Lightweight, Actionable Analytical Tools Based on Statistical Learning for Efficient System Operations

Devesh Tiwari, Saurabh Gupta, and Christian Englemann

Focus area: Integration of Measurement and ModSim; ModSim Methods and Tools
Modeling and simulation scope: performance, reliability and power efficiency for system operations
Target environment and application: System events and logs generated by production systems

Modeling and simulation community has always relied on accurate and meaningful system data and parameters to drive analytical models and simulators. HPC systems continuously generate huge amount system event related data (e.g., system log, resource consumption log, RAS logs, power consumption logs), but meaningful interpretation and accuracy verification of such data is quite challenging. This challenge forces us to make suboptimal choices for both the future and current HPC systems. For example, lack of accurate and meaningful analysis of the RAS data may lead to suboptimal RAS support in the future hardware systems. Similarly, lack of useful analysis of the power consumption behavior can lead to inefficient power and cooling for the future HPC systems, leading monetary loss.

There are multiple challenges in performing meaningful and reliable analysis of system data generated by different hardware and software instrumentation. First, the amount of data generated is often hard to manage efficiently, and perform analysis on. Second, the analysis process for such data is often not portable because different system vendors produce data in different logging formats, undesirably quite cryptic in many cases. Third, the turning on system instrumentation points and performing analysis on the produced data imposes significant overhead, making it unattractive to use it as an actionable tool.

We propose a novel, principled approach that the modeling and simulation community can take to address these challenges. We propose that we should identify a set of metric that are most useful for application performance, scientific productivity, and system operations. The first step the proposed research is to use modeling and simulation tools for understanding the impact of different performance, power, and RAS features on system and application performance. Unfortunately, the space of such parameters can be fairly large for systematic exploration, and is often the reason for “want-all, get-none”. We argue that execution profile traces of applications captured on production machines can prune the vast parameter space. Production system snapshots can identify important features that are representative set to understand the behavior of overall systems.

Based on these pruning of parameters, novel parameter-driven analytical models and stochastic modeling can drive the next steps, i.e., quantifying the impact of such parameters on application and system performance under different operating conditions. We have shown that novel statistical learning techniques can be developed to translate machine-dependent parameters to machine-independent features for better performance, power, and reliability exploration. The next step is to drive simulation based studies based on the insights gained from the models and statistical learning. Simulation can further refine and validate the analytical models and statistical learning based analytical tools. This is an iterative process to obtain better accuracy and increase the effectiveness of model based results and results.

Finally, based on our modeling and simulation step, we decide on a set of parameters and metric that should be monitored online and novel actionable tools that utilize these measured data to make decisions for improving system operations and application performance. Note that analytical tools are developed and evaluated rigorously during the offline step, this reduces the need for tweaking them dynamically during online phase. In our experience, one of the most important achievement of this process has been reducing the amount of data that needs to be analyzed online by developing novel statistical learning based analytical tools. We have developed several strategies that enables us to make reasonably well-guided decisions without having to process huge amount of data in real-time, this makes the analytical tools take actions and provide guidelines in near real-time. Interestingly, there are several useful by-products of this approach. This approach reduces the need to keep all the system generated data for long-term since the statistical tools can be used to extract only the reduced set of relevant data and patterns that capture the system behavior. This approach is very useful in standardizing the common set of features and their logging formats across systems and over time to make the analysis tools portable. This particular problem is more involved and challenging for many possibly non-technical reasons, nevertheless, this approach can be helpful in achieving some standardization or common set of features/metric that we argue could be portable across systems. Last, but not least, this process also works as a great feedback mechanism for developing system benchmarking and stress-test suite, especially for resilience regression test-suite.

Preliminary results obtained from following our previously described approach has been very encouraging. Through an iterative process, we have identified an important set of RAS and power consumption metric that are important for application performance and system operations. We intelligently pruned the set of RAS and system performance metric to reduce the performance overhead of online measurement and monitoring of these metric. We are continuing to develop actionable analytical tools that applying statistical learning technique to improve application performance and system operations. Some recent successful examples include Lazy Checkpointing [1], Quarantine Scheduling Technique [2], Power-capping aware Checkpointing [3], GPU reliability assessment tools [4,5], demystifying temperature/power interactions with system provisioning [6,7]. We will discuss some of these examples in detail to demonstrate how modeling and simulation techniques have been to used to directly improve HPC system operations. This will talk bring unique perspective and experience in demonstrating how modeling & simulation based research can actually be translated into production systems. We will discuss the short-term opportunities for modeling & simulation community to increase the impact and effectiveness of our analytical tools, “dos and don’ts”, long-term challenges and opportunities.

[1] **DSN 2014:** Devesh Tiwari, Saurabh Gupta, and Sudharshan S. Vazhkudai, "Lazy Checkpointing: Understanding and Mitigating Checkpointing Overheads on Extreme-Scale Machines", In proceedings of IEEE Conference on Dependable Systems and Networks (DSN 2014), June, 2014.

[2] **DSN 2015:** Saurabh Gupta, Devesh Tiwari, Christopher J. Jantzi, James H. Rogers, and Don Maxwell, "Understanding and Exploiting Spatial Properties of System Failures on Extreme-Scale HPC Systems", In Proceedings of the 45th IEEE Conference on Dependable Systems and Networks (DSN 2015), June, 2015.

[3] **DSN 2016:** Kun Tang, Devesh Tiwari, Saurabh Gupta, Ping Huang, Qi Lu, Christian Engelmann and Xubin He, "Power-capping Aware Checkpointing: On the Interplay among Power-capping, Temperature, Reliability, Performance, and Energy", To appear in the 46th IEEE Conference on Dependable Systems and Networks (DSN 2016), June, 2016

[4] **HPCA 2015:** Devesh Tiwari, Saurabh Gupta, Jim Rogers, Don Maxwell, Paolo Rech, Sudharshan Vazhkudai, Daniel Oliveira, Dave Londo, Nathan Debardeleben, Philippe Navaux, Luigi Carro, and Arthur Buddy Bland, "Understanding GPU Errors on Large-scale HPC Systems and the Implications for System Design and Operation", In Proceedings of 21st IEEE Symposium on High Performance Computer Architecture (HPCA 2015), February 2015

[5] **SC 2015:** Devesh Tiwari, Saurabh Gupta, George Gallarno, Jim Rogers and Don Maxwell, "Reliability Lessons Learned from GPU Experience with the Titan Supercomputer at Oak Ridge Leadership Computing Facility", In Proceedings of Supercomputing 2015 (SC15), Austin, TX, November, 2015.

[6] **HPCA 2016:** Bin Nie, Devesh Tiwari, Saurabh Gupta, Evgenia Smirni, and James H. Rogers, "A Large-Scale Study of Soft-Errors on GPUs in the Field", In Proceedings of the 22nd IEEE Symposium on High Performance Computer Architecture (HPCA 2016), Barcelona, Spain, March, 2016.

[7] **ICAC 2016:** Jaimie Kelley, Christopher Stewart, Devesh Tiwari and Saurabh Gupta, "Adaptive Workload Profiling for Power Efficient HPC", To appear in the 13th IEEE International Conference on Autonomic Computing (ICAC 2016), July, 2016