GDB - May 9th -2018
G. Bagliesi

CNAF incident: CMS view

CompOPS and Facilities team
On behalf of CMS Collaboration

The incident and the recovery

- CNAF is the second largest Tier-1 in CMS, providing in 2017 (Rebus) between 12 and 16% of the total resources in CMS Tier-1 federation
- November 9th center was flooded
  - Power distribution heavily damaged
  - Tape library: 150 wet tapes/7000
  - 40 tapes belonging to CMS
- Disk arrays/servers lower in the racks damaged (but recoverable thanks to raid systems)
- CPU: Worker Nodes not damaged but off until center recovered
- At the moment of the flood, CNAF was providing to CMS 66 kHS06 of CPU, 4.2 PB of occupied disk, and 14.5 PB of occupied tape.

- CMS followed very closely the evolution of CNAF, implementing actions in order to minimize the impact on its operations.
  - Our strategy was driven by the expected recovery timescale O(3 months)
- CNAF went back online in February
  - Recovery and re-commissioning operations in February quite smooth
  - most of the components working as expected without reiterated interactions.
- Overall: LHC experiments have been “less affected” thanks to the intrinsic redundancy of the distributed Computing Model

Effect of missing CPU

- The missing CPUs, while very valuable to CMS operations, constitute 13% of the computational resources pledged at Tier-1s

- The deficit was partially alleviated by the offer from other centers (Tier-0, Tier-1s and Tier-2s) to provide additional resource
  - Incident was close to the end of 2017 Run
  - Tier-0 and HLT resources became available shortly, mitigating the problem.

Effect of missing data

- Larger effects on CMS dataset accessibility and data custodiality
  - CMS has reduced along Run 2 the number of disk copies to essentially one (excluding the very small MiniAOD data tier)
  - a great part of the CNAF disk content was not accessible on other sites
  - tape copy of all important data is usually kept
    - in most cases recovery from tape was possible (apart from the cases where the tape copy was also sitting at CNAF).
- In relatively few cases it was necessary to trigger the re-generation of samples
  - small resource utilization, but expensive procedure for computing operators
- in February, when the disk came back online, we could recover all the files

Effect of missing tapes

- CMS operated under the assumption that the large majority of tape cartridges were safe and will be accessible again within few months.
- Special actions taken for the 40 cartridges that suffered direct water contact:
  - 6 cartridges holding RAW data (~60 TB):
    - content replicated from CERN to other Tier-1s in order to re-establish official CMS custodial policy (2 full copies of all RAW data); this happened in 1-2 weeks from the flood.
  - 34 cartridges holding derived / reproducible datasets (~300 TB):
    - no action was taken, since we were reassured data would have been probably recoverable anyhow
    - to date, we know the recovery process is proceeding without problems
- If recovery fails for some tapes, we will need to plan a difficult reprocessing program:
  - old software releases (5 or more years)
  - sizeable effort in terms of manpower
- At the moment the risk seems really limited, and we are still confident all the data will be available again shortly.
  - Full recovery expected by the end of May

Final considerations

- The cost for CMS was mostly in terms of operation manpower
  - Analyses and Upgrade activities were not delayed in a particular way, but operators had to identify missing samples, recover them from tape when possible, and eventually reprocess them
- A scenario in which data would have been permanently lost is much more difficult, since it would need reprocessing of data with old releases
  - We have the technology to do it in principle, but we never exercised in at large scale.
- The impact during data taking, when we would have missed a tape and disk endpoint, would have been larger
  - CNAF, given the size of its storage and the available bandwidth, is one of the Tier1s able to accept the largest data streams we collect, which are too big for other Tier-1s.
  - Our transfer/tape archiving bandwidth is generally heavily used, and a large Tier-1 being offline stresses the remaining, small number of tape sites, especially during data taking