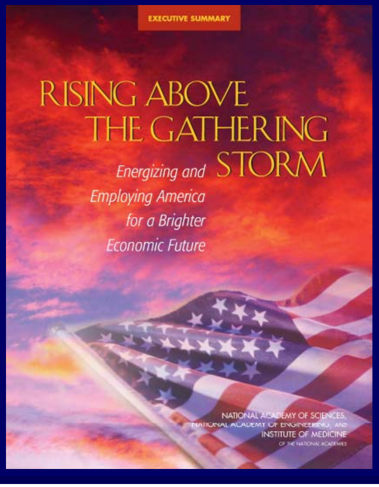
Ranger: Providing a Path to Petascale Computing In Texas!

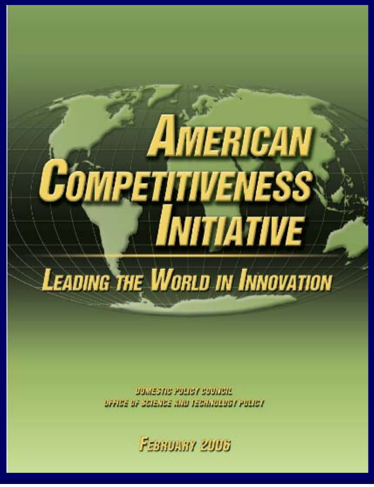
Jay Boisseau, Director
Texas Advanced Computing Center
The University of Texas at Austin

HiPCAT Meeting
February 29, 2008

Context: The Case for More Powerful Computational Science Capabilities

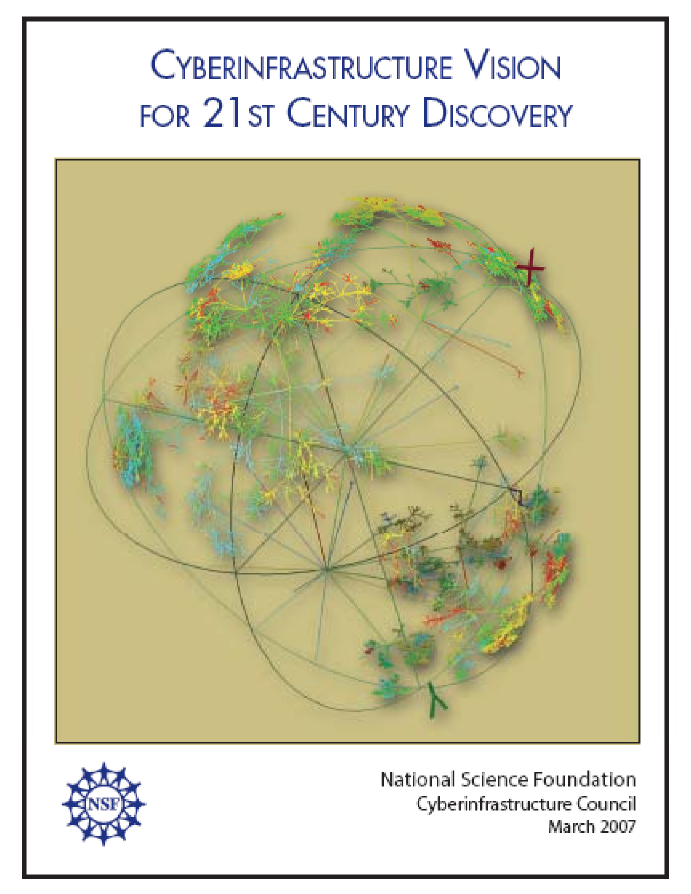
- National Academies' "Rising Above the Gathering Storm" report urges reinvestment in Science/Technology/Engineering/Math
- American Competitiveness Initiative calls for doubling of NSF, DOE/SC, NIST budgets over 10 years; largest federal response since Sputnik
- NSF 5-year Strategic Plan fosters research to further U.S. economic competitiveness by focusing on fundamental science & engineering


EXECUTIVE SUMMARY
RISING ABOVE THE GATHERING STORM
Energizing and Employing America for a Brighter Economic Future
NATIONAL ACADEMY OF SCIENCES, NATIONAL ACADEMY OF ENGINEERING, AND INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES


AMERICAN COMPETITIVENESS INITIATIVE
LEADING THE WORLD IN INNOVATION
DOMESTIC POLICY COUNCIL
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
FEBRUARY 2006

Context: The 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!



CYBERINFRASTRUCTURE VISION
FOR 21ST CENTURY DISCOVERY

NSF

National Science Foundation
Cyberinfrastructure Council
March 2007

available for download at NSF web site

First NSF Track2 System: 1/2 Petaflop!

- TACC selected for first NSF ‘Track2’ HPC system
  - $30M system acquisition
  - Sun Constellation Cluster
  - AMD Opteron processors
  - Expandable configuration

- Project includes 4 years operations and support
  - System maintenance
  - User support
  - Technology insertion
  - $29M budget

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 Center for Advanced Computing: 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: Chief System Engineer (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 - 123 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 / ~3 μsec latency
  - Sun InfiniBand-based switches (2) with 3456 ports each
  - Full non-blocking 7-stage Clos fabric
  - Mellanox ConnectX IB cards

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
- Texas support:
  - UT Austin providing power: up to $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 users to conduct world-class science every day!*
- Texas cost: NSF allowed TACC to allocate 5% of cycles to Texas higher education

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
  - Intel
  - Sun Studio

- Batch System
  - *Ranger* using SGE (Grid Engine) instead of LSF
  - 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
  - **netcdf**, portable I/O routines
  - **hdf**, portable I/O routines
  - **fftw**, open-source fft routines
  - **scalapack/plapack**, linear algebra
  - **slepc**, eigenvalue problems
  - **sprng**, parallel random number generators
  - **papi**, performance application programming interface

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
SCRATCH	800 TB	Large temporary storage; not backed up, purged periodically
WORK	200 TB	Large allocated storage; not backed up, quota enforced
PROJECTS	2 TB	Repository for TeraGrid Community Software
HOME1	50+ TB	Permanent user storage; automatically backed up, quota enforced
HOME2	50+ TB	Permanent user storage; automatically backed up, quota enforced
HOME3	50+ TB	Permanent user storage; automatically backed up, quota enforced

Technology Insertion Plans

- Technology Identification, Tracking, Evaluation, and Insertion are crucial
  - 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.

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

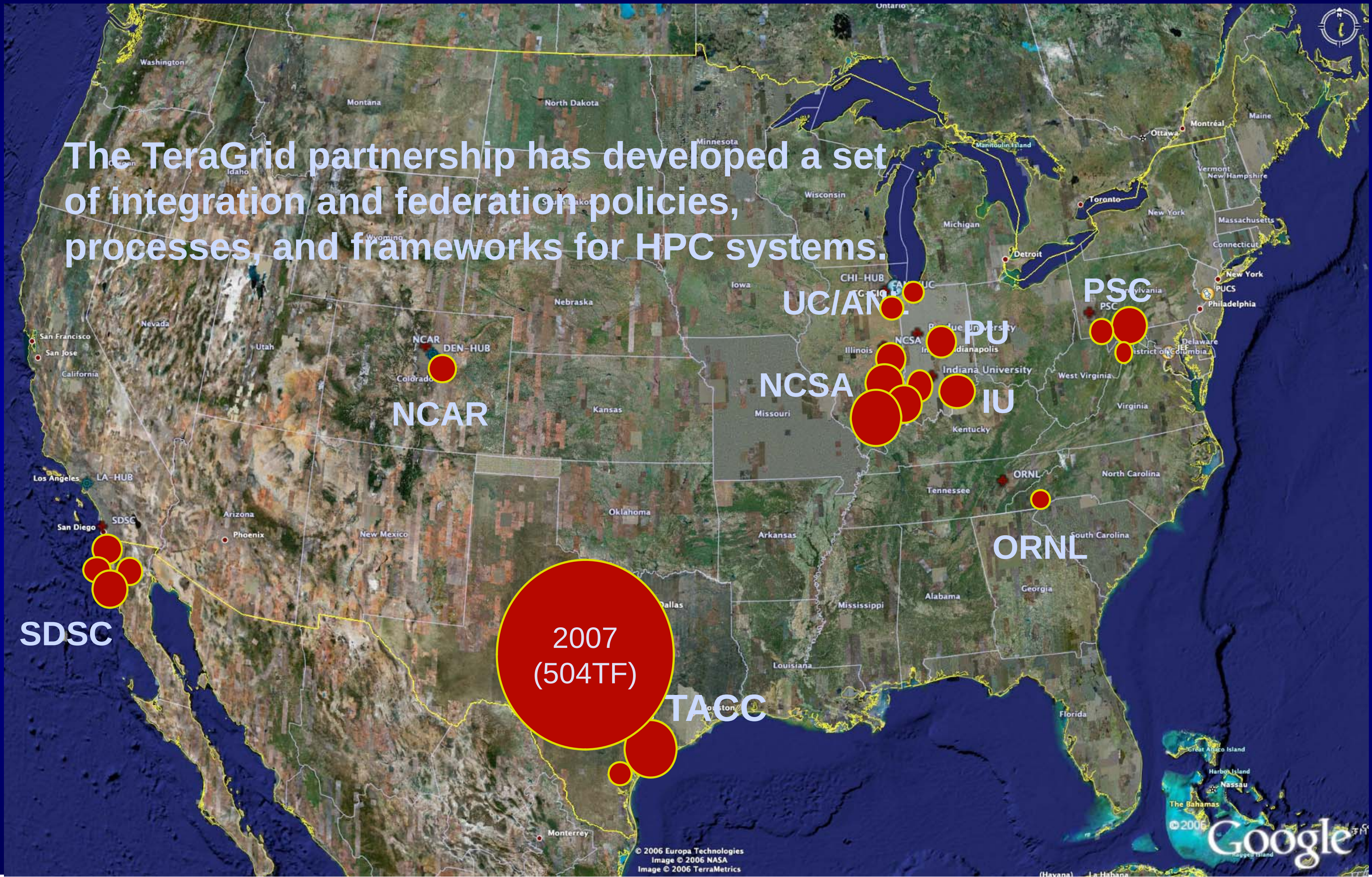
Impact in TeraGrid

- 472M CPU hours to TeraGrid
  - more than sum of *all* current TG HPC systems
- 504+ Tflops
  - 5x current top system
- Enable unprecedented research
  - *Jumpstart progress to petascale for entire US academic research community*
  - Re-establish NSF as a leader in HPC

Current TeraGrid HPC Systems

QuickTime™ and a
decompressor
are needed to see this picture.

TeraGrid HPC Systems plus Ranger



The TeraGrid partnership has developed a set of integration and federation policies, processes, and frameworks for HPC systems.

© 2006 Europa Technologies
Image © 2006 NASA
Image © 2006 TerraMetrics

● Computational Resources (size approximate - not to scale)

Impact on Science

- TeraGrid resources are available to all researchers at US institutions in all disciplines
- Ranger will enable researchers to attack problems heretofore much too large for TG
- Already seeing applications in astronomy, biophysics, climate/weather, earthquake modeling, CFD/turbulence, and more scale to 1000s of cores
- Just went into production on Monday Feb 4--much more to say very soon!

How Does This Help Texas?

- TACC may allocate *up to 5%* of the cycles (26M CPU hours!) to Texas higher ed institutions
- Allocations requests must be submitted to TACC
- Review/decisions will be based on
  - Research/education merit
  - Team capability/expertise for using system
  - Opportunity for impact in Texas
  - Level of support needed

How Do Texans Apply?

- Apply through the TACC User Portal (portal.tacc.utexas.edu) after March 3 but before March 21
- Future deadlines will be one month before beginning of quarter (March 1, June 1, September 1, December 1)
- Instructions are on the TACC user portal

What Kinds of Allocations?

- Research
  - Default: Up to 500K CPU hours
  - Last for one year
  - Can request up to 1M by special arrangement
- Education
  - Up to 100K hours
  - Last for 2 quarters
- Startup
  - Up to 50K hours
  - Last for 1 quarter
  - Used for gaining expertise, preparing larger requests
  - May be repeated once

What Kind of Support?

- Ranger documentation available on TACC web site and via user portal
- Training
  - TACC teaches classes in Austin
  - Can teach classes at remote location if enough students, adequate facilities
  - Online training available in March on portal, from Cornell
- Helpdesk support available via TACC Consulting system on portal
  - There is no funding for extra support for non-TeraGrid usage--we're having to take it out of our hide, so be gentle!

Summary

- NSF determined to be a leader in petascale computing as component of world-class CI
- TACC determined to be a leading in providing advanced computing technologies to national community, but with emphasis on Texas!
- Ranger is available for Texas researchers on April 1 (no joke!), with requests accepted after March 3 and due by March 21