



Friedrich Miescher Institute
for Biomedical Research
Part of the Novartis Research Foundation



Massively Scalable Storage Environment

Dean Flanders
Novartis Research Foundation division FMI

HMK Open Day 2009

January 28, 2009

© Copyright 2009 FMI

Outline

- Who is FMI
- The Storage Challenge
- Storage Requirements
- Coping with Storage Growth
- Issues with Traditional Solutions
- SAM-FS
- COPAN Systems
- SAM-FS + COPAN Systems at FMI
- Additional information
- Conclusion



Who is FMI?



Founded in 1970
Located in Basel, Switzerland

Friedrich Miescher Institute (FMI)

- Internationally recognized as a center of excellence in biomedical research with a strong record of innovation in the molecular biology of disease.
- Devoted to fundamental biomedical.
- Current research focuses on the study of heritable changes known as epigenetics, growth control and neurobiology.
- Part of the Novartis Research Foundation.

The Storage Challenge

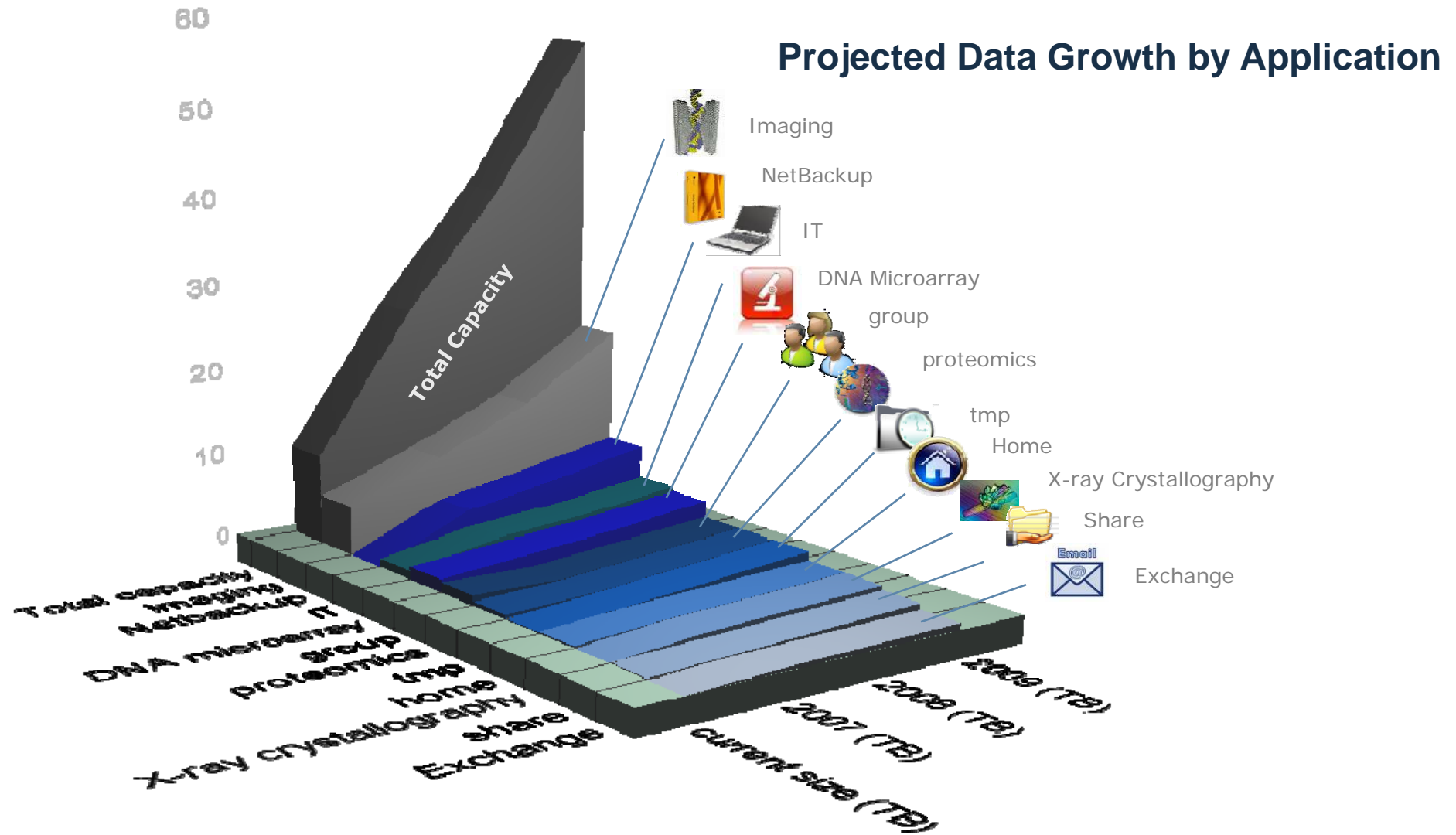
- Higher resolution instruments and new techniques require a highly scalable and a high performance storage solution.
- Terabytes of research data must be easily stored and remain easily accessible for the long term.
- Needed a cost-effective and realistic solution for backing up and restoring multi-terabyte file systems, while coping with limited power, cooling, and space resources.

Storage Requirements

- Highly scalable and easily expandable (*no chopping up file systems or re-designs in 1-2 years as with typical file servers*)
- Fast backup and recovery (*multiple scenarios*)
- Low cost (*initial and maintenance*)
- Keep terabytes of information online and available to the users (*tape too slow and file servers not practical*)
- Simplicity (*not too many moving parts or pieces*)
- Space, energy and cooling efficient (*easily overlooked*)

Storage Forecast at FMI

(already will exceed this)



A new way forward...

(storage continents)

Transactional Databases and Analysis (application data, block level, virtualization)

- I/O intensive
- Random read/write
- Small files
- Modest storage growth
- Steady growth rates
- Mission Critical
- Block-level virtualization
- Structured data (mostly)

Persistent Data Files, Data Protection and Archive Data (user data, file level, abstraction layer)

- Large files
- Very large storage
- Infrequent access
- Event driven
- Reference content
- Business vital
- Created but not modified
- Data accumulation
- Data integrity
- Long-term retention
- Explosive growth

*Within 30 days the majority
of data becomes persistent data*

Analysis



20GB

Database



8MB

Backup



10MB

Replication



20MB

Maps



60MB

Video



300MB

Imaging



48GB

Document



80KB

Transactional data

Persistent data

Amount of data in the typical enterprise

DR Copy

Challenges with Tape Libraries

- *No online data access*
- *Data integrity and security concerns*
- *Slower access times*
- *Space constraints*



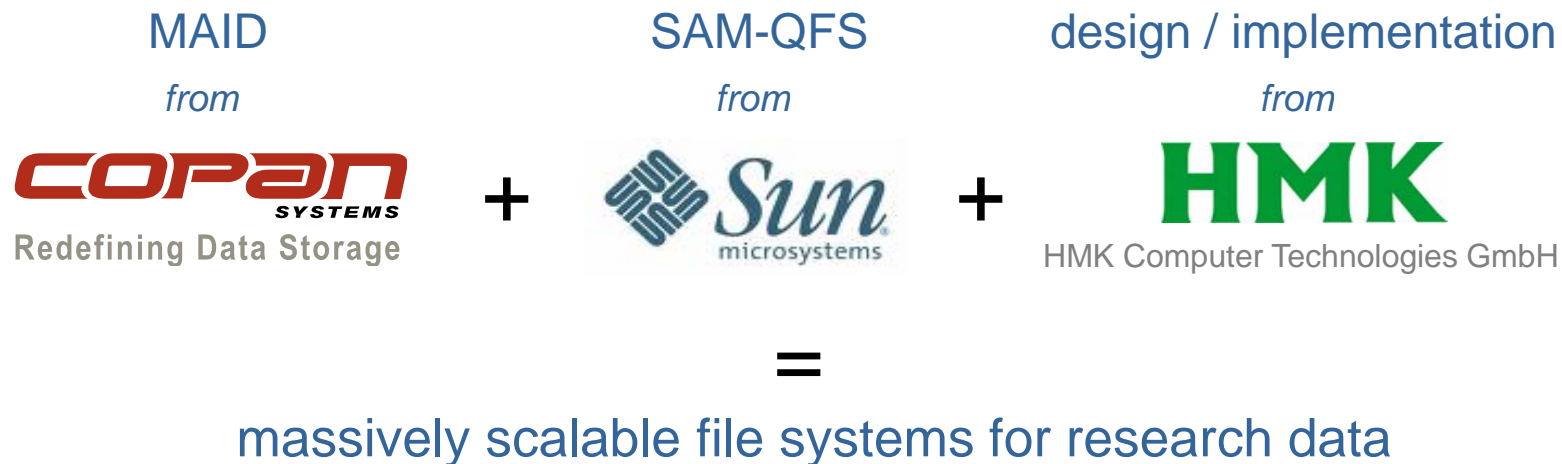
Challenges with Traditional Disk Arrays

- *Power & space inefficient*
- *Cost prohibitive*
- *Significant cooling requirements*
- *Substantial provisioning of power*
- *Breaking up file systems into chunks*
- *Backup / restores of large file systems*
- *Snapshots are not backups...*



Persistent Data

- Shared HSM File System SAM-FS from Sun Microsystems with Enterprise MAID technology from COPAN Systems
 - A new massive file system (no need to break file systems into chunks)
 - No stress on FMI's existing data center infrastructure
 - Immediate access to all data
 - Ability to restore multi-terabyte file systems in minutes



SAM-FS

(file virtualization / CDP)



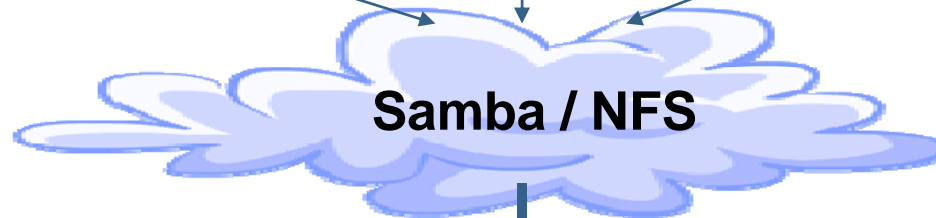
applications



network drives



scratch areas



Why SAM-FS?

- **Data Protection**

- Copy of files to target 1, target, 2, etc. can be set to happen as early as the close of the file, providing instant archiving and backup.

- **Caching**

- The file system serves as a cache for files, immediately copying files to the archive and backup. It works on the “high watermark” principle, it will release data from the file system when the mark is reached (leaving them only in the archive and backup), freeing space nearly instantly. When data is accessed it can be copied back to the cache or streamed directly from the archive or backup.

- **File System Recovery**

- Recovery of the file system requires only restoring the information pointing to the archive (inodes). Restoring takes only seconds, irrespective of the amount of data originally on the file system. Data in the cache at the time of the inode backup can be migrated back transparently and logs can be examined to find files that were stored after the backup of the inodes.

- **Metadata Server Corruption**

- Metadata can be rebuilt from the inodes in the file system.

- **Transparency**

- Integration with Samba (Windows file access server) and quotas allows users to see directly from Macintosh and Windows (NFS for UNIX) their files. A user can see how much data is being stored online and in the archive. As a result, volumes can be any size.

- **Expertise**

- Working with a HMK Data who has done approximately 300 installations of SAM-FS and have worked with the product since 1995 (they are the original re-sellers in Europe).

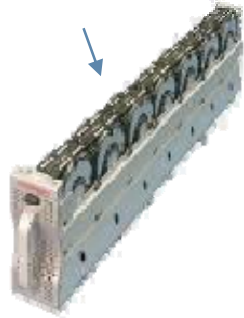
- **Simplicity**

- Very simple system, very few moving parts, very well tested.

COPAN Systems

COPAN
SYSTEMS
Redefining Data Storage

14 X SATA drives



112 X SATA drives
per shelf, can expand to
8 trays with 886 drives total



- **Disk based storage which gives nearly instant access to files**
(tape can take 60-90 seconds)
- **Scales to 896TB per system**
(1 TB drives, drive sizes can be mixed, compression can potentially double capacity)
- **25% of the disks spinning at one time**
(reduce heat and power consumption)
- **2.75 TB per hour transfer for 300T**
(4 fiber channel interfaces)
- **5.25 TB per hours for 300TX**
(8 fiber channel interfaces)
- **Presents itself as a VTL**
(De-dupe and replication, with CIFS, NFS, native MAID, etc. coming soon...)
- **Much higher density than other archive solutions**
(e.g. EMC Centera to the left)



COPAN |-----| EMC |-----|

Why COPAN Systems?



← Initial FMI configuration – 1 Shelf (40TB useable)

Hypothetical Expansion Scenario

(can be done without re-design / no downtime / no power or cooling changes)



- 2006 – Shelf of 500 gig drives (56TB raw)
- 2008 – Shelf of 1000 gig drives (112TB raw)
- 2010 – Shelf of 1500 gig drives (168TB raw)
- 2012 – Shelf of 2000 gig drives (224 TB raw)

- 2014 – Shelf of 2500 gig drives (280TB raw)
- 2016 – Shelf of 3000 gig drives (336 TB raw)
- 2018 – Shelf of 3500 gig drives (392 TB raw)
- 2020 – Shelf of 4000 gig drives (448TB raw)

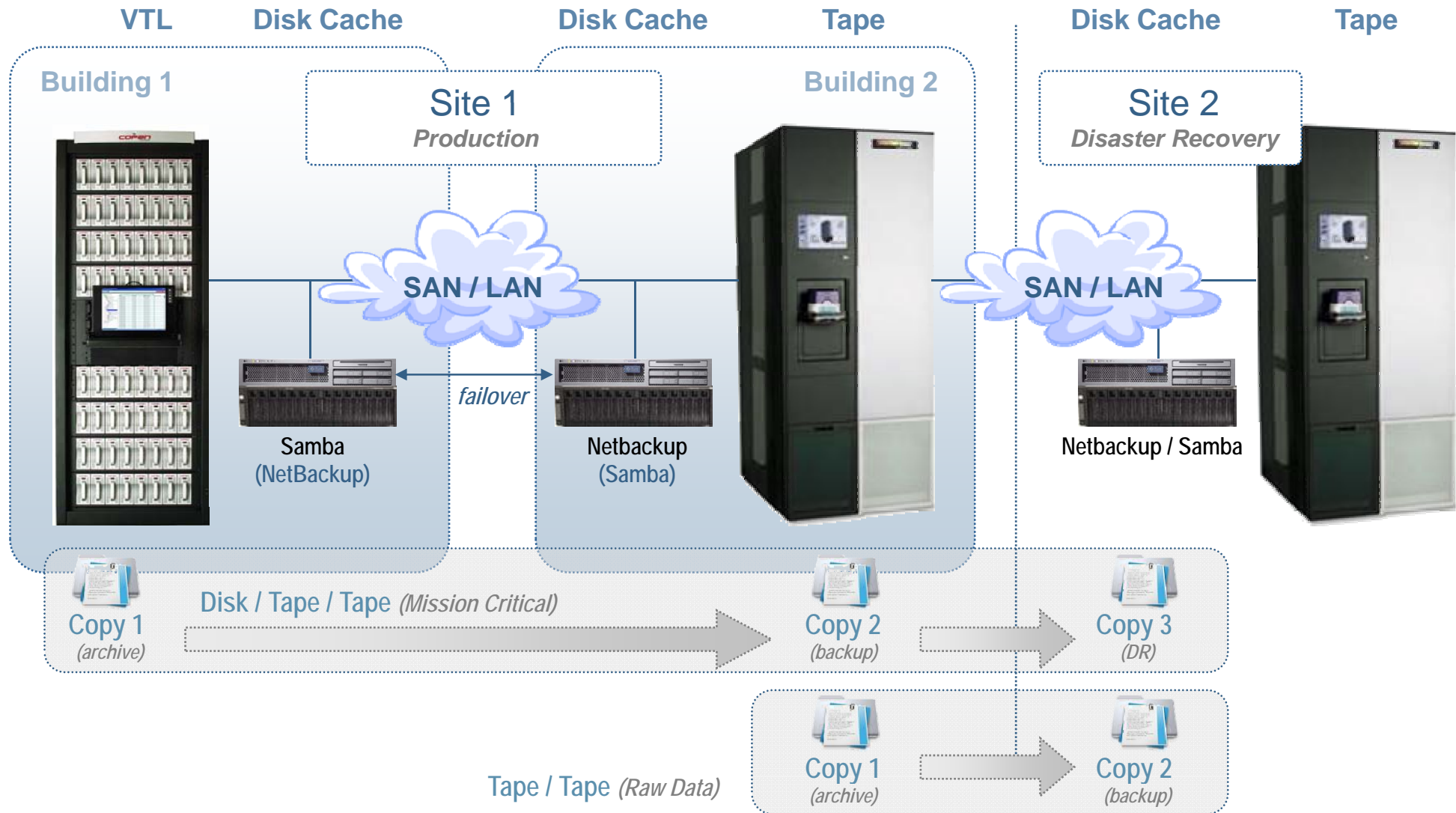
2016 TB



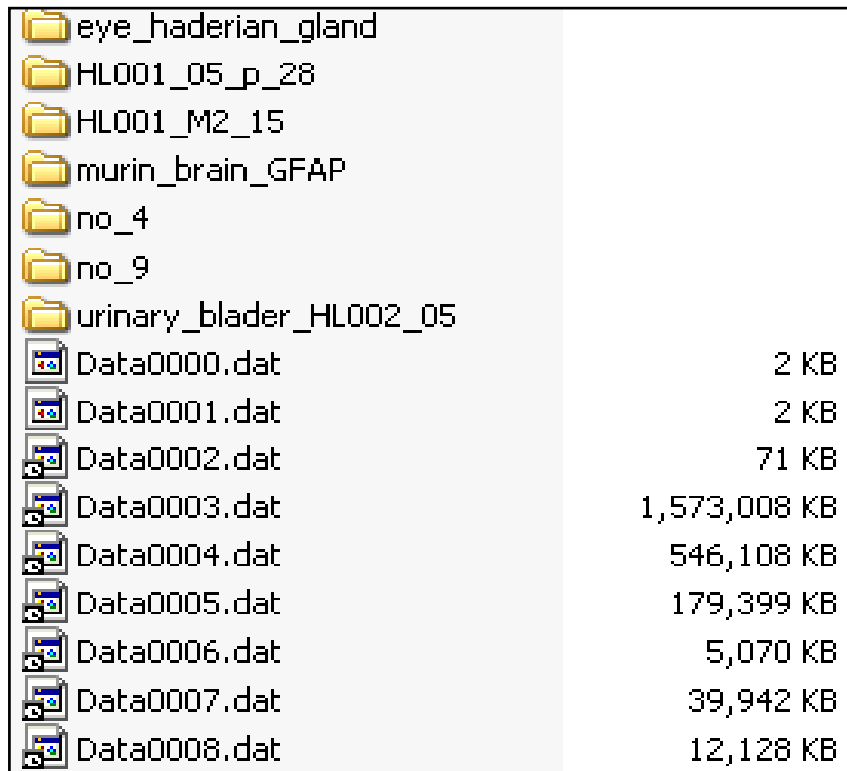
2022 data from shelf 1
migrated to shelf 8,
shelf 1 upgraded
to 4.5TB drives (504 TB raw), etc.

Storage Solution @ FMI

(Next Month)



How the user experiences SAM-FS / COPAN Systems

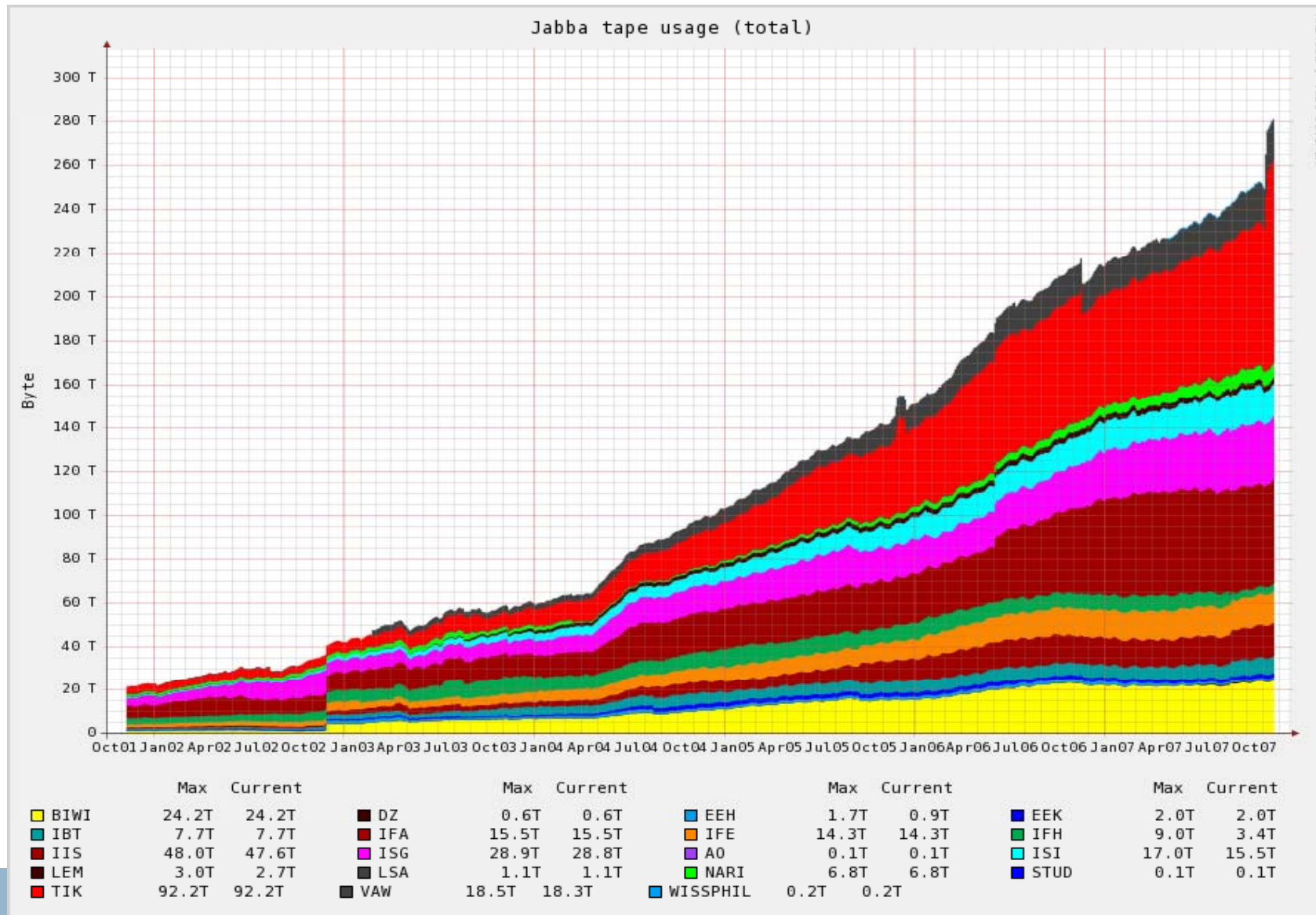


eye_haderian_gland	
HL001_05_p_28	
HL001_M2_15	
murin_brain_GFAP	
no_4	
no_9	
urinary_blader_HL002_05	
Data0000.dat	2 KB
Data0001.dat	2 KB
Data0002.dat	71 KB
Data0003.dat	1,573,008 KB
Data0004.dat	546,108 KB
Data0005.dat	179,399 KB
Data0006.dat	5,070 KB
Data0007.dat	39,942 KB
Data0008.dat	12,128 KB

Benefit to user:

- Nearly infinite storage
- First byte to stream ~30 seconds (file copy starts during staging)
- Subsequent files like disk.
- Much better at repositioning of files than tape (VTL)
- ability to have terabytes of storage available quickly (requirement these days)
- transparent view into storage spaced used (disk and tape), *use of quotas with dfree in samba gives view in Explorer*
- tape / disk areas
- NFS / samba access

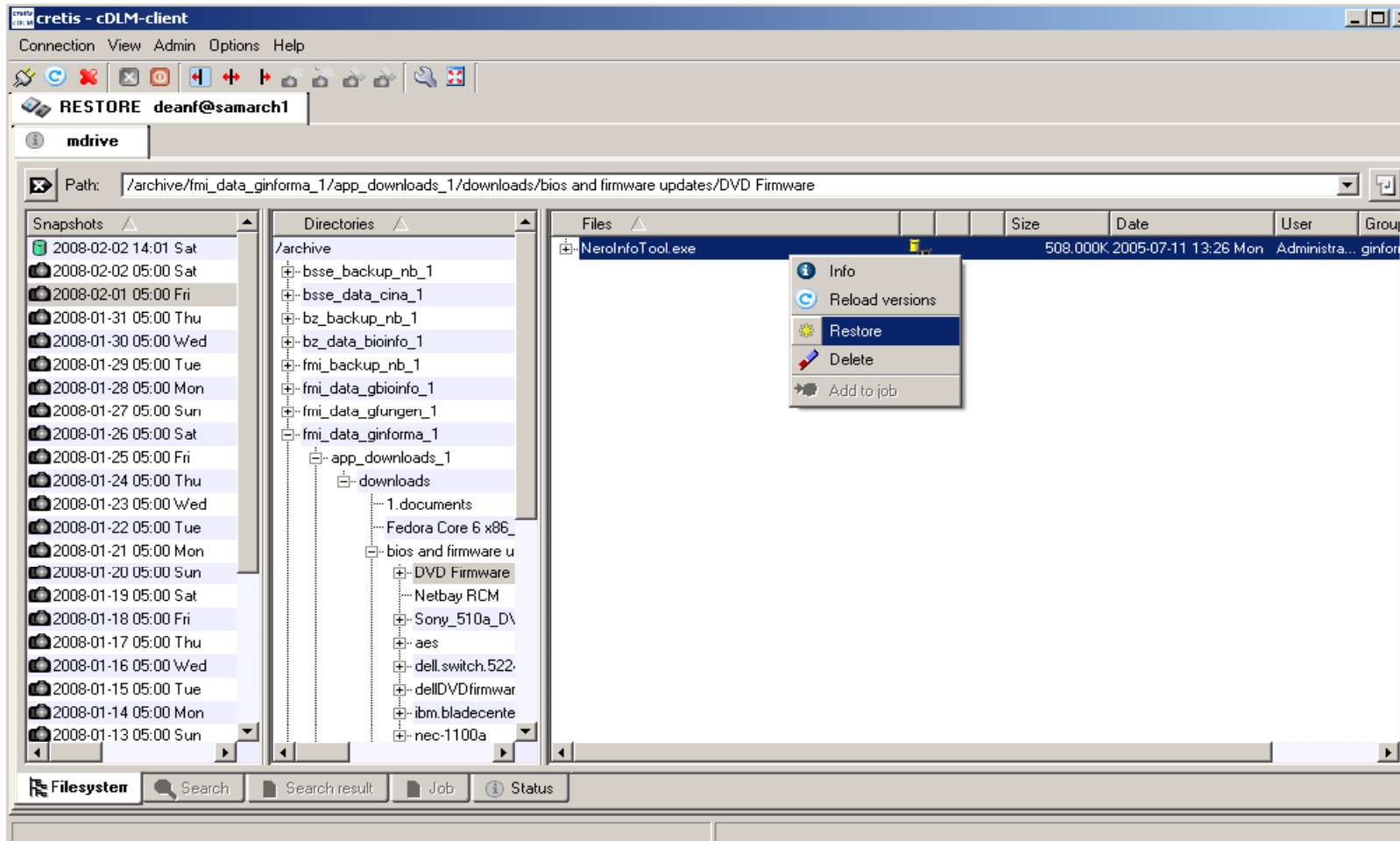
Scalability of SAM-FS @ ETH Zurich



- ***SAMnet***
enhanced samba for SAM-FS
- ***cDLM***
data life cycle management tool that allows a user to see versions for files
- ***SAM fast media refresh / recycle***
allows for 5-10 times faster media recycling than typically in SAM-FS
- ***Open source tools***
Samba enhancements, du2rrd, etc.

cDLM tool

(file / versioning / “snapshot”)



Delegation of file restores or user self-service.

Important Considerations

- **Single stream performance of Copan**
(read and write, but user does not detect)
- **Bottlenecking at the VTL**
(resolve with cache)
- **Cache sizing**
(need to be able to write one entire tape out into cache, helps with issues above)
- **Tape only copies**
(saves you money)
- **Number of files**
(can kill performance on any file system)
- **ACLs in SAM-FS**
(can be worked out)
- **No DVE support in SAM-FS**
(can be worked out)
- **Have a third copy of critical data...**
(cover your bases)

- ***Cost Effectiveness***

The proposed solution covers primary storage and archive/backup at a very reasonable cost for this scale.

- ***Scalability***

Extreme scalability in the high performance and archive/backup area.

- ***Recoverability***

Servers and terabytes of data can be recovered in minutes and easy recovery is possible from multiple failure scenarios.

- ***Simplicity***

It is not a complex solution. Standard network protocols are being used. Storage can be quickly provisioned either in the primary storage area or the archive area.