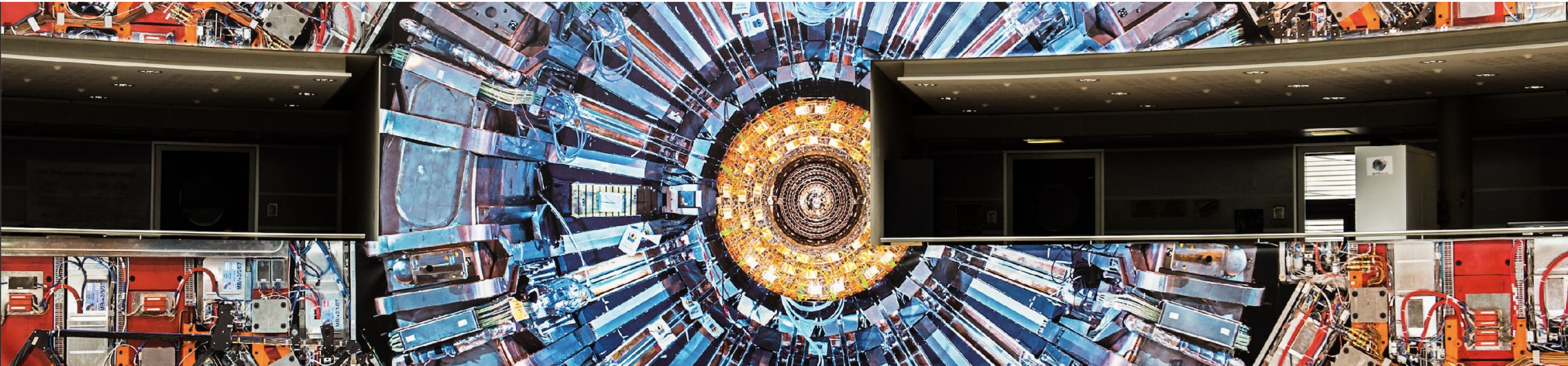
Storage in CMS: Status and Plans

J. Letts (UCSD), D. Piparo (CERN) - WLCG/HSF Workshop - 19-NOV-2020



Introducing CMS Storage

- **Disk space: 172.1 PB***
  - Disk space is allocated for central operations and physics analysis activities:
    - e.g. buffers for operations, storage for active set of analysis data, and part of the AOD set
  - High quality custodial storage and JBODs used for caches
- **Tape: 319 PB provided by 7 Tier-1s and the Tier-0***
  - e.g. custodial storage for RAW, archival for AOD
- CMS, the sites, and other groups are already now exploring new ways of deploying storage:
  - e.g. caches, data lakes - more in the later slides

** 2020 Pledge, from CRIC*

This talk

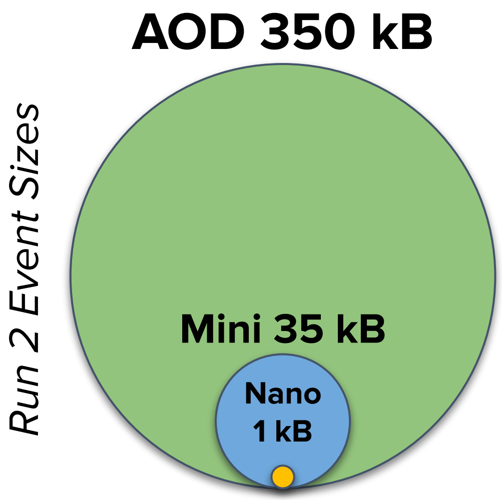
- Access Patterns
  - Reducing storage needs with innovative analysis formats and active network usage
  - Caches
  - Possible QoSes
- Storage-less sites
- Data Lake
- Storage-intensive workflows at HPCs
- SRM and tokens
- Third party copy

Access Patterns and Implications on Future Storage

- **Phase-2 access patterns: no fundamental changes wrt today**
  - Potential exception: innovative patterns to speed up analysis at analysis facilities (AF)
  - Careful assessment needed as AF prototypes become more realistic: is it affordable? How intrinsically fast is a realistic analysis application?
- **The new scale will have profound implications for storage. Cost may be driving**:
  - A more active usage of archival storage for all formats but the analysis ones, e.g. cold storage used not only for long term custodial storage
  - An even more active usage of network and less replication
  - More integrated central data processing workflows, e.g. better control of input/output files on staging spaces

Small Analysis Formats

- **Two small analysis formats, MiniAOD (35 kB/evt) and NanoAOD (1-2 kB/evt)**
  - One single central flavour of Mini and Nano, content centrally managed
  - Persistified models: Mini - OO, Nano - Fundamental types, arrays thereof
  - Mini: in production, adopted for all analyses in Run 2 (except very specific detector studies needing full AOD)
  - Nano: in production, 30% of analyses adopted it. Target is 50% by the beginning of Run 4, hopefully more.
  - **Run 3: mix of Mini and Nano, increasing portion of analyses relying on Nano towards Run 4**
- Future AFs: prototyping and R&D is ongoing
  - AF architecture, relation with the Lake, adequate storage layer
  - Software: no explicit (event) loops, optimisations behind the scenes, plug into data science tools, thread/process based parallelism



Reducing Storage Needs with Network

Today, in production, CMS reduces the storage needs balancing network and storage usage.

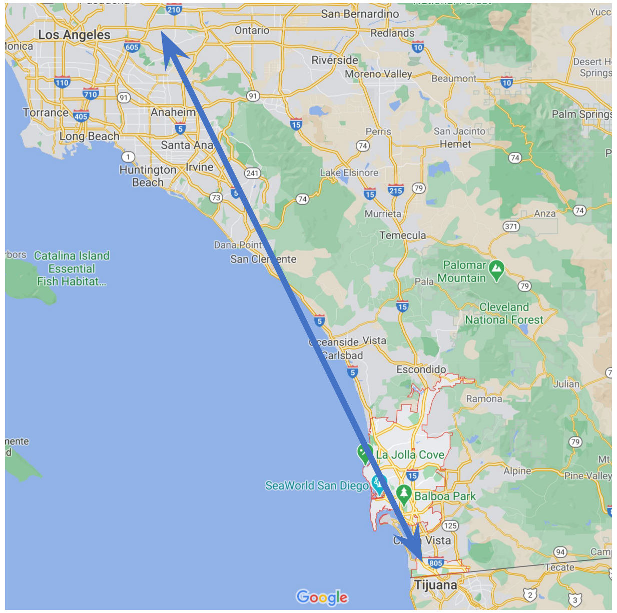
1. Remote reads via AAA (Any data, Anytime, Anywhere), user and production jobs
   - Generic XRootD service, redirecting file opening
   - Possible thanks to CMSSW and ROOT developments
   - More efficient than a full copy, only desired/needed columns (branches) read
2. *Premixing*: a strategy of simulating pileup. Overlaying 1 “pileup only” event from a big library onto a hard scatter event instead of N minbias events (*Classical Mixing*).
   - Typically place PU libraries at FNAL and CERN
   - Run mixing at Tier-X centres and read remotely the PU events.
   - Reduces LAN network bandwidth and CPU needs wrt to classical mixing, trading off WAN bandwidth.
     - CMS is interested in active network management.
3. Caches: a strategy to provide data to process/analyse allowing to save operations/storage cost
   - Custodial storage is somewhere else
   - Example use cases: multiple sites (E.g. Tier-2s), storageless sites (e.g. HPC)

- **Plan to rely on 1, 2 and 3 during Run 3**
- 1. and 3. Potentially folded into data lake implementations in Phase-2
- CMS expects to gain more quantitative experience over time

Disk storage needs can be reduced with an active use of network

Caches

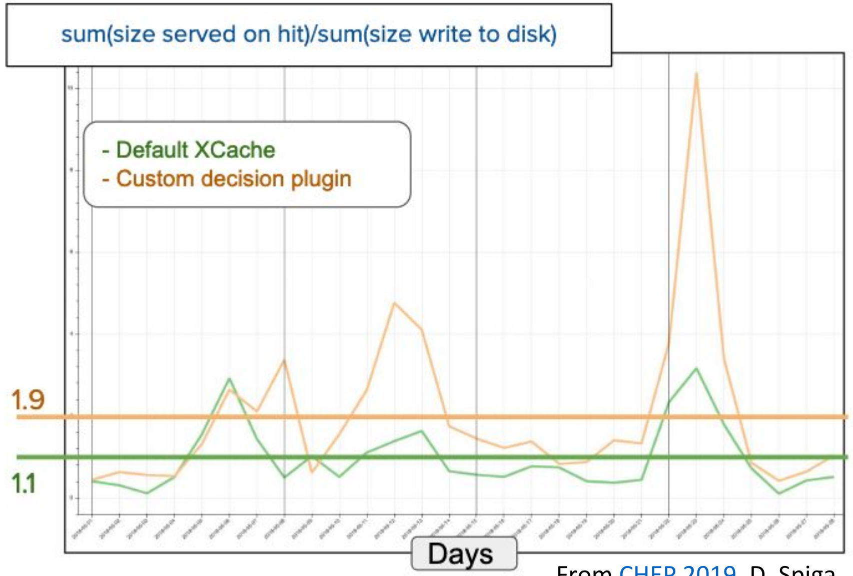
- Caches are a reality in CMS, since a while, e.g. XCache implementation
  - Can be orchestrated via k8s
  - Rely on inexpensive JBODs
  - Easy to build on-demand
- Examples in production:
  - UCSD + Caltech, merged namespace
  - SSD data cache @CNAF to support I/O of CINECA
- R&D: INFN distributed cache model
  - Perfectioning the approach: smart decisions about what to evict, also using ML techniques
  - Not CMS specific
- CMS plans to increase its experience with caches during Run 3 to be ready for Phase-2
  - Started to profit from them already now!



From CHEP 2019, E. Fajardo (Google Maps)

UCSD-Caltech link:

- **120 Miles**
- **100 gbps**
- **< 3ms**



From CHEP 2019, D. Spiga

Potential Future Storage Implementations

- Presently CMS relies on three QoSes, materialised in tape, disk, and caches
  - Custodial archival storage - tape
  - Custodial storage holding analysis/production datasets - disk (usually replicated at the filesystem level)
  - Inexpensive, non-redundant storage to read analysis/production datasets - caches
    - Watching the market closely: pricing for solid state devices could become attractive
  - User space - disk (often replicated >2x at the filesystem level)
- Implementations currently considered for Phase-2:
  1. **Custodial archival storage** - what today we address with tape
  2. ***Data Origin* space inside the lake providing immediate access**: the location where data arrives at that is meant to remain immediately accessible for file open without delays.
  3. ***Data Origin* space inside the data lake NOT providing immediate access:** e.g. for optimising costs with cold storage
  4. **Transient caches and buffers** - JBODs
  5. **User space for analysis data**
  6. **User space for custom analysis data**, not backed up. (Almost) exclusively analysis non-OO data formats
  7. **Space for random access of columnar data** that is fundamentally transient - JBODs?
- Need to acquire experience in this area to further refine categories.

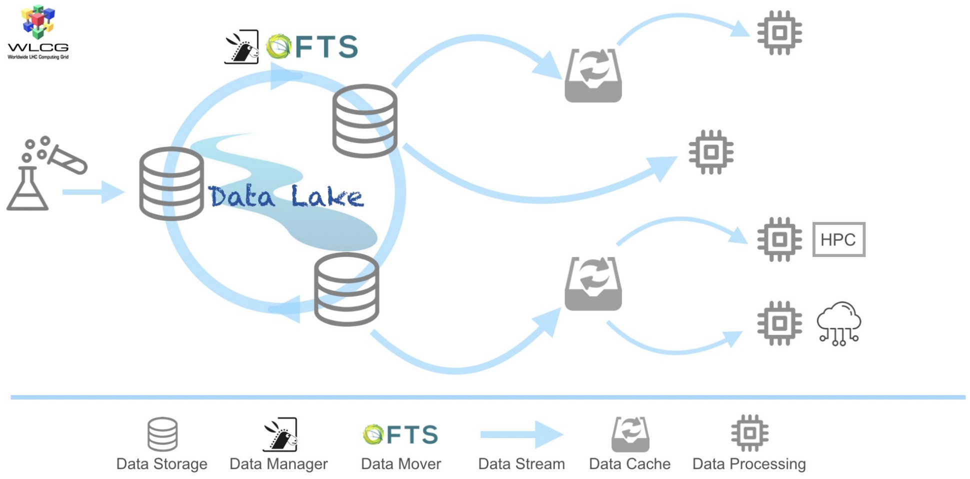
Exploring new QoSes to accommodate future workflows optimising costs

Storage-less Sites

CMS's position is sufficiently well described in the WLCG DOMA document, *Access* section

Phase-2: Lake model

- Storage is not necessarily co-located with compute in the Lake model.
- Content delivery network capable of caching and minimising impact of network latency.



Xavier Espinal (CERN)

Run 3: Towards the Lake approach

- CMS became very tolerant about site configuration/design
  - Don't refuse reasonably usable resources: expand the resource base as much as possible
  - Substantial progress made in the last 2 years, e.g. exploitation HPCs for full spectrum of production workflows, from generation, simulation, digitisation-mixing, reconstruction and (partially) analysis
  - Prepare for the scenario where we do not control and can partially influence the configuration of some sites
- In this context, caches are key

Lakes and their interface

- One lake per region (continent?)
  - Different lakes may make different choices
- **CMS expects the Lake defined by its interface**
  - e.g. to express the different QoSes available, reduce overall cost
- **With a clear interface, the need for knowledge of the internals would vanish**
  - Leave room to Lake devoperators/architect to re-organise internal structure
  - Easier to accommodate evolutions in storage technology, different internal configurations, new QoSes

Workflows with Heavy Storage Requirements @ HPCs

- **CMS runs all data processing steps at HPCs**
  - Gen, Sim, Digi, Mix, Reco.
  - Analysis: making progress
- Work is necessary to commission new machines, e.g.
  - Lack of common transfer protocol interface for HPCs globally
  - Unwillingness of centres to support 3rd party copy via Dav
  - Direct access to files in their WAN-reachable storage from the WNs
- **Caches as edge services do play a role**
  - Experience so far very promising, e.g. CINECA
  - Possible further optimisations, e.g. capability to manage caching policies
- Run 4 HPCs might significantly contribute to LHC and CMS processing needs
  - Challenge to move exabytes of RAW data to these sites for processing

SRM and Tokens

SRM:

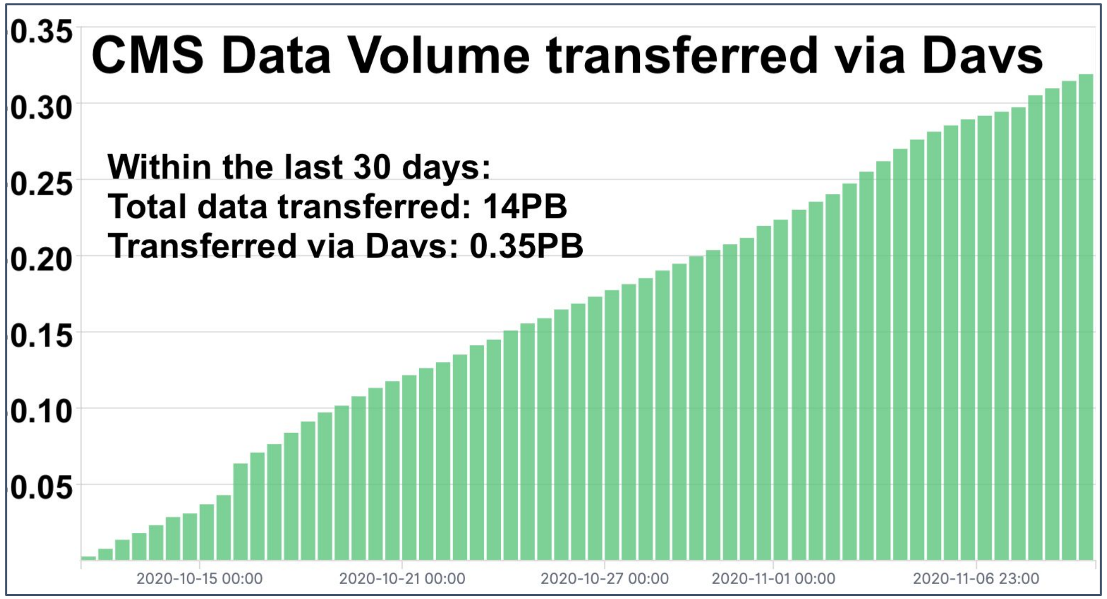
- As long as SRM service is supported, no reason for sites to switch to gridftp/gsiftp
- Potential issue on tape endpoints
  - Transition to CTA at Tier-0 and RAL good news in this sense

Token-based authentication:

- Ready to fully engage with OSG, WLCG, and EGI on a timescale for transitioning to tokens.
  - Only schedule presently known from OSG
  - Ongoing work also in the data lake prototype within ESCAPE
- Opportunity for WLCG and EGI to agree on a schedule?
- Development IAM instance for CMS already exists see this talk for more details
- Grid Middleware that CMS uses (HTCondor & GlideinWMS) has integrated token-based authentication, although we are still using x509-based authentication.
- Many CMS sites have integrated xRootD third-party copy (see next slide) as a step on the path to migrate away from GridFTP.

Third-Party Copy

- Substantial work ongoing and much progress made in the area of HTTP-TPC
  - Risked to be blocked next month when we migrate DM by a Rucio bug, fixed now!
- 13 sites in Production (dCache, XRootD, StoRM) using HTTP-TPC
  - T1s: FNAL, KIT, JINR, IN2P3, and CNAF (5/7 CMS T1 sites!)
  - T2s: DESY, MIT, Florida, Caltech, Wisconsin, Nebraska, Purdue, UCSD
  - Others already making progress, e.g. CSCS, PIC, TIFR, UCL, Brunel, London_IC
- Production (PhEDEx) traffic over davs:
  - 2.5% of overall traffic
  - Up to 40% in and out of Nebraska & UCSD (two big T2s)


CMS Data Volume transferred via Davs

Within the last 30 days:
Total data transferred: 14PB
Transferred via Davs: 0.35PB

Diego Davila

More @ CHEP 2019

Conclusions

- **Access patterns will not dramatically change in Run 3, the scale will have a serious impact on future storage**
- Approaches to reduce storage needs already in production, e.g.
  - Network management
  - “Kilobyte per event”- range analysis data formats
  - Caches of JBODs
- **Use them in Run 3 to acquire expertise in view of Run 4**
- Substantial progress with TPC, token-based auth, usage of storage-less sites
- Plenty of opportunities to learn together with other experiments
- Interested in active network management