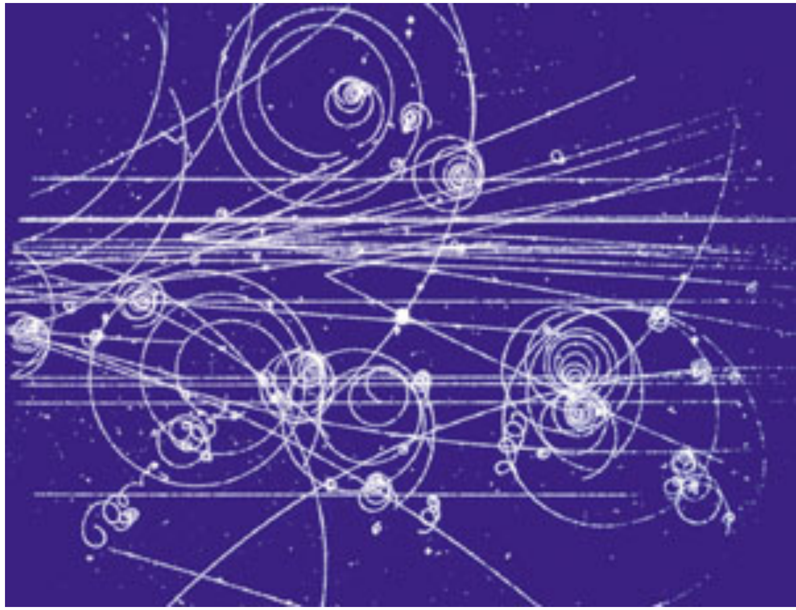
Science and Technology Facilities Council

The High Energy Physics Software Stack

Katy Ellis, PPD RAL
Efficient Computing for HEP, 17/02/2020

Contents

- Personal introduction
- HEP terms and experiments
- Upcoming challenges
- HEP software stack
- Distributed computing
- Efficiencies



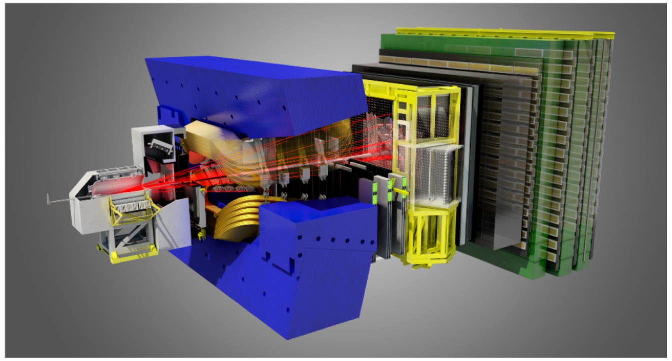
Science and Technology Facilities Council

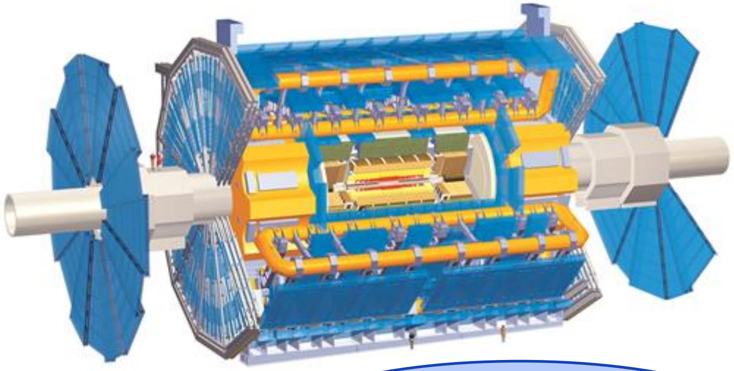
Personal introduction

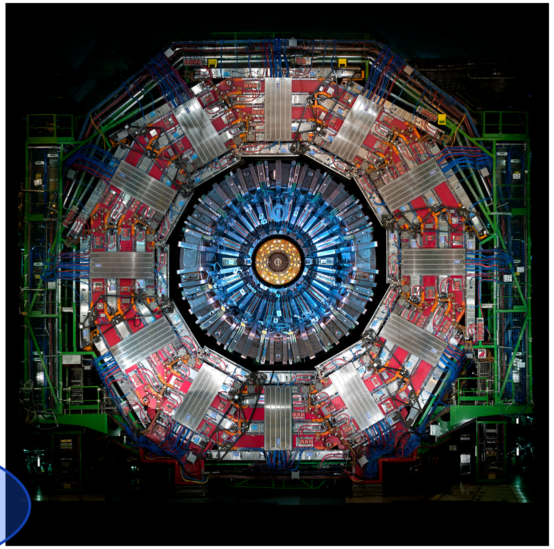
- CMS / Tier 1 Liaison at the Rutherford Appleton Laboratory (RAL)
  - Ensures that computing for the CMS experiment is performing well, particularly at RAL
  - Sits on the interface between CMS physicists and Tier 1 service administrators
- PhD in Experimental Particle Physics
- Various (computer) modelling jobs in industry and at UK labs
  - Stealth Scientist at Qinetiq
  - Tester of oil and gas reservoir simulation software at Schlumberger
  - Nuclear fusion power station modeler at Culham

What is HEP?

- High Energy Physics, aka Particle Physics, is the study of subatomic matter particles and force-carrying particles.
- Several different types of experiment, many on a grand scale:
  - Accelerators (such as the Large Hadron Collider)
  - Neutrino





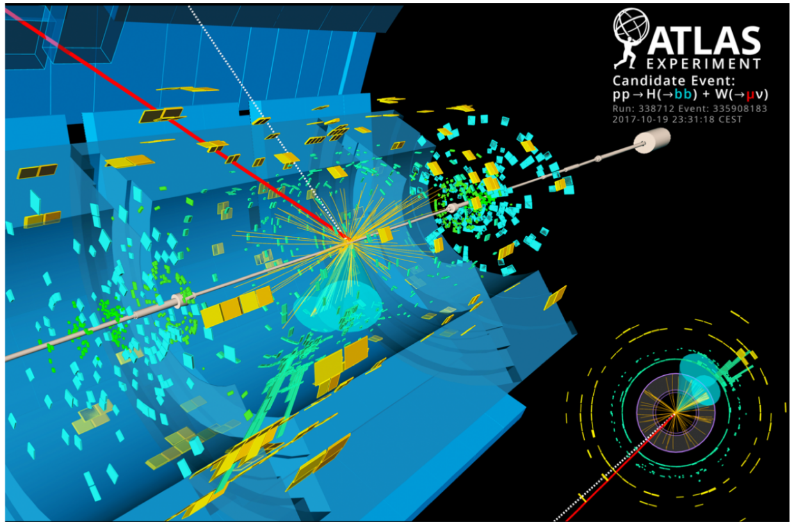


~100 million channels

Science and Technology Facilities Council

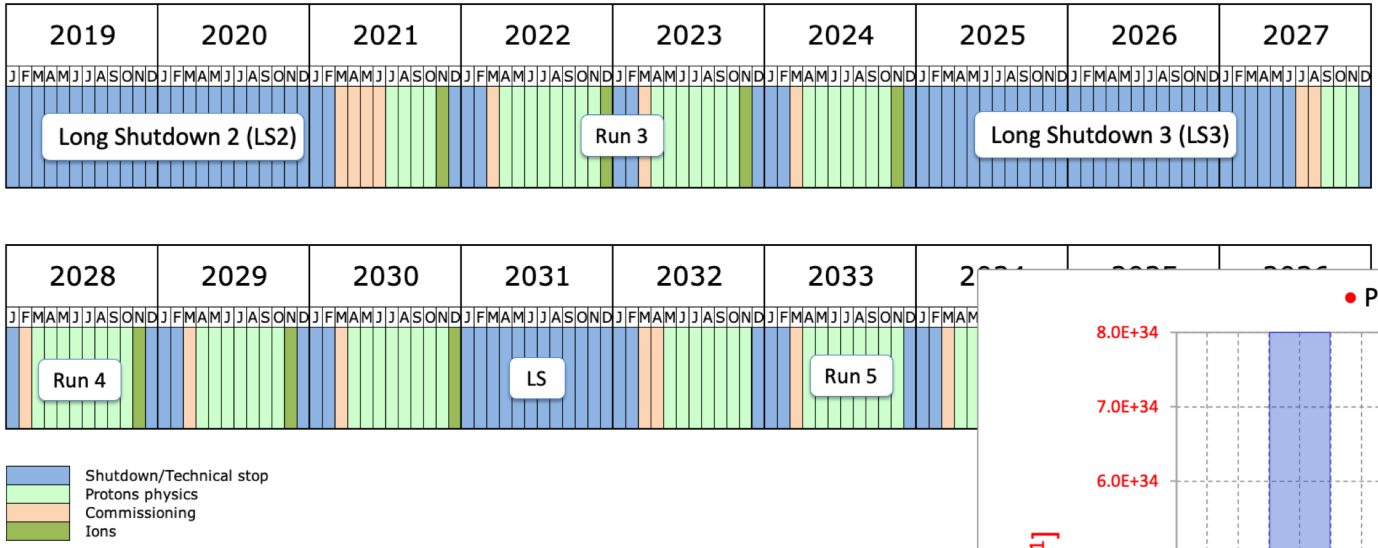
Events

- ‘Event’ is used in HEP to represent a snapshot of the detector at a particular moment
  - Ideally this captures something interesting!
- The same word ‘event’ also describes a theoretical interaction created by Monte Carlo-based computer modelling
- As events are processed by different parts of the software stack they are stored in different formats.
  - RAW 1-2 MB/event, all the hits from the detector
  - AOD 100 kB/event, physics objects for analysis
  - Derived smaller formats

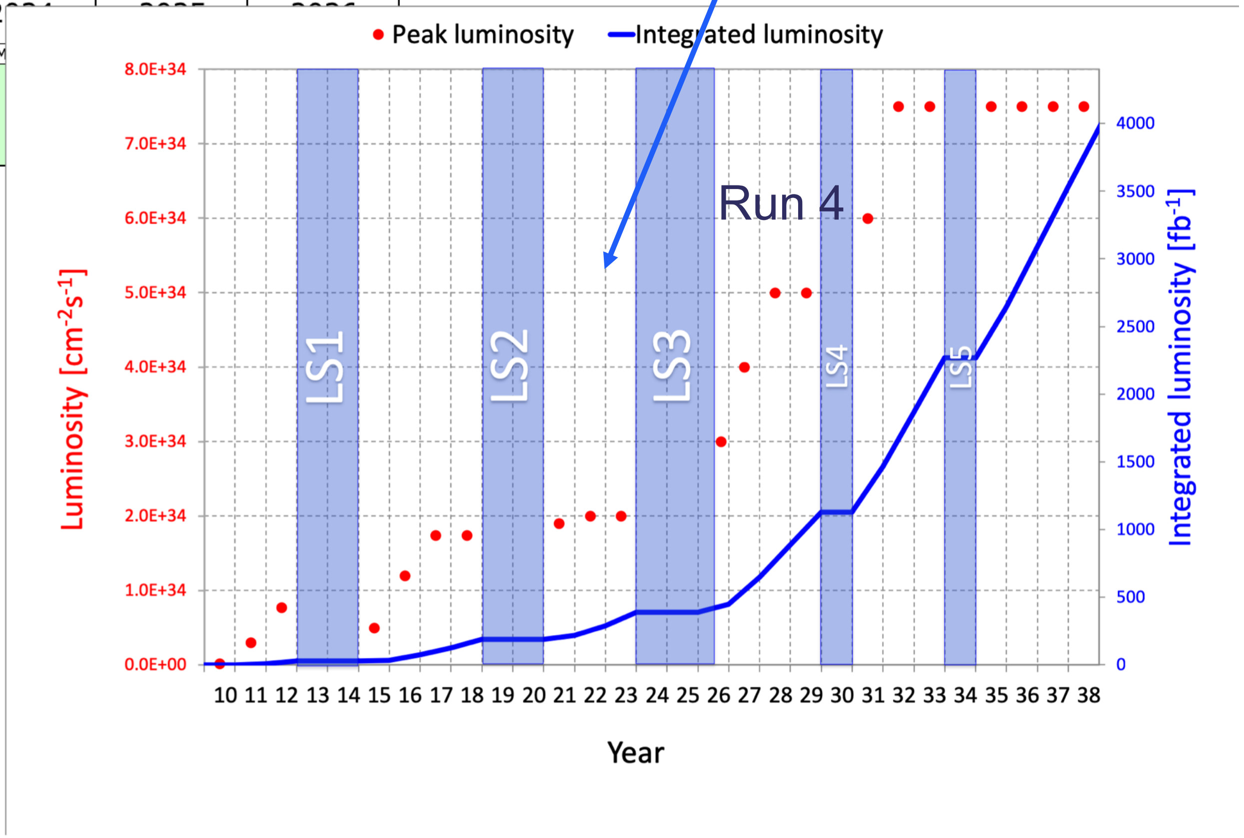


Science and Technology Facilities Council

Increase in LHC data rate



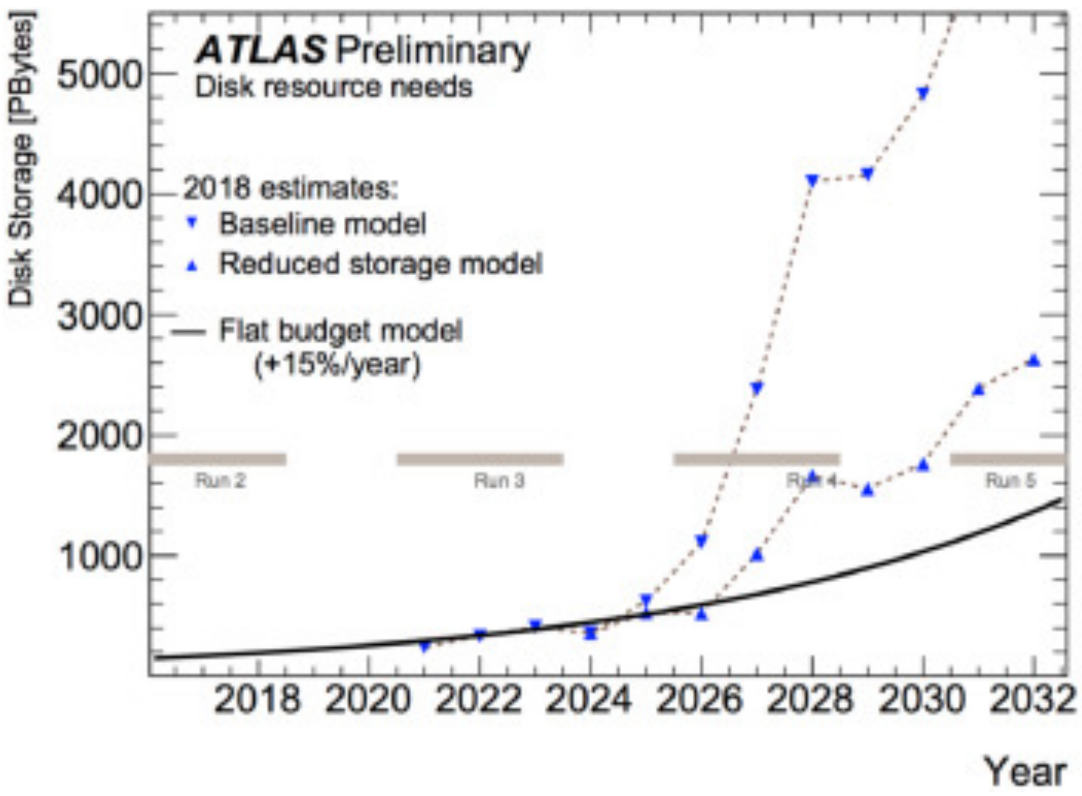
Plot has not been updated with delayed Run 3 schedule



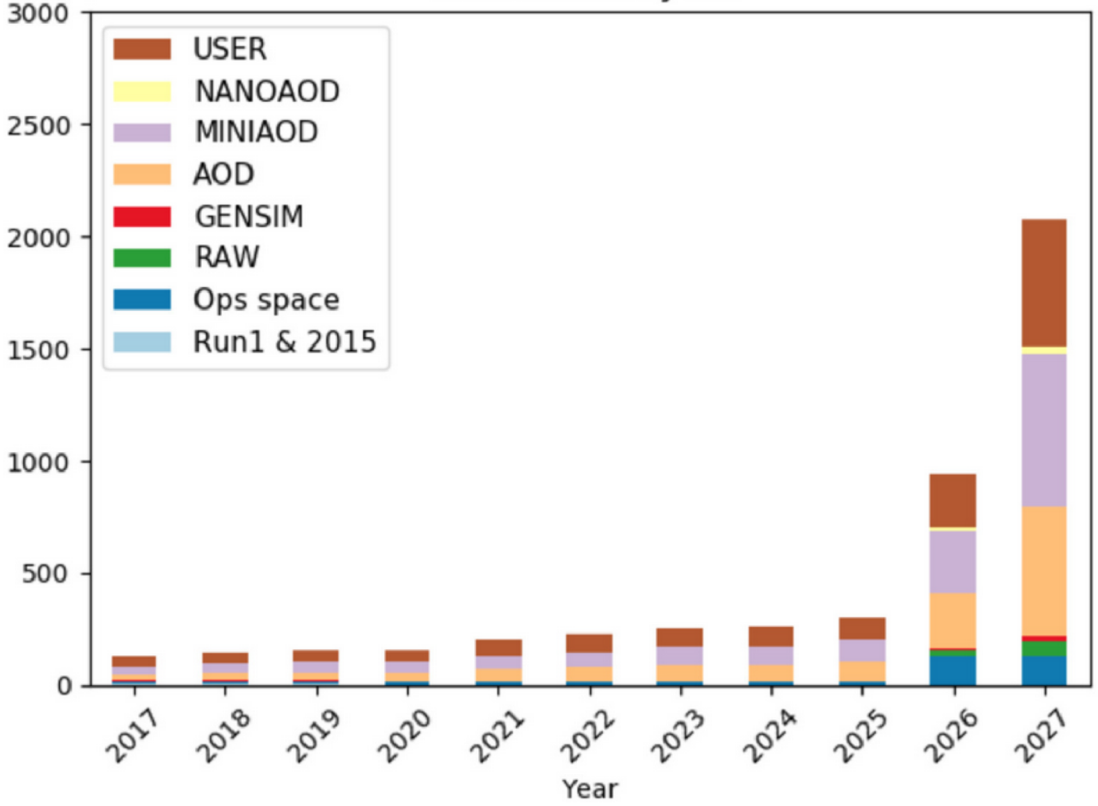
- Increased pile-up (number of collisions per bunch crossing)
- Increased detector resolution
- Increased recorded event rate

Science and Technology Facilities Council

Disk resource estimation

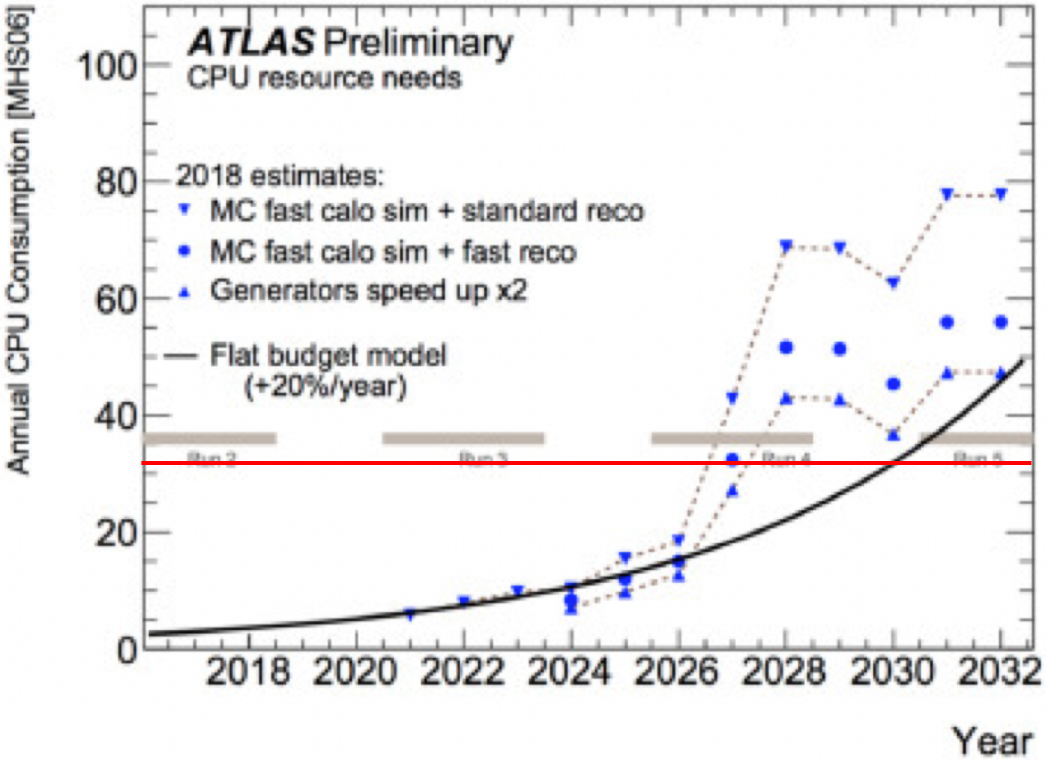


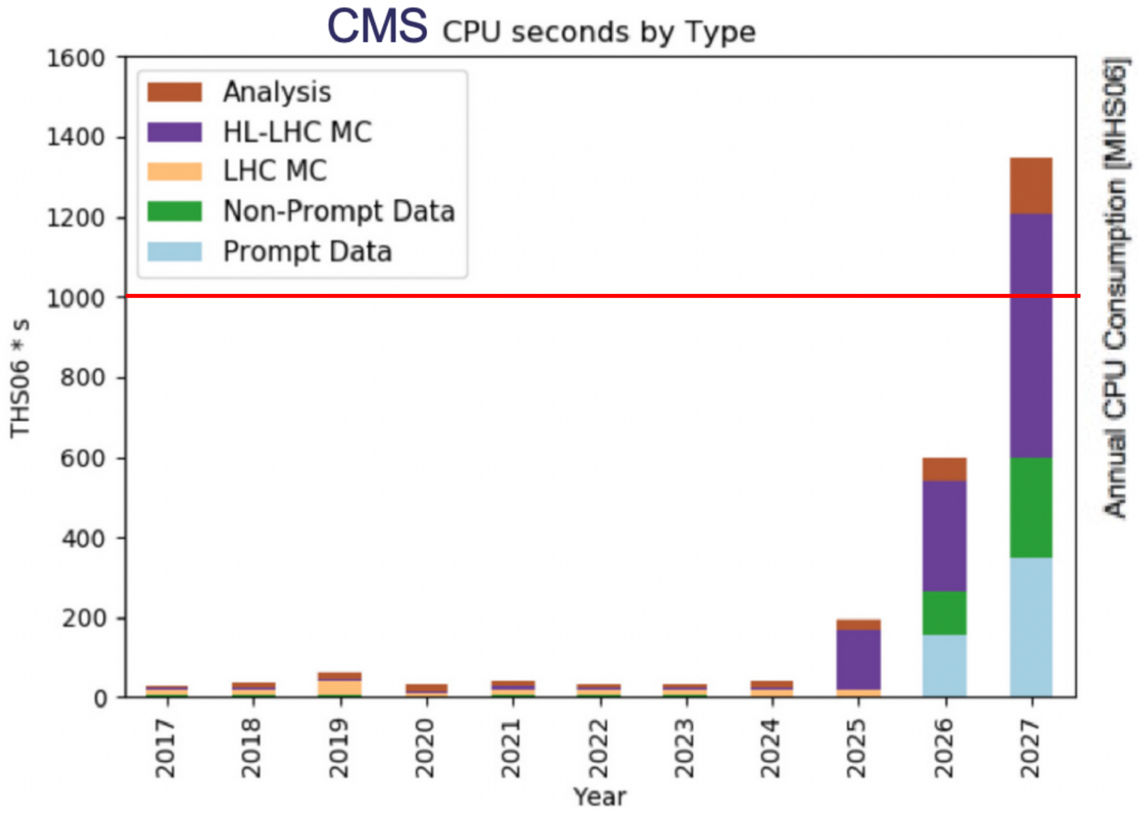
CMS Data on disk by tier



Science and
Technology
Facilities Council

CPU resource estimation





Red Lines are equivalent capacities

You can't expect ATLAS and CMS to use the same units…

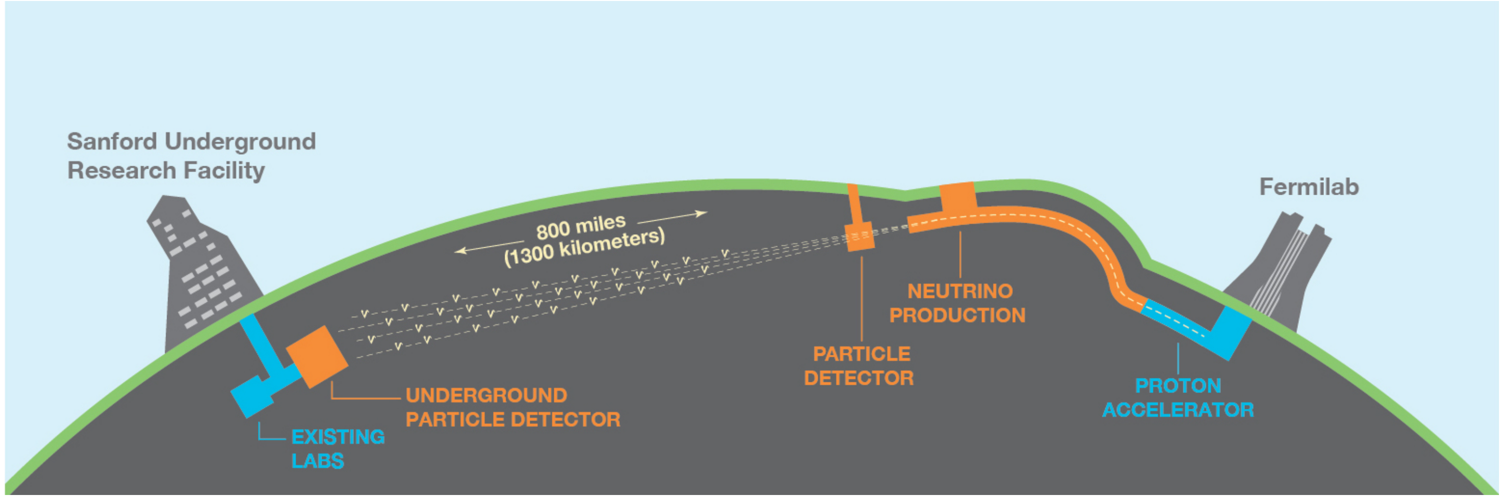
Science and Technology Facilities Council

Tape

- 2 copies of the LHC experiments RAW data are archived on tape; one copy at CERN while the other is spread across the Tier 1 sites.
  - Other data (normally infrequently used) is also archived on tape.
- Cheap (~1/4 cost of disk) and reliable long term storage.
  - Experiments using tape to prevent them from filling up disk
  - Disadvantages if you want to read the data as it is slow to retrieve
- Tape will be critical for a cost effective computing model.
  - Projects like the ATLAS Tape Carousel aim to optimize recalling data from tape.

Science and Technology Facilities Council

DUNE – Deep Underground Neutrino Experiment



- Pre-trigger data rate is ~ 6 TB/s.
- Limit on the output data rate is estimated to be about 30 PB/year or 8 Gbit/s steady state rate

July 2017
Groundbreaking for LBNF/DUNE

Autumn 2018
ProtoDUNE detectors online at CERN

2019
Begin main cavern excavation in South Dakota

2022
Begin installing the first DUNE detector

2026
Fermilab's high-energy neutrino beam to South Dakota operational with two DUNE detectors online

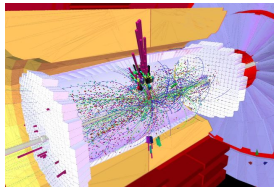
Science and Technology Facilities Council

HEP workflows

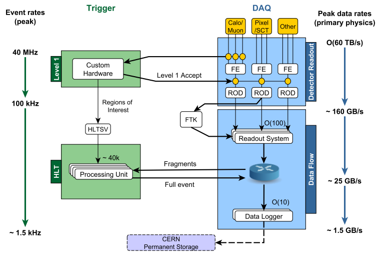
"Physics analysis"

Data reduction

Reconstruction



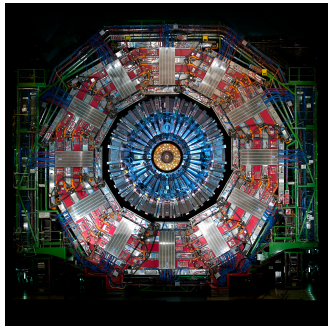
Simulation&DIGI



Trigger system

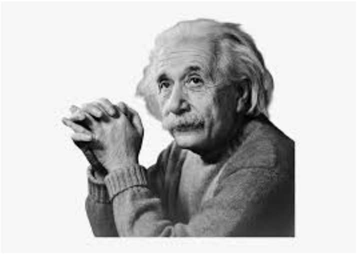
MC generators

Experiment data



Theory

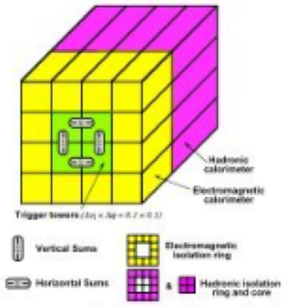
Experiment



Science and Technology Facilities Council

Event rate and trigger

- The number of proton-proton **bunch-crossings** taking place in CMS and ATLAS is ~ 40 million per second
- Trigger system determines which data events to keep
  - Decision must be made autonomously within an average time of a third of a second
- Experiments are currently limited by available computing to recording ~1000 events/s.
  - First level on the detector
  - High Level Trigger run on ~ 60k cores (CMS)
  - Size of an event: 1-2 MB

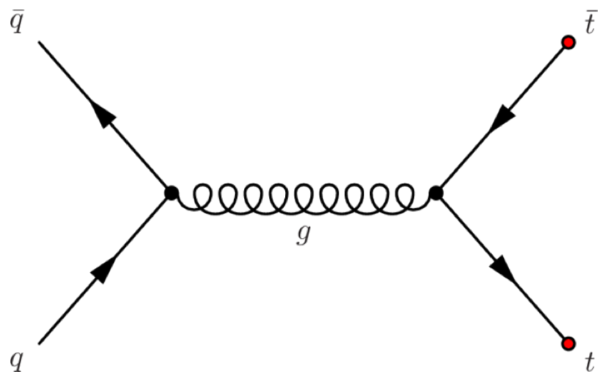


Science and Technology Facilities Council

Monte Carlo generation

- Computer program to model:
  - The “hard scatter”
  - “Final state” particles
  - Including “remnant” and background effects
- Outputs ‘4-vectors’ (p_x, p_y, p_z, E)
- More complex events require significantly more computing resource.

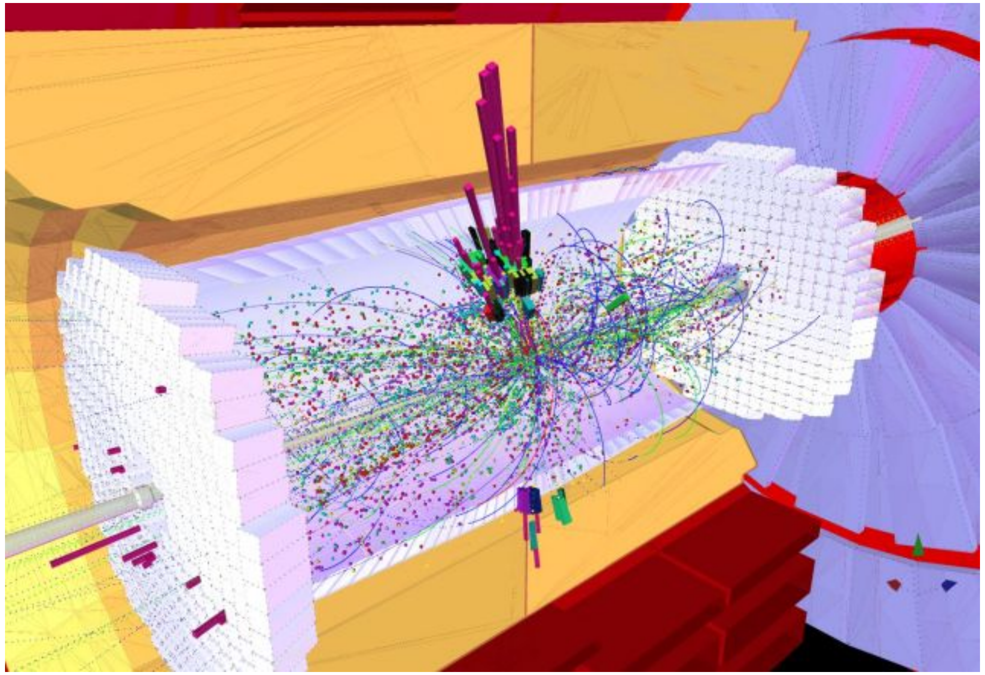
- Size of event: ~1 MB
- Require typically 1-10+ times as many events as measured experimentally.
- 10+ billion / year



Science and Technology Facilities Council

Simulation & Digitization

- The MC generator-level data is input to detector-modelling software to simulate e.g. magnetic fields, particle/detector interactions.
- Output is digitized so it resembles detector output, e.g. individual calorimeter cell hits, known detector inefficiencies.
- Including MC gen. step, uses ~40% of computing resources
- Size of event: 1.5 MB

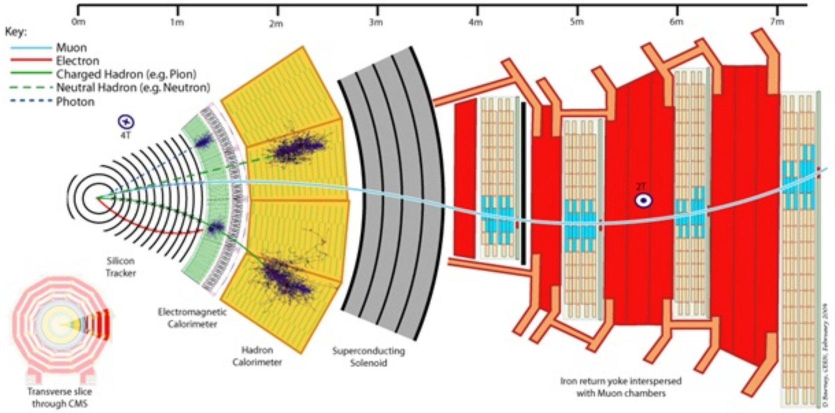


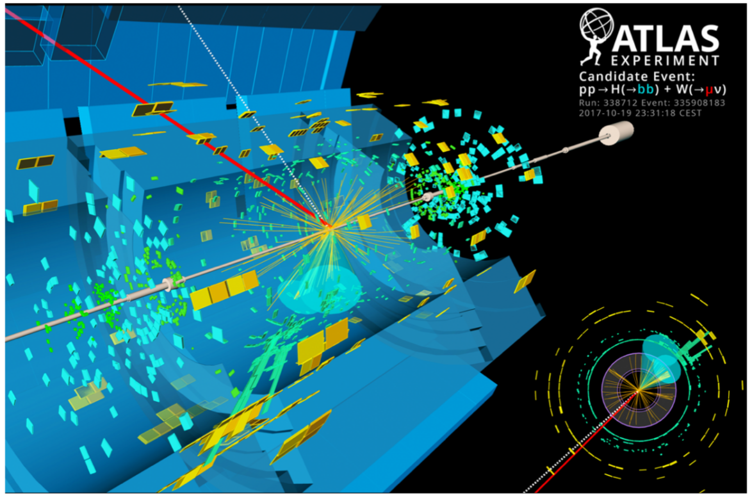
Science and Technology Facilities Council

Reconstruction

- Uses ‘hits’ from the detector or the digitized simulation.
- Using information about where in the detector the hits occurred, energy deposited, track bending, etc.
- Reconstruction algorithms define the ‘physics objects’
  - Electrons, muons, taus, jets, photons, etc.

- Size of event: ~1.5 MB



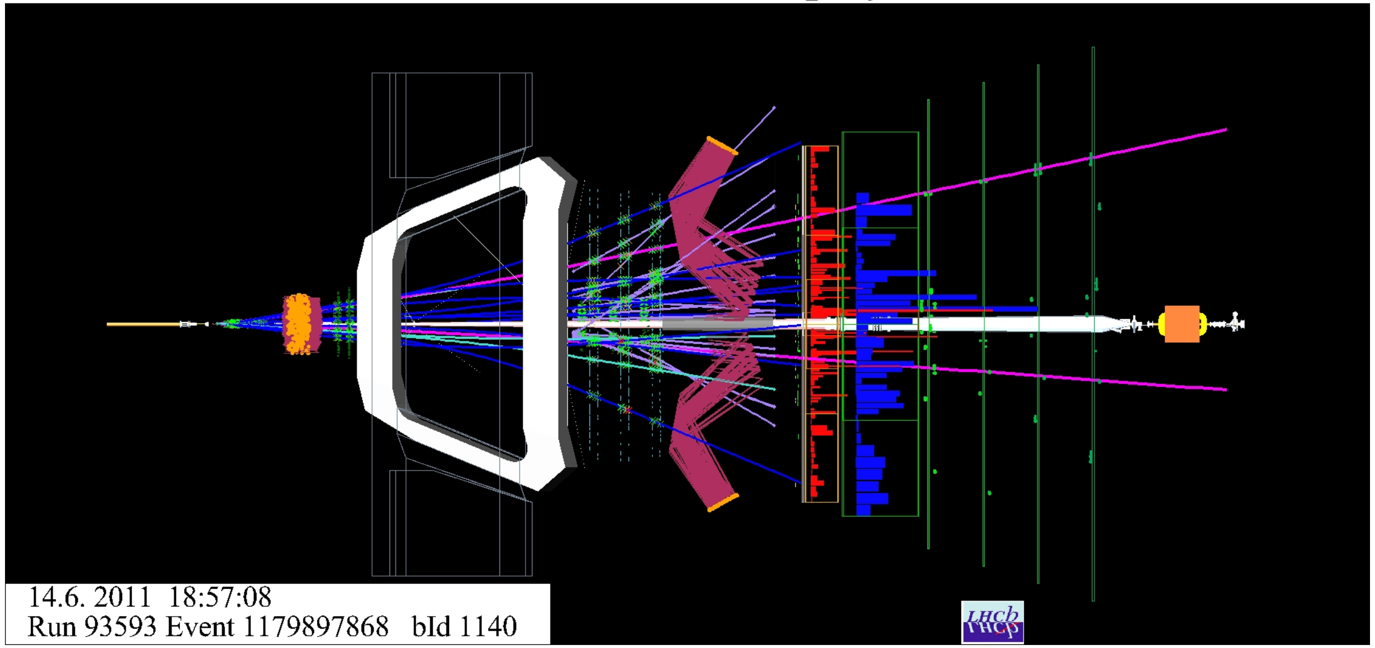


Science and Technology Facilities Council

Reconstruction - LHCb

- LHCb do something different – calibrate and do reconstruction 'live', no hardware trigger
  - Smaller amount of data, not possible with CMS/ATLAS

LHCb Event Display



Science and Technology Facilities Council

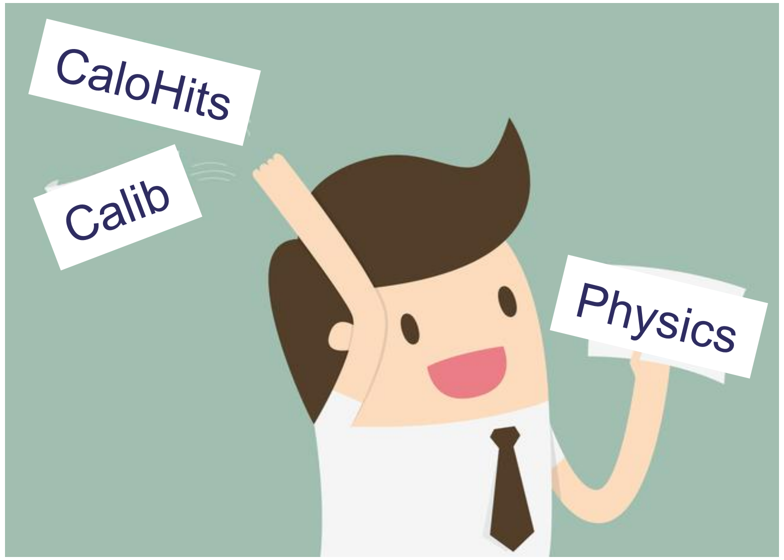
Reprocessing?

- Better understanding of detector
  - Improved calibration
  - Ageing of detector
- Better software
- Produce updated event format

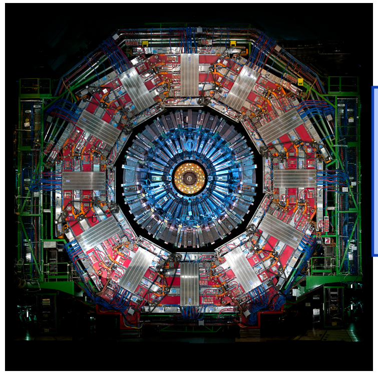
Data reduction

- Lot of information is recorded, but not required for analysis
- An AOD (Analysis Object Data) file is a distilled data format, containing physics objects as a minimum.
- Data formats are shrinking

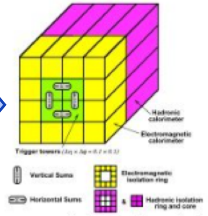
- Size of event: ~0.1 MB
- Size of CMS ‘nanoAOD’ format: ~ 0.001 MB



Summary of data reduction



40 million/s



100,000/s * 1.5 MB
150GB/s

1000/s * 0.75 MB
750 MB/s

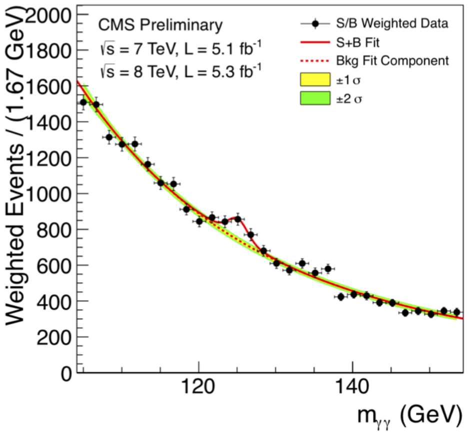
1000/s * 1.4 MB
1.4 GB/s

Permanent storage

“Online”

“Offline”

??? * 0.1 MB



Science and Technology Facilities Council

Full HEP software stack

- Mostly talked about the top two layers… but these sit on infrastructure.
- Experiment computing teams coordinate jobs.
- The Worldwide LHC Computing Grid runs the rest at labs and universities.

Layer	Responsible	Experiment 1	Exp. 2	Exp.3
6	Physicist end users	Selecting data, writing analysis code	…	…
5	Physics coders	Analysis frameworks, reconstruction code, calibration code	…	…
4	Computing teams	Central job submission, transfer management	…	…
3	WLCG	Setting up and running middleware, experiment support		
2	WLCG	Software infrastructure for CPU and storage		
1	WLCG	Installing, maintaining and decommissioning physical hardware		

Science and Technology Facilities Council

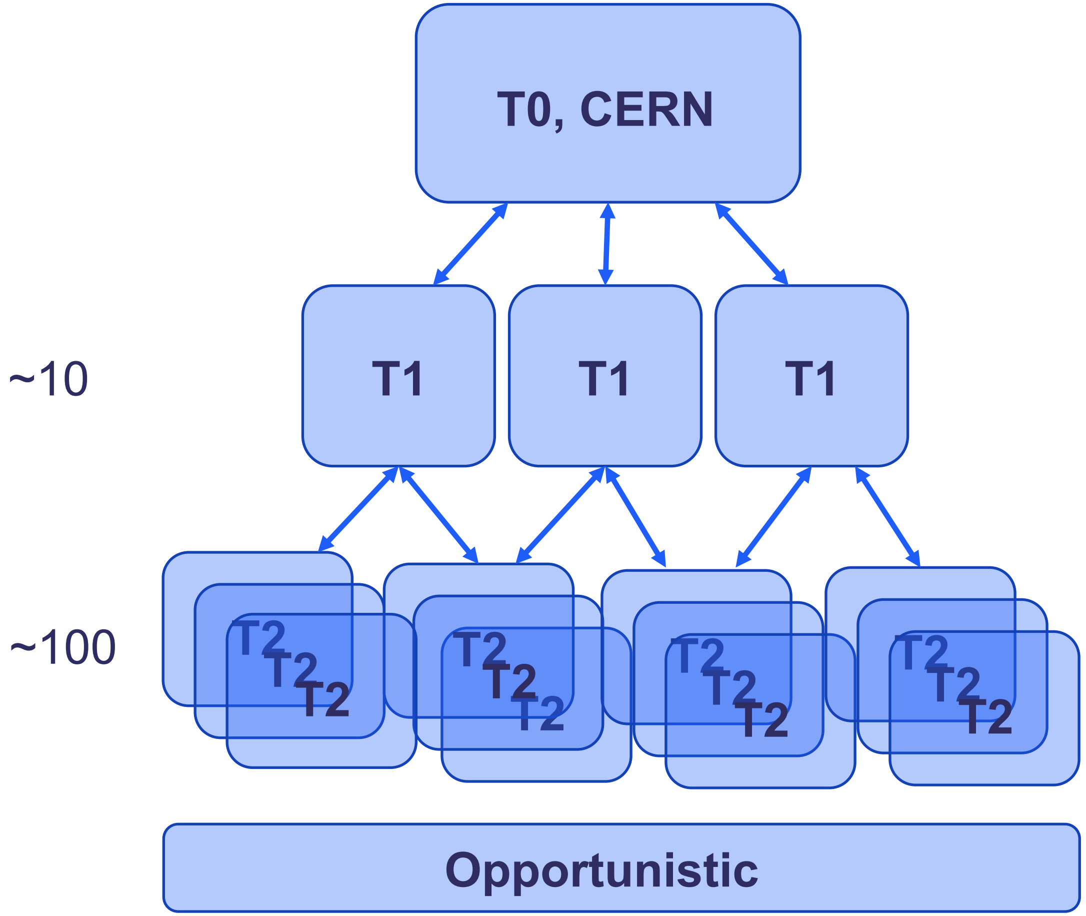
1 – Physical hardware

- Due to the way funding works we have computing resources spread across many sites.
  - Universities often provide a lot of additional resources/support.
- This is the opposite of the industry model which uses a small number of very large data centres.
  - The WLCG is consolidating storage at fewer sites but CPU will remain at a large number of sites.
- The result:
  - Resource provision is very heterogenous.
  - Hardware in production for a long time (very slow to move to new tech).

Science and Technology Facilities Council

https://www.gridpp.ac.uk/wiki/HEPSPEC06

Computing is distributed



- 150k CPUs HLT
- 100k CPUs, 100 PB disk, 270 PB tape
- 200k CPUs, 230 PB disk, 530 PB tape
- 240k CPUs, 230 PB disk

Science and Technology Facilities Council

Numbers are for all LHC experiments.
https://wlcg-rebus.cern.ch/apps/pledges/requirements/

Network and data access

- Major sites connected via 100 Gbit/s networks
  - But many smaller sites are 10-40 Gbit/s
- Move towards increased number of jobs accessing remote data
  - Can the network support such an increase in traffic?
  - Especially during data-taking?

Science and Technology Facilities Council

2, 3 – Software infrastructure and middleware

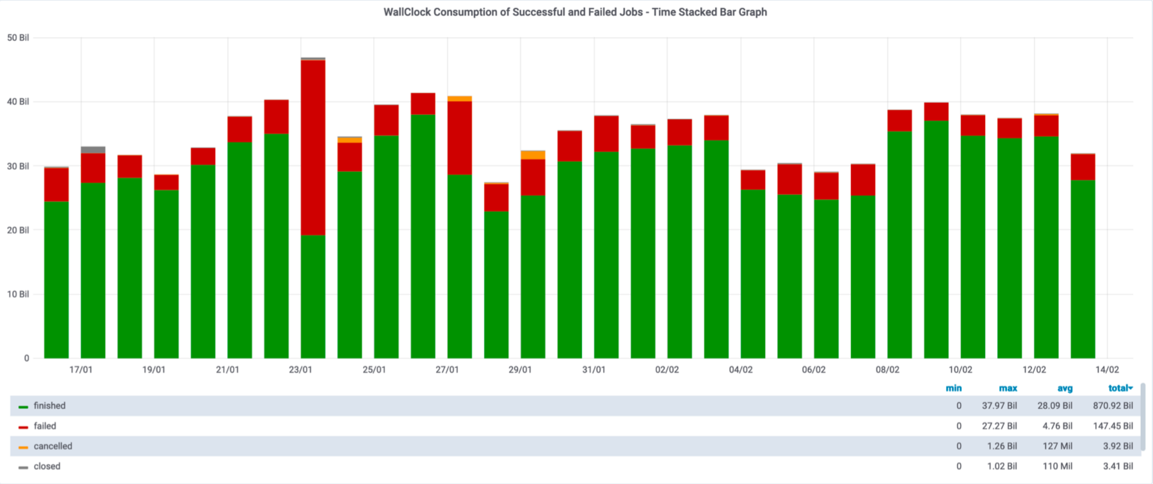
- Software infrastructure is the services run directly on top of the physical hardware
  - Storage infrastructure (e.g. dCache, Ceph)
  - Worker node infrastructure (e.g. HTCondor)

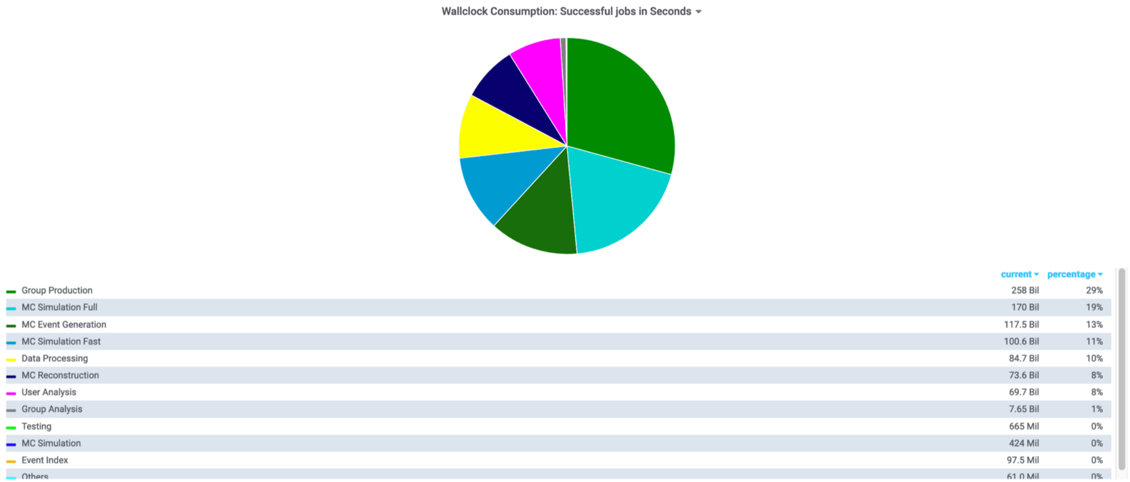
- Middleware provides access to the sites resources.
  - CVMFS for software access
  - Frontier/Squid for conditions data
  - CE for access to the batch systems.

4 – Experiment distributed computing

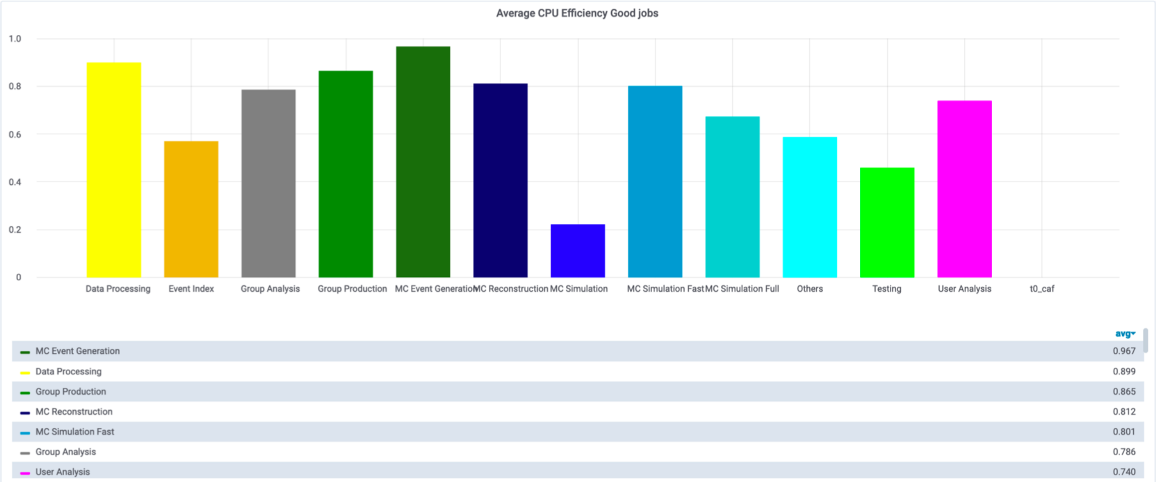
- Workflow management systems like Panda, WMAgent/CRAB and DIRAC/Gaudi run millions of jobs every day on the Grid.
  - Processing/production teams convert physics requests into jobs and check that they produce valid results.
- Data management systems like Rucio and FTS ensure data is correctly located.
- Operations team monitor systems to look for failures.

ATLAS Jobs



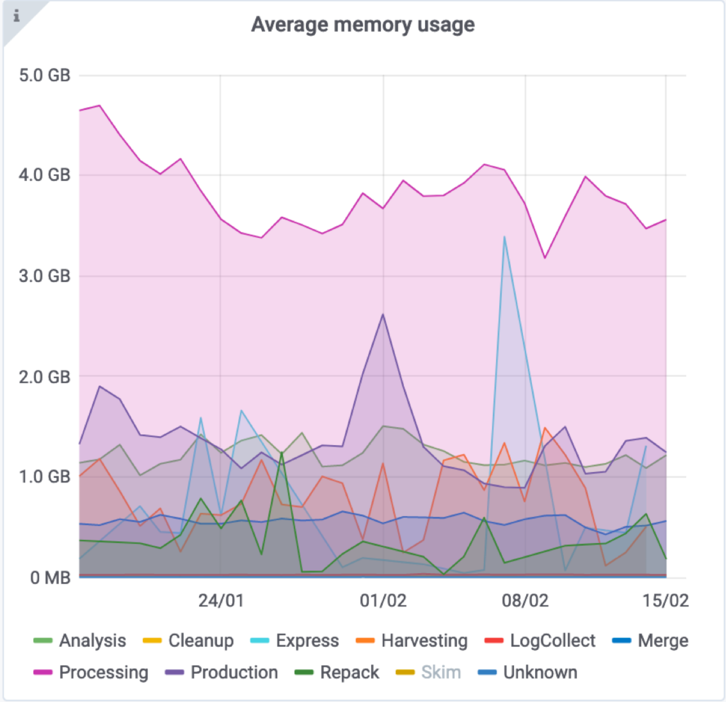


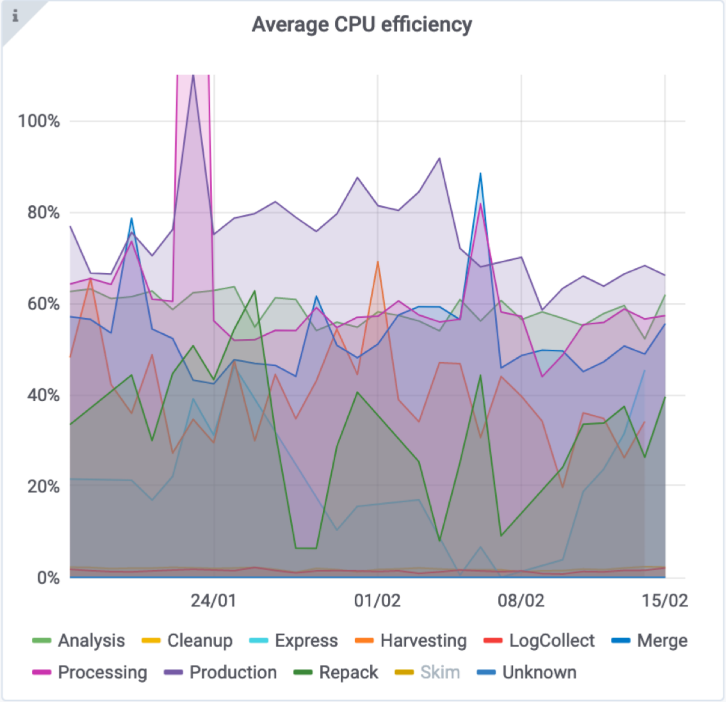
- Plots show ATLAS jobs run in the last 30 days.
- ~13% CPU Wall time lost due to failing jobs.
- Further CPU time lost due to job inefficiencies.

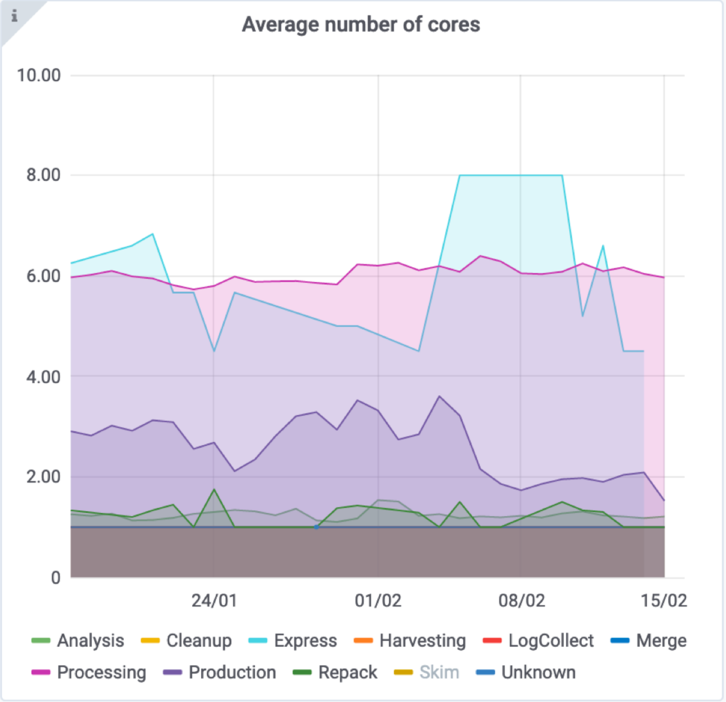


Science and Technology Facilities Council

CMS Jobs





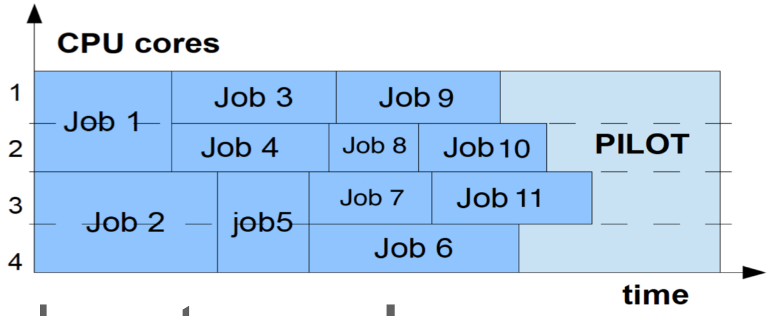


- Plots show the range of different requirements of CMS jobs.

Science and Technology Facilities Council

Further causes of inefficiency

- Jobs run in containers (sometimes multiple layers)
  - Lots of advantages on heterogeneous hardware and environments
  - Overhead – what is the impact on efficiency, can it be improved?
- Jobs run in ‘pilots’
  - Easy allocation of single/multi core jobs
  - Is time wasted?



- Job mix, don’t run all the data intensive jobs at once!
- Are jobs accessing data efficiently?
  - What is lost by accessing data offsite?

Science and
Technology
Facilities Council

Final words

- HEP has a complex software stack, with a number of well-defined roles.
- High-throughput regime, not HPC
- Amount of data will increase slightly in the next few years, and then significantly after 2027.
- Both processing and volume need huge reductions.
  - Lots of places where efficiencies must be made.

Science and Technology Facilities Council

Thank you

Facebook: Science and Technology Facilities Council

Twitter:@STFC_matters

YouTube: Science and Technology Facilities Council

References

- **Computing models in high energy physics:**
  - https://www.sciencedirect.com/science/article/pii/S2405428319300449
- **ATLAS Computing and software public results:**
  - https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults
- **LHCb Tier-1 Resource review:**
  - https://indico.cern.ch/event/862097/contributions/3631848/attachments/1973409/3284824/20200122-mcnab-lhcb-ral-review.pdf