Low-Leakage Asymmetric-Cell SRAM

Navid Azizi, *Student Member, IEEE*, Farid N. Najm, *Fellow, IEEE*, and Andreas Moshovos, *Associate Member, IEEE*

***Abstract*—We introduce a novel family of asymmetric dual-V_t static random access memory cell designs that reduce leakage power in caches while maintaining low access latency. Our designs exploit the strong bias toward 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 extent. A novel sense amplifier, in combination with dummy bitlines, allows for read times to be on par with conventional symmetric cells. With one cell design, leakage is reduced by 7× (in the zero state) with no performance degradation, but with a stability degradation of 6%. Another cell design reduces leakage by 2× (in the zero state) with no performance or stability loss. An alternative cell design reduces leakage by 58× (in the zero state) with a performance degradation of 1% and an area increase of 2.4% and no stability degradation.**

***Index Terms*—Asymmetric static random access memory (SRAM) cell, dual-threshold voltage, leakage current, static memory.**

I. 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-submicrometer (<0.18 μm) processes. Moreover, it is expected to grow by a factor of five every chip generation [1]. For processors, it is estimated that, in 0.10-μm technology, leakage power will account for approximately 50% of the total chip power [2].

Since leakage power is proportional to the number of transistors, and given the projected large memory content of future system-on-chip (SOC) devices, it becomes important to focus on static random access memory (SRAM) structures such as caches, which 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, e.g., [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) [4]–[6]. The mechanisms that identify which cache parts will be unused and that enable leakage reduction incur considerable power and performance overheads that have to be amortized over long periods of time. 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 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 perfect knowledge of which cache parts will be left unused for long periods of time is known beforehand [7].

Traditional SRAM cells are symmetrically composed of transistors with identical leakage and threshold characteristics. Our *asymmetric* SRAM cell designs offer low leakage with little or no impact on latency. In our asymmetric SRAM cells, selected transistors are “weakened” to reduce leakage when the cell is storing a zero (the common case). In this study, 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 0.13-μm 1.2-V CMOS technology. The two best designs offer different performance/leakage characteristics. With one cell design, leakage is reduced by 7× (in the zero state) with no performance degradation. An alternative cell design reduces leakage by 70× (in the zero state) with a sense time degradation of 8.5% (the total read cycle time is degraded by only 2%). Four other cells with improved stability relative to a standard SRAM cell are also designed; one reduces leakage by 2× in the preferred state with no loss in performance or stability, and other reduces leakage by 58× with no loss of stability and a sense time degradation of 3%. By comparison, the use of an all high-V_t cell reduces leakage by approximately 70×, but increases sense times by 43%.

We make the following contributions.

1) We introduce a novel family of asymmetric SRAM cells.
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 of-

Manuscript received August 15, 2002; revised January 7, 2003. This work was supported in part by the Semiconductor Research Corporation under Contract SRC 2001-HJ-901.

The authors are with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON M5S 3G4, Canada.

Digital Object Identifier 10.1109/TVLSI.2003.816139

fers drastic leakage reduction while maintaining high performance and comparable noise margins and stability.

The remainder of this paper is organized as follows. In Section II, we present our asymmetric cell family. In Section III, we present the sense amplifier. In Section IV, we present the simulation results of a SRAM using the different asymmetric cells. Section V includes a discussion on architectural level techniques to leverage the asymmetric nature of the cells. Finally, we offer conclusions in Section VI.

II. Asymmetric SRAM Cells

Early research performed to reduce the power consumption of SRAMs consisted of reducing the dynamic power dissipation through changes in the peripheral circuitry. Due to the large number of transistors contained in SRAM arrays, the static power dissipation within the array has become a large fraction of the total power dissipation. Ideally, a SRAM cell should be fast and should dissipate low leakage power. This is increasingly at odds with the fundamental technology tradeoff between transistor speed and leakage. Conventional high-performance SRAM cells use a symmetric configuration of six transistors with identical threshold voltages. One can reduce leakage by using higher V_t transistors, but unfortunately using an all-high-V_t transistor cell degrades performance by an unacceptable margin. Our asymmetric SRAM cells reduce leakage while maintaining high performance based on the following approach: select a preferred state and weaken only those transistors necessary to drastically reduce leakage when the cell is in that state. These cells exhibit asymmetric leakage and access behavior. Fortunately, their asymmetric access behavior can be exploited to maintain high performance while reducing leakage.

A. Technology

All results reported in this paper are HSPICE simulation results produced at 110 °C using SPICE models of a commercial 0.13-μm 1.2-V CMOS technology. Furthermore, throughout this paper, the following convention will be used. A high V_t (HV) transistor is obtained from the basic 0.13-μm 1.2-V transistor (referred to herein as the regular V_t (RV) transistor) by artificially increasing the V_t by 0.2 V using the HSPICE in-line parameter DELVTO. This value of 0.2 V was chosen because it leads to a difference of approximately 10× between the leakage currents of HV and RV transistors, which is typical of dual-V_t technology.[1] Finally, the sizes of the transistors comprising the basic SRAM cell, which forms the starting point for our design variations below, are part of the technology specification for the 0.13-μm process. As such, these sizes cannot be disclosed.

B. Nine Asymmetric Cells

As shown in Fig. 1, a SRAM cell comprises two inverters, i.e., (P2, N2) and (P1,N1), and two pass transistors, i.e., N3 and N4. In the inactive state, the wordline (WL) is held low so that the two pass transistors are off isolating the cell from BL and BLB. At this stage, the bitlines are also typically charged at $V_{\rm DD}$ (e.g., logic "1"). Cells spend most of their time in the inactive state. In this state, most of the leakage is dissipated by the transistors that are: 1) off and that 2) have a voltage differential across their drain and source. The value stored in the cell (i.e., the cell state) determines which transistors these are. When the cell is storing a "0," as in Fig. 1, the leaky transistors are P1, N4, and N2. If the cell was storing a "1," then transistors P2, N1, and N3 would dissipate leakage power.

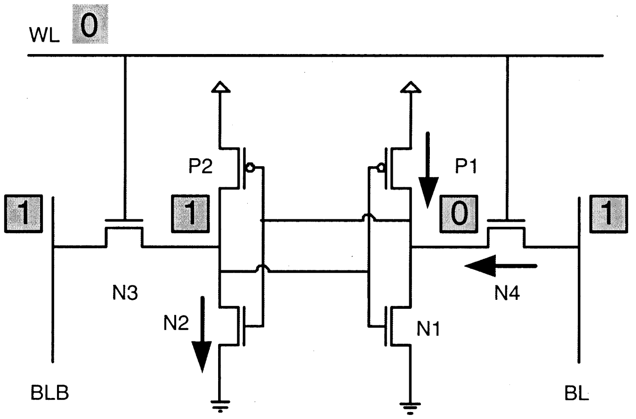


Fig. 1. Transistors that dissipate leakage when a SRAM cell is holding a "0."

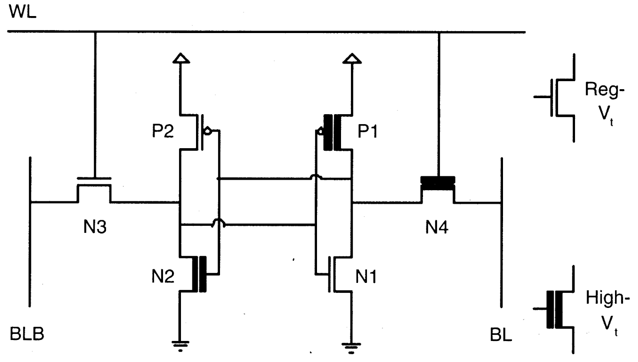


Fig. 2. Basic asymmetric SRAM cell.

A simple technique for reducing leakage power would be to replace all transistors with high-V_t ones, to obtain the HV cell, but this unacceptably degrades the bitlines discharge times by 61.6%.[2] Since ordinary programs exhibit a strong bias in cache-resident bit values [8], 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 HV transistors, as seen in Fig. 2. This basic asymmetric (BA) cell was simulated and it exhibits the same leakage as the RV cell when holding a logic "1," but its leakage is reduced by 70× when holding a logic "0."

The read access time of the BA cell is, however, degraded. Due to N2's and N4's higher threshold voltage, the bitline discharge takes longer. The discharge times for BLB and BL are

[1] If increased length were used to reduce leakage, the transistors' area would need to increase by at least a hundredfold, thus, a dual-V_t approach is preferable.

[2] Discharge time is defined as the time from when the WL is raised to when one of the bitlines reduces to 90% of its precharge value. 90% was chosen due to it being a appropriate differential signal for sense amplifiers to trigger.

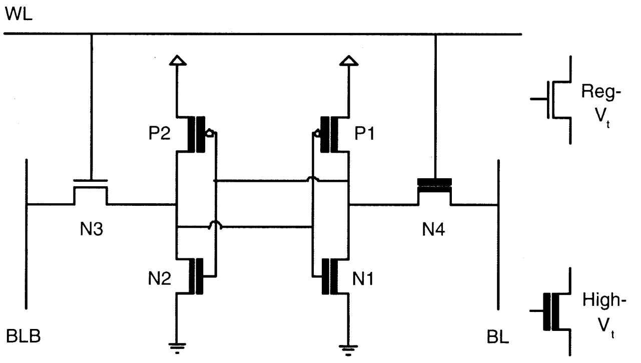


Fig. 3. Leakage improved three cell (the LE cell).

12.2% and 46.4% longer than the discharge time 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 amplifier,[3] as discussed in Section III.

Since p-channel metal–oxide semiconductor (pMOS) transistors have very little effect on the cell's read access time (the role of pulling down the bitlines is played by the two n-channel metal–oxide semiconductor (nMOS) transistors on the side of the cell storing the "0"), a better asymmetric cell consists of the BA cell with P2 also set to high-V_t. This cell, referred to as leakage improved (LI) 2, has the advantage of partially reduced leakage in the high leakage state. When the cell is holding a logic "1," its leakage is reduced by 1.6× relative to the RV cell, and when holding a logic "0," its leakage is reduced by 70×. The discharge times for BLB and BL are 12.2% and 46.4% longer than the discharge times for the RV cell, respectively, the same as the BA cell's discharge times. One further improvement is possible because, due to the sense amplifier (described below), which matches the read time on the slow side of the cell to the fast side, there is no need for N1 to be low-V_t. This leads to the cell in Fig. 3, referred to as LI3. This cell further reduces leakage in the high leakage state so that its leakage relative to the RV cell is reduced by 7× in the "1" state and by 70× in the "0" state. The BL discharge time is now 61.6% longer than the discharge time for the RV cell, but that is of minor importance due to the novel sense amplifier design, as we will see below.

The two asymmetric cells, i.e., LI2 and LI3, take the BA cell and improve its leakage performance while not affecting its read access time. Another design front is to take the BA cell and try to improve its read access time, while keeping some of the leakage benefits found in the BA cell.

To eliminate the speed penalty incurred in the BA cell due to both pull-down paths having one high-V_t transistor, both N2 and N3 are made low-V_t. This cell, i.e., speed improved (SI)1, now has discharge times for BLB and BL, which are 0% and 46.7%, respectively, longer than the RV cell. Thus, one side of the cell is just as fast as the RV cell. However, this cell suffers from higher leakage than the BA cell, with a leakage reduction of 2× relative to RV when holding a "0," and no leakage reduction when holding a "1."

The same transformations performed on BA to improve its leakage performance can also be performed on the SI 1 cell. First, P2 is made high-V_t, and then N1 is also made high-V_t. These two new cells are named SI2 and SI3, respectively. SI3 is shown in Fig. 4. SI2 has leakage reductions of 2× and 1.6× when storing a "0" and "1," respectively, while SI3 has leakage reductions of 2× and 7×.[4] These two cells have no read access time degradation compared to the RV cell along BLB, but have a 46.5% and 61.6% degradation along BL, respectively. Once again, the degradation along BL is of minor importance due to the novel sense amplifier.

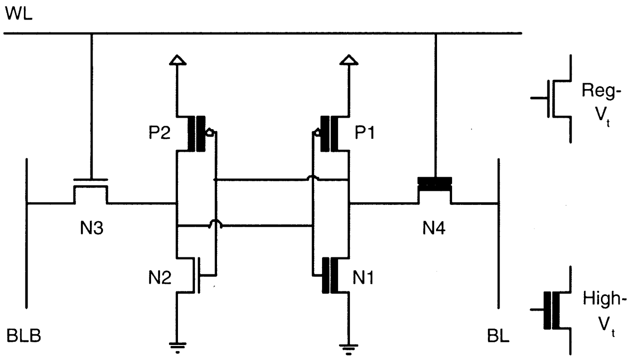


Fig. 4. Speed improved three cell (the SE cell).

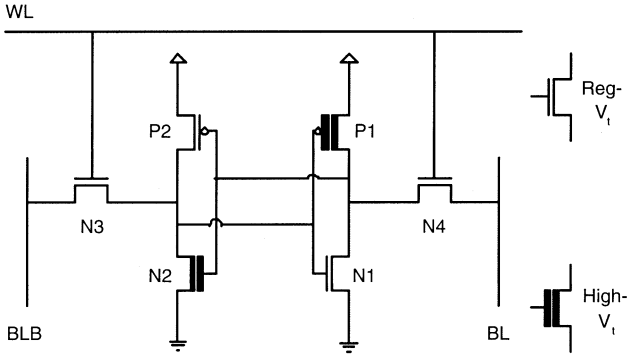


Fig. 5. Special precharge cell.

One would like to get the very low leakage of the LI2 and LI3 cells and a very small read access delay. A final asymmetric cell can meet these objectives, but it requires a different read operation. In the steady state, instead of keeping BL precharged to $V_{\rm DD}$, keep it at ground. Now, N4 can be kept low-V_t for the preferred "0" state. This special precharge (SP) cell is shown in Fig. 5. Before a read, BL may have to be raised to "1," or a new sensing scheme may have to be developed, which may be power hungry. This cell requires changes to the peripheral circuits of the SRAM array, and further work is required to develop this concept. Nevertheless, the results for this cell are presented for completeness: leakage is reduced by 83.3× in the "0" state, while the "1" state shows no leakage reduction. The SP cell shows the possible leakage reduction when only two transistors are dissipating leakage and both are high-V_t. Furthermore, bitline discharge times are degraded by 12.2% and 0%.

[3] The new sense amplifier does not improve sense times for symmetrical SRAM cells.

[4] Note that SI3 reverses the preferred leakage state to the state when the cell is holding a "1." All further references to this cell will have the "1" state as the preferred state so that the cell language remains in conformity with all other cells, but it should be noted that, in practice, the cell bitlines can be flipped to allow for "0" to be the preferred state without affecting any of the performance or stability results shown here.

TABLE I
SUMMARY OF LEAKAGE REDUCTION FOR ALL ASYMMETRIC CELLS RELATIVE TO THE RV CELL

Asymmetric	Storing a '0'			Storing a '1'		
Cell	Leakage Reduction	# of leaky RV	# of leaky HV	Leakage Reduction	# of leaky RV	# of leaky HV
BA	69.50X	0	3	1.00X	3	0
LI2	69.50X	0	3	1.61X	2	1
LI3	69.50X	0	3	6.96X	1	2
SI1	2.04X	1	2	1.00X	3	0
SI2	2.04X	1	2	1.61X	2	1
SI3	2.04X	1	2	6.96X	1	2
SP	83.33X	0	2	1.00X	3	0
HV	69.50X	0	3	69.50X	0	3

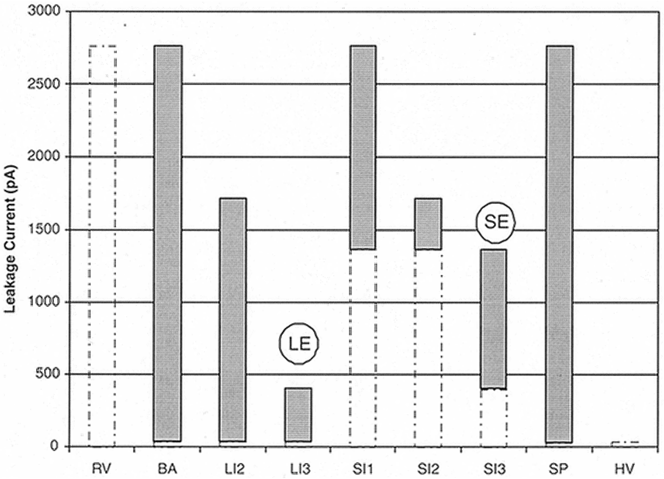


Fig. 6. Graphical representation of asymmetric leakage characteristics of all cells.

TABLE II
SUMMARY OF BITLINE DISCHARGE TIMES FOR ALL ASYMMETRIC CELLS RELATIVE TO THE RV CELL

Asymmetric Cell	% Increase of BLB discharge time	% Increase of BL discharge time
BA	12.25%	46.73%
LI2	12.25%	46.50%
LI3	12.09%	61.64%
SI1	0.00%	46.73%
SI2	0.00%	46.51%
SI3	0.00%	61.69%
SP	12.26%	0.00%
HV	61.64%	61.64%

A summary of the leakage reduction while holding a "0" and "1" can be seen in Table I. Fig. 6 shows the asymmetry between the leakage when each cell is holding a "0" or a "1." The bitline discharge times are summarized in Table II, which shows the distinction between the LI and SI cells. All LI cells show a near 12% increase in bitline discharge times, while the SI cells show no increase in bitline discharge times. Furthermore, Fig. 6 shows that the BA and LI2 cells show no advantage to LI3 since the LI3 cell has the same speed performance as the BA and LI2 cells, but with better leakage performance. The LI3 cell will be referred to henceforth as the leakage enhanced (LE) cell. Also, SI3 is clearly the best design from within the SI cells. The SI3 cell will be referred to henceforth as the speed enhanced (SE) cell.

TABLE III
SUMMARY OF WRITE TIMES FOR ALL ASYMMETRIC CELLS

Asymmetric Cell	Percent Increase over RV cell
BA	51.9%
LI2	41.2%
LI3	36.0%
SI1	56.1%
SI2	45.5%
SI3	40.2%
SP	11.5%
HV	69.4%

Until now, only the bitline discharge times of the different cells have been compared, and write times have been ignored. The write times of the cells are less important because stronger write drivers can be designed to drive the bitlines, and write drivers are a small portion of the total SRAM. The write times of the asymmetric cells all lie within the write times of the RV and HV cells. The precise numbers can be seen in Table III. The LE and SE cells show the smallest increase in their respective groupings.

Since the LE and SE are the two best designs from the two sets of asymmetric cells, only these two cells, and variations on them will be further discussed in the following.

C. Stability

Another major consideration with the cell design is its stability. There are two interrelated issues: read stability and noise margins [3], [9]. Intuitively, read stability indicates how likely it is to invert the cell's stored value when accessing it, and was computed as the ratio of $I_{\rm trip}/I_{\rm read}$, where $I_{\rm trip}$ is the current through the pull-down NMOS when the state of the cell is being reversed by injecting an external current $I_{\rm test}$ and where $I_{\rm read}$ is the maximum current through the pass transistor during a 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 [10]. In our case, the stability of all cells was measured by simulation (HSPICE) via both the SNM and $I_{\rm trip}/I_{\rm read}$ methods. Under both stability tests, the stability was first measured under nominal conditions, assuming no process variations.

To measure stability under process variations, two sets of tests were then performed. First, the SNM and $I_{\rm trip}/I_{\rm read}$ tests were performed on 59 049 combinations of different V_t and length variations for all six transistors in the cell. The combinations included modifying by $\{-3\sigma, 0, +3\sigma\}$ the nMOS transistors' V_t and length values and the pMOS transistors' V_t value, thus giving $3^8 = 59\,049$ combinations. The worst case value for various cells was found, and compared to the worst case value obtained for the RV cell.

Second, Monte Carlo analysis was performed to obtain a distribution for the SNM and $I_{\rm trip}/I_{\rm read}$. For each cell, 500

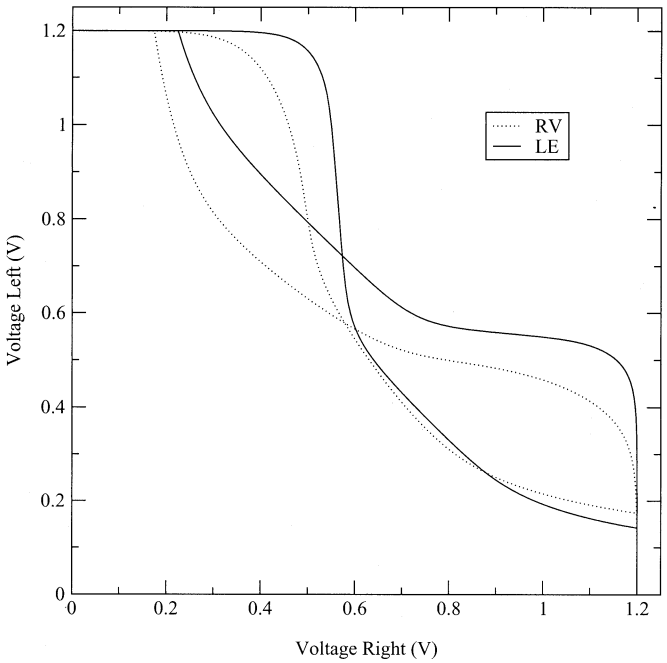


Fig. 7. SNM of the LE cell.

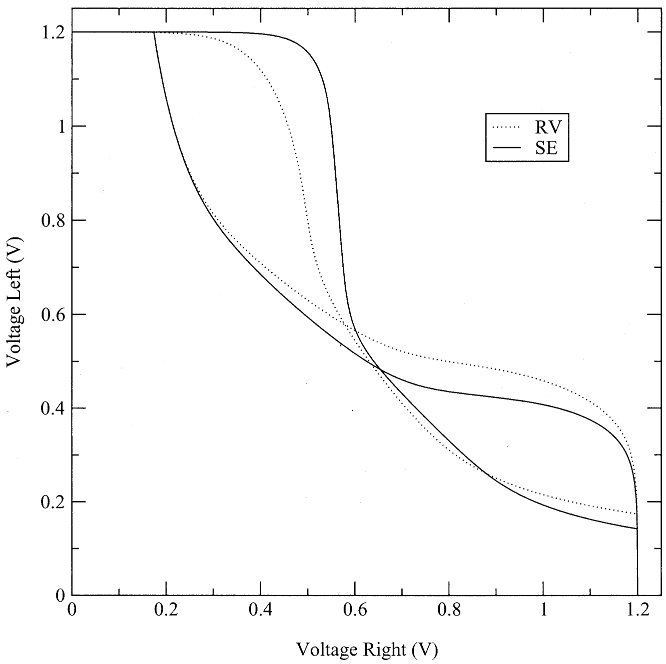


Fig. 8. SNM of the SE cell.

scenarios for V_t and length were randomly generated, consistent with their joint distributions, and simulated. The mean of the distribution was estimated using the unbiased estimator in (1), and the variance was estimated by using the unbiased estimator in (2). Furthermore, the normal scores method was used to graphically determine the distribution type [11]. Given the distribution type, mean, and variance, the probability of the failure for various cells was then computed

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \tag{2}$$

1) SNM: The SNM of the LE and SE cells were computed through simulation. The SNM of the RV cell was also computed to be used as a reference. Under nominal conditions, the SNM of the LE and SE cells were 0.246 V and 0.221 V, respectively, while the SNM of the RV cell was 0.250 V. Thus, the LE and SE show a decrease in SNM of 1.6% and 11.7%. One would expect that by using HV transistors in the design, the SNM of the cells would increase, but the asymmetry of the cells skews the lobes of the butterfly curve and decreases the SNM, as will be explained below.

First, let us examine the SNM of the cells when the wordline is not active. During this state, the SRAM cell is not as vulnerable as when it is being read, but a study of this case helps to understand the decrease in the SNM when the cell is being read. When the WL is off, the only transistors that affect the SNM are the four transistors comprising the back-to-back inverters.

Since the four internal transistors of the LE cell are all high-V_t, the cell has equal low and high noise margins of 0.685 V, a 22.6% increase over the standby SNM of the RV cell, 0.559 V. However, when the SNM of the cell is being measured during a read, as seen in Fig. 7, the cell has high SNM in one state, 0.363 V, and low SNM in the other, 0.246 V.

TABLE IV
WORST CASE SNM

Cell	Worst-Case SNM(V)	% Increase over RV cell
RV	0.091	–
LE	0.088	-3.73%
SE	0.065	-28.79%

The asymmetry in the LE butterfly curve is due to the mismatch between the strength of the pass–gate (N3) and pull-down (N2) transistors. During a read, the N3 pass transistor, due to it being low-V_t, has a higher conductivity than N2 and raises the voltage at the storage node to a higher voltage than if the two NMOS were of equal strength.

For the SE cell, the internal inverter pair are not identical. Thus, the standby (i.e., with the wordline off) SNM of the cell has asymmetric lobes with noise margins of 0.535 and 0.727 V, in the worst case, a 4.2% decrease in noise margin compared to the RV cell. The source of this mismatch is the V_t difference between N1 and N2, which causes one of the transfer characteristics to commence its transition in the SNM plot from "0" to "1" later than normal. During a read, the mismatch between the size of the lobes becomes exaggerated because it is as if a constant is subtracted from the noise margin on each side of the cell since each side of the cell has equal strength pass transistors and pull-down transistors. While being read, the SE cell has low and high noise margins of 0.222 and 0.365 V, respectively. The SNM plot of the SE cell is shown in Fig. 8.

As explained in Section II-C, process variations were analyzed by two methods. First, by sweeping over 59 049 cases, the worst-case SNM was found for each cell and is summarized in Table IV.

The asymmetric cells stability performance degrades compared to that of the RV cell. Since process variations induce an asymmetry in the butterfly curve, the original asymmetry inherent in the butterfly curves for the LE and SE allows one

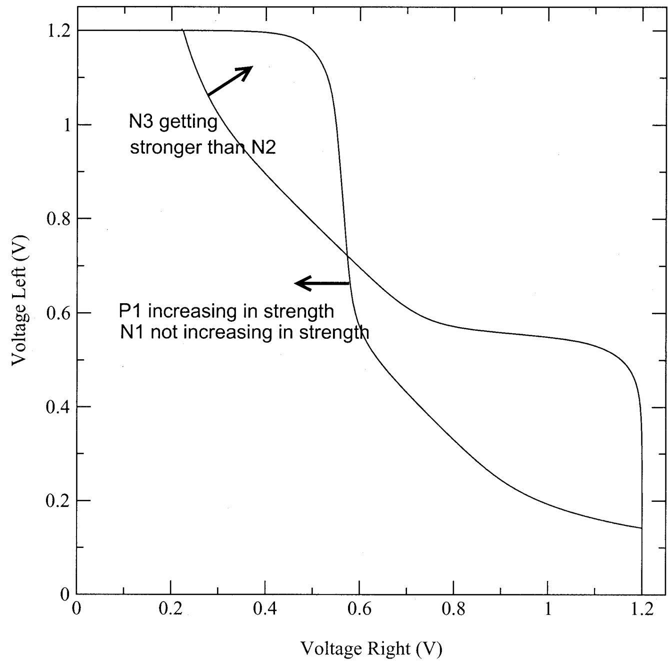


Fig. 9. SNM of LE cell pinching off.

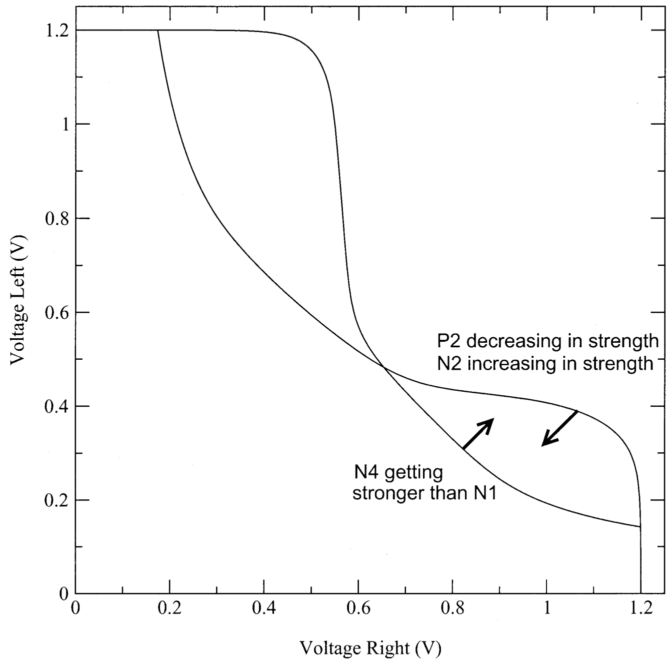


Fig. 10. SNM of SE cell pinching off.

lobe of the butterfly curve to become pinched off even further and lose stability. For the LE cell, the butterfly curve becomes pinched off when N3 becomes stronger than N2 and P1 increases in strength, while N1 does not. Fig. 9 shows the effect graphically. The worst case for the SE cell occurs at a different process corner. The butterfly curve becomes pinched off when P2 decreases in strength and N2 increases in strength, and N4 gets stronger than N1, as shown in Fig. 10.

TABLE V
MEAN AND σ DURING MONTE CARLO ANALYSIS FOR SNM

Cell	Mean(V)	Standard Deviation(V)
RV	0.231	0.0182
LE	0.247	0.0244
SE	0.218	0.0249

TABLE VI
NOMINAL ${}_{\rm trip}/I_{\rm read}$

Cell	Nominal(V)	% Increase
RV	2.26	–
LE	2.10	-7.31%
SE	2.60	14.86%

TABLE VII
WORST CASE $I_{\rm trip}/I_{\rm read}$

Cell	Worst-Case(V)	% Increase
RV	1.71	–
LE	1.58	-7.81%
SE	1.82	6.12%

Monte Carlo analysis was also performed on the RV, LE, and SE cells. Table V summarizes the mean and standard deviation of the SNM. Furthermore, the normal scores method showed that the distributions for all cells were Gaussian. Due to their very small standard deviation, the SNM of all cells remains very close to their respective mean. Thus, the mean of the SNM becomes a very important measure, and is a better reflection of the stability than the nominal or worst case SNM. Using the mean as a measure of stability, the LE has a 7% increase in SNM and the SE has a 5.8% decrease.

2) $I_{\rm trip}/I_{\rm read}$: Using the SNM as a measure of stability showed that the LE cell was comparable to the RV cell while the SE cell showed a marginal decrease in stability. When $I_{\rm trip}/I_{\rm read}$ is computed by simulation, it is seen that the SE outperforms the RV cell and the LE suffers. Table VI shows the results.

The LE cell has a lower $I_{\rm trip}/I_{\rm read}$ value due to the V_t mismatch between the pass transistor and pull-down transistor on one side of the cell. The $I_{\rm trip}$ values from both sides of the cell show a drop compared to the $I_{\rm trip}$ value from the RV cell due to both pull-down transistors becoming high-V_t. However, with N3 remaining low-V_t, $I_{\rm read}$ on the fast side of the cell does not suffer the same drop, and $I_{\rm trip}/I_{\rm read}$ falls compared to that of the RV cell.

The SE, due to it having the same strength pull-down and pass transistors on each side of the cell, does not experience the same problem as the LE cell. On the slow side of the cell, both $I_{\rm trip}$ and $I_{\rm read}$ fall compared to the RV cell, but $I_{\rm read}$ falls by a larger amount, thus increasing the $I_{\rm trip}/I_{\rm read}$. On the fast side of the cell, $I_{\rm read}$ does not change compared to the RV cell, but $I_{\rm trip}$ increases slightly. In the RV cell, the reduction in voltage (due to leakage) at the stored "1" node degrades the current sinking capacity of the pull-down NMOS. In the SE cell, because of the high-V_t transistors on the "1" side of the cell, there is no degradation in the current sinking capacity of the pull-down transistor and, thus, $I_{\rm trip}$ increases leading to a larger $I_{\rm trip}/I_{\rm read}$.

A total of 59 049 different corner cases of process variations were simulated and the worst case $I_{\rm trip}/I_{\rm read}$ was noted in each cell, and is summarized in Table VII. The LE and RV cells achieve their worst case $I_{\rm trip}/I_{\rm read}$ for the same process corner:

TABLE VIII
MONTE CARLO ANALYSIS FOR $_{\mathrm{trip}}/I_{\mathrm{read}}$

Cell	Mean	σ
RV	2.20	0.0765
LE	2.10	0.0898
SE	2.53	0.1100

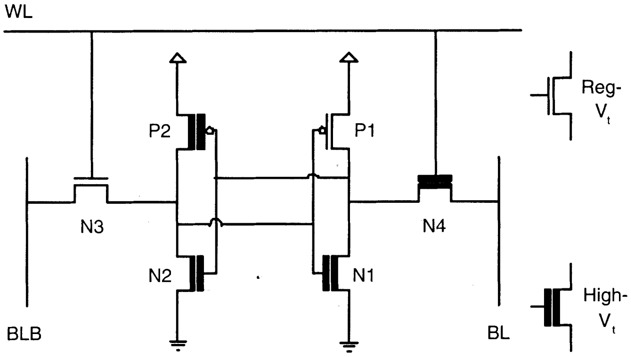


Fig. 11. SLE cell.

when the difference in strength between N2 and N3 is amplified with N2 becoming weaker, and N3 becoming stronger. The SE cell, however, suffers its worst case $I_{\mathrm{trip}}/I_{\mathrm{read}}$ when N4 becomes stronger than N1.

The Monte Carlo analysis show that $I_{\mathrm{trip}}/I_{\mathrm{read}}$ is also Gaussian from the linear plots obtained from the normal scores method. Table VIII shows the mean and standard deviation of the three cells. Notice, once again, the standard deviation is very small and, thus, most cells will be very near the mean where the LE shows a 4.35% decrease and the SE cell shows a 14.84% increase in $I_{\mathrm{trip}}/I_{\mathrm{read}}$.

D. Improved-Stability Cells Through Threshold Voltage

As seen in the previous section, the SE and LE cells have either a lower stability in the SNM or $I_{\mathrm{trip}}/I_{\mathrm{read}}$ tests. In many cases, the stability of the cell is a critical factor to obtain a desired yield and to lower the cost of the chip. In that regard, two derivative cells, one from the LE cell and one from the SE, have been developed that improve upon their SNM, but do not decrease the leakage as much as the SE and LE cells. The two new cells are named stability-leakage enhanced (SLE) and stability-speed enhanced (SSE).[5]

One way to improve the SNM of the cells under process variations is to try to make the size of the lobes of the butterfly curve symmetric. For the LE cell, the lobes can be made more symmetric by making N2 low-V_t, but this new cell would just be the SE cell. Another option is to make P1 low-V_t. This change, seen in Fig. 11, has the opposite effect of the lower arrow in Fig. 9, and makes the lobes of the butterfly curve more symmetric. The SNMs are now 0.360 and 0.283 V instead of 0.363 and 0.246 V. To make the SE cell's SNM plot more symmetric, P2 can be made low-$V_{(t)}$ to have the opposite effect as the top arrow in Fig. 10. By doing this, the SNM plot, shown in Fig. 14, has SNMs of 0.256 and 0.362 V instead of 0.222 and 0.366 V. The SSE cell is shown in Fig. 12. The change in SNM can be seen in Figs. 13 and 14.

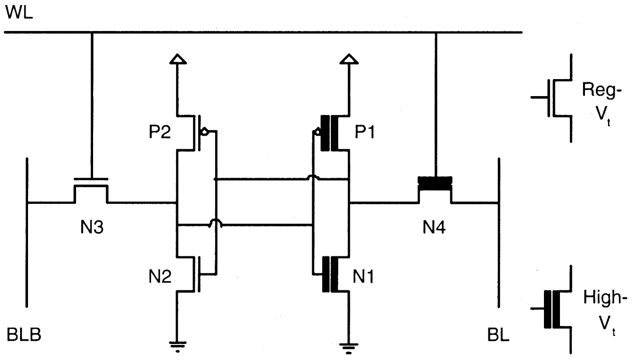


Fig. 12. SSE cell.

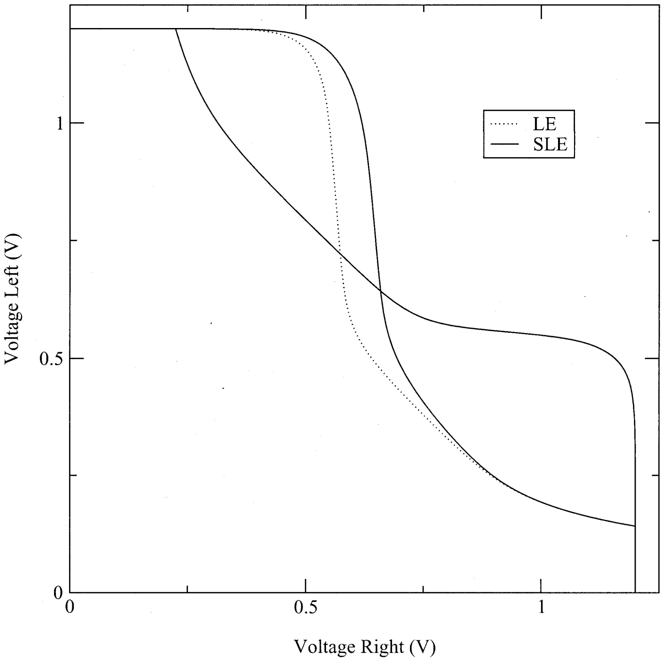


Fig. 13. SNM for SLE cell.

For these stability-improved cells, all the previous tests for leakage, performance, and stability can be performed to compare them to the cells they were derived from, as well as to, the RV cell.

1) Leakage: The leakage performance of the stability-improved cells falls off, as is expected due to one transistor in the LE and SE being re-converted to a low-V_t transistor. For the SLE cell, the leakage reduction when holding a "1" remains unchanged at a 6.96× reduction relative to RV, but the leakage reduction when holding a "0" changes from 69.5× to 2.5×. For the SSE cell, when it is holding a "0," the leakage reduction stays at 2.04×, but when it is holding a "1," the leakage reduction changes from 6.96× to 1.91×.

2) Performance: Since the PMOS transistors do not play a large role in discharging the bitlines, it would be expected that the discharge time for the stability-improved cells to be very close to the cells they derived from. Through simulation, it is seen that the discharge times along BL and BLB remain almost

[5] The $I_{\mathrm{trip}}/I_{\mathrm{read}}$ of the cells is not improved since the only method to improve the LE cell's value considerably is to make the pull-down NMOS on the fast side of the cell low-V_t, which would make it the same as the SE cell. The SE cell already has a better $I_{\mathrm{trip}}/I_{\mathrm{read}}$ than RV under nominal, worst case, and mean conditions.

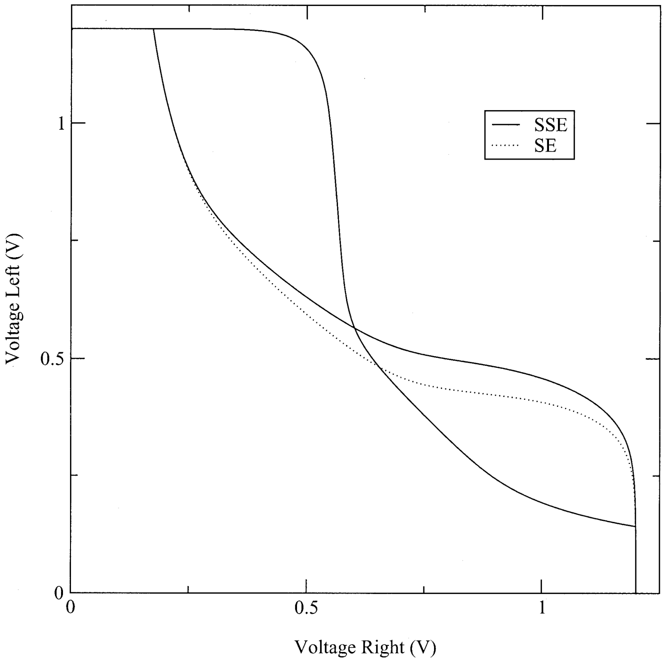


Fig. 14. SNM for SSE cell.

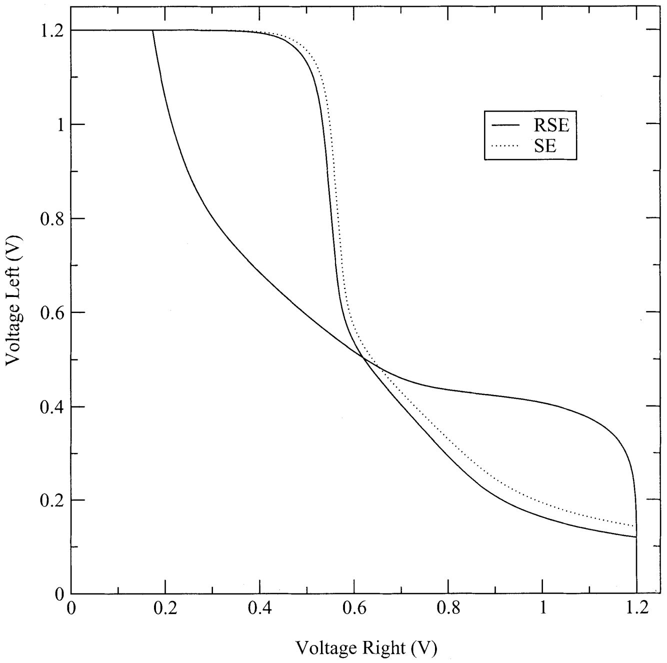


Fig. 15. SNM for RSE cell.

TABLE IX
%IMPROVEMENT OVER RV CELL FOR IMPROVED-STABILITY CELLS UNDER SNM TEST

Cell	Worst-Case SNM % improvement	Mean SNM % improvement
SLE	35.41%	22.63%
SSE	7.80%	9.29%

TABLE X
%IMPROVEMENT OVER RV CELL FOR INCREASED-STABILITY CELLS UNDER $I_{\mathrm{trip}}/I_{\mathrm{read}}$ TEST

Cell	Worst-Case $I_{\mathrm{trip}}/I_{\mathrm{read}}$ % improvement	Mean $I_{\mathrm{trip}}/I_{\mathrm{read}}$ % improvement
SLE	-10.50%	-6.78%
SSE	4.06%	13.43%

constant. As for the write times, SLEs write time decreases to a 33.15% increase over RVs write time from LEs 35.95% increase. The SSE write time jumps to a 49.22% increase over RVs write times.

3) Stability: The stability analysis has also been performed on the derivative cells for both the SNM and $I_{\mathrm{trip}}/I_{\mathrm{read}}$. Both derivative cells perform better than the RV cell in the worst case and under Monte Carlo analysis. The results are shown in Table IX.

Under the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ method, there is very little change because $I_{\mathrm{trip}}/I_{\mathrm{read}}$ depends strongly on the NMOS transistors, which have not been changed, but the stability-improved cells perform slightly worse than the cells from which they were derived. The results are summarized in Table X.

E. Improved-Stability Cells Through Transistor Sizing

From the previous section, it can be seen that when stability is recovered through a change in threshold voltage of the PMOS transistors, a large portion of the leakage benefits of the asymmetric cells are lost. Furthermore, the low $I_{\mathrm{trip}}/I_{\mathrm{read}}$ of the LE cell could not be improved by threshold voltage assignment.

Another way of improving stability is to resize some of the transistors to reclaim the conductance lost due to the high-V_t assignment. This change does not have a large effect on the leakage characteristics because leakage increases exponentially with reduced threshold voltages, but increases only linearly with transistor size. Moreover, the low $I_{\mathrm{trip}}/I_{\mathrm{read}}$ of the LE cell can be improved by transistor resizing.

The lobes of the SNM plot for the SE cell can be made more symmetric by making N1 wider. In our case, we increased the width of this transistor by 26%, leading to a new cell, which we refer to as resized speed enhanced (RSE). The SNM for the RSE cell is comparable to that of the RV cell and the change in N1's size leads to an increase of only 2.9% in cell area. The change in the SNM plot can be seen in Fig. 15, where the margins are now 0.253 and 0.347 V instead of 0.222 and 0.366 V. The RSE cell's nominal value for $I_{\mathrm{trip}}/I_{\mathrm{read}}$ does not change much compared to the nominal value for the SE cell. On the slow side of the cell, which had the higher $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value for the SE cell, the increase in N1's size allows for I_{trip} to become larger and increases the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value. The fast side of the cell, however, which has the limiting $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value, has a reduced I_{trip} that reduces the final value of $I_{\mathrm{trip}}/I_{\mathrm{read}}$ to 2.53. The reduction in I_{trip} is due to the "1" storage node having a slightly lower voltage due to the increased leakage through N1. Nevertheless, the RSE cell's $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value is still 11.8% better than that of the RV cell.

For the LE cell, increasing the width of N2 allows the conductance of N2 to approach that of N3, which leads to an increase in I_{trip}, thus increasing $I_{\mathrm{trip}}/I_{\mathrm{read}}$. By increasing N2's width by 22% (leading to an only 2.4% increase in cell area), the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value of the new RLE cell was made to be 2.28, which is comparable to the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value of 2.26 of the RV

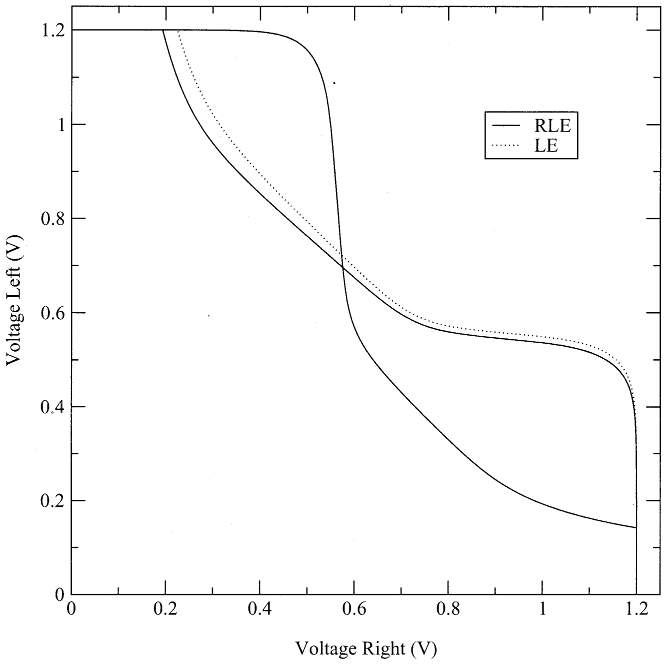


Fig. 16. SNM for RLE cell.

cell. The increase in N2's width also increases the SNM of the RLE cell where the margins are now 0.349 and 0.280 V instead of 0.363 and 0.246 V. The SNM plot for this cell can be seen in Fig. 16.

For these resized cells, all the previous tests for leakage, performance, and stability were performed to compare them to the cells they were derived from, as well as to, the RV cell.

1) Leakage: As expected, the leakage performance of the resized cells is better than that of the SLE and SSE cells. For the RLE cell, the leakage reduction when holding a "1" remains unchanged at a 6.96× reduction relative to RV, but the leakage reduction when holding a "0" only slightly reduces from 69.5× to 57.9×. (The SLE cell's leakage reduction when holding a "0" was only 2.5×). When the RSE cell is holding a "0," the leakage reduction stays at 2.04× relative to RV, and when it is holding a "1," the leakage reduction only changes from 6.96× to 6.79×. This change is also minimal when compared to the SSEs leakage reduction of 1.91×X.

2) Performance: Due to the increased size of the pull-down NMOS transistors, the resized cells have the potential of improving the read-access time of the cell. For the RLE cell, the discharge time along BLB remains at a 61.1% increase over the RV cell's BLB discharge time, but the BL discharge time is now only 3.7% longer than the RV cell's discharge time. As noted previously, only the BL discharge time is important due to the timed read based on the new sense amplifier. For the RSE cell, the discharge time along the fast side of the cell, BL, does not change, but the discharge time along BLB is reduced from the SE cell's 61.7% increase over RV to a 49.2% increase over RV. This extra performance along BLB plays no important role in the cell's performance. As for the write times, RLEs write time increases to a 39% increase over RVs write time from LEs 35.95% increase. The RSE write time jumps to a 45% increase over RVs write times.

TABLE XI
PERCENTAGE IMPROVEMENT OVER RV CELL FOR RESIZED CELLS UNDER SNM TEST

Cell	Worst-Case SNM % improvement	Mean SNM % improvement
RLE	36.65%	21.4%
RSE	9.89%	7.9%

TABLE XII
PERCENTAGE IMPROVEMENT OVER RV CELL FOR RESIZED CELLS UNDER $I_{\mathrm{trip}}/I_{\mathrm{read}}$ TEST

Cell	Worst-Case $I_{\mathrm{trip}}/I_{\mathrm{read}}$ % improvement	Mean $I_{\mathrm{trip}}/I_{\mathrm{read}}$ % improvement
RLE	2.51%	5.04%
RSE	11.6%	15.37%

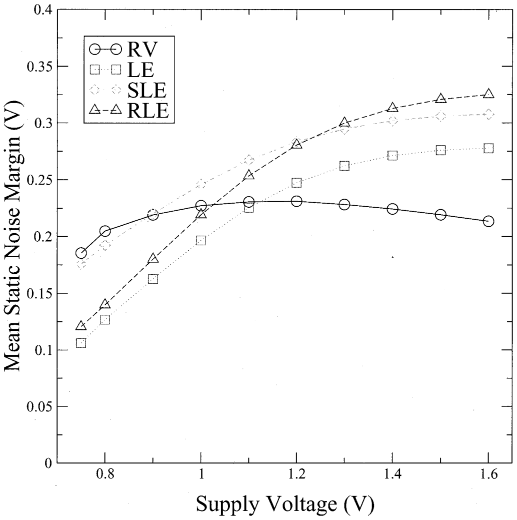


Fig. 17. Mean SNM under different supply voltage for LE derived cells.

3) Stability: The stability analysis has also been performed on the resized cells for both the SNM and $I_{\mathrm{trip}}/I_{\mathrm{read}}$ tests. Both resized cells perform better than the RV cell in the worst case and under Monte Carlo analysis for the SNM. These results can be seen in Table XI. Under the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ test, the RLE cell now performs better than RV both in the worst case and on average. The increase in N1's size accomplishes the higher $I_{\mathrm{trip}}/I_{\mathrm{read}}$. The RSE cell's $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value also increases slightly under all tests, even surpassing the SE cell's $I_{\mathrm{trip}}/I_{\mathrm{read}}$ value in the worst case. With a larger pull-down transistor, the process variations do not have as much an effect on the RSE cell's stability. These results are shown in Table XII.

F. Stability at Different Supply Voltages

Another figure of merit for the different cells is their stability under different supply voltages. For the technology being used, the nominal supply voltage is 1.2 V. Monte Carlo analysis has been performed for the RV, LE, SLE, RLE, SE, SSE, and RSE cells for supply voltages ranging from 0.75 to 1.6 V, for which the mean SNM is shown in Figs. 17 and 18.

From the plot it can be seen that for voltages above 1.2 V, LE, SLE, and RLE improve their SNM advantage over the RV cell. With a higher V_{GS}, the difference in conductance between the pass–gate (N3) and pull-down (N2) transistors, which was the

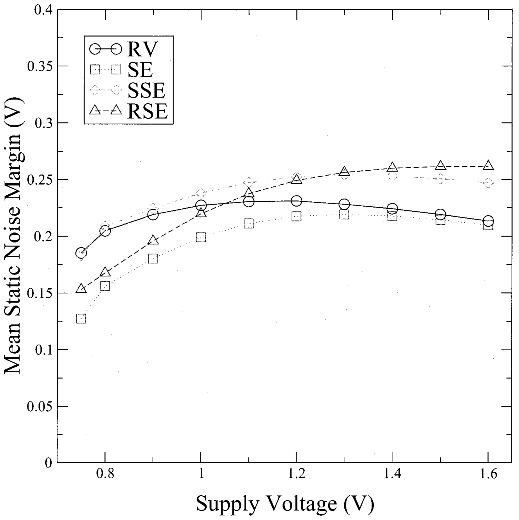


Fig. 18. Mean SNM under different supply voltage for SE derived cells.

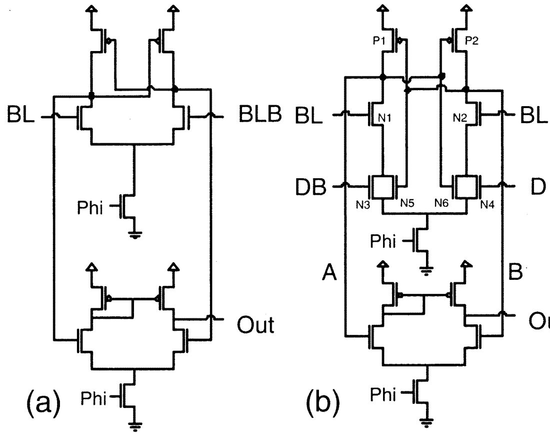


Fig. 19. (a) Conventional sense amplifier. (b) New sense amplifier.

root cause of the low stability at 1.2 V, diminishes. At higher voltages, the SNM of the SE and SSE cells starts to diminish just as the SNM of the RV, but at a lower rate. The SNM of the RSE cell levels off at higher voltages.

With lower supply voltages, the SNM of the asymmetric cells starts to suffer. For the LE, SLE, and RLE cells, the SNM decreases rapidly, but SLE's SNM remains comparable to that of RV, while RLE's SNM becomes comparable to that of LEs. This decrease in stability is caused by the difference in conductance between RV and HV transistors at low V_{GS}'s. Furthermore, at low V_{GS}, the extra conductance of the larger transistor in the RLE cell does not have a large effect since the transistor is not fully on. The SNM of SE, SSE, and RSE also decreases, but not as fast as that of LE. Again, this decrease in SNM is due to the difference in conductance at low V_{GS}'s.

Based on [12], the voltage regulator and power distribution network in microprocessors must maintain the supply voltage to within ±5% of nominal. Therefore, the reduced stability at low voltages for the asymmetric cells might not be a big concern, except perhaps during any chip testing that may need to be performed at low voltage.

The same tests were performed for the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ method with the result that the curves for all cells are much better behaved. The SE and SSE cells have a near 24% advantage over the RV cell at 0.75 V and a 8% advantage at 1.65 V. The LE and SLE cells have approximately a 16% decrease in $I_{\mathrm{trip}}/I_{\mathrm{read}}$ at 0.75 V and are comparable at 1.65 V to the RV cell. The resized cells behave slightly differently, with the RSE cell having an 11.7% improvement at 1.65 V and a 32.2% improvement at 0.75 V. The RLE cell has a 9.6% improvement at 1.65 V and a 4% decrease at 0.75 V.

III. Sense Amplifier

A conventional sense amplifier, shown in Fig. 19(a), is not suitable in our design due to the slow access time when the cell is storing a "0." To obtain fast read times regardless of the data value, a new sense amplifier has been designed and is shown in Fig. 19(b). The design of this sense amplifier is based loosely on MOS current-mode logic (MCML) ideas presented in [13]. Compared to the conventional sense amp, the new sense amplifier has four additional transistors and an area increase of roughly 0.229 μm^2 or 14.4%.

In addition to BL and BLB, the sense amp has two new inputs, i.e., D and DB. These are connected to a dummy column of cells that store "1" at all times, but which are otherwise exactly identical to all other cells in the array. This dummy column extends the full length of the SRAM array so that, during every read operation, one of the dummy cells will have its workline asserted. Since the dummy cells always store a "1," they are always fast on the discharge (as fast as the fast side of any other cell), and they are used to provide something like a timer signal. This is achieved by connecting the dummy bitlines to the sense amp in a reverse way (D connected to the right side, where BLB is connected, and DB connected to the left side, where BL is connected), so that D and DB trigger a fast read of a "0" result when the cell being read has "0" content.

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. While initiating the sensing for a "0" is delayed, the combined effect of the dummy cell and the slow side of the asymmetric cell makes the sensing process itself much faster once initiated so that the end result becomes available at about the same time as it would when sensing a "1."

The detailed operation of the sense amplifier is as follows. Initially, the bitlines are precharged and all four amplifier inputs rise to V_{DD}. During this phase the sense amplifier is being reset and nodes A and B are reset to an intermediate value. During a read operation, either BLB will discharge (cell has a "1," fast discharge from the fast side) or BL will discharge (cell has a "0," slow discharge from the slow side). Furthermore, the signal DB, which is on the fast side of the dummy cell, will be discharged since the dummy cells permanently hold a logic "1." The D signal will always remain near V_{DD}.[6] If BLB is being discharged (a logic "1" is being sensed), then the differential pair composed

[6] The D signal should not be tied to V_{DD} since common-mode noise sources, such as clock feedthrough, need to appear on D to achieve high noise immunity.

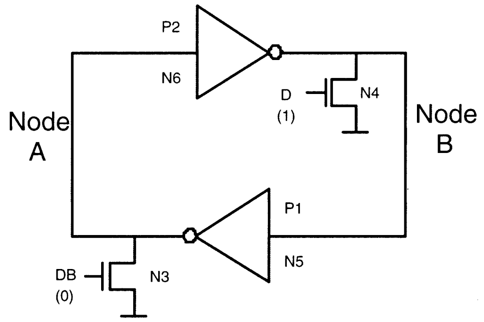


Fig. 20. Approximate sense amplifier circuits when reading a “0.”

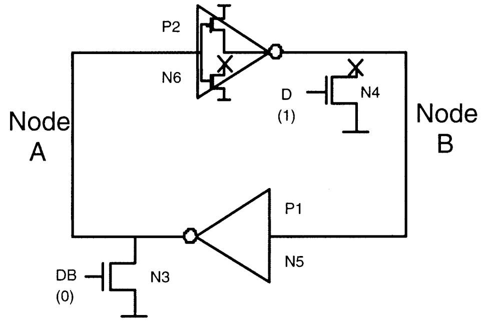


Fig. 21. Approximate sense amplifier circuits when reading a “1.”

of N1 and N2 causes increased current to pass through the left branch, thus increasing the voltage at node B and decreasing the voltage at node A. Through the positive feedback loop of P1, P2, N5, and N6, the rate of change for nodes A and B are increased to achieve quick sensing. When BL is being discharged (a logic “0” is being sensed), 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 N3 and N4, initiate the sensing of a logic “0.” Through the combined effect of the DB bitline being discharged and BL being discharged, albeit at a slower rate, approximately symmetric sense times are achieved.

The operation of the sense amplifier can also be understood from the following two circuit simplifications. First, consider the situation when a “0” is being read and the slow side of the cell is discharging. For simplicity, assume that the fast signals from the dummy column are much faster than the slow signal from the read column. In this case, the gates of N1 and N2 are at $V_{\rm DD}$ due to precharging, and the transistors are on. Since the transistors are conducting, they can be replaced by short circuits, and the sense amplifier simplifies to the circuit shown in Fig. 20. When DB has not been discharged, nodes A and B are being pulled low by N3 and N4. Once DB starts to discharge, N3 turns off and there is nothing keeping node A low and the back-to-back inverters latch a “0.”

When the sense amplifier is reading a “1,” the normal and dummy bitlines are equally fast, and N1 and N2 cannot be simplified away. However, the circuit can still be simplified to the circuit shown in Fig. 21. In Fig. 21, node B has its pull-down path blocked due to BLB discharging. Node B is then pulled high by P2 since N2 has cut off N4’s and N6’s ability to pull the node low. The circuit is then forced to latch the correct value.

For this sensing scheme to achieve reliable results, it must allow for adequate time for BLB to discharge before initiating a logic “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. Furthermore, since N3 and N4 are in parallel with N5 and N6, they cannot constrain the current along the branch as much as N1 and N2 can. We have also performed sensitivity analysis of this sense amplifier, and it performs on par with the conventional sense amplifier.

To limit the sense power, the sense amplifiers are clocked, as in [14]–[16]. The sense clock turns on the amplifiers and sets them up in their high gain region before the sensing occurs. To improve yield and ensure low-power operation, the clock path must be matched to the data path. This matching is achieved by using an extra set of dummy bitlines to match the bitline delay and clock the sense amplifiers at the appropriate time, as in [16].

IV. SRAM

A 4-kB SRAM was designed, extracted, and simulated to measure the read times of the new sense amplifier. The SRAM was composed of eight sub-arrays that each contained 64 rows and 64 columns. The addition of the dummy column increases the area of the SRAM by less than 1/64th of the original area. The layout of the SRAM is shown in Fig. 22. The SRAM was simulated at a temperature of 110 °C with the RV, LE, SE, SLE, SSE, RLE, RSE, and HV cells. Furthermore, the RV and HV cells were also simulated with a conventional sense amp, and these results were used as a reference for the new sense amplifier. The following details the read and write times of the SRAM as well as some performance characteristics of the new sense amplifier.

Fig. 23 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 dummy cells (their contribution is negligible, however, given the size of the SRAM). The leakage trends seen above for the single cell remain true for the complete SRAM, where LE and SE offer a reduction of 70× and 2× while storing a “0” and a reduction of approximately 7× when storing a “1.” The stability improved cells, and the resized cells also show the same leakage trends from the single-cell experiments.

The total SRAM read access time includes the following four components:

1) input register propagation delay and hold times;
2) address decoding delay;
3) delay for WL, bitline, and sensing;
4) output register setup time.

The simulation results showing these components for the various SRAMs composed of different asymmetric cells are shown

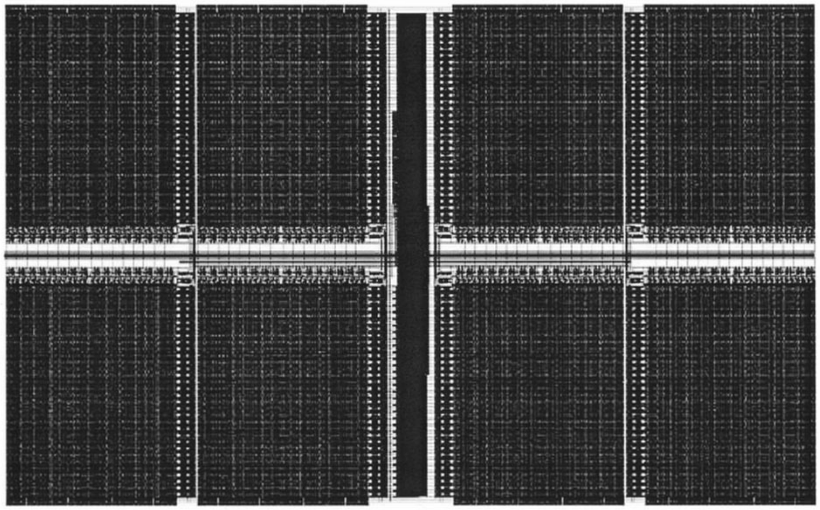


Fig. 22. Layout of simulated SRAM.

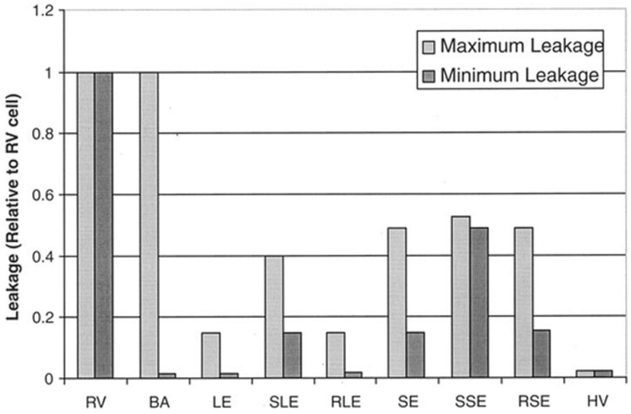


Fig. 23. Maximum and minimum leakage current attribute to cells.

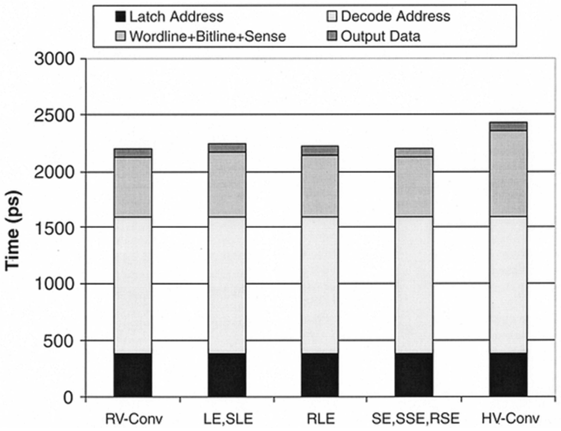


Fig. 24. Breakdown of memory access time.

in Fig. 24. Notice that only the sense times (the third component of Fig. 24) 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. 25 shows the sense times in more detail. It can be seen that the worst case sensing times of the asymmetrical cells with the new sense amplifier are now on par with the RV cell with a conventional sense amplifier. In comparison with the RV cell using a conventional sense amp, the LE cell is 8.5% slower (although the effect on the total read time is an increase of just under 2%, as seen in Fig. 24). The SE cell, on the other hand, has sense times that are within 1% of the sense times of the RV cell. The RLE cell has a sense time that is only 3.4% longer than that of the RV cell, and the RSE cell has a worst case sense time that is equal to that of the SE cell. It is interesting to note that the HV cell with a conventional sense amplifier is 43% slower.

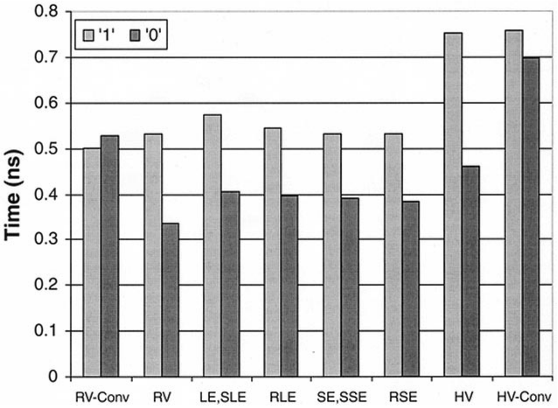


Fig. 25. Sense times during a read cycles, i.e., the third component of Fig. 24.

An important side comment to be made is that the new sense amplifier does not speed up the sensing for RV and HV when compared to the sensing with the conventional sense amplifier. Indeed, the RV and HV cells with the new sense amplifier have worst case sense times, which are 1% slower than the sense times with the conventional sense amplifier. Thus, in comparing the speed of the new cells with the new sense amp to the conventional cells with the conventional sense amp, the comparison is *fair* and valid because the new sense amplifier on its own does not speed up the read access time of the conventional cells.

When a "1" should be read, there is a race between BLB and DB to read a "1" and "0," respectively. If the signal along DB is faster than usual, while the signal along BLB is slower than usual, there is a possibility that an incorrect value may be read by the sense amp. Therefore, a margin has to be designed into the SRAM array; more specifically, the discharge time along

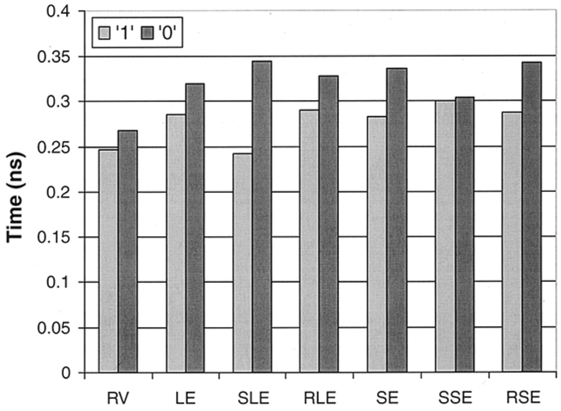


Fig. 26. Write times.

DB must be slower than BLB so that an incorrect value is not read. The built-in margin must also consider the variations in discharge times along BLB and DB.

Monte Carlo analysis was performed on the different asymmetrical cells to find the mean and standard deviation of the read current from each side of the cell. The sense amplifier was then tested assuming the read current along bitlines could vary by $\pm 3\sigma$ from its mean value. The worst case possibility when reading a "1" is when the signal along BLB is 3σ slower and the signal along DB is 3σ faster. The worst case possibility when reading a "0" is when the signal along both BL and DB are 3σ slower. Under all conditions, the sense amplifier sensed the correct value.

To determine the margin that existed until the sense amplifier would sense the incorrect value, the signal along DB was incrementally made faster and BLB made slower until the sense amplifier would output a "0" when BLB was discharging. When DB was 3.74σ faster than its nominal value and BLB was 3.74σ slower than its nominal value, the sense amplifier read the incorrect value. Noting that the capacitance along the dummy bitlines was 13.6% larger than the capacitance along the normal bitlines, the signal along the dummy bitline can be up to 17.4% faster than the signal along BLB before the sense amplifier reads the incorrect value in this design. Thus, the sense amplifier will perform correctly as long as (3) is satisfied as follows:

$$\frac{I_{\mathrm{read}_{\mathrm{dummy}}}}{C_{\mathrm{dummy}}} \leq 1.174 \frac{I_{\mathrm{read}_{\mathrm{fastbitline}}}}{C_{\mathrm{bitline}}}. \tag{3}$$

Furthermore, the capacitance along the dummy bitline C_{dummy} can become a design consideration. C_{dummy} can be controlled by connecting the dummy bitlines to different numbers of sense amplifiers and not routing the bitlines through a column multiplexer. By lowering C_{dummy}, the "0" read time will become faster, and the "1" read time will become slower. At the same time, the probability of a sense failure will increase. Conversely, by increasing C_{dummy}, the "0" read time will increase, and the "1" read time will decrease. The probability of a sense failure will also decrease. In this design, C_{dummy} can be increased to make the read times between a "1" and a "0" more symmetric and, at the same time, increase the margin before an incorrect value is read.

Finally, the write times for the different cells are shown in Fig. 26. The write times range from a 13.4% increase over the RV cell's write time for the SSE cell to a 27.6% increase for the RSE cell. The increase in write times is of minor importance since the write times are shorter than the read times associated with the cells and, therefore, the speed of the SRAM is dependent on the read time.

V. Architectural Enhancements

We investigated two cache organizations that use asymmetric cell designs, i.e., *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 [8], the SPEC2000 programs we studied exhibit a strong bias toward zero. Specifically, we observed that a level-1 data cache had an average 78.7% zeros in the data stream, and a level-1 instruction cache had an average of 62.9% zeros. Given this, the statically biased cache with the SE cells reduces leakage by 4.5× and 3.8× for an instruction and a data cache, respectively, compared to conventional symmetric-cell caches. The caches are 32-kB four-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 regular-V_t 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 tradeoffs apply to this design. An investigation of these issues is part of our ongoing and future work.

VI. Conclusion

In this paper, we have 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 toward 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 six best asymmetric cells offer different performance/leakage/stability characteristics. The SE cell reduces leakage power by at least 2× and by 7× in the preferred state. It is as fast as the conventional RV SRAM cell. By comparison, the LE cell reduces leakage by at least 7× and by approximately 70× in the preferred state. Its total read time is only 2% higher than the SE and RV cells. These latter two cells have lower stability than LE under both the SNM *and* the $I_{\mathrm{trip}}/I_{\mathrm{read}}$ tests. Four other cells that compensate for stability were designed, two by choosing

TABLE XIII
SUMMARY RESULTS FOR ALL THE CELLS

Cell	Leakage (0)	Leakage (1)	Δ Delay	Δ Stability (SNM)	Δ Stability ($I_{\mathrm{trip}}/I_{\mathrm{read}}$)	Δ Area
RV	100%	100%	0%	0%	0%	0%
LE	1%	14%	2%	7%	-5%	0%
SE	14%	50%	0%	-6%	15%	0%
SLE	14%	43%	2%	23%	-7%	0%
SSE	50%	53%	0%	9%	13%	0%
RLE	2%	14%	1%	22%	5%	2%
RSE	15%	49%	0%	8%	15%	3%

TABLE XIV
SUMMARY RESULTS FOR ALL THE CELLS, SHOWING EXPECTED LEAKAGE

Cell	Expected Leakage	Δ Delay	Worst Δ Stability	Δ Area
RV	100%	0%	0%	0%
LE	5%	2%	-5%	0%
SE	25%	0%	-6%	0%
SLE	23%	2%	-7%	0%
SSE	51%	0%	9%	0%
RLE	6%	1%	5%	2%
RSE	25%	0%	8%	3%

different combinations of threshold voltages for the cell transistors, and two by changing some transistor sizes. The SSE cell reduces leakage power by 1.9× and 2× in the preferred state with no performance degradation, and the SLE cell reduces leakage power by 2.5× and 7× in the preferred state with only a 5% increase in read access times. The SSE and SLE cells have comparable stability to the RV cell. The RLE cell reduces leakage by 58× in the preferred state and by 7× in the other state with only a 1% increase in read access time, and an area increase of approximately 2.4%. The RSE cell reduces leakage by approximately 7× in the preferred state and 2× in the other state. It has no performance degradation, but has an area increase of approximately 2.9%. The RLE and RSE cells have comparable stability to the RV cell. By comparison, an all high-V_t cell reduces leakage power by approximately 70×, while its bitline discharge time is 43% slower than the SE and RV cells.

We also proposed two cache organizations that used either a static bias toward zero, or dynamic, selective inversion to maximize the number of cache bits that are zero. While the reduction possible with either technique depends on application behavior, for the SPEC2000 benchmarks that we considered, the statically biased cache with the SE cells reduces leakage by 4.5× and 3.8× for an instruction and a data cache, respectively, compared to conventional symmetric-cell caches.

A summary of the results pertaining to all the cells, relative to the RV cell, is shown in Table XIII. Here, "Leakage (0)" and "Leakage (1)" refer to the leakage when the cell is storing a "0" and a "1," respectively, "Delay" refers to the total read access time and "Area" refers to the total cell layout area. If we work with the observed average of approximately 70% "0" 's and 30% "1" 's, we can give projections for the "Expected Leakage" for the long-term average leakage of an SRAM array (accounting only for the cell leakage), as shown in Table XIV, where the column for "Worst Δ Stability" gives the worst case between the two columns of Table XIII corresponding to the SNM and $I_{\mathrm{trip}}/I_{\mathrm{read}}$ tests. If one had to single out *the best* cases, it is perhaps the case that SSE and RLE combine the best features. SSE has less than half the original leakage of the RV cell with no loss of either performance, stability (mean at nominal voltage), or area. RLE has only 6% of the original leakage of the RV cell with a performance loss of only 1%, no loss of stability, and an area increase of only 2%.

In future technologies, gate leakage will become a considerable portion of total static power dissipation at room temperature. The asymmetric cells presented in this paper do not address the gate leakage component and, thus, their total leakage savings will be less. The asymmetric cells, however, are expected to considerably decrease the static power dissipation at high operating temperatures where subthreshold leakage will still be the dominant leakage component. Furthermore, as V_{DD} decreases in successive technologies, the V_t difference between the regular- and high-V_t transistors must be reduced to keep the asymmetric cells stable; thus, the leakage savings will be reduced. In future technologies, as V_{DD} is decreased, the stability of the asymmetric cells may decrease.

REFERENCES

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, pp. 23–29, July–Aug. 1999.

[2] T. Kam, S. Rawat, D. Kirkpatrick, R. Roy, G. S. Spirakis, N. Sherwani, and C. Peterson, "EDA challenges facing future microprocessor design," *IEEE Trans. Computer-Aided Design*, vol. 19, pp. 1498–1506, Dec. 2000.

[3] F. Hamzaoglu, Y. Ye, A. Keshavarzi, K. Zhang, S. Narendra, S. Borkar, M. Stan, and V. De, "Dual V_t-SRAM cells with full-swing single-ended bit line sensing for high-performance on-chip cache in 0.13 μm technology generation," in *Proc. Int. Low Power Electronics and Design Symp.*, July 2000, pp. 15–19.

[4] S. Kaxiras, Z. Hu, and M. Martonosi, "Cache decay exploiting generational behavior to reduce leakage power," in *Proc. 27th Int. Computer Architecture Symp.*, Goteborg, Sweden, July 2001.

[5] S. H. Yang, M. D. Powell, B. Falsafi, K. Roy, and T. N. Vijaykumar, "An integrated circuit/architecture approach to reducing leakage in deep-submicron high-performance I-caches," in *Proc. 7th Int. High-Performance Computer Architecture Symp.*, Jan. 2001, pp. 147–157.

[6] H. Zhou, M. C. Toburen, E. Rotenberg, and T. M. Conte, "Adaptive mode control: A static-power-efficient cache design," in *Proc. Int. Parallel Architectures and Compilation Techniques Conf.*, Sept. 2001, pp. 61–70.

[7] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 1996.

[8] N. Azizi, A. Moshovos, and F. N. Najm, "Asymmetric-cell caches: Exploiting bit value biases to reduce leakage power in deep-submicron, high-performance caches," TR-01-01-02, ECE Dept., Univ. Toronto, Toronto, ON, Canada, 2002.

[9] E. Seevinck, Sr., F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SSC-22, pp. 748–754, Oct. 1987.

[10] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, pp. 658–665, Apr. 2001.

[11] E. A. Amerasekera and F. N. Najm, *Failure Mechanisms in Semiconductor Devices*, 2nd ed. New York: Wiley, 1997.

[12] E. Grochowski, D. Ayers, and V. Tiwari, "Microarchitectural simulation and control of di/dt-induced power supply voltage variation," in *Proc. 8th Int. High-Performance Computer Architecture Symp.*, 2002, pp. 2–12.

[13] J. M. Musicer and J. Rabaey, "MOS current mode logic for low power, low noise CORDIC computation in mixed-signal environments," in *Proc. Int. Low Power Electronics and Design Symp.*, July 2000, pp. 102–107.

[14] T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. M. K. Tsutsumi, Y. Nishimura, Y. Kohno, and K. Amani, "A 20 ns 4 MB CMOS SRAM with hierarchical word decoding architecture," in *IEEE Int. Solid-State Circuits Digital Tech. Papers Conf.*, 1990, pp. 132–133.

[15] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in low-power RAM circuit technologies," *Proc. IEEE*, vol. 83, pp. 524–543, Apr. 1995.

[16] B. S. Amrutur and M. A. Horowitz, "A replica technique for wordline and sense control in low-power SRAMs," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1208–1219, Aug. 1998.

Navid Azizi was born in Tehran, Iran, on July 24, 1978. He received the B.A.Sc. degree (with honors) in computer engineering, the M.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2001 and 2002, respectively, and is currently working toward the Ph.D. degree in electrical engineering at the University of Toronto.

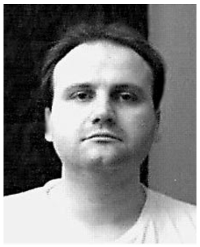
From May to September of 2000 and 2001, he was a Software Engineer with the Altera Corporation, Toronto, ON, Canada. His current research interests include low-power memory and digital integrated-circuit design.

Mr. Azizi was the recipient of the 2001–2003 Postgraduate Scholarship Award presented by the National Sciences and Engineering Research Council (NSERC) of Canada.

Farid N. Najm (S'85–M'89–SM'96) received the B.E. degree in electrical engineering from the American University of Beirut (AUB), Beirut, Lebanon, in 1983, and the M.S. and Ph.D. degrees in electrical and computer engineering (ECE) from the University of Illinois at Urbana-Champaign (UIUC), in 1986 and 1989, respectively.

From 1983 to 1984, he was an Electronics Engineer with AUB. In 1989, he joined Texas Instruments Incorporated, Dallas, TX, as a Member of Technical Staff with the Semiconductor Process and Design Center. In 1992, he joined the ECE Department, UIUC, as an Assistant Professor, and then became a Tenured Associate Professor with UIUC in 1997. In 1999, he joined the ECE Department, University of Toronto, Toronto, ON, Canada, where he is currently a Professor of ECE. He coauthored *Failure Mechanisms in Semiconductor Devices, 2nd Ed.* (New York: Wiley, 1997). His research interests are in the general area of computer-aided design (CAD) tool development for low-power and reliable very large scale integration (VLSI) circuits, including power estimation and modeling, low-power design, power grid analysis and verification, and reliability analysis and prediction.

Dr. Najm is an associate editor for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and the IEEE TRANSACTIONS ON VERY LARGE SCALE (VLSI) SYSTEMS. He served as general chairman for the 1999 International Symposium on Low-Power Electronics and Design (ISLPED'99), and as Technical Program co-chairman for ISLPED'98. He has also served on the Technical Committees of international conferences on computer-aided design (ICCAD), design automation conference (DAC), custom integrated circuits conference (CICC), international symposium on quality electronic design (ISQED) and international symposium on low-power electronics and design (ISLPED). He was the recipient of the 1992 IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS Best Paper Award, the 1993 National Science Foundation (NSF) Research Initiation Award, and the 1996 NSF CAREER Award.



Andreas Moshovos (S'94–A'98) received the B.Sc. degree in computer science from the University of Crete, Crete, Greece, in 1990 and the Ph.D. degree in computer sciences from the University of Wisconsin–Madison, in 1998.

He is currently an Assistant Professor with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada. His research interests are computer architecture and microarchitecture for high-performance, low-cost, and power-aware microprocessors.

Prof. Moshovos is a Member of the Association for Computing Machinery (ACM). He was a recipient of the National Science Foundation (NSF) CAREER Award