Cost-Aware Prediction of Uncorrected DRAM Errors in the Field

Isaac Boixaderas*, Paul M. Carpenter*, Petar Radojković*, Eduard Ayguadé*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {isaac.boixaderas, paul.carpenter, petar.radojkovic, eduard.ayguade}@bsc.es

***Keywords*—Memory system, Reliability, Error prediction, Machine learning, Random forest, Cost–benefit analysis.**

I. Extended Abstract

One of the main causes of hardware failure in large-scale clusters is an uncorrected error in main memory [1]–[4]. Node failures are especially problematic in high-performance computing (HPC), where a single tightly-coupled job may execute for days on thousands of nodes. If any node fails, the whole job is terminated, typically wasting all CPU hours since the last checkpoint. Memory system reliability is therefore an important limit on the ability to scale to larger systems.

This abstract summarizes our study, published in SC20 [5], which aims to increase effective use of HPC systems by reducing the compute time lost due to memory system failures. Firstly, our study presents and evaluates a method to predict DRAM uncorrected errors (UEs) that can enable the system to take active mitigation measures, e.g. checkpointing or live job migration. We concentrate on uncorrected errors (UEs), which cause the node to fail, rather than corrected errors (CEs), which do not have a direct connection to UEs.

Secondly, we discuss and clarify several aspects of methodology, relating to cost–benefit analysis and potential sources of bias, that are essential for such a prediction method to be useful in practice. We show that standard metrics for data prediction, such as precision, recall and F1-score, are not correlated with saved compute time or mitigation costs, and therefore are insufficient to decide whether and for which model parameters the prediction is useful in practice. Instead, we use a cost–benefit analysis which directly compares the system resources needed for training, failure prediction and failure mitigation against the saved compute time due to successful failure prediction and mitigation [6].

Overall, our open source method [7], reduces lost compute time by up to 57%, a net savings of 21,000 node–hours per year for a real production job distribution. We encourage the community to adopt our methodology for pre-processing, model training, parameter exploration and evaluation, so that future DRAM error prediction methods are also free from training bias and supported by a cost–benefit calculation.

A. Environment description

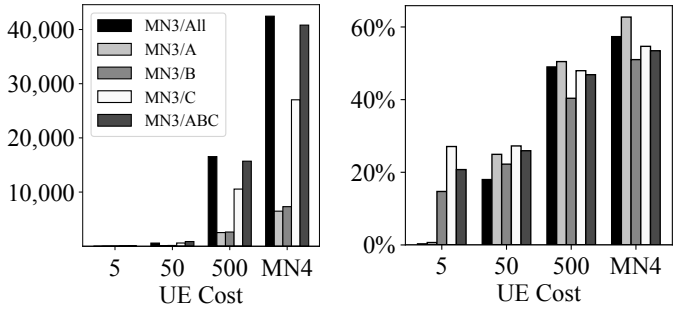
Our prediction method is trained and evaluated on memory error logs from the PRACE Tier-0 MareNostrum 3 supercomputer [8] over a production period from October 2014 to November 2016. At the time, MareNostrum 3 comprised 3056 compute nodes and more than 25,000 DDR3-1600 DIMMs from all three major memory manufacturers. These manufacturers have been anonymized, and are referred to as *Manufacturer A* (6694 DIMMs), *B* (5207 DIMMs) and *C* (13,419 DIMMs).

B. Prediction method and methodology

We perform uncorrected DRAM error prediction using a random forest classifier. The classifier makes one prediction for each DIMM as to whether or not it will experience a UE in the upcoming "prediction window". The features were obtained directly from the CE and event logs. We use offline learning, with hyperparameter tuning and time series cross-validation.

C. Results

Figure 1 summarizes the results of the cost–benefit analysis for the prediction model and mitigation. The x-axis is the UE cost: fixed at 5, 50 and 500 node–hours or calculated using the job size distribution from production MareNostrum 4 HPC job logs. The y-axis in Figure 1a is the saved node–hours, which is the reduction in lost compute time due to UE prediction and mitigation compared with the baseline system. The y-axis in Figure 1b is the saved node–hours normalized to the number of node–hours lost in the baseline system. Results are shown for MareNostrum 3 as a whole (MN3/All) and its different subsystems corresponding to DRAM manufacturer: MN3/A, MN3/B and MN3/C. Bars MN3/ABC show the overall results for the whole system treated as the sum of its three DRAM manufacturer subsystems. The results show that the effectiveness of the method is similar across all these scenarios but that the cost–benefit calculation is strongly influenced by the average UE cost. For a small UE cost of 5 node–hours, UE prediction has zero effect on the saved compute time. For medium and large UE costs, however, savings are seen in Figure 1a, of 586 node–hours (medium) and 16,541 node–hours (large). These are reductions of 18% and 49% respectively (Figure 1b). The node–hours savings computed based on the production job logs reach 57% which is equivalent to 42,000 node–hours or 21,000 node–hours per year.



(a) Number of saved node–hours (b) Percentage saved node–hours

Fig. 1: The model cost-efficiency depends on the UE cost. For large UE cost, the savings are significant, measured in thousands of node–hours over the two-year production period.

D. Conclusions

This paper summarized our method to predict DRAM uncorrected errors and our cost–benefit methodology. We see that the effectiveness of our prediction scheme is highly dependent on system and workload characteristics, pointing the way to future work on adaptive resiliency techniques. The full paper [5] provides full details on the prediction method, the features used for prediction and the detailed cost–benefit methodology. It also analyzes the effect of parameters such as prediction window, prediction frequency, and decision threshold. It evaluates the method also with standard data prediction methods, explores the relative importance of prediction features and compares the random forest approach with five other machine learning classifiers. Overall, we hope that future researchers will build on our work to improve the throughput of production HPC systems as demonstrated by a clear cost–benefit calculation.

Acknowledgment

This work was supported by the Spanish Ministry of Science and Technology (project PID2019-107255GB / AEI / 10.13039/501100011033), Generalitat de Catalunya (contracts 2014-SGR-1051 and 2014-SGR-1272), the Ministry of Economy and Competitiveness of Spain (Ramon y Cajal fellowships YC2018-025628-I and RYC2017-23269) and the European Union's Horizon 2020 research and innovation programme (grant agreement No 754337, EuroEXA).

References

[1] HP, "How memory RAS technologies can enhance the uptime of HPE ProLiant servers," Hewlett Packard Enterprise, Technical white paper 4AA4-3490ENW, Feb 2016.

[2] I. Giurgiu, J. Szabo, D. Wiesmann, and J. Bird, "Predicting DRAM Reliability in the Field with Machine Learning," in *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track*, 2017, pp. 15–21.

[3] B. Schroeder, E. Pinheiro, and W.-D. Weber, "DRAM Errors in the Wild: A Large-scale Field Study," in *Proceedings of the International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2009, pp. 193–204.

[4] A. A. Hwang, I. A. Stefanovici, and B. Schroeder, "Cosmic rays don't strike twice: understanding the nature of dram errors and the implications for system design," *ACM SIGPLAN Notices*, vol. 47, no. 4, pp. 111–122, 2012.

[5] I. Boixaderas, D. Zivanovic, S. Moré, J. Bartolome, D. Vicente, M. Casas, P. M. Carpenter, P. Radojković, and E. Ayguadé, "Cost-aware prediction of uncorrected dram errors in the field," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–15.

[6] P. Radojkovic, M. Marazakis, P. Carpenter, R. Jeyapaul, D. Gizopoulos, M. Schulz, A. Armejach, E. Ayguade, F. Bodin, R. Canal, F. Cappello, F. Chaix, G. Colin de Verdiere, S. Derradji, S. Di Carlo, C. Engelmann, I. Laguna, M. Moreto, O. Mutlu, L. Papadopoulos, O. Perks, M. Ploumidis, B. Salami, Y. Sazeides, D. Soudris, Y. Sourdis, P. Stenstrom, S. Thibault, W. Toms, and O. Unsal, "Towards Resilient EU HPC Systems: A Blueprint." European HPC resilience initiative. White paper, April 2020. [Online]. Available: https://resilienthpc.eu/blueprint2020

[7] I. Boixaderas, D. Zivanovic, S. Moré, J. Bartolome, D. Vicente, M. Casas, P. M. Carpenter, P. Radojković, and E. Ayguadé, "UEPREDICT: A method for predicting DRAM Uncorrected Errors and evaluating its model's performance," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3872777

[8] Barcelona Supercomputing Center, *MareNostrum 3 User's Guide*, Apr. 2016.

Isaac Boixaderas is a Research Engineer at Barcelona Supercomputing Center (BSC). He received his BSc degree in Computer Science from Universitat Politècnica de Catalunya (UPC) in 2016. Currently, he is pursuing a MSc in Data Science at Universitat Oberta de Catalunya (UOC). Prior to starting his career as a researcher, he worked as a Software Engineer at Universitat Internacional de Catalunya (UIC) and Inbenta Holdings Inc. in Barcelona.