Addressing data challenges in scientific computing at extreme scale

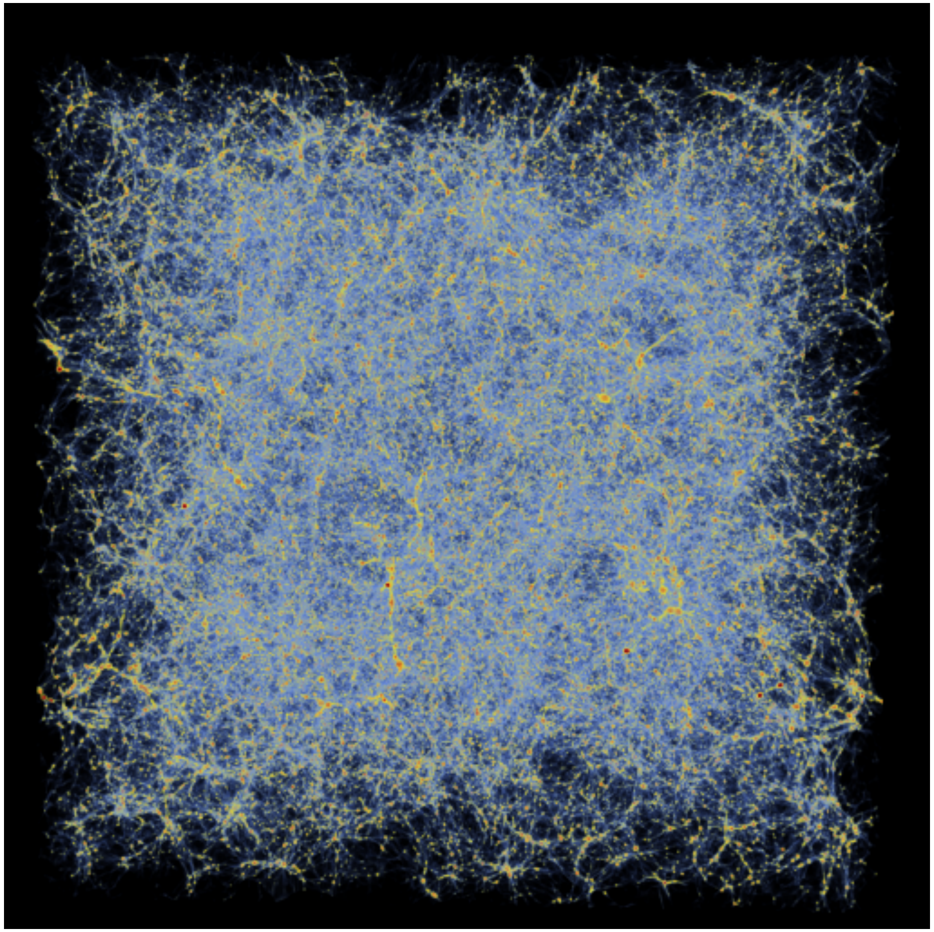
Venkatram Vishwanath

Argonne National Laboratory

venkatv@mcs.anl.gov

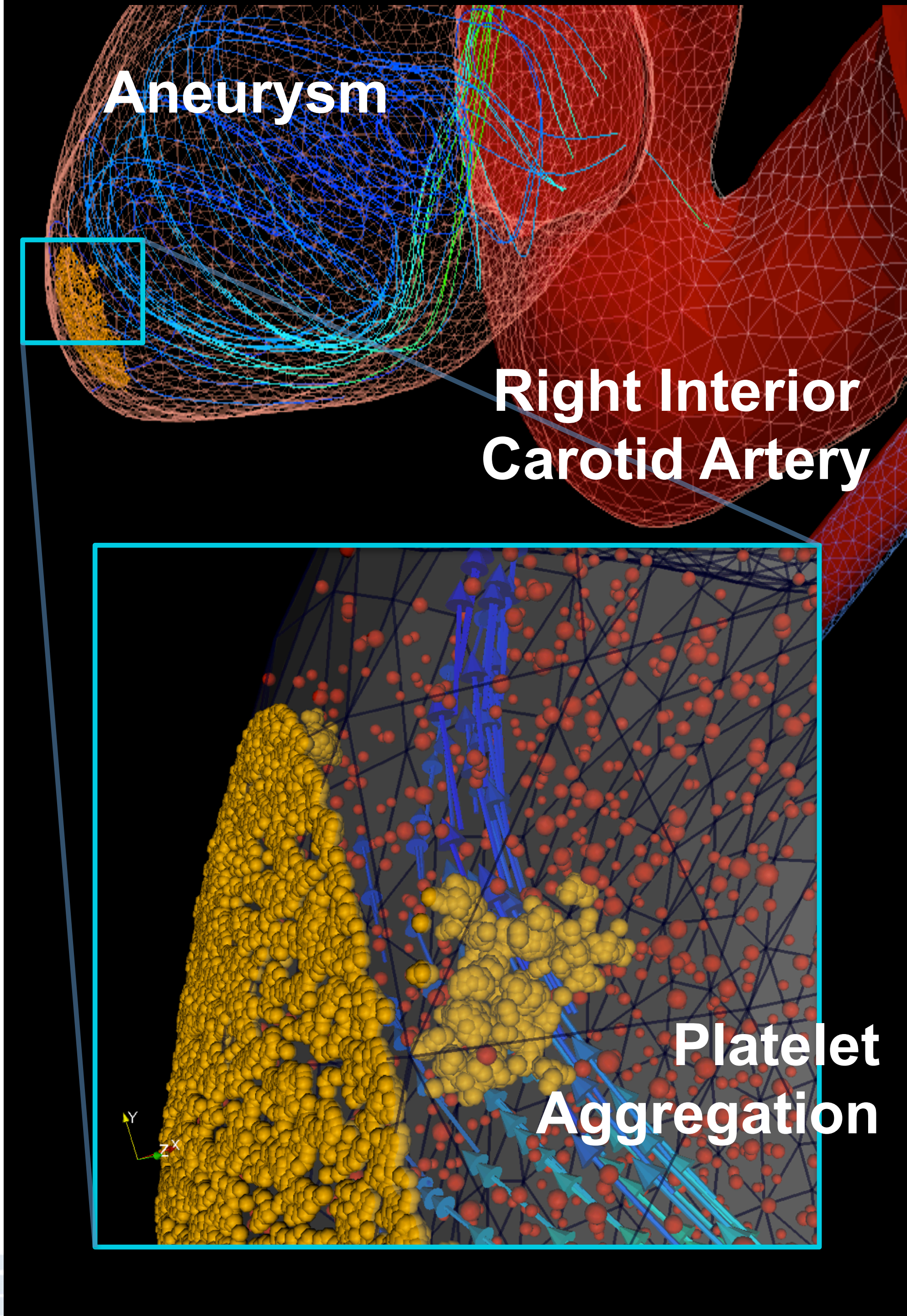
Data scale and requirements

- ENZO BAO Simulation
  - Volume 1 billion light years on a side
  - Shortly after Big Bang to half the present age of the universe
  - 64 billion cells
  - 4096^3 grid resolution
  - 148 Terabytes of data (~570 steps, 1 variable, 256GB/step)
  - 4 Million CPU hours



Dataset Complexity

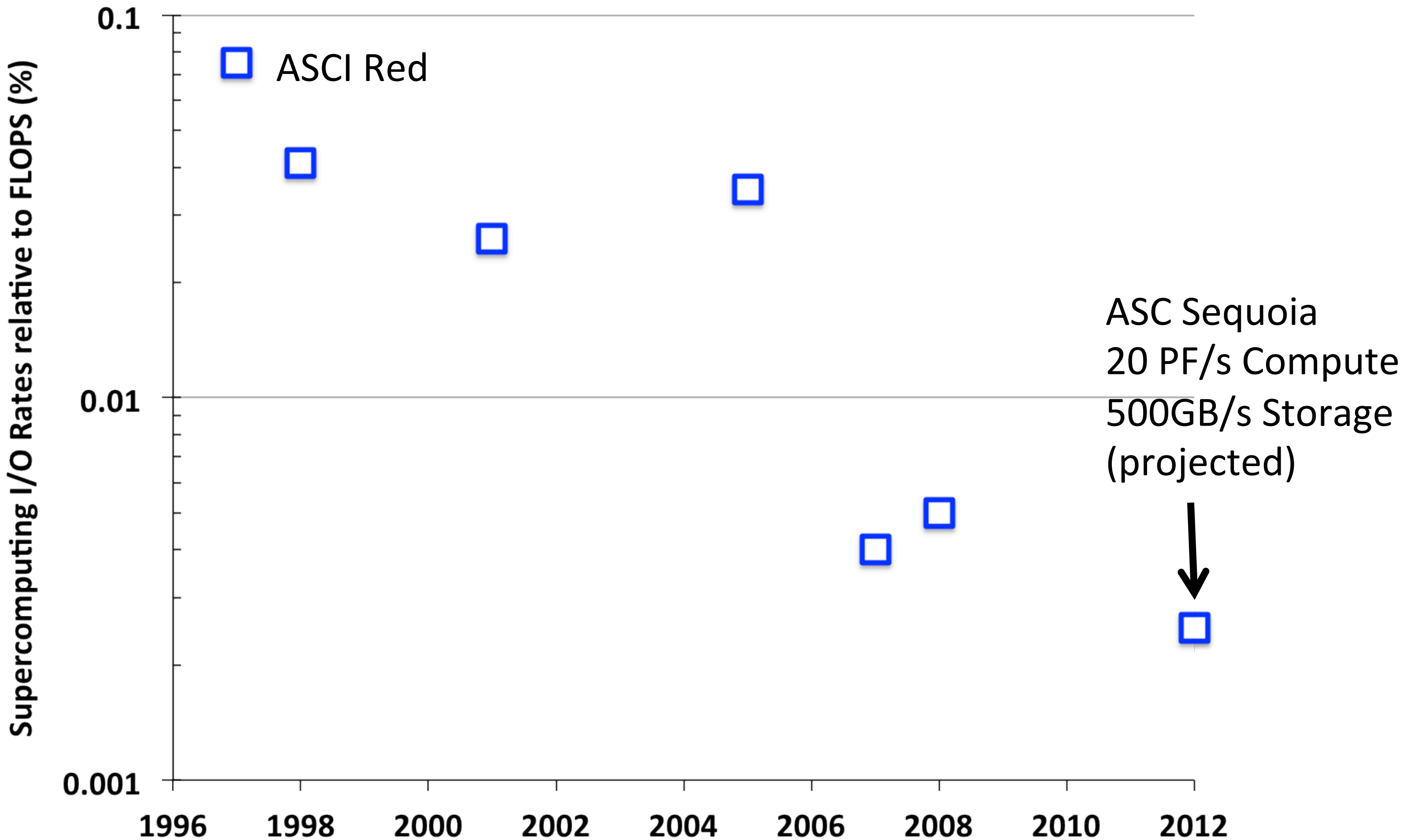
- Complexity as an artifact of science problems and codes:
  - Coupled multi-scale simulations generate multi-component dataset.
  - Atomistic data representations for plasma, red blood cells, and platelets from MD simulation.
  - Field data for ensemble average solution generated by spectral element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



Scale and Complexity of Systems

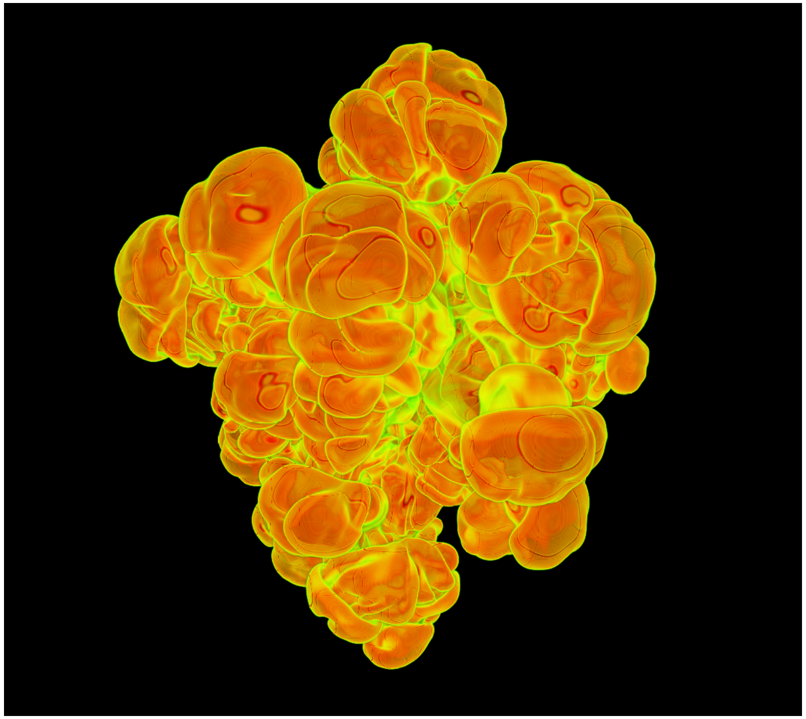
System	Blue Gene/Q	K Computer	Tianhe-1A	
Peak Perf # of Racks	20 PF 96	11.3 PF 864	4.7 PF 112	
# of cores	1,572,864	705,024	202,752	
Processor	PowerPC	SPARC 64	Xeon X5670	NVIDIA M2050
Mem per core (Flops/byte)	1 GB 4.9	8 GB 1	1 GB 0.75	0.21 GB 3
Interconnect	5D Torus	6D Torus	Fat Tree	
Power	6 MW	12.7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

Storage vs Computation Trends



Storage Systems have not kept up with the computing trends, and the gap appears to be widening

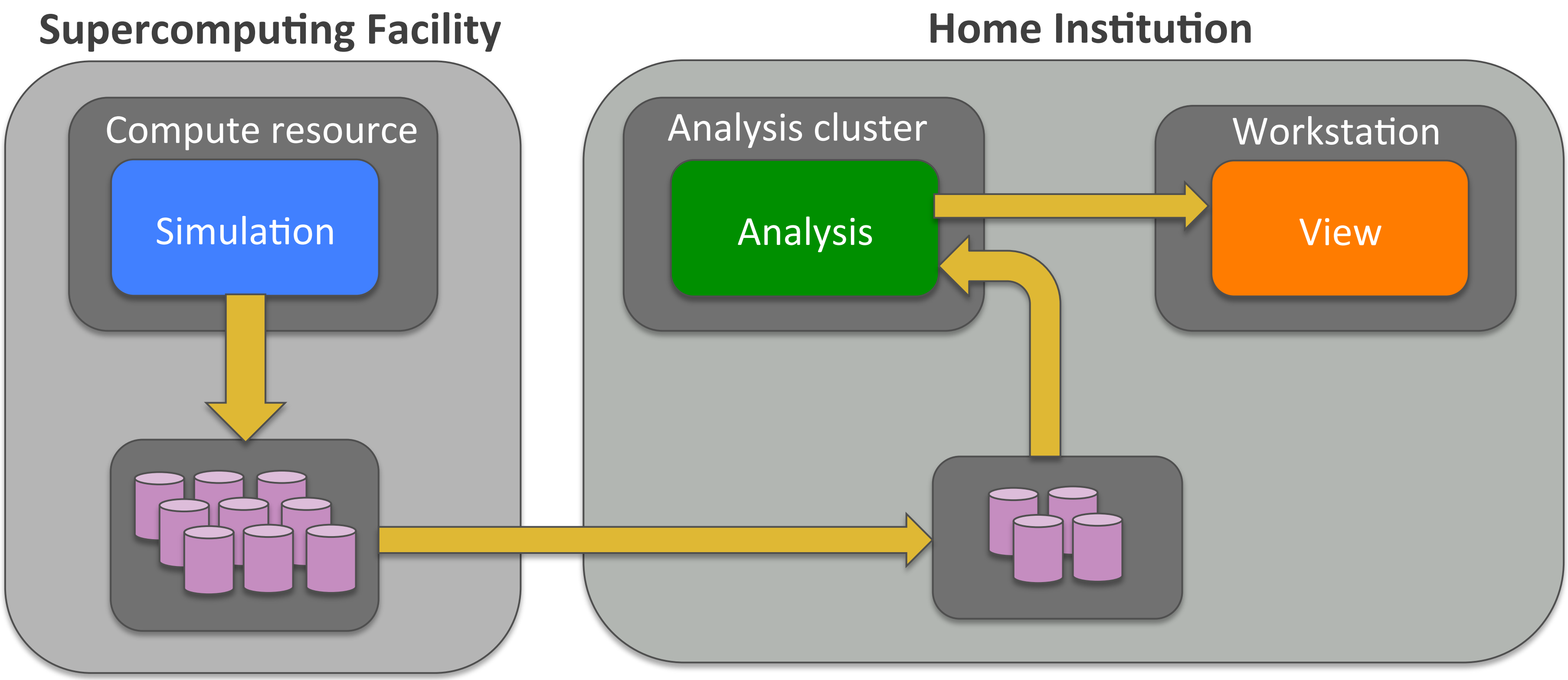
FLASH Astrophysics I/O performance



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

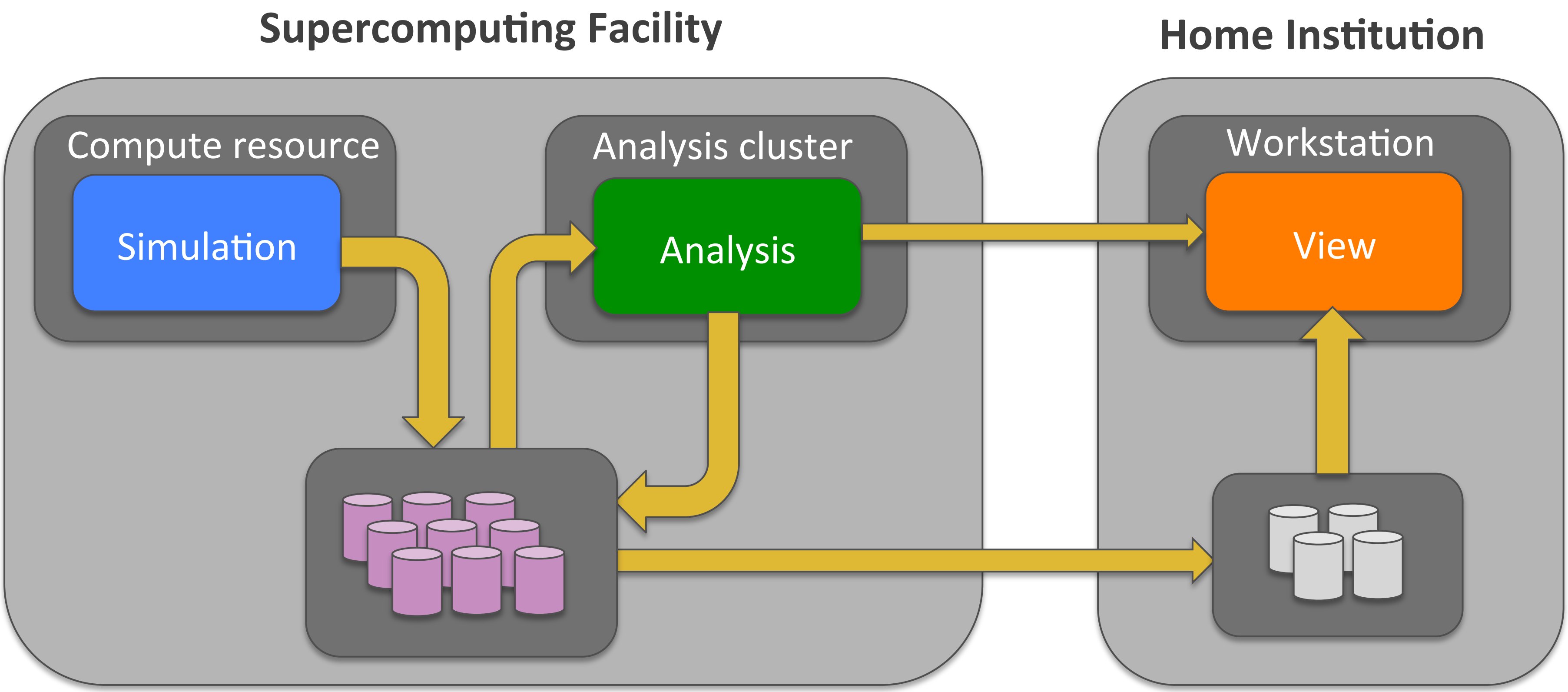
During large-scale capability runs, up to 30% of time spent in I/O

Traditional Science Pipeline



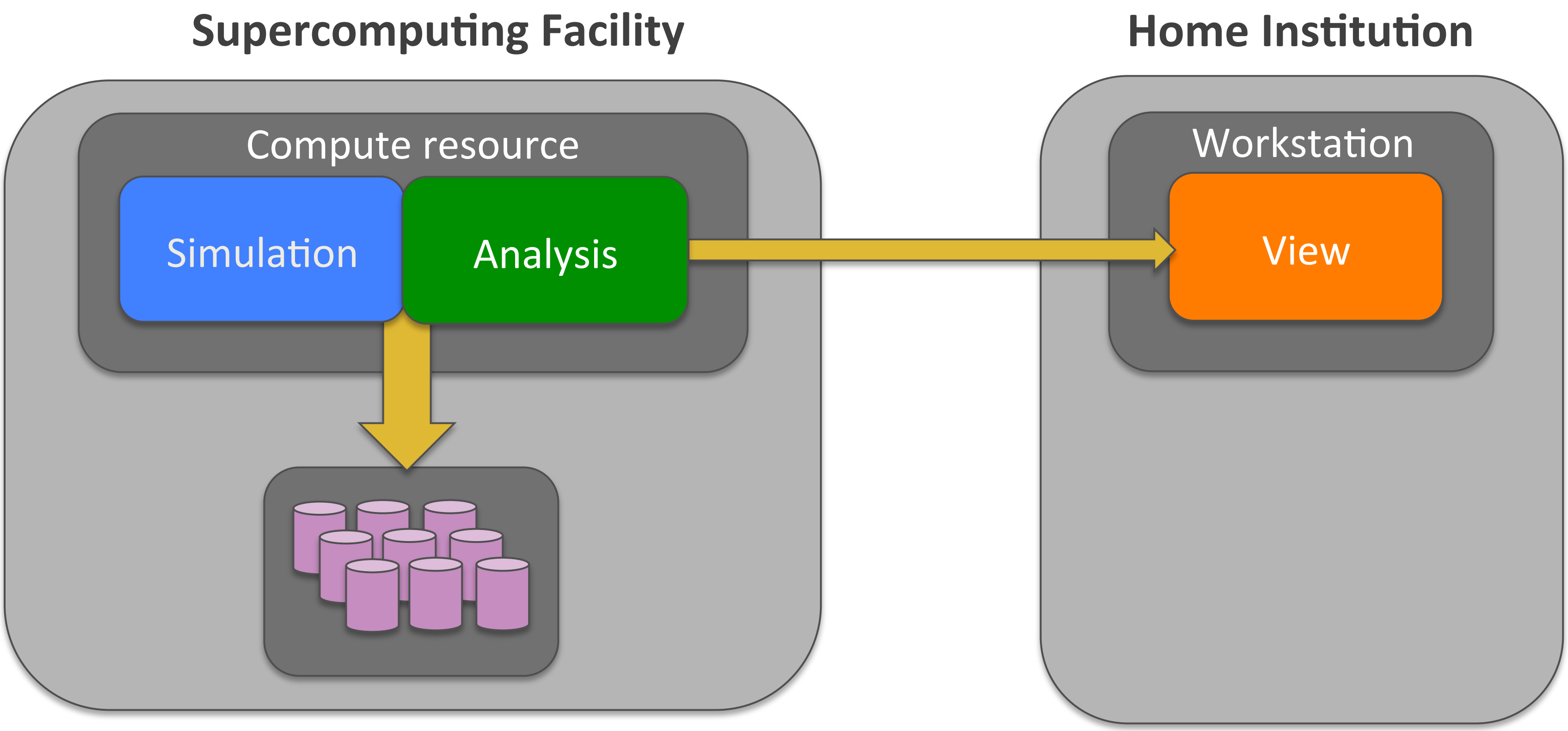
Time to discovery is high as we are moving data to and from storage

Post Processing Pipeline in HPC



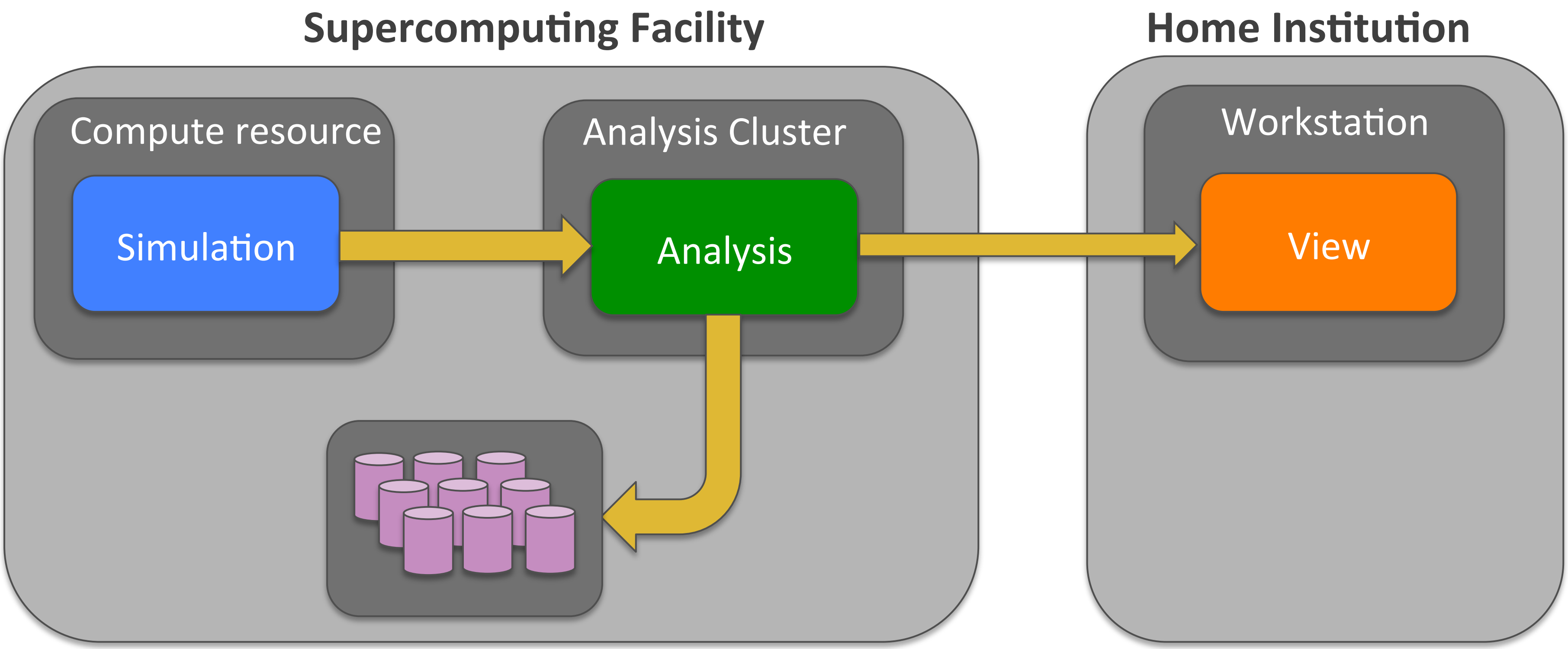
Storage systems are currently unable to cope with extreme scale data sizes in a cost-effective way and this will only get worse in future

in situ - Simulation Time Analysis on the Compute Resource



Analysis occurs during simulation time on the compute resource

co-analysis - Simulation time analysis on a direct attached analysis resource



- Compute resource and Analysis resource **are directly connected over an ultra high-speed network**
- **Data is moved to** the analysis resource **memory**

in situ versus co-analysis

in situ

Pros

- Uses simulation data structures
- No additional hardware resource required

Cons

- Time-varying and memory-intensive analysis is extremely difficult

co-analysis

Pros

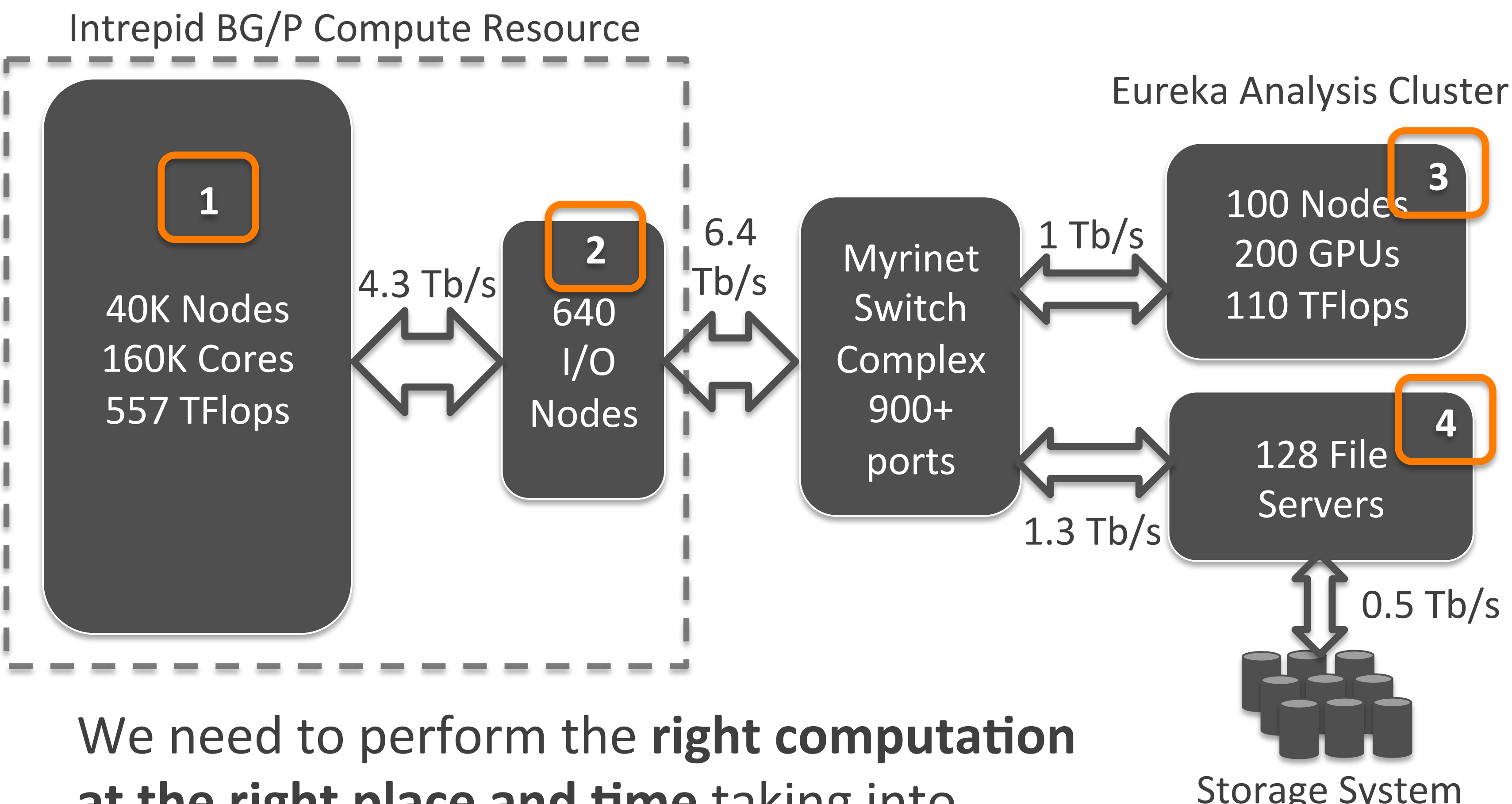
- Extremely flexible analysis including time-varying analytics
- Does not require precious simulation resources

Cons

- Requires a co-scheduling infrastructure

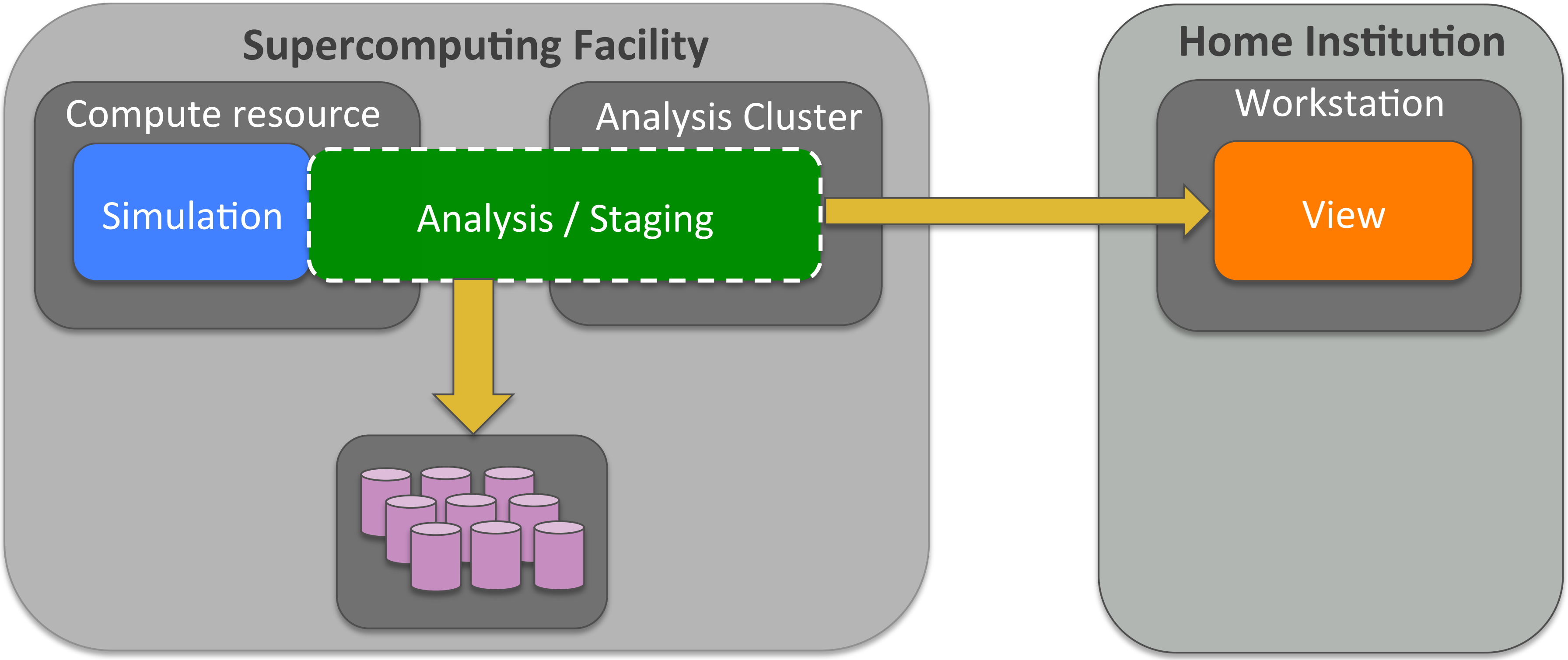
Proposed solutions require modification to the simulations code and a flexible approach is needed

Simulation-time Analysis Opportunities on the Argonne Leadership Computing Facility

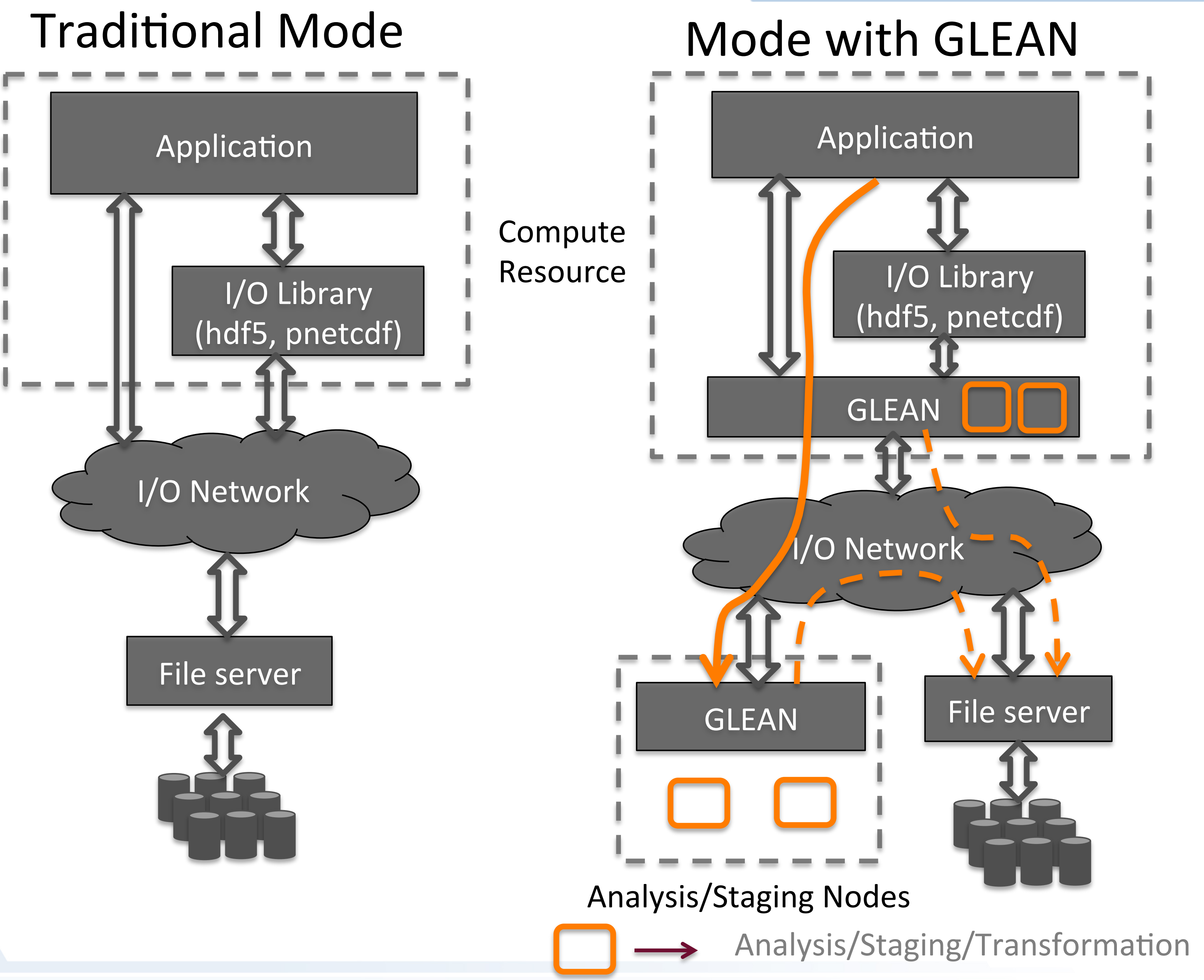


We need to perform the **right computation at the right place and time** taking into account the characteristics of the simulation, resources and analysis

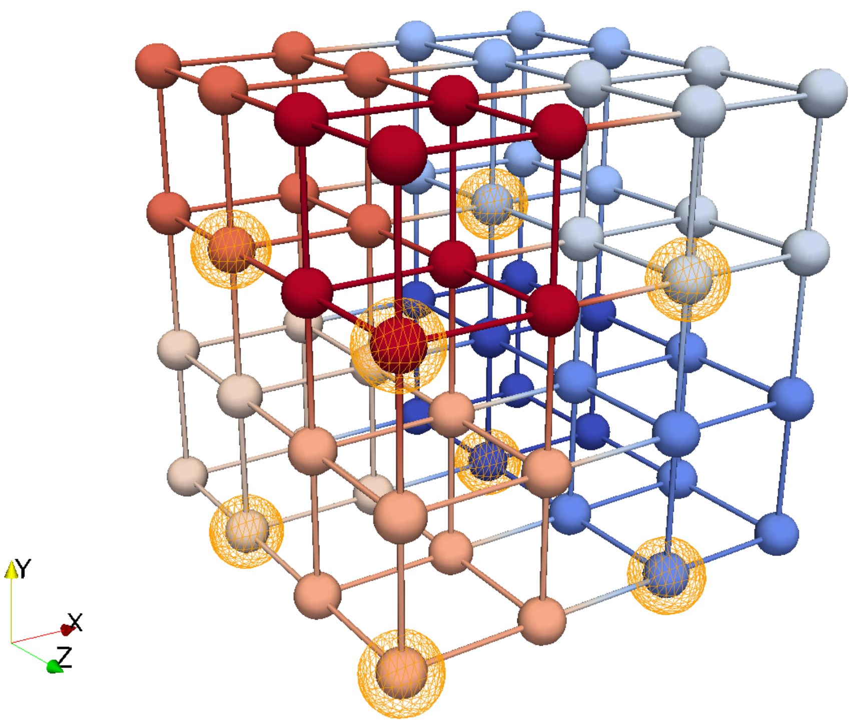
Our approach - GLEAN



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**.

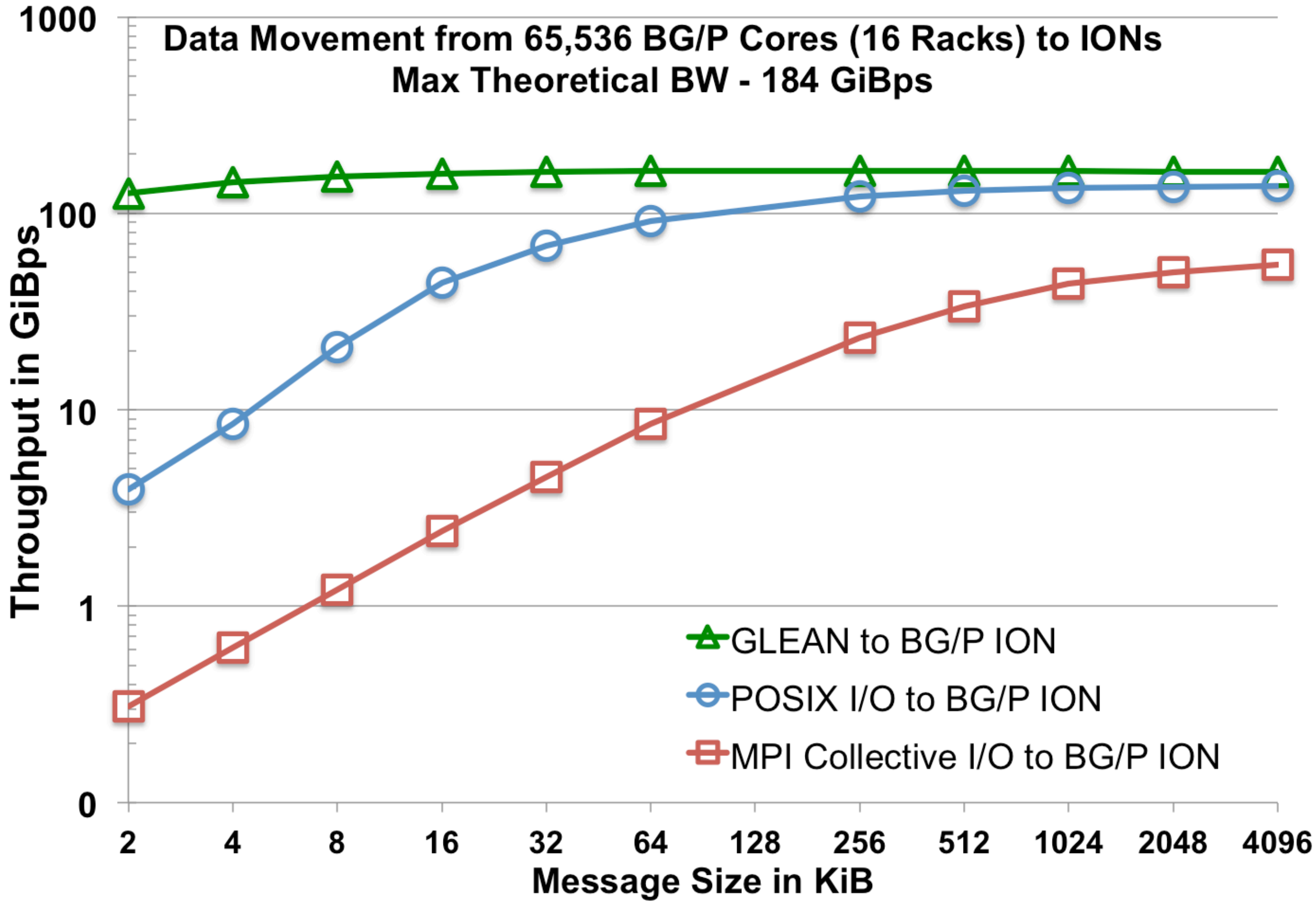

Traditional Mode
Application
I/O Library
(hdf5, pnetcdf)
I/O Network
File server
Compute
Resource
Mode with GLEAN
Application
I/O Library
(hdf5, pnetcdf)
GLEAN
I/O Network
GLEAN
File server
Analysis/Staging Nodes
Analysis/Staging/Transformation

Exploiting the BG/P Topology for I/O



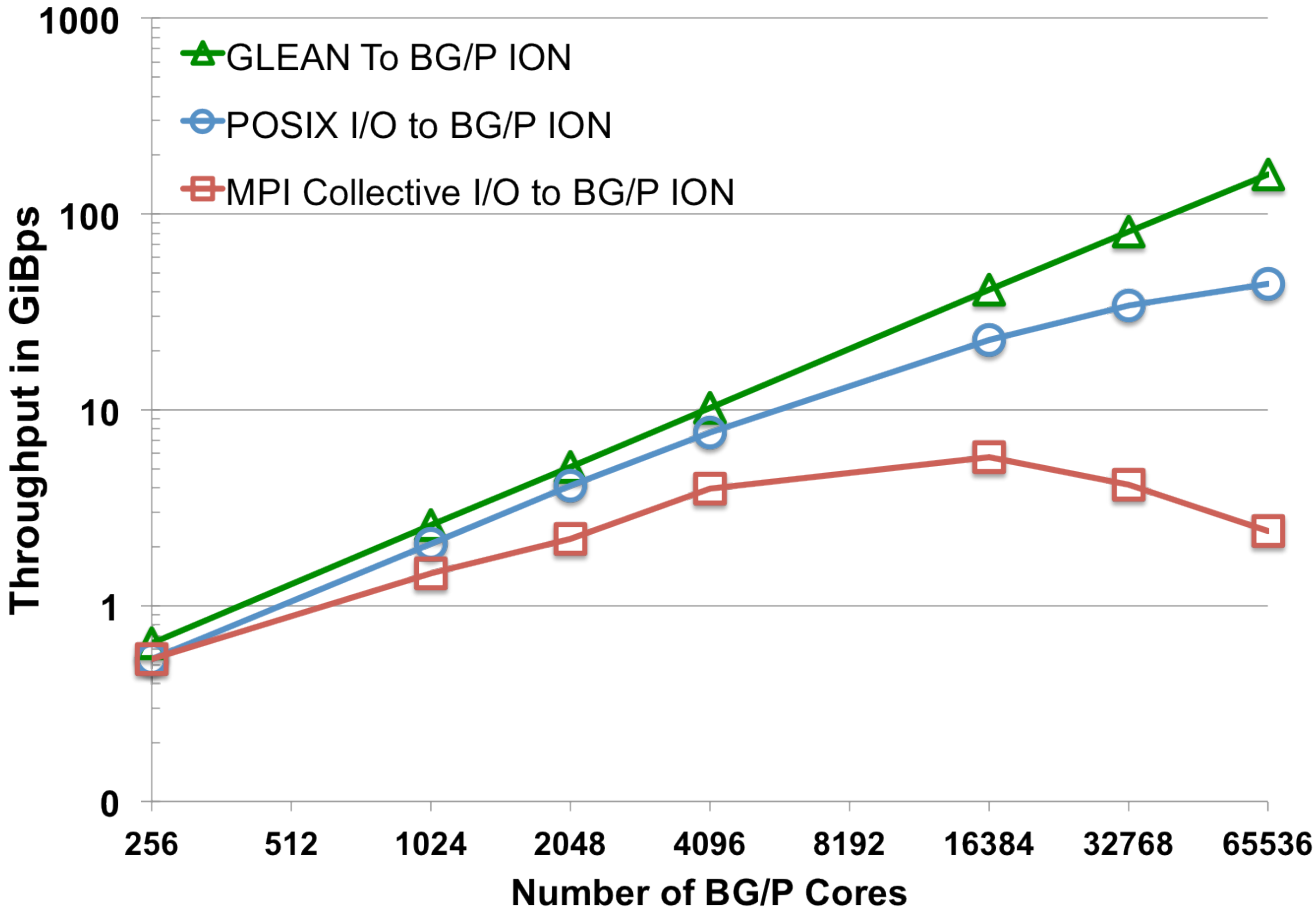
- Aggregator groups formed by exploiting the BG/P *personality* information
- Restrict aggregation traffic to a pset
- Exploit both 3D torus and tree network for data movement
- Dynamic # of aggregators based on message size

Data movement from 16 BG/P racks (64K cores)



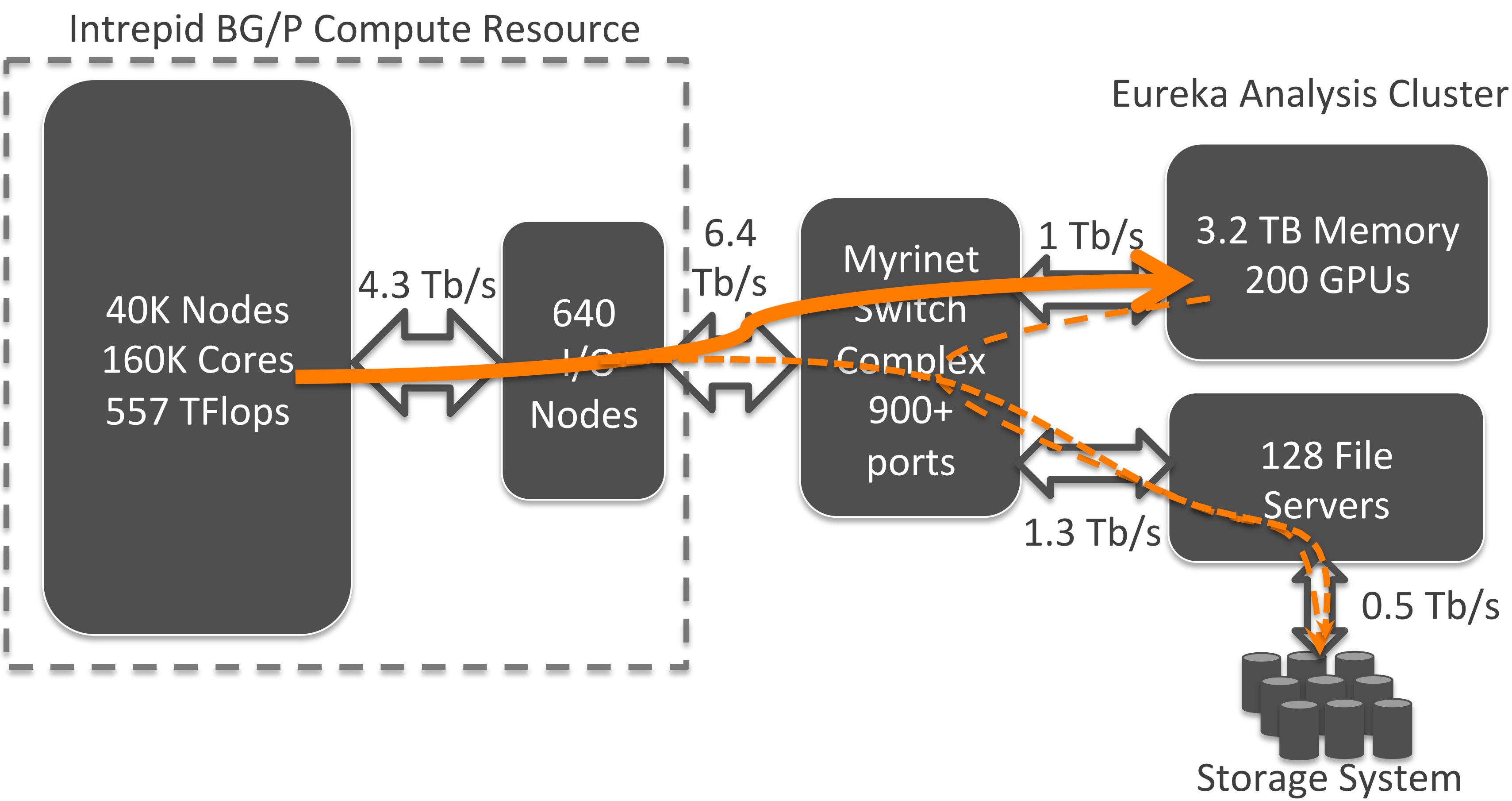
GLEAN demonstrates scalable performance for both small messages and large messages, and achieves up to **90% of the peak aggregate** throughput out of BG/P

Strong scaling performance to write 1GiB



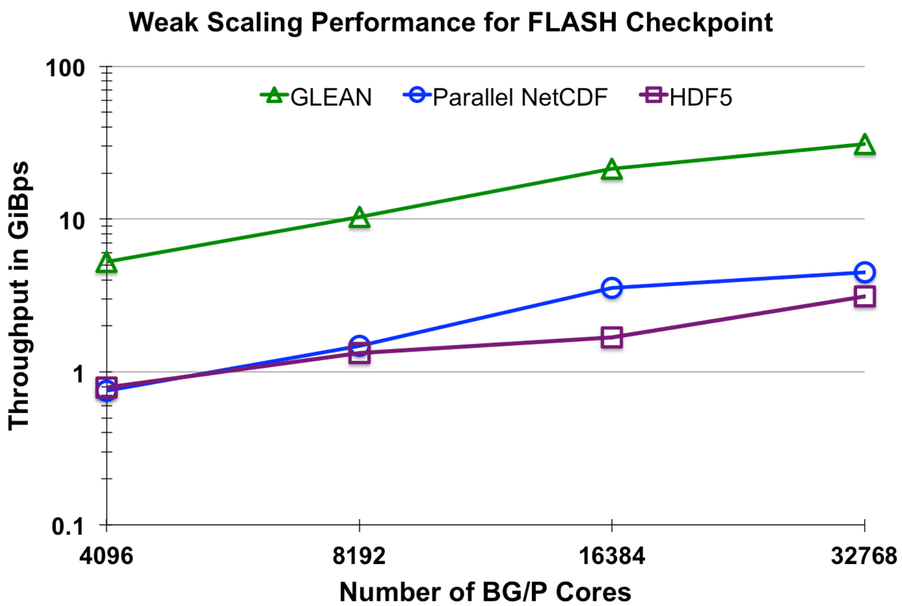
Strong scaling is critical as we move towards future systems with lower memory per core

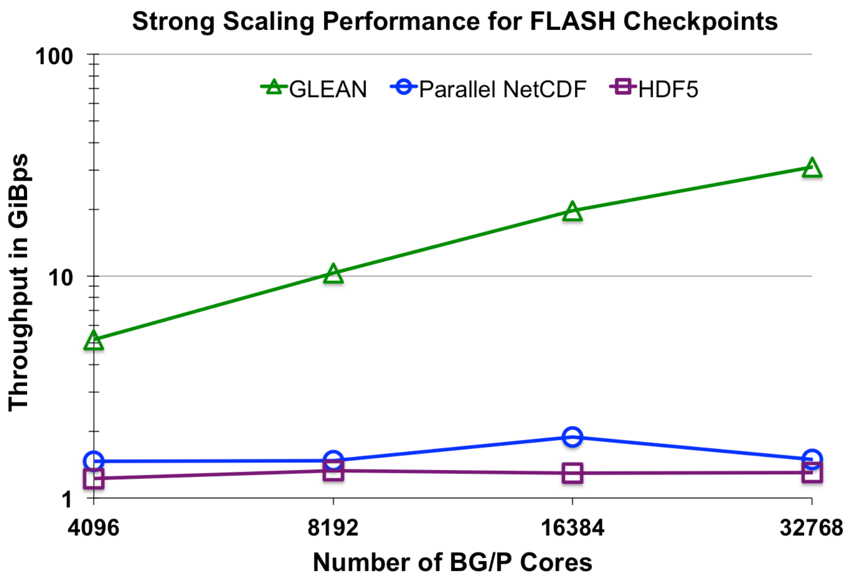
Data Staging on ALCF Resources



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

Performance for FLASH checkpoints

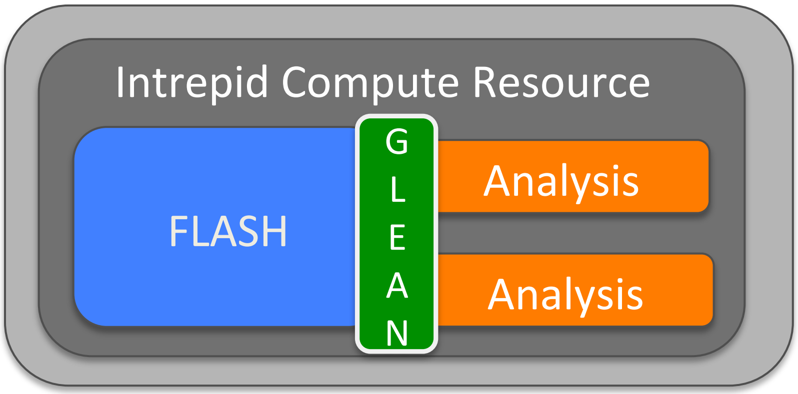




- For weak scaling at 32,768 cores, GLEAN sustains 31 GiBps and achieves an observed speedup of **10-fold** over pnetcdf and hdf5
- For strong scaling at 32,768 cores, GLEAN sustains 27 GiBps and achieves an observed speedup of **15-fold** over pnetcdf and hdf5
- 16.3 GiBps to Storage at 32K cores

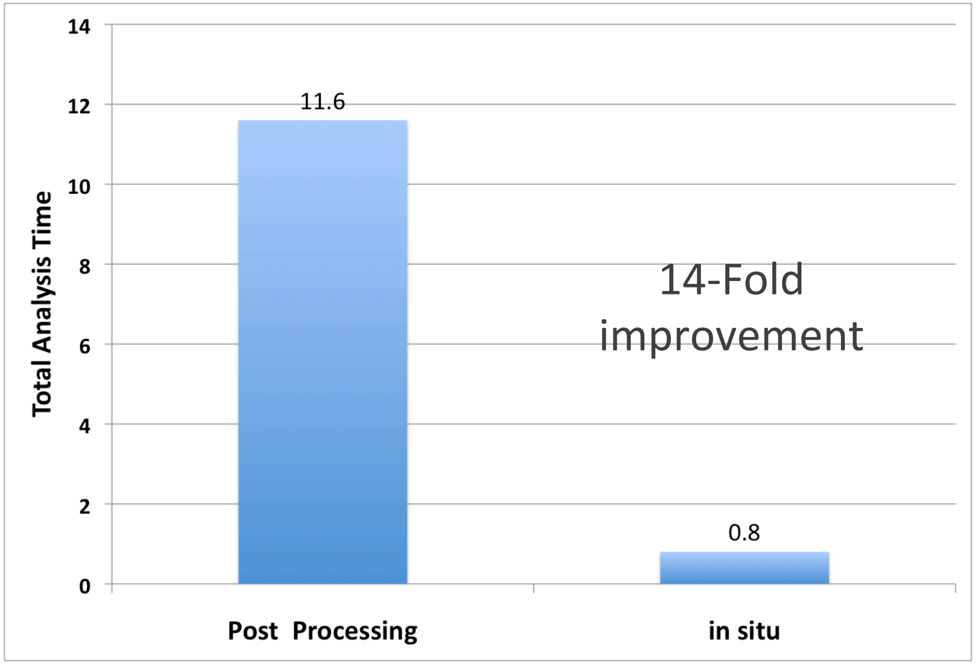
in situ analysis of FLASH using GLEAN

ALCF Facility

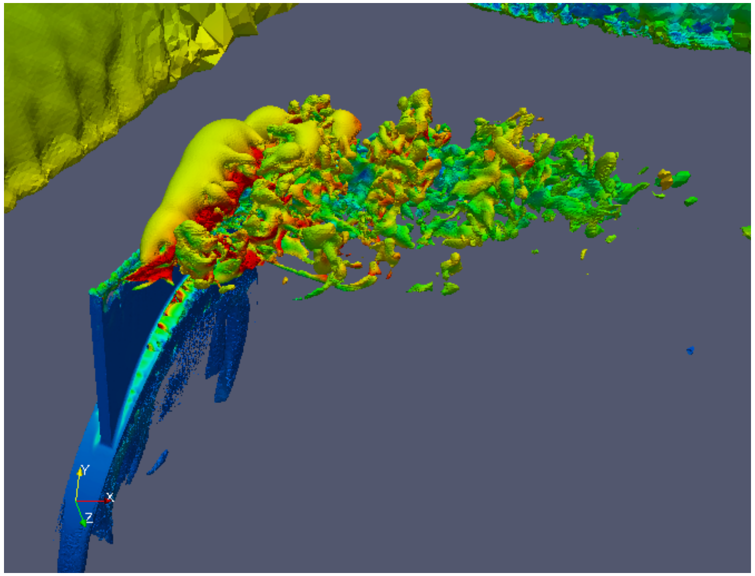


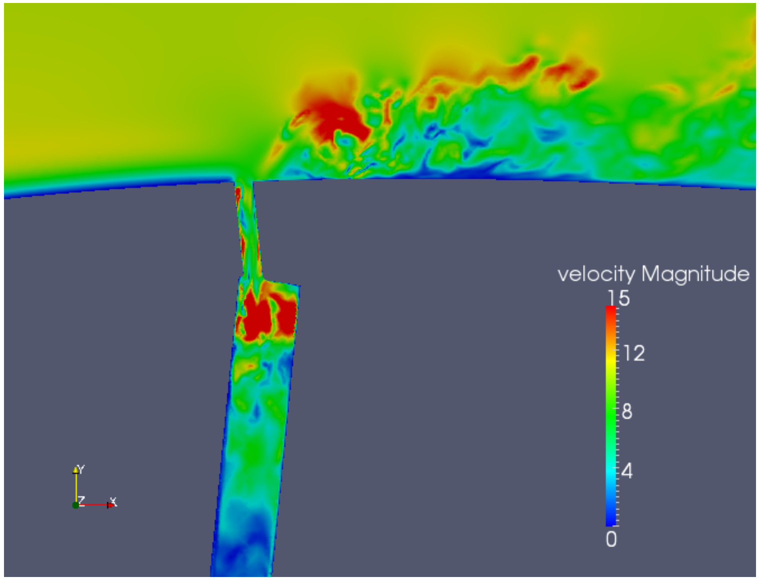
- in situ analysis to compute fractal dimension for 5 variables of a FLASH simulation on 2048 BG/P processors

- Fractal Dimension illustrates the degree of turbulence in a particular time step as well as within a sub-region of the domain



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 GBps** sustained throughput for data movement enabling simulation-time analysis

GLEAN- Enabling simulation-time data analysis and I/O acceleration

- Scaled to **entire ALCF (**160K BG/P cores + 100 Eureka Nodes)
- Provides I/O acceleration **by asynchronous data staging and topology-aware data movement** and achieved up **to 350-fold improvemen**t for FLASH and S3D at 32K cores (SC'10, SC'11[x2], LDAV'11)
- Leverages **data models** of applications including adaptive mesh refinement and unstructured meshes

Infrastructure	Simulation	Analysis
Co-analysis	PHASTA	Visualization using Paraview
Staging	FLASH, S3D	I/O Acceleration
In situ	FLASH	Fractal Dimension, Surface Area, Histograms
In flight	MADBench2	Histogram

- Design of algorithms for scalable data analytics
- Autonomic data movement infrastructure that takes into account node topology and system topology, is network aware and cognizant of application needs

Acknowledgements

- DOE Office of Advanced Scientific Computing Research
- Argonne Leadership Computing (ALCF) Resources
- ANL - Mike Papka, Mark Hereld, Joseph Insley, Eric Olson, Aaron Knoll, Tom Uram, Jiayuan Meng, Vitali Morozov, Kalyan Kumaran, Rob Ross, Tom Peterka, Rob Latham, Phil Carns, Kevin Harms, Kamil Iskra, Susan Coughlan, Ray Loy, Ilya Safro and ALCF team
- FLASH Center – Chris Daley, George Jordan, John Norris, Anthony Scopatz, Milad Fatnejad, Carlo Graziani and Don Lamb
- IIT – Zhiling Lan, Wei Tang and Ziming Zheng
- Kitware - Pat Marion and Berk Geveci
- Univ of Colorado– Ken Jansen, Michel Rasquin and Ben Matthews
- Univ. of Utah – Valerio Pascucci, Shusen Liu, Sidharth Kumar
- Texas A&M – Valarie Taylor, Adrian Salazar, Xingfu Wu

venkatv@mcs.anl.gov