17.3 A Reconfigurable Dual-Port Memory with Error Detection and Correction in 28nm FDSOI

Mahmood Khayatzadeh[1], Mehdi Saligane[1], Jingcheng Wang[1], Massimo Alioto[2], David Blaauw[1], Dennis Sylvester[1]

[1]University of Michigan, Ann Arbor, MI,
[2]National University of Singapore, Singapore, Singapore

SRAM is a key building block in systems-on-chip and usually limits their voltage scalability, due to the major impact of process/voltage/temperature (PVT) variations at low voltages [1]. Assist techniques to extend SRAM operating voltage range improve the bit cell read/write stability [1-5], but cannot mitigate variations in the internal sensing delay that is needed to develop the targeted bitline (BL) voltage. Hence, large guard bands and performance margins are still needed to ensure correct operation. These margins increase as supply voltage is lowered (Fig. 17.3.1) and must be addressed especially when the SRAM is coupled with margin-less processor designs (e.g., Razor).

This paper introduces a Razor-style [6] error detection and correction (EDAC) technique for SRAM arrays. As commonly applied to logic paths in microprocessors, EDAC schemes eliminate guard bands by dynamically adjusting to PVT variations, thus substantially improving either energy efficiency at fixed frequency or performance at a fixed voltage. We propose the use of dual sampling to speculatively complete an SRAM read access in one cycle for the common case, and in a larger number of cycles for the small number of bit cells that are particularly slow due to random variation (Fig. 17.3.1). Hence, the read access time is fast for most bit cells within a die, whereas the extended cycles can accommodate for the slowest bit cells within the die. Measurements at 0.7V show throughput gains up to 5.6× compared to a conventionally margined SRAM.

Razor SRAM can operate in four different modes (Fig. 17.3.2): normal dual-port mode (2R/2W), Razor read/normal dual-port write (RR/2W), merged-wordline write (2R/1W), Razor read/merged-wordline write (RR/1W). Typically, 2R/2W is used at a voltage close to nominal, whereas the other modes are used at low voltage to improve the robustness against variation, as discussed below. In RR/□2W mode, write is performed as in a conventional dual-port (2R/2W) SRAM, whereas read is performed with the dual sampling approach shown in Fig. 17.3.1. In most cases, the read output is available after one clock cycle, and is then confirmed to be correct through comparison with a second sample taken in the second cycle (Fig. 17.3.2), assuming for simplicity that error detection is performed in a single cycle as in Fig. 17.3.1. For very slow bit cells (those in the statistical tail), the second sample will be found to be different from the first one, and an error will then be flagged. Since these slow bit cells at the tails trigger errors infrequently, the average latency remains very close to one cycle, thus halving the clock cycle compared to a conventional array. Conversely, Razor SRAM is able to accommodate a substantially larger worst-case sensing time t_{sense}. Where t_{sense} is the worst-case delay required to complete the bitline voltage development and the subsequent sensing when compared to a conventional array at same clock cycle T_{CK}. For a typical ratio of t_{sense}:T_{CK} = 1:3, Razor SRAM extend the sensing time to 4/3.T_{CK} (i.e. 4×) compared to a conventional array (T_{CK}/3) as shown in Fig. 17.3.1. Such speed improvement adds to the speed-up inherently offered by margin elimination. Hence, Razor SRAM improves its throughput despite the requirement to double the number of cycles per read.

In 2R/1W mode, the two wordlines are driven as one during write, thus increasing the effective size of the access transistors and hence improving the bit cell's writability (Fig. 17.3.2). In this case, the second address decoder is turned off to reduce energy by 3.9%. 2R/1W, is in effect, a write assist technique that enables more robust operation at low voltages, at the cost of reduced number of write ports (from two to one). The 1RR/1W mode combines the RR/2W and 2R/1W modes, which respectively speed up the read and strengthen write access (i.e., write can be correctly performed in a shorter time). Razor Read (RR) allows the removal of the conservative margin that is traditionally introduced to compensate for the few failing bit cells at the tail of the statistical distribution of t_{sense}. In this case, margin-less RR enables a considerable speed-up compared to a conventional read. This is particularly useful in the near-threshold regime where variations causes a long tail in the read current.

After being detected, each error flagged by the Razor SRAM is managed by the processor (error control block in Fig. 17.3.3) through any existing roll-back mechanism available in processor (i.e. via branch misprediction) or using the EDAC capability in Razor-class processors [6]. A Razor-class SRAM memory controller sets the operation mode and handles the following special events (Fig. 17.3.3): read during read (RDR), incoming read in the same row as on-going read (RSR), write during read (WDR). The correct content of the register file is preserved by inhibiting write-back during error detection via the *stabilize* register (Fig. 17.3.3), and managing errors flagged in Razor register files within the same block.

The 28nm FDSOI Razor SRAM test chip (Fig. 17.3.7) includes a 32kb array with four 128×64 sub-banks, 4:1 column multiplexing, and on-chip testing harness. To prevent the sense amplifier from affecting the bitline voltage in the second read cycle, a current latch sense amplifier is used (Fig. 17.3.2). The only overhead compared to a conventional SRAM is due to the memory controller, whose area and power are 8.7% and 6.8% of the whole memory, respectively. Larger capacity SRAM further amortize the memory controller overhead. Measurements of 20 die (Fig. 17.3.4) indicate that the Razor read mode RR/2W always improves the maximum operating frequency compared to when it is disabled (2R/2W), and its advantage increases at lower voltages due to the larger variations (Fig. 17.3.4). At 0.7V (1V), the Razor SRAM average speed-up is 1.73× (1.16×) due to timing speculation. From Fig. 17.3.4, the clock cycle advantage of Razor SRAM reduces by 2.79× (1.8×) on average, when compared to the clock cycle target of a guard-banded SRAM design. The guard-banded design accounts for 3σ timing variation due to process variation (σ empirically measured over 20 chips, see Fig. 17.3.4), a 10% voltage drop, and temperature corners (worst among –20 and 85°C). Note that comparisons for guard-banded SRAM are made at 0.7V rather than 0.6V since the very large PVT margins in the latter case lead to impractically large performance margins. In the case of 0.6V, Razor SRAM provides even larger potential savings via margin elimination. The merged-wordline scheme (2R/1W) speeds up write time by 3.7× at 0.6V since it improves the severely degraded write-ability at low voltages, and hence shortens the time needed to perform a correct write.

Figure 17.3.5 shows the speed-up of Razor SRAM compared to the non-guard-banded SRAM at 0.6V, as they can operate correctly at up to 190MHz and 40MHz, respectively. The power consumption is approximately the same for both cases, due to the small Razor power overhead. Similar considerations hold at 0.7V, although the speed-up of the Razor SRAM is less pronounced due to smaller variation. The energy-throughput curve in Fig. 17.3.5 shows that the Razor SRAM offers 57% energy reduction at iso-throughput compared to the guard-banded SRAM, thanks to the speed-up enabled by the Razor mode when aggressively scaling the supply voltage.

Figure 17.3.6 shows that Razor SRAM substantially improves throughput compared to an SRAM designed with the previously mentioned PVT margins. At 0.7V, the average throughput gain of Razor SRAM is 2.7×, 4.48×, and 5.1× while accounting for P, PV, and PVT variations, respectively. Guard band elimination enables a throughput improvement by 5.6×, 5.2×, and 3.5× for the best, average, and worst dies at 0.7V. In terms of energy efficiency, the proposed Razor SRAM enables up to 57% improvement under aggressive voltage scaling at iso-performance with 6.8% area overhead.

Acknowledgements:
The authors gratefully acknowledge STMicroelectronics for chip fabrication.

References:
[1] N. Planes et al., "28nm FDSOI technology platform for high-speed low-voltage digital applications," *IEEE Symp. VLSI Circuits*, pp. 133-134, Jun. 2012.
[2] E. Karl, et al., "A 0.6V 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology," *ISSCC Dig. Tech. Papers*, Feb. 2015.
[3] J. P. Kulkarni et al, "A 409GOPS/W Adaptive and Resilient Domino Register File in 22nm Tri-Gate CMOS Featuring In-Situ Timing Margin and Error Detection for Tolerance to Within-Die Variation, Voltage Droop, Temperature and Aging," *ISSCC Dig. Tech. Papers*, Feb. 2015.
[4] E. Karl et al., "A 4.6GHz 162Mb SRAM design in 22nm trigate CMOS technology with integrated active VMIN enhancing assist circuitry," *ISSCC Dig. Tech. Papers*, pp. 230-232, Feb. 2012.
[5] F. Frustaci, et al, "A 32kb SRAM for error-free and error-tolerant applications with dynamic energy-quality management in 28nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 244-245, Feb. 2014.
[6] S. Das et al, "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance," *IEEE J. Solid-State Circuits*, vol. 4, no. 1, pp. 32-48, Jan. 2009.

 978-1-4673-9467-3/16/$31.00

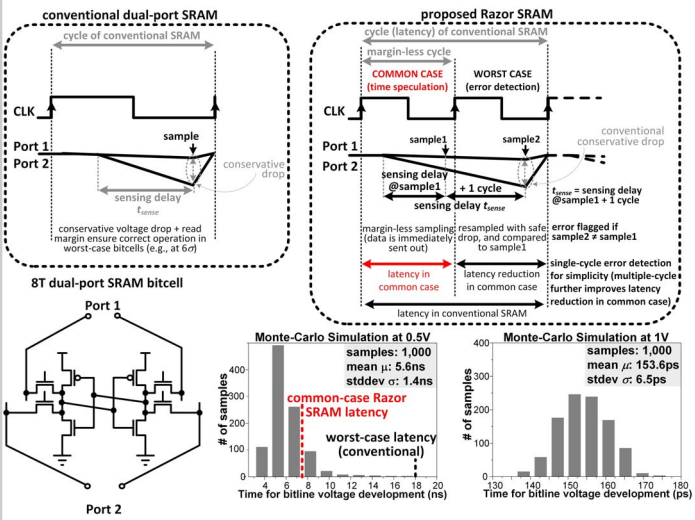

Figure 17.3.1: (l) 2-port SRAM cell and read timing. (r) Razor SRAM sensing BL at margin-less (speculative) and worst-case read instants.

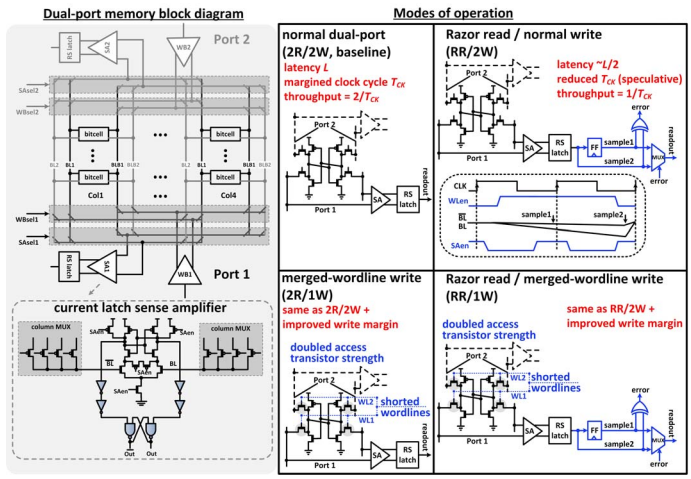

Figure 17.3.2: Razor SRAM array organization (top left) and sense amplifier circuit (bottom left). SRAM configurations are illustrated in four different operation modes (right).

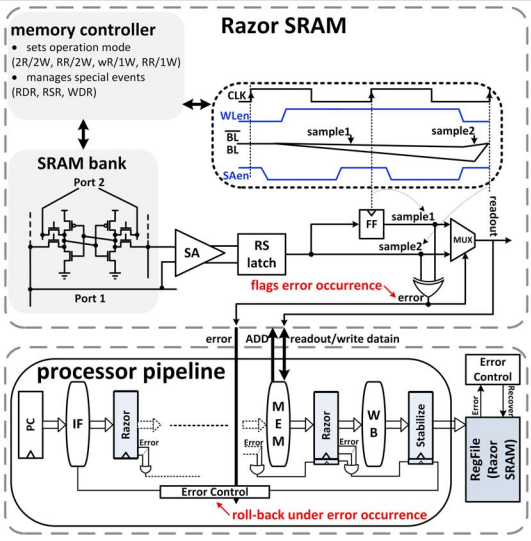

Figure 17.3.3: Top view of Razor SRAM (top) and Razor-class processor pipeline (bottom).

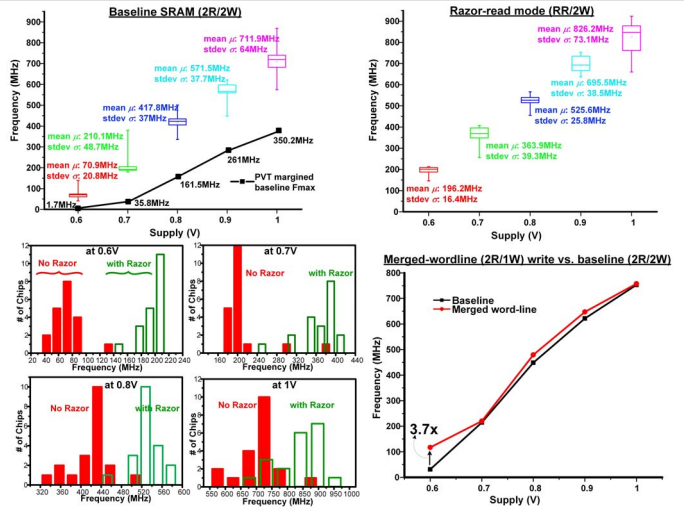

Figure 17.3.4: Shmoo plot (20 dice): Fmax of baseline and RR mode (top). Fmax histograms at different voltages (bottom left). Merged-wordline scheme speed improvement at low voltage (bottom right).

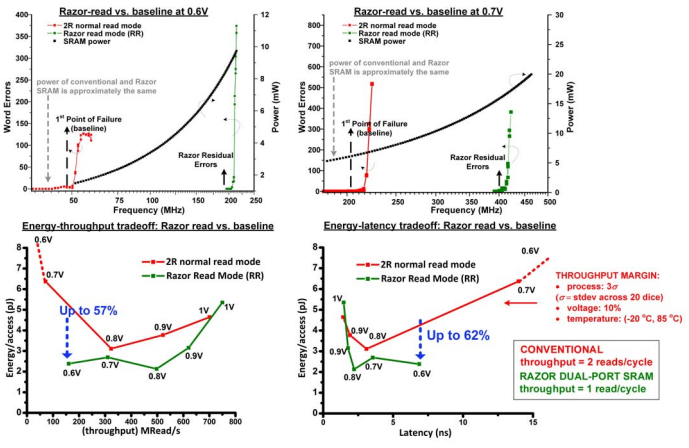

Figure 17.3.5: Read access measurements: number of word errors and power versus frequency in non-margined baseline and RR mode at 0.6V and 0.7V (top). Razor energy reduction iso-throughput (bottom).

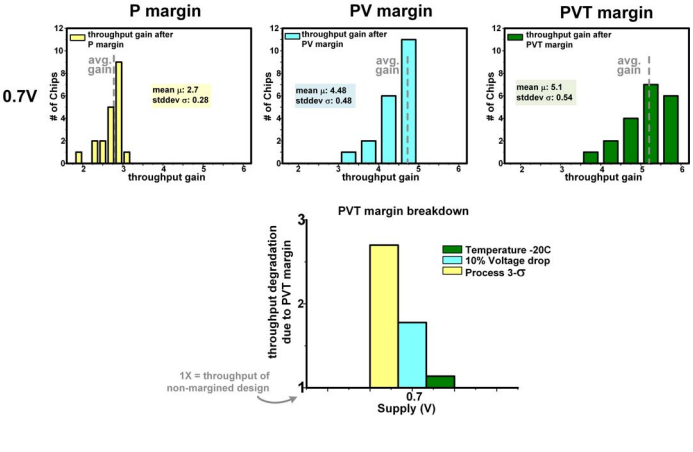

Figure 17.3.6: Distributions measured across 20 dice at room temperature, 0.7V (top): throughput gain of Razor read compared to baseline margined for P (left), PV (center) and PVT (right). PVT guardband breakdown is shown (bottom).

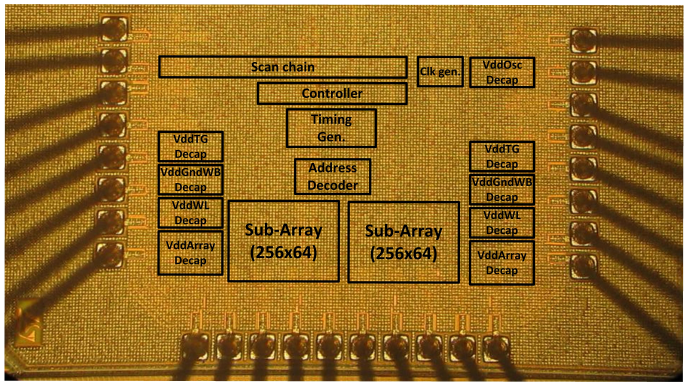

Figure 17.3.7: Die micrograph.

978-1-4673-9467-3/16/$31.00