replacement will become complete. Despite the work being done on bubble memories, it is difficult to define exactly what these are, thus it is not known what effect these will have. The large magnetic backup storage units are not so easily replaced and may well remain. In the 1973 era the first versions of 10^{12} bit memories arrived, but it is difficult for me to imagine the amount of information in 10^{12} bits, so their proper use is, to me at least, unclear.

5. Input and Output Devices

It is well known that it is the dramatically cheaper components of the central processor, which have reduced the cost of computers in the past, as well as making them more reliable. The cost of peripheral equipment has not dropped in price as much nor increased as greatly in reliability, and it seems that input-output equipment will remain the troublesome part of computer systems.

Output in the form of hardcopy (printed paper) is highly desirable. V.D.U. devices are fine when you only want to make notes of what is going on, but you can not always go back to construct exactly things for which you have no hard copy. Without hard copy it is very hard to do scientific research because you do not have a firm path back once ideas have caused you to try several things. Certain amounts of scrolling can be implemented, but these must always be finite. However, cathode ray tubes are playing an increasing role for display purposes and are adequate if not ideal. Xerox-type machines may help solve the problem, and it is likely that voice output will be very useful because sound is omni-directional. Thus voice output can inform the user of occurences (such as termination of a job) for which he does not require a permanent record. However, spoken input on a general basis seems more remote in time.

6. Computer Architecture

A computer is the assemblage of individual components into a useful combination; individual switches are put together to form multipliers and so on. Using the earlier discussion it becomes difficult to believe that 10^{14} multiplications per second will ever be achieved, 10^{13} being perhaps close to an upper bound, and 10^{12} being a reasonable ultimate goal to hope for. At present we are somewhere near 10^{8}, which suggests the remarkable fact that in 25 years we have progressed almost two thirds of the way, measured of course in the exponent of growth, of all that it is reasonable to hope for in multiply speed. The three physical principles on which this figure is based are: the finite velocity of light and the assumption that nothing can signal faster (there may be exotic particles which move faster, but we do not know how to use them); we must use very small detectors and emitters (of some form of electromagnetic radiation) and these must be built of molecules (it is difficult to envisage building out of smaller particles); and we have to get rid of heat. It is difficult to escape these three physical limitations; therefore we can be reasonably confident of the predictions of the order of 10^{12}. Comparing this figure with the extrapolation from Los Alamos where the figure was $\propto 3 \times 10^{9}$ they differ by a factor of 300. The truth is probably somewhere between the two; the evolution of computing cannot go on for another 25 years and see another increase of 10^{8} in pulse rate. We are rapidly approaching the end of a saturation curve.

This is for a single processor, without the problems of parallelism. However architecture includes the organisation of parts into a whole computer so it is wise to review this trend. In the early von Neumann type machine everything went through the arithmetic unit - it was the buffer for all information transmissions. From the almost completely sequential machine it was quickly shown where pieces of parallelism would speed up the whole. Index registers were one of the first parallel processing units, and they saved the loading and unloading of the arithmetic unit. Then small computers placed in connection with the input-output units greatly speeded up the main machine. Thus gradually what was conceived of as a sequential machine attained a reasonable degree of parallelism, with smaller local control units hidden within the main computer. Parallelism probably has still more to offer, but it raises two nasty problems. It is not so much a question of what can be done, but how economically it can be done; that is to say, is it better to put many machines into one box to make a very fast machine, or just make several smaller machines? Secondly, how reliably can it be done, not so much in hardware terms but in the area of recovery; how do you track a problem in several small machines?

There are two schools of thought on how to structure a maximally fast computer; one favours extreme simplicity of architecture, while the other tries to overwhelm the difficulties through the use of very many components. The future will probably use both approaches, with simplicity tending to dominate.

7. Computer Size

The four size classes of machine had different sources of inspiration. The biggest machines, the maxis, lie on the frontier of development and, in the beginning of electronic computing almost all the machines were a kind of maxi. These gave rise to the midis which have been widely used in science and business. The minis came from the need for computers to interact with the outside world and required reliability for process control. Their design is centered around this pressing need for reliability. The so-called micros, or hand-held machines, are now becoming very widespread and come out of the integrated circuit technology. However hand-held is not the essential element, though it has selling attractions, but rather it is easy portability that matters.

The essential problem is that the more flexible and powerful the computer (in some real sense) the more different are the problem calculations it can perform, and therefore the more information that must be supplied to specify which one should be done. Even in the 'better' hand-held machines each button can represent up to three different things, depending on the position of another button; therefore things are beginning to get a little awkward. Earlier, the desirability of hard copy was mentioned: a reasonably labelled record of what was done. It is this 'reasonably labelled' which causes problems since it indicates the alphabet must be available. Thus it would seem that the micro machine will grow to the size of a fairly large book, having room for a reasonable typewriter keyboard and a roll of hard copy paper, but still being very portable, probably plugged into the mains. This appears to be the type of machine which will become very popular - a portable typewriter with several chips added.

While there are definite limits to the speed of a machine, the same is less true of price. Price has a habit of falling very rapidly under mass production and competition (take the case of hand-held calculators). Thus, though we may have achieved two thirds of the exponent of speed, we are probably nowhere near two thirds of the exponent of cost; cost will go down much further per 'operation done'.

In the architecture of computers, networks have a role to play and these can be of large or of small machines. For psychological reasons the networks of small machines will probably become the most popular. The reason being that if someone needs to do a job that is of vital importance to himself he would prefer to have control over the machine rather than being 'just another user' who may be overlooked when modifications are made. The owner of a machine can always revert to the previous software if modifications should cause his program to stop running; a large bureau is apt to be less sympathetic. However, the value of being connected into a wide range of input, output and storage devices means that many smaller machines will be arranged into networks of computers and have the ability to reach out and use such facilities, without, at the same time, outside influences being able to reach into the small machine.

8. Summary

We make predictions both for economic and for scientific reasons; we like to find out what the future holds for us. One prediction, based upon the extrapolation of past trends, is about 10^9 operations per second by the year 2000. Using physics (the velocity of light, the size of molecules and heat generation) the upper limit is about 10^{12} operations per second and suggest a slightly higher value, say 10^{10} multiplications per second by the year 2000. Other predictions are the solid-solid state device; the fact that machines will be small of necessity and the great problems which will occur with heat removal. There is an assumption behind all these predictions that the world will continue on a smooth course (with no great revolutions, or atomic wars) and society will continue to evolve, as in the past 25 years, with its minor ups and downs. Any great catastrophe is bound to invalidate any detailed prediction such as this is intended to be.

9. Discussion

Professor Page pointed out that the prediction of a date (the year 2000) was based purely on technical constraints and wondered what effects economic constraints would have on this date. That is to say if the extrapolation of the proportion of the gross national product spent on computer development was taken into account, what would be the effect? Professor Hamming quoted Seymour Cray (builder of the Cray 1 at Los Alamos) who remarked that the cost of the Cray 1, 7.5 million dollars, has long been the most anyone would pay for a computer; thus he concluded the fastest machine in the year 2000 would also cost 7.5 million dollars. As a more direct answer he said that the costs of development of large machines, such as the IBM 360, were so high that they will not be made again, and therefore concluded that the percentage of gross national product for computer development would drop in the future.

Professor Randell noted that the only time software had been introduced was when parallelism was mentioned, and that it had then been a passing reference to Fortran compilation. Was this a relevant problem for the year 2000? In reply Professor Hamming said that, in the year 2000, Fortran would still be there. It may look different and have acquired many desirable features, but it would still be called Fortran.

Staying with the theme of parallelism, Mr. Laver asked if this was the ultimate escape from the speed of light; and were there any problems to which this technique would not apply? Professor Hamming repeated that parallelism would probably not speed up the compilation process very much. To which Dr. Hartley commented that 256 separate compilations could be done at the same time on a 256 processor machine. He said that the most efficient use of parallelism was to do that which was most naturally parallel; in this case 256 persons working simultaneously. Professor Hamming wondered what advantage parallelism had over 256 separate machines and pointed out the great software problems involved in controlling the parallelism.on 256 different problem compilations or on a single compilation. However he believed that the running of large problems, such as the weather problem, would definitely require a degree of parallelism.