Engineering, Operations & Technology
Information Technology

Shared Object-Based Storage and the HPC Data Center

Jim Glidewell
High Performance Computing
Enterprise Storage and Servers

Computing Environment

Engineering, Operations & Technology | **Information Technology** **Enterprise Storage and Servers**

- Cray X1
  - 2 Chassis, 128 MSPs, 1TB memory
  - 46 TB storage managed by ADIC StorNext HSM (5.5 TB online)
  - 8 TB of direct-attached short-term storage

- Linux Clusters
  - Systems:
    - 2 128 node dual-Xeon (32 bit) clusters
    - 2 128 node dual-Opteron clusters
    - 3 256 node dual-Opteron clusters
    - More on the way…
  - All Clusters share access to a second ADIC StorNext HSM

History and Current Issues

- Used DMF (Data Migration Facility) since the early-90's to manage disk space
- With the Cray X1, DMF was not an option
- An HSM was deemed essential
- Selected ADIC StorNext based on Cray support and recommendation
  - Initially for the Cray X1 only
  - Soon after, chosen for Linux cluster as well
- The I/O demands of the cluster were severely underestimated, as was the cluster growth rate
- As our clusters have grown, StorNext has developed significant performance problems

Hierarchical Storage Management - Pros & Cons

- Pros
  - Reduces storage costs
  - Makes highly efficient use of disk space
  - Allows users to view the storage available as “unlimited”
    - Eliminates need for user quotas
    - Reduces day to day storage maintenance issues
  - Simplifies detailed storage capacity decisions
  - Reduces backup requirements
- Cons
  - Administration is complex and time-consuming
  - User delays waiting for file retrieval
  - Data tends to build without bounds
  - Serious cleanup only occurs when a system is retired
  - Moving data from one HSM to a new one is very time consuming

Strategy for Shared Storage

- **Situation**
  - **HPC storage was tied to computing platform**
  - **No common storage for all HPC systems**
  - **Duplication of data as user processes use multiple platforms**
  - **Current Cluster SAN unable to deal with increasing load**
- **Needed a storage system**
  - **To serve as a shared repository for HPC data**
    - **Preferred direct access from cluster, NFS option**
    - **High-performance NFS from Cray X1**
  - **To serve as a high-performance replacement for cluster SAN**
- **Wanted a solution to serve *both* functions**
  - **Shared HPC permanent directory**
  - **Cluster home directory**
  - **Shared HPC temporary storage (7 - 30 days)**
  - **Cluster temporary storage (7 - 30 days**

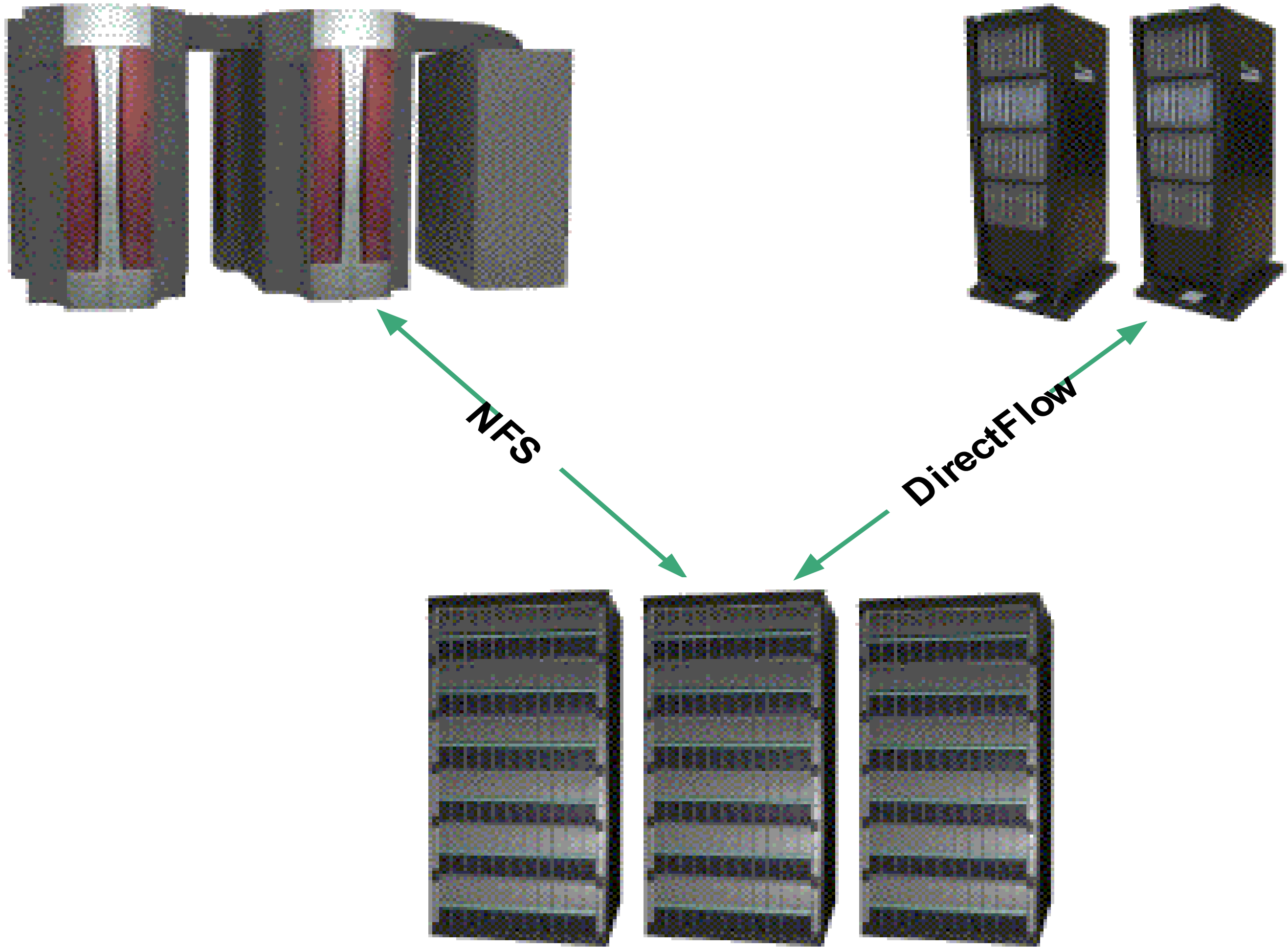
Storage System Selection Criteria

- Accessibility from all HPC systems
  - NFS from the Cray X1
  - Direct client access from Linux clusters preferred
- Availability
  - 24 by 7 uptime
  - Concurrent storage system maintenance
  - Reliability, resiliency, and redundancy
- Performance
  - Ability to operate with a large number of clients
  - High single-node performance
  - High aggregate bandwidth
  - Scalable performance
- Manageability
  - Ability to grow volumes seamlessly
    - No dump & reload
    - No performance penalty
  - Simple interface for management

Utilizing Panasas in Boeing HPC

- **Used for multiple functions**
  - **Linux user home directories**
  - **Shared HPC user storage**
  - **Linux high-speed temporary storage**
  - **Shared temporary storage**
- **Panasas directory for each user**
  - **Linux home directory is a subdirectory**
  - **Cray home directory remains on X1**
- **Shared home directory between systems not desirable**
  - **Different binaries, shell init scripts, etc.**
- **Common absolute path for permanent and temporary storage on all HPC systems**
- **DirectFlow access from Linux clusters, NFS from Cray X1**

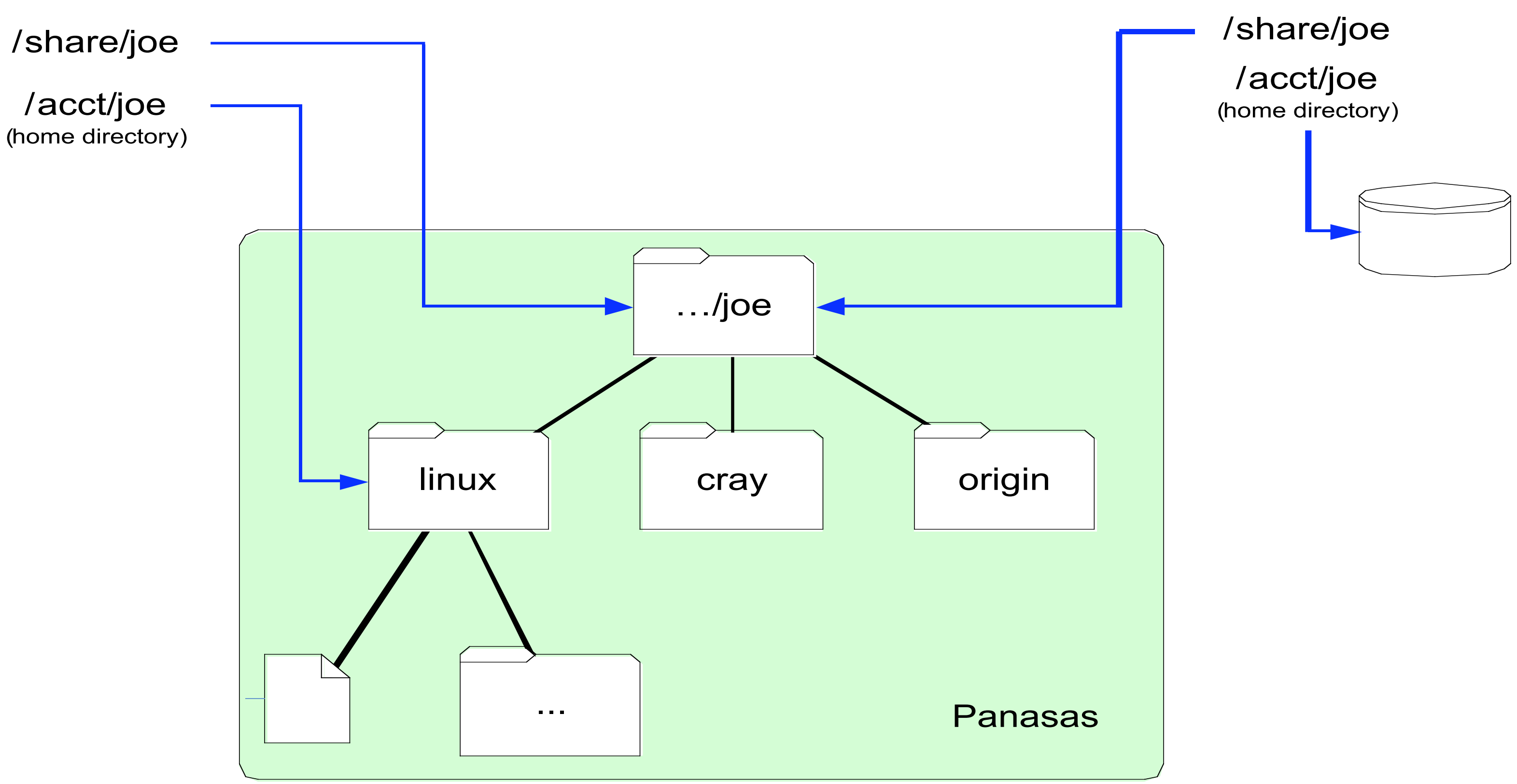
Panasas Access Methods



User Directory Structure

Linux Clusters

Cray X1



/share/joe

/acct/joe
(home directory)

/share/joe

/acct/joe
(home directory)

…/joe

linux

cray

origin

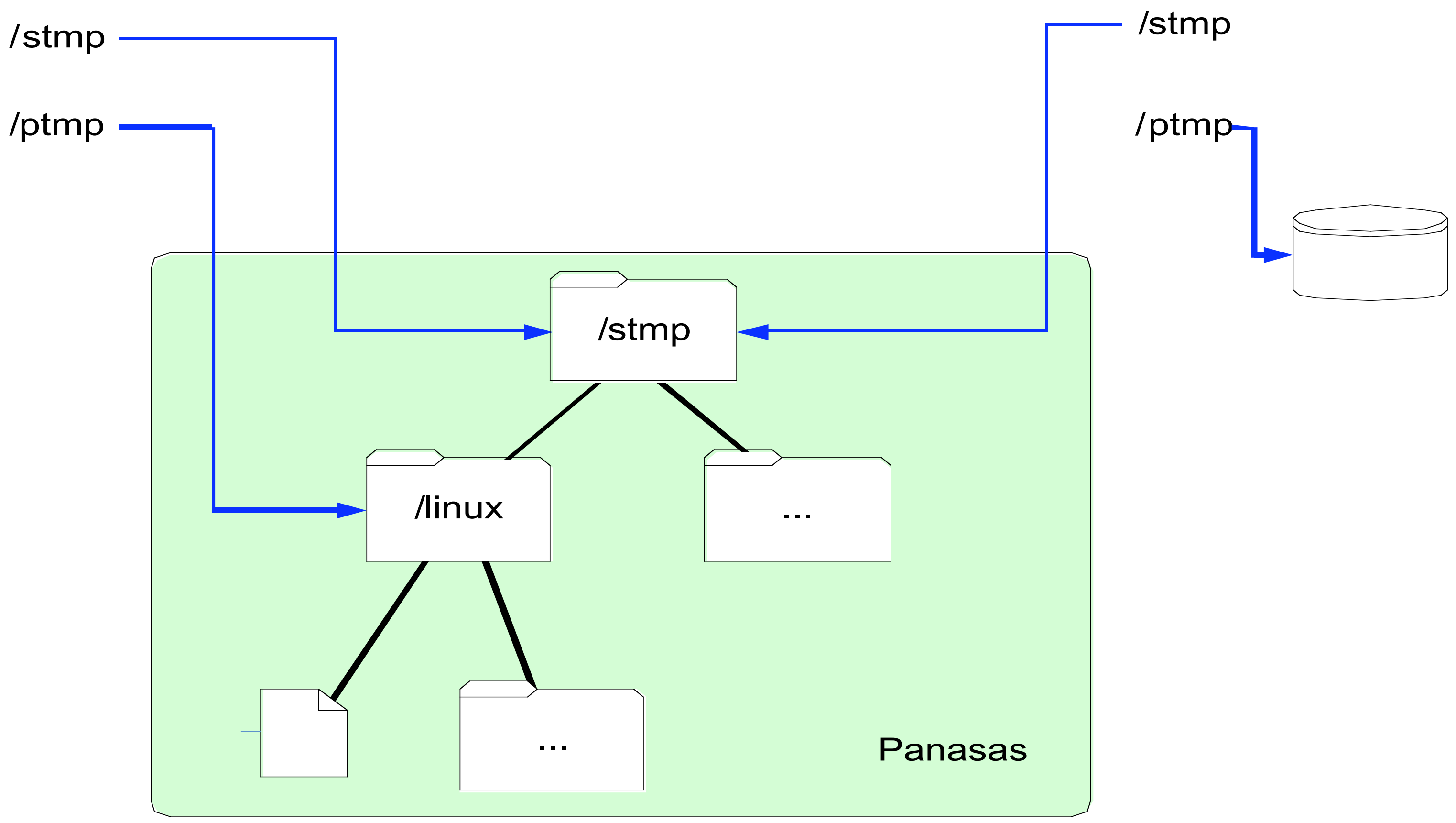
…

Panasas

Temporary Directory Structure

Linux Clusters

Cray X1



/stmp

/ptmp

/stmp

/ptmp

/stmp

/linux

...

...

Panasas

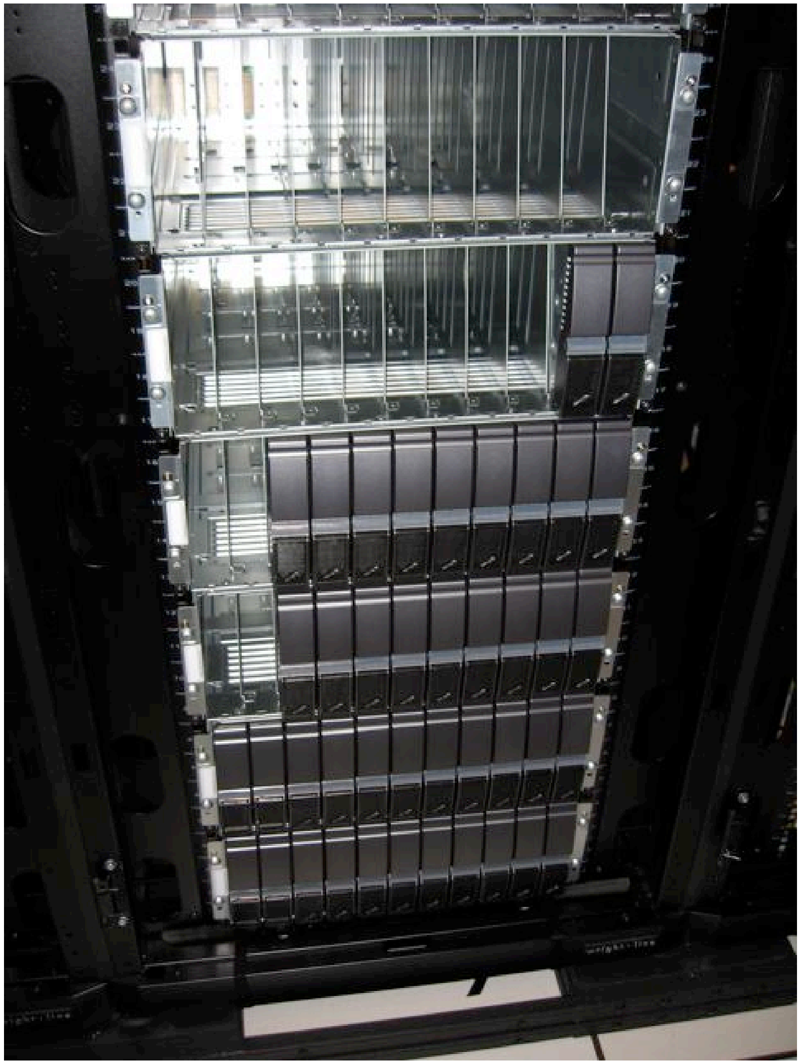
What is “Shared Object-Based Storage” ?

- ANSI Standard OSD-1 r10 defines the Object-based Storage Device (OSD) interface
- Multiple Vendors and Options
  - Lustre
  - Panasas
  - EMC
  - HP
- Files exist as one or more objects, rather than groups of blocks
- Storage is intelligent and can move these objects around for redundancy and/or performance
- Design goals are robustness, scalability, flexibility
- Storage interface is standardized, but metadata handling is proprietary

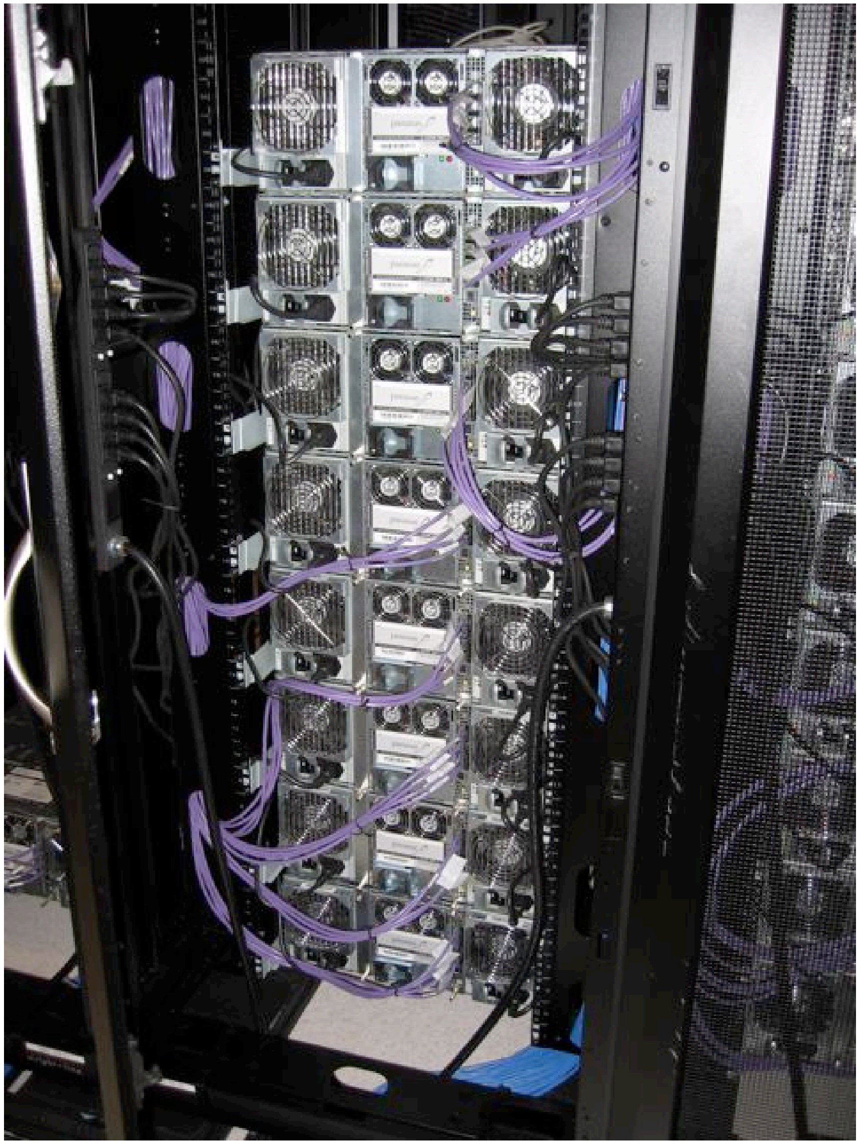
The Panasas Storage System

- **Realms, Bladesets, Volumes**
  - **Logically, Panasas presents itself as:**
    - **a single realm, containing**
    - **one or more bladesets, each containing**
    - **one or more volumes**
- **Shelves, Blades**
  - **Panasas hardware is delivered in**
    - **shelves (rack-mounted), which each contain**
    - **11 blades**
  - **Blades come in two types:**
    - **Director blades - manage metadata traffic, NFS access**
    - **Storage blades - contain drives & intelligent controller**
- **Access**
  - **DirectFlow client on Linux**
  - **NFS and SMB from other clients**

Panasas Hardware



Panasas Hardware (continued…)



Our Panasas Installation

- **Production - 52 Shelves of Panasas 3000 Storage**
  - **Each shelf contains 2 director blades and 9 storage blades (2+9)**
  - **500 gigabytes per storage blade**
  - **4.5 terabytes per shelf raw capacity**
  - **Seven racks, total of 234 terabytes raw capacity**

- **Evaluation System**
  - **3 1+10 shelves, 800 GB blades**
  - **Used for initial evaluation**
    - **Administrator training and familiarization**
    - **Validated bladeset expansion process**
    - **Very rigorous testing**
  - **Retained to test Panasas 3.x software**

HPC Panasas Storage

Engineering, Operations & Technology | **Information Technology** **Enterprise Storage and Servers**

Panasas Performance

- **Performance is a function of multiple factors**
  - **Network speed**
  - **Concurrent usage**
  - **Number of shelves in bladeset**
  - **Access method**
    - **DirectFlow for Linux clients**
    - **NFS/CIFS for other clients**
- **NFS speed from Cray X1**
  - **35 Mbytes/second**
- **Single Stream from dual-Opteron node (gigabit link)**
  - **Up to 85 MBytes/second**
- **Single shelf bandwidth**
  - **~ 300MBytes/second**
- **20 clients, 4 shelves**
  - **1.2 GBytes/second (60 MBytes/sec. average per client)**
- **Total aggregate bandwidth**
  - **Over 10GBytes/second - limited by network bandwidth**

Panasas Issues

- **Bugs reported and resolved**
  - **Evaluation system was extensively tested**
  - **A large number of support cases were opened**
    - **Gathering needed debug data was time-consuming**
    - **The vast majority of these cases were closed quickly**
- **System limitations**
  - **Needed to split realm - too many director blades**
  - **Unable to mix blade disk sizes within a bladeset**
  - **Scaling issues regarding administration**
    - **Time to reboot realm with new software**
- **Outstanding enhancement requests**
  - **Management of multiple realms by a single GUI**
  - **Site-defined metadata**
  - **Tool to get stat() data in bulk (similar to SGI_FS_BULKSTAT)**
  - **ACLs**

Backup Issues

- Storage growth is having a big effect on backup
- Disk and RAID systems capacity growth exceeds that of tape
- Traditional “Base + incrementals” backup strategy is becoming impractical
- Evaluated using the enterprise backup service
  - Adding our storage would double weekly backup
  - Required significant upgrade to their hardware
  - Weekly base dumps were not practical
  - “Synthetic base dumps” were an untried option
  - Analysis showed that after 12 months, >75% of all data being written to tape was data that had already been backed up
- HSM as a backup server…

HSM as a Backup Server

- **Basic Backup Strategy**
  - **User storage is not managed by HSM**
  - **HSM contains volumes and directories that match that of user storage**
  - **One-way file synchronization is done nightly**
    - **From user storage to HSM**
    - **Can be done on a volume or directory basis**
    - **Disk to disk copy**
    - **Uses “rsync” command**
  - **HSM migrates data to tape over time**
  - **HSM-aware backup facility**
    - **xfsdump -a …**
    - **Backs up inode information only**
    - **Data is on HSM-managed tapes**
- **HSM is not directly user accessible**

The Boeing HPC HSM Backup System - Specs

- **Hardware**
  - **SGI Altix 450**
    - **16 cores**
    - **48 gigabytes memory**
    - **24 fibre channel ports**
  - **40 Terabytes of DDN-based storage (InfiniteStorage 6700)**
  - **SUN/STK SL8500 Automated Tape Library**
    - **1500 tape slots**
    - **6 T10000 Drives**

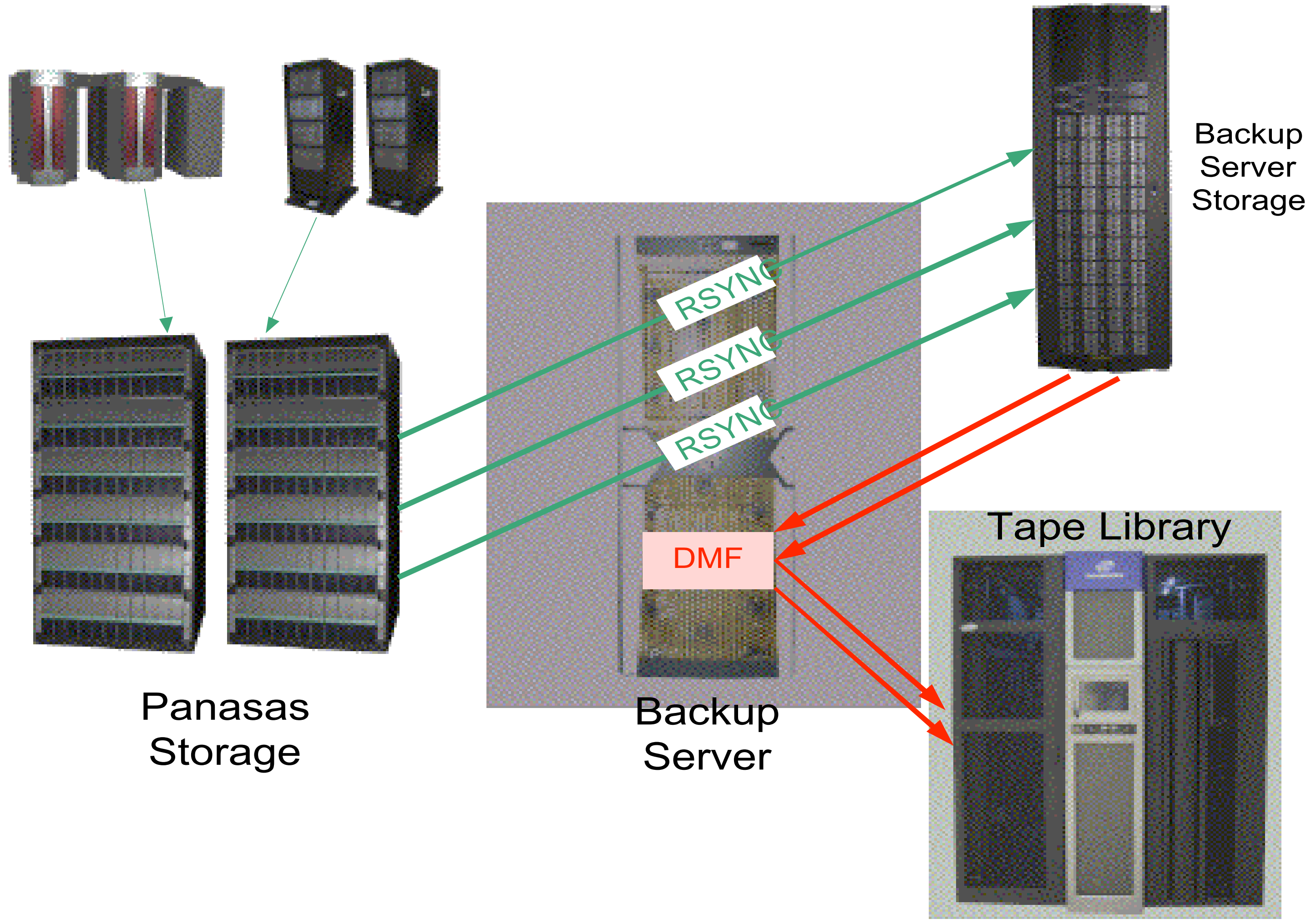
- **Software**
  - **SLES 9 + SGI ProPack 4**
  - **DMF 3.6**
  - **TMF**

The Boeing HPC HSM Backup System - Hardware

SGI Altix 450

STK SL8500

The Boeing HPC HSM Backup System



HSM as a Backup Server - Benefits

- **HSMs have proven functionality**
  - **Mature and robust products**
  - **HPC group has years of experience with DMF**
- **Data is written *once* to tape**
- **Optimized usage of tape media, drives**
- **HSM manages tape merges and “soft-deleted” data**
- **Fast recovery option in case of catastrophic failure of primary storage**
  - **Suspend all work**
  - **Mount HSM system in place of production storage**
  - **Resume production**
- **Option to use (part of) the backup server as a true HSM**

Summary

- Panasas has met our needs for a central HPC storage facility
- Performance via DirectFlow client is very good, NFS access from the Cray is more than adequate
- Panasas has provided very good support, and was very responsive to bug reports
- Evaluation system was very helpful tool for familiarization and testing
- The use of an HSM as a backup server has been a great success for us
- Users have been very happy with performance