“How To Build the Fastest Academic Supercomputer in America-- Twice in One Year”

Jay Boisseau, Director
Texas Advanced Computing Center
The University of Texas at Austin
October 3, 2007

“It Ain’t Braggin’ If It’s True”

- TACC has emerged as one of the leading academic advanced computing centers in the nation & world in the past six years
- TACC’s world-class resources & services advance science nationwide through the NSF TeraGrid
- TACC R&D projects produce computational technologies and techniques that augment resources, users, and science
- *“Everything’s bigger in Texas:”* TACC will deploy the most powerful general-purpose* supercomputer in the world in November 2007

*Maybe. And I’m not counting BlueGene/L as general purpose….

TACC

But:

- This title was Henry's, not mine
- He took advantage of me being busy to pick my title without much argument, long ago
- I did argue, a little (hence the modifiers 'academic,' 'in America,' etc.)
- What the hell, I'm a Texan (now)!
- This may be all the braggin' I get to do with petascale systems cropping up like weeds in 2008
- By the way, it's really 14 months, not one year

How to Make Such a Presentation…

- There are really two options:
  1. Stick to the title, strictly
  2. Don't
- I chose option #2
  - I'll talk about *how we got to where we are, which is really the answer to the more specific question*
  - I'll discuss Ranger a lot, but give props to Lonestar as well

TACC

Outline

1. All the crap I just said….
2. TACC Overview
3. Onset of the Terascale & Petascale Eras
4. Ranger: A Bridge to Petascale Computing
5. Closing Thoughts and Summary

TACC Mission

To enable scientific discovery and enhance society through the application of advanced computing technologies.

TACC

TACC Strategic Approach

- Resource & Services
  - Evaluate, acquire & operate world-class resources
  - Provide expert support via leading technology expertise
- Research & Development
  - Produce new computational technologies and techniques
  - Collaborate with researchers to apply advanced computing technologies
- Education & Outreach
  - Inform public about impact of advanced computing
  - Educate students to increase participation in advanced computing careers

TACC Technology Focus Areas

- High Performance Computing (HPC)
  - Performance benchmarking, analysis, optimization
  - Linear algebra, solvers
  - CFD, computational chemistry, weather/ocean modeling, computational biomedicine
- Data & Information Analysis (DIA)
  - Scientific visualization
  - Data collections management
  - Data analysis & mining
- Distributed & Collaborative Computing (DCC)
  - Portals & gateways
  - Middleware for scheduling, workflow, orchestration

TACC Applications Focus Areas

- UT Austin is a big university--so we support all disciplines
- NSF supports all science & engineering users--so again we support all disciplines
- But three areas receive extra attention:
  1. Geosciences: crucial to UT Austin, to Texas, and to society
  2. Life sciences: big in UT System, important to society, well funded by NIH
  3. Emergency situation assessment & response: crucial to society, leverages lots of applications and technology expertise at UT Austin

TACC HPC & Storage Systems

LONESTAR

Dell Linux Cluster
2900+ dual-core CPUS,
>11 TB memory, >60 Tflops peak

CHAMPION

IBM Power5 System
96 Power5 CPUs,
192 GB memory, ~1 teraflop

ARCHIVE

STK PowderHorns (2), managed by Cray DMF
2.8 PB max capacity

GLOBAL DISK

Sun SANs and
Data Direct Disk
> 50TB

TACC

Lonestar: World-Class HPC System

Lonestar is the one of the most powerful supercomputing systems for academic research in the US/world **now** (production for one year)

- Key specs
  - 1460 Dell PowerEdge 1955 blade servers
  - 2920 Intel Xeon dual-core processors / 5840 cores at 2.66 GHz
  - Cisco InfiniBand interconnect
- Performance/capabilities
  - *62 teraflops peak performance*
  - *11 TB total memory*
  - *69 TB disk in parallel file system*
  - *10 gigabit/sec bandwidth, < 5 μsec latency*

Lonestar is available nationally via the NSF TeraGrid

TACC

TACC Advanced Visualization Systems

- *Maverick:* Sun Terascale Visualization System
  - 128 UltraSparc 4 cores, ½ TB memory
  - 16 GPUs, > 3 Gpoly/sec
- Also: SGI Prism, Dell Cluster, Workstations
- Immersive and tiled displays
  - 3x1 semi-cylinder immersive environment
    - immersive capabilities with head/motion tracking
  - 5x2 large-screen, 16:9 panel tiled display
  - 3x3 tiled 30” LCD display

TACC

TACC System Hosting

- TACC purchases scalable systems: expandable
- Many researchers contribute funding for expansion in return for guaranteed allocations
- Advantages:
  - ✓ Better pricing, more flops/$
  - ✓ Much better utilization through allocations (no lost idle cycles) - much better flops/$
  - ✓ Professional operations with 24x7 support - less $
  - ✓ TACC software, archival system, networking, etc. - more capability, less $
  - ✓ Much greater scientific capability of larger system - bigger science/$
- No fee, but must provide 10% of cycles to community

TACC

TACC Support Services

- Technical documentation
  - Available via TACC User Portal and main web site
- Consulting
  - User submit issues via the TACC User Portal
- Training
  - Taught on-site, sign up at TACC User Portal
  - Can be taught at campuses with 10+ attendees, facilities
- System selection/configuration consulting
  - Can provide guidance on purchasing of local resources

TACC R&D – High Performance Computing

- Scalability, performance optimization, and performance modeling for HPC applications
- Evaluation of cluster technologies for HPC
- High performance linear algebra, solvers
- Climate, weather, ocean modeling collaboration and support of DoD
- Computational fluid dynamics
- Computational chemistry

TACC

TACC R&D – Data & Information Analysis

- Feature detection / terascale data analysis
- Remote interactive visualization
- Performance characteristics and capabilities of high-end visualization technologies
- Hardware accelerated visualization and computation on GPUs
- Hosting scientific data collections and analysis services (sorta new)
- Scientific data mining & statistical analysis (new)

TACC R&D – Distributed & Collaborative Computing

- Web-based grid-enabled portals and science gateways
- Grid scheduling, job orchestration and workflow tools
- Large-scale distributed computing
- Overall grid deployment and integration

TACC

“Scientific Computing Curriculum” Academic Classes

- Teach *applied* advanced computing technologies and techniques
- Comprehensive five-course curriculum:
  - Scientific programming and computing
  - Parallel computing, visualization, grid computing
- Taught through UT CS department but also *cross-listed in science/engineering departments*
  - New Division of Statistics & Scientific Computation will list in Fall08
- Some class materials available for download now
- Will record and post lectures in 2008, teach remotely in 2009(?)

TACC Scientific Computing Courses

- Undergrad
  - Introduction to Scientific Programming (soon)
  - Scientific/Technical Computing
- Undergrad and grad
  - Parallel Computing for Science & Engineering
  - Distributed and Grid Computing for Science and Engineering
  - Visualization and Data Analysis for Science and Engineering (soon)

Strategic Focus Activities in 2008+

- Petascale Computing
  - Integration, management & operation of leadership-class systems
  - Performance optimization for multi-core processors
  - Fault tolerance for applications on large systems
  - Achieving extreme performance & scalability: algorithms, libraries, community codes, frameworks, programming tools, etc.
- Petascale Visualization & Data Analysis
  - 'In-simulation' visualization, HPC visualization applications
  - Remote & collaborative visualization
  - Feature detection techniques
  - Data collections hosting with layered analysis services
- Remote/Collaborative Usage of Petascale Resources
  - Tools for scheduling, orchestrating work on petascale resources
  - Tools for integrating local, remote resources
  - Portals for communities, community applications

How We Built Lonestar (the Third)

- ✓ Acquired previous clusters and proved value, impact, manageability, etc.,
- ✓ Became experts in large-scale clusters with previous Lonestar
- ✓ Vision and plan executed to transform TACC into a ***center*** with R&D and EOT, not just a facility with resources and services (and vis, grid as well as HPC)
- ✓ Got national rep through TeraGrid award
- ✓ Hired awesome HPC staff, like Karl Schulz and Tommy Minyard
- ✓ Begged Dell for a good deal (call George Jones)
- ✓ Convinced UT System to fund new Lonestar

Onset of the Terascale & Petascale Eras

The Terascale Era

- June 1997:
  - "ASCI Red" (Sandia) entered TOP500 list of most powerful supercomputers at #1 with 1.07 TF/s
  - Held position until Nov. 2000
- As of Nov. 2006:
  - #1 is LLNL (NNSA/DOE) IBM Blue Gene: 367 Tflops peak, 280 Tflops HPC
  - #500 machine nearly 5 Tflops peak (nearly 3 Tflops HPL)
  - 11 countries represented in top 50
- IBM ThinkPad T43p would have been in the top 100 on first TOP500 list (1993)

The Case for More Powerful Computational Science Capabilities

- Many recent federally-commissioned reports have urged sustained, long-term U.S. investments in HPC to realize full benefits of computational science:
  - NSF: Cyberinfrastructure (2003)
  - DOE: Facilities for the Future of Science (2003)
  - NIH: Biomedical Information Science and Technology Initiative
  - Council on Competitiveness: Supercharging U.S. Innovation and Competitiveness (2004)
  - Interagency: High End Computing Revitalization Task Force (2004)
  - DOE: Capability Computing Needs (2004)
  - NAS: Getting up to Speed: The Future of Supercomputing (2005)
  - PITAC: Report on Computational Science (2005)
  - NSF: Simulation-Based Engineering Science (2005)

We Need Supercomputing

The Petascale Era

- DOE, NSF, and other US agencies now aggressively pursuing programs to deploy
  - peak petaflops systems *now*
  - *sustained* petaflops systems in the next 4 years
- A few US petascale projects
  - NSF Track2 systems deployed annually (2007-11)
  - DOE NNSA Roadrunner system @ LANL (1 PF+, 2008/09)
  - DOE Office of Science systems @ ORNL, ANL (1 PF, 2008/09)
  - NSF Track 1 Petascale Acquisition (10-20 PF, 2011)
- Cost of hardware/ & operations for NSF and DOE *sustained* petaflops systems alone: >$1B

Petascale Computing Opportunities

- Petascale will be here next year: up to science & engineering communities to make effective use
- Modeling and simulation can contribute significantly to making headway on many of the 'grand challenge' problems facing ***society*** as well as science:
  - future energy, climate change, environmental sustainability, clean water, natural disasters, neuroscience, drug design, predictive medicine, intelligent manufacturing, supply chain management, first principles materials design, etc.
- Petascale systems preset unprecedented capabilities, opportunities to make headway on many of the societal grand challenges

Petascale Computing Opportunities

- Raw throughput/memory will permit many enhancements to current “terascale” simulations:
  - Increased resolution
  - Greater fidelity of physics models
  - Inverse problem (a.k.a. model calibration, parameter estimation, data assimilation)
  - Uncertainty quantification
  - Optimization (design and control)
- Ultimately: simulation-based decision-making under uncertainty
  - Likely an exascale (zetascale, yottascale) computing problem for terascale deterministic forward problems

The Billion Dollar Question:

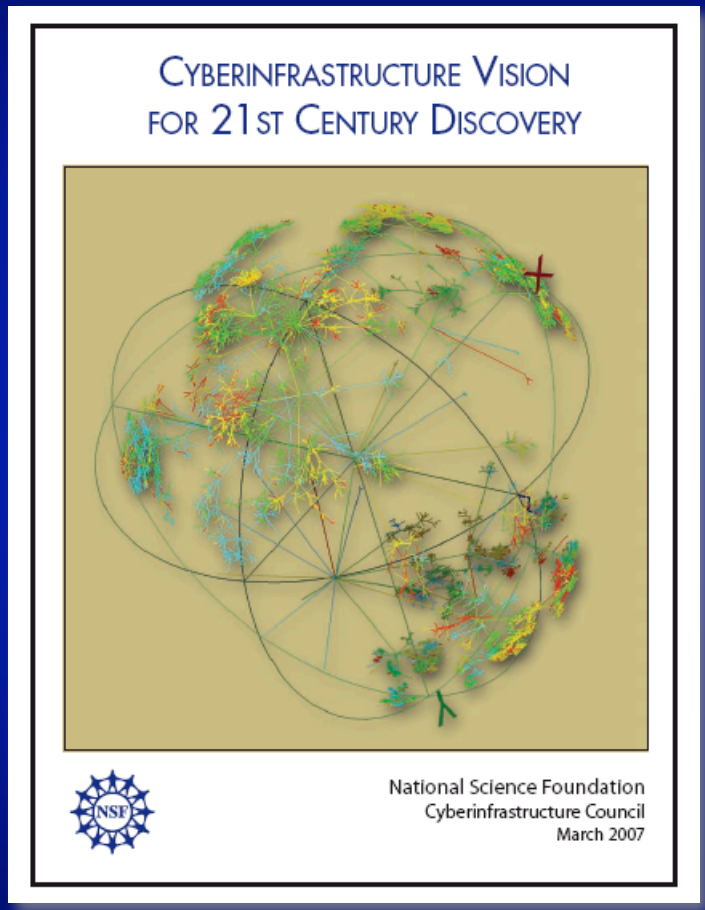
Will we be able to make effective use of PF systems?

- Enormous challenges for petascale computational science:
  - Mathematical models
  - Numerical approximations
  - Scalable numerical algorithms
  - Scalable geometric algorithms
  - Scientific visualization and data management
- Petascale computing challenges been underappreciated at agency levels for the past 15 years, still remain to be solved
  - Major troubles ahead unless sufficient resources are aimed at creating "scalable computational science"
- Indications of change - Example: NSF is planning a 5-year, $0.75B program: Cyber-enabled Discovery and Innovation (CDI)
  - Starting small though--first year only $26M

Ranger, a Bridge to Petascale Computing

NSF Cyberinfrastructure Strategic Plan

- NSF Cyberinfrastructure Strategic Plan released March 2007
  - Articulates importance of CI overall
  - Chapters on computing, data, collaboration, and workforce development
- NSF investing in world-class computing
  - Annual "Track2" HPC systems ($30M)
  - Single "Track1" HPC system in 2011 ($200M)
- Complementary solicitations for software, applications, education
  - Software Development for CI (SDCI)
  - Strategic Technologies for CI (STCI)
  - Petascale Applications (PetaApps)
  - CI-Training, Education, Advancement, Mentoring (CI-TEAM)
  - Cyber-enabled Discovery & Innovation (CDI) starting in 2008: $0.75B!

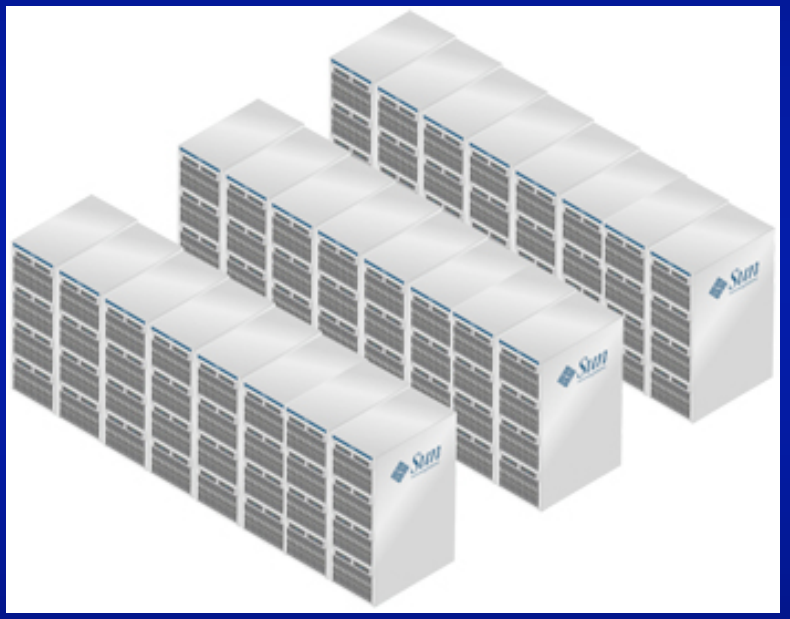


http://www.nsf.gov/od/oci/CI_Vision_March07.pdf

TACC

First NSF Track2 System: 1/2 Petaflop!

- TACC team won *first* NSF 'Track2' HPC system
  - $30M Sun system
  - 15,000+ quad-core AMD Opterons
  - 504 Tflops peak performance
  - 125 TB memory
  - 1.7 PB disk
  - 2.3 max μsec MPI latency
- TACC & ICES at UT Austin, plus Sun, Cornell & ASU operating/supporting system for 4 years ($29M)
- #1 HPC system in world in Dec 2007
  - Up to 5% allocable to Texas higher ed institutions (potentially)



TACC

Team Partners & Roles

- Institutions
  - TACC / UT Austin: project leadership, system hosting & operations, user support, technology evaluation/insertion, applications support
  - ICES / UT Austin: applications collaborations, algorithm/technique transfer and support
  - Cornell Theory Center: large-scale data management & analysis, on-site and remote training and workshops
  - Arizona State HPCI: technology evaluation/insertion, user support
- Roles
  - Project Director: Jay Boisseau (TACC)
  - Project Manager: Tommy Minyard (TACC)
  - Co-Chief Applications Scientists: Karl Schulz (TACC), Omar Ghattas (TACC), Giri Chukkapalli (Sun)
  - Chief Technologist: Jim Browne (ICES)

Ranger System Summary

- Compute power - 504 Teraflops
  - 3,936 Sun four-socket blades
  - 15,744 AMD Opteron “Barcelona” processors
    - Quad-core, 2.0 GHz, four flops/cycle (dual pipelines)
- Memory - 125 Terabytes
  - 2 GB/core, 32 GB/node
  - 132 GB/s aggregate bandwidth
- Disk subsystem - 1.7 Petabytes
  - 72 Sun x4500 “Thumper” I/O servers, 24TB each
  - ~72 GB/sec total aggregate bandwidth
  - 1 PB in largest /work filesystem
- Interconnect - 10 Gbps / 2.3 μsec latency
  - Sun InfiniBand-based switches (2) with 3456 ports each
  - Full non-blocking 7-stage Clos fabric
  - Mellanox ConnectX IB cards

TACC

Ranger System Summary (cont.)

- Management nodes - Sun x4600 4U Servers
  - 4 login nodes
  - 1 Rocks master
  - 2 SGE servers
  - 2 N1SM managers
  - 2 Subnet Managers
  - 6 Lustre Meta-Data Servers
  - 4 Archive data-movers
  - 4 external GridFTP servers
- Ethernet Networking - 10Gbps
  - Two external 10GigE networks: TeraGrid, NLR
  - 10GigE fabric for login, data-mover and GridFTP nodes, integrated into existing TACC network infrastructure
  - Force10 S2410 and E1200 switches

Ranger I/O Subsystem

- Disk Object Storage Servers (OSS) are based on Sun x4500 “Thumper” servers
  - Each server has 48 SATA II 500 GB drives (24TB total) - running internal software RAID
  - Dual Socket/Dual-Core Opterons @ 2.6 GHz
  - Downside is that these nodes have PCI-X - raw I/O bandwidth can exceed a single PCI-X 4X Infiniband HCA
  - 72 Servers Total: 1.7 PB raw storage
- Metadata Servers (MDS) based on Sun Fire x4600s
- MDS is Fibre-channel connected to 9TB Flexline Storage
- Target Performance
  - Aggregate bandwidth: 70+ GB/sec
  - To largest $WORK filesystem: 40 GB/sec

Design:
Top loading Disks
Front to rear airflow
Redundant fans
Passive Backplane
No wires in box

Reliability/Availability
Enterprise class SATA disks
1M hours MTBF
RAID 0, 1, 5, 10
Redundant Power
Hot-swap FRUs

TACC

Ranger Speeds & Feeds

	Originally Proposed System	Final System
Compute Node Metrics		
Total # of Compute Nodes	3288	3936
Total # of Processing Cores	52,608	62,976
Total Peak Flops	421 TFlops	529 TFlops
Parallel Filesystem Metrics		
Total Raw Disk Capacity	1.73 PB	1.73 PB
Disk I/O Bandwidth	40 GB/s	40 GB/s
Distributed Memory Metrics		
Total Memory	105 TB	125 TB
Total Memory Bandwidth	110 TB/s	132 TB/s

Note: peak performance target now 504 Tflops

System ratios	Bisection BW (B:F)	0.015
	Memory BW (B:F)	0.238
	Memory size (B:F)	0.233

TACC

Ranger Space, Power and Cooling

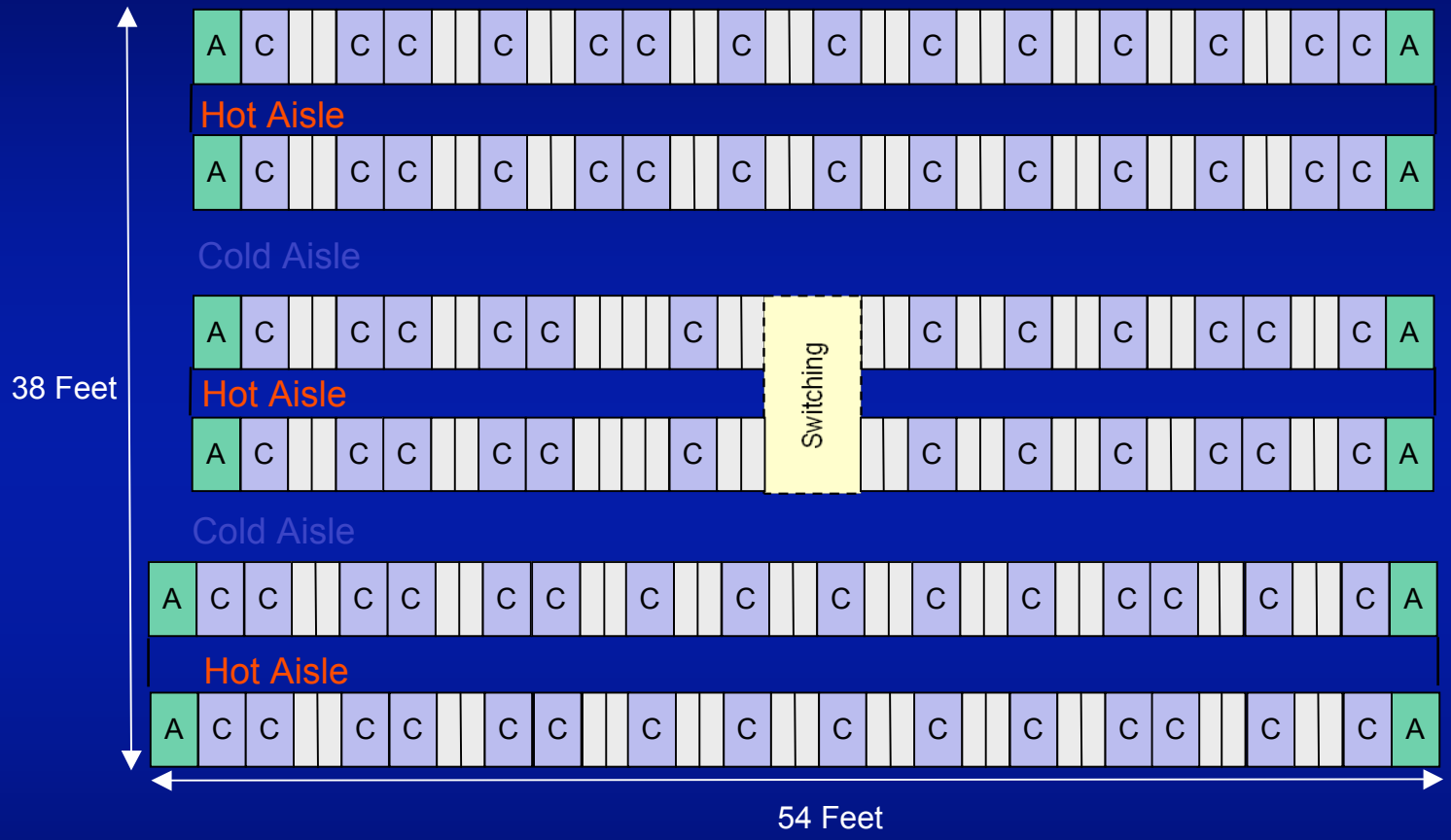
- Total Power: 3.4 MW!
- System: 2.4 MW
  - 94 racks--2 compute, 12 support, plus switch
  - 120 in-row cooling units
  - 2054 sqft total footprint (~4000 sqft including PDUs)
- Cooling: ~1 MW
  - In-row units fed by three 400-ton chillers (N+1)
  - Enclosed hot-aisles
  - Supplemental 280-tons of cooling from CRAC units
- Observations:
  - Space less an issue than power
  - Cooling > 25kW per rack difficult
  - Power distribution a challenge, more than 1284 circuits

Ranger: In-Row Coolers and PDUs

- APC in-row coolers (black) are used for primary cooling. Compute racks will slide in between the coolers, which are heat exchangers that draw from the hot aisles and circulate ambient into the cold aisles
- 20+ PDUs (beige) to power the 80+ racks

TACC

Ranger System Layout



C — 82 Compute Racks
3,936 Blades,
15,744 CPUs , 126 TB RAM

A — 12 Disk Racks
72 X4500 (1.7 PB)
25 X4600 (Support/Metadata)

APC In-Row Coolers

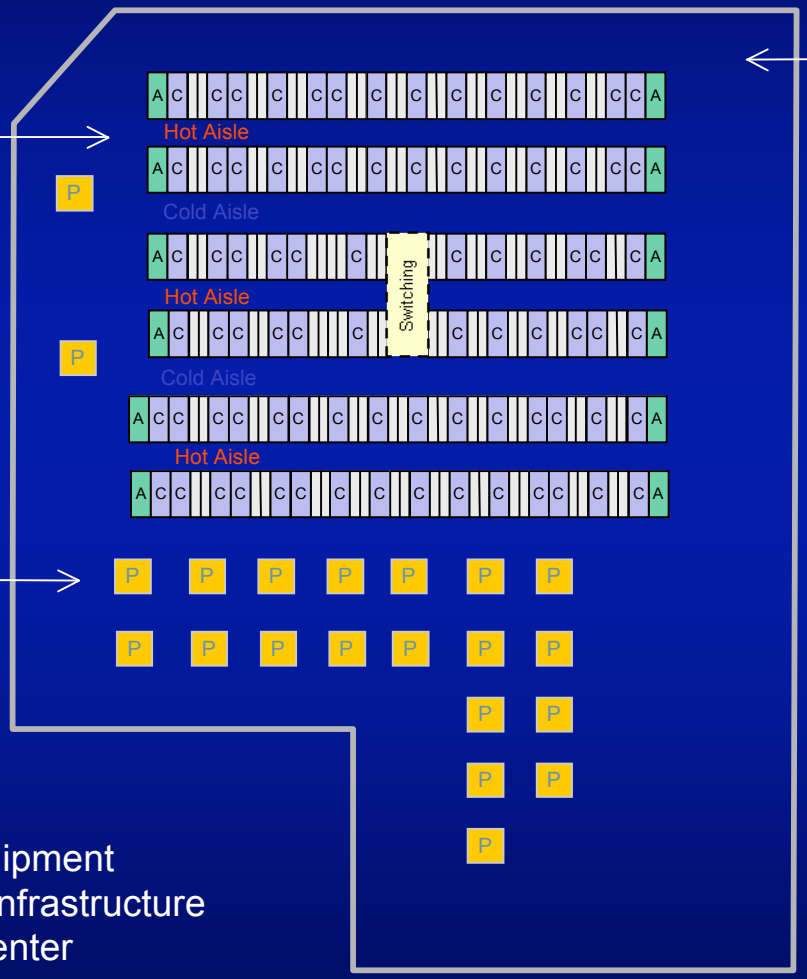
TACC

Ranger Data Center

APC Hot-aisle containment system

3 x 400 Ton Chillers
7 x 40 Ton CRAC

21 x PDUs
64 x 20A Circuits
~1,300 Circuits total



30” Raised Floor delivers:
- Power, networking, and chilled water
-Base cooling and humidity control

30 Opteron Cores, 60 GB of RAM, and 850GB disk per sq-ft

1.2kW per sq-ft (including cooling equipment)

~ 2,000 sq-ft for equipment
~ 4,500 sq-ft for all infrastructure
~ 6,400 sq-ft data center

TACC

Ranger Project Costs

- NSF Award: $59M
  - Purchases full system, plus initial test equipment
  - Includes 4 years of system maintenance
  - Covers 4 years of operations and scientific support
- UT Austin providing power: $1M/year
- UT Austin upgraded data center infrastructure: $10-15M
- TACC upgrading storage archival system: $1M
- Total cost $75-80M
  - Thus, system cost > $50K/operational day
  - Must enable user to conduct world-class science every day!

TACC

Ranger User Environment

- ***Ranger*** user environment will be similar to ***Lonestar***
  - Full Linux OS on nodes
    - 2.6.18 is starting working kernel
    - hardware counter patches on login and compute nodes
    - *Rocks* used to provision nodes
  - Lustre File System
    - $HOME and two $WORK filesystems will be available
    - Largest $WORK will be ~1PB total
  - Standard 3rd party packages
  - InfiniBand using next generation of Open Fabrics
  - MVAPICH and OpenMPI (MPI1 and MPI2)

Ranger User Environment

- Suite of compilers
  - Portland Group PGI
  - Sun Studio
  - PathScale
  - *Possibly Intel compilers*
- Batch System
  - One main difference will be the batch system environment - *Ranger* using SGE (Grid Engine)
  - Providing standard scheduling options: backfill, fairshare, advanced reservations
- Baseline Libraries
  - **ACML**, AMD core math library
  - **GotoBLAS**, high-performance BLAS
  - **PETSc**, sparse linear algebra
  - **metis/pmetis**, graph bisection
  - **tau/pdtoolkit**, profiling toolkit
  - **sprng**, parallel random number generators
  - **papi**, performance application programming interface
  - **netcdf**, portable I/O routines
  - **hdf**, portable I/O routines
  - **fftw**, open-source fft routines
  - **scalapack/plapack**, linear algebra
  - **slepc**, eigenvalue problems

Ranger System Configuration

At this scale, parallel file systems are universally required
Lustre and Sun X4500's are used for all volumes

Logical Volume Name	Estimated Raw Capacity	Target Usage
WORK1	1 PB	Large temporary storage; not backed up, purged periodically
WORK2	~500 TB	Large allocated storage; not backed up, quota enforced
PROJECTS	2 TB	Repository for TeraGrid Community Software
HOME1	2 TB	Permanent user storage; automatically backed up, quota enforced
HOME2	2 TB	Permanent user storage; automatically backed up, quota enforced
HOME3	2 TB	Permanent user storage; automatically backed up, quota enforced

TACC

User Support Challenges

- NO systems like this exist yet!
  - Will be the first general-purpose system at ½ Pflop
  - Quad-core, massive memory/disk, etc.
- NEW user support challenges
  - Code optimization for quad-core, 16-way nodes
  - Extreme scalability to 10K+ cores
  - Petascale data analysis
  - Tolerating faults while ensuring job completion

User Support Plans

- User support: ‘usual’ (docs, consulting, training) plus
  - User Committee dedicated to this system
    - Active, experienced, high-end users
  - Applications Engineering
    - algorithmic consulting
    - technology selection
    - performance/scalability optimization
    - data analysis
  - Applications Collaborations
    - Partnership with petascale apps developers and software developers

User Support Plans

- Also
  - Strong support of ‘professionally optimized’ software
    - Community apps
    - Frameworks
    - Libraries
  - *Additional* Training
    - On-site at TACC, partners, and major user sites, and at workshops/conferences
    - Advanced topics in multi-core, scalability, etc
    - Virtual workshops for remote learning
  - Increased communications and technical exchange with all users via a TACC User Group

Technology Insertion Plans

- Technology Identification, Tracking, Evaluation, and Insertion are crucial to improving a $50K/day system!
  - Cutting edge system: software won't be mature
  - Four year lifetime: new R&D will produce better technologies
  - Improve system: maximize impact over lifecycle
- Chief Technologist for project, plus supporting staff
  - Must build communications, partnerships with leading software developers worldwide
  - Grant doesn't fund R&D, but system provides unique opportunity for determining, conducting R&D!
  - Targets include: fault tolerance, algorithms, next-generation programming tools/languages, etc.

Ranger Project Timeline

Sep06	award, press, relief, beers
1Q07	equipment begins arriving
2Q07	facilities upgrades complete
2Q-3Q07	construction of system
4Q07	early users
Dec07	production, many beers
Jan08	allocations begin

TACC

Applications Performance Notes

- Obviously, no data for final system
  - We finally have parts, but it's not built yet! (~30 days)
- Performance predictions can be made from previous & pre-production versions, prototypes, single procs and blades, etc.
- Applications performance projections for NSF benchmarks look **very** promising
- Expect some applications to sustain 50-100+ Tflops
  - On very large problem sizes: up to 100 TB!
- *Much more to say at SC07!*

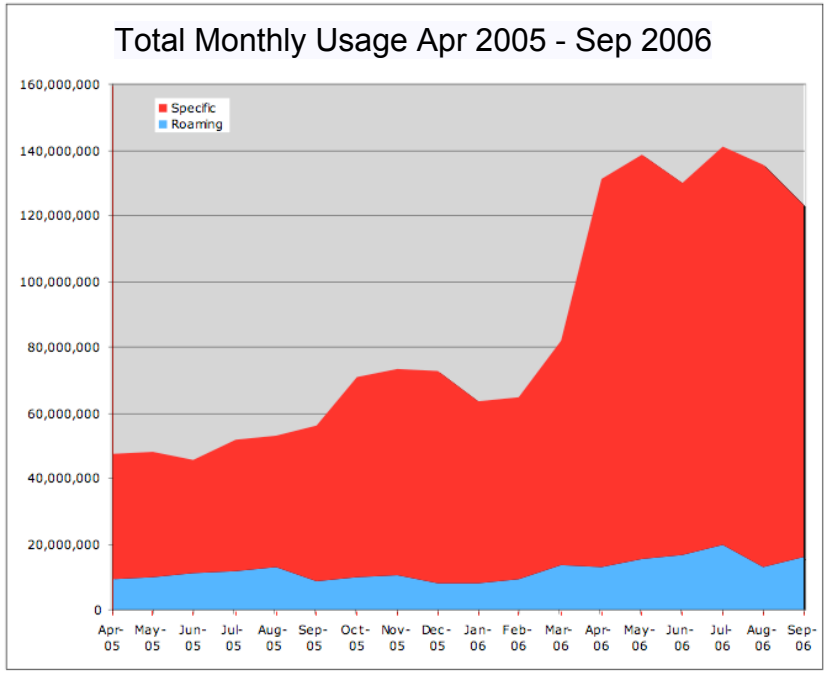
TACC

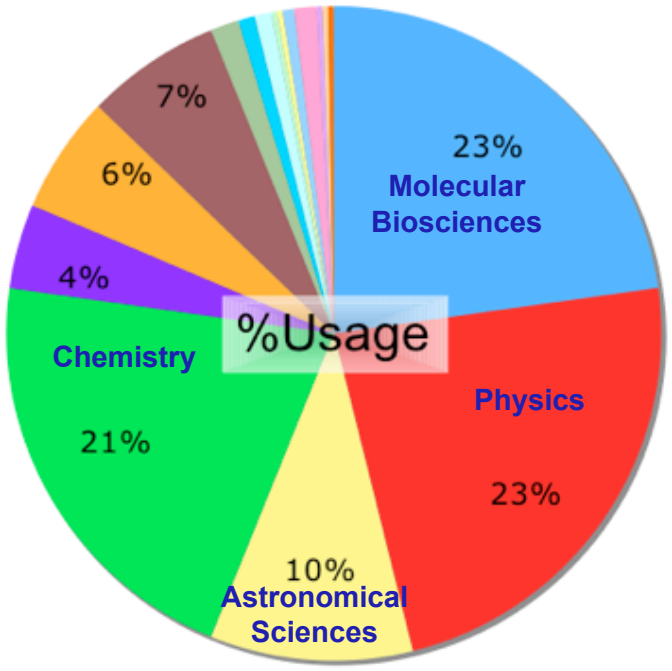
Impact in TeraGrid, US

- 472M CPU hours to TeraGrid: more than double current total of all TG HPC systems
- 504 Tflops : over 5x current top system
- Enable unprecedented research
- Re-establish NSF as a leader in HPC
- *Jumpstarts progress to petascale for entire US academic research community*
- *But it's a step-function machine--will take time for users to scale up*

TACC

Who Might Use Ranger?
Recent TeraGrid HPC Usage by Domain



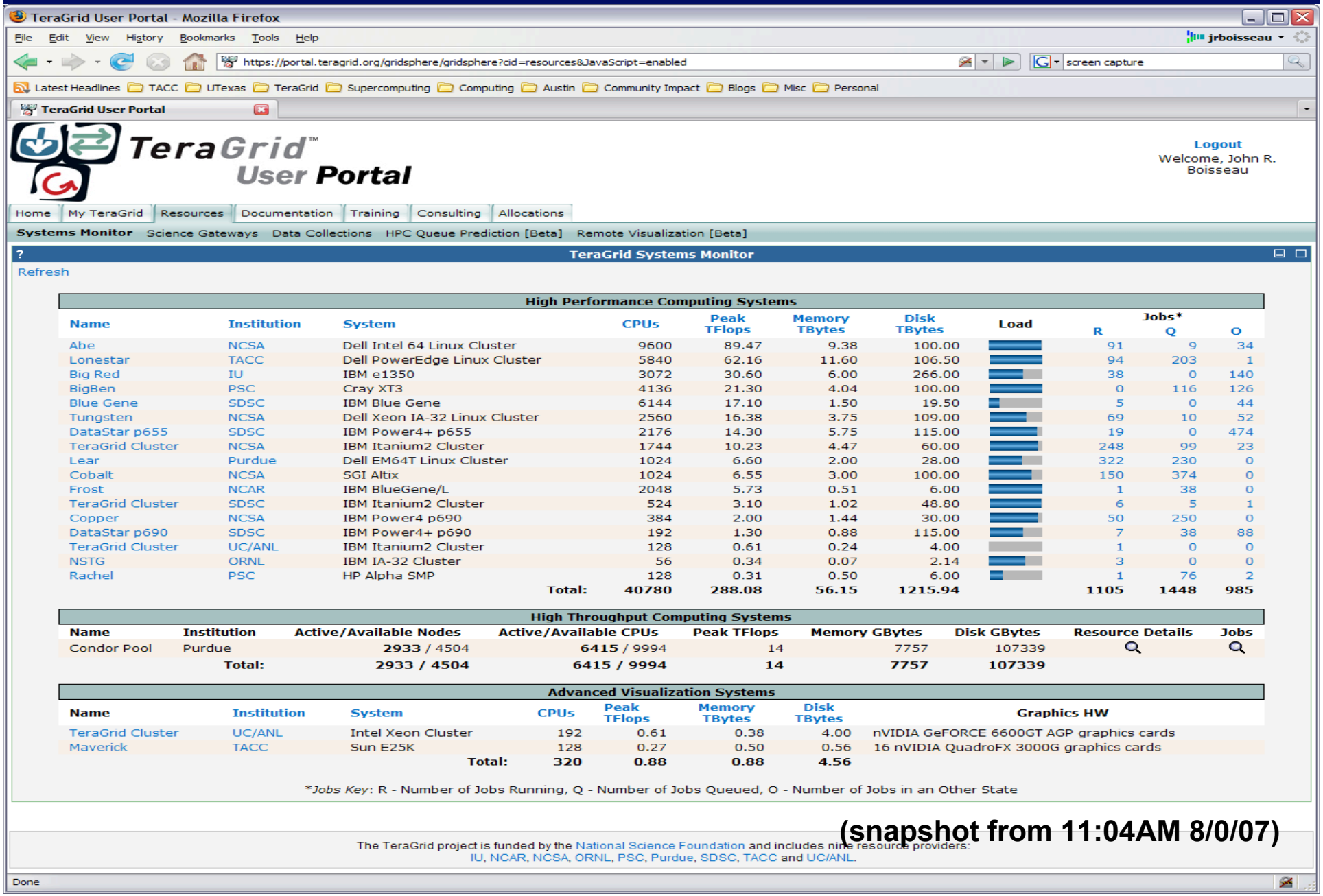


- Molecular Biosciences
- Physics
- Astronomical Sciences
- Chemistry
- Materials Research
- Chemical, Thermal Systems
- Atmospheric Sciences
- Advanced Scientific Computing
- Earth Sciences
- Biological and Critical Systems
- Ocean Sciences
- Cross-Disciplinary Activities
- Computer and Computation Research
- Integrative Biology and Neuroscience
- Mechanical and Structural Systems
- Mathematical Sciences
- Electrical and Communication Systems, Design and Manufacturing Systems, Environmental Biology

TACC

1000+ projects, thousands of users

Current TeraGrid HPC Systems



TeraGrid User Portal - Mozilla Firefox

File Edit View History Bookmarks Tools Help — jrboisseau

https://portal.teragrid.org/gridsphere/gridsphere?cid=resources&JavaScript=enabled

Latest Headlines | TACC | UTexas | TeraGrid | Supercomputing | Computing | Austin | Community Impact | Blogs | Misc | Personal

TeraGrid User Portal

TeraGrid™ User Portal

Logout
Welcome, John R. Boisseau

Home | My TeraGrid | Resources | Documentation | Training | Consulting | Allocations

Systems Monitor | Science Gateways | Data Collections | HPC Queue Prediction [Beta] | Remote Visualization [Beta]

TeraGrid Systems Monitor

Refresh

High Performance Computing Systems

Name	Institution	System	CPUs	Peak TFlops	Memory TBytes	Disk TBytes	Load	Jobs* R	Q	O
Abe	NCSA	Dell Intel 64 Linux Cluster	9600	89.47	9.38	100.00		91	9	34
Lonestar	TACC	Dell PowerEdge Linux Cluster	5840	62.16	11.60	106.50		94	203	1
Big Red	IU	IBM e1350	3072	30.60	6.00	266.00		38	0	140
BigBen	PSC	Cray XT3	4136	21.30	4.04	100.00		0	116	126
Blue Gene	SDSC	IBM Blue Gene	6144	17.10	1.50	19.50		5	0	44
Tungsten	NCSA	Dell Xeon IA-32 Linux Cluster	2560	16.38	3.75	109.00		69	10	52
DataStar p655	SDSC	IBM Power4+ p655	2176	14.30	5.75	115.00		19	0	474
TeraGrid Cluster	NCSA	IBM Itanium2 Cluster	1744	10.23	4.47	60.00		248	99	23
Lear	Purdue	Dell EM64T Linux Cluster	1024	6.60	2.00	28.00		322	230	0
Cobalt	NCSA	SGI Altix	1024	6.55	3.00	100.00		150	374	0
Frost	NCAR	IBM BlueGene/L	2048	5.73	0.51	6.00		1	38	0
TeraGrid Cluster	SDSC	IBM Itanium2 Cluster	524	3.10	1.02	48.80		6	5	1
Copper	NCSA	IBM Power4 p690	384	2.00	1.44	30.00		50	250	0
DataStar p690	SDSC	IBM Power4+ p690	192	1.30	0.88	115.00		7	38	88
TeraGrid Cluster	UC/ANL	IBM Itanium2 Cluster	128	0.61	0.24	4.00		1	0	0
NSTG	ORNL	IBM IA-32 Cluster	56	0.34	0.07	2.14		3	0	0
Rachel	PSC	HP Alpha SMP	128	0.31	0.50	6.00		1	76	2
		Total:	**40780**	**288.08**	**56.15**	**1215.94**		**1105**	**1448**	**985**

High Throughput Computing Systems

Name	Institution	Active/Available Nodes	Active/Available CPUs	Peak TFlops	Memory GBytes	Disk GBytes	Resource Details	Jobs
Condor Pool	Purdue	**2933** / 4504	**6415** / 9994	14	7757	107339		
	Total:	**2933 / 4504**	**6415 / 9994**	**14**	**7757**	**107339**		

Advanced Visualization Systems

Name	Institution	System	CPUs	Peak TFlops	Memory TBytes	Disk TBytes	Graphics HW
TeraGrid Cluster	UC/ANL	Intel Xeon Cluster	192	0.61	0.38	4.00	nVIDIA GeFORCE 6600GT AGP graphics cards
Maverick	TACC	Sun E25K	128	0.27	0.50	0.56	16 nVIDIA QuadroFX 3000G graphics cards
		Total:	**320**	**0.88**	**0.88**	**4.56**	

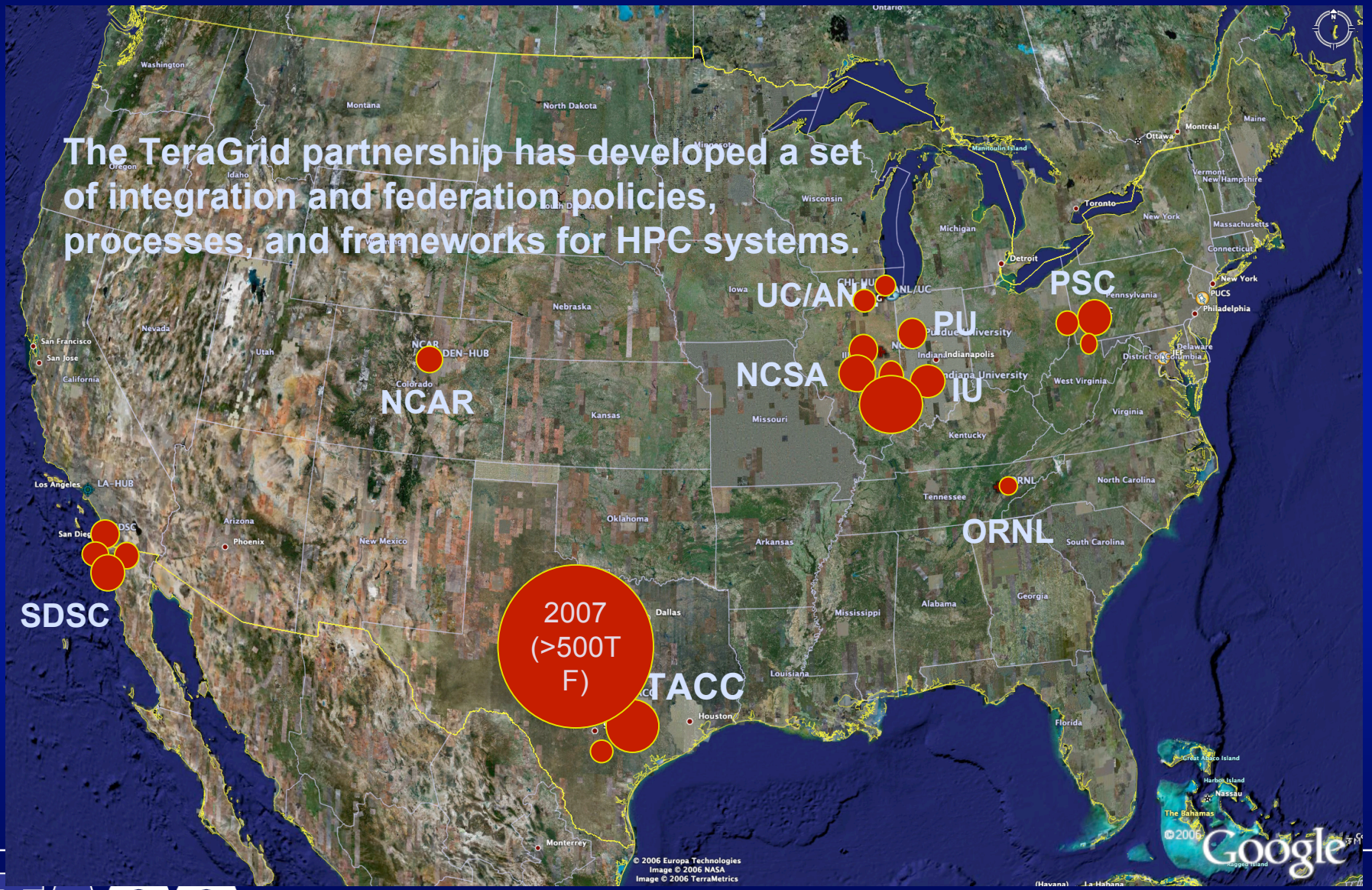
**Jobs Key*: R - Number of Jobs Running, Q - Number of Jobs Queued, O - Number of Jobs in an Other State

The TeraGrid project is funded by the National Science Foundation and includes nine resource providers:
IU, NCAR, NCSA, ORNL, PSC, Purdue, SDSC, TACC and UC/ANL.

Done

(snapshot from 11:04AM 8/0/07)

TeraGrid HPC Systems plus Ranger



How We Built Ranger (almost done)

- ✓ Acquired previous clusters and proved value, impact, manageability, etc.,
- ✓ Became experts in large-scale clusters with previous Lonestar
- ✓ Continued execution of vision and plan to transform TACC into a ***leading center*** with R&D and EOT, not just a facility with resources and services (and vis, grid as well as HPC)
- ✓ ***Earned*** reputation and experience through TeraGrid award execution
- ✓ Hired more awesome HPC staff
- ✓ Collaborated with awesome leaders like Jim Browne and Omar Ghattas/UT ICES, Dan Stanzione/ASU, Dave Lifka/Cornell
- ✓ Partnered with Sun for great, unique system (call Giri Chukkapalli)
- ✓ Convinced NSF panel that we are now **leaders**

TACC

Closing Thoughts and Summary

Some “Peta-Challenges”

- Achieving performance on many-core
  - Processor/memory bandwidth gap increasing
- Scalable algorithms
  - To 10K-100K+ cores
  - also, must be effectively implemented
- Scalable programming tools
  - debuggers, optimization tools, libraries, etc.
- Fault tolerance
  - Increased dependence on commodity (MTBF/node not changing) and increased number of nodes -> uh oh!

Petascale Data Analysis Challenge

- Data analysis ‘in the box’
  - Data will be too big to move (network, file system bandwidths not keeping pace)
  - Analyze in simulation if able
  - Or at least analyze while data still in HPC parallel file system
  - Must develop CPU-based scalable techniques
  - Or must develop better packaging for GPUs, include on more nodes

What's Next for NSF HPC?

- Track2 system awards every year
- Track1 award made this year for 2011-2016
  - To ***sustain*** 1 PF (so peak > 10 PF)
- SDCI and STCI awards for HPC software tools
- PetaApps awards for accelerating petascale applications development
- CI-TEAM awards to encourage participation in high-end computational science
- DataNet awards for hosting large scientific collections
- CDI starting in 2008: *major* NSF-wide initiative!

Petascale Power Challenge

- Power constraints--generation and distribution--limit number and location of petascale computing centers
  - Remember: flops/watt is getting better, but we're building much larger systems!
  - Petascale system power budgets will be more than staffing!
  - But HPC expertise becomes even more important than hosting expertise due to other challenges
- Upshot: petascale HPC systems won't be everywhere; consolidation, location make sense for cost of operations

Some Predictions

- Next NSF Track2 is also homogeneous, but 3rd or 4th will not (some Cell, GPGPU, or…)
  - But not solely Cell or GPGPU at petascale!
  - Los Alamos building hybrid petascale Opteron-Cell system in 2008!
- Commodity switches will increase in port count greatly (thousand-way+) very soon (2008?)
- Serious *community* efforts on optimizing Linux for many-core compute nodes (not just vendor-specific)
- Lightweight checkpoint restart for Linux clusters
- Leading centers limited by location, infrastructure, but become islands: host compute, data, vis, etc.

Summary

- Push to petascale is driving HPC vendors like the push to 1 GHz drove AMD, Intel
- NSF is again a leader in HPC, as a component of world-class cyberinfrastructure
- *Ranger and other petascale systems will enable unprecedented high-resolution, high-fidelity, multi-scale, multi-physics applications*

Thanks To…

- The National Science Foundation (NSF) for giving TACC the opportunity to deploy Ranger and help the science community move to petascale computing
- Omar Ghattas, Charlie Catlett, Karl Schulz and Tommy Minyard for many contributions to this presentation
- The awesome TACC staff for making this wild ride incredibly fun (and thus tolerable)
- You for your tax dollars, attention, and hopefully future support/collaboration!

TACC

The University of Texas at Austin
Distinguished Lecture Series in Petascale Computation

- Web accessible: http://www.tacc.utexas.edu/petascale/
- Past Lectures
  - “Petaflops, Seriously,” Dr. David Keyes, Columbia University
  - “Discovery through Simulation: The Expectations of Frontier Computational Science,” Dr. Dimitri Kusnezov, National Nuclear Security Administration
  - “Modeling Coastal Hydrodynamics and Hurricanes Katrina and Rita,” Dr. Clint Dawson, The University of Texas at Austin
  - “Towards Forward and Inverse Earthquake Modeling on Petascale Computers,” Dr. Omar Ghattas, The University of Texas at Austin
  - "Computational Drug Diagnostics and Discovery: The Need for Petascale Computing in the Bio-Sciences,” Dr. Chandrajit Bajaj, The University of Texas at Austin
  - "High Performance Computing and Modeling in Climate Change Science,” Dr. John Drake, Oak Ridge National Laboratory
  - "Petascale Computing in the Biosciences - Simulating Entire Life Forms," Dr. Klaus Schulten, University of Illinois at Urbana-Champaign
- Suggestions for future speakers/topics welcome

TACC