“BLACKCOMB2: HARDWARE-SOFTWARE CO-DESIGN FOR NONVOLATIEL MEMORY IN EXASCALE SYSTEMS”

FINAL TECHNICAL REPORT FOR
DOE GRANT NUMBER DE-SC0012295

PERIOD OF PERFORMANCE: 06/15/14 – 06/14/17

REGENTS OF THE UNIVERSITY OF MICHIGAN

PRINCIPAL INVESTIGATOR: TREVOR MUDGE

Final Report

The Exascale supercompting program has set a goal of producing an exaFLOP-class computer within a power budget of 20MW—today's supercomputers perform in the petaFLOPS, consuming ~10MW. Exascale machines are projected to require at least 100PB of main memory capable of sustaining a bandwidth of 100PB/s (0.1B/FLOP). Main memory in today's Petascale systems consumes 30% of the total power, and projections show that a simple scaling of today's DDR3-based memory to 100PB will result in a power consumption of 52MW. Additionally, such a memory will suffer from so and hard errors so frequently that rollback will take longer than the mean-time-to-failure. DDR4, 2x improvement in efficiency which also fails to meet this target. While mobile DRAM standards like LPDDR2/3 provide larger improvements (6-7x), they offer 3-5x less bandwidth compared to DDR4, thus requiring a much larger number of chips for the same performance. This in turn increases their fault tolerance requirements. DRAM technology is offering 8Gb chips currently with at least two generations of shrinkage in the future. The success of 3D die stacking exemplified by Micron’s Hybrid Memory Cube suggest DRAM coupled with 3D die stacking is a relatively low risk option for Exascale machines compared to unproven emerging memory technologies. Accordingly, our work in the last period examining 3D die-stack DRAM memories and their challenges.

Initial Work

Our initial work on stacked DRAM focused on energy and reliability for Exascale memories. We co-optimized error resilience costs, access energy and refresh power to arrive at an energy-efficient and resilient 3D-stacked memory for Exascale computing. In addition to its area and power advantage, 3D integration allows us to stack conventional bitcells fabricated in a then-existing DRAM technology (50 nm) over a 28nm CMOS logic die. Our studies employed conservative design rules (50 vs. 20nm and 28 vs. 14nm) in part because these were the nodes for which we could get details. Also, our goal in this work was to show that Stacked DRAM limits the design risk to just the stacking technology (already demonstrated in commercial products) and is an alternative to more speculative low-power non-volatile memory technologies, as noted. In order to address power and reliability, we made the following key contributions:

1. Reduced DRAM refresh power by restructuring by restructuring subarrays to minimize bitline capacitance. In a 100PB memory built using DDR3 chips, refresh power alone can be as high as 3-4MW, consuming 20% of the total power budget for the system. Our proposed technique achieves ~5x savings in refresh power with a ~10% increase in subarray area.
2. Optimized the energy/area overhead of including stronger resilience mechanisms. We proposed Subarraykill—a fault tolerance mechanism implemented on subarrays in a DRAM bank. It protects against soft errors and hard errors (such as multi-bit faults along columns/rows and 3D technology-specific faults such as TSV failures). A novel feature is that we used rotational Single Byte Error Correction Double Byte Error Detection

(SBCDBD) ECC with 4-8b per byte to reduce these overheads instead of conventional SECDED codes. Accessing a 128b data word in a 4kb page using a (144, 128) SBCDBD2 (B=4b) ECC decoder instead of 4x (39, 32) SECDED decoders reduces access energy by 26% and check-bit storage and refresh power overheads from 21.9% to 12.5% without decreasing error coverage.

3. Include the impact of data locality on the optimal page size. Access energy primarily results from activating rows of bitcells with a RAS (Row Address Strobe). Reducing the page size decreases the energy spent per RAS by activating fewer subarrays. On the other hand, if workloads exhibit good data locality, larger pages are desirable as higher reuse of the page contents reduces the number of RASs and results in greater energy savings. We include this tradeoff in the optimization study by simulating our DRAM model with NEK5000 benchmarks representing anticipated Exascale applications. Our proposed solution was a 32Gb 3D-stacked DRAM with a page size of 4kb, access energy of 5.1pJ/bit and standby power of 0.75pW/bit. For 100PB, the total power consumption is ~4.7MW at a data bandwidth of 100PB/s. This is an improvement of ~6.5x over DDR4 DIMM-based solutions and ~1.8x over the first genera on HMC. This would leave 15MW for processors, interconnect, cooling and the other sources of power loss in an Exascale system.
4. Other results from our resiliency investigations. Our studies of resiliency have also resulted in results that apply to conventional packaging as a corollary. We summarize them below. A paper detailing our findings has been submitted to the MEMSYS conference (see uploaded papers—this work was joint with the DARPA PERFECT program). Most server-grade memory systems provide Chipkill-Correct error protection at the expense of power and/or performance overhead. In the MEMSYS paper we present low overhead schemes for reliable commodity DRAM systems that have better power and IPC performance compared to Chipkill-Correct solutions. Specifically, we propose two erasure and error correction (EECC) schemes for x8 memory systems that have 12.5% storage overhead and do not require any change in the existing memory architecture. Both schemes have superior error performance due to the use of a strong ECC code, namely, RS(36,32) over GF(28). Scheme 1 activates 18 chips per access and has stronger reliability compared to Chipkill-Correct solutions. If the location of the faulty chip is known, Scheme 1 can correct an additional random error in a second chip. Scheme 2 trades off reliability for higher energy efficiency by activating only 9 chips per access. It cannot correct random errors due to a chip failure but can detect them with 99.9986%, and once a chip is marked faulty due to persistent errors, it can correct all errors due to that chip. Synthesis results in 28nm node show that the RS (36,32) code results in a very low decoding latency that can be well hidden in commodity memory systems and, therefore, it has minimal effect on the DRAM access latency. Evaluations based on SPEC CPU 2006 sequential and multi-programmed workloads show that compared to Chipkill-Correct, the proposed Schemes 1 and 2 improve IPC by an average of 3.2% (maximum of 13.8%) and 4.8% (maximum of 31.8%) and reduce the power consumption by an average of 16.2% (maximum of 25%) and 26.8% (maximum of 36%), respectively.

Employing NV-Memory for Checkpointing

Future Exascale supercomputers are at risk of high failure rates due to the sheer number of devices they will contain. To reduce the impact of failures, checkpointing to non-volatile storage is typically employed. We investigate using NAND flash memory for checkpointing because it is a mature technology that has very high density at low cost. However, there are challenges in using NAND flash because programming it is slow and it has limited endurance. These constraints combined with the high failures rates of DRAM at the Exascale level, which necessitates frequent checkpoints, result in high checkpoint overhead and high-energy consumption. To overcome these challenges, we have developed a two-part solution to increase flash bandwidth and reduce DRAM failure rates. First, we present an architectural solution that leverages multiple levels of parallelism in flash to increase bandwidth. Using plane-level parallelism and die-level parallelism, we increase flash bandwidth by 2x and 16x, respectively. Second, we explore error-correcting algorithms for DRAM and propose a two-tiered error correction code (ECC) that reduces the DRAM failures rates significantly. Our two-part solution enables us to checkpoint faster and less frequently, leading to 63% reduction average checkpoint energy and 88% reduction in average checkpoint overhead.

We reported our findings in a recent paper as follows:

1. We investigated the design of a large memory node with several types of commodity DRAMs. We found that 3D stacked or high bandwidth DRAMs are the most energy efficient for use in future Exascale systems.
2. We studied the failure rates of DRAM and show how they lead to very short mean me between failures at the Exascale level. We present the challenges of checkpointing an Exascale system amid high failure rates and large checkpoint sizes.
3. We proposed using NAND flash to make local checkpoints in an Exascale node. We use different types of parallelism to increase the bandwidth to flash and reduce the checkpoint time.
4. We use strong ECC schemes and show that they decrease DRAM failure rates and increase the MTBF. The longer MTBF reduce the requirement for frequent checkpoints and allows checkpointing with lower energy and less me. Fewer writes to the NAND flash also extends the lifetime of flash.
5.

Our studies showed that a two-tiered ECC scheme with RS(36,32) and XCC extends the MTBF by 50Kx with only 14% additional storage overhead. Longer MTBF allows us to make less frequent checkpoints and we reduce the energy spent on checkpointing by 63% across all applications we studied. With plane-level parallelism, we gain 2x increase in bandwidth, and with die-level parallelism we gain a 16x increase in bandwidth. By checkpointing faster and less often, we reduce the overhead of checkpointing by 88%. Less frequent checkpointing and fewer writes allows us to maintain the lifetime of NAND flash up to 4 years.

Mitigating Risk

In our latest work (described in the MEMSYS paper cited above) we have pulled together ideas from the work summarized above to show that the next generation of exascale systems with hundreds of petabytes of memory can be constructed without relying on more speculative, futuristic memory technologies will be unnecessary, at least for the next generation. Building exascale supercomputers requires resilience to failing components such as processor, memory, storage, and network devices. Checkpoint/restart is a key ingredient in attaining resilience, but providing fast and reliable checkpointing is becoming more challenging as the amount of data to checkpoint and the number of components that can fail increase in exascale systems. To improve the speed of checkpointing, emerging non-volatile memory (phase change, magnetic, resistive RAM) have been proposed. However, using unproven memories to create checkpoints will only increase the design risk for exascale memory systems. In this paper, we show that exascale systems with hundreds of petabytes of memory can be constructed with commodity DRAM and SSD flash memory and that newer non-volatile memory are unnecessary, at least for the next generation. The challenge when using commodity parts is providing fast and reliable checkpointing to protect against system failures. A straightforward solution of checkpointing to local flash-based SSD devices will not work because they are endurance and performance limited. We present a checkpointing solution that employs a combination of DRAM and SSD devices. A Checkpoint Location Controller (CLC) is implemented to monitor the endurance of the SSD and the performance loss of the application and to decide dynamically whether to checkpoint to the DRAM or the SSD. The CLC improves both SSD endurance and application slowdown; but the checkpoints in DRAM are exposed to device failures. To design a reliable exascale memory, we protect the data with a low latency ECC that can correct all errors due to bit/pin/column/word faults and also detect errors due to chip failures, and we protect the checkpoint with a Chipkill-Correct level ECC that allows reliable checkpointing to the DRAM. Using our proposed approach, the SSD lifetime increases by 2×—from 3 years to 6.3 years. Furthermore, the CLC reduces the average checkpointing overhead by nearly 10× (47% from a 420% slowdown), compared to when the application always checkpointed to the SSD.