THIC Inc.

The Premier Advanced Recording Technology Forum

NCAR

Almost

The NCAR MSS – Evolution of a Petabyte Archive

Gene Harano
National Center for Atmospheric Research
1850 Table Mesa Dr., Boulder, CO, 80305-5602
Phone:+1-303-497-1203 FAX: +1-303-497-1848
E-mail: snow@ucar.edu

Presented at the THIC Meeting at the
National Center for Atmospheric Research

Boulder CO 80305-5602

June 11-12, 2002

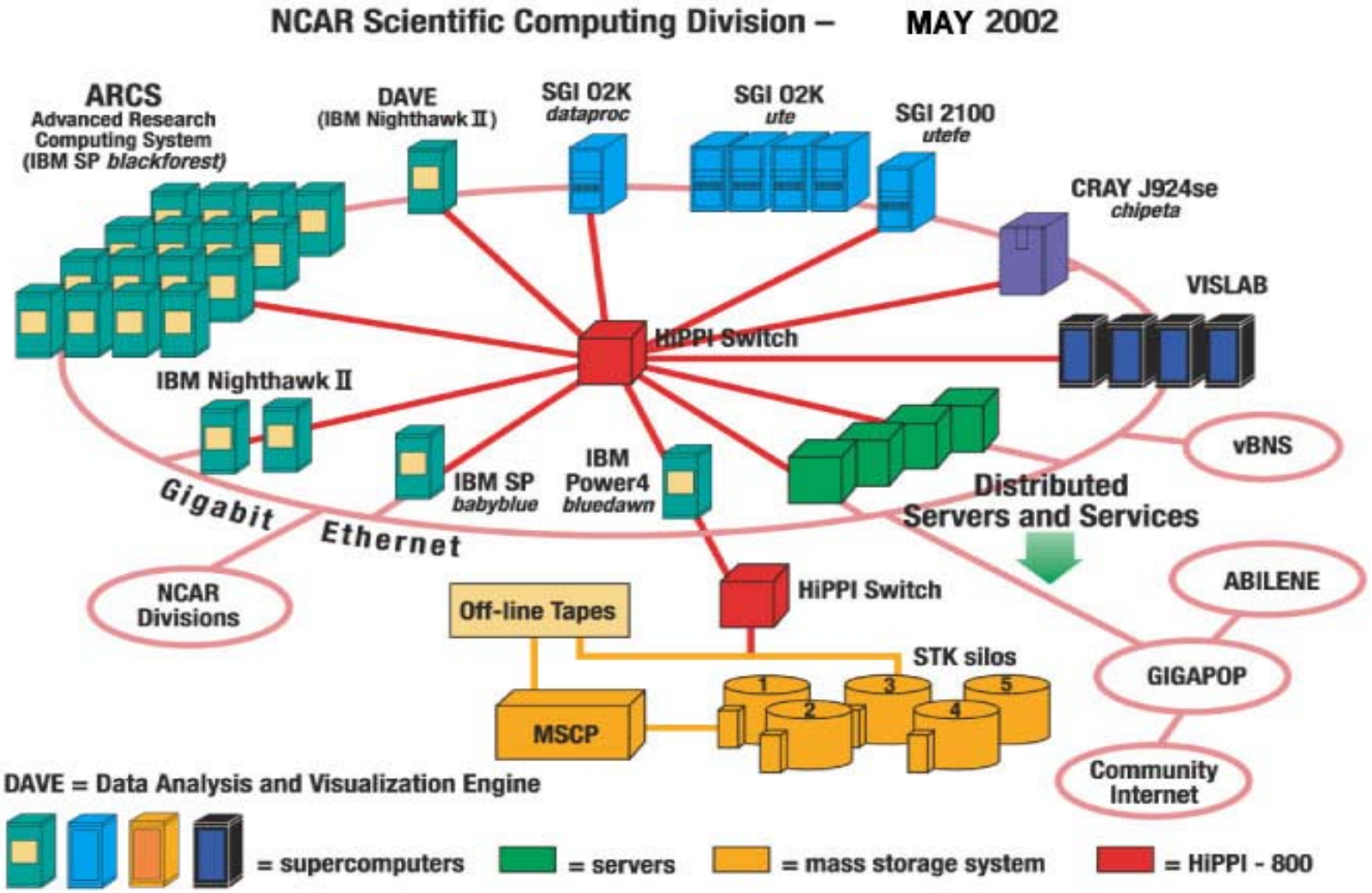
National Center for Atmospheric Research

Primary Data Source: NCAR HPC

- Computer Modeling
  - Coupled Global climate
  - Ocean, Atmosphere, Sea Ice
- Weather
  - Mesoscale and microscale prediction
- > 2 TFLOP peak; ~85 GFLOPS sustained
- 80% of the MSS data

Other Data Sources

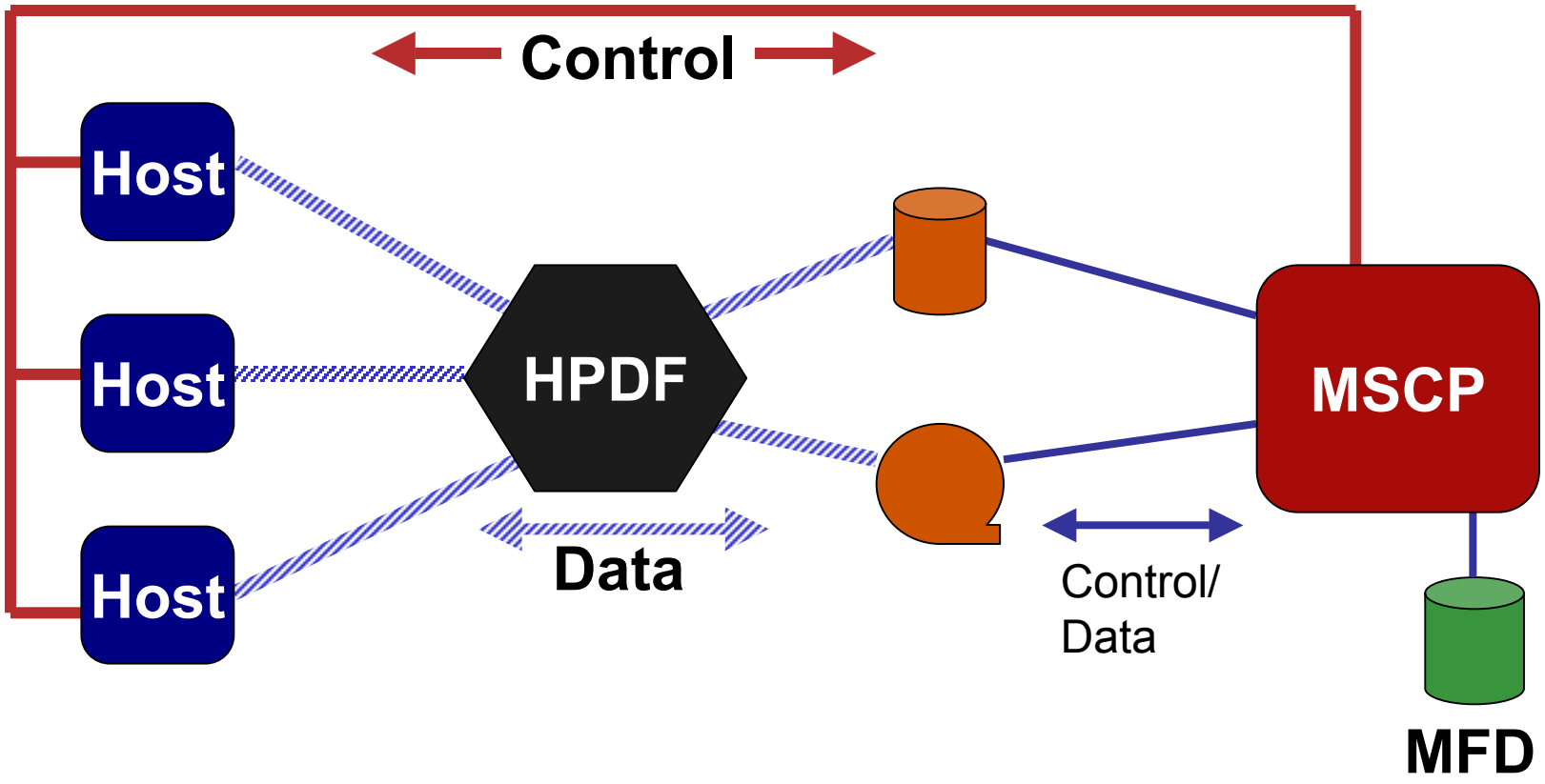
- Remote Sensing and Instrumentation
  - Satellite instrumentation
  - Airborne instrumentation and aircraft facilities
  - Ground and non-aircraft sensors
  - Radar
- 12% of the MSS data



NCAR MSS – Introduction

- Custom system operational since 1986
- Tape based archive with a small amount of disk cache for small files (internal storage hierarchy)
- Based on the IEEE Mass Storage Model Version 2
- rcp model – transfers files
- Implemented 3rd party transfers to optimize data movement

3rd Party Transfer



High Performance Data Fabric (SAN)

- NCAR MSS has used a SAN since 1986.
- Direct tape and disk access via the SAN
- 20+ hosts (including 2nd campus)
- HiPPI and ESCON based
- Moving away from HiPPI/ESCON toward FC & GigE
- Separation of control (IP) and data paths (HiPPI)

HPDF

- High Performance Data Fabric – direct tape access in addition to disk access.
  - 1986 – Manual Mounted tapes
  - 1987 – Disk Cache for small files
  - 1989 – Robotic mounted tapes
- No need to stage tape files to disk

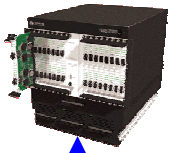
HPDF

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

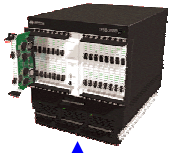
MSS Command/Control Path
to/from
Supercomputers and servers

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)

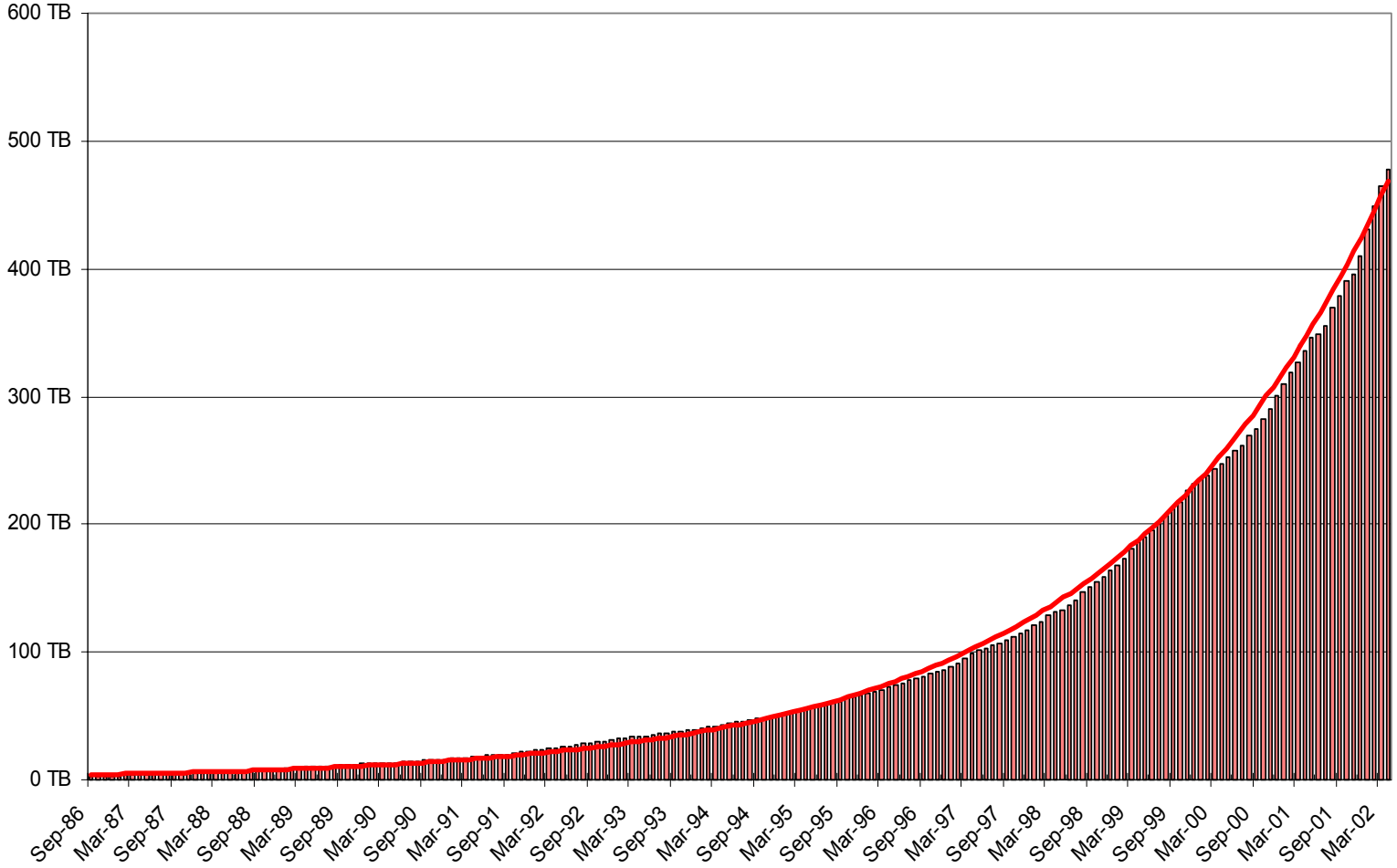
Current Statistics

- 884 TBs total, 490 TBs of unique data. (as of 27 May 02)
- 13.3 M files
- 20 TB net growth rate per month (10 TB purged, 30 TB new data)
- 1.5 TB moved per day on behalf of user requests
- 1.5 TB moved per day for internal system migration, multiple copies, compaction, ooze, etc.
- 4700 robotic tape mounts per day

Scalability

- Has scaled from 2 TB in 1986 to 884 TB today using the same basic core design.
- Has scaled from one Cray 1A to over 20 hosts representing over 2 TFLOPS peak.
- Keys:
  - HPDF (SAN)
  - Modular software design allows replication of functional pieces

NCAR MSS Growth



Data Longevity

- Data "ooze" – keeping the data alive. Media migration
  - 1986 –Ampex TBM system -> 3480
  - 1991 – 3480 -> 3490E
  - 1997 – StorageTek SD-3 (Redwood)
  - 1999 – StorageTek 9840
  - 2001 – StorageTek 9940

Data Ooze

- Plan for obsolescence
  - What is the useful lifetime of the technology?
    - What is the expected shelf-life of the media?
    - Will a drive/system be available to read the media?
    - What is the useful lifetime of the data?
  - Start migration allowing sufficient time before the system becomes obsolete
  - Worst case - lose half the useful system lifetime

Data Availability

- Maintain user access to the data while it is being oozed
  - Background process - Don't want to impact normal user access.
  - Takes longer to complete the migration unless additional resources are committed.

Perpetual Ooze

- Can large archives be completely migrated before the end of the useful lifetime of the technology?
  - Must have backward compatible drives
  - Concurrent migration from multiple technology sources

What does the future hold?

- Tape will continue to be the main storage media for the NCAR MSS.
- Size of the NCAR MSS disk cache will increase.
  - MSS log based simulator: 20 TB disk cache
  - Reduce tape recall which are extremely high latency and allow the use of higher density higher latency tapes.

Future - continued

- SAN vs NAS
- NAS will take over at the host connect level
- SAN will be used in the backend behind NAS servers.
- Because the “Common File System” will be a long time coming, if ever.

Future - continued

- Will tape die?
- Tape is still the best archival media.
- Disk will take over the backup/restore role
- The need for high performance high capacity HSM systems combining disk and tape will grow.

Questions?