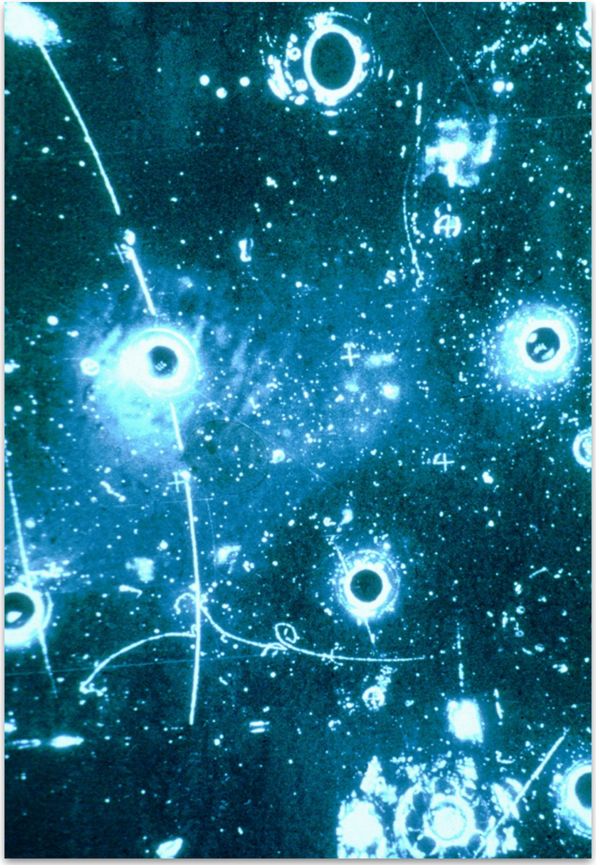
Preparing for the HL-LHC computational challenge

D. Piparo, CERN - tCSC 2024

Outline

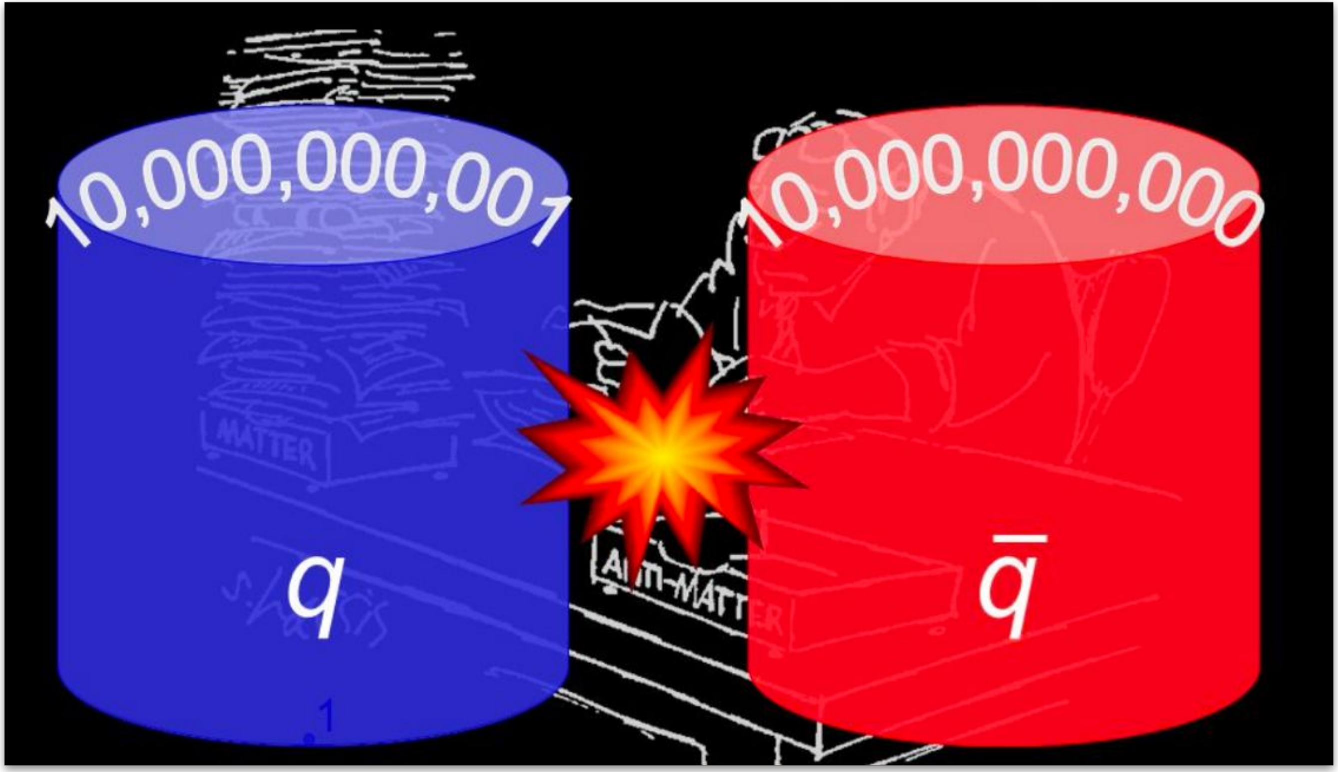
In this talk we will introduce some basic concepts related to HEP data processing and analysis workflows, seeing them in action in the context of LHC experiments. We'll also talk about the evolution of the LHC accelerator and experiments. We'll characterise at a high level what are the consequences of those upgrades for the HEP data processing software, in particular in the context of an evolving hardware and computing infrastructure.

Understanding the Universe



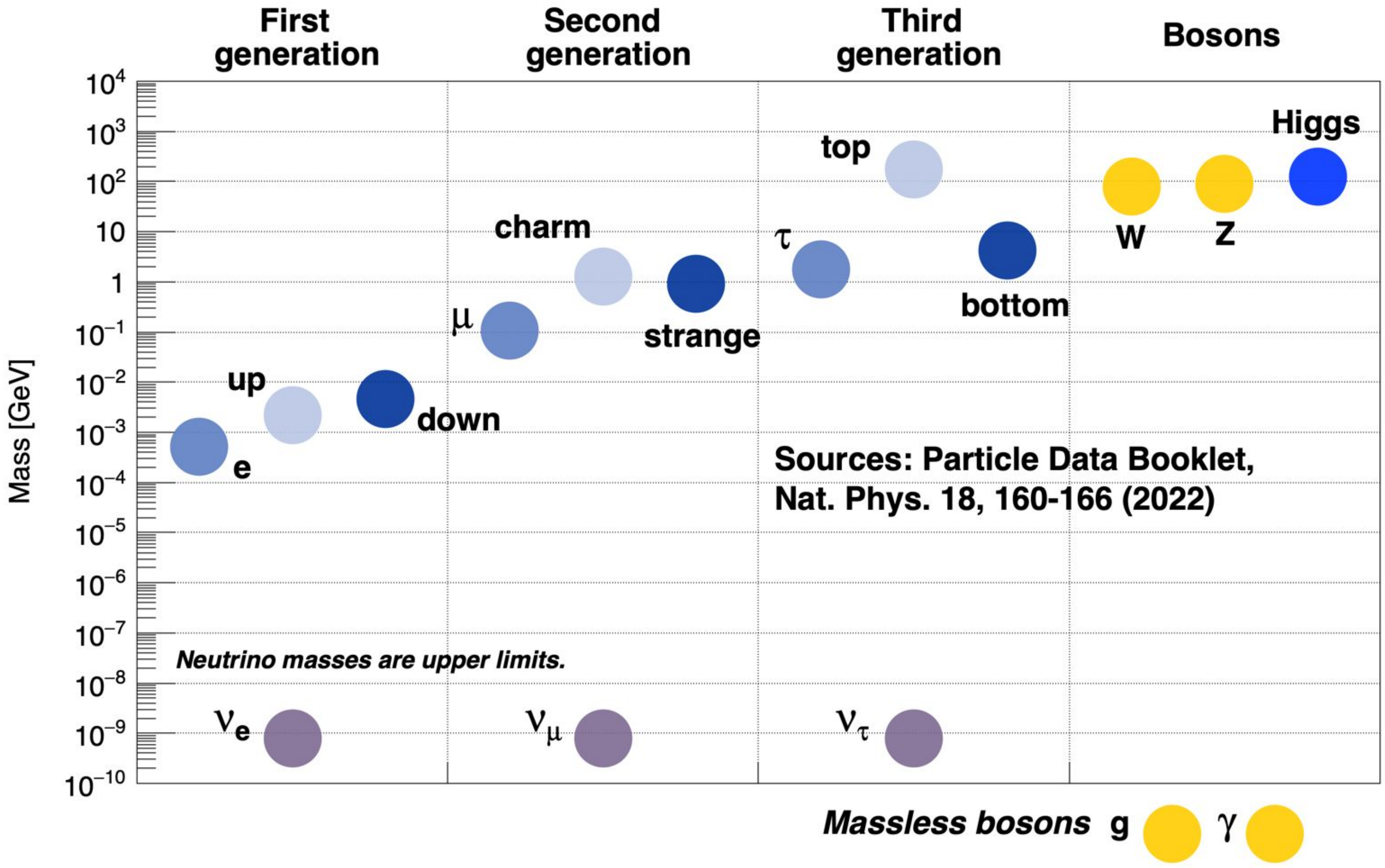
http://cds.cern.ch/record/39468

Some Unanswered Questions

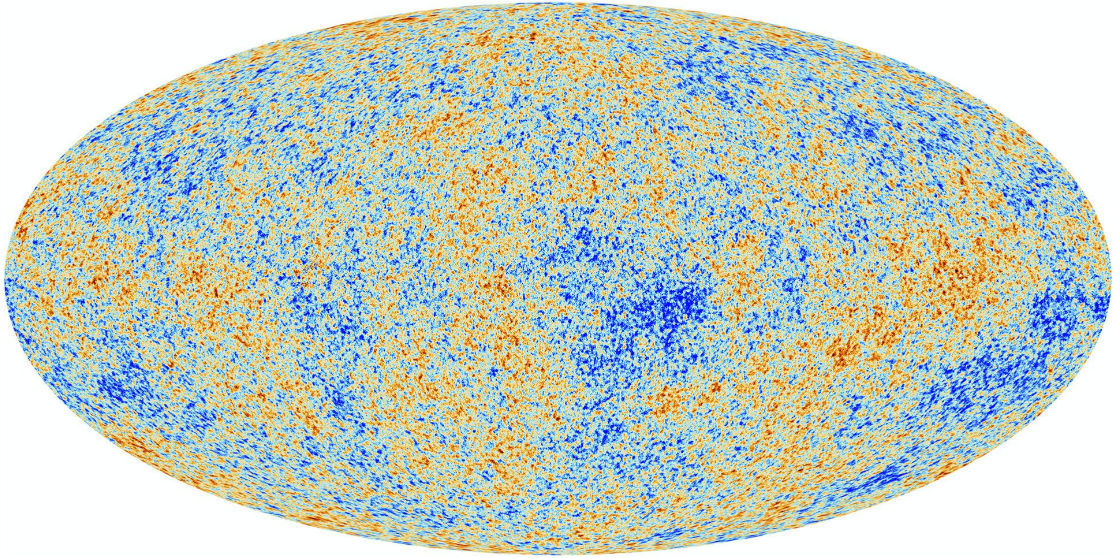


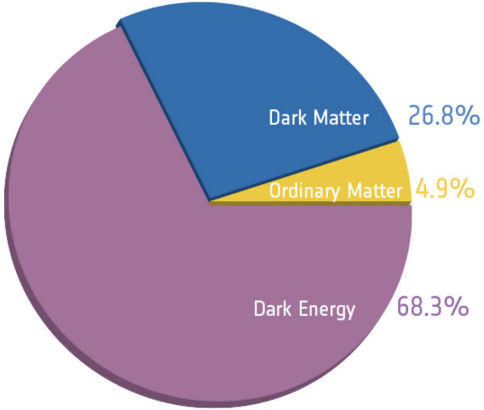
A. Hoecker, https://indico.cern.ch/event/34851/

Some Unanswered Questions



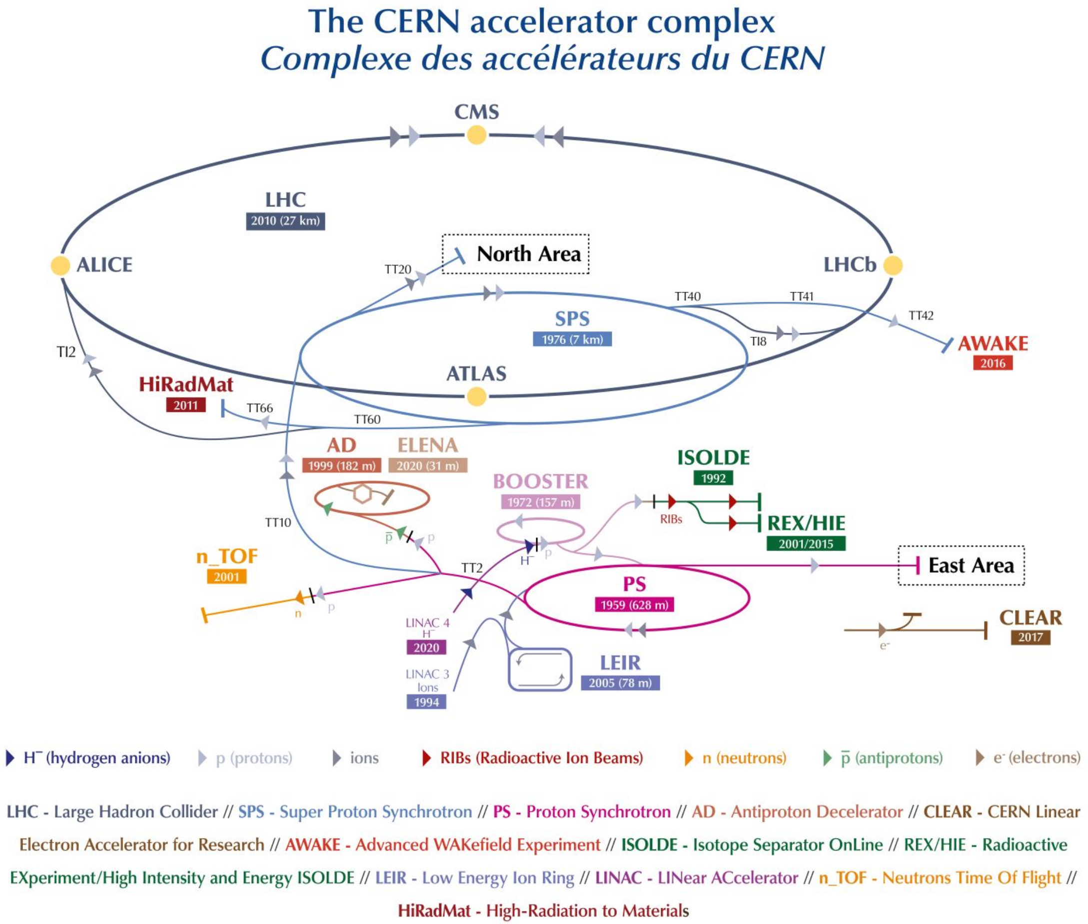
Some Unanswered Questions





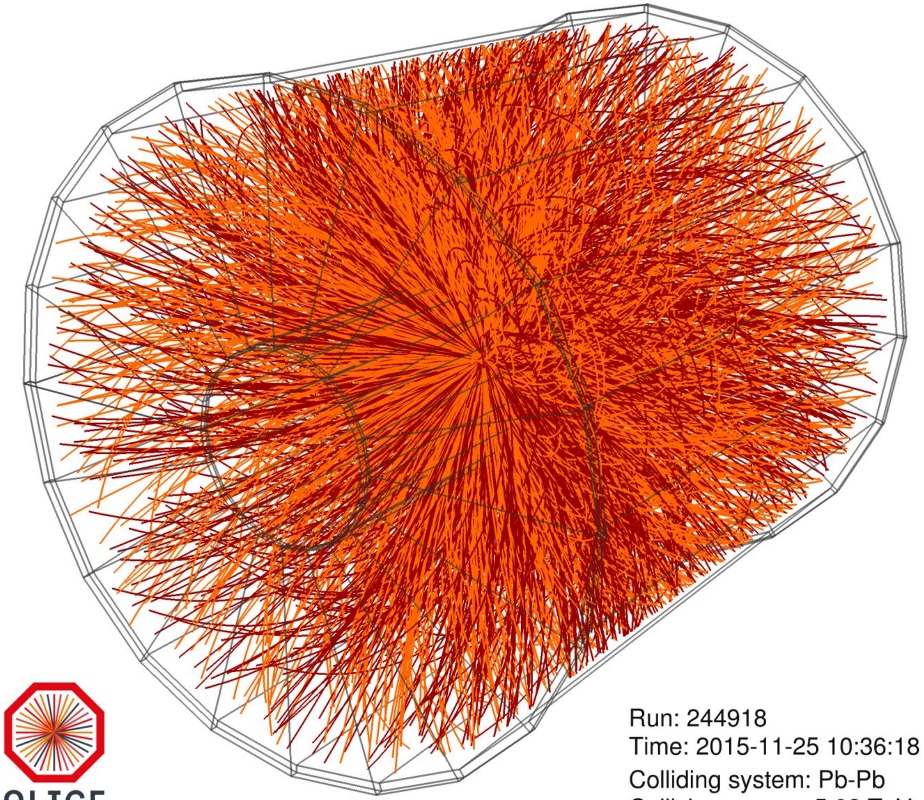
http://www.esa.int/Science_Exploration/Space_Science/Planck/Planck_reveals_an_almost_perfect_Universe

Accelerate

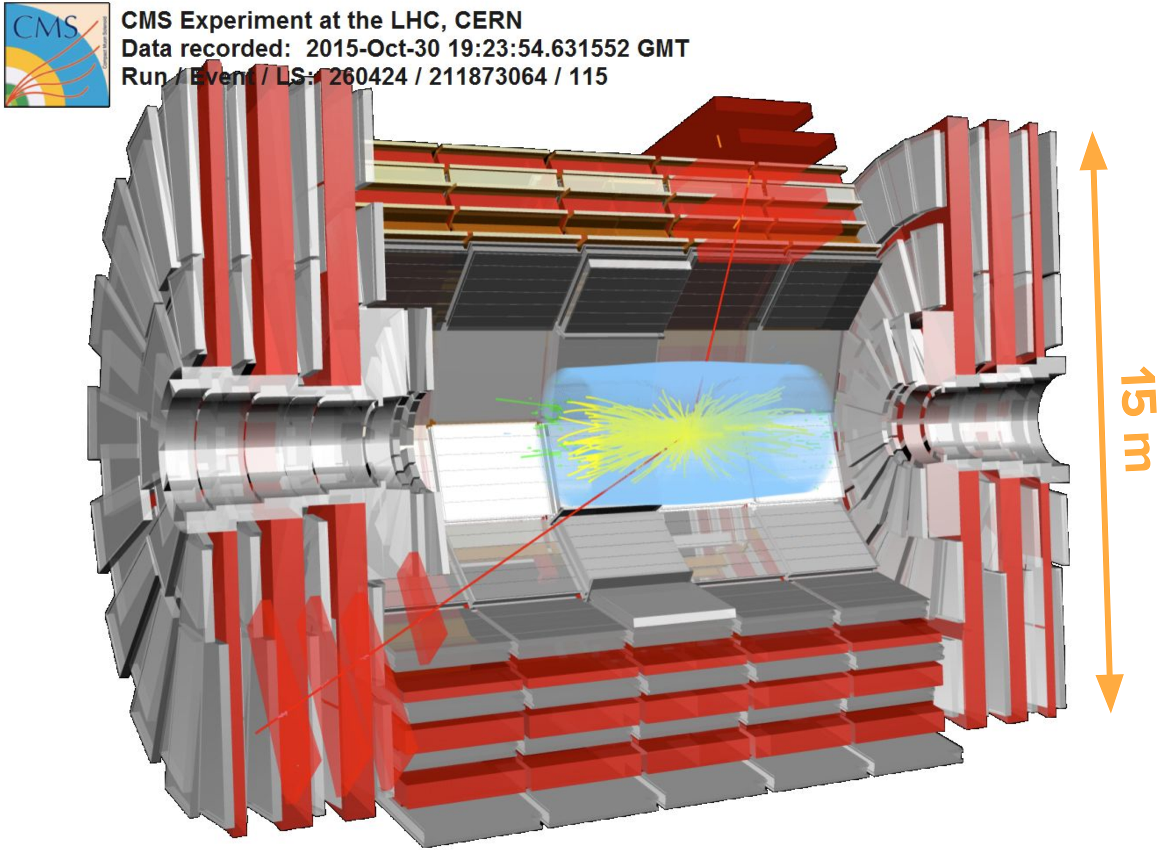


The CERN accelerator complex
Complexe des accélérateurs du CERN
CMS
LHC
2010 (27 km)
ALICE
North Area
LHCb
TT20
TT40
TT41
TT42
SPS
1976 (7 km)
TI8
AWAKE
2016
TI2
HiRadMat
2011
TT66
ATLAS
TT60
AD
1999 (182 m)
ELENA
2020 (31 m)
ISOLDE
1992
BOOSTER
1972 (157 m)
RIBs
REX/HIE
2001/2015
TT10
East Area
n_TOF
2001
TT2
PS
1959 (628 m)
CLEAR
2017
LINAC 4
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE - Radioactive EXperiment/High Intensity and Energy ISOLDE // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials

Collide

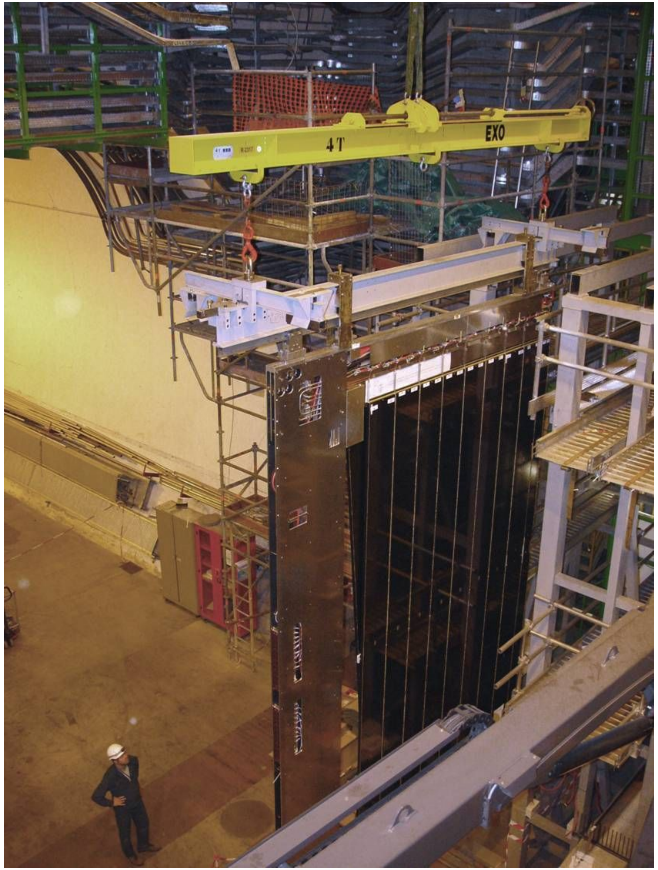

Run: 244918
Time: 2015-11-25 10:36:18
Colliding system: Pb-Pb
Collision energy: 5.02 TeV

Detect



CMS
CMS Experiment at the LHC, CERN
Data recorded: 2015-Oct-30 19:23:54.631552 GMT
Run / Event / LS: 260424 / 211873064 / 115
15 m

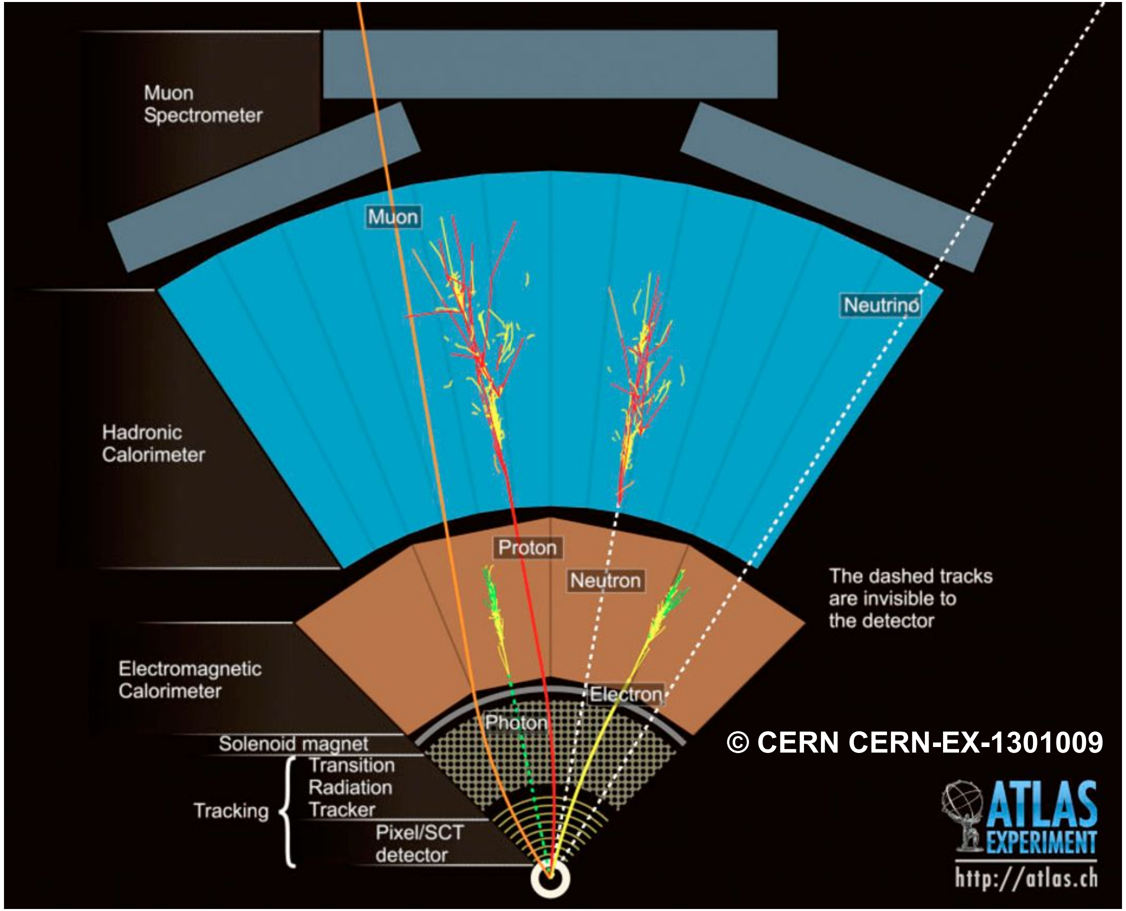
Detect



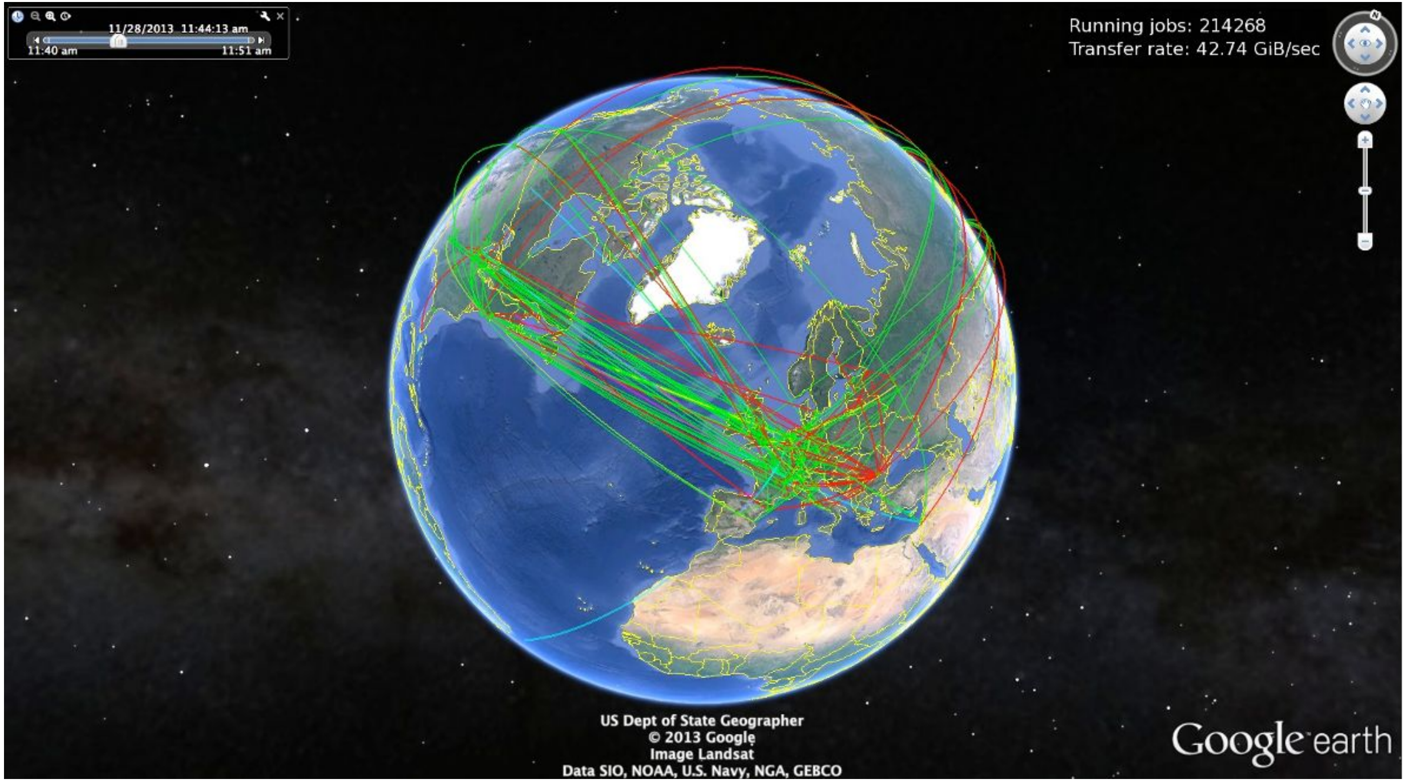
© CERN **LHCb-PHO-OUTTRACK-2006-001-1**

© CERN **LHCb-PHO-SILTRACK-2008-002-1**

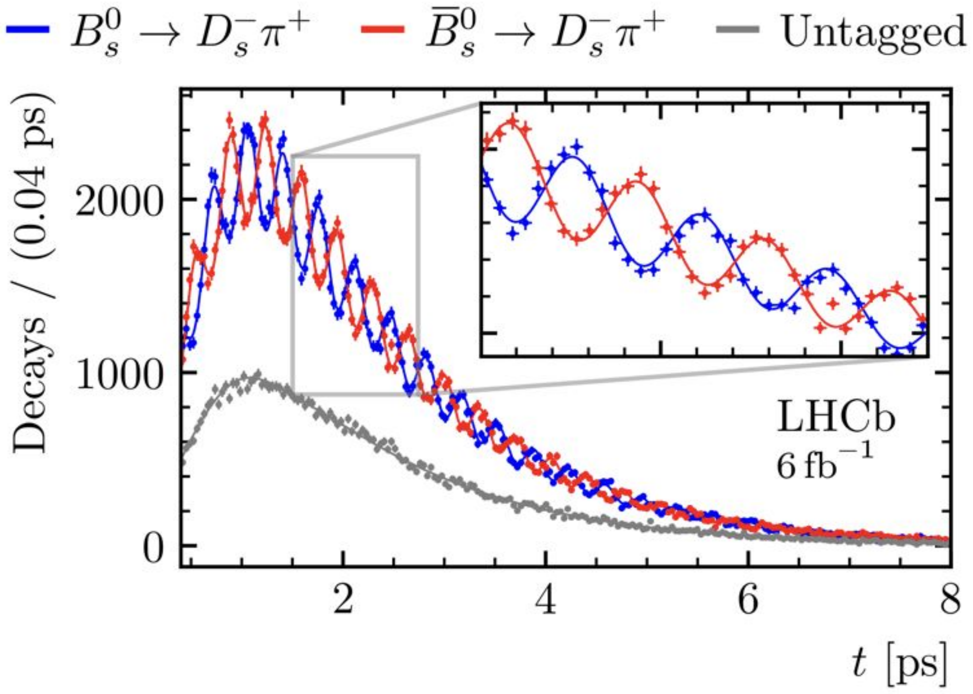
An Onion Structure



Process


11/28/2013 11:44:13 am
11:40 am
11:51 am
Running jobs: 214268
Transfer rate: 42.74 GiB/sec
US Dept of State Geographer
© 2013 Google
Image Landsat
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google earth

Analyse



https://lhcb-public.web.cern.ch/Welcome.html#BsOsc21

Discover!

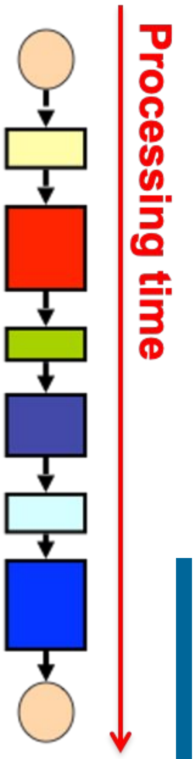
At CERN on 4 July 2012, the ATLAS and CMS collaborations present evidence in the LHC data for a particle consistent with a Higgs boson, the particle linked to the mechanism proposed in the 1960s to give mass to the W, Z and other particles. (Image: Maximilien Brice/Laurent Egli/CERN)

https://home.cern/science/physics/higgs-boson

Data acquisition and processing in a nutshell

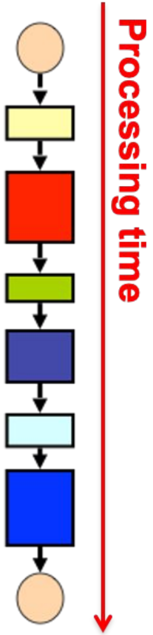
- **HEP main data: statistically independent* Events (particle collisions)**
- **Simulation, Reconstruction and Analysis: process "one Event at the time"**
    - **Event-level parallelism** (success of the Grid!)
    - Landscape is changed: advent of parallel data processing frameworks
- **Applications composed of several algorithms to:**
    - **Select and transform measured/simulated "raw" event data into "particles"**
    - **Create simulated "raw" event data** (event generation+simulation of passage of particles through matter+simulation of detector response to such energy depositions)
- **Final result: statistical data (histograms, distributions, etc.)**
    - Typically: **comparison between simulation and data**
- **All of these algorithms:**
    - **Are mainly developed by "Physicists"**
    - May require additional "detector conditions" data (e.g. calibrations, geometry, etc)

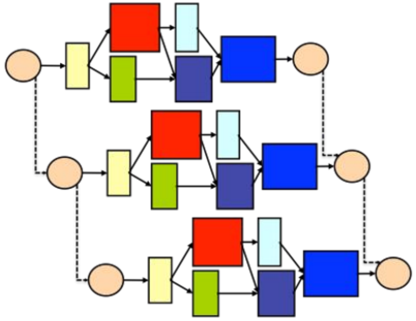
Raw data "definition": readout of the ADC of the subdetectors' frontends



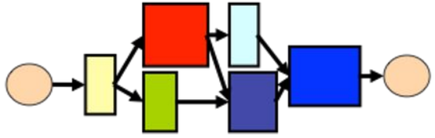
* Continuous readout would deserve more details.

Data acquisition and processing in a nutshell





Parallelism at all levels



Parallelism within an event

How Much Data?

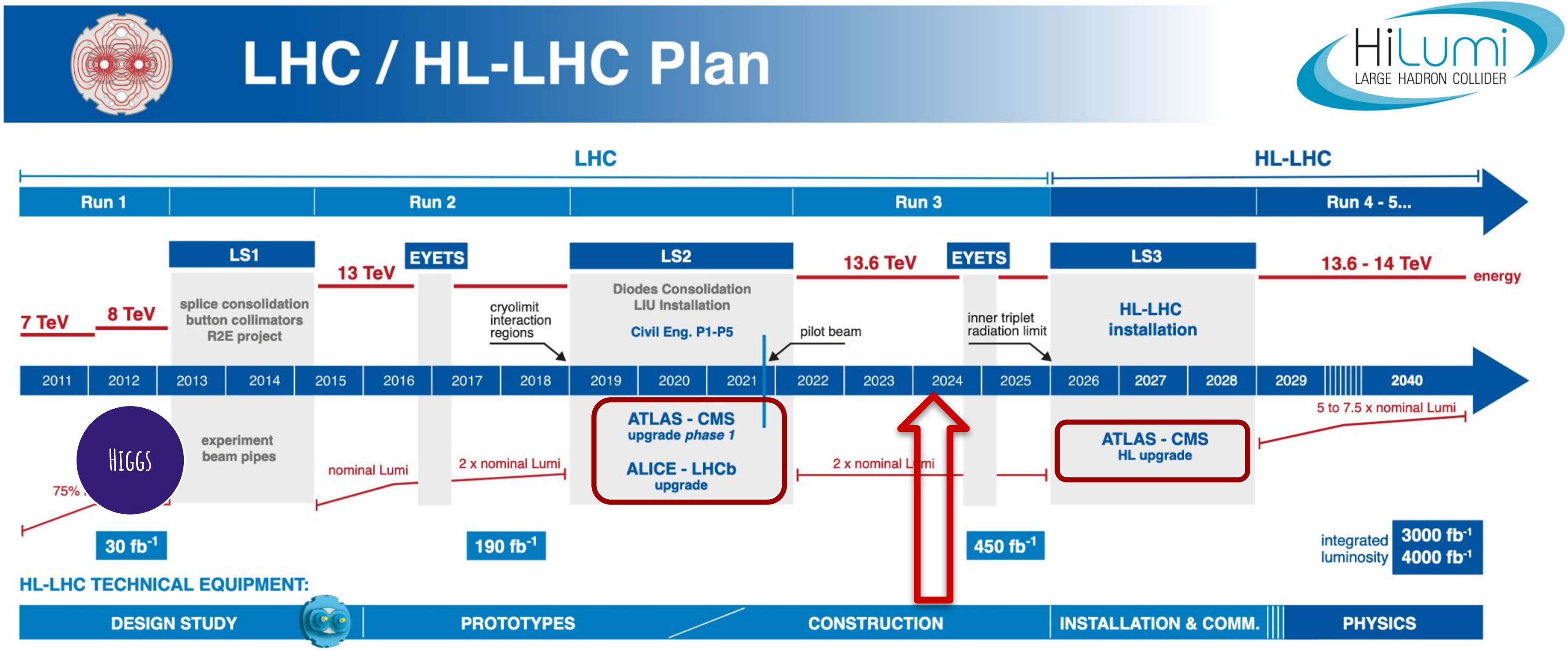
Experiment	2024 Disk Pledge [PB]*	2024 Tape Pledge [PB]*	2024 Disk+Storage Pledge [PB]*
ALICE	199	283	482
ATLAS	406	667	1073
CMS	304	673	977
LHCb	93	250	343
TOTAL	1002	1875	**2875**

** Numbers approximated to the nearest unit*

https://wlcg-cric.cern.ch/core/vopledgereq/listcomp/?year=2024

The Computational Challenges Ahead

How can we achieve all that in the Future?



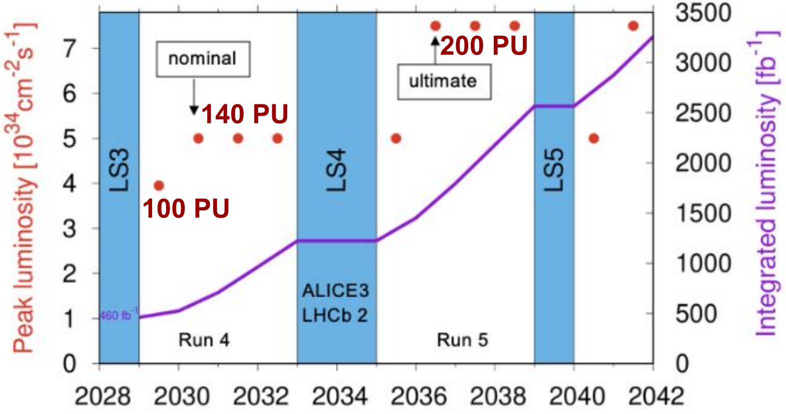
Upgraded detectors + improved accelerator
Alice and LHCb: Upgraded already!

https://project-hl-lhc-industry.web.cern.ch/content/project-schedule

How can we achieve all that in the future?

- **Upgraded detectors and an upgraded LHC**
- **Much more granular measurements, larger RAW events, larger number of objects to be considered during reconstruction**
- All this in the context of rapidly evolving computing platforms!

Ultimate HL-LHC performance

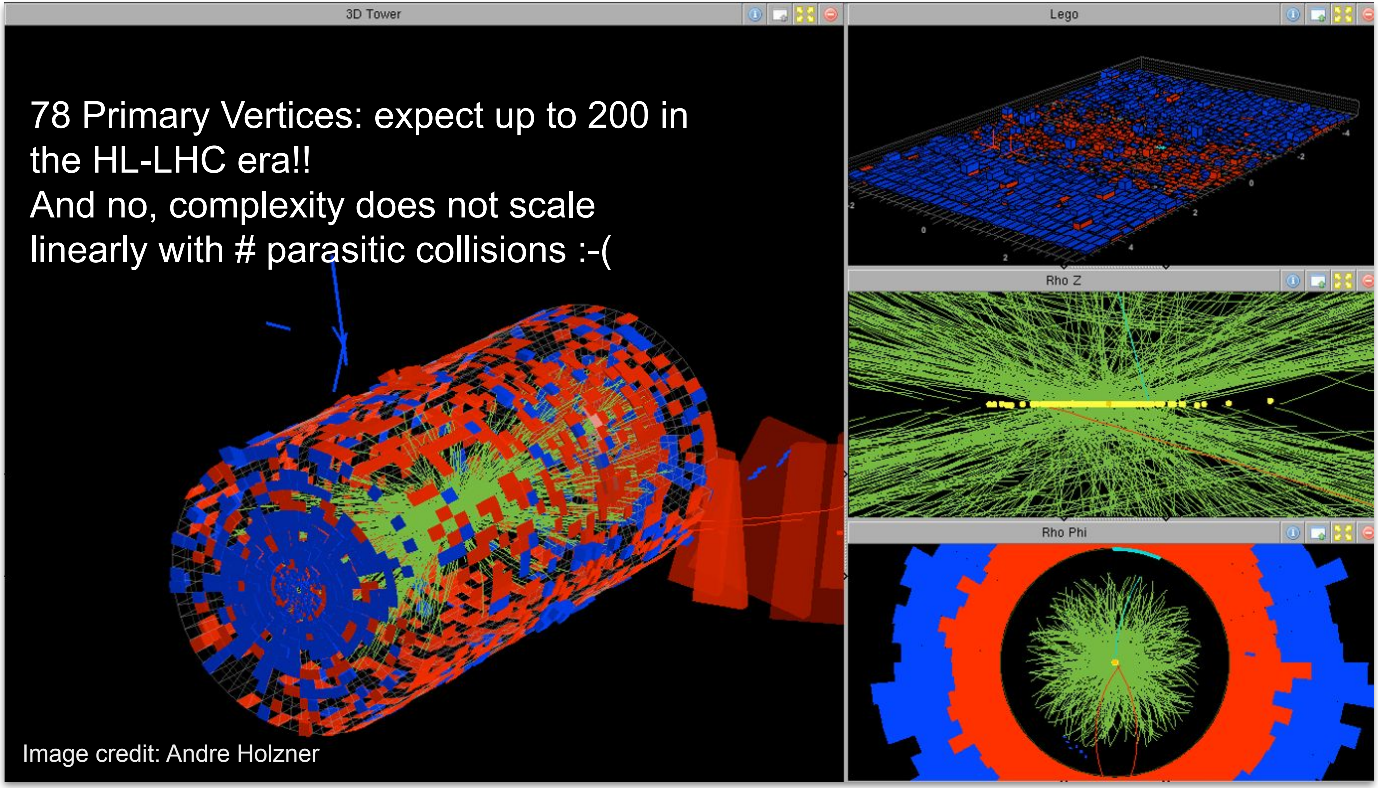


F. Gianotti, 13-1-2022

HL-LHC upgrades have to be paired to an equally ambitious upgrade of data processing software, computing tools and services.

By delivering more integrated luminosity, more parasitic collisions per bunch crossing are generated, more complexity to be handled by our algorithms!

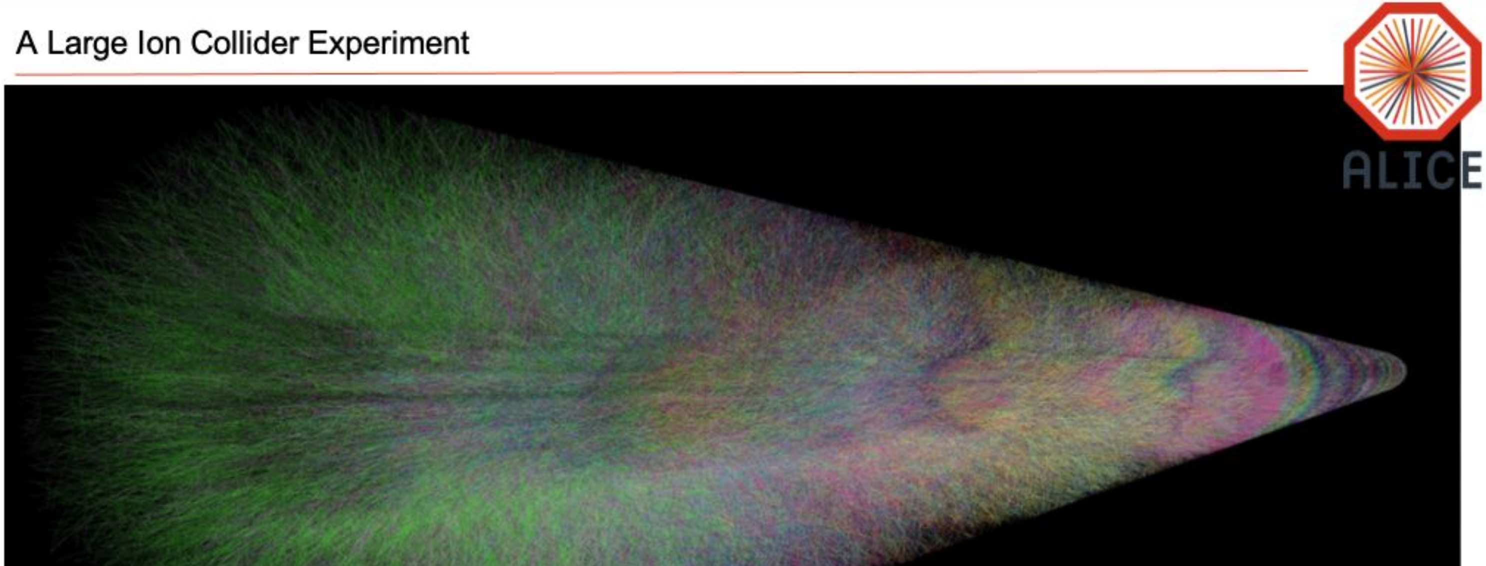
Vertices



https://cms.cern/news/reconstructing-multitude-particle-tracks-within-cms

ALICE, in Run 3

A Large Ion Collider Experiment

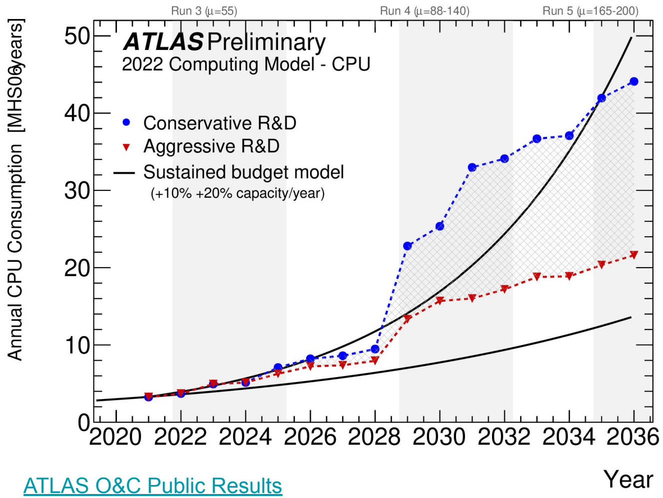


Visualization of **2 ms of 50 kHZ Pb-Pb** data as expected in the ALICE TPC in LHC Run 3

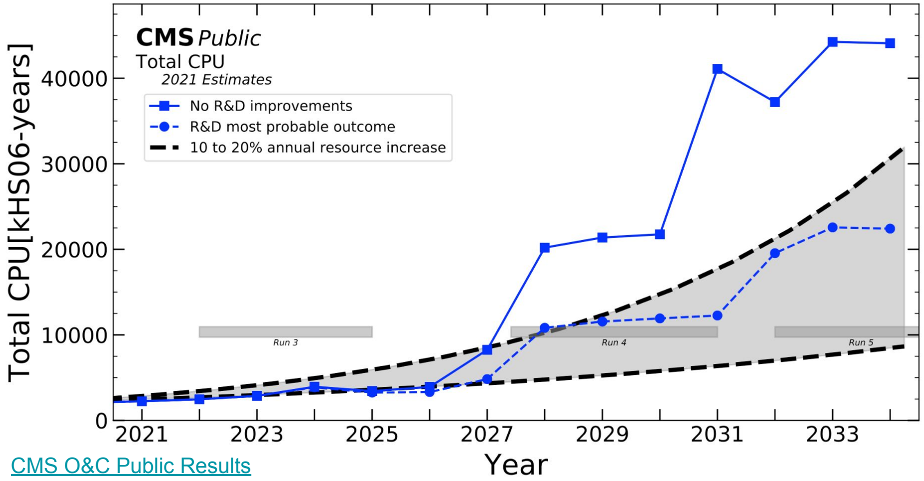
- Basic processing unit: **Time Frame** (~10 ms of data ~500 collisions @ 50 kHZ Pb-Pb)
- Whole TF reconstructed in one shot
- In absence of triggers (reference for drift-time estimate) z position of clusters is not defined

S. Piano, https://agenda.infn.it/event/26182

Offline projected computing needs in the HL-LHC era



[ATLAS O&C Public Results]



[CMS O&C Public Results]

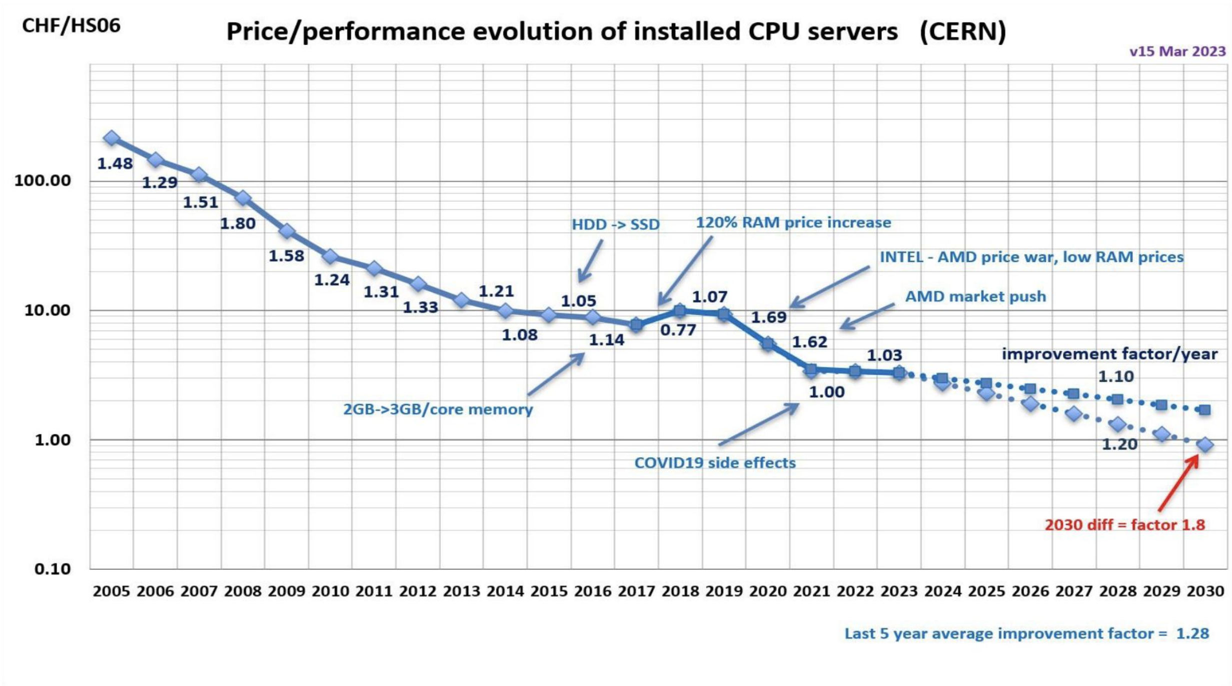
Online rates for upgraded detectors

- ALICE: 1 kHz → 50 kHZ
- ATLAS: 1 kHz → 10 kHz
- CMS: 1 kHz → 7.5 kHz
- LHCb: 12.5 kHz → 30 MHz

What a big gap between projected needs and availability!

It can be reduced with talented and curious physicists-developers :) **A single person can make the difference!**

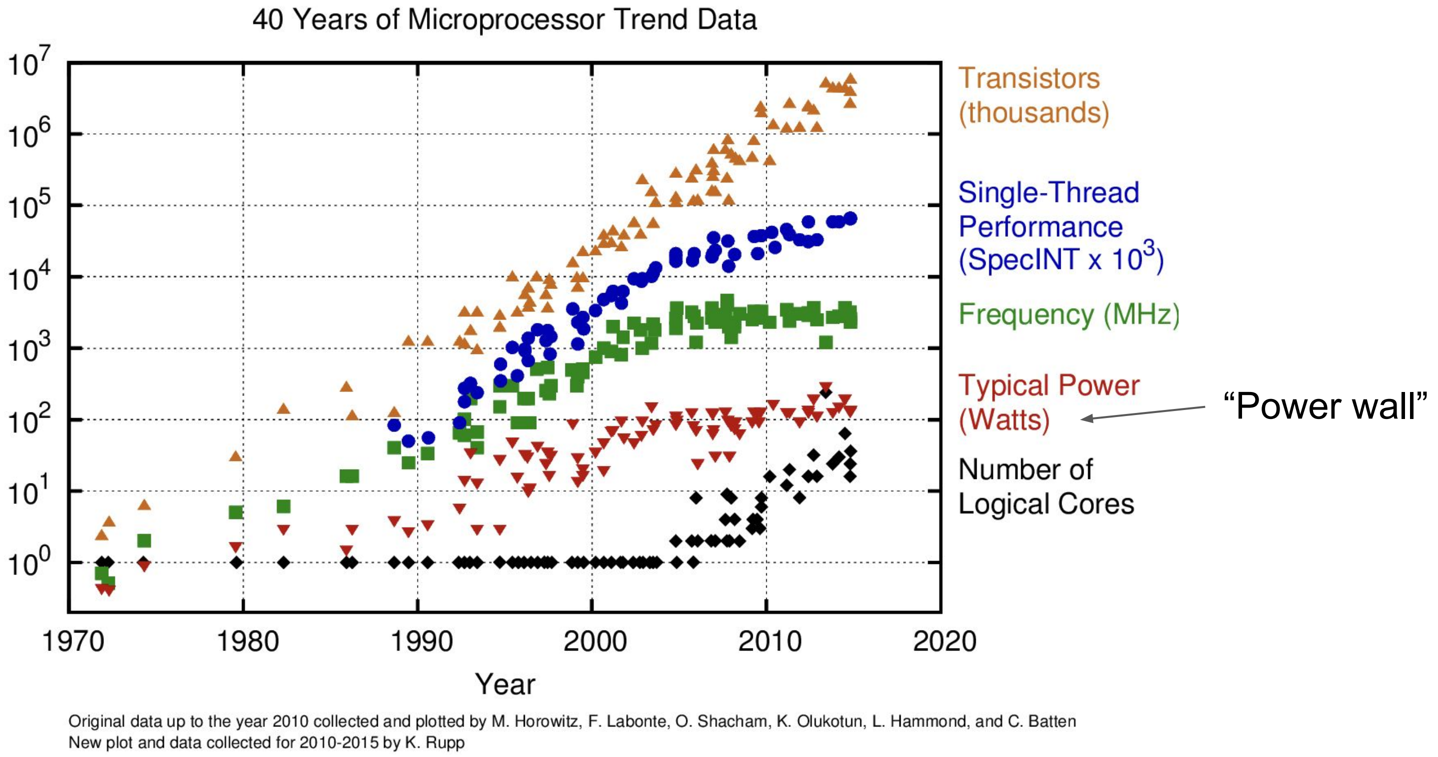
Can technology alone come to rescue?



- Predictions affected by large uncertainties
- Cannot count on technology alone
- Fierce competition among CPU manufacturers
- In the HL-LHC era, **computing power might not come only from CPUs, but also from accelerators**

Plot by B. Panzer

Parallelism 1/2



https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/

Parallelism 2/2

- Expressing parallelism at the data and task level will be key
- Clock speed of CPUs basically ceased to develop
- Moore’s law still valid: more and more transistors in our processors
- These materialised as:
    - Vector units: allow to apply the same instruction on multiple data (“data parallelism”)
    - CPU-cores: allow to formulate the problem in tasks executed concurrently (“task parallelism”)

Heterogeneity

- For two decades HEP benefited from a very uniform hardware landscape, dominated by the x86 CPU architecture
- {Data, High Performance} Computing centres feature today various CPUs, x86 but also ARM (and RISC-V in the future?)
- And heterogeneous architectures, too: most of the computing power is expressed through accelerators, e.g. GPUs
- Challenging devices to program, radically different way of thinking to the memory, e.g. offload calculations, asynchronous traffic between host and device
- Not embracing heterogeneous platforms could mean making HEP data processing “legacy computing”

High Performance Computing 1/2

- HPC: an asset to address scientific challenges and increase the competitiveness of industry
- Substantial national and supra-national investments have been made in this area
- HPCs are likely to be part of the future HEP computing infrastructure
- HPCs approaching the Exascale (10^{18} floating point operations per second capacity)
    - A veritable race, the first exascale HPC will be there well before Run 4
- How much is an EFlop?
    - LHC needs today the equivalent of ~30 PFlops
    - A single Exascale system enough to process 30 “today” LHC
    - A single Exascale system could process the whole HL-LHC with no software improvement or computing model change
    - Of course unrealistic, but just to give you an idea

High Performance Computing 2/2

- There are hurdles to overcome to use HPCs for HEP
    - HEP and HPC: **language spoken by experts can be different**
    - **Data access** (access, bandwidth, caches …): HEP has data processing applications (HTC)
    - **Submission of tasks** (MPI vs Batch systems vs proprietary systems)
    - **HPC: Handful computational kernels VS HEP: thousands of small kernels**
    - Environment **less open than Grid** one (OS, access policies, …)
    - **Node configuration** (low RAM/Disk, …)
    - Primary **architecture** (x86_64, Power9, ARM, proprietary, …)
    - Relationships between Tier-{1,2} providers and **HEP experiments are decades long**
- **And virtually all HPCs offer accelerators, e.g. GPUs**
    - Not being able to use GPUs might prevent HEP to even access HPCs because of existing policies

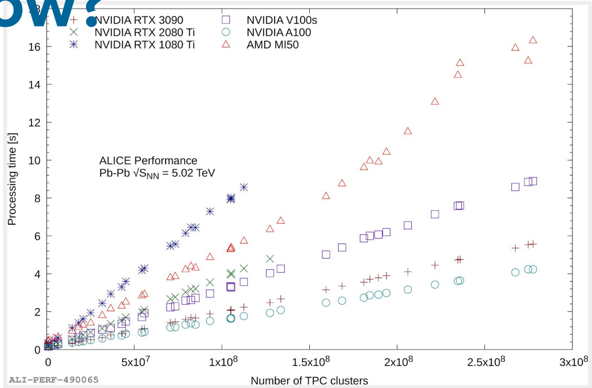
Solving those hurdles requires also investing in people

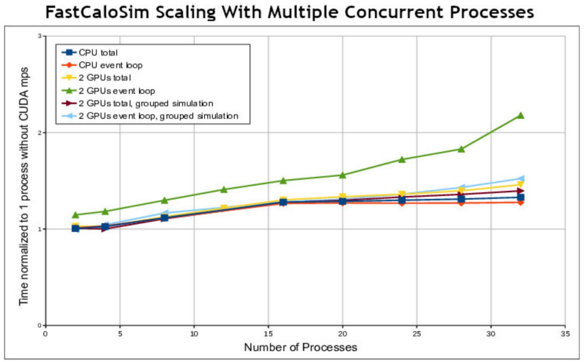
GPUs: Where are we now?

Slide inspired by V. Pascuzzi, LHCP 2021

Alice, continuous TPC readout; LHCb, fully software-based triggers

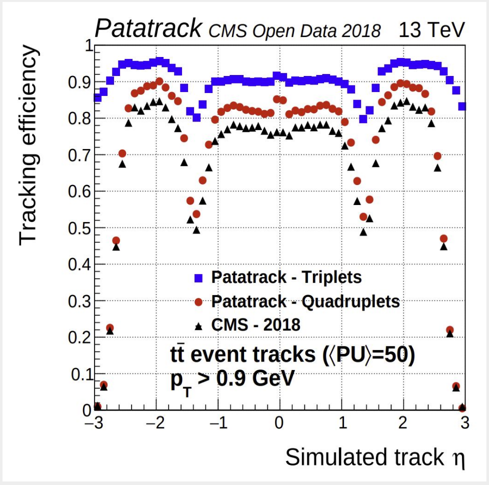
- ALICE leveraging AMD GPUs for real-time processing, online data reduction and compression
- LHCb's Allen framework using ~200 NVIDIA GPUs for HLT1 to handle 30 MHz input rate

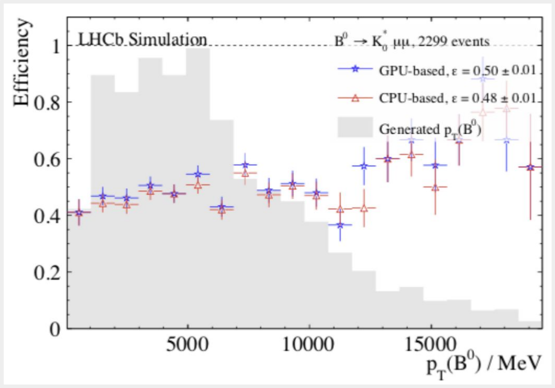




CMS using NVIDIA GPUs for part of the HLT sequence

- Higher throughput and equivalent results to CPU
- Equal or better physics performance
- Used for Run 3 at HLT





ATLAS

- Promising LAr Calorimeter reconstruction porting to GPU targeted code

Bocci, Kortelainen, Innocente, Panteleo, Rovere (2020), arxiv:2008.13461, *Bozzi* (2021) Web seminar , *Rohr* (2020) arxiv:2006.04158, Leggett (2021) vCHEP

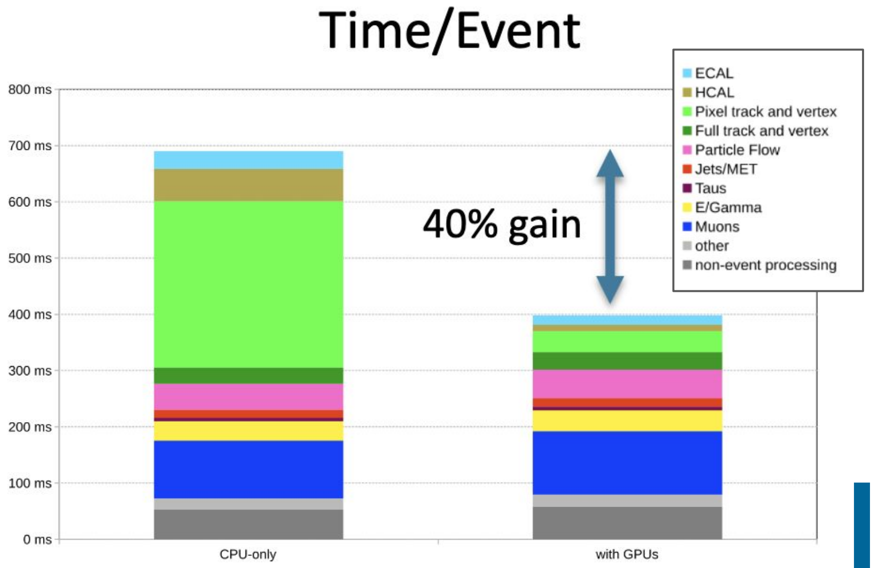
A Word About Energy Savings

An example from the CMS experiment

High Level Trigger Farm equipped with GPUs, fo

- 70% better event processing throughput
- 50% better performance per kW
- 20% better performance per cost

In some configurations, GPUs can mean less energy used.



Plot by A. Bocci CHEP2023

How to evolve from here

- A bright future of HEP ahead of us: many questions to answer, discoveries made accessible thanks to the HL-LHC program
- More events, more complex, at a finer granularity
  - Processing and analysing them is a challenge for us
  - To be faced with a constantly evolving hardware landscape
- Two elements which help succeeding: parallelism and heterogeneity
- Interesting times for motivated scientists-engineers!
  - A single person can make the difference
- A lot has been achieved, but a lot still needs to be done
- This tCSC is an opportunity for getting nearer to the solution, as a community