SDAV's Data-intensive Management Techniques and Tools

D. Boyuka (NCSU), S. Byna (LBNL) P. Carns (ANL), J. Dayal (GA Tech), G. Eisenhauer (GA Tech), Z. Gong (NCSU), T. Jin (Rutgers), S. Kumar (Utah), S. Klasky (ORNL), S. Lakshminarasimhan (NCSU), R. Latham (ANL), Q. Liu (ORNL), M. Parashar (Rutgers), V. Pascucci (Utah), N. Podhorszki (ORNL), R. Ross (ANL), N. Samatova (NCSU/ORNL), K. Schwan (GA Tech), E. Schendel (NCSU), A. Shoshani (LBNL), Q. Sun (Rutgers), V. Vishwanath (ANL), M. Wolf (GA Tech/ORNL), J. Wu (LBNL), F. Zhang (Rutgers), X. Zhang (GA Tech)

Simulations are generating an unprecedented amount of data, facilitated by the rapidly increasing computational capabilities of leading compute resources. Additionally, the systems on which these simulations execute are becoming increasingly complex. Organizing, transforming, indexing, and reducing data to enable effective analysis, as well as carefully managing data movement and orchestrating data analysis in complex system architectures, are significant challenges in extreme scale, data intensive scientific discovery.

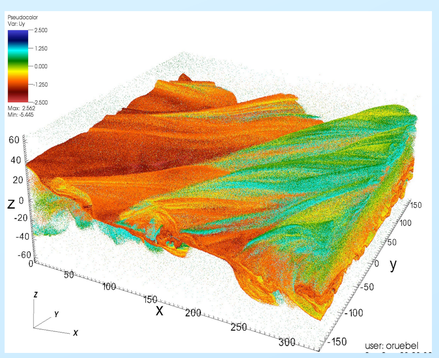
FastBit Indexing

FastBit is an open-source data processing library following the spirit of NoSQL movement. It offers a set of searching functions supported by compressed bitmap indexes. It treats user data in the column-oriented manner similar to well-known database management systems such as Sybase IQ, MonetDB, and Vertica.

FastBit allows one to quickly find patterns of interest in large datasets. Applications include:
- Improving energy-efficient combustion by tracking ignition
- Identifying regions of special significance for next-generation fusion reactors
- Designing better drugs through use of molecule docking software by companies such as BioSolveIt
- Detecting cyber attacks on computer systems
- Analysis of many terabytes of web log traffic by Yahoo!
- Analysis stock market data in the financial sector

Collisionless magnetic reconnection, a space weather phenomenon such as reaction of Earth's magnetosphere to solar winds, is simulated by the VPIC code. FastQuery, a parallel API built on FastBit indexing was used to locate highly energetic particles of a trillion-particle simulation.

This research has led to several discoveries including confirmation of preferential acceleration of particles along magnetic field, existence of agyrotropy near the reconnection hot-spot, and correlation of energetic particles with flux ropes (right).



Gene Context Analysis is a new way of determining its function based on the genes in a neighborhood. The most expensive part of Gene Context Analysis is to compare the query organism against other organisms from the same family. FastBit is also being applied to dramatically increase the rate of comparison in this field.

Flexpath Messaging

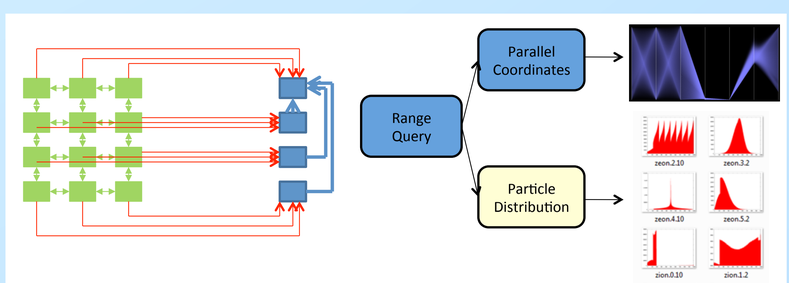
Efficient execution of parallel I/O pipelines is critical across a number of scenarios, including:
- data staging methods for running analytics and visualization
- data streaming and the online QoS control of such data streams
- aggressive use of source-based data reduction and filtering
- convenient ways to carry out remote data visualization

To meet these challenges, we have developed Flexpath, an event-based messaging middleware, which brings an active messaging approach to I/O. We use **active messages** to perform a number of tasks, such as format exchanges, metadata exchanges, scheduling of I/O, and transmission of the data itself.

Flexpath's active messaging provides a publish/subscribe paradigm, allowing for the decoupling of the interacting components that enables greater scalability, an ability to customize data streams, and improved fault management.

Flexpath is built on top of the EVPath event-based messaging layer, and using dynamic code generation (via CoD), users can create and deploy data conditioning plug-ins that can perform filtering and transformation operations on live data streams. The widely adopted ADIOS is used as an interface into Flexpath. In addition to operating on sockets and ENET, EVPath is built on top of the NNTI interface and thus can operate on a number of lower level transports: Gemini, Portals, Infiniband.

The figure below shows the reduction tree that implements the range query over GTS particles, resulting in the visualizations at right.

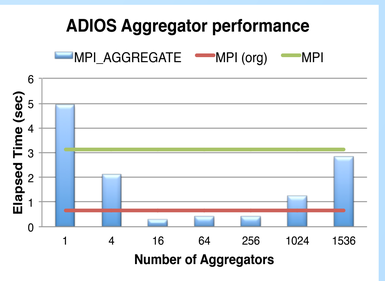


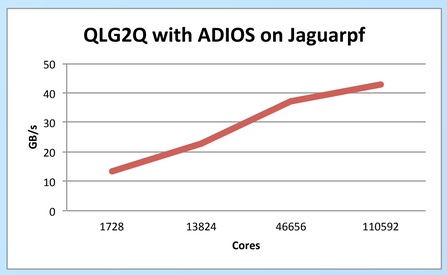
Adaptable I/O System (ADIOS)

ADIOS is an adaptable, scalable, and high-performance I/O system software based using I/O plugins, a parallel binary-packed file format, and a unified framework for accessing data from both persistent storage and data streams.

Approach:
- Unified I/O framework for both file and stream processing
- Developed scalable parallel file format (BP)
- Aggregate I/O method (for files) for small/mid data size
- Streaming Read API
- Rich set of data selection types for analysis needs

Using ADIOS aggregation, checkpoint writing for the XGC1 fusion particle code was accelerated by 11.5x as compared to MPI-IO with and without Lustre stripe-aligned writes (right).





For the QLC2Q quantum turbulence code, ADIOS has enabled runs of up to 110K processes writing at upwards of 40 GB/sec (left).

ADIOS also serves as a framework for integration of many other tools into application I/O paths, such as the Flexpath and DataSpaces work described here.

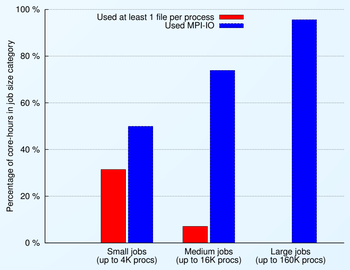
SDAV
Scalable Data Management, Analysis, and Visualization

U.S. DEPARTMENT OF ENERGY | Office of Science

The Scalable Data Management, Analysis, and Visualization Institute

Darshan I/O Characterization

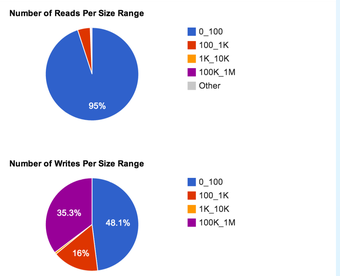
Darshan is a lightweight, scalable I/O characterization tool that transparently captures I/O access pattern information from production applications. It is currently deployed in production to automatically collect data on the Mira IBM Blue Gene/Q and Intrepid IBM Blue Gene/P systems at the Argonne Leadership Computing Facility as well as the Hopper Cray XE6 at the National Energy Research Scientific Computing Center. The broad scope of data collected with Darshan can be used to tune applications, guide I/O research activity, and inform future procurement decisions.



This figure illustrates the prevalence of key I/O characteristics across three job size categories on Intrepid. This data covers 13,613 unique jobs spanning the first five months of 2013. On Intrepid we see that file-per-process workloads (in which each application process opens a unique file) become increasingly rare at larger job sizes, while MPI-IO usage becomes increasingly common. These trends highlight how application I/O behavior can vary at different scales in order to improve performance or simplify data management.

IO Summary from Darshan

Start	End	Wallclock (secs)	MB Read	MB Written	Estimated I/O Rate (MB/sec)	Estimated Percent Time Spent in I/O
04-05 16:04:05	04-05 16:06:51	166	590.3	597.6	355.52	2.01%



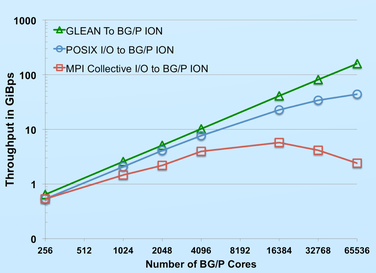
Example of NERSC web portal data available to Hopper end-users once a job has completed. Darshan data is used to automatically produce a graphical summary indicating the access size distribution, the amount of data accessed, estimated performance, and percentage of time spent performing I/O. Users can use this data as a starting point to evaluate I/O performance and validate application behavior.

GLEAN I/O Framework

GLEAN is a flexible and extensible framework for simulation-time data analysis and I/O acceleration taking into account application, analytics, and system characteristics to perform **the right analysis at the right place and time**.

Key Features of GLEAN include:

- Data staging, reduced synchronization semantics, and asynchronous I/O to improve I/O performance.
- Topology-aware data movement to improve I/O performance on current and future generation supercomputers characterized by high-radix interconnects
- Leverages application data models prevalent in HPC including unstructured meshes, structured grids, and AMR grids.
- Support for in situ, in transit, and co-analysis.
- Seamless and non-intrusive integration with applications using an interposer mechanism.

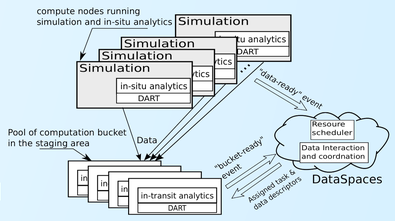


Strong scaling performance for 1GiB data movement off ALCF Intrepid Blue Gene/P compute nodes. GLEAN provides 30-fold improvement over POSIX I/O at scale. Strong scaling is critical as we move towards systems with increased core counts.

See the SDAV Cosmology poster for more on GLEAN.

DataSpaces

DataSpaces is a programming system targeted at current large-scale systems and designed to support dynamic interaction and coordination patterns between scientific applications. DataSpaces provides a semantically specialized shared-space abstraction using a set of staging nodes. This abstraction derives from the tuple-space model and can be associatively accessed by the interacting applications of a simulation workflow (right).

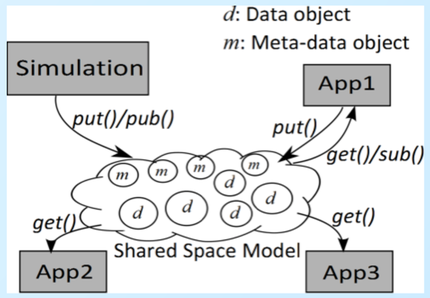


Approach:

- Virtual shared-space programming abstraction
- Distributed, associative, in-memory object store
- Adaptive cross-layer runtime management
- Efficient, high-throughput/low-latency asynchronous data transport

Features:

- **Mapping and Scheduling**: this service manages the in-situ/in-transit placement of online data processing operations as part of the coupled simulation-analytics workflows.
- **Scalable Messaging**: the service enables publish/subscribe/notification type messaging patterns to the scientists (below). The messaging system allows scientists to (1) dynamically subscribe to data events in regions of interest, (2) define actions that are triggered based on the events, and (3) get notified when these events occur.



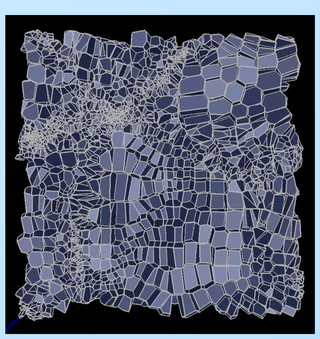
Parallel NetCDF (PnetCDF)

Parallel netCDF is a library providing high-performance I/O while maintaining file compatibility with Unidata's netCDF. This provides file portability, interoperability with the wide array of tools that support netCDF, and support for very large netCDF files. PnetCDF is used across a variety of domains, for these reasons.

In the HACC cosmology code, Parallel netCDF for two main tasks: periodically storing particle data and other outputs, and producing a Voronoi tessellation (under development). These products support analysis tasks; other methods handle checkpoint data. When using Parallel netCDF, the HACC code achieves a significant fraction of theoretical storage bandwidth while still maintaining the benefits of a self-describing portable file format.

For example, peak parallel I/O bandwidth to 128 Object Storage Targets (OSTs) on the NERSC Hopper machine is 26.7 GiB/sec. HACC's Parallel netCDF approach, which stores index data alongside raw particle arrays, has been shown to deliver 56 percent of this rate in test runs (below).

PnetCDF is also being evaluated for storing output of in situ Voronoi tesselation calculation (right). This type of data is useful for exploring gravitational lensing and the morphology of cosmic structures.

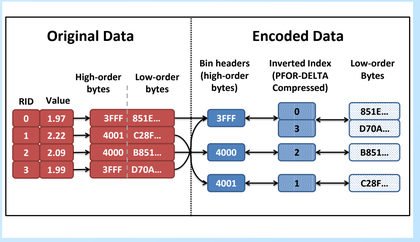


DIRAQ

DIRAQ is a parallel, in situ framework for unobtrusively indexing and compressing distributed scientific data at simulation runtime. DIRAQ employs a novel data encoding scheme that enables overall I/O and storage reduction while also including a storage-light-weight, aggregated index structure in support of efficient query processing.

Features:

- An aggressive index compression technique that takes advantage of spatial locality in data values in each process (right).
- An in-network indexing layout scheme for large-scale indexes that results in aggregated, defragmented indexes on disk, ideally suited for global-context analysis.
- A data- and memory-aware, adaptive aggregation algorithm, built on GLEAN's process/network topology awareness.



DIRAQ produces indexes that are up to 6x smaller than those of start-of-the-art indexing techniques, with the defragmented indexes resulting in an order of magnitude improvement in query performance. Overall, the integrated data compression and indexing method in DIRAQ results in a storage footprint of only 55-90% of the raw data size (a net reduction, including the added query index).

MLOC

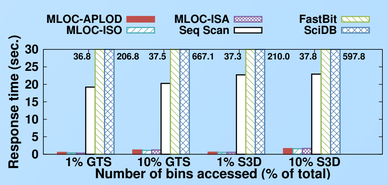
MLOC is a Multi-level Layout Optimization framework for Compressed scientific datasets, which optimizes queries over complex and heterogeneous access patterns without the need for replicating the original datasets. MLOC provides a flexible, layered architecture for applying layout optimizations hierarchically, based on analysis requirements.

For accelerating queries on heterogeneous access patterns, MLOC utilizes the following optimizations:

- Chunking and Hilbert space-filling curve mapping for space-constrained sub-volume access.
- Binning for value-constrained range queries.
- Byte-level division for precision-based multi-resolution access.

MLOC has demonstrated up to 600x speed-up of overall query response time for heterogeneous access patterns compared to state-of-the-art techniques. MLOC, driven by compression in the backend, can offer 8-82% savings in storage over currently used scientific database management techniques.

Query response time (sec.) on 8 GB datasets. For region-retrieval queries, no space constraints are set and value selectivity is 1% and 10%. MLOC-APLOD (APLOD precision-level layout), MLOC-ISO (ISOBAR lossless compression) and MLOC-ISA (ISABELA lossy compression) are three MLOC-based approaches.



Contact Us

Data Management Area Leads
Scott Klasky <klasky@ornl.gov>
Rob Ross <rross@mcs.anl.gov>

ADIOS and ADIOS Data Staging
Norbert Podhorszki <pnorbert@ornl.gov>

DataSpaces
Manish Parashar <parashar@rutgers.edu>

Flexpath
Karsten Schwan <schwan@cc.gatech.edu>

FastBit
John Wu kwu@lbl.gov

GLEAN
Venkat Vishwanath <venkatv@mcs.anl.gov>

PIDX
Valerio Pascucci <pascucci@sci.utah.edu>

APLOD, DIRAQ, ISOBAR, and MLOC
Nagiza Samatova <samatova@csc.ncsu.edu>

Darshan
Phil Carns <carns@mcs.anl.gov>

Parallel netCDF
Rob Latham <robl@mcs.anl.gov>
Wei-keng Liao <wkliao@ece.northwestern.edu>

SDAV
Scalable Data Management, Analysis, and Visualization

U.S. DEPARTMENT OF ENERGY | Office of Science