HEP Computing in the 20s (and beyond)

16 April 2018

Davide Costanzo

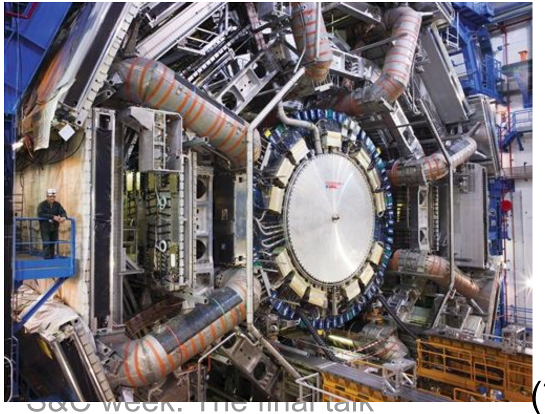
University of Sheffield

(d.costanzo@sheffield.ac.uk)

The role of Computing and Software

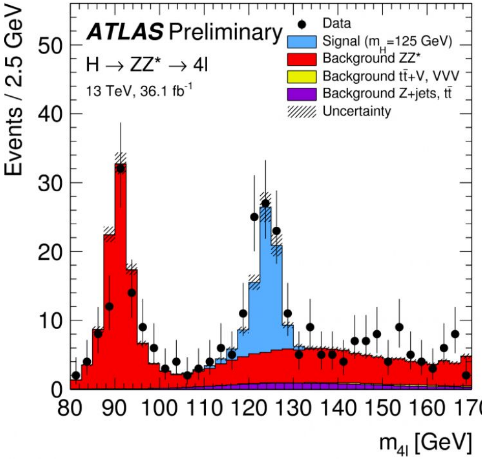
Accelerators

Detectors

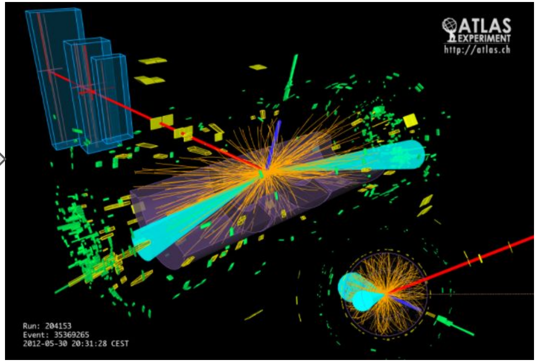


Data

Analysis

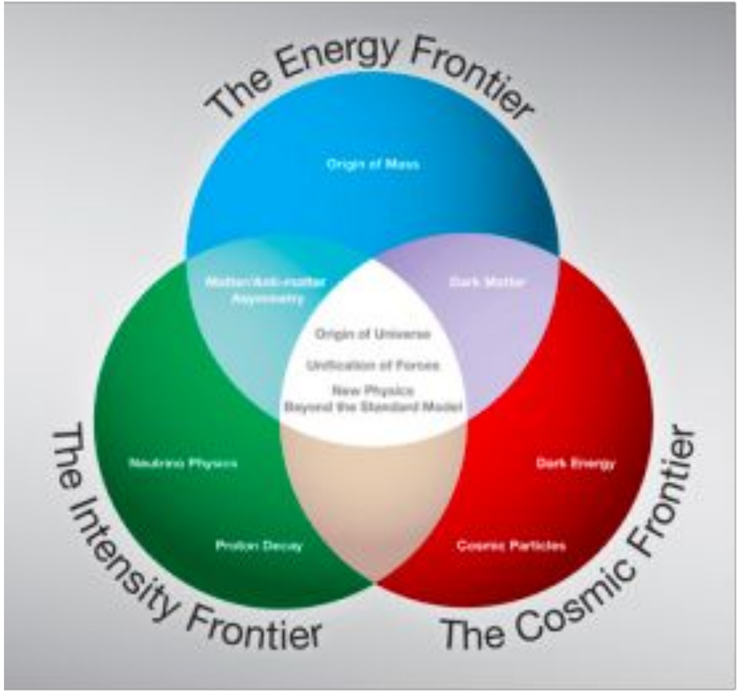


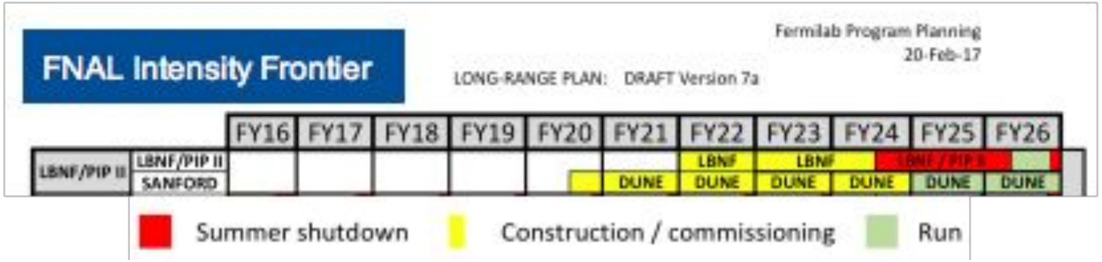
Sw & Computing



(*replace ATLAS by your favourite experiment)

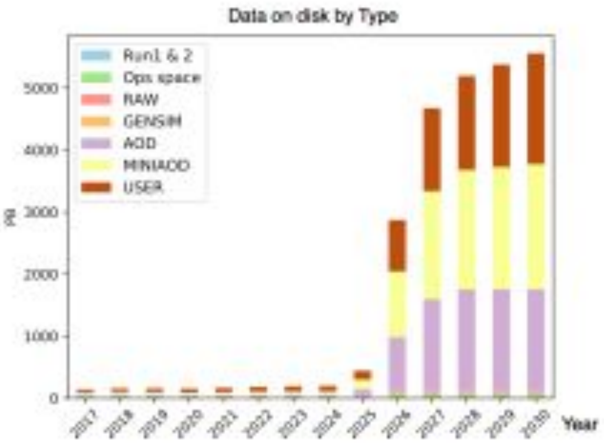
HEP plans → Big data in the 2020s





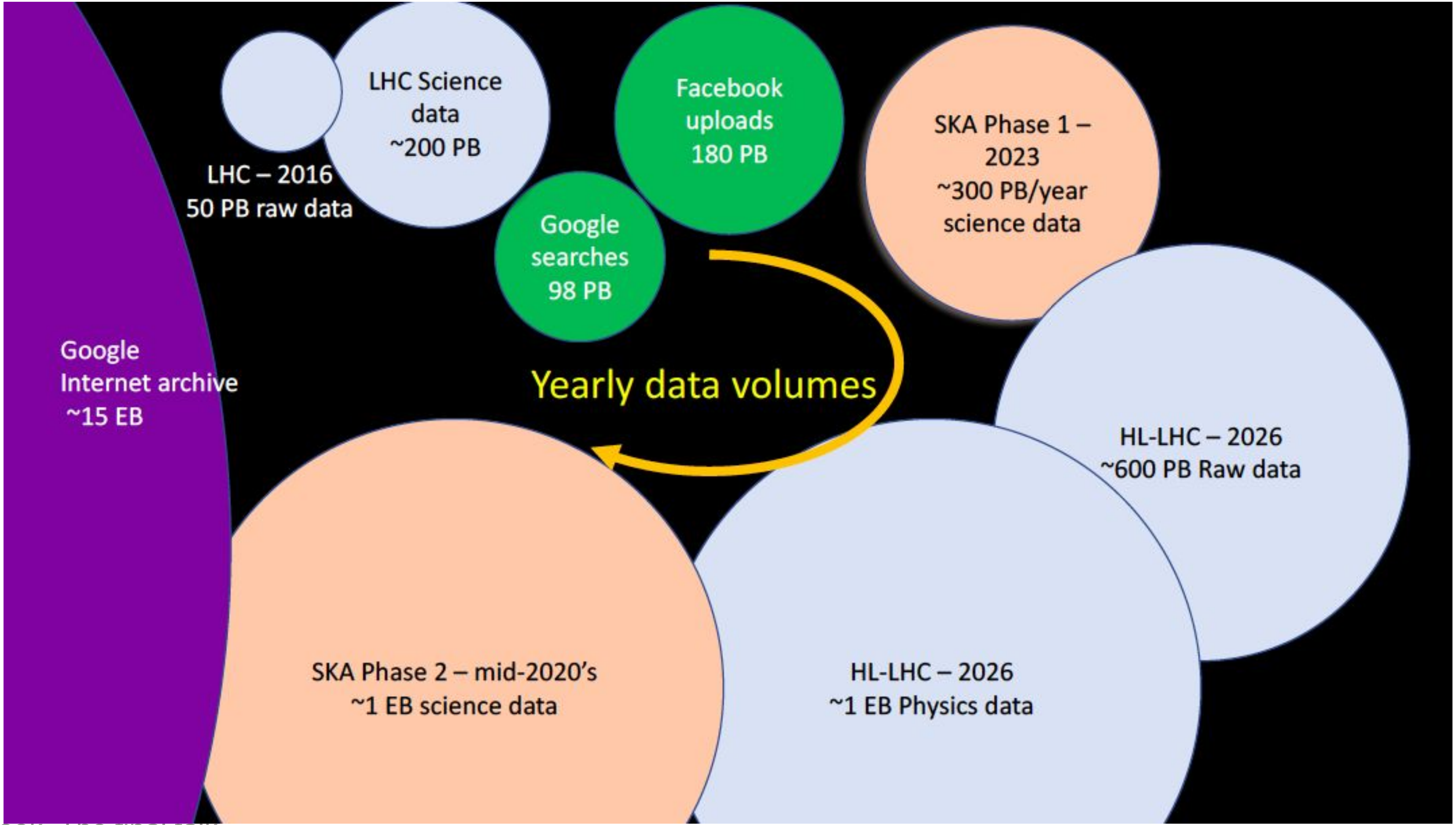
LHC and HEP Software and Computing dimension

- HEP Physics programme supported thanks to a vast investment in Software
  - Estimated to be around 50M lines of C++
  - Which would cost around 500 MCHF to develop commercially
- Critical component of the physics production pipeline
  - From Trigger to physics analysis, simulation and final plots
- HEP experiments use about 600k CPU cores every hour
  - At CERN. Tier1s, Tier2s, Tier3s, Cloud, High-Performance Computers (HPC)
- LHC stores 400 PB of data on disk and 600 PB on tape
  - In 2018, we are already at the “Exabyte” scale



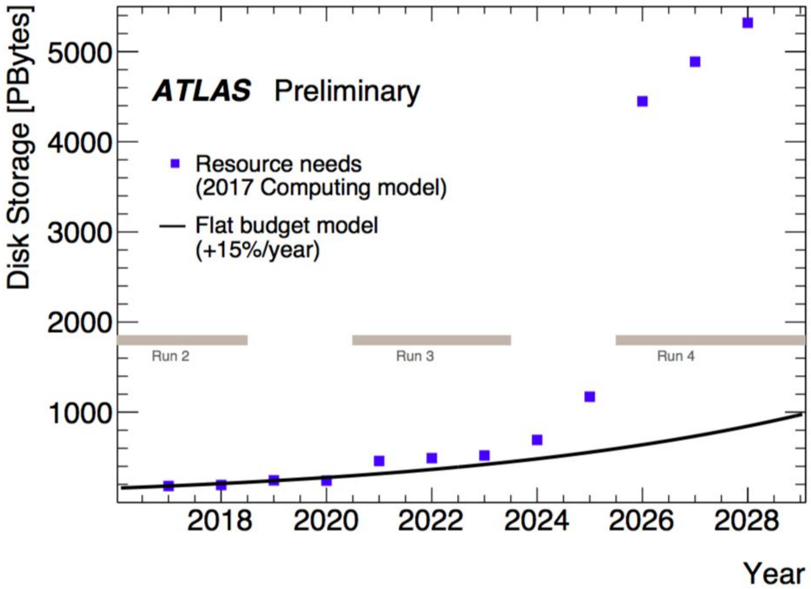
- Huge challenge for HL-LHC

International data needs (2016+)



Software and Computing Challenges. HL-LHC and beyond

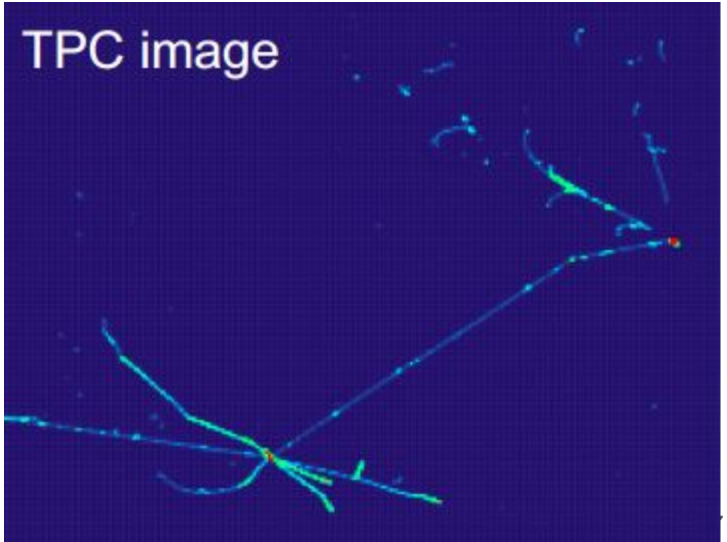
- Computing and Software for High-Luminosity LHC is still an open problem
  - Future colliders (HE-LHC, FCC, …) will deliver large datasets
  - Extrapolation of current model is simply not affordable!
- Belle II and Dune face similar challenges
  - And non-HEP experiments (SKA, LSST, LIGO, …)



- HEP Software Foundation (HSF) established to facilitate coordination and common efforts in Software and Computing in HEP in general
  - Several workshop took place and a community white paper was recently published

DUNE, LArSoft and neutrino challenges

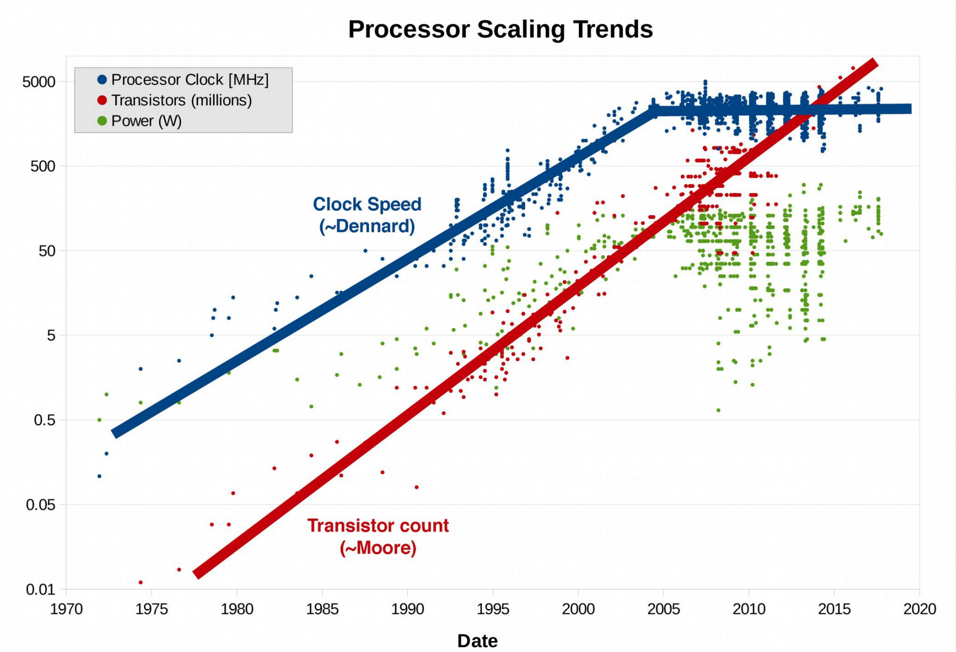
- LArSoft
  - A collaboration of experiments, labs, university groups and software projects
  - Several challenges on the algorithmic side.
  - Reconstruction is basically a 3D image pattern recognition
- Performance improvements needed
  - Faster, less memory hungry
  - While improving the physics performance!
- Computing challenges
  - Potentially 400 PB/year
  - Supernovae and proton decay searches need a continuous stream (with radioactive decay bckd)
- Trigger, zero suppression

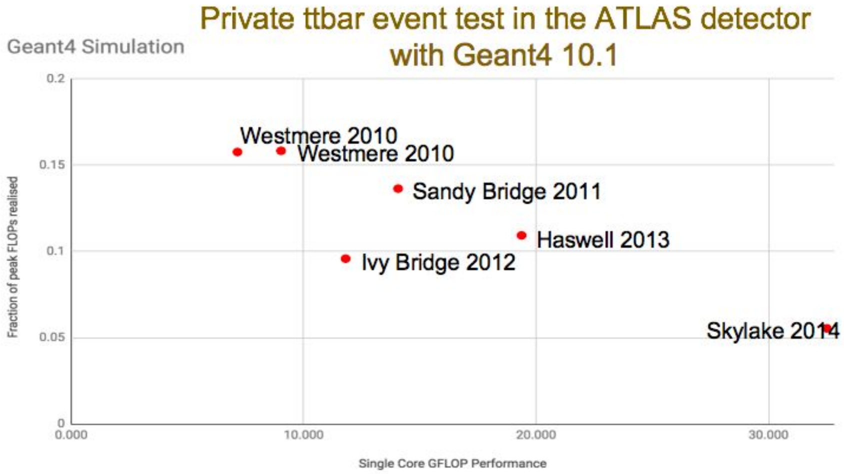


[(*More info)]

Technology evolution. Needs to upgrade

- ➢ Moore law -> increase of transistor density with time
- ➢ Clock speed stopped at ~3GHz
- ➢ Limited by power (Wm^{-2} limit)
- ➢ HEP software designed to execute one Instruction set at a time
- ➢ Improvements from SW vectorisation





- ➢ Multiple processors per chip
  - ○ Multithreaded applications
  - ○ Memory shared by multiple threads

HEP Software Foundation - HSF

arXiv:1712.06982v3 [physics.comp-ph] 11 Feb 2018

A Roadmap for
HEP Software and Computing R&D
for the 2020s

HEP Software Foundation[1]

ABSTRACT: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

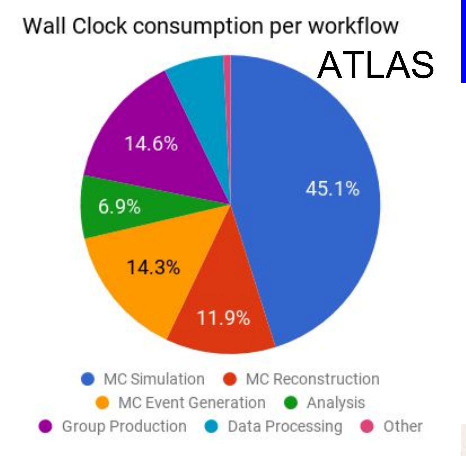
(for more info ask Graeme Stewart for a longer talk)

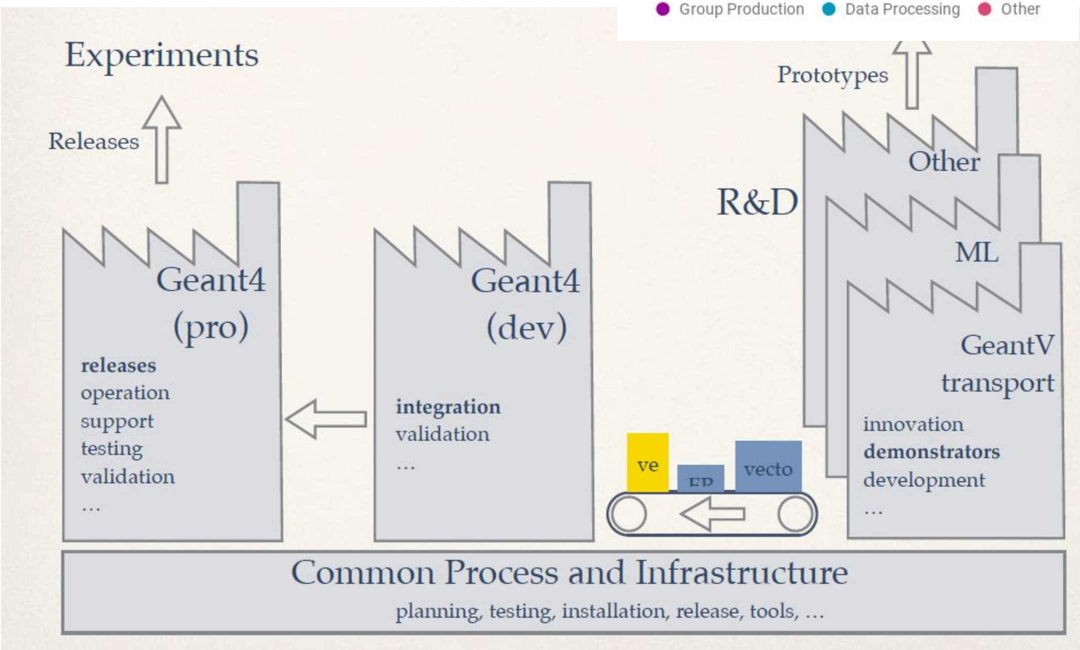
HEP Software Foundation

- ➢ Community white paper available on arXiv
- ➢ Summary on the CERN Courier
- ➢ Latest community workshop, 2 weeks ago
  - ○ Joint with the wLCG workshop, ~150 participants
- ➢ 13 topical sections, 70 pages, 300 authors from 120 institutes
  - ○ summarising R&D in a variety of technical areas for HEP Software and Computing
  - ○ Separate in-depth papers on each topic
- ➢ 1 section on training and careers

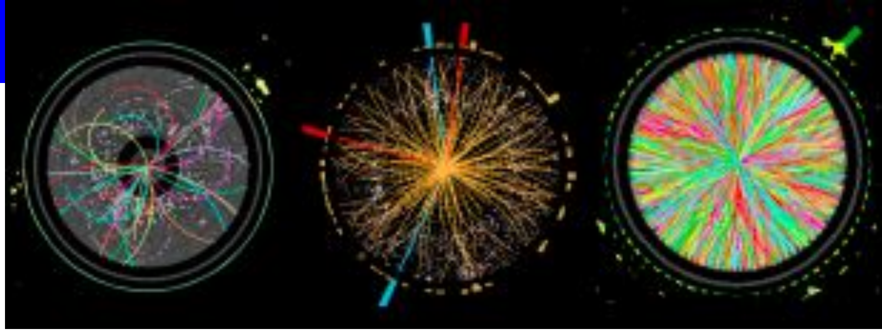
Future challenges. Simulation

- ➢ Simulating our detectors consumes huge resources today
- ➢ Remains a vital area for HL-LHC and intensity frontier
  - Geant4 as a common tool
- ➢ Main R&D topics
  - Improved physics models for higher precision at higher energies (HL-LHC and then FCC)
  - Adapting to new, in particular vectorised, computing architectures
  - Fast simulation: develop a common toolkit for tuning and validation
  - Geometry modelling: easier and efficient modelling of complex detectors



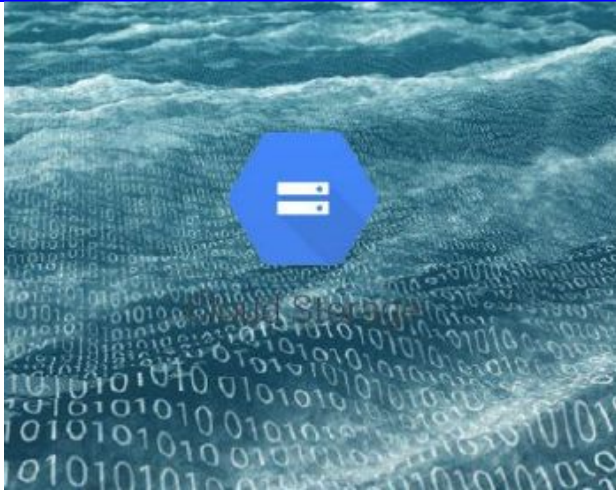


Trigger and event reconstruction



- Very experiment specific
  - Commonalities for collider experiments and neutrino experiments
- Move to software triggers is already a key part of the program for LHCb and ALICE already in Run 3
  - ‘Real time analysis’ increases signal rates and can make computing much more efficient (storage and CPU)
  - Robust validation techniques when information will be discarded
- Main R&D topics
  - Controlling charged particle tracking resource consumption and maintaining performance at very high pile-up (of 200 or even 1000!)
  - Detector design itself has a big impact (e.g., timing detectors, track triggers)
  - Improved use of new computing architectures, extending the use of GPGPUs and possibly FPGAs

Data management



- ➢ Data storage costs are a major driver for HEP today
  - HL-LHC will bring a step change in the quantity of data
  - LHCb already there for this year’s run 3
- ➢ Main R&D topics
  - Adapt to new needs driven by changing algorithms and data processing needs, e.g., fast access to training datasets for Machine Learning, high granularity access to event data, rapid high throughput access for a future analysis facilities…
- ➢ Consolidate storage access interfaces and protocols
- ➢ Efficient hierarchical access to data, from high latency tape and medium latency network
- ➢ Common solutions towards a “data lake” or “data ocean”
  - ATLAS data management tool (Rucio) used by other communities and being considered by CMS, Fermilab, SKA, …
  - Collaboration with google - Rucio endpoint in google storage

Machine learning

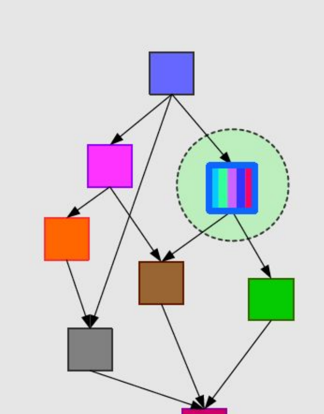
- ➢ Not a challenge for HEP, per se, rather a tool set providing solutions
  - ○ ML techniques applicable to many different areas
- ➢ Neural networks and Boosted Decision Trees used in HEP for a long time
  - ○ e.g., particle identification algorithms
- ➢ The field has been significantly enhanced by new techniques (Deep Neural Networks), enhanced training methods, and community-supported packages
  - ○ Very good at dealing with noisy data and huge parameter spaces
  - ○ A lot of interest from our community in these new techniques, in multiple fields
- ➢ Main R&D topics
  - ○ Speeding up computationally intensive pieces of our workflows (fast simulation, tracking)
  - ○ Enhancing physics reach by classifying better than our current techniques
  - ○ Improving data compression by learning and retaining only salient features
  - ○ Anomaly detection for detector and computing operations
- ➢ Good links with the broader Machine Learning and Data Science communities required

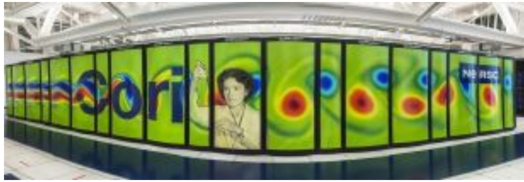
High Performance Computers - HPC

- Large facilities for a broader Science community
  - Require highly parallelised workflows
  - More recent ones use Intel Knightlanding CPUs

- Require dedicated attention to run HEP jobs
  - Software installation, I/O, memory usage, ...
  - ATLAS simulated ~9% of Geant4 events on HPC

- So far mostly US facilites
  - Some exist in Europe, collaborations with CSCS PizDaint
  - Bid under prep for time on Prace in EU (Knightlanding)





- Future challenge. Heterogeneous computing
  - Use GPU or FPGA part of HPCs
  - Ship threads to different facilites and reassemble events back together

alk

Conclusions

- ➢ Computing and Software is an essential component of the future of HEP
  - Consolidation where possible for better economy of scale
- ➢ Computing infrastructure needs to evolve with technology
  - Investments in CPU, Storage (disk, tape, …), network, need to continue
  - Moore law helps, but it’s not for granted!
  - Role of HPCs unclear, but likely to change our SW paradigm. We need our say in future facilities
- ➢ Development and improvement of Software is of critical importance
  - Training of new Software users. Large students’ base in HEP experiments
  - Software professionals need to develop a career on par to detector builders
  - The days when we could patch a bit of code together and make a plot are long gone!
  - Personal opinion: US are ahead of Europe on this
- ➢ Data and analysis preservation, open access
  - Need to preserve our precious data for the future
  - And allow for public access. (most software stack is now public)