# Discussion 2: use of alternative hardware



Maurizio Pierini




European Research Council

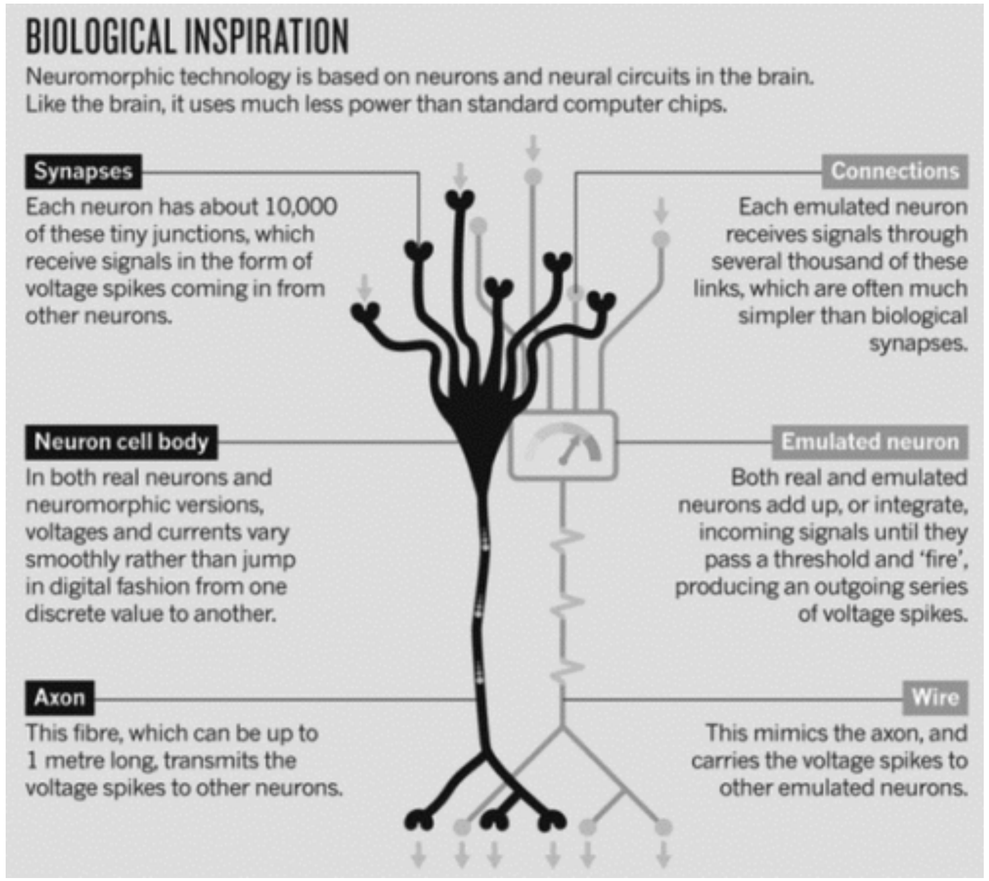
# What this is about

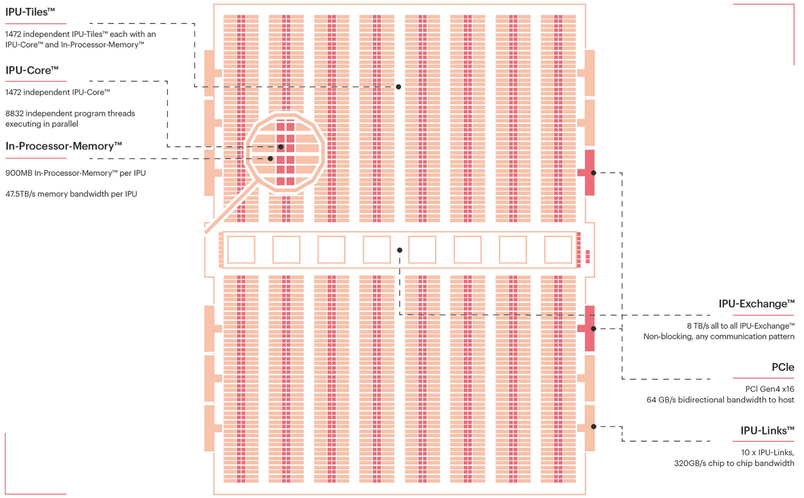
*With new technologies emerging (e.g., deep learning), scientific computing environment are becoming more and more heterogeneous. Several parallel computing devices (GPGAs, GPUs, etc) can be exploited to accelerate traditional algorithms and include deep learning components in the processing workflows. A wide set of dedicated computing devices (TPUs, IPUs, ...) are targeting specific use cases, e.g., convolutional neural networks, recurrent networks, graph networks. Neuromorphic computing opens the possibility to use spiking neural networks for signal processing. On a longer term, quantum computing might offer interesting alternatives to solve large combinatoric problems. With private companies dictating the direction followed by innovation, big scientific collaborations might have to adapt their data processing to follow this trend, which could offer specific advantages.*

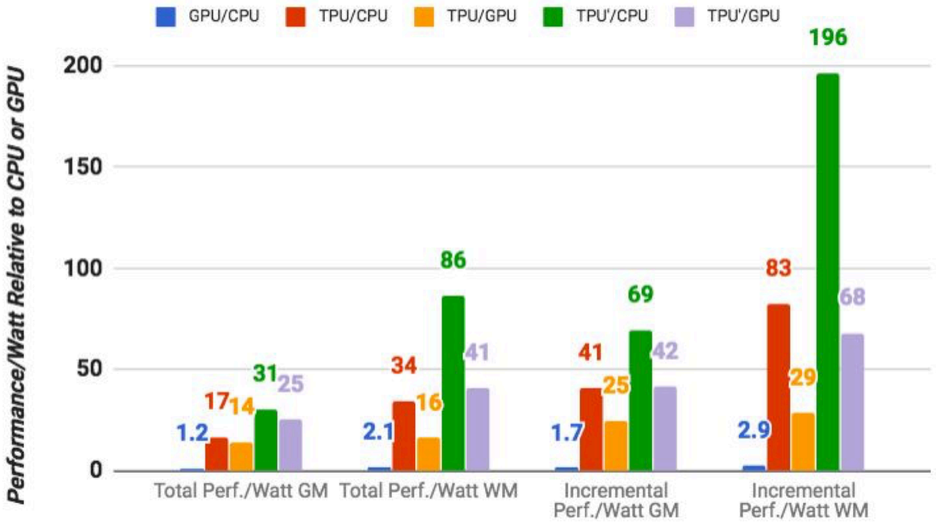
# The hardware landscape is changing

- *In the past, CPUs were the main working horses used for Big Science application*
- *For real-time applications, ASICs on custom electronics were a typical choice*
- *Then parallel computing changed the landscape*
  - *FPGAs replaced ASICs as a common choice for RT (at least in HEP)*
  - *GPUs (and FPGAs?) offer now speed up opportunity thanks to parallelism*
  - *Big investments towards Deep Learning and Quantum computing will reshape the landscape around us and we should adapt*

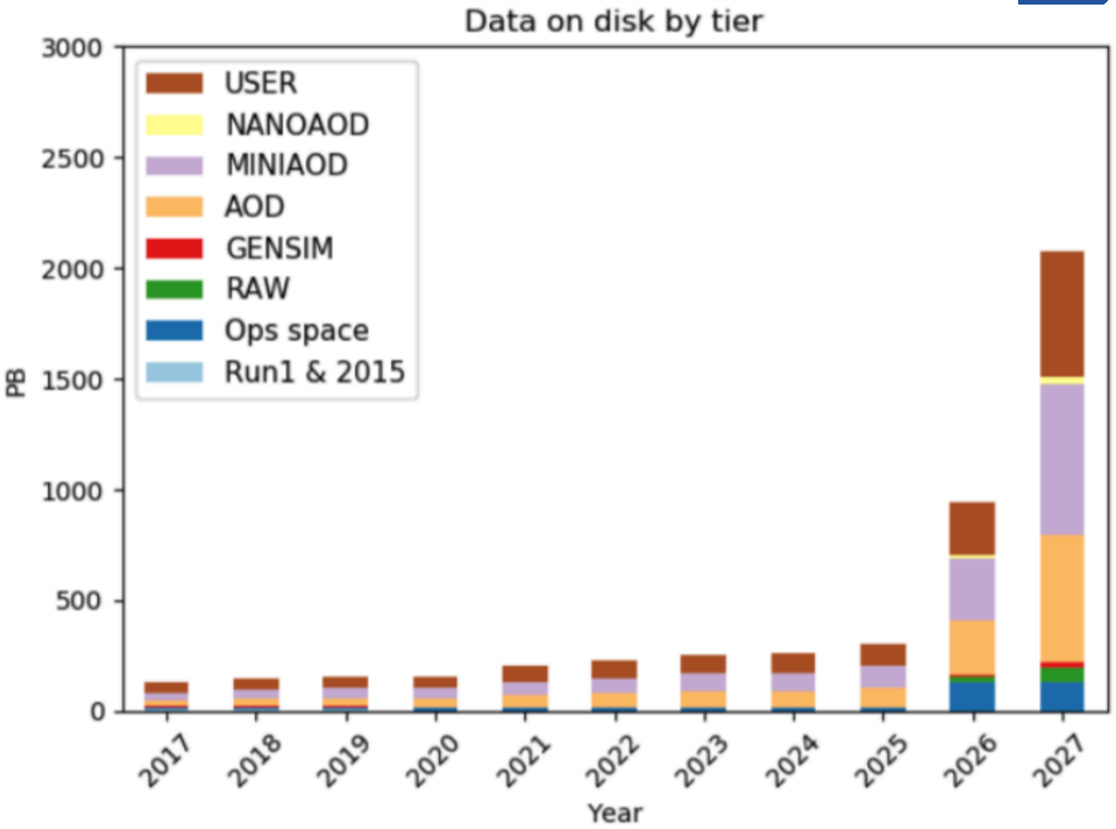
# New Hardware: examples

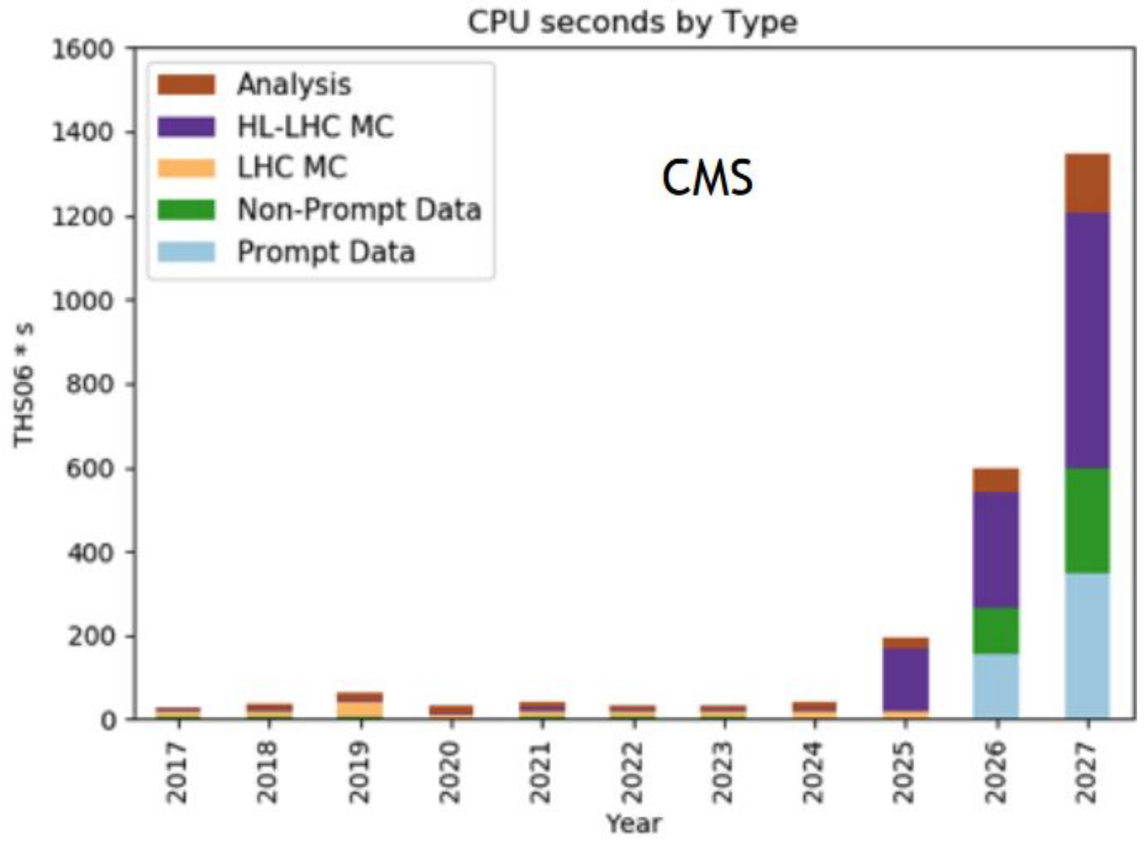


# The big problem

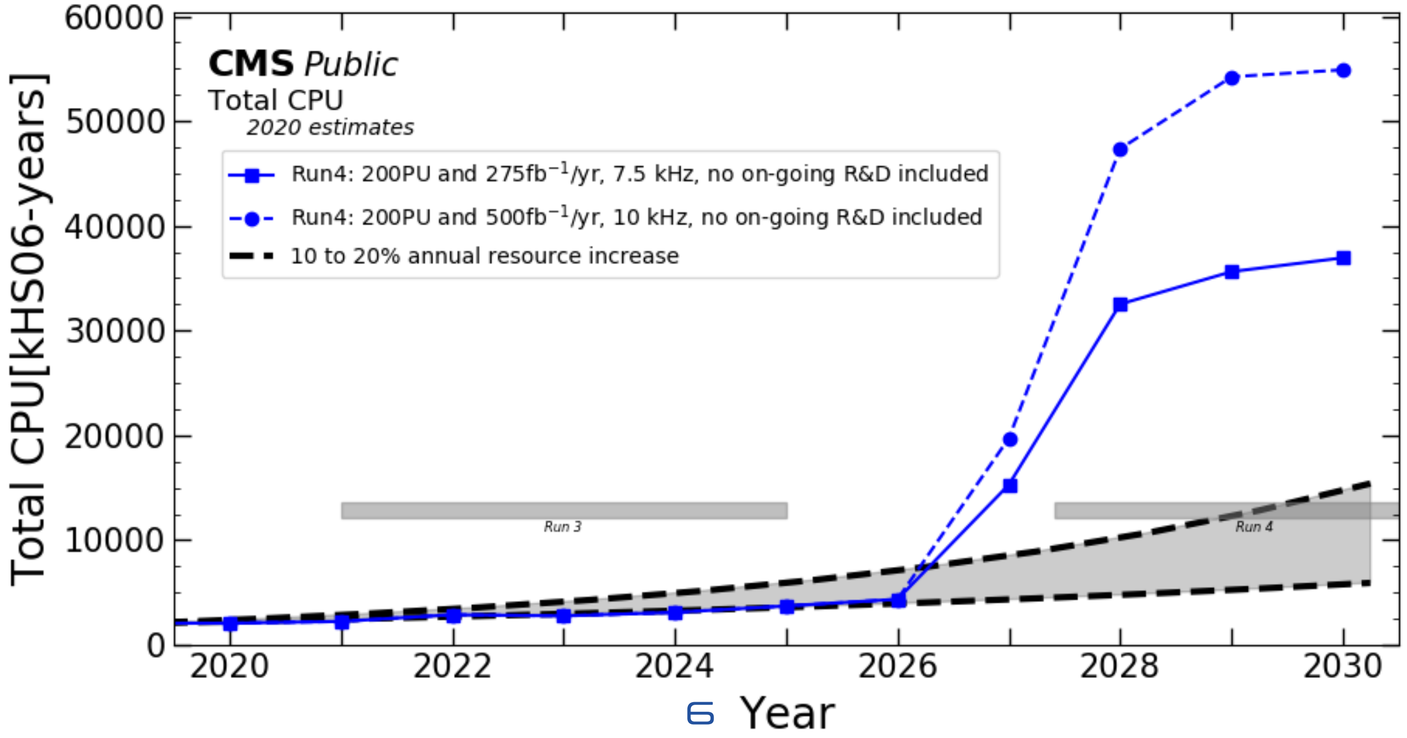



- *We could take more data, but we don't*
  - *because we cannot process them (CPU taken by MC)*
  - *because we could not generate the corresponding MC*
  - *because we cannot store them*
  - *because we cannot store the corresponding MC*

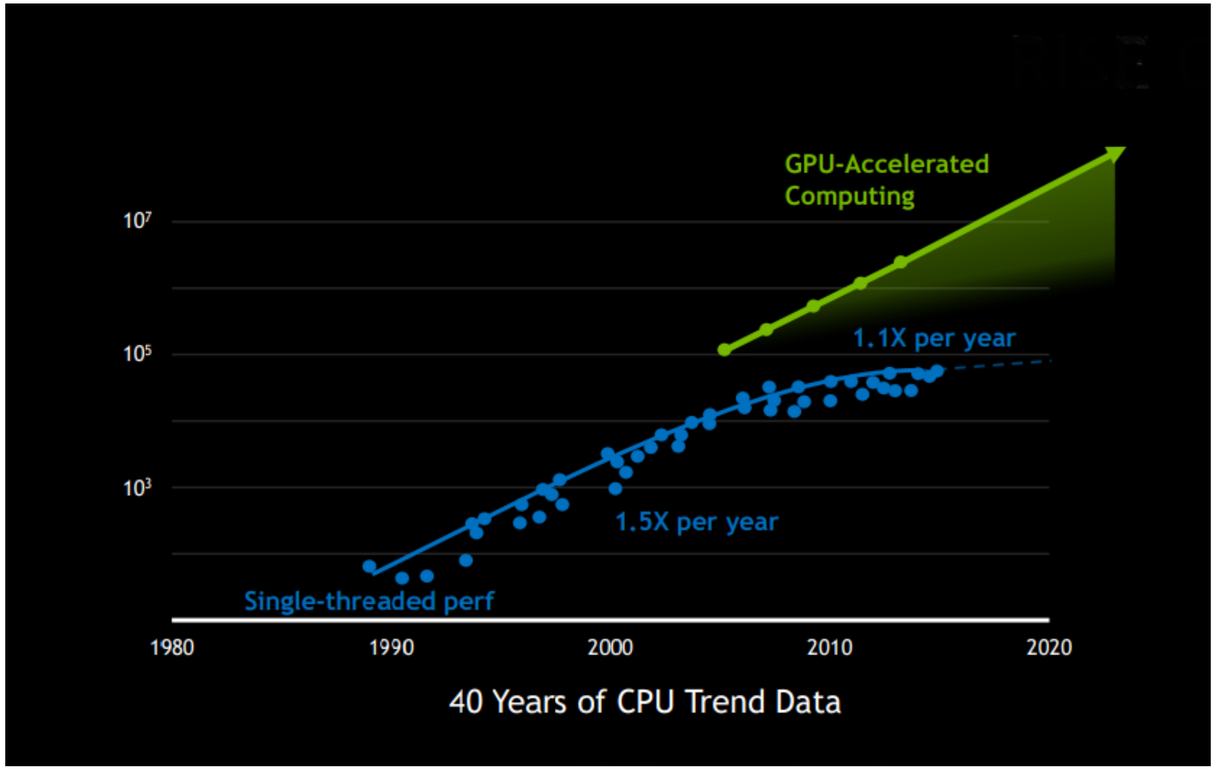
# Our resources don't scale up

- *Scientific challenges grow in complexity*
- *Brute-force computing power increase at fixed cost worked so far*
- *It will not in the past (Moore's law saturation)*
- *Example: projected HL-LHC computing power is x10 off wrt needs*
- *We need to look at new resources outside*
- *Those resources don't target us as main customers*



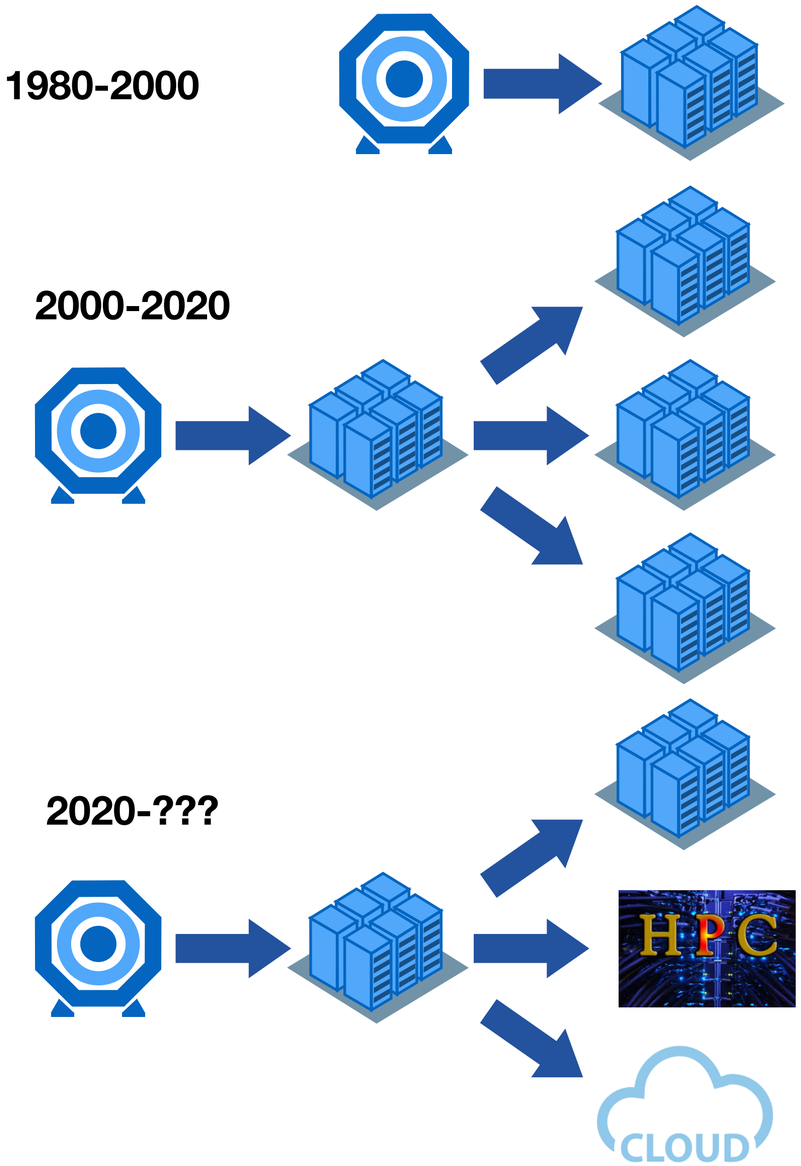
# HPC sites evolution

- *Computing environment around us is evolving*
  - *HPC sites*
  - *Cloud platforms*
- *These sites are investing massively on GPUs*
- *Cloud providers have even more heterogenous environments*
  - *FPGAs, TPUs, IPUs*
- *How can we take advantage of that?*

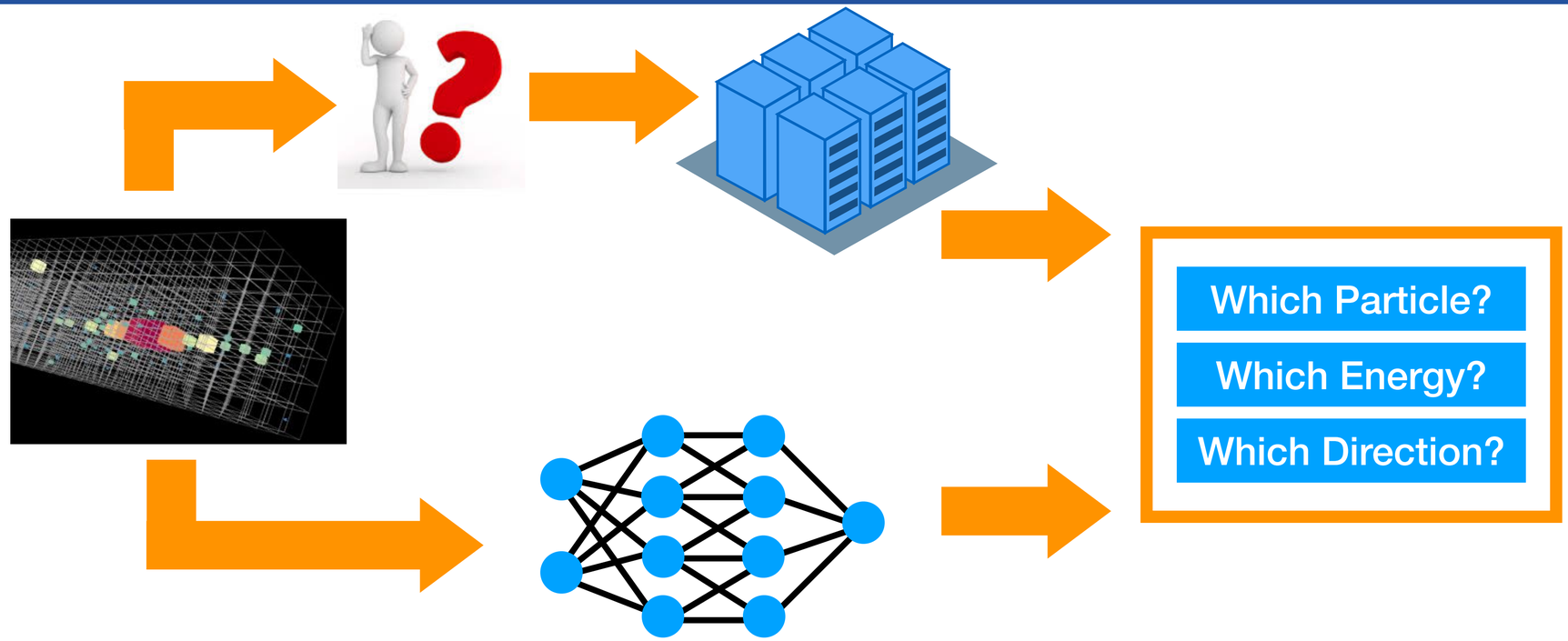


# The (HEP)_computing evolution

- *We are transitioning from edge to cloud computing*
- *We are not anymore the leading figure (compare CERN computer centre to cloud providers)*
- *It is likely that scientific computing in the future will be largely outsourced to HPC sites and (opportunistically?) to Cloud providers*
  - *e.g., Amazon cloud integrated to LHC Computing Grid as a virtual extension of FNAL Tier*
- *This opens the room to new workflows, which would exploit hardware heterogeneity*

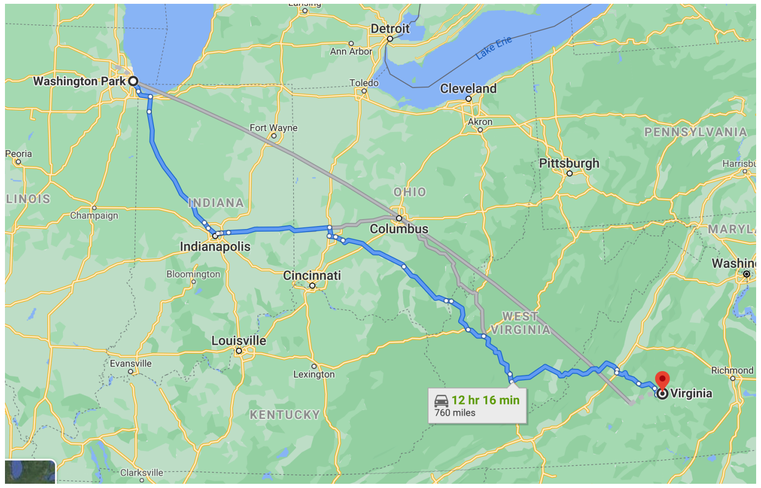


# Deep Learning is central to this

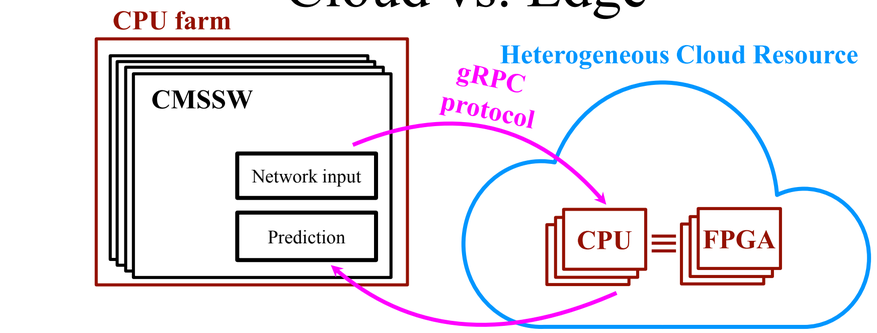


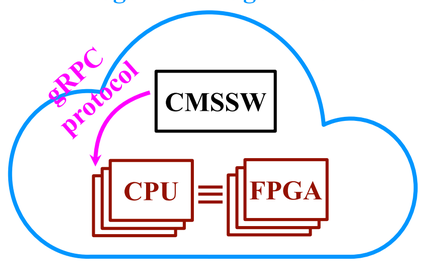

- *Big push to dedicated infrastructures*
- *GPU/CPU ratio in HPC sites increasing to accommodate more and more training & inference needs*
- *We might have to make ourselves more DL compliant to benefit of these resources*

# An example: remote computing in real time



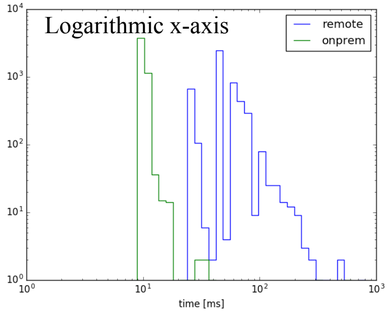

## Cloud vs. Edge

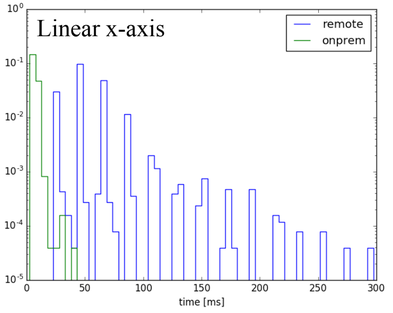



- Cloud service has latency
- Run CMSSW on Azure cloud machine → simulate local installation of FPGAs ("on-prem" or "edge")
- Provides test of ultimate performance
- Use gRPC protocol either way

## SONIC Latency




- Remote: cmslpc @ FNAL to Azure (VA), ‹time› = 60 ms
  - Highly dependent on network conditions
- On-prem: run CMSSW on Azure VM, ‹time› = 10 ms
  - FPGA: 1.8 ms for inference
  - Remaining time used for classifying and I/O

# Towards the discussion

- *These are only a few considerations to start with*
- *Taken from my personal experience with HEP and in particular LHC*
- *I would be very interested to hear about experience from other fields*
  - *long-term needs (storage and computing)*
  - *migration to heterogenous environments*
  - *edge computing vs cloud computing*
  - *the role of Deep Learning in all of this*
- *Based on this, we should come out with a list of interesting hardware options*