Reliability Models for Double Chipkill Detect/Correct Memory Systems

Nathan DeBardeleben,
Sean Blanchard, HPC-5;
Rakesh Kumar,
Stevenson Jian,
University of Illinois;
Vilas Sridharan, Advanced Micro Devices

Chipkill correct memories are an advanced type of error correction memories used on many modern high-performance computing (HPC) systems and high-end servers. While previous work has shown chipkill to be extremely beneficial, the existing analytical reliability models have focused on detecting and correcting a single bad symbol per codeword. However, chipkill technology that can detect and correct two bad symbols per codeword already exists. In this work, we propose a reliability model for double chipkill detect/correct memory (DCC) systems. Additionally, we present a Monte Carlo simulation that tracks very well with the analytical model. We use this to look at what this model says about meantime to error as an HPC system ages with several different memory types. We also examine how differing chipkill technologies impact the meantime to replacement of a dual in-line memory module (DIMM).

Chipkill correct is an advanced type of error correction in memory that significantly improves the reliability of memory by allowing continued memory operation in the event of device-level failures in memory. Large-scale studies show that chipkill correct reduces the uncorrectable error rate of memory by 10× [1] to 42× [2] compared to Single Error Correction, Double Error Detection (SECDEC) error correcting codes (ECC). As the size of memory continues to increase, the demand for higher reliability in memory increases as well. As a result, chipkill correct memories have become very popular among HPC systems and high-end servers with large memory capacity.

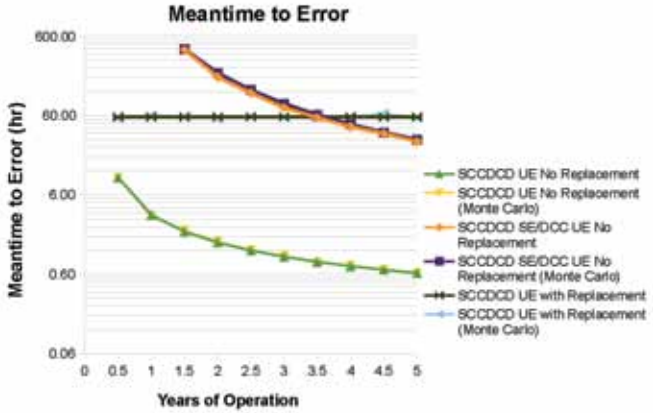

Fig. 1. Comparison of the mean time to different types of errors as predicted using the analytical models to that observed by Monte Carlo simulation for the evaluated memory organization.

As chipkill correct memories become commonplace, there will be a greater desire from system designers to predict the reliability of their system for a particular strength of chipkill correct memories, a particular memory organization such as capacity per DIMM and ranks per channel, etc., and a desired maintenance schedule of how often to perform memory scrubbing and to replace failing DIMMs in memory. Existing measurement of uncorrectable error rates for chipkill correct memories [1,2] are for specific memory organizations and therefore do not provide a way to extrapolate the reliability of memory for different memory organizations. Although a large body of work on the correctable error rate of error protected memory exist in literature [3-5], they only consider uncorrectable error rates on an individual codeword by codeword basis and ignore the correlation of faults between codewords due to the device-level faults such as row, column, and subbank faults. However, the driving factor for the increased reliability of chipkill correct memories is that they can correct against device-level faults that affect a large number of cells in memory per fault, some of which are not correctable by SECDED. As a result, by ignoring device-level faults, previous work on the reliability of ECC memories is adequate for chipkill correct memories.

The purpose of our work is to model the reliability of chipkill correct memories for an arbitrary memory organization and to calculate how often DIMMs need to be replaced for a certain type of chipkill correct memory. Such information will help guide the design of the next generation of HPC systems.

Our work creates an analytical model of memory. We make many simplifying assumptions and are careful to explain why these assumptions are realistic and still make the results useful and interesting. The model allows us to calculate the probability of uncorrectable error SCCDCD, probability of silent data corruption for SCCDCD, probability of uncorrectable and silent data corruption for DCC, and the probability of error with DIMM replacement. These involve many complex equations that are not appropriate for this publication and are not given here.

In addition to our analytical model, we have also designed a Monte Carlo simulation. The Monte Carlo simulation considers a single channel at a time. Each run of Monte Carlo simulates how long it takes for the corresponding condition (replacement for SCCDCD or DCC or uncorrectable fault for SCCDCD) to occur. In each simulated time interval, a random floating point number is generated for each type of fault. If a random number is less than the probability of its corresponding type of fault for the current time interval, a fault of that type is injected into the simulated memory channel. When the condition of unrecoverable error (UE)/SE/replacement is met, the total time span between the start of the simulation to the failing time interval is reported. After a number of time intervals, whose total span equals a scrubbing period,

For more information, contact Nathan DeBardeleben at ndebard@lanl.gov.

Funding Acknowledgments
DOE, NNSA, Advanced Simulation and Computing Program

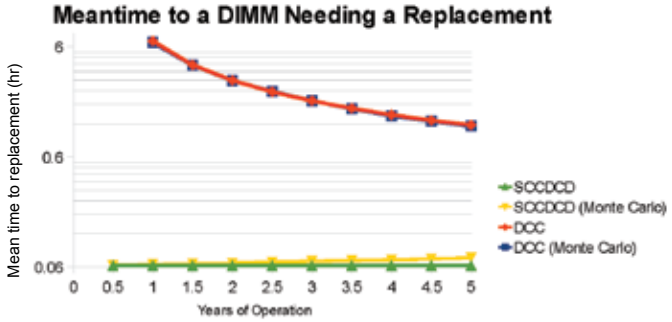

Fig. 2. Comparisoon of the mean time to a DIMM needing replacement as predicted using the analytical models to that observed by Monte Carlo simulation for the evaluated memory organization.

have passed, all the row and symbol faults are cleared to simulate the effect of scrubbing. Similarly, after a number of time intervals, whose total span equals a DIMM replacement interval, have passed, all the remaining faults are cleared to simulate the effect of DIMM replacement. The total number of Monte Carlo runs is equal to the total number of channels in the target memory system. The outputs from each run are grouped into different bins depending on which time interval they fall into.

***Table 1**. Simulated Memory Organization*

Size of Row in a Bank	8 kilobytes
Banks Per Rank	8
Capacity Per Device	2 gigabytes
Data Devices Per DIMM	32
DIMMs Per Channel	2
Total Memory Size	64 petabytes
Number of Scrubs Per Week	1

Table 1 shows the memory organization used for evaluation. Note that we are simulating an early exascale system design, so newer simulations would exacerbate DIMM errors due to larger capacities of more recent exascale system design predictions. Figure 1 compares the mean time to different types of errors as predicted using the analytical models to that observed by Monte Carlo simulation for the evaluated memory organization. A couple of important observations can be made from Fig.1. First, the comparison between SCCDCD UE with no replacement versus SCCDCD SE/DCC UE with no replacement show that the reliability of tolerating two bad symbols is almost two orders of magnitude greater than the reliability of tolerating only one bad symbol per codeword. The second observation derived by comparing SCCDCD UE with replacement to SCCDCD UE without replacement is that with DIMM replacement the meantime to uncorrectable error can be dramatically increased even with infrequent DIMM replacement (the replacement schedule is once per week as given in Table 1). In other words, DIMM replacement need not be immediate and but can be done at a convenient pace. Finally, DCC may not require any DIMM replacement until much later in the operation of the machine.

Figure 2 compares the mean time to a DIMM needing replacement as predicted using the analytical models to that observed by Monte Carlo simulation for the evaluated memory organization. The visible difference between the Monte Carlo model and the analytical model for SCCDCD is due to the fact that in the Monte Carlo model a DIMM that needs to be replaced is taken out of the total pool of DIMMs instead of actually being replaced, which causes the total number of DIMMs simulated to decrease over time. Since a relatively large number of DIMM replacements are required, the deviation from the analytical model can become large. On the other hand, since the number of DIMM replacements is very small for DCC, the deviation from the analytical model is not noticeable.

We are taking this work in several directions. On current systems the community does not have a good understanding of DIMM device faults and how that should drive DIMM replacement strategies. This can result in either a failure prone environment or be wasteful in disposal of (as we have shown here, under some circumstances) usable DIMMs. Furthermore, we are working to model actual replacement strategies including the amount of time it takes to replace a DIMM and the cost associated with replacement so that we can make some very practical recommendations. Additionally, we are looking at some next generation triple chipkill technologies and developing representative models and simulations. We are working with AMD and SNL to compare results from DRAM errors on Cielo with the results shown in [2].

[1] Schroeder, B. et al., "DRAM Errors in the Wild: A Large-Scale Field Study," *Sigmetrics* ACM, New York, NY 193 (2009).

[2] Sridharan, V. and D. Liberty, International Conference on High Performance Computing Networking, Storage and Analysis, Article 76, Supercompuing 2012.

[3] Mkherjee, S. et al., "Cache Scrubbing in Microprocessors: Myth or Necessity?" *Pacific Rim International Symposium Dependable Computing (PRDC)* (2012).

[4] Saleh, A.M. et al., *IEEE Trans Reliab* **39**(1), 114 (1990).

[5] Schiano, L. et al., "Markov models of fault-tolerant memory systems under SEU," *IEEE Workshop Memory Technology, Design, and Testing* (2004).