Argonne NATIONAL LABORATORY

Data Analysis Approaches at Extreme Scale

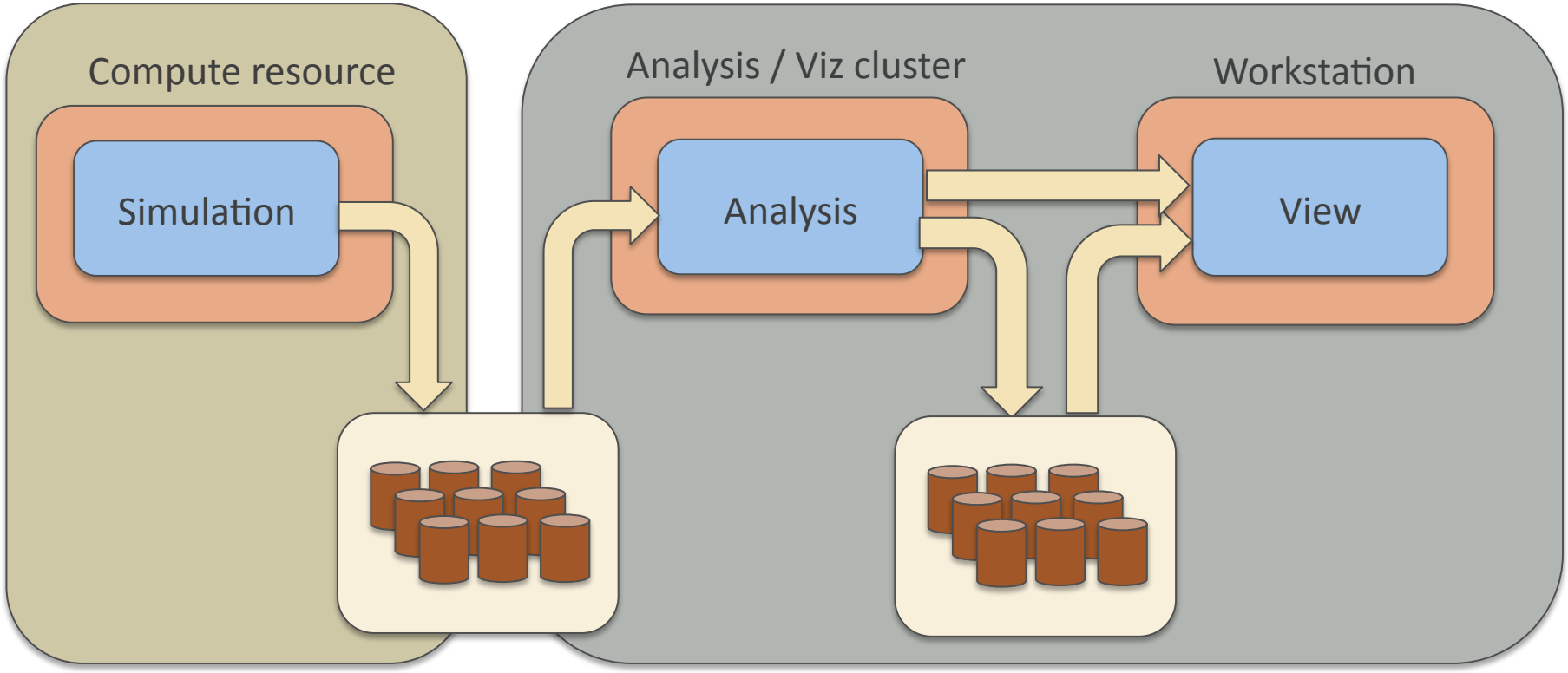
Venkatram Vishwanath

Argonne National Laboratory

venkatv@mcs.anl.gov

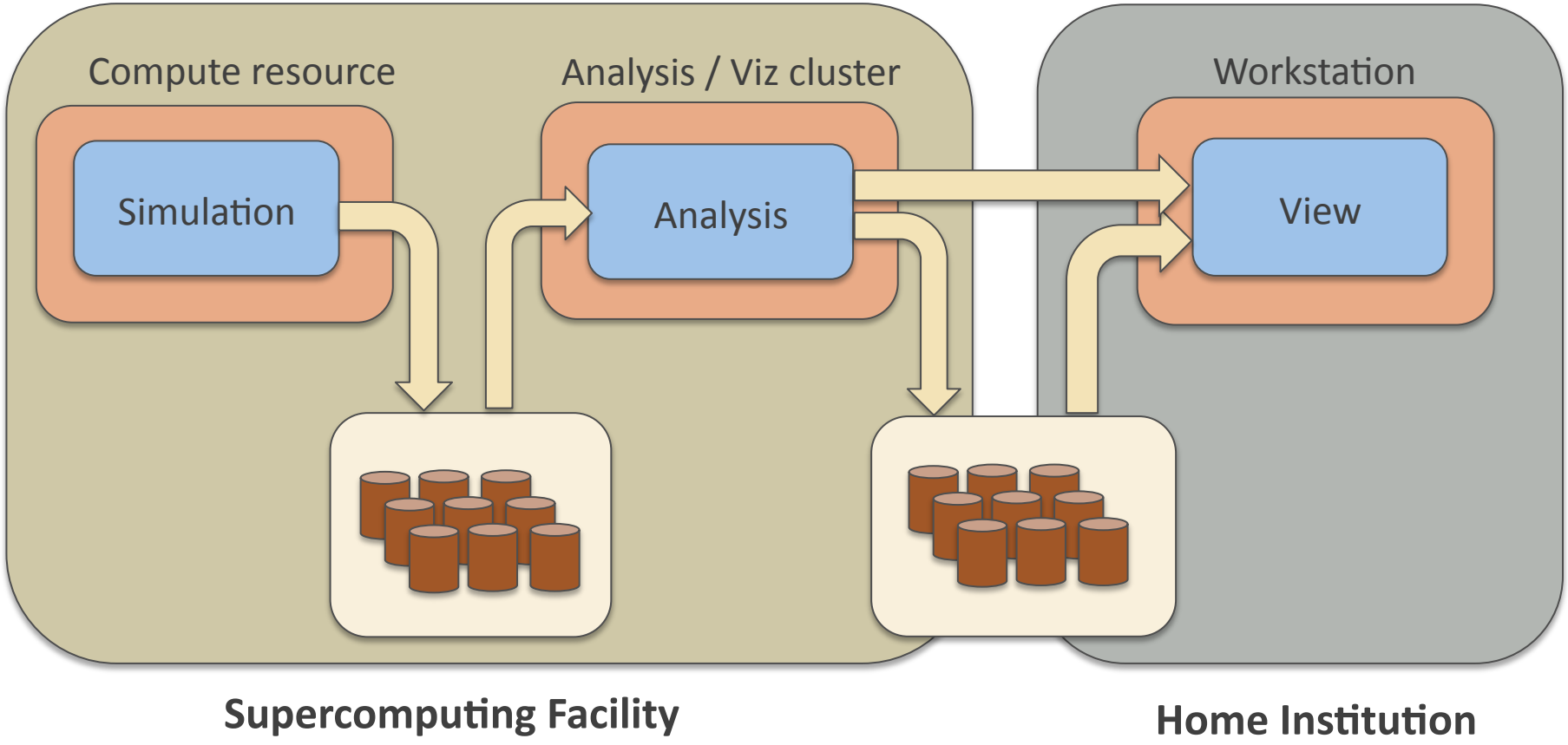
U.S. DEPARTMENT OF ENERGY

Traditional Science Pipeline



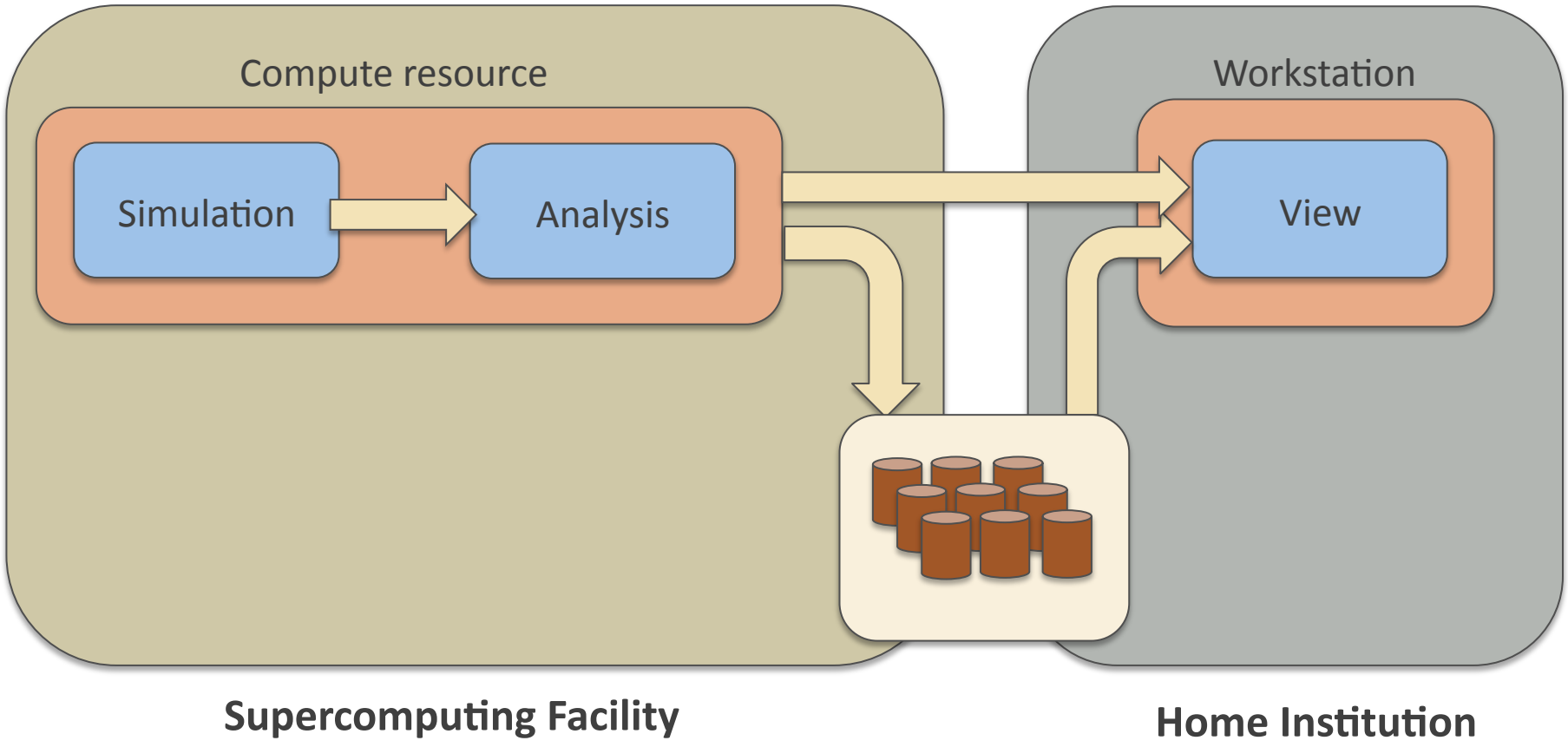
- Time to discovery is high as we are moving data to and from storage.
- Due to the data size, moving data is a challenge, and, we need ultra-high-speed networks to enable this.

Post Processing Pipeline in HPC



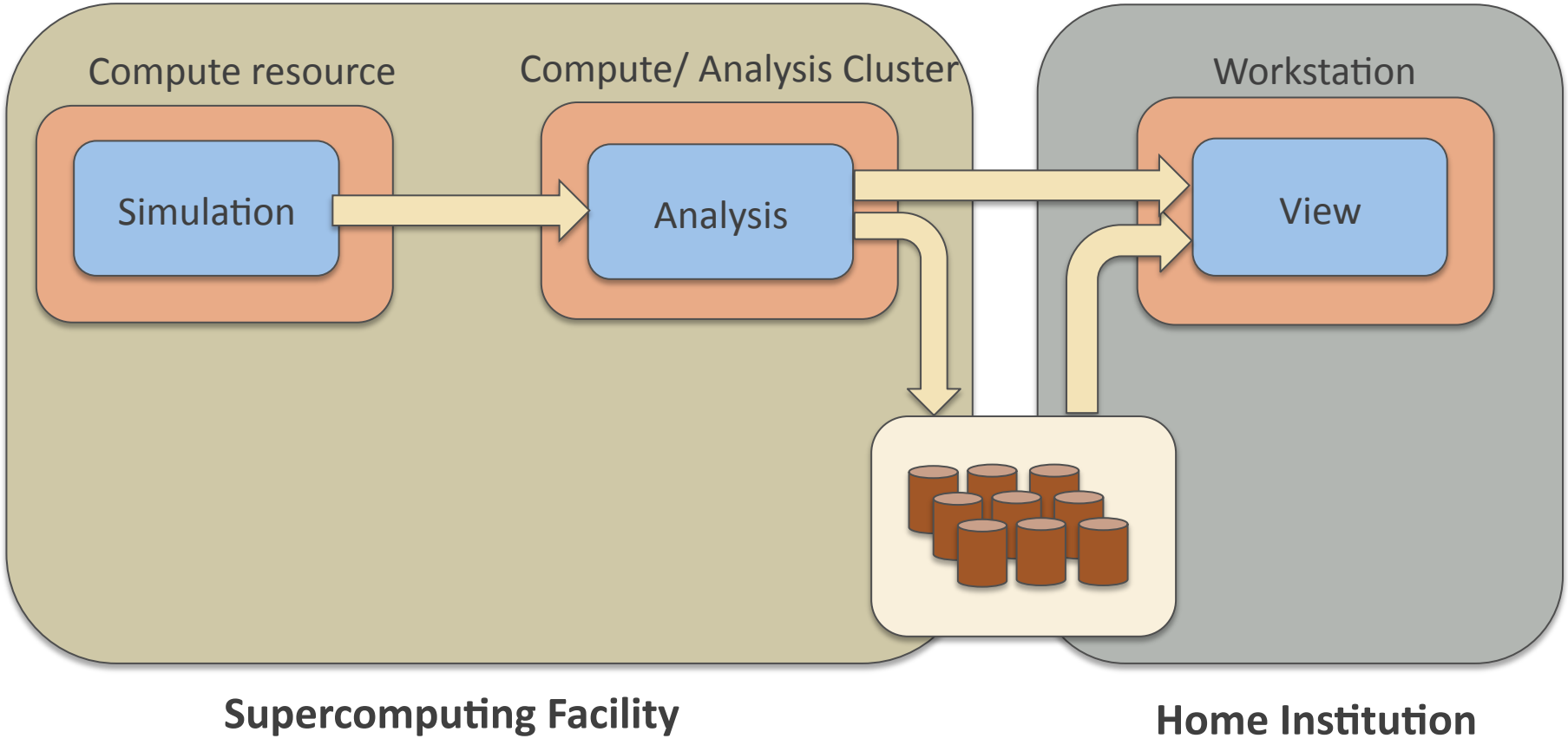
We are currently unable to cope with the data sizes at extreme scales and this will only get worse in future

Simulation Time Analysis on the Compute Resource - *in situ*



Simulation computation and analysis occur on the compute resource

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



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

Comparing simulation time in situ and simulation time co-analysis

In situ

- ✓ uses simulation data structures
- ✓ no additional hardware resource required
- x uses simulation data structures, time and memory
- x time-varying analysis is extremely difficult

Co-analysis

- ✓ extremely flexible
- ✓ time-varying analytics
- ✓ may not require precious computational resources
- x requires a co-scheduling infrastructure

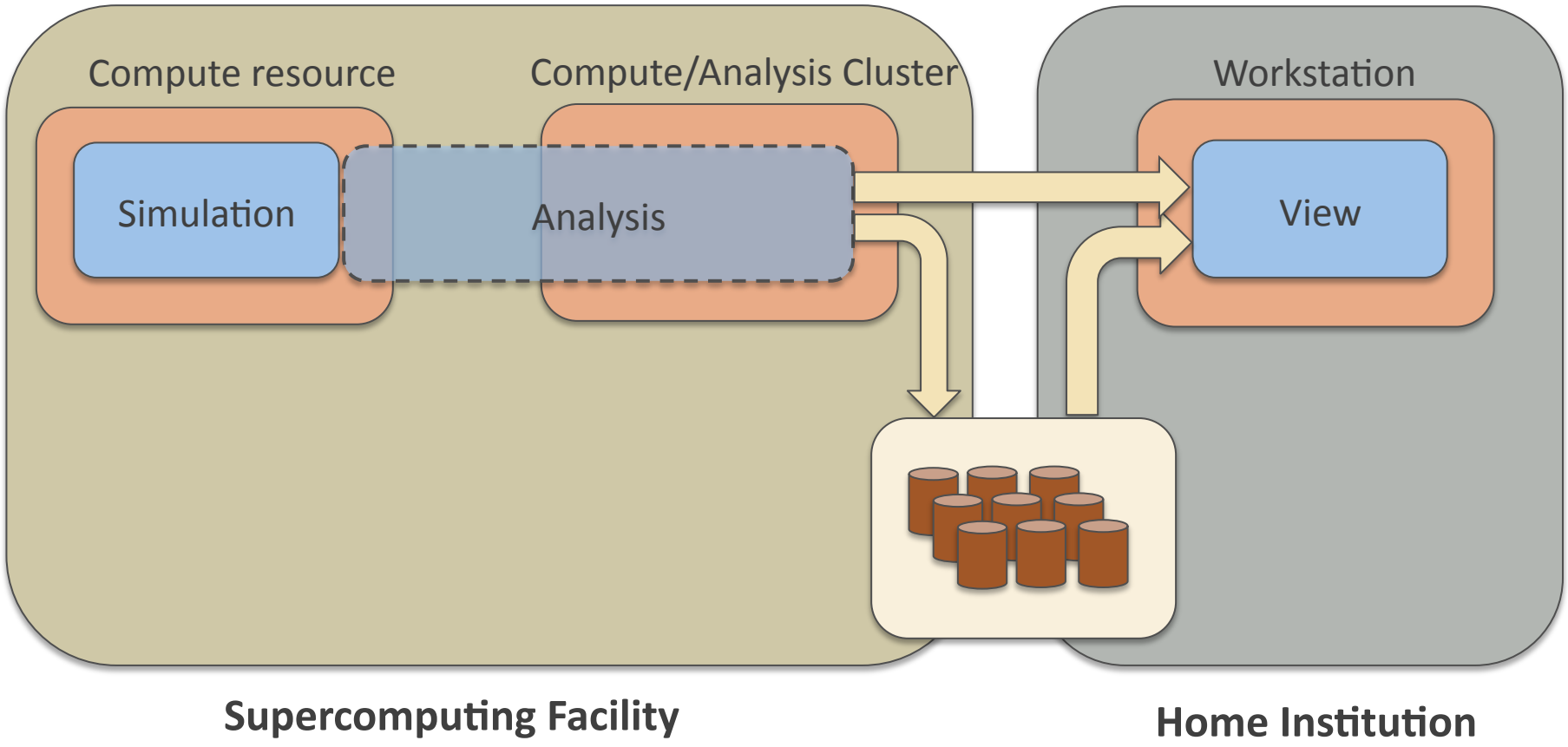
How do I analyze my data?
In situ, Co-Analysis or Post Processing?

There are multiple answers!!

A solution needs to consider, among others:

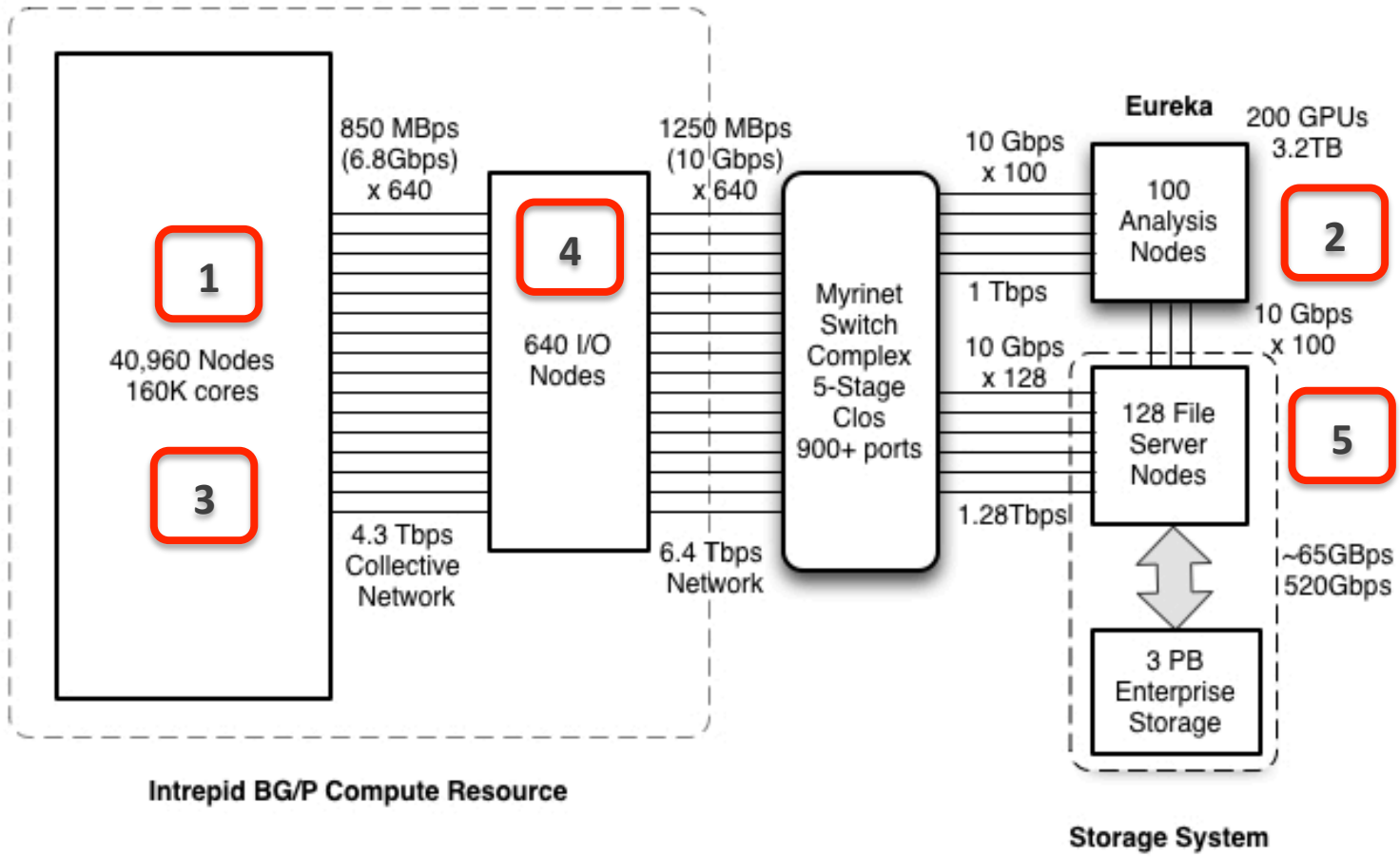
- Simulation characteristics
- Analysis characteristics
- System characteristics
- Be flexible enough to meet the needs of the scientist

Our Approach



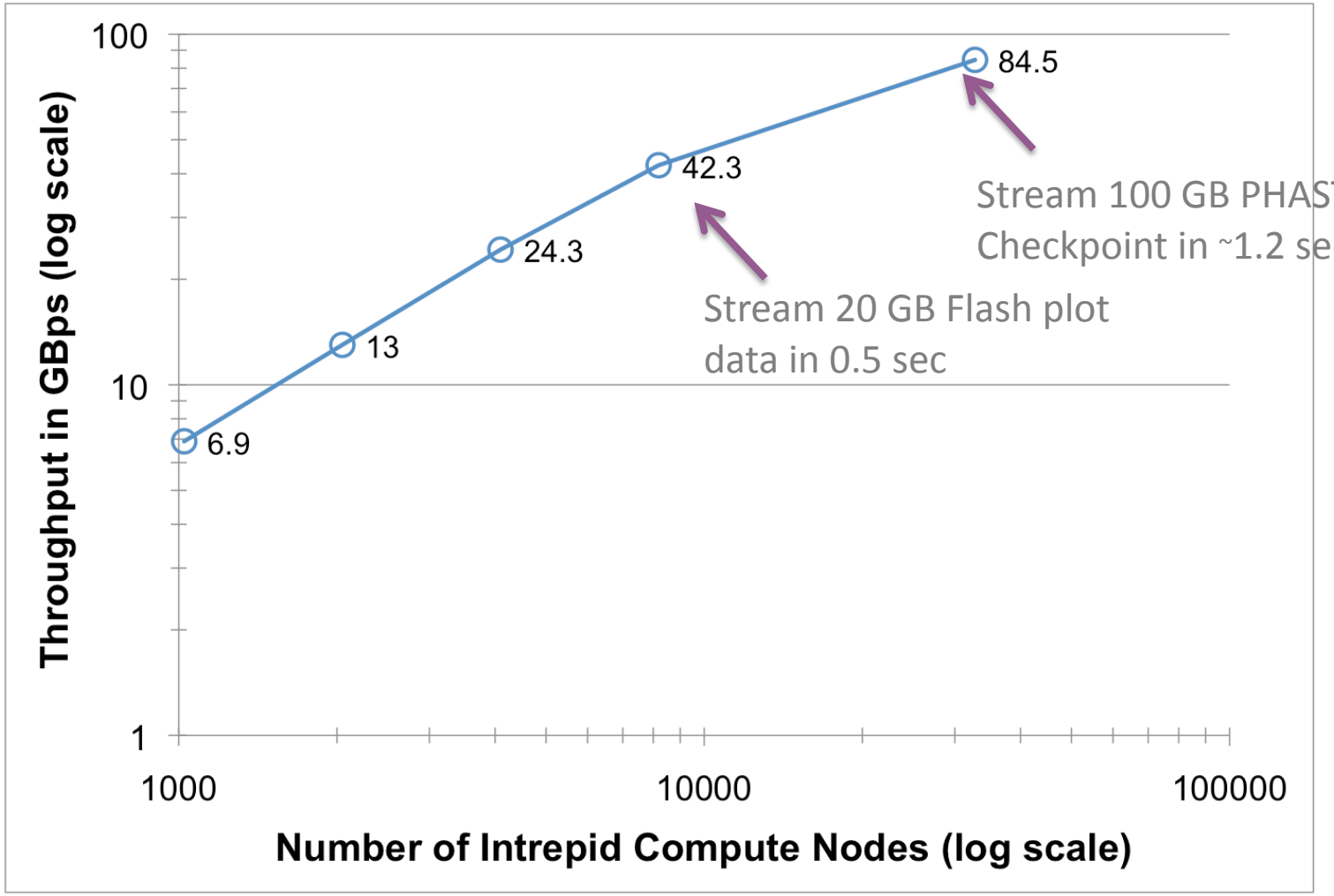
A flexible, hybrid approach taking into account application, analytics and system characteristics

Simulation-time Analysis Opportunities on ALCF Intrepid



The analysis abstractions are applicable to other leadership class infrastructures including systems at ORNL, NICS, TACC and NCSA

Data Streaming from Intrepid



V. Vishwanath et al., " *Accelerating I/O Forwarding in IBM Blue Gene/P Systems*", To appear at SC 2010, New Orleans, November 2010

Datasets: Application view vs Storage view

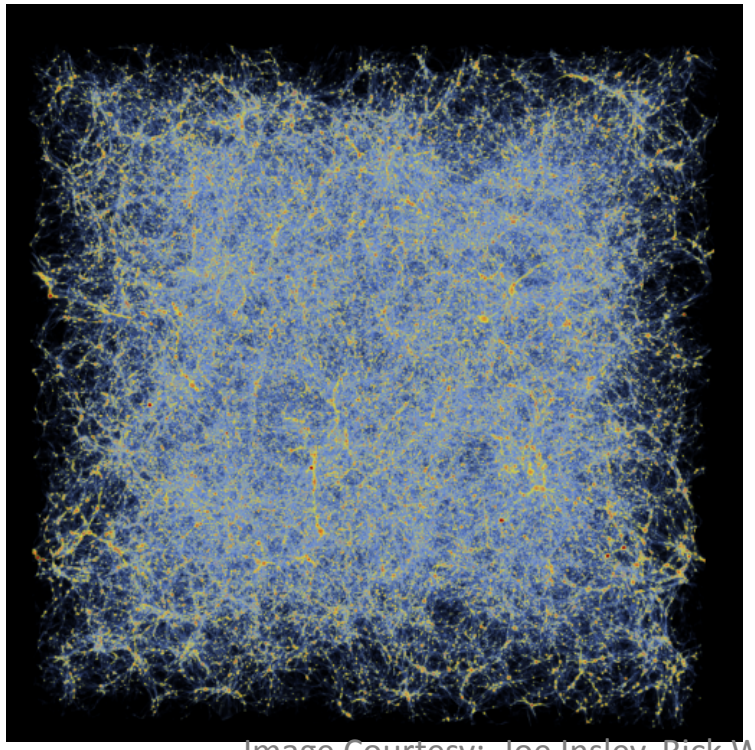
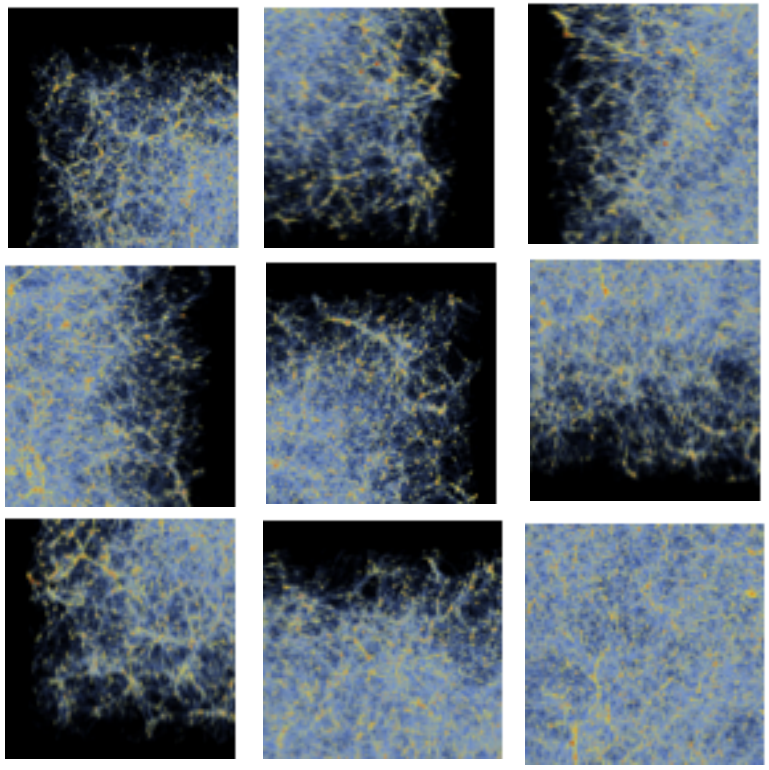


Image Courtesy: Joe Insley, Rick Wagner

Application View
$(4096)^3$ pressure



Storage View
4096 files 64MB each

There is a growing disparity between an application's data models and how the data is mapped onto storage

Enhancing productivity by capturing the workflows of scientists

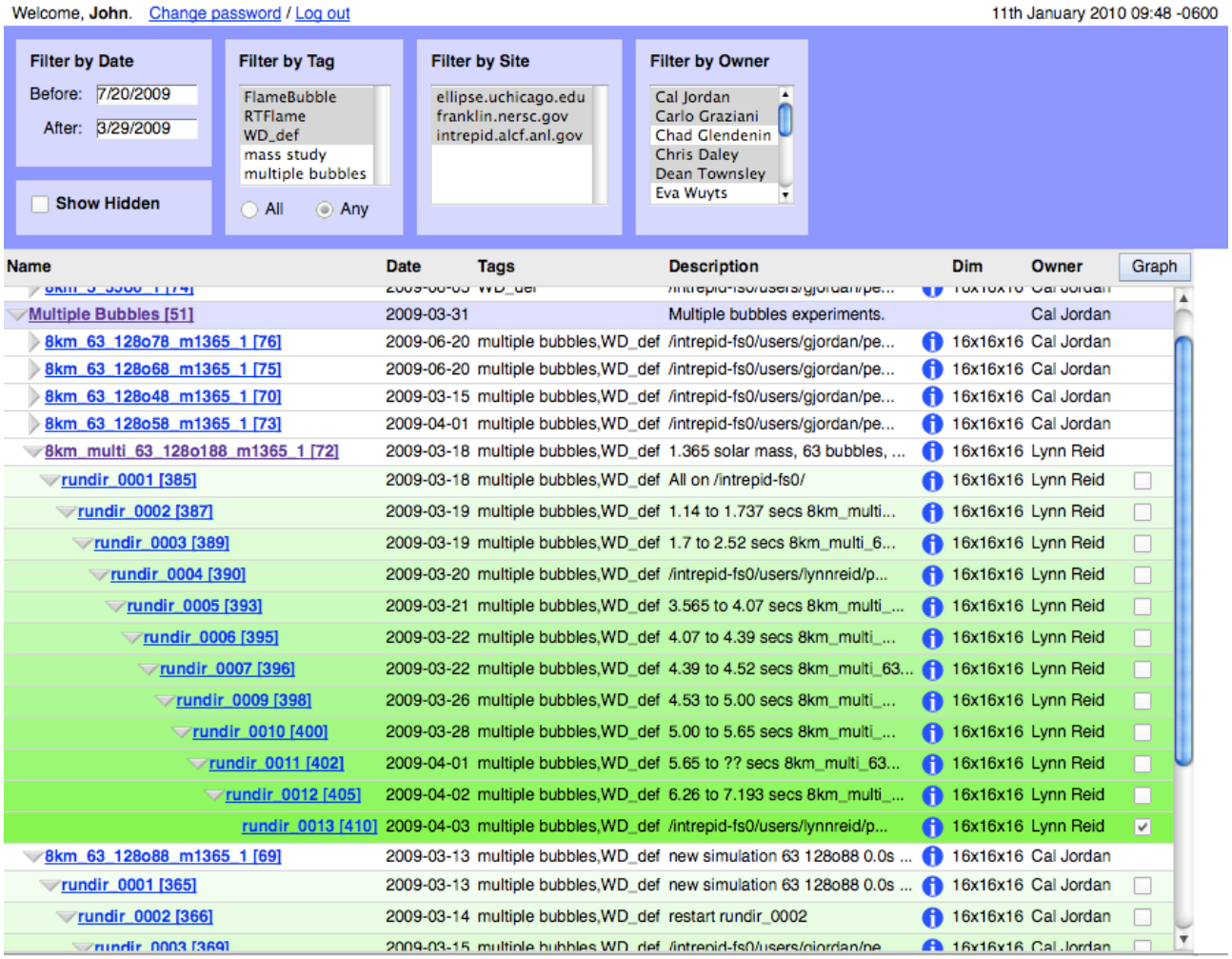
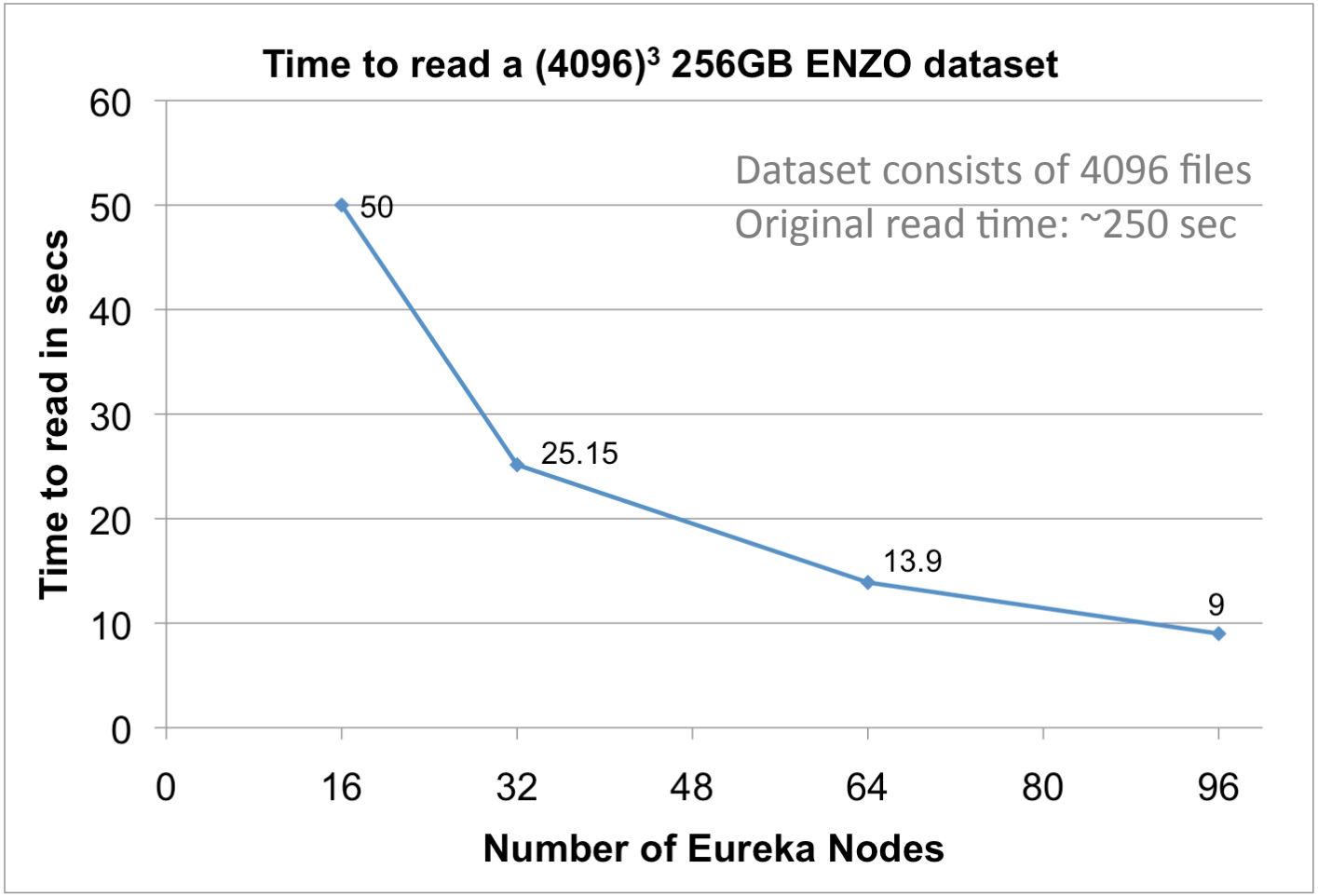


Image Courtesy: Randy Hudson, Flash Center

Accelerating ENZO Analysis by Leveraging Data Semantics



Data semantics will be of extreme importance at exascale

Parting Thoughts

- Simulation time data analysis at extreme scales is critical in order to reduce the amount of data written.
- There are multiple solutions for simulation time analysis. A solution needs to take into account resource, application and analysis characteristics.
- Leveraging data models will be key in order to deal with future data sizes

Acknowledgements

- DOE Office of Advanced Scientific Computing Research
- ANL Director's Fellow Award
- Mike Papka, Mark Hereld, Joe Insley, Eric Olson, Tom Uram, John Norris, Randy Hudson, Rob Ross, Tom Peterka, Rob Latham, Phil Carns, Kamil Iskra, Dries Kimpe, Vitali Morozov and Kazutomo Yoshii
- Rick Wagner, Robert Harkess and Mike Norman for the ENZO data
- Argonne Leadership Computing (ALCF) Resources and the ALCF Team