Big Data Challenges In Leadership Computing

20 Years of Excellence in Computational Science

OLCF

OAK RIDGE LEADERSHIP COMPUTING FACILITY

1992–2012

Presented to:

Data Direct Network's SC 2011 – Technical Lunch

November 14, 2011

Galen Shipman

Technology Integration Group Leader

U.S. DEPARTMENT OF ENERGY | Office of Science

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Computing at ORNL: Driven by Open Science

Innovative and Novel Computational Impact on Theory and Experiment

- Seeks computationally intensive, large-scale projects to significantly advance science and engineering
- Encourages proposals from universities, other research institutions and industry

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

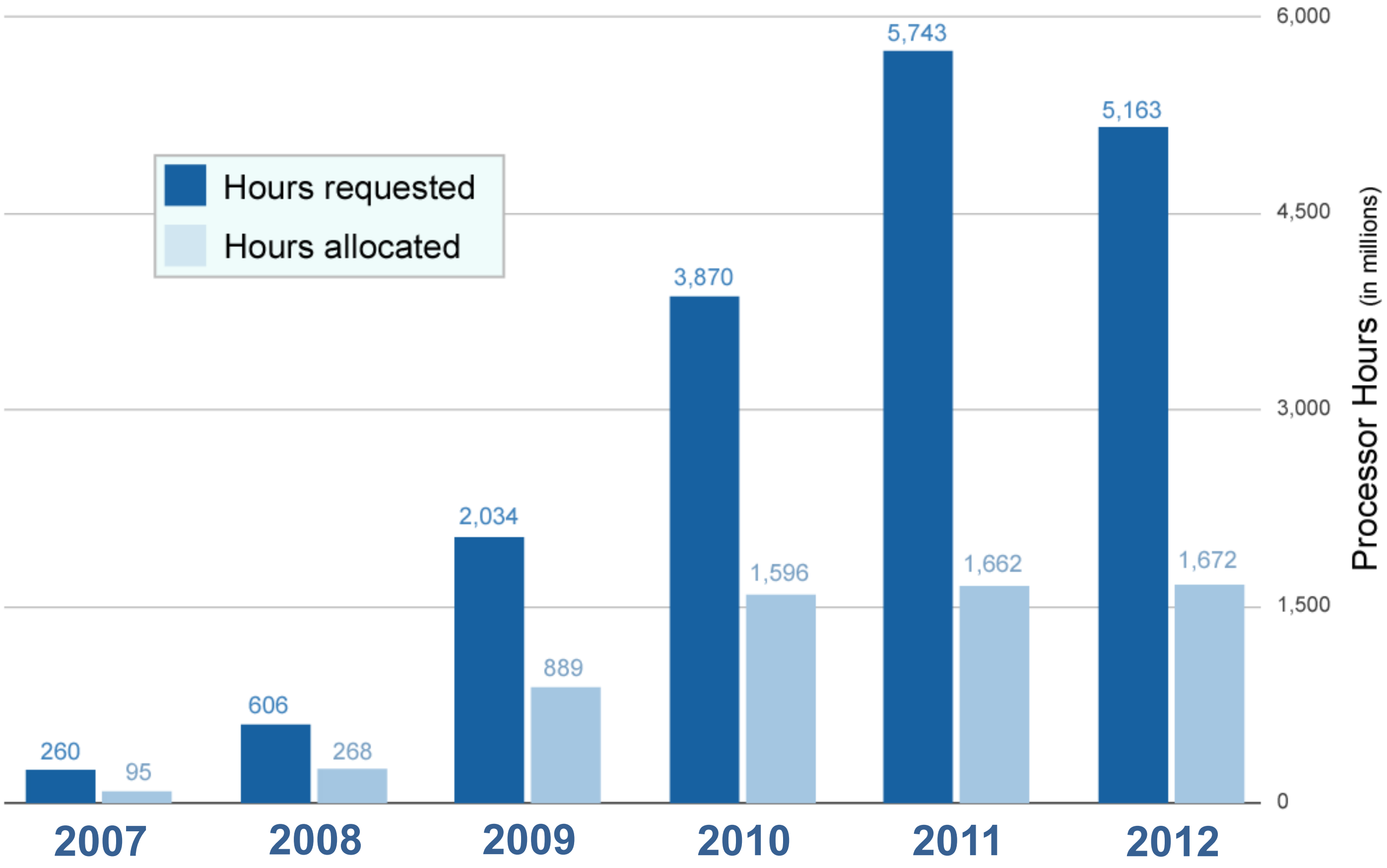
Awards made annually

Peer-review of proposals for impact, computational readiness

Allocations are from 1 to 3 years

Average 10+ million compute hours per year

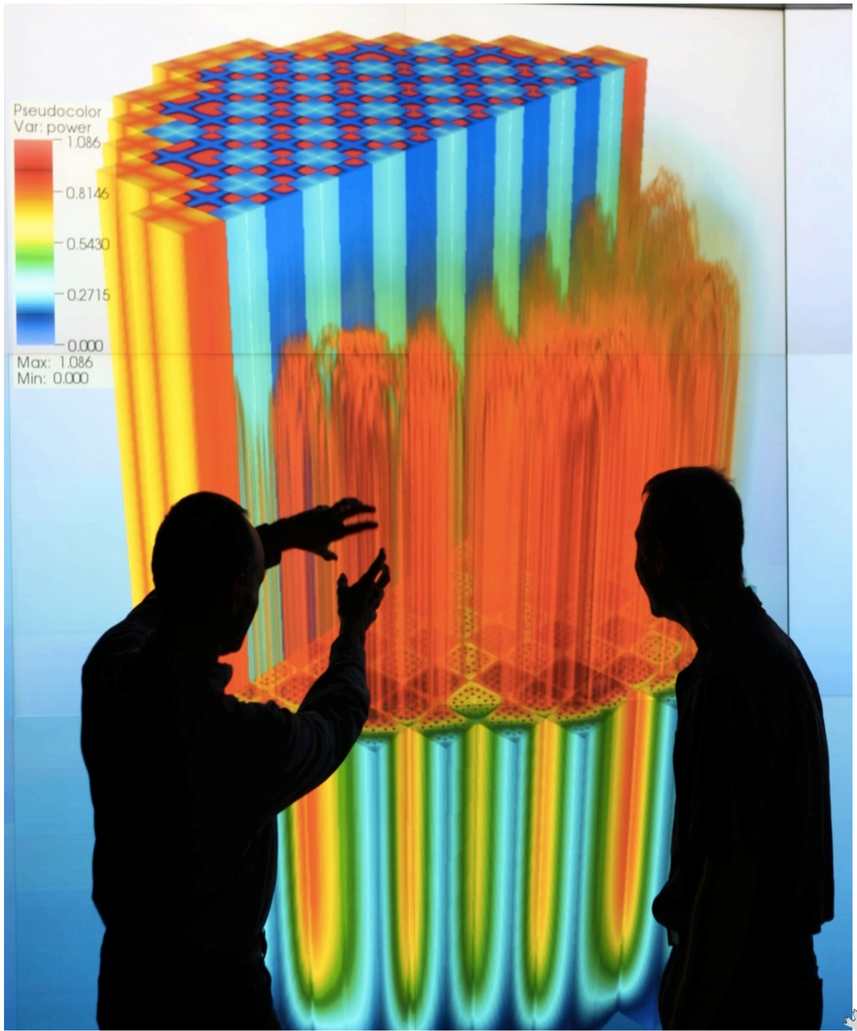
But… INCITE is 2.5 to 3.5 Times Oversubscribed



Meeting this demand for computational resources requires continued investment in leadership computing

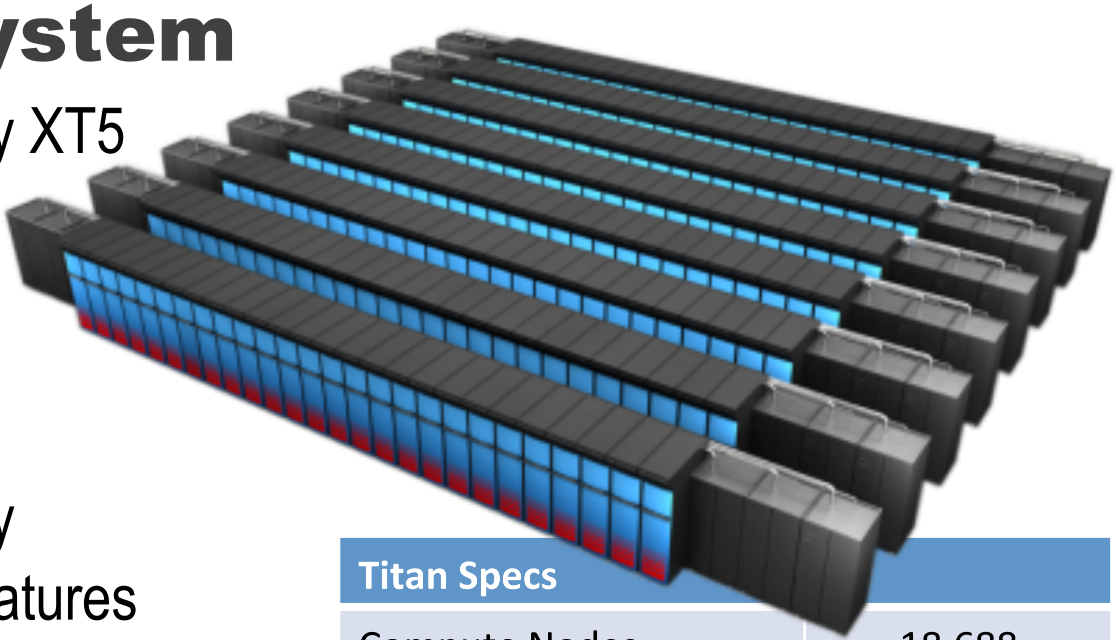
The OLCF-3 Project Will Help Meet The Shortfall in Computational Resources for INCITE

- The next phase of the Leadership Computing Facility program at ORNL
- An upgrade of Jaguar from 2.3 Petaflops today to between 10 and 20 PF by the end of 2012 with operations in 2013
- Built with Cray's newest XK6 compute blades
- When completed, the new system will be called Titan



ORNL's "Titan" System

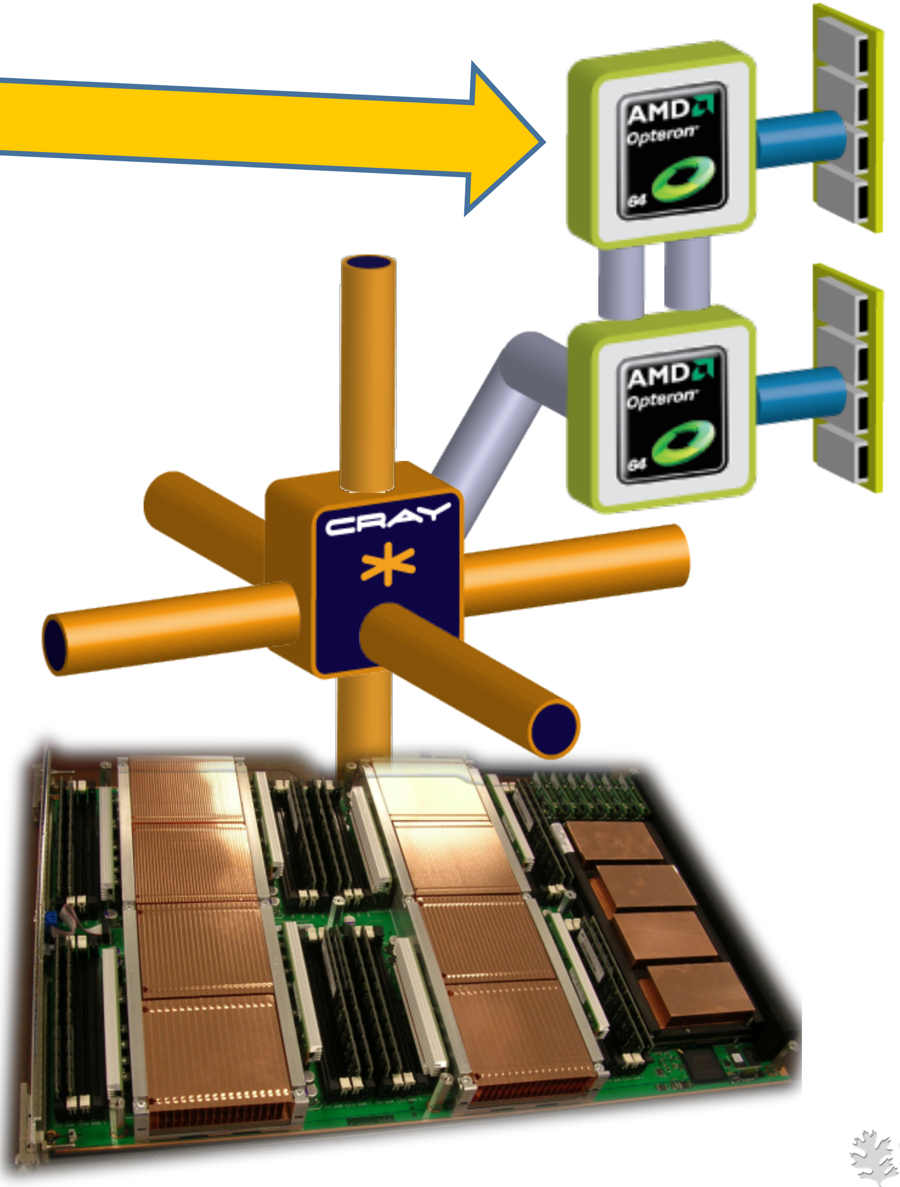
- Upgrade of existing Jaguar Cray XT5
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA M2090 "Fermi" GPUs
  - 2012: 10-20 PF NVIDIA "Kepler" GPUs
- 10-20 PFlops peak performance
  - Performance based on available funds
- 600 TB DDR3 memory (2x that of Jaguar)



Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
NVIDIA "Fermi" (2011)	665 GFlops
# of Fermi chips	960
NVIDIA "Kepler" (2012)	>1 TFlops
Opteron	2.2 GHz
Opteron performance	141 GFlops
Total Opteron Flops	2.6 PFlops
Disk Bandwidth	~ 1 TB/s

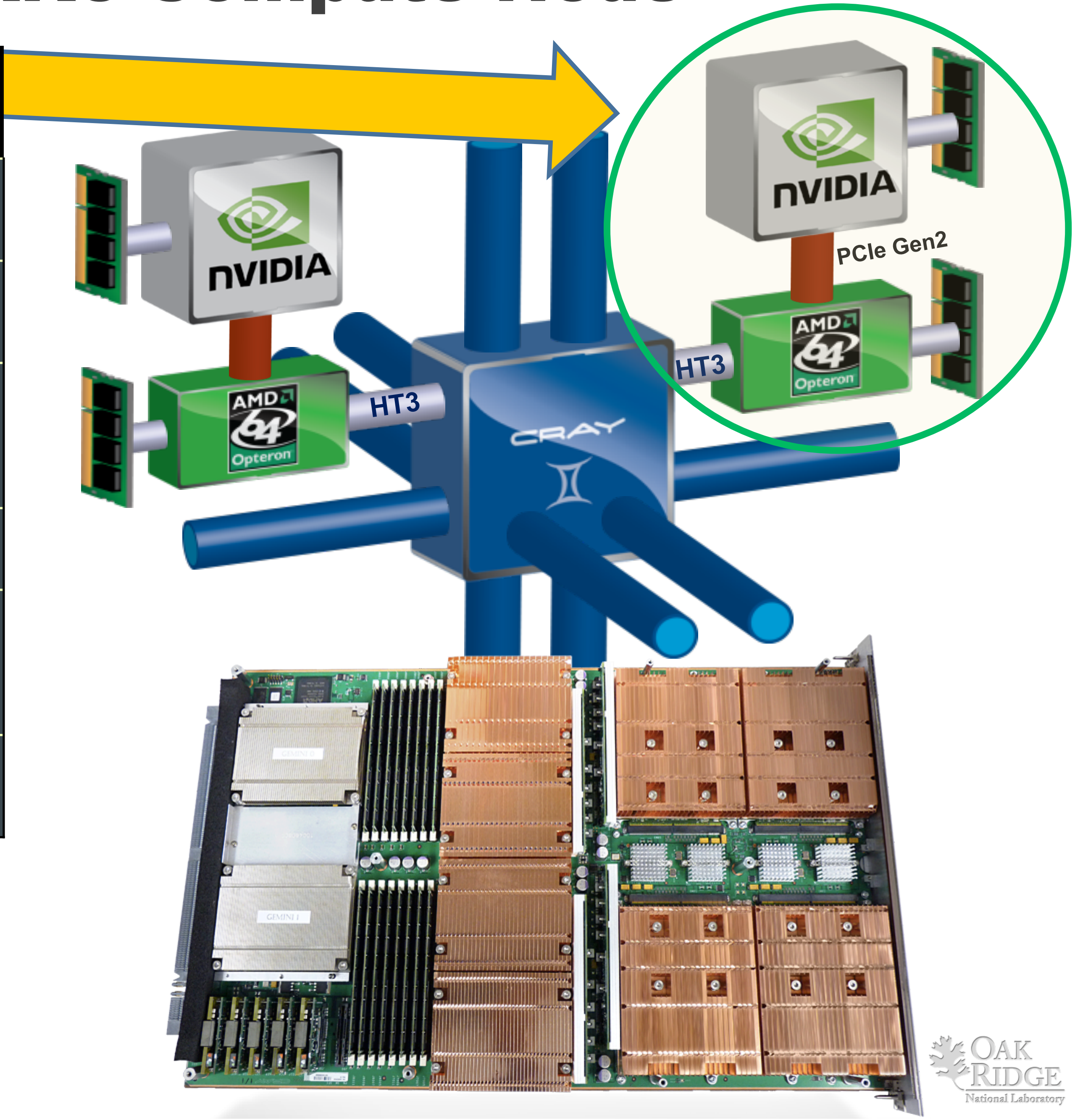
Jaguar’s Cray XT5 Compute Node

XT5 Compute Node Characteristics
Two AMD Opteron Istanbul 6 core processors @ 2.6 GHz
Host Memory 16GB 800 MHz DDR2
SeaStar2+ High Speed Interconnect
Four compute nodes per XT5 blade. 24 blades per rack



Titan's Cray XK6 Compute Node

XK6 Compute Node Characteristics
AMD Opteron 6200 Interlagos 16 core processor @ 2.2GHz
Tesla M2090 @ 665 GF with 6GB GDDR5 memory
Host Memory 32GB 1600 MHz DDR3
Gemini High Speed Interconnect
Upgradeable to NVIDIA's next generation Kepler processor in 2012
Four compute nodes per XK6 blade. 24 blades per rack



Titan XK6 Builds Upon The Proven Cray XT Series of Systems

Highly Integrated Packaging

Provides more compute per rack while maximizing reliability through custom engineered airflow

480 V power

More efficient than converting from 480 V to 208 V to 48 V

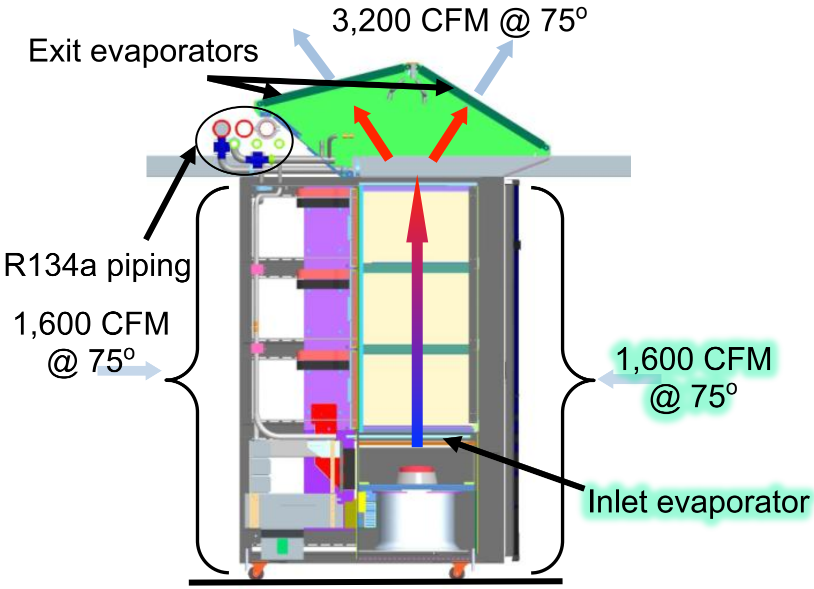
ECOphlex liquid cooling

Liquid-cooled design exhausts heat to R134a before it leaves the cabinet.

Replaces 100 CRAC units!

Saves about 900 KW

of power in air movement alone



What About the File System?

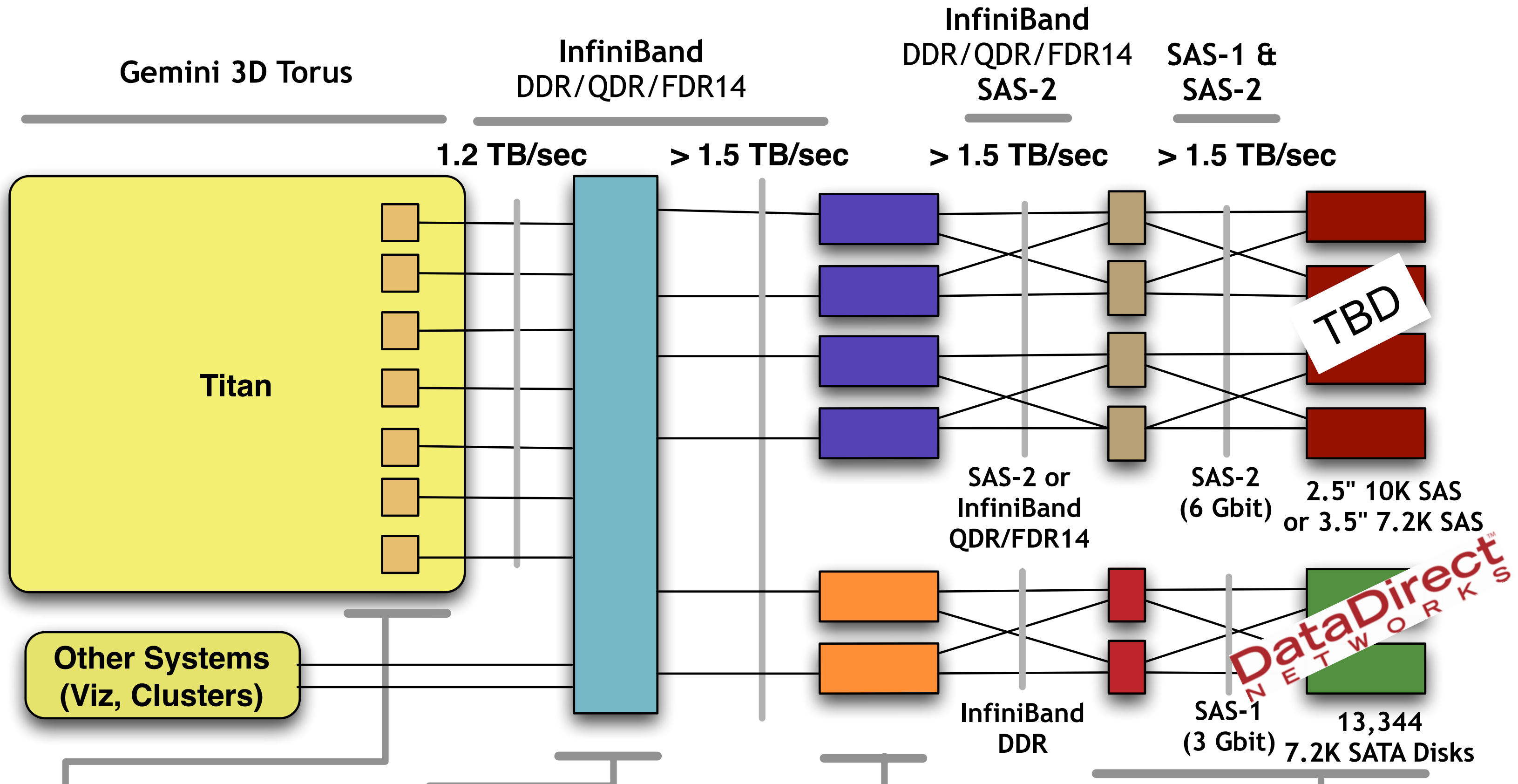
Open Scalable File Systems, Inc.

- We will continue to use Lustre™ as our file system
- Expand our Spider file system infrastructure
  - We expect to increase capacity by 10 – 30 Petabytes
    - (depending on storage technology)
  - We expect to increase bandwidth by up to 1 TB/sec
- Targeting Lustre 2.x
  - Enhanced metadata performance and resiliency under development with Whamcloud (NRE contract)
  - Leveraging OpenSFS activities
    - Community engagement
    - Next-generation feature development
    - Support of the canonical Lustre source tree
- The only file system that meets our requirements today

Lustre Improvements for Titan

- ORNL funded work
  - Improved resiliency through Imperative Recovery
  - Improved metadata performance for large-scale file creation workloads – critical to many of our leadership-class applications
- OpenSFS funded work
  - Single-server metadata performance
  - Distributing metadata among multiple servers
  - Online consistency checking
- Targeting Lustre 2.2 and 2.3 for initial phase of most work
  - Depends on completion before feature-freeze

Spider Phase 2



Lustre Router Nodes
Forward I/O operations
from Compute Clients
to Storage Nodes
384 "XIO nodes"

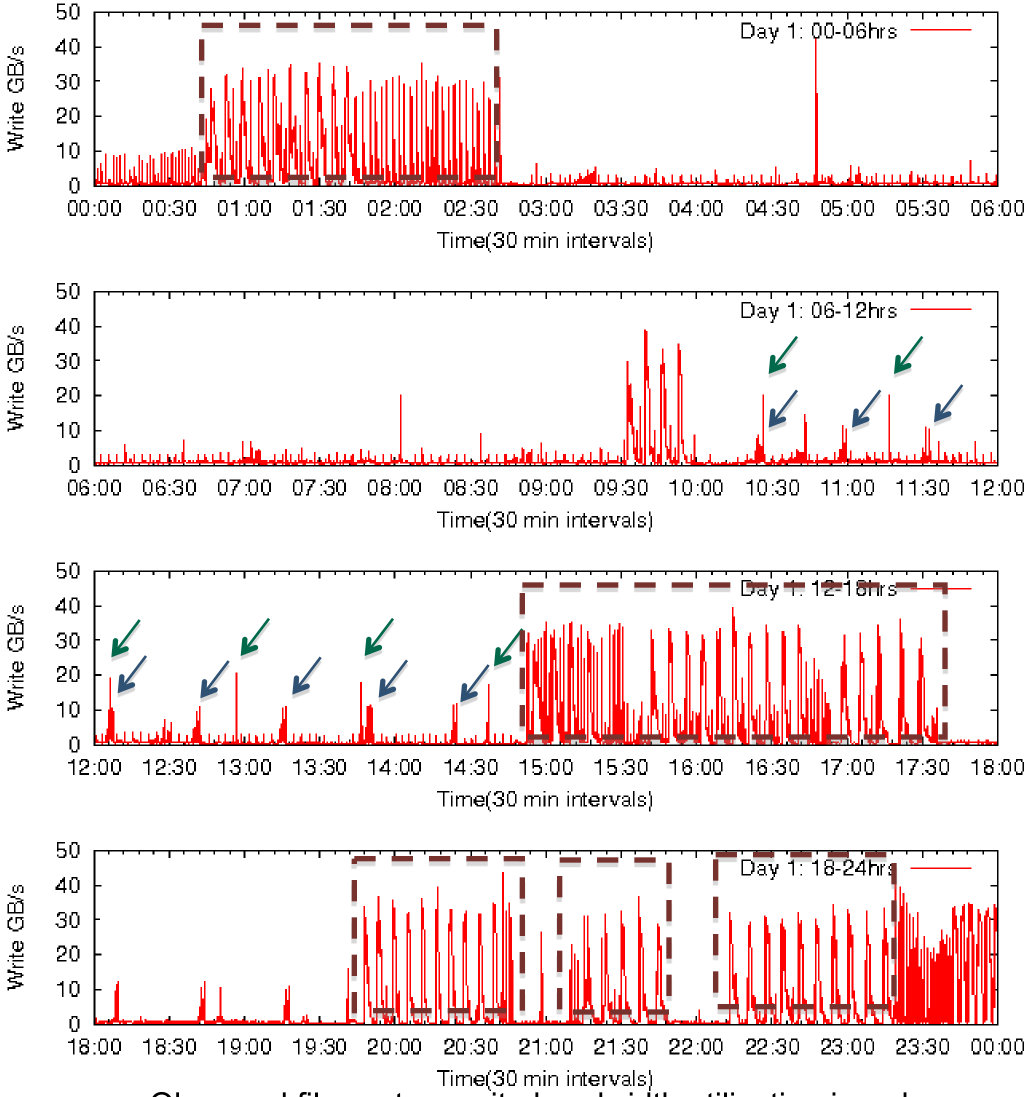
InfiniBand
switch complex
1200+ port 32 or 54 Gbit/sec
3000+ port 20 Gbit/sec
InfiniBand switch
complex

Storage Nodes
run parallel file system
software and manage
incoming FS traffic.

Enterprise Storage
controllers and large
racks of disks are connected
via SAS

Quantifying Total System Performance

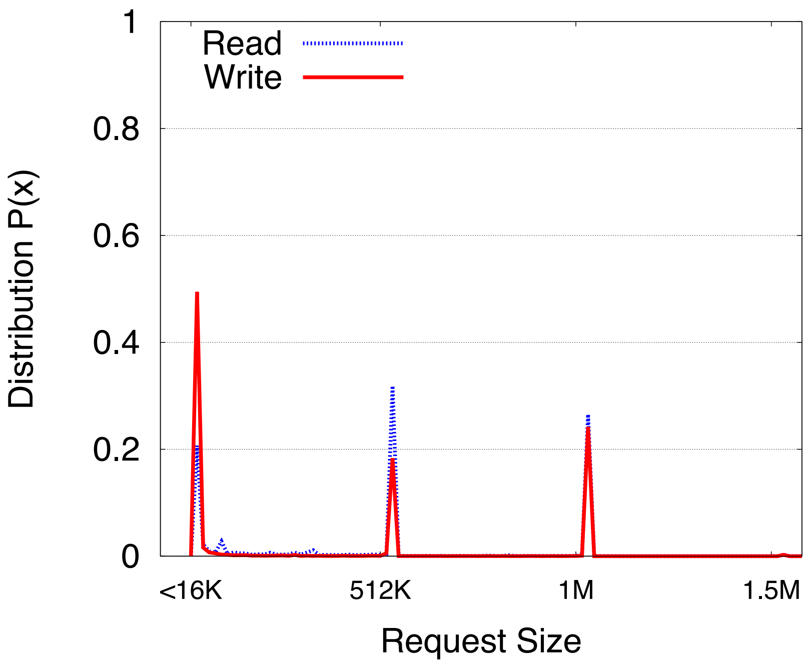
- Drive bandwidth in isolation is a very poor metric and may not translate to accelerated application performance
- Our workloads are are extremely varied from large block sequential I/O to small randomized read/write workloads
- To better understand our workloads we have developed a number of tools for monitoring and analyzing our system
- Based on these results we have developed a fairly extensive benchmarking suite to assess storage system performance
- We expect "system analytics" to play a large role in future system planning, optimization, and a feedback loop for dynamic systems



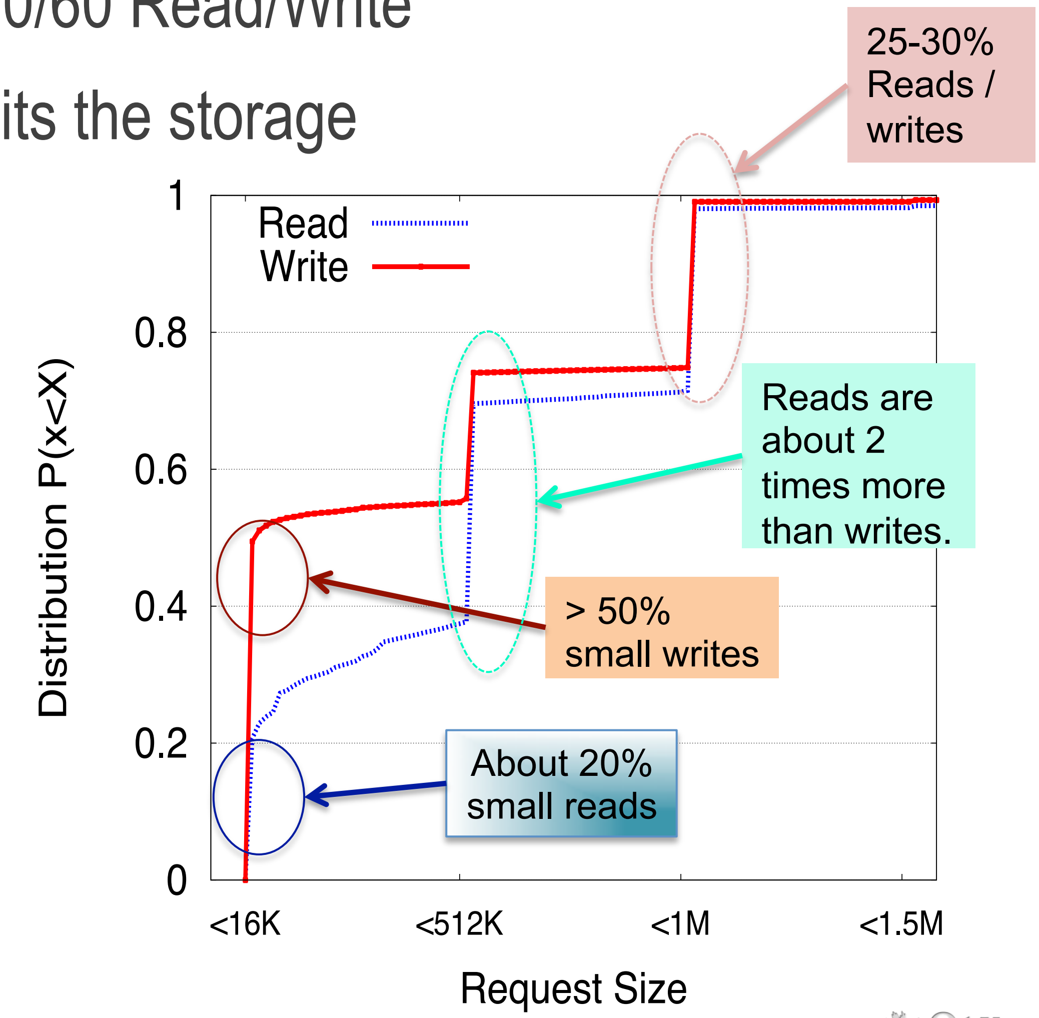
Observed file system write bandwidth utilization in a day

Observation From Spider Phase 1

- Many requests are small (< 16K)
- Workload is approximately 40/60 Read/Write
- Most I/O is random once it hits the storage



Majority of request size (>95%)
- <16KB
- 512KB and 1MB



25-30% Reads / writes

Reads are about 2 times more than writes.

> 50% small writes

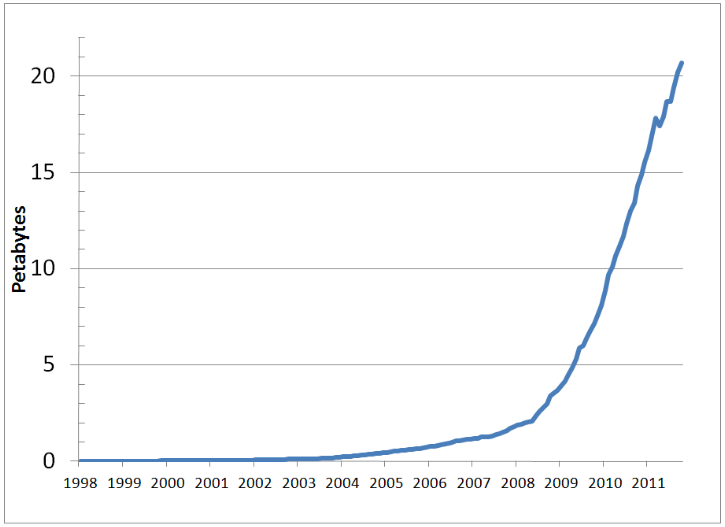
About 20% small reads

Spider Phase 2 – Disk Architecture

- Expect marginal improvements in disk drive performance
  - Latency has flat-lined
  - Bandwidth improvements are modest
  - Considering high-performance SAS rather than SATA to meet our performance targets
    - Dramatically higher IOPs and substantial bandwidth improvements
    - May map better to our expected workloads, providing better realized application performance
    - Capacity requirements are not substantially higher than those of our current system (10PB) and may allow high-performance SAS
- What about flash based storage?
  - Unlikely to deploy a pure flash based storage for Spider Phase 2
  - Integration of flash as an element in the caching hierarchy may provide benefits
    - Provide a burst buffer and may allow further sequentialization prior to destaging

Longer Term Challenges in Big Data

- Tens of thousands of disk drives
- Tens of thousands of tapes
- Over 25 Petabytes of data
- Managing ¼ - ½ billion files is difficult
  - At 200K threads we can generate millions of persistent objects in a single application invocation
  - One user has over 400 TB of data in 8M files
  - One project has over 700 TB of data in 19M files
- Managed with very little information
  - User ID of owner
  - Group ID of owner
  - Total size in bytes
  - Time of last access← current figure of merit!
  - Time of last modification
  - Time of last status change

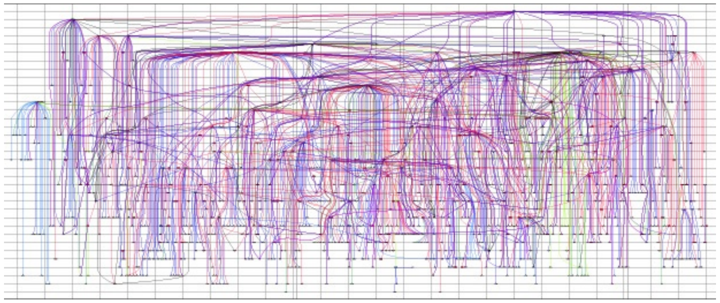


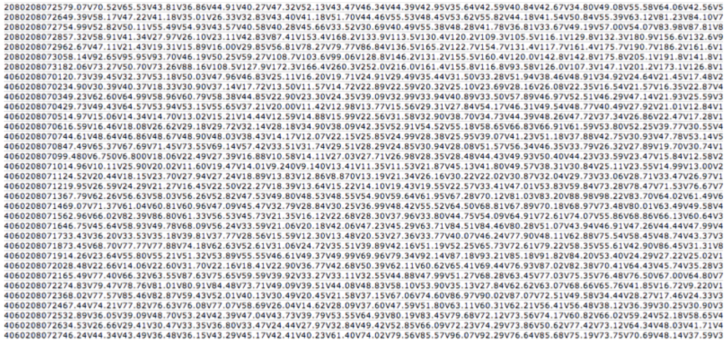
Over 20 Petabyes of data in archival storage alone

Visual Analytics of large-scale ensemble workloads

Exascale Will Exacerbate This

- Managing trillions of objects will be daunting
  - Exascale systems with O(10^9) threads could generate tens of billions of persistent objects in a single application invocation
  - How will data be organized?
    - Blocks of bytes, structured data models, self describing objects (reflection/introspection)
    - Enable automated analysis and data aggregation at the storage level by imparting data structure to the storage system
  - Does a file system namespace even make sense?
    - Companies such as Google manage over 70 PB in a single BigTable instance (2010)
    - Extended attributes may provide flexibility for alternative approaches (tagging, virtual directories)





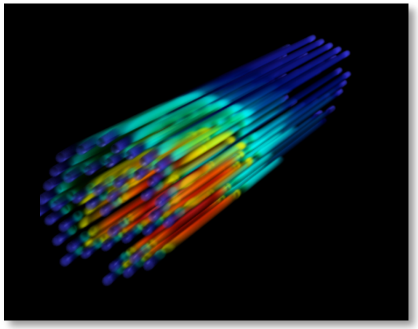
```
2080208072579.07V70.52V65.53V43.81V36.86V44.91V40.27V47.32V52.13V43.47V46.34V44.39V42.95V35.64V42.59V40.84V42.67V34.80V49.08V55.58V64.06V42.56V5
2080208072649.39V58.17V47.22V41.18V35.01V26.33V32.83V43.40V41.18V51.70V44.46V55.53V48.45V53.62V55.82V44.18V41.54V50.84V55.39V63.12V81.23V84.10V7
2080208072754.99V52.82V50.11V55.49V54.93V43.57V40.58V40.28V45.66V33.52V30.69V40.49V55.38V48.28V41.78V36.81V33.67V49.19V57.00V54.07V83.98V87.81V8
2080208072857.32V58.91V41.34V27.97V26.10V23.11V42.83V87.41V153.4V168.2V133.9V113.5V130.4V120.2V109.3V105.5V116.1V129.8V132.3V180.9V156.6V132.6V9
2080208072962.67V47.11V21.43V19.31V15.89V16.00V29.85V56.81V78.27V79.77V86.84V136.5V165.2V122.7V154.7V131.4V117.7V161.4V175.7V190.7V186.2V161.6V1
2080208073058.14V92.65V95.95V93.70V46.19V50.25V59.27V108.7V103.6V99.06V128.8V146.2V131.2V155.5V160.4V120.0V142.8V142.8V175.8V205.1V191.8V141.8V1
2080208073182.06V73.27V50.70V73.26V88.16V108.5V127.9V172.3V166.4V260.3V252.0V216.0V161.4V155.8V116.8V93.58V126.0V107.3V147.1V201.2V173.1V126.8V1
4060208070120.73V39.45V32.37V53.18V50.03V47.96V46.83V25.11V16.20V19.71V24.91V29.49V35.44V31.50V33.28V51.94V38.46V48.91V34.92V24.64V21.45V17.48V2
4060208070234.90V30.39V40.37V18.33V30.90V37.14V17.72V13.50V11.57V14.72V22.89V22.59V20.32V25.10V23.69V28.16V26.08V22.35V16.54V21.57V16.35V22.87V4
4060208070349.23V62.60V64.99V58.96V60.79V58.38V44.85V22.90V23.30V24.35V39.09V32.99V33.94V40.89V33.50V57.89V46.97V52.51V46.29V47.14V21.93V25.59V3
4060208070429.73V49.43V64.57V53.94V53.15V55.65V37.21V20.00V11.42V12.98V13.77V15.56V29.31V27.84V54.17V46.31V49.54V48.77V40.49V27.92V21.01V12.84V1
4060208070514.97V15.06V14.34V14.70V13.02V15.21V14.44V12.59V14.88V15.99V22.56V31.58V32.90V38.70V34.73V44.39V48.26V47.72V37.34V26.86V22.47V17.28V1
4060208070616.59V16.46V18.08V26.62V29.18V29.72V32.14V28.18V34.90V38.09V42.35V52.91V54.52V55.18V58.65V66.83V66.91V61.59V53.80V52.25V39.77V30.55V4
4060208070744.61V48.64V46.86V48.67V48.90V48.03V38.43V14.17V12.07V22.15V25.85V24.99V28.38V25.95V39.07V41.23V51.18V37.88V42.75V30.93V47.78V53.14V5
4060208070847.49V65.37V67.69V71.45V73.55V69.14V57.42V33.51V31.74V29.51V28.29V24.85V30.94V28.08V51.57V56.34V46.35V33.79V26.32V27.89V19.70V30.74V1
406020807099.480V6.750V6.800V18.06V22.49V27.39V16.88V10.58V14.11V27.03V27.71V26.98V28.35V28.48V44.43V49.93V50.40V44.23V33.59V23.47V15.84V12.58V2
4060208071014.96V10.11V25.90V20.02V11.60V19.47V14.01V9.240V9.140V13.41V11.35V11.53V21.87V45.13V41.80V49.57V38.31V30.84V25.11V23.55V14.99V13.00V2
4060208071124.52V20.44V18.15V23.70V27.94V27.24V18.89V13.83V12.86V8.870V13.19V21.34V26.16V30.22V22.02V30.87V32.04V29.73V33.06V28.71V33.47V26.97V1
4060208071219.95V26.59V24.29V21.27V16.45V22.50V22.27V18.39V13.64V15.22V14.10V19.43V19.55V22.57V33.41V47.01V53.83V59.84V73.28V78.47V71.53V76.67V7
4060208071367.79V62.26V56.63V58.03V56.26V52.82V47.53V49.80V48.53V48.55V54.90V59.64V61.95V67.28V70.12V81.03V83.20V88.98V98.22V83.70V64.02V61.49V6
4060208071469.07V71.37V61.04V60.81V60.96V47.09V45.47V32.79V28.84V30.25V36.99V48.42V55.52V64.50V68.81V67.89V70.18V68.97V73.48V80.01V63.49V49.58V4
4060208071562.96V66.02V82.39V86.80V61.33V56.53V45.73V21.35V16.12V22.68V28.30V37.96V33.80V44.75V54.09V64.91V72.61V74.07V55.86V68.86V66.13V60.64V3
4060208071646.75V45.64V58.93V49.78V68.09V56.24V33.59V21.06V20.18V42.06V47.23V45.29V63.71V84.51V84.46V80.28V51.07V43.94V46.91V47.26V44.44V47.99V4
4060208071733.43V36.20V33.53V35.18V39.81V37.77V28.56V15.59V12.30V13.48V20.53V27.36V33.77V40.07V46.24V77.90V48.11V62.88V75.54V58.45V48.74V43.37V3
4060208071873.45V68.70V77.77V77.88V74.18V62.63V52.61V31.06V24.72V35.51V39.89V42.16V51.19V52.25V65.73V72.61V79.22V58.35V55.61V42.90V86.45V31.31V8
4060208071914.26V23.64V55.80V55.21V51.32V53.89V55.55V46.61V49.37V49.99V69.96V79.34V92.14V87.18V93.21V85.18V91.82V84.20V53.40V24.29V27.22V25.02V1
4060208072028.48V22.66V14.06V22.60V31.70V22.16V18.41V22.90V36.77V42.68V50.39V62.11V60.62V65.41V69.44V76.93V87.02V82.38V70.41V64.43V45.74V35.28V3
4060208072165.49V77.40V66.32V63.55V87.63V75.65V59.59V39.92V33.27V33.11V32.55V44.88V47.99V51.27V68.28V63.45V77.03V75.35V76.48V76.50V67.00V64.80V7
4060208072274.83V79.47V78.76V81.01V80.91V84.48V73.71V49.09V39.51V44.08V48.83V58.10V53.90V35.13V27.84V62.62V63.07V68.66V65.76V41.85V16.72V9.220V1
4060208072368.02V77.57V85.46V82.87V59.43V52.01V40.13V30.49V20.45V21.58V37.15V67.06V74.60V86.97V90.02V87.07V72.51V49.58V34.44V28.27V17.46V24.33V3
4060208072467.44V74.21V77.82V76.63V76.08V77.07V58.69V26.04V14.62V28.09V37.60V47.59V51.80V63.11V60.31V62.21V56.41V56.48V38.12V36.39V30.25V30.90V3
4060208072532.89V36.05V39.09V48.70V53.24V42.39V47.04V43.73V39.79V53.55V64.93V80.19V83.45V79.68V72.12V73.56V74.17V60.82V66.02V59.24V52.18V58.65V4
4060208072634.53V26.66V29.41V30.47V33.35V36.80V33.47V24.44V27.97V32.84V49.42V52.85V66.09V72.23V74.29V73.86V50.62V77.42V73.12V64.34V48.03V41.71V4
4060208072746.24V44.34V43.49V36.48V36.15V43.29V45.17V42.41V40.23V61.40V74.02V79.56V85.57V96.07V92.29V76.64V85.68V75.19V73.75V70.69V48.14V37.59V3
```

The POSIX Interface and Metadata

- A proven interface for human interaction
  - Hierarchical directories provide organization
  - Filenames provide a mechanism for identification
    - Augmented with standard attributes
  - But how often do you rely upon “spotlight” over “finder”?
- Widely used to support non-interactive “batch” workloads
  - We often see over 100 thousand files in a single directory
  - Applications may use file naming strategies based on combinations of rank, timestep, variable identifier
  - Often very little information is conveyed in this organization and naming to a human
  - Understanding of this structure is often limited to a single researcher or a small cohesive team

Structured Data in an Unstructured Data Store

- The POSIX write/read/seek model is extremely flexible, supporting any number of data models
- This extreme flexibility often comes at the cost of understandability and performance
  - POSIX is a poor API for scientific data models
  - Limits scalability
- Scientific simulations often rely upon well known data models
  - But… this model is not imparted to the storage system
- Scientific datasets often have complex relationships that are not captured in scientific data models or storage systems
  - Climate land model experiment – land cover forcing – multiple scenarios
  - These datasets may comprise hundreds of thousands of files representing multiple model configurations with individual files spanning time and/or space



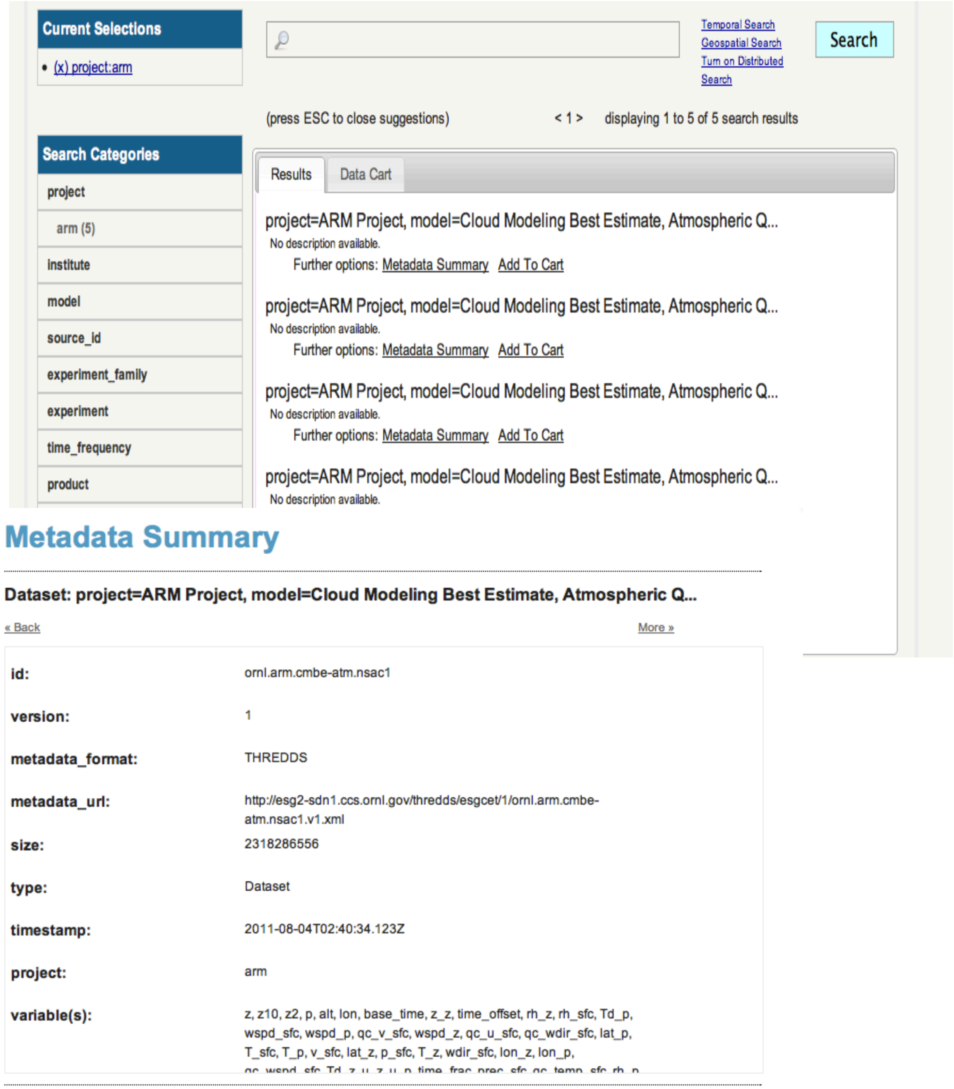
Denovo
Radiation Transport – used in a variety of nuclear energy and technology applications.

How Do We Impart Meaning Using File Systems Today?

- The climate community is an exemplar in data management for simulation data using existing (often antiquated) technologies
- Data Reference Syntax (DRS) and Controlled Vocabularies
  - "atomic datasets" – granules mapped to individual variables representing the entire spatial-temporal domain
  - Variable names are defined by the Climate and Forecast Metadata convention
  - File names encode additional metadata:
    - filename = <variable name>_<MIP table>_<model>_<experiment>_<ensemble member>[_<temporal subset>].nc
  - Atomic datasets are then organized using directory structure
    - <activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable name>/ <ensemble member>/

How Is Data Shared?

- Metadata from climate simulation datasets is then harvested into one or more THREDDS catalogs
- Search and discovery is enabled through Apache SOLR or Sesame RDF
- Data delivery is enabled through GridFTP or Data Mover Light



Current Selections

- (x) project:arm

Temporal Search
Geospatial Search
Turn on Distributed Search

Search

(press ESC to close suggestions)   < 1 >   displaying 1 to 5 of 5 search results

Search Categories

project
arm (5)
institute
model
source_id
experiment_family
experiment
time_frequency
product

Results | Data Cart

project=ARM Project, model=Cloud Modeling Best Estimate, Atmospheric Q...
No description available.
Further options: Metadata Summary Add To Cart

project=ARM Project, model=Cloud Modeling Best Estimate, Atmospheric Q...
No description available.
Further options: Metadata Summary Add To Cart

project=ARM Project, model=Cloud Modeling Best Estimate, Atmospheric Q...
No description available.
Further options: Metadata Summary Add To Cart

project=ARM Project, model=Cloud Modeling Best Estimate, Atmospheric Q...
No description available.

Metadata Summary

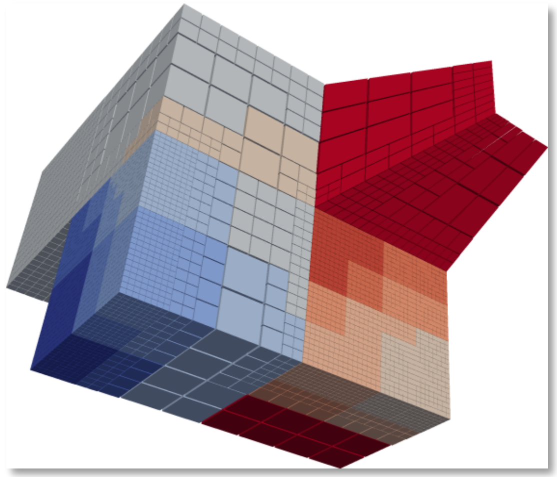
Dataset: project=ARM Project, model=Cloud Modeling Best Estimate, Atmospheric Q...

« Back   More »

id:	ornl.arm.cmbe-atm.nsac1
version:	1
metadata_format:	THREDDS
metadata_url:	http://esg2-sdn1.ccs.ornl.gov/thredds/esgcet/1/ornl.arm.cmbe-atm.nsac1.v1.xml
size:	2318286556
type:	Dataset
timestamp:	2011-08-04T02:40:34.123Z
project:	arm
variable(s):	z, z10, z2, p, alt, lon, base_time, z_z, time_offset, rh_z, rh_sfc, Td_p, wspd_sfc, wspd_p, qc_v_sfc, wspd_z, qc_u_sfc, qc_wdir_sfc, lat_p, T_sfc, T_p, v_sfc, lat_z, p_sfc, T_z, wdir_sfc, lon_z, lon_p, qc_wspd_sfc, Td_z, u_z, u_p, time_frac_prec_sfc, qc_temp_sfc, rh_p

Lots of Work to Impart Meaning in an Unstructured Data Store

- Can we impart structure and relations to better capture metadata directly within the data store? What is needed?
  - Need the ability to model complex relationships between data elements
  - Support for multi-dimensional data and metadata
  - Sparse data support
  - Flexible search capabilities
  - Distributed and parallel
- Exemplars exist: BigTable and Cassandra
  - How can we leverage these blank-sheet of paper approaches in designing a data management system for Science?



Extreme-Scale AMR
Scaling difficult Adaptive Mesh Refinement techniques to over 224,000 cores on **Jaguar** demonstrating excellent scaling.

How to Address These Challenges

- Develop a generalized scalable object store as the foundation
  - Leverage existing technologies where possible
- Research in alternative persistent storage semantics and services
  - Leverage the experience of other big-data communities
    - Ideally we identify common needs and thereby share long-term costs
  - Identify a base level of semantic and services required by simulation and analysis workloads and the scientific data models they use
- Establish partnerships with the vendor community to productize technologies developed

Join Us At The Open

Join Us for the Open Source File Systems BOF – Transitioning from Petascale to Exascale

- Tomorrow (11/15) From 12:15 – 1:15

Questions?

Galen Shipman
Email: gshipman@ORNL.Gov

The research and activities described in this presentation were performed using the resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.
We gratefully acknowledge the support of the Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under contract DE-AC05-00OR22725 with Oak Ridge National Laboratory.