# Grid Enabled Mass Storage System (GEMSS): The Storage and Data Management System used at the INFN Tier1 at CNAF

Pier Paolo Ricci, on behalf of INFN CNAF Tier1 Storage
pierpaolo.ricci@cnaf.infn.it

CHEP 2012

New York University

# Summary

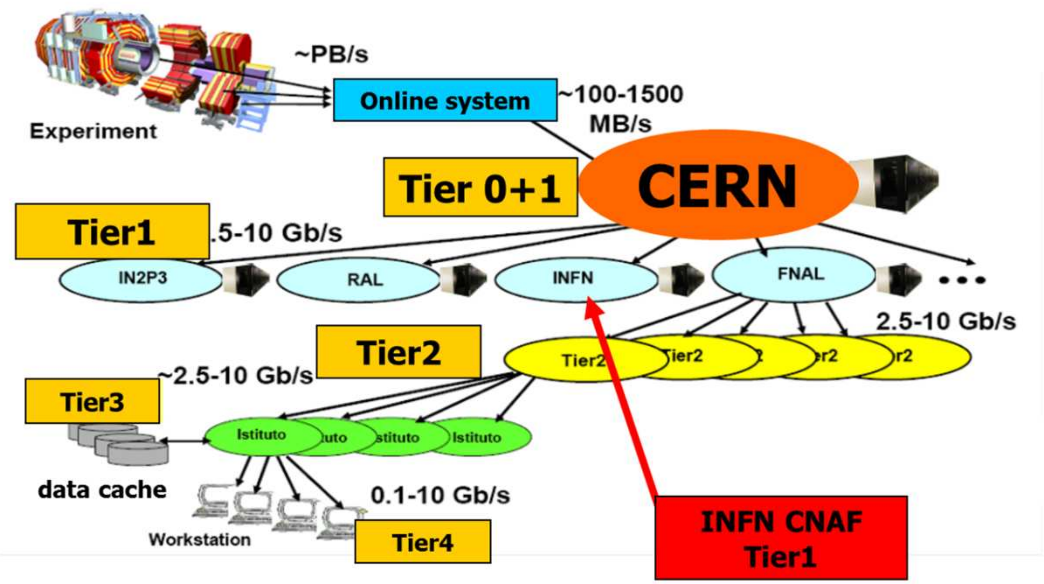
- Description of INFN CNAF Tier1 resources
- The GEMSS system: development history and description
- GEMSS software overall details
- GEMSS Latest improvements: optimization administration tools and monitoring
- Experiments activity of the last years and analysis of relevant user cases (ATLAS, CMS and LHCB)

# INFN CNAF and LHC computing

- The INFN CNAF hosts the INFN (Italian Nuclear Physics Institute) main computing centre, a WLCG Tier1
- Supporting High Energy Physics experiments of LHC at CERN: ALICE, ATLAS, CMS, LHCb

Also resources for other Physics experiments:

BABAR (SLAC), CDF (FNAL), VIRGO (Cascina), ARGO (Tibet), AMS (Satellite), GLAST/FERMI (Satellite), PAMELA (Satellite), MAGIC (Canary Islands telescope)…



# Tier 1 Storage resources

**8.4 PB of used on-line (net) disk space (GEMSS)**

- 7 **DDN** S2A 9950 => 7 PB
- 7 **EMC²** CX3-80 + 1 **EMC²** CX4-960 => 1.4 PB

… and under installation

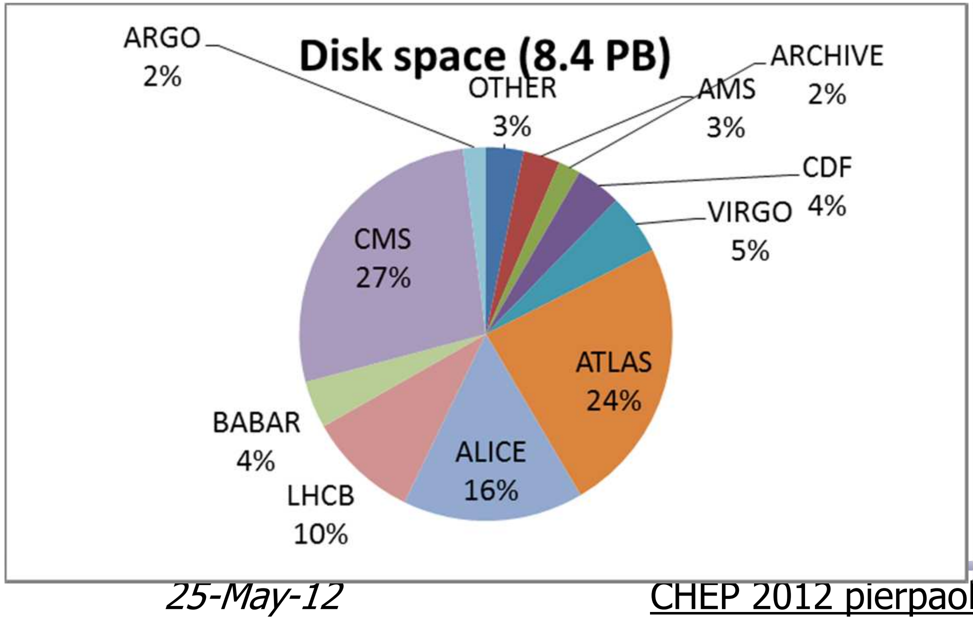
3 Fujitsu Eternus DX400 (3 TB SATA) : + **2.8 PB**

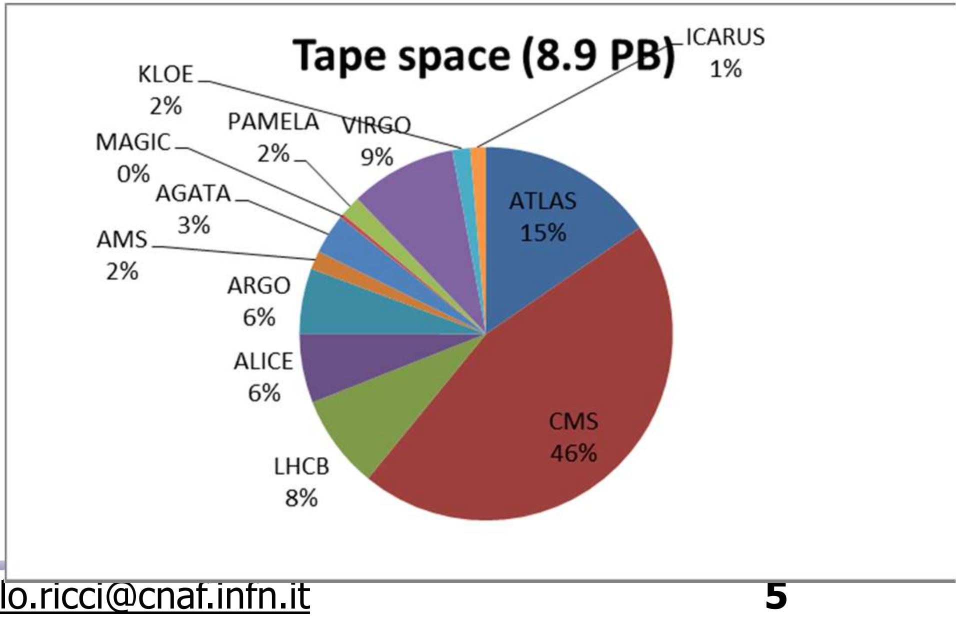
**(TOTAL ~11.2 PB)**

- Access using diskservers over Fibre Channel (SAN)
  - ~40 disk servers (10 Gb/s ethernet) on DDN
  - ~90 disk servers (1 Gb/s ethernet) on EMC²
- Tape library **Sl8500 9 PB + 5 PB (just installed)** on line with **20 T10KB** drives and **10 T10KC** drives
  - 9000 tapes x 1 TB tape capacity, ~ 100 MB/s of bandwidth for each drive
  - 1000 tapes x 5 TB tape capacity, ~ 200 MB/s of bandwidth for each drive
  - Drives interconnected to library and servers via dedicated SAN (TAN).
  - 13 Tivoli Storage manager HSM nodes access to the shared drives.
  - 1 Tivoli Storage Manager (TSM) server common to all GEMSS instances.
- All storage systems and disk-servers are on SAN (4 Gb/s or 8 Gb/s)

# WHAT IS GEMSS?

- A full HSM (Hierarchical Storage Management) integration of GPFS, TSM and StoRM (StoRM is the SRM interface, *see http://storm.forge.cnaf.infn.it/*)
- Minimize management effort and increase reliability:
  - Very positive experience for scalability so far;
  - Large GPFS installation in production at CNAF since 2005 with increasing disk space and number of users;
- The whole disk space partitioned in several GPFS clusters served by ~130 diskservers + ~9 PB of used tape space.
  - The full system is easily managed (only 2 FTEs);
  - All experiments at CNAF (LHC and non-LHC) agreed to use GEMSS



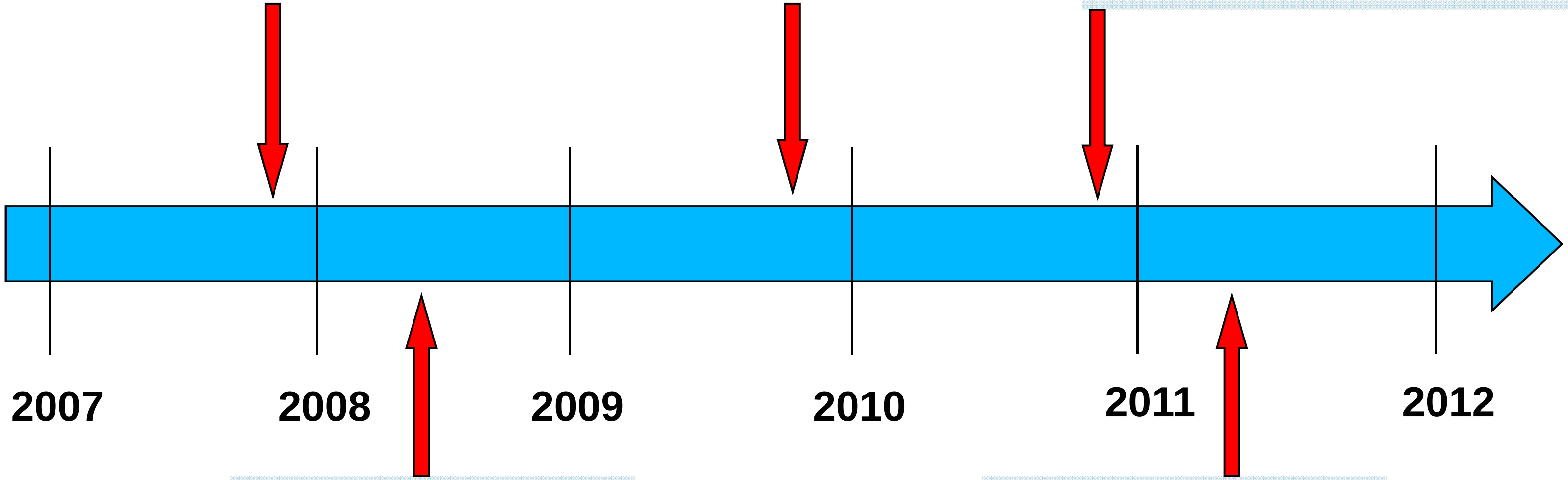


# GEMSS Development TimeLine

D1T0 Storage Class implemented with StoRM/GPFS **for LHCb and ATLAS**

D0T1 Storage Class implemented with StoRM/GPFS/TSM **for CMS**

**ATLAS, ALICE, (CMS) and LHCb experiments, together with all other non-LHC experiments (Argo, Pamela, Virgo, AMS) use GEMSS in production!**

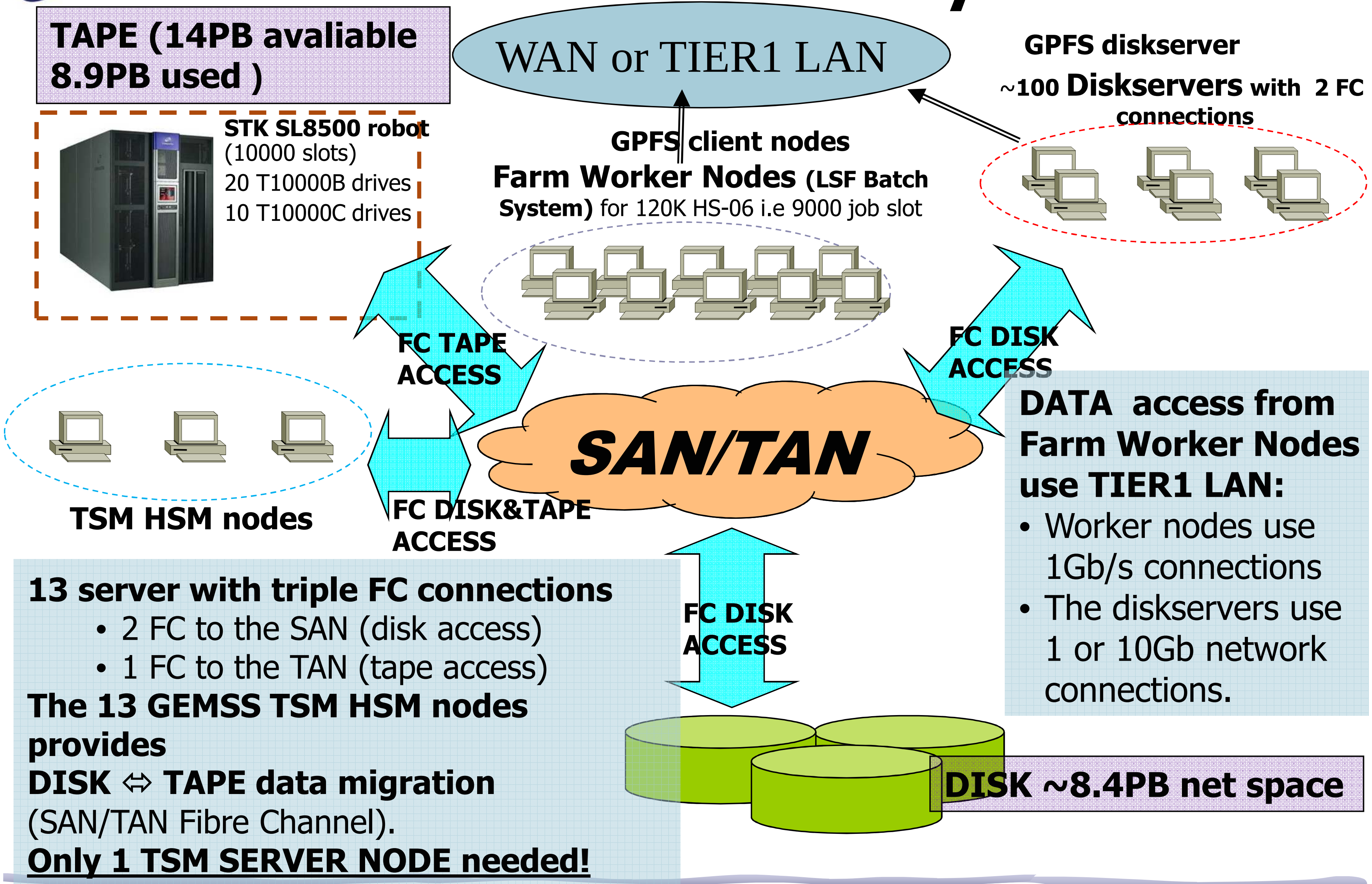


D1T1 Storage Class implemented with StoRM/GPFS/TSM **for LHCb**

Introduced DMAPI server (to support GPFS 3.3/3.4

**GEMSS is used by all LHC and non-LHC experiments in production for all Storage Classes**

# GEMSS resources layout



**TAPE (14PB avaliable 8.9PB used )**

**STK SL8500 robot**
(10000 slots)
20 T10000B drives
10 T10000C drives

WAN or TIER1 LAN

**GPFS diskserver**

~100 **Diskservers with 2 FC connections**

**GPFS client nodes**

**Farm Worker Nodes (LSF Batch System)** for 120K HS-06 i.e 9000 job slot

**FC TAPE ACCESS**

**FC DISK ACCESS**

**SAN/TAN**

**TSM HSM nodes**

**FC DISK&TAPE ACCESS**

**FC DISK ACCESS**

**DATA access from Farm Worker Nodes use TIER1 LAN:**

- Worker nodes use 1Gb/s connections
- The diskservers use 1 or 10Gb network connections.

**13 server with triple FC connections**

- 2 FC to the SAN (disk access)
- 1 FC to the TAN (tape access)

**The 13 GEMSS TSM HSM nodes provides**
**DISK ⇔ TAPE data migration**
(SAN/TAN Fibre Channel).

**Only 1 TSM SERVER NODE needed!**

**DISK ~8.4PB net space**

# GEMSS implementation

- **GPFS** clusters
  - provide fast and reliable filesystems with direct access (posix file protocol) from the Worker Nodes Farm => all the WNs access to the GPFS filesystem as local!
  - use Block level I/O interface over network
  - use parallel I/O over all diskservers for optimizing the performance
  - cluster means no-single-point-of-failure in the disk access layer
- **TSM**
  - migrates data to tape and provide further access to them
  - uses SAN/TAN for data transfers instead of LAN
- **STORM**
  - provides data access using the LCG grid tools.
  - is compliant with the standard SRM interface version 2.2

**GEMSS** is a software layer for GPFS-TSM interaction for optimization and administration, providing a complete solution for storage access

# GPFS to TSM data flow

- GPFS Information Lifecycle Management (ILM) implements data movements between storage pools. GPFS uses an SQL-like policy language to define data migrations rules.
- GPFS “external storage pool” extends the use of policy driven migration and recall system to the tape storage (HSM tape extension of GPFS)
- External pool “rule” defines script for migrating/recalling files.
- GPFS policy engine automatically builds candidate lists and passes them to external pool scripts.

To generate a list of file candidates for migration, ILM scans the file-system building a result set of file attributes and pathnames that matches the search criteria specified

```
RULE 'premigrate from data1 to tape cms preprod'
MIGRATE FROM POOL 'data1' THRESHOLD(0,100,0)
WEIGHT(CURRENT_TIMESTAMP-ACCESS_TIME)
TO POOL 'TAPE PREMIGRATION CMS_PREPROD'
FOR FILESET('CmsData','CmsMc') WHERE
PATH_NAME LIKE '%/%x_preproductionx_%/%'
ESCAPE 'x'
```

**EXAMPLE: All files with *“preproduction”* keyword in PATH NAME goes to the specific tape pool “CMS PREPROD” i.e. a specific set of tape cartridge**

# GPFS to TSM data flow (2)

The data flow system from GPFS to TSM use standard the GPFS features

- **"pre-migration"** stands for the action of copying a file to tape, but keeping the original copy also on disk. Done "as soon as possible" and the number of disk-tape streams is configurable for each filesystem
- **"migration"** is the action of copying the file to tape and removing the content of the file from disk and keep a so-called "stub" file as a normal file. This usually occurs at a specific threshold (garbage collector)

- DMAPI (Data Management API) are used in GPFS/TSM and extended attributes are added during the migration phase to the stub file as a identification key for the file.

Each LHC VO has a dedicated number of redundant TSM HSM nodes for migration (and recall) shared over the GPFS filesystem. The number of disk-tape stream threads is configurable for each node.

- e.g. a VO with 2 HSM nodes dedicated and a number of 3 thread can use at maximum a number of 6 drive /stream.

# TSM to GPFS data flow

Accessing data that are only premigrated is immediate! (data already on disk)

… and for recalling data that are already migrated?

2 distinct recall methods

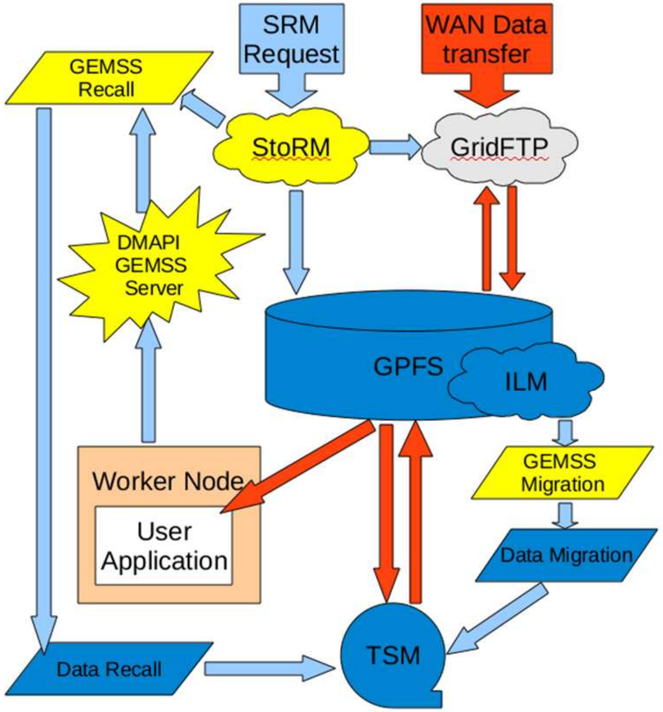
- **Selective recalls**. The user asks for a file to be recalled from tape prior to the first access using SRM commands (StoRM). When the file has been recalled the access or transfer (using gridftp for WAN access) is performed
- **Transparent recalls.** The file is accessed by a read operation (usually from user jobs) without a distinction between premigrated (still on disk) or migrated (only the stub file is on disk). In the last case the read operation triggers via DMAPI the recall of the file from tape. When the recall is over the job continues

**The recall operation is a tricky issue**:

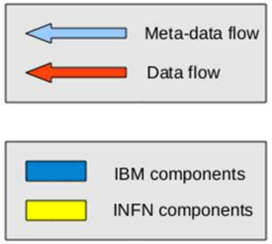
1. Recall requests should be collected in a specific time lapse
2. Requests should be ordered (i.e. sorted) according to the files distribution on tape(s) to minimize number of mount/dismount operation in the tape library

# GEMSS schema

- New component in GEMSS: DMAPI Server
  - Used to intercept READ events via GPFS DMAPI and re-order recalls according to the files position on tape;
  - “Preload library” is not needed anymore (it was used in order to transform on the client side a transparent recall into a selective recall in prev. GEMSS version)
  - Available with GPFS v.3.x



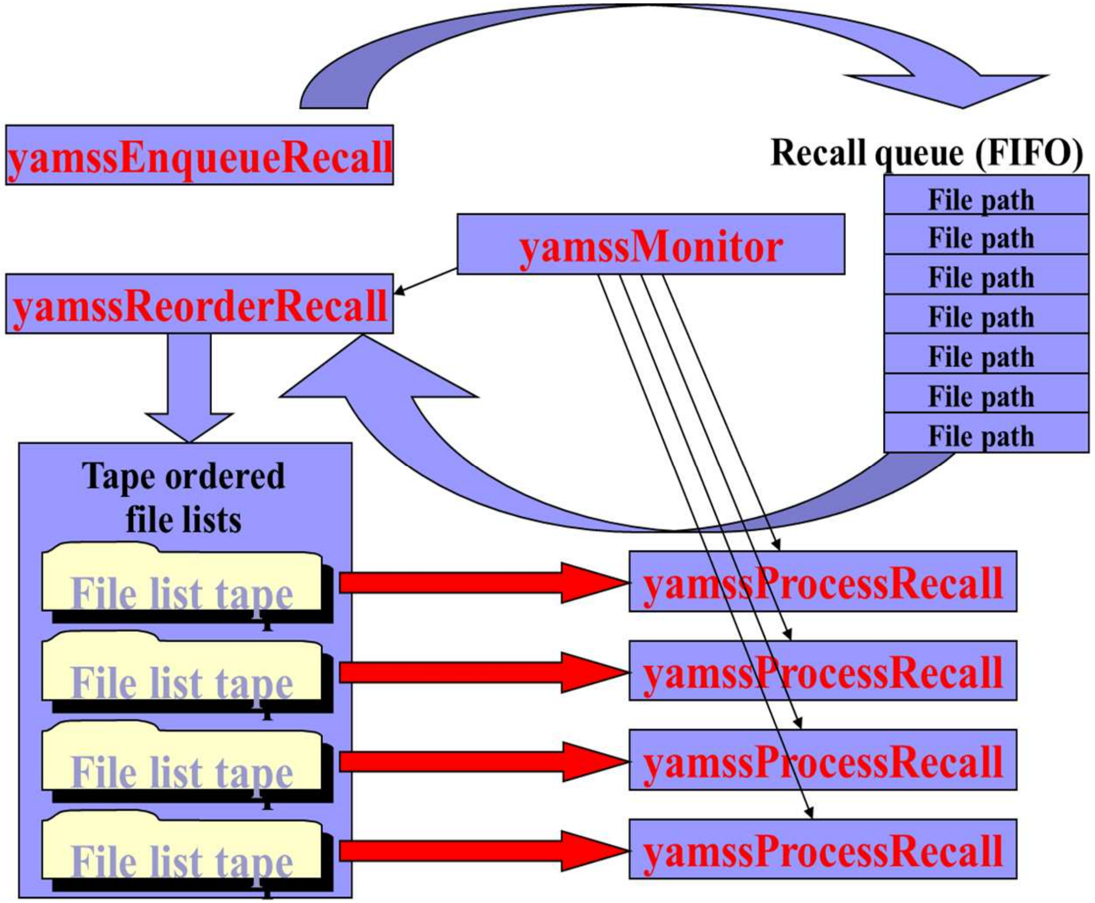
- **GPFS**: disk-storage software infrastructure
- **TSM**: tape management system
- **StoRM**: SRM service
- **TSM-GPFS** interface
- **Globus GridFTP**: WAN data transfers



# GEMSS recall system

- Selective recall system in GEMSS use 4 processes:
  yamssEnqueueRecall
  yamssMonitor,
  yamssReorderRecall
  yamssProcessRecall

- yamssEnqueueRecall & yamssrReorderRecall manage a FIFO queue with the files to be recalled, fetches files from the queue and builds sorted lists with optimal file ordering.



yamssProcessRecall actually creates the recall streams, perform the recalls and manages the error conditions (i.e. retries file recall failures…)

yamssMonitor is the supervisor of the reorder and recall phases

# GEMSS interface

- Set of administrative commands have been also developed, (for monitoring,  stopping and starting migrations and recalls, performance reporting).
- Almost 50 user interface commands/daemon

some examples…

- yamssEnqueueRecall (command)
  - Simple command line to enqueue into a FIFO the files to recall from tape
- yamssLogger (daemon)
  - Centralized logging facilty. 3 log files (for migrations, premigrations and recalls) are centralized for each YAMSS-managed file system
- yamssLs (command)
  - "ls"-like interface, but in addition prints status of each file: premigrated, migrated, disk-resident.

- RPM package for installation/distribution
- STAT files for collecting accurate statistic:

i.e. statistic file for recall:

Time stamps

filename

```
REC OK 1336415461 1336415466 1336415625 1296093154
/storage/gpfs_tsm_cms/cms/store/data/Commissioning10/ZeroBias/RECO/Dec22ReReco_from_V4_v1/0164/CE8F6F53-B529-
E011-ADBE-E0CB4E19F986.root 8328848523 8328848523 tsm-hsm-3.cr.cnaf.infn.it T01076
```

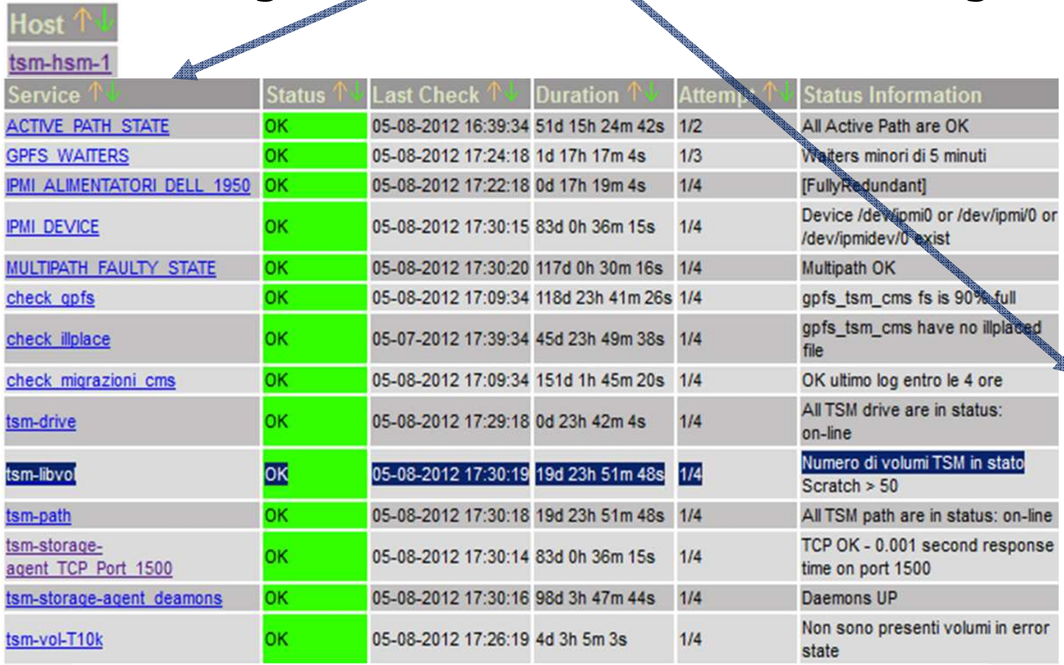
filesize

HSM node

Tape Label

# GEMSS monitoring

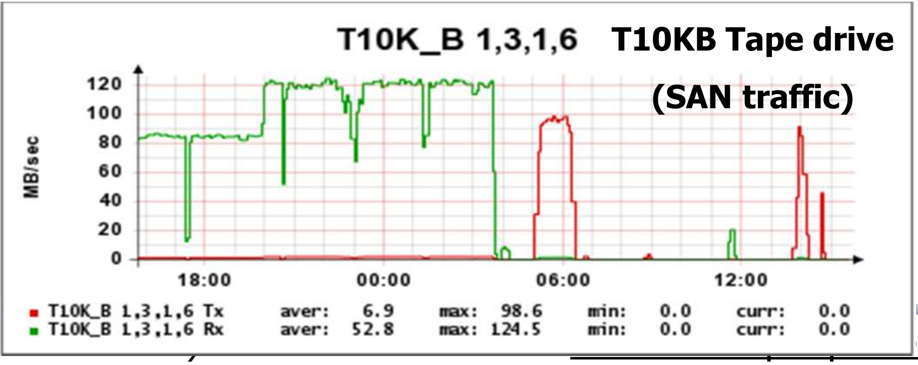
- Integration with NAGIOS for alert system, notification and automatic actions (i.e. restarting of failed TSM daemons).
- Integration with LEMON monitoring.

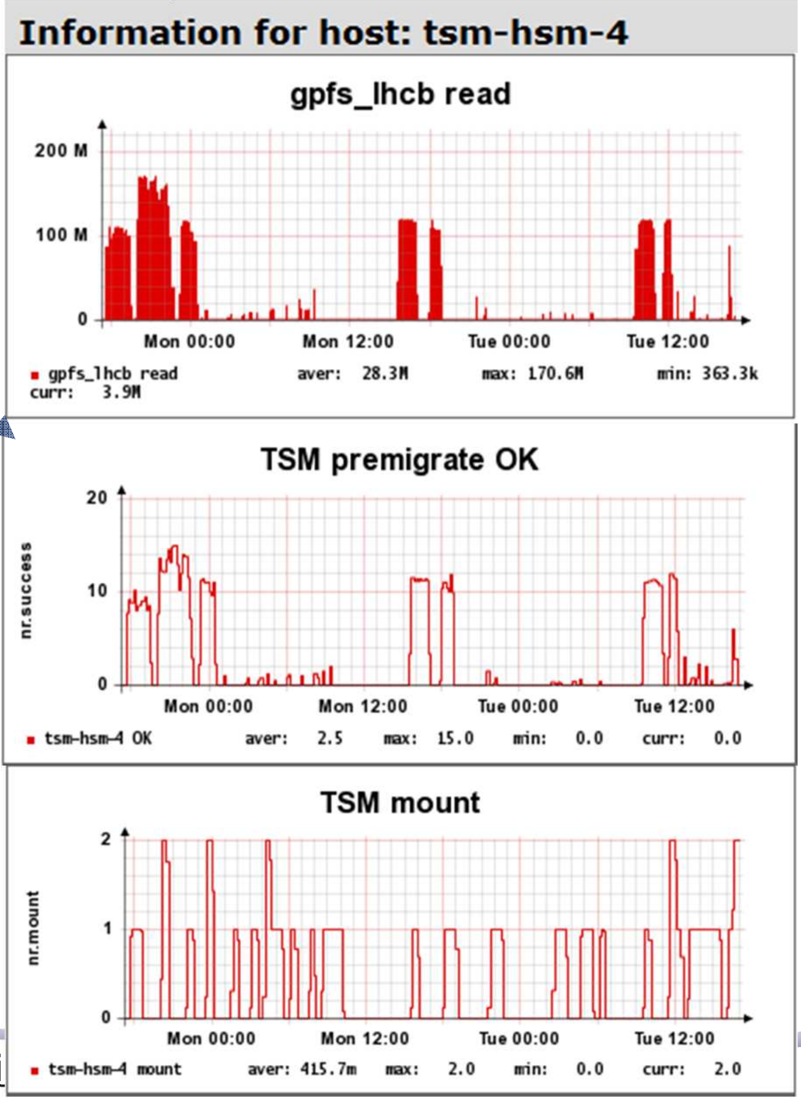


Host ↑↓

tsm-hsm-1

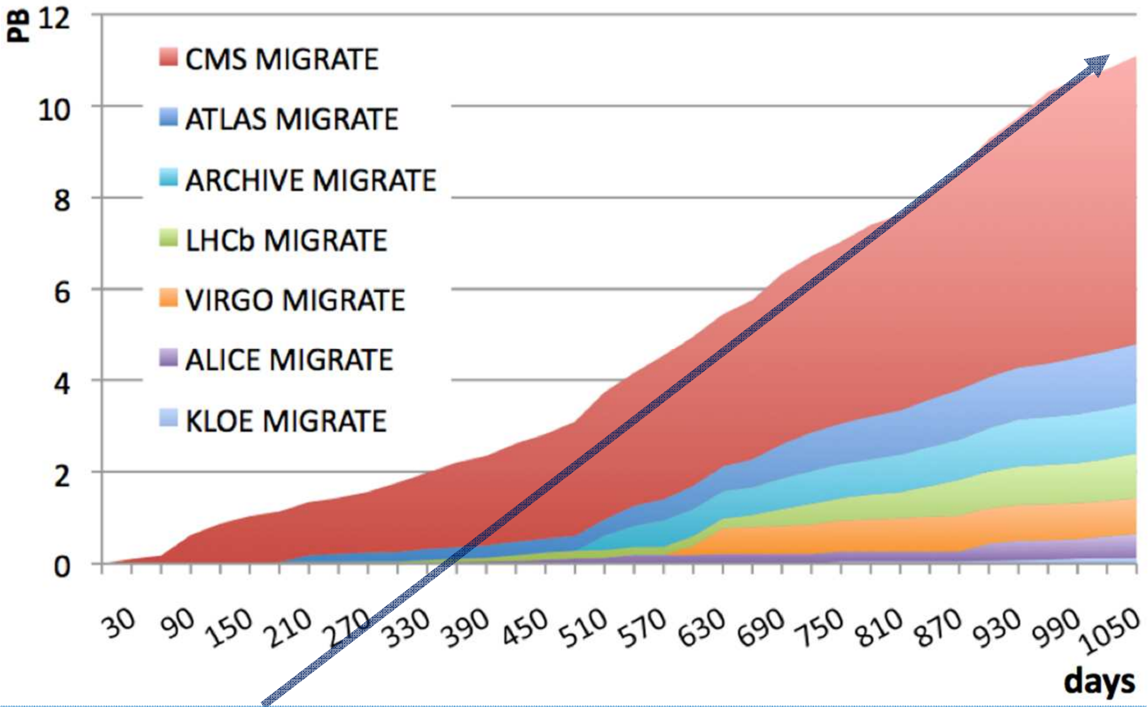
| Service ↑↓ | Status ↑↓ | Last Check ↑↓ | Duration ↑↓ | Attempt ↑↓ | Status Information |
|---|---|---|---|---|---|
| ACTIVE_PATH_STATE | OK | 05-08-2012 16:39:34 | 51d 15h 24m 42s | 1/2 | All Active Path are OK |
| GPFS_WAITERS | OK | 05-08-2012 17:24:18 | 1d 17h 17m 4s | 1/3 | Waiters minori di 5 minuti |
| IPMI_ALIMENTATORI_DELL_1950 | OK | 05-08-2012 17:22:18 | 0d 17h 19m 4s | 1/4 | [FullyRedundant] |
| IPMI_DEVICE | OK | 05-08-2012 17:30:15 | 83d 0h 36m 15s | 1/4 | Device /dev/ipmi0 or /dev/ipmi/0 or /dev/ipmidev/0 exist |
| MULTIPATH_FAULTY_STATE | OK | 05-08-2012 17:30:20 | 117d 0h 30m 16s | 1/4 | Multipath OK |
| check_gpfs | OK | 05-08-2012 17:09:34 | 118d 23h 41m 26s | 1/4 | gpfs_tsm_cms fs is 90% full |
| check_illplace | OK | 05-07-2012 17:39:34 | 45d 23h 49m 38s | 1/4 | gpfs_tsm_cms have no illplaced file |
| check_migrazioni_cms | OK | 05-08-2012 17:09:34 | 151d 1h 45m 20s | 1/4 | OK ultimo log entro le 4 ore |
| tsm-drive | OK | 05-08-2012 17:29:18 | 0d 23h 42m 4s | 1/4 | All TSM drive are in status: on-line |
| tsm-libvol | OK | 05-08-2012 17:30:19 | 19d 23h 51m 48s | 1/4 | Numero di volumi TSM in stato Scratch > 50 |
| tsm-path | OK | 05-08-2012 17:30:18 | 19d 23h 51m 48s | 1/4 | All TSM path are in status: on-line |
| tsm-storage-agent_TCP_Port_1500 | OK | 05-08-2012 17:30:14 | 83d 0h 36m 15s | 1/4 | TCP OK - 0.001 second response time on port 1500 |
| tsm-storage-agent_deamons | OK | 05-08-2012 17:30:16 | 98d 3h 47m 44s | 1/4 | Daemons UP |
| tsm-vol-T10k | OK | 05-08-2012 17:26:19 | 4d 3h 5m 3s | 1/4 | Non sono presenti volumi in error state |





# GEMSS and GPFS/StoRM activity

- Details of experiment activity of relevant user cases is reported in the next slides:
- ATLAS
- CMS
- LHCB



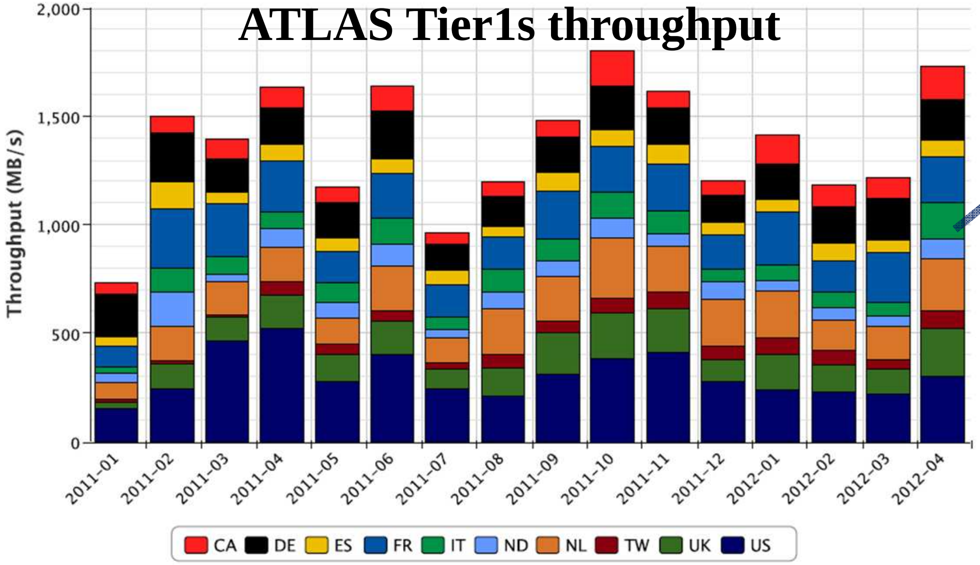
TOTAL of ~11 PB of data have migrated since the start of GEMSS official production

(some data was deleted by user=> now 8.9PB used)

# ATLAS activity GEMSS StoRM

- The INFN CNAF Tier1 is the only site with tape facilities (managed by GEMSS) for ATLAS (other Italian sites use Storm and GPFS i.e. Milano-Tier2, and other Tier3s)
- Tape at INFN-T1 are used to store RAW-data coming from Tier0 and simulation HITS Data.
- Data on stored on tape are accessed for reprocessing only
- In 2011-2012 the CNAF Tier1 storage system performed ATLAS data management activities according to the other ATLAS Tier1s average
- The INFN CNAF Tier1 ATLAS storage setup:
  - TOTAL DISK CAPACITY: 2.7 PB
  - TOTAL TAPE CAPACITY: 3.6 PB

# ATLAS storage data transfer

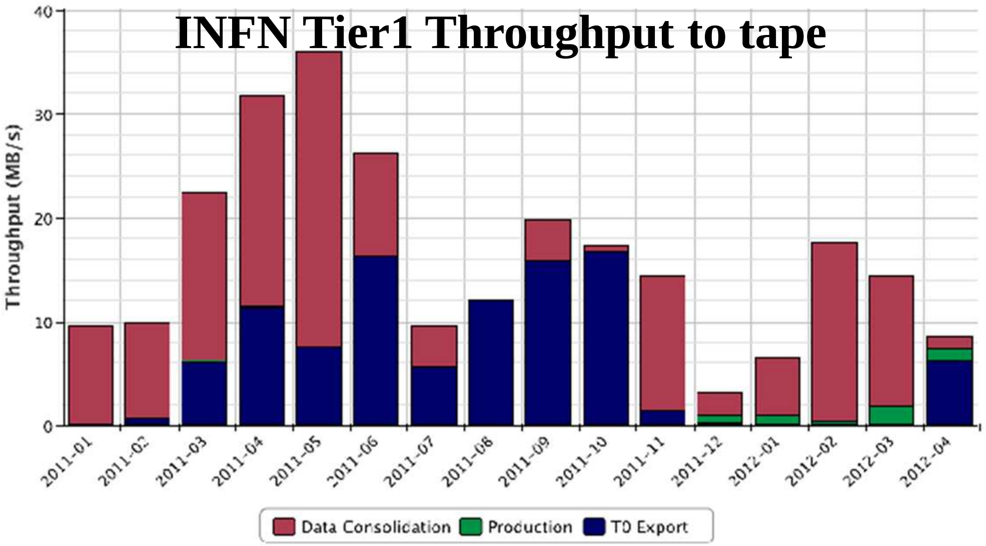

ATLAS Tier1s throughput

**INFN-T1 receives 10% of the ATLAS DATA**

Data transfer efficiency 94% (on the same level of the other Tier1s)

**INFN-T1 StoRM GPFS data transfer performances:**

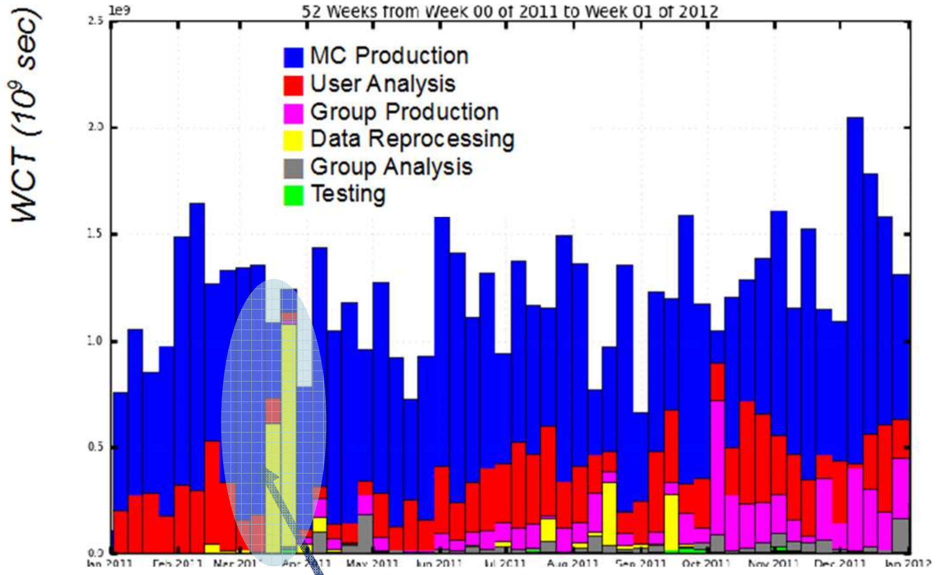
Peak 1800 MB/s (10 min. resolution)
Daily peak 800 MB/s
Average 84 MB/s (1GB/s worldwide)


INFN Tier1 Throughput to tape

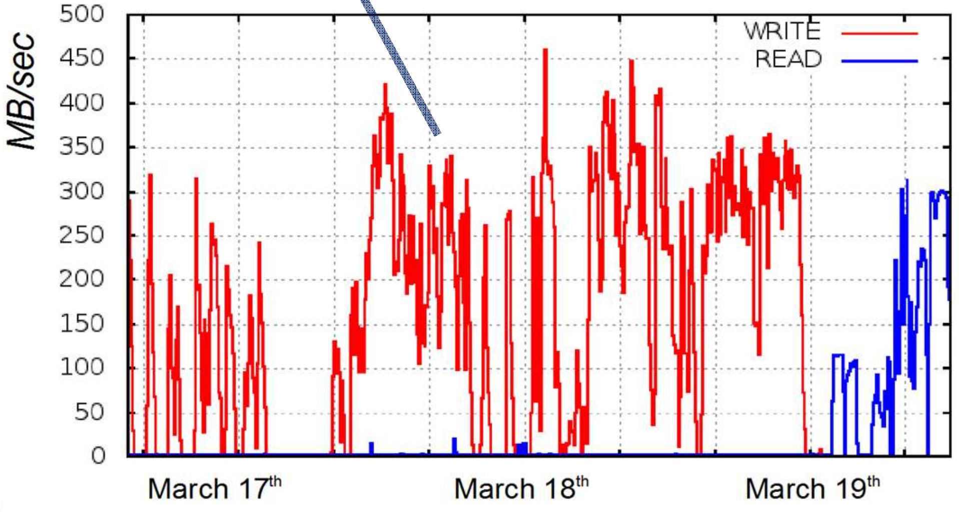
**INFN-T1 ATLAS incoming throughput written on tape**
Average 16 MB/s, Peak 500 MB/s
Data from Tier0
Data consolidation (preplaced primary data from reprocessing)

# ATLAS 2011 data reprocessing

2011 Data Reprocessing (yellow)



GPFS <=> TSM traffic



**4,20% of total processing activity at T1 (170 TB)**

Reprocessing is the only activity in ATLAS Computing involving high data recall from tape

The data is recalled on disk buffer before jobs execution => jobs access the data from wns use posix file protocol (GPFS local access)

High efficiency (99% successfull jobs)

Few days needed to complete reprocessing activity (on average with other Tier1s)

**GPFS <=> TSM traffic**

**write:** recalls for tape to disk for reprocessing

**read:** write to tape from TIER-0 (raw data flow)

Good performance for simultaneous read/write access

# CMS test activity

**GEMSS successfully in production since end of 2009 for CMS**

preparatory tests were successful, in terms of <u>rates and quality of transfers</u> using standard CMS workflows:

1. **Transfer tests with the PhEDEx 'LoadTest' infrastructure**

system could handle typical CMS rates (e.g. up to 300 MB/s and 500 MB/s in export to several dozens of sites, sustained for several hours)

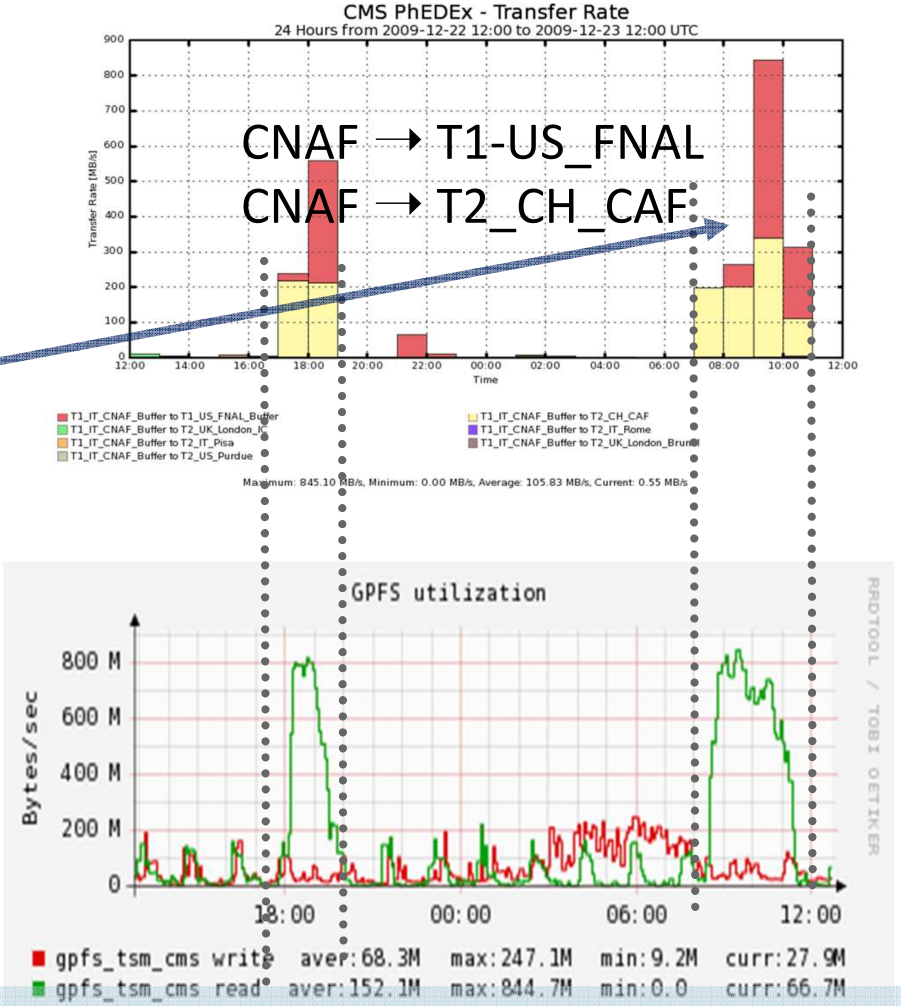
2. **Tests with Job Robot jobs**

The CMS workflows efficiency was not impacted by the change of storage system (CASTOR => GEMSS)

3. **Tests with real CMS jobs (CNAF farm)**

The disk storage could serve data at ~1.2 GB/s to the nodes with a ~100% success rate

**<u>As from this experience, CMS gave a very positive feedback on the new system, and agreed to migrate over to it</u>**



CNAF outbound traffic: a T1 (FNAL, US) and a T2 (CAF, CERN) tested simultaneously on the new system, and the corresponding GPFS utilization (Lemon monitoring)

# CMS transfers in production

## CNAF imports:

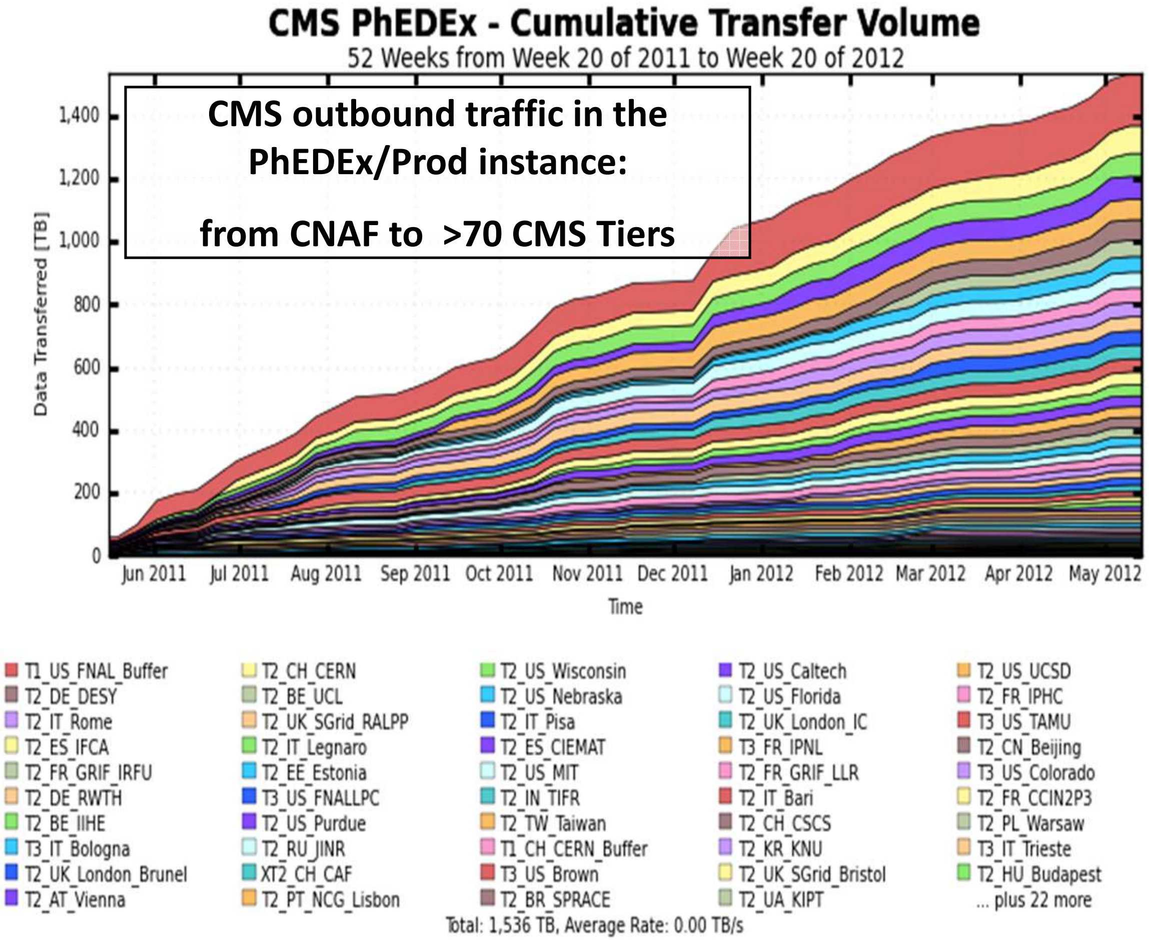
mostly from CERN and T1s, and from INFN T2 centres

- **of the order of ~50 TB/week on average (sustained since years)**

## CNAF exports:

to T1s, T2s and several T3s

- **~1.5 PB over last year**



CMS PhEDEx - Cumulative Transfer Volume

52 Weeks from Week 20 of 2011 to Week 20 of 2012

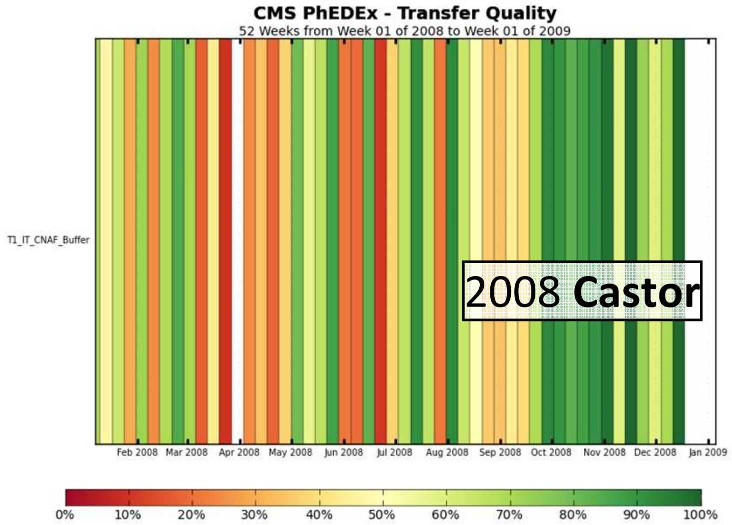
**CMS outbound traffic in the PhEDEx/Prod instance:**

**from CNAF to >70 CMS Tiers**

**With a massive transfer activity, the CNAF storage efficiently serves the needs of several dozens of CMS Tiers.**

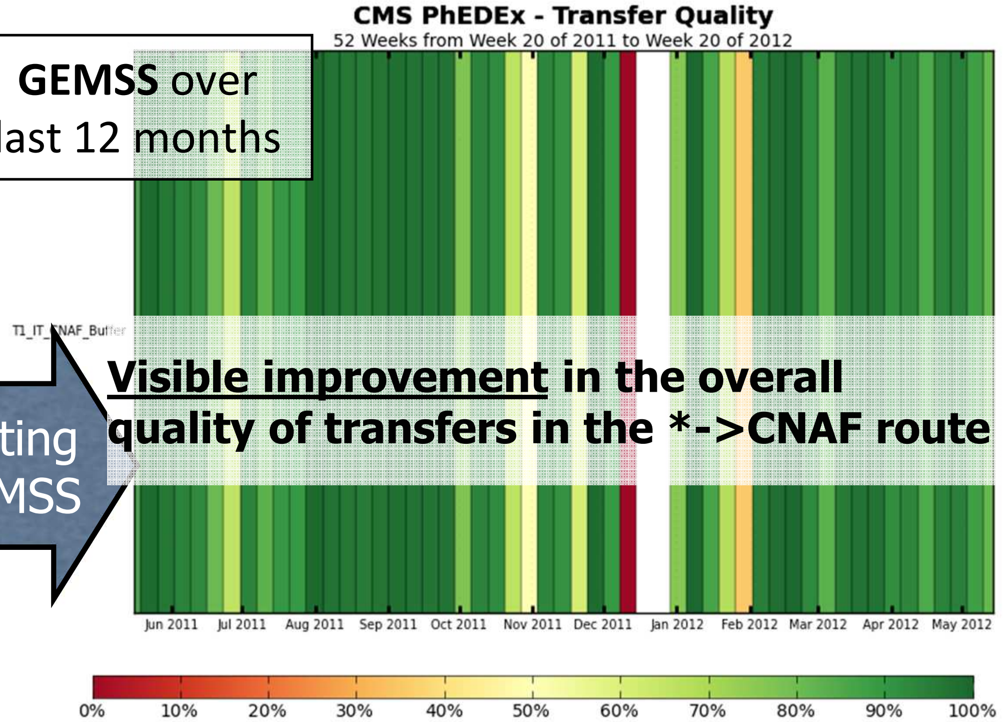
# CMS quality of transfers

The quality of transfers (successes over attempts) is an interesting observable to estimate "is a storage system good for Ops?"



2008 **Castor**

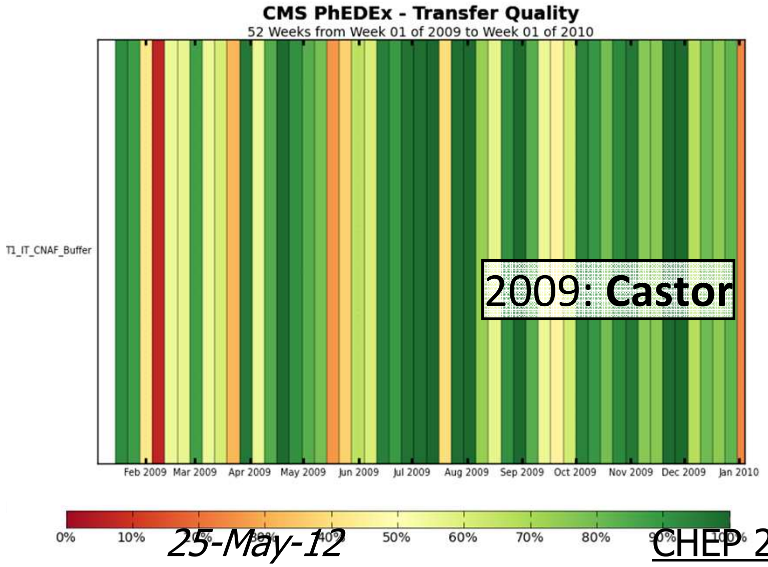
<u>CMS PhEDEx provides this in its standard monitoring</u>



**GEMSS** over last 12 months

migrating to GEMSS

**<u>Visible improvement</u> in the overall quality of transfers in the *->CNAF route**



2009: **Castor**

After moving to GEMSS…

**<u>CNAF storage was definitely more stable for CMS</u>**

# LHCb real data processing

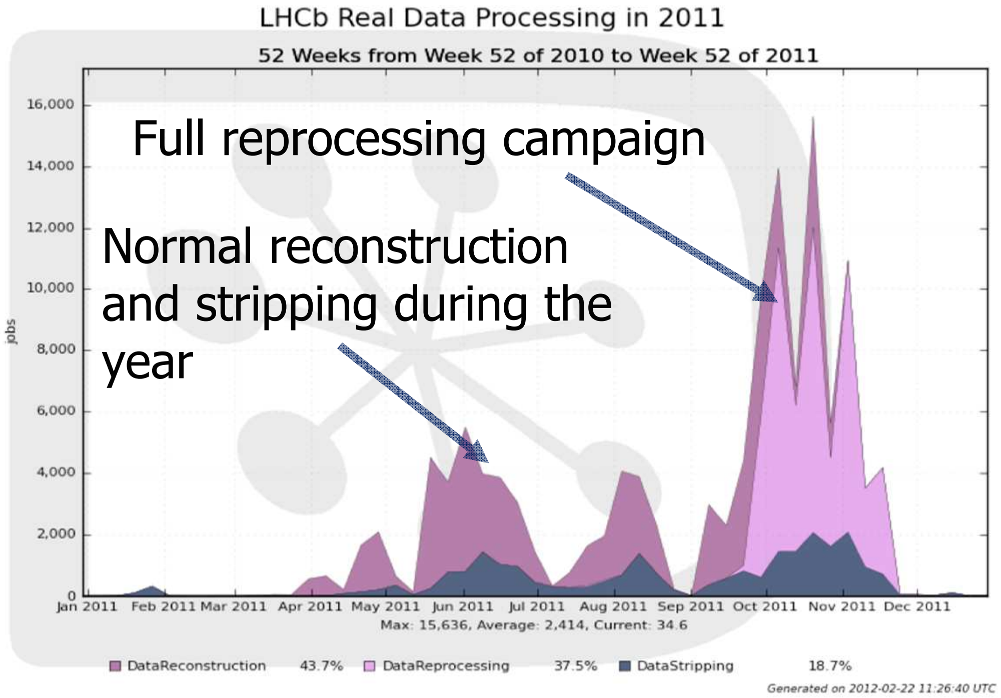
LHCb performs a “near-online” distributed reconstruction of raw data

- Data acquired by the detector is uploaded to CERN and then immediately distributed to the Tier1s

**- Raw data (RAW) and Reconstructed data (SDST) are stored on Tier1 tape**

As third category, data files used for analysis also go to tape storage for **long term archival (ARCHIVE)**

NOTE: It is important to keep for a very long time datasets that were used in order to produce published physics results

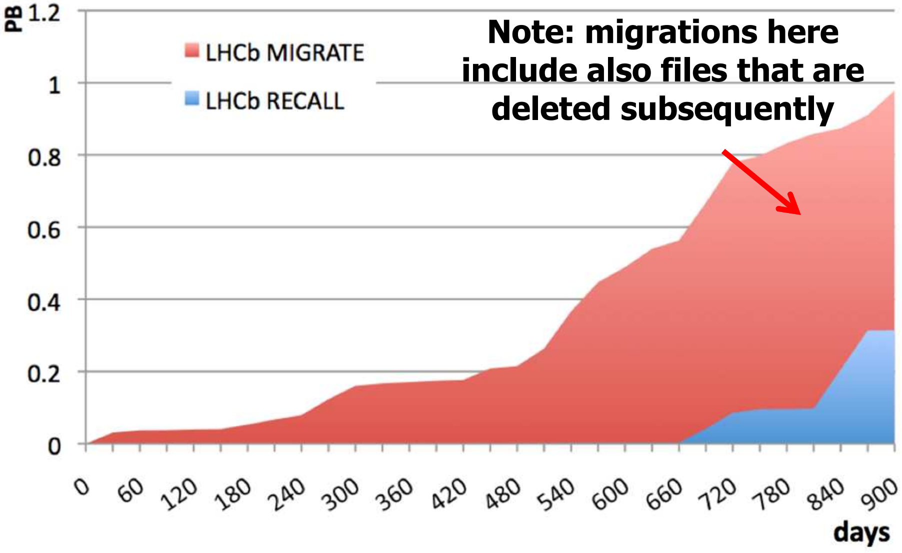


Full reprocessing campaign

Normal reconstruction and stripping during the year

# Present usage of tape resources for LHCb at CNAF

- **A total of about 0.75 PB of tape space in use in GEMSS**
  - RAW data: 0.17 PB
  - SDST data: 0.17 PB
  - ARCHIVE data: 0.41 PB

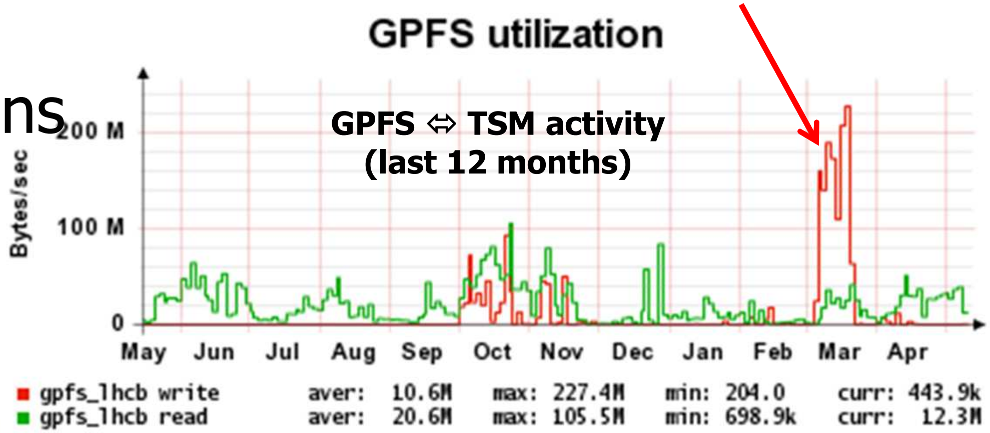


**Note: migrations here include also files that are deleted subsequently**

- **Few resources needed thanks to GEMSS design**
  - 2 dedicated HSM node

  No dedicated drives
  - Max 4 drives for migrations
  - Max 6 drives for recalls

**Massive tape recalls during reprocessing campaign**



**GPFS ⇔ TSM activity (last 12 months)**

# Disk stage area for LHCb tape

- **Tape stage** area shared with **Pure disk** areas
  - **LHCb currently has 0.76 PB in a dedicated GPFS area**
    - 40 TB at maximum are guaranteed for staging area
    - Remaining usable for analysis and for users disk space
  - **If more free space in the filesystem is avaliable, the stage area is dynamically expanded**
    - For example at present there are about 120 TB in use
  - **The minimum staging area is relatively small but shared across all the available volume (thanks to GPFS)**
    - maximal throughput performance (about 2 GB/s!)
    - allows to avoid wasting a lot of disk space in staging areas with the aim of providing a large sustainable throughput

# Conclusion

- The recent improvements of GEMSS have increased the level of reliability and performance of the storage access.
- GEMSS is the storage solution used in production in our Tier1 as a single integrated system for ALL the LHC and no-LHC experiments.
- Results from the experiment perspective of the latest years of production shows the system reliability and high performance with moderate effort. We are happy of our system!

…questions?