BOEING

# X1 StorNext SAN

Jim Glidewell
Information Technology Services
Boeing Shared Services Group

# Current HPC Systems

- Two Cray T-90's
- A 384 CPU Origin 3800
- Three 256-CPU Linux clusters
- Cray X1

# Longtime DMF Site

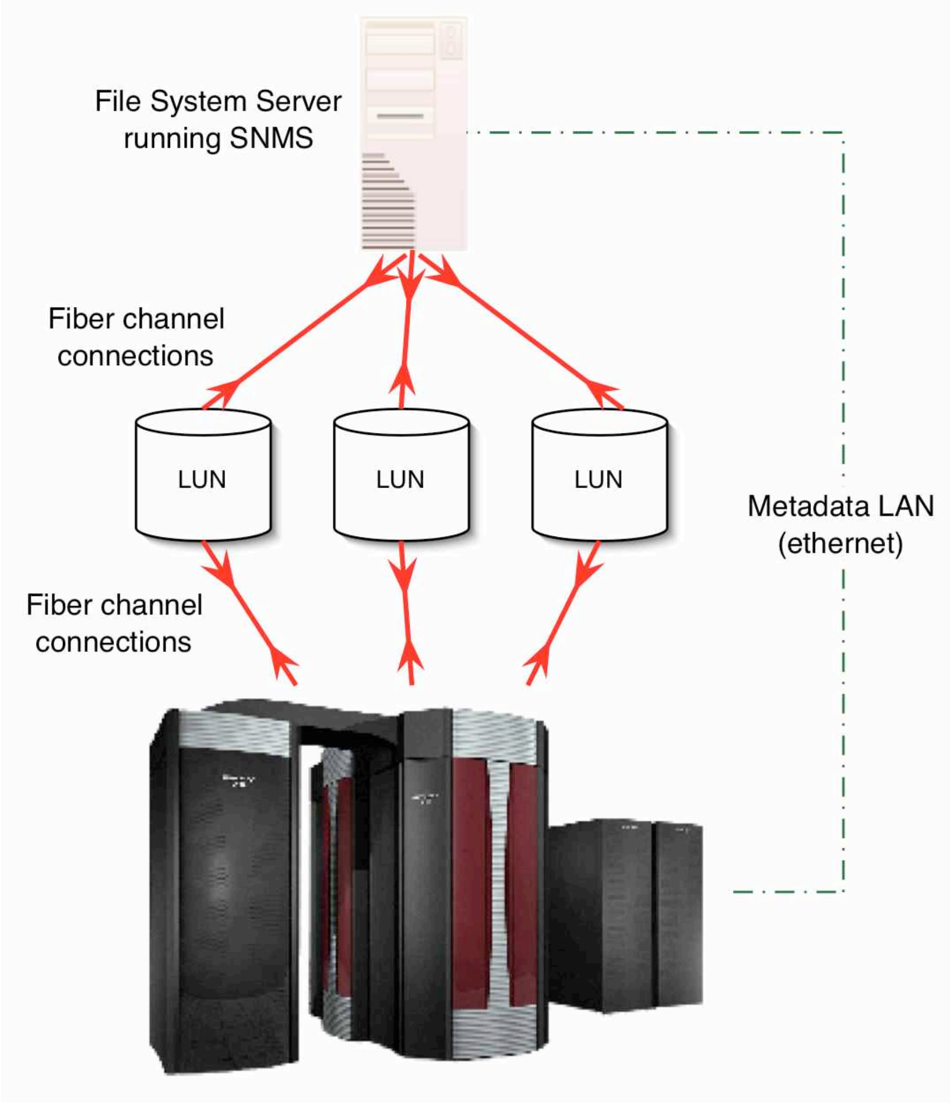
- Started using DMF in 1990 on Cray systems
  - 16 EScon STK 9840A drives shared between two T-90s
  - Past drives include 4490, 4490E, and Timberline
  - 37 terabytes of migrated T-90 data
- Using DMF since 1998 on SGI Origins
  - Six fiber-channel STK 9840B
  - Previously used Redwood drives
  - 38 terabytes of migrated data

# Hardware

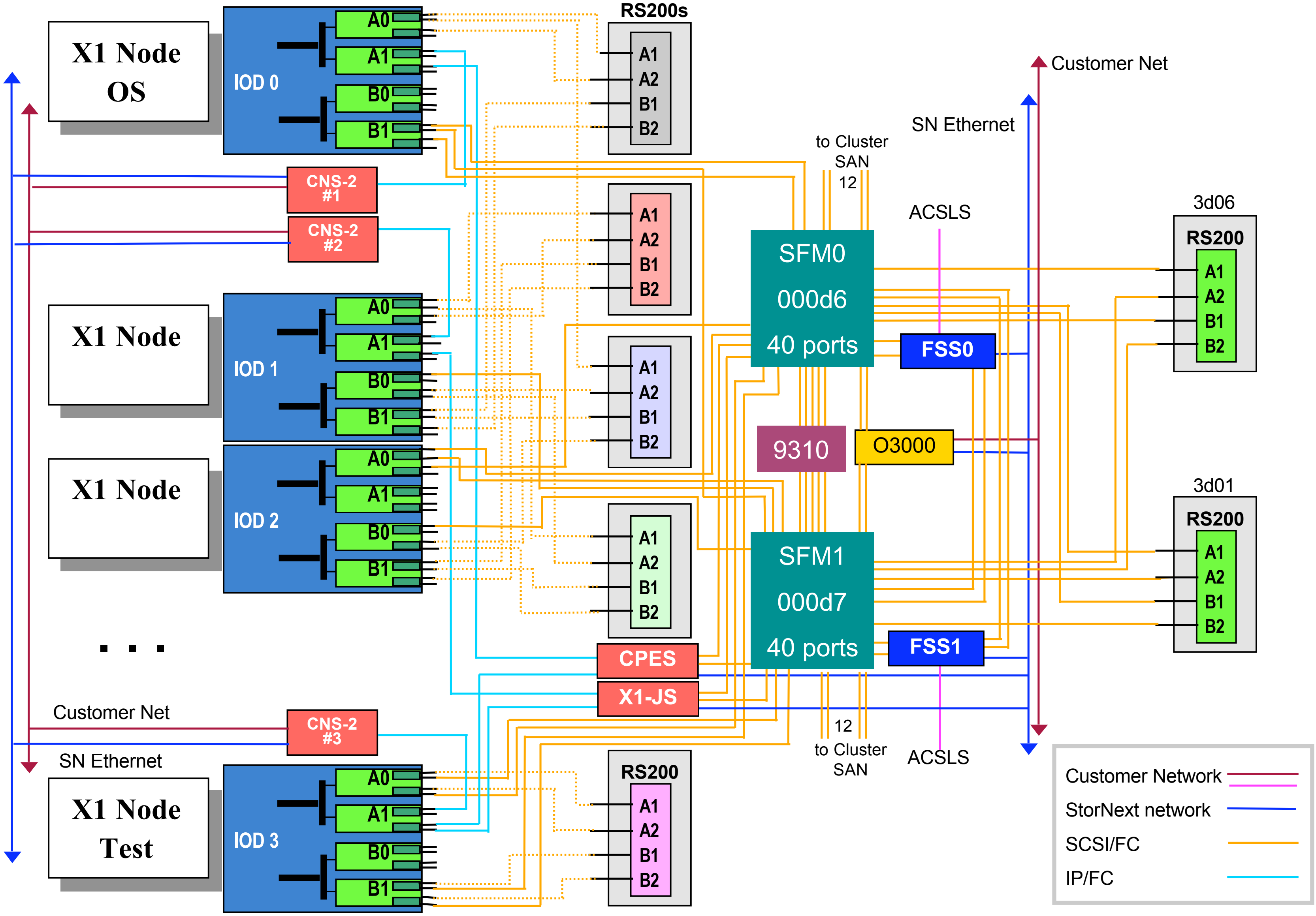
- Fully configured single chassis X1 (64 MSPs, 512GB)
- 26 TB of LSI RAID disk
  - 7TB currently configured into SAN
  - Roughly 1600MB/sec total bandwidth across all controllers
- Three Storagetek 9310 Powderhorn silos
  - Using ACSLS 7.0
  - Serving two T-90s, Origin 3800, Linux cluster, and X1
- ADIC SNMS SAN
  - Hosted on Dell 2650 servers running Linux
  - Six fiber-channel STK 9840C drives
  - STK drives selected for compatibility with other services in the datacenter
  - Chose 9840C over 9940C for faster load time
  - ADIC tape libraries used at other Boeing locations

# SAN Hardware

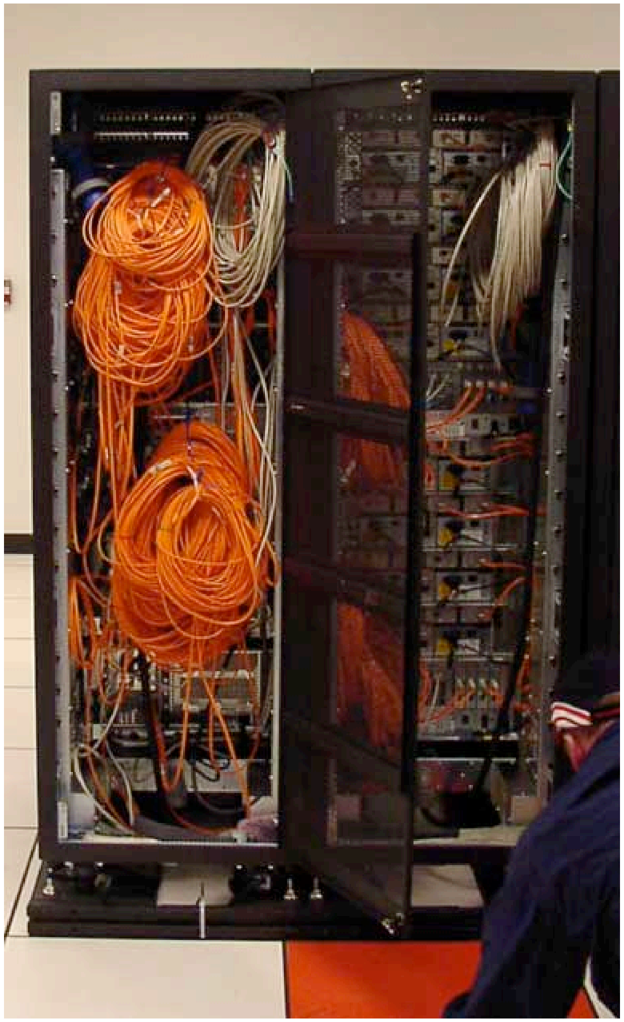
# Basic StorNext SAN Architecture



# Boeing SAN Cabling



# SAN Cabling During Installation



# Terminology

- SNMS - StorNext Management Suite
  - Manages migration, backup, monitors space, etc.
- SNFS -StorNext File System
  - Journaled file system - server and client
  - Appears as local file system to host
- Division between SNMS and SNFS seems “fuzzy”
  - SNMS is essential for a functioning file system
- Terms such as migrated, online, offline, etc. are same as DMF
- When files go offline, they are “truncated” by SNMS

# Setup and Configuration

- Goal was maximal performance with heterogeneous mix of file sizes, data access, and access patterns
- We did not have adequate guidance during our initial attempts at configuration
  - Our original plan had over 20 filesystems defined
  - We were told after the machine arrived that a reasonable limit was 8, four if failover was desired
- Setup took months - even with a Cray SAN analyst on site most of the time
- A permanent Cray test bed would have been useful
- Tuning guides, both general and client specific, would be helpful
- Training was adequate, but not in the depth we felt we needed
- Initial performance was less than expected…

# Performance

- Performance (especially writes) less than expected
  - Much lower than same disks directly attached
  - Appears to be an I/O bottleneck
- Mount options have hidden some of the performance problems
  - Same mount options on Linux made performance worse...
- X1 client very slow to remove large number of files
  - Other clients do not have this problem
- Deviated from Cray standard configuration to improve performance
  - Enabled CTQ (Command Tag Queueing)
  - Turned on write caching on LSI RAID controllers

# SAN vs. Direct Attached Disk

| | Reads | | Writes | |
|---|---|---|---|---|
| | Direct | SAN | Direct | SAN |
| cp | 130 MB/s | 250 MB/s | 120 MB/s | 30 MB/s |
| cvcp | 367 MB/s | 175 MB/s | 300 MB/s | 97 MB/s |

# Operational Issues

- Not as resilient as we had hoped
  - SNMS needs to be running for mv's to work
  - Recycling the MSM component may abort outstanding requests
  - Huge log files can fill /usr, requiring an eventual reboot
- The SNMS <->ACSLS (STK library) interface needs work in order to meet our needs
  - SNMS wants to control entire library (which is shared...)
  - Disaster recovery should not require whole library to be audited
  - Tape mounts pause as we enter tapes into the library
  - Still trying to understand the process of adding/removing tapes

# Operational Issues continued…

- fsmedcopy is slow and awkward
  - More testing of this feature is needed
- If file system server rebooted, clients sometimes fail to recover
  - Remount the SAN filesystem
  - Reboot the host system (!)
- Tested many failure modes before production
- Long list of SAN-related SPRs

# User Interface Issues

- No dmget equivalent  on host
  - This could be a significant problem in the future
  - Request for this functionality has been made to ADIC
- GUI access is preferred method of user control
- Slow response to metadata actions

# Production Experience So Far

- SAN production started April 25th
- No service interruptions in the first two weeks of production
- Most of our users never noticed
  - A good sign in a transition of this type…
  - No user complaints
- Despite some operational issues, StorNext provides valuable features for our site
  - Ability to manage data in excess of our physical disk capacity
  - Trickle backup
  - Option of sharing file systems  between heterogeneous hosts
  - Familiar “DMF-like” characteristics
  - Support for duplicate copies of backup media
- Cray and ADIC support was a key factor in our decision to move forward with production

# Linux server platform

- Linux, as a StorNext server platform, appears immature for Boeing needs
- StorNext only recently released for Linux
- Linux I/O is somewhat primitive for the needs of a StorNext file server
- Limited number of failover paths (max_scsi_lun)
- Good long-term strategy

# Conclusions

- Setup was significantly more difficult than expected
  - For both Cray and us
  - Took months instead of weeks
  - The lack of a StorNext test bed had serious impacts
  - Earlier direct involvement by ADIC would have been beneficial
  - Lessons learned at our site are likely to benefit others
- Boeing would have benefited from a more "turnkey" solution to managed storage
- Performance issues (particularly writes) need to be addressed for our environment
- Linux as a server platform is in line with long-term HPC strategy, but is still immature
- StorNext appears to be a solid product, but needs more polish for our needs
  - Ease of setup
  - HPC features (dmget, ACLs, etc.)
  - Handling of exceptions
- We have been in production since April 25th
- Cray has made significant efforts to make this installation succeed

# Coming soon…