Practical Reliability Analysis of GPGPUs in the Wild: From Systems to Applications

Evgenia Smirni
College of William and Mary
Williamsburg, VA, USA
esmirni@cs.wm.edu

ABSTRACT

General Purpose Graphics Processing Units (GPGPUs) have rapidly evolved to enable energy-efficient data-parallel computing for a broad range of scientific areas. While GPUs achieve exascale performance at a stringent power budget, they are also susceptible to soft errors (faults), often caused by high-energy particle strikes, that can significantly affect application output quality. As those applications are normally long-running, investigating the characteristics of GPU errors becomes imperative to better understand the reliability of such systems. In this talk, I will present a study of the system conditions that trigger GPU soft errors using a six-month trace data collected from a large-scale, operational HPC system from Oak Ridge National Lab. Workload characteristics, certain GPU cards, temperature and power consumption could be indicative of GPU faults, but it is non-trivial to exploit them for error prediction. Motivated by these observations and challenges, I will show how machine-learning-based error prediction models can capture the hidden interactions among system and workload properties. The above findings beg the question: how can one better understand the resilience of applications if faults are bound to happen? To this end, I will illustrate the challenges of comprehensive fault injection in GPGPU applications and outline a novel fault injection solution that captures the error resilience profile of GPGPU applications.

The presented work is done in collaboration with Adwait Jog (College of William and Mary), Devesh Tiwari (Northeastern University), Ji Xue (Google), and my Ph.D students Bin Nie and Lishan Yang. The interested reader is directed to [1–4] for details.

CCS CONCEPTS

• **Computer systems organization → Parallel architectures**; **Single instruction, multiple data**; **Reliability**; *Multiple instruction, multiple data.*

KEYWORDS

HPC systems; GPGPUs; workload characterization; fault injection; machine learning models for reliability analysis; resilience

ACM Reference Format:
Evgenia Smirni. 2019. Practical Reliability Analysis of GPGPUs in the Wild: from Systems to Applications. In *Tenth ACM/SPEC International Conference on Performance Engineering (ICPE '19), April 7–11, 2019, Mumbai, India.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3297663.3310291

ICPE '19, April 7–11, 2019, Mumbai, India

ACM ISBN 978-1-4503-6239-9/19/04.
https://doi.org/10.1145/3297663.3310291

BIO

Evgenia Smirni is the Sidney P. Chockley Professor of Computer Science at the College of William and Mary, Williamsburg, VA, USA. Her research interests include queuing networks, stochastic modeling, Markov chains, resource allocation policies, storage systems, data centers and cloud computing, workload characterization, models for performance prediction, and reliability of distributed systems and applications. She has served as the Program co-Chair of QEST'05, ACM Sigmetrics/Performance'06, HotMetrics'10, ICPE'17, DSN'17, SRDS'19, and HPDC'19. She also served as the General co-Chair of QEST'10 and NSMC'10. She is an ACM Distinguished Scientist, a senior member of IEEE, and a member of the Technical Chamber of Greece.

ACKNOWLEDGMENTS

This work has been partially supported by NSF grants CCF-1649087 and CCF-1717532.

REFERENCES

[1] Bin Nie, Devesh Tiwari, Saurabh Gupta, Evgenia Smirni, and James H. Rogers. 2016. A large-scale study of soft-errors on GPUs in the field. In *2016 IEEE International Symposium on High Performance Computer Architecture, HPCA 2016, Barcelona, Spain, March 12-16, 2016.* 519–530. https://doi.org/10.1109/HPCA.2016.7446091

[2] Bin Nie, Ji Xue, Saurabh Gupta, Christian Engelmann, Evgenia Smirni, and Devesh Tiwari. 2017. Characterizing Temperature, Power, and Soft-Error Behaviors in Data Center Systems: Insights, Challenges, and Opportunities. In *25th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2017, Banff, AB, Canada, September 20-22, 2017.* 22–31. https://doi.org/10.1109/MASCOTS.2017.12

[3] Bin Nie, Ji Xue, Saurabh Gupta, Tirthak Patel, Christian Engelmann, Evgenia Smirni, and Devesh Tiwari. 2018. Machine Learning Models for GPU Error Prediction in a Large Scale HPC System. In *48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2018, Luxembourg City, Luxembourg, June 25-28, 2018.* 95–106. https://doi.org/10.1109/DSN.2018.00022

[4] Bin Nie, Lishan Yang, Adwait Jog, and Evgenia Smirni. 2018. Fault Site Pruning for Practical Reliability Analysis of GPGPU Applications. In *51st Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2018, Fukuoka, Japan, October 20-24, 2018.* 749–761. https://doi.org/10.1109/MICRO.2018.00066