A 128Kb CMOS static random-access memory

by J. L. Chu
H. R. Torabi
F. J. Towler

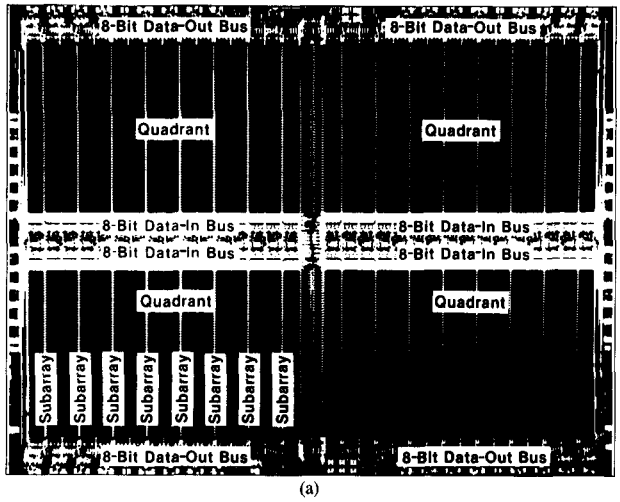
This paper describes an all-CMOS 128Kb static random-access memory (SRAM) with emitter-coupled-logic (ECL) I/O compatibility which was designed for the air-cooled Enterprise System/9000™ processors. Access time of 6.5 ns is achieved using 0.5-μm channel length and 1.0-μm minimum geometry. Pipelining and self-resetting circuit techniques permit the chip to operate with cycle time less than access time. To achieve the high-reliability requirement in the TCM environment, a novel technique utilizing a sacrificial substrate is used to "burn in" chips prior to their attachment to the TCM.

Introduction

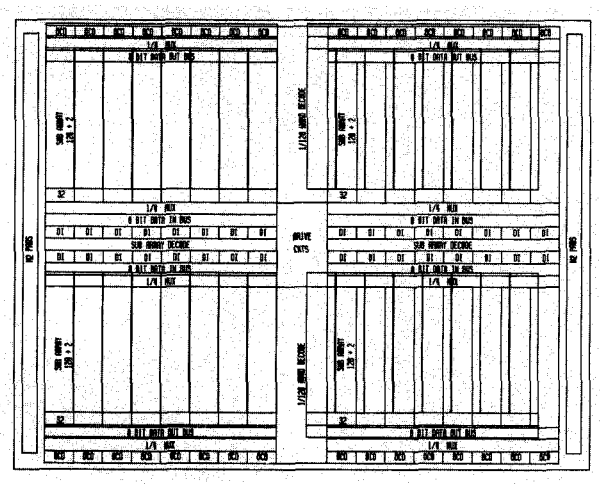
Widely used in both memory and logic chips, CMOS technology is generally recognized for its high circuit density and low power dissipation. In static RAM applications such as control store and level 1 or level 2 of the memory hierarchy, performance in the sub-10-ns range is required. These applications have traditionally been dominated by high-power bipolar circuit designs. Recent developments in technology and photolithography have made possible the fabrication of CMOS devices with channel lengths of 0.5 μm or less. With the use of these process advances and innovative circuit techniques, an all-CMOS static RAM chip with a typical access time of 6.5 ns and cycle time of 5 ns has been developed and manufactured. The fast cycle time, combined with a 32-bit-wide data interface, makes the chip capable of providing more than six billion bits/s [1]. The chip also has complete I/O compatibility with emitter-coupled logic (ECL) [2, 3], which allows it to communicate directly with bipolar devices.

The chip described in this paper was designed for the air-cooled mid-range of the new IBM Enterprise System/9000™ (ES/9000™) family of processors [4]. It has twice the circuit density of its bipolar counterpart, thus providing a much higher packing density at the thermal conduction module (TCM) level and reducing external communication among TCMs. In addition, the power dissipation is one-sixth that of the chip's bipolar equivalent, which allows the TCM to be air-cooled. Pipelined architecture using a self-resetting circuit technique allows the chip to function over a wide range of device parameters and operate with cycle time less than access time.

While circuit techniques and process technology permit the fabrication of high-performance and high-density chips, reliability must keep pace with the ever-increasing circuit density and system complexity. Historically, static "burn-in," in which the chip is subjected to operating power and temperature conditions over a period of time, has been used to verify reliability. The recent development of *in situ* burn-in—tests performed during burn-in to ensure that the chips are experiencing proper stress and to identify failures



(a)



(b)

Figure 1

128Kb CMOS static RAM: (a) photomicrograph; (b) physical outline drawing.

which may potentially be corrected—have decreased failure rates by a factor of 30 or more. For single-chip modules, the burn-in process usually involves the insertion of the fully packaged chip into a socket on a burn-in board. For TCM application, the process is not as straightforward, and for TCMs containing mixed-circuit technologies, the situation becomes even more complex. CMOS chips used in a multichip package along with bipolar chips cannot use the *in situ* burn-in process to achieve reliability objectives. For ES/9000 processors, a novel chip burn-in technique was developed which allowed the *in situ* burn-in of the CMOS chips before they were mounted on the TCM substrate.

Architecture and circuit operation

- *Chip layout and architecture*

As in any high-performance chip, circuit layout is a major part of the design. A well-organized circuit layout not only conserves valuable area on the chip surface, but also reduces signal noise coupling and enhances circuit performance.

Array

The chip is divided into four identical 32Kb quadrants; each quadrant comprises eight 4Kb subarrays, as shown in the photomicrograph of **Figure 1(a)** and the physical outline drawing of **Figure 1(b)**. The 4Kb subarrays each contain 128 word lines (vertical) and 32 bit-line pairs (horizontal). For any particular read/write cycle, one subarray per quadrant is selected to provide eight bits of information per quadrant. Three column addresses are used for the subarray selection. The word-line selection is accomplished by decoding seven word addresses for a "1 out of 128" decode. Two bit addresses are used to select eight out of the 32 bit-line pairs. Thus, a total of 32 bits (eight per subarray) are available to the data-out circuitry. If a $\times 8$ (i.e., a chip operating as a 16Kb $\times$ 8 memory) operation is desired, two additional section addresses are used to select two out of eight available data bits (per quadrant) before sending them to the off-chip drivers (OCDs). The data-in control for a write operation is similar. For write $\times 32$ (a chip operating as a 4Kb $\times$ 32 memory), eight bits per quadrant are written into each of the selected subarrays. The two section addresses are used to select two out of eight bits per quadrant in the case of write $\times 8$.

Word-line decoder and driver

The word-line decoders and drivers, located in the vertical-center section of the chip, decode the seven word addresses to select one out of the 128 word lines. Each word line runs across all subarrays on both sides of the chip. The combination of a selected word line and the output of the column decoder selects the cells within that subarray.

Data-in circuitry

The 32 data-in receivers, conveniently located at the horizontal-center portion of the chip for easy access to one end of the subarrays, provide input signals to the selected subarray during a write operation. To further conserve power, only eight data-in receivers are active during operation in the $\times 8$ mode. Eight write-select signals select the bit that is to be written into the memory array.

Sense amplifier and data-out circuits

During any read/write cycle, eight bits of information per quadrant are sensed and amplified by the sense amplifier.

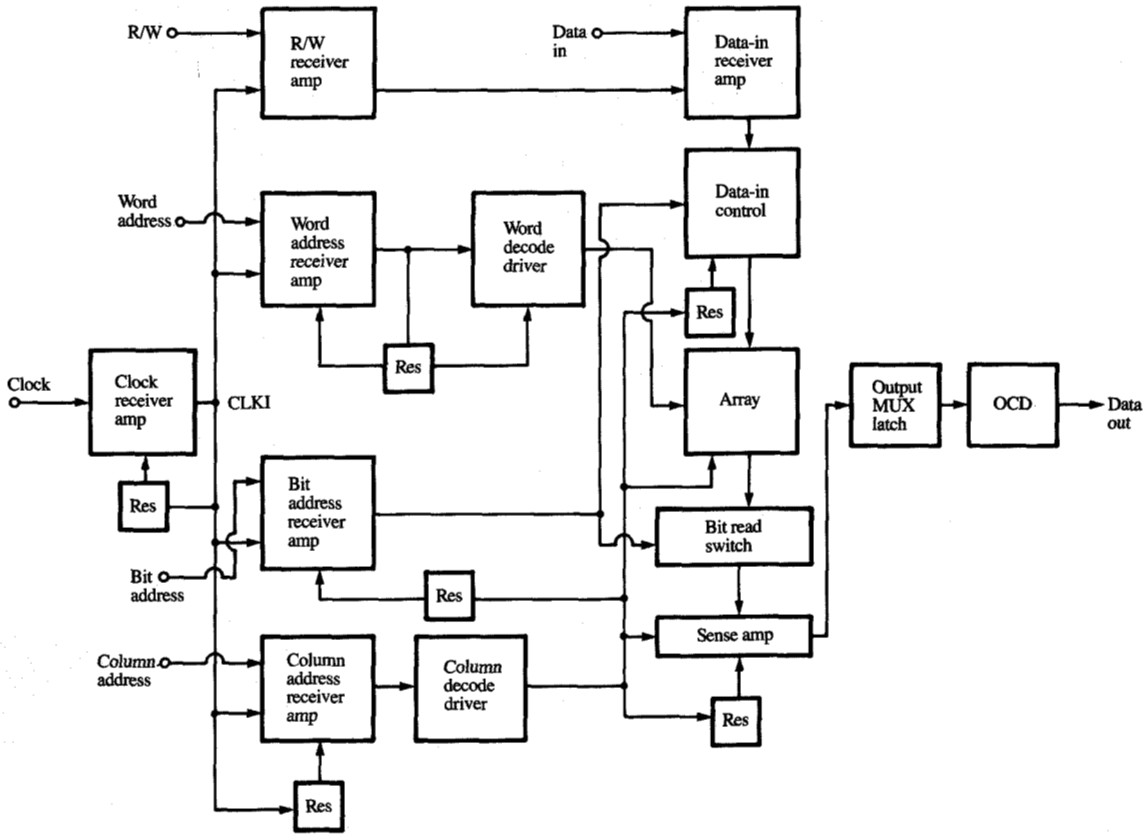


Figure 2

Simplified chip block diagram.

The output of the sense amplifiers (eight per subarray) in each quadrant is fed into an eight-bit data bus. Further manipulation of eight data bits, using two additional section addresses and dc controls in the data-out multiplexer (MUX) circuit, provides ×8 (two bits per quadrant) or ×8 with copy functions. All data bits (8 or 32) are finally fed to the off-chip drivers to provide line-driving capability. The sense amplifiers and the data-out circuitry are located along the outer edge of the chip.

- *Circuit operation*

Simplified block diagram

A simplified block diagram is illustrated in **Figure 2**. The downward transition of the clock input initializes all chip operations. Each functional block is independently timed with internal feedback and reset. As each block completes its function, its circuitry is reset and is ready to accept information from the next cycle, achieving operation with cycle time less than access time.

Circuit descriptions

The functional blocks of the chip are designed into a chain of self-resetting circuit macros, initiated by a single clock. The internal clock (CLKI) is generated from the ECL-level clock by using an asynchronous differential amplifier triggering a single-shot driver; the internal clock then triggers the amplifiers for all other inputs. A local reset is generated in each block, permitting cycle time to be reduced below access time. By using locally generated timing pulses, the need for a centrally located timing chain and the wiring associated with it are eliminated, reducing the loads on the reset paths. This simplifies the physical layout and improves cycle time.

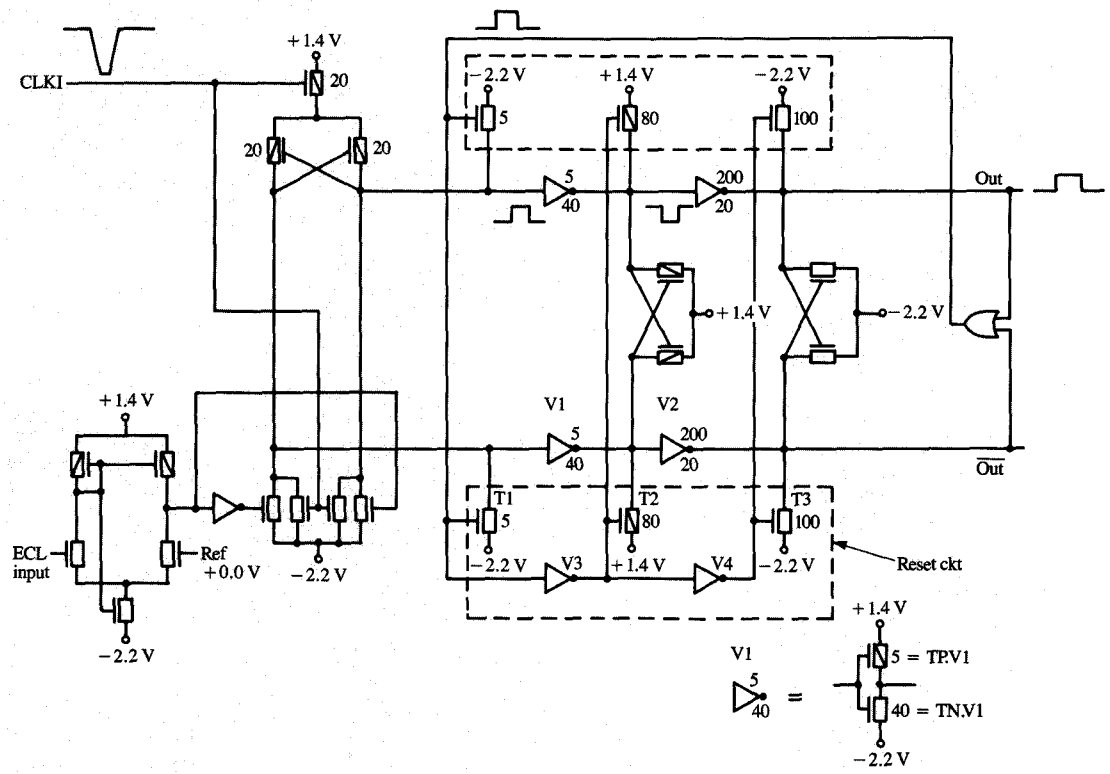


Figure 3

Address buffer (all sizes relative).

ECL-to-CMOS-level conversion is achieved by using a receiver with less than 100 mV sensitivity [5]. The receiver output is a pair of differential signals; the same receiver is used on all inputs to the chip.

A unique, clocked address buffer circuit (**Figure 3**) provides the fast cycle capability necessary for pipelined operation. The differential outputs of the ECL-to-CMOS convertor are applied to the inputs of the true and complement (T/C) self-resetting driver. An improvement in speed over conventional CMOS circuits is gained by separating the rising-edge and falling-edge performances. The critical path is primarily loaded by driver devices TN.V1 and TP.V2, resulting in one or the other output going high after the CLKI pulse falls. The NOR of the outputs initiates fast resetting of the critical path. The reset loop is designed to be longer than the width of the CLKI pulse to prevent a false pulse from the driver. Devices TP.V1 and TN.V2 are small devices that maintain the standby state. The input-triggered, self-resetting circuit technique is used throughout the chip.

Further chip-operation details are shown in **Figure 4**. CLKI triggers the address buffer and terminates the NOR decode precharge. The higher-order address bits discharge all deselected NORs, while the least significant address bit (LSB/LSBN) is delayed for safe decoder operation. The delay allows the redundant word decoders to resolve and block the LSB/LSBN from firing if a redundant word line is selected. The AND of the NOR output and the rising LSB (LSBN) at device T10 (T10N) results in the selected word line (WLODD or WLEVEN) being driven low. The LSB/LSBN rising triggers the reset of the address buffers and the precharge of the NOR decoder. As soon as the NOR decoder has been precharged, the word system can be triggered again to start another cycle.

Subarray operation is shown in **Figure 5**. The output of the column address NOR decoder selects a subarray and starts the subarray cycle. The bit-line precharge is controlled by WB0–WB3 signals. The selected bit line will have the precharge devices (internal to the Restore Control block) turned off, allowing either a read or write operation.

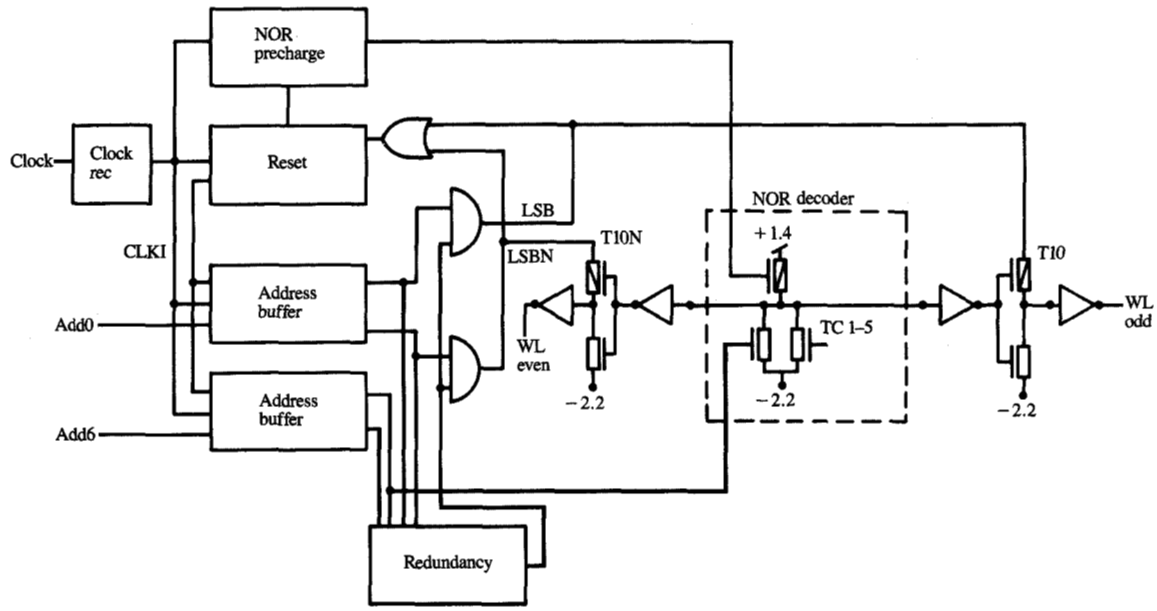


Figure 4

Word system block diagram.

On a write, one side of the local data-in bus pair (DB/DBN) is driven to −2.2 V, which pulls down the true or complement side of the bit line through devices Q1 or Q1N. For a read, the data-in bus floats and three of the read-bit decode lines (RB0–RB3) switch high, turning off the PFETs in three of four read-bit switches and disconnecting the three unselected bit lines from the sense amplifier. The selected cells discharge the true or complement side of the bit line at a nominal rate of 268 mV/ns. The half-selected cells with an active word line and unselected bit lines held at the 1.4-V level produce the worst condition for cell stability. Extensive stability analysis showed that the cell is stable under the worst-case application condition. A dummy word line drives a load designed to match the 64 word-line devices on a normal word line. The load is discharged, producing a delay which allows for a typical 237 mV of signal at the sense amplifier when the set signal (SSA) is applied. The signal at the sense amplifier is amplified and buffered to drive the 2-pF global-read data bus (GRDB0-3), which runs the width of a quadrant.

The eight global read data buses feed the multiplexer, which determines the data latches to be set. Three conditions are possible: the ×8 mode, the ×8 mode with four copies of the eight bits of data, and the ×32 mode with 32 unique data bits. These modes are controlled by two user-selectable dc chip inputs. In the ×32 mode, the signal ×32 is dc-active high and the negative-going data bus discharges node Z or ZN, setting the data latch. For the ×8 mode, the dc line labeled ×8 is active, along with a byte (BYTEA-D) decoded from two section addresses. The path through the MUX is now GRDB0-M-Z to the latch or GRDB0N-MN-ZN. The four copies of the data are provided by switching ×8×4 line active and allowing the signal on the M line to set three additional latches. An additional feature of the data latch is the output disable (ODIS) line, which forces all data latches to a "0" state; this is used in the control-store application.

The off-chip driver is attached to the output of the data latch. It can be switched to a tri-state mode by using an externally accessible control line. The driver achieves ECL output levels by the combination of the on-chip NFET or PFET and the terminating resistor on the receiving chip. The voltage division between the "on" FET and the terminating resistor limits the voltage levels to the required maximum ±700 mV.

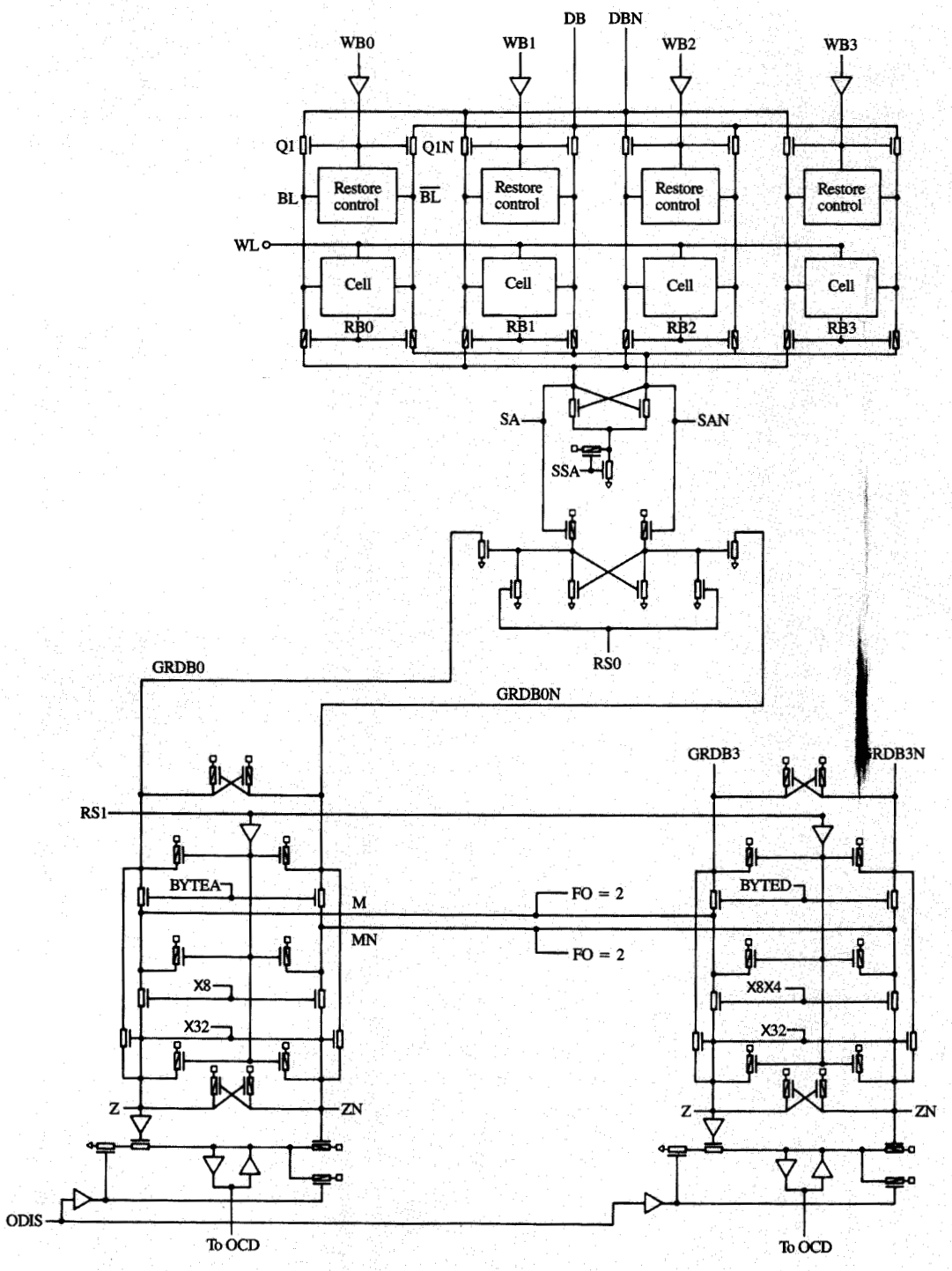


Figure 5

Simplified subarray circuit schematic.

Chip specifications and applications

• *Operating modes and chip specifications*
There are seven operating modes designed to accommodate unique system requirements, as shown below. Each operating mode is controlled by external control pins and can be changed by applying the appropriate bias to those pins.

- Read/Write ×8—The chip is organized as 16Kb × 8.
- Read/Write ×32—The chip is organized as 4Kb × 32.
- Read ×8 with copy—Same as Read ×8, except that each data-out bit goes to three additional OCDs to increase fan-out and drive capability.
- WSELECT control—This function is to gate the number of bits (from 1 to 8 in Write ×8 or group of 4 bits in Write ×32) written into the array.
- Combination of Read ×8, Write ×32 or Write ×32, Read ×8.
- Output select mode—This function forces all OCDs to the low state.
- High-Z output—This function forces all OCDs to a high-impedance state, allowing dot-OR capability.

The overall chip application specifications are shown in **Table 1**.

• *Applications*
The minimal power dissipation of the CMOS static RAMs allows all TCMs to be air-cooled instead of water-cooled. Identical power supplies and "footprint" (the geometrical array of C4 connections) are used for both the logic and CMOS array chips, significantly reducing the complexity of the TCM design and providing placement flexibility by avoiding unique sites for the memory chips. The TCM substrate is a multilayered ceramic (MLC) with 63 layers; it has a maximum capacity of 121 logic and memory chips [4]. The following five TCM modules are the basic components of the ES/9000 Models 190 through 480:

1. Processor module, where the instruction microcode is stored.
2. Buffer control module, where the CMOS array is used in the high-speed L1 memory hierarchy.
3. System control module, which controls the processor and subsystem.
4. Channel control module, which provides channel management function.
5. Vector module, for scientific and engineering computations.

Reliability

• *Chip burn-in considerations*
For high-density DRAMs and SRAMs, the best and most efficient way to improve reliability and decrease failure rates is the *in situ* burn-in of a single-chip module. Chips placed on ceramic substrates by using a solder-reflow chip-joining process are capped, tested, and burned in. However, for TCM applications, the chips cannot be singly packaged in modules, nor can they be burned in after controlled collapse chip connection (C4) joining to the TCM substrate. In the newly developed reduced radius removal (R3) process, the chips are attached to a temporary carrier (burn-in substrate) for burn-in and test and are then removed (mechanically sheared off) for the next assembly level. Chips are normally joined to ceramic substrates via a solder-reflow operation. To make a very strong bond, the contact area between the chip's C4s and the copper lines on the substrate must basically be the same size (**Figure 6**). Typically, if the chip's C4s are 6.0 mil in diameter, the maximum area of contact to the Cu conductor is also 6.0 mil. In the R3 process, the area of opening is limited to a fraction of the C4 diameter (**Figure 7**); for example, 2.0 mil versus 6.0 mil. Even with this restricted area, the strength of the bond between the chip and substrate proved to be more than adequate to support the burn-in and test operations. At the conclusion of the burn-in and test process, chips are mechanically sheared off; since the smallest area of contact between chip and substrate is at the C4-to-substrate join, C4s break very close to the substrate with no physical damage to the chip or substrate. A high-temperature hydrogen reflow process reshapes the C4 balls to their original form. Chips are then visually inspected for any physical damage prior to shipment. The burn-in substrates can be reused, and more than 20 repetitive cycles have been demonstrated.

Table 1 Chip specifications.

Density	128Kb SRAM
Organization	16Kb × 8
	4Kb × 32
Cell type	6-D static
Cell size	13.2 × 17.8 μm (234.96 μm^2)
Chip size	8.54 × 6.59 mm
Performance T_a/T_c	6.5/5 ns typical, 9/8 ns worst case
Power supplies	1.4 V, GND, −0.7 V, −2.2 V
Temperature	T_j = 65°C max., 25°C min.
I/O interface	ECL (±0.272-mV logic levels)
Power dissipation	2 W at 8-ns cycle

• *Observations and data*
To ensure that no additional failure mechanisms such as chip-join yield to the small contact areas, physical damage to the chip during the shear, or C4 volume loss were introduced, data were generated on more than 10 000 chips; the results are summarized as follows:

- *C4 fracture location* All C4s broke very close to the burn-in substrate.

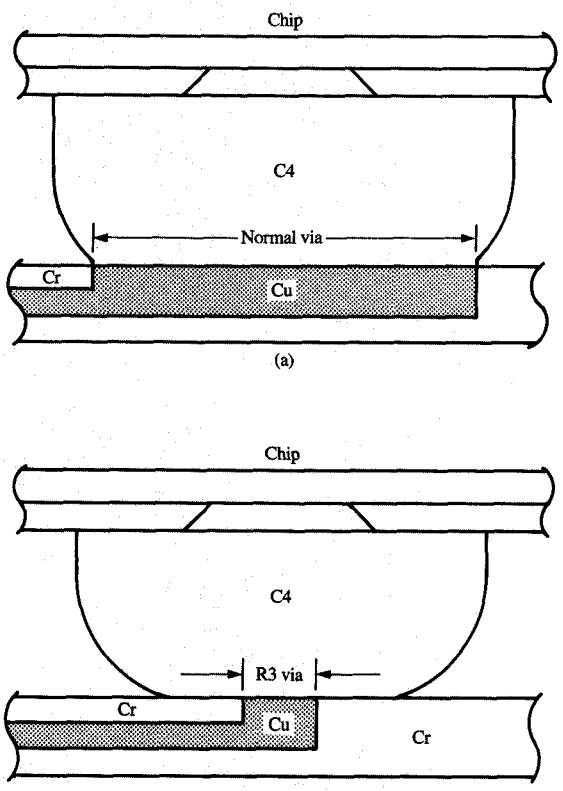


Figure 6

(a) Normal substrate; (b) burn-in substrate.

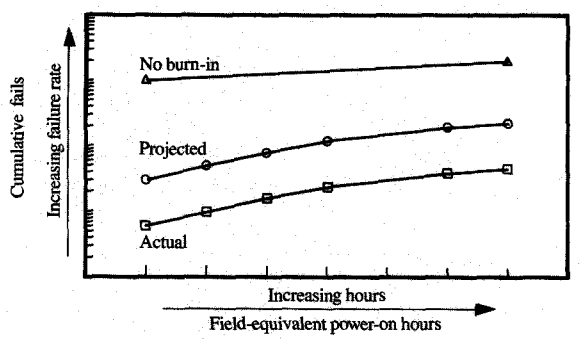


Figure 7

128Kb CMOS static RAM reliability performance vs. projection.

- *C4 deformation* No C4 deformation was seen on chips joined to the R3 substrates.
- *C4 % Sn content* Minimal Sn content change was observed within the specification.
- *C4 joints* Chip–substrate join yield was as expected.
- *Chip damage* No chip damage was seen due to the removal process.
- *Low-volume pads* C4 volume did not change to a measurable amount.
- *Substrate life* A sample substrate was used repeatedly for over 20 cycles.
- *C4 fatigue life* C4 life on chips treated with the R3 process was as good or better than on chips without the R3 process.
- *Functional feasibility* As long as the chips were properly joined, functionality remained excellent throughout the R3 process.

- *Summary*

This chip burn-in technique is a useful method of stressing/testing chips with C4s before they are assembled onto a multichip module. The feasibility of the R3 process was demonstrated by data taken from more than 10 000 chips in the field; the data also indicated that no new failure mechanisms had been introduced. Post-burn-in reliability during the early life of the chip also improved more than 30 times (Figure 7).

Process technology

A 1.2-μm CMOS technology [6] with 0.5-μm channel lengths was used to fabricate the chip. A double-ion-implanted lightly doped drain (DILDD) technique is used for fabricating the NFETs. Diffusions and polysilicon gates are salicided with a self-aligned Ti salicide to reduce resistance. Three Al–Cu metal wiring levels are used. The key process features are listed below:

- 1.2-μm average feature size.
- 0.5-μm channel length.
- N-well CMOS.
- DILDD NMOS.
- Ti salicide diffusions and polysilicon.
- 1-μm × 1-μm contacts.
- Three levels of Al–Cu metal.

Conclusions

A unique all-CMOS ECL-compatible SRAM has been described. Novel clocked circuitry providing high performance, combined with a flexible architecture, has allowed one chip to fulfill multiple applications in the mid-range processors of the ES/9000 family. ECL I/O compatibility is achieved by using an all-CMOS differential amplifier with less than 100 mV sensitivity. This allows the SRAM chips to interface directly with other bipolar chips

in the TCM, resulting in a significant improvement in system performance. Fast access has been achieved, along with a cycle-time less-than-access-time capability at modest power levels, which in turn allows the TCM to be air-cooled. Innovative techniques to allow for burn-in of chips before they are assembled into a multichip package have been developed and demonstrated.

Acknowledgments

The authors wish to thank individuals who helped to develop and manufacture the 128Kb SRAM described in this paper: for development (IBM Burlington), Mark Yungfleisch, Russ Houghton, Jack Romanoski, Ted Selfridge, and Ron Rossi; (IBM Thomas J. Watson Research Center) Terry Chappell and Barbara Chappell; for reliability and R3 process development (IBM Burlington), Rich Garcia, Bruno Aimi, Andy Ionta, and Andy Forcier.

References

1. F. Towler, J. Chu, R. Houghton, P. Lane, B. A. Chappell, T. I. Chappell, and S. E. Schuster, "A 128K 6.5 ns Access/5 ns Cycle CMOS ECL Static RAM," *ISSCC Digest of Technical Papers* **32,** 30–31 (1989).
2. S. Miyaoka, M. Odaka, K. Ogive, T. Ikeda, M. Suzuki, H. Higuchi, and M. Hirao, "A 7ns/350mW 64K ECL Compatible RAM," *ISSCC Digest of Technical Papers* **30,** 132–133 (1987).
3. T. Awaya, K. Todoya, O. Nomura, Y. Nakaya, K. Tanaka, and H. Sugawara, "A 5ns Access Time 64K ECL RAM," *ISSCC Digest of Technical Papers* **30,** 130–131 (1987).
4. V. L. Gani, M. C. Graf, R. F. Rizzolo, and W. F. Washburn, "IBM Enterprise System/9000 Type 9121 Model 320 Air-Cooled Processor Technology," *IBM J. Res. Develop.* **35,** No. 3, 342–351 (1991, this issue).
5. B. A. Chappell, T. I. Chappell, S. E. Schuster, H. M. Segmuller, J. W. Allan, R. L. Franch, and P. J. Restle, "Fast CMOS ECL Receiver with 100-mV Worst Case Sensitivity," *IEEE J. Solid-State Circuits* **23,** No. 1, 59 (1988).
6. D. T. Wong, R. D. Adams, A. Bhattacharyya, J. Covino, J. A. Gabric, and G. M. Lattimore, "An 11-ns 8K×18 CMOS Static Ram," *IEEE J. Solid-State Circuits* **23,** No. 5, 1095–1103 (1988).

Received January 21, 1991

Jeff L. Chu *IBM General Technology Division, Burlington facility, Essex Junction, Vermont 05452.* Mr. Chu received a B.S. degree from Bucknell University in 1962 and an M.S. degree from the University of California at Berkeley in 1964, both in electrical engineering. He joined IBM in East Fishkill in 1968, transferring to the General Technology Division in Burlington in 1971. Mr. Chu has worked in various design and management positions in the development of DRAM, CCD, and SRAM chips, and is currently an Advisory Engineer in a high-performance SRAM development area.

Hamid R. Torabi *IBM General Technology Division, Burlington facility, Essex Junction, Vermont 05452.* Mr. Torabi received a B.S. degree from Norwich University in 1981 and an M.S. from the University of Vermont in 1985, both in electrical engineering. He has been with IBM since 1983 as a reliability engineer and is at present an Advisory Engineer in a reliability engineering area. Mr. Torabi has received informal awards and an IBM Outstanding Technical Achievement Award in 1990 for work on the 128Kb CMOS chip; he has also filed for a patent and published two invention disclosures.

Fred J. Towler *IBM General Technology Division, Burlington facility, Essex Junction, Vermont 05452.* Mr. Towler received his B.S.E.E degree from Northeastern University in 1977, joining IBM that same year. He is an Advisory Engineer currently working in a high-performance CMOS development department. Mr. Towler's previous work experience with IBM includes work in bipolar and MOS development.