Invited Review

Trends in computers

Ralph E. GOMORY
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

pp. 330–340
Invited Review

Trends in computers

Ralph E. GOMORY
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

Keywords: Computers, technology, systems, storage, software

I am really pleased to have this chance to talk to this group, which includes so many people of whom I have such very pleasant memories.

I will talk about the evolution of computers, which of course, is something I have been spending a lot of time on for the last fifteen years. This evolution, I also believe, is one of the great events of our time. It certainly is clear to all of you. I think, the effect that it already has had on management science and I think to some extent, the effect that it has already had on society at large. One thing I am sure you are going to see as I go through this talk in more detail, is that there is every reason to believe that the present rapid progress of computers and therefore their effect on management science, on society, will continue.

My talk is structured as follows. First, I am going to go through the elements of computer technology: logic and memory, storage, and a few words about software technology. Next, I am going to talk about the impact which the progress in those areas has on the computer systems that are built out of those parts. Then after I finish this technical part of the talk, I am going to talk briefly about the national and international reaction to this very rapid evolution of computers.

There is one very peculiar characteristic about computers which, although probably many of you are conscious of it, I want to stress, because I believe it is the fundamental thing that lies behind technical progress. You may have observed that computers seem to get better every year—although there are usually no breakthroughs. Yet there is this steady progress. The basic reason for this can be summarized in one sentence, which is that a small bit is just as good as a big bit.

You have to remember that when we are working with computers we do no work in the physical sense. We are not like automobile manufacturers, we do not have to make something that has a capability to do physical work. It does not have to carry people up hills. All computers have to do is make and move distinguishable marks. In general, if we can make that very small mark, let us call it a bit, smaller, it usually becomes much cheaper and the logic usually, though not always, goes faster when the bits are closer together. So since we are not doing any physical work and since we are only moving marks, we do have available to us a most remarkable royal road of progress, which is simply: Make everything a little bit smaller every year.

And that is what is happening and what, as far as I can tell, will continue to happen as far into the future as one can see. Everyone has a different idea about how far into the future they can see, so I will say for these purposes with great definiteness for the next ten years and with a considerable definiteness thereafter.

At some point we may have to change technology. The present dominant technology of course is the silicon technology, whose progress I am about to describe to you in more detail. The scale on which we work in the silicon technology is mea-
sured today by units of microns or micrometers or
millions of a meter. However, we know that by
other means it is possible to record information
basically at the molecular level and all of us do
that automatically in our own bodies, so that
beyond the whole scale of silicon technology there
is no reason to think that, even if the silicon
technology should falter, the business of recording
bits ever finer has any fundamental limitations, at
least down to the molecular level.

I think we are all going to have to live with the
kind of progress that we have seen in the past and
which I am about to project out for the future.
The dimensions at which people can expect to be
working in silicon will be moving from today's
dominant two-micron technology forward through
one-and-a-half, one-and-a-quarter, one micron by
the end of this decade, and slightly below that in
practice. All of this is done without anything that
one would call a breakthrough. It calls for a great
deal of intelligent work. A great amount of detail
has to be worked out to make the various layers of
silicon thinner but free of all holes, to make the
interfaces between the layers have the correct
electrical properties, and so on and so forth. That
work calls for great intelligence, great ingenuity,
and the passage of a certain amount of time; but
there is absolutely no reason to think that it
cannot be done, and we can very confidently
expect to be working at these dimensions in the
years indicated.

By the way, in all the projections which I am
going to make, I want to stress three things. First
of all, you are going to see a lot of straight line
projections. However, none of them are done by
taking the past and making a straight line extrapo-
lation from it. They are really based on technical
analyses of what will be possible in the various
years; in other words, the technical work has been
done to analyze the problems. Second, all the
projections I give are what I would call 'surprise
free'. I have made no attempt to speculate or to
incorporate the possibilities of totally new events
or of breakthroughs of one sort or another. Any
such event would cause even more rapid progress
than is shown here. A similar surprise that has
been excluded would be the possibility of these
apparently straightforward technical evolutions
running into an unexpected barrier. That, too, has
been known to happen. We might call that a
'negative' surprise. My projections are surprise
free in both directions. However, I will give you a
subjective evaluation, which is that positive
surprises in these technologies are more likely
than negative ones. What you will be looking at, I
think, is a conservative projection.

Now, if the technology evolves into micron and
submicron dimensions, it will make possible a
very large number of transistors on each silicon
chip. Figure 1 is a partly historical and partly
projected count. All the named chips here exist in
various degrees of reality, and the rest of the
projection, which is based on our technological
capability to work at these dimensions, indicates
something like a million transistors towards the
end of the decade. The way you reach this, as I
say, is not through straight line extrapolation, but
by looking at the processes you expect to have by
the year 1990, what dimension the individual
transistors will be, what yield can reasonably be
expected, and therefore how many per chip. The
purpose of the historical aspect of the chart is to
show that this projection is a continuation of an
already existing tendency.

If you take those transistors, you can then
count up the number of circuits that would be
available and from that, and from a circuit count,
you can consider the sort of central processing
unit, or CPU, that could be made with that number
of transistors. Today, although spread out
over many, many, many chips, we know what sort
of a central processing unit you can design with
150000 circuits or how much faster a machine will
go if it has 300000 circuits instead of 150000
circuits. This is because the speed of computers is
dependent on the individual circuit speed on one

![Figure 1. FET evolution](image-url)

---

Figure 1. FET evolution
hand and on the number of circuits which you are willing to lavish on the central processing unit on the other hand. Both of these are ingredients of speed.

The sorts of speed that can be obtained from a one-chip central processing unit are approximately 5 to 10 MIPS in 1988 and 10 to 20 MIPS in 1991. Everything has to be taken a little bit approximately because there are various degrees of availability of technologies: they can be in the laboratories, they can be in limited production, they can be in full production, and then there is a dispute to how many years after that you can expect to see actual machines being sold based on those technologies; so it depends on how you imagine the whole process working.

And again, these MIPS numbers (or millions of instructions per second numbers) are obtained from thinking through the processes on the chip, how many circuits are available, what sort of machine you can then design out of that number of circuits, and that results finally in an estimate of speed. I think the main point that you should get from this is that a one-chip CPU (you can think of a workstation or personal computer type of machine) will, towards the end of this decade, be somewhere in the five or ten or possibly twenty MIPS range, so that you should think in terms of having on your desk the sort of computing power that today is associated with the very largest mainframes. This evolution is not anything like a science fiction type, but is the most straightforward and the most conservative possible projection. So I think we are going to have it.

So far I have talked only about the single chip central processing unit. Today, the larger machines and the faster machines use many chips to make available a very large number of circuits which again, translates into more speed. This will continue to be the case. The larger machines with multichip CPUs will, in the same time period towards the end of the decade, probably be giving something like one hundred MIPS or perhaps a little bit more.

The technical evolution in this case is the evolution of the so-called 'package', which, although it does not capture the public imagination in the way that the evolution of chips does, is as technically difficult and as necessary to the evolution of large machines. What you are looking at here in Figure 2 is a package, a first level package, the package that supports the chips directly. The little
back squares on them are chips, and the package has been cut away to show the interconnecting wiring that allows the signals to pass from one chip to another. I can only assure you that the evolution of very fast machines, for very large scale scientific computing or for very large scale commercial use, depends as much on the evolution of these packaging, which become extremely complex, as it does on the evolution of the chip technology. Today’s high end packages consist of 33 layers of specially fabricated ceramic, which enable you to have 33 layers of interconnecting wiring of a very fine scale. That whole evolution, together with the problem of extracting the heat from this multichip machine, will also go forward and give us something like a hundred MIPS per CPU in the larger machines.

The same technology and the same observations hold for memory. Memory will move forward using the same technologies to multimillion bit chips, so that memory, too, is carried forward by the same evolutionary path. However, most data and most large programs are in fact stored not in even the large memory but in disc, and it is a very legitimate question to ask whether the disc technology, which is fundamentally a mechanical technology, not unlike a phonograph record, can reasonably be expected to progress.

The answer is both yes and no. What is quite remarkable and surprising is that the density on disc, that is, the number of bits per square unit, per centimeter of surface, can reasonably be expected to go forward, due to the same process of miniaturization: That is, it becomes possible to make smaller and smaller bits. However, the access time to move the arm to the bits does not change. It is still the same swing of the arm year after year, so that in fact the technological evolution in this area will be towards cheapness (cheapness is basically governed by the number of bits per square inch) but the speed, unlike the semiconductors, does not go up with the density. The speed, therefore, of disc technology should remain approximately constant over the next decade but the cheapness will evolve almost as rapidly as in the semiconductor case.

That this is possible is quite remarkable because the fundamental limiting factor in disc technology is the necessity of having the head get so close to the disc surface that it can detect a single bit uncontaminated, so to speak, with the signal from the adjacent bits. The height at which that head flies above the surface of the disc is today a fraction of a micron. It is actually a finer distance than the dimensions in which we work in semiconductors, so that you have this disc going at 3600 revolutions per minute and the head flying less than a wavelength of light above it. In order to make the progress projected here, that distance will have to decrease so that the ever smaller bits can be individually detected.

It is somewhat surprising that this can be done. If you were to blow this whole thing up, I think you would see the surprising nature of it, because if you blew the head up to be the size of an airplane you would find that the problem is to fly the airplane over the ground at full speed six inches off the ground. It is true that there are bumps, and this is one of the fundamental difficulties in this area.

Nevertheless, there is every reason to think that this will evolve as indicated in Figure 3, and therefore that the cheap discs will accompany the cheap memory and the cheap CPU towards the end of this decade. However, what you will see is —because of the speed of the CPU at that time and the lack of speed of the disc—an ever increasing speed gap between the time to get to, say, a register on a CPU and the time to get to bits on disc.

For that reason, the memory systems of the computers of this period will become more complicated and you will probably be looking at something of the following sort: That is, that behind the registers there will be a first level

![Graph](Image)
cache, behind the first level cache a second level cache, behind that main memory, behind that a paging memory, possibly behind that still another level called disc cache. But somehow, that enormous and growing gap must be filled up by stages of memory in order that the data can be staged down and the data that is currently needed will be available at a pace that is commensurate with the speed of the CPU. This calls for two things: The development of the proper staging algorithm, and, more inherently, that the problems should have a certain inherent locality of reference so that they are not jumping all over the possible data.

This figure (Figure 4), by the way, gives you some feeling for the dimensions at which one works under disc technology, because Figure 4 shows a portion of the head. This is the sensitive portion of the magnetic head that flies up over the surface here. All those coils that you see in Figure 4 are about two microns or one-and-a-half microns wide and are fabricated by the same techniques, basically lithographic techniques, that make it possible to fabricate very small semiconductor structures.

So again, the fundamental thing that is driving the disc technology is the possibility of making, through lithography, very tiny structures. Those tiny structures then are used to detect very tiny bits. The only hope for getting away from the technical difficulties in this area lies in the devel-

Figure 4. Thin film head
development of the optical technology. The real technical difficulty in discs lies in the closeness of the head to the disc. And the need for closeness comes from the fact that no one has ever succeeded in beaming a magnetic field like an optical beam. It is too bad you can not seem to beam magnetism. If you could, it would be wonderful. But, in fact you cannot and so there is possibility of having basically the same structure but with beams of light being used to detect the bits on the surface. The main advantage of this is that it enables the head to be pulled back from the surface. This optical technology is in a state of rapid development. At the moment, at the predictive level, it has the characteristic that you can only write once on the surface because of the nature of the change which you induce into this surface: you burn a little hole in it, and that is irrevocable. However, there are variations on this that have been worked on which would have the characteristic of being erasable, so I think one way or another we will manage to get the disc technology to keep up in cheapness with the rest of the machines, but not in speed.

Let me say a few words about software. One of the troubles with software, as a matter of fact, is that it does not lend itself well to slides. I am going to talk first about systems software. The system programming on which we are all unfortunately somewhat dependent, because the machines cannot really do their functions without it, does not show tremendous signs of progress. There are many tools basically of a sort where you keep various versions of the system software and you are able to have many people work on the same version and keep their updates separated and keep the updated versions themselves easily accessible and without confusion. All of that sort of thing is possible, but the more exciting possibilities, for instance to be able to specify the operating system in a high level language which is relatively intelligible and expresses in a semi-transparent way your ideas and then to have that translate into executable code seem to be quite far away. What we have seen in the past is slow progress and what we are likely to see in the future in this area—and I specify systems programming—is probably more of the same.

On the other hand, application software, which is of particular interest to this group I think, has many, many more favorable attributes. For one thing, there is the availability of workstations. Anyone, I think, who has had the experience of programming on a one-person workstation is aware of the simplicity that goes with a very simple operating system, made for one person, the quick response, and the many other rather fundamental characteristics that make application program development easier on a workstation. And one of the things I will comment on later is the possibility, or the very real possibility, that what we will evolve to are easy-to-program workstations connected in a transparent fashion to large machines in order to invoke the large machine, when and only when that large machine power or data is actually necessary.

Application programs themselves, however, that is, the software technology, has gone through a considerable evolution. When we used to program a long time ago, as some of you can perhaps remember, everything was in the program, the program proper, so to speak, the data, the I/O, all of that you wrote in one great indistinguishable mass. In some sense, the evolution of applications programming has been to segregate out portions of this and to pass some of this function over to the operating system or, in other ways, get it written once and for all. What we are seeing today is a further step in that direction. We have already separated I/O and commands. We are moving in the area of knowledge-based programming, into separating out part of the knowledge, so that you have, so to speak, separated the knowledge or rules from the data, and what is left to be programmed is well structured.

In the area of logic programming, as is indicated here in this PROLOG slide, Figure 5, part of the applications programming task has been taken over by something the programmer does not have to deal with. For example, in PROLOG you, the programmer or application developer ask a question. You do not write the search routine. The search for the answer goes on automatically and is not something that you have to program. So that again, the progress consists of two dimensions, one in segregating the various parts of the program more thoroughly, and a second in pushing more of the work away from the programmer at the expense of machine cycles. In other words, the search is not conducted with the skill and intelligence which it could be if you, the programmer, programmed an algorithm. However, it is left to a
hidden inference engine to do that search for you. The disadvantage of this is that the expense of this computation is often very, very large, and this had led people to the search for special machines to execute in this direction more rapidly. That, in essence, is what the Japanese fifth generation program is all about.

I have talked about the technology, including software technology, and now I want to mention the types of systems that will evolve from this.

In this period we will see workstations hooked together, perhaps through local area nets, those in turn hooked to large systems with special capabilities, we will see many computers netted together, so that the communication aspect of whatever machine you are on will be very, very important.

A typical workstation of this period, and again this is a technological projection, will be very powerful. The typical or average workstation, let us say, will be of the ten MIPS variety, with the correct sort of display and printer, and the high end workstation will be like that, only more of the same.

I am ignoring in this talk most of the intricacies of display and printer. However, I would point out that what you want in the way of a workstation printer is not what you have today, because you would really want high quality and silence as opposed to what is traditional with computer printers, which is speed, low quality and noise. In the evolution of printers and displays the royal road of progress by miniaturization is not open, because it is necessary in the end to make marks that are visible to human eyes. Nevertheless, Fig-
Figure 6. Color ink-jet printing (original in color)

ure 6 indicates that the necessary progress will be made, and that by this time you should expect a more desirable sort of printer. What I am showing you here is a picture made by a printer, it is in fact in this case and ink-jet printer in one of our research laboratories, which shoots drops of ink at the paper, is inherently very, very quiet, and as this figure shows, is capable of superb quality and color. I am not going to advocate the ink-jet technology versus some of the other competing technologies, because it is very early in the technological race for workstation printers. However, the main point I want to make is that this or several other new printer technologies should provide the appropriate printers by the end of this decade. In addition to your ten to twenty MIPs you will have, of course, the high resolution color display, and this quality of printing, and all for the sort of price that you pay today for today's run-of-the-mill workstation printer and displays. Again, this is a conservative projection. In fact, all the projections that I have ever tried to make in this area have erred on the side of conservatism. The demand seems to bring forth invention at a remarkable rate.
The typical workstation would have MIPS to burn, for ease of use. In other words, you have to consider that the cost of a machine consists of the cost of the CPU, the cost of the memory, and the cost of the. Let us say, disc, in approximately equal thirds. With a very high speed CPU provided at low cost you do not save all that money if you give the CPU the appropriate memory and the appropriate disc; there is only one third to be saved. However, if you scale down memory and disc so that you have a CPU speed that is out of proportion to memory and disc, you have in some sense cycles to burn. And the way to burn them is in the direction of making ease of use for people. So I think we are going to see cheap machines with powerful, somewhat disproportionately powerful CPU's, using those cycles for ease of use.

There are many dimensions to ease of use, but I will speak briefly about some of the more human characteristics, that is, the recognition of speech or natural language input. Speech recognition today is very limited but the algorithms are probably reasonably good. What is lacking is the sheer computer power to do the computing, and by ten years from now, approximately, I think that a very reasonable speech capability far greater than today's, and probably at the level of continuous speech can be attained. The numbers of MIPS that are required to do continuous speech processing in real time are in the 200 to 300 MIPS range. You can do continuous speech processing today actually, but, you know, you say a few words to the computer and then it chugs away for half an hour and then it recognizes the words. The number of MIPS here far exceeds the number of MIPS which I projected would be available in a single workstation CPU. This would probably be an area in which special purpose chips would be designed, special signal processing or speech processing CPU's and several of them in parallel would be a reasonably cost effective way to provide the capability that is required here. Similarly, I can only say that natural language input such as English or Danish, will be possible on a fairly large scale by this time. I think the algorithms are there. It is just they call for an enormous amount of processing, and processing is one thing we will be able to do.

The other aspect of these workstations is that they will be connected (and I think we start to see that evolution today) to larger machines to share data, which is so important in the commercial world. I think the picture you should have is that at your workstation you will call for data and you will not wonder where it is, whether it is local or remote, and it will be the responsibility of this large coordinated system to find it for you. That is not easy but it can be done.

The large systems themselves will continue to provide both a pool of expertise, because they are evolving from today's central machines, and very large MIPS, and a general back-up capability plus special printing and scientific capabilities. Most of all, they will provide access to shared data. This should evolve in such a way that you are not conscious of the existence of this tremendous power. The structure of the large systems themselves should change. I think I have alluded to the fact that the individual central processing units would probably evolve towards a level of some one hundred MIPS. However, the demand for MIPS from central processors is actually growing at a rate far greater than we can build the central processing units to cope with them, and so we are seeing—and this should continue—a definite evolution towards multi-CPU large machines. Thus, the typical large machine for the future will not be a single central processing machine even in the commercial environment but will be a multi-processor. Therefore within a single system, by this time you will be seeing a computing capability on the order of a thousand MIPS. Because of the multi-CPU configuration, this machine should stay up, even though individual elements go down.

One of the interesting directions of research is to push parallelism much further than the levels I have mentioned so far. I have these rather schematic drawings, in Figure 7, in which the processors and memories are connected in different sorts of hypothetical ways, by switches. This is a very exciting and and active area of exploration. The question is, why not have a thousand chips all lined up together, performing some remarkable degree of execution? The answer to why not is: it is very hard. It is very hard to configure a single problem for multiple executions, and I think the exploration of parallel algorithms is one of the most exciting research areas relating to computers. However, even if the various machines are working on different problems, you may still make up some price performance from a configuration of
this sort. The problems of the interaction of various CPU’s with each other are very nontrivial and need attention.

Minicomputers, too, will evolve very rapidly. They may use approximately the same CPUs as the workstations, but would be configured very differently with much more I/O and with both software and hardware that allow them to be coupled to each other. You may see multiprocessor minisystems competing with the large systems for database and other applications of that sort.

I have some remarks here about the operating systems of minis. It would be important to keep them relatively simple in order to foster the interconnection and therefore the additive power of the mini-hardware. However, the real message throughout is interconnection, that is, a lot of the utility of these machines will be enormously fostered if the software and the communications protocols and the communications software allow the effective utilization, from one workstation, of both the data and the computing power scattered throughout the network.

Already today these capabilities are evolving, even in the heterogeneous systems, that is to say systems consisting of both the engines of different manufactures and different communication nets. Data retrieval is going on across a network consisting of the machines of many manufactures and two different communication net protocols. We all know it is easier to do this in, say, a configuration of all minis, all of homogeneous architecture. However, it can be done and it will be done in this heterogeneous style as well.

Let me summarize the technical developments. First of all, at the fundamentals, which is providing the logic and memory on which everything else is built, there is almost no reason to project anything worse than progress at the rate which has characterized the industry for the last ten years. This will lead us to the sorts of MIPS numbers that I have described, through no breakthrough, but through the sheer, gritty process of making everything a little smaller. There is every reason to think that systems will become increasingly interconnected and, through the use of all those MIPS and software progress, easier to use. Because of the great MIPS power and the cheapness therefore of computation, computers will become ever more pervasive, both in business and slowly as a part of everyday life. They will become an always more powerful force.

Because of this enormous importance, computers and the evolution of computers has become increasingly subject to national and political pressures. For example, in Japan, there has been a whole series of national programs aimed at making the Japanese computer industry a world class industry. Many people have suddenly become aware of fifth generation computers, but that program is only part of a traditional Japanese approach to industry which has been followed in many other industries and very effectively.

However, as this has become more and more a public issue and a public relations issue, various other governments have decided that they, too, must do something about this vital industry, so
that we now see the Western government reactions to the work that is being done in Japan. My personal belief is that the strength of the Japanese computer industry is due to the excellent work of the Japanese individuals, who work very hard and are very good engineers, much more than to government sponsorship. However, everyone has his own beliefs on this matter.

We have recently seen in Europe the very extensive ESPRIT proposal, which is not aimed narrowly at responding to the fifth generation, which is basically special machines for logic programming which I alluded to earlier, but a more general and large scale effort to bolster the European standing in computers, which is a very natural and reasonable thing to do.

In the United Kingdom, the Alvey program is on a slightly lesser scale than the European and aimed somewhat more heavily in the direction of logic programming. And again in the United States, the United States Government, in its Department of Defense arm called DARPA, has launched a program very specifically aimed at logic programming and on a scale of the same general magnitude.

So with the increasing importance of computers, and you noticed that I project that they will become much more pervasive than they are, the governments are becoming increasingly concerned that the various national and geographical units must develop their own capabilities. This concern extends also not only to the industrialized countries where we tend to think about it, notably in Japan, Europe, the United Kingdom and the United States, but also to the newly developed or less industrialized countries, where their reaction to computers is sometimes very negative. These countries sometimes see the computer not as a device that will contribute to raising their standard of living (which it will), but as a device that will widen the gap between the industrialized and the less industrialized countries because of the head start that Europe, the United States and Japan already have.

They have reacted to this with a great deal of concern. The reactions range all the way from such things as special studies in Argentina that will be reporting to the President, to market reservation legislation in several countries. A market reservation means that you do not allow the import of foreign computers into your country, and there are various forms of it. You can have market reservation which allows for the manufacture in that country of computers, such a company being partly foreign owned, or in the more extreme cases you can demand that the company be wholly nationally owned. There are a great many experiments now being conducted in both the industrialized and in the less industrialized countries, as they try to struggle with the question of how to deal with this enormous event, and how to make it something that each country participates in.

Let me summarize. I think we have seen that almost without doubt computers will continue the present rapid evolution, which means that the computers of today will look as weak and as pale in comparison to what we can expect in ten years, as those of ten years ago look today. I project that this will continue far longer. Because of this technical evolution, mostly based on the simple fact that bits can be made smaller and smaller, computers will become extraordinarily more useful, more significant, more pervasive, and easier to use. We have seen a reaction to this a level of national concern, which is largely by the way a reaction to the present level of technology, a level which is going to change.

Whether this change will intensify the national unease, which is, I suppose, what everyone would immediately think, or whether as the computer becomes more familiar and better understood, it will lose some of its mystical and frightening properties and be regarded more as a commodity like corn and approached in that fashion, I do not know. I personally think that when it comes to discussing the impact of computers on the world, on people, on family life, as well as on nations, the non-computer expert speculates as expertly in this matter as the computer expert. Or perhaps I should say more precisely, I think all persons are equally bad at this.

I believe then that we are but a small way down a long evolutionary path. We will continue a rapid evolution of such depth and importance that its long term effects defy the imagination. We will see the consequences unfold year by year.

Thank you very, very much.