Towards New Metrics for High-Performance Computing Resilience

Achievement: Developed new metrics for quantifying resilience of HPC applications that quantify the impact of various types of faults, errors and failures in HPC environment. Designed outcome-based metrics that measure resilience in terms of the impact of events on application reliability and performance. Demonstrated value of the new metrics for quantifying the combined impact of diverse fault types. Also demonstrated measurement of composite improvement in application resilience from implementing a multilayered resilience solution.

Significance and Impact:
We have designed outcome-based metrics that quantify the impact of faults, errors and failures on an application's reliability and performance. The proposed metrics have been designed to quantify the combined impact of these solutions on an application's ability to produce correct results and to evaluate their overall impact on an application's performance in the presence of various modes of faults. Resilience solutions are designed to deal with various types of faults, errors and failures that may result in outcomes ranging from having no effect on correctness or performance, performance degradation without affecting correctness, limited loss in precision to catastrophic failures. The proposed metrics are valuable for understanding the scope of impact, effective coverage and performance efficiency of existing and emerging resilience solutions.

Research Details:

- Demonstrated a set of new resilience metrics that provide a standardized way to measure the impact of fault events on an application's correctness and the performance efficiency independent of the nature of the fault.
- Developed the Resilience Factor (RF_{PE}), which captures performance efficiency (PE) of resilience solutions, i.e., the impact on performance on account of solutions employed for dealing with fault, error, failure events. The RF_{VE} is a value efficiency (VE) metric that captures the impact of events on output values, i.e., the application's scientific outcome.

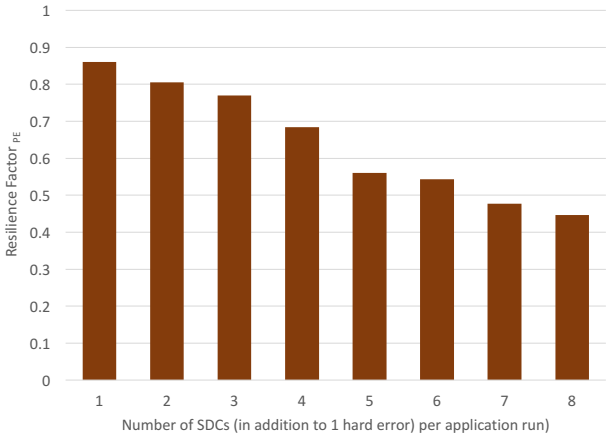

Figure 1: Using RF to measure resilience of a linear solver to multiple fault modes (silent data corruptions and process failures)

- Defined a Resilience Factor Yield (RY) for aggregating several RFs to quantify the aggregate resilience characteristics of an application.

Sponsor/Facility: This work was performed at Oak Ridge National Laboratory (ORNL). It was sponsored by the US Department of Energy Office of Science Early Career Research Program.

PI and affiliation: Christian Engelmann -- Computer Science and Mathematics Division (CSMD), Oak Ridge National Laboratory (ORNL)

Team: Saurabh Hukerikar, Rizwan A. Ashraf and Christian Engelmann

Publications:

Saurabh Hukerikar, Rizwan A. Ashraf, Christian Engelmann, Towards New Metrics for High-Performance Computing Resilience. In Proceedings of FTXS'17, ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC), Washington, DC, USA, June 26, 2017, DOI: http://dx.doi.org/10.1145/3086157.3086163

Overview:
Various types of faults are common in modern HPC environments, and the complexity of the hardware architectures and software environment make it difficult to trace the root cause and impact on application's outcome. HPC practitioners have long relied on dependability metrics for availability and system reliability, such as mean time to failure (MTTF) and its variants to measure resilience. These metrics impose a service-based view on system operation, which measure continuous service accomplishment in terms of the time to fatal failure. These metrics are incompatible for quantifying impact of fault events on the application resilience, a property which is concerned with keeping the HPC applications on the path to acceptable solutions despite the presence of faults, errors and failures. Resilience solutions allow applications to embrace a range of outcomes, including performance degradations without loss of correctness, losses in precision, wildly incorrect answers on account of silent errors to continuing execution after recovery from fatal crashes. We have defined outcome-based metrics called the Resilience Factor (RF) to measure the impact of various types of faults on the reliability and performance attributes of an HPC application, independent of the nature of the fault and the types of resilience solutions used by an application. We have demonstrated the value of using this metric to quantify resilience of a linear solver that is subjected to soft errors that result in silent corruptions and hard errors that result in process failures, and is protected using an algorithm-based mitigation solution in concert with a MPI-layer failure recovery solution.