New Memory Technologies

Jan A. Rajchman

The multifarious capabilities that have made the computer the great success of our age are due to exploitation of the high speed of electronic computation by means of stored programs. This process requires that intermediary results be stored rapidly and furnished on demand for long computations, for which high speed is worthwhile in the first place. Storage devices or memories must have capacities sufficient not only for the intermediary results but also for the input and output data and the programs. The demand for fast access and large capacity has grown constantly. Today there are memories accessible in tens of nanoseconds and memories with more than a billion bits, but in general fast memories have small capacities and memories with large capacities have slow access.

Ideal would be a single device in which vast amounts of information could be stored in nonvolatile form suitable for archival record-keeping and yet be accessible at electronic speeds when called for—perhaps a shoe-box device containing 10^{12} bits accessible at random in nanoseconds. So far there is no way to realize this ideal. Fortunately, the benefits of large capacity and rapid access can be obtained by use of a hierarchy of different types of storage devices of decreasing capacity and increasing speed. The main hierarchy today comprises, on one hand, large-capacity magnetic recording devices, which are accessed mechanically and serially (reels of tapes, disks, and drums), and on the other hand, fast electronic memories (the core memory and various types of transistor memories). The user may be unaware of the hierarchy, as the distribution of information between the various devices is built and programmed into the system through remarkable software artifices in computer architecture, memory organization, and internal operating programs. An outstanding shortcoming of the present state of the art is the gap (*1*) between the mass-memory storage devices and the electronically addressed memories. The spanning of the gap by software is extremely clever—but it entails not only labor and complexity but elaborate internal programs that preempt much of the memory capacity and tend to slow the processing.

This article deals with electronic memories and new developments in this area (*2*, *3*). In particular, the challenge of developing an electronically addressed memory that fills the gap between the present mechanically accessed magnetic disk memories and fast transistor memories is considered.

Core Memory

One of the first electronic memories was a circulating delay line, a signal transmission device in which the output, properly amplified and shaped, was fed back into the input. Although it was economical, it had the inherent drawback of serial access: the greater the capacity, the longer the average access time. What was really needed was selective access to any stored data in a time that was both as short as possible and independent of the data address or any previous access. This is known as random access, so named to emphasize the total freedom of accessing and therefore of branching (following one or another part of a program), which is indispensable in the execution of stored programs. The first random-access memories (RAM's) were electrostatic storage tubes. In the early 1950's the core memory (*4*, *5*) replaced these early devices, providing a solution to the need for random access that truly fired the emerging computer industry into its fabulous growth.

A core is a tiny ring-shaped piece of magnetic material in which a bit of information is stored by magnetizing it in a particular direction, that is, to one or the other of its two remanent states. The core memory consists of an array of cores strung on fine wires. To retrieve information from a core a current is sent through it that causes it to be magnetized in a given direction that is the same as or opposite to the direction in which it had been left. If the magnetic state changes, a voltage is induced that provides the readout and at the same time a signal for restoring the core to its original state. To select one core in an array of many, two current pulses are applied simultaneously to one row and one column of the array, and the core responds only to the coincidence of the two excitations. This is possible because the ferrite cores used have a square hysteresis loop and are very uniform. Individual cores are made and sorted automatically, hundreds per second, and then are strung on wires, generally manually. During about a decade (1955 to 1965) many interesting magnetic structures were developed in which the elements were made by a batch process rather than handled individually (*6*). It turned out, however, that the ratio of electronics to magnetics in the system had a far greater influence on performance and cost than the nature of the magnetic structure. Hence the established and constantly improving core technology prevailed.

The core memory has become the main internal computer memory and was used universally until challenged recently by semiconductor memories. Typical are memories with 1 million words of 30 to 60 bits each, randomly accessible in 1 microsecond. The core memory has also been extended to very large capacities, of the order of 100 million words, and has proved the benefits of filling the gap in large systems, that is, software simplification and more efficient computation (*6a*). However, this gap-filling solution is too expensive (about 1 cent per bit) to be of general use. The core memory is still dominant today. From 1966 to 1971 the annual production of cores in the United States increased from 10 to 100 billion, and it has probably again increased tenfold from 1971 to 1976. The market outside IBM is estimated to be more than 300 billion cores.

Semiconductor Memories

Since the mid-1960's the spotlight has shifted from the core memory to semiconductor memories. This has resulted from the development of large-scale integration (LSI) techniques for silicon that permit the mass fabrication of microscopically scaled arrays of transistors and all their interconnections through a single set of batch processes. This revolutionary development has had a profound influence on the whole elec-

The author is retired vice president for information sciences, RCA Laboratories, Princeton, New Jersey 08540, and visiting professor at the Electrical Engineering and Computer Science Department, University of California, Berkeley 94704.

tronics industry and is the subject of another article in this issue (*7*).

It took many years for the semiconductor industry to reach its present proficiency. The first impact on memories was the replacement of tubes by transistors in circuits associated with core memories, which resulted in savings in power, space, and eventually cost. This led to modern core memories with high ratios of electronics to magnetics—and incidentally to general knowledge of memory systems that is still basic today. The first integrated semiconductor memories appeared in the mid-1960's (*8*). Early transistor memories used static cells with bistable states, each made of a flip-flop circuit with two branches so connected that the "on" state of one ensured the "off" state of the other. At first, relatively high power (about 1 milliwatt) bipolar transistors were used, with which very short access time (less than 100 nanoseconds) could be reached. Hence bipolar memories were used in conjunction with cores, as fast "scratch pads," a hierarchy concept that has been expanded in today's architecture, where fast "cache" memories are used with slower main semiconductor memories. However, the real impetus in the area of semiconductor memories came with the development of unipolar field-effect transistors of the metal oxide semiconductor (MOS) type, which made possible low-power cells (typically 50 microwatts today). In bipolar transistors conduction is by both electrons and holes, whereas in MOS's it is either by electrons (n-type transistors) or by holes (p-type transistors). (Complementary MOS's, however, are made of both p-type and n-type transistors.) The LSI technologies for the two involve similar processing and pattern-forming steps. In general, bipolars are used for high-speed and MOS's for high-capacity memories.

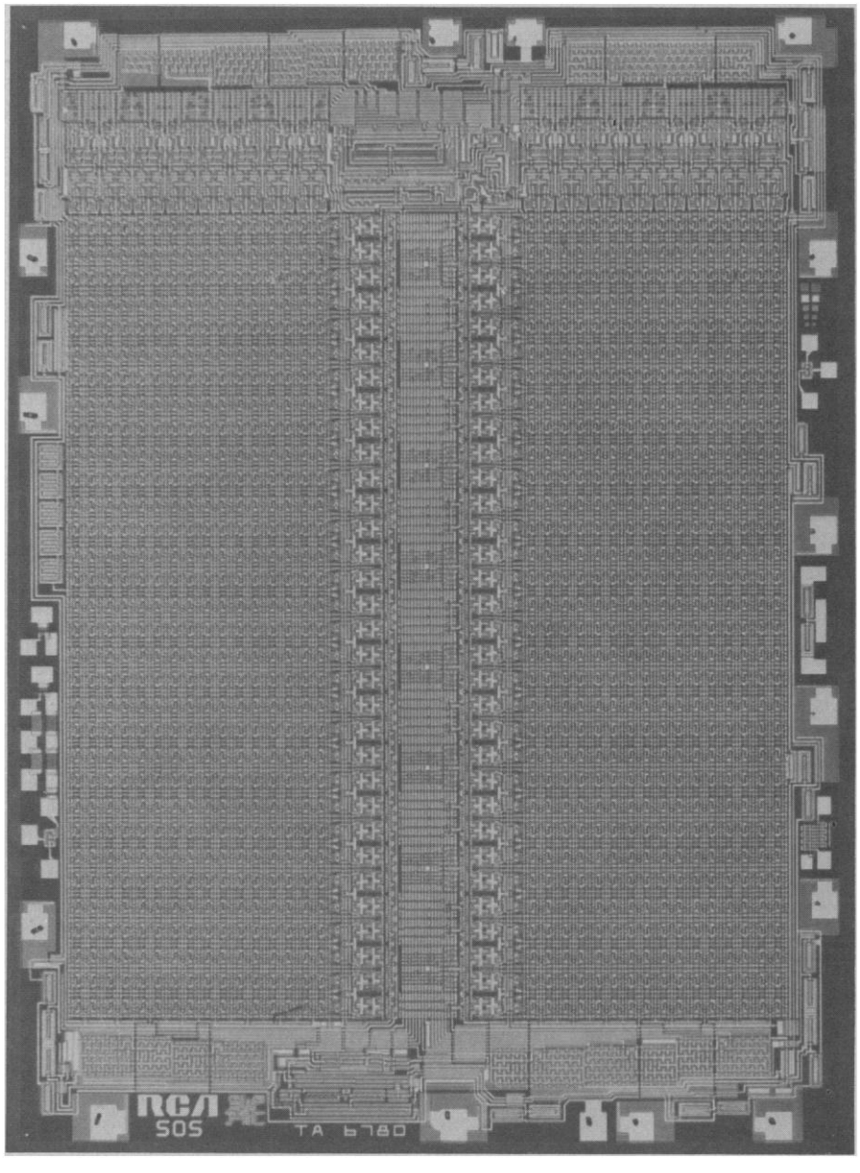

Fig. 1. Example of a 1K RAM on a chip measuring 37 by 51 mm. This RAM has static six-transistor cells, an access time of about 100 nsec, and a total power of 35 mw. High speed and low power in its MOS circuits are achieved through the silicon-on-sapphire (SOS) technology, in which silicon is grown on a sapphire substrate in transistor areas only so that the connecting leads directly on the sapphire have a very low capacitive loading. [Source: RCA Laboratories]

In LSI technology the active transistors and all their connections are simultaneously made by unified batch processes. Consequently, the larger the chips cut from the original silicon wafer (which is 5 to 10 centimeters in diameter) the greater the economy, since there are relatively fewer costly connections to the outside. On the other hand, the difficulty of making perfect chips grows with their area. An optimum size turns out to be about 7 to 8 millimeters on the side, for which the yield is only about 20 percent. This puts a great premium on the effective use of the silicon area and has led to a complete reversal of emphasis in circuit design. Whereas efforts were made previously to minimize the active devices (tubes and later single transistors), in LSI the transistors occupy a small area compared to that needed for connections and their spacings, hence ingenuity in design is centered on minimizing requirements for connections and simplifying geometrical layouts (see Fig. 1).

For memory cells, one figure of merit is the bit area, kW^2, where k is a measure of cell complexity and W is the width of line or space attained with a particular technology (W is typically 5 micrometers). Another is the number of lines and contacts necessary per cell (*9*). Static bipolar cells show low merit under these criteria, as they require two or more transistors and two resistances. The same is generally true of static MOS cells made of n-type or p-type transistors or of both types in complementary symmetry. These MOS cells have various relative advantages, but all require four to ten transistors. It was found that much better merit factors are achievable in dynamic cells, in which the bit is stored on a capacitance and is periodically refreshed. At first, three MOS transistors were needed for these cells, but modern designs do with one. Dynamic cells occupy an area of only about 25 by 25 μm. This dense bit packing is obtained at the expense of a relatively small lengthening of average access time, which is taken up by the refresh cycle. However, because of the need to refresh information, the dynamic cell favors periodically timed or

synchronous systems, which are less flexible than asynchronous systems. Hence the static cell with its uncompromising performance has not lost out.

In general, a memory chip carrying N bits is organized $N \times 1$; that is, it has one read-write bit channel and N addresses. Usually N is a power of 2 such as 1024, 4096, or 16384, in which cases the chip is referred to as 1K, 4K, or 16K. Internally, the N storing cells are physically arranged in an array and are selected by the two array lines on which they lie. The chip carries two decoders that select these two lines from a binary code. Hence the number of address lines to the chip, $\log_2 N$, is relatively small (10 for 1K, 12 for 4K, and 14 for 16K). The chip also carries read-write circuits, timing circuits, and regeneration circuits when dynamic cells are used. These peripheral circuits preempt two-thirds of the chip area in a typical MOS RAM 4K chip. Chips may be packaged in separate cans, which typically have 16 pins, and the cans may be mounted and interconnected on printed boards. Alternatively, chips can be mounted directly on the board.

Whatever technique is used for mounting chips, the interconnections are the dominant cause of failure. The mean time between failures (MTBF) per component is typically more than 1000 years, after an initial weed-out period of 10 to 200 hours during which it is not unusual for 1 percent of the components to fail (*9*). Despite this long component MTBF, a system with hundreds of thousands of components could well have an MTBF below several weeks. Fortunately, error-correcting codes with redundant additional digits can generally increase the MTBF; in fact, the probability of error can be reduced almost at will with a sufficient increase in system complexity. Single-bit errors, which are the major type by deliberate system design, can usually be corrected by relatively simple codes that entail a modest system cost increment (less than 15 percent).

To complete the memory system, of course, power supplies, containing boxes, and coupling and checking circuits are necessary. The resulting system per bit cost is in general almost ten times greater than the chip cost per bit.

The main memory of the IBM 370/168 may be taken as illustrative of the state of the art: it has a capacity of 64 million bits (64 Mb), uses 32,000 2K chips with static cells, has a cycle time of 2 μsec, costs 2.4 cents per bit, and achieves an MTBF of 10^5 hours (with error-correcting codes). Also standard in the last few years have been 4K, 16-pin dynamic cell packages with access times of about 200 to 300 nsec for 0.2 to 0.4 cent per bit (*10*). This year 16K RAM's have been appearing, and—which is indicative of the rapid progress being made—they have the same speed and power dissipation (½ watt) and pin count as the 4K RAM's (*11*). Another indicator of progress is a recently announced (*12*) bipolar technology known as integrated injection logic (I^2L), which has been used to make one-transistor dynamic cells on 4K chips that outperform their MOS counterparts.

Semiconductor RAM's may well fill the gap between present-capacity RAM's and mechanical disks by simply extending their capacity. A good case for this is made by Hodges (*9*), who envisages that in 5 years quarter-billion-bit memories will be available that are accessible in 2 μsec and cost 0.04 cent per bit. He assumes a reasonable extrapolated improvement in yield but no deviation from conventional technology. Other extrapolations confirm the increase in capacity and project even greater cost reductions.

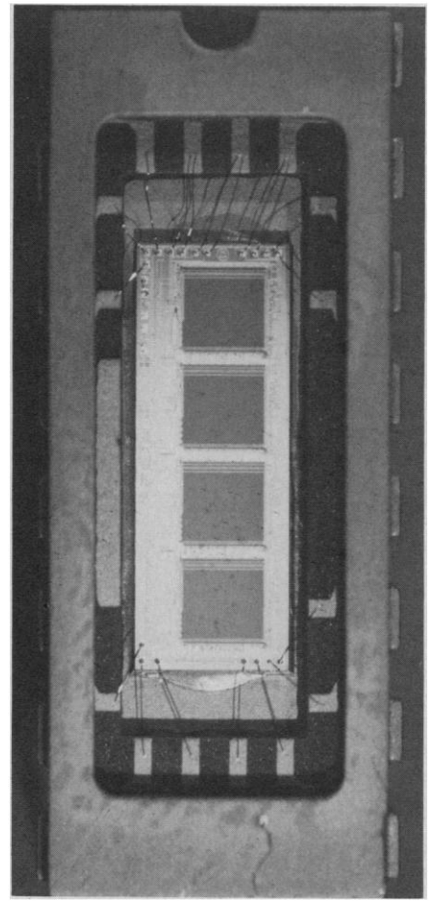

Fig. 2. A 16-kb block-addressed charge-coupled memory device. Each of the four 4K chips is organized in a series-parallel-series array and carries its own decoders. Operated at a data rate of 1 Mhz, the mean access time is 2 msec and the on-chip power dissipation only 1.5 μw per bit. The memory is fabricated by a simple MOS n-channel technique (*17*). [Source: Bell Telephone Laboratories]

At the heart of LSI techniques is photolithography, which makes it possible to conveniently create all the necessary microgeometrical patterns, including the delineation of areas for doping and for metallization. An arsenal of photographic reducing techniques, photoresistive coatings, and etching techniques has evolved. Diffraction of light at visible wavelengths limits the width of well-defined lines to about 2 μm, although at present widths of 4 to 5 μm are typical. A dramatic reduction to submicrometer dimensions can be achieved through the use of electron beams rather than photolithography. An electron beam focused to a fraction of a micrometer traces the desired pattern under computer control, either directly on the wafer or on a mask. The mask pattern is then replicated by x-ray exposure. A great deal of work and capital have been devoted to electron-beam lithography (*13, 14*), which already provides greater reproducibility in masks of conventional scale and in future will undoubtedly be producing scaled-down structures with high bit densities. Experimental structures with line widths of less than 1 μm have been demonstrated for some time.

Charge-Coupled Devices (CCD's)

One way to increase storage density is to avoid having a direct contact to each storage location, by shifting packets of charge at the silicon oxide interface in MOS devices, as was first described by Boyle and Smith in 1970 (*15*). Charge, injected or not at the beginning of the line, is shifted by applying phased pulses to the clocked electrodes so that the pattern of presence and absence of charges appears at the end of the line. Here it is amplified and reinjected at the beginning, making out of the CCD shift register line a serial memory with N circulating bits, much like the early delay lines. This return from RAM's to the older concept with its long latency time is motivated by the higher bit densities obtainable. As it turns out, the rules of design of RAM and CCD cells are the same, and there is not much to be gained with CCD's, whose register lines have to be spaced in the same way as the cells of RAM's and whose elements occupy roughly the same length along the line. However, with CCD's it is possible to use a much smaller charge with smaller elements in a scaled-down structure, be-

cause the stored charge is available at the end of the line to produce the readout signal, whereas in RAM cells it is divided by the capacitive loading of other cells on the line. This also leads to simpler readout circuits, and as the access circuits are inherently simpler the peripheral circuits preempt only half rather than two-thirds of the chip area. The net result is a fourfold increase in bit density, a 16K CCD being made on the same size chip as a 4K RAM.

To mitigate the latency time problem, a number of loops are organized in various ways. Each loop can carry its own regeneration, and the loops can be randomly addressed. Or the loops can share their regeneration circuits and be arranged in a series-parallel-series combination with short and long loops. Shared circuits with random loop selection are used in line-addressable RAM's (Larams), taking advantage of some inherent storage in unclocked CCD's. With these arrangements any block of information can be addressed randomly, but the information within the block flows serially, albeit at very high bit rates. While various trade-off's of latency time and complexity can be obtained, the average access time is several hundred microseconds.

There has been intensive work on CCD's (*15–18*) since their inception, and in 1975 and 1976 four manufacturers announced 9K, 16K, and 65K memories (see Fig. 2). Several more 16K CCD's are due in 1977. The CCD's benefit from all the advances that have been made in MOS technology, and hence their great bit density immediately translates into a correspondingly lower cost. On the other hand, their access times are almost 100 times longer than those of RAM's, and they are constrained by their particular data block size and organization and hence do not enjoy the great versatility of RAM's. It is likely that both CCD's and RAM's will be used as gap-filler technologies in the near future.

Whatever their success as memories, CCD's have many other applications well suited to their serial operation. An important area is signal processing. With elements sensitive to light, CCD's become a solid-state TV camera that already has characteristics rivaling those of its vacuum tube vidicon counterpart.

Bubble Memories

The control of cylindrical domains or bubbles in certain magnetic materials, first described in 1967 by Bobeck (*19*), has led to extremely interesting possibilities for memories. The bubbles are formed in a platelet by applying a biasing magnetic field normal to its surface. For a certain range of bias, cylindrical domains of reverse magnetization are formed. The domains can be moved in the plane of the platelet by relatively small field gradients. To produce a shift register, these gradients can be created by the combination of a rotating magnetic field over the whole platelet and an overlay of suitably patterned permalloy. The pattern is usually a succession of T-bars and I-bars, which become tiny magnets that periodically attract and repel bubbles and cause them to propagate along the line of bars. By connecting the two ends of a shift register a storage loop is obtained. At some location in that loop there is a tiny looped conductor, which when energized in one polarity generates a bubble and when energized in the opposite polarity annihilates a passing bubble if there is one. In this way a pattern of bubbles or no bubbles can be made to circulate and be stored in the loop. To read out the stored pattern, at some location in the loop there is a bubble replicator that creates two bubbles out of a passing one, one which continues in the loop and another one which is steered into the detector. Here it is expanded and its presence is sensed by the magnetoresistance of the permalloy guiding structure itself (see Fig. 3) (*20*).

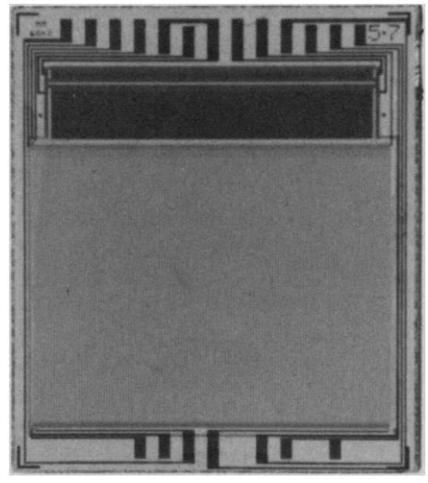

Fig. 3. Photograph of a 66-kb magnetic bubble chip (4 by 5 mm) organized in major and minor loops and having 3-μm bubbles. The large black rectangular area is a chevron expander-detector; here the bubbles are greatly expanded and cause an appreciable change in electrical resistance of the permalloy magnetic guiding structure. [Source: Bell Telephone Laboratories]

Fabrication of a bubble device starts with a substrate of gadolinium-gallium garnet (G^3). On these G^3 wafers, which are 5 to 7.5 cm in diameter, an epitaxial layer of doped yttrium-iron garnet (or more recently garnets with calcium and germanium), is grown. Next, an overlay of permalloy and conductors, patterned by standard photolithography, is applied on the wafer. Finally the wafer is cut into chips. Bubble devices require similar but fewer and simpler manufacturing steps than integrated semiconductor devices. A barium ferrite plate provides the necessary biasing permanent field and preserves on the system level the inherent nonvolatility of the bubble memory. Bubble chips can be packed into standard packages, which also hold the necessary magnetic-field generating coils and bias magnets as well as shields from unwanted external fields (see Fig. 4). A 32-pin package has been designed that houses four chips with a total capacity of 280 kilobits (kb).

Because bubble devices are inherently serial, various memory organizations with many storing loops are used, which resemble the CCD organizations that were developed later. Long registers, major and minor loop arrangements, and decoders correspond to CCD serpentine shift registers, series-parallel-series registers, and Larams. Much of the address logic switching necessary within these composite organizations is obtained with the bubble technology itself. Advantage is taken of the local fields created by either a current loop or a bubble that influences the path another bubble will take. Therefore, while semiconductor circuits are still necessary to amplify the sense signals and furnish the inputs and the currents for the rotating magnetic field, they represent a significant but not a controlling economic factor.

A decade of intensive research and development work on bubbles in many laboratories throughout the world has resulted in a basic knowledge of the physics and materials, as well as many ingenious devices, in what has been one of the most important innovations in magnetism (*20, 21*). Large-capacity memories are in pilot production, and it is likely that several manufacturers will announce products in 1977. Chips are likely to carry 32 to 100 kb. Bubble memories intended mostly for gap-filler applications are called files, and they clearly compete with CCD's. Their chief disadvantage is a bit rate of only several hundred kilobits per second, compared to several megabits per second in CCD's. Their chief advantage is nonvolatility and added reliability, although the latter point is being debated. While they might be less expensive because of their simpler processing, this is an uncertain advantage in view of the experience

and volume production achieved in the manufacture of CCD's, which have a head start of several years.

Recent laboratory developments may give bubbles a greater competitive edge. With the minimum photolithographic line width of 2 μm, the permalloy bar file (PBF) used so far has a period of 28 μm and 5-μm bubbles. This has been bettered by a newly demonstrated pattern of asymmetrically shaped disks with a period of 16 μm, 3-μm bubbles, and a chip density of 68 kb (*20*). Extension to 8-μm periods and 250-kb chips is foreseen. Another ingenious innovation, the contiguous disk file, has a gapless structure and tolerates a relatively coarse photolithography; in this file 1-μm bubbles have been propagated along the periphery of contiguous 5-μm disks (*21*).

Perhaps the most interesting development is the bubble lattice file, in which bubbles are packed in a hexagonal lattice determined by bubble-bubble interactions and the densities achieved are four times greater than those of a PBF employing bubbles of the same size (*22*). Information is no longer coded by the presence or absence of a bubble. Instead, use is made of the fact that the magnetization in going radially out of the bubble reverses by rotation within the domain (Bloch) wall, and that the sense of that rotation can be either the same all the way around the bubble or else in one direction in one half and the other in the other half. Entire lines of the lattice can be translated by electrical conductors parallel to the lines but spaced many lines apart. An extracted line at the end is read out seriatim, as in PBF's, by taking advantage of the fact that bubbles in one state move along a field gradient and those in the other state move normally to it.

While their low bit rates and debatable economy make bubble files of current design marginal competitors as gap-filling memories, there is great promise in the newest developments. Bubbles have other applications. For instance, their serial access and their ability to store without holding power make them a natural device for recording audio signals, since storage capacities for minutes of recording seem already economically feasible. The same properties render them valuable in space applications. There are also various uses in telephone systems.

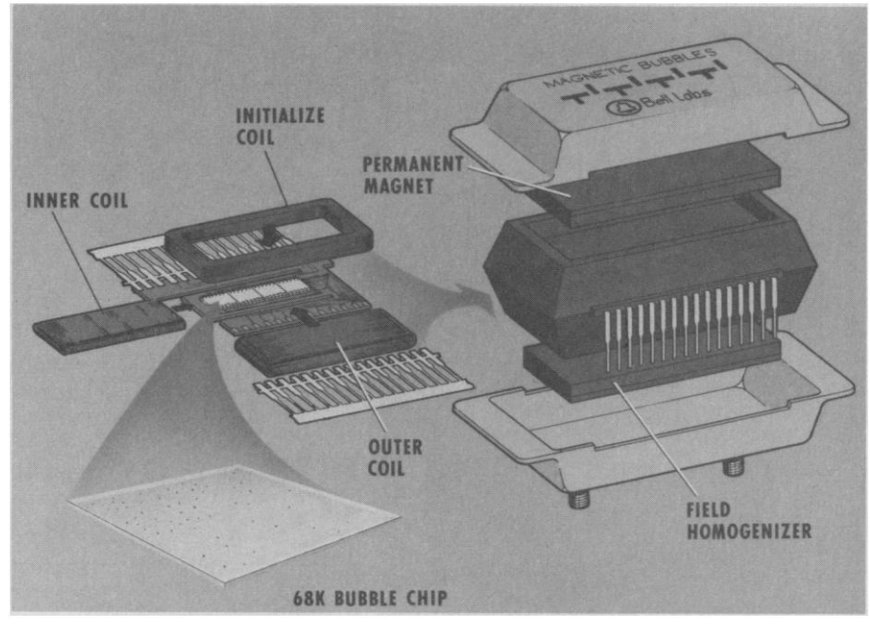

Fig. 4. Scale drawing of a magnetic bubble package. The four-chip package stores 270 kb and contains the necessary biasing permanent magnet and coils for the rotating field. [Source: Bell Telephone Laboratories]

Electron Beam–Accessed Memories

The electron beam is an addressing pointer of high definition and energy density that can easily be deflected. In storage tubes of the 1940's there were severe limitations to such addressing because of the use of surface charge storage and inadequacies in focusing and deflecting the beam. Two recent innovations, storage within a semiconductor and compounded deflection, may bring us closer to realizing the inherent potential of beam addressing. Two groups, one at General Electric working on the beam-addressable MOS (Beamos) (*25*), and the other at Micro-Bit Corp., have memory systems nearing production. For example, consider the Beamos tube with 32 Mb.

The target of this tube is a silicon wafer that has no geometrical pattern. On a *p*-type substrate there is an epitaxially grown *n*-type layer, then a thin layer of silicon dioxide, and finally a thin layer of aluminum. The thicknesses are such that a 10-kev electron penetrates through the metal and SiO_2 layers and generates electron-hole pairs in the *n*-layer. In order to write, a positive bias is applied to the aluminum layer, and holes that are generated in the SiO_2 are trapped near the Si-SiO_2 interface. To read out, the element is interrogated with a beam of lower intensity and no bias voltage. Many electron-hole pairs are produced in the *n*-layer, and if there are trapped charges many holes diffuse to the *n-p* junction before recombining. Here they are detected as a current in the external circuit. If there are no trapped charges, the holes recombine and produce no output. The output is typically 1000 times greater than the beam current because each high-energy electron generates many electron-hole pairs. Some erasure occurs on readout, and after a number of cycles the information must be refreshed. (However, with the beam off, charge will remain for as long as 1 month.) Thus the silicon target provides storage in depth, not on the surface, as well as local amplification of the signal. Its only drawback is a fatigue effect that reduces the readout, and this may be cured by relatively infrequent annealings of the tube.

The addressing is in two parts. First, the beam is deflected by a short conical structure of low aberration and strikes normally one of the apertures of a matrix of lenslets. The matrix is made up of two metal plates that have an array of holes (an 18 by 18 array on 1.5-mm centers) and are maintained at different potentials. Each precisely aligned pair of holes forms an immersion lenslet that focuses the beam to 2 to 4 μm. Second, the beam is deflected by bars running along rows and columns between the holes of the matrix. No matter which lenslet is reached, the reduced beam will be subject to the second deflection. In this compounded deflection the accuracy and stability at each step need only be a small fraction of what would be required with a single step. Data can be laid out at will within a lenslet field, and data blocks can vary over wide limits. However, guard bands are required between blocks and lenslets fields. In practice, the best operation is achieved with random access to a block of information and serial access

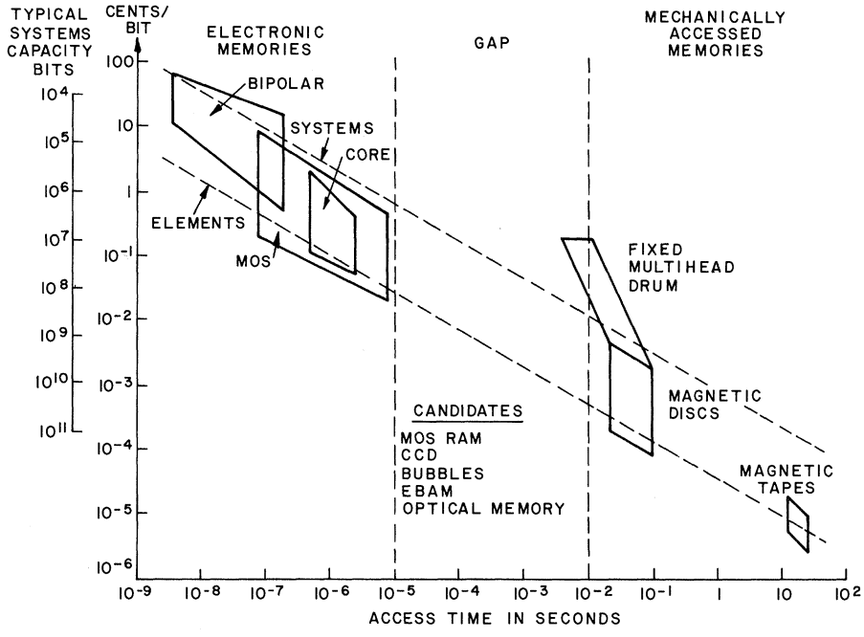

Fig. 5. Cost per bit and capacity plotted against access time for memories in use today. Costs at the element and system levels are indicated by dashed lines.

within the block. The access time is typically 30 μsec and the bit rate 30 Mb/sec.

In a system of large capacity a number of tubes can be configured in various ways. Tubes can be addressed singly and many circuits shared between them for low cost. Alternatively, many tubes can be addressed in parallel for high data rates.

The resort to vacuum technology has clear benefits: it eliminates all geometrical steps from silicon processing, and it leads to very high bit densities. A bit spacing of 4 μm has been demonstrated and it may be possible to bring this down to 1 μm. Costs should be lower than with LSI technology, which is encumbered with complex patterns and a multitude of connections. On the other hand, the tubes are complex, prone to fatigue, and require a high-voltage supply, and the analog circuits needed for deflection are not as foolproof as digital circuits. Despite many years of work and extraordinary achievements, this technology has not appeared on the market and its future is still uncertain.

Optical Memories

Modern developments in optics such as the laser, holography, and various electrooptical effects, have led to new technologies for information storage that are having success in mass storage devices involving mechanical motion (*26*). Of interest here is the unique possibility they offer for making a single memory with the capacity of disk memories and the accessibility of transistor memories (*27*). Briefly, the idea is as follows.

A laser beam is deflected in two directions and passes through an array of light valves in such a way that the image of the array is focused on a storage medium at some location x,y. Some of the light from the laser bypasses the array and strikes the medium at the same location x,y. Thus a hologram is produced that stores the pattern of openings of the light valves, or one page of information. Because of the redundancy inherent in holograms, the medium does not need to be perfect at the microscopic level of every bit, but merely at some reasonable fraction of a page area. To read from a stored hologram, a reference beam is again directed at it. The real image that created the hologram reappears on the array of light valves. In the array, in addition to a light valve at each point, there is also a light sensor and a semiconductor flip-flop or one bit of a transistor memory. The array, or "latrix," can be considered as the main inner transistor memory of the computer (or part thereof), and the whole arrangement as a way of "photographing" the contents of the internal memory on the mass storage medium and vice versa. In this photographing, light transmission is exploited to transfer information over a very wide channel. Many thousands of bits travel between the storage medium and the transistor memory—all at once. This is very important, as this is the first concept for mass storage that involves such close coupling with the inner transistor memory that it not only avoids the fantastic speeds necessary to bridge the gap to conventional mechanically accessed memories over narrow channels, but provides a whole system that operates as a true random-access memory.

A feasibility model (*28*) has shown the validity of all the principles of the concept: access by compounded selection of page composition and location, wide-channel optical information transfer, and real-time writing and reading of holograms. In addition, great progress has been made in many laboratories in light deflection (by sonic waves and electro-optic effects) and page composers (based on electro- or magneto-optic effects or the distortions of a thin metallic membrane), as well as the appropriate optical and electronic systems. This progress has shown that fairly reasonable holographic systems could be designed. The main difficulty is the storage medium.

There simply is not a storage medium that is sufficiently sensitive to work at reasonable speeds with reasonable laser powers. Many materials were explored. Curie-point writing in thin films of materials such as manganese bismuth provided fine holograms but required inordinate power. More promising is lithium niobate, which has a unique combination of semiconductor and electrooptical effects. Its sensitivity was increased 10,000-fold in recent years, but it needs to be further increased by a factor of 100 to 1000 to be practical. The difficulty can be appreciated by considering that the sensitivity needed is comparable to that of conventional silver halide photography, which depends on a chemical gain mechanism for its high sensitivity. This suggests the possibility of a storage medium with local amplification, where light would merely trigger an effect energized by an auxiliary power source. In fact, devices made of a sandwich of a photoconductor and an elastomer, working in this manner, exhibited good sensitivities. However, these devices suffer from great technological difficulties.

Concluding Remarks

Faster stores cost more and consequently have smaller capacities. A convenient, although oversimplified, way to illustrate the hierarchy is to plot cost per bit and capacity in typical systems in use against average access time, a mixed parameter that takes into account random-access time and bit rates. The plot (Fig. 5) shows a speed gap of three orders of magnitude between electronically and mechanically addressed memories, and the overlapping areas of usage of

bipolar, MOS, and core memories. It also shows that there is usually an order of magnitude difference between costs at the element and the system levels.

How is the gap likely to be filled? The case for LSI silicon technology is very strong, and the revolutionary nature of that technology cannot be overemphasized. In the first place, there are fundamental reasons why it is ideal for memory matrix arrays. In any such array the peripheral access and signal-amplifying circuits will inevitably be made out of transistors, so that also using transistors for the array avoids the complications of heterogeneous technologies and the many connections needed between them. More significantly, such an array is nearly ideal in the sense that it is connection-dominated, the active storing cell occupying a negligible fraction of the area and contributing negligibly to failures. It is therefore difficult to imagine any other array randomly addressed through connections that can compete with LSI RAM's. The LSI technology also offers another good contender, the CCD, that achieves denser and hence cheaper elements, although with some sacrifice in accessibility.

However, the effects of LSI technology lie beyond these considerations. Throughout the history of computers it was considerably more expensive to manipulate bits for computing or logic switching than for storage. A core cost a few cents, but a tube and its components a few dollars. Central processing units with large memories were a natural architectural consequence. Today, for the first time, logic and storing costs are comparable. Thus it makes sense to associate much more logic with memory, which is precisely what is done in microprocessors. Microprocessors may change our thinking about computer system organization and computing methods. For example, the practice of multiprogramming with a large computer may be replaced by single programming by many small computers. In these smaller computers electronic memories may have sufficient capacity, and there may be much less emphasis on the gap. Microprocessors have also given a new dimension to the intriguing possibilities of simultaneous programming by many computers.

To sum up, the LSI technology is the prime contender for filling the gap. It has also brought about microcomputers, which may change our views on the storage hierarchy itself. The bubble technology can provide sufficient capacity, but its economic advantage has yet to be proved on the market and it has performance limitations with respect to LSI RAM's and CCD's. However, it has a great potential for improvement. Despite great progress and further possibilities, electron beam–addressed memories have a difficult road to the market. Optical techniques offer the best solution conceptually, but much research is necessary, particularly in storing materials.

This article has focused on new memory technologies for filling the gap between electronically and mechanically addressed memories—that is, on electronic memories in the range 10^6 to 10^9 or 10^{10} bits. There is another gap between mechanically accessed disk memories and manually accessed records, such as huge stacks of magnetic tapes. The Tetrabit, IBM 3850, and Unicon memory systems fill that gap by mechanically addressing 10^{12} bits. Thus there has been progress in making electronic what was mechanical and mechanical what was manual. However, we are still far from the ideal shoe-box device with 10^{12} bits accessible in nanoseconds, and still farther from the capacities of 10^{15} bits needed for many already well-defined applications. Although much can still be expected from LSI and magnetic recording techniques, these greater goals may require radically new approaches.

Many laboratories are looking into basic principles. Superconductive memories based on the Josephson effect may be able to operate in picoseconds on miniscule power. The boundaries within the walls of magnetic domains, exploited in the bubble lattice devices, are also used in a so-called cross-tie memory that may provide nonvolatile storage memories on LSI chips (*29*). Other possibilities being explored are storage by electrolytic deposition in an array of cells, novel sources of electrons and ions for beam devices, and applications of biomolecular properties. Most of this research is in an early laboratory stage and the eventual system aims are not sharply defined. Very high speed and very low power memories rather than large capacity may well be the benefits of some of these approaches.

Any radical improvement in memory technology will ultimately greatly affect our way of life, as previous innovations have shown. The challenge is a fascinating one as it reaches into basic knowledge and demands imaginative invention.

Summary

Computers today use a hierarchy of large-capacity, relatively slow mechanically accessed memories in conjunction with fast electronically accessed memories of relatively small capacity. While the gap between these is spanned by ingenious organizations and programming, it would be highly desirable to fill it instead by some device of sufficient capacity and speed. Candidates for gap-filling memories include metal oxide semiconductor (MOS) random-access memories (RAM's) made by large-scale integration (LSI); charge-coupled devices; magnetic bubble devices based on cylindrical domains of magnetization; electron beam–addressed memories; and optical memories based on lasers, holography, and electrooptical effects. At present, the MOS RAM is the prime contender. Its natural evolution and the evolution of magnetic-recording techniques on which mass storage is based are likely to continue to shape the future as they have for more than a decade. On the other hand, radically new technologies, still at an early laboratory stage, are aimed at a more ideal solution than today's hierarchy.

References

1. E. W. Pugh, *IEEE Trans. Magn.* **7**, 810 (1971).
2. A comprehensive review of digital storage systems appears in *Proc. IEEE* **63** (No. 8) (1975).
3. G. C. Feth, *IEEE Spectrum* **13**, 37 (June 1976).
4. J. W. Forrester, *J. Appl. Phys.* **22**, 44 (January 1951).
5. J. A. Rajchman, *RCA Rev.* **13**, 183 (June 1952).
6. ———, *Sci. Am.* **217**, 18 (July 1967).

6a. F. P. Brooks, Jr., *IEEE Trans. Magn.* **15** (No. 3) (September 1969).

7. R. Noyce, *Science* **196**, 1102 (1977).
8. J. D. Schmidt, *Solid State Design* (January 1965), pp. 21–25.
9. D. A. Hodges, *Proc. IEEE* **63**, 1136 (1975).
10. R. A. Abbott, W. M. Regitz, J. A. Karp, *IEEE J. Solid State Circuits* **SC-8**, 299 (October 1973).
11. C. Kuo, N. Kitagawa, D. Ogden, *Electronics* **49**, 81 (May 1976); also commercial announcements of various manufacturers in 1976.
12. W. B. Sander, W. H. Sheperd, R. D. Schinelle, *ibid.* **49**, 99 (August 1976).
13. A. N. Broers and M. Hatzakis, *Sci. Am.* **227**, 34 (November 1972).
14. R. Bakish, Ed., *Proceedings of the Sixth International Conference on Electron and Ion Beam Science and Technology* (Electrochemical Society, Princeton, N.J., 1976).
15. W. S. Boyle and G. E. Smith, *Bell Syst. Tech. J.* **49**, 587 (1970).
16. L. M. Terman and L. G. Heller, *IEEE J. Solid State Circuits* **SC-11**, 4 (February 1976).
17. A. M. Mohsen, M. F. Tompsett, E. N. Fuls, E. J. Zimanny, Jr., *ibid.*, p. 40.
18. W. F. Kosonocky, *ONR Contract Rep. Contract NO0014-76-C-0371* (1976); pp. V-520 to V-534.
19. A. H. Bobeck, *Bell Syst. Tech. J.* **46**, 1901 (1967).
20. ———, P. J. Bonyhard, J. E. Gevsic, *Proc. IEEE* **63**, 1176 (1975).
21. M. S. Cohen and H. Chang, *ibid.*, p. 1196.
22. P. J. Bonyhard and J. L. Smith, *IEEE Trans. Magn.* **12**, 614 (1976).
23. G. S. Almasi, G. E. Keefe, Y. S. Lin, R. J. Hendel, R. F. McGovey, paper presented at the INTERMAG Conference, London, April 1975.
24. O. Voegli, B. A. Calhoun, E. L. Rosier, J. C. Slonczewski, *AIP Conf. Proc.* **24**, 617 (1975).
25. W. C. Hughes, C. Q. Lemmond, H. G. Parks, G. W. Ellis, G. E. Possin, R. H. Wilson, *Proc. IEEE* **63**, 1230 (1975).
26. D. Chen and J. D. Zook, *ibid.*, p. 1207.
27. J. A. Rajchman, *J. Appl. Phys.* **41**, 1376 (1970).
28. W. C. Stewart, R. S. Mezrich, L. S. Cosentino, E. M. Nagle, F. S. Wendt, R. D. Lohman, *RCA Rev.* **34**, 3 (March 1973).
29. L. J. Schwee, H. R. Irons, W. E. Anderson, *IEEE Trans. Magn.* **23**, 608 (1976).