



## **Integrating Fibre Channel Storage Devices into the NCAR MSS**

**John Merrill**

**National Center for Atmospheric Research  
1850 Table Mesa Dr., Boulder, CO, 80305-5602  
Phone: +1-303-497-1273 FAX: +1-303-497-1848  
E-mail: [jhm@ucar.edu](mailto:jhm@ucar.edu)**

**Presented at the THIC Meeting at the STK  
Bldg 8 Auditorium, 1 Storage Tek Dr.**

**Louisville, CO 80027-9451**

**July 22-23, 2003**



# National Center for Atmospheric Research





## Outline

- Brief history of Mass Storage at NCAR
- Overview of current MSS
- Current statistics
- Problems/limitations of current system
- Future plans for incorporating new devices



# MSS-I (1971)

- 1600-bpi tapes
- 200 Gbytes of data
- CDC 6600 and 7600
- Over 1,000 tape mounts/day



## MSS-II (1975 – 1985)

- AMPEX Terabit Memory System (TMS-4)
- Up to 2.5 Tbytes by 1985
- 2 Cray 1-A's
- rcp model – moves entire files to/from host at user request



## MSS-III (1985 – present)

- Re-implementation of MSS-II on newer hardware
- MSCP running on IBM platform, under MVS
- 3<sup>rd</sup> party transfers
- Initially only supported Bus/Tag devices
- First Powderhorn silo installed in 1989

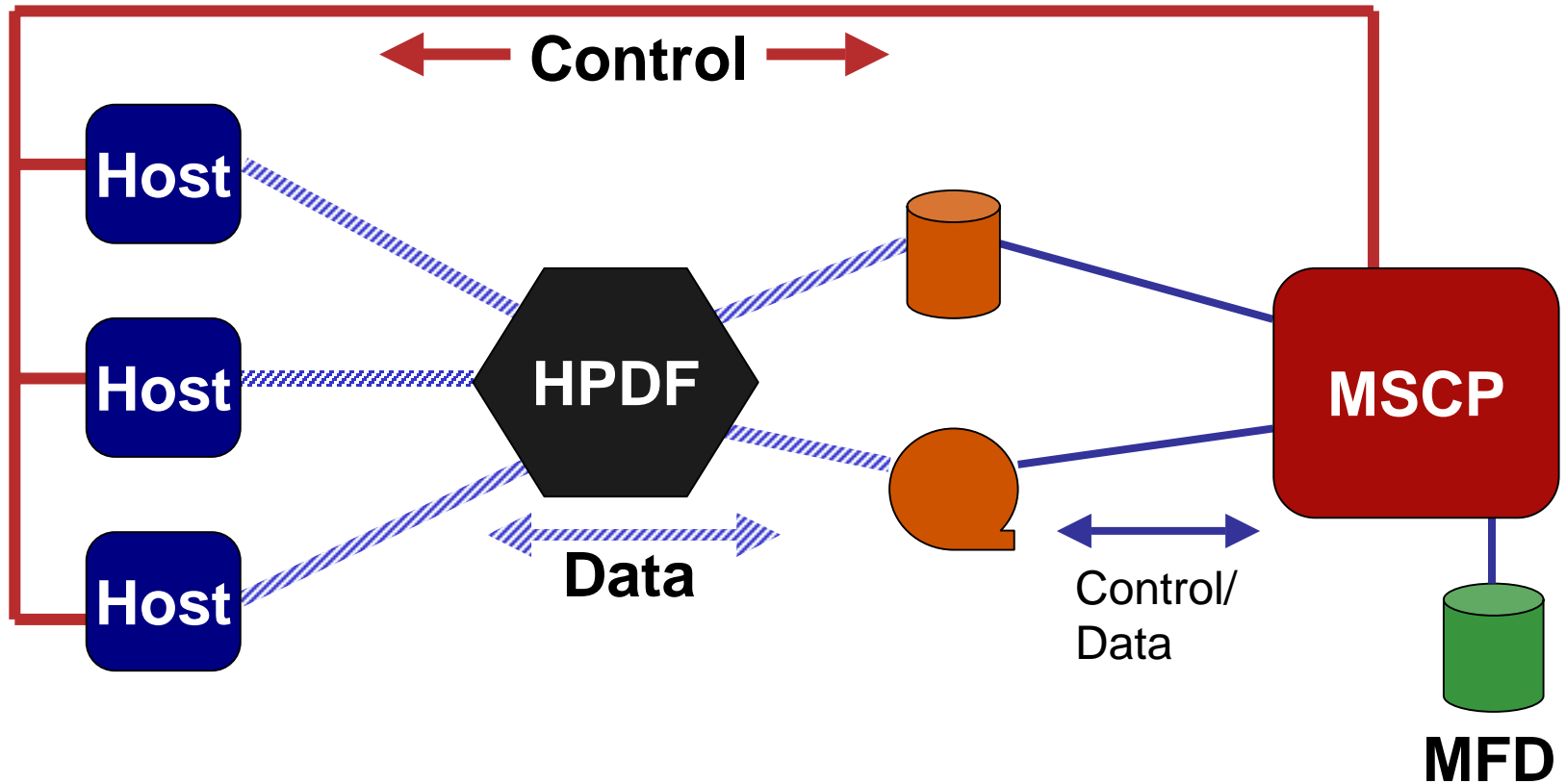


# Summary of Device Types used at NCAR

- Ampex TBM tapes
- IBM 3480, 3490E; STK 4490
- IBM 3390 disk farm
- STK SD-3 (Redwood) – ESCON
- STK 9840A – ESCON
- STK 9940A – ESCON



## 3<sup>rd</sup> Party Transfer





# HPDF

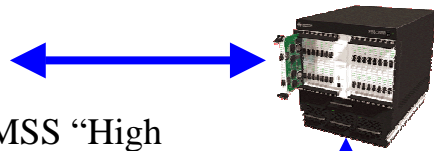
- 1986 – Network Systems Corp. Hyperchannel™ 50. Non-blocking switch using A-series adapters with multiple trunks.
- 1992 – HiPPI replaced Hyperchannel™ 50 (encapsulation)
- RDM/E emulates a host 370 FIPS-60 or ESCON channel.



MSS Command/Control Path  
to/from  
Supercomputers and servers

MSCP (IBM 9672)

HIPPI Switches



MSS "High  
Performance Data  
Fabric" to/from  
Supercomputers  
and servers

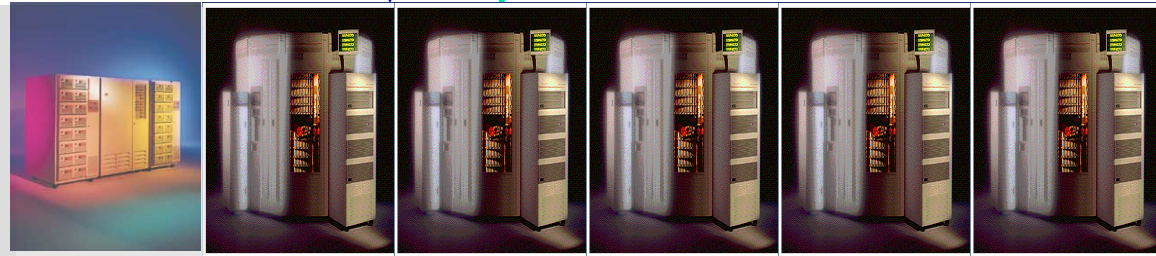


RDM/E

Data Migration  
And  
Import/Export  
server

Online Archive

Offline Archive



5 StorageTek 9310 Tape Libraries (9840 and 9940)  
Total capacity 1.8 PetaBytes with 60 GB cartridges

150K Tapes  
(3490E, Redwood, 9840, 9940)



# Current Statistics

- 1.35 PBs total, 791 TBs of unique data. (as of 1 July 2003)
- 18.8 M files
- 25 TB net growth rate per month (10 TB purged, 35 TB new data)
- 2.5 TB moved per day on behalf of user requests
- 2.5 TB moved per day for internal system migration, multiple copies, compaction, ooze, etc.
- > 5000 robotic tape mounts per day



# Sustained aggregate bandwidth

- 5 TB/day = 58 MB/sec (sustained 24x7)
- 8 TB/day = 93 MB/sec (Jan 2003 during Redwood offload)
- Currently total aggregate Bandwidth = 120 MB/sec.
- Peaks can easily max out the system. Need at least twice the current available BW.



# Problems and limitations of current system

- HiPPI is losing vendor support. Network Systems adapters are OLD technology, already at end-of-life.
- Only a limited number of supercomputers and servers have HiPPI interfaces
- Newest/highest capacity tape drives not being offered with ESCON interfaces. ESCON is too slow anyway. FC appears to be the way to go.



# Problems we had to solve to allow us to use FC

- MVS system unable to access FC devices, but still controls silo mounts/dismounts
- Didn't want to have to have FC interfaces on all MSS host machines, for a number of reasons



# Enter the “Storage Manager”

- Portable software that runs on most Unix platforms
- Can be distributed to improve throughput
- MSS hosts only need IP connection (GigE is preferable, but 100baseT will work)
- Talks to MVS system via IPC message passing software
- Controls access to all FC-attached devices



# Walk-through of an MSS write to STMGR disk cache

- Host contacts MSCP (MVS system) with request to write an MSS file
- MSCP validates the request, and contacts the STMGR
- STMGR allocates space on FC-RAID, and starts a task to handle the data transfer
- Session details are sent back to host via MSCP



## Walk-through (cont.)

- IP-mover on host contacts STMGR with info to identify the session via TCP socket
- IP-mover sends the file across the socket connection, STMGR task reads from socket, writes to FC-RAID (STK D178)
- Host notifies MSCP when transfer is complete. MSCP ties up loose ends, and updates the metadata catalog



# Differences for FC-tape

- MSCP may need to mount a tape (could be manual or robotic). Reads header record on tape via STMGR call.
- MSCP passes info about the transfer to the STMGR
- STMGR does tape positioning, writes header record for file being written
- Same as disk transfer at this point, except that a trailer record will be written at the end



# Current status of STMGR

- Disk cache version completed – runs on an SGI Origin 2100 running IRIX 6.5
- Host software completed and tested
- Started friendly user testing of 240 Gbytes of disk cache. Will replace existing disk farm.



# What's next?

- Put disk farm replacement in production
- Expand to include files up to 50 MB
- Complete changes to allow FC-tape devices
- Upgrade to 9940-B drives, start copying data from 9940-A to 9940-B.



# Future enhancements

- Run migrator natively on STMGR, and utilize direct FC connection to devices
- Expand disk cache to handle all user reads and writes, of all size files. Will need 20-40 TB of cache to do this.
- Move control of silo onto STMGR



# Questions?