Petascale System Infrastructure

Presented by

Galen M. Shipman

Group Leader, Technology Integration
National Center for Computational Sciences

SC11 Seattle, WA

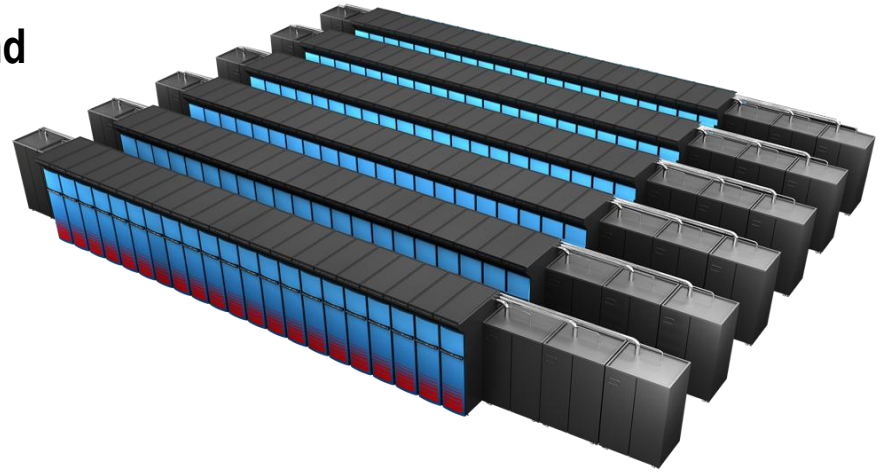
OAK RIDGE National Laboratory

A demanding computational environment

Jaguar XT5 (upgrading to XE6)	18,688 Nodes	224,256 Cores	300+ TB memory	2.3 Pflops
Titan	10-20 PF system – scheduled deployment in 2012			
Frost (SGI Ice)	128 Node institutional cluster			
Smoky	80 Node software development cluster			
Lens	30 Node visualization and analysis cluster			

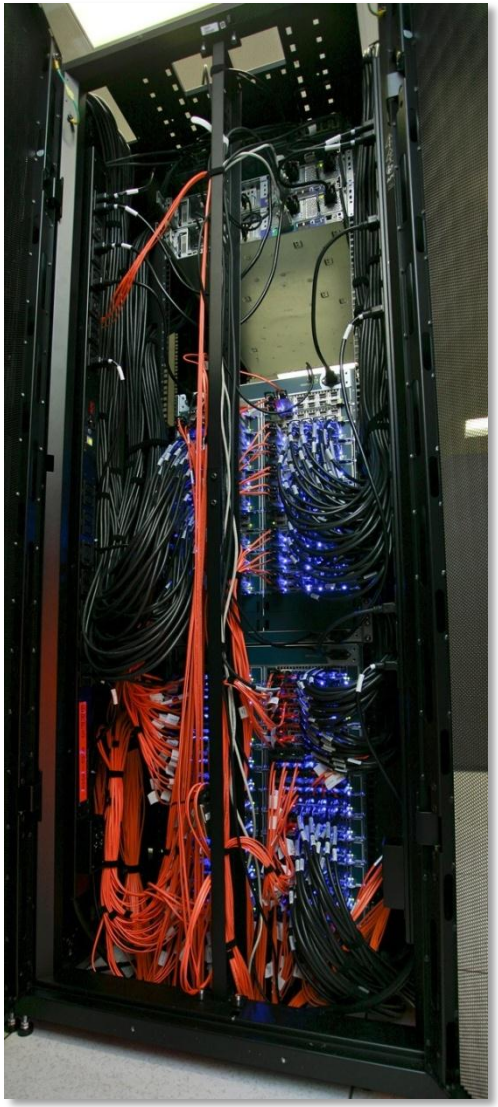
ORNL's "Titan" system goals

- Similar number of cabinets, cabinet design, and cooling as Jaguar
- Operating system upgrade of today's Linux 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
- 10-20 PF peak performance
  - Performance based on available funds
- Larger memory - more than 2x more memory per node than Jaguar



Titan Specs	
Nodes	18,688
Memory per node	32 GB + 6 GB
NVIDIA "Fermi"	665 GFlops
# of Fermi chips	960
NVIDIA Kepler	NDA
Opteron	2.2 GHz
Opteron performance	141 Gflops
Total Opteron Flops	2.6 Pflops

OLCF networking

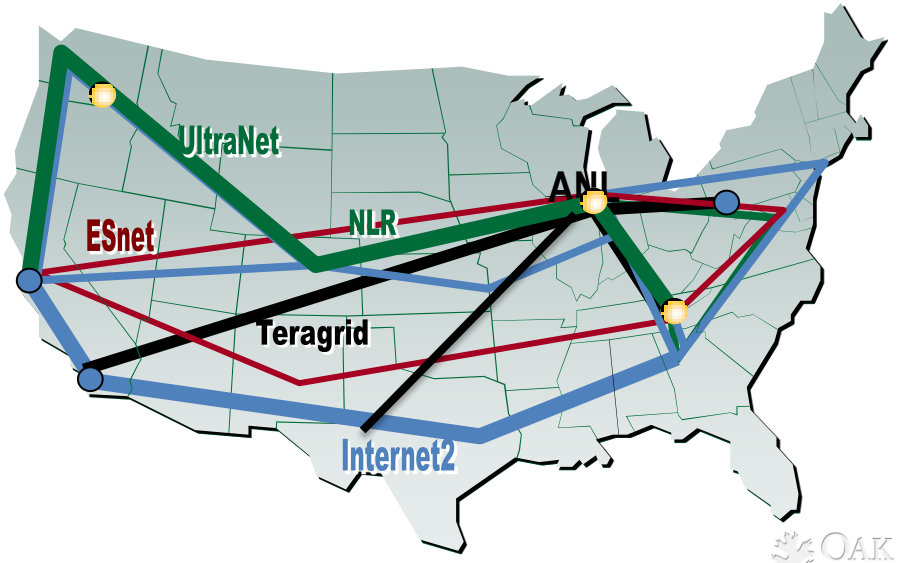


- A service-rich computational environment interconnected via our InfiniBand System Area Network
  - Over 3,000 InfiniBand ports
  - Over 3 miles of cables
  - Scales as computational environment grows
  - An InfiniBand-based network helps meet the bandwidth and scaling needs for the center at reasonable costs
- A robust Ethernet network
  - Ubiquitous connectivity for systems management
  - Over 750 1-GBe and 150 10-GBe ports
- Wide-area networking via ESNet and others

High bandwidth connectivity to NCCS enables efficient remote user access

Connected to Major Science Networks ORNL-owned dark fiber and DWDM equipment linking ORNL to Chicago, Nashville, and Atlanta	
OC192 to ESNET with backup OC48 - Added 10 Gb SDN dynamically reconfigurable (layer 2)	10 Gb to NSF Teragrid
10 Gb to Internet2	2 x 10 Gb UltraScienceNet
4 x 10 Gb to National Lambda Rail	10 Gb Futurenet to NSF Cheetah net

- ORNL participating in the Advanced Networking Initiative (ANI)
  - 100 Gb native optical network in a loop that includes OLCF, ALCF, and other facilities in the northeast as well as a spur from Chicago to the west coast



Spider

One of the Fastest Lustre file systems in the world

Demonstrated bandwidth of 240 GB/s on the center-wide file system

Largest scale Lustre file system in the world

Demonstrated stability and concurrent mounts on all major OLCF systems

- Jaguar XT5
- Opteron Dev Cluster (Spider)
- Visualization Cluster (Lens)
- End-to-end Cluster (Sith)

Over 19,000 clients mounting the file system in production

Over 282,000,000 files

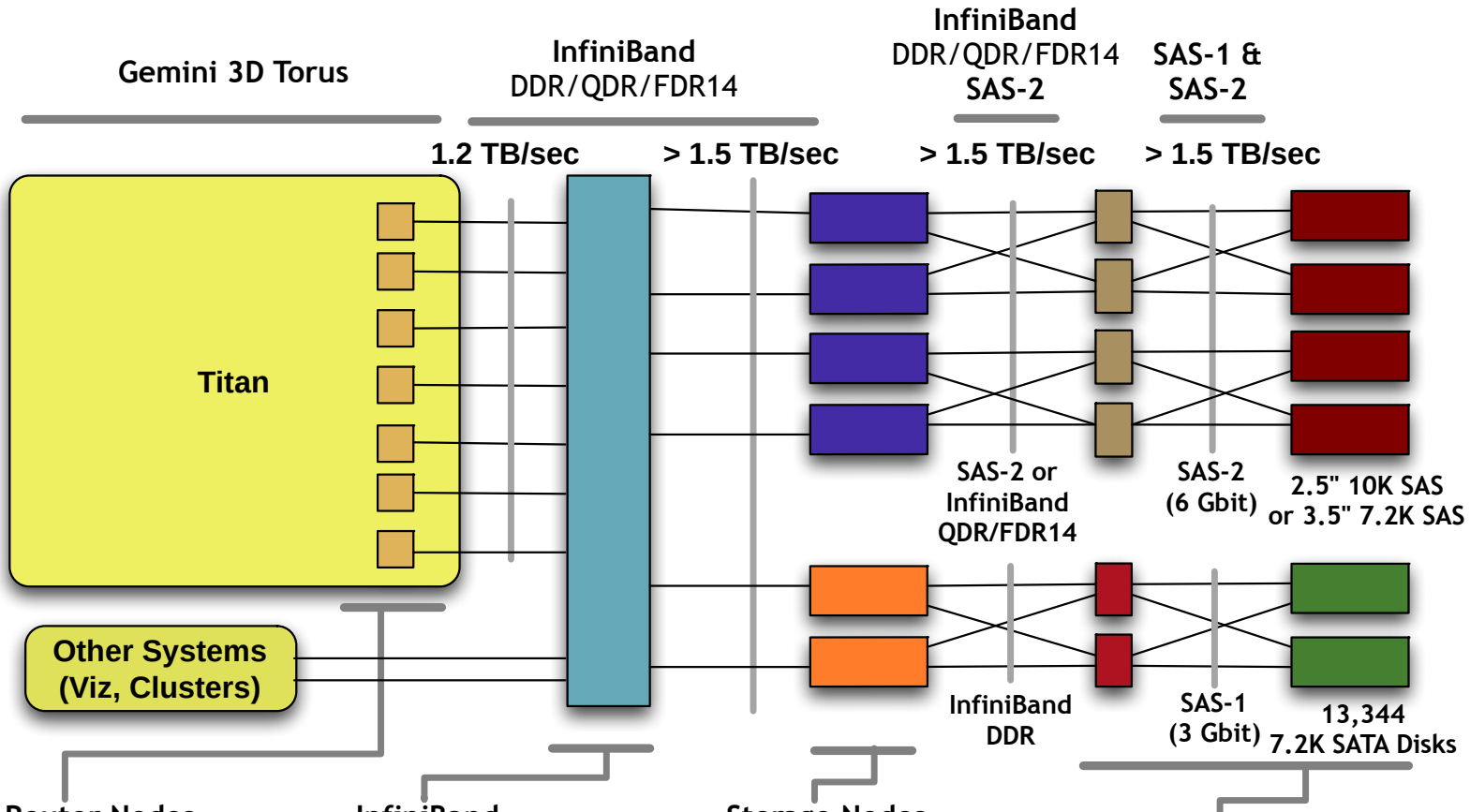
Multiple petabytes of data stored

Cutting edge resiliency at scale

Demonstrated resiliency features on Jaguar

- DM Multipath
- Lustre Router failover

A next generation file system for Titan



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

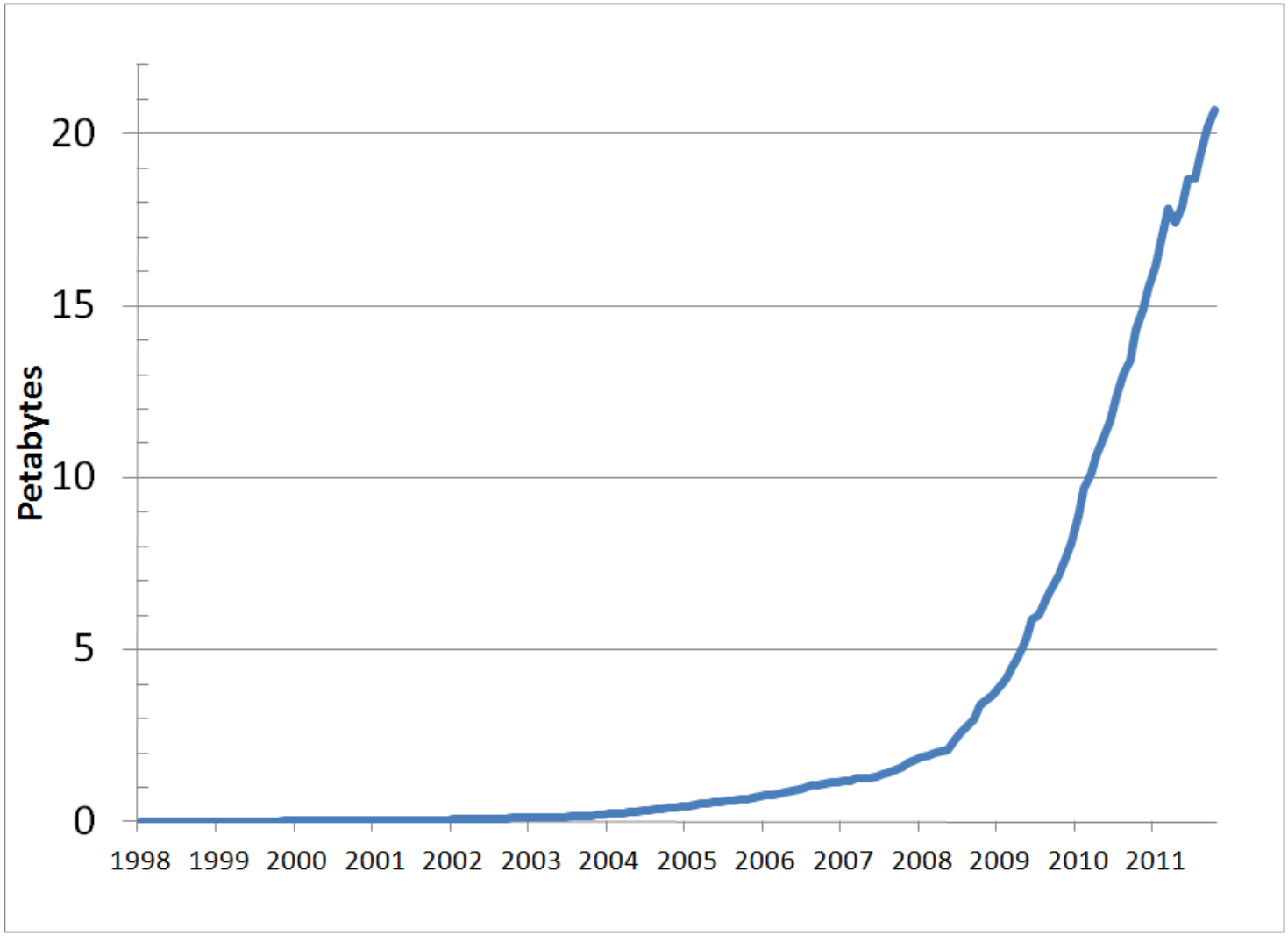
Archival storage infrastructure

HPSS

U.S. DEPARTMENT OF ENERGY | Office of Science

- HPSS provides archival storage for all systems (60 PB capacity; easily expandable)
- HPSS has been upgraded with two additional tape libraries to add additional capacity and bandwidth
- HPSS software has already demonstrated ability to scale to many petabytes
- Capacity and bandwidth on both tapes and disks are scaled to maintain a balanced system
- Utilize new methods to improve data transfer speeds between parallel file systems and archival system (transfer agent, LFM)

HPSS – Managing Exponential Growth in Storage



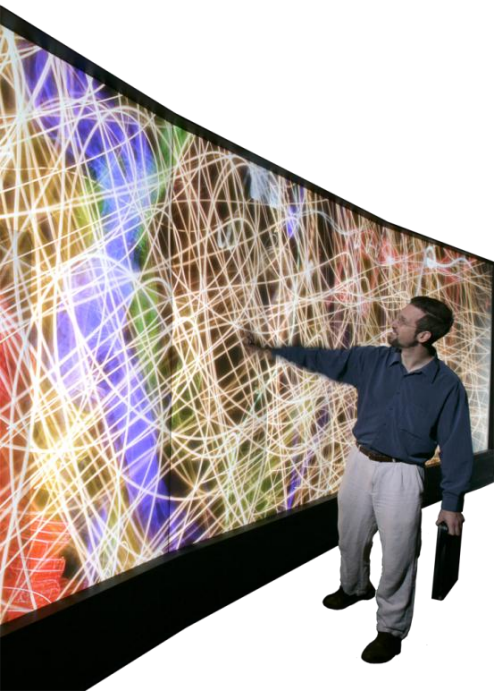
HPSS Growth:	
PB	Months
1	101.5
2	20.3
3	6.2
4	4.3
5	3.1
6	2.6
7	2.8
8	2.1
9	1.4
10	1.4
11	2.1
12	1.8
13	1.5
14	1.7
15	1.5
16	1.5
17	1.4
18	3.2
19	2.4
20	1.3

HPSS is managing 20+ PB and growing at more than 30 TB per day.

Visualization and data analysis resources

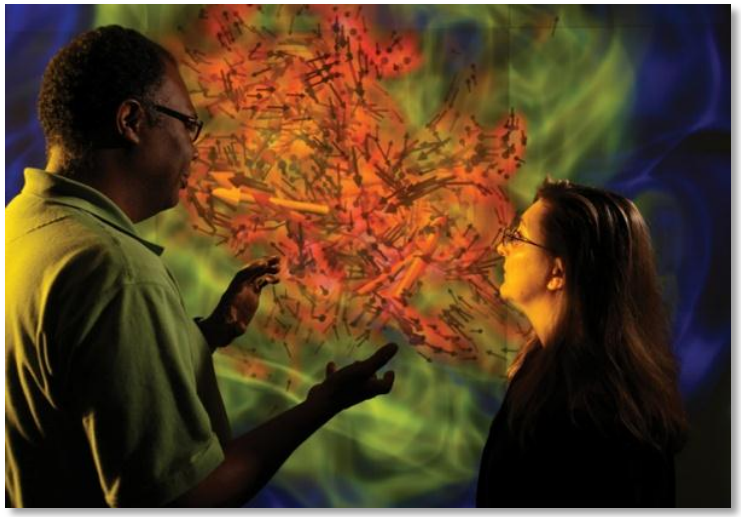
Hardware

- Everest Powerwall
  - 30 ft by 8 ft 35 megapixel display wall
- Lens Cluster
  - 32 nodes each with 64 GB and four quad-core Opterons w/ GPUs
- Everest Cluster
  - 18 nodes to drive display wall



Software

- VisIT
- EnSight Gold and DR
- ParaView
- AVS/Express
- R MPI
- IDL
- SCIRun
- Xmgrace, Gnuplot, Kepler



Contact

Galen Shipman

Technology Integration
National Center for Computational Sciences
(865) 576-2672
gshipman@ornl.gov