Edition dedicated to the memory of our dear Vicente Hernández

Evolving WLCG towards HL-LHC - J. Flix

IBERGRID 2023

Benasque (Spain)

25-29 September 2023

PIC port d'informació científica

Who am I?

PIC
port d'informació
científica

First of all… **Apologies!** It's been a while since I don't show up in IBERGRID Conf!

Exp. HEP Physics PhD (doing **HEP Computing** since 2006)

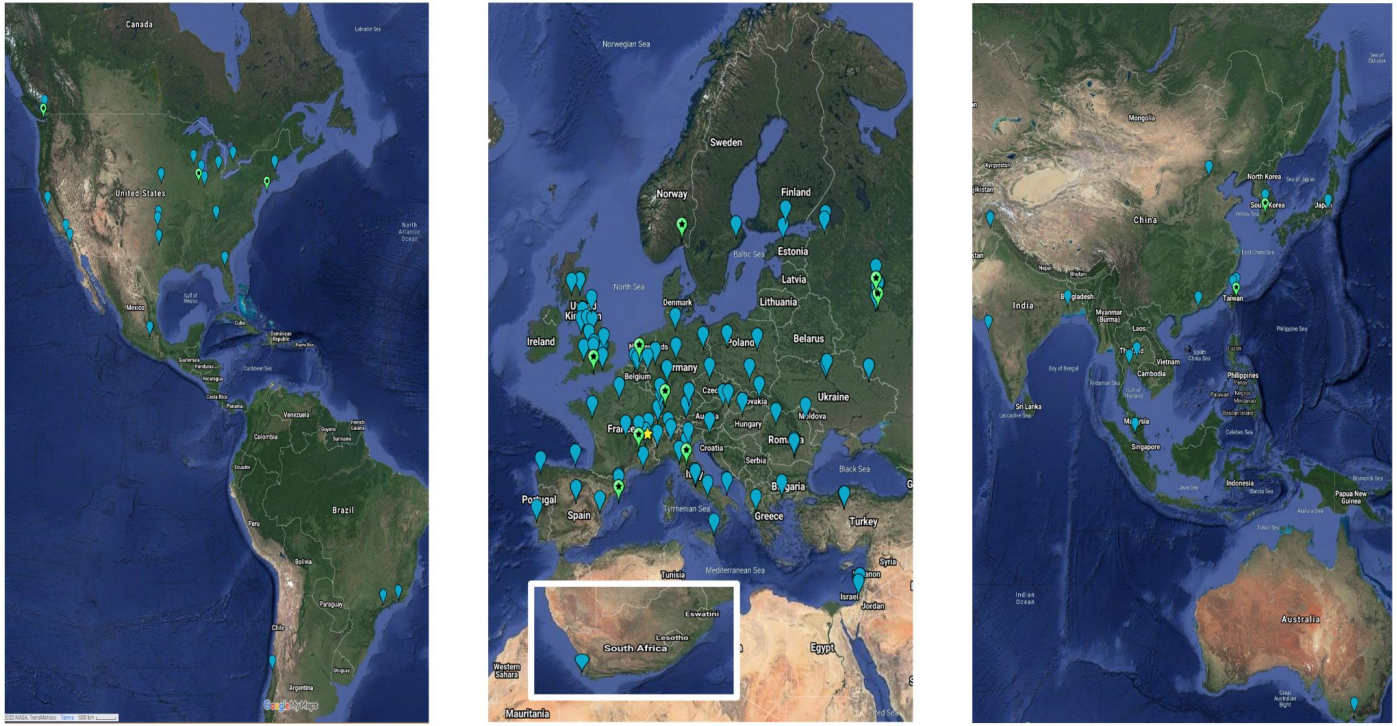
I am the manager (coordinator P.I.) of the **WLCG Spanish Tier-1** deployed in PIC (Barcelona), that supports ATLAS, CMS and LHCb experiments, +25% of **Spanish ATLAS Tier-2** (PIC) and +75% of **Spanish CMS Tier-2** (Madrid) → This represents ~65% of LHC computing resources deployed in Spain

Member of the **CMS Collaboration**, with computing mgt responsibilities

Chair of the **WLCG Grid Deployment Board (GDB)**

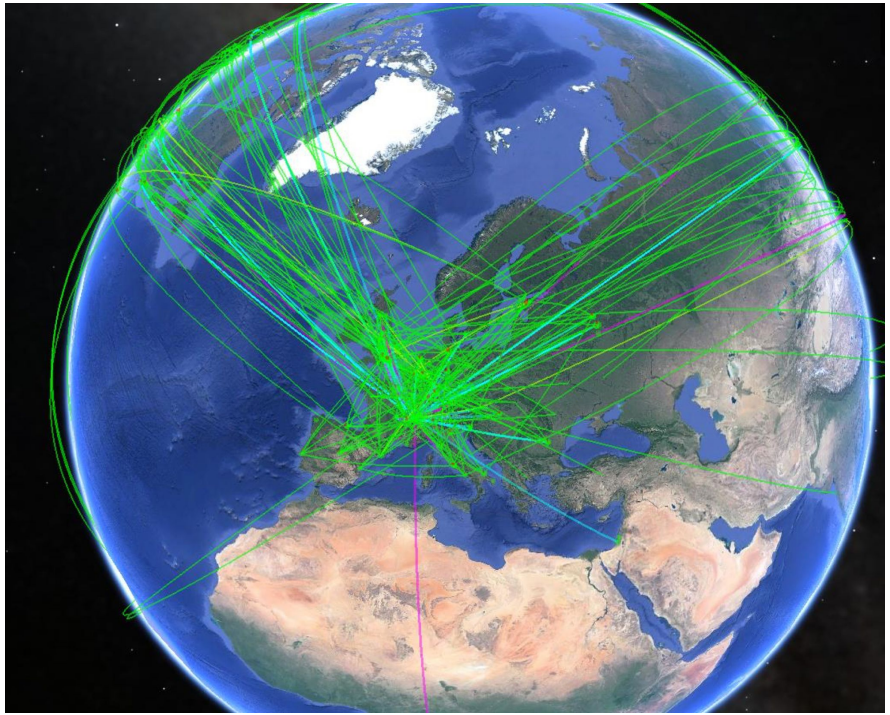
WLCG today

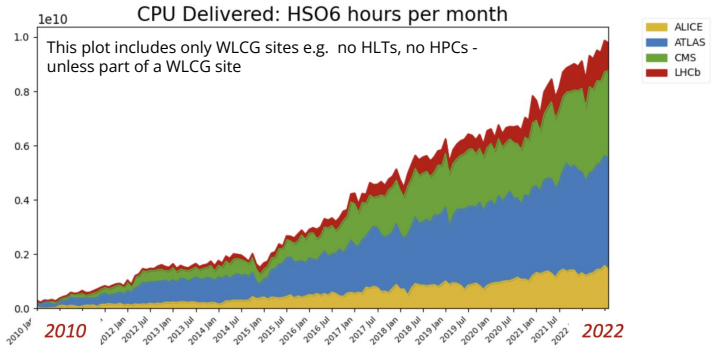
WLCG was commissioned/deployed over **15 years ago** - worldwide scientific computing effort!
It extends across **165 sites** in **42 countries**, bound by **66 MoUs**

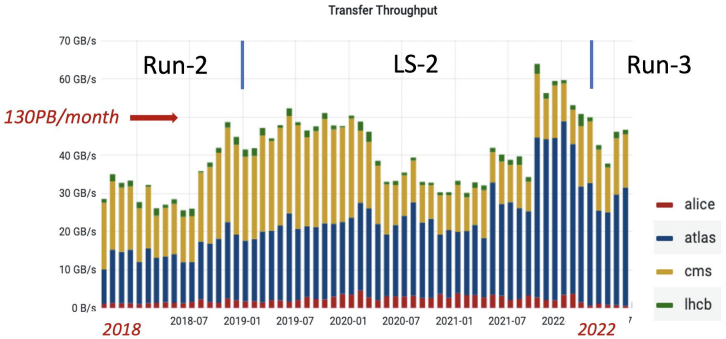


WLCG today

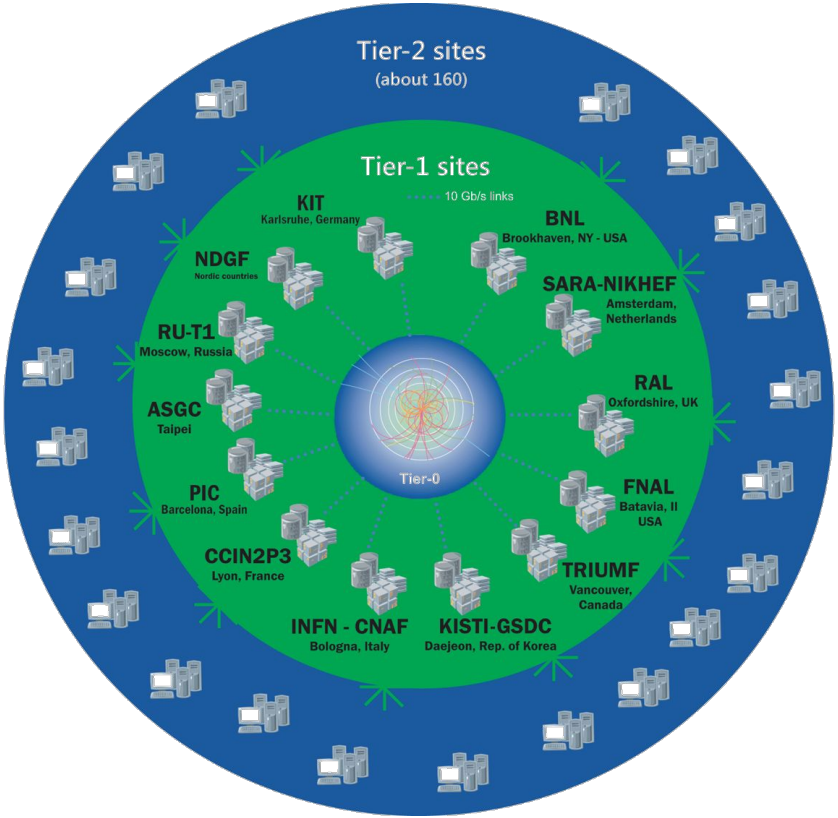
Worldwide **data storage and processing 24/7**. It is at the scale of **1+ million cores** and **1+ Exabytes of storage**







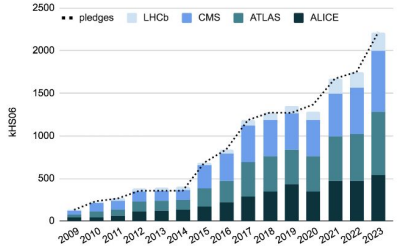
WLCG tiered structure



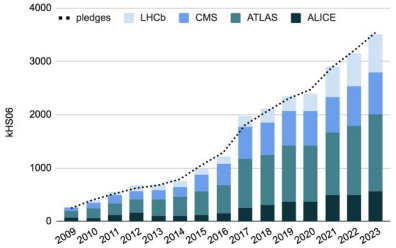
WLCG pledges evolution (MoU)

PIC
port d'informació
científica

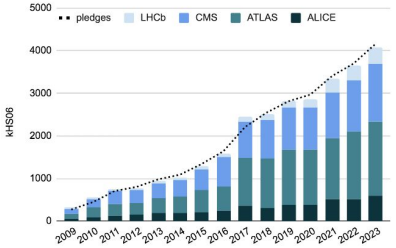
CPU: Tier-0



CPU: Tier-1



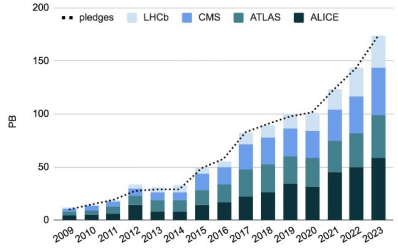
CPU: Tier-2



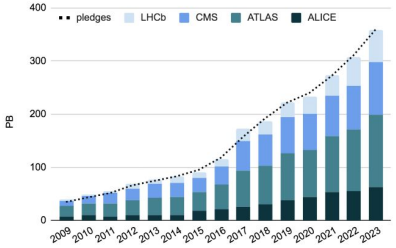
~9.9M HS06 (~750k CPU cores)

+ additional ~40% opportunistic

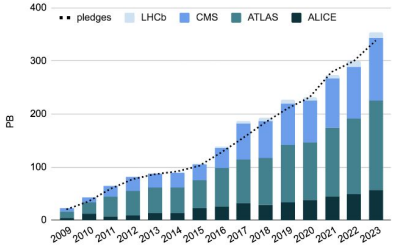
Disk: Tier-0



Disk: Tier-1

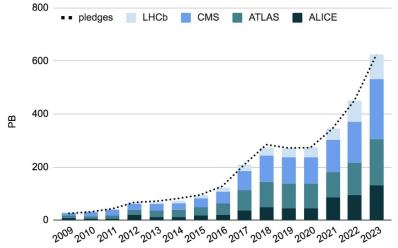


Disk: Tier-2

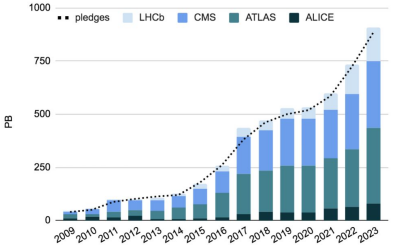


~870 PB in Disk

Tape: Tier-0



Tape: Tier-1



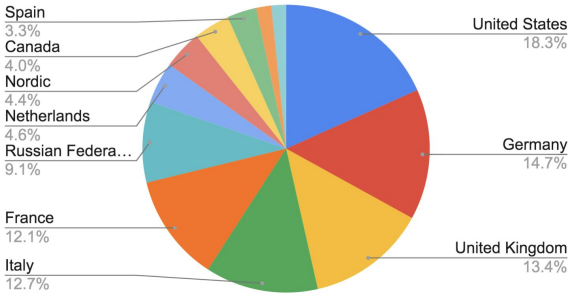
~1500 PB in Tape

Collaborative effort from 42 countries!

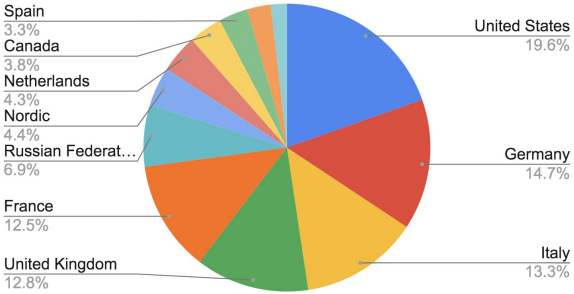
WLCG 2023 pledges by Country

PIC
port d'informació
científica

CPU Tier-1



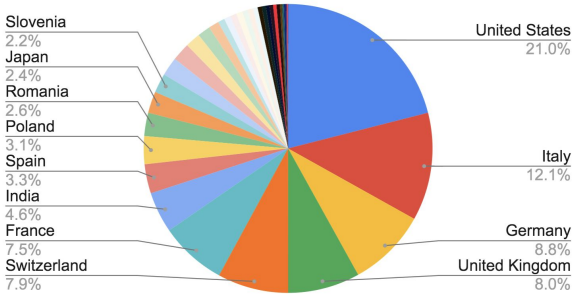
Disk Tier-1



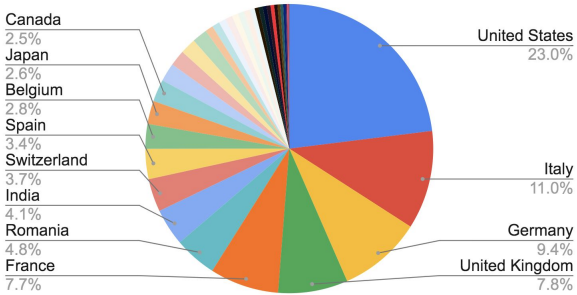
Spain currently pledges **4% of the WLCG resources** at Tier-1 and Tier-2, for the ATLAS, CMS, and LHCb experiments

Portugal currently pledges ~**0.5% of the WLCG resources** at Tier-2, for the ATLAS and CMS experiments

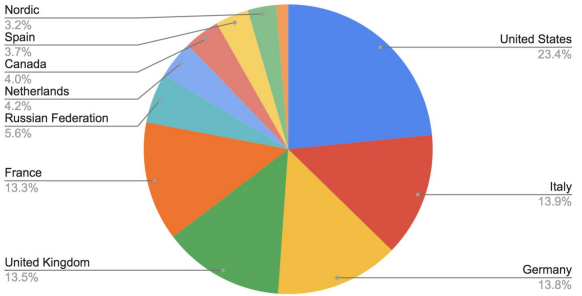
CPU Tier-2



Disk Tier-2

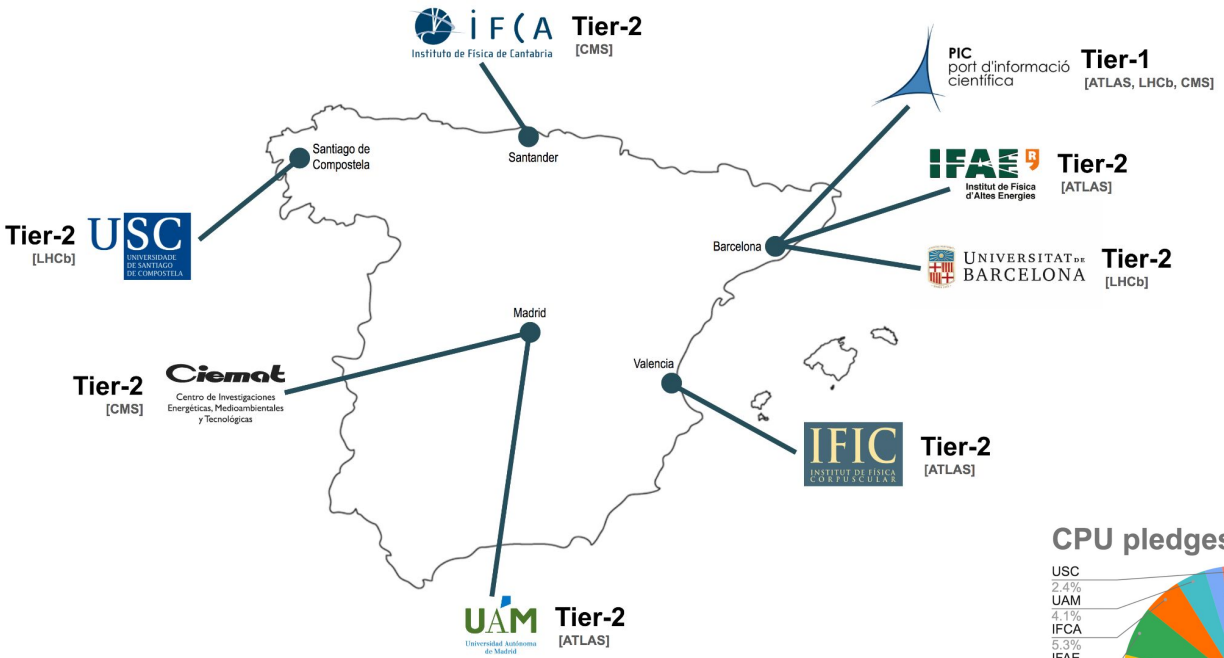


Tape Tier-1

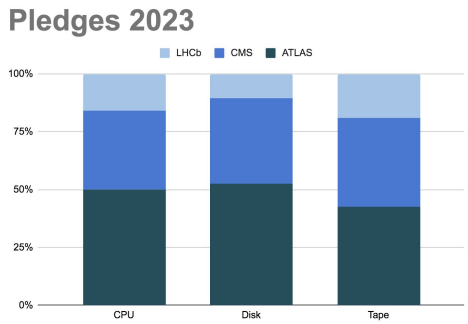


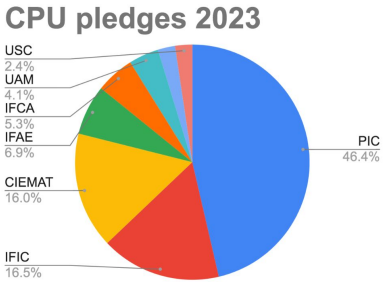
e.g. WLCG in Spain

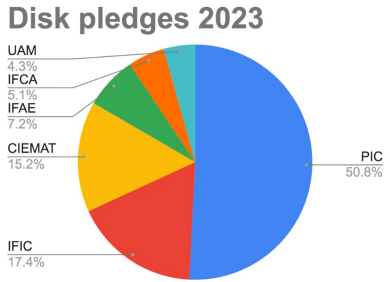
Apologies Portuguese colleagues!



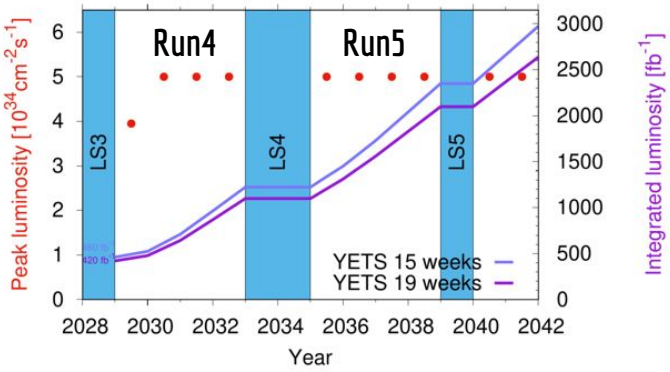
~20,000 CPU cores
23 PB disk
33 PB tape

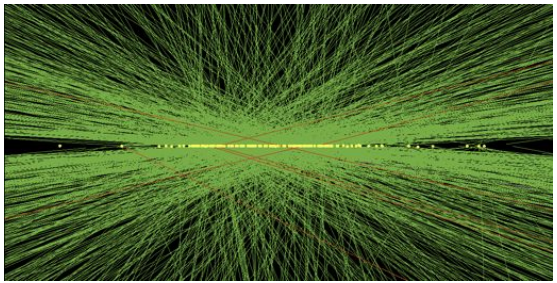






The HL-LHC computing challenge





Run-3/4/5, pile-up estimated to average 54/140/200

The challenge is to handle **increasing**

- Data **volumes**
- Data **complexity**

In the **context** of

- Constrained funding
- Sustainability concerns
- Power shortages?
- Global politics
- Security and trust
- Increasingly heterogeneous resources

Driven by ATLAS and CMS, though ALICE and LHCb will also be a challenge, particularly in Run-5

Towards HL-LHC computing

2017: HEP Software Foundation Community Whitepaper: a bottom-up exercise. Identify the areas of work to address the HEP challenges of the 2020s

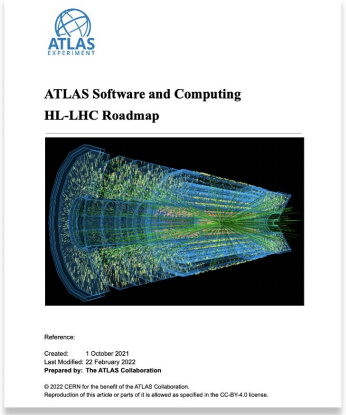
2018: The first WLCG strategy toward HL-LHC document: a top-down high-level prioritization of the whitepaper, for the LHC needs

The LHCC review series of HL-LHC computing were initiated in **2019**: a multistep process tracking the progress towards HL-LHC:

2020: review of ATLAS and CMS plans, Data Management (DOMA), offline software, the WLCG collaboration and infrastructure. Documents

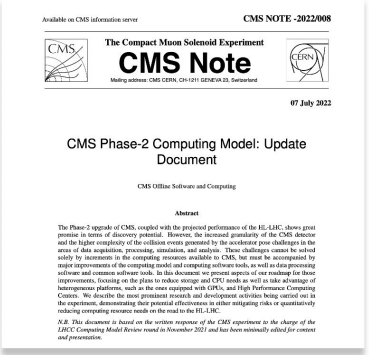
2021: update from ATLAS and CMS, common software activities (generators, simulation, foundation software, analysis, DOMA). Report

Two more reviews to come…


ATLAS Software and Computing
HL-LHC Roadmap

Reference:

Created: 1 October 2021
Last Modified: 22 February 2022
Prepared by: The ATLAS Collaboration


Available on CMS information server

CMS NOTE -2022/008

The Compact Muon Solenoid Experiment

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

07 July 2022

CMS Phase-2 Computing Model: Update Document

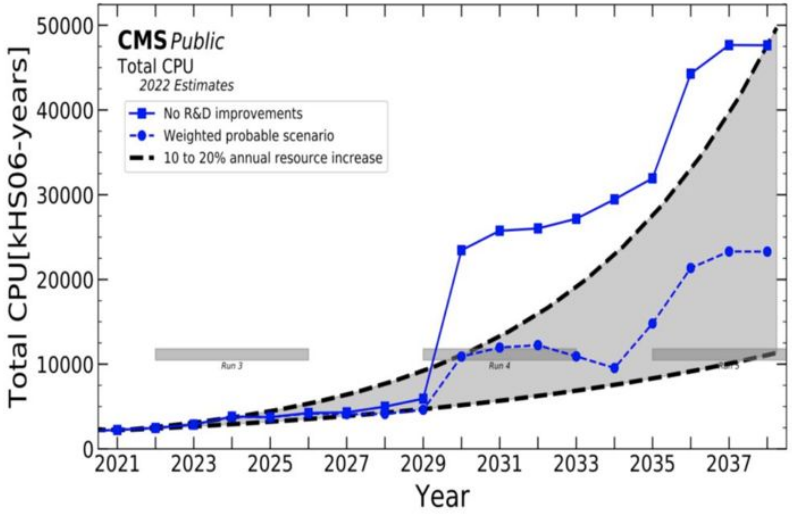
CMS Offline Software and Computing

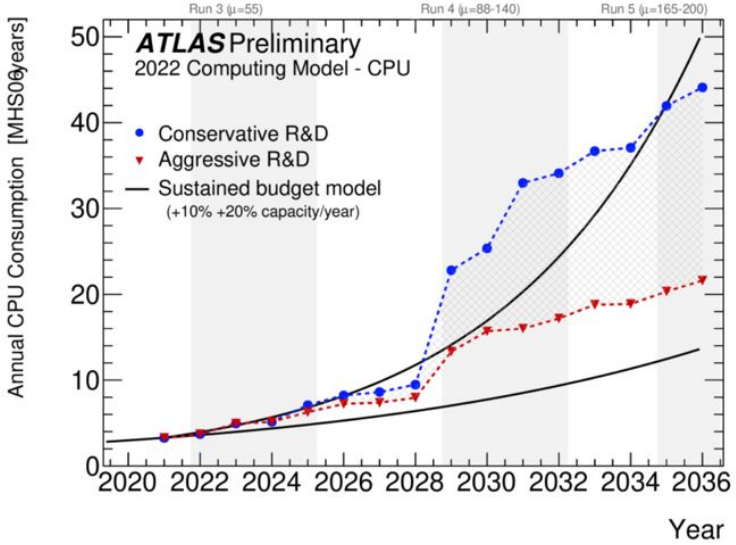
Abstract

The Phase-2 upgrade of CMS, coupled with the projected performance of the HL-LHC, shows great promise in terms of discovery potential. However, the increased granularity of the CMS detector and the higher complexity of the collision events generated by the accelerator pose challenges in the areas of data acquisition, processing, simulation, and analysis. These challenges cannot be solved solely by increments in the computing resources available to CMS, but must be accompanied by major improvements of the computing model and computing software tools, as well as data-processing software and common software tools. In this document we present aspects of our roadmap for those improvements, focusing on the plans to reduce storage and CPU needs as well as take advantage of heterogeneous platforms, such as the ones equipped with GPUs, and High Performance Computing Centers. We describe the most prominent research and development activities being carried out in the experiment, demonstrating their potential effectiveness in either mitigating risks or quantitatively reducing computing resource needs on the road to the HL-LHC.

N.B. This document is based on the written response of the CMS experiment to the charge of the LHCC Computing Model Review round in November 2021 and has been minimally edited for content and presentation.

Future requirements on CPU



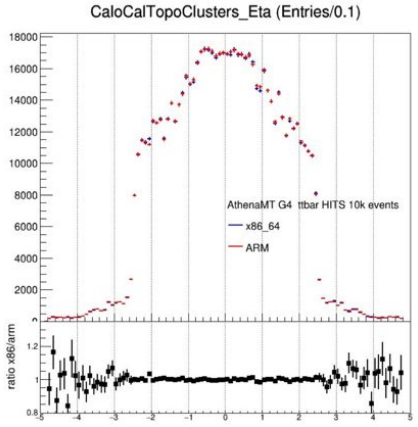


Historically, an **additional ~40% of CPU capacity** has been made available opportunistically to the Grid (as compared to pledges). *This may be impacted in future by the increase of cost of energy*

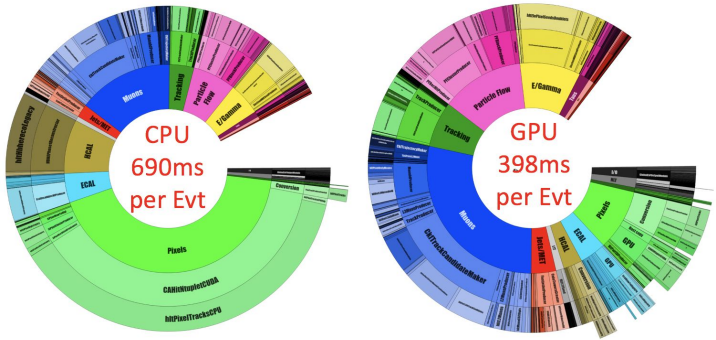
Additional capacity provided from **HPC**, and **Cloud** facilities, which speed up the physics output. This has required developments to use **different systems and architectures**

Investing in **R&D activities** and **Software Engineering**, is essential… Significant uncertainties [e.g. HW costs?]

Use of new architectures



ATLAS Simulation Physics Validation on ARM



CMS Online Reconstruction with GPUs

For much of the last two decades, HEP computing was dominated by **x86 architecture** (INTEL and AMD)

Alternative architectures can offer **advantages of speed or power consumption** and are often used in HPC machines

Moving code from x86 processors to RISC processors such as ARM and Power, **requires physics-validation**

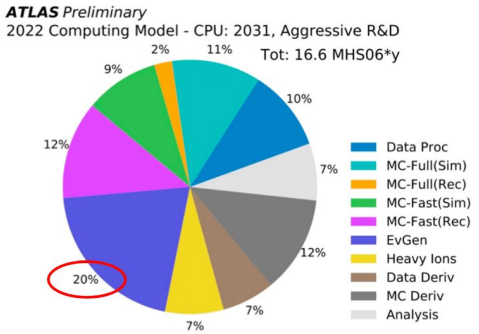
Moving code to accelerators such as GPUs requires completely **restructuring/re-writing** the algorithms to take advantage of the parallelism offered

Such heterogeneous architectures are already playing a critical role in Run 3 in most online systems of the LHC experiments

This is a **major effort** requiring expertise from the experiments' and new programming skills, but it is very important to rejuvenate our code and make it more sustainable

Speed and Power

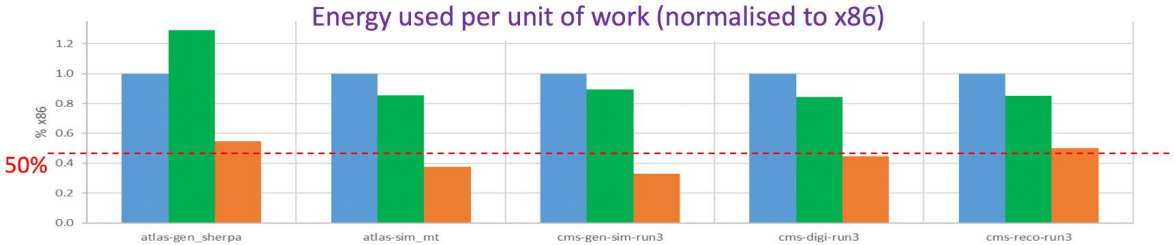
GPUs process some types of workload much faster, reducing the amount of CPU required



		madevent		
CUDA grid size		8192		
$gg \to t\bar{t}ggg$	MEs precision	$t_{TOT} = t_{Mad} + t_{MEs}$ [sec]	N_{events}/t_{TOT} [events/sec]	N_{events}/t_{MEs} [MEs/sec]
Fortran	double	1228.2 = 5.0 + 1223.2	7.34E1 (=1.0)	7.37E1 (=1.0)
CUDA	double	19.6 = 7.4 + 12.1	4.61E3 (x63)	7.44E3 (x100)
CUDA	float	11.7 = 6.2 + 5.4	7.73E3 (x105)	1.66E4 (x224)
CUDA	mixed	16.5 = 7.0 + 9.6	5.45E3 (x74)	9.43E3 (x128)

NVidia V100, Cuda 11.7, gcc 11.2

ARM (used in mobiles) offer the potential of reducing the amount of energy needed



x86
Dual x86
ARM

Reducing carbon footprint!

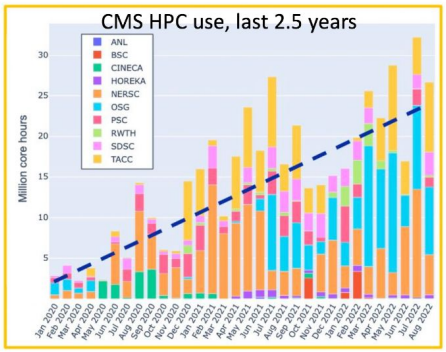
Use of HPC centers

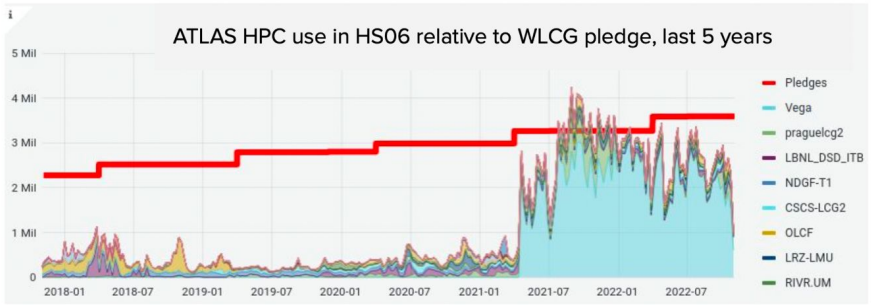
Each HPC brings its own set of **challenges and constraints**, due to the diversity in access and usage policies, the services available, and the system architectures. Various barriers have been reduced over the years, **some still exist…**

Even when these are addressed, HPCs live for a relatively short time (3-5 years) compared to WLCG sites, and then the work might need to be done again. Allocations are as well short (bureaucratic work)

WLCG computing is done economically on the sort of hardware used on the Grid. Some **national computing priorities** may intend to **complement HTC pledges with HPC resources** at some point soon

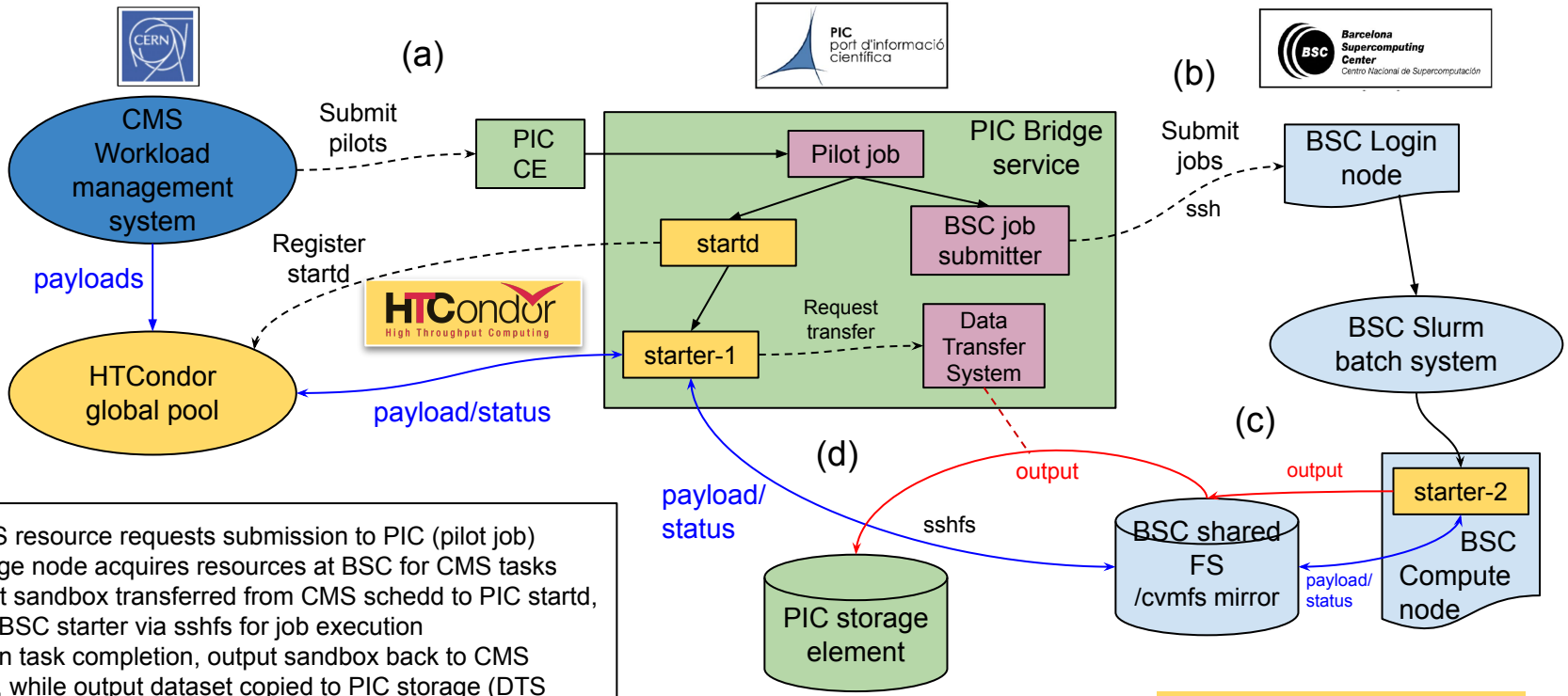
HPC will be a **significant part of the future WLCG** in some countries





Use of the BSC by CMS through PIC

Spoiler of tomorrow's talk!



(a) CMS resource requests submission to PIC (pilot job)
(b) Bridge node acquires resources at BSC for CMS tasks
(c) Input sandbox transferred from CMS schedd to PIC startd, then to BSC starter via sshfs for job execution
(d) Upon task completion, output sandbox back to CMS schedd, while output dataset copied to PIC storage (DTS acting as third party copy manager)

→ **At CHEP2021 proceedings (link)**
→ **At ISGC 2022 (link)**
→ **At CHEP2023 (link)**

PIC and HTCondor team collaboration to **use a shared FS as control path for HTCondor**

Use of Cloud resources

Experiments have been using **commercial cloud resources** through special projects since many years, sponsored by big providers and EU projects

So far, resources on Grid premises have been less expensive… Is this expected to change? **Risks around costs**, TCO studies ongoing (e.g. ATLAS Google Cloud grant)

Computing: cloud procurements can make a lot of sense (**economic sense to burst-to-cloud**)

Storage: custodial data must remain at Tier-0 and Tier-1 sites. Data replicas in the cloud could help speed up processing and analyses (if data egress charges are **affordable**)

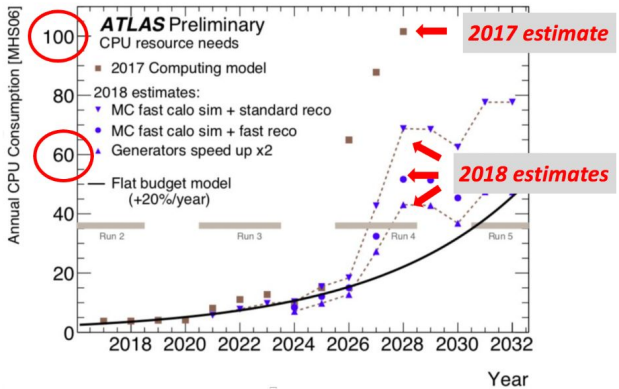
Integrating such resources has **security implications**: their CAs are not in IGTF. The **WLCG Resource Trust Evolution TF** is looking into sustainable recipes

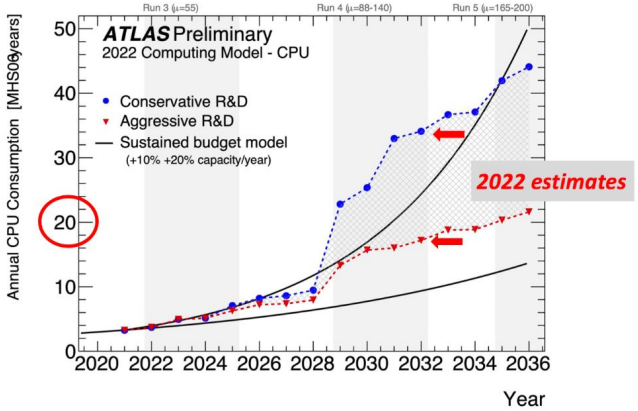
Software

Investment in software (at all levels) is an **important aspect**

Software development is **essential to reduce hardware requirements**:

- Enables portability to new architectures (GPU, ARM, etc)
- Optimises algorithms to improve efficiency
- Allows development of common tools and libraries
- Facilitates modernisation (for security reasons; or because things are deprecated; or because there are performance/operational advantages)





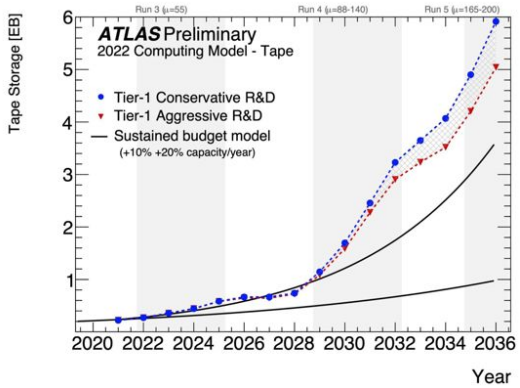
Future requirements on Tape

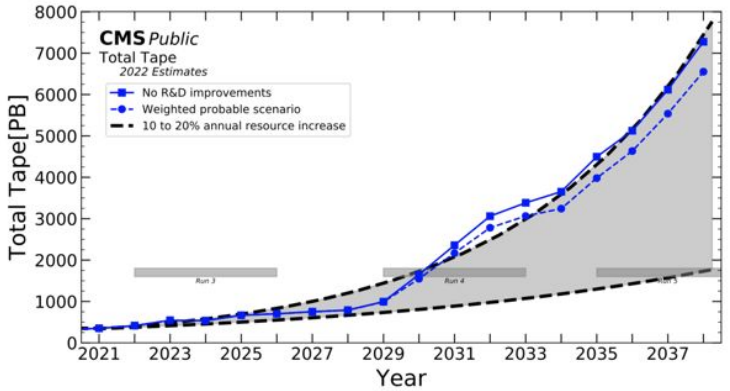
For HL-LHC, tape storage usage will be dominated by **RAW and AODs** (data and MC)

But **tape is not just archival**: data is regularly recalled and processed, mitigating the higher costs of disk

Data Carousels provide organised, scheduled, and sequential recall of data for processing, optimising use of bandwidth from tape

Costs are hard to predict (limited number of players; steps in technology) - use disk as a tape-backend? (KISTI use case)





Future requirements on Disk

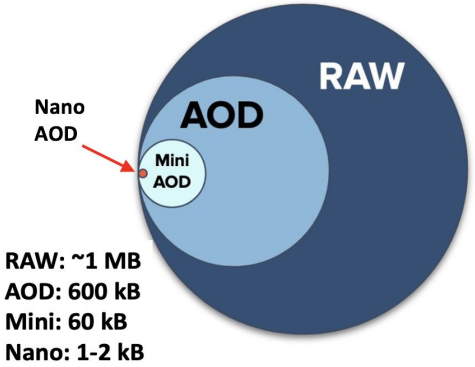
Disk storage used for **“reconstructed” data that is being analysed.** Various different formats and versions of both real and simulated data

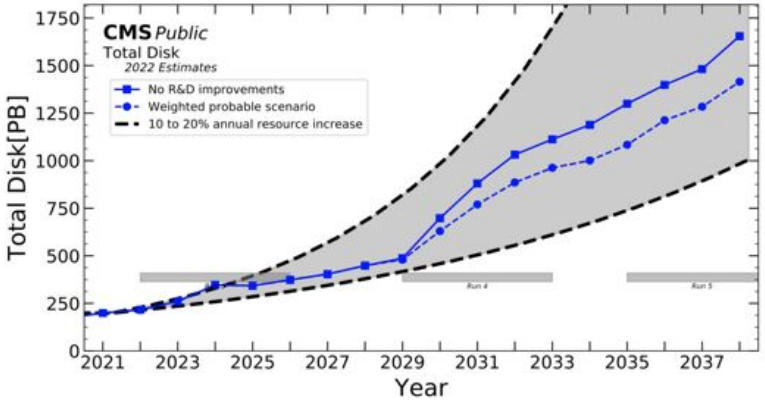
Sometimes **staged** from tape, e.g. Data Carousels

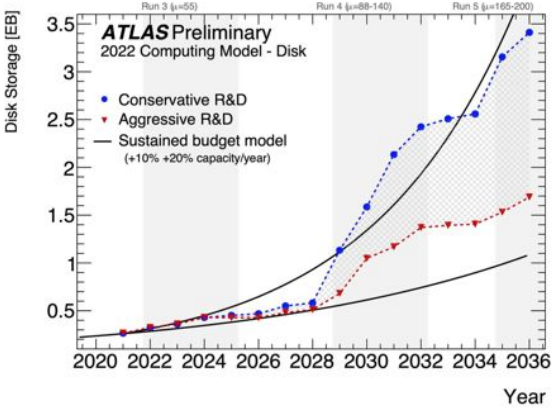
Big progress has been made in developing and using **smaller data formats**

Envisioning the use of **data caches** close to sites focused only on CPU resources

CMS Data Formats







Analysis Facilities

HL-LHC computing models foresee **Analysis Facilities** as a new type of resources

Some WLCG facilities may want to offer such resources instead of traditional set-ups, or even in addition to those resources

- Proper recognition needed in pledges and accounting

Input data must not too quickly become the bottleneck

- **Fast local storage and network are needed** (cf. HPC burst buffers)
- Possibly in the form of **caches**, i.e. local data losses can be tolerated

Profiting again from **alternative architectures**: ARM, GPUs

Users may access such services through **Jupyter notebooks**... auto-scaling from the laptop to the grid!

Networks

Networks are the backbone of WLCG, which supports the core functions of data acquisition archival and processing → the 4th resource!

Dual-stack deployment of storage has reached ~94% at the T2 sites

IPv4 will need to remain supported at least a few more years for legacy workflows that cannot handle IPv6

IPv6 networks allow new features to be taken advantage of

- Dynamic provisioning through Software Defined Networks
  - Allowing better use of available capacity instead of relying on overprovisioning
- Packet marking and flow labeling
  - Allowing activities to be monitored separately, both between and within VOs

New L3 VPNs like LHCONE are expected to be set up for **partner projects**

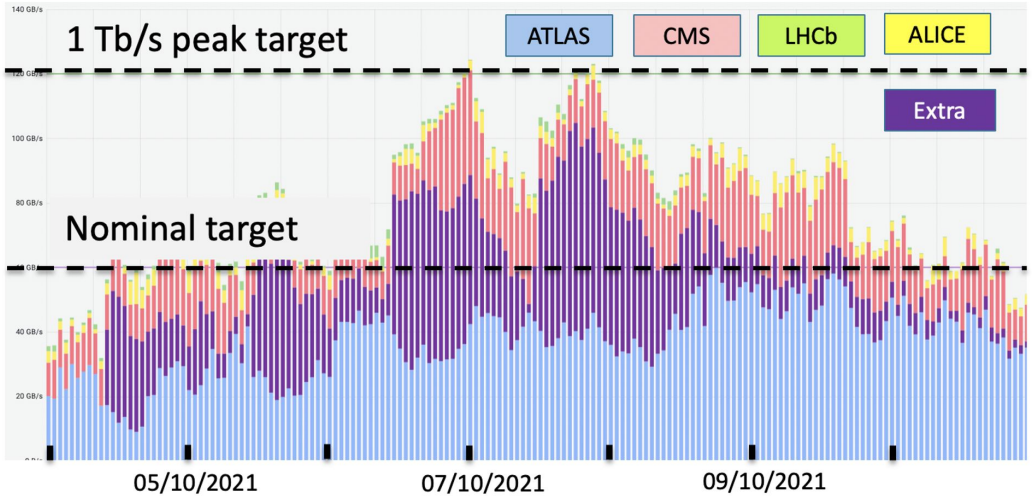
Networks also need to be interfaced with **HPC centers and Cloud providers**

Data Challenges

A series of four **data/network challenges before HL-LHC** to demonstrate full capability

The first “10%” challenge took place in October 2021, in preparation for Run-3. The 2nd challenge (**Feb. 2024**) will be to at least double this. Preparation workshop in Nov. 2023

Challenges are end-to-end tests including storage, protocols, networks, and data management services (e.g. Rucio, FTS)



Run-3 targets met:

- Nominal transfer rate sustained
- Peak transfer rate reached

Authentication and authorization

Since a few years, WLCG has started **transitioning from X509 user certificates and VOMS proxies towards using WLCG tokens** instead

- Arguably the biggest change ever! Will also allow to phase out the remaining Globus dependencies

A timeline with tentative milestones was published in August 2022

Computing: in production for OSG & WLCG, work in progress for EGI Check-in tokens

Data: discussions ongoing between stakeholders

- Experiments, CTA, dCache, DIRAC, Echo, EOS, FTS, IAM, Rucio, StoRM, XRootD, …
- Rates, lifetimes, scopes, …
- Plans for Data Challenge 2024 (12th-23rd February) are shaping up!

End game (~2026): users no longer need X509 certificates!

Monitoring

Though many systems are used by experiments and sites, **complete monitoring overviews at the WLCG level are tricky and sometimes incomplete**

To **improve** we need to **measure many metrics**...

For example, **data traffic monitoring** **improvements** (in particular for the **XRootD** protocol) concern a number of stakeholders

- Experiments
- SE and FTS developers
- CERN MONIT and IAM teams
- WLCG, OSG and EGI Operations

MONIT foreseen to be used more for consolidation. **CRIC** used as WLCG information system

New systems keep **emerging** and gaining popularity in many places: Prometheus, ...

Collaboration with other HEP experiments

PIC
port d'informació
científica

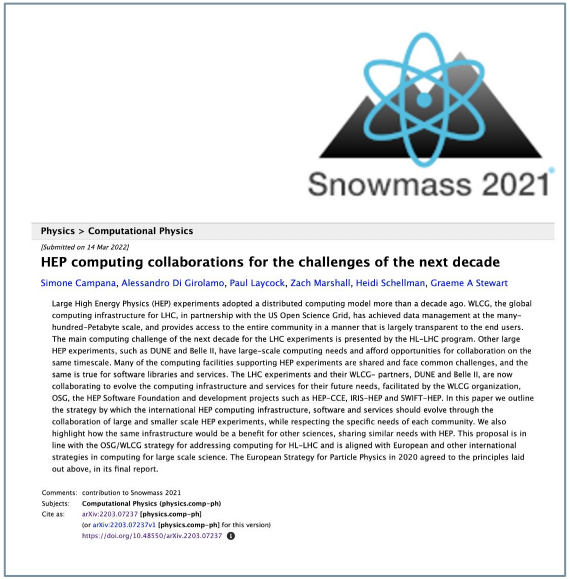
WLCG presented a joint paper with **DUNE** and **Belle-2** to the Snowmass 2021 process, which detailed the strategic directions needed to address the computing challenges of the experiments over the next decade

It complements the WLCG contribution to the European Strategy for Particle Physics in 2019

Three strategic areas:

- Strengthen the backbone of core services and policies
- Evolution of the infrastructure to integrate modern technologies and facilities
- **Broadening the scope of the WLCG collaboration to create partnership with other HEP experiments**

Today DUNE,Belle-2 and JUNO are WLCG “observers” and share many services with WLCG (including some LHCOPN/LHCONE network infrastructure)



Collaboration with other sciences

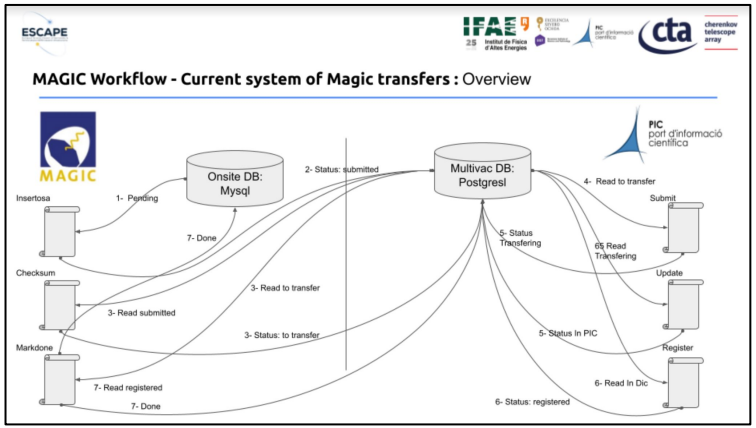
The **ESCAPE project** (SKA, CTA, KM3Net, EST, ELT, HL-LHC, FAIR, CERN, ESO, JIVE) implemented a prototype data infrastructure across Europe based on many of the WLCG building blocks and on top of many WLCG facilities

e.g. one of the demonstrated applications was the delivery of large datasets produced by Gamma ray telescopes, adopting Rucio to stream files from the telescopes to a Data Lake for permanent storage and access

Used **PIC** and **MAGIC** and **LST** for the demonstrator:

* Satisfactory deployment of **DIRAC+Rucio** in the **PIC Kubernetes** cluster and integrated in **CTA-Grid**

As other sciences grow, WLCG will increasingly need to **cohabit** (networks, sites) and **collaborate** (tools, services, security)… **this will take time and effort**



Stones in the road…

Security incidents may affect part of our infrastructure at times

- Security experts from our infrastructure provider projects and sites collaborate to minimize the fallout → for example through the **SAFER trust group**
- WLCG sites can collaborate on incident response and prevention through the **Security Operations Center WG**
- Sites can contribute through passive DNS SOC services

WLCG may be affected by t**echnological disruptions**, e.g. ORACLE EOL for tape technology… or **positively affected by technological advances!**

WLCG may be affected by **upstream licensing or support changes**

- Case in point: the new RHEL source code policy
- With consequences for AlmaLinux and Rocky Linux

Conclusion and Outlook

PIC
port d'informació
científica

Since many years, the **Worldwide LHC Computing Grid** has successfully provided the distributed computing infrastructure for the CERN LHC experiments

During that time, the WLCG has seen **evolution in technologies** as well as **growth**, to deal with ever **increasing data rates**

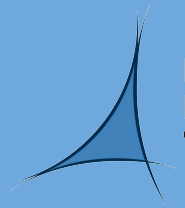
Those trends need to be made to continue, to allow the WLCG to take on **High-Luminosity LHC** data volumes as of 2029

The **future of WLCG is about**: scale, functionality, sustainability, and efficiency

Improvements across a wide range of **services and software** have been described, some of which already bring benefits as of today

Collaboration with other experiments, **projects**, and **sciences** with similar computing challenges. **Cooperate to success!!** much of the infrastructure is shared

Thank you!

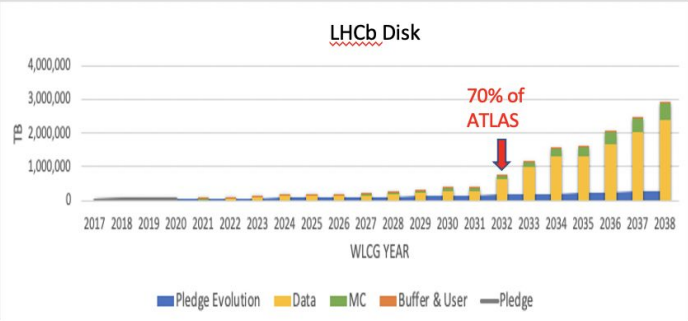


PIC
port d'informació
científica

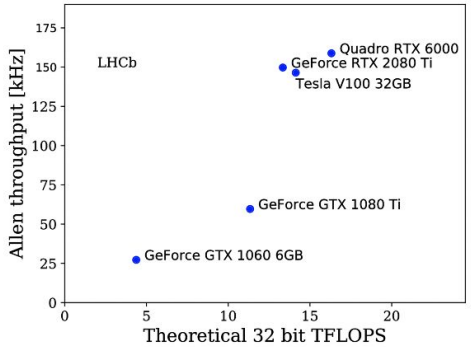
And thanks to D. Britton and M. Litmaath for inputs!

LHCb: Computing Challenges and Heterogeneity

- LHCb already completed phase-1 upgrade and trigger must process 32 Tb/s in software in Run3 (same as ATLAS/CMS High Level Triggers in HL-LHC, but 6+ years earlier!)
- This will increase by another factor 7.5x for further upgrades for Run-5.
- One of the biggest data challenge in HEP in next decades!



- LHCb moving towards heterogeneous architectures: GPUs to complement CPUs
  - For partial detector reconstruction in Run 3
  - For full detector reconstruction in Run 4?
  - For fast detector simulation in Run 4 or 5?
- Development of "Allen" computing framework as a natively heterogeneous partner for GAUDI. Enable efficient exploitation of available heterogeneous architectures across the world, not only online.

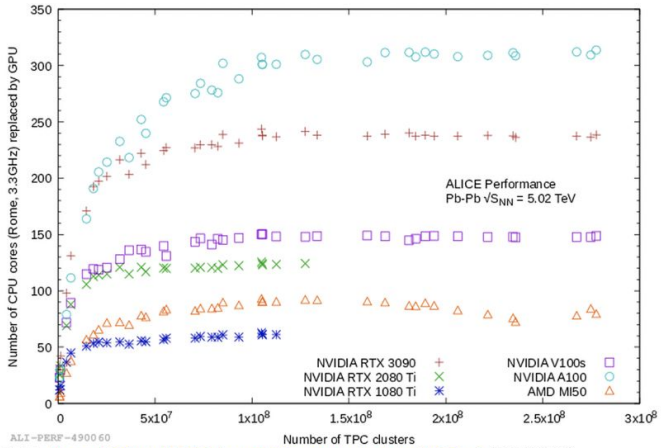


ALICE: Computing Challenges and Heterogeneity

- ALICE will record 100x more collisions in Runs 3&4, compared to Runs 1&2 (Pb-Pb collisions up to 50kHz the 100x higher data rate enabled by novel detector technology).
- Substantial inter bunch crossing pileup; (untriggered) continuous readout.
- Speed up of synchronous processing (real time) from GPU usage + from algorithmic improvements + tuning of CPUs.

- GPU based solution gives a large saving on hardware and operating costs.

- Porting of asynchronous (offline) reconstruction code to GPUs well advanced thanks to common online-offline (O2) framework



D.Rohr, Usage of GPUs in ALICE Online and Offline processing during LHC Run 3 (vCHEP 2021)