Project StarGate
An End-to-End 10Gbps HPC to User Cyberinfrastructure
ANL * Calit2 * LBNL * NICS * ORNL * SDSC

Report to the

Dept. of Energy Advanced Scientific Computing Advisory Committee

Oak Ridge, TN

November 3, 2009

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

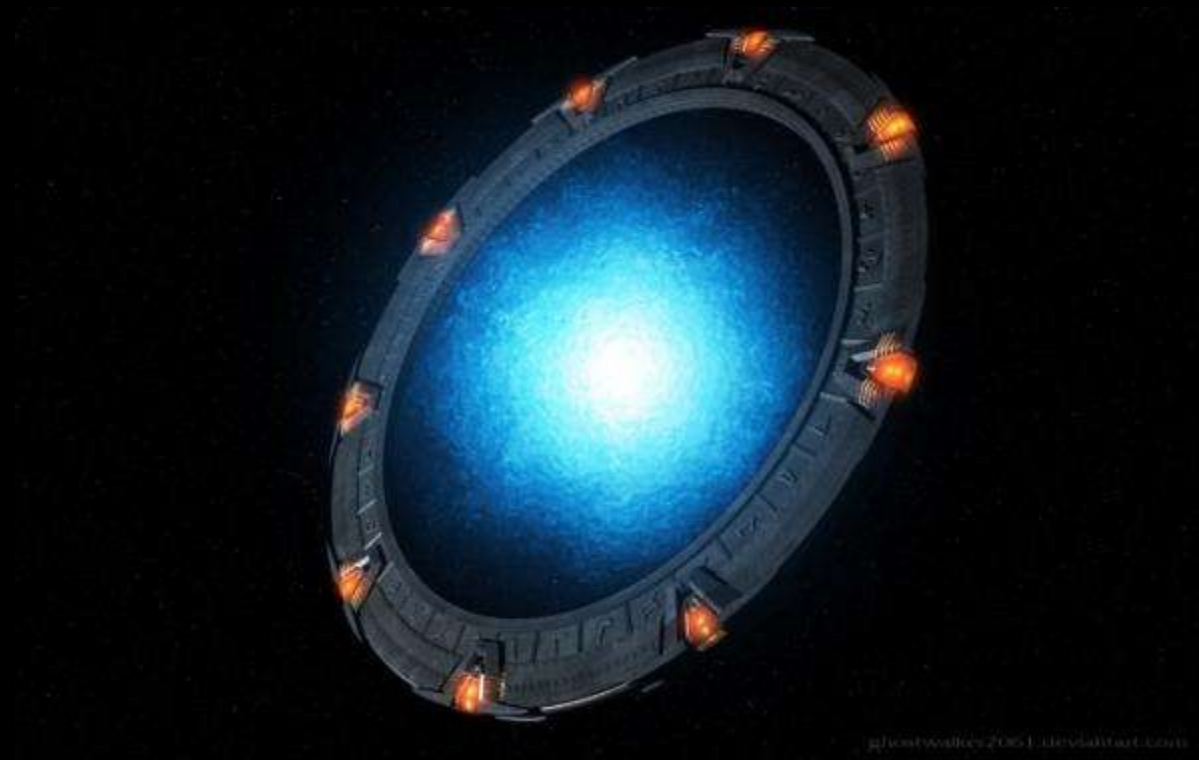
Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

Twitter: lsmarr

Project StarGate



ANL * Calit2 * LBNL * NICS * ORNL * SDSC

Credits

Argonne National Laboratory

Network/Systems

- Linda Winkler
- Loren Jan Wilson

Visualization

- Joseph Insley
- Eric Olsen
- Mark Hereld
- Michael Papka

Lawrence Berkeley National Laboratory (ESnet)

- Eli Dart

National Institute for Computational Sciences

- Nathaniel Mendoza

Oak Ridge National Laboratory

- Susan Hicks

Calit2@UCSD

- Larry Smarr (Overall Concept)
- Brian Dunne (Networking)
- Joe Keefe (OptIPortal)
- Kai Doerr, Falko Kuester (CGLX)

San Diego Supercomputer Center

Science application

- Michael Norman
- Rick Wagner (coordinator)

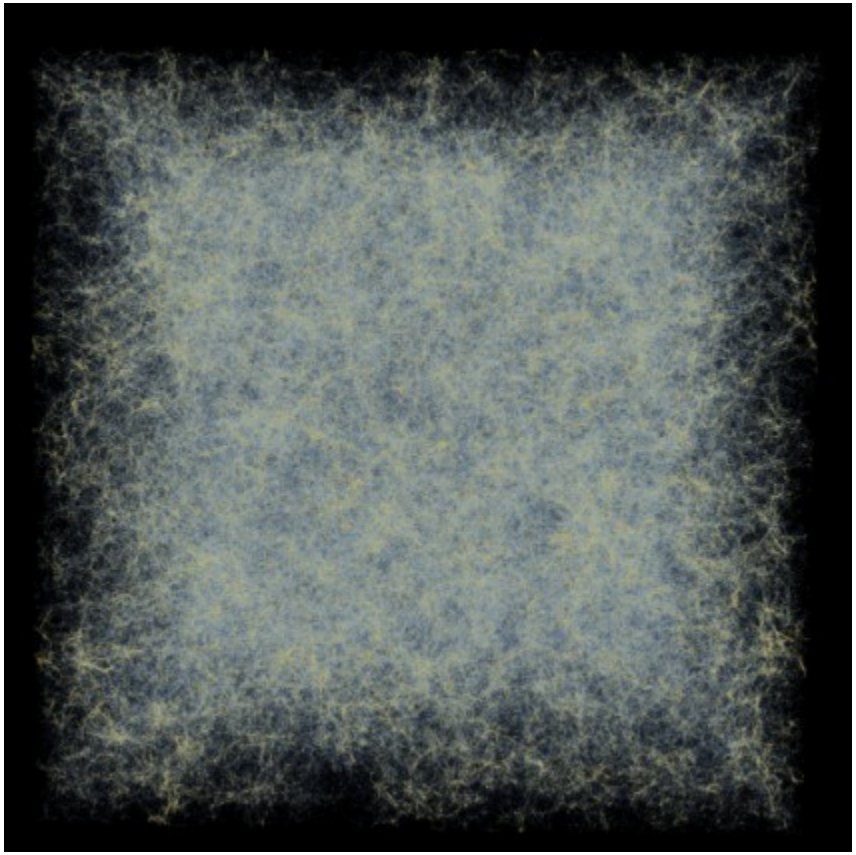
Network

- Tom Hutton

- ANL * Calit2 * LBNL * NICS * ORNL

Exploring Cosmology With Supercomputers, Supernetworks, and Supervisualization

Intergalactic medium on 2 Glyr scale



- 4096^3 particle/cell hydrodynamic cosmology simulation
- NICS Kraken (XT5)
  - 16,384 cores
- Output
  - 148 TB movie output (0.25 TB/file)
  - 80 TB diagnostic dumps (8 TB/file)

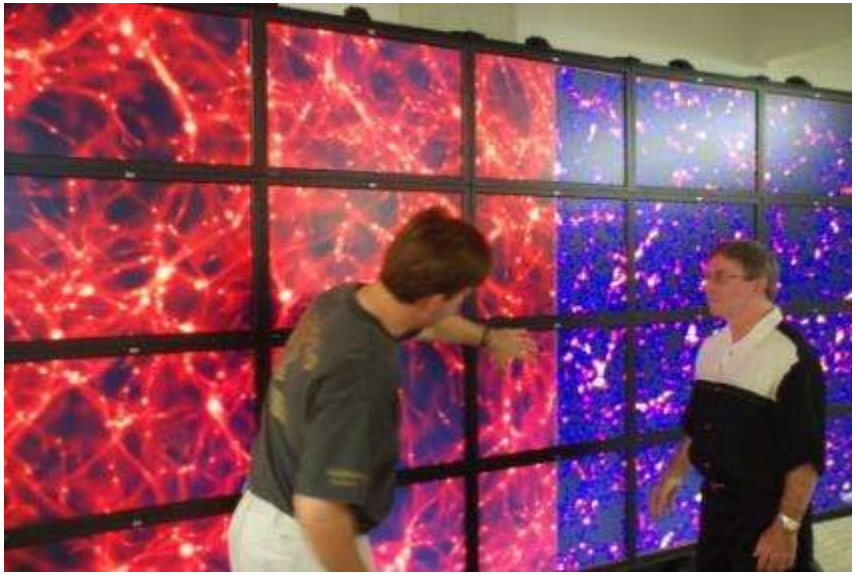
Science: Norman, Harkness,Paschos SDSC
Visualization: Insley, ANL; Wagner SDSC

- ANL * Calit2 * LBNL * NICS * ORNL

Project StarGate Goals

- Explore Use of OptIPortals as Petascale Supercomputer "Scalable Workstations"
- Exploit Dynamic 10 Gbs Circuits on ESnet
- Connect Hardware Resources at ORNL, ANL, SDSC
- Show that Data Need Not be Trapped by the Network "Event Horizon"

OptIPortal@SDSC

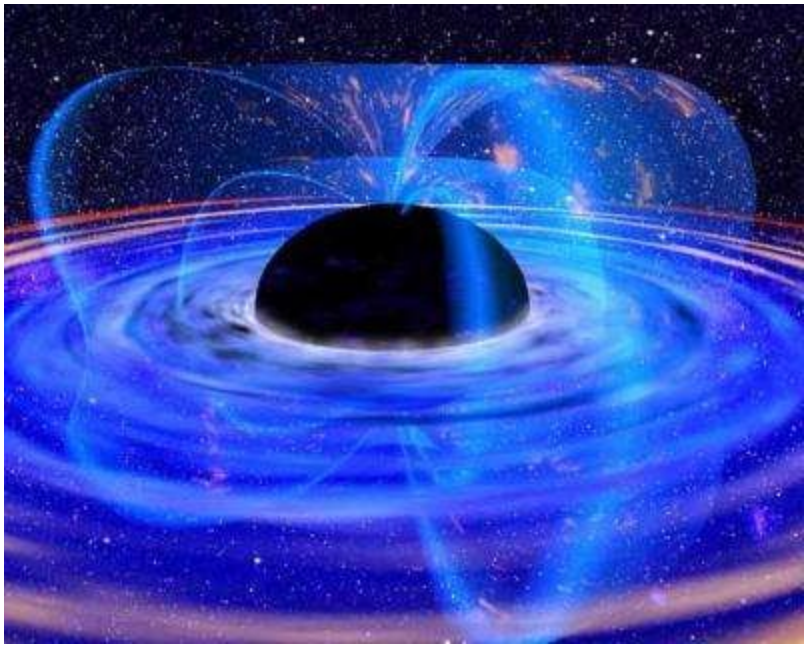


Rick Wagner Mike Norman

- ANL * Calit2 * LBNL * NICS * ORNL

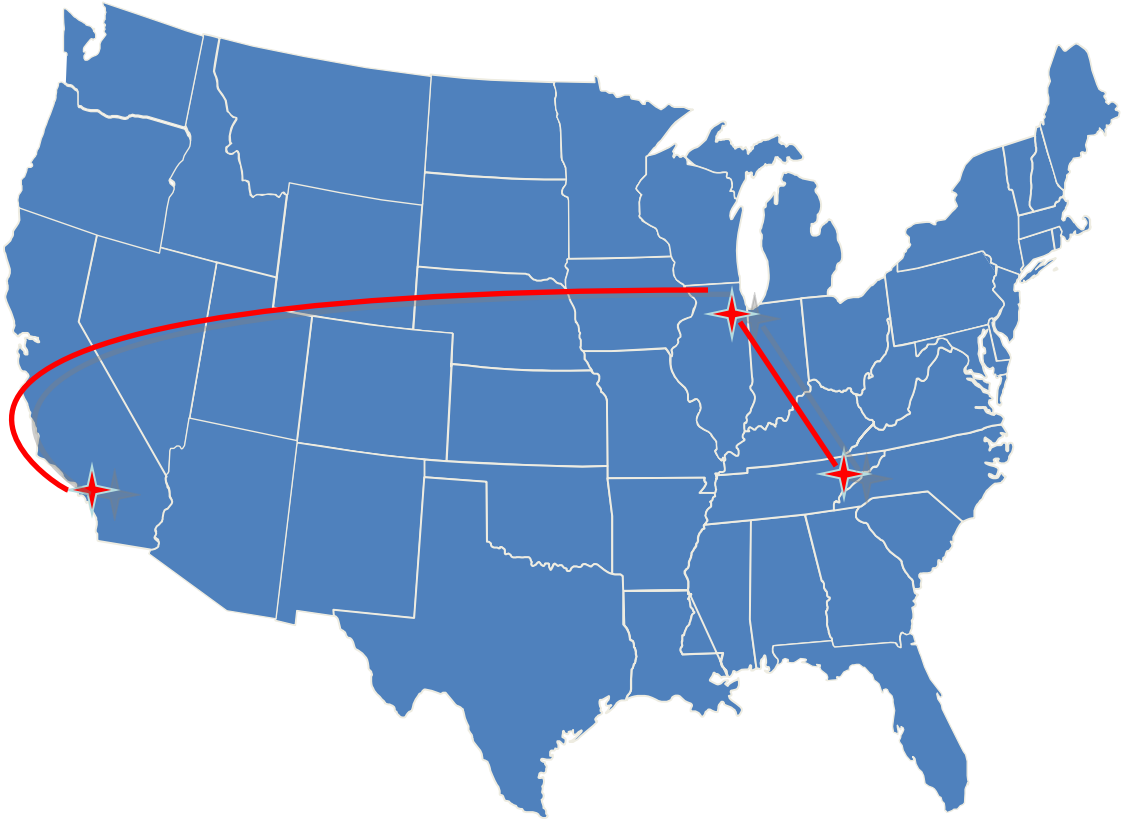
Why Supercomputer Centers Shouldn't Be Data Black Holes or Island Universes

- Results are the Intellectual Property of the Investigator, Not the Center Where it was Computed
- Petascale HPC Machines Not Ideal for Analysis/Viz
- Doesn't Take Advantage of Local CI Resources on Campuses (e.g., Triton) or at other National Facilities (e.g., ANL Eureka)



- ANL * Calit2 * LBNL * NICS * ORNL

Opening Up 10Gbps Data Path
ORNL/NICS to ANL to SDSC



End-to-End Coupling of User with DOE/NSF HPC Facilities

StarGate Network & Hardware

ESnet

Science Data Network (SDN)
> 10 Gb/s fiber optic network
Dynamic VLANs configured using OSCARS

ALCF

DOE Eureka
100 Dual Quad Core Xeon Servers
200 NVIDIA Quadro FX GPUs in 50 Quadro Plex S4 1U enclosures
3.2 TB RAM



SDSC

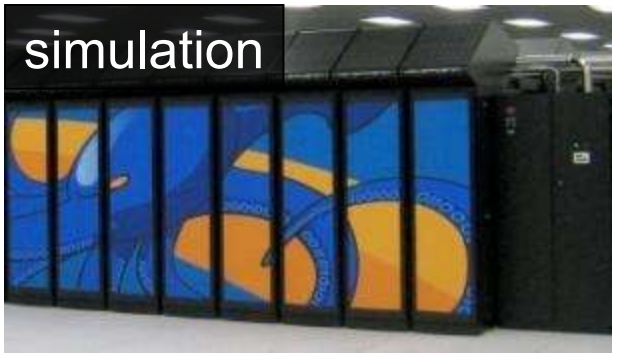


Calit2/SDSC OptIPortal1
20 30” (2560 x 1600 pixel) LCD panels
10 NVIDIA Quadro FX 4600 graphics cards > 80 gigapixels
10 Gb/s network throughout

Challenge: Kraken is not on ESnet

NICS

NSF TeraGrid Kraken
Cray XT5
8,256 Compute Nodes
99,072 Compute Cores
129 TB RAM



- ANL * Calit2 * LBNL * NICS * ORNL

StarGate Streaming Rendering

ESnet

ALCF Internal

3 A media bridge at the border provides secure access to the parallel rendering streams.

gs1.intrepid.alcf.anl.gov

SDSC

ALCF



5 Updated instructions are sent back to the renderer to change views, or load a different dataset.

2

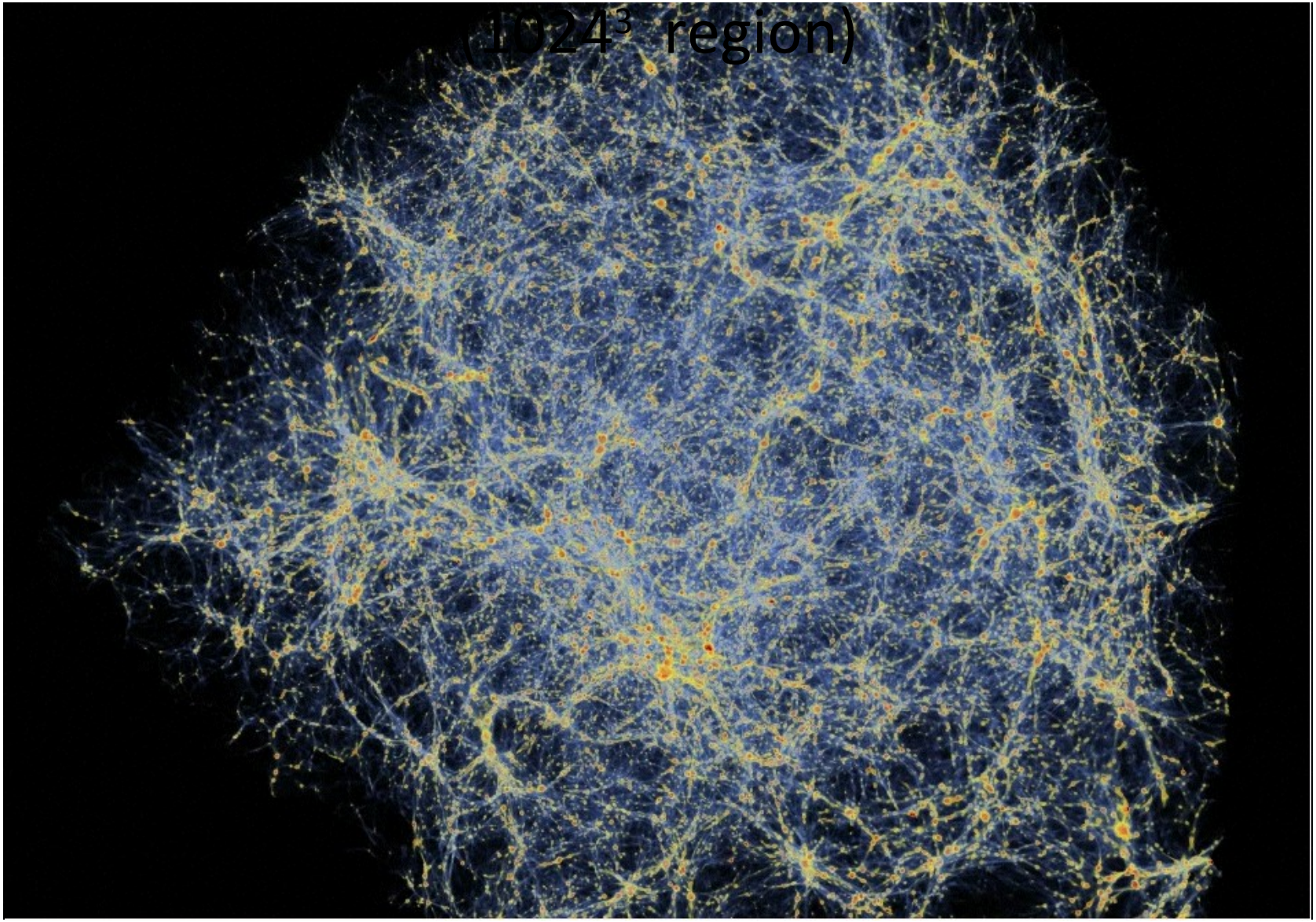
The full image is broken into subsets (tiles). The tiles are continuously encoded as a separate movies.

4 flPy, a parallel (MPI) tiled image/movie viewer composites the individual movies, and synchronizes the movie playback across the OptIPortal rendering nodes.

1 Simulation volume is rendered using vl3 , a parallel (MPI) volume renderer utilizing Eureka's GPUs. The rendering changes views steadily to highlight 3D structure.

- ANL * Calit2 * LBNL * NICS * ORNL

Test animation of 1/64 of the data volume (1024^3 region)



www.mcs.anl.gov/~insley/ENZO/BAO/B4096/enzo-b4096-1024subregion-test.mov

- ANL * Calit2 * LBNL * NICS * ORNL

Data Moved

- ORNL to ANL data transfer nodes
  - 577 time steps ~148TB
  - Peak bandwidth ~2.4Gb/s
    - Disk to disk
    - GridFTP, Multiple Simultaneous Transfers, Each with Multiple TCP Connects
    - Average Aggregate Bandwidth <800mb/s, Using Multiple Transfers
- Additionally
  - Pre-Transfer: Data was Stored in ORNL HPSS, Had to be Staged to Disk on Data Transfer Nodes
  - One Moved to HPSS Partition, Cant Move Data Back
  - Post-Transfer: Each Time Step was a Tar File, Had to Untar
- Moving Forward, will Need Direct High-Bandwidth Path from Kraken (NICS) to Eureka (ALCF)

- ANL * Calit2 * LBNL * NICS * ORNL

ANL Eureka Graphics Cluster

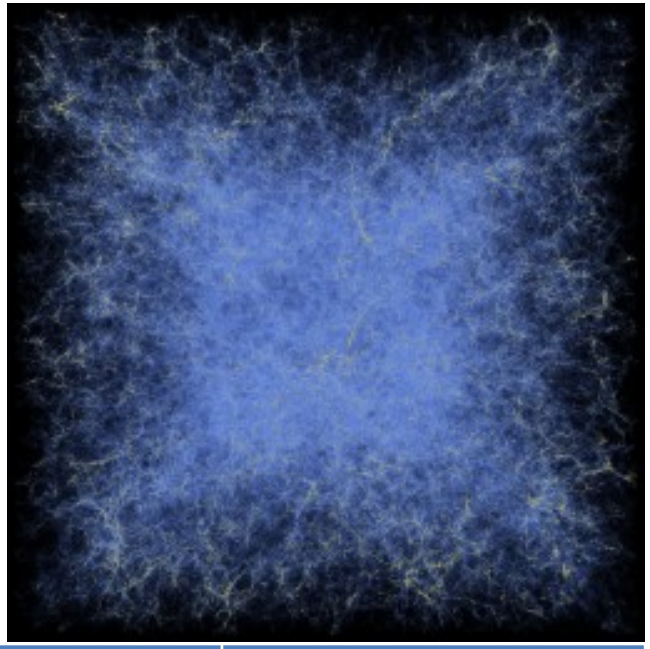
- Data Analytics and Visualization Cluster at ALCF
- (2) Head Nodes, (100) Compute Nodes
  - (2) Nvidia Quadro FX5600 Graphics Cards
  - (2) XEON E5405 2.00 GHz Quad Core Processors
  - 32 GB RAM: (8) 4 Rank, 4GB DIMMS
  - (1) Myricom 10G CX4 NIC
  - (2) 250GB Local Disks; (1) System, (1) Minimal Scratch
  - 32 GFlops per Server

Visualization Pipeline

- vl3 – Hardware Accelerated Volume Rendering Library
  - 4096^3 Volume on 65 Nodes of Eureka
- Enzo Reader can Load from Native HDF5 Format
  - Uniform Grid and AMR, Resampled to Uniform grid
- Locally Run Interactively on Subset of Data
  - On a Local Workstation, 512^3 Subvolume
- Batch for Generating Animations on Eureka
- Working Toward Remote Display and Control

- ANL * Calit2 * LBNL * NICS * ORNL

vl3 Rendering Performance on Eureka



- Image Size: 4096x4096
- Number of Samples: 4096

Data Size	Number of Processors/ Graphics Cards	Load Time	Render/Composite Time
2048^3	17	2min 27sec	9.22 sec
4096^3	129	5min 10sec	4.51 sec
6400^3 (AMR)	129	4min 17sec	13.42sec

Note Data I/O Bottleneck

- ANL * Calit2 * LBNL * NICS * ORNL

Next Experiments

- SC09 - Stream a 4Kx2K Movie From ANL Storage Device to OptIPortable on Show Floor

- Mike Norman is a 2009 INCITE investigator
  - 6 M SU on Jaguar
  - Supersonic MHD Turbulence Simulations for Star Formation
  - Use Similar Data Path for This to Show Replicability

- Can DOE Make This New Mode Available to Other Users?

- ANL * Calit2 * LBNL * NICS * ORNL