Semiconductor data storage offers many advantages. These include:
Speed — systems have been built with cycle times below 20 ns.
Low Level Modularity — memory need no longer be centralized, but can be distributed through the system at no cost penalty.
A Broad Spectrum of Compatible Cost — performance combinations are available.
Ease of combination with logic in the same structure simplifying decoding and sensing as well as offering new functional capability such as associative structures.
A common systems assembly technology.

An overriding consideration with respect to the employment of semiconductor memories, however, relates to the cost — for memory is an exceedingly cost-sensitive application. This obsession has influenced considerably the direction in which semiconductor random access memories are being pursued. This paper will review the status of semiconductor random access memories pointing out some of the important changes that have taken place in the last year or so that make it increasingly attractive from a cost point of view, as well as for the advantages listed above.
The Impact of Price Drops

In order to show the magnitude of the changes that have taken place in this field, Figure 1 is a reproduction of a plot of the then existing situation that was used in a talk in March 1969. The minimum cost for semiconductor memories was about 40¢ per bit, making cores a cheaper alternative for any memory larger than about 1,000 bits. At that time the semiconductor industry had a considerable problem in convincing the user that the situation would change by some time in the 1970's to that depicted in Figure 2, where semiconductor costs were projected below those of core systems for memories of all sizes. Of course, this extrapolation assumed that core memory prices would not change significantly, since cores represented a relatively mature technology for this application.

What has happened is shown in Figure 3. The erosion in semiconductor prices combined with the introduction of new products that have been specifically designed for low cost random access memory applications have reduced the present cost for semiconductor storage elements to below that projected in Figure 2. The resulting effect has been far more rapid than would have occurred just through the evolutionary decreases in costs that integrated circuit users have come to expect. Core memory prices have also dropped dramatically, partially in response to the emergence of a competing technology.

The advances responsible for this dramatic cost decrease do not depend on any new device or physical phenomenon, but strictly upon applying conventional integrated circuit technology and clever circuit design to the problem at hand. Both bipolar and MOS structures have been made for this application. To date, the use of bipolar in large systems is evidently more advanced, since at least two major systems are committed to the incorporation of medium or large bipolar random access memories. The ILLIAC IV utilizes many parallel processors, each with a 256K bit, 188 nsec cycle bipolar RAM. Recently IBM has announced the 370/145 computer which will employ half million byte modules of 200 nsec bipolar RAM. Probably these were first because the performance required justified semiconductor memory even at a premium in cost. The speed ranges are beyond those obtainable with conventional core memories. It is in such high cost per bit applications that semiconductor storage is most easily competitive. It was the safest place for first commitment to semiconductor memories.

Storage elements in these memories are simple cross coupled flip-flops. Very simple structures can be employed for these bits because all of the related addressing, decoding and sensing circuitry is integrated on the same semiconductor chip. This ability to put the appropriate peripheral circuitry close to the storage cells greatly simplifies the cell structures that can be employed.

Figure 1 — Cost comparison for semiconductors and core memories of various capacities as the situation existed in early 1969.

Figure 2 — Similar cost comparison as projected in early 1969 for sometime in the 1970's. Both Figure 1 and 2 were taken from a New York IEEE Convention talk in March 1969.
In the case of the ILLIAC memory, the semiconductor chip contains 256 bits selected by means of a 3 out of 6 code. Such a chip fits in a conventional 16-pin dual in-line package. The organization of the chip is 256 words by one bit. In the system, these are combined with a decoder/driver chip that converts ordinary binary address bits to the special 3 out of 6 code and has sufficient drive that relatively large memories can be constructed with a minimum of additional drivers. For example, a memory system composed of a similar 256 bits bipolar memory device is shown in Figure 4. This 2K x 18 memory system fits on a single printed circuit board about a foot square. The center portion contains the 144 storage packages. On the bottom by the connector are the data buffers. This system operates at 120 nsec cycle time. Several such boards can be combined into a larger system with no loss in performance.

While such a memory is competitive in cost with speeds of 200 nsec or below, it is significantly more expensive than slower core systems because of the large chips required for the storage elements. To be competitive in this range, it is necessary to pack the bits more densely on the silicon wafer. This results from the basic economics of the semiconductor industry. The cost of processing a wafer through a particular process sequence is essentially independent of what is contained on the silicon wafer. Only the photo masks are changed from one structure to another.

To minimize silicon cost, it is important that a maximum number of good bits be obtained from a given wafer. This consideration when extended to the completed part is somewhat modified by packaging and testing costs, since these go up less than linearly with the number of bits in the semiconductor die. In any case, high packing density is of paramount importance. It relates directly to the number of possible bits that can be placed on a wafer. In addition, since yield relates strongly to die area, for smaller bits the percentage yield of dice with a given functional complexity is increased.
Dynamic Minicells

As a net result for memory structures, the cost per bit drops rapidly with the area per bit. Obviously, this results in a considerable incentive to achieve minimum memory cell size. Accordingly, many ways have been investigated to make smaller cells. Some of the cells in current semiconductor memories are compared in Figure 5. It is apparent from Figure 5 that the smallest cells are achieved utilizing MOS rather than bipolar memory structures. This is generally true. The functional packing density with MOS runs from 3 to 10 times that achievable with bipolar structures. It is a result both of the fact that the MOS transistor is self-isolating and that certain logic configurations are more easily achieved with them. As additional incentives toward the MOS structures, the wafer process cost is generally somewhat lower than in the bipolar case, since the processing sequence has fewer major steps; and a lower effective defect density can be achieved, resulting in higher yields for a given die size. Thus, the MOS technology offers a very important approach to achieving low cost semiconductor RAMS. As an example of a currently important application of MOS, consider the small cell in Figure 5.

Figure 6 shows a circuit diagram for the dynamic cell shown in Figure 5. This simple MOS storage cell utilizes three minimum size transistors. The storage is on the gate capacitance C of Q1. In operation, the data on the data in line is written onto C by selecting the particular cell through the write select gate, Q3. The data is read by charging the data out line then turning on Q2. If a negative charge (i.e., a "one") is stored on the gate of Q1, the data output line is discharged, while if a zero (no charge) is stored on Q1, the data out line is not discharged.

Of course, the charge on capacitance C leaks off through the reverse-biased diode associated with the drain junction on transistor Q3. Thus, it requires periodic refreshing to retain the stored information. Hence, this type of storage cell is referred to as a dynamic memory. Every few milliseconds the information must be rewritten into the cell to maintain the level of charge. While for an isolated cell this could prove to be a considerable disadvantage, when several of these are combined into the same monolithic structure with the necessary circuitry for amplifying and rewriting the output, it becomes overall a very efficient means of semiconductor storage. This cell itself can be realized with silicon gate MOS technology in an area less than 6 sq. mils, compared with about 25 sq. mils for the 256 bit bipolar.

Figure 7 shows a photomicrograph of a 1024 bit array of such cells including the necessary refresh amplifiers and decoders. This particular chip is arranged as a 32 x 32 array of storage elements. The X and Y decoders form a cross through the middle of the array. The refresh amplifiers are at the bottom. The entire semiconductor die is 113 x 139 mils or slightly more than 15 sq. mils per bit overall. 32 refresh amplifiers are required, since selecting a particular X address during any read cycle refreshes all 32 bits having that address. Thus, in order to refresh the entire array, it is necessary that each of the 32 possible X addresses be selected. The read cycle automatically writes the information into the bit location. Hence every few milliseconds it is necessary that each of the 32 address lines be read. In a system it is still possible to arrange the structure so that a total of 32 cycles is capable of refreshing the entire memory, irrespective of how large the system might be. While this commitment of a certain number of memory cycles to refresh the stored information introduces a constraint, the decrease in the cost-per-bit of the storage because of the much higher packing density that can be obtained by the dynamic cell makes it a good economic compromise.
Refresh Methods

There are several ways that the refresh can be accomplished on a system basis. In cases where synchronous operation is important or where speed of operation is not critical, it is possible to double the cycle time so that every memory cycle has a refresh period also. Alternatively, it is possible to steal a fraction of the available cycles so that every location is refreshed as necessary. For example, in a 500 nanoseconds cycle memory requiring refresh every two milliseconds, some 32 cycles out of 4,000 or something less than 1% of the available cycles must be committed to the refresh. In such asynchronous operation, on occasions an effective memory access may prove to be twice as long as the average memory access. This can be improved somewhat by keeping track of which addresses have been accessed so that only those that have not been addressed in a given two millisecond period are refreshed. However, this additional bookkeeping is only appropriate in quite large systems, since it results in a significant increase in the overhead for a relatively small improvement in the average access time. It is also possible to use such procedures as to shut the memory down every two milliseconds and accomplish an entire refresh in a 16 microsecond period. For real time applications it is often not necessary to keep the data longer than the refresh time of the memory, so refreshing is completely unnecessary.

A memory system built from such 1024 bit memory elements is shown in Figure 8. This system is complete except for power supplies. It is organized as a 4096 words by 16 bits. For a system of this size, the overhead cost can readily be made less than 0.2¢ per bit — much less than for a comparable sized core memory.

This particular system operates comfortably at a 650 nanoseconds cycle and 400 nanoseconds access when made with memory components that are specified as 600 ns cycle. It can be expanded to very large systems with no further degradation in the performance.

These two examples represent the first generation of bipolar and MOS RAMS designed for computer main frame memories. They have already demonstrated many of the advantages claimed for semiconductor RAM and are cost competitive with alternative techniques. Considerable evolutionary progress can be expected as new generations incorporate improved devices, circuits and system design. The flexibility of semiconductor integrated circuit technology offers very significant room for future progress.