

A 2-ns Cycle, 3.8-ns Access 512-kb CMOS ECL SRAM with a Fully Pipelined Architecture

Terry I. Chappell, Barbara A. Chappell, *Member, IEEE*, Stanley E. Schuster, *Fellow, IEEE*, James W. Allan, Stephen P. Klepner, Rajiv V. Joshi, *Member, IEEE*, and Robert L. Franch

Abstract—This paper describes a 512K CMOS SRAM with ECL interfaces which has a 2-ns cycle time and a 3.8-ns access time, both of which are valid for random READ/WRITE operations. The 2-ns cycle time is achieved without degrading access time or operating margins by using a fully pipelined architecture incorporating self-resetting circuit blocks. The CMOS process features a 0.8- μm average feature size, self-aligned TiSi_2 , triple-level metal, and a 0.5- μm L_{eff} . Details of the pipelined architecture are described along with several examples of the self-resetting circuit blocks with emphasis on features key to high-speed operation, fast cycle time, and robust design.

I. INTRODUCTION

IMPROVEMENTS in architecture, circuits, and process technology have lowered the access and cycle times of high-capacity (> 64 kb) SRAM's to 5 ns in recent years [1]–[3]. This paper describes the first sub-5-ns high-capacity SRAM, a 512-kb CMOS ECL SRAM with 2-ns cycle time and 3.8-ns access time for fully random READ/WRITE operations [4]. The 2-ns cycle time was obtained through more aggressive use of pipelining than previously reported [2], [5], [6]; the pipelining is transparent to the user. This fast cycle time permits a very high memory bandwidth; with a $\times 16$ organization, the SRAM achieves a 1-gigabyte/s data rate, the largest reported for a high-capacity SRAM. The chip has ECL interfaces; the ECL-to-CMOS receivers have < 100 -mV worst-case sensitivity [7], while the CMOS-to-ECL off-chip drivers achieve ± 90 -mV level control.

Fundamentally, the minimum cycle time of a pipelined SRAM is the subarray cycle time—the time needed for a READ/WRITE word-line pulse and a precharge pulse sufficient to restore the bit lines after a WRITE. This SRAM achieves this minimum cycle time without sacrificing operating margins or access time. Fast subarray cycle time is facilitated by the use of a full-CMOS six-transistor memory cell that can be written and cycled quickly without sacrificing soft error rate, stability, or standby power. All

other circuit blocks provide cycle times less than or equal to that of the subarray.

A key benefit of a full-CMOS six-transistor memory cell in a SRAM with a fast cycle time is the fast WRITE time it provides. From the time the word line rises, the six-transistor cell used in this SRAM requires only 1 ns to switch and to have the high side of the cell recover to the full power supply voltage. In contrast, the full recovery of the high side of a four-transistor cell (with either resistor loads or thin-film p-channel transistor loads) requires many nanoseconds. Thus, operating a four-transistor cell at a 2-ns cycle time would result in a severe degradation of the high side of cells, which are written in one cycle and then read in the next cycle. Insuring the READ stability of these cells would be difficult without either increasing the ratio of their latch device size to access device size, at the cost of area, or increasing their pull-up current, at the cost of standby power for true resistor loads or at the cost of area for thin-film p-channel transistor loads.

The characteristics of the SRAM are summarized in Table I. The chip features a fast 2-ns cycle time, which is valid for any combination of READ/WRITE operations. As will be shown, it was found that the fast cycle time also facilitated a 3.8-ns chip-select access time including word-line redundancy. The chip was designed for synchronous operation without the need for address-setup time. The high-performance CMOS interfaces provide ECL-level I/O signals using differential-amplifier receivers previously reported [7] and a new off-chip driver circuit that will be described. The logical organization of the chip places 32 kb behind each of 16 data pins. At 200 MHz, the active power of the chip is 3.5 W. The power supplies given in Table I were used because they are IBM mainframe compatible. The same performance can be obtained with a single 3.6-V power supply. To facilitate testing at the minimum cycle time, the chip includes test circuitry which can provide 8-b-deep ECL-level patterns to 12 input pads. The pattern rate can be varied from 1.4 ns per cycle to over 20 ns per cycle using an on-chip voltage-controlled oscillator.

After a brief discussion of the CMOS technology and the physical organization of the chip, the pipelined archi-

Manuscript received April 30, 1991; revised June 27, 1991.
The authors are with IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.
IEEE Log Number 9102710.

TABLE I
512-kb CMOS SRAM CHARACTERISTICS

Organization	32 kb $\times$ 16
Timing	CS synchronous with addresses
Cycle Time	2 ns
CS Access Time	3.8 ns
Address Setup Time	0 ns
Power ($T_{\text{cycle}} = 5$ ns)	3.5 W
Power Supplies	$V_{DD} = 1.4$ V, $V_{REF} = 0$ V, $V_{SS} = -2.2$ V
Redundancy	8 word lines
I/O Interface	0 ± 0.5 V (IBM ECL)
Cell Size	7.5×11.4 μm
Chip Size	11.07×10.06 mm

TABLE II
CMOS TECHNOLOGY FEATURES

Average Feature Size	0.8 μm
n- and p-Channel L_{eff}	0.5 μm
Gate Oxide Thickness	12.5 nm
n-Channel Junction Depth	0.15 μm
p-Channel Junction Depth	0.24 μm
n^+ / p^+ Polycide Sheet Resistance	5 $\Omega / \square$ (Self-aligned TiSi_2)
Poly Pitch	1.5 μm
Metal 1 Pitch	1.9 μm
Metal 2 Pitch	2.5 μm
Metal 3 Pitch	3.3 μm
Cell Type	6T CMOS

texture of the chip will be described. Next, detailed measurements of internal chip waveforms demonstrating 2-ns cycle time operation will be presented. Then, the impact of wire RC delays on performance will be discussed. Circuit examples that demonstrate the implementation of the pipelined architecture are included in the next section. Following this is a section showing measurements of operating margins, access time, and cycle time. Finally, key results will be summarized.

II. CMOS TECHNOLOGY FEATURES

The CMOS process used in fabricating the 512-kb CMOS SRAM is summarized in Table II [8]. The three-level metal process has an average feature size of 0.8 μm and an aggressive 0.5- μm L_{eff} . Self-aligned TiSi_2 was used to reduce the sheet resistance of the source/drain diffusion regions and the polysilicon layer to 5 Ω/sq . Essentially all local wiring was done on the metal 1 and metal 2 layers. Metal 2 was also used for global wiring. Metal 3 was used only for global wire routing and power distribution.

III. PHYSICAL ORGANIZATION

The chip photomicrograph is shown in Fig. 1. For high performance, the chip is physically organized into 32 subarrays of 16 kb. Each subarray consists of 132 word lines by 128 bit-line pairs with eight bit switches connected to each of 16 sense amplifiers. The outside subarrays appear darker because the global word lines on metal 3, which emanate from the center of the chip, do not pass over the tops of these subarrays but are terminated at their connections at the edge of these subarrays.

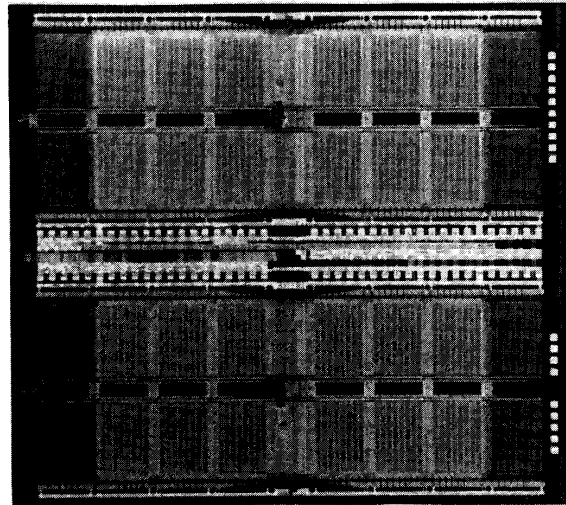


Fig. 1. Photomicrograph of 512-kb CMOS ECL SRAM.

Eighty pads are provided in the center horizontal alley of the chip for 49 I/O and 30 power supply connections. One pad is unused. Each I/O signal connects to only one pad. The 49 I/O signal pads are used as follows: 16 data-in, 16 data-out, seven X addresses, four Y addresses, four Z addresses, one write enable, and one chip select. The pads have been placed for C4 flip-chip packaging. Other pad placements that use only perimeter pads could be accommodated with < 100 -ps degradation in access time and with no degradation in cycle time.

To help match the delays of the word-line, bit-switch, and data-in signals, they are routed together, first across the horizontal center of the chip, then up the center vertical alley in the chip, and finally horizontally across metal-3 global wires to the subarrays. To help insure that a subarray is selected before these signals arrive, the subarray select signals follow a simpler path, which goes horizontally across the middle of the chip and then directly up or down to a subarray.

Data-out circuits are located at the far extremes of the center horizontal alley of the chip. Pads and circuits that facilitate testing of the word-line redundancy circuits and that provide on-chip generation of an 8-b-deep pattern for fast cycle time testing are located in the vertical alley on the right-hand side of the chip.

IV. DESCRIPTION OF PIPELINED OPERATION

Fig. 2 illustrates pipelined operation of the SRAM at a 2-ns cycle time. The start of four cycles and the completion of three of these cycles are shown. These operations can be any combination of READ or WRITE and they can be to any mixture of addresses. To allow pipelining, the circuitry in the critical path of the chip is divided into eight pipe segments. Each of these pipe segments is an input-triggered, self-resetting circuit block. The measured delay of each of these blocks is indicated approximately

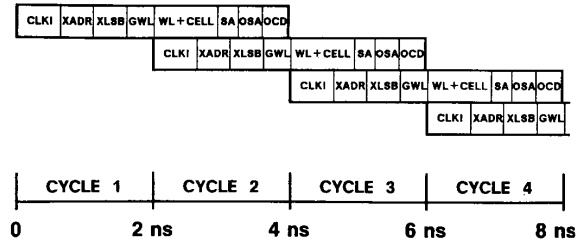


Fig. 2. Pipelined operation of 512-kb SRAM at 2-ns cycle time.

by the length of the labeled boxes. Each block can accept a new input every 2 ns. The pipe segments consist of: CLKI, which generates an on-chip clock pulse from the chip-select input signal; XADR, which amplifies all but one of the address input signals; XLSB, which amplifies the remaining address input signal and selects a normal or redundant global word line; GWL, which drives a global word line; WL+CELL, which contains the decoders/drivers and memory cells of a subarray; SA, which senses and amplifies the bit-line signal; OSA, which OR's together the outputs from the sense amplifiers in one half of the chip; and OCD, which provides CMOS-to-ECL level conversion of the output signals.

In detail, the pipelining in the chip works as follows. The first READ or WRITE operation begins at the start of Cycle 1. By the end of Cycle 1, the CLKI block will have completed its reset and a second READ or WRITE operation can begin at the start of Cycle 2. Meanwhile the first operation begun in Cycle 1 will continue to propagate through the various pipeline blocks of the chip. By the end of Cycle 2, the CLKI block will have completed its second reset and a third READ or WRITE operation can begin at the start of Cycle 3. Also at the end of Cycle 2, the READ or WRITE operation begun in Cycle 1 will have rippled through the eight pipeline blocks and the resulting data will appear on the data-out pins of the chip. Meanwhile the second operation begun in Cycle 2 will continue to propagate through the various pipeline blocks of the chip. This process continues for subsequent cycles. Like the CLKI block, all other circuit blocks shown in Fig. 2 operate at a 2-ns cycle time to support pipelined operation. Due to the pipelining shown in Fig. 2, the cycle time of the chip is significantly reduced, in this case to one-half of the READ access time.

V. PIPELINED ARCHITECTURE AND DESIGN FEATURES

The basic pipelined architecture described above has several benefits. With a cycle time half of the access time, the bandwidth of the SRAM is double that of a conventional chip with cycle time equal to access time. In addition, this bandwidth doubling can be achieved without sacrificing access time. In fact, it was found that improving the cycle time helped improve the access time, as will be shown in the circuit examples which follow. Also,

achieving the high bandwidth was not done by using a complex distributed clocking scheme. The SRAM uses only simple clocking of the input stages where data and addresses enter the chip. There is no separate distribution of a clock or other handshake signal to following pipeline stages. Instead, each stage communicates by sending a pulse to the next stage in the pipeline. This technique is like wave pipelining [9], [10], but instead of propagating a single transition at a time through the circuitry, a fully formed pulse is propagated. Each block in the pipeline is designed to respond quickly only to the leading edge of the pulse it receives. This significantly increases the operating speed of the SRAM while simultaneously relaxing the matching constraints on convergent paths due to the need to match the propagation delay of only one, not both, transitions through the circuit networks. We call this technique bubble pipelining [11].

The subarray is the cycle time bottleneck. Its cycle time is limited by the minimum word-line pulse width needed for a valid READ or WRITE operation plus the minimum precharge pulse width needed to fully restore the bit lines. When designing these pulse widths, all combinations of READ and WRITE operations need to be considered. The most difficult combination is a WRITE operation followed by a READ operation. This is due to the long precharge time needed after the WRITE to fully restore the completely discharged bit line before a valid READ can begin. A 2-ns cycle time for this combination was obtained.

Fast subarray cycle time is facilitated by the use of a full-CMOS, six-device memory cell that can be written and cycled quickly without sacrificing soft error rate, stability, or standby power.

All circuits external to the subarray provide cycle times less than the subarray cycle time. In order to keep pace with the fast cycle capability of the subarray, the chip critical path circuitry is divided into eight pipe segments, each of which is a self-resetting circuit block. A fast cycle time and minimum delay for each block is achieved by using fast forward amplification of the leading edge of the pulsed input signals, followed by quick self-resetting of all nodes back to their standby state. Forward amplification uses large devices in the critical path along with smaller devices for holding the standby state and for ensuring clean power-up initialization. Self-reset uses large devices, which subsequently turn off before a new cycle begins. This combination of devices results in each block being responsive to only one edge of an input pulse and in each block producing a pulsed output signal. The use of these devices will be further detailed in the circuit examples below.

At least four separate factors contribute to ensuring good operating margins of the pipeline circuit blocks. First, the output-pulse width of each block is designed to be three times that needed to fully switch the input of the next pipeline block. Second, wherever pulses intersect, an adequate operating margin is obtained by either providing a relatively wider pulse or by capturing the earlier arriving

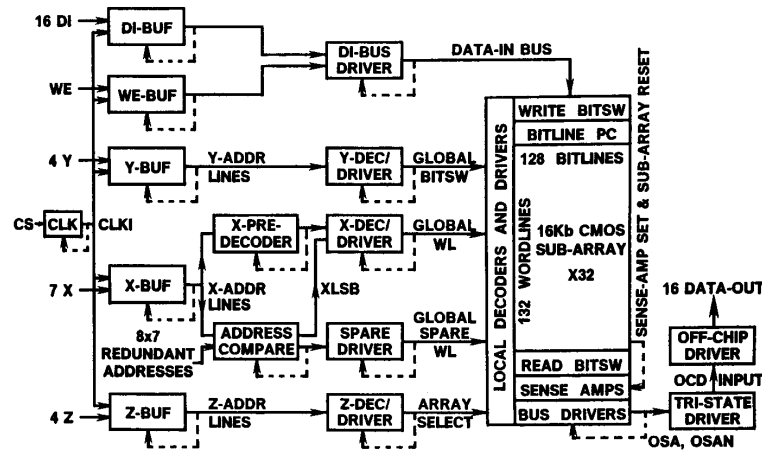


Fig. 3. Block diagram of 512-kb CMOS ECL SRAM.

pulse with a latch or by dynamic storage. The captured signal is cleared during reset, so as to not interfere with a subsequent cycle. Third, the standby devices described above are made small enough to have minimal impact on delay and output-pulse width but are large enough to overwhelm noise sources. Finally, the 512-kb CMOS array provides over 20 nF of on-chip decoupling capacitance, which greatly helps to minimize power bus noise. This capacitance is distributed over the entire chip and was carefully tied together into a low-impedance net using all three layers of metallization.

VI. BLOCK DIAGRAM OF OPERATION

The pipelined architecture of the SRAM is shown in the block diagram of Fig. 3. The memory array is divided into 32 subarrays of 16 kb, each organized as 132 rows (four are for redundancy) by 128 columns, and containing a 16-b data I/O path. Word-line redundancy circuits with negligible impact on chip performance are included. The chip circuitry is organized into six major systems: data-in, X row select, Y column select, Z subarray select, subarray, and data-out. Successive pipeline stages in these systems are not clocked; the CLKI signal generated from chip select is used only to set the input buffers.

Self-timed reset, indicated by the dashed lines, is integrated into each major block in every system. The only exception to this is in the data-out system, where the output of the tri-state driver circuit is captured with a latch to provide a data window equal to cycle time.

The chip operates in a fully pulsed mode. Each block is activated by the leading edge of its input signal and then begins its self-timed reset off the leading edge of its output signal. The trailing edge of the output pulse turns off the reset devices inside the block in preparation for the next cycle. Pulses are typically 0.6 to 1 ns wide.

For adequate operating margin, the data-in bus pulse is 1.5 ns wide, which ensures its intersection with the X and Y selection pulses at the subarray. The Z subarray select

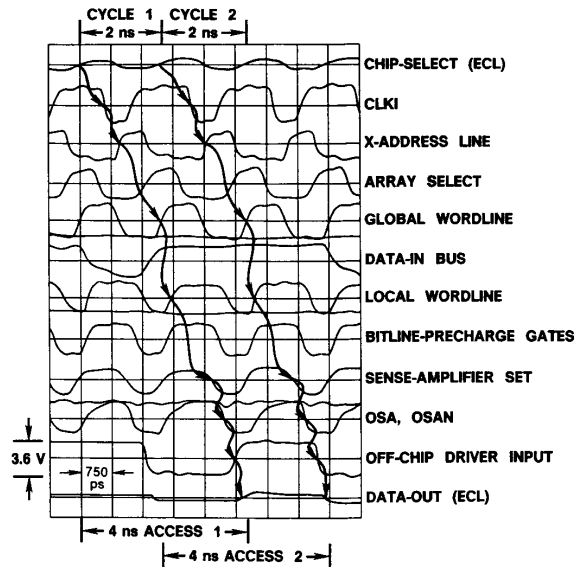
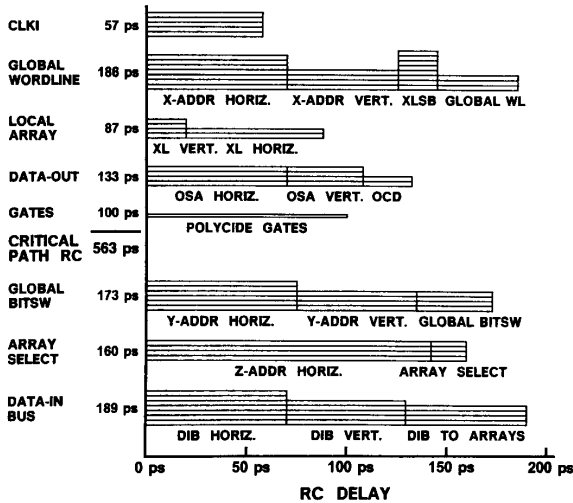


Fig. 4. Measured waveforms of self-resetting circuit blocks at 2-ns cycle time.

pulse arrives earlier than these pulses and is captured by dynamic storage on a 6-pF bus inside the subarray.

VII. MEASURED PIPELINED WAVEFORMS AT 2-NS CYCLE TIME

The above described characteristics are evident in the measured internal signal waveforms, shown in Fig. 4, of key circuit blocks operating at a 2-ns cycle time. These waveforms were taken with a 3-GHz picoprobe feeding an 18-GHz sampling oscilloscope. As can be seen, even at a 2-ns cycle time, every waveform has a well-defined active and standby period. Two pipelined READ cycles are highlighted by the pointers. The access time for each of these

Fig. 5. Calculated wire *RC* delay components.

READ cycles is indicated at the bottom of Fig. 4 and is less than 4 ns. WRITE operations, indicated by the wide, low level on the data-in bus, precede and follow the READ cycles. As can be observed, the data-in bus signal surrounds the local word-line signal, which ensures a good WRITE margin for the cell. The data-in bus measurement was taken at the end of the wiring net and the voltage drop due to discharging the bit line is evident in the waveform.

The data-in bus, subarray select, and global word-line pulses shown in Fig. 4 intersect at the subarray. To insure adequate operating margins for these signals, a wide data-in bus pulse is used while the earlier arriving subarray select pulse is captured with dynamic storage.

This figure illustrates the extensive pipe segmentation used in the peripheral circuitry to keep pace with the fast cycle capability of the CMOS array.

VIII. IMPACT OF WIRE *RC* DELAYS

The calculated wire *RC* delays in all of the systems on the chip are shown in Fig. 5. The length of the bars shows the *RC* delay contributions of every major wire in each system. The number of stripes in a bar is proportional to the physical wire width.

The top five bars show the delay components in the critical path of the chip. These add up to a total of 563 ps or just under 15% of the measured access time. To achieve this value, it was necessary to consider wire *RC* delays in the early chip floor planning and then to use larger than minimum width wires on all three levels of metallization. Also, interdigitated devices with a large number of fingers were used wherever possible to help minimize intrinsic gate delay. Finger lengths of 30 μm or less were used to reduce the total intrinsic gate delay to 2.5% of the chip-select access time.

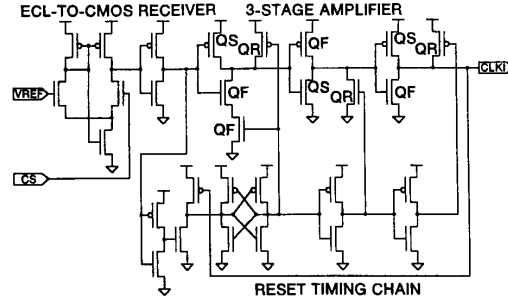


Fig. 6. Chip-select clock circuit.

The longest four bars show the delays in the *X*, *Y*, *Z*, and data-in systems which feed the subarray. These four systems have nearly equal *RC* delay components. This helps minimize the variation in the arrival times of the pulses from these systems at the subarray.

Since the critical path of the chip is divided into eight pipe segments, the impact of wire *RC* delays on cycle time is much less than it is on access time. Only *RC* delays which are within the reset loop of a pipe segment will have an impact on cycle time. These *RC* delays are small and do not appreciably increase the cycle time of this SRAM.

IX. EXAMPLES OF SELF-RESETTING CIRCUIT BLOCKS

A. Chip-Select Clock

The chip-select clock block shown in Fig. 6 is used to synchronously clock all other inputs. This circuit is an example of a self-resetting circuit block for amplifying a single-rail signal. The circuit provides two-stage, asynchronous, ECL-to-CMOS conversion of the chip-select input signal with 100-mV worst-case sensitivity [7] and further three-stage amplification to drive a 6-pF on-chip load in 650 ps. In this circuit example and the two to follow, the critical path nodes have large devices for forward drive (*QF*) and for reset (*QR*) in order to minimize delay and cycle time, respectively, and smaller devices (*QS*) to hold the standby state. The CLKI output signal is looped back through a timing chain to reset the three-stage amplifier and produce a 650-ps-wide CLKI pulse. To minimize the impact of variations in p-to-n tracking on CLKI pulse width, the reset loop uses five inversions rather than three so that the numbers of p- and n-channel devices in the loop are more closely balanced. The latch in the reset circuitry insures that only one CLKI pulse is generated when chip select is held high.

Since the chip-select block is self-resetting, an aggressive taper is used in the forward path of the three-stage amplifier to minimize delay without degrading cycle time. Without self-reset, the aggressive taper used in the amplifier would limit its cycle time to over 4 ns. Then achieving the modest goal of making cycle time equal to access time would require reducing the taper of the amplifier at the

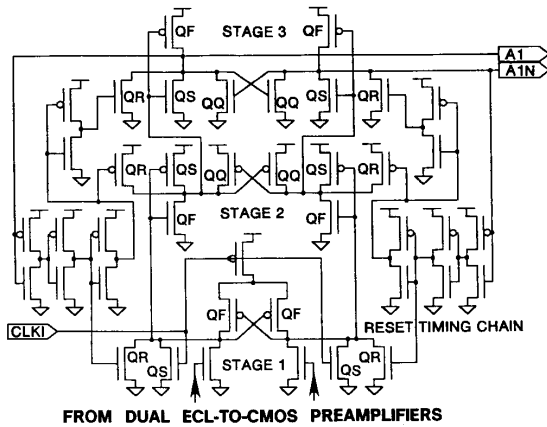


Fig. 7. ECL-to-CMOS address buffer circuit.

cost of an increase in access time. Thus, the pipelined architecture using self-resetting circuit blocks helps to reduce access time while it improves cycle time.

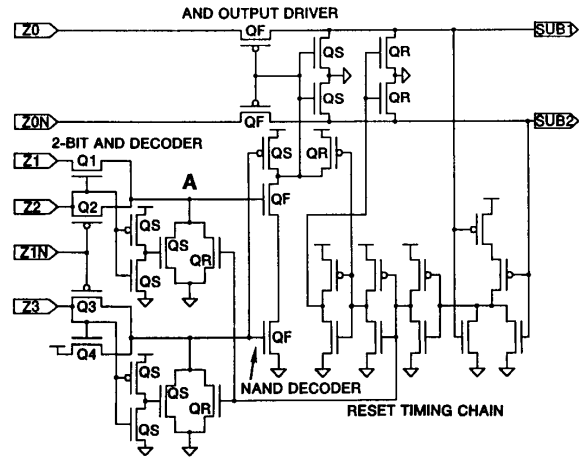
B. ECL-to-CMOS Address Buffer

The ECL-to-CMOS address buffer shown in Fig. 7 is an example of a self-resetting circuit block that provides complementary outputs. It is used to buffer the *DI*, *WE*, *X*, *Y*, and *Z* inputs shown in Fig. 3. This circuit consists of a pair of asynchronous ECL-to-CMOS preamplifiers (like the one shown in Fig. 6) which provides a differential-input signal to the dual-rail, three-stage, self-resetting, clocked amplifier shown in Fig. 7. When the *CLKI* input signal falls, the differential input signal is amplified to drive a 2-pF on-chip load in 350 ps. In addition to pairs of *QF*, *QR*, and *QS* devices, the critical nodes in the amplifier also contain cross-coupled pairs (*QQ*) for improved active noise margin. For minimum cycle time and a completely symmetric layout, each side of the amplifier has its own reset loop with five inversions.

Because the address buffer is self-resetting, over 90% of the *CLKI* input signal current is devoted to clocking the cross-coupled p-latch in Stage 1 and less than 10% is needed to maintain the standby state of the buffer. In contrast, a previous nonpipelined RAM [12] devoted over 50% of the *CLKI* signal to resetting the buffer and maintaining its standby state. This improvement can be used to reduce the total delay through the *CLKI* and address buffer blocks. This is another example of how the use of pipelining to improve the cycle time of the chip also helps to reduce the chip-select access time.

C. Z Decoder

The *Z* decoder is an example of a self-resetting circuit block incorporating a 4-b decode function and is shown in Fig. 8. This block is used to select one of 16 subarrays on each side of the chip. The *Z* decoder has a delay of only 500 ps driving a 4-pF on-chip load. Besides providing a

Fig. 8. 4-b *Z*-decoder circuit.

well-controlled pulse width and a fast cycling capability using the techniques described above, this circuit exploits the complementary pulses from the address buffer for high-speed decoding using a novel 2-b AND for the first stage.

The 2-b AND decoder works as follows. The inputs to the decoder are low in standby, with either the true or complement of each input pulsing high for selection. If both *Z1* and *Z2* are pulsed high, then *Q1* and *Q2* act like a transmission gate, rapidly driving the gate of the *QF* device to a high selected state. This device combination also prevents a false select due to any address skew causing *Z2* to rise before *Z1N* rises, since *Q1* is made large enough to clamp node *A* low until *Q2* turns off.

The outputs from two of these AND decoders are inputs to the second-stage NAND decoder. The output of the NAND decoder is input to the final 1-out-of-2 selection stage. While the AND and NAND decoders are resolving, the *Z0* or *Z0N* signal falls from its high selected standby state to deselect either the *SUB1* or *SUB2* output, respectively. A 5-b decoder can be made by routing the drain of *Q4* and the gate of *Q3* to the true or complement of a fifth input bit.

D. CMOS-to-ECL Off-Chip Driver

The CMOS-to-ECL off-chip driver (OCD) is shown in Fig. 9. An easily overridden input latch captures the CMOS-level signal of the tri-state OCD input line. The two-stage feedback amplifier then converts this signal to an ECL-level output signal capable of driving a 50-Ω transmission line load in less than 360 ps. To improve level control, the negative feedback device is switched to match the conductivity type of the output driver. By overlapping the switching of the feedback devices, undershoot and overshoot on node *A* are reduced, helping to limit the maximum slew rate on the *DO* output to 2 V/ns. The off-chip driver provides a nominal ± 500 -mV

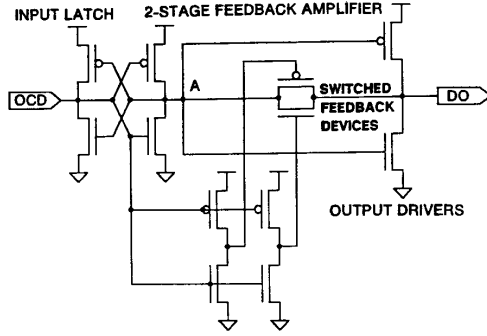


Fig. 9. CMOS-to-ECL off-chip driver circuit.

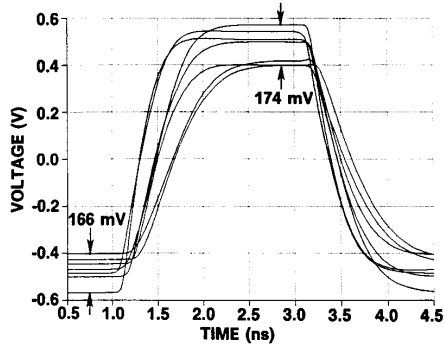
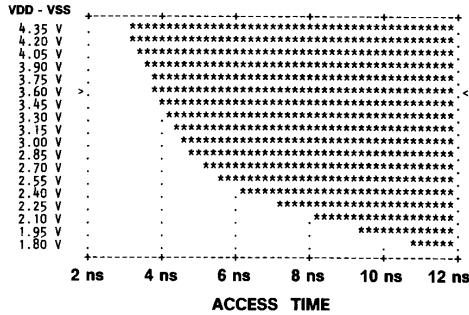
Fig. 10. Simulated level control of CMOS-to-ECL off-chip driver when driving a 50- Ω load terminated to V_{REF} (0 V). Simulation results are shown for worst-case combinations of $\pm 30\%$ L_{eff} , ± 150 -mV threshold voltage, and $\pm 10\%$ power supply variations.

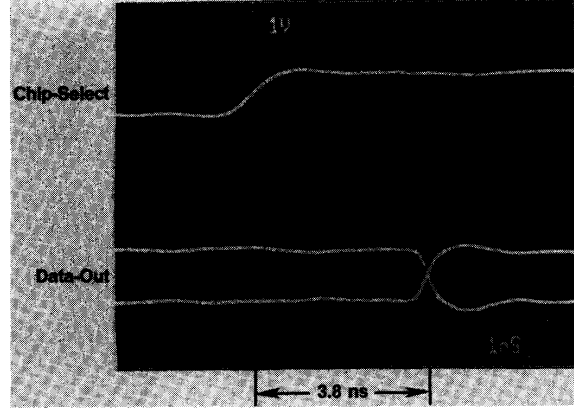
Fig. 11. Schmoo plot of chip-select access time.

output swing with a simulated ± 90 -mV level control (or 180-mV spread) for worst-case combinations of $\pm 30\%$ L_{eff} , ± 150 -mV threshold voltage, and $\pm 10\%$ power supply variations, as shown in Fig. 10.

X. MEASUREMENT RESULTS

A. Schmoo Plot of Chip-Select Access Time

The room-temperature Schmoo plot shown in Fig. 11 demonstrates that proper chip operation was maintained

Fig. 12. Chip-select access time of 512-kb CMOS ECL SRAM with 50- Ω output termination.

over a wide power supply range, from less than 2 V to over 4.3 V. Schmoo plots at 85°C showed similar good results.

Proper chip operation was also confirmed with N and N^2 pattern tests at cycle times of 2.5 and 5 ns, respectively, as limited by the available commercial tester. At a 2.5-ns cycle time, proper operation was maintained when power supplies were reduced more than 20% from nominal.

B. Chip-Select Access Time

The measured chip-select access time with all word-line redundancy activated is 3.8 ns for the worst-case path through the chip, as shown in Fig. 12. This fast access time is achieved even at the fastest cycle time, as shown in the next section.

C. 2-ns Cycle Time WRITE / READ Waveforms

Evaluation of chip operation under minimum cycle conditions uses on-chip test circuitry, which can provide 8-b-deep ECL-level patterns to 12 input pads. The pattern rate can be varied from 1.4 ns per cycle to over 20 ns per cycle, using an on-chip voltage-controlled oscillator. Fig. 13 shows measurements of chip operation at 515 MHz, or slightly less than 2-ns cycle time, using the on-chip test circuitry to first write and then read the ZERO and ONE logic states of two different memory cells (A and B) on one bit line. This particular sequence, which places WRITE's and READ's of opposite logic states next to each other, demonstrates that the bit line is adequately re-stored following a WRITE operation when the chip is cycling at less than 2 ns.

XI. SUMMARY AND CONCLUSIONS

In summary, this 512K CMOS SRAM has achieved a 2-ns cycle time and a 3.8-ns chip-select access time, both of which are valid for fully random READ/WRITE operations. This high performance is due to the combination of

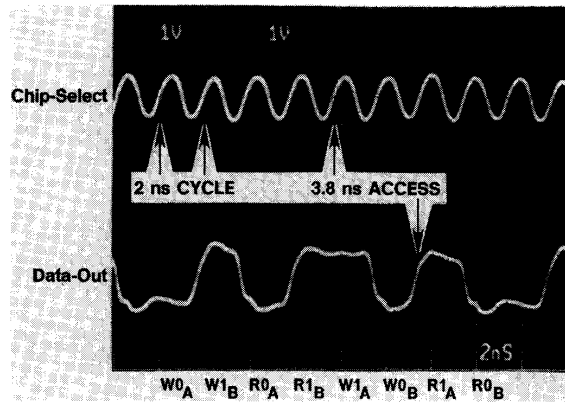


Fig. 13. Measured waveforms of chip select and data-out at 2-ns cycle time. Chip select, data-in, write enable, and $X0$, provided by on-chip pattern generators, are sequenced to write and read the ZERO and ONE logic states of two memory cells (A and B) on one bit line.

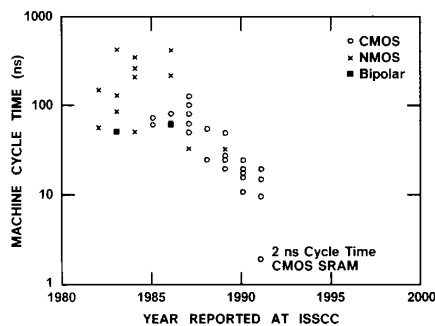


Fig. 14. Comparison of silicon microprocessor cycle times reported at ISSCC to the 2-ns cycle time of the 512-kb CMOS SRAM.

an aggressively pipelined architecture having eight pipe segments, which are implemented with self-resetting circuit blocks, and a $0.8\text{-}\mu\text{m}$ CMOS technology having a $0.5\text{-}\mu\text{m}$ L_{eff} . A fast cycle time for each of the self-resetting circuit blocks was achieved by making extensive use of fast forward amplification followed by quick self-resetting of all nodes back to their standby state. Careful minimization of wire RC delays reduced their impact to less than 15% of the chip-select access time.

The chip has good operating margins due to at least four factors. First, only simple clocking of input stages was used. Second, the internal pulses generated by the self-resetting circuit blocks are made much wider than needed to fully switch the following block, and where pulses intersect they are made even wider or are captured with a latch. Third, devices that hold the standby state are made small enough to have minimal impact on delay and output-pulse width but are large enough to overwhelm noise sources. And fourth, the CMOS array provides over 20 nF of on-chip decoupling capacitance, which greatly reduces the noise on internal power buses. Due to these factors, reliable chip operation was observed from less than 2 V to over 4.3 V at temperatures up to 85°C .

To put the 2-ns cycle time of this SRAM in perspective, its cycle time is compared in Fig. 14 with the cycle time of silicon microprocessors reported at ISSCC. As can be seen, SRAM cycle time need not limit microprocessor cycle time through the end of this decade.

ACKNOWLEDGMENT

The authors appreciate the support of L. Terman, T. Ning, R. Isaac, H. Yu, and the Yorktown Silicon Facility Staff.

REFERENCES

- [1] M. Odaka *et al.*, "A 512Kb/5 ns BiCMOS RAM with 1kG/150 ps logic gate array," in *ISSCC Dig. Tech. Papers*, Feb. 1989, pp. 28–29.
- [2] F. Towler *et al.*, "A 128K 6.5 ns access/5 ns cycle MOS ECL static RAM," in *ISSCC Dig. Tech. Papers*, Feb. 1989, pp. 30–31.
- [3] M. Takada *et al.*, "A 5 ns 1 Mb ECL BiCMOS SRAM," in *ISSCC Dig. Tech. Papers*, Feb. 1990, pp. 138–139.
- [4] T. Chappell *et al.*, "A 2 ns cycle, 4 ns access 512Kb CMOS ECL SRAM," in *ISSCC Dig. Tech. Papers*, Feb. 1991, pp. 50–51.
- [5] D. Wendell *et al.*, "A 3.5 ns, 2K $\times$ 9 self timed SRAM," in *1990 Symp. VLSI Circuits*, June 1990, pp. 49–50.
- [6] J. Beyers *et al.*, "A 32b VLSI system," in *ISSCC Dig. Tech. Papers*, Feb. 1982, pp. 128–129.
- [7] B. Chappell *et al.*, "Fast CMOS ECL receivers with 100-mV worst-case sensitivity," *IEEE J. Solid-State Circuits*, vol. 23, no. 1, pp. 59–67, Feb. 1988.
- [8] L. Wang *et al.*, "0.5 micron gate CMOS technology using E-beam/optical mixed lithography," in *1986 Symp. VLSI Technology Dig. Tech. Papers*, May 1986, pp. 13–14.
- [9] S. Anderson *et al.*, "The IBM system/360 model 91 floating point execution unit," *IBM J. Res. Develop.*, pp. 34–53, Jan. 1967.
- [10] L. Cotten, "Maximum rate pipelined systems," in *AFIPS Proc. Spring Joint Computer Conf.*, 1969, pp. 581–586.
- [11] M. Beakes, private communication, May 1991.
- [12] T. Chappell *et al.*, "A 3.5-ns/77 K and 6.2-ns/300 K 64K CMOS RAM with ECL interfaces," *IEEE J. Solid-State Circuits*, vol. 24, pp. 859–868, Aug. 1989.



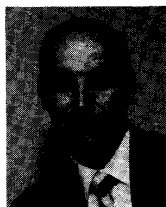
Terry I. Chappell received the B.S. degree in physics from Harvey Mudd College, Claremont, CA, in 1972 and the Ph.D. degree in electrical engineering from the University of California at Berkeley in 1978.

In the course of his doctoral research, he was employed in the Integrated Circuit Engineering Department at Tektronix in Beaverton, OR, in the Computer Aided Design Group at Sandia Laboratories in Albuquerque, NM, and in the Silicon Photovoltaic Device Group at the RCA David Sarnoff Research Center in Princeton, NJ. In 1978 he joined IBM at the Thomas J. Watson Research Center, Yorktown Heights, NY, where he is currently a Research Staff Member, working primarily in the field of VLSI memory design and systems.



Barbara A. Chappell (M'85) received the B.S.E.E. degree from the University of Portland, Portland, OR, in 1977 and the M.S.E.E. from the University of California at Berkeley in 1981.

In 1978 she joined IBM at the Thomas J. Watson Research Center, Yorktown Heights, NY, where she is currently a Research Staff Member, working primarily in the field of VLSI memory design and systems. Her previous employment included ten years with the Custom IC Department at Tektronix, Beaverton, OR.

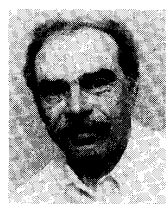


Stanley E. Schuster (S'61-M'65-SM'89-F'90) received the B.S. and M.S. degrees in electrical engineering from New York University, New York, in 1962 and 1969, respectively.

From 1962 to 1963 he was with the Bendix Corporation, and from 1963 to 1965 he served as an officer in the U.S. Army Signal Corps. In 1965 he joined the IBM Research Division, Yorktown Heights, NY, working initially on the effort that led to IBM's NMOS technology for main memory. His work in the early 1970's on

redundancy was instrumental to IBM's implementation of redundancy in DRAM products and included the development of laser personalization techniques which were also used for fast turnaround time in the personalization of PLA's. Since the late 1970's he has worked on high-speed NMOS and CMOS SRAM's. Currently he is Manager of the SRAM group in the VLSI logic and memory area.

Mr. Schuster is an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS and was Guest Editor for the Special Issue on Logic and Memory in October 1986. He served on the ISSCC Program Committee from 1985 to 1988. He has received six IBM Outstanding Invention and Contribution Awards and the ninth level IBM Invention Achievement Award.



James W. Allan joined IBM in 1964 at Poughkeepsie, NY, where he worked in the Ferrite Core Development Group. In 1970 he transferred to IBM in East Fishkill, NY, to do logic design and system test with the MLC Test Equipment Support Group. Since 1976 he has been doing chip physical design in CMOS, bipolar, and Josephson technology at the IBM Thomas J. Watson Research Center in Yorktown Heights, NY.



Stephen P. Klepner was born in New York, NY, on April 15, 1942. He received the B.S.E.P. and Ph.D. degrees from New York University, New York, in 1962 and 1969, respectively.

In 1969 he joined the IBM Corporation in East Fishkill, NY, and worked on aspects of bipolar and MOS technology. In 1976, at Yorktown Heights, NY, he became involved in Josephson processing and testing. In 1983 he resumed his work on MOS processing.



Rajiv V. Joshi (M'89) received the B.Tech. degree from the Indian Institute of Technology, Bombay, the M.S. degree from the Massachusetts Institute of Technology, Cambridge, and the Ph.D. degree from Columbia University, New York, NY.

From 1981 to 1983 he worked on the evaluation and development of carbide, nitride, and ceramic cermets at GTE Laboratory, Waltham, MA. He joined IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in 1983

and since then has worked in systems, science, and technology. He worked on the tungsten-silicide process used in IBM's 1.25- μm NMOS and 0.5- μm CMOS DRAM technologies. He developed a high-pressure CVD W process, which is currently used in several IBM products and in the semiconductor industry. His recent work includes process integration for CMOS and bipolar technologies utilizing inorganic and polyimide dielectrics. His research interests include thin-film processing, advanced metallization schemes, and development of back-end technologies for VLSI and ULSI circuits. He has authored and co-authored over 40 research papers, presented four invited talks, and holds seven U.S. patents.

Dr. Joshi is a member of ASME. He has received four IBM invention achievement awards.



Robert L. Franch received the B.S.E.E. degree from the Polytechnic Institute of New York, Brooklyn, in 1980.

In 1980 he joined IBM, East Fishkill, NY, in a Bipolar Device Reliability Group, where he worked on accelerated life testing of bipolar logic, memory, and test vehicle chips. In 1984 he joined IBM Research in Yorktown Heights, NY, as a member of the Test Systems Group. He has since been engaged in the functional testing of NMOS, CMOS, and bipolar logic and memory

chips developed at IBM Research.