INFN CNAF Data Center

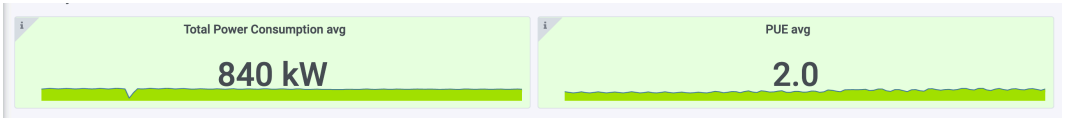
Luca dell'Agnello

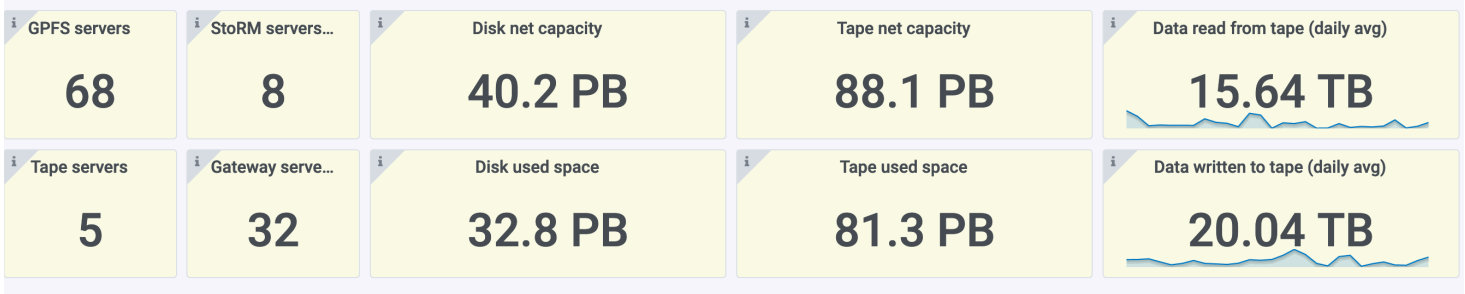
CTSC review

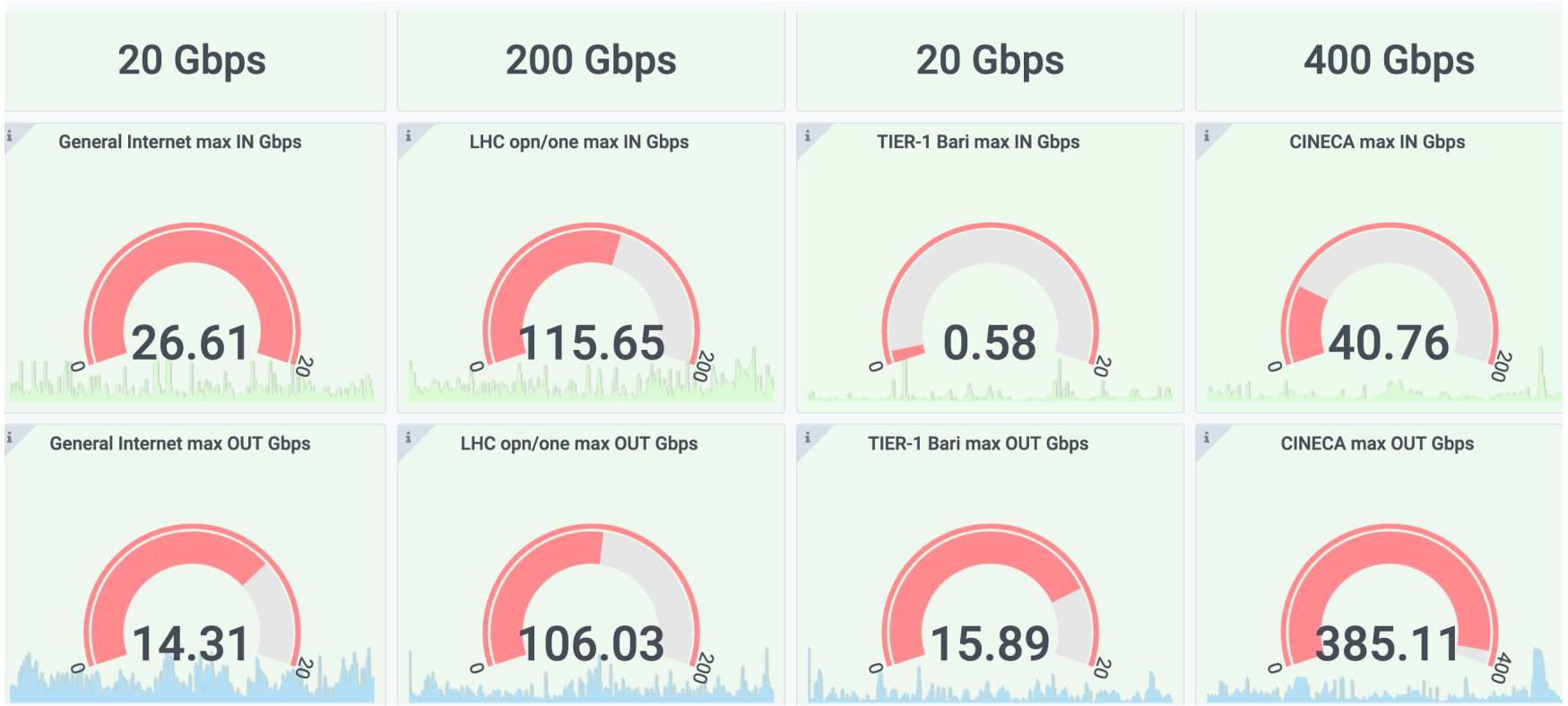
Bologna, October 9 2020

Multiple KPIs have been defined, would be nice to summarize the facility performance in a handful list of KPI, is there a plan? Which KPIs?

- An extensive set of KPIs has been defined for all aspects of the data center. A subset is presented at the same page of the monitoring (https://t1metria.cr.cnaf.infn.it/)
- We are reconsidering all these in order to reduce to those really meaningful to give the performances and state of the data center






20 Gbps
200 Gbps
20 Gbps
400 Gbps
General Internet max IN Gbps
26.61
0
20
LHC opn/one max IN Gbps
115.65
0
200
TIER-1 Bari max IN Gbps
0.58
0
20
CINECA max IN Gbps
40.76
0
200
General Internet max OUT Gbps
14.31
0
20
LHC opn/one max OUT Gbps
106.03
0
200
TIER-1 Bari max OUT Gbps
15.89
0
20
CINECA max OUT Gbps
385.11
0
400

CR7: long term preservation of Data and software is not mentioned.

- Funding for computing resources of experiments is negotiated between each collaboration and a referral committee (external to CNAF)
- On the other hand, the analysis of the experiments' requirements in terms of needed performances (e.g. Quality of Service for the storage, needed bandwidth, etc.) is done before the actual allocation of the pledges, with the collaboration of User Support and of the CNAF technical groups.
- Unless explicitly requested (and funded) LTDP is meant as bit preservation only
  - Only 4 experiments have been ceased the production phase so far
  - For CDF we have the complete framework for analysis up and running (and used!)

- StoRM is evolving in line with the requirements emerging from the WLCG and other communities served by the CNAF data center (Storm is also used in more than 20 other WLCG sites).
- StoRM through GEMSS provides a flexible and efficient system for tape management for which we have not found yet a valid alternative.
- Even moving from GPFS to CEPH is not an easily viable solution to get rid of GEMSS (we should repack O(100) PB of data out of TSM to the new system)

- Stopping the Tier-1 service for months during Run3 would be unacceptable for our users. Moreover, besides the WLCG experiments, we provide resources to several other experiments and the stop would require, in many cases, moving their activity (and the data!) to other centers (most of the experiments rely on Tier-1 only!).
- Furthermore, to migrate the storage, in any case, we would need to backup or copy the data in advance on some storage installed in the new data center before actually moving the hw.
  - Considering that we have ~6000 HDDs, the probability of getting RAID systems back on-line without significant losses is quite low. Moreover, none of storage providers would guarantee data integrity on any system which has been moved from one location to another.
  - This is the most delicate and long operation (in any case!)

Timescale of migration by services (how to migrate storage? Tape library?

- The most difficult part is to move the storage (in any case data should be copied in advance), we do not see a particular delay due to doing this operation without a down.
- First step: installation of the network systems and services (core router, DNS and alike)
  - DCI Tbit/s connection to CNAF
- Interconnection to Leonardo (CINECA pre-exascale machine)
  - Leonardo will provide almost all the needed computing power for Tier-1
  - Most of the WNs installed at CNAF will be simply switched off (the newest part of the farm will be mostly installed at CINECA in 2022).
  - Farm services (CEs, UIs etc.). will be moved one by one (they are mostly virtual services).
- Moving the libraries will require the prior installation of replicate HSM services at the new data center.
  - Moving a library, one at a time, we will be able to accept data from the Tier-0.
  - From the experience of the flooding, the time needed to unmount and remount a tape library is ~1-2 weeks.
- To move storage, the plan is to install a large buffer of disk (~20 PB) at the new data center and copy onto it the data from the storage systems at CNAF one by one (then moving the emptied hw to be used as part of the buffer).
  - One hp is to delay the acquisition of part of 2021 storage (the replacement part) and install it directly into the new data center.
  - From our experience to move data from a storage system to a new one (we perform this operation routinely without service interruption when replacing hw) can take less tan 2 weeks for a system of ~ 5 PB net.
  - In 2022 we will have to move ~ 50 PB of data (~20 weeks for copying the data)
  - Only a subset of the hw systems will be moved from CNAF
- We think we should complete the migration of storage in ~6 months (contingency including)
- The overall process should be completed in 7-8 months

Is it a transition (no interruption of service) or a migration (interruption) ?

- It is a transition (see previous answer)

Impact of covid on 'Tecnopolo renovation' schedule? Is there contingency?

- The schedule for renovation foresees that the hall for CNAF data center will be ready in Q2 2022 (while Leonardo will be installed some months before). The impact of Covid pandemic has been the delay of almost one year of Leonardo installation. This still fits within our schedule.

Storage: Why Ceph? Which configuration?

- *Ceph is a widely adopted and supported by a large community including within WLCG. Its features match well both requirements from "traditional" applications and from the cloud world.*
- *Our testbed is composed by:*
  - *8 servers (2x10 Gbps)*
  - *4 jbod with 30 disks each*
  - *Erasure coding 6+2*
- *Moreover we will gain experience with the system we will install at the end of this year (to be put in production Q1 2021)*

Storage: Timeline of evaluation wrt migration to Technopolo

- *20% of capacity out from the current storage tender will be configured as CEPH: we should be able to draw some definitive conclusions (mostly in terms of operational procedures and reliability) before the migration.*
- *In any case, we are also negotiating with IBM for the renewal of support of GPFS licenses: the first indications are very promising.*

Storage: Tape system: bandwidth far from maximum capacity (need for GEMSS?)

- *The quoted figure (400 MB/s) is the maximum bandwidth a single drive can achieve. As for the overall throughput, we are far from the maximum due to several factors:*
- *The new library still has few used slots*
- *The activity is quite limited*