BLUE WATERS

BREAKING THROUGH THE LIMITS

NCSA's Petascale Computing Storage System

Michelle Butler

Technical Program Manager

Storage Enabling Technogies Group

mbutler@ncsa.uiuc.edu

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTING

NCSA

IBM

NCSA’s largest system today

- **Abe: 1955 blade cluster**
  - 2.33 GHz Cloverton Quad-Core
    - 1,200 blades/9,600 cores

    Lincoln – 192 quadcore with 96 NVIDIA
    - 152 TF; 18.9 TB RAM; 580 TB disk
    - Perceus management; diskless boot
    - Red Hat Enterprise Linux 4 (Linux 2.6.9) – soon to be updated
  - Cisco Infiniband
    - 2 to 1 oversubscribed
    - OFED-1.2 w/ HPSM subnet manager
  - Lustre over IB
    - 22 OSTs
    - 4 9500 DDN controllers direct FC
    - 10 FasT controllers on SAN fabric
    - 12.4GB/s sustained
    - 22 OSTs and 6 MDS
  - Power/Cooling
    - 500 KW / 140 tons

- Production date: July 2007
- #8 on Top 500 (June 07)

- User Environment
  - Torque/Moab
  - Sofenv
  - Intel Compilers
  - MPI: MVAPICH, VMI-2, etc.

NCSA Facility - ACB

- Advanced Computation Building
  - Three rooms, totals:
    - 16,400 sqft raised floor
    - 4.5 MW power capacity
    - 250 kW UPS
    - 1,500 tons cooling capacity

  - Room 200:
    - 7,000 sqft – no columns
    - 70” raised floor
    - 2.3 MW power capacity
    - 750 tons cooling capacity

NCSA's Other Systems

- Distributed Memory Clusters
  - Mercury (IBM, 1.3/1.5 GHz Itanium2):
    - 1,846 processors
    - 10 TF; 4.6 TB RAM; 90 TB disk

- Shared Memory Clusters
  - Cobalt (SGI Altix, 1.5 GHz Itanium2):
    - 2 x 512 processors
    - 6.6 TF; 1 TB or 3 TB RAM; 250 TB disk

NCSA Storage Systems

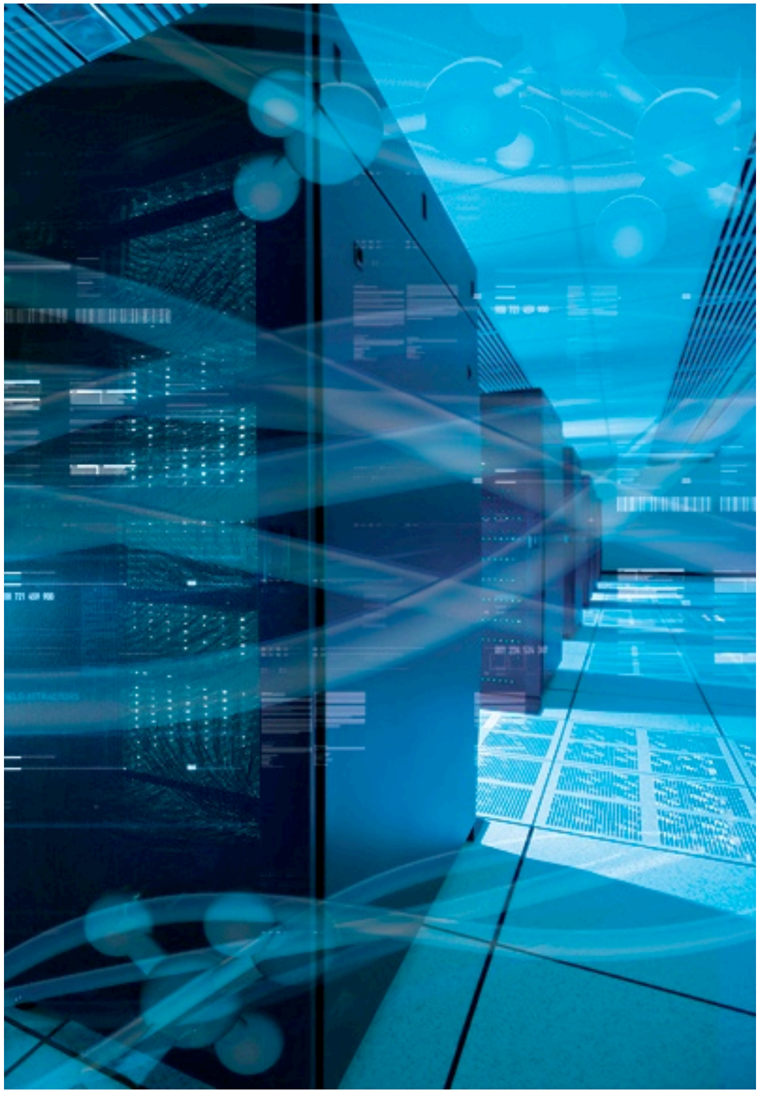
- **Archival: SGI/Unitree (5 PB total capacity)**
  - 160TB disk cache; 50 tape drives
  - currently 3.8PB of data in MSS
    - 80TB – 100TB monthly ingestion rate
    - licensed to support 5PB resident data(copy0)
  - ~30 data collections hosted

- **Infrastructure: 280TB Fiberchannel SAN connected**
  - Fiberchannel SAN connected; FC and SATA environments
  - Lustre, AFS & NFS filesystems

- **Databases:**
  - 8 processor 12GB memory SGI Altix
    - 30TB of SAN storage
    - Oracle 10G, mysql, Postgres
  - Oracle RAC cluster
  - Single-system Oracle deployments for focused projects

Coming in 2011—Blue Waters

- First sustained-petascale system for open science in the world
- Unparalleled national asset will revolutionize scientific research with significant societal impact
- Hundreds of times more powerful than today’s supercomputers
- Comprehensive project includes software, application development and optimization, education, and industry interactions
- New, energy-efficient facility
- National collaboration with UI, IBM, and Great Lakes Consortium
- Supported by $208 million grant from National Science Foundation



Details of what I can say:

System Attribute	Abe	Blue Waters
Vendor	Dell	IBM
Processor	Intel Xeon 5300	IBM Power7
Peak Performance (PF)	0.090	
Sustained Performance (PF)	0.005	≥1
Number of Cores/Chip	4	
Number of Processor Cores	9,600	>200,000
Amount of Memory (TB)	14.4	>800
Amount of Disk Storage (TB)	100	>10,000
Amount of Archival Storage (PB)	5	>500
External Bandwidth (Gbps)	40	100-400

Green Building Initiative: Silver Rating

- **20MW of power into the new building**
  - 4 5MW feeds
- **12000 tons of cooling**
- **20,000ft computer room space**
- **NCSA has own chilled water towers and from campus 12,000 tons and will use outside air 60% of the year to keep cool**
  - BW racks are water cooled based on power 7
- **Cost from the University**
  - Years past all power and cooling were “free”
  - 58% overhead charged to all non-hardware grants
  - New policy for computer labs to pay for power/cooling
    - Ahhhhhh!

What CAN we talk about?

- File system is GPFS! It's going to be really big and fast, but I can't say how big or how fast. ☺
- Disk drive choices with failures
  - 7 failures a day = file system failure every 100 days… ☹
- Stay tuned to Roger Haskin's talk this afternoon.
  - Vdisk, declustered RAID6+, policy data management engine
- Tightly coupled with HPSS as archive server. So tight that HPSS will not have it's own disk cache (except for working areas such as databases and htar)
  - Users access GPFS only, HPSS is in the background as tape only.
- Policy engines… small files will be migrated, but won't be purged (GHI componets)
  - We will be pushing the GHI to it's max.

Talk about continued:

- **Data management – working with file sets so that a single change such as chown reflects to entire set of files; (group stage..etc)**
- **HPSS – RAIT development – duplicate copy is too expensive for .5EB archive size**
- **Inport/export system – queueing for data retrieval into and out of BW machine and bringing data from tape if required   System won't start a job without all data in house and on spinning disk.**
- **375 tape drives - whew!**
    - Library technology not set in stone, it will be 2years before entire archive purchase is done to eek out as much performance from new drives as I can. (8TB tapes!)

Talk about continued:

- 10 yrs for the system without technology refresh (at least that is what is going on now)

- Questions?