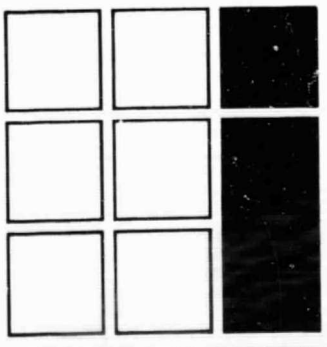
EDT-1

CASE FILE COPY



INTERMETRICS

Final Report

Engineering Study
for
A Mass Memory System
For Advanced Spacecrafts

By: Alex L. Kosmala, John P. Green, and
Fred H. Martin

August 1970

Prepared under Addition #1 to Contract
NAS 9-9763 by

INTERMETRICS, INC.
380 Green Street
Cambridge, Massachusetts 02139

Foreword

This document is the final report of an engineering study of mass memory systems for advanced manned spacecrafts. The work was performed as an extension of a Multiprocessor Computer Systems study, which was previously reported. This effort was sponsored by the National Aeronautics and Space Administration's Manned Spacecraft Center, Houston, Texas under contract NAS 9-9763. It was performed by Intermetrics, Inc., Cambridge, Massachusetts under the technical direction of Alex L. Kosmala.

The study program covered the period from December 1, 1969 through July 1, 1970. The Technical Monitor for the Manned Spacecraft Center was Mr. James P. Ledet.

- -

Table of Contents

		Page
Chapter 1.	Objective and Summary of Study	1
1.1	Introduction	1
1.2	Technology Review	2
1.3	Memory Requirements	4
1.4	Recommendations	5
1.5	Final Word	5
Chapter 2.	Review of Memory Technologies	7
2.1	Introduction	7
2.2	Ferrite Cores	9
2.3	Planar Films	28
2.4	Plated Wire	44
2.5	Block Oriented Magnetic Memory Techniques	61
2.6	Semiconductor Techniques	76
2.7	Beam Accessed Memory Technologies	115
2.8	Cryogenic Memory Technology	143
Chapter 3.	Mass Memory Requirements	148
3.1	Multiprocessor Design Introduction	148
3.2	Memory Performance Requirements	154
3.3	Memory Physical Requirements	160
3.4	Environmental and Operational Requirements	166
Chapter 4.	Evaluation and Recommendations	179
4.1	General Evaluation Criteria	179
4.2	Evaluation Strategy	182
4.3	Comparison of Memory Technologies	183
4.4	Comparison of Technologies for M1	183
4.5	Comparison of Technologies for M2	188
4.6	Comparison of Technologies for M3	193
4.7	Summary of Recommendations	200

List of Figures

Figure		Page
Chapter 2.		
2.2.1	Ferrite Hysteresis Loop	10
2.2.2	Ferrite Core Array	10
2.2.3	Checkerboard Wiring Pattern	12
2.2.4	3-D Ferrite Core Memory Organization	14
2.2.5	2-D Ferrite Core Memory Organization	16
2.2.6	Transition from 2-D to 2½-D Organization	18
2.2.7	2½-D Ferrite Core Memory Organization	19
2.3.1	Thin Film Hysteresis Loop	29
2.3.2	Thin Film Switching Astroid	29
2.3.3	Effect of Creep and Skew	32
2.3.4	Split Film Memory Element	37
2.3.5	Mated Film Memory Element	38
2.3.6	Coupled Film Memory Element	39
2.3.7	Post and Film Technique	41
2.4.1	Plated Wire Memory Element	45
2.4.2	Write, Read Pulse Sequence	47
2.4.3	Plated Wire Manufacturing Process	51
2.4.4	Plated Wire Tunnel Structure	54
2.4.5	Plated Wire Woven Array	55
2.5.1	Propagation of Domain in Magnetic Wire	63
2.5.2	Construction of Dynabit Element	64
2.5.3	Propagation of Magnetic "Bubble" Domain	67
2.5.4	Sonically Scanned Film Memory Schematic	71
2.5.5	Hysteresis Loop Behavior Under Strain	72
2.5.6	Rotation of Easy Axis Under Strain	73

Figure		Page
2.6.1	Basic Construction of MOS Transistor	80
2.6.2	Source-to-drain Conductance Characteristics for P-Channel MOS Transistor	81
2.6.3	Sequence of Processing Steps for CMOS Integrated Circuit	85
2.6.4	Principle of Charge-Coupled Device	87
2.6.5	Construction of P-Channel MNOS Transistor	89
2.6.6	Conductivity of Silicon Oxide and Nitride as a Function of Electric Field	91
2.6.7	MNOS Threshold Voltage Characteristic	95
2.6.8	Bipolar Memory Cell	95
2.6.9	MOS Inverter Circuit	97
2.6.10	Static MOS Memory Cell	98
2.6.11	Dynamic MOS Memory Cell	100
2.6.12	CMOS Inverter and Transmission Gate	100
2.6.13	CMOS Memory Cell	101
2.6.14	Connections to 1,024 Bit Array	104
2.6.15	Memory Cost Versus Chip Complexity	106
2.6.16	Cost Prediction for Semiconductor Memories	112
2.7.1	Elements of an Optical Memory System	117
2.7.2	Read-Only Holographic Memory System	118
2.7.3	Mechanical Beam Deflection System	130
2.7.4	Acoustooptic Deflection System	131
2.8.1	Superconducting Memory Element	144
2.8.2	Writing a '1' into Superconducting Loop	145
Chapter 3.		
3.1	Intermetrics Multiprocessor Series	149
3.2	Normalized Word Rate (W) vs M2/M1 Access Time Ratio (R)	155

List of Tables

Table		Page
Chapter 2.		
2.2.1	Core Memory Electronics Requirements	20
2.3.1	Characteristics of Some Thin Film Memories	42
2.4.1	Characteristics of Some Plated Wire Memory Systems	59
2.6.1	Current IC Performance Characteristics	102
2.7.1	Characteristics of Recording Media	127
Chapter 3.		
3.1	M2 Access Time (t_2) as Function of M1 Access Time (t_1) for 1 MIPS	156
3.2	t_2 as a Function of t_1 for 10 MIPS	159
3.3	1975 Memory Requirements	164
3.4	1980 Memory Requirements	165
3.5	Inner Zone Charged Particle Distributions	168
3.6	Outer Zone Charged Particle Distributions	169
3.7	Effect of Shielding in Synchronous Orbit	170
3.8	Absorbed Radiation from Inner Zone Particles	172
3.9	Absorbed Radiation from Outer Zone Particles	172

Chapter 1

Objective and Summary of Study

1.1 Introduction

Designs for the next generation of manned space vehicles are currently being formulated. They involve advanced computer systems performing comprehensive tasks of guidance, control, navigation, monitoring, data reduction, and communication. The typical aerospace computer that was employed for each of these roles in the past was of modest performance capability and possessed a limited storage facility which was considered to be just sufficient for each function. However, it is expected that the new generation of on-board computing systems will demand the on-line availability of vastly increased storage resources. The chief characteristics of future space missions that will contribute to the need for massive on-board storage facilities are:

a) The variety and complexity of the systems and the activities which the computing system is expected to service. Examples of these are vehicle control, life support, inventory management, scientific experiments, communications with the crew, with other spacecraft and the ground, etc.

b) The vast amounts of data that are expected to be generated on-board must be reduced and stored. Estimates of the bulk of data generated by the systems aboard an orbiting space base have been as high as 10^{11} bits per day [1].

c) The long duration of the planned missions of the future will require the computing system to be capable of performing leading roles in a large variety of missions, and of providing support to numerous and varied programs of scientific exploration. These tasks will involve a vast amount of supporting software which must be available on-board.

These aspects have been discussed in the Final Report to the first part of this contract. A computer system was proposed to enable these tasks to be accomplished. A feature of the design was its ability to be gradually improved and enlarged in an orderly fashion to satisfy the increasing computational demands of a developing space program. The

design, defined the necessity for a hierarchy of memory systems to span the conflicting requirements of nanosecond access times and billion bit capacities; it did not seem very likely that these requirements could be met by one memory system based on a single technology, at least in the foreseeable future. It was the purpose of the study summarized by this report to:

a) determine the characteristics required of each category of memory in the necessary hierarchy to enable it to be implemented by specific technologies;

b) determine from a detailed search and review which of the existing, or proposed, memory technologies would be the most suitable for implementation of each category;

c) perform these tasks in order to determine the feasibility of operational memory systems for advanced manned space vehicles for two distinct and separately considered time periods, namely 1975 and 1980.

This report is presented basically in three parts:

a) the review of memory technologies;

b) the derivation of requirements for the various categories of memory;

c) a comparison of technologies for each category, resulting in recommendations for 1975 and 1980.

Short summaries of the three sections are now given.

1.2 Technology Review

Chapter 2 presents descriptions of a number of technologies that have been used to implement memory systems. The greatest depth of detail has been reserved for those that are considered to show the greatest promise for space vehicle applications, especially in terms of the hierarchical requirements detailed in Chapter 3. The technologies reviewed are summarized below.

a) Ferrite Cores. This is the predominant technology today, accounting for over 95% of all computer memories. Despite the existence of a large supporting industry, and continued improvements in density, speed and cost, ferrite cores may lose their dominant position to other techniques at least for space applications, because of basic speed, size and power limitations.

b) Laminated Ferrites. A batch-oriented ferrite element with promising performance potential, but severe manu-

facturing difficulties.

c) Planar Thin Films. A concept with a long history of attention because of its high speed and potential for batch fabrication. Production difficulties with variations on the theme intended to improve density, power and noise levels have discouraged widespread use, except in some military applications.

d) Plated Wire. A more recently pursued variation of thin film technique, plated wire has been very actively developed by industry and government. Speed, low power and moderate density combined with automated production give this technique a strong potential.

e) Serial Memories. A collection of thin film and magnetic domain techniques which aim at high densities, low power and low cost at the expense of a block oriented organization. Some show commercial promise on account of low cost, but only the "bubble" domain concept has real promise for future mass storage in space.

f) Semiconductors. This large field is examined in some detail. Semiconductor integrated circuits will constitute a large proportion of any computing system, but only complementary MOS and the non-volatile variable-threshold MNOS are seen to have mass storage potential. Major problems are yield at the higher chip densities, and substrate interconnection techniques. A wide capacity, speed and power capability is covered by this field, and it is believed to become the major memory technology by the end of the decade.

g) Beam Technologies. Beam accessing is being investigated for the very large capacity systems of over a billion bits. The principal features are low inertia, high resolution and a departure from the discretely fabricated and interconnected storage medium. There are severe problems in the areas of erasable storage, of deflection systems and in the generation, detection and conversion of optical information. Most designs are at the conceptual rather than implementational stage. They involve considerable overhead equipment such as lasers, optical path control, CRT devices, and therefore must demonstrate high storage densities (up to 10^8 bits/in^2) to compensate for the otherwise inefficient use of space.

h) Cryogenic Memories. High storage densities and low noise levels are the incentive to superconducting memory cell investigation. Cumbersome overhead equipment and difficult environmental conditions have depressed interest in this field.

1.3 Memory Requirements

From a consideration of performance specifications of 1 million instructions per second (MIPS) for 1975, and 10 MIPS for 1980, the following definitions for three categories of memory are presented in Chapter 3:

a) M1. A high speed, random access, read and write, buffer memory, addressable to the level of one 30-40 bit word.

	1975	1980
Capacity	10^5 bits	10^6 bits
Access time	100-200 ns	40-60 ns
Data rate	5 x10^8 bits/sec	10^9 bits/sec
Volume	300 in^3	
Weight	15 pounds	
Power	15 watts	

b) M2. A large capacity read and write memory, randomly addressable to a 250-400 bit block.

	1975	1980
Capacity	10^7 bits	10^8 bits
Access time	2-4 μs	0.5-1.0 μs
Data rate	5 x10^8 bits/sec	10^9 bits/sec
Volume	1500 in^3	
Weight	100 pounds	100-200 pounds
Power	85 watts	

c) M3. A very large capacity, block-oriented read and write memory, addressable to a 10^4 to 10^5 bit block.

	1975	1980
Capacity	10^9 bits	10^{10} bits
Access time	10-100 ms	10 ms
Data rate	10^5-10^6 bits/sec	10^7 bits/sec
Block size	10^4 bits	10^5 bits
Volume	2-4 ft^3	
Weight	200-400 pounds	
Power	200-400 watts	

The physical and operational requirements imposed by the environment of a space application are also examined in this chapter.

1.4 Recommendations

A category by category evaluation of the technologies reviewed in Chapter 2 is presented in Chapter 4, and recommendations for the two time periods are made. The results are summarized below.

Category	Year	Technology
M1	1975	Plated wire
	1980	CMOS
M2	1975	Plated wire
	1980	MNOS/CMOS
M3	1975	1. Optical drum 2. "Bubble" domain
	1980	Holographic

1.5 Final Word

The prediction of technological trends, for even a short journey into the future, must at times take on the aspect of guesswork. For the innocent inquirer there are false signposts along the route pointing in directions that eventually turn into blind alleys. Even researchers in the field have pursued a memory phenomenon for as long as fifteen years before finally concluding that it would, in fact, never yield its promised potential.

The disparity among the estimates for the success of a particular memory technique, even from persons actively engaged in its development, indicates the uncertainty in this game. For example, a number of workers in the plated wire field, when asked in 1968 for their forecasts for 1975, gave estimates for this technology ranging from a 3% to 60% share of the memory market [2].

In our recommendations we have attempted to avoid being misled by the well known tendency among technology forecasters to make their predictions for the near future too optimistic, and for the long term, too conservative.

References for Chapter 1

1. Progress Review, Space Station Program: Phase B Definition, Document PDS70-1212, North American Rockwell, February 19, 1970.

2. Fedde, G.A. and Chong, C.F., "Plated Wire Memory - Present and Future", IEEE Trans. Magnetics, MAG-4, No. 3, September 1969, pp. 313-318.

Chapter 2

Review of Memory Technologies

2.1 Introduction

This chapter presents a detailed review of the major technologies that have, at one time or another, been investigated for their application to computer memories. No particular regard of the specific mass memory requirements that are presented in Chapter 3 is taken at this time, and no technology trade-offs with respect to these requirements are made until Chapter 4.

Space is devoted to material of a tutorial nature where this helps to promote an understanding of the technology. Several of the technologies are treated in considerable detail. This is sometimes due to their outstanding potential for mass memory implementation and sometimes, as in the case of planar thin film, because the principles and problems of the technology are shared by, and have relevance to, others.

The following technologies are described in this chapter:

Ferrite cores
Laminated ferrites
Planar thin films
Plated wire
Serial, block oriented techniques
Semiconductors
Beam accessed techniques
Cryogenic techniques

This list covers all technologies that were considered to have application to the implementation of massive storage systems for space. The complete range of variations on each theme is not always covered, but sufficient indication of the overall potential of each technique is presented. Some technologies are known to have been omitted. Among these are the following:

a) Acoustic Delay Lines. Interest continues to be shown in quartz and glass delay lines as small memories for CRT displays and desk calculators, but the potential for massive storage is considered to be near zero.

b) Ovonics. Some memory properties are claimed for the amorphous semiconductor, but the apparently very low densities achievable, the early stage of development, and the reluctant general acceptance of its claimed advantages rule it out of consideration. Ovonic devices appear to have great resistance to nuclear radiation, and this property is being studied under government sponsorship.

c) Magnetic Drum and Disks. This is the traditional mass storage technology. It was felt that the principles, capabilities, and limitations of moving magnetic medium devices were too well known to deserve a detailed technical review in this section. Some reference to the possibilities of space-qualified drums and disks is made in the evaluation of technologies in Chapter 4.

2.2 Ferrite Cores

2.2.1 Introduction

The principle of coincident current selection of the stored magnetization of toroidal ferrites was first patented by MIT in 1956 [1]. Since then the ferrite core has been continuously developed as a storage medium for computers. Its attractive properties have led to its dominance over all other technologies, to the extent that "core" has almost become synonymous with mainframe "memory". Over 95% of all computer memories manufactured today employ ferrite cores, and the capabilities of this technology form the standard against which competing methods must be measured.

Among the reasons for this state of affairs are [2]:

1) The ferrite core is an excellent storage device: it can be written into and read out of very rapidly, it is non-volatile and has a near-zero standby power requirement; it has become inexpensive to make with high degrees of reproducibility and uniformity.

2) Memory systems have been devised and turned into practical products with no drawn-out development stage.

3) Having been first on the scene, the ferrite core has had a comparatively long competition-free maturing phase.

The following sections give a brief description of the principles of operation. Also included is a review of the basic memory organizations that have been developed to address and sense the information from arrays of ferrite cores. These organizations have relevance for all memory techniques that rely on a matrix approach to gain access to desired information.

2.2.2 Principle of Operation

Some ferrites have a magnetization (M v. H) characteristic that is almost "square". This property can provide both memory and logic function. In the absence of a driving current (or field H) two remanent magnetizations of opposite sign M_a, M_b are possible, as in Figure 2.2.1. It is necessary to exceed a critical, or switch current I_f to cause the magnetization to change sign, say from M_a to M_b. A current I_p is insufficient to switch the magnetization; its only effect is to leave a slightly diminished remanent flux, M_a', due to the

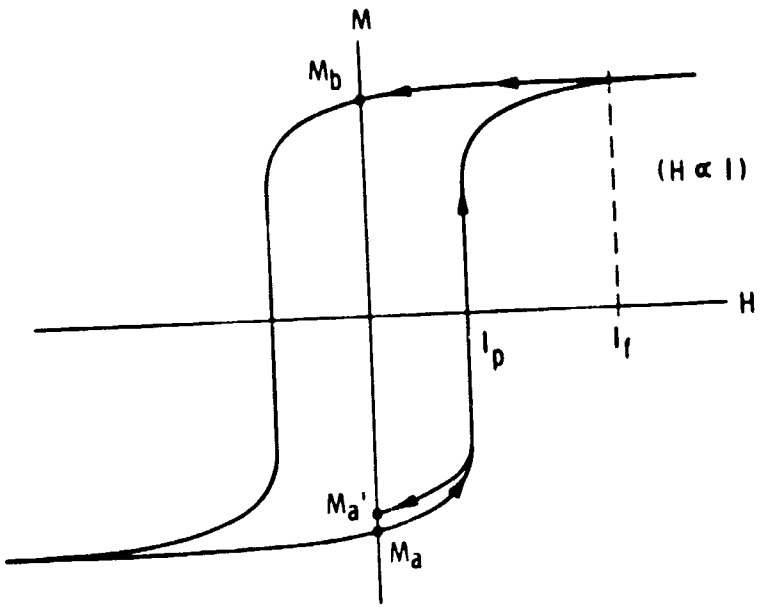


Figure 2.2.1 Ferrite Hysteresis Loop

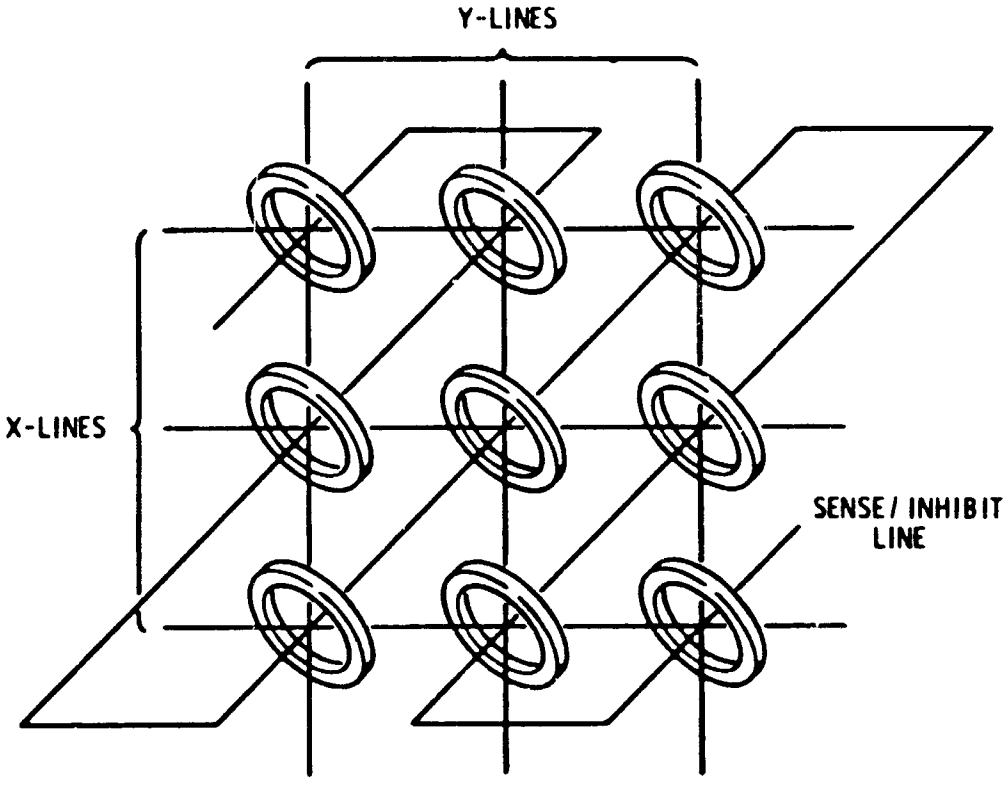


Figure 2.2.2 Ferrite Core Array

"non-squareness" of the hysteris loop. The material can therefore distinguish between "full" and "partial" select currents.

The ferrite core is normally an annulus, which provides a stored, closed, circumferential flux path. Early cores were over 100 mil O.D. (1 mil = .001 inch), but over the years reductions in size have occurred (to reduce the energy of magnetization) until today memories have been constructed of 12/7 mil (O.D./I.D.) cores [3]. Curiously, each new "standard" size that has been established can almost pass through the hole in its predecessor.

Switching speed is proportional to drive current, if the current is well above the threshold value. It is a complex function of drive current duration, because short pulses involve different magnetic mechanisms in the material than continuous currents. A 20 mil core switches its state in approximately 100 ns when excited by the field from a wire carrying about 1 amp.

Cores are arranged in rectangular arrays, with at least two wires threading each at right angles, as in Figure 2.2.2. The core's property of not switching in response to partial currents is the basis for selective access. The desired core location must be at the intersection of partially driven X and Y lines, a unique location where the fields associated with the currents combine to exceed the switching field. A third wire can be used, threading all cores in an array, to detect the large change in flux through a core as it switches in response to the "coincident" partial select currents. As mentioned before, a consequence of imperfect "squareness" in the B-H characteristic causes each partially selected core to contribute a small flux change, which is detected by the sense wire as noise. To prevent individual cores from additively contributing noise, the cores and sense wire are often arranged in a cancelling "checkerboard" pattern (Figure 2.2.3).

To detect a '1' a core must be switched from its '1' state to the '0' state. Information is therefore destroyed on read, and must be restored with an immediate write. To write a '1' it is necessary to switch the core from its '0' state, with currents of opposite polarity to those required to read a '1'. Bipolar drive circuits are obviously necessary. To write a '0' the core must be inhibited from changing its '0' state; this can be accomplished by opposing the full write select current with a partial inhibit current. Sometimes this is done with a fourth, inhibit line, and sometimes the sense line is used with additional external logic.

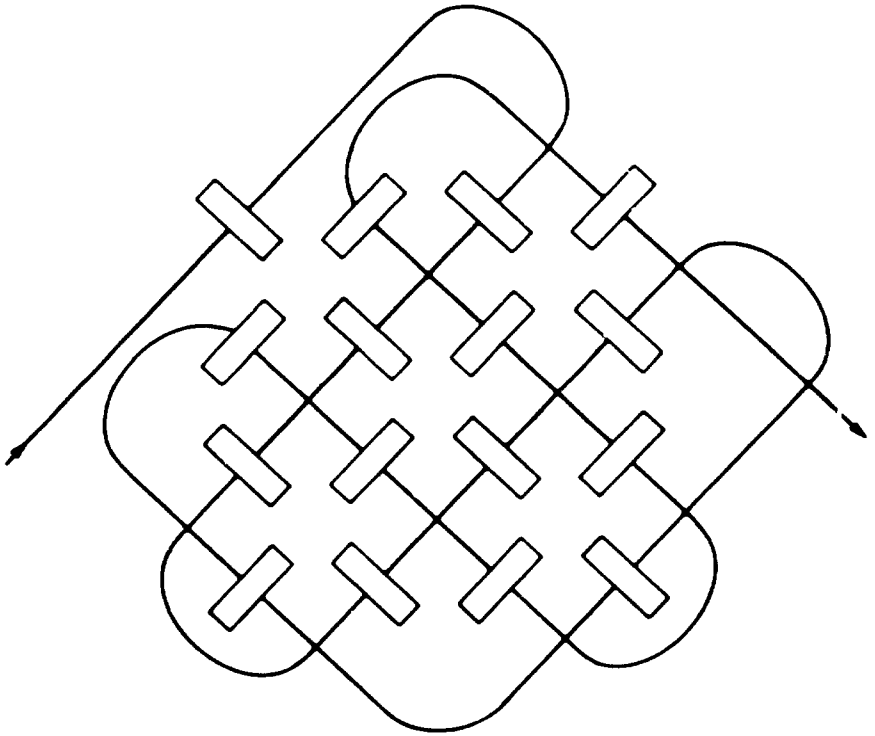


Figure 2.2.3 Checkerboard Wiring Pattern

2.2.3 Memory Organizations

A basic problem is how to get to the desired core locations to read or write information. Three different approaches to addressing core arrays have been devised to serve the conflicting requirements of speed, drive circuit complexity and power.

2.2.3.1 3-D Organization

This requires the least amount of address and drive circuitry and for medium performance memories is the most commonly used. The memory is organized into M square planes (where M = bits/word). Plane M_1, for example, contains bit 1 of all words in the memory (total number of words = N/M, where N = total capacity in bits). Each plane therefore contains N/M cores, and is addressed by X and Y word drive lines, where XY = N/M. Bit 1 of a particular word could be read by selecting the X and Y drive lines of array M_1 that intersect at the location corresponding to the address of the word. Connecting corresponding X and Y lines of each of the M memory planes together allows one X,Y selection to read in parallel all M bits in a particular word. Because it is possible to access an individual bit in one plane, the 3-D approach is sometimes called a bit-organized memory. Other configurations deliver more than one bit at a time from each array. Figure 2.2.4 illustrates a 3-D memory. A selection matrix is shown driving the $\sqrt{N/M}$ X- and $\sqrt{N/M}$ Y- lines of the memory. Such an added level of address selection is often used in large memories to reduce the amount of drive circuitry, although each level contributes a delay which slows down overall response. In addition to the X-wire and the Y-wire, a sense wire to detect the switching of the selected core during a read operation must be provided. The sense wire can double as the inhibit line during a write cycle, but a fourth wire is sometimes added. Since this wire threads all cores, it is long and resistive. It dissipates heat during the inhibit cycle, when it must carry a half select current, and it is the major cause of delays in the 3-D memory because transients take longer to decay in the inhibit-sense line than any other.

The chief properties of the 3-D organization are:

a) Coincident current selection is performed for both read and write: coincidence of current on two wires is required for a read, and on three wires for a write.

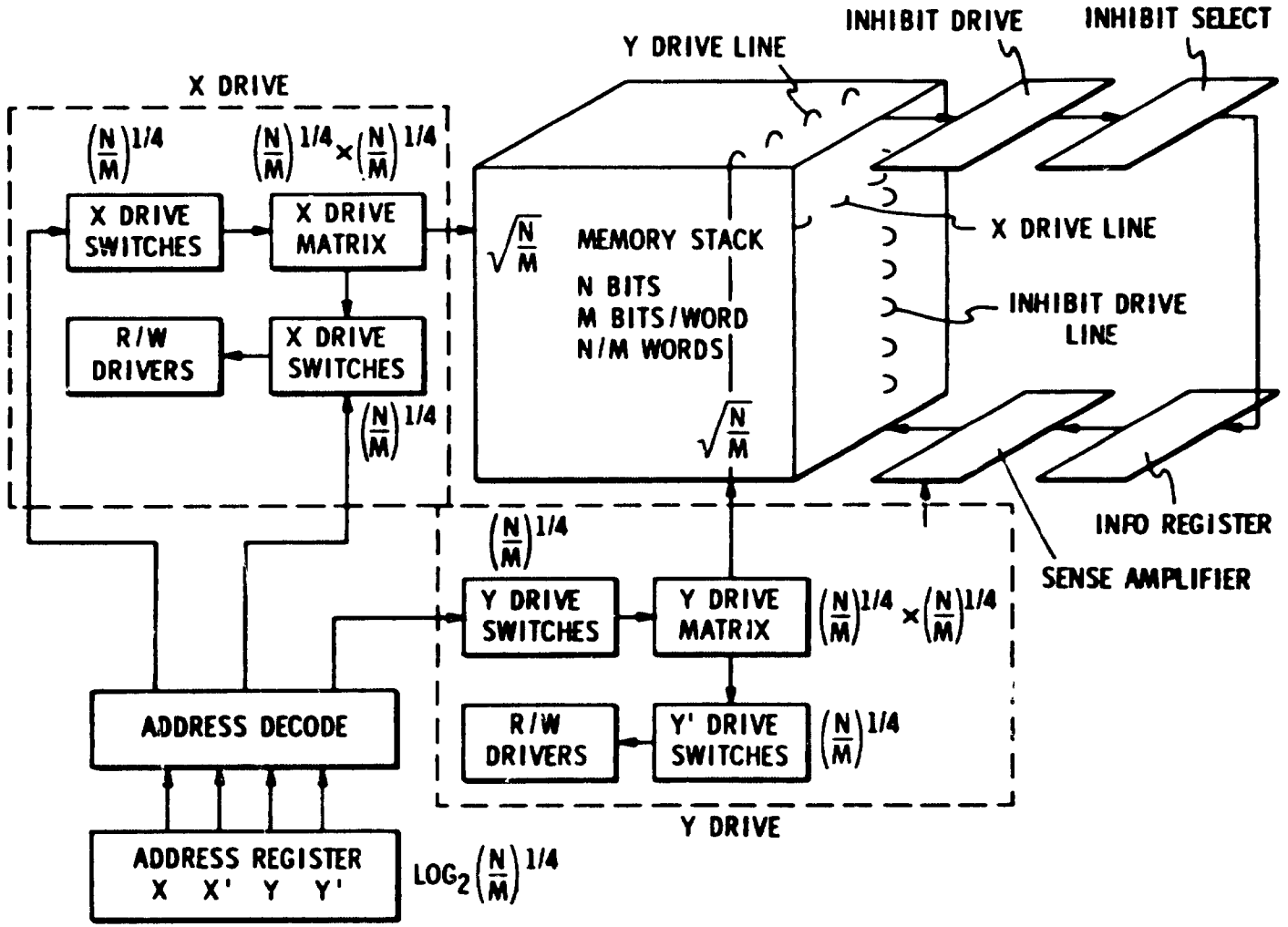


Figure 2.2.4 3-D Ferrite Core Memory Organization

b) The X,Y drive lines are used only for addressing, and the M (or Z) lines are used only for sensing and/or inhibiting.

c) Maximum use of the inherent decoding property of partially addressed cores is made to reduce circuit complexity, e.g., total number of XY drive circuits is $4(N/M)^{\frac{1}{2}}$, which becomes a really significant factor with large capacity memories (see Figure 2.2.4).

Its drawbacks are:

a) Relatively tight drive current tolerances are necessary to provide satisfactory core selection throughout the memory by the coincidence principle.

b) The limit on the switching current imposed by the maximum partial select current does not allow the full speed of the core to be utilized.

c) The lengthy sense/inhibit line contributes a major portion of the total system delay.

2.2.3.2 2-D Organization

This configuration is dedicated to speed. Cores are arranged in a rectangular plane; one side is length M, the number of bits in a word, and the other is N/M, the number of words. Access to the contents of a word is gained by pulsing one of the N/M word lines with a full select current. All cores on that line that were '1' switch to '0', inducing corresponding voltages in the bit/sense lines, and all that were '0' stay '0'. As mentioned before, switching speed is proportional to drive current. In the 2-D (or linear select) memory coincidence is not a requirement on read, and therefore the full potential of the core can be exploited by overdriving the word current. For writing, the word line drive is reversed in polarity, and all uninhibited cores are switched back to '1'. Those required to stay '0' are inhibited by an opposing coincident current on the appropriate bit/sense line. The chief characteristics of this organization are:

a) Coincident currents on two wires are required only for writing; for reading, cores are treated purely as storage elements and have no selection function.

b) Only two wires per core are needed. This allows smaller cores to be used, with further advantages in speed.

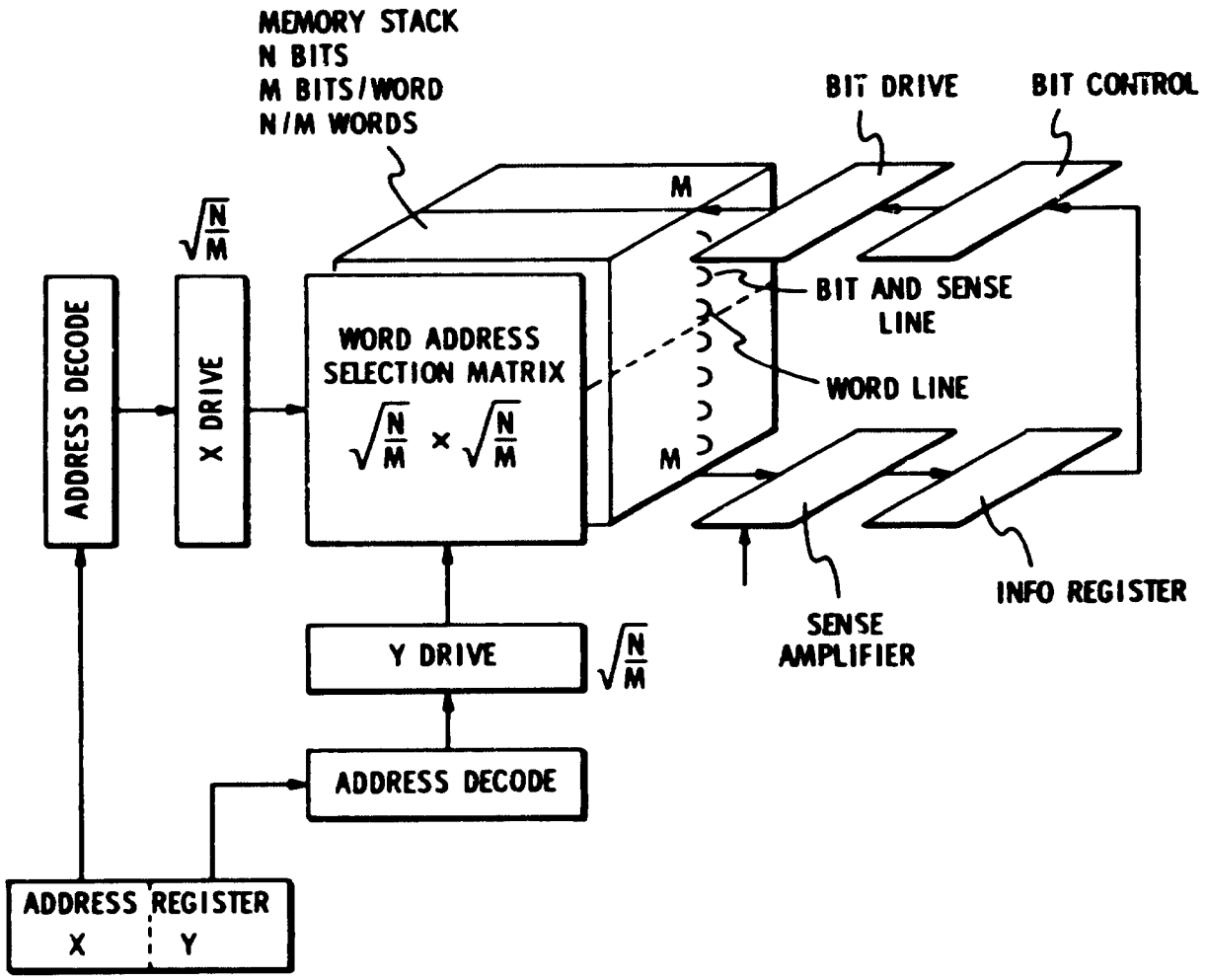


Figure 2.2.5 2-D Ferrite Core Memory Organization

c) The word lines are used only for addressing and the bit lines are used only for sensing and inhibiting.

d) The sense wires in a 2-D array thread only one excited core: this means the signal-to-noise ratio in a 2-D memory is inherently better.

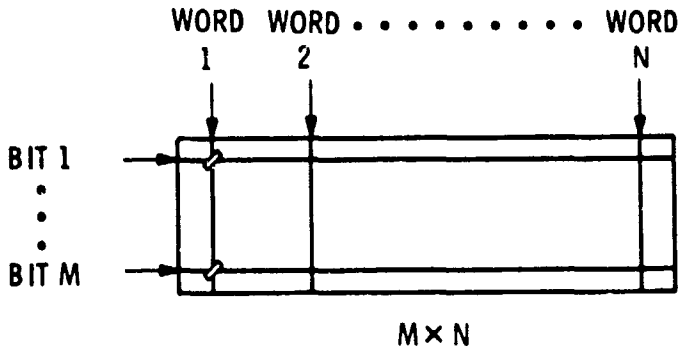
The very oblong symmetry of the 2-D organization is disguised by folding the long word side into a rectangular matrix with equal rows and columns of $(N/M)^{\frac{1}{2}}$ elements, as illustrated in Figure 2.2.5. Address decoding then necessitates only $2(N/M)^{\frac{1}{2}}$ drive circuits, which is, however, for large memories, a considerably greater complexity than for the 3-D configuration.

2.2.3.3 2½-D Organization

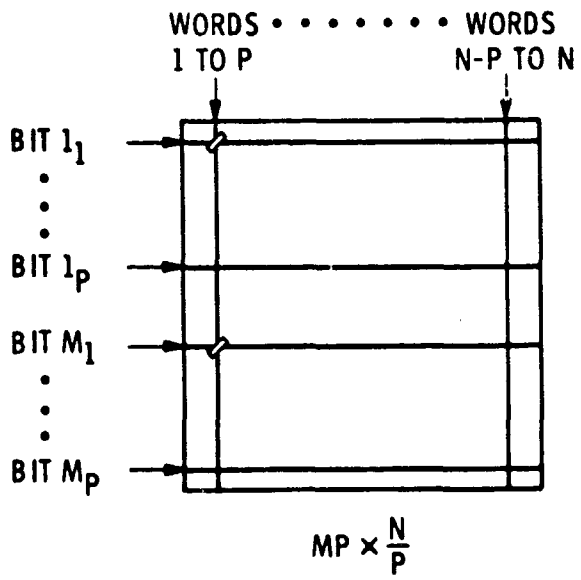
This organization attempts to combine the better points of both the 3-D and 2-D approaches (hence the name), without incurring any faults of its own. The technique dates from 1954 but was not used as a solution to the performance/complexity problem until IBM revived it for their Large Core Storage.

It is essentially a two-dimensional approach, but relieves the drive circuit complexity by "squaring up" the geometry of the pure 2-D. Figure 2.2.6 illustrates the transition. The significant characteristic of the resulting array is that the Y or bit dimension handles both address and data information, unlike the pure 3-D or 2-D. The planar representation of Figure 2.2.6 is "up-ended" for practical reasons, to give the organization of Figure 2.2.7. The memory is organized into M bit planes, each array having an aspect ratio W/L, where WL = N/M.

Reading is accomplished by coincident currents of $-I_p$ on the appropriate W line and on each of M appropriate L lines. Cores at the M intersections are set to the '0' state. Effects of flux changes are read out by a sense wire common to all cores on a bit plane; or by the same L line by separating out sensed signals from drive noise. The latter is a slower technique, but allows 2½-D, 2-wire construction with its consequent higher packing density, and is the method favored for mass stores where speed is of less essence. The real advantage of the 2½-D approach lies in its fast write. There is sufficient circuit redundancy in the y-direction to provide both address selection and data insertion. A word line is pulsed with current $+I_p$. The y-address and bit information are logically combined and sent to the y-drivers. To cause a '1' to be stored a $+I_p$ current is added to the full word



a) 2-D schematic



b) 2½-D schematic

Figure 2.2.6 Transition From 2-D to 2½-D Organization

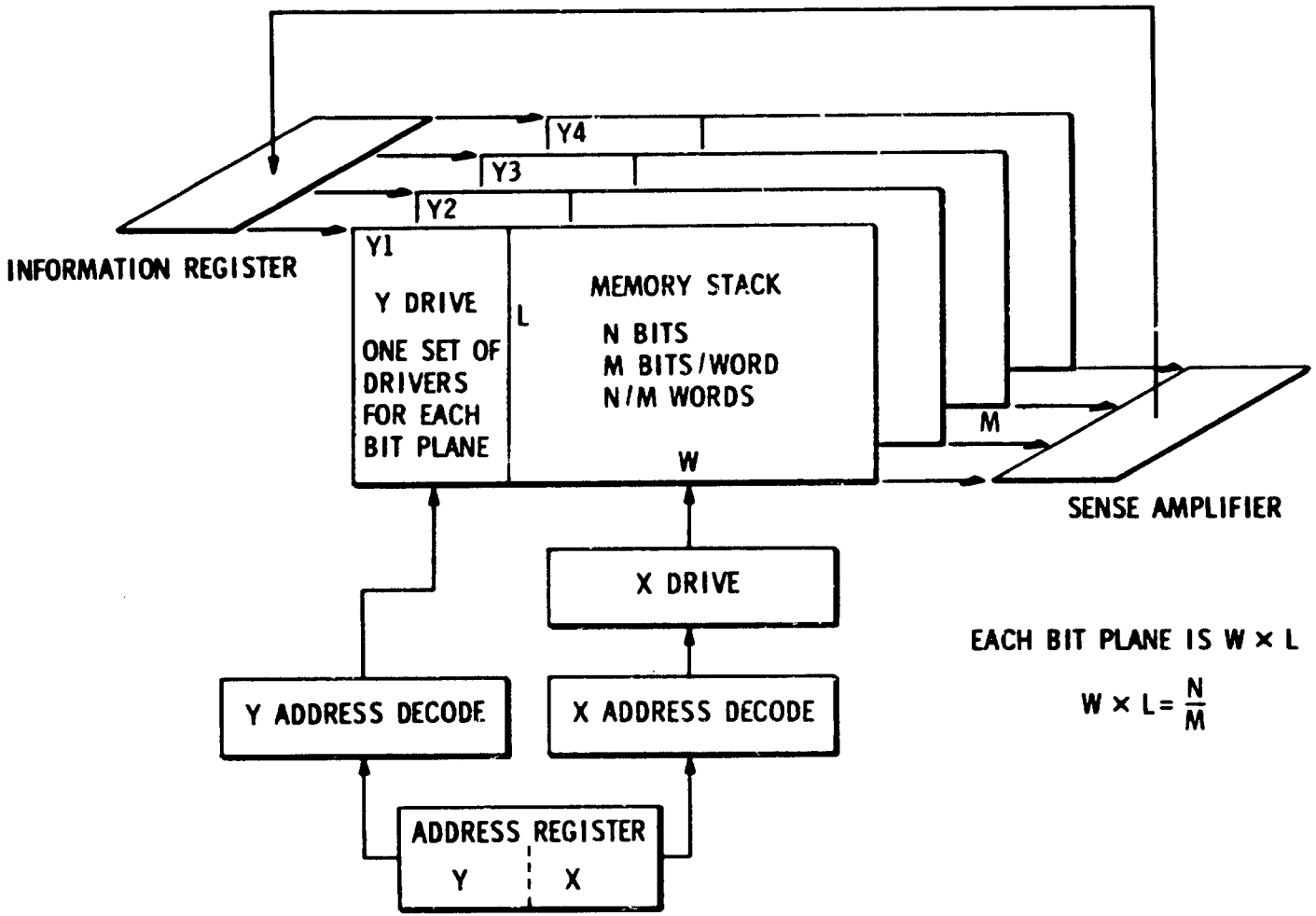


Figure 2.2.7 2½-D Ferrite Core Memory Organization

current in the appropriate L line in each bit plane. Intended zeros are left unselected. This provides for faster switching than for the additive half-select currents of the 3-D configuration, because the latter requires built-in delays to ensure "bracketing" of the X and Y pulses by the inhibit pulse.

The 3-D inhibit line recovery problem is eliminated since there is no inhibit line. Of the four states possible by combining X and Y $\pm I_D$ drives, only two are used: read and write. Practical memories use the other two to effect a x2 word address circuit economy, by a physical rearrangement of the bit planes [5]. Shorter drive and sense lines than for the equivalent 3-D or 2-D memory can be employed.

2.2.3.4 Summary of Drive Electronics

The quantity of drive electronics, for each of the above configurations is summarized in Table 2.2.1. The switch generates the bipolar drive current, and the driver routes the current to the appropriate wire in the memory.

	Word		Bit	
Configuration	Switch	Driver	Switch	Driver
2-D	$2(N/M)^{\frac{1}{2}}$	$2N/M$	M	-
$2\frac{1}{2}$-D (1)	$2\ W^{\frac{1}{2}}$	W	$2M\ L^{\frac{1}{2}}$	ML
3-D	$4(N/M)^{\frac{1}{4}}$	$2(N/M)^{\frac{1}{2}}$	M(2)	-

Notes: (1) For $2\frac{1}{2}$-D: $W \times L = \frac{N}{M}$

N = bits total
M = bits/words

(2) For 3-D this is an inhibit switch

Table 2.2.1 Core Memory Electronics Requirements

2.2.3.5 Configuration Capabilities

The 3-D organization of ferrite cores is used where performance and/or capacity is not critical, but the cost

associated with electronics is of prime importance. For very high speeds and/or large capacities, high drive currents, small cores and high densities are required, necessitating consideration of the 2-D and 2½-D approaches. 2-D is employed where high speed is paramount and/or where moderate capacity does not pose an impossible penalty in addressing, drive, and sense electronics costs. 2-D memories have most frequently been used as high-speed scratch pad storage of less than a few thousand words.[5] However, an experimental ferrite core memory of 5.9×10^5 bits built to explore the feasibility of very high speeds achieved a 110 ns cycle time using a 2-D, 2 wires per core, 2 cores per bit configuration[3].

The 2½-D does not incur the drive electronic expense of the 2-D, nor does it require the careful tolerancing and compensation of drive currents necessary with the 3-D approach. It is therefore popular in the larger system, especially since it allows the use of 2-wire arrays. Some sacrifice in cycle time is inevitable when the third, sense wire is omitted, since delays must be allowed for transients to decay; but for large core storage this is not of prime importance. An example of 2½-D core memory organization is the IBM 2361 Large Core Storage which contains 16.8×10^6 bits and operates with cycle times of 8 μs.

2.2.4 Ferrite Core Status and Trend

2.2.4.1 Current Status

Cores have steadily decreased in size since their inception. Today's standard sizes as used in commercially available arrays are 22, 20, 18 and 16 mil outside diameter; cores down to 12 mil O.D. and 7.5 mil I.D. have been quantity produced and incorporated in operational memory systems [3].

The trend towards smaller cores is due to the following interrelated reasons:

1) Smaller cores can be made to switch faster, without increasing the drive current, because the shorter magnetization path length results in a higher switching field.

2) The smaller cross sectional area reduces the inductance of the drive line, because the contributions from individual cores form the major portion of the total inductive load. The result is lower transmission delay, back voltage, and power dissipation.

3) The smaller dimensions, coupled with 2) above, enable denser arrays to be constructed. The reduced drive line dimensions compensate in part for the increased resistivity of the smaller diameter lines.

4) The increased surface-to-volume ratio of the smaller core improves its ability to dissipate the heat generated during switching.

The overall memory speed is strongly influenced by an improvement in core response, because the core switching time is a large fraction, up to 50% to 75%, of the basic memory cycle. Since core switching is destructive, a full memory read cycle must include a restoration of the destroyed information. The resulting full read cycle is therefore three to four times longer than the core switching time. Small cores present manufacturing problems. The ferrite core owes its popularity to massive automation of the processes of manufacture, testing, and assembly. For smaller cores to become standard practice, this machinery needs to be developed to a higher degree of performance.

The magnesium-manganese ferrite core is fairly temperature sensitive and requires compensation of drive currents for even narrow temperature ranges (25°C - 50°C). Lithium ferrite is now being used, especially in military systems [7]. In this material monovalent lithium displaces the usual trivalent manganese in the complex ferrite molecular system. It has a higher Curie temperature: 400°C versus 150°C to 200°C, and consequently a lower temperature coefficient: -0.16% per °C against -0.36%. In this respect nickel-iron memories have an advantage over ferrite core, since permalloy is even less temperature sensitive: about -0.07% per °C.

Drive currents are typically 400 ma to 600 ma, half select. Drive line impedances are about 120 ohms, which means 50-70 volts must be switched. This is outside the scope of integrated circuitry at the present time. Work is being directed at lower switching levels and wider temperature capability, but the penalty is switching time: 4μsecs appears to be the limiting speed.

The core/wire assemblies must be supported so as to be capable of:

1) Surviving the mechanical environmental factors of shock, vibration and acceleration. This requirement is obviously more serious in military and aerospace memories, and has contributed to a reluctance to employ advanced core techniques in these applications.

2) Conducting away the heat generated in switching cores and current carrying conductors, and providing a uniform temperature distribution throughout the stack (local hot spots defeat the temperature compensation of drive currents). Some high performance stacks employ a cooling liquid circulating in a hollow frame [3], but conduction and air cooling are the usual mechanisms.

3) Allowing the stacking of arrays into a module. Currently planes can be stacked at about 12 to the inch.

2.2.4.2 Ferrite Core Trends

Most of the trends have been indicated already:

1) Smaller cores for higher speeds and densities.

2) Lower drive currents for IC compatibility and reduced power consumption.

3) New materials to achieve higher switching speeds, wider temperature tolerance and easier manufacture [8].

Today's core memory has a barrier at about 500 ns cycle time. Of this 100-200 ns is absorbed by the ferrite core itself, and the remainder is contributed by geometry and circuitry. Exotic techniques, i.e., partial switching of the magnetization by very short, hard pulses, and two cores per bit to improve the signal to noise ratio in a fast, noisy system have been employed to achieve a 110 nanosecond cycle time [3]. But these are expensive devices, which together with the necessary, but electronics-consuming, 2D-2W organization, and cooling problems, are the penalties of higher speeds.

Packing densities in core arrays available today are in the range 1000-3000 bits/sq.in., depending on core size and array organization. Experimental arrays have been built from 12 mil cores spaced at 7.5 mil in the bit direction, and 15 mil in the word direction [3], with a resulting density of 8,500 cores/sq.in. Even though the volumetric efficiency reported in reference [3] deteriorated to 3,900 bits/cu.in. when the plane arrays were assembled, densities in excess of 50,000 bits/cu.in should be realizable if the requirement for very high speed were to be relaxed.

The one unavoidable drawback of the simple toroidal core is the destructive nature of the information readout. Significant future developments to obviate this are very unlikely,

because the necessary complex flux linkage must be provided by a multiapertured geometry, exemplified by the Raytheon BIAX, which is difficult to manufacture in a miniaturized form.

2.2.5 Planar Ferrite Technology

2.2.5.1 Introduction

Because of the overwhelming success of the toroidal ferrite core as a memory element, attempts have been made to develop a batch fabrication technique for this technology. The incentive to batch fabrication is common to other technologies, i.e., lowered costs due to a non-discrete manufacturing process, and a higher reliability promised by fewer processing steps and reduced parts counts. However, only one or two attempts have been made to develop a planar ferrite technique to the point where an experimental memory system could be built [9]. Efforts have been directed in two directions, which differ according to the manner in which the word and digit conductors are arranged. The earlier technique provided apertures in the ferrite sheet through which plated wiring could be lead [10], but was not pursued. The other technique is to sandwich orthogonal word and digit lines between laminates of ferrite; this has received considerable attention from at least two manufacturers, and will be described in some detail.

2.2.5.2 Laminated Ferrite Technique

Reference [11] describes how a ferrite laminate is formed. The structure consists of two orthogonal sheets of ferrite, each with a pattern of parallel, noble-metal (platinum, palladium, gold) conductors on it, separated by an insulating sheet 0.5 to 1.0 mil thick. The overall thickness of the finished laminate is typically 6-8 mils. Conductor dimensions vary with manufacturing method from 2.4 to 6 mil in width, 0.5 to 1.5 mil in thickness, and are spaced from 10 to 20 mils apart. The conductors are formed in the ferrite either by rolling a ferrite slurry over a conductive paste pattern on a glass substrate and allowing it to dry, or by filling the grooves formed in a ferrite sheet by phonograph record techniques or by embossing, with conductive paste. The laminate is formed by subjecting the ferrite sandwich to a pressure of 2000 p.s.i. at 90°C, after which it is fired for 10 hours at 1260°C. The ferrite system reported in reference [11] was $0.38\ MgO\text{-}0.19MnO\text{-}0.05ZnO\text{-}0.38Fe_2O_3$. This composition was found to possess the best compromise of the several characteristics cited as desirable for low power, high density ferrite memories:

1)	Squareness of the hysteresis loop (Br/Bm)	≥ 0.9
2)	Remanent flux density (magnetization)	≥ 1000 gauss
3)	Coercive force	≤ 1.5 oersted
4)	Curie temperature	$\geq 300°C$
5)	Resistivity	$\geq 10^6$ ohm-cm
6)	Grain size	≤ 10 μ

Low grain size and high resistivity are desirable in order to achieve minimal conductor widths and spacings, without incurring excessive current leakage between adjacent conductors or unwanted cross-couplings. Unfortunately, grain size and coercivity are related inversely, so that great care must be taken to fire the laminate at the optimum temperature, of which the grain size is a function.

Ferrite sheets formed from a wet slurry shrink in thickness by a factor of about 6 upon drying, and shrink in linear dimensions by a further 20% on firing at high temperature. Neither of these changes can be exactly predicted or controlled, and the latter shrinkage is a limiting factor in the production of laminates. There is a tendency for the conductors to fracture under the action of the differential changes in geometry during firing; the resulting conductor spacings, being variable and non-predictable, make batch-oriented connection techniques for laminated ferrite planes very difficult to implement.

Ferrite sheets cannot be made greater than about one square inch in area, which limits arrays to about 64x64 intersections for a 15 mil conductor spacing. To establish connections the conductors must be bared by air-abrading away the surrounding ferrite at the edges of each sheet, a process which is liable to result in damage and does not lend itself readily to automation.

Formation of ferrite sheets by pressure sintering could minimize the geometry change during firing; this process would also allow the use of solid, prefabricated conductors which could project beyond the edge and would therefore eliminate the airbrasion step of manufacture.

2.2.5.3 Laminated Ferrite Array Performance

The conductors of the laminated ferrite array are about 1 inch in length and have a resistance of about 2.5 ohms. But the heavy inductive and capacitative loading imposed by the

surrounding ferrite causes the line to exhibit a characteristic impedance of about 200 ohms at 10 MHz. The output voltage is of the order of 1-2 millivolts, but the unusually high self inductance and capacitance cause back voltages of up to 20 millivolts per intersection, which poses a problem to the read electronics. However, it is possible to achieve adequate operation with word write/read currents of only 60/100 ma, making operation with MOS I.C.'s feasible. Access times are determined by digit/sense line decay times, which in turn are influenced by data patterns and termination impedances: values between 100 ns and 250 ns have been measured.

Laminated ferrite arrays suffer from a number of performance disadvantages:

1) For reliable operation and good signal to noise ratios, two conductor intersections/bit must be used, which diminishes the possible packing density.

2) The read out is, like the toroidal ferrite core, destructive.

3) Being basically a 2-wire system, only 2-D operation is possible with a consequent heavy demand (in high capacity systems) for driving and sensing electronics.

4) Continuous high speed operation of the same cross-over point causes localized heating and local changes in electrical and magnetic characteristics.

The output signal amplitude and the switching speed can be improved by NRZ digiting (in which only state changes produce current switching), and by partial switching of the ferrite, but both techniques add to the electronic complexity.

2.2.5.4 Summary

Laminated ferrite techniques have achieved arrays which exhibit high speed (100-200 ns cycle time), low power (30/100 ma digit/word drive current), and densities between 2000 and 4000 bits/sq. in. The arrays cannot, however, exceed about 4000 bits in capacity because materials difficulties limit the size of ferrite sheets to about a square inch. The feasibility of constructing systems of 10^7 to 10^8 bits, with speeds of 2 μs, dissipations of 100 watts, and a volume of 300 cu. in. has been suggested [9], but as yet no system of this size has been implemented.

This technique has one or two significant problems. First, there are severe manufacturing difficulties; there is a strong dependence of the performance characteristics on the control of a large number of processing variables, from preparation of the

ferrite compound through to the finished laminated array. Some of the processing steps are tricky, e.g., exposure of electrical connections by airbrasion. Secondly, the poor control that can be exercised over physical dimensions such as conductor spacing, conductor separation at cross-overs, etc., makes mass assembly techniques nearly impossible, reduces yield and maintains a high level of cost. New techniques, such as pressure sintering, have been suggested to alleviate some of these problems, but as yet have not been proved.

References for Section 2.2

1. Forrester, J.W., U.S. Patent No. 2,736,880, February, 1956.

2. Russel, L.A., et al, "Ferrite Memory Systems", IEEE Trans. Magnetics, MAG-4, June 1968, pp. 134-145.

3. Werner, G.E., and Whalen, R.M., et al, "A 110-Nanosecond Ferrite Core Memory", IBM Journal, March 1967, pp. 153-161.

4. Brown, J.R., "First-and Second-Order Ferrite Memory Core Characteristics and Their Relationship to System Performance", IEEE Trans. on Electronic Computers EC-15, No. 4, August 1966.

5. Gilligan, T.J. and Persons, P.B., "High Speed Ferrite 2½-D Memory", Proc. FJCC, Vol. 27, 1965, pp. 1011-1021.

6. Brown, J.L., et al, "IBM System/360 Engineering", Proc. FJCC, Vol. 26, 1964, pp. 205-231.

7. Rickard, B.W., "Military Masses Its Cores for Battlefield Conditions", Electronics, June 9, 1969, pp. 122-124.

8. Agajanian, A.H., "Li-Mn Ferrites for High-Speed Computer Memory Applications", IEEE Trans. Magnetics MAG-6, March 1970, pp. 90-95.

9. Shahbender, R., "Laminated Ferrite Memories - Review and Evaluation", RCA Review, June 1968, pp. 180-198.

10. Rajchman, J.A., "Ferrite Apertured Plate for Random Access Memory", Proc. IRE., Vol. 45, No. 3, 1957, pp. 325-334.

11. Gordon, I., et al, "An MOS-Transistor-Driven Laminated Ferrite Memory", RCA Review, June 1968, pp. 199-229.

2.3 Planar Films

2.3.1 Introduction

The rotational switching of the remanent magnetization of thin anisotropic magnetic films has been investigated since 1955; the main incentive to their development was the achievement of a batch fabrication process that could rival the ferrite core memory in cost. Early core memories were difficult to produce: ferrite toroids were tricky to fabricate and test in large numbers, and wire stringing was not automated to the extent it later became. This motivation changed, because core technology developed to become an industry within the computer industry, and has always maintained a price-performance superiority over all other techniques. Work on films has continued, however, spurred by the promise of a superior performance and the hope of eventual commercial viability.

Many of the principles and problems of planar and cylindrical anisotropic magnetic films are similar, so much of the following explanation is applicable to the plated wire technique described in a later section.

2.3.2. Basic Principles of Thin Film Memory

Thin magnetic films can exhibit anisotropic behavior; for example the characteristic B-H hysteresis loop changes as the film is rotated in the applied field. The magnetization in a thin film element has preference for a particular orientation, and this direction is referred to as the "easy" axis, with the "hard" axis being orthogonal to it. The easy axis magnetic properties are affected by the application of hard axis fields. Figure 2.3.1 illustrates the changing behavior of the easy-axis B-H characteristic of a thin film element as an increasing hard-axis field is applied. The variation in the shape of the hysteresis curve represents flux changes which can be used to induce voltage signals in sense windings placed near the element to determine the presence and sign of a remanent magnetization. The hard axis field is usually controlled by word selection logic and the easy axis parameters are associated with the bit or digit organization of a memory. If the element is originally magnetized as at point (a) in Figure 2.3.1, the application of a word field H_w decreases the remanent magnetization to point (b). The corresponding change in easy axis magnetization flux will induce a voltage in a sense wire given by Faradays Law:

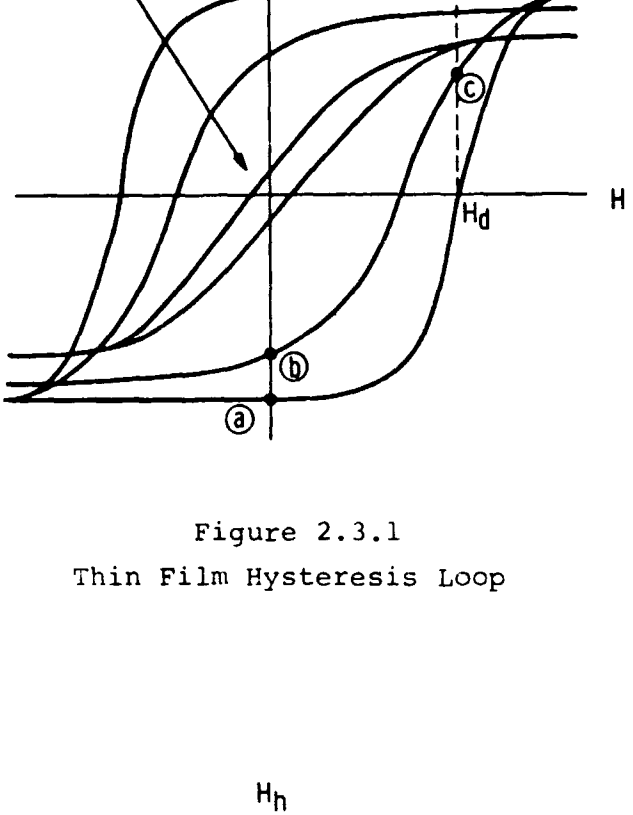


Figure 2.3.1
Thin Film Hysteresis Loop

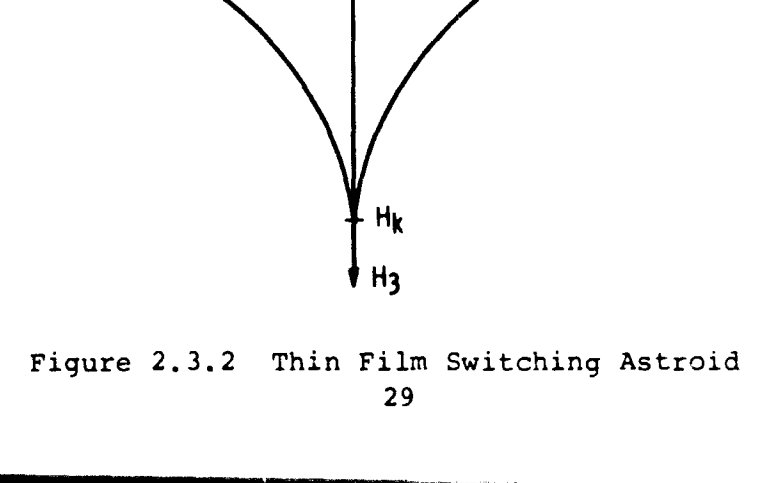


Figure 2.3.2 Thin Film Switching Astroid

$$V(t) = -N \frac{d\phi}{dt} .$$

Removal of the word field returns the element to point (a). This is an example of a non-destructive readout of information (NDRO). If a digit field H_d is simultaneously applied, the operating point moves from (a) to (b) to (c). Removal of the word field then shifts the magnetization of the element to (d). Finally, removal of the digit field leaves the remanent magnetization at (e), a situation which corresponds to the writing of a bit into the element, of opposite sign to that originally stored.

Note that the easy axis field necessary to switch an element decreases with increasing hard axis field. The magnitudes of these fields, in the case of an ideal film, are related by:

$$H_e^{2/3} + H_h^{2/3} = H_k^{2/3}$$

where H_k is defined as the anisotropy field constant. This relationship generates the switching astroid shown in Figure 2.3.2. For a typical 1000 Å Ni-Fe film H_k is about 5 oersted. If the vector sum of the applied fields exceeds the astroid switching boundary, e.g. H_1 in Figure 2.3.2, the remanent field will align itself with the easy axis field H_{e1}, no matter where it was before. If the sum does not exceed the boundary, e.g. H_2, no switching takes place. A hard axis field that exceeds H_k, e.g. H_3, will reduce the remanent field to zero. This last is the destructive read-out case (DRO), which is usually employed for its speed and greater freedom from non-ideal film characteristics than NDRO.

The magnetic thin film's main attraction has been its high switching speed. For films less than 3,000 Å the switching mechanism is by magnetization vector rotation rather than domain wall motion, and occurs in 1 to 3 nanoseconds. Actual memory response time is therefore almost totally determined by geometry and drive electronics delays.

The amount of magnetic energy switched is low, approximately 10^{-10} joules; and owing to fairly loose coupling of the flux change into the sense conductors, typical output signal voltages are low, usually 0.5 to 1.5 millivolt. The currents in word and drive currents for the simple, open element geometries are fairly high; 600-1000 ma and 50-100 ma respectively. The word currents are outside the range of today's integrated circuits.

Much research into films is devoted to bringing about integrated circuit compatibility.

A memory is organized by arranging the individual film elements in an array upon some substrate, with orthogonal word and digit conductors intersecting at each bit location. The elements need not be discrete areas of magnetic film, but can be locations in a continuous sheet. Many variations on this theme have been explored, most of them attempts to overcome or circumvent one or more of the problem areas discussed in the next section.

2.3.3 Thin Film Problems

The departure of magnetic thin films from the ideal has been extensively studied and reported on in the literature over the last 15 years.

2.3.3.1 Creep

The thin film element does not exhibit homogeneous magnetic properties [1]. When a bit is written into a discrete memory element, the edges of the element do not switch together with the center, unless the same bit is repeatedly written into that location. Microscopic flaws in the film, irregularities in the film edges and in the substrate cause local sensitivity to field disturbances, which then act as nucleation centers for magnetic domain formations. These domains tend to grow and propagate under the influence of varying fields. Repeated switching of an element with the same bit, say a '1', can therefore change its anisotropic behavior, and unsymmetrical or even erroneous output can be obtained when a '0' is then cycled in the same position. This characteristic is called creep, and is especially serious in continuous film memories as, for example, plated wire, because adjacent bit locations can be encroached upon by the errant domain wall motion under the influence of disturbing fields. The effect of creep is to reduce the boundary for unwanted switching from that of the ideal astroid. This second boundary, termed the creep threshold and illustrated in Figure 2.3.3, is the limiting field which, even if applied an infinite number of times, will not cause switching to occur. A good memory design assures that the combined effects of all fields not intended to result in switching i.e.:

1) Unselected bit fields

2) NDRO word fields

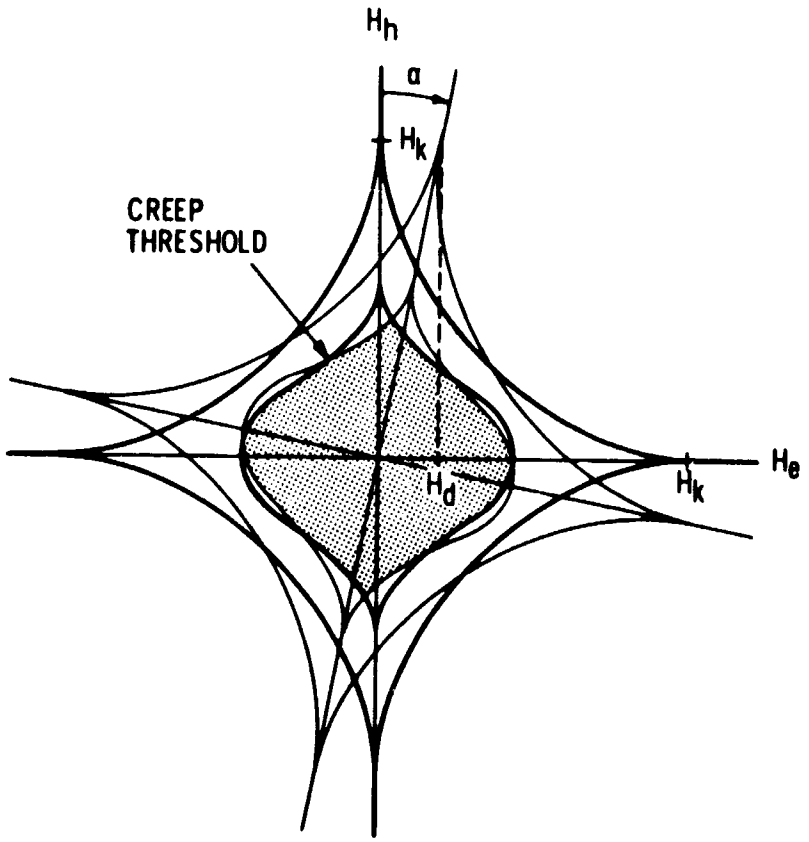


Figure 2.3.3 Effect of Creep and Skew

3) Stray dynamic and static disturbance fields

are contained within the creep threshold.

2.3.3.2 Skew and Dispersion

The easy axes of an array of elements are only nominally parallel to the geometrical axis, and to each other. The degree of average misalignment of the array is called skew. Figure 2.3.3 illustrates the effect of a skew of α on the switching boundaries. Since a component of the word field now appears along the element's easy axis, the bit field to switch the element in the presence of skew must exceed:

$$H_d > H_k \sin \alpha$$

implying increased bit currents, with consequent increase in driver dissipation and disturbance fields.

Dispersion is the variation of the easy axis skew within a discrete element, or over the surface of a continuous sheet. It is usually represented by $\alpha 90$, which is defined as the maximum dispersion angle of the individual magnetization contributions from 90% of the area under consideration. Both effects can be considered additive and are due to the polycrystalline nature of the magnetic film and the failure of all the crystals to line up perfectly. Their combined effect is to reduce the area within the effective creep threshold (see shaded portion of Figure 2.3.3) and to reduce the stray field margin. The effects of skew, dispersion, the demagnetizing field and other stray fields combine to set an upper limit to the NDRO word field. This has been discussed at length in Ref. [3]. Briefly, there is a limit to the closeness with which the remanent magnetization may be aligned to the hard-axis during a read operation, beyond which the above effects can combine to prevent the magnetization from relaxing to its previous position.

2.3.3.3 Magnetostriction

Magnetostriction has been quoted as the most serious cause of low yield of planar film arrays [2]. Although the deposited film is closely controlled to yield a zero-magnetostrictive composition, approximately 80% nickel and 20% iron, small deviations from ZMC are very sensitive to mechanical strain. The effect is a large increase in the skew angle with consequent decrease in the disturb threshold. Magnetostriction presents problems for substrate mounting, temperature equalization, and protection against acceleration, shock and vibration.

2.3.3.4 Disturbance Fields

A thin film memory element is subject to a number of disturbing fields:

1) The demagnetization field

2) Stray fields due to skew, dispersion

3) Fields from neighboring elements

4) Unaddressed location fields

5) Fringe effects from currents in adjacent conductors

6) Fields from pulses in word, bit lines induced by adjacent conductors

7) Adjacent element switching fields

8) The earth's magnetic field

9) Eddy currents in ground planes

The summed effect of all these must not exceed the local disturb threshold, or unwanted switching will occur.

Attempts to improve the density and speed of magnetic film memories have always run into problems associated with one or more of these disturbance effects, and different techniques of dealing with them have been developed.

2.3.4 Planar Thin Film Techniques

Thin film memory arrays differ principally in 1) the geometry of the thin film elements, 2) the geometry of the word and digit conductors, and 3) the choice of substrate.

The objective of the element design is usually to reduce the demagnetization field of the simple rectangular thin film segment without diminishing the output signal or increasing the disturb thresholds.

The shape of the magnetic element influences the demagnetization field. Oblong shapes with the easy axis in the longer dimension have lower demagnetization fields; a diamond shape with 3:1 ratio of axes has been used. Domain nucleation centre effects are reduced by tapering the edges of the element. Thinner films have smaller demagnetization fields, but are more difficult to control in manufacture, and yield a smaller output

signal when switched. For continuous films sufficient element separation must be assured so that only part of the element area is rotationally switched, thus avoiding problems associated with edge anisotropy and nucleation which would interfere with adjacent locations. The influence of stray fields can be decreased by increasing the disturb threshold of the elements, and by reducing the magnitude of the fields.

The composition of the magnetic film may be altered by the addition of ternary cobalt, copper or manganese [5]. H_K and the coercivity H_C can be increased and α 90 decreased by an addition of up to 15% cobalt. A higher coercivity provides an increased resistance to unwanted switching, but it also implies higher drive currents, which in turn will aggravate the stray field problem. The inverse approach of lowering H_K and the switching threshold, so that lower drive currents can be used, has also been considered. Stray fields can be reduced by:

1) Increased separations

2) Lowered drive currents

3) More sophisticated conductor geometry

4) Use of permeable keepers.

Arranging current conductors in complementary pairs, or creating multiple turns for a single conductor can help to concentrate the local magnetic field and attenuate the spread. The former has been used in a thin film array which is described later, and the latter is ideally suited to the plated wire film technique. Both naturally lead to a lower element density, but also to reduced drive currents.

Ferrite or permalloy keepers have been used to surround the elements and the conductors. Their effect locally is to reduce the demagnetizing field and to reduce drive current requirements. Between elements the keeper's low reluctance absorbs the stray fields and inhibits their interaction with adjacent locations. A continuous keeper will also reduce eddy currents in conducting substrates or ground planes, which create disturbing fields and incidentally prolong delays in drive current pulses. Ferrite has the advantage that it is electrically non-conducting, but its permeability decreases with frequency and it loses its effect for sharp current pulses. Permalloy has a wide frequency range, but is effective only for closely spaced, wide conductors.

Several different implementational approaches will now be described.

2.3.4.1 Split Film

This arrangement pairs two identical 15 mil x 45 mil film segments on either side of a set of rectangular intersecting word and digit lines. Segments are spaced 20 mil in the word and 50 mil in the sense-digit direction. The assembly is constructed of several substrates. Figure 2.3.4 illustrates the geometry. The magnetization field of the two segments, which are spaced 5 mils, reinforce each other and their distribution is nearer that of a closed flux linkage. Less dependence is placed on element shape and thickness. Burroughs has used this arrangement in the 8.5×10^5 bit film memory for the B8500 computer and the 1.3×10^5 bit memory for the ILLIAC IV. They claim that demagnetizing fields are reduced 50%. Bit density is 800 bits/sq.in. Manufacture is reasonably straightforward.

2.3.4.2 Mated Film

UNIVAC employs multiple depositions on a single substrate. The magnetic material surrounds the digit line (Figure 2.3.5). The complementary word lines, two per element, are orthogonal to the array and pass through holes in successive planes, in the so-called "Solid Stack" assembly. Magnetic keepers are deposited as part of the element to contain the word field. The complex shapes and alignment tolerances restrict the density to 200 bits/sq.in., somewhat defeating the purpose of the sophistication; but UNIVAC claims an increased yield for the process, and relaxed tolerances on creep and sense noise [2]. This technique is applied to the UNIVAC 1832 airborne multiprocessor memory, which is a 10^6 bit, 750 ns cycle time unit.

2.3.4.3 Coupled Film

IBM has developed a high density element by depositing a 1000 Å film on rectangular cross section, intersecting word and digit lines (see Figure 2.3.6) [4]. The word-line film provides flux closure for the sense-line film elements. Separations are 0.5 to 1 mil. Packing density is very high: 10^4 bits/sq.in., and word drive currents are low: 200 ma. The manufacturing problems are, however, fairly severe.

2.3.4.4 Post-and-film

This technique uses a continuous permalloy film evaporated onto a glass substrate. The elements are defined at the cross points of orthogonal word and sense lines. The unique feature is that the fields are localized by a continuous

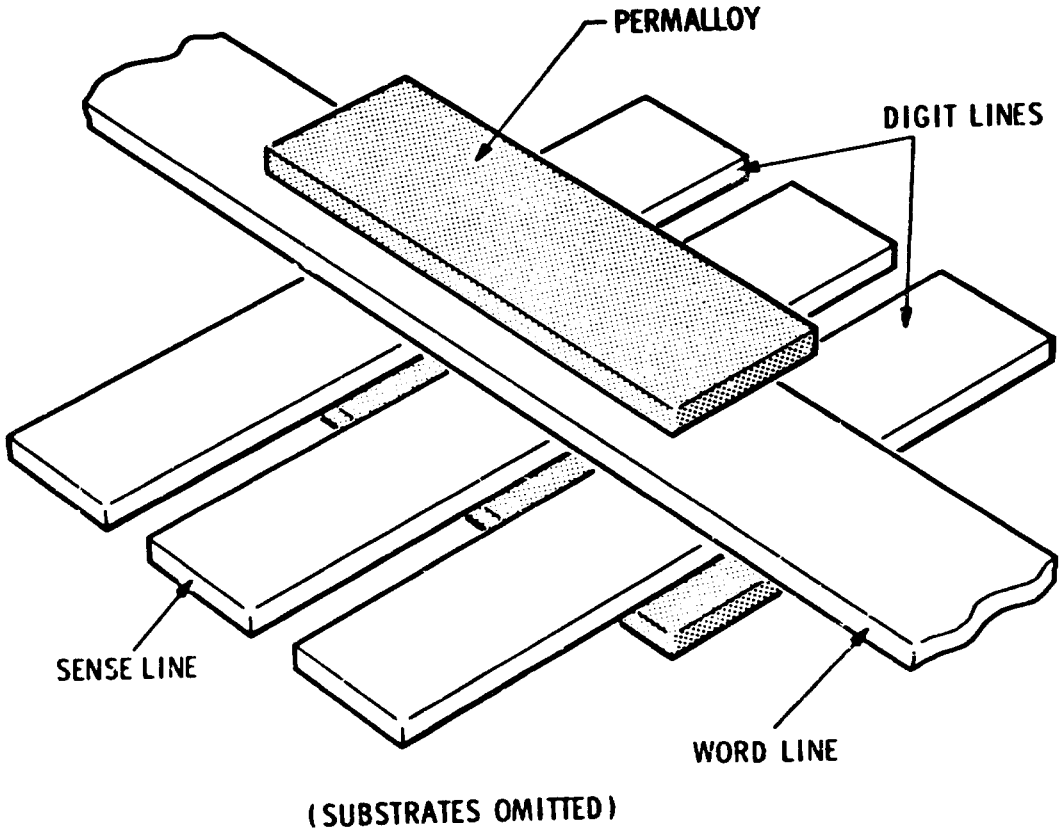


Figure 2.3.4 Split Film Memory Element

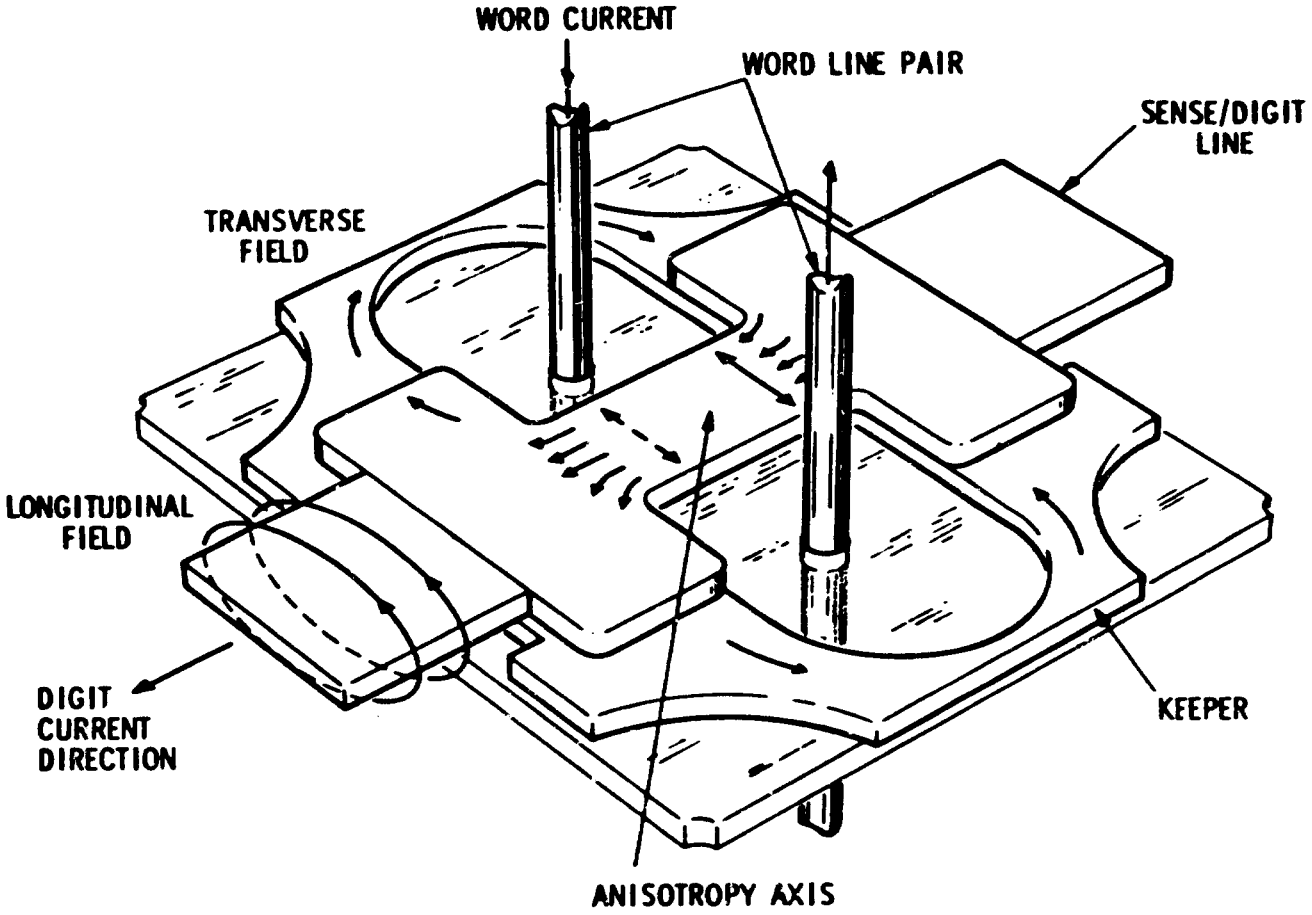


Figure 2.3.5 Mated Film® Memory Element

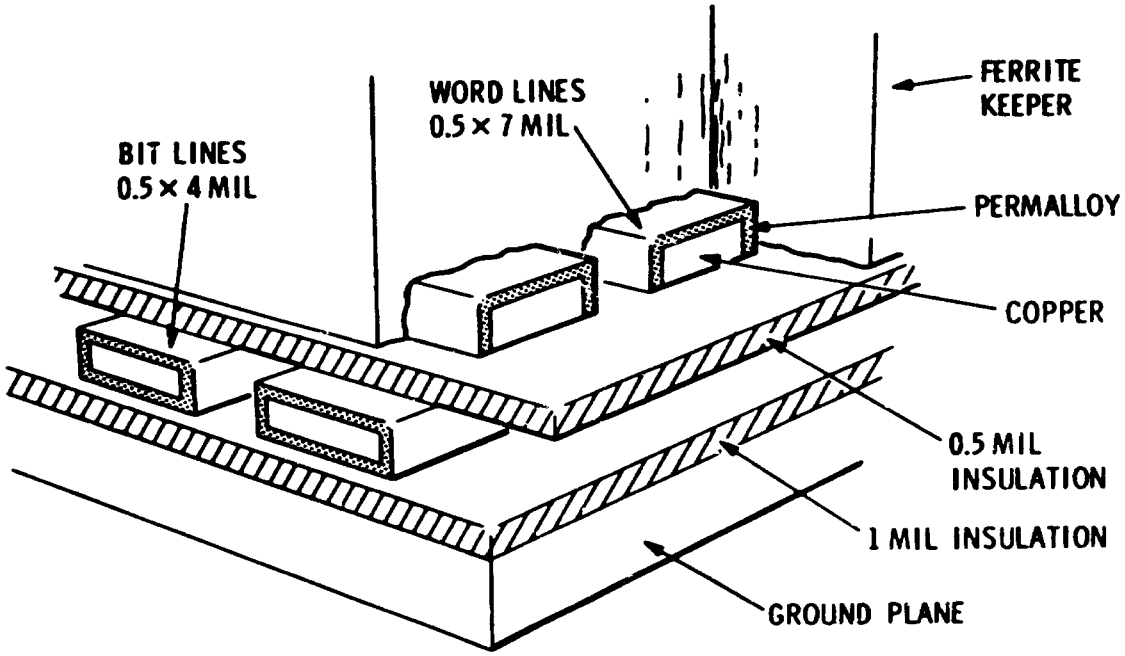


Figure 2.3.6 Coupled Film Memory Element

ferrite keeper plane which has been scribed to receive the conductors, leaving an array of mesas or "posts", hence the name (see Figure 2.3.7). Groove spacing is 15 mil, but two interconnections per bit are used, yielding a density of about 2000 bits/sq.in. No etching is done which, proponents claim, boosts the yield. Drive currents are 250 ma and 40 ma for word and digit lines respectively; a 1.3 x 10^5 bit memory dissipates 22 watts [7]. A double film version using soft and hard magnetic films can demonstrate NDRO performance without critical drive currents. Variations on this technique, including the "waffle-iron" memory, so-called because of the shape of the ferrite keeper, are being investigated by several manufacturers in the U.S., Europe and Japan. It offers some manufacturing advantages, being batch oriented with few processing steps, but achieves no breakthrough in packing density.

2.3.5 Thin Film Memory Application

Although thin film memory development has been pursued for nearly 15 years, adequate manufacturing techniques have not been devised to produce yields which would make this technology commercially attractive. Apart from one or two outstanding exceptions, use of thin films in computer memories has been restricted to laboratory research vehicles, or to special applications requiring high speed or limited capacity, e.g. scratch pad stores and airborne military computer memories.

Thin film techniques are still being very actively investigated in industrial laboratories, although emphasis is now on plated wire research. Some planar memories are still being constructed, but not usually greater than 10^5 to 10^6 bits. The inherent high speed of the switching mechanism still provides an incentive for continuing effort. A 60 ns cycle time, NDRO, 1.5 x 10^5 bit thin film memory has been described [6]. This sort of performance is still a challenge to semi-conductors. But activity in the mass storage field above 10^6 bits has been slight. The high per-bit costs, and the problems of integrating large numbers of 2-D arrays with inherent low output and sensitivity to noise, have been the cause. It is significant that a high percentage of operational thin film memories operate in the DRO mode, despite the speed and electronics advantages inherent in NDRO, because DRO requires far less caution regarding thresholds. Of even greater significance to the possible future of planar film mass memories was the decision by Burroughs (which has been committed to film memories for their commercial computers for 10 years), to replace the ILLIAC IV thin film memory with a semiconductor version.

A list of some thin film memory systems that have been assembled is presented in Table 2.3.1.

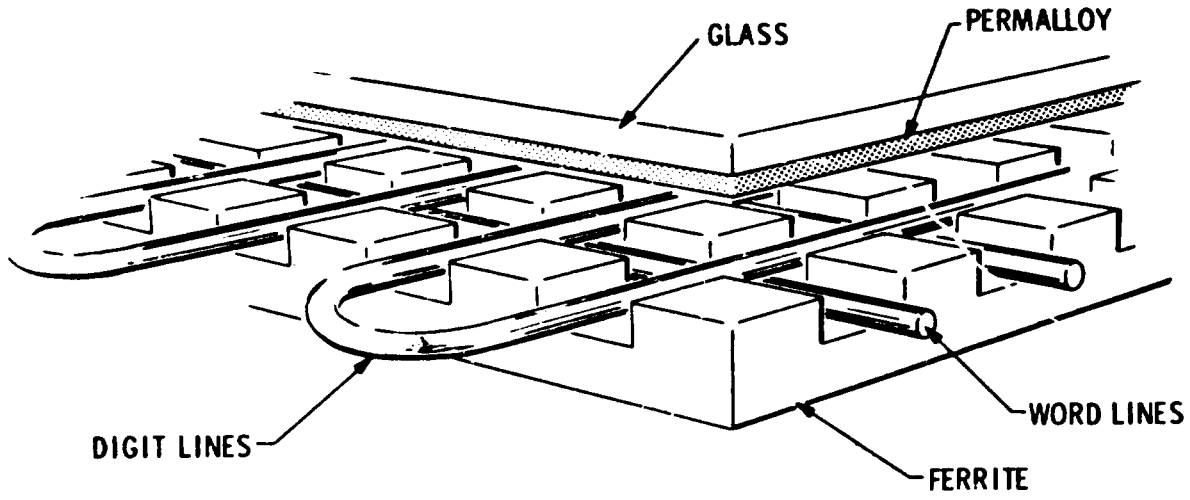


Figure 2.3.7 Post and Film Technique

	Capacity (bits)	Volume (cu.in.)	Weight (lbs)	Access/Cycle	Power (watts)
Burroughs B8500	8.5×10^5			250ns/500ns	845
Burroughs ILLIAC IV	1.3×10^5			120ns/240ns	
UNIVAC 1832	5.0×10^5	1100	49	225/500ns	190
MIT Lincoln	3.6×10^5				200

Table 2.3.1 Characteristics of Some Thin Film Memories

References for Section 2.3

1. Overn, W.M., "Controlling Creep and Skew in Thin-Film Memories", Electronics, September 15, 1968, pp. 125-128.

2. Overn, W.M., "Status of Planar Film Memory", IEEE Trans. Magnetics Mag-4, September 1968, pp. 308-312.

3. Logue, C.L., "Effects of Skew and Demagnetizing Fields on NDRO Performance", IEEE Trans. Magnetics MAG-5, March 1969, pp. 10-13.

4. Chang, H., "Coupled-Film Memory Elements", J. App. Physics, Vol. 38, No. 3, March 1967, pp. 1203-1204.

5. Freedman, J.F., "Soft Magnetic Thin Film Memory Materials", IEEE Trans. Magnetics MAG-5, December 1969, pp. 752-764.

6. Kohn G., et al, "A Very-High-Speed, Nondestructive Read Magnetic Film Memory", IBM Jour. of Res. and Develop., March 1967, pp. 162-168.

7. Veith, R.F., and Womack, C.P., "Post-and-Film Memory Delivers NDRO Capability....", Electronics, January 19, 1970, pp. 102-106.

2.4 Plated Wire

2.4.1 Introduction

The plated wire memory technique depends on the establishment of an anisotropic magnetic field in a thin, cylindrical film of permalloy, a nickel-iron alloy. It was first suggested as a computer memory element as long ago as 1958 [1], but has not risen to prominence as a competitor to the ferrite core until the last two years or so, due mainly to problems associated with wire production and array manufacture which have rendered the technique uneconomical. That these appear to have been overcome is evidenced by a surge of commercial activity in this field [2].

2.4.2 Description of Technique

The basic principle of operation is illustrated in Figure 2.4.1. The permalloy film is deposited electrolytically in a continuous layer of thickness 5,000-10,000 Å on a beryllium-copper wire of typically 5 mils diameter. Electrodeposition occurs in the presence of a magnetic field maintained by a current in the wire, so that an "easy" or preferred orientation of magnetization in the circumferential direction is established. Information may then be stored in the magnetic film according to the sense of the magnetization: a clockwise direction could represent a "1" as in Figure 2.4.1, and anti-clockwise a "0". A particular bit of information is located in the region of intersection of the orthogonal word "strap" and the continuous plated wire. This means that a plated wire memory is not physically discrete to the bit level as is for example, a core memory. The stored information is read out by applying a pulse of current along the word strap. The word field so produced deflects the stored magnetization away from its rest position along the easy axis. The resulting flux change induces a voltage in the plated wire, typically 5 mv, of sign dependent on whether a "1" or "0" is stored. Note that both states generate an explicit signal. The sense amplifiers are therefore required to cope with bipolar signals, which implies a higher complexity than presence, absence sensing. The magnetization will, for a properly chosen word current, relax to its original rest position when the word pulse terminates, thus affording a non-destructive read-out. The NDRO capability is a strong attraction of the plated wire technique over its big rival, the ferrite core, and results in an economy of drive and sense electronics. Information is written into the bit region by applying a sense line current pulse at the same time as the word pulse, so that the fields combine to steer the magnetization toward the desired easy axis direction. When the pulses terminate, usually with the bit pulse lagging

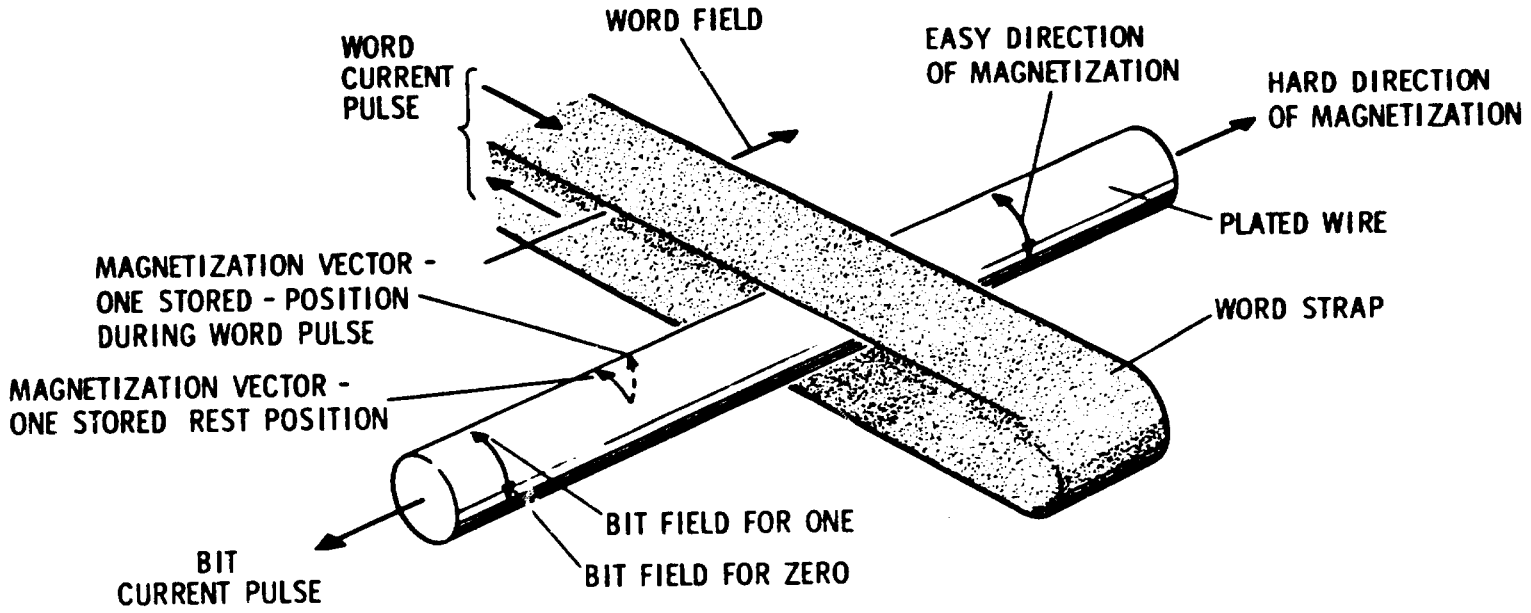


Figure 2.4.1 Plated Wire Memory Element

the word pulse by some 50 ns as illustrated in Figure 2.4.2, the magnetization will remain in the new rest position. Word currents are typically 300 ma to 850 ma, and bit currents are 30 ma to 50 ma. The closed flux path for the stored magnetization eliminates the usual problem of demagnetization of physically small linear magnetic elements, such as those employed in planar thin film techniques. It also ensures localization of the effects of the stored field, which is beneficial to the achievement of a high packing density.

2.4.3 Advantages of Plated Wire

The main advantages of the plated wire technique are [3]:

1) High speed at low power consumption, which is a consequence of the small amount of magnetic material in each bit, typically 2.5×10^{-8} cu.in. and the small amount of energy to be switched, on the order of 0.5×10^{-9} joules.

2) Good output signal, 5-10 mv, due to the very close coupling of the magnetic material and the sense line.

3) NDRO capability, which allows memory arrays to be organized with a minimum of driving and sense electronics without penalizing speed.

4) It is a non-volatile memory technique, implying a zero standby power requirement and permanent storage.

5) The storage medium is a polycrystalline structure, and its magnetic properties are essentially unaffected by radiation environments.

6) The plating material has a high Curie temperature (approx. 500°C), and a wide range of operating temperatures (drive current compensation is about -0.07/°C). However, plated wire suffers from accelerated aging of some of its characteristics at elevated temperatures, as will be described later.

2.4.4 Problems and Limitations

An information bit in a plated wire memory element is formed where the wire intersects the word strap. This results in a less precisely bounded physical location for the bit than in other memory technologies. The difficulties of defining and localizing the area of magnetic activity, by geometry and by controlling the properties of the materials, form the main drawbacks to the plated wire technique.

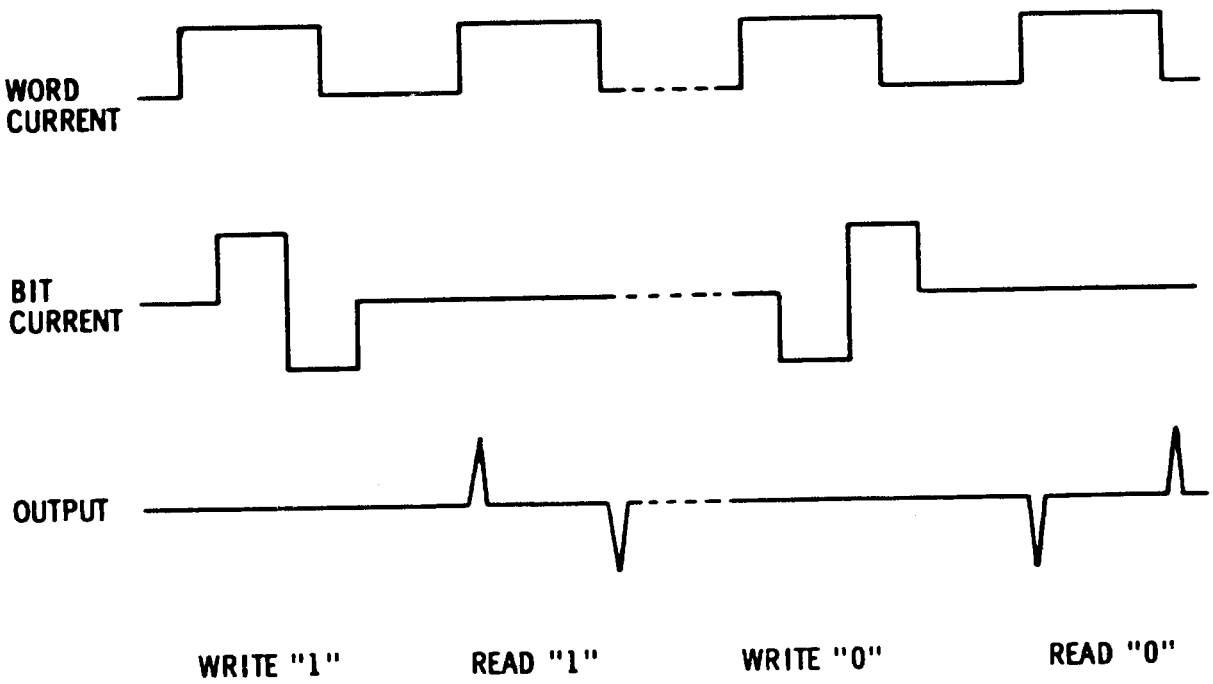


Figure 2.4.2 Write, Read Pulse Sequence

2.4.4.1 Size of a Bit

This depends on

1) the geometry of the driving conductors and

2) the magnetic properties of the film.

Calculations have shown [4] that for typical conductor sizes and geometries the word drive field at 50 mils from the center of the word strap is about 25% of its value at the center of the bit, and is almost independent of word strap width.

As for planar magnetic films, the demagnetizing field of the film spreads the applied word field on either side of the word line. The distance is given by:

$$\lambda = \lambda\sqrt{rt(M/H_k)}$$

where r = radius of the film

t = thickness of the film

M = magnetization

H_k = anisotropy field

i.e., it is independent of word line geometry. Wire is superior to planar film in this respect, since λ can be reduced by the use of smaller diameter wire, without impairing the output signal amplitude. This kind of analysis is only true for ideal film behavior, however, and in practical cases many factors affect the effective bit length.

In practice bit density is limited to between 20 and 25 bits to the linear inch, for a simple orthogonal array. The introduction of a magnetic medium around the word strap has been used in some plated wire memory designs to localize the word field and thus reduce the adjacent bit disturbance, allowing a closer bit packing to be achieved. A powdered nickel-zinc ferrite has been employed [5] to achieve a higher permeability (μ=6) between and around bit positions. The higher dielectric constant of the material (k=5) had a deleterious effect on the response times, but this was almost offset by the reduced dimensions made possible by the higher bit density. A secondary benefit of the magnetic keeper is the reduction in word current required to achieve a given change in the magnetization of the film, because the demagnetizing effect of the strong local field produced by the word current is absorbed by the keeper. This allows integrated circuit technology to be contemplated for the drive electronics. Power dissipation considerations limit IC currents to about 250 ma. In ref. [5] the word current

was reduced from 400 ma to 270 ma by introduction of the ferrite keeper; word line delay time increased from 3.2 ns to 5.4 ns; bit time delay from 1.8 to 2.0 ns. Packing density improved by 50%.

2.4.4.2 Thin Film Characteristics

1) The disturb threshhold of a plated wire memory element is a measure of the stability of the magnetization in the presence of disturbing fields from adjacent drive lines, eddy currents, static and dynamic fields from adjacent locations, misalignment of the easy axis, etc. The most serious manifestation of bit disturbance is creep, since it diminishes the distinction of the boundary between adjacent bits. Creep is caused by domain wall mobility under the influence of repeated word field application in the presence of a digit write field. Several tactics can be employed to minimize its effect:

   a) The domain wall motion coercive force can be maximized by controlling the surface roughness of the wire prior to the plating of the permalloy [6].

   b) The keeper described in Section 2.4.4.1 can be effective in maintaining the word field below the creep threshold at the adjacent bit location.

   c) The digit write current can be pulsed in a bipolar fashion, as shown in Figure 2.4.2. The mechanism for the effectiveness of this device is not yet fully understood [7]. The second, complementary pulse apparently reverses the domain wall motion due to the first pulse.

2) Skew and dispersion of the anisotropic axes affect the magnitude and the symmetry of the output, but are less serious effects than creep. Dispersion is typically 2°-5°.

3) Aging of the plated wire can change the characteristics of the memory. Tests have been conducted on 12,000 Å 80/20 Ni-Fe plated 5 mil wire under a controlled, elevated stress environment [8]. They show that aging results in

   a) Reduced output voltages

   b) An eventual inability to write, probably caused by an increase in the easy axis dispersion.

   The effects of aging can be minimized by post-deposition annealing in the presence of a magnetic field. The importance of minimizing the tendency to age is obvious in a long term

space vehicle application. Current indications are that plated wire life is at least 10-100 years, even at 50°C [9].

4) Zero magnetostriction is an essential requirement for the deposited permalloy, to avoid anomalous output voltages induced by manufacturing strains or environmental stresses. A composition of approximately 80% nickel, 20% iron exhibits this property; the achievement of this ratio is an objective of the plating process. An accuracy of ± 0.2% in the composition of the permalloy must be maintained in order to achieve manufacturing yields above 50% [9].

2.4.5 Manufacturing Techniques

The attractiveness of the plated wire memory technique from a commercial point of view stems from the possibility of a more continuous and batch-oriented, and therefore cheaper, manufacturing process than for ferrite cores. Wire manufacturing and testing has been under intensive investigation and development since about 1965.

2.4.5.1 Plated Wire Manufacture

Wire is usually a beryllium copper alloy which, although of higher resistivity than pure copper, exhibits an improved tensile strength. The most commonly used size is 5 mil diameter (36 gage), although recently the use of wire of 2 mil diameter has been developed by Honeywell for a military computer memory. The wire undergoes a continuous processing on its way to becoming a plated wire memory element. The sequence of electrocleaning (to reduce surface imperfections), copper plating (to achieve a controlled finish [6]), magnetic alloy plating, annealing, testing and cutting is illustrated in Figure 2.4.3. In order to optimize the various magnetic characteristics discussed in the previous sections close control over the process variables indicated in Figure 2.4.3 is necessary. Although there is apparent scope here for closed loop process control, in practice the relationship between the measurable, desirable wire properties and the controllable variables of the plating process is either not straightforward, or does not lend itself to continuous implementation. Some degree of automatic closed-loop feedback control has, however, been achieved for control of plating thickness and magnetostriction [9]. The following are the significant factors of manufactur

1) Plating.The thickness is a function of the plating current density, cathode efficiency, and time. The zero-magnetostrictive properties are sensitive to bath temperature, electrolyte flow rate, current density and time. The addition of cobalt or copper ions to the plating solution decreases

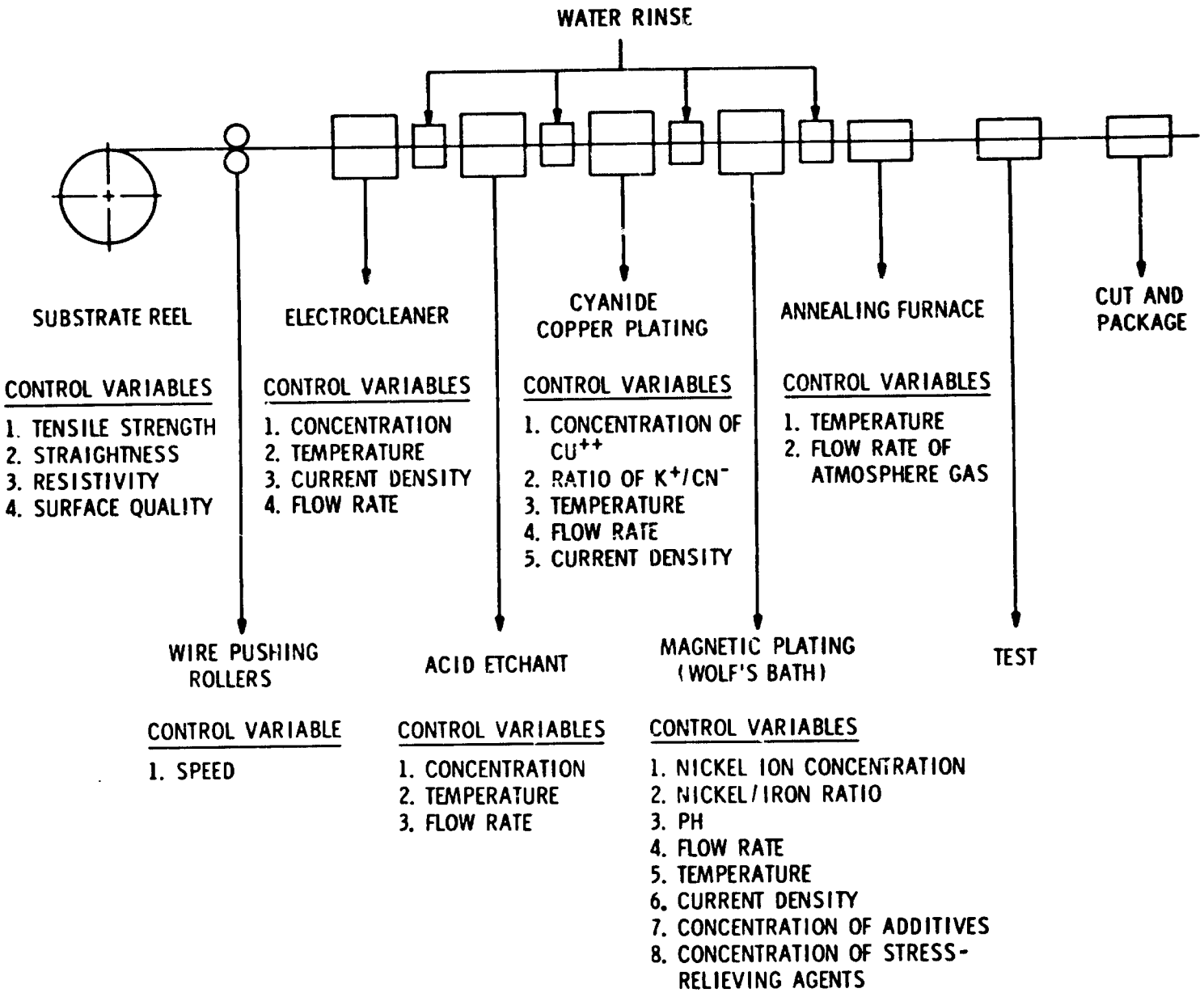


Figure 2.4.3 Plated Wire Manufacturing Process

the sensitivity of the nickel-iron ratio to temperature. Since the effect is also to increase the anisotropy magnetization (and consequently, the NDRO limit) for the plated wire memory, cobalt is often added to the nickel-iron alloy of plated wire for commercial memories. Cobalt has no apparent effect on the zero magnetostrictive property of the plating.

The anisotropy is produced by the circumferential field induced by a current flowing in the wire during the plating process. A field of up to 32 oersteds is sufficient to realize an easy-axis magnetization with a skew of less than .02 oersted.

In order to ensure that even the small variations in the Ni-Fe ratio that occur do not produce wire with permanent magnetostrictive properties the wire is usually pushed rather than pulled through the plating solution. The possibility of plating upon twisted wire, which will later untwist and develop skew, is thereby avoided.

2) Testing is perhaps the most critical part of the wire manufacturing process. The memory element is a continuous medium and must be checked microscopically. Some tests, for example, checking for correct B-H plot characteristics can only be done off-line because of their complexity and the time required; they are performed at intervals only as a check on the wire's uniformity. On-line tests are usually oriented toward worst case memory operation; they involve the cycling of individual bit elements through millions of typical read and write cycles in the presence of known disturbing fields, and then testing for loss of information and the inability to write correctly. 32,000 read and write cycles may be performed for every 4 mil of wire (i.e. every 1/10 bit) as the wire moves by at 30 inches/minute [10]. One method which claims to reduce the number of tests by three orders of magnitude makes worst case tests for the specific plated wire characteristics known to cause information loss, e.g. creep, skew, low disturb threshold, etc. [11]. The output of the test process is the activation of a rejection mechanism which isolates sections of wire in which faults have been detected.

3) Summary. Wire is manufactured in a continuous, closely-monitored process. Yields for plated wire lengths of about 12 inches are 35% - 60%. Longer lengths, up to 10 feet or so, can be made but the yield falls sharply. Long sense lines are usually segmented. The chemical and mechanical parameters of the plating process can be individually closed-loop controlled, but what is really needed to make the production of high yield, high quality wire commercially competitive over the ferrite core is the closing of the

loop from the determination of the wire's magnetic characteristics back to the plating variables.

2.4.5.2 Memory Array Assembly

Two main techniques have been employed to integrate plated-wire sense lines and word straps into an array. They differ in the arrangement of the word conductors, but both keep the plated wire lines as straight and stress-free as possible. Figure 2.4.4 illustrates one method, in which the sense lines are housed in thermally formed tunnels in a plastic or resin compound. The tunnels can be formed in two halves analogously to a phonograph record, or by withdrawing monel wire patterns, 1-3 mils larger than the plated wire, after thermo-molding the plastic substrate around them. The word lines are formed on the outer surfaces of the plastic by photo resist and etching. The critical factors are:

1) Tunnel pitch: a 3% change in pitch can cause a 4% change in operating current conditions across the array [9]

2) Tunnel parallelism and the alignment of the sense lines with their appropriate termination mountings: errors here can cause anomalous output voltages through local non-zero magnetostrictions. The word lines can be arranged on both sides by careful registration of front and back, to provide full turn word fields, or word currents can be returned via a common ground plane. In the latter case eddy currents contribute up to 40% of the field [9], but can cause additive word pulse trailing edge delays of up to several microseconds for ground plane thickness of 1-3 mil.

In the second method, shown in Figure 2.4.5, a word line mat is first woven around 5 to 8 mil diameter wires that are later withdrawn. The sense lines are then inserted into the mat. Permalloy keeper wires can be woven alternately with the word wires, which are usually driven in alternate directions to minimize the adjacent word disturbance. The principle advantage here is the ability to make multi-turn word lines, and to interconnect word lines in more complex patterns to reduce the disturbance field still further. Multiple turn word lines allow a marked reduction in the required word drive current, which must be achieved before integrated circuit technology can be implemented throughout the memory system.

Both these techniques are in commercial volume production.

In large memory planes it is necessary to provide spare plated wires at the edges of the array to ensure that all

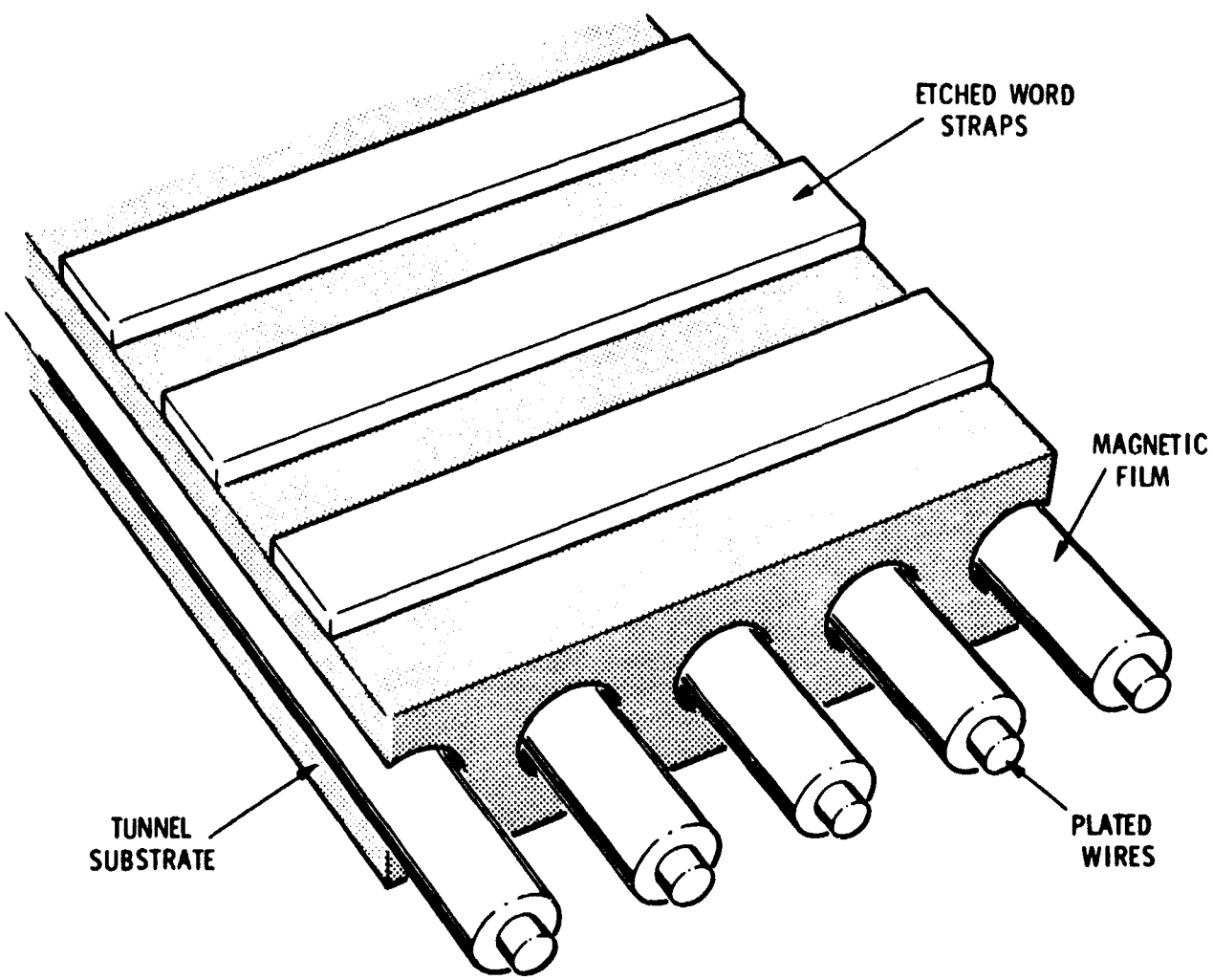


Figure 2.4.4 Plated Wire Tunnel Structure

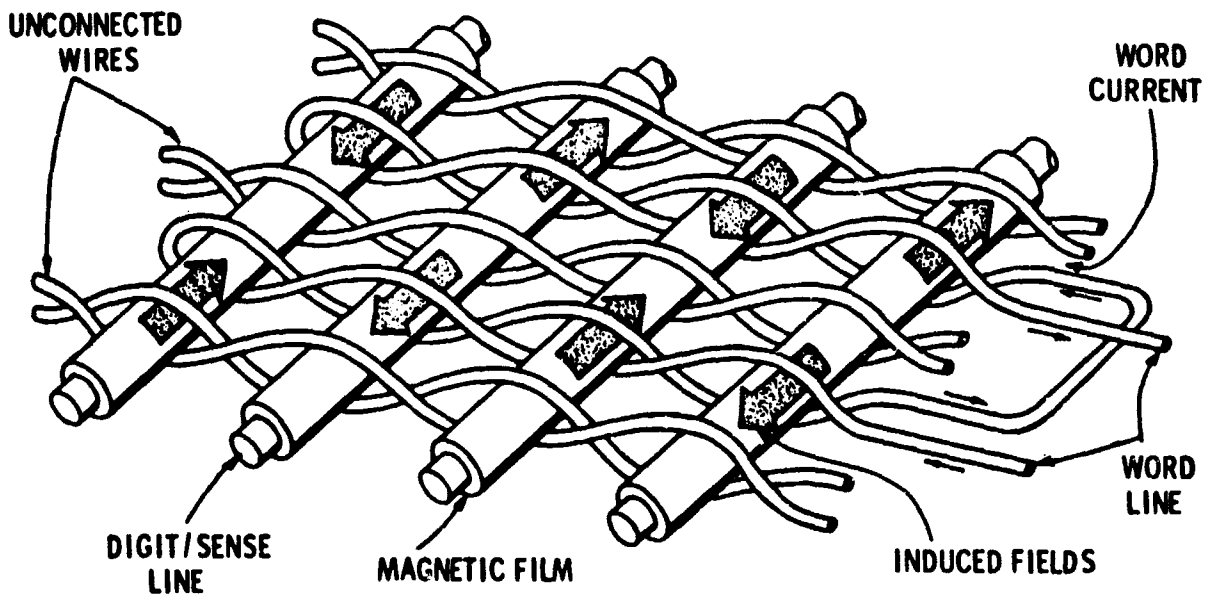


Figure 2.4.5 Plated Wire Woven Array

sense wires experience a uniform magnetic environment. Spare noise cancellation wires are often inserted at intervals in the sense wire array. Both practices tend to lower the overall bit density.

2.4.6 Plated-Wire Memory System Characteristics

2.4.6.1 Speed

The basic plated wire switching mechanism is attributable to the rotation of the magnetization vector and is therefore very fast, on the order of a few nanoseconds. Most of the total system delay accumulates, as with most matrix-organized systems, from the need to drive long lines with inherent CV^2 energy losses and from the delays associated with address decoding, drive and sense circuitry. The amount of magnetic energy released when a bit switches is low, 10^{-9} to 10^{-10} joules; but close coupling of the sense lines still maintains a good output signal, with a range of 2 to 5 millivolts. The net result is an overall speed advantage over ferrite core memories, by a factor of between 2 and 10, with access times from 75 ns to 1 μs depending on capacity and application.

2.4.6.2 Organization

The plated wire element is fundamentally a 2-wire device, so 2-D organization of arrays is commonly used. For memories larger than a few thousand bits, the plated sense wires become too long and the organization must be "squared up" with more equal numbers of bits and words on each side. The NDRO property of plated wire can be used to simplify the decoding of a large parallel output of several words that then results from the energization of a long word line. A simple decoding matrix suffices to select the desired word; the restoration of the unwanted data is not necessary since the information is not destroyed by the read-out process. In this way 2½-D-like operation can be achieved at less circuit complexity than for ferrite cores.

2.4.6.3 Data Rate

The data rates available from plated wire memories are very much a function of the organization, the degree of parallelism, and consequently the volume of electronics and the power available. Since the basic technology has speed, rates of over 10^8 bits/sec into large parallel interfaces of 50-100 bits are possible.

2.4.6.4 Density

Since bits cannot be spaced closer than 40-50 mils by today's standards, and sense line spacing is limited to 20-50 mils, the planar packing density at the bit level is limited to 500-1000 bits/sq.in. The volume density of the arrays does not exceed about 6000 bits/cu.in. Plated wire memory systems have overall bit densities that are generally less than 1000 bits/cu.in. The use of plated wire for very large spaceborne memories of 10^8 bits is therefore doubtful: a design study for a 10^8 bit plated wire memory by UNIVAC yielded a size estimate of 30 cu. ft. Owing to the need for rigid support of plated wire arrays, their physical density is on the order of 0.1 lb/cu.in., and a 10^8 memory might therefore weigh several thousand pounds.

2.4.6.5 Environmental Sensitivity

Plated wire is relatively insensitive to the effects of radiation, and memories can readily be constructed to be resistant to shock, acceleration and vibration. These properties make their use in military systems attractive. Compensation of drive currents is employed to combat temperature variations, but extremes of temperatures can accelerate the aging processes in the magnetic film.

2.4.6.6 Future Trends

Most research is concentrating on improving the bit density, lowering the necessary drive currents, increasing the speed of operation, and improving the manufacturing process for higher yield and lower costs. Current estimates by workers in the field [9] indicate that in 3-5 years we shall see densities of 50 bits/inch, word drive currents of 200 ma, and speeds of 100 ns for capacities of up to 0.6×10^6 bits. Costs will be less than 1¢ per bit. These figures are desirable from a commercial point of view and will serve to keep this field very active, but higher bit densities and improvements in size and weight are necessary if this technique is not to remain restricted to the small or medium sized space vehicle application of 10^5 to 10^6 bits.

An interesting variation on the constructional technique described in Section 2.4.5.2 has been reported by Japan's Electrotechnical Laboratory [12] which claims a performance that already exceeds the above predictions. A ferrite keeper substrate 5 cm x 7 cm is scribed with circular grooves spaced at 10 mils to accommodate 0.13 mm plated sense wires and 0.07 mm

copper word wires. In an orthogonal array word wires are placed in alternate grooves, and two crossings per bit are used. This yields a density of about 3000 bit/sq.in. Word current of 100 ma, and 100 ns access time are claimed.

2.4.6.7 Other Plated Wire Techniques

A non-anisotropic cylindrical magnetic film has been used in an axially switched mode to provide a high speed main frame memory element [13]. A 4000 Å film of 97%-3% iron-nickel is deposited on 10 mil beryllium copper wire, with a coercivity of 16 oersted. The wire is cut into 7" rods which are wound into word coils on a 5 mil pitch, co-axially with a bit sense coil on a 70 mil pitch. The technique claims that it is less dependent on the strict control of magnetic parameters during manufacturing, and that it is less sensitive to the anisotropic discrepancies of magnetostriction, skew, dispersion, etc. Access times of 300 ns for a 10^6 bit memory were claimed. However, its linear bit density at 10 bits/inch is less than half that of the other plated wire technique, and it has not been pursued by other manufacturers than NCR. It will not be discussed further.

2.4.7 Current Application of Plated Wire

Plated wire memories are specified for the mainframe of the UNIVAC 9300 series of computers. Those manufactured in this country employ the tunnel construction for the memory arrays and those manufactured under license in Japan incorporate memories which are supplied by Toko and utilize the woven technique pioneered by that company. Various small-to-medium aerospace and military computers are using the plated wire technique, for the reasons already stated. Brief statistics of some plated wire systems are presented in Table 2.4.1. The activity in the plated wire field is, however, indicative that many more commercial designs will incorporate this technique into computer systems in the near future.

Manufacturer	Capacity (bits)	Size (cu.in)	Weight (lbs)	Access/ Cycle	Power Watts & duty
UNIVAC	10^7	4.3×10^4			100
Honeywell (Minuteman)	5×10^5	460[1]		350ns/ 1.4μs	56 @ 30%
Honeywell (2 mil wire)	2×10^5	120	4.5	250ns/ 500ns	18
Librascope	8×10^4	40	4.25		0.25
Honeywell (Poseidon)	2.4×10^4	84[1]		300ns/ 1.00μs	10 @ 40%
Toko[2]	3.2×10^5	2.5×10^3	12	250ns/ 500ns	150 @ 100%

TABLE 2.4.1 Characteristics of Some Plated Wire Memory Systems

Note 1: with electronics

Note 2: DRO, non-military

References for Section 2.4

1. Gianola, U., "Nondestructive Memory Employing a Domain Oriented Steel Wire", J. Appl. Phys., Vol. 29, 1958, p. 849.

2. "The Year of Plated Wire", Electronic News, Fairchild Publications, January 26, 1970.

3. Fedde, G.A., "Plated-Wire Memories: UNIVAC's Bet to Replace Toroidal Ferrite Cores", Electronics, May 15, 1967.

4. Dove, D.B., "Demagnetizing Fields in Thin Magnetic Films", Bell Sys. Tech. J., September 1967, pp. 1527-1559.

5. Furuoya, T., et al, "Plated Wire Memory with Flexible Ferrite Keeper", IEEE Trans. Magnetics MAG-4, September 1968, pp. 375-378.

6. Richards, H.D., et al, "Topography Control of Plated-Wire Memory Elements", IEEE Trans. Magnetics MAG-4, September 1968, pp. 351-355.

7. Waaben, S., "High Speed Plated Wire Memory System", IEEE Trans. Elect. Comp. EC-16, June 1967, pp. 335-343.

8. Lutes, O.S., "Evaluation of Aging in Electrodeposited Permalloy Memory Wires", IEEE Trans. Magnetics MAG-4, September 1968, pp. 371-375.

9. Matthias, J.S. and Fedde, G.A., "Plated-Wire Technology: A Critical Review", IEEE Trans. Magnetics MAG-5, December 1969, pp. 728-751.

10. Chernow, G., "Boosting Plated-Wire Yield: Which Knob To Adjust", Electronics, September 1, 1969, pp. 95-96.

11. Lukianov, R., "Testing Continuously Plated Magnetic Wire", IEEE Trans. Magnetics MAG-4, September 1968, pp. 390-391.

12. ----------, "Plated-Wire Memory Packs in the Bits With Crossed Double-Decked Grooves", Electronics, December 22, 1969, pp. 167-168.

13. Meier, D., "Rods Look Like Wires, Act Like Cores", Electronics, November 11, 1968, pp. 128-131.

2.5 Block Oriented Magnetic Memory Techniques

2.5.1 Introduction

The term "block oriented" is associated with the organization of a memory rather than the technology employed, and we will discuss under this heading some of the varied techniques that have been employed for this type of magnetic memory device. In this memory organization the smallest addressable item of data is a block. A block is a group of bits, which may be organized into various bytes, words, etc. within the block, but are not individually accessible. The size of a block can be anywhere from 10 bits to 10^5. The reason for this organization is a desire to minimize the quantity of addressing required in really large memories. In operation it is analogous to rotating magnetic devices, such as discs and drums, in that information is written into and read out of blocks in a serial fashion. Access to a particular piece of data requires the whole block, or record, to be processed.

Block organized magnetic memories differ according to whether the data is moved in serial fashion through the magnetic medium and past the interrogation elements, as in a shift register, or whether the data remains relatively stationary within the medium and some agency, be it a magnetic or electric field, or a stress wave, moves through the medium, serially interrogating the data locations and communicating their contents to a read-out element. These two classes will now be described with examples.

2.5.2 Magnetic Shift Registers

The first class of BORAM* described above is really a magnetic shift register. The information exists in the form of magnetic domains which are caused to move through the medium under the influence of a changing magnetic field. The data bits are defined by domain walls separating regions of opposite magnetization. Shift register techniques differ in the physical form of the medium and the technique used to shift the domains.

2.5.2.1 Magnetic Wire

Domains of opposite axial magnetization can be formed in a uniform magnetic wire, and then moved along the wire under the influence of a moving axial magnetic field created by sets of dual conductors. The currents in adjacent pairs of conductors are phased to ± 90° with respect to each other, so that the domains can be moved backward or forward according to which field leads the other. Domains are created at one end of the wire by a write coil and detected at the other by read coil.

---

* Block Oriented Random Access Memory

Domain motion relies on the fact that in an anisotropic magnetic wire with a longitudinal easy axis the energy required to reverse the magnetization of a domain can be greater than that required to move it along the wire. Figure 2.5.1 illustrates the sequence of creation of a domain and the four phases on the conductors during one cycle of the driving current.

This method has been developed by Hughes for their Dynabit memory system. The solid magnetic wire is wound on a 1" former which carries longitudinal copper conductors resembling the commutator on an electric motor (see Figure 2.5.2). In one version that has been produced [1] the former is 8" long, is wound at about 30 turns to the inch and contains about 25 bit domains per turn, each spaced by a guard band, yielding a total of over 8,000 bits in a volume of about 6 cubic inches. The density of this device is very low, therefore: about 1300 bits/cu. in. The power consumption for a readout rate of 125 kilobits/sec is 8 watts per element of 8,000 bits. Currents of 2 amps are needed in the conductor, with rise times of 250 nanosecs. Output signals are typically 5 mv. More advanced versions are planned which will exhibit a x10 increase in bit capacity and density,with power requirements increasing by a factor of four. Readout rate is likely to be limited to less than 1 megabit/sec, because of limitations in domain wall propagation speed and the linear dimensions of a bit.

2.5.2.2 Magnetic Film

Other magnetic domain shift registers have employed parallel strips of magnetically soft film bordered by magnetically hard film to constrain the motion of the domains in the soft material. A multiphase magnetic field provides the driving force to the domains, whose direction of motion is predetermined by the geometrical design of the channels. Domains prefer to grow into larger dimensions of lower energy, and shaped "necks" in the channel rely on this to steer the domains towards the read head. A commercial product using this technical incorporates 1024 bit shift registers consuming 6w at a bit rate of 200 KHz, and will be announced sometime in 1970.

2.5.2.3 Metallic Shift Register Summary

An inherent feature of the propagation of magnetic domains is that it is essentially a loss-less process. The energy of creation remains with the domain, and regeneration is not required at intervals within a block of domains. Losses are only due to ohmic losses in conductors and eddy currents induced into support structures. Although this gives these techniques an advantage over magnetostrictive and acoustic devices,

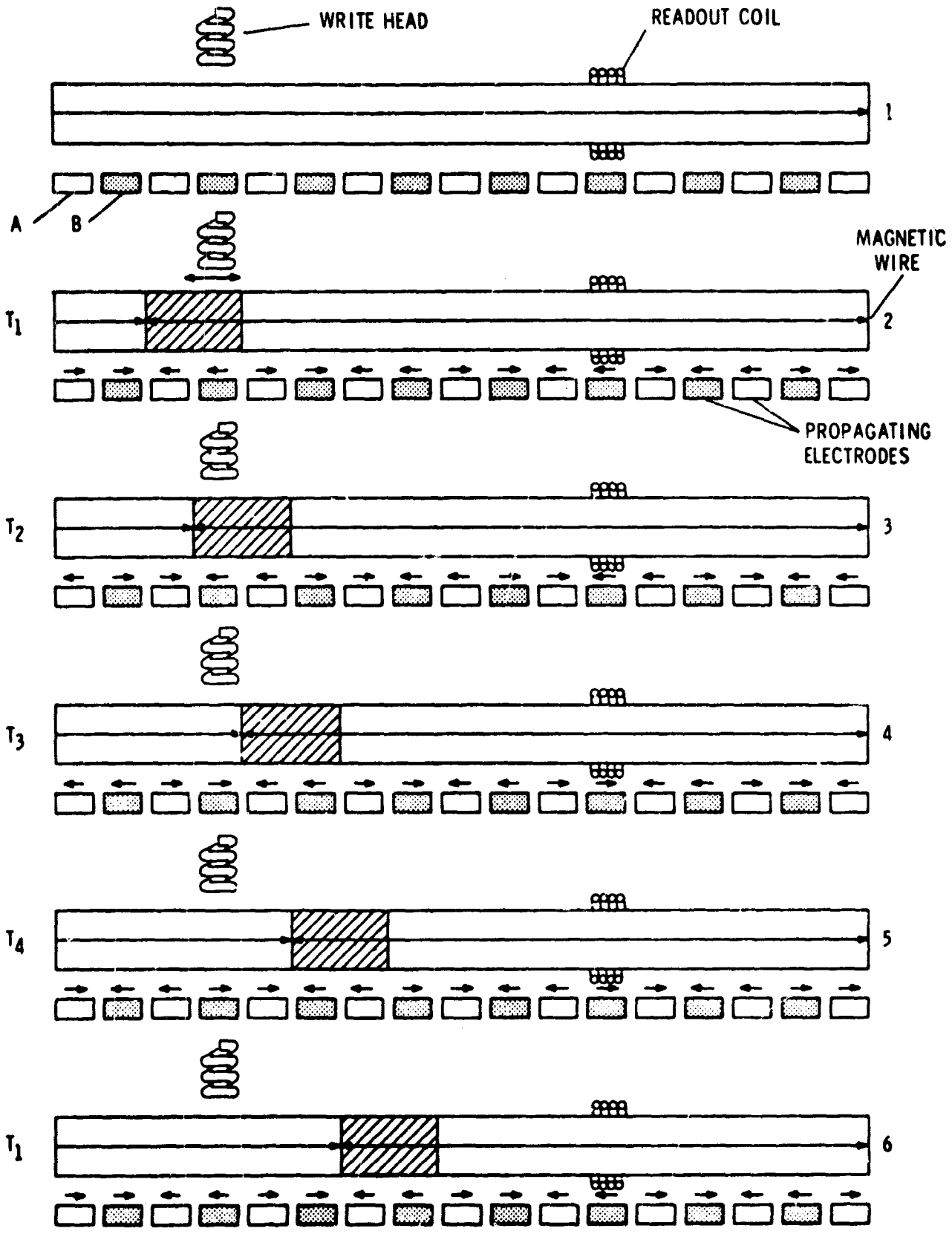


Figure 2.5.1 Propagation of Domain in Magnetic Wire

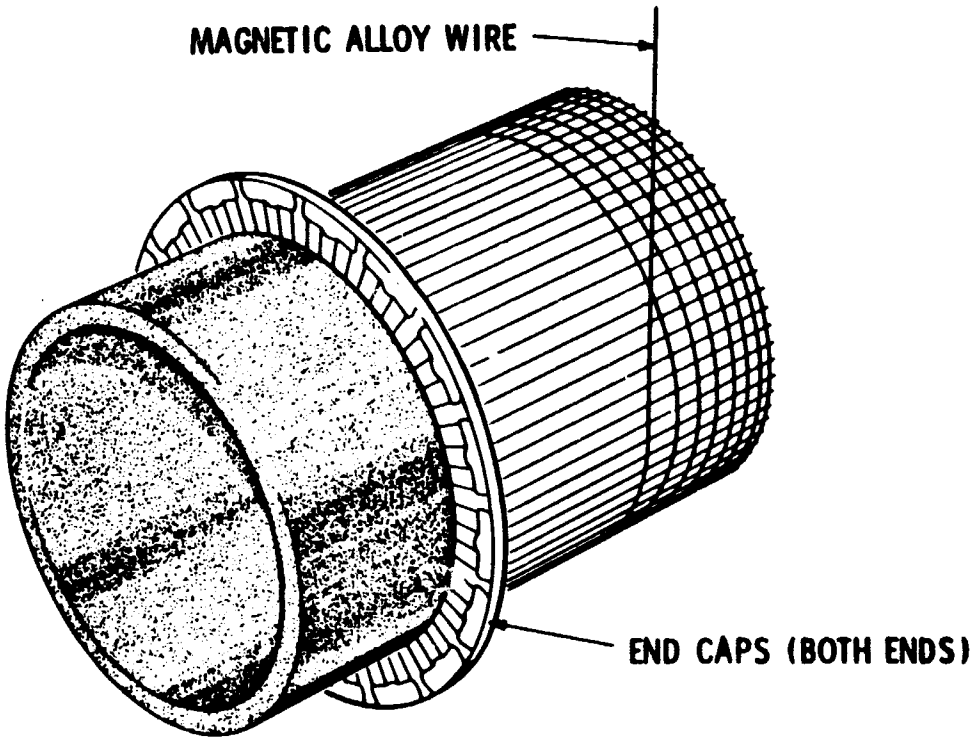


Figure 2.5.2 Construction of Dynabit ® Element

in general they suffer from poor bit densities of less than 1000 bits/sq. in., poor output signals, high driving currents with consequent noise and power dissipation problems, and the need for close magnetic material control to yield adequate ratios of nucleation to driving field.

These techniques have the advantage of allowing stop, start operation. A memory system which incorporates a degree of "look ahead" can cycle a block to position the correct data under the read head and then hold it until the system is ready for it. This serves to reduce the latency time, or average wait, for the desired data which is the usual penalty paid by adopting a BORAM organization. In most shift registers retention is lost when the data ceases to move. Also they allow NDRO, and are non-volatile.

2.5.2.4 Non-metallic Magnetic Domain Shift Registers

The study of cylindrical domains in sheets of orthoferrites has been reported for some time [2]. These domains have certain attractive features that give them an advantage over those in metallic materials: bit densities of 10^4 to 10^5 per sq. in. are possible; the domains are of constant stable size determined for a given material by a fixed bias field; they are able to move in any direction in the plane of the material; and they can be readily detected by optical, electrical or magnetic means. Shift registers employing this kind of domain have been experimentally established.

The magnetic "bubble" memory makes use of a physical property of certain materials that a thin plate or film of the material will support small cylindrical magnetic domains and allow these domains to be moved about in the material under the action of externally applied control fields. The presence or absence of a domain at a particular location determines the "1" or "0" state of the variable assigned to that location.

The materials which exhibit this property include the orthoferrites, which have the general formula $RFeO_3$ (where R is any rare earth or yttrium [2]), and some of the garnets, particularly gadolinium iron garnet [3] and yttrium indium garnet [4]. Thin plates or films of these materials are manufactured in such a way that an easy axis of magnetization of the material is normal to the plate or film. This results in the direction of magnetization in the material being normal to the surface. At any point, there are then two possible magnetizations for the material corresponding to the two possible normal unit vectors.

The magnetization in these plates naturally tends to establish itself in domains, or areas having the same magnetization and separated from other domains having the opposite magnetization by domain walls or boundaries. The size and shape of these domains is a function of many factors, such as the wall energy, wall motion coercivity, saturation magnetization of the material, plate thickness, bias magnetic field, etc. and has been studied in considerable detail [2, 3, 5, 6]. The magnetic domains respond to two opposing forces. One is the magnetostatic energy, which tries to increase the domain's surface energy, while the other is the sum of the bias field energy and the domain wall energy which try to reduce the domain's volume and wall area. The domain tries to assume a shape which minimizes the net energy [7].

In the materials which are suitable for magnetic bubble memory applications, the domains succeed in taking on the shape of small circular cylinders of approximately uniform size when an external biasing magnetic field of an appropriate strength is applied. The diameter of these domains is dependent on the strength of the biasing field and the plate thickness. It will vary from approximately the thickness of the plate (when the plate thickness is properly matched to the properties of the material and the magnetic field is relatively high), to many times this size if the field is lower or the thickness of the plate does not match the properties of the material. Application of a field higher than that required to maintain the minimum size domain will cause the domains to collapse and vanish permanently. Applications of a field too low to maintain the largest circular shape will result in the domains enlarging into long, snakelike strips of arbitrary length.

One of the necessary properties of materials for bubble memory is that they have a low domain wall coercivity, so that the domains are easily moved by creating local variations in the bias field. The variations can be generated by passing current through conductors or by the action of permalloy elements deposited on the surface of plate. Shifting, replication of domains, and logic functions can all be accomplished by proper design of the domain control elements [5].

One of the schemes for shifting the domains by the use of wedge shaped permalloy elements is illustrated in Figure 2.5.3. This scheme makes use of the fact that, because of magnetostatic energy considerations, a domain can move easily off the pointed end of a narrow wedge of permalloy but will have difficulty moving off the blunt end [7]. The permalloy pattern consists of a series of wedges placed sharp end to blunt end along the path the domain is to take. The size of a domain is determined by the applied magnetic bias and in Figure 2.5.3a, a single domain is located between the first and second wedges.

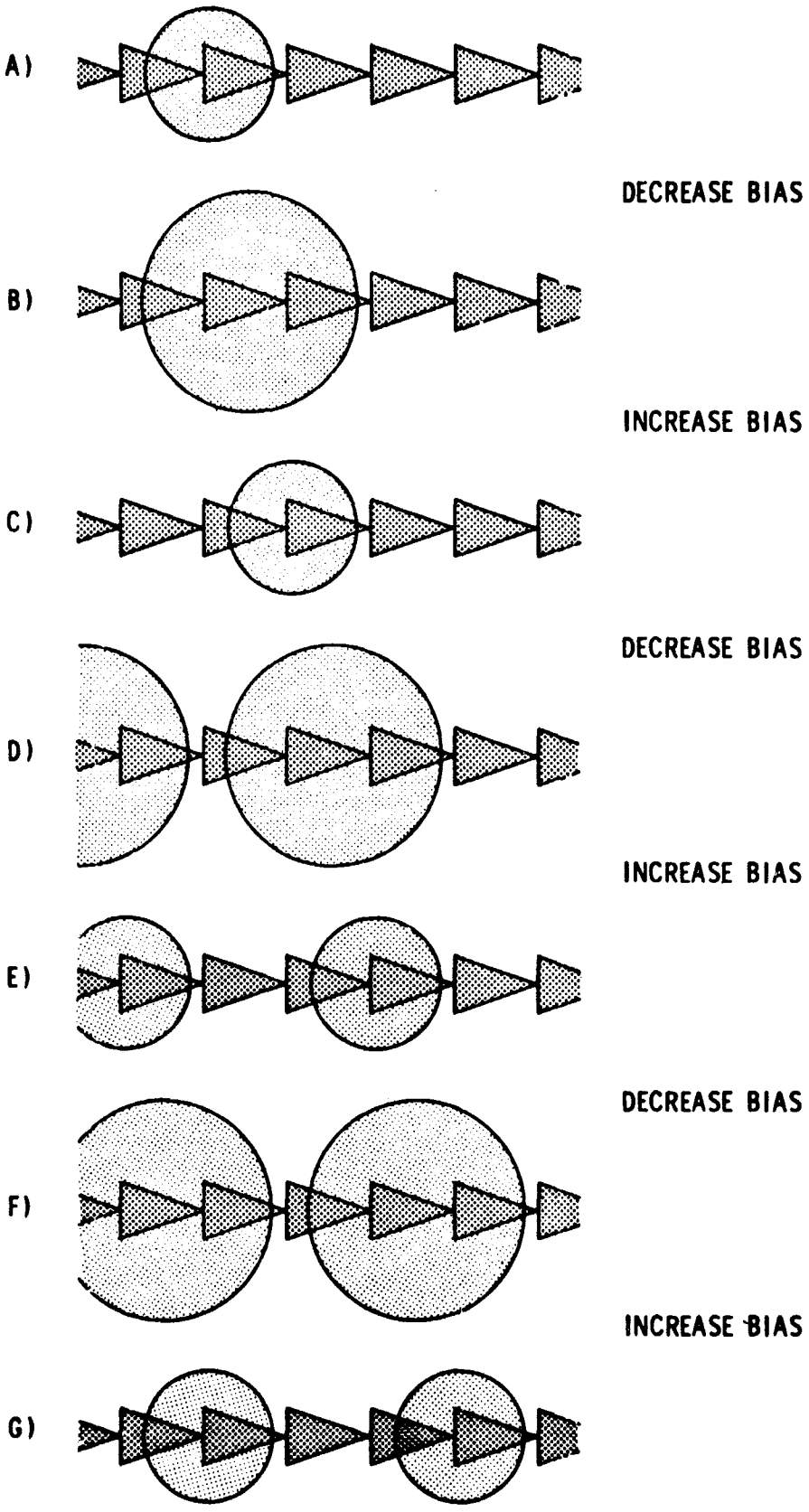


Figure 2.5.3 Propagation of Magnetic "Bubble" Domain

Propagation of the domain is accomplished by modulating the strength of the bias field. First the field strength is reduced, which causes the domain to grow until it straddles the first, second, and third wedges as shown in Figure 2.5.3b; the bias field is then increased again and the domain shrinks but as it does it pulls away from the first wedge and sits between the second and third wedges, as in Figure 2.5.3c. Another cycle of the bias field modulation moves the domain to between the third and fourth wedges, and a third cycle moves it to between the fourth and fifth wedges (Figure 2.5.3g). Also shown in Figures 2.5.3d through g is another domain entering from the left. This second domain is three wedges behind the first domain, and illustrates the point that domains can not follow one another by less than about three normal domain diameters. This restriction appears to hold for all of the shifting schemes.

Many of the materials that are useable for these memories are at least partially transparent and have the property that they will rotate the plane of polarization of polarized light as a function of their magnetic state (Faraday effect). It is therefore easy to observe the domains under polarized light, and when viewed in this manner the domains look like bubbles running around in a liquid, hence the name "bubble" memories.

The presence or absence of a domain in a particular location can be sensed by induced voltage, by the Hall effect, or by optical means. The induced voltage method makes use of a sense line and sense amplifier combination, and involves the intentional collapse of the bubble. The read out signal is quite small. Under carefully controlled conditions, an output of 1 millivolt per microsecond has been reported [7]. Both the Hall effect and the optical readout require special sensing attachments that are more complicated than a simple sense line at the location of the bubble, and these attachments add to the complexity of the memory.

The plates and films for bubble memories can be manufactured by a number of methods. For orthoferrite memories, Bell Telephone Labs uses a bulk growing technique. The necessary rare earths and a ferrite are placed in a crucible and melted. The melt is slowly cooled back to room temperature and the result is a single crystal of orthoferrite. This crystal is then sliced and polished into plates a few mils thick [7]. Recently, Nippon Electric Co. in Japan has reported considerable success in growing orthoferrite crystals using a floating zone process, which they claim is much faster and gives better crystals than the Bell method [8]. Bell is also working with garnet memories. They have reported using the flux method to grow memory plates of $Gd_{2.3}Tb_{0.7}Fe_5O_{12}$ which have storage capabilities of 10^7 bits per in^2.

Autonetics uses a different method in their manufacture of garnet memories. They use chemical vapor deposition to create epitaxial thin films of yttrium indium garnet which are from 1 to 50 microns thick. Autonetics claims this to be a less expensive technique and that it produces a larger, more homogeneous magnetic medium than the Bell orthoferrites. Autonetics also claims that their garnet memories will be much less temperature sensitive than the orthoferrite memories [4].

The optimum size for a domain varies from a few mils in some of the orthoferrite [2] to a few microns for an yttrium indium garnet thin film on a gadolinium gallium garnet substrate [4]. It is therefore possible to conceive of memories with storage densities of 10^5 to 10^8 bits per square inch after allowing for 3 positions for each bit during the shifting operation. Shifting can take place at rates exceeding 3×10^6 shifts per second and the energy required to perform 10^{12} operations per second has been estimated at 0.040 watts [7].

Magnetic bubble circuits require many of the same manufacturing process - such as masking, thin film depositing, chip handling, etc. - as semiconductors, and a semiconductor device or integrated circuit manufacturer should be able to manufacture magnetic bubble devices with comparatively little difficulty.

The electronic overhead for a magnetic bubble memory will have to include a permanent magnet capable of maintaining a field intensity of the order of 20 to 40 Oersteds normal to the plane of the memory material. In addition, circuitry will be required to control the shifting operation, and detectors or sense amplifiers will be needed for the readout.

Since magnetic bubble technology is built around the shifting of domains along predetermined paths in two dimensions, it is not well adapted to random access memory operations. It is particularly adaptable to long shift register memories, however, or possibly to memories having a relatively small number of parallel shift registers. The contents of one or more of these shift registers, read in series or series/parallel would then be the storage block in a block organized mass memory. At the present time, such devices are just emerging from the concept stage to the laboratory model stage, and several more years of development will be required before magnetic bubble memories appear in working computers.

2.5.3 Scanned Block Oriented Magnetic Memories

In the second class of memory defined in Section 2.5.1 the data remains relatively stationary and is scanned by a moving

agency. Two examples will be described, one using an acoustic stress wave and the other a travelling domain wall to interrogate data locations.

2.5.3.1 Sonically Scanned Memory

This type of storage device depends on the anisotropic behavior of a magneto-strictive material under the influence of physical strain. Long strips of 10-20 mil wide magnetic film of about 1000 Å thickness are deposited on a fused silica substrate and overlayed with a pattern of conductors to sense the changes in magnetic flux (see Figure 2.5.4). An ultrasonic transducer is fixed to one end of the substrate and an absorbing medium is attached to the other to suppress reflections. The film material is of a strain sensitive composition, about 60% nickel, 25% iron and 15% cobalt. As a sonic pulse generated by the transducer passes along the substrate the resulting shear strain changes the local anisotropic behavior of the film as indicated in Figure 2.5.5. Coincident read and write pulses in the overlaid conductors sense or change the local remanent magnetization. Reading depends on the reversible rotation of the easy axis magnetization toward the direction of strain. This effect is greatest when the strain is at right angles to the easy axis (see Figure 2.5.6). Writing can be performed at strained locations because the switching threshold is then reduced.

The transducer is the principal source of difficulty. A 4 MHZ PZT transducer allows a square sonic pulse of low dispersion to be transmitted into fused quartz, but owing to the high velocity of 3000-4000 m/sec, and the finite switching time of the material, only about 25 bits per inch are possible. Higher frequency materials such as CdS enable 50 MHZ and more than 100 bits/inch to be realized, but they present processing difficulties. At the higher frequencies pulse dispersion occurs and this limits distinguishable bit separation to about 20 mil [9]. The film must be uniformly composed and deposited to avoid anomalous magnetostrictive effects. Its strain-sensitivity should yield a 2:1 change in threshold for a 10^{-5} strain [10]. The H_k should be < 15 Oe; "inverted" films in which the $H_k < H_c$ ensure rotational rather than the slower domain wall switching (see Section 2.3).

This type of memory is non-volatile, affords NDRO, and is capable of high data rates of up to a few hundred MHZ. The density is, however, fairly low, and is unlikely to exceed 10,000 bits/cu. in.

There has been considerable industry investigation of the acoustically scanned magnetic film technique, notably by RCA and Sylvania, but no complete memory has been offered commercially.

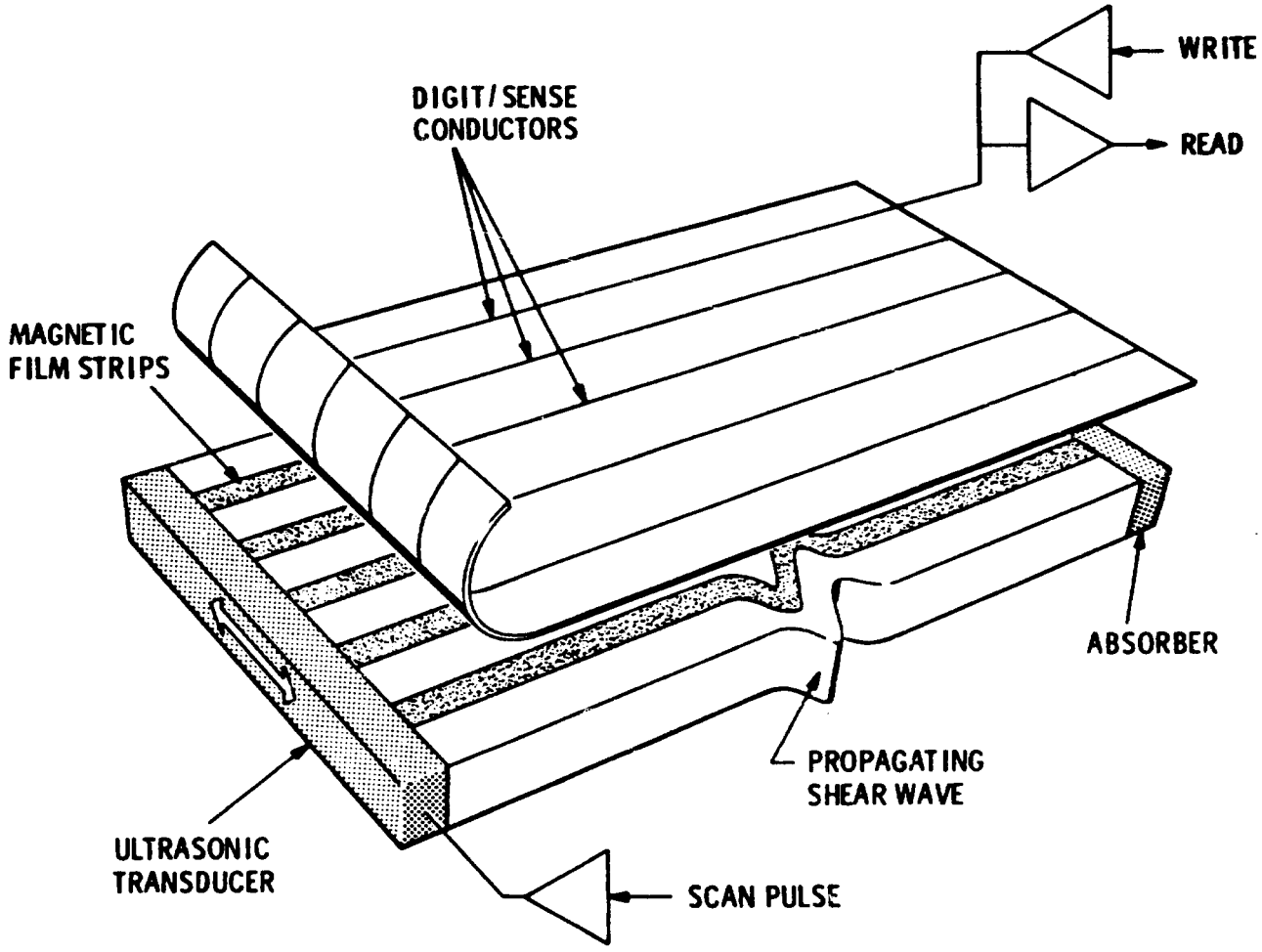


Figure 2.5.4 Sonically Scanned Film Memory Schematic

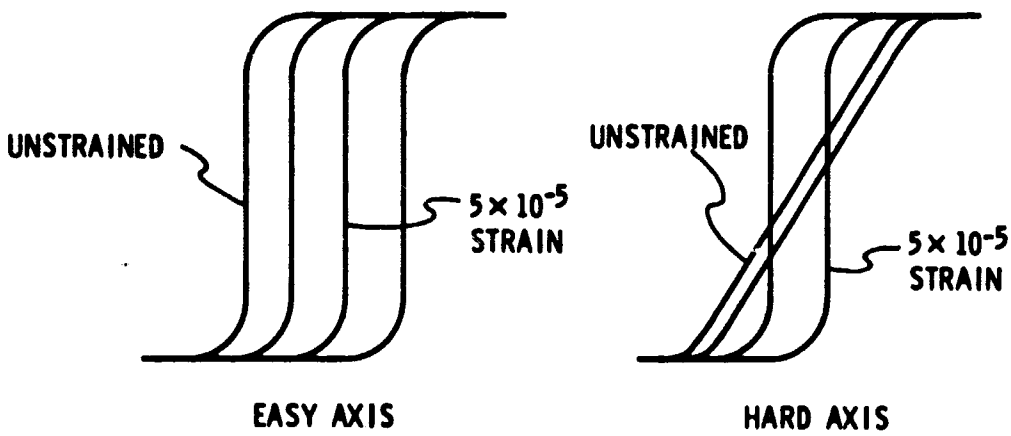


Figure 2.5.5 Hysteresis Loop Behavior
Under Strain

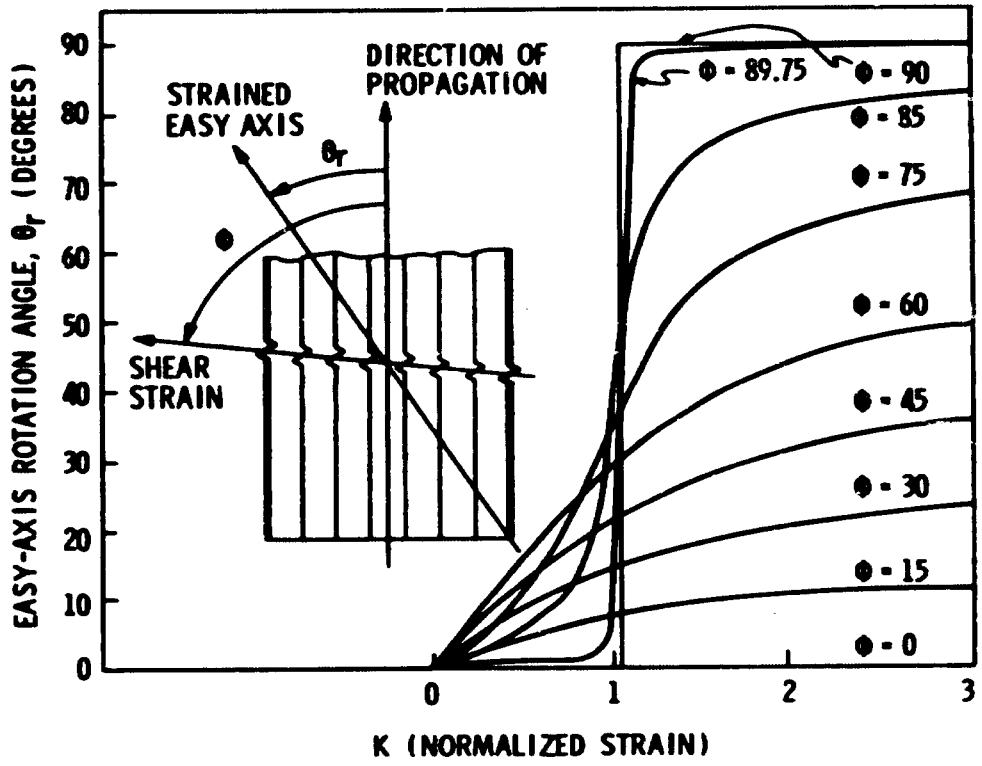


Figure 2.5.6 Rotation of Easy Axis Under Strain

2.5.3.2 Travelling Domain Wall Scanning

A straight domain wall structure travelling transversely in a thin magnetic film has been used by Burroughs [11] to scan the local magnetization in a second film. The first film is a zero magneto-strictive Ni-Co-Fe film of 1100 to 1300 Å, and the second is typically of the composition for planar magnetic film described in Section 2.3. Domain wall motion is induced by a local pulsed field; for speeds above 1000 to 1500 m/sec the domain wall is dynamically stable and about 0.1 mm wide. Speeds are proportional to the 2nd to 5th power of the driving field and can attain 10,000 m/s. An essentially unipolar scanning field is required for detection, and is induced by eddy current action in a conducting ground plane placed between the films. Bit densities of about 4000 bits/sq. in. have been attained at bit rates up to 10 MHZ per scanned track. This technique has had no reported operational application to date.

References for Section 2.5

1. "Dynabit (T) Storage Element for Mass Memory and Buffer Applications", Technical Data Sheet, Hughes Aircraft Company, Fullerton, California.

2. Bobeck, A.H., et al, "Application of Orthoferrites to Domain-Wall Devices", IEEE Trans. Magnetics, MAG-5, September 1969, pp. 544-553.

3. Gianola, U.F., et al, "Material Requirements for Circular Magnetic Domain Devices", IEEE Trans. Magnetics, MAG-5, September 1969, pp. 558-561.

4. ----------, "Breaking the Bubble", Electronics, June 8, 1970, pp. 39-40.

5. Perneski, A.J., "Propagation of Cylindrical Magnetic Domains in Orthoferrites", IEEE Trans. Magnetics, MAG-5, September 1969, pp. 554-557.

6. Thiele, A.A., "The Theory of Cylindrical Magnetic Domains", Bell System Tech. J., December 1969, pp. 3287-3335.

7. Karp, H.R., "Magnetic Bubbles - A Technology in the Making", Electronics, September 1, 1969, pp. 83-87.

8. ----------, "Floating Bubbles", Electronics, July 20, 1970, p. 150.

9. Shahbender, R., et al, "Sonic Film Memory", RCA Review, March 1970, pp. 39-59.

10. Onyshkevych, L.S., "Strain Sensitive Thin Magnetic Films", Journal Appl. Phys., 1 February 1968, pp. 1211-1213.

11. Shafer, P.E., "Dynamically Stable Transverse Domain-Wall Structures, and Application to a Serial Access Thin-Film Memory", IEEE Trans. Magnetics, MAG-5, September 1969, pp. 639-640.

2.6 Semiconductor Techniques

2.6.1 Introduction

The techniques that have been developed over the last ten years for the integration of large numbers of semiconducting devices on a monolithic substrate have recently been applied to the fabrication of memory arrays [1,2,3]. There are several good reasons for this development.

Firstly, semiconducting devices can exhibit attractive characteristics as memory elements. They are capable of high speed, high density, high level output, and low power consumption. The availability of two basically different semiconductor technologies provides flexibility and diversity that permits broad tradeoffs of speed,power, density, complexity, and cost to be made to suit a variety of applications. It is possible to combine the function of logic and storage, so that address decoding can be included on the same substrate as the memory cells and the concept of associative memories can be readily implemented. Information can be read out in a non-destructive manner which is not dependent on rates of change of current, unlike the inductive mechanisms of memory techniques based on magnetic phenomena. Most semiconductor storage devices contain active, amplifying elements, which together with the continuous output capability, result in a good inherent signal-to-noise ratio.

Secondly, memory technology is based on manufacturing processes that have had many years of development and use, and whose products have achieved high degrees of reliability and economy. Integrated circuit techniques promise the attainment of batch fabrication techniques, which will ultimately yield low costs that are not as dependent on "per bit" factors as discrete techniques.

Thirdly, almost without exception, the drive, sense, control, and interface logic of any memory device is implemented with semiconductors, as is the system of which the memory is a part. It is obviously of great benefit if the memory technology presents to the system designer and implementer problems of design, integration, and operation that have been faced and solved in the rest of the computer system,and which do not require a separate, perhaps totally new, expertise.

Several factors have until now served to keep semiconductor memories from being cost competitive in comparison to other technologies, and there also are some technical limitations. In order to understand these factors, the steps being taken to eliminate them, and to make estimates of future trends, an outline of semiconductor memory techniques is given in the

next section, followed by sections giving detailed descriptions of the more promising technologies, the practical factors of implementation, and indications of future trends.

2.6.2 Semiconductor Memory Techniques

2.6.2.1 Basic Semiconductor Technologies

Two major integrated circuit technologies have been developed: one based upon the bipolar junction transistor, and the other on the metal-oxide-semiconductor (MOS) field effect transistor. Each possesses unique properties that have been exploited for different applications. The bipolar transistor is a low impedance, current control device capable of switching speed down to the sub-nanosecond region. Circuits incorporating bipolar IC's are used for high speed logic operations and in high speed control or scratchpad memories of a few hundred bits, where the characteristic larger device area and higher power consumption are of less consequence than short switching times. The MOS transistor depends on the movement of surface charges in silicon and is characterized as a high impedance, voltage controlled, bi-directional switch. It is smaller by a factor of 4 or 5 than a bipolar IC transistor, dissipates less power, requires a half to a third as many processing steps, and is considerably cheaper per bit; but it switches more slowly because of lower charge mobility, and its high impedance and high voltage requirement (up to 30V) make MOS circuits speeds very vulnerable to stray capacitance, on and off the chip.

MOS technology is being actively developed because of its high density and simple fabrication. The problems of high threshold voltage (which makes interfacing to bipolar circuitry difficult), and low speed are being tackled with some success. At the same time bipolar memory developments are yielding devices of smaller size and lower dissipation. Thus the two technologies will probably coexist for some time to come. However, from current trends it seems unlikely that bipolar techniques will be considered for memory systems in excess of a few hundred thousand bits, but it is probable that MOS techniques will penetrate the mass storage regions around 10^8 bits.

2.6.2.2 Semiconductor Storage Element

Except for read-only memories and the new MOS techniques, the memory storage element is composed of a number of individual devices, i.e., transistors, diodes, capacitors, and resistors.

The most commonly used circuit element is the back-coupled flip-flop, which usually consists of six or more transistors. Bipolar read and write memories employ the flip-flop exclusively, each typically occupying a 50-100 sq. mils cell in the memory array. A different principle is possible with the MOS transistor, because its inherent high impendance allows one to store information as a charge on the gate capacitance. Input and output gating can be accomplished with only three or four transistors, which are smaller than their bipolar equivalents, so that an array cell occupies 10-15 sq. mils.

The MNOS device exhibits storage by trapping charge at the interface of two insulators of differing dielectric constant that form the gate dielectric of a transistor. A single device serves to store each bit, and cccupies less than 1 sq. mil. This technique is described in some detail in a later section.

One major drawback of semiconductor memory elements relying on voltage feedback and/or charge regeneration is that the information is volatile and evaporates when the power supplies drop below a minimum level. Information loss may be prevented by a self-contained power supply, i.e., battery or storage cell, but this is a theoretical rather than practical solution. The one exception is the MNOS charge storage mechanism, which has shown retention times up to two years, and it is this that makes it exceptionally promising for future memory applications.

The read-only memory employs a single device for each bit, usually a diode or transistor, which is selectively unconnected during manufacture according to the desired pattern of ones and zeros, by custom masking or by laser vaporization after metallization. Densities for ROMs are therefore higher than the equivalent read and write store by a factor of about four, depending on capacity and the amount of on-chip decoding [7].

2.6.2.3 Large Scale Integration

Large scale integration is the concept that makes the notion of the semiconductor memory array feasible. It is the placing of a large number of individual devices, currently up to 5,000, on a single silicon substrate. The largest area of silicon that is being worked today as a monolithic substrate is about 30,000 sq. mils. Both the decreasing yield and the increasing difficulty of testing more dense and complex structures currently make the production of larger chips uneconomical. Commercially available LSI memory chips today are typically 100-150 mils square and may contain up to 1024 bits in a 1 microsecond random access MOS dynamic circuit organization [5]; bi-

polar LSI is available in chips containing up to 256 bits with 70 ns access time [6]. The reasons for the variation in density between 10^4 and 10^5 bits/sq. in., speed, and technology will be discussed later. Much higher device densities and larger chip sizes are being achieved in experimental quantities.

One of the major problems in the application of LSI techniques to memories is the interconnection of the individual chips, which contain only 100-1000 bits, to form larger arrays. Various techniques (such as beam leads, "spider" bonds, solder bumps, etc.) are currently competing for acceptance as the method for bonding chips to intermediate substrates. These in turn must be mounted on cards, or stacked in some way, to become the modules that are assembled to form the memory package. It appears that each level of interconnection reduces the area bit density by a factor between 2 and 10; it becomes imperative to maximize the number of circuit functions per circuit termination, e.g. by on-chip address decoding, serialized data transfer, or multiplexing techniques, so that the number of individual modules can be minimized.

2.6.3 MOS Technology

The surface charge-separation effect in bulk silicon on which MOS LSI is based predates the invention of the junction transistor by 22 years [8], but it is only in the last two or three years that manufacturing processes have developed sufficiently to allow its commercial exploitation.

2.6.3.1 Basic Principle

A MOS transistor consists of a silicon oxide dielectric sandwiched between an aluminum gate and a silicon substrate. A voltage on the gate controls the conductance of a channel layer in the substrate between two highly doped source and drain regions (see Figure 2.6.1). Figure 2.6.2 demonstrates that source-to-drain conductance does not occur until the gate-to-source potential exceeds a certain threshold, and illustrates the pentode-like I_{DS} versus V_{DS} characteristic.

The transistor is categorized as p-channel or n-channel depending on whether the majority carriers in the channel are holes or electrons. It operates in depletion or enhancement-mode according to whether channel conductance does, or does not, occur in the absence of gate bias, this being a function of the substrate doping level. The p-channel enhancement mode transistor with a threshold of -3.5 to -5 volts is the most common MOS device today. Although n-channel transistors can achieve faster switching and smaller device geometry, by virtue of the higher mobility of electrons over holes (by a factor of two), they require tighter control over impurities and contamination, are more difficult to produce as enhancement

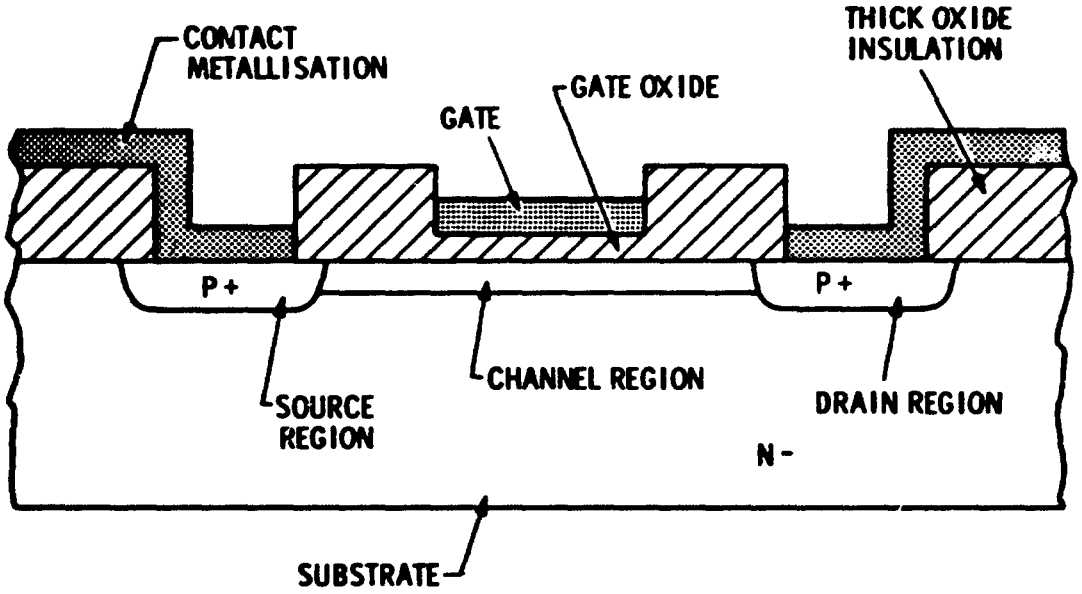


Figure 2.6.1 Basic Construction of MOS Transistor

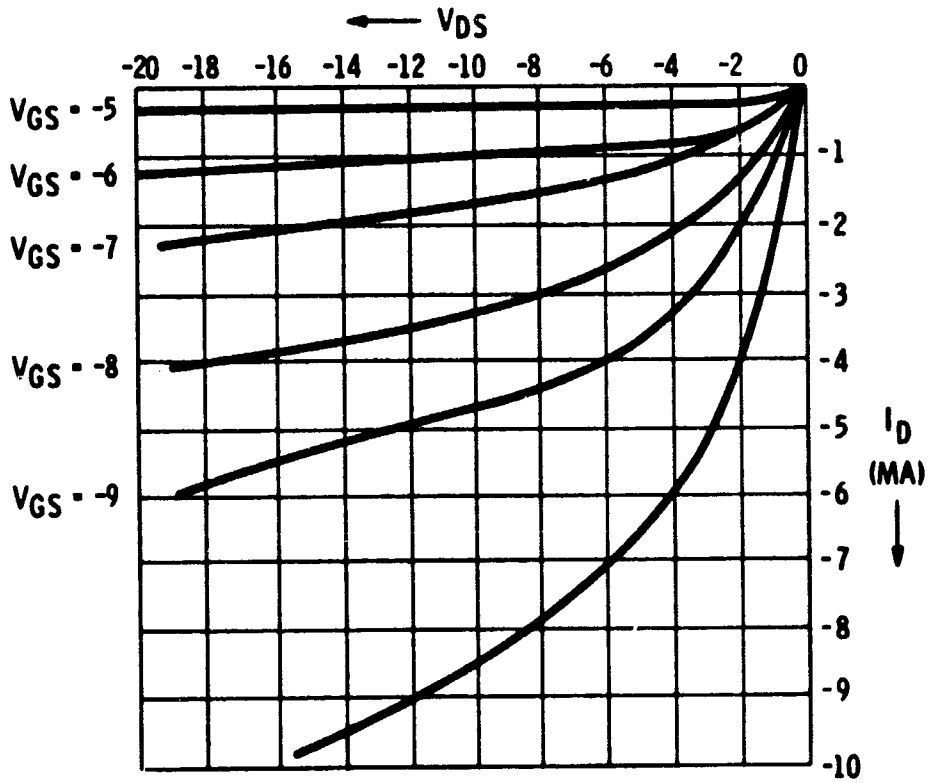


Figure 2.6.2 Source-to-Drain Conductance Characteristics for P-Channel MOS Transistor

mode devices, and are thus relatively unpopular. The combination of p and n-channel devices on the same substrate, the so-called complementary MOS principle, has attractive properties of high speed and low power and will be elaborated on later.

The source and drain of the MOS transistor are usually back-biased with respect to the substrate, and there is no interaction between adjacent devices. This obviates the need for the isolation regions that are required to separate devices on a bipolar integrated circuit chip, and which account for up to 30% of the substrate area.

Gate impedance is of the order of 10^{12} megohms; a MOS transistor does not load the previous stage, and there is no logic fan-out limitation. The gate capacitance-leakage resistance time constant is about 10 millisec, and forms the basis for the dynamic MOS memory cell.

The major drawbacks to the MOS transistor are:

1) The high threshold of 3.5 to 5 volts. This prevents direct interfacing with bipolar TTL circuits, which conventionally operate with a 4.5 volt logic swing. It also forces the use of high MOS circuit voltages to maintain adequate logic level ratios.

2) The high input gate and Miller feedback capacitances caused by gate overlap with the source and drain regions, together with the natural high impedance of the device (10K-100K ohms "on" resistance for small geometry transistors) make for slow switching speeds.

2.6.3.2 Reducing Threshold Voltage and Input Capacitance

The field-effect threshold voltage is given by [10]:

$$v_t \propto \frac{t}{k} (Q_{ss} + Q_b) + \phi_{ms} \qquad (2.6.1)$$

where

t = gate dielectric thickness

k = gate dielectric constant

Q_{ss} = surface state charge

Q_b = channel charge per unit area

ϕ_{ms} = gate-substrate work function

The following methods of controlling these parameters to achieve a minimum v_t are being investigated:

1) N-doped silicon of 111 crystal lattice orientation is commonly used as a substrate because its high Q_{ss} and threshold minimize the influence of manufacturing processes on threshold drift [9]. But improving techniques are lowering contamination levels sufficiently to allow the use of 100 silicon, whose Q_{ss} is one quarter that of 111 orientation. However, with 100 silicon there is a danger that conductors evaporated onto the thick oxide layer cause parasitic field effect inversions in the substrate. This can be reduced by thicker oxide insulation, thinner gate dielectric, and/or lower circuit voltages, each of which has compensating drawbacks.

2) Silicon nitride has a dielectric constant of 7.5 compared to silicon oxide's 3.8 and has been used as the gate insulator to reduce the threshold voltage to 1.5 to 2 volts. It is a more difficult material to process, and requires a thin layer of oxide between it and the substrate to prevent anomalous surface effects.

3) The work function term ϕ_{ms} can be lowered by using silicon rather than aluminum as the gate material [11]. Polycrystalline silicon is deposited on the gate oxide, and is then diffused with boron p-type impurity to increase its conductivity. This operation also defines the source and gate regions. The silicon gate process is self-aligning, since it uses the gate as a diffusion mask, and therefore produces lower gate capacitances, and allows smaller geometries. This technique offers low thresholds of 1.5 to 2 volts without restricting allowable circuit voltage levels to less than 15 to 18. Furthermore, because diffusion is performed after the high temperature step of gate oxide growth, the diffusion process tolerances may be relaxed, and device yield is higher.

4) The channel charge per unit area, Q_b of Eq. (2.6.1), can be controlled to yield desired thresholds from 4 down to 1 volt, by the process of ion implantation [12]. The channel region is selectively doped by bombardment with boron ions accelerated by 40-300 kilovolts through the gate oxide prior to gate metallization. By implanting after metallization and using the gate as a mask [12], the process yields self-aligning gate structures. Implantation can be conducted at room temperature. A high degree of control over the doping levels and implanted area is claimed. Radiation damage to the substrate is minimized by post-implantation annealing. Disadvantages of the method are that multiple wafer implantation is currently not possible, and that the implantation equipment is complex and therefore expensive.

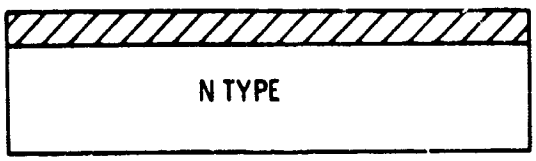
Most of these techniques are not available in high volume commercial devices, and the question of which process, or combination of processes, will receive widespread acceptance over the next 2-3 years is not clear. But it is probable that over the next five years MOS improvements will yield devices of high performance, density and reproducibility and that memory applications will provide the greatest impetus for large-scale production.

2.6.3.4 Complementary MOS

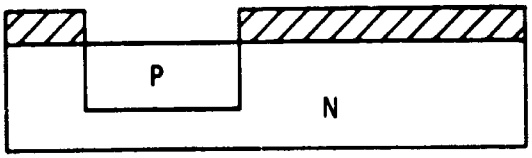
Complementary pairs of n and p-channel MOS transistors provide a highly non-linear circuit element: both are amplifying devices, one is off when the other on. The arrangement provides for high speed, high noise immunity, and very low power consumption [13]. In the past severe processing difficulties, such as the need for oppositely-doped regions on a common substrate, and the high sensitivity of the n-type channel transistor to contamination, have discouraged CMOS development. However, several manufacturers have been successful in making this process tractable, and CMOS LSI arrays are being produced in commercial quantities. Figure 2.6.3 illustrates the major processing steps used by RCA [14]. The large p-region that acts as the substrate for the n-channel transistor results in a larger cell geometry than for single channel transistors. Because all circuit elements are non-linear, logic level ratios can be less than for single p-channel logic: low threshold CMOS can operate with supply levels well below 10 volts. Provision of the proper biases between the p and n-regions and the substrate is a problem for the circuit designer, in that n-channel thresholds are more sensitive to this bias than p-channel.

The use of an insulating crystalline substrate, such as sapphire or spinel, is being explored for its application to CMOS fabrication [15]. A film of monocrystalline silicon is selectively doped with p and n-regions and then etched away to isolate the individual CMOS devices. Parasitic reactances between evaporated wiring and the substrate are effectively eliminated, leakeage currents between devices are zero, and the diffusion process tolerances can be relaxed. The substrate does not, however, provide an epitaxial surface to the silicon, and this forms the major processing difficulty. Devices built on sapphire have shown very high speed: the access time of experimental memory chips with full decoding is down to 30 nanoseconds, and standby power dissipation is a few nanowatts/bit.

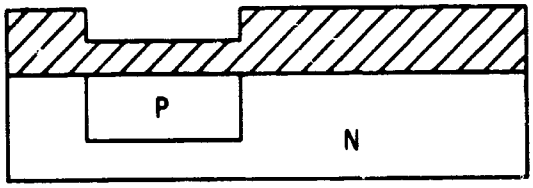
Although no CMOS memory has been built to exceed a few thousand bits, a currently available 288 bit random access



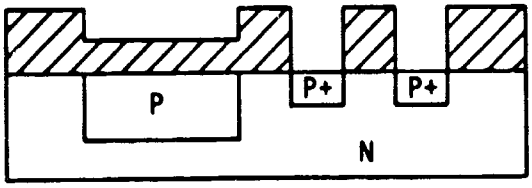
1) OXIDE WAFER



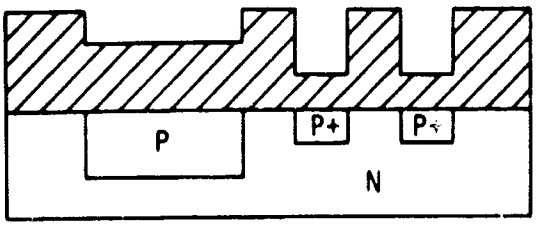
2) OPEN OXIDE AND DIFFUSE P WELL



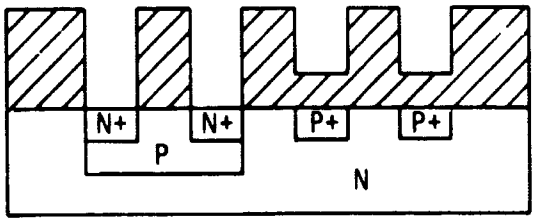
3) RE-OXIDIZE WAFER



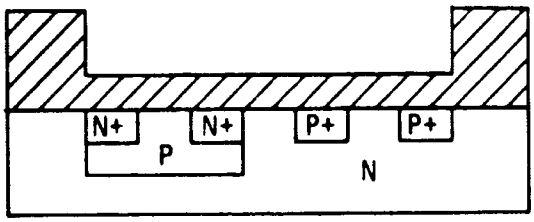
4) OPEN OXIDE AND DIFFUSE P+ S&D



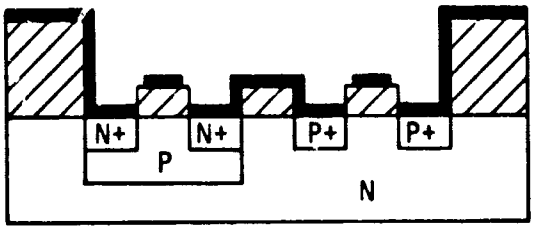
5) RE-OXIDIZE WAFER



6) OPEN OXIDE AND DIFFUSE N+ S&D



7) REMOVE OXIDE FROM ACTIVE REGIONS
AND GROW CONTROLLED CHANNEL OXIDE



8) OPEN CONTACTS AND METALIZE

Figure 2.6.3 Sequence of Processing Steps for CMOS Integrated Circuit

CMOS chip is being used to build a 0.5 million bit module for a memory that will be expandable to 5 million bits, with a power requirement of 2 watts at 100,000 bits/sec, and 0.25 watts in standby [16]. CMOS is very likely to be developed for low-power, high-speed logic functions, such as address decode, word drive and bit sense, because the more refined techniques described above are helping CMOS to break into the nanosecond ranges formerly the preserve of bipolar logic. It also has the ability to share the same substrate as other MOS devices more easily than bipolar circuits.

A number of industrial concerns are strongly developing CMOS memory techniques under government sponsorship, notably RCA and Westinghouse. CMOS LSI is being produced in large quantities for the computer peripheral and calculator markets. These are indications that complementary symmetry metal-oxide-semiconductor techniques will become a significant fraction of the MOS industry.

2.6.3.5 Charge Coupled Devices

This recently announced derivative of MOS technology appears to have some application to memory techniques [25]. The charge coupled device stores data by trapping minority charge carriers (holes in n-type silicon) in deep depletion regions at the surface of a silicon substrate (see Figure 2.6.4). The depletion region is formed by applying a voltage to a conductor evaporated onto the silicon dioxide insulation layer. Minority carriers are created by reverse-biased p-n junctions, by avalanching, or by photon interaction of light falling on the bottom of the substrate. The presence of a charge is detected by measuring the change in the effective capacitance, or the potential of the conductor.

The structure of a storage element is even simpler than that of an MOS transistor: no p or n-type diffusions are necessary, and no multiple etching of oxides is required. However, to provide storage and access in an integrated memory array, the only mechanization appears to be that of a serial shift register. For this three conductors are required to provide the 3-level potential wells necessary to transfer one data bit to an adjacent location (a similar problem to the shifting of the bubble memory's domain regions). The speed of transfer is less than 100 ns; but as in other types of MOS arrays, speed is likely to be limited more by conductor geometry than by charge mobility. As the charge transfers, it becomes attenuated, because a fraction, about 2%, is not transferred. Data regeneration thus becomes a necessity. The time constant for a stationary charge to decay is several seconds, so data must be kept moving. The current corresponding to the minority carrier motion is only 10^{-7} amperes

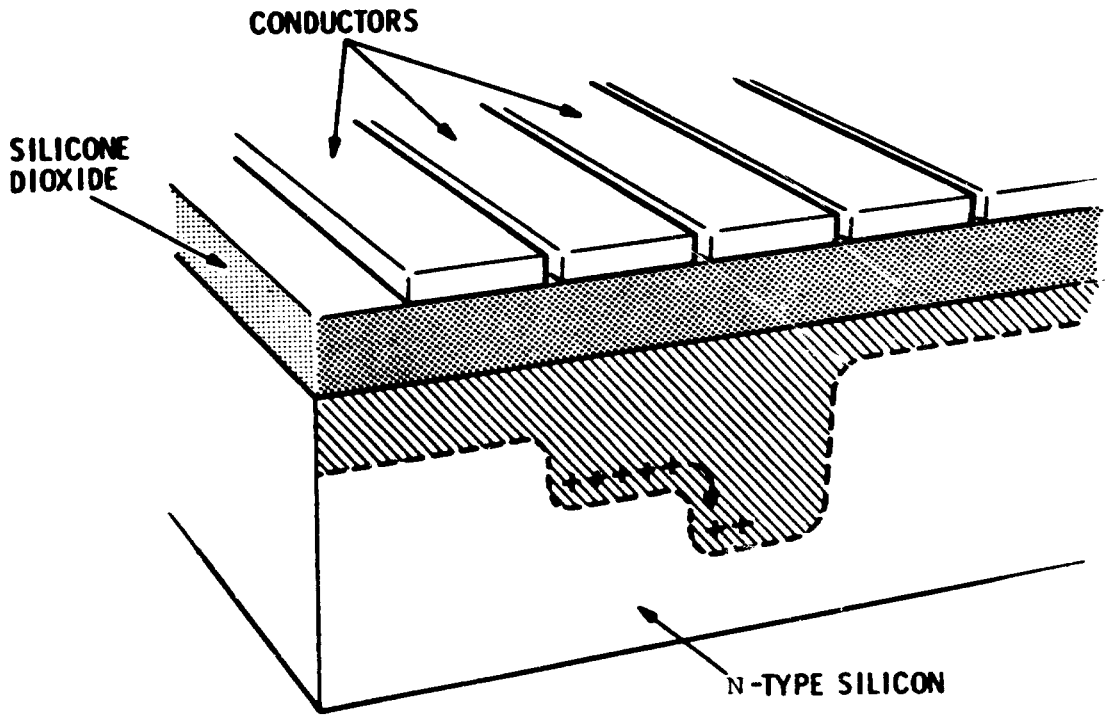


Figure 2.6.4 Principle of Charge-Coupled Device

at a frequency of one megahertz, but it is well above the thermal noise level of 10^{-10} amps.

The charge coupled device as a bulk storage element faces competition from other MOS technologies because:

1) it does not provide arrays as dense as MNOS, nor is it non-volatile (all potential wells and their contents disappear when the potentials are removed).

2) It is available only in a non-random access, shift register form.

It is thus unlikely to be a serious contender for a mass storage technique. However, an interesting possible application of the CCD may be as a light detecting array in optical memories.

2.6.4 The Variable Threshold MNOS Transistor

The threshold voltage of an MNOS device can be altered by applying a gate potential of an appropriate amplitude and duration [20]. The change has long term characteristics, with a decay time constant that has been observed to be as long as several months. It can easily be interrogated non-destructively, requires no power to maintain its condition, and may be reset by a suitable gate potential of opposite polarity. The MNOS structure is similar to the MOS transistor, and since only one device is required per stored bit, high density can be achieved. The electrical characteristics of the MNOS transistor are very similar to those of the regular MOS device. MNOS devices can be fabricated by a well developed process, that of MOS technology.

These characteristics make the MNOS transistor a very attractive memory element, and are the reasons for the industry-wide investigation of its properties.

2.6.4.1 The Storage Mechanism

The MNOS transistor differs from the regular MOS structure in that the gate dielectric consists of two layers of dissimilar insulating materials (see Figure 2.6.5). Silicon nitride on silicon dioxide is the most common combination, although alumina (aluminum oxide) has been used [21]. For suitable values of the electric field across the dielectric layers (about 3×10^6 volts/cm) free electrons will leave the bulk silicon, to be transported to the nitride-oxide interface, where they remain upon removal of the field. Two models have been proposed for the transportation and retention of charge

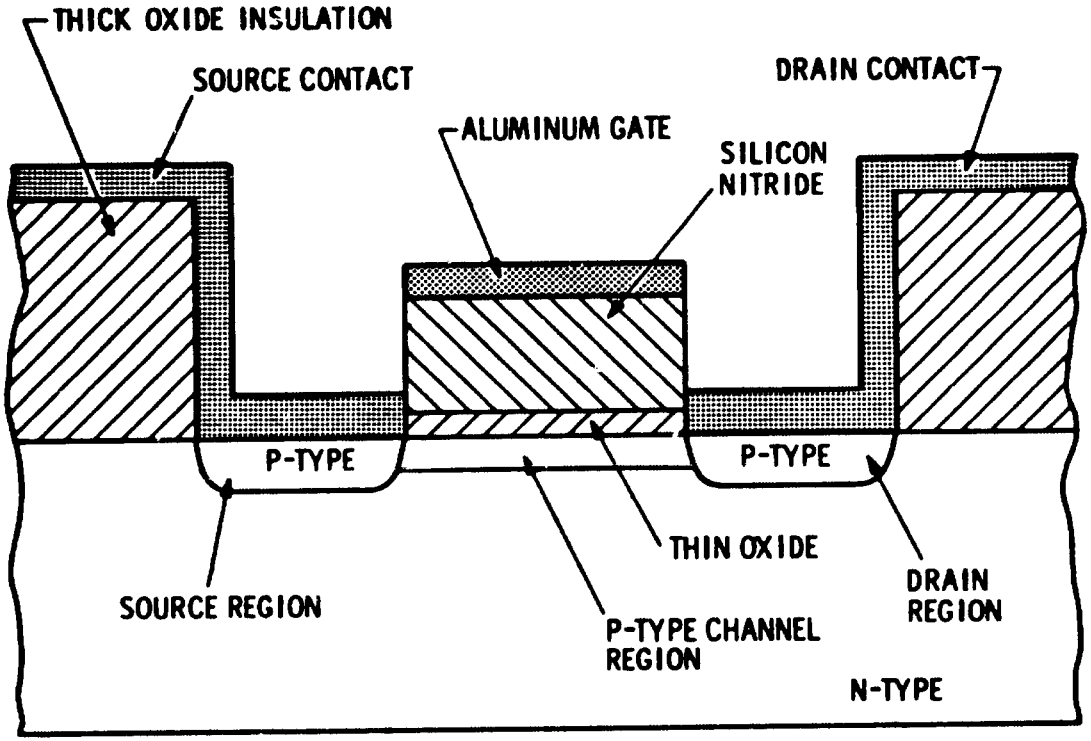


Figure 2.6.5 Construction of P-Channel MNOS Transistor

in the dielectric layer according to whether thick or thin silicon oxide layers are used. The first [22] involves the concept of differential electrical conductivity within the insulating layers. Figure 2.6.6 illustrates that the conductivities of silicon nitride and oxide vary dramatically with the impressed electric field. The conductivity theory suggests that the nitride-oxide interface can be charged rapidly by conduction through the two layers under the influence of a high "write" field. Very little charge is conducted away by the lower "read" field, and even less leaks away due to the field created by the charge itself during standby. The other theory [23] proposes that energy level "traps" at the oxide-nitride interface can capture electrons from and release them to the bulk silicon by tunnelling through very thin oxide layers of less than 50 Å under the influence of positive and negative electric fields. The probability of electron movement in and out of these traps diminishes rapidly as the electric field strength is reduced.

For a nitride thickness of about 1000 Å, and an oxide thickness of 100 Å, a gate potential of approximately -30 volts will typically change the threshold of a p-channel device by about -5 volts. A +30 volt pulse will restore the threshold to its former value. The gate voltage, source current curves for the two threshold values is shown in Figure 2.6.7. The amount of shift is dependent on the gate voltage amplitude and duration, the insulation thicknesses, and the substrate doping levels. For certain doping levels the transistor can even be changed from an enhancement to a depletion mode device by the application of positive gate write voltages (i.e. the threshold is shifted through zero to become positive, and conduction will occur for zero gate bias). This is undesirable for memory application since it constitutes a current drain in the uninterrogated mode.

The stored information is sensed by interrogating the device with a gate voltage midway between the threshold limits. If channel conduction occurs, the cell is in the lower threshold condition representing a "zero", and if it does not, it indicates the higher threshold, representing a stored "one".

2.6.4.2 Performance Characteristics

For a given transistor structure there is a relationship between the duration of the write pulse required to alter the threshold voltage and the period over which this threshold shift is retained. A typical ratio for these times which is being achieved in working devices today is 10^{12}. For example, if a gate write pulse of one microsecond is applied, the resulting threshold voltage shift will decay with a time constant of $10^{-6} \times 10^{12} = 10^{6}$ seconds, or about 12 days. Ratios of 10^{13} are being demonstrated on an experimental basis, and one in-

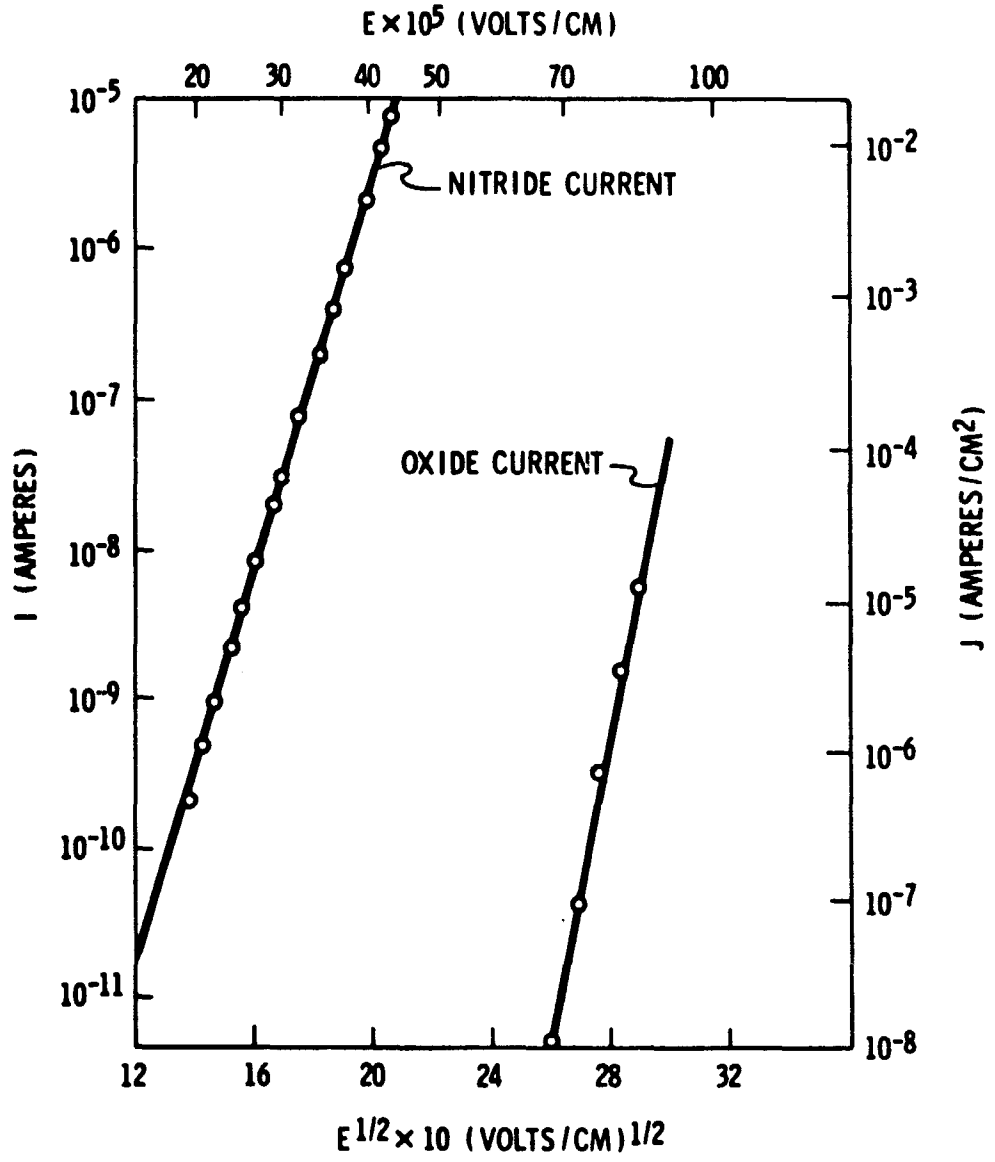


Figure 2.6.6 Conductivity of Silicon Oxide and Nitride as a Function of Electric Field

vestigator, Sperry-Rand, claims that 10^{16} should in theory be achievable. As the written-in charge decays, the threshold voltages corresponding to a stored "zero" and a stored "one" will tend to approach each other. A separation in the two states of 2 volts is generally taken as the minimum indication of reliable storage.

The required amplitude of the write pulse in early MNOS devices was as high as 100 volts, well outside the range of I.C. driver capabilities. But this has now been brought down to the 30 to 40 volt level, and work continues to lower this to the 20 to 30 volt range that is within the capability of regular low-threshold MOS devices.

The greatest impetus to MNOS research comes from the prospect of the electrically-alterable read-only memory (EAROM) rather than the RAM. Consequently more emphasis is being placed on extending the retention time than in shortening the write time. The effect of the environment on retention time is being studied, e.g. temperature and radiation. Retention time is affected by high temperature: Ref. [23] reports that at 150°C a threshold shift of 2V was observed at 1000 hrs, whereas none was observed after 4000 hrs at room temperature. The radiation resistance is good for a field effect transistor because the gate dielectric is mainly silicon nitride (see Chapter 3.0). Moderate radiation doses, such as experienced in earth orbital environments, produce a positive space charge in the gate dielectric layers. The effect is to produce a negative shift in the threshold voltage, which must be kept negligible compared to that produced by the basic MNOS mechanism.

The read characteristics are similar to those of regular MOS devices, and read-out is accomplished in the same way as for MOS arrays. The switching time for the channel is of the order of a few tens of nanoseconds. But the access time of a MNOS array of more than a few bits is determined to a greater extent by the gate and conductor-to-substrate capacitances, by the driver transistor characteristics, and by the output sensing arrangement, than by device switching delay, just as in the case of a pMOS or CMOS array.

2.6.4.3 MNOS Memory Arrays

Since only one MNOS device is required to store a bit, the cell is very small, typically 0.3 to 0.7 sq. mil. Array density is limited more by the finite conductor widths and spacings of about 0.2 mil than by device size. Chips containing 1024 bits have been made to modest dimensions: 50 by 80 mils. Much larger arrays are possible: 8,192 bits and even 16,384 bits on a 120 x 120 mil chip are considered to be very feasible and would not involve state of the art improvements in photolitho-

graphy, according to Sperry-Rand. The indications are that the MNOS approach to semiconductor arrays using today's MOS processing techniques will permit data densities of well over 10^6 bits/in^2 to be realized. Processing improvements, such as the use of an electron beam for defining the geometries of the devices and the conductors, may allow conductor widths down to 1 μm (=.04 mil) to be achieved. This would make bit densities in excess of 10^7 bits/in^2 possible.

The major consideration with the larger arrays is that on-chip decoding becomes essential to minimize the number of external connections. MNOS transistors arrays present one or two unique problems to the logic designer: the write voltages are somewhat higher than the 20-25 volts that are current MOS practice; the writing of ones and zeros requires high gate voltages of opposite polarity; and in some instances the writing of "ones" and "zeros" involves pulses with an assymmetry of amplitude and duration. To preserve correct biasing with respect to the substrate, some of the address decoding and array driving logic must be isolated from the memory array. This can be done by separate chip connections, or preferably by diffusing isolation regions on the chip. Fairly complex voltage switching to the gate, the source and drain, and the substrate are necessary to achieve the correct polarities and amplitudes for reading and writing ones and zeros. To date the experimental arrays that have been built have employed partial on-chip decoding by pMOS transistors, and off-chip sensing by bipolar circuits. Large arrays for memory systems will need to contain full decoding and on-chip sensing if the interconnection of several thousand chips is to be accomplished. A compatible MOS circuit technique would appear to be the most natural medium in which to implement this logic, as it could then be accommodated on the same chip with a minimum of additional processing steps.

2.6.4.4 Problems Yet to Be Solved

The intensive investigation of the MNOS transistor by the semiconductor industry over the last two to three years has resulted in a high level of understanding and controllability of the storage phenomenon; the major thrust must now be to integrate the device into memory systems larger than the individual chips that have been produced to date.

Among the system problems that have to be tackled is the question of the MNOS storage mechanism's long term volatility. Although refined processing techniques may eventually increase the retention times to many thousands of hours, it will be necessary to devise data refreshing techniques for memory systems built during the next two or three years. The principle is analogous to that employed to prevent evaporation of data

in dynamic MOS memory cells, except that the refresh periodicity will be of the order of tens or hundreds of hours rather than a millisec or two. This problem is more severe for read-only applications, as random access memories can be organized to keep data moving, either through normal operation or by system design, thus obviating the need for special refresh circuitry.

The establishment of firm device reliability data is essential to the exploitation of MNOS. Early samples showed poor life times under conditions of repeated writing, but it is thought this was due mainly to failure of the thin gate dielectric under the stress of the fields produced by the formerly high gate potentials. Some threshold voltage drift is still being observed after 10^9 writings of alternate ones and zeros. The dielectric layers apparently require a more careful processing to achieve repeatable MNOS characteristics than the regular MOS structure. The long-term reliability of stored data in the continuous radiation environment of an orbital space station must be established. An additional function of the refresh mechanism described above could be to provide the MNOS memory system a degree of immunity from the accumulative effects of radiation fields.

Reliability data can only be accumulated by the continuing investigation and application of the MNOS mechanism over a number of years.

MNOS device performance improvements are expected to continue and perhaps even accelerate under the impetus of its application to operating memory systems. Random access memory applications will in general demand faster write times, on the order of the read time capability, i.e. less than 1 μs. ROM application will demand long retention times, upwards of 10,000 hrs, and a high degree of non-volatility under all conditions. In addition device design will have to accommodate to the decoding, driving, and sensing logic circuit problems outlined in section 2.6.4.3.

Several of the problems that require solution before MNOS memory systems of significant capacity can be realized are common to those of other semiconductor memory techniques:

1) Manufacturing processes must be refined to achieve a level of yield of good chips that is sufficient to make the technique economically attractive.

2) Chip mounting and bonding, and module interconnection must be effective, fast, reliable, and economical.

3) The integration of large numbers of memory cells in a dense package must be accomplished without sacrificing speed and signal-to-noise ratio, and without incurring

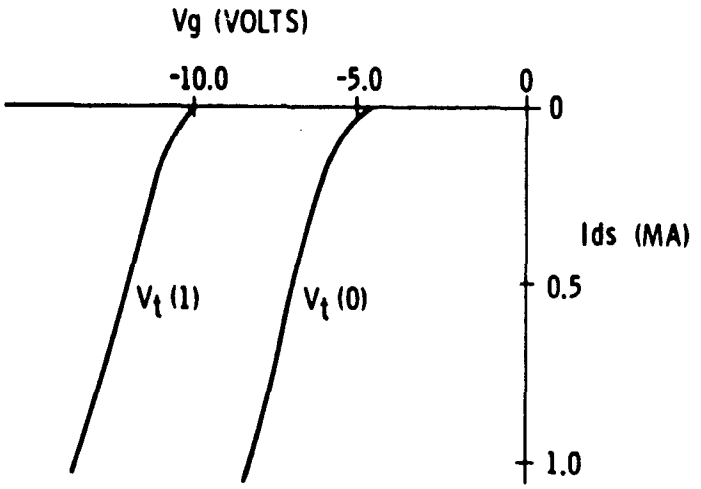


Figure 2.6.7 MNOS Threshold Voltage Characteristic

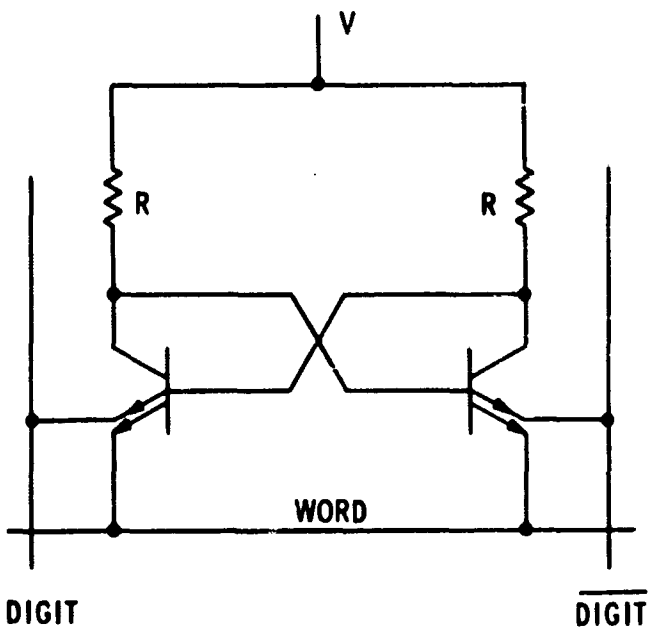


Figure 2.6.8 Bipolar Memory Cell

excessive power dissipation.

The construction of a 10^6 bit MNOS memory system is currently being undertaken by Sperry-Rand for the Air Force. The specifications have been written as a drum replacement for airborne computers, and the emphasis is on density and retention time rather than speed. The main purpose of the project is to develop insight into the problems of MNOS memory system design.

2.6.5 Practical Considerations of Semiconductor Arrays

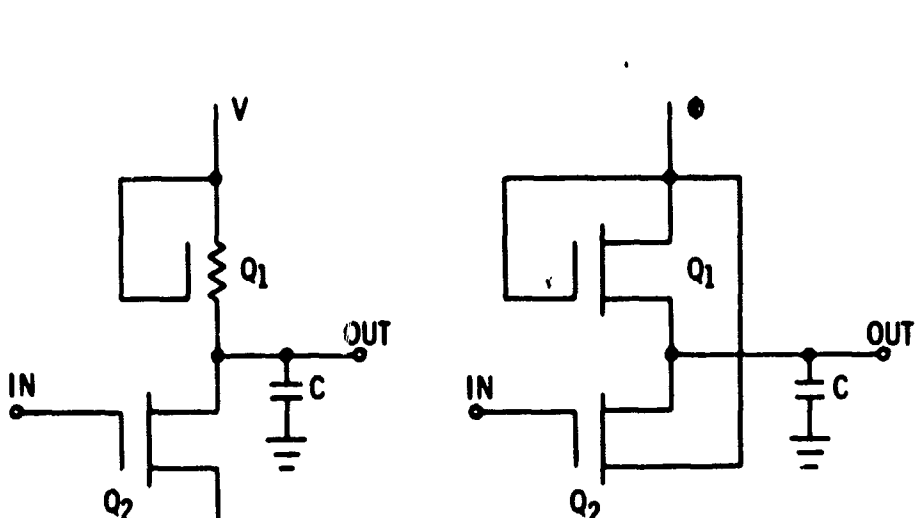
2.6.5.1 Circuit Design

Bipolar memory arrays are constructed of back-coupled flip-flop circuits, in which gating is usually performed by multiple-emitter devices. THE ILLIAC IV memory array element [6] of Figure 2.6.8 is a simple example. A single 5 volt supply is required. Power dissipation is typically 1 to 2 milliwatt per bit. Resistors R are formed by doping substrate regions to the required resistivity; they are much larger than equivalent MOS resistances by two orders of magnitude.

MOS arrays are based on the inverters of Figure 2.6.9. In the static inverter of Figure 2.6.9a Q_1 is the "on" resistance of a small geometry transistor, and Q_2 is a switch. The output logic level ratio is

$$\frac{R_{Q_2}}{R_{Q_1} + R_{Q_2}} .$$

Different geometries are required in this design, and output trailing edges are much slower than leading edges, since $R_{Q_2} < R_{Q_1}$. If this inverter is back-coupled it forms the static memory cell of Figure 2.6.10 which can retain its state as long as V is applied. It typically dissipates about 10^{-3} watt, whether operating or in standby. If supply voltage V is replaced by a clocked source ϕ operating between V and ground, and correctly phased with respect to the input, Q_1 can pull the output voltage on C rapidly to V. When ϕ returns to ground, Q_1 cuts off, and the voltage on C representing a '1' decays slowly with a time constant around 10 ms. An input turns on Q_2, which grounds C to form a '0'. This circuit can use equal geometry transistors, is independent of resistance ratios, and provides faster operation. It is the basis of the dynamic MOS memory cell of Figure 2.6.11. Since the output is only held for a few millisec, C must be periodically recharged. In a large random access array this is done by a refresh amplifier typically servicing 32 cells, at about 2 ms intervals. For a 500 ns cycle time memory of 1024 bits, a total of 32 cycles



(a) STATIC (b) DYNAMIC

Figure 2.6.9 MOS Inverter Circuit

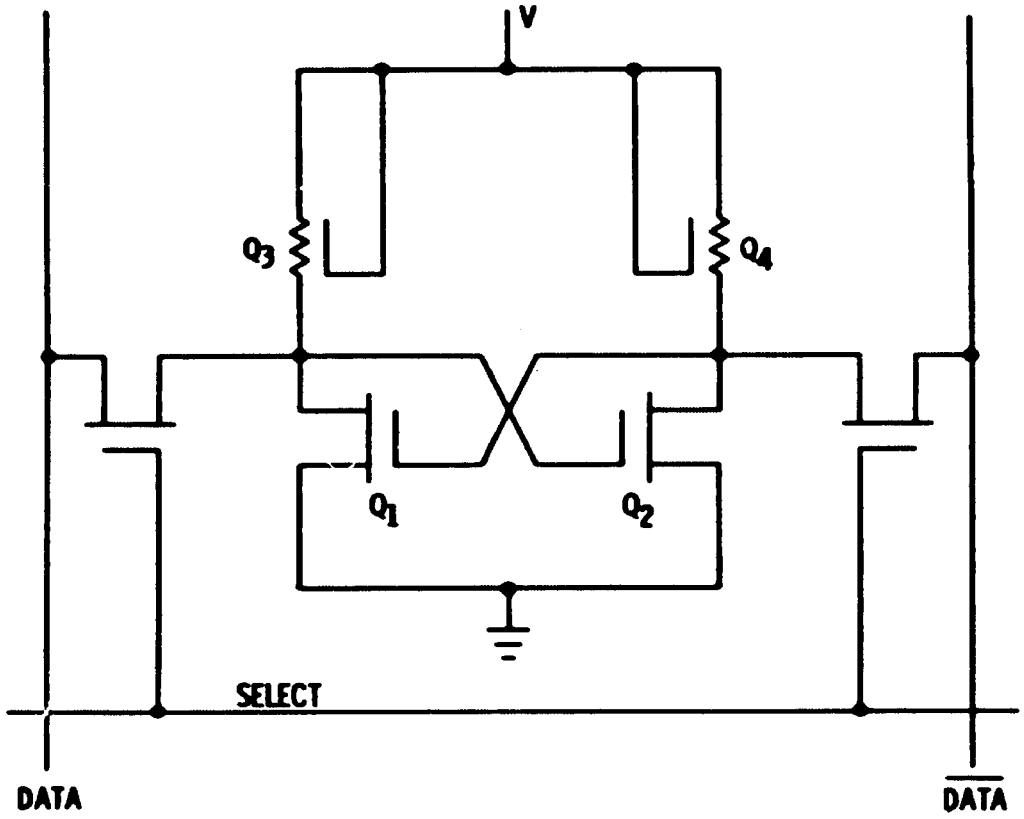


Figure 2.6.10 Static MOS Memory Cell

at 500 ns each, i.e. 16 μs out of 2 ms or less than 1% of the memory cycles are devoted to refreshing the data. Shift registers based on the dynamic cell refresh their data during normal operation. Standby power of the MOS dynamic RAM array is about 10^{-6} watts/bit, and operating power is about 10^{-4} watts/bit. Power is dissipated to a large extent by the CV^2 losses incurred by the charging and discharging of drive and sense line capacitances. For example, a typical 100 pf total chip capacitance draws 200 milliwatts from a 20 volt supply if charged and discharged every microsecond. The same cause contributes significantly to the access times of MOS arrays, which for static p-channel arrays are about 1000 nanosecs and for dynamic, 300- 600 nanoseconds.

The two basic CMOS circuits are the inverter and transmission gate of Figure 2.6.12. Since Q_1 and Q_2 are always in opposite states, the CMOS inverter provides rapid charging and discharging of C through driver transistors, just as the dynamic MOS inverter, but without multiple clocking. In addition CMOS is insensitive to supply variations and noise. Standby power dissipation is V times the device "off" leakage current, typically 10 volts times 1 nano amp, or 10^{-8} watts. Operating power is dissipated almost entirely in the charging of the stray capacitances during state transitions. A simple CMOS flip-flop is shown in Figure 2.6.13. The transmission gate is bi-directional, and performs set-reset and output gating functions depending on the states of the W and D-lines. Device area on the chip is typically 50 sq. mils, i.e. 3-4 times the size of a MOS dynamic cell (see Section 2.6.3), and write cycle time for one cell is less than 20 nanosecs. CMOS shift registers have been designed and run at frequencies up to 100 MHz.

A summary of typical performance figures for currently available semiconductor memory arrays is given in Table 2.6.1.

2.6.5.2 Array Organization

One of the most important factors in the integration of semiconductor chips to form larger arrays is the minimization of the number of external interconnections. Various on-chip array configurations are possible to effect this. A bit-organized random access array of N cells requires only one data output (which may mean four pins, if complement data, in and out, is to be available). The array may then be squared off, and the bit selected by coincident conductor geometry. This takes $2\sqrt{N}$ address connections (or $4\sqrt{N}$ if complements are required). Power, ground, chip-select, read/write, and clock lines add 5-10 more pins. The minimum pin count of a 1024 bit chip without address decoding would therefore be 70 to 80, and for an 8192 bit chip, over 200. This is impractical for the interconnection of thousands of chips in a large memory. Full binary address decoding on the chip can reduce the $2\sqrt{N}$ address

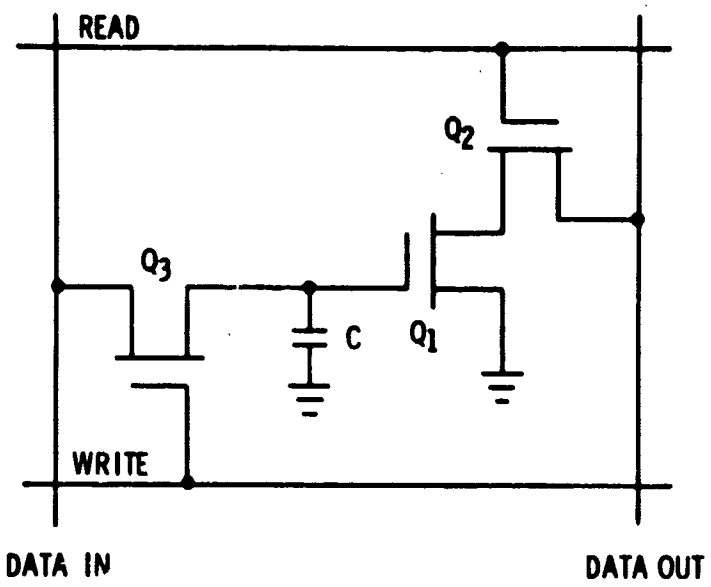


Figure 2.6.11 Dynamic MOS Memory Cell

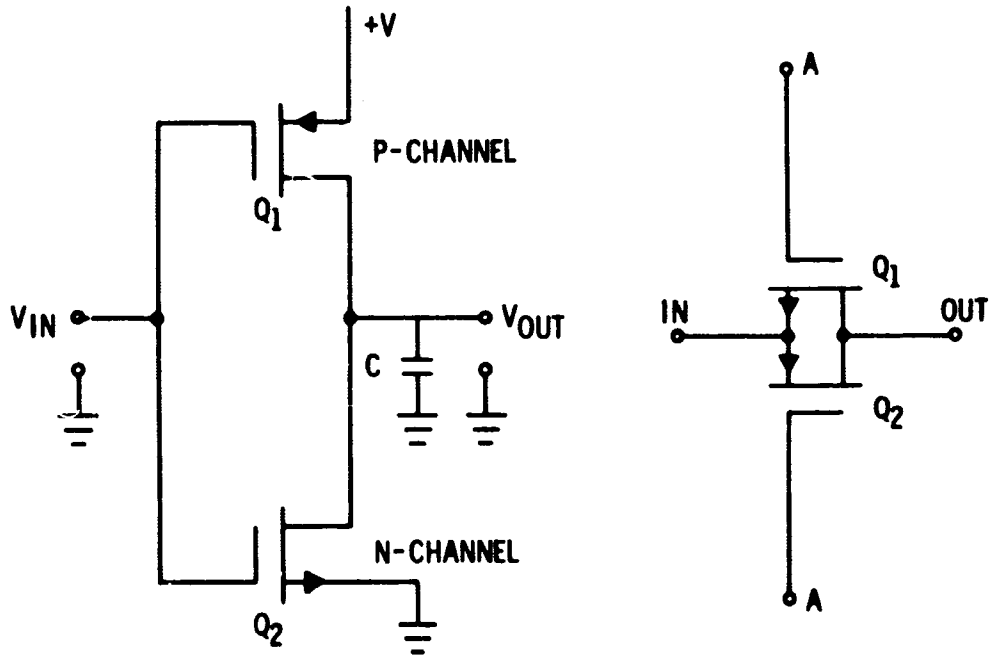


Figure 2.6.12 CMOS Inverter and Transmission Gate

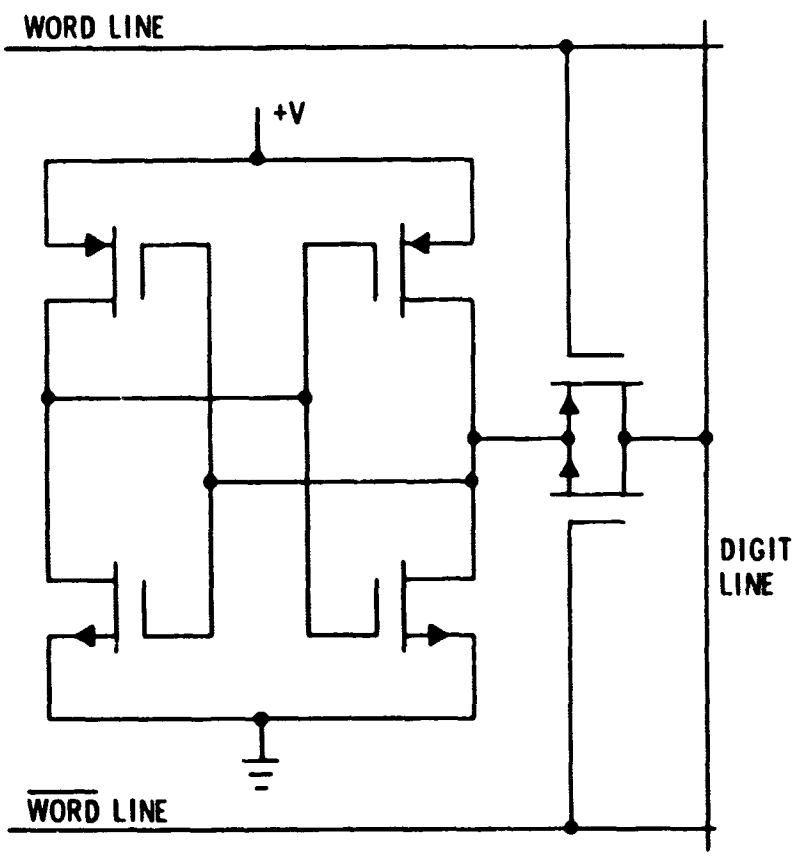


Figure 2.6.13 CMOS Memory Cell

	Cell Area (sq. mil)	Bits/Chip (with decode)	Standby Dissipation (watts/bit)	Chip Cycle Time (nanosec)	Cost/Bit (100 qty)
Bipolar	50-100	256	$> 10^{-3}$	70	50¢
P-channel MOS					
Static	25-50	256	$< 10^{-3}$	1000	15¢
Dynamic	10-15	1024	$< 10^{-6}$	300-600	5¢
CMOS	50	288	$< 10^{-7}$	50-100	-
MNOS	1	1024	zero	$10^{3} - 10^{6}$	-

Table 2.6.1 Current IC Performance Characteristics

pins to $\log_2 N$. However, this is done at the expense of substrate area to accommodate the decoding circuits, which generally do not pack as densely as the regular array elements, and incurs a major increase in the access time contributed by decoding logic delays. High speed bipolar logic can minimize these delays, but it is very difficult to integrate with MOS devices on the same substrate. CMOS can match bipolar speeds at a far lower power dissipation, and has compatibility with single-channel MOS manufacturing processes. The combination of CMOS decoding and single-channel MOS memory cells would seem to have attractive possibilities.

Most semiconductor RAM chips available today are bit-organized with on-chip address decoding. Figure 2.6.14 shows the 20 pin definitions for a typical 1024 word x 1 bit dynamic pMOS RAM. To maintain speed, current sensing is often used for output, since it minimizes voltage swing and CV^2 losses. Bit organization helps to minimize the necessary interface circuitry, which cannot always be accommodated on the chip. A memory of a given capacity will, independent of organization involve a given total number of chips. Bit organization achieves the lowest number of interconnections and therefore contributes directly to the reliability of the system. An additional indirect consequence of bit organization is that a chip failure produces only one error bit in each word, unlike word organization, in which a failure will affect two or more bits per word. One error is easily picked up by simple parity checking systems, and may be corrected by single correcting Hamming redundancy codes.

Shift registers do not provide access to individual data locations, and therefore do not require as many pins as the equivalent capacity RAM. Here only input and output, power and control connections are required. There is a direct penalty in access time, which averages $\frac{1}{2}N/f$, where f is the shift frequency. In dynamic MOS shift registers, however, f may vary from the maximum near 5 MHz to as slow as the 1 KHz refresh rate; this allows information look-ahead to be performed and a shorter effective access time to be achieved.

2.6.5.3 Interconnection and Packaging of LSI

The highest packing density for a semiconducting memory would be achieved by placing the whole array on one chip. Only connections for power, input/output, and addressing would be required. For capacities above 10^6 bits this would involve a substrate area of tens of square inches. The probability of producing a fault-free slice of these dimensions is almost infinitesimal. Large slice LSI has been produced from a single 1½ or 2 inch wafer, by a process of isolating the faulty regions or elements [24]. After the manufacturing process the

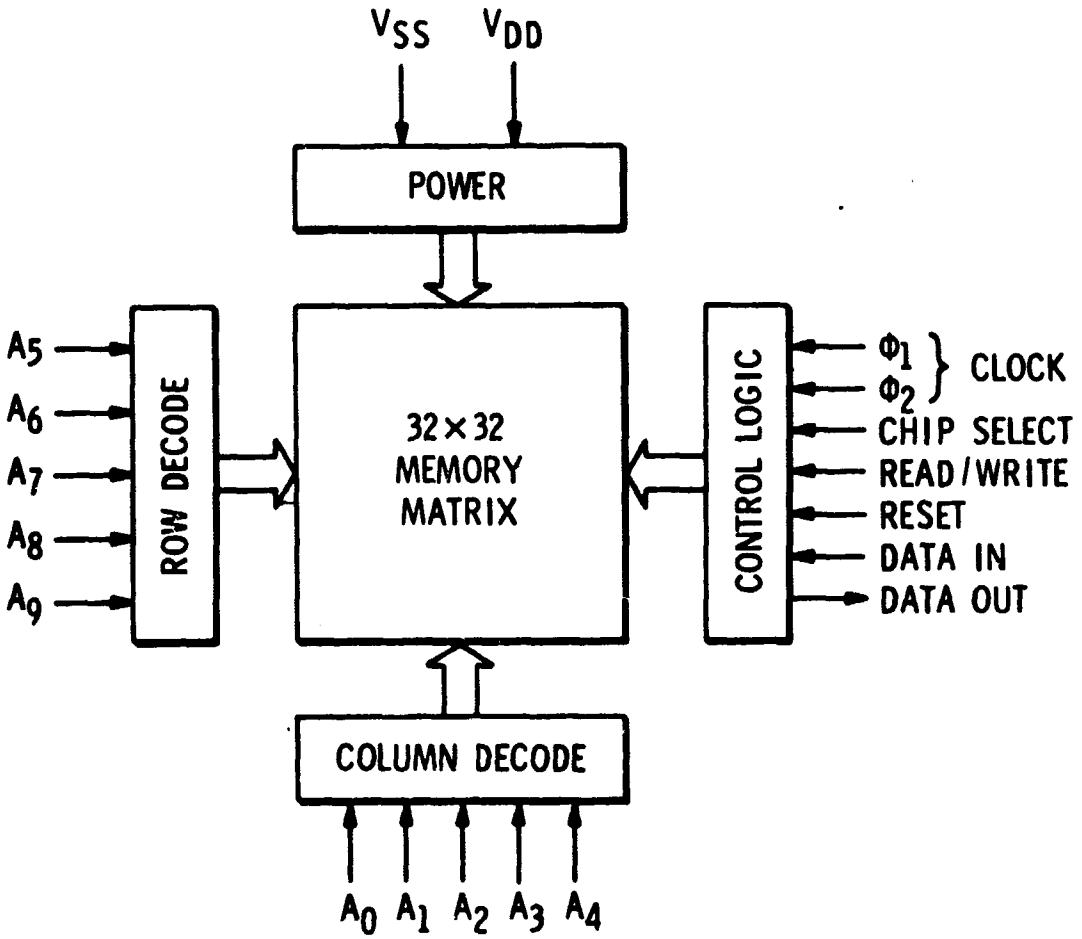


Figure 2.6.14 Connections to 1,024 Bit Array

completed slice is probe-tested, and the positions of the good elements are defined to an automatic program that generates a metallization pattern to interconnect them to achieve the desired function.

There is a tradeoff between the number of chips in a system and their complexity. The complex chip is obviously expensive, but only a few are required. The simple chip is cheaper, but the interconnection costs grow with the number of chips. For a particular application there will be a minimum cost for the system (Figure 2.6.15). The number of chips to achieve the minimum costs can obviously not be generalized for all systems.

In order to create larger logic systems than can be realized within the limited area of a silicon integrated circuit chip, a method of mounting and interconnecting chips must be found. A very desirable characteristic of the method should be that the high packing density achieved on the chip not be drastically degraded in the process. Secondly, the method should provide good mechanical support and reliable bonds. Thirdly, and perhaps most important, the method must be economically attractive or else it will not become accepted.

Standard techniques today show a very poor packing efficiency, because so far only the second and third points above have received much attention. Chips are most commonly mounted in ceramic "dual in-line" packages or DIPs, and these are soldered to printed circuit boards. Of the total volume of a typical board containing 70-80 DIPs, up to 98% is devoted to interconnection and support structure, and only 2% is represented by the DIPs. The chips themselves account for less than .02% of the total volume. Before memory systems of up to 10^8 bits based on semiconductor LSI can be realized some improvement in the packaging efficiency must occur. The first problem is the establishment of a satisfactory chip mounting and connection technique.

The current status of LSI chip bonding and packaging is in a high degree of flux. It is conceded by industry that this is the major snag in the widespread acceptance of LSI [17]. Up to now the majority of integrated circuit bonding to larger substrates, or "headers", has been by flying leads. Chips are attached face up on a substrate, and fine wires are connected between bonding pads on the chip and the substrate surface. The operation is performed manually, one wire at a time, and the two bonds per wire are inspected visually. The method is well understood, allows cross-overs, and does not require pads to be located at the edges of chips. But it is time consuming, and costly, in proportion to the number of connections, and has reliability problems.

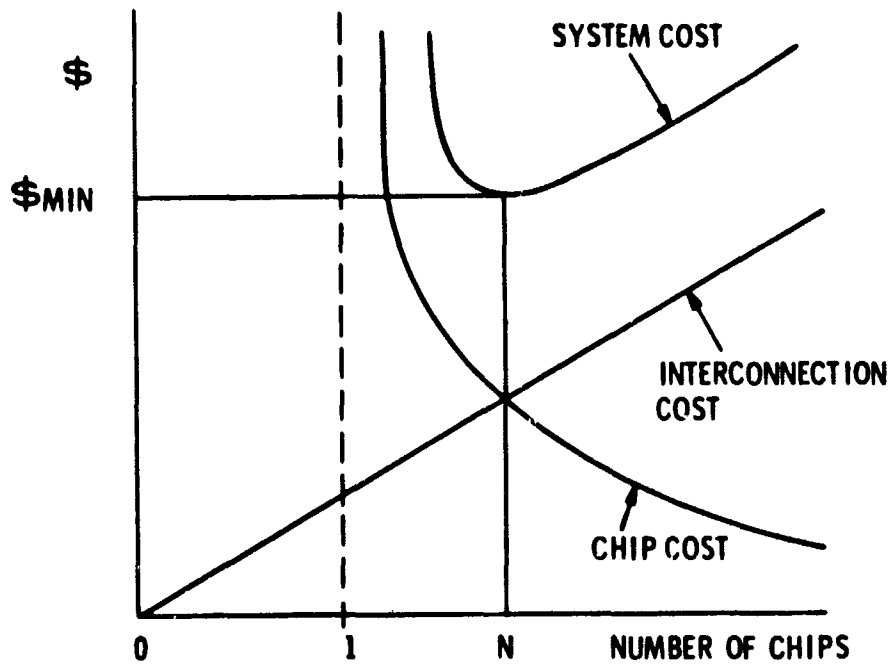


Figure 2.6.15 Memory Cost Versus Chip Complexity

Other techniques attempt to achieve processes that allow automatic, simultaneous bonding of many connections [18]. The main contenders are described below:

1) Beam Leads. In this technique, pioneered by Bell Laboratories, rectangular leads of gold, 0.5 mil thick and 3 mil wide, overhang the edge of the chip by 5 mils. The leads are formed on the substrate by several stages of evaporation and etching in addition to the regular IC process. Extra care is necessary to prevent gold migration and to assure good adherence and contact. Chips are passivated with silicon nitride before metallization and coated with glass afterward to protect the leads during handling. This, it is claimed, makes hermetic sealing unnecessary. Chips are separated from the wafer by an etching process that leaves the leads projecting over the edge of the chip. After separation they retain their orientation, and are therefore adaptable to automatic handling equipment. Eventually the chip is placed on its substrate, face down, and all beams are bonded to their respective contacts in one operation. Costs are 20%-30% higher than chip-and-wire, but improved yields are expected.

2) Bumped Bonds. In this technique chips receive additional treatment to create raised contacts, or "bumps", on the chip surface. The material used depends on the bonding technique. The bumped chip is placed on its substrate, face down, and is bonded to the contact areas by thermo-compression or application of heat. Thermo-compression bonding requires the bumps to possess plastic flow properties. Suitable materials are gold, silver, aluminum, or copper. Heat requires a low melting point alloy such as eutectic lead-tin. Conducting epoxy has also been used in a cold process.

3) Spider Bonding. The leads are preformed in a one-piece 2 mil aluminum frame which is attached to the chip in one operation by ultrasonic or thermo-compression bonding. The leads are then separated from the frame, and from this point the process is similar to beam lead bonding, but without the problems of fragility and thin nitride insulation failure.

All the foregoing techniques are "flip-chip" processes and suffer from a number of disadvantages: the possibility of trapped material between the chip and substrate, poor heat dissipation, and the difficulty of visual inspection of bonds.

4) Beam Laminate. The beams in this technique are formed

on an insulating film of Kapton with the conducting pattern etched on it. Holes to accommodate the chips are etched out, leaving connecting beam leads cantilevered over the edges. The chips are attached with good thermal contact to a separate substrate, in registration with the holes in the film. The film is then laid over the array of the chips, and the beams bonded to the contact areas on the chips. The method allows face-up bonding, and provides good heat dissipation. Motorola is offering an 8,192 bit pMOS memory system [19] constructed of aluminum modules, each containing 9 chips connected by beam laminate bonding. The modules form 13% of the total volume of 6 cubic inches, but the chips are still only about 0.5% of the total.

Which of these techniques becomes the standard for second-level interconnection is impossible to tell. Each is receiving considerable attention from the industry. Beam leads appear to have a slight edge, and are being offered commercially by Raytheon. Bump bonds have been used for IBM's Solid Logic Technology in the 360 series. It is without question a problem that must be resolved before the full promise of LSI is realized in a reliable, economical memory system.

2.6.5.4 Environmental Factors

The semiconductor integrated circuit chip itself is a rugged device. It is relatively insensitive to the effects of acceleration, vibration and shock, by virtue of its monolithic nature, small size and low inertia. Fragility and unreliability in an IC system is contributed by the second-level substrates upon which the ICs are mounted, and by the IC and substrate interconnection methods. Estimates of failure rates of interconnection schemes in the environment of a space vehicle are between 10^{-3}% and 10^{-4}% per 1000 hrs [26]. Developing techniques (see 2.6.5.3) that provide automatic, mass-bonding capability are expected to improve these figures to 10^{-7}% per 1000 hrs because of their greater precision, uniformity, and reproducibility.

The silicon IC can withstand wide temperature variations.

The sensitivity of MOS ICs to surface contamination can be minimized by passivation during the manufacturing process with the deposition of highly inert silicon nitride over the finished surface. Glass encapsulation, and hermetic sealing of the 2nd-level substrate further increase the resistance to hostile environments.

The one environmental hazard to which ICs are relatively vulnerable compared to other memory technologies is radiation. A detailed description of the nature of this sensitivity is

given in Chapter 3. The bipolar IC can be made fairly insensitive to even severe radiation by the technique of dielectric isolation. The diffused isolation region which normally forms the largest junction around each device on the silicon substrate is replaced by silicon oxide in an additional series of processing steps, or by an insulating substrate, e.g. sapphire. The first technique results in a lower bit density while the second involves an expensive material and difficult processing. In the benign radiation environment of non-military space application the bipolar IC is fairly immune and such isolation will not be necessary. The MOS transistor, however, is very sensitive to the ionizing effect of even moderate radiation in the gate dielectric, which is manifested as a shift in the threshold voltage. A silicon nitride gate dielectric can greatly decrease this sensitivity because its denser structure inhibits ion movement, while affording other advantages (see 2.6.3.2), and a miniminal amount of shielding can be quite effective (see Chapter 3). To date not very much information is available on the long term stability of MOS or MNOS devices in sustained mild radiation environments.

2.6.6 Semiconductor Memory Status and Trends

2.6.6.1 Current Status

Although there is a great deal of activity throughout the semiconductor industry in the application of LSI techniques to the production of memories and memory systems, relatively few examples of large (greater than 10^6 bits) semiconductor memories are extant. The reason is probably an economic one. In order for semiconductors to gain a footing in the mainframe memory field, a cost effectiveness advantage over cores must be demonstrated. For costs to decrease appreciably, a high volume demand must exist. Each factor, cost decrease and demand, must precede the other; this is plainly a paradoxical situation, and one which will only slowly be resolved. However, the almost inexorable improvements in MOS performance, production methods and costs, and the fierce competition among semiconductor manufacturers will eventually break the deadlock. At the present time manufacturers are aiming at small memory applications, such as the desk calculator, and are concentrating on refining and consolidating their IC manufacturing techniques. The result is the ready availability from a large number of semiconductor houses of small single chip memory arrays (from 16 bit RAMs to 4096 bit ROMs) with a variety of performance characteristics.

Some manufacturers are offering assemblies of chips integrated into memory modules or subsystems of up to few thousand

bits. These usually consist of a printed circuit board on which the individual array packages are mounted together with the necessary decoding, driving, sensing, clock and power distribution circuits. These units are aimed at the user who would normally opt for a small core memory, and therefore offer a good price-for-performance ratio, rather than high capacity or density.

Only two or three manufacturers have constructed systems over 10^5 bits. IBM has had a sufficiently large internal market to be able to develop a 2.5 x 10^5 bit, 60 nanosecond cycle time bipolar memory for the "cache" memory of the 360/85, and similar, though smaller, memories for the 360/195 and the new 370 series. The goal of 360/85 system is very high performance, and consequently less emphasis has been given to density (less than 200 bits/in^3) and power dissipation. Fairchild is producing a 1.5 x 10^5 bit, 200 ns bipolar memory; 64 of these units make up the mainframe of the ILLIAC IV computer being built by Burroughs.

A different application, that of high speed disc or drum replacement, is the intent of the Advanced Memory Systems' SSU mass memory. The basic building block of this system is a packaged chip containing dual 256-bit dynamic MOS shift registers operating at a shift frequency that ranges from a maximum of 1 MHz to a minimum of 10 KHz. These units are assembled by printed circuit board techniques into a basic system module of over 8 million bits. The modules can be combined to provide expansion to over 64 million bits of semiconductor storage. This represents by far the largest assembly of LSI devices into a single memory system to date. At the maximum configuration the system contains 16,384 chips, occupies 75 cubic feet, and consumes 2.3 kilowatts. The bit density is about 300 bits/in^3 and the dissipation is over 20 milliwatts/bit.

These three examples of large semiconductor systems are all based on commercial grade chips mounted in standard packages (ceramic DIP and TO-5 can), which are in turn mounted on conventional printed circuit boards. The performances in each case are superior to those of the more usual technology that they replace. But the low packaging density, relatively high power dissipation and conventional construction attest to the very conservative approach being taken to achieve these performances.

2.6.6.2 Future Trends

The most significant trend that is being forecast for the next five to ten years is the gradual invasion and eventual takeover of the computer memory field by semiconductor techniques.

The rate and extent of this change will be paced by economical rather than performance factors. Today semiconductor memories are about twice as expensive as core memories of equivalent performance. Semiconductor manufacturing costs are expected to fall to about one quarter that of core five years from now, but the price may remain at the core memory level for a while, with consequent large profit margins. Eventually, an increasing penetration of the field will bring about fiercer competition among semiconductor manufacturers, and this will force the prices down, as illustrated in Figure 2.6.16 [27].

Manufacturing improvements will result in larger chip areas and finer photo-lithographic tolerances being economically feasible. Fully decoded random access arrays in 2048 bit, 300 ns MOS and 512 bit, 80 ns bipolar forms are estimated by a conservative observer of the subject to be available in commercial quantities by 1975 [2]. Complete commercial memory systems of 10^6 bits and 500 ns full cycle time are expected to be available within five years. The larger computer manufacturers with in-house memory capability will probably be incorporating even larger and faster semiconductor memories into their products.

It is probable that computer organizations will evolve to take advantage of the greater flexibility of the semiconductor memory; memory organizations of modest capacity that combine the functions of storage with search, comparison, decision and other logical operations can be expected to appear.

As a consequence of the increasing domination by semiconductors of the commercial memory field, the smaller and more specialized market place of the military and government-sponsored semiconductor memory is expected to witness considerable improvement over today's performance and packaging capabilities. The newer technologies of complementary MOS and the variable threshold transistor can be expected to yield fully decoded arrays of over 10^4 bits on a single chip, with near-zero standby power and operating dissipation in the 10^{-6} to 10^{-7} watt/bit region. Packaging improvements will enable the densities of larger memory systems to approach 10^6 bits/in^3. Improved manufacturing, packaging and testing procedures are expected to yield reliabilities that will make large semiconductor memory systems of up to 10^8 bits feasible.

Developments such as these will require a long time, five to ten years, before they become a general capability of the industry; the main reason being that the sophisticated aerospace application that sponsors them represents a relatively limited demand, and attracts only mild competition. In any case, for these developments to take place at all depends primarily on the success of the semicodnuctor industry in fulfilling the predictions being made for it in the commerical memory market.

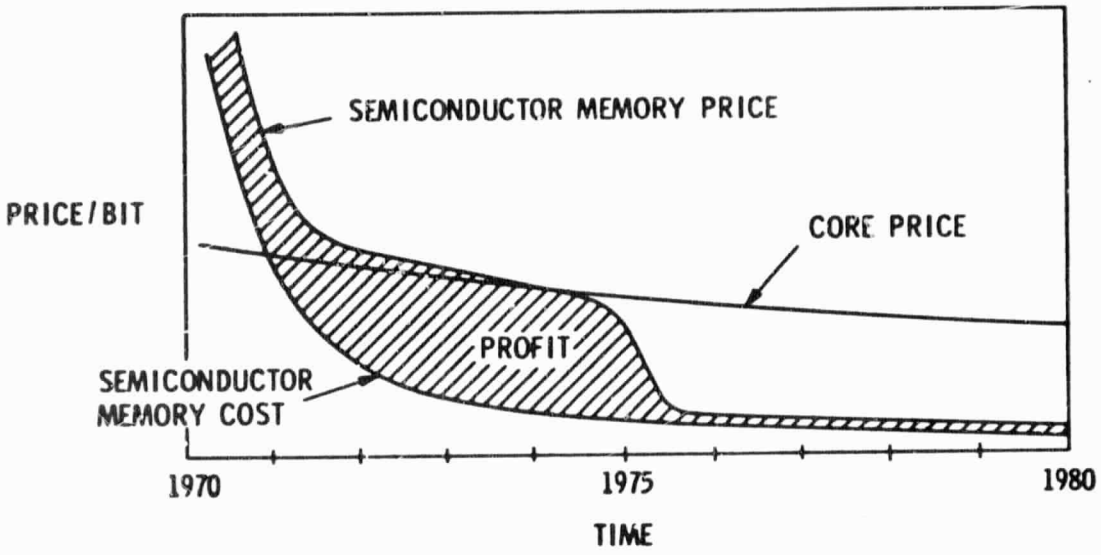


Figure 2.6.16 Cost Prediction for Semiconductor Memories

References for Section 2.6

1. Graham, R.F., "Semiconductor Memories: Evolution or Revolution?", DATAMATION, June 1969, pp. 99-104.

2. Brenner,J.W., "A Survey of Mainframe Semiconductor Memories", Computer Design, May 1970, pp. 63-71.

3. Elson, B.M., "Semiconductor Memories Evolve", Aviation Week and Space Technology, June 15, 1970, pp. 38-40.

4. Hoff, M.E., "MOS Memory and Its Application", Computer Design, June 1970, pp. 83-87.

5. Boysel,L., et al, "Random Access MOS Memory Packs Move Bits to the Chip", Electronics, 16 Feb. 1970, pp. 109-115.

6. Greene, F.S., et al, "Megabit Bipolar LSI Memory Systems", 1970 IEEE International Solid-State Circuits Conference Digest, pp. 40-41.

7. Bergh, A., et al, "Design Considerations for a High-Speed Bipolar Read-Only Memory", 1970 IEEE International Solid-State Circuits Conference Digest, pp. 66-67.

8. Lilienfeld, J.E., U.S. Patent No. 1,900,018, 1926.

9. Seely, J.L., "TTL Compatibility is Here", Electronic Design, Sept. 1969, pp. 70-73.

10. Op. cit., p. 71.

11. Faggin, F., et al, "A Faster Generation of MOS Devices With Low Thresholds is Riding the Crest of the New Wave, Silicon-Gate IC's", Electronics, Sept. 29,1969, pp. 88-94.

12. McDougall, K., et al, "Ion Implantation Offers a Bagful of Benefits for MOS", Electronics, June 22, 1970, pp. 86-90.

13. Herzog, G.B., "COS/MOS: The Best of Both Worlds", Electronics, February 17, 1969, pp. 109-113.

14. Herzog, G.B., et al, "Large-Scale Integrated Circuit Arrays", Technical Report AFAL-TR-69-352, RCA Laboratories/AF Avionics Lab, WPAFB, January 1970.

15. Burns, J.R., et al, "Silicon on Sapphire Complementary MOS Circuits for High-Speed Associative Memory", Proc. FJCC, Vol. 35, November 1969, pp. 469-477.

16. Trepp, J., Westinghouse Aerospace Division, Baltimore, Maryland, private communication, March 1970.

17. Brinton, J., "The Broken Promise of LSI: Packaging", Electronics, March 30, 1970, pp. 123-131.

18. ----------, "New Chips Pose Problems - From Wafer Processing to Hybrid Assembly", Electronic Design, Oct. 1969, pp. 68-73.

19. Hart, T.W., et al, "A Main Frame Semiconductor Memory for Fourth Generation Computers", Proc. FJCC, Vol. 35, 1969, pp. 479-488.

20. Wegener, H.A.R., et al, "The Variable Threshold Transistor, a New Electrically-alterable, Nondestructive, Read-Only, Storage Device", International Electron Device Meeting, Oct. 1967.

21. Nakanuma, S., "A Read-Only Memory Using MAS Transistors", 1970 IEEE International Solid-State Circuits Conference Digest, Feb. 1970, pp. 68-69.

22. Sewell, F.A., "Charge Storage Model for Variable Threshold FET Memory Element," Appl. Phys. Lett. Vol. 14, 15 Jan. 1969, pp. 45-47.

23. Ross, E.C., et al, "Theory of the Switching Behavior of MIS Memory Transistors", RCA Review, June 1969, pp. 366-381.

24. Calhoun, D.F., "The Pad Relocation Technique for Interconnecting LSI Arrays of Imperfect Yield", Proc. 1969 FJCC, Vol. 35, November 1969, pp. 99-109.

25. Boyle, W.S., et al, "Charge Coupled Semiconductor Devices", Bell Sys. Tech. Jour., April 1970, pp. 587-593.

26. Curtis, D., Quarterly Technical Report No. 3, Mass Memory Applications, A.D. Little/NASA-ERC Contract No. NAS12-2185, April 1970, p. 38.

27. Norman, R., "Semiconductor Memories: Make or Buy?", EEE, July 1970, pp. 42-44.

2.7 Beam Accessed Memory Technologies

In beam memories, a beam of photons or a beam of electrons is used to alter and/or sense some physical property of a storage medium. The storage medium is usually constructed in the form of a thin film, and the physical property is a magnetic, photographic, ferroelectric, etc. effect that can be readily altered or sensed.

The attraction of these memories is the extremely high recording densities that can be obtained. For example, an optical beam can be focused to a diffraction-limited spot having a diameter of the same order as the wavelength of the light. Visible light wavelengths are about half a micron, so a recording system having a spot spacing of five spot diameters would have a storage capability of 10^8 bits per square inch.

In comparison, present magnetic recording techniques in which reading is done by inductive pickoff can accommodate densities of about 10^5 bits per square inch. Technical improvement might increase this to 10^6 bits per square inch but the density is limited by signal amplitude and noise, which is in turn determined by such factors as trackwidth, track spacing, surface to head distance, and head wear.

Two different types of recording can be associated with beam memories. The first of these is bit discrete storage in which a physical location in the storage medium is reserved for each bit. The second type of storage is holographic in which a block or page of bits is stored in the form of a hologram, and no particular location within the hologram can be said to correspond to a particular bit.

Two separate and distinct properties of beams are exploited in these different types of storage. These are the ability to focus the beam to a very small spot to achieve a sufficiently high power density to record and/or read out the data at the required data rate and second, and the property of phase coherence which allows information to be stored and recovered by holographic techniques.

Both electron and photon (laser) beams can be focused to small spots having a high energy density, and memory schemes using only this property can be implemented with either type. The electron beam is easily generated, and is easily deflected by magnetic or electrostatic fields (which may or may not be an advantage); but the storage medium must be located in an evacuated enclosure, which not only creates many technical problems but makes replacement of the medium difficult. The photon beam is more difficult to generate and deflect but the storage medium need not be located in a vacuum. On balance, the advantages of the photon beam appear to outweigh its dis-

advantages, and for this reason the laser appears to be the preferred beam source for bit discrete beam memories. For holographic memory use, the laser appears to be the only practical source of a phase coherent beam. A successful effort to devise a laser design suited to a spaceborne application has been reported. A He-Ne cw laser of dimensions 16.5 x 8 x 3 inches, weight 14 lbs, power requirement of 33 W, with a 5 milliwatts output has demonstrated up to 19,000 hours continuous operation without failure [60].

The major shortcomings of the bit discrete optical memories is their vulnerability to imperfections and dust on the recording surface and their requirement for extreme accuracy of alignment (of the order of the spot size). These factors are much less critical for systems based on holographic techniques.

In holographic memories, the state of each bit is stored over the whole extent of the hologram. Scratches, dust particles, or other defects in the recording surface will reduce the signal to noise ratio, but cannot obliterate an individual bit without so destroying the entire hologram that the malfunction is easily detectable. Holograms are also much less sensitive to misalignments of the system and have the further advantage that no optical elements are required between the hologram and the readout plane [2].

Holographic memories are essentially block or page organized. A single hologram will contain a large number of bits, and illumination of the hologram will cause all of these bits to be displayed at the output plane. The memory is so organized that the output from a number of different holograms can be projected on the same output plane to allow a number of blocks to be readout.

Figure 2.7.2 shows the organization of a read only holographic memory which is being developed by Bell Telephone Labs [2,3]. The memory consists of a rectangular array of 32x32 photographic holograms each 1.2 mm in diameter and located on 2 mm centers. Each hologram contains 4096 bits which can be projected onto a 64x64 element photodetector array by deflecting the output beam of a laser. The reconstructed bits at the photoconductor array are 150 microns in diameter and lie on 200 micron centers. A major increase in the number of bits in a hologram appears to be on the horizon. Researchers in Japan are working on a 201x201 element detector array on an 8mm x 8mm substrate which would make 40,000-bit holograms feasible.

If a hologram having appreciable thickness is substituted for the thin hologram of the system discussed above, an additional addressing technique comes into play. The readout of the hologram now depends strongly on the angle of incidence of the readout beam, and multiple holograms can be stored in the same volume of storage medium by varying this angle. This technique

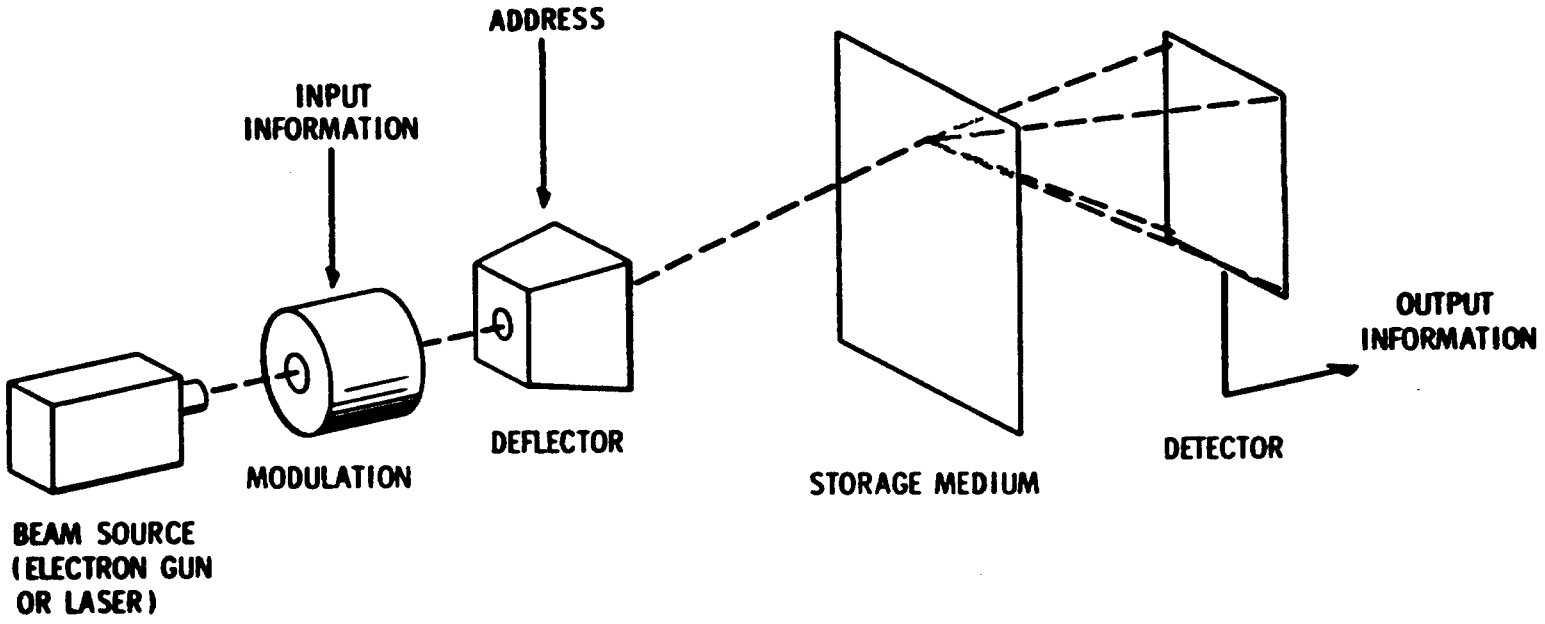


Figure 2.7.1 Elements of An Optical Memory System

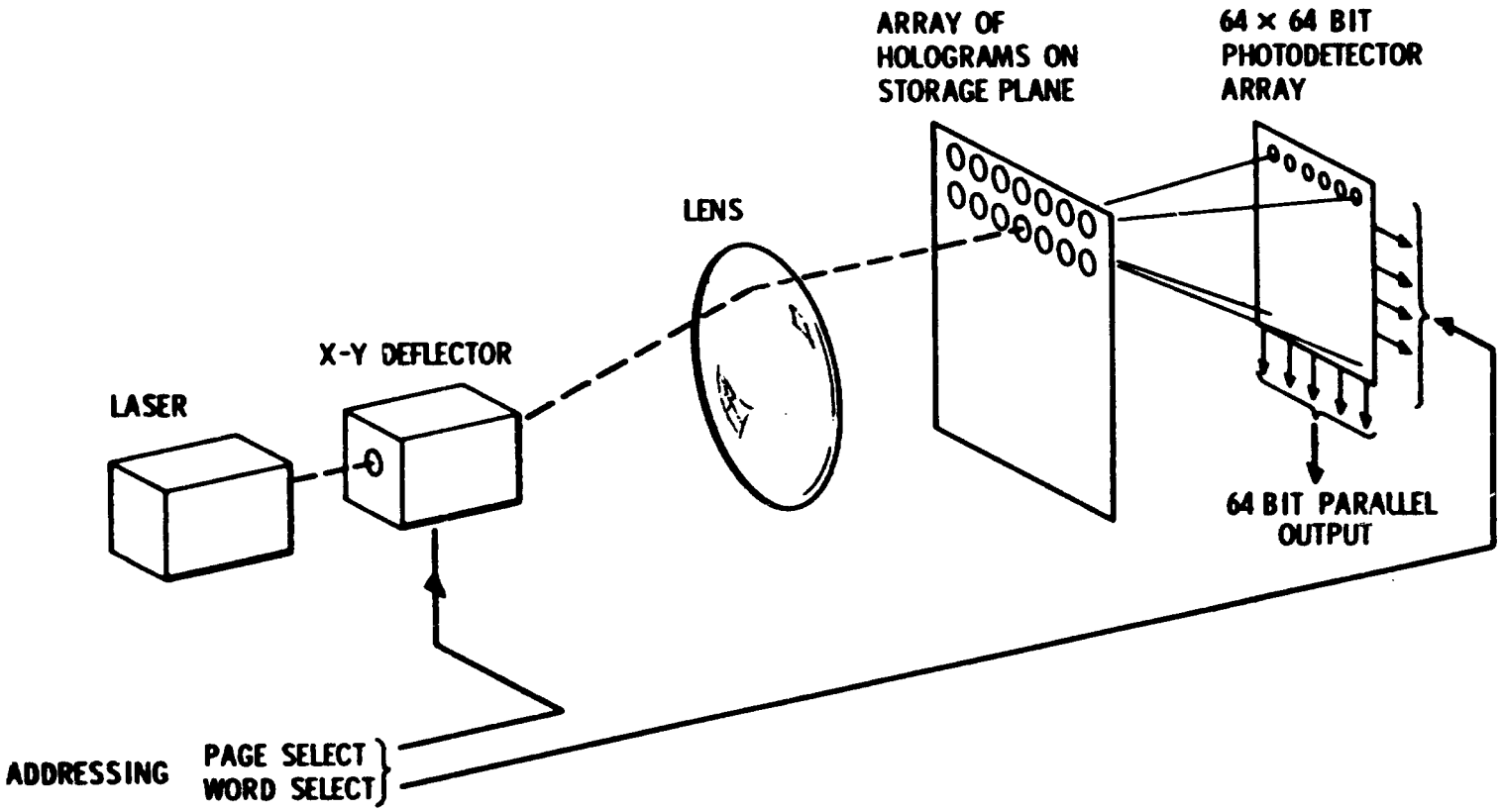


Figure 2.7.2 Read-Only Holographic Memory System

is being investigated by a number of researchers [4,5,6,7,8].

Optical and beam memories can be categorized by the functions - writing, reading, erasing, and rewriting - that they are able to perform. The most elementary memory would be the read only memory. This memory would not have the modulator functions shown in Figure 2.7.1. Instead, information is incorporated in the storage medium before it is placed in the memory and all the memory can do is read it out.

A more complicated form of memory is the write-once memory. These memories have all of the components shown in Figure 2.7.1, but the storage medium is such that the information cannot be erased once it is written in. Typically, the information can be read out as often as desired.

While memories of this type are obviously less flexible than memories where the data can be erased and rewritten at any time, it is still possible to change data after it has been stored. One means of doing this is to store an index to the memory in a separate small random access memory. When it is desired to delete or modify a record in the memory, the record data is pulled out, the data modification made, and the entire record rewritten at a new vacant location on the recording medium. The new address is then placed in the index in place of the obsolete address, and this data item thereafter accessed at the new location in the mass store. In bit discrete memories another scheme can be used where a data record has continuing transactions which do not obsolete the remainder of the record. At the time of initially placing this record into the memory, blank spaces may be left, and a local address stored in the index. Then when data is to be added to a record this local address is read from the index, incremented by one and returned, and the new data is written into the blank spaces in the mass memory. This may be repeated until the local address regions initially allocated to this record are all filled. Then, either a transfer address to a new mass store track may be utilized for continuation of the record, or the procedure of completely transferring the record to a new location in the mass file may be used. For some applications, arrangements such as these are adequate.

The third category of memories would have complete write, read, and rewrite capability. This is the category of memory that is really required for use in such applications as this study is directed towards. Much research and development effort has gone into producing such a memory but to date no complete system has been successfully demonstrated.

A fourth possible category would be the write only memory. This would be a memory in which data could be recorded but not read back. The recording medium would be removed and read out elsewhere.

Such a memory would be used for such applications as recording experimental data for later reduction. Some memory systems currently in existence almost fall into this category. An example is a photographic recorder produced by Synergistics. This records 36 channels of data on an 8 mm photographic film and is designed as a replacement for experimental data tape recorders. Although recording and readout are performed in the same unit, the film must be removed from the recorder and processed before readout can be accomplished.

2.7.2 Beam Memory Storage Media

2.7.2.1 Photographic

Optical memory systems employing photographic techniques for read-only storage were built by Bendix, IBM and others as long ago as 1962 [9]. This form of storage has many advantages: photographic emulsions are sensitive, with a typical exposure energy requirement for microfilm of 20 to 100 $x10^6$ J/cm^2 [10], and they are useable for both bit discrete and holographic storage. Recording densities of greater than 10^9 bits per square inch have been achieved [1]. The Bell Telephone Labs holographic memory described above used a photographic store.

Photographic memory stores are permanent, which can be an advantage or disadvantage depending on the application. There is also a frequently inconvenient delay necessitated by the requirement for development of the latent image, so the information is not available for immediate use, and the handling associated with the development process can cause damage to the film and is undesirable in most operating memory systems. The two principal emulsions used for photographic memories are silver halide and dichromated gelatin. The silver halide emulsions are faster, more stable, and more permanent than dichromated gelatin, but the dichromated gelatin has a higher diffraction efficiency for holographic work [3].

Among the photographic memories that have been developed are the Bendix Photostore, a Photographic Plate Memory by International Computers and Tabulators, and the 36 channel tape recorder replacement by Synergistics.

2.7.2.2 Vaporization Storage

The development of the laser, with its ability to generate high energy densities in very small areas has made possible the use of recording techniques that require more recording energy than photographic film. One of these, which produces a permanent record similar to photographic film but does not require chemical developing, is the vaporization of

a thin film of opaque material on a suitable substrate.

The desirable characteristics for the material of such a film are a low thermal conductivity, to permit high resolution recording, and a low boiling point to keep the energy requirements low. Bismuth is one material which exhibits these characteristics. Successful recordings of thin holograms on Bi films 75 to 200 Å thick on a glass substrate have been reported [11]. These holograms had resolutions of nearly 100 lines per mm and nearly 6% efficiency which is very close to the theoretical maximum for this type of hologram. The recording energy was less than 50 $x10^3$ Joules/cm^2 for a 100 Å film.

The vaporization technique can also be used for bit discrete memories. The use of carbon films in a bit discrete memory was studied in some detail by Cavley [12]. At least two commercial bit discrete memory systems are now in production. These are the Unicon Laser Recording System and the Foto-Mem FM 390, both of which are described in a later section.

2.7.2.3 Photochromic Storage

Photochromic materials are materials which change color when exposed to light of an appropriate wavelength. In some, the color change is permanent, but in others it is reversible. These materials can be used as an erasable optical memory storage medium [13]. In general, illumination of a reversible photochromic material by light of a short wavelength will cause the material's absorption spectrum to shift toward longer wavelengths, while illumination with long wavelength light will cause the absorption spectrum to shift back towards shorter wavelengths. Thus a material that is transparent to visible light can be darkened by exposure to ultraviolet light and bleached by visible or infrared light [1].

Photochromic materials suffer from thermal fatigue and the organic photochromics exhibit a fatigue or reduced sensitivity after cycling.

Photochromic materials tend to be less sensitive to light than photographic emulsions, but since the light sensitive effects are at the atomic level, they have a higher resolution and do not exhibit the graininess of silver halide emulsions [13,14].

Among the photochromic materials that have been investigated for memory use are strontium titanate ($SrTiO_3$) doped with Fe, Fe/Mo, Ni/Mo, or Co/Mo and calcium flouride (CaF_2) doped with La, Ce, Gd, or Tb [3,4,13], KBr [8], and silver halide crystallites in borosilicate glass [15].

Most of the work that has been done with photochromic memories has been directed towards either thick or thin hologram storage, although there does not appear to be any reason why bit discrete recording could not be accomplished if desired. The techniques used vary. Some researchers start with the material fully bleached and darken it to write information while others start with a colored material and bleach to write [14]. Writing is accomplished using one laser source and erasing uses a second laser. Readout might use a third laser [3,15] or the write laser at reduced power [1]. Other combinations are also possible [8].

The energy required for photochromic recording is one to three orders of magnitude greater than that required for photographic recording. The energy required to expose KBr has been reported as 1×10^{-3} J/cm^2 [8], and CaF_2 or $SrTiO_3$ require 50×10^{-3} J/cm^2.

2.7.2.4 Ferroelectric Storage

In the photochromic materials, the absorption of the material is modified by the action of writing. This is an amplitude process. In another class of materials, the index of refraction can be altered. These materials are called phase materials, and consist of ferroelectric crystals such as lithium niobate ($LiNbO_3$) and barium titanate ($BaTiO_3$) [3].

The mechanism of the effect appears to be the photo-excitation of electrons from the valence band to the conduction band where they are then free to drift and change the index of refraction of the material through the electrooptic effect. Erasure of the image is accomplished by heating the material or illuminating it with short wavelength light. The image has a finite lifetime which is a strong function of temperature. In $BaTiO_3$, at room temperature it is about 0.3 sec, but in $LiNbO_3$ it is several weeks [3].

The resolution of the materials appears to be good. Chen, et al [6] reported a limiting resolution in excess of 4000 lines per mm in $LiNbO_3$. The required writing powers reported to date all appear to be higher than those required for photochromic materials, although further research may turn up a lower power material. $BaTiO_3$ requires 0.1 Joules/cm^2 [16]. Thaxter reported using as much as 1.2 Joules/cm^2 to get an image in $Sr_{0.75}Ba_{0.25}Nb_2O_6$ [17] (this material also requires a DC electric field of several KV/cm before an image can be recorded). $LiNbO_3$ requires 1 w/cm^2 for 100 sec to obtain a maximum image [6]. This corresponds to 100 Joules/cm^2. However, the material was observed to be highly nonlinear: the image strength was proportional to the square root of the light intensity, a point to be kept in mind when attempting to reduce the write time

by increasing the intensity.

Electron beam writing on ferroelectric material has also been demonstrated [18]. The material was $Bi_4Ti_3O_{12}$. Resolution of the order of 1 to 2 microns was reported and the domains were observed by transmitting polarized light through the material.

2.7.2.5 Magnetooptic Recording

The storage of information by altering the magnetic state of a material has been practiced successfully for years. Obvious examples of such storage are ferrite cores and magnetic tape. The major drawbacks of these classic storage methods for the very large memory applications are the technical and economic factors that are a result of the limitations on minimum core size, minimum recording head gap, etc.

The development of optical techniques for creating magnetic domains having dimensions of the order of a micron makes possible the magnetic recording of information at much higher densities. These techniques involve the use of a laser to heat a thin film of magnetic material under conditions that cause a change in the magnetic state of the material. The two most commonly proposed methods of accomplishing this change of state are Curie Point writing and Compensation Point writing.

Curie Point writing makes use of the property of ferromagnetic materials whereby they gradually lose their magnetization as the temperature is increased until, at the Curie temperature, they lose their ferromagentic properties and become paramagnetic. The transition is quite sharp, and is reversible. A thin film of the material is uniformly magnetized normal to the surface of the film. A laser is then used to heat a spot on the film to above the Curie temperature. The material in the spot loses its magnetization, and on cooling below the Curie temperature the material is remagnetized, but in the opposite direction, by the action of the magnetic field of the surrounding unheated material [19,20], or by an external field [21].

Compensation Point writing makes use of the property that in some materials the net magnetization of the material is the difference between two strong sublattice magnetizations which are temperature dependent. At the compensation point, these two magnetizations are equal and the net magnetization is zero [22]. This causes a sharp peak in the external magnetic field that must be applied in order to alter the magnetic domains in the material.

In a memory application, a thin film of the material is magnetized in a known direction and held at the compensation temperature. A laser is used to heat a spot on the film a few degrees above the compensation temperature. An external magnetic field of the proper strength can then be applied to the film to cause only the heated portion of the film to switch.

Readout of the state of a spot on one of these films is accomplished by examining the film under polarized light. If the material is transparent, the Faraday effect can be used. This effect is the rotation of the plane of polarization of the light on passing through the film as a function of the magnetic state of the film. If the material is not transparent, the Kerr effect can still be used. This is a similar rotation of the phase of polarization of the reflected wave, but is a weaker effect.

Numerous materials have been proposed for use in magneto-optic memory systems. For Curie point writing, MnBi is the most common suggestion. This material has a large Faraday rotation of 5×10^5 deg/cm [23] for ease of readout, but its Curie temperature of 350°C implies a large temperature change for writing. This is partially offset by its large absorption coefficient, so that the calculated energy to write a 3 micron diameter spot is 10^{-9} Joules. This could be accomplished, for example, with a 50 ns pulse from a 20 mw laser (assuming no losses) [22].

Europium oxide (EuO) is an attractive material for magneto-optic systems because of its even lower write energy requirements of 10^{-10} Joules for the same size spot. Unfortunately the memory must operate on the range of 20°K to 77°K [21], making continuous refrigeration necessary.

Other Curie point materials that have been suggested include MnAs [24]; CrTe [25]; $Fe_{1.67}Ge$, $Mn_{(1.5+x)}S [0 \leq x \leq 0.3]$, and $La_{0.7}Sr_{0.3}MnO_3$ [26]; and hexagonal Co [22,27]. Research into these materials is aimed at realizing a low Curie temperature, specific heat and thermal conductivity, to minimize the heating power requirement, and a high magnetooptic effect to provide a strong, noise-free readout capability, while avoiding special environmental conditions such as cryogenic temperatures.

Most of the experimental work with compensation point writing has been done with Gadolinium Iron Garnet (GdIG) because its compensation temperature is close to room temperature. Sometimes the GdIG is doped with some element such as Al to modify the compensation temperature [7].

The minimum domain size for compensation point memories using GdIG appears to be about 100 microns in flat crystals

and about 30 microns in crystals that have been scribed to create predetermined domain boundaries [28]. This compares to domains of less than 1 micron diameter in MnBi. The writing power required for a compensation point memory is dependent on the strength of the magnetic field that is used for switching the heated material and the thickness of the material [29]. Goldberg [28] has estimated that a GdIG compensation point memory having a 1 μsec cycle time would require a 100 mw laser. The packing density would be 10^6 bits/cm^2 on a 10 micron wafer. MacDonald has given a writing energy of 125.10^{-3} Joules/cm^2 for a 1 micron thick GDIG film [29]. This high value appears to be due to the large temperature change (80°C) required by his mechanization.

Many other techniques of changing the magnetic state of a material have been investigated and more will probably be investigated in the future. Kump and Chung [30] studied the magnetization of permalloy by applying a magnetic field and inducing a strain in the permalloy by heating it with a laser beam. Others have studied more complicated structures, such as multiple layer films consisting of Pd-Co over Permalloy[31], or Permalloy over Ni-Co [32].

The signal to noise ratio for the readout has been given considerable thought also. Various signal enhancement schemes and modulation methods have been studied in an effort to obtain better memory operation [33,34]. Most studies of magnetooptic memories have been directed towards memories in which each bit is individually written on the film by the laser. Some work has also been done on holographic recording of arrays or pages of bits [3,35]. These schemes of necessity require more energy per record because more area is being heated (and more information recorded) at a time. Rajchman [35] has proposed a memory which would store 10^4 bits in a hologram of 0.5 mm^2 area using a 10 KW pulse power of 15 to 20 nanoseconds duration. The storage medium was to be a 700 Å MnBi film. The holographic scheme would eliminate the problem caused by micron sized flaws and dust particles in the recording film.

2.7.2.6 Thermoplastic Storage

Another recording medium, which has some very interesting advantages for optical memory use, is the thermoplastic surface. The essence of this memory system is that the surface of the thermoplastic is deformed by some means to record data, and the deformation is removed to erase it.

While this technique could probably be adapted to bit by bit storage, the work reported to data has been directed towards holographic storage of pages of data. One method of accomplishing this, which has been reported in some detail [3,36], is a xerographic process in which a photoconductor and a thermoplastic layer are deposited on a glass substrate. To make a

recording, an electric charge is placed on the surface of the thermoplastic, and the film is then exposed holographically. The light striking the photoconductive layer generates hole-electron pairs and the electrons drift to the photoconductor-thermoplastic interface. A second charging step is performed and then the thermoplastic is heated. The surface of the thermoplastic deforms under the effects of electrostatic attraction to form a relief hologram. Erasure is accomplished by heating the thermoplastic in the absence of a charge pattern.

Thermoplastic recordings are reported to have less optical noise than silver halide photographic holograms [36] and to require exposures in the range of 4.10^{-3} Joules per cm^2 [3], which is less than that required for photochromic or ferroelectric recordings. The drawbacks that have been noted are a tendency for old images to reappear [9] and an uncertainty in the number of times the films can be reused [3,9,37].

2.7.2.7 Photoconductive Memory Schemes

Several photoconductive film memory schemes have been proposed [38,39]. These all involve a multiple layer film structure in which a material having multiple stable conduction states is placed under a photoconductive layer and a transparent electrode. This multiple-layer structure combines the functions of the modulator and/or the detector with that of the storage medium. A laser is used to address a portion of the lower film by making the photoconductive layer conduct, and the state of the lower film is altered (for write) or sensed (for readout) by way of the transparent electrode. Very little work on these schemes beyond the idea stage has been reported.

2.7.2.8 Other Approaches

There has also been some work directed towards obtaining a direct interaction between a laser beam and a magnetic storage film [40] or a ferroelectric material [41] without the intervening thermal effects described earlier in this section. Very little published information on these schemes has been found. Although the work done to date has been only exploratory it appears to be interesting and should be pursued further if the physics of the materials in question is such that the write energy can be kept down to reasonable levels, such as 10^{-2} J/cm^2.

2.7.3 Auxiliary Equipment

The actual storage medium in an optical memory is a negligible part of the total hardware that is required to produce an operating memory system. The majority of the memory

	Write Energy J/cm^2	Density bit/in^2	Notes
Photographic			
Silver halide	$20\text{-}100 \times 10^{-6}$	$>10^9$	requires image developing, can not be erased, permanent
Dichromated gel	3×10^{-3}	$>10^9$	
Vaporization (Bi)	50×10^{-3}	6×10^8	can not be erased, permanent
Photochromic ($SrTiO_3$)	50×10^{-3}	$>10^9$	erasable, finite retention time, requires multiple wavelength beams
Ferroelectric ($LiNbO_3$)	100	1.5×10^9	erasable, finite retention time, requires high write energy
Magnetooptic			
Curie Pt. (MnBi)	$15\text{-}30 \times 10^{-3}$	$>6 \times 10^8$	erasable, does not decay
Comp. Pt. (GdIG)	$<125 \times 10^{-3}$	1.5×10^6	erasable, does not decay, temperature sensitive
Thermoplastic	4×10^{-3}	N/A	limited life films, incomplete erasure
Photoconductive	N/A	N/A	

Table 2.7.1 Characteristics of Recording Media

consists of the auxiliary equipment required to generate and deflect the read and write beams, modulate the write beam with the information to be recorded, and process the information being read out of the memory. The following sections describe the hardware required for these auxiliary functions.

2.7.3.1 Read and Write Beams

Two distinct properties of beams are used by the bit discrete and the holographic storage schemes. In the bit discrete schemes, data is stored at discrete points in the storage medium, and the property of interest is the ability to deflect and focus the beam accurately to these points. For most of the storage media, the write process requires that the beam supply appreciable energy to the spot. For this mode of operation, both light (laser) beams and electron beams [35] have been proposed. The laser beams are easy to focus, but the deflection is difficult and the laser for generating the beam is bulky and expensive, especially when higher powers are required. The electron beam, on the other hand, is easily generated and deflected, but the system must be operated in an evacuated environment, which makes changing of the memory element inconvenient. Furthermore, very few of the storage mechanisms interact well with an electron beam for readout, so a separate means of readout must be provided.

An interesting scheme for a dual beam readout, using both a light beam and an electron beam has been suggested [34,42,43]. This scheme uses the electron beam to address the bit and the optical beam for the readout. The electron beam is amplitude modulated at a convenient frequency and this modulates the temperature of the magnetic storage material at the location of the bit. This in turn modulates a diffracted and rotated light beam in such a way that the signal is easily recovered. The advantage of this scheme is that the light beam does not have to be tightly focused or accurately aimed at the bit, and the output signal to noise ratio is improved.

For the holographic memories, the property in the beam that is of particular interest is the coherent character of laser light. This is the property that makes holographic recording and readout possible. It is a property of the output of a laser,and is not shared by electron beams.

2.7.3.2 Addressing Schemes

One of the major problems to be overcome in the design of an optical beam memory is the provision of means for causing the beam to reach the desired location in the storage medium.

a) Mechanical Deflection. The earliest, and still most commonly used, method of addressing different locations is through mechanical motion, using arrangements such as tapes, drums, disks, scanners, etc. which are the counterparts of the methods used in magnetic recording. Examples of these devices would be the rotating disk and drum memories developed by Bendix in the early 60's [9], the rotating mirror scanner [44], and the rotating head scanners for use with optical tape storage systems [45]. While these systems are capable of much higher recording densities and total storage capacities than are available from conventional magnetic recording systems, and data transfer rates of well over 10^6 bit per second are possible, these systems suffer from most of the same problems of slow access, narrow bandwidth, and mechanical difficulties as their magnetic recording counterparts.

Wider bandwidth mechanical systems have been developed using piezoelectric effects. One system attributed to Schlafer and Fowler [46] uses a piezoelectric transducer in a shear mode to tilt a mirror. Deflections of 0.05 degree and bandwidths of 17 KHz for a single mirror are reported. Another scheme, shown in Figure 2.7.3 uses a piezoelectric element to change the geometry of a prism by causing a plate to contact the reflecting surface of the prism [9]. The element is essentially digital in nature and switching times of <35 μsec are anticipated.

b) Acoustooptic Deflection. Deflection of a light beam by non-mechanical means has been achieved by several methods. The most commonly used of these are the acousto-optic and the electrooptic devices.

The acoustooptic deflector makes use of the periodic change in index of refraction that can be produced in many materials by an acoustic wave. The period of the change in index of refraction is the same as the wavelength of the acoustic wave, and this periodic change in index of refraction produces a grating which can deflect light by diffraction as shown in Figure 2.7.4. The acoustic wave is generated by a transducer whose frequency controls the grating spacing, and thereby the deflection [35].

One common compound which exhibits this property is water. Dual axis deflection arrays using water and having 16x16 deflection positions are reported as typical [35]. Such an array would have an access time of 6 μsec, a transmission efficiency of 80% and an input power requirement of 1 watt. Unfortunately, the sound absorption in water limits the frequencies that can be used to the order of 30 MHz, which produces very small deflections.

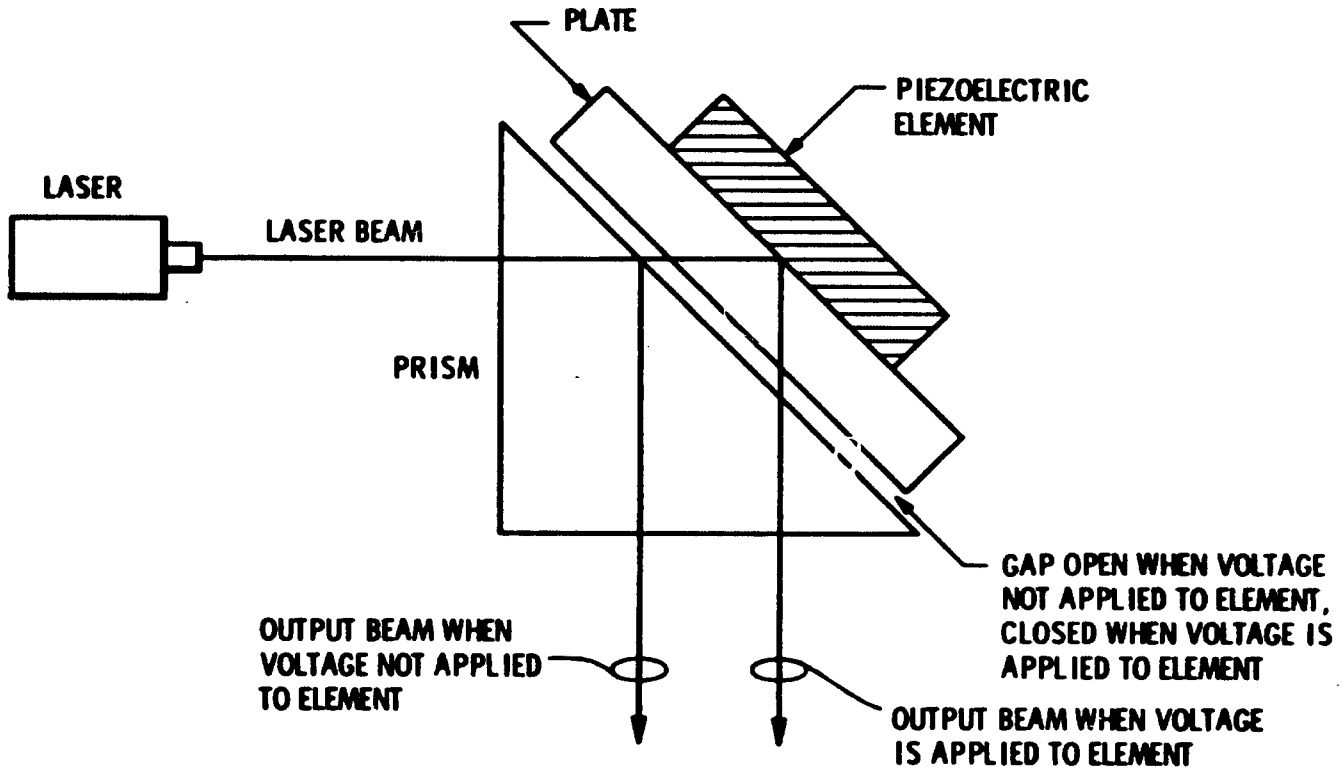


Figure 2.7.3 Mechanical Beam Deflection System

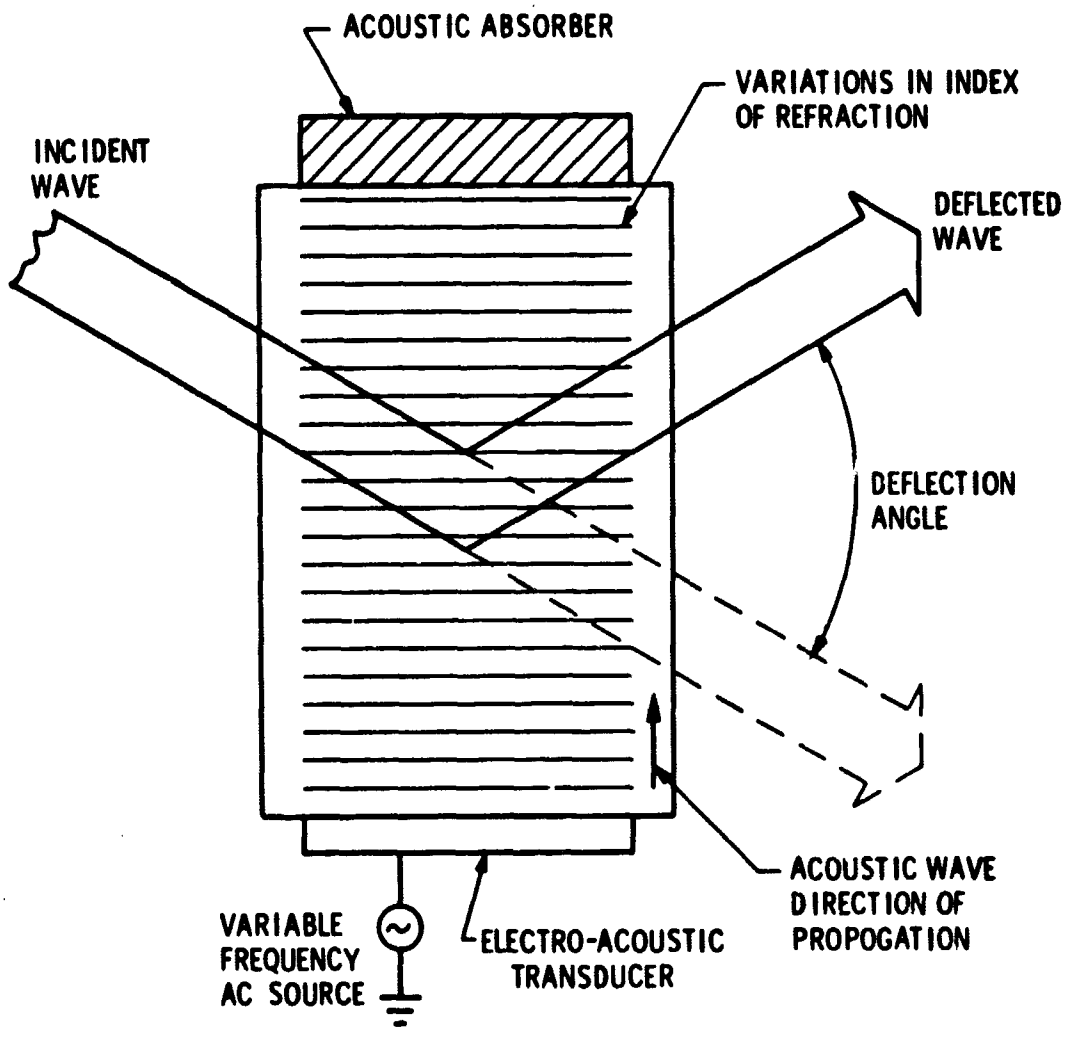


Figure 2.7.4 Acoustooptic Deflection System

If solid alpha iodic acid (α-HIO_3) or crystalline lead molybdate ($PbMoO_4$) are substituted, frequencies up to 1 GHz can be used and deflections of 7° have been achieved. Pinnow, et al, have reported achieving a 2 μsec access time for a 64x64 deflection array using $PbMoO_4$ [35].

The photoelastic effect, which makes acoustooptic deflection possible, occurs in all states of matter and in particular in crystalline media of all symmetry classes. Thus the range of possible materials for acoustooptic applications is virtually unlimited [47]. Among the other materials that have been studied for acoustooptic deflection systems are TeO_2 [48] and $PbWO_4$ [49].

c) Index of Refraction Gradient Deflection. A light beam will be deflected on passing through a material when there is gradient in the index of refraction in the material perpendicular to the direction of propagation. Control of this gradient can be used to control the deflection. Experiments which accomplish this using the quadratic electrooptic effect and other means have been reported [46]. Boer [50] has reported an experimental system capable of resolving 100 positions in a two axis setup using thermally induced gradients near heater wires suspended in Dow Corning Silgard 15 Dielectric Gel. The response time of this system was not reported.

d) Electrooptic Deflection. The electrooptic deflection technique is based on the use of birefringent materials, such as $CaCo_3$, $NaNO_3$, or KH_2PO_4 [51]. These materials deflect light that is polarized in one direction relative to the crystalline axes of the material differently from light that is polarized in a perpendicular direction. By proper design of the elements, parallel outputs displaced by a prescribed separation can be generated.

An electrooptic switch preceding the element can then be used to select between the two output beams. This switch makes use of the longitudinal electrooptic effect of potassium dideuterium phosphate (KD*P) crystals (Pockels effect). The Pockels switch requires light to be transmitted throught the electrodes. It would be preferable to use the transverse electrooptic effect but large, good quality crystals showing this effect are not yet available. In either case, the switch has the ability to allow polarized light to pass through unaltered or with the plane of polarization rotated by 90°, depending on whether the voltage applied to electrodes on the crystal is zero or the half wave voltage $V_{\lambda/2}$ [51].

A number of these switch and birefringent crystal pairs can be arrayed in series to obtain multiple deflections.

An electrooptic deflection system has been reported [52] which has six binary deflections or a total of 64 positions in each of two directions to give 4096 accessible locations. Two additional deflections allow any of four such arrays to be interrogated, giving a total of 16384 locations. The optical transmission of this system is in excess of 40% and the access time less than 2 μsec.

Kulcke, et al [51], have proposed a system which would have $2^{10}x2^{10}$, or over 10^6, positions, and other authors [35,46] have stated that deflection systems of this size should be available in the foreseeable future.

The access times of these birefringent deflection methods are determined by the switching speed of the polarization switches, which is in turn ultimately limited by the power dissipation in the electrooptic crystals. Presently available electrooptic switches appear to be limited to about 10^6 deflections per second [46].

When many deflection stages are placed optically in series to obtain a large number of output addresses, then the degree of aberration and background light generated in each stage becomes important [51]. This can cause an unacceptable amount of light in the undesired output positions. Other problems that must be considered are 1) optical distortion of the phase front in passing through the deflection system, 2) optical losses, largely at the transparent electrodes of the KDP electrooptic switches, 3) variations in deflection with wavelength, especially with systems requiring more than one wavelength for recording, erasing, etc., and 4) the size and uniformity of elements required by the displacement of the beam in the later stages of a series system [46,51].

e) Magnetooptic Deflection. The magnetooptic system is similar to the electrooptic system except for the substitution of magnetooptic polarization rotation, using such effects as the Faraday Effect, for the switching element [53].

f) Laser Internal Deflection. All of the previously mentioned deflection systems involve deflection of the beam outside of the laser cavity. There are also methods for controlling the direction of the beam emitted by a laser. In one such system [54] a crossed array of electrooptic switches, a linear polarizer and a birefringent bias plate within the laser cavity are used to select one of the many possible modes in which the laser could oscillate.

The same author postulated a system having about 7400 possible output beams and an output of 1 watt CW for a 2.5 KW Kr lamp excitation. The switching speed would be in excess

of 4 MHz. A number of other, similar, switchable laser schemes have been reported [55].

2.7.3.3 Modulators

All of the optical memory schemes which have the ability to record as well as read require some means of modulating the intensity of the light beam during the record process. The severity of the problems associated with this modulation are a direct function of the number of bits that must be recorded simultaneously. For memories which record data a single bit at a time in serial fashion, or a few bits at a time in parallel-serial, the modulator is relatively easy to arrange since both the electrooptic and magnetooptic effects can be applied as modulators. The acoustooptical effect can also be used for modulation under certain circumstances, by a scattering type of interaction [56] although this effect may not be of much use in optical memory work. And there are other effects, such as the change in scattering as a function of applied electric field in birefringent ferroelectric ceramics [56,57] (this effect is really another manifestation of electrooptic interaction). In some cases, just modulating the input power to the laser is sufficient.

When large numbers of bits must be recorded simultaneously, the problem is much more difficult. This is the case with the page organized bit discrete memories and with all of the holographic memories. Here the whole page of bits must be read onto some form of a page composer which is then used to modulate the light beam (laser beam) in "parallel". A page composer could be made up of an array of electrooptic or magnetooptic elements, and such an array of 10^6 electrooptic elements has been proposed [8].

One of the few authors to really discuss this problem [35] has suggested the use of recently reported liquid materials called nematic materials which have some properties of crystals. These materials contain rodlike molecules which are normally in parallel alignment making the material transparent. Upon application of an electric field, the molecules ionize and slide past one another, and the resulting turbulence causes variations in the local index of refraction which results in scattering of light [58]. The time constants of this effect are however, of the order of milliseconds.

2.7.3.4 Readout

Readout from most of the optical beam memory systems is accomplished by illuminating the storage medium and detecting some change in the beam as it is either transmitted through or reflected from the storage medium. The exceptions to this statement are primarily the photoconductive memory systems

which use a photoconductive layer and a storage medium located between a pair of electrodes. In these cases the readout system is integrated with the storage medium and the light beam is used only for addressing. These memories were discussed briefly earlier, in the section on storage mediums, and will not be considered farther here.

The effects that the storage medium can have on the beam are attenuation, change in polarization, or in the case of holographic schemes, deflection or phase shift. In each case, these changes are eventually converted into an amplitude image of some kind of a photodetector.

As in the case of modulators, the difficulties associated with the readout are directly related to the number of bits that must be read simultaneously. When a single bit or a few bits are read out of a bit discrete memory, there are comparatively few problems in obtaining a fast reading at a high signal to noise ratio with very low optical input power. A representative analysis is that by Chen, et al [19] for readout from an MnBi film by the Faraday rotation method. They assumed a 1 MW laser and allowed 10 db for losses in the deflection system and 30 db in the MnBi film. They concluded that the resulting 10^{11} photons per second arriving at the detector would give a signal to noise ratio of 30 at a readout rate of 10^9 bits/sec.

With page readout systems such as holographic memories, the optical input power required to obtain a given output bit rate may or may not be greater than that required for bit by bit readout. The optical energy is divided among the bits in the page and the bandwidth of each detector can be reduced by the same factor. The determining factor will be the time over which the photodetector can integrate the rate of arriving photons. In any case, the energy required to read out will be less than the energy required to write.

There are serious practical problems associated with the design of a readout device which can receive a page containing many bits of data and deliver it in a suitable format to the output bus of the memory. Only recently has the development of LSI techniques made possible the fabrication of complete readout arrays on a single chip. RCA [59] has built a 256 element detector having a photodiode and two MOS FETs in each element. The output is in the form of 16 words of 16 bits each. The whole array including the word select lines and digit output lines is on a single chip 0.76x0.76 cm and having 65 electrical leads.

Bell Telephone Labs has an array built up of 64 eight by eight bit subarrays mounted on a glass substrate for a total of 4096 bits. And it has been reported [9] that a Japanese

research team has developed a 201x201 bit array on an 8 mm square silicon chip to give over 40,000 bits.

When it is desired to change any of the information in one of the pages or blocks, the contents of the whole block must be accessed and read out by the detector. When the desired changes have been incorporated, the whole block must then be transferred to the page composer to allow it to be re-recorded. This might involve the transfer of 10^4 to 10^6 bits just to change one bit. Rajchman [35] has proposed an interesting mechanization for the memory in which the detector and page composer are combined into a single LSI structure called a LATRIX (for Light Accessible Transfer Matrix). A LATRIX is an array of elements, each of which has three basic components:

1) a memory cell, such as a bistable flip-flop,

2) a photodetector, which can be controlled to set the flip-flop (for the read process), and

3) a light valve which is controlled by the flip-flop (for the write process).

Transfer of a page from the detector to the composer takes place in a bit-parallel fashion, obviating the need for external buffers and/or slow serial techniques. Output and input can be performed in a random access fashion. The LATRIX acts as a small random access semiconductor buffer memory. In this mechanization, the LATRIX is placed between the deflector and the storage medium. For writing, the deflected beam is modulated by the page composer and reacts with a reference beam at the storage location to form the hologram of the page. For reading, a read reference beam is deflected to the desired storage location.

Using this scheme, a hologram containing a block of data is read out to the LATRIX. The particular bits that are to be read out or changed are then read in or out of the LATRIX as with a random access semiconductor memory. The block is then recorded back onto the storage medium if the new data is to be saved.

Schemes such as this LATRIX offer the possibility of greatly reduced bit handling at the expense of a much more complicated composer-detector combination. The reduction of such schemes to practice is probably many years into the future.

2.7.3.5 Available Systems

A number of read-only photographic memories were developed during the 1960's. Several of these were mentioned earlier. They are based on optical sensing from drums, disks, or plates in

in much the same way as magnetic recording is sensed off a drum, disk, or tape [9].

In the last year, several optical memory systems having a write-once capability have come on the market. These are the Unicon Laser Recording System manufactured by Precision Instrument Company, the Foto-Mem FM390 Memory System, and the Synergistics PDR-5 Laser Recorder. None of these systems is directly adaptable to the requirements of a spaceborne mass memory, primarily because of their inability to erase and rewrite data.

The Unicon System records data in the form of holes burned into a metallic thin film on a 7 mil polyester base by a laser. Thus the memory falls into the vaporization storage classification. Readout is accomplished by detecting the reflection from the metallic film.

An area of three by eight microns is allowed for each bit so the storage density is about 41,600 bits per mm^2 or 2.8×10^7 bits per in^2. The data is recorded on strips 31.25 inches long having a recorded width of 3.5 inches. The capacity of each strip is 2.9×10^9 bits. 400 strips are used to achieve total memory capacity of more than 10^{12} bits.

During a read or write operation, the strip is mounted on a spinning drum and the tracks on the strip are accessed by a moving optical system. Access times of 200 msec within a strip and 5 sec strip to strip are claimed. The Unicon 690 Laser Mass Memory System is currently in production and at least two installations are in progress.

The Foto-Mem FM390 is also a vaporization storage system employing a laser for both reading and writing, although it differs in detail from the Unicon System. The data is stored in a circular area on a 4"x6" coated celluloid card. The reason for this particular form factor is historical, since it allows the cards to be handled by equipment designed to be used with 4"x6" microfiche cards, which are also manufactured by the same company.

Each card holds up to 5×10^6 bits and 100 such cards form a data cell. A maximum of 500 such cells can be used to give a memory of 2.5×10^{11} bits capacity.

The cards are spun during recording and readout, and have an access time of 50 msec for data on a card that is already in position. Maximum access time for information that is on a card in a cell that must be fetched is about 9 sec.

Accessing of the data on a card is accomplished by the rotation of the card and track selection. The latter function makes use of a deflection scheme that has been described as a

digital electrooptic system, but the manufacturer would not give any more information.

Readout is accomplished by measuring the light transmitted through the film instead of the reflected light as in the case of the Unicon System. For both read and write, the data rate is 800,000 bits per sec in 8 bit parallel bytes.

The Synergistics PDR-5 Laser Recorder is a 36 channel photographic system intended as a replacement for a magnetic data tape recorder. Direct read back of the information is not possible since the film must first be processed. Information is stored in parallel on 36 tracks spaced 0.007 in. apart on an 8 mm film. The maximum recording density is 5000 bits per inch per track, giving a density of 7×10^5 bits per in^2.

During recording, light from a laser is divided into 36 beams and each beam is amplitude modulated with the data for that channel. An optical system then focuses the beams onto the 8 mm film. For read back, the same system is used but with no input to the modulator. An array of 36 photodetectors behind the film provides the output.

The maximum data rate is over 10^7 bits per second at 60 inches/sec film speed, and 4×10^9 bits can be stored on a 10½ inch diameter reel containing 2000 feet of film.

References for Section 2.7

1. Blue, M.D. and Chen, D., "Optical Techniques Light the Way to Mass-Storage Media", Electronics, March 3, 1969, pp. 108-113.

2. Anderson, L.K., "Holographic Optical Memory For Bulk Data Storage", Bell Lab. Record, November 1968, pp. 319-325.

3. LaMacchia, J.T., "Optical Memories: A Progress Report", Laser Focus, February 1970, pp. 35-39.

4. Bossomworth, D.R., et al, "Thick Holograms in Photochromic Materials", Applied Optics, Vol. 7, No. 1, January 1968, pp. 95-98.

5. Shankoff, T.A., "Phase Hologram in Dichromated Gelatin", Applied Optics, Vol. 7, No. 10, October 1968, pp. 2101-2105.

6. Chen, F.S.,et al, "Holographic Storage in Lithium Niobate", Appl. Phys. Lett., October 1968, pp. 223-227.

7. Mikaeliane, A.L., et al, "Design Principles of Holographic Devices", IEEE Jour. Quant. Electr. QE-6, April 1970, pp. 193-198.

8. Lackner, H.G., "Crystal Memory", NASA/MSFC memo R-TEST-ID, (No date).

9. Weitzman, C., "Optical Technologies for Future System Design", Computer Design, April 1970, pp. 169-175.

10. Peters, C.J., et al, "Signal Recorder/Reproducer Using a Coherent Light Source and A Photographic Film Record", Proc. IEEE, Vol. 58, June 1970, pp. 886-893.

11. Amodei, J.J., et al, "Holograms in Thin Bismuth Films", Appl. Phys. Lett. Vol. 15, No. 2, July 1969, pp. 45-46.

12. Cavley, G.L., "Feasibility of Using a Focussed Laser for Mass Information Storage", PhD Thesis, MIT Dept. Of Electrical Engineering, September 1968.

13. Kiss, Z.J., "Photochromic Materials for Quantum Electronics", IEEE Jour. of Quant. Electr., QE-5, No. 1, January 1969, pp. 13-17.

14. Mikaeliane, A.L., et al, "F-1-Holograms on Photochromic Films", IEEE Jour. Quant. Elect. QE-4, November 1968, pp. 757-762.

15. Megla, G.K., "Optical Properties and Applications of Photochromic Glass", Applied Optics, Vol. 5, No. 6, June 1968, pp. 945-960.

16. Townsend, R.L., et al, "Optically Induced Refractive Index Changes in $BaTiO_3$", J. Opt. Soc. Am., Vol. 59, October 1969, p. 1530.

17. Thaxter, J.B., "Electrically Controlled Holographic Storage in Strontium-Barium Niobate", NEREM Record, Vol. 11, November 1969, pp. 10-11.

18. Cummins, S.E., "Electron-Beam Writing of Ferroelectric Domains in $Bi_4Ti_3O_{12}$ Single Crystals", Proc. IEEE (letter), June 1970, pp. 938-939.

19. Chen, D., et al, "MnBi Films: Physical Properties and Memory Applications", Jour. Appl. Phys., Vol. 39, No. 8, July 1968, pp. 3916-3927.

20. Lewicki, G.W., "Curie-Point Switching in Mn-Bi Films", IEEE Trans. Magnetics, MAG-5, September 1969, pp. 298-299.

21. Fan, G.Y., et al, "Low-Temperature Beam-Addressable Memory", Jour. Appl. Phys., Vol. 39, No. 2, February 1968, pp. 1216-1218.

22. Eschenfelder, A.H., "Magneto-Optic Materials", Jour. Appl. Phys., Vol. 41, March 1970, pp. 1372-1376.

23. Chen, D., et al, "Laser Beam Writing, Erasing and Reading on Manganese Bismuth Films", Jour. of Quantum Electronics, Report on Int. Quant. Electr. Conf., May 1968, pp. 378-379.

24. Stoffel, A.M., "Thin Film Magneto-Optic Read-Write Memory Element", IBM Tech. Disc. Bull., Vol. 12, No.1, June 1969, p. 209.

25. Comstock, R.L., et al, "Magneto-Optic Properties of CrTe Films Prepared by Sequential Evaporation", Jour. Appl. Phys., Vol. 41, No. 3, March 1970, pp. 1397-1398.

26. Sadagopan, V., "Magneto-Optic Materials", IBM Tech. Disc. Bull, Vol. 12, No. 10, March 1970, p. 1586.

27. Treves, D. et al, "Laser Beam Recording on a Magnetic Film", Jour. Appl. Phys., Vol. 40, No. 3, March 1969, pp. 972-973.

28. Goldberg, N., "A High-Density Magneto-Optic Memory", IEEE Trans. Magnetics, MAG-3, No. 4, December 1967, pp. 605-611.

29. MacDonald, R.E., et al, "Magneto-Optics", Jour. Appl. Phys., Vol. 40, No. 3, March 1969, pp. 1429-1435.

30. Kump, H.J., et al, "Thermostrictive Recording on Permalloy Films", IBM Technical Jour., May 1966, pp. 255-260.

31. deBouard, "Feasibility Study of a Large Capacity Magnetic-Film Optical Memory", IEEE Jour. Quant. Electr., QE-4, May 1968, p. 378.

32. Smaller, P., "Amplitude-Modulated Magneto-Optic Readout", Jour. Appl. Phys., Vol. 39, No. 2, February 1968, pp. 1218-1219.

33. Smith, D.O., "Magnetic Films and Optics in Computer Memories", IEEE Trans. Magnetics MAG-3, No. 3, September 1967, pp. 433-452.

34. Smith, D.O., "Proposal For A Magnetic-Film Memory Accessed by Combined Photon and Electron Beams", IEEE Trans. Magnetics, MAG-3, No. 4, December 1967, pp. 593-599.

35. Rajchman, J.A., "Promise of Optical Memories", J. Appl. Phys., Vol. 41, No. 3, March 1970, pp. 1376-1383.

36. Urbach, J.C., et al, "Thermoplastic Xerographic Holography", Appl. Opt.,Vol. 5, 1966, p. 666.

37. Lin, L.H., et al, "Write-Read-Erase in Situ Optical Memory Using Thermoplastic Holograms", J. Opt. Soc. Am., Vol. 59, October 1969, p. 1529.

38. Blanchard, J.G., "Beam Addressable NDRO Memory", IBM Tech. Disc. Bull., Vol. 11, No. 4, September 1968, p. 359.

39. Anacjer, W., "Beam Addressable Memory Cell", IBM Tech. Disc. Bull., Vol. 12, No. 10, March 1970, p. 1591.

40. ----------, "Optically Exciting a Magnetic Memory: A Feasibility Study", Computer Design, July 1969, p. 34.

41. ----------, "Electro-Optic Film Functions at Speed of Light", Computer Design, May 1970, p. 28.

42. Smith, D.O., "Review of Proposals for Photon- and Electron-Beam Accessed Magnetic-Film Memory Systems", IEEE Trans. Magnetics, MAG-5, No. 3, September 1969, pp. 641-642.

43. Smith, D.O., "Methods for Improving the Signal-to-Noise Ratio of Photon and Electron Beam Accessed Magnetic Film Memory Systems", IEEE Trans. Magnetics, MAG-4, No. 4, December 1968, pp. 634-639.

44. Mee, C.D., "A Proposed Beam-Addressable Memory", IEEE Trans. Magnetics, MAG-3, No. 1, March 1967, pp. 72-76.

45. Hunt, R.P., et al, "Optical Scanning for a Magnetooptic Memory", J. Appl. Phys., Vol. 41, No. 3, March 1970, pp. 1399-1400.

46. Fowler, V.J., Schlafer, J., "A Survey of Laser Beam Deflection Techniques", Proc. IEEE, Vol. 54, No. 10, October, 1966, pp. 1437-1444.

47. Pinnow, D.A., "Guide Lines for the Selection of Acoustooptic Materials", J. Quant. Electr. QE-6, No. 4, April 1970, pp. 223-2

48. ----------, "Optical Deflection Addressing Moves Holographic Memory Closer to Reality",Electronics, 5 January 1970, pp. 71-72.

49. Pinnow, D.A., et al, "Lead Molybdate - A Melt Grown Crystal With a High Figure of Merit for Acoustooptic Applications", Appl. Phys. Lett., Vol. 15, No. 3, August 1969, pp. 83-86.

50. Böer, K.W., "Deflecting of Light Beams Using a Gradient of the Index of Refraction in Semiconducting Glasses", Appl. Phys. Lett., Vol. 16, No. 6, March 1970, pp. 228-230.

51. Kulcke, W., et al, "Digital Light Deflectors", Proc. IEEE, Vol. 54, No. 10, October 1966, pp. 1419-1429.

52. Chang, D.C., et al, "Electrooptical Light Deflection For a Holographic Memory", NEREM Record, November 1969, pp. 8-9.

53. Nagill, P.J., "Magnetooptic Deflection Method", IBM Tech. Disc. Bull., Vol. 11, No. 7, December 1968, p. 810.

54. Dakss, M.L., "A Fast Digitalized Scan Laser", IEEE J. Quant. Electr. QE-4, No. 5, May 1968, pp. 375-376.

55. Pole, R.V., et al, "Bidirectional Electrically Switched Laser", Appl. Opt., Vol. 4, No. 1, January 1965, pp. 119-121.

56. Land, C.E., "Electrooptic Ceramics for Information Storage and Display", NEREM Record, November 1969, pp. 12-13.

57. Land, C.E., "Ferroelectric Ceramic Electrooptic Materials and Devices", Proc. IEEE, Vol. 57, No. 5, May 1969, p. 751.

58. Heilmeier, G.H., et al, "Dynamic Scattering: A New Electrooptic Effect in Certain Classes of Nematic Liquid Crystals", Proc. IEEE, Vol. 56, No. 7, July 1968, pp. 1162-1171.

59. Assour, J.M., et al, "A Photodetector Array for Holographic Optical Memories", RCA Review, December 1969, pp. 557-566.

60. Bridges, W.B. and Kolb, W.P., "The First Space Qualified Laser", Laser Focus, October 1969, pp. 39-42.

2.8 Cryogenic Memory Technology

2.8.1 Introduction

A low temperature environment for computer logic and memory elements has long been contemplated as desirable, because noise phenomena are diminished, allowing low level, low power operation, and the use of superconductivity offers prospects of high speed, high density memory devices. Device patents were granted as long ago as 1955 [1], and investigations of the superconducting switch and memory element have been going on ever since [2].

Superconductivity offers attractive properties for the implementation of computer memories: the zero electrical resistance provides loss-less paths for driving and sensing, and in the form of recirculating currents makes feasible the concept of a non-dissipative storage element. Furthermore, the perfect diamagnetism serves to inhibit inductive interaction between storage elements, so that very high packing densities can theoretically be realized [2]. Batch fabrication, using thin film techniques, is feasible.

However, the difficulty of working with the extremely low temperatures needed to ensure superconductivity (the highest known superconducting material, Nb_3Sn, requires 18°K) has discouraged all but a few from realizing working cryogenic memory systems. There have been serious attempts by RCA and Siemens [3] to determine the problems associated with cryogenic memories, but most other investigations have been on a laboratory scale, and have confined themselves to explorations of various phenomena and devices based on them.

Most investigations have been concerned with the control of recirculating currents by the introduction of local resistivity, and have differed mostly in the geometry of the elements and the technique used to drive and sense and decode the information. Earlier attempts than the one to be described below integrated the superconducting cryotron switch with the memory arrays, but this lead to difficulties with non-uniformity of switch performance, high drive currents and testing problems [4]. Some investigations into application of the Josephson switch [5] have been made, but none is known to have resulted in a workable memory element.

2.8.2 The Superconducting Core Memory Element

This element stores information by the presence or absence of a continuously circulating current in a closed path of a thin film of tin (see Figure 2.8.1). Figure 2.8.2 illustrates the sequence of writing-in a "one". The hole in the lead substrate locally reduces the decoupling effect of

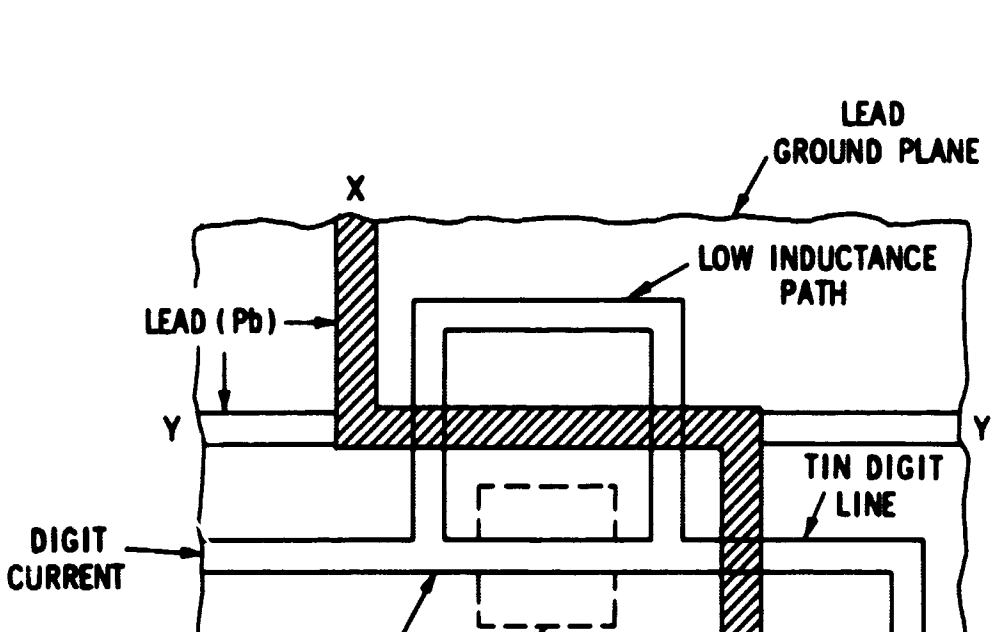


Fig. 2.8.1 Superconducting Memory Element

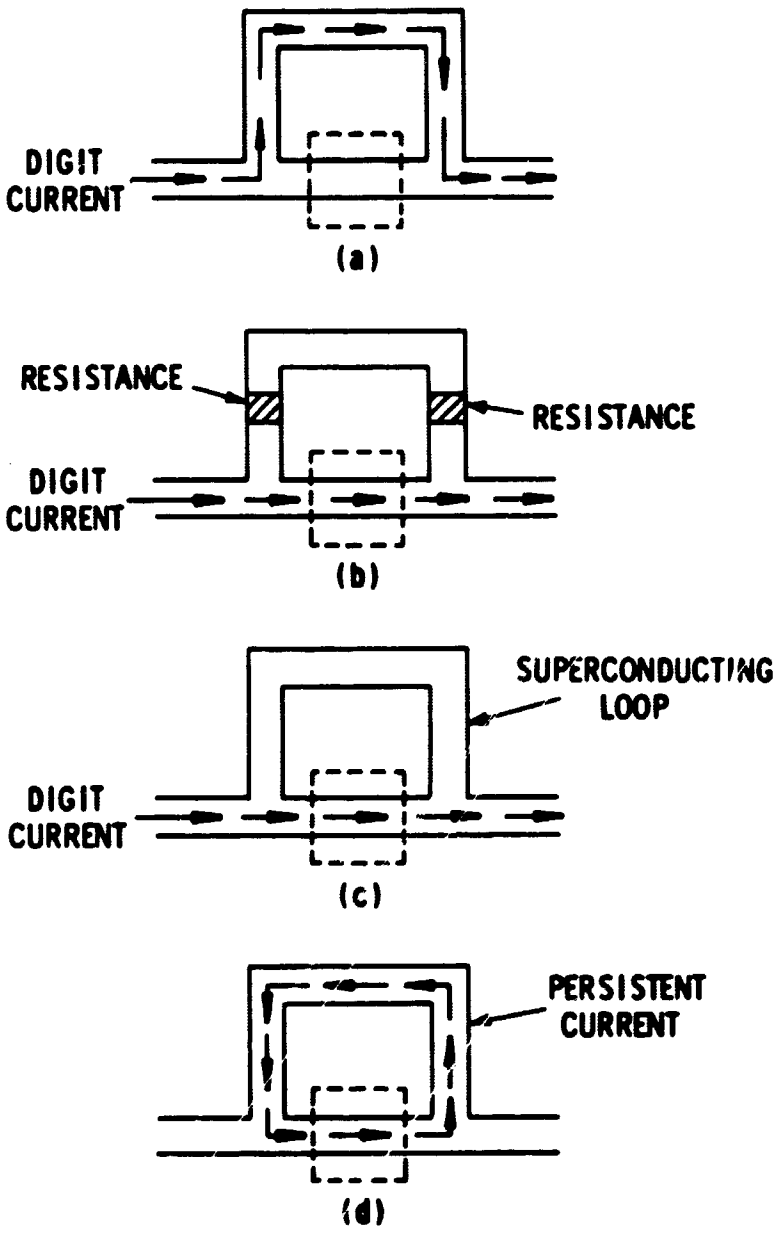


Fig. 2.8.2 Writing a '1' into Superconducting Loop

the diamagnetism and contributes to the higher inductance in that path. This forces the current normally to flow around the longer arm of the closed loop. Resistance is introduced into the low-inductance arms by coincident currents in the X and Y conductors which produce a local magnetic field high enough to destroy the superconductivity. This "shuttling" of the flux between the arms of the core is also the basis for a non-destructive interrogation of the state of the element, but the induced output voltages are only on the order of 200 microvolts per element. The advantages claimed for this element over previous thin-film devices are:

1) The introduction of structure into the design makes the operation depend only on the bulk properties of the material, as opposed to previous attempts with continuous sheet superconductors.

2) Operation that ensures insensitivity of the current levels to the properties of the material.

3) Separation and isolation of the storage and drive/sense currents.

The decode circuits are mounted outside of the cryogenic area required by the arrays themselves. Most of the system dissipation takes place in the connecting leads. However, only an estimated 8,704 wires would be required for a practical 10^8 bit memory [3], because a theoretical 10^7 bits can be driven by one line [4]. The memory array can therefore be regular and of high density: RCA has constructed ¼ million bit arrays at a density of over 10^4 bits/sq.in. [3]. The principal problem of this design is the achievement of adequate yield of good memory planes. The large area, up to 24 square inches, which is possible with this technique, and must be exploited to reduce interconnection problems, makes the substrate very vulnerable to defects during processing, and it is also easily damaged by subsequent careless handling [2].

Read-write cycle times of 4 microseconds have been achieved, with major delays contributed by the external circuitry. Device speeds are determined mostly by geometry, since response is determined by the L/R time constant, where R is the induced non-superconductive resistance, which is a function of the shape and thickness of the thin film element. The superconducting transition itself takes place in less than 10^{-9} seconds.

2.8.3 Disadvantages of Cryogenic Memory Systems

The outstanding problem is, of course, the need for a low temperature environment. The lead/tin element described

previously requires 3.5°K, which can only be achieved by liquid helium techniques. (All other combinations would require at least liquid hydrogen temperatures.) Closed cycle helium liquefiers are commercially available, but their considerable weight and volume would add undesirable overhead to the memory system. This type of equipment is inefficient from a power consumption viewpoint, contains moving parts, requires maintenance to stay within operating levels, and is of unknown reliability in a space environment. In the event of a breakdown in refrigeration the memory becomes entirely volatile as soon as the temperature rises above the superconducting point. Regaining the operating temperature upon correction of the malfunction can take hours rather than seconds. These are formidable snags to be offset by the attractions of the actual storage component. A spacecraft environment, ironically, could provide some assistance to the cryogenic problem. Liquid helium, hydrogen and nitrogen have been used in the propulsion systems of manned space vehicles and could thus be available for the computer memory system. Furthermore, the hard vacuum of space could be used to boil off a fixed amount of liquid helium at a controlled vapor pressure to provide an operating environment for a superconducting memory for a finite period. An efficiently insulated system might be sustained for days by such a passive system with a few litres of liquid helium.

On the whole, the prospect of a cryogenic mass memory system for an orbiting space vehicle is very dim.

References for Section 2.8

1. Buck, D.A, "Magnetically Controlled Gating Element", U.S. Patent 2,832,897, filed July 27, 1955.

2. Gange, R.A., "Cryoelectric Hybrid System For A Very Large Random Access Memory", Proc. IEEE, Vol. 56, No. 10, October, 1968, pp. 1679-1690.

3. Gange, R.A., "Cryoelectric Memories: Best Hope For Large and Fast Storage Units", Electronics, March 17, 1969, pp. 108-112.

4. Gange, R.A., "Taking Cryoelectric Memories Out of Cold Storage", Electronics, April 17, 1967, pp. 111-120.

5. Matisoo, J., "Josephson-Type Superconductive Tunnel Junctions and Applications", IEEE Trans. Magnetics, MAG-5, No. 4, December 1969, pp. 848-873.

Chapter 3

Mass Memory Requirements

3.1 Multiprocessor Design Introduction

Advanced planning for the data handling capabilities aboard the next generation of manned space vehicles has centered around capacity, reliability and flexibility. The Intermetrics Multiprocessor Final Report proposed a multiprocessor computer organization because:

a) the computing capacity required for a large orbiting space station exceeds that attainable from a single processor.

b) the reliability exceeds that of a single processor.

c) it is desirable to allow incremental expansion of the system without overhauling it.

The full detail of the multiprocessor operation is presented in the Final Report to Contract NAS 9-9763. The intention here is to use this configuration as a basis for developing memory requirements. Figure 3.1 is a schematic of the proposed configuration.

3.1.1 Description of Memory Hierarchy

Fundamentally, at least three memory levels are defined:

a) M1: a high speed random access memory. Each processor in the multiprocessor interfaces intimately with an M1 memory. This interface is dedicated: M1 can only be accessed by other processors via the common data bus, with severe reductions in the data rate. Such access will only occur in the event of a processor failure, or for very slow speed operations such as status monitoring. The size of M1 is determined by the scope of the multiprocessor, because M1 serves mainly to provide sufficient capacity to limit the data rate on the common bus by reducing the necessary traffic between processors and the larger memories. (For a more detailed explanation, see the Final Report to the first part of this contract.) The M1/processor interface is designed to facilitate exploitation of the high access speed of M1. Being dedicated it will not have to cope with the constraints imposed on the common data bus, namely: restricted data

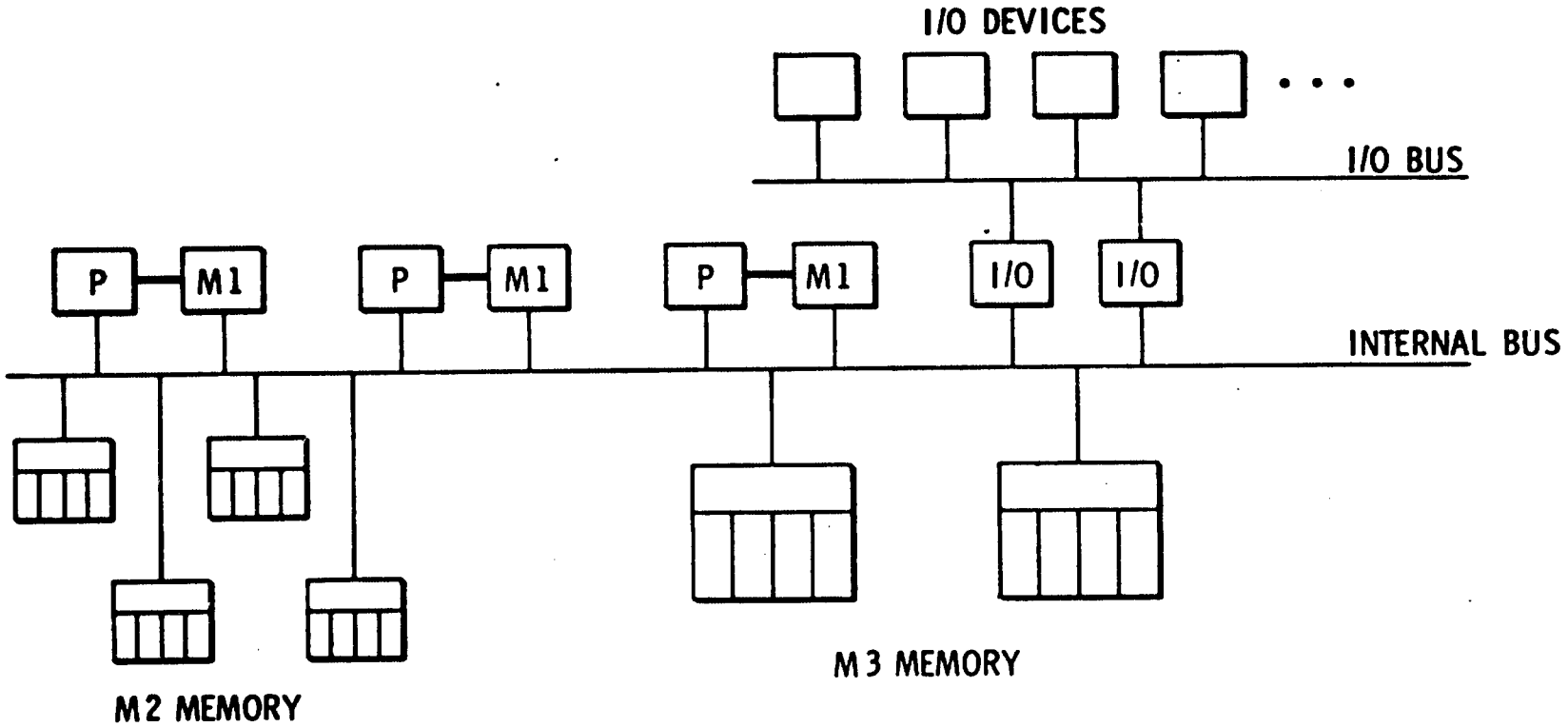


Figure 3.1 Intermetrics Multiprocessor Series

width, reduced rates due to long data paths and their inevitable delays, multiplexing of data, serial/parallel conversion, etc. These problems must be faced by the other interface that M1 must serve, that of the common data bus. The processor interface will provide parallel random access to one or more words in M1 at a time (a word is taken to be of the order of 32 bits), in a manner common to large high speed computers such as the IBM 360/85. The bus interface on the other hand handles groups of, say, 8 words together with all their control and identification bits in a narrow data path of about 10 lines, and must accommodate to all the constraints enumerated above. These widely differing access requirements must be met by the control logic associated with M1. Its task is made more tractable by the fact that the two interfaces do not constitute a true "two-port" mechanism, in that data transfer does not occur across both simultaneously. It may be necessary for M1 to keep a continuous "ear" to the bus so that it may respond to the calls of other control mechanisms, but data transfers are sequential.

b) M2: a moderately fast read and write memory which will hold all the operational programs and data for the tasks at hand. This memory is of sufficient size to require only occasional access to the "bulk store". The ratio of M1 to M2 size can be adjusted for efficient bus utilization. Note that this organization is adaptable to different computing performance requirements. If the processor is relatively slow then M1 will be small and information will be required from M2 quite often, establishing the bus usage. If the processor is faster, a larger M1 is required. The calls to M2 will be relatively less frequent and the bus traffic can be maintained fairly uniform, i.e. independent of processor speed.

M2 is not dedicated to a particular M1/processor pair, and there may be a number of M2 units in the multiprocessor system. These units will contain the prime copies of all programs that are in operation at one time. For example, an M1 memory may demand a copy of a particular program from an M2 known to contain it. The program copy is transferred via the common bus, but the "official" version remains in M2. If during the course of operation the program or its data are modified, then only the necessary updates are communicated back to M2. Thus "read" traffic between M1 and M2 is expected to exceed "write" by a wide margin. This could have a beneficial impact on the design of the M2 memory. Data in M2 is organized into small groups of about 8 words which constitute the smallest addressable units. Each group must, however, be randomly accessible to M1. A measure of "block" orientation in the structure of M2 is therefore possible. The M2/M3 interface is some-

what different even though it takes place over the same data bus as the M1/M2 communication. Although the M2 will accommodate the full set of programs that are operational at any time, it will not provide it a permanent home. This is the function of M3. At times when the nature of the mission changes significantly, operational program swapping occurs over the M2/M3 interface in large blocks of 10^4 bits or more. Since this is fairly infrequent, say every few hours, there is no need to perform the changeover in microseconds, and unload/reload times of up to a few seconds will be allowed, thus contributing a comparatively light load to the bus.

c) M3: This memory will be capable of holding all the necessary operational programs for all vehicle activities as well as providing for collection and storage of experiment data. Two types of manned spacecraft can be envisioned: logistical and laboratory. The primary function of the logistical craft is the transportation of men and supplies; e.g., a space shuttle. Its bulk storage requirements will be minimal: a high capacity may only be necessary to accommodate an unusually large number of operational programs if the logistical vehicle is designed for wide-ranging mission applications. On the other hand, the laboratory vehicle's primary function will be the collection and analysis of experimental data. Of course, a variety of operational and life-sustaining functions will be necessary but the business of the craft, or station, will be the long term observation, collection, analysis, and transmission of experimental data. Undoubtedly, large bulk digital storage capacity will be required.

The data in M3 will be organized into large blocks of around 10^3 words (i.e., 10^4 to 10^5 bits). M3 must be capable of read and write operations. It can be fully block oriented, allowing a shift register or page-oriented type of modular construction. It cannot however, be serially organized as, for example, a tape recorder. Its very large capacity would impose a serial mode access time of hundreds or thousands of seconds, which would be unacceptable as a changeover transient.

As the ultimate receptacle for all operational programs and data M3 must provide highly non-volatile and permanent storage. It must provide a means for updating, changing or adding to this body of software. It would be highly desirable to be able to do this by physical replacement of parts of the storage medium. In this way new information can be prepared off-line at a ground facility where it can be subjected to a more rigorous qualification, and then be shipped via the periodic logistics vehicle for insertion in the computer system. In addition, such a

concept would allow the expansion of the total information capacity available to the on-board computer by creation of an off-line archival storage facility. In the best of all worlds such an archival facility could possibly eliminate the separate serial tape recorders otherwise needed for the collection of the vast quantities of digital data that are expected to be generated by other spacecraft systems and scientific experiments [3,4].

The multiprocessor will include three more small, but identifiable, memory sections comprising microprograms, the stack and provisions for dead-start recovery. All may physically be part of M1 although the dead-start capability should be stored in a separate read-only memory for maximum reliability. Microprograms represent the computer instruction set and are stored in very fast, alterable memory. The significant advantage of microprogramming is flexibility in that additional instructions may be created or instructions changed, after computer manufacture. The "stack" serves as dedicated "scratch pad" storage (e.g., a push-down list) and is unaffected by information swapping among the M1, M2, and M3 memories.

Manned spacecraft application demands that the computer be able to bootstrap itself from a dead-start. Thus, a set of micro-instructions must be provided which will restore a minimum set of computer instructions and initiate a start-up sequence. In addition, certain diagnostic functions must be included to isolate probable malfunctions and at least indicate why full operation is being prevented.

The microprogram control store and the dead-start storage will be considered to form the only requirement for read-only memory in this hierarchy. No need for massive amounts of read-only storage can be defined. In general read-only memories can achieve higher speeds and data densities than read and write memories because of the absence of a write mechanism, and their long term reliability is probably somewhat superior. But their great drawback is the permanence of the stored information. It is inconceivable that large amounts of data and programs will be able to survive for lengthy periods without eventually requiring updating or modification. Changes must be effected either by physically replacing the read-only memory in whole or in part, or, if its basic design allows, by electrically altering its contents. An electrically alterable memory may not have to meet the high-speed write requirements of a read-write memory, but the addition of a write mechanism makes its advantage in terms of speed, size and security of data rather marginal. A memory that requires replacement to effect an update poses a logistics and maintenance problem that is not an acceptable penalty in view of the vagueness of the ROM's advantages. A final argument for the elimination of massive read-only storage is that no large body of programs or data can be defined that

is long-term invariant and also requires a high level of security.

In summary, qualitative memory requirements are presented below for each memory type.

Function	Speed	Size	Type
Microprogram	very fast	moderate	alterable read-only
Dead-start	fast	small	read-only
Stack	very fast	small	read and write
M1: dedicated	fast	moderate	random access
M2: main storage	moderate	large	random access
M3: bulk	slow	very large	block oriented random access

3.1.2 Multiprocessor Performance Characteristics

The access time seen by a processor attempting to fetch instructions and data directly influences the maximum operating speed. In operation, the processor issues requests to the dedicated memory M1. If the addressed word is currently contained in M1, it is sent to the processor; otherwise, M1 initiates a fetch of a group of words from M2 and retains the group for future use. The performance then is a function of the speed ratio between the access times for the two memories and the probability that a given access finds the word not in M1. It can be shown that the normalized word rate, defined as

$$W = \frac{\text{access time if word in M1}}{\text{actual access time}}$$

is given by

$$W = \frac{1}{1 + (R-1)\,p}$$

where R ≡ access time ratio M2/M1

p ≡ probability of word not in M1.

For example, for R = 10 and p = .05, W = .69; i.e., for a miss probability of 5% the system "appears" to operate at 70% of M1 speed even though M2 is only 1/10 as fast. The expression for W is plotted in Figure 3.2 with the probability as a parameter. The Intermetrics multiprocessor takes advantage of this hierarchical relationship between the M1 and M2 memories. It is characteristic of typic[illegible]ograms that accesses to instructions and data tend to be localized, over short time intervals. As a result, a split-level memory can be used to great advantage to provide, at one level, a very high speed modestly-sized store, and at the other level, large capacity at readily attainable speed.

The role of M1 is similar to that of the "cache" memory in the IBM 360/85. A striking result of IBM studies [1,2] on cache performance was that over a variety of different application programs occupying varying amounts of storage, a cache of 16 kilobytes (128K bits) was adequate to attain an average miss probability of 3.2%. The cache represented only about 1-2% of the total 360/85 storage. Referring to Figure 3.2, if the IBM main/cache speed ratio were 10 then the effective speed of the 360/85 would be close to 80% of cache speed. IBM experience also indicates that system performance is not very sensitive to cache size above a critical threshold. That is, once the cache is big enough to eliminate most of the calls to the main memory, doubling its size yields only marginal increase in performance.

3.2 Memory Performance Requirements

In order to implement the multiprocessor design the memory performance requirements are separated into two time periods: 1975, 1980. For each period the necessary memories are logically divided between working memory (M1 and M2) associated intimately with the processor,and the bulk storage (M3) facility. For each case, the memories are large by present day flight standards and are legitimate "mass-memories".

3.2.1 The 1975 Period

3.2.1.1 Working Storage (M1, M2)

Preliminary space station studies by IBM [3] and North American Rockwell [4] indicate necessary processor speeds of 1-10

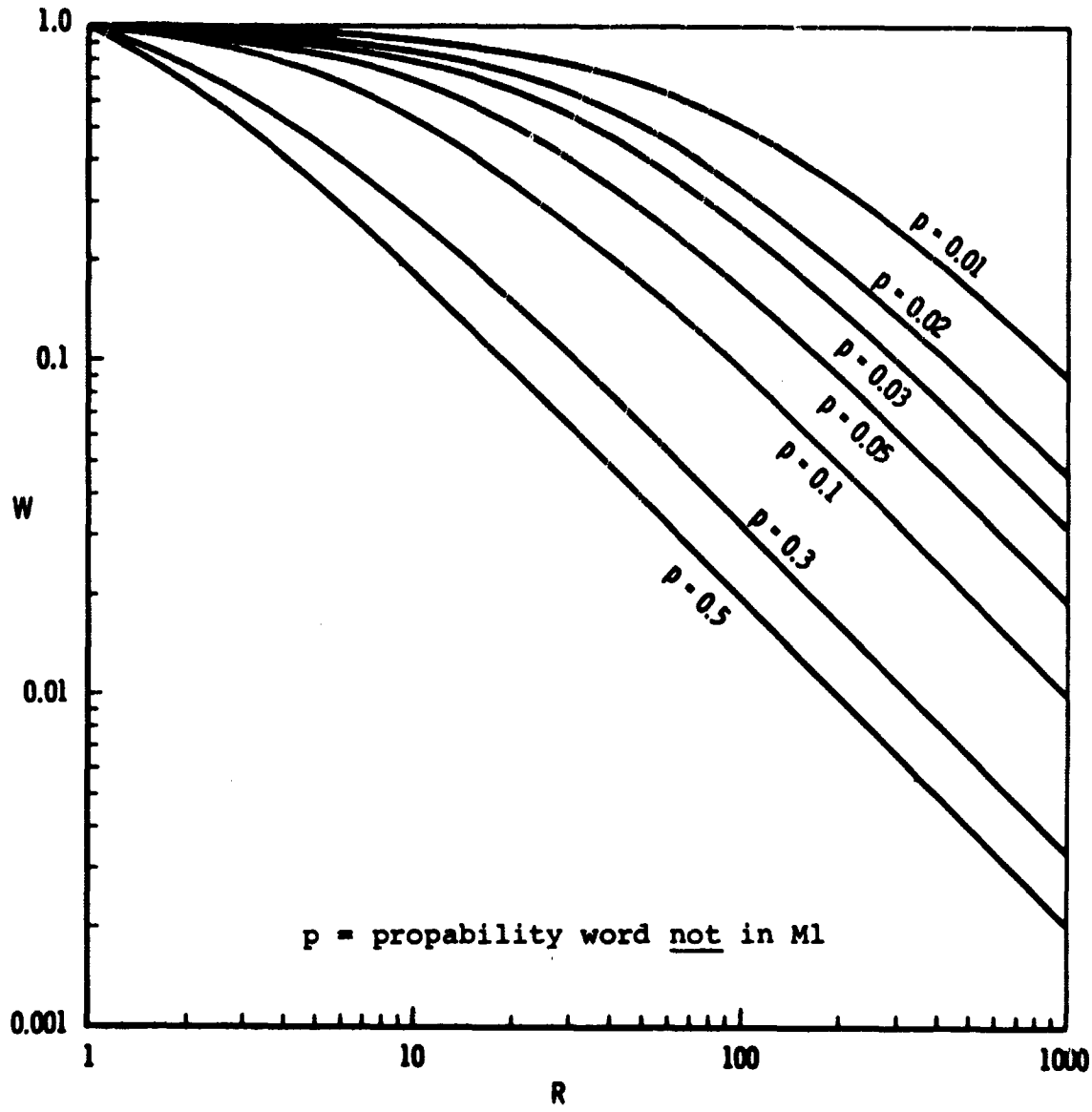


Figure 3.2 Normalized word rate (W) vs. M2/M1 access time ratio (R)

million instructions per second (MIPS). (These are commonly referred to as "equivalent add instructions".) The processor speed requirements for 1975 will be set at 1 MIPS, or an effective access time of 1 μs. The effective access time is achieved through the hierarchy of M1, M2 as previously described and illustrated in Figure 3.2. Initially (1975) requirements for memory size will be set similar to the IBM 360/85 experience. Thus, let M1 = 10^5 bits and M2 = 10^7 bits (100 times as large). The IBM analysis for cache memories of about 10^5 bits capacity indicated that miss percentages varied from 36% to 4% for a wide variety of actual programs. If the M1 and M2 access times are represented by t_1 and t_2, in order to achieve an operating speed of 1 MIPS, the factor $W/t_1 = 10^6$. The resulting required values for t_2 are tabulated below against different M1 access times, for miss probabilities of .02, 0.1, 0.25.

		p = .02		p = 0.1		p = 0.25	
t_1 (ns)	W	R	t_2 (μs)	R	t_2 (μs)	R	t_2 (μs)
50	.05	951	47.6	191	9.55	77	3.85
100	.1	451	45.1	91	9.10	37	3.70
200	.2	201	40.0	41	8.20	17	3.40
500	.5	51	25.5	11	5.50	5	2.50

Table 3.1 M2 Access Time (t_2) as Function of M1 Access Time (t_1) for 1 MIPS

This table indicates that very modest speeds are required of M2 in order to satisfy an instruction frequency of 10^6 per second at even high miss probabilities. However, three other factors not accounted for to this point will demand that M2 exhibit a higher performance:

1) M2 acts as a working memory to a number of processor/M1 pairs, and must have sufficient margin of speed to service simultaneous accesses for short periods of time.

2) No account has been made of M2 storing activities. Read requests are expected to exceed stores by some factor that is a function of the program running.

3) Formatting, serial-parallel conversion, and other data transmission operations must be included in the M2 access time.

For these reasons the M2 access time requirements will be set in the region 2 μs - 4 μs, which provides a factor of 2 to 4 margin over the minimum speed predicted by access requirements alone. M1 speeds for the 1975 period will be set at 100 ns - 200 ns. The capacity and speed of M1 and M2 for a 1975 operational date will therefore be specified as:

	M1	M2
Capacity	10^5 bits	10^7 bits
Access time	100 ns-200 ns	2 μs-4 μs

(Note: it can be presumed that the microprograms, stack and dead-start storage will total $<10^5$ bits and can be similar to M1, or perhaps somewhat faster; e.g. 10^4 bits @ 50 ns. These smaller memory categories will not be considered further.)

3.2.1.2 M2 Data Width and Bit Rate

If a fetch is required from M2 to M1 then efficiency would dictate that more than one word should be transferred. The Multiprocessor Final Report suggests that M2 be arranged in 8 word blocks (32 bits/word). Accounting for error checking, identification and addressing it may be presumed that M1 could accept one 40 bit word per 100 ns, or approximately a 400 bit block every 800 ns. The data rate out of M2 is therefore 400 bits/800 x 10^{-9} = 5 x 10^8 bits/sec. It is reasonable to assume that with state-of-the-art electronics, a serial transfer rate of 5 x 10^7 bits/sec can be achieved. Therefore, the data width must be at least

$$\text{Width} \quad \frac{5 \times 10^8}{5 \times 10^7} = 10 \text{ lines}$$

For reasons of reliability the transmission of data should occur over as few paths as possible. However, a data width of this magnitude will not in general be the most natural for the accessing mechanism of the M2 memory, or for storing in

M1. Most high speed memory systems prefer to handle data in words of up to several hundred bits at a time. These incompatibilities must be resolved by interface logic at either end of the data bus, which must reformat the data to suit the control mechanisms of the individual memories.

3.2.1.3 Bulk Storage (M3)

Estimates of required bulk storage for manned space operations vary widely; preliminary estimates range from 10^9 bit [3] to 10^{12} bits [4]. An analysis by North American Rockwell indicates that the space station cumulative data processing requirements will be 300-400 x 10^9 bits/day with experiments accounting for 90% of the total. It is expected, by Intermetrics, that the computer system will process and compress this data so that only a small percentage of it need be maintained in the files at any one time. For 1975, 10^9 bits of bulk storage would seem adequate and quite ambitious.

The memory will be block oriented so that 1/1000 of M2 memory can be replaced with a single fetch request. Thus, with M2 = 10^7 bits, each M3 block will consist of 10^4 bits and there will be 100,000 blocks. The access time and data rate will depend upon the desired speed of block transfer. This will not be a critical design parameter. Bulk storage should not experience frequent accessing, and in addition, the operating system should be willing to wait reasonable lengths of time in order to write or read bulk storage. Rather arbitrarily, suppose it is desired to transfer a 10^4 bit block of data between M2 and M3 memory in 0.1 sec (that is, all of M2 could be rewritten in 100 sec). Under these circumstances the M3 access time would be 0.1 sec and the bit rate 10^5 bits/sec. Allowing for a serial transfer rate of only 10^7 bits/sec because of longer M3 to M2 lines, transmission could easily be handled serially over a single line.

If the requirements on total M2 replacement are tightened, say to 10 sec, then a 10^4 bit block must be accessed in 10 ms and the bit rate rises to 10^6 bits/sec. Let the following set be defined as the M3 1975 requirement:

	M3
Capacity	10^9 bits
Access time	10 ms - 100 ms
Block size	10^4 bits
Bit rate	10^5 - 10^6 bits/sec

3.2.2 The 1980 Period

For 1980 the required performance numbers of the previous sections are increased 10-fold. It is felt that at these levels the data handling and storage requirements for an earth orbiting space station through 1985 can be satisfied.

3.2.2.1 Working Storage (M1, M2)

The desired processing speed is set at 10 million instructions per second, or an effective access time of 100 ns. Both M1 and M2 are increased in size, to 10^6 and 10^8 bits respectively. An effective access time of 100 ns implies that $W/t_1 = 10^7$. The corresponding t_2 values for miss probabilities of .02, .05, 0.1 are tabulated below. Although the composition and average mix of programs is not expected to change significantly between 1975 and 1980, the size of M1 has been increased 10-fold. The result will be a lower average probability that a desired word is not in M1, and for this reason only ranges of p up to 0.1 are tabulated.

		p = .02		p = .05		p = .1	
t_1	W	R	t_2 (μs)	R	t_2 (μs)	R	t_2 (μs)
20	0.2	201	4.02	81	1.62	41	0.82
40	0.4	76	3.04	31	1.24	16	0.64
60	0.6	34	2.04	14.3	0.86	7.7	0.46
80	0.8	13.5	1.08	6	0.48	3.5	0.28

Table 3.2 t_2 as a Function of t_1 for 10 MIPS

Using the same kind of reasoning as in Section 3.2.1.1 for 1975, t_2 must be adjusted downward. However, with larger size memories for M1 and M2, and an only slightly increased level of bus traffic, stores and time sharing should have a less serious effect. Consequently the following requirement for M1 and M2 for 1980 will be defined as

	M1	M2
Capacity	10^6 bits	10^8 bits
Access time	40 ns - 60 ns	0.5 μs - 1 μs

3.2.2.2 M2 Data Width and Bit Rate

Continuing with the 8-word block at approximately 40 bits/word, and a 50 ns access time yields

$$\frac{400 \text{ bits}}{8 \times 50 \times 10^{-9}} = 10^9 \text{ bits/sec}$$

Using the serial rate of 5×10^7 bits/sec

$$\text{Width} \quad \frac{10^9}{5 \times 10^7} = 20 \text{ lines}$$

Thus, the data width for M2 should be increased to 20 lines, unless the data transmission frequency can be developed upwards to 500 megabits/sec.

3.2.2.3 Bulk Storage (M3)

The capacity of M3 for 1980 will be set at 10^{10} bits, and the access time will remain at the lower bound set for 1975, namely 10 ms. The block size will be increased to 10^5 bits, which imposes a data rate requirement of 10^7 bits/sec, allowing M2 to be completely replaced in 10 seconds. This data rate should be readily handled by a single transmission line between M2 and M3. These specifications are summarized below.

	M3
Capacity	10^{10} bits
Access time	10 ms
Block size	10^5 bits
Data rate	10^7 bits/sec

3.3 Memory Physical Requirements

The memory system will affect and be affected by its environment primarily with regard to its volume, weight, and power dissipation. For application to an advanced manned space vehicle it is obviously desirable to have available the smallest, lightest, most economical memory system which provides the highest capacity and the shortest access times. However, when looking toward an earth orbiting space station designed to be

able to sustain fifty to a hundred persons for extended periods of time, it is of questionable justification to specify a 1/10 cubic foot rather than a 1 cubic foot, or even 10 cubic foot mass memory unit. Although it is obvious that "the smaller it is the better", the specification of physical limitations for the memory involves factors of existing or future technology, of development time, reliability and maintainability, all of which can be given the common denominator of cost. If specifications are written around the ultimate promise of a technology there is a great risk that even after considerable expenditure of time and money the objective may prove to be unattainable.

Most aerospace electronic equipment today is packaged in units that are 1 to 1½ cubic feet in volume, weigh 50 to 100 lbs., and dissipate 100 to 300 watts. These characteristics are approximately those of a simple ATR box. It would seem reasonable that the physical specifications of the memory system should be in line with other packaged electronic equipment in an advanced spacecraft. Not only would this approach enable a more reasonable demand to be made of the technology, but the more mundane problems of equipment handling, mounting, and maintenance, of the repair, replacement and storage of spare modules, the design of controls, cables, plugs, etc. would be similar to those of other equipment in the spacecraft.

Consequently we will postulate the dimensions of the elements of the memory system in terms of cubic feet rather than the cubic inches theoretically realizable with some of the candidate technologies.

3.3.1 Physical Criteria for M1 and M2

For purposes of physical sizing we will combine the functions of M1 and M2: since M2 has a capacity one hundred times that of M1, but is of a similar organization, it is likely to overwhelm M1 in terms of physical characteristics. The specification of 1 cubic foot for the combined volume of M1 and M2 imposes an overall M1/M2 packing density requirement of about 6000 bits/cubic inch for the 1975 capacity figure of 10^7 bits: this density is within the reach of several of today's technologies. The physical density attained by current electronic packaging practices is about that of water, i.e. a 1 cubic foot unit weighs 50 to 100 lbs. This order of weight allows internal structural modularity to be somewhat non-uniform and provides space around heat-dissipating components for the circulation of cooling air. The 1975 allowed weight for the M1/M2 unit will be 100 lbs. Heat dissipation will be limited to 100 watts, which is small enough to allow it to be conducted away from the external surface of the unit without massive thermally conducting structures. Passage of

air through the unit is undesirable because of the risk of contamination of the active memory elements by water, dust, oil, etc. Furthermore, it is highly desirable that individual units be hermetically sealed, so that variations in external pressure do not affect the operation of the memory. This aspect is discussed in Section 3.5.

For 1980 the M1/M2 required capacities increase tenfold to 10^8 bits and the access time decreases from 4 μs to 1 μs. It will be assumed that the major thrust of technology will be to realize this improved performance without corresponding increases in size and power dissipation. It is expected that the physical density will increase somewhat as more efficient packaging and structural techniques are developed. Accordingly the 1980 physical requirements will be set at volume: 1 cubic foot, weight: 100-200 lbs, power dissipation: less than 100 watts.

3.3.2 Physical Criteria for M3

The M3 memory represents a capacity increase over M2 by a factor of 100. The concessions of a lower access speed and a block-oriented organization however, permit the data to be stored at a higher density, and a corresponding physical increase of the same order is not necessary. Techniques that would be applicable to the storage of 10^9 to 10^{10} bits will probably involve a considerable amount of "overhead" equipment, i.e., apparatus or structure which is not directly involved in the storage of data. For example, in the case of optical beam memories the laser and its associated control and detection equipment would fall into this category. The very high densities of the actual storage medium used in such mass memories compensates for the otherwise inefficient use of space. No packaged memory of M3 performance capabilities has been developed for space-borne use to date, and therefore very little precedent is available upon which to base a reasonable physical specification.

On the assumption that storage medium densities of 10^8 to 10^9 bits per cubic inch are achievable, it would seem that the addressing, reading and writing mechanism would dominate the physical characteristics of an M3 memory. Accordingly a somewhat arbitrary size range of 2 to 4 cubic feet will be chosen for M3 as a reasonable volume in which to package a space qualified M3 mechanism. Since the physical density is unlikely to be high, a weight range of 100 to 200 lbs. will be specified. A power dissipation of 200 to 400 watts will be allowed, with the same restriction for entry of cooling air as for M1/M2.

In view of the unavailability of practical data for this type of memory, no adjustments are made to the specifications in going from the 1975 requirement of 10^9 bits to the 10^{10} bits of 1980.

3.3.3 Summary of Performance and Physical Requirements

In the previous sections an attempt has been made to characterize a memory system that provides the necessary storage facility for the kind of advanced manned space vehicle computer defined in the Final Report to the first part of this contract. The performance specifications were derived by considering the requirements that would be placed on such a computer by a space station application.

It is possible to build a system to meet these specifications today, with storage devices that are available as off-the-shelf commercial items, or with equipment that can be built using standard parts and familiar construction techniques. The factor that turns this into something that is currently beyond the state of the memory art is, of course, the exigencies of a space borne environment.

The constraints of space, weight, and a limited source of power are too severe for the massive rotating magnetic storage devices of a typical computer facility or for the dissipative, high speed, working memory whose brow must sometimes be cooled by a closed circulating fluid system, as is the case in the IBM 360/85. What these constraints might be specifically was postulated in the previous sections. However, the numbers given for volume, weight and power can only be viewed as being reasonable and approximate and are not to be taken too definitively. For example, a 1975 bulk memory of 10^9 bits occupying 2-4 ft^3 is specified. If this can be realized in 1 to 6 ft^3 then the requirement must be considered satisfied. If the memory cannot be accommodated in less than 10 to 50 cubic feet, then the intent of the requirement is violated. The data presented can be useful for prediction, estimation and evaluation, but only to the extent that judgement and perspective are exercised.

The specifications suggested for 1980 are, of course, of greater speculation than for 1975. The effects of unexpected technological breakthroughs could make the speed, volume, weight and power guidelines ludicrously conservative. On the other hand, it is also possible that the expected performance improvements over the next ten years cannot be realized because of a fundamental impasse, which would serve to make the physical predictions unattainable.

Tables 3.3 and 3.4 provide a summary of the performance characteristics of M1, M2 and M3 for the two periods 1975 and 1980. The sections that follow will provide a discussion of environmental and operational factors that will also impact the design of space-borne memory systems, but probably in a more subtle fashion than performance.

	M1	M2	M3
CAPACITY	10^5 bits	10^7 bits	10^9 bits
BLOCK* SIZE	30-40 bits	250-400 bits	10^4 bits
ACCESS TIME	100-200 ns	2-4 μs	10-100 ms
DATA RATE	5 x10^8 bits/sec	5 x10^8 bits/sec	10^5-10^6 bits/sec
VOLUME	1 ft^3		2-4 ft^3
WEIGHT	100 lbs		200-400 lbs
POWER	100 watts		200-400 watts

Table 3.3 1975 Memory Requirements

* Smallest addressable unit

	M1	M2	M3
CAPACITY	10^6 bits	10^8 bits	10^{10} bits
BLOCK SIZE	30-40 bits	250-400 bits	10^5 bits
ACCESS TIME	40-60 ns	0.5-1.0 μs	10 ms
DATA RATE	10^9 bits/sec	10^9 bits/sec	10^7 bits/sec
VOLUME	1 ft^3		2-4 ft^3
WEIGHT	100-200 lbs		200-400 lbs
POWER	100 watts		200-400 watts

Table 3.4 1980 Memory Requirements

3.4 Environmental and Operational Requirements

3.4.1 The Rigors of Space Flight

The launch, operation and recovery of a space craft encompass a variety of environmental conditions which will stress both men and equipment. Space craft and trajectories are designed to achieve a maximum of human safety and as such mechanical and electrical devices on-board will benefit from the necessary benign operating conditions.

For the majority of the time the physical operating conditions do not impose undue stress on the equipment. Temperature and pressure of future manned spacecrafts will be controlled to provide a "shirt-sleeve" environment. The accelerations during orbital flight will occur for only short periods of, at the most, a few minutes, and will not be intense, probably less than 1 g. However, for infrequent events such as launch, re-entry, or emergency situations the physical environment of the mass memories will assume intense and fairly severe characteristics for which due provision must be made in the design.

The computer mass memories for manned spacecraft will provide the bulk and archival storage for long-term space observation and experimentation, perhaps over a period of several years. Under these conditions two other factors will be of importance in the design specification of a mass memory: its ability to successfully survive the short and long-term accumulative effects of radiation, and the ability to achieve the necessary lifetimes and levels of reliability required by the chosen maintenance procedures.

The mass memory environmental requirements will next be discussed under the headings of the physical environment, the radiation environment and reliability requirements.

3.4.2 Physical Environment

The mass memory device must be constructed so as to survive the accelerated portions of flight. These may be characterized as follows: less than 10 g's during launch, generally less than 1 g during orbital maneuvers and less than 10 g's during entry. In addition, levels of vibration are anticipated during these periods. Vibration may be attributable to engine performance and to structural bending and/or fuel sloshing induced by interactions among the space craft, its control system, and in the case of atmospheric flight, the environment. Sinusoid oscillations of up to 25 g's and more can be experienced at frequencies greater than a few hundred hertz. In general an adequate design specification would be the MIL-SPEC, MIL E-5400K for electronic equipment, which covers the vibration regimes for

aircraft and helicopters.

Although future spacecraft are expected to be designed for normal ambient pressures and temperatures, nevertheless for emergency reasons it will probably be necessary for the cabin areas to be capable of being exposed to the vacuum of outer space and yet remain operational. Consequently it is highly desirable that should the computer system be expected to remain functional the memories will have to withstand a zero pressure environment. This implies that perfect hermetic sealing, which is traditionally a very difficult thing to achieve, must be provided for units up to several cubic feet in volume; alternatively the internal components themselves must be able to function in a vacuum. The latter capability is obviously easier to realize for solid state devices than for those which employ moving parts. Even if operation is not required in the condition of zero pressure, the memory must obviously not suffer permanent failure as a result of its exposure to it. An outcome of this requirement is that cooling of dissipating units should not be achieved by circulating air around the exterior surfaces. Cooling must be accomplished by conduction into plates, jackets, etc., to which the unit is thermally bonded. Heat must be removed by coolants contained in hermetically sealed closed systems, by thermoelectric methods, or by radiation.

3.4.3 Radiation Environment

A spacecraft in earth orbit, or for that matter, on an interplanetary trajectory will experience doses of radiation. Radiation can cause transient and/or permanent damage to electronic equipment and can materially affect memory performance depending upon the techniques employed. Of the computing memory technologies the crystalline structure of semiconductors is the most susceptible to radiation. In this section certain features of the near-earth radiation environment and their effect on semiconductor material are described.

The primary belts of radiation in the vicinity of the earth have been designated as the Inner and Outer Zones. The Inner Zone extends roughly from altitudes of 1000Km to 10,000Km, and the Outer Zone from there to 80,000Km. Semiconductor damage is a result of radiation absorption and as such depends on particle density as well as incident particle energy. Four types of radiation are of interest: protons, neutrons, gamma rays and electrons [5]. (Gamma rays refer to electromagnetic radiation with energies higher than typical x-rays.) The types as they occur near the earth have a wide spread in intensity, energy and time distributions but the significant radiation consists mainly of trapped electrons and low energy protons. Table 3.5 lists the particle distributions in the Inner Zone in terms of particle energy, incident rate and approximate yearly dosage.

Particles	Energy	Distribution Particles/sec/cm^2	Total Dose/Yr. Particles/cm^2
Trapped	>20KeV[1]	$<2 \times 10^{10}$	6×10^{17}
Electrons	>200KeV	$<10^8$	3×10^{15}
	>600KeV	$<10^7$	3×10^{14}
Protons	>10MeV	$<10^7$	3×10^{14}
	>40MeV	$<2 \times 10^4$	6×10^{11}

[1] 1 electron-volt (eV) = 1.6×10^{-19} joules, or 1.6×10^{-12} ergs.

Table 3.5 Inner Zone Charged Particle Distributions

The data on Outer Zone radiation is listed in Table 3.6 and represents the environment at synchronous orbit altitudes (approximately 18,000 - 20,000 N.M.).

Particles	Energy	Distribution Particles/sec/cm^2	Total Dose/Yr. Particles/cm^2
Trapped Electrons	<5eV	3×10^9	10^{17}
	>200eV	3×10^9	10^{17}
	>10KeV	7×10^7	2.1×10^{15}
	>100KeV	3.5×10^7	1.1×10^{13}
	>1MeV	5×10^5	1.5×10^{13}
Trapped Protons	<5eV	3×10^9	10^{17}
	>200eV	3×10^9	10^{17}
	>1.8KeV	2×10^8	6×10^{15}
	>100KeV	1×10^7	3.8×10^{14}
	>1MeV	2×10^3	5.4×10^{10}
Solar Protons	>1MeV	3×10^3	1×10^{11}
	>10MeV	4×10^2	1.3×10^{10}
	>100MeV	1.5×10^1	4.6×10^8
Solar α-Particles (Max)	>1MeV	1.5×10^3	5×10^{10}
	>10MeV	2×10^2	6.5×10^9
	>100MeV	7	2.3×10^8
Galactic Cosmic Rays	>100MeV	5	1.7×10^8

Table 3.6 Outer Zone Charge Particle Distribution [6]

The most intense regions of trapped radiation are in the Inner Zone. The Outer Zone trapped protons do not present much of an electronic hazard because of their low energy of 1 MeV or less [7]. However, solar protons which occur during solar flares and times of sun spot activity can contribute significant levels of radiation. Shielding can be very effective in reducing semiconductor damage. Table 3.7 illustrates the residual radiation for two shielded devices in synchronous orbit.

Particles	Total Dose/Year Particles/cm	
	6mil SiO_2[1]	7-mil Kovar[2]
Electrons	3.9×10^{14}	1.3×10^{14}
Protons	2×10^{10}	8×10^{9}
αParticles	2.5×10^{9}	1×10^{9}

[1]minimum shielding applied to silicon solar cell

[2]minimum shielding provided by can or cover of hermetically sealed semiconductor device

Table 3.7 Effect of Shielding in Synchronous Orbit [6]

The discussion here has been confined to trapped or other natural radiation phenomena. Radiation hardening of semiconductor devices has received much attention in the literature in connection with military applications, that is, the effects of nuclear weapons radiation. Naturally, this produces radiation of much higher incident energies consisting to a great extent of high energy neutrons. In the context of mass memory requirements for advanced manned spacecraft, it is presumed that the consideration of savage radiation environments is beyond the scope of this report.

3.4.3.2 Effect of Radiation on Semiconductor Material

There are two fundamental effects of radiation in semiconductor material, displacements and ionization [5]. Displacement refers to the physical damage to a crystal lattice

produced by knocking an atom from its normal position. Ionization is the freeing of orbital electrons to form ionized atoms and free electrons. These two effects cause three types of damage: displacement damage, transient damage and surface damage.

Electrons and protons are the important types of damaging radiation in the near-Earth environment. Both can cause displacement, ionization, and transient damage. However, energy magnitudes are so low that transient effects can be neglected and displacement damage is only significant for solar cells. Therefore space radiation effects can be attributable primarily to surface damage caused by ionization. The insulated gate field effect transistor (IG FET or MOS transistor) is a device that operates by surface charge phenomena. Ionization and charge accumulation within the silicon dioxide gate dielectric layer seriously affect the performance, acting like an incremental gate voltage. The silicon dioxide serves as an insulator and ionization effects can be minimized by improving the insulation through use of other materials with a denser crystalline structure, e.g., silicon nitride [5].

The relationship between the incident particle radiations listed in the tables presented above and the resultant ionization in the semiconductor material is complex. Absorbed radiation is measured in rads-silicon where 1 rad is defined as the amount of radiation that deposits 100 erg of energy per gram of silicon.

The absorption coefficient for silicon is fairly independent of the level of incident energy and a reasonable value for energy conversion from surface to volume for low energy particle radiation is

$$4 \times 10^5 \text{ rads} = 1 \text{ cal/cm}^2 \text{ of incident energy}$$

This may be expressed in the more convenient approximations of:

$$10^5 \text{ rad} = 1 \text{ joule/cm}^2$$

or $$1 \text{ rad} = 100 \text{ ergs/cm}^2$$

or $$1 \text{ rad} = 6 \times 10^{13} \text{ ev/cm}^2$$

Absorption data can now be added to the tables of the previous section by noting the incident energy/cm^2 and converting to numbers of rads and silicon. These conversions are presented in Tables 3.8 and 3.9. The absorbed energies (and energy rates) can be used to compute resulting charge concentrations (apparent gate voltages) and leakage currents.

Particles	Energy	Energy Rate (rads/sec)	Total Dose/ Year (rads)
Trapped Electrons	20KeV	6.7×10^{-1}	2.1×10^{7}
	200KeV	3.3×10^{-1}	1×10^{7}
	600KeV	1×10^{-1}	3.1×10^{6}
Protons	10MeV	1.6	5×10^{7}
	40MeV	1.3×10^{-2}	4.1×10^{5}

Table 3.8 Absorbed Radiation From Inner Zone Particles

Particles	Energy	Rads/Sec	Total Dose/ Year Rads
Trapped Electrons	<5eV	2.5×10^{-4}	8×10^{3}
	200eV	1×10^{-2}	3.1×10^{5}
	10KeV	1×10^{-2}	3.1×10^{5}
	100KeV	5×10^{-2}	1.6×10^{6}
	1MeV	1×10^{-2}	3.1×10^{5}
Trapped Protons	<5eV	2.5×10^{-4}	8×10^{3}
	20VeV	1×10^{-2}	3.1×10^{5}
	1.8KeV	5×10^{-3}	1.6×10^{5}
	100KeV	1.6×10^{-2}	5×10^{5}
	1MeV	3×10^{-5}	9.5×10^{2}
Solar Protons	1MeV	4.5×10^{-5}	1.4×10^{3}
	10MeV	6.7×10^{-5}	2.1×10^{3}
	100MeV	2.5×10^{-5}	8×10^{2}
Solar Particles (Max)	1MeV	2.3×10^{-5}	7.2×10^{2}
	10MeV	3.4×10^{-5}	1.1×10^{3}
	100MeV	1.3×10^{-5}	4×10^{2}
Galactic Cosmic Rays	100MeV	1×10^{-5}	3.1×10^{2}

Table 3.9 Absorbed Radiation From Outer Zone Particles

The absorption rates shown are quite low but the total yearly doses are large enough to cause surface damage to MOS devices. It can be shown [5] that 4 $x10^{13}$ carriers/cm^3 are produced per rad-silicon. These carriers may recombine, be extracted at the oxide boundaries, or become trapped in trapping centers within the silicon dioxide films covering the silicon-device surfaces. The charge distribution is then established under the influence of the applied gate voltage and in fact, the resulting damage is somewhat dependent on this voltage during irradiation. The net space charge is observed to be positive with the consequent negative charge induced at the oxide boundary [8]. The electrical properties of silicon dioxide films determine the degree of surface degradation from exposure to ionizing radiation.

The MOS transistor is a pure surface device. For an n-channel device, the space charge induced by radiation tends to make the surface more n-type, i.e., more conducting,and alters the effectiveness of the applied gate voltage [9]. Experimental data taken on a number of MOS devices illustrates the sensitivity of gate threshold voltage as a function of radiation absorption,with applied gate voltage during radiation as a parameter. For gate biases between 10 and 20 volts, gate threshold voltages vary by factors between 2½ and 3½ to 1 after 10^6 rads of absorption [10]. Barry and Page [8] suggest that doses as low as 10^7 rads can produce significant voltage shifts and that the shift is a linear function of the dose up to about 10^7 rads after which saturation gradually takes place. Gordon's data would indicate that saturation occurs at higher dose levels, perhaps closer to 10^7 rads. The effects seem to be dose-rate independent, at least for rates less than 30 rads/sec, and show little recovery after several months at room temperature [8].

It is apparent from the literature that MOS devices are sensitive to radiation-induced surface damage at the dosage levels expected in near-Earth space flight. Therefore, they are unsuitable as reliable storage media. However, MOS can be "protected" by shielding and/or radiation hardening. Shielding with aluminum reduces the impinging low energy electron and proton populations and, depending on the number of grams/cm^2, allows the less frequent higher energy radiation to pass. Gordon [10] shows that the threshold voltage variation can be reduced from 5 to 1 to 1.25 to 1 by the use of 2.5 gm/cm^2 shielding (for an annual IMP-satellite radiation dose). Radiation hardening takes advantage of the dependence of radiation effects on applied gate voltage. In particular, moderate amounts of radiation at large positive voltages (say +30 volts) reduces the sensitivity to radiation at lower voltages [8]. A typical device can be made to withstand radiation doses two orders of magnitude greater than an unhardened device, before

an arbitrary small change in bias point occurs. This would represent a worthwhile improvement in radiation hardness. Barry tested a p-channel MOSFET which required re-zeroing after only a few hundred rads. With hardening, the device required no adjustment after 2×10^4 rads. Since the surface effect is dependent on the applied gate voltage, eventually the quantity and spatial distribution of the trapped space charge will change to comply with the environment of the operating gate voltage, and the hardening is lost. For this reason marked improvement through hardening can only be expected if doses do not exceed a few tens of thousands of rads.

The most promising approach to reducing radiation sensitivity is through a basic modification of the device structure; i.e., the use of a better insulating material such as silicon nitride in place of silicon dioxide. The nitride material is less susceptible to charge buildup induced by radiation. The test results reported by Gordon are indeed impressive. The radiation performance of metal-silicon nitride-silicon devices were compared against two types of MOSFETs; all were p-channel devices. The radiation doses ranged up to 10^5-10^6 rads of 1.5 MeV electrons at three sets of gate voltages. In each case, the nitride threshold voltage varied only a fraction of a volt for zero to full radiation. As radiation was increased further, up to 10^8 rads, the threshold voltage exhibited more variation but the maxima still were constrained to about ±1 volt. The conclusion drawn was that MNS devices offer practical radiation tolerance for space application and other insulators should be sought for further improvement.

In view of the foregoing discussion, the following requirements are established for semiconductor use in a mass memory:

1) If MOSFETs are used then shielding is required such that the device itself will receive less than 10^4 rads/year. It is felt that at this level the requirement on threshold voltage can be met.

2) Semiconductor devices must be radiation tolerant to the extent that gate threshold voltage variations must be held to within ±1 volt over the useful lifetime (say five years).

3) It is recommended that nitrides or other improved insulators be used and that shielding be added if necessary to enable the devices to survive doses of the order of 10^8 rads/year.

3.4.4 Reliability Requirements

This section will briefly discuss factors associated with the long-term aspects of advanced space vehicle operations, and their impact on a computer memory system.

The outstanding characteristic of projected future space missions is their prolonged operational life of up to ten years. This is a new factor to consider along with the traditional requirement for the utmost in equipment reliability associated with manned space flight. No equipment of any sophistication has been produced with proved failure-free lifetimes on the order of 100,000 hours per unit. It is very unlikely that revolutionary techniques of computer construction will evolve that will enable MTBFs of this order to be designed to and met. Even if such a method were claimed to be developed, it would be very difficult to subject it to realistic evaluation! In other words, space-borne systems with a high probability of continuing operation will probably be developed by application of strategies that are well known today. The crucial factor is of course the "continuing operation" of the system. This can be achieved without imposing a zero failure rate requirement, by combining the failure prone components of the system into a redundant organizational structure, and by imposing a maintenance regime that ensures the highest probability of system operation.

3.4.4.1 Component Reliability

There is, of course, nothing like starting with reliable components in the first place. A system's basic reliability is determined largely by the number and reliability of its components, and partly by the environmental and operating conditions it experiences. Computer memories tend to be regularly structured from a large number of identical elements. Where these elements are physical entities, as for example in the case of the ferrite core, the reliability of the system hinges on the performance of the element. This is especially true in the case of massive memories for which the ratio of peripheral circuit components to storage elements is small. The ferrite core is a simple, passive device for which manufacturing processes have been developed over the years to yield high quality and uniformity. The failure rate for ferrite cores has been estimated to be 2×10^{-7} per 1000 hours [11], which is probably better than the reliability achievable by any other discrete technique. However, memory technologies that employ batch fabrication of the elements will, other things being equal, enjoy an advantage in terms of reliability over discrete techniques.

All memory systems will be composed, in whole or in part, of semiconducting devices. Being active components of much

higher complexity than cores they are more vulnerable to variations in manufacturing and operating tolerances. Failure rates for integrated circuits used in the Minuteman missile and Apollo vehicle guidance systems indicated values between 10^{-4} and 10^{-5} per 1000 hours [12]. These rates are expected to improve as IC technology matures.

A large number of components in a memory system implies a large number of electrical interconnections. The unreliable joint has always been a major source of failure in electronic equipment. It appears that failure rates of between 10^{-7} and 10^{-9} per 1000 hours have been experienced in operational equipment with soldered, wrapped and ball-bonded connections [13]. Much emphasis is currently being placed on the development of batch processing techniques of electrical interconnection (see Chapter 2, Section 2.6.5.3) with the objective of eliminating the variable quality associated with the individually made connection. None of the new techniques has amassed many operational hours from which estimates of reliability can be derived.

To summarize, it is very desirable that a chosen mass memory technology employ a fabrication technique that maximizes the number of elements created in a single manufacturing step, and that the design of the memory arrays and associated circuitry minimizes the number of discrete electrical connections for a given logic or memory function.

3.4.4.2 Organization for Reliability

Several identical unreliable components can be redundantly combined to form a system that has a smaller probability of failure than an individual component, at least for a time. The possible organizations of redundant components and their properties has been discussed in Chapter 3 of the Final Report to the first part of this contract. Briefly, redundant systems can be "open", meaning that failed components are replaced and the system restored to full capability, or "closed", in which failed components are excommunicated and the population of good components is reduced. The closed system has been analyzed at length in Reference [14]. Results showed that a 240% increase in the number of components improved the five year reliability of a modularly organized computer from .018 to 0.67, assuming a logic gate failure rate of 10^{-5} per 1000 hrs, and a standby-to-active failure rate ratio of 0.1. The "open" system is inherently capable of higher reliability, but requires an inventory or maintenance service.

For a memory system consisting largely of a regular, repeated structure, redundant modularity is a natural approach. The question is whether to employ powered-down "standby" components which are switched on upon detection of a failure, or

"always-on" full redundancy. The former takes advantage of the higher reliability of the standby mode, but requires no extra error detection and reconfiguration logic. Full redundancy makes for simpler error detection because it allows majority voting of the outputs of three or more identically operating components. But it has only two levels of operation, making the concept of "graceful degradation" impossible, and requires one more redundant component.

For memory systems error detection can be fairly simple; some of the more sophisticated codes even allow degrees of error correction. Redundant memory modules cannot be left in a "cold standby" without incurring a large initialization transient at turn-on due to the lack of current program data. For the same reason, "graceful degradation" is a difficult concept to implement within a memory system since it involves a complex interface between hardware and software. A good example of a memory system that is configured for reliability without imposing a software penalty is that of the NASA/MIT SIRU computer. One of the two identical memory modules functions as a standby, but contains current programs and data, which are continuously kept up to date with all "write" commands. Error checking is performed on the other, operational, module. Switchover occurs in the space of one instruction cycle as soon as an error is detected in the operational module.

The degree of complexity and duplication within a memory system is a function of the overall space vehicle reliability criterion. The ability to survive one failure to a component in a system and remain operational implies at least dual redundancy at the component level with separate error detection, as in the SIRU computer, or triple mode redundancy with majority voting. Two failure survival requires two spares, or quadruple mode redundancy, and so on.

To summarize, a highly desirable attribute of a mass memory technology would be to allow a highly modular construction with as few, often repeated, elements as possible. Systems involving large overhead items such as, for example, an electron beam generator will be very difficult to configure for long-term reliability, whereas a semiconductor array would be more amenable to such a design.

References for Chapter 3

1. Conti, C.J., Gibson, D.H., and Pitkowsky, S.H., "Structural Aspects of the System/360 Model 85: General Organization", IBM Sys. Jour., Vol. 7, No. 1, 1968, pp. 2-14.

2. Liptay, J.S., "Structural Aspects of the System/360 Model 85: The Cache", IBM Sys. Jour., Vol. 7, No. 1, 1968, pp. 15-21.

3. IBM, Draft of Space Station DMS Requirements, S.O.W. item no. 2.1.2-b, c, d, e; undated, informally transmitted, January 1970.

4. North American Rockwell, Space Division, Space Station Program Phase B. Definition February Progress Review, PDS70-212, 19 February 1970.

5. Larin, F., Radiation Effects in Semiconductor Devices, John Wiley & Sons, New York, 1968.

6. Stanley, A.G., Ryan, J.L.,"Charged Particle Radiation Environment in Synchronous Orbit", MIT Lincoln Laboratory Technical Report #443, 15 May 1968.

7. ----------, "Models of the Trapped Radiation Environment", 5 Volumes, NASA SP-3024, 1966.

8. Barry, A.L., Page, D.F., "Radiation Hardening of MOS Transistors for Low Ionizing Dose Levels", IEEE Trans. On Nuclear Science NS-13, December 1966, pp. 255-261.

9. Hughes, H.L., "Surface Effects of Space Radiation on Silicon Devices", IEEE Trans. on Nuclear Science NS-12, December 1965, pp. 53-63.

10. Gordon, F., Wannemacher, H.E., "The Effects of Space Radiation on MOSFET Devices and Some Application Implications of Those Effects", IEEE Trans. on Nuclear Science NS-13, December 1966, pp. 262-272.

11. Harold, G.M., "Trends in Computer Technology", Proc. 24th Annual Meeting, Inst. of Navigation, June 19-21, 1968.

12. Workman, W., "Failure Modes of Integrated Circuits and Their Relationship to Reliability", Microelectronics and Reliability, Vol. 7, 1968, pp. 257-264.

13. Norris, K.C., et al, "Reliability of Controlled Collapse Interconnections", IBM J. Res. and Dev., Vol. 13, May 1969, pp. 266-271.

14. Erwin, F.D., et al, "Modular Computer Architecture Strategy for Long Term Missions", Proc. FJCC, 1969, pp. 337-345.

Chapter 4

Evaluation and Recommendations

The objective of this chapter is to review the results of the technology investigation of Chapter 2 against the specific requirements of a mass storage system for advanced manned space vehicles defined in Chapter 3; to select and evaluate several technologies for comparison against the specific categories of memory that were identified in Chapter 3; and finally to make recommendations to NASA as to which technologies would prove the most fruitful for further development in each category. In order to come to the final conclusion several guiding factors will be taken into account in addition to the functional requirements of Chapter 3, and some of these are discussed next.

4.1 General Evaluation Criteria

4.1.1 The Development Schedule

The objective of this memory study is to arrive at a recommendation of technologies that can be developed to meet specified performance criteria in a fully operational, space qualified form, in the year 1975, and again in the year 1980. Although these dates appear to be well into the future, experience in the past has shown that in order to produce equipment that must demonstrate advanced performance, possess high degrees of reliability and be able to survive successfully the exigencies of a space vehicle environment, considerable developmental lead time must be scheduled. An example, admittedly fairly conservative, that illustrates this statement is the case of the Apollo guidance computer's read-only memory. The intial conception of the core rope memory took place at MIT's Lincoln Laboratory in the mid-1950's; and it was then developed into prototype form by the Instrumentation Laboratory in the late 1950's. Its use for the Apollo mission was proposed when MIT undertook the guidance task in the fall of 1961, at which time the first operational test was scheduled to occur about two years later, in late 1963. The schedule of test flights leading to the successful lunar landing did not, in fact, start until the fall of 1966, but some measure of the effect of an 8-10 year development cycle is provided by the resulting fault-free performance of the memory system during each subsequent mission. Other instances of development of a

new technology for a specific application can be quoted in which similar levels of reliability have been attained in about half that time.

The implementation of a mass memory system for the next generation of manned space vehicles must be able to fit a minimal development schedule, which will be approximately defined as follows:

1)	From conception of the basic device technology to demonstration of its suitability for a large memory system:	1-2 years
2)	From feasibility demonstration to working prototype memory system:	2-3 years
3)	From prototype to fully space-qualified operational memory system:	2-3 years
	Total	5-8 years

The longer development times will be demanded for the newer and/or more complex memory techniques that may be considered, and the shorter ones for those which are less complex or more directly related to established techniques and upon whose expertise their development can draw. For the 1975 operational date candidate technologies must obviously be in an already conceived state today; furthermore, those which are expected to generate considerable developmental difficulty must be even farther along, perhaps well toward a prototype stage.

Some of the performance requirements for 1980 estimated in Chapter 3 are considerably in advance of the state of the memory art today (or the immediate, foreseeable future). Realization of this performance may have to depend on the results of prolonged development cycles based on techniques that are still very embryonic. Success will depend on the ability to establish significant capabilities at intermediate performance milestones along the development route. For example, if it is found feasible to develop a 10^8 bit semiconductor memory system from storage elements whose capabilities have only been examined in hundred-bit modules, it will be essential first to establish a functioning system of, say, 10^4 bits; from there the program can go on to 10^6 bits, using the results of analysis and tests on the 10^4 bit unit, and eventually work up to the targeted 10^8 bit system. The necessity for a graduated development cycle like this will serve to prolong the period from laboratory to space station for some of the newer techniques that have not, as yet, seen service in even a demonstration memory system. For example, the read/write holographic, the magnetic "bubble" domain,

and the MNOS transistor memory concepts are technologies that fall into this category; their feasibility for implementation by 1975 or 1980 will be judged accordingly.

4.1.2 Commercial Aspects

One strong influence that must be considered in the evaluation of a technology is the interest shown in that technology by the computer industry as it pursues its commercial activities. The fact that memory represents about 50% of the total cost of manufacturing a commercial computer system means that the development of memory techniques for higher performance and lower cost is vital to the computer maker. If development of the chosen memory technique for advanced space vehicles provides to the industry an enhanced capability in the commercial field, generates marketable technical fall out, or produces greater control over, or insight into, current manufacturing processes, then certain intangible benefits accrue to the space application that cannot otherwise be bought:

a) Cost. The cost to NASA of developing a system that is related to the needs of industry will be less in the long run than the cost of wholly supporting the necessary research, development, test and procurement phases of a one-off design uniquely tailored to a specific space requirement. Memory techniques with commercial relevance will benefit from the cost-sharing by industry in terms of mechanized and automated production and extensive test and qualification facilities. Such facilities, it has also been established, will greatly enhance the uniformity, quality and hence reliability of the product, which will be, perhaps, the most important aspect of massive storage facilities for manned space.

b) Reliability. A memory technique that is closely related to one that is already widely used benefits from the availability of operational data recording failure rates, maintenance and repair problems, and so on. This is significant, because the establishment of life time and failure rate information for a novel technique is next to impossible for the long term missions that are being specified. It has been shown [1] that it takes 10 years to demonstrate the attainment of a 10^{-7}/hour failure rate by determining the time to five failures per 1000 devices. To shorten this time to 1000 hours the population of devices must increase to 90,000. Space station failure rate requirements for some components may be several orders more severe than 10^{-7}/hour, and consequently much higher populations are demanded for reliability demonstration. Clearly, it will be prohibitively expensive to assure this re-

liability in the case of unique designs that have no precedent. However, good correlation of field service data may make reliability estimation possible for space systems related to equipment that has seen considerable service.

Finally, commercial interest will serve to maintain a momentum in the development of a particular technique that cannot be assured merely by the availability of adequate funding. It will spark the competitive element that often spurs on a promising investigation and helps to shorten the period between feasibility demonstration and prototype development. Absence of such competition has the opposite effect: even promising avenues of research may eventually peter out. This has been the case, for example, with the acoustic film memory, where the one remaining developer, Sylvania, is questioning the advisability of continuing its effort. Competition has always served to keep down the customer's costs. The burgeoning semiconductor memory field is a good example of how the jostling activity of the many potential vendors has been the cause of a monotonic decline in the price per bit, even though the performance, in terms of capacity, speed and density, has continued to improve.

4.2 Evaluation Strategy

In order to determine which of the many technologies reviewed in Chapter 2 would be most suitable for development to meet the requirements for mass storage detailed in Chapter 3, some kind of evaluation strategy must be employed. The most natural sequence of decision making would seem to be the following:

1) For each of the memory categories defined select a number of candidates which from their performance characteristics and developmental status would appear to be the most promising for the two time periods;

2) Conduct trade-off comparisons between technologies competing for each memory category, using a more specific set of evaluation criteria, to determine the most suitable;

3) Make a final judgement by weighing the more promising techniques against the less tangible criteria that were outlined in the previous section, for example, commercial attractiveness, cost, etc.

Although this sounds like an orderly sequence, in practice it breaks down somewhat because of disparities among both the requirements and the characteristics of candidate technologies, because what is a significant evaluation criterion

for one technology has no meaning for another, because sometimes the intangible judgement criteria have overwhelming importance, and so on. The strategy to be employed in the next section is to take each category of memory, M1, 2 and 3, and attempt to apply the above sequence of evaluations wherever possible. However, it is expected that most of the technologies will be assessed, each on its own merits, with evaluation being based on outstanding characteristics which may, in some cases, be unique to that technology.

4.3 Comparison of Memory Technologies

In the following sections, each of the three memory categories that were defined in Chapter 3 will be addressed in turn and a most suitable technology for each will be assessed.

The criteria for judgement fall into the two categories of specific and qualitative. Specific parameters are those to which numerical, or definitive values can be given. Examples of numerical quantities are absolute capacity, speed, weight, volume, power consumption, and density in terms of bits, watts, and pounds per cubic inch; non-numerical but specific properties are volatility, destructive or non-destructive read-out, moving parts or solid state, and so on. Qualitative parameters are those for which only a high, medium, low or a good, bad or indifferent evaluation can be given. Examples of these are resistance to environmental stresses (temperature, pressure, acceleration, radiation, etc.), reliability, maintainability, flexibility, and so on.

4.4 Comparison of Technologies for M1

The high-speed memory M1 as defined in Chapter 3 is a random-access buffer between the processor unit and the larger working storage of M2. It is capable of read and write operations; it must service an interface several words in width with the processor, and another, narrower interface with a high speed common bus; it must be addressable to the level of a single word. Volatility and permanence are of low significance; high speed and random access are essential.

4.4.1 M1 in 1975

M1 performance requirements for 1975 are repeated here in summary:

Capacity:	10^5 bits
Access speed:	100-200 nanoseconds
Data rate:	5 $x10^8$ bits/sec

The physical parameters quoted in Chapter 3 include both M1 and M2. Both are of similar organization, i.e., randomly accessed; but, although M1 is only one hundredth the capacity, it is twenty times as fast. Assuming therefore that volume and power are linearly related to the speed-capacity product, a five-to-one division of space and power will be postulated. Approximate physical limitations for M1 are therefore:

Volume:	300 cubic inches
Power:	15 watts
Weight:	15 pounds

The significant parameters for the technology now become:

Data density:	330 bits/in^3
Dissipation:	1.5 $x10^{-4}$ watts/bit

Although the density figure represents an overall figure including space for circuitry and logic to perform address decoding, selection, driving, sensing and control, it is still a fairly easy achievement for most technologies. A far more restrictive requirement is the dissipation limit. The power consumption of a memory system can vary between very wide limits, depending on speed and utilization. For the M1 memory a fairly constant duty cycle of between 50% and 75% is assumed reasonable; it is at this level of utilization that M1 should not dissipate in excess of 15 watts or so. Since volatility and retention of data are of little consequence here, a low standby power is of no value.

There are today only a few memory technologies that are able to penetrate the region of access times below 200 ns. Of these only two, thin magnetic film and bipolar integrated circuit,are known to have been incorporated in computer memory systems with a capacity of 10^5 bits or more, and cycle times less than 200 ns. Examples of such applications are the thin film memories designed into many of the Burroughs computers, and the semiconductor buffer memories specified for the IBM 360/85 and the current 370 series of computers. However, in all instances the volume of the memory system was not a critical factor, and space in terms of cubic feet rather than cubic inches was consumed. Furthermore, since the applications were commercial, power dissipation was of less consequence than performance. The power dissipations

were up to two orders of magnitude in excess of the 15 watts that is specified in the case of M1.

The specified packing density and power dissipation have been approached, but at not quite the specified access times, by a number of militarized memory systems in plated wire and the "post and film" variant of thin-film techniques. These have exhibited cycle times between 0.5 μs and 1.5 μs with access times as low as 250 ns (see Sections 2.3 and 2.4).

A number of candidate technologies, therefore, are capable of being developed to the level of performance specified here for an operational memory by 1975. Of the magnetic techniques, ferrite cores must be dismissed, because the combination of smaller dimensions and partial switching necessary to attain speeds below 200 ns are considered to be disadvantageous in comparison to the possibilities of planar and cylindrical magnetic films. Furthermore, the low power requirement makes NDRO almost a necessity, and this is more easily achieved by thin film switching than by cores. The low power requirement also favors techniques which couple the storage element very closely to the field generated by the drive lines, in a manner exemplified by plated wire. Some planar film variations, such as the coupled film and the post and film techniques, also possess efficient field utilization, but are considerably more complex in design and therefore expensive to manufacture. Ferrite in planar form has been used as a keeper to obtain a more concentrated effect of the driving fields by reasons of its higher permeability (Sections 2.2.3, 2.2.4), and as a storage element (Section 2.2.5). But it is considered that memory arrays with a high ferrite density suffer from the loading effect which the increased dielectric constant and permeability have on the capacitance and inductance of the drive and sense conductors, and this route to low power will find difficulty in achieving the specified access times.

Of the semiconducting techniques it is probable that neither the single channel MOS nor the bipolar integrated circuit array will be able to achieve 100 ns at 10^{-4} watts per bit. The speed is determined to a large extent by the distributed L and C of the current conductors, whose effect can be minimized by low driving and terminating impedances. However, the dissipation of the back-coupled flip-flop circuit commonly used as a storage element increases as the value of the load resistor is decreased to lower the RC time constant. With current bipolar and MOS integrated circuit practices such circuits must dissipate between 10^{-3} and 10^{-2} watts per bit to achieve 100 ns system cycle times. Although dynamic MOS storage cells can dissipate less than

10^{-4} watts per bit, the speed capability will probably remain around 0.5 μs to 1.0 μs. Lower distributed capacitances will only be achieved by smaller device and conductor geometries such as may be made possible by electron beam techniques. It is doubtful whether such practices will be standardized in time to be established by 1972 or 1973, which would be necessary to achieve a 1975 operational memory system of acceptable reliability and cost. The circuit power dissipation can be greatly improved by reducing the operating potential, since the power loss incurred in charging and discharging line capacitances is proportional to CV^2. Further improvement is effected by making the circuit impedances non-linear: low for the charge/discharge sequence, and high for the intervening quiescent periods. Both these factors can be achieved by complementary transistor logic circuits (see Section 2.6.3.4), which have the additional desirable attribute of considerable noise immunity. Complementary MOS is therefore a very attractive candidate for this application. However, although MOS and CMOS memory techniques are being actively investigated and applied at the present time, fundamental device behavior and satisfactory processing techniques have still some way to go before becoming as universally accepted and reliable as some of the magnetic techniques.

4.4.2 Recommendations for M1 for 1975

From the discussion of the previous section it appears that magnetic techniques possess a slight edge over semiconductors in terms of their lower power dissipation. It is considered very feasible that the speeds of which plated wire and other planar films are capable will match the required 100-200 ns within two to three years, whereas it is considered less likely that semiconductor dissipation will drop to the required 10^{-4} watts/bit at these speeds (except for the less well-developed CMOS technology).

Of the thin-film techniques first choice must go to plated wire. The strong background of industry investment and the expected widespread application make for a very healthy and cost-effective situation for NASA. The development of the smaller 2 mil wire by Honeywell is a very promising route to high speed, low power and high density. The NDRO capability allows a pseudo-2½D organization to be used for the basically 2-D technique, and this promises a reduction of power dissipation in the sense electronics, besides eliminating the write-back operation of DRO. A second choice would be the planar post and film technique, primarily because of its advanced status (being currently in production for a military computer) and relatively simple batch-oriented construction. Whether the extensive planar ferrite keeper will prevent access times of 100 ns from being achieved is the problem to be faced by this approach. No semiconductor

technique is recommended for a 1975 implementation primarily because there is doubt whether the specified speeds can be achieved at the required limited power dissipation. Only complementary MOS techniques are seen as a realistic solution to this problem but are considered to be lagging the comparable magnetic thin film device technology and are not recommended for a 1975 operational 10^5 bit memory.

4.4.3 M1 for 1980

For a 1980 operational date the M1 performance requirements are:

Capacity:	10^6 bits
Access speed:	40-60 nanoseconds
Data rate:	10^9 bits/sec

Volume and power dissipation are specified to remain unchanged from the 1975 figures. The resulting significant technology parameters then become:

Data density:	3,300 bits/in^3
Dissipation:	1.5 x10^{-5} watts/bit

As in the case of the 1975 parameters, the more significant item is the power dissipation per bit. Of the technologies discussed in the previous section only complementary MOS integrated circuits appear to be capable of microwatt-per-bit average dissipation figures at the specified access times. At capacities in excess of a million bits, power losses caused by the charging and discharging of drive and sense line capacitances becomes even more significant than for 10^5 bits. CMOS is able to offer an attractive solution to this problem. The development of dielectric isolation techniques such as silicon-on-insulating-substrate processes (as described in Section 2.6.3.4) will help to reduce the on-chip capacitances, minimize leakage currents, and assist CMOS in achieving the specified high speed and low dissipation.

No inductively sensed magnetic technique can be visualized as realizing these performances.

4.4.4 Recommendation for M1 for 1980

Of the known memory techniques it is recommended that the complementary metal oxide semiconductor technology be developed in conjunction with techniques to reduce on-chip capacitances and quiescent leakage currents.

4.5 Comparison of Technologies for M2

M2 is the main working memory of the hierarchical system defined in Chapter 3. It is characterized as a large, random access, read/write storage element. The minimum randomly addressable unit is a block of several words totalling 250 to 400 bits. Since M2 is the primary store for all operating software, non-volatility of the stored information is moderately important.

4.5.1 M2 in 1975

The performance requirements for M2 in 1975 are summarized from Chapter 3 as follows:

Capacity:	10^7 bits
Access time:	2 μs - 4 μs
Data rate:	5 $\times 10^8$ bits/sec

The physical specifications, after the calculations of the previous section defining the dimensions of the M1 unit, are given by:

Volume:	1500 cubic in. (approx.)
Weight:	85 pounds
Power consumption:	85 watts

These figures yield the following significant technology parameters:

Data density:	6700 bits/in^3
Dissipation:	8.5 $\times 10^{-6}$ watts/bit

As in the case of M1, the bit density is not a severe requirement even by today's standard practices. If memory planes can be stacked at 5-10 to the linear inch, then the required area density is only 700-1500 bits/sq.in., which can readily be achieved by a number of techniques.

The dissipation limitation is, however, a severe requirement to place on this size of memory. It represents an improvement of between one and two orders of magnitude over the implementation of an equivalent system by today's ferrite core technology. Other technologies that could be considered for the implementation of M2 by 1975 are listed below with estimates of the power dissipation per bit that each would demonstrate in such a memory today.

Technology	Power watts/bit
Ferrite core	3×10^{-4}
Plated wire	10^{-4}
Planar film	10^{-4}
Laminated ferrite	10^{-5}
Bipolar IC	2×10^{-3}
Dynamic MOS	10^{-3}
CMOS	10^{-4}
Sonic BORAM	10^{-6}

It is obvious that M2 cannot be implemented at the level of development that most of these technologies have reached today. Although a serially accessed, block-oriented technique is included in the above listing to demonstrate its low dissipation in comparison with the others, it is not conceivable that this type of memory can achieve the random access times of 2 μs to 4 μs required here. These speeds imply shift frequencies in excess of 10^8 Herz, or a complex structure of short shift registers, either of which would erode the BORAM's advantage over other techniques.

Since there are no obvious candidates today, the recommendation of a technology for M2 is obviously a matter of selecting that which is considered to be the most promising, most amenable and most cost-effective for development to the required level of performance.

Of the magnetic techniques ferrite cores will again be dismissed from consideration because faster techniques with lower power requirements are available, because NDRO is not possible in a straightforward fashion, and the large drive currents needed make integrated circuit operation a difficult, if not an impossible task. Since the access time requirement is greatly relaxed compared to the M1 evaluation, the thin film methods, which were discounted for M1 because of their reliance on use of ferrite keepers to reduce drive current requirements, may prove to be more promising contenders for M2. However, the planar techniques tend to suffer either from poor density or from a complicated (and therefore expensive) structure. They must again be rated lower than the more promising plated wire. The laminated ferrite technique, which was reviewed in Section 2.2.5, must be considered a contender, since it has been developed with a view to implementing a 10^7 bit, 2 μs memory with dissipation less than 100 watts, very close to the 1975 specification for M2.

Of the semiconducting techniques bipolar ICs must be eliminated, because access time requirements measured in microseconds rather than nanoseconds provide MOS techniques a far easier path to the achievement of dissipations less than 10^{-5} watts/bit. Dynamic MOS might be a contender, because of its low standby dissipation, but CMOS is considered to possess the greatest potential of development for lower dissipations of all the current semiconductor memory techniques. The overwhelming factor which will tell against most semiconductor approaches, however, is the importance of non-volatility to the operation of M2. Although manufacturing proponents of MOS arrays counter the argument of volatility with suggestions of self-contained batteries as a back-up in case of power loss, this is not considered to be a satisfactory solution for reasons of lowered reliability, increased maintenance, and less resistance to environmental stress. There is one semiconductor technique, however, the MNOS concept, that offers the prospect of a non-volatile semiconducting storage element, besides certain other advantages, but it is felt that for 1975 this technique will not be sufficiently well established to bear successful comparison to other techniques.

Magnetic techniques have an advantage over semiconducting arrays in terms of their resistance to radiation effects, which is an important factor for M2 if lack of it jeopardizes the integrity of the stored information. Semiconductors are, of course, a constituent of any memory system, but their vulnerability to radiation becomes a serious hazard to long term reliability when they are also employed as the storage element.

The drawback of the magnetic (or rather, non-semiconducting) approach is that there is no possibility of integrating the address decoding logic with the storage elements. This implies a large number of interconnections and subsequent increases in physical complexity and overall unreliability. This consideration would tend to favor the technique that yields the largest number of bits per manufactured element. For example it has not been found possible to produce laminated ferrite arrays in larger than one square inch planes. One plane can only accommodate about 4000 bits in a 64x64 coincident array, and requires at least 128 external connections. Plated wire arrays on the other hand have been manufactured in arrays with four times as many bits, and only twice as many connection points. All other factors being equal, this would constitute an advantage over laminated ferrites. Serial block-oriented devices have a great advantage in terms of a minimal number of connections, but none is known to be fast enough to provide random access to a 250 to 400 bit block in 2 μs to 4 μs.

4.5.2 Recommendations for M2 for 1975

In summary, a magnetic technique is considered to be a more feasible prospect than semiconductors for development as the 1975 M2 technology, because of the volatility, vulnerability to radiation, relatively high power dissipation, and embryonic status of the existing semiconducting techniques. Of the magnetic technologies plated wire is considered to be the most promising candidate for the following reasons:

1) The performance of plated wire memories is closing in on the required levels specified for M2. The advent of smaller 2 mil wires promises reduced physical dimensions, lower drive currents, and IC operation for the supporting electronics. The technique certainly has a more than adequate speed margin. The task will be to convert this into an attribute for power reduction, possibly by the extensive employment of higher permeability keepers.

2) Manufacturing techniques for plated wire have been developed over a considerable period of time and are now well established. Continuous automatic production and testing will ensure high quality, uniformity and hence reliability. In this respect plated wire scores over other techniques which might otherwise be equally promising. For example, laminated ferrites have good performance potential but present severe production problems (see Section 2.2.5).

3) Plated wire is being energetically developed by government and industry for both commercial and military applications, and this provides the desirable momentum and cost effectiveness indicated in Section 4.1.

4.5.3 M2 for 1980

The 1980 performance requirements for M2 are:

Capacity:	10^8 bits
Access speed:	0.5 - 1.0 microseconds
Data rate:	10^9 bits/sec

The volume and power dissipation figures remain unchanged from 1975, and therefore yield the following:

Data density:	7×10^4 bits/in^3
Dissipation:	8.5×10^{-7} watts/bit

This set of specifications is probably the most formidable of

any to be considered during the course of this study. The volumetric packing density is the least demanding of the requirements, and yet it imposes an area density of between 10^3 and 10^4 bits/in^2, depending on the array stacking capability, which is beyond the reach of today's inductively coupled magnetic techniques (excepting perhaps that of the magnetic domain, or "bubble", shift register). It is, however, well within the capability of MOS semiconductor arrays, and all of the optically accessed storage media. The "bubble" concept is eliminated for the same reason that the sonically scanned BORAM was for 1975, namely that it is basically a serial device, and access in one microsecond to individual blocks as small as 400 bits demands either an extremely high shift frequency, or a very complex structure of short shift registers. Beam-scanned media must also be ruled out, in spite of their attractive storage densities, because the 10^9 bits/sec data rate implies a parallel read out capability of several hundred bits at a time. It is considered very undesirable, if not impossible, to modulate a single beam at 10^9 hertz, and in addition perform series-to-parallel conversions of the accessed data at that frequency.

4.5.4 Recommendations for M2 for 1980

The most promising approach to this problem is felt to be the combination of the MNOS storage element with CMOS decode, drive and sense circuitry, preferably on the same chip. MNOS storage cells are already achieving read times of 50-100 nanoseconds, and dimensions of less than 1 square mil (10^{-6} in^2). Semiconductor processing improvements are expected to realize device geometries down to 10^{-1} sq. mil. Single-chip arrays of well over 10^4 bits are being confidently forecast within the next year or two, so that by 1980 modules of this capacity should have achieved very acceptable levels of yield and reliability. The resulting high area density is not of importance in terms of the overall volumetric efficiency, but it makes feasible reduced cell- and conductor-to-substrate capacitances, and higher numbers of cells per drive line. These are essential factors in reducing the power consumption while retaining the capability of accessing data in 0.5 μs to 1.0 μs. Development of insulating substrate techniques will further help to minimize power requirements by maintaining minimal stray reactances.

The most significant factor in favor of a semiconducting approach to a 10^8 bit random access memory is the possibility of integrating the decoding logic with the storage element. CMOS is chosen as the logic circuit technology for the reasons of high speed and low dissipation discussed in the previous section on M1. This memory will accommodate up to 400,000 addressable pieces of information, which means that the address

decoding logic circuitry will make a significant contribution to the memory cycle time, and it must therefore be capable of logic stage delays of only a few nanoseconds. The common heritage of MNOS and CMOS will be of advantage in exploiting their coexistence on the same chip. Although there are differences in the processing steps of CMOS and MNOS fabrication, there is sufficient similarity to make their integration quite feasible. The insulating substrate technique could further facilitate amalgamation of these processes by minimizing the interaction between the memory and logic devices. Not only will the resulting array exhibit lower delay times, but the reduction in the number of interconnections will be a major contribution to system reliability.

The MNOS/CMOS combination also possesses the advantages of non-volatility. MNOS devices are already exhibiting retention times measured in thousands of hours. By 1976 to 1978 this could very well be increased to a few years, without increasing write times beyond a microsecond, and would in effect constitute virtual non-volatility. MNOS has the additional property of being relatively radiation resistant, due to the silicon nitride gate dielectric (see Section 3.4.3.2). It is felt that these attractive prospects make this technology an outstanding candidate for M2 by 1980.

4.6 Comparison of Technologies for M3

M3 was defined in Chapter 3 as a read and write, block oriented, massive storage device of moderate access speeds. It serves as the permanent repository of both programs and data, and as a recording device for the vast amounts of data expected to be generated by future advanced manned space vehicle systems. A highly desirable feature of an M3 technology would be ability to allow the physical removal and replacement of portions of the stored information. Permanence and integrity of the stored information are of the utmost importance.

4.6.1 M3 in 1975

M3 operational performance requirements for 1975 were defined in Chapter 3 as:

Capacity:	10^9 bits
Access time:	10-100 milliseconds
Data rate:	10^5-10^6 bits/sec
Block size:	10^4 bits

The physical requirements were defined as:

Volume:	2-4 cubic feet
Power:	200-400 watts
Weight:	200-400 pounds

Expressed in terms of the significant technology parameters, these values yield:

Data density:	3-6 $x10^5$ bits/in^3
Dissipation:	2-4 $x10^{-7}$ watts/bit

Limiting the power dissipation to the submicrowatt/bit region confines consideration of the electrically or magnetically accessed storage methods to those which are block oriented. It is improbable that any but the magnetic "bubble" techniques will be able to achieve overall densities up to a million bits per cubic inch. The latest work on the latter technology has reported a breakthrough to area densities of up to 10^6 bits/in^2. Shift frequencies up to ten megahertz have also been reported, making possible the 10 ms access times. These factors contribute to making the "bubble" a strong contender for this task. However, the technique is very much in an early developmental stage, and no application to an operational memory has been reported to date. Furthermore, there appear to be severe manufacturing problems which restrict the size of acceptable wafers to less than a square inch, and therefore raise the problem of interconnections. Once the materials problem has been brought under control, this technique will be able to inherit many of the well-developed production methods of the semiconducting industry, due to strong similarities in processing, handling, and interconnection. Accordingly, the "bubble" domain shift register will be considered as a strong back-up for a 1975 M2, in the event the recommendation that follows proves unrealizable.

In the past memory systems of capacity in excess of 10^9 bits have invariably gained access to required blocks of data by a combination of moving storage media (e.g., rotating disks and drums) and moving read-write mechanisms. A fixed-head drum such as the IBM 2301 unit can provide access to information in less than 5 milliseconds, but it is limited in capacity to 3 $x10^7$ bits, and exceeds by orders of magnitude the physical parameters specified for M3. A moving head design adds 50-60 milliseconds to the access times, most of it contributed by the need to accelerate and decelerate the arm supporting the read/write head, but a x10 increase in storage capacity is thereby realized. The IBM 2314 can, in its full configuration of eight 2316 disk pack units, provide access

to 2 x10^9 bits in about 60 milliseconds. This would satisfy the M3 performance requirement; but the 2314's volume of several hundred cubic feet, weight of thousands of pounds, its kilowatt power requirement, and the need for a concrete floor and an airconditioned room are not encouraging indications of the potential of this technology for a space environment.

The basic reasons for this bulk is the limitation of inductively sensed magnetic storage. Head gaps and head-to-surface spacing are limited to about 0.1 mil: dimensions less than this bring about an increase in surface and head wear. The resulting linear recording density is about 2,200 bits/inch. The accuracy with which moving heads can be positioned limits track spacing to about 100 to the inch. The net result is a limiting area data density of about 2 x10^5 bits per square inch [2]. Since the storage medium must be rotated rapidly and very accurately, implying massive support and drive mechanisms, and because it is necessary to provide access to moving heads, the volumetric efficiency of this technique is very low.

The most desirable arrangement for memories of this capacity is to accommodate all the information in such a manner that access to it can be effected by some interrogating device of low physical or electrical inertia without involving motion of the storage medium. Photon and electron beams provide the capability of very low inertias; they can be modulated at megacycle frequencies; they are able to resolve objects 1 to 10 microns in diameter. These are attractive properties. The application of beam technology to the accessing of high density information has its problems, however. The information is most easily accessed by beam techniques in a two-dimensional arrangement. If bits can be spaced on 10 μ centers, an area density of about 6 x10^6 bits/in^2 could be realized. Even so, the accommodation of 10^9 bits would require 170 square inches of storage medium. It is extremely difficult to design solid-state deflection schemes to allow a photon or electron beam to scan an area as large as this with micron accuracies (see Section 2.7). An electron beam is simpler to deflect than a light beam, but it is basically an analog operation, limited to 10^{-1}% to 10^{-2}% accuracies. Light beam deflection is very difficult, and its techniques are still in early stages of development. The combination of electron and photon beams, to avoid the drawbacks of both, has been proposed, but has not seen widespread acceptance.

A relaxation of the deflection problem is obviously brought about if the packing density could be increased, thus reducing the dimensions of the storage medium that must be scanned. This can be achieved by going to three-dimensional storage, with 10-100 micron spacing in all dimensions, or by decreasing the two-dimensional spacing by a factor of, say, 10. The multidimensional storage of images in photochromic

crystals has been experimentally studied (see Section 2.7.2.3), but has met with little success, and is not seen as a technique in the immediate future. The second approach is a feasibility, because the recording of spots only a few wavelengths of light in diameter has been proved possible. However, if bit spacing down to 1 μ is considered, then the memory structure must possess a dimensional stability measured in microinches, and fabrication would require machining accuracies of the same order. It is doubtful if such equipment could be built, let alone be qualified to fly in a space vehicle by 1975. Holographic techniques, with their property of providing a degree of immunity to dimensional variations, may in time provide a solution to this problem. Holographic memory systems have been proposed, and some read-only versions have been actively developed for mass storage, but no read and write holographic memory is known to have been constructed as yet. There are severe difficulties in generating and storing holograms of digital information. The necessary "page composer" required to convert a 10^4 to 10^5 bit block of data into the optical form from which a hologram can be created has not been satisfactorily implemented to date.

In summary, the combination of very high data density (up to 10^8 bits/in^2), solid-state deflection, and holographic techniques is an impossibility for a 1975 operational status.

4.6.2 Recommendations for M3 for 1975

A more conservative approach than the all solid-state optical memory is considered to be a wiser course of action for an implementation of M2 by 1975. The step from inductively-sensed to optically-sensed magnetic recording is considered essential in order to realize higher densities. But it is proposed that mechanical rotation of the storage medium be retained to alleviate the addressing problem. Although the undesirability of moving parts for a space-borne memory system has been established, it may be that for 1975 it is a case of moving parts, or no M3!

There is considerable precedent for the fixed-head rotating disk in memory devices for aerospace applications. Many have been produced, albeit with moderate capacities of a few million bits, to survive the environments of military vehicles. A single axis drum, continuously rotating at a uniform rate, suspended by an enclosed gas bearing might be expected to possess a lifetime of tens of thousands of hours. The problem of head and surface wear to which high density inductive devices are prone, would be eliminated by an optical sensing technique: a highly directable light beam would allow the head-to-surface separation to be increased.

Of the techniques investigated for optical, erasable, two-dimensional storage of digital information, the Curie point writing of thin manganese bismuth films seems to have potential for this application. It allows spots approaching 1 μ to be defined, it has moderate write energy requirements, and it provides permanent storage. It can be read by the Faraday or Kerr effect, for both of which it shows high coefficients (10^5 and 10^4 degrees per cm).

The overall M3 volume of 2 to 4 cubic feet could accommodate a rotating surface of several hundred square inches. This allows a recording density of 10^6 to 10^7 bits/in^2, which is just over an order of magnitude denser than current inductive magnetic recording. Data could be separated by 0.25 to 0.5 mil along the track and tracks could be separated by about 0.5 mil. Each track can be completely scanned in 10 millisecs, at 100 revolutions per second.

A major difficulty will be the deflection of a light beam to scan over the recording area. In this case, however, deflection in only one dimension is called for, since rotation provides access in the other. Although an attractive approach might be the employment of integrated arrays of light emitting diodes, one diode to each track, with sensors on the same substrate, MnBi may require too much write energy to allow reliable diode operation. Other, lower energy Curie point materials do exist, for example, EuO. This material requires low temperature operation, however, which is considered highly undesirable. The alternative source of photon energy is, of course, the laser. Some versions have been specifically developed for long term space usage, and no special difficulty is expected with fragility or unreliability.

Deflection systems can be of very low inertia since no massive heads are involved; piezoelectric deflectors have been investigated, and have shown responses of less than one millisec.

The chief problem is still one of dimensional accuracy and stability. The motion of storage medium and its substrate relative to the interrogating beam will have to be constrained to less than 10^{-4} inch. Some degree of servo positioning by using control tracks on the rotating surface is a possibility.

To summarize, the implementation of a 10^9 bit, read and write memory, with access times to 10 millisec, in a volume of 2 to 4 cubic feet, is without precedence. The inadequacy of inductive magnetic recording is clear. Of the many optical phenomena and techniques being investigated, it is considered to be inadvisable to bring more than one new technique to an operational status the first time optical storage is attempted, especially for the difficult environment of space. Accordingly, one new magnetooptic technique, that of Curie point writing by

a modulated light beam, and Faraday (or Kerr) effect reading in manganese bismuth is proposed. It is proposed to reduce the deflection problem to one dimension by mounting the storage media on a rotating surface of several hundred square inches. Deflection speeds and accuracies can be higher than for the equivalent moving head discs of inductive magnetic devices, because the inertia of the deflecting mechanism can be much lower. Reliability and freedom from maintenance is expected to be high because of simple construction and the absence of wearing processes.

The use of optical techniques to achieve densities of 10^6 to 10^7 bits/in^2 has been established; commercial products are available using a laser beam to create micron holes in a plastic film. But the other elements of the proposed optical memory have not seen system operation. This recommendation is not therefore based on an evaluation of existing equipment, and must be vulnerable to the criticism of potential manufacturers.

A second choice for a 1975 M3 will be the magnetic "bubble" domain shift register. Althought neither this technique nor the optical memory have been implemented to date, it is felt that the optical approach presents less of a technological breakthrough in the available time than the satisfactory development of "bubble" materials and methods.

4.6.3 M3 for 1980

M3 operational performance requirements for 1980 were defined as:

Capacity:	10^{10} bits
Access time:	10 milliseconds
Data rate:	10^7 bits/sec
Block size:	10^5 bits

Volume, weight and power were unchanged from 1975, yielding the following:

Data density:	3-6 x10^6 bits/in^3
Dissipation:	2-4 x10^{-8} watts/bit

It is probable that the specified dissipation per bit will exclude all but the optical technologies. It will be difficult even for the "bubble" domain shift register, which must operate by the control of local fields generated by current-carrying conductors. Their distributed inductance

and capacitance will make operation at shift frequencies of tens of megacycles,without exceeding the required 10^{-8} watts/ bit dissipation levels,very difficult.

Technological forecasts for a period ten years ahead must amount more to conjecture than deduction, especially when the present stage of development is no better than a concept. This is the case for holographic memory techniques, which by arguments put forward in the previous sections, must be proposed for the implementation of this class of memory.

Breakthroughs on several fronts must occur during the next three to five years to allow realization of 10^{10} bits in 4 cubic feet by holographic techniques. The least of them is the ability to store the information so densely that it does not occupy more than a few tens of square inches, i.e., the achievement of densities better than 10^8 bits/in^2 is required. Secondly, the capability of deflecting a coherent beam of light to scan the stored information must be developed. This will demand positional accuracies of at least 1 in 10^4, and to conserve space, deflection angles up to 90° must be made possible. Thirdly, a technique must be found of implementing the "page composer", so that blocks of 10^5 bits can be converted into holographic images and stored in the medium. This device has had no known implementation to date. (Magnetic "bubbles" in translucent orthoferrites have been suggested as a possible technique for this purpose.) Fourthly, a detection array must be developed to convert the real image formed from the decomposed holographic record into electrical signals. Some work on integrated circuit photo diode arrays has been done by Bell Telephone Laboratories. It would appear that the charge-coupled device described in Section 2.6.3.5 has application as an image detector. If this element were to be combined with the page composer it would obviate the need for an auxiliary buffer to accommodate the block during write operations; and it could, incidentally, enable considerably faster access times than the specified 10 ms to be achieved.

This is a formidable array of necessary progress. The detailed recommendation of individual implementation is felt to be not meaningful, because the conjectural nature of the whole system would require an overall design study to evaluate the most effective individual techniques.

4.6.4 Recommendations for M3 in 1980

As stated in the previous section, only the concept and not the specific technology of the holographic memory will be recommended for the implementation of M3 by 1980. Recommendation of the individual technologies for the elements of the

system is felt to require design exercise of wider scope than this study.

4.7 Summary of Recommendations

The following is a tabulated summary of the results of the foregoing evaluation.

Memory Category	Operational Date	Significant Characteristics	Recommended Technology
M1	1975	10^5 bits; 100-200 ns; 15 watts	Plated wire
	1980	10^6 bits; 40-60 ns; 15 watts	CMOS
M2	1975	10^7 bits; 2-4 μs; 85 watts	Plated wire
	1980	10^8 bits; 0.5-1.0 μs; 85 watts	MNOS/CMOS
M3	1975	10^9 bits; 10-100 ms; 200-400 watts	1. Optical drum 2. "Bubble" domain
	1980	10^{10} bits; 10 ms; 200-400 watts	Holographic

References for Chapter 4

1. See reference 12, Chapter 3.

2. See reference 22, Section 2.7.