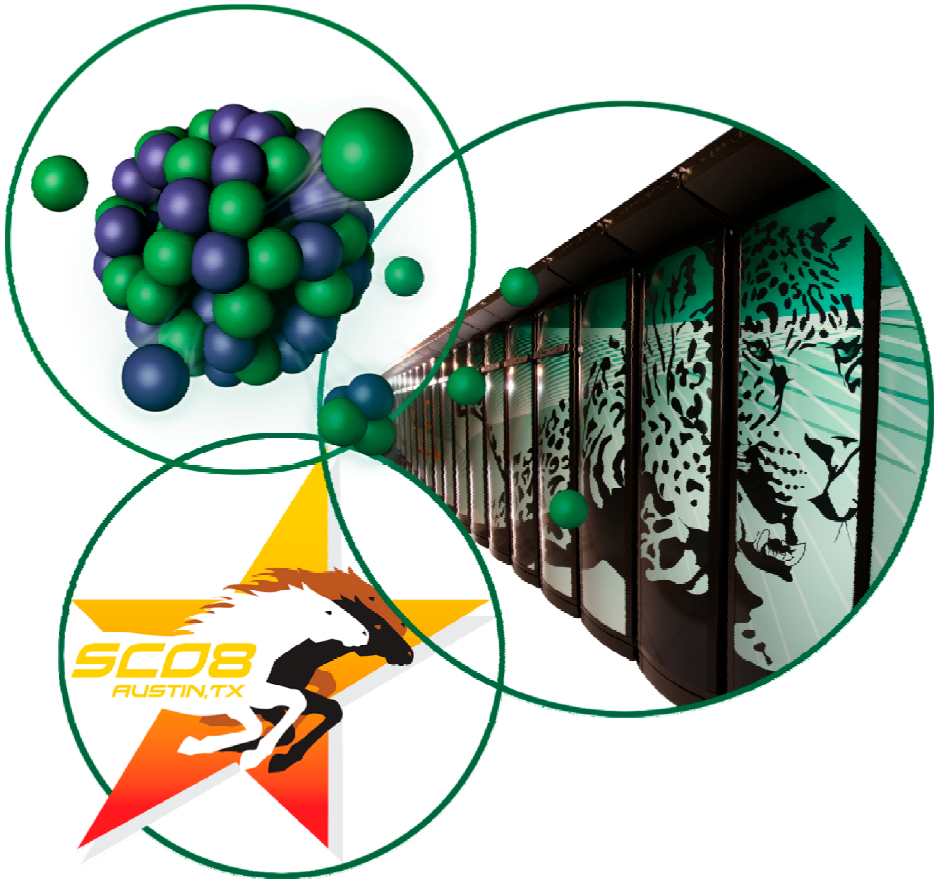
Infrastructure



Presented by

Galen M. Shipman

Technology Integration
National Center for Computational Sciences

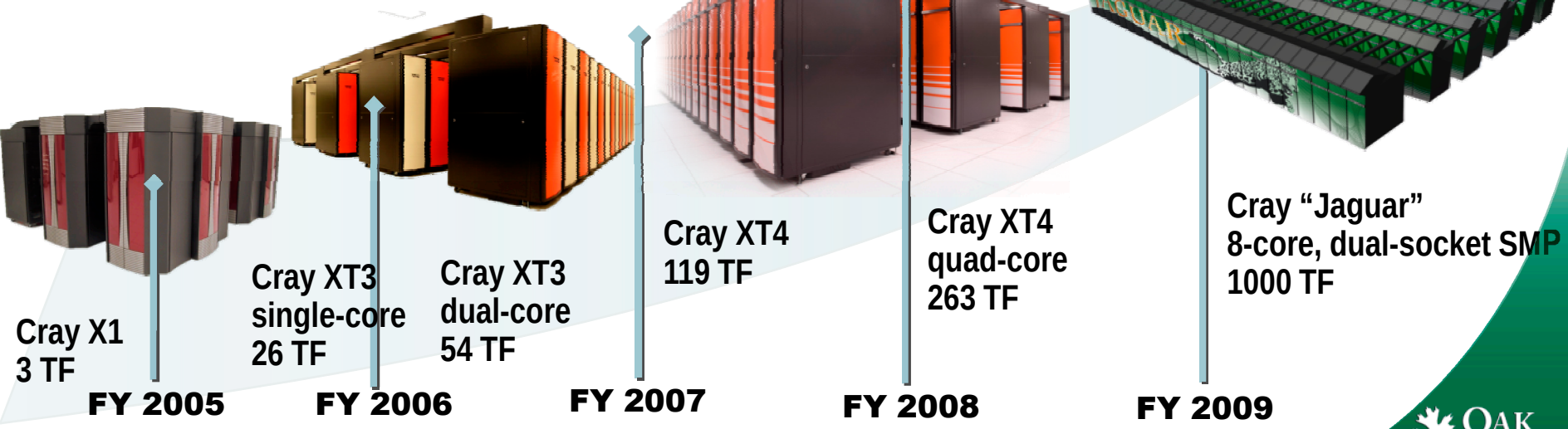
Summary

- LCF roadmap
- Infrastructure for the petascale
  - Networking
  - File systems
  - Archival storage
  - Data analytics

We have increased system performance 300× times since 2004

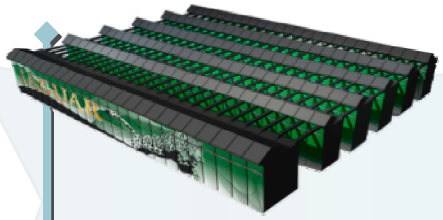
Hardware scaled from single-core through dual-core to quad-core and dual-socket SMP nodes	Scaling applications and system software is the biggest challenge

• NNSA and DoD have funded much of the basic system architecture research
  • Cray XT based on Sandia Red Storm
  • IBM BG designed with Livermore
  • Cray X1 designed in collaboration with DoD
	• SciDAC program is funding scalable application work that has advanced many science applications
	• DOE-SC and NSF have funded much of the library and applied math as well as tools
	• Computational liaisons are key to using deployed systems

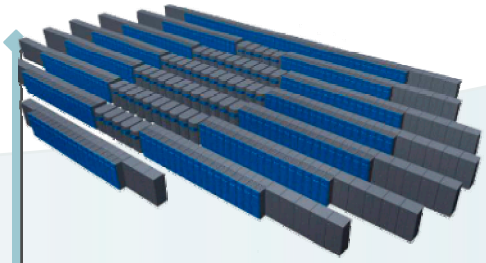


We will advance computational capability by 1000× over the next decade

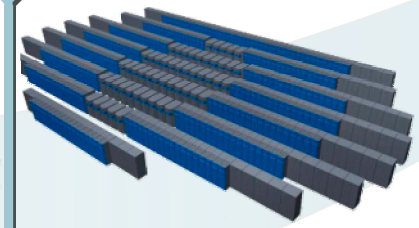
Mission: Deploy and operate the computational resources required to tackle global challenges	Vision: Maximize scientific productivity and progress on the largest-scale computational problems
• Deliver transforming discoveries in materials, biology, climate, energy technologies, etc. • Ability to investigate otherwise inaccessible systems, from supernovae to energy grid dynamics	• Providing world-class computational resources and specialized services for the most computationally intensive problems • Providing stable hardware/software path of increasing scale to maximize productive applications development



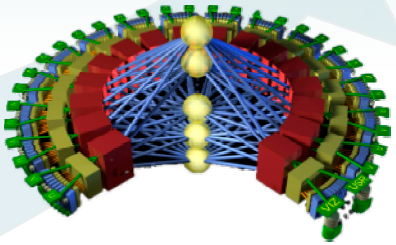
Cray "Jaguar": 1 PF
Leadership-class
system for science



DARPA HPCS: 20 PF
Leadership-class
sustained PF system



100–250 PF

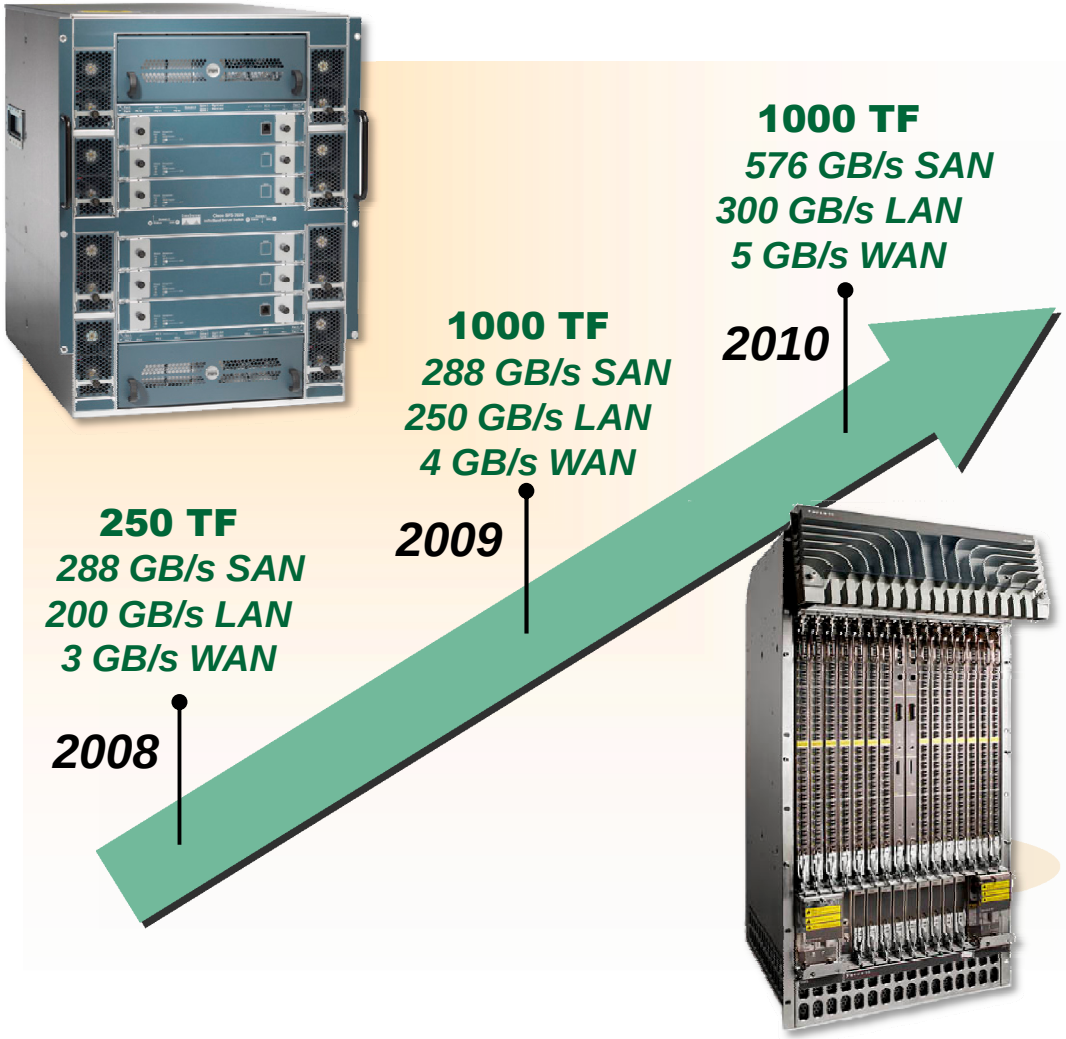


Future system: 1 EF

FY 2009 **FY 2011** **FY 2015** **FY 2018**

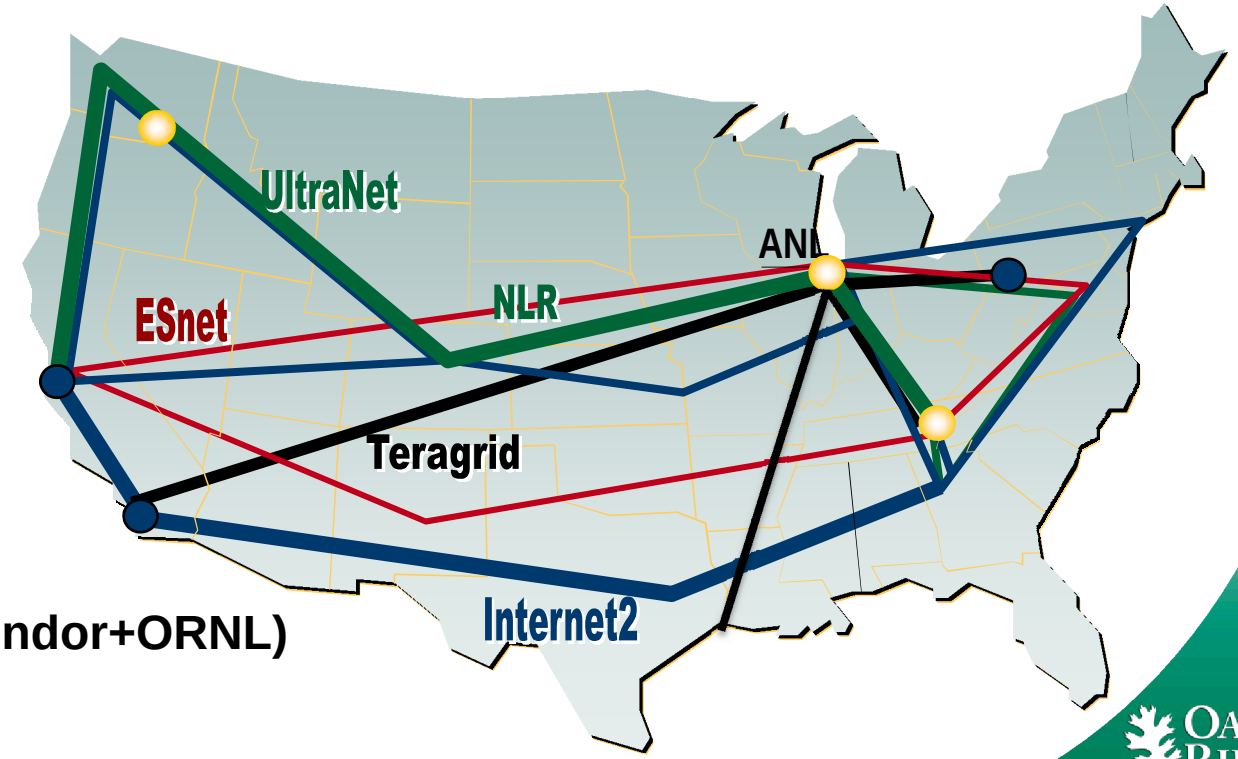
Network

- Shifting to a hybrid InfiniBand/Ethernet network
- InfiniBand-based network helps meet the bandwidth and scaling needs for the center
- Wide-area network will scale to meet user demand using currently deployed routers and switches



High bandwidth connectivity to NCCS enables efficient remote user access

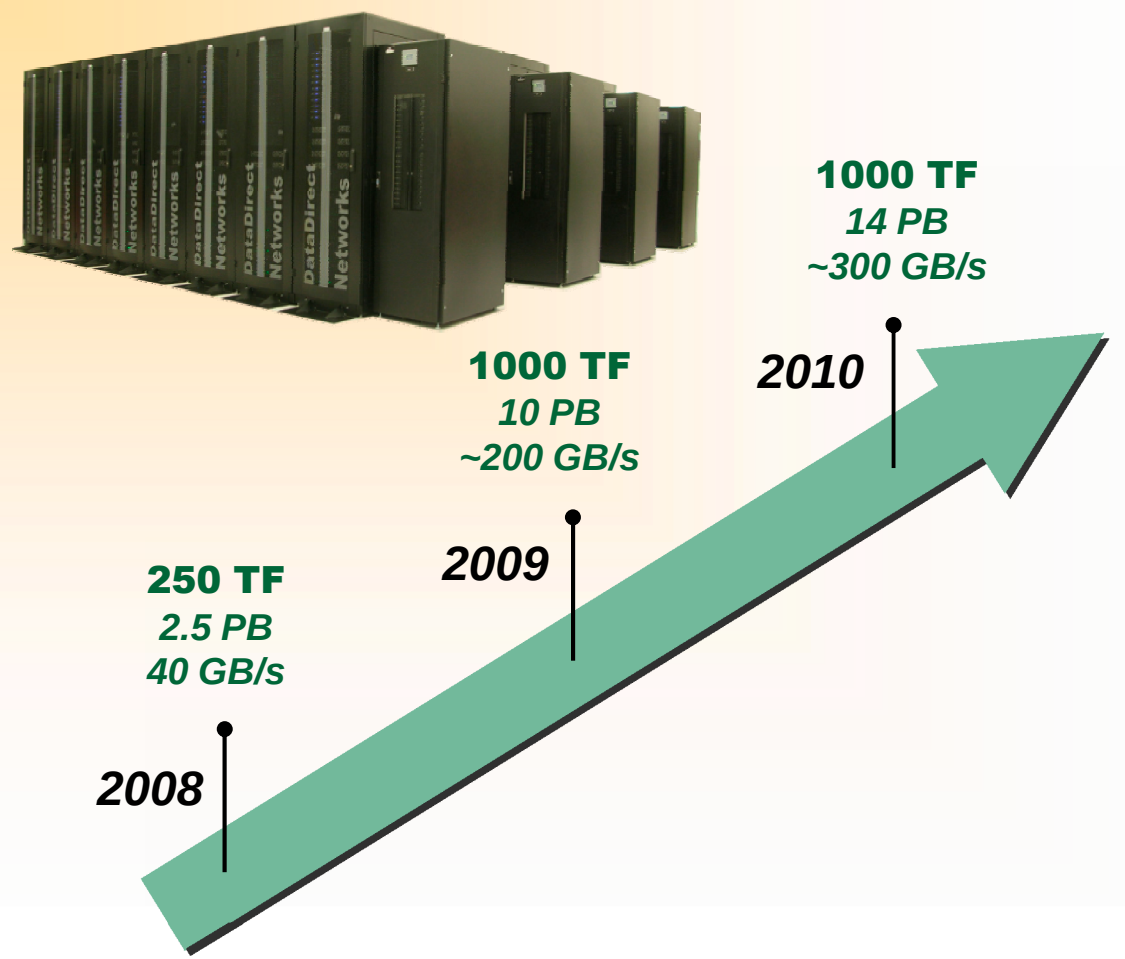
Connected to Major Science Networks We own dark fiber to Atlanta and Chicago	
OC192 to ESNET with backup OC48	1 - 4 x 10 Gb to NSF Teragrid
10 Gb to Internet2	2 x 10 Gb UltraScienceNet
4 x 10 Gb to National Lambda Rail	10 Gb Futurenet to NSF Cheetah net



Target
100Gb/lamda (NRL+Vendor+ORNL)

Centerwide file system (aka Spider)

- Increase scientific productivity by providing single repository for simulation data
- Connect to all major LCF resources
- Connected via our InfiniBand SAN
- Becomes *the* file system for the 1000 TF system



Storage Infrastructure

Challenge:
Management of over 14,000 disk drives

Center -wide shared file systems

- **"Spider" is being installed to provide a shared, parallel file system for all systems**
  - **Based on Lustre file system**
- **Bandwidth of over 200 GB/s with 10 petabytes capacity**
- **HPSS provides archival storage for all system. 40 PB capacity; easily expandable.**
- **HPSS has been upgraded with two additional tape libraries to add additional capacity and bandwidth**

HPSS

Office of Science
U.S. DEPARTMENT OF ENERGY

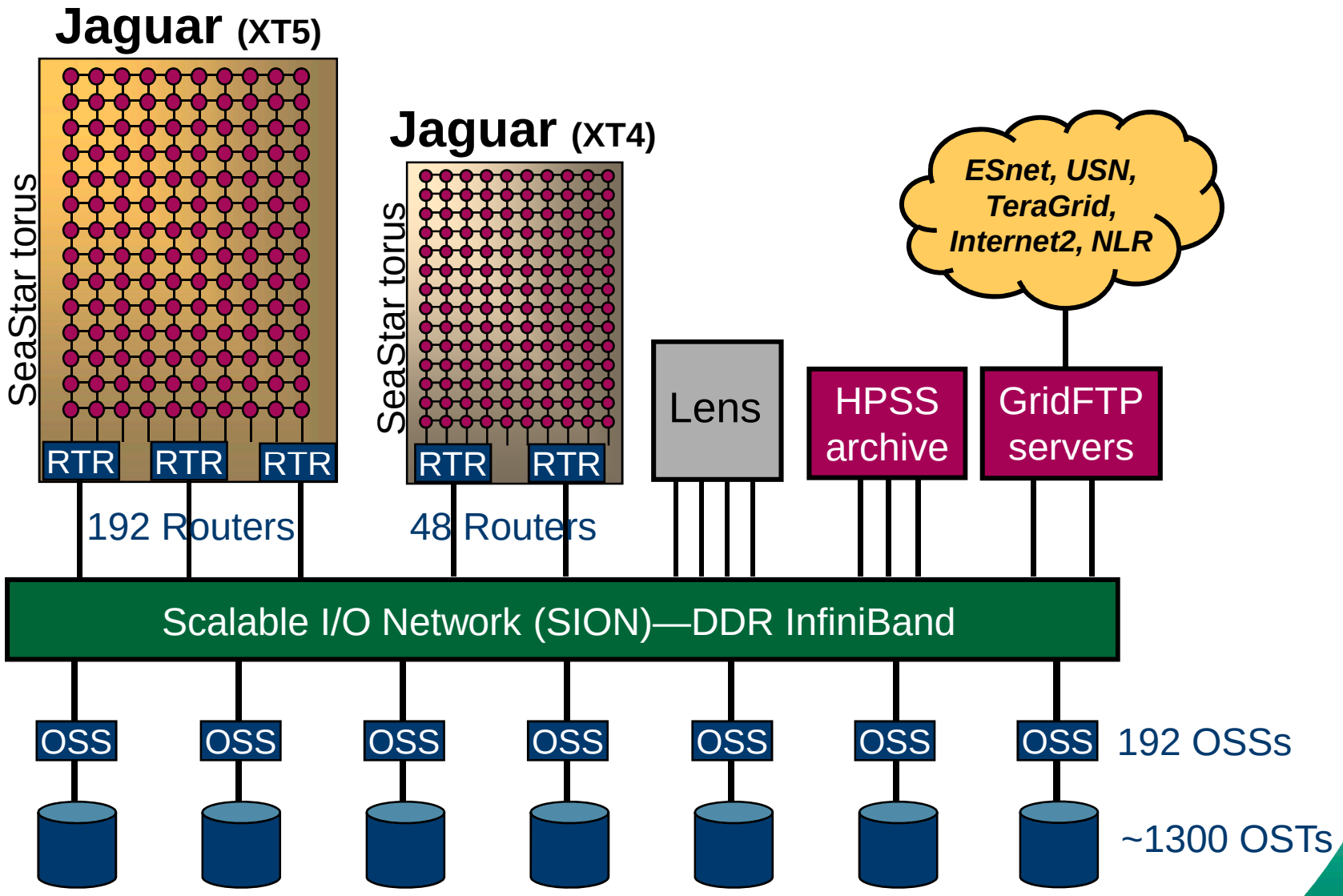
Spider

- Lustre-based file system
- Can natively utilize the InfiniBand network
- Already proven on today's XT4 at 10K+ clients

Production and test-beds	FY 2008	FY 2009	FY 2010
OSS	48	192	272
Controller pairs	12	48	68
Capacity (PB)	2.5	10	14
Bandwidth (GB/s)	40	~200	~300

- External-based system will utilize routers that are part of the transport protocol used in Lustre (route between IB and SeaStar)
- External system already demonstrated on current XT systems

Spider



Total storage capacity = 10 petabytes

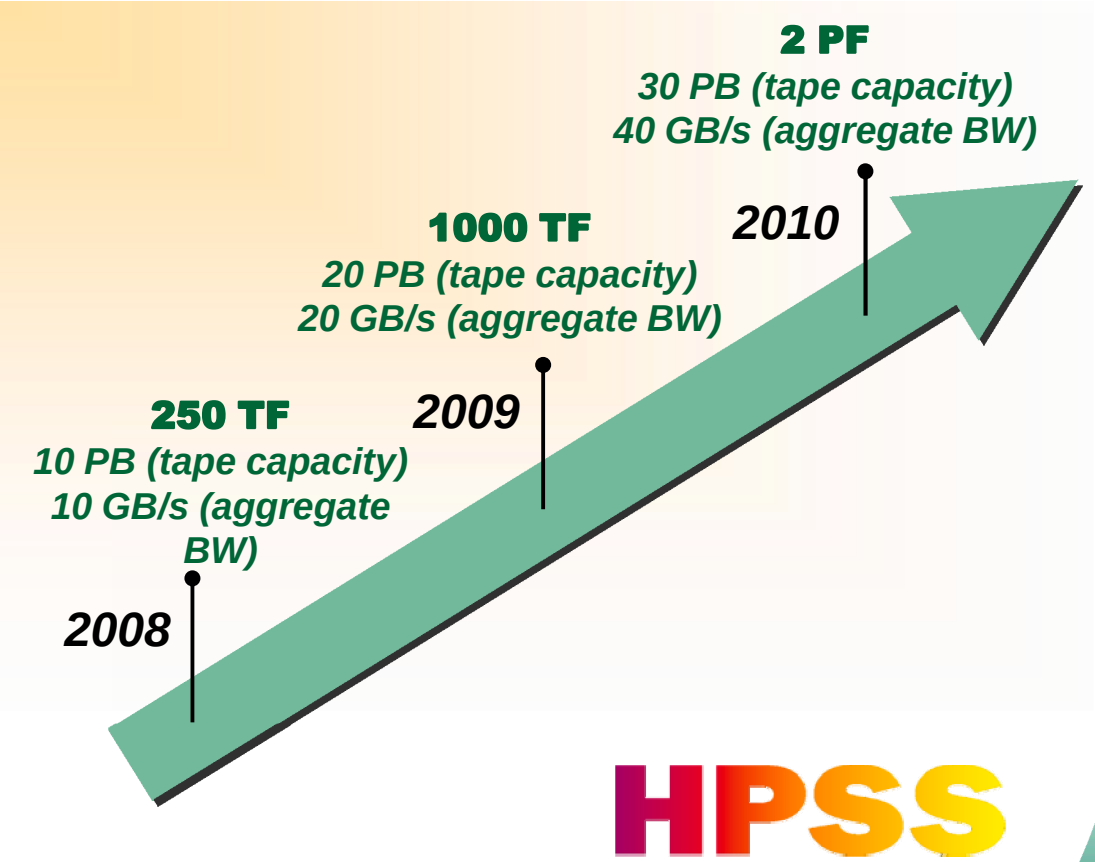
Target bandwidth = 200 gigabytes/second

Spider facts

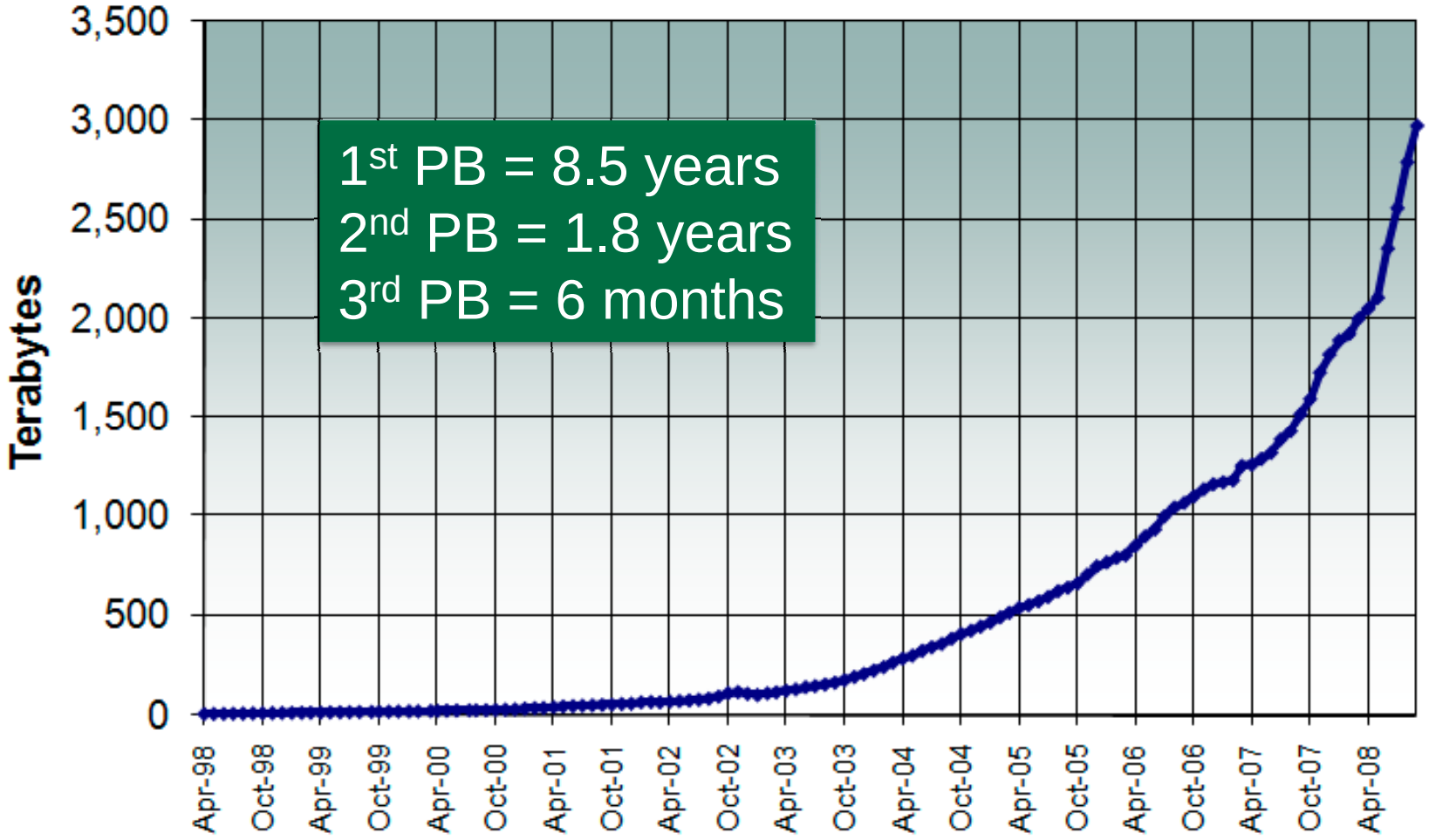
- >200 GB/s of aggregate bandwidth
- 48 DDN 9900 couplets
- 13,440 1 TB SATA drives
- Over 10 PB of RAID6 capacity
- 192 storage servers
- More than 1000 InfiniBand cables
- ~0.5 MW of power
- ~20,000 lb of disks
- Fits in 32 cabinets using 572 ft^2

Archival storage

- HPSS software has already demonstrated ability to scale to many PB
- 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)



Archival storage—past usage

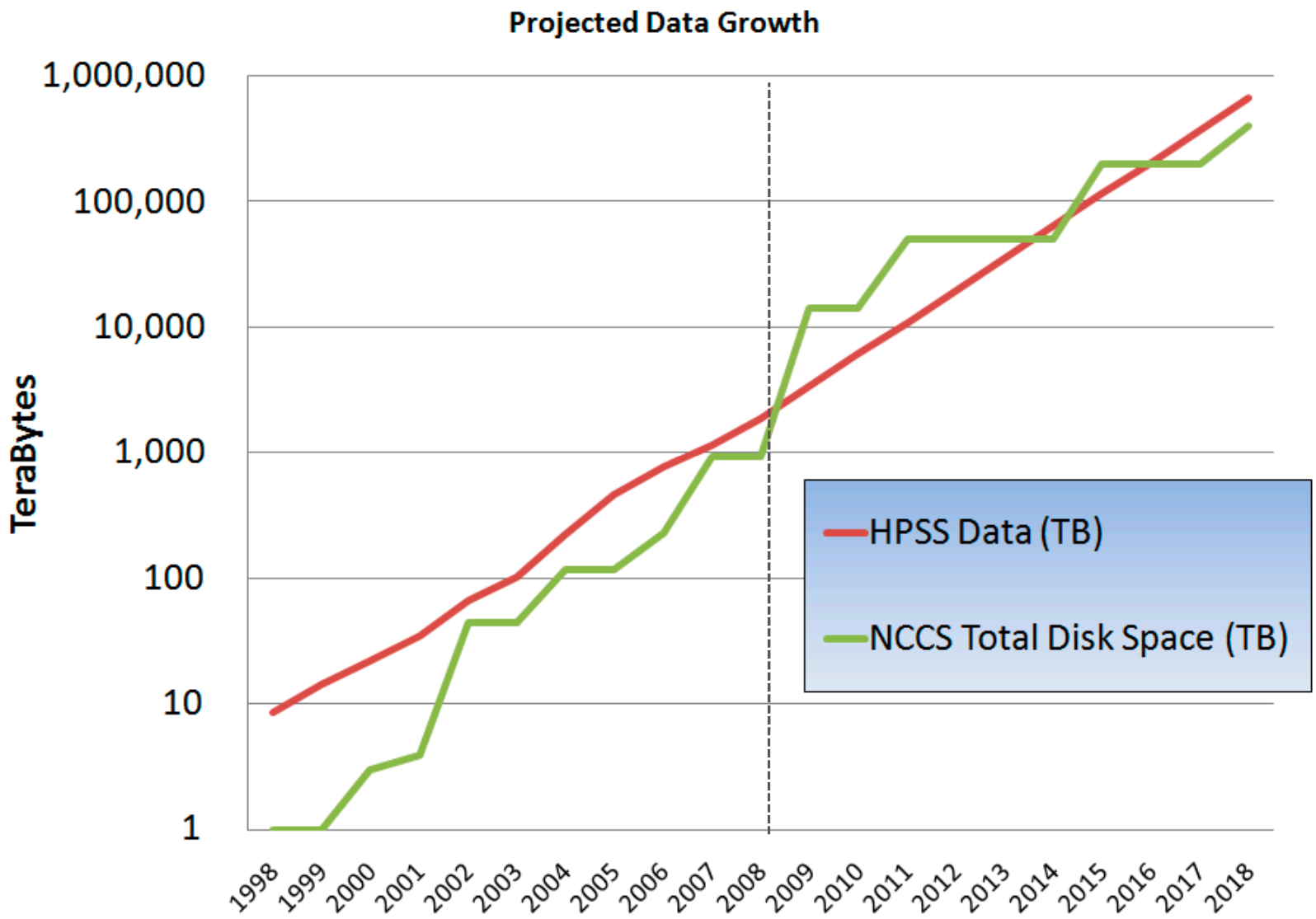


Data are Doubling Every Year!

Archival storage

	FY 2008	FY 2009	FY 2010
Silos	4	4	5
Tape drives	40	56	70
Tape capacity (PB)	10	20	30
Disk cache (TB)	500	1500	3000
Tape bandwidth (aggregate GB/s)	3.8	7.6	12
Disk bandwidth (aggregate GB/s)	10	20	40

Exponential growth in data capability



Visualization and data analysis resources

Hardware

- Everest Powerwall
  - 30’x8’ 35 megapixel display wall
- Hawk Cluster
  - 58 Opteron nodes with NVIDIA GPUs
- Lens Cluster
  - 32 nodes each with 64 GB and 4 quad-core Opterons w/ GPUs
- Everest Cluster
  - 18 nodes to drive display wall
- Ewok Cluster
  - 81 node dual-core Xeons

Software

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

Data analytics—strategies

Jaguar XT4 (250 TF)

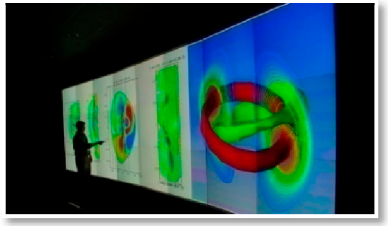
- **Utilize portion of system for data analysis (50 TF/20 TB)**

Jaguar XT5

- **Utilize Jaguar as analysis resource (250 TF/60 TB)**
- **Provision fraction of Jaguar XT5 for analysis**

Operations infrastructure systems
Now and future estimates

Archival Storage	FY 2008	FY 2009	FY 2010
Capacity (PB)	10	20	30
Bandwidth (GB/s)	10	20	40



Viz/End-to-End	FY 2008	FY 2009	FY 2010
IO B/W	15	40	200
Memory (TB)	2+	60	200

Central Storage	FY 2008	FY 2009	FY 2010
Capacity (PB)	1	10	14
Bandwidth (GB/s)	40	200	300

Networking	FY 2008	FY 2009	FY 2010
External B/W (GB/s)	3	4	5
LAN B/W (GB/s)	200	250	300
SAN B/W (GB/s)	288	288	576

Contact

Galen Shipman

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