The CODES Project

Enabling Co-Design of Multi-Layer Exascale Storage Architectures

Robert Ross (Technical Lead), Phil Carns, Argonne National Laboratory
Christopher Carothers, Rensselaer Polytechnic Institute

Overview

The data demands of science and the limited rates of data access place a daunting challenge on the designers of exascale storage architectures. *Co-design* of these systems will be necessary to find the best possible design points for exascale systems. Designers must consider performance and reliability in the context of the I/O patterns and requirements of applications and analysis tools at exascale. Meeting these constraints will require the development of a multi-layer hardware and software architecture incorporating devices that do not yet exist. The most promising approach for co-design of such systems is simulation.

The goal of this project is to enable the exploration and co-design of exascale storage systems by providing a detailed, accurate, and highly parallel simulation toolkit for exascale storage. We will develop models to realistically represent application checkpoint and analysis workloads. These models will be joined together using the Rensselaer Optimistic Simulation System (ROSS), a discrete-event simulation framework that allows simulations to be run in parallel, decreasing the simulation run time of massive simulations to hours. Building on our prior work in highly parallel simulation and using our new high-resolution models, our system will capture the complexity, scale, and multi-layer nature of exascale storage hardware and software, and it will execute in a time frame that enables “what if” exploration of design concepts.

With this new toolkit we will investigate design options and trade-offs related to improving the reliability and performance at scale of potential exascale storage architectures. We will work with industry, DOE computing facilities, and the computer and computational science communities to refine our models and to encourage the use of this powerful tool in the design of future extreme-scale storage systems.

Impact

The project will advance the goals of the ASCR mission through development of an exascale storage simulation framework, deployment of the framework to storage system researchers, vendors, and experts in the community, and most importantly, through discovery of novel storage architectures, storage software algorithms, and scalable application interfaces to storage. Through better I/O concurrency and system resilience, this research will enable the scalability of scientific applications of interest to the Department of Energy and ultimately, to advance scientific discovery.

In the near term, storage architectures are likely to evolve to include storage in the compute system, but otherwise will look similar to those seen on systems at Oak Ridge and Argonne today. Accurate simulation of the systems will help us better understand potential bottlenecks and allow us to compare to revolutionary storage system alternatives.

Rensselaer

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

The CODES Project

Recent and Current Activities

Exascale Network Modeling

Effective simulation and co-design of exascale storage system is dependent upon accurate models for a variety of HPC system components, including I/O workloads, compute nodes, interconnect networks, I/O forwarding, file servers, and storage devices. We have recently enhanced our interconnect modeling capability by developing high-fidelity models for two candidate network topologies for exascale systems: the multidimensional torus network (as used on the Blue Gene/Q architecture) and the dragonfly network (as used on the Cray XC30 architecture). We used ROSS create a high-fidelity flit-level simulation of a million-node torus and dragonfly network. We validated our results against empirical measurements on Blue Gene/P and Blue Gene/Q systems as well as existing cycle-accurate (but less scalable) dragonfly simulators. We then explored how tradeoffs in dimensionality, link speed, routing algorithms, and router configuration affect a variety of large-scale network patterns for each network topology.

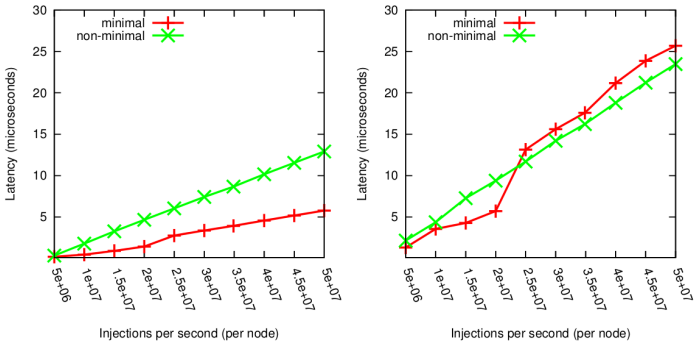

Figure 1. Impact of routing algorithm on average and maximum latency for a 1.3M node dragonfly with bisection traffic.

We have also developed a simulation abstraction layer, known as "modelnet", that allows multiple network models to be used interchangeably as pluggable components of larger scale system models. This capability will allow us to rapidly evaluate candidate storage architectures with minimal changes to the modeling framework.

Simulator Configuration and Load Balancing

In the course of our work we have recognized the importance of developing reusable model components that can be combined in different ways for architecture evaluation and co-design. To that end we have developed an extension to the ROSS parallel discrete event simulation framework that provides a common framework for composing reusable sub-models into end-to-end system models at run time. This framework also provides a mechanism to control the mapping of model components to parallel discrete event simulation processes. We are currently exploring how to leverage this capability for improved load balancing of simulation computation and discrete event traffic.

I/O Workload Models

I/O workload models are another key component of exascale co-design. In previous work we simulated workloads using a custom I/O workload description language. While this technique is highly flexible, constructing workloads based on existing applications is a manual process. We have therefore generalized the interface for injecting I/O workloads into our simulations to allow for interchangeable use of multiple, and in some cases automated, sources of workload data. We are currently developing techniques for translating instrumentation of production HPC applications into model workloads for high-fidelity simulation of large-scale application workloads.

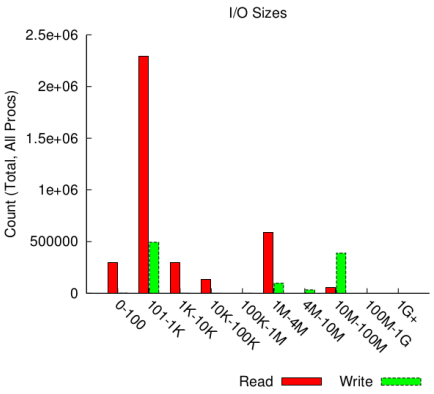

Figure 2. Example of I/O statistics gathered from an HPC application using Darshan. We can leverage data from Darshan and other similar tools to generate realistic I/O workload models.

Contacts

For more information on the project, please contact Robert Ross <rross@mcs.anl.gov>. For additional information the ROSS simulation system and possible uses, please contact Christopher Carothers <chrisc@cs.rpi.edu>.

Rensselaer

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

The CODES Project