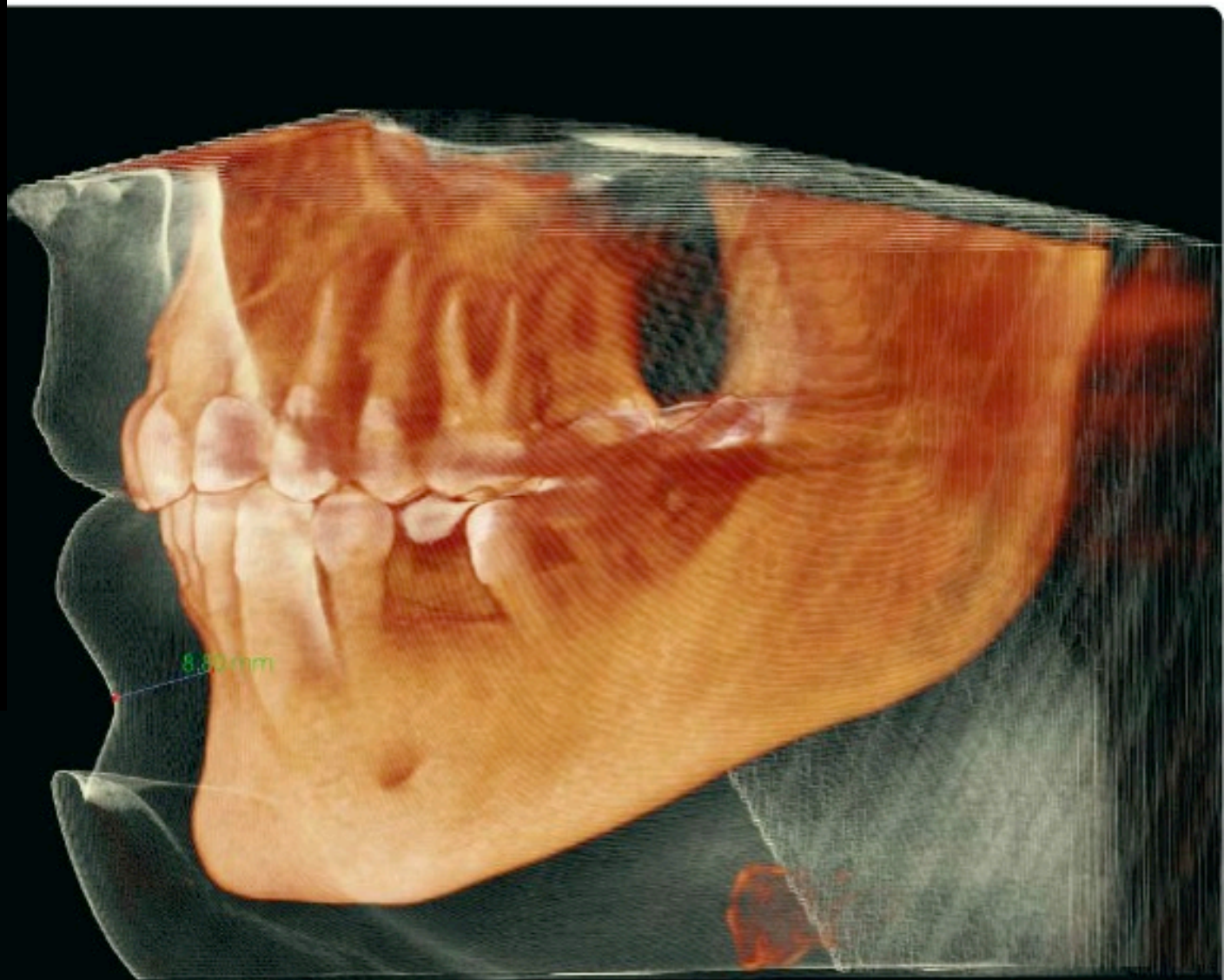
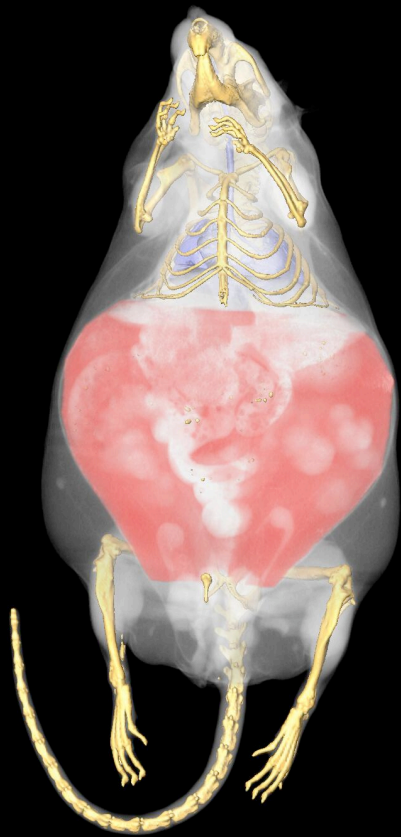
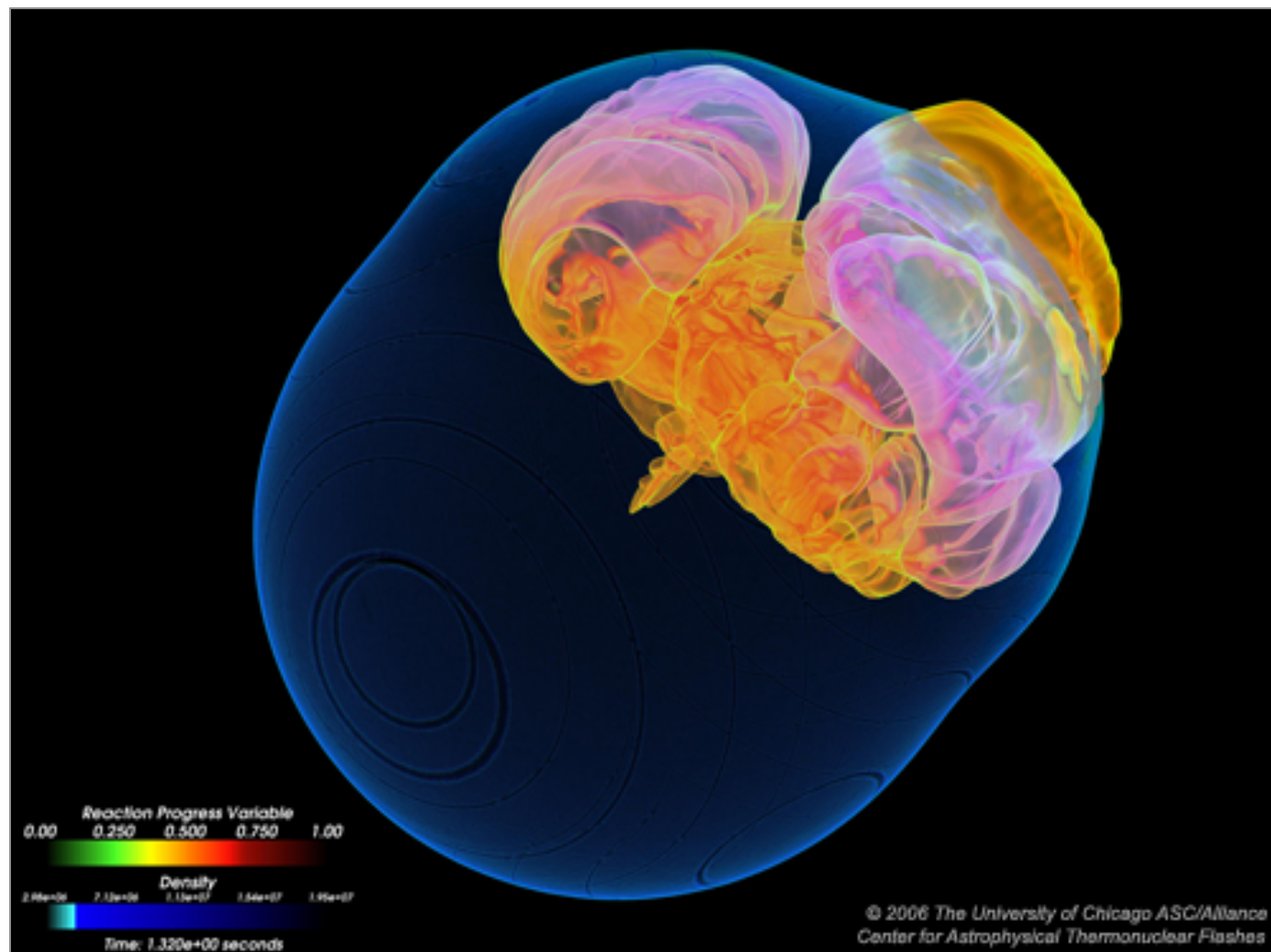


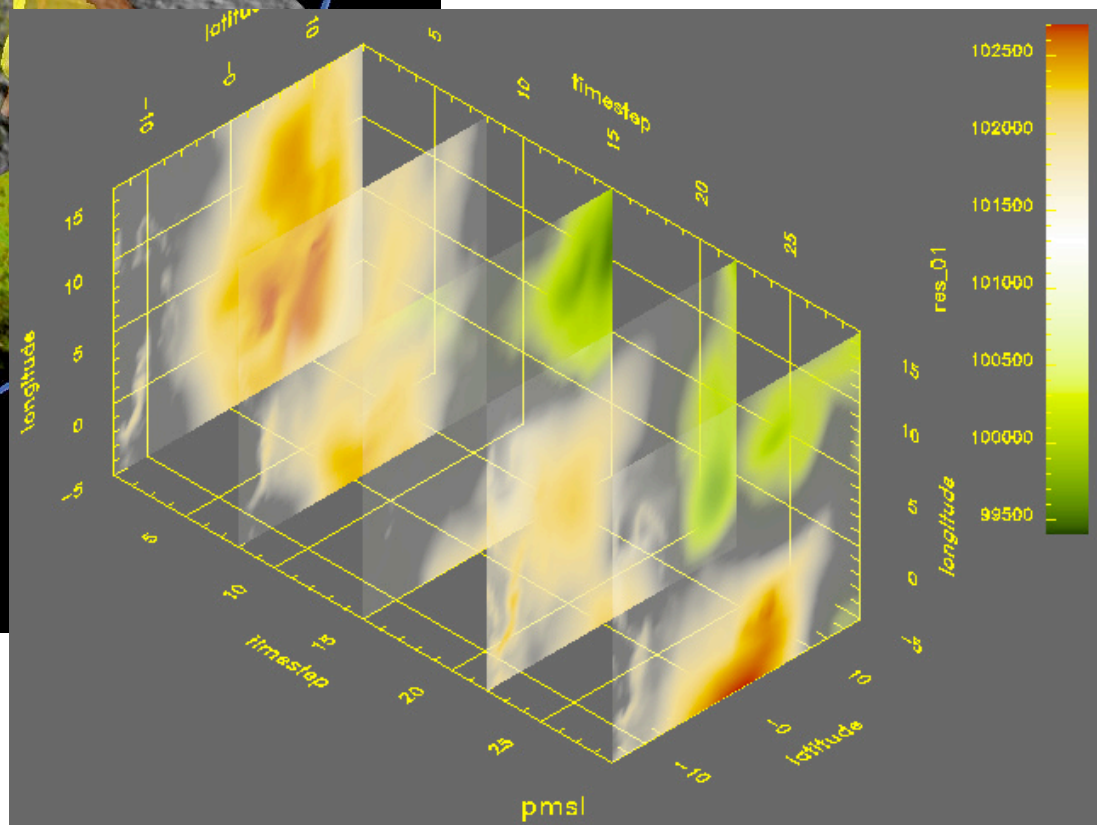
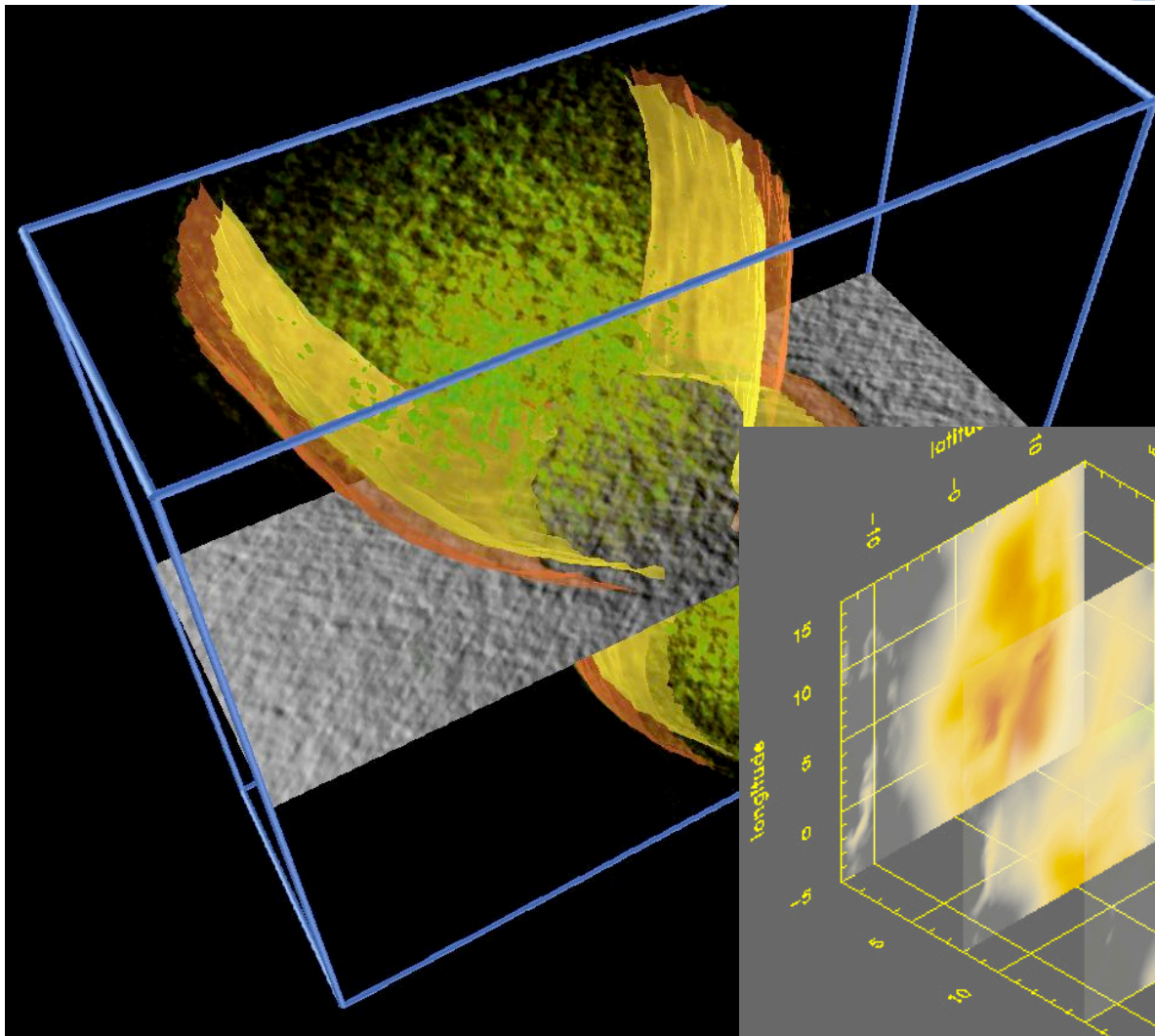
Data Analysis and Visualization

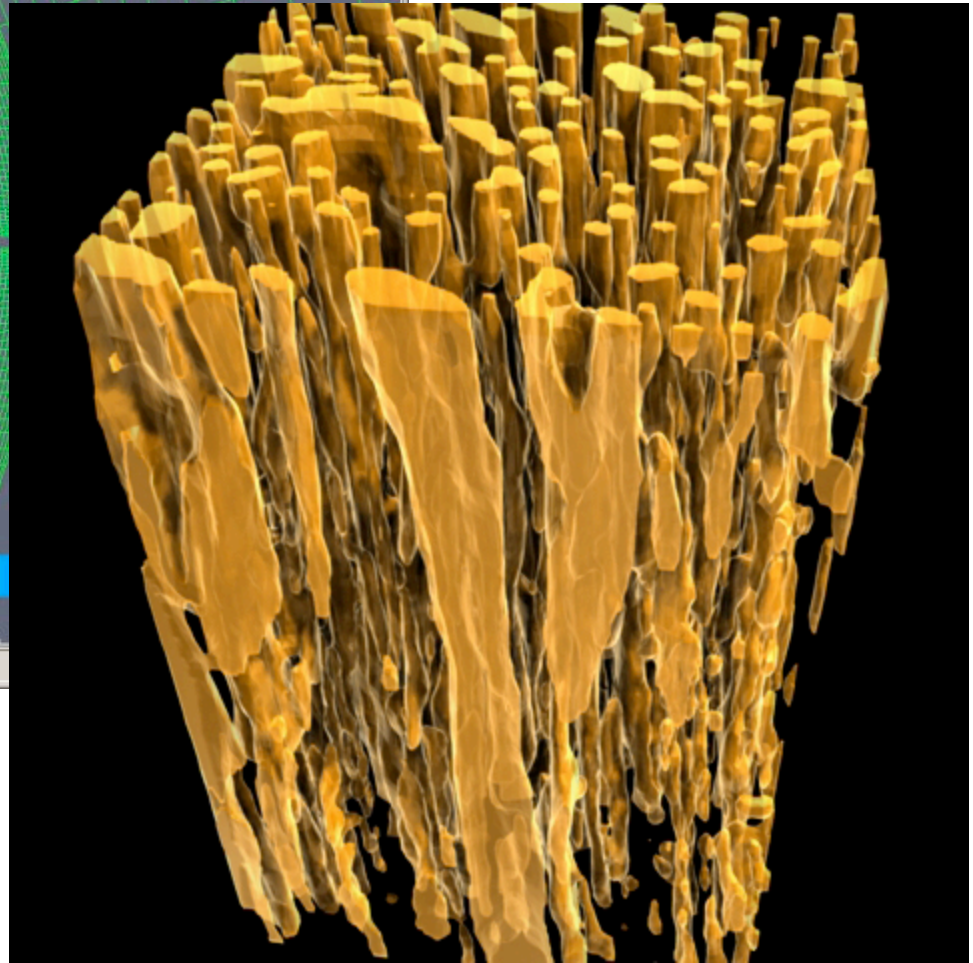
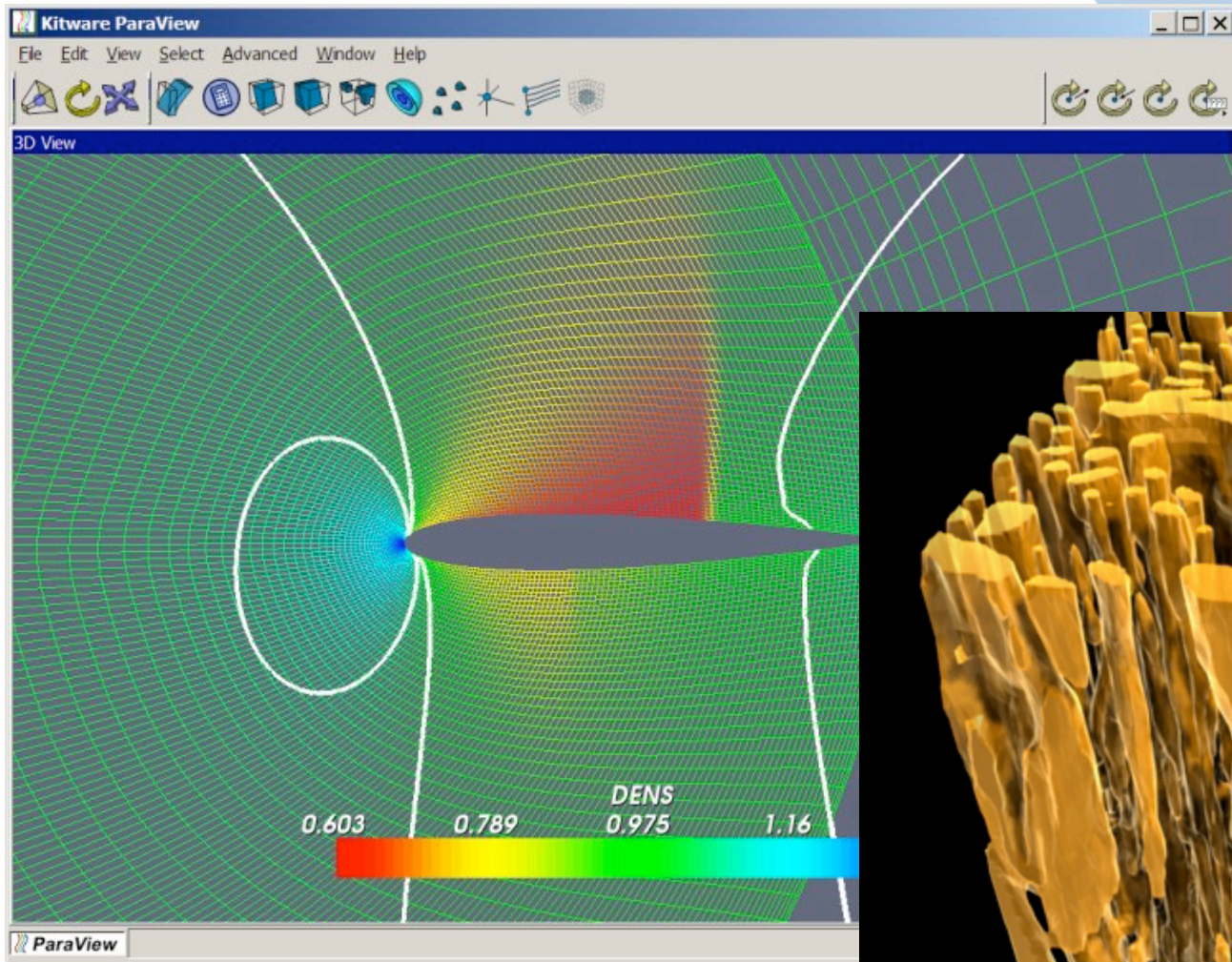
CScADS 2011

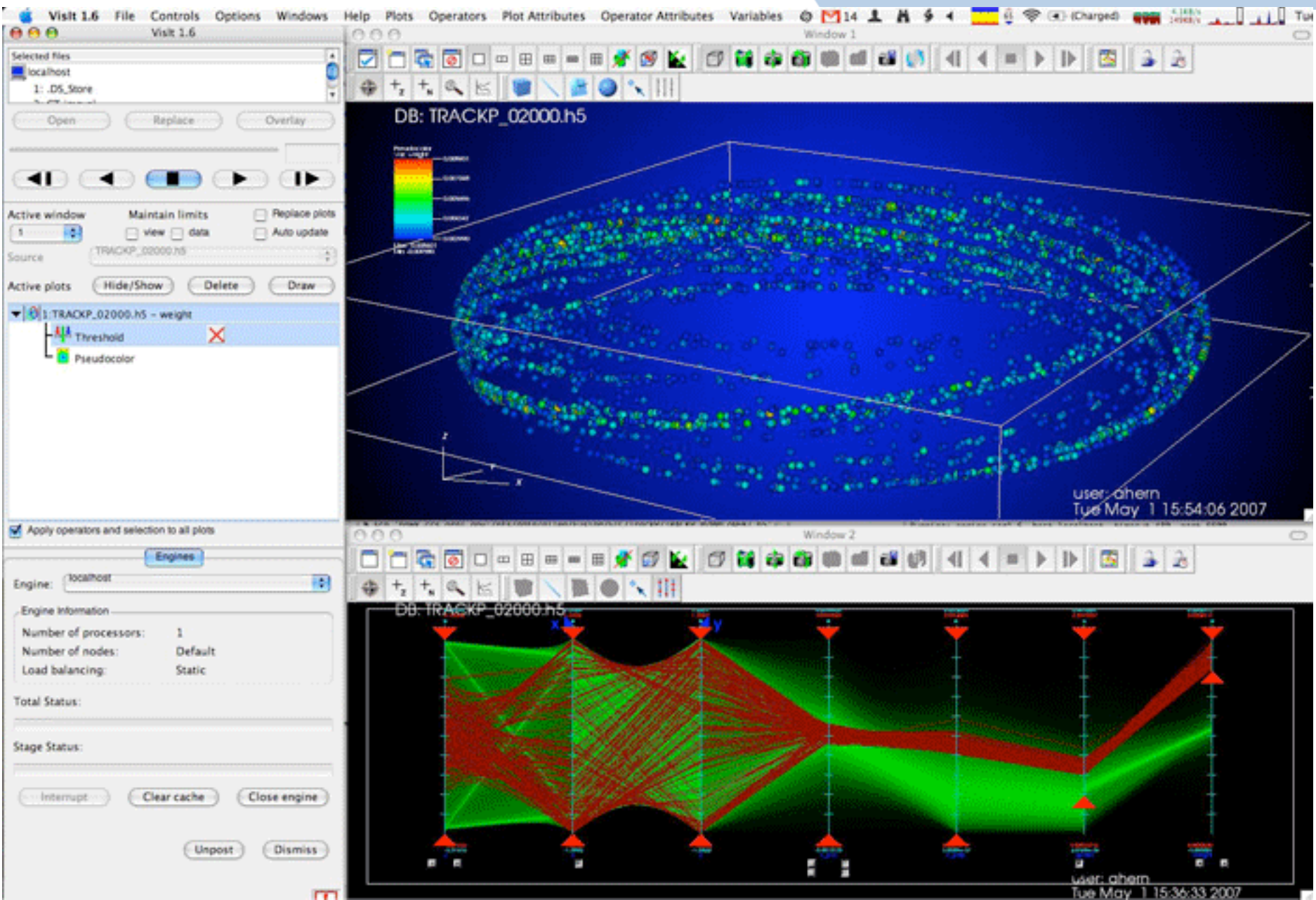
Venkatram Vishwanath











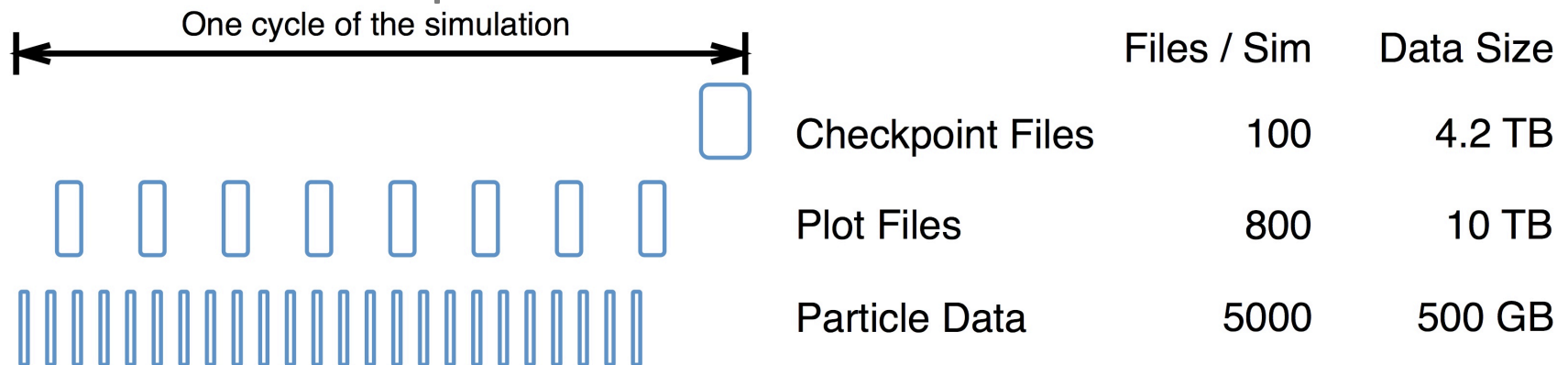
Your Goals

- What do you look to get out of analysis process?
- What analysis tools are you currently using?
 - What are the limitations?
- Do you do real-time exploration or batch processing?
 - What is the role of real-time exploration?
- Batch?
 - Percentage of your analysis time spent in either mode?
- Do you look at images, movies or graphs?
 - What is role of each (e.g. graphs for science, images publications, movies for talks)



Your Footprint

- How long do your simulations run?
 - Is the result a time series?
 - How many files does that produce?
- How much of your simulation time is I/O?
- What are your dataset sizes?
 - checkpoint files, variables, species, analysis files
- How long do you spend on analysis, what is the fraction of compute versus human?

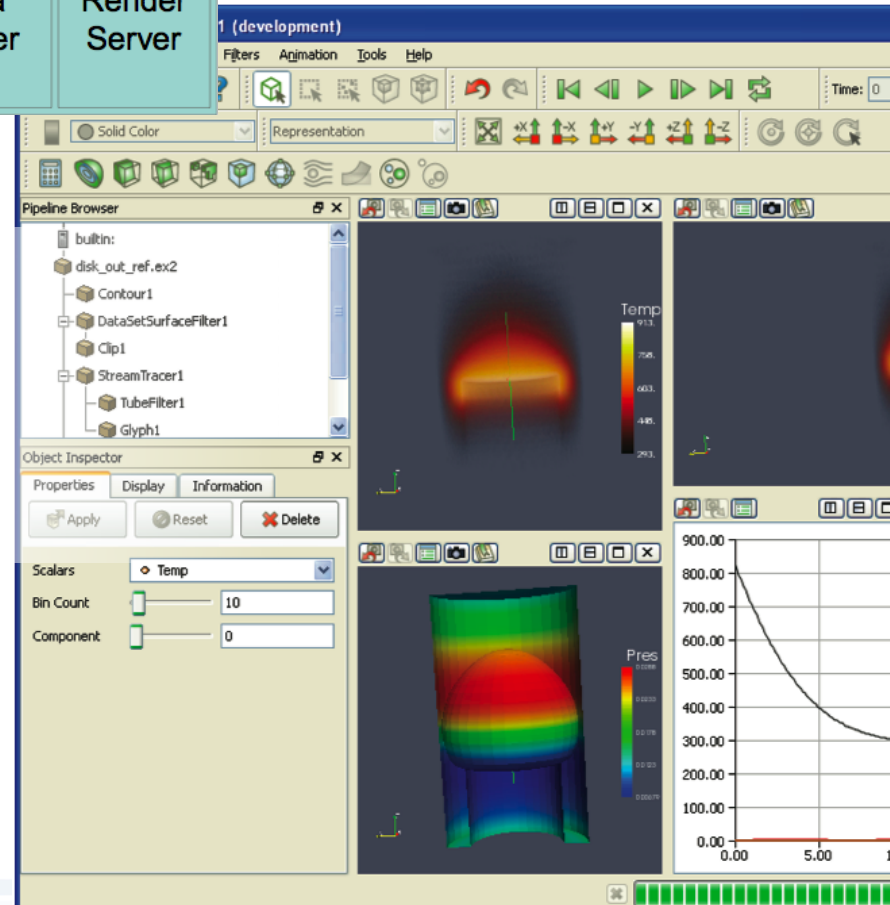
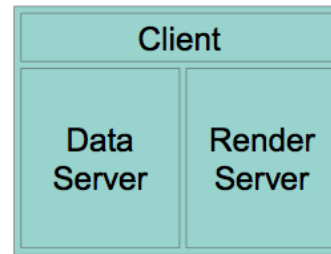
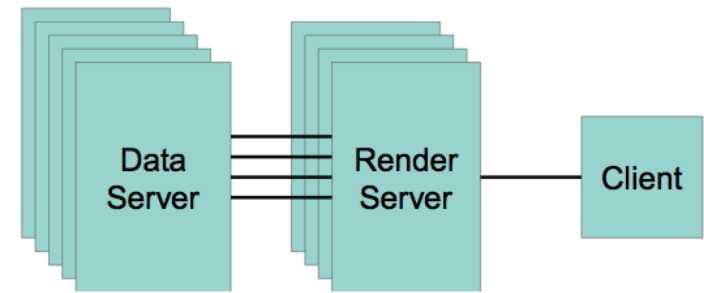


All Sorts of Tools

- Visualization Applications
 - VisIt
 - ParaView
 - EnSight
- Domain Specific
 - PyMol, RasMol
- APIs
 - VTK: visualization
 - ITK: segmentat & registration
- GPU performance
 - vl3: shader-based vol ren
 - Scout: GPGPU acceleration
- Analysis Environments
 - Matlab
 - Parallel R
- Utilities
 - GnuPlot
 - ImageMagick
- Visualization Workflow
 - VisTrails

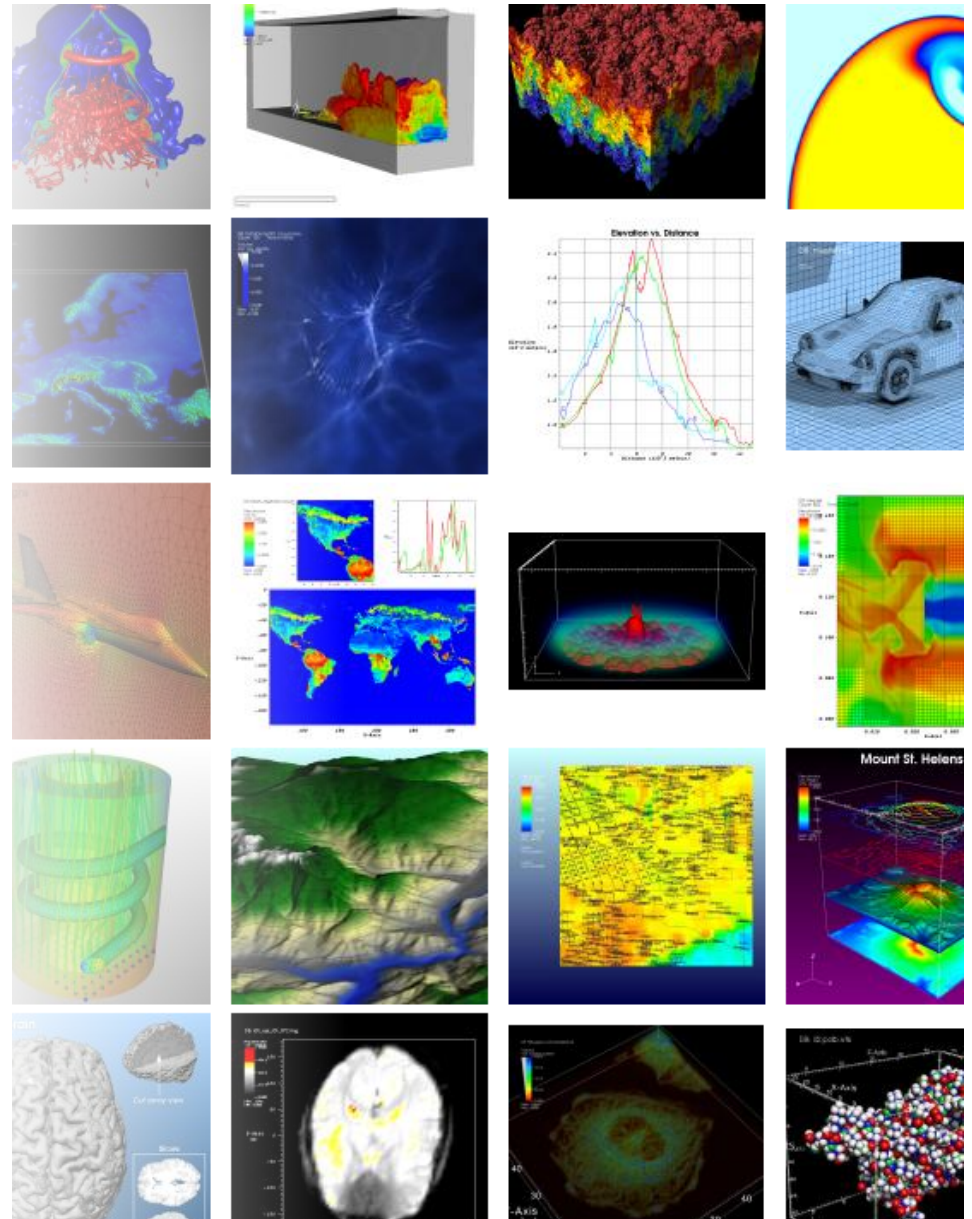
ParaView Overview

- Parallel Visualization Application
- Open source
- VTK + Tcl
- Python scripting
- Interactive and batch
- About
 - Kitware, Sandia National Labs, CSimSoft, LANL, Army Research ...and community
 - <http://www.paraview.org>



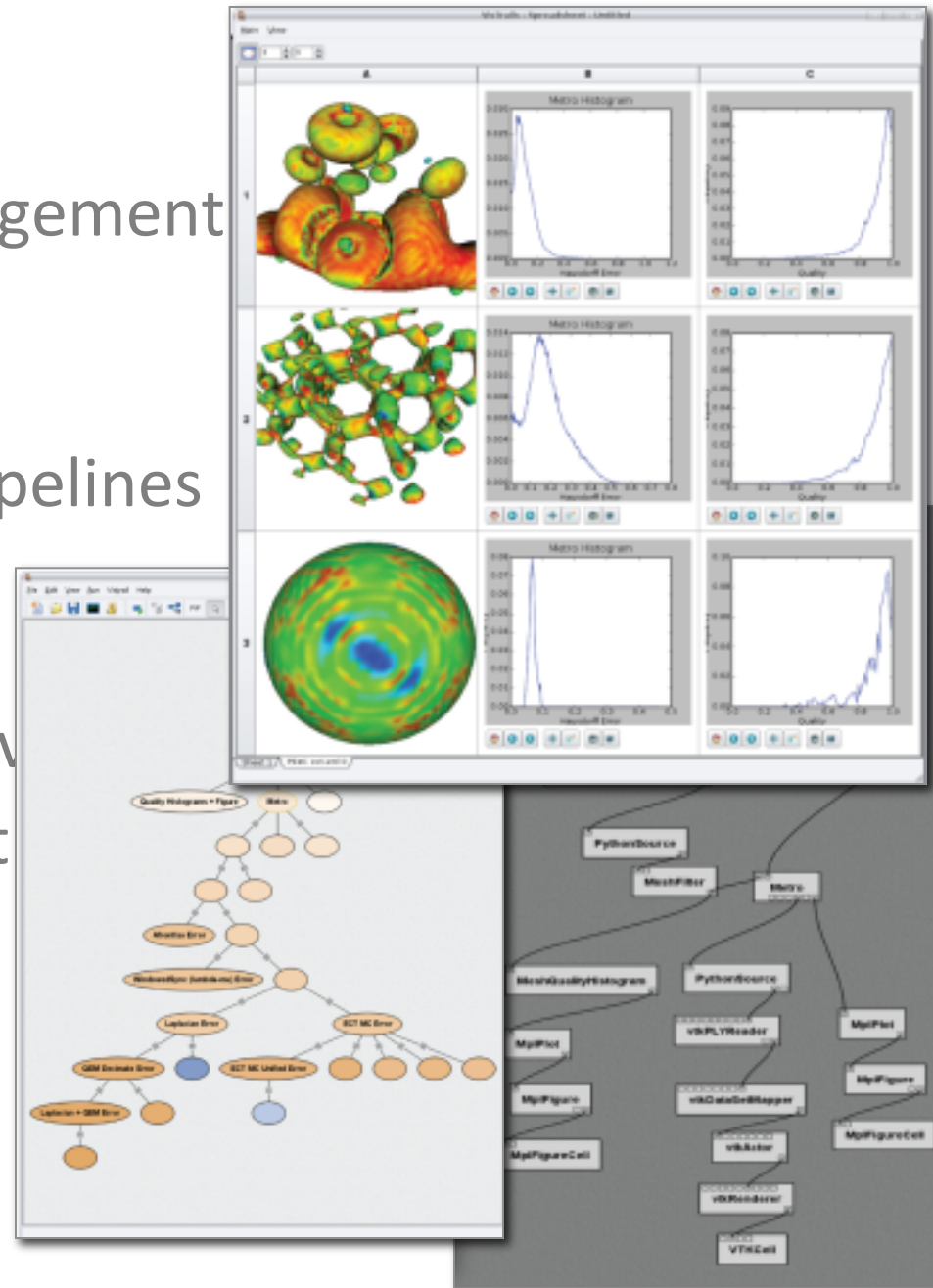
VisIt Overview

- Parallel interactive visualization application
- About
 - DOE ASCI
 - <https://www.llnl.gov/visit>



VisTrails

- Scientific workflow management for visual data analysis
- Construct and execute pipelines
 - Visual programming
 - VTK, ITK, and Matplotlib
- History tree captures provenance
- Visualization spreadsheet
- About
 - <http://www.vistrails.org>



ParaView Hands On

Tutorial

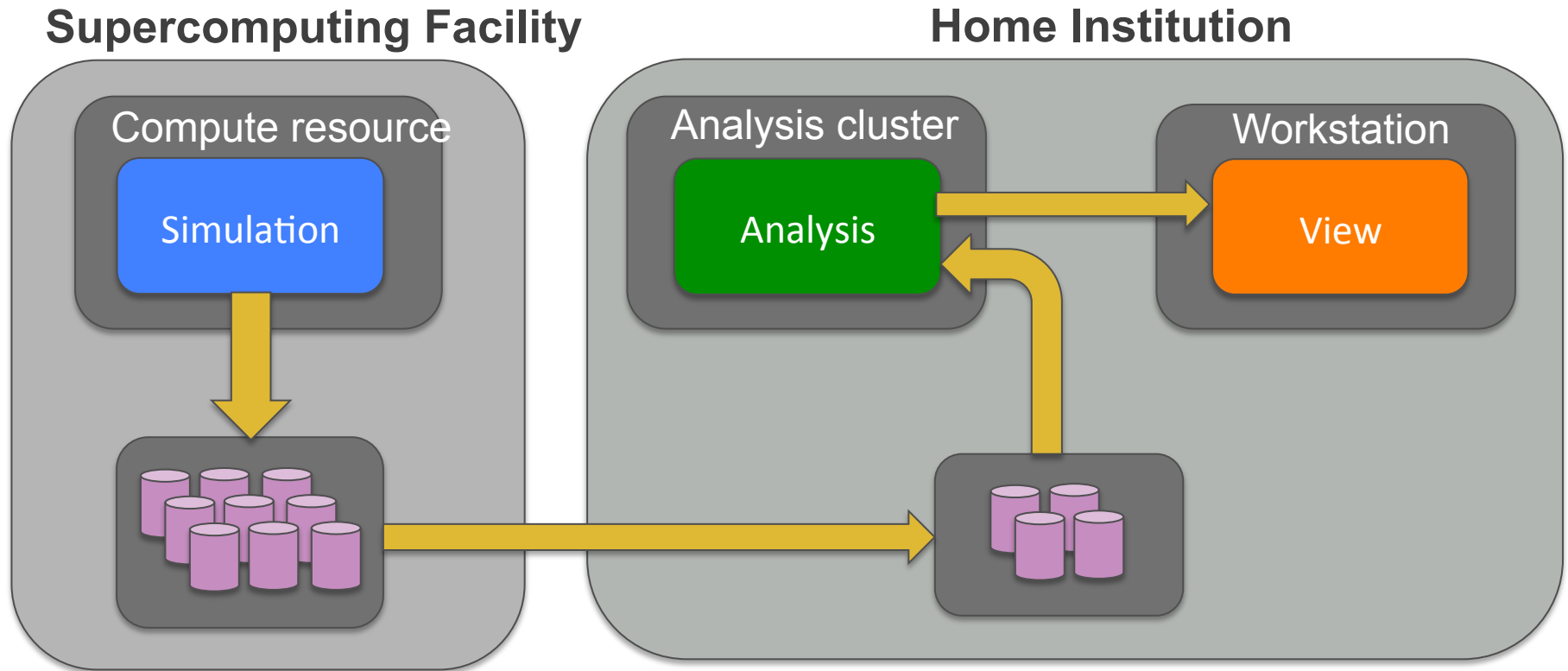
https://wiki.alcf.anl.gov/index.php/ParaView_Red_Blood_Cell_Tutorial

Download ParaView at:

<http://www.paraview.org/paraview/resources/software.html>



Traditional Science Pipeline

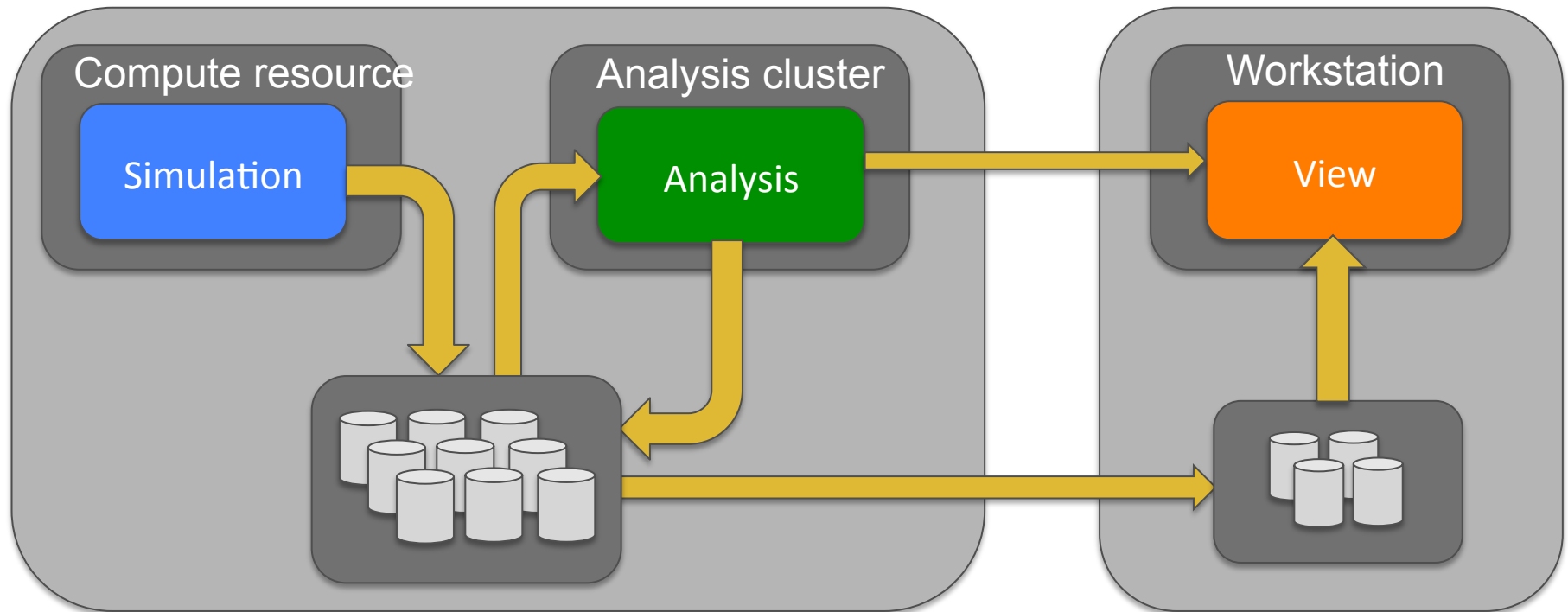


Time to discovery is high as we are moving data to and from storage

Post Processing Pipeline in HPC

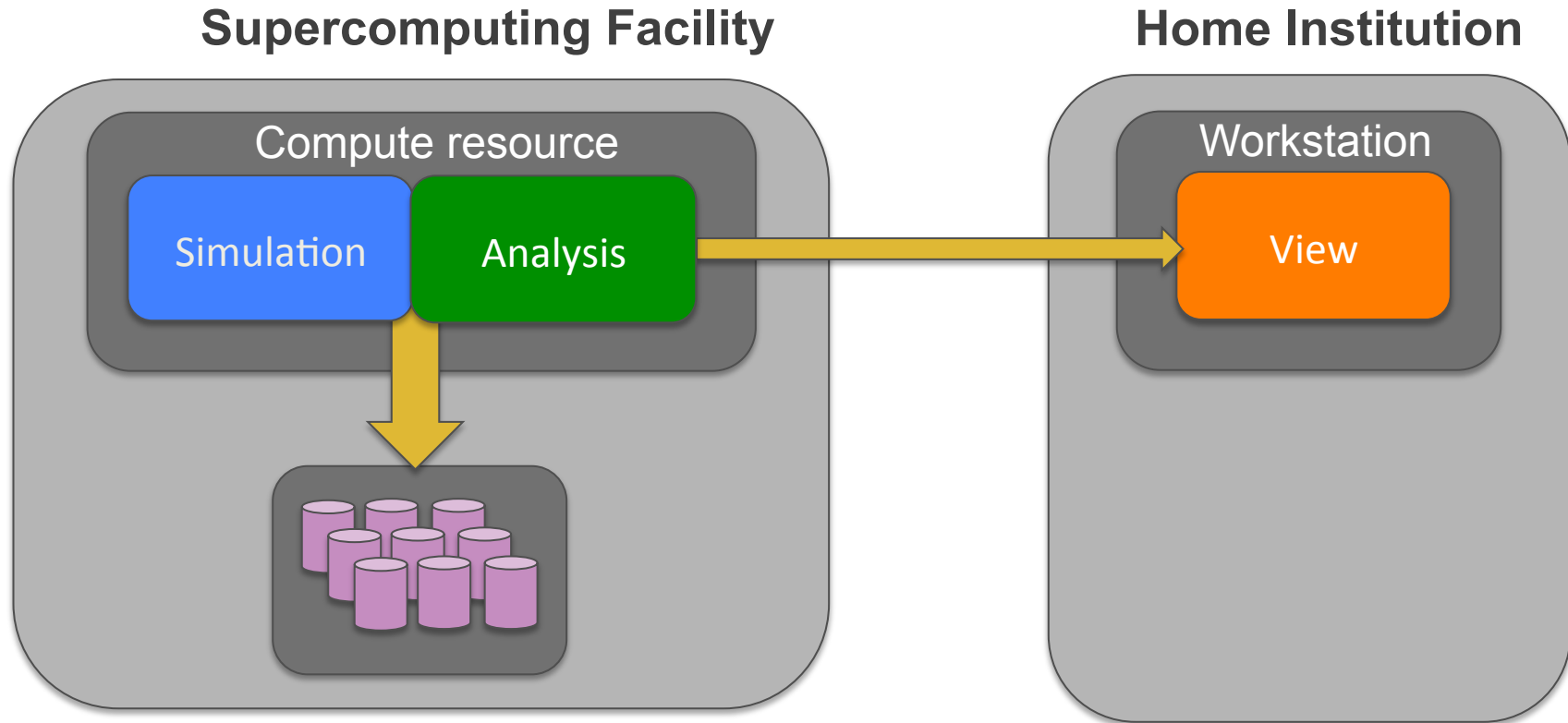
Supercomputing Facility

Home Institution



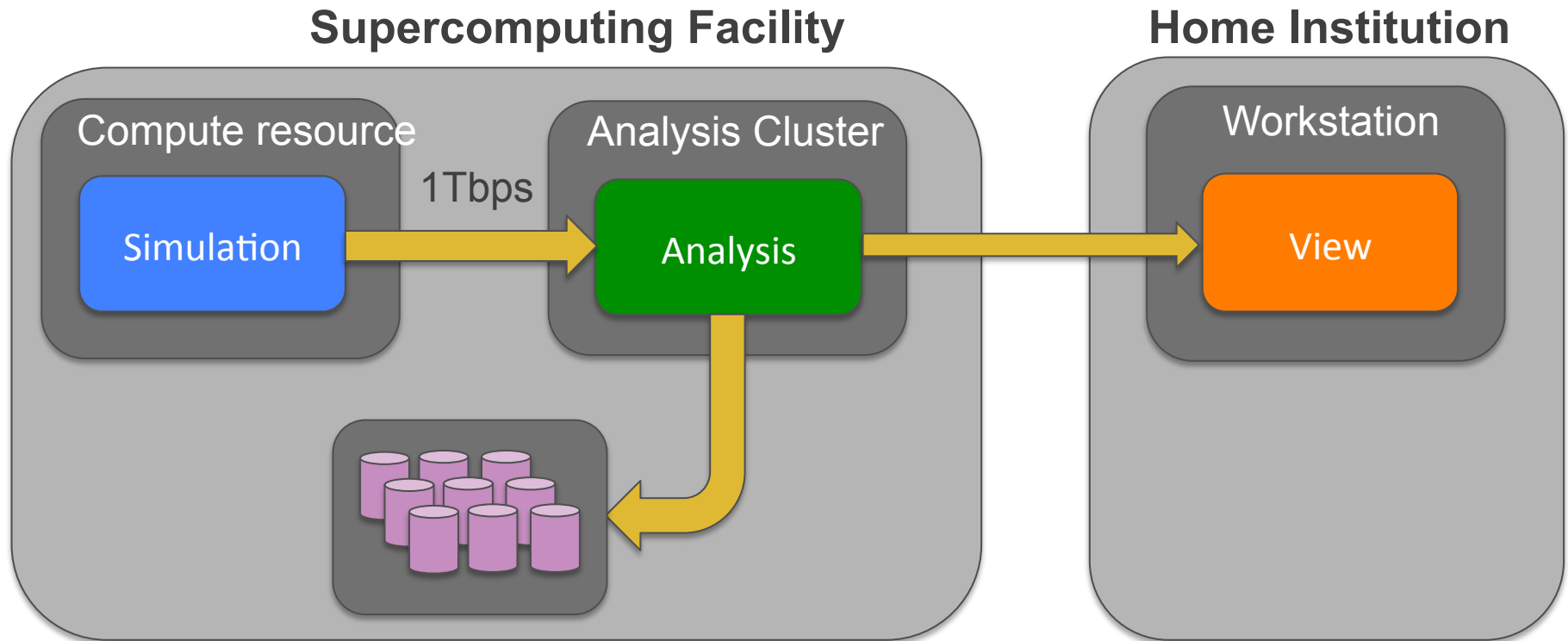
Storage systems are currently unable to cope with extreme scale data sizes in a cost-effective way and this will only get worse in future

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



How do I analyze my data? in situ, co-analysis or post processing?

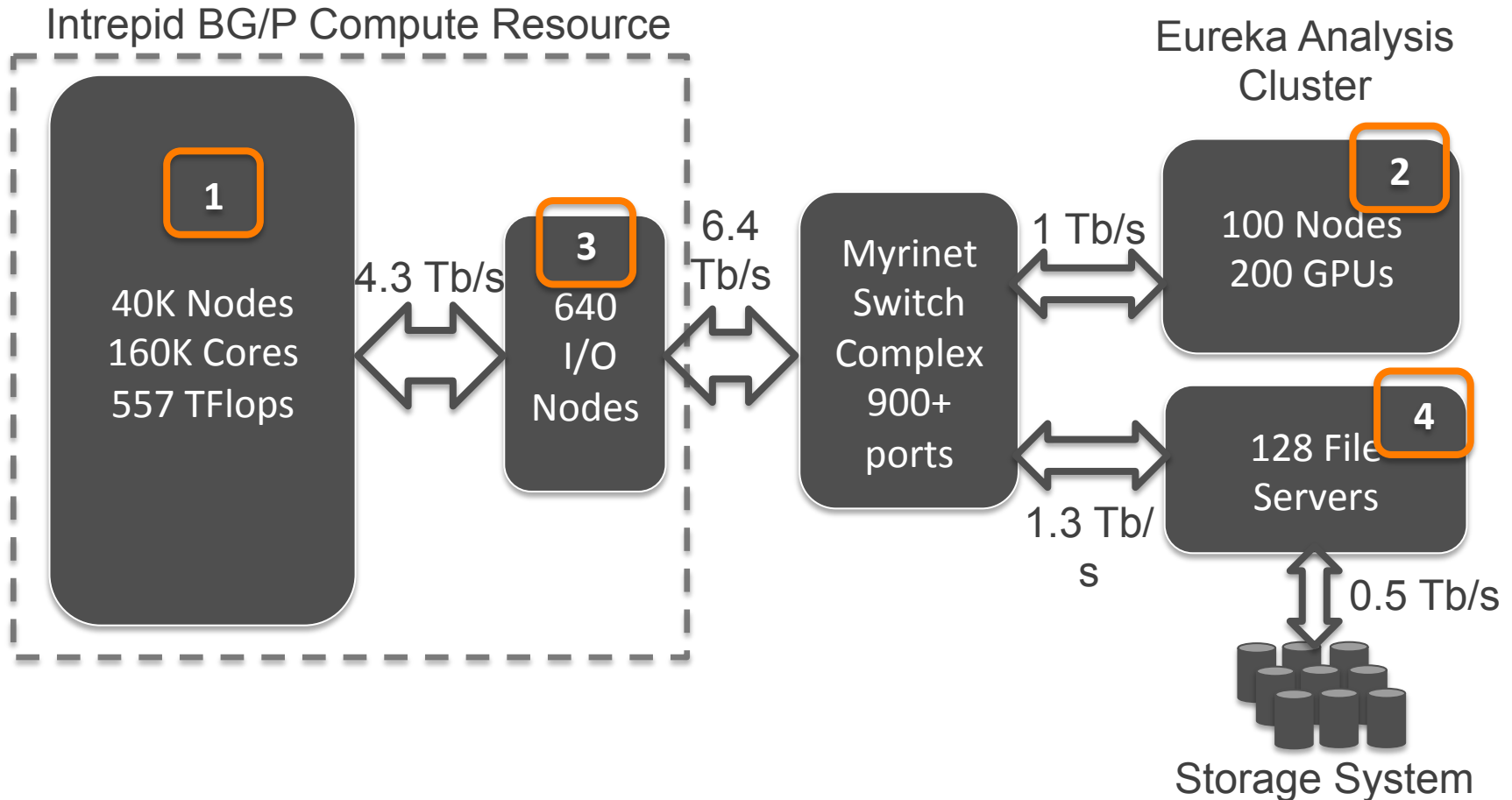
There are multiple answers!!

A solution needs to consider, among others:

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



Argonne Leadership Computing Facility Infrastructure and Analysis Opportunities



The analysis abstractions are applicable to other leadership class infrastructures