1-1-2022

Algorithm-Based Fault Tolerance at Scale

Joshua Dennis Booth
University of Alabama in Huntsville

RCEU 2022 Project Proposal

Project Title

Algorithm Based Fault Tolerance at Scale

Faculty Information

Name: Joshua Dennis Booth, PhD

Status: Assistant Professor

Department/Program: Computer Science

College: College of Science

Phone: 256.824.6433

UAH Email: joshua.booth@uah.edu

Proposal ID RCEU22-CS-JDB-01

RCEU 2022 Project Proposal

I. Project Description

One of the fundamental design principles for large distributed systems is fault tolerance. Fault tolerance is the ability for a system to recover or have some form of resiliency from errors. Fault tolerance methods are critical for scalable long-running applications and simulations (e.g., climate simulations). Two board categories of fault tolerance methods exist: application-specific and application-agnostic. Most current application-specific are algorithm specific and require detailed information about how the algorithm is implemented on hardware. On the other hand, application-agnostic may completely ignore any execution patterns and relies on expensive checkpointing strategies. However, a third middle-ground approach may exist based on observations that many of the scalable long-running simulations have a similar iterative update execution pattern. Though this observation is not as detailed as application-specific methods, it does provide room to base new fault tolerance methods that are adaptive and self-learning.

This work proposes the exploratory work related to building new scalable fault tolerance methods around the iterative nature found in large partial differential equations (PDEs) simulations. In particular, the work looks at how to parameterize the fault tolerance space and how to adapt these parameters during runtime to achieve some level of fault tolerance. In particular, the parameterization would be based on the initial work-related to gradient-based metrics of the solution vector to capture the intrinsic multi-scale resiliency of iterative applications. The work in developing an initial metric was started by Dr. Sun (University of Kansas), Dr. Raghavan (Vanderbilt University), and Dr. Gainaru (Oakridge National Lab). Dr. Booth, Dr. Sun, and Dr. Raghavan have continued to start parameterizing and automating this approach.

The resulting findings will have an impact across multiple areas related to scalable long-running simulations and may even be used by others at UAH.

II. Student Duties, Contributions, and Outcomes

a. Specific Student Duties

The student's duties will include reviewing literature about iterative simulations, current fault tolerance methods, and parallel computing. In the first three weeks (week 1-3), the student will be asked to review and summarize three journal papers written about scalable fault tolerance, and the student will be asked to learn the following under the guidance of several tutorials posted by Lawrence Livermore National Labs and XSEDES: login, learn to submit jobs, compile already written code, review basic Linux operations for our target hardware. In the next three weeks (week 4-6), the student will review one journal paper per week and will move into the testing phase. The testing phase will include running simulations written in MATLAB and C using MPI. The runs will allow for collecting vector solution trace data to help build models. The data will be collected in CSV format and analyzed using Microsoft Excel. In the next two weeks (week 7-9), the student will build models using the collected data and start to implement fault tolerance methods using the models. This will require modifying code in MATLAB and the low-level language of C. IN the last two weeks (weeks 9-10), the student will produce a Latex write-up and prepare a power in Microsoft PowerPoint explaining their work.

b. Tangible Contributions by the Student to the Project

The student will produce an online repository (GitHub) with programs, data, and writeups. The goal is to turn this poster into a piece to submit at either SuperComputing's ACM undergraduate

Poster Section or SIAM PP's poster section. The choice of venue will be determined by the outcome of the experiments.

c. Specific Outcomes Provided by the Project to the Student

The student will gain a deep insight into the importance of fault tolerance and the iterative applications that are common in high-performance computing simulations. They will also gain experience working on high-performance supercomputers. Learning to write good code and use these systems is an important milestone for anyone looking to enter the field of high-performance computing. Additionally, they will learn the important lesson of how to read and review research journals, and how to start writing their own in Latex.

III. Student Selection Criteria

The student must have completed course work in either CS 221 or CPE 212 and MA224. They must be comfortable working in a terminal environment (e.g., bash shell).

IV. Project Mentorship

The student will meet with Dr. Booth twice a week to discuss the progress of the project. Ad-hoc meetings will be set up as needed if issues come up. Dr. Booth will regularly monitor the student's repo to judge if the student is having issues. Dr. Booth currently has an undergraduate research student that will be moving into their MS under Dr. Booth, and he will ask this student to aid in some of the "getting started" questions about the computer systems. Ideally, the student will be able to work in a lab in the CS department, while Dr. Booth works in his office. If questions come up that need a quick response, the student can pop into his office.

Safety and Contingency Plan: The student will be required to take the XSEDE video tutorial on good practices for shared resource computing. The tutorial explains that these devices are shared and that all work is monitored, i.e., there is no expectation of privacy. Additionally, any illegal behavior, such as doing calculations for nuclear devices, running spamming attacks, password cracking, etc, will be reported to the proper authority. Additionally, all repositories the student writes to will not be open to public view (only the student and Dr. Booth will have access) to keep the privacy of the student. Ideally, research will happen on campus in Tech Hall. However, if face-to-face meetings are not able to happen, meetings will be done via Zoom or Microsoft Teams. Overall, this will have no real impact on the project, though may slow it down and be inconvenient for both the student and Dr. Booth.