CMS storage usage

- Basic concepts
- A few examples
- Use cases can not be worked out in detail in advance of usage
- Discussion internal to CMS is starting
- CMS contact persons at Tier1 will be part of that discussion, as it always in the past

Disk and Tape

- CMS Computing Model and Computing TDR talk about disk requirements and tape requirements
- They were never meant to be separate worlds

- The distinction is mostly practical and based on experience and history
  - Some data need “backup” (RAW)
  - Some data needs fast access (AOD)
  - Some data is not accessed frequently (older RECO versions)
  - Tape still less expensive then disk when data integrity has to be guaranteed
  - We do not predicate success on WLCG being able to keep al data on disk

Why tape

- Only data on tape can be thought to be “safe”
- Tape provides cheap disk overflow for unfrequently used data

- Hence Tier2 (no tape) disks are only caches

- Hence Tier1 share custodial (safekeeping) responsibility for RAW and SIM data and keep tape copy of data that are expensive to reproduce (RECO, AOD)

- The need to get important data from another site’s tape is another expensive operation. Hence some important data may be replicated on tape locally at many sites (previous AOD versions)

Why disk

- Ideally all data is safe on tape, and only cached on disk to "bring it closer to CPU". With exceptions…
  - Intermediate production output (delete tomorrow)
  - Not-Validated-Yet datasets (possibly delete next week)
  - Transit buffers for impedance matching (WAN-tape-WAN)
  - User data

- In practice this cache has to be large

- Hence computing model has some arbitrary assumptions on what is on disk, what is on tape, in order to give guidance to sites and funding agencies about cost and deployment issues
  - It low-latency access to tape would exist, disk could be small

Data movement

- Data goes on tape
  - To be safe
  - Because there is no other place (simulation output, previous reconstruction versions)

- Data comes back on disk
  - To be processed

- Tape data may also need to be exported, via disk pool

- Making detailed examples is of little utility
  - Design Tier1 for reasonable rates
  - Usage will adapt and fit
  - Read patterns are very difficult to predict

Examples

- Will try to work out some numbers from Computing TDR
- Aimed at show disk/tape rates and cache needs

- Very preliminar exercise

- Do not be surprised to see these numbers changes as we better understand, and change again as we hit reality

- Computing Model and Computing TDR are not a day-to-day operation guidebook. Things will be different

- Get involved in our discussion

Nominal rates: write

- All data that enters a Tier1 go on tape
  - But must be immediately accessible on disk as well
  - Also more data goes the same way

- Tier0 output RAW-RECO + AOD (250+50 MB/sec)
  - To each T1 1/7th of RAW-REC + all AOD ➔ 45 MB/sec)
  - Simulation from Tier2 ➔ 20MB/sec
  - Re-processing, new RECO+new AOD ➔ 30MB/sec
    - ☞ One pass in a month
  - From other T1's re-reco versions of AOD ➔ 150MB/sec peak
- Tier1 writes to tape at 100MB/sec flat. Peak at 250
  - Recover from backlog etc.

- Large portion of big disk pool is dynamic
  - New version comes in
  - Old version goes out (stays on tape)

Nominal rates: read

- Reprocessing of MC events, one pass 4 month ➔ 50MB/sec
  - Need to go faster ? Alternate, not overlap, with RAW reco
- Access previous version of a dataset (access RECO) ➔ 20 MB/sec
  - 10TB in a week
- Access to SIM-RECO data (only 10% on disk) ➔ 20MB/sec
  - 10% data in a week = 10 TB/week

- Multiple such activiy cohexist
  - Ten groups make one pass each to all data each month, but who uses previous version ? Who uses simulation ?
  - Usage can adapt to resources, limiting freedom to access data
  - Why one week ?

- Tier1 reads tape at ~100MB/sec (large uncertainties)
  - E.g. if money gets really short, may need to put some RAW on tape, and read them back

Disk caches

- There is no explicit disk cache sizing for access to tape data in CMS Computing Model
- But, we indicated the Tier1 as places of organize activity
  - Non-cahotic access could fit small (TB's) cache
    - ☞ 5GB files, 1TB can feed 100 processors as long as data in/out is well synchronized with CPU activity
    - ☞ would e.g. apply to large reprocessing
  - For user/group access to one archived dataset prestage-and-lock of the full dataset may be easier: few TB per dataset
  - A lot will depend on how good our data/workflow tools can be, including capacity to properly schedule CPU
- Data access can be easily organized by fileblocks: 1~2 TB
- Idea was that several 10's TB cache is not worth spelling out separately in front of a 1.2 PB disk for a nominal Tier1

- Trivial statement: more efficient stage-in/out = smaller cache

Storage classes

- We have a clear case for Disk1Tape0 as (small) scratch space
  - Not quantified in CMS Computing Model
- We had to call all our data Disk0Tape1 (megatable)
  - Active tape, *not* backup-and-never-retrieve
  - Tier0 data that goes to Tape must also be accessible from disk
  - The computing model indicates the size of a large disk cache to avoid cache miss for the most frequent access:
    - raw data, most recent version of RECO and AOD

- Tape read rate at least same as write one (reprocessing)
  - Larger read needs possible

- Access to tape must be efficient: here you set the requirements
  - A lof of data move in/out of tape and disk, good to hand to dedicated system, rather then Tier1 depend on CMS DM
  - Data should be organized on tape so that processing a consistent chunk of data (all RECO of a dataset) makes minimum tape mounts