Enhancing Performance and Energy Consumption of HER Caches by Adding Associativity

Vicente Lorente[1], Alejandro Valero[1], and Ramon Canal[2]

[1] Department of Computer Engineering
Universitat Politècnica de València
Valencia, Spain
[2] Departament d'Arquitectura de Computadors
Universitat Politècnica de Catalunya
Barcelona, Spain

Abstract. Unlike other previous techniques, the recently proposed Hard Error Recovery (HER) fault-tolerant cache provides 100% fault-coverage in L1 data caches. This full coverage makes the HER cache appropiate for fault-dominated future technology nodes.

An n-way set-associative HER cache implements one cache way with fast SRAM banks and the remaining ways with eDRAM banks to address power and area. Since the number of eDRAM cache blocks used in a specific HER cache organization depends on the cache associativity (i.e., the implemented number of ways), we expect that the performance and energy consumption provided by a given HER cache design strongly depends on the cache geometry.

In this work we study the behavior of the HER cache design when applied to a highly associative L1 cache like those found in some modern microprocessors. In particular this work explores a 32KB 8-way associative L1 data cache such as the one used in Intel Haswell microarchitecture.

Experimental results show that, at low-power modes compared to a conventional cache with the same storage capacity and number of ways, area, leakage power, and dynamic energy savings of a 4-way HER cache are by 25%, 85%, and 62%, respectively. These percentages are further improved (by 40%, 89%, and 68%, respectively) when the cache associativity is increased to 8 ways, while the performance loss with respect to both an 8-way conventional cache and the 4-way HER cache is minimal.

1 Introduction

As transistor features continue shrinking in future technologies, fabrication process variations make cache memory cells more unreliable at low voltages. If the voltage is lowered beyond a given reliable level, namely Vccmin, the probability of failure increases exponentially due to the higher sensitivity of the cell to parametric variations. The Vccmin of a given cache structure (e.g., L1 or L2 caches) is defined by the highest Vccmin of its memory cells, and it typically determines the Vccmin of the processor as a whole.

On the other hand, current microprocessors support different power modes. High-performance modes focus on speedup the workload, which requires high frequency and voltage levels. On the contrary, low-power modes work at minimum frequency/voltage levels to improve energy savings. However, a high Vccmin limits the ability to support

D. an Mey et al. (Eds.): Euro-Par 2013 Workshops, LNCS 8374, pp. 454–464, 2014.

lower power modes in order to further exploit the trade-off between performance and energy consumption.

Process variation affects in a different way the behavior of the cache memory cell depending on the technology. In fast Static Random-Access Memory (SRAM) cells, it induces Static Noise Margin (SNM) variability which causes errors [1,2,3] in some cells when working below Vccmin. Different approaches have been devised to deal with this problem [4,5,6]. Most existing proposals provide a fault-coverage (percentage of faults that can be detected/corrected) rather low (see Section 2). Since the failure probability increases as the transistors shrink, this coverage is insufficient for future technology nodes [7].

In contrast, in embedded Dynamic RAM (eDRAM) cells, which have been also used in some modern processors [8], errors basically lump into the cell retention time instead of altering the stored value. The worst case of these device variations determines the refresh period for the whole eDRAM array. In other words, eDRAM variation problems can be straightforwardly addressed by increasing the refresh rate. This fact is used in [9] to propose a Hard Error Recover (HER) cache, which combines SRAM and eDRAM technologies to deal with hard-error recovery at low-power mode while sustaining the performance at high-performance mode.

HER caches are implemented with one SRAM fast bank and *n-1* eDRAM banks, corresponding to the *n* ways in *n-way* associative cache. Since the distribution of SRAM and eDRAM cache blocks in a HER cache is defined by the cache associativity, it is expected that HER cache performance and energy consumption will be strongly affected by the specific cache geometry. In this paper, we study the behavior of the HER cache design when applied to a highly associative L1 cache like the one used in Intel Haswell microarchitecture [10].

Experimental results show that an 8-way HER cache saves area, leakage power, and dynamic energy with respect to a 4-way HER cache with the same storage capacity. Compared to a conventional cache, the 4-way cache working at low-power mode reduces area, leakage, and dynamic consumption by 25%, 85%, and 62%, respectively, whereas these percentages are up to 40%, 89%, and 68% for the 8-way HER approach. These benefits are obtained with minimal performance degradation with respect to the conventional scheme.

The rest of this paper is organized as follows. In the next section the most relevant approaches that allow the system to work below Vccmin are summarized. In section 3 a brief review of the HER cache is given. Experimental results, including area, performance and power, are presented in section 4. Finally, the authors give our conclusion in section 5.

2 Related Work

Due to inter and intra-die process parameter variations, memory cells that are marginally functional during manufacturing tests can undergo runtime failures due to voltage-thermal noise, soft errors, or aging effects. Depending on the impact of these effects, different segments of a memory array may move to different reliable design corners that can be determined using post-fabrication characterization. In [11], authors divide the cache in memory blocks and classify them into three types of blocks depending on the threshold voltage variation of their transistors (NMOS and PMOS). Then, different error correcting codes (ECC) are applied to each block according to this classification.

In [12] it is proposed an adaptive cache design that uses up to half the data array to store ECC information at low voltage to reduce energy. In high-performance mode, the whole data array is enabled. They add some hardware structures which are monitored by the operating system to select the desired reliability level. In low-power mode, some physical ways are used to store ECC information. For instance, to support *only* 4-bit error correction for each 64 bits segment at a 520mV supply voltage, the number of ways devoted to ECC can be as high as half the cache ways. In this case, performance degrades by 10% with respect to a defect-free cache.

In [5] it is presented a variation-aware cache architecture, which adaptively resizes the cache to avoid accessing to faulty blocks. When a faulty block is accessed, the bitmap information is used to select a non-faulty block in the same row. The cache implements a self-test circuitry, which tests the entire cache and detects faulty cells. Tests are conducted whenever the operating conditions change. Performance losses are by 1.5% and 5.7% in data and instruction caches, respectively.

In [4] authors propose two architectural techniques that enable microprocessor caches to operate at voltages below Vccmin. These techniques, namely Word-disable and Bit-fix, reduce the effective cache storage capacity by 50% and 25%, respectively. The word-disable scheme combines two consecutive cache lines in low voltage mode to form a single cache line without failing words. The bit-fix scheme uses a quarter of the ways to store positions and fix bits for failing bits in other ways of the set. A test is performed at boot time to identify those segments of the cache that fail at low voltage. The required circuitry increases area by 8% and reduces performance by 10%.

The Reconfigurable Energy-efficient Near Threshold (RENT) cache architecture [13] consists of a near threshold tolerant cache way and several conventional SRAM cache ways. The former way is implemented with large error resilient 8T cells to reduce energy consumption, whereas the remaining cache ways are implemented with conventional 6T SRAM cells to maintain the performance. The near threshold tolerant way is accessed first and, in case of miss, the conventional ways are then accessed. Unlike the above works, bit failures are avoided by choosing an appropriate supply voltage for each way. This scheme reduces energy by 86% at the expense of area due to the use of 8T cells.

Finally, refresh power potentially represents a large fraction of the overall system power, particularly during low-power states when the processor is idle. In [14] the cache refresh power is reduced by increasing the refresh time from 30μs (worst-case) to 440μs. This increase substantially mitigates power but causes errors due to capacitor discharges. This problem can be solved by using costly ECC codes. For instance, for the larger refresh time, authors need to apply the 5EC6ED ECC algorithm, which requires 51 extra bits for each 64B line (10% area overhead). To reduce ECC area overhead, the proposed solution uses a huge 1KB line size for the L3 cache instead of typical 64B lines. This technique reduces power by 93% compared to an eDRAM with no error correction, and by 66% compared to an eDRAM implementing single error correcting codes (SECDEC).

In general, these solutions allow the system to work below Vccmin by disabling those segments of the cache where one or more bits can fail, thus reducing the effective storage capacity. The highest fault coverage is achieved by [11], which can correct up to 6 bit-errors for each 64-bit block. Nevertheless, this is still less than 10% coverage. Providing higher coverages in these proposals could become prohibitive in terms of area, delay, or power.

In short, the presented state-of-the-art proposals will not match the coverage requirements of future technologies. Moreover, in [7], it is further argumented why less than 100% fault-coverage is unsuitable for fault-dominated future technology nodes.

3 HER Caches

The HER cache combines eDRAM and SRAM technologies in order to achieve 100% fault-coverage. For the data array, the HER cache uses k cache banks to implement an n-way set-associative cache, where k/n banks are implemented with SRAM technology and the remaining $k - k/n$ with eDRAM technology.

At high-performance, the HER cache works using the entire storage capacity and architectural decisions are devised to address performance losses. To achieve good performance, in the HER cache the MRU block of each cache set is always stored in an SRAM bank. Moreover, HER caches are designed with no refresh mechanism to reduce energy overhead. Instead, architectural mechanisms are introduced to ensure that no useful data is lost due to capacitor discharges.

At low-power, HER caches provide 100% fault-coverage in set-associative L1 data caches while reducing area and power with respect to a conventional cache. This is achieved by using eDRAM banks to keep copies of SRAM banks data (i.e. replicas). As process variation impacts eDRAM cells retention time but does not change their contents, these replicas enable the processor to recover from any number of SRAM bit failures due to manufacturing imperfections. In consequence, HER caches support 100% fault-coverage of errors due to process variation imperfections, which is a major design concern to be addressed in future technology nodes.

More details about HER caches rationale and implementation can be found in [9].

4 Experimental Results

The HER cache has been modeled on top of the SimpleScalar (with Alpha ISA) simulation framework [15] to obtain the execution time and memory events required for estimating dynamic energy (i.e., cache hits, misses, swaps, writebacks, etc.). All the bank contention induced by these events has also been modeled. The 8-way HER cache has a single cache bank built with SRAM technology, while the remaining banks are implemented with eDRAM. Cache access time, capacitance, area, leakage power, and dynamic energy per access type were estimated with CACTI 5.3 [16] for 45nm. The overall dynamic energy was calculated combining the results of both simulators.

All these results have been obtained for each operation mode, from now on referred to as *hp* (high-performance) and *lp* (low-power) modes, respectively. For the *hp* mode, it has been assumed a voltage/frequency pair of 1.3V/3GHz and no bit failures. Two

Table 1. Architectural machine parameters

Microprocessor core	
Issue policy	Out of order
Branch predictor type	Hybrid gShare/Bimodal: gShare has 14-bit global history plus 16K 2-bit counters Bimodal has 4K 2-bit counters Choice predictor has 4K 2-bit counters
Branch predictor penalty	10 cycles
Fetch, issue, commit width	4 instructions/cycle
ROB size (entries)	256
Memory hierarchy	
L1 data cache	32KB, 4-way and 8-way, 64 byte-line
L1 data cache access time	SRAM: 2-cycle in *hp*, 1-cycle in *lp* eDRAM: 4-cycle in *hp*, 2-cycle in *lp*
L2 data cache	512KB-8way, 64 byte-line, 10-cycle
Main memory access time	100-cycle

different voltage/frequency pairs have been studied in *lp* mode, referred to as *lp1* and *lp2*, respectively. The former assumes 0.7V/1.4GHz and the latter 0.5V/500MHz similar to [12]. The probability of failure for a 45nm node was calculated as 20% and 50% of the SRAM bits, respectively. Although some defective bits can be in the same line, the results presented assume the worst case, that is, all defective bits are located in different cache lines.

For comparison purposes, a conventional SRAM cache has also been modeled. In this scheme, no error failures are assumed regardless of the operation mode. At *hp* mode, the scheme is referenced as Conv cache and at *lp* as ZfConv (zero-failure conventional) cache.

Table 1 summarizes the main architectural parameters. Both integer (Int) and floating-point (FP) SPEC benchmarks [17] were run using the *ref* input sets. After skipping the initial 1B instructions, statistics were collected during 500M instructions.

4.1 Area

This section analyzes the area savings of the HER scheme compared to a conventional SRAM cache of the same storage capacity and number of ways. Table 2 shows the results (in mm^2) for a 45nm technology node, where SRAM and eDRAM cells occupy

Table 2. Tag array and data array areas (in mm^2) of both Conv cache and HER cache schemes

Cache	Tag array (mm^2)		Data array (mm^2)		Data array area	Total area
ways	Conv	HER	Conv	HER	savings (%)	savings (%)
4	0.015	0.030	0.498	0.355	29	25
8	0.015	0.030	0.850	0.490	42	40

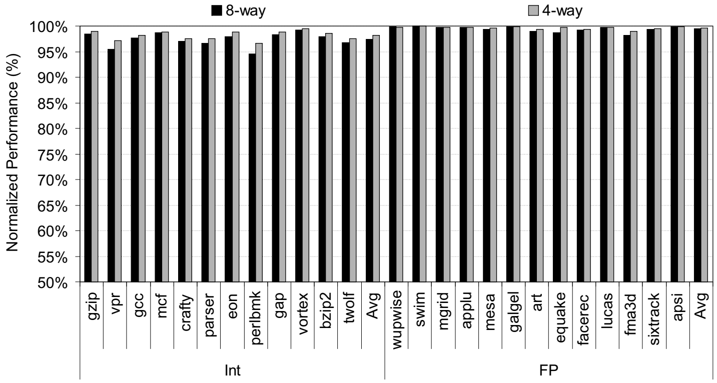


Fig. 1. Normalized performance (%) of the HER caches with respect to the 8-way Conv cache in *hp* mode

an area of $0.296\mu m^2$ and $0.062\mu m^2$, respectively, according to CACTI. The HER cache significantly reduces area thanks to the design is partly implemented with eDRAM technology. As expected, this reduction increases with the cache associativity. Compared to the Conv approach, the area reduction of the data array can be as much as 29% and 42% for 4-way and 8-way HER caches, respectively.

On the other hand, the HER cache uses special error-free SRAM cells to build the tag array, whose size is twice as large as the size of conventional SRAM cells [18]. Even considering both tag and data arrays, area savings are as large as 40% for the 8-way cache.

4.2 Performance Evaluation

Figure 1 shows, for each application, the normalized performance in *hp* mode of HER caches implemented with an *infinite* retention time (i.e. perfect capacitors) with respect to an 8-way conventional cache with the same storage capacity. As observed, the IPC loss is higher in the 8-way HER cache because it stores more blocks in *slow* eDRAM technology. Nevertheless, the performance degradation for Int benchmarks is on average only by 2.63% with respect to the Conv cache; and much lower (by 0.58%) for FP benchmarks. These percentages are by 1.93% and 0.38%, respectively, for the 4-way HER cache.

As real capacitors lose their contents over time, we calculated (as done in [9] for the 4-way HER cache) which is the minimum capacitance to match the performance of HER caches with an infinite retention time. In 8-way HER caches, at least 4fF, 14fF, and 62fF capacitors are required in *hp*, *lp1*, and *lp2* modes, respectively. Unfortunately, current trench capacitors only allow capacitances up to 30fF [19,20]. However, this limitation only affects *lp2* mode, and our experiments show that, compared to an 8-way ZfConv cache, the performance degradation is by 2.69% for Int benchmarks. For the 4-way HER cache, this percentage is by 2.54%.

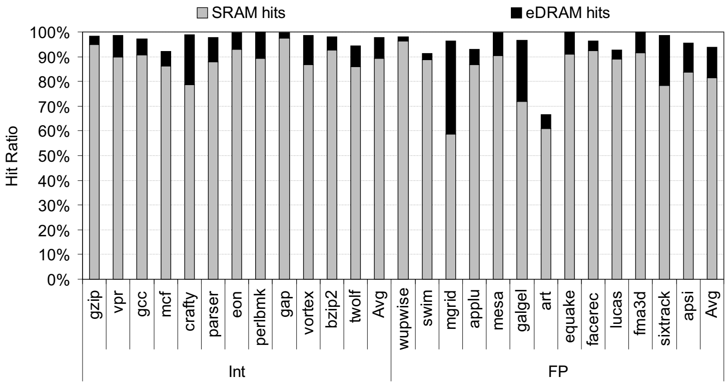


Fig. 2. Hit ratio (%) split into SRAM and eDRAM hits per benchmark for the HER cache in *hp* mode

Table 3. Hit ratio (%) broken down into SRAM, eDRAM, and eDRAM replica hits for the cache schemes across the studied operation modes

Bench. type	Hit ratio (%)	8-way				4-way			
		Conv	HER			Conv	HER		
		hp	*hp*	*lp1*	*lp2*	*hp*	*hp*	*lp1*	*lp2*
Int	SRAM	97.9	89.4	71.6	44.6	97.8	92.6	73.4	42.1
	eDRAM	–	8.5	8.2	8.1	–	5.2	4.7	4.7
	eDRAM replica	–	–	17.9	45.1	–	–	19.3	50.7
	total	97.9	97.9	97.8	97.8	97.8	97.8	97.5	97.5
FP	SRAM	93.7	81.6	66.0	41.2	93.7	86.7	68.8	41.9
	eDRAM	–	12.1	11.9	12.0	–	7.0	6.0	6.0
	eDRAM replica	–	–	15.7	40.4	–	–	17.9	44.8
	total	93.7	93.7	93.6	93.6	93.7	93.7	92.7	92.7

Figure 2 plots the cache hit ratio for each application in *hp* mode for the 8-way HER cache with an infinite retention time. The hit ratio has been split into hits in the SRAM and eDRAM banks. Most cache accesses (by 89% and 82% on average for Int and FP, respectively) are performed in the fast SRAM bank that holds the MRU data, so confirming the effectiveness of the HER cache's swap mechanism even in 8-way set-associative L1 caches.

Table 3 shows the average hit ratio of both cache schemes across the studied operation modes. For *lp* modes, hits in the eDRAM replica are also presented. The eDRAM hit ratio is higher in 8-way caches than in 4-way caches since the former have more eDRAM blocks. In contrast, for *lp* modes, the percentage of hits in the eDRAM replica way is lower in 8-way caches. This is because the data that potentially needs backup (SRAM blocks) are reduced to the half in such caches. For the same reason, the impact on the overall hit ratio of losing an eDRAM way for backups of the SRAM way is less remarkable in 8-way caches.

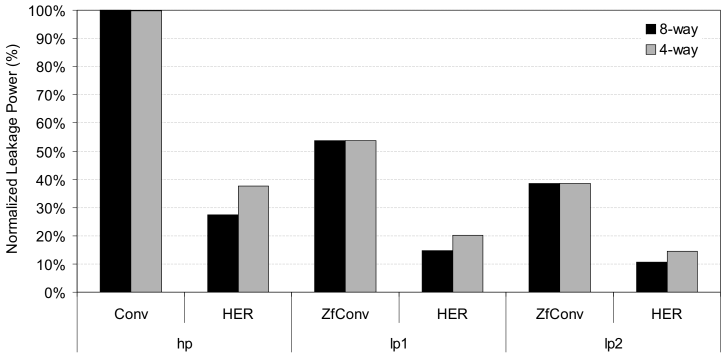


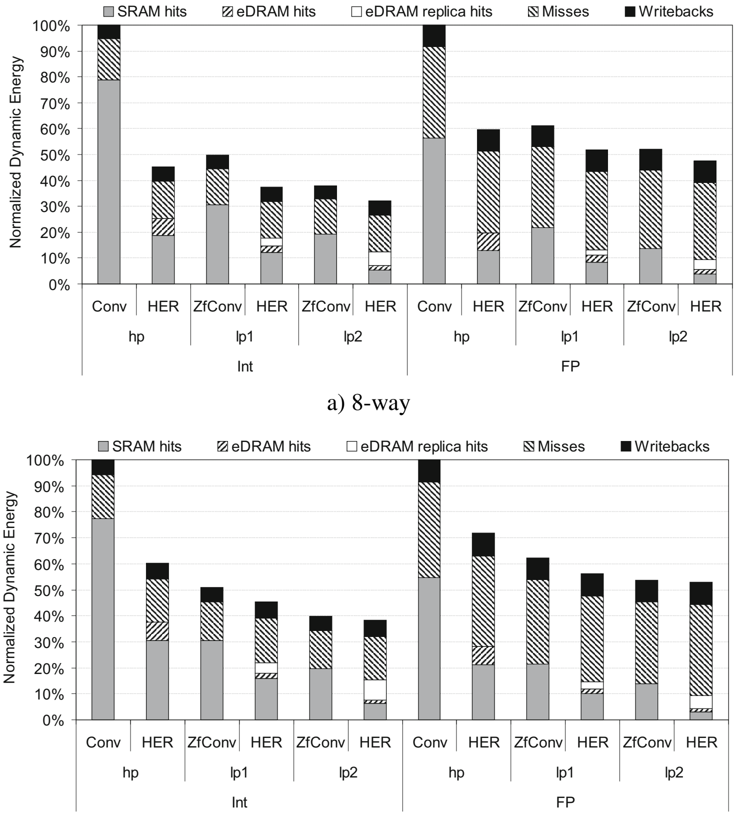
Fig. 3. Normalized leakage power (%) of the studied caches with respect to the Conv cache

4.3 Power and Energy Consumption

Figure 3 illustrates the normalized leakage with respect to the 8-way Conv approach. For HER caches, leakage currents decrease with the number of eDRAM ways since this technology minimizes leakage by design. In contrast, leakage consumption is roughly the same in both 4-way and 8-way conventional (Conv and ZfConv) approaches. This is due to these memories are only SRAM-based and they just reduce leakage because of the lower supply voltage in *lp* modes. Overall, leakage savings provided by the 8-way HER cache are as high as 73% in *hp* mode and 89% in *lp2* mode. These percentages are reduced to 62% and 85%, respectively, in the 4-way cache.

Figure 4(a) and Figure 4(b) plot the dynamic energy for 8-way and 4-way caches, respectively, normalized with respect to a conventional cache with the same storage capacity and number of ways. This consumption has been divided into five categories: SRAM hits, eDRAM hits, eDRAM replica hits, misses, and writebacks. The SRAM hits category includes the access to the eDRAM replica and the access to all the SRAM ways; the eDRAM hits category includes accessing both the SRAM way and the target eDRAM way. In addition, it also considers the energy due to swaps (unidirectional transfers in *lp* mode); the eDRAM replica hits category also takes into account the access to the SRAM faulty lines; the misses category includes not only the access to the SRAM part but also the unidirectional transfers in *hp* mode; and finally, the two latter categories include the energy consumed by both L1 and L2 cache accesses.

As can be seen, for a given operation mode and type of benchmark, the 8-way HER cache reduces the expenses in the SRAM hits category with respect to the 4-way HER cache. This is mainly due to the former accesses to a smaller part of the data array. The results in both eDRAM replica hits and misses category can be similarly reasoned. In addition, the energy due to misses is higher in 4-way caches because the number of such cache events decreases with the associativity. Minor differences can be seen in the writebacks category. In contrast, the consumption related to eDRAM hits increases in 8-way caches. This is due to in such caches more accesses are performed in eDRAM data, which in turn trigger swap operations. Nevertheless, the overall dynamic savings are larger in 8-way caches than in 4-way caches. For integer applications in *hp* mode,



a) 8-way

b) 4-way

Fig. 4. Normalized dynamic energy (%) *categorized* with respect to the Conv cache

these benefits are by 55% with respect to Conv, while this percentage drops down to 40% for 4-way caches. In *lp2* mode, the dynamic energy reduction is by 66% and 62% in 8-way and 4-way HER caches, respectively.

5 Conclusions

In order to study the impact on performance and power consumption of the cache geometry on HER cache designs, this paper extends the work in [9] by increasing the associativity of the HER cache scheme.

Results show that in low-power mode, when compared to a 32KB-4way HER cache, a 32KB-8way HER cache reduces leakage and dynamic energy by 4.7% and 9.6%, respectively. When compared to a 32KB-4way low-power conventional SRAM cache, these values are by 89% and 68%. These energy savings are mainly due to the higher percentage of eDRAM banks present in the 8-way HER design.

In high-performance mode, when compared to the conventional SRAM cache, the performance losses of an 8-way HER cache are by 2.63% and 0.58%, for integer and floating-point benchmarks, respectively. These results are just a little worse than those of a 4-way HER cache (1.93% and 0.38%).

Finally, this paper also includes an analysis of the area savings provided by the 8-way HER cache. With respect to the conventional SRAM cache, it achieves by 42% data array area savings.

Acknowledgments. This work was supported by Generalitat de Catalunya (2009SGR1250), by FP7 program of the European Commission (TRAMS-248789), by Spanish Ministerio de Economía y Competitividad (MINECO) and by FEDER funds under Grant TIN2012-38341-C04-01 and TIN2010-18368.

References

1. Bhavnagarwala, A.J., et al.: The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability. IEEE Journal of Solid-State Circuits 36(4), 658–665 (2001)
2. Mukhopadhyay, S., et al.: Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 24(12), 1859–1880 (2005)
3. Shirvani, P.P., McCluskey, E.J.: PADded Cache: A New Fault-Tolerance Technique for Cache Memories. In: Proceedings of the 17th IEEE VLSI Test Symposium, pp. 440–445 (1999)
4. Wilkerson, C., et al.: Trading off Cache Capacity for Reliability to Enable Low Voltage Operation. In: Proceedings of the 35th Annual International Symposium on Computer Architecture, pp. 203–214 (2008)
5. Agarwal, A., et al.: Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture. IEEE Journal of Solid-State Circuits 40(9), 1804–1814 (2005)
6. Ansari, A., et al.: Archipelago: A Polymorphic Cache Design for Enabling Robust Near-Threshold Operation. In: Proceedings of the 17th International Symposium on High Performance Computer Architecture, pp. 539–550 (2011)
7. Nomura, S., et al.: Sampling + DMR: Practical and Low-overhead Permanent Fault Detection. In: Proceedings of the 38th Annual International Symposium on Computer Architecture, pp. 201–212 (2011)
8. Sinharoy, B., et al.: IBM POWER7 multicore server processor. IBM Journal of Research and Development 55(3) (2011)
9. Lorente, V., et al.: Combining RAM technologies for hard-error recovery in L1 data caches working at very-low power modes. In: Proceedings of the Design, Automation, and Test in Europe Conference, pp. 83–88 (2013)
10. Kanter, D.: Intel's Haswell CPU Microarchitecture, "Real World Technologies" (November 13, 2012), `http://www.realworldtech.com/haswell-cpu/`
11. Paul, S., et al.: Reliability-Driven ECC Allocation for Multiple Bit Error Resilience in Processor Cache. IEEE Transactions on Computers 60(1), 20–34 (2011)
12. Alameldeen, A.R., et al.: Adaptive Cache Design to Enable Reliable Low-Voltage Operation. IEEE Transactions on Computers 60, 50–63 (2011)
13. Dreslinski, R.G., et al.: Reconfigurable Energy Efficient Near Threshold Cache Architectures. In: Proceedings of the 41st Annual IEEE/ACM International Symposium on Microarchitecture, pp. 459–470 (2008)

14. Wilkerson, C., et al.: Reducing Cache Power with Low-Cost, Multi-bit Error-Correcting Codes. In: Proceedings of the 37th Annual International Symposium on Computer Architecture, pp. 83–93 (2010)
15. Burger, D., Austin, T.M.: The SimpleScalar Tool Set, Version 2.0. ACM SIGARCH Computer Architecture News 25(3), 13–25 (1997)
16. Thoziyoor, S., et al.: CACTI 5.1. Hewlett-Packard Laboratories, Palo Alto, Technical Report (2008)
17. spec2000: Standard Performance Evaluation Corporation, http://www.spec.org/cpu2000
18. Kulkarni, J.P., et al.: A 160 mV, Fully Differential, Robust Schmitt Trigger Based Subthreshold SRAM. In: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, pp. 171–176 (2007)
19. Keeth, B., et al.: DRAM Circuit Design. Fundamental and High-Speed Topics. John Wiley and Sons, Inc., Hoboken (2008)
20. Mueller, W., et al.: Challenges for the DRAM Cell Scaling to 40nm. In: IEEE International Electron Devices Meeting 4, pp. 336–339 (2005)