Circuit and architectural techniques for minimum-energy operation of SRAM-based caches

Brian Zimmer, Pi-Feng Chiu, Krste Asanović and Borivoje Nikolić
University of California, Berkeley, CA, USA

Abstract – **To continue reducing voltage in scaled technologies, both circuit and architecture-level resiliency techniques are needed to tolerate process-induced defects, variation, and aging in SRAM cells. Many different resiliency schemes have been proposed and evaluated, but most prior results focus on voltage reduction instead of energy reduction. At the circuit level, device cell architectures and assist techniques have been shown to lower V_{min} for SRAM, while at the architecture level, redundancy and cache disable techniques have been used to improve resiliency at low voltages. This paper presents a unified study of error tolerance for both circuit and architecture techniques and estimates their area and energy overheads. Optimal techniques are selected by evaluating both the error-correcting abilities at low supplies and the overheads of each technique in a 28nm. The results can be applied to many of the emerging memory technologies.**

Keywords— SRAM, V_{min}, Low Power, Low Voltage, Cache, Processors.

I. SUMMARY

Improving energy efficiency is critical to improving computing capability, from mobile devices operating with limited battery capacity to server designs operating under thermal constraints. A common technique in modern systems is the use of dynamic voltage-frequency scaling (DVFS) to trade performance for energy efficiency. Lowering supply voltage improves energy efficiency, which reaches a maximum at a supply near the threshold voltage in CMOS technology. Operation at low supply voltages, however, reduces design margins and increases the probability of errors. SRAM bitcells are the most sensitive to errors because their small size increases the impact of random variations, and the large number of cells in typical systems increases the likelihood of extreme variations. Hence the recent upswell of interest in solutions that cope with the unreliability of SRAM cells at low voltages to increase the available energy-efficiency improvements from DVFS.

The source of an error can be broadly categorized as either hard or soft. Hard faults can occur during manufacturing or appear as the system ages and are generally permanent. Soft errors happen rarely to random devices and are generally transient. It is important to distinguish between hard faults and soft errors as different modeling and mitigation techniques are required for each category. Hard faults define the yield of a system while soft errors define the failures-in-time (FIT) of a system. These error metrics, yield and FIT, are treated as constraints on acceptable system designs. Because of process variation, an SRAM cell might work at a higher voltage but fail at a lower voltage, so yield is also parameterized by operating point. The probability that a bitcell fails increases exponentially with reducing voltage, and the minimum operating voltage of a system with acceptable SRAM yield is referred to as V_{min}.

Resiliency refers to the ability to tolerate process variation and prevent device non-idealities from causing system failure. Many different resiliency schemes that have been proposed both at the circuit level and the microarchitecture level. Circuit-level techniques, such as assist circuits that change wordline [2], bitline [3] [4], or cell supply voltages [1] on a cycle-by-cycle basis to strengthen or weaken particular devices during each operation, have been shown to significantly reduce Vmin. However, circuit-level techniques must be re-evaluated for each process node, and do not entirely eliminate failures. Architecture-level resiliency techniques use redundancy or error correction to repair or avoid bitcell failures. Redundancy-based techniques guarantee working memory cells to compensate for failing cells. Manufacturing faults are commonly handled with row or column redundancy that can correct a few cells per SRAM array [11]. To lower V_{min}, many proposed microarchitecture-level schemes attempt to identify cells that fail at lower voltages, and then keep the cache working at lower-voltage operating points by reconfiguring the cache to avoid these failing cells, albeit with reduced capacity [5], [6], [7]. Error-correction-based techniques encode data with extra bits that are used to detect and correct bit flips when they occur [8], [9], [10], [12].

In general, existing studies have focused on a single layer of abstraction, comparing circuit solutions to other circuit solutions and architecture techniques to other architecture techniques (with a few exceptions, such as [12]). Accurately accounting for interactions between circuit and architecture-level techniques is critical. Circuit-level requirements can be significantly relaxed if a small amount of redundancy is assumed, and the effectiveness of architecture-level techniques depends on the sensitivity of bitcell error rate to voltage.

In this paper, we compare and analyze many prior resiliency schemes using a new holistic error model, and consider the effects of circuit-level design assumptions on the results.

This paper adds to the existing body of work in four main categories.

1. A generic error model is proposed that can intuitively evaluate many of resiliency techniques with common evaluation metrics and assumptions. The proposed hierarchical combination of binomial distributions can accurately quantify the effectiveness of a wide variety of resiliency schemes.
2. The sensitivity of SRAM error events to circuit-level assist techniques is summarized in a unified fashion, for both 6T and 8T cells.
3. The sensitivity of microarchitecture-level resiliency techniques to circuit-level assumptions is analyzed. We show how previous works used a particular dataset that is very optimistic about the benefits of voltage scaling and leads to aggressive design points that require resiliency techniques with unnecessarily high complexity.
4. Previously published techniques are evaluated and compared for energy-efficiency improvements and implementation overheads.
5. Common trends from the analysis are summarized in the form of generic design guidelines for resilient cache design.

The analysis is performed assuming the 28nm CMOS technology, but is applicable to finer technology nodes as well. A general framework of analyzing the impact of architectural techniques on array resiliency is applicable to a broad range of emerging memory technologies.

References

[1] E. Karl, Y. Wang, Y.-G. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mis- try, and M. Bohr, “A 4.6 GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active Vmin-enhancing assist circuitry,” in Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International, pp. 230–232, IEEE, 2012.

[2] M. Sinangil, H. Mair, and A. Chandrakasan, “A 28nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6V,” in Int. Solid-State Circuits Conf. Dig. Tech. Papers, pp. 260–262, 2011.

[3] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, “A 45nm 0.6V cross-point 8T SRAM with negative biased read/write assist,” IEEE Symp. VLSI Circuits Dig., 2009.

[4] H. Pilo, I. Arsovski, K. Batson, G. Braceras, J. Gabric, R. Houle, S. Lamphier, C. Radens, and A. Seferagic, “A 64 Mb SRAM in 32 nm High-k Metal-Gate SOI Technology With 0.7 V Operation Enabled by Stability, Write-Ability and Read-Ability Enhancements,” IEEE J. Solid-State Circuits, vol. 47, pp. 97 –106, Jan. 2012.

[5] T. Mahmood, S. Kim, and S. Hong, “Macho: A failure model-oriented adaptive cache architecture to enable near-threshold voltage scaling,” in Proceedings of the 2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA), pp. 532–541, 2013.

[6] C. Wilkerson, H. Gao, A. R. Alameldeen, Z. Chishti, M. Khellah, and S.-L. Lu, “Trading off cache capacity for reliability to enable low voltage operation,” in Computer Architecture, 2008. ISCA’08. 35th International Symposium on, pp. 203–214, IEEE, 2008.

[7] J. Abella, J. Carretero, P. Chaparro, X. Vera, and A. González, “Low Vcc_{min} fault-tolerant cache with highly predictable performance,” in Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 42, pp. 111–121, 2009.

[8] Z. Chishti, A. R. Alameldeen, C. Wilkerson, W. Wu, and S.-L. Lu, “Improving cache lifetime reliability at ultra-low voltages,” in Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 42, pp. 89–99, 2009.

[9] A. R. Alameldeen, I. Wagner, Z. Chishti, W. Wu, C. Wilkerson, and S.-L. Lu, “Energy-efficient cache design using variable-strength error-correcting codes,” in Proceedings of the 38th annual International Symposium on Computer Architecture, ISCA ’11, pp. 461– 472, 2011.

[10] R. Naseer and J. Draper, “DEC ECC design to improve memory reliability in sub-100nm technologies,” in Electronics, Circuits and Systems, 2008. ICECS 2008. 15th IEEE International Conference on, pp. 586–589, IEEE, 2008.

[11] A.Ohba, S.Ohbayashi, T.Shiomi, S.Takano, K.Anami, H.Honda, Y.Ishigaki, M.Hatanaka, S. Nagao and S. Kayano, “A 7-ns 1-Mb BiCMOS ECL SRAM with shift redundancy,” IEEE J. Solid-State Circuits, vol. 26, pp. 507–512, Apr. 1991.

[12] N. S. Kim, S. C. Draper, S.-T. Zhou, S. Katariya, H. R. Ghasemi, and T. Park, “Analyzing the Impact of Joint Optimization of Cell Size, Redundancy, and ECC on Low-Voltage SRAM Array Total Area,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 20, no. 12, pp. 2333–2337, 2012.