BLUE WATERS

SUSTAINED PETASCALE COMPUTING

The Blue Water's File/Archive System

Data Management Challenges

Michelle Butler
(mbutler@ncsa.illinois.edu)

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

NCSA is a…

- World leader in deploying supercomputers and providing scientists with the software and expertise needed to fuel discoveries in science and engineering
- Unique partnership among the University of Illinois, state of Illinois, and federal government
- Home to more than 250 computing experts and students
- Key partner in the National Science Foundation's TeraGrid project
- Home to Blue Waters, expected to be the most powerful computer for open scientific research when it comes online in 2011

TeraGrid™

NCSA's current computing power

- 4 production systems
- More than 155 teraflops (155 TRILLION calculations every second)
- About 1,500 users nationwide
- Researchers receive time at no cost through peer review
- Archive environment at 6PB and growing at 75%/year

Let's get Blue Waters specific!

Diverse Large Scale Computational Science

Science areas	Multi-physics, Multi-scale	Dense linear algebra (DLA)	Sparse linear algebra (SLA)	Spectral Methods (FFT)s (SM-FFT)	N-Body Methods (N-Body)	Structured Grids (S-Grids)	Unstructured Grids (U-Grids)	Data Intensive
Nanoscience	X	X	X	X	X	X		
Chemistry	X	X	X	X	X			
Fusion	X	X	X			X	X	X
Climate	X		X	X		X	X	X
Combustion	X		X			X	X	X
Astrophysics	X	X	X	X	X	X	X	X
Biology	X	X					X	X
Nuclear		X	X		X			X
System Balance Implications	General Purpose balanced System	High Speed CPU, High Flop/s rate	High Performance Memory	High Interconnect Bisection bandwidth	High Performance Memory	High Speed CPU, High Flop/s rate	Irregular Data and Control Flow	High Storage and Network bandwidth

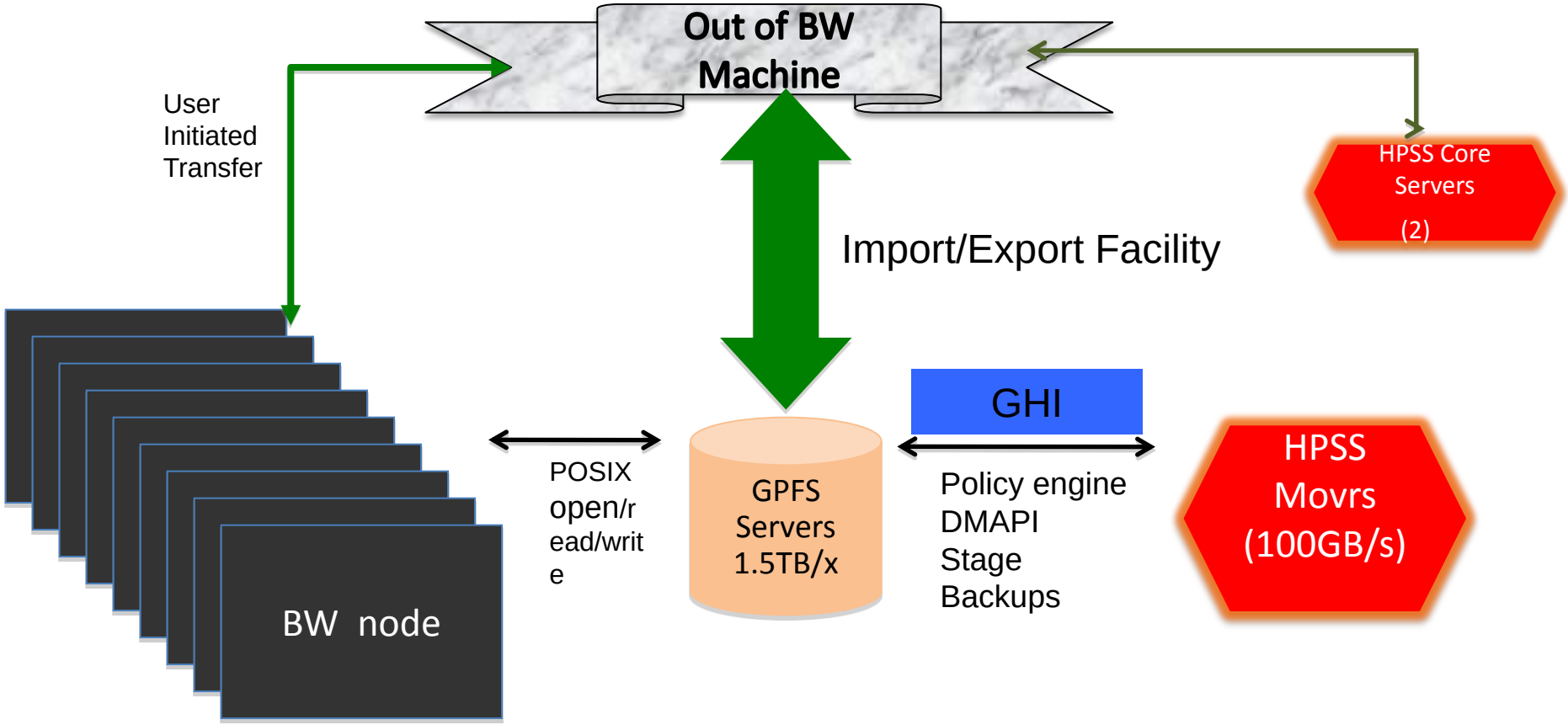
Blue Waters Petascale Computing System

Blue Waters Computing System

System Attribute	Typical Cluster (NCSA Abe)	Track 2 (TACC)	Blue Waters*
Vendor	Dell	Sun	IBM
Processor	Intel Xeon 5300	AMD	Power 7
Peak Perf. (PF)	0.090	0.58	~10
Sustained Perf. (PF)	~0.005	~0.06	~1.0
Number of cores	9,600	62,976	>300,000
Amount of Memory (PB)	0.0144	0.12	>1.0
Amount of Disk Storage (PB)	0.1	1.73	>18
File system Performance (GB/s)	11	30	>1500
Amount of Archival Storage (PB)	6	2.5	~500
External Bandwidth (Gbps)	40	10	100-400

* *Reference petascale computing system (no accelerators).*

I/O Software environment



Storage Management – BW Approach

Have the right data at the right place at the right time

- Blue Waters will proactively use the new storage functions to implement a new state of the practice in HPC storage management (hours to fill diskcache, day to write to tape)
- Goal – no pain (for users anyway ☺) !
  - To have one extremely large storage space – with on-line and near-line limits
  - Approach storage as with virtual memory
    - Large virtual storage
    - Limited work sets of data
    - Try to keep the data with the most temporal locality in the highest (fasted) levels of storage when it is needed
  - Goal vs reality needs to be explored
  - Fall back is to implement a more standard

Possible Layout

Usage	File System	On-line Usable Capacity	Near-Line Capacity	Managed	Quota	Backup
User Home Directories	Midperf	4PB	20 PB?	Yes	Yes on-line & near-line	Yes - via GHI – relatively rapid backup (> 24 hours? residency) All files > 1MB Metadata backedup weekly
High Performance Large Files	Highperf	14 PB	480PB?	Yes	Yes on-line & near-line	Yes – via GHI Longer (> 7 day residency?) Metadata backuped up before upgrades. Alternative is to subdivide with and without GHI
Large Scale Test	Test	.2PB	4	No	No	For new system testing

On going research

- Batch jobs –
  - users tell us ahead of time
    - what objects need to be online before job can be started
    - how much storage space is needed for the job
  - NCSA behind the scenes will move up the data from near-line(on-demand stage) or from across country (gridftp)
  - Using attributes in GPFS to “lock” files on disk so that they don’t get “punched or purged” before all the data is on-line.

On going research -

- What files need to stay on disk for further anaylsis? (post analysis)
  - what can go to archive immediately (safe keeping),
  - what can be deleted? Checkpoints?
  - Post job data management step

On going research -

- For retrieval: how will the files need to be associated together
  - Using GPFS filesets for the PRAC projects
  - Researching the filesets environments
    - so policy scans can be run in parallel over filesets
    - quotas implemented at fileset level
    - use HPSS family of files for project from GPFS filesets

National Petascale Computing Facility at a Glance

www.ncsa.illinois.edu/BlueWaters

EYP MCF/
Gensler
IBM
Yahoo!

- 88,000 GSF over two stories—45' tall
  - 30,000+ GSF of raised floor
  - 20,000+ unobstructed net for computers
  - 6' clearance of raised floor
- 24 MW initial power feeds + backup
  - Three 8 MW feeds + One 8 MW for backup
  - 13,800 volt power to the each
- 5,400 Tons of cooling
  - Full water side economization for 50%+ of the year
  - Automatic Mixing of mechanical and ambient chilled water for optimal efficiency
  - Adjacent to (new) 6.5M gallon thermal storage tank
- 480 Volt distribution to computers
- Energy Efficiency
  - PUE - ~1.02 to <1.2 (projected)
  - USGBC LEED ~~Silver~~ Gold (Platinum?) classification targe

Questions? See me

Michelle Butler
NCSA/University of Illinois
Technical Program Manager
mbutler@ncsa.illinois.edu- http://www.ncsa.uiuc.edu/BlueWaters