A 15-ns CMOS 64K RAM

STANLEY E. SCHUSTER, MEMBER, IEEE, BARBARA A. CHAPPELL, MEMBER, IEEE, ROBERT L. FRANCH,
PAUL F. GREIER, STEPHEN P. KLEPNER, FANG-SHI J. LAI, MEMBER, IEEE,
PETER W. COOK, MEMBER, IEEE, ROBERT A. LIPA, MEMBER, IEEE, REGINALD J. PERRY,
WILLIAM F. POKORNY, AND MICHAEL A. ROBERGE

Abstract —This paper describes a 64K CMOS RAM with an access time of 15 ns. The RAM was built using a technology with self-aligned $TiSi_2$, single-level metal, an average minimum feature size of 1.35 μm, and a minimum effective channel length of 1.1 μm. An access of 10 ns is possible with the word line stitched on a second level of metal and some minor redesign. High speed is achieved through innovative circuits and design concepts. New CMOS circuits include a sense-amp set signal generator, a row decoder, and an input circuit. These circuits feature use of CMOS devices to an advantage for high-speed safe operation. A layout-rule-independent graphics tool was used for the artwork design.

I. INTRODUCTION

THE DRAMATIC reduction that is taking place in memory access time can be clearly seen in the plot of access time versus year for SRAM's presented at the ISSCC [1]–[16] shown in Fig. 1. FET memories at high levels of integration have moved into the very high-performance area. This downward trend in access time should continue into the foreseeable future. At the 1984 ISSCC we presented a 20-ns 64K NMOS design [5]. Also included on the plot is a 0.78× scaling of that design presented at the 1985 International Symposium on VLSI Technology, Systems, and Applications which gave access times as fast as 11 ns [10]. In this paper we will describe a 64K CMOS RAM with measured access times of under 15 ns and simulated access times of 10 ns with the addition of a second level of metal and some minor redesign.

The characteristics of the 64K CMOS RAM are given in Table I. The high speed of this CMOS RAM is due to a combination of technology and innovative CMOS peripheral circuitry. After a brief description of the technology, three of the key circuits will be described: the sense-amplifier set generator, the row decoder, and the input circuit. In each case, the advantageous use of CMOS devices for high speed while maintaining low-power safe

Fig. 1. Plot of access time versus year for SRAM's presented at the ISSCC.

TABLE I
64K CMOS RAM CHARACTERISTICS

Organization	64K (4K x 16)
Cell Type/Area	4-D *NMOS*/210μm²
Access Time	15ns
Cycle Time	≤ 1.5 T_{ACC}
Supply	5V

operation will be featured. In addition, the use of this chip to demonstrate a layout-rule-independent physical design tool will be discussed. The use of a high-performance memory chip as a test vehicle served as a challenging demonstration of the potential of the tool.

II. TECHNOLOGY

The RAM was built using a relatively straightforward CMOS technology with only a single level of metal [17]. Process parameters are given in Table II. A cross section

TABLE II
CMOS TECHNOLOGY

- Single level metal

- Self-Aligned TiSi$_2$ (5 Ω's/$\square$) over polysilicon and n and p diffusions.

- 22.5 nm gate insulator thickness

- 1.1 μm and 1.2 μm L_{eff} for n- and p-channel devices respectively

- 1.35 μm average minimum feature size

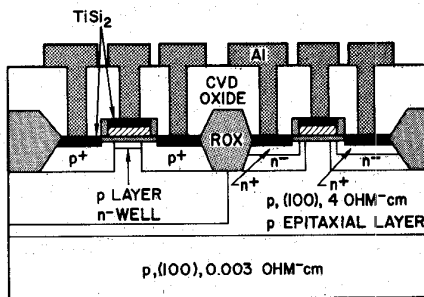- Junction depth 0.25 μm for n-channel device and 0.30 μm for p channel device.

Fig. 2.   Cross section of the CMOS structure (from [17]).

of the CMOS structure is shown in Fig. 2. The main features of the technology include:

1) a 1-MeV ion-implanted retrograde n-well;
2) arsenic–phosphorous double diffused n^+/n^- junctions for the n-channel devices to improve the drain breakdown voltage and hot-electron reliability;
3) a self-aligned TiS$_2$ process with a nitride spacer to reduce the sheet resistances of both polysilicon gates and diffusions; and
4) a 4-μm-thick p-type epitaxial layer grown on a very heavily doped substrate to increase latch-up immunity.

The cell array for this chip was taken from a previous 64K NMOS design (see [5] for a cell layout drawing). It is a four-device cell to which resistors could be added on a second level of poly. The addition of resistors would make the cell fully static and would require no change in the physical or electrical design of the four-device portion of the cell. Thus the RAM performance would be unaffected by the addition of load resistors to the cell. The cell stability and soft error rate with and without load resistors were simulated using conservative assumptions and an analysis methodology that includes transient effects which are important whether dynamic storage or high-resistance loads are used [18]. The cell stability and soft error rate were found to be adequate for several important system applications without the addition of cell load resistors.
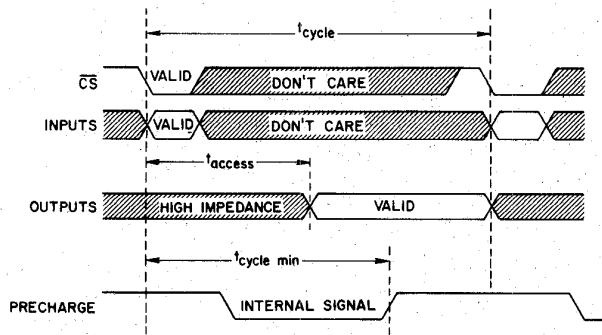
Fig. 3.   Waveforms showing 64K CMOS RAM operation.

III.   CHIP OPERATION

Chip operation, as shown in the waveforms of Fig. 3, differs from the more conventional approaches which use address-transition detection to initiate the timing chain. In this design a cycle is initiated only by $\overline{CS}$ falling. The inputs are sampled for a short period of time, then all inputs, including $\overline{CS}$, are disconnected until output data have become valid and the precharge of the chip has begun. A longer than minimum cycle is shown. For a minimum cycle, inputs would have to be valid when the precharge of the chip has begun and a new cycle is initiated, as indicated on the figure. The approach offers a number of advantages, listed below, that typically are not available with more conventional approaches.

1) The chip can be operated at minimum cycle time since inputs may be changed during an access.

2) The chip is insensitive to glitches on the inputs once the short sampling period at the beginning of a cycle ends and the inputs are disconnected from the internal chip circuitry.

3) Data outputs are always latched in a valid state or are in a high-impedance state, except when they are in transition.

4) Precharging of internal nodes is automatically initiated at the end of an access.

5) The chip has the same cycle time for any combination of READ and WRITE operations even if data-in and data-out pins are shared.

IV.   KEY CIRCUITS

The development of new CMOS peripheral circuitry was key to the high-speed access which was a major objective of the 64K CMOS RAM design. Fig. 4 shows a simplified block diagram of the access path, with the delay through each block indicated. In most of the access path, data simply ripple from block to block, with one block activating the next one. Care was taken to achieve a uniform distribution of delay throughout the critical path. Three of the more important new circuits developed for this design will be described: the self-timed array and sense-amplifier circuitry; the row decoder which uses an innovative two-stage NOR and NAND decoder, and the address buffer
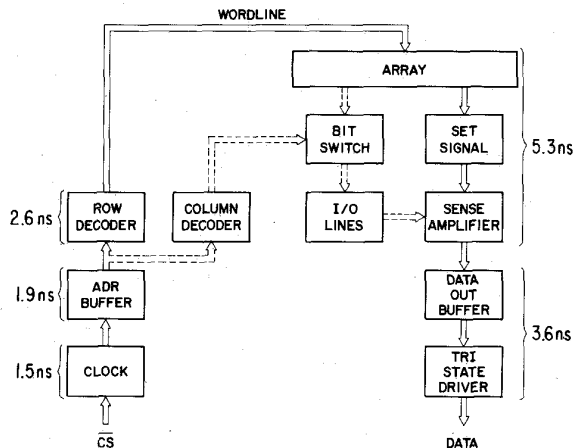
Fig. 4. Simplified block diagram and delay of the 64K CMOS access path.

Fig. 5. Array and self-timed sense-amplifier circuitry.

which uses a nonlinear front end and a self-referencing CMOS latch.

A. Sense-Amplifier Circuitry

The sense amplifier has several unique features including:

    a) a sense-amplifier setting waveform that has two distinct slopes for a slow and fast set;

    b) a technique for generating the setting signal so it is timed for the accessed word line using p- and n-channel devices; and

    c) p-channel decoupling devices between the sense amplifier and the I/O lines for faster setting.

The unique features of the sense-amplifier design result in both very high performance and reliable operation over wide parameter variations. This has been confirmed by simulations and actual hardware results.

The array and sense-amplifier circuitry are shown in Fig. 5. During a READ or WRITE operation, a row and a column decoder will be selected. The selected row decoder will cause its associated word line to go high and the selected column decoder will turn on the gates of the n and p complementary parallel bit-switch devices. The use of dual bit switches is necessary to avoid threshold drops in propagating the cell signal from the bit lines to the I/O lines during a READ or in propagating the signal in the reverse direction during a WRITE. Since the bit lines and I/O lines are high at the start of a READ cycle, the p-channel device forms the best path for conducting the signal. When an I/O line is set to a low level during a WRITE, the n-channel bit switch provides the best path for discharging the bit line to a good low level.

At the end of each word line is the sense-amp set signal generator circuit. As the selected word line rises, it turns on all the memory cells along its length and its set signal generator. A differential voltage builds up across the bit-line pairs as a result of the memory cells turning on. On one of the bit lines the differential voltage propagates
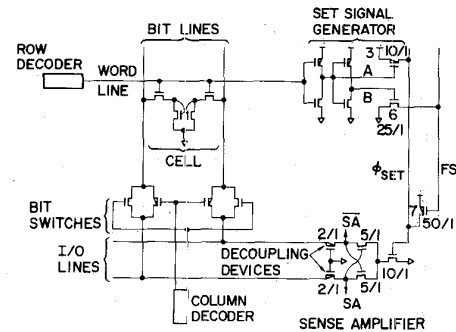
through the selected bit switch onto the I/O lines and sense amplifier. As adequate voltage across the sense-amplifier nodes develops, the fast and slow signal from the set signal generator causes the sense amplifier to latch.

The set signal generator of Fig. 5 is connected both to the ϕ_{SET} line and the FS line. Prior to a word line rising, the ϕ_{SET} line is precharged low and the FS line is precharged high. As the word line rises, the output (node A) of the first inverter stage of the set signal generator falls. Node A falling turns on the 10/1 p-channel device 3, which causes ϕ_{SET} to rise in its slow set mode of operation. A short time later the output (node B) of the second stage of the set signal generator will rise, causing n-channel device 6 to turn on, which in turn discharges the FS line to a low level. Device 7 (a large 50/1 p-channel device) connects the FS and ϕ_{SET} lines. When the FS line discharges, device 7 turns on and causes ϕ_{SET} to rise in its fast set mode of operation. The slow and fast set slopes and the delays between them can be adjusted by changing the sizes of the devices in the set signal generator and device 7.

In addition to the slow and fast set signal and self-timing from the accessed word line, high-speed operation is further improved by the small p-channel decoupling devices between the small capacitance nodes of the sense amplifier (SA and SAN) and the high capacitance I/O lines. These decoupling devices make it possible to set the sense amplifier much faster for the same differential signal compared to a sense amplifier without decoupling devices. The SA and SAN nodes are directly connected to the data-out buffer for further amplification before the signal is driven off-chip.

Simulated sense-amplifier waveforms are given in Fig. 6. The two distinct slopes of the ϕ_{SET} signal can be clearly seen. The smooth transition in slope from slow to fast occurs in conjunction with the increased differential voltage build-up across the sense-amplifier nodes. It can also be seen that the small p-channel decoupling devices make it possible to set the sense amplifier as ϕ_{SET} rises without having to discharge the large bit-line or I/O line capacitances.

Extensive simulations have demonstrated that the design of the set signal generator results in reliable performance even if there are substantial parameter variations. Since the set signal is generated from the selected word line, sensitiv-
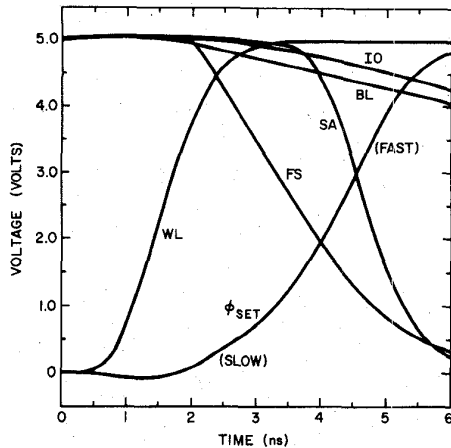
Fig. 6.   Simulated sense-amplifier waveforms assuming the use of a second level of metal to stitch the word line.
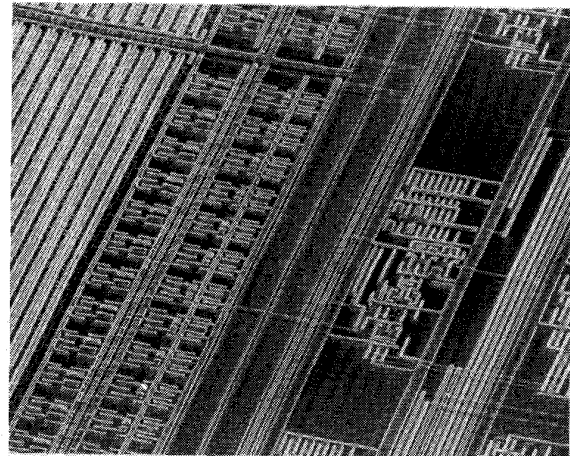

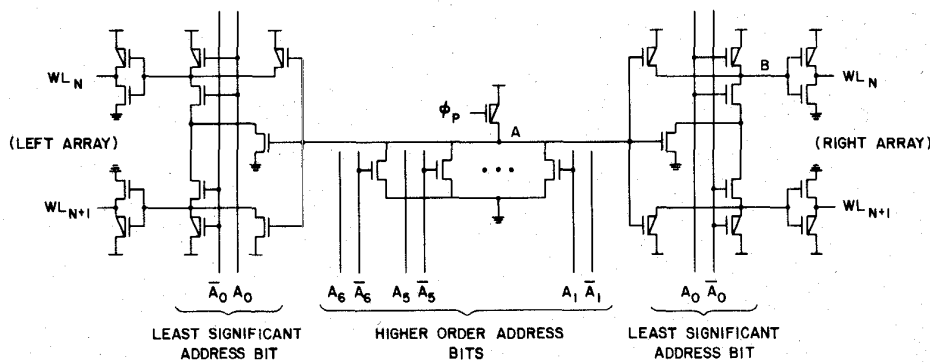Fig. 7.   SEM of array and sense-amplifier circuitry.

Fig. 8.   Row decoder with two stages of decoding.

ity to timing skews is limited to the path through the set signal generator to the sense amp relative to the path through the array to the sense amp. Within these sensitive paths, timing variations due to parameter variations are limited by a number of compensating factors. The use of both p and n devices in the set generator and in the array signal path (n cell access device, p bit switch, p decoupling device) tends to compensate for shifts in p thresholds relative to n thresholds. The use of a double inversion in the set signal generator tends to compensate for shifts in the supply voltages. The relatively small device count in the set generator helps to contain sensitivity to errors between devices of the same type on the same chip. Errors due to on-chip variations in capacitances can be compensated by designing the *FS* line and the ϕ_{SET} line to have capacitance components similar to those of the bit lines.

An SEM of the array and sense-amplifier circuitry is shown in Fig. 7. The set signal generators are at the end of the word lines. As can be seen, the ϕ_{SET} line and *FS* line run the entire length of the array. The sense-amplifier layout is symmetrical and balanced. This layout was generated with the layout-rule-independent physical design tool, and the symmetry was retained as layout rules were changed.

B. Row Decoder

The CMOS row decoder of Fig. 8 is a key block in the access path. It is very fast while also minimizing voltage overshoots and undershoots on the internal nodes of the decoder. Minimization of voltage overshoots and undershoots was a critical factor in the choice of circuits during the design of the 64K CMOS RAM. Conventional CMOS decoder circuits with series connected devices can have internal nodes that may be capacitively coupled well below ground or above the power supply voltage. With this stacked device type of circuit, adjustment of physical design and device sizes to damp the capacitive coupling may result in increased delay for decoder selection. In the decoder circuitry in this design, devices stacked more than two deep were not used. Also, stacking large numbers of devices was avoided elsewhere in the chip. As a consequence of this and other factors, voltage overshoots and undershoots, which could cause charge injection into the substrate and possibly trigger latch-up, were kept to under 0.25 V on all internal nodes of the chip.

The row decoder circuit has two stages of decoding. The first stage is a NOR decoder with the true or complement of the higher order address bits as inputs. The second stage is a two-input NAND decoder with the output of the NOR as
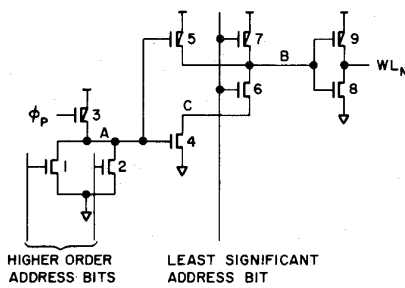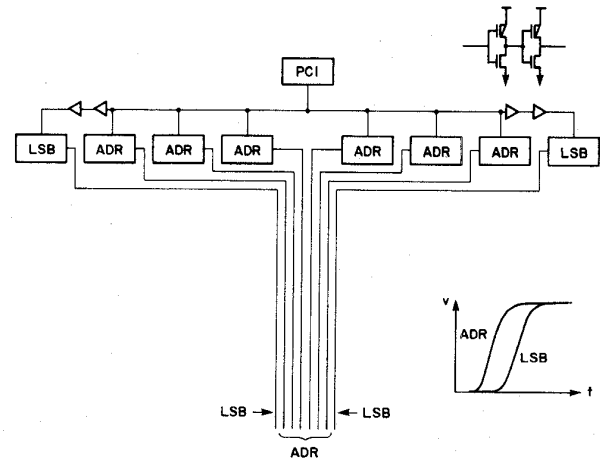
Fig. 9.   Simplified row decoder.

Fig. 10.   Delay of least significant address bit from higher order address bits.

Fig. 11.   Simplified input circuit and nonlinear voltage characteristic.

one of its inputs and either the true or complement of the least significant address bit (LSB) as the other input. In the 64K CMOS design, since the decoders are in the center of the chip, a single NOR decoder can drive four word lines.

The simplified row decoder circuit of Fig. 9 shows only a single word line to facilitate description of the circuit operation. In standby, all the address lines are low and the output (node A) of the NOR decoder is held to a high voltage through device 3. The word line is in the unselected low state, since the LSB is low, causing the NAND output (node B) to be high. The initiation of an access causes the precharge device 3 to be turned off and the address buffers to drive high either the true or complement of the address inputs to the decoder circuits. A word line is selected only if all the higher order address inputs to its NOR stage remain low and the LSB input to its NAND stage goes high. This results in the NOR decoder output (node A) staying high, the NAND output (node B) going low, and thus the word line going high. Following the selection of a word line and the setting of the sense amps, the selected NOR is discharged due to circuitry not shown on Fig. 8, thereby causing the selected word line to fall. The result is a well-controlled word-line pulsewidth, independent of the cycle time that occurs in an actual application. At the end of an access, all address lines are returned to a low state and the precharge device 3 is turned on. Consequently, the dynamic storage time on the NOR decoder node is small and well controlled.

The word line will remain in the unselected state during an access if the associated LSB remains low or if any of the higher order address inputs go high, causing node A to go low. Unselected word lines, and all word lines during standby, are actively held to ground. However, a momentary bounce on an unselected word line could occur if the NOR output (node A) did not discharge to a low level before the rising LSB turned on device 6 in the NAND stage. Any possibility of an unselected word-line bounce is eliminated by providing two stages of delay of the LSB rising to the higher order address bits rising, as shown in Fig. 10. Address line skew is contained by careful physical placement of the address buffers and address lines, by use of identical layouts for all address buffers, and through design of the address buffer circuit (see Section IV-C). Even with the very conservative bounce protection delay of 0.8 ns, the row decoder is still very fast, with a nominal delay from the higher order address bits rising to the word
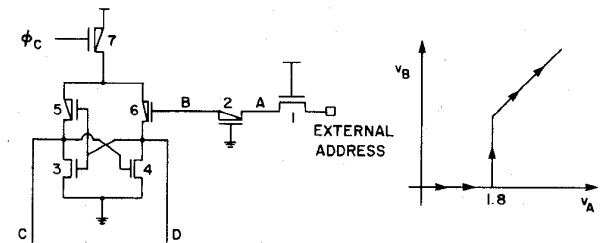
C. Input Circuit

The circuit for input of TTL addresses and data has high speed, low power dissipation, and safe operation. It is shown in a simplified version in Fig. 11, with the complete schematic shown in Fig. 12. Activated by the clock input falling, the circuit converts TTL levels to CMOS on-chip drive, latches the input state, and then disconnects the external input from the internal circuitry during an access. Following an access and the rise of the clock input, the circuit is designed to quickly precharge the internal nodes and the address lines for cycle time minimization. The power dissipation and delay skew as a function of TTL variations and device parameter variations is well contained by this circuit design, which also provides very high speed. The delay through the circuit from the rise of the clock input until the rise of the large capacitance address lines is only 1.9 ns. As will be described in this section, CMOS devices are key to the high-speed safe operation of this circuit—especially as used in the two distinctive portions in Fig. 11: the nonlinear front end and the self-referencing latch.

A salient feature of the high-speed input circuit is the nonlinear front end, which gives the voltage characteristic shown in Fig. 11. Because of the body-effected threshold voltage of p-channel device 2, a solid ground is provided at node B over the full range of low-input TTL signal levels. This can be seen in the voltage characteristic where the voltage at node B versus the voltage at node A is plotted.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.

fastcase®
Smarter legal research.