USCMS Tier 1 Update on Use of IBRIX Fusion FS

Lisa Giacchetti

HEPiX Spring 2005

Outline

- Brief History of Global Filesystem/Storage Evaluation
- What is IBRIX Fusion FS?
- USCMS Tier 1 Configuration of IBRIX
- USCMS T1 Deployment Experiences

Brief History

- Fermi CD started evaluation of Global Filesystem/Storage solutions Fall of 2002
- USCMS a key contributor as they were seeking a solution for “user disk” - highly reliable space for code, work area common area or compressed dst
- 4 vendors evaluated

  Panasas, Zambeel, Spinnaker, IBRIX

  Most included a hardware component as well as a global filesystem component.
- Standard set of tools developed to maintain consistency across evaluations (see HEPiX fall 02 report by Chris Brew)

  Combination of bonnie, IOZone and reader/writer (2 simple “C” programs written at FERMI

Brief History

CMS Decided to Implement IBRIX solution – Why?

No specialized hardware required

- Enabling us to redeploy hardware if this solution did not work
- Made cost of product less than others with hardware components

Provided NFS access

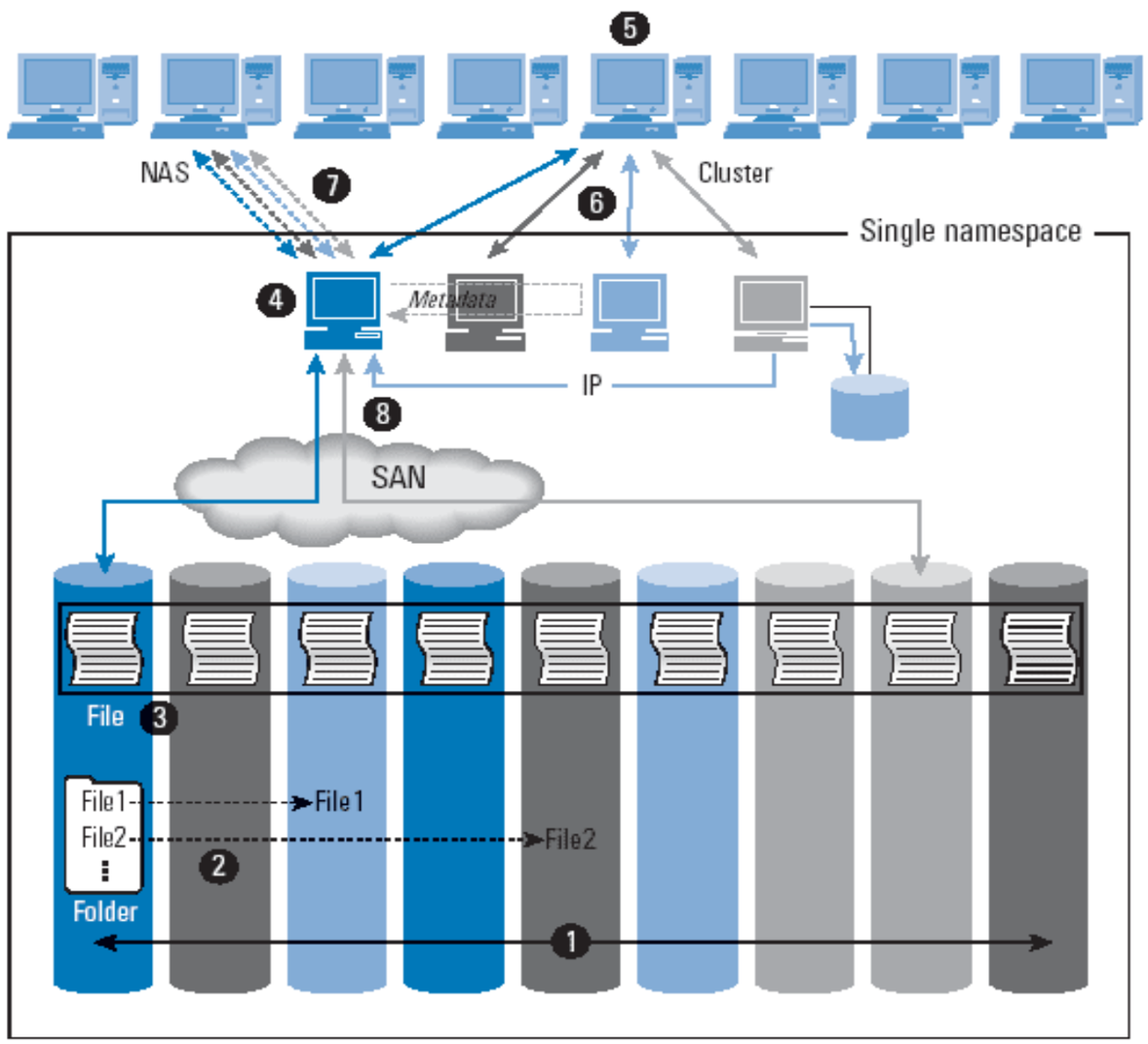
- A universal protocol which required no client side code
- Initial decision to only use NFS access because of this

IBRIX very responsive to our requests and to issues we found during the evaluation

What is IBRIX Fusion FS?

- Purely software solution: no specialized hardware dependencies
- Comprised of:
  - Highly scalable POSIX compliant parallel file system
  - Logical volume manager based on LVM
  - High-availability features
  - Comprehensive management interface which includes a gui
- Features: www.ibrix.com/features.php
  - Distributed and modular metadata management
  - Dynamic load balancing
  - Automatic server failover
  - Partial online operation
  - Scalable multi-node locking with no central lock manager
  - Uses **any** combination of San attached or direct attached storage

What is IBRIX Fusion FS?



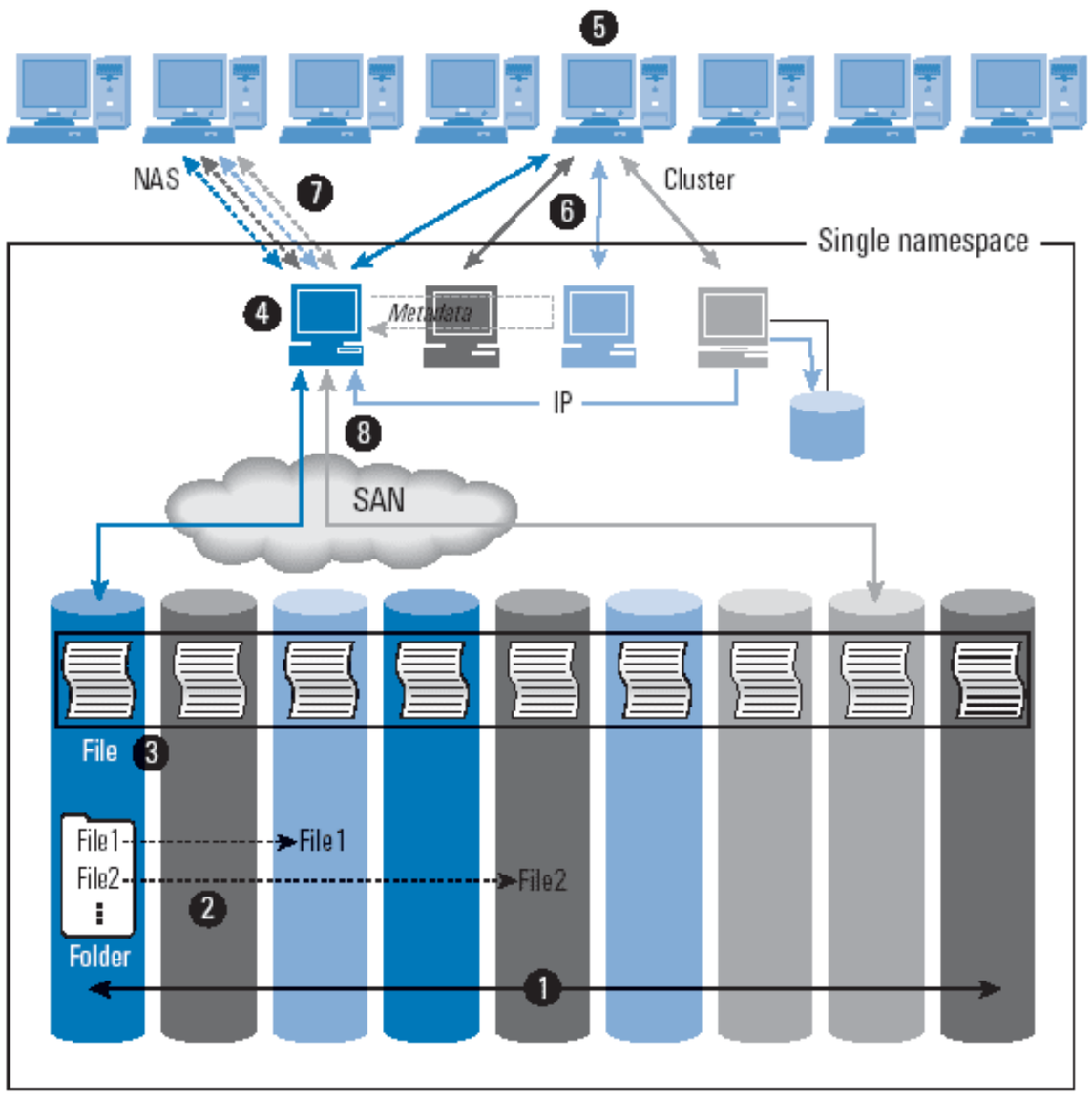
- (1) File space = collection of is a collection of segments. Each segment a repository for files and directories. No implicit namespace relationship among segments. Segments can be of any size
- (2)Location of files and dirs w/in particular segment independent of their respective and relative location in namespace. So a dir could be on different segment than the files within it.

  allocation of files/dirs dynamic and policy determined by admin
- (3) Individual files can be distributed across multiple segments

Reprinted from Dell Power Solutions, February 2005.

Copyright C2005 Dell Inc. All rights reserved

What is IBRIX Fusion FS?

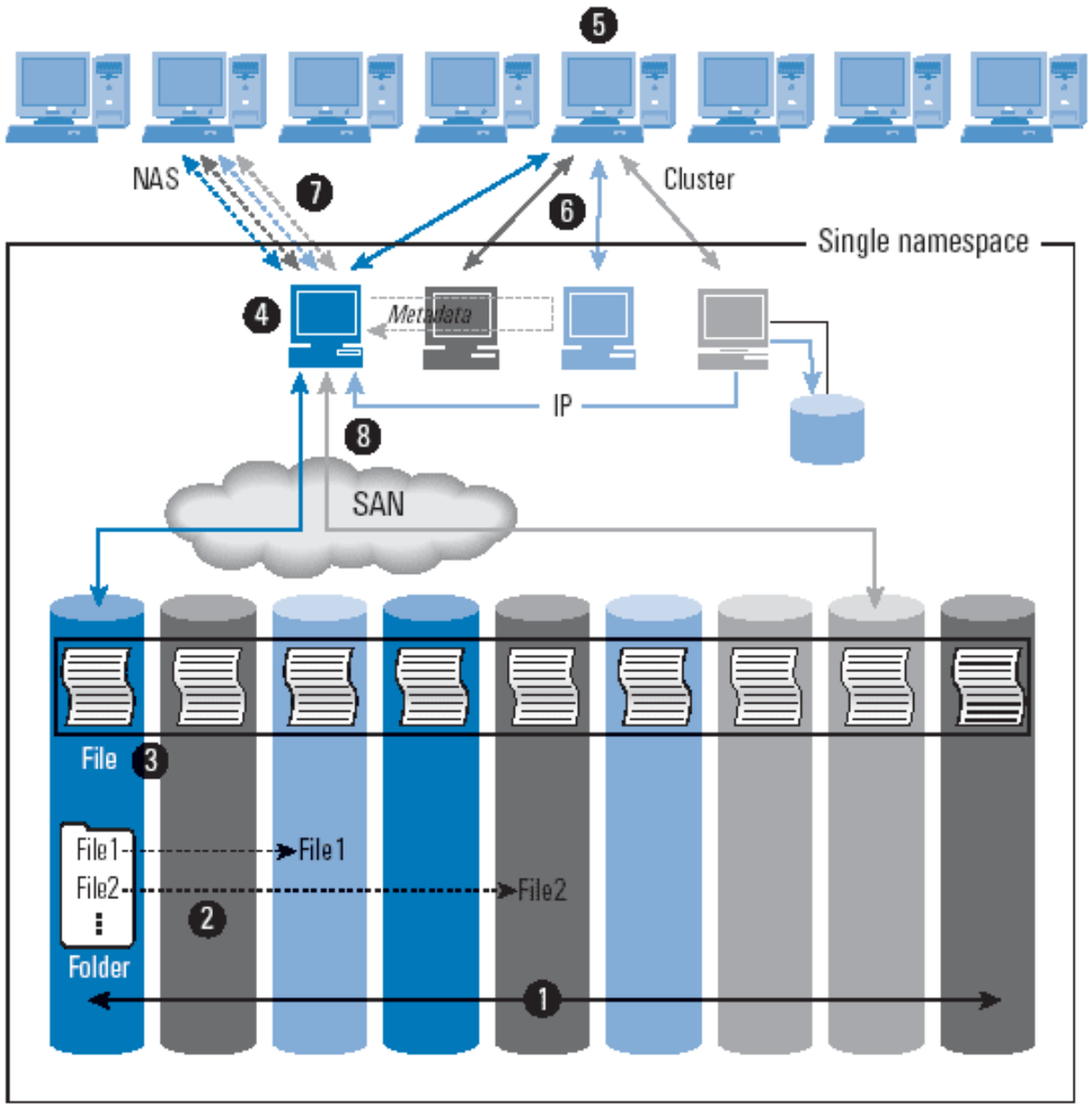


- (4) Each segment assigned to server. Servers can “own” multiple segments. Online migration of segments, adding of servers and adding of segments. Failover protection for servers available.
- (5) Clients run applications that use the file system. Access via locally mounted cluster file system (IbrixLite) or via NFS/CIFS
- (6) IbrixLite advantage

  IBRIX driver aware of segmented architecture and can route requests directly to correct segment server.

Reprinted from Dell Power Solutions, February 2005.

Copyright C2005 Dell Inc. All rights reserved



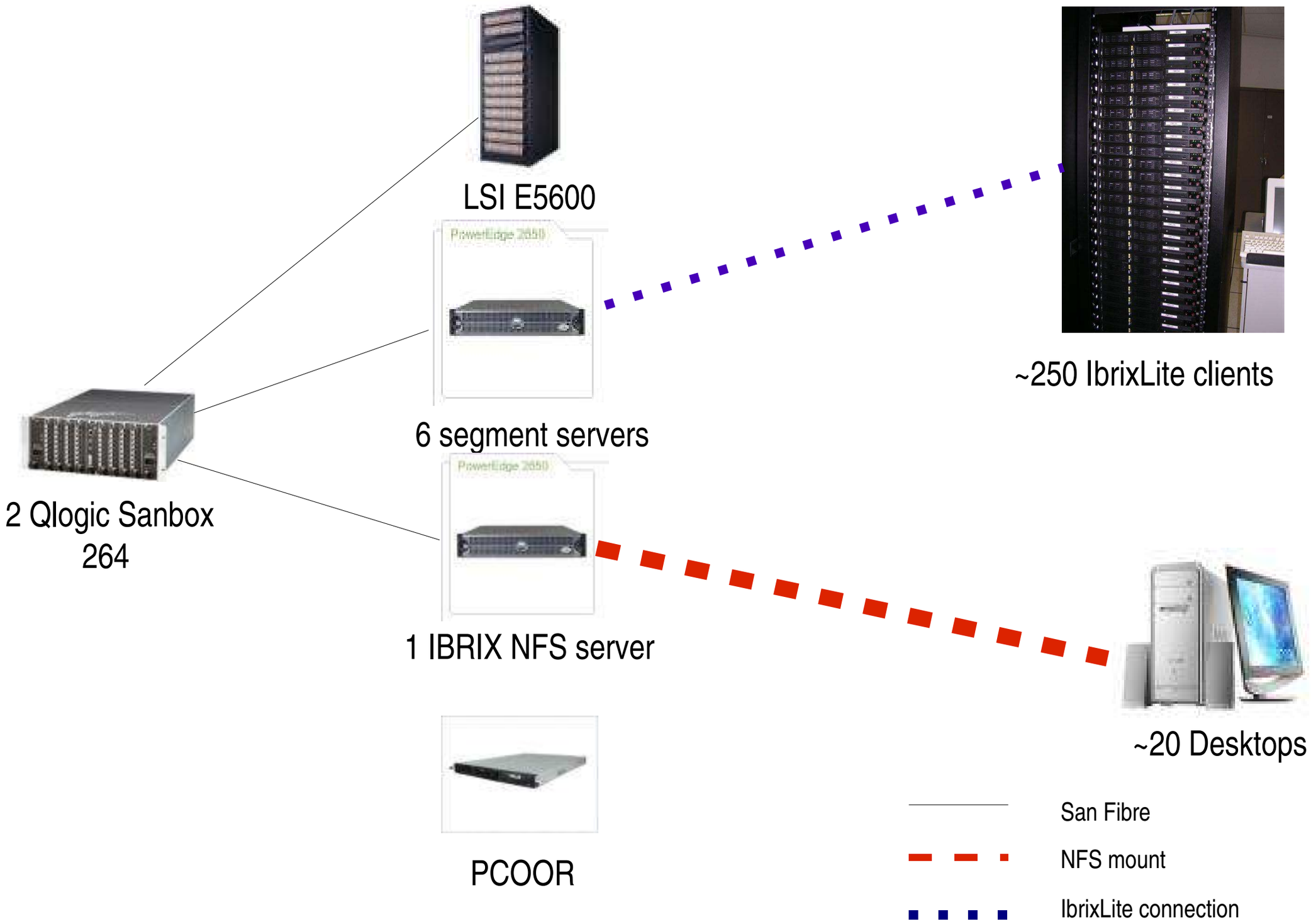
- (7) When NAS protocols used, all communication done via the server file system mounted from on client. Benefit = multi-platform support, less cost
- (8) A requests at a segment server could be for a file on a segment that is either owned by the server, owned by another server but accessible via the SAN or owner by another server and not accessible via SAN.

  scenario 2: server gets metadata from owning server and does IO directly on SAN

  scenario 3: IO performed through owning server over IP network

Reprinted from Dell Power Solutions, February 2005.

Copyright C2005 Dell Inc. All rights reserved


LSI E5600
~250 IbrixLite clients
6 segment servers
2 Qlogic Sanbox 264
1 IBRIX NFS server
~20 Desktops
PCOOR
San Fibre
NFS mount
IbrixLite connection

USCMS T1 Configuration of IBRIX

- IBRIX file system layout

  4 LSI trays configured into 10 RAID5 sets of 5 disks each

  Each RAID set is then configured as an IBRIX segment

  5 segments in each of 2 file systems totaling 2.7TB usable disk per file system

  4 segments servers own 2 segments each, other 2 only own 1 segment

- Each segment server has a fail-over host assigned to it
- The NFS server also has a fail-over server.

  NFS fail-over requires a second nic on the fail-over server so it can assume the IP/mac address of the nic used for NFS traffic on the failed server.

- All servers have IBRIX monitoring from pcoor enabled.
- Expect to grow to 20TB before beginning of data taking.

USCMS T1 Configuration of IBRIX

/uscms/

home/<user_name>/ (2GB Quota per user)

work/

private/ (permissions on this dir restricted to user)

nobackup -----> /uscms_data/<d#>/<username>

physics_grp/<physics_grp>/ (5GB quota per physics group)

code/

work/

nobackup -----> /uscms_data/<d#>/<physics_grp>

METADATA/

ups/

prod/

grid/

software/ (link in here to web area on cmsxfer?)

sw/

/uscms_data/d1/

<user_name>/ (5GB quota per user)

<physics_grp>/ (250GB quota per physics group)

grid_home/ (30Gb quota for OSG users)

USCMS Deployment Experiences

- First installation Dec of '03

  4 segment server, 1 pcoor and 3 trays of LSI disk

  Decision made for all client access to be over NFS

  - Less impact on clients
  - More universal access

  First couple of months:

  - Figuring out the best hardware configuration of disk/segments
  - Figuring out Qlogic preferred pathing configuration on segment server
  - Testing file system features, management, failover, performance

USCMS Deployment Experiences

- March '04 file system put in to limited use for USCMS Production cluster
  - Just a couple of users running production jobs as uscms01
  - Low level interactive use
- April '04 added access to IBRIX from the USCMS User Analysis Facility (UAF)
  - Users told file system was in a pre-production state
  - Encouraged feedback
  - Disk available for interactive and batch use
- June '04 upgrades
  - Server OS and IBRIX code
  - More segment servers and disk

USCMS Deployment Experiences

Aug '04 Issues were:

Many issues with NFS access to file system

- Some inherent to NFS

Fail-over mechanisms had bugs or certain fail-over scenarios were not recognized

NFS client issues with recovery after a failure

Quota mechanism not returning over-quota signal to user program

Uncovered several usage cases that had not been considered in the IBRIX file system design which caused failures

- Very long paths with symbolic links at the end
- Editing of a file with emacs from 2 different segment servers

- IBRIX support and engineering extremely responsive and provided fixes quickly
- In the end the NFS access issues were the largest block to moving forward

USCMS Deployment Experiences

- August '04 Due to stability issues IBRIX recommends that we:

  Wait for v1.3 of server code

  Re-deploy using thinclient instead of NFS

  - They had seen much better stability with this configuration

  USCMS agrees with this and temporarily shelves use of IBRIX

  - Were able to re-deploy IBRIX disk as regular NFS mounted disk while IBRIX resolved problems our large scale actual user uncovered
- Jan '05 v1.3 beta of IBRIX installed

  Extensive testing shows file system is much more reliable/stable
- Feb '05 IBRIX v1.3 production release installed

  271 thinclients on RH73, SL3(F) intel and athlon

  6 segment servers

  1 pcoor

  Clusters have been running very successfully

USCMS Deployment Experiences

- Current status

  Thinclient is working very well and system is stable

  User and group quotas are working as part of our managed disk plan

  - A few requests for enhancement to quota sub-system to improve quota management

  Working on getting users to migrate off of ALL nfs mounted work area or data disk

  IBRIX file systems via NFS in limited numbers but has been stable

  Refining admin documentation and knowledge

  Will add in 2 more segments servers and another 2.7TB of disk by mid-May

  - Plan for 20TB by start of data taking

  Thinclient rpm's are kernel version dependent; IBRIX committed to providing rpm updates for security releases in a timely fashion

Credits

- IBRIX technology information courtesy of IBRIX and Dell Power Solutions Magazine: www.ibrix.com