CNAF experience in dynamic resources allocation

Enrico Fattibene

INFN-CNAF

pre-GDB - Tape Evolution – February 9, 2021

Outline

- Nearline storage @ CNAF
- Traditional drives allocation
- Dynamic sharing of tape drives
- Future works

Nearline storage @ CNAF - libraries and co.

- 96 PB of tape storage installed (85 PB used)
  - 1 Oracle-StorageTek SL8500 library
    - Almost full
    - 16 T10000D tape drives (scientific data), shared among experiments
  - 1 IBM TS4500 library
    - In production since February 2020
    - 6200 slots -> 120PB virtual capacity
    - 750 slots filled -> 15PB total space
    - 6 PB used
    - 19 TS1160 tape drives (scientific data), shared among experiments

Nearline storage @ CNAF - servers

- IBM Spectrum Protect (TSM) servers
  - 1 for HSM service (scientific data) - and 1 standby
- HSM servers
  - 5 active servers (1 for each LHC experiment and 1 for the others)
  - Each server can manage one or more GPFS FS, in HSM mode
  - Running TSM-HSM services and GEMSS
    - GEMSS provides optimization in migration/recall management
  - 1 standy-by server can be put in production in case of unavailability of one of the active ones

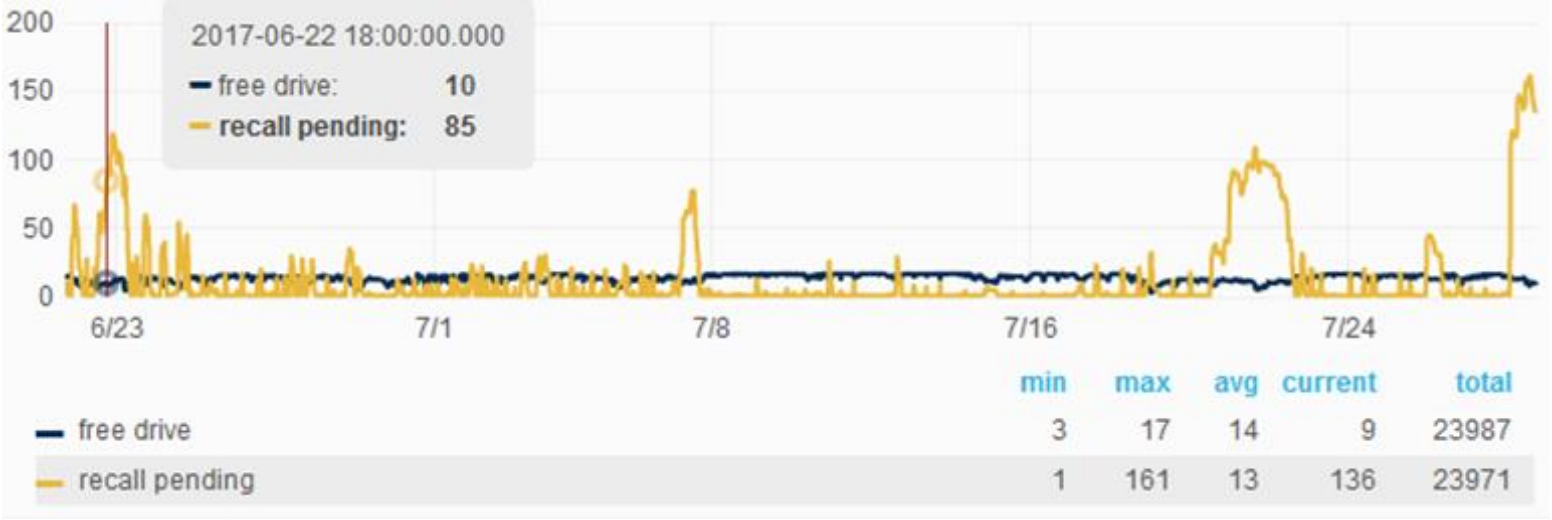
Traditional drive allocation

- Tape drives are shared among experiments
- Each experiment could use a maximum number of drives for recall or migration, statically defined in GEMSS
- In case of scheduled massive recall or migration activity these parameters were manually changed by administrators
- Administrative tasks (reclamation, repack) could interfere with production
- We noticed cases of free drives that could be used by pending recall threads

Traditional drive allocation

Total number of recall threads pending and free drives

- June-July 2017
- In several cases a subset of free drives could be used by recall threads



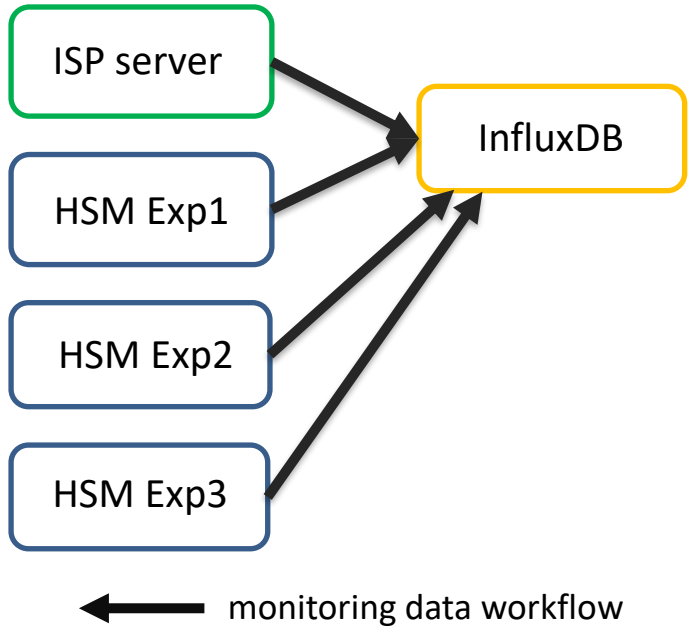
Motivations for a dynamic drive allocation

- Minimizing users waiting time for recalls
- Performing administrative tasks without interfering with production
- Optimizing tape drives usage results in saving costs
  - 20 drives used at 80% of time do (more or less) the same job of 40 drives used at 40% of time
- Data on tape are foreseen to grow of 20%/year until 2025
  - 200 PB by 2025
  - A more intense recall activity is expected

Dynamic allocation of tape drives

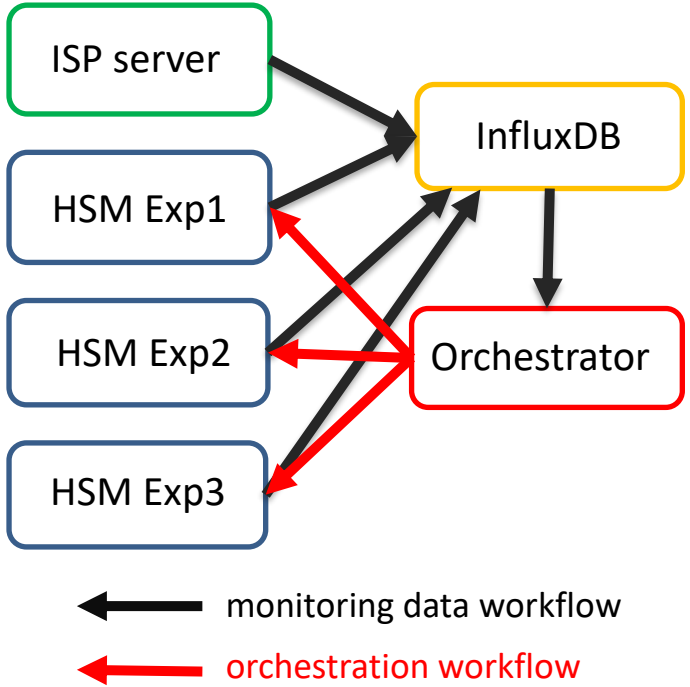
- Software solution to dynamically allocate drives to experiments for recalls

- InfluxDB stores monitoring information on:
  - number of free drives, from ISP server
  - number of recall threads running and number of pending recalls from each HSM server (e.g Exp1, Exp2, Exp3)



Dynamic allocation of tape drives

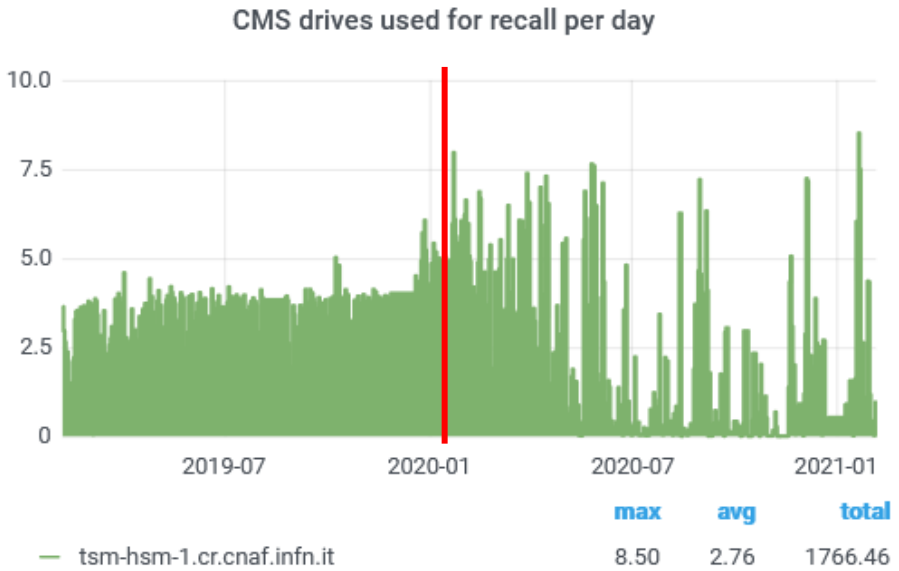
- Orchestrator:
  - performs comparison between pending recalls and free drives
  - in case of free drives and pending recalls, changes GEMSS parameter for maximum number of recall threads on the HSM server, to reach a maximum configurable value
  - can start reclamation processes when free drives are over a desired threshold

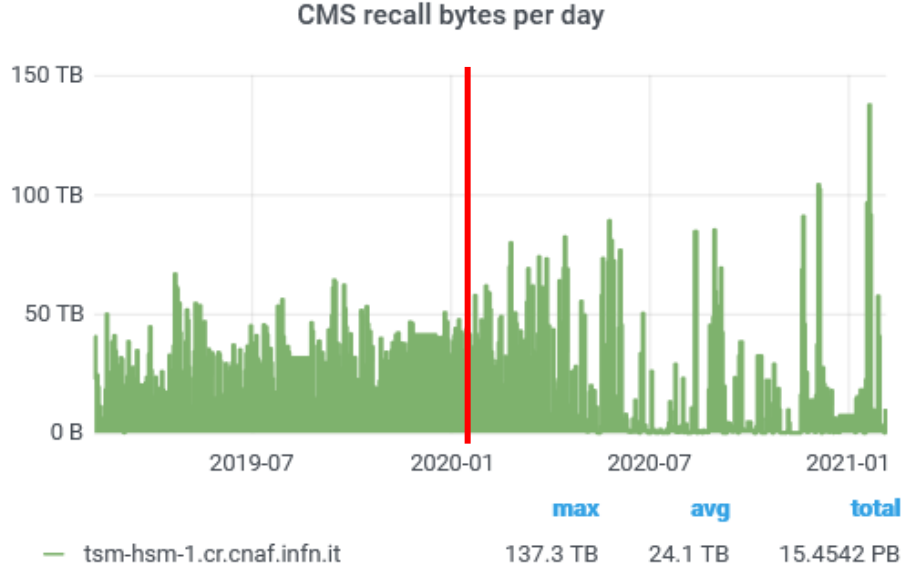


Dynamic allocation effects - CMS

Last 2 years CMS recalls. Dynamic allocation in prod since 20 Jan 2020

All data in SL8500 library (16 drives)





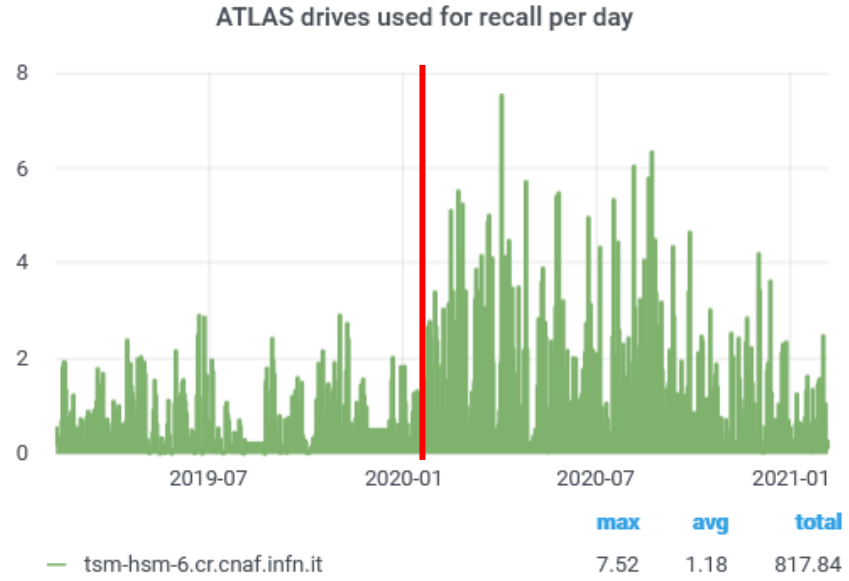
Data read first year (Feb19-Jan20): CMS 9.2 PB – All exp 16 PB

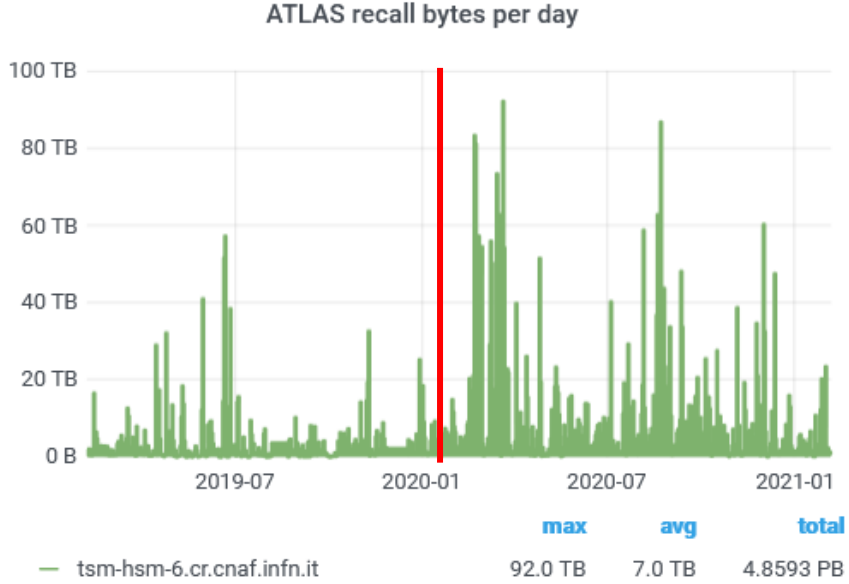
Data read second year (Feb20-Jan21): CMS 6.2 PB – All exp 13 PB

Dynamic allocation effects - ATLAS

Last 2 years ATLAS recalls. Dynamic allocation in prod since 20 Jan 2020

All data in SL8500 library (16 drives)





Data read first year (Feb19-Jan20): ATLAS 1.3 PB – All exp 16 PB

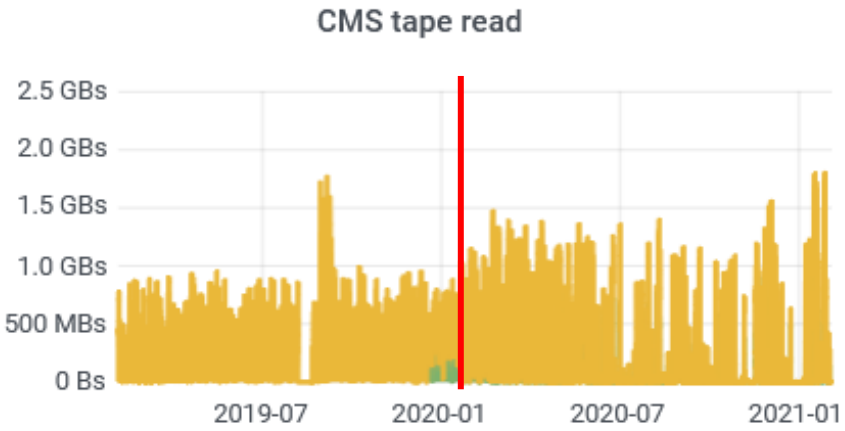
Data read second year (Feb20-Jan21): ATLAS 3.5 PB – All exp 13 PB

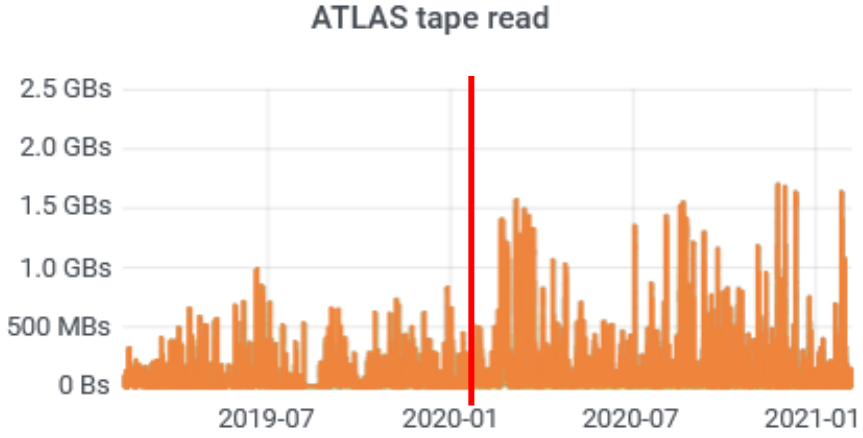
Recall throughput

Last 2 years recall thoughput CMS/ATLAS

Dynamic allocation in prod since 20 Jan 2020

All data in SL8500 library (16 drives)





Traditional vs dynamic allocation

Sample comparison: real CMS bulk recalls. Similar number of files and TB read

Traditional

Recall period: 18-23 Apr 2019

Duration: 138 hours

Number of files: 98k

Data read: 319.5 TB

Avg drives used: 3.7

Avg throughput: 650 MB/s

Dynamic

Recall period: 17-19 Jan 2021

Duration: 72 hours

Number of files: 92k

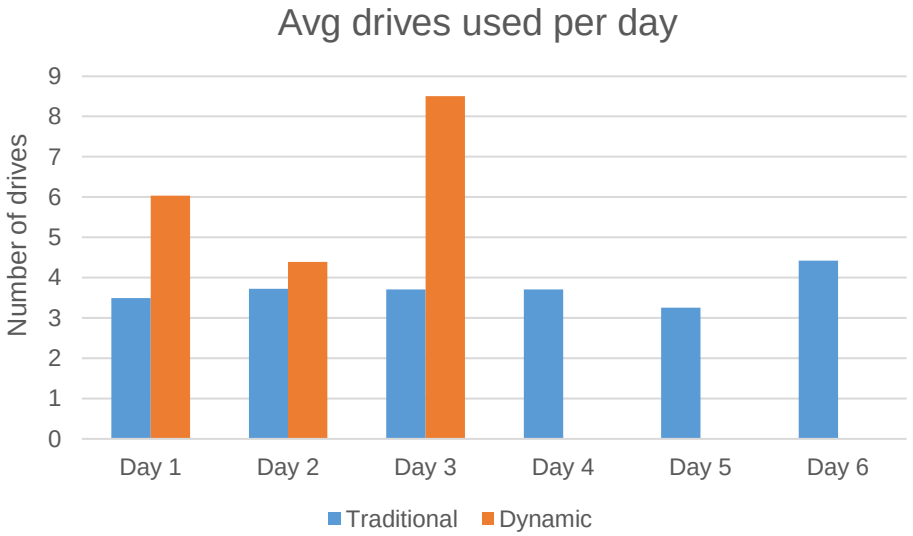
Data read: 313.5 TB

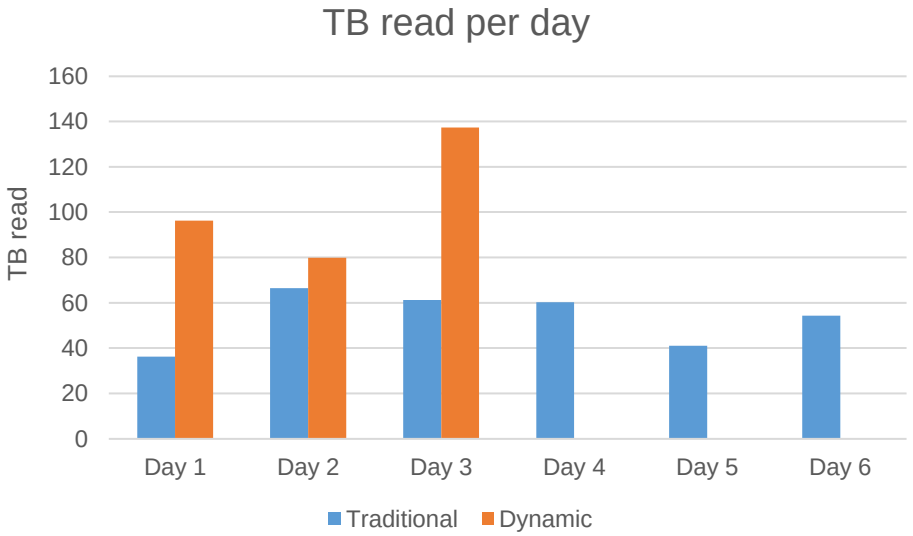
Avg drives used: 6.3

Avg throughput: 1.2 GB/s (+85%)

Traditional vs dynamic allocation

Sample comparison: CMS real bulk recalls. Similar number of files and TB read





Max drives used per day:

- 4.4 traditional vs 8.5 dynamic

Max TB read per day:

- 66.5 traditional vs 137 dynamic

Future works

- We will adapt this sytem to distinguish recalls involving tapes of different libraries
  - GEMSS is already able to put recall requests in different queues
- Optimization of migrations
  - Setting number of threads (i.e. tape drives) and number of files per thread on the basis of:
    - Available space on buffer
    - Number of files and amount of data to migrate
    - Number of free drives

Conclusions

- Dynamic drive allocation allows us to
  - Decrease users waiting time for recalls
  - Performing administrative tasks without interfering with production
- Compared to traditional allocation
  - Throughput peaks: 1 GB/s -> 1.8 GB/s
  - Data read per day peaks: 60 TB -> 100 TB
  - Thoughput improvement (sample comparison): 85%