A One Transistor RAM for MPC Projects

by

James J. Cherry and Gerald L. Roylance

Massachusetts Institute of Technology

545 Technology Square, Cambridge, MA

Abstract: Many MPC projects, such as video frame buffers, need a large memory subsystem. A one transistor per bit dynamic memory using Mead-Conway design rules is being designed with this purpose in mind. The memory cell size is 16.5λ by 8λ (about the same size as a 1975 4K RAM cell with λ = 2.5 microns).

While a complete high density memory subsystem has not been designed, two chips have been designed to test its major components. One chip is a 1K memory array that tests the sense amplifier, column decoder/driver, and read/write logic. This chip lacks a timing generator and clock drivers. The second chip tests some low power bootstrapped clock drivers. These test chips are currently being fabricated.

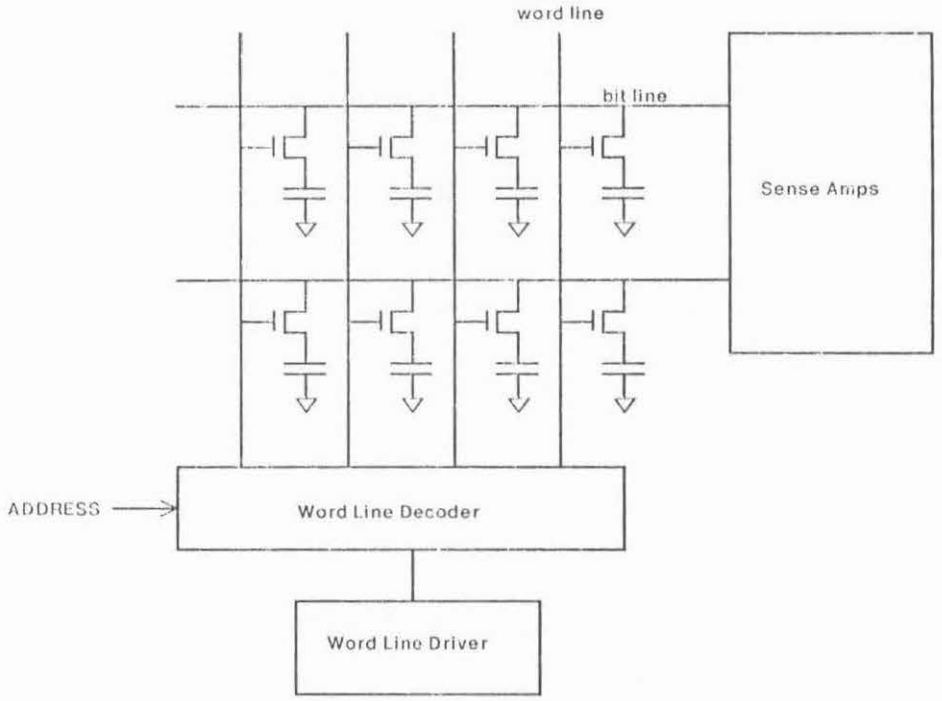
This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Laboratory's V. L. S. I. research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract number N00014-80-C-0622 and in part by the Advanced Research Projects Agency under Office of Naval Research contract N00014-75-C-0643.

1. Design Considerations for a One Transistor RAM

Many VLSI projects need a moderately large memory module. A one transistor dynamic memory has been designed as a subsystem usable in the Multi-Project Chip (MPC) designs being undertaken by several universities. Because the memory is intended for MPC projects, it follows the conservative Mead-Conway design rules and uses only a single layer of polysilicon; consequently, the achievable memory density suffers. Furthermore, the memory design must tolerate wide process variations because many different fabrication lines are used for MPC.

Throughout the RAM design we have decided in favor of simplicity in order to give the RAM the best chance of working. When the choice was between speed and density, we chose density because we feel density is more important to the average project.

A block diagram of the memory is shown in the first figure. The major components are the memory array, sense amplifiers, column address decoder, and word line (column) driver. The bits of the memory are stored as a voltage (0 or 2 volts in our case) on the capacitors in the array. All of the memory cells in a column are read at once. Addressing is done by the word line decoder; the decoder takes a positive pulse from the word line driver and steers that pulse to the column selected by the address lines. That word line pulse turns on all of the pass transistors in the selected column, thus connecting the memory cell capacitor to the horizontal bit lines and to the sense amplifiers where the binary value held in the cell is determined. The sense amplifier must be sensitive to signals on the order of a hundred millivolts because the memory cell capacitance and the (much larger) stray bit line capacitance form a voltage divider. In the present design this attenuation is a factor of 15.



There are several references on one transistor RAM design. Barnes [1] gives a short description of several sense amplifier designs. We used the charge transfer sense amplifier first reported by Heller [5, 6]. We did not use the more

sophisticated version of this amplifier described in Heller [7] because we felt there would be a better chance of making the simpler version work. The charge transfer sense amplifiers have more sensitivity than amplifiers that directly sense the voltage difference on the bit lines and are also tolerant of bit line capacitance and transistor threshold voltage variations.

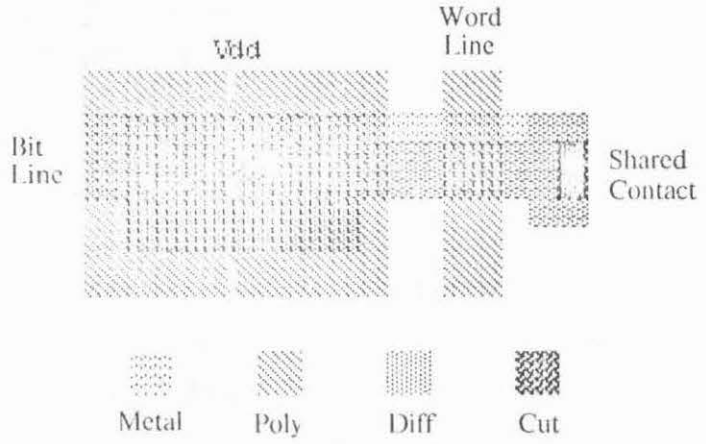
The input-output (I/O) circuits [not shown] follow those of Gray [4]. The I/O is done on one side of the array (rather than in the center of the the array) so many bits (for example, 32 bits) can be read or written at once, allowing higher memory bandwidth than that available from commercial parts that are limited to reading from 1 to 8 bits at a time.

The high voltage (7.5 volt) bootstrap driver used to precharge the bit lines is based on one described by Chan [2] The word line decoder is derived from one given by Tzou in an article on CCD memories [11].

2. Memory Cell Design Considerations

Several different memory cell layouts were tried in the search for highest possible array density. The densest one we found (shown below) uses metal bit lines and polysilicon word lines. The memory capacitor is actually an enhancement mode transistor whose source and drain are tied together to make one terminal and whose gate forms the other terminal; this latter terminal is connected to V_{DD}.

For a bit line capacitance to storage capacitance ratio of 15 to 1 (with 64 cells on each bit line), the cell size is 16.5 λ by 8 λ. This cell is half the size (comparing square λ's) of a three transistor RAM cell designed by Dick Lyon at Xerox PARC. With $\lambda = 2.5\ \mu$, this cell is the same size as the INTEL 2104, a commercial dynamic RAM that came out in 1975.



Memory Cell Layout

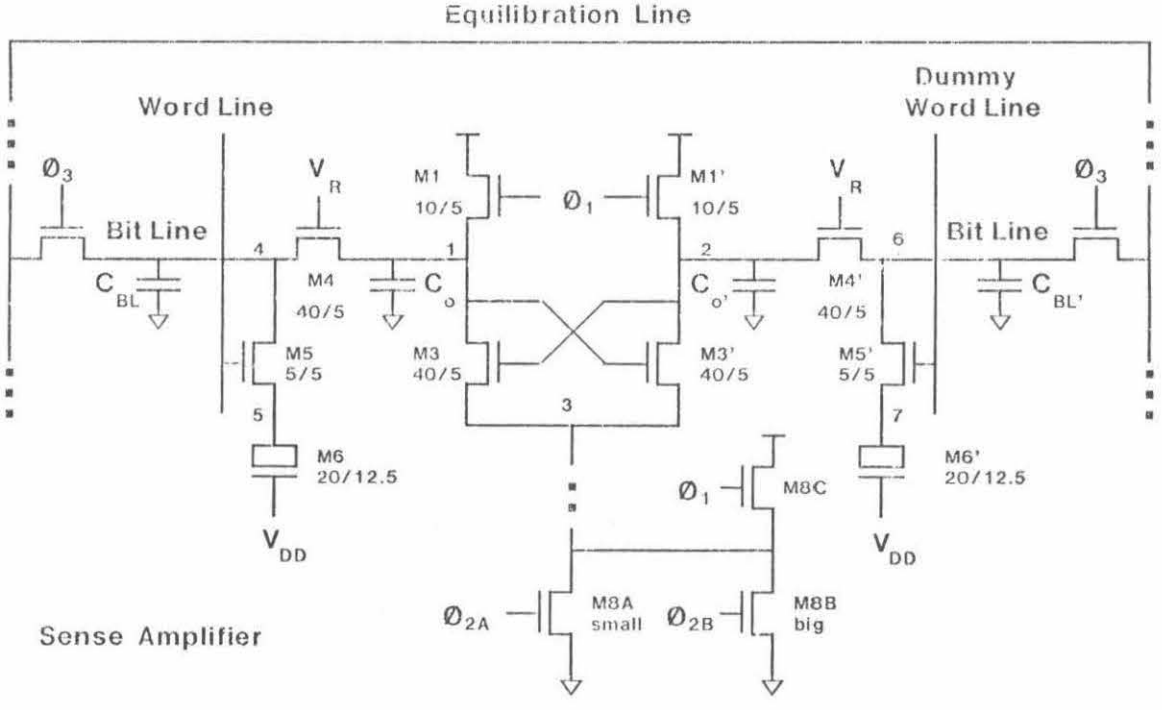
Cell size is not the only size consideration. The sense amplifiers and the word line decoders must also fit in the pitch determined by the cell size. While there was little problem with the decoder pitch, the tightest pitch we could give the sense amplifier was 11 λ (when we needed 8 λ). Although a bit line pitch of 11 λ could be used, it would cause a significant increase in cell area. In a two level polysilicon process (which we do not have), the area penalty can be avoided by using a folded bit line [10] or a staggered bit line [8].

We were able to use the minimum area cell (with its 8 λ pitch) by placing two amplifiers side by side and fitting the pair to a 16 λ pitch. This layout is not symmetrical and some effort is needed to balance the stray coupling from control lines onto the bit lines or the memory will not work. The noise injected onto the bit lines by these control lines is one-half of the signal that the amplifier is trying to sense!

3. Sense Amplifier

The heart of every sense amplifier design is a cross-coupled differential pair (M3 - M3' in the figure below). A small initial voltage difference on nodes 1 and 2 is amplified when node 3 is pulled down by M8. The main difference between the many different sense amplifiers used in dynamic RAMs is how the cross coupled latch is connected to the bit lines [1]. In order to sense a small differential voltage quickly, the large capacitance of the bit lines must be isolated from the internal nodes of the differential amplifier.

The next figure shows the basic *charge transfer sense amplifier* that we use and that was first described by Heller [5, 6]. (Much of the peripheral circuitry that we use comes from Gray [4]).



The memory cell capacitors are FET's M6 and M6'. φ_1 is a high voltage pulse (above V_{DD}) that precharges nodes 1, 2, and 3 to V_{DD}. V_R is a supply voltage less than V_{DD}, so nodes 4 and 6 charge to $V_R - V_T$ (V_T is the transistor threshold voltage) as M4 and M4' reach cut-off. Since $V_R - V_T$ is the highest the bit lines can

go, this voltage is used to store a logical *one* in the memory cells. A logical *zero* is represented by storing zero volts in the memory cell capacitor. A voltage half way between these two limits is stored on a dummy cell (M6') to provide a reference voltage for the sense amplifier. Each bit line is populated with 64 memory cells and one dummy cell. When reading a column of bits from one side of the array, the dummy cell on the opposite side of the array is addressed. Thus, the figure above depicts the the situation encountered when reading a bit out of the left half of the memory array.

When a read sequence starts, the precharge line (φ_1) is turned off and the word line and dummy word lines are brought high to connect a memory cell (M6) and the dummy cell (M6') to opposite sides of the differential sense amplifier. Consider the case in which a zero is stored in M6. The **drop** in voltage on the bit line will be:

$$\Delta V_4 = (V_R - V_T)\, C_S / (C_S + C_{BL})$$

where C_S is the capacitance of the memory cell, and C_{BL} is the bit line capacitance. This takes M4 out of cut-off and into saturation with a gate drive of ΔV_4. M4 will remain on until the bit line is charged to back to $V_R - V_T$. The charge necessary to charge $C_S + C_{BL}$ back to $V_R - V_T$ must come from C_o, the parasitic capacitance on node 1. Thus,

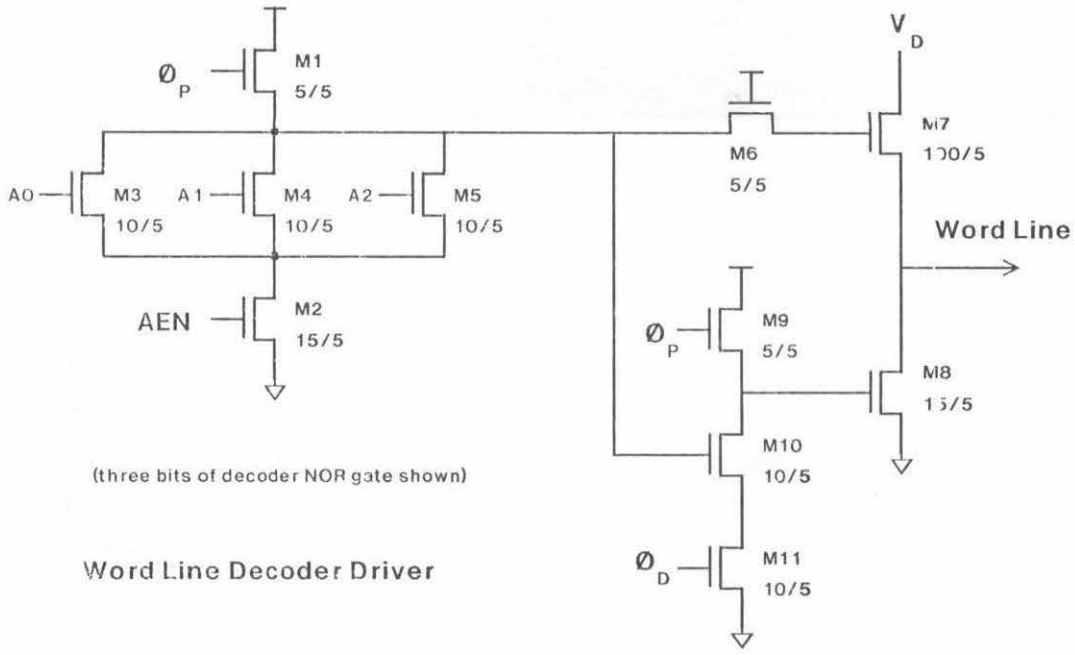
$$\Delta V_2 = (V_R - V_T)\, C_S / C_o$$

The voltage change at node 3 will be one half of this. Note that the differential voltage $\Delta V_2 - \Delta V_3$ is independent of the bit line capacitance if enough time has elaspsed [5, 6]. φ_{2A} comes on and starts pulling node 3 toward ground, eventually turning one of M3 or M3' on. Node 3 is pulled down slowly enough to prevent the other latch transistor from turning on [9]. As the V_{GS} of the on transistor increases, so does it's g_m. φ_{2B} is then turned on to quickly pull node 2 to ground. The bit that was stored on M6 has now been refreshed, so the word line is turned off. φ_{2A} and φ_{2B} are turned off. φ_3 is turned on to short **all** of the bit lines together; since half of the bit lines are at the logic zero level (0 volts) and the other half are at logic one ($V_r - V_t$), the bit lines go to a voltage half way between the two levels. This voltage is stored in the dummy cell by dropping the dummy word line and φ_3.

A memory write cycle proceeds exactly like that of a read, but instead of letting the contents of the memory cell determine the fate of the bit line, a circuit at the end of the bit line either pulls it to ground or pushes it up with a capacitor approximately the same size as a memory cell [4].

4. Word Line Decoder

It is impractical to use a separate high capacitance driver for each of the 128 word lines, so a single driver must be shared by many word lines. The word line decoder does both the memory addressing and the multiplexing of the high capacitance driver. The word line decoder also clamps the unselected word lines to ground to minimize some problems related to subthreshold conduction of the FET's that isolate the unselected memory capacitors to the bit lines [2, 4]. The word line decoder circuit show belowis a modified version of the one described by Tzou [11] . The basic idea is to bootstrap pass transistor M7 with its own gate-source capacitance so that all of the voltage developed by the word line driver is delivered onto the word line.



Operation of the decoder driver starts with φ_P precharging the gates of M8 (the clamp transistor) and M7 (the pass transistor) while V_D (word line driver), φ_D, and AEN are all low. When the address lines (A0-A6) have settled, φ_P goes low followed by AEN going high. If all of the address lines inputs of the NOR gate decoder are low, then the word line is selected and the pass transistor will allow the word line pulse to pass through. In that case, when φ_D comes on it will discharge the gate of M8, allowing the output to rise when V_D comes along. As V_D rises, isolation transistor M6 turns off allowing M7 to bootstrap.

If one of the address lines in the NOR decoder is high, then the word line is not selected. When AEN goes high, the gate of M7 is discharged, turning it off and isolating the word line from the driver. φ_D is prevented from discharging the gate of M8, the clamp transistor, by M10.

Tzou's design does not include isolation transistor M6. Without this transistor, much of the bootstrap charge on M7 is lost in charging the diffusion capacitance of the decoder logic. Another addition to Tzou's design is M2, which provides a simple means of disabling the address lines.

The capacitance associated with the drain of M7 is a surprisingly large load to the word line driver. While the word line capacitance for a 16Kbit version of the RAM for a typical process would be only 2pF, the combined drain capacitances of 128 decoders is several times higher (about 8pF). Our first layout of M7 used a straight gate and had a capacitance from the drain to ground of .1pF. Bending the gate into a rectangle (a suggestion of Pat Bosshart) eliminated the sidewall capacitance and reduced the capacitance to .068pF per drain. Another layout, suggested by Tom Knight but not used by us, has an effective capacitance of .056pF.

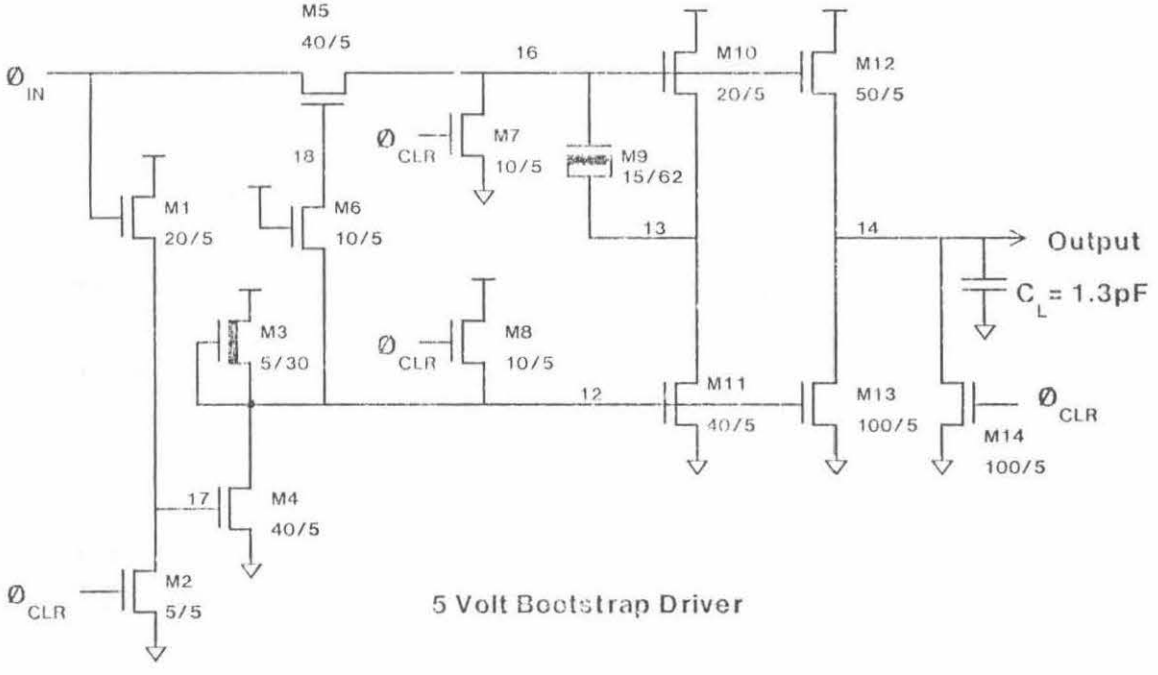
5. Bootstrap Drivers

In a typical application the memory array is large and consequently the clock signals are large capacitive loads which require driver circuits. We designed two bootstrap drivers: one is a 5 volt driver for the address inputs of the word line decoder and φ_3 of the sense amplifier and the otheris a 7.5 volt driver for precharging the bit lines (φ_1). A third driver which is for the word lines (V_D)has not been designed. The 5 volt circuit takes 5nS to switch a 1.3pF load (128 5μ x 5μ gates); the 7.5 volt circuit takes 15nS to drive the same load.

Neither of the driver circuits is connected to the memory array so that the drivers and the memory can be tested independently. Super buffers are used for the address buffers in the current memory array, but all other clocks are simply bonded out to pads. Normal depletion load inverters and super buffers are not acceptable for the word line driver because their logic zero output is not 0.0 volts. Super buffers consume more standby power than a dynamic bootstrap driver and so super buffers are less desireable in a large memory.

The 5 volt bootstrap driver is modeled after one used in the INTEL 2118 16K dynamic RAM (see figure). φ_{CLR} precharges the gates of M11 and M13 high, turning them on. When φ_{IN} starts to rise, it charges capacitor M9 and starts to turn M10 and M12 on. M6 isolates node 18, allowing that node to bootstrap and keep M5 turned on hard. M1 and M4 form a comparator that notices when φ_{IN} has gone above 2 threshold drops. When this happens, M4 turns on and pulls nodes 12 and

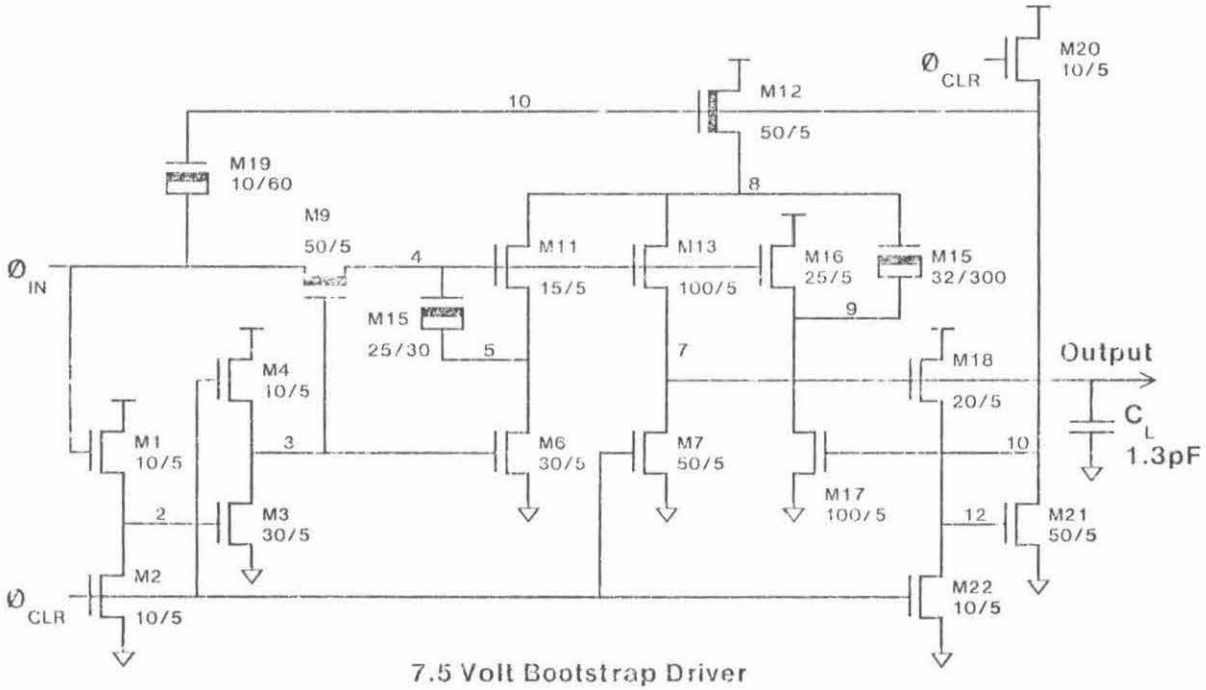
18 down to ground. M11 and M13, which had been holding down nodes 13 and 14, now turn off, letting those nodes rise. Capacitor M9 bootstraps node 16 (which was isolated by M5 when M5's gate fell), turning M10 and M12 on hard. M12 pulls the output node voltage up. φ_{IN} can now fall without affecting the rest of the circuit because M5 is off. φ_{CLR} turns M11 and M13 on and turns off M10 and M12, forcing the output low and resetting the circuit. The bootstrap capacitor M9 is driven from node 13 and not from node 14 to get more gate drive on M12 which significantly improves the output rise time.



5 Volt Bootstrap Driver

The high voltage bootstrap driver (after Chan [2]) is basically two of the 5 volt drivers that have been merged together. M15 bootstraps in much the same way as M9 did in the previous circuit. When M15 is bootstrapping, node 10 (which was pushed high by φ_{IN}) connects node 8 to V_{DD}. As node 4 rises, the output (node 7) also rises. M18 and M21 form another comparator that notices when the output has exceeded two V_t; then M21 pulls down node 10, turning off M17 which had been holding one terminal of capacitor M15 down at ground and also turning off M12 which had been holding the other terminal at V_{DD}. M16 has been turned on because node 4 has been bootstrapped high earlier. M15 now bootstraps node 8.

M11 is still on, so node 5 follows node 8 which pushes node 4 still higher. Node 4 makes M13 and M16 stay turned on; thus, the output follows node 8. The clever feature of this circuits is that M15 does not charge share with the load capacitance until the load capacitance has been charged up to two V_t (ie, until the comparator trips).



7.5 Volt Bootstrap Driver

6. Conclusion

We designed two project chips to test our designs. One chip is a 128 by 8 (1K) array of memory cells with sense amplifiers, word line decoders (implemented with super buffer address drivers), and multiplexed read/write logic. This array size provides a reasonable feasibility test for building a 16K subsystem. No clock or timing generation is included on the chip because we did not have enough time. We expect access times of 150ns and cycle times of 250ns. A second, separate project chip test the 5 volt and the 7.5 volt bootstrap drivers. Additional circuitry is included on that chip for measuring the capacitive loading on the drivers when a

low capacitance probe is connected to their outputs.

The original goal of this effort was to develop a high density memory subsystem that could be treated as a "black box" by designers with little or no analog background. We severely underestimated the magnitude of such a task. As it stands, our results can only be viewed as a first cut towards that goal. We found that many design challenges lie in the peripheral circuitry such as drivers and decoders: *there is much more to a one transistor RAM than sense amplifiers.*

We thank Mark Johnson for telling us about laying out high frequency gate oxide capacitors and for spotting the undesirable control line coupling in the sense amplifier. We thank Tom Knight for general guidance and moral support. The RAM was designed as a term project for an MOS analog circuit design course taught by Prof. Yannis Tsividis of Columbia University while visiting MIT. Prof. Tsividis has given both of us a better understanding of using MOSFETs in both analog and digital design.

7. Bibliography

1. J. Barnes, "A High Performance Sense Amplifier for a 5V Dynamic RAM", **IEEE Journal of Solid-State Circuits**, Vol SC-15, No. 5, October 1980, pp 831-839.

2. J. Y. Chan, et al, "A 100nS 5V Only 64Kx1 MOS Dynamic RAM", **IEEE Journal of Solid-State Circuits**, Vol SC-15, No. 5, October 1980, pp 839-846.

3. Electronic Design, "Circuit Techniques Tune Up for Production of 64-K RAMs", **Electronic Design**, October 25, 1980, pp 31-32.

4. K. Gray, "Cross-Coupled Charge-Transfer Sense Amplifier and Latch Sense Scheme for High-Density FET Memories", **IBM Journal of Research and Development**, Vol 24, No. 3, May 1980, pp 283-290.

5. L. G. Heller, D. P. Spampinato, Y. L. Yao, "High-Sensitivity Charge-Transfer Sense Amplifier", **1975 IEEE International Solid-State Circuits Conference**, pp 112-113.

6. L. G. Heller, D. P. Spampinato, Y. L. Yao, "High Sensitivity Charge-Transfer Sense Amplifier", **IEEE Journal of Solid-State Circuits**, Vol. SC-11, No. 5, October 1975, pp 596-601.

7. L. G. Heller, "Cross-Coupled Charge-Transfer Sense Amplifier", **1979 IEEE International Solid-State Circuits Conference**, Feb 1979, pp 20-21.

8. T. C. Lo, R. E. Scheuerlein, R. Tamlyn, "A 64K Dynamic Random Access Memory: Design considerations and Description", **IBM Journal of Research and Development**, Vol. 24, No. 3, May 1980, pp 318-327.

9. W. T. Lynch, H. J. Boll, "Optimization of the Latching Pulse for Dynamic Flip-Flop Sensors", **IEEE Journal of Solid-State Circuits**, Vol SC-9, No. 2, April 1974, pp 49-55.

10. F. J. Smith, R. T. Yu, I. Lee, S. S. Wong, M. P. Embrathiry, "A 64 kbit MOS Dynamic RAM with Novel Memory Capacitor", **IEEE Journal of Solid-State Circuits**, Vol. SC-15, No. 2, April 1980, pp 184-189.

11. A. Tzou, et al, "A 256K-Bit Charge-Coupled Device Memory", **IBM Journal of Research and Development**, Vol 24, No. 3, May 1980, pp 328-338.

12. Y. S. Yee, L. M. Terman, L. G. Heller, "A 1 mV MOS Comparator", **IEEE Journal of Solid-State Circuits**, Vol. SC-13, No. 3, June 1978, pp 294-297.