The Path to Petascale at Oak Ridge National Laboratory

Presented by

Philip C. Roth
Future Technologies Group
Oak Ridge National Laboratory

SC06 POWERFUL BEYOND IMAGINATION

The ORNL Future Technologies Group

- **Led by Jeffrey S. Vetter (vetter@ornl.gov)**
- **Founded October 2004**
- **Computer Science and Mathematics Division**
- **Mission: identify and understand core technologies for improving performance, efficiency, reliability, and usability of future generations of high-end computing platforms**
- **Use measurement, modeling, and simulation**

Group Members		***Main Collaborators***
Sadaf Alam	Philip Roth	Patrick Worley
Richard Barrett	Olaf Storaasli	Pratul Argawal
Nikhil Bhatia	Weikuan Yu	Jacob Barhen
Jeremy Meredith	Micah Beck (UT-K)	Hong Ong
Collin McCurdy	David Bader (GaTech)	Many vendors
Kenneth Roche		

FutureTech I/O Research

- **Performance measurement and prediction**

Preliminary I/O Summary

			Mechanism				Example Problem	
App	Version	Use	Fortran I/O	MPI-IO	NetCDF	Scheme	Size	Frequency
GYRO	3.0.0	Read input files	x			F: Rank 0	< 1 MB	Initialization
		Write checkpoint files	x	x		F: Rank 0	87.5 MB	Once per 1000 time-steps
		Write logging/debug files	x			Rank 0	~150 KB	File-dependent
POP (standalone)	1.4.3/2.0	Read input files	x		x (2.0 only)	Parallel, rank 0		Initialization
		Read forcing files	x		x	Parallel, rank 0		Every few time-steps
		Write 3d field files	x		x	Parallel, rank 0	1.4 GB	Several per simulation-month
CAM (standalone)	3.0	Read input files	x			Rank 0	~300 MB	Initialization
		Write checkpoint files	x			Rank 0		Once per simulation-day
		Write output files	x			Rank 0	~110 MB	Termination
AORSA2D		Read input files	x			All ranks	~26 MB	Initialization
		Write output files	x			Rank 0	~10 MB	Termination
VH-1		Read input files	x			Rank 0		Initialization
		Write timestamp files	x			All ranks, post-processed into single file	28 GB / timestep	Hundreds per run

Oak Ridge National Laboratory
U. S. Department of Energy

UT-Battelle

1

- **Deploying systems software and storage testbed in the Experimental Computing Laboratory (ExCL)**

Oak Ridge National Laboratory

Center for Computational Sciences

- Established in 1992
- In 2004, designated by Secretary of Energy as site of nation's Leadership Computing Facility
- http://www.nccs.gov

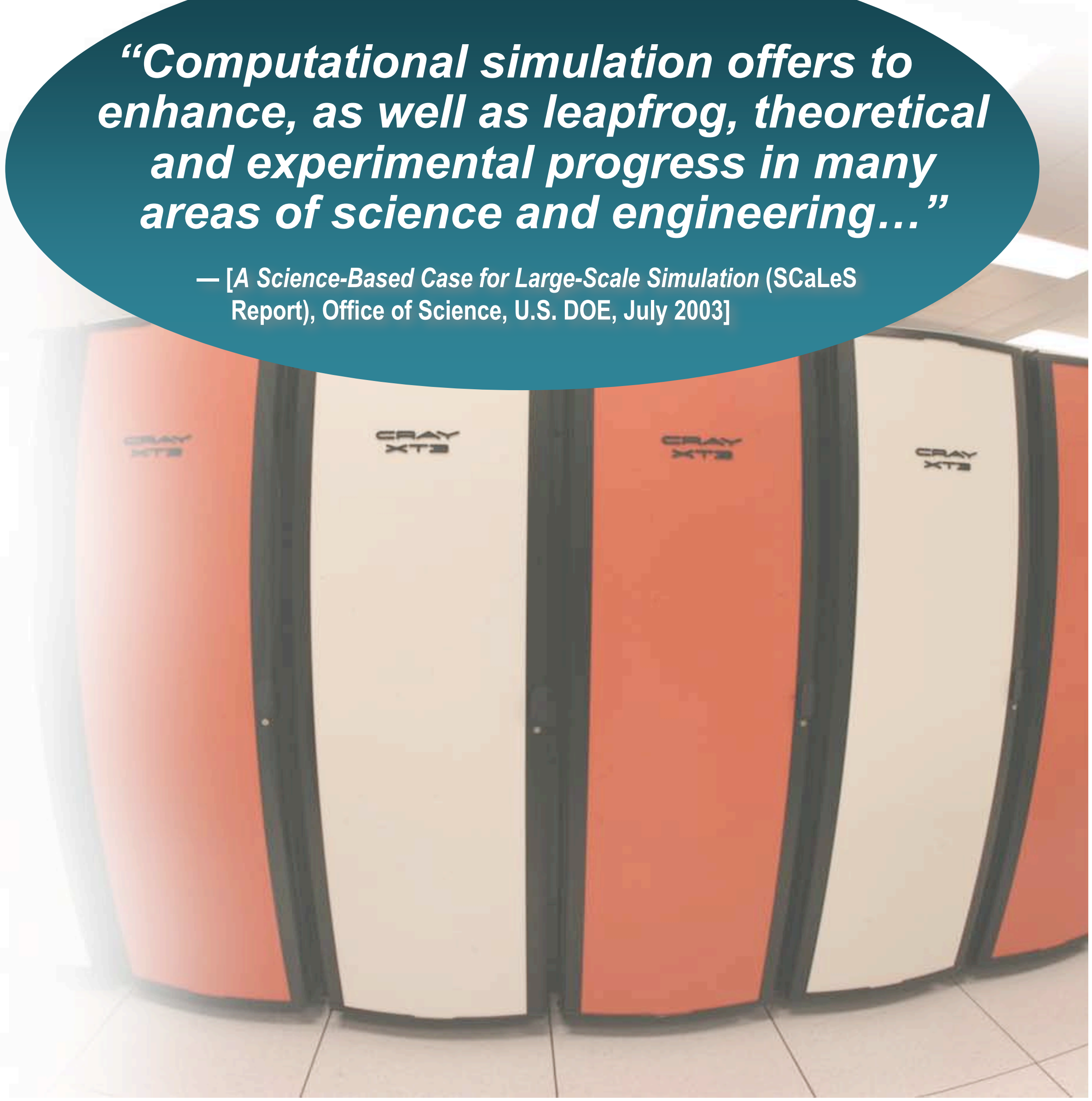
OAK RIDGE National Laboratory

Science Drivers

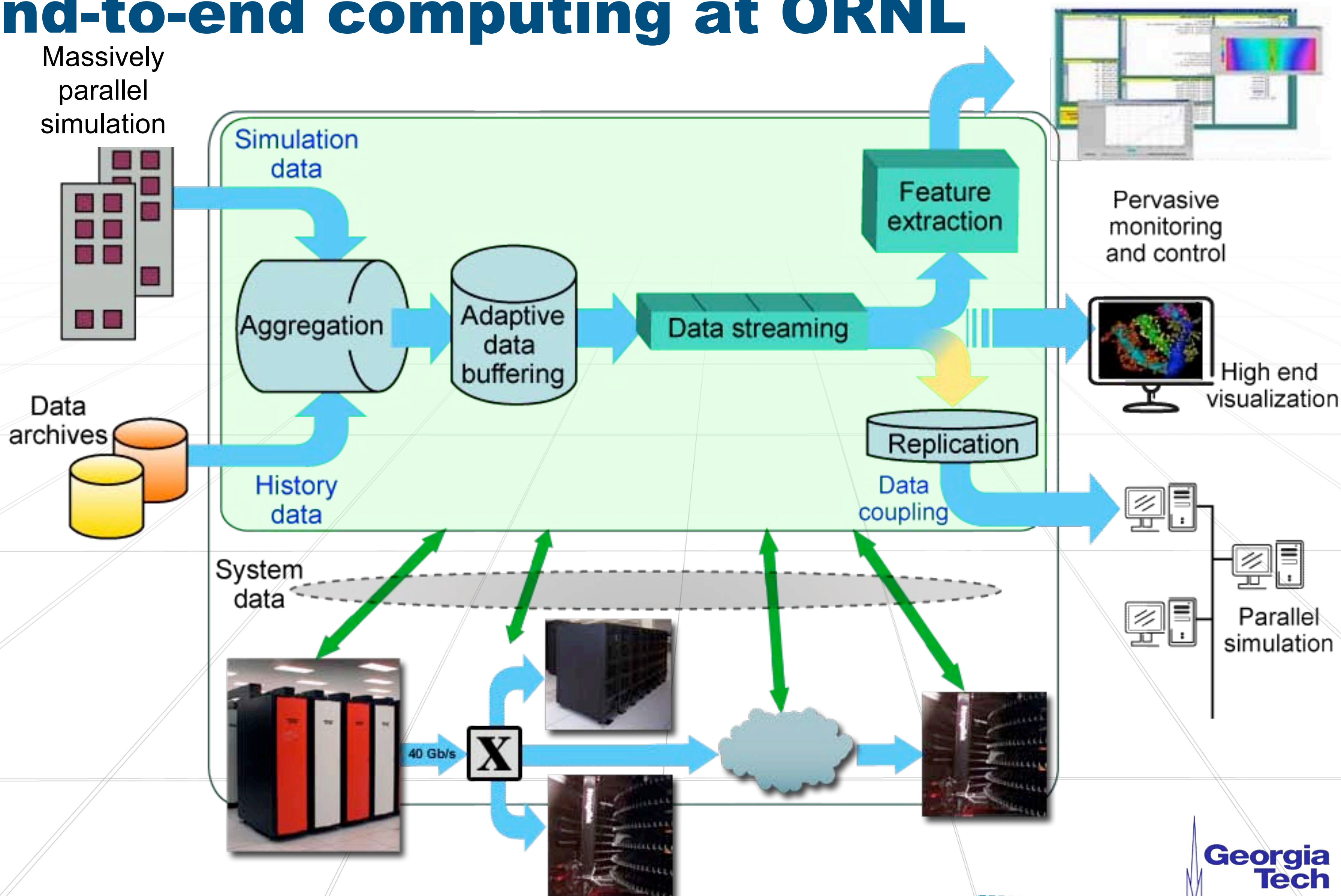
- **Advanced energy systems (e.g., fuel cells, fusion)**
- **Biotechnology (e.g., genomics, cellular dynamics)**
- **Environmental modeling (e.g., climate prediction, pollution remediation)**
- **Nanotechnology (e.g., sensors, storage devices)**

"Computational simulation offers to enhance, as well as leapfrog, theoretical and experimental progress in many areas of science and engineering…"

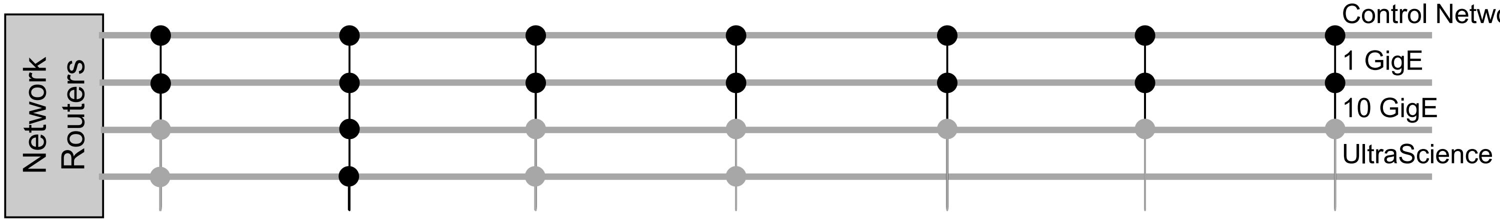
— [*A Science-Based Case for Large-Scale Simulation* (SCaLeS Report), Office of Science, U.S. DOE, July 2003]



End-to-end computing at ORNL



Current CCS Resources



October 2006 Summary

7 Systems

Supercomputers
13,834 CPUs
26TB Memory

74 TFlops

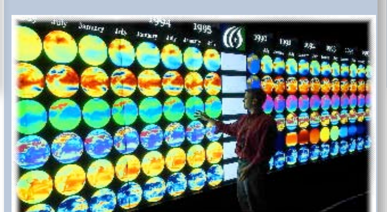
Cray XT3 Jaguar	Cray X1E Phoenix	SGI Altix Ram	IBM SP4 Cheetah	IBM Linux NSTG	Visualization Cluster	IBM HPSS
(10,506) 2.6GHz 21TB Memory	(1,024) 0.5GHz 2 TB Memory	(256) 1.5GHz 2TB Memory	(864) 1.3GHz 1.1TB Memory	(56) 3GHz 76GB Memory	(128) 2.2GHz 128GB Memory	Many Storage Devices Supported
120 TB	44 TB	36 TB	32 TB	4.5 TB	9 TB	5 TB

Total Shared Disk 250.5 TB

Evaluation Platforms
- 144-processor Cray XD1 with FPGAs
- SRC Mapstation
- Clearspeed
- BlueGene (at ANL)

Test Systems
- 96-processor Cray XT3
- 32-processor Cray X1E*
- 16-processor SGI Altix

Scientific Visualization Lab
35 megapixels Power Wall



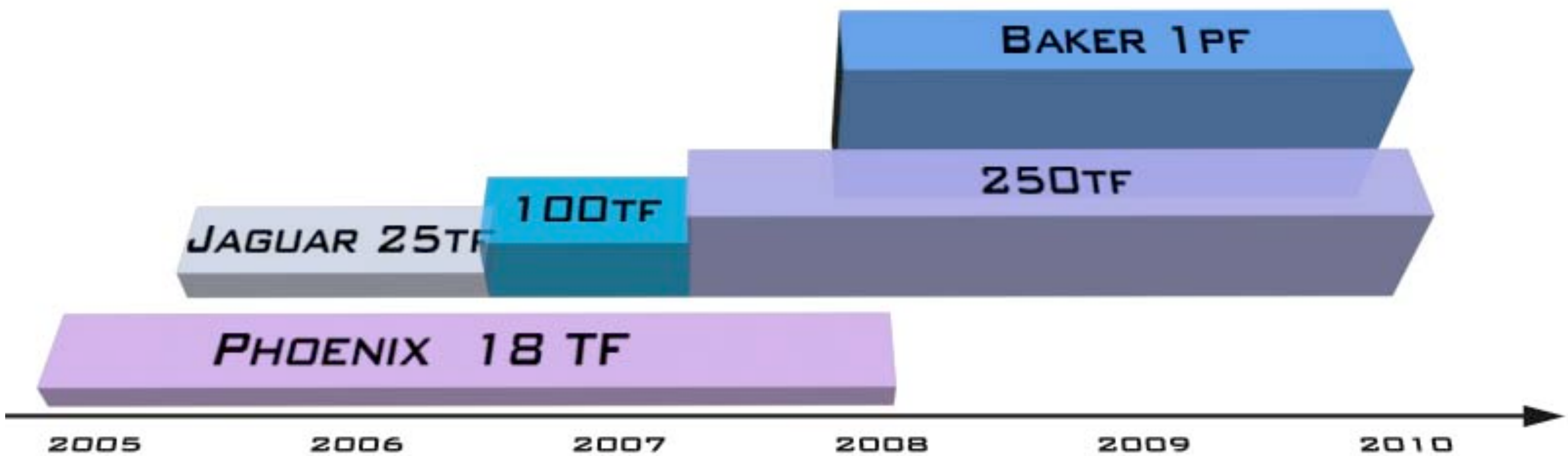
Backup Storage

5PB

5 PB Data Storage

OAK RIDGE National Laboratory

CCS Roadmap



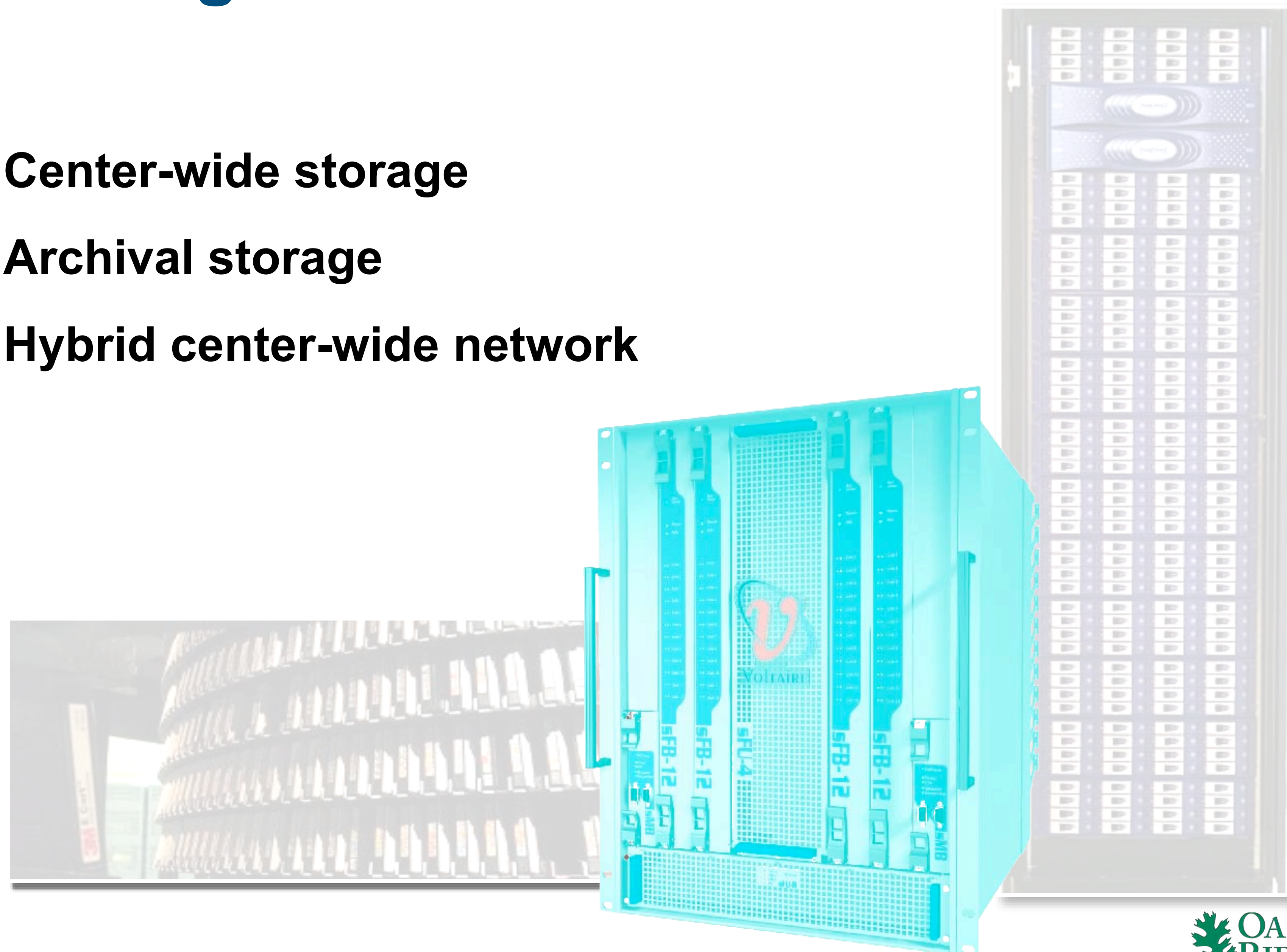
BAKER 1PF
250TF
100TF
JAGUAR 25TF
PHOENIX 18 TF
2005
2006
2007
2008
2009
2010

System Specifications

	54 TF	100 TF	250 TF	1000 TF
Compute Processors	5,212 Dual-Core 2.6 GHz Opteron	Adds 6,296 Dual-core 2.6 GHz Opterons	Replaces 6,296 dual-core with quad-core Opterons	22,400 quad-core Opterons
SIO Processors	82 Single-core 2.4 GHz Opteron	116 dual-core 2.8 GHz Opteron	Total 198 Opteron	544 quad-core Opteron
Memory per Socket/Total	4 GB / 20 TB	4 GB / 45 TB total sys	8 GB / 69 TB total sys	8 GB / 175 TB
Interconnect Bandwidth per Socket	Seastar 1 1.8 GB/s	Seastar 2 4.0 GB/s	Seastar 2 4.0 GB/s	Gemini
Disk Space	120 TB	Adds 780 TB total 900 TB	Total 900 TB	5 - 15 PB
Disk Bandwidth	14 GB/s	Adds 41 GB/s total 55 GB/s	Total 55 GB/s	240 GB/s

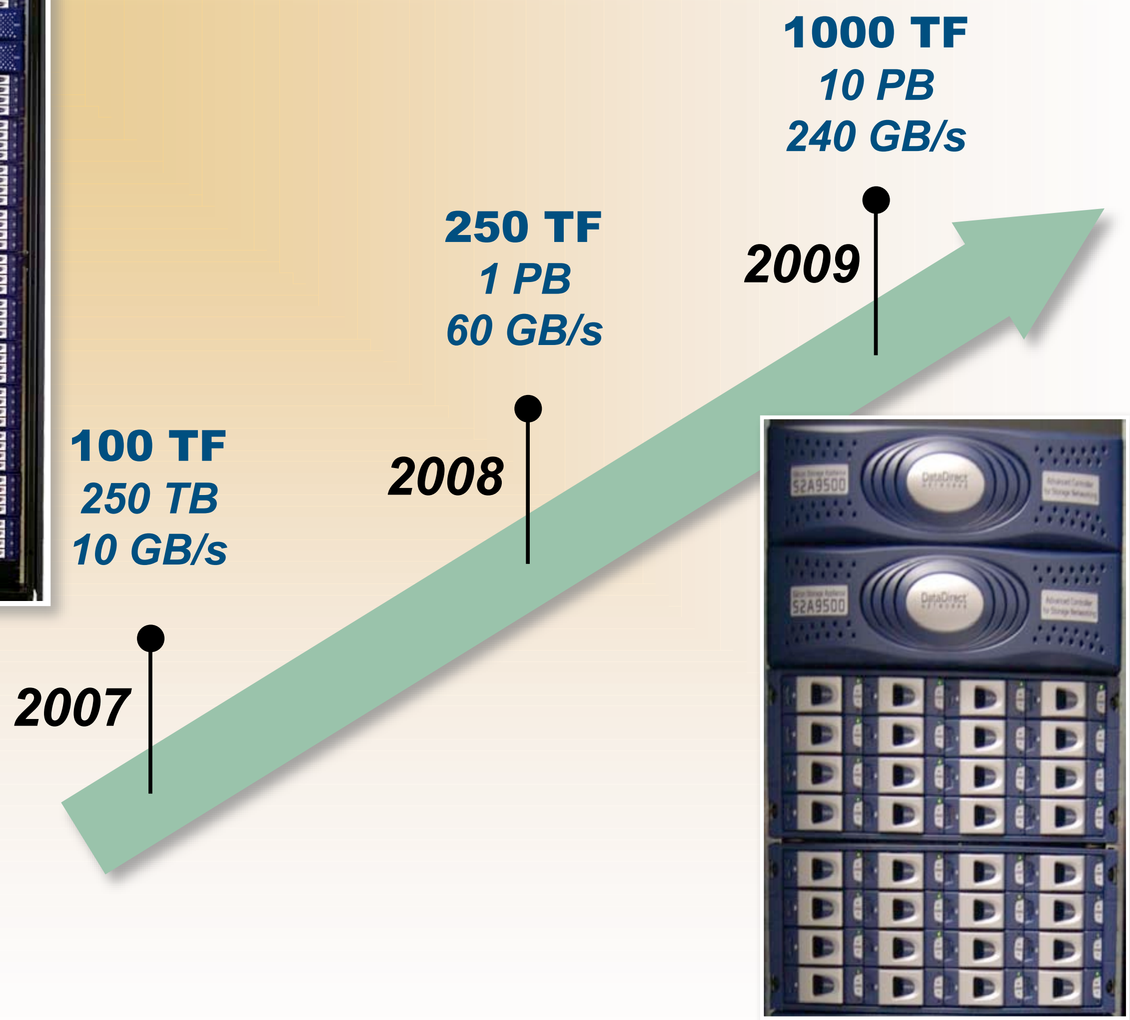
Evolving CCS Infrastructure

- Center-wide storage
- Archival storage
- Hybrid center-wide network



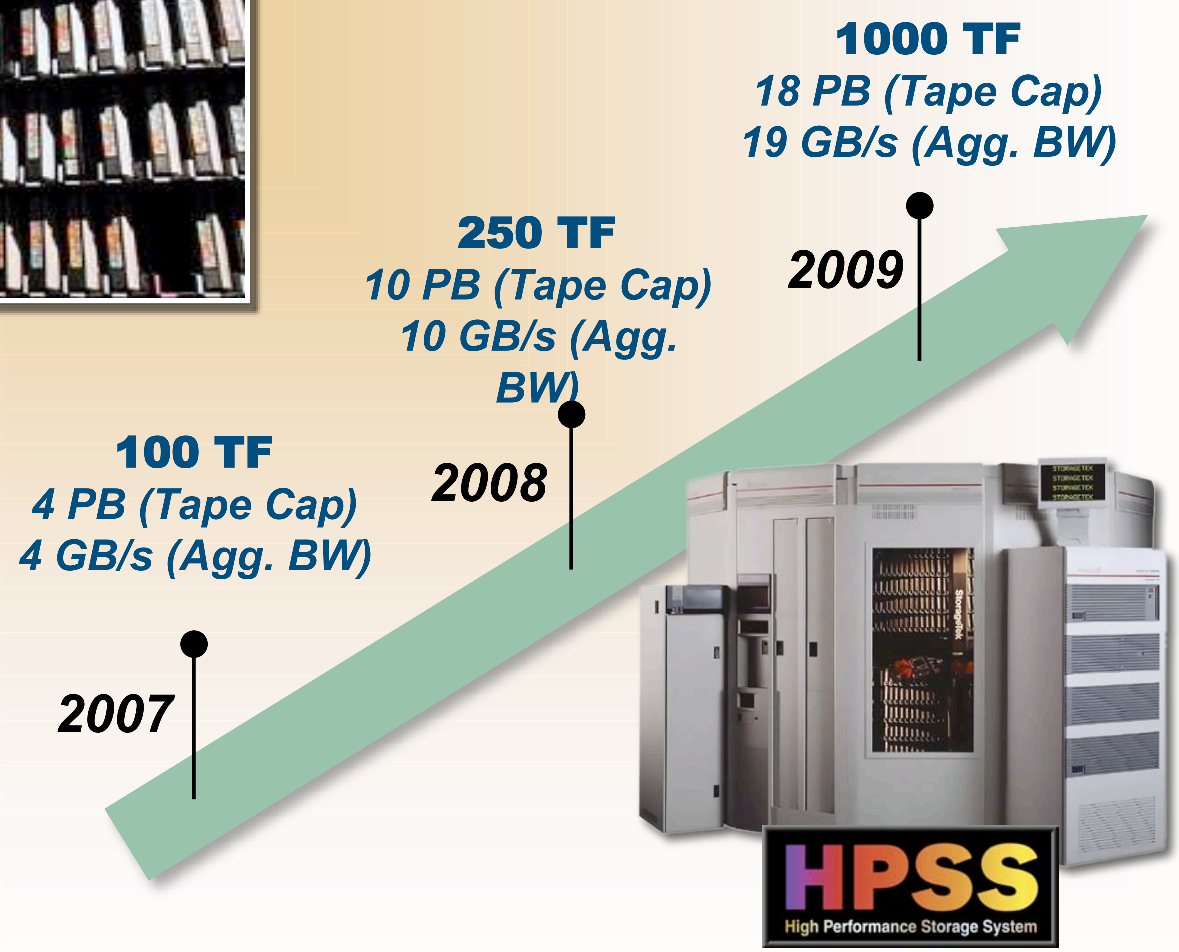
Center-Wide File System (Spider)

- **Increase scientific productivity by providing single repository for simulation data**
- **Connect to all major LCF Resources**
- **Connected to both InfiniBand and Ethernet networks**
- **Potentially becomes *the* file system for the 1000 TF System**



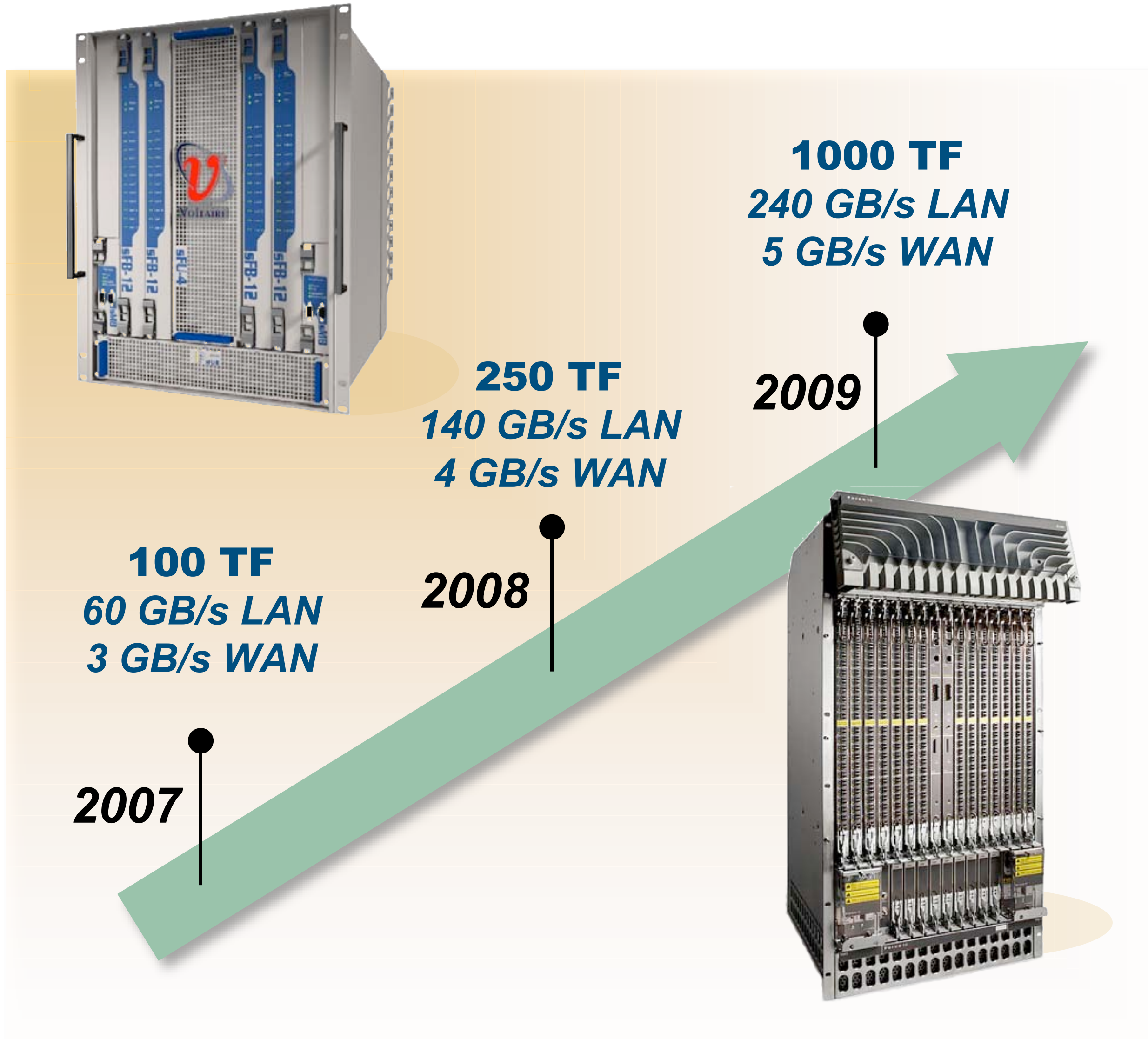
Archival Storage

- HPSS Software has already demonstrated ability to scale to many PB
- Add 2 Silos/Year
- Tape Capacity & Bandwidth, Disk Capacity and Bandwidth are all scaled to maintain a balanced system
- Utilize new methods to improve data transfer speeds between parallel file systems and archival system



Network

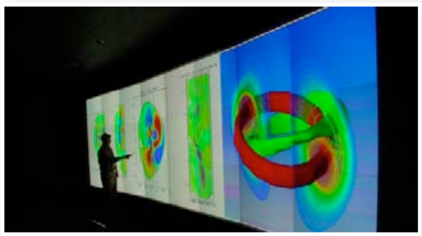
- Shifting to a hybrid InfiniBand/Ethernet network
- InfiniBand based network helps meet the bandwidth and scaling needs for the center
- Wide-Area network will scale to meet user demand using currently deployed routers and switches



OAK RIDGE National Laboratory

Operations Infrastructure Systems
Now and Future Estimates

Archival Storage	FY07	FY08	FY09
Capacity (PB)	4	10	18
Bandwidth (GB/s)	4	10	19
Resources ($M)	3.6	6.0	3.2



Viz/End-to-End	FY07	FY08	FY09
IO B/W	10	20	60
Memory (TB)	0.5	20	69
Resources ($M)	0.15	0.82	0.35

Central Storage	FY07	FY08	FY09
Capacity (PB)	0.5	1.0	10.0
Bandwidth (GB/s)	10	60	240
Resources ($M)	3.4	11.4	4.0

Networking	FY07	FY08	FY09
External B/W (GB/s)	3	4	5
LAN B/W (GB/s)	60	140	240
Resources ($M)	1.5	1.4	1.1

Oak Ridge National Laboratory

Centers of Excellence at ORNL

- **Cray Center of Excellence**
  - Assists science teams in achieving desired performance on Cray platforms
  - Scaling and tuning libraries and codes
- **Lustre Center of Excellence**
  - Announced November 14, 2006
  - Enhance scalability of Lustre File System to meet performance requirements of petascale systems
  - Build Lustre expertise through training and workshops
  - Assist science teams in achieving desired I/O performance

Contacts

Philip C. Roth

Future Technologies Group
Computer Science and Mathematics Division
rothpc@ornl.gov

Arthur S. Bland

Leadership Computing Facility Project Director
Center for Computational Sciences
blandas@ornl.gov

R. Shane Canon

Technology Integration Group
Center for Computational Sciences
(865) 574-2028
canonrs@ornl.gov