Fast Machine Learning

Sergo Jindariani (Fermilab)

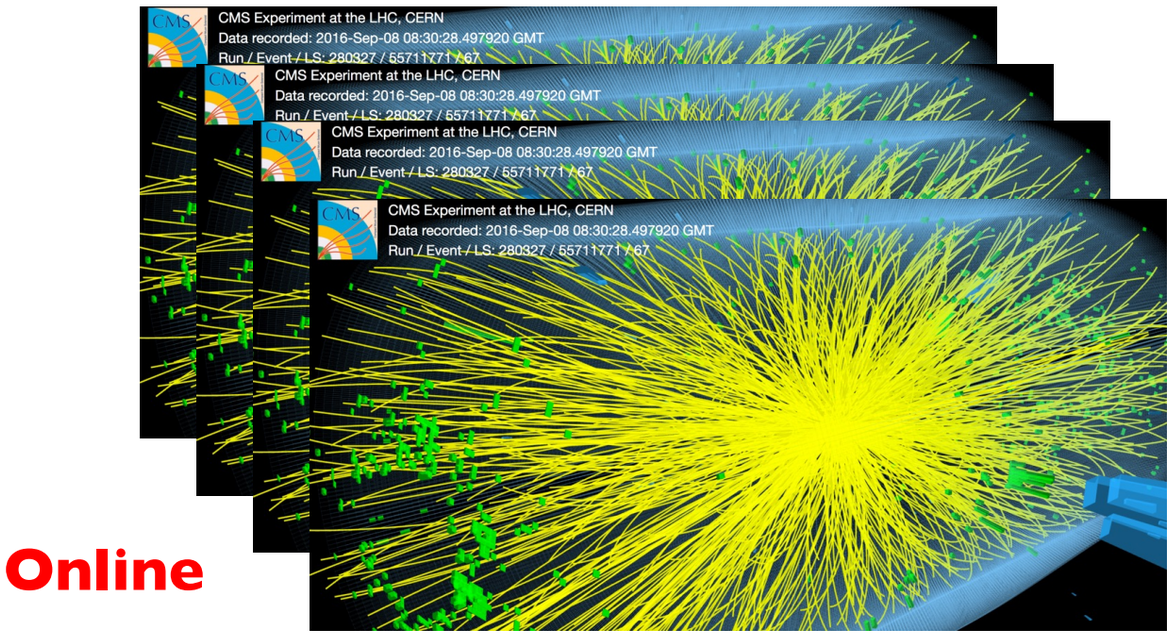
On behalf of many people from many institutions working on accelerated ML in physics

https://fastmachinelearning.org/

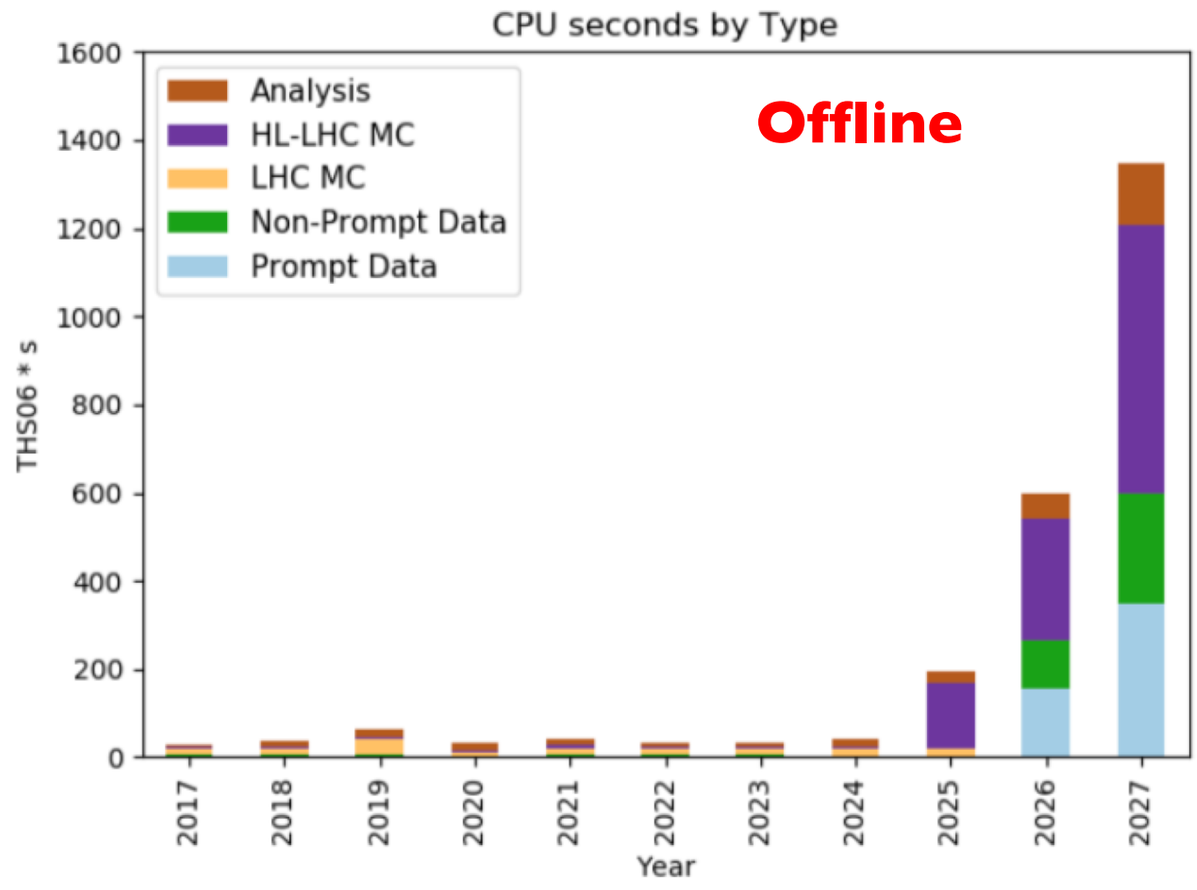
The Challenge: consider LHC example

Detector are becoming bigger and more complex (x10-20 number of channels)

Data rates are increasing (x10 collisions per second)



40 Million times per second

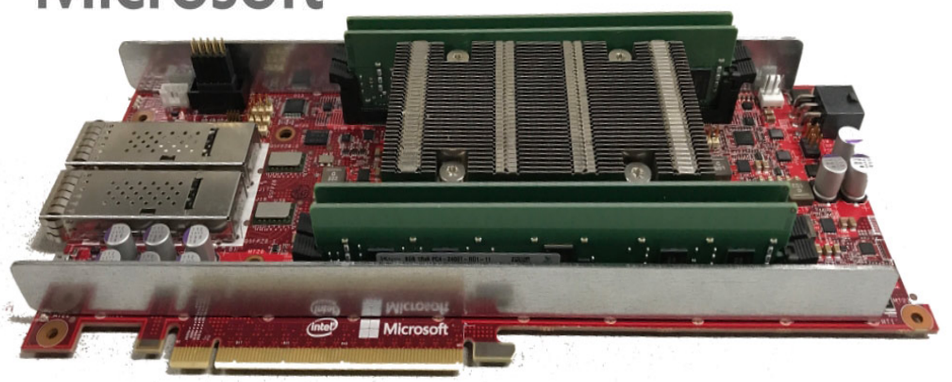


- Traditional approach with CPUs cannot keep up with this scaling of Data*Complexity
- Can we solve the problem by using ML algorithms in specialized Hardware?

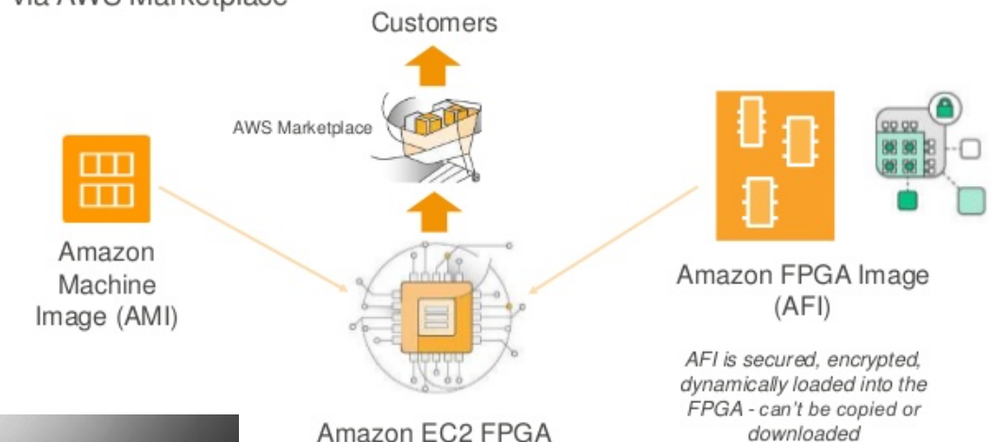
Machine Learning in hardware

Industry is already deeply invested into specialized hardware for ML (based on FPGAs, ASICs)

Microsoft

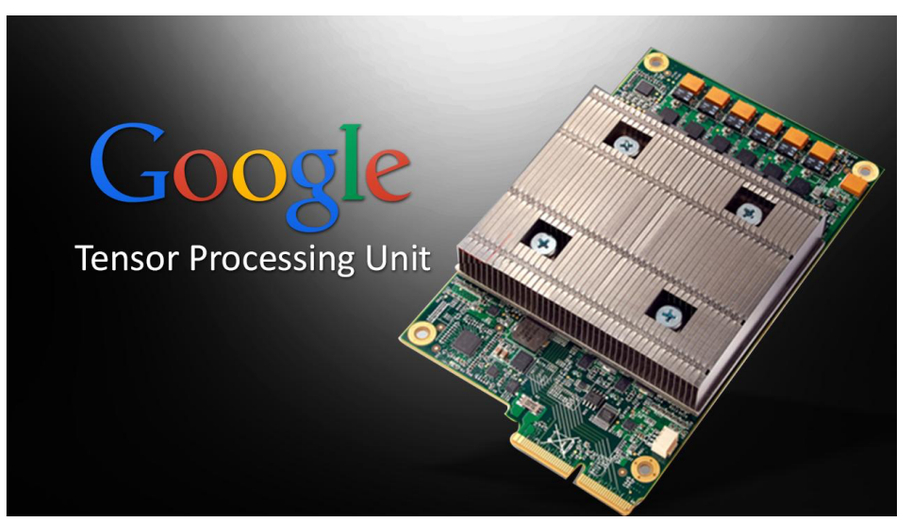


Delivering FPGA Partner Solutions on AWS
via AWS Marketplace



INTEL® FPGA ACCELERATION HUB

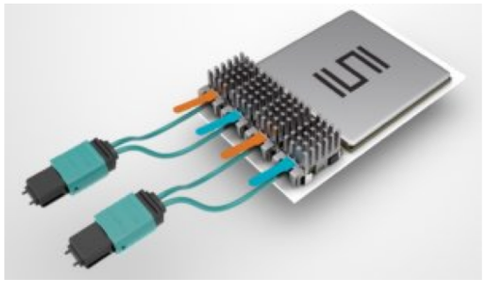
The Intel® Xeon® Acceleration Stack for FPGAs is a robust framework enabling data center applications to leverage an FPGA's potential to increase

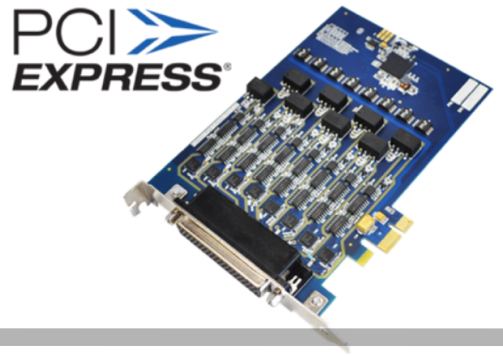


Understanding Timescales

Optical to electrical to fpga pins

Stream through PCIexpress





<1μs

1ms-1s

All FPGA systems

Hybrid GPU/CPU+FPGA systems

LHC Physicists

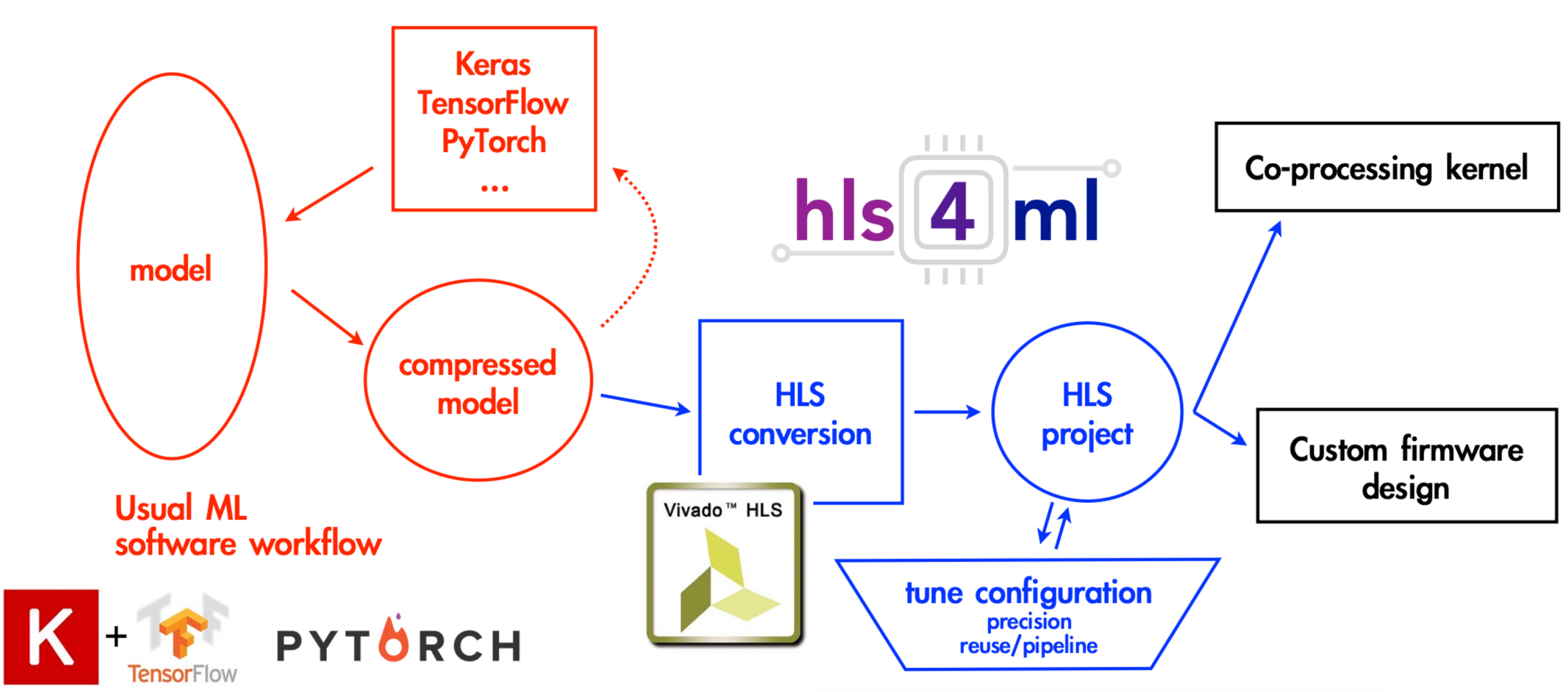
Other Physicists

Industry

What is hls4ml

User-friendly tool to build and optimize ML models for FPGAs:

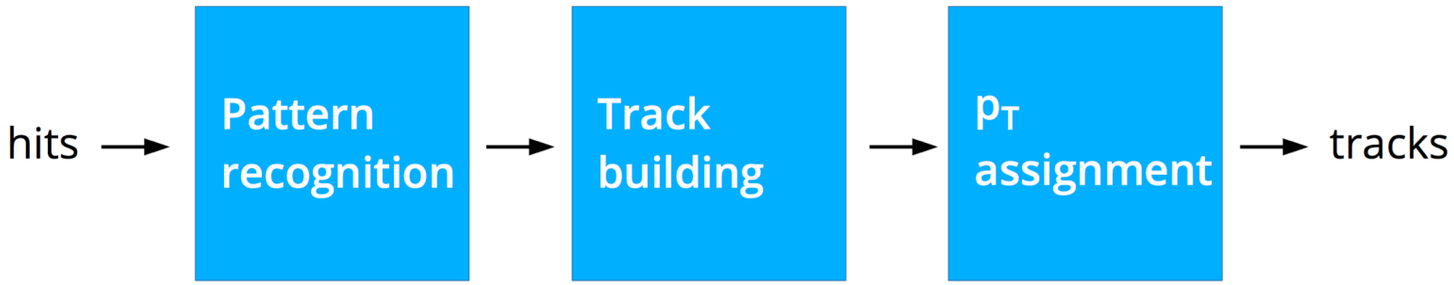
- Reads as input models trained with standard ML libraries
- Uses Xilinx HLS software
- Comes with implementation of common ingredients (layers, activation functions, binary NN ...)



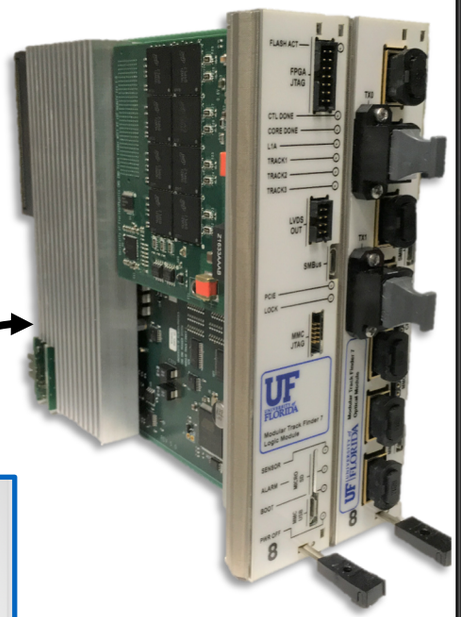
https://fastmachinelearning.org/hls4ml/

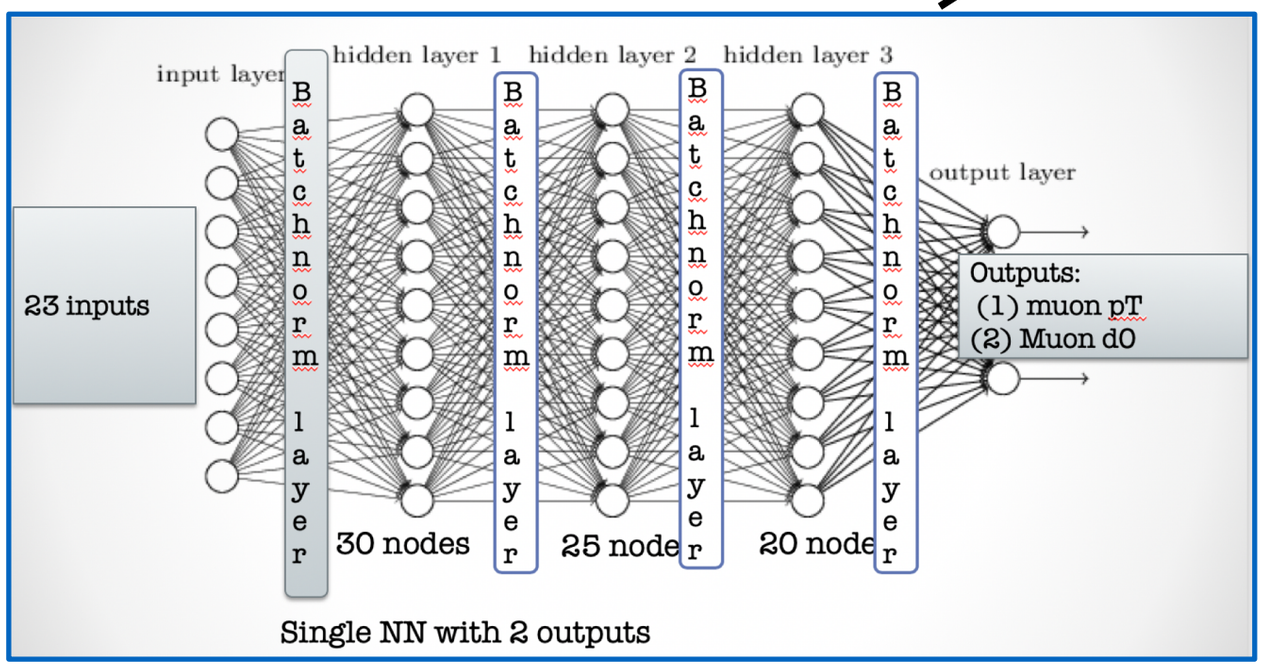
Deep Learning in Level-1

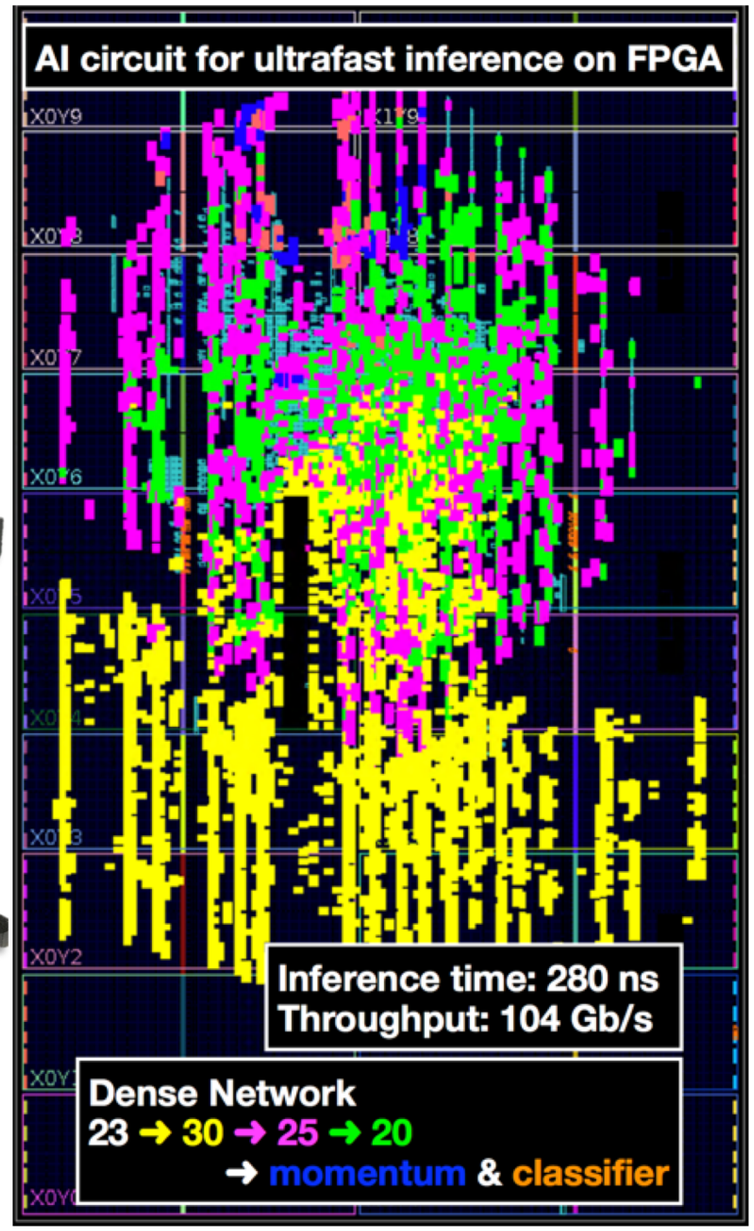
Level-1 Muon Reconstruction Steps



Implementation in Virtex-7 FPGA in MTF-7 uTCA boards



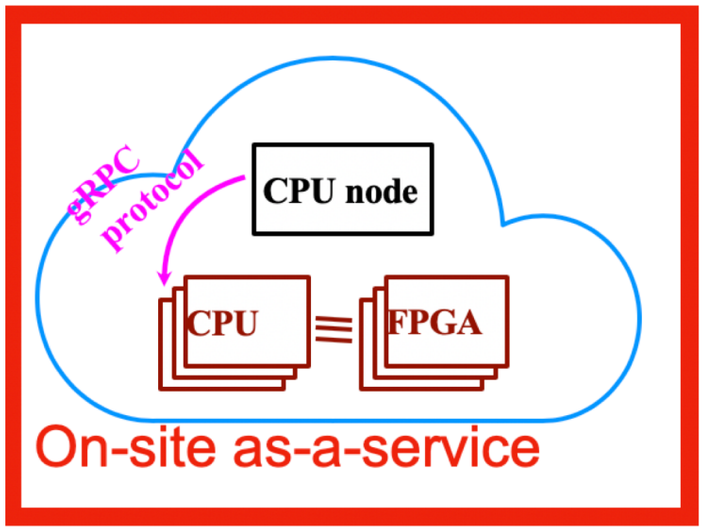




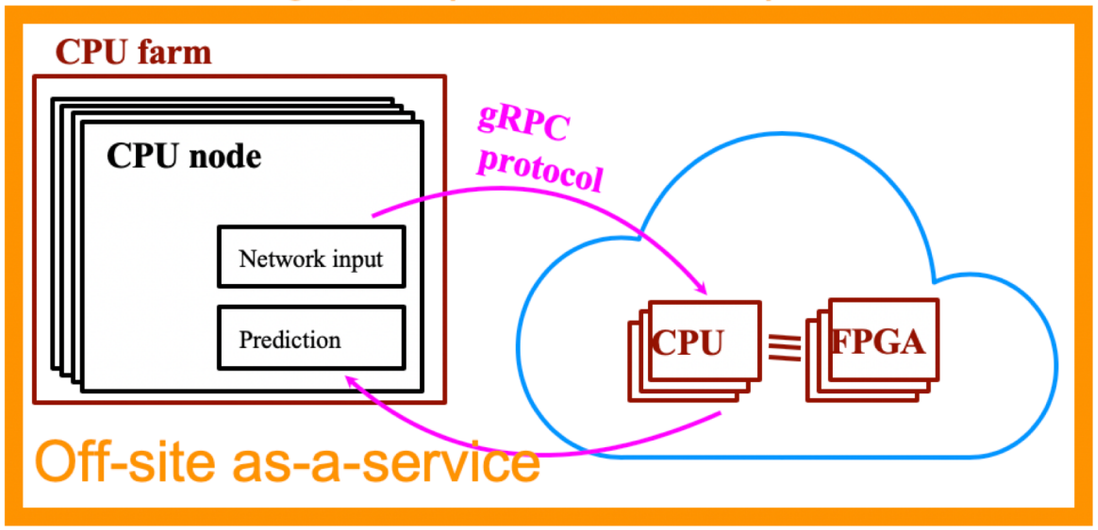
Deep Learning in HLT/Offline

- timescales at the level of milliseconds or seconds
- Previous systems have been CPU only
  - New systems will likely be heterogeneous (FPGAs/GPUs…)

Low latency Triggering



Larger latency but still large throughput (future slides)



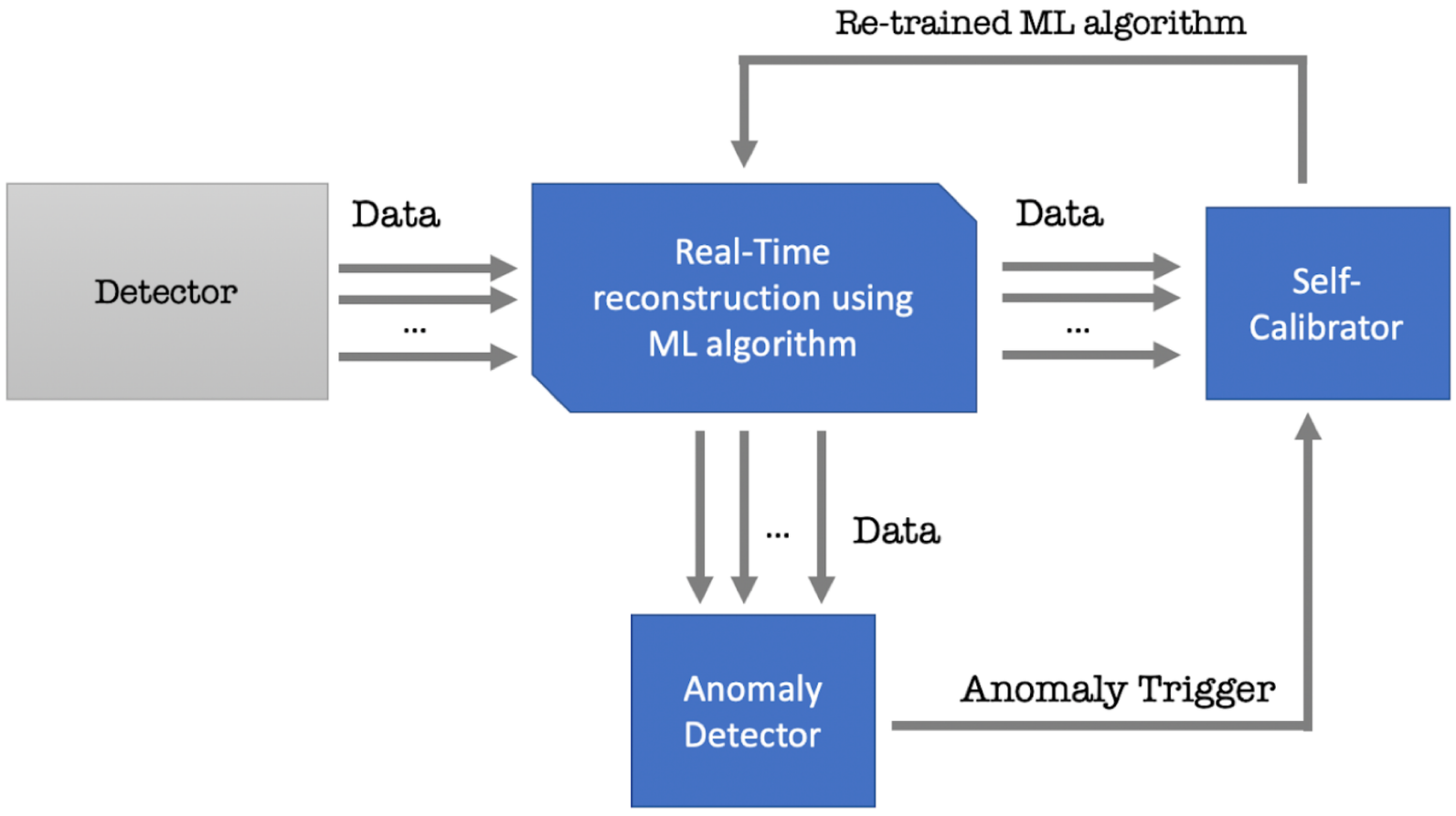
When latency not critical element : can go off-site to the cloud

For timescales in <100ms, this is not an option

Test	Inference time
local	• **10 ms** (~2 ms on FPGA + classifying, I/O) • Meets HLT latency requirement
remote	• **60 ms** (includes travel latency) • (4/10/100) faster than CPU-only computations

System Controls

Dedicated presentation about Accelerator Controls
Expanding into detector controls and self-sustainable detectors.



Major work needed on both algorithms and hardware side to understand the reach and the limitations of this approach.

Quickly Growing Community

- This effort has started to address challenge of data reconstruction at the LHC

- Since then, we are quickly identifying other cases with the same issues
  - Neutrino Event Reconstruction
  - Fixed Target Experiments
  - Observational Cosmology
  - GW detection
  - Accelerators

- New ideas:
  - New architectures, algorithms, hardware platforms
  - Develop ML ASICs tailored for high-rad or cold-temp environments?

- Have extended our collaboration to incorporate everybody
  - Inaugural workshop can be found here https://indico.cern.ch/event/822126
  - You too can join our Fast Machine Learning effort