Software-based Resilience for Extreme-Scale Systems

Dorian Arnold, Patrick Bridges, Ron Brightwell, Kurt Ferreira, Guy Feldman, Alireza Goudarzi, Torsten Hoefler, Dewan Ibtisham, Scott Levy, Darko Stefanovic, Bryan Topp and Patrick Widener

1Sandia National Laboratories • 2Purdue University • 3ETH Zürich
Contact: Dorian Arnold, darnold@cs.unm.edu

Resilience at Exascale

Increased HPC system size and complexity will render extremely low mean times between failures (MTBF). If current checkpoint/restart mechanisms are used for future extreme scale systems, system utilization will suffer. We aim to enhance fault tolerance by:

- Studying the viability, limits and cost/benefit of checkpoint compression;
- Correlating and predicting failure events;
- Exploiting address-space similarities;
- Using simulations to study fault tolerance performance;
- Exploring the feasibility of HPC benchmarks for resilience research.

Compression-Based Checkpoint-Restart Optimizations

While perhaps an obvious optimization, checkpoint compression is not pervasive. Compression is viable if its benefits outweigh its costs. Using the Mantevo minapps, LAMMPS, BLCR and common compression utilities, we study checkpoint compression viability and evaluate application makespan improvements from compression.

- We observe that compression viability is independent of checkpoint sizes, process counts and user-level versus system-level checkpointing.
- We conclude that checkpoint compression is an effective strategy for improving the time and space efficiency of checkpoint/restart protocols.
- We also compare checkpoint compression against other software and hardware-based optimizations and explore its impact on power and energy usage.

Failure Event Correlation and Prediction

We study how HPC event analysis can be used for fault-tolerance:

- We use partial correlations to observe statistical significance and conditional correlations in HPC event data;
- We use information theory to understand the fundamental predictive power of failure event data;
- Using information theory results, we study reservoir computing and Bayesian classifiers for failure event prediction; and
- Based on partial correlation results, we use principal component analysis to understand how dimensionality reduction can apply to HPC event data.

This work contributes an understanding of the theoretical limits of HPC event prediction, the shortcomings of only relying on direct correlations, accurate failure prediction using only small windows of failure event history and the observation that principal component analysis can be used to significantly reduce event data volumes without loss of relevant information.
Exploiting Memory Content Similarities to Repair DRAM Errors

DRAM ECC failures are one of the most frequently observed sources of node failure in large-scale distributed systems. Because the rate at which failures occur increases proportionally to the number of processors, larger, more powerful systems will experience more frequent failures. We can prevent some ECC DRAM errors from leading to node failure by exploiting similarities in main memory. When a memory error occurs on a page that is similar to one or more other pages in the address space of an application, we can use the contents of the similar page to reconstruct the contents of the faulted page without needing to terminate the affected application. Our preliminary results (see figure) indicate that significant similarity exists in the memory of several important HPC applications. As a result, this approach may enable us to correct many errors that would otherwise result in node failures.

Simulating Exascale Resilience

Effective evaluation of fault tolerance mechanisms for exascale systems has been challenging; current systems are significantly smaller and typically have different architectural features (e.g., interconnect, persistent storage) than next-generation systems. Additionally, accurate analytical models do not yet exist for many emerging resilience techniques. As a result, we have developed a simulation framework for efficiently evaluating the performance of resilience techniques on future systems. Using this simulator we have demonstrated, for example, that we can accurately simulate a 128K node, 10 hour, production run of LAMMPS with a speed-up of over 10,000x. The efficiency of our simulator is due to three key observations: (1) like OS noise, fault-tolerance mechanisms and even faults themselves can be modeled as CPU detours; (2) we need only simulate those application and system features relevant to fault-tolerance; (3) we can forgo the overhead and complexity of cycle-accurate simulation because of the coarse-grained nature of resilience operations.

Benchmarks for Resilience

Classical HPC benchmarks conceived in an era with infrequent failures and emphasizing floating-point performance, will not provide the right type of information needed to evaluate resilience strategies for future exascale applications and systems. We show that even when using a more appropriate metric such as efficiency, the overhead characteristics of micro-benchmarks and proxy applications that compute similar problems can differ noticeably under uncoordinated checkpointing. These results indicate that benchmarking approaches for HPC applications and systems should be reconsidered for a new HPC regime where resilience concerns trump CPU performance.

Publications

[SC '12] Dewan Ibtesham, Dorian Arnold, Kurt Ferreira and Ron Brightwell, "Comparing GPU and Increment-based Checkpoint Compression”, (Poster.)  
[Resilience '12] Kurt Ferreira, Rolf Riesen, Dorian Arnold, Dewan Ibtesham and Ron Brightwell, "Viability of Compression to Decrease Message Log Sizes.”