DESIGN OF AN EMBEDDED FULLY-DEPLETED SOI SRAM †

Raymond J. Sung, John C. Koob, Tyler L. Brandon, Duncan G. Elliott, Bruce F. Cockburn

Department of Electrical and Computer Engineering
University of Alberta
Edmonton, AB T6G 2G7, Canada
{sung|jkoob|brandon|elliott|cockburn}@ee.ualberta.ca

Abstract

We describe the design of an embedded 128-Kb Silicon-On-Insulator (SOI) CMOS SRAM, which is integrated alongside an array of pitch-matched processing elements to provide massively-parallel data processing within one integrated circuit. An experimental 0.25-μm fully-depleted SOI process was used. The design and layout of the SOI memory core and results from calibrated circuit simulations are presented. The impact of the floating body effect is investigated for both memory reads and writes. We describe threshold mismatch effects in the sense amplifier that result from the floating body voltage. Floating body effects are compared against simulated results for an SRAM designed in a 0.25-μm partially-depleted SOI process.

1. INTRODUCTION

Silicon-On-Insulator (SOI) CMOS integrated circuits are known to have significant potential performance advantages over their bulk silicon complementary metal-oxide semiconductor (CMOS) counterparts [1]. In silicon-based SOI technology, the field effect transistors (FETs) are manufactured as isolated islands on top of an insulating layer of silicon dioxide. Thus the FETs are electrically isolated from each other and from the underlying bulk silicon. Many advantages stem from the SOI structure. Layout is simplified because there is no need for wells, well contacts, or isolation trenches between devices. Latch-up is avoided because the source and drain regions are surrounded by insulator. The lack of diode junctions around the source and drain regions results in reduced leakage currents and junction capacitances, and hence permits faster switching speed and lower power consumption.

In the case of memories, the reduced source/drain capacitances afforded by SOI are especially attractive because the bitlines are connected to typically hundreds of drains [2]. Reducing the junction capacitance thus directly reduces a major component of the bitline capacitance, which is a critical parameter limiting memory performance. For example, in the study presented in [2], the junction capacitance in the bulk silicon SRAM model was estimated to contribute 42.1% of the total bitline capacitance at a temperature of 25 C. Moving to an SOI design reduced the junction capacitance by 74.5%, which in turn reduced the bitline capacitance by 31.4%.

There are two major types of SOI processes: fully-depleted and partially-depleted [1]. In a fully-depleted SOI (FD-SOI) process, an ultrathin silicon film (typically less than 50 nm) is present on top of a thick insulator layer. Active components are formed by appropriate n- and p-type doping, and are then completely isolated from each other by etching away the intervening silicon. Also, the silicon film is thin enough that the depletion layer extends through the entire thickness of the film. A floating body is a body region under the gate that has not been provided with a contact that connects it to a suitable potential. SOI circuits with floating bodies are subject to so-called floating body effects, such as kinks in the I-V characteristic caused by charge collection in the body and associated reductions in the threshold voltage. In a partially-depleted SOI (PD-SOI) process, the top silicon layer is thicker, which simplifies the processing but introduces some unwanted circuit behaviors, such as increased history dependence and parasitic bipolar currents. However, PD-SOI has the advantage of lower leakage currents and higher transconductance [3].

The FD-SOI process used for this design is a 0.25-μm fully-depleted SOI CMOS process made available to our project by MIT Lincoln Labs [4, 5]. Features of this process include mesa-etched isolation, a 170-nm thick buried oxide (BOX), and a 47-nm thick silicon film. As it is a research technology, some aspects of the process, such as the pitch of all three aluminum metal layers (1.2 μm), are not optimized for high-density logic. Nevertheless, although published results are available for partially-depleted SOI SRAMs [2], this paper may be the first to describe the design of a fully-depleted SOI SRAM.

Our SRAM is to be used in a logic-in-memory single instruction stream, multiple data stream (SIMD) architec-

†This work was supported by Micronet R & D, by MOSAID Technologies Inc., the Natural Sciences and Engineering Research Council of Canada (NSERC), the Alberta Informatics Circle of Research Excellence (iCORE), and the Canadian Microelectronics Corporation (CMC). Access to FD-SOI fabrication is being provided by MIT Lincoln Labs.

ture, where each processor in a linear array is pitch-matched to one or more adjoining memory columns. In our implementation, each processor element (PE) is pitch-matched to two columns of memory. Background on the resulting Computational RAM (C-RAM) co-processor is available in [6]. The C-RAM architecture can provide several orders of magnitude speed-up on problems that map efficiently to the SIMD model, such as many multimedia and data compression operations.

This paper focuses on the design of an SRAM for floating body effects in a fully-depleted SOI process. The rest of the paper is organized as follows: The next section discusses the architecture of the SRAM and gives layout details of the core. The third section presents design considerations related to phenomena specific to FD-SOI designs, such as floating body effects in the memory core and sense amplifiers. The fourth section contains simulation results for our FD-SOI SRAM. Finally, we conclude with a summary of our design results.

2. SRAM ARCHITECTURE AND LAYOUT

The proposed test chip includes an SRAM array with 512 rows and 256 columns, a size constrained by available test die area. Off-chip read and write accesses occur via a conventional column decoder and a databus running parallel to the wordlines. A second memory access method is for the pitch-matched PEs to access, in parallel, local memory locations for reads or writes. Each PE sees one column of memory as its local memory, with all processors using the same row address offset. To accommodate the PEs being twice the width of a memory column, the PEs are staggered at alternate ends. To evaluate the FD-SOI implementation of embedded SRAM, the array was designed and simulated with the PEs omitted. A block diagram of the memory column schematic is shown in Figure 1.

As is frequently done in SOI designs, area in the memory cell is saved through the use of abutted NMOS-PMOS drains [7]. The transistors forming the inverters in the 6-T SRAM cell, shown as M6, M7, M10 and M11 in Figure 2, are connected by simply abutting them.[1] The salicide on the drain regions provides a low resistance path rather than a diode junction. The ability to abut n- and p-type diffusion is not available in typical bulk CMOS logic processes without local interconnect. While the metal pitch and contact sizes in this process are not optimized for logic or memory, memory cell area can be saved through abutting. In the conventional layout in Figure 3, the cell dimensions are 7.65 μm by 4.5 μm while the dimensions of the cell with abutted drain diffusion regions in Figure 4 are 6.1 μm by 5.0 μm.[2] Not only is the area reduced by 3.93 μm^2 (11.4%), but the larger width of the latter cell simplifies PE pitch-matching.

[1] Cell stability during a read cycle was guaranteed by sizing the NMOS pulldown devices to be three times as large as the access devices.

[2] Note that some process layers, such as horizontal paths in Metal 3 and body implants, have been omitted from the layout plots for clarity.

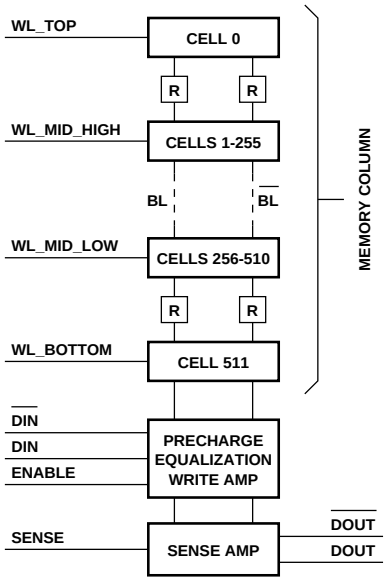

Figure 1: Block Diagram of an SRAM Column

This cell retains the strapped wordline, and the modified transistor arrangement permits a greater number of shared contacts. Manhattan layout style was used to avoid problems when processing the CAD and layout data.

The sense amplifier schematic for our embedded SRAM [8] is shown in Figure 5. This particular sense amplifier does not require bitline isolation transistors since the bitline inputs are decoupled from the data outputs. This reduces the number of critical edges required during a read cycle. To read a memory cell, transistors M4 and M5 form a differential pair that senses the voltages on the bitlines. A sense operation begins by asserting signal SET to turn on M6. The D_{out} and $\overline{D_{out}}$ nodes start to fall from the precharged value of V_{dd}. If the voltage on BL is slightly higher than $\overline{BL}$, the voltage on the $\overline{D_{out}}$ node falls faster than the complementary node voltage. Node D_{out} is eventually pulled back up to V_{dd}, and the sensed value is latched. Increasing the gate widths of M4 and M5 and minimizing the gate width of M6 improved sense amplifier sensitivity. Such sizing does slow down the latching speed, but the experimental nature of the SOI process led us to be conservative in our design.

The write amplifier and precharge circuit (see Figure 6) consists of tri-state inverters, as well as clocked precharge and equalization transistors. The inverters are enabled with stacked transistors rather than NAND-NOR gates to reduce layout area. Figure 11 depicts the corresponding layout. The write amplifier is on the right, while the sense amplifier, precharge and equalization circuits are on the left.

3. DESIGN CONSIDERATIONS RELATED TO FD-SOI PHENOMENA

Design considerations for PD-SOI memories have been discussed in several published works [2, 9, 10]. For PD-SOI

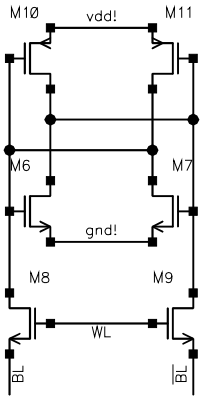

Figure 2: 6-T SRAM Cell Schematic

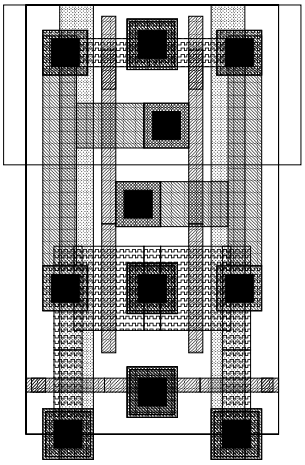

Figure 3: Conventional SRAM Cell Layout

RAMs it has been established that asymmetric bitline loading caused by the floating body voltage introduces small variations in the time needed for a sufficient bitline differential to develop during a read. Furthermore, floating body parasitic bipolar currents and increased subthreshold leakage during the write cycle have to be accounted for in any design since they increase the time that it takes for the bitlines to swing rail-to-rail. While these effects may be less noticeable in FD-SOI, they can still be observed in circuit simulations and need to be duly considered during design. As well, in the sense amplifier, the floating body voltage can cause threshold voltage mismatches if the same data is sensed repeatedly. The usual solution to this problem is to introduce body contacts to tie the sense amplifier transistor bodies to their source terminals or to the supply rails. However, body contacts are less effective in FD-SOI than in PD-SOI, and many circuit simulators do not support them, so characterization of the floating body sense amplifier is required.

Accurate simulation of FD-SOI circuits has been difficult in the past because of the lack of proper models. Re-

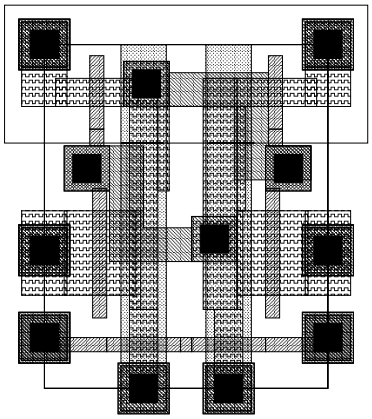

Figure 4: Layout of SRAM with Abutted Drains

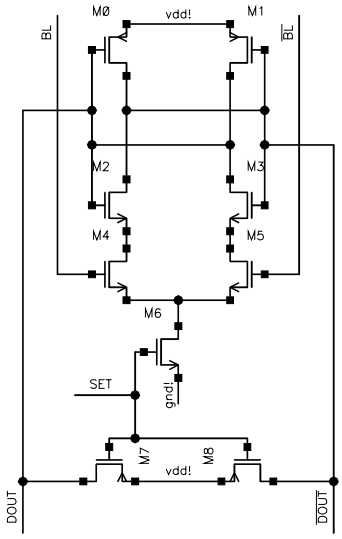

Figure 5: Sense Amplifier Schematic (from [8])

cently, however, more reliable physical-based models for FD-SOI transistors have been developed [11]. Our simulations are based primarily on the University of Florida SOI physical model, which was incorporated into HSPICE as "Level 58". Parameter evaluation and process-based calibration for our particular technology have been undertaken [12] and should provide us with accurate results that encompass issues peculiar to SOI. One of the most important concerns is modeling the parasitic bipolar transistor effect, which can be selected by setting an HSPICE model parameter. We also pursued the option of performing similar simulations with an HSPICE "Level 49" model originally intended for bulk CMOS, but calibrated to closely approximate our FD-SOI process [5]. This bulk CMOS model takes into account many aspects of our process, such as subthreshold leakage, but unfortunately does not take into account floating body effects.

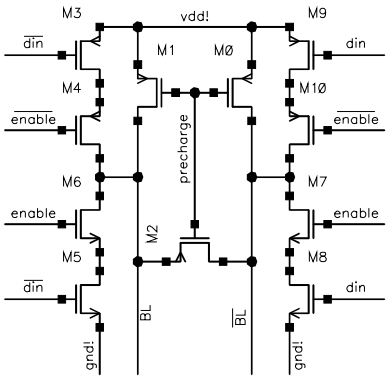

Figure 6: Write Amplifier and Precharge Schematic

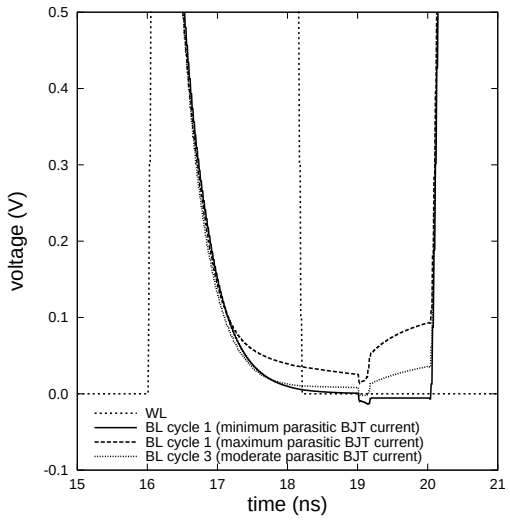

Figure 7: Write Cycle with Parasitic Bipolar Effects

4. SIMULATION RESULTS

Extracted simulations were performed on a schematic that represented a column of 512 SRAM cells. To reduce simulation time, blocks of cells in the column were created that multiply transistor loading and parasitic capacitances by a specified factor (see Figure 1). Since extraction of parasitic resistances was not supported by the simulator, resistive elements (shown labeled "R") were added to the bitlines in a T-model arrangement. The linear capacitance required by the T-model was simply the sum of existing parasitic bitline capacitances. Note that the central block of SRAM cells in Figure 1 was split to support simulation of bitline load imbalance. Cell characterization was done for the cell that is the furthest from the sense and write amplifiers. All simulation measurements in the time domain were made from one-half V_{dd}. In addition, the design was simulated using transistors from TSMC's 0.25 μm bulk CMOS process in order to demonstrate the speed advantage of SOI.

During a write cycle, unselected cells may present themselves as bitline leakage sources due to transient parasitic

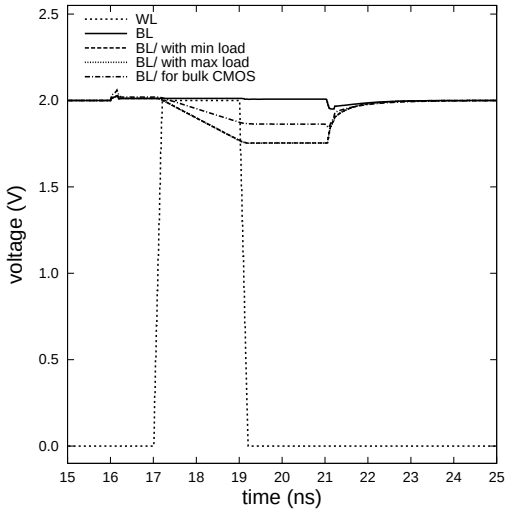

Figure 8: Bitline divergence during a read operation

bipolar currents and higher than normal subthreshold leakage. While the column of memory is idle, the bodies of the access transistors are electrically floating and drift to a potential between the drain and source voltages [9]. The charged body forms the base of a transient parasitic bipolar NPN transistor with the source as the emitter. When the source is pulled down a diode drop below the base, bipolar currents will flow until the floating body is discharged. Furthermore, raised body potentials result in lower transistor threshold voltages and increased subthreshold leakage. This effect is transient in nature during the write cycle since the raised V_{bs} will eventually be restored to a lower value upon removal of the body charge. Thus, the first write cycle after a long period of dormancy will have the highest unselected cell leakage.[3] When the bitlines are precharged to V_{dd}, the situation of maximum parasitic bipolar currents and subthreshold leakage occur on the true or complement side of the bitline when trying to "write 0" to a polarized column of all 1's or a "write 1" to a polarized column of all 0's.

In our simulations, we tested for the case of minimum and maximum bitline parasitic BJT currents during a write cycle. The memory cells were left dormant for a period of 1 ms to allow adequate body charging to occur. Minimum bitline parasitic BJT current during a write cycle was tested by writing a 0 to a polarized column of all 0's. In this case, the bitline reaches V_{ss} before leveling off, as shown in Figure 7. Maximum parasitic BJT current was tested by writing a 0 to a polarized column of all 1's so that the bipolar currents add up on the true side bitline. Figure 7 illustrates how the bitline with maximum parasitic BJT current levels off approximately 40 mV higher than the case for minimum parasitic BJT current. Figure 7 also shows the case of the

[3] Note that a read does not constitute a period of activity in this situation since reads do not pull the source of the access transistors low enough to discharge the body. Therefore, parasitic bipolar currents are not noticeable during read operations.

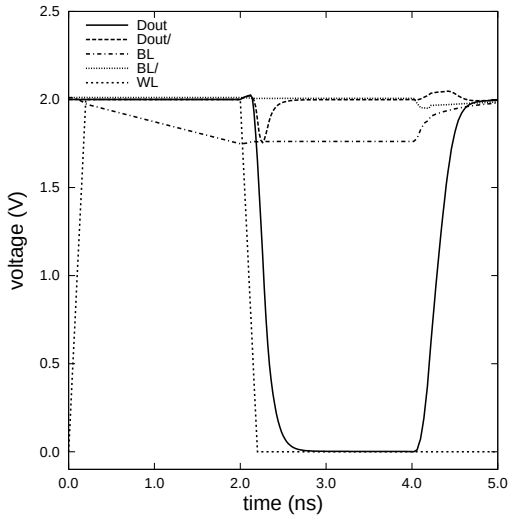

Figure 9: Reading "0" for the First Time

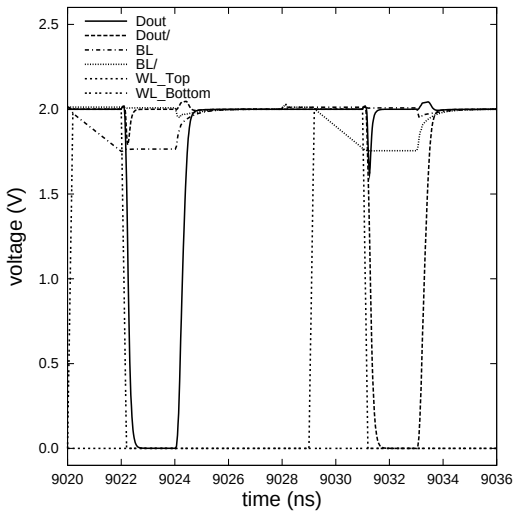

Figure 10: Reading "1" after Reading "0" 1000 Times

third consecutive write cycle following the case of worst-case parasitic bipolar current. The third cycle exhibits moderate parasitic bipolar current since the body charges have been reduced by the preceding two write cycles. In a PD-SOI SRAM, the bitline with maximum parasitic bipolar effect levels off about 200 mV higher than the bitline exhibiting minimum parasitic bipolar effect [2]. Therefore, our results show that the parasitic bipolar effect is approximately five times higher in PD-SOI than in FD-SOI.

Write disturb concerns for unselected cells due to parasitic bipolar currents is not a problem in this design since the unselected cell does not toggle state even under the conditions of maximum bipolar currents. Memory write time was measured from activation of the word line to the cell state falling to one-half of V_{dd}. The worst case time to write a cell was measured as 630 ps. No noticeable variation in write times was observed between maximum and minimum parasitic BJT currents for our FD-SOI SRAM. However, for the PD-SOI SRAM, the write times varied by 20% between the maximum and minimum parasitic BJT currents [2]. For comparison, a write of a cell in the TSMC bulk process required 1.28 ns. This speedup is again the result of reduced device drain capacitance on the bitlines.

During a read operation, the floating bodies of the cell access transistors may cause the device capacitance contribution to the bitlines to vary with the stored value. That is, the programmed "1" side of the cell can contribute slightly more capacitance to its bitline than the programmed "0" side. To measure this effect, a polarized column of all 1's, except for the selected cell, was simulated against a polarized column of all 0's except the selected cell. The former case represents minimum loading on the complement bitline while the latter case represents maximum loading.

As shown in Figure 8, there is no difference, within the limits of simulation error, in the developed offset between the maximum and minimum loaded bitlines even after 1 ms of memory column dormancy. The offset difference of 11 mV in a PD-SOI SRAM was minimal [2]. The measured bitline differential is approximately 241 mV for a wordline strobe of 2 ns. For the 0.25-μm bulk process, the developed differential was only 151 mV for the same wordline duration. This demonstrates the benefits of reduced junction capacitance on dual bitlines and enables early firing of the sense amplifier for more aggressive designs. For our memory array, the access time measured from wordline assertion until the data outputs fall to half V_{dd} at the sense amplifier was 2.21 ns.

The sense amplifiers of an array column should contain closely matched devices in order to discriminate small voltage differentials across the true and complement bitlines [13]. However, with the bodies of the sense amplifier transistors floating in SOI, devices may become mismatched due to varying threshold voltages. This unwanted hysteresis effect causes the same logical value to be more easily read in subsequent cycles. Reading the same logical value over hundreds or thousands of cycles can cause the sense amplifier to develop a preferential bias toward reading that value. Since HSPICE "Level 58" does not support simulation of body contacts, we chose to leave the sense amplifier transistor bodies floating and to develop a larger differential on the bitlines in order to overcome possible worst-case preferential bias.

Discounting floating body loading effects, the wordlines were normally strobed so that the developed offset voltage is approximately 241 mV before activation of the sense amplifier. This effect was simulated for a read of 1000 zeros followed by a read of a one (see Figures 9 and 10). The sense amplifier was fully functional for this case of extreme read bias. Bias in the sense amplifier transistors could be seen as an decrease in the voltage sag of the $\overline{D_{out}}$ node before it evaluates high. For the first read of a 0 at 20 ns, $\overline{D_{out}}$ sags 1.753 V, whereas for the one-thousandth read of a 0 at 9020 ns, $\overline{D_{out}}$ sags only to 1.790 V. Furthermore, when a 1 is read at 9029 ns, D_{out} sags to 1.621 V because it has to

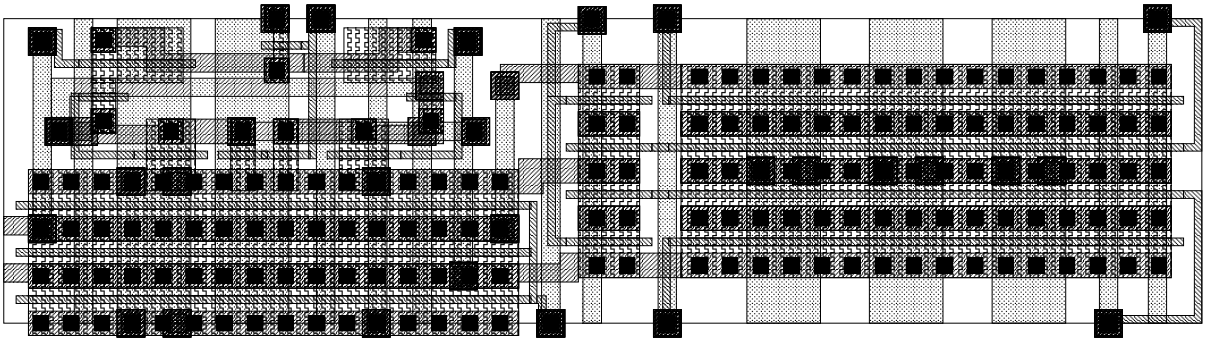

Figure 11: Layout of Write Amplifier and Precharge Circuit

overcome sense amplifier hysteresis.

5. CONCLUSION

In this paper, we investigated several potential challenges in the design of an embedded SRAM in an experimental, fully-depleted SOI process for use in a logic-in-memory application. Bitline parasitic bipolar transistor currents caused by floating body effects were present in our FD-SOI SRAM, but to a lesser extent than in a PD-SOI SRAM. The maximum parasitic bipolar current has no effect on the worst-case memory write time of 630 ps. We considered asymmetric bitline loading during the read cycle and found that this effect was negligible for the sensing operation. The worst-case memory access time was 2.21 ns for the SRAM core. In the absence of the ability to simulate body contacts, worst-case floating body effects had to be carefully modeled and accounted for in the design of the sense amplifier. Another practical challenge to FD-SOI designers is the current difficulty in acquiring accurate models and CAD tool support. Despite these challenges, we have demonstrated the practicality of designing embedded SRAMs in an academic FD-SOI process.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the invaluable technical assistance provided by Dr. Lisa G. McIlrath of 3D-IC, Dr. Jim Burns of Lincoln Laboratories, Massachusetts Institute of Technology, and Dr. Jerry G. Fossum of the SOI Group, Dept. of Electrical and Computer Engineering, University of Florida (Gainesville, FL). Dr. McIlrath made our project possible by arranging for access to Lincoln Labs' experimental FD-SOI process.

7. REFERENCES

[1] S. Cristoloveanu, "Silicon on Insulator Technology," ch. 4 in *The VLSI Handbook*, W.-K. Chen (ed.), CRC Press LLC, 2000.

[2] J. B. Kuang *et al.*, "SRAM Bitline Circuits on PD SOI: Advantages and Concerns," *IEEE J. of Solid-State Circuits*, vol. 32, no. 6, Jun. 1997, pp. 837-844.

[3] S. Park *et al.*, "A 0.25-μm, 600-MHz, 1.5-V, Fully Depleted SOI CMOS 64-Bit Microprocessor," *IEEE J. Solid-State Circuits*, vol. 34, no. 11, Nov. 1999, pp. 1436-1445.

[4] J. A. Burns *et al.*, "Performance of a Low Power Fully-Depleted Deep Submicron SOI Technology and its Extension to 0.15 μm," *Proc. 1996 IEEE Int. SOI Conf.*, Oct. 1996, pp. 102-3.

[5] J. A. Burns *et al.*, "Design Criteria for a Fully-Depleted 0.1-μm SOI Technology," *Proc. 1997 IEEE Int. SOI Conf.*, Oct. 1997, p. 78-9.

[6] D. G. Elliott *et al.*, "Computational RAM: Implementing Processors-In-Memory with Low Area and Power Overhead," *IEEE Design & Test of Computers*, Jan./Mar. 1999, pp. 32-41.

[7] K. Kumagai *et al.* "A New SRAM Cell Design Using 0.35μm CMOS/SIMOX Technology," *Proc. 1997 IEEE Int. SOI Conf.*, Oct. 1997.

[8] A. Chandrakasan *et al.*, "Register Files and Caches," ch. 14 in *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.

[9] P. F. Lu *et al.*, "Floating Body Effects in Partially Depleted SOI CMOS Circuits," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, Aug. 1997, pp. 1241-1253.

[10] A. G. Aipperspach *et al.*, "A 0.2-μm, 1.8-V, SOI, 550-MHz, 64-b PowerPC Microprocessor with Copper Interconnects," *IEEE J. Solid-State Circuits*, vol. 34, no. 11, Nov. 1999, pp. 1430-35.

[11] J. G. Fossum and S. Krishnan, "Physical Modeling Needed for Reliable SOI Circuit Design," *IEICE Trans. on Electronics*, vol. E80-C, no. 3, Mar. 1997, pp. 388-93.

[12] M.-H. Chiang *et al.*, "UFSOI Model Parameter Evaluation: Process-Based Calibration," U. Florida Tech. Rep. FL 32611-6130, Nov. 1998.

[13] K. Bernstein *et al.*, "SRAM Cache Design Considerations," ch. 6 in *SOI Circuit Design Concepts*, Kluwer Academic Publishers, 2000.