Computing at CERN: challenges and opportunities

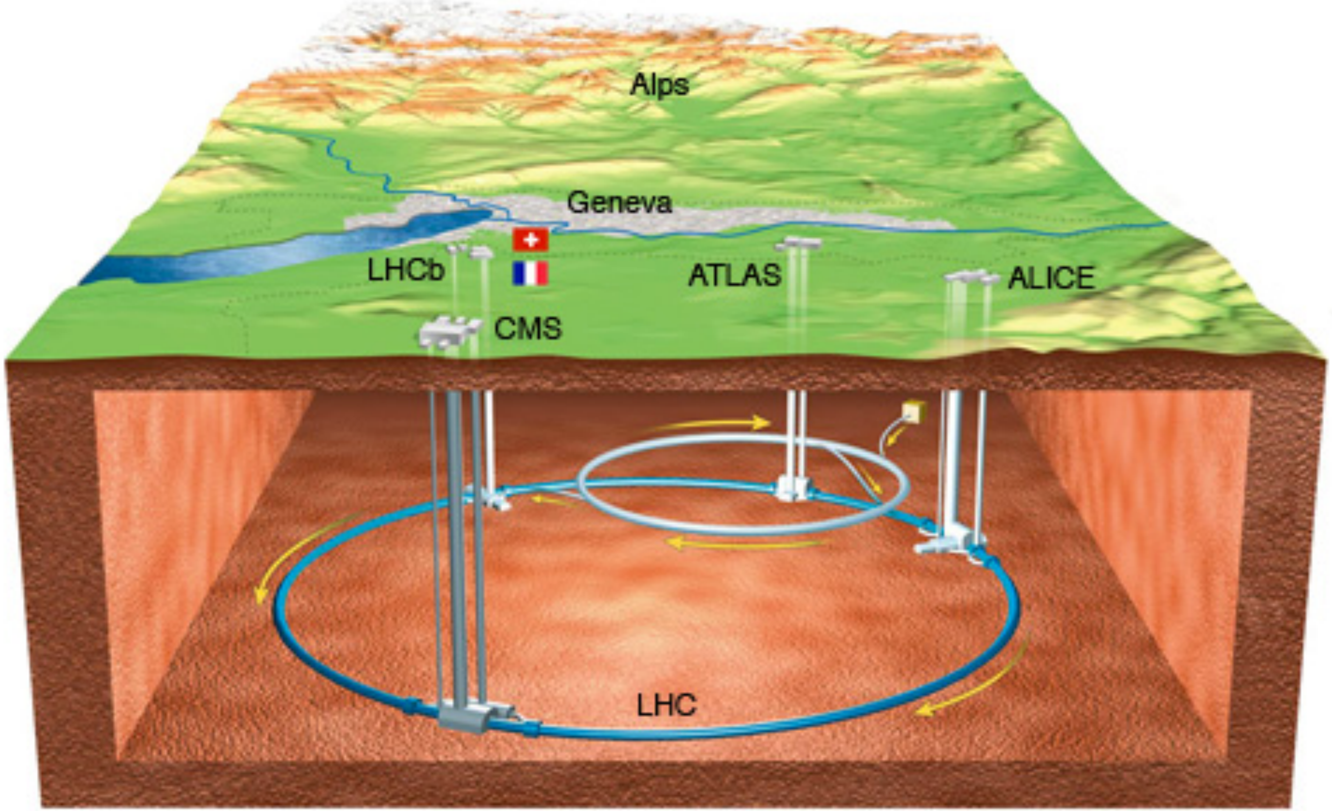
28.6.2017

PASC 2017

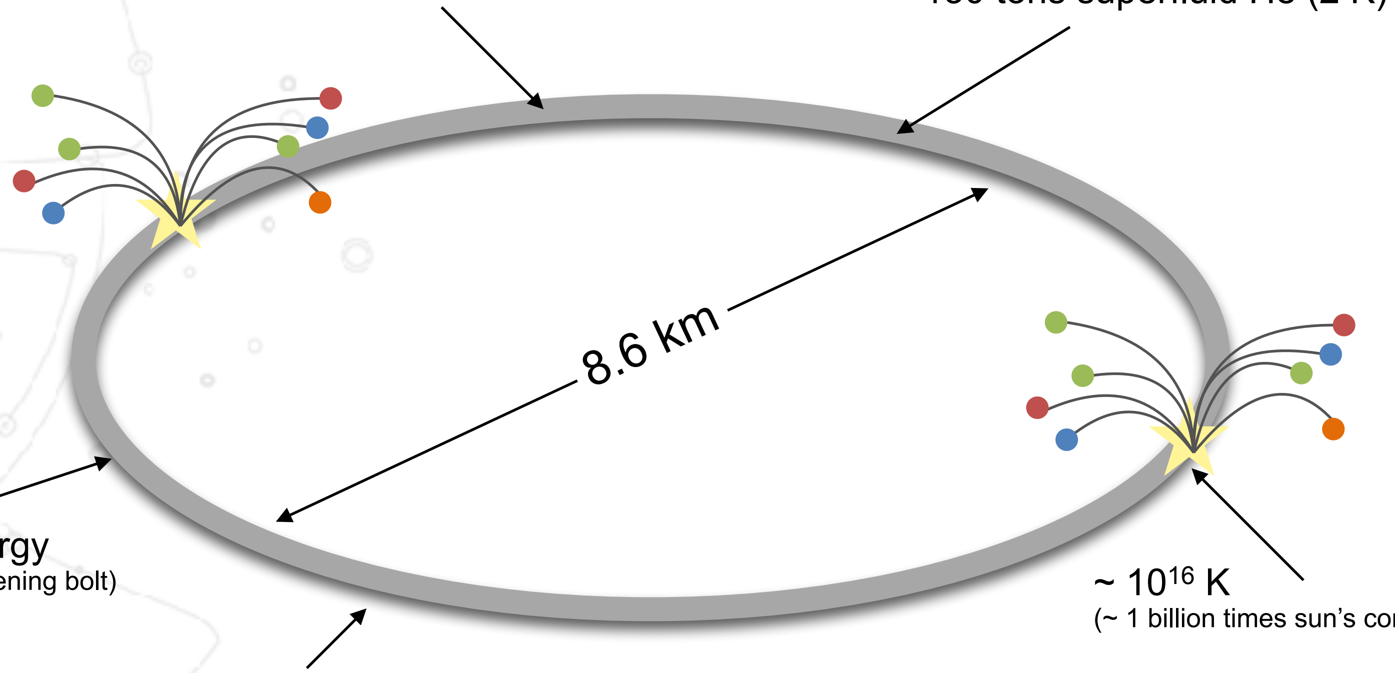
Omar Awile (omar.awile@cern.ch),

CERNopenlab

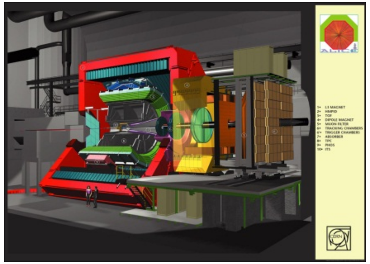
CERN: The **L**arge **H**adron **C**ollider



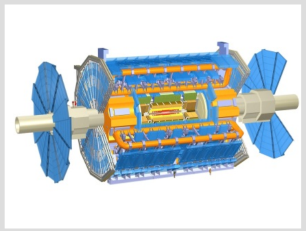
The Particle Accelerator



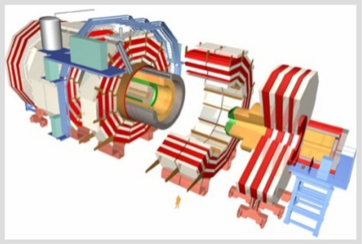
The experiments



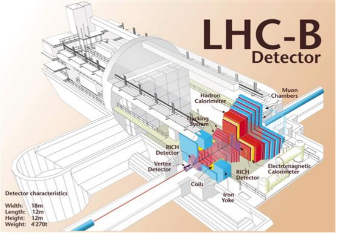
- ALICE – “A Large Ion Collider Experiment”
  - Size: 26 m long, 16 m wide, 16m high; weight: 10000 t
  - 35 countries, 118 Institutes
  - Material costs: 110 MCHF



- ATLAS – “A Toroidal LHC ApparatuS”
  - Size: 4 6m long, 25 m wide, 25 m high; weight: 7000 t
  - 38 countries, 174 institutes
  - Material costs: 540 MCHF



- CMS – “Compact Muon Solenoid”
  - Size: 22 m long, 15 m wide, 15 m high; weight: 12500 t
  - 40 countries, 172 institutes
  - Material costs: 500 MCHF

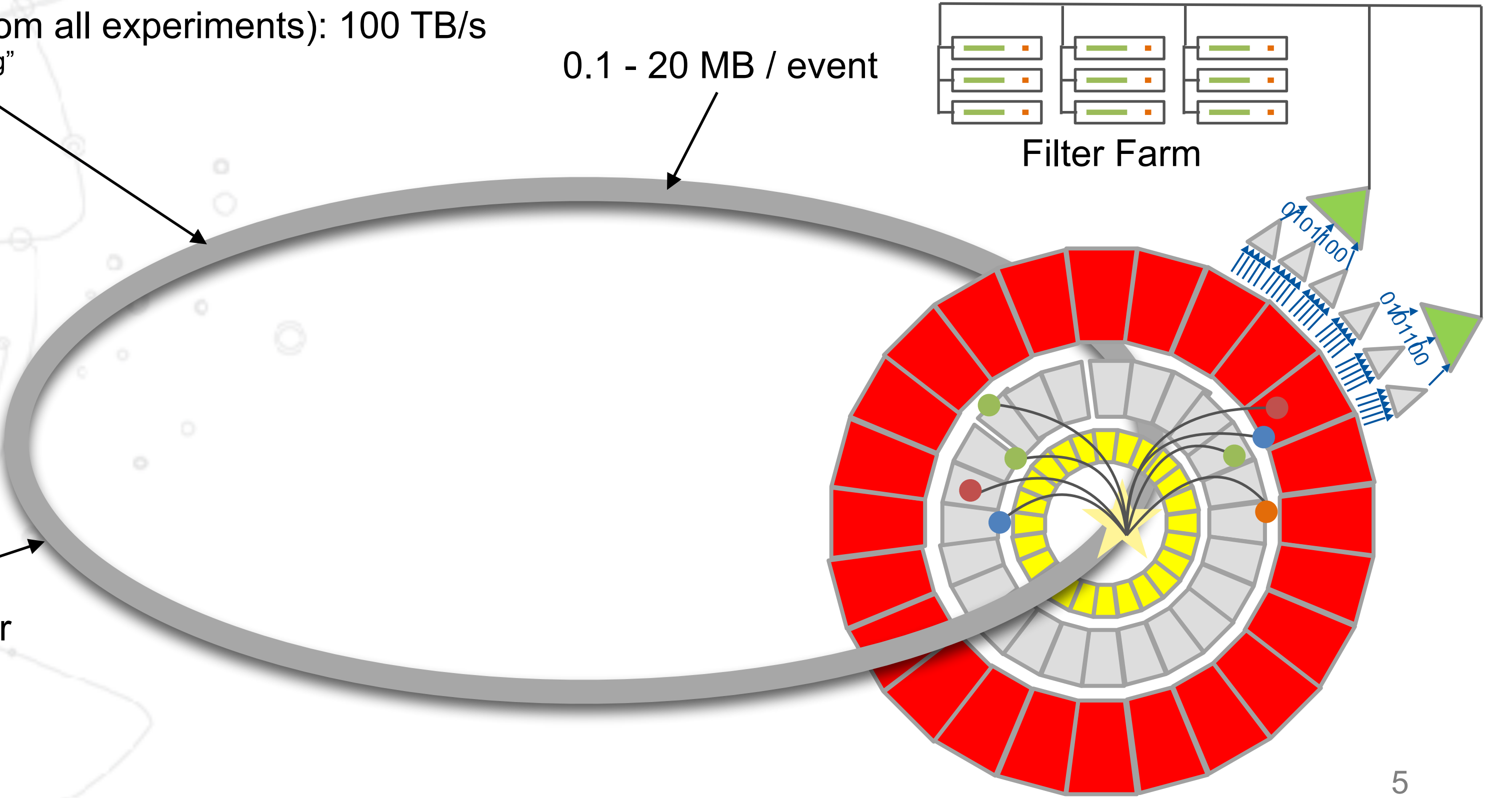


- LHCb – “LHC beauty”
  - Size: 21 m long, 13 m wide, 10 m high; weight: 5600 t
  - 15 countries, 52 Institutes
  - Material costs: 75 MCHF

Data Acquisition (DAQ) & Hardware Triggering

total readout rate (from all experiments): 100 TB/s

1 in 10^7 events is “interesting”



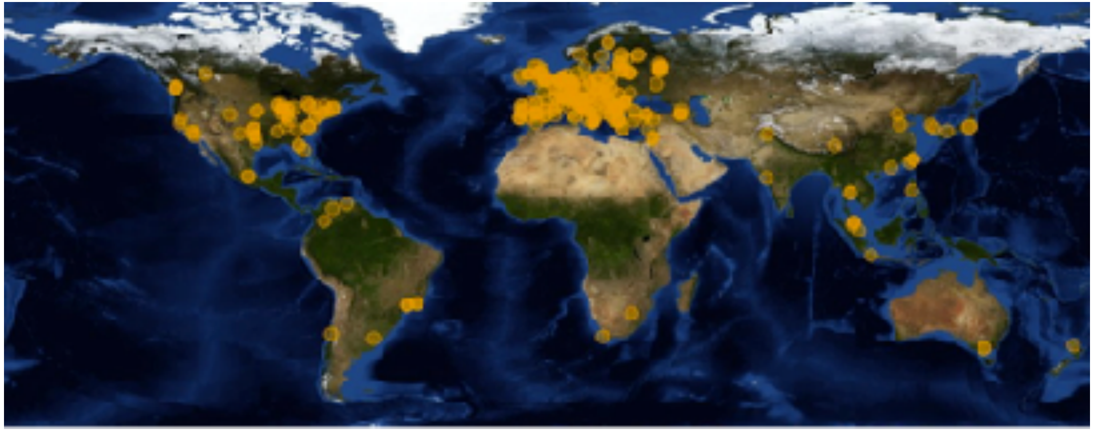
40 Million crossings per second

The High-Level Trigger

- Triggering (filtering) is done in two steps
    1. Filter data in hardware (low-level L1 trigger) incoming rate $4*10^7$
       reduction factor 400 - 10’000
    2. Filter the pre-filtered data using an in-software high-level trigger
       reduction factor: 10 - 2’000
- Filtering has to be performed online in near-realtime.
- HLT reconstructs all particle trajectories and decays
    - takes decision based on reconstructed event data
- Remaining data is stored for offline analysis

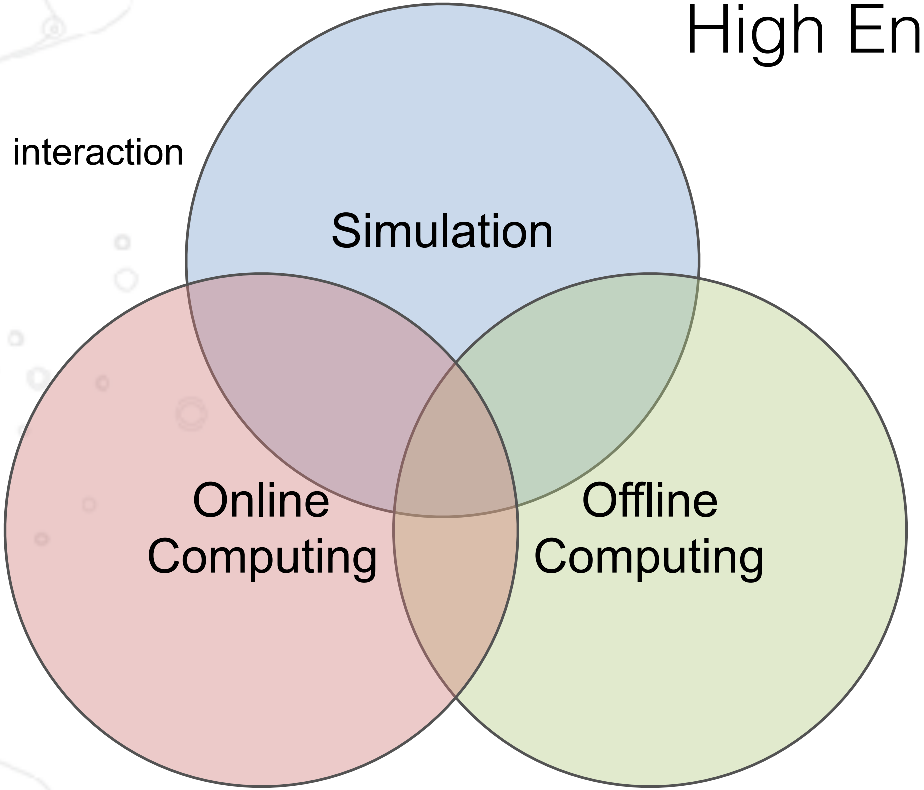
The Worldwide LHC Computing Grid

- The remaining events after filtering are sent to the CERN datacenter for tape archival and disk buffer
    - Redistributed to 13 Tier-1 and 155 Tier-2 sites
- Up to 700k cores (2016) committed for up to 400k jobs in distributed computing grid
- Analyses are run offline



The Three Areas of Computing in experimental High Energy Physics

detector and
particle/material interaction



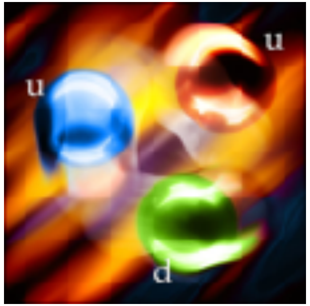
High throughput, near realtime
data filtering

virtualized, distributed, grid computing,
data analysis

High Performance Computing vs. High Throughput Computing

High Performance Computing	High Throughput Computing
Usually **large** data sets (grids, particles, PiCs) **decomposed** over many nodes	Many **small** datasets (events) **distributed** over many nodes
Tightly coupled, i.e. some kind of (local, or global) synchronization needed over runtime —> local / global comm.	Embarrassingly parallel, each event is processed independently except for event building or aggregation (global comm.)
Parallelization using MPI+X (or global address space)	Parallelization using multithreading (pthreads, TBB) + process manager
Execution time as short as possible, but no hard time limits	Triggering: near realtime (< 100 ms) Offline Analysis: no hard time limits
fault tolerance important but difficult (MPI)	data loss costly, but fault tolerance easy to deal with

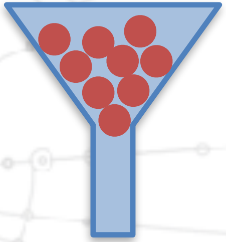
What about HPC at CERN?



- lattice QCD codes
  - Method for solving quantum-chromodynamics theory. Simulations used in theoretical particle physics
  - CERN runs one (small) HPC cluster specifically for lattice QCD codes.
- accelerator simulation
  - simulating the particle beam
  - simulating physical properties of beam pipe and interactions between beam and its environment
  - structural simulations of beam pipe
  - Need for HPC resources and expertise increasing

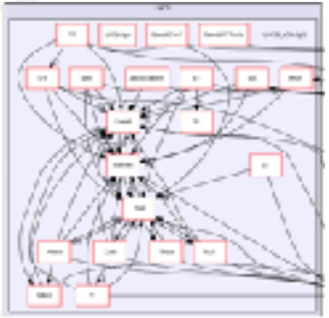
Computing Challenges Today

10^9 events/s
100 Tbit/s



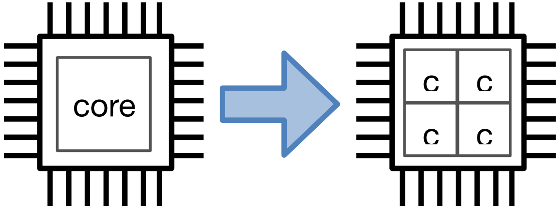
~ 45 PB / yr

reduction factor up to 10^4

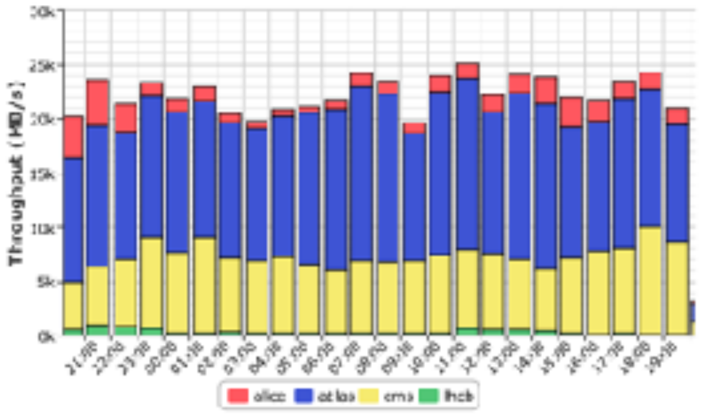


Complex codes
100k - 1M LOCs

many algorithms
flat performance profile
10 - 100 ms to decide if event is to be kept

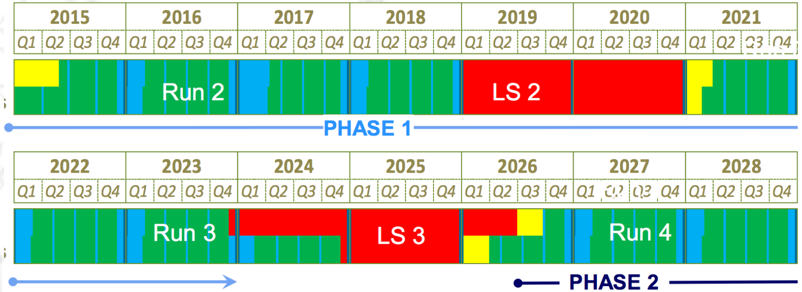


Codes were developed before multicore / SIMD architectures became “mainstream”

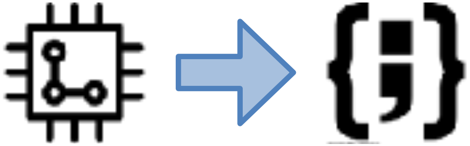


Very complex systems, variety of network and compute hardware,
Analysis code written by scientists with *casual* programming experience
big fluctuations in job queue size

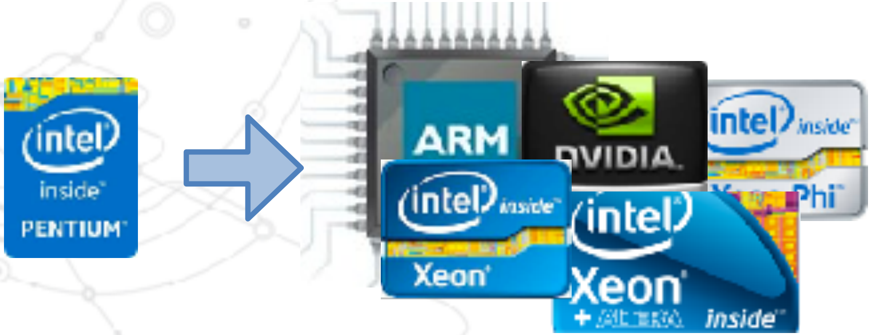
Computing Challenges in the Future



Expected increase 50x in data throughput after LS 3 (upgrade)



Experiments will try to implement triggering entirely in software - reduction factor must be achieved entirely in software!



hardware (compute & interconnect) is becoming more heterogeneous. Programming models and software must adapt

How should data be stored in the future? Produced data extremely valuable!

Thank you!

and now…

- Gerhard Raven, (VU University Amsterdam, Netherlands)
  HEP Realtime Analysis: Scaling Beyond Embarrassingly Parallel

- Daniel Hugo Campora Perez, (University of Seville, Spain)
  High Performance Computing meets High Energy Physics

- Felice Pantaleo, (CERN, Switzerland)
  Heterogeneous Event Selection at the CMS Experiment