Active and Passive Elec. Comp., 1998, Vol. 21, pp. 147-165
Reprints available directly from the publisher
Photocopying permitted by license only

SEMICONDUCTOR MEMORIES

H. VIRANI

Kassam Bros. Electrical LTD Unit M 20 Mineola Rd. East Mississ Auga L5G 4N9 CANADA

(Received 30 September, 1996; In final form 3 January, 1997)

1 INTRODUCTION

There is a vast amount of development work proceeding in research establishments of competing semiconductor manufacturers, the aim of which is to discover the ideal electronic memory device. It would be cheap, small, fast, use little power, and retain its data indefinitely in the event of power supply removal. Some companies have a commitment to bipolar technology because of historical development of the company and past investment of capital whereas others are concentrating on developing unipolar technology. This has lead to competing claims of the superiority of a technology for a given application.

Table I summarises the present (1996) position of the competing types of logic circuit.

In general it will be seen that the desirable attributes are not obtained in one type of device and compromises must be made for a given application.

Although bipolar technology has the advantage in speed, MOS is generally superior with respect to power consumption, functional density, and process complexity. The advantages of MOS are particularly important in achieving very large-scale integration (VLSI). Typical upper limits of bipolar processes, such as transistortransistor logic (TTL) medium speed, are about 60 gates/chip, whereas in custom dynamic p-channel MOS (PMOS) logic, densities of 250 gates/ chip are common. VLSI offers excellent benefits toward achieving high reliability, small size, and light weight.

TABLE I

Parameters							
Process Type	*Circuit Type*	*Nominal Stage Delay (ns)*	*Power Dissipation per Gate @ 1 MHz (mW)*	*Noise Immunity (V)*	*Area per Gate (sq mils)*	*Process Steps*	*Technological Malurity*
TTL	Low Power Static	25	1.2	0.4	–	60.65	Mature
TTL	Medium Speed Static	8	7.5	0.4	104–115	60–97	Mature
TTL-Schottky	Static	3	18	0.3	126	60–74	Mature
High Thrshld PMOS	4-Phase	90–150	0.4	2	10–15	30–40	Mature
Low Thrshld PMOS	4-Phase	90–150	0.2	1	10–15	30–40	Mature
PMOS Silicon-Gate	2-Phase	40–50	0.16–0.3	0.5–0.8	10–15	29	Mature
NMOS on-Sapphire	Static	15	0.3	1	5–10	30–40	Relatively now
NMOS Silicon-Gate	2-Phase	40–80	0.17–0.28	0.4	12–14	23–34	Mature
CMOS	Static	12–17	0.3	3.5–4	15–35	37	Mature
CMOS-on Sapphire	Static	6–10	0.15–0.2	3.5–4	15–25	30–40	Relatively now

Note: Data are based on 2-input NANO gate with fanout of 3.

Reliability of MOS is improved in two ways. First, VLSI reduces the number of inter-chip connections – a major factor contributing to integrated circuit (IC) failure rates. Second, the low power attributes of MOS tend to keep chip temperatures low for a given gate density, minimizing adverse reliability effects of high-temperature operation.

For semiconductor memory systems, high packing densities are required. In bipolar devices, the main development effort is being concentrated on increasing the packing density of devices on a silicon slice. MOS transistors require fewer manufacturing processes but their greatest advantage is that they are automatically isolated from other transistors on the slice, since if the substrate of the device is connected to the most negative voltage (for an N-channel device) all the individual N-diffusions are reverse biassed. To summarize, for semiconductor memory systems, MOS technology is the obvious choice because:

a. smaller number of production processes and hence a possibility of a greater yield than with bipolar devices;
b. a smaller area on the chip is needed for the active device;
c. the transistors are self-isolating.

Due to (b) and (c), the packing density of the basic cells on a single substrate is much greater than is at present possible using bipolar technology.

The two main drawbacks of unipolar systems are they are slower in operation than bipolar, and they require higher operating voltages, which make them incompatible with bipolar elements unless some form of interfacing is used.

Because of the advantages of MOS technology the rest of these notes will concentrate on MOS types.

2 CLASSIFICATION OF MEMORIES

All the memories to be discussed are Random Access Memories in the sense that any location can be accessed at random taking the same time as any other location. It is accepted practice, probably due to the historical development of these devices, to reserve the term "Random Access Memory (RAM)" for the fastest read/write memories and to use other terms for other types.

Semiconductor memories are classified according to the relative speed of reading and writing. At the two extremes are the read-only memory (ROM) and the read-write memory (RWM). In between is the range in which much development work is at present concentrated, the read-mostly-memories (RMM) otherwise known as electrically-alterable-read-only memories (EAROM). The names have evolved as technology has developed, accounting for the quaintness of some of them.

A fundamental limitation of semiconductor memories is that the data are volatile, that is, can be lost if the power supply is interrupted.

It is a fact of life that a memory that has the fastest write time, and so is suitable for use in high-speed systems, loses its data in the shortest time when the power supply is removed. This means that if a fast semiconductor memory were to be used as a replacement for core in the main store of a high-speed computer, it would be essential to protect it against power supply interruption, either unintentional or at shut-down of the machine, by using a battery stand-by supply for example.

It is worth remenbering that problems occur with core memory also if the power supply is lost unexpectedly since the read-write cycle essential to restore information read from core may be interrupted. Normal shut

down of the machine involves a specified sequence of operations that leave the data intact in core memory, otherwise extensive re-programming may be necessary.

Consideration of these points has lead some computer manufacturers to accept electronic memory as a satisfactory alternative to core, and the main criteria currently arethose of cost and speed.

For some slower systems, such as in industrial control, the slower types of unipolar memory (with longer storage time in the absence of power supply) have been used as main memory.

Table II compares various types of unipolar (field effect) memory devices.

TABLE II

Write Time per word	*Memory type*	*Erasure Method*	*Storage Time without power supply*	*Classification*
–	ROM (Mask Programmed)	–	Infinite	ROM
1 sec	PROM (Fusable link etc.)	–	Infinite	PROM
6 secs	F.A. MOS	U.V. irradiation	1000 years	ROM/RMM
10 m sec	MNOS (grade A)	electrical	100 years	RMM
100 μ sec	MNOS (grade C)	electrical	1 year	RMM
1 μ sec	MNOS (grade E)	electrical	1 day	RMM/RAM
350 nS	MOS (RAM)	electrical	2 m sec	RAM

Bipolar memories are available also, and are classified in a similar manner. Bipolar memories have faster access times and use ECL (less than 25 nS) or Schottky TTL (less than 70 nS) technology. The disadvantages of bipolar memories are power dissipation, size, and cost, as indicated previously.

3 OPERATION OF MOS SEMICONDUCTOR MEMORIES

The following descriptions refer to p-channel types. Each type has a high resistivity, (e.g. 10 Ω cm) n-type substrate with heavily doped p+ type source and drain diffusion.

MOS

The transistor has a metal-gate electrode, either aluminium or polycrystalline silicon, that overlaps the source and drain diffusions, and which is separated from the semiconductor by a silicon-dioxide layer, about 100 nm thick. A negative charge on the gate attracts holes to the oxide-semiconductor interface, and these provide a conducting channel between source and drain. Either this gate capacitance or the diffusion capacitance can be used as the storage capacitor in an MOS memory.

MNOS

The transistor has the metal-oxide-semiconductor sandwich replaced by a metal-silicon-nitride-silicon-oxide-semiconductor sandwich. The oxide is much thinner than in the MOS transistor, usually only 1–2 nm, i.e., three to six atoms wide. At this thickness, charge can tunnel through the oxide. The presence or absence of a charge at the oxide-nitride interface defines the two states of the device. As in the MOS transistor, the charge is detected by its effect on the source-drain resistance. Charge is stored or removed by biasing the gate electrode positively with respect to the substrate.

FAMOS

The transistor has a metal electrode similar to the gate in the MOS type but it is completely surrounded by oxide and is electrically isolated. The presence or absence of charge on this electrode is detected in the same way as in the MOS and MNOS devices. The charge is stored by putting the drain electrode into avalanche breakdown, and it is erased by exposure of the device to ultraviolet radiation, which gives the charge sufficient energy to escape from the electrode.

4 READ-ONLY-MEMORIES

Typical applications are code conversion, look-up tables (e.g., sine tables for electronic calculators), fixed control logic sequences, microprogramming stores, etc. ROMs are used where it is required to provide a

pre-determined bit pattern of logic 1's and O's at the data output lines corresponding to a particular address at the input. They can be considered as combinational logic circuits.

There are two main types of read-only-memories:

4 (i) Mask Programmable

Here a fixed program is inserted into the store during manufacture by making connections between appropriate lines of the matrix.

The basic requirements for a ROM are essentially the same as those for RAMs:

a. Readout speed should be high.
b. Power dissipation should be low.
c. The basic cell size should be small.
d. The number of input and output lines should be as few as possible.

Some of these requirements lead to conflicting design criteria, i.e., high switching speeds lead to higher device dissipation. Fewer input/ output lines require signal decoding to be performed within the chip, which results in increased design complexity, chip area, and device dissipation.

Operation of a ROM

Fig. 1. shows part of a typical ROM, a 2560 bit dot pattern matrix character generator. For the description which follows only column 1 output for rows 1 to 8 will be examined for the 64th character in store. This 64th character relates to the ASCII code 111111, which is normally the symbol "?".

After insertion of the required character code, the last driver MOST in the 64 bit array for each of the eight row circuits and each of the five column units, are energized making a total of 40. The required row is then selected by addressing the three input row address lines, giving a total number of 8 rows. If now the 64th character is selected, and the condition for row 1 and column 1 is required, the input signals to the ROM would be:

Character Select	111111
Row Select	000

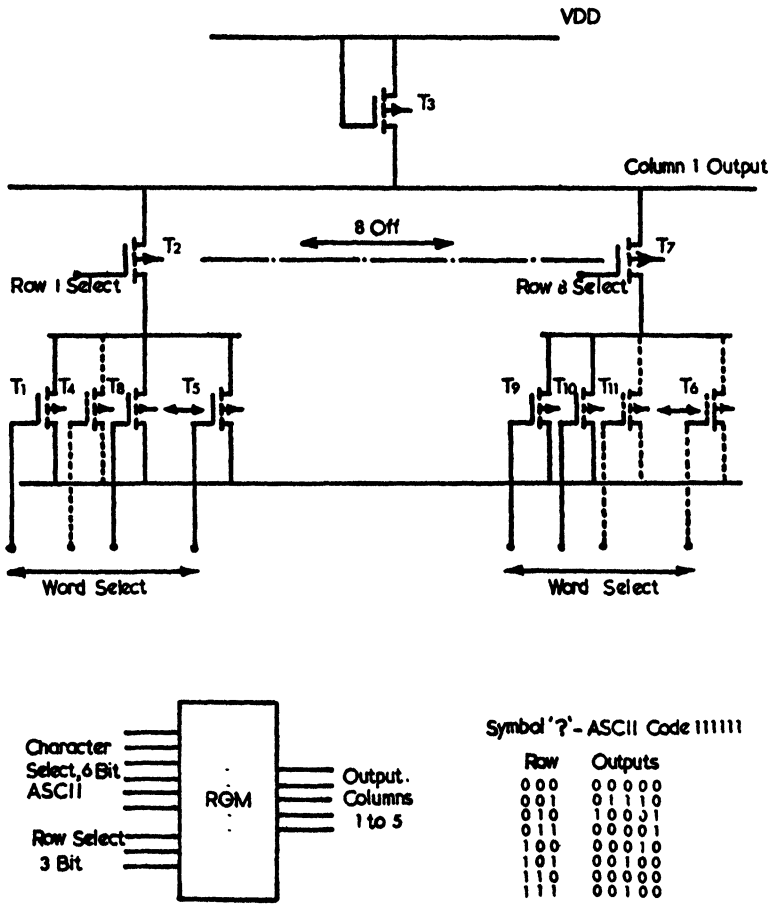

FIGURE 1

Transistors 5 and 2 would then be turned ON, thus giving a LOW or Logic 0 output for column 1. The row select inputs would then be clocked through a normal binary sequence until the input signals would be:

Character Select 111111

Row Select 111

This time, transistor 7 will be energised, but transistor 6 has been omitted from the array. Conduction cannot therefore take place through the load MOST, transistor 3, and therefore the output for column 1, row 8 would be HIGH, Logic 1.

Storage in a mask programmable read only memory is therefore accomplished by the presence or absence of the driver MOST at a particular address location. For the design suggested here, a driver MOST being present represents a logic 0, and the absence of the driver MOST represents a logic 1.

The operation of this circuit has assumed some form of address decoding network to perform the following:-

a. accept the 6-bit ASCII code and to convert it into 64 discrete signals;
b. accept the 3-bit row select binary code and produce the eight separate row select signals.

Decoding trees consisting of series/parallel connected MOS transistors are often used in MOS memory arrays for this purpose.

4. (ii) Electrically Programmable (PROM)

These were originally developed to avoid the high cost of developing the masks for small quantity production (e.g., memory of 256 $\times$ 10 bit words costs about \$300 for masking followed by a cost of about \$10 per unit). In small quantities, PROM work out cheaper than mask programmable types but due to the large, cell size required the packing density is reduced so that they are economic only for prototype work and small production runs. The PROM is originally manufactured with logic 0 at every address, and the customer programs it by feeding a current pulse into the output lead of the device with the location addressed. The driver transistor at the selected address is destroyed so that a logic 1 now remains. Once programmed in this way the information stored in a PROM cannot be changed.

5 READ-MOSTLY-MEMORIES

These devices, which are currently under development, have a basic cell design similar to the mask programmable devices. The gate insulating layer of the driver MOST in this case is made from silicon nitride. This material, when inserted between the gate and substrate, has the ability to store charge by becoming electrically polarized, hence shifting the threshold voltage of the driver MOST by up to 10 Volts. Readout is accomplished in the normal way, so long as the input voltage is less than the

minimum value required to polarize the gate insulation layer. The driver MOST will then only conduct if the shift in the threshold produced by the stored charge is such that Vin > V_{T° In general, these devices are slow to re-program and are therefore used in applications where data is read out of the system more often than it is changed, hence the name, read mostly memories.

6 RANDOM ACCESS MEMORIES (OR READ/WRITE MEMORIES)

These consist of a matrix of memory cells, bistable elements arranged in x rows and y columns. Each cell may be separately addressed and its contents read out under the control of a "read" line, or over-written under the control of a "write" line. Two types of bistable element are used:

i. Static, in which the element consists of two cross-coupled inverter circuits. This type is based on the well-proven bipolar design for a bistable circuit or flip-flop, but is modified to use unipolar devices.
   This type of circuit relies on d.c. input to establish d.c. output levels, so that the cell absorbs current from the power supply when in the conducting (ON) state, thus resulting in power dissipation. Power dissipation can be reduced by increasing the effective resistance of the load MOST, but at the expense of reduced operating speed and increased chip area to accommodate the larger transistor.
ii. Dynamic, in which the element relies on the capacitance of the gate electrode storing charge. Since the leakage resistance at the gate electrode is very high, the charge will remain for some time and can be used as memory. It is however, necessary to "refresh" the stored data at regular intervals to prevent the charge leaking away and the data being lost.

Advantages of dynamic over static cells are:

a. ratioless design (amplification factors of load and driver MOSTs are equal) resulting in greater packing density;
b. due to ratioless design the turn-on and turn-off switching times are similar resulting in a much faster circuit;
c. current flows during the switching period only, resulting in a great reduction of power dissipation.

Typical circuit operation

6. (i) Static RAM

Consider Fig. 2. The bistable action will be performed so long as the output voltage levels of the inverters, when in the low state, are less than V_T.

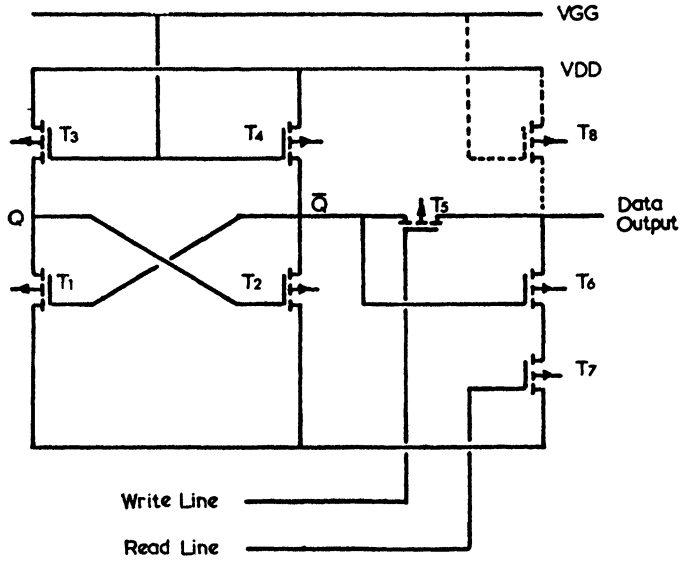

FIGURE 2

In order to write data into a unit such as this, the data output line, under the control of the external logic circuitry, is either taken high, logic 1, or low, logic 0, depending upon the state of the data that is to be stored. The stored data may then be retrieved by simply sensing the state of the data output line.

To be able to operate successfully in this mode, each bistable element must have associated with it the necessary read/write gating and an output buffer stage to prevent the loss of stored data during a read cycle. Consider in more detail the circuit shown. Transistors 1 and 2 are the drivers associated with the two inverters, and transistors 3 and 4 are the two load devices. Transistor 5 is the device required for writing data into the cell, and provides direct access to the cell from the data output terminal. Transistor 6 acts as a buffer to prevent the read device, transistor 7, from overwriting any stored data.

During a read cycle, transistor 7 is turned on, transistor 6 is therefore in the conducting or non-conducting state depending upon the state of the

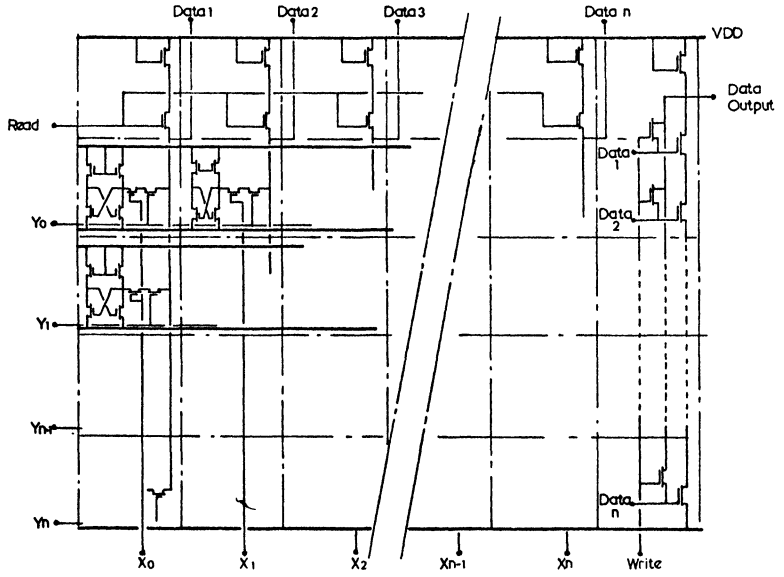

FIGURE 3

stored data at the $\overline{Q}$ terminal. A common load device, transistor 8, is used to define the logic levels on a common data line. If $\overline{Q}$ is in the low state, i.e., transistor 2 is conducting, then during a read cycle, transistor 6 will be off, and hence the data output will be high. On the other hand, if $\overline{Q}$ had been high, transistor 6 would have been turned on, producing a low output state.

To read information into the cell, transistor 7 is turned off and transistor 5 turned on. The data line is then held at either a high or low state depending upon the information that is to be inserted.

This basic cell may now be incorporated into a large array suitable for the storage of a large quantity of information. There are many configurations available, but the example illustrated in figure 3 utilizes X-Y select in the form of series switching transistors with the common load system thus producing an AND operation.

Consider the operation of the circuit shown during a read cycle. With the read terminal held at logic 1, and no address inserted, all of the series transistors in the output stage will be switched on, thus producing a low, or logic state. If address Yo and Xo is now inserted, the cell (XY)o is connected to data line 1. If the information stored in this cell were logic 1, then the output from data line 1 would be high, thus keeping the data out-

put line in a low state. If the information contained in this cell had been a logic 0, however, the output from data line 1 would become low, thus turning off the series transistor in the output array, which is associated with data line 1; the output voltage will therefore rise to logic 1.

To write data into a cell, the read line is turned to logic 0 and the write line held high at logic 1. The required state may now be inserted into the required cell, via the data output line.

NOTE In this very simple design, which is just being used as an illustrative example, the information read out from a particular cell is the inverse of the data that was stored at that point.

These devices require an input decoding network similar to that used with the ROM.

6 (ii) Dynamic RAM

a. Ratio circuit: figure 4. shows a four transistor cell based on the fundamental static cell design. Data are stored in the cell in the parasitic capacity C_1 and C_2. In order to read data from the cell, the word select line is taken negative so that transistors 3 and 4 are turned ON. Current may then be sensed flowing in the two data lines, thus giving an indication of the state of the stored data.

   In order to write fresh information into the cell, the word select line is again taken negative, thus turning on transistors 3 and 4. If one of the data lines is now taken to a negative potential whilst the other is kept at zero volts, the cell is forced, by transferring charge to C_1 and C_2, to the new desired state.

   To refresh the data, which may already exist in the cell, the two data lines as well as the word select line, are taken to a negative potential. The bistable action already described will then ensure the charge associated with the capacitors C_1 and C_2 is refreshed providing a ratio of about 20:1 exists between the gain factors of T_1 and T_3 as well as T_2 and T_4, i.e., if C_1 had been charged and C_2 had not been charged, during a refresh cycle T_1 is turned ON and T_2 will remain OFF. The charge on C_1 will then be refreshed to a negative potential, whilst the charge associated with C_2 will remain at zero.

   Although this design gives some of the advantages of dynamic systems, one of its major disadvantages is that a ratio must exist between

the gain factors of the load and driver MOST, thus affecting the cell size and operating speed of the system.

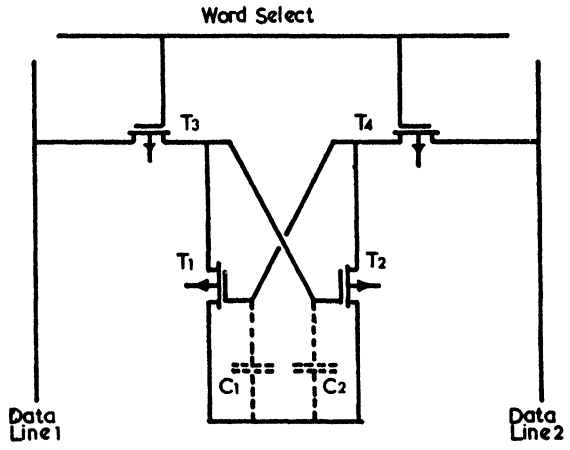

FIGURE 4

b. Ratioless circuit: figure 5. shows a three transistor dynamic cell, the operation of which is as follows. Information is stored in the capacitor C, which can be either altered or refreshed via transistor 3. In order to read the data from the cell, the read data line is first pre-charged to a negative potential. If a logic 1 is now stored in the capacitor C, the read data line will be discharged as soon as the read transistor is switched ON. If the

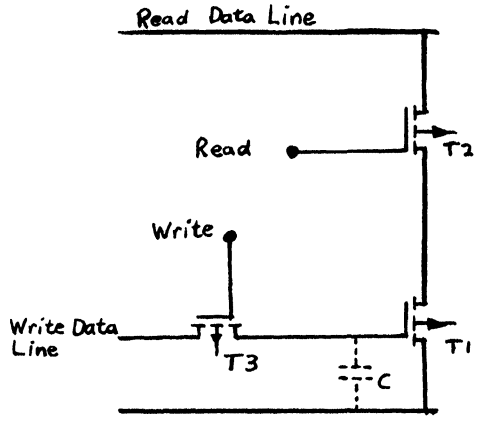

FIGURE 5

stored data had been a logic 0, the line would not have been discharged and a logic 1 would have remained on the line, therefore the output from a cell such as this is the complement of the stored data and it could be used for refresh purposes if inverted and fed back to the write data line.

The successful operation of this circuit does not rely upon a ratio of amplification factors between the transistors used, and therefore "unit cells" may be used throughout the system. This circuit, therefore, has all the advantages of a dynamic system and is the most popular circuit in use today giving a typical access time of about 150nS and needing a data refresh cycle about every 2mS.

The dynamic MOS RAM shown in figure 6. is organized on the same X-Y matrix system as the static RAM. This means that any of the memory cells within the system may be accessed by addressing their X-Y co-ordinates.

As with the other three transistor dynamic cells, a pre-charge pulse charges the parasitic capacitors associated with transistors T_6 and T_7. The X read select line is then enabled, addressing one complete row of cells, which results in the data appearing on the read line being the inverse of the data stored in that particular cell. This data is then made available at the output during a refresh cycle during which the write enable line is activated, followed by a write select pulse. In this way, the output data for a complete row of memory cells is fed back to the relevant input. Hence, by addressing one cell within an array such as this, not only is the data associated with that cell refreshed, but the data associated with all the cells in the same row is also refreshed. Therefore, the burden placed upon external circuitry to ensure an adequate refresh cycle time is reduced.

Consider the operation of this circuit in more detail. If the capacitor associated with transistor T_2 is charged to a logic 1, then during the read select cycle, the precharge associated with the read line (Transistor T_6), is discharged to earth thus holding the gate of transistor T_5 at 0 volts. When the write enable line is raised to logic 1, transistor T_5 will be off, and hence the precharge associated with T_7 will remain and will be seen as a logic 1 at the output terminal. Finally, the write select line is raised to logic 1, thus turning transistor T_1 ON and transferring the charge associated with the write data line to transistor T_2. In this way the stored data has been refreshed. This operation is seen clearly on the Timing Diagram, which illustrates the operation of the cell when storing a logic 1 and a logic 0, figure 7.

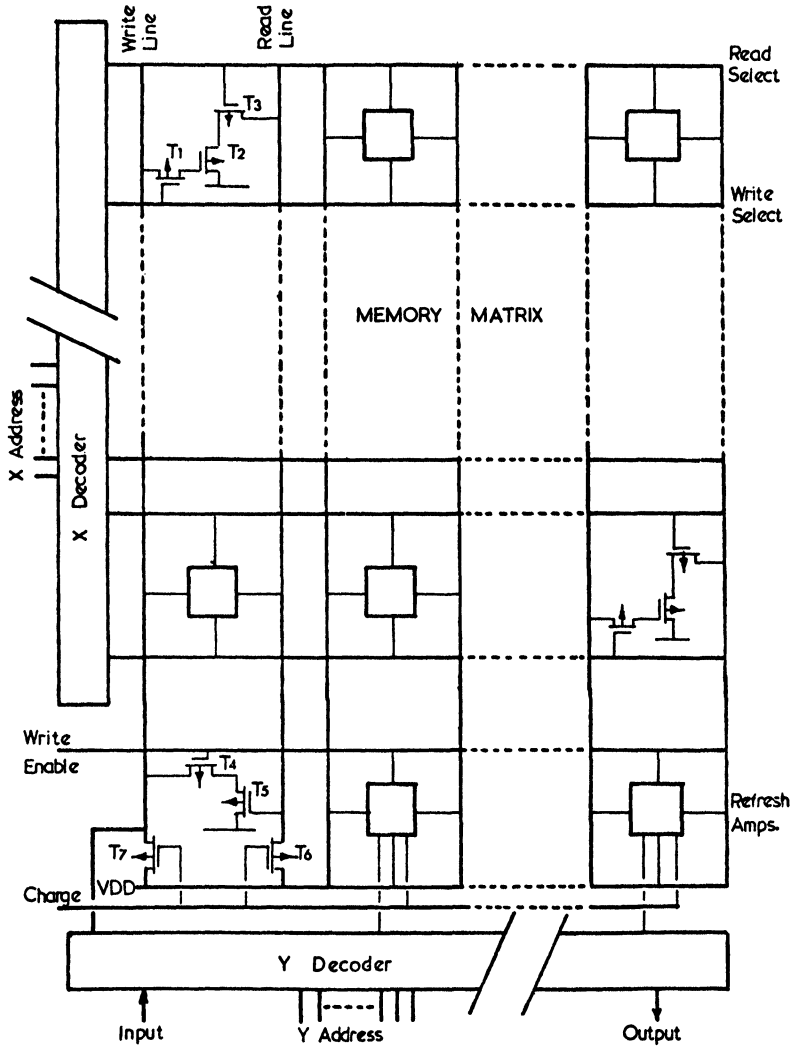

FIGURE 6

In order to write new data into the cell, transistor T_4 is held off, and the new data is inserted from the write line via transistor T_1 into the cell. The system used here as an example, requires a four phase clock system, the phase relationship of which is shown on the timing diagram of figure 7.

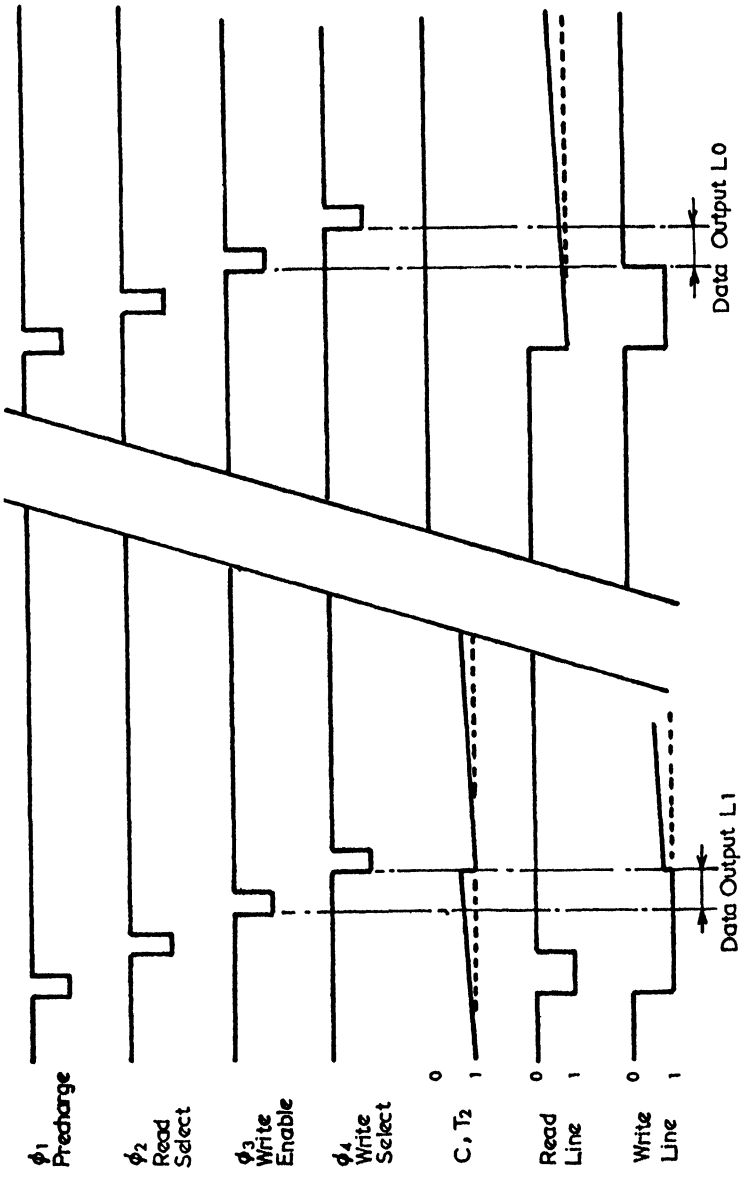
ϕ_1 Precharge
ϕ_2 Read Select
ϕ_3 Write Enable
ϕ_4 Write Select
C, T_2
0
1
Read Line
0
1
Write Line
0
1
Data Output L0
Data Output L1

FIGURE 7

7 COMPLEMENTARY MOS (CMOS) DEVICES

In MOS circuits, there is a power/speed compromise due to the resistance presented to the output circuit by the load MOST. This results in a choice having to be made between high switching speeds and high power dissipation or low power dissipation and low switching speed.

Complementary MOS devices attempt to overcome this problem by using a combination of P and N channel transistors. Figure 8 shows the arrangement of a CMOS static inverter.

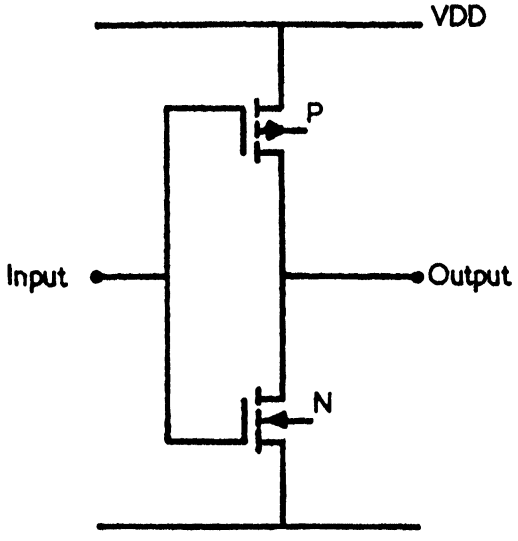

FIGURE 8

If 0 Volts represents Logic 0, and 1 Volt represents Logic 1, then with a Logic 0 applied to the input of the inverter, the N channel device is turned OFF and the P channel device ON, as a result of which the output is at Logic 1. Similarly, if a Logic 1 was applied to the input, the N channel device would be turned ON, and the P channel OFF, resulting in a Logic 0 at the output. This illustration shows that at no time does a d.c. path exist from the power supply rail to ground, and therefore in circuits such as this, the d.c. power dissipation is negligible.

Although the major advantage of these types of circuits lies in their low quiescent power dissipation, current transients do occur during a switching cycle, when both devices are on for a short period. As a result, the power

dissipation of CMOS gates increases rapidly with frequency as shown in Fig. 9 . CMOS circuits, however, are capable of much greater operating, speeds than conventional MOS devices, since the output node capacitance is always charged and discharged through an ON transistor to either the power supply rail or earth.

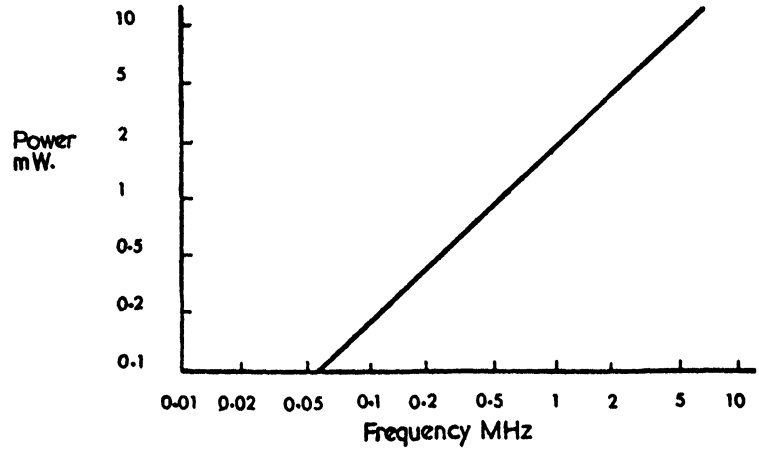

FIGURE 9

8 PROGRAMMABLE LOGIC ARRAYS

A ROM can be considered as a universal combinatorial logic gate, where the addresses are the gate inputs, and the output of the ROM is the gate output. By programming the memory, any combinatorial logic function can be generated. The ROM can therefore replace a collection of logic gates, and it has the further practical advantage that the delay is independent of the particular inputs.

In many applications, not all the possible input combinations are required, so that a smaller and cheaper ROM may be used, and the particular input combination can be programmed in the same way as the output. Such device is called a programmable logic array (P.L.A.). The outputs from a PLA may be connected back to the inputs directly, or through flip-flops and in this way complete logic sub-systems may be designed for quite different applications by choosing the programming pattern for the

P.L.A. This programming can be carried out automatically by computer, once the designer has defined the truth table or logic equations.

MOS and bipolar PLA are both now being marketed. The bipolar P.L.A. will revolutionise the design of logic systems that are made today by using T.T.L. Fig. 10 shows a graph of cost per sub-system plotted against the number of sub-systems produced per year. This graph is based on a study of a hypothetical '1000-gate' sub-system and it assumes that there are 10 gates per package in a typical T.T.L. system. Until recently, this system would have been designed with custom L.S.I. or with a carefully engineered mix of off-the-shelf T.T.L. devices, depending on the number of systems to be built. Now, however, the PLA approach provides the cheapest solution, making the engineered T.T.L. approach uneconomic. The main cost saving in using PLA is in the initial design and layout stages.

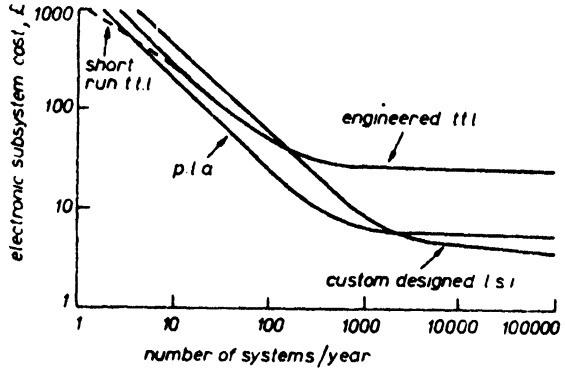

FIGURE 10

9 CHANGING TRENDS

Semiconductor memories will eventually replace magnetic cores in computer main-frame memories. The low costs and simplicity of use will also lead to the widespread use of semiconductor read-write, read mostly, and read only memories in applications far from computers.