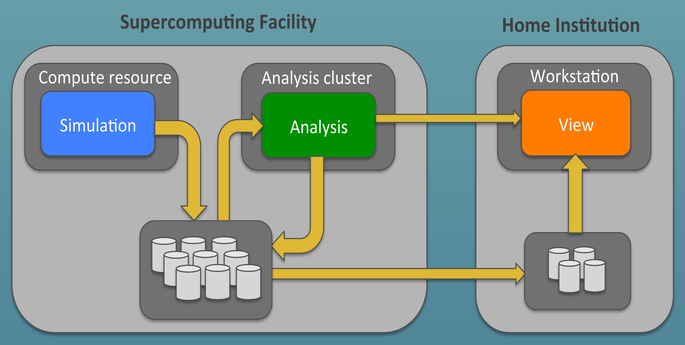
Accelerating I/O and Analytics at Extreme Scales

Venkatram Vishwanath, Mark Hereld, Joseph Insley, and Michael E. Papka

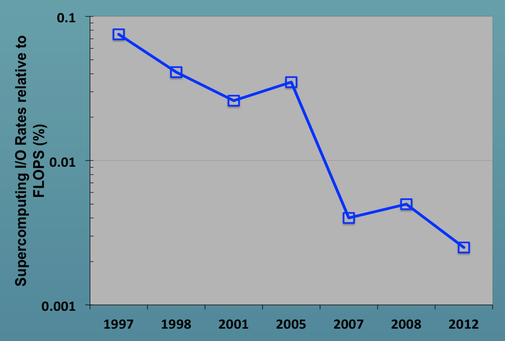
Argonne Leadership Computing Facility, Mathematics and Computer Science Division
Argonne National Laboratory
venkat@anl.gov

Traditional Science Pipeline



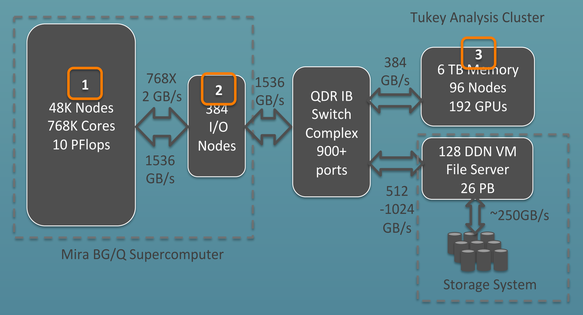
- Post processing is commonly used for data analysis
- Storage systems have been unable to meet I/O requirements
- Time to discovery is high as we constantly need to access storage

Storage Trends



Storage rates relative to the Compute rates of the #1 supercomputer in the Top 500 list over time. Storage rates have failed to keep up with computational trends, and this is expected to be even worse in future

Promising Alternatives



in situ

- uses simulation data structures
- no additional resources required
- increases memory and simulation time
- requires changes to the simulation

co-analysis

- extremely flexible analysis including time-varying and exploratory possible
- no memory or compute overhead
- requires additional hardware resources

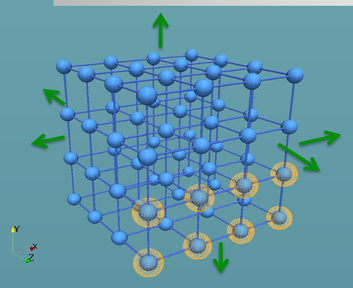
There are multiple possible approaches. Our goal is to design solutions that best fit the science needs to perform the right analysis at the right place and time

GLEAN is a framework to enable simulation-time data analysis and I/O acceleration taking into account system architecture, application and analysis requirements, and data semantics

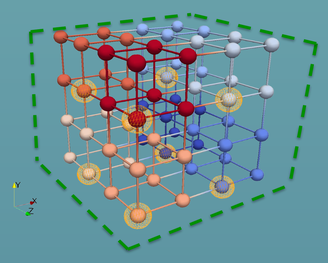
- Scaled to the entire ALCF-2 Mira BG/Q system consisting of **768K** BG/Q cores
- Used for production science and facilitated > 15 PBs of I/O to date
- Algorithms developed as part of GLEAN are being integrated into MPI-IO

GLEAN Characteristics

Exploit Network Topology and Minimize Synchronization

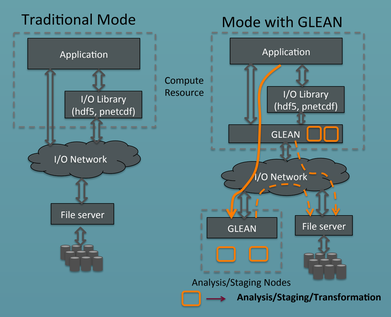


MPI-IO: Global Comm



GLEAN: Pset Bounded Comm

Seamless Integration



Key characteristics of GLEAN include asynchronous data staging, leveraging data semantics, exploiting parallelism in end-systems, burst-buffer, reduced synchronization, and ability to embed simulation-time analysis

Infrastructure	Simulation	Method
Co-analysis	PHASTA	Visualization using Paraview
I/O Acceleration	FLASH, HACC	Staging and Direct I/O
In situ	FLASH	Fractal Dimension
In flight	MADBench2	Histogram

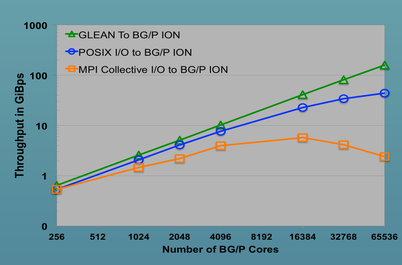
Impact

- Multi-fold improvement in I/O for multiple applications
- Simulation-time analysis of applications enabling faster science insights and helping overcome I/O bottlenecks
- Spawned rethinking of Analysis at simulation time
- Data architectures to meet the needs of computational science

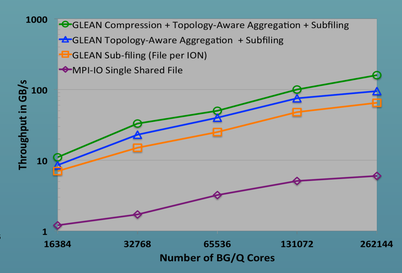
Future Work

- Scaling to Edison and Titan
- Data services including layout, analytics, and indexing
- Work closely with science teams to refine GLEAN

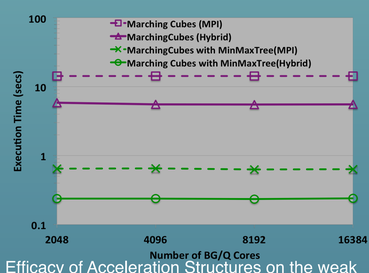
Accomplishments



Strong scaling performance of the various I/O mechanisms to move 1 GiB of data from BG/P compute nodes to the BG/P I/O Nodes



Efficacy of GLEAN subfiling, compression, and topology-aware aggregation on the achievable I/O Performance on BG/Q for HACC



Efficacy of Acceleration Structures on the weak scaling performance of computing marching cubes using MPI and with Hybrid Parallelism (MPI +OpenMP) with a per process volume of 256x256x256 as we scale from 2048 cores of BG/Q (4096x4096x2048) to 16,384 cores of BG/Q (8192x4096x4096)

- Achieved 160 GB/s for HACC production simulations on Mira BG/Q system, enabled Gordon Bell science runs, and written and read multi-PBs of data.
- Demonstrated simulation-time analysis at scale for PHASTA using ParaView

Publications:

- H. Bui, V. Vishwanath, J. Leigh, H. Finkel, S. Habib, K. Heitmann, M. Papka, "*Scalable Parallel I/O on the BG/Q Supercomputer using Compression, Topology-Aware Data Aggregation, and Subfiling*", In Euromicro PDP, 2014
- S. Habib, V. Morozov, N. Frontiere, H. Finkel, A. Pope, K. Heitmann, V. Vishwanath, et al., "*HACC: Extreme Scaling and Performance Across Diverse Architectures", In ACM/IEEE Supercomputing 2013 (Gordon Bell Finalist)*
- S. Lakshminarasimhan, D. Boyuka, S. Pendse, X. Zou, J. Jenkins, V. Vishwanath, M. Papka, N. Samatova, "*Scalable in situ Scientific Data Encoding for Analytics Query Processing", In HPDC 2013 (Best Paper Award)*
- V. Morozov, V. Vishwanath, K. Kumaran, J. Meng, and M. E. Papka, "*Early Experience on the Blue Gene/Q Supercomputing System*", In IPDPS, 2013
- V. Vishwanath, M. Hereld, V. Morozov and M. Papka,"*Topology-Aware Data Movement and Staging for I/O Acceleration on Leadership Computing Systems*", In IEEE/ACM Supercomputing, 2011
- V. Vishwanath, M. Hereld and M. Papka,"*Toward Simulation-time Data Analysis and I/O acceleration on Leadership systems*", In IEEE Large Data Analysis and Visualization (LDAV), 2011

Collaborators: MCS-Argonne, LCF-Argonne, HACC HEP-Argonne, FLASH Center (UChicago), Valerio Pascucci (University of Utah), Nagiza Samatova (NCSU), Jason Leigh (UIC), Kwan Liu-Ma (Davis), Ken Jansen (Colorado), Kitware Inc., CESAR Co-design Center,

Acknowledgements: Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357, an Argonne National Laboratory, and DOE-supported ASC / Alliance Center for Astrophysical Thermonuclear Flashes at the University of Chicago under contract B523820