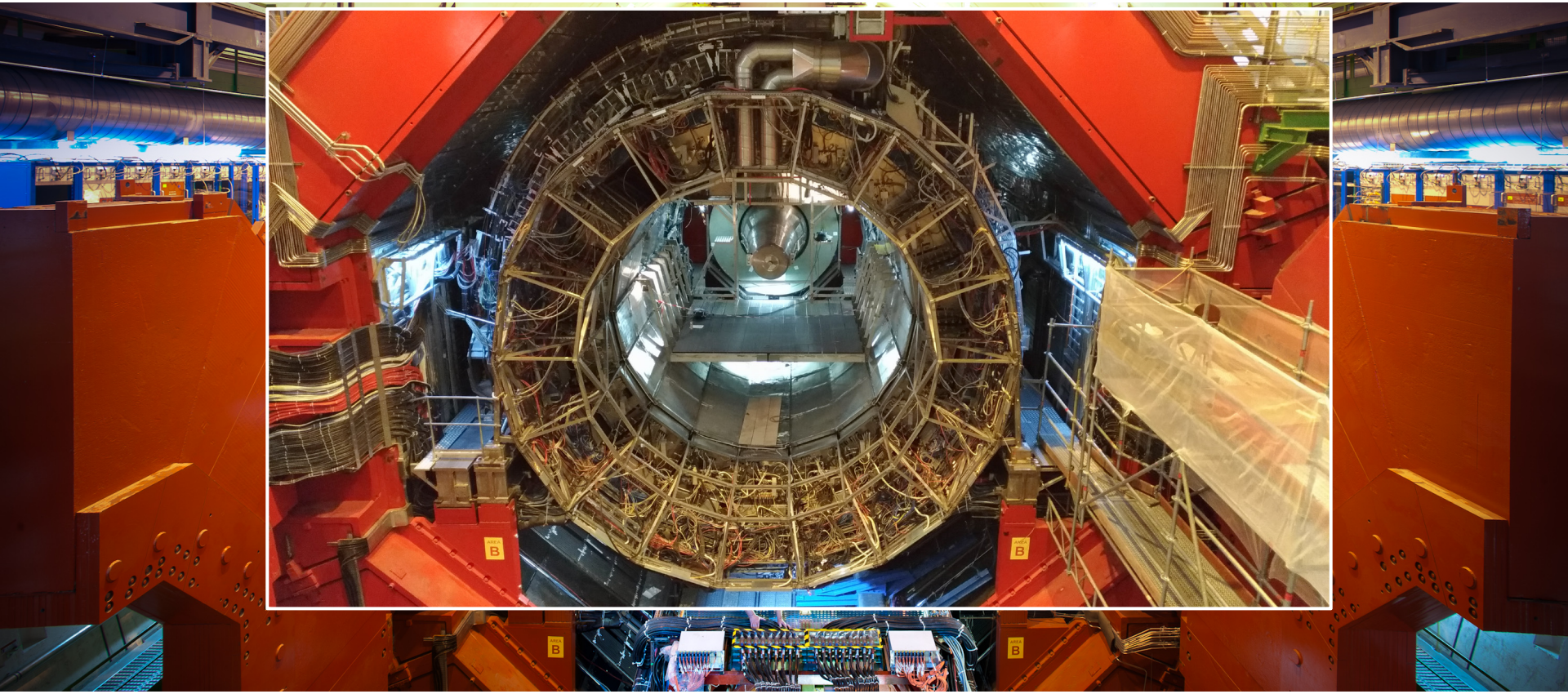
Operations in Italy and plans for INFN sites

Stefano Piano – INFN sez. Trieste
Stefano Bagnasco, Francesco Noferini



Operations Organization

- ALICE-IT Computing Coordination:
  - Stefano Piano
- Deputy:
  - Stefano Bagnasco
- Tier-2 Operations Coordination:
  - Francesco Noferini
  - Monthly phone conference for coordination and performance monitoring
  - Yearly face-to-face workshop (2012 @ CT, 2013 @ TS, 2014 @ LNF, 2015 @ CNAF, 2016 @ PD, 2017 @ BA, 2018 @ TO)
- Responsible for Tier-2:
  - D. Elia (BA), G. Andronico (CT), M. Biasotto (PD-LNL), S. Lusso (TO)
- Operations at Tier-1: Francesco Noferini and Nicolò Jacazio
- Monthly Tier-1 Management Board at CNAF

Outline

- Available and expected resources in 2019

- The Tier-1 at INFN-CNAF
- Performance of the Italian sites
- Status of Italian Tier-2 sites

- R&D activities

Resource available for ALICE

- **Tier-1 at CNAF, Bologna**
  - Shared with other LHC experiments and a large number of others
- **4 official Tier-2 centers**
  - Bari, Catania, Padova-LNL and Torino
  - “Official” means “directly funded by INFN according to plans and pledges”
- **Additional (minor) center**
  - Trieste (Cagliari dismissed last year)
  - Local resources, different creative funding, mostly out of pledge
- Projects providing resources in the ALICE INFN sites over the last years
  - **ReCaS (BA and CT)**, sizeable contribution to 2014 and 2015 pledges
  - **PON IBISCO (BA and CT),** contribution from 2019 to 2021 pledges

Networking
Tier-1 and Tier-2's

- Tier-1 at 100 Gbps LHCONE + 100 Gbps LHCOPN
  - Dedicated link with CINECA: 500 Gbps ($\Rightarrow$ 1.2 Tbps) VPN ready
- All Tier-2's connected to LHCONE with at least 10 Gbps
  - Through GARR-X
  - All Tier-2's easily upgradable to 40 Gbps
  - Bari, Catania, Padova-LNL already at 20 Gbps
  - Bari dedicated network link: 20Gbps to Naples and 20 Gbps to CNAF
- IPV6
  - All INFN Tier-2 sites (but Trieste) have being acting coordinately (IPv6 link and DNS server with IPv6 by April 2018, storage configured with dual-stack by June 2018, last site on December 2018)

Computing resources at INFN
Current Status

- Tier-1 2019 pledge already available:
  - CPU: 60600 HS06 already installed
  - DISK: 6148 TB SE - 448 TAPE buffer = 5700 TB (5184 TiB) already deployed
  - TAPE buffer: 448 TB + 608 TB on loan (total buffer 961 TiB) to process 2018 Pb-Pb data (from December 2018 to April 2019: 1.3 PiB)
  - TAPE: 11687 TB almost full (11506 TB used)
- Tier-2 2019 pledge: **CPU**: 70845 HS06 / **DISK**: 6659 TB

	Bari	Catania	LNL-Padova	Torino	Total
HS06	16043	15125	14393	15235	60796
TB	1504	1204	1210	1396	5314

  - Still below the 2018 pledge (cpu almost there, disk -20%)
  - CPU price higher than expected (monetized), compensation in 2019!
  - Two long storage procurement tenders, lasted 18 months and 9 months!

Computing resources at INFN
Expected @ Tier-2

		Bari	Catania	Padova-LNL	Torino	Totale
10/05/2019	HS06	16043	15125	14393	15235	60796
2019 procurement	HS06	2000	2900	3700	2800	11400
2019	**HS08**	**18043**	**18025**	**18093**	**18035**	**72196**
10/05/2019	TB	1504	1204	1210	1396	5314
2017 procurement	TB	-	350	300	-	650
2018 procurement	TB	250	136	185	177	748
2019	**TB**	**1754**	**1690**	**1695**	**1573**	**6712**

Pledge
2018
61050 HS06
2019
70845 HS06

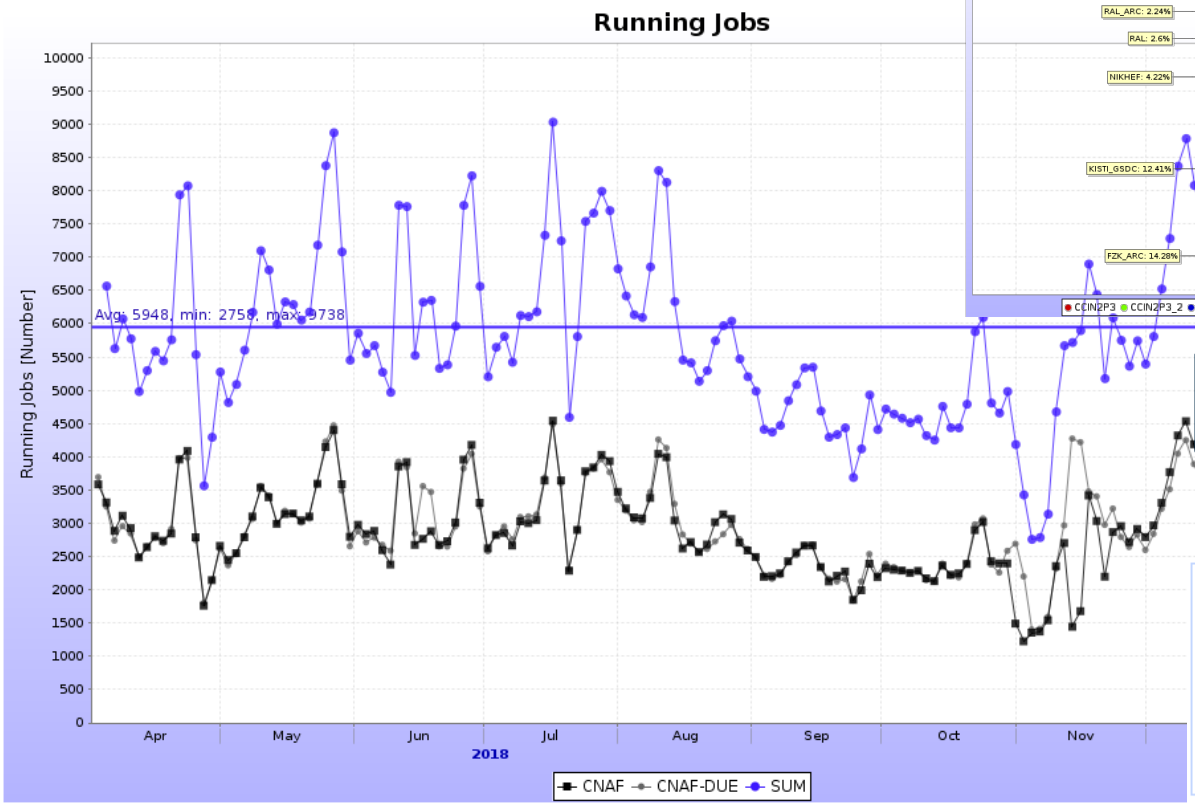
Pledge
2018 - 2019
6659 TB

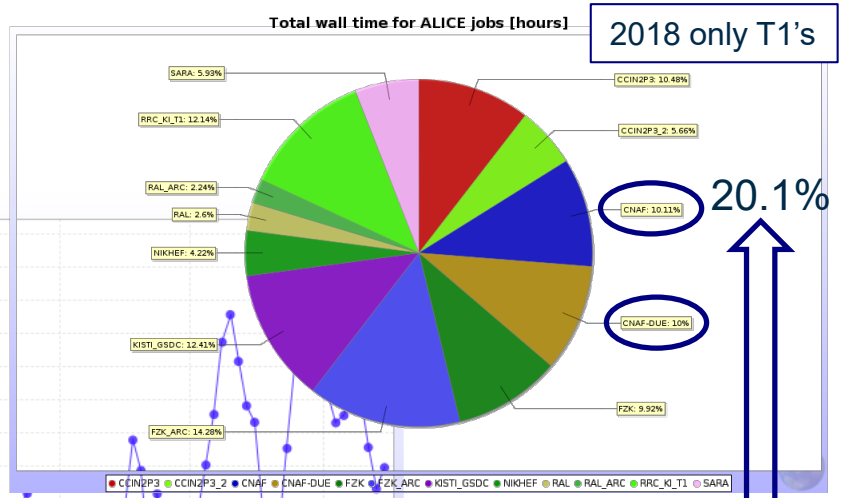
CPU: purchase on going, deployment expected by July

Storage: CT, PD-LNL and TO delivered and installed, acceptance tests just finished
BA and TO 2017 procurement already deployed, BA 2018 ETA by July

Performance of the Italian sites: CNAF

CNAF 2018 pledge: 52 kHS06 CPU, 5140 TB disk and 13530 TB tape storage





2018 only T1's

20.1%

CNAF ranks second of the ALICE Tier1 sites despite the flooding incident !

Jobs efficiency (cpu time / wall time)

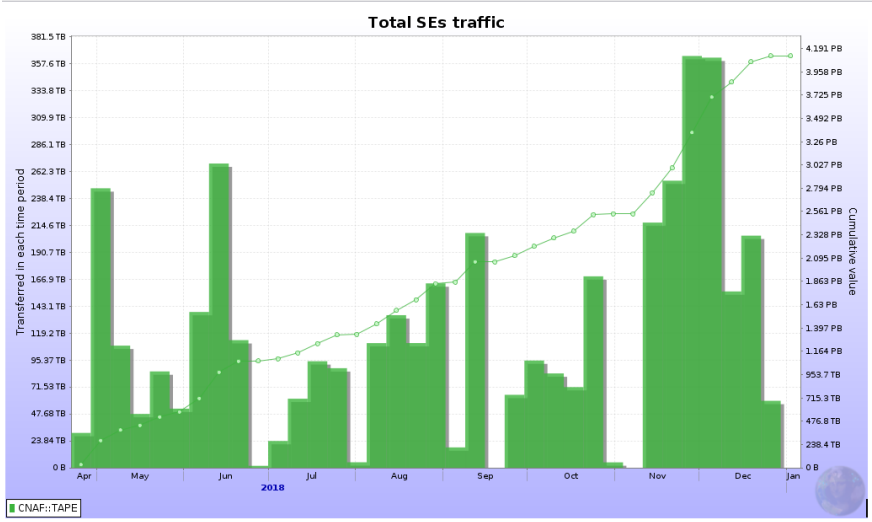
	Series	Last value	Min	Avg	Max
1.	CNAF	90.59	4.119	85.17	100
2.	CNAF-DUE	90.97	0	85.2	100
	Total	90.78		85.18	

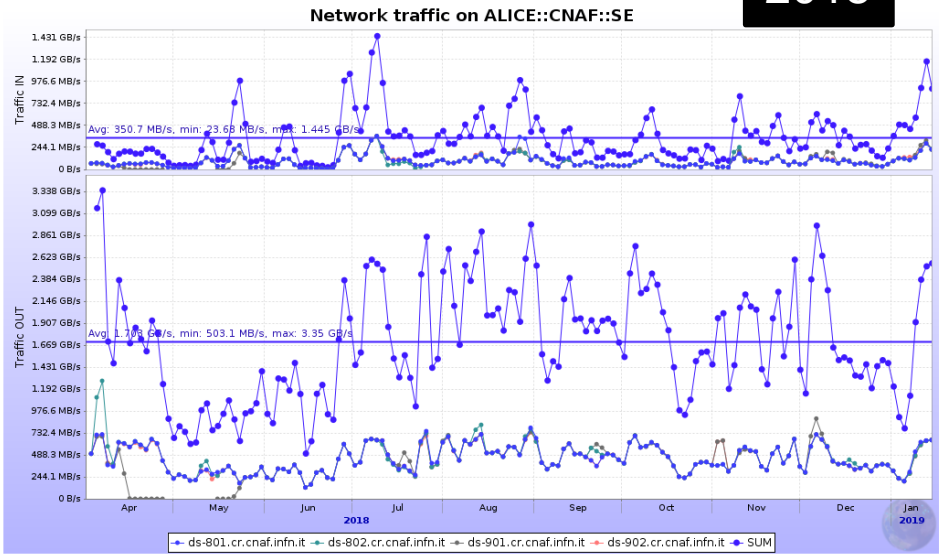
Performance of the Italian sites: CNAF

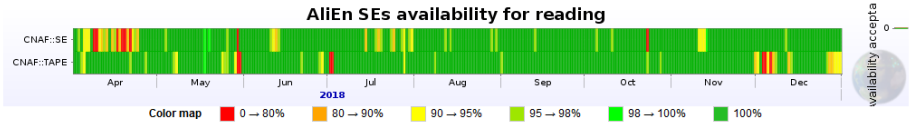
2018

The excellent FS performances allow to analyze data from SE with high efficiency:
average throughput of about 2 GB/s
peak throughput > 4 GB/s

Network connectivity: the 100 Gb/s of the WAN links allowing ALICE to sustain a total traffic up to 1.5 GB/s >> 600 MB/s from buffer to tape (360 TB of raw data per week from Tier0 to CNAF)







Statistics

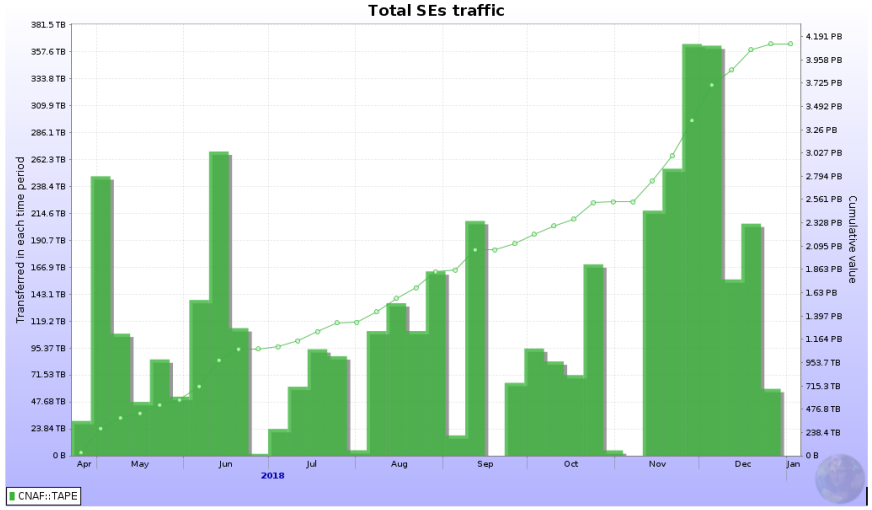
Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
CNAF::SE	01 Apr 2018 00:05	31 Dec 2018 23:35	6429	141	97.85%	97.92%
CNAF::TAPE	01 Apr 2018 00:06	01 Jan 2019 00:18	6460	122	98.15%	98.18%

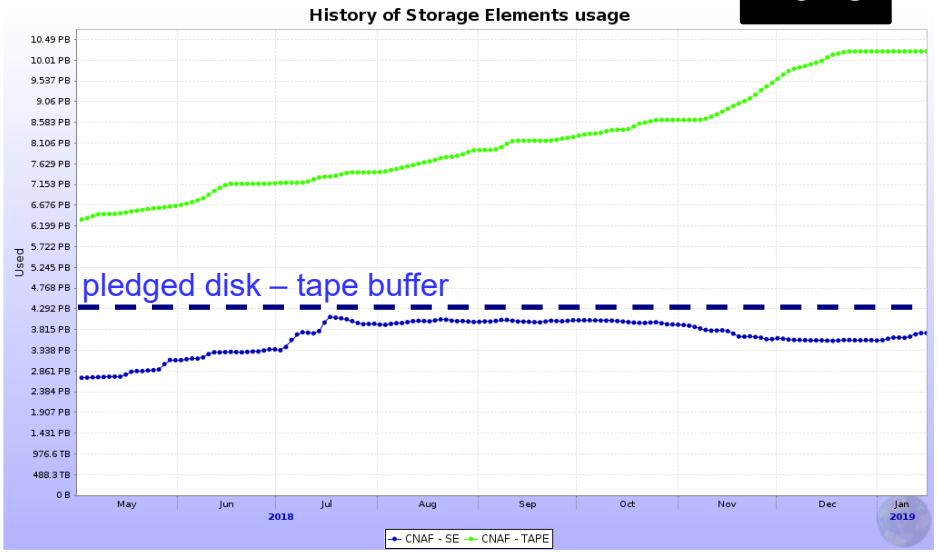
Performance of the Italian sites: CNAF

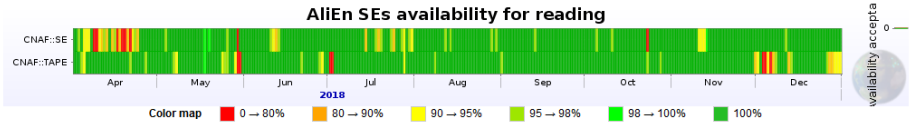
2018

The excellent FS performances allow to analyze data from SE with high efficiency:
average throughput of about 2 GB/s
peak throughput > 4 GB/s

Network connectivity: the 100 Gb/s of the WAN links allowing ALICE to sustain a total traffic up to 1.5 GB/s >> 600 MB/s from buffer to tape (360 TB of raw data per week from Tier0 to CNAF)





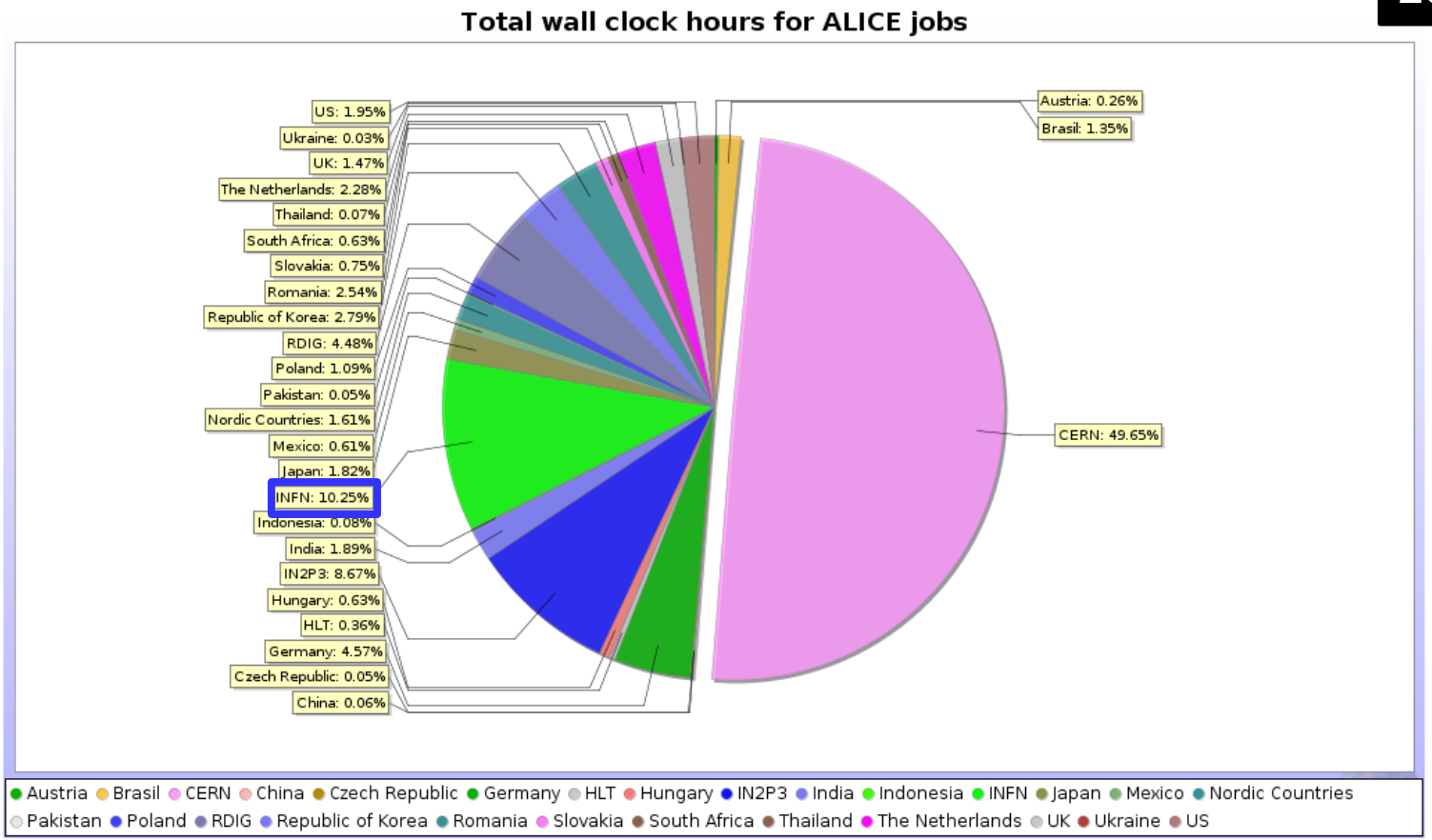


Statistics

Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
CNAF::SE	01 Apr 2018 00:05	31 Dec 2018 23:35	6429	141	97.85%	97.92%
CNAF::TAPE	01 Apr 2018 00:06	01 Jan 2019 00:18	6460	122	98.15%	98.18%

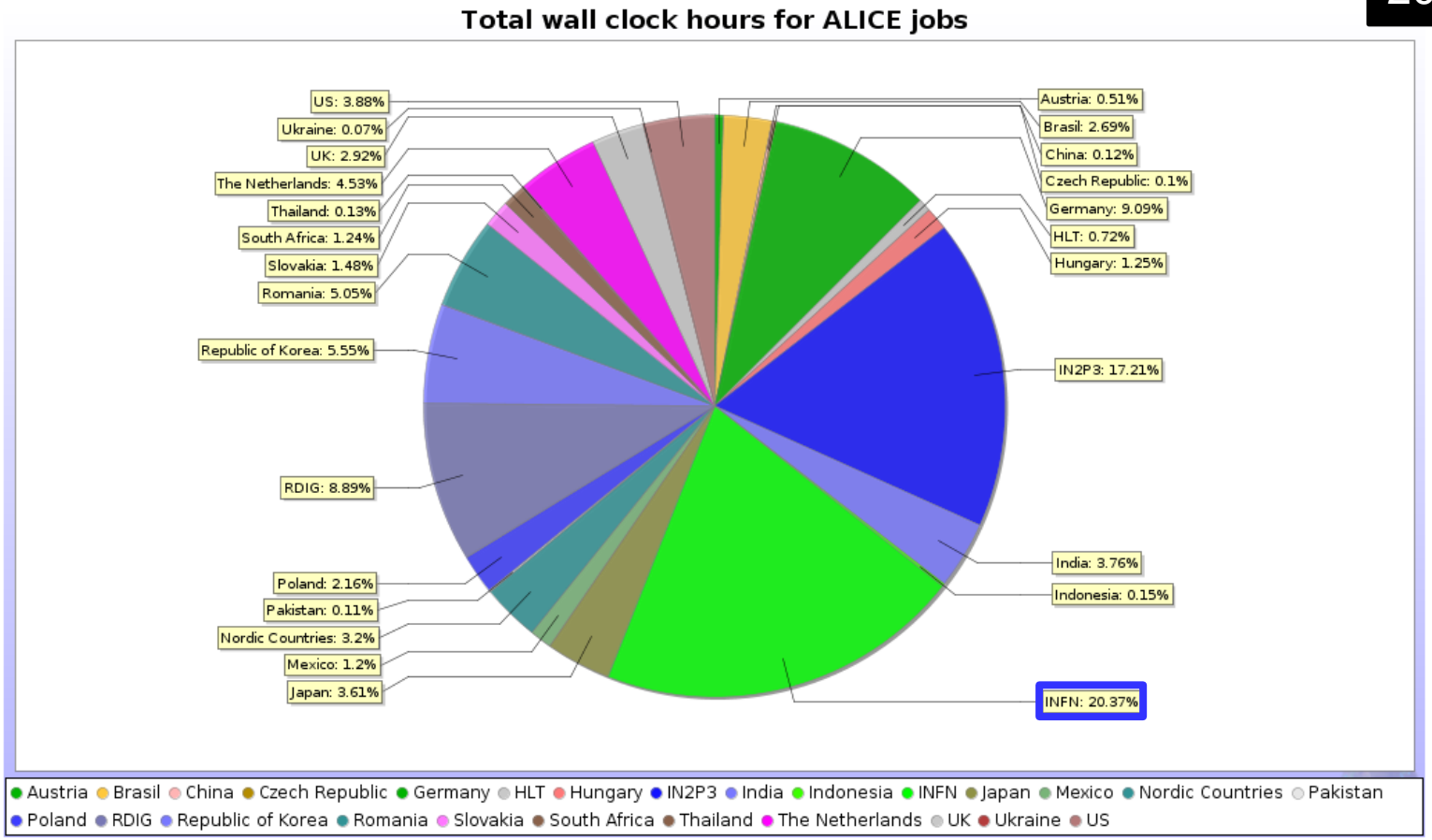
Performance of the Italian sites

2018

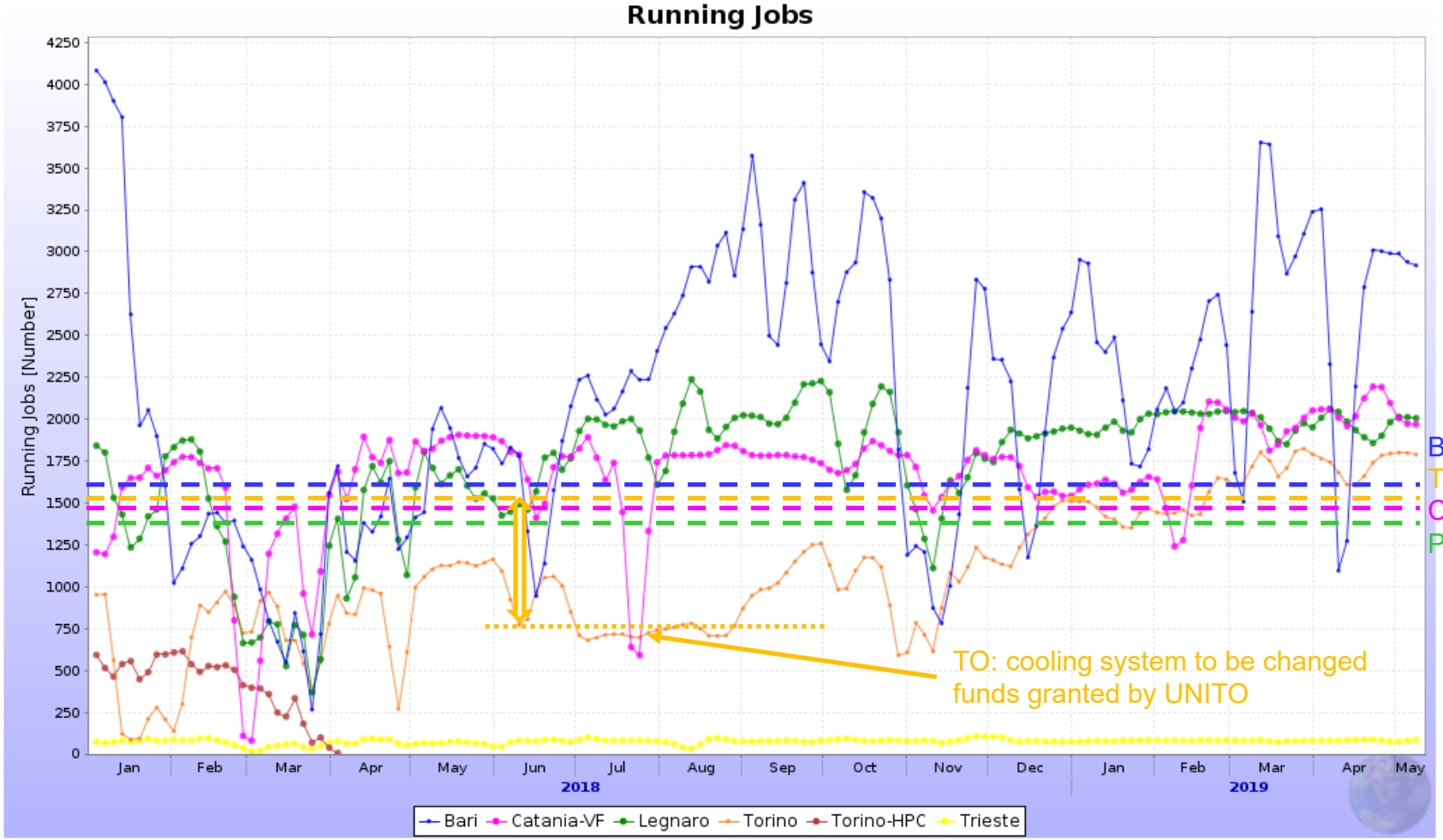


Performance of the Italian sites

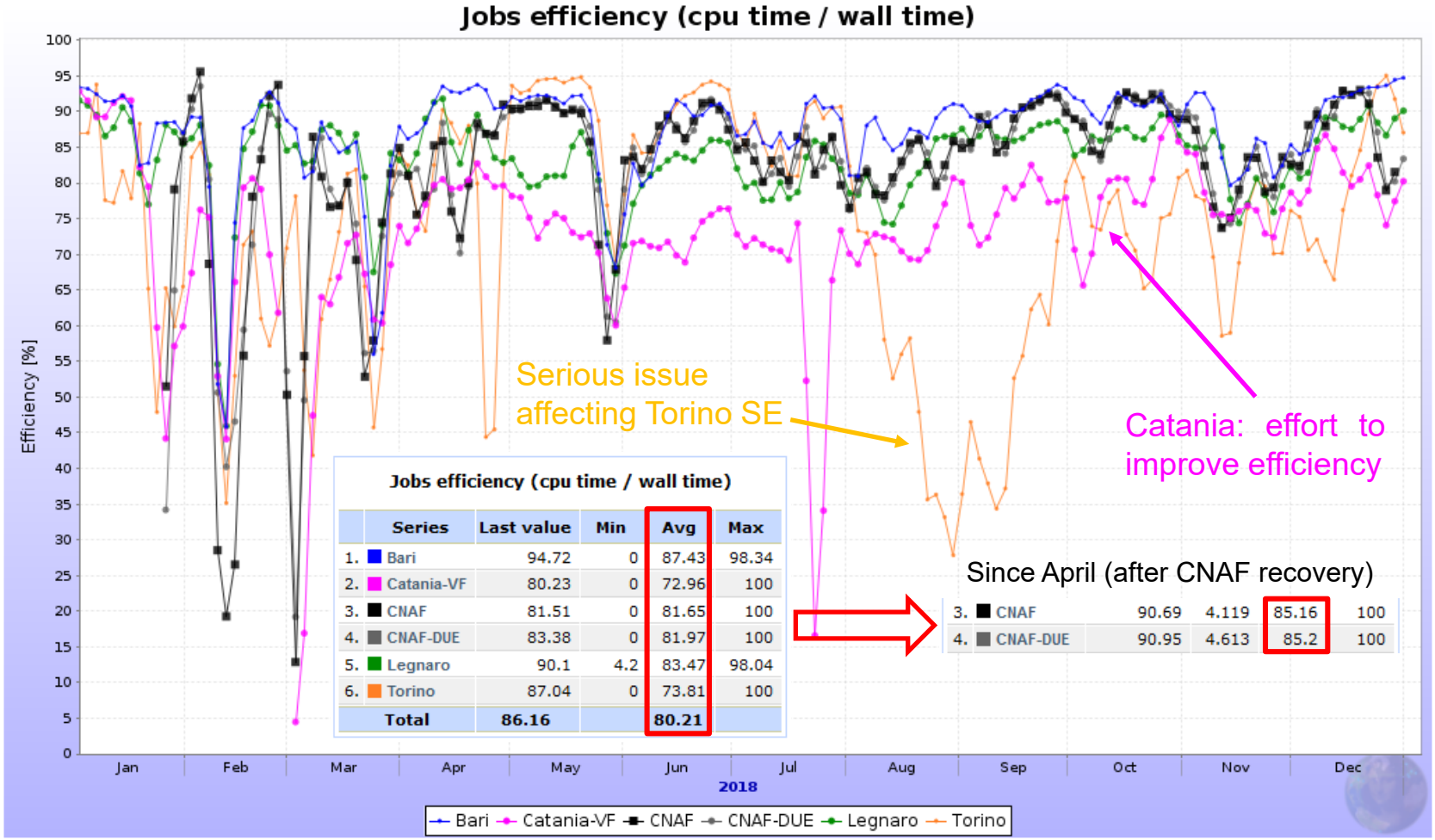
2018



Performance of the Italian sites: T2 CPU



Performance of the Italian sites: T2 Eff.

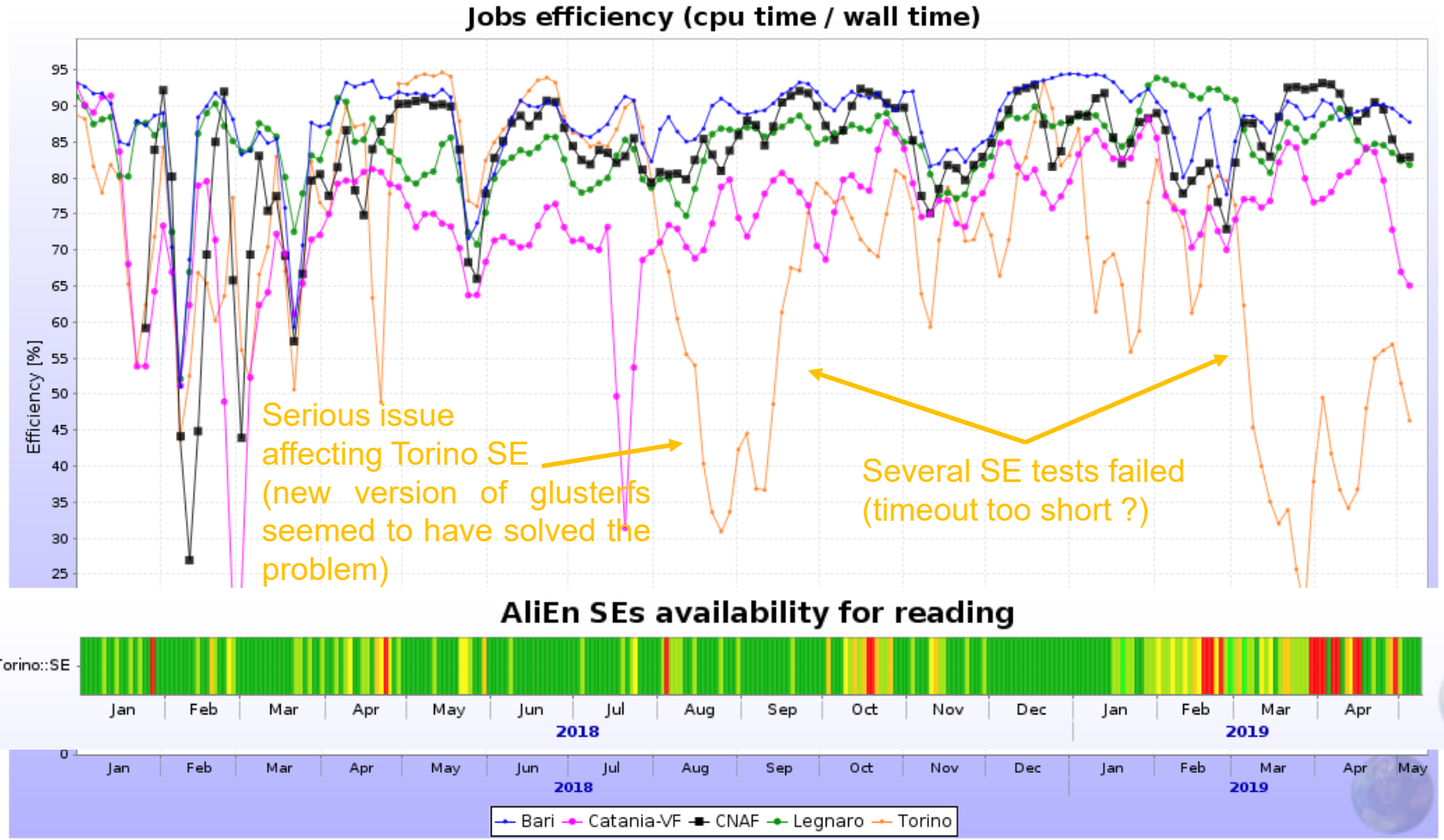


Jobs efficiency (cpu time / wall time)

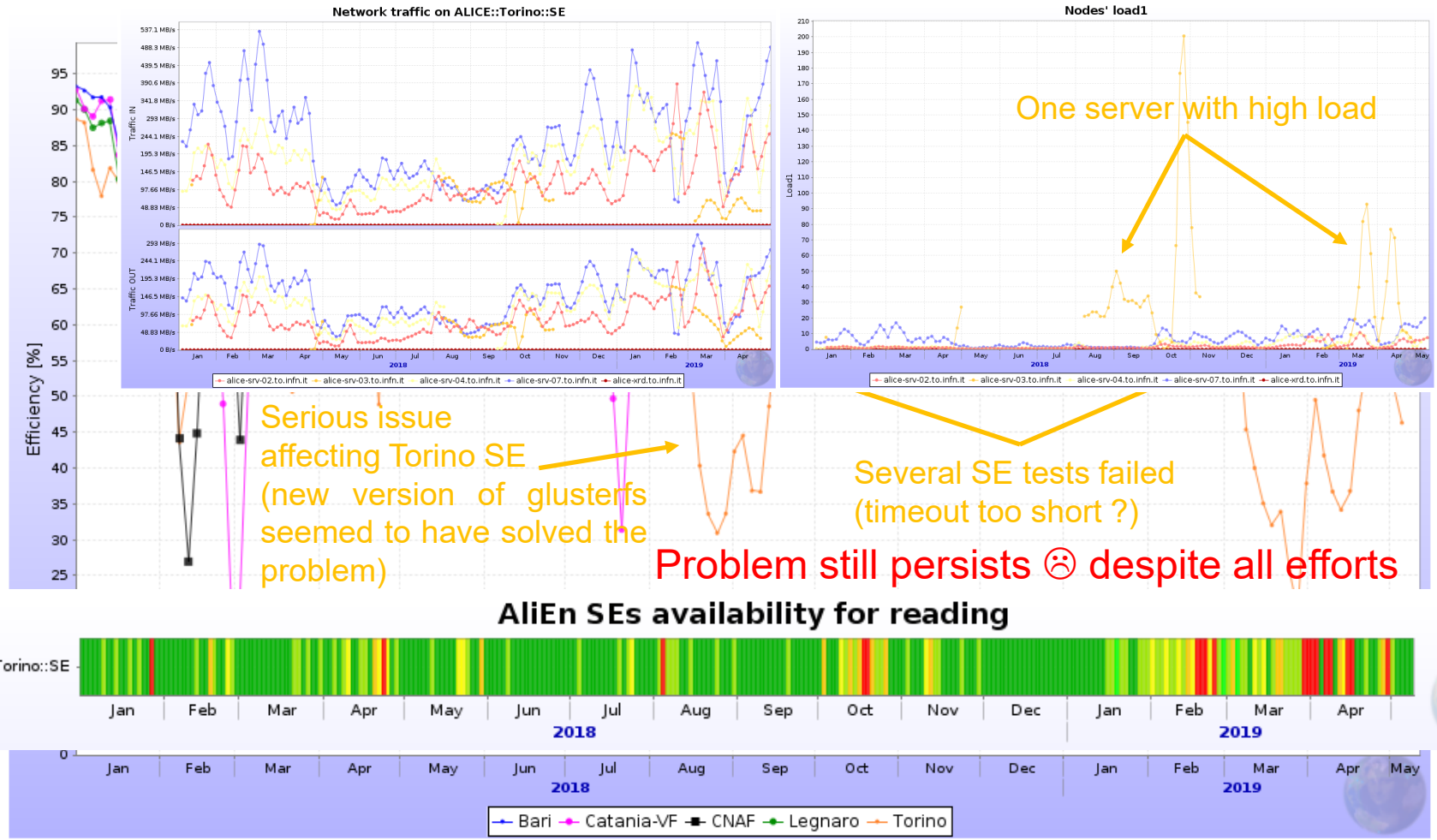
	Series	Last value	Min	Avg	Max
1.	Bari	94.72	0	87.43	98.34
2.	Catania-VF	80.23	0	72.96	100
3.	CNAF	81.51	0	81.65	100
4.	CNAF-DUE	83.38	0	81.97	100
5.	Legnaro	90.1	4.2	83.47	98.04
6.	Torino	87.04	0	73.81	100
	Total	86.16		80.21	

3.	CNAF	90.69	4.119	85.16	100
4.	CNAF-DUE	90.95	4.613	85.2	100

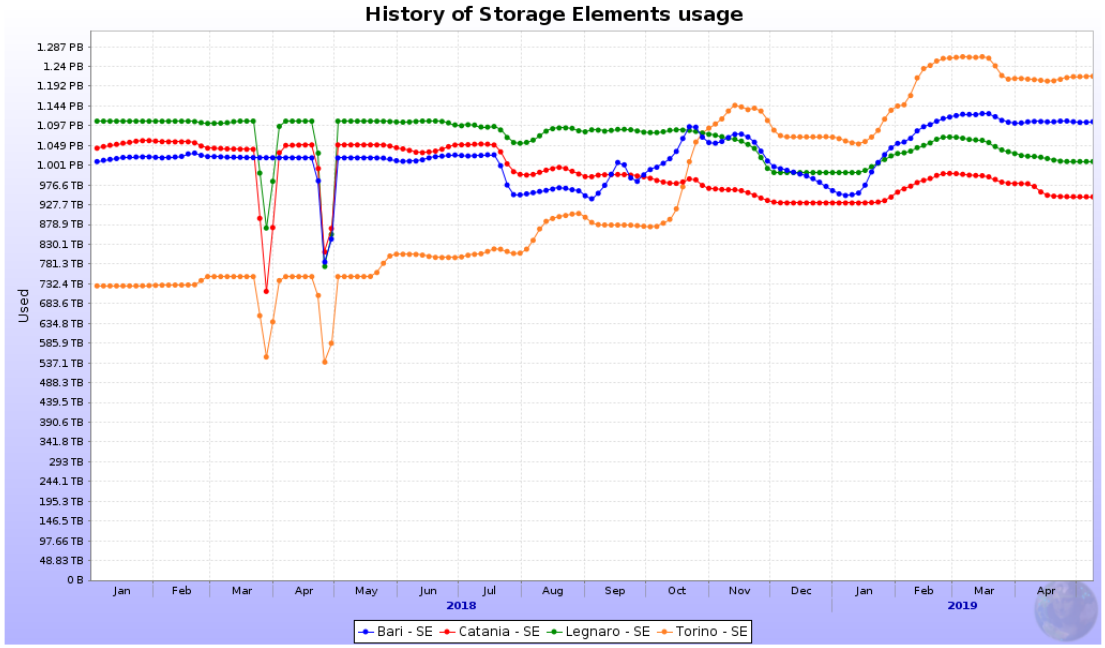
Performance of the Italian sites: T2 Eff.



Performance of the Italian sites: T2 Eff.

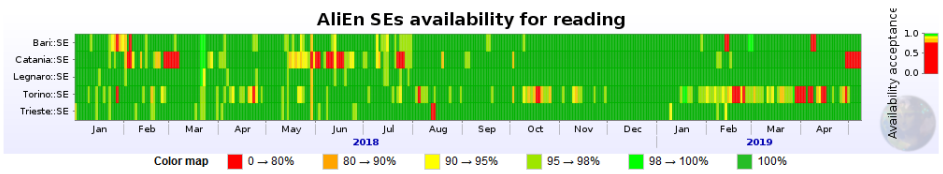


Performance of the Italian sites: T2 SE



	Used [TiB]	Size [TiB]	Used %
BA	1132	1368	83%
CT	947	1095	87%
PD-LNL	1034	1100	94%
TO	1245	1270	98%

BA, PD-LNL SE very reliable > 98%, PD-LNL and TO almost full! Waiting for the full deployment of 2017 and 2018 procurement…



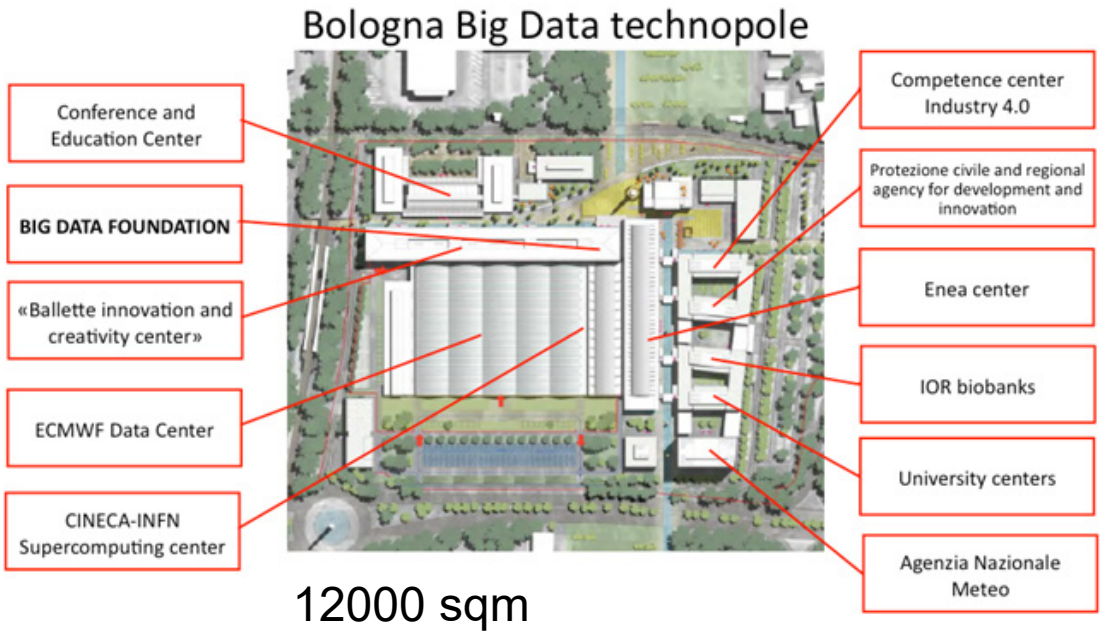
Statistics

Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
Bari::SE	01 Jan 2018 00:00	08 May 2019 23:42	11743	187	98.43%	98.43%
Catania::SE	01 Jan 2018 00:02	08 May 2019 23:42	10995	712	93.92%	93.26%
Legnaro::SE	31 Dec 2017 23:13	08 May 2019 23:40	11922	23	99.81%	99.83%
Torino::SE	31 Dec 2017 23:22	08 May 2019 23:44	11272	520	95.59%	95.68%
Trieste::SE	31 Dec 2017 23:22	08 May 2019 23:47	11707	82	99.30%	99.31%

Search for a new location for the Tier-1

- ECMWF center will be hosted in Bologna from 2019 in the Technopole area

- Possibility to host in the area also:
  - INFN Tier-1
  - Cineca computing center

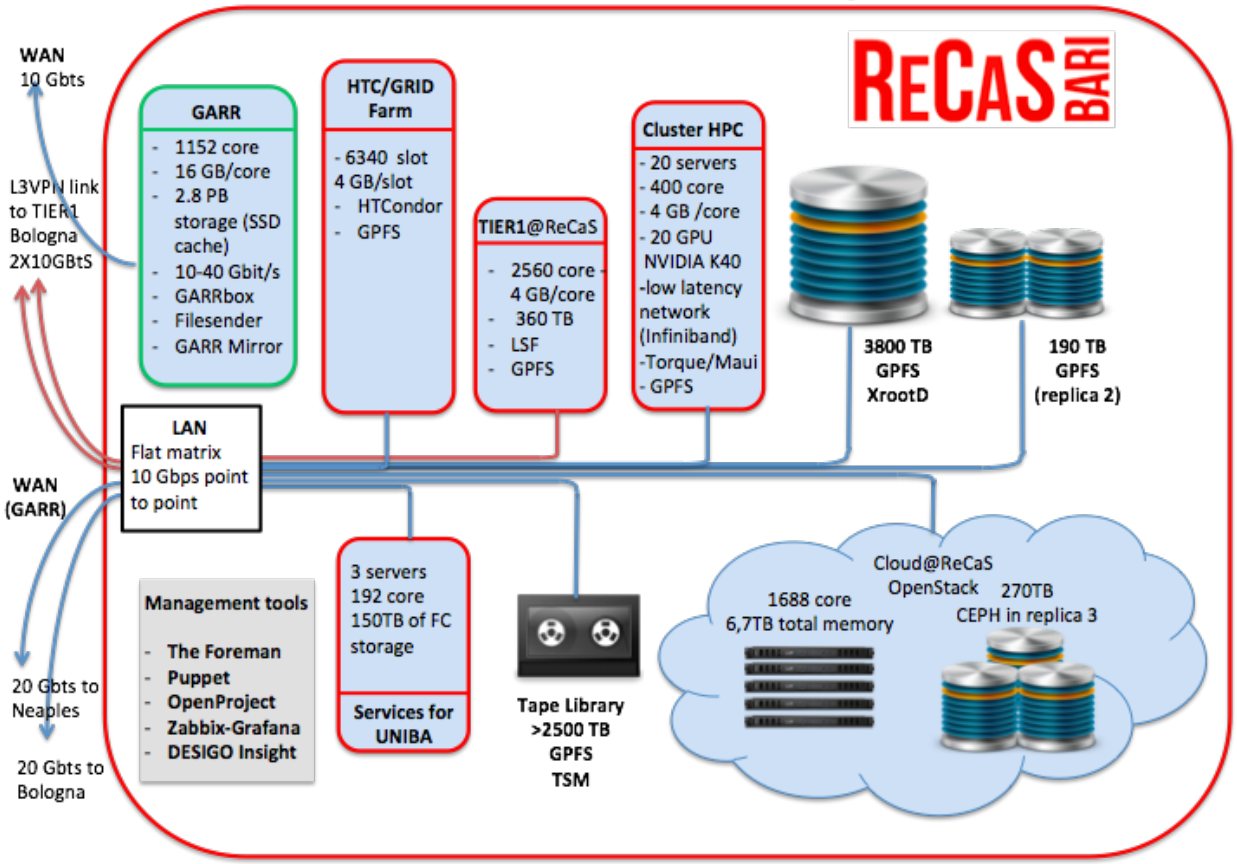


- Already allocated 40 M€ for the Italian government to refurbish the area. Looking for extra budget for INFN & CINECA

INFN-Bari Tier2 @ReCaS

Istituto Nazionale di Fisica Nucleare

ALICE



- 420 m2 with 4x20 racks
- 10 Gbps point-to-point
- 6 Computer Room Air Conditioners
- UPS 800kW x 7 min + GP1650
- Dedicated network link: 10Gbps x2 to GARR, 20Gbps to Naples and 20 Gbps to Bologna
- Cloud platform: OpenStack
- Batch system: HTCondor
  - 184 Worker Nodes
  - 350+ network connections
- 8192 cores (2304 INFN - 5888 UNIBA)
- 3552 TB DELL (1152 INFN - 2400 UNIBA)

- IBM System Storage TS3500 Tape Library (UNIBA): 2.5 PB Tape storage
- HPC cluster (800 core Intel, Infiniband and 20 NVIDIA K40 boards) (UNIBA)
- HPC Cluster composed of 20 servers

INFN-Bari Tier2 @ReCaS

Available resources:

Network: 20 Gbps link GARR-X/LHC-ONE (ready for upgrade to 40 Gbps)

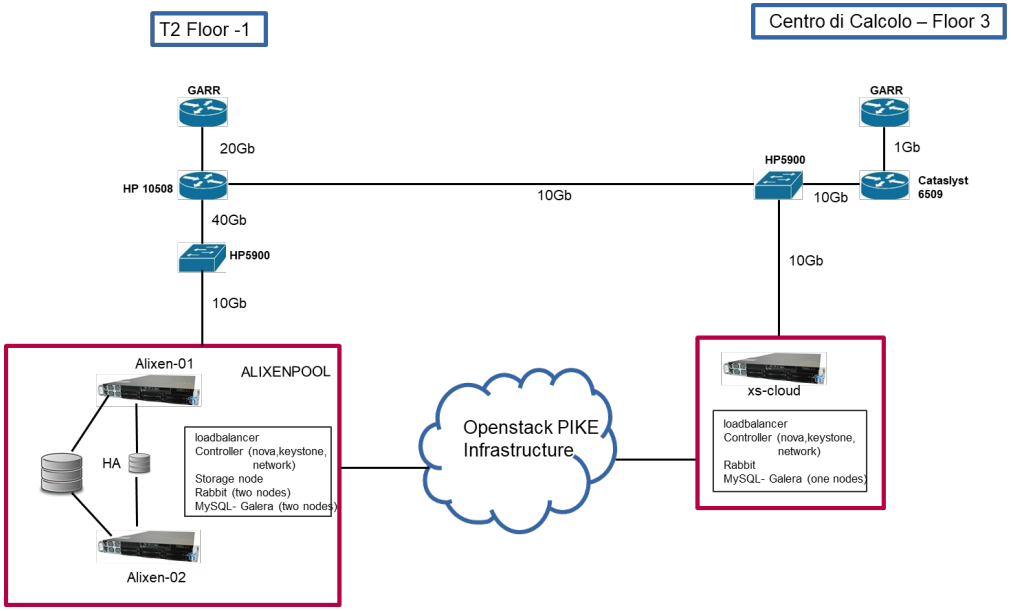
Storage:

- currently online: ~1.474 PB (2017 pledge)
- usage: 1.13/1.47 TB (76.6%), due to recently added 280 TB
- needed for 2018-2019 pledge: additional 270 TB, expected online before summer break
- software: native XRootD version 4.9.1 and IPv6-compatible

CPU:

- currently available: ~16000 HS06 (2018 pledge)
- usage: largely above pledge (opportunistic):
  - max running jobs: 5545
  - avg running jobs: 2338
- needed for 2019 pledge: ~2000 HS06

INFN-Catania VirtualFacility



ALICE INFN-CATANIA T2 Virtualization:

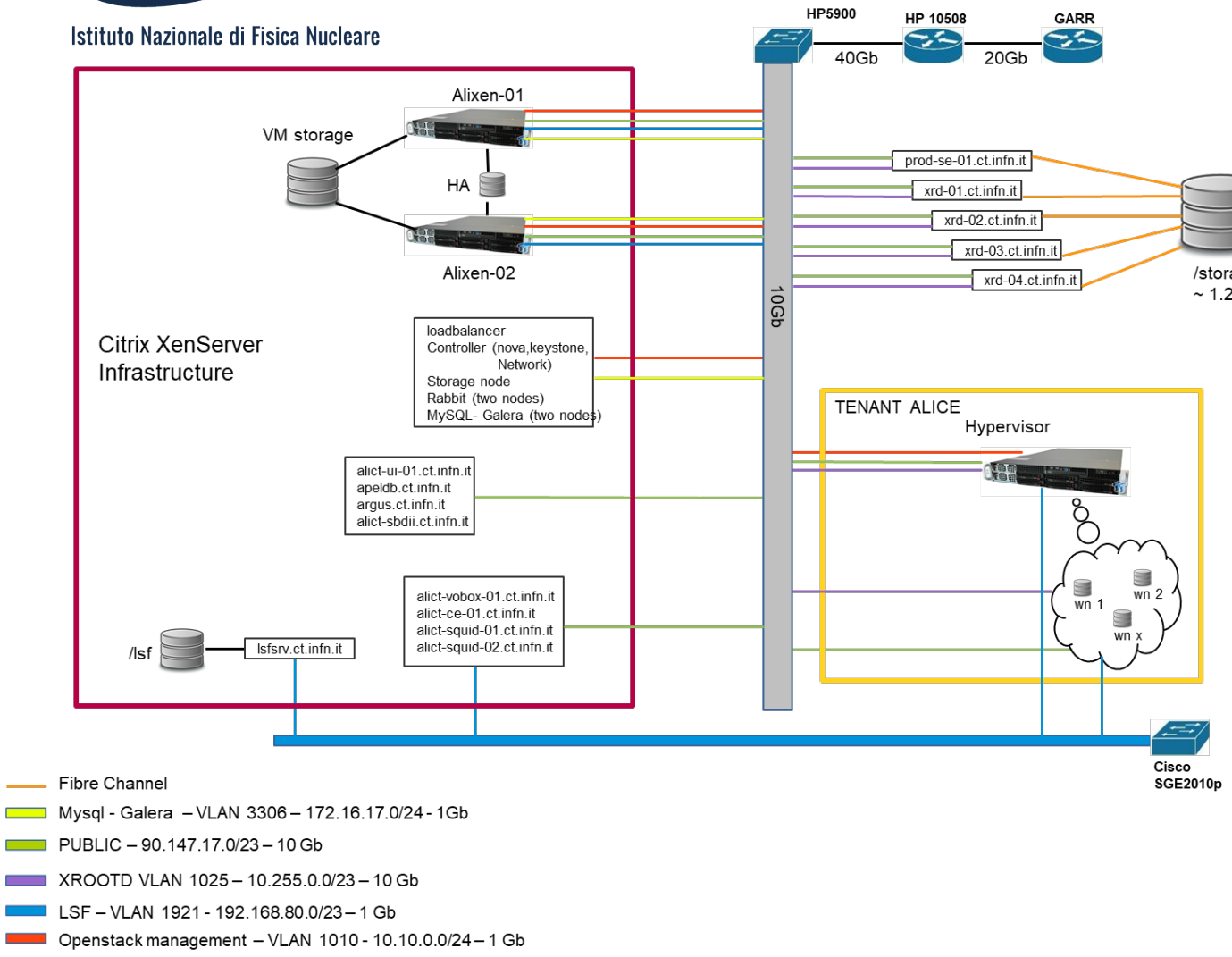
- Citrix XenServer pool
- running VMs for central services
  - CE (LSF Batch Server)
  - VOBOX, site BDII, UI, ARGUS, APELDB, CVMFS SQUID
- VMs per WNs
  - 5 VMs for each hypervisor
  - 14(8) core, 56(24) GB RAM 100GB disk
  - 160 VMs providing 2048 virtual cores

OpenStack in HA + Zen LB

- Controller Node, RabbitMQ, MariaDB (Galera Cluster), Keystone
- NetworkNode configured using L2 linuxbridge
- OS release Pike

INFN-Catania

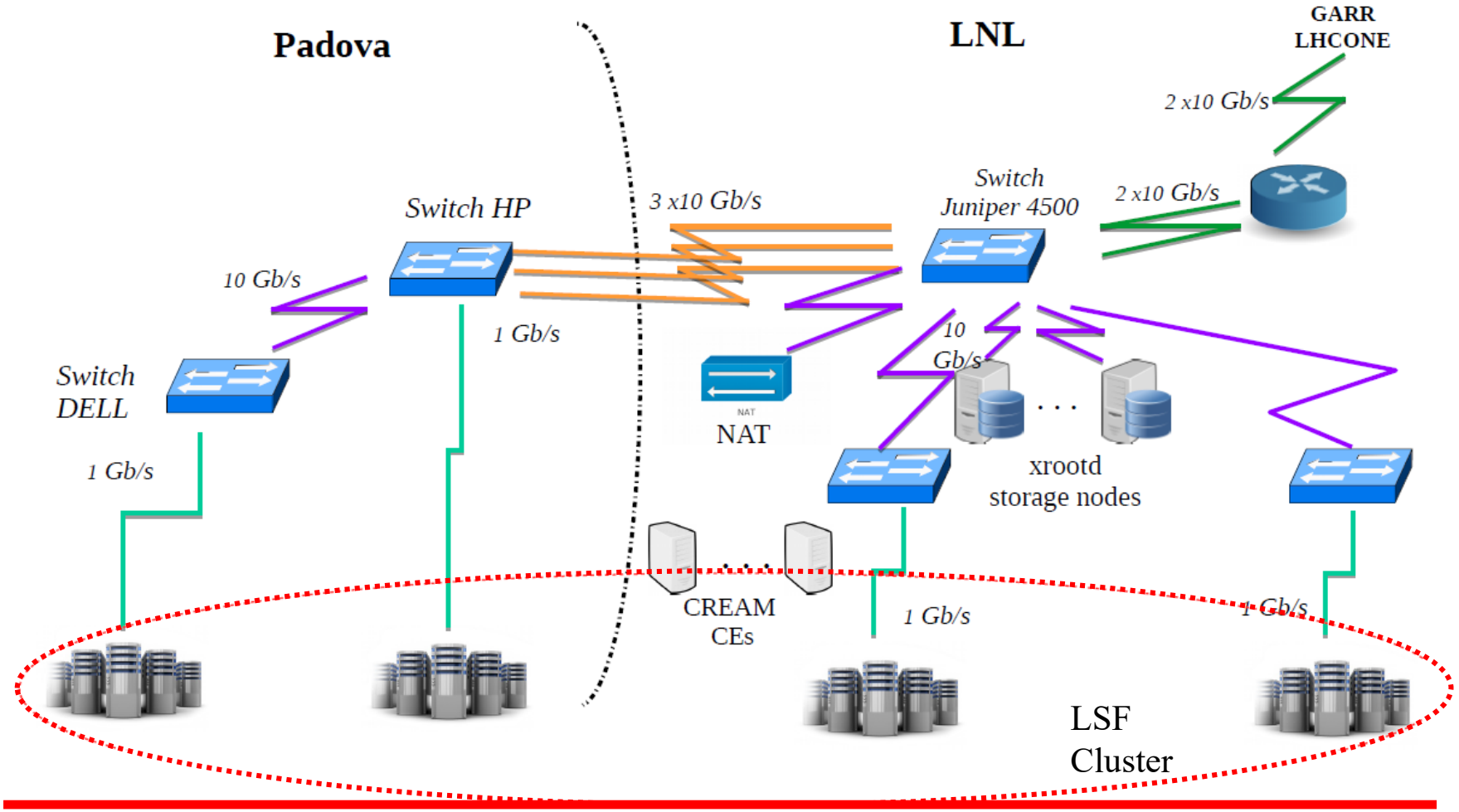
Istituto Nazionale di Fisica Nucleare



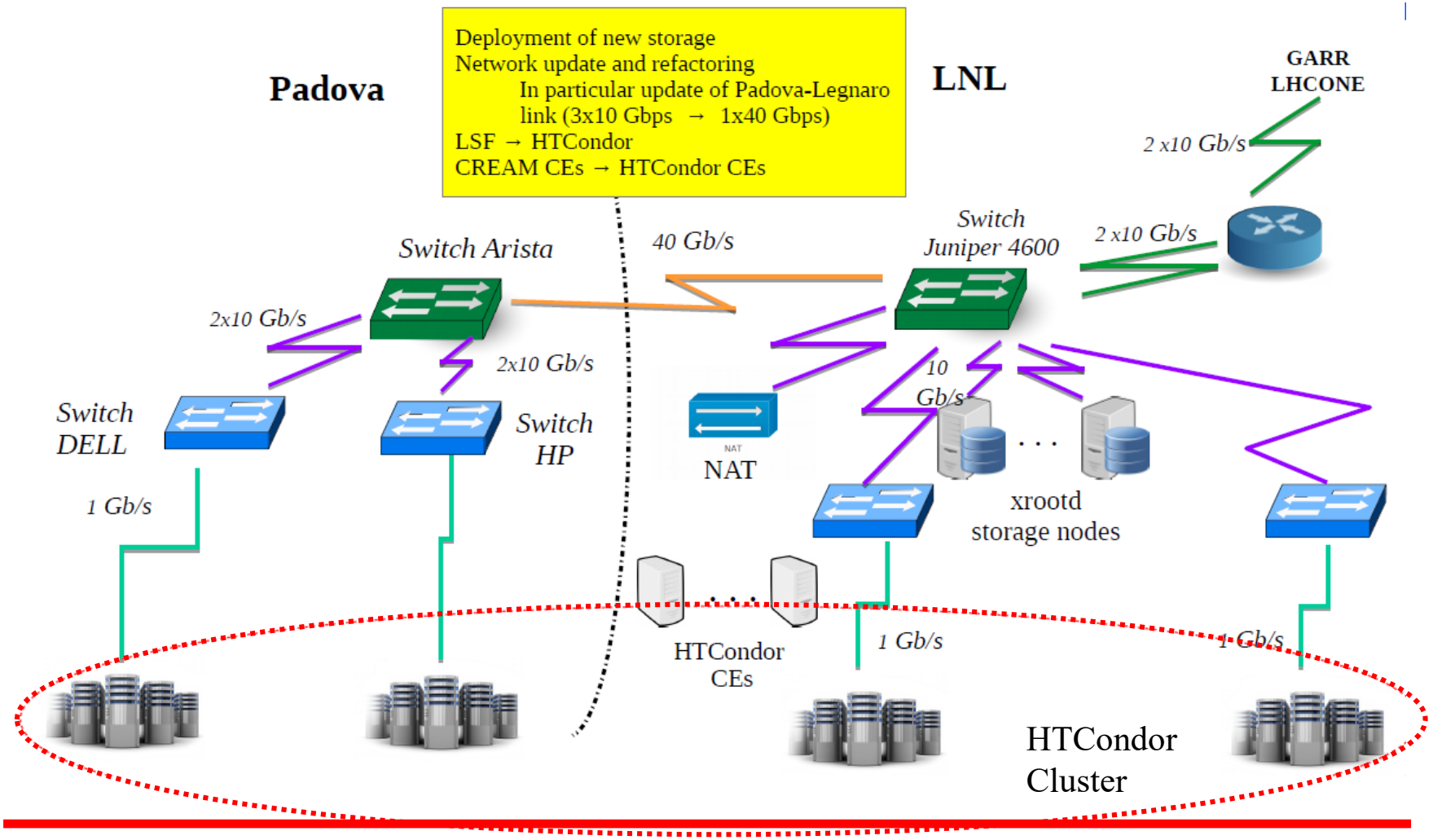
ALICE::CATANIA::SE:

- ~1.2PB (92% used)
- GPFS: v3.5.0.10
- 5 xrootd server (1 redirector)
  - xrootd v4.8.4
  - network bandwidth 50Gbps (10 Gbps x 2 for each server)
- 20 Gbps link GARR-X/LHC-ONE

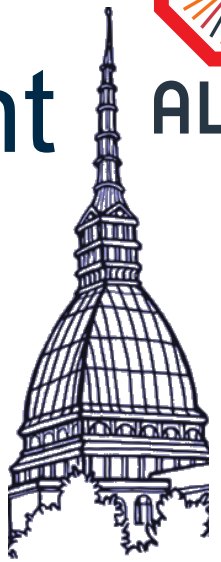
INFN-Padova LNL State of the art



INFN-Padova LNL Foreseen evolutions



INFN-TORINO Computing Element



Torino

- alibox2.to.infn.it submits to WLCG CE (t2-ce-01.to.infn.it)
- 224 Virtual WN 8 job slot —> 1792
- 7 Physical WN 16 job slot —> 112
- Total 1902 (1904 -2) max allowed running jobs
- NOT Alice exclusive

Torino-HPC

- occam-10.to.infn.it submits to opportunistic CE (occam-10.to.infn.it)
- Fully docker containerized services (VO-Box, Squid proxy, HTCondor queue, Plancton, workers)
- The VO-Box and, CE queue and the Squid proxy coexist on the same host
- 35 Physical WN
- Plancton daemons running on disused nodes launch worker processes that catch a (unique) job from task queue
- Currently only aliprod jobs are allowed
- The number of docker containers (jobs) running depends on max_memory (e.g. 4GB)
- Memory swap is allowed (old rotational HDD)
- Up to 600 concurrent jobs
- Alice exclusive

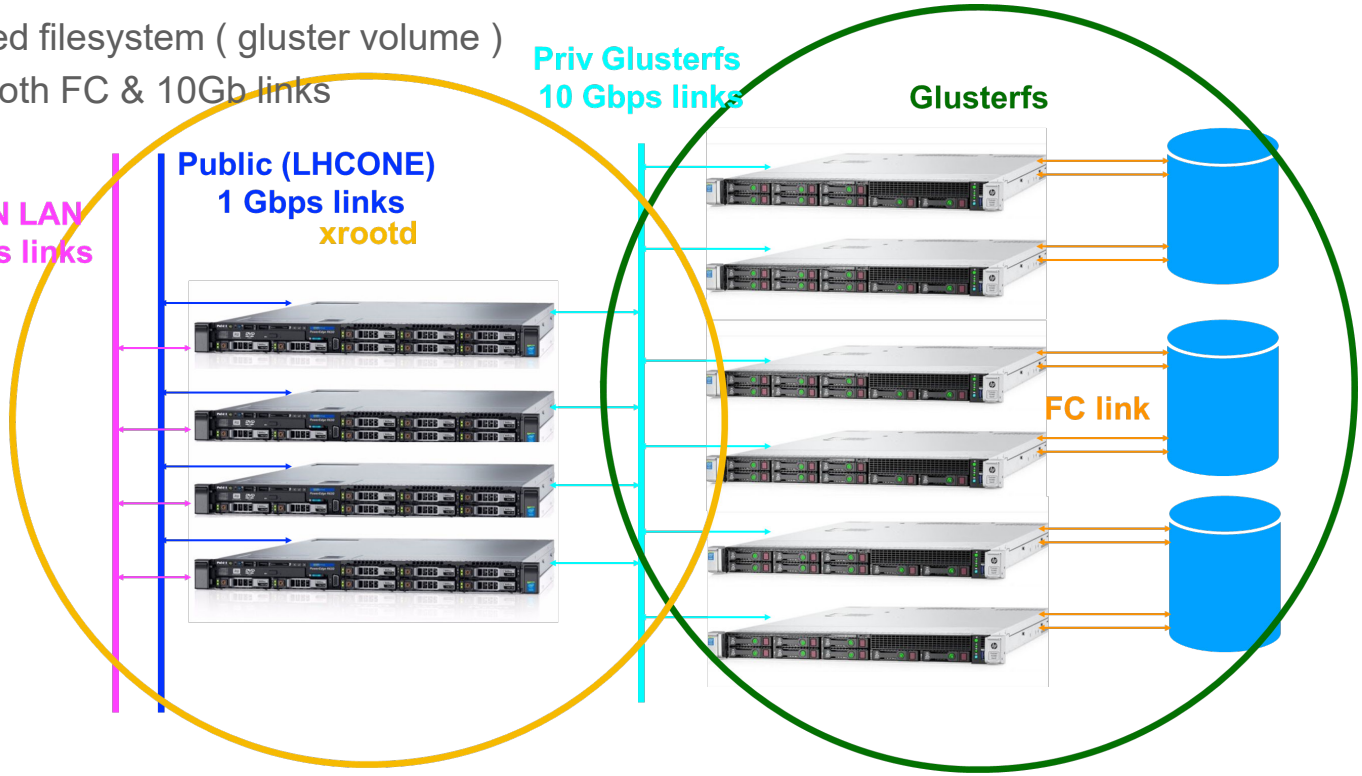
INFN-Torino Storage Element

ALICE::Torino::SE

- alice-xrd.to.infn.it xrootd v4.3.0
- 1 redirector + 3 xrootd servers
- Glusterfs 3.10-3 shared filesystem (gluster volume)
- 14 disk servers with both FC & 10Gb links
- 55 bricks

Totaling:

- 1440 TB
- (754 TB Used)
- (686 TB Avail)

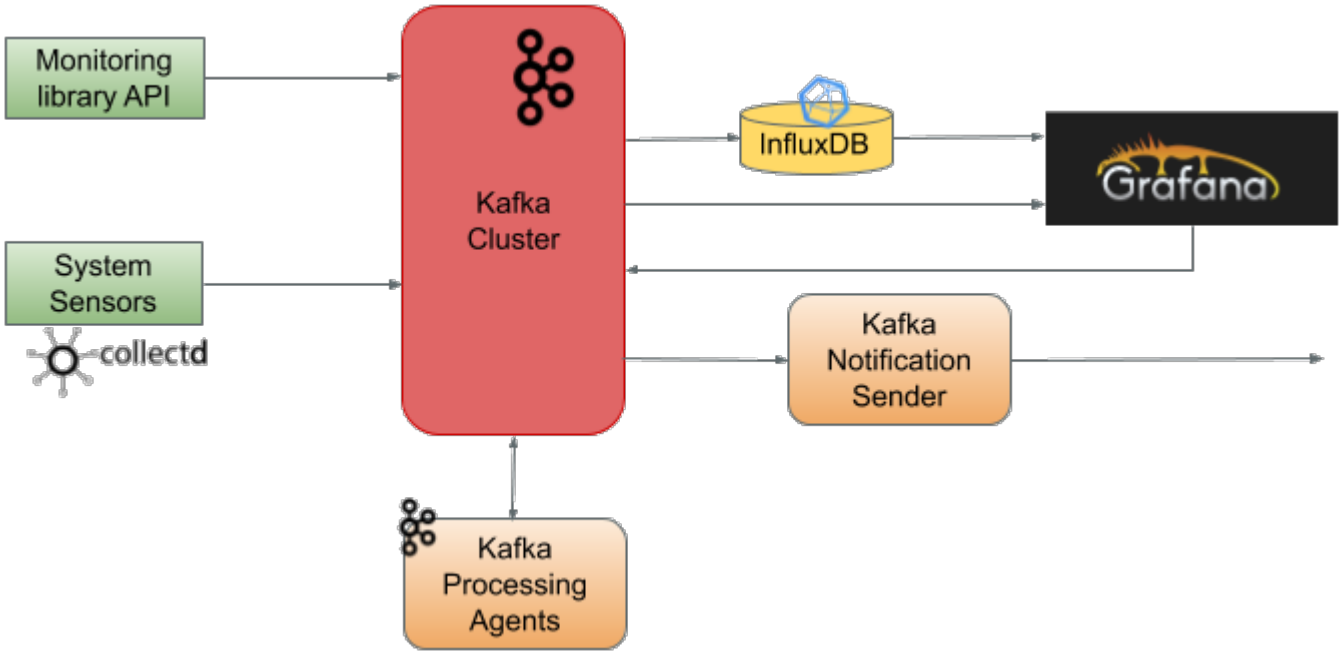


ALICE computing R&D activities in Italy

- 3 fellowship contracts provided by the INFN in 2017 + 1 in 2019 for the LHC computing developments towards Run3 and Run4:
- **G. Vino – Bari:** development of the new ALICE monitoring system for the O^2 farm at CERN
  - based on modular stack solution mostly relying on the Apache Kafka streaming platform
- **D. Berzano – Torino-CERN:** development of new analysis framework for the O^2 system
  - novel workflow management solutions on large workloads and datasets across heterogeneous platforms
- **S. Vallero – Torino:** new strategies in the analysis algorithms
  - developing a machine learning (ML)-as-a-Service toolkit that will enable the experiment to integrate the usage of ML algorithms
- **M. Concas – Torino:** development and implementation of unsupervised reconstruction algorithms on GPUs for the ALICE ITS upgrade

ALICE computing R&D activities in Italy

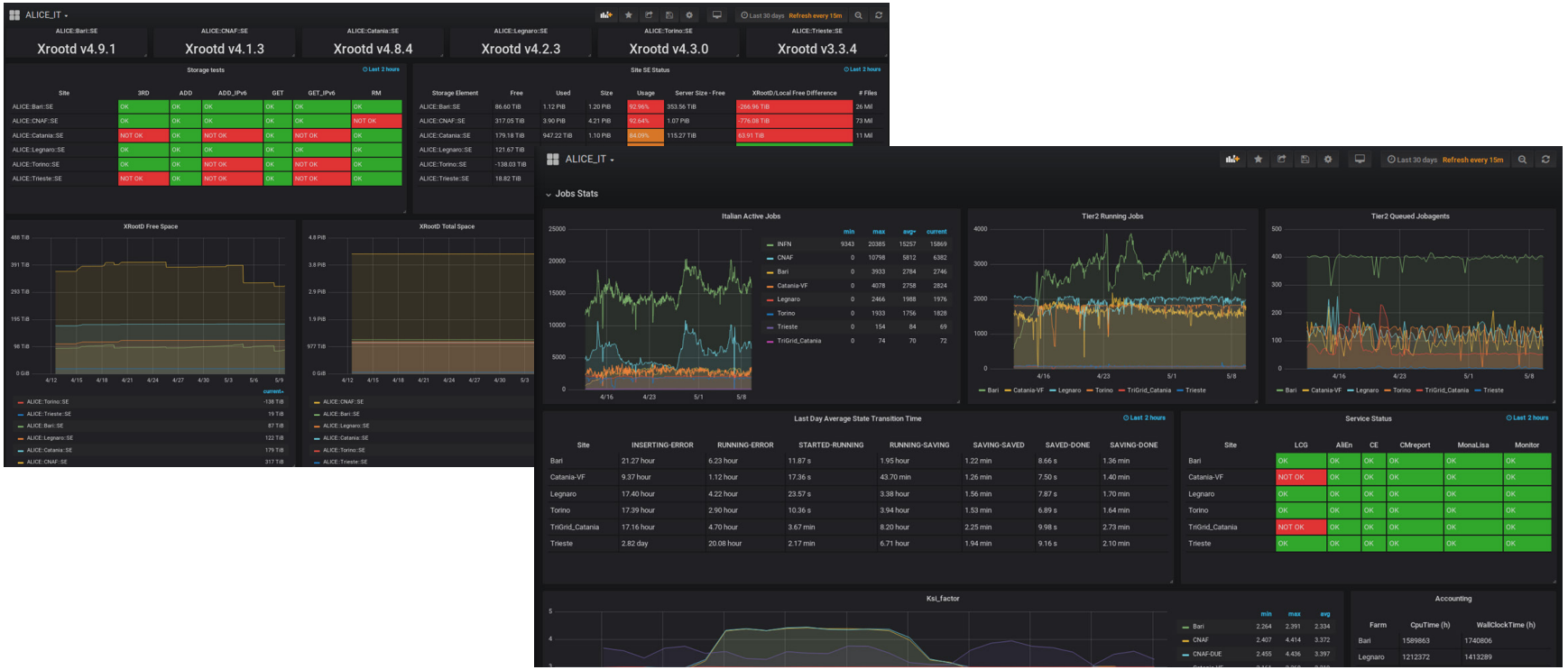
ALICE monitoring system for the O^2 farm at CERN:



ALICE computing R&D activities in Italy

ALICE-IT Dashboard:

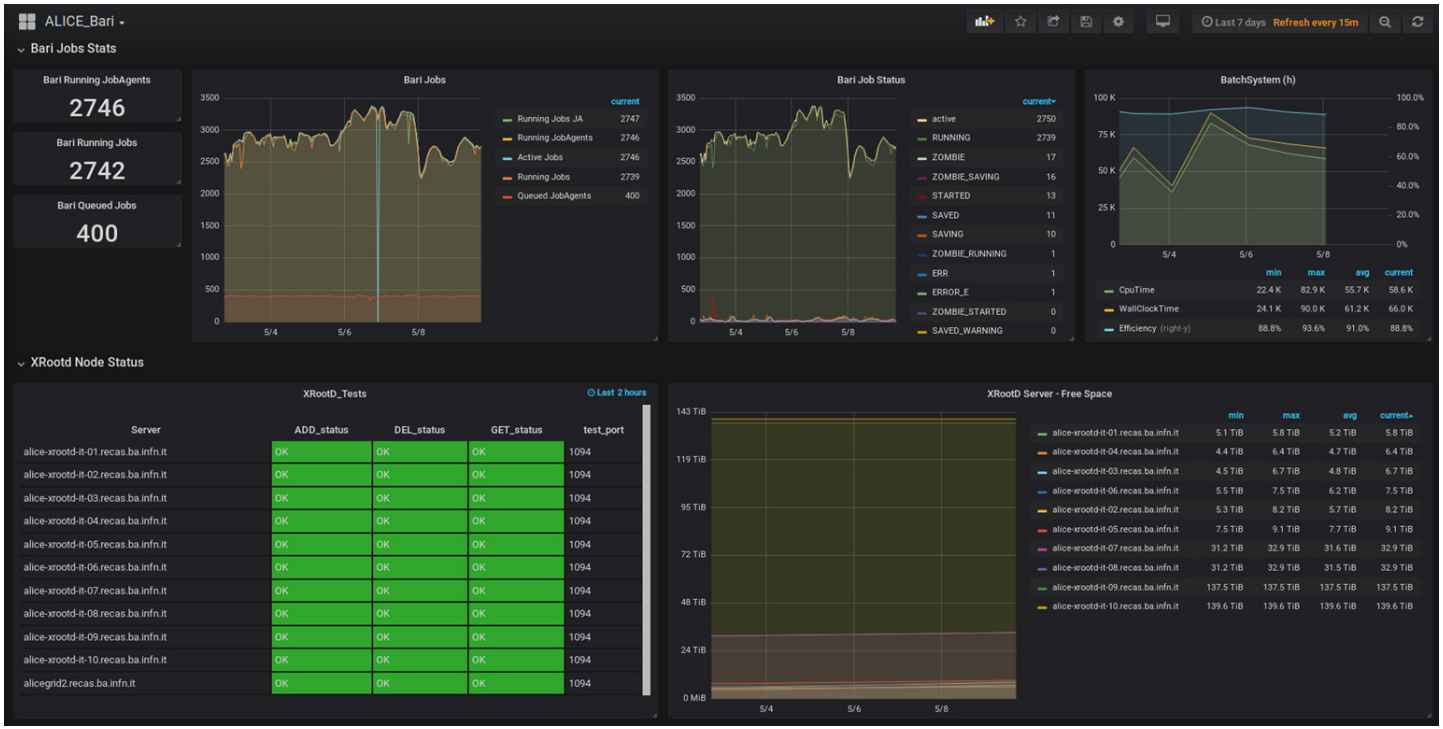
Initially developed in Bari, actually exported to all the ALICE-IT sites

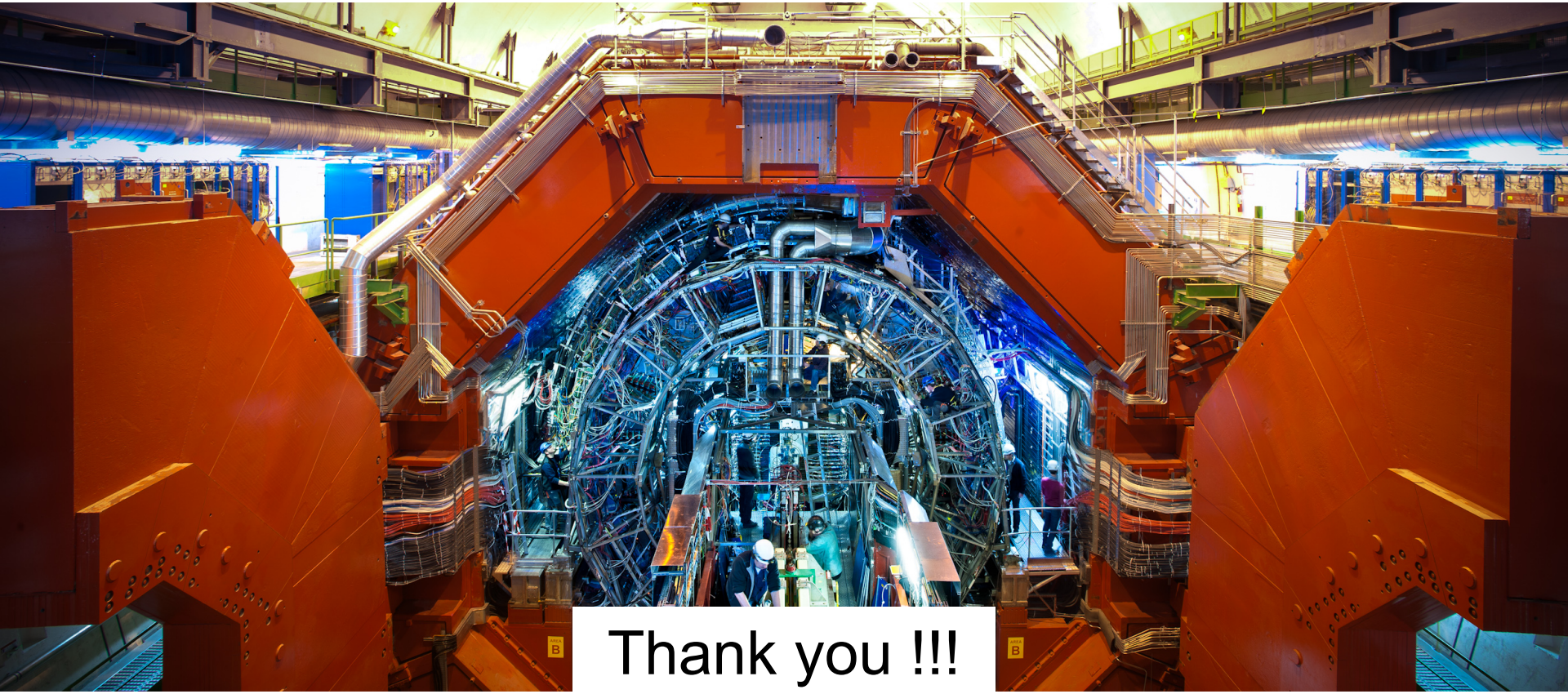


ALICE computing R&D activities in Italy

ALICE-IT Dashboard:

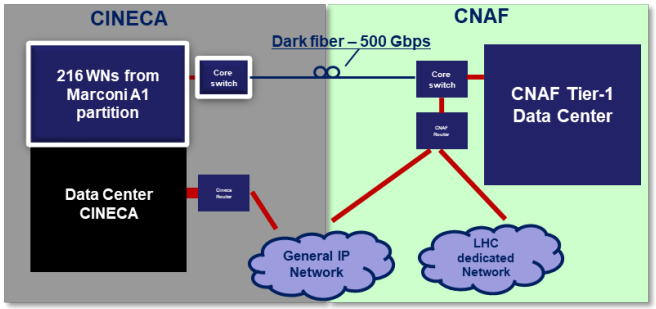
Initially developed in Bari, actually exported to all the ALICE-IT sites




Thank you !!!

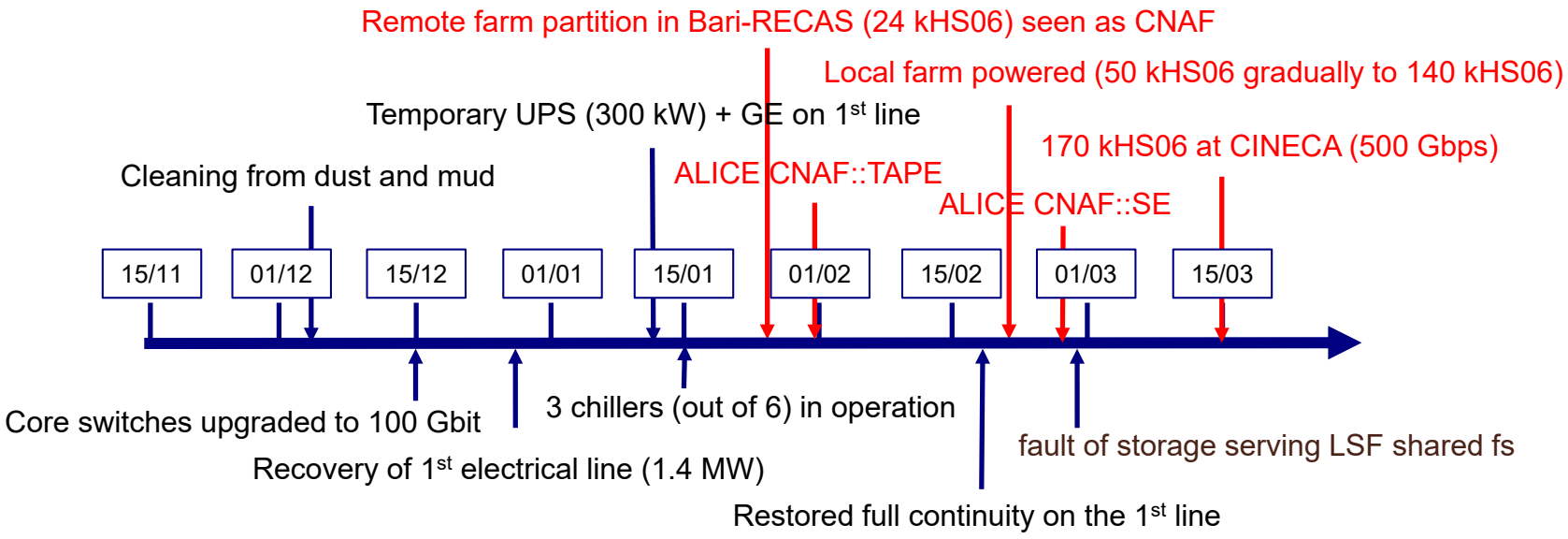
CNAF Farm remote extensions

- ~13% of CPU resources (~25 kHS06) pledged to WLCG (2017) experiments are located in Bari-RECAS data center
  - Transparent access for WLCG experiments
  - Similar to CERN/Wigner extension
  - 20 Gbps VPN
  - Disk cache provided via GPFS-AFM
- Since March 2018 ~170 kHS06 provided by CINECA
  - 500 Gbps (⇒ 1.2 Tbps) VPN ready
  - No disk cache, direct access to CNAF storage
  - CentOS7 with Singularity

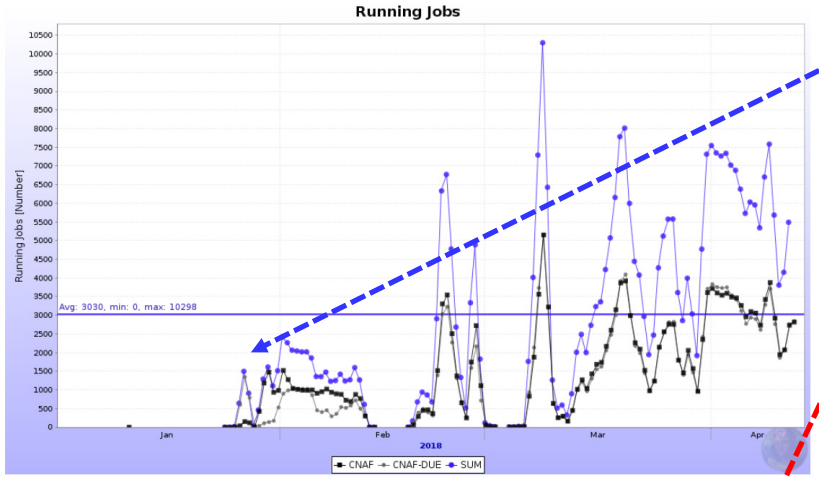


CNAF recovery timeline

- The CNAF team has been working tirelessly to restore the centre
- Weekly phone conference since November
- ALICE has been collaborating with the CNAF experts on re-establishing a normal operation at the center



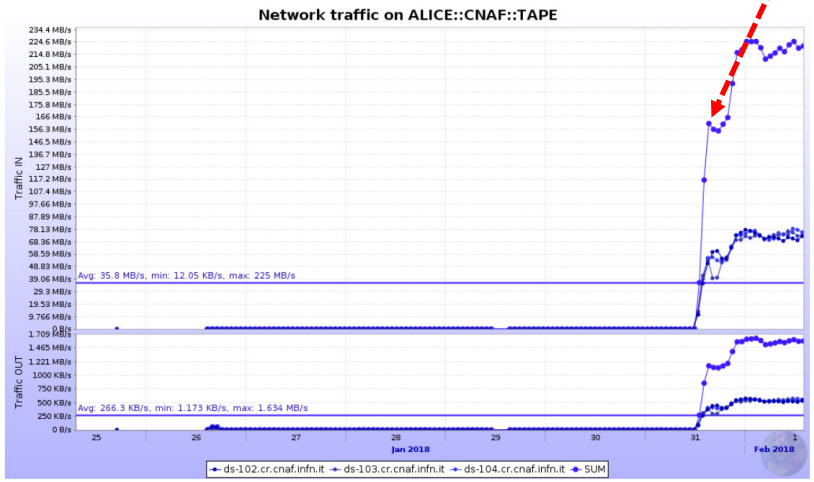
Prompt ALICE reaction to CNAF restart

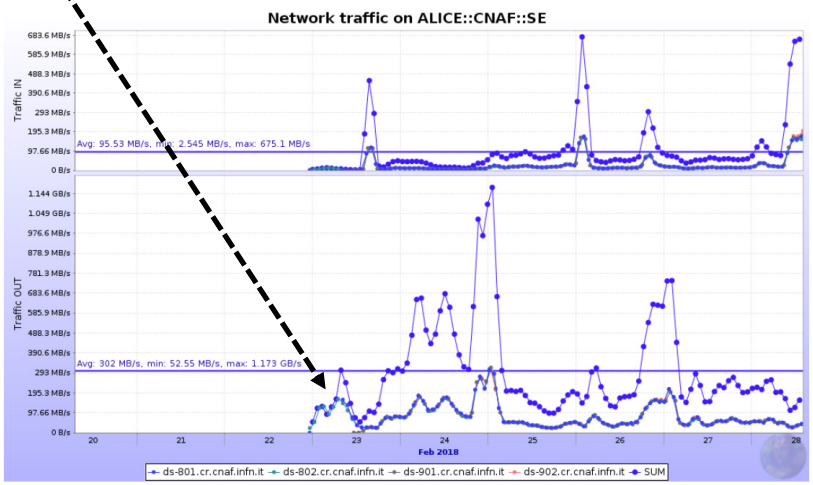


Since Jan 25th ALICE jobs have been running on CNAF queues! At the beginning only simulation jobs, reconstruction and analysis jobs turned on March (full functionality of SE is needed).

Tape available for ALICE on January 31st

Since Feb 23th SE accessible on R/W





CNAF accident 9/11/2017 Mitigation

- No large detrimental effect from the loss of CPU thanks to additional CPU offered at GridKA and CERN (and Bari-RECAS)
- The loss of 10% of ESDs (unique files) resulted in reduction of statistics for certain analyses (possible to re-generate ESDs)
- The AODs were replicated at other storages, in the order of importance for analysis, to preserve the turnaround speed of the analysis and to avoid bottlenecks
- The data on the damaged tapes has one replica at CERN:
  - high cost of tape restoration ➔ ALICE can replicate again the data once the tape service at CNAF is restored
- In addition about 1 PB of 2017 RAW data were not replicated to CNAF and the replication of the data was postponed to 2018.
  - no replication at other T1s to keep balanced RAW at the T1s