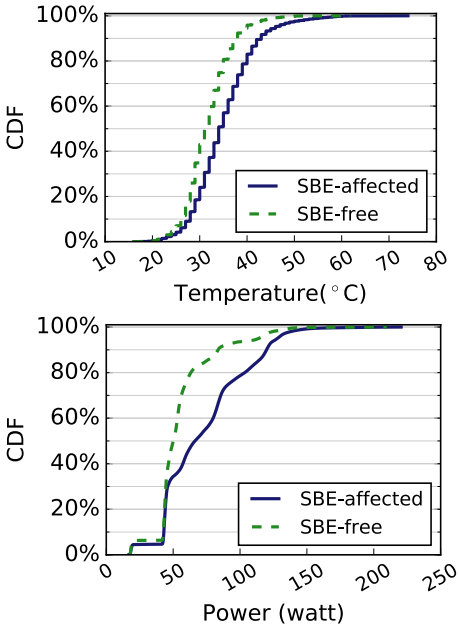
Characterizing Temperature, Power, and Soft-Error Behaviors in Data Center Systems

Achievement: Studied and characterized the interplay between temperature, power consumption, and reliability for general-purpose graphics processing units (GPGPU) nodes on ORNL's Titan supercomputer.

Significance and Impact: This work offers an understanding of GPGPU reliability in today's super-computers and paves the way for maintaining efficient and correct operation of future exascale systems.

Research Details:

- Conducted an in-depth study of GPGPU temperature, power consumption, and soft-error reliability on ORNL's Titan supercomputer.
- Developed an understanding of the differences in central processing unit (CPU) and GPGPU temperature and power consumption behavior.
- Obtained knowledge about the higher single-bit error susceptibility of GPGPUs under higher temperatures and under higher power consumption.
- Utilized the obtained knowledge to successfully predict higher single-bit error rates under higher temperatures and under higher power consumption.



Cumulative distribution functions of temperature and power consumption of single-bit error (SBE) offender GPGPU nodes during SBE-affected and SBE-free periods, showing that temperature and power consumption affect GPGPU reliability.

Sponsor/Facility: This work was performed at ORNL in collaboration with College of William and Mary and Northeastern University. The work was sponsored by DOE ASCR (Program manager: Lucy Nowell).

PI and affiliation: Christian Engelmann – Oak Ridge National Laboratory

Team: B. Nie, D. Tiwari, J. Xue, S. Gupta, C. Engelmann, and E. Smirni.

Publication: B. Nie, D. Tiwari, J. Xue, S. Gupta, C. Engelmann, and E. Smirni. **Characterizing Temperature, Power, and Soft-Error Behaviors in Data Center Systems: Insights, Challenges, and Opportunities**. In *Proceedings of the 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS) 2017*, Banff, AB, Canada, September 20-22, 2017.

Overview:

We conducted an in-depth study on the GPGPU nodes in the Titan supercomputer to understand the interplay between temperature and GPGPU reliability, and more specifically GPGPU soft errors. We also investigated how different levels of power consumption affect the data center's operations and reliability. We believe that a thorough understanding on the relationship between temperature, power consumption and GPGPU errors is key to improving the operational efficiency of data centers. Beyond characterizing the conditions that may lead to GPGPU errors, we also exploited the observed insights and implications for error probability prediction. This work made the following three contributions.

First, we analyzed large amounts of measured system data to understand the characteristics of the temperature distribution on Titan. In particular, we investigated how the GPGPU temperature distribution varies in time and space across the system. Furthermore, we compared GPGPUs with other components, such as CPUs and dual in-line memory module (DIMM), and studied how their temperature distributions differ from one another over time. We also studied how frequently Titan nodes become extremely hot and

for how long they stay in such a hot state. We discovered that the retention time histogram in the hot state and in the normal state varies significantly between CPUs and GPGPUs. We also found that GPGPUs switch in and out of the hot and cold states more frequently compared to CPUs and stay in these states for a shorter period of time. We observed that, surprisingly, the retention time in the hot state remains similar for cabinets from different temperature zones.

Second, we showed that there exists an interconnection between temperature, power consumption, and GPGPU soft-errors, SBEs specifically. It is challenging to quantitatively exhibit and exploit this relationship. We point out that predicting future GPGPU soft errors based on past temperature and power is simply inconclusive as contrasting conclusions may be reached.

Third, we elaborated on how to exploit these observations for GPGPU soft-error probability prediction. We proposed a machine-learning-based technique that leverages observations of past system measurements to predict soft error probability in GPGPUs. We showed that temperature and power consumption are of almost equal importance in GPGPU soft-error probability prediction, and together with a host of other factors including resource utilization, node location, and application type, may determine whether an upcoming application execution on a set of nodes will result in GPGPU soft errors. We evaluated our technique under various scenarios to demonstrate its effectiveness and robustness.