

Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance Content-Addressable Memories

Igor Arsovski and Reid Wistort

IBM Silicon Solutions, Essex Junction, e-mail: arsovski@us.ibm.com

Abstract - A memory sense-amplifier self-calibrates during sense-line precharge to reduce the required signal development and minimize data capture timing uncertainty caused by random device variation. When compared to conventional single-ended sensing, this method reduces sense time by 70% and decreases sense-power by 40%. The self-referenced sensing scheme (SRSS) is used to implement the search operation in Content-Addressable Memory (CAM) testchip. Fabricated in 1V 65nm CMOS, this scheme achieves a 0.6ns search time on a 70bit sense-line while consuming only 0.99 fJ/bit/search. Measured search access time on a five bank 64x240bit ternary CAM including selective precharge is 2.2ns. Measured power consumption at 450MHz is 10mW. Hardware shows robust search operation over a voltage range of 0.6V to 1.7V.

I. INTRODUCTION

As CMOS device geometries scale below 100nm, random variation in device parameters create a major challenge in yielding high performance designs [1]. As device geometry shrinks, variation in length, width, and device threshold grows, introducing timing uncertainty for data-arrival and data-capture [2]. This is especially evident in semiconductor memories where the critical path delay is dominated by the memory sensing operation, in which highly-variable minimum size memory devices develop signal on highly capacitive sense-lines. The variable sense current produced by these devices causes a wide range of slew rates which in turn produce large timing uncertainty for output data arrival. Fig. 1 illustrates this problem through a simplified single-ended sensing model in which I_{SENSE} models the sense-current of a single memory cell, C_{SENSE} models the sense-line capacitance, and V_{TH} models the threshold of the single-ended sense-amplifier (SA).

To detect I_{SENSE} the SA shown in Fig. 1 needs to quickly determine whether the sense-line is floating in a high impedance state, or being pulled low by I_{SENSE} . The conventional single-ended sensing scheme performs this differentiation by precharging the sense-line to VDD, selecting a memory cell, and then allowing the cell to develop signal by discharging the sense-line capacitance. If the cell is storing a '0', $I_{SENSE} = 0$, and the sense-line remains precharged at VDD. If the cell is storing '1', $I_{SENSE} > 0$, and the sense-line is slowly discharged to GND. The single-ended SA simply determines whether the sense-line remained above V_{TH} ('0' state) or discharged below V_{TH} ('1' state).

However, as random variation in both I_{SENSE} and V_{TH} increases with shrinking device sizes, so does the data-arrival window for the sense output (OUT). Fig. 1 illustrates this random device variation through the Gaussian distributions for I_{SENSE} and V_{TH} in a commercial 65nm CMOS process. I_{SENSE} variation (ΔI_{SENSE}) causes variable sense-line slew rates which when supplied to SAs with variable V_{TH} (ΔV_{TH})

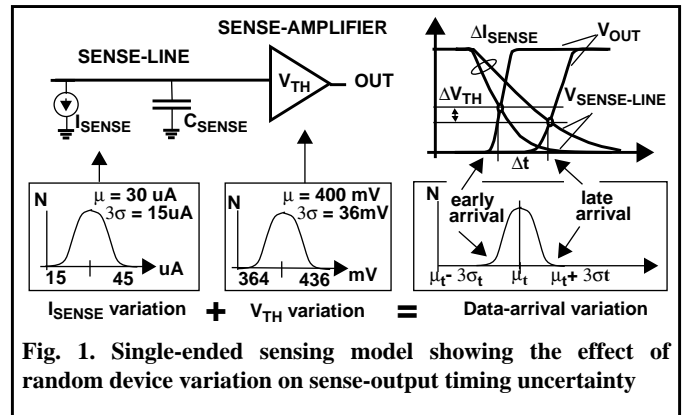


Fig. 1. Single-ended sensing model showing the effect of random device variation on sense-output timing uncertainty

produce two extremes for output data arrival, early arrival and late arrival. In Fig. 1, this is represented through the output data arrival distribution showing the mean arrival as (μ_t) and the standard deviation for arrival as (σ_t). For a three-sigma design the minimum sense-delay can vary across different sensing schemes, and is given by ($\mu_t + 3\sigma_t$).

With I_{SENSE} varying by as much as 3X across-chip, the timing uncertainty ($3\sigma_t$) in the conventional sensing scheme [3] can exceed 50% of μ_t . The same is true for the power-improved current-race sensing scheme [4], which precharges the sense-line to GND and senses the impedance as the sense-line is charged towards VDD. To bound the large data-arrival timing window, current memories use large data-capture margins which increase sense-delay and reduce performance. The self-referenced sensing scheme (SRSS) presented in this paper improves performance by reducing the required signal development and minimizing the timing uncertainty caused by random device variation. Instead of using a common precharge voltage for all SAs [3,4], the SRSS allows each individual SA to precharge its sense-line to a unique voltage level slightly above its own SA threshold. As described in Section III, this self-referenced precharge reduces signal development by 90% for an overall sense-time improvement of 70% when compared to [3], and 60% when compared to [4]. By reducing the voltage swing on the sense-line this method also improves sense-power by 40%.

Although the implementation focus of this paper is Content Addressable Memory (CAM), this scheme can be applied to ROMs, multi-port memories, single-ended register arrays, and other single ended sensing applications.

II. SELF-REFERENCED SENSING SCHEME

In the CAM context, SRSS is used to implement the performance-critical search operation. Fig. 2 shows the CAM search architecture highlighting the circuit detail of both the CAM sense-line, also known as Match-Line, and the SRSS SA. The search operation is performed by supplying the search data on the Search-Lines (SLs), comparing this search

data to the stored data in every CAM word in parallel, and developing the search results on the Match-Lines (ML).

This parallel word comparison is performed with dedicated search hardware attached to each ML. As the ML architecture in Fig. 2 illustrates, any bit-mismatch between the search data provided on the SLs (SL₀ to SL_n), and the stored data ('d') in the memory cells ('m') will create a path from ML to GND through the mismatched bit-compare circuits. Similar to the simplified model in Fig. 1, the ML is either in high-impedance state, or the ML is being pulled to GND by at least one bit-compare circuit. For the purposes of this paper ML with no mismatches will be denoted as ML₀, and an ML with the hardest to detect mismatch, where only one bit-compare circuit is mismatched, will be denoted as ML₁. In general, an ML with 'n' bit-misses will be denoted as ML_n.

Prior to sensing, the search data is supplied on the SLs while (MLRST=1) is used to reset the MLs to GND. Focusing on the SRSS SA, this ML state keeps the SA signals CS high, SN low, and MLOUT high. The sensing is executed in two stages: precharge and evaluate.

In the precharge stage (where MLRST=0, PRE=0) P1 devices in each SA start precharging the ML toward VDD. Multi-bit mismatched MLs drain this precharge current and keep the ML firmly at GND, while ML₀ and ML₁ ramp across the threshold of I1. As soon as the ML voltage crosses the threshold of I1, the CS voltage drops to a level that biases N1 in the cut-off region and stops the ML precharge. At this point each ML₀ and ML₁ is precharged to a unique level slightly above the threshold of its own SA. This self-referenced precharge maintains a small and constant delta between the precharge and sense voltage for each SA across the chip. As soon as the ML voltage reaches this level the precharge current from P1 is channeled to SN quickly charging it to VDD and switching MLOUT low. This completes the precharge stage. The PRE=0 portion of Fig. 3 shows the voltage development on the main nodes of the SRSS SA for both an ML₀ and ML₁ sense case on a 70bit ML. It is evident that ML₁ develops a slightly lower voltage than ML₀, and its corresponding CS₁ node is driven higher than CS₀ to compensate for the current drained by the one-bit mismatch.

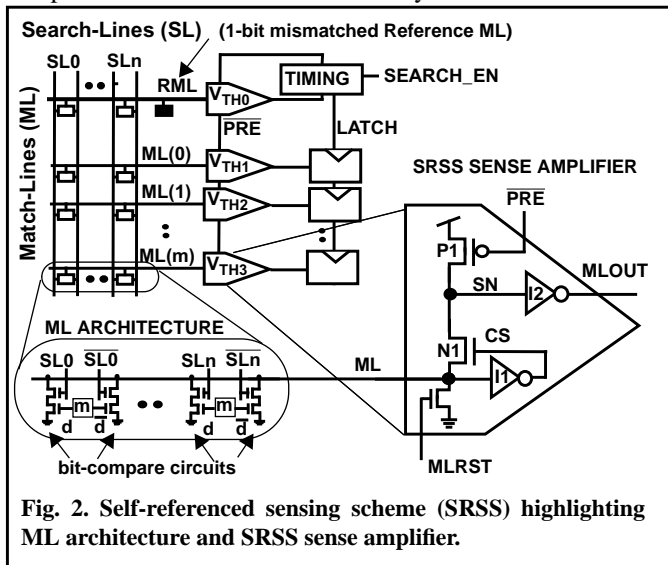


Fig. 2. Self-referenced sensing scheme (SRSS) highlighting ML architecture and SRSS sense amplifier.

To ensure adequate precharge-time for all MLs, a reference ML (RML) in a constant ML₁ state, is also included. The RML shown in Fig. 2 ensures more-than-adequate precharge time and tracks the MLs across process, temperature and voltage conditions. As soon as the RML's MLOUT drops low (similar to MLOUT₁ in the Fig. 3), the precharge stage is completed and the evaluate stage begins.

In the evaluation stage (where PRE=1) ML₀ acts as a capacitor, keeping its precharge state, while ML₁ starts to discharge through the mismatched bit-compare circuits. As ML₁ discharges it quickly trips I1 causing SN and ML to quickly equalize and then discharge, bringing MLOUT₁ back to its high state. At this point the sense data is ready to be detected. The PRE=1 portion of Fig. 3 shows the high-impedance ML₀ keeping its precharge state while the hardest to detect mismatch ML₁ starting to discharge. As soon as P1 stops providing current, CS₁, which was keeping N1 biased to compensate for the pull-down current of the one-bit miss, now causes SN₁ and ML₁ to equalizes and switch MLOUT₁ back to the high state. For fast MLOUT transition the threshold of I2 is designed higher than that of I1. To ensure adequate data capture timing over a wide process window, the LATCH signal shown in Fig. 3 is also generated by the RML. When the RML's MLOUT switches high it generates a LATCH signal which is used to capture the search results. Any mismatch between RML and regular MLs is compensated by delaying the LATCH signal before it is sent to capture the search results. When the search data is latched, the MLRST goes high to reset all MLs to GND and prepare the array for the next search operation. As shown in the next section, the LATCH delay required to bound the data-arrival for SRSS is greatly reduced when compared to [3,4], significantly improving sensing speed. Power performance is also improved by reducing the ML voltage swing to roughly 1/2 VDD level.

III. TOLERANCE TO RANDOM DEVICE VARIATION

The SRSS reduces both signal development time and timing uncertainty for data capture. By generating a ML precharge voltage relative to each SA threshold, this scheme cancels out

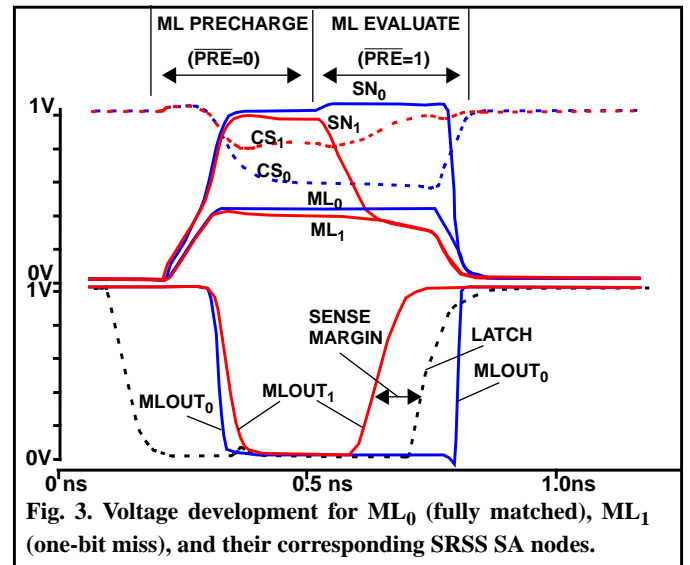


Fig. 3. Voltage development for ML₀ (fully matched), ML₁ (one-bit miss), and their corresponding SRSS SA nodes.

the timing uncertainty caused by random device variation in the SA devices. Variation in P1, N1, and I1 is cancelled out by the self-referenced precharge voltage, while timing uncertainty caused by the threshold variation of I2 is significantly reduced by the SN slew rate which is $\sim 5X$ faster than the slew rate of the ML. By precharging the ML slightly above the threshold of each SA, this scheme also reduces the timing uncertainty caused by variation in I_{SENSE} . The reduced signal development time also reduces the integration of I_{SENSE} variation on the ML capacitance to further reduce the timing uncertainty.

To show the performance improvement of SRSS, Fig. 4 compares waveforms generated by Monte Carlo simulation of both the conventional [3] and SRSS ML sensing scheme. For a fair comparison the simulation was performed on identical 70bit MLs under identical environmental conditions. For explanation purposes the precharge pulse (\overline{PRE}) was artificially extended to separate the precharge and sense phase. Focusing on the ML precharge, it is clear that the SRSS has a 60% reduced voltage swing over the conventional scheme which results in a direct reduction in ML power. The band of precharge voltage levels for the SRSS ML illustrates the self-referenced nature of the SRSS SA. Since the ML precharge voltage is relative to the SA threshold and the SA threshold is unique in every Monte Carlo run, each run for the SRSS SA results in a distinct ML precharge voltage. In contrast, the conventional ML sensing scheme treats all SAs equally by precharging all MLs to identical voltage level. As soon as the precharge stops, a bit-compare circuit starts to discharge the capacitive ML with highly variable currents creating a wide envelope of ML slew rates. In the conventional scheme this causes the high-going output MLOUT to arrive over a large timing window requiring a large bounding margin to capture all data-arrival times. In the SRSS, the ML precharge voltage is close to the SA threshold of each SA, producing an almost 90% reduction in required signal development and a much tighter timing uncertainty. The overall sense-time and power comparison of the SRSS scheme compared to the two most commonly used ML sensing schemes is shown in Fig. 5. For a

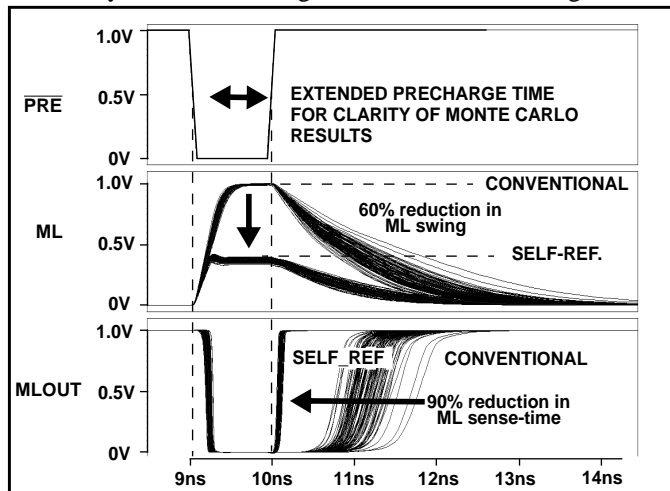


Fig. 4. Comparing voltage development on ML and MLOUT for both conventional ML sensing [3] and SRSS.

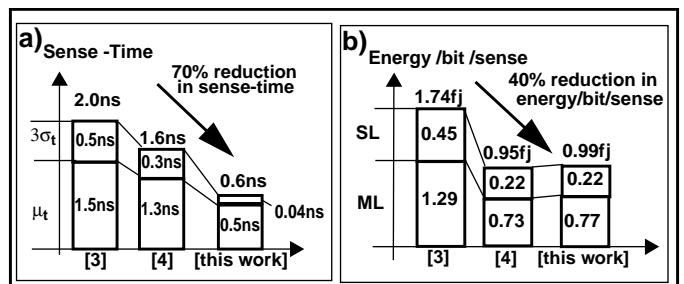


Fig. 5. a) Comparison of ML sense-time for SRSS and two commonly used sensing schemes on a 70bit ML b) CAM performance metric showing SRSS energy/search reduction

fair comparison these schemes were all simulated in 65nm CMOS process on identical MLs for identical environmental conditions. Fig. 5(a) shows a significant reduction in both μ_t and $3\sigma_t$ providing a 70% reduction in sense-time. In addition to speed improvement, this scheme also improves power. Fig. 5(b) uses a well known CAM power performance metric to report the energy required to perform a single search operation for a single bit of data. Similar to [4], the SRSS achieves a reduction in both SL and ML energy for a total of 40% lower energy/bit/search than the conventional sensing scheme[3].

IV. TOLERANCE TO NOISE AND SUBTHRESHOLD LEAKAGE

The main concerns when operating a single-ended SA close to its threshold are noise and sub-threshold leakage. To improve SRSS noise margin, the difference between the ML precharge voltage and the SA threshold can be increased using two methods. First, by reducing the strength of I1 it takes longer to shut-off the ML precharge through N1, increasing the precharge voltage and extending the noise margin. Second, an additional keeper device, omitted in Fig. 2 for clarity, is added to increase the noise margin. This device shown in Fig. 6 is gated by the CS node and turns-on slightly during the precharge stage to increase the ML voltage beyond the SA threshold, also increasing the noise margin. The effect of this keeper device can be seen in Fig. 3 during the precharge stage,

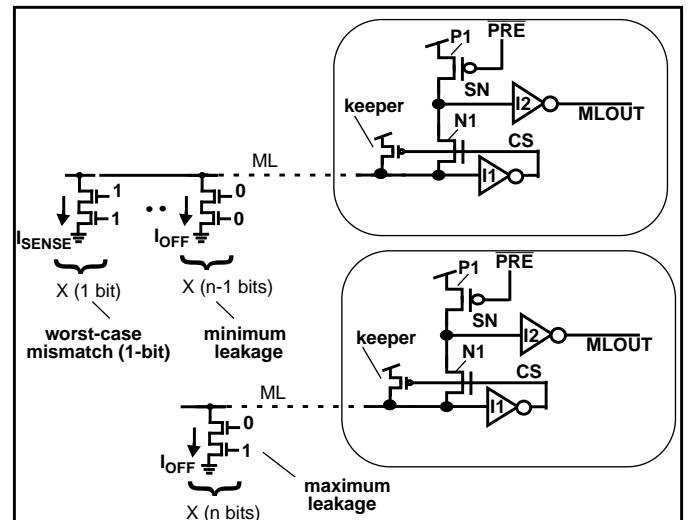


Fig. 6. Worst-case data-pattern for testing the maximum effect of subthreshold leakage on SRSS sensing

which shows a slowly decreasing CS_0 . By extending the ML precharge the keeper device, which turns-on during the precharge stage, increases the precharge voltage and reduces the susceptibility to noise. To further reduce noise susceptibility during the brief sensing stage the ML is shielded from noisy signals through layout techniques.

To address ML sub-threshold leakage this scheme limits the cell-count on the ML. Furthermore, the keeper device used for noise reduction is also used for sub-threshold leakage cancellation. This device is sized to produce current weaker than a single turned-on bit-compare circuit but stronger than the maximum leakage current on the ML. Fig. 6 shows the worst case data-pattern for subthreshold leakage. In this pattern the SRSS tries to differentiate between an ML_1 with minimum leakage (all inactive bit-compare circuits are turned off with '0' on both NFETs) and an ML_0 which has all bit-compare circuits in the maximum leakage state ('0' on the top and '1' on their bottom NFET device). Monte Carlo simulation of this worst case data-pattern under extreme environmental conditions (1.7V, 125C, best case process corner), shows sensing margin that is more than 300% of the sense-time. To accommodate any model to hardware discrepancies, four margin signals were also added to provide three additional precharge-time extensions and three additional latch-time extensions. None of the margin pins had to be used during hardware testing.

V. HARDWARE RESULTS

The SRSS is implemented in 64x240bit ternary CAM testchip and fabricated using 1.0V, 65nm CMOS process. The testchip microphotograph, having a total CAM area of 115um x 783um, is shown in Fig. 7. Table I summarizes the CAM testchip features and hardware measurement results. Measurements confirm a fully functional search operation across 0.6V - 1.7V voltage range using the minimum sense margin settings. Total search access for the five-bank 64x240bit ternary CAM including selective-precharge evaluation was measured at 2.2ns. This McLeod loop hardware measurement [7] supports simulation results of 0.6ns search time on a single 70bit ML. Power measurements show a 10mW of power consumption at the maximum cycle frequency of 450Msearches/second. This significant improvement in performance is achieved with a minimal area overhead that is 10% smaller compared to [4].

VI. SUMMARY

Small devices in advanced CMOS processes experience large variations in parameters such as width, length, and most importantly threshold voltage [1]. In semiconductor memories, these random device parameters cause variation in both memory-cell sense current and sense-amplifier threshold voltage. This variation coupled with low slew-rate on capacitive sense-lines causes data-arrival timing uncertainty which requires large bounding margins, significantly degrading performance. This paper presented a novel single ended sensing scheme which minimizes the timing uncertainty caused by random device variation. This scheme uses a self-referenced sense-line precharge to cancel sense-amplifier device variation. In addition, this precharge reduces signal development time, which in turn reduces the timing effects of

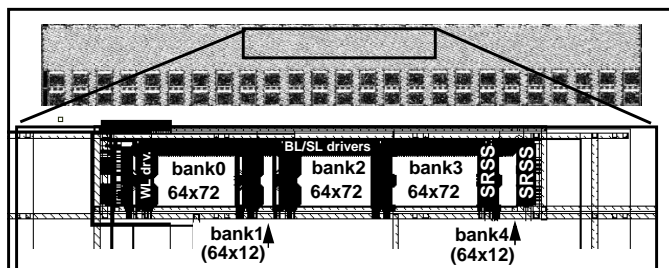


Fig. 7. Microphotograph of CAM testchip showing lower metal layers through chip layout

memory device variation. When compared to the conventional sensing scheme the SRSS achieves 70% faster sense-time and 40% lower power. When compared to the low-power current race scheme this scheme achieves a 60% faster sense-time for a 5% increase in power consumption. These improvements are achieved through a robust design implementation verified in hardware.

Organization	3 banks x 64words x 72 bits 2 banks x 64words x 12 bits
Performance (1.0V,25C)	
- Search time (72bit bank)	0.6ns
- Access time (240bit)	2.2ns
- Power (at 450MHz)	10mW at (1V, 450MHz)
Technology	1.0V, 65nm CMOS
Functional Voltage	0.6-1.7V
Ternary CAM Cell Size	2.3 μ m ²
Test Chip Size	115um x 783um

Table 1: CAM features and hardware results summary

ACKNOWLEDGMENTS

The authors thank M. Fragano, L. Wissel, R. Nadkarni, H.Pilo, and J. Oppold for insightful discussions, M. Willete for hardware testing, M. Ziegerhofer and M. Lestrangre for technical editing, and M. Boudreaux, J. Chickanosky, and M. Merrill for encouragement and support.

REFERENCES

- [1] H. Masuda, S. Okawa, M. Aoki "Approach for physical design in sub-100 nm era," ISCAS 2005, Vol. 6, pp. 5934 - 5937.
- [2] P.S.Zuchowski, P.A. Habitz, J.D.Hayes, and J.H. Oppold, "Process and environmental variation impacts on ASIC timing," ICCAD-2004, pp. 336 - 342
- [3] P. Lin and J. Kuo, "A 1-V 128-kb Four-Set-Associative CMOS Cache Memory Using Wordline-Oriented Tag Compare(WLOTC) Structure with Content-Addressable Memory (CAM) 10-Transistor Tag Cell," IEEE JSSC, Vol. 36, No. 4, pp. 666-676, Apr. 2001.
- [4] I. Arsovski, T. Chandler, A. Sheikholeslami, "A Ternary Content-Addressable Memory (TCAM) Based on 4T Static Storage and Including a Current-Race Sensing Scheme," IEEE JSSC, Jan. 2003.
- [5] C. A. Zukowski and S.-Y. Wang, "Power reduction in large fan-in CMOS gates in logic arrays using selective precharge," Great Lakes Symposium on VLSI, 1997, pp. 83-87
- [6] H. Miyatake, M. Tanaka, and Y. Mori, "A design for high-speed low-power CMOS fully parallel content-addressable memory macros," IEEE JSSC, Vol 36, No. 6, pp. 956-968, June 2001.
- [7] M. McLeod, O. Wager, and A. Vogel. "A new method for improved delay characterization of VLSI logic," European Solid State Circuit Conference, September 1982