THE RAMDISK STORAGE ACCELERATOR — A METHOD OF ACCELERATING I/O PERFORMANCE ON HPC SYSTEMS USING RAMDISKS

By

Timothy B. Wickberg

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Approved:

Christopher D. Carothers, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

November 2011
(For Graduation December 2011)

ABSTRACT

I/O performance in large-scale HPC systems has failed to keep pace with improvements in computational performance. This widening gap presents an opportunity to introduce a new layer into the HPC environment that specifically targets this divide.

A *RAMDISK Storage Accelerator* (RSA) is proposed; one that leverages the high-throughput and decreasing cost of DRAM, while providing an application-transparent method for pre-staging input data and committing results back to a persistent disk storage system. The RSA is constructed from a set of individual RSA nodes; each with large amounts of DRAM and a high-speed connection to the storage network. Memory from each node is made available through a dynamically constructed parallel filesystem to a compute job; data is then asynchronously staged on to the RAMDISK ahead of compute job start, and written back out to the persistent disk system after job completion.

The RAMDISK thus provides for very-high-speed, low-latency access that is dedicated to a specific job; the asynchronous data staging frees the compute system from time that would otherwise be spent waiting for file I/O to finish at the start and end of execution.

To support this asynchronous data-staging model requires an method of scheduling this new capability alongside that of the traditional task of scheduling compute resource access for each job. A proof-of-concept implementation based on the SLURM job scheduler is presented, and demonstrates an operational 16-node RSA system connected to a 1024-node IBM Blue Gene/L.

This thesis presents a case in favor of this RAMDISK Storage Architecture along with specific work done to implement a proof-of-concept system demonstrating the feasibility of this mode of operation.