Reliability and Energy Data Analysis and Modeling for Extreme Scale Systems

Li Yu and Zhiling Lan
Illinois Institute of Technology
lyu17@hawk.iit.edu, lan@iit.edu

***Abstract*—Reliability and energy are two of the top major concerns in the development of today's supercomputers. To build a powerful machine while at the same time satisfying reliability requirement and energy constraint, HPC scientists continue to seek a better understanding of system and component behaviors. Toward this end, modern systems are deployed with various monitoring and logging tools to track reliability and energy data during system operations. Since these data contain important information about system reliability and energy, they are valuable resources for understanding system behaviors. However, as system scale and complexity continue to grow, the process from collecting system data to extracting meaningful knowledge out of overwhelming reliability and energy data faces a number of key challenges. To address these challenges, my work consists of three parts, including data preprocessing, data analysis and advanced modeling.**

I. Introduction

Reliability and energy have become two major concerns as we move towards exascale high performance computing (HPC). To build systems with effective resilience mechanisms and high energy-efficiency, an in-depth understanding of system and component behaviors is required. Because of this, modern systems are deployed with various monitoring and logging facilities to track reliability and energy data during system operations. For example, the environmental monitors deployed on IBM Blue Gene systems can collect data like temperatures, clock frequency, fan speeds, and voltages from the underlying hardware devices and generate RAS (Reliability, Availability, and Serviceability) events when abnormal readings are encountered; the OVIS monitoring tool developed from Sandia National Lab can collect various state variables (e.g., temperature, CPU utilization, fan speed) and user-specified variables (e.g., aggregated memory errors over the life span of a job) on various large-scale clusters.

Although these data are regarded as valuable resources for understanding system behaviors, *extracting meaningful knowledge from them and in turn facilitating system design* remain a challenging process whose difficulty has been rapidly escalated by the ever growing system scale with unprecedented complexity. The huge data volume and the great data complexity not only cast a heavy burden on monitoring and logging facilities but also make it difficult to gain useful information from them. To address these challenges, my thesis work is dedicated to reliability and energy data analysis and modeling on extreme scale systems. It contains three key component: *data preprocessing*, *data analysis* and *analytical / stochastic modeling*. Data preprocessing is a refining process of the raw system data, through which less useful information can be filtered out. Data analysis extracts useful information from the refined data to formalize domain knowledge. Analytical / stochastic modeling further leverages the knowledge gained from data analysis to provide a comprehensive view of system behaviors and guide system design. The overview of my work is shown in Fig. 1.

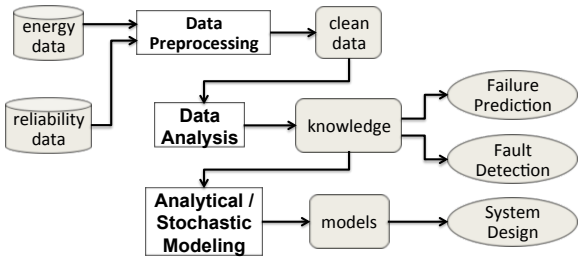

Fig. 1. The overview of my thesis work.

II. Data Preprocessing

With respect to data preprocessing, we have focused on improving the efficiency of data collection in modern supercomputers. In [1], we present a 2-dimensional online data filtering mechanism to remove noisy and redundant data horizontally (via feature selection) as well as vertically (via instance selection). Our design comprises two major components. Feature is feature selection, where data filtering is conducted horizontally, meaning representative features are selected and the rest is dropped. Second is instance selection, where data filtering is performed vertically along the time axis, meaning representative instances are selected and the rest is dropped.

To demonstrate the effectiveness of our filtering design, we conduct two case studies on real environmental data collected from two production supercomputers (an environmental log from the Blue Gene/P system at Oak Ridge National Lab and an OVIS log from a cluster at Sandia National Lab). We examine the amount of disk storage that can be reduced by applying our online filtering mechanism. We also compare the effects of our filtering mechanism as against random filtering. Further, we study whether the proposed filtering mechanism brings a positive or negative impact on failure prediction and diagnosis. The use of multiple data is to ensure the presented mechanism is not biased to any specific system or log and thus is general for providing filtering service for a variety of log data. For the Blue Gene/P environmental data, our method can reduce $\sim 85.6\%$ disk space without losing prediction accuracy and root cause information. For the OVIS log, we can achieve $\sim 99.7\%$ disk space saving without losing

both prediction accuracy and root cause information. To the best of our knowledge, we are among the first to explore 2-dimensional filtering (i.e., horizontally as well as vertically) for reducing system logs like environmental data and utilize the filtered data for better failure prediction and diagnosis on large-scale systems.

III. Data Analysis

With respect to data analysis, we have focused on developing automated data analysis solutions for extracting valuable reliability knowledge (e.g., failure pattern and propagation) out of potentially overwhelming volumes of data. The knowledge can be used for both fault detection and prediction.

In [2], we analyze and compare the impact of *observation window* and *lead time* on two commonly used prediction approaches, namely period-based and event-driven approaches. The objective is two-folded: one is to show which prediction approach provides better accuracy and is more suitable for practical use in reality, and the other is to provide some guidance in terms of the design of proactive failure management. The major contributions include: (1) we develop an online Bayesian-based failure prediction method and implement it via both period-based and event-driven approaches and (2) we evaluate these prediction approaches under a variety of testing parameters. In particular, we examine the sensitivity of observation window and lead time on both prediction mechanisms. To the best of our knowledge, this work is the first to study the time characteristics of these two commonly used prediction approaches in large-scale systems. Experimental results show that the period-based Bayesian model and the event-driven Bayesian model can achieve up to 65.0% and 83.8% prediction accuracy, respectively. Furthermore, our sensitivity study indicates that the event-driven approach seems more suitable for proactive fault management in large-scale systems like Blue Gene/P.

In [3], we present a scalable, non-parametric anomaly detection framework for large scale systems. The purpose of this work is to address both the *scalability* issue and *practical use* issue for anomaly detection in large-scale systems. *Our design is built on two key techniques*. First, we adopt a grouping strategy, through which the big problem involving the analysis of a large system (e.g., thousands or hundreds of thousands of nodes) is divided into many small problems involving the analysis of a small group of nodes (e.g., tens to dozens of nodes). Second, we develop a novel non-parametric clustering based method, which does not rely on any pre-defined clustering groups and thus is capable of handling various cases. In addition to the design of the non-parametric method, in this paper we also present an analytical study to quantitatively prove that the proposed diagnosis method can provide better diagnosis accuracy than existing parametric methods. The preliminary results indicate that our decentralized design is highly scalable and outperforms existing diagnosis methods by up to 34% in terms of diagnosis accuracy. Our framework incurs very low runtime overhead. In our experiments, it merely takes about 11 ms to detect various anomalies under different workload and anomaly combinations.

IV. Analytical & Stochastic Modeling

With respect to analytical & stochastic Modeling, we have focused on providing analytical and stochastic models for quantitatively analyzing performance under the reliability and energy constraints.

In [4], we present a set of reliability-aware speedup models to extend Amdahl's law and Gustafson's law by considering failure impact. The goal of this work is to provide a more accurate measurement of application speedup in a practical failure-prone environment. More importantly, we consider both Exponential and Weibull failure distributions in the model construction. To tackle the challenge of dynamic failure rate inherent in Weibull distribution, we derive lower and upper bounds of application speedup under Weibull failure distribution. By means of real failure data from various production systems, we have demonstrated that these analytical models can better represent application performance and speedup in the presence of failures. Moreover, our results clearly show that Weibull based models outperform Exponential based models in terms of characterizing application speedup in the presence of failures.

In [5], we explore a new stochastic modeling approach by presenting a colored Petri net named PuPPET (Power Performance PETri net) for quantitative study and predictive analysis of different power management mechanisms on extreme scale systems. Although colored Petri nets have been widely used for modeling large scale systems like biological networks, *to our knowledge this is the first attempt of applying CPN for quantitative power-performance modeling in high performance computing*. Using the system traces (i.e., workload log and power data) collected from the production 10-petaflops IBM Blue Gene/Q system named Mira at Argonne National Laboratory, our experiments show that PuPPET can effectively model batch scheduling and system energy consumption with a high accuracy, e.g., an error of less than 4%. The emulation of executing four-month jobs on Mira took a couple of minutes on a local PC. Given that Mira is a petascale machine with $49,152$ nodes, this result demonstrates that PuPPET is highly scalable. In the two case studies, we find that although DVFS seems to cause less impact on system performance, it could significantly impact hardware lifetime reliability, as high as up to 3X higher failure rate. Hence, unless in case of a tight power cap and high system utilization, power-aware allocation is a preferred power capping solution due to its comparable performance with DVFS and no impact to system reliability

References

[1] L. Yu, Z. Zheng, Z. Lan, T. Jones, J. Brandt and A. gentile,"Filtering Log Data: Finding the needles in the Haystack", *In Proc. of DSN*, 2012.

[2] L. Yu, Z. Zheng, Z, Lan and S. Coghlan, "Practical Online Failure Prediction for Blue Gene/P: Period-based vs Event-Driven", In Proc. of Proactive Failure Avoidance, Recovery, and Maintenance Workshop (PFARM), 2011.

[3] L. Yu and Z. Lan, "A Scalable, Non-Parametric Anomaly Detection Framekwork for Hadoop", In Proc. of the ACM Cloud and Autonomic Computing Conference (CAC), 2013.

[4] Z. Zheng, L. Yu, and Z.Lan, "Reliability-Aware Speedup Models for Parallel Applications with Coordinated Checkpointing/Restart", To appear in the IEEE Trans. on Computers, , 2014.

[5] L. Yu, Z. Zhou, S. Wallace, M. E. Papka and Z. Lan, "Quantitative Modeling of Power Performance Tradeoffs on Extreme Scale Systems", Technical Report, Illinois Institute of Technology, 2014.