Low-Leakage Asymmetric-Cell SRAM[1]

Navid Azizi
ECE Department
University of Toronto
Toronto, Ontario, Canada
M5S 3G4
nazizi@eecg.toronto.edu

Andreas Moshovos
ECE Department
University of Toronto
Toronto, Ontario, Canada
M5S 3G4
moshovos@eecg.toronto.edu

Farid N. Najm
ECE Department
University of Toronto
Toronto, Ontario, Canada
M5S 3G4
f.najm@utoronto.ca

ABSTRACT

We introduce a novel family of asymmetric dual-V_t SRAM cell designs that reduce leakage power in caches while maintaining low access latency. Our designs exploit the strong bias towards zero at the bit level exhibited by the memory value stream of ordinary programs. Compared to conventional symmetric high-performance cells, our cells offer significant leakage reduction in the zero state and in some cases also in the one state albeit to a lesser extend. A novel sense-amplifier, in coordination with dummy bitlines, allows for read times to be on par with conventional symmetric cells. With one cell design, leakage is reduced by 7X (in the zero state) with no performance degradation. An alternative cell design reduces leakage by 40X (in the zero state) with a performance degradation of 5%.

Categories and Subject Descriptors

B.3.1 [**Memory Structures**]: Semiconductor memories

General Terms

Design

Keywords

SRAM, Low-leakage, Low-power, Dual-V_t

1. INTRODUCTION

As a result of technology trends, leakage (static) power dissipation has emerged as a first-class design consideration in high-performance processor design. Historically, architectural innovations for improving performance relied on exploiting ever larger numbers of transistors operating at higher frequencies. To keep the resulting switching power dissipation at bay, successive technology generations have relied on reducing the supply voltage. In order to maintain performance, however, this has required a corresponding reduction in the transistor threshold voltage. Since the MOSFET sub-threshold leakage current increases exponentially with a reduced threshold voltage, leakage power dissipation has grown to be a significant fraction of overall chip power dissipation in modern, deep-submicron ($< 0.18\mu$) processes. Moreover, it is expected to grow by a factor of five every chip generation [1]. For processors, it is estimated that in 0.10μ technology, leakage power will account for about 50% of the total chip power [2].

Since leakage power is proportional to the number of on-chip transistors, much of recent work in reducing leakage power has focused on SRAM structures such as the caches that comprise the vast majority of on-chip transistors. Existing circuit-level leakage reduction techniques are oblivious to program behavior and trade off performance for reduced leakage where possible [3]. Combined circuit- and architecture-level techniques reduce leakage for those parts of the on-chip caches that remain unused for long periods of time (thousands of cycles). These methods are not effective when most of the cache is actively used.

We present a family of novel *asymmetric* SRAM cell designs that lead to new cache designs which we refer to as the *Asymmetric-Cell Caches* (ACCs). ACCs offer drastically reduced leakage power compared to conventional caches even when there are few parts of the cache that are left unused. ACCs exploit the fact that in ordinary programs most of the *bits* in caches are *zeroes* for both the data and instruction streams. It has been shown that this behavior persists for a variety of programs under different assumptions about cache sizes, organization and instruction set architectures, even when assuming perfect knowledge of which cache parts will be left unused for long periods of time [4].

Traditional SRAM cells are symmetrically composed of transistors with identical leakage and threshold characteristics. Some recently proposed SRAM cells use symmetric configurations of transistors with different leakage and threshold characteristics [3]. These cells are either optimized for access latency or leakage power but not both. Our *asymmetric* SRAM cell designs offer low leakage with little or no impact on latency. In our asymmetric SRAM cells, selected sets of transistors are "weakened" to reduce leakage when the cell is storing a zero (the common case). In this work, we achieve the weakening by using higher-V_t transistors, however, this may also be possible by appropriate transistor sizing. We evaluate our designs by simulation, based on a commercial 0.13μ, 1.2V CMOS technology. The two best designs offer different performance/leakage characteristics. With one cell design, leakage is reduced by 7X (in the zero state) with no performance degradation. An alternative cell design reduces leakage by 40X (in the zero state)

[1]This project was supported in part by the Semiconductor Research Corporation (SRC 2001-HJ-901).

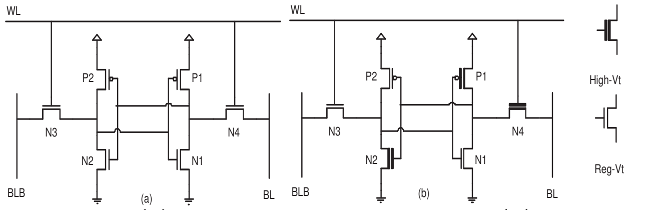

Figure 1: (a) Symmetric SRAM cell. (b) Original Asymmetric (OA) SRAM cell

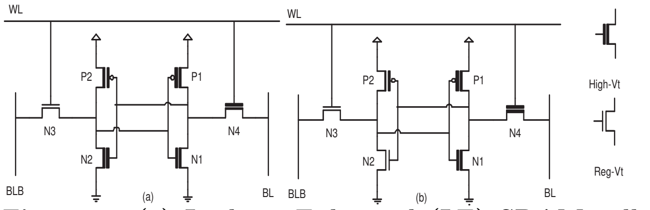

Figure 2: (a) Leakage Enhanced (LE) SRAM cell. (b) Speed Enhanced (SE) SRAM cell

with a sense time degradation of 10% (the total read cycle time is degraded by only 5%). By comparison, the use of an all high-V_t (HV) cell reduces leakage by about 40X but increases sense times by 26%.

We make the following contributions: (1) We introduce a novel family of asymmetric SRAM cells. No previous work on designing asymmetric SRAM cells exists. (2) We introduce a novel sense amp design that exploits the asymmetric nature of our cells to offer cell read times that are on par with conventional symmetric SRAM cells. (3) We evaluate a cache design that is based on ACCs and demonstrate that compared to a conventional cache, it offers drastic leakage reduction while maintaining high performance and comparable noise margins and stability.

The rest of this paper is organized as follows: In section 2, we present our asymmetric cell family. In section 3, we present the sense amplifier. In section 4, we present the simulation results of an SRAM using the different asymmetric cells. Section 5 includes a discussion on architectural level techniques to leverage the asymmetric nature of the cells. Finally, we conclude the paper in section 6.

2. ASYMMETRIC SRAM CELLS

Fig. 1(a) shows a conventional SRAM cell. In the inactive state, when the cell is not being written to or read from, most of the leakage power is dissipated by the transistors that are *i)* off and that *ii)* have a voltage differential across their drain and source. In Fig. 1(a), if the cell were storing a '0', transistors P1, N2 and N4 would dissipate leakage power. A simple technique for reducing leakage power would be to replace all transistors with high-V_t ones, but this unacceptably degrades the bitlines discharge times.

Since ordinary programs exhibit a strong bias in cache-resident bit values [5], another possibility to reduce leakage power, but at the same time keep read access times short, is to choose a preferred stored value and to only replace those transistors that contribute to the leakage power in this state with high-V_t transistors, as seen in Fig. 1(b).

This original asymmetric cell (OA) cell was simulated (at 110°C using SPICE models of a commercial 0.13μ, 1.2V CMOS technology and it exhibited the same leakage as the all regular-V_t (RV) cell when holding a logical '1', but it decreased leakage by 40X when holding a logical '0.' *Throughout this paper, we will use the following convention.* A high-V_t (HV) transistor is obtained from the basic 0.13μ, 1.2V, transistor (referred to herein as the *regular-V_t* transistor) by artificially increasing the V_t by 0.2V using the HSPICE in-line parameter DELVTO. It is understood that one may question the specific choice of 0.2V in practice. However, one can argue that the conclusions of this work, namely the feasibility and utility of using an asymmetric cell to reduce leakage, are valid irrespective of the specific value used for DELVTO. This specific value was selected, in our case, because it leads to a difference of about 10X between the leakage currents of HV and RV transistors, which is typical of dual-V_t technology.

The read access time of this OA cell is degraded. Due to N2's and N4's higher threshold voltage, they increase the bitline discharge time. The discharge times for the BLB and BL are 12.2% and 46.4% longer than the discharge times for the RV cell respectively. Read times can be made to match the faster read time by using a set of dummy bitlines and a novel sense amp, as is discussed in Section 3.

2.1 Two Improved Asymmetric SRAM Cells

Starting with the asymmetric cell of Fig. 1(b), we have investigated a total of 9 meaningful variations that offer different leakage and performance characteristics. In the interest of space, we present the two best designs, which are shown in Fig. 2.

The leakage enhanced (LE) cell in Fig. 2(a) offers better leakage behavior than that of Fig. 1(b) because it also dissipates reduced power when holding a logical '1' since N1 and P2 have been made high V_t. Compared to the RV cell it decreases leakage by 40X and 7X when holding a logical '0' and '1' respectively. The discharge times for this cell are 12.2% and 61.2% longer on BLB and BL respectively compared to the RV cell, but again dummy bitlines and a new sense amplifier allow the read times to match the fast side of the cell regardless of the data being stored (as will be seen in section 3).

The speed enhanced (SE) cell in Fig. 2(b) dissipates higher leakage compared to the cell in Fig. 2(a) but it allows for read times that are virtually identical to that of the RV cell. Compared to the all RV cell the SE cell decreases leakage by more than 2X and 7X when holding a logical '0' and '1' respectively. The discharge time along BL is 61.2% longer compared to the RV cell, but dummy bitlines allow for quick sensing.

2.2 Supply Voltage Analysis

Leakage power is becoming increasingly important given the trend of decreasing the supply and threshold voltages in successive technologies [7]. We have tested our asymmetric cells with different supply voltages, and appropriately scaled threshold voltages, to measure leakage, discharge times, and cell flip times (the time required to flip the cell state). Fig. 3 shows the leakage while holding a '0' and '1' for all cells under different supply voltages. The figure shows that the leakage savings incurred by using the asymmetrical cells continues for lower supply voltages, and becomes more important as the leakage current rises exponentially with smaller supply voltages.

The bitline discharge times on the fast side of the cell and flip times for all cells are shown in Fig. 4(a) and (b) respectively. While the discharge time of the LE cell is slightly

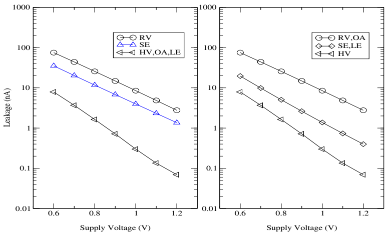

Figure 3: (a) Leakage when holding 0 (b) Leakage when holding 1

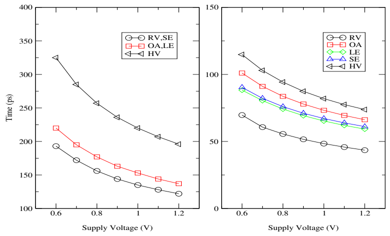

Figure 4: (a) Bitline discharge times (fast side) (b) Flip Times

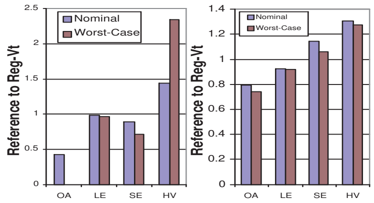

Figure 5: (a) Noise Margins (b) I_{trip}/I_{read} Stability

longer than that of the RV cell it is much shorter than that of the HV cell. The discharge time for the SE cell is virtually unchanged from that of the RV cell.

The cell flip times of the asymmetric cell all lie in between the cell flip times of the RV and HV cells, but, as seen in Fig. 4, are just a fraction of the discharge times.

2.3 Stability Analysis

Another major consideration with the cell design is its stability. There are two interrelated issues: read stability and noise margins [3][6]. Intuitively, read stability indicates how likely it is to invert the cell's stored value when accessing it, and was measured as the ratio of I_{trip}/I_{read} [3]. The static noise margin (SNM) of an SRAM cell is defined as the minimum dc noise voltage necessary to flip the state of the cell [8]. We have performed stability analysis on all the cells reported in this paper at a supply voltage of 1.2 V. Process variations were accounted for by performing the stability analysis under 59,049 combinations of different V_t and length for all six transistors in the cell. Fig. 5 shows the noise margins and stability of the cells normalized to those of the RV cell under nominal conditions and for the worst-case condition. While the OA cell fails under process variations, the LE and SE have comparable or better SNM and stability.

3. SENSE-AMPLIFIER

A conventional sense amplifier, shown in Fig. 6(a), is not suitable in our design due to the slow access time if the cell is storing a '0.' To obtain fast read times regardless of the data, a new sense amplifier was designed and is shown in Fig. 6(b). Compared to the conventional sense amp, the new sense amplifier has 4 extra transistors and an area increase of roughly 0.229 μm^2 or 14.4%. In addition, the sense amplifier uses a set of *dummy bitlines*, which are always fast (as fast as the fast side of the asymmetric cell), to trigger the reading of a logical '0' thus achieving fast access times when the slow bitline is discharging. Each pair of dummy bitlines are tied to the D and DB terminals of one column of dummy cells which all store a '1'. During every read operation one of the dummy cells will have its wordline asserted.

Sensing a '1' is as fast as a conventional sense amp since this is done by sensing a discharge of BLB due to the action of the fast side of the cell. Sensing a '0' is *initiated* at a later time than it would be in a conventional sense amp. This is done to allow sufficient time for the fast side to trigger the sense amp if it has to do so.

The sense amplifier operates as follows: Initially, the bitlines are precharged and all four amplifier inputs rise to VDD. If, during a read, BLB is being discharged (cell's fast side), then the differential pair composed of MN1 and MN2 causes increased current to pass through the left branch, thus increasing the voltage at node B and decreasing the

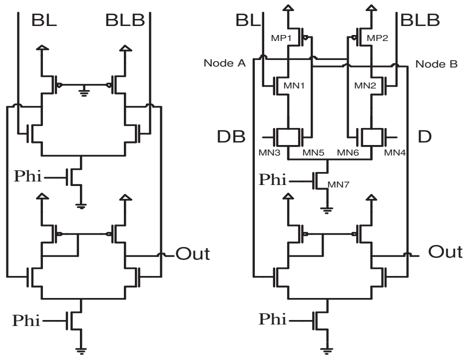

Figure 6: (a) Simple Sense Amplifier. (b) New Sense Amplifier

voltage at node A.

When BL is being discharged, then it does so at a slower rate since it is being discharged from the slow side of the asymmetric cell. To achieve fast sensing in this case also, the dummy bitlines, which are connected to the differential pair of MN3 and MN4, initiate the sensing of a logical '0.'

For this sensing scheme to achieve reliable results it must allow for adequate time for BLB to discharge before initiating a logical '0' read. This safety factor is achieved in two ways. First, the dummy bitlines are connected to all sense amps and therefore have a slightly higher capacitive load compared to real bitlines leading to a slower discharge on DB compared to BLB. The extra capacitive loading does not slow the sense time when BL is discharging because of the concerted effort between BL and DB to sense the same value. Second, the transistors connected to the bitlines are wider than the transistors connected to the dummy bitlines leading to a higher transconductance. This leads to a higher gain from the bitlines to the output than from the dummy bitlines. We have also performed sensitivity analysis of this sense amplifier, and it performs on par with the conventional sense amplifier.

4. SRAM

Using the above cells and the sense amplifier presented in section 3, a 32-Kbyte SRAM was designed and simulated to measure leakage, and read and write times. Each of the 128 SRAM sub-arrays contains 64 cells along each bitline, and 32 cells along each wordline. The SRAM was simulated at a temperature of 110°C with the RV, OA, LE, SE and HV cells. Furthermore the RV and HV cells were simulated with a conventional sense amp, and these results were used as a reference for our design.

Fig. 7 shows the total leakage within the SRAM attributable to the SRAM cells when the SRAM is either holding all '0's or all '1's. The leakage includes the leakage needed for the two sets of dummy cells. The leakage trends for the single cell in section 2 continue for the com-

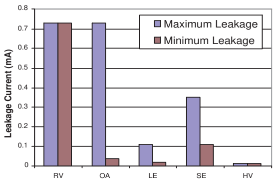

Figure 7: Max and Min Leakage Attrib. to Cells

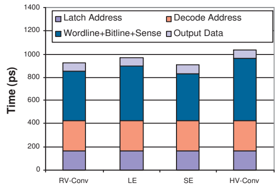

Figure 8: Breakdown of Memory Access Time

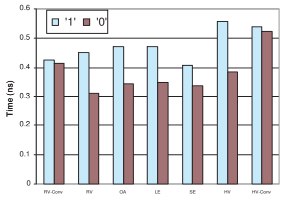

Figure 9: Sense times during a read cycle

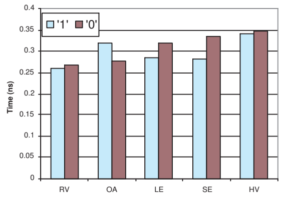

Figure 10: Write times for cells

plete SRAM where the LE and SE offer a reduction of 40X and 2X while storing a '0' and offer a reduction of about 7X when storing a '1.'

The total SRAM read access time includes four components: 1) input register propagation delay and hold times, 2) the address decoding delay, 3) the delay for wordline, bitline and sensing, and 4) the output register setup time. Our simulation results showing these components for the various SRAM arrays are shown in Fig. 8. Notice that only the 3rd component is affected by the cell design. Specifically, this time is the time period from when precharging is complete to when the sense amplifier has reached 90% of its swing.

Fig. 9(a) shows the sense times (the 3rd component of Fig. 8) for all cells. It can be seen that the worst-case sensing times are now on-par with the RV cell with a conventional sense amplifier. Compared with the RV cell with a conventional sense amp, the LE cell is 10% slower (although the total read time increases by about 5%, as seen in Fig. 8), but the SE cell is slightly faster (note this is not because the sense amplifier is quicker, but because the bitline discharge time for the SE cell is 50ps quicker than that of the RV cell, which is a byproduct of the asymmetry of the SE cell). The HV cell with a conventional sense amp would be 26% slower.

The write times for the different cells are shown in Fig. 10. The LE and SE cells show an increase of 19% and 25% respectively over the RV cell. The increase in write times is of minor importance since the write times are all shorter than the read times of the associated cells and therefore the speed of the SRAM is dependent on the read time.

5. ARCHITECTURAL ENHANCEMENTS

We investigated two cache organizations that use asymmetric cell designs: *statically biased* and *dynamic inversion*. In the *statically biased* cache, the cells are simply replaced with asymmetric ones. This cache is *statically biased* to dissipate low leakage power only when it stores the preferred bit value ('0'). What makes this cache successful is *typical* program behavior: as we show in [5], the SPEC2000 programs we studied exhibit a strong bias towards zero. The statically biased cache with the SE cells reduces leakage by 4.5X and 3.8X for an instruction and a data cache, respectively, compared to conventional symmetric-cell caches. The caches are 32Kbyte 4-way set associative caches. While programs with a higher fraction of '1's than '0's may exist, our SRAM would still dissipate much lower leakage power compared to the RV cell cache.

In *selective inversion*, the values stored within a block can be inverted at a byte granularity (other granularities are possible). In this design, if a byte contains five or more ones it is inverted prior to storing it in the cache. This cache needs an additional *inversion flag* cell per byte that holds information on which bytes were inverted. Inversion happens at write time. Since stores are typically buffered in a write buffer and are only sent to the data cache on commit, there is plenty of time to decide and apply inversion if necessary. Additional area, dynamic power and performance trade-offs apply to this design. An investigation of these issues is beyond the scope of this paper.

6. CONCLUSION

In this paper, we proposed a novel approach that combines both circuit- and architecture-level techniques. Our approach drastically reduces leakage power dissipation. The key observations behind our approach are that cache-resident memory values of ordinary programs exhibit a strong bias towards zero or one at the bit level.

We introduced a family of high-speed asymmetric dual-V_t SRAM cell designs that exploit this bit-level bias to reduce leakage power while maintaining high performance. The *speed enhanced* cell reduces leakage power by at least 2X and by 6X in the preferred state. It is as fast as the conventional, regular-V_t SRAM cell. By comparison, the *leakage enhanced* cell reduces leakage by at least 6X and by about 40X in the preferred state. Its sense time is 10% higher than the speed enhanced and the regular-V_t cells (total read time is only 5% higher).

7. REFERENCES

[1] S. Borkar, "Design challenges of technology scaling," *IEEE MICRO*,vol.19, no.4, pp. 23–29, July–Aug. 1999.

[2] T. Kam et. al., "EDA challenges facing future microprocessor design," in *IEEE Transactions on Computer-Aided Design*, vol. 19, no. 12, Dec. 2000.

[3] F. Hamzaoglu et. al., "Dual-Vt SRAM cells with full-swing single-ended bit line sensing for high-performance on-chip cache in 0.13um technology generation," in *Proc. 2000 Intl. Symp. on Low Power Electronics and Design*, July 2000.

[4] John L. Hennessy and David A. Patterson, Computer Architecture: A Quantitative Approach (2nd edition), Morgan Kaufman, 1996.

[5] N. Azizi, A. Moshovos, F. N. Najm, B. Falsafi, "Asymmetric-cell caches: exploiting bit value biases to reduce leakage power in deep-submicron, high-performance caches," *ECE Computer Group Technical Report TR-01-01-02*, Univ. of Toronto.

[6] E. Seevinck Sr et. al., "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 748–754, Oct. 1987.

[7] V. De, and S. Borkar, "Technology and Design Challenges for Low Power and High Performance Microprocessors," in*Proc. 1999 Intl. Symp. on Low Power Electronics and Design*, 1999.

[8] A. Bhavnagarwala et. al., "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," in *IEEE J. of Solid-State Circuits*, vol. 36, Apr. 2001.