Real Time Deep Learning for Future Detectors

Sergo Jindariani, Nhan Tran (Fermilab)

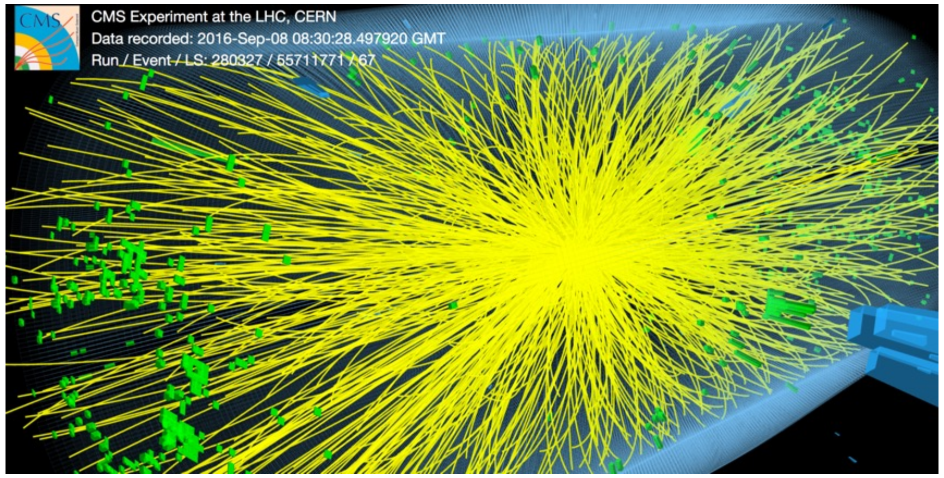
Snowmass white paper planning meeting: Detectors, May 2020

Disclaimer:

- the subject of this talk is about more than just colliders. It is a capability that HEP has and can be used elsewhere. However, we will use collider examples to illustrate various points.

- This is a rapidly growing effort with many contributors: www.fastmachinelearning.org

Challenge: Rates and Complexity

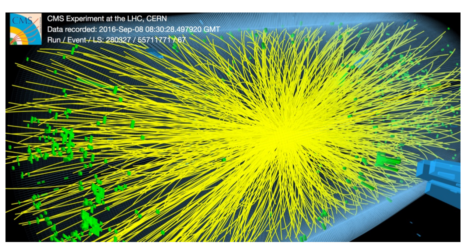


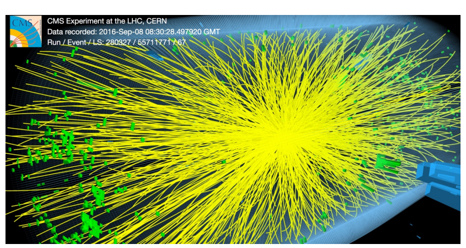
In HL-LHC the average number will go up to <PU>=200

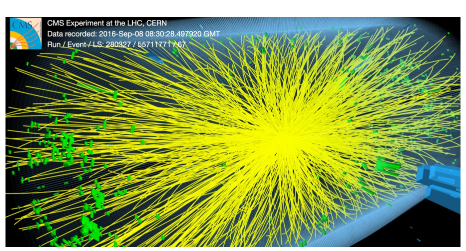
FCC-hh <PU> = 800-1000

Number of channels

	CMS Calorimeters	Pixels
'Phase-0'	~80k	66M
'Phase-1'	~90k	123M
'Phase-2'	6.5M	2B



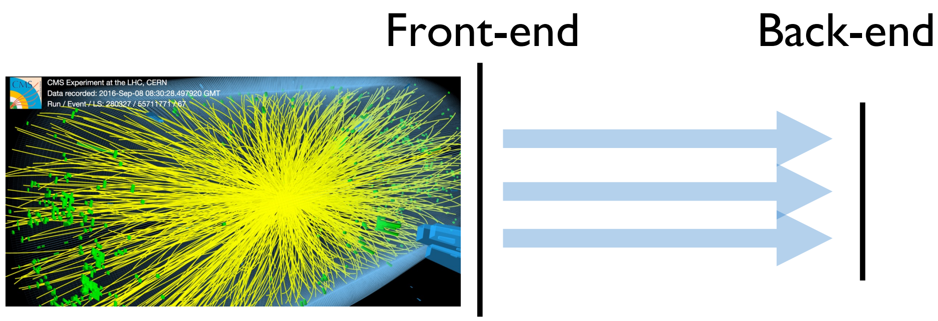




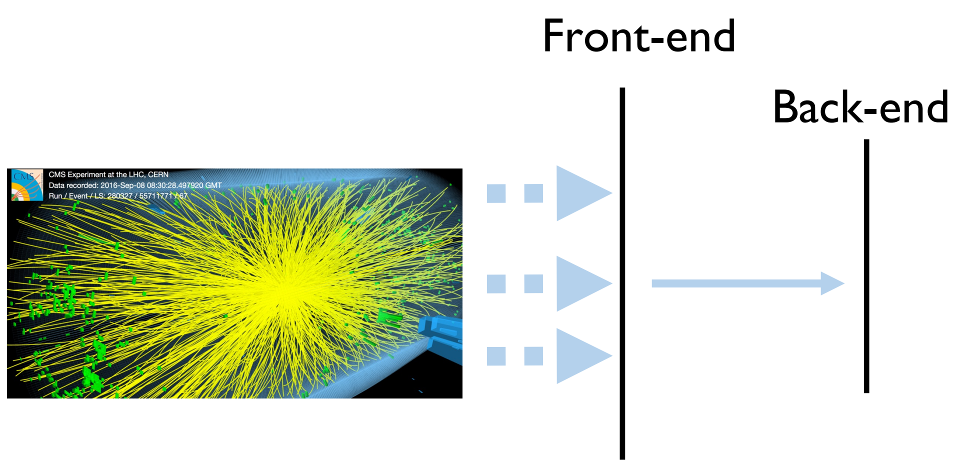
- At Level-1 Trigger new event coming every 25 ns
- 100s of Tbps
- Total latency budget ~10us

Future proton colliders will significantly exceed these requirements

Solution?



Option 1: Send more data out
- requires very high bandwidth, low power, rad hard links
- requires ultra fast data reconstruction

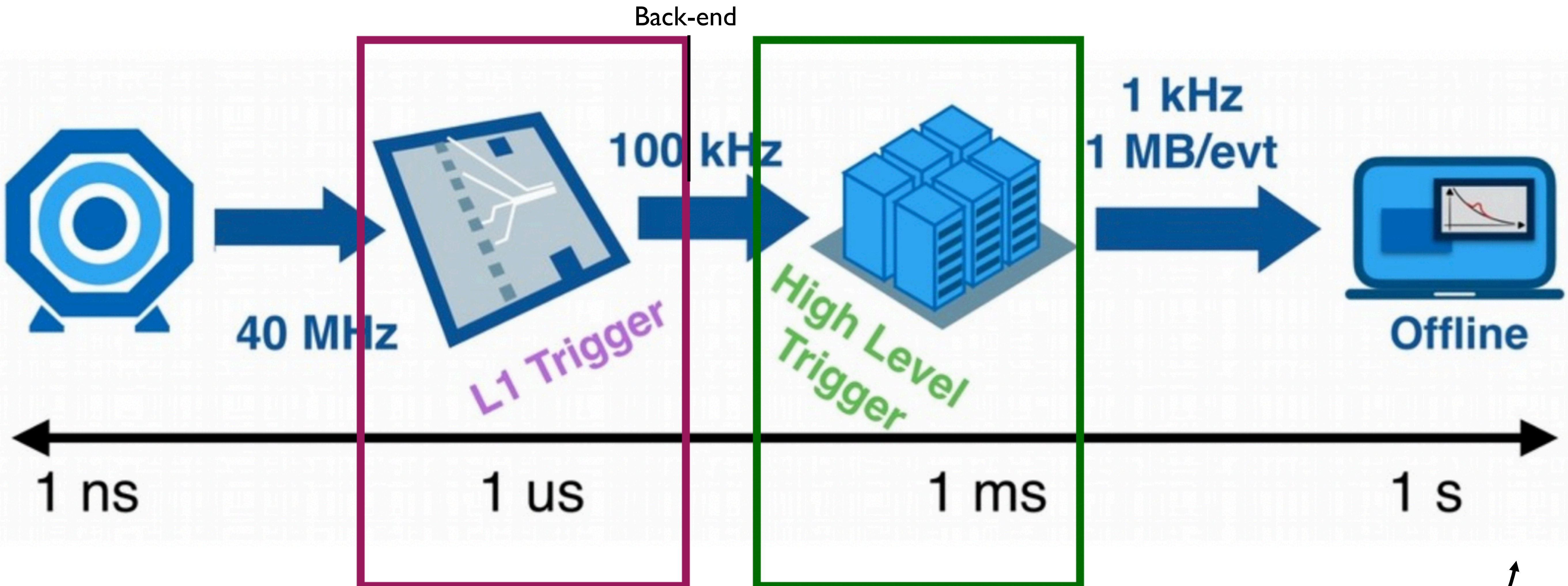


Option 2: put more logic to frontend
- Local tracking/clustering, efficient data encoding
- More logic = more power. Again, has to be rad hard or cold (e.g. LAr)
- No FPGA => ASIC

Can machine learning help with one or both options?

(in principle ML algorithms are naturally suited for this job)

Data Processing



Employ ML algorithms in Level-1 hardware

I will mostly focus on this

Rewrite CPU-intensive HLT/Offline tasks using ML

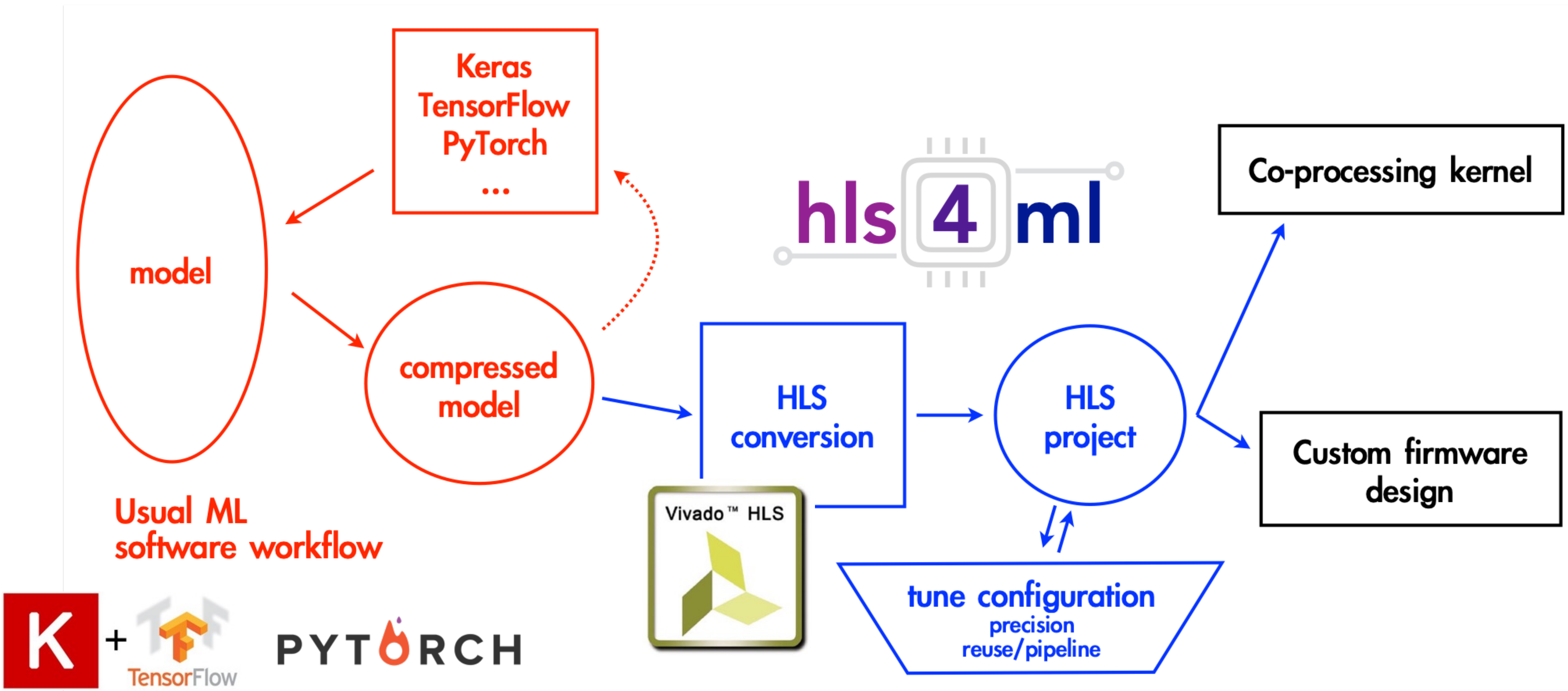
Accelerate using co-processors (e.g. FPGAs)

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 ...)

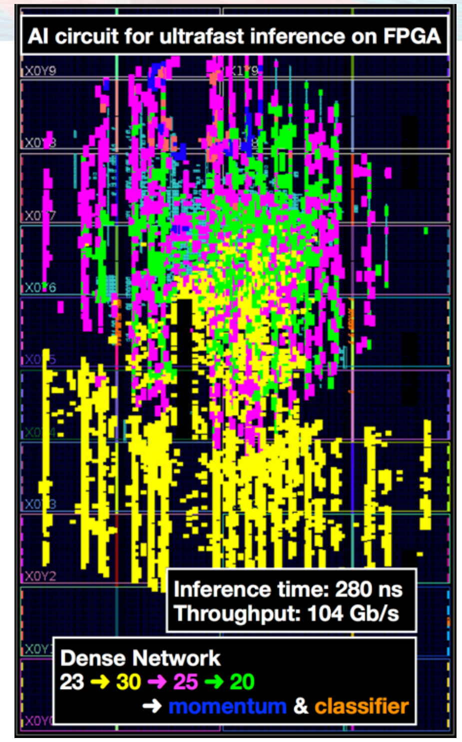


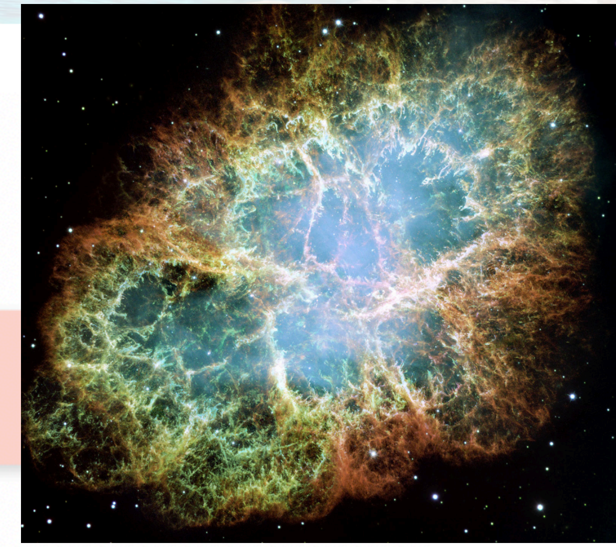
ONNX

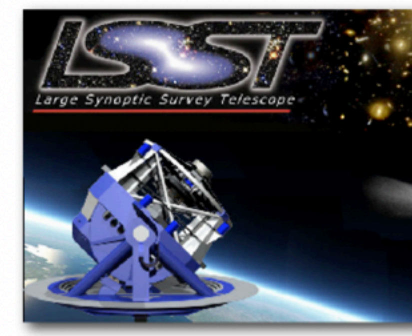
https://fastmachinelearning.org/hls4ml/

Quickly Growing Community

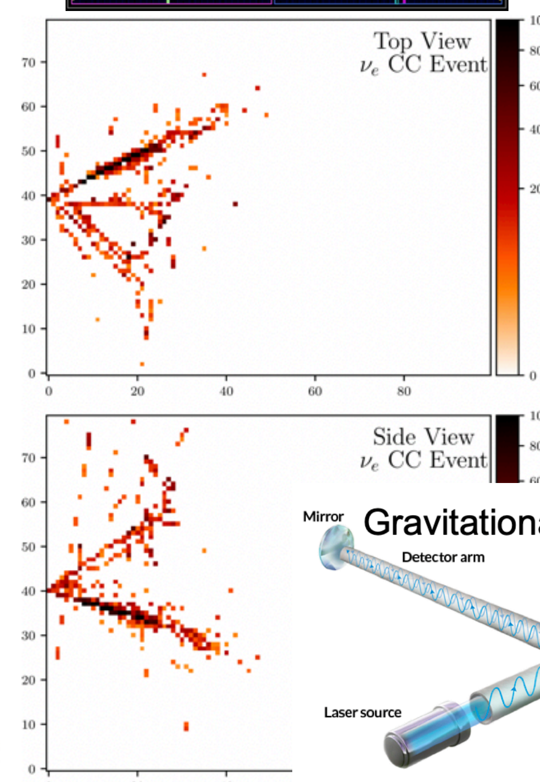
- LHC Efforts:
  - Prompt and Displaced Muons
  - Calorimeter Clustering
  - Tau Identification
  - Jet Substructure/Tagging
  - Anomaly detection

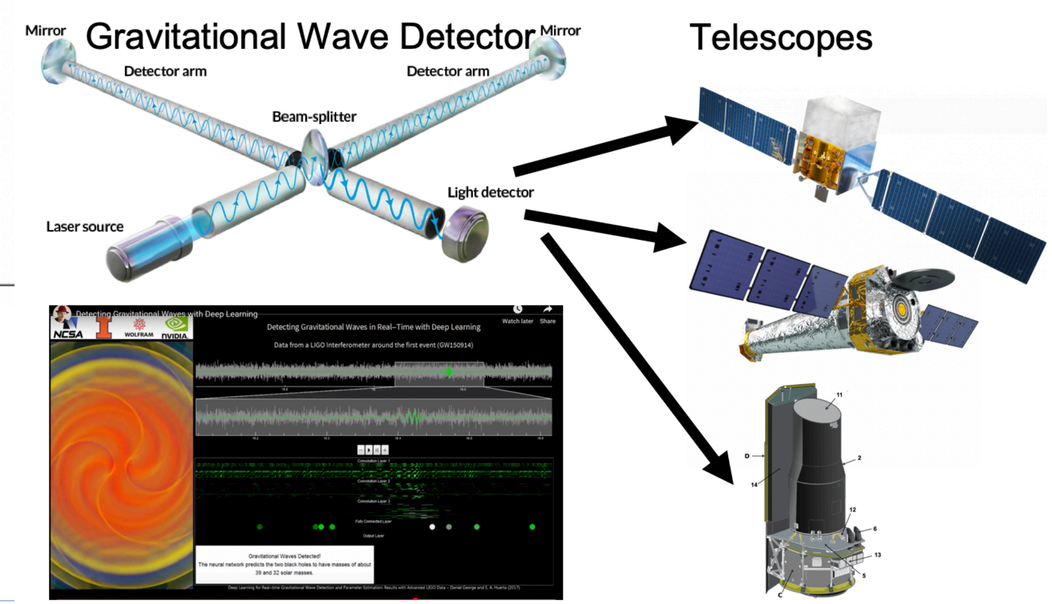


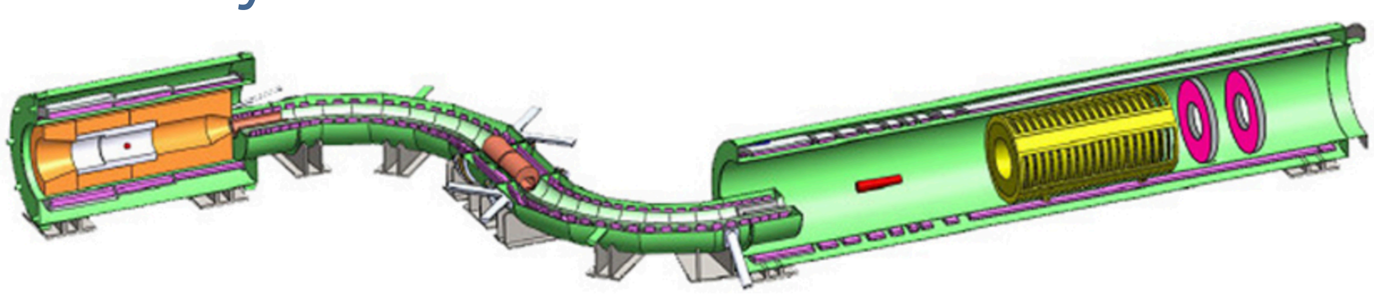




- Beyond-LHC efforts
  - Neutrino Event Reconstruction
  - Fixed Target Experiments
  - Observational Cosmology
  - GW detection
  - System controls





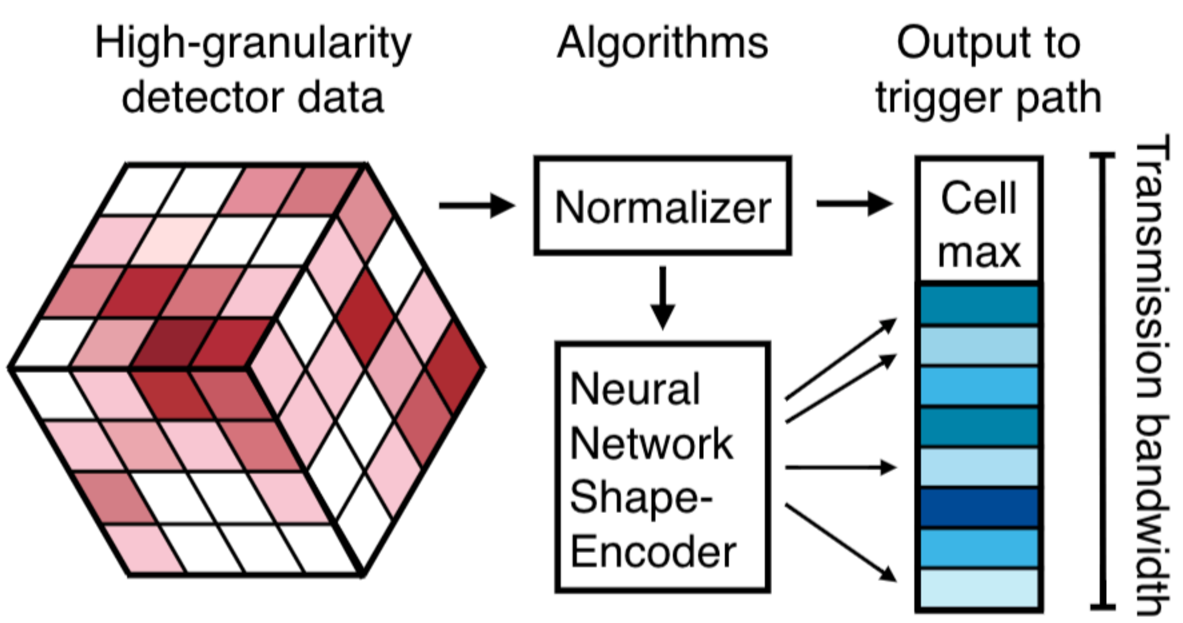


hls4ml in ASIC

Future: Can we implement more sophisticated ML algorithms in the on-detector ASICS?

- smarter data compression
- local tracking/clustering
- anomaly detection (channel sync, time drift, other issues?)

There must be other applications…



Technology: LP CMOS 65nm
Power: 280mW (@25ns this is 7nJ per inference) — 10x better than FPGA
Network: 4448 multiplications, 2286 parameters
Area: 2.5 mm^2

System Controls

Can we have ML based accelerator and detector controls?

- Example use case pursued for the Fermilab accelerator complex

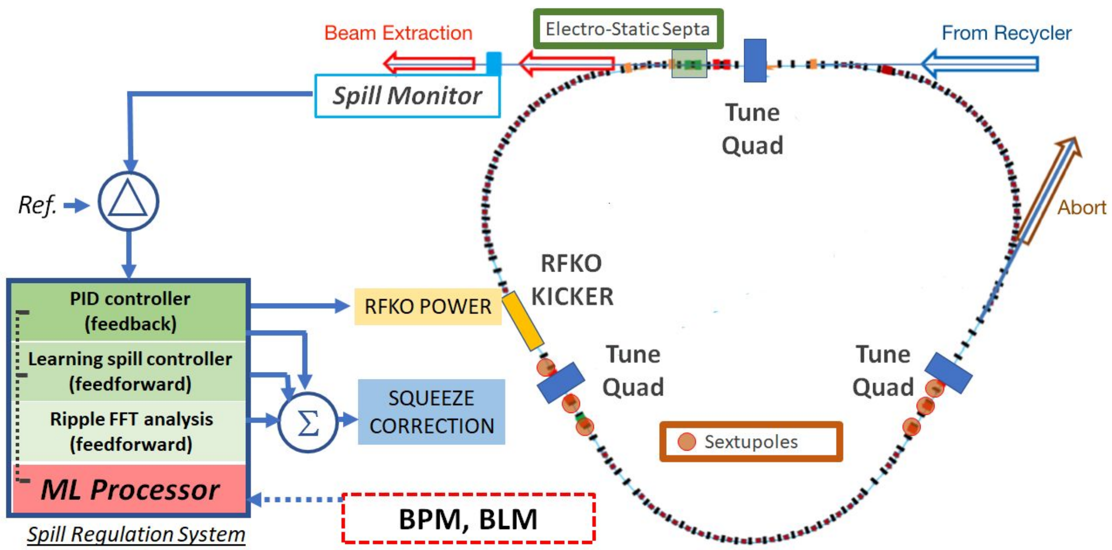
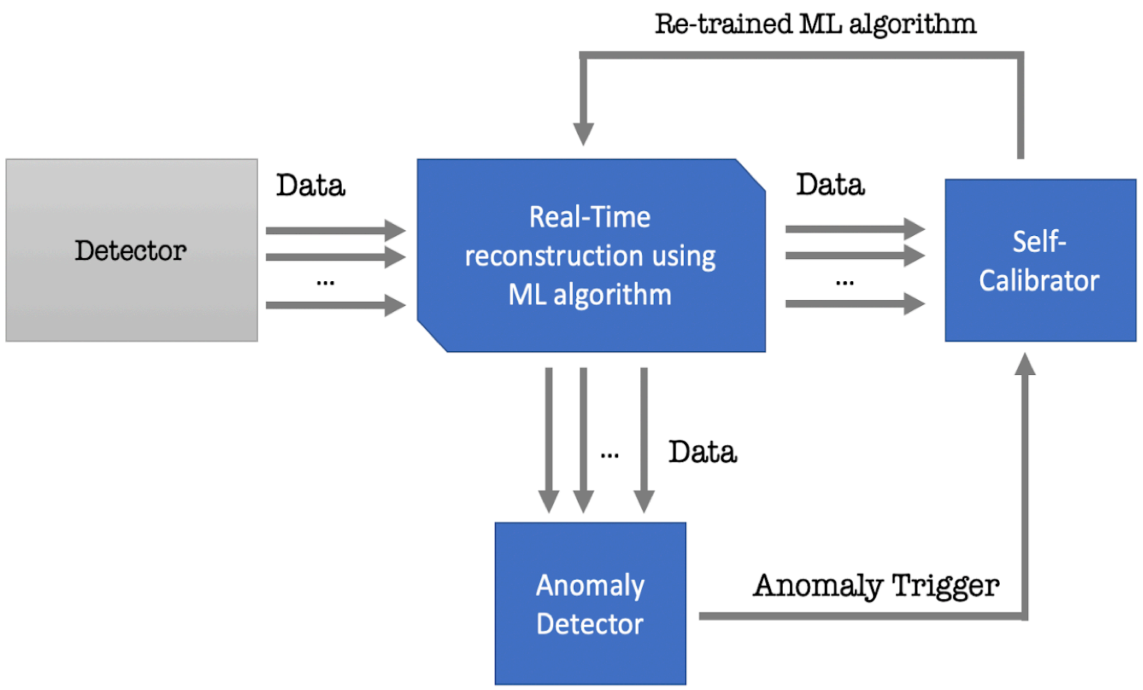


Figure 2: Delivery Ring and control loop for the Spill Regulation System

Take it further: the idea of future intelligent systems

- Detect an anomaly using online data and re-calibrate the response
  - Requires online training
  - many challenges on both algorithm and HW side



Goes beyond HEP!

Summary

- The size, rate and complexity of future detectors presents a significant challenge for online and offline data reconstruction

- Machine Learning has a potential to help solve some of these problems
  - Note that I focused almost entirely on inference, fast learning is a whole different topic, but ties to autonomous systems

- Foundation for this effort has been built and we are looking to extend it in many directions across HEP and beyond it

Extras

A quick how-to

- Easy to **install** via `pip`: `git clone … && cd hls4ml && pip install .`
- Easy to **configure** through yaml config file

  `Inputs:` your trained model
  `Precision:` inputs, weights, biases, …
  `ReuseFactor:` how much to parallelize
  `Strategy:`
  `Resource` for large NN
  `Latency` for pipelined-based code
  for small NN

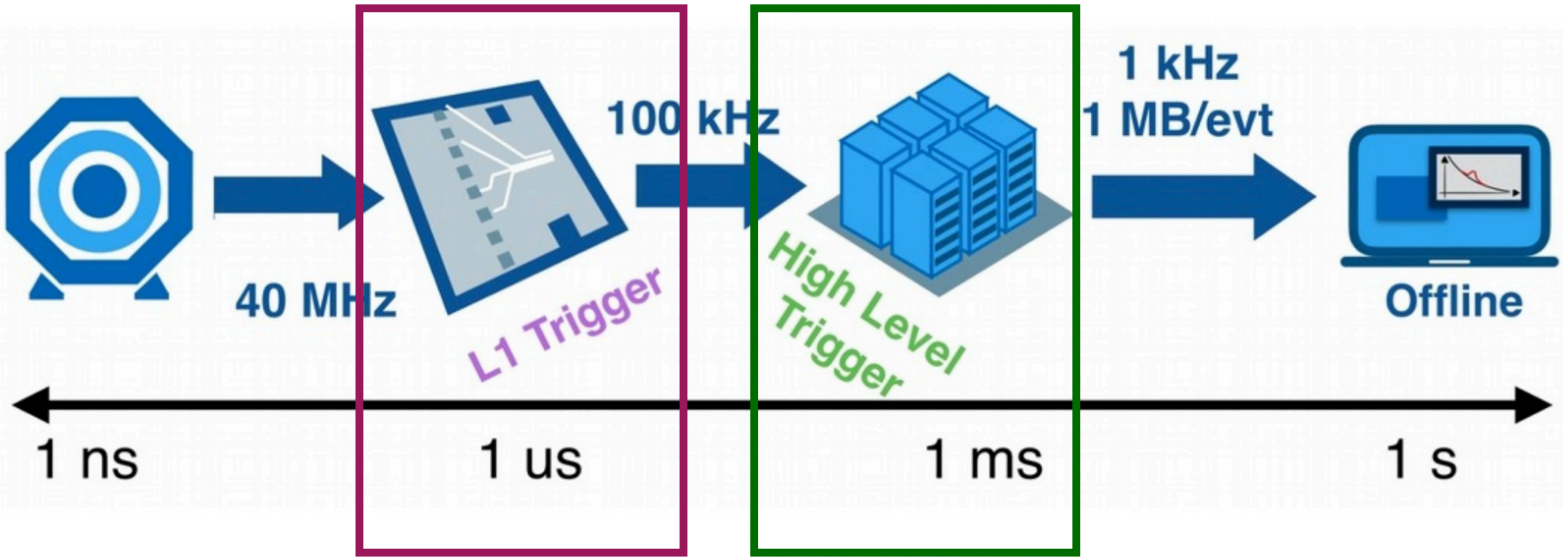
```yaml
KerasJson: keras/KERAS_3layer.json
KerasH5:   keras/KERAS_3layer_weights.h5
OutputDir: my-hls-test
ProjectName: myproject
XilinxPart: xcku115-flvb2104-2-i
ClockPeriod: 5

HLSConfig:
  Model:
    Precision: ap_fixed<16,6>
    ReuseFactor: 1
    Strategy: Latency #Resource
  LayerName:
    dense1:
      ReuseFactor: 2
      Strategy: Latency #Resource
      Compression: True
```

keras-config.yml

- Easy to **run**:

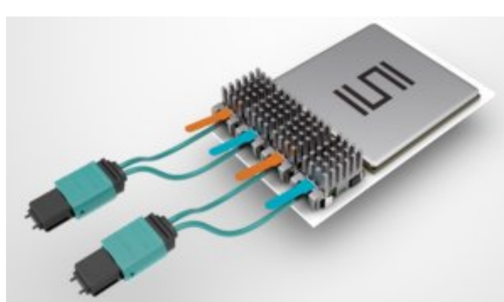
  Conversion: `hls4ml convert -c keras-config.yml`
  Build: `hls4ml build -p my-hls-test -c -s -r`
  Help: `hls4ml -h / hls4ml command -h`

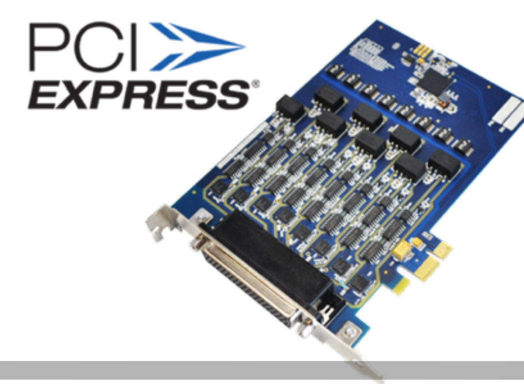
Data Processing



Optical to electrical to fpga pins

Stream through PCIexpress





<1μs

1ms-1s

The Team

MEET THE COLLABORATORS

(click on name for more info)

CERN
Vladimir Loncar (PhD, Computer Science); Jennifer Ngadiuba (PhD, Physics); Maurizio Pierini (PhD, Physics); Sioni Summers (PhD, Physics);

Columbia University
Giuseppe Di Guglielmo (PhD, Computer Science)

Fermilab
Christian Herwig (PhD,Physics); Burt Holzman (PhD,Physics); Sergo Jindariani (PhD,Physics); Thomas Klijnsma (PhD,Physics); Ben Kreis (PhD,Physics); Mia Liu (PhD,Physics); Kevin Pedro (PhD,Physics); Ryan Rivera (PhD,EE); Nhan Tran (PhD,Physics)

Hawkeye 360
EJ Kreinar (Computer Science)

MIT
Jack Dinsmore (Undergraduate, Physics); Song Han (PhD, EECS); Phil Harris (PhD, Physics); Sang Eon Park (Graduate, Physics); Dylan Rankin (PhD, Physics);

UC San Diego
Javier Duarte: PhD, Physics, Caltech

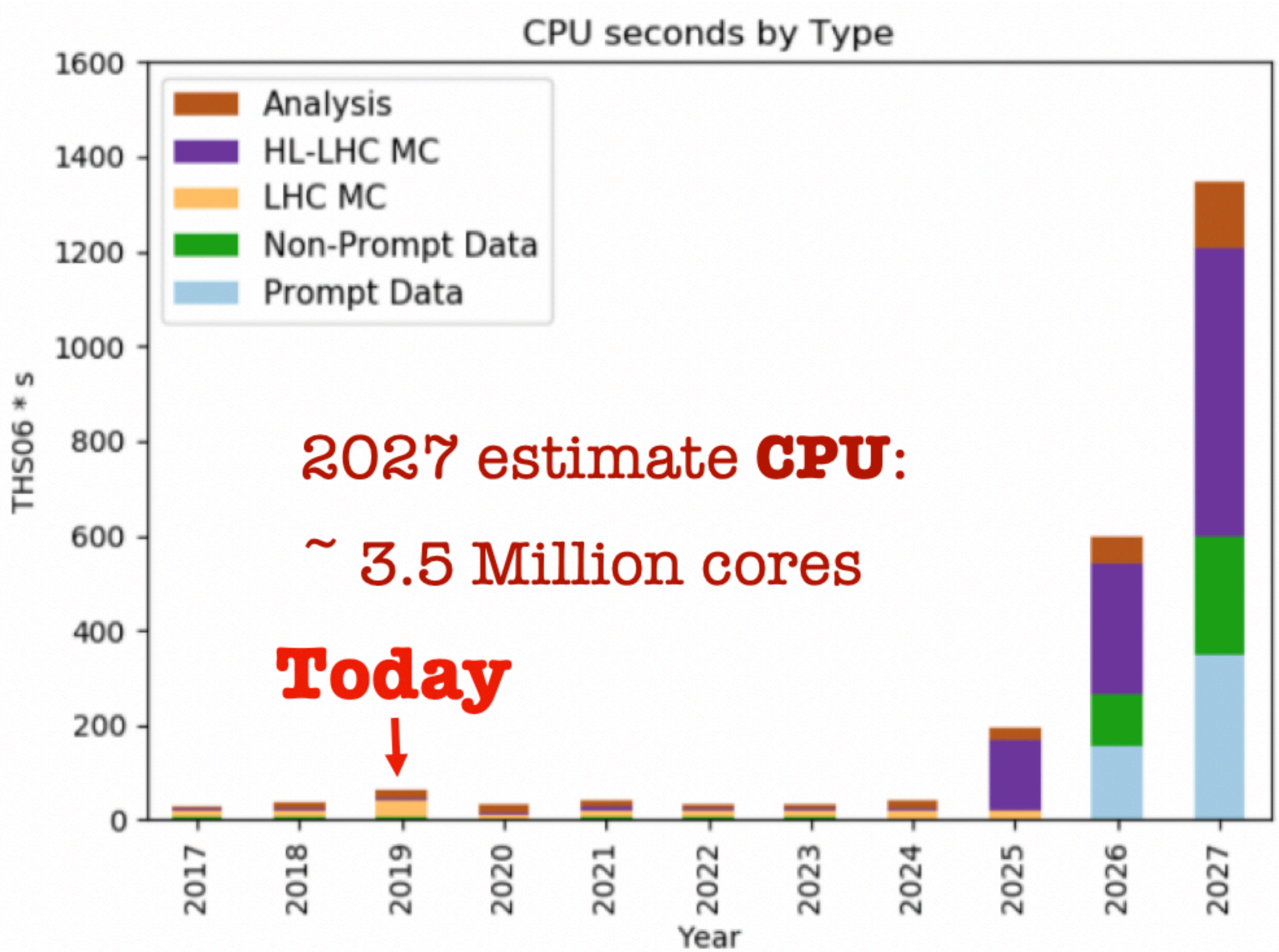
University of Illinois Chicago
Zhenbin Wu (PhD, Physics);

University of Illinois Urbana-Champaign
Markus Atkinson (PhD, Physics); Mark Neubauer (PhD, Physics);

University of Washington
Scott Hauck (PhD, EECS); Shih-Chieh Hsu (PhD, Physics);

Challenges for HLT/Offline

10x larger events * 5x the rate * 10 years of data-taking



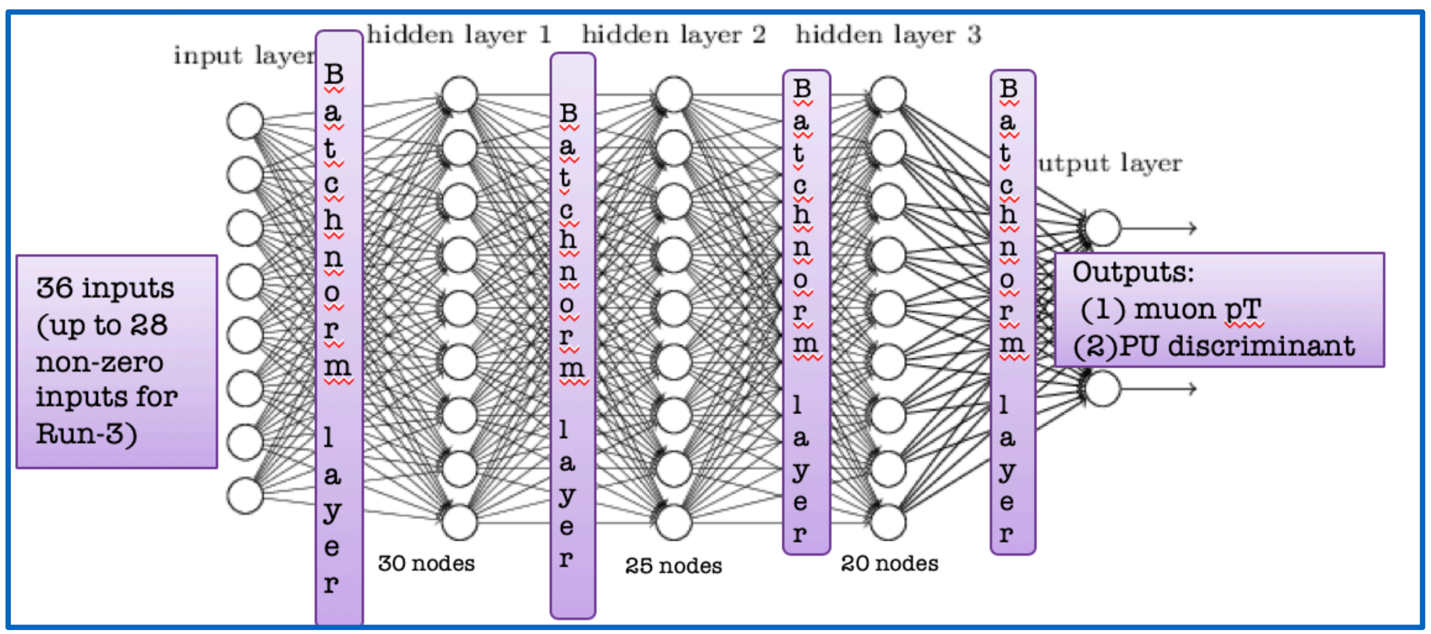
And what if we need to expand the physics program?

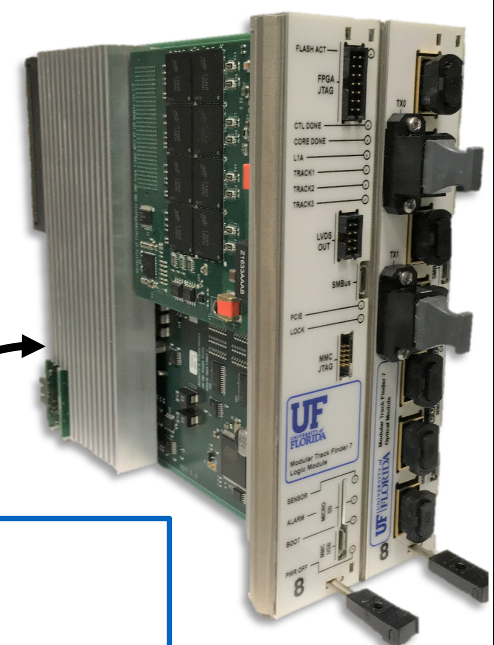
Deep Learning in Level-1

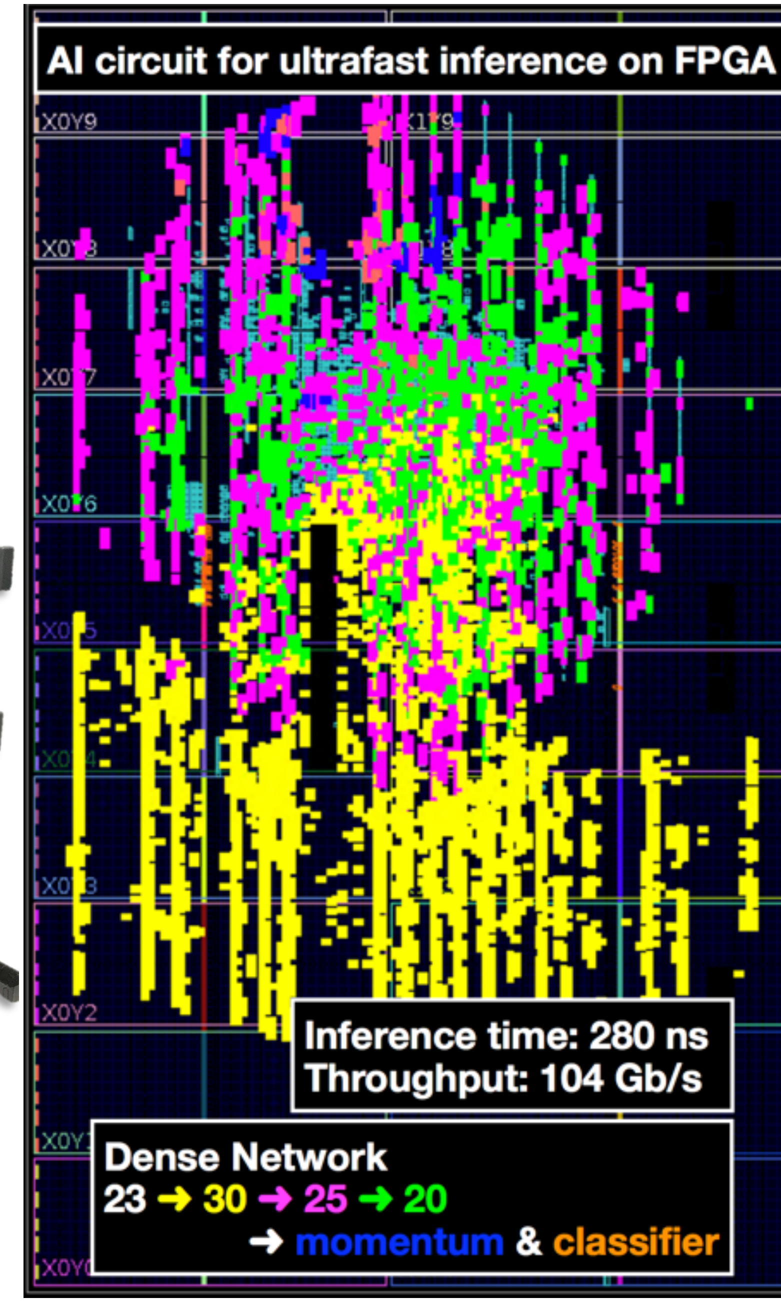
CMS Muon Reconstruction Steps

hits → Pattern recognition → Track building → p_T assignment → tracks

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

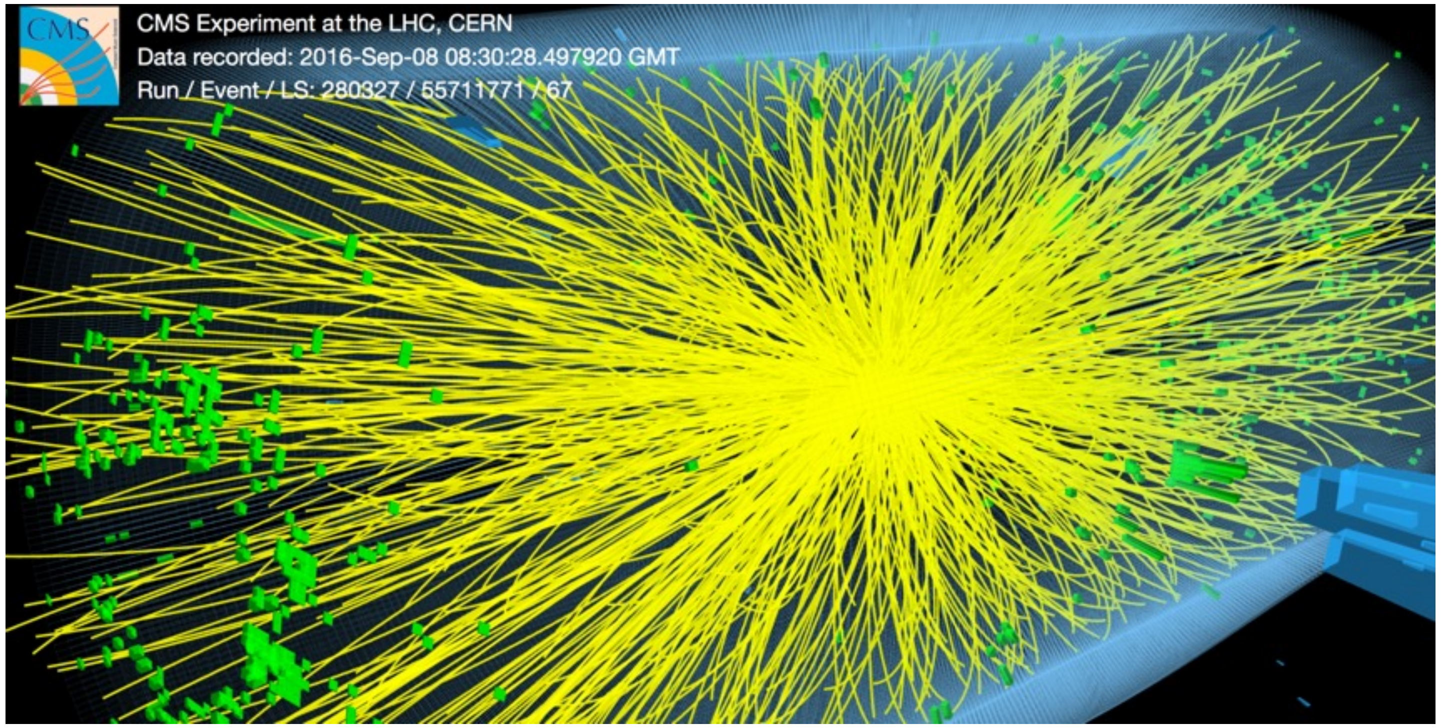






The Challenge

This is an event collected in 2016. Interaction region is ~10 cm in Z



- More data => more physics, but also more PileUp.
- Currently up to 70 collisions per event.
- In HL-LHC the average number will go up to <PU>=200
- FCC-hh <PU> = 800-1000