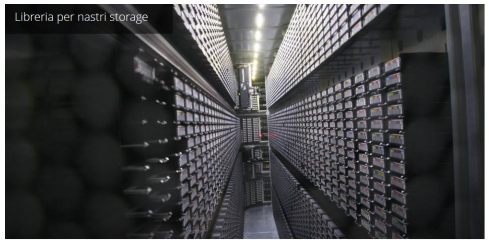
CNAF e CSN2

Luca dell'Agnello, Daniele Cesini – INFN-CNAF

CCR WS@LNGS - 23/05/2017

Il Tier1@CNAF – numbers in 2017

- 25.000 CPU core



- 27PB disk storage
- 70PB tape storage

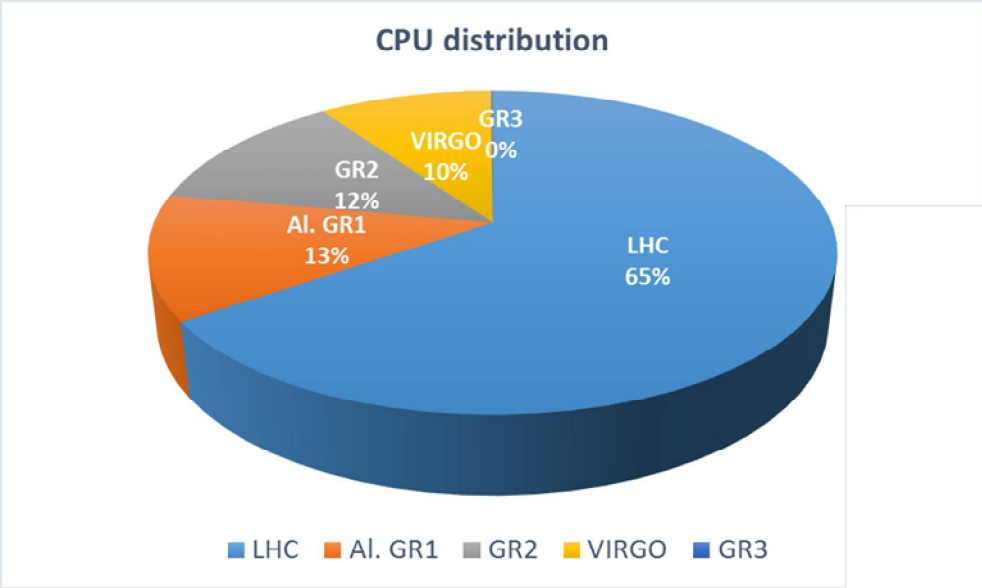
- 1.2MW Electrical Power
- PUE = 1.6

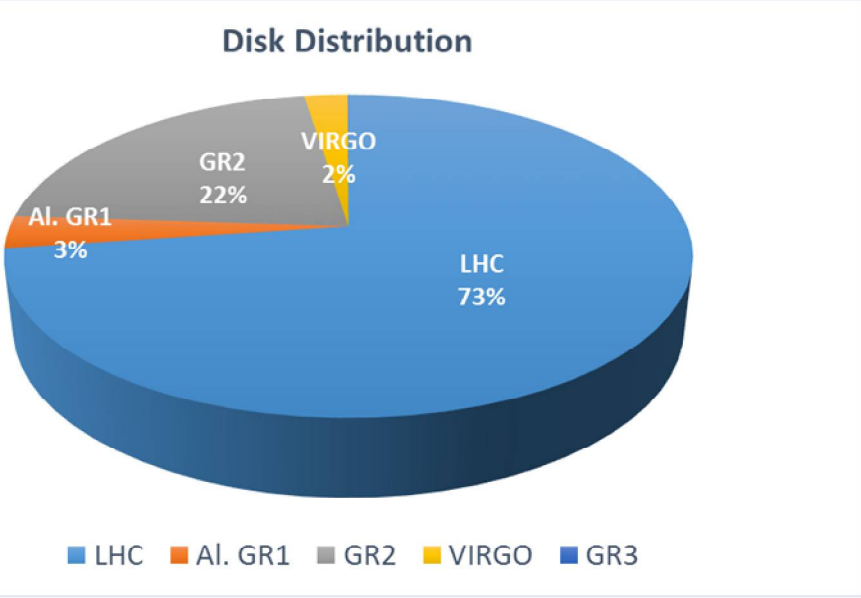
Experiments @CNAF

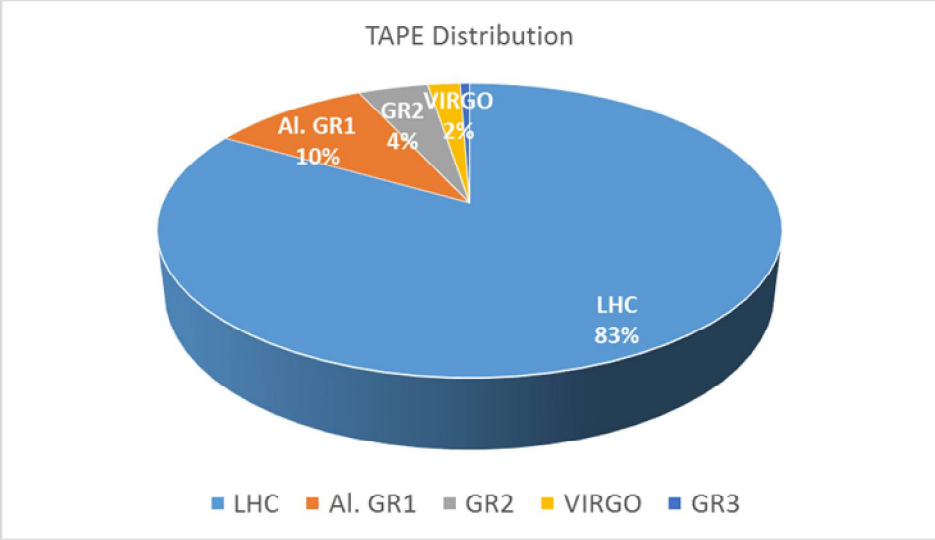
- CNAF-Tier1 is officially supporting 38 experiments
  - 4 LHC
  - 34 non-LHC
    - 22 GR2 + VIRGO
    - 4 GR3
    - 7 GR1 non LHC

- Ten Virtual Organizations in opportunistic usage via Grid services (on both Ter1 and IGI-Bologna site)

Resource distribution @ T1

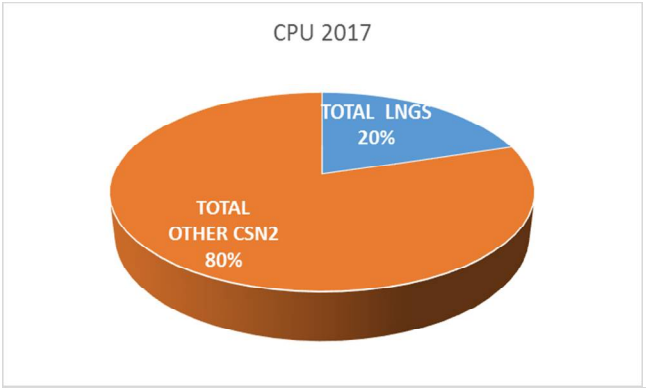


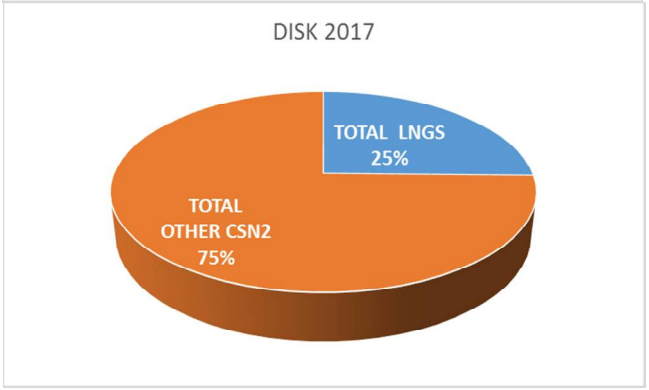


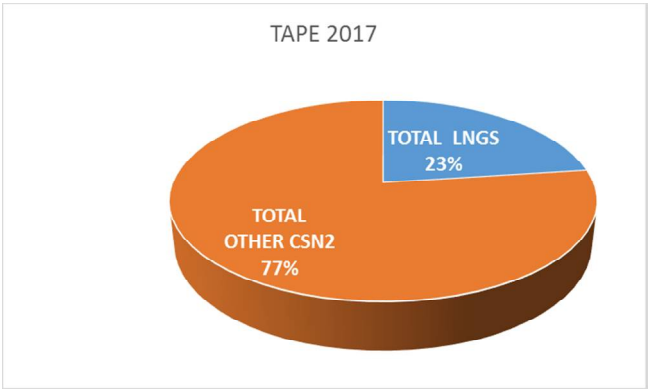


LNGS Experiments @ T1

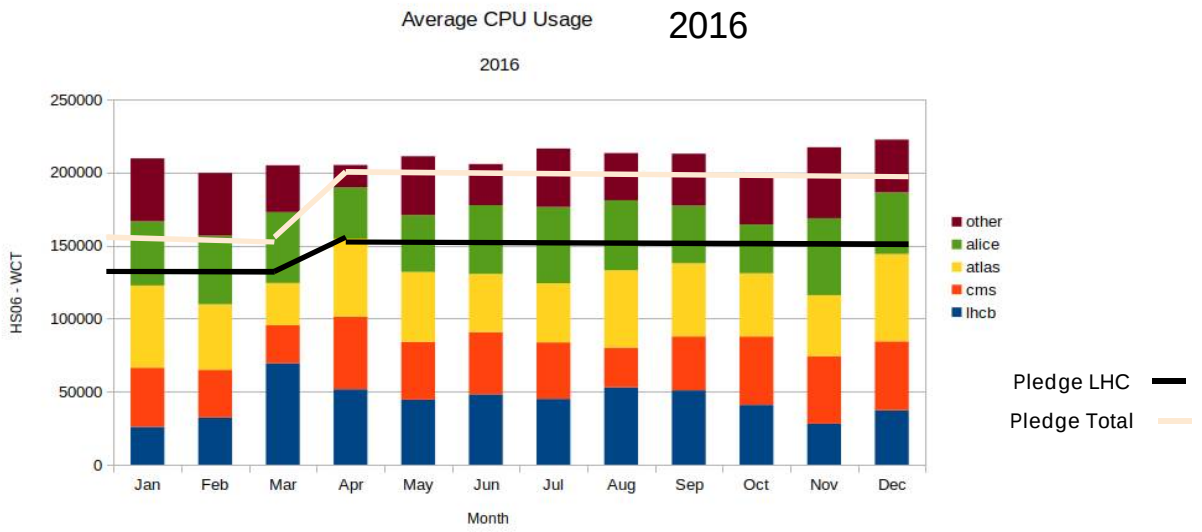
Experiment LNGS	2017		
	CPU HS06	DISK TB-N	TAPE TB
ICARUS	0	0	330
XENON	700	110	0
Borexino	1500	169	10
Gerda	40	45	40
DARKSIDE	4000	860	300
CUORE	1400	262	0
CUPID	100	15	0
COSINUS	0	0	0
LSPE	1000	7	0
TOTAL	8740	1468	680

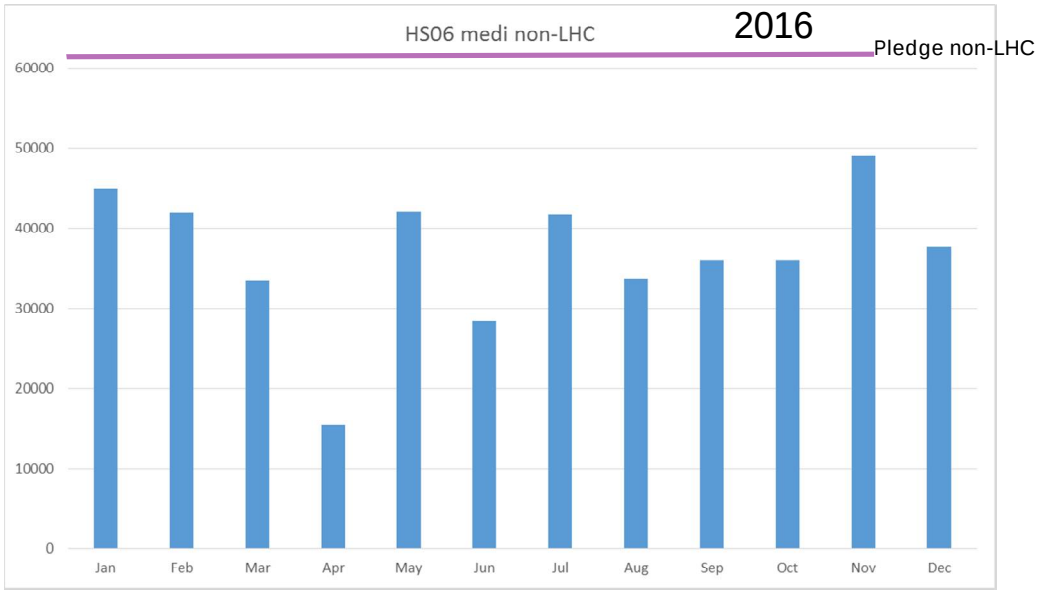






CPU usage

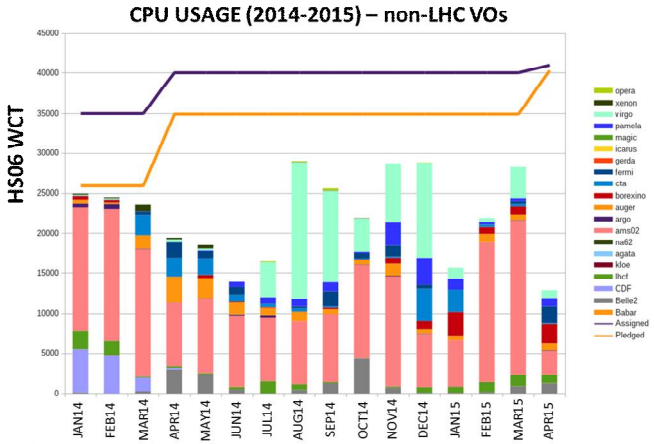




Under-usage, but typical burst activity for these VOs

- Allow to cope with overpledge requests

2014-2015



Extrapledge management

- We try to accommodate extrapledge requests, in particular for CPU
  - Handled manually
    - RobinHood
    - Identification of temporary inactive VOs
    - One offline rack that can be turned on if needed
  - Much more difficult for storage
- Working on automatic solution to offload to external resources
  - Cloud
    - ARUBA,Microsoft
    - HNSciCloud project
  - Other INFN sites
    - Extension to Bari

Resource Usage – number of jobs

LSF handles the batch queues

Pledges are respected changing dynamically the priority of jobs in the queues in pre-defined time windows

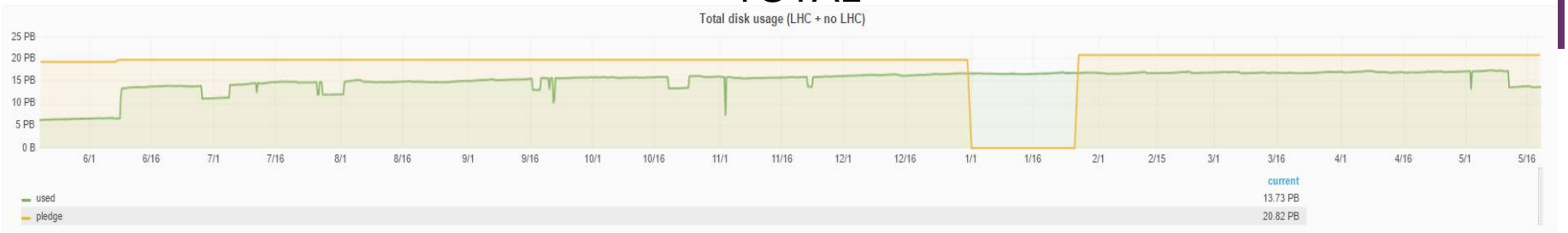
Short jobs are highly penalized

Short jobs that fail immediately are even more penalized

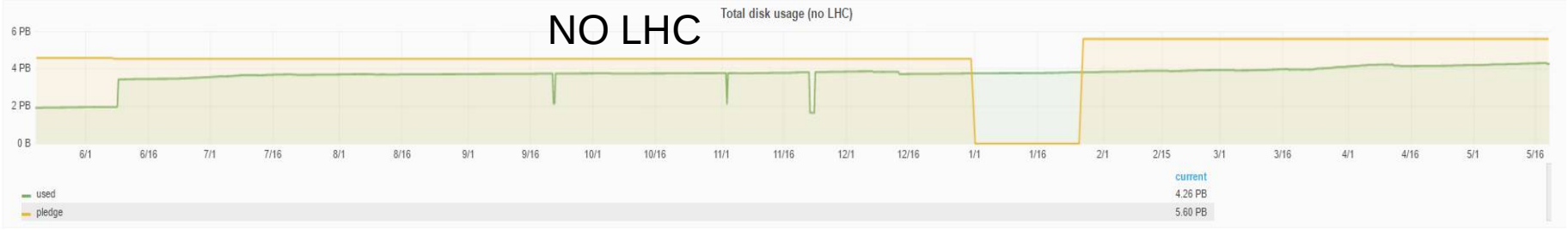
Disk Usage

Last 1 year

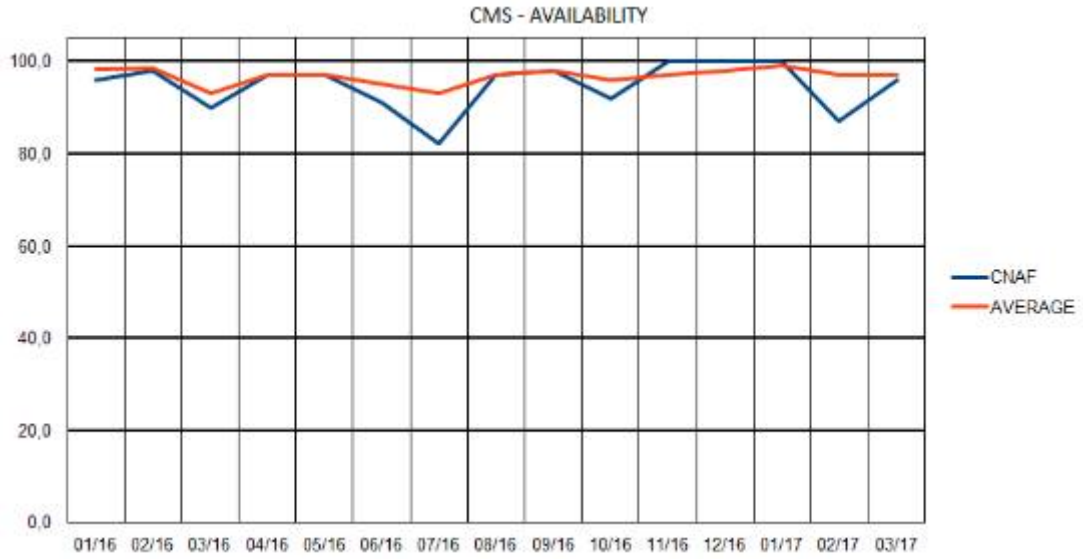
TOTAL

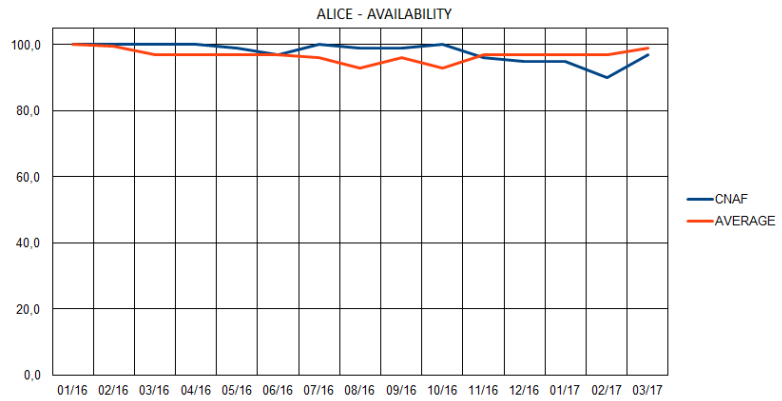


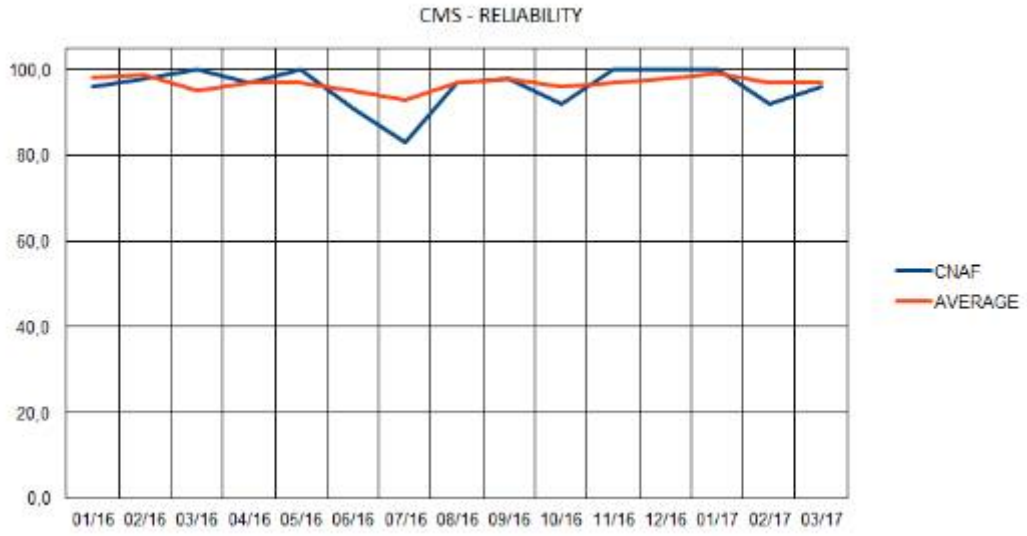
NO LHC

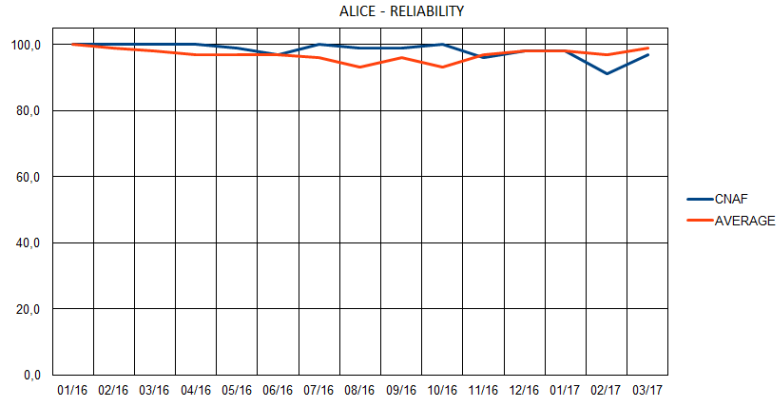


Tier1 Availability and Reliability

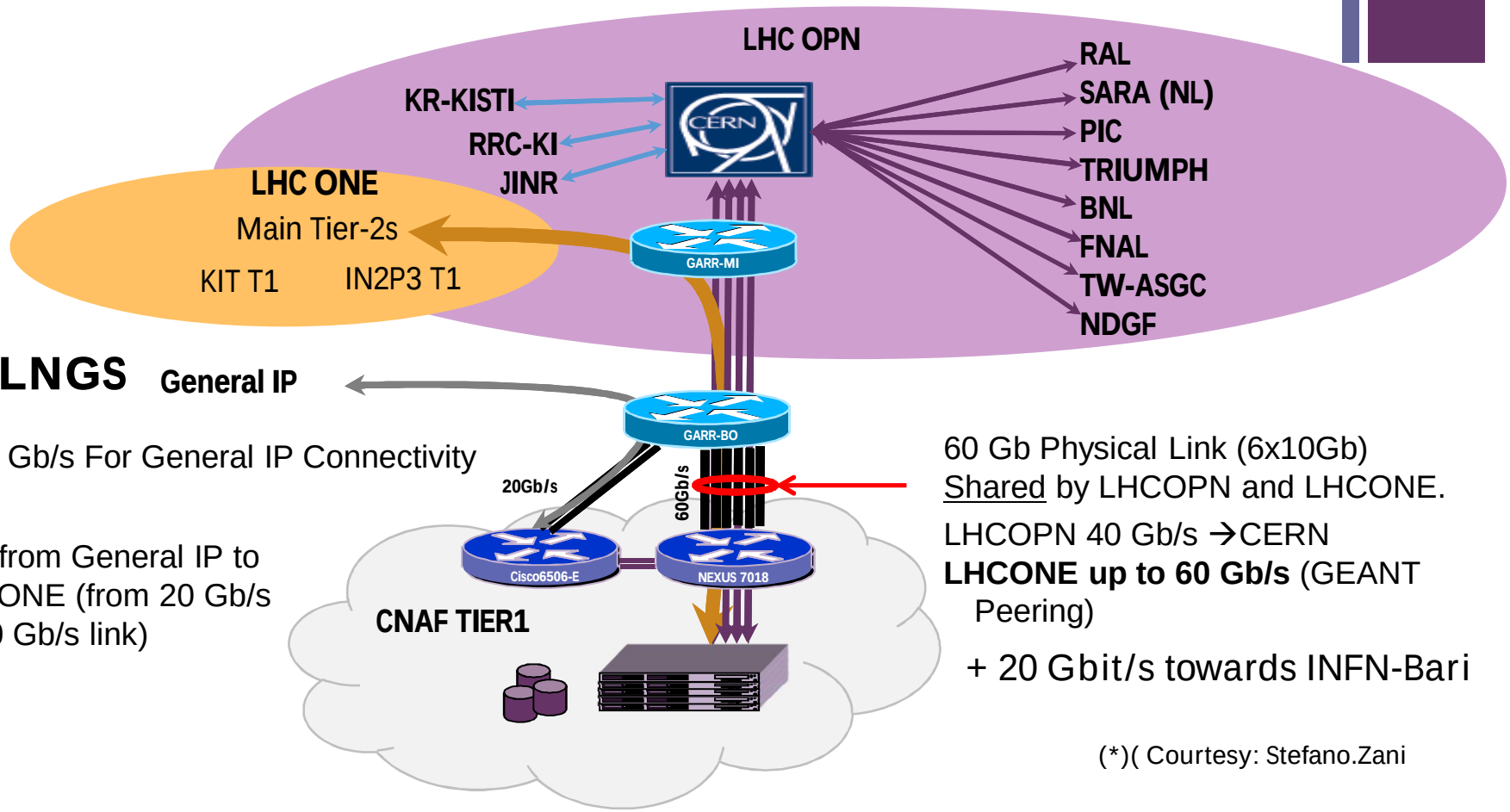








WAN@CNAF (Status)



20 Gb/s For General IP Connectivity

LNGS from General IP to LHCONE (from 20 Gb/s to 60 Gb/s link)

60 Gb Physical Link (6x10Gb) Shared by LHCOPN and LHCONE.

LHCOPN 40 Gb/s →CERN

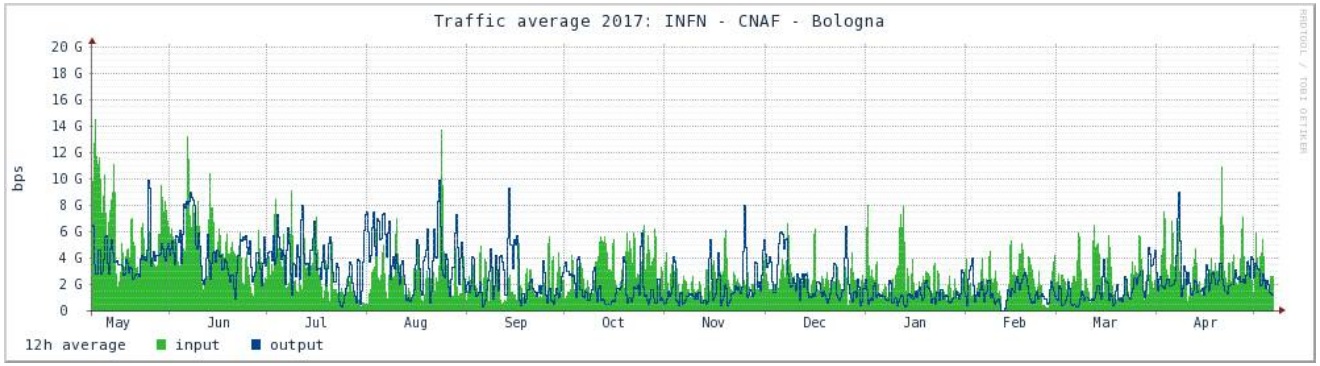
LHCONE up to 60 Gb/s (GEANT Peering)

+ 20 Gbit/s towards INFN-Bari

(*)(Courtesy: Stefano.Zani

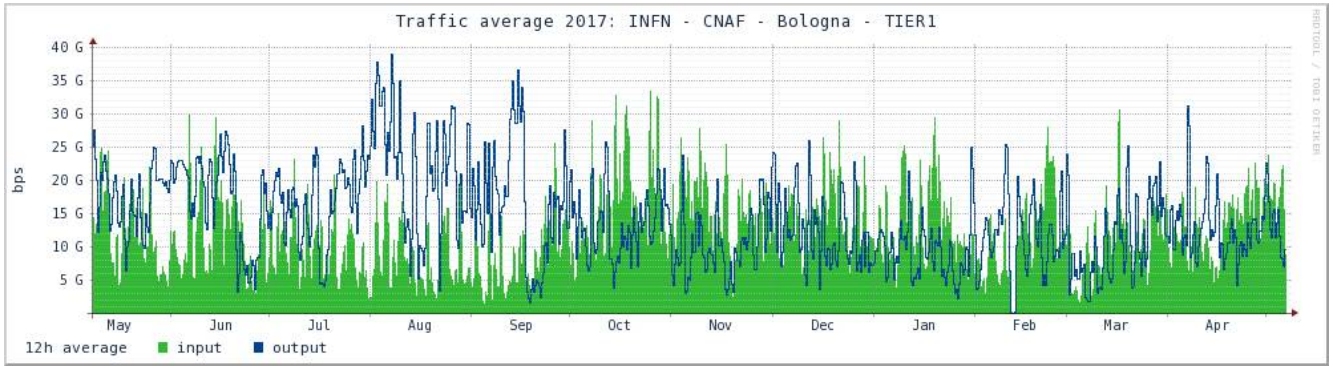
Current CNAF WAN Links usage
(IN and OUT measured from GARR POP)

GENERAL IP



AVG IN:2,7 Gb/s
AVG OUT:1,8 Gb/s
95 Perc IN: 6,54 Gb/s
95 Perc Out:4,6 Gb/s
Peak OUT: 19,5 Gb/s

LHC OPN + ONE (60 Gb/s)

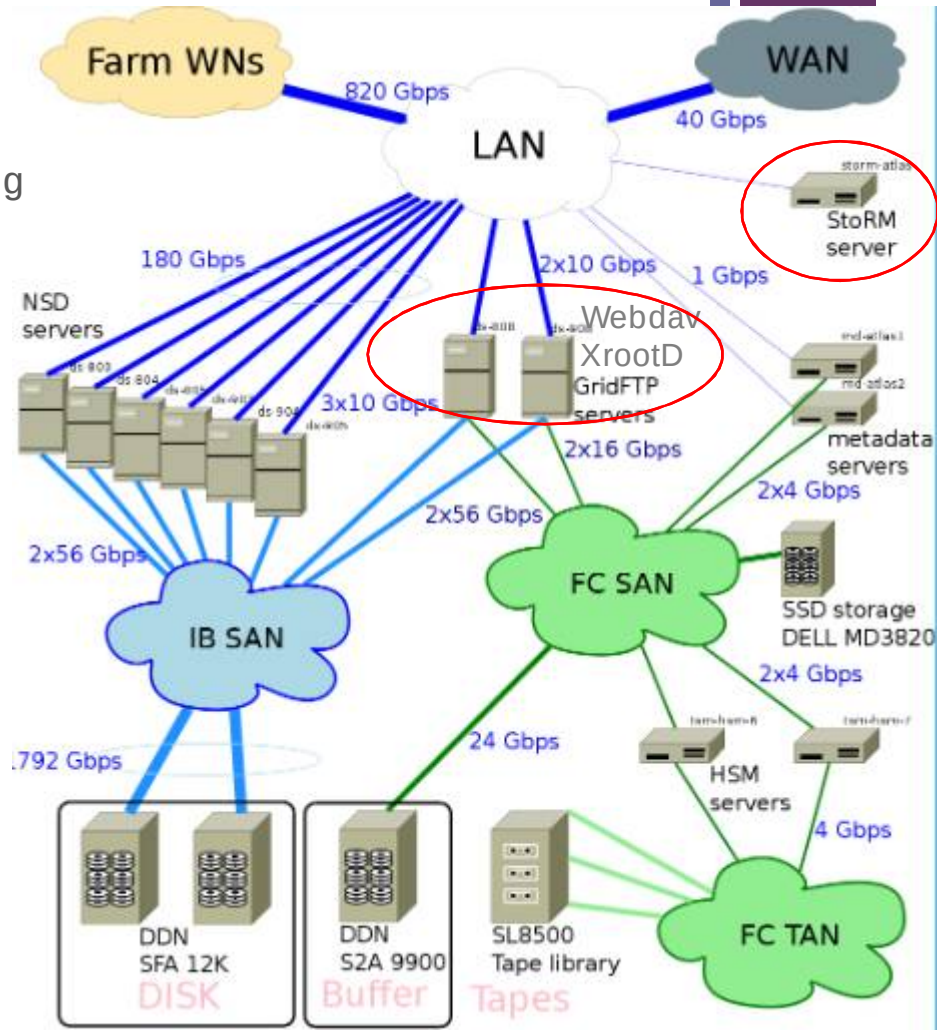


AVG IN:12,1 Gb/s
AVG OUT:11,2 Gb/s
95 Perc IN: 24,3 Gb/s
95 Perc Out:22,4 Gb/s
Peak OUT: 58,3 Gb/s

(*)(Courtesy: Stefano.Zani

Available data management and transfer services

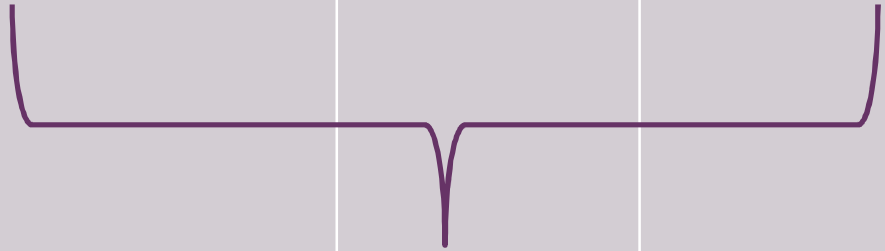
- StoRM (INFN SRM implementation)
  - Need a personal digital certificate
  - Need to belong to a VOMS managed Virtual Organization
  - Gridftp and Webdav transfer service
  - Checksum stored automatically
  - It is the only system to manage remotely the Bring Online from Tape
- Plain-Gridftp
  - Need only a personal certificate, no VO
  - No checksum stored automatically
- Dataclient
  - In-house development, it is a wrapper around gridftp
  - Lightweight client
  - Only need CNAF account
  - Checksum automatically stored
- Xrootd
  - Auth/AuthZ mechanism can be configured
- Custom user services can be installed
  - Even if we prefer to avoid



(*) Numbers refer to the 2016 ATLAS SAN/TAN

Data transfer services

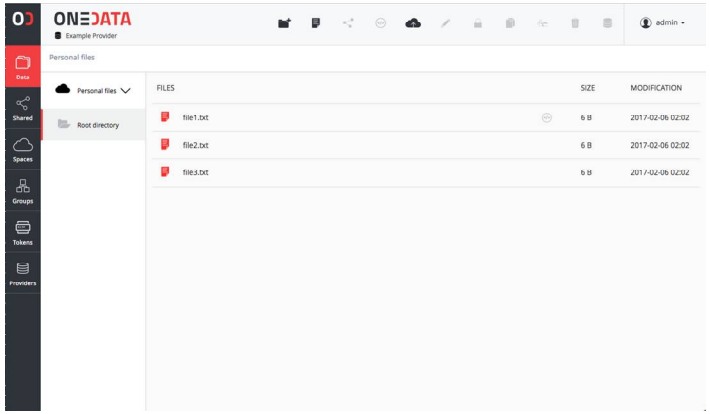
StoRM +gftp	StoRM +webdav	Gridftp plain	Dataclient	Xrootd	Custom
4 LHC Virgo AMS AGATA ARGO AUGER BELLE CTA GERDA GLAST ICARUS ILDG MAGIC NA62 PAMELA THEOPHYS XENON COMPASS PADME JUNO	LHCb ATLAS BELLE	DARKSIDE DAMPE BOREXINO KLOE	KM3 CUORE CUPID NEWCHIM	4 LHC DAMPE	VIRGO/LIGO LDR LIMADOU …

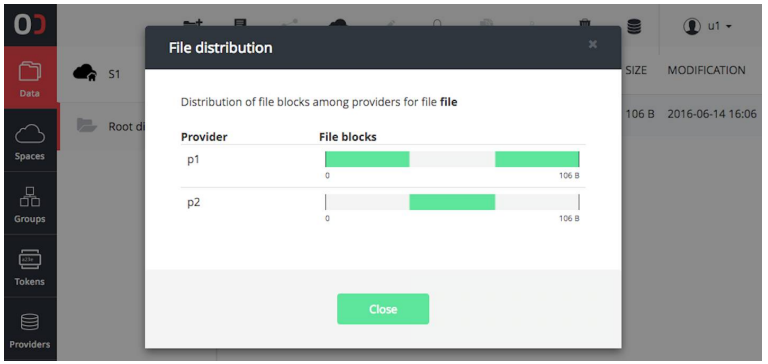


Unification??

Data management services under evaluation

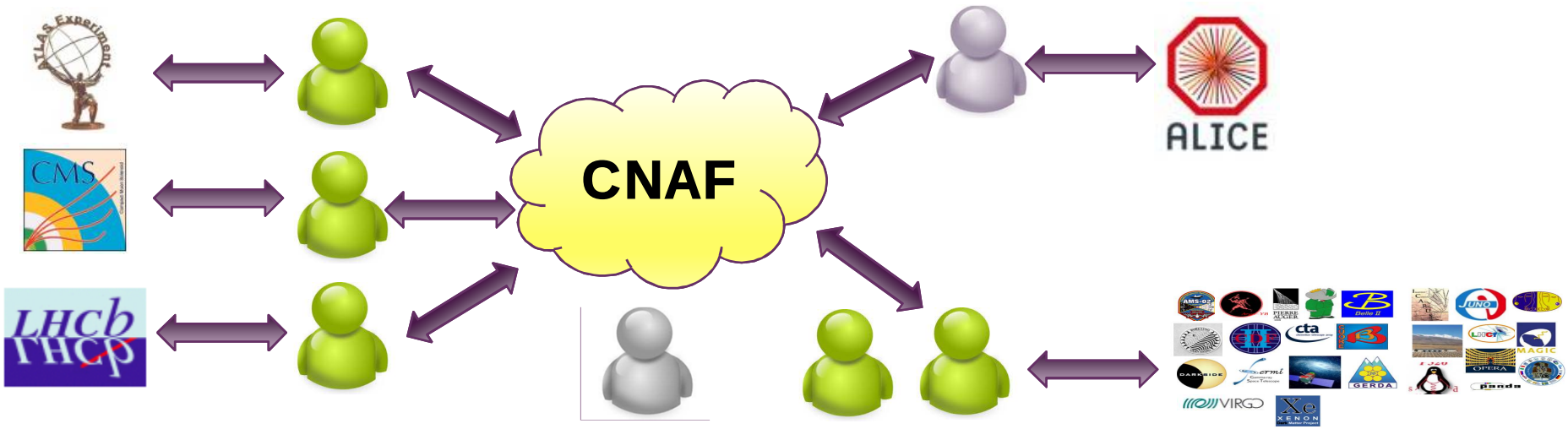
- ONEDATA
  - Developed by CYFRONET (PL)
  - INDIGO project data management solution
  - WEB interface + CLI
  - Caching
  - Automatic replication
  - Advanced authentication
  - Metadata Management
- OwnCloud/NextCloud + storm-webdav or storm-gridftp
  - Add a smart web interface and clients to performant DM services
  - Need developments and non trivial effort





User support @ Tier1

- 5 group members (post-docs)
  - 3 group members, one per experiment, dedicated to ATLAS, CMS, LHCb
  - 2 group members dedicated to all the other experiments
- 1 close external collaboration for ALICE
- 1 group coordinator from the Tier1 staff



User Support activities

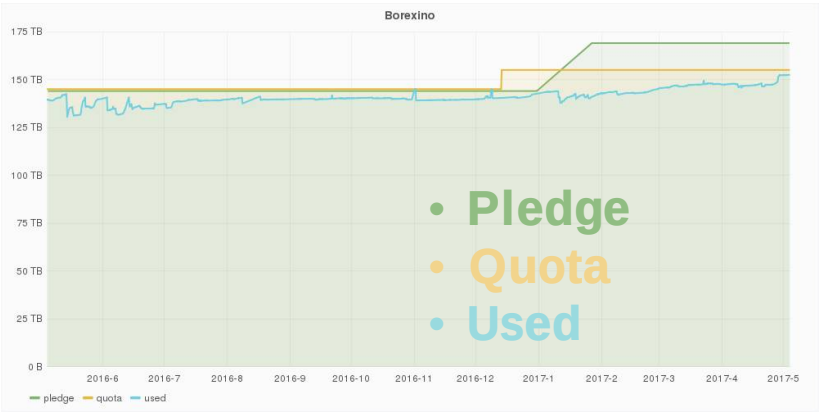
- The group acts as a first level of support for the users
  - Incident initial analysis and escalation if needed
  - Provides information to access and use the data center
  - Takes care of communications between users and CNAF operations
  - Tracks middleware bugs if needed
  - Reproduces problematic situations
    - Can create proxy for all VOs or belong to local account groups
- Provides consultancy to users for computing models creation
- Collects and tracks user requirements towards the datacenter
  - Including tracking of extrapledges requests
- Represent INFN-Tier1 in WLCG coordination daily meeting (Run Coordinator)

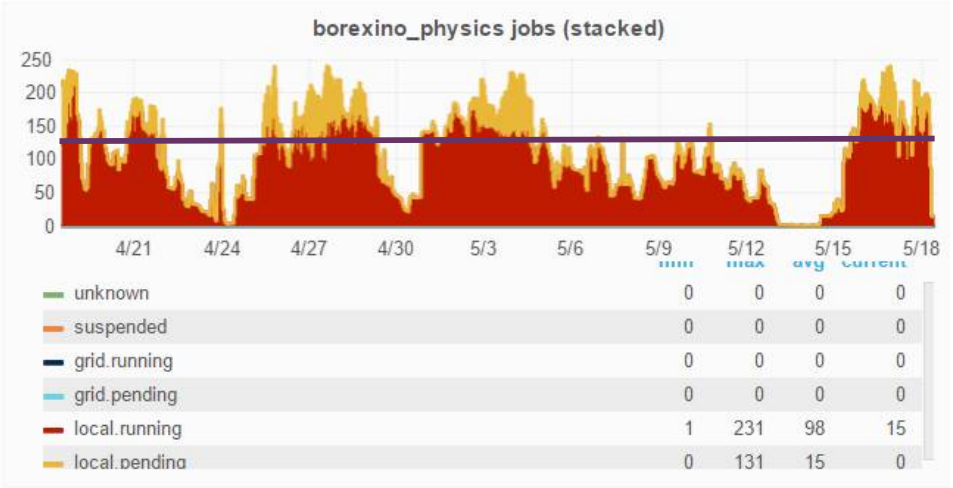
Contacts and Coordination Meetings

- GGUS Ticketing system: https://ggus.eu
- Mailing lists: user-support<at>cnaf.infn.it, hpc-support<at>cnaf.infn.it
- Monitoring system: https://mon-tier1.cr.cnaf.infn.it/
- FAQs and UserGuide
  - https://www.cnaf.infn.it/utenti-faq/
  - https://www.cnaf.infn.it/wp-content/uploads/2016/12/tier1-user-guide-v7.pdf
- Monthly CDG meetings
  - https://agenda.cnaf.infn.it/categoryDisplay.py?categId=5
  - No meeting, no problems
  - Not enough: we will organized ad-hoc meeting to define the desiderata and complaints

Borexino

- At present, **the whole Borexino data statistics** and the user areas for physics studies are **hosted at CNAF**.
- Borexino standard **data taking requires a disk space increase of about 10 TB/year** while a **complete MC simulation requires about 7 TB/DAQ year**.
- **Raw data** are hosted on disk for posix access during analysis. A backup copy is hosted on CNAF tape and at LNGS.





CUORE

- Raw data are transferred from the DAQ computers to the permanent storage area at the end of each run also to CNAF. In CUORE **about 20 TB/y of raw data** are expected.

- In view of the start of the CUORE data taking, **since 2014 a transition phase** has started to move the CUORE analysis and simulation framework to CNAF. The framework is now installed and it is ready for being used.

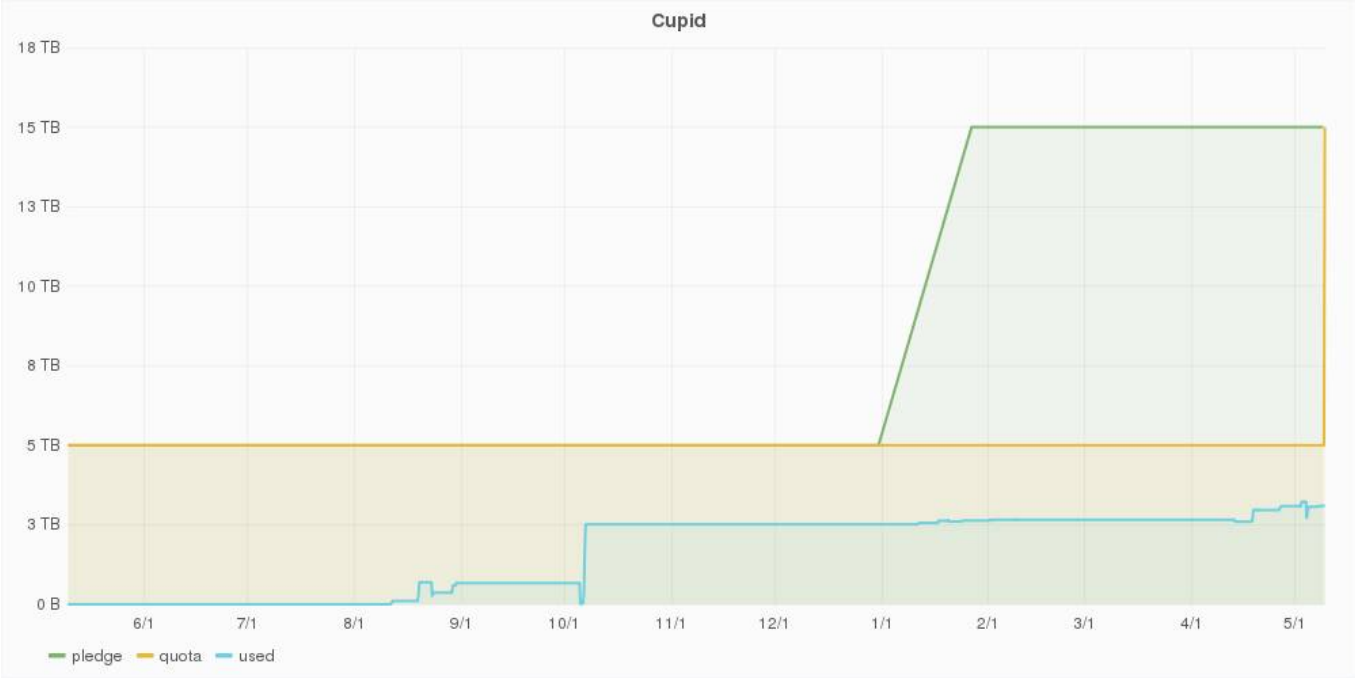
- **In 2015 most of the CUORE Monte Carlo simulations were run at CNAF**, and some tests of the data analysis framework were performed using mock-up data.

Data transferred using Dataclient

- **Pledge**
- **Quota**
- **Used**

CUPID-0

- Data already flowing to CNAF, including raw data
- Data transferred using Dataclient



- **Pledge**
- **Quota**
- **Used**

DarkSide-50

- Darkside-50 raw data:
- Production rate: 10TB/month
- Temporarely stored at LNGS before being transferred to CNAF
- Transferred to FNAL via CNAF
- Transfers via gridftp

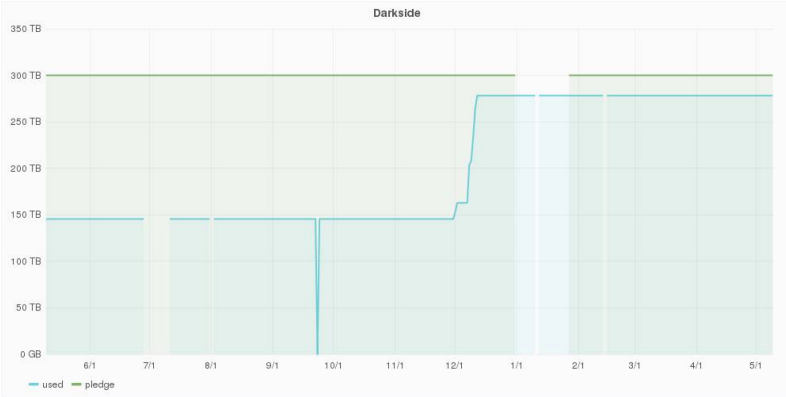
Reco data:

- Reconstruction at FNAL and LNGS
- Transferred to CNAF from both site (few TB/year)

MC:

- Produced in several computing center (including CNAF)

- **Pledge**
- **Quota**
- **Used**



Gerda

- Gerda uses the storage resources of CNAF, accessing the GRID through the virtual organization (VO) . **Currently the amount of data stored at CNAF is around 10 TB**, mainly data from Phase I, which is saved and registered in the Gerda catalogue.

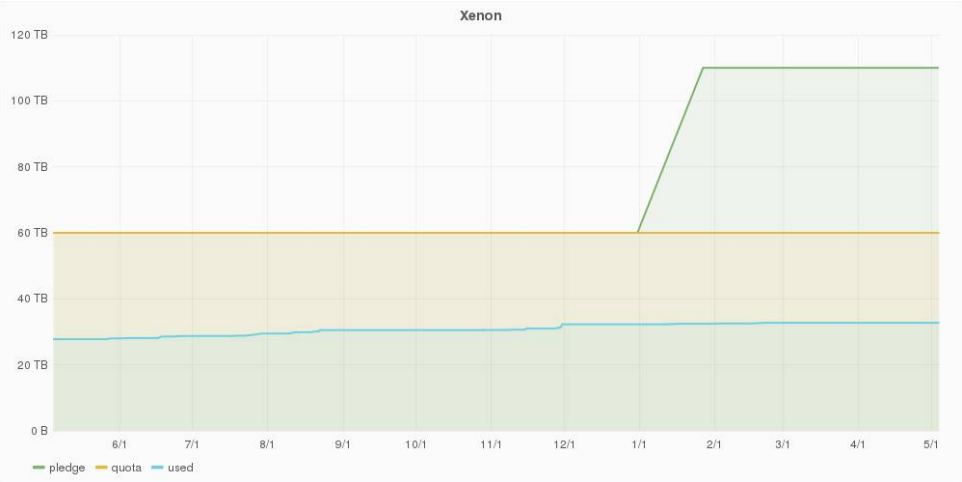
- Data is stored at the experimental site in INFN's Gran Sasso National Laboratory (LNGS) cluster. The policy of the Gerda collaboration requires three backup copies of the raw data (Italy, Germany and Russia). **CNAF is the Italian backup storage center,** and raw data is transferred from LNGS to Bologna in specific campaigns of data exchange

- The Gerda collaboration uses the CPU provided by CNAF to process higher level data reconstruction (TierX), and to perform dedicated user analysis. Besides, CPU is mainly used to perform MC simulations.

XENON

- CNAF mainly access trough VO and Grid services
- Raw data not stored at CNAF
  - plans to change
- Need to increase interaction

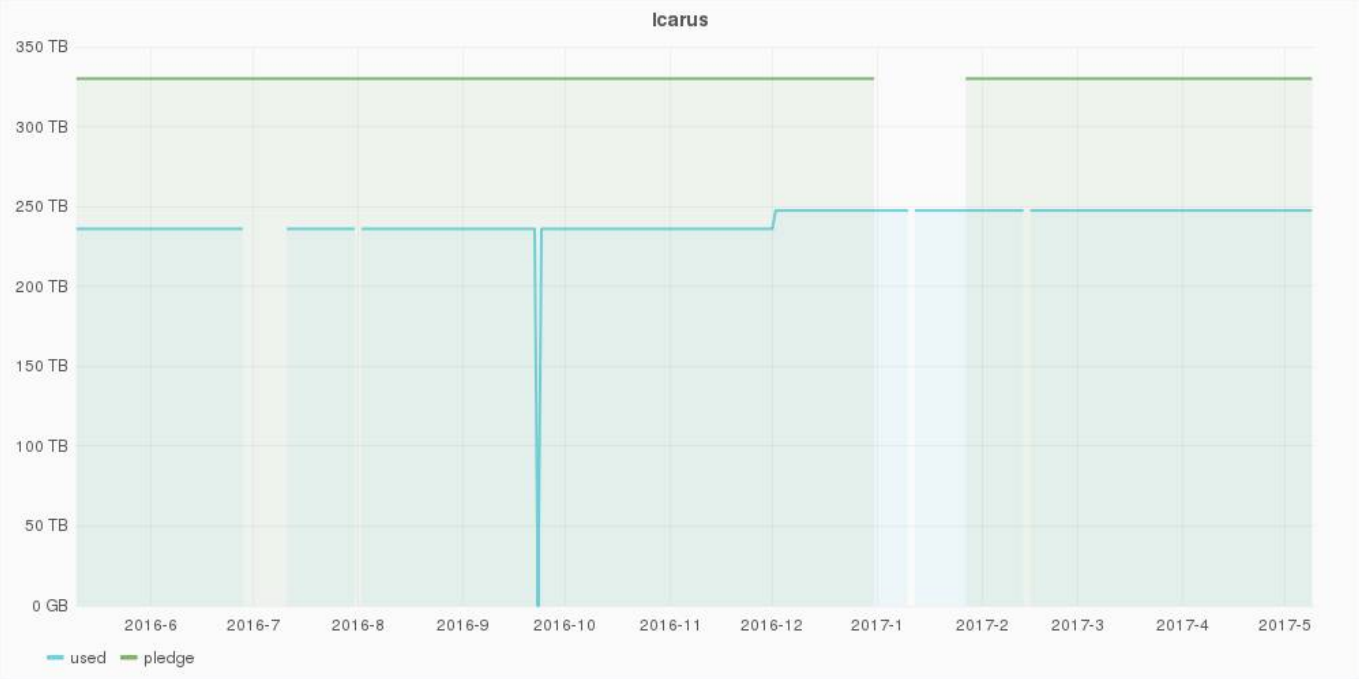
Disk Usage



- **Pledge**
- **Quota**
- **Used**

ICARUS

- Raw data copy via StoRM
- Only tape used + powerful UI for local analyses when needed
- Huge recall campaigns handled manually
  - Need CPU to analyze them (not pledged)



LNGS T0@CNAF ?

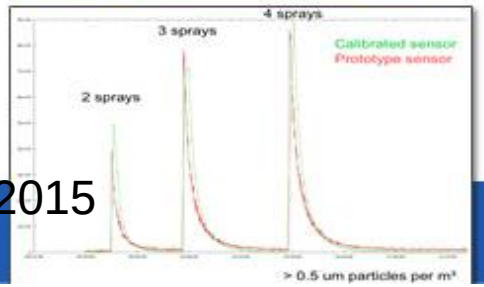
- All LNGS experiments@CNAF already store a backup copy of raw data at T1
  - With the XENON exception
- It seems that there are no network bandwidth issues
- Data Management services available to reliably store data on tape, disk or both
- But…in any case there should be another copy outside CNAF

LNGS T0@CNAF ?

- All LNGS experiment (supported by CNAF) already store a backup copy at T1
  - With the XENC
- No network bar
- Data Managem
  disk or both
- But…in any cas

Dust incident

- Identified 13 tapes in one library affected by concrete or foam particles
- Isolated incident by verifying all other tapes in the building
- Recovered 94% files with custom low-level tools and vendor recovery; 113 files lost
- Fruitful exchanges with other tape sites on CC protective measures (access and activity restrictions, special clothing, air filters etc)
- Library cleaning by specialist company envisaged
- Prototyped a dust sensor to be installed inside libraries, using cheap commodity components, achieving industrial precision and reaction time

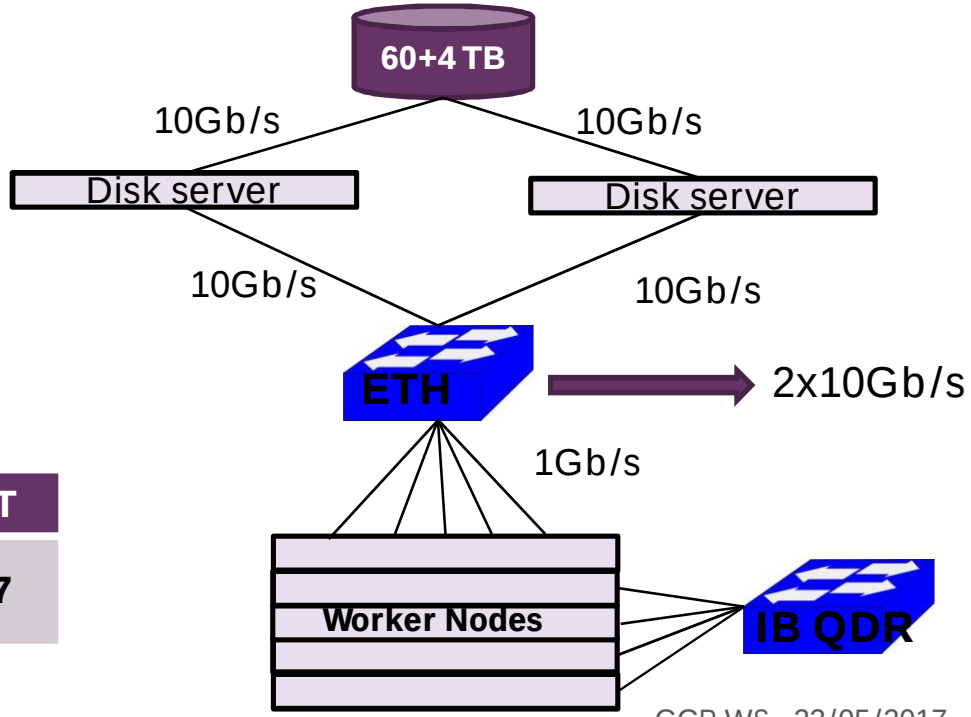


Eric Cano (CERN) @ CHEP2015
14/4/2015

2014 HPC Cluster

- 27 Worker Nodes
  - CPU: 904HT cores
    - 640 HT cores E5-2640
    - 48 HT cores X5650
    - 48 HT cores E5-2620
    - 168 HT cores E5-2683v3
  - 15 GPUs:
    - 8 Tesla K40
    - 7 Tesla K20
    - 2x(4GRID K1)
  - 2 MICs:
    - 2 x Xeon Phi 5100

- Dedicated STORAGE
  - 2 disks server
  - 60 TB shared disk space
  - 4 TB shared home
- Infiniband interconnect (QDR)
- Ethernet interconnect
  - 48x1Gb/s + 8x10Gb/s



	CPU	GPU	MIC	TOT
TFLOPS (DP - PEAK)	**6.5**	**19.2**	**2.0**	**27.7**

2014 Cluster - 2

- Main users:
  - INFN-BO, UNIBO, INAF theoretical physicists
- New users can obtain access
  - AMS for GPU based applications development and testing
  - BOREXINO
  - CERN accelerators theoretical physicists
- Cannot be further expanded
  - Physical spaces on racks
  - Storage performance
  - IB switch ports

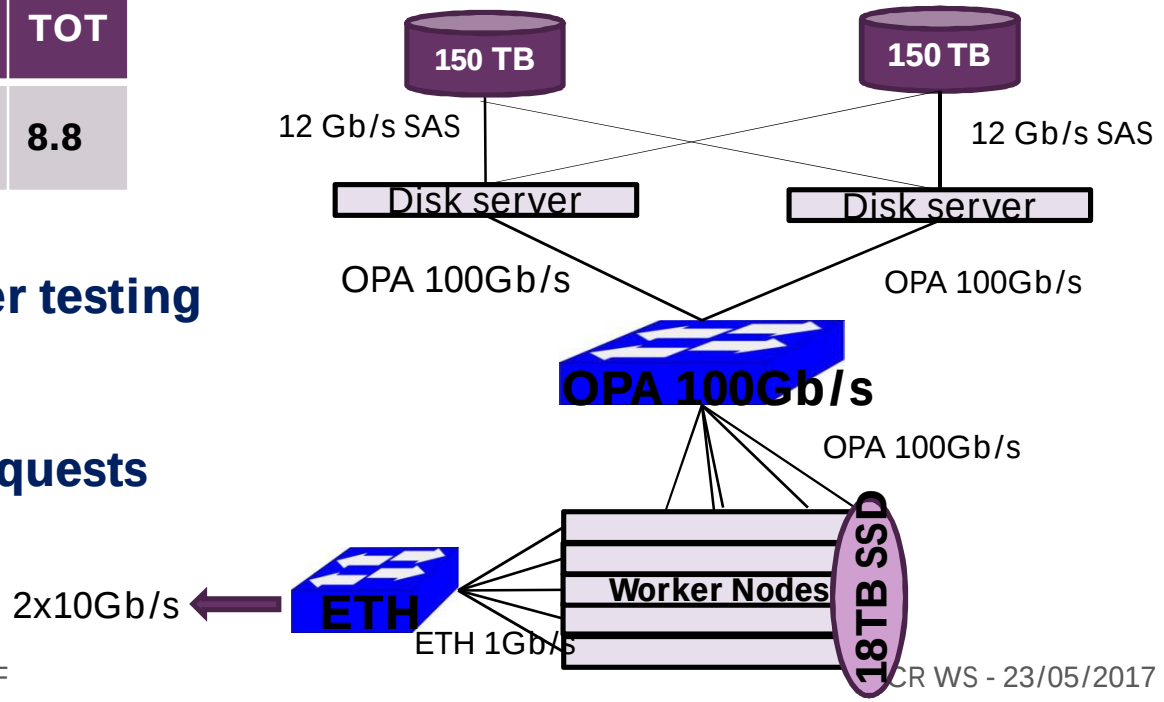
2017 HPC Cluster

- 12 Worker Nodes
  - CPU: 768 HT cores
    - Dual E5-2683v4 @2.1 GHz (16 core)
  - 1 KNL node:
    - 64 core (256HT core)

	CPU	GPU	MIC	TOT
TFLOPS (DP - PEAK)	**6.5**	**0**	**2.6**	**8.8**

- Dedicated STORAGE
  - 2 disk servers + 2 JBOD
  - 300 TB shared disk space
  - (150 with replica 2)
  - 18 TB SSD based file system using 1 SSD on each WN – used for home dirs
- OmniPath interconnect (100Gbit/s)
- Ethernet interconnect
  - 48x1Gb/s + 4x10Gb/s

- WN installed, storage under testing
- Will be used by CERN only
- Can be expanded
  - **i.e. LSPE/EUCLID requests**



Conclusion

- The collaboration between CNAF and CSN2 experiment is proceeding smoothly
  - However we should increase the frequency of interactions with the user support group and at the CDG
    - collect feedback and new requests
    - computing models checkpoints
- Data management services and IP connectivity make it possible to have the T0 at CNAF for some of the LNGS experiments
- Small HPC clusters are avalaible for developments, tests and small productionPiccoli cluster hpc sono disponibili per prove e sviluppo
  - It is possible to expand the newer HPC cluster to cope with new requests if needed (i.e. LSPE, EUCLID)