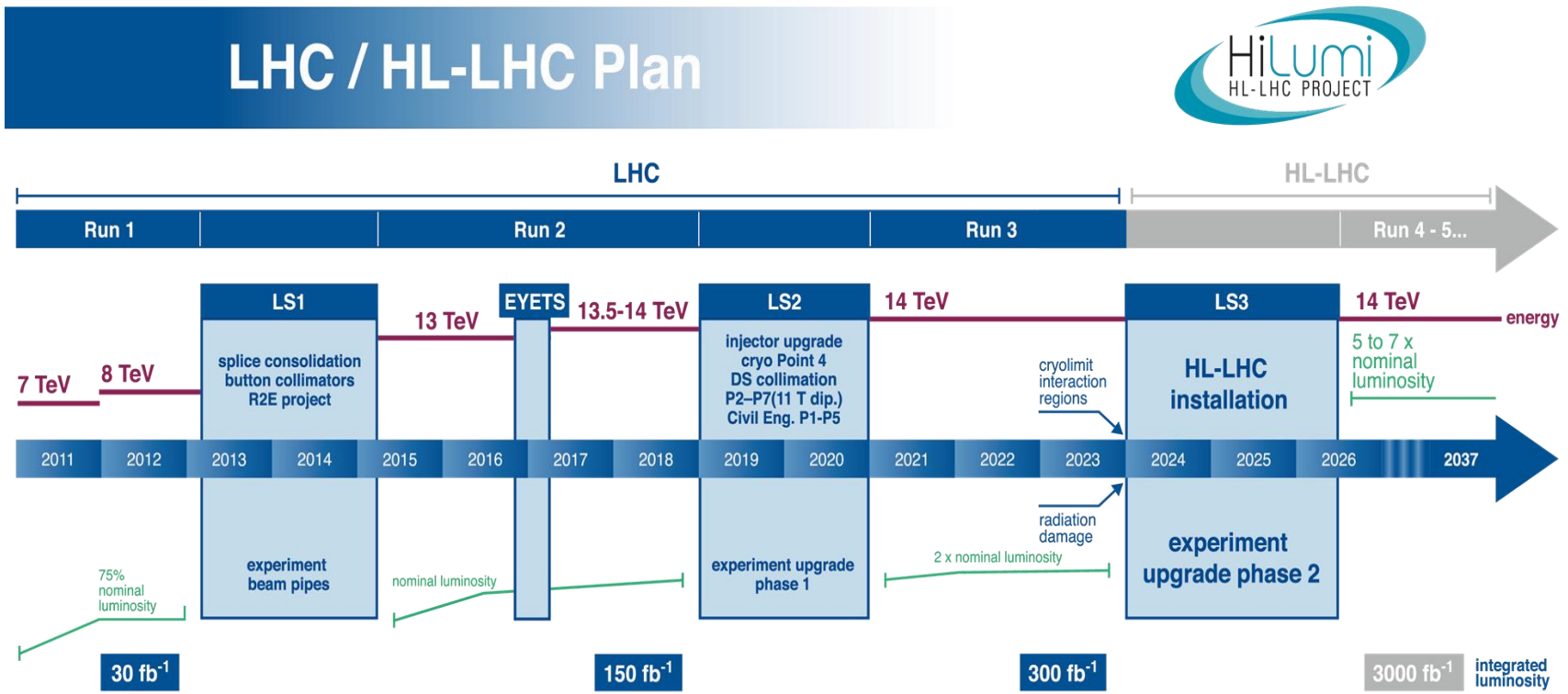
Accelerated GPU computing with E4

Felice Pantaleo

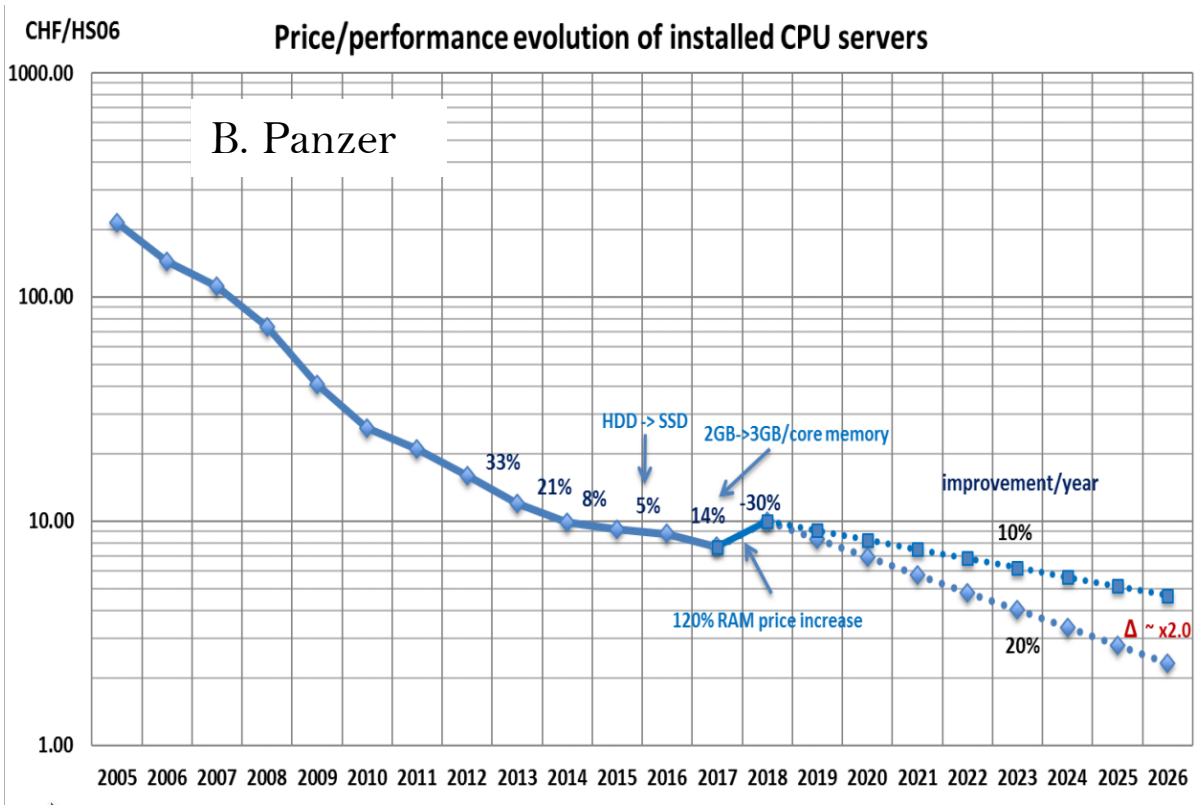
(CERN – Experimental Physics Department)

felice@cern.ch

LHC Upgrade Schedule



CPU trends



- CPU evolution is not able to cope with the increasing demand of performance
- Depending on the application, GPUs can provide better performance and energy efficiency

The Times They Are a-Changin'

Achieving sustainable HEP computing requires change

Long shutdown 2 represents a good opportunity to embrace a paradigm shift towards modern heterogeneous computer architectures and software techniques:

- Heterogeneous Computing
- Machine Learning

Algorithms and Frameworks

Algorithms and Frameworks

The acceleration of algorithms with GPUs is expected to benefit:

- Online computing: decreasing the overall cost/volume of the event selection farm, or increasing its discovery potential/throughput
- Offline computing: enabling software frameworks to execute efficiently on HPC centers and saving costs by making WLCG tiers heterogeneous
- Volunteer computing: making use of accelerators that are already available on the volunteers' machines

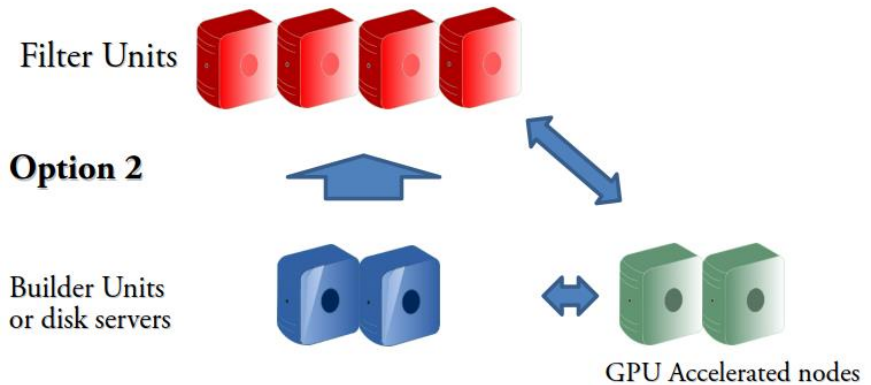
Online: LHCb

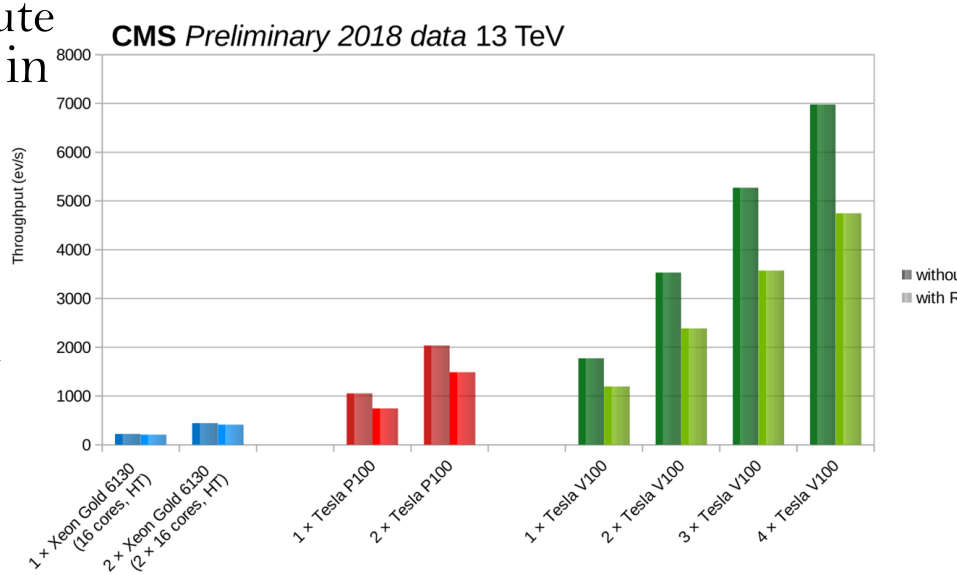
- Big changes during LS2
- Event selection based on two software-based HLT levels
- HLT-1 reduces 5TB/s input to 130GB/s:
  - Track reconstruction, muon-id, two-tracks vertex/mass reconstruction
  - GPUs can be used to accelerate the entire HLT-1 from RAW data
- HLT-2 reduces 130GB/s to 10GB/s:
  - Full offline quality reconstruction, alignment and calibration

More info: D. vom Bruch, V. Gligorov, D. Campora Perez

Online & frameworks: CMS - Patatrack

- Demonstrated advantage of heterogeneous reconstruction from RAW to Pixel Vertices at the CMS HLT
  - 1 order of magnitude both in speed-up and energy efficiency wrt full Xeon
  - Running within the CMS software framework
  - Benchmarks executed at Flatiron institute
- Parallelization of more algorithms to run in production during Run 3 and Run 4
- Performance portability
- Definition of a composable farm approach with remote offload





More info: A. Bocci, F. Pantaleo

Offline & Volunteer: Beam Dynamics

Simulation of dynamics of particle beams in accelerators

- With single-particle and with multi-particle interaction
- Embarrassingly parallel problem limited by memory bandwidth (in the case of multi-particle interaction)
- Numerical stability
- Running on LHC@Home and on server-grade machines
- Goals:
    - Incorporate advanced non-linear tracking algorithms
    - Improve performance portability CPU/GPU and avoid code duplication
    - Improve integration

More info: R. De Maria, L. Mether, A. Oeftiger

Machine Learning

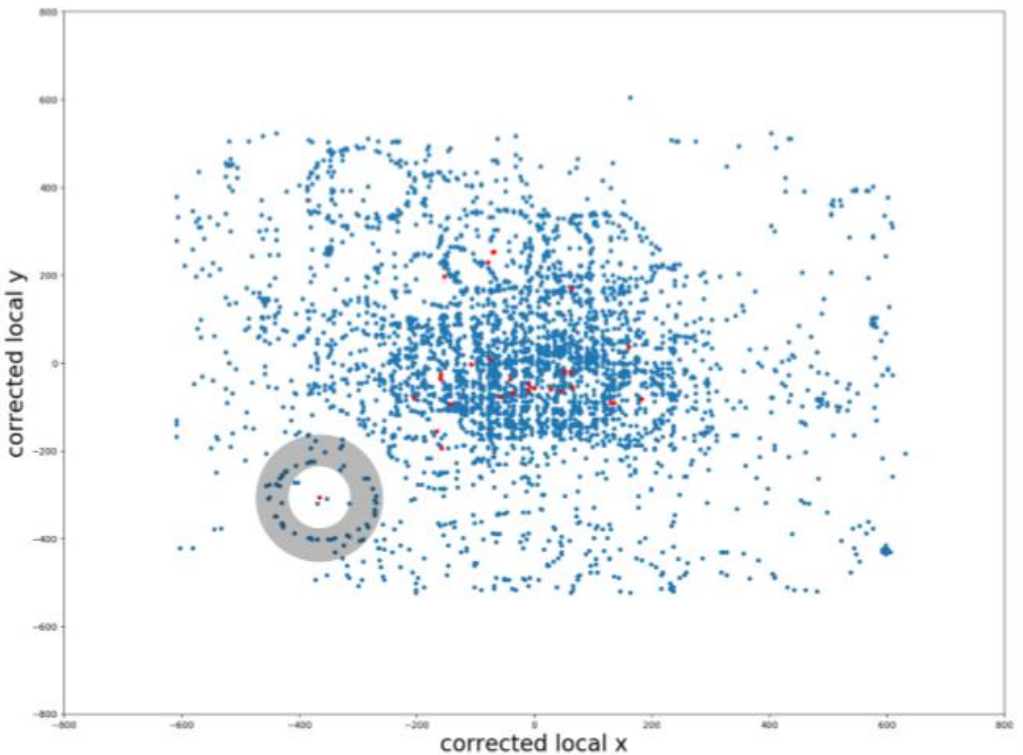
Machine Learning

- NNs are becoming more complex:
  - dramatically increasing training time
  - decreasing productivity due to higher turnaround time
  - increasing the infrastructural cost

- GPUs have proven to be of invaluable for data scientists in the last decade by accelerating by factors the training and inference times of neural networks

PID at the LHCb RICH detector

- Tracks extrapolated to the RICH surface
- Hit pattern in ring around the extrapolation point is fed to a CNN
- Classification problem
- Classical solution computationally expensive

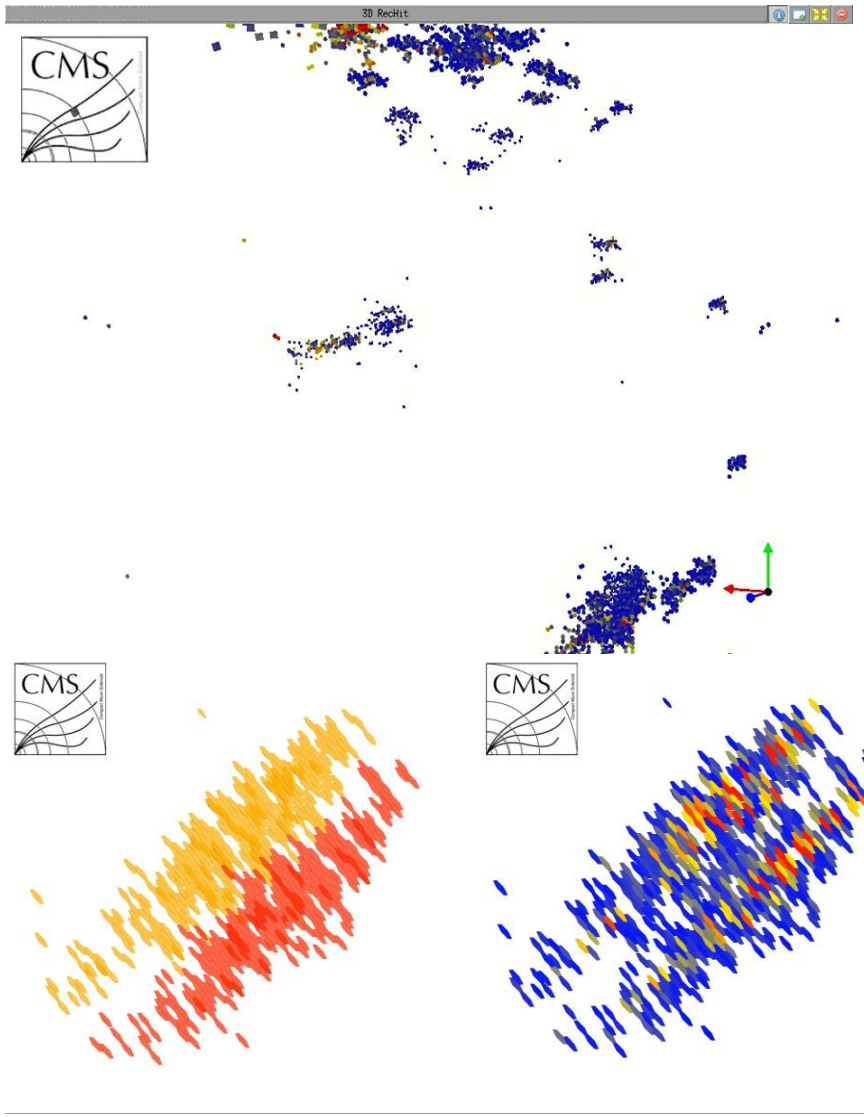


More info: D. Campora Perez

Reconstruction in CMS HGCal

Starting from Run4, High Granularity Calorimeter, a 5-D detector

- usage of unsupervised clustering algorithms
    - optimized to arrange the recorded particle hits into candidate particle showers
- usage of supervised algorithms to accomplish tasks such as particle identification and energy regression on the clustered showers
- Development of de-noising techniques
- Integration of fast inference in CMSSW



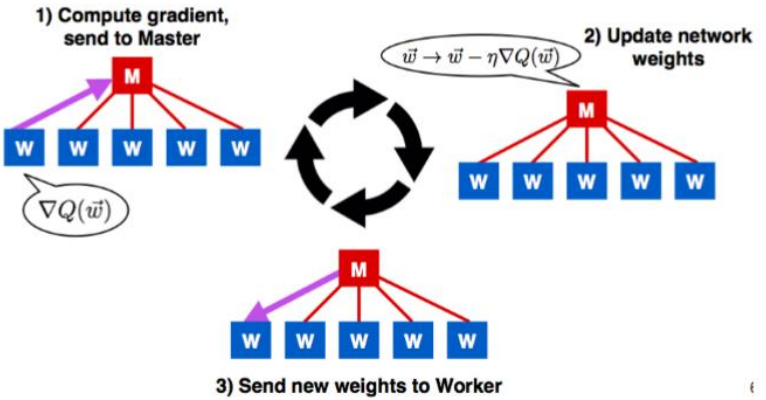
More info: [M. Rovere, V. Innocente, J. Kieseler]

Generative models for HGCal

- Stick to classic calorimeter generation problems
- Use HGCAL as benchmark
- Explore alternative/complementary directions to computing vision
    - Graph Networks for GANs
    - VAEs (with Graph or recurrent models)
- Understand feasibility
- Optimise the model
- Customize solutions

More info: M. Pierini, F. Pantaleo

Distributed Training and Optimization



Full parameter scan is resource/time consuming.

Hence looking for a way to reach the optimum hyper-parameter set for a provided figure of merit (e.g. loss)

Utilize all directions of parallelism in model training and model optimization

Enable exploitation of large HPC facilities as multi-node training machines

More info: J-R. Vlimant, M. Pierini

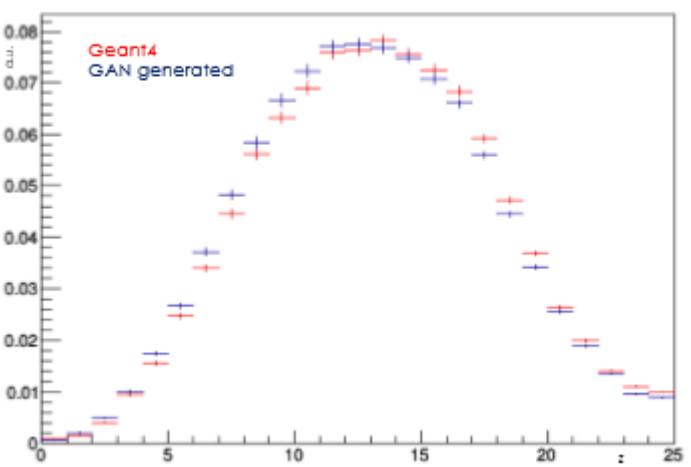
Fast simulation with GANs

Classical electromagnetic shower simulation can be very demanding

Demonstrated three orders of magnitude better performance by running inference on a GAN wrt to classical approach

Quite accurate physics performance

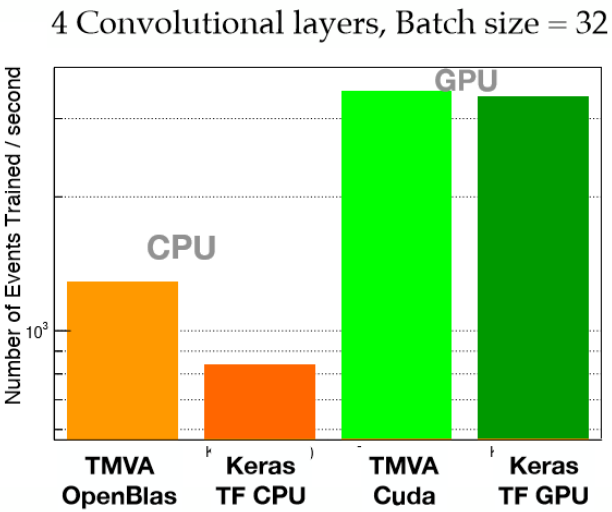
Shower longitudinal section



More info: [F. Carminati, G. Khattak, S.Vallecorsa]

TMVA Deep Learning on GPU

- support for dense, conv. and recurrent layers
- Parallel on CPU (tbb) and GPU (CUDA)
- Excellent performance and high numerical throughput
- ROOT has direct access to data, so no conversion is needed and a ML model can be directly evaluated (1.5x over specialized libraries)
- Plan to make better use of CUDA SDK libraries to improve training and inference
- Plan to provide support for Recurrent layers and GANs on GPUs
- GPU could also be used later for other use cases within the ROOT software: e.g. fitting and modelling for statistical and parameter estimations



More info: L. Moneta

Conclusion

- We are working in an exciting era for HEP and computing
- R&D is needed now more than ever as new ideas can really make the difference, by enabling to reach their full physics potential (or more) in a sustainable fashion
- Looking forward to gaining more expertise to operate/virtualize heterogeneous farms, out of our usual comfort zone (thanks to IT-CM)