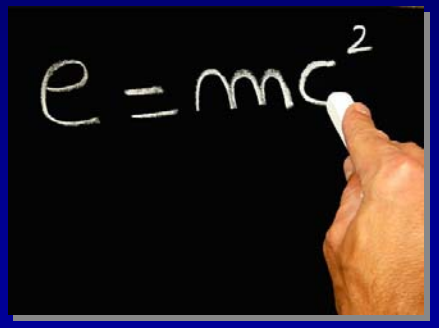
Managing Terascale Systems and Petascale Data Archives

February 26, 2010

Tommy Minyard, Ph.D.
Director of Advanced Computing Systems

Motivation: What's all the high performance computing fuss about?



- It's about that pesky stuff you learn in math and physics class
- As the Universe expanded and cooled, atomic particles were created, the forces of nature 'split,' and the galaxies and stars formed.
- The resulting Universe is governed by mathematical equations.
- Understanding the Universe means being able to describe it and predict its behavior.
- Our mathematical 'language' for describing and predicting the behavior of physical systems is ***calculus and differential equations***.
- Determining the theories and governing equations requires observation or experimentation, and testing hypotheses.

THE GRAND CHALLENGE EQUATIONS

$$B_i A_i = E_i A_i + \rho_i \sum_j B_j A_j F_{ji}$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\vec{F} = m\,\vec{a} + \frac{dm}{dt}\vec{v}$$

$$dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV$$

$$\nabla \cdot \vec{D} = \rho$$

$$Z = \sum_j g_j\, e^{-E_j/kT}$$

$$F_j = \sum_{k=0}^{N-1} f_k e^{2\pi i jk/N}$$

$$\nabla^2 u = \frac{\partial u}{\partial t}$$

$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}$$

$$p_{n+1} = r\, p_n\, (1 - p_n)$$

$$\nabla \cdot \vec{B} = 0$$

$$P(t) = \frac{\sum_i W_i\, B_i(t)\, P_i}{\sum_i W_i\, B_i(t)}$$

$$-\frac{h^2}{8\pi^2 m}\nabla^2 \Psi(r,t) + V\,\Psi(r,t) = -\frac{h}{2\pi i}\frac{\partial \Psi(r,t)}{\partial t}$$

$$-\nabla^2 u + \lambda\, u = f$$

$$\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla)\,\vec{u} = -\frac{1}{\rho}\nabla p + \gamma\,\nabla^2 \vec{u} + \frac{1}{\rho}\vec{F}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f$$

• NEWTON'S EQUATIONS • SCHROEDINGER EQUATION (TIME DEPENDENT) • NAVIER-STOKES EQUATION •
• POISSON EQUATION • HEAT EQUATION • HELMHOLTZ EQUATION • DISCRETE FOURIER TRANSFORM •
• MAXWELL'S EQUATIONS • PARTITION FUNCTION • POPULATION DYNAMICS •
• COMBINED 1ST AND 2ND LAWS OF THERMODYNAMICS • RADIOSITY • RATIONAL B-SPLINE •

[Courtesy of San Diego Supercomputer Center]

TACC

Supercomputing and Society

- Supercomputing is not just for science any more!
- Every one of you benefited from supercomputing today in multiple ways.
- Supercomputing is used daily in automotive and aerospace engineering, pharmaceutical development, financial modeling, medical research, homeland defense… and film!

TACC

TACC Mission & Strategy

The mission of the Texas Advanced Computing Center is to enable scientific discovery and enhance society through the application of advanced computing technologies.

To accomplish this mission, TACC:

Resources & Services

- Evaluates, acquires & operates advanced computing systems
- Provides training, consulting, and documentation to users

Research & Development

- Collaborates with researchers to apply advanced computing techniques
- Conducts research & development to produce new computational technologies

TACC

TACC Background

- Operating as an Advanced Computing Center since 1986, Organized Research Unit of UT Austin
- More than 80 Employees at TACC
  - 15 Ph.D. level research staff
  - 5 Senior system administrators with > 15 years experience running HPC systems
  - Graduate and undergraduate students
- Currently support thousands of users on multiple production systems

TACC

TACC Resources are Terascale, Comprehensive and Balanced

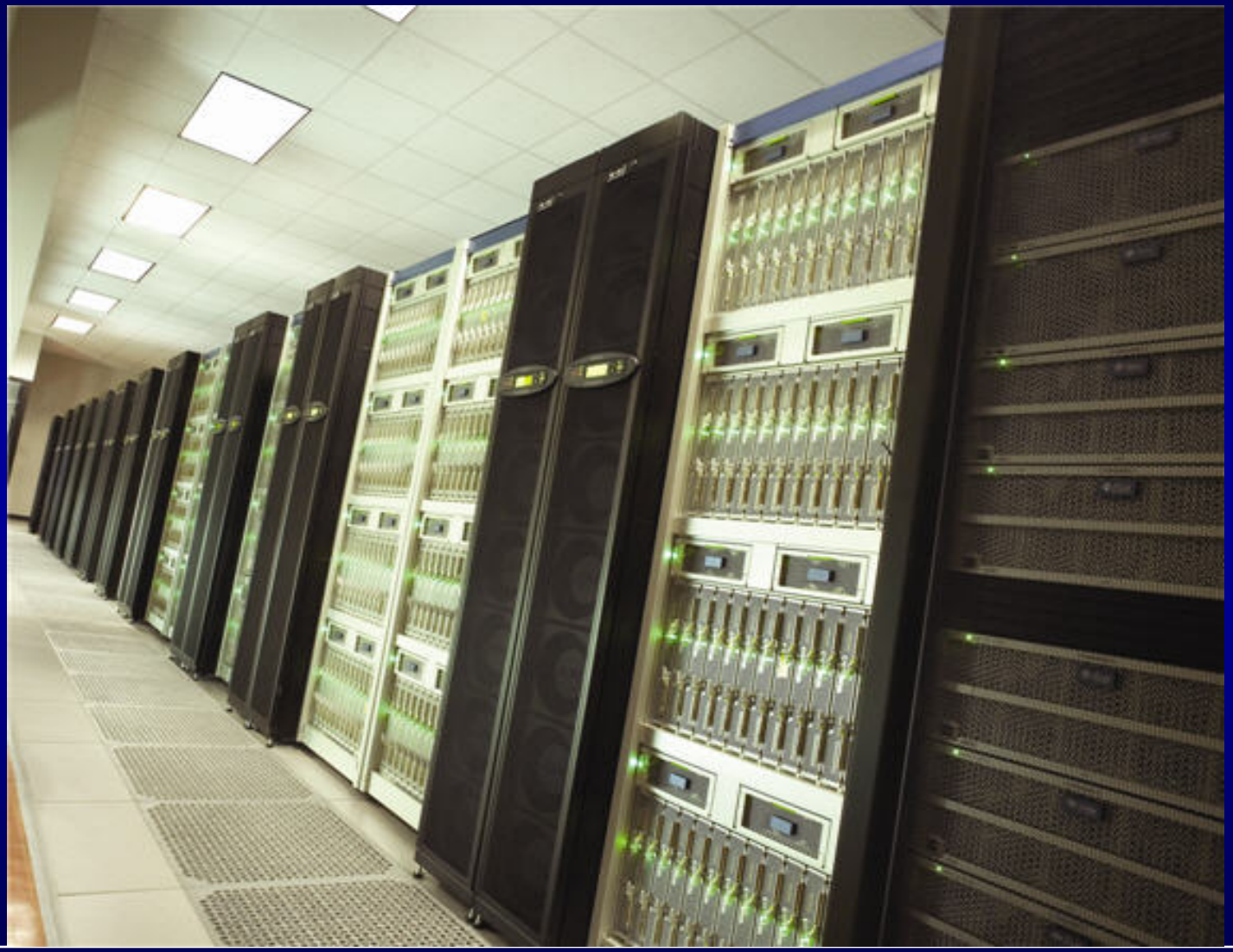
- **HPC systems** to enable larger simulations analyses and faster turnaround times
- **Scientific visualization resources** to enable large data analysis and knowledge discovery
- **Data & information systems** to store large datasets from simulations, analyses, digital collections, instruments, and sensors
- **Distributed/grid computing servers & software** to integrate all resources into computational grids
- **Network equipment** for high-bandwidth data movements and transfers between systems

TACC

Current HPC Systems

- Ranger – 3936 four-socket quad-core AMD Barcelona nodes, InfiniBand network
- Lonestar – 1460 two-socket dual-core Intel Woodcrest, InfiniBand interconnect
- Stampede – 220 node, quad-core, serial throughput and grid computing cluster
- Longhorn – 256 node, two-socket quad-core Intel Nehalem, InfiniBand QDR, NVIDIA Quadroplexes
- Discovery – 60 node benchmark system with variety of processors, InfiniBand DDR

Ranger



NSF

Sun microsystems

AMD Smarter Choice

THE UNIVERSITY OF TEXAS AT AUSTIN

TACC

Ranger: What is it?

- Ranger is a unique instrument for computational scientific research housed at UT
- Results from over 2 ½ years of initial planning and deployment efforts beginning Nov. 2005
- Funded by the National Science Foundation as part of a unique program to reinvigorate High Performance Computing in the United States
- Oh yeah, it's a Texas-sized supercomputer

TACC

Ranger System Summary

- Compute power – 579.4 Teraflops
  - 3,936 Sun four-socket blades
  - 15,744 AMD “Barcelona” processors
    - 2.3GHz quad-core, four flops/clock cycle
- Memory – 123 Terabytes
  - 2 GB/core, 32 GB/node
  - 132 TB/s aggregate memory bandwidth
- Disk subsystem – 1.7 Petabytes
  - 72 Sun x4500 “Thumper” I/O servers, 24TB each
  - 50 GB/sec total aggregate I/O bandwidth
  - 1 PB raw capacity in largest filesystem
- Interconnect – 1 GB/s, 1.6-2.9 μsec latency, 7.8 TB/s backplane
  - Sun InfiniBand switches (2), up to 3456 4x ports each
  - Full non-blocking 7-stage fabric
  - Mellanox ConnectX InfiniBand

Ranger Space, Power and Cooling

- Total Power: 3.4 MW!
- System: 2.4 MW
  - 96 racks – 82 compute, 12 support, plus 2 switches
  - 116 APC In-Row cooling units
  - 2,054 sq.ft. footprint (~4,500 sq.ft. including PDUs)
- Cooling: ~1 MW
  - In-row units fed by three 350-ton chillers (N+1)
  - Enclosed hot-aisles by APC
  - Supplemental 280-tons of cooling from CRAC units
- Observations:
  - Space less an issue than power
  - Cooling > 25kW per rack a challenge
  - Power distribution a challenge, almost 1,400 circuits

External Power and Cooling Infrastructure

TACC

Hot aisles enclosed and racks in place

TACC

Core InfiniBand Switches and Cables

TACC

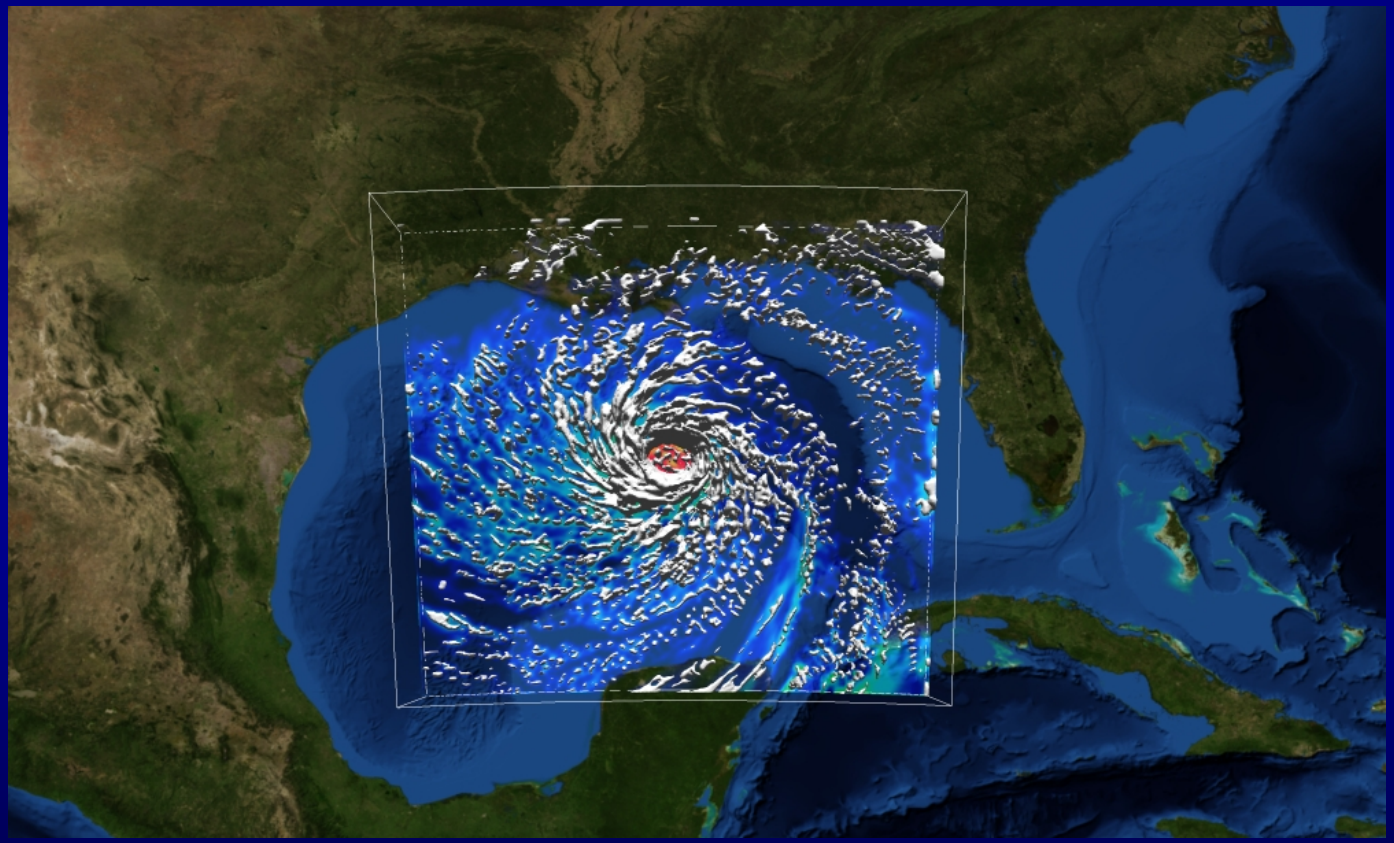
TACC Lonestar System

Dell Dual-Core 64-bit Xeon Linux Cluster
5840 CPU cores (62.1 Tflops)
10+ TB memory, 100+ TB disk

TACC

Weather Forecasting

- TACC worked with NOAA to produce accurate simulations of Hurricane Ike, and new storm surge models



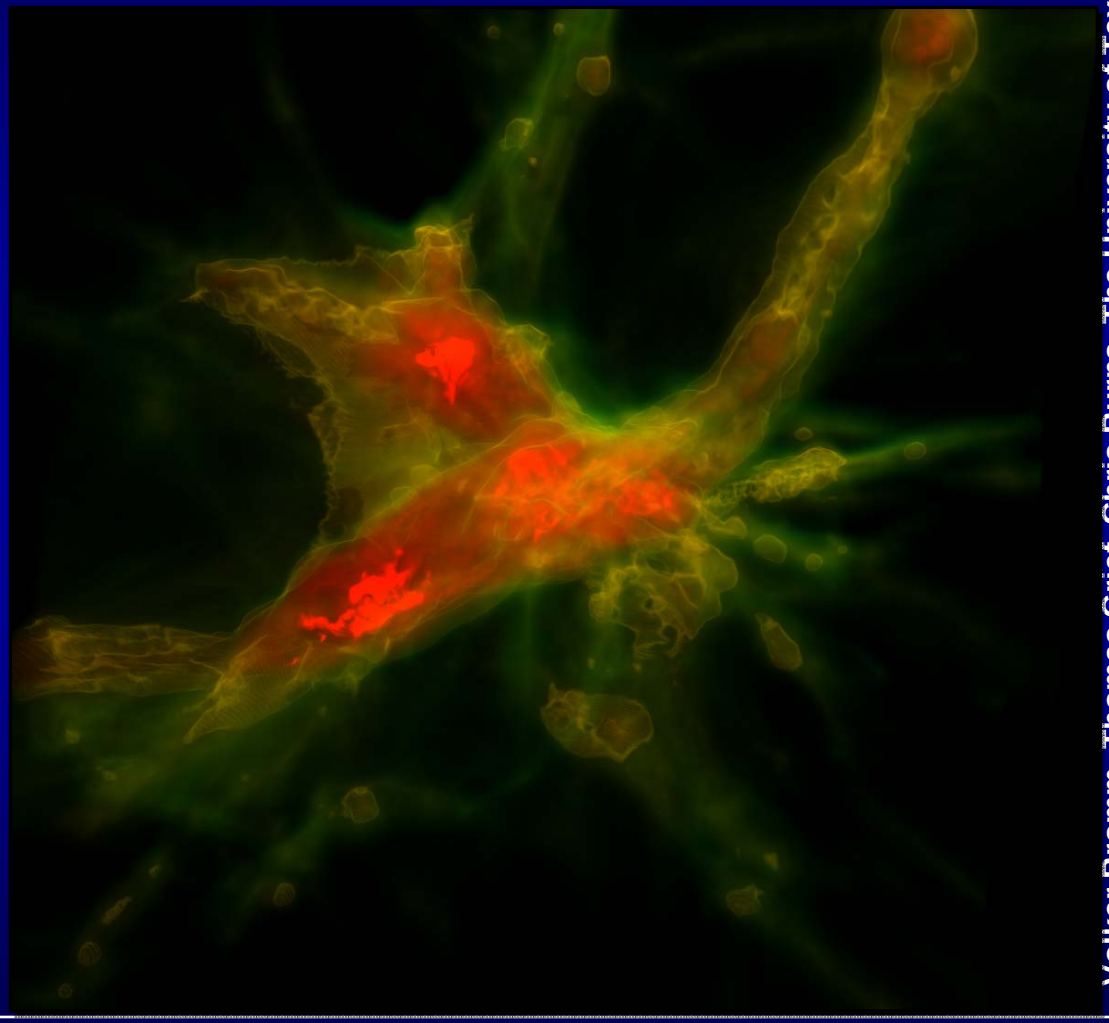
Using up to 40,000 processing cores at once, researchers simulating both global and regional weather predictions received on-demand access to Ranger, enabling not only ensemble forecasting, but also real-time, high-resolution predictions.

TACC

Researching the Origins of the Universe

Volker Bromm is investigating the conditions during the formation of the first galaxies in the universe after the big bang.

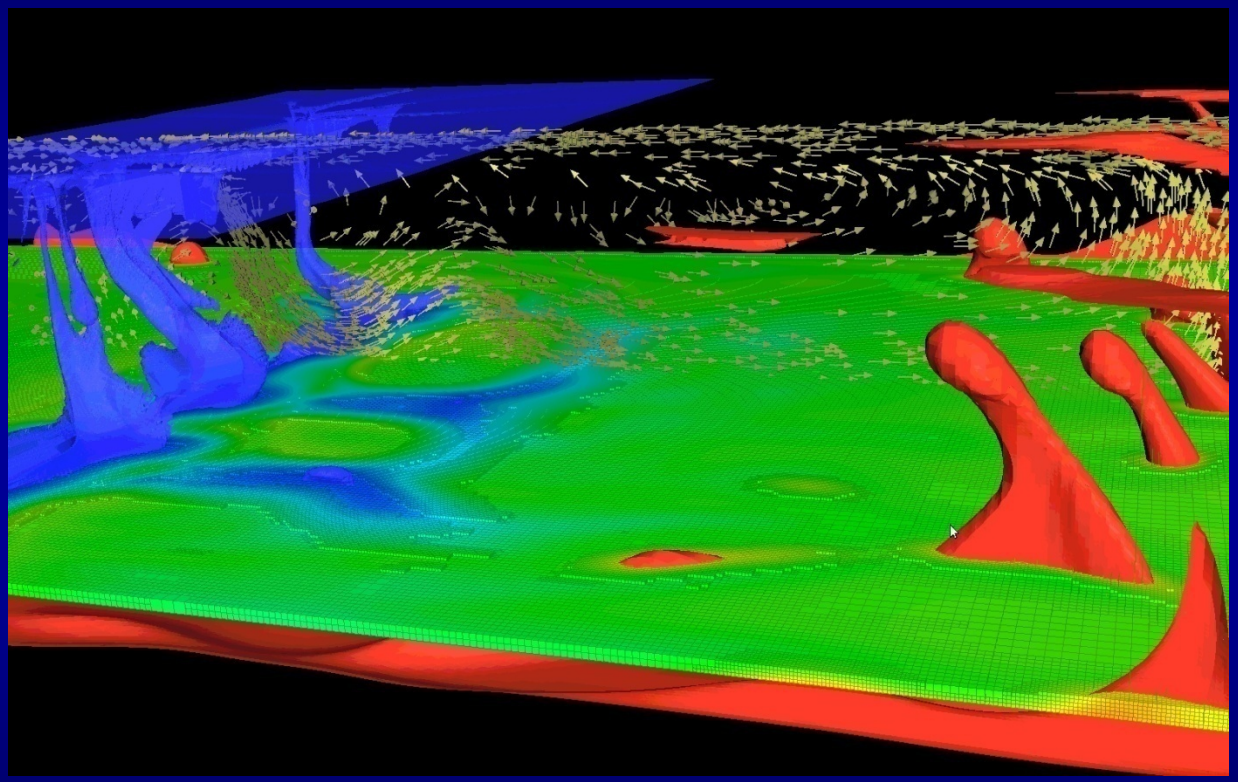
This image shows two separate quantities, temperature and hydrogen density, as the first galaxy is forming and evolving.



Volker Bromm, Thomas Grief, Chris Burns, The University of Texas at Austin

TACC

Computing the Earth's Mantle

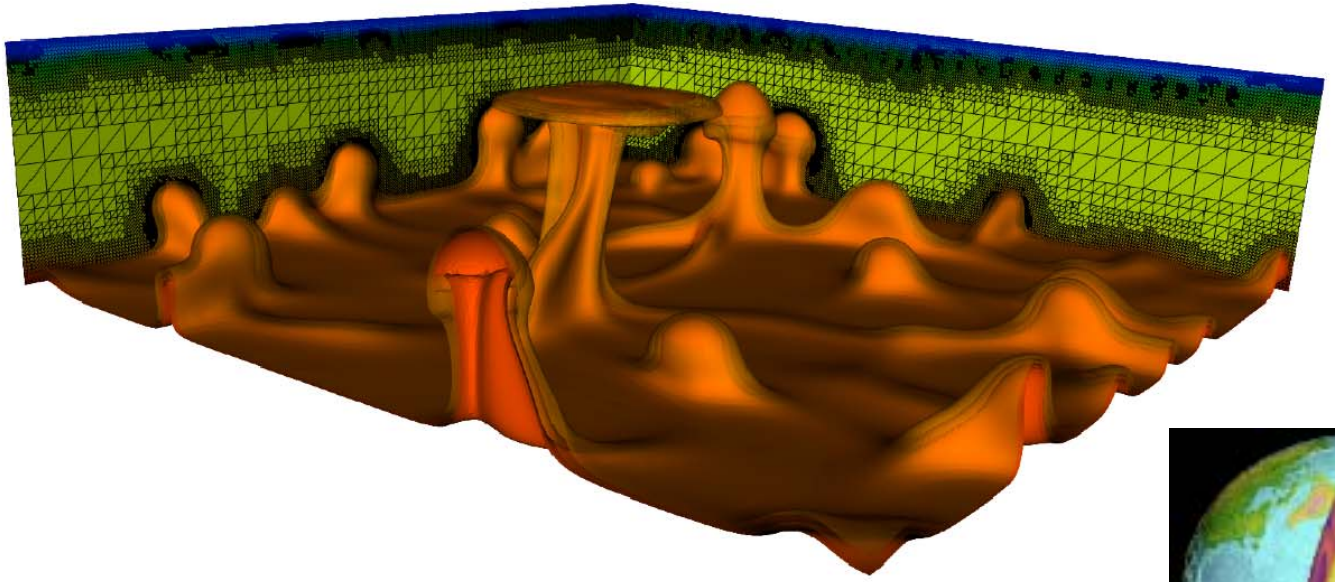


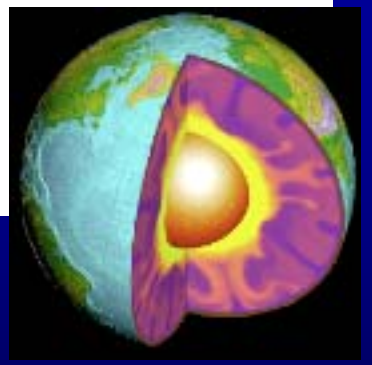
Carsten Burstedde, Omar Ghattas, Georg Stadler, Tiankai Tu, Lucas Wilcox, The University of Texas at Austin

- Omar Ghattas is studying convection in the Earth's interior. He is simulating a model mantle convection problem. Images depict rising temperature plume within the Earth's mantle, indicating the dynamically-evolving mesh required to resolve steep thermal gradients.
- Ranger's speed and memory permit higher resolution simulations of mantle convection, which will lead to a better understanding of the dynamic evolution of the solid Earth

TACC

Application Example: Earth Sciences Mantle Convection, AMR Method

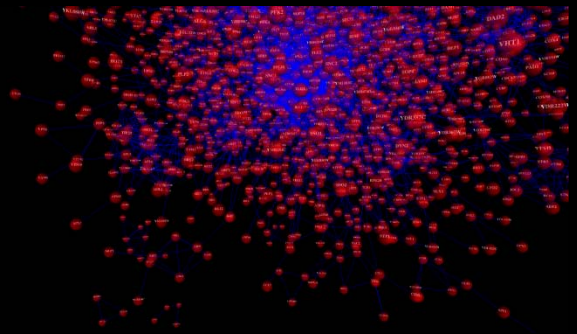




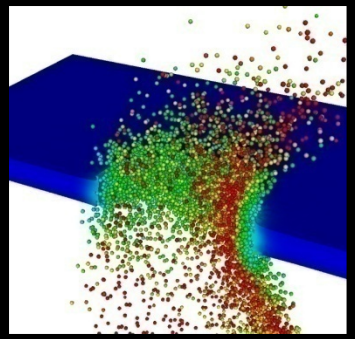
TACC

Courtesy: Omar Ghattas, et. al.

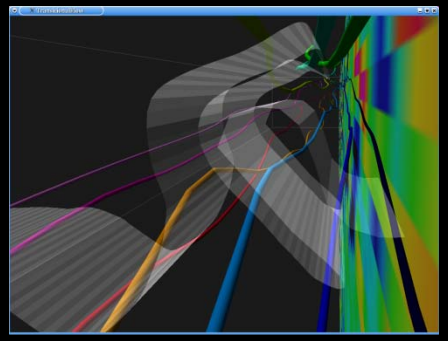
TACC provides visualization resources and services to a national user community



Bioinformatics

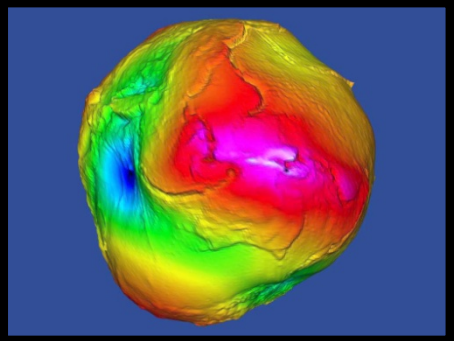


Orbital Debris

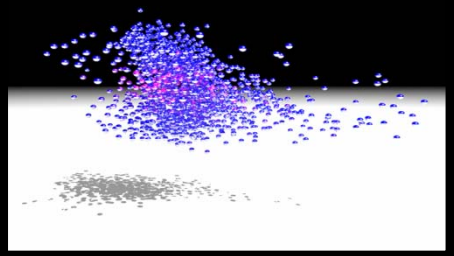


Turbulent Flow

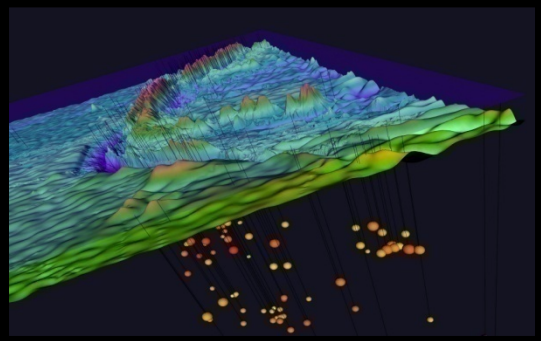
CT Models



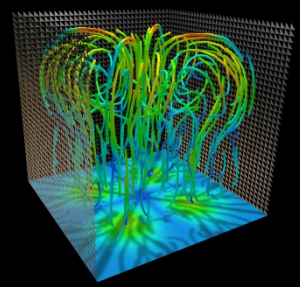
Gravity Map



Quantum Chemistry



GeoSciences



Natural Convection

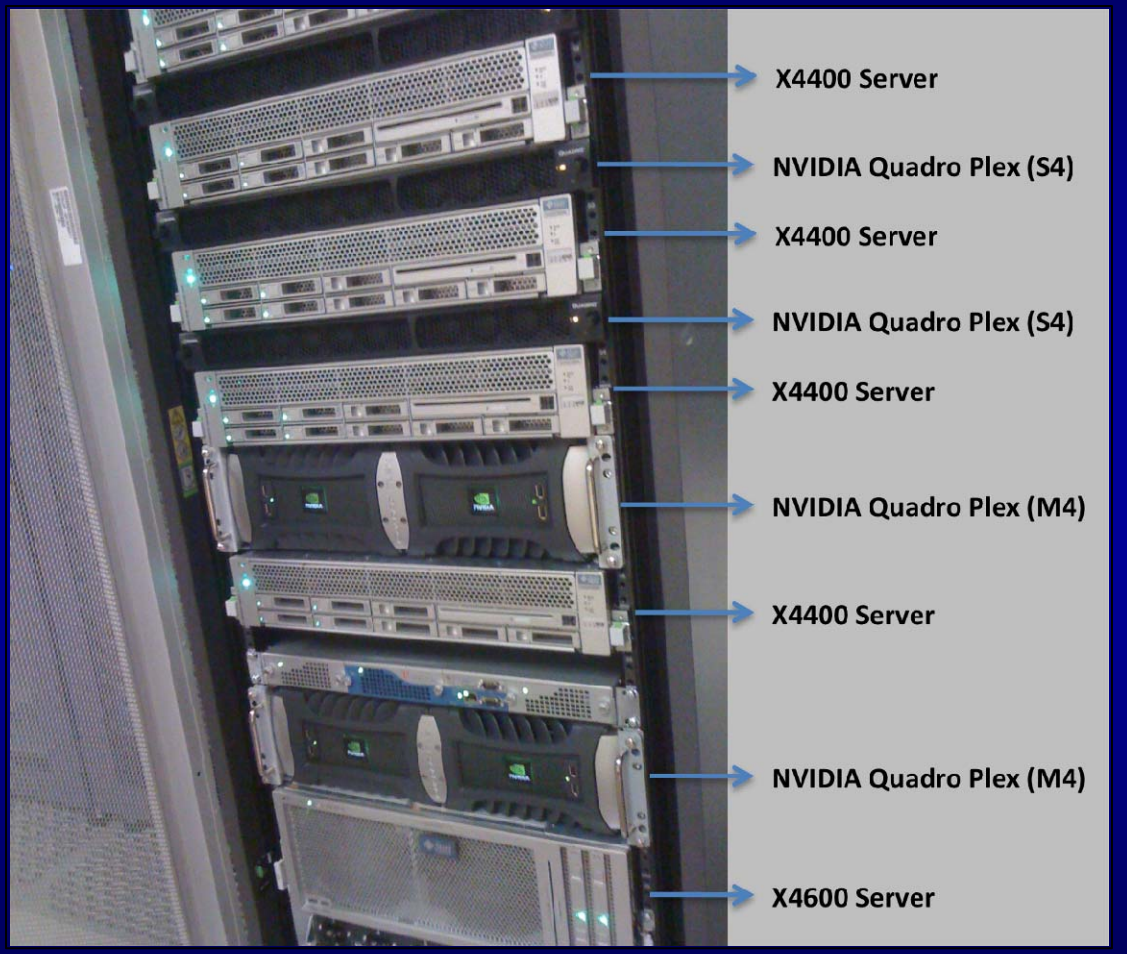
TACC

Remote and Collaborative Visualization

- TACC has been providing remote and collaborative visualization resources to the national community since 2004
- First remote and collaborative resource was **Maverick** – Sun SMP with network-attached graphics processing units (GPUs)
- Currently providing **Spur** – cluster with fat memory nodes and attached Nvidia Quadroplexes, directly connected to Ranger

Spur - Visualization System

- 128 cores, 1 TB distributed memory, 32 GPUs
- Sun Fire X4600 server
  - 8 AMD Opteron dual-core CPUs @ 3 GHz
  - 256 GB memory
  - 4 NVIDIA FX5600 GPUs
- Sun Fire X4440 server
  - 4 AMD Opteron quad-core CPUs @ 2.3 GHz
  - 128 GB memory
  - 4 NVIDIA FX5600 GPUs
- Conntect to Ranger's InfiniBand fabric and accesses its filesystems



TACC

TACC XD Vis Resource -- **Longhorn**

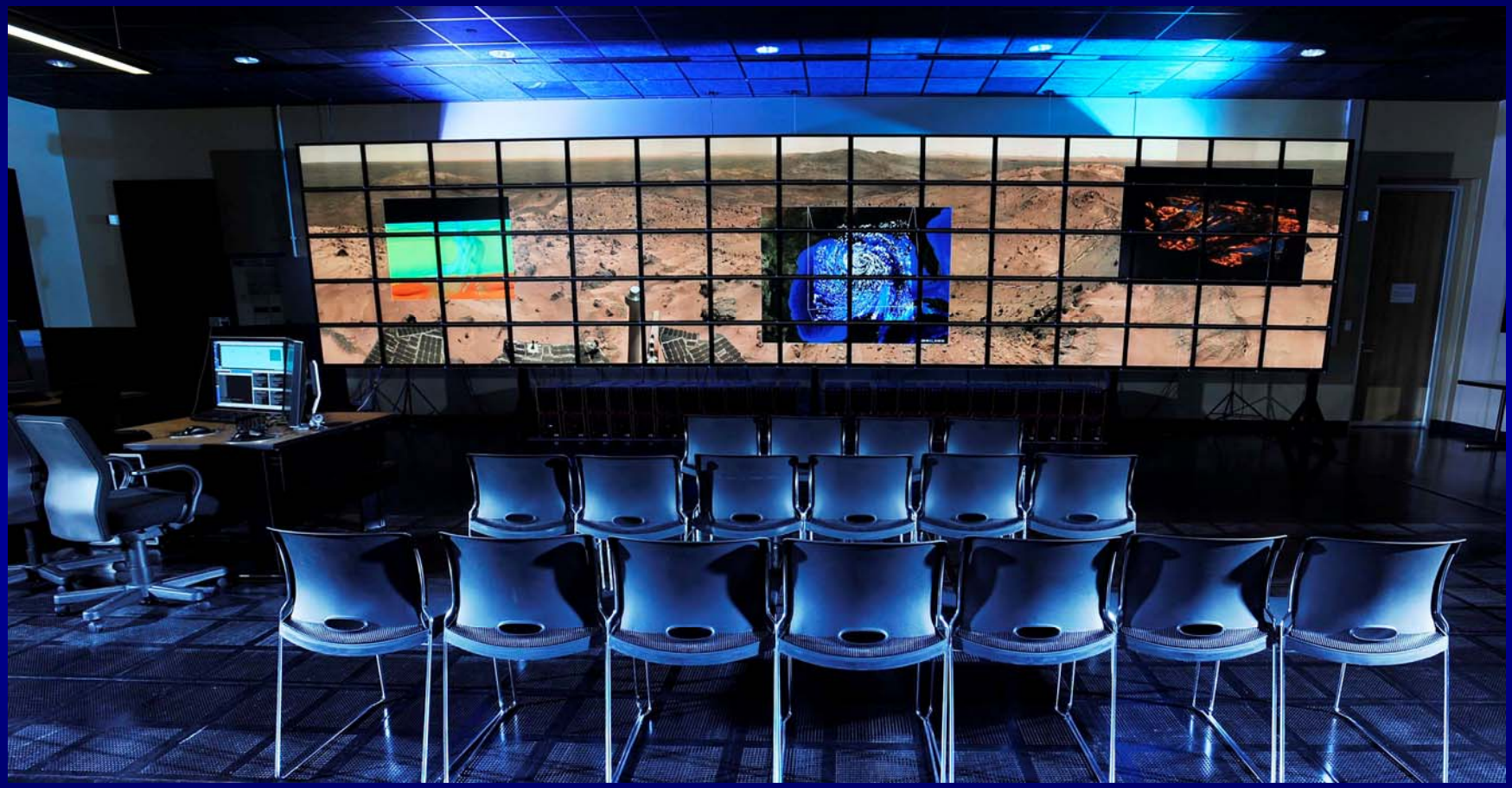
- 256 Dell Quad Core Intel Nehalem Nodes, 2045 cores
  - 240 Dell R610 Nodes
    - Dual socket, quad core per socket: 8 cores/node
    - 48 GB shared memory/node (6 GB/core)
    - 2 Nvidia Quadro FX5800 GPUs/node
  - 16 Dell R710 Nodes
    - Dual socket, quad core per socket: 8 cores/node
    - 144 GB shared memory/node (18 GB/core)
    - 2 Nvidia Quadro FX5800 GPUs/node
  - 14 TB aggregate memory
- QDR InfiniBand Interconnect
- 200TB Lustre Parallel File System
- Jobs launched through SGE

TACC

Stallion

- 15x5 tiled display of Dell 30-inch flat panel monitors
- 307M pixel resolution , 4.7:1 aspect ratio
- 100 processing cores with over 36GB of graphics memory and 108GB of system memory
- 6TB shared file system

Stallion



TACC

TACC Storage Systems

- Corral – Lustre global filesystem
  - 1.2 PB DataDirect Networks disk array
    - InfiniBand based S2A9900 controller
    - 1200 1TB SATA drives, 20 shelves 60 drives each
  - 10 Dell PE1950 servers with 10GigE and IB
- Ranch – Solaris SAM/QFS Archival System
  - Sun x4600, 8-socket, 32-core system, 32GB
  - Sun Storagetek SL8500 tape library
    - 10,000 tape slots, 1TB tapes
    - 14 T10000 tape drives

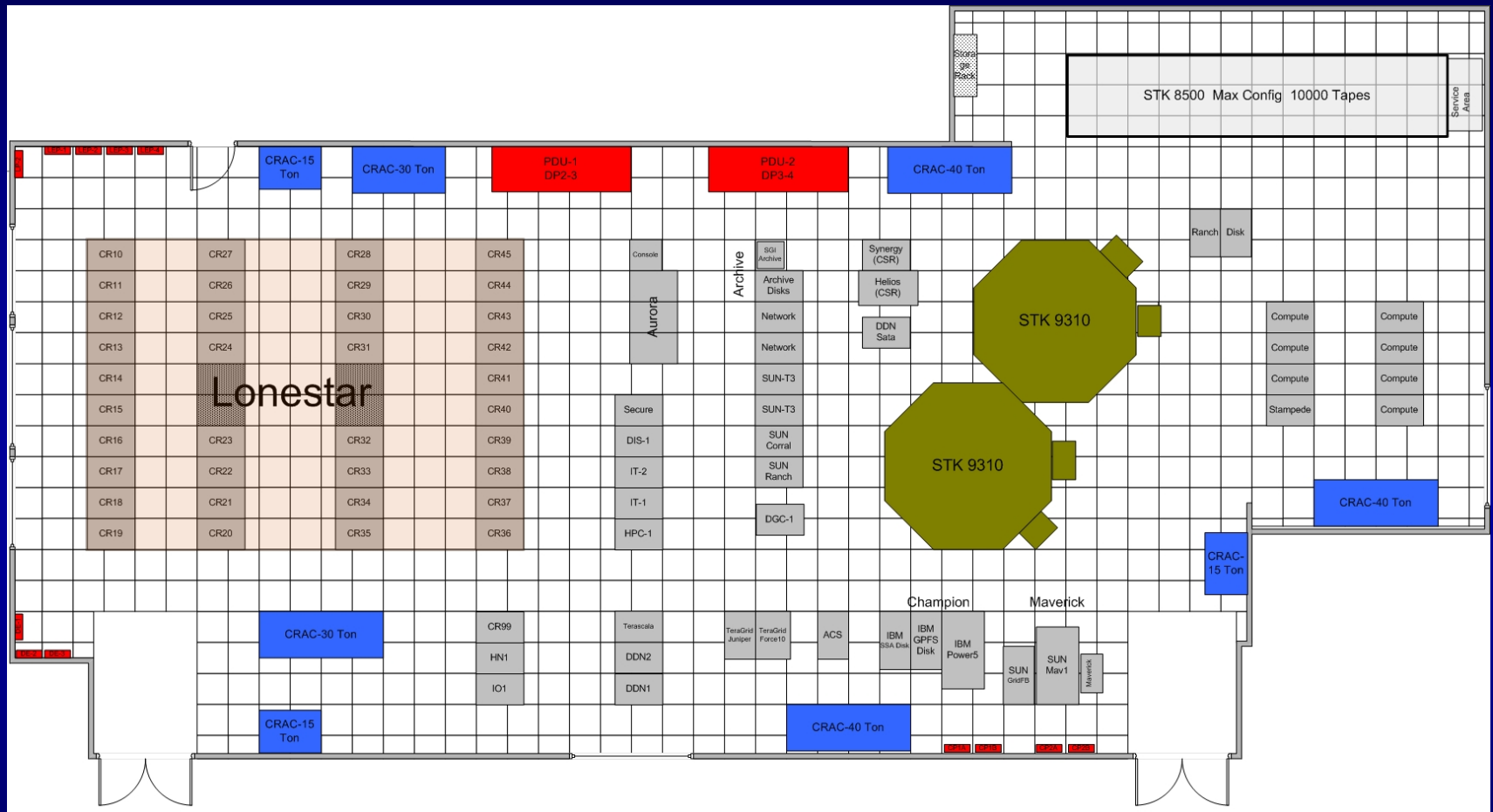
Storage Observations

- SAN replaced by Corral Lustre global filesystem, faster and cheaper per TB than previous fibre channel solutions
- Storage capacity and long-term needs grow proportionally with performance of HPC systems
- On Ranger, we have to purge approximately 1PB worth of files every three months of operation
- Ranch ingests about 15 TB of data per day, but only serves out less than 1TB
- Exponential rate of growth of data, 2/3 of all data stored on Ranch is less than 2 years old

TACC Commons Data Center

- Current specifications:
  - 3,800 sq. ft of 18” raised floor
  - 750 kW power, primarily 208V 20amp circuits
  - 210 tons cooling from Liebert CRACs, mix of 40, 30 and 15 ton units
- Currently occupied by Lonestar, Stampede, Ranch and miscellaneous support systems

CMS Data Center Layout



CMS Computer Room

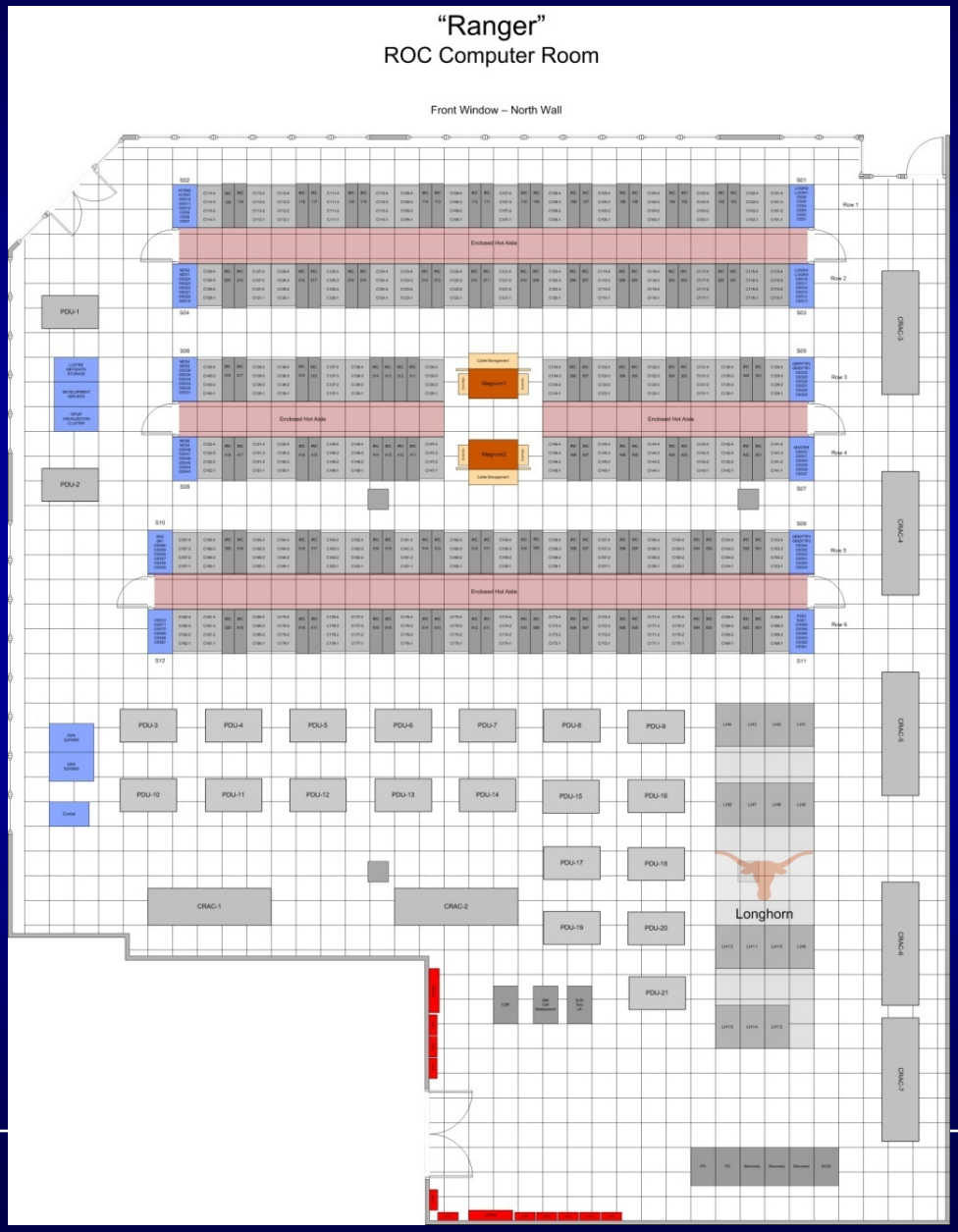
ROC Data Center

- In 2004, new building planned for TACC
- Requested new data center as CMS was already getting close to full
- Requested specifications: 12,000 sq. ft., 5MW power and associated cooling
- Design specifications
  - 6,200 sq. ft. of 30” raised floor
  - 1MW power available through wall circuits
  - 280 tons of cooling from seven 40-ton CRAC units

Ranger Data Center Challenges

- Each rack rated at 28.8kW, each APC In-Row Cooler (IRC) provides 22kW of max cooling
- Disadvantage of IRC units, increased system footprint with IB cable length restrictions
- Even with 30-inch raised floor, cabling an extreme challenge to work around chilled water pipes with thick copper cables

ROC Data Center Layout



Summary

- Computational science demands ever increasing data storage and management
- TACC resources 3-5 years ahead of commercial enterprise systems
- Tape still best option for long-term archival data storage and retention in terms of cost and reliability

More About TACC:

Texas Advanced Computing Center

www.tacc.utexas.edu

info@tacc.utexas.edu

(512) 475-9411