SAND2018-10808

FY18 ASC CSSE L2 Milestone 6362: Local Failure Local Recovery (LFLR) Resiliency for Asynchronous Many Task (AMT) Programming and Execution Models

Executive Summary

Author: Keita Teranishi and Robert L. Clay, 08/30/2018

Introduction

The overall goal of this work was to perform an in-depth analysis of resilience schemes adapted to the Asynchronous Many-Task (AMT) programming and execution model with the goal of informing the Sandia Advanced Simulation and Computing (ASC) program's application development strategy for next generation platforms (NGPs).

Milestone Description

As written in the Milestone Reporting Tool (MRT), the milestone description is as follows:

"Resilience is an imminent issue of next-generation platforms (NGP) because of reliability concerns based on the architectural complexity and the constraint in system power budget. Under such unreliable computing systems, introducing failure mitigations at the runtime and application layers is essential for executing ASC applications. Our study involves the development of a prototype resilient AMT system to enable multiple task replay, replication strategies, and other AMT-based failure mitigations. The prototype code will facilitate 1) evaluating the scalability, performance, and costs for different AMT resilience options, and 2) assessing the accuracy/cost tradeoffs of the application-specific failure detection and mitigation schemes in the AMT context. Our empirical study will be performed with emulated failures either on a large-scale HPC system or simulated architecture to represent NGP. The outcome is expected to provide definitive guidance to the design roadmap of the AMT programming framework (e.g., ATDM-DARMA) development as well as the resilience-driven co-design of NGP. Additionally, the milestone activity will provide initial capability for AMT resilience on ASC apps.

Specific Deliverables:

1. Prototype implementation of resilience schemes for the asynchronous many-task model, including task replication and replay.
2. An analysis of the accuracy-cost tradeoffs, scalability, performance and costs for multiple AMT resilience options.
3. A report to inform the code development roadmap guiding the Sandia/ASC strategy for AMT resilience for NGP."

The Exit Criteria for this Milestone states: This milestone will be completed when the above-mentioned analyses are complete and a report is provided summarizing the key results and recommendations concerning AMT programming framework development for NGP.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Impact Statement

The outcome of the work indicates that (1) resilience techniques implemented within an AMT framework allow efficient and scalable recovery under frequent failures (approximately 100 detected soft failures per second), that (2) the abstraction of task and data instances in the AMT programming model enables readily usable and generic Application Program Interfaces (APIs) for resilience to meet a variety of error and failure scenarios, and that (3) this abstraction enables predicting the performance of resilient-AMT applications with a simple simulation infrastructure.

Summary of Work Done

We have completed the three deliverables mentioned above through the following efforts. Our resilient-AMT prototype enables multiple of AMT-based application recovery schemes including task replay, replication, algorithm-based fault tolerance (ABFT) and task-based checkpoint/restart. For analytical study, we have built a resilient-AMT simulator that reflects the resilience techniques and AMT runtime scheduler behavior implemented in our resilient-AMT prototype. Our resilient-AMT prototype and simulator were used to evaluate the tradeoffs between task replay and task replication as well as the performance on our AMT runtime on high performance computing (HPC) platforms.

Path Forward

The outcome of this work will help the design of resilience and fault tolerance capabilities of Sandia's existing parallel programming model effort such as DARMA and Kokkos. Furthermore, this new resilient programming model allows the unreliable computing system paradigm more explicitly, which in turn provides more flexible options in system co-design of NGPs in the era of post-Moore's Law computing.

The successful completion of this effort is given in the committee review memo (SAND2018-9668 R) and the final report (SAND2018-9672).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.