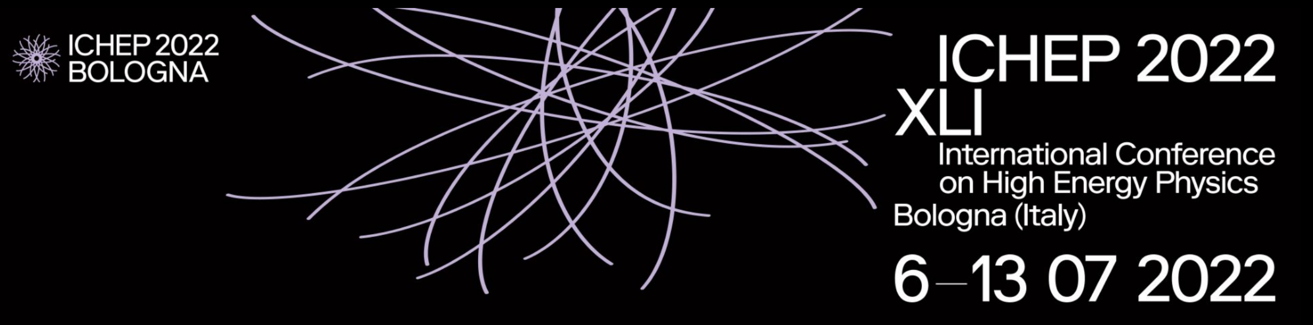
Computing infrastructure: status and outlook

Stefano Piano – INFN sez. Trieste
Computing Resource Coordinator of the ALICE collaboration

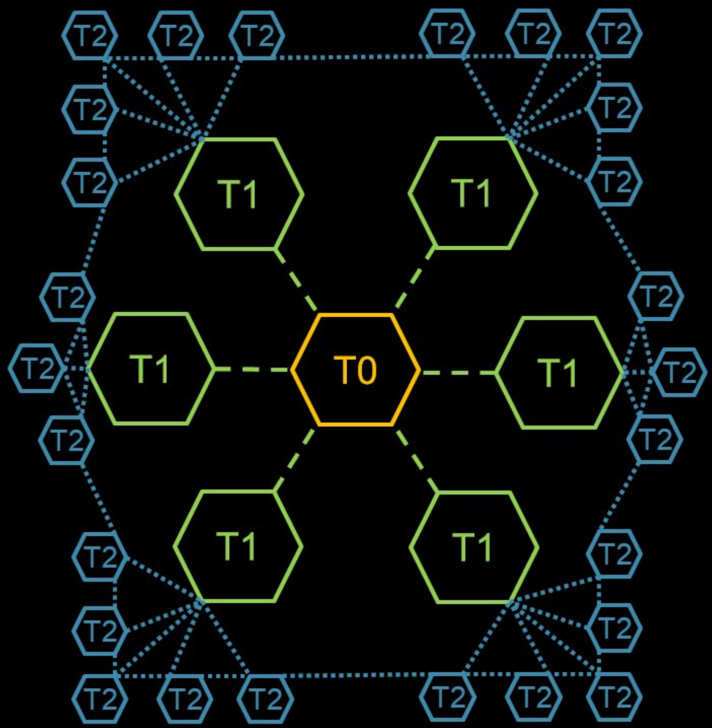


Outline

- Status of computing infrastructure:
  - Resource evolution
  - Scale of computing needs today
- Outlook:
  - Run 3: ALICE and LHCb
  - Run 4 and beyond:
    - Computing needs for HL-LHC
    - Challenges and possible solutions
  - Collaboration with other exp's and other sciences

Computing infrastructure

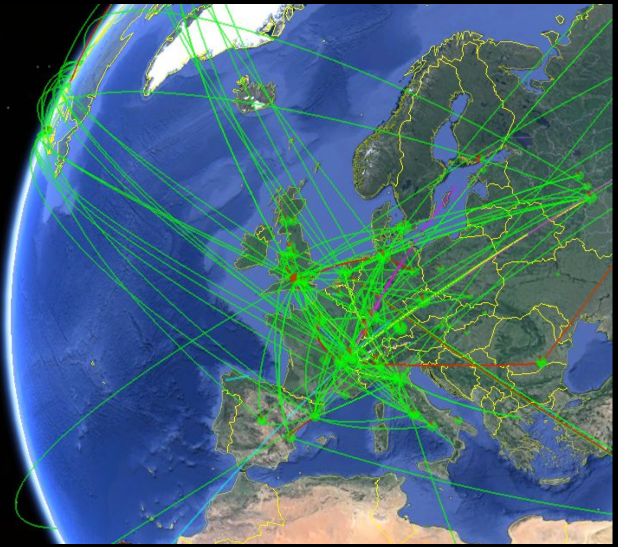
- Globally distributed system of computing centres:
  - Configured in a tiered architecture that functions as a single coherent system:
    - Tier 0 – Tier 1 – Tier 2 – Tier 3
  - Each centre provides Grid-enabled gateways to CPU and storage, which are used by VOs middleware to interact with the centres in a structured way; some of the centers also provide Analysis Facilities
  - Extensive high-quality network allows for communication among all computing centres



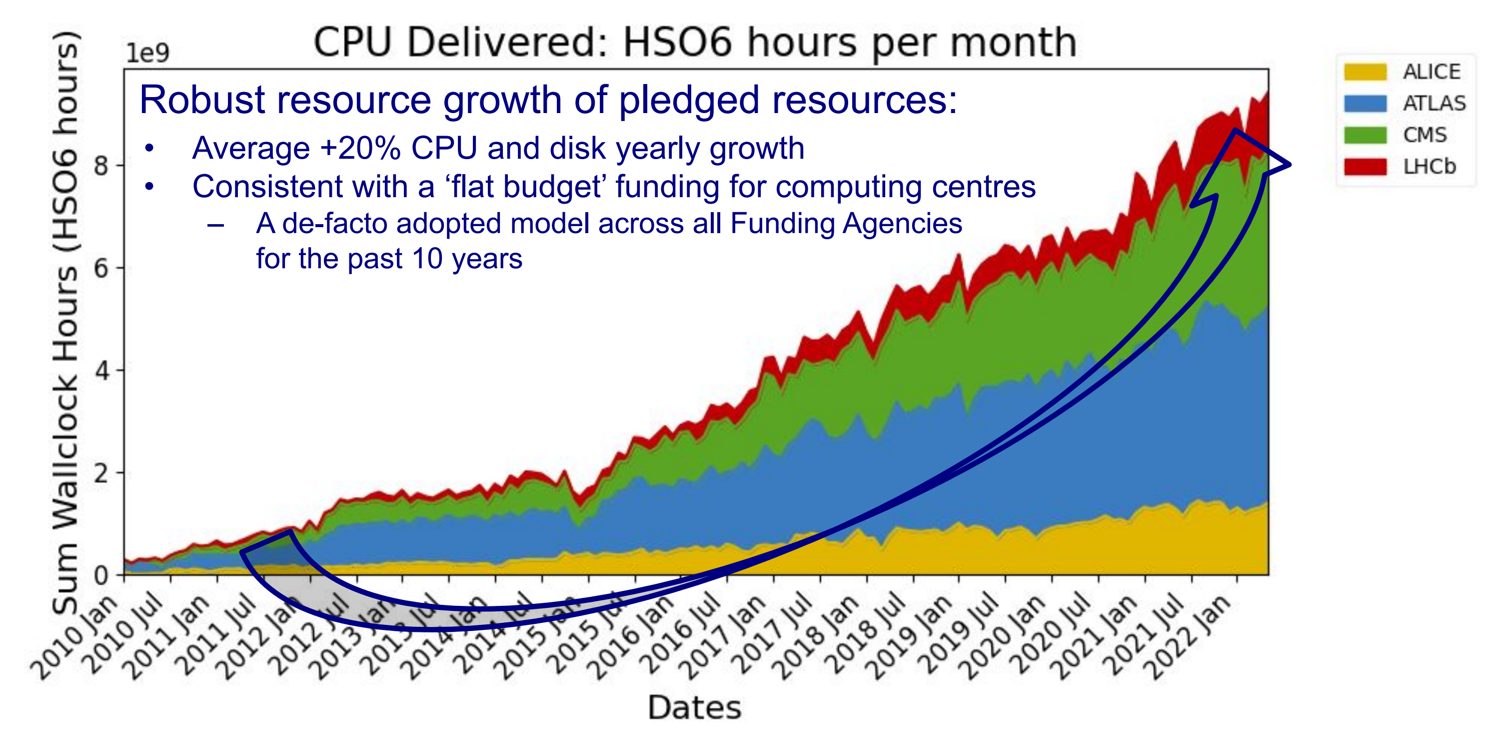
Computing infrastructure

WLCG
Worldwide LHC Computing Grid

- Globally distributed system of computing centres:
  - Configured in a tiered architecture that functions as a single coherent system:
    - Tier 0 – Tier 1 – Tier 2 – Tier 3
  - WLCG: Tier0 + 14 Tier1's + >150 Tier2's in more than 40 countries



WLCG resource evolution



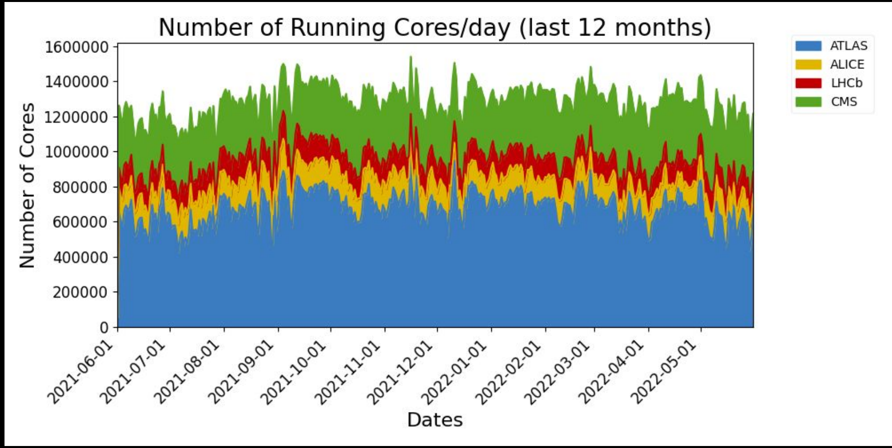
Robust resource growth of pledged resources:

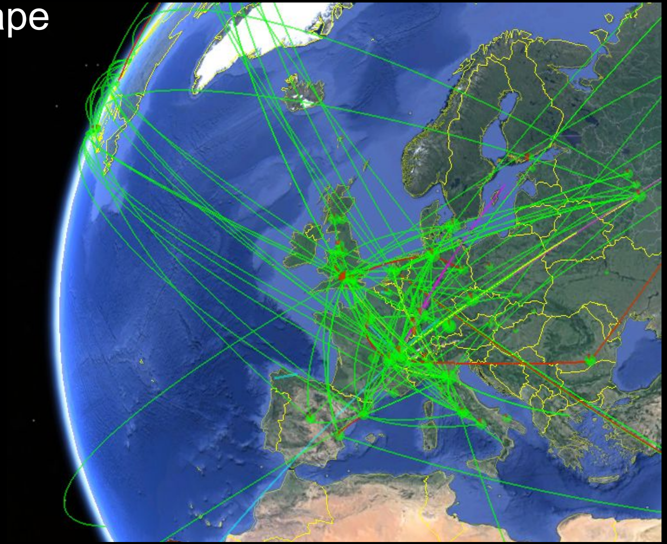
- Average +20% CPU and disk yearly growth
- Consistent with a 'flat budget' funding for computing centres
  - A de-facto adopted model across all Funding Agencies for the past 10 years

Scale of LHC computing needs today

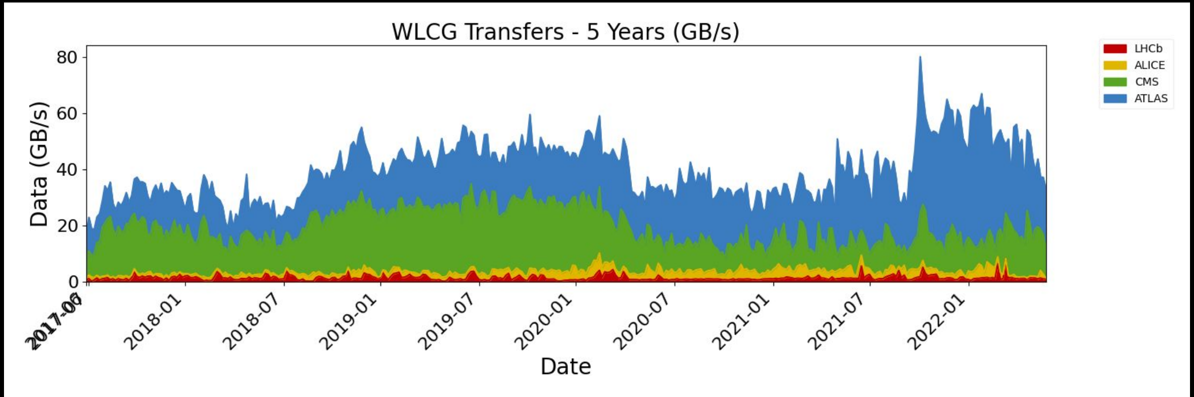
- CPU: > 1 million cores fully occupied (pledges+opportunistic resources)
- Firmly in the Exabyte-scale data:
  - 2022 pledges for all LHC exp's: 0.8 EB disk and 1.2 EB tape
  - Data ingestion tens of EB/year

WLCG
Worldwide LHC Computing Grid

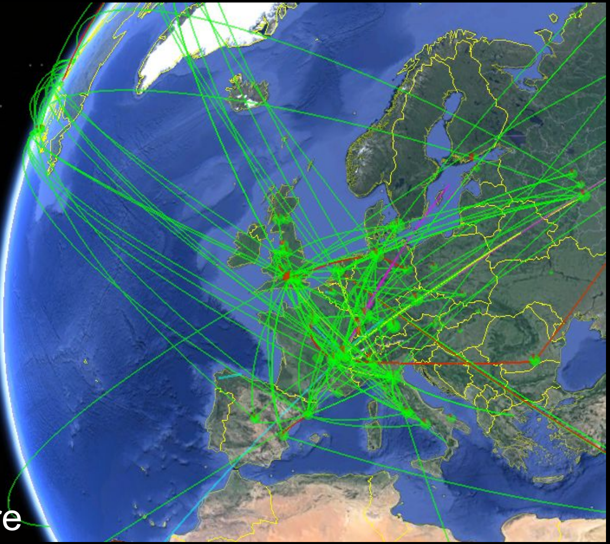




Scale of LHC computing needs today

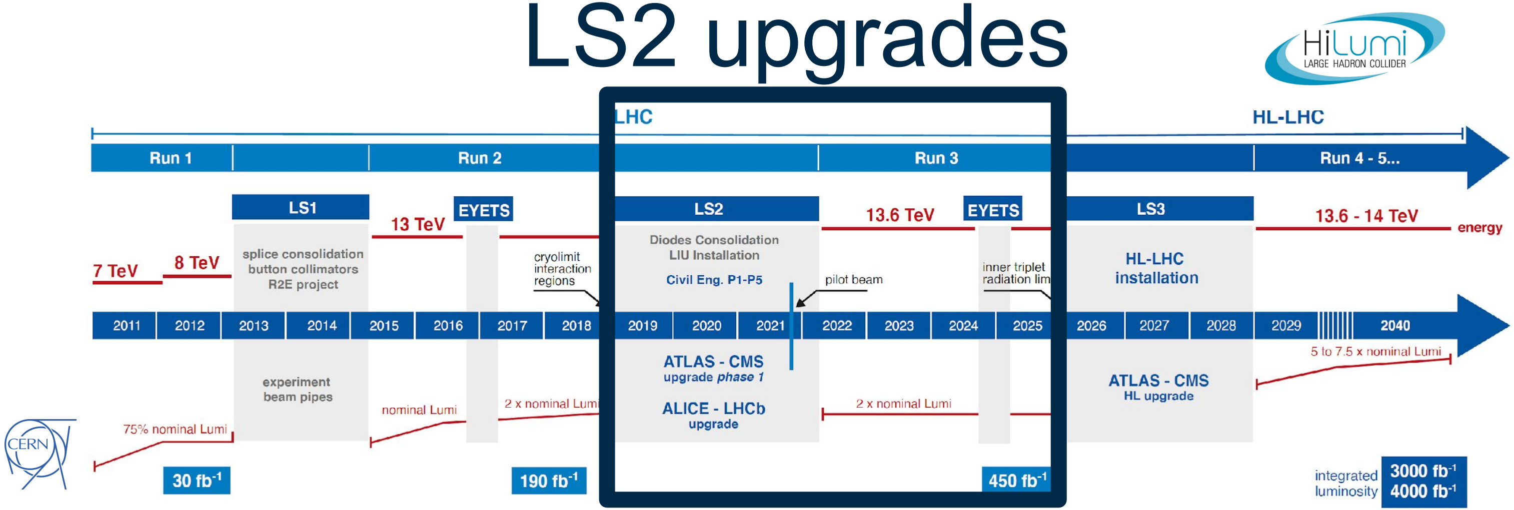


WLCG
Worldwide LHC Computing Grid



- Networking:
  - LHCOPN >1 Tbps from Tier 0 to 14 Tier1's
  - LHCOne overlay of 10-100 Gbps networks to connect
    - Tier 1 – Tier 2 – Tier 3
  - Other HEP experiments share a part of this infrastructure
    - Belle II, Dune, Pierre Auger, NovA, XENON, JUNO
  - Other sciences will use much of the same computing infrastructure

LHC/HL-LHC plan: LS2 upgrades



ALICE upgrade

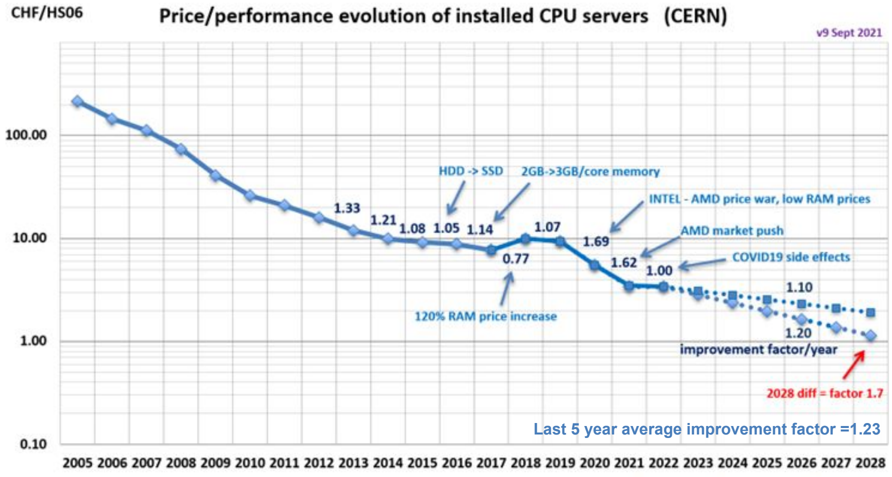
- From 1 kHz to 50 kHz Pb-Pb interaction rate
- Collect 13 nb^{-1} of Pb-Pb collisions at 5 TeV (inspected ~1 nb^{-1} during Run 1 and Run 2)
- 100x more recorded MB events wrt Run 1 and Run 2

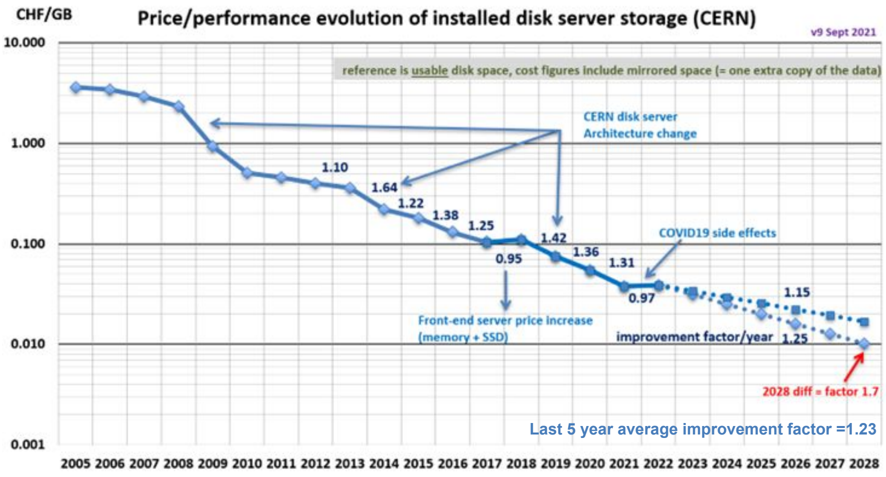
LHCb upgrade

- Raising the instantaneous luminosity by a factor five to $2 \times 10^{33} cm^{-2}s^{-1}$
- Implementing a full software trigger to overcome limitation of L0 hardware trigger
- At least 10x more recorded events

Hardware cost evolution

- Hardware cost is more and more dominated by market trends rather than technology and science has no influence on these markets
- For the same budget we have been assuming +20%/year in storage and CPU capacity, but we have to deal with large fluctuations and a flat budget constraint





S.Campana, [Computing - challenges and future directions] (ECFA 2021)

Hardware cost evolution

- Hardware cost is more and more dominated by market trends rather than technology and science has no influence on these markets
- For the same budget we have been assuming +20%/year in storage and CPU capacity, but we have to deal with large fluctuations and a flat budget constraint

- In the last years semiconductors shortage and negative trends
  - Situation improving but not back to normal before end of 2022
  - CPU CERN procurement: Q1 2022 +82% wrt 2020, 6/8 months delivery time
- The cost of energy is a major concern at several countries
  - The main impact is on CPUs (e.g. CERN T0 consumption: 80% CPU vs 20% disk)
  - In many cases, less opportunistic CPUs might be expected at the facilities

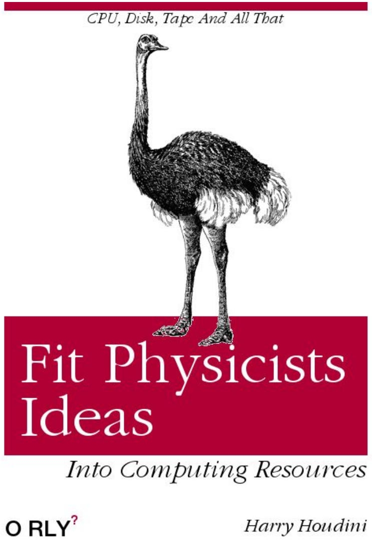
- New AMD and Intel processors this summer:
  - General improvement in energy efficiency (x2 in last 5 years)

S.Campana, Computing - challenges and future directions (ECFA 2021)

D. Fazzini, Friday
R. Münzer, Friday

A big challenge in data handling

- Projections assume constant funding every year for LHC computing
- Technology improvements will bring ~20% more resources every year i.e. computing increases by factor 5 in 10 years (“flat budget” scenario)
- ALICE @ Run 3 and 4: 100x more recorded collisions
- Need to gain factor 20 (disk and CPU) through smarter strategy and algorithms maintaining (or better improving) the physics performance
- Similar challenge for LHCb: 30x increase in throughput from the upgraded detector (10x physics event rate x factor 3 increase in average event size due to larger pile-up)
- Keep data volumes under control: aggressive compression (ALICE), selective persistence (LHCb), optimized data formats
- Simulation and reconstruction optimization (see next talk!)

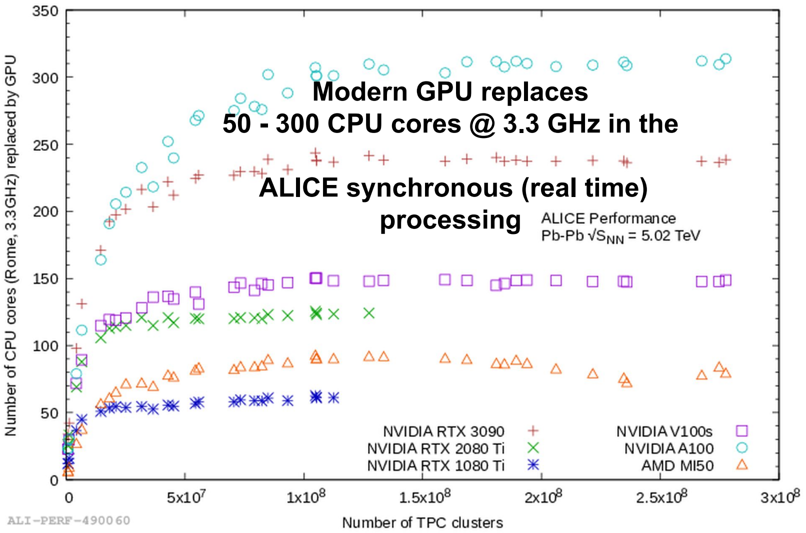


Courtesy of LHCb collaboration

C.Bozzi, Software e computing in LHCb: la sfida di Run3 (e oltre) (seminar @ CNAF 2021)
S.Piano, The ALICE O² computing model for Run 3 and 4 (seminar @ CNAF 2021)

Heterogeneous architectures

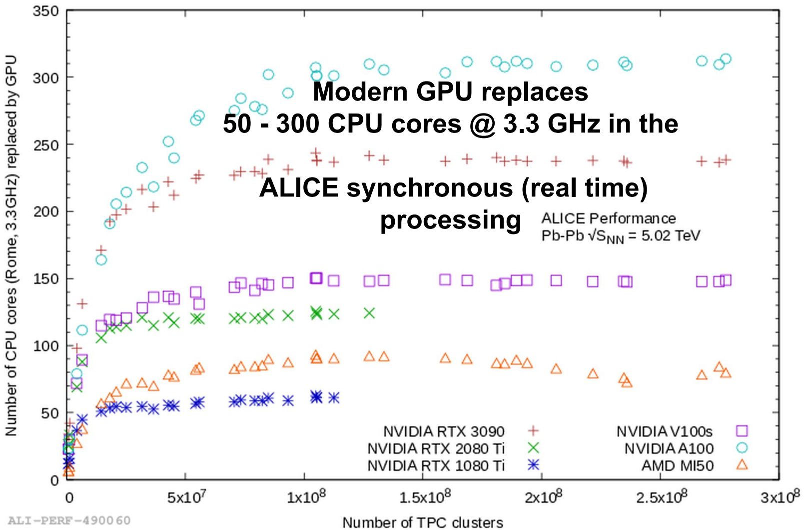
- Heterogeneous architectures: complementing CPU capacity with accelerators (e.g. GPUs)
  - GPUs offer more theoretical FLOPS in a compact package
  - Lower cost than CPUs per theoretical FLOPS
- Playing a fundamental role in Run 3 already, in most online systems. Non exhaustive examples:
- **ALICE:** Speed up from GPU usage + from algorithmic improvements + tuning on CPUs
  - porting of asynchronous (offline) reconstruction code to GPUs well advanced thanks to common online-offline framework
- **LHCb:** exploitation of heterogeneous architectures, thanks to Allen framework:
  - for partial reconstruction in Run 3 (HLT1)
- **CMS:** Patatrack Pixel Reco + ECAL and HCAL:
  - 30% of the online Run 3 reconstruction is offloaded to GPUs



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

Heterogeneous architectures

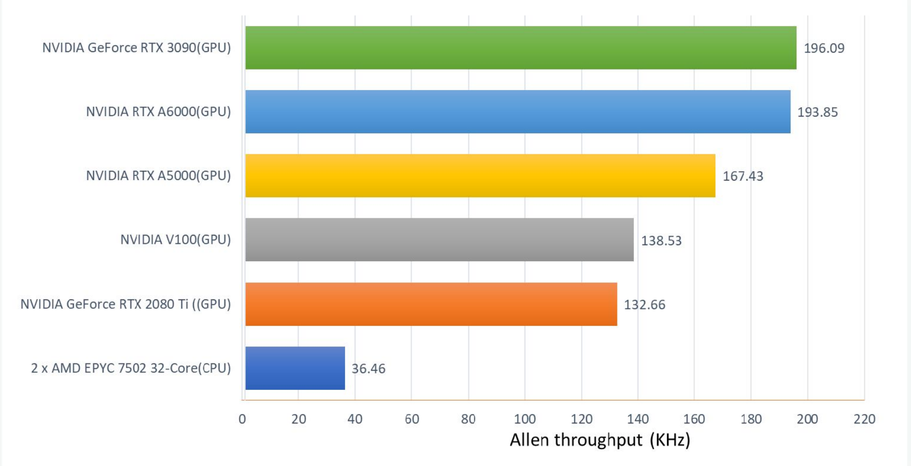
- Heterogeneous architectures: complementing CPU capacity with accelerators (e.g. GPUs)
  - GPUs offer more theoretical FLOPS in a compact package
  - Lower cost than CPUs per theoretical FLOPS
- Playing a fundamental role in Run 3 already, in most online systems. Non exhaustive examples:
- **ALICE:**
  - Without GPU 1800 Event Processing Nodes:
    - 2 CPUs x 32 cores per EPN (115 kcores)
  - With GPU 250 Event Processing Nodes
    - 2 CPUs x 32 cores + 8 GPUs per EPN
  - GPU based solution strong impact on hardware and operating cost savings



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

Heterogeneous architectures

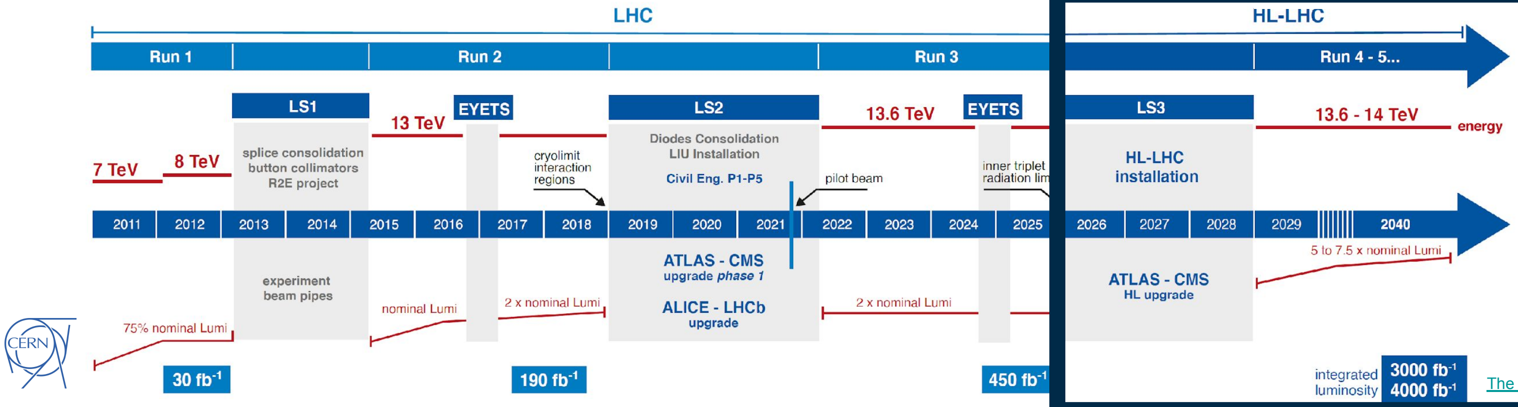
- Heterogeneous architectures: complementing CPU capacity with accelerators (e.g. GPUs)
  - GPUs offer more theoretical FLOPS in a compact package
  - Lower cost than CPUs per theoretical FLOPS
- Playing a fundamental role in Run-3 already, in most online systems. Non exhaustive examples:
- **ALICE:**
  - Without GPU 1800 Event Processing Nodes:
    - 2 CPUs x 32 cores per EPN
  - With GPU 250 Event Processing Nodes
    - 2 CPUs x 32 cores + 8 GPUs per EPN
  - GPU based solution strong impact on hardware and operating cost savings
- **LHCb:**
  - Full detector read-out at 40 MHz (visible: 30MHz)
  - HLT1 running on GPUs on ~170 EB servers:
    - Cost savings: less EB servers and no need for high-speed network from EB to HLT2 farm
  - GPU: more opportunities for future performance gain



LHCb-FIGURE-2022-010

LHC/HL-LHC plan: Run 4 and beyond

V. Cavaliere, Friday
T. Tomei, Friday
B. Radburn-Smith, Saturday



The HL-LHC project

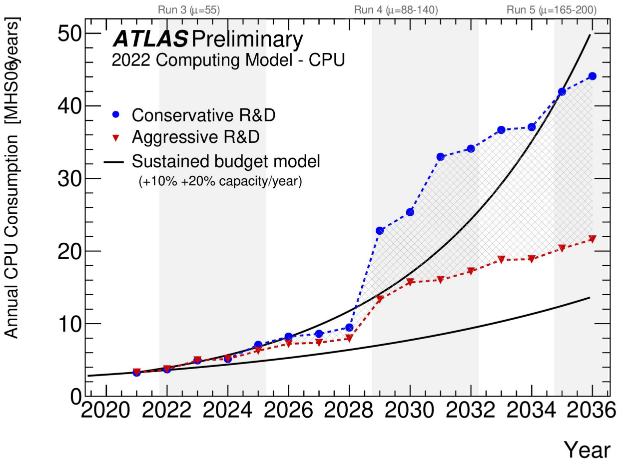
HL-LHC project

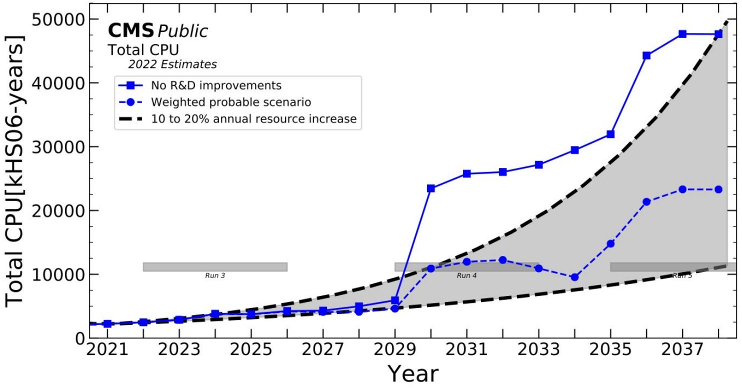
- Expected to be operational from 2029
- Objective is to increase the integrated luminosity by a factor of 10 beyond the LHC's design value
- During Run 1/2 ~200 fb^{-1} delivered to both ATLAS and CMS, expected to provide 4000 fb^{-1} by 2040 (> 20x)

ATLAS and CMS HL upgrades

- ATLAS and CMS are facing the challenge of an instantaneous luminosity increase from $2x10^{34}$ $cm^{-2}s^{-1}$ (end of Run 2) to $7.5x10^{34}$ $cm^{-2}s^{-1}$ (Run 5)
- While for ALICE and LHCb Run 4 will be at a similar scale than Run 3. New plans for Run 5 and beyond.

Expected CPU needs for HL-LHC



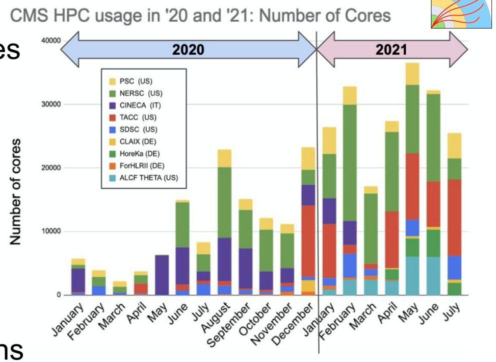


- The gap between available and needed resources can be filled up, assuming the main R&D activities are successful. There are still large uncertainties
- Investing in person power now is crucial to ensure we will be ready for HL-LHC
- Investing in hardware must be done in close cooperation with the R&Ds

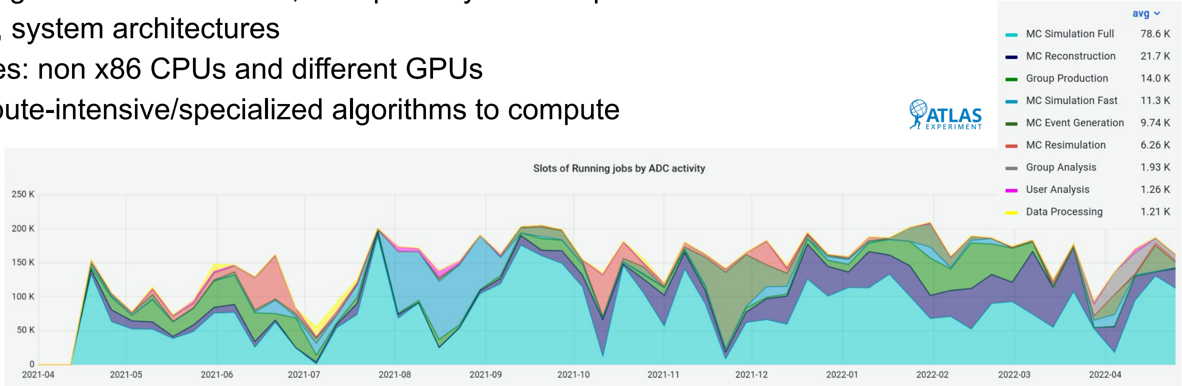
ATLAS Collaboration, Computing and Software - Public Results
CMS Collaboration, Offline and Computing Public Results

High Performance Computing

- A welcome addition of resources comes from opportunistic CPUs
  - HPC centers: unique opportunity - substantial investment of national and international entities
  - HEP engagement with DOE & NSF in USA and with PRACE and EuroHPC in Europe
  - In 2021: CMS used 10x more capacity at HPC sites wrt 2019,
  - and ATLAS exploited 2.2 MHS06 HPC CPU capacity



- Accessing and using resources at HPC centers comes with different challenges:
  - Need to incorporate HPCs keeping integration costs in check, transparently for the operations
  - Diversity in access and usage policies, system architectures
  - Heterogeneous computing architectures: non x86 CPUs and different GPUs
  - CPU remains central, but offload compute-intensive/specialized algorithms to compute accelerators
  - Opportunistic CPU usage but no opportunistic disk
- HPC comes with a cost
  - Requires significant investment in training and development
  - Much harder for smaller collaboration



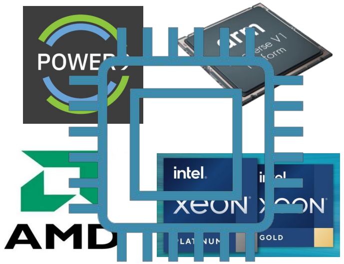
Harnessing a supercomputer for ATLAS (CERN NEWS 2022)

Heterogeneous Computing

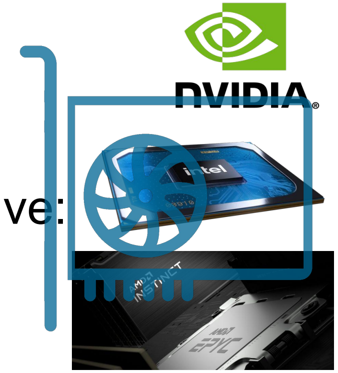
GPUs:
Several talks on ML,DL,(G)NN,AI
F. Legger, Saturday

Generators with GPUs:
A. Valassi, Friday
E. Bothman, Saturday

FPGA:
M. Lorusso, Friday
K. Tadome, Friday



- Today we use opportunistically some types of computing system, in particular HPC systems, and HLT
- In future, this heterogeneity will expand; we must be able to make use of all types: Non-x86 architectures, GPUs, HPCs, clouds, HLT farms, FPGA?
- Requires:
  - Common provisioning mechanisms, transparent to users
  - Facilities able to control access (cost), efficient use
- HPC storage is transient, cloud storage is still prohibitively expensive:
  - Must be able to deliver data to them when they are in active use
  - Data delivery will become crucial!



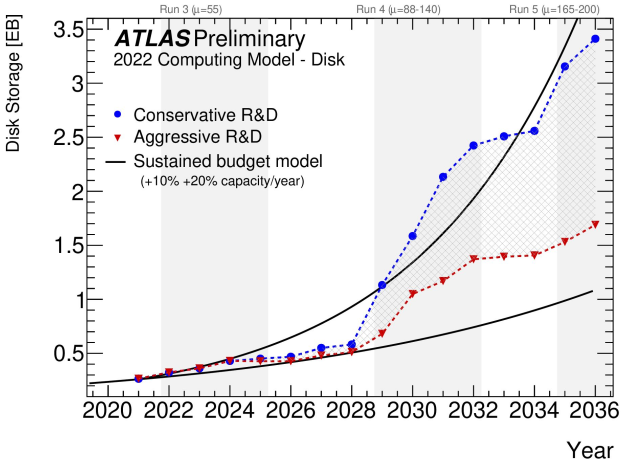
W. Guan, Friday
L. Morganti, Saturday

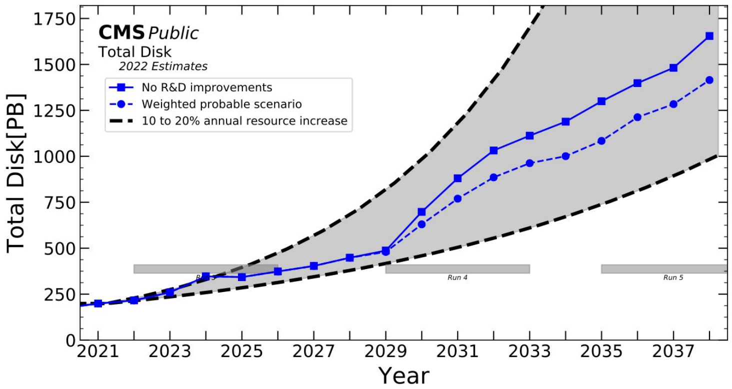
Infrastructure and services sustainability

- The HL-LHC challenge is not just about resources
- The sustainability of the infrastructure over the next 15 years is also challenging
- WLCG strategy toward HL-LHC:
  - Focused activity towards tackling the data challenge within DOMA (Data Organisation, Management & Access) R&D programme:
    - Progress towards a more flexible model integrating a heterogeneous landscape of facilities, new distributed storage and content delivery mechanisms, integration of HPC storage, commercial cloud storage
    - Authentication and Authorization Infrastructure update to industry-standard tokens
    - Network R&D activities, data transfer optimization, new data transfer protocols and mechanisms, monitoring

S.Campana, Computing - challenges and future directions (ECFA 2021)

Expected disk needs for HL-LHC

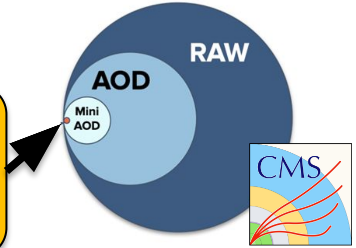




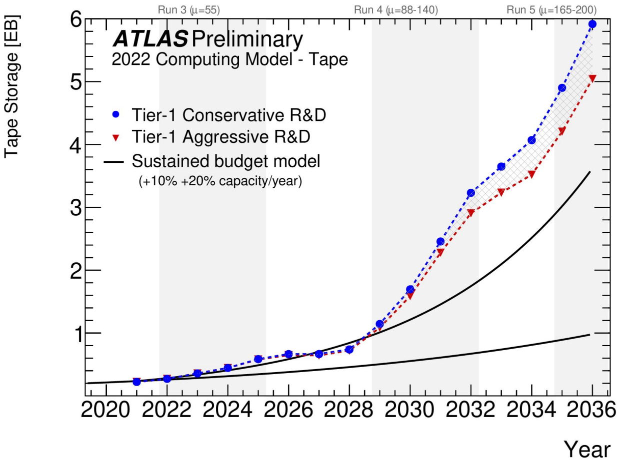
- Disk storage needs are dominated by the amount of reconstructed data, in different formats and versions. The strategy focuses on
  - Reduced analysis formats
    - Larger formats will not be generally needed on disk
  - Active use of tape

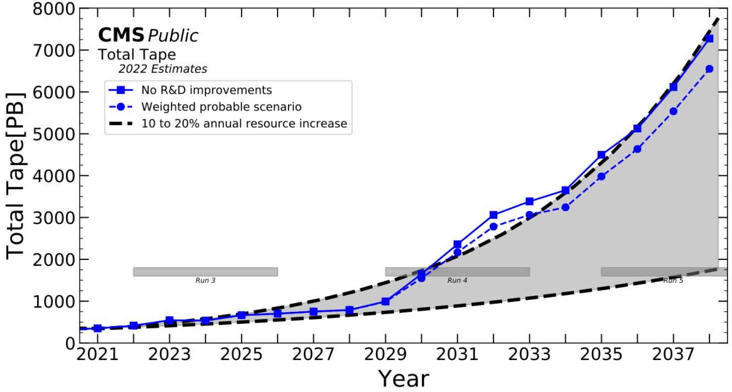
ATLAS Collaboration, [Computing and Software - Public Results]
CMS Collaboration, [Offline and Computing Public Results]

NanoAOD:
- 1 kB in Run 3
- <5 kB in HL-LHC



Expected archive storage needs for HL-LHC



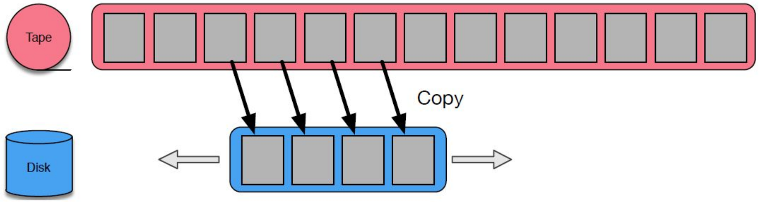


- Archive storage needs (tape) for RAW, AOD and MC data
  - Possible gains with compression/suppression, but moderate
  - Be prepared to invest in the needed tape volume and optimize the rest
- Management of exabyte-scale data with common tools:
  - Orchestration of common infrastructure

S.Campana, Computing - challenges and future directions (ECFA 2021)
ATLAS Collaboration, Computing and Software - Public Results
CMS Collaboration, Offline and Computing Public Results

Data Carousel

- Keep data (AOD) on tape with disk used as an operational cache

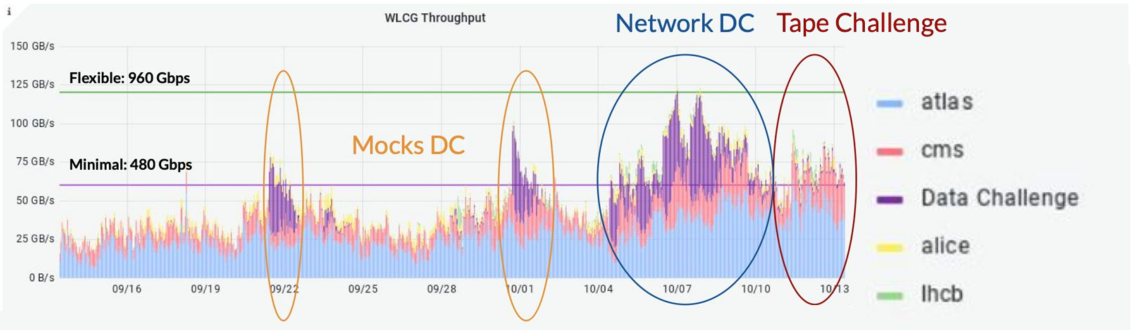


- Leverage the lower cost of tape MEDIA w.r.t. disk
- Rely on the capability of tape systems to recall data fast enough
- Tape in infrastructure is not just an archive media. Data is recalled regularly for further processing
- Organized and marshalled activities, nothing to do with the old use of hierarchical mass storages
- Active use of tape will play an even more important role in HL-LHC
- Tape bandwidth has also a cost that needs to be monitored and optimized

ATLAS Collaboration, The ATLAS Data Carousel Project Status (vCHEP2021)

Expected network needs for HL-LHC

- ATLAS and CMS will produce 350 PB of RAW data per year each
  - To be exported in ~ real time from CERN to the T1s
  - +200 Gbps for ALICE and LHCb
  - 4.8 Tbps minimal scenario (9.6 flexible) from CERN to T1s by the time of HL-LHC
- To be reprocessed in ~3 months largely at the T2s
  - 400 Gbps/experiment target for T1 to T2s traffic
- WLCG Data Challenges: incremental process toward HL-LHC, through a regular dialog between the network providers, the experiments and the facilities:
  - 10% of the target in 2021: matched the Run 3 needs
  - 30% in 2023
  - 60% in 2025
  - 100% in 2027



A. Held, Saturday
I. Cruceru, Friday

Analysis Facilities

- Run 3 AFs: high-throughput grid sites aiming at testing and tuning of the analysis tasks
  - Fast analysis of limited statistics data samples
  - ALICE plans to offload ~10% of the analysis workload from Grid to AFs in Run 3
- But … analysis needs for HL-LHC will require an (r)evolution in our AFs and tools
- Several community efforts among exp's and several projects (HSF/IRIS-HEP)
- 3 types of AFs are emerging
  - Large sites with all data local, CERN Lake:
    - SWAN: interactive, Jupyter notebooks + batch system + GPU
    - Accessible by all the systems: EOS storage (data, user output), CVMFS (software)
  - Remote access of storage providers, ESCAPE Data Lake:
    - Federated data infrastructure, concepts evolved from DOMA
    - Data Lake as a Service: Jupyterhub + rucio plugin + scratch space
  - Distributed storage among set of sites with caches, US ATLAS and US CMS
    - Three shared Tier 3 computing spaces at BNL, SLAC, and UChicago; prototype at Tier 2 Nebraska
    - All users are allowed access to the facilities
    - Tools specific for user analysis shared via CVMFS, data caching via Xcaches at each site

A. Forti, Impact on Analysis Facilities in the context of DOMA evolution (Analysis Ecosystems Workshop II 2022)

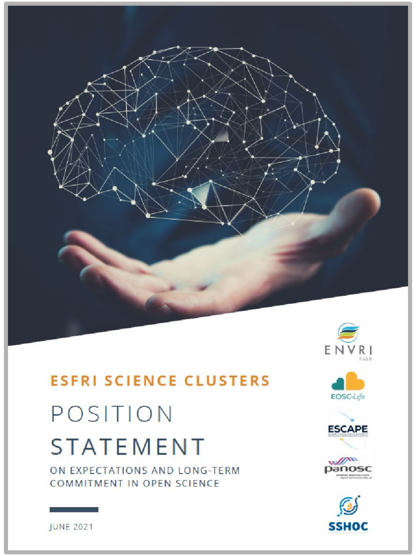
Sustainability through collaboration


CERN COURIER

Aug 11, 2017

SKA and CERN co-operate on extreme computing

Big-data co-operation agreement



The Belle-2 and DUNE HEP experiments leverage the same infrastructure as WLCG (and they are “associate members”)

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

Belle II

Evolution of Scientific Computing in the next decade: HEP and beyond

WLCG Overview Board

17th December 2018

Contact: Ian Bird (Ian.Bird@cern.ch), Simone Campana (Simone.Campana@cern.ch)

Astroparticle Physics European Consortium (APPEC)

APPEC

APPEC Contribution to the European Particle Physics Strategy

December 17, 2018

Editorial Board:
S. Katsanevas, A. Masiero, T. Montaruli, J. de Kleuver, A. Haungs

Contact Person:
T. Montaruli (APPEC Chair from Jan. 1, 2019)
Email: teresa.montaruli@unige.ch

Website: http://www.appec.org

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

EOSC-Life

S.Campana, [Computing - challenges and future directions] (ECFA 2021)

Summary

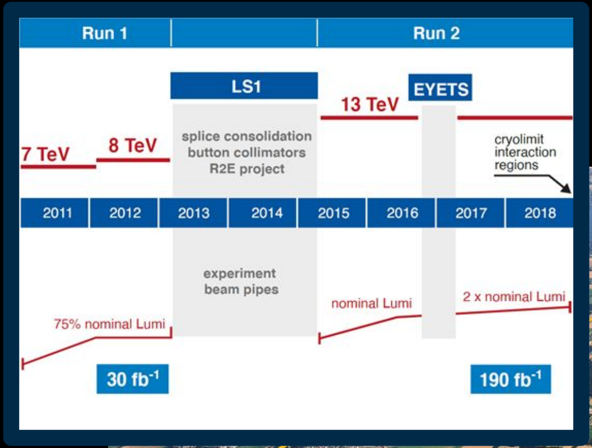
- LHC computing was very successful in providing the global environment for HEP physics
  - The resources growth was robust (+20%) as well as their use
  - Worked extremely well during Run 1 and Run 2, solid foundation for Run 3 and HL-LHC
- HL-LHC presents major challenges for LHC computing
  - Management and analysis of exabyte-scale data
  - Keeping the computing needs within the fixed flat investment
- How can these challenges be overcome?
  - Fully exploit the features offered by modern HW architectures
    - Execution of codes and validation of outputs across various compute resources
  - Towards a more flexible and sustainable infrastructure
  - Synergies and collaborations across scientific disciplines and with Industry partners
- Getting performant software and computing infrastructure requires significant investment in programming and computing skills
  - Training, sustained support and career paths for computing experts

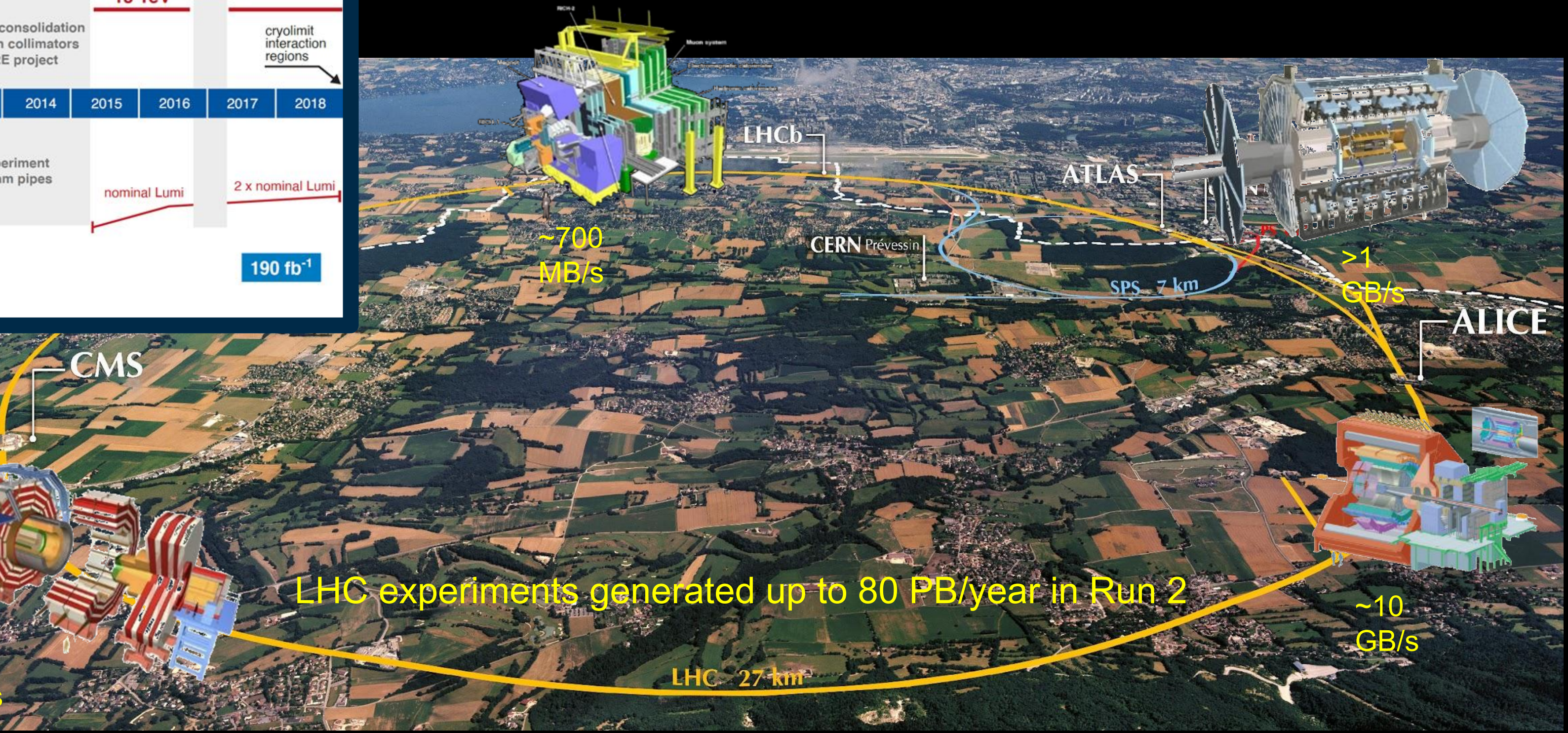
Acknowledgments

Special thanks to those who provided material and suggestions for these slides:

- Simone Campana (WLCG project leader)
- Mario Lassnig and Christoph Wissing (DOMA co-coordinators)
- Alessandro Di Girolamo and Zach Marshall (ATLAS CC’s)
- James Letts and Danilo Piparo (CMS CC’s)
- Concezio Bozzi and Ben Couturier (LHCb CC)
- Latchezar Betev (ALICE Grid Op’s)

Run 2 data volume





I.Bird, WLCG: new challanges and collaboration with other science projects (LHCOPN/LHCONE 2020 workshop)

W. Guan, Friday
L. Morganti, Saturday

R&D on data delivery

- R&D programmes in place to:
  - Look at the present data flows
  - Try to understand where the actual deficiencies are
  - At the same time, natural improvements in software that allow to try out more flexible computing models

- Objectives are:
  - Minimisation of data required to travel over the network
  - Offloading of data from “expensive” storage to “cheap” storage

- Under the constraint to keep the current processing throughput or in the best case even improve it