The NoLoSS Project

Investigating the roles of node local storage in Exascale systems

Kamil Iskra (Technical Lead), Rob Ross, Argonne National Laboratory
Bronis R. de Supinski, Maya Gokhale, Kathryn Mohror, Lawrence Livermore National Laboratory

Overview

The international computational science community is on a path to build exaFLOP-capable systems by the year 2018. These exascale systems will enable transformative science discoveries in a number of areas, including climate, combustion, nuclear energy, and national security. A key exascale barrier is the need for scalable storage of persistent state: one that provides the necessary I/O bandwidth and capacity without overwhelming the power, cooling, and cost budgets of an exascale system. Traditional global storage system approaches simply cannot scale to meet these requirements.

With the development of inexpensive, nonvolatile memory technologies such as flash memory and phase change memory, it is feasible to include solid state persistent memory on every node in a future exascale system – enabling in-system storage (also referred to as node local storage). In-system storage augments the memory hierarchy, potentially reducing DRAM requirements and thus the node's power requirements. It streamlines and simplifies checkpointing, increasing system reliability. In-system storage reduces the peak bandwidth requirements of a global exascale storage system, offering a scalable checkpoint/restart solution. However, there remain considerable research challenges to realizing these potential benefits, especially if one wants to hide the complexity introduced by another layer in the storage hierarchy from the user.

The goal of the Node Local Storage Systems (NoLoSS) project is to conduct a detailed assessment of the potential roles and benefits of in-system storage in exascale computational science. We are exploring existing hardware options for NLS and assess the software mechanisms that best exploit them based on a detailed analysis of existing Office of Science applications. We are implementing important examples of those mechanisms and determining how modifications to the existing hardware mechanisms could better support them. We will continue this three-pronged, iterative process throughout the project's lifetime, including anticipating how our successes will alter I/O usage patterns of emerging exascale applications.

Impact

Our project combines the unique strengths of two national laboratories. Through our research collaborations, our connections to major Department of Energy computing centers, and our participation in exascale computing activities in the community, we are interacting with application teams, vendors, and system software experts to share our ideas and to gather additional input. In-system storage is a low-risk solution to the serious issue of I/O access at exascale.

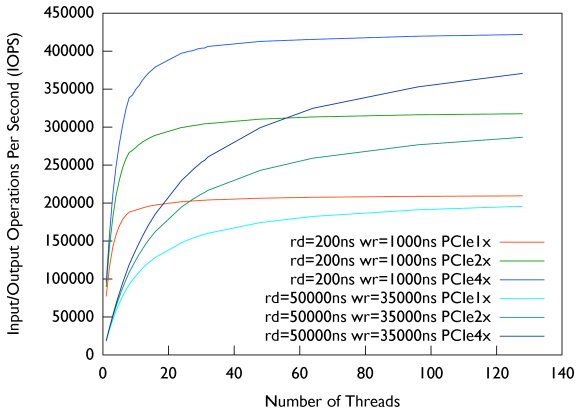
We are leading the co-design of successful in-system storage solutions, ensuring that the best designs are available to vendors for inclusion in exascale architectures, with software ready to use them.

U.S. DEPARTMENT OF ENERGY
Lawrence Livermore National Laboratory
Argonne NATIONAL LABORATORY

The NoLoSS Project

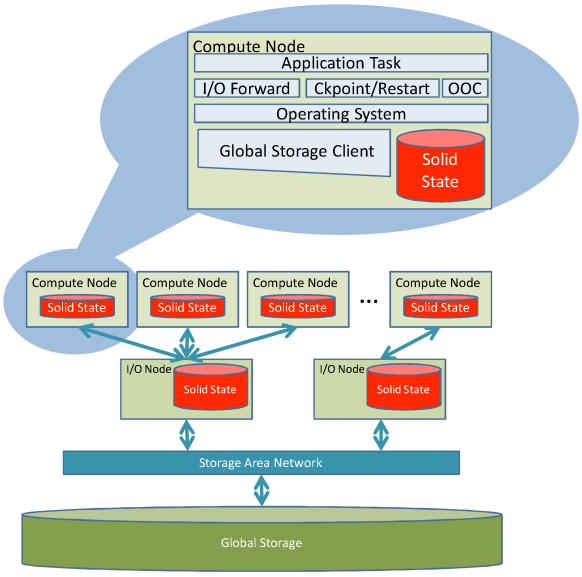
Current Activities

We are developing a simulator to model the delay function for accessing application data stored in a node local non-volatile memory. The simulator will allow us to analyze quantitatively the impact of using a variety of permanent memory technologies and associated interconnect options on applications, libraries, and system software. Based on a device driver derived from the Linux block RAM disk (RD), the simulator will be able to model delays that are consistent with today's PCI-E flash cards and the predicted delays for PCM-based solutions attached via a high-speed peripheral bus. The delay model includes asymmetric read/write of the requested payload and a pipeline model to characterize burst requests.



Simulated performance for a multi-threaded benchmark with varying device read/write latencies and I/O bus speeds in a 50-50 read/write random workload using 4K pages.

The Scalable Checkpoint/Restart Library (SCR) addresses the problem of prohibitively high checkpoint times for large-scale applications. SCR is a multi-level checkpointing library that checkpoints to storage on the compute nodes in addition to the parallel file system. Its low-cost checkpoints to in-system storage are 100–1000x faster than the parallel file system and reduce the load on the file system by a factor of two. In addition, we have found that checkpoints stored locally protect against 85% of observed system failures. Current research on SCR includes developing methods for checkpoint compression and coordination of asynchronous writes of checkpoints from in-system storage to the parallel file system.



System overview. In-system storage used for data staging on the compute nodes and I/O nodes is marked in red.

We are working on a scalable data staging infrastructure. Based on the I/O Forwarding Scalability Layer (IOFSL) project, this infrastructure will provide write-behind buffering on in-system storage for burst traffic such as checkpoint files. By staging data on compute nodes and possibly also I/O nodes and moving it to global storage later, we can flatten I/O bandwidth peaks by spreading them in time, generating a more desirable I/O signature. We expect this infrastructure to enable us to meet I/O requirements with a less capable, less expensive I/O system, yet allow applications to return to computations more quickly.

We are also coordinating with the CODES project with the goal of assisting in the incorporation of more accurate solid state storage models and architectures into the CODES simulation system and eventually using this system to better understand future architectures.

Contacts

For more information on the project, please contact Kamil Iskra <iskra@mcs.anl.gov>. For additional information on the SCR system and possible uses, please contact Kathryn Mohror <mohror1@llnl.gov>.

U.S. DEPARTMENT OF ENERGY
Lawrence Livermore National Laboratory
Argonne NATIONAL LABORATORY