# INFN-T1 site report

Andrea Chierici, Vladimir Sapunenko
On behalf of INFN-T1 staff

HEPiX spring 2012

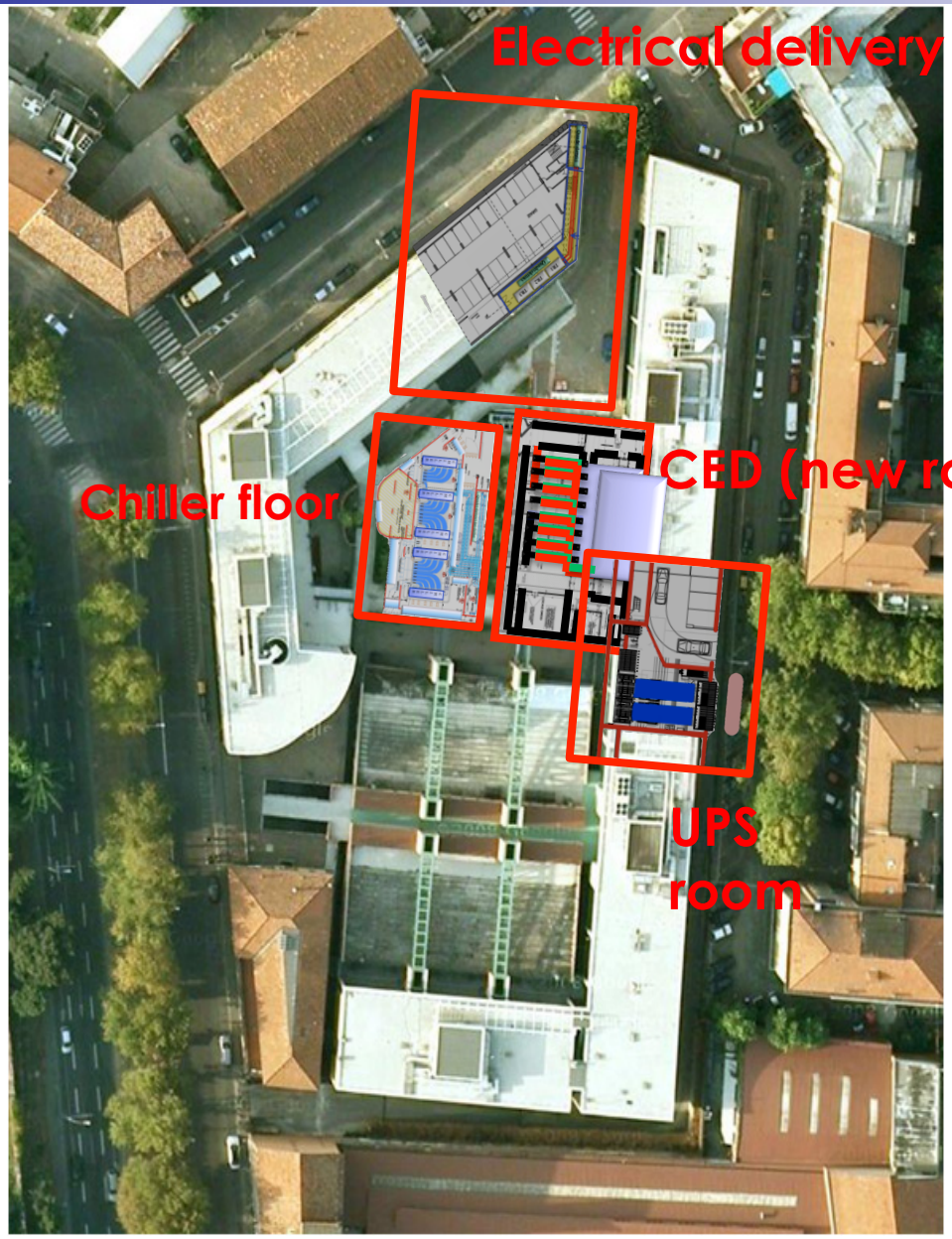
# Outline

- Facilities
- Network
- Farming
- Grid and Middleware
- Storage
- User experience

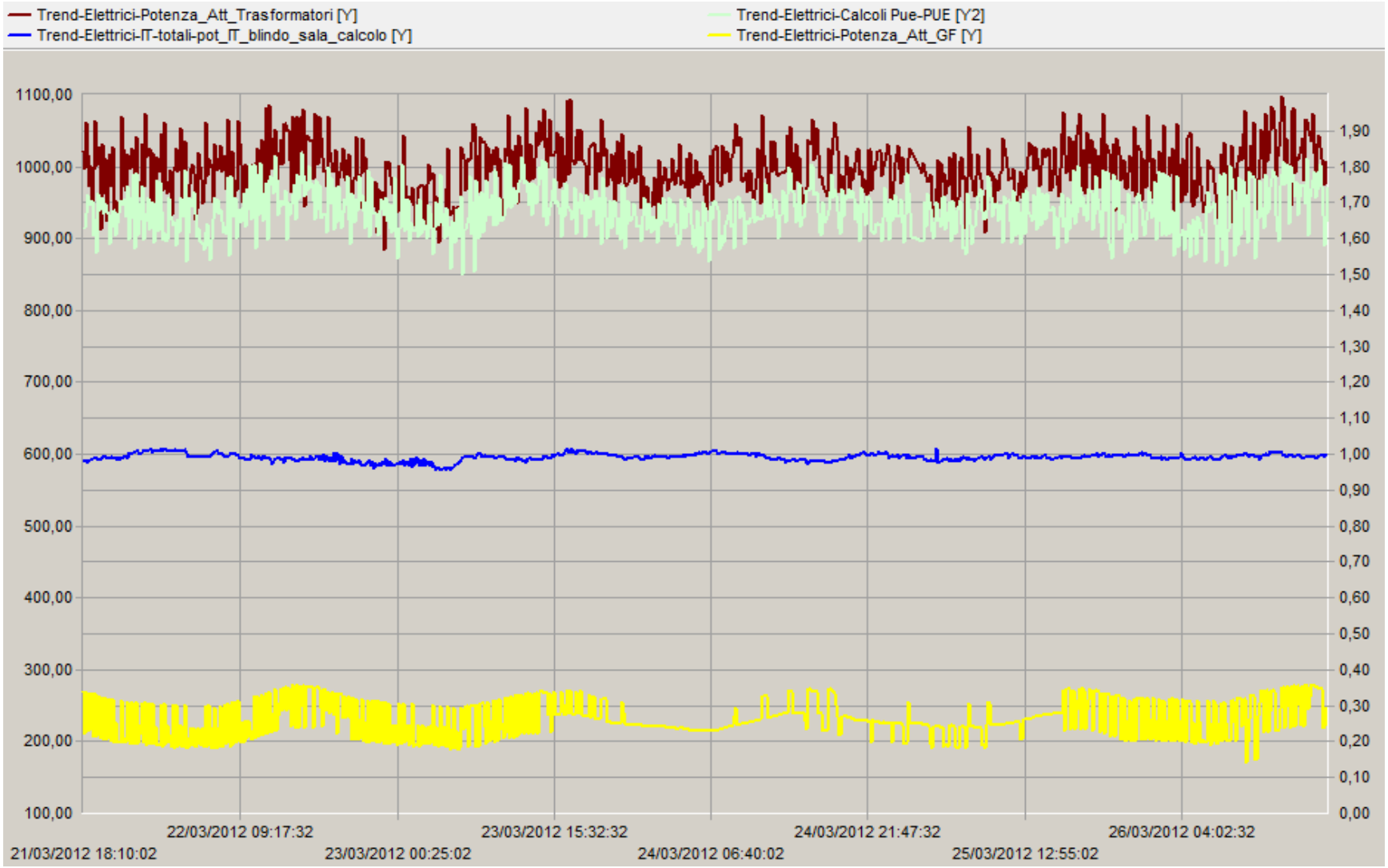
# INFN-Tier1: numbers

- **1000 m$^2$** room with capability for more than 120 racks and several tape libraries
  - **5 MVA** electrical power
  - Redundant facility to provide 24hx7d availability
- Within May (thanks to new 2012 tenders) even more resources
  - **1300 servers** with more than **10000 cores available**
  - **11 PBytes** of disk space and **14 PBytes** on tapes
    - Aggregate bandwith to storage: ~ 50 GB/s
- WAN link at **30 Gbit/s**
  - **2x10 Gbit/s over OPN**
  - With forthcoming GARR-X bandwidth increase is expected
- > 20 supported experiments
- ~ 20 FTE (currently evaluating some positions)

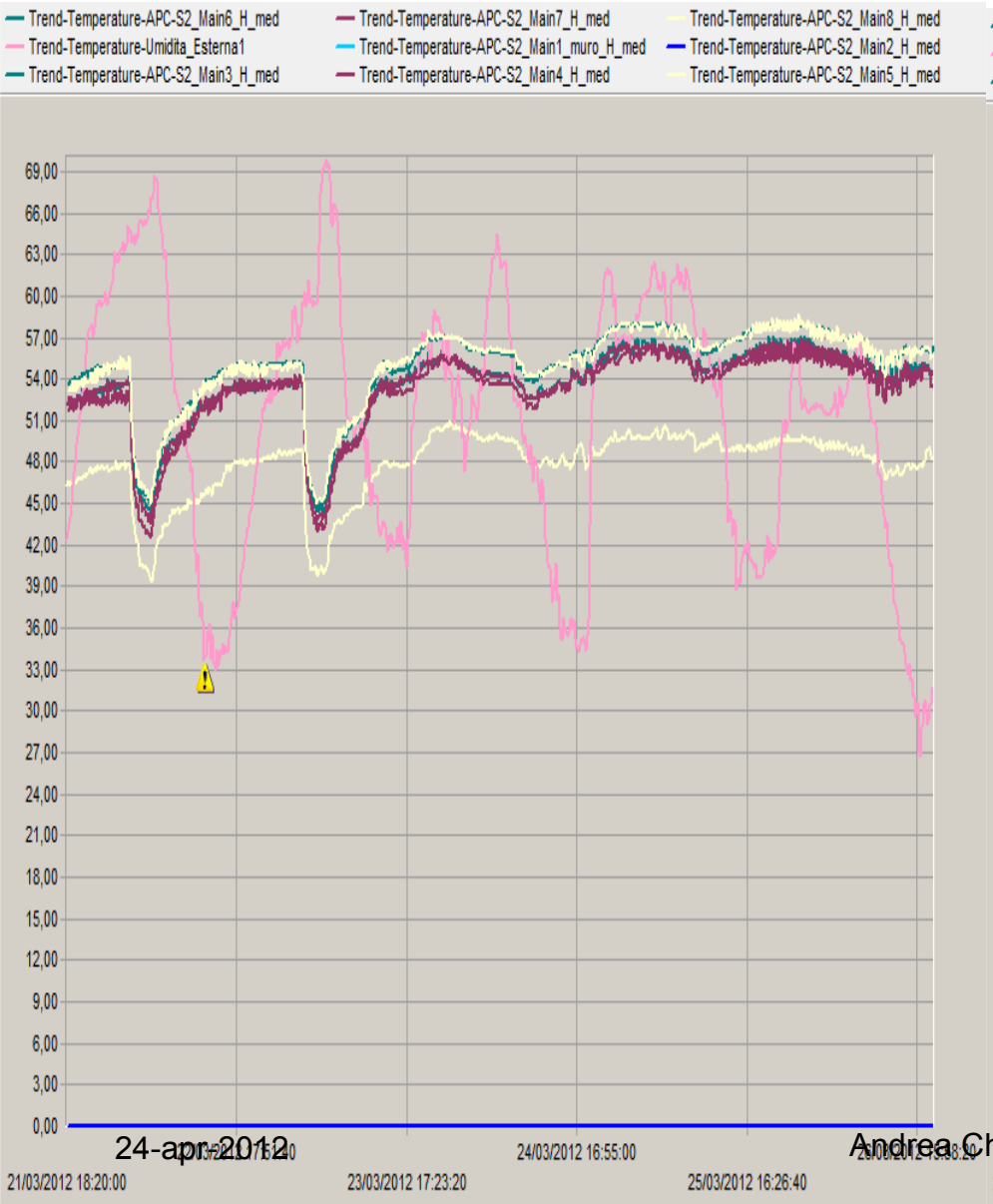
# Facilities

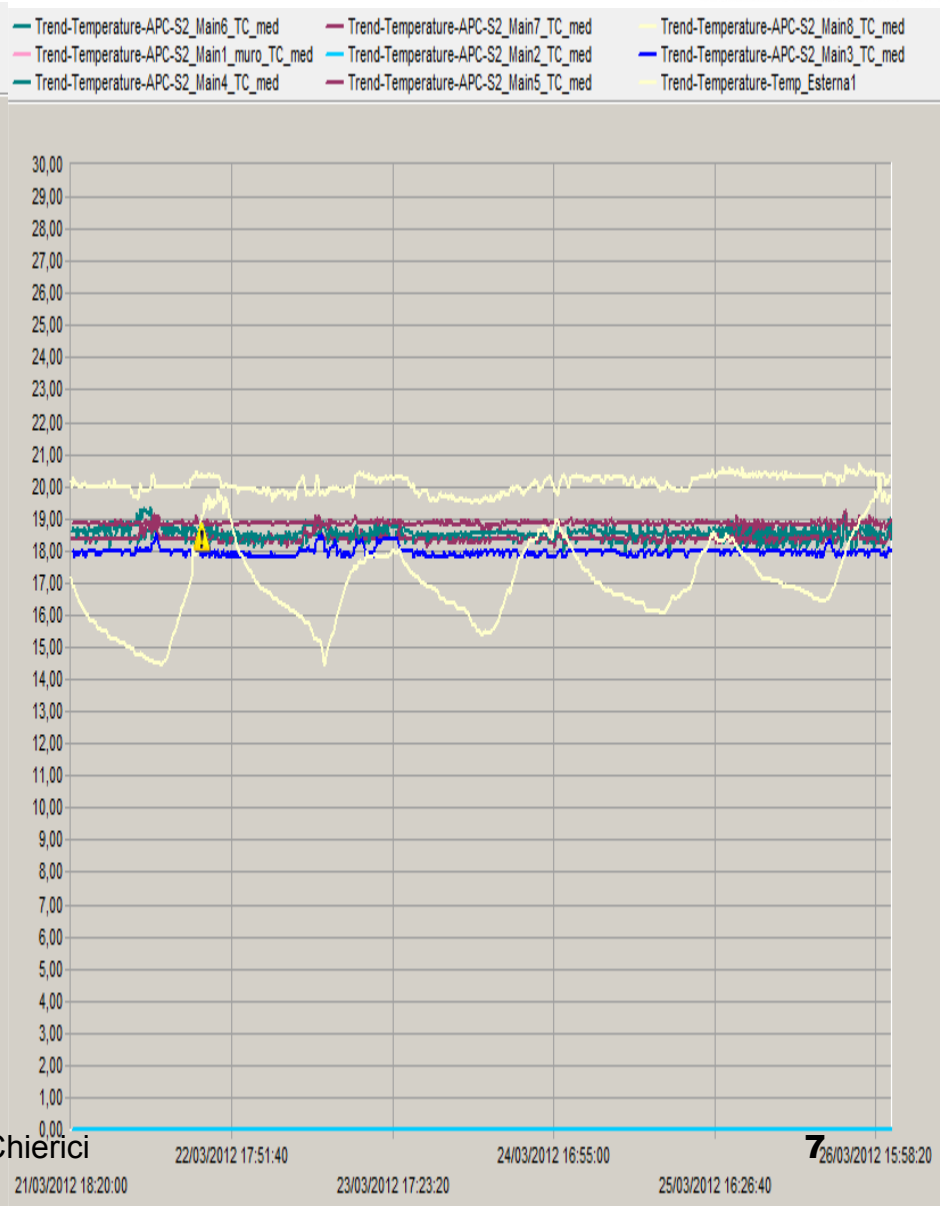

Electrical delivery
Chiller floor
CED (new room)
UPS
room

# Tier1 power (5 days)



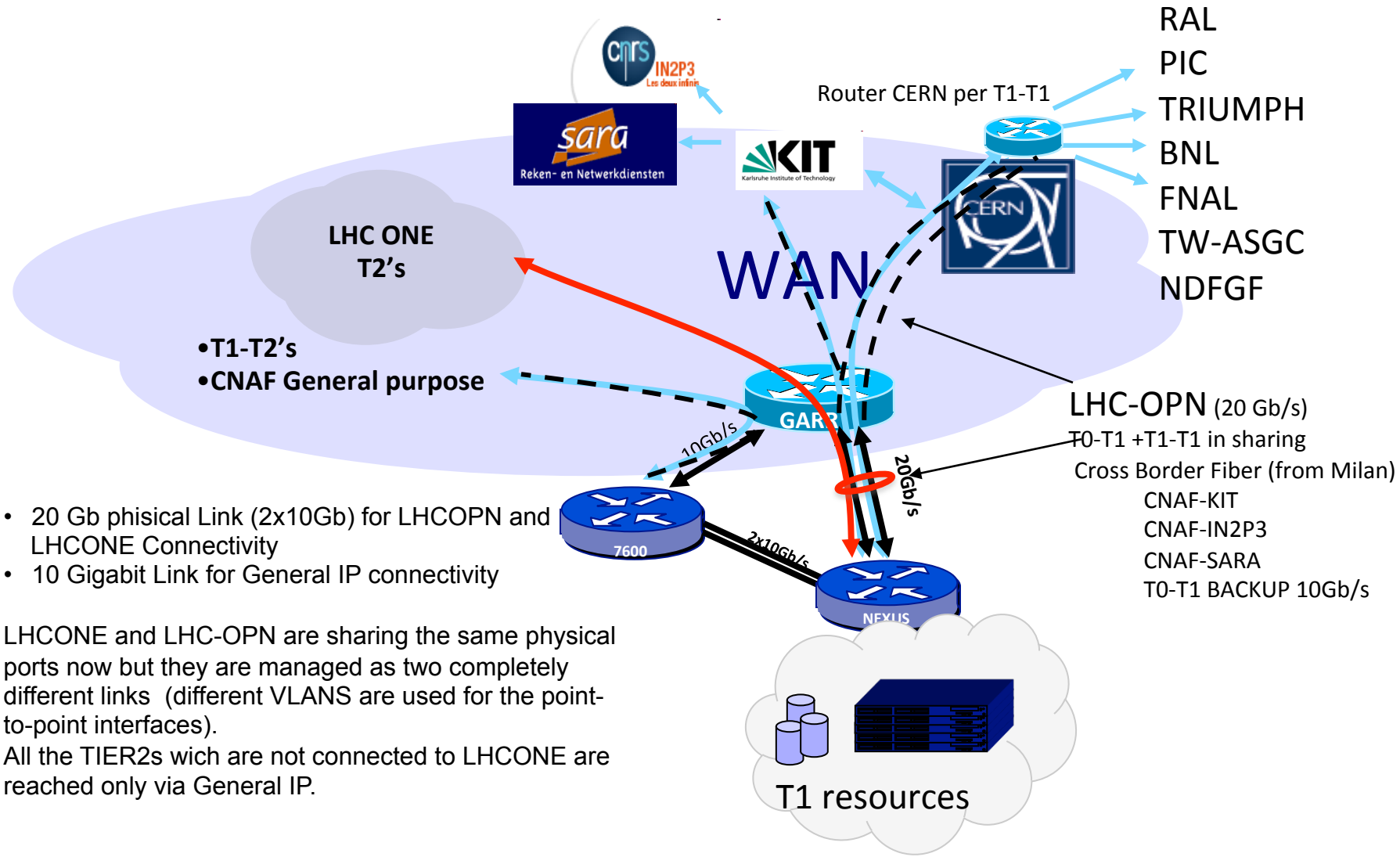
# APC blocks temp./hum. (5 days)



# Network

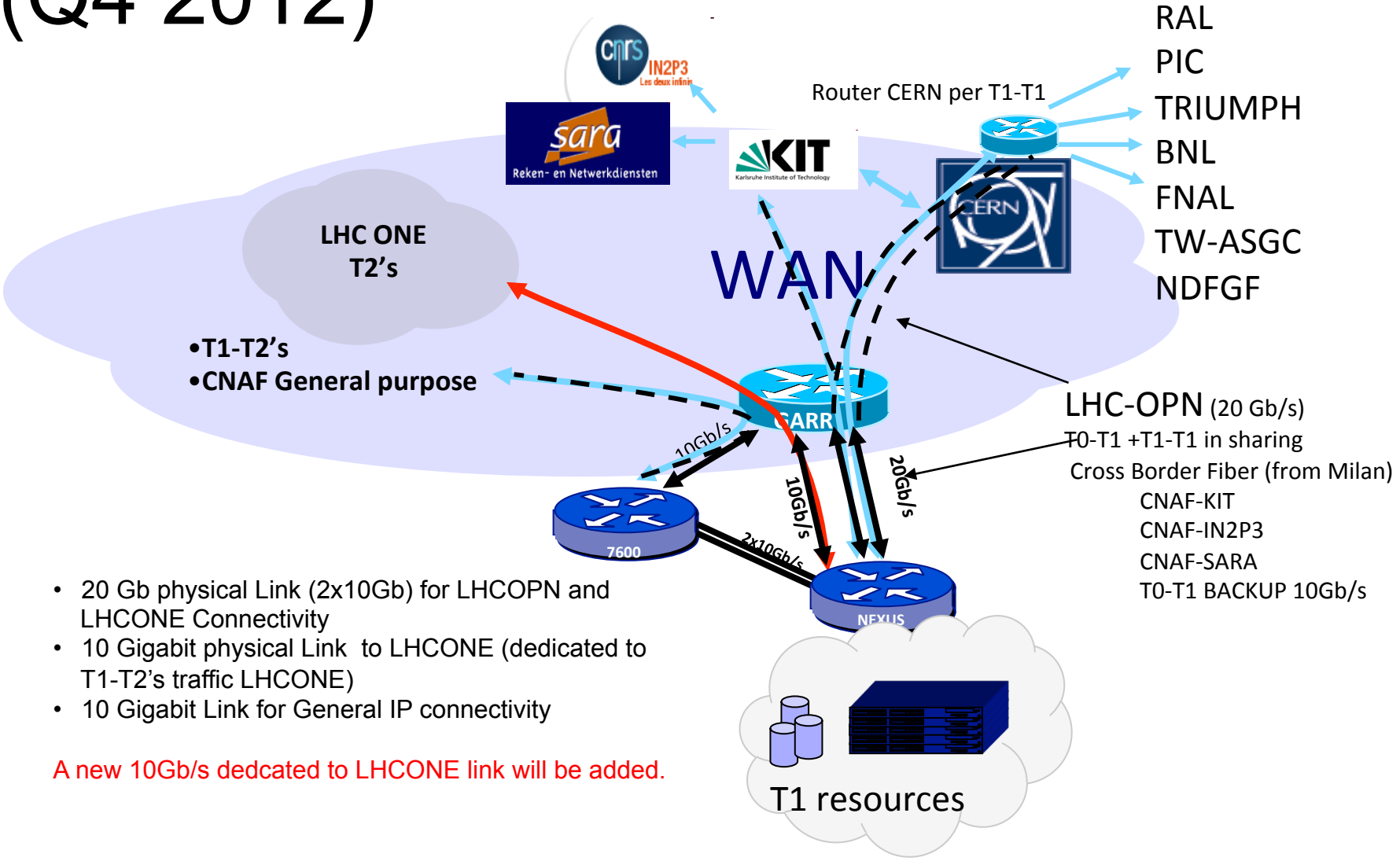
# Current WAN Connections (Q1 2012)




- 20 Gb phisical Link (2x10Gb) for LHCOPN and LHCONE Connectivity
- 10 Gigabit Link for General IP connectivity

LHCONE and LHC-OPN are sharing the same physical ports now but they are managed as two completely different links (different VLANS are used for the point-to-point interfaces).
All the TIER2s wich are not connected to LHCONE are reached only via General IP.

# Forthcoming WAN Connection (Q4 2012)




- 20 Gb physical Link (2x10Gb) for LHCOPN and LHCONE Connectivity
- 10 Gigabit physical Link to LHCONE (dedicated to T1-T2's traffic LHCONE)
- 10 Gigabit Link for General IP connectivity

A new 10Gb/s dedcated to LHCONE link will be added.

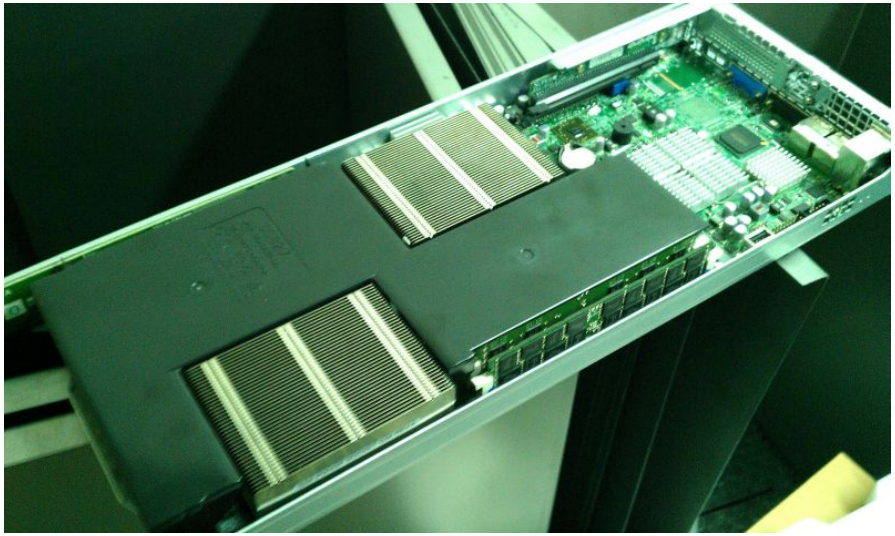
# Farming

# Computing resources

- Currently 120K HS-06
  - ~9000 job slots
- New tender will add 31,5K HS-06
  - ~4000 potential new job slots
  - 41 enclosures, 192 HS-06 per mobo
- We host other sites
  - T2 LHCb
  - T3 UniBO

# New tender machine

- 2U Supermicro Twin square
  - Chassis: 827H-R1400B
    - (1+1) redundant power supply
    - 12x 3.5" hot-swap SAS/SATA drive trays (3 for each node)
    - Hot-swappable motherboard module
  - Mobo: H8DGT-HF
    - Dual AMD Opteron™ 6000 series processors
    - AMD SR5670 + SP5100 Chipset
    - Dual-Port Gigabit Ethernet
    - 6x SATA2 3.0 Gbps Ports via AMD SP5100 controller, RAID 0, 1, 10
    - 1x PCI-e 2.0 x16
- AMD CPUs (Opteron 6238) 12 cores, 2,6Ghz
- 2 2tb sata hard disks 3,5”
- 40GB Ram
- 2 1620w 80gold power supply



# Issues for the future

- We would like to discuss problems with many-cores architectures
  - 1 job per core
  - RAM per core
  - Single gigabit ethernet per box
  - Local disk I/O
  - Green Computing: how to evaluate during tender procedures

# Grid and Middleware

# Middleware status

- Deployed several EMI nodes
  - UIs, CreamCEs, Argus, BDII, FTS, Storm, WNs
- Legacy glite-3.1 phased-out almost completely
- Planning to completely migrate glite-3.2 nodes to EMI within end of summer
- Atlas and LHCb switched to cvmfs for software area
  - Facing small problems with both
- Tests ongoing on cvmfs server for SuperB

# WNoDeS: current status

- WNoDeS is deployed in production for two VOs
  - Alice: no need for direct access to local storage
  - Auger: needs a customized environment requiring a direct access to a mysql server.
- Current version of WNoDeS deployed in production is using the GPFS/NFS gateway.

# WNoDeS: development

- WNoDeS will be distributed with EMI2
- New feature called mixed mode
  - Mixed mode avoids to statically allocate resources to WNoDeS
  - A job can be executed on a virtual resource dynamically instantiated by WNoDeS
  - The same resources (the same hv) can be used to execute standard jobs
  - No resources overbooking
  - No hard resource limit enforcement which will be provided using cgroup
- General improvements

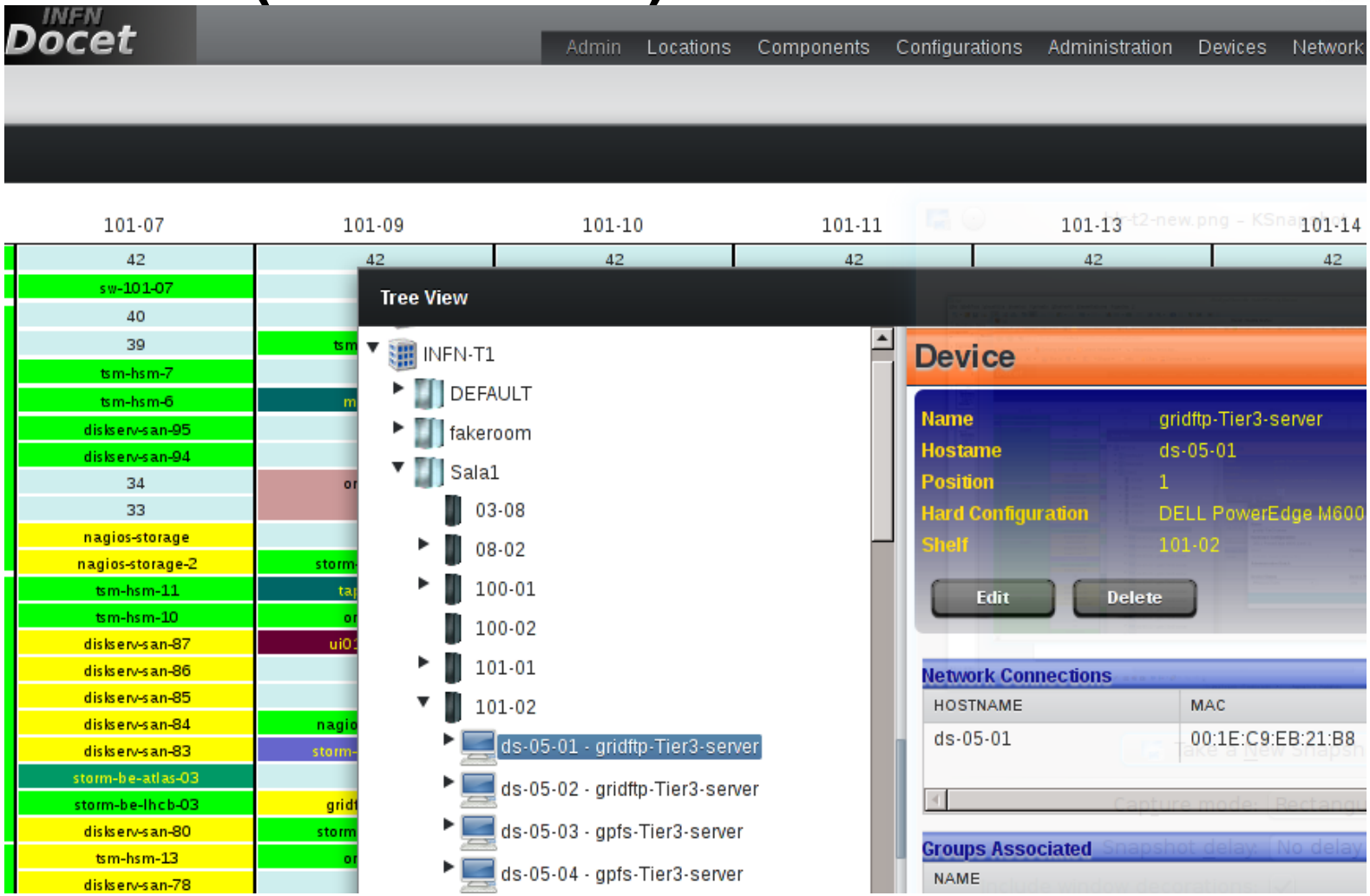
# Docet (Data Operation Center Tool)

- DB-based webtool designed and implemented internally
- In use at various INFN sites.
  - PostgreSQL, tomcat+java
  - Cmdline tool (rely on XMLRPC python webservice)
- Provides inventory for HW, SW, Actions & docs
- Initially populated by grabbing and cross-relating info from heterogeneous authoritative sources (Quattor, DNS, DHCP, XLS sheets, plaintext files) through a bunch of custom python scripts

# Docet (T1 tools)

- Cmdline Tools:
  - dhcpd.conf handling: preserves consistency (update from/to docet DB to be activated)
  - Cmdline query/update tool. It provides:
    - re-configuration data for nagios
    - add/list/update HW failures
    - Support for batch operations (insert new HW, remove dismissed HW)
  - We plan to deploy more docet-based configuration tools

# Docet (T1 tools)



# Storage

# Storage resources

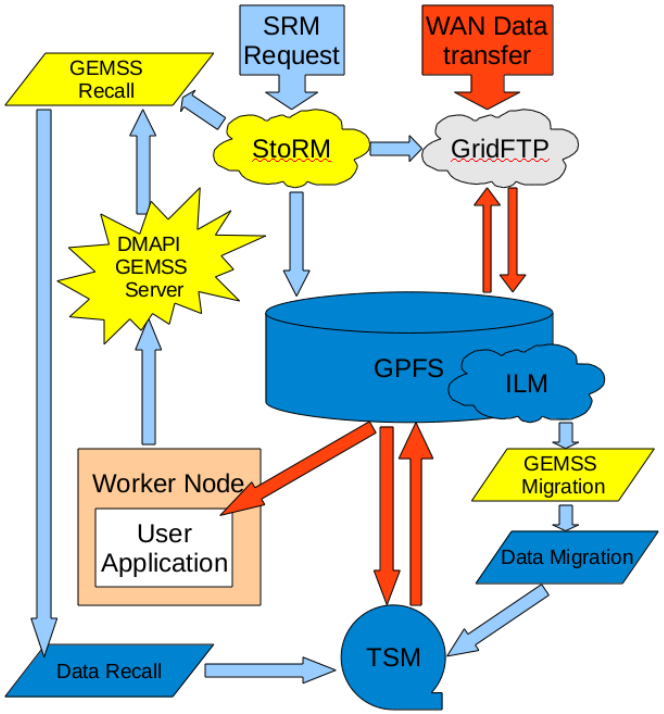
- 8.4 PB of on-line disk with GEMSS
  - 7 **DDN** S2A 9950
    - 2 TB SATA for data, 300 GB SAS for metadata
  - 7 **EMC²** CX3-80 + 1 **EMC²** CX4-960 (1 TB disks)
  - 2012 acquisition: 3 Fujitsu Eternus DX400 S2 (3 TB SATA)
- Servers
  - ~32 NSD servers (**10 Gbps** ethernet) on DDN
  - ~60 NSD servers (1 Gbps ethernet) on EMC²
- Tape library **SI8500** (14 PB on line) with **20 T10Kb** drives and **10 T10Kc** drives
  - 9000 x 1 TB tape capacity, 1 Gbps of bandwidth for each drive
  - 1000 x 5 TB tape capacity, 2 Gbps of bandwidth for each drive
  - Drives interconnected to library and tsm-hsm servers via dedicated SAN (TAN)
  - TSM server common to all GEMSS instances
- All storage systems and disk-servers are on SAN (FC4/FC8)

# GEMSS: Grid Enabled Mass Storage System

- Integration of GPFS, TSM and StoRM
- Our choice is driven by need **to minimize management effort**:
  - Very positive experience for scalability so far;
  - Large GPFS installation in production at CNAF since 2005 with increasing disk space and number of users;
- Over 8 PB of net disk space partitioned in several GPFS clusters served by less than 100 disk-servers (NSD + gridFTP);
  - 2 FTE employed to manage the full system;
  - All experiments at CNAF (LHC and non-LHC) agreed to use GEMSS as HSM

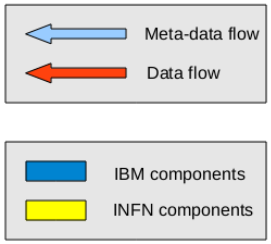
# GEMSS evolution

- 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;
  - Available with GPFS v.3.x

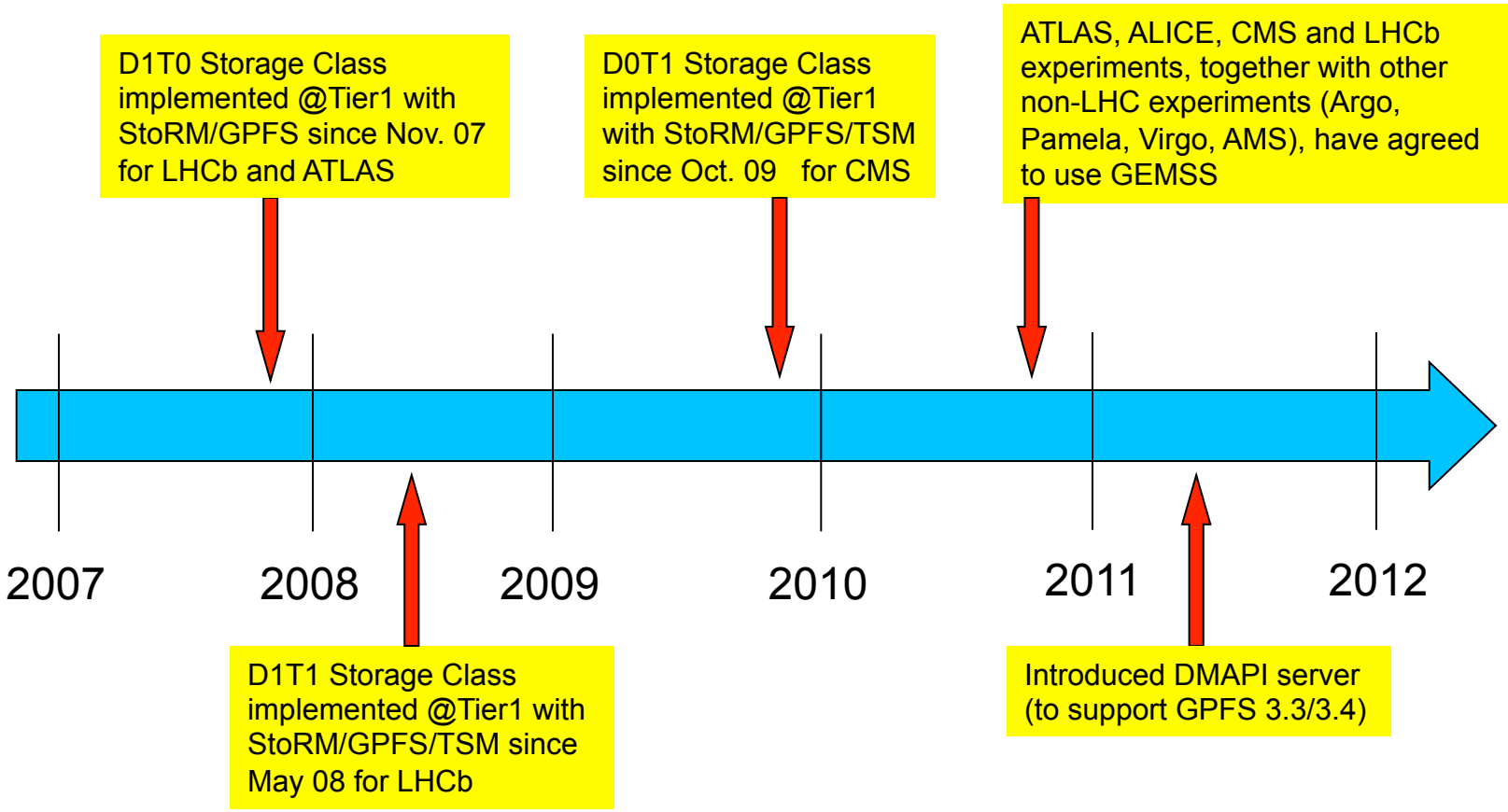



Disk-centric system with five building blocks

1. GPFS: disk-storage software infrastructure
2. TSM: tape management system
3. StoRM: SRM service
4. TSM-GPFS interface
5. Globus GridFTP: WAN data transfers



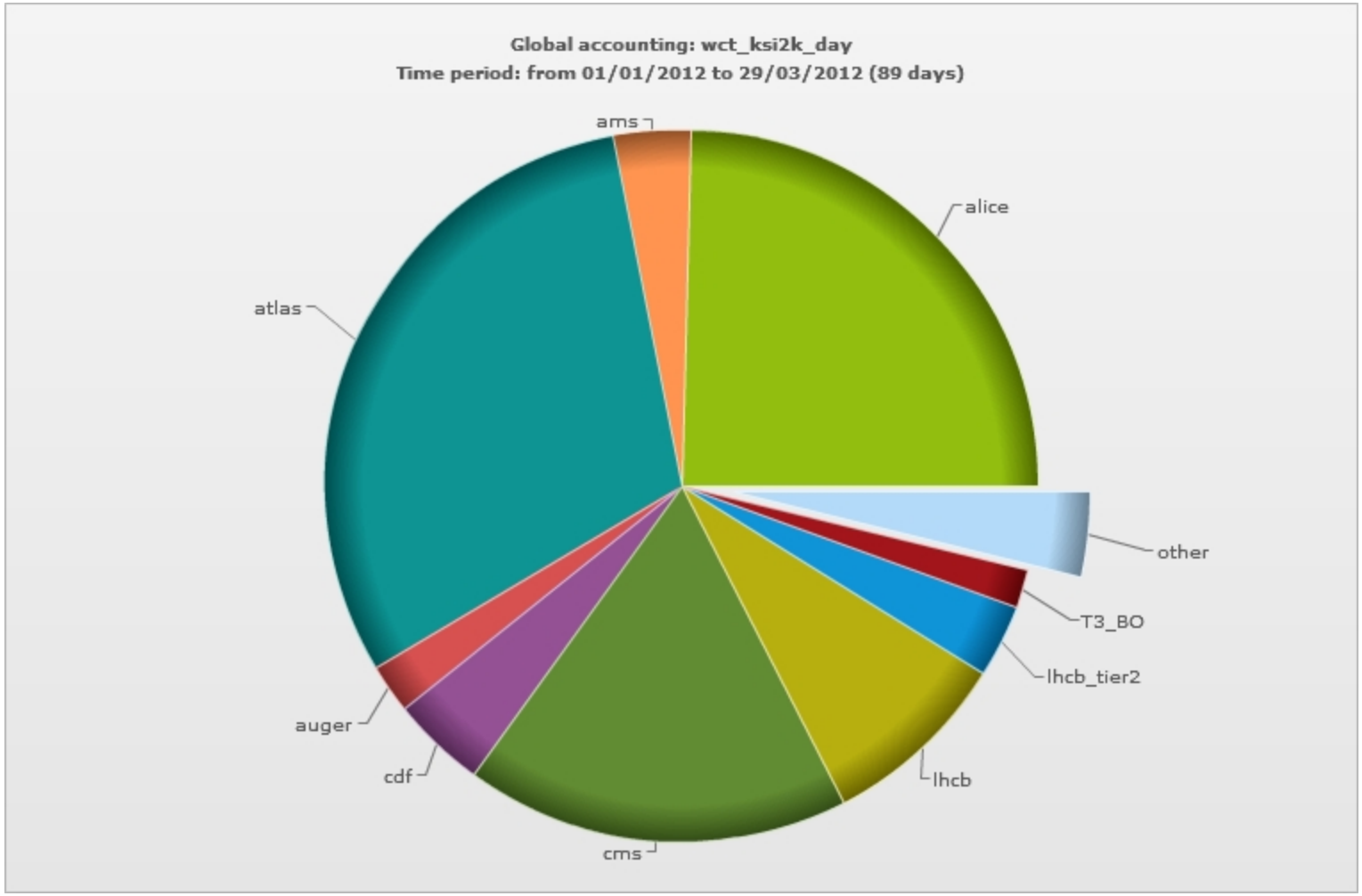
# GEMSS: Timeline



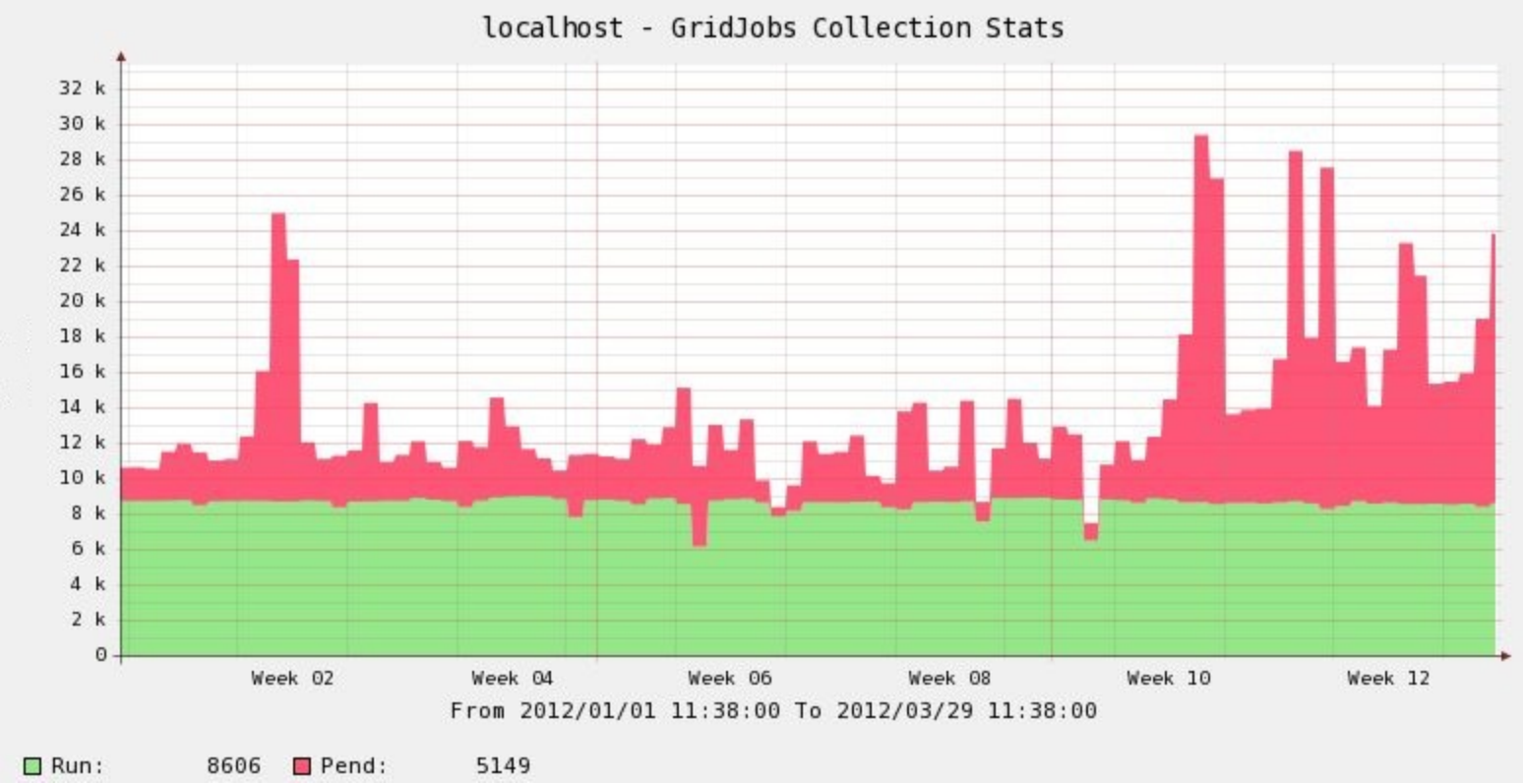

**GEMSS is now used by all LHC and non-LHC experiments in production for all Storage Classes**

# User experience

# Resource usage per VO



# Jobs


localhost - GridJobs Collection Stats
32 k
30 k
28 k
26 k
24 k
22 k
20 k
18 k
16 k
14 k
12 k
10 k
8 k
6 k
4 k
2 k
0
Week 02
Week 04
Week 06
Week 08
Week 10
Week 12
From 2012/01/01 11:38:00 To 2012/03/29 11:38:00
Run: 8606
Pend: 5149

# LHCb feedback

- More than 1 million jobs executed
  - Analysis: 600k, Simulation 400k
  - User analysis not very efficient (about 50%): too large bandwidth requested
    - available bandwidth for LHCb will be significantly raised with 2012 pledges
- Stability of the services during last year
  - Small fraction of failed jobs
  - Good performance of data access, both from tape and disk

# CMS feedback

- 5 Kjobs/day  average (50 Kjobs/day peak)
- Up to 200 TB data transfers per week both in and out
- Recent tests proved the possibility to work with 10 GB files in a sustained way

# Alice feedback

- $6.93 \times 10^6$ KSI2k hours consumed in one year of very stable running
- The cloud infrastructure based on WNoDes provided excellent flexibility in catering to temporary requirements (e.g. large memory queues for PbPb event reconstruction).
- CNAF holds 20% of ALICE RAW data (530 TB on tape)
- All data are accessed through an XROOTD interface over the GPFS+TSM underlying file system

# Questions?