Bridging the memory access gap

by DENNIS E. SPELIOTIS
Micro-Bit Corporation
Lexington, Massachusetts

INTRODUCTION

The rapid growth of electronic data processing over the past two decades has been characterized by an almost insatiable appetite for larger and faster memories. In fact, over this period of time, on-line storage capacity has increased about three times as much as CPU power.[1] Yet, in spite of this impressive growth rate, memory represents perhaps the most limiting area in the development of more advanced computer systems. Furthermore, the increasing diversity of storage devices and the wide disparity in the price-performance of these devices, present a difficult challenge to the system designer, and account for a large part of the complexity of the software and hardware system to manage the storage facilities.

Viewed simplistically, memory devices can be classified into two basic categories: electronically accessed main memory and electromechanically accessed peripheral memory. The former is fast and relatively expensive, while the latter is very slow (typically a factor of 10^4 to 10^5 slower access) and relatively inexpensive (by about a factor of 10^{-1} to 10^{-3} in price per bit). Between these vastly separated device technologies we have the famous memory access gap,[1] which has persisted essentially unchanged over the last twenty years even though the boundaries on either side of the gap have moved toward faster access by about an order of magnitude over this same period of time.

The absence of a bridging technology is in no way the result of a lack of effort and intensive search to develop such a technology. Suffice it to mention cryogenics, thin magnetic films, thermoplastics, and magneto-optics as a partial list of the most spectacular but unfruitful endeavors in this domain. In recent years we find a tremendous effort concentrated on magnetic bubbles and CCD (charge coupled devices). However, the technology with the greatest promise and potential in bridging the access gap is the three quarters of a century old electron beam, which ironically also provided the access means for the very first random access memory way back when it all began.[2]

The purpose of this paper is to set forth the fundamental arguments on "why electron beams", then to describe the particular desirable attributes of an electron beam addressable memory system, and finally to describe the achievements to date and the expectations for the future.

WHY ELECTRON BEAMS

Peripheral memories

Peripheral memories in digital computer systems have been dominated by magnetic recording devices, such as disks, drums, and tapes. The spectacular success and growth of magnetic recording storage devices derive from certain inherent characteristics of the technology:

(a) Very low cost storage media based on homogeneous (non-discrete) magnetic surfaces.
(b) A means of accessing which allows tens of millions of bits to share one write-read transducer and encode-decode-sense channel.
(c) The fundamental limits of the technology were far beyond the demands placed on the technology, thus allowing plenty of room for growth and expansion in device capability and performance. Using areal density as a measure of device sophistication, disk systems for example have gone from 2.2×10^3 bits/in^2 in RAMAC I introduced in 1956 to 2.24×10^6 bits/in^2 in the CDC 9762 introduced in 1974, a factor of one thousand improvement in packing density!

The fundamental deficiency of disks is their slow electro-mechanical access. System designers have resorted to a number of techniques in order to partially mask the long access time, but all these approaches are costly and less than satisfactory at best. They include:

(a) Queuing and look-ahead to minimize disk arm motion. This is difficult to optimize in multiprocessing environments and at high priority interrupt frequencies.
(b) Transfer large blocks to minimize the frequency of accesses to the peripheral device. This requires expensive buffering, and depending on the data transfer bandwidth could actually result in lengthening of the average access time in certain processing environments.
(c) Employ fixed head disks and drums to eliminate arm motion and head positioning time. This reduces the average access time by a factor of three to five,

but increases the per bit price by one to two orders of magnitude.

(d) Employ more than one head per arm, like in the IBM Winchester disk system, or more than one head per track, like in the IBM 2305-I disk system—both of which are rather expensive propositions for only a small improvement in access time.

It is clearly evident that the slow access time of the electromechanically-accessed peripheral memories is a very serious bottleneck in improving system performance, and the techniques that are being used to shorten it provide only small relief at considerable cost. Furthermore, the evolution of computing systems toward timesharing, virtual storage, multiprocessing, and network processing tends to aggravate the situation and places additional emphasis for a technology that bridges the access gap. Until such a technology becomes available, we will continue to see an accelerated growth in the size of main memory—and a corresponding increase in total system cost—as a necessary prerequisite for efficient system performance.

Main memories

Main memories on the other hand have price-performance characteristics which are essentially the opposite of peripheral memories. Their outstanding advantage is their very fast access time achieved by direct wired access to each bit. Their disadvantage is the high cost per bit, and the reasons for this are several:

(a) The bits are physically discrete entities. This has very important reflections on the cost to introduce the discreteness and obtain satisfactory yields.

(b) Wired access to each bit also reflects on the cost to introduce or install the wiring, and to make the many thousands of interconnections needed for a sizable memory. It also affects the yield and the reliability of the devices, particularly as the bit packing density and the total capacity of the memory increase.

(c) The number of bits that can share a sense amplifier is typically a few thousand as contrasted to tens of millions in the case of peripheral memories.

It can be argued that the cost of main memory will continue to decrease as more integration and automation are introduced into device fabrication. However, the above arguments still apply, and diminishing returns will tend to dampen cost improvements, particularly for mature technologies.

Searching for a gap-filling technology

Ideally what we need is a new technology which can approach on the one hand the access time of main memories, and on the other the per bit price of peripheral memories. Clearly, such a technology must employ electronic accessing. It should also incorporate many of the other attributes which contribute to the low per bit cost of peripheral memories, such as high bit packing density ($\geq 10^6$ bits per square inch), avoid structure or discreteness for defining the bits, employ a minimum number of interconnections, and allow for the sharing of a very large number of bits by a sense amplifier. Many different schemes have been tried and a much greater number of techniques have been proposed for a memory technology which may satisfy the above requirements. Generically they fit into two categories with different philosophy of accessing:[3]

Moving the bits to the sensor

This category includes all the shift register types of devices which electronically propagate one or more series of bits to a sense amplifier through a fixed propagating structure. The prime examples of such technologies are charge coupled devices (CCD) and magnetic bubble memories, the latter of which in particular is receiving a tremendous amount of attention currently.

If we examine the potential of bubble memories on the basis of the criteria outlined above for peripheral and main memories, we can make the following observations:

(a) Even though the storage medium is homogeneous, the propagating structure is not.

(b) Material perfection requirements in moving bit devices are tough, and this reflects on yields and cost.

(c) Very small bubbles which would permit high bit packing densities have been observed only in amorphous films and bubble lattice structures, both of which have their own materials and processing complexities.

(d) Bubble propagation speeds are rather slow, and even though paralleling is straightforward, it adds rapidly to the cost of driver-sense electronics.

(e) Major-minor loop organizations to facilitate the sharing of a large number of bits by one sense amplifier in order to reduce cost tend to degrade their access time.

The most significant advantages of this technology are non-volatility, low power requirements, and volumetric compactness.

The above observations on magnetic bubbles essentially apply also to the CCD technology, but with some significant differences. The access time and propagation time of CCD is faster by about one order of magnitude than that of magnetic bubbles. Offsetting this advantage are certain disadvantages, which include volatility and the need for very frequent refreshes of the data pattern (a factor which can have significant repercussions on error rates even when the memory is not being accessed), higher power requirements, and the need for a large amount of additional electronic circuitry for selecting, driving, and refreshing the much shorter data loops.

Clearly, for both CCD and magnetic bubbles, the fabrication and processing complexity, the level of discreteness in the definition of the bits, the areal bit density, the number of interconnections, and the number of bits that can share one write-read channel, fall in the intermediate area between main memory and disk-type peripherals, as does their access time. Therefore, their price-performance potential would place them in the "classical" access gap, where they should offer cost competitiveness with small (up to 20 megabits) fixed-head disk or drum systems, but with an all solid-state technology and with significant improvements in performance. More specifically, we would expect CCD to penetrate the small auxiliary storage sector where performance is paramount, whereas magnetic bubble memories will be used in special applications where moderate performance is acceptable, but ruggedness, reliability, non-volatility, compactness, and low power consumption are emphasized (aerospace systems, process control, word processing, numerical control, and telecommunications).

Moving the sensor to the bits

In this category we place technologies that employ an inertialess access mechanism which interacts with a stationary storage medium for the writing and reading of information. Consequently, our discussion is restricted to devices which employ high energy beams of either sound, light, or electrons as the accessing mechanism.

Sonic beams

Fundamental limitations in focusing and deflecting a sonic beam preclude the feasibility of using it as a random access addressing mechanism. Serial access utilizing a magnetostrictive film as the storage medium has been demonstrated,[4] and even though the transfer rates are very attractive the bit densities are not, which would tend to exclude this approach as a serious contender for general purpose beam addressable memory applications.

Light beams

Spurred on by the development of the laser and holography, a tremendous amount of attention has been given to the development of optically accessed memories. Unfortunately today, more than a decade later, the prospects appear less than exciting for various reasons the most important of which are:

(a) The most fundamental problem is the development of a suitable, nonvolatile, erasable, optical storage medium. Many different materials and interaction modes have been investigated, including magneto-optics,[5,6] thermoplastics, photochromics, photodichromics, electro-optics, and amorphous semiconductors. All of these storage materials have one or more basic shortcomings which limit their application and usefulness (such as low sensitivity requiring enormous beam energy densities for writing and erasing particularly for holograms, limited reversibility, low diffraction efficiency, need for cryogenic operating temperatures).

(b) The other fundamental problem is the development of high speed, high repetition rate, low cost digital deflectors which can address a large number of resolution elements. Because of the overhead required for the generation, modulation, focusing and deflection of a light beam, the minimum capacity for an economical memory system would have to be about 10^8 bits. For bit by bit recording, this would require a deflector system capable of accessing $10^4 \times 10^4$ resolvable spots, a requirement far exceeding the capabilities of acousto-optical and electro-optical deflectors. Such deflectors are adequate for a page organized holographic memory, but this approach is impractical today due to the unavailability of suitable materials (except for read-only photographic emulsions) and page composers[7] for input data formatting prior to exposure of the holograms.

(c) Even if the deflector limitations to bit by bit recording were to be eliminated through some breakthrough in development, still the addressability of a field of 10^8 bits would present formidable problems due to diffraction, depth of field, depth of focus, aberrations in the optical components, and accuracy and stability requirements in the deflection electronics. Proposals[6] to get around these problems by incorporating mechanical motion of the storage medium are unattractive, because the achievable areal storage density could only be slightly higher than for magnetic recording while the performance would be comparable and the cost much higher.

Electron beams

Electron beams possess very attractive properties which render them far superior to light as a memory access mechanism.

(a) Like light, they can be formed into high energy density, high resolution beams, but since their diffraction limit is several orders of magnitude beyond that of visible light they are inherently capable of much greater resolution, depth of focus, and depth of field.

(b) Unlike light beams, deflection, modulation, and scanning of electron beams is exceptionally simple and fast.

(c) Because electron beams are strongly interactive with both electric and magnetic fields, we can envision a variety of materials as a possible storage medium, including ferroelectric, magnetic, semiconductor, thermoplastic and insulator films.

(d) Achievable spot size and energy density (current density) into the spot are closely interrelated, and are ultimately limited by the brightness of the source, the physical size of the electron-optical components, and the aberrations introduced by these components—particularly the deflector. On the other hand, the limits on relocating the spots are imposed by the accuracy and stability of the deflection electronics and by thermal and mechanical considerations. These constraints combine to set some practical limits on the number of spots that can be randomly addressed reliably in a single lens-deflector field, and that number is 10^7-10^8. Brighter sources such as field emitters would certainly allow much higher current into the spot, but would not appreciably alter this limit until more accurate deflection electronics became available. Even with thermal electron sources, however, we can conservatively project several million bits for a single lens-deflector field and available materials, corresponding to areal densities of 2×10^7-7×10^7 bits per square inch, which are indeed very impressive.

(e) The total field that can be accessed by a single electron beam can be expanded by several orders of magnitude by employing two stage deflection and an array of lenses known as the fly's eye[8] configuration. This approach removes the constraints imposed by deflection electronics and opens the road to the development of memories with capacities of several gigabits, access time of a few microseconds, and costs comparable to those of large disk files. It is precisely this unlimited potential and the tremendous versatility and capability of electron beams that makes them the most powerful developmental technology in the race to bridge the access gap.

DESIGN CRITERIA FOR AN ELECTRON BEAM MEMORY

In this section we discuss the components of an electron beam addressable memory (EBAM) system, the available choices in the design and configuration of the system, and some of the reasons for selecting a preferred embodiment in the practical implementation of the system. The discussion is aimed toward the general concepts of the design rather than the specific analytic details.

Electron optics

The electron source should be a dispenser-type cathode, which has long life of up to 50,000 hours at loadings of over 1 A/cm^2, and is readily available and inexpensive. Such cathodes incorporated into well designed guns operating at moderate beam energies (about 10 kV) can provide excellent brightness. Much higher brightness can be obtained with field emission cathodes which, however, need additional engineering development. Since brightness increases rapidly with increasing beam voltage, the selection of that voltage must be made as a compromise between brightness, operational requirements of the storage target, and the complexity and cost for insulation, power supplies, and deflection amplifiers, which increase with increasing voltage. A good choice would appear to be a beam voltage of about 10 kV.

For beam focusing and deflection, electrostatics is definitely a clear choice over magnetics for practical as well as fundamental reasons. Rapid random access requirements would exclude magnetic structures due to hysteretic or inductive limitations. Cost, size, and weight would also favor the electrostatic approach. Furthermore, the difficulty of confining magnetic fields would inhibit the close packing of a cluster of tubes in a system sharing power supplies and deflection electronics—a highly desirable configuration to reduce system cost and increase throughput bandwidth. Finally, magnetic focusing and deflection would be totally inapplicable for array electron-optical structures of the fly's eye type.

Storage medium

The storage target must be a stable material, compatible with high vacuum and bakeable to at least 350°C, possess some physical property which can be reversibly, rapidly, and efficiently altered by a high resolution electron beam, be capable of good "one"-"zero" discrimination and large signal-to noise ratio at very high bit packing densities, and have a long life under continuous operation. A crucial additional requirement is that the target be homogeneous and structureless in a plane perpendicular to the beam axis so that the bits are located at the points of beam incidence, rather than for the beam having to find predetermined bit locations related to a specific structure in the target plane. This requirement is imposed not only by cost considerations in introducing the structure and the resultant yields, but also by the previously mentioned need to share electronics for a matrix of tubes, and to accommodate some inherent residual aberrations and differences among tubes.

The lack of a suitable target that satisfies the above requirements has certainly been the most constraining limitation in the development of EBAM systems. Thermoplastic materials[9] have very limited reversibility due to polymerization induced by cross-linking. Similarly limited are amorphous to crystalline phase transitions in chalcogenide glasses.[10] A comparison between magnetic and electrostatic storage targets results strongly in favor of the latter. Lorentz interactions either with the magnetization or with the external fringing field of magnetic films are so weak as to preclude readout from such films at anywhere near the desired bit densities and speeds.[11] Also, if readout was based on using the energy of the beam to thermally disturb the cooperative coupling of assemblies of dipoles, electrostatics would again be favored over magnetics because of higher energy density. Ferroelectric thin films[12] are potentially powerful contenders for storage targets in EBAM systems, but additional materials development is required to fully realize their potential. Even tougher materials problems have impeded progress toward the development of a photoconductor-ferroelectric sandwich target.[13]

A material which satisfies very well the requirements for an EBAM storage medium is the silicon-silicon dioxide system, and it is indeed an extraordinary phenomenon that all known current approaches[14–17] to write/read electron beam memories are based in some way or another on silicon technology and electrostatic charge storage. Some of these approaches use surface charge storage in a finely etched structure on the oxide which is selectively charged by the beam for writing and serves to modulate either the secondary electron emission[14] or the transmissivity of a low energy beam[15] for reading; other approaches[16,17] use beam induced imbedded charge storage in the oxide and near the silicon dioxide-silicon interface, which modulates the depletion region in the silicon and serves to separate the beam induced charge carriers in the silicon for read out. The latter approaches are superior in that they employ structureless targets (for planes perpendicular to the beam axis), use beam voltages more compatible with high gun brightness and resolution, are directly usable in array optical configurations, and provide large local amplification during the reading and the writing operations. Even though such targets have their problems and limitations, they are based on a wide ranging and fast advancing technology, which can confidently provide the needed improvements. A more detailed description of the imbedded charge storage mechanism is given in the last section of this paper.

Other considerations

We have already mentioned the importance of sharing some of the electronics overhead among several tubes in order to amortize costs over a large storage capacity. Random access electron-optical memories require very stable and highly regulated power supplies, and a very fast and accurate deflection electronics system, both of which contribute in very significant proportion to the overall system cost, and hence the need for sharing. Using common deflection for a number of parallel channels (like 18), also results in the added advantage of greatly increasing system throughput and bandwidth. This, of course, implies that there exists a minimum total capacity below which electron optical memories are less cost effective, and this minimum is in the range of two to eight megabytes depending on whether the systems are used as main memory extensions or as high performance peripherals.

The access time of a few microseconds reflects primarily the deflection settling time. Consequently, even though EBAM systems can be used to access a single bit at a time, a more efficient utilization would result from a block addressable organization of the memory, where the beam settling delay is incurred only for the starting location of the block. Block organization also offers the added advantage of data encoding, and permits the reading of transitions rather than absolute charge levels, thus relaxing the requirements on signal uniformity.

The demise of the Williams tube and the subsequent slow progress in the development of electron beam memories was due largely to the basic deficiencies of the inadequately developed underlying technologies that are used in the making of such memories, and the unavailability until much later of the very large computers that are needed for design simulation of off-axis high resolution electron optics. Within the last few years, however, the confluence of significant advances and improvements in all the supporting technologies have made the realization of large and superior EBAM systems possible. These include:

(a) precise, stable, fast, and inexpensive electronic components
(b) silicon-based targets of excellent perfection and processing control
(c) long life dispenser-type cathodes and superior electron optical components
(d) advanced materials, fabrication techniques for low cost, high precision parts and assemblies, superior cleaning and handling techniques, and improved vacuum technology.

THE MOS ELECTRON BEAM MEMORY

Electron beams and MOS targets can be combined to create a new and powerful memory technology. The basic operation of the memory is illustrated in Figure 1. Information is stored in a pattern of positive charge in the

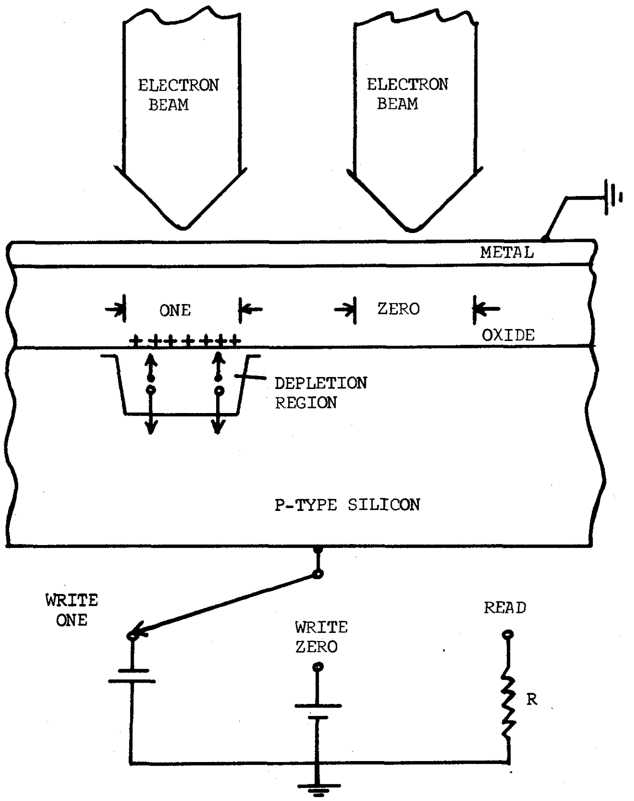

Figure 1—Storage and readout mechanism of the MOS electron beam memory (not to scale)

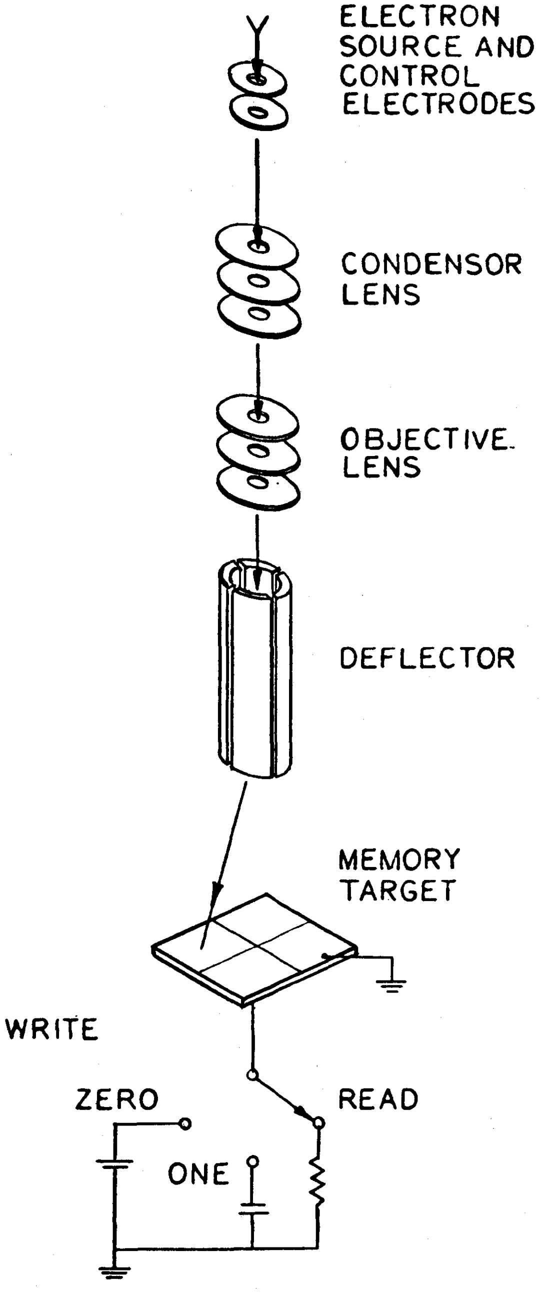

Figure 2—Schematic description of electron beam memory tube using single channel optics

oxide, near the interface to the semiconductor. The storage of positive charge corresponding to a "ONE" is accomplished by exposure to the beam under positive gate bias, while the removal of positive charge corresponding to a "ZERO" is accomplished by exposure to the beam under negative gate bias. The stored positive charge depletes the underlying regions of the semiconductor and drives the surface of the p-type substrate into inversion.

The energy of the beam (10 kV) and the thickness of the gate and the oxide are such that the beam penetrates several thousand Angstrons into the silicon. The reading of the stored information is accomplished in the silicon by using the same beam as for writing. The electron-hole pairs generated by the penetrating read beam in the semiconductor are separated by the strong field in the depleted regions under the ONES and a current is detected in the sensing resistor R, while the absence of a field under the ZEROS permits most of the generated charge carriers to recombine. Since the creation of an electron-hole pair requires an energy loss of about 3.7 eV, and the beam enters the semiconductor with an energy of a few keV, the readout signal is greatly amplified over the

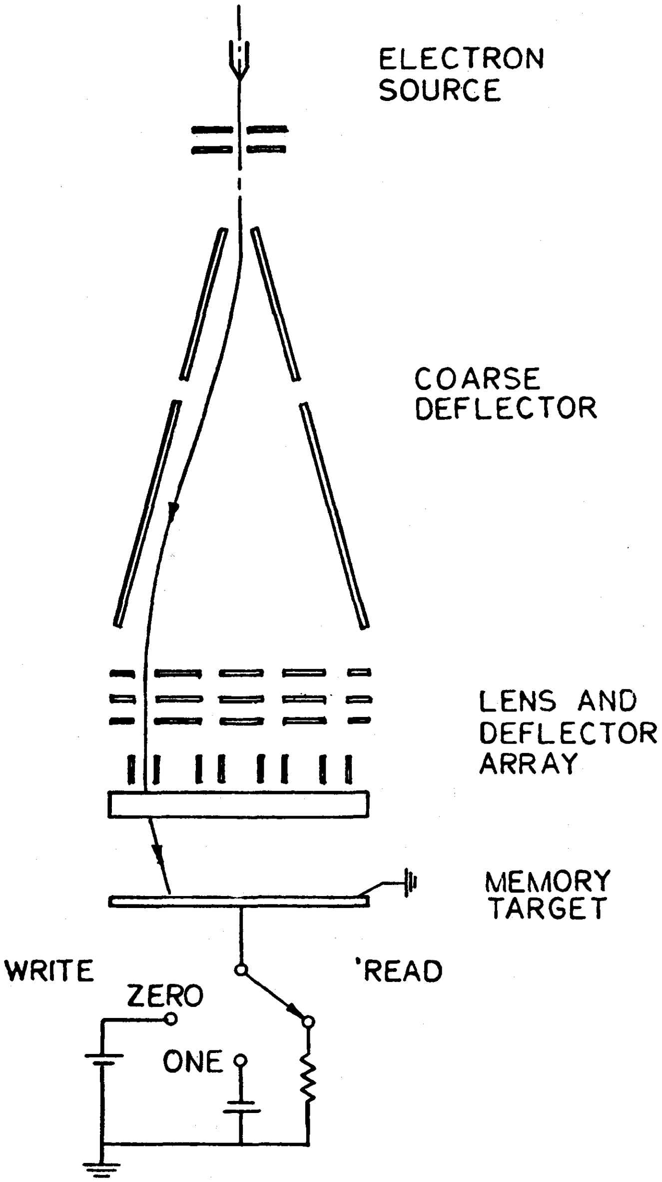

Figure 3—Schematic description of electron beam memory tube using array optics

read beam current by a large local and noiseless gain; a similar but much smaller gain also occurs during the writing process. This gain represents one of the most outstanding advantages of the MOS storage medium.

A schematic description of an EBAM tube under development using single channel optics is shown in Figure 2. Such a tube stores 4.2 megabits in 1 cm^2 target, using a 30 nA beam into a 2.5 micron spot, and is limited principally by deflection electronics accuracy and stability and by deflector aberrations. A memory system consisting of 18 such channels in parallel will have a capacity of 75 megabits, access time to any block of under 10 microseconds, and read/write throughput rates of 38 and 5 megabits/sec, respectively.

The limitations on the number of bits per tube imposed by deflection can be removed by using two stage deflection in the array optics configuration shown in Figure 3. This approach can start with tubes of 32 megabits capacity while the upper limits may be in the range of 0.5 to 5 gigabits per tube corresponding to bit densities of the order of 10^8 to 10^9 bits/cm^2! Because array optics also delivers a much higher current density into a given spot on the target, the throughput of the memory will also be considerably faster. A memory system consisting of 18 parallel channels of 32 megabits each, will have a capacity of about 600 megabits, access time to any block of under 12 microseconds, and read/write throughput rates of 90 and 40 megabits/sec, respectively.

Electron beam memories will initially be cost-competitive with fast auxiliary storage devices, and eventually with all on-line random access peripheral memories, but with far superior performance. Equally important, however, is their potential use as main memory extensions, in combination with a semiconductor cache, where they will have a large price advantage at comparable performance. This application is envisioned in Figure 4, which shows price-performance comparisons for EBAM and other memory technologies through the end of this decade.

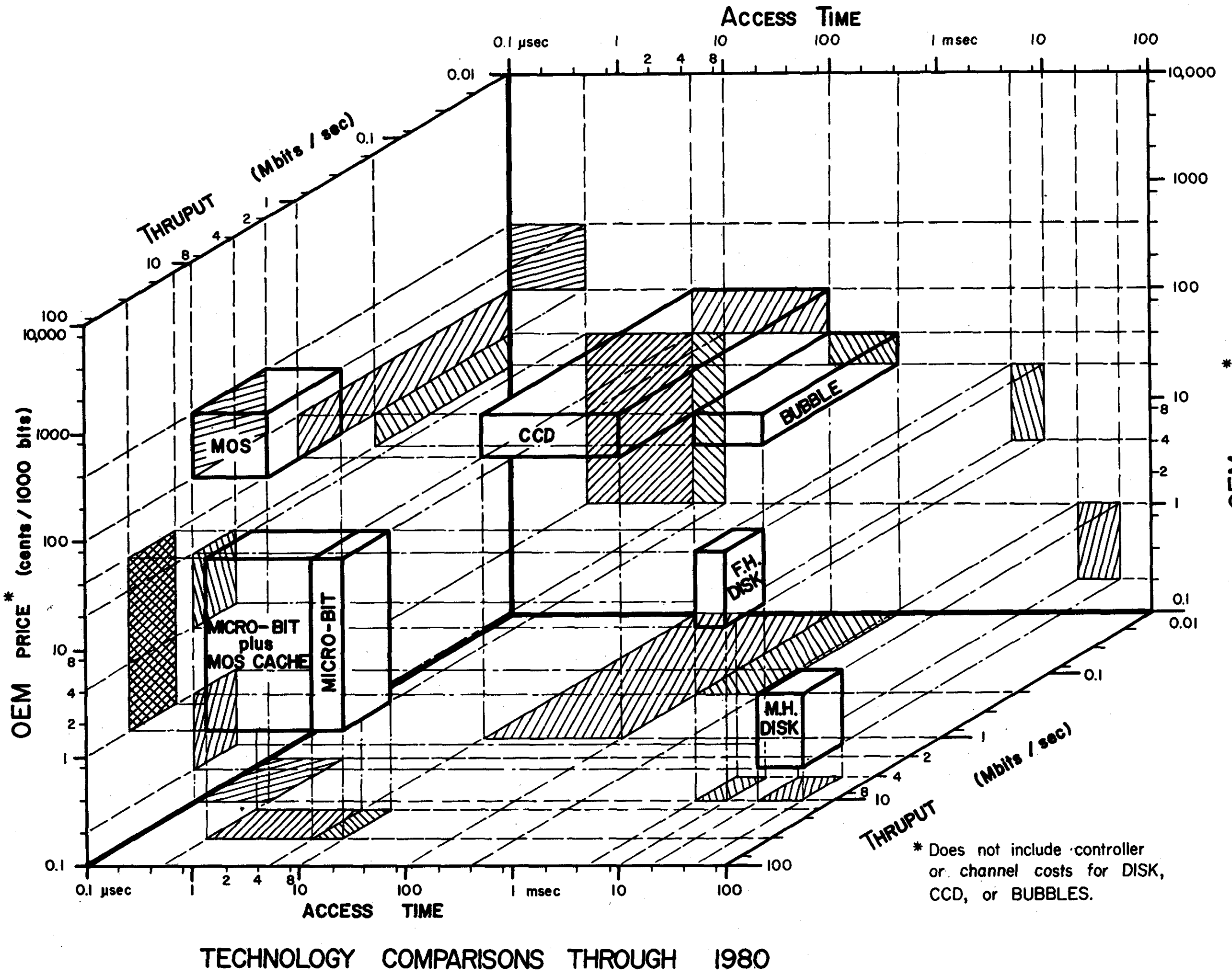

Figure 4—Price-performance projections of various memory technologies through 1980. Transfer rates for the shift register devices are along each track

Electron beam memories have the power not only to occupy a strong place in the memory access gap, but also they have the potential to completely eliminate the gap and bring about a drastic simplification in system architecture and a large improvement in cost-performance.

ACKNOWLEDGMENT

It is a pleasure to acknowledge the many helpful discussions with my colleagues K. J. Harte and D. O. Smith.

REFERENCES

1. Pugh, E. W., "*IEEE Trans. Magnetics*, MAG-7, 810, 1971.
2. Williams, F. C., and T. Kilburn, *Proc. IEE* (London) Part III, 96, 81, 1949.
3. Smith, D. O., *Ann. N.Y. Acad. Sci.*, 189, 298, 1972.
4. Shahbender, R., R. Jerkart, H. Kurlansik and L. Onyshkevych, *IEEE Trans. Magnetics*, MAG-5, 427, 1969.
5. Chen, D., J. F. Ready, and E. Bernal G., *J. Appl. Phys.*, 39, 3916, 1968.
6. Eschenfelder, A. H., *J. Appl. Phys.*,41, 1372, 1970.
7. Rajchman, J. A., *J. Appl. Phys.*,41, 1376, 1970.
8. Newberry, S. P., T. H. Klotz, Jr., and E. C. Buschmann, *Proc. Nat. Electrn. Conf.*,23, 746, 1967.
9. Glenn, W. E., *J. Soc. Motion Picture Television Engrs.*, 69, 577, 1960.
10. Chen, A. C. M., *IEEE Trans. Electrn. Dev.*,ED-20, 160, 1973.
11. Cohen, M. S., *IEEE Trans. Mag.*,MAG-4, 639, 1968.
12. Smith, D. O., K. J. Harte, M. S. Cohen, S. P. Newberry and D. E. Speliotis, U.S. Patent No. 3,710,352, January 9, 1973.
13. Chapman, D. W., *Proc. IEEE Int. Computer Group Conf.*,56, 1970.
14. Kelly, J., J. S. Moore and P. R. Thornton, *NAECON '74 Record*, 55, 1974.
15. Heiman, R. and A. Waxman, *Digest IEEE Intl. Conv.*, New York, 152, 1970.
16. Cohen, M. S. and J. S. Moore, *J. Appl. Phys.*, 45, 5335, 1974.
17. Ellis, G. W., G. E. Possin and R. H. Wilson, *Appl. Phys. Let.*, 24, 419, 1974.