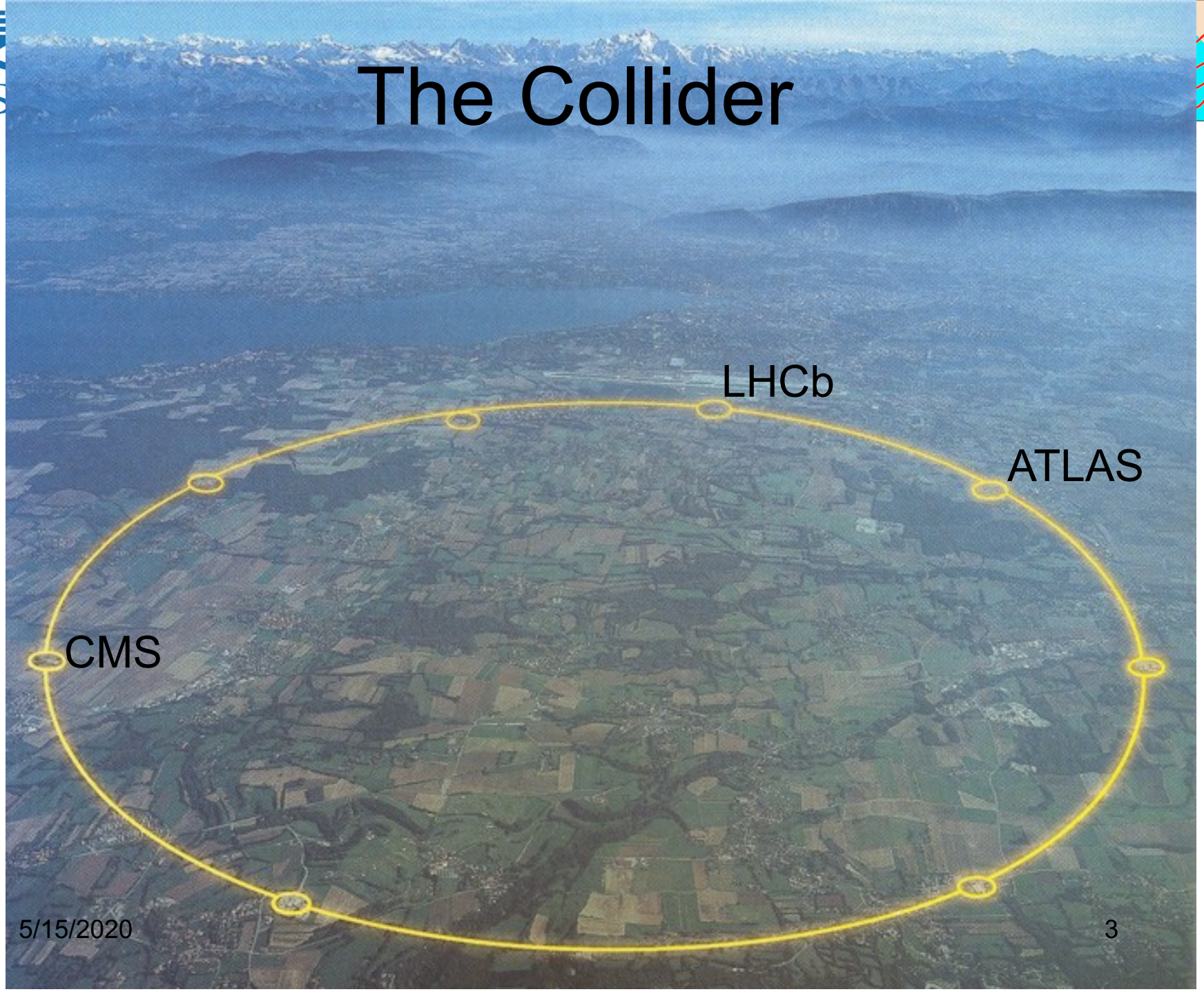
Data use by the CMS experiment at the LHC

Frank Würthwein
SDSC/UCSD
ESNet Seminar May 15th, 2020

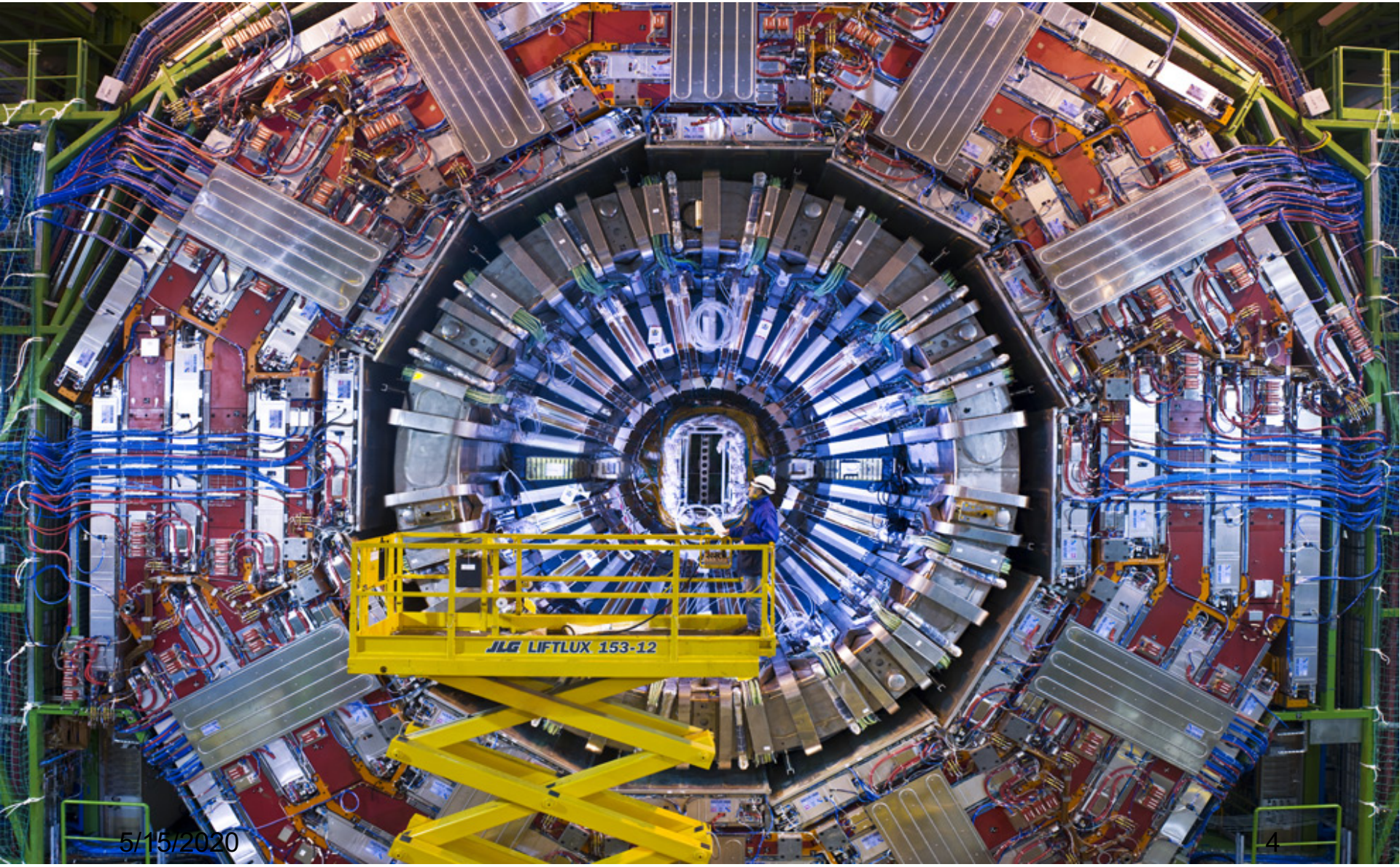
Disclaimer:

Any Opinions expressed here should not be misinterpreted as official opinions of the CMS collaboration.

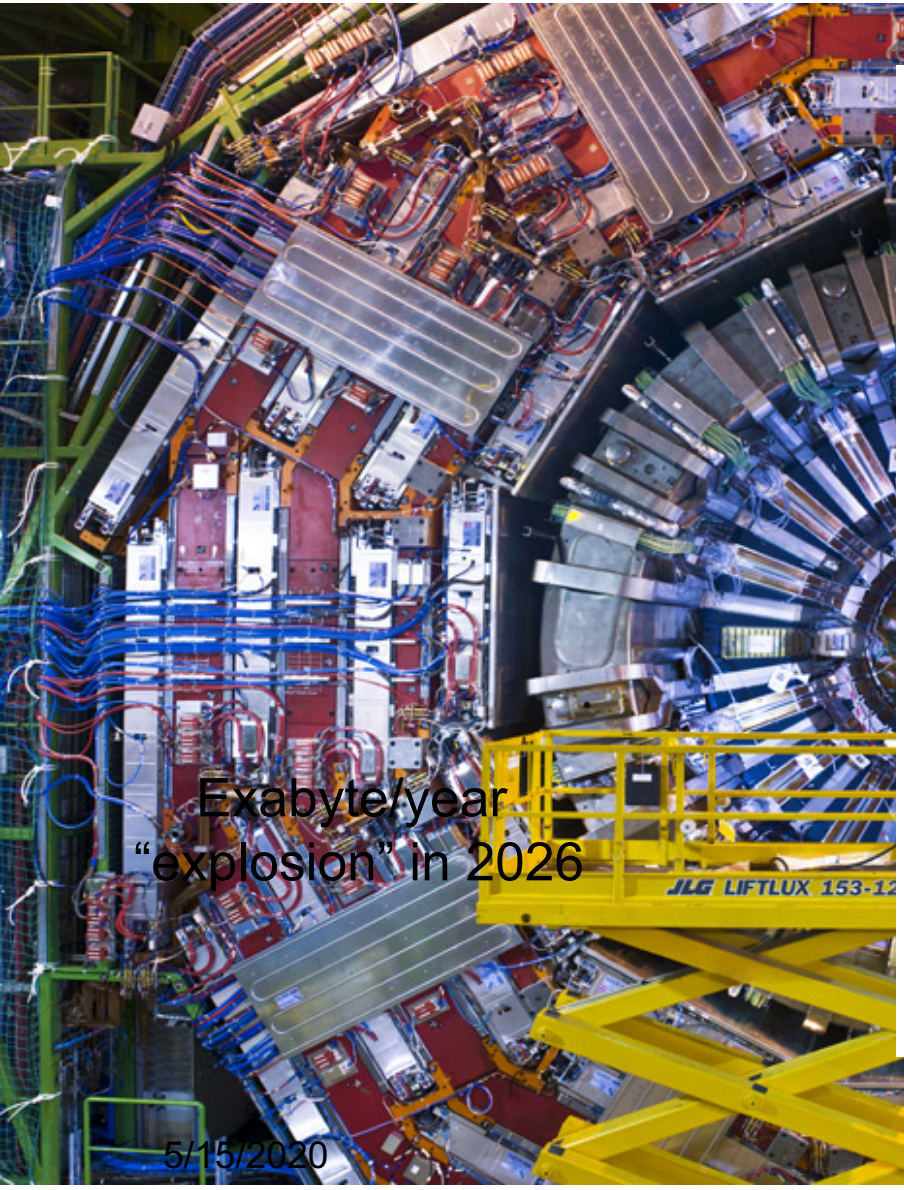
The Collider



The CMS Experiment (R-Φ view)



The CMS Experiment



Exabyte/year "explosion" in 2026

- **80 Million electronic channels**
  - x 4 bytes
  - x 40MHz
  - ---
  - ~ **10 Petabytes/sec** of information
  - x 1/1000 zero-suppression
  - x 1/100,000 online event filtering
  - ---
  - ~ 1000 Megabytes/sec raw data to tape
  - **~10 Petabytes of raw data per year written to tape, not counting simulations.**
- 4000 Scientists (1200 Ph.D. in physics)
  - ~ 200 Institutions
  - more than 40 countries
- 12,500 tons, 21m long, 16m diameter

Annual Data Volume

	# of collissions	# of events simulated	RAW event size [MB]	AOD event size [MB]	Total per year [PB]
Today	9 Billion	22 Billion	0.9	0.35	~20
HL-LHC	56 Billion	64 Billion	6.5	2	~600

The beams get “brighter” by x6
Data taking rate goes up by x6
Simulations go up by x3

Primary Data volume per year goes up by x30

This talk is about R&D strategies to keep the cost the ~same despite a x30 increase in data volume per year.

Will motivate the R&D via a detour on how science is done.

The Science

(oversimplified view)

This is what we measure

Brightness of the colliding beams

corrected Yield = Cross Section times Luminosity

This is the science

Corrected yield = observed beam crossings that pass selections divided by selection efficiency

Selections are developed, and their efficiency is determined from simulations.

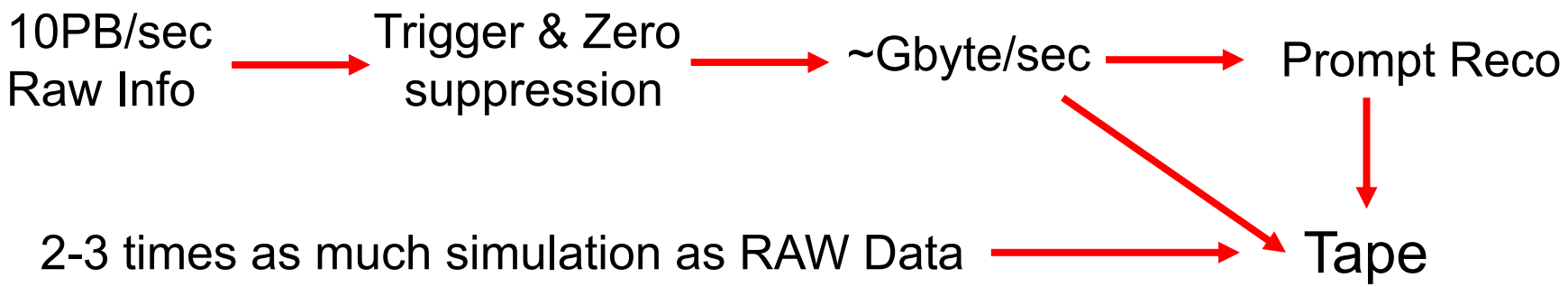
“Big bang” in the laboratory

- We gain insight by *colliding protons at the highest energies* possible to measure:
  - Production rates
  - Masses & lifetimes
  - Decay rates
- From this we *derive the “spectroscopy” as well as the “dynamics” of elementary particles*.
- Progress is made by going to higher energies and more proton proton collisions per beam crossing.
  - More collisions => increased sensitivity to rare events
  - More energy => probing higher masses, smaller distances & earlier times

Spectroscopy and Dynamics

- Spectroscopy:
  - What are the particles that exist ?
  - What are their properties ?
- Dynamics:
  - What are the forces ?
  - How do the particles couple to the forces ?
  - How do these depend on energy and angular momentum ?

Basic Data Flow Today



All subsequent processing and analysis starts with decisions of what data stays on disk, and for how long.

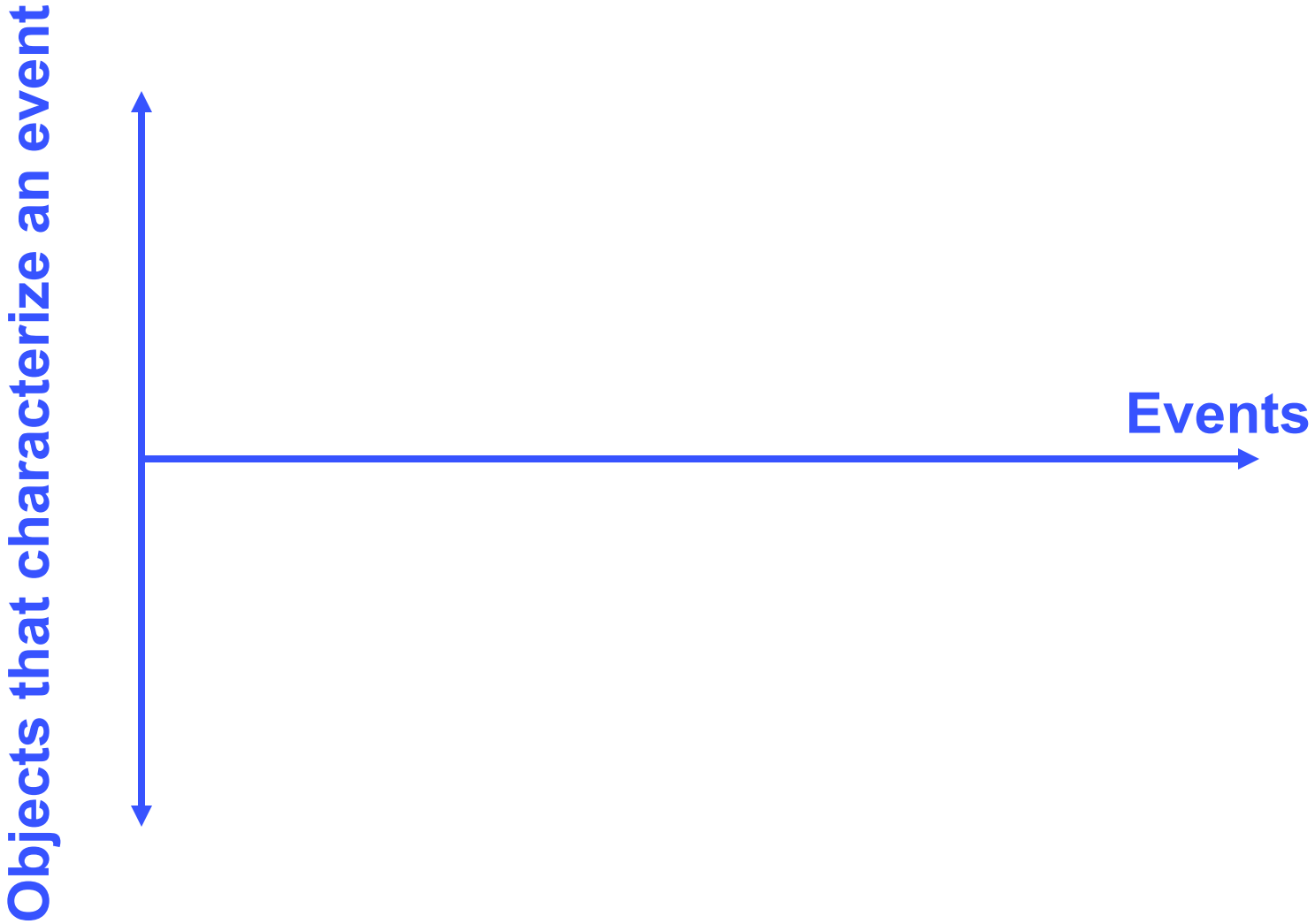
150 Petabytes Global disk based Data Federation

Any data anywhere on disk can be accessed from anywhere with an internet connection.
Something like a Content Delivery Network for Science.

~200,000 core processing capacity across 50-100 clusters

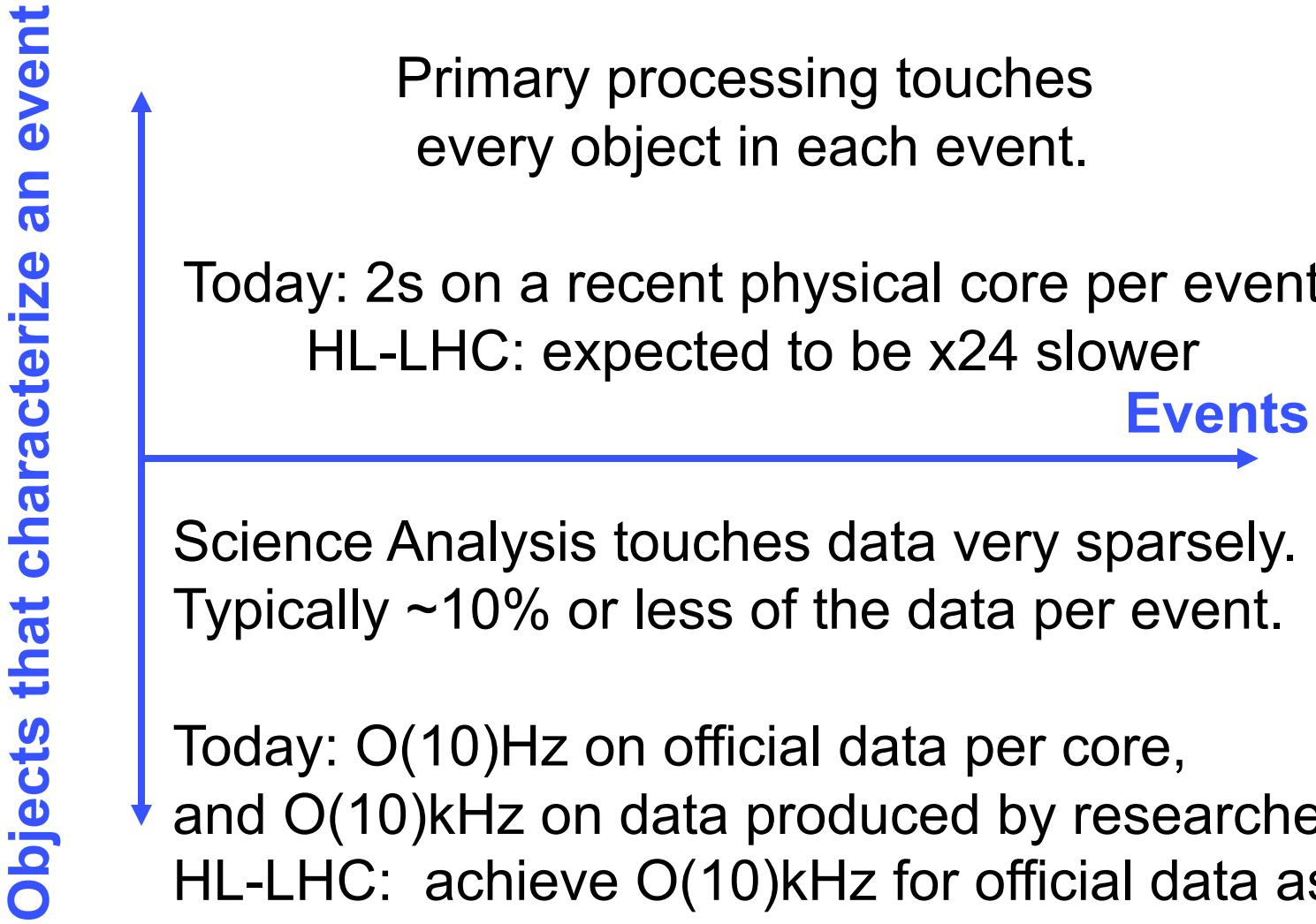
The Nature of Data at the LHC

Think of it as a 2D Problem



Let’s contrast some processing use cases in this context

Think of it as a 2D Problem



Primary processing touches every object in each event.

Today: 2s on a recent physical core per event.
HL-LHC: expected to be x24 slower

Science Analysis touches data very sparsely.
Typically ~10% or less of the data per event.

Today: O(10)Hz on official data per core,
and O(10)kHz on data produced by researchers.
HL-LHC: achieve O(10)kHz for official data as well.

The HL-LHC really has 2 data problems that have very little to do with each other !!!

- Primary processing of O(1) Exabytes at O(100) seconds per event per core.
  - All data in event is accessed
- Science Analysis of TB to PB at O(10)kHz event rate per core.
  - Small fraction of data per event is accessed.

The HL-LHC Primary Processing Problem

A more detailed discussion can be found here

RAW Data Processing at HL-LHC

- Each of ATLAS and CMS want to do their annual processing campaign of previous years data and simulations during the first ~100 days of the new year.
  - 1 Exabyte in 100 days => 10PB data/day => 1Tbit/sec
- All data resides on tape across the T1 centers worldwide.
  - Roughly 40% of it at FNAL and BNL combined, i.e. in USA
  - US portion of processing is ~ 400Gbit/sec for 100 days straight.
    - Even if you restrict processing to just the RAW, and consider only one experiment at a time, this is still ~100Gbit/sec non-stop for 100 days in the US alone !!!

Technical Challenges

- Tape recall
  - How much bandwidth can we achieve from tape?
    - annual processing is unlikely to be the only tape archive activity for those 100 days.
  - What's reasonable for buffer sizes in front of tape archive?
- Manage the limited disk buffer at archival T1
  - Tape recalls will be carousel style, i.e. buffer much smaller than the exabyte dataset.
- Manage 1Tbit/sec network to one or more HPC center, plus probably many smaller center.
  - Network bandwidth needs to be managed with tools like SENSE and AutoGOLE
  - We will want to transfer in bursts >> than steady state requirement.
- Manage the disk buffer at the HPC center
- Bring outputs back the same way, including storing in tape archive(s).
- Co-schedule processing and all of the above.
  - Probably feed computing steady state while data transfers are bursty.

Aside on Transatlantic Networking

- In early LHC days, each T1 center was responsible for processing of data it archives.
  - Assumed that networks are not good enough to move data around the world as needed.
- Today, ATLAS and CMS transfer data globally to where there is processing capacity.
  - Not just at T1s but also T2s, and not just data in the region where it is archived.

The early HL-LHC may be more like the early LHC

The HL-LHC Analysis Data Challenge

Analysis Challenge

Each of ATLAS and CMS has more than 1000 scientists from a few hundred institutions in more than 50 countries that want to exercise their academic freedom to analyze this data to their hearts contents.

Innovation & science success depends on academic freedom

Collectively produce Primary datasets

Compete against each other for best ideas/results

Converge on publishable results that all can agree with.

Competition drives innovation

Event Sizes

	RAW [MB]	AOD [MB]	MINI [MB]	NANO [MB]
Today	0.9	0.35	0.035	0.001
HL-LHC	6.5	2.0	0.250	0.002

CMS produces different size data formats for different purposes.

RAW -> AOD -> MINI -> NANO

Each can be produced form the previous.

Most flexible
slowest
~50s per event per core

Easiest to use
fastest
10kHz events/core

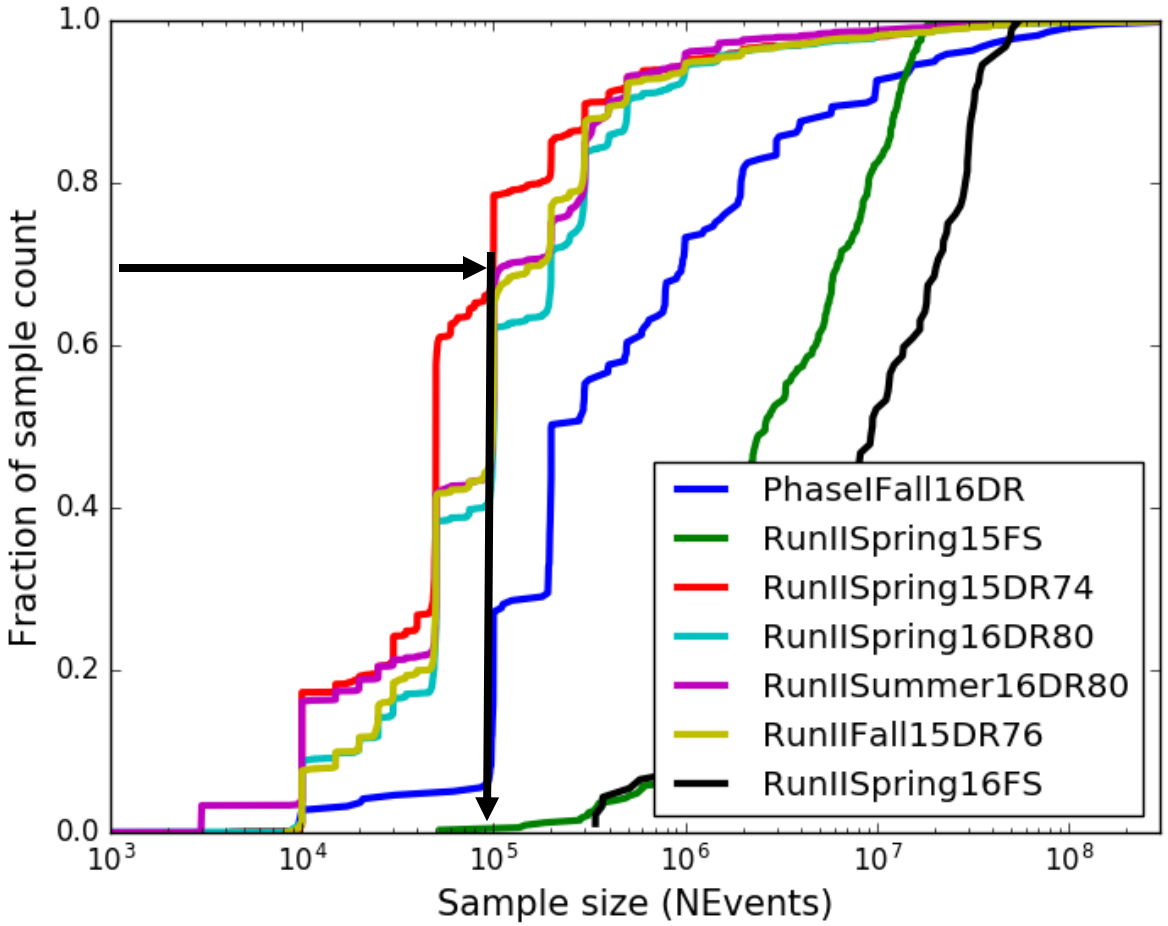
(speed expectation for HL-LHC data)

Aside on Simulations

- The science program is exceedingly diverse.
  - All imaginable physics that could be produced, and searched for needs to be simulated.
    - Thousands of small samples of limited physics interest.
  - A few dominant physics processes are copious background processes for many searches
    - A much smaller number of very large samples.
- Dynamic range of distinct physics samples ranges in size over orders of magnitudes.

Annual repetitions for simulations to correspond to annual data releases.

Size vs fraction of sample count



Example Campaign:

RunIISummer16DR80

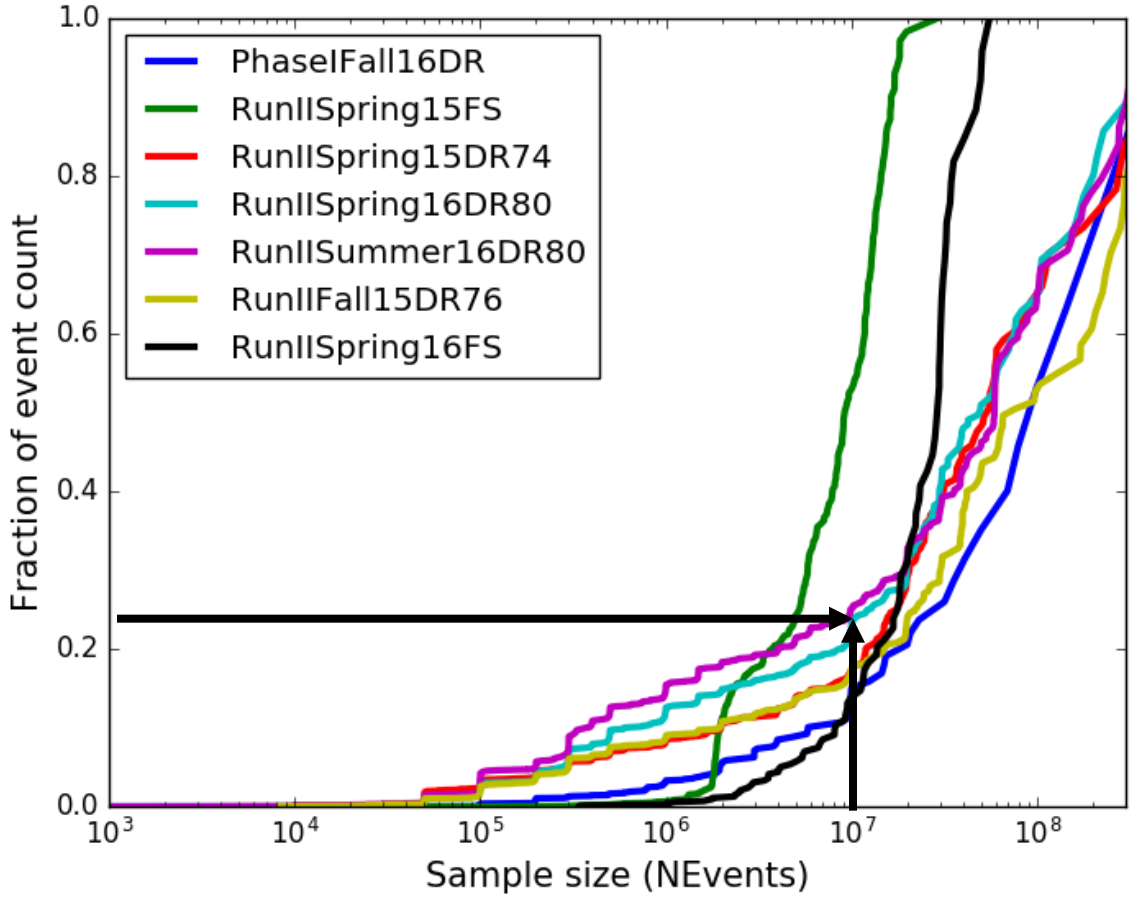
~70% of samples with < 100,000 events per sample

This campaign has 12 Billion events total

It was the main physics campaign for 2016 data taking

14,000 samples in this simulation campaign

Size vs fraction of total evts



Example Campaign:

RunIISummer16DR80

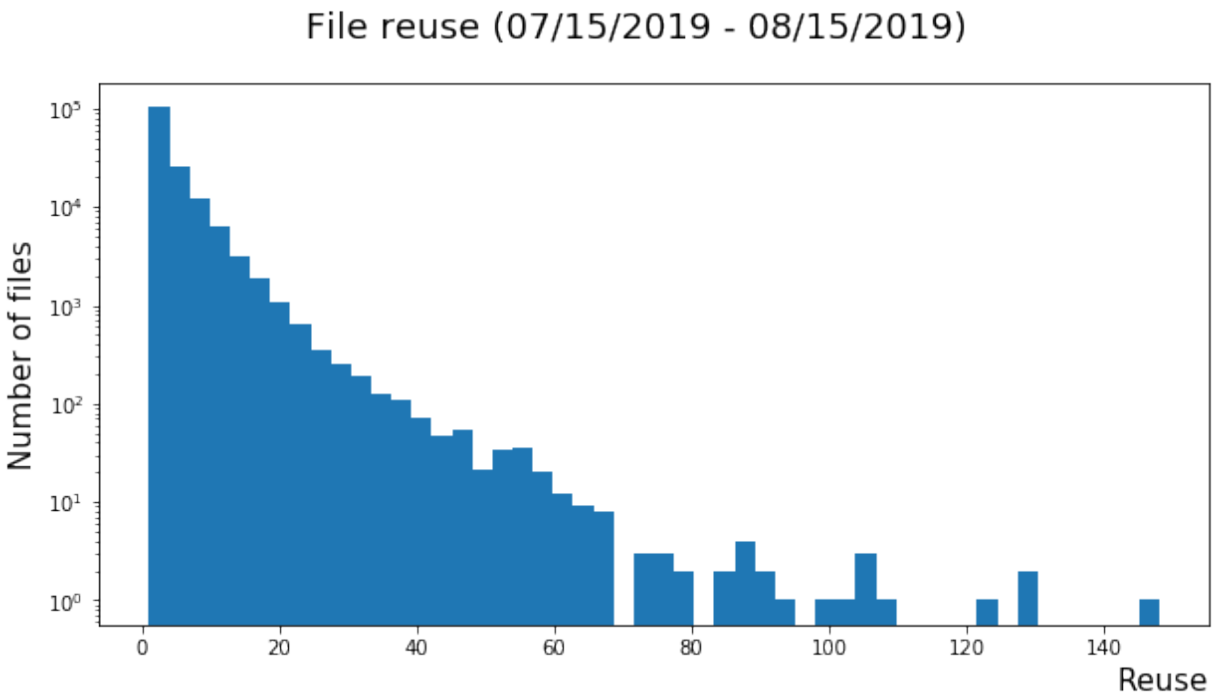
~80% of the total # of events are in samples > 10 M events/sample

From previous page:

~70% of samples with < 100,000 events per sample

We should expect that most samples are rarely accessed.

Measured File Reuse



Measured # of times each file at UCSD was accessed during 7/15-8/15 2019.

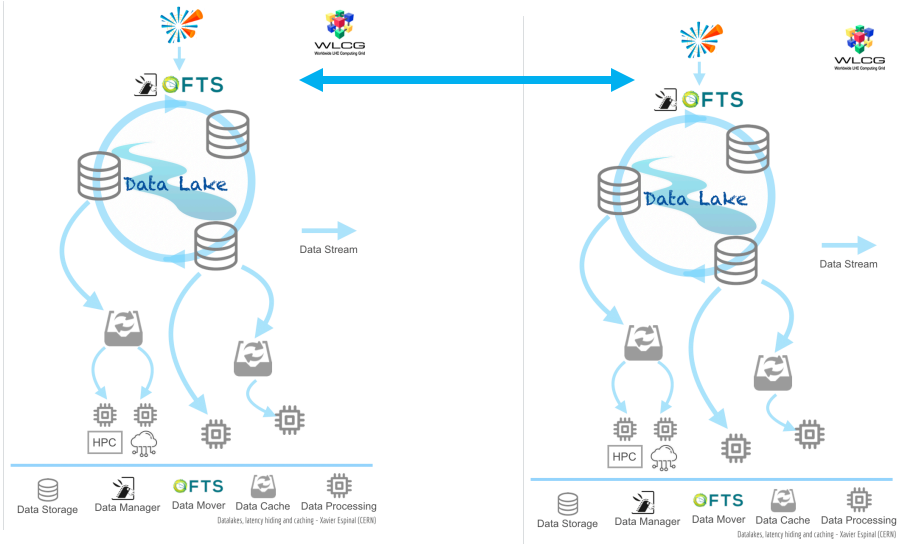
Hyper-exponential Distribution

The HL-LHC Data Lakes Model

LHC Data Lakes Model

- More than one lake globally
  - E.g. USA as one lake per experiment seems plausible.
  - “Federation of lakes”
- Centrally managed replication between lakes.
- Intra lake data access via mix of:
  - Top-down placement, e.g. as part of workflows
  - Bottoms-up placement for cache misses
  - Streaming for remote file open

Cross lake transfers



Start exploring features via mix of R&D pilots and production pilots.

Aside on Networking

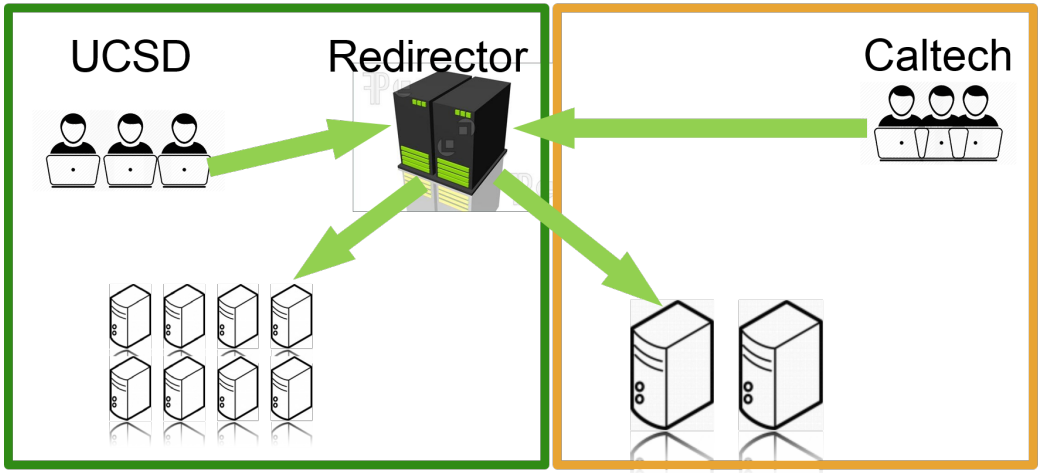
- Data Lakes Model implies that CMS manages its use of the transatlantic link. (no streaming)
  - E.g. 90% of the link use could be under SENSE control for HL-LHC, while the remaining 10% is wild west and best effort without performance guarantees.
- Within the US, CMS will likely want to tag flows according to broad use categories.
  - Be able to assign priorities based on category.

Production Scale Caching Pilot

Caltech & UCSD operate a joint PB disk cache.

Southern California (SoCal)

(Roughly 20,000 cores across Caltech & UCSD … half typically used for analysis)



CPU in both places can access storage in both places.

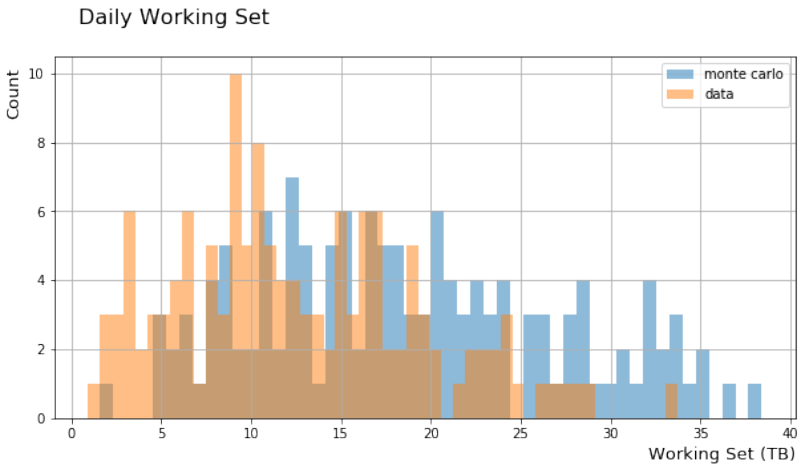
How much disk space is enough?

Cache MINI and measure working set accessed.

Working Set (WS) Definitions

WS = sum of sizes of all files accessed in a time period.

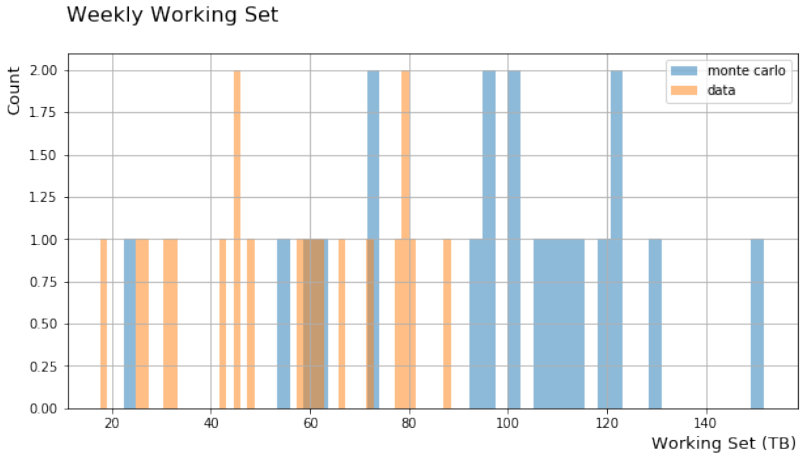
For SoCal cache prototype we measured:

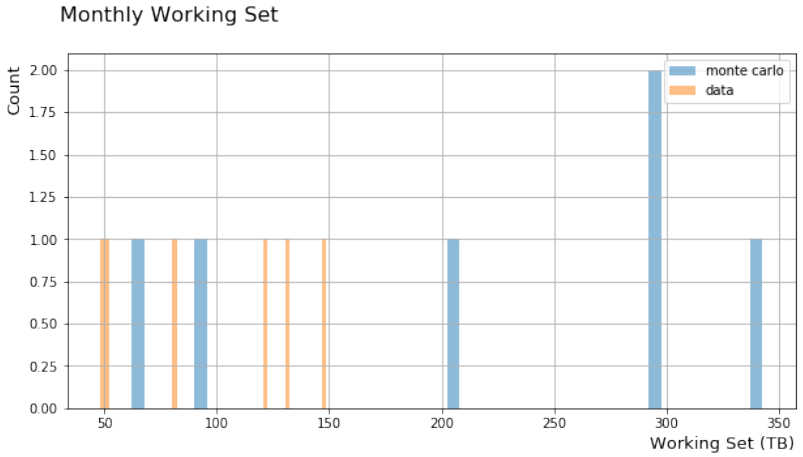


Few tens TB daily
Few hundreds TB monthly

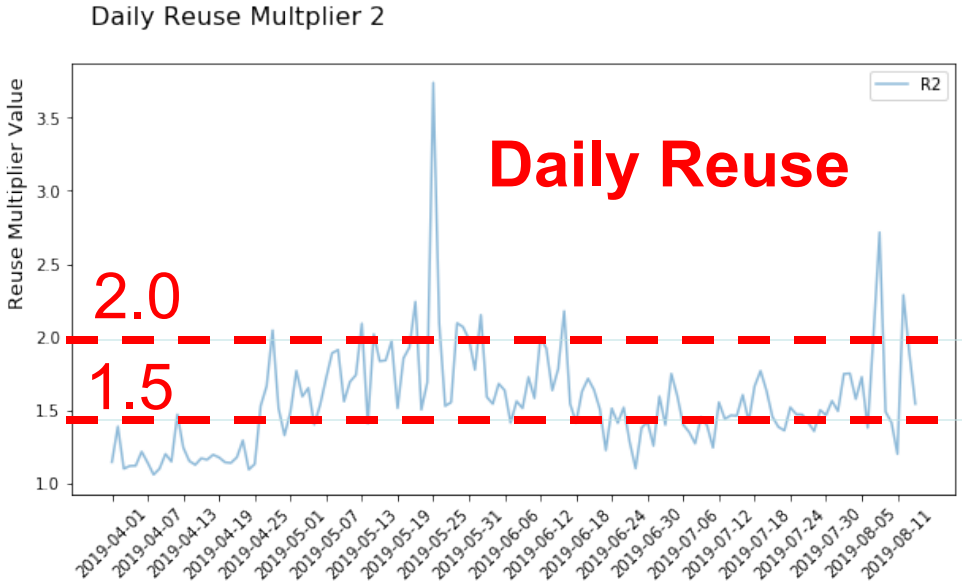
An obvious x10 trade-off between disk space and network use.

SoCal WS for 10/19 = 451 TB





File reuse Measurement



Numerator sums over
all files accesses per day

$$\frac{\sum_i size(f_i)}{\text{Working set}}$$

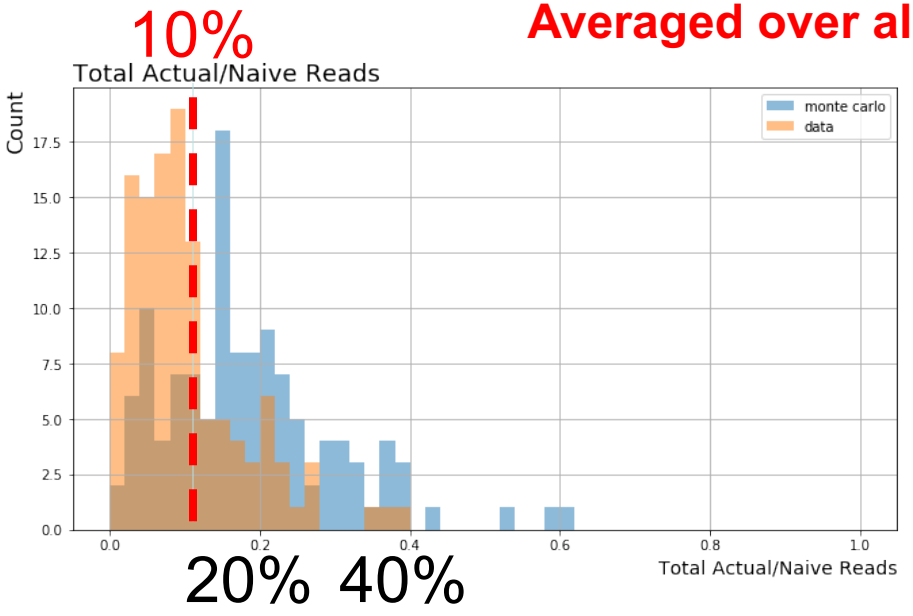
small files matter less than large files

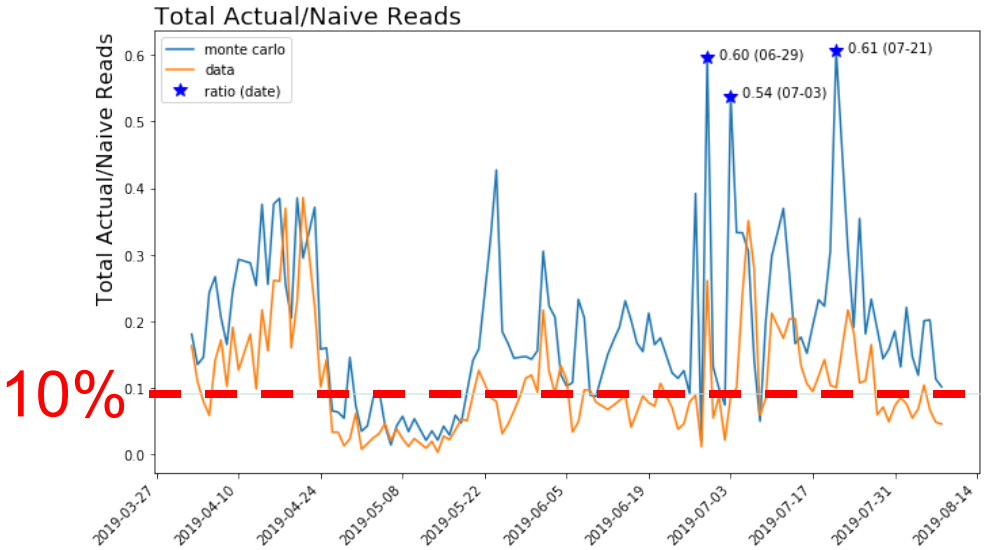
If all files accessed are accessed once only then Reuse = 1.
If each file in working set is accessed 2 times then Reuse = 2.

Modest average reuse of files in cache even daily!

Partial File Reads

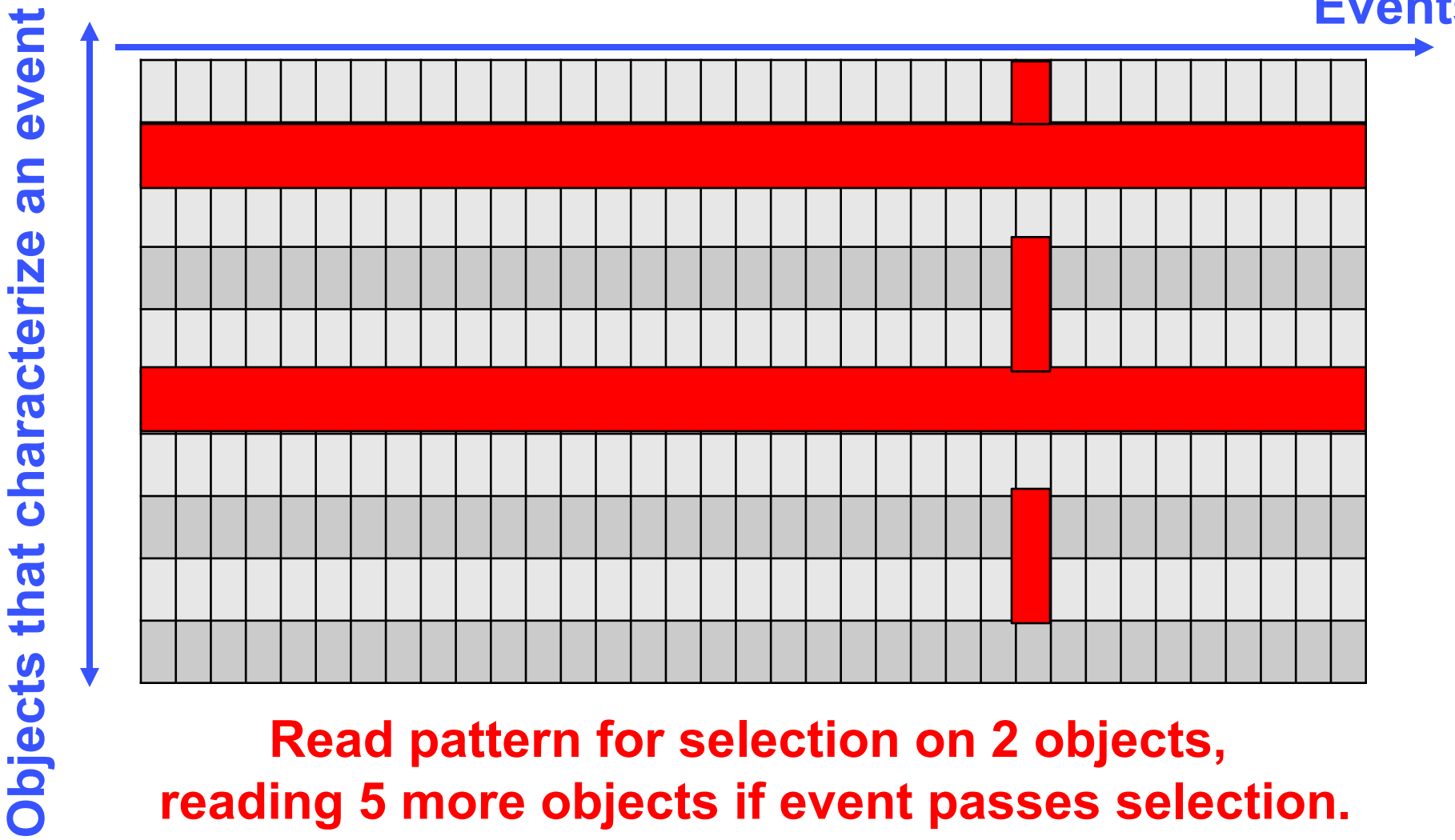
Averaged over all file reads per day





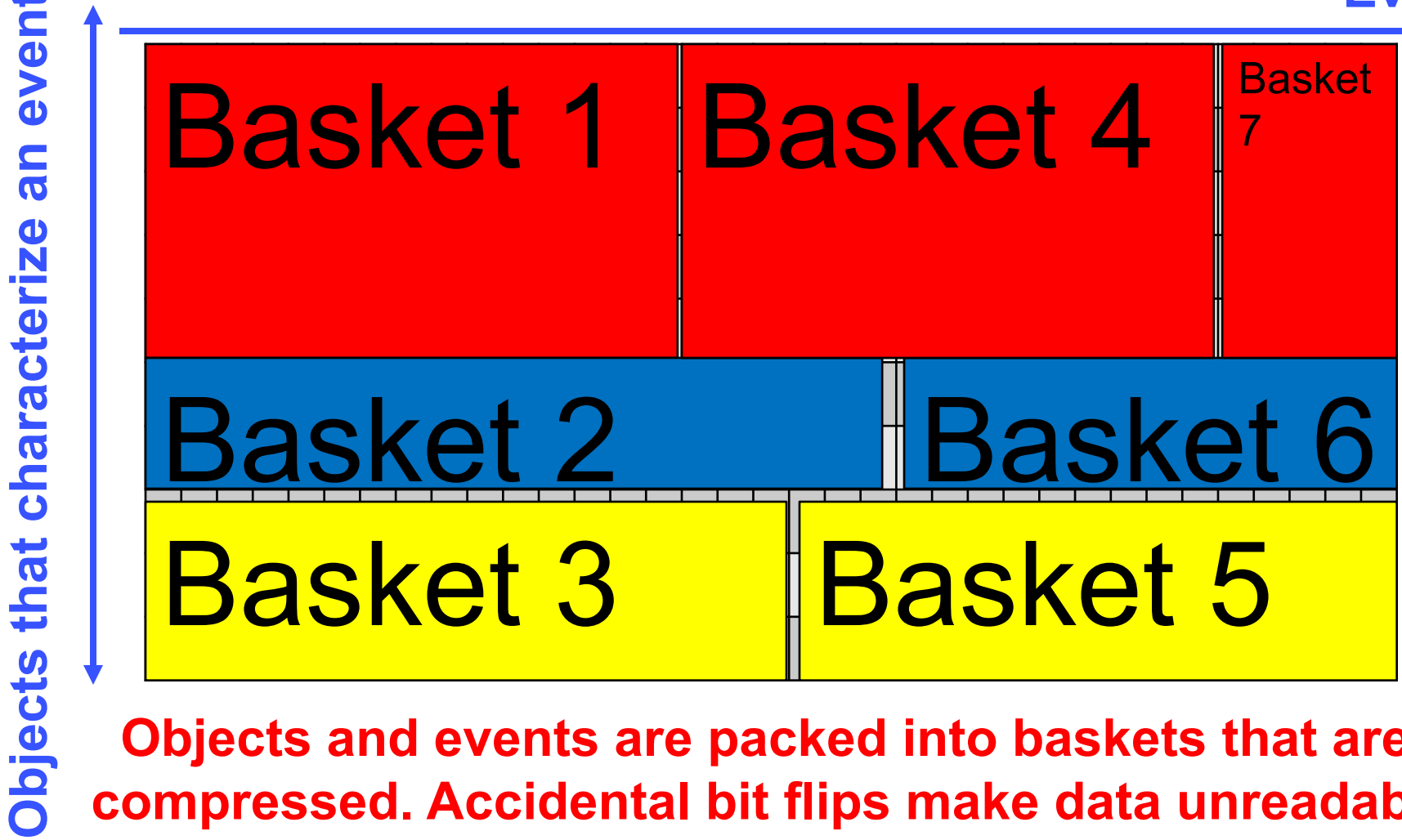
The data formats of CMS for Analysis are designed to support partial file reads.

Recall 2D structure of data



Read pattern for selection on 2 objects,
reading 5 more objects if event passes selection.

Aside on Data Integrity



Objects and events are packed into baskets that are compressed. Accidental bit flips make data unreadable

Length and width of baskets are part of data format definition. They are designed with read patterns in mind.

3 R&D Topics as next steps

- Better Cache Miss Algorithms than LRU
- How large can a region be?
- Can we exploit partial file reads to save disk space?

Better Cache Miss Algorithms

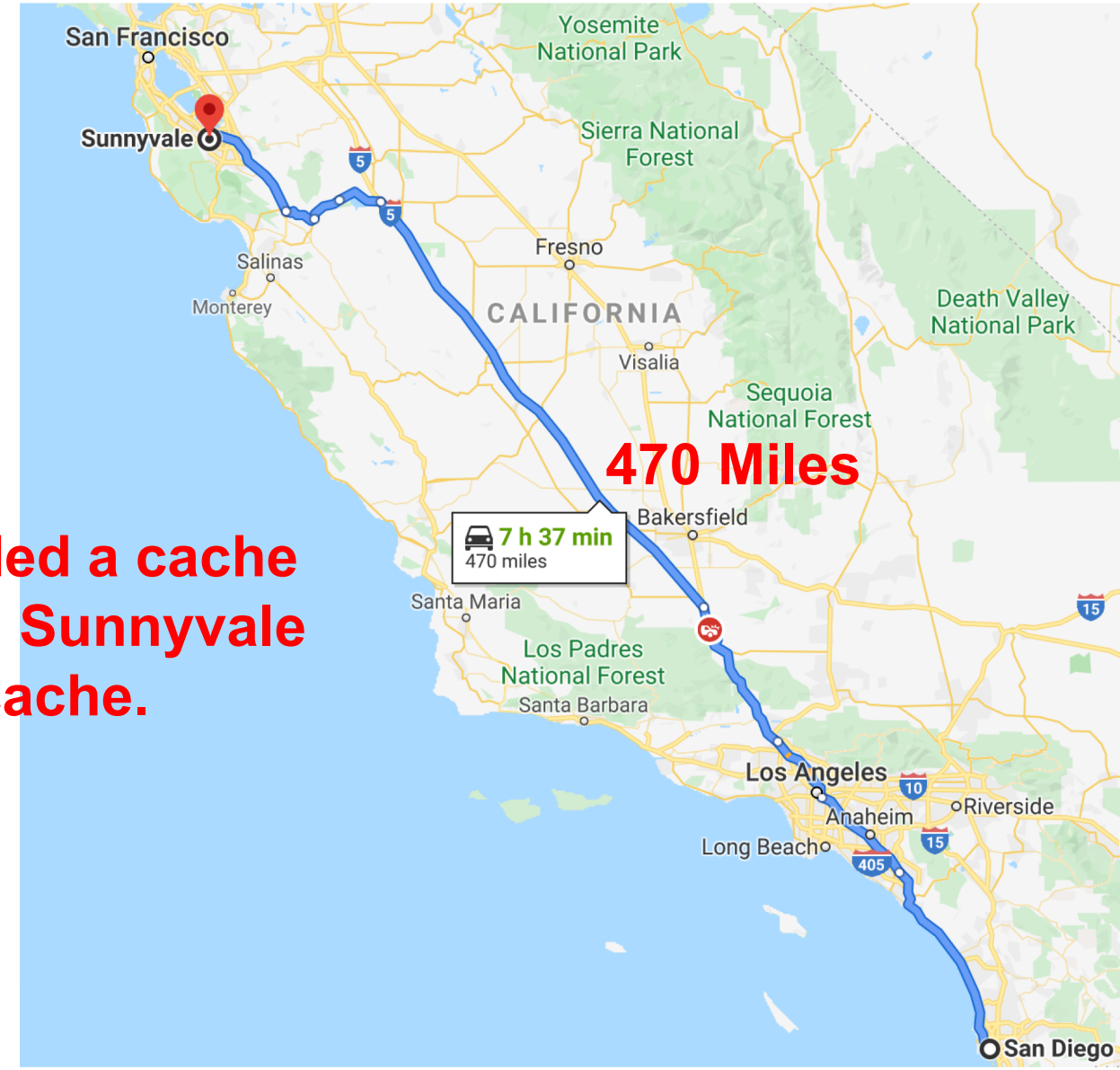
- Information in sample metadata:
  - physics content of sample
  - Processing campaign of sample
- This information ought to be useful to predict future use much better than LRU.
  - Detect when new campaign becomes more popular than old campaign for same physics content.
  - Learn what physics is done where, and cache accordingly.
- A group at INFN is developing an AI algorithm for cache misses along these lines.

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. => Regional Caches span countries in EU

Added ESNet Cache to SoCal

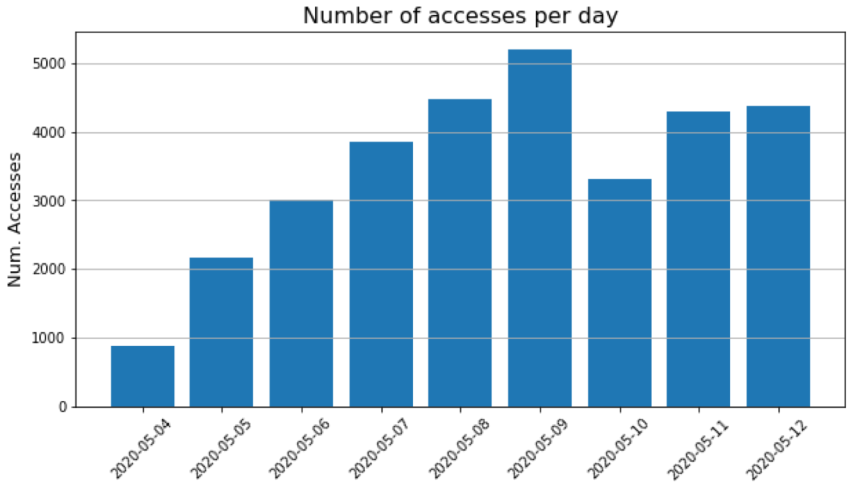
In early May, we added a cache at the ESNet POP in Sunnyvale to the SoCal cache.



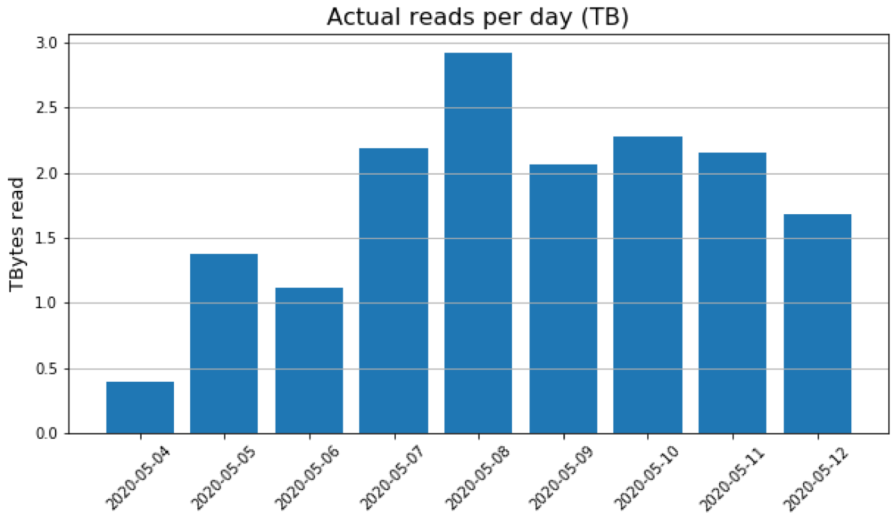
5/15/2020

Initial Use of ESNet Cache

~ 5000 accesses per day



~2TB data read per day



The objective is to gain experience of operating a regionally (~500 miles) distributed cache.

E.g.: We will be comparing CPU time / wall time for the full range of jobs we see in production system.

Exploiting partial file reads

- Researchers analyze multiple datasets for an analysis with the same executable.
  - Typical physics publication requires a couple dozen datasets.
- Can we predict access patterns, and exploit them for partial file caching?
  - learning in realtime ?
  - user defined data filters when read into caches at analysis facilities ?

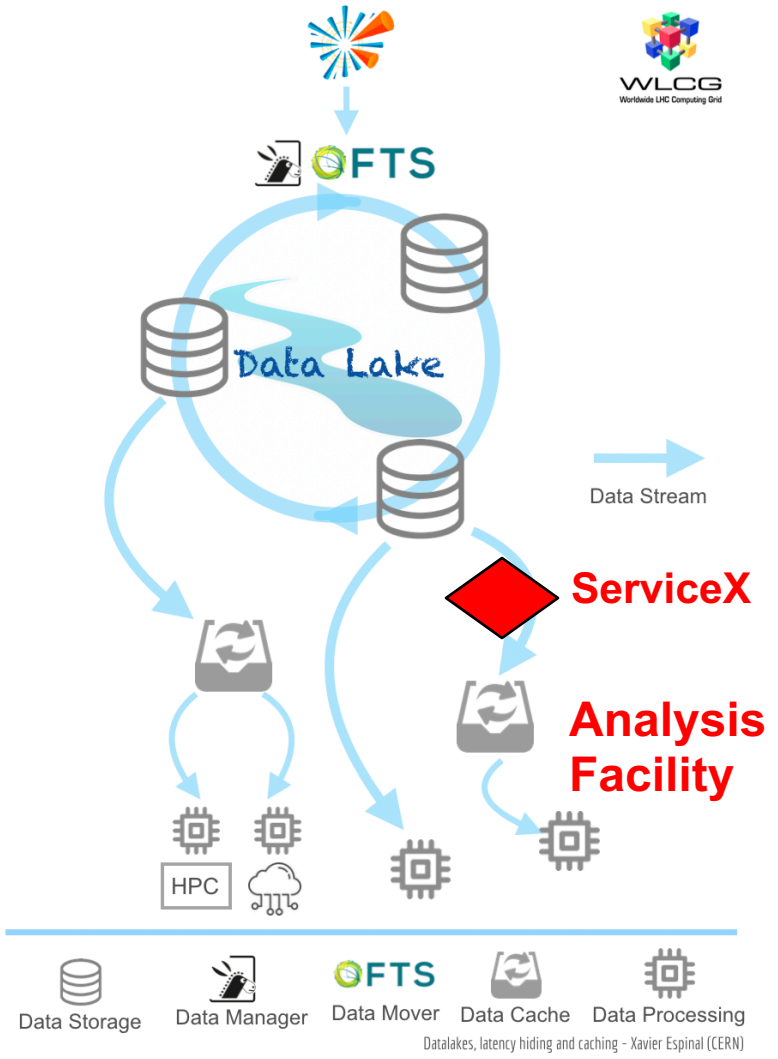
Unclear today how best to exploit partial file reads

Aside on Analysis Facilities

- The LHC community is considering a paradigm shift from analyzing its data sequentially along the events axis, to “declarative programming”.
  - Define the selections and let the loop over events be implemented by a “compiler” & infrastructure.
  - Selections define relevant object axis
- This opens up more predictive and speculative analytics on what in the 2D plane of objects vs events to store and cache.
  - Initial pilot projects show orders of magnitude speedups.

R&D for ServiceX

- ServiceX allows transformations to be applied on the data as it enters the cache of the analysis facility.
  - Filtering on events?
  - Reformatting to change bucket structure?
  - User specified code?



Unclear today what's the "killer app"

Summary & Conclusions

- HL-LHC expects Exabyte/year data by 2028.
  - x30 increase over today !!!
- This has lead over the last couple years to careful re-examination of what we do with our data, and why.
- Identified promising R&D to save on disk space needs via a mix of network, caching, and tape archiving.
  - Combination of top-down, bottoms-up, and direct access to data.

Questions ?