Engineering limits on computer performance

The astounding progress in microelectronics, which has fueled the rapid growth in computer performance, is finally beginning to approach fundamental physical limits; henceforth computer architecture will become more important.

Charles L. Seitz and Juri Matisoo

Although the title of our article addresses engineering *limits*, we do not mean to convey a grim outlook. Over the past 20 years, the technology of computer systems has advanced dramatically in terms of performance, cost and reliability. There is every reason to expect this advance to continue, at a rate almost as shocking as we have experienced to date. However, the advance already achieved has pushed the mechanisms of switching, storage and communication close enough to fundamental physical limits to bring into awareness for the first time limitations in the engineering of high-performance systems.

Thus, fundamental limits may permit the size of the transistors on a silicon chip to be reduced by another factor of 10—but probably not a factor of 100. Although, in principle, a reduction by a factor of 10 would lead to an order of magnitude improvement in computer performance, mundane problems, such as how to provide wiring to interconnect these smaller devices, could prevent us from realizing much of this potential improvement.

The three basic functions required in computing systems are: *switching* (non-linearity and amplification in addition to logical operations such as provided by a transistor), *storage* of information in electronic form (a variation in stored energy, such as the quantity of charge stored on a capacitance that must be large enough to assure reliability) and *communication* of information (bringing operands into proximity, normally accomplished by some form of wire that unavoidably interposes several kinds of degradation of signals, including delay, bandwidth limits and noise). The "greed for speed" is now pushing the mechanisms providing all three of these functions to their limits in the face of competing requirements for perfect reliability of computations (if not for machines) and low cost.

The different technologies on which computer microelectronics are based have different limitations. We cannot discuss here all possible technologies, but focus instead on semiconductor microelectronics, which is today and for the foreseeable future the principal means of providing high-speed switching, storage and communication. Semiconductors are not, however, a single technology, but include both the mainstream silicon technology, and an alternative that is promising for high-performance systems, gallium arsenide (see figure 1). Even within silicon technology one finds that the engineering situation is not simple; the evolution and limitations of systems or circuits based, for example, on the bipolar transistor or on the field-effect transistor switches are quite different. "Exotic" technologies such as Josephson junction microcircuits have been demonstrated in laboratories, but are not likely to be exploited commercially in the immediate future.

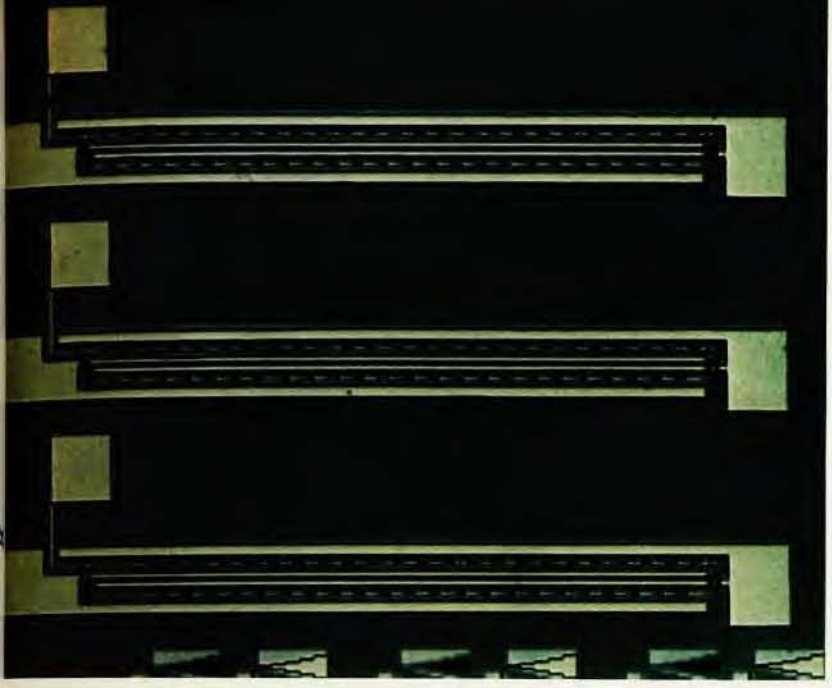
Different computer architectures also place very different demands on technology and engineering. All high-performance digital technologies provide switching speeds and energies that are comparable to the communication delays and energies. Hence it is no coincidence that the other articles in this special issue focus on concurrent (or parallel) algorithms, architectures and programming. It is not simply that computers with more parts should be able to address larger problems in less time. Computers that are organized so that their parts are less tightly coupled place less strigent demands on communication and synchronization than conventional sequential computers. The execution of each instruction on a sequential computer may require signals to traverse the entire machine. Techniques for speeding up these machines (such as instruction prefetching, execution pipelines and cache memories) allow faster interpretation of sequential programs through concurrency and localization of communication. Concurrent architectures and algorithms are approaches to organize computations to provide markedly higher levels of concurrency and localization of communications.

Semiconductor microelectronics

How does one explain the dramatic progress in the cost and performance of computers in a way that might reveal

Charles L. Seitz is associate professor of computer science at California Institute of Technology, Pasadena, California, and Juri Matisoo is manager of the silicon technology department at the Thomas J. Watson Research Center, International Business Machines Corporation, Yorktown Heights, New York.

0031-9228 / 84 / 0500 38- 08 / $01.00



Gallium arsenide MESFET ring oscillator exhibits electron mobility superior to devices manufactured in silicon. When material problems are solved, room-temperature GaAs systems are expected to outperform silicon systems by a factor of 2. Figure 1

where this technology will lead us? For the past 25 years, advances in computer hardware have been tied directly to advances in microelectronics. The products of this technology—integrated circuits, or "chips"—are manufactured by processes that, while they demand high degrees of cleanliness and control and very large investments in equipment, are extremely efficient. The entire process requires a sequence of about 100 steps, many of which are similar and use the same equipment. Each sub-sequence of several steps produces all of the transistors or wires of a given type at once over the entire chip. Many chips, in turn, are processed simultaneously on a wafer, such as the one shown on this month's cover. This highly parallel manufacturing process (similar to the manufacture of circuit boards with photolithography, but at a much finer scale) is in marked contrast to the serial assembly—component-by-component and wire-by-wire—of traditional electronic systems.

Gordon Moore[1] made the empirical observation that the number of elementary components on the most complex chips commercially available *doubles annually*. "Moore's law" held true for about 15 years, taking the single isoplanar transistor on which the early digital ICs were based as the 1959 origin. Since 1974, by which time chips with 32K components were marketed, the growth in complexity of single chips has declined to approximately a doubling each two years. It was also in the early 1970s that the character of the most complex chips changed from logical components such as gates, registers, adders and so on to system parts such as processors and memory.

Random-access memory chips are generally the leaders in the number of components on a chip for two reasons. First, they are often designed at the same time as the process that is used to manufacture them and thus have earlier access to the most advanced processes. The second reason is that their regularity allows relatively more components to be laid out per unit area than can be achieved in less regular chips such as instruction processors. RAM chips gradually displaced magnetic core storage during the 1970s. Thereafter, except for electromechanical parts such as disks, printers and keyboards, all system functions were accomplished with microelectronic parts.

RAM chips storing 256 kilobits use about half a million components and, true to the trend of a doubling in complexity each two years since 1974, were first marketed in 1982. Megabit chips with about 2 million components are expected to come into commercial use within the new year or two. Advanced instruction processors, such as the Hewlett-Packard "Focus" chip[2] introduced in 1983, do not lag far behind, with some 450 000 transistors.

Will this growth in complexity of single chips continue? By all indications from the semiconductor processes and designs being developed in research laboratories, it will continue for many more years, with some inevitable leaps and plateaus. This remarkable exponential growth in complexity is a direct consequence of a steady reduction in both the size of the "features"—wires and transistors—on chips and the density of defects. Both of these trends are easily quantified.

The number of components that can be packed into a given area is inversely quadratic in the feature size. The feature size of the earliest integrated circuits was about 25 microns. The typical feature size of today's microprocessor and storage chips is about 2 microns—a reduction in area by a factor of more than 100 for the same function (the actual gains achieved are larger, owing to the evolution of circuit and layout design styles). Functioning circuits with minimum feature size down to about 0.1 micron have been demonstrated in laboratories.

Yield varies approximately as e^{-DA}, where D is the defect density and A is the chip area. If systems could be partitioned into chips without regard for functional modularity, one would select an optimal chip area (die size) to minimize the sum of the cost of the chips themselves and their packaging. Small chips cost very little because of the large number produced on each wafer and the high yield, but they would require more packages and external connections than the same system implemented with larger chips. Very large chips have such poor yield, sometimes less than one working chip per wafer, that the chip itself can be very expensive. The earliest ICs were less than 1 mm^2, while optimal chip area today, depending on the process, is

typicaly 25 to 50 mm^2. The trend to larger chip areas may take a leap to something like wafer size if experiments with redundancy techniques to increase chip area are successful. In these experiments chips are manufactured with redundancy for all components, and schemes are devised to identify and interconnect those components that are not defective.

The product of increased circuit density and chip area, together with a lot of clever engineering in system, logic, circuit and device design, is what has led us from small-scale integrated circuits (a few gates or a "flip-flop" on a chip) to the very-large-scale integrated system, best represented by the computer on a chip. The shrink, or "scaling," in feature size also has many interesting physical advantages that we will discuss separately for the metal oxide semiconductor field-effect transistor technologies and for the bipolar technologies.

The key to reducing both feature size and defect density is advances in processing, particularly in photolithography. The large dimensions of features on early chips allowed fairly simple contact exposure, analogous to contact printing in photography, of the photoresist-covered wafer through an optical mask. Also, the tolerance for misalignment was well within the dimensional stability of the wafer.

Feature size in commercial processes is now pushing toward 1 micron, quite close to the limits of optical lithography imposed by diffraction. This dimension is only about 3 wavelengths of the violet light used to expose the photoresist. Moreover, feature dimensions must be reproduced accurately on a surface made uneven by previous layers of circuitry. The tolerable misalignment of the successive layers of a 1-micron process, not more than 0.5 micron, say, over a wafer 100 mm or more in diameter, is not comfortably within the dimensional stability of the wafers through many steps in processing. Thus, while the wafer remains the handling unit for processing, the exposure of patterns in processes approaching 1-micron feature size is accomplished by systems that step along the wafer, typically realigning on each chip, and covering total areas approaching 1 cm^2. These advances in lithography—the reduction both in feature size and defect density—have also been greatly assisted by trends away from high-temperature processing in diffusion ovens to ion-implant machines, and from "wet" etching to plasma and reactive-ion etching.

Microcircuits with submicron features are regularly produced in laboratories by electron-beam and x-ray lithography (see figure 2), and we expect these techniques to come into commercial use within a few years. It is quite easy to envision a future x-ray manufacturing technology in which scanning electron-beam systems produce x-ray masks (as indeed they are widely used today for optical masks) with the patterns transferred to the wafer by a stepping alignment and exposure system. The transition from "combat-proven" optical lithography to x-ray systems might well create a short pause in the progressive scaling of feature size; however, soft x rays of about 0.4 nm wavelength provide such excellent contrast and depth of focus that one might anticipate a leap to feature sizes of a fraction of a micron. The engineering problems of manufacturing electrically functioning systems with such small features are formidable, but we believe they will be overcome.

Beam line for x-ray lithography installed by IBM on Port U6 at the National Synchrotron Light Source at Brookhaven National Laboratory. Figure 2

Scaling MOS technology

The active devices of MOS technologies are field-effect transistors, which were at one time fabricated as a metal control electrode, called the gate, over a thin oxide layer on the surface of the semiconductor, nominally single crystal silicon (see figure 3). Polycrystalline silicon doped to be a conductor has replaced the metal in modern processes as the next level of wiring above the silicon, with one or two levels of metal wiring above the "poly" level. To discuss the consequences of scaling the dimensions of a MOS transistor, one must understand that it is built along the surface of the silicon, rather than down into it as with bipolar devices. In particular, the length of the "channel" under the gate is determined by the feature size of the process.

Let us examine the consequences of shrinking such a device and the wires attached to it by a factor of 10—from a feature size of 25 microns (typical of early integrated circuits) to 2.5 microns (typical of today), or from 2.5 microns to a future technology of 0.25 micron. The same scaling factor is applied in all three dimensions, retaining the same relative flatness of the surface, and the electrical field patterns within the device are simply scaled down. A simple analysis[3] shows that the voltage (V) and maximum current (I) both scale linearly: $V \rightarrow V/10$, $I \rightarrow I/10$. This result is very fortunate because then the power per device (VI) scales down quadratically ($VI \rightarrow VI/100$) while the number of devices per unit area scales up quadratically, and the power per unit area remains constant. The transit time (τ) scales linearly ($\tau \rightarrow \tau/10$)—that is, the smaller switch is also faster.

The fundamental figure of merit for switching devices, the *switching energy* (E_{sw}) is the energy required per switching event. This quantity is clearly the same as the product of the power per device (when switching at maximum speed) and its delay, and hence is also called the power–delay product. Total power is an excellent predictor of system cost, while each switching delay contributes to the reciprocal of performance. Thus the switching energy of the technology used to build a computing system is a good predictor of cost/performance of the system in which the switches are used. More fundamental-

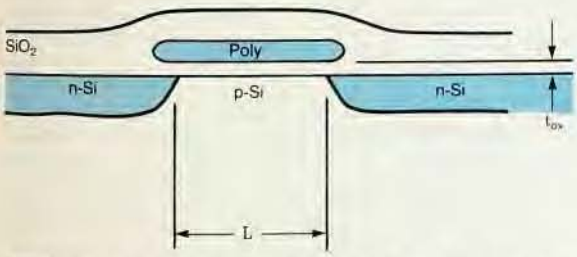
ly, it is proportional to the cost/performance of a computation averaged over numerous elementary switching events that make up the computation.

We can now appreciate fully the underlying reason for the remarkable advances in the cost/performance of systems based on MOS technologies, and the promise inherent in the next factor-of-10 reduction in feature size. The switching energy (the product of a quadratic scaling of the power per device and linear scaling of the delay) scales as the third power of feature size $E_{sw} \rightarrow E_{sw}/1000$ (assuming the electric fields remain constant in the scaling). Our example of a future MOS technology with 0.25-micron feature size is very much a replica—scaled in size, energy and speed—of today's MOS technology. However, this technology is close enough to fundamental limits that another ten-fold decrease in feature size and voltage is questionable. One would start encountering switching energies that are too small a multiple of the thermal energy kT, as well as appreciable tunneling through the very thin gate oxide layers, and statistical fluctuations in threshold voltages across many devices—owing to a statistically small number of the impurity ions that determine the threshold voltage.

One may achieve more or less than the full advantage in performance per cost predicted by this idealized scaling depending on a number of important details, but they are largely details. Overall, we can thus expect an improvement in the switching energy by three orders of magnitude from a decrease in feature size by a single order of magnitude.

The second-order effects, both good and bad, are often important in the practical engineering of systems. For example, most of the signal energy in typical MOS systems goes to drive the parasitic capacitance of wires rather than the transistor gates. When a decrease in feature size allows one to consolidate a system (such as an instruction processor and a number of storage units onto a single chip), the area and power required to drive the parasitic capacitance of all of the package pins and interchip wires is reduced. On the other hand, shrinking the wires on a chip reduces their cross section (quadratically) and increases their resistance per unit length (quadratically), while the capacitance per unit length remains approximately constant. At submicron feature size, the resistance of a few mm of metal wire is so large that the wire cannot be treated as an equipotential; rather, diffusive propagation of signals becomes the rule. One can solve this problem either with repeater amplifiers on long wires, or by using additional thicker layers of metal interconnectors. Another curious problem is that leakage and subthreshold currents do not scale with the other currents, but as $e^{-(V_t/kT)}$, which makes some of the MOS techniques of storage by isolating charge on a capacitance relatively less attractive. One would have to scale temperature to make the scaling exact.

In addition, the reader should not conclude that this future 0.25-micron MOS technology allows one to achieve a gain of 1000 in performance at constant cost with the same computer designs. The cube-law scaling of the switching energy is the product of a quadratic scaling of area and power per function, and only a linear scaling of transit time. One finds in reality that this linear transit-time scaling is very difficult to achieve in larger systems because of delays imposed by long wires, unless system parts can operate concurrently. For conventional sequential systems it is accordingly much easier to exploit advances in MOS technology to reduce cost than to enhance performance.



MOS field-effect transistor in cross section. *L is the channel length, t_{ox} is the gate oxide thickness. Millions of these devices are* manufactured at once on a wafer by first oxidizing the p-type silicon to a thickness several times t_{ox} *and then patterning channels in which* the oxide is t_{ox} *thick. The oxide channels become the first layer of* interconnection and form two of the transistor terminals. Polycrystalline silicon ("poly") is then deposited and patterned to form the second layer of interconnection and the gate electrode of the transistor. Next, a phosphorous implant or diffusion converts the channel areas to n-type conduction, except under the poly, thus forming a "self-aligned" transistor. Finally alternate layers of oxide and metal are deposited; the metal layers are patterned as wiring layers and are interconnected through arrays of holes in the oxide layers. Figure 3

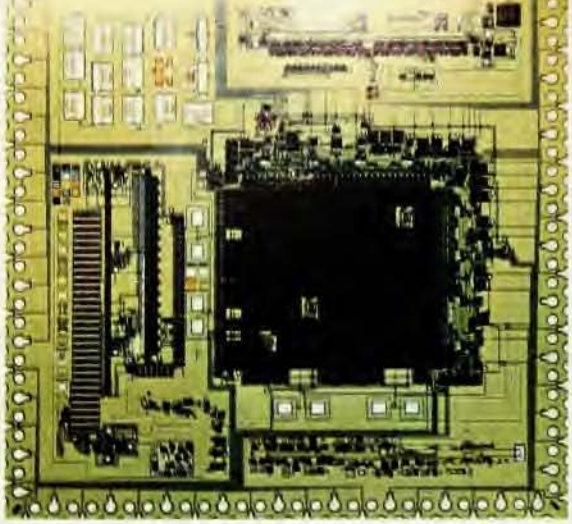
Scaling bipolar technology

On the basis of equivalent feature size, high-performance bipolar circuits have an advantage in speed of about a factor of five over MOS technology. However the integration levels of bipolar chips historically have lagged behind that of MOS chips by about an order of magnitude. Some of this lag arises from problems in cooling. The remainder of the lag results from the differing sensitivities the two technologies have to defects and to defect types.

Consequently, bipolar devices and circuits are used in those portions of computers in which performance is a prime requirement. Generally, these are the processor, the cache (which is the highest-performance portion of the storage subsystem), and that part of the communication system in which high data rates are required (the channels). Indeed, bipolar technology is currently pervasive in all computing systems that emphasize performance, and nearly all midrange systems. As feature size continues to shrink for both MOS and bipolar technologies, one expects the MOS technology to make continued inroads from the low-performance end into the midrange arena, which is currently the province of bipolar technology.

High-performance computers achieve their performance by using both high-speed circuitry for executing functions as rapidly as possible and, at the same time, a prodigious number of circuits to perform complex and concurrent operations. This design strategy, coupled with the relatively low level of bipolar integration, leads to the use of a large number of chips in a typical high-performance computer. The number of chips places strong demands on packaging technology and leads to a design style amenable to extensive automation in logic-circuit placement, wiring and testing. This style, in which a matrix or array of devices is fabricated, has been in common use for more than a decade in the bipolar world and is now also becoming popular in the MOS world. State-of-the-art bipolar gate arrays contain several thousand circuits on a chip with logic delays under loaded conditions of typically about 500 psec (or 250 psec for the circuit with essentially no load). A chip of this size dissipates about 5 watts.

Bipolar technology can be scaled to much narrower linewidths than the current feature size of 2.5 microns,



Memory cache array of Josephson junctions fabricated using a lead-alloy technology having a minimum dimension of 2.5 micron. Figure 4

much in the same way as MOS technology; however, the details of the scaling differ. The major difference results from a fundamental difference in operation of MOS and bipolar devices—namely, the bipolar voltage levels are effectively set by the silicon band gap, so that bipolar scaling leaves the circuit voltage levels essentially unchanged. The scaling principles of bipolar devices and circuits have only recently been elucidated.[4] The scaling procedure begins with an optimized high-performance bipolar circuit, such as a current-switch emitter follower. As the lithographically determined horizontal dimensions are shrunk, the vertical dimensions of the emitter and base are also shrunk. To avoid problems with device operation it is necessary to increase the base doping substantially and to adjust the collector doping in proportion to the emitter current density. As the linewidth under the coordinated horizontal and vertical scaling is shrunk from the present 2.5 microns to 0.25 micron, the relative delay contributions of various components remain approximately constant and the overall circuit delay reduces by approximately one order of magnitude, going from about 250 psec to 25 psec. Thus, from the point of view of performance, there is still considerable mileage left in the silicon bipolar technology.

To obtain a corresponding increase in the level of integration one must address the critical problems of device isolation and interconnection wiring. Fundamentally, these issues are the same for the MOS and bipolar technologies. From the point of view of wiring, the problems are those of signal propagation and electromigration (atomic motion caused by the flow of electric current) which can lead to catastrophic opens in conductors. Both of these problems are alleviated by retaining as large a cross section for the wires as possible, which runs counter to the need to have a large number of wires to enable all of the thousands, indeed hundreds of thousands, of circuits to be interconnected to perform the appropriate functions. The natural solution is to stack many wiring planes on top of one another. For example, current bipolar arrays of several thousand circuits use four levels of interconnection wiring.[5] One envisions an overwhelming need to increase this number greatly to make usable the levels of circuit integration one postulates from device scaling alone. Indeed, the major technological stumbling blocks to very high levels of integration are likely to be these seemingly mundane problems of interconnection wiring and device isolation, rather than the more intellectually appealing issues of device and circuit scaling.

Underlying all the projected improvements are major technological problems. Lithography must advance, and processing techniques for ever-shallower vertical structures must be devised. None of these issues is trivial, and major worldwide efforts are required to continue to satisfy Moore's law, or even to continue along a shallower curve. The economic forces underlying the computer industry are likely to assure an adequate effort in the further development of silicon technology and, to a lesser degree, in the search for alternatives.

Systems of chips

A host of major engineering problems arises when many chips are assembled to form a computing system. For performance, it is desirable to pack these chips into as small a volume as possible to minimize the distance, and thus the delay, for signals. However, the need for adequate cooling, power distribution and interconnection tends to force volumes in the opposite direction. The end result is the usual engineering compromise, largely driven by cost comparisons of alternative solutions.

A large variety of geometrical arrangements for packaging a collection of chips currently exists in the computer industry. At the very-high-performance end, the choices are somewhat more limited and two types of arrangements are prominent. One arrangement places single chips into a carrier, referred to as the module; these in turn are mounted on a card, which in turn plugs into a board. Such a configuration is commonly found in personal computers. The second approach, typifed by current IBM high-end processors, places about 100 bare chips on a module that provides cooling, power and interconnections. These large modules, containing about 100 000 circuits (and thus likely to be major functional units of the computing system), are in turn plugged into a board that provides for module-to-module signal interconnection and distributes the necessary power.

One must provide an adequate number of terminals to transmit signals from one chip to another. The physical dimensions of these terminals should be kept small, not only because several hundred may be required, but also because the inductance discontinuities presented by these structures must be kept as small as possible. A common IBM practice is to cover the chip with an array of small solder balls that provide the devices with both structural support and electrical connections. Other manufacturers employ peripherally distributed input–output pads and connect these pads to the module carrier by a variety of techniques. The chip interconnection wiring is also accomplished in a variety of ways. IBM, for example, uses ceramic sheets onto which is deposited interconnection wiring of relatively gross dimension. These sheets are then stacked in as many as 30 to 40 layers to provide sufficient wiring to interconnect the 100 or so chips. The major technology issue is wiring yield and provision of adequately low resistance and a controlled characteristic impedance.

The chips are supplied with current by a power distribution network capable of carrying several thousand amperes at the board level at a few volts. The major engineering problem results from the varying load presented to the power supply by the switching of the

myriad logic circuits. A large number of circuits switching at the same time on any given chip can significantly depress the voltage level of the power supply. Portions of the computing system will become inoperative until the power-supply voltage recovers to its design value. The design problem, therefore, is to make the inductance of the power distribution system as low as practical, within the constraints imposed by the technology and the physical dimension, and to provide as much local "ballast" as possible in the form of capacitive energy storage in the vicinity of the chip. The power supplied is converted into heat, which then must be removed to keep the chip temperatures within tolerable operating limits.

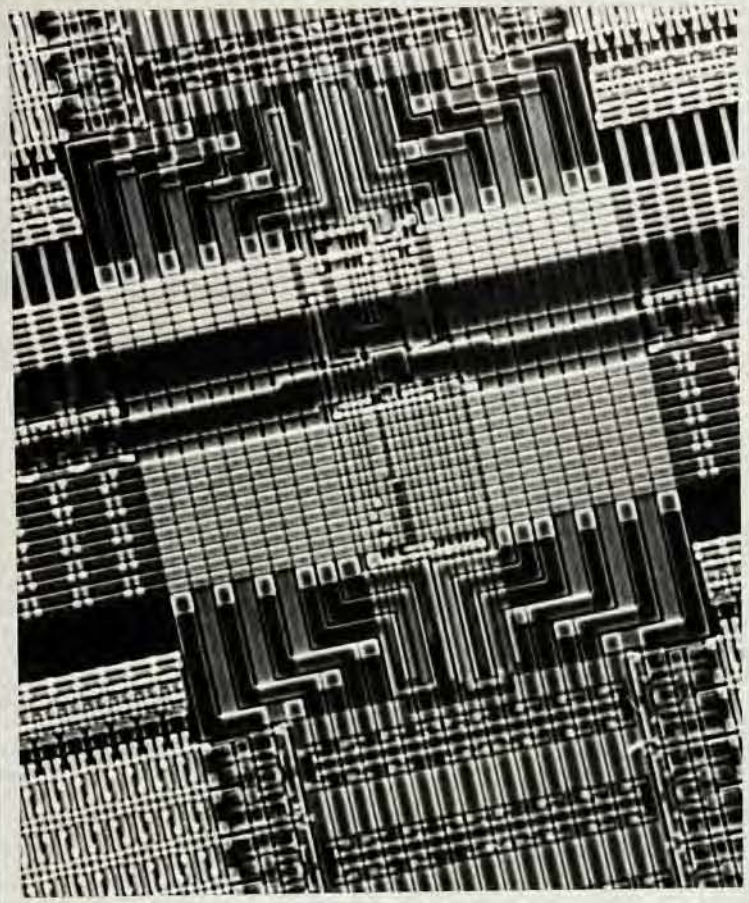
Roughly, a few watts per chip can be removed simply by air cooling; greater power levels require special cooling techniques. Again, there is a vast array of approaches. In the piston cooling approach used in the IBM module, heat is conducted through the silicon chip to a piston that conducts the heat to a water-cooled plate. This approach is capable of removing on the order of 10 watts per chip. More exotic techniques[6] (not yet practical) in which cooling channels are etched into the silicon chip and coolants are circulated through these channels have the demonstrated capability of dealing with several hundred watts per chip with minimal temperature increase.

One additional problem is worth noting. In a physically extensive system whose dimensions are large when measured in units of pulse risetime, width or even period, the problem of assuring that timed events occur precisely is not a minor issue. All present-day computing systems require synchronization to assure that certain events will occur essentially simultaneously. Synchronization is conventionally achieved by distributing timing pulses from a central source. All necessary timing pulses must arrive at their destination within required time intervals. The clock-signal distribution system requires careful design and an adequate safety margin.

Thus the assembly of multi-chip systems allows for a large variety of engineering solutions to a common set of problems. Unlike the tidy world of silicon, a material much studied and about which a great deal is known, packaging can make use of a very large set of materials and processes and historically has done so. In the future, chip assembly will likely follow directions already pioneered by the silicon-chip technology.

Storage subsystem

Many computationally demanding problems are limited as much by the size of the working set (data) on which a



Central Intersection (magnified 400 times) of new experimental, high-speed chip providing 64-k bits of memory.

Experimental high-speed memory chip

Typical of ongoing advances in microelectronics is the new 64 kilobit computer memory chip developed on an experimental basis at IBM's T. J. Watson Research Center. The new chip provides access times of 16–20 nanoseconds, compared to 70–300 nsec for other 64K chips. The improved performance is achieved through new design concepts and innovative circuitry, including:

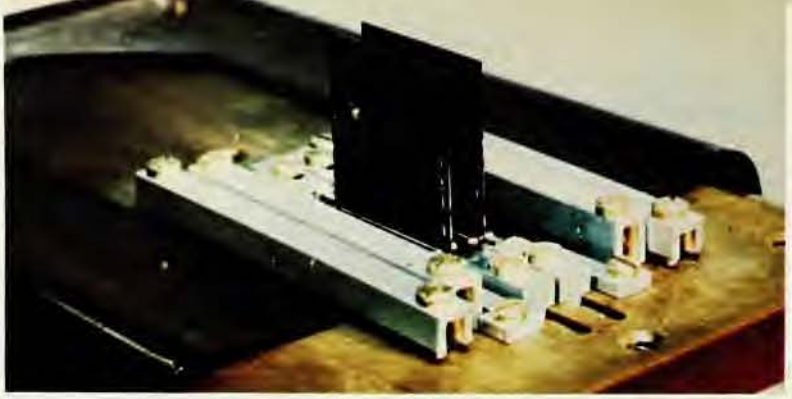
- A self-timed sensing circuit that permits much faster chip operation than is possible with conventional circuits.
- An improved address buffer that allows the chip to read information quickly. The address buffer detects the location of data to be retrieved from the chip.
- Simpler timing and a reduced number of clocks required to keep the chip operating properly.

In addition to its faster access time, the new chip can discharge information in blocks of 16 bits at a time. Most 64K chips have an output of one to eight bits at a time. Also, the new chip uses a four-device memory "cell" instead of the commonly used single-device cell. Although the four-device cell is larger than the conventional unit, it is faster.

The chip is 4.5×7.2 mm (roughly $\frac{1}{4}$ inch on a side). The average minimum feature size is 1.7 micron, and the channel length is 1.2 micron. Memory cells are 292 microns square. The technology used is n-channel metal oxide semiconductor, field-effect transistor with a single level of metal and a single level of polycide.

Physically, the chip consists of four 16K blocks, or quadrants, of memory cells with associated row and column decoders in adjacent blocks. During access, each 16K block has one row decoder and four column decoders activated, providing an *"x–y" grid to* access bit locations.

There are four self-timed sense amplifiers and data-out buffers per quadrant—one for each pair of input/output lines. They are located on the periphery of the chip. The bit lines are separated from the I/O lines and sense amplifiers only by the bit line switches. The sense amplifier is set from the accessed wordline. This self-timing sense amplifier minimizes timing skews that would result if a separate timing chain were used.



Josephson test device is composed of four circuit chips assembled on two cards plugged into a microsocket board in a high-density, 3-dimensional, card-on-board package. The device was successfuly operated with a minimum cycle time of 3.7 nsec. Figure 5

program operates as by the number of operations per second that the machine can deliver. One can reasonably expect that advances in microelectronics will continue to decrease the cost per bit of the primary random-access storage of mainframe computers; however, this figure is far from the whole story.

The storage of mainframe computers, as well as of the less-conventional concurrent computers, is usually organized into a hierarchy of increasing size and access time. Typically, the data that can be accessed fastest are in a computer's internal registers. Most high-performance computers next use one or more levels of "cache" memory, typically a few thousand words (see figure 4). The computers are organized to intercept those accesses to primary storage locations that are duplicated in the cache and store or return data in a fraction of the primary storage cycle time. This technique is effective in making the average storage-access time close to the cache-access time as long as the "miss rate" is much less than the ratio of cache time to primary storage time. (The miss rate is the fraction of times the cache memory does not contain a needed storage location, because of a branch in the program or other cause of a large discontinuity in storage location sequence.) There are many strategies for deciding which data to keep in the cache, all of which depend on regularity or locality of access. Fast storage is expensive per bit, and the cache scheme allows one to achieve the effect of a large primary memory whose cycle time is on average only slightly more than that of the cache.

Next one finds the primary storage (typically a million or more words) that could well be thought of in systems using "virtual memory" as a cache for the computer's true primary storage, almost invariably disks. (In a system using virtual memory, the logical memory address is related to the physical address by a scheme enabling that substitution of slow memory space for fast memory space in ways that present a logical fast memory to the programmer that is much larger than the physical fast memory.) The ratio of access times between these levels is so large (on the order of 10^{-4} to 10^{-5}) that the miss rate to primary storage must be much lower than people can achieve in general-purpose systems. As a result, a substantial fraction of the programming effort for many large computations is directed at spooling data in and out from disks so it will be available at the right moment, and treating boundary problems when the storage will not accommodate the entire problem at once and "regions" must be dealt with in sequence. What is really required is a storage medium that bridges the gap in cost and access time between the primary random-access memory and disks. However, we know of no help in sight to satisfy this requirement.

Several technologies are involved in the memory or storage subsystem of the computing system. Generally, the cache employs the highest-performance technology—bipolar. Cache sizes tend to be relatively small because of the high cost per bit. The next level of the memory hierarchy is therefore a cheaper and slower semiconductor store, usually consisting of MOS static or dynamic memory chips. Indeed, there may be a number of such levels in the hierarchy, each progressively slower in yielding the stored information and each progressively larger in size until one reaches the permanent disk storage.

Storage reliability is also an issue that requires tradeoffs between technological and organizational solutions. For example, information stored as charge on a capacitance built into a microelectronic chip is less liable to destruction by an alpha particle or high-energy neutron if the capacitance is larger. But larger capacitance requires more energy and, at the same power, more time to switch.

Alternatively, one might accept a certain rate of "soft errors," and encode stored information with an error-correcting code. (Soft errors refer to temporary losses of information, such as those caused by ionized particles.) Error correction requires that more bits be stored, and error-correction logic introduces delays and higher costs.

Many detailed engineering tradeoffs revolving around the cost of available technologies are made in the design of a storage subsystem for a particular machine. Ideally, one would like to have a very-high-capacity, nonvolatile storage subsystem capable of yielding information to the processor on each cycle. Currently, such a storage would be prohibitively expensive. Instead, a number of technologies are now used to optimize cost/performance.

Novel technologies

The silicon technologies—bipolar and MOS—have been, are and will continue to be the major integrated-circuit technologies, and they can be extended from the current state of the art, both in performance and level of integration. However, the search continues for alternative approaches that may offer better ultimate performance. Gallium arsenide is a semiconductor material that exhibits electron mobilities substantially superior to those in silicon. This intrinsic advantage can be translated to a performance advantage at the device and circuit level. Indeed, there have been numerous demonstrations of the capability of gallium arsenide circuits in various device configurations (see figure 1).[7,8] Unloaded delays of about 10 psec have been demonstrated in an exotic variant of the gallium arsenide technology, called the high-electron-mobility transistor, which is made from GaAs/AlGaAs structures.

The intrinsic performance advantage of GaAs devices over silicon devices is smaller on a loaded circuit basis and smaller still when systems are assembled from a large collection of chips. The hope is that room-temperature GaAs systems will ultimately have a performance advantage of about a factor of 2 over silicon technology. Whether a GaAs technology actually achieves this potential performance strongly depends upon the rate of technological development and, in particular, the rate at which the level of integration (which is currently low) can approach that of silicon technology. A considerable number of materials-related problems stand in the way. There is some cause for optimism based on recent progress and the increased resources for development the technology has attracted.

An even more exotic technology is the Josephson junction, based on tunneling phenomena observed in a metal-insulator-metal sandwich when the metals are superconducting and the insulator is sufficiently thin to permit electron tunneling. Over the last decade this technology has received considerable development, primarily at IBM, and has reached the state where practical logic circuits as well as a packaging approach have been demonstrated experimentally (figure 5). On the logic-circuit level, (loaded) logic delays as small as 10 psec have been attained. In addition to the incredible performance of the Josephson switches, the technology offers the less obvious but equally important advantage of superconducting transmission lines as lossless wires. From the point of view of performance, the Josephson technology is perhaps the ultimate.

The major difficulty with the technology revolves around the need to refrigerate the circuits to 4 K, rendering single-chip applications prohibitively expensive. Consequently, one is faced with an "all or nothing" situation in which the first computing system developed must contain all of the subsystems, placing the considerable burden of a very large entry price on the technology. Consequently, commercial exploitation of the technology for high-performance computing systems appears feasible only in the relatively distant future.

High-performance architectures

How can the expected advances in circuit technology be put to work in high-performance computers? Because each technology offers its own advantages and inflicts its own limitations and idiosyncrasies, one must expect that there are many possibilities.

It is not a radical extrapolation to predict that the performance of present-day scientific supercomputers, say that of the Cray-1, will become available at the price of common mainframes, or someday even as desktop models. The Cray-1 is extraordinarily well thought-out and engineered; it attains its high performance partly by fast circuitry and partly through concurrencies in the process of interpreting sequential programs, including pipelining of the arithmetic. Except for its primary storage, the Cray-1 employs about 300 000 bipolar circuit packages of relatively low complexity. Because the design is already very well partitioned into functional elements, we can envision that a Cray processor of similar performance could be built at reduced cost using a higher-complexity bipolar or a gallium arsenide technology. In fact, a machine in the same style as the Cray-1 could be achieved even in MOS technology with a feature size of 0.25 micron. The performance of the switching elements in this ultimate MOS technology is quite comparable to that of the present Cray-1, and the complexity of single chips in this technology would be so high that only 10 to 20 would be required.

Improvements in microelectronics are more readily translated into reductions in cost than in advances in system performance. The prospects of achieving, say, one, two or three orders of magnitude advance in computing performance over today's supercomputers are good, but will certainly require progressively more radical approaches than repackaging existing system designs in new technology. An additional single order of magnitude is probably about as much as one can reasonably expect for sequential computers. From the standpoint of physical engineering limitations, one can see that this particular evolutionary path is beginning to experience diminishing returns in performance for cost and design effort. This expectation was recently documented[9] by Bill Buzbee of Los Alamos, who notes that the rate of improvement in computations per unit time in high-performance computers has been decreasing smoothly to an asymptote at about 3 billion floating-point operations per second.

Even to achieve this asymptotic performance, sequential supercomputers must be made still smaller, because the time for signals to propagate along wires has become a major fraction of the clock or operation cycle. Of course, supercomputers built from chips of higher complexity and with more compact packaging are steps in this direction. However, the design and engineering problems are formidable.

The other route to high performance is to exploit concurrency more overtly—not only in the interpretation of sequential programs but also in employing concurrent algorithms for the solution of large scientific problems. Concurrent machines can bypass engineering difficulties in that large, high-performance machines are composed of smaller machines of relatively lower performance but with an attractive ratio of performance to cost.[10] This approach places simpler demands on the engineering at the expense of system and application software.

The architectures that are the most direct extrapolation of present supercomputers are shared-storage multiprocessors. When these machines are relatively tightly coupled through the storage, the complexity of the switch between multiple processors and multiple storage elements does not scale attractively with the number of elements, and the number of processors is limited to a fairly modest number, say 16 to 64. Existing multiprocessors follow this scheme. However, much larger machines are possible if each processor includes enough local memory or cache to allow the switch a latency exceeding the processor instruction cycle. Denelcor's Heterogeneous Element Processor and Bolt, Beranek and Newman's Butterfly machine are commercial examples of this architecture.

Machines based on a message-passing model of computation (the components communicate like computers in a network) are still more loosely coupled, so as to have few engineering limitations on size even up to millions of computing elements. The practical size limit of these machines may well depend on reliability—the mean time between failure and mean time to repair must be small enough to permit the system to compute for a useful fraction of the time. Such machines operate efficiently only for problems for involving a comparably large concurrency in the computation and requiring sparse or localized communication. VLSI-inspired architectures such as the Caltech ensemble machines[11] and systolic arrays[12] are examples.

Thus, there are no fundamental obstacles for concurrent—as opposed to sequential—supercomputers to achieve an essentially open-ended range of performance. These machines can exploit advances in the circuit technology particularly easily, so that their performance will be cost-effective for those problems that lend themselves to concurrent execution.

References

1. G. E. Moore, *Proc. Caltech Conf. on Very Large Scale Integration*, California Institute of Technology, Pasadena, California (1979).
2. M. Canepa, E. Weber, H. Talley, VLSI Design Magazine **4**, Jan/Feb 1983.
3. C. Mead, L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, Mass. (1980).
4. T. H. Ning, D. D. Tang, P. M. Solomon, Technical Digest, International Electron Devices Meeting, 1980, page 61.
5. S. Brenner *et al*, *Proc. 1983 IEEE International Solid-State Circuits Conference*, IEEE, New York, NY (1983).
6. D. B. Tuckerman, R. F. W. Pease, IEEE Electr. Dev. Lett. **EDL-12**, 126 (1981).
7. R. C. Eden, A. R. Livingston, B. M. Welch, IEEE Spectrum **20**, December 1983, page 30.
8. H. Morkoc, P. M. Solomon, IEEE Spectrum **21**, Feb. 84, page 28.
9. B. L. Buzbee, IFIP Congress, Paris, Sept. 1983. Los Alamos National Laboratory, #LA-UR-83-1392.
10. C. L. Seitz, *1982 Conf. on Advanced Research in VLSI, MIT*, Artech Books, Dedham, MA, 1982.
11. C. L. Seitz, J VLSI and Computer Systems **1**, no 2 in press.
12. H. T. Kung, Computer Magazine, January 1982, page 37. □