ABSTRACT

Islam, Tanzima Zerin Ph.D., Purdue University, May 2013. Reliable and Scalable Checkpointing Systems for Distributed Computing Environments. Major Professor: Saurabh Bagchi.

By leveraging the enormous amount of computational capabilities, scientists today are being able to make significant progress in solving problems, ranging from finding cure to cancer – to using fusion in solving world's clean energy crisis. The number of computational components in extreme scale computing environments is growing exponentially. Since the failure rate of each component starts factoring in, the reliability of overall systems decreases proportionately. Hence, in spite of having enormous computational capabilities, these ground breaking simulations may never run to completion. The only way to ensure their timely completion, is by making these systems reliable, so that no failure can hinder the progress of science.

On such systems, long running scientific applications periodically store their execution states in *checkpoint files* on stable storage, and recover from a failure by restarting from the last saved checkpoint file. Resilient high-throughput and high-performance systems enable applications to simulate scientific problems at granularities finer than ever thought possible. Unfortunately, this explosion in scientific computing capabilities generates large amounts of state. As a result, todays checkpointing systems crumble under the increased amount of checkpoint data. Additionally, the network I/O bandwidth is not growing nearly as fast as the compute cycles. These two factors have caused scalability challenges for checkpointing systems. The focus of this thesis is to develop scalable checkpointing systems for two different execution environments – high-throughput grids and high-performance clusters.