

## 12.0 Computer Hardware and Architectures

### 12.1 Circuits and Devices at the “Bottom”

The late Nobel Prize physicist Richard Feynman once wrote a provocative visionary paper that discussed the “room at the bottom.” He predicted, indeed championed, a revolution in our understanding and manipulating of micro things.

For more than two decades, the computer world has been finding and using a great deal of room at the bottom. The lines that circuit makers engrave on their silicon (and other strata) have gone from many microns wide to quarter-micron wide. Consequently, the number of transistors on a chip has risen from the tens of thousands to the tens of millions. Since small also means fast (because of speed-of-signal-propagation considerations), these circuits have become many factors of ten faster. The story is familiar to all: next year’s Pentium is 1.6 times the speed of last year’s Pentium, barring the fact that memory speed doesn’t keep up. But the net is still impressive with a conservative doubling every 3 years.

Not so familiar, except to consumers who may have the insight to wonder why they are getting their new PC disk drives so cheaply, is the revolution in storage density on rotating magnetic and optical media. Here the gains have been tracking the gains in the chip arena. Both have seen a doubling of performance per dollar every eighteen months or sooner. The machine that is being used to write this is three years old. It came standard with an 80 MB hard drive. The 1995 version comes with 500 MB.

The changes are revolutionary because they are exponential. With exponentials you can forget about the past. The main impact of exponential change is new structures, not the evolutionary change we’ve seen with the relatively constant priced PC of 1981.

Exponential change is revolutionary because it moves things from minute to immense in a short time. Our slow-moving minds, organizations, and infrastructure, accustomed to at most linear change, are overwhelmed.

Does the bottom have a hard floor somewhere? For years, people have been predicting a real floor for silicon circuit technology, and none of those predictions have come true. The latest prediction is that at about one tenth micron, circuit characteristics will become unstable and error-prone. The year may be 2003-2006. New materials will have to be found or new computer system architectures will need to be used to preserve the historical growth curve. These will be discussed later.

Magnetic disk technology is an exquisitely refined technology, driven to that state by intense competition in the industry. Magnetic bit density advance may stop its exponential growth within the next decade. Optical disk technology is not yet the replacement, in cost or performance, primarily because it has not yet received the major development that intense competition and high volume bring. Both of these technologies for secondary storage share a common architectural future: parallelism. Parallelism fits awkwardly into the traditional modes of software writing. But parallelism fits data base filing and searching very well, providing another avenue for continuation of the growth curve. We discuss this later.

### **12.1.1 New Materials**

Beyond electronic switches, what phenomena in nature can provide reliable computing (i.e., behave in a “machine-like way”)?

A molecule of DNA can be thought of as “computing” the proteins whose manufacture it guides. Experiments have been recently reported in which a classical mathematical problem was encoded as DNA’s base sequences. A modest quantity of DNA (containing almost a million million DNA molecules) performed what is in effect an analog computation. The solution was retrieved by standard genetic engineering techniques. While primitive at this stage, this kind of molecular-level computation probably will open the door to a future of immense parallel computations. The search is on for other molecules that are effective “computers.” (See discussion in High Assurance Systems, Chapter 13.)

Physicists are exploring as another form of “analog computation,” certain phenomena that are implicit in the equations of quantum mechanics. The specific class of computations under study for the so-called quantum computers is factoring. Factoring might be considered to be a rather narrow and perhaps unimportant class of computations, until one recalls that modern strong encryption methods rely on the difficulty of factoring!

Molecular computing and quantum computing are long shots—long in odds and (will be) long in coming. However, the parallelism and speed they could bring would make even today’s revolutionary change seem evolutionary.

Some effort is being invested in optical computing—information processing on “bits of light” using “light switches”. There is also work on hybrid electro-optical computing, where the physical bit representation is moved from the circuit domain to the light domain and back, to capture the advantages of the technology of switching circuits. The gains in speed that are possible using light as the medium are potentially very large, but progress here is slow. We may not see optical or electro-optical computing during our forecast period.

## **12.2 Computers-on-chips and Their Architectures**

There are two key concepts underlying predicting the future of computer systems:

1. The economics of volume production
2. Parallelism

### **12.2.1 Economics**

Why is Intel like the New York Times? Is the analogy farfetched? No. Each produces its principle product by a kind of printing process. Each achieves low cost of the product by large volume production. The revolution in computers has been essentially a printing revolution (shades of the revolutionary technology of Gutenberg).

The costs associated with producing microcomputers are extremely large. These costs form a big barrier to entry into this business. Over the next decade, we will see the number of microcomputer makers shrink to a literal handful; and most of the microcomputers made will be made by Intel.

The cost of chips will range from several dollars (for smaller or older designs) to several hundreds of dollars (for the latest, fastest). Since these costs are small compared with the value of the end product, *many* will be used *together* in various configurations to satisfy the different “sizing” needs of computer architects. Think of them as the prefab modules of a semi-custom house.

Large numbers of the latest and fastest will be organized together on very-high-speed, in-the-box networks to form a supercomputer. Much smaller numbers will provide the computing power for an engineering or graphics-arts workstation (same network-in-the-box idea, however). One or a few microcomputers will inhabit the various types of personal computers we will have. Scaled computers will have common software and I/O compatibility.

### **12.2.2 Parallelism**

Techniques of parallel processing inhabit all levels of computer design. Only the end-user is shielded (more or less) from their complexities. Parallelism is now one of the most important themes in university computer science departments and research labs.

At the chip level, microcomputer architects decompose and rearrange (with remarkable ingenuity) the basic instruction and communication processes to obtain as much concurrency of processing as possible.

As discussed earlier, the systems of the future will be networks of microcomputers built mostly out of merchant-standard components. The microcomputers will often be clustered, with very high speed communication within a cluster, and lower inter-cluster rates of communication.

The concept is very simple and the technology is likely to evolve to be relatively simple to do, once microcomputer chips are in hand. This implies that it will be relatively easy for potential adversaries to put together virtually any size computing engines that they need for command and control or for weapons systems control. Beyond military, the same is true of course in commerce.

What network architectures will be used in this style of architecture? ATM for LANs shows great promise, and experiments are being done using ATM for the internal “LANs” as well. But it will be many years (but not many decades) before the cost/benefit ratio for ATM is as low as today’s conventional networking technologies.

## **12.3 Pictures and Sound**

The foregoing discussion was built around the concept of “computer” as we have known it for 50 years. Even the PC has been a kind of “little mainframe”. But the concept is changing. The computer of the next few decades will process bits at literally billions per second, but will not be doing much “computing” (i.e., numeric or symbolic). The bits will be digital video bits, digital audio bits, and the bits from sensors (see Personal Computing, Chapter 3 and Human-Computer Interaction, Chapter 4).

Handling pictures (with speeds and qualities that people want to buy) will be the “killer app” of the next few decades of “computing.” In the future Communication Sciences Departments

will replace today's Computer Science Departments. MIT has had one of these for a long time --the visionary Media Laboratory.

Our communications have been dominated by *telephones* and *television*. In the future the device will be the *telecomputer*. This is a fairly strong prediction. It says that the device to which this era of so-called "convergence" will converge is the computer-become-video rather than the cable TV set-top box (i.e., Microsoft, SGI, and Oracle rather than Time Warner, TCI, etc.).

For architects, this view of the future implies new architectures. The chips will be largely devoted to video I/O and sensor I/O. Most of the bits/sec flowing in or out will be representing pictures, not text.

A company like Intel is predicting that the majority of its revenues in the year 2000 will not be microcomputer revenue but rather revenue from a variety of chips to service video (and other data) streams! The fiber lines to homes and offices will enable this part of the information revolution.

The new "computer/communications" architectures are only now being worked out. So the design is not yet clear—in fact there will be an evolutionary competition for best surviving design. What role will parallelism play in the new designs? What will be the special features of the telecomputer architecture that will make it easy for the software developers to innovate and keep applications flowing?

## 12.4 Some Raw Performance Guesses

These guesses are "raw" in two senses:

- a. In IT, it is difficult to make quantitative predictions too far out.
- b. As important, the concepts themselves are "raw" as we move away from familiar territory to the new systems. For example, the traditional measure, Megaflops (concerned with the speed of calculation), makes less and less sense tomorrow as the percentage of bit-processing devoted to calculation decreases drastically.

That said:

1. A billion operations per second (BOPS) in desktop workstations in 1996-97 (expensively) and 2000 cheaply. Unfortunately, using standard "rules of thumb" one gigabyte of memory is needed to keep the machine balanced. Unless we see a dramatic change in memory pricing, such a machine seems impractical.
2. Four BOPS in high performance workstations by 2000, PCs by 2002-3.
3. Supercomputers using a "network of workstations" architecture achieve a thousand billion OPS (TeraOPS) by 2000 or sooner. Such machines have a commercial and economic realization by 2010. What will hold up such designs is the relatively small demand for such large machines. Microsoft is dedicated to creating an "upsizing" environment based on clusters of commodity PCs and ubiquitous networks such as ATM.

4. Having defined the end-points of the spectrum, almost any capability in between will easily be possible because of the plug-and-play nature of network of workstations (NOW) architectures (discussed earlier). The marketplace will abound with options for customers.
5. In 10-15 years digital video on “telecomputers” will be as common as television is today.