Xcache Initiatives and experiences

Pre-GDB Meeting

June 8th 2019

Recap of what to expect for the HL-LHC from CMS

Start with CMS Data Formats and their expected use

Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

Primary Processing:
RAW -> AOD -> Mini -> Nano

Another way of looking at it:
80+160 Billion events/year (Data+MC) = 240B events/year
$\Rightarrow$ 7.4MB x 8e10 ~ 6e11 MB ~ 0.5 Exabytes/year of RAW
$\Rightarrow$ 2.0MB x 2.4e11 ~ 5e11 MB ~ 0.5 Exabytes/year of AOD
$\Rightarrow$ 0.2MB x 2.4e11 ~ 0.5e11 MB ~ 50 Petabytes/year of Mini
$\Rightarrow$ 0.004MB x 2.4e11 ~ 0.01e11 MB ~ 1 Petabyte/year of Nano

Data formats span x1000 in size per event.
Files in large data formats are touched at most twice a year.

Buffers & Caches

- It seems very likely that the size and expected use of data at HL-LHC motivates a clear distinction between disk use for “buffers” and “caches”.
- All processing done via buffers
- All analysis done via caches

Buffer

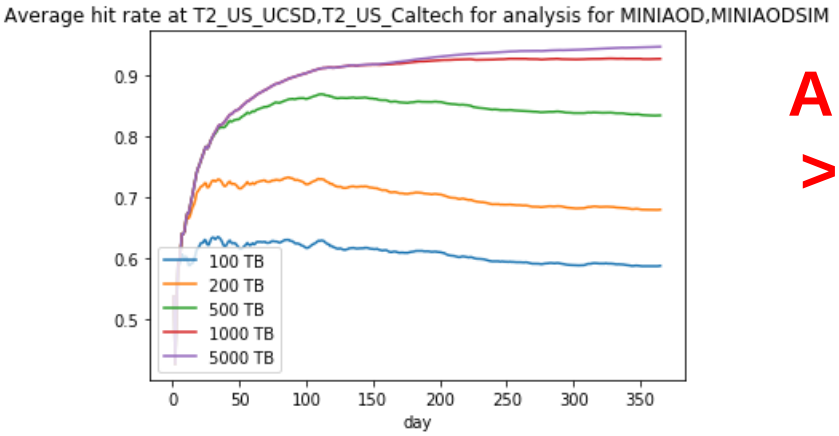
- Managed as an integral part of workflows from tape recall to tape store of output.
- Minimize disk for buffers by tight integration of processing and buffer management
  - Technical challenges lie in workflow integration with data handling.
- Processing requires lots of CPU and very little disk space because reconstruction is slow.
- Special case:
  - Re-making MINIAOD from AOD is a processing step that is entirely tape recall bandwidth limited.
    - Again, modest disk space requirement for this buffer.

Caches

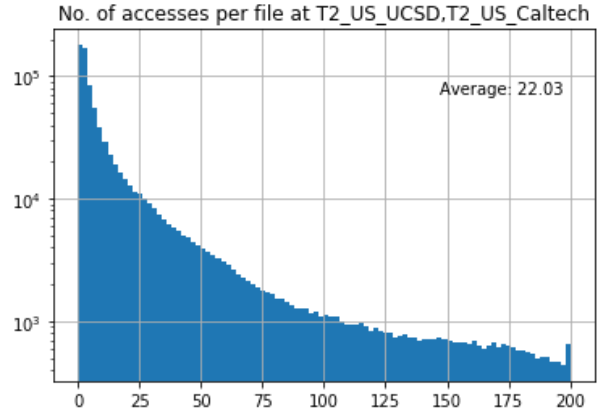
- Analysis use of data is:
  - Heavy reuse of the same data by many people
  - Transient in that data is versioned and has a life cycle going in and out of “fashion”.
  - Dominated by MINI and NANO, and thus modest in size.
- **A priori, this is a perfect use case for “caching”.**
  - How exactly cache misses are handled is a detail we’ll get back to later.

Cache “Simulation”

- Andrea Sciaba analyzed the CRAB use of data for 2018. (See backup for details of methodology)



A 1 Petabyte cache leads to >90% hit rate in SoCal T2s.



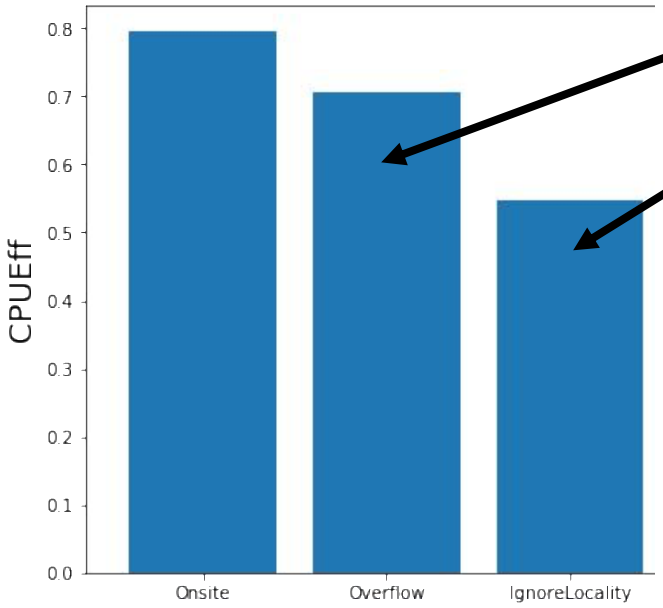
Distribution of # of times a file is read peaks at 0 and has an average of 22.

Very wide distribution !

Remote Reads

- CMS invested a lot of effort into optimizing our IO stack such that we can do remote reads for analysis without significant loss of CPU efficiency.

Details of Ciangottini's study.



Overflow = Regional remote read in US or Italy

IgnoreLocality = global remote read

Caches can span regions to maximize CPU per disk.

Distances in EU

Good goal to set for IO stack to be sufficiently latency tolerant to lose less than 10% in CPU time for access distances of 500-1000 Miles.

Risk Assessment

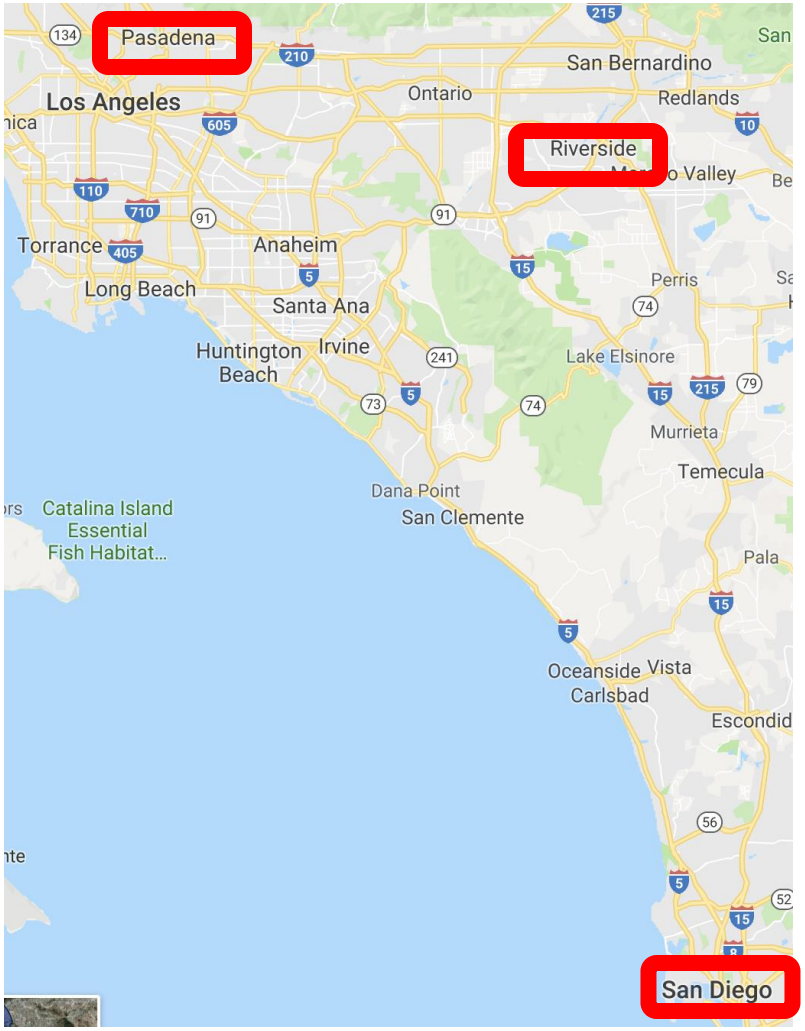
- In Run2, the average IO rate for analysis with CRAB in CMS was very low.
  - All analysis execution was single threaded.
  - Average less than 10Hz per batch slot for event sizes of 400kB/event.
- There are discussions about “Analysis Facilities” that could radically change the IO requirements for analysis.
  - Much higher IO, much more latency sensitive, making distributed caches impractical.

Southern California

Regional Perspective of T2s at Caltech and UCSD.

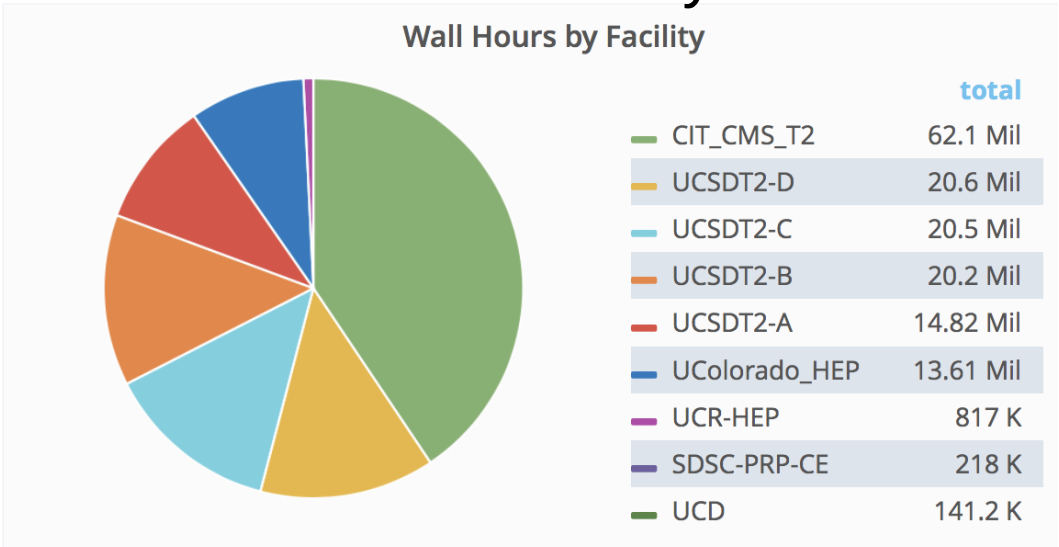
Both are CMS only sites.

Southern California (SoCal)



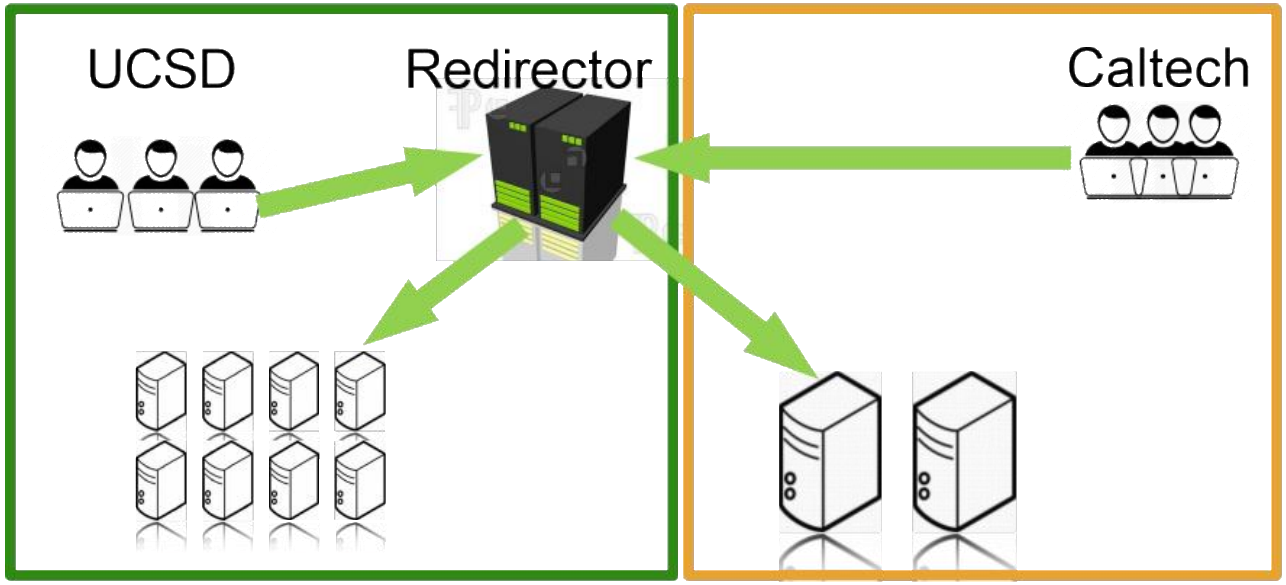
T2_US_UCSD
T2_US_Caltech
T3_US_UCR
… and maybe some more …

Wall hours last year



SDSC has ~70k x86 cores in house.
NERSC to San Diego ~ 500 miles.

SoCal XRootD Cache



	UCSD	Caltech
Nodes	11 (10 more coming)	2
Disk Capacity per node	12x2TB = 24TB	30x6TB (HGST Ultrastar 7K600)
Network Card per node	10 Gbps	40 Gbps
Total Disk Capacity	264 TB	360TB

Last Year

In process to expand cache space to 1Petabyte.

Caching all of MINI and NANO, data and sim.

T2 wish list (I)

- Want CMS to switch to Buffer & Cache mode.
  - Buffer that assumes nothing in buffer needs to stay there for longer than a week, to keep buffer small.
- Want to operate only JBODs
- Want CMS to be responsible for dealing with data losses due to disk losses.

Overall, want to decrease total cost of ownership.

T2 wish list (II)

- All CPU in CA to benefit from SoCal disks.
  - Why replicate if data can be accessed via the network?
  - Caltech/UCSD/UCR and maybe HPC at SDSC and NERSC could all read our cache directly?
- Want Xrootd cache software that works well
  - Is operationally stable & performs well enough
  - Accounting that I can use to understand system performance & use.

Overall, want to spend T2 funds on increasing events/sec we can process. We think this means buying more CPU less disk.

Handling Cache Misses?

- Not my problem as a site operator.
- Am willing to start with what XRootD provides out of the box.
  - Gain experience with model of less disk per site.
  - Disk bought now will still be in use by HL-LHC !!!
- Adjust later to more clever schemes.
  - Have Rucio manage cache misses via XRootD plug-in ?
    - Have block replication scheme as proposed by Lammel talk in DOMA Access on Ja. 29th 2019.

Analysis Facility

- Any HDD we buy today, could be obsolete by HL-LHC if the experiment switches to analysis facility based on NVME deployments of storage & NANO.
- This seems to also argue for being conservative with our HDD purchases going forward.