20/05/2024

Sabine Crépé-Renaudin, LPSC Grenoble CNRS/IN2P3

Computing for HL-LHC

Corfu 2024 Workshop on Future Accelerators

Note

Computing covers a wide range of very interesting subjects

- Preparation for HL-LHC concerns all of the computing components
  - ➜ Impossible to cover everything

Objective of this talk

- Start with a brief status of LHC computing to understand from where we start
- Give the main challenges imposed by HL-LHC data taking conditions
- Describe the overall context since it has a large impact on our developments
- Describe the main activities to take up the HL-LHC challenges

➜ so that you have a good overview of the main hot topics

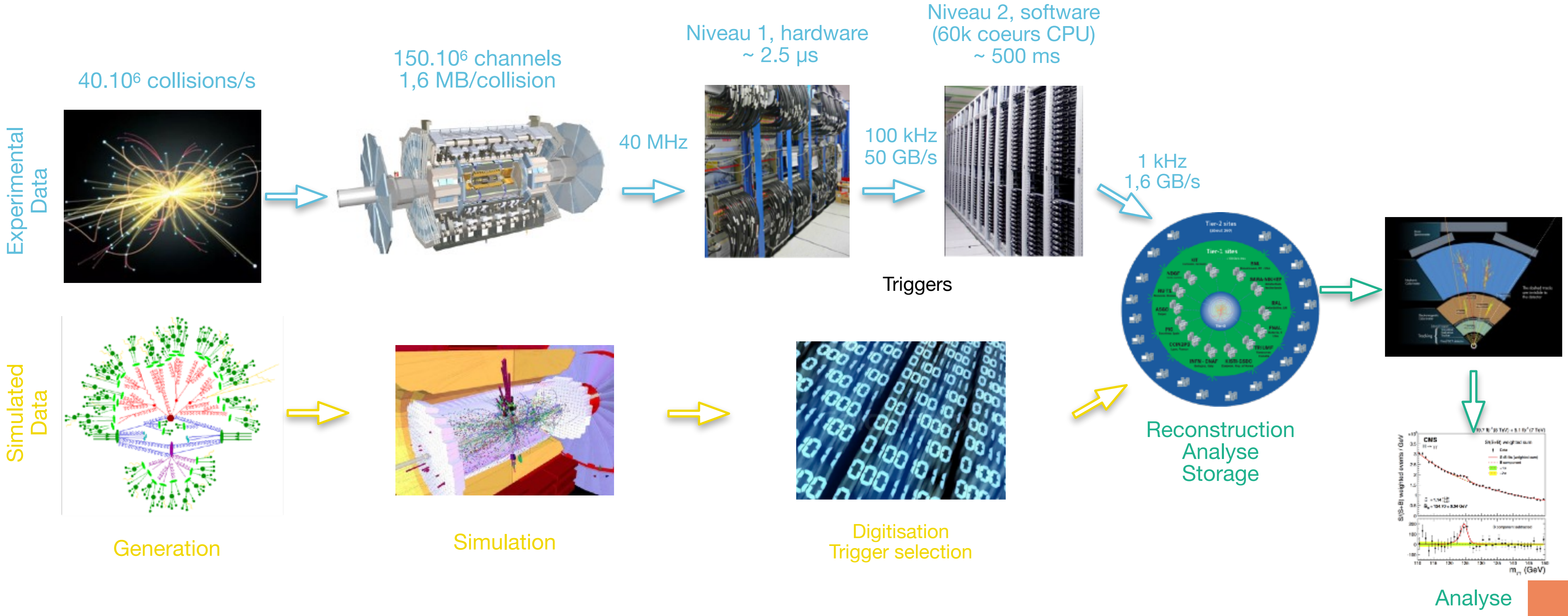
➜ and hopefully that this triggers your interest to look more closely to these developments and all the ones that I could not cover

Main Sources of this talk in

- CHEP 2023 conference: https://indico.cern.ch/event/1369601
- WLCG/HSF 2024 Workshop: https://indico.cern.ch/event/1369601
- HEPiX Spring 2024 Workshop: https://indico.cern.ch/event/1377701
- JENA 2023 computing workshop: https://agenda.infn.it/event/34738

From where we start : key ingredients

Typical Computing in HEP



Experimental Data

40.10^6 collisions/s

150.10^6 channels
1,6 MB/collision

40 MHz

Niveau 1, hardware
~ 2.5 µs

100 kHz
50 GB/s

Niveau 2, software
(60k coeurs CPU)
~ 500 ms

Triggers

1 kHz
1,6 GB/s

Simulated Data

Generation

Simulation

Digitisation
Trigger selection

Reconstruction
Analyse
Storage

Analyse

Typical for LHC run 1-2

Main components

Hardware

- Trigger
- WLCG grid
  - grid sites : CPU, storage disks and tapes
  - network

Middleware, interware and databases

- Data management and distribution: FTS, xrootd, webdav, DDM, Rucio, DIRAC, …
- Computing task management : Panda, DIRAC, …
- Software distribution: CVMFS
- Databases: conditions database, detectors, softwares, datasets, sites etc
- Monitoring : sites, computing tasks, storages, transfers, networks
- Communication, tickets systems: GGUS, JIRA

Softwares

- Trigger
- MC generation
- Simulation
- Reconstruction
- Analyse
- Monitoring

➜ 10s millions of code lines

Skilled people

- for all the components
- technicians, engineers and physicists

dCache

ROOT Data Analysis Framework

XRootD

CTA (Tape Archive)

EOS (File System)

CVMFS (Software Distribution)

CRIC (Information System)

FTS (File Transfer System)

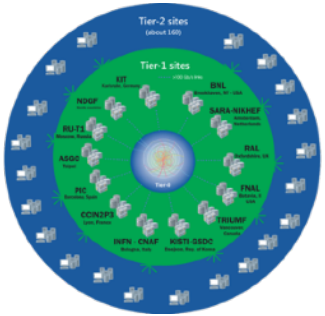
RUCIO

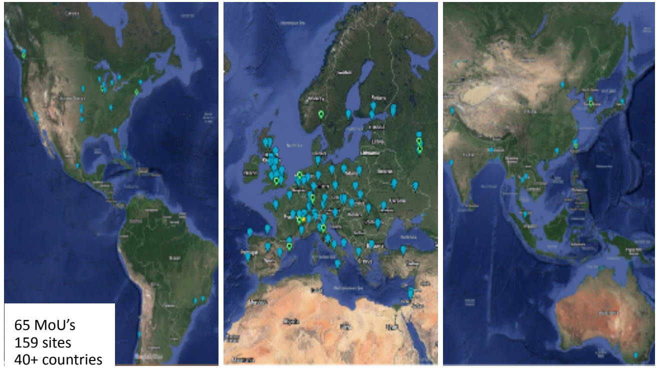
DIRAC (Workload Management)

IAM

…

WLCG

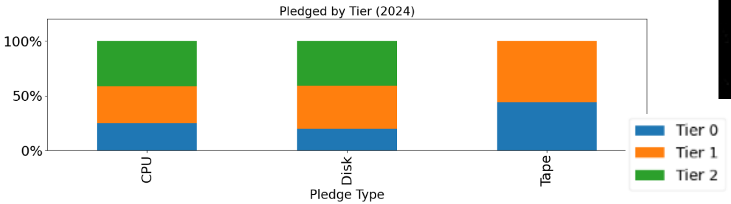


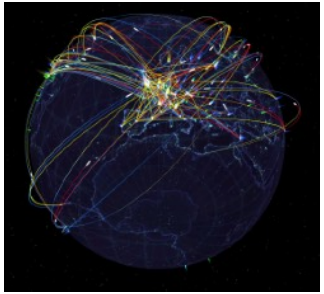


65 MoU's
159 sites
40+ countries

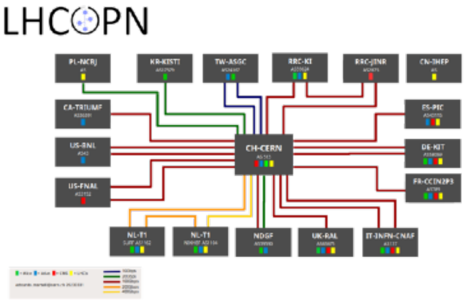
WLCG today

- 159 sites in 40 countries
    - Tier 0 at CERN, 14 Tier 1, Tier 2, Tier 3
- ➜ much less hierarchical than at the beginning,
- ➜ much more dynamic and automatised





2 dedicated networks

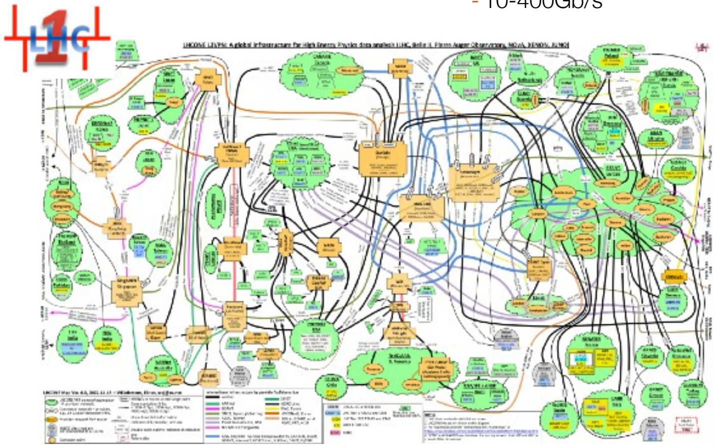


LHCOPN

- Tier 1 and Tier 0
- 13 countries, 3 continents
- 2.1 Tb/s to Tier 0

LHCONE

- 31 network providers
- 117 interconnected sites
- 5 continents
- 10-400Gb/s



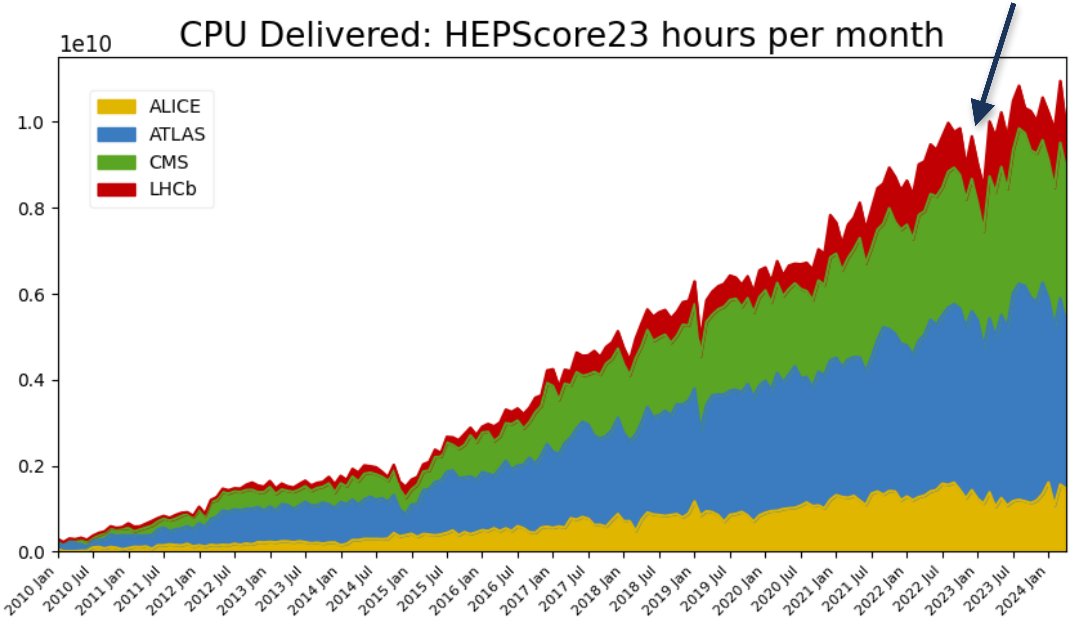
Computing

Total: > 1 million CPU cores

Evolution of delivered CPU (HS23 hours/month)

➜ 2010 – present

Drop in winter 2022/23: energy crisis



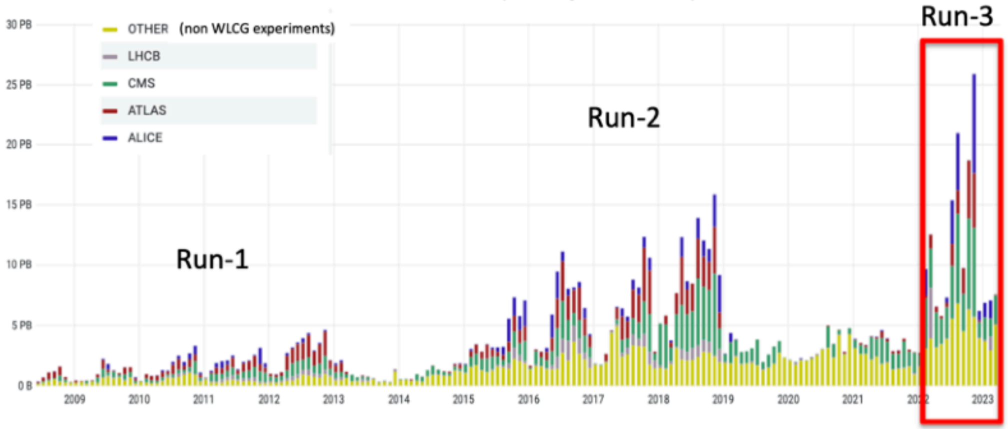
Used CPU capacity

- WLCG T0+T1+T2 and 3
- Pledged or opportunistic
- Note: Cloud/HPC/HLT/ Volunteer Computing resources are not accounted for in this view, unless they are part of or integrated with a WLCG site

Simone.Campana@cern.ch - April 2024 RRB

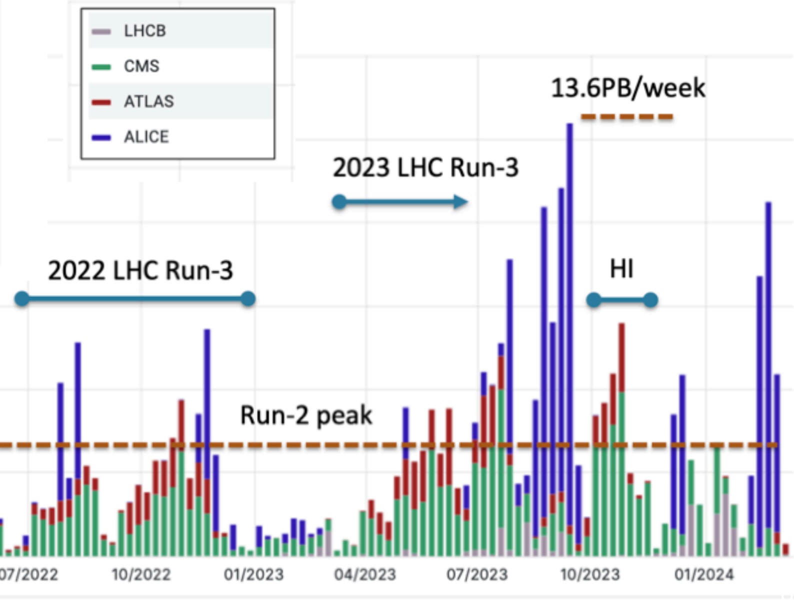
T0 data archiving

Evolution of stored data per month to CERN tape storage



Run 3 zoom

➜ data/week : peak at 13,6 PB/week !

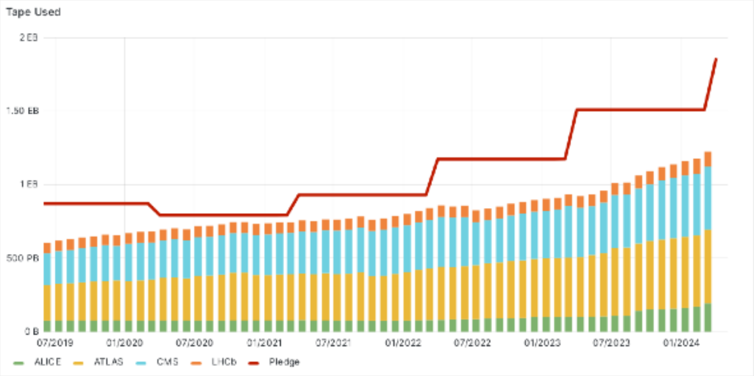


Simone.Campana@cern.ch - April 2024 RRB

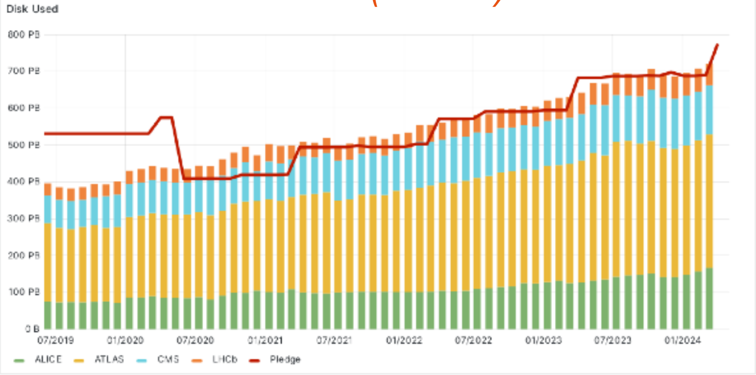
Data

Data Total: *1,9 ExaByte of data (disk + tape)*

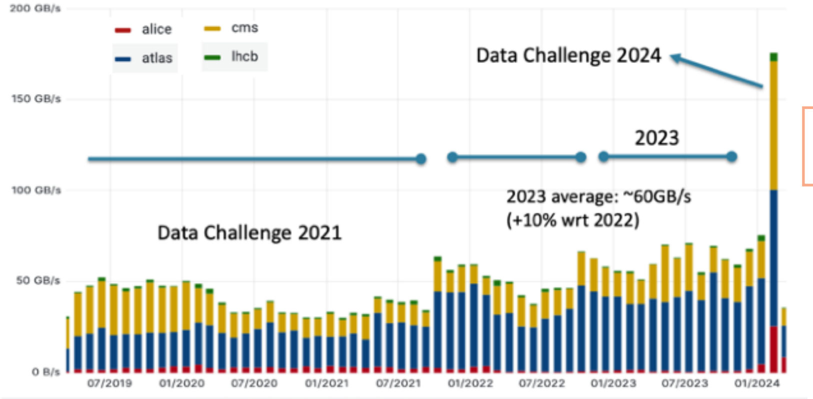
Stored data on tape (TO+T1) ➜ 1,2 EB



Stored data on disk (all Tier) ➜ 720 PB



Transferts between all sites / month



Data Challenge 2024

2023

2023 average: ~60GB/s (+10% wrt 2022)

Data Challenge 2021

Average Transfert Rate 2023: 60 GB/s

Overall

A notable success

- Needs of the LHC experiments successfully met by the construction of a distributed Exascale federation with resources, services, software and dedicated teams worldwide

Within a flat budget

Collaborations strong enough to go through major crises:

- Covid pandemic: WLCG ran smoothly during the pandemic period
- Ukraine war: Russian ressources mostly compensated, Ukrainian sites back online
- and consequences:
  - cope with delay in delivery => Ex: 1 year delay to get network components, other hardware ~x2 wrt to pre-COVID times
  - energy cost increase

What are the HL-LHC needs ?

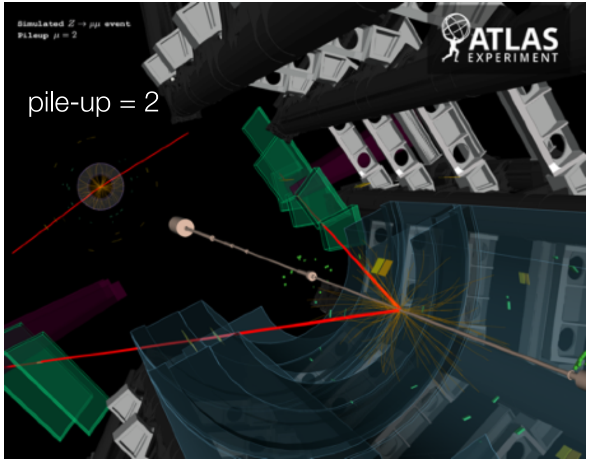
HL-LHC computing challenges (I)

HL-LHC data

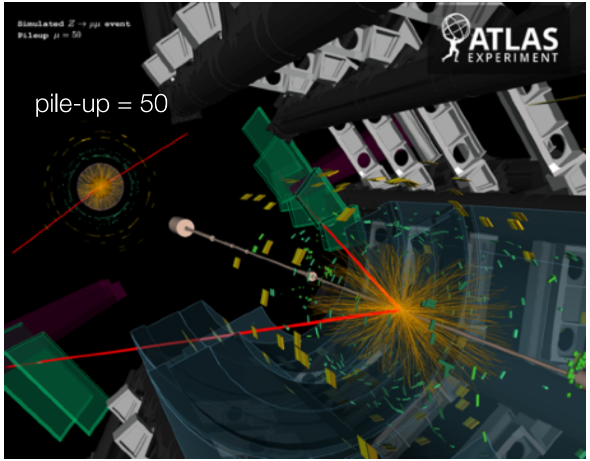
- integrated luminosity = 20 x run 2
- average pile-up ~200 = 6 x run 2

➜ more data and more complex data

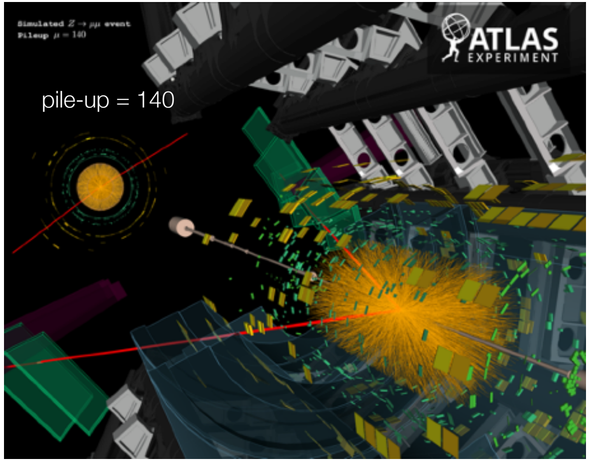
LHC run 1	LS1	LHC run 2	LS2	LHC run 3	LS3	HL-LHC run 4	LS4	HL-LHC run 5
2009 2010 2011 2012	2013 2014	2015 2016 2017 2018	2019 2020 2021	2022 2023 2024 2025	2026 2027 2028	2029 2030 2031 2032	2033 2034	2035 2036 2037 2038
7 TeV 8 TeV		13 TeV		13.6 TeV		13.6-14 TeV		13.6-14 TeV
Pile-up ~20		Pile-up ~35		Pile-up ~45		Pile-up ~ < 140		Pile-up ~ < 200
Trigger ~1 kHz		Trigger ~ 2 kHz		Trigger ~ 4k Hz		Trigger ~10 kHz		Trigger ~10 kHz
29 fb^{-1}		156 fb^{-1}		~350 fb^{-1}		~900 fb^{-1}		~1200 fb^{-1}



pile-up = 2



pile-up = 50



pile-up = 140

HL-LHC computing challenges (II)

HL-LHC throughput

- Data throughput from detector back-ends:

  LHC: 100 GB/s => HL-LHC 1-15 TB/s
- Typical LHC “live-time”: 5Ms/year

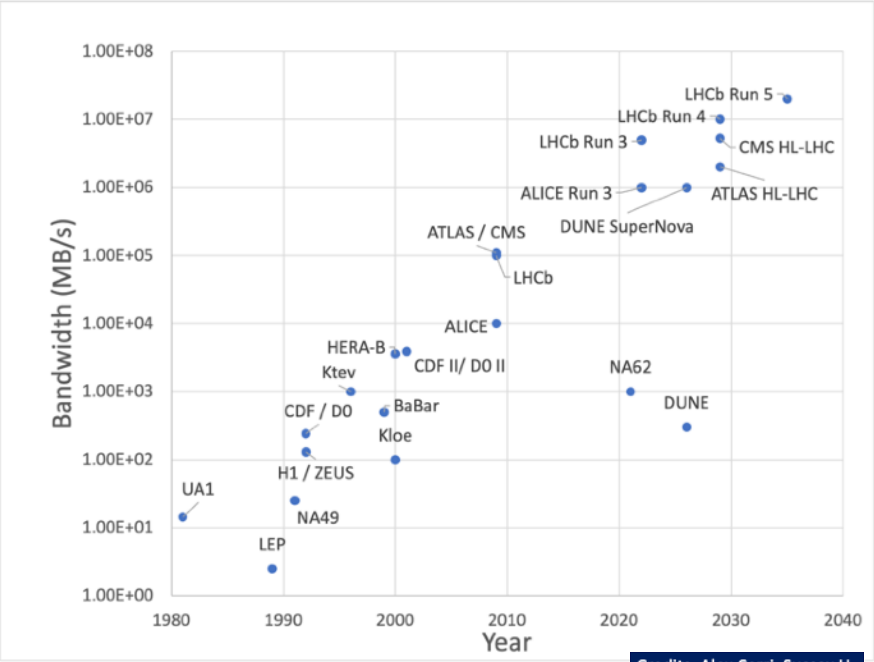
➜ Data volumes: 5-50 EB/year

➜ Triggers will have to significantly reduce data volumes

➜ Expected data throughput from CERN to Tiers1 = 4.8Tb/s !

Also an analysis challenge

	LHC run1-2	HL-LHC run
analysis dataset	10 TB	1000 TB
analysis resource	laptop	analysis facility

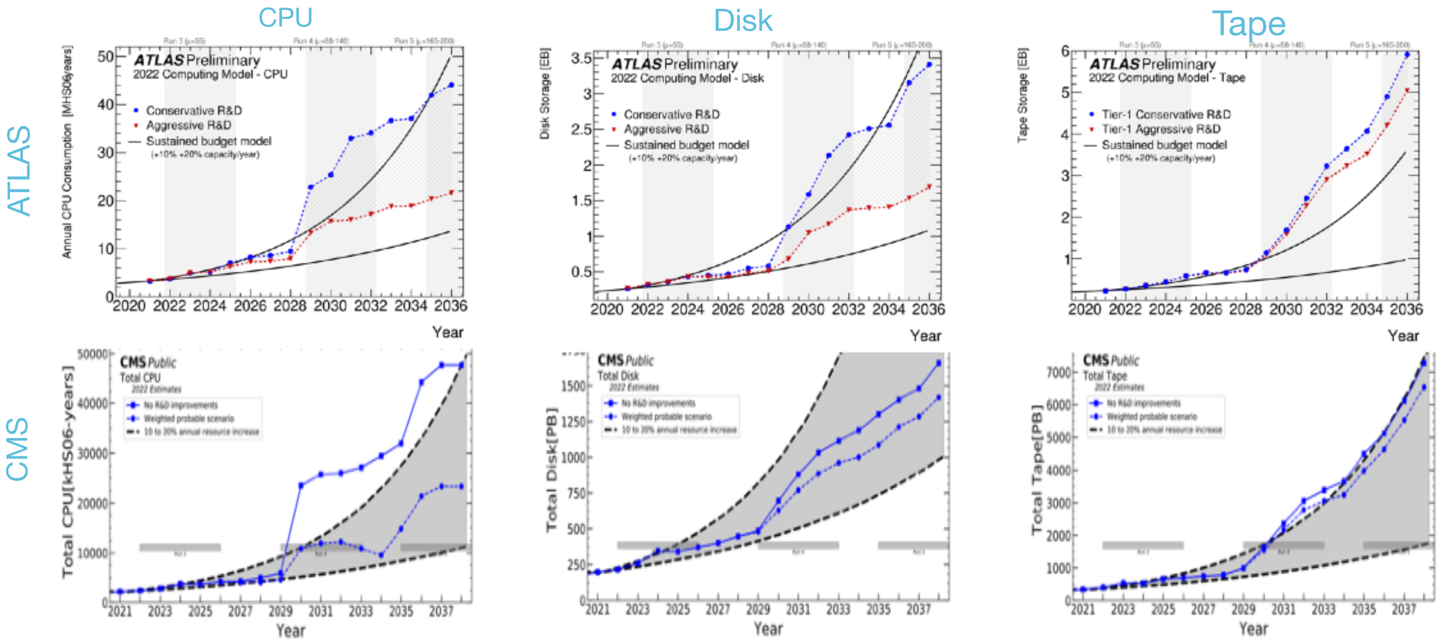


Credits: Alex Cerri, Sussex U.

HL-LHC projections

First projection in 2015 ➜ *Needs = 10x what the flat budget allows !!*

Latest evaluation in 2022 after a lot of work ➜ *Flat budget allows to cover ~ the needs but only in the most optimistic scenario*



A moving context: challenges and opportunities

Hardware costs

Increase of WLCG ressources within flat budget

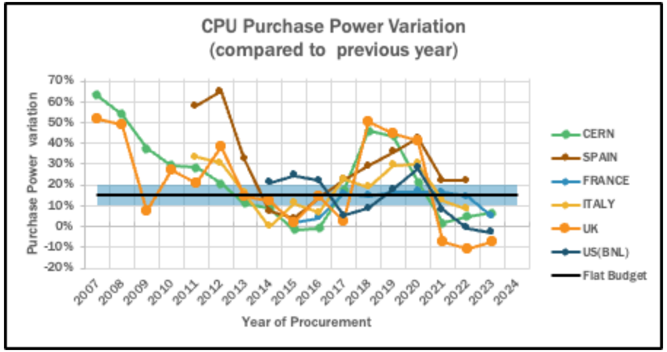
- WLCG model relies on decrease of CPU, disk and tape cost => need ~15% increase of ressource every year for same budget to fulfil experiments needs
- Works well since beginning of LHC (one exception in 2017 with the outstanding performance of LHC)
- Observed average ressource increase in last 5 years (several countries report the evaluation)
    - CPU: +14%/year, disk: +15%/year, tape media: +20%/year

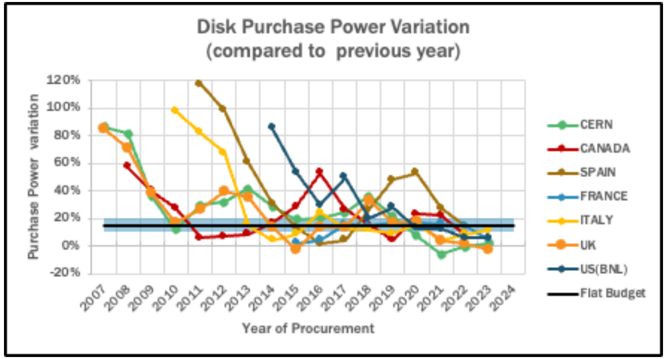
Start to follow up GPU

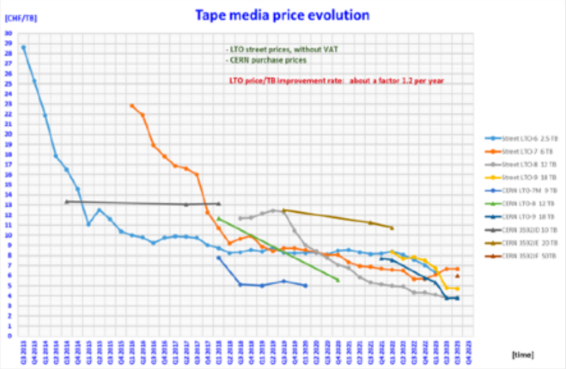
- Trend in GigaFLOPS/USD is favourable for video games but flattened for HPC cards price
    - Very volatile markets, long procurement times (52+ weeks of delivery time). High demand worldwide (AI/ML/ChatGPT)

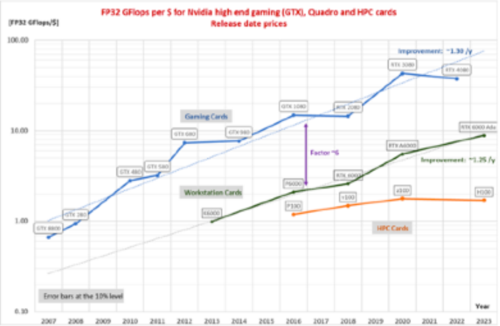
Concerns for the next years

- Much more fluctuations of the market => less predictable prices
- Increase of demand and crises tend to increase prices (in 2024 CPU price increased wrt 2023) and hardware time delivery (x2 these last years)









Hardware evolution

Computing architecture evolution

- Moore's Law: transistor density still doubles every two years
- Clock speed stalls since 2000 too much power used
- Dennard's scaling: power used by silicon device is independent on the number of transistor but proportional to the transistor area

New processors

- GPU: multi-core servers with co-processors and complex memory configuration
  - in 2023 70% of Top500 machine power is from accelerators
  - power consumption controlled
  - multiple competing infrastructure and different programming langage
- non-x86 CPU architectures share increasing (AMD, ARM)
  - more energy efficient than x-86 CPU
  - HPC FUGAKU is based on ARM

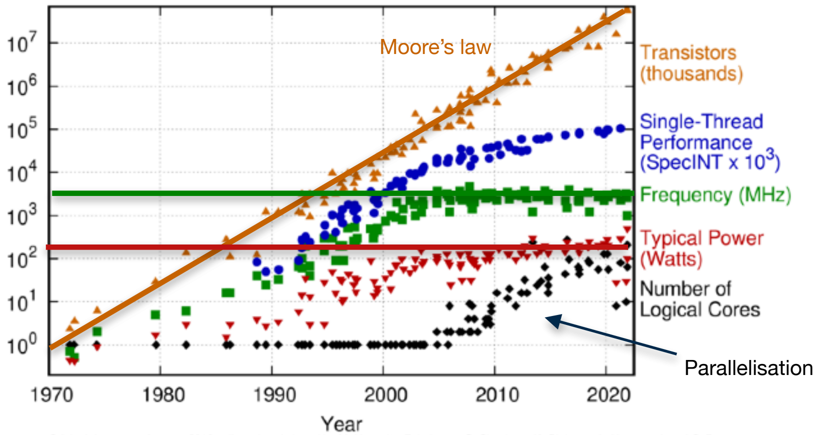
Consequences

➜ evolution towards more parallelism

➜ evolution towards heterogeneous system

Next revolution: Quantum Computing

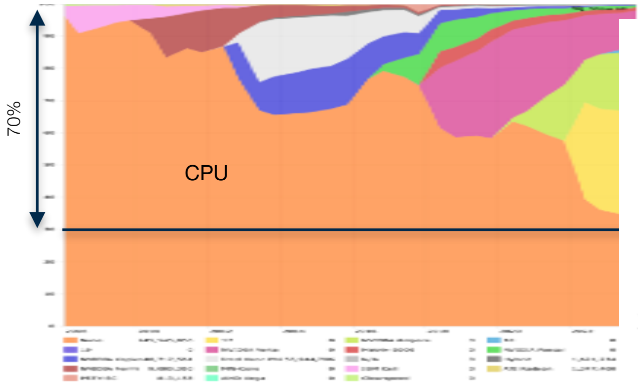
50 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp

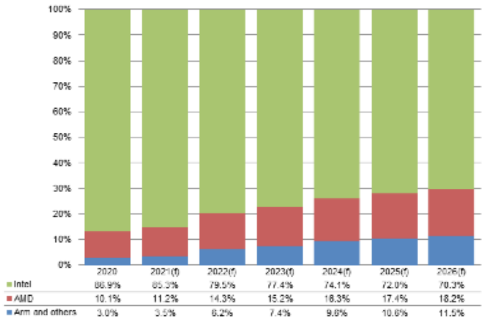
https://github.com/karlrupp/microprocessor-trend-data?tab=readme-ov-file

Accelerator performance share in Top 500



https://www.top500.org/statistics/overtime/

CPU share



	2020	2021(f)	2022(f)	2023(f)	2024(f)	2025(f)	2026(f)
🟩 Intel	86.9%	85.3%	79.5%	77.4%	74.1%	72.0%	70.3%
🟥 AMD	10.1%	11.2%	14.3%	15.2%	18.3%	17.4%	18.2%
🟦 Arm and others	3.0%	3.5%	6.2%	7.4%	9.6%	10.6%	11.5%

https://www.digitimes.com/news/a20211007GS400.html&chid=2

Heterogenous resources

HPC

- Huge investments of countries in HPC machines, entering exascale area
- Challenges
  - very heterogeneous in hardware and policies
  - mostly GPU now
  - not generally suited for data-intensive processing
    - also a network issue !
  - security policies

Commercial clouds

- Clouds flexible, large ressources available
- Challenges:
  - interfaces
  - networking
  - procurement, economic model and vendor locking
- Cost effectiveness ? potentially interesting for special tasks or peak needs

Different hardware

- CPU, GPU, FPGA, ASIC
- Vendors: Intel, AMD, ARM, Power, NVIDIA
- and different programming libraries !

➜ Need portable code

- Portability libraries with abstraction layer to hide the backend implementation and use their parallelism efficiently

kokkos

SYCL™

alpaka

Growing computing needs

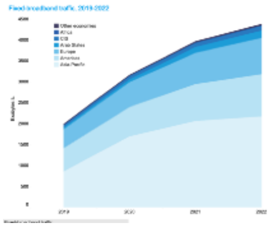
In our research fields

- Several experiments have or will have non negligible computing needs in the near future
- Also in astro-particle and cosmology: Vera Rubin/LSST, CTA, KM3NET, Euclid, ET and SKA and multi-messanger approach
- Nuclear physics: dynamic structuration ongoing, EuroLabs, FAIR, EIC, Lattice QCD…

In other research fields and in society

- Importance of data growing everywhere: improved technologies easier to use and computing ressources more easily available, developments of AI etc

Internet Traffic



https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-internet-traffic/

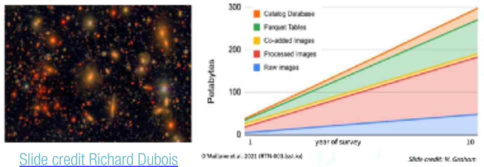
Taking into account Open Science movement

- Open source software and FAIR (findable, accessible, interoperable, reproducible) data
- Important investments for instance in Europe with the construction of the EOSC (European Open Science Cloud

VERA C. RUBIN OBSERVATORY LSST

The final LSST 10-year sky map will be like having ~3 million of these, tiled over the entire southern sky.

The Rubin Observatory's total data holdings will start (after DR1) at ~40 PB and grow to ~300 PB over the 10-year LSST.

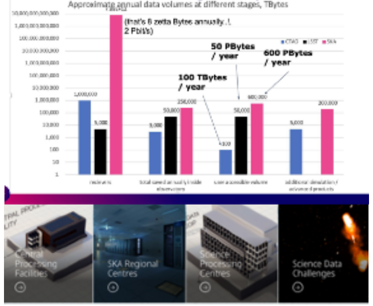


Slide credit Richard Dubois

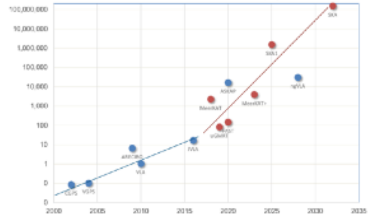
SKAO Rosie Bolton @ CHEP 2023

Annual Data volumes through the systems

Approximate annual data volumes at different stages, TBytes



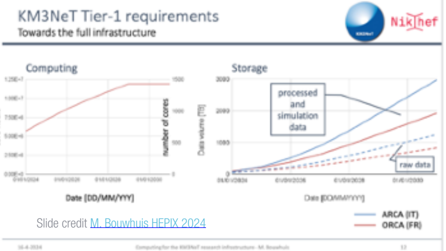
RADIO ASTRONOMY DATA VOLUMES (TB)



Taylor R. et al. Big Data Research Infrastructure Collaboration Toward the SKA (BRICSKA). doi: 10.1590/0001-3765202120201027. PMID: 34076205.

KM3NeT Tier-1 requirements

Towards the full infrastructure



Slide credit M. Bouwhuis HEPIX 2024

16-4-2024 Computing for the KM3NeT research infrastructure - M. Bouwhuis 12

Computing from Nuclear structure, reactions & astrophysics

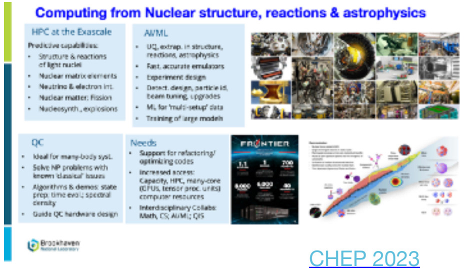
HPC at the Exascale

Predictive capabilities:
- Structure & reactions of light nuclei
- Nuclear matrix elements
- Neutrino & electron int.
- Nuclear matter: fission
- Nucleosynth., explosions

AI/ML
- UQ, extrap. in structure, reactions, astrophysics
- Fast, accurate emulators
- Experiment design
- Detect. design, particle id, beam tuning, upgrades
- ML for 'multi-setup' data
- Training of large models

QC
- Ideal for many-body syst.
- Solve NP problems with known 'classical' bases
- Algorithms & demos: state prep; time evol.; spectral density
- Guide QC hardware design

Needs
- Support for refactoring/ optimizing codes
- Increased access: Capacity, HPC, many-core (GPUs, tensor proc. units) computer resources
- Interdisciplinary Collabs: Math, CS; AI/ML; QIS



CHEP 2023

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT Slide credit Heidi Schellman

Projected disk needs

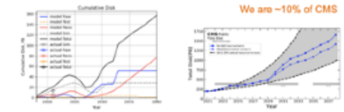


Figure 6.4: Estimated size of various disk samples in PB for DUNE and CMS at the HL-LHC for comparison. This estimate includes retention policies and multiple copies. The points show actual use in 2021 which was lower than planned due to delays in distributing second copies of samples to remote sites.

Artificial Intelligence

Use of AI since decades in HEP

- signal/background separation, particle identification, multivariate analysis… ex BDT widely used since 90’

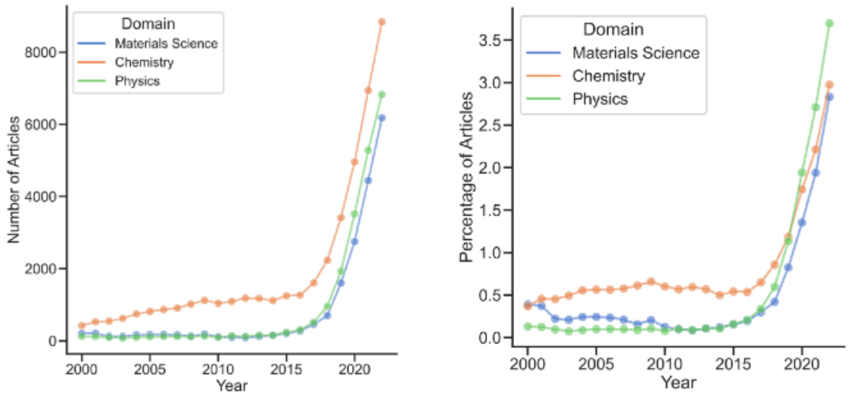
Modern AI and computing ressources allow new usage

- generative models, unsupervised classification, low latency inference
- CPU, GPU, FPGA implementation
- AI assisted code generation

➜ AI usage and developments at all stages of computing in our fields

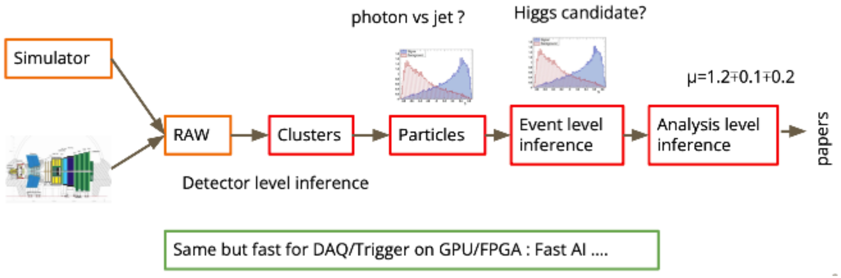
➜ See Tobias Golling talk

ML Publications in Science



Ben Blaiszik, "2021 AI/ML Publication Statistics and Charts". Zenodo, Sep. 07, 2022. doi: 10.5281/zenodo.7057437.

CHEP 2023



photon vs jet ?

Higgs candidate?

Simulator

RAW

Clusters

Particles

Event level inference

Analysis level inference

papers

$\mu=1.2\mp0.1\mp0.2$

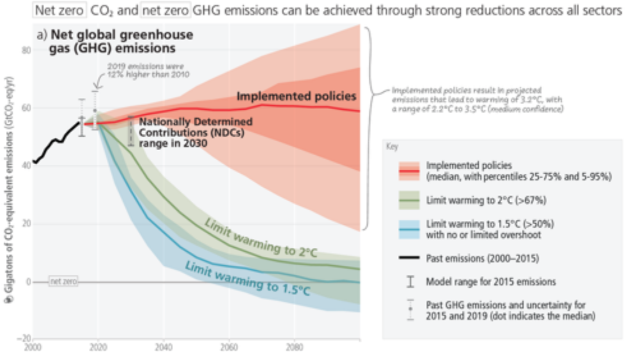
Detector level inference

Same but fast for DAQ/Trigger on GPU/FPGA : Fast AI ....

© David Rousseau

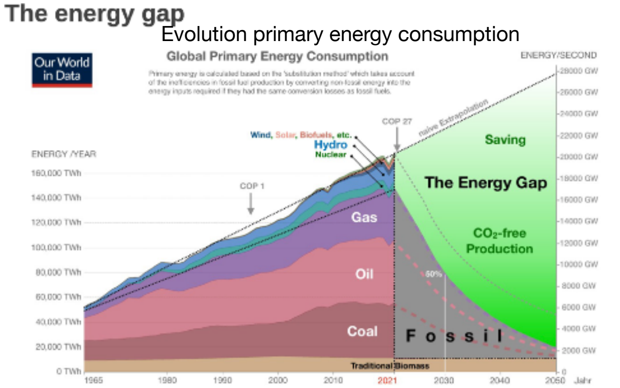
Sustainability

Evolution of Greenhouse Gaz Emissions



IPCC 2023 Synthesis Report, Ref. [7]

Evolution primary energy consumption



https://ourworldindata.org/grapher/electricity-prod-source-stacked

The contribution of data-centers to greenhouse-gas emission is sizeable and growing, no exception in our fields of research

Paris agreement

- Zero emission around 2100 => -50% by 2030 !
- To get the 50% in 2030 you have equivalently
  - to expand CO2 free energies by a factor 12
  - to increase energy efficiency by a factor 2
  - to save energy by a factor 2
- will have a mix of these

➜ What translation in our computing models and infrastructure ?

➜ See for instance HECAP+ initiative

The way to HL-LHC

Development plan

Infrastructure

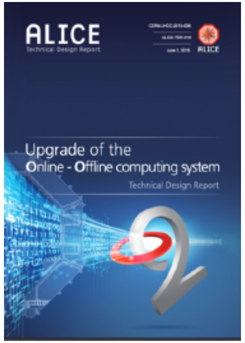
- new developments while keeping the infrastructure in production
  - ➜ scale-up tests only possible on top of the production grid
  - ➜ Data Challenge to test and push the infrastructure closer to HL-LHC request
- continuous developments since run 2 => validated new tools integrated straight away
  - ➜ some work for HL-LHC already in use in run 3

Software as a key element

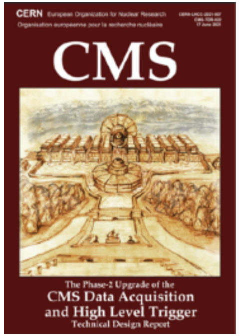
- is one of our best lever arm
  - ➜ need to tackle all elements: trigger, generation, simulation, reconstruction, analysis
  - ➜ make the best use of new technology and technics

Towards HL-LHC computing models


UPGRADE TDR
LHCb Computing Model
Technical Design Report


ALICE
Technical Design Report
Upgrade of the Online - Offline computing system
Technical Design Report


ATLAS HL-LHC Computing Conceptual Design Report


CERN European Organization for Nuclear Research
Organisation européenne pour la recherche nucléaire
CMS
The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger
Technical Design Report

A collaborative effort beyond HL-LHC

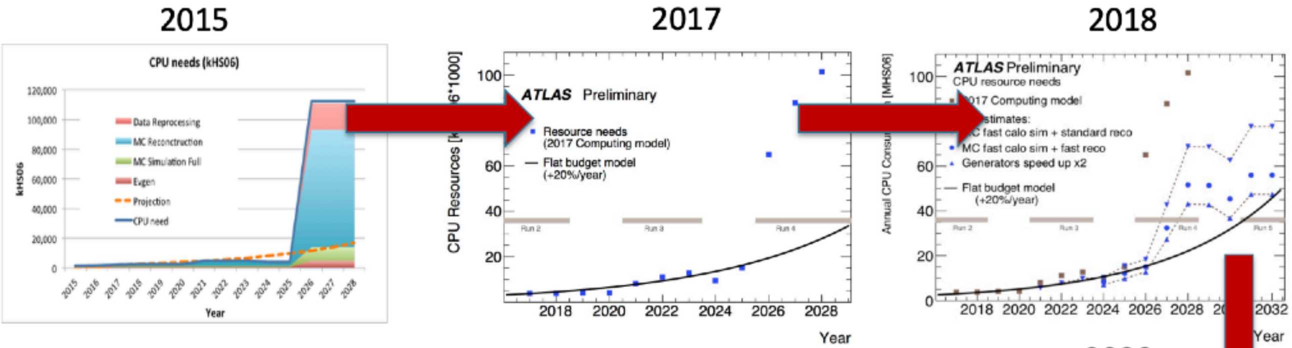
Enlarge collaborations

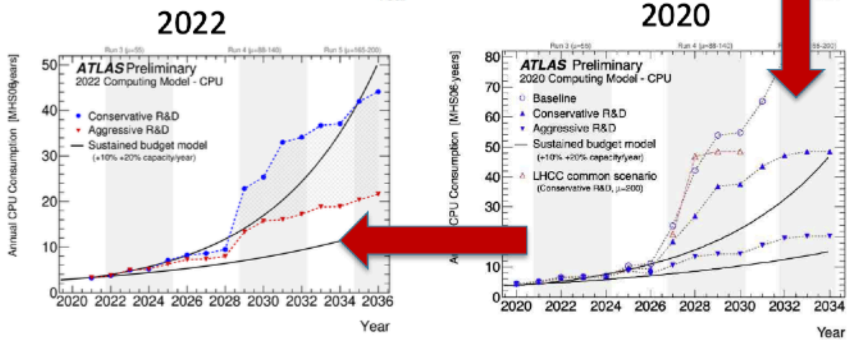
- **Common developments** allow to share expertise, to share effort and find common solution when possible
  - ➜ ease deployments and allow shared ressources and infrastructure
- Eased by international collaborations and programs
  - HEP Software Foundation (HSF), IRIS-HEP for software developments
  - WLCG/DOMA for WLCG infrastructure expands beyond LHC with DUNE, Belle-2, JUNO and VIRGO as WLCG observers
  - European programs for the development of the European Open Science Cloud (EOSC) => ESCAPE project for HEP, astronomy and nuclear physics, EVERSE project for software
  - JENA computing workshop to discuss synergies across the 3 communities
    - 5 working groups created to draft a report by the end of the year to be discussed at next year JENA symposium with Funding Agency
      - HPCs, SW and heterogeneous architectures, Data Management and FAIR data, ML, Training – Dissemination – Education

DOMA

Results already visible !

HL-LHC computing resource needs evolution





2015 projections: resource needs = 10x more than budget allows

2022 projections: resource needs compatible with budget (optimistic scenario)

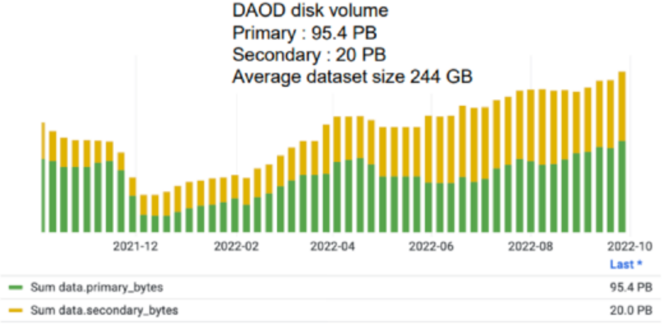
Simone Campana CHEP 2023

Infrastructure

Storage and data management

Difficult challenge

- No opportunistic storage and storage hard to operate => storage needs are the hardest ressource to fulfil
- Constant work to optimise data distribution [ATLAS CHEP 2023]()
    - lifetime model, regular/automatic cleaning of unused or superseded data
- Early R&D and test through DOMA collaboration and ESCAPE EU project and experiment

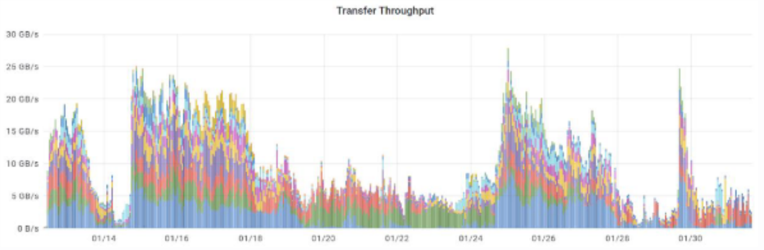


DAOD disk volume
Primary : 95.4 PB
Secondary : 20 PB
Average dataset size 244 GB

Better usage of tape: data carousel

- tape driven workflow to allow job to get data directly from tape

  => disk space saving
- close collaboration between sites, experiments workflow and data management system
- in production in ATLAS with peak rate 20-25 GB/s
- develop now derived reconstruction data on demand
- Tape smart writing: find intelligent algorithms for file placement on tape



Tape staging from T0/T1s in 2023-01 (ATLAS DDM dashboard)

[CHEP 2023]()

Storage and data management

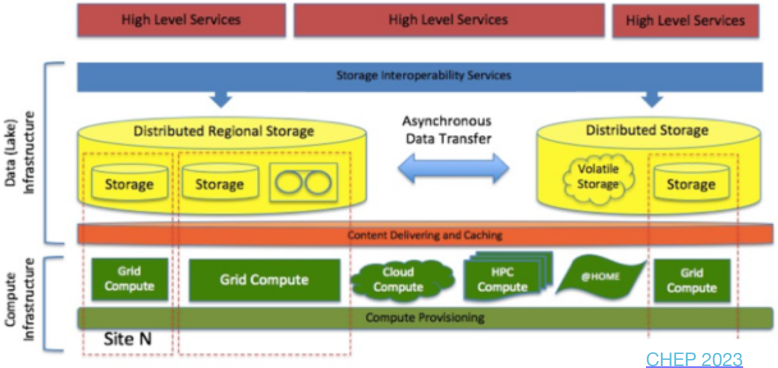
Data Lake

- Put data in a distributed federated storage that minimise replication, assure availability, with different storage quality level depending on the needs
- Serve data to remote (or local) compute ressources => grid, cloud, HPC, ..
- Simple caching is all that is needed at compute site (or none, if fast network)

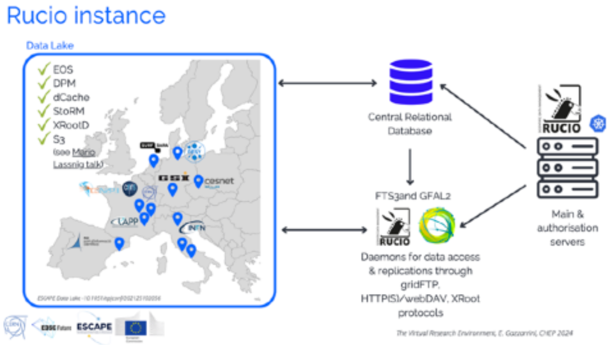
ESCAPE Data Lake

- Based on CERN File Transfer Service (FTS), Rucio data management and orchestration open-source tools
- Security assured through token based AAI
- HEP development benefited to astronomy and nuclear physics collaborations, now collaborative developments within ESCAPE new collaboration (CTA, SKA, LSST, KM3NET, LOFAR, FAIR, HL-LHC etc)
- Part of EOSC-Future Virtual Research Environment, successfully exploited with ESCAPE Dark Matter Science Project CHEP 2023 CHEP 2023

➜ See also Data lake implementation in the Nordic countries CHEP 2023



CHEP 2023



The Virtual Research Environment, E. Gazzarrini, CHEP 2024

CHEP 2023

Data Challenges

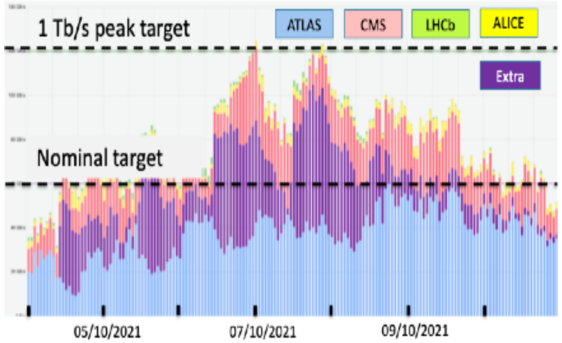
Objective

- progressively test the infrastructure by increasing data rate on top of production
- allows to test concurrently in real condition on a large set of sites end-to-end transfer capacities:
  - network, storage, middleware
- allows to test new technologies

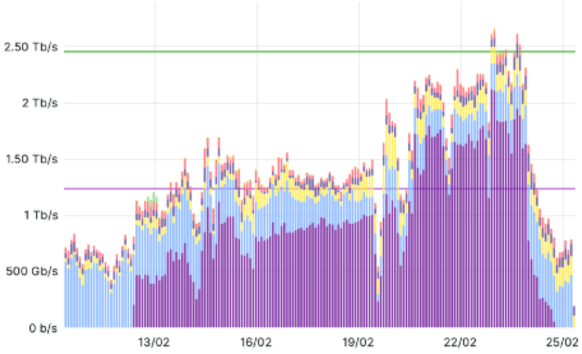
Campaigns

- DC21: target 10% of HL-LHC needs
  - 1Tb/s reached ! and lessons learned
  - [2021 Data Challenges Wrap Up](https://indico.cern.ch/)
- DC24: target 25% of HL-LHC needs
  - 1,2 Tbps **minimal** scenario, 2,4 Tbps **flexible** scenario
  - including the 4 LHC experiments + Belle2 + DUNE
  - New monitoring following the DC21 post mortem, tokens usage tests
  - Token AAI tested

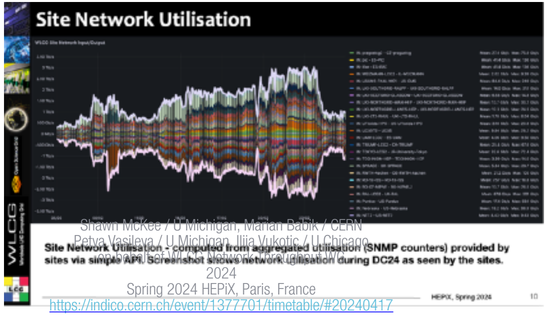
DC21 WLCG data transfers (Gbps) – 15 days



DC24 WLCG data transfers (Gbps) – 15 days



Site network monitoring

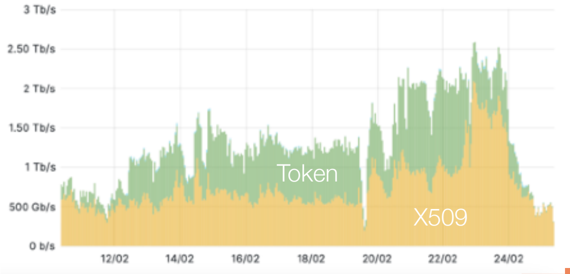


Site Network Utilisation - computed from aggregated utilisation (SNMP counters) provided by sites via simple API. Screenshot shows network utilisation during DC24 as seen by the sites.

HEPiX, Spring 2024 10

Shawn McKee / U Michigan, Marian Babik / CERN
Petya Vasileva / U Michigan, Ilija Vukotic / U Chicago
on behalf of WLCG Network Throughput WG
2024
Spring 2024 HEPiX, Paris, France
https://indico.cern.ch/event/1377701/timetable/#20240417

Token utilisation in transfers



[Introduction to token](https://indico.cern.ch/)

[IAM datachallenge WLCG-HSF 2024](https://indico.cern.ch/)

Heterogenous computing

GPU usage already in run 3 (I)

ALICE continuous readout

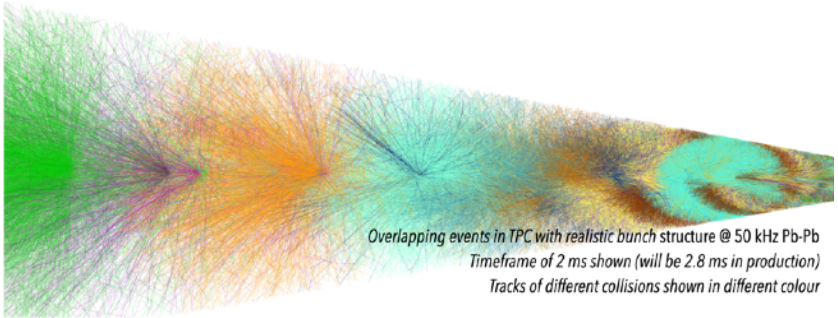
- Heavy ion data taking rate increased from 10 to 50 kHz at run 3
- With the high TPC latency, decision to operate in continuous readout mode **without trigger.**
  - ➜ Time-frame (TF) of 2.5-20ms instead of event acquisition

O2 facility

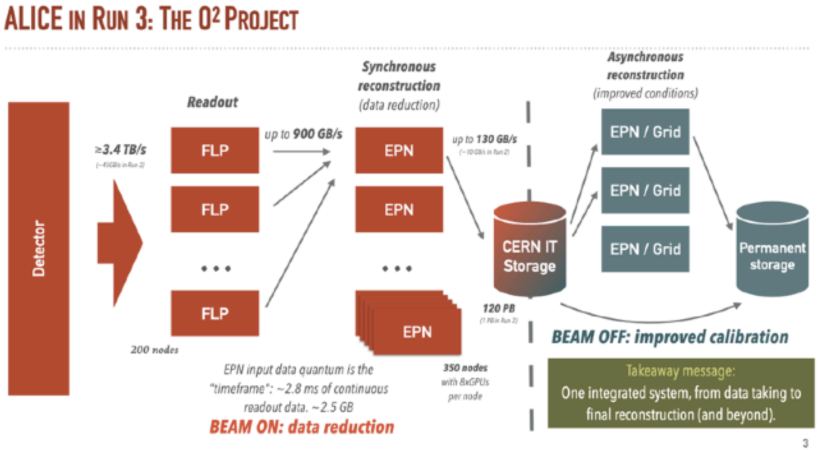
- First Level Processing (FLP): Readout + FPGA corrections to compress data and build the time-frames
- Event Processing Node (EPN): 8 GPUs AMD MI50 – 2 CPUs 32 core AMD Rome 7452
  - online synchronous reconstruction and data reduction when beam on
  - asynchronous reconstruction when beam off
    - Offloading 60% of the processing to the GPU induces a speedup ~2.5x, expect 5x with 80%
    - 2023 first use of EPN farm: 1/3 of data processed on EPN (CPU+GPU) and 2/3 on GRID (CPU)

Performance

- 55 CPU 3.3GHz Cores equivalent to a single AMD-Mi50 GPU



Overlapping events in TPC with realistic bunch structure @ 50 kHz Pb-Pb
Timeframe of 2 ms shown (will be 2.8 ms in production)
Tracks of different collisions shown in different colour



ALICE IN RUN 3: THE O² PROJECT

EPN input data quantum is the "timeframe": ~2.8 ms of continuous readout data. ~2.5 GB

BEAM ON: data reduction

BEAM OFF: improved calibration

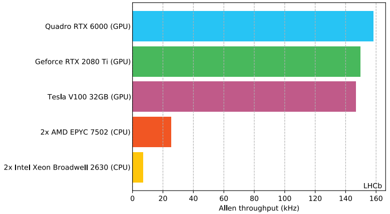
Takeaway message:
One integrated system, from data taking to final reconstruction (and beyond).

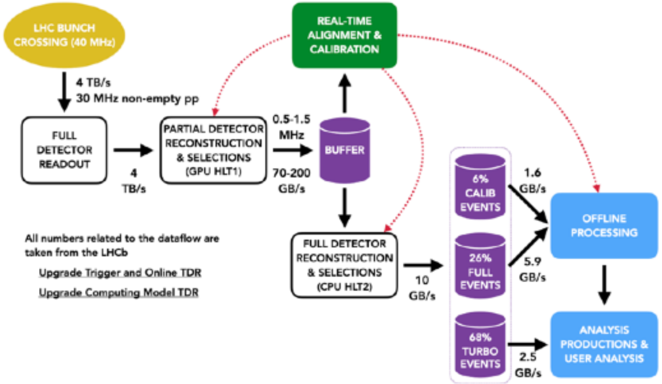
[WLCG-HSF 2023](#) [CHEP 2023](#)

GPU usage already in run 3 (II)

Full Software trigger at LHCb

- For Run 3, LHCb is getting 5 × more collisions per second => trigger becomes the bottleneck
- Goal evolved from rejecting (trivial) background to categorise different 'signals'
- => Drop the L0 trigger, reconstruct 30 MHz of events before making trigger decision





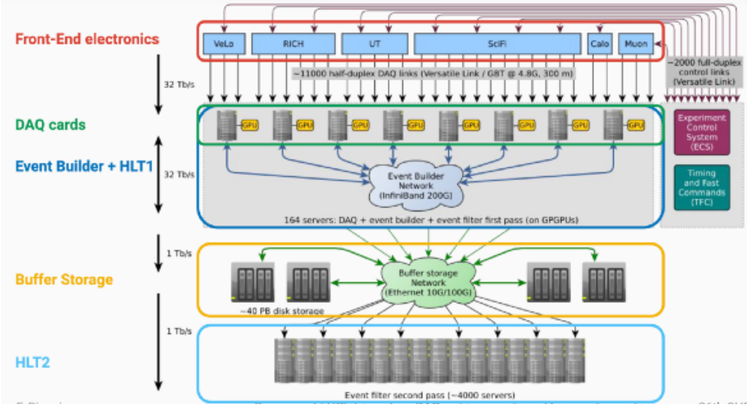
All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR

Upgrade Computing Model TDR

RTA: Real-Time Analysis

- FPGAs-based clustering for Silicon Pixel detector
- HLT1: GPU based reconstruction
  - 163 RTX A5000s (one per Event Builder node)
  - software HLT1 sequence has been implemented in CUDA
- HLT2: CPU-based full reconstruction



Successfully commissioned

- In 2022: 40 Tbits/s network throughput achieved ➜ equivalent to 4% of the internet
- Achieving 30 MHz with less than 200 GPUs !

CHEP2023 - 1 and CHEP 2023 - 2

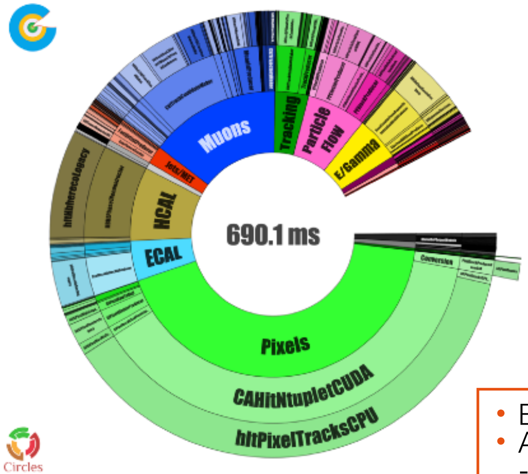
GPU usage already in run 3 (III)

CMS heterogenous HLT farm

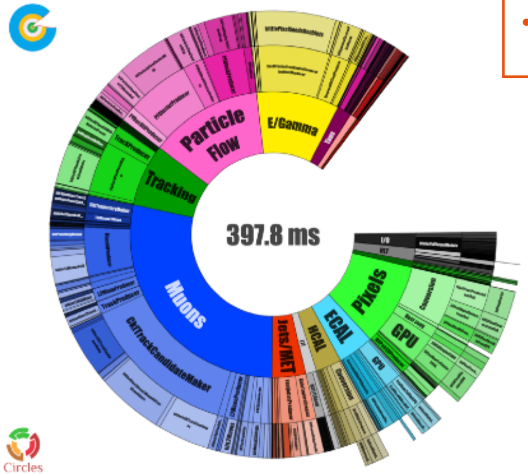
- HL-LHC will require ~one order of magnitude faster trigger
  - L1T 100kHz -> 750kHz (7.5x), pile-up 60 -> 200 (~3x) + more complex detectors
- => include GPUs at the HLT already for run 3
  - 200 nodes 2CPU+2GPU = ~26k CPU cores AMD Milan 7633 + 400 NVIDIA T4 GPUs
- HCAL, ECAL, pixel local reconstruction and pixel tracking on GPU => a lot of work needed to optimise software on these heterogenous architectures
- Allow scouting strategy = increasing event rate (lower thresholds) and decreasing event size to perform analysis directly with objects reconstructed at the HLT (no RAW data is stored

R&D in Performance Portability

- Porting Heterogenous Code to Alpaka (performance portability library) – to reduce dependency on hardware; kokkos and sycl also studied
- Aiming to offload 10% of (Run-3 and Phase-2) offline reconstruction by end of 2023



- Event throughput +80%
- Average time per event -40%
- Power consumption: -30%
- Experience gained for offline computing



WLCG-HSF 2023

CMS CHEP 2023

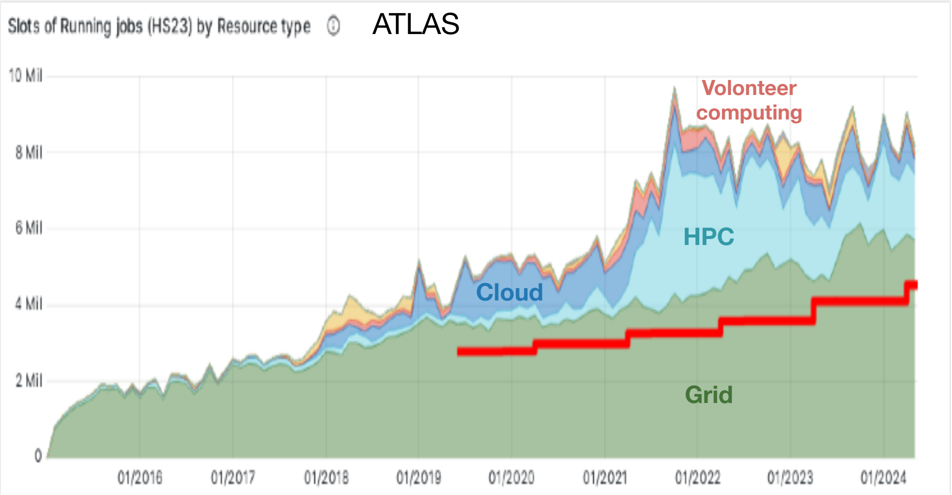
HPC usage

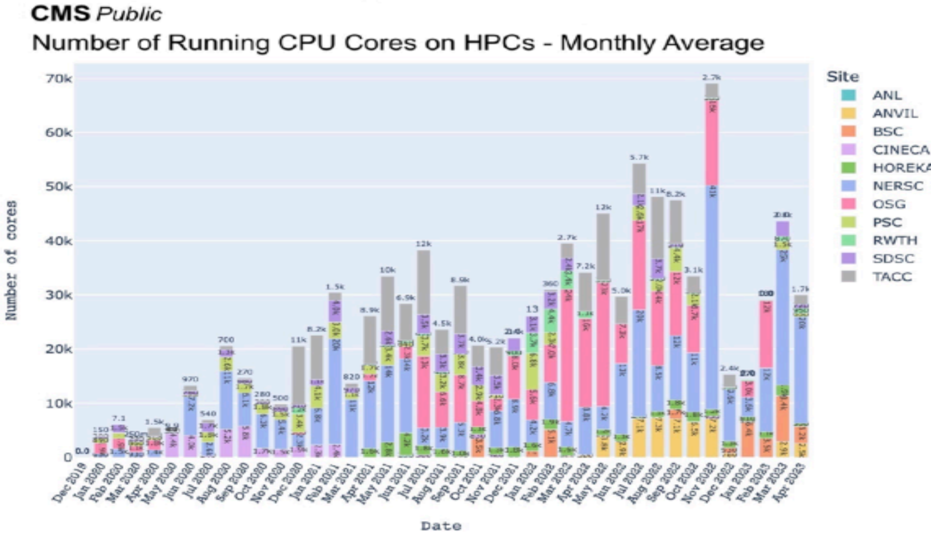
Increased use of HPC ressources

- Mostly for MC simulation, some MC reconstruction
- Exploiting HPCs could be
  - relatively easy for those which have policies similar to HEP sites ones
  - very complicated In other cases: workload management systems had to be creatively adapted in order to use these resources

Challenges

- low network connectivity on the worker nodes
- security rules prevent access to standard software repositories (CVMFS)
- Ingress/egress requirements
- ressources booking


Slots of Running jobs (HS23) by Resource type ATLAS


CMS *Public*

Number of Running CPU Cores on HPCs - Monthly Average

Cloud ressources

➜ Clouds in WLCG sites have been used for a long time

Commercial Cloud usage

- Clouds have a lot of different ressources and allows for flexibility, on demand ressources
- Some challenges: interface, network, provider specificities, vendor locking
- Also procurement and economic model
  - ➜ Cost effectiveness to be demonstrated

ATLAS Google project

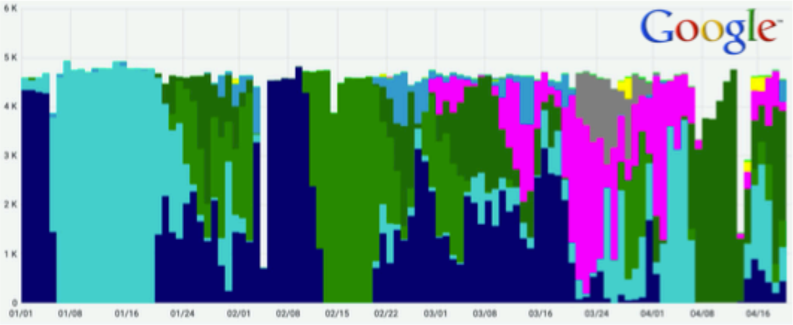
- Full integration of Google into ADC activities and infrastructure: important work to adapt data and job management tools (Panda, Rucio)
  - PanDA & Rucio developments are cloud-independent
  - Available for user analysis with Dask & Jupyter

➜ Work to understand cost of commercial cloud with respect to owning our ressources

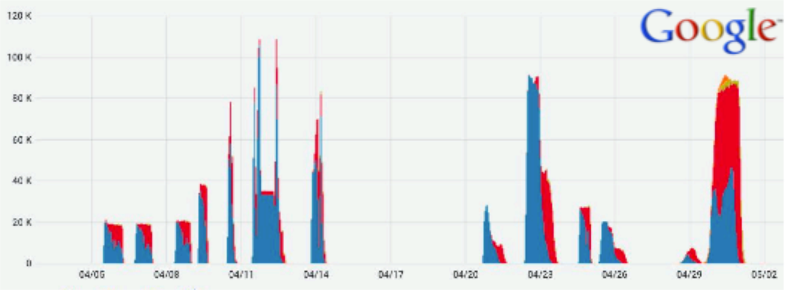
➜ Look for interesting use cases

- Cloud bursting: Dynamic/on-demand slot allocation
- Network offloading: Use Google network for transfers
- GPUs: On-demand GPU hardware
- Special analysis workflows: ML, fitting, special MCs

Running various w/flows on 5k job slots at the Google site



Bursting up to 100k slots for simulation and reconstruction jobs



ATLAS CHEP 2023

Software: generation, simulation, reconstruction, analyse

Event Generation

➜ will represent 10-20% of HL-LHC Computing

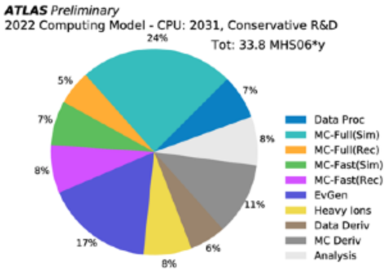
R&D in HSF Physics Event Generator WG

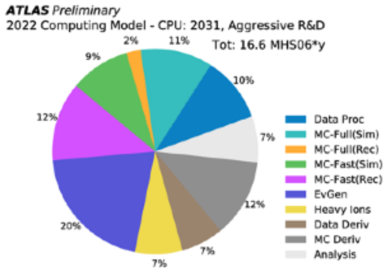
- Main Objectives [Ref]
  - Accounting and profiling to understand CPU costs
  - Move to GPUs and vectorised code
  - Optimise phase space sampling and integration algorithms, including AI use
    - ➜ see MadJax (differentiable per-event MEs)
  - Research on reducing the cost associated with negative weight events
    - ➜ see New Sherpa configuration by ATLAS that achieves 50% reduction of negative weights [Ref]
  - Promote collaboration, training, funding and career opportunities

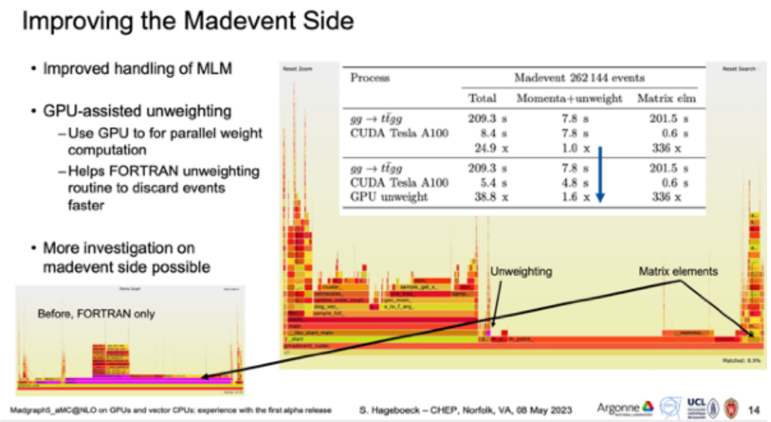
Example: MadGraph

- Matrix Element calculation in event generator can be efficiently parallelised using vector CPU or GPUs
  - Vector CPU: Speed ups of 6x for AVX512
  - GPU: 20x to 80x depending on the process using Tesla A100

➜ reengineering of MG5aMC is a functional alpha release for LO QCD / EM processes, but weak interactionsneed more work







Improving the Madevent Side

- Improved handling of MLM
- GPU-assisted unweighting
  - Use GPU to for parallel weight computation
  - Helps FORTRAN unweighting routine to discard events faster
- More investigation on madevent side possible

Process	Madevent 262 144 events		
	Total	Momenta+unweight	Matrix elm
$gg \to t\bar{t}gg$	209.3 s	7.8 s	201.5 s
CUDA Tesla A100	8.4 s	7.8 s	0.6 s
	24.9 x	1.0 x	336 x
$gg \to t\bar{t}gg$	209.3 s	7.8 s	201.5 s
CUDA Tesla A100	5.4 s	4.8 s	0.6 s
GPU unweight	38.8 x	1.6 x	336 x

Madgraph5_aMC@NLO on GPUs and vector CPUs: experience with the first alpha release S. Hageboeck – CHEP, Norfolk, VA, 08 May 2023 14

MadGraph CHEP 2023

Simulation

Simulation impact

- Simulations are dominating Run 3 CPU usage: ALICE 50%, LHCb 90%, ATLAS 50%

➜ Lot of R&D towards new technique and use of accelerators

- Need
  - to improve accurate full simulation for detector performance and training of last simulation
  - to make more use of fast simulation

Fast simulation

- originally based on parameterisations of detector response, (generative) ML techniques now widely used

➜ ATLAS AtlFast3, CMS FastSim and FlashSim

➜ Porting fast simulations on heterogeneous resources, exploiting portability models Cf FastCaloSIm

Full simulation

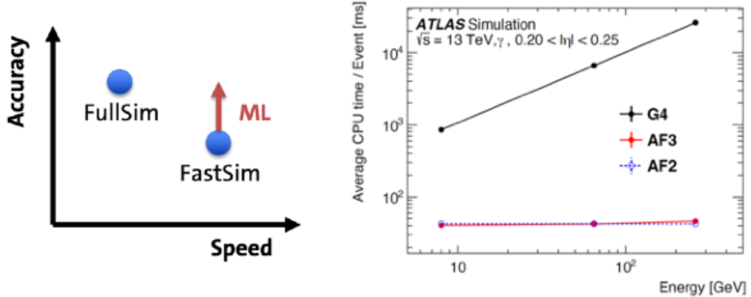
- Continuous progress to speed up Full Simulation:

➜ same accuracy in the physics description, but faster

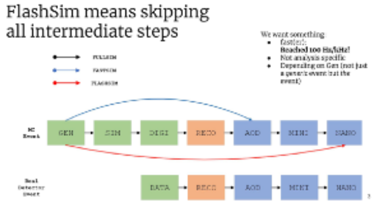
➜ recent 20% improvement for ATLAS simulation (Woodcock Tracking of gamma – avoid to stop at volume boundaries)

- R&D with GPU prototypes: EM shower simulation on GPU
  - AdePT (Accelerated demonstrator of electromagnetic Particle Transport), Celeritas

Full vs fast Simulation



FlashSim

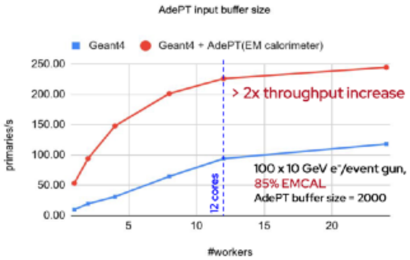


FlashSim means skipping all intermediate steps

We want something

- faster!:
  Reached 100 Hz/kHz!
- Not analysis specific
- Depending on Gen (not just a generic event but the event)

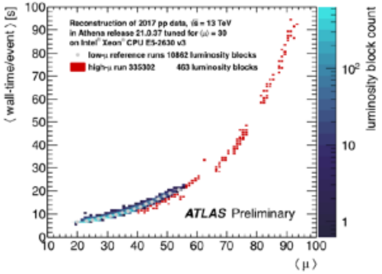
Adept

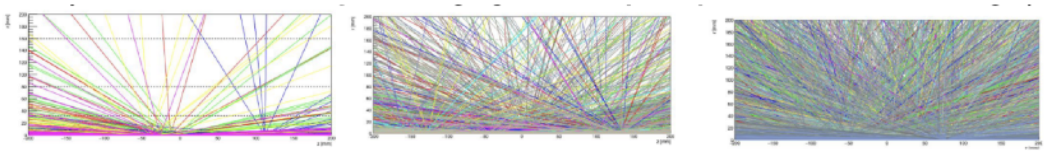


Reconstruction

R&D on track reconstruction with high pile-up

- non linear increase of needed computing ressource with event complexity for classical reconstruction algorithms

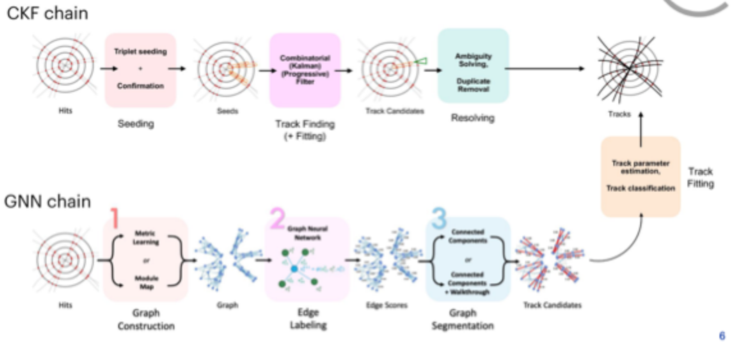




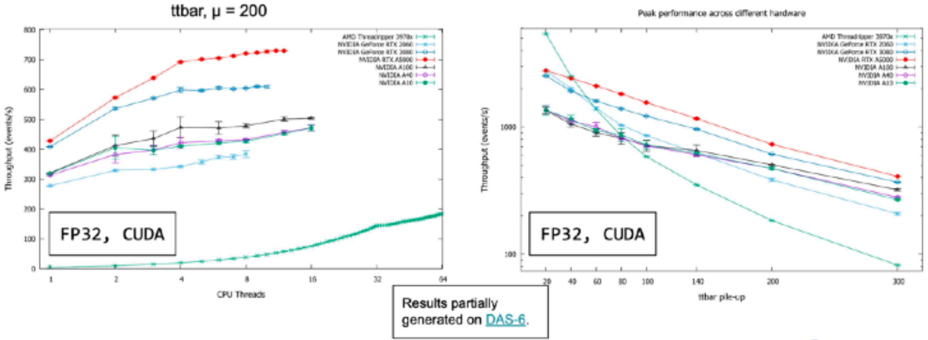
2010 <μ> ~ 5 2018 <μ> ~ 40 2029 <μ> ~ 200

ACTS project

- ACTS open source project: toolbox for charged particle reconstruction for HEP/NP
- R&D on method and algorithm, 2 main chains developed: Kalman Filter or Graph Neural Net (GNN)
- R&D on parallelisation and performance portability
- On track to be used by ATLAS for track reconstruction by the LHC's Run-4



Event throughput for ttbar



Results partially generated on DAS-6.

GPUs become competitive at high pile-up. Highest performance observed on NVIDIA® workstation GPUs so far.

CHEP 2023

Analyse

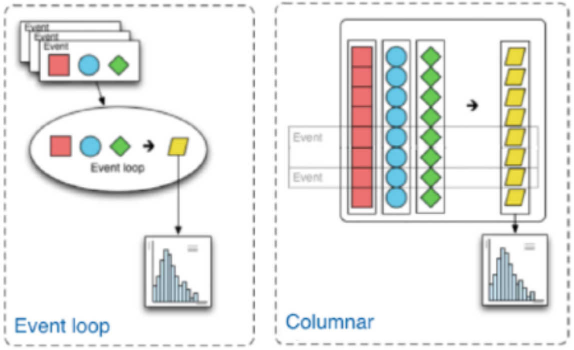
Analysis challenge

- Typical dataset size 10 TB (run 1-2) => 1000 TB (run 4)
- Current workflows do not scale:
  - disk/tape are limited/expensive and analysis data are projected to occupy ~30% (ATLAS model)
  - will need to go from Laptop to Analysis Facility

Analysis R&D

- Analysis workflow evolution: Columnar Analysis going from event loop to array programming and more use of Python ecosystem and industrial standards
  - Coffea: python ecosystem -> ScikitHEP, Awkward array, DASK
  - ➜ array programming solution for quick insights delivery
- Developments of more compact formats and reducing data copies and intermediate formats
  - ➜ NanoAOD and ATLAS PHYSLITE
  - ➜ target ~10kB/event for fast and lightweight analysis
- ROOT's Tree -> RNTuple, RDataFrame high level API for data analysis
  - encodings and I/O upgrade of the event data file format and access
  - Access to novel and future storage technologies: native support for HPC and cloud, object stores, direct disk-to-GPU data transfer
  - Collaboration with experiment for Event Data Model integration and optimisation
  - ➜ 10-20% smaller files, x3 to x5 better single core performance
  - ➜ enables fast adaptation to modern technologies, like object stores
  - ➜2024 goal: release RNTuple 1.0 supporting the EDM of all experiments

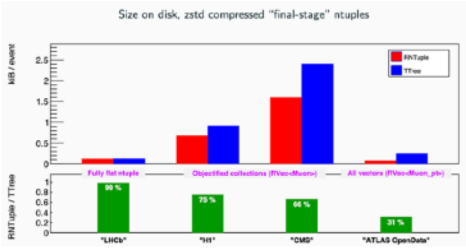
Columnar Analysis



Python ecosystem

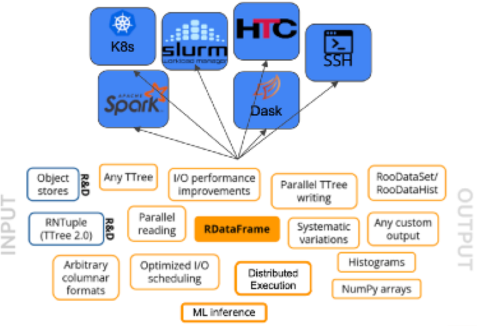
Coffea

File size



CHEP 2023

RDataFrame



WLCG-HSF2024

Analysis facility

➜ Several initiatives

IRIS-HEP Analysis Grand Challenge Link

- binned analysis, reinterpretation and end-to-end optimisation of physics analysis use cases
- includes the development of the required cyber infrastructure

Analysis Facility prototype at CERN Link

- both batch and interactive
- DASK, HT-Condor integration
- open for user tests (documentation)

Virtual Research Environment Link

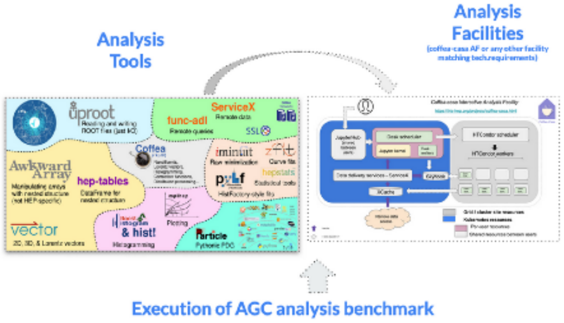
- development in EU-founded project in the line of EOSC building: ESCAPE
- open source platform with access to all the digital content needed to develop, share and reproduce an end-to-end scientific result in compliance with FAIR (findable, accessible, interoperable, reproducible) principles
- All the building blocks from storage to computing with AAI and notebook to analysis preservation

➜ Reana: reproductible research analysis platform Link

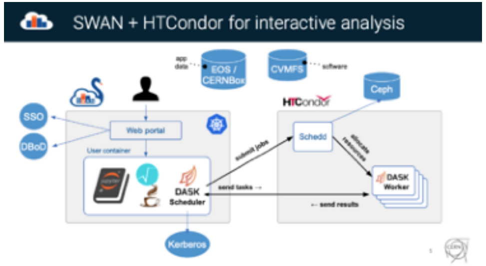
  - developed at CERN
  - allow to preserve, distribute and reproduce
  - flexible scalable and reusable => container

And more like INFN Napoli initiative Link

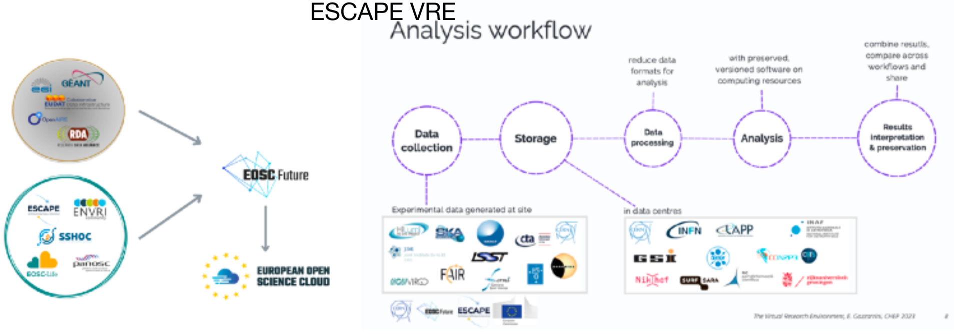
IRIS-HEP Analysis Grand Challenge



CERN facility software component



ESCAPE VRE



Sustainabiity

Sustainability

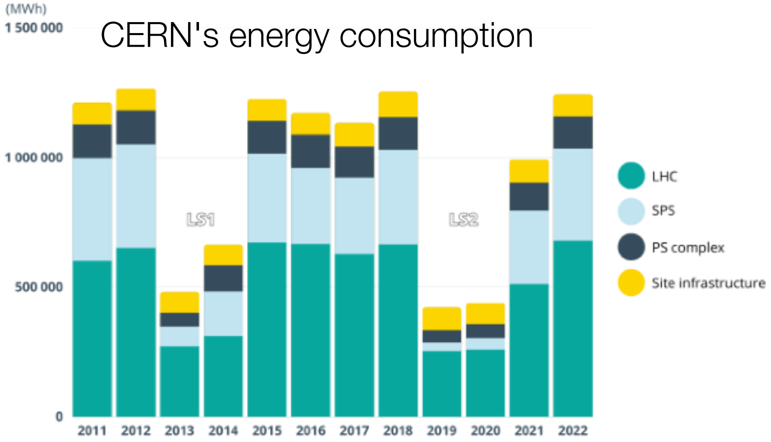
Remember: Paris agreement goal

- Zero emission around 2100 => -50% by 2030 !
- To get the 50% in 2030 you have equivalently
  - to expand CO2 free energies by a factor 12
  - to increase energy efficiency bay a factor 2
  - to save energy by a factor 2

➜ A lot of initiatives to decrease our computing carbon footprint

- Understand our current and future footprint
- Reduce everywhere as soon as we know how to do it
  - reduce use of older nodes
  - reduce CPU clock speed when bad electricity footprint
- R&D and studies

- And we are big consumers!


CERN's energy consumption

https://hse.cern/content/energy-management

- The CERN energy consumption is ~1.25TWh/year during LHC runs
- The CERN IT infrastructure contributes <5% in average
- CERN provides ~20% of the WLCG resources

CHEP 2023

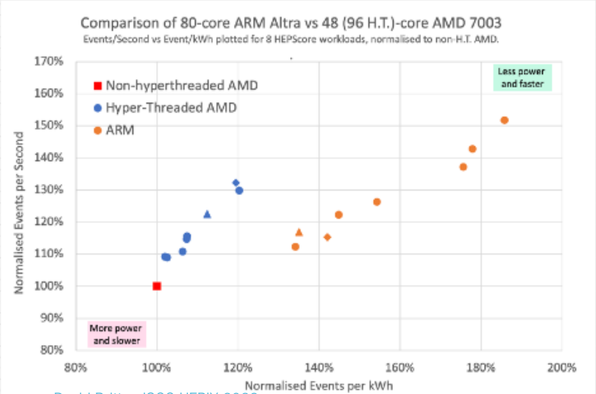
Heterogenous computing and Carbon footprint: ARM

ARM architecture

- Promising architecture with low power consumption
- some sites planning to provide resources
- => most of the LHC workloads ported to ARM. ATLAS software successfully validated and running. Ongoing validation for the 3 other LHC experiments

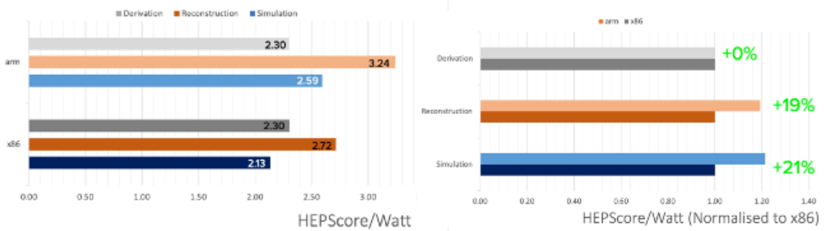
➜ In depth studies at Gridpp Glasgow site with very interesting results

ARM and AMD comparison

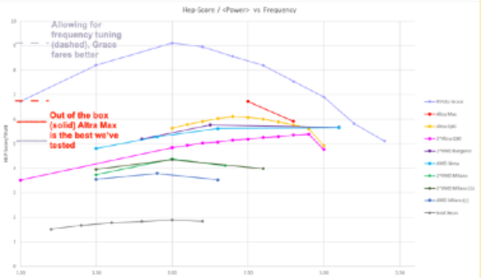


David Britton ISCG HEPIX 2023

HEPScore/Watt workflow comparison

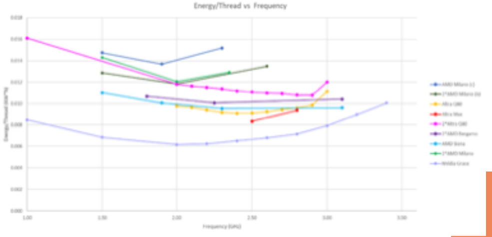


HEP-Score/Power vs frequency



Dwayne Spiteri0 HEPIX Spring WS 2024

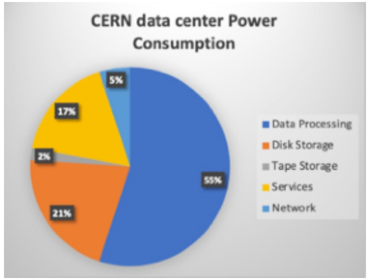
Energy/thread vs frequency



Estimating the WLCG footprint for HL-LHC

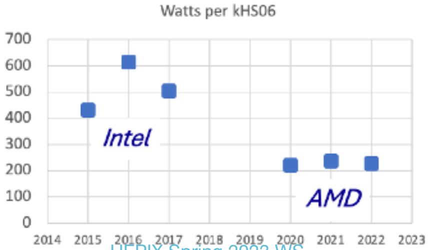
Simone Campana CHEP 2023

Distribution of power consumption



+

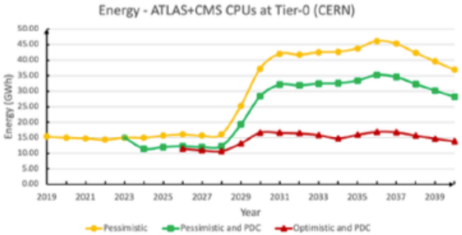
Watt/HS06



HEPIX Spring 2023 WS

+

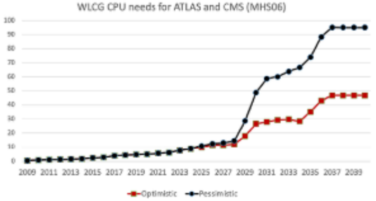
Power Usage Effectiveness (PUE) of site



➜ Effect of new computing center @Prevessin

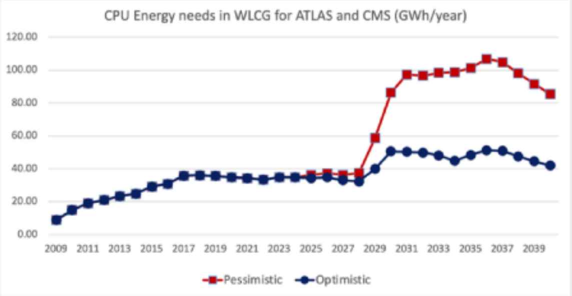
+

CPU needs

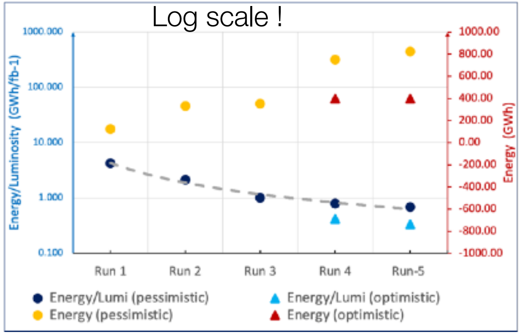


➜ Otimistic/pessimistic R&D scenarii

CPU energy needs per year



GWh and GWh/fb^{-1} estimation for WLCG



➜ Energy needs in Run-4 and Run-5:

- +100% compared to Run-2 in the pessimistic scenario
- +10% in the optimistic scenario

Human

Skilled and motivated people

Recognition

- Computing and software activities are fundamental in our research activities
  - they have huge impact on our physics results
  - their cost is similar than the one of the experiment
  - they are part of our experiments and are no more a « service »
- We need skilled and motivated people
  - Their work should be shared and publicised
  - Not only in dedicated meeting and conference !
  - Their expertise should be recognise also when physicists

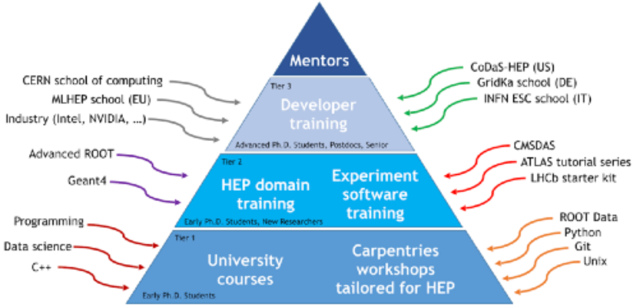
Training

- Students often lack of software and computing skills
- Computing and software are evolving quickly => continuous learning is needed

➜ tutorials, training are important and there are lot's of opportunities:

- ➜ tutorials, school and training locally, by their institution, in collaborations, organised by the HEP Software Foundation (HSF)…
- ➜ HSF as a forum to build the community and share knowledge

- Software Training in HEP



Mentors

Tier 3

Developer training

Advanced Ph.D. Students, Postdocs, Senior

Tier 2

HEP domain training

Experiment software training

Early Ph.D. Students, New Researchers

Tier 1

University courses

Carpentries workshops tailored for HEP

Early Ph.D. Students

CERN school of computing

MLHEP school (EU)

Industry (Intel, NVIDIA, ...)

CoDaS-HEP (US)

GridKa school (DE)

INFN ESC school (IT)

Advanced ROOT

Geant4

CMSDAS

ATLAS tutorial series

LHCb starter kit

Programming

Data science

C++

ROOT Data

Python

Git

Unix

M. Ballroom CHEP 2023

HSF iris hep **Building a community** HSF Training – Kilian Lieret for HSF Training

Weekly meetings

Monthly Hackathons

Platforms

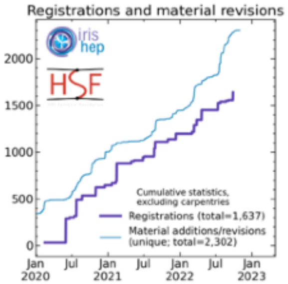
Community pages

Increasing our reach

Recognition

How-to guides

@hsftraining



Conclusion

Conclusion

➜ *Software and Computing are key elements of our experiments*

Starting with a success

- Computing for LHC run1 and 2 with a complete and complex set of software, middleware and WLCG infrastructure at the Exa-scale successfully allows to store, process and analyse LHC data, leading to wonderful scientific results

Take up the HL-LHC software and computing challenge

- Infrastructure and software will have to cope for high luminosity (200 x run 2) + pile-up up to 200 + throughput ~10TB/s
- In a context that is rapidly evolving: hardware, technologies, cost ➜ constraints and opportunities
- Taking into account Open Science and effort to reduce our carbon footprint

Huge amount of work and R&D already done and still a lot ahead

- New developments integrated in production as soon as validated
- Software, hardware and computing models need to be adapted for heterogenous ressources, more parallelisation => flexibility needed to adapt to ressources not build for and by us
- Make best use of new technologies and technics: AI, progress done outside our field
- Be prepared to the next (r)evolution like Quantum Computing
- Training is very important as proper recognition for the work done in the computing and software fields