“Data Lake”
A Site Perspective

Frank Würthwein
SDSC/UCSD

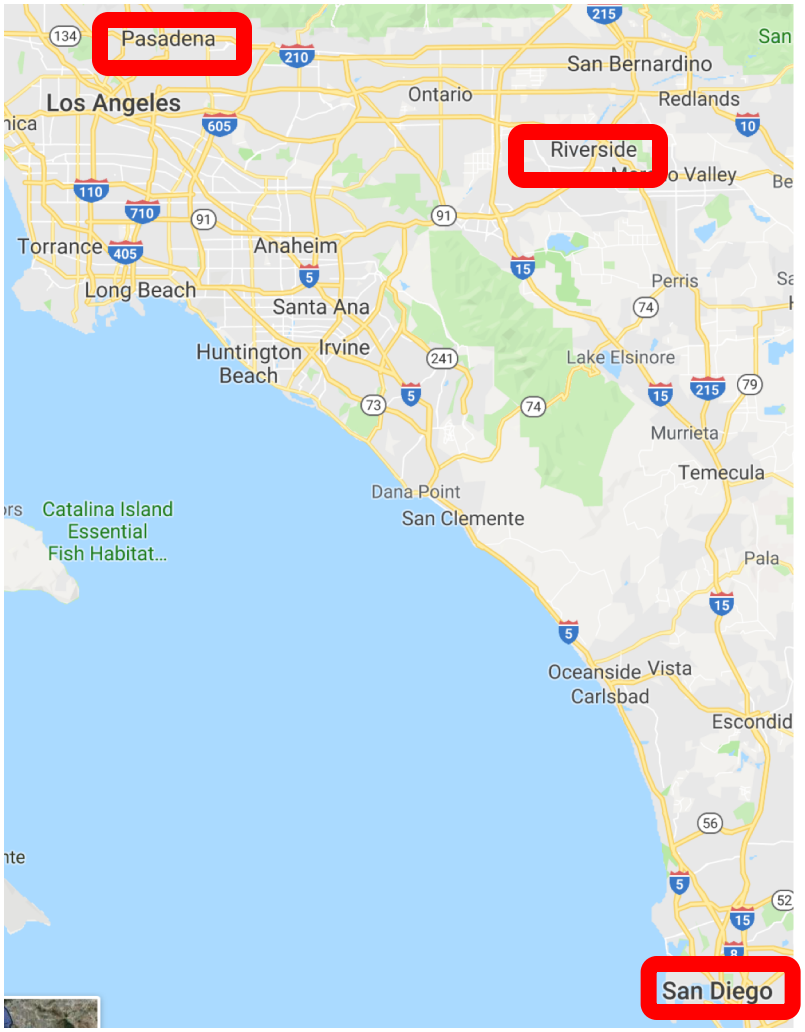
DOMA Access

June 4th 2018

Purpose of this talk

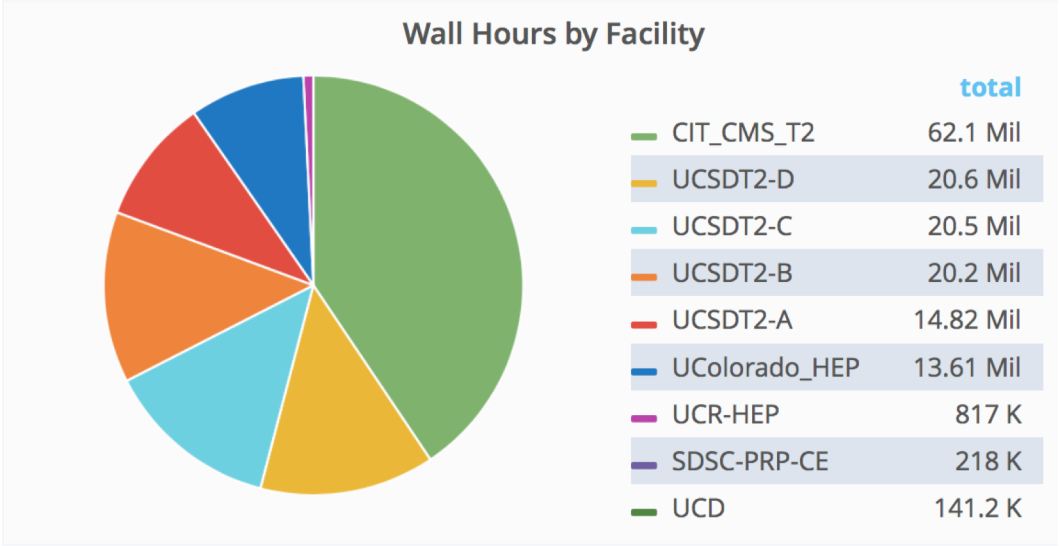
- There was a lively debate during the HOW DOMA session when Xavier presented the “Straw Proposal”
- It seemed that people would benefit from a presentation that looks at this entirely from the perspective of a site like UCSD, or a region like the South Western United States.
- So my purpose today is to present such a perspective.

Define Southwest of US



T2_US_UCSD
T2_US_Caltech
T3_US_UCR
… and maybe some more …

Wall hours last year



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

Start with 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.

Expectations for our T2

- There is some disk space we need to offer as a buffer space for processing.
  - All primary processing will be driven centrally in CMS.
  - If CMS has its act together then this ought to be tightly connected with the workflows to avoid wasting buffer space at T2.
    - Bring input data from archive, move it to T2, process it, delete input data, move output back to archive.
    - Time of data in buffer at T2 should be less than a week.
- There is some disk space that will be reused heavily for data analysis.
  - We can understand reuse patterns based on Run2 patterns of use.

Use of disk in Run2 (I)

- UCSD deploys HDFS with replica == 2
  - Every file is sliced into 128MB blocks
    - Every 2GB file will get distributed across 16 servers.
  - We have 182 servers
    - If we loose any two disks across two different servers in a power outage, we corrupt many files.
    - We often loose more than 2 disks in a power outage.
- **We can not afford replication high enough to not worry about data corruption after power outages.**
- 1 PB RAW translates into 450TB useable, and still we are dealing with corruptions after every power outage.

Use of Disk Space Run2 (II)

MINIAOD and MINIAODSIM in XCache

	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

Datasets	Size (TB)
/*/Run2016*-03Feb2017*/MINIAOD	182.8
/*/RunIISummer16MiniAODv2-PUMoriond17_80X_*/MINIAODSIM	502.5
/*/*RunIIFall17MiniAODv2*/MINIAODSIM	211
/*/*-31Mar2018*/MINIAOD	137.9
Total	1041

Use ~620TB of
RAW Disk to cache
1Petabyte of logical data

In comparison, UCSD alone has 2.4PB RAW disk used with replications controlled by PhEDEx that serves mostly AOD that is little used.

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 spend the T2 hardware funds on our ability to increase the total Hz rate of events we can provide processing for. We think this means buying more CPU and less disk.

T2 wish list (II)

- Want all CPU that is close to us in terms of CMS application latency tolerance to benefit from our disks.
  - Why replicate if data can be accessed via the network?
  - Caltech/UCSD/UCR and maybe U.Colorado can share each others disk.
    - UCR and U.Colorado don't need to manage disk at all maybe?
    - I'd rather spend hardware funds on CPU to increase the Hz processing rate we have collectively available.

How do deal with 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.
- Adjust later to more clever schemes.
  - Have Rucio manage cache misses via XRootD plug-in as ATLAS proposes.
    - Have block replication scheme as proposed by Lammel talk in DOMA Access.

Implied R&D Items (I)

- XRootD xcache software that works well
  - is operationally stable.
  - performs well enough.
  - has good accounting of performance that I can easily access for my site.
    - Working set, total reads, average re-read, and all of that for total and per namespace subgrouping.
    - Exit code accounting so I see what jobs fail at my site that access the cache.

The above will lead to low cost of ownership

Implied R&D Items (II)

- CMS would need to shape up with its production processing capabilities.
  - Better integration of tape archive
  - Data spends less time in buffer at T2
- CMS should work long term on better cache miss handling, and better integration of job placement given knowledge of cache content.

Summary & Conclusions

- I clearly see advantages for my T2 operations from the Data Lake straw proposal presented at DOMA session at HOW.
  - Less operational burden.
  - Less money spent on disk that is rarely accessed.
- We can get started immediately with existing Xcache software.
  - Slowly increase the money spend on CPU vs disk, thus reversing the opposite trend.
- There is a lot that can be improved going forward.
  - Smarter treatment of cache misses.
  - Smarter placement of jobs given knowledge of cache content.
  - Better production workflows such that data spends less time in T2 buffers.

Use Run3 to put the improvements in place over time!