Extra Bits on SRAM and DRAM Errors - More Data from the Field

Nathan DeBardeleben*, Sean Blanchard*, Vilas Sridharan†, Sudhanva Gurumurthi‡,
Jon Stearley§, Kurt B. Ferreira§, and John Shalf¶

*Ultrascale Systems Research Center[1]
Los Alamos National Laboratory
Los Alamos, New Mexico
{ndebard, seanb}@lanl.gov

†RAS Architecture, ‡AMD Research
Advanced Micro Devices, Inc.
Boxborough, Massachusetts
{vilas.sridharan, sudhanva.gurumurthi}@amd.com

§Scalable Architectures
Sandia National Laboratories[2]
Albuquerque, New Mexico
{jrstear, kbferre}@sandia.gov

¶Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, California
{jshalf}@lbl.gov

***Abstract*—Several recent publications have shown that memory errors are common in high-performance computing systems, due to hardware faults in the memory subsystem. With exascale-class systems predicted to have 100-350x more DRAM and SRAM than current systems, these faults are predicted to become more common. Therefore, further study of the faults experienced by DRAM and SRAM is warranted. In this paper, we present a field study of DRAM and SRAM faults in Cielo, a leadership-class high-performance computing system located at Los Alamos National Laboratory.**

Our DRAM results show that vendor choice has a significant impact on fault rates. We also show that command and address parity on the DDR channel, a required feature of DDR3 memory, is beneficial to memory reliability. For SRAM, we confirm that altitude has a significant impact on SRAM fault rates, and that the majority of these SRAM faults are caused by high-energy particles. We also show, contrary to what might be expected, that the majority of uncorrected SRAM errors are due to single-bit strikes. Finally, we examine the impact of fault and/or error rates when scaling node capacities to a potential future exascale-class systems.

I. Introduction

Recent studies have confirmed that memory errors are common in memory systems of high-performance computing systems [1], [2], [3]. Moreover, the U.S. Department of Energy (DOE) currently predicts an exascale supercomputer in the early 2020s to have between 32 and 128 petabytes of main memory, a 100x to 350x increase compared to 2012 levels [4]. Similar increases are likely in the amount of cache memory (SRAM). These systems will require comparable increases in the reliability of both SRAM and DRAM memories to maintain or improve system reliability relative to current systems. Therefore, further attention to the faults experienced by memory sub-systems is warranted. A proper understanding of hardware faults allows hardware and system architects to provision appropriate reliability mechanisms, and can affect operational procedures such as DIMM replacement policies.

In this paper, we present a study of DRAM and SRAM faults for two large leadership-class high-performance computer systems. Our primary data set comes from Cielo, an 8,500-node supercomputer located at Los Alamos National Laboratory (LANL) in the U.S. state of New Mexico. A secondary data set comes from Jaguar, a 18,688-node supercomputer that was located at Oak Ridge National Laboratory (ORNL) in Tennessee. For Cielo, our measurement interval is a 15-month period from mid-2011 through early 2013, which included 23 billion DRAM device-hours of data. For Jaguar, our measurement interval is an 11-month period from late 2009 through late 2010, which included 17.1 billion DRAM device-hours of data. Both systems were in production operation and heavily utilized during these measurement intervals.

Both systems use AMD Opteron™processors manufactured in a 45nm silicon-on-insulator (SOI) process technology. The processors employ robust error detection and correction to minimize the rate of detected, uncorrectable errors (DUE) and reduce the risk of silent data corruption (SDC).

This paper presents the following novel contributions:

- In Section III, we examine the impact of DRAM vendor and device choice on DRAM reliability [2][3]. We find that overall fault rates vary among DRAM devices by up to 4x, and transient fault rates vary by up to 7x.
- In Section IV, we show that DDR command and address parity, a required feature of DDR3 devices, dramatically increases main memory reliability. We are aware of no other work that studies the potential value of this feature.
- In Section V, we examine the impact of altitude on SRAM faults [2][3]. As expected, altitude has a significant

[1] A portion of this work was performed at the Ultrascale Systems Research Center (USRC) at Los Alamos National Laboratory, supported by the U.S. Department of Energy contract DE-FC02-06ER25750. The publication has been assigned the LANL identifier LA-UR-14-21089.

[2] Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. This document's Sandia identifier is SAND 2014-0408C.

[3] The data in Section III and Section V-A was published at SC'13. We include it in this paper because we believe it will also be of interest to the SELSE audience.

effect on the fault rate of SRAMs in the field. We also compare SRAM fault rates in the field to rates derived via accelerated testing and show that most SRAM faults are caused by high-energy particles.

- Also in Section V, we show that in 45nm (SOI) technology, most SRAM uncorrected errors result from single-bit rather than multi-bit upsets.
- In Section VI, we present an analysis of scaling fault and/or error rates to potential exascale-class systems sizes.

II. Systems Configuration

As stated previously, we examine two large-scale systems in this paper: Cielo, a supercomputer located in Los Alamos, New Mexico at around 7,300 feet in elevation; and Jaguar, a supercomputer located in Oak Ridge, Tennessee, at approximately 875 feet in elevation.

Cielo contains approximately 8,500 compute nodes. Each Cielo *node* contains two 8-core AMD Opteron processors, each with eight 512KB L2 and one 12MB L3 cache. The processors use a 45nm SOI process technology. Each Cielo compute node has eight 4GB DDR3 DIMMs for a total of 32GB of DRAM per node.

Cielo contains DRAM modules from three different vendors. We anonymize DRAM vendor information in this publication and simply refer to DRAM vendors A, B, and C. The relative compositions of these DRAM manufacturers remain constant through the lifetime of Cielo.

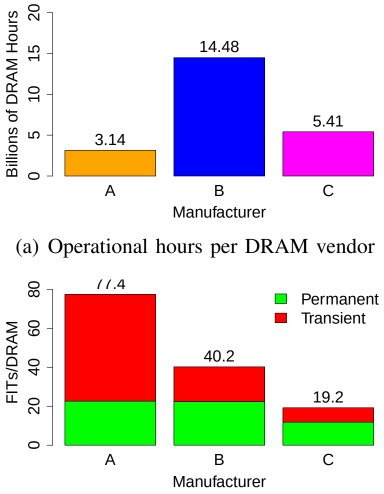
During our measurement interval, Jaguar (which was upgraded in 2012 and now is named Titan) contained 18,688 nodes. Each node contained two 6-core AMD Opteron processors, each with six 512KB L2 caches and one 6MB L3 cache, and built in 45nm SOI technology. Each Jaguar node has eight 2GB DDR2 DIMMs for a total of 16GB of DRAM per node. We do not have DRAM vendor information for Jaguar.

III. DRAM Corrected Errors

A. DRAM Vendor Effects

Figure 1(a) shows the aggregate number of DIMM-hours per DRAM vendor for Cielo during our observation period. Our observation period consists of 3.14, 14.48, and 5.41 billion device-hours for DRAM vendors A, B, and C, respectively. Therefore, we have enough operational hours on each vendor to make statistically meaningful measurements of each vendor's fault rate.

Figure 1(b) shows the fault rate experienced by each vendor during this period, divided into transient and permanent faults, meaning that errors were observed in one versus multiple scrub intervals respectively (see [2] for more details). The figure shows a substantial difference among vendors. Vendor A has a 4.0x higher fault rate than Vendor C. This figure also shows that the permanent fault rate varies by 1.9x among vendors, from 22.6 FIT to 11.8 FIT per DRAM chip, while the transient fault rate varies by more than 6x among vendors, from 54.8 FIT to 7.4 FIT per DRAM device. The figure also shows that Vendor A's transient fault rate is larger than its permanent fault rate, while the other two vendors have higher permanent than transient fault rates.



(a) Operational hours per DRAM vendor

(b) Fault rates per DRAM vendor

Fig. 1. Operational hours and fault rate by DRAM vendor.

In Cielo, more than 50% of the faults are transient. This contrasts with previous studies that pointed to permanent faults as the primary source of faults in modern DRAM [1][3]. Our data indicates that this conclusion depends heavily on the mix of DRAM vendors in the system under test. Specifically, vendor A had a higher rate of transient faults than permanent faults, but the other vendors had slightly more permanent than transient faults.

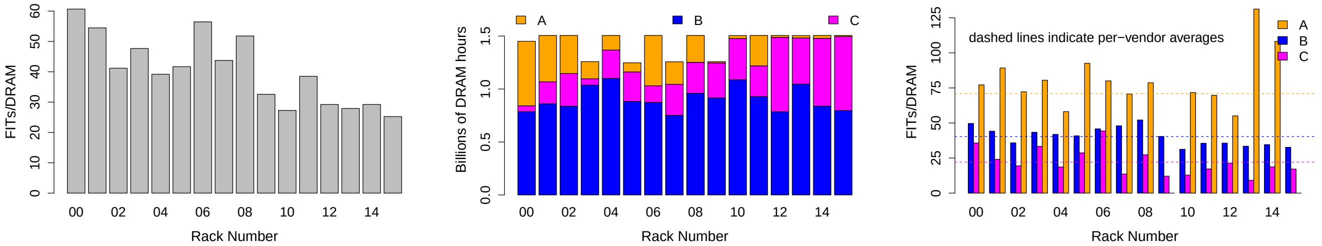
Another interesting result is that transient and permanent fault rates appear to vary together–vendors with the highest transient fault rate also have the highest permanent fault rate. It is unclear why this should be the case, but may indicate shared causes of transient and permanent faults.

B. Location in the Data Center

The physical conditions in a large machine room can vary widely. For instance: poor cooling may lead to hot spots, or an improperly installed circuit may lead to voltage spikes. The LANL data center is designed carefully and monitored heavily to minimize such effects. We examined Cielo fault data with respect to physical location to verify there were no facilities-based effects.

Most observed variances across physical location in the LANL machine room were uninteresting or statistically inconclusive. However, there is one notable exception to the lack of variance, shown in Figure 2(a). Lower-numbered racks show significantly higher DRAM fault rates than higher-numbered racks (with faults aggregated across rows). Without any further information, this trend could be attributed to temperature or other environmental differences across racks.

However, when examining operational hour data by vendor in Figure 2(b), it is clear that lower-numbered racks had significantly more operational hours from Vendor A than higher-numbered racks, which had more operational hours from Vendor C than lower-numbered racks. (Racks 3, 5, 7,



(a) Fault rate per rack

(b) Operational hours per rack

(c) Fault rate per rack by vendor

Fig. 2. DRAM fault rate positional effects by rack.

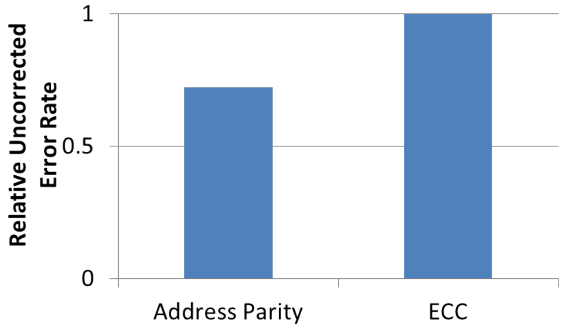


Fig. 3. Rate of DRAM address parity errors relative to uncorrected data ECC errors.

and 9 show fewer operational hours because they contain visualization nodes with different hardware, which we omitted from this study.) As shown in Figure 1, Vendor A has a higher overall fault rate than Vendor C. Therefore, racks with DRAM from Vendor A will naturally experience a higher fault rate than racks with DRAM from Vendor C. In the case of Cielo, this translates to lower-numbered racks having higher fault rates than higher-numbered racks.

When we examine the by-rack fault rates by vendor (Figure 2(c)), the by-rack fault rate trend essentially disappears. There is a slight trend in the fault rates across racks for Vendor B that is currently unexplained, but this is a very weak effect and may be due to statistical variation rather than a true effect.

IV. DDR Command and Address Parity

A key feature of DDR3 (and now DDR4) memory is the ability to add parity-check logic to the command and address bus. Though command and address parity is optional on DDR2 memory systems, we are aware of no other study that examines the potential value of this parity mechanism.

The on-DIMM register calculates parity on the received address and command pins and compares it to the received parity signal. On a mismatch, the register signals the memory controller of a parity error. The standard does not provide for retry on a detected parity error, but requires the on-DIMM register to disallow any faulty transaction from writing data to the memory, thereby preventing possible corruption.

The DDR3 sub-system in Cielo includes command and address parity checking. Figure 3 shows the rate of detected command/address parity errors relative to the rate of detected, uncorrected data ECC errors. The figure shows that the rate of command/address parity errors was 75% that of the rate of uncorrected ECC errors. Our conclusion from this data is that command/address parity is a valuable addition to the DDR standard. Furthermore, increasing DDR memory channel speeds may cause an increase in signaling-related errors. Therefore, we expect the ratio of address parity to ECC errors to increase with increased DDR frequencies.

V. SRAM Faults

In this section, we examine fault rates of SRAM. First, we compare the fault rates of Cielo and Jaguar to extract any effect caused by the $6,500$-foot difference in altitude between the two systems. Second, we compare the SRAM fault rate in Jaguar to accelerated testing and confirm that, as expected, most SRAM faults can be explained by high-energy particles.

All SRAM in both systems is from a single vendor. All errors from faults discussed in Sections V-A and V-B were corrected by the processor's internal ECC.

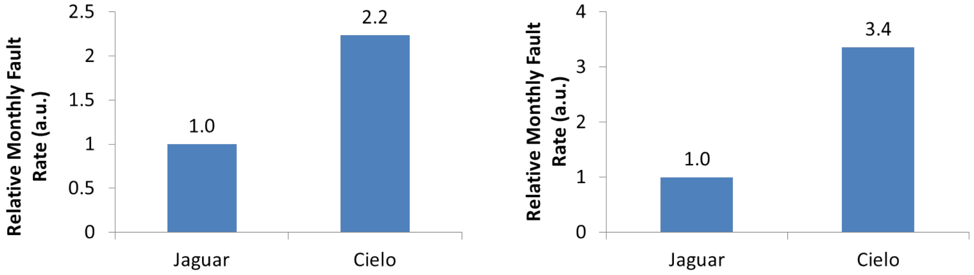
A. *Altitude Effects*

It is known that a data center's altitude has impacts on machine fault rates. The two primary causes of increased fault rates at higher altitude are reduced cooling due to lower air pressure and increased cosmic ray-induced neutron strikes. While the first can be corrected by lower machine room temperatures and higher airflow, data centers typically do not compensate for cosmic ray neutrons directly.

Figure 4 shows the average SRAM transient fault rate on Cielo relative to the average SRAM transient fault rate on Jaguar. Ninety-five percent of the data falls within one standard deviation of the mean fault rate. Figure 4 shows that Cielo experiences a 2.3x increase in the SRAM transient fault rate relative to Jaguar in L2, and a 3.4x increase relative to Jaguar in L3. The average flux ratio between LANL and ORNL without accounting for solar modulation is 4.39 [5]. Therefore, we attribute the increase in SRAM fault rates to the increase in particle flux experienced at LANL. The fact that Cielo's increase in fault rate relative to Jaguar is less than that predicted by altitude alone indicates that there may be additional sources of SRAM faults, such as alpha particles [6].

B. *Comparison to Accelerated Testing*

Accelerated particle testing is used routinely to determine the sensitivity of SRAM devices to single-event upsets from energetic particles. To be comprehensive, the accelerated testing must include all particles to which the SRAM cells will



(a) L2

(b) L3

Fig. 4. SRAM transient faults in Cielo and Jaguar (different arbitrary units for each plot).

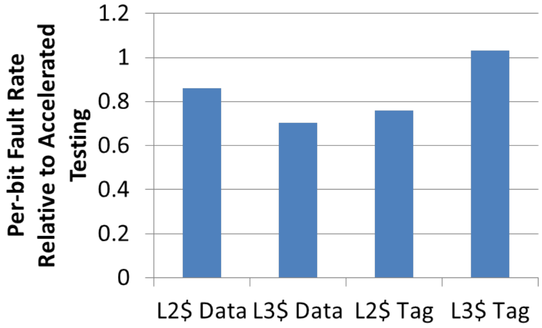


Fig. 5. Rate of SRAM faults in Jaguar compared to accelerated testing.

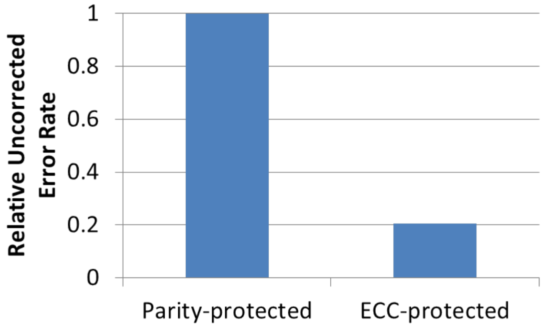


Fig. 6. Rate of SRAM uncorrected errors in Cielo from parity- and ECC-protected structures.

be exposed to in the field (e.g. neutrons, alpha particles) at rates that approximate real-world conditions. Therefore, it is important to correlate accelerated testing data with field data to ensure that the conditions are in fact similar.

Accelerated testing on the SRAM cells used in Cielo and Jaguar was performed using a variety of particle sources, including the high-energy neutron beam at LANL and an alpha particle source. This testing was “static” testing, rather than operational testing. SRAM cells were initialized with known values, exposed to the beam, and then compared to the initial state for errors.

The L2 and L3 caches employ hardware scrubbers, so we expect our field error logs to capture the majority of bit flips that occur in the L2 and L3 SRAM arrays. Figure 5 compares the per-bit rate of SRAM faults in Jaguar’s L2 and L3 data and tag arrays to results obtained from the accelerated testing campaign on the SRAM cells. The figure shows that accelerated testing predicts a higher fault rate than seen in the field in all structures except the L3 tag array. The fault rate in the L3 tag is approximately equal to the rate from accelerated testing. The per-bit fault rate seen in the field is expected to be somewhat lower than the rate seen in accelerated testing due to several effects, such as certain cases in which faults are overwritten without being detected.

Overall, the figure shows good correlation between rates measured in the field and rates measured from static SRAM testing. Our conclusion from this data is that the majority of SRAM faults in the field are caused by known high-energy particles. While an expected result, confirming expectations with field data is important to ensure that parts are functioning as specified, to identify potential new or unexpected fault modes that may not have been tested in pre-production silicon, and to ensure accelerated testing reflects reality.

C. SRAM Uncorrected Errors

In this section, we examine uncorrected errors from SRAM in Cielo and highlight some key findings. Note, the conclusions of this study are specific to current and near-future process technologies, and do not account for novel technologies, such as ultra-low-voltage operation.

Figure 6 shows the rate of SRAM uncorrected errors on Cielo, as before, in arbitrary units. The figure splits the data into two categories: uncorrected errors from parity-protected structures, and uncorrected errors from ECC-protected structures. This includes structures in the core, all caches, and a variety of non-core arrays and FIFOs. ECC-protected structures include the L2 and L3 caches, which comprise the majority of the die area in the processor.

The figure shows that the majority of uncorrected errors in Cielo came from parity-protected structures, even though these structures are far smaller than the ECC-protected structures. Parity can detect, but cannot correct, single-bit faults, while ECC can correct single-bit faults. Therefore, our primary conclusion from this study is that the majority of SRAM uncorrected errors in Cielo are the result of single-bit, rather than multi-bit, faults.

This result, combined with the results from the previous subsection, implies that the best way to reduce SRAM uncorrected error rates simply is to extend single-bit correction (e.g. ECC)

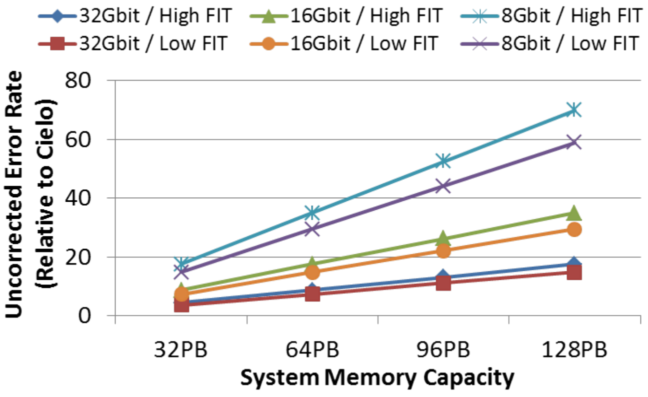


Fig. 7. Rate of DRAM uncorrected errors in an exascale supercomputer relative to Cielo.

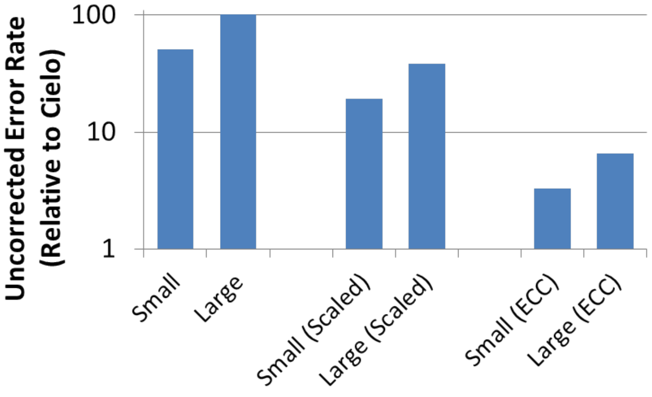


Fig. 8. Rate of SRAM uncorrected errors relative to Cielo in two different potential exascale computers.

through additional structures in the processor. While a non-trivial effort due to performance, power, and area concerns, this solution does not require extensive new research. Once single-bit faults are addressed, however, addressing the remaining multi-bit faults may be more of a challenge, especially in highly scaled process technologies in which the rate and spread of multi-bit faults may increase substantially [7].

VI. Projecting to Exascale Supercomputers

In this section, we examine the impact of DRAM and SRAM faults on a potential exascale supercomputer. Our goal is to understand whether existing reliability mechanisms can cope with the challenges presented by likely increases in system capacity and fault rates for exascale systems. To accomplish this, we scale our observed DRAM and SRAM error rates to likely exascale configurations and system sizes. We also model the impact of different process technologies such as FinFET transistors.

All analysis in this section refers to detected errors. We do not analyze any potential undetected errors. We only examine error rate scaling for currently-existing technologies and do not consider the impact of new technologies such as die-stacked DRAM or ultra-low-voltage CMOS.

A. DRAM

Exascale supercomputers are predicted to have between 32PB and 128PB of main memory. Due to capacity limitations in die-stacked devices [8], most of this memory is likely to be provided in off-chip memory, either DRAM or some type of non-volatile RAM (NVRAM). The interface to these off-chip devices is likely to resemble current DDR memory interfaces. Therefore, it is critical to understand the reliability requirements for these off-chip sub-systems.

Prior work has shown that DRAM vendors maintain an approximately constant fault rate per device across technology generations, despite reduced feature sizes and increased densities [9]. Therefore, we expect that per-device fault rates in an exascale computer will be similar to those observed in today's DRAM devices. Our goal is to determine whether current error-correcting codes (ECCs) will suffice for an exascale supercomputer. Therefore, we assume that each memory channel in an exascale supercomputer will use the same single-chipkill code in use on Cielo and Jaguar. As a result, we project that the per-device uncorrected error rate in an exascale supercomputer will be approximately the same as the per-device error rate in Cielo and Jaguar.

Because we do not know the exact DRAM device capacity that will be in mass production in the exascale timescale, however, we sweep the device capacity over a range from 8Gbit to 32Gbit devices. A larger DRAM device can deliver a specified system memory capacity with fewer total devices.

Figure 7 shows the results of this analysis. The figure plots the system-wide DRAM uncorrected error rate for an exascale system relative to the DRAM uncorrected error rate on Cielo. The figure plots two per-device error rates at each point, one taken from Jaguar data and one from Cielo data. The figure shows that uncorrected error rate for an exascale system ranges from 3.6 times Cielo's uncorrected error rate at the low end to 69.9 times Cielo's uncorrected error rate at the high end.

At a system level, the increase in DRAM uncorrected error rates at the high end of memory capacity is comparable to the 40x reduction in DRAM uncorrected errors achieved when upgrading from SEC-DED ECC to chipkill ECC [1]. If we assume that SEC-DED ECC provides insufficient reliability for today's memory subsystems, our conclusion from these results is that higher-capacity exascale systems may require stronger ECC than chipkill.

B. SRAM

A socket in Cielo contains 18MB of SRAM in the L2 and L3 cache data arrays. Exascale-class processors are projected to see a substantial increase in processing power per socket, and thus will contain significantly more SRAM per node. Taking into account technology trends and reported structure sizes for CPU and GPU processors [10], we project that an exascale socket will contain over 150MB of SRAM, or an 8-10x increase in SRAM per socket over current supercomputers. The increase in SRAM relative to today's systems is less than the increase in DRAM because of the switch to general-purpose graphical processing units (GPGPUs) and/or accelerated processing units (APUs), which rely less on large SRAM caches for performance than traditional CPU cores.

Assuming that SRAM fault rates remain constant in future years, the per-socket fault rates will increase linearly with SRAM capacity. Therefore, an exascale processor will see 8-10x the number of SRAM faults experienced by a current processor, and potentially 8-10x the rate of uncorrected SRAM errors. This translates to a system-level uncorrected error rate from SRAM errors of 50-100 times the SRAM uncorrected error rate on Cielo, depending on the system size. This is shown in the first group of bars of Figure 8, labeled "Small" for a low node-count system and "Large" for a high node-count system.

Per-bit SRAM transient fault rates have actually trended downwards in recent years [7]. If this trend continues, the SRAM uncorrected error rate per socket will be lower than our projections. For instance, according to Ibe et al. the per-bit SER decreased by 62% between 45nm and 22nm technology. If we see a corresponding decrease between current CMOS technologies and exascale technologies, the exascale system SRAM uncorrected error rate decreases to 19-39 times the uncorrected error rate on Cielo, shown in the second group of bars of Figure 8.

Finally, as noted in Section V-C, the majority of uncorrected SRAM errors are due to single-bit faults. If these errors are eliminated in an exascale processor (e.g. by replacing parity with ECC protection), the exascale system SRAM uncorrected error rate would be only 3-6.5 times Cielo's uncorrected error rates, shown in the third group of bars in Figure 8.

Our conclusion from this analysis is that, vendors should focus aggressively on reducing the rate of uncorrected errors from SRAM faults. However, large potential reductions are available through reductions in SER due to technology scaling, and much of the remainder may be possible through expanding correction of single-bit faults. Once practical limits are reached, however, more advanced techniques to reduce the rate of multi-bit faults may be needed.

VII. Conclusion

This paper presented a field study of DRAM and SRAM faults across two large high-performance computing systems. Our study resulted in several primary findings:

- We found a significant inter-vendor effect on DRAM fault rates, with fault rates varying up to 4x among vendors. This shows the importance of accounting for vendor effects in DRAM studies.
- We found that command and address parity can have a significant positive impact on DDR memory reliability, which is a result not yet shown in previous work.
- We demonstrated that SRAM faults are primarily due to high-energy particle strikes, and that SRAM uncorrected errors mostly stem from single-bit faults in 45nm SOI technology.
- We presented an analysis of scaling fault and/or error rates to potential future exascale-scale class systems.

References

[1] V. Sridharan and D. Liberty, "A study of DRAM failures in the field," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12. Los Alamitos, Calif., USA: IEEE Computer Society Press, 2012, pp. 76:1–76:11. [Online]. Available: http://dl.acm.org/citation.cfm?id=2388996.2389100

[2] V. Sridharan, J. Stearley, N. DeBardeleben, S. Blanchard, and S. Gurumurthi, "Feng shui of supercomputer memory: Positional effects in DRAM and SRAM faults," in *Proceedings of SC13: International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. New York, NY, USA: ACM, 2013, pp. 22:1–22:11. [Online]. Available: http://doi.acm.org/10.1145/2503210.2503257

[3] B. Schroeder, E. Pinheiro, and W.-D. Weber, "DRAM errors in the wild: a large-scale field study," *Commun. ACM*, vol. 54, no. 2, pp. 100–107, Feb. 2011. [Online]. Available: http://doi.acm.org/10.1145/1897816.1897844

[4] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, and K. Yelick, "Exascale computing study: Technology challenges in achieving exascale systems, Peter Kogge, editor & study lead," 2008.

[5] "Flux calculator," http://seutest.com/cgi-bin/FluxCalculator.cgi.

[6] R. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," *IEEE Transactions on Device and Materials Reliability*, vol. 5, no. 3, pp. 305–316, Sept. 2005.

[7] E. Ibe, H. Taniguchi, Y. Yahagi, K. i. Shimbo, , and T. Toba, "Impact of scaling on neutron-induced soft error in srams from a 250 nm to a 22 nm design rule," in *Electron Devices, IEEE Transactions on*, Jul. 2010, pp. 1527–1538.

[8] J. Sim, G. H. Loh, V. Sridharan, and M. O'Connor, "Resilient die-stacked dram caches," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ser. ISCA '13. New York, NY, USA: ACM, 2013, pp. 416–427. [Online]. Available: http://doi.acm.org/10.1145/2485922.2485958

[9] L. Borucki, G. Schindlbeck, and C. Slayman, "Comparison of accelerated DRAM soft error rates measured at component and system level," in *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, 2008, pp. 482–487.

[10] AMD, "AMD graphics cores next (GCN) architecture." [Online]. Available: http://www.amd.com/us/Documents/GCN_Architecture_whitepaper.pdf