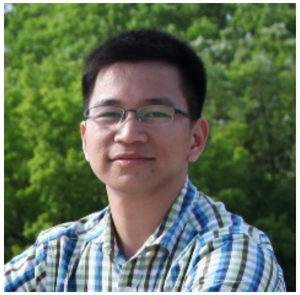
Addressing I/O Bottlenecks and Simulation-time Data Analytics at Extreme Scale

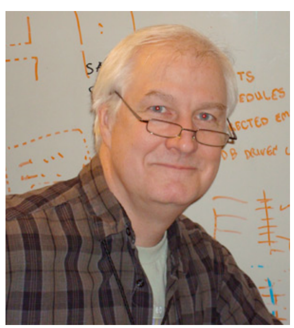
Venkatram Vishwanath

Argonne National Laboratory

venkat@anl.gov

Contributors include





Acknowledgements

- DOE Office of Advanced Scientific Computing Research
- Argonne Leadership Computing (ALCF)
- ANL - Mike Papka, Mark Hereld, Joseph Insley, Silvio Rizzi, Tom Uram, Jiayuan Meng, Vitali Morozov, Eunsung Jung, Kalyan Kumaran, Phil Carns, Rob Ross, Rob Latham Kevin Harms, Jeff Hammond Susan Coughlan, Katrin Heitmann, Salman Habib, Hal Finkel, Adrian Pope, Todd Munson, Sven Leyffer, Raj Kettimuthu, Steve Crusan and ANL team
- FLASH Center – Chris Daley, George Jordan, Anshu Dubey, John Norris, Randy Hudson, Carlo Graziani and Don Lamb
- Kitware - Pat Marion and Berk Geveci
- Univ of Colorado– Ken Jansen, Michel Rasquin and Ben Matthews
- Univ of Utah– Valerio Pascucci and Sidharth Kumar

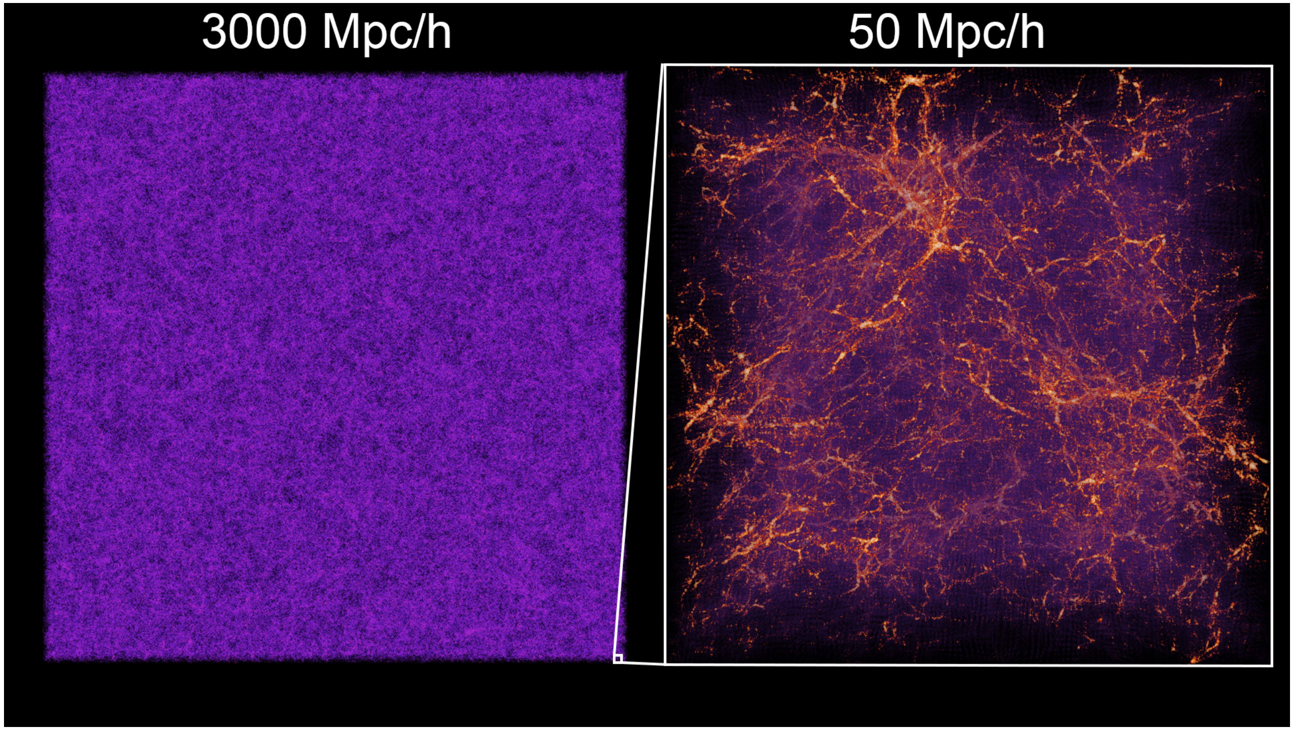
venkat@anl.gov

Acknowledgements

- UC Davis– Nick Leaf and Kwan-Liu Ma
- NCSU– Nagiza Samatova, Sriram L., Drew B.
- LBNL- John Wu, Prabhat, Suren Byna
- NTU Taiwan– Jerry Chou, Albert Chiu
- IBM – Paul Coffman
- HDF5 Group – Quincey Koizol

venkat@anl.gov

Data Scale and Requirements



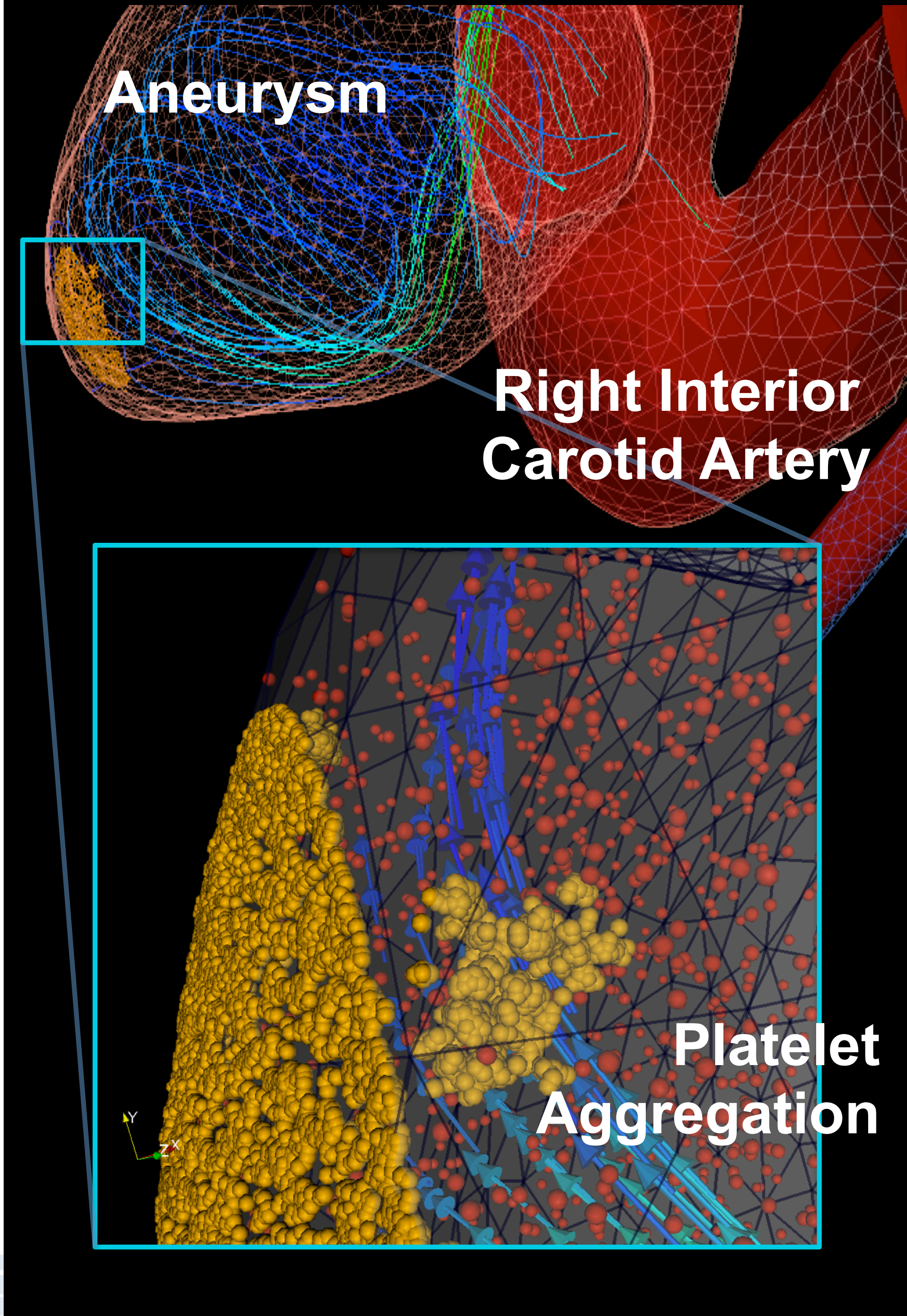
HACC Cosmology Simulation

– 14 Pflops sustained performance on 1.6 Million cores
– 20 PB and counting on Mira
– Checkpoints files are 400TB, and analysis outputs are 10s TB

Dataset Complexity

- Complexity as an artifact of science problems and codes:
  - Coupled multi-scale simulations generate multi-component dataset.
  - Atomistic data representations for plasma, red blood cells, and platelets from MD simulation.
  - Field data for ensemble average solution generated by spectral element method hydrodynamics code

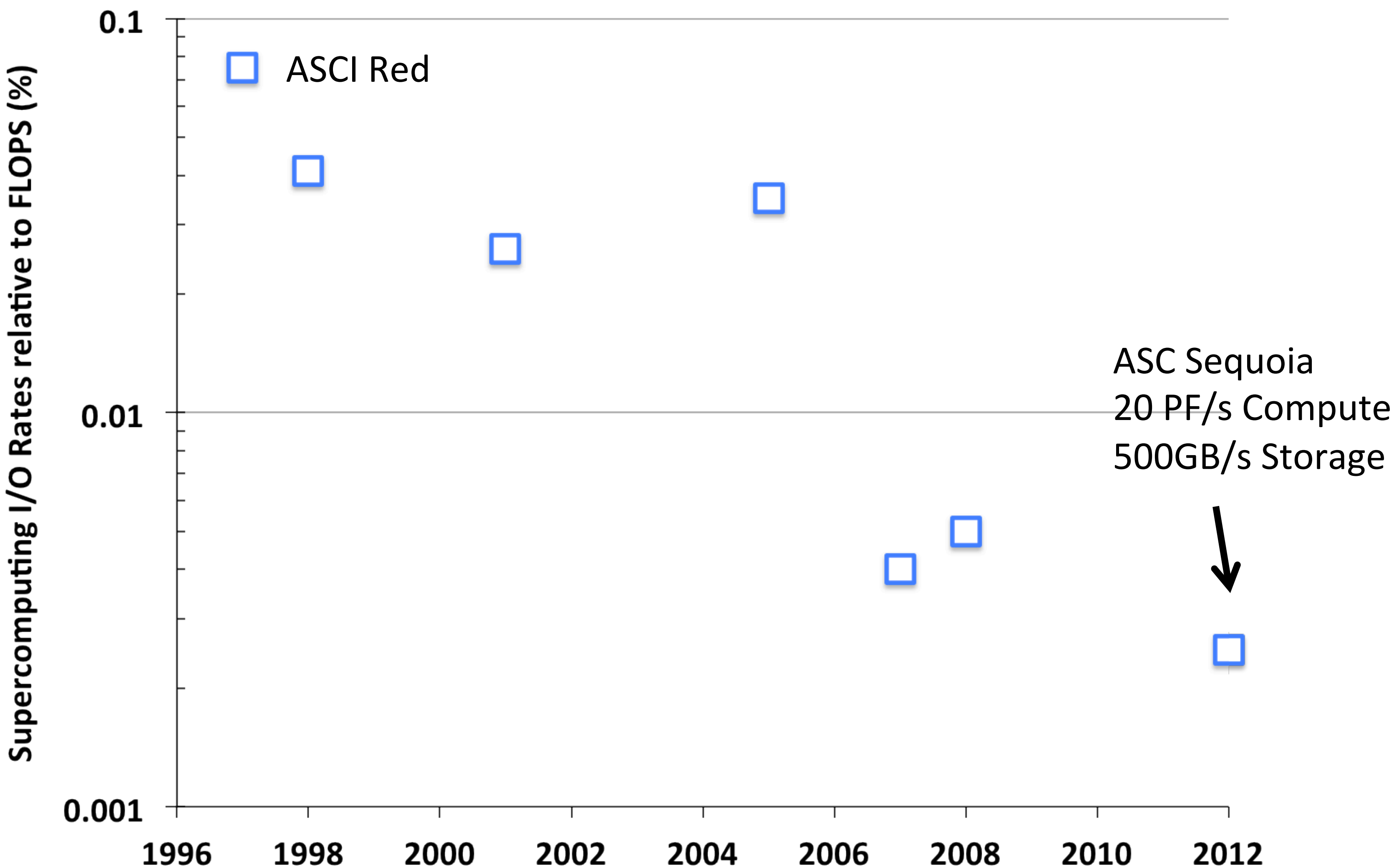
SC 2011 Gordon Bell Winner



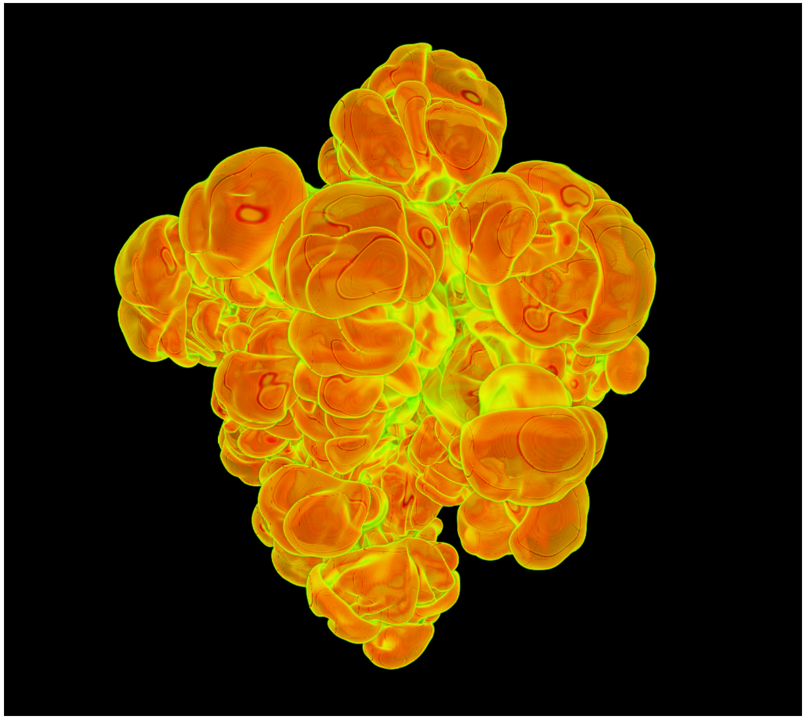
Scale and Complexity of Systems

System	Blue Gene/Q	K Computer	Tianhe-1A	
Peak Perf **# of Racks**	20 PF 96	11.3 PF 864	4.7 PF 112	
# of cores	1,572,864	705,024	202,752	
Processor	PowerPC	SPARC 64	Xeon X5670	NVIDIA M2050
Mem per core **(Flops/byte)**	1 GB 4.9	8 GB 1	1 GB 0.75	0.21 GB 3
Interconnect	5D Torus	6D Torus	Fat Tree	
Power	6 MW	12.7 MW	4.04 MW	
Gflop/watt	3.4	0.19	1.2	

Storage vs Computation Trends



FLASH Astrophysics I/O performance



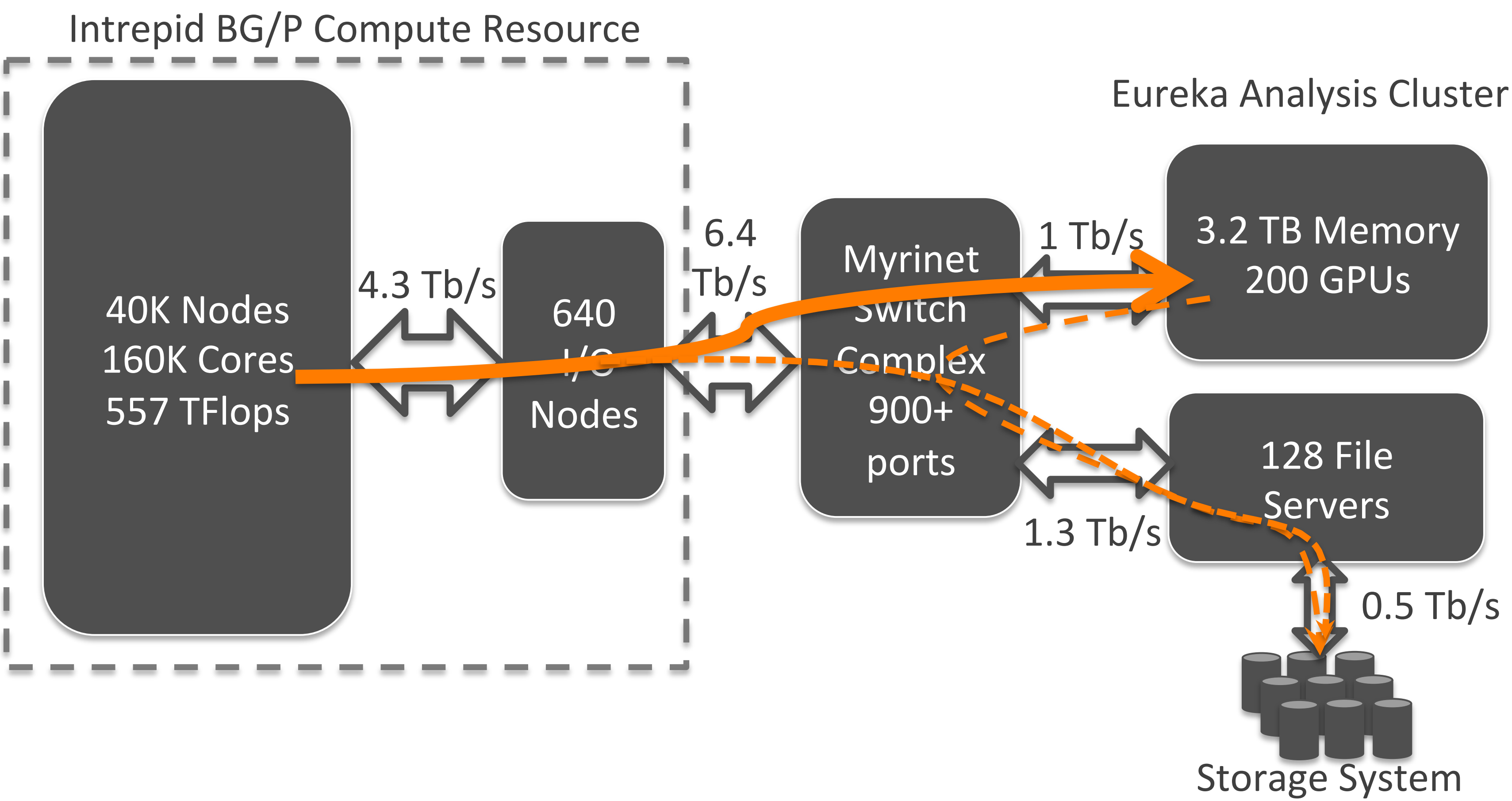
System Peak	65 GiB/s
IOR benchmark	35 GiB/s
FLASH Checkpoint	1 GiB/s
FLASH Plot files	0.2 GiB/s

During large-scale capability runs, up to 30% of time spent in I/O

Approaches to Address Data Challenges

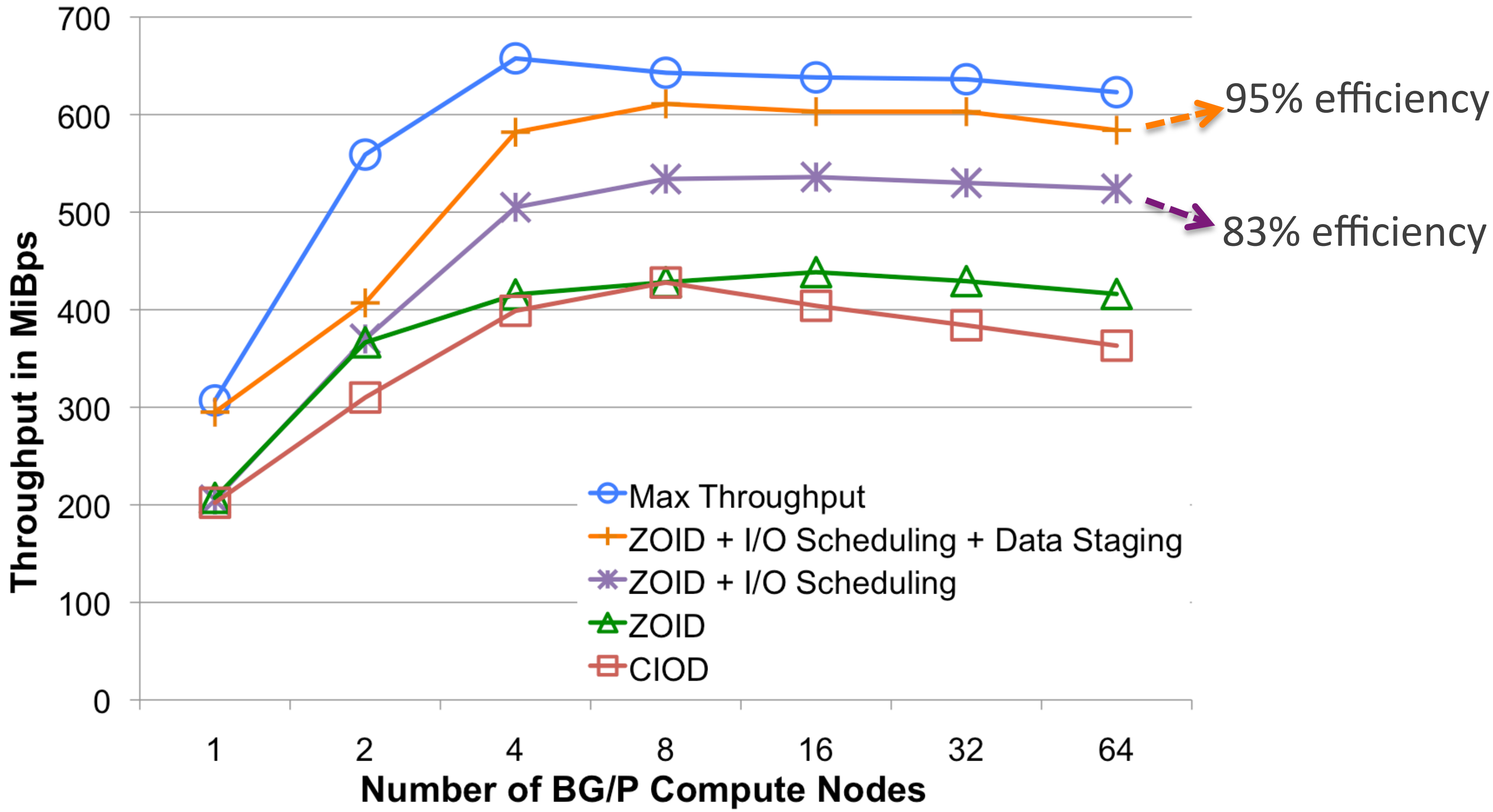
- Developing novel infrastructures via data staging and simulation-time analysis
- Leveraging application data models
- Scalable algorithms using reduced synchronization semantics and topology-aware data movement
- Exploiting data layouts
- Scalable analysis and visualization algorithms
- Work with applications and demonstrate at scale

Data Staging to improve I/O performance



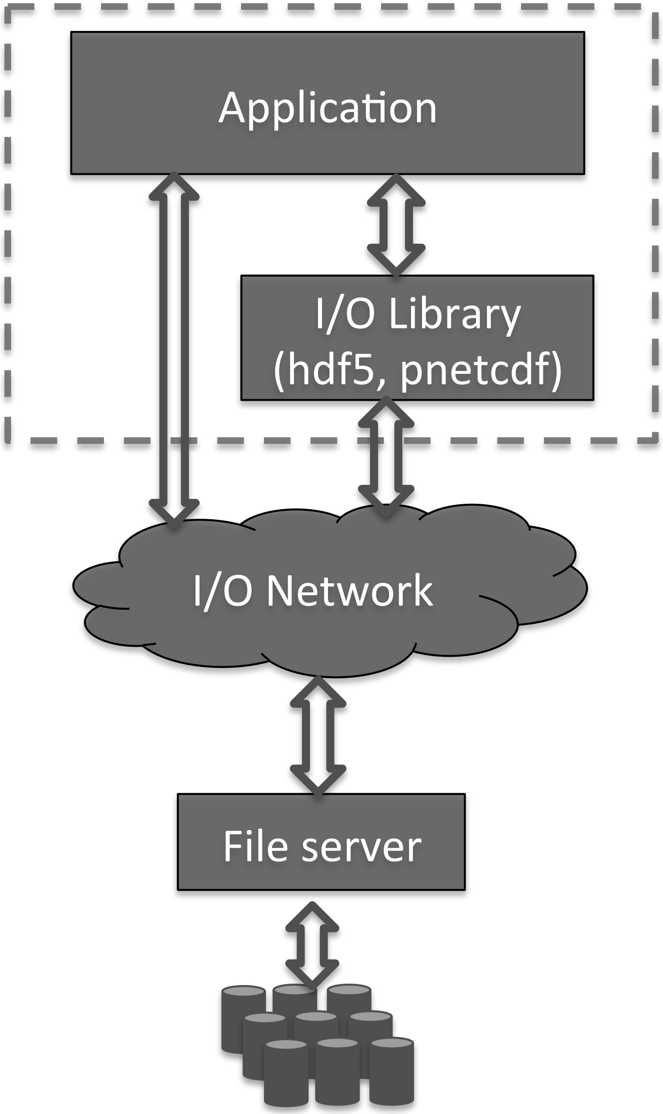
Staging enables the application I/O to be written out asynchronously while enabling the simulation to proceed ahead, and helps sink bursty I/O

Data Staging on I/O Forwarding Nodes (SC'10)

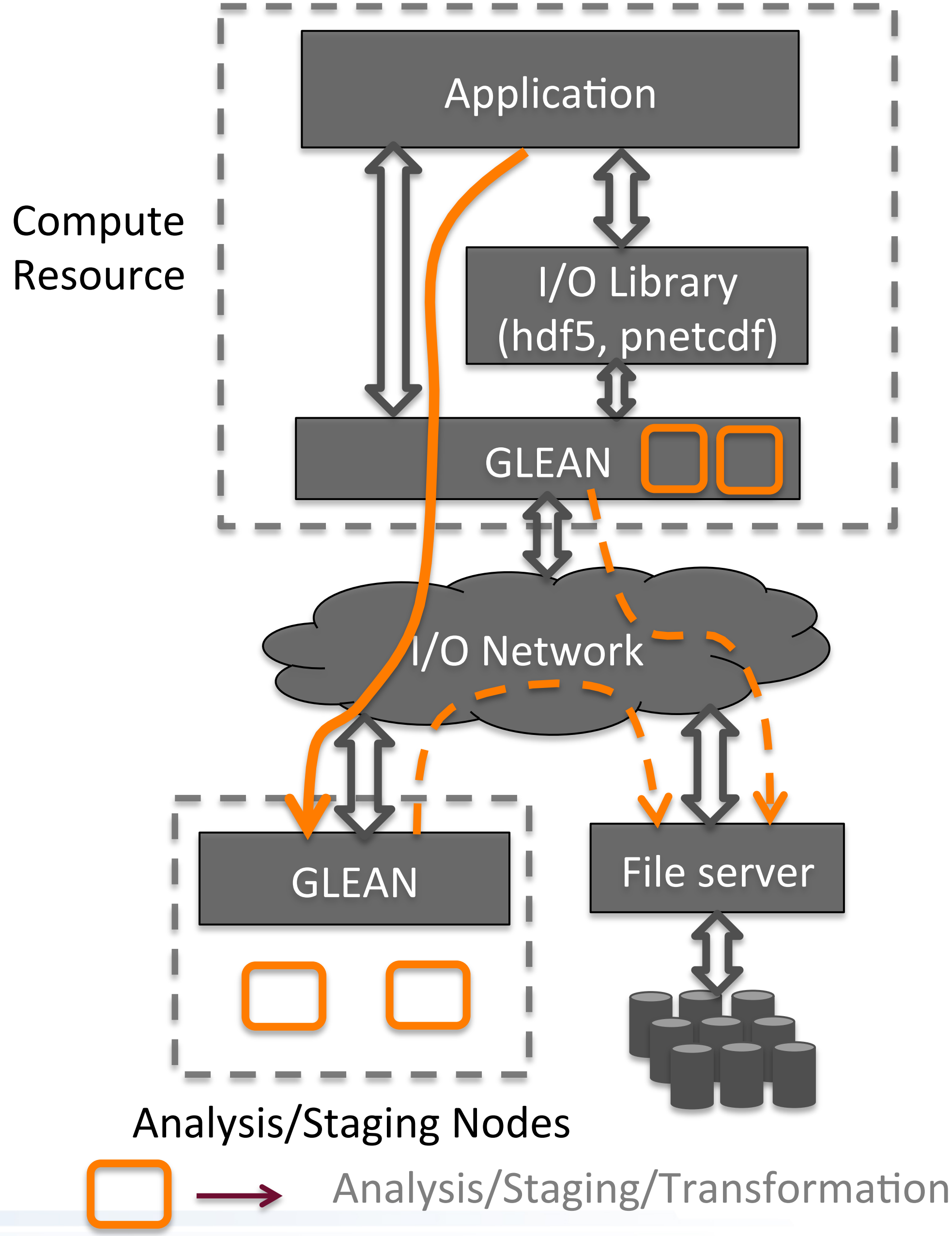


As we move towards exascale systems consisting of 1000s of low-power cores, effective I/O scheduling and data staging mechanisms will be of critical importance **(SC'10)**

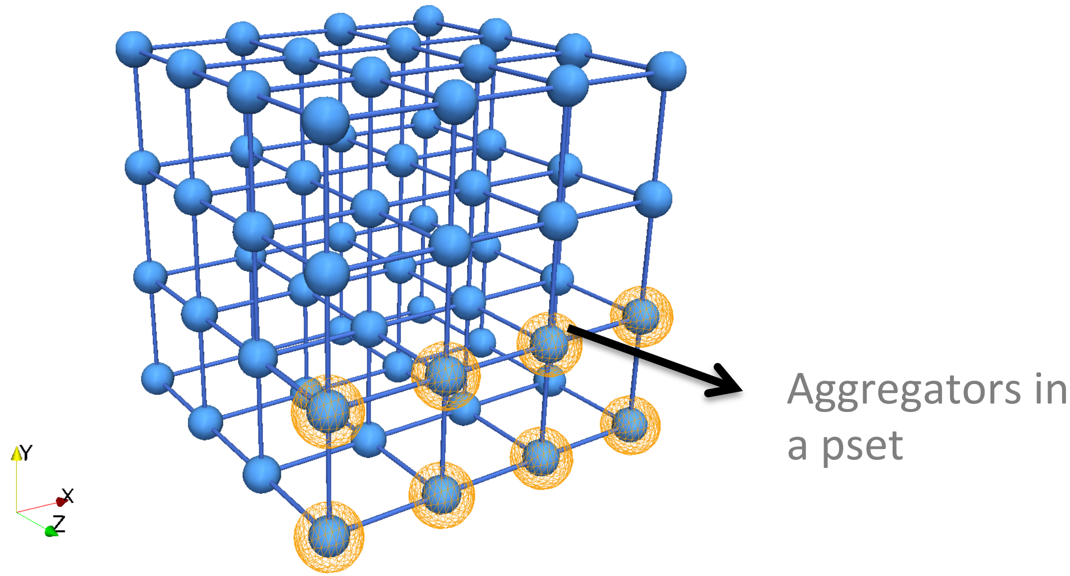
Traditional Mode



Mode with GLEAN



MPI Collective I/O on BG/P

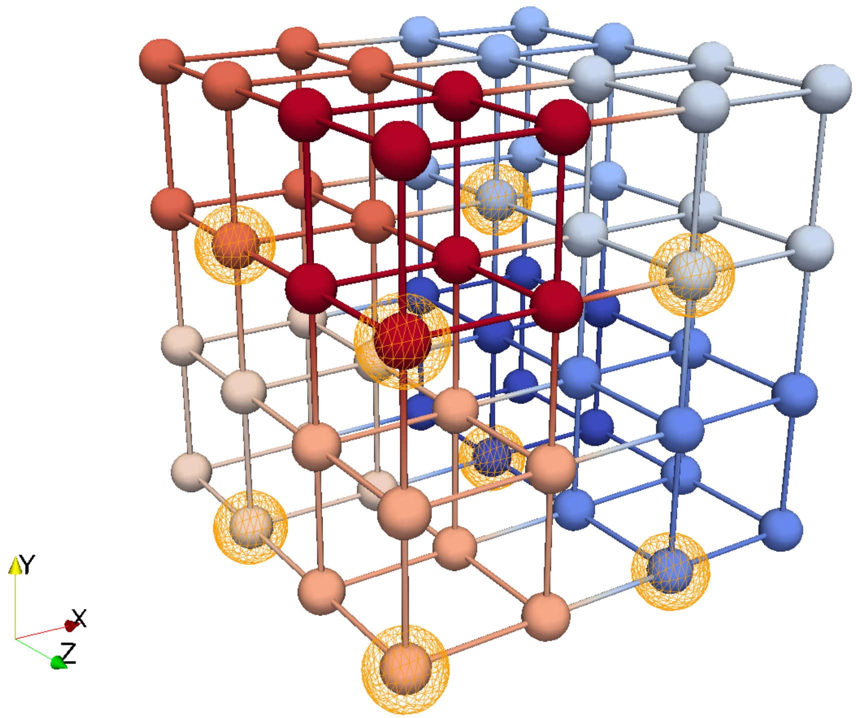


MPI collective I/O has 3 phases:

- Exchange of offsets and sizes using MPI_Alltoallv over the collective network
- Exchange of data to the aggregators
- Write the data out over the collective network

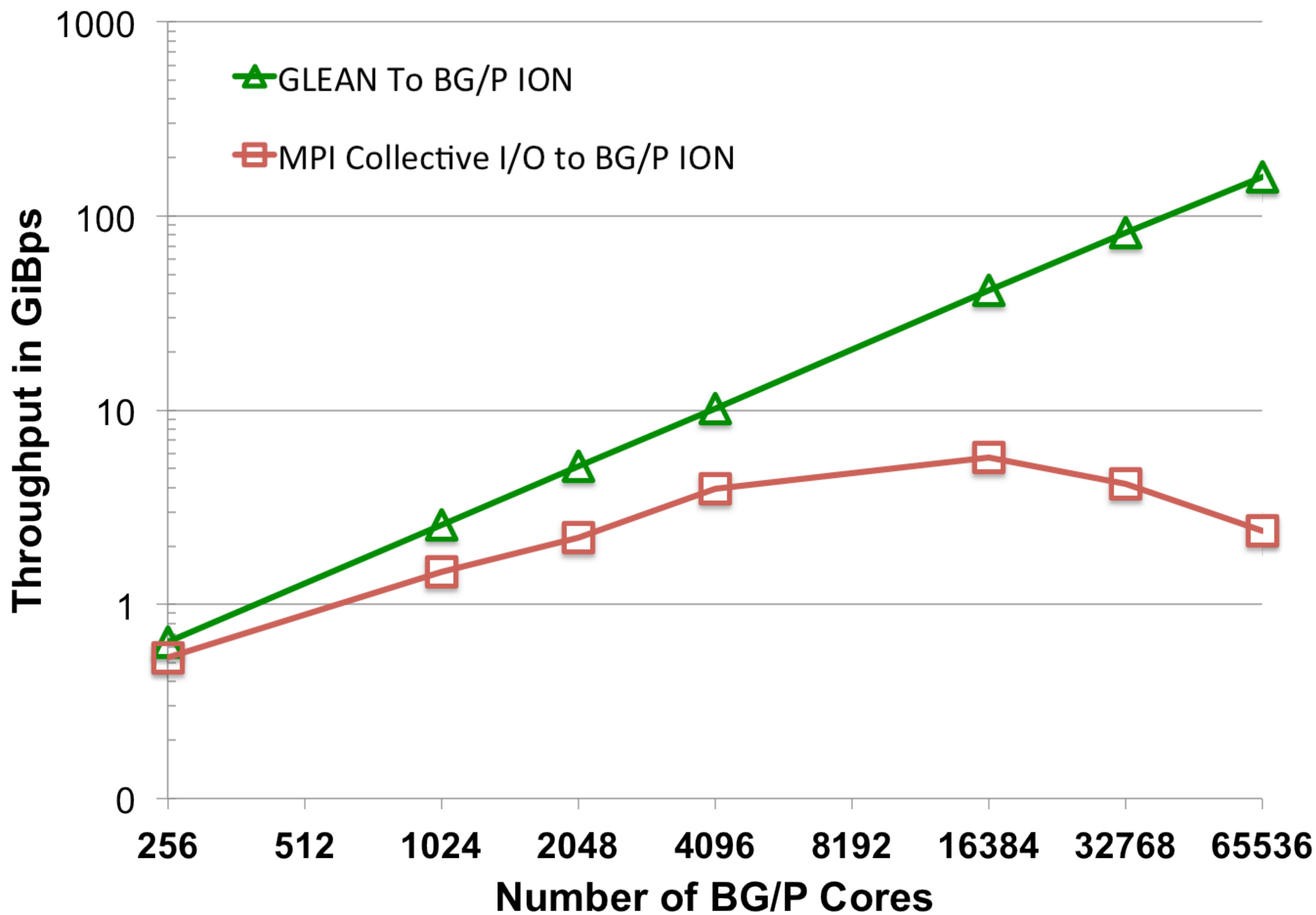
Designated aggregator node could be in a different pset - several hops away

Exploiting Topology for I/O Acceleration



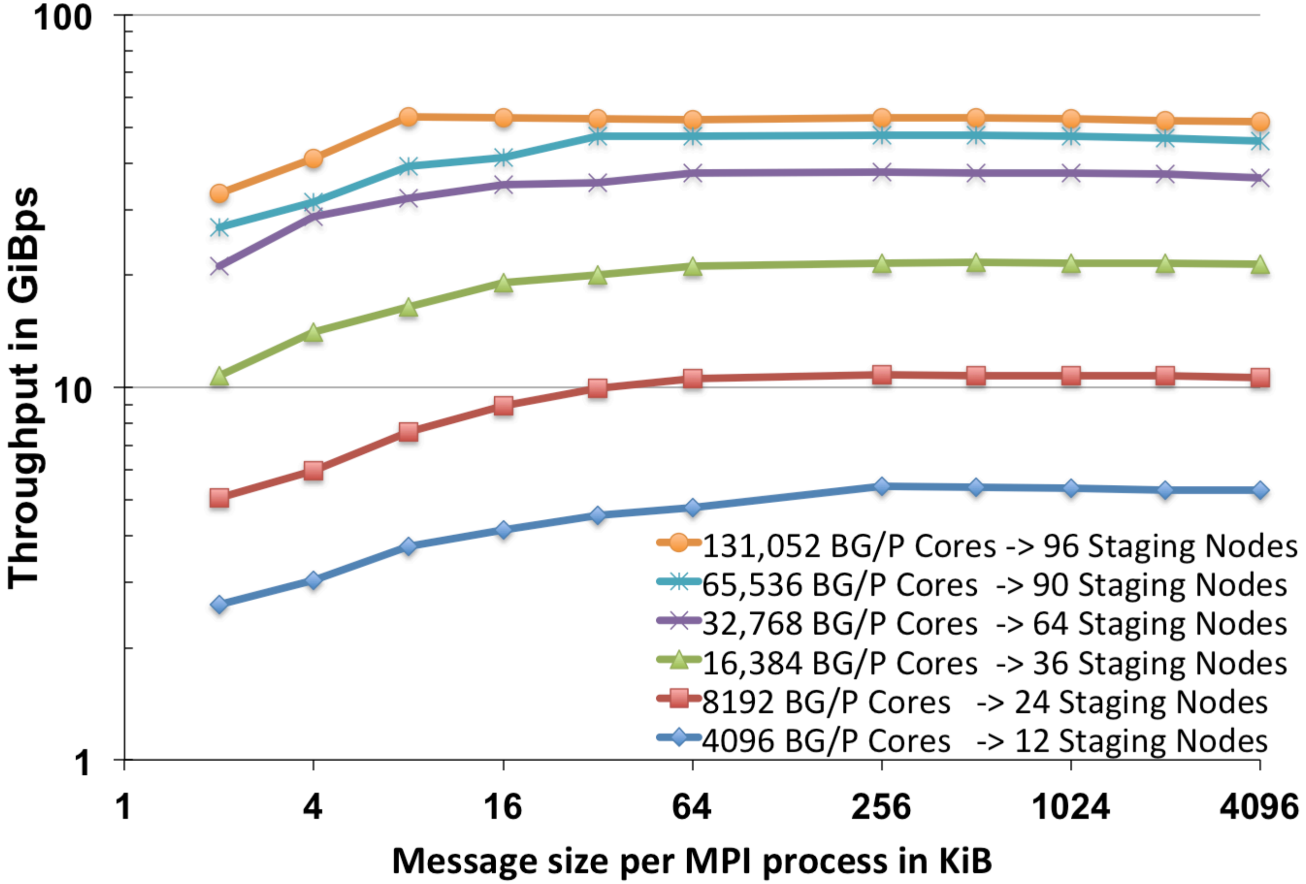
- Aggregator groups formed by exploiting the BG/P *personality* information
- Restrict aggregation traffic to a pset
- Exploit both 3D torus and tree network for data movement
- Dynamic # of aggregators based on message size

Strong scaling performance to write 1GiB



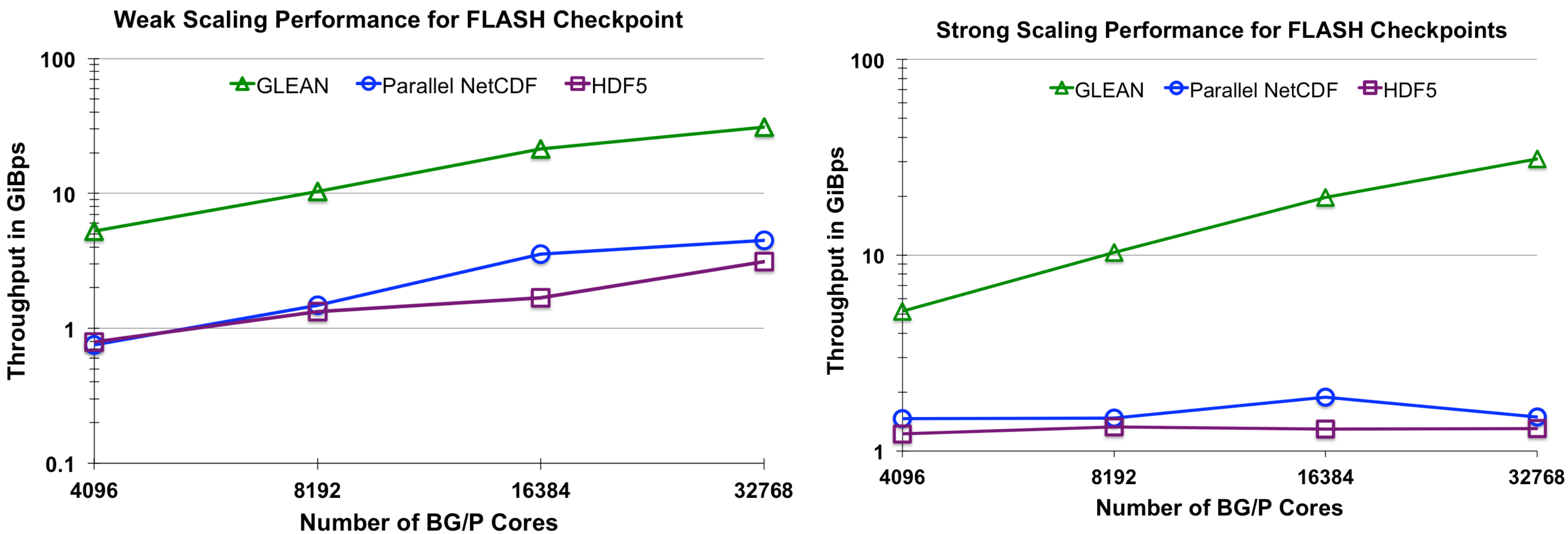
Strong scaling is critical as we move towards future systems with lower memory per core (SC'11)

End-to-end data movement performance scaling to 131,052 Intrepid cores (32 racks)



GLEAN sustains **54 GiBps** of aggregate throughput at 131,052 cores (80% of the entire system) with 96 Eureka nodes

Performance for FLASH checkpoints



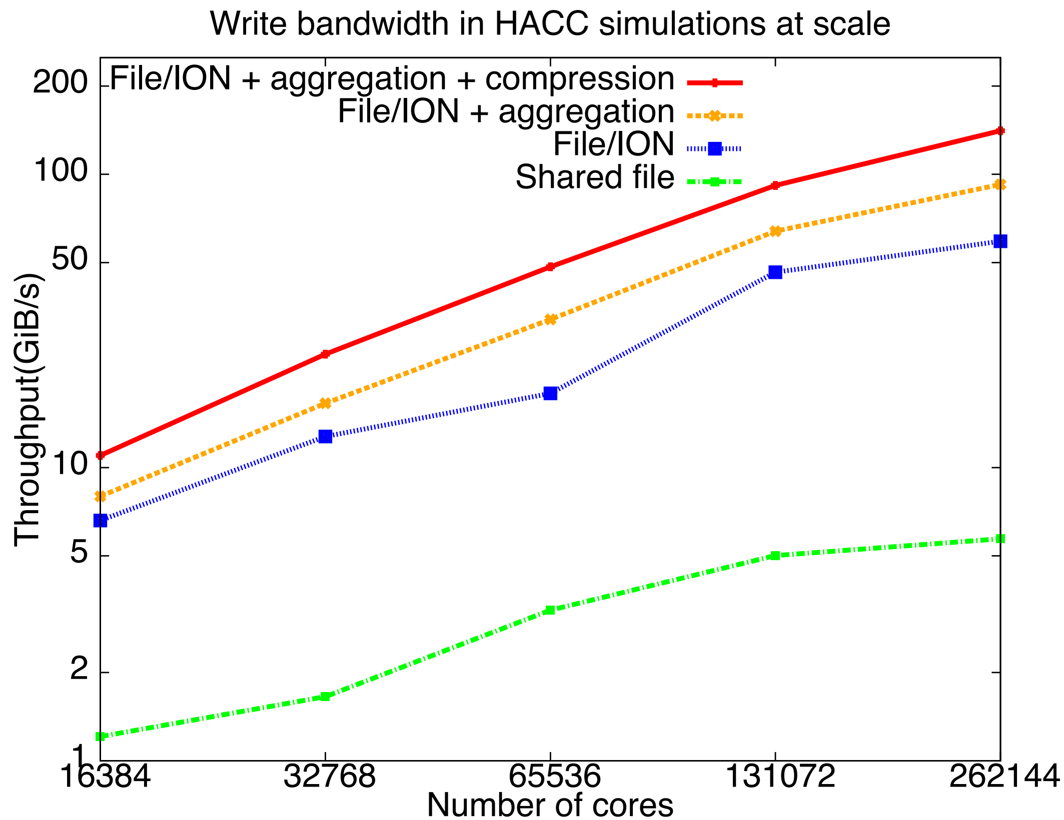
- For weak scaling at 32,768 cores, GLEAN sustains 31 GBps and achieves an observed speedup of **10-fold** over pnetcdf and hdf5
- For strong scaling at 32,768 cores, GLEAN sustains 27 GBps and achieves an observed speedup of **15-fold** over pnetcdf and hdf5
- 16.3 GBps to Storage at 32K cores.

Scalable I/O at 768K cores with GLEAN

- Joint work with HACC team
- Scaled to the entire **768K** cores of Mira BG/Q system
- Integrated with HACC Cosmology production simulation runs and enabled the Gordon Bell runs
- Used in production on BG/Q (Mira) and Cray (Hopper)
- Achieved **~180 GB/s** for HACC I/O and up to ~16X improvement over the previous I/O mechanism on Mira
- Written and read **~20 PB** of data on Mira (and counting)
- Used for all HACC inputs and outputs of production runs including particle, cosmo, and halo data
- Parallel lossless data compression with custom pre-conditioner, and parallel checksums (fletcher64 and crc64)

SC 2013 Gordon Bell Finalist

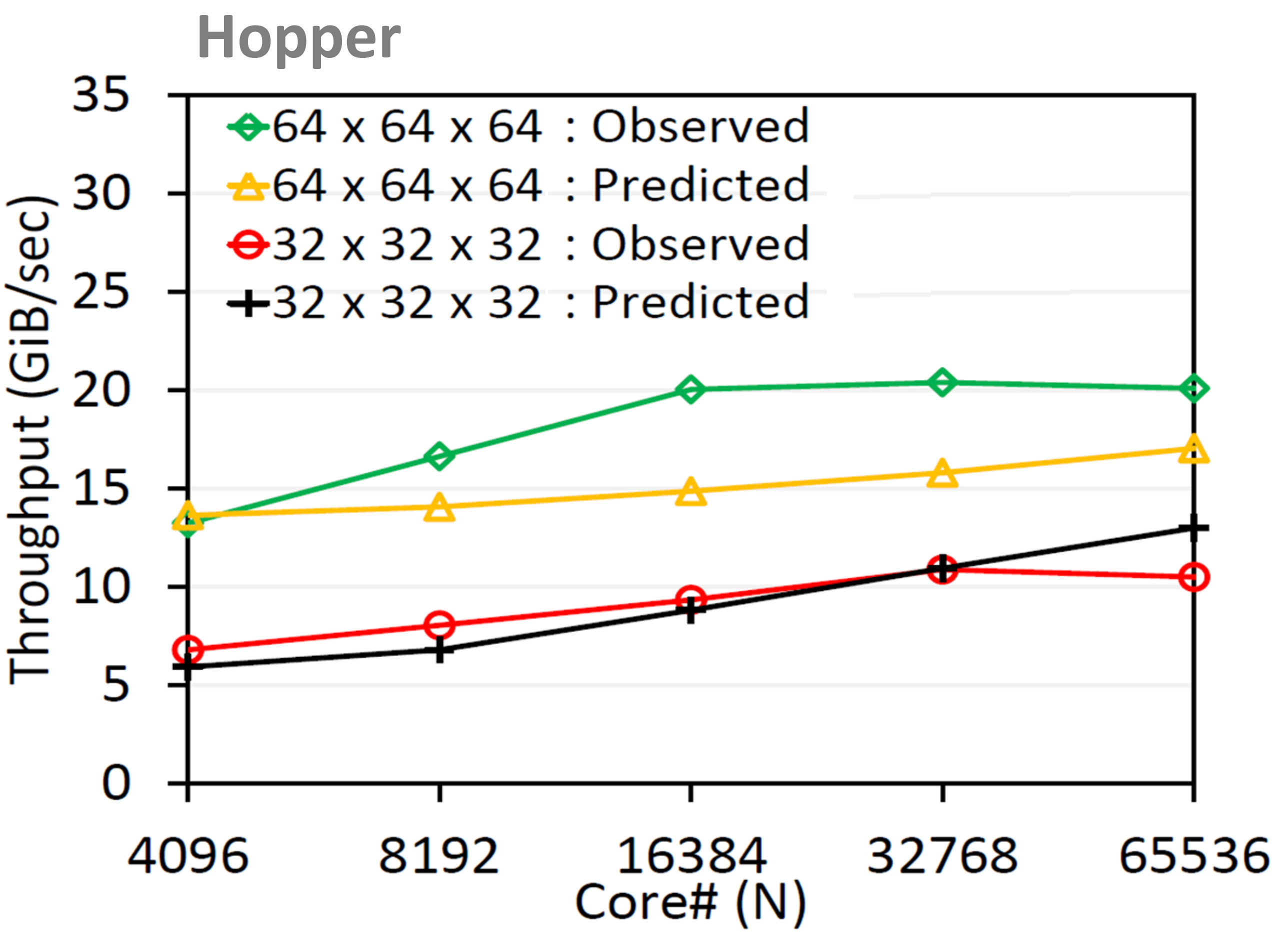
Scalable I/O using GLEAN for HACC simulations



- Multi-fold improvement for writing out HACC analysis outputs
- Subfiling, topology-aware data movement, and compression are key for I/O performance at scale (PDP'2014)

H. Bui, V. Vishwanath, H. Finkel, K. Harms, J. Leigh, S. Habib, K. Heitmann, M. E. Papka. "Scalable parallel I/O on the Blue Gene/Q supercomputer using compression, topology-aware data aggregation, and subfiling," In the Proceedings of the 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2014), Turin, Italy, February, 2014.

Using Regression Models for I/O Tuning (SC'13)



I/O is a challenging problem with several parameters needed to be tuned for performance. Our approach helps identify these to mitigate I/O bottlenecks.

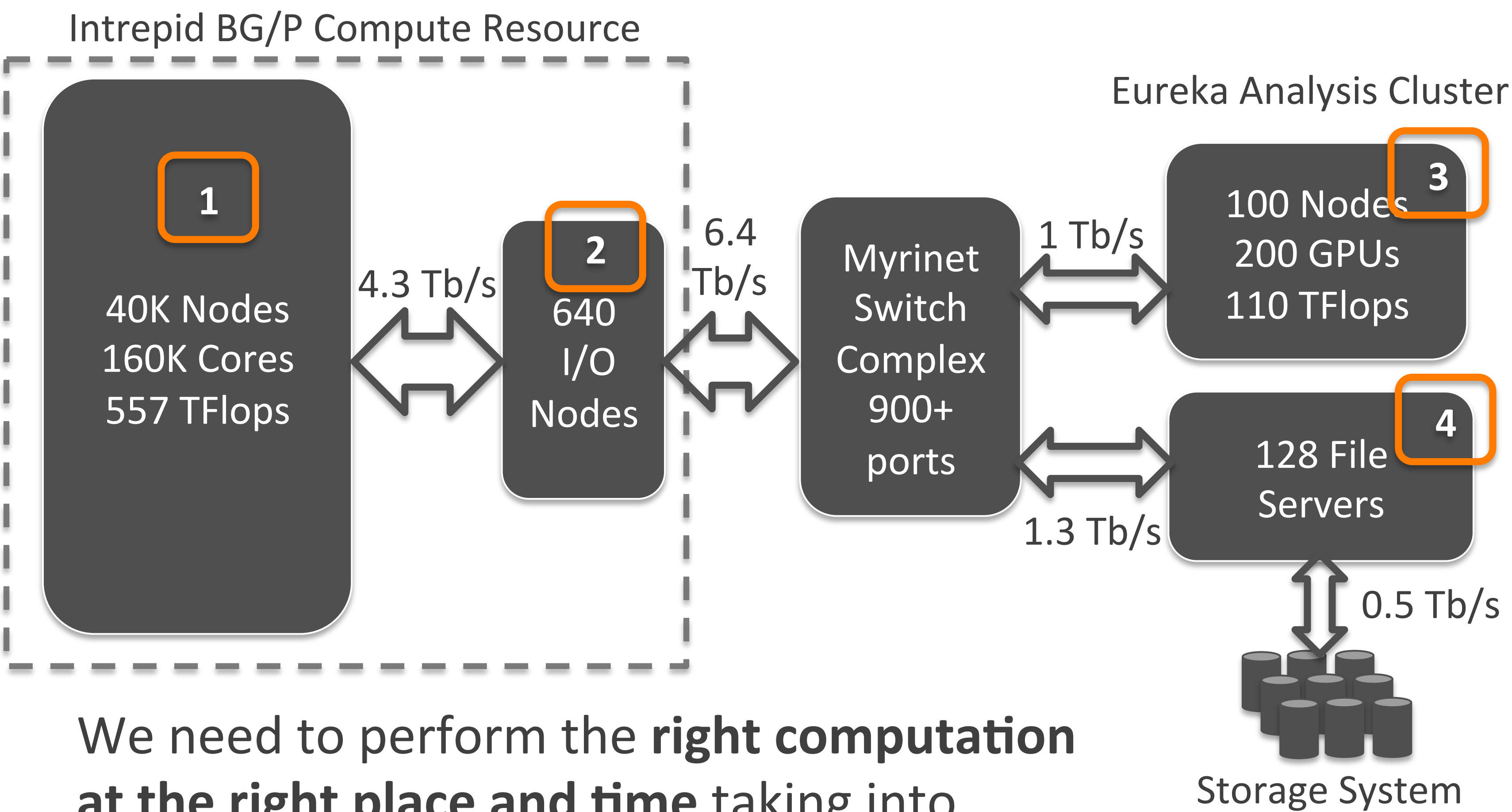
We Compare Performance of Different Models On Intrepid and Choose the Best

Model	Error (in %)	Model	Error (in %)
Linear Reg	19.6	SVM Reg (Lin)	21.2
Ridge Reg	20.2	**Decision Trees**	**9**
Lasso	18.9	SVM Reg (Poly)	16
Lars	20.34	Gaussian Processes	13
Elastic Net	21.68	**Random Forests**	**8.2**
SGD	16.7	**GBDT**	**8.1**

Tree based Models exhibits least error

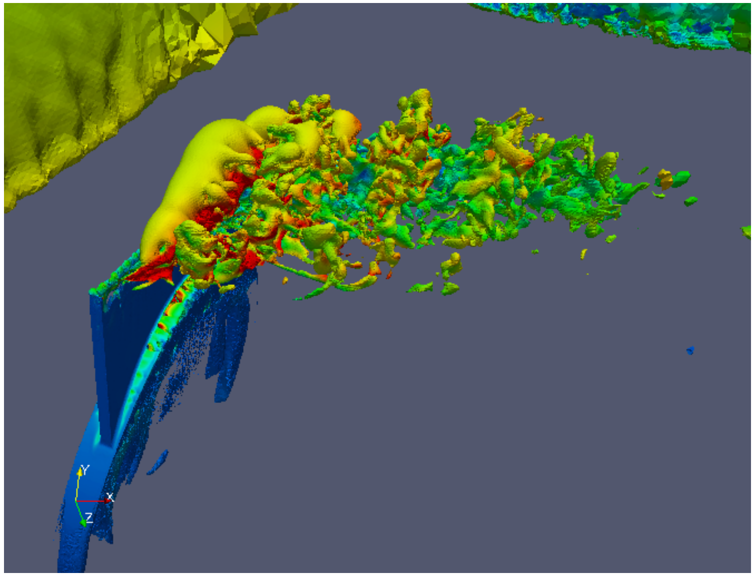
- Tree-based models are simple and intuitive to understand
- The decision at each step (node of the tree) is based on a single parameter of the dataset, which involves a quick look-up operation along the depth of the tree.
- Other models solve a complex optimization problem, making it difficult to judge the relative usefulness of specific dataset attributes.

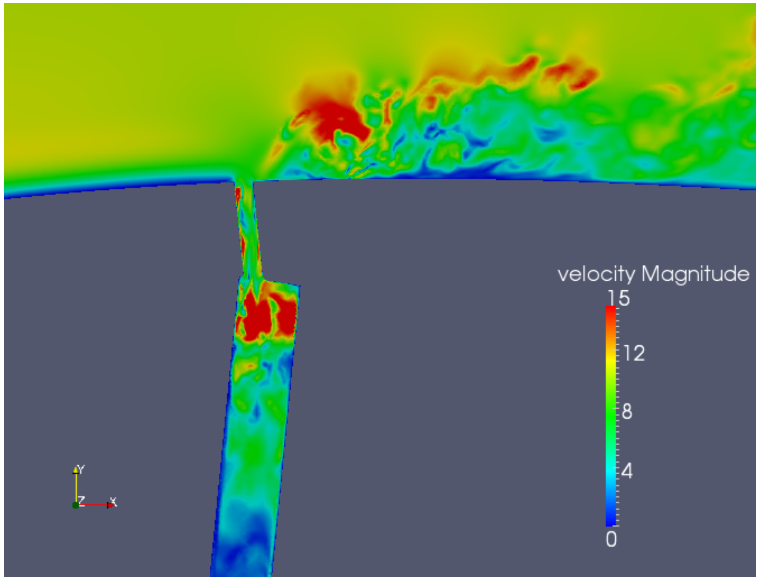
Simulation-time Analysis Opportunities on the Argonne Leadership Computing Facility



We need to perform the **right computation at the right place and time** taking into account the characteristics of the simulation, resources and analysis

Simulation-time analysis of PHASTA on 160K Intrepid BG/P cores

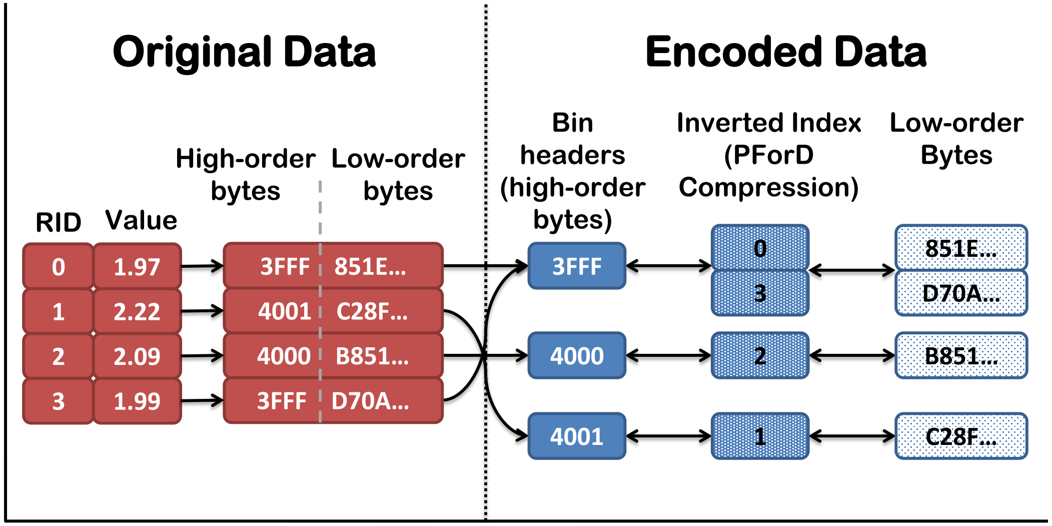




Isosurface of vertical velocity colored by velocity and cut plane through the synthetic jet (both on 3.3 Billion element mesh). *Image Courtesy: Ken Jansen*

- Visualization of a PHASTA simulation running on **160K cores** of Intrepid using ParaView on 100 Eureka nodes **enabled by GLEAN**
- GLEAN achieves **48 GiBps** sustained throughput for data movement enabling simulation-time analysis

Database Indexing to Accelerate Queries in HPC



- Indexing is commonly used to in databases accelerate search queries.
- In Data-centric HPC, with indices, a scientist can interactively explore the dataset. The challenge is in dealing with index generation and index sizes

- Data Indexing and Reorganizing for Analytics-induced Query processing
- Scaled to BG/P and Cray XE-6 system
- Demonstrated with FLASH and S3D via GLEAN

(Best paper award at HPDC'13)

Other Relevant Threads

- SKOPE - Language for performance modeling
- Heterogeneous multi-site workflow scheduling
- Modeling end-to-end parallel storage transfers
- HPDF Project – Programmable parallel network and storage infrastructure for improved performance
- ExaHDF5 Project & Concerted Flows Project
- I/O Optimization on Cray systems
- Scheduling
- Scalable Visualization and Analytics

Relevant Papers

- 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", In Proceedings of the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), Seattle, USA, November 2011.
- V. Vishwanath, M. Hereld, K. Iskra, D. Kimpe, V. Morozov, M. Papka, R. Ross, and K. Yoshii, "Accelerating I/O Forwarding in IBM Blue Gene/P Systems", In Proceedings of the IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2010), pp. 1--10, November 2010.
- V. Vishwanath, M. Hereld, and M. E. Papka, "Simulation-time data analysis and I/O acceleration on leadership-class systems using GLEAN", In Proceedings of the IEEE Symposium on Large Data Analysis and Visualization (LDAV), Providence, RI, USA, October 2011.
- M. Rasquin, P. Marion, V. Vishwanath, B. Matthews, M. Hereld, K. Jansen, R. Loy, A. Bauer, M. Zhou, O. Sahni, J. Fu, N. Liu, C. Carothers, M. Shephard, M. E. Papka, K. Kumaran, B. Geveci,"Co-visualization of full data and in situ data extracts from unstructured grid CFD at 160K cores", In Proceedings of the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), Seattle, USA, November 2011.
- S. Lakshminarasimhan, D. A. Boyuka II, S. V. Pendse, X. Zou, J. Jenkins, **V. Vishwanath**, M. E. Papka, N. F. Samatova, "*Scalable In Situ Scientific Data Encoding for Analytical Query Processing",* In the proceedings of the 22nd International ACM Symposium on High Performance Parallel and Distributed Computing **(HPDC 2013)**, New York City, New York, June 2013. ***[Best Paper Award]***
- E. Schendel, S. Harenberg, H. Tang, **V. Vishwanath**, M.E. Papka and N. Samatova, "*A Generic High-performance Method for Deinterleaving Scientific Data",* In the 19th International European Conference on Parallel and Distributed Computing (**EuroPar**), Aachen, Germany, August 2013.
- S. Habib, V. Morozov, N. Frontiere, H. Finkel, A. Pope, K. Heitmann, K. Kumaran, V. Vishwanath, T. Peterka, J. Insley, D. Daniel, P. Fasel, Z. Lukic, "HACC: Extreme Scaling and Performance Across Diverse Architectures", In the Proceedings of the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2013), Denver, Colorado, USA, November 2013 (Gordon Bell Finalist).
- H. Bui, V. Vishwanath, H. Finkel, K. Harms, J. Leigh, S. Habib, K. Heitmann, M. E. Papka. "Scalable parallel I/O on the Blue Gene/Q supercomputer using compression, topology-aware data aggregation, and subfiling," In the Proceedings of the 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2014), Turin, Italy, February, 2014.

Thank You!!