Memory on a chip: a step toward large-scale integration

Already in production, a read-only memory made of MOS transistors on a single chip of silicon is fast, inexpensive and easily stores any combination of 256 bits in less than 1/200th square inch

By Lee Boysel

Fairchild Semiconductor Division, Fairchild Camera & Instrument Corp., Mountain View, Calif.

Read-only memories, increasingly used to control instruction execution in general-purpose computers, have many potential applications in displays, controls and telemetry systems. However, the lack of suitable batch-fabrication methods has limited the size of high-speed, all-semiconductor types such as diode arrays because their cost per bit rises sharply with size.

Now, a read-only memory on a monolithic chip measuring 60 by 80 mils is in production at the Fairchild Camera & Instrument Corp.'s Semiconductor division. A density of 256 bits on this chip has been achieved through the use of metal oxide semiconductor technology. Active MOS devices serve as both the memory elements and as logic elements for address decoding. The memory's access time of 1 microsecond is longer than that of some other forms of memory, but is more than adequate for most applications.

Previous attempts to fabricate monolithic memory arrays from bipolar devices have met with only limited success because bipolar memory cells exhibit parasitic capacitance and must therefore be isolated from one another. To provide this isolation, together with word-decoding logic and output buffers, requires a complex succession of processing steps that appear to restrict the potential for reducing memory size and cost.

Magnetic memory systems, whether they be electromechanical drums and disks or all-magnetic thin films, wires or ferrite cores, cannot be batch-fabricated. Decoders and buffers require technologies different from that of the storage element. And the electromechanical forms are very slow.

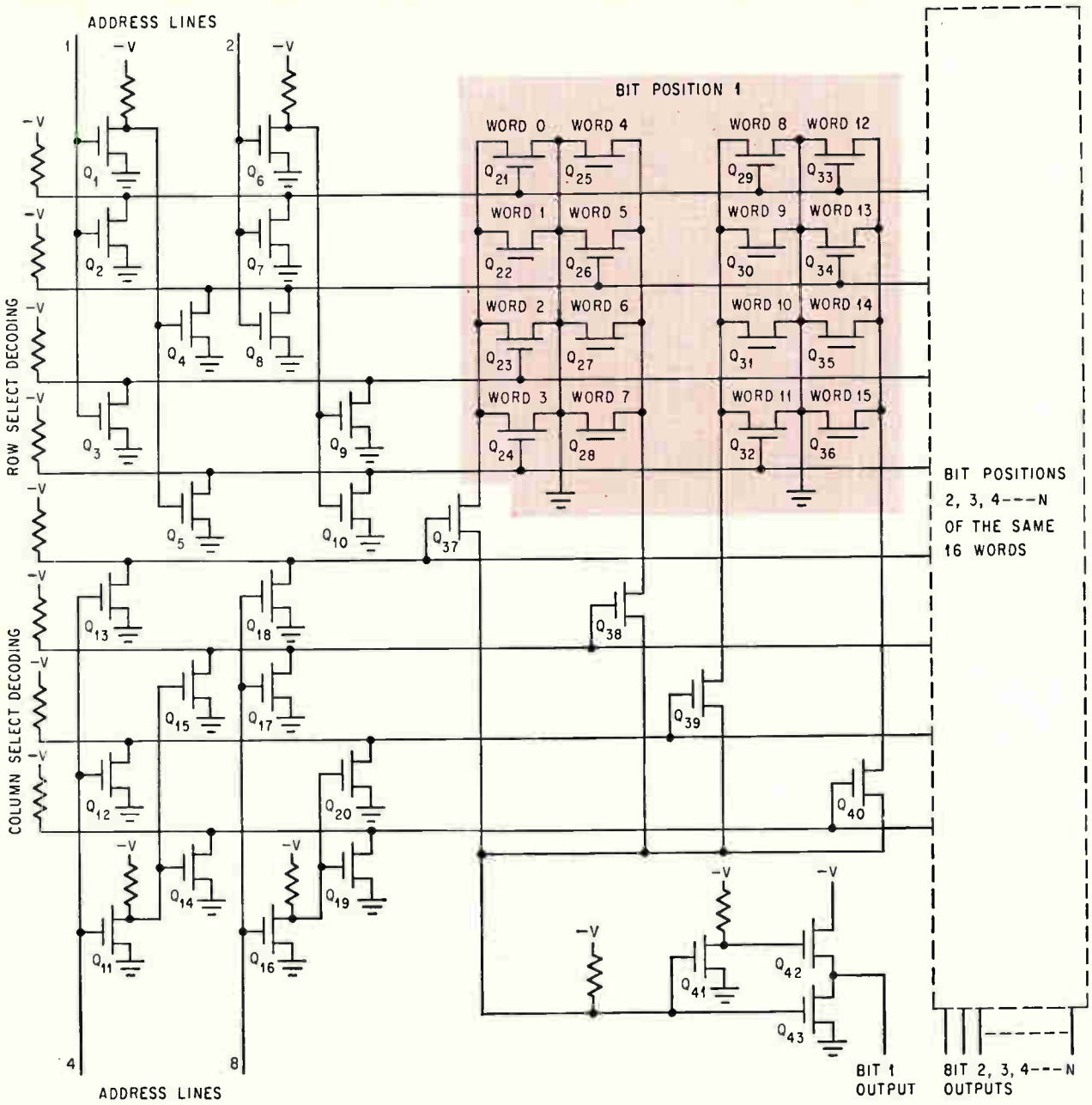
A memory built exclusively with active MOS devices has none of these disadvantages. The entire storage matrix, with address decoders and output buffers, can be formed simultaneously on a monolithic substrate using standard MOS techniques. Isolation isn't necessary, and component density can therefore be greatly increased. This, in turn, yields improved reliability and economy by permitting the packaging of complete functional entities together. External connections for internal purposes are not necessary. For instance, a 1,024-bit memory containing 256 four-bit words requires only 14 leads: eight binary input lines for the address, four output lines for the data and one lead each for supply voltage and ground reference.

Switching with MOS

The basic circuit depends on the application of a MOS transistor as a switch. The memory matrix stores a binary 1 wherever a channel can be established between a source and a drain; the oxide layer is made thin enough at these points to open the channel with an electrical signal. Where a binary 0 is to be stored, the oxide insulating meterial is left sufficiently thick so that the channel cannot be opened. The address decoder is made of MOS NOR gates that select only one word and gate all bits of that word to the output buffer. Load resistors are made of MOS transistors having their sources short-circuited to their gates. [See "Logic functions in MOS," p. 96.]

In the most straightforward design, a one-dimensional decoder selects a single word line. This line gates as many MOS transistors as there are bits in

One chip, greatly enlarged in this photo, stores 64 four-bit words with 1-microsecond access time.



A bit position of a 16-word memory, with the output buffer for that position and the address decoder for all 16 words. Storage elements of the memory array are tinted.

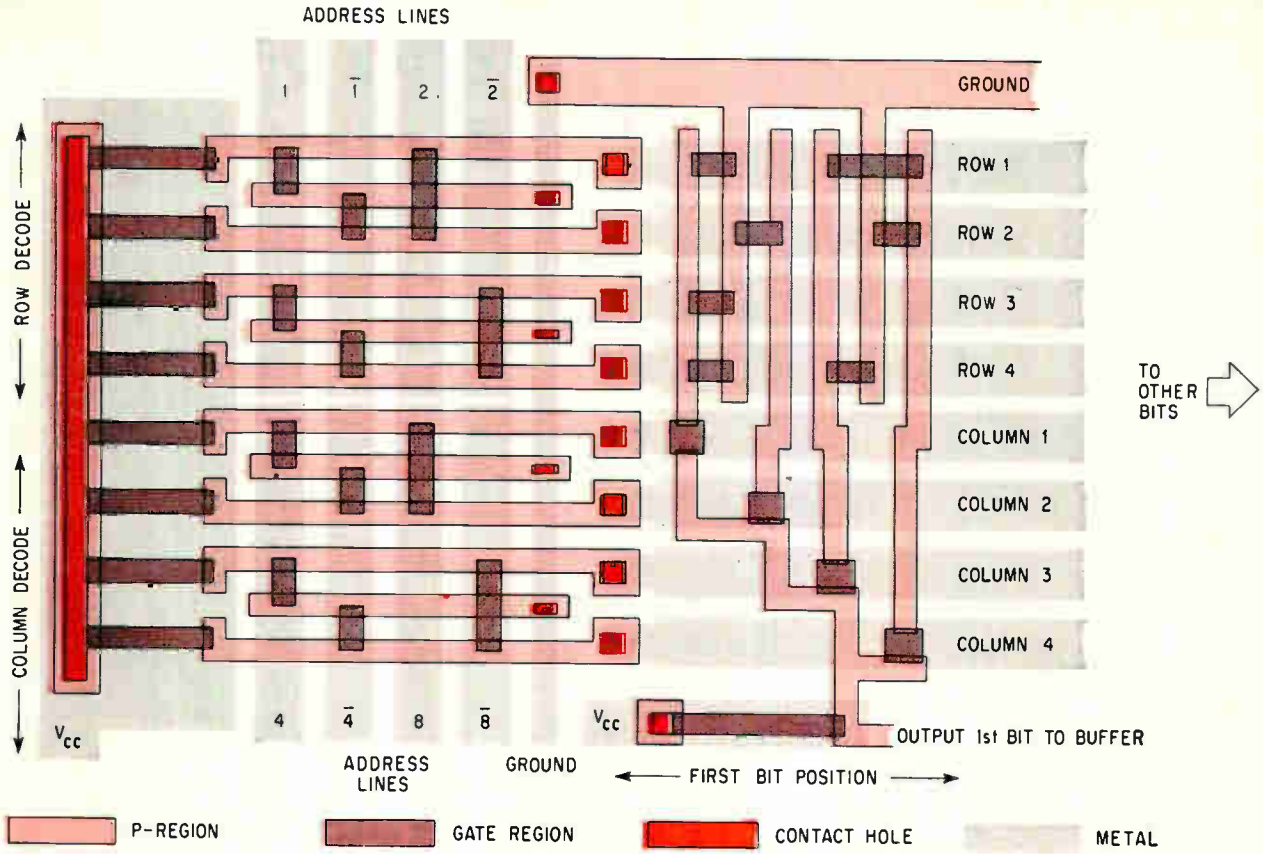
a single word in the memory. The opened gates pass current from the source to the drain, and the voltage drop across the load resistor produces a pulse through the buffer for each bit.

Each bit line in this straightforward design is the drain connection for many MOS transistors, no more than one of which can be on at any one time. The large number of inactive connections represents a substantial capacitive load on the bit line, slowing its operation.

A better way

The Fairchild read-only memory uses a two-dimensional decoding scheme that reduces the number of decoding gates and therefore the amount of parasitic capacitance. The schematic shown above depicts part of a 16-word memory with an indeterminate number of bits in each word. A four-bit address can select any of the 16 words. A single bit position common to all 16 words is organized in a 4-by-4 array; two address bits select a row and the other two select a column.

Suppose word 5 is to be read out of the memory. The address 5 in binary form is 0101. The address lines labeled 8 and 2 in the schematic are therefore at a positive potential, and address lines 4 and 1 are negative. The three gates controlled by each of the two positive lines are closed, and the gates with negative levels are open. Because Q_6 and Q_{16} are closed, no current passes through their load resistors and the potential holds gates Q_9, Q_{10}, Q_{19} and Q_{20} open. And because Q_1 and Q_{11} are open,



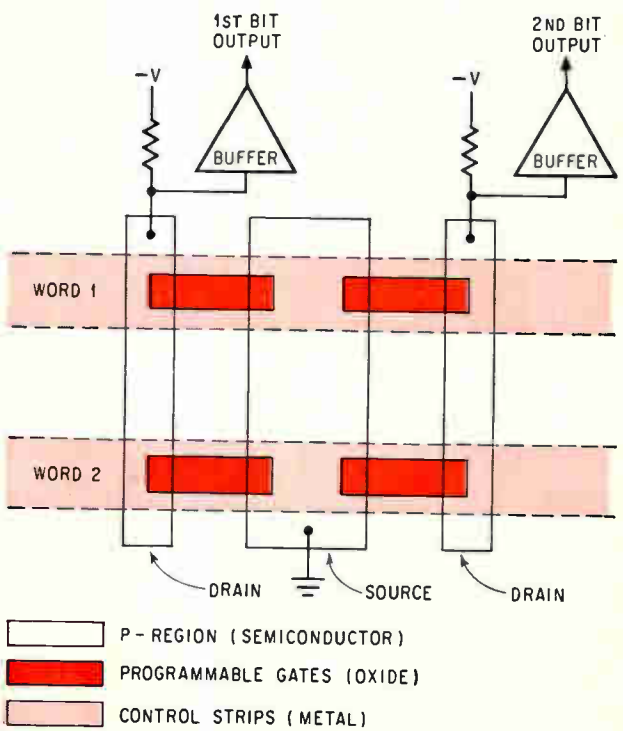
Layout of the 16-word memory in the schematic diagram on the opposite page. The actual 256-word memory being produced is laid out similarly. The output buffer is not shown in this layout.

their load resistors produce a voltage drop that holds gates Q_4, Q_5, Q_{11} and Q_{15} closed.

Because gates Q_4 and Q_8 are both closed, the row line that is their source is negative; transistors Q_{22}, Q_{26}, Q_{30} and Q_{34} are gated on, together with transistors in other bit positions of words 1, 5, 9 and 13 in the memory. Each of the other three row lines is connected to at least one open gate; the load resistor therefore has a voltage drop, the row lines are positive and all transistors controlled by these lines are off.

Similarly only the column line that is the source for transistors Q_{15} and Q_{17} is negative, and Q_{38} is gated on. The other three column lines are positive and Q_{37}, Q_{39} and Q_{40} are off. Gates Q_{37} through Q_{40} have a common load resistor (at the bottom of the schematic), but since only Q_{38} is open, current can be supplied only to transistors Q_{25}, Q_{26}, Q_{27} and Q_{28}. Of these four, Q_{26} is on, corresponding to one bit of word 5 in the memory. Transistors Q_{41}, Q_{42} and Q_{43} buffer the voltage drop across the load resistor. Current passes through Q_{42} from the external circuit—presumably some functional part of a computer or other digital assembly. Current coming from outside represents the reading out of a binary 1 from the memory.

In the same way, suppose word 9 is to be read out. The address is 1001; the row and column decoders work as described previously so that gates Q_1, Q_8, Q_{12} and Q_{20} are closed and the correspond-



Relationship of various layers in the MOS read-only memory. The p-region is deposited first, then the gates and finally the metal control lines. The read-only memory's organization is more sophisticated than this simplified diagram would indicate, but the layers are deposited in the same order.

Logic functions in MOS

Metal oxide semiconductor transistors can be combined into NOR gates from which any logic function can be constructed. A basic two-input NOR gate is at the right; gates with three or more inputs can be put together similarly. The two MOS transistors in the diagram share a common load resistor made from a third MOS transistor with its source connected to the gate.

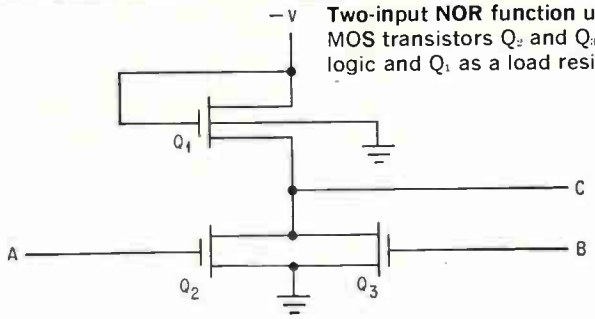
The design of a logic function using MOS transistors is simply topological control of the transistors' transconductance. The drain current is proportional to the transconductance, which in turn is proportional to the device geometry. When a MOS transistor is conducting, its transconductance, together with that of its load resistor, make a voltage divider that determines the output voltage level. When the transistor turns off, the output voltage becomes very nearly the same as the supply voltage.

Simple NOR logic functions can be implemented in MOS devices by connecting the transistors directly, as at the right. The function illustrated is $\overline{AB + C} = D$, read NOT [(A AND B) OR C] equals D. The OR part of the function is represented by the two parallel paths; the AND part by the two transistors in series. The NOT stems from the inverting effect of the transistors. Both parallel paths must have the same total resistance to maintain the proper transconductance ratio between the logic transistors and the load transistor. And this is where a penalty is levied against logic designers using this method.

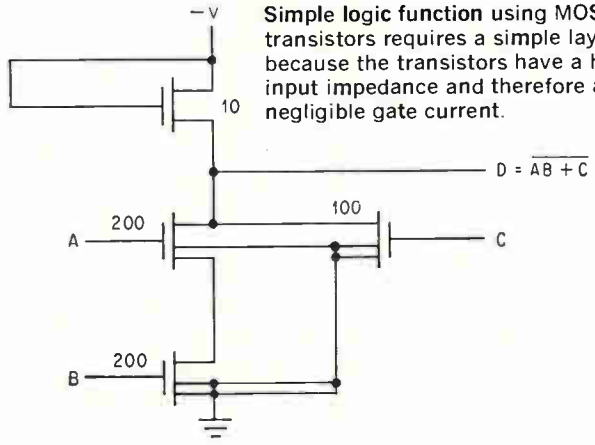
The resistance of each of the series transistors must be half that of the parallel transistor, so that their total resistance is the same. If their resistance is half, then their transconductance is double; therefore the area of the mask is doubled, as in the mask layout at the right. This presents a size penalty and shows that logic functions should be designed with as many parallel paths as possible and with a minimum of series paths to keep the total area of the mask to a minimum. This approach corresponds to using many OR's and very few AND's.

In the read-only memory, a single decoding transistor must be capable of pulling a coordinate line sufficiently positive to cut off the appropriate transistors in the memory array or in the column gates. But since more than one transistor may be turned on, the coordinate lines may have any of several (in the 16-bit schematic, either of two) positive levels capable of ungating a transistor.

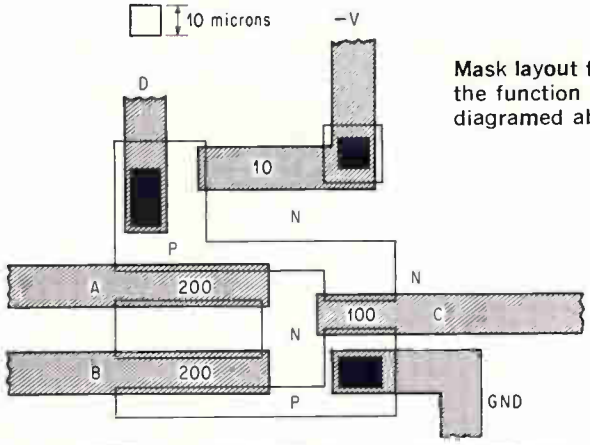
—W.B.R.



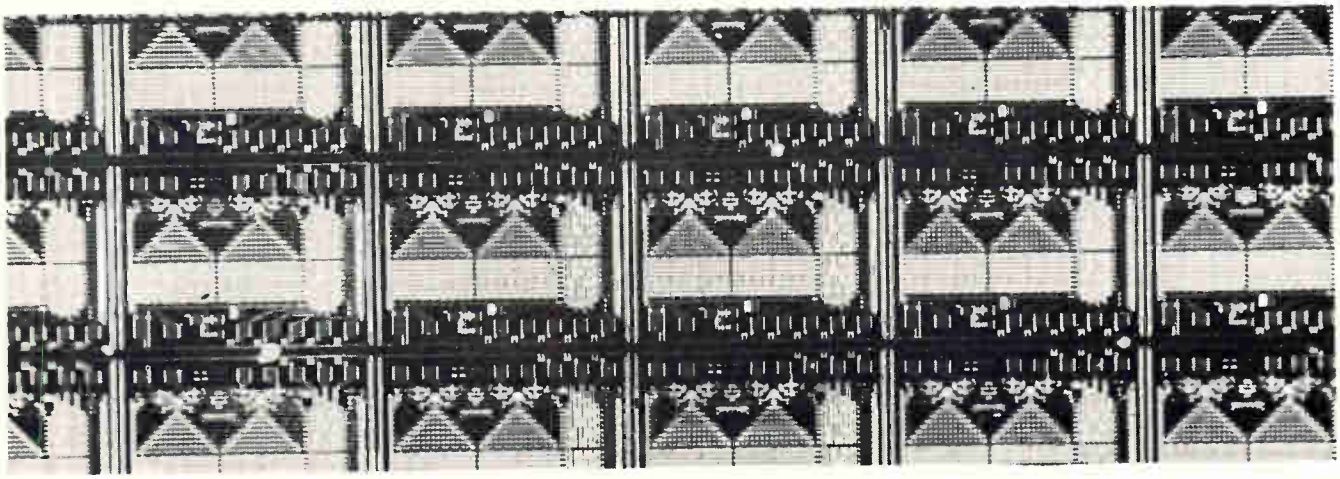
Two-input NOR function uses MOS transistors Q_2 and Q_3 for logic and Q_1 as a load resistor.



Simple logic function using MOS transistors requires a simple layout because the transistors have a high input impedance and therefore a negligible gate current.



Mask layout for the function diagramed above.



Not houses on a hillside, but part of a wafer containing row upon row of memory chips.

ing row and column lines are negative. All other row and column lines are positive. Through Q_{39}, current is supplied to Q_{29}, Q_{30}, Q_{31}, and Q_{32}. Of these, only Q_{30} can accept the current, but it has a thick oxide layer and is not turned on even though its gate is negative. There is no voltage drop across the load resistor. Current passes through Q_{43} to the external circuit, representing the readout of a binary 0 from the memory.

Twice or thrice the root

The two-dimensional decoder approach adds no logic stage delays but results in a much lower parasitic capacitance than the straightforward one-dimensional decoder. Each bit line is the source connection for four transistors at most in the memory array plus four transistors in the column decoding network (Q_{37} through Q_{40} in the schematic), for a total of eight—a 50% reduction from the straightforward design.

Larger arrays would show an even greater contrast between the two approaches. The number of connections to each bit line in the two-dimensional design is twice the square root of the memory size, or 92% less than the number of connections in the one-dimensional design for a 1,024-bit memory.

Three-dimensional stacking, using a decoding network divided into rows, columns and a third set of lines, can reduce the number of connections and decoding gates and the amount of parasitic capacitance by an even larger amount—three times the cube root of the memory size. In large memories, this produces a very great improvement in the speed-power product, a figure of merit for memories. Such stacks are easily implemented in IC's; the arrays don't actually occupy three dimensions in space, but rather involve a third set of decoders that establish a selective ground connection instead of the common ground bus shown in the schematic.

Depositing and etching

In fabricating the read-only memory, long source and drain diffusion strips of p-type material are deposited on the n-type substrate. A thick layer of oxide insulating material is laid on top of the p-type strips. Wherever a gate is to be established, a hole is etched through the oxide clear down to the p-layer and a thin layer of the same oxide is laid in the bottom of the hole, a process called "thinning the oxide." In general, the gates are narrow rectangles whose long dimension is at right angles to the long dimension of the strips of p-material.

The final step is the deposition of metal gate control strips on top of the oxide. However, at certain points these strips must make contact with the p-region as well as with the gates, so before the metal is deposited, more holes are etched through the oxide down to the p-region.

This arrangement of perpendicular strips permits high density of memory cells and retains a simple processing procedure. The actual cell density reaches a million bits per square inch, even when tolerances are kept relatively loose for ease of manufacturing. No current design techniques for devices requiring isolation can approach this extraordinary density.

Simple redesigning

Since each memory requires a new configuration, depending on the data it stores, each memory requires a complete new design. The redesign is reduced to a minimum with a master pattern for the oxide-etching mask. This pattern, if unmodified, would produce a mask that would insert a binary 1 in every bit position in the memory. To insert binary 0's, a small piece of layout tape is placed over the corresponding openings in the pattern. The mask made from the modified pattern is then used with the other masks in a standard process for manufacturing MOS devices.

The read-only memories are easily tested. A copy or test reference of the memory is built of bipolar integrated circuits and a matrix of discrete diodes. Every word in the memory being tested is addressed, and its output compared with that of the corresponding word in the test reference. The comparison is made at high speed and a complete 64-word memory can be tested in a few milliseconds. A batch of identical memories is tested on the wafer before the wafer is diced. The test reference can be modified for different memories by adding or deleting discrete diodes.

The standard read-only memory now being produced contains 64 words of four bits each. It has p-channel devices in which the electric field created by the gate converts the n-type substrate to p-type. The memory's layout follows that shown for the 16-word example in the diagram on page 95 [a photograph of the complete monolithic memory is on p. 92].

The Fairchild memory has an access time of about a microsecond. Much shorter access times would be attainable in a similar memory built with complementary devices—that is, with p-channel transistors in the memory array and n-channel transistors in the decoder, or vice versa.

New concepts and new approaches to logical organization are needed if the full potential of large-scale integration in MOS technology is to be realized. Circuit design and logic design are no longer important; the function labels in a block diagram have taken their place.

The author

Lee Boysel is in charge of all standard MOS product design at Fairchild Semiconductor. Previously, he designed large MOS arrays for data acquisition at the IBM Corp.