A Low-Power High-Performance Embedded SRAM Macrocell

A. M. Fahim, M. Khellah, and M. I. Elmasry
VLSI Research Group, Department of Electrical and Computer Engineering
University of Waterloo, Waterloo
Ontario, CANADA

Abstract

A new approach to modeling the decoding hierarchy in a hierarchical word line (HWL) SRAM architecture using integer-linear programming (ILP) is introduced. Using this approach, the HWL architecture is shown to be inadequate for very large SRAM sizes. Alternatively, a new low-power high-speed SRAM architecture is described. This architecture is shown to have fairly constant speed and power dissipation for sizes ranging between 32kb to 4Mb. Low-power is achieved by a voltage boosting technique not requiring a two-step voltage [7], and by a new method of tristating memory cells during a write operation. The SRAM was implemented in a 0.35μm CMOS technology operated at 150MHz while dissipating only 10mW.

1 Introduction

The growth of data intensive portable applications has led to aggressive developments in high-speed and low-power DSPs with large embedded SRAMs. The issue of finding architectural transformations to achieving low-power and high performance SRAMs has been often addressed by using the HWL architecture [1]. In this architecture the SRAM is broken up into smaller blocks and connected by three levels of decoding. For SRAM sizes larger than 1Mb, the HWL architecture may produce an excessively large number of blocks. Capacitive loads of 30pF on address lines have been reported for multi-Mb SRAM sizes [2]. This imposes a severe bottleneck for high performance SRAMs.

Recently, several low-power high-performance SRAM architectures have been proposed. One such architecture is the DSL cellarchitecture [3]. It features low-power write by tristating the memory cells' V_{SS} line during a write operation. High-speed read is achieved by pulling memory cells' V_{SS} line to a negative value during a read. This increases in the effective V_{GS} of the access and drive transistors, which decreases the read access time while operating with low subthreshold leakage currents. As has been reported in another study [4], the DSL architecture suffers from a power inefficient negative power supply.

Another low-power high-performance SRAM architecture, named "Over-Vcc Grounded Data Storage (OVGS)", has recently been proposed [4]. This architecture features a single block 1Mb SRAM. An ultra-low supply voltage of 0.5V has been used to achieve 100MHz operation while dissipating only 5mW. This has been achieved through the use of a multiple threshold CMOS (MTCMOS) technology, and by using of a boosted supply voltage. A V_{DD}=0.5V has been set to limit the subthreshold current to the desired level. The boosted supply voltage was set to satisfy the speed requirements. Charge recycling was also used to reduce the power dissipated in the bit lines of the unaccessed memory cells of the activated row. The disadvantage of this architecture is that it requires large voltage boosters since both the word line and bit lines are boosted. It also requires large capacitors per 4 words to perform charge recycling.

In this work, a new SRAM architecture enabling low-power write in a single-V_T 0.35μm CMOS technology is reported. An integer-linear programming (ILP) SRAM model is reported in Section 2. In Section 3, the new SRAM architecture and its associated circuitry is reported, followed by its performance evaluation in Section 4. Finally, a summary of the results is given in Section 5.

2 SRAM ILP Model

The HWL SRAM architecture is shown in Fig. 1. This architecture consists of a number of groups (G), memory blocks per group (BPG), words per block (WPB), and columns per block (CPB). The global decoder enables the address lines to be sent to only one group of memory blocks. The block decoder enables the address lines to be sent to only one memory block. This is effective in reducing the load capacitance on the address lines. The optimization problem now becomes one of finding the optimal values of the four parameters stated above given that delay must be minimized under energy constraints.

There are two types of delays: gate delay and wiring delay. A simple linear current model was used for gate delay and is given as [8]

$$t_{decoder} = \frac{C_L V_{DD}}{k(V_{DD} - V_T)} + k_2 N \qquad (1)$$

where k depends on device geometry, N is fan-in, and k_2 is the delay for a 2-input AND gate. Note that both N and C_L can be expressed in terms of the four model parameters G, BPG, WPB, and CPB. This linear delay model is valid for deep submicron CMOS devices and can be considered as

0-8186-8409-7/98 $10.00 © 1998 IEEE

an overestimate for larger CMOS devices. Wiring capacitance is simply given as

$$t_{wire} = R_{wire} \cdot C_{wire} \cdot \qquad (2)$$

Note that both R_{wire} and C_{wire} also depend on the 4 model parameters. The energy of the decoders is simply given as

$$E_{decoder} = \sum \alpha \cdot C_L V_{DD}^2 \qquad (3)$$

where the switching activity, α, may be assigned arbitrary values, depending on the memory addressing behavior of the code being executed. For modeling purposes, a switching activity of 0.5 has been assumed and a data width of 32 bits has also been assumed.

A conventional SRAM block has been assumed for the purposes of model construction. Linear delay and power models have been constructed for the conventional SRAM. The delay is given as

$$t_{block} = k_3 CPB + k_4 WPB \qquad (4)$$

The first component represents the delay due to word line driving and the second component represents the delay associated with the bit lines. The CPB parameter has been lower limited to the word size of 32 bits. Using smaller column lengths for an SRAM block has been shown to result in a less energy efficient SRAM [10]. Both k_3 and k_4 are constants. The energy of the SRAM block is given by

$$E_{block} = k_5 V_{DD} \tau + k_6 WPB \cdot V_{swing} \cdot V_{DD} \qquad (5)$$

where k_5 is any DC current consumption and k_6 is a constant related to the power dissipated along the bit lines. In SRAM design, the power due to driving the word line is very small in comparison to the power dissipated in the bit lines [5]. In this implementation, for example, the bitline capacitance of a 32K x 32b SRAM is 500x larger than the word line capacitance. For this reason, only the bit line dependence on power is taken into account. The first term represents DC power that may be dissipated in the peripherals, and τ is the average duration that DC current is dissipated per clock cycle.

This ILP model was simulated in GAMS [9] and the results of these simulations are shown in Figs. 2-3. Fig. 2 shows the effect of increasing the SRAM size on delay and energy. Clearly, the performance varies quite drastically even if an efficient architecture such as HWL is used. Attached to each point is the optimal configuration reported as $G-BPG-(WPB \times CPB)$. Fig. 3 shows the effect of constraining the SRAM's energy on the minimum delay and optimal configuration for a 256Kb SRAM block. Since bit line capacitance is larger than that of the word line, CPB must be maximized for minimum energy. The CPB parameter has been upper limited to 256 bits for these simulations. Exhaustive HSPICE simulations confirmed that the SRAM model is accurate within 20%.

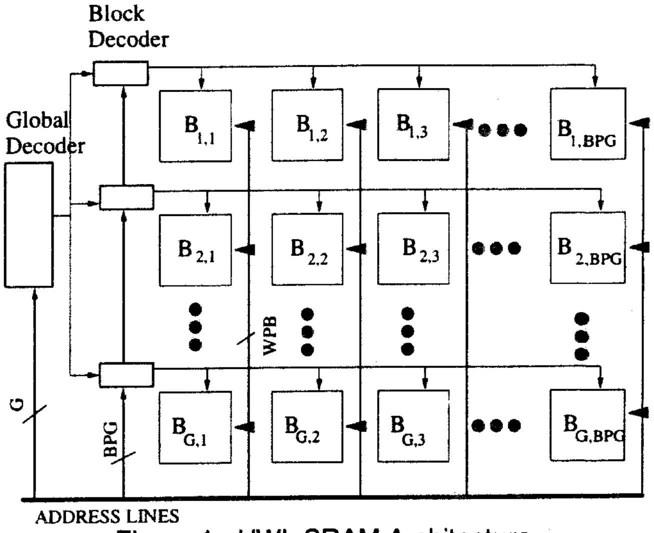

Figure 1. HWL SRAM Architecture

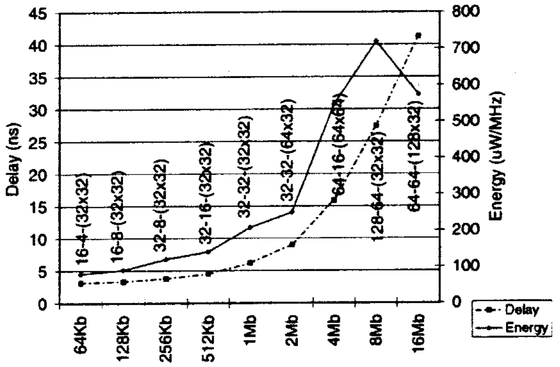

Figure 2. ILP model - minimum delay with no constraints

In all cases, the HWL SRAM architecture does seem to have a performance bottleneck when approaching the multi-megabit range. This is due to the large number of SRAM blocks present in the HWL array. Clearly, the method to minimize delay and power is to make the SRAM block more efficient. In terms of the model described above, the optimization problem now can be written as

$$\min \sum_{i=1}^{3} \{t_{decoder,i} + t_{wire,i}\} + t_{block} \qquad (6)$$
$$\text{s.t.} \quad E_{decoder} + E_{block} < E_{budget}$$

where i represents the hierarchical decoding level, and E_{budget} is the energy constraint value which is set by the SRAM's energy budget. Close examination of equations (1) to (5) reveal that the optimization problem of equation (6) is equivalent to minimizing constants $k_3 - k_6$.

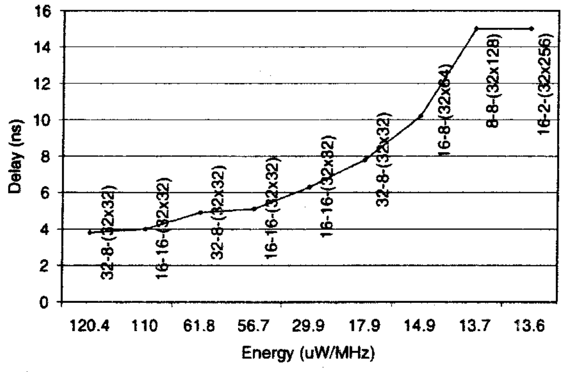

Figure 3. ILP model: minimum delay with energy constraint

3 SRAM Architecture and Circuit Design

As stated in the previous section, energy of an SRAM block is minimized if the CPB parameter is maximized. This, however, leads to large delays. The only way that constants k_3 – k_6 can be simultaneously minimized is to break the dependence of the SRAM's power & delay on the SRAM's vertical (WPB) and horizontal (CPB) dimensions.

Such a task would open way to maximizing the SRAM's block size, which may eliminate excessive decoding wire capacitance for multi-Mb SRAMs. As stated earlier, most of the energy dissipated in an SRAM is due to bit line toggling. Two types of operations access the bit line: read operations and write operations. For read operations, the bit-line clamped current sense amplifier (BLC-ISA) [6] was used. As with other current sense amplifiers, the BLC-ISA enables the sensing operation to be fairly independent of the length of the bit lines. The BLC-ISA has also been demonstrated to work well for low supply voltages. For convenience, the BLC-ISA is reproduced here and shown in Fig. 4.

For write operations, a new SRAM architecture, shown in Fig. 5 was used. This new architecture features voltage boosting during write. As Fig. 6 shows, the static noise margin (SNM) of the memory cell is drastically reduced when the word line voltage (V_{WL}) is above V_{DD}+V_T. This means that very small bit line swings are now achievable using this feature.

Two main problems exist with V_{WL} boosting approach. As reported in [7], boosting the V_{WL} is not a practical approach due to the reduced stability of the other words in the same row as the word to be accessed. Instead, the V_{WL} was limited to a short boosted voltage pulse followed by a V_{DD} voltage level, which limits the advantage of word line boosting. This limitation is eliminated in our architecture by having a different voltage booster for each set of word columns. This is similar to the partial word line activation technique [5]. The voltage level on the virtual V_{DD} line is controlled through Y-decoding. If, for example, a word in the Y0 block (see Fig. 5) is accessed, then its voltage booster would supply high V_{DD} (HV_{DD}) to the selected word, whereas the equivalent word in the Y1 block would receive just V_{DD}. Note that power consumption due to toggling the virtual V_{DD} line is recycled by discharging the boosted voltage back to the power supply line.

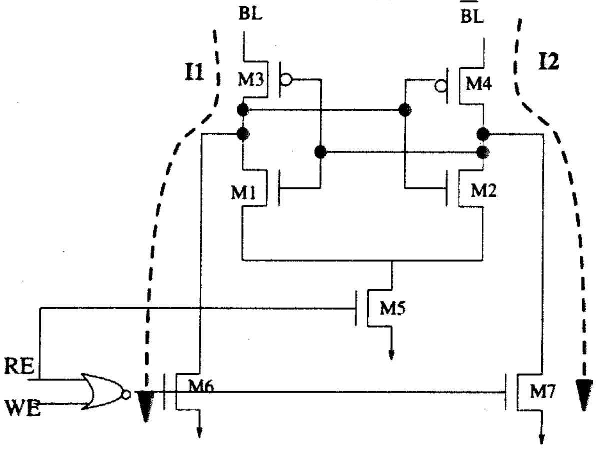

Figure 4. The bit-line clamped current sense amplifier

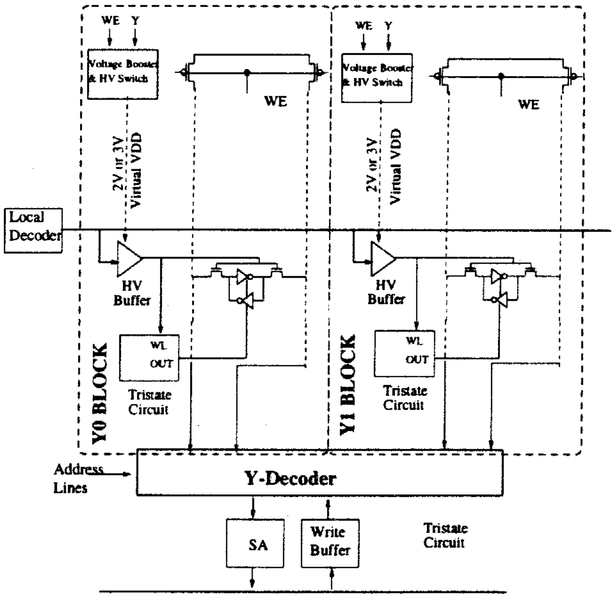

Figure 5. Low-power Write Architecture

To handle voltages higher than V_{DD}+V_T, a special high voltage (HV) buffer is needed. This is necessary to avoid high dc currents caused by both the PMOS and NMOS devices in an inverter being turned on simultaneously during a write operation as demonstrated in Fig. 7. The schematic of the HV buffer is shown in Fig. 8. The feedback PMOS device is necessary to bring the gate voltage up to HV_{DD} in case the word is not accessed. The coupling capacitor is needed to prevent HV_{DD} from feeding prior gates. The size of the capacitor must be large enough as to allow sufficient voltage to toggle the inverter receiving the high voltage. The capacitor size is also upper bounded by the need for the feedback PMOS device, P2, to toggle quickly in order to limit the short circuit current, without leaking charge to node B. To avoid these two conflicting requirements, the capacitor is sized sufficiently large as to charge up node A. A bypass NMOS device, N2, is used to discharge node A. Transistor N3 helps in bringing WL to zero as quickly as possible as to immediately turn on P2 in order to avoid excessive leakage current through inverter P1-N1.

Another problem associated with the voltage boosting technique is due to node conflict of the memory cells with the write buffers. Due to WL voltage boosting, a small swing is theoretically needed on the bit lines during a write. Hence, the write buffer sizes may be reduced. If the write buffer sizes are comparable to the memory cells' latch sizes, this would create a node conflict causing large DC current flow. This also causes the voltages on both bit lines to temporarily drop (shown in Fig. 9) causing excessive power dissipation. In order to prevent this, the write buffers have to be sized up. This, however, would offset the power savings of boosting the V_{WL} to above $V_{DD}+V_T$.

Another method of dealing with this problem is to tristate the memory cells during a write operation. This was used successfully in the DSL architecture. The signal used to tristate the memory cells during a write operation was derived from the write-enable (WE) signal. This means that the WE signal must be routed throughout the entire SRAM. This may prove to be a power expensive operation. Another method of detecting whether a write operation is being performed on a particular word is to use a differential amplifier with one input being the word line and the other being $\frac{HV_{DD}+V_{DD}}{2}$. This method was employed in this study. Such a circuit is shown in Fig. 10 and placed in the SRAM architecture as shown in Fig. 5.

The H_{VDD} input node is actually the word line and not the virtual V_{DD} line. This reduces the load on the virtual V_{DD} line, and hence, increasing the efficiency of the voltage booster. The other input to the differential pair is maintained at V_{DD} when the word line is not activated. When the word line is chosen, this node is raised to $\frac{HV_{DD}+V_{DD}}{2}$ by using a small coupling capacitor. Hence, if a read operation is chosen, the output of the differential pair would be logic zero, and the output of the detector is 0. On the other hand, if a write operation is chosen, the output of the differential pair would be logic one, and the output of the detector is 1. An output of 1 would tristate all memory cells in the selected word. To cut down DC power consumption, the current source of the differential pair is controlled by a word-line-pulse enable signal, WL.

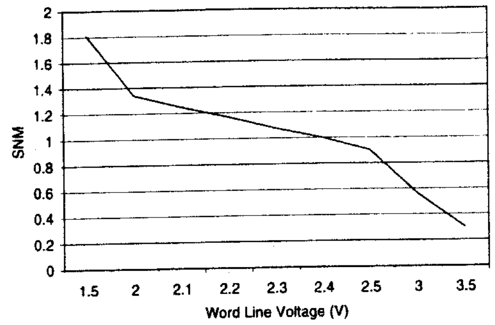

Figure 6. Effect of V_{WL} on SNM (V_{DD}=2V)

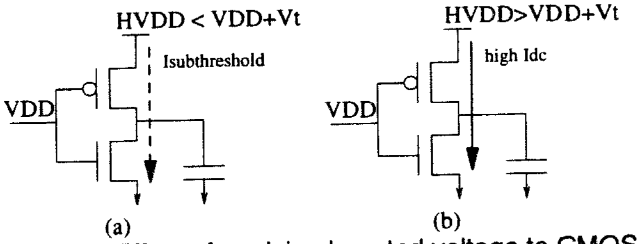

Figure 7. Effect of applying boosted voltage to CMOS inverter

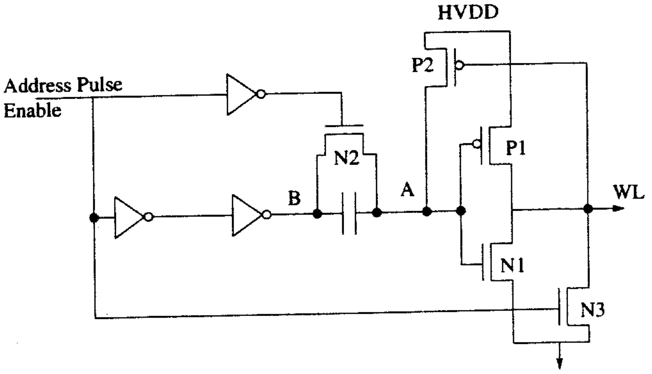

Figure 8. Schematic of the HV Buffer

Finally, decoding circuitry was constructed. Unlike the HWL architecture, decoding energy is not a dominant factor of the total power dissipation due to reduced wiring capacitance. For this reason, logic families, such as CMOS, dynamic CMOS (domino), pass transistor logic (PTL), and dynamic pass transistor logic (dynPTL), were evaluated in terms of their delay only. As Fig. 11 shows, dynPTL has the least delay. A 3 stage decoding hierarchy was implemented using the dynPTL gate. The CLK input of the dynPTL AND gate (Fig. 12) is actually the enable signal from the upper level in the decoding hierarchy.

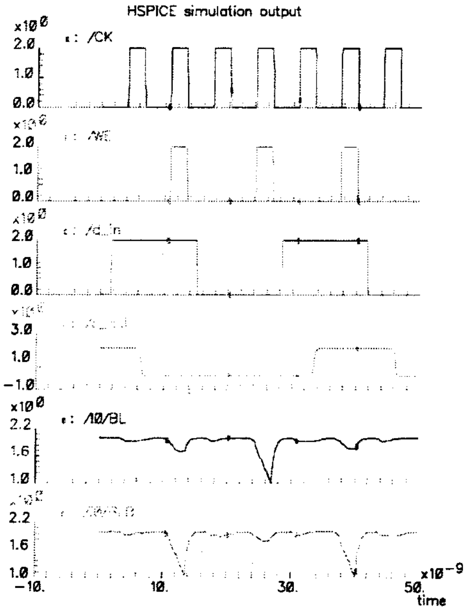

Figure 9. HSPICE output of SRAM demonstrating drop of both BL and $\overline{BL}$ for a 1Kx32 SRAM

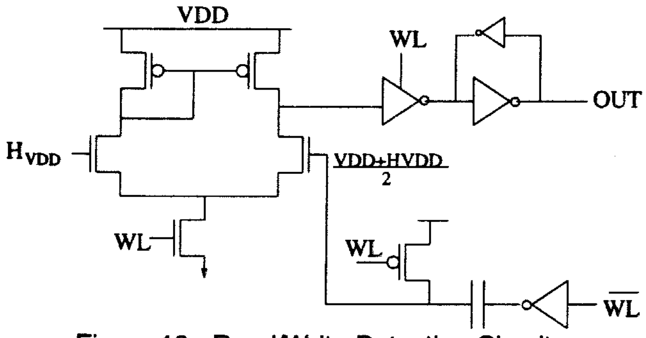

Figure 10. Read/Write Detection Circuitry

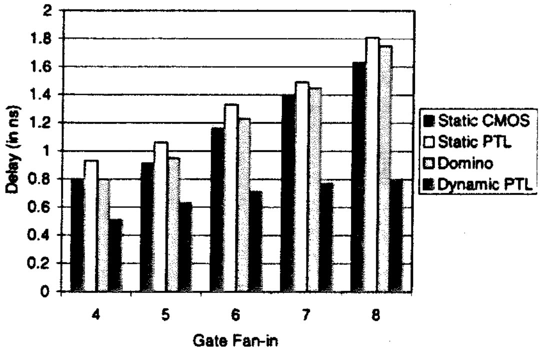

Figure 11. Comparison of Decoder Logic Families

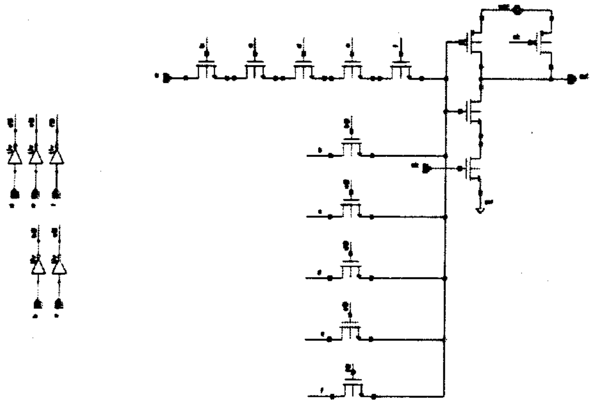

Figure 12. 6-input dynPTL AND gate

4 SRAM Performance

The SRAM has been implemented in a 0.35μm CMOS technology. The supply voltage of 2V and boosted V_{WL} of 3V was used. The minimum supply voltage was 1.5V (limited by the technology's $V_{TN}+V_{TP}$). The delay and power versus SRAM size is shown in Fig. 13. As desired, both the delay and power of the SRAM are fairly independent of size for a range of 8kb to 4Mb. The SRAM is compared to the DSL and OVGS architectures in terms of delay and energy in Fig. 14. Energy is measured in pW/MHz/bit. As the SRAM's power breakdown in Fig. 15 shows, the decoding power is now only a small fraction of the total power consumption.

Fig. 16 shows an HSPICE simulation of a 1Mb (16K x 64b) SRAM. The test vectors consisted of alternating reads and writes. Note the voltage boosting of the word line to 3V during a write operation. Voltage scaling of the SRAM from 1.5V to 2.5V is shown in Fig. 17. Note that a maximum voltage of 2.5V was set due silicon reliability. The performance of the SRAM at 1.5V is 70MHz.

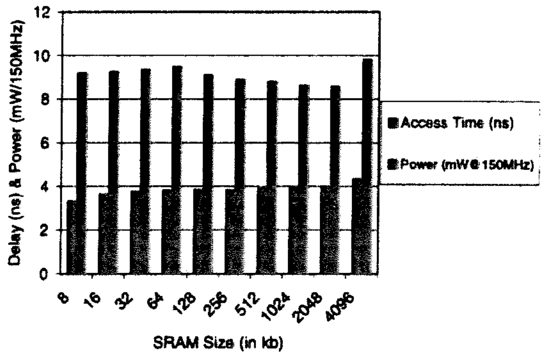

Figure 13. Delay and Power versus SRAM size

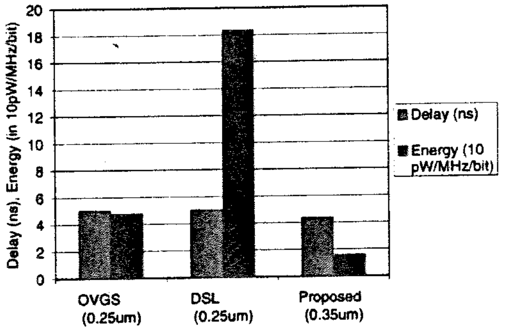

Figure 14. Delay & Energy of SRAM designs. Delays are for max size (OVGS=1Mb, DSL=8kb, proposed = 4Mb).

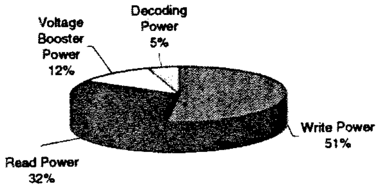

Figure 15. Breakdown of power in SRAM

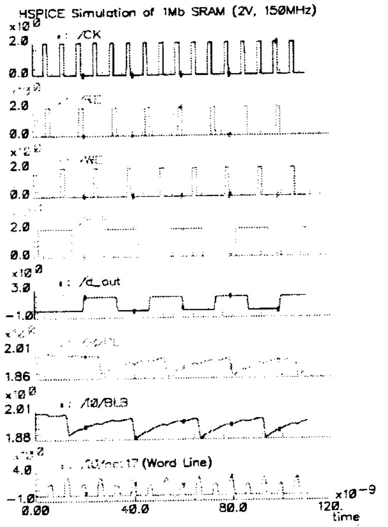

Figure 16. HSPICE simulation of a 1Mb SRAM

5 Summary

An ILP model has been constructed for the HWL SRAM architecture. The model reveals that for a conventional SRAM core, the HWL architecture fails to meet the energy and speed requirements of modern DSP and microprocessor applications. The ILP model of the SRAM has revealed the power and delay bottlenecks of an SRAM employing the HWL architecture. Using these results, a low-power high-performance SRAM architecture has been designed. This architecture features independence of power and delay from SRAM size for a range of 8kb to 4Mb. Low-power techniques included voltage boosting, charge recycling, and memory cell tristating. Performance of this architecture implemented in a 0.35um CMOS technology is 150MHz, 10mW at V_{DD}=2V.

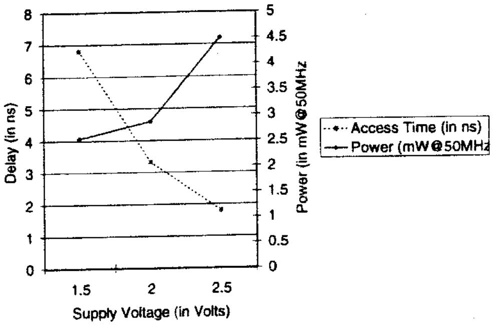

Figure 17. Delay & energy of a 1Mb SRAM with voltage scaling

References

[1] T. Hirose, et al., "A 20-ns 4-Mb CMOS SRAM with Hierarchical Word Decoding Architecture," *JSSC*, pp. 1068-1074, October 1990.

[2] K. Sasaki et al., "A 9-ns 1-Mbit CMOS SRAM," *JSSC*, vol. 24, no. 5, pp. 1219-1224, October 1989.

[3] H. Mizuno and T. Nagano, "Driving Source-Line Cell Architecture for Sub-1-V High-Speed Low-Power Applications," *JSSC*, pp. 552-556, April 1996.

[4] H. Yamauchi, et al., "A 0.5V / 100MHz Over-Vcc Grounded Data Storage (OVGS) SRAM Cell Architecture with Boosted Bit-line and Offset Source Over-Driving Schemes," *ISLPED*, pp. 49-54, 1996.

[5] K. Itoh, et. al., "Trends in Low-Power RAM Circuit Technologies," *ISLPED*, pp.84-87, 94.

[6] T. Blalock, "A High-Speed Clamped Bit-Line Current-Mode Sense Amplifier," *JSSC*, vol. 26, no. 4, April 1991.

[7] K. Ishibashi, et al., "A 1-V TFT-Load SRAM Using a 2-Step Word-Voltage Method," *JSSC*, pp. 1519-1524, Nov. 1992.

[8] M.I. Elmasry, *Digital BiCMOS Integrated Circuits*, IEEE Press, pp. 3-14, 1995.

[9] A. Brook, D. Kendrik, A. Meeraus, *GAMS: A User's Guide*, Scientific Press, Mass., 1992.

[10] R. Evans and P. Franzon, "Energy Consumption Modeling and Optimization for SRAM's," *JSSC*, vol. 30, no. 5, pp. 571-579, May 1995.