INFN-T1 site report

Luca dell'Agnello

On behalf ot INFN-T1 staff

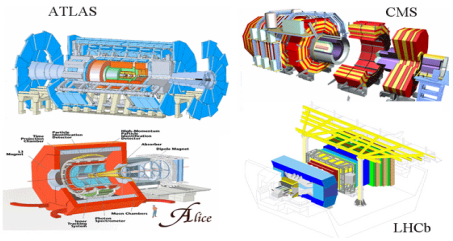
HEPiX Spring 2013

Outline

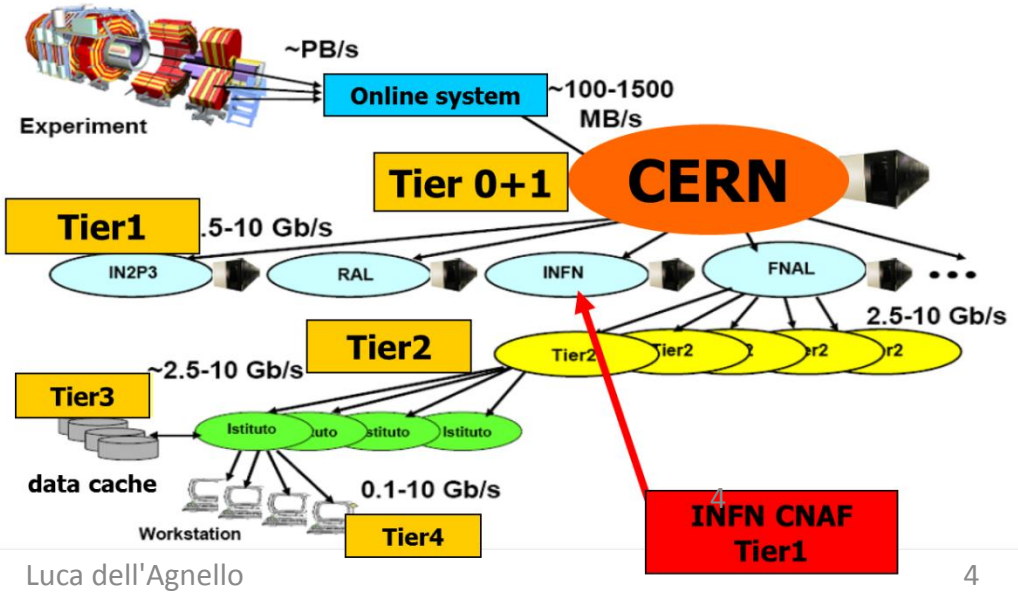
- Introduction
- Network
- Farming
- Storage
- Grid and Middleware
- User experience

INFN-Tier1

- CNAF, is the Italian Tier-1 computing centre for the LHC experiments ATLAS, CMS, ALICE and LHCb....
- … but also one of the main Italian processing facilities for several other experiments



- BaBar and CDF
- Astro and Space physics
  - VIRGO (Italy), ARGO (Tibet), AMS (Satellite), PAMELA (Satellite) and MAGIC (Canary Islands)
- And other (e.g. Icarus, Borexino, Gerda etc…) for a total of more than 20 experiments

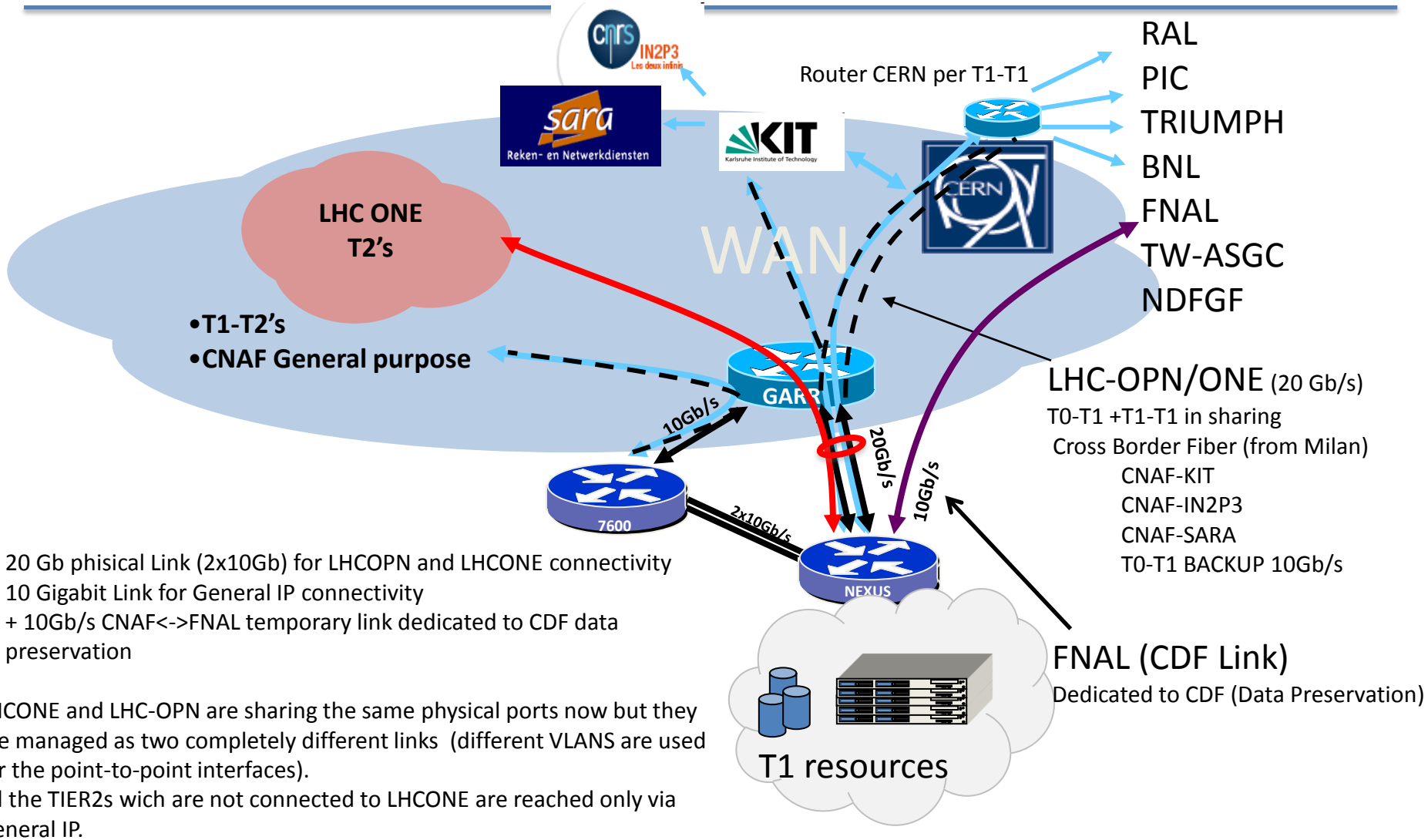


INFN-Tier1: some figures

- **1000 m²** room with capability for more than 120 racks and several tape libraries (only one in production)
  - **5 MVA** electrical power
  - Redundant facility to provide 24hx7d availability
- 1300 server with about 13800 cores available (~ 135 KHS06)
- **~11.4 PBytes of disk space** for high speed access and **~16 PBytes** on tapes
  - tape library capacity currently scalable up to 50 PB
  - Aggregate bandwidth to storage: ~50 GB/s
- 10 Gbit technology on LAN
  - 3x10 Gbps WAN connections (to be upgraded to 5x10 Gbps)
- ~20 FTEs to manage facilities, network, databases, storage and farm services (including MSS, CEs, SRM end-points, FTS ….)

Network

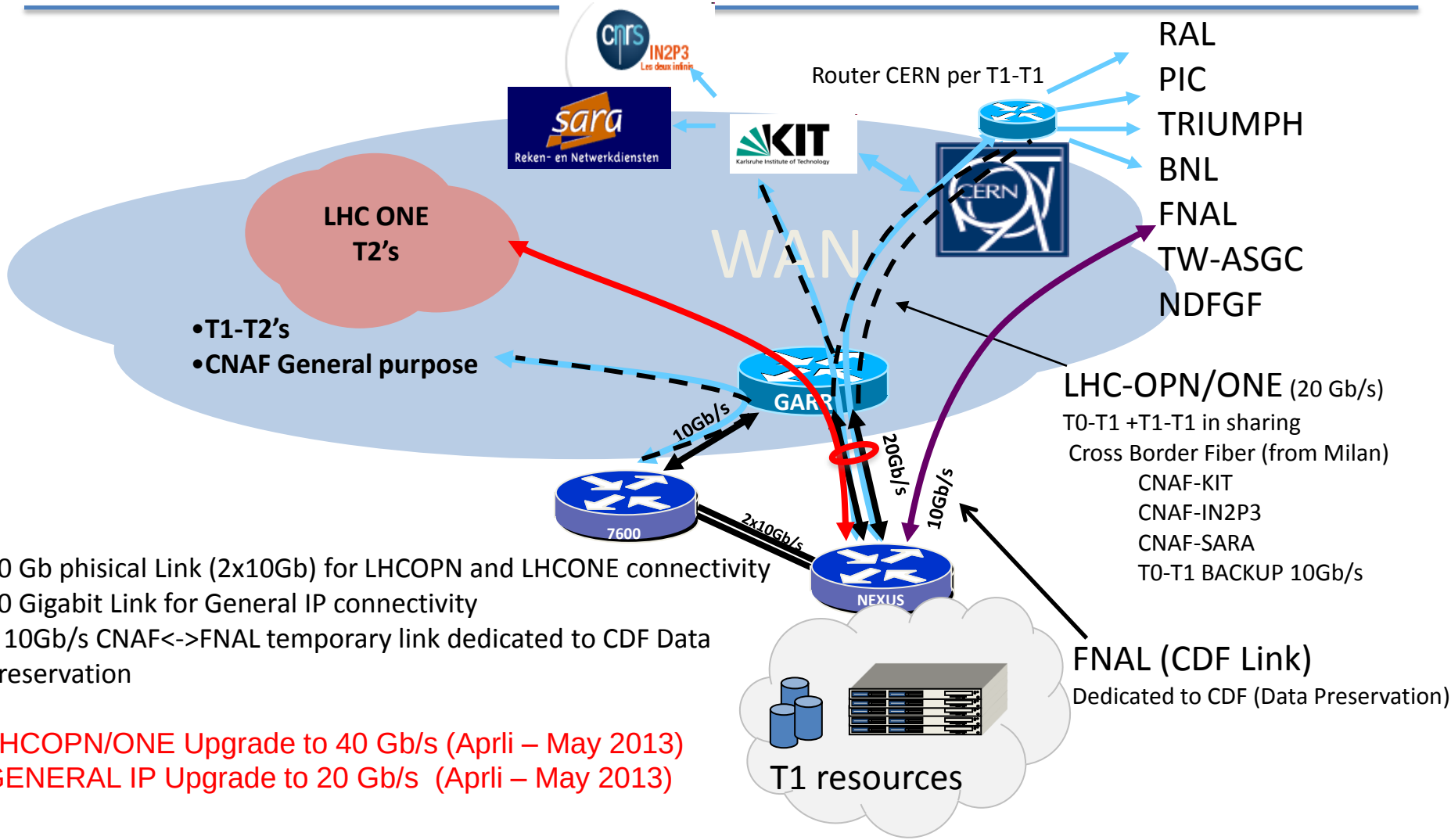
WAN Connections (Q1 2013)



- 20 Gb phisical Link (2x10Gb) for LHCOPN and LHCONE connectivity
- 10 Gigabit Link for General IP connectivity
- + 10Gb/s CNAF<->FNAL temporary link dedicated to CDF data preservation

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.

WAN Connection (Q2 2013)



- 20 Gb phisical Link (2x10Gb) for LHCOPN and LHCONE connectivity
- 10 Gigabit Link for General IP connectivity
- + 10Gb/s CNAF<->FNAL temporary link dedicated to CDF Data Preservation

- LHCOPN/ONE Upgrade to 40 Gb/s (Aprli – May 2013)
- GENERAL IP Upgrade to 20 Gb/s (Aprli – May 2013)

New activities

- Almost completed transition to Jumbo Frame on LAN
  - 10gbit disk and gridftp servers already using JF since 2010
  - Farm just reconfigured
  - Reconfiguration of other services on going
- IPV6 activation on General IP (Q2 2013)
  - Testing dual stack on dedicated services (gridftp, xrootd, CEs, StoRM end-points….)
- IPV6 activation on LHCOPN (Q4 2013)
  - Coordinated effort with HEPIX IPv6 wg
  - done if the previous testing activity gives good results
- SDN (First steps..)
  - OpenFlow Layout test (Q2-4 2013)
    - Interoperability test on physical switches (on trial)

Farming

Computing resources

- Currently 135K HS-06 in the farm
  - 13.800 job slots
  - HT activated only on subset of the Intel based part of the farm (~10 % of the wns)
- New tender will add 45K HS-06
  - ~4000 new job slots
  - 65 enclosures, 172 HS06 per mobo
  - Oldest nodes will be switched off this year

2013 CPU tender

- 2U Supermicro Twin square
  - Chassis: 827HQ-R1620B
    - (1+1) redundant power supply
    - 12x 3.5" hot-swap SAS/SATA drive trays (3 for each node)
    - Hot-swappable motherboard module
  - Mobo: H8DGT-HLF
    - Dual AMD Opteron™ 6000 series processors
    - AMD SR5690 + 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 6320) 8 cores, 2.8Ghz
- 2.2 TB SATA hard disks 3,5”
- 64GB Ram DDR3-1600 ECC REG
- 2x1620W 80+ platinum power supply
- 172HS06 on 32bit, SL6

New activities

- Investigating Grid Engine as an alternative batch system (to LSF)
  - Comparison with SLURM (being tested by INFN BARI)
- Job Packing
  - See Stefano's presentation
- Study of power consumption optimization
  - New CPU architectures (Atom, ARM etc..)
  - Software control (DCM by Intel)

Storage

Storage resources

TOTAL of 11.4 PB on-line (net) disk space

- 7 **EMC²** CX3-80 + 1 **EMC²** CX4-960 (~1.9 PB) + 100 servers (2x1 gbps connections)
- 7 **DDN** S2A 9950 + 1 DDN SFA 10000 (~9.4 PB) + ~60 servers (10 gbps)
- Aggregate bandwidth: 50 GB/s

- Tender for additional disk (+ 1.9 PB-N) completed
  - DDN SFA 12k + 24 servers
- Tape library **SI8500 ~** 16 PB on line with **20 T10KB** drives and **10 T10KC** drives (to be increased)
  - 8800 x 1 TB tape capacity, ~ 100MB/s of bandwidth for each drive
  - 1200 x 5 TB tape capacity, ~ 200MB/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
- Tender to expand library capacity in Q2 2013
- All storage systems and disk-servers are on SAN (4Gb/s or 8Gb/s)

Storage configuration

- All disk space is partitioned in several GPFS clusters served by ~150 servers
  - One cluster per each main experiment
  - GPFS deployed on the SAN implements a full HA system
  - System scalable to tens of PBs and able to serve thousands of concurrent processes with an aggregate bandwidth of tens of GB/s
- GPFS coupled with TSM offers a complete HSM solution
- Access to storage granted through standard interfaces (posix, srm, xrootd and webdav)
  - Fs directly mounted on wns

New activities (LTDP)

- Long Term Data Preservation for CDF
  - ~5 PB of data to be copied from FNAL
  - Starting copy in May
  - 5 Gbps dedicated link CNAF ←→ FNAL
  - Also “code preservation” issue to be addressed
- Participation to interdisciplinary LTDP project (PideS)
  - Proposal submitted

Grid and Middleware

Grid Middleware status

- EMI-2 update status
  - UIs, Argus, BDII, WNs
  - Many nodes at SL6
  - Some issues with Cream EMI-2/SL6
    - Upgrading to EMI-3/SL6
  - Storm will be deployed on EMI-3
  - EMI-1 phasing-out
  - Only VOBOX left at glite-3.2
- Separate queue to test WNs SL6/EMI-2
  - One VO left to validate

WNoDeS status

- (Virtual) WNs on demand
- Distributed with EMI middleware
- Deployed on part of INFN-T1 farm
  - Mixed mode allows both virtual and real jobs on the same hv
  - Auger VO main user for this solution, requiring a direct access to a dedicated mysql server
  - Also investigating cloud computing

User experience

Supported communities

- Besides the LHC experiments CNAF supports other scientific communities
  - Directly, through the CNAF User Support Service
    - CDF, Babar, (SuperB,) LHCf, Argo, AMS2, Auger, Icarus, Fermi/GLAST, Pamela, Borexino, Xenon, Gerda, Virgo, CTA, local theoretical physics and astronomy groups
  - Indirectly, through the Grid projects
- Different level of computing expertise and heterogeneous computing tools used
- Trying to attract more user communities

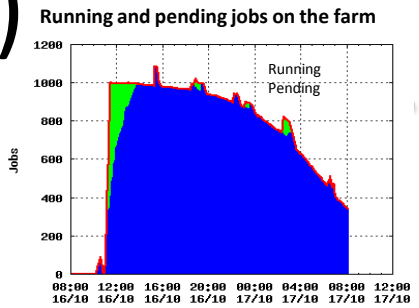
User Support Service

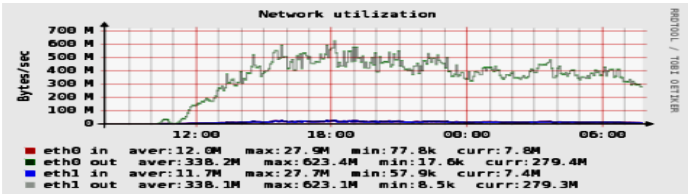
- Team of seven skilled young people plus a coordinator
- Used to assign support people to experiments, now migrating to a model with more work done in common by the team
  - Facilitates the transfer of know-how from big to small user communities
  - Facilitates the integration of new team members
    - Typically 2 to 4 years contracts
- Independent of the Tier-1 team but working in close contact

Backup slides

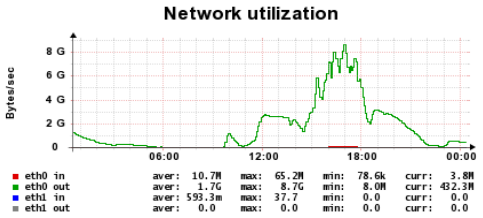
GEMSS in production (1)

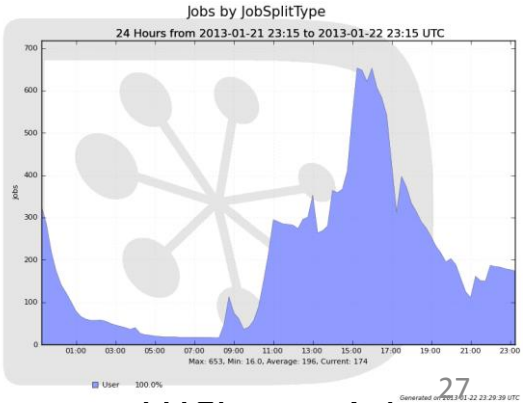
- *Gbit technology* (2009)
  - Using the file protocol (i.e. direct access to the file)
  - Up to 1000 concurrent jobs recalling from tape ~ 2000 files
    - 100% job success rate
    - Up to 1.2 GB/s from the disk pools to the farm nodes
- *10 Gbit technology* (since 2010)
  - Using the file protocol
  - Up to 10000 concurrent jobs accessing files on disk
    - Up to 10 GB/s from the same fs to the farm nodes
  - WAN links towards saturation



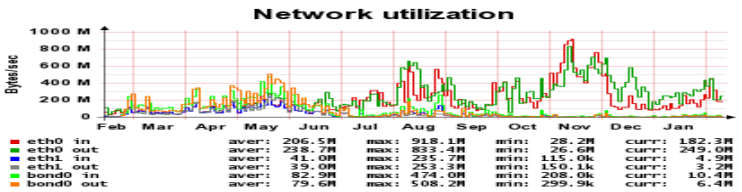


Aggregate traffic on eth0 network cards (x2)





LHCb user jobs



Building blocks of GEMSS system

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