Systems and Software Overview at NCSA

Terascale systems and TeraGrid

Mike Showerman

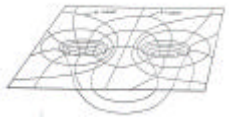
Cluster Software and Tools

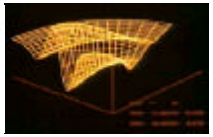
mshow@ncsa.uiuc.edu

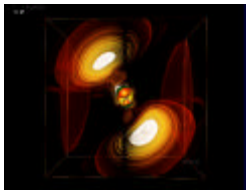
National Center for Supercomputing Applications

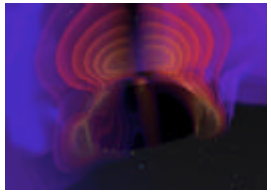
Black Hole Collision Problem

1,800,000,000X









1963
Hahn and Lindquist
IBM 7090
One Processor
Each 0.2 MF
3 Hours

1977
Eppley and Smarr
CDC 7600
One Processor
Each 35 MF
5 Hours

1999
Seidel and Suen, *et al.*
NCSA SGI Origin
256 Processors
Each 500 MF
40 Hours

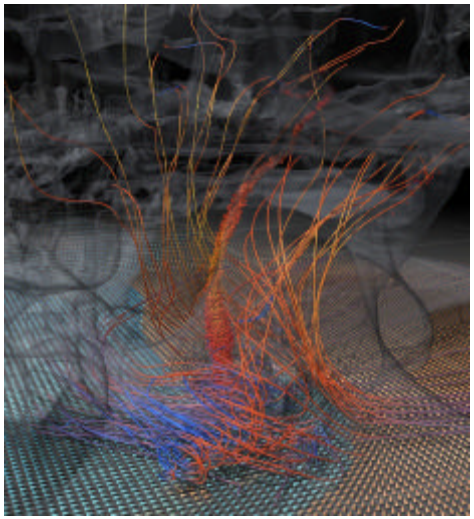
2001
Seidel *et al*
NCSA Pentium III
256 Processors
Each 1 GF
500,000 Hours total
plus 500,000 hours at NERSC

300X

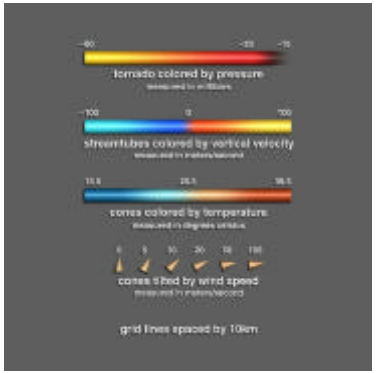
30,000X

~200X

Tornado Modeling – Data to Knowledge



Wilhelmson and Cox



Large complex datasets require data analysis and visualization in the search for understanding

NCSA/Alliance Multiphase Strategy

- **Multiple user classes**
  - ISV software, hero calculations, data intensive analysis
  - distributed resource sharing, parameter studies
- **Four computing approaches**
  - shared memory multiprocessors
    - 12 32-way IBM IBM p690 systems (2 TF peak)
    - large memory and ISV support
  - TeraGrid Itanium2 clusters
    - 64-bit Itanium2/Madison (10.6 TF peak)
    - ETF partners
  - IA-32 clusters (>17 TF peak)
    - 32-bit systems for hero calculations
    - dedicated sub-clusters (3 TF each)
      - To be allocated for weeks or longer to specific teams
  - Alliance Technology Grid & Condor resource pools
- **Complemented by large-scale archives**
  - ~500 TB secondary and 2 PB tertiary storage

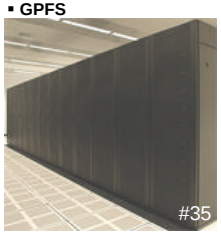
NCSA Control Room

NCSA Computing Environment — 32 TF

Platinum
- Intel Pentium III 1 GHz IBM cluster
- 1,024 processors
- 1 TF peak performance
- GPFS

Titan
- Intel Itanium 800 MHz IBM cluster
- 320 processors
- 1 TF peak performance
- NFS

Copper
- IBM POWER4 p690 systems
- 384 processors
- 2 TF peak performance
- GPFS, 24 TB



Mercury, phase 1 TeraGrid
- Intel Itanium 2 1.3 GHz IBM cluster
- 512 processors + head nodes
- 2.662 TF peak performance
- GPFS, 60 TB
- Production Jan 2004

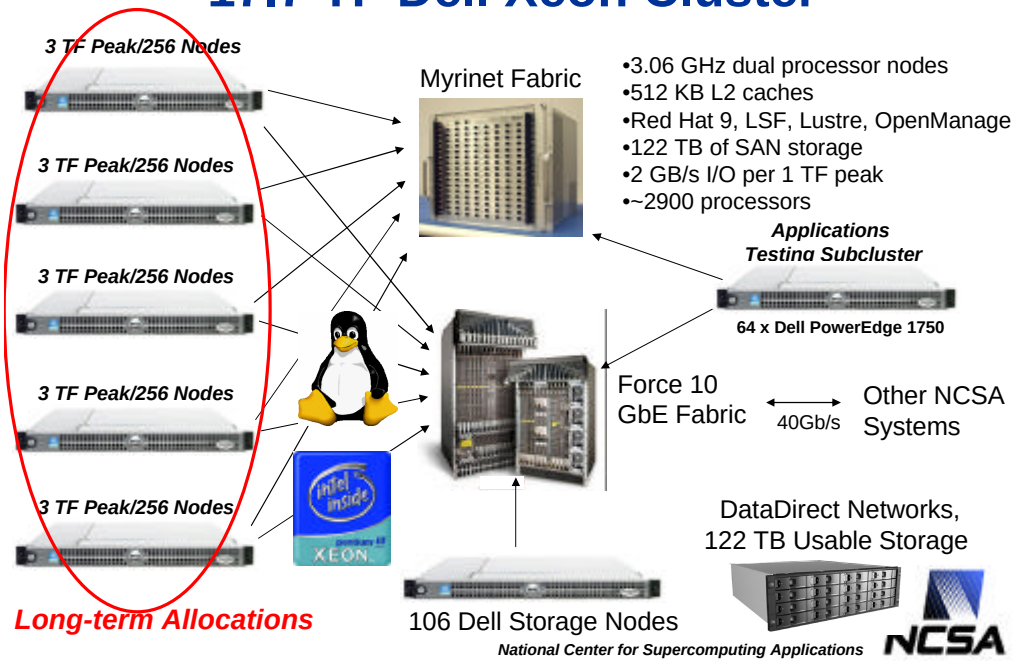
Mercury, phase 2 TeraGrid
- Intel Itanium 2 1.5 GHz IBM cluster
- 1,334 processors
- 8 TF peak performance
- GPFS, 170TB
- Production 2Q 2004

Tungsten
- Intel Xeon 3.0 GHz Dell cluster
- 2,560 processors + IO, debug node
- 17.7 TF peak performance
- Lustre, 140 TB

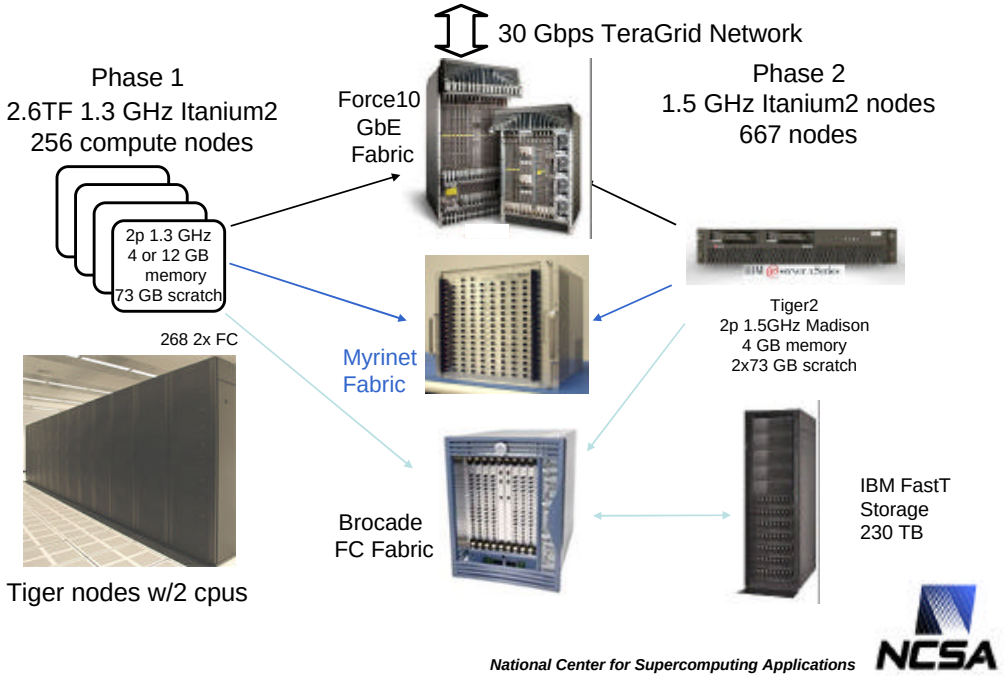
ıpercomputing Applications

NCSA

17.7 TF Dell Xeon Cluster

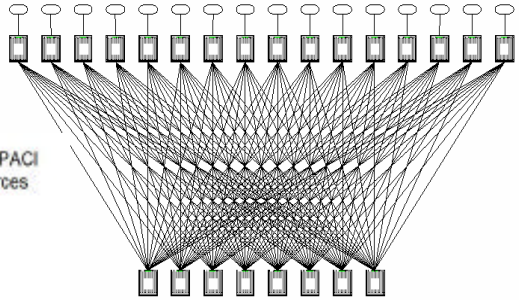


NCSA ETF 10.6 TF Itanium2 and 230 TB Disk

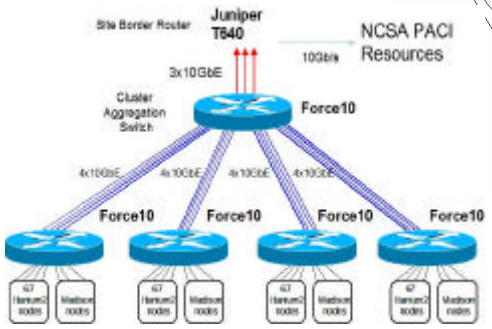


NCSA TeraGrid Network Configs

Myrinet Fabric



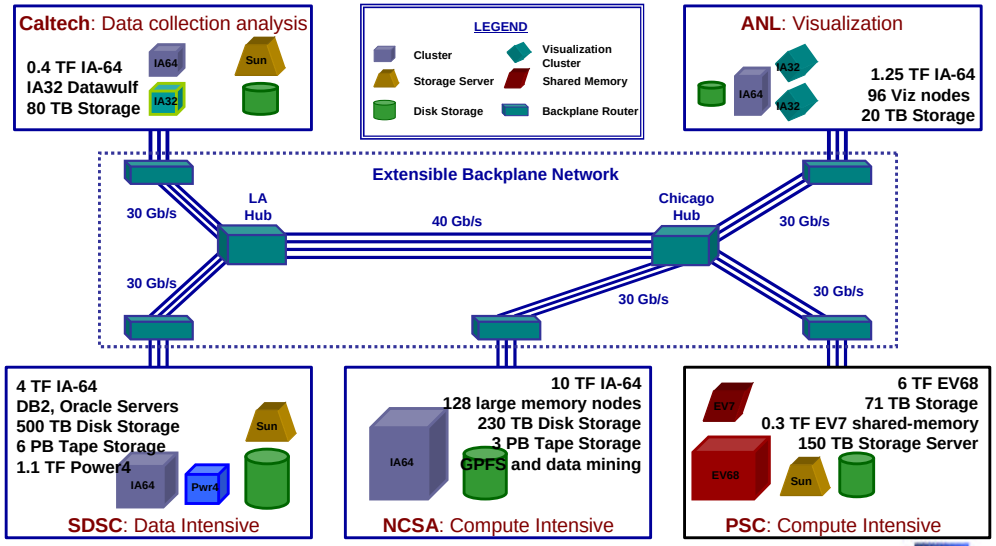
GbE Fabric



940 Itanium2/Madison Nodes

+ SAN Storage networks

Extensible TeraGrid Facility



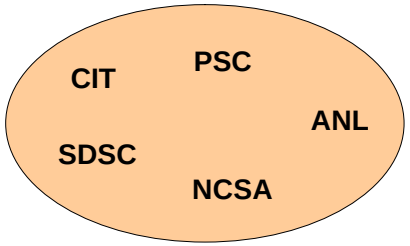
+ ETF2 sites: TACC, IU, Purdue and ORNL in FY04

NCSA

"TeraGrid Roaming"

Nearly eliminate the barrier to entry

- **Example:**
  - Develop application at ANL, run at NCSA
  - Run at Caltech with data from SDSC
  - Run large job across all sites
- **This requires**
  - Consistent interoperable environments
  - Unified allocations, accounting and billing
  - Predictable levels of service, easy to find help desk



Src: TeraGrid

Common TeraGrid Software Stack

- **A social contract with the user:**
  - LORA: Learn Once, Run Anywhere
- **Precise definitions:**
  - Services
  - Software
  - User Environment
- Reproducibility
  - Standard configure, build, and install
  - Single CVS repository for software
- **Code developed for TeraGrid must be Open Source**

Src: TeraGrid

Core CTSS Components Across All Platforms: Linux, AIX, Tru64

- atlas
- blas
- condor-g
- gcc
- globus-2.4.3-gcc
- gsi-ncftp
- gsi-openssh
- gx-map
- hdf4
- hdf5
- Java_COG
- mpich-g2-gcc
- mpich-p4-gcc
- myproxy
- openssh
- openssl
- petsc-gcc
- python
- softenv
- srb-client
- Tcl

Additional component sets include:

Intel compilers, IA64 (Myrinet, BIOS, etc), Linux kernels & patches

Src: TeraGrid

Security

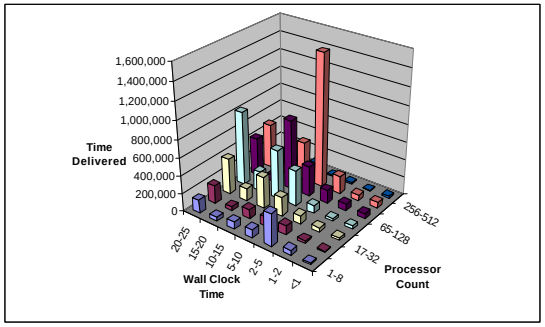
- **We are a small community!**
- **Local root exploits are more deadly**
  - Security model can not rely on trusted users
  - Security fixes happen faster then vendors can adjust
- **Secure Systems can be compromised from an administrative desktop.**
- **TeraGrid has been a Target**
- **Attacks have been organized**
  - Adapting to our reactions
- **Long road ahead.**

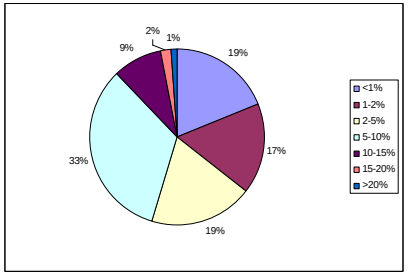
NCSA Performance Engineering

- **Dedicated applications support group within NCSA**
  - Work directly with applications and scientific teams
  - Optimize and analyze applications performance
- **Integrated NCSAbench test suite**
  - Benchmarks and application kernels proven to be useful and relevant
  - Used to test and evaluate systems
    - HPL has repeatedly been invaluable for this on multiple systems
    - Used single node and cluster wide to help identify problems
  - Support for procurements and development of acceptance tests
    - Packaged and made available to vendors
  - Open source codes
- **Expansion/extension for I/O and network performance**
  - Data intensive applications and I/O performance growing in significance
  - Plans include data and I/O benchmarks

Aggressive Performance Analysis

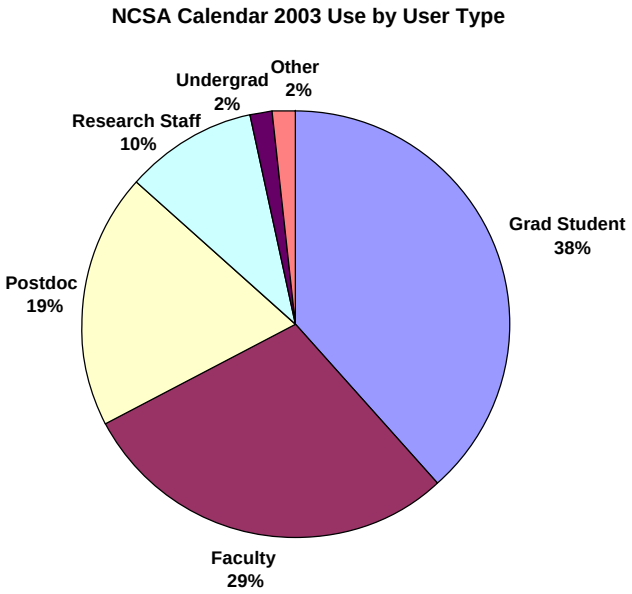
- **Hardware counter capture on IA-32 and Itanium clusters**
  - Perfsuite uses PAPI and NCSA tools
  - guide to system assessment and application tuning
- **Performance expedition tool applications**
  - 2X improvement in GenIDLEST
  - 20-50% improvement in WRF/NCOMMAS





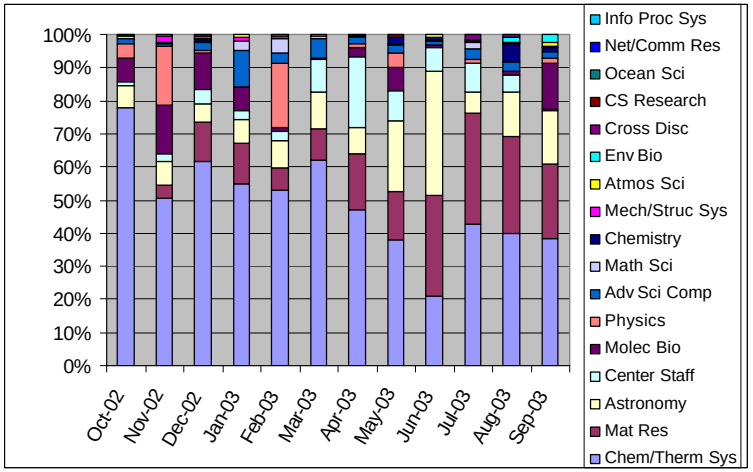
Platinum Performance

Alliance 2003 Users by Type

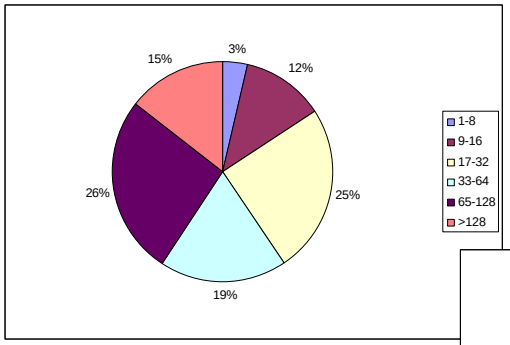


Titan Production Usage Patterns

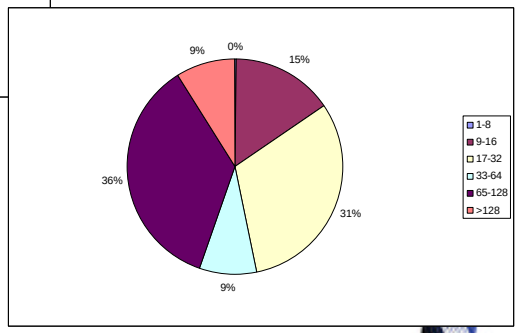
Percentage Usage by Field of Science



Titan Parallel Usage

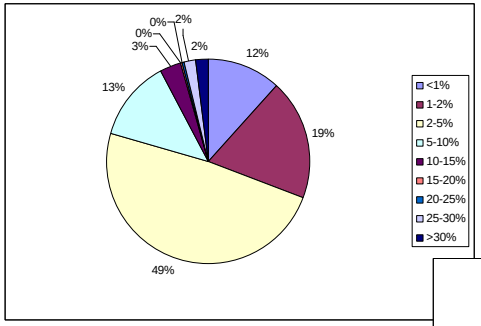


Percentage of time used for increasing node counts for all production runs in FY03



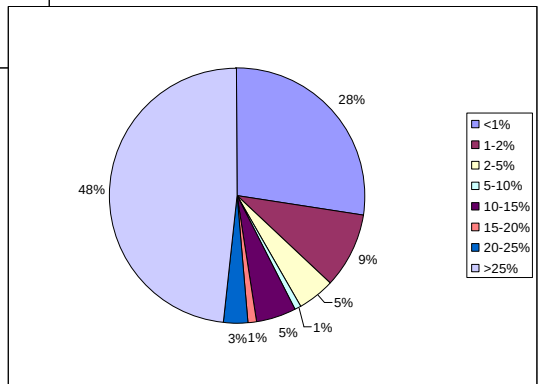
Percentage of time used for increasing node counts for top 5 users (65%) in FY03

Titan Efficiency



Efficiency as a percentage of time used for production runs in FY03.
Data from perfsuite

Efficiency for a Materials Science code in use by a top 5 user



Machine Room Expansion
(Old news - over 2 years old)

- **7500 Square Feet Usable**
- **Additional 750 tons of cooling**
- **Additional 2500 KVA**
- **Basement + 2 stories**
  - 1 usable floor
  - 6 foot raised floor
  - FULL!
- **Designed for machine trends 10 years out**

New Building

- **30 million, 142,000-square-foot building**
- **Occupation in 2005**
- **Staff moves to 1 building (currently in ~7)**
- **NCSA entering a new era**

Summary

- **Clustered systems are very performance competitive**
  - Complex hardware and software systems and environments need to be simplified and integrated
- **Grid environments are maturing and stabilizing**
  - Grid-based applications require extra effort on the part of support and apps developers/users
- **NCSA is in a period of computational and facilities growth**