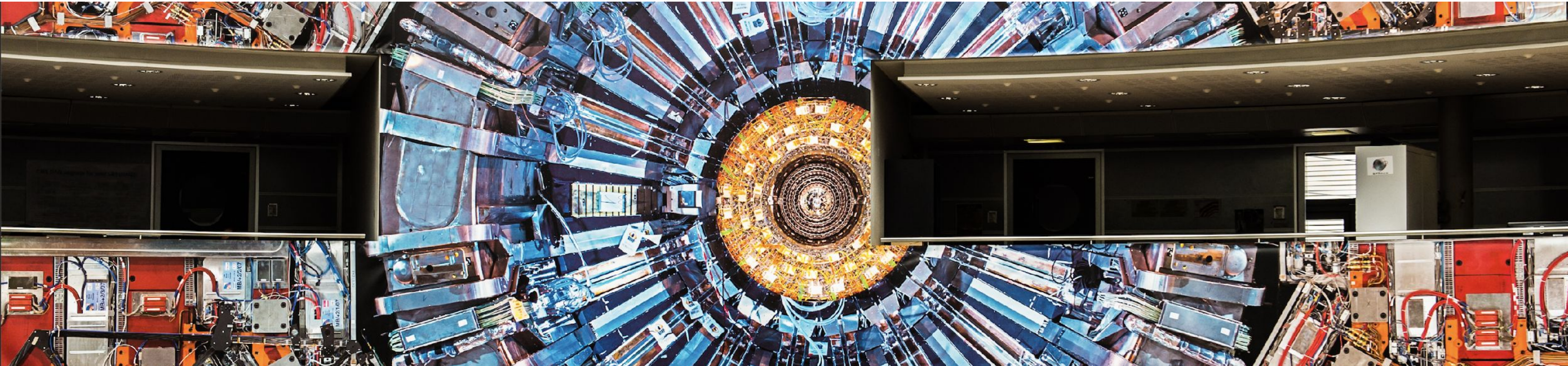
Looking Forward to HL-LHC
CMS O&C Viewpoint

CMS O&C

Offline Software and Computing

K. Ellis (STFC), J. Letts (UCSD), D. Piparo (CERN) - WLCG Workshop 2022 - November 7, 2022



Strategic Goals of CMS Offline Software & Computing (O&C)

Our mission is simple… to enable the physics program of CMS by:

- Efficiently using all of the computing resources available to us, while
  - Minimizing computing resource needs
  - Maximizing throughput
  - Minimizing job failures
  - Minimizing manual operations (effort)
- Completing requests in short, predictable amounts of time (*** not completely under the control of O&C ***)

The Run-3 Computing Model

(many elements of which are the same for HL-LHC)

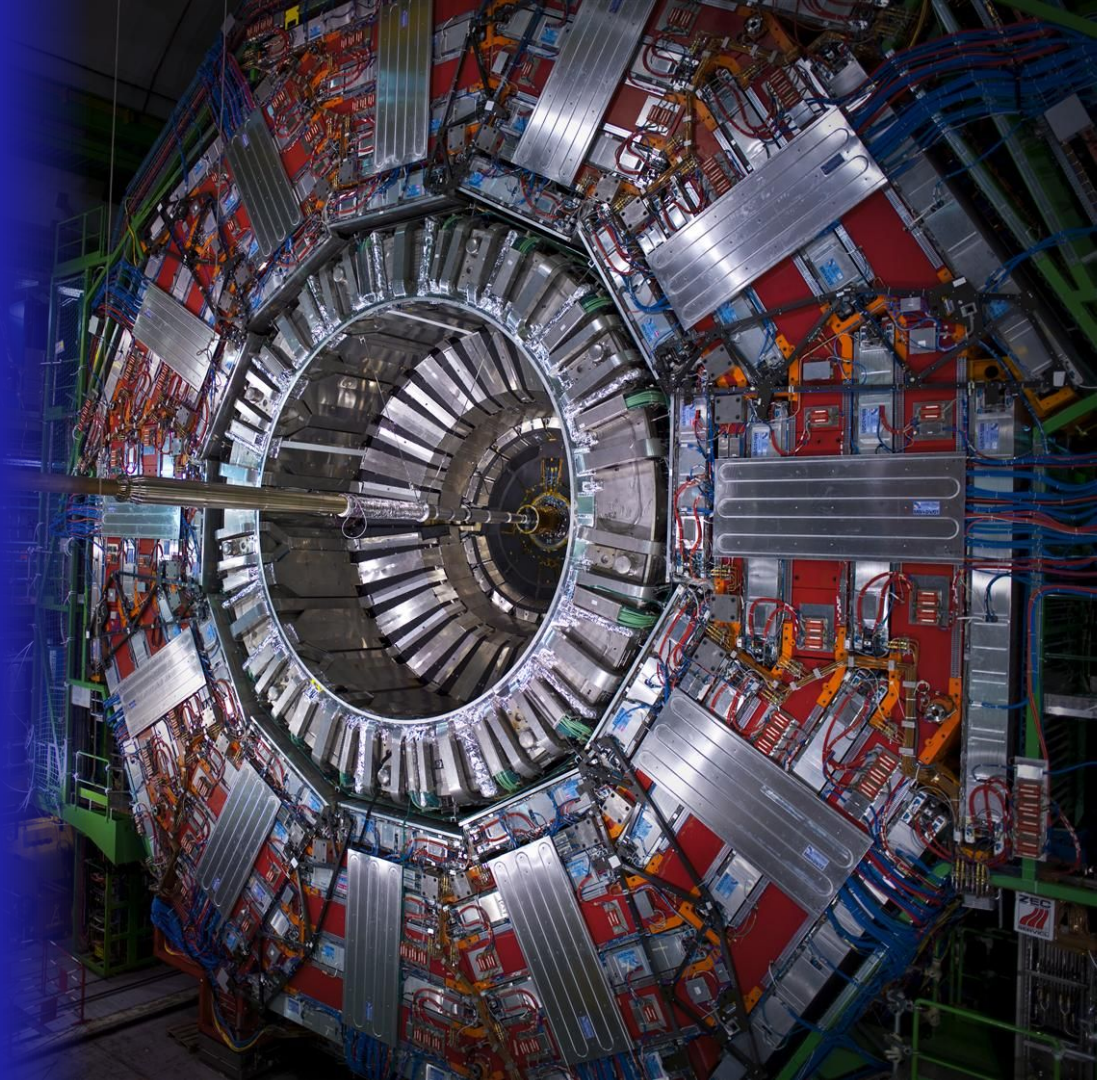
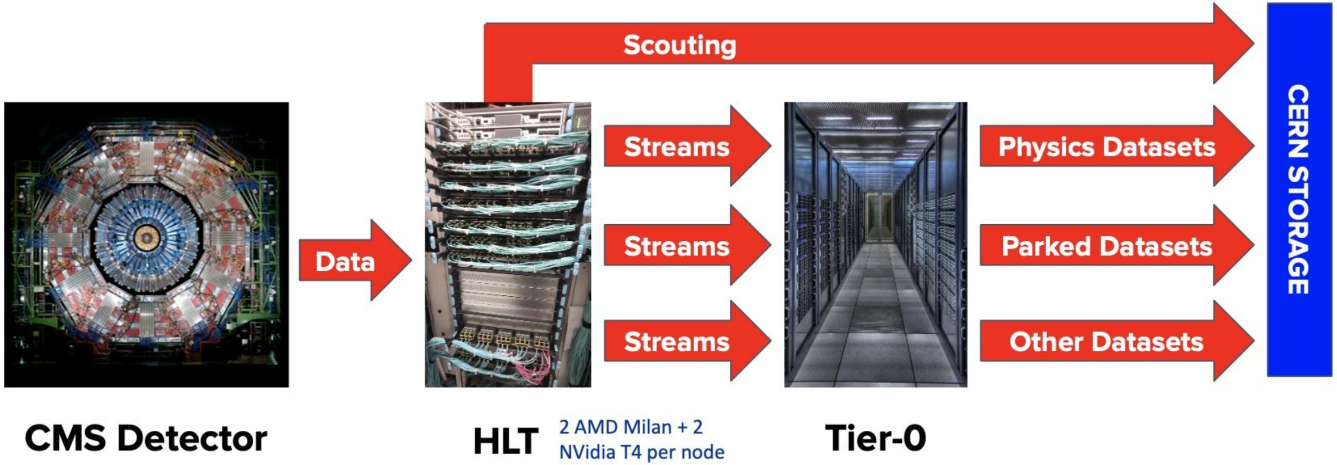


Image: © CERN

Data Flow from the Experiment to the Tier-0 from 10 Km

Data from the detector flows to the High-Level Trigger farm (HLT), which for Run 3 is equipped with NVidia T4 GPUs - increasing overall throughput by +70% in Run 3 already. The HLT sends streamer files of data separated according to the various trigger categories. Offline GPU workflows are being developed now (~10% runtime offloaded)!



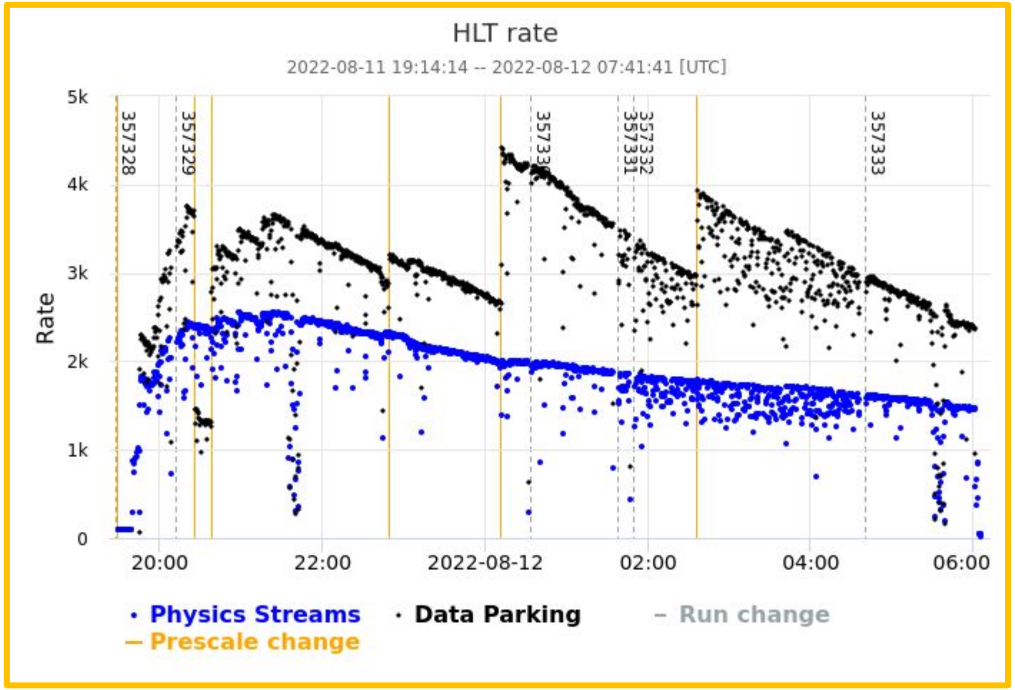
The Tier-0 processes the streamer files and outputs RAW and other specialized streams data to CERN storage, divided into “Primary Datasets” (PDs), e.g. dimuon, e-gamma, JetMET.

- One archival copy of the RAW physics data is made at CERN.
- A second, working copy is transferred to one of the CMS Tier-1 sites (except for “parked” and “scouting” data set types e.g.)

Run 3 is not “More of the Same”

… rather an opportunity to aggressively maximize the exploitation of the LHC program through different data taking modes:

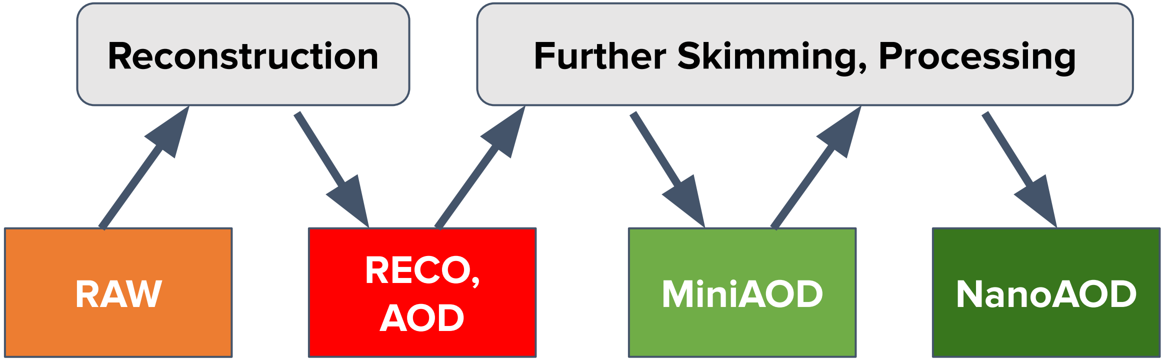
- **Prompt reconstruction: 1.3 kHz** performed at the Tier-0. MiniAOD & NanoAOD produced.
  - 2 RAW copies: one at CERN, the other at Tier-1’s
- **Data Parking: 3 kHz**. Single copy of the RAW on CERN tape. Mainly focussed on B-physics triggers.
  - Also (exceptionally) promptly reconstructed at the Tier-0 in 2022.
  - Peak prompt + parked >5kHz, i.e. HL-LHC rates (but not event sizes) already today.
- **Scouting: 30 kHz**. Data directly reconstructed at the HLT, also using GPUs. Overcome statistical limitations in particular regions of the phase-space.
  - No RAW data saved, custom data format (~10 kB/evt), partial reprocessing possible.



Reconstruction of CMS RAW Data

- The reconstruction sequence turns the binary output event data (RAW) into sets of physics quantities ready for data analysis (e.g. positions, energies etc.)
- All physics PDs are “promptly reconstructed” at the Tier-0
  - In 2022, we managed to also promptly reconstruct the parked data.
  - Not the final calibrations or alignment were used
    - Need data to derive them!
  - Typically CMS re-reconstructs the data at the end of the year, using the working copy at the Tier-1 sites, using improved calibrations and alignment information.
- Heavy Ion runs are particularly challenging: sustained maximal rates, stressing the infrastructure to the maximum level.

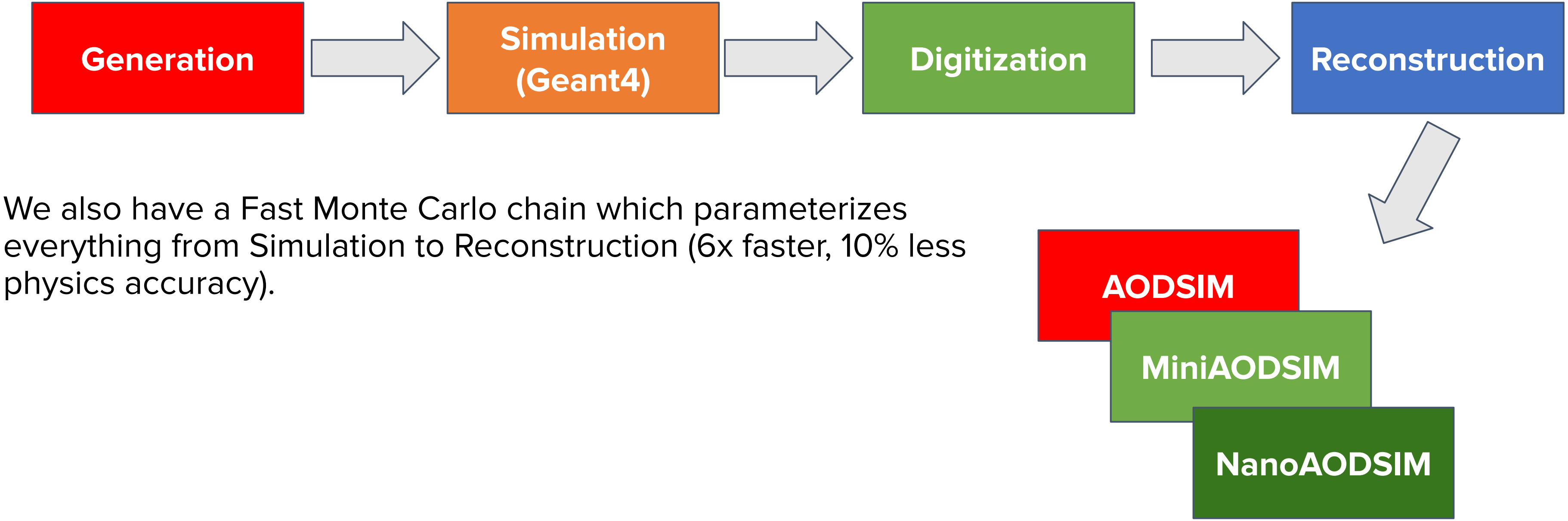
Not clear if we will be able or need to do this in Run 4-5, but it is in the model.



- Slim data formats are produced already as a step in the prompt reconstruction at the Tier-0
- NanoAOD can be produced in ROOT's RNtuple format already today.

Monte Carlo Simulations in Production

CMS has a multithreaded framework, and thread-safe code for all steps of Monte Carlo production except for some event generators (which are therefore encapsulated in processes).



Generation

Simulation (Geant4)

Digitization

Reconstruction

AODSIM

MiniAODSIM

NanoAODSIM

We also have a Fast Monte Carlo chain which parameterizes everything from Simulation to Reconstruction (6x faster, 10% less physics accuracy).

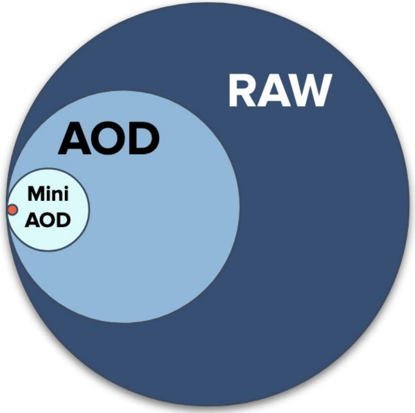
CMS Data Formats

CMS has been working with slim data formats for analysis for several years: *We saw the wall coming!*

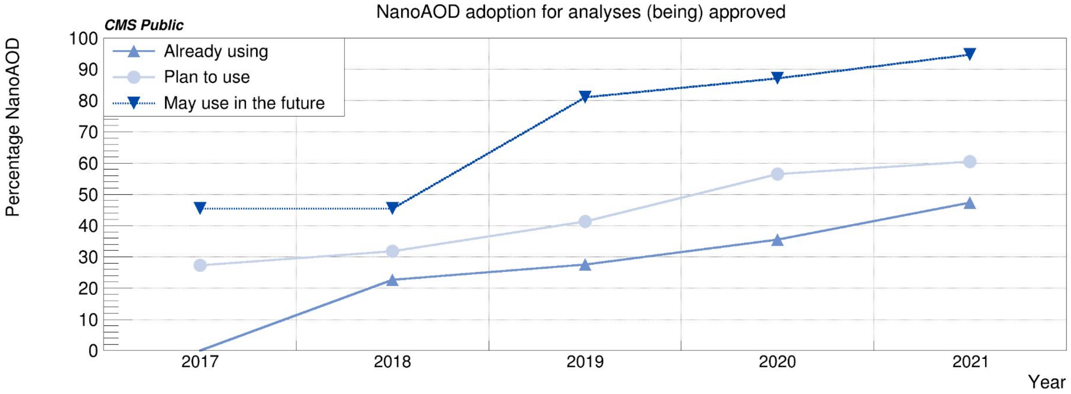
- NanoAOD: used by more than 50% of analyses today. Fundamental types and arrays.
- MiniAOD: covers almost all other analysis needs. Object-oriented data model.
  - Also adopted for Heavy Ion studies in Run 3
- Scouting format: custom, object-oriented, ~10kB/evt.

Analysis is currently done on the Grid, driven by users analysing the slim data formats. How this will evolve depends on the use cases that CMS needs to solve - to be discussed in the upcoming Conceptual Design Report (CDR) next year:

- Focus on finding solutions to well-posed problems, not *vice versa*.



RAW: ~1 MB
AOD: 600 kB
Mini: 60 kB
Nano: 1-2 kB



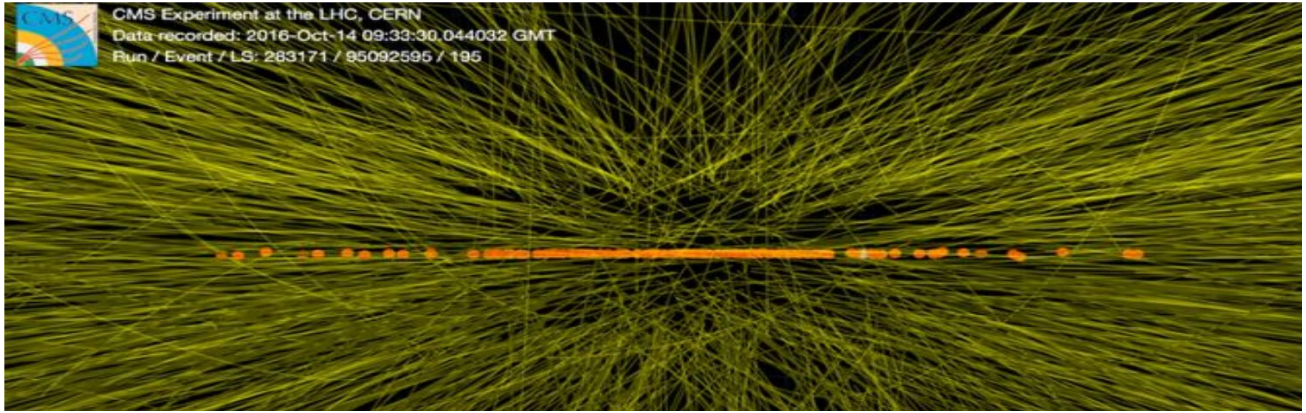
Pileup Simulation

One important aspect of Monte Carlo simulation is the overlaying of the minimum bias events, which are essentially a background to the main physics processes:

- During Run 3 in 2022, 54 pileup (PU) events on average
- By Run 4 PU=140, and for Run 5 PU=200.

Rather than simulate individual pileup events and overlay them, CMS creates *premixed* pileup libraries:

- One premixed PU event per physics event
- Libraries are hosted at CERN and Fermilab (CMS' largest Tier-1 site) and **read remotely with XRootD**
- Network is utilized to conserve disk space at Tier-1 and Tier-2 sites - PU libraries are large O(PB).

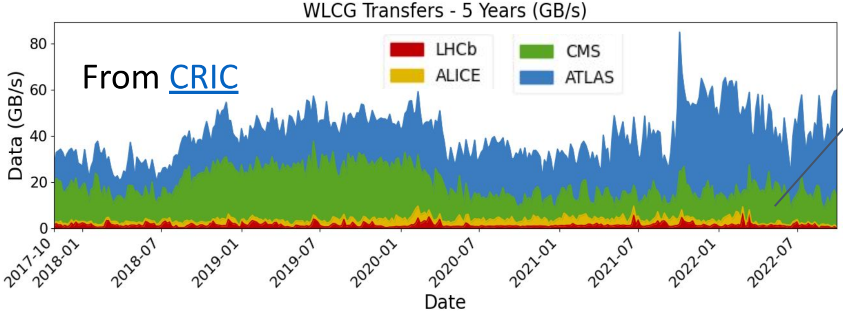

CMS Experiment at the LHC, CERN
Data recorded: 2016-Oct-14 09:33:30.044032 GMT
Run / Event / LS: 283171 / 95092595 / 195

Network Utilization: 2016-2022

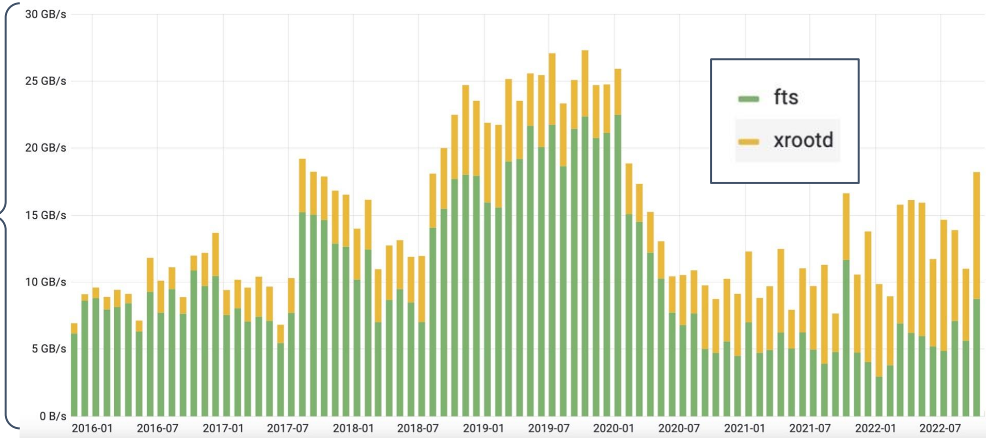
In 2020, important changes affecting workflow and data management were made:

- Transition to Rucio for scheduled data transfers - opportunity to review our data placement models
- "StepChain" workflows: all steps of the simulation carried out on the same workernode as a chain with no transfer of intermediate data products.
  - A use case conceived for HPCs, became universally used on the Grid.
- Improving XRootD monitoring (and network monitoring generally) is a big interest of CMS.

Network utilization under control so far in Run 3. Next year will be very interesting!

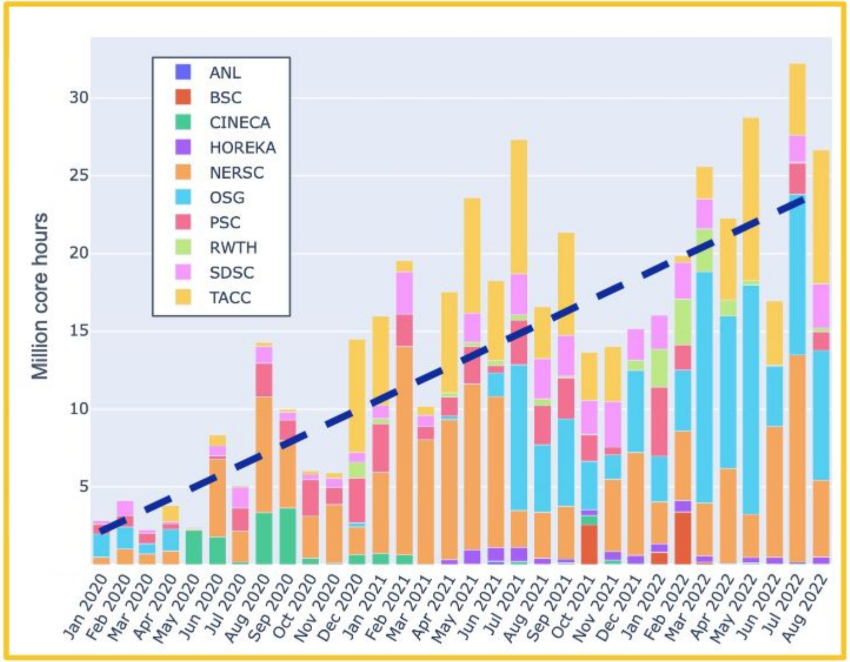


From CRIC



HPC Utilization

- HPC usage continues to grow in CMS. Two main models:
  - Transparent site extensions, e.g. RWTH, HOREKA, Marconi.
  - Resources accessed through a service, e.g. HEPCloud, OSG: used primarily for Monte Carlo production.
- We also use the old Run-2 HLT Cloud offline, as well as beyond-pledge contributions from sites.
- Full physics validation has been carried out on the POWER9 CPUs at Marconi 100, opening this architecture for CMS production.



CMS is preparing for HPC usage to grow throughout Run 3 and beyond.

Computing Resource Needs in Run 3 and HL-LHC

Image: © CERN

Computing Resource Needs in Run 3 and HL-LHC

Baseline computing resource needs are expected to grow modestly during Run 3, so far as we understand the running conditions from the LPC *today*.

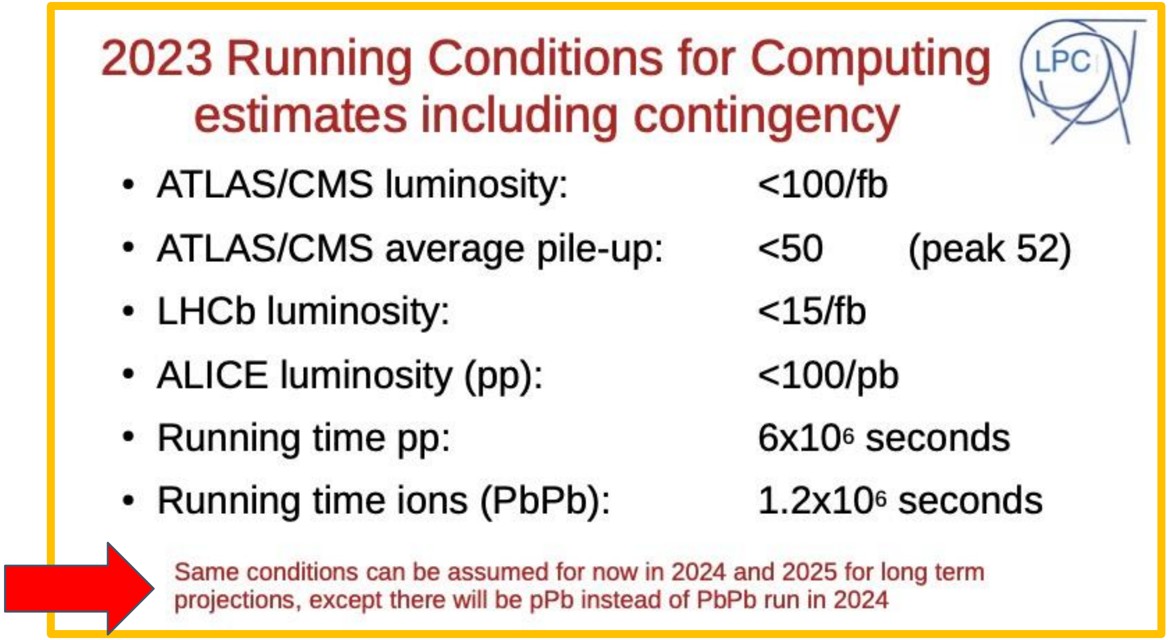
- HPC usage (which is significant and has been growing steadily for 3 years) allows CMS to expand the physics program beyond the baseline.
  - Many allocations have high latency to provision, and some with restrictions.
- Use of accelerators is an essential component of HLT workflows *online*.
  - Each node of the Run 3 HLT farm has 2 Nvidia T4 GPUs installed.
  - Offload (especially of tracking) to GPUs results in a 70% speed-up.
  - Higher offload targets for the HLT in Run 4 (50%) and Run 5 (80%).
  - Offline workflows using GPUs are being prepared, but CMS does not expect to make any formal requests for accelerators during Run 3, but will be able to make use of them opportunistically.
  - Performance portability choice for Run 3: Alpaka
- “Small” GPUs seem to be more suited for CMS
  - E.g. NVidia T4 or A10

Sites deploying GPU models with good scientific data processing performance and which Alpaka supports (or plans to support soon) would be particularly useful to CMS.

LHC Running Conditions for Computing

Each year we get updated guidance for computing (which may be different from the guidance for physics) from the LPC for the running conditions.

- Important input for the resource requests.
- Currently preparing the computing resource requests for 2024.



2023 Running Conditions for Computing estimates including contingency

• ATLAS/CMS luminosity:	<100/fb
• ATLAS/CMS average pile-up:	<50 (peak 52)
• LHCb luminosity:	<15/fb
• ALICE luminosity (pp):	<100/pb
• Running time pp:	6x10^{6} seconds
• Running time ions (PbPb):	1.2x10^{6} seconds

Same conditions can be assumed for now in 2024 and 2025 for long term projections, except there will be pPb instead of PbPb run in 2024

Lots of uncertainty: shorter running years, longer HI run in 2023, no HI run in 2022.

Bottom line: Even with the uncertainties, the computing resource needs during Run 3 will not change drastically from the original projections.

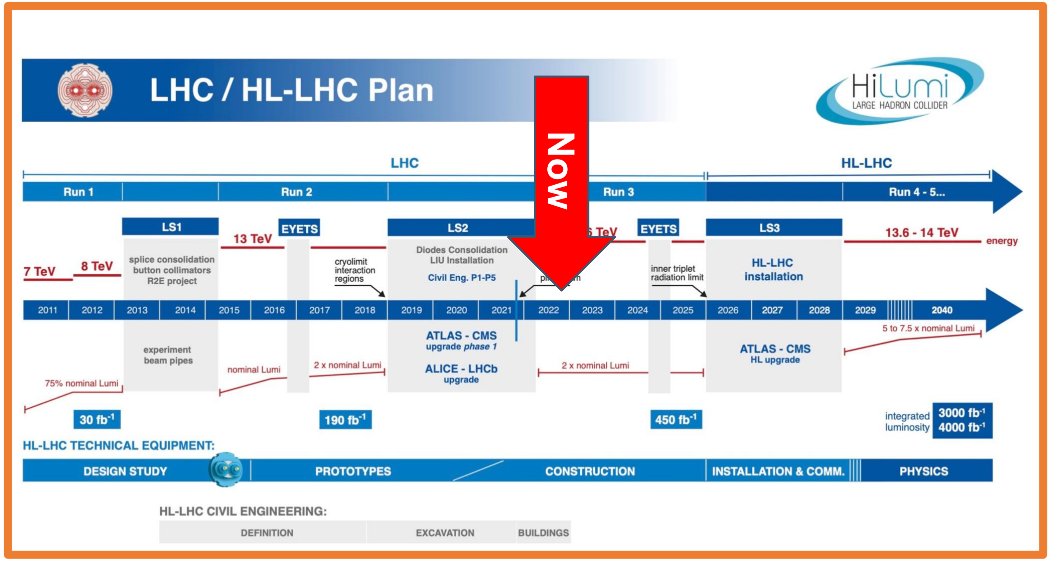
Inputs to the Computing Model for HL-LHC

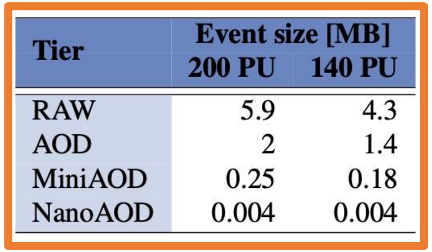
The original schedule for HL-LHC had pileup = 200 already in 2027, a large jump.

- Revised schedule from January 2022 is much more “adiabatic”
- First full year of HL-LHC running at PU=140 is 2030.
- PU=200 reached only in Run 5

Other CMS-specific inputs to the computing model:

- Prompt trigger rates from the HLT: 5 kHz in Run 4 and 7.5 kHz in Run 5 (from the DAQ/HLT TDR)
- RAW and derived data tier event sizes (still under study!)



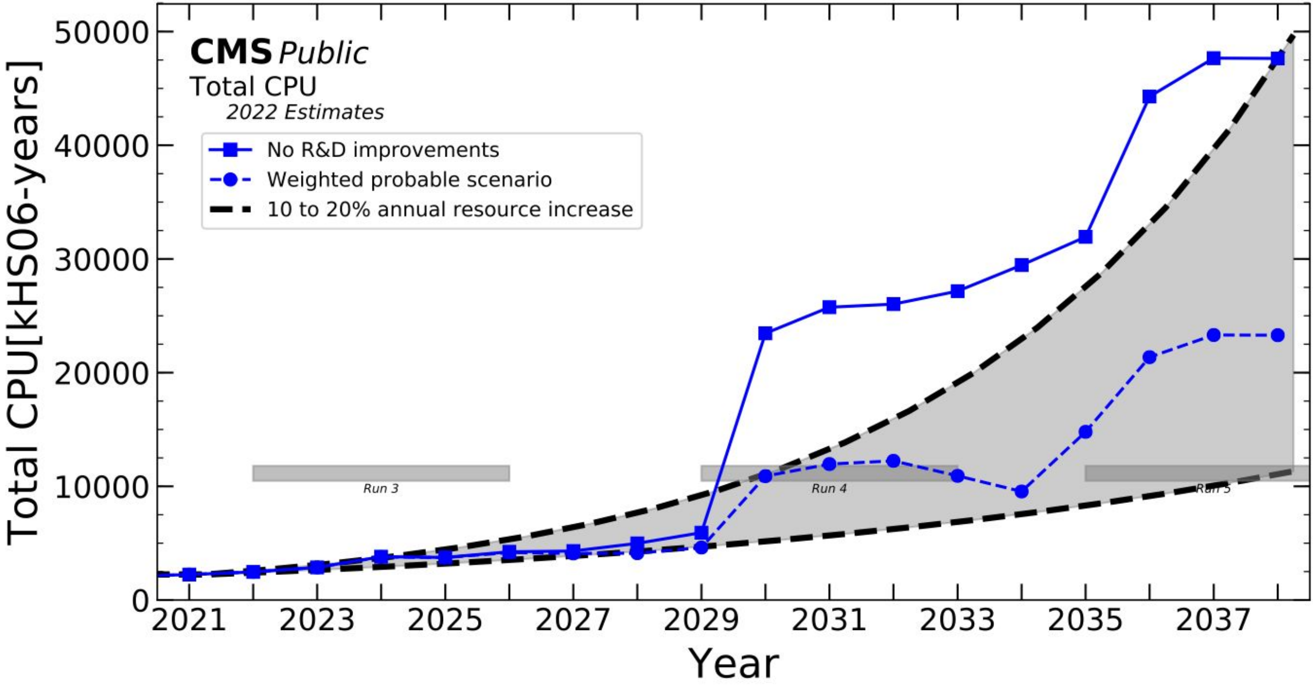


Tier	Event size [MB]	
	200 PU	140 PU
RAW	5.9	4.3
AOD	2	1.4
MiniAOD	0.25	0.18
NanoAOD	0.004	0.004

HL-LHC CPU Needs Projections

R&D lines and activities which can reduce needs between now and Run 4 include code optimizations in all processing steps (GEN, SIM, RECO), tracking improvements (improved p_T cuts, mkFit - new tracking introduced for Run 3), just to name some examples.

N.B. The lines have a different meaning in the ATLAS (connservative and aggressive R&D) and CMS (no R&D and most probable, i.e. conservsative, R&D outcome) plots.



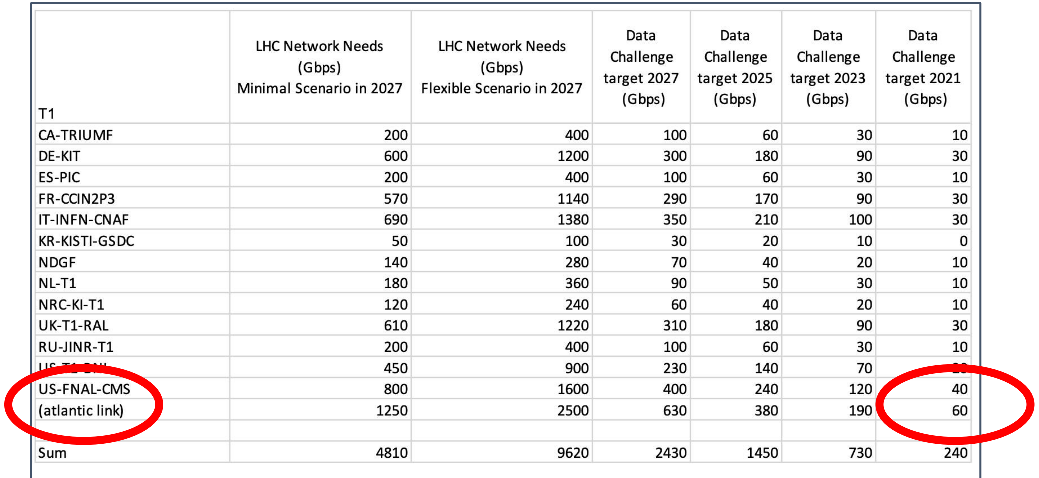
Storage projections shown in the backup slides.

Assessing Future Needs

As a baseline, WLCG & experiments did back-of-the-envelope estimates of HL-LHC needs by extrapolating Run 2 network usage by the experiments to PU=200 scales. A lot has changed since then:

- Run 4 start has slipped from 2027 to 2029, with the first full production year 2030 with PU=140 instead of PU=200.
- PU=200 will be reached in Run 5, more than a decade from now.

Still, it's a very good starting point:



T1	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	200	400	100	60	30	10
DE-KIT	600	1200	300	180	90	30
ES-PIC	200	400	100	60	30	10
FR-CCIN2P3	570	1140	290	170	90	30
IT-INFN-CNAF	690	1380	350	210	100	30
KR-KISTI-GSDC	50	100	30	20	10	0
NDGF	140	280	70	40	20	10
NL-T1	180	360	90	50	30	10
NRC-KI-T1	120	240	60	40	20	10
UK-T1-RAL	610	1220	310	180	90	30
RU-JINR-T1	200	400	100	60	30	10
US-T1-BNL	450	900	230	140	70	20
US-FNAL-CMS	800	1600	400	240	120	40
(atlantic link)	1250	2500	630	380	190	60
Sum	4810	9620	2430	1450	730	240

Table 2: data challenge target rates.

The capacity and performance of the transatlantic link is especially important for CMS.

CMS will update these projections in the upcoming Conceptual Design Report (CDR) for software and computing by late 2023.

Conclusions

- CMS has a flexible computing model based on the pillars of processing, storage, and networking, and a solid plan to reach HL-LHC scales.
  - Updated plans and projections in a CDR about a year from now.
- Run 3 is in many ways a testbed for HL-LHC, maximizing the physics reach.
  - Prompt reconstruction (1.3 kHz), parking data (3 kHz), and scouting (30 kHz)
- We are aware of our use of network: Improving monitoring is a priority!
  - Example: PU premixing uses XRootD over the network to conserve storage.
- CMS already offloads much of the online processing to GPUs, offline will soon.
- Efficient use of HPCs is expanding: goal to have this be transparent to central operations.

References to Public CMS Documents & Plots

- CMS Phase 2 Computing Model: Update (CERN-CMS-NOTE-2022-008)
https://cds.cern.ch/record/2815292/files/NOTE2022_008.pdf

- CMS Offline Software and Computing Public Results (2022):
https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

- The Phase-2 Upgrade of the CMS Data Acquisition and High-Level Trigger TDR (2021)
https://cds.cern.ch/record/2759072/files/CMS-TDR-022.pdf

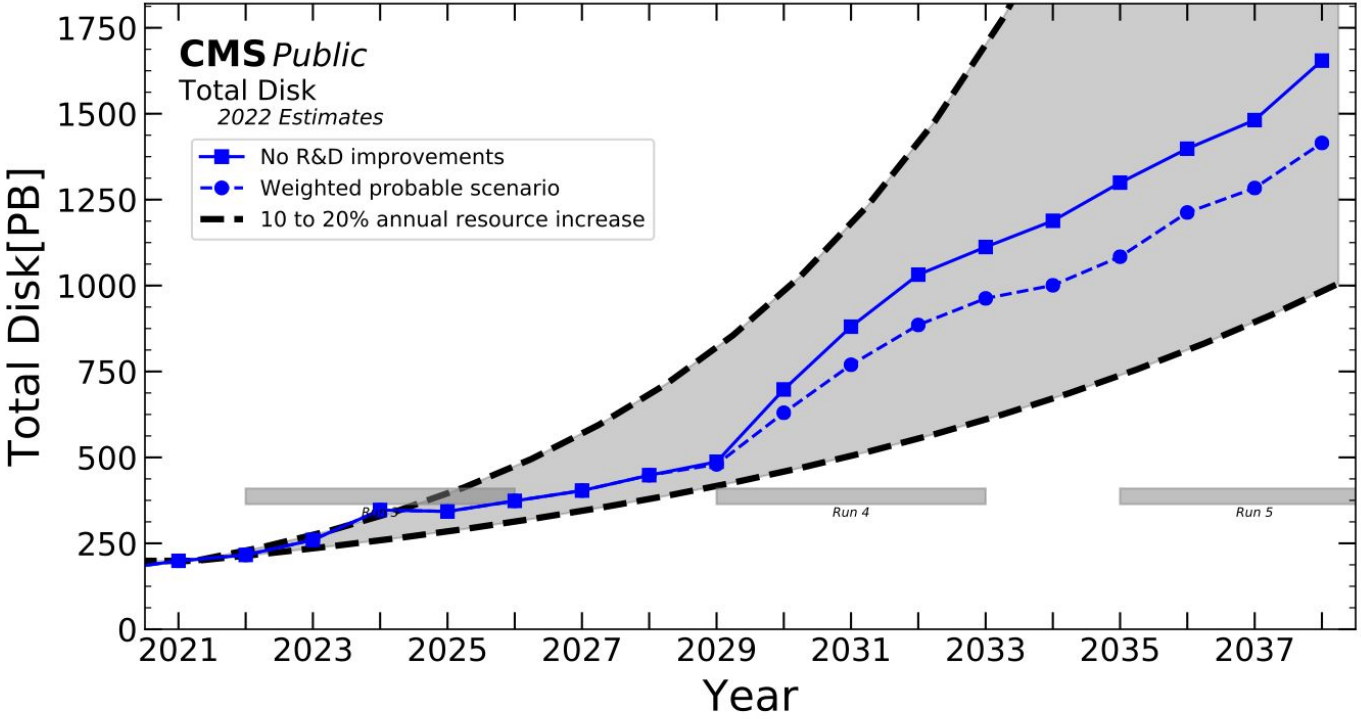
- CMS Update to the LHCOPN-LHCONE Meeting #49 (October 2022)
https://indico.cern.ch/event/1146558/contributions/5063563/attachments/2534544/4361711/221025-CMS-LHCOPN-LHCONE.pdf

Backup Slides

Image: © CNAF

Computing Needs Projections for HL-LHC

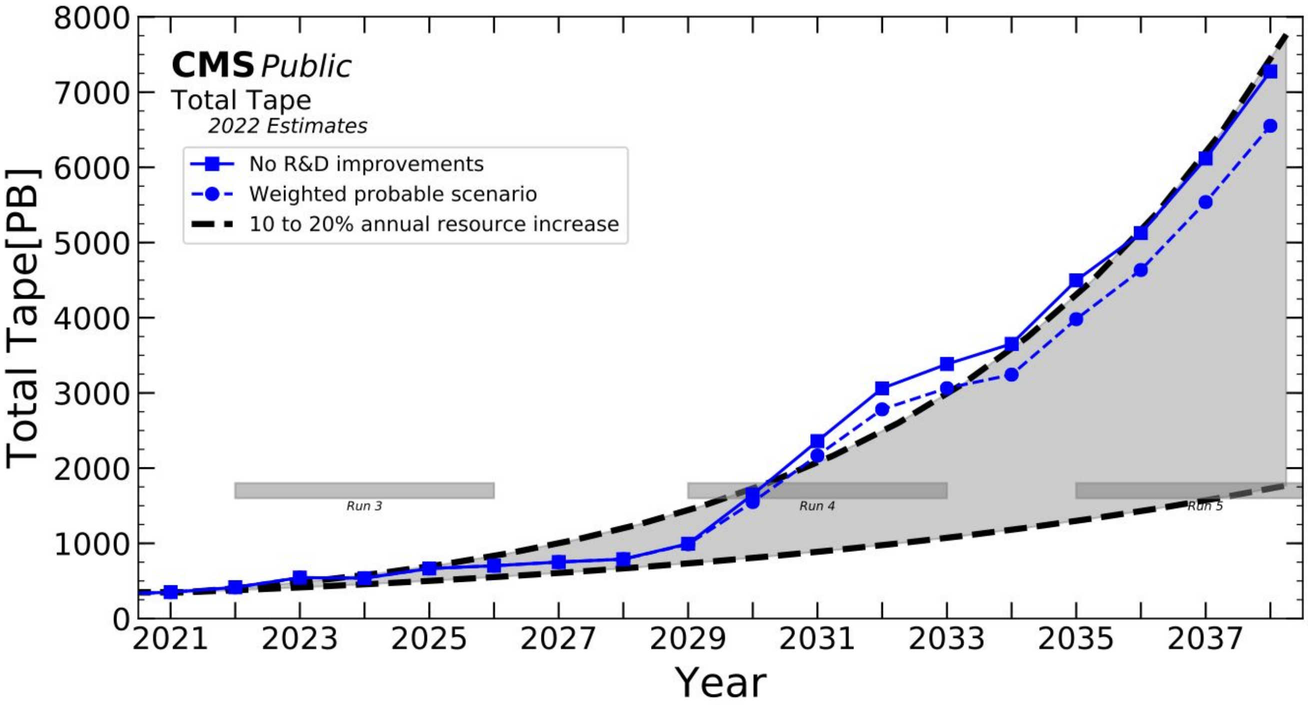
Disk storage (or whatever QoS this evolves into by Run 4) is one of the most precious resources for the experiment.



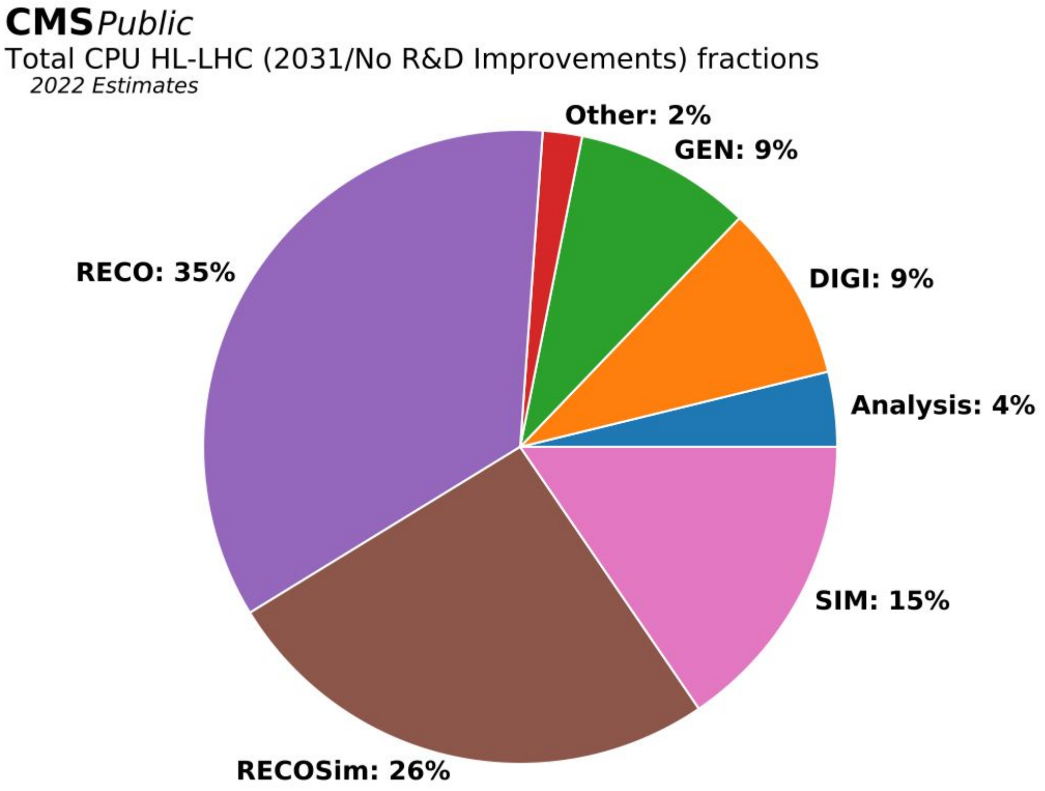
K. Ellis, J. Letts, D. Piparo, WLCG Workshop, November 7, 2022

Computing Needs Projections for HL-LHC

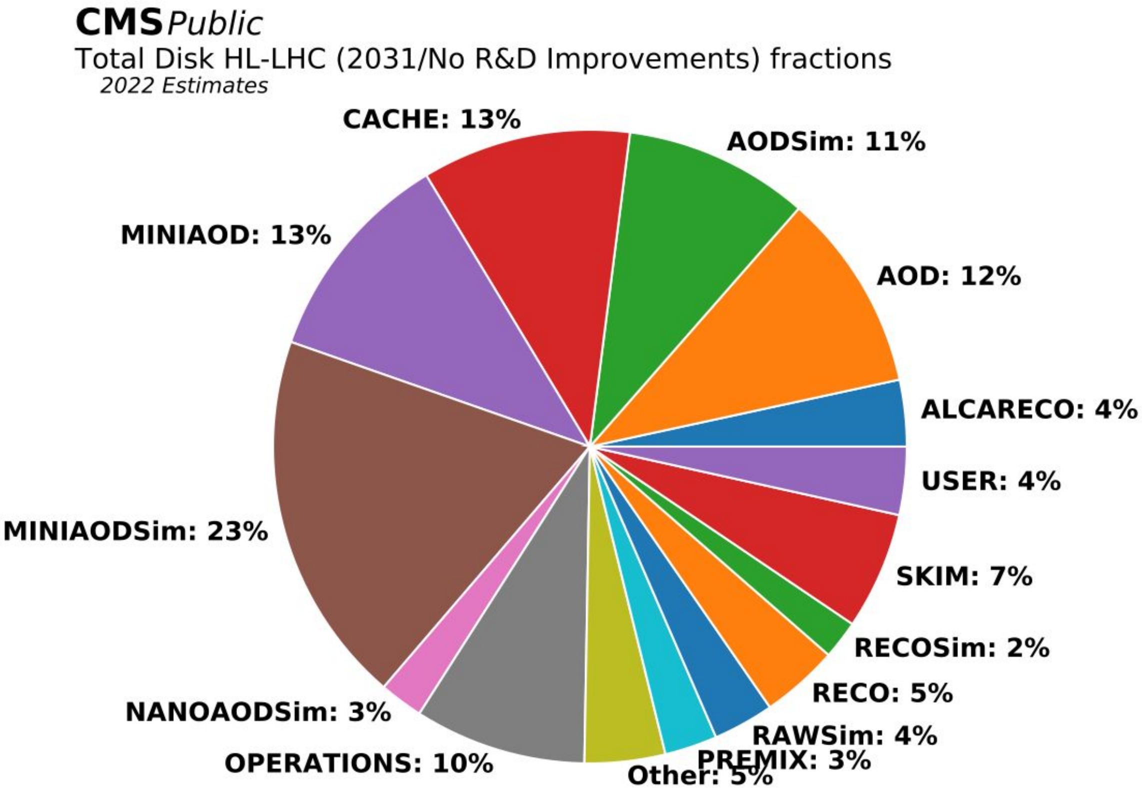
Tape storage (or whatever QoS this evolves into by Run 4) needs partly driven by the volume of the RAW data (archival & working copies), as well as cold storage of larger simulated data formats.



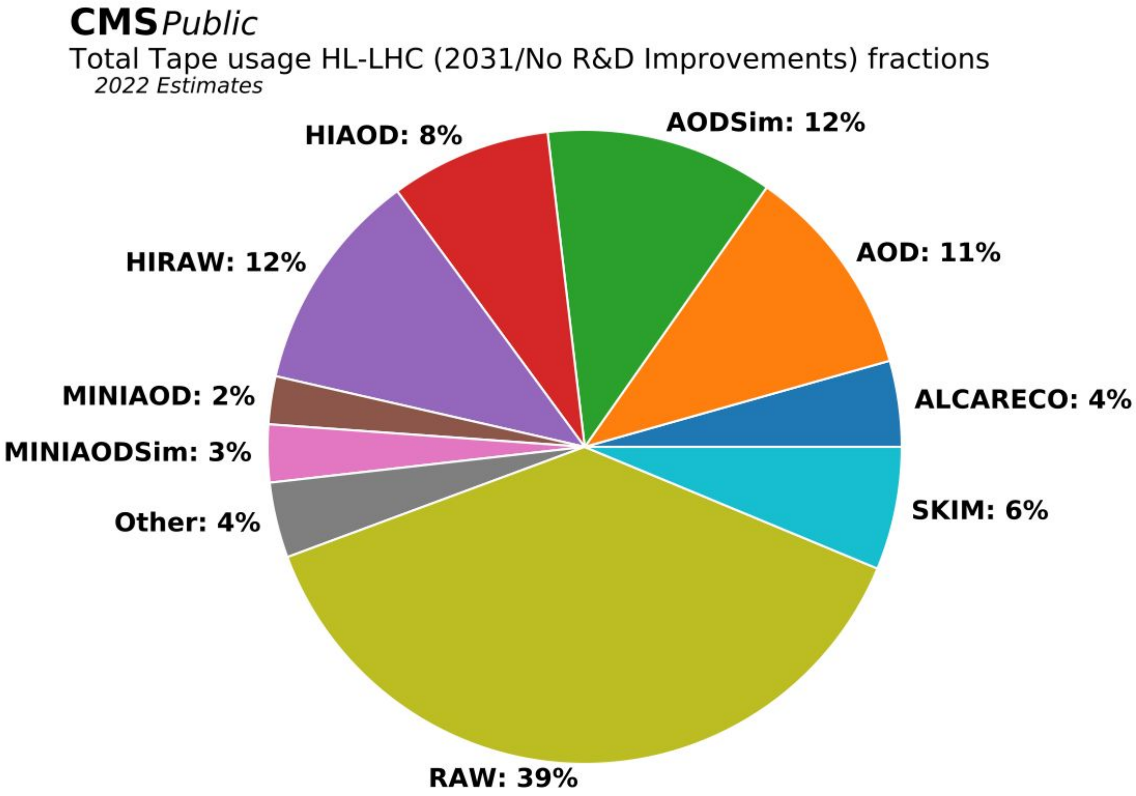
HL-LHC CPU Utilization Pie Chart



HL-LHC Disk Utilization Pie Chart



HL-LHC Tape Utilization Pie Chart


CMS Public
Total Tape usage HL-LHC (2031/No R&D Improvements) fractions
2022 Estimates
HIAOD: 8%
AODSim: 12%
HIRAW: 12%
AOD: 11%
MINIAOD: 2%
ALCARECO: 4%
MINIAODSim: 3%
SKIM: 6%
Other: 4%
RAW: 39%