Project Summary

RELIABILITY, AVAILABILITY, AND SERVICEABILITY FOR PETASCALE HIGH-END COMPUTING AND BEYOND

Grant ID: DE-FG02-08ER25836
Reporting period: 6/8-5/11
PI: Chokchai “Box” Leangsuksun
Computer Science Program
College of Engineering and Sciences
Louisiana Tech University

1. Progress Narrative

Our project is a multi-institutional research effort that adopts interplay of RELIABILITY, AVAILABILITY, and SERVICEABILITY (RAS) aspects for solving resilience issues in high-end scientific computing in the next generation of supercomputers. We aim to address various HPC reliability challenges by researching and offering solutions such as:

- An HPC RAS framework concept with combinations of proactive and reactive fault-tolerance (FT) mechanisms using a battery of approaches to adapt to various checkpoint and virtualization techniques
- Analysis and modeling of failures to steer runtime towards improved fault handling and to reduce the FT overhead required by reliability-aware checkpointing and resource management.
- An advancement in computer reliability, availability and serviceability (RAS) management systems to work cooperatively with the OS/R to identify and preemptively resolve system issues.
- Explore advanced monitoring and modeling techniques to further improve application performance and predictability of system interruptions.

The overall goal of the research conducted at Louisiana Tech University (LATECH) is to improve HEC/HPC resilience by investigating techniques and tools for RAS. We have collaborated with PI Scott from ORNL and PI Muller from North Carolina State University in various aspects as well as an Ovis group in Sandia National Lab. However, during the 3rd year, the highlighted LATECH-led research results lie in the following tracks:

- Failure prediction in a large scale HPC
- Investigate reliability issues and mitigation techniques including in GPGPU-based HPC system.
- HPC resilience runtime & tools

1.1 Reliability modeling and log and failure analysis techniques in a very large scale system.

In year 1, we investigated techniques and critical factors required to better understand the failure characteristics of very large HPC systems and to perform the reliability model analysis of system dependability. System- and application-level failures and other important events can be characterized by mining relevant log files and performing statistical analysis on the provided information. The resulting data may then be used in any number of future developments and studies on the corresponding computational architecture, including fields such as failure prediction, fault tolerance, performance modeling and power awareness.

This research track provides a statistical analysis of the application- and system-level failures encountered and logged by the LLNL IBM Blue Gene/L supercomputing system over a six-month period. Our findings suggest that the Blue Gene/L log files provide a wealth of information, but much of it is to no use of those interested in providing resilience to it and similar technologies, without first filtering the data and performing appropriate statistical analysis to arrive at correct and logical values.

We also studied the sensitivity of repair or recovery time with respect to the mean time to failure (MTTF) under an assumption that the 2nd failure following the 1st one before the system recovery will make no impact to the system. The results suggested that the Blue Gene/L system has a MTTF of 5.89 hours – roughly 4 times a day – which is a very close approximation of the actual MTTF of the LLNL Blue Gene/L. The full report can be found in [1]. Further studies will include by-percent load analysis and non-uniform application load failure analysis. Failures may occur as often as 10.6 minutes apart, or, likewise, one could go days (in the case of the 10 minute MTTR assumption – 3.8 days) without observing a system breakdown.

1.2 Develop near-optimal checkpoint approach

The incremental checkpoint concept has long been investigated as a means of reducing the overhead required by regular (full) checkpointing. Presently, however, there are no production-grade incremental checkpoint systems active within the HPC arena. We anticipate that, in the near future, there will exist systems which support HPC incremental checkpoint/restart, which in turn will further reduce overhead requirements. To gain the advantages provided by the incremental checkpoint technique, we proposed and studied an optimal checkpoint frequency function that globally minimizes the expected wasted time of the incremental checkpoint mechanism. Also, the re-computing time coe□cient used to approximate the post-restart re-computing time is derived. Moreover, to relieve complexity within the recovery state, full checkpoints are performed from time to time. As such, we provide an approach to evaluate the appropriate number of incremental checkpoints completed within two consecutive full checkpoints. However, presently the number of incremental checkpoints is constant, while the checkpoint interval derived from the proposed model varies corresponding to the failure rate of the system.

In this track, the existence and uniqueness of the multitude of incremental checkpoints have been proved. Moreover, the optimal checkpoint frequency function has been derived to be able to support any distribution of the time-to-failure random variable. The checkpoint time is illustrated in the case of a Weibull distribution which can be simplified to the exponential case. The optimality of the expected wasted time can be achieved theoretically by the proposed checkpoint frequency function and the derived number of incremental checkpoints.

1.3 Tunable Holistic Resiliency Approach for HPC systems

This track was led by ORNL and supported by the LATECH and NC State teams. We propose a tunable holistic resiliency approach for HPC to allow extreme-scale HPC machines to run in the event of frequent failures in such a manner that their capability is not severely degraded. At its core, a generic HPC RAS framework concept enables various combinations of proactive and reactive fault-tolerance (FT) mechanisms using a customization approach to adapt to individual system properties and application needs. Both proactive and reactive FT mechanisms provide an e□cient and complete coverage for all failures in a HPC system.

1.4 Towards failure prediction.

Our overarching goal is the ability to predict failures and to quantify how well failure prediction techniques must perform. We investigated techniques and critical factors to measure quality of failure prediction and proposed two novel metrics [1] measuring goodness of two major types of failure prediction: time-to-failure (TTF) prediction (regression type) and failure occurrence (FO) prediction (classification type). Instead of using balanced measurements such mean-squared error (MSE), we defined new metrics concerning the directional aspect of predicted failure time relative to the actual time. We also discovered that traditional metrics (MSE, MAD, precision and recall) were sometimes misleading. Further details of these new measurements can be found in

the publication [1]. We continue our work on failure prediction and hope to exploit these new metrics to derive new prediction techniques and further improve existing ones.

1.5 GPU-enabled checkpoint protocol and modeling

Due to the fact that the reliability and availability of a large scaled system inverse to the number of computing elements, fault tolerance has become a major concern in high performance computing (HPC) including a very large system with GPGPU. We introduced a two-phase checkpoint protocol [2], which presents a novel model of checkpoint/restart utilizing CUDA streamed and GPU virtualization. Basically, the protocol must first checkpoint GPU kernel states to CPU memory and then from a source CPU memory to a target memory system or reliable storage. We also focus on the latency hiding checkpoint/restart protocol on the GPU side and the models in which we can further improve overheads in data movement between the host and GPU. Our experimental results reveal the performance improvement due to the latency hiding checkpoint/restart mechanism in three different aspects; checkpoint overhead, restart overhead, and wasted time. Our simulation shows the influence of the latency hiding checkpoint/restart mechanism to the performance of the long-running application. We hope that our finding will be a foundation of new works that will be beneficial to recent DOE investments in large scale GPU-enabled systems and the future systems.

1.6 New HA-OSCAR

In March 2004, PI Leangsuksun & Scott introduced HA-OSCAR [3], an HA software stack that aims to improve resilience in an HPC cluster environment to open source community. Last year, we reintroduced a re-architected and re-implemented HA-OSCAR supported Ubuntu Linux distribution. The current version expands more platforms and environment supports including Redhat/CentOS, cloud/grid computing and distributed storages. The new HA-OSCAR [5,6] provides three HA-enabled foundation services, namely; monitoring, failure-over & user-defined FT action and replication services. The project has been used for a HA solution for important applications such as The Electronic Health Record Exchange[4]. The code repository and more details can be founded in [5].

1.7 Baler: Deterministic, lossless log message clustering tool

The rate of failures in HPC systems continues to increase as the number of components comprising the systems increases. System logs are one of the valuable information sources that can be used to analyze system failures and their root causes. However, system log files are usually too large and complex to analyze manually. There are some existing log clustering tools that seek to help analysts in exploring these logs, however they fail to satisfy our needs with respect to scalability, usability and quality of results. To address these issues we have developed a tool for log message pattern extraction (or log clustering), called “Baler”[7]. We seek to overcome the complex and volume issues while retaining the usability feature of condensing log files into a small number of informative sequences which can be expanded to reveal the underlying detail if desired. To this end we tokenize the data and perform clustering based on token attributes. Our methodology requires only single pass processing of the input data and is able to incrementally process log files without loss of global information. The patterns resulting are deterministic messages from different input datasets will always be represented by the same output patterns and clusters, thus allowing explicit comparison of results over long time periods.

2.4 Broader Dissemination and Educational Outreach

Four publications and talks in international conferences and workshops as well as an open source release.

- Organizing a 2009 workshop -- Resilience 2009 -- in Munich, Germany, on June 9, 2009. PI was a co-chair and primary organizer of the workshop that provided a venue for the discussion of ongoing research in HPC and Resilience issues.
- Organizing a 2010 workshop -- Resilience summit 2010 -- in Santa Fe, New Mexico, on October 13, 2010. PI was a co-chair and primary organizer of the workshop that provided a venue for the discussion of ongoing research in HPC and Resilience issues. In addition, the open source HA-OSCAR project has been adopted by other application specific project such as healthcare industry.

3.4 Discussion of any changes from the original proposed approach

We have explored fault issues in a GPGPU environment. This change is motivated by the large system that will be deployed at ORNL and others with the increasing popularity of GPGPU in HPC systems. We have produced results in this new area.

4.4 List of publications, talks, and project websites

[1] C. Leangsuksun et al, "*Using Log Information to Perform Statistical Analysis on Failures Encountered by Large-Scale HPC Deployments*", In Proceedings of the 4rd International Conference on Availability, Reliability and Security (ARES) 2009, Fukuaka, Japan, March 16-19, 2009.

[2] C. Leangsuksun, N. Naksinehabooon, R. Nassar, M. Paun, S. L. Scott, N. Taerat, "*Incremental Checkpoint Schemes for Weibull Failure Distribution*", submitted to International Journal of Foundations of Computer Science (IJFCS), April 2009.

[3] S. Scott, C. Engelmann, G. Vallee, T. Naughton, A. Tikotekar, G. Ostrouchov, C. Leangsuksun, N. Naksinehaboon, R. Nassar, M. Paun, F. Mueller, C. Wang, A. Nagarajan, J. Varma , "*A Tunable Holistic Resiliency Approach for High-Performance Computing Systems*", refereed poster at 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 09), Feb 2009, Raleigh, North Carlorina.

[5] M. Paun, C. Leangsuksun, R. Nassar, R. Gottumukkala, "RELIABILITY OF A SYSTEM OF K NODES FOR HIGH PERFORMANCE COMPUTING APPLICATIONS", *IEEE Transactions on Reliability* Vol 59, Issue 1, 2010..

[6] N. Naksinehaboon, N. Taerat, C. Leangsuksun, C. Chandler and S. Scott, "Benefits of Software Rejuvenation on HPC Systems", submitted to *The 2010 IEEE International Symposium on Parallel and Distributed Processing*, Taipei, Taiwan, September 6-9 2010

[7] HA-OSCAR Project website: http://xcr.cenit.latech.edu/ha-oscar.

[8] Petashare project website, www.petashare.org

[9] New HA-OSCAR code repository, http://github.com/okoye/HA-OSCAR

[10] HA-OSCAR: High Availability Computing Enabler, invited talk in *The 9th DOSAR Workshop, held in Pilanesberg, Sun City, South Africa, April 5-9 2010,* http://indico.cern.ch/conferenceDisplay.py?confId=85358.

[11] N. Taerat, C. Leangsuksun, C. Chandler and S. Scott, "Proficiency Metrics for Failure Prediction in High Performance Computing", The 2010 IEEE International Symposium on Parallel and Distributed Processing with Applications, Taipei, Taiwan, September 6-9 2010..

[12] S Laosooksathit, N Naksinehaboon, C Leangsuksun, to appear in THE 9TH ACS/IEEE INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS (AICCSA 2011), December 27th - December 30th 2011, Sharm El-Sheikh, Egypt

[13] HA-OSCAR Project website: http://xcr.cenit.latech.edu/ha-oscar.

[14] A. Kawtrakul, et al, Towards Development of a National Blueprint for Better and Smarter Healthcare Services in Thailand, SRII Global Conference 2011, San Jose, California, March 29 - April 2, 2011

[15] New HA-OSCAR code repository, http://github.com/okoye/HA-OSCAR

[17] N. Taerat, J. Brandt, A. Gentile, M. Wong, and C. Leangsuksun. Baler: Deterministic, lossless log message clustering tool. In Proceedings of the International Supercomputing Conference (ISC), Hamburg, Germany, June 19-23 2011. Springer Verlag, Berlin, Germany.