Experience in dynamic tape drive allocation to manage scientific data

Tape facility at INFN CNAF

INFN CNAF provides storage resources for 4 LHC experiments (ALICE, ATLAS, CMS, LHCb) and ~30 non-LHC collaborations

- ✓ ~ 50 PB on disk
- ✓ ~ 90 PB on tape

Tape infrastructure components:

- ✓ **1 tape library Oracle-StorageTek SL8500**
- ✓ **1 tape library IBM-TS4500**
- ✓ **16 T10KD, 19 TS1160 tape drives** for scientific data
- ✓ **GEMSS** (Grid Enabled Mass Storage System) software developed by INFN that provides a full **HSM** (Hierarchical Storage Management) integration of:
  - ✓ **StoRM** (Storage Resource Manager): software released by INFN based on SRM (Storage Resource Management) interface to access storage resources
  - ✓ **IBM Spectrum Scale:** the disk storage software infrastructure
  - ✓ **ISP (IBM Spectrum Protect)** software: the tape system manager

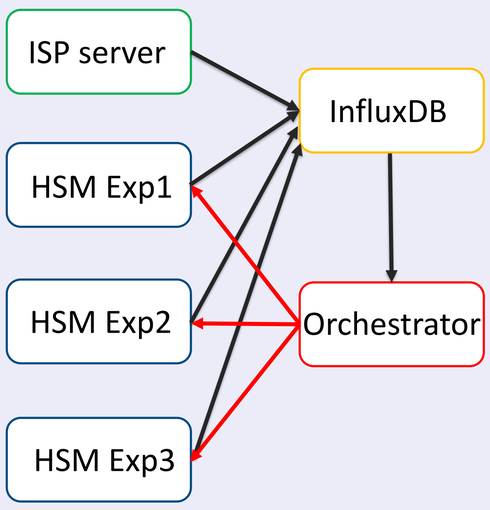
Traditional tape drive allocation

All tape drives are shared among experiments

Before 20 January 2020

- ✓ Each experiment could use a maximum number of drives for recalls or migrations, 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 drives which were free and instead could have been used by pending recall threads

Dynamic tape drive allocation



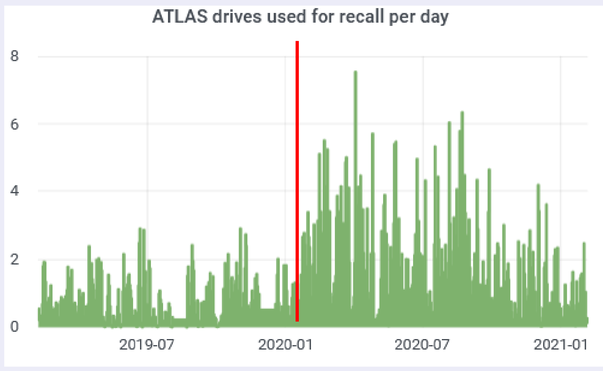
- ✓ Software solution to dynamically allocate drives to experiments for recalls
- ✓ InfluxDB stores monitoring information on:
  - ✓ number of free drives
  - ✓ number of recall threads running and number of pending recalls from each tape server (e.g HSM Exp1, Exp2, Exp3)
- ✓ Orchestrator:
  - ✓ performs 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 tape (HSM) server, to reach a maximum configurable value
  - ✓ can start reclamation processes when free drives are over a desired threshold

Dynamic allocation in production

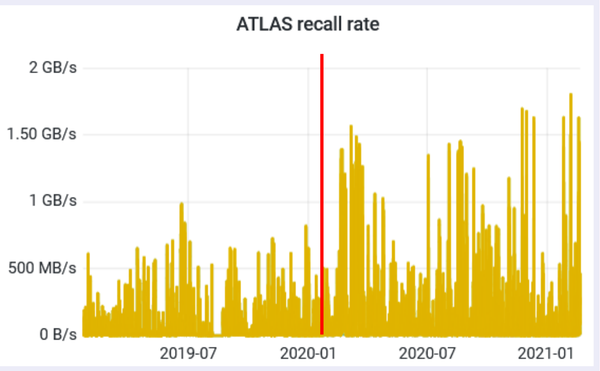
In production since 20 Jan 2020

ATLAS example

- ✓ 2 years period recalls (Feb19-Jan21)
- ✓ All data in SL8500 library (16 drives)
- ✓ Data read first year: ATLAS 1.3 PB – All exp 16 PB
- ✓ Data read second year: ATLAS 3.5 PB – All exp 13 PB



Drives used for recalls (average per day)



Recall rate (average per hour)

Traditional vs dynamic allocation

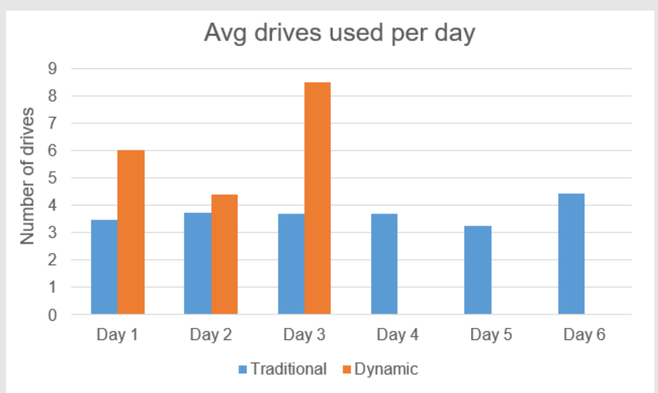
Sample comparison: real CMS bulk recalls. Similar number of files and TB read

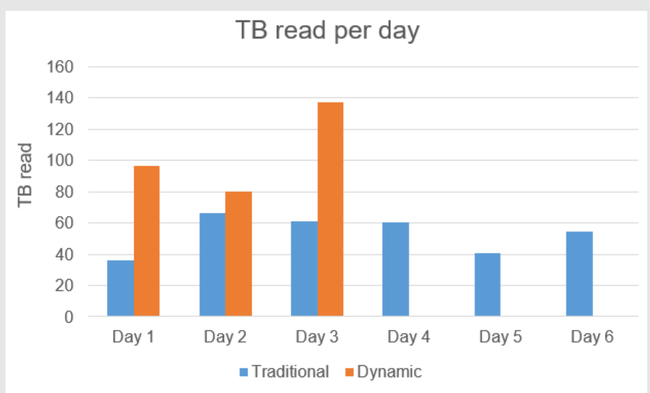
Traditional

- ✓ 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%)**





Highlights

- ✓ Dynamic drive allocation allows us to
  - ✓ Decrease users waiting time for recalls
  - ✓ Perform administrative tasks without interfering with production
- ✓ Compared to traditional allocation
  - ✓ Throughput peaks: 1 GB/s -> 1.8 GB/s
  - ✓ Data read per day peaks: 65 TB -> 135 TB
  - ✓ Throughput improvement (sample comparison): 85%
- ✓ Manage different recall queues for different sets of tape drives
  - ✓ By working with multiple libraries
- ✓ Possible extension to migration processes
  - ✓ Setting number of threads (i.e. tape drives) and number of files per thread considering:
    - ✓ Available space on buffer
    - ✓ Number of files and amount of data to migrate
    - ✓ Number of free drives

A. Cavalli, D. Cesini, A. Falabella, E. Fattibene, F. Fornari, L. Morganti, A. Prosperini, V. Sapunenko
INFN-CNAF Bologna (Italy)
enrico.fattibene@cnaf.infn.it