U.S. Department of Energy Best Practices Workshop on
File Systems & Archives
San Francisco, CA
September 26-27, 2011
Position Paper

Venkatram Vishwanath, Mark Hereld and Michael E. Papka
Argonne National Laboratory
<venkatv, hereld, papka>@mcs.anl.gov

ABSTRACT

The performance mismatch between the computing and I/O components of current-generation HPC systems has made I/O a critical bottleneck for scientific applications. It is therefore crucial that software take every advantage available in moving data between compute, analysis, and storage resources as efficiently as networks will allow. Currently available I/O system software mechanisms often fail to perform as well as the hardware infrastructure would allow, suggesting that improved optimization and perhaps adaptive mechanisms deserve increased study.

We describe our experiences with GLEAN – a simulation-time data analysis and I/O acceleration infrastructure for leadership class systems. GLEAN improves the I/O performance, including checkpointing data, by exploiting network topology for data movement, leveraging data semantics of applications, exploiting fine-grained parallelism, incorporating asynchronous data staging, and reducing the synchronization requirements for collective I/O.

INTRODUCTION

While the computational power of supercomputers keeps increasing with every generation, the I/O systems have not kept pace, resulting in a significant performance bottleneck. The *ExaScale Software Study: Software Challenges in Extreme Scale Systems* explains it this way: ``Not all existing applications will scale to terascale, petascale, or on to exascale given current application/architecture characteristics" citing ``I/O bandwidth" as one of the issues. On top of this, one often finds that existing I/O system software solutions only achieve a fraction of quoted capabilities.

We have developed an infrastructure called GLEAN [1,2] to accelerate the I/O of applications on leadership systems. We are motivated to help increase the scientific output of leadership facilities. GLEAN provides a mechanism for improved data movement and staging for accelerating I/O, interfacing to running simulations for co-analysis, and/or an interface for in situ analysis via a zero to minimal modification to the existing application code base. GLEAN has scaled to the entire infrastructure of the Argonne Leadership Class Facility (ALCF) comprising of 160K Intrepid IBM Blue Gene/P (BG/P) cores and demonstrated multi-fold improved with DOE INCITE and ESP applications. We discuss some of the lessons learned which could be considered for best practices on file systems and archives.

OUR POSITION

Based on our experiences with GLEAN, we believe the useful components to improve the I/O performance on leadership class systems include topology-aware data movement, leveraging data semantics, incorporating asynchronous data staging, leveraging fine-grained parallelism, and non-intrusive integration with applications. We briefly elucidate these.

Topology-aware Data Movement: As we move towards systems with heterogeneous and complex network topologies, effective ways to fully exploit their heterogeneity is critical. The IBM BG/P has five different networks with varying throughputs and topologies. The 3D torus interconnects a compute node with its six neighbors at 425 MB/s over each link. In contrast, the tree network is a shared network with a maximum throughput of 850 MB/s to the I/O nodes. The tree network is the only way to get to the I/O nodes in order to perform I/O. BG/Q is expected to have a more complex network topology. Similarly, several other Top-500 supercomputers have complex topologies. As seen in Figure 1, by leveraging the various network topologies, in GLEAN, we achieve up to 300-fold improvement in moving data out from the BG/P system. Another critical aspect is that our data movement mechanism uses reduced synchronization mechanisms wherein only neighboring processes need to co-ordinate their I/O. This is critical as we move towards future systems with millions of cores.

Fine-grained Parallelism: GLEAN's design employs a thread-pool wherein each thread handles multiple connections via a poll-based event multiplexing mechanism. This is critical in future many-core systems with low clock-frequency per core, where multiple threads are needed to drive the 40 Gbps and higher network throughputs per node to saturation.

Asynchronous data staging refers to moving the application's I/O data to dedicated nodes and next writing this out to the filesystem asynchronously while the application proceeds ahead with its computation. Asynchronous data staging helps satisfy the bursty nature of application I/O common in computational science and blocks the simulation's computation only for the duration of copying data from the compute nodes to the staging nodes. Data staging also significantly reduces the number of clients seen by the parallel filesystem, and thus mitigates the contention including locking overheads for the filesystem. Staging mitigates the variability in I/O performance seen in shared filesystems on leadership systems when accessed concurrently by multiple applications.

Leveraging Application Data Models: I/O system software typically use stream of bytes and files to deal with an application's data. A key design goal in GLEAN is to make application data models a first-class citizen. This enables us to apply various analytics to the simulation data at runtime to reduce the data volume written to storage, transform data on-the-fly to meet the needs of analysis, and enable various I/O optimizations leveraging the application's data models. Toward this effort, we have worked closely with FLASH, an astrophysics application, to capture its adaptive mesh refinement (AMR) data model. We have interfaced with PHASTA, which uses an adaptive unstructured mesh, to make unstructured grids supported in GLEAN, and with S3D, a turbulence simulation, to capture it's structured grid model. We have worked with many of the most common HPC simulation data models ranging from AMR grids to unstructured adaptive meshes.

Non-Intrusive Integration with Applications: Application scientists are very interested in I/O solutions wherein they can get the added performance improvements without having to change their simulation code (or with minimal changes). To achieve this, we have mapped Parallel-netCDF and hdf5 APIs, commonly used high-level I/O libraries in simulations, to relevant GLEAN APIs, thus enabling us to non-intrusively interface with simulations using pnetcdf and hdf5.

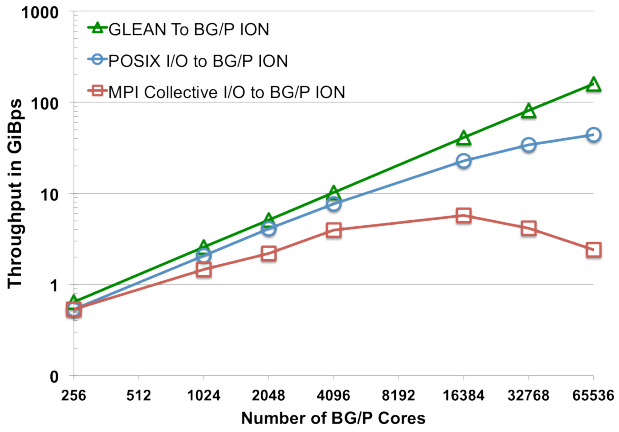

Figure 1: Strong scaling performance of the I/O mechanisms to write 1 GiB data to the BG/P IONs (log-log scale) on ALCF infrastructure

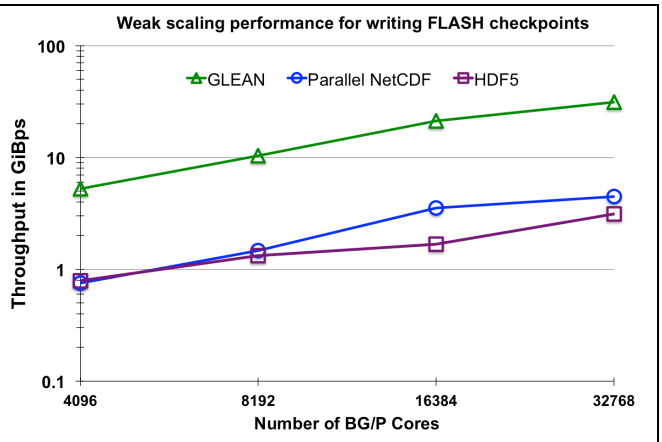

Figure 2: Weak scaling results for writing FLASH checkpoint data

REFERENCES

1.V. Vishwanath, M. Hereld, V. Morozov, and M. E. Papka, "*Topology-aware data movement and staging for I/O acceleration on Blue Gene/P supercomputing systems*", To appear in IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), Seattle, USA, November 2011.

2.V. Vishwanath, M. Hereld, and M. E. Papka, "*Simulation-time data analysis and I/O acceleration on leadership-class systems using GLEAN*", To appear in IEEE Symposium on Large Data Analysis and Visualization (LDAV), Providence, RI, USA, October 2011.