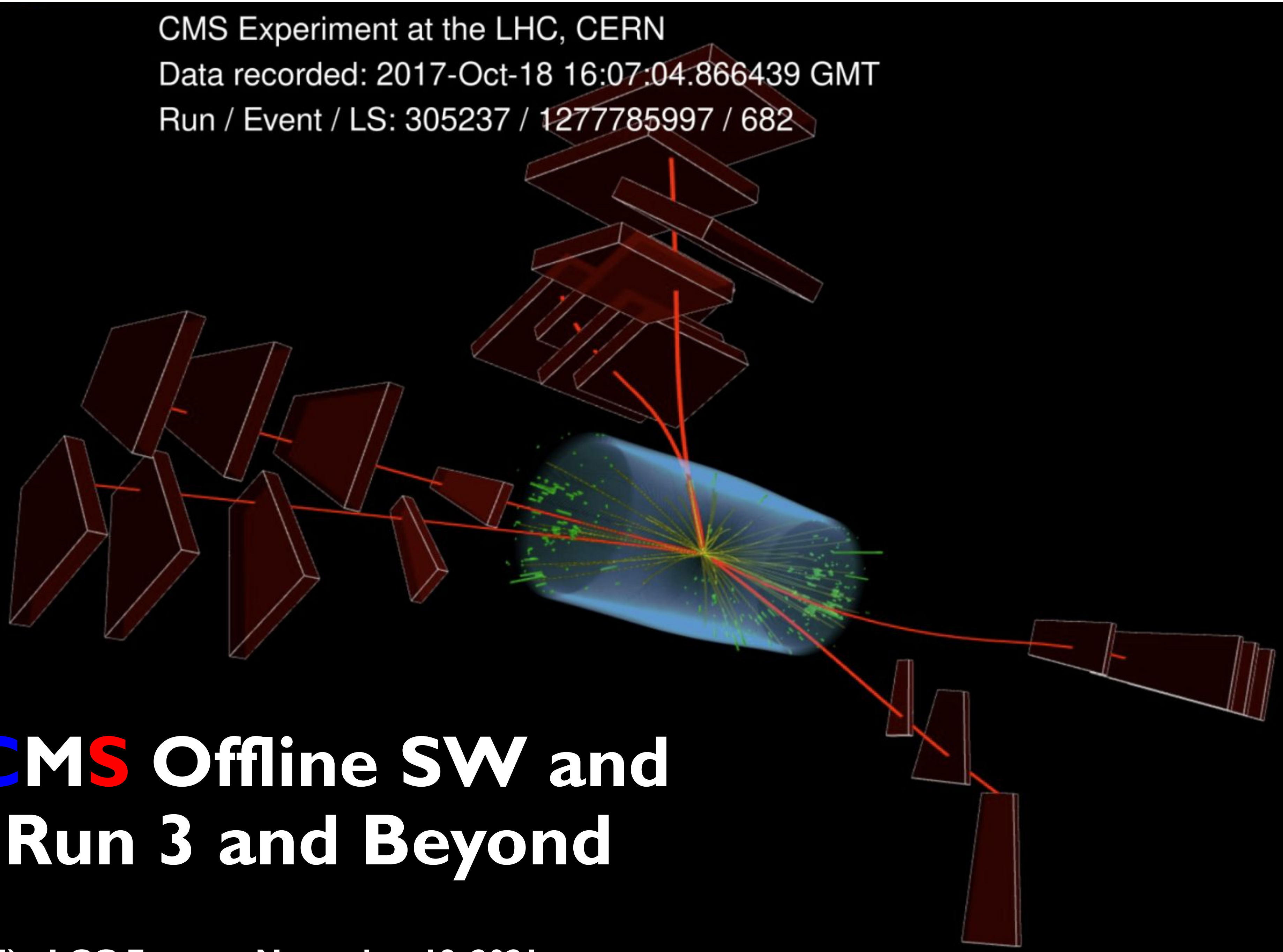


Highlights of CMS Offline SW and Computing in Run 3 and Beyond

J. Letts (UCSD), D. Piparo (CERN) - LCG France - November 18, 2021

Topics

- Offline Software and Computing in CMS
  - The meaning of Run 3 for us
- Computing resources in 2022 and preview of 2023
- Highlights of innovations delivered during LS2
  - Computing tools and (common) software

CMS Offline Software and Computing

Image: © CERN

The Goal of O&C

Deliver datasets to enable the CMS Physics Programme and the software to produce, process and analyse them

Many, many interesting **activities at the bleeding edge of software and hardware technologies** stem from this simple formulation!

We strive to make our computing model more and more flexible to be able to adapt to future price fluctuations of computer hardware

The O&C Area

We are on Mattermost, come and chat with us!

Coordinators

Devops approach

Core Software	Computing Ops	Dyn. Res. Provisioning	Facility Services
Simulation	Workload/Data Mgmt Devel	Reconstruction	Resource Management
Monitoring & Analytics	Release Planning Ops	Upgrade Software	Submission Infra
Analysis Infra & Support	Machine Learning *	Upgrade R&D and TDR	Web Services & Security
Generators *	L1 Software **	DPOA ***	

* Joint with Physics
** Joint with L1 DPG
*** Joint with CB

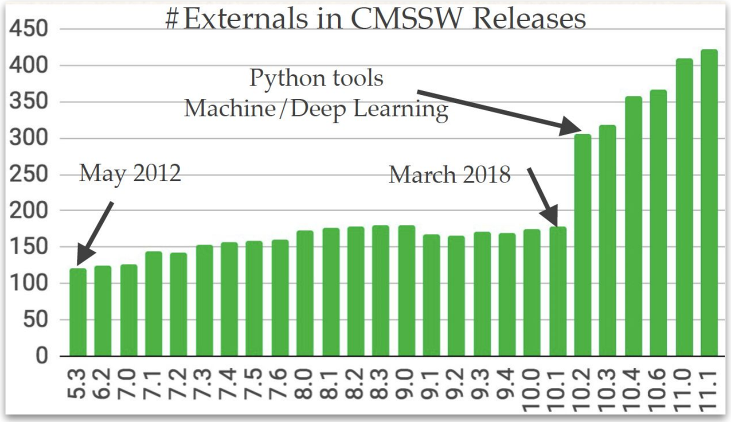
Computing Resources Board

19 groups, a very broad set of expertises

Since a long time, we had no group coordinator affiliated with a French institute. Unique case among countries pledging Tier-1 resources to CMS

Offline Software

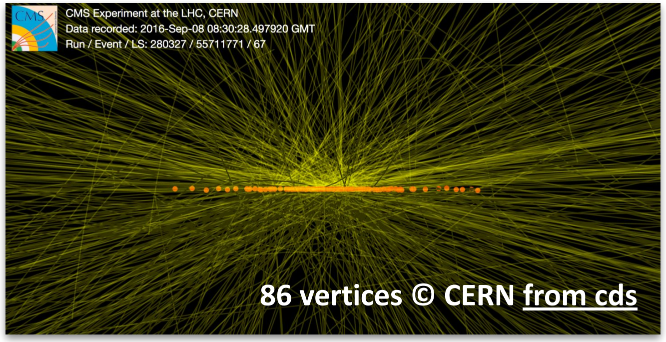
- **A crucial asset, built during the years**, condensing invaluable (detector) expertise
  - 1100+ commits/month, 100+ committers/month
- Same codebase for HLT & Offline, of today & HL-LHC
  - Big advantage for CMS
- CMSSW is [on Github](#) since 2012
  - And was open source since the start
- ~5M of C++ + other languages
  - ~400 external packages supporting that code



CMS Software: an asset to be preserved and grown with dedicated human effort

Run 3

- Run 3: our top priority
- **Challenging running parameters, e.g. lumi levelling @40 PU in 2022 and @50/55 in 2023/2024**
  - Approaching HL-LHC scenarios
  - Run 4 planned to start with ~100 PU for 1-2 years (see HL-LHC ultimate lumi projections here)



Very ambitious physics programme by CMS: not just more of the same!

- High rate prompt reconstruction: ~1.5 kHz
- Data Parking: write on tape at Tier-0, re-construct at the end of the year
  - Very successful during Run 1 and Run 2, e.g. 10B unbiased B decays recorded in 2018
- Data Scouting: stream reconstructed at HLT to explore phase space otherwise not accessible, e.g. low mass resonances
- Heavy Ions: very high statistics, virtually no pt cut, push our infrastructure to the limit
  - E.g. 10 GB/s to be recorded on tape at CERN

Computing and Software in Run 3

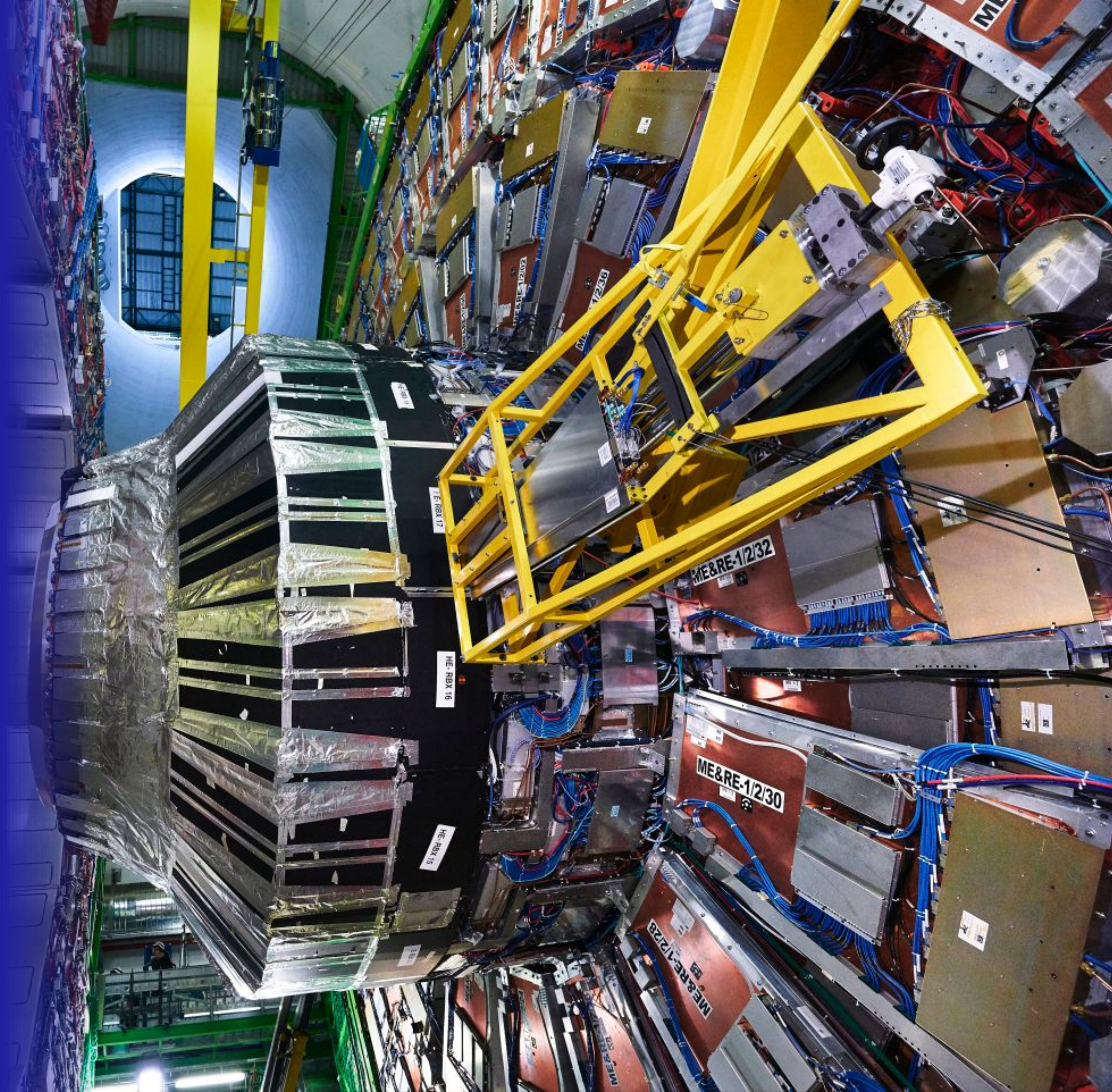
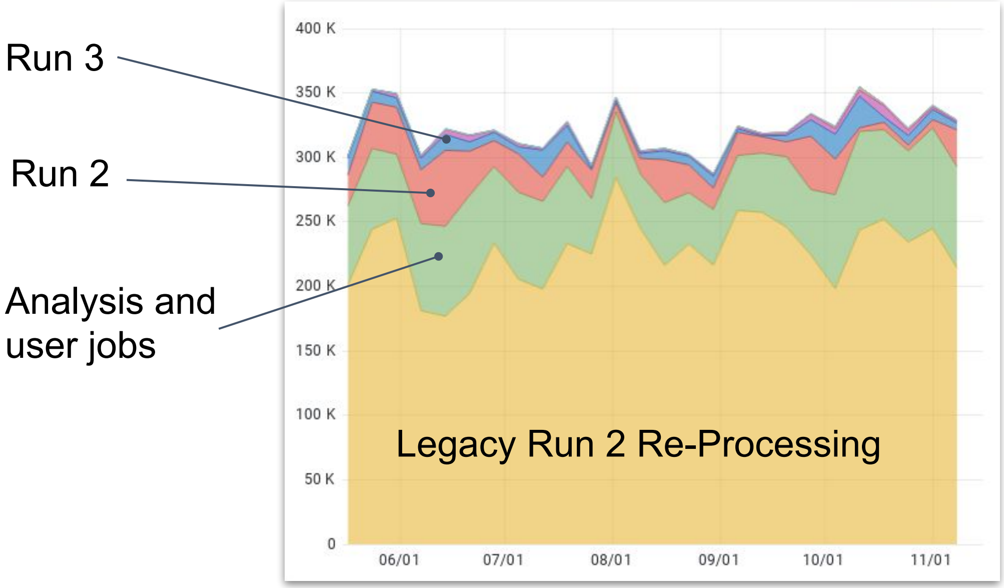


Image: © CERN

Usage of Resources and the Role of France

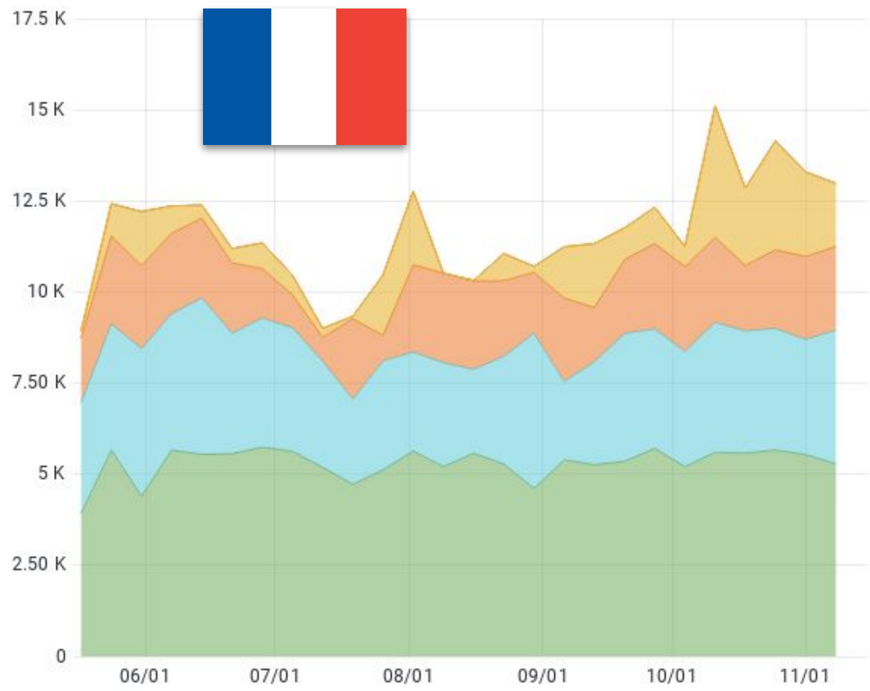
Running Cores



French contributions:

- **Tier-1: 9.8% of all 3 resources**
- **Tier-2 resources (CRIC)**
  - **Disk: 6.1%**
  - **CPU: 6.5%**

+ Excellent network!



	min	max	avg	current	total
TOTAL	8.93 K	15.1 K	11.6 K	13.0 K	302 K
T1_FR_CCIN2P3	3.93 K	5.75 K	5.31 K	5.29 K	138 K
T2_FR_GRIF_LLR	2.15 K	4.29 K	3.24 K	3.66 K	84.1 K
T2_FR_IPHC	655	2.46 K	1.96 K	2.30 K	51.0 K
T2_FR_GRIF_IRFU	0	3.59 K	1.10 K	1.73 K	28.5 K

France provides a fundamental contribution that allows to enable the Physics Programme of CMS

Tier-1: not only a key resource for CMS, but allows to participate hands-on to ongoing R&D activities for Run 4!

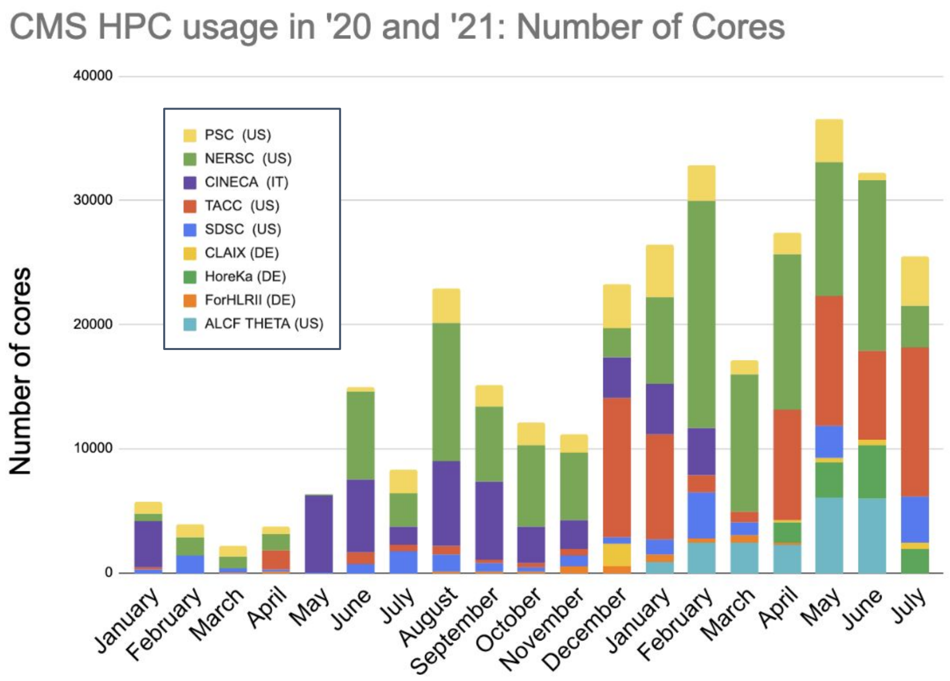
HPCs and Challenges

- Substantial national and supranational investments in HPCs globally: they are there to stay!
  - Exascale machines will be well available by HL-LHC
- Being able to **use accelerators helps leveraging HPCs**
  - **But is not sufficient**
- There are other hurdles to overcome to use HPCs for HEP
  - HEP and HPC: **language spoken by experts can be different**
  - **Data access** (access, bandwidth, caches ...): HEP has data processing applications (HTC)
    - HPCs are “storageless sites”
  - **Submission of tasks** (MPI vs Batch systems vs proprietary systems)
  - Environment **less open than Grid** one (OS, access policies, ...)
  - **Node configuration** (low RAM/Disk, ...)
  - Primary **architecture** (x86_64, Power9, ARM, proprietary, ...)
  - Relationships between providers and **CMS are decades long**

HPCs: Current Status

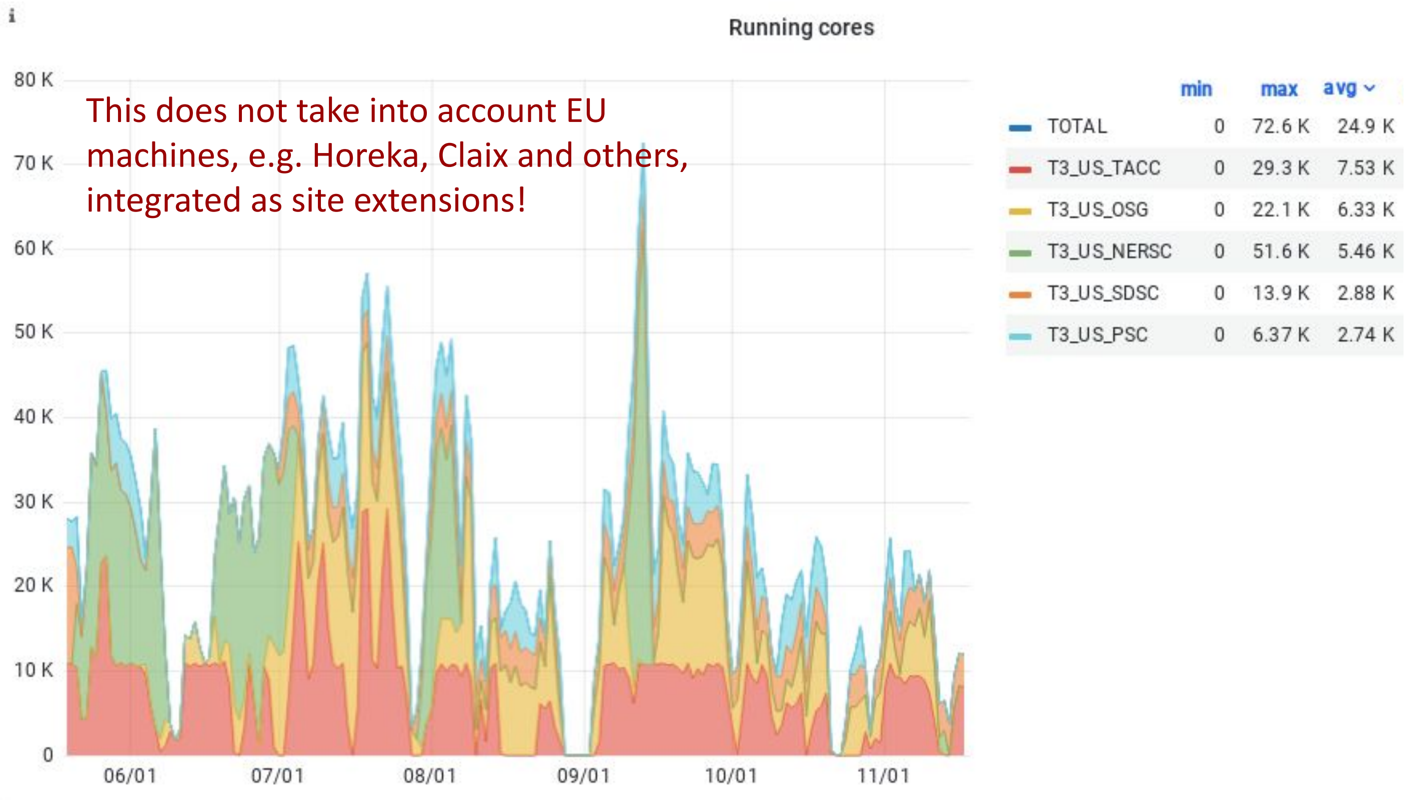
- CMS uses HPCs in production for all steps of the processing: gen, sim, digi/pu-mimx, reco, mini/nano creation: not a prototypal utilisation!
  - **Capacity used by CMS at HPCs tripled in 2020 wrt 2019 and tripled again in 2021 (so far) wrt 2020**
- Our philosophy: **integrate HPCs at no cost for computing operations**. Two main approaches:
  - HepCloud: a single entry point to all US HPCs, for operations effectively it is a single site.
  - **Site extension: preferred solution in EU, success stories in Italy and Germany. Transparent to operations.**

HPCs: an additional opportunity for France to contribute to CMS?



(Some) HPC Usage in the last 6 Months

This does not take into account EU machines, e.g. Horeka, Claix and others, integrated as site extensions!



Beyond x86 CPUs

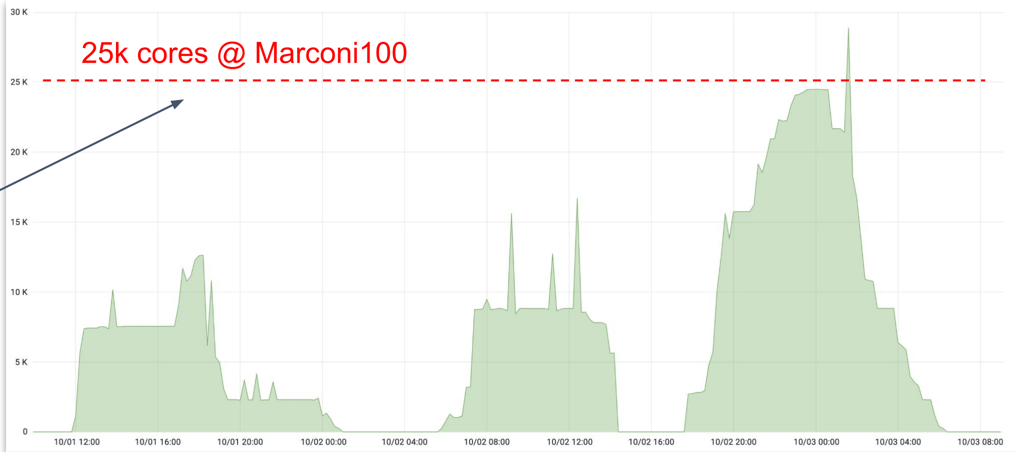
- Working on the integration of **Marconi 100 at INFN CINECA (IBM Power 9 + NVidia V100s)**
- INFN got 3.5 MCoreH in 2021 to:
  - Enable multi-arch support for CMS prod/analysis jobs
  - Perform physics validation on Power 9 for CMS
- Achieved so far:
  - **Full CMS SW stack for Power (since 2016)**
  - Established a complete integration of the CMS Workload Management (both central production and user jobs)
  - Technical test of analyses and release validation workflows successful
- **This first attempt was really promising.**
- Physics validation ongoing
  - Large samples, physics objects and analysis experts evaluating the physics performance of the produced samples with respect to a known reference (same sw run on x86)

MARCONI - 100

Rank	System
11	Marconi-100

Nodes: 980
Processors: 2x16 cores IBM POWER9 AC922 at 3.1 GHz
Accelerators: 4 x NVIDIA Volta V100 GPUs, Nvlink 2.0, 16GB
Cores: 32 cores/node
RAM: 256 GB/node
Peak Performance: ~32 PFlop/s
Quick startup guide

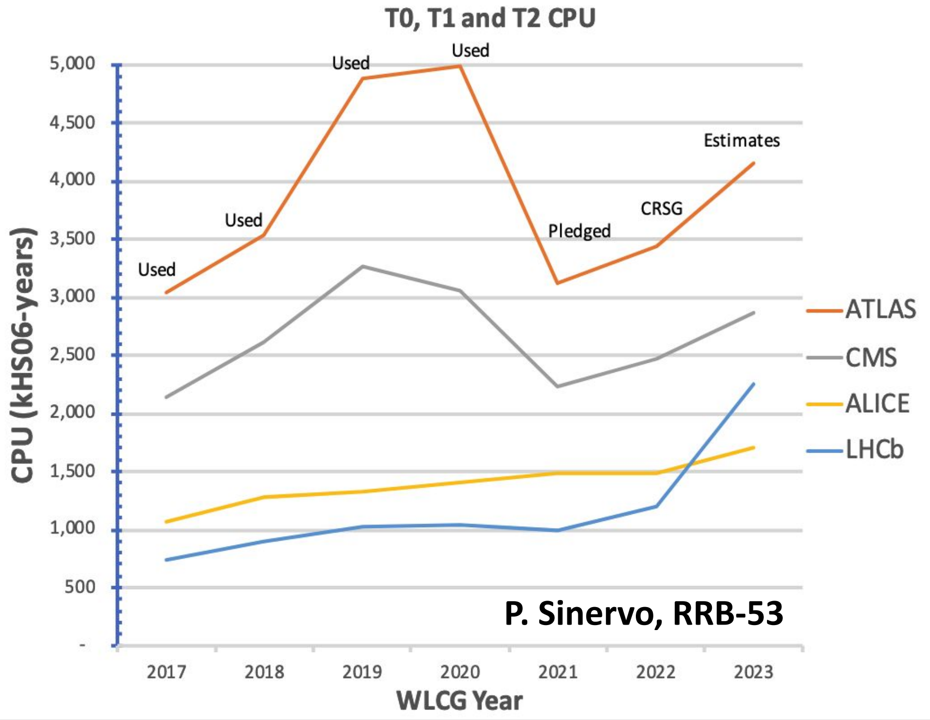
Nov20 top500.org
A "Small Summit"



Preparing for the future acquiring more integration expertise

Computing Resources in 2023: Preliminary request

Information from RRB-53 plenary (public)



CMS		2021 C-RSG recomm.	2021 Pledged	2022 Request	2022 req. /2021 C-RSG	2022 C-RSG recomm.	2023 Preliminary Request	2023 req. /2022 C-RSG
CPU	Tier-0	500	500	540	108%	540	720	133%
	Tier-1	670	764	730	109%	730	800	110%
	Tier-2	1070	1151	1200	112%	1200	1350	113%
	HLT	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Total	**2240**	**2415**	**2470**	**110%**	**2470**	**2870**	**116%**
	Others							
Disk	Tier-0	30.0	30.0	35.0	117%	35.0	45.0	129%
	Tier-1	77.0	76.0	83.0	108%	83.0	98.0	118%
	Tier-2	92.0	96	98.0	107%	98.0	117.0	119%
	Total	**199.0**	**202**	**216.0**	**109%**	**216.0**	**260.0**	**120%**
Tape	Tier-0	120.0	120.0	155.0	129%	155.0	228.0	147%
	Tier-1	230.0	219.0	260.0	113%	260.0	316.0	122%
	Total	**350.0**	**339**	**415.0**	**119%**	**415.0**	**544.0**	**131%**

CMS-1 The C-RSG finds the CMS resource projections justified, as they are based on the currently known 2023 parameter seems justified. The C-RSG expects that these parameters will become more firm by Spring 2022, resulting in more refined resource requests for 2023.

CMS-2 The C-RSG recognizes CMS for its continued effort to support non-X86 architectures, as this is expected to increase the robustness of its software as well as prepares for the future hardware landscape.

CMS-3 The C-RSG supports CMS plans to adapt lossy compression algorithms for heavy-ion data, algorithms that promises a high level of compression without sacrificing the accuracy of the physics data.

R&D work ongoing to evaluate this approach for pp

CMS-4 The C-RSG applaud CMS for its continued effort to decrease the size of the analysis data and especially its plans to have 50% of all analysis to be on nano-AODs by the end of Run 3.

Highlights of Innovations During LS2

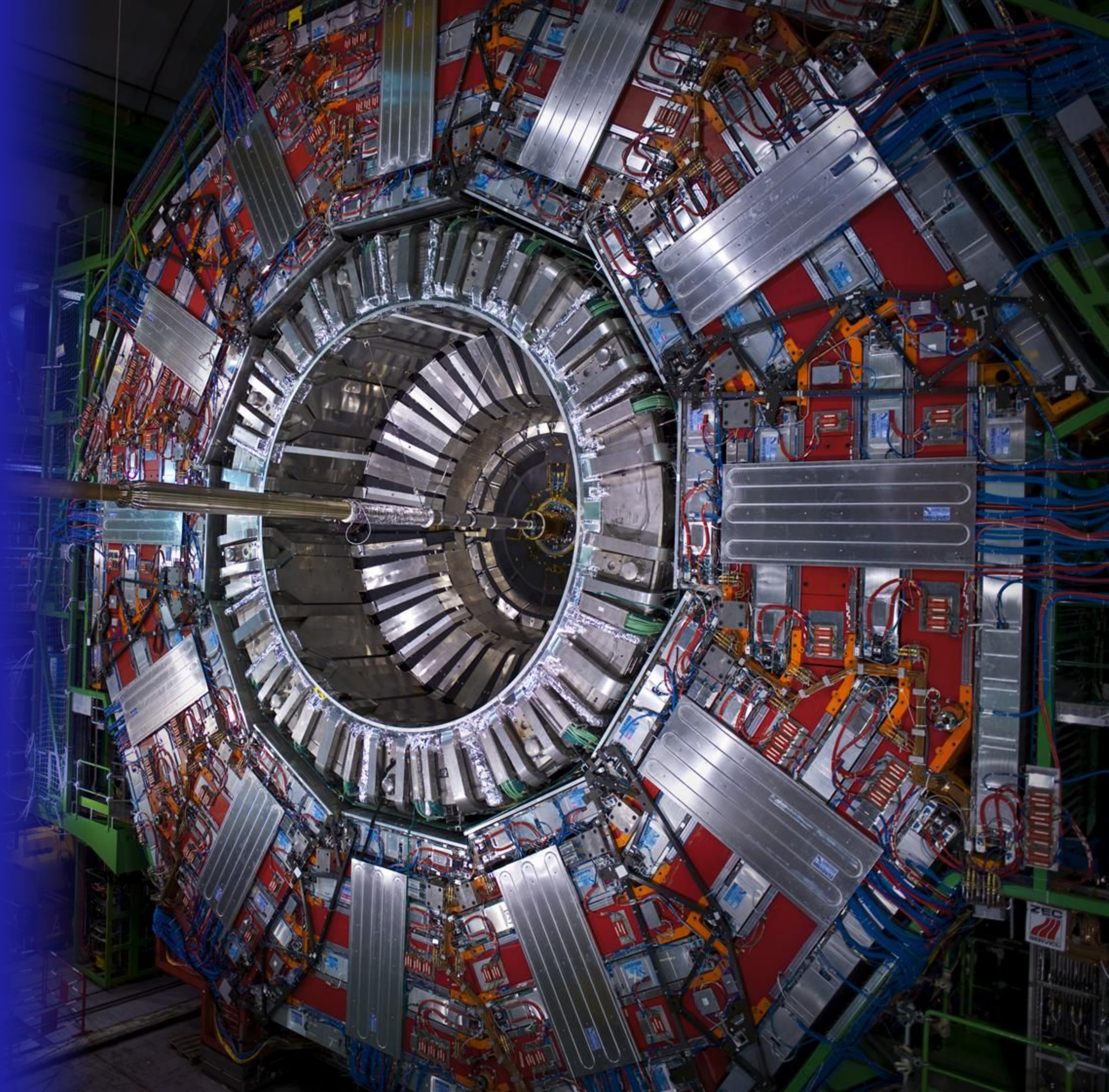


Image: © CERN

Innovations during LS2

- **The CMS software stack and comp. tools were adequate for needs in Run 2**, and then some
- **No real hint that Run 3 would pose irresolvable problems either**; **but, since Phase-2 could be a different story**, CMS planned to try and test any disruptive technology already in Run 3
- Example innovations that happened during LS2
  - Offload to accelerators
  - CRIC: Grid resource catalogue (click here to see the public resource requests of experiments!)
  - DD4Hep: Geometry description tool
  - Rucio: data management tool
  - WebDav protocol for data transfers
  - Migration of internal CMS portfolio of services to k8s
  - NanoAOD
  - EOS advanced features
- **Common solutions with other experiments are a way to mitigate the support cost**

Accelerators Support in the CMSSW Framework

- CMSSW software framework: the orchestrator of CMS data processing units (“Modules”)
  - A powerful engine that makes data processing very efficient
- **CMSSW supports multithreaded execution**
  - All data processing steps: gen, sim, digi/mix, reco, Mini/Nano production, HLT, Tier-0) are multithreaded since Run 2 start
- **During LS2, support for “external work” was added: a generic mechanism to offload calculations**
  - Keep CPUs busy during offload if needed

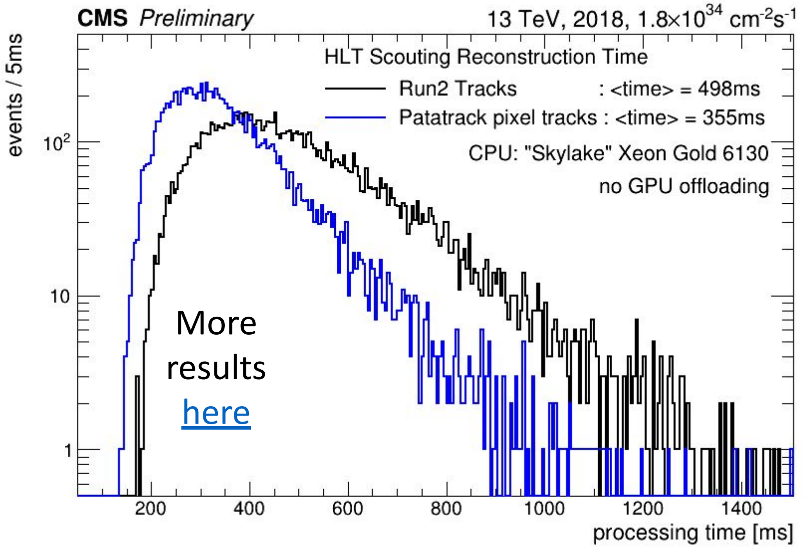
Usage of accelerators today:

- Non-ML: offload of CUDA code on GPU on the same wn (Solution chosen for the Run 3 HLT)
- ML:
  - Offload through Tensor Flow or ONNX on GPU on the same wn
  - Offload through SONIC (Services for Optimized Network Inference on Coprocessors) on accelerators mounted on a different node. A very promising R&D, potentially giving even greater flexibility to CMS computing model

Offloading non-ML code on GPUs

- **GPUs will be in production at the HLT in 2022 already: ~30% of the runtime of the HLT sequence offloaded to GPUs**
- **We are actively working to expand the usage of GPUs for offline computing already for Run 3**: not only for the lower cost per unity capacity, but also for the flexibility of our computing model
  - Allocations on HPCs may become possible only if GPUs are used
- **Exciting times for Physicists-developers**
  - Lots of opportunities in CMS for working on GPU related projects!
- **Early to make any statement about needs for pledged GPUs**

In production in 2022



CRIC

- Common catalogue for all LHC experiments (click on the image to access CRIC)

CRIC
Computing Resource Information Catalog

In production

Site Topology ▾ Pledges ▾ Accounting ▾ Downtime ▾ Admin ▾ Logs ▾ Help ▾ Login

Export | Columns 8/10 | Filter | Reload | Pledge comparison (VO) View | Pledge comparison (Federation) View

VO Requirement list

Show All entries

Tier	Pledge Type	Year	ALICE	ATLAS	CMS	LHCb	Sum
filter by Tier	filter by Pledge Type	2021	filter by ALICE	filter by ATLAS	filter by CMS	filter by LHCb	filter by Sum
0	Disk	2021	45500	29000	30000	18800	123300
0	Tape	2021	86000	95000	120000	43800	344800
0	CPU	2021	471000	525000	500000	175000	1671000
1	Tape	2021	57000	235000	230000	75900	597900
1	Disk	2021	53300	105000	77000	37600	272900
1	CPU	2021	498000	1170000	670000	574000	2912000
2	Disk	2021	44800	130000	92000	7300	274100
2	CPU	2021	515000	1430000	1070000	321000	3336000
Tier	Pledge Type	Year	ALICE	ATLAS	CMS	LHCb	Sum

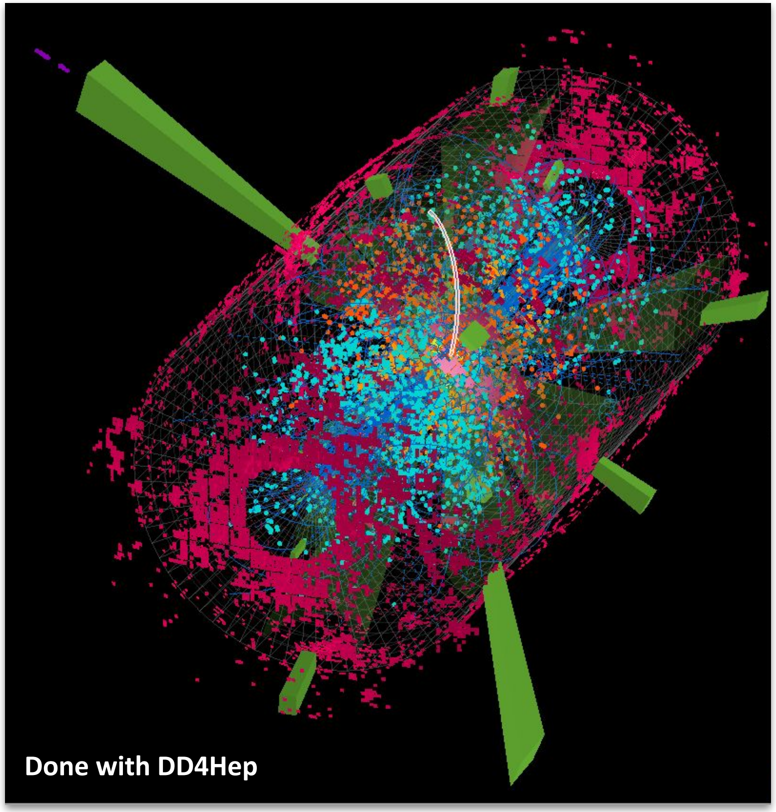
Showing 1 to 8 of 40 entries

Previous 1 Next

Geometry Description: DD4HEP

- Until Run 2 geometry description done with a in-house tool: DDD
- **LS2: transition to the community tool DD4HEP**
  - Used, among the others, by LHCb and FCC studies
  - Natively integrated with ROOT and Geant 4
- Some advantages for CMS:
  - **A more sustainable software stack for Run 3 and beyond**
  - More **modern, thread friendly geometry description**
  - An opportunity to **review our geometry**, converging on an improved description!
  - Stringent **battery of unit tests** developed
- In the process, **contributions and improvements delivered to DD4Hep and ROOT**
  - Not only benefits for CMS, but also for common software
- Run 3 in production, Phase-2 almost done, Run 2 and then Run 1 to migrate next

In production



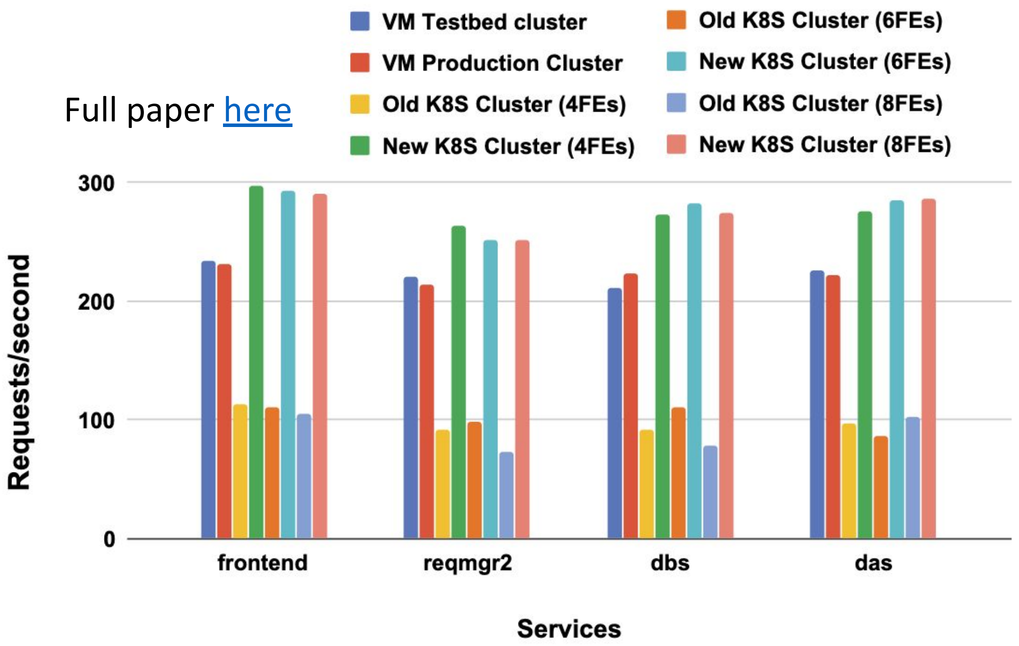
K8s Backend for CMS Internal Services

- CMSWEB: internal CMS portfolio of services
  - E.g. Workload/data mgt, dataset catalogue
- Previously running on VMs
  - Only relatively flexible, obliged us to release monolithic new versions of the portfolio, load balancing difficulties
- Moved to CERN IT's K8s service
  - Excellent support and collaboration with IT
  - Better performance of services
  - Deploy new versions of individual components
  - Better usage of resources
  - Operations easier than before

K8s for internal services: better usage of human and computing resources

In production

Full paper here



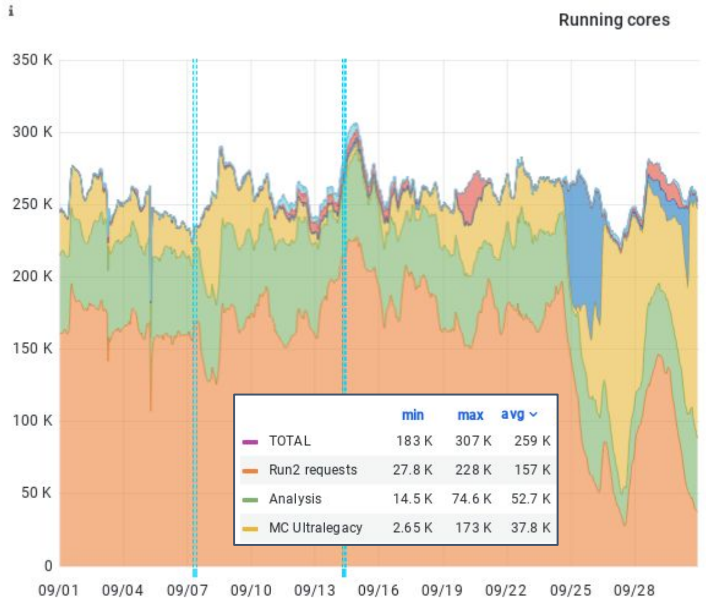
RUCIO: The Data Management tool of CMS

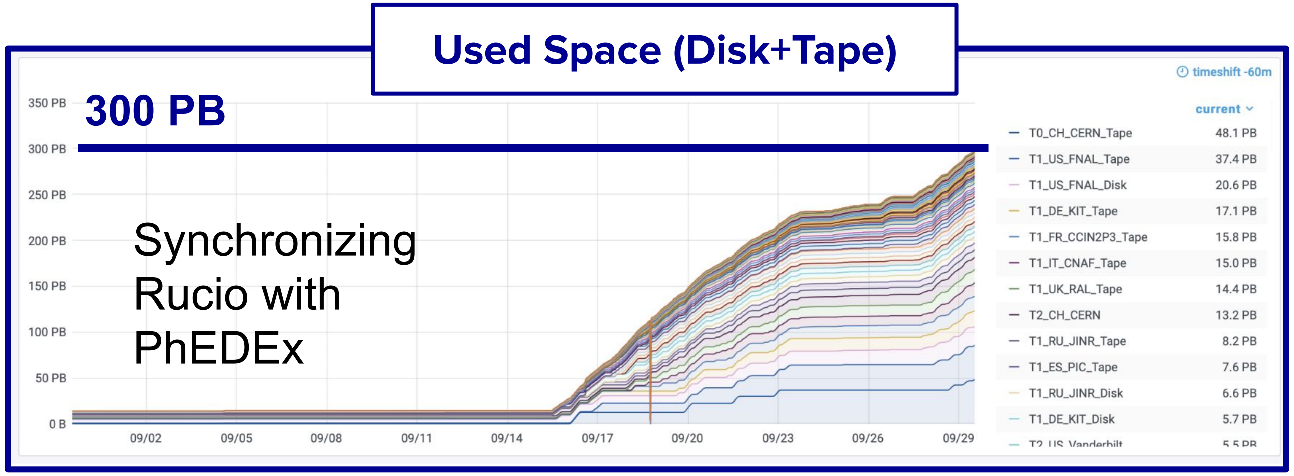
- **CMS transitioned from PhEDEx + Dynamo to Rucio for data management**
  - PhEDEx (data movement) and Dynamo (Dynamic Disk Manager) were custom CMS services
  - Rucio: community supported, shared with ATLAS and other experiments
- Transition: coordination of many moving parts e.g. workload management
- **Could not afford downtime or interruption in any CMS computing service**
  - Data management, production, or analysis
  - Preparatory work ongoing for months
- Extremely smooth transition
  - RUCIO deployed on Kubernetes from day zero

A well planned transition, no disruption during the switch, more sustainable sw stack

In production

No degradation at all of CPU usage during the switch!





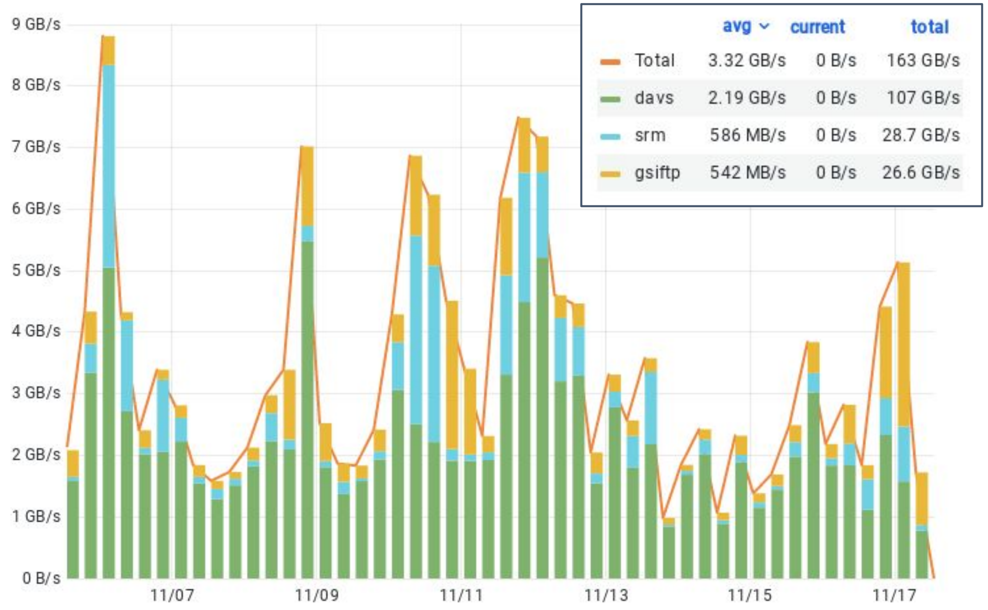
WebDav Protocol

- Migrate away from GridFTP to WebDav protocol for our transfers
- **A milestone of the roadmap which will lead us to the usage of token based authentication**
- Migration started in collaboration with our sites during Q1 2021
- Tier-1's and Tier-2's basically migrated
  - Working now on Tier-3's
- French Tier-1, Tier-2's all migrated: Thanks!
  - T3_FR_IPNL working on it
- 65% of transfer volume between sites through WebDav
  - It was 20% in August

In production

Migration to WebDav: a successful joint effort of CMS and our sites

FTS Transfer throughput excluding same site transfers

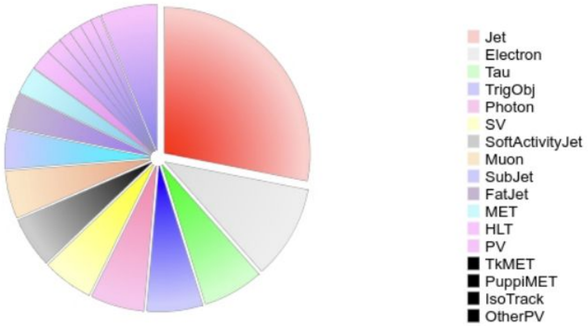


NanoAOD

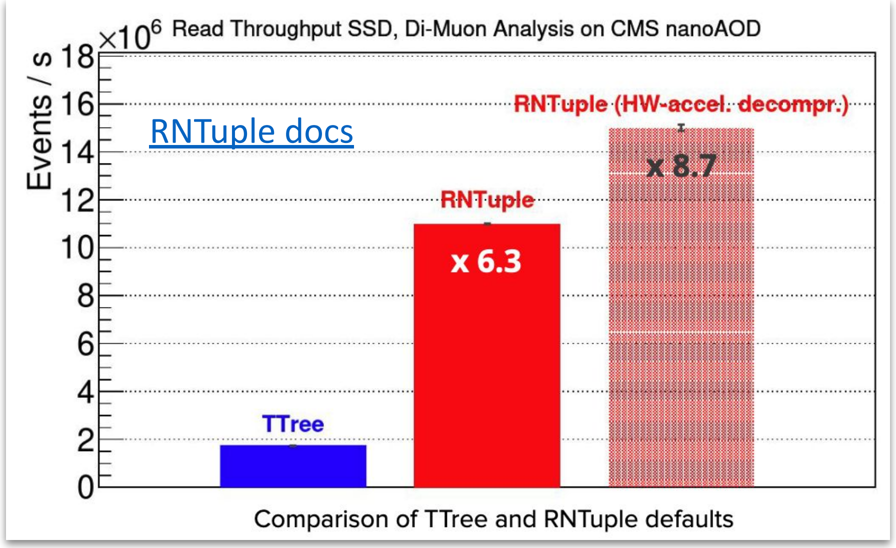
- **CMS created two small analysis formats MiniAOD (~50 kB/evt) and NanoAOD (1/2 kB/evt)**
- MiniAOD: used throughout Run 2 by the vast majority of analyses
  - Adopted by HI for Run 3!
- **NanoAOD: adopted by 30% of the analyses**
  - Target 50% by the end of Run 3
  - Official CMS Ntuples: columns of fundamental types and arrays thereof
- Will be produced at the Tier-0 for prompt reconstruction
  - Looking at data will be fast and easy
- **Crucial ingredient to face the HL-LHC storage challenge**

In production

(8.379 Mb, 10000 events, 0.86 kb/event)



collection	items/evt	kb/evt	b/item
Jet	5.46	0.164	30.8
Electron	0.66	0.061	94.9
Tau	0.64	0.039	63.0
TrigObj	2.93	0.036	12.7
Photon	0.85	0.035	42.0
SV	1.09	0.033	30.7
SoftActivityJet	5.82	0.033	5.8
Muon	0.48	0.031	66.3
SubJet	1.08	0.026	24.3
FatJet	0.60	0.022	38.0
MET	1.00	0.017	17.9
HLT	1.00	0.013	13.6



J. Blomer

NanoAOD: a powerful way to meet the HL-LHC storage challenge, in production today

CMS can produce natively NanoAOD in RNTuple format, the successor of TTree in ROOT.

EOS Advanced Features

- EOS: storage technology at CERN with a veritable world wide community behind
- Created for HEP (even if used also elsewhere): **actively testing new features for the benefit of CMS at the Tier-0 in collaboration with CERN IT. Examples: Erasure Coding, prioritised writings**

Erasure Coding (EC)

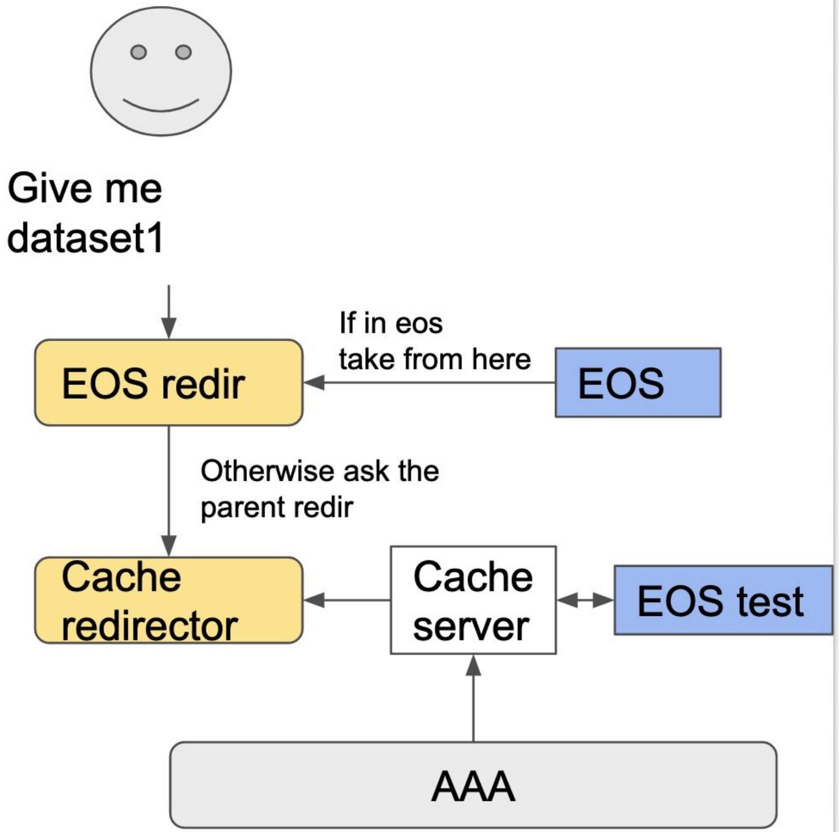
- More logical space for the same raw disk: divide data in blocks and add parity blocks to ensure recovery (from replica 2x to replica 1.2-1.4x)
- Providing new EOS space to analysis groups through nodes with EC enabled: promising results so far

IO priorities

- Give more priority to some writes wrt to others
- Very useful when analysis/grid and data taking activities are ongoing at the Tier-0
- Planning a test with Tier-0 dominantly saturating the EOS bandwidth

Data Caches: More Flexibility to the Computing Model

- Cache: non-custodial storage space used in several ways in CMS
  - An additional QoS on top disk and tape, *de facto*
- SOCAL Cache serving UCSD and Caltech Tier-2s, used in production (since 2019)
  - 200 Km, 100 gbps, below 3ms
- CNAF Cache to sustain IO from Marconi HPC @ CINECA to CNAF Tier-1 storage, used at the time in production
- Experimentation at CERN with a cache dedicated to mini/nano not stored on EOS
- Potential way to serve storageless sites
- Useful building block for future analysis facilities



Conclusions

Conclusions

- Run 3 is the priority for CMS O&C: ambitious physics program to enable
- Several innovations were put in production during LS2, targeting Run 3 and beyond
  - In the area of HPCs, data management, non-x86 architectures, caches...
  - Common software solutions adopted: reduce the cost of our sw toolset
- **A successful Run 3 can only happen thanks to the support of our sites**
  - **France: substantial and reliable Tier-1 and Tier-2 resources provided to CMS - fundamental to enable the physics programme of CMS**
- Beyond resource provision, many opportunities to contribute to O&C activities:
  - Innovative algorithms, HPC integration, data management...
- **Exciting times for curious physicist/developers/integrators/computing experts: a single person can make the difference!**