Mass Storage at CERN

GDB meeting, 12. January 2005

Tape Storage Installation

- StorageTek (STK)
- 10 Powderhorn Silos, distributed over two physical sites (5 silos each)
- Each silo with 6000 tape slots
- Today 10 PB maximum capacity
- 9940B tape drives
  - 30 MB/s read/write speed
  - 200 GB cartridges
- 54 drives installed
  - 23+23 for physics production
  - and 8 for Backup procedures
- Still 10 * 9840 drives in production

Disk Storage Installation

- NAS disk server connected via Gigabit Ethernet 9single connection per server)
- 1 – 3 Terabyte per node (usable space)
- One large file system (striped+RAID5) or multiple file systems with mirrored disks

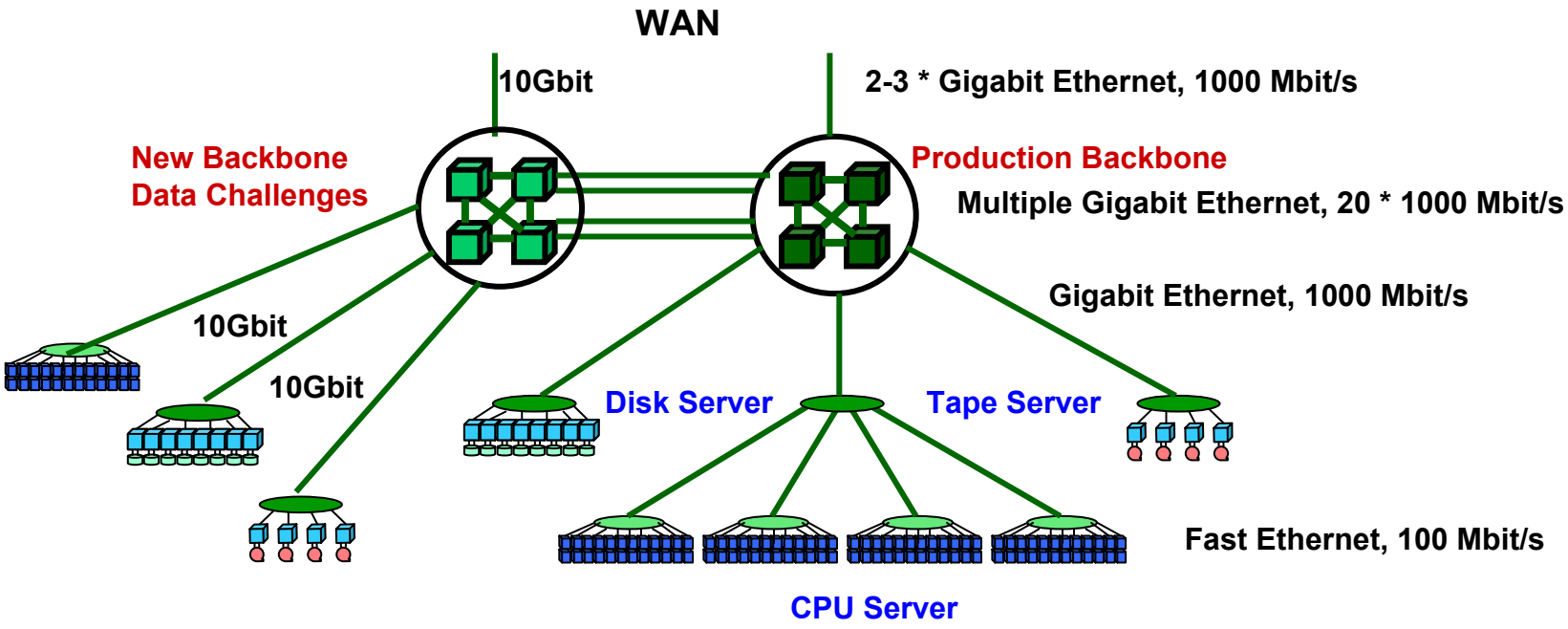
- ~400 disk server nodes
- ~450 terabytes of space (usable, RAID config)
- ~6000 single disks (75 – 200 GB)
- Spread over ~80 separated disk pools (aggregation of file systems and server)

Mass Storage Software

- CASTOR
- CERN development
- new version in final testing phase
- Today ~30 million files and ~4 PB of data (15% LHC experiments)
- ~4.5 FTE development team
- ~7 FTE operation team (disk, tape, Castor) high level, not sysadmin

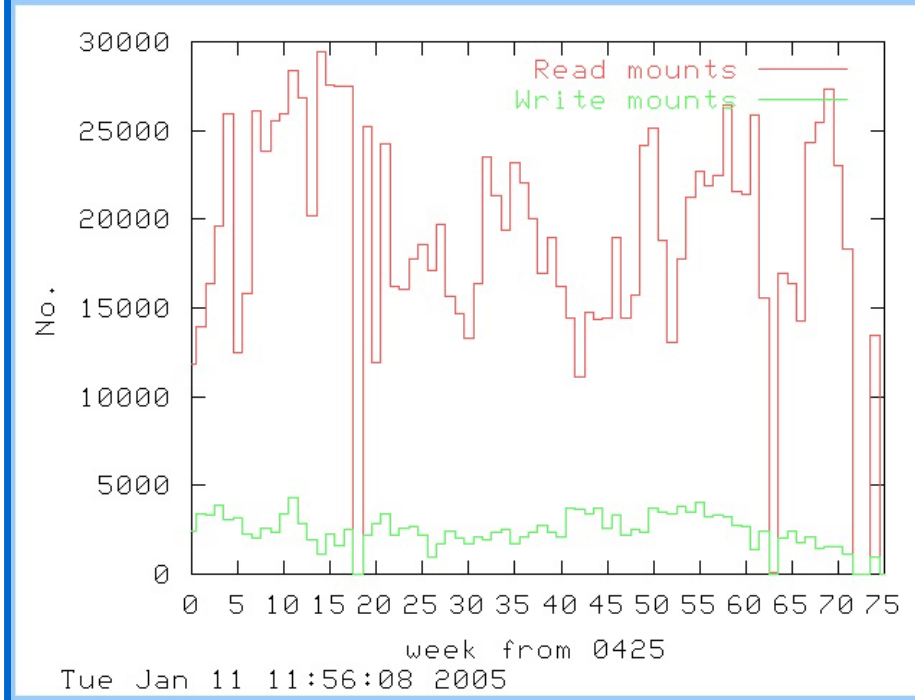
Network Installation

There is a parallel installation of networks, which are interconnected



Tape System performance (I)

Fig. 1 ACLS4 Read and Write mounts per week. (Weeks 0425 to 0501)



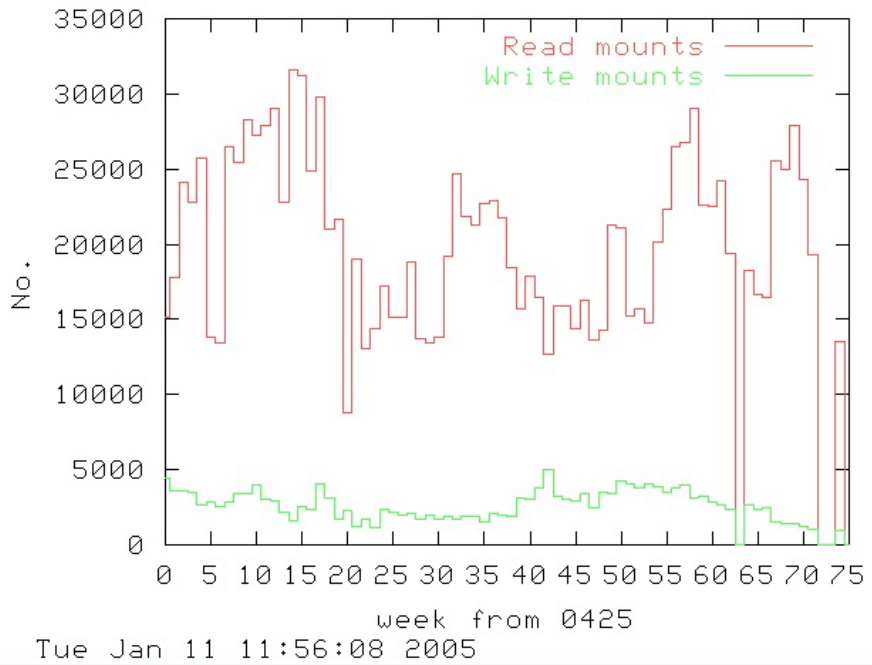
Number of tape mounts per week in the two silo installations

>50000 per week for an average of 40 active tape drives

reaching the limits of the robotics

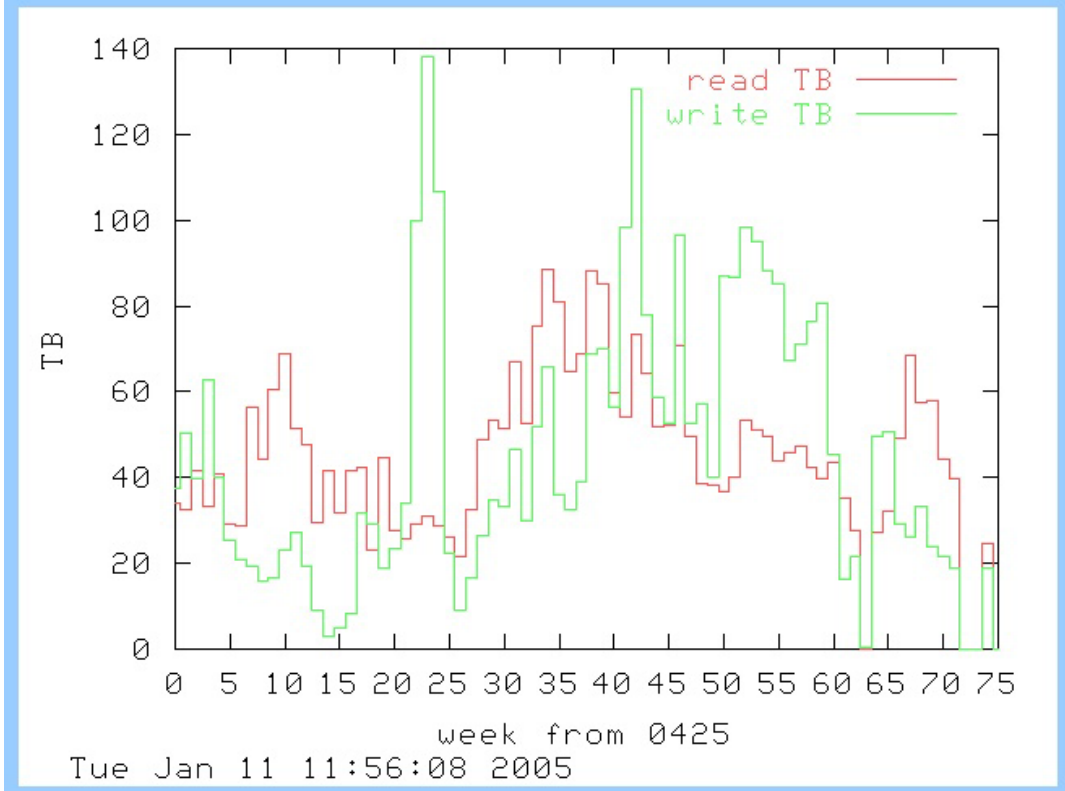
locality of tape drives and cartridges in connected silos

Fig. 2 ACLS5 Read and Write mounts per week. (Weeks 0425 to 0501)



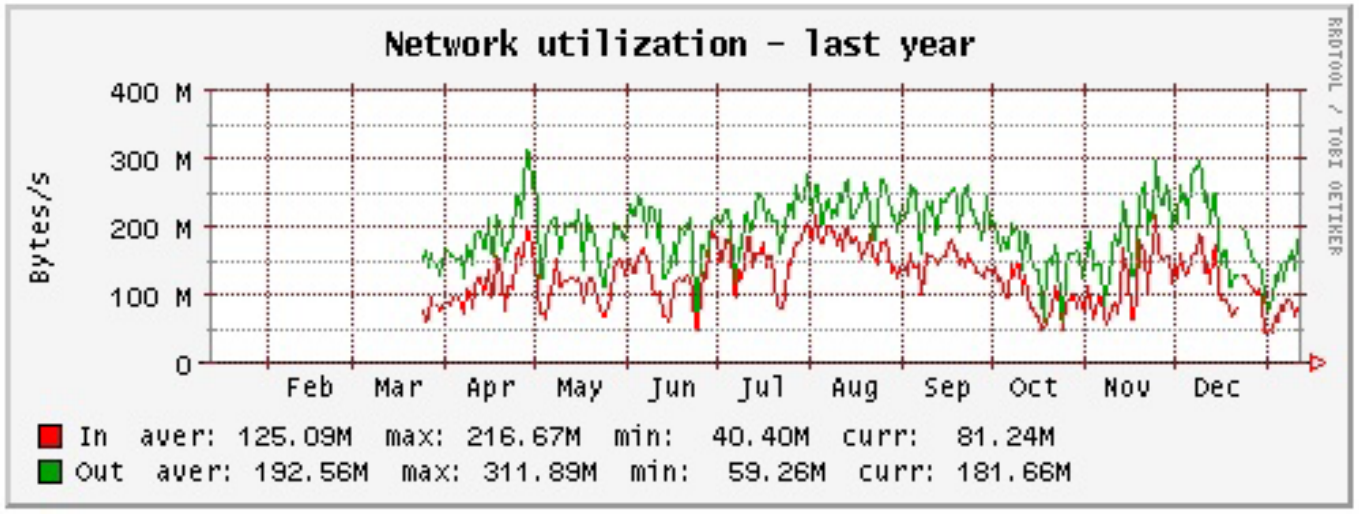
Tape System performance (II)

Fig. 5 Total read/write data volume (TB) per week for both libraries. (Weeks 0425 to 0501)



relating this to the #mounts plots
--> writing tapes is much more efficient than reading them
factor 10 less mounts for a writing a TB relative to reading a TB

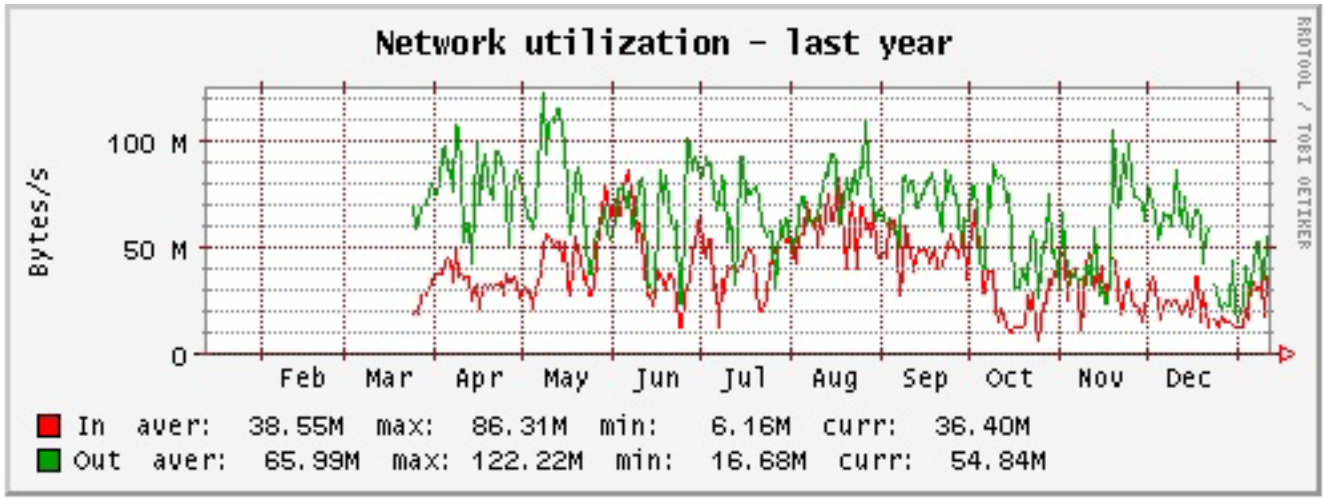
Storage performance



short time period peaks are averaged-out here. there were half day peaks of twice the shown speeds

some disk pools used at 80% and some at 5%

aggregate (~300 disk servers) read and write speed over the last 10 month



the existing nodes are not used to their full capacity

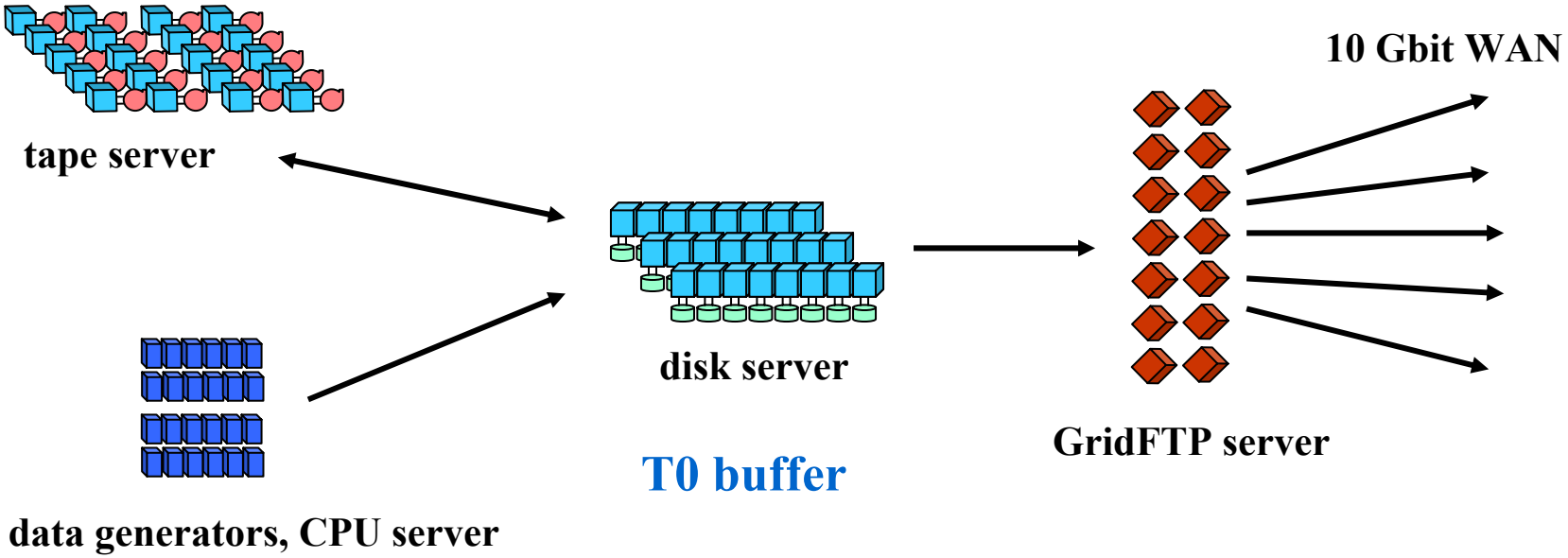
inefficiency problems
--> application features
access patterns

aggregate (~50 tape server) read and write speed over the last 10 month

2005 main activities

- **Run the service data challenges : 300 MB/s and 500 MB/s WAN ←→ Mass Storage disk only and tape storage included**
- **Run the local ALICE-IT data challenge, 750 MB/s disk+tape**
- **Handle the experiment production load continuously over the year → impact of service data challenges, no extra tape resources in 2005**
- **Understand much better the limitations and boundary conditions for tape-disk storage systems**
- **Prepare the purchase of a new tape storage system in 2006**

Resources for the 500 MB/s DC at CERN



To run with 500 MB/s one needs about :

- **25 tape server (one tape drive each) → 60% efficiency (= large files)**
- **25 disk server with each 2 TB space → 24h buffer)**
- **15 gridFTP server → 50% efficiency**
- **20 data generator nodes**

the system needs to be part of the standard activities, so that the 'extra' man-power needed is acceptable.

Setup

parallel setup of tests and production :

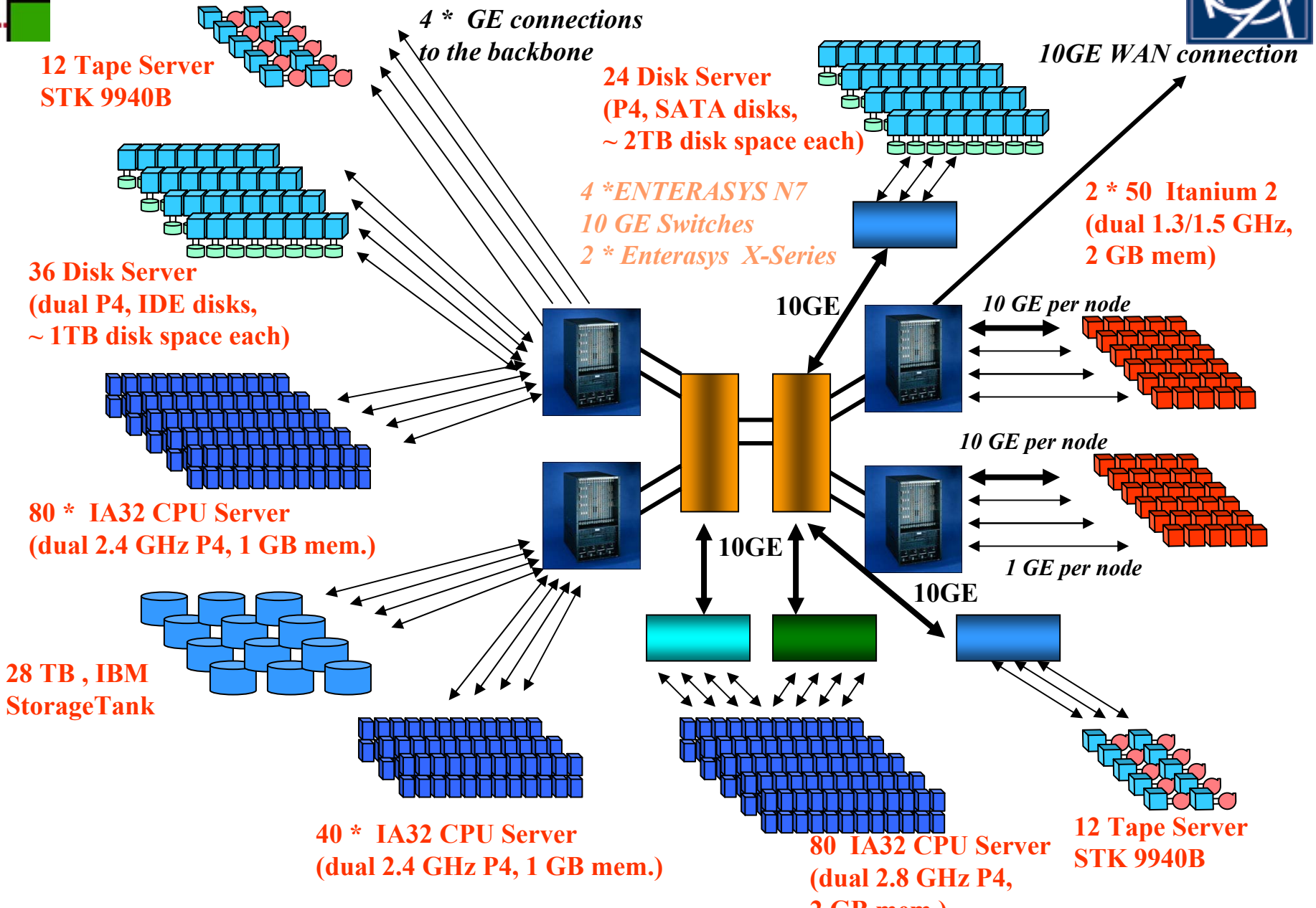
1. **high throughput cluster for data challenges with 10 GBit connections our new backbone construction will be 'folded' in during Q3 2005**

2. **the current production system (GridFTP service) will continue in our production backbone (up to ~ 2 Gbit)**

3. **when the new backbone is stable, point 2 with enhanced network capacity will move over**

production and test setups will essentially exist 'forever' in parallel

High Througput Prototype (openlab + LCG prototype)



2005 main activities

- Run the service data challenges : 300 MB/s and 500 MB/s
  WAN ←→ Mass Storage disk only and tape storage included
- Run the local ALICE-IT data challenge, 750 MB/s disk+tape
- Handle the experiment production load continuously over the year
  → impact of service data challenges, no extra tape resources in 2005
- Understand much better the limitations and boundary conditions for tape-disk storage systems
- Prepare the purchase of a new tape storage system in 2006

Mass Storage Performance

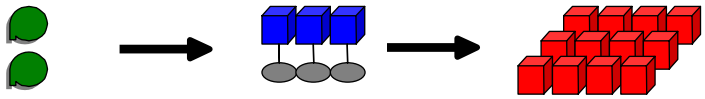
First set of parameters defining the access performance of an application to the mass Storage system

speed of the robot
distribution of tapes in silos
(at the time of writing the data)

number of running batch jobs
internal organization of jobs (exp.)
(e.g. just request file before usage)
priority policies (between exp. and within exp.)
CASTOR scheduling implementation

CASTOR
database performance

tape drive speed
tape drive efficiency



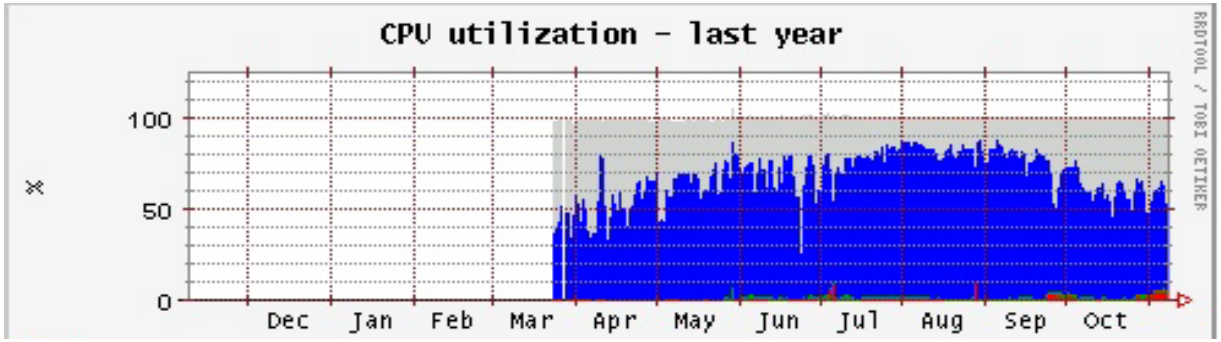
disk server
filesystem
OS + driver

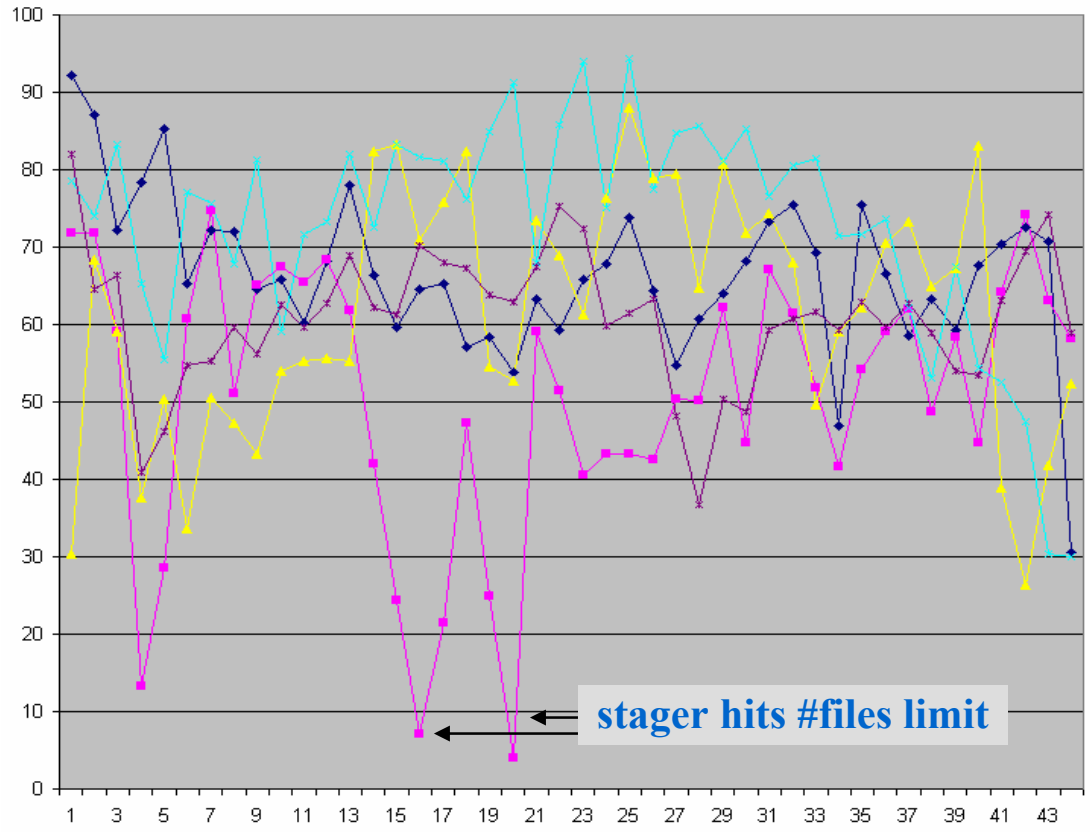
CASTOR load balancing mechanism
monitoring
Fault Tolerance
disk server optimization

data layout on disk
exp. policy
access patterns (exp.)
performance overall
file size

bugs and features

Tape efficiency effects





Analyzing the Lxbatch inefficiency trends
→wait time due to tape queues

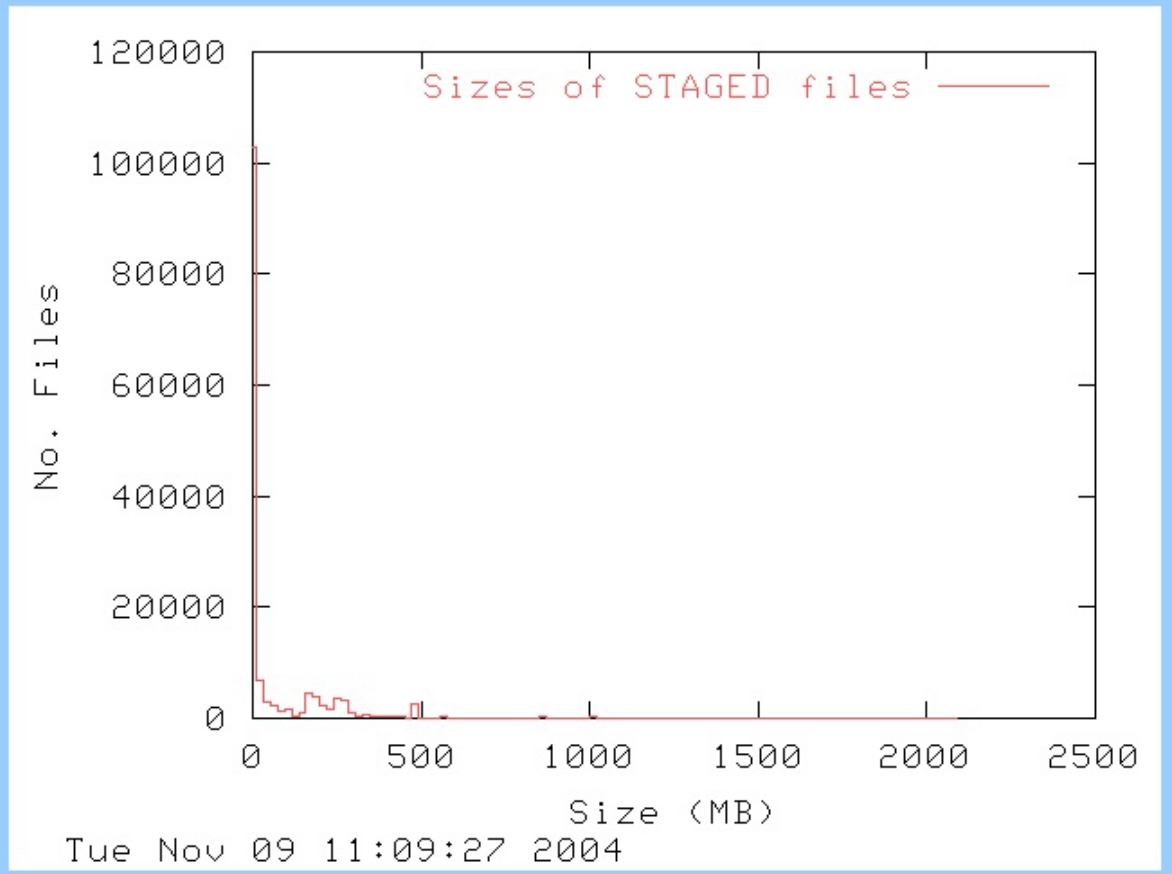
Example : File sizes

Average file size on disk

ATLAS	43 MB
ALICE	27 MB
CMS	67 MB
LHCb	130 MB
COMPASS	496 MB
NA48	93 MB

large amounts < 10MB

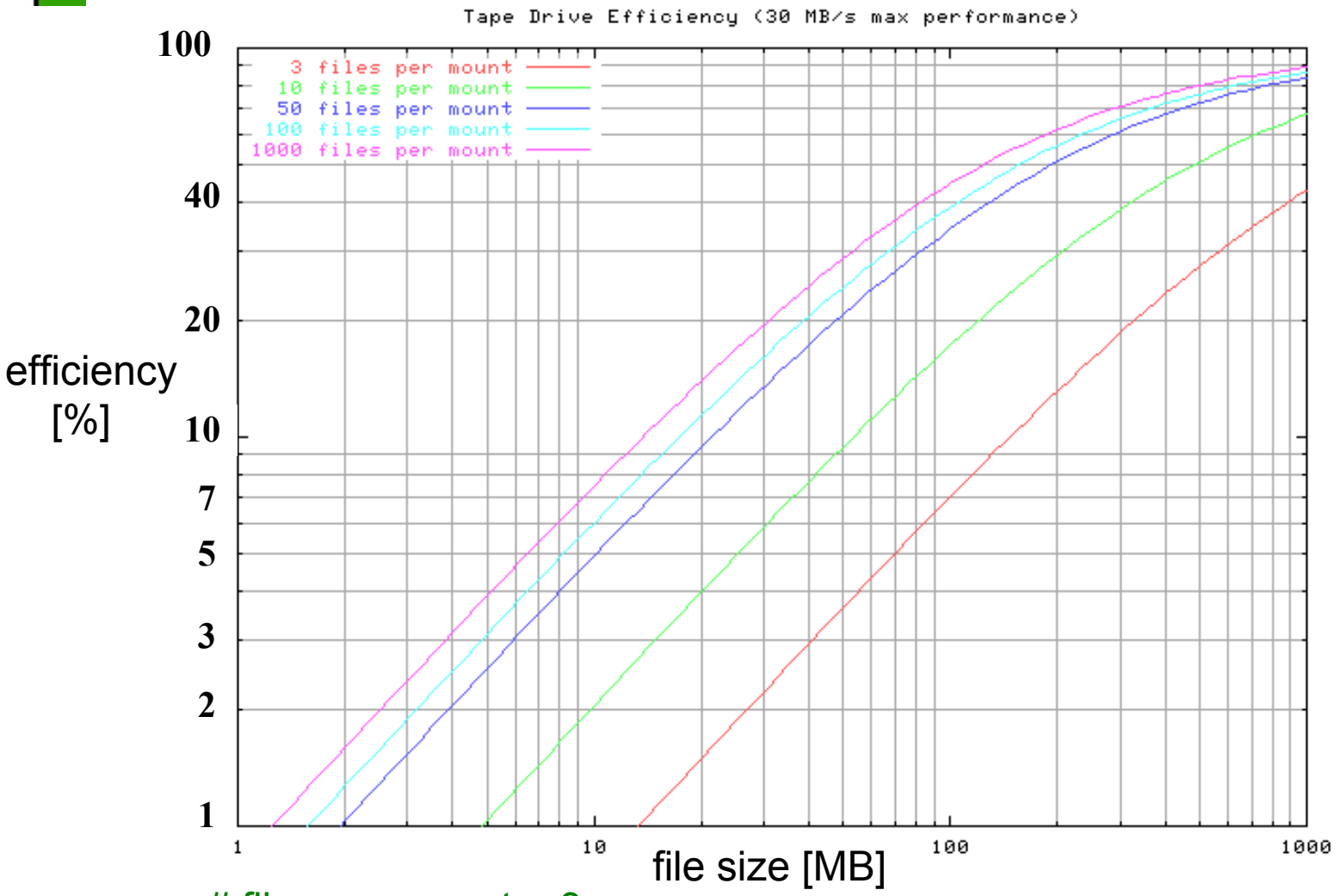
Fig 4. Size distribution of files currently STAGED



Total contents of above histogram is : 146847.00

File sizes: minimum: 0.0, maximum: 2000.0 and average: 71.3 (MB)

Analytical calculation of tape drive efficiencies



average # files per mount ~ 3
large # of batch jobs requesting files, one-by-one

tape mount time ~ 120 s
file overhead ~ 4.4 s

Efficiency improvements

combination of problems **example : small files + randomness of access**

possible solutions :

- **concatenation of files on application or MSS level**
- **extra layer of disk cache, Vendor or 'home-made'**
- **hierarchy of fast and slow access tape drives**
- **very large amounts of disk space**
- **…….**

Currently quite some effort is put into the analysis of all the available monitoring information to understand much better the influence of the different parameters on the overall performance.

the goal is to be able to calculate the cost of data transfers from tape to the application

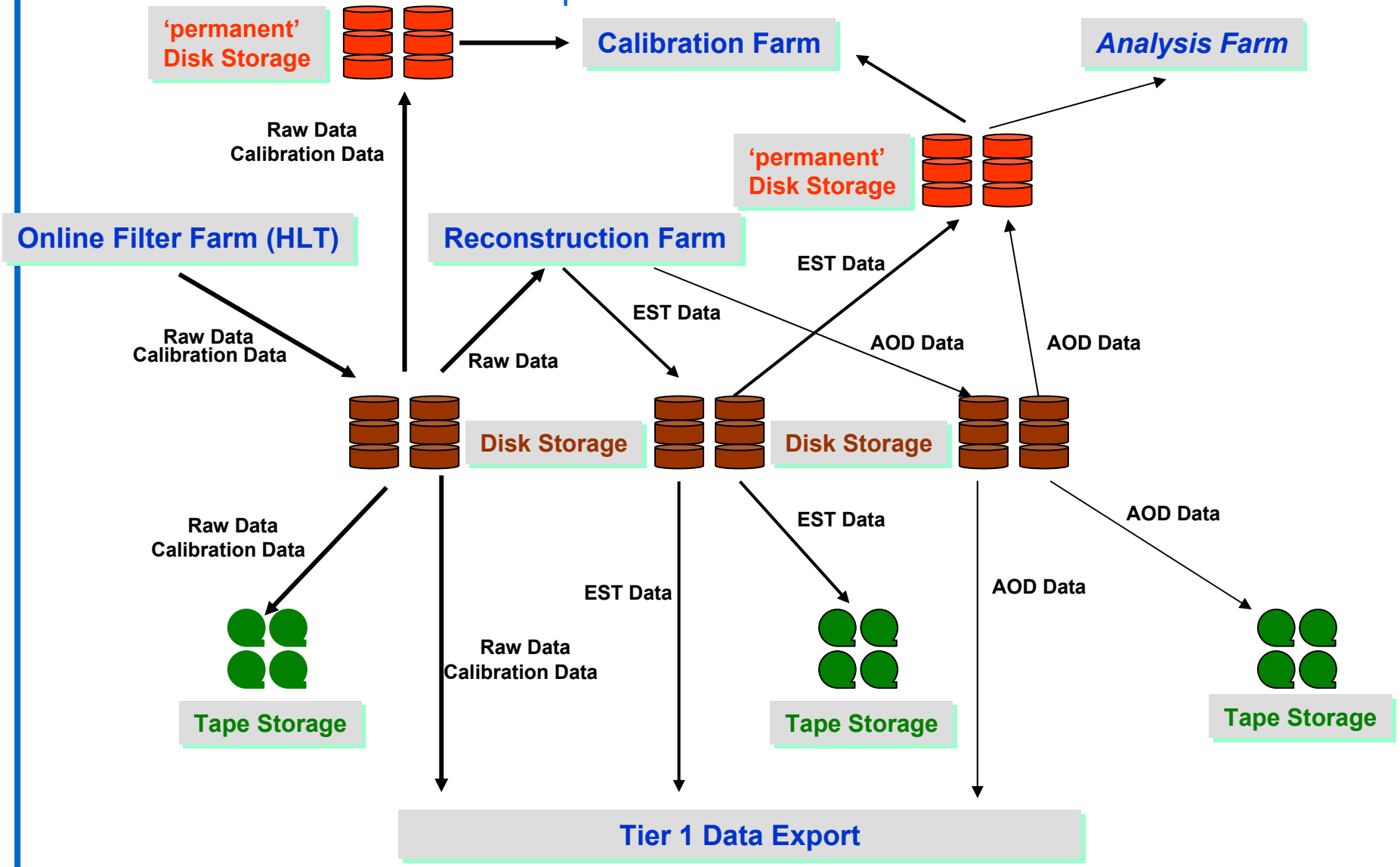
→ CHF per MB/s for volume of X TB

2005 main activities

- Run the service data challenges : 300 MB/s and 500 MB/s
  WAN ←→ Mass Storage disk only and tape storage included
- Run the local ALICE-IT data challenge, 750 MB/s disk+tape
- Handle the experiment production load continuously over the year
  → impact of service data challenges, no extra tape resources in 2005
- Understand much better the limitations and boundary conditions for tape-disk storage systems
- **Prepare the purchase of a new tape storage system in 2006**

Dataflow local CERN Fabric 2007

Complex organization with high data rates (~10 GBytes/s) and ~100k streams in parallel



Tape storage for 2006- (I)

Concentrate on linear tape technology , not helical scan (Exabyte, AIT..)

Today's choices :

IBM
3592 drives → 40 MB/s and 300 GB cartridges
25 KCHF and 0.7 CHF/GB

StorageTek
9940B drives → 30 MB/s and 200 GB cartridges
35 KCHF and 0.6 CHF/GB

LTO consortium (HP, IBM, Certance)
LTO-2 drives → 20 MB/s and 200 GB cartridges
15 KCHF and 0.4 CHF/GB (decreased by factor 2 during last 18 month)

LTO-3 drives are available since about 3 weeks,
40-60 MB/s and 400 GB cartridges → 0.4 CHF/GB

the standalone drive costs about 6 KCHF, modified robotic drives are available
3-6 month later and the modifications (fibre channel, extra mechanics,etc.)
adds another ~ 10 KCHF to the cost

the error on the drive costs is up to 50%, media costs varies by 10%

Tape storage for 2006- (II)

STK : 9940B, 200 GB cassettes , 30 MB/s speed, today
STK : STKA, 500 GB cassettes , 120 MB/s speed, mid 2005
STK : STKB, 1000 GB cassettes , 240 MB/s speed, beg. 2008 ?

IBM : 3592A, 300 GB cassettes , 40 MB/s speed, today
IBM : 3592B, 600 GB cassettes , 80 MB/s speed, beg. 2006

LTO : LTO2, 200 GB cassettes , 20 MB/s speed, today
LTO : LTO3, 400 GB cassettes , 60 MB/s speed, mid 2005
LTO : LTO4, 800 GB cassettes , 120 MB/s speed, beg. 2008

Boundaries :

- would like to 'see' the drive for about one year in the market
- 5 years max lifetime for a drive technology
- one year overlap of old and new tape drive installations
- have a new service ready for 2007

→ not easy to achieve

Tape storage for 2006- (III)

There are very little choices in the large robotic tape storage area.

6500 cartridge silo including robotics ~ 1 MCHF +- 30 %

(the best , flexible in the market is currently probably the STK 8500 tape library)

an old STK powderhorn silo (5500 slots) costs about 200 KCHF, but does not support LTO or the new IBM drives

to be considered :

single large installation or distributed, separate installations
locality of data, load balancing for reading is defined at writing time
regular physical movement of tapes is not really an option
the pass-through mechanism between silos has still 'locality' restrictions
→ one move of a tape = one mount of a tape

prices of robots and drives have large error margins (50%), because these are non-commodity products and depend heavily on the negotiations with the vendors (level of discount)

more details here:
http://lcg-computing-fabric.web.cern.ch/LCG-Computing-Fabric/presentations/tape_storage_issues_phase2_01.dec.2004.ppt

Summary

Lots of challenges……..

Need to do plenty of things in parallel which are correlated (new backbone, new tape storage, understand and test MSS tape←→ disk relation, data challenges, production….).

Service data challenges must be part of the 'standard' IT production schemes as fast as possible to reduce the need for larger extra man-power (they need of course extra resources, as these are new services…).

The data challenges will in 2005 interfere heavily with productions, as large amounts of tape resources are needed over longer periods.

Understanding the tape-to/from-application dataflow is key for the costing and sizing of the systems in 2007, need to be 'fixed' in 2005
→ TDR, experiment computing models