Timing Error Correction Techniques for Voltage-Scalable On-Chip Memories

Eric Karl, Dennis Sylvester, and David Blaauw
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI
{ekarl,dennis,blaauw}@eecs.umich.edu

***Abstract*—We describe new DVS-capable SRAM circuit techniques and sensing schemes that enable timing error correction for memories. The sensing scheme and circuit techniques enable aggressive voltage scaling and eliminate conventional design margins considering (i) inter- and intra-die process variations, (ii) local supply voltage variations, and (iii) temperature fluctuations. The proposed techniques enable the exploitation of address and data-dependent memory access delays, allowing additional voltage scaling within a given error recovery energy budget. Applications allowing a fraction of latent operations enable voltage-scaling below a critical voltage. Below this critical voltage point, occasional temperature, voltage and process variations induce timing errors in critical paths which are detected and corrected by the proposed circuits. Simulation results indicate that the techniques enable aggressive supply voltage-scaling to obtain power savings from 12 to 35%.**

I. Introduction

Increasing demand for large, fast on-chip memories has placed growing importance on designing high-speed memories with minimal power consumption while delivering high yields in an era of probabilistic device performance. Leakage power in these large on-chip memories is significant and designers must struggle to minimize leakage during both standby *and* active modes of operation. Leakage power decreases roughly cubically with reduction in supply voltage [1]; therefore dynamic supply voltage scaling systems provide a powerful mechanism to control dynamic and leakage power with reasonable complexity and area overhead (i.e., no need for multiple voltage supplies within the memory array).

This work presents an approach to dynamic voltage scaling (DVS) for SRAM-based memories that allow aggressive scaling of supply voltage to reduce active power and gate and subthreshold leakage power with a simple, non-invasive sensing scheme. Conventional DVS techniques are limited to a conservative critical voltage that includes overly pessimistic margins for worst-case process variations and temperature fluctuations [3]. In addition to eliminating voltage margins and process and environmental fluctuations, the proposed circuits also provide a measure of protection from uncertainty due to SRAM access device leakage currents. Exponential growth in leakage current variability is detailed in [1] and threatens to dramatically reduce the number of SRAM access devices per bitline in aggressively scaled process technology.

The proposed approach dynamically converges to a minimum operating voltage through an embedded timing error detection and correction circuit. The SRAM voltage is adjusted at run-time by monitoring the rate of timing errors detected, even allowing operation at sub-critical voltages for tradeoffs of error rate vs. supply voltage scaling. A differential voltage is developed on the bitlines in the SRAM array and a standard sense amplifier is triggered speculatively by an enable signal generated from a clock edge. After a delay, a second sense amplifier re-samples the bitline to confirm the value, relying upon a larger voltage differential to provide greater confidence in the measurement. If a timing error (an error in the circuit due to insufficient time to evaluate) is detected, the correct data is available one cycle later from the conservatively-clocked sense amplifier. This technique is particularly advantageous considering technology scaling, since (i) intra-die and ambient variations lead to greater safety margins, (ii) interconnect leads to increased delay variability between SRAM banks, and (iii) data-dependent bitline leakage variability reduces certainty in effective read currents. These factors combine to result in overly-pessimistic worst-case design [2].

II. Timing Error Detection and Correction

The general requirement to ensure proper operation in a sequential system is to guarantee the proper values are stored and propagated through the intermediate storage elements. Commonly used sequential storage elements rely upon sampling input values at clock edges (clock boundaries). In the proposed single-cycle SRAM, there are two clock boundaries that require delay-error detection and correction circuitry: 1) at the clocked storage element near the I/O interface and 2) at the sense amplifier. Existing RAZOR latch and flip-flop circuit structures from [3] can detect and correct timing errors in the signal at the I/O interface clock boundary. The RAZOR latching mechanism consists of an additional delayed-clock latch that re-samples the final data to detect transient timing and voltage errors and returns the correct result with a one cycle penalty in the case of an error.

At the sense amplifier boundary on the read bitpath, two standard differential latch-type (DLT) sense amplifiers [4] are used to double-sample the bitline during a read operation in the SRAM. In Fig. 1, the rising edge of the EN1 signal is generated from the falling edge of the clock to trigger the

0-7803-8834-8/05/$20.00 ©2005 IEEE.

original sense amplifier. The output of this original sense amplifier is immediately stored in an unclocked S-R latch to guarantee stability of the static output bus during the precharge phase of the SRAM cycle. The rising edge of EN2 is delayed from EN1, while the bitlines in the bank continue to develop more significant differential voltage. As supply voltage is scaled down in the SRAM, the effective read current is decreased and the word-line pulse arrives later due to eroding performance in the decode logic. These effects combine to result in a reduced differential voltage present on the bitlines given a fixed amount of time from the beginning of the read cycle. Re-sampling the bitline voltage with a delay from the original sense enable signal allows additional time for the bitlines to discharge and overcome process-variation induced offset voltages, data-dependent leakage currents, and activity-dependent internal voltages in the sense amplifier.

Enable signal pulses are generated from the falling edge of the clock using an inverter delay chain connected to a NAND gate as shown in Fig. 2. A point on the delay chain is tapped to generate the falling edge of the precharge clock (PCLK) using a NAND gate with the unlatched bank select signal. This ensures that the delayed pre-charge phase will begin as soon as the enable pulse (EN1) for the sense amp is de-asserted. EN1 is also used to clock the data bus mux enable to prevent XOR glitches from increasing latency. If the output of the error detection XOR is high when the enable signal is de-asserted, the NAND gate in Fig. 1 will select the output of the second sense amp to be driven on the data bus. The rising edge of EN1 presets the mux to select the output of the main sense amplifier to minimize delay impact.

When a speed path failure is detected at the sense amplifier, the correct value is muxed onto the static data bus. This result will not reach the I/O interface within the clock cycle, but the memory element at the I/O interface is capable of re-sampling the data bus and propagating the correct value via the latching mechanism proposed in [3]. If an error is detected at the latch column or at the sense amplifier block, the SRAM can return a signal similar to a cache miss in a hierarchical memory system. The corrected data is forwarded to the system at the end of the second cycle after the request rather than in one cycle. Many existing systems support hierarchical memory models and would require few changes to include such a timing speculative SRAM.

We have found that delaying the pre-charge phase of the SRAM to accommodate the shadow sense amp requires a 2X increase in pre-charge device and driver sizes. Total area overhead is projected to be less than 8% for an SRAM with similar block sizes and organization. The area overhead is localized near the sense amplifiers; therefore, the overhead is highly dependent upon the cells/sense amplifier. A greater ratio of cells to sense amplifiers reduces the fractional area overhead due to the proposed dual-sensing scheme. The overall structure of the proposed 32-bit SRAM is detailed below in Fig. 3. The 64 kB SRAM is divided into 16 rectangular banks subdivided into four 1kB blocks. The buses for the SRAM were routed to minimize wire-length in the routing channels between banks.

In order to explore address-dependent delay, a pre-charged dual-rail address bus is used to prevent glitching and false evaluate paths without requiring an additional clock boundary in the decode logic. Traditional designs have relied upon clocked decode networks or arrival pulse propagation alongside the address bus to prevent glitching

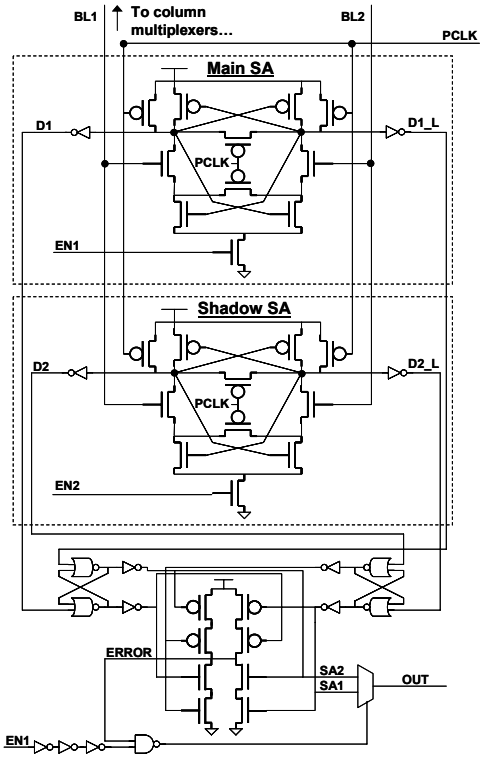

Fig. 1. Main and shadow sense amplifiers latched and compared via XOR, multiplexed to data bus

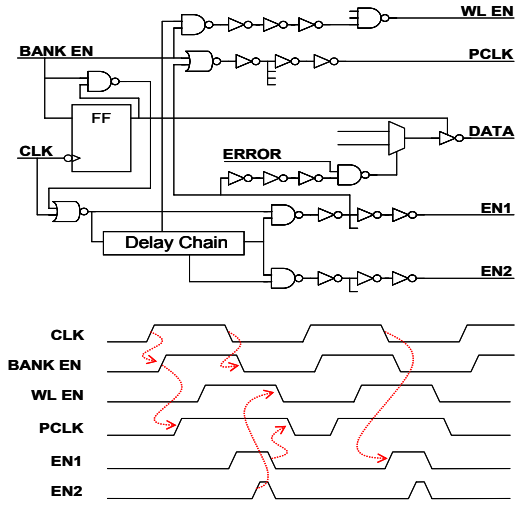

Fig. 2. Timing signal generation and waveforms

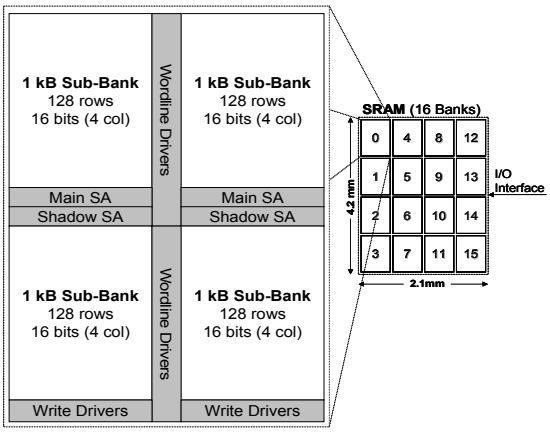

Fig. 3. SRAM bank floorplan

and initiate the read/write sequence. Using the dual rail bus allows simple arrival pulse generation at each bank *as the data arrives*. Generating the pulse and propagating it along the address bus is another option, but simulations revealed less variability between enable pulse and data arrival when using local pulse generators with the dual-rail bus.

III. Exploiting Address/Data Dependence

Recent commercial SRAM designs consist of many banks spread over great distances on chip. Signal propagation and address decode delays typically dominate cycle times in large SRAM designs. A combination of repeaters and long wire segments throughout the address bus are responsible for substantial variation in the path delay between different banks in the SRAM. This path delay variability causes portions of the address space in the SRAM to develop timing errors at widely varying supply voltages. Generating the arrival/enable pulses as the data arrives allows each bank to complete the read/write operation in the minimum cycle time rather than limiting quicker banks to a slower cycle. This benefits applications with high data access locality. When the SRAM is accessed, most accesses will target addresses from a few banks, allowing the SRAM to tune the supply voltage to the process and interconnect characteristics of the dominant banks in the active data set.

Leakage variability in advanced processes is quickly becoming a critical issue in determining the yield of high-performance components. In the context of the SRAM, bitline leakage depends upon the data stored in each cell connected to the bitlines. In addition to process-related variability, leakage currents dependent upon stored data are influencing the effective read currents of SRAM cells [5]. The bitline re-sampling technique allows designers to avoid margining for infrequent transient states that lead to worst-case read current. The shadow sense amplifier allows a speculative sensing phase with the detection/correction to maximize the cycle times of the SRAM in the presence of leakage-induced read current variability.

IV. Results

To evaluate the effectiveness and explore the trade-offs inherent to the shadow-sense amplifier, a 64 kB single-cycle SRAM was designed in 0.18um CMOS technology. Simulation results confirm that the SRAM operates at 250 MHz with worst-case device and interconnect models at 85C with a 10% margin on 1.8V VDD. Under typical process conditions at 25C the design approaches 400 MHz operation. The shadow sense amplifier is clocked to reliably detect timing errors down to 1.3V in the worst-case corner at 250 MHz. Increasing the delay between sense amplifier enable signals increases the minimum safe operating voltage. Additional delay beyond 500ps between sense amplifier enable signals requires pre-charge logic differentiating read and write operations, allowing early pre-charge following a write.

Detailed variability simulations including parasitic models of the SRAM considering physical design were used to determine failure voltages for each SRAM bank. The variability model was adapted from industry-provided SPICE models and includes individual intra-die Gaussian distributions for W, L, and Vth and inter-die Gaussians for W, L, Vth, Rdsw, μ0, and Tox. Each bank of the SRAM was simulated with a fixed set of inter-die variation (-1σ was used) and intra-die variation for devices (-2σ to +2σ). Gaussian variations were generated for each relevant device in each bank and combined to compose a "chip" for simulation purposes.

Figs. 4 and 5 detail the timing error-rate vs. supply voltage with the SRAM used as a direct-mapped cache running memory traces obtained from 11 SPEC2000 benchmarks run on the SimpleScalar/Alpha v 3.0 toolset [6]. The memory traces consisted of 10 million simulation cycles taken from an optimal point within the program execution to deliver typical memory access patterns using the Early SimPoint method [7]. Fig. 4 displays the error-rate on a trace of the *gap00* benchmark for four different

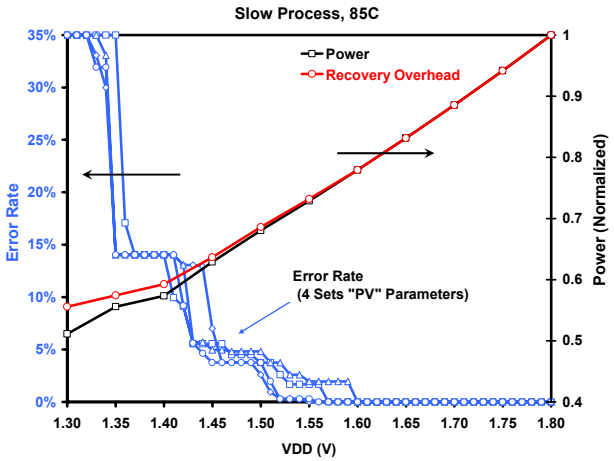

Fig. 4. Error rate of memory as VDD reduces for gap00 memory trace. Four traces represent sets of intra-die variation with inter-die fixed at slow corner. Power both with and without recovery overhead is shown.

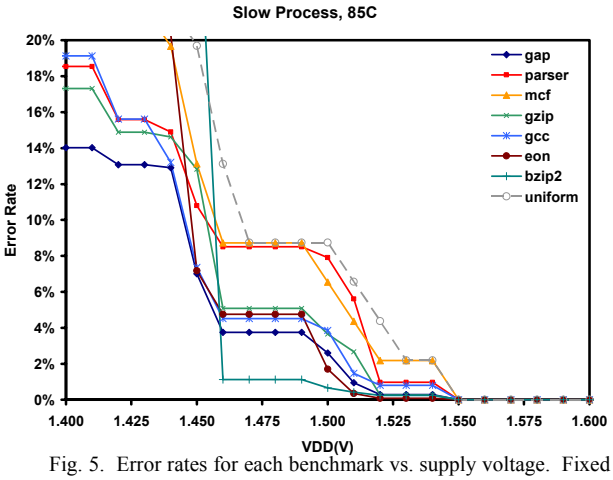

Fig. 5. Error rates for each benchmark vs. supply voltage. Fixed "slow corner" inter and intra-die variation.

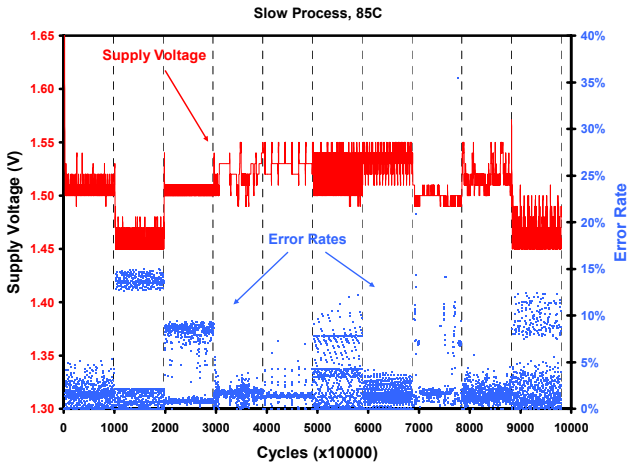

Fig. 6. Instantaneous supply voltage and error rate during 10M cycle DVS simulation with varying workload. Uses error counter with ± 10mV VDD update every 10000 cycles over 10 benchmarks.

worst-case inter-die SRAMs defined by the Gaussian intra-die variability model described above. Fig. 5 shows the impact of memory access patterns on error rate for a given chip. Applications with high data access locality (e.g., *gzip*) will exhibit dramatic fluctuations in error rate, while applications with lower data access locality (e.g., *gap*) will demonstrate frequent gradual increases in error rate as the voltage is lowered. Fig. 6 is a trace of 100M cycles of SPEC benchmarks showing instantaneous voltage and error rates using a simple voltage control algorithm that updates every 10000 cycles. The supply voltage ranges over 100mV during operation for a target error rate of 2%. Error rates surpass 2% in many instances due to the simple ± 10mV control algorithm.

In Table I, the error detection/correction circuitry allows an SRAM design at the worst-case inter-die corner to operate at 1.55V at 85C with a zero error rate and 1.5V with 5% error-rate, as compared to the 1.8V supply for an SRAM without error detection and with a voltage safety margin, saving 12% and 17% power respectively after considering the overhead of the additional circuitry. Without an additional safety margin on voltage, the SRAM operates at 1.62V in the worst-case corner. At operating temperatures less than 85C, the SRAM adjusted to 1.45V at 50C or 1.39V at 30C, saving 23% and 29% power, respectively, over a margined design operating at a fixed 1.8V. A typical part (i.e., not at the worst-case inter-die corner) can be operated at 1.3V VDD, achieving up to 35% power savings over a conservatively margined SRAM. Static power, while not appreciable for this technology, is sensitive to supply voltage with both gate and subthreshold leakages benefiting from lower operating voltages and corresponding reduced ambient temperatures [1].

TABLE I. POWER CONSUMPTION AT VARIOUS OPERATING POINTS

Operating Point	VDD	Total	Static
Single Sense Amp, WC, 85C	1.80V	76.5 mW	16.3 uW
Single Sense Amp, WC, 85C	1.62V	64.3 mW	13.6 uW
Dual Sense Amp, WC, 85C (Zero Error Rate)	1.55V	67.8 mW	12.6 uW
Dual Sense Amp, WC, 85C (~5% Error Rate)	1.50V	63.8 mW	11.9 uW
Dual Sense Amp, WC, 50C (Zero Error Rate)	1.45V	58.9 mW	5.56 uW
Dual Sense Amp, WC, 30C (Zero Error Rate)	1.39V	54.0 mW	3.92 uW
Dual Sense Amp, TYP, 85C (Zero Error Rate)	1.30V	49.4 mW	16.0 uW

All power results for dual sense amplifiers include overhead for the required signal generation, circuitry and increased pre-charge device sizes.

V. CONCLUSION

New DVS SRAM circuit techniques enable the elimination of conservative margins stemming from inter/intra-die process variation and temperature fluctuations while enabling address-dependent and data-dependent timing error detection. The described technique is relevant for coping with increased address decode delays in large SRAMs and data-dependent leakage current variability in advanced process generations. Implemented in a 64 kB cache composed of 16 4kB banks, the technique incurred ~8% area overhead and minimal cycle time penalties while realizing power savings of up to 35%.

REFERENCES

[1] R. K. Krishnamurthy, *et. al*, "High performance and low-power challenges for sub-70nm microprocessor circuits." *IEEE Custom Integrated Circuits Conf.*, pp. 125-128, May 2002.
[2] A. Bhavagnarwala, *et. al*, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability." *IEEE J. Solid-State Circuits*, pp. 658-665, Apr. 2001.
[3] D. Ernst, *et al.*, "RAZOR: A low-power pipeline based on circuit-level timing speculation." *MICRO Conf.*, 2003.
[4] A. Chrisanthopoulos, *et al.*, "Comparative study of different current mode sense amplifiers in submicron CMOS technology." *IEE Circuits, Devices and Systems*, v. 149, iss. 3, 2002.
[5] K. Agawa, *et al.*, "A bitline leakage compensation scheme for low-voltage SRAMs." *IEEE J. Solid-State Circuits*, pp. 726-733, May 2001.
[6] T. Austin, *et. al*, "SimpleScalar: An infrastructure for computer system modeling." *IEEE Computer*, Feb. 2002.
[7] T. Sherwood, *et. al*, "Automatically characterizing large-scale program behavior." *ASPLOS-X*, Oct. 2002.