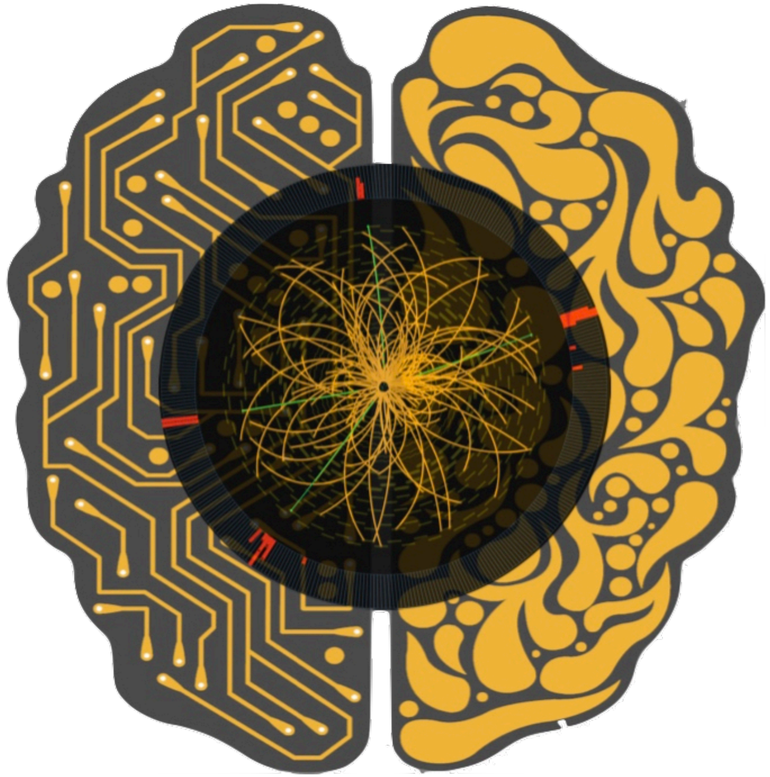
Activities in Fast ML co-processor group

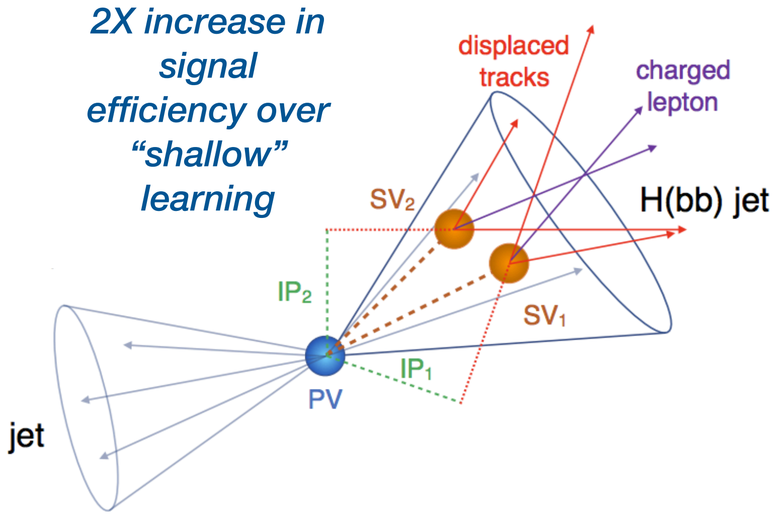
Mia & Markus

Mini-workshop on Portable Inference

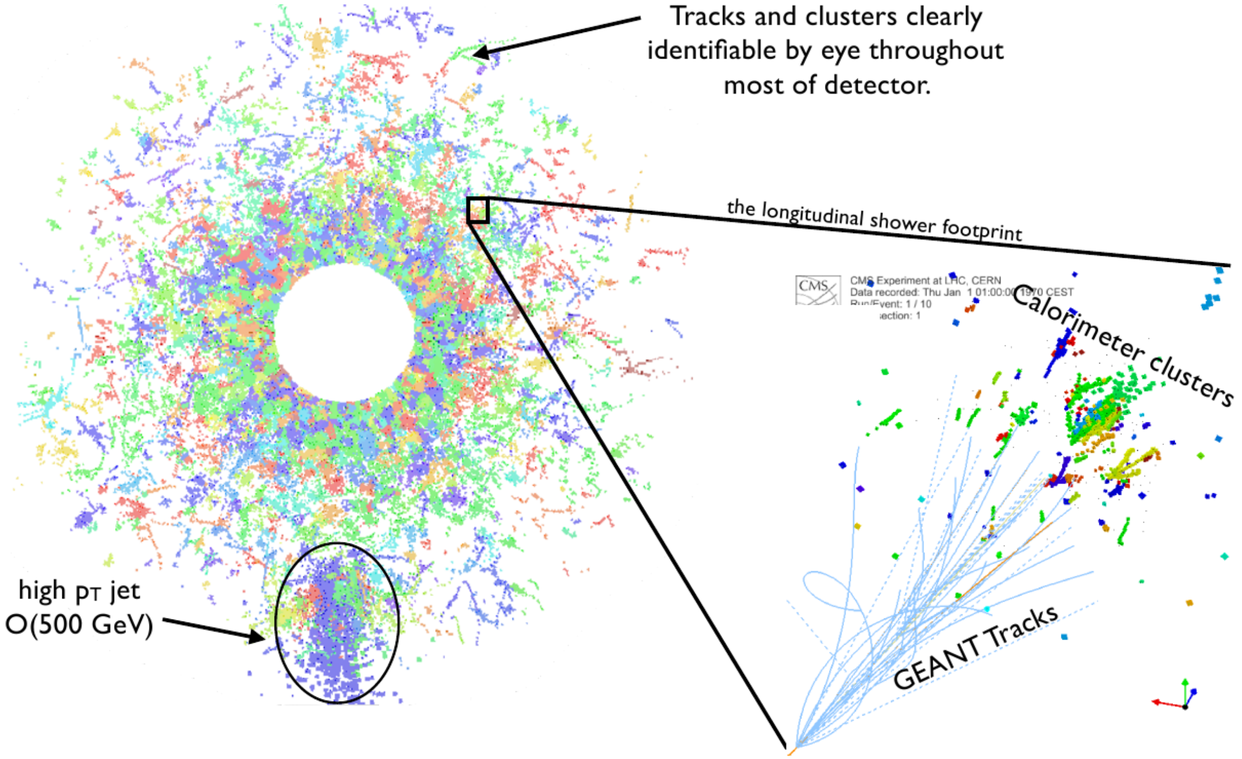
Dec.4. 2020



A quest for accelerated Machine Learning inference



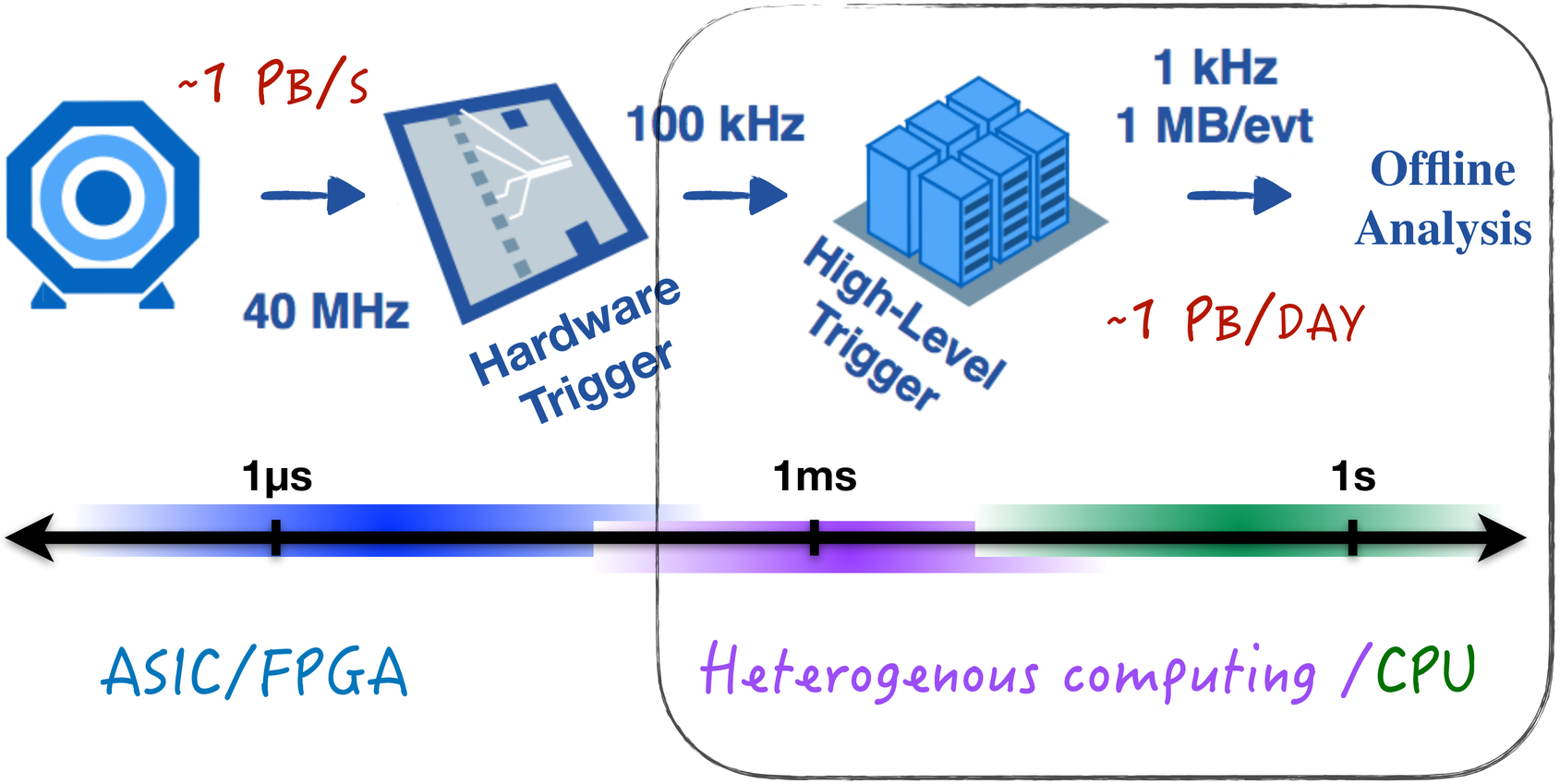
Heavy flavor jet tagging



CMS High Granularity Calorimeter Reconstruction

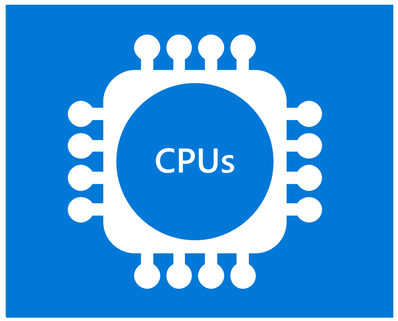
Accelerated machine learning opens up AI application domain in real-time system and offers novel solutions to computing challenges. See white paper fast ml workshop 2019.

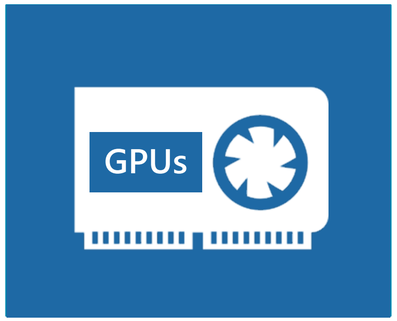
Accelerated ML for HLT/offline

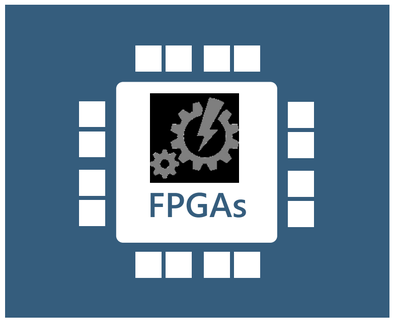


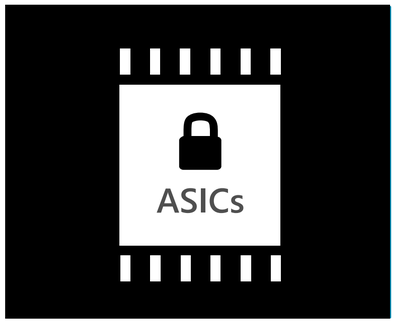
Discussed in the fast ML co-processor group

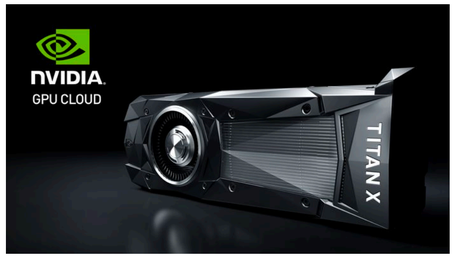
#Trending in Industry: Heterogeneous Computing

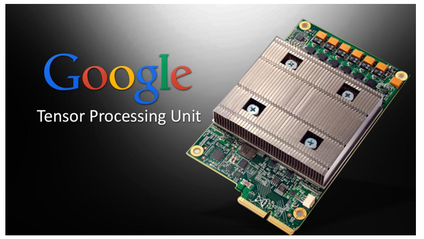


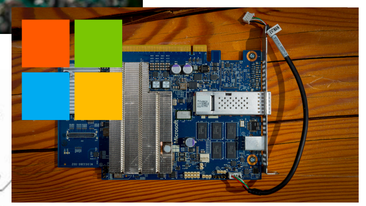






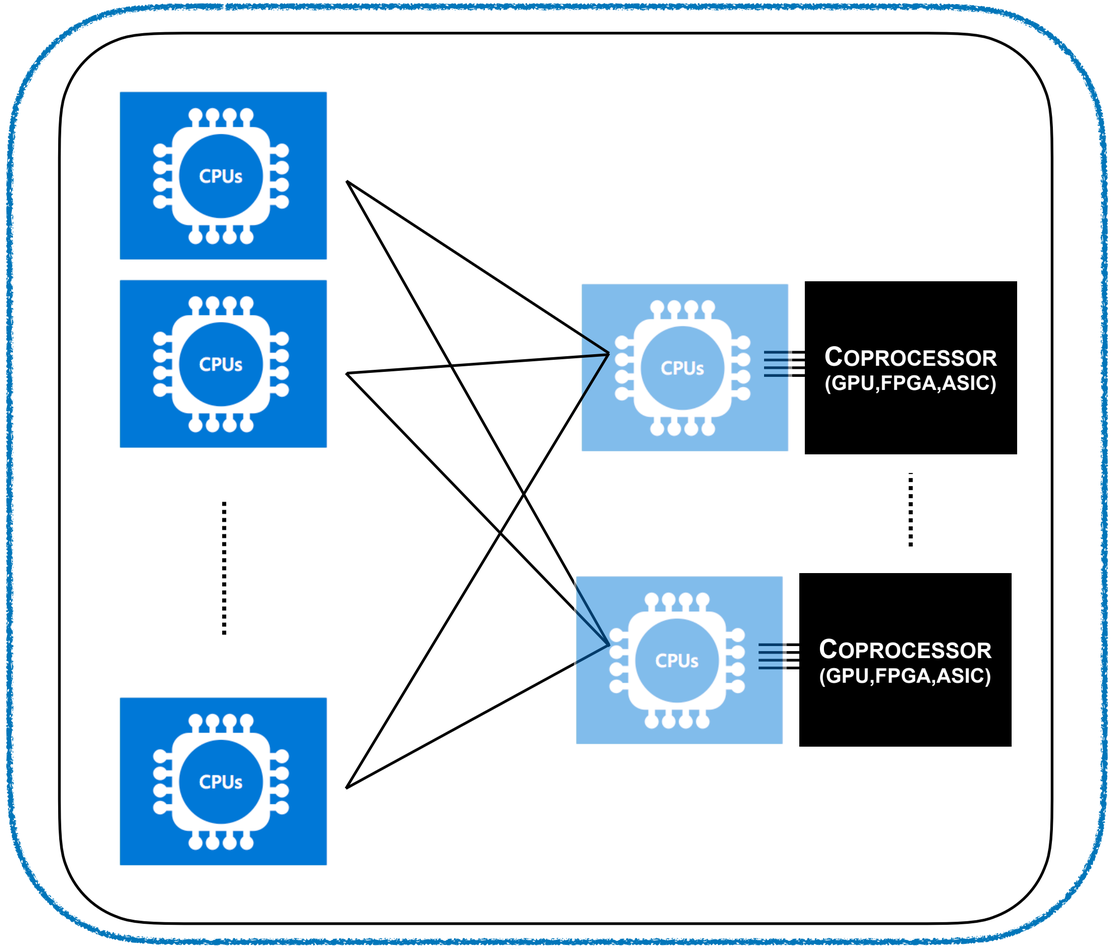






Advances driven by
big data explosion
& machine learning

#Heterogeneous Computing Paradigm

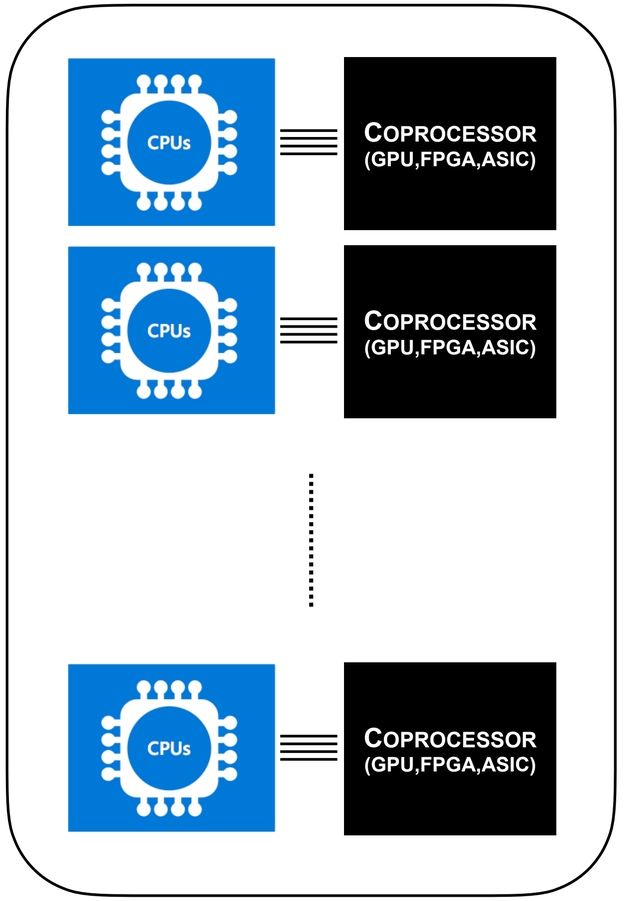


Pros:

scalable algorithms

scalable to the grid/cloud

Heterogeneous heterogeneity (mixed hardwares)

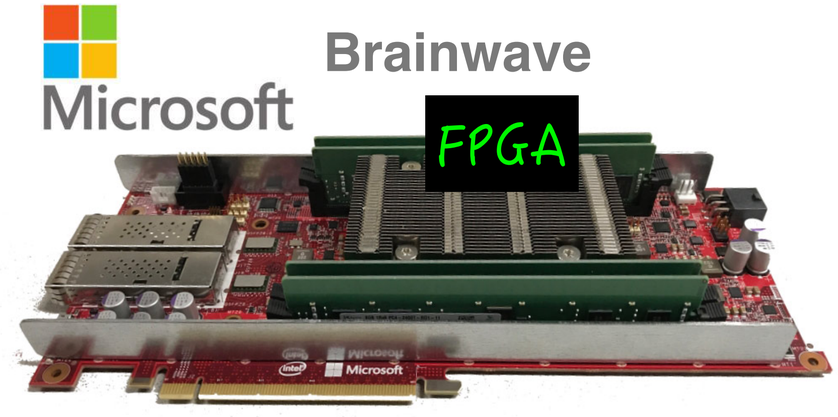


Pros:

less system complexity

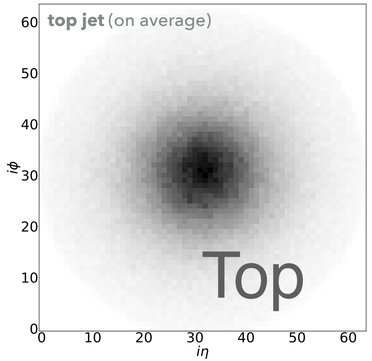
no network latency

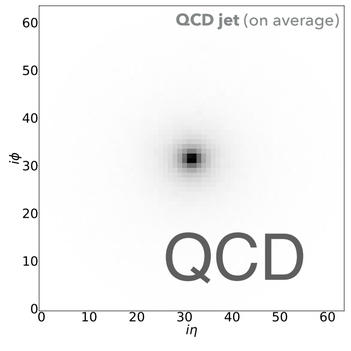
Services for Optimized Network Inference on Co-processors

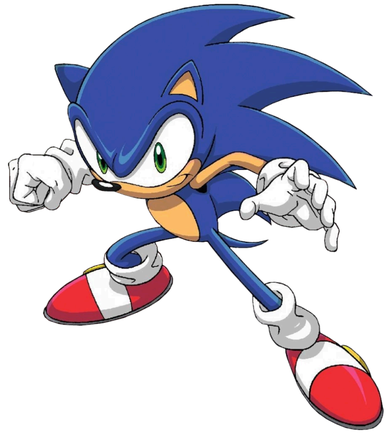


Question:

Can we/How can we take advantage of **heterogenous computing as-a-service** for our big data problems?



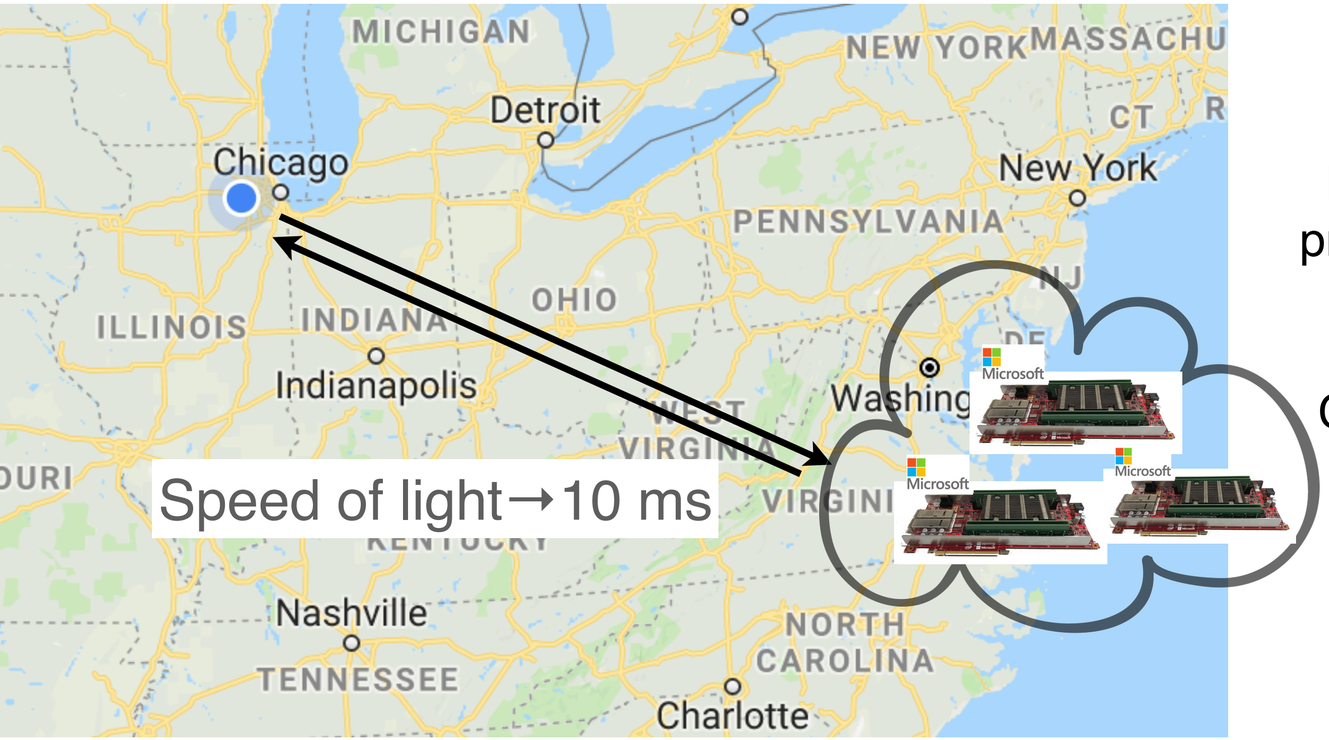




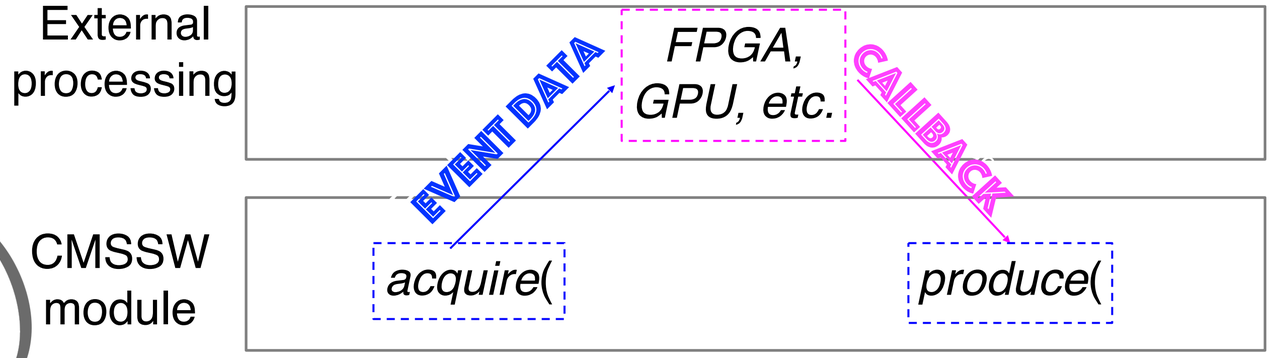
SONIC

Published in CSBS

Proof of concept: as-a-service approach



Test integrated in CMS software stack



SONIC: latest explorations in the fast-ml group

GPU-as-a-service at the LHC

https://arxiv.org/abs/2007.10359

Hardware platforms

GPU-as-a-service for DUNE

https://arxiv.org/pdf/2009.04509.pdf

SONIC: latest explorations in the fast-ml group

FACILE

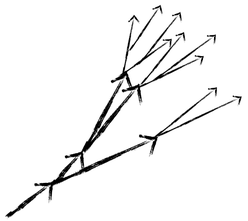
CMS Hadronic Calorimeter channel regression

DeepCalo*

ECAL cluster regression

ResNet

top quark image classification



GPU-as-a-service

https://arxiv.org/abs/2007.10359

Hardware platforms

2k parameters ⟷ **10 M parameters**

GPU-as-a-service for DUNE

https://arxiv.org/pdf/2009.04509.pdf

Algorithm complexity

More benchmarks driven by use cases to test scaling for HLT/offline

SONIC: latest explorations in the fast-ml group

GPU-as-a-service
https://arxiv.org/abs/2007.10359

GPU-as-a-service for DUNE
https://arxiv.org/pdf/2009.04509.pdf

Hardware platforms

FPGA-as-a-service Toolkit

Open source tools: flexibility

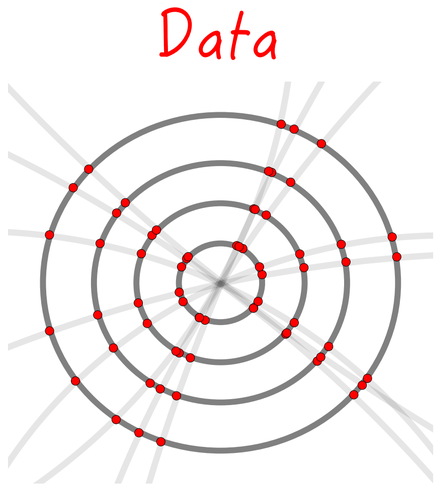
hls4ml

Algorithm complexity

More benchmarks driven by use cases to test scaling for HLT/offline

Future prospects

Co-processor group: Integrating heterogeneous computing resources across HEP experiments



Next:

More proof-of-concept studies:

SONIC in High Performance Computers (HPCs)

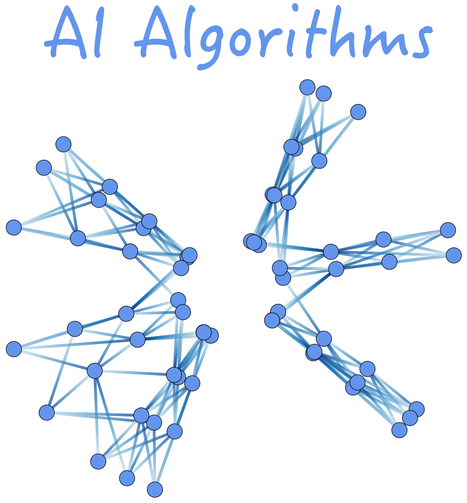
: Most studies performed using Cloud services/on-premises clusters

Non-ML based approach with SONIC

: the as-a-service approach can be extended

Expect more discussions on integration in experiment production systems

: processing for full-scale protoDune-SP reconstruction, integration in CMS software and production system, identify bottlenecks in software, tools & resources.

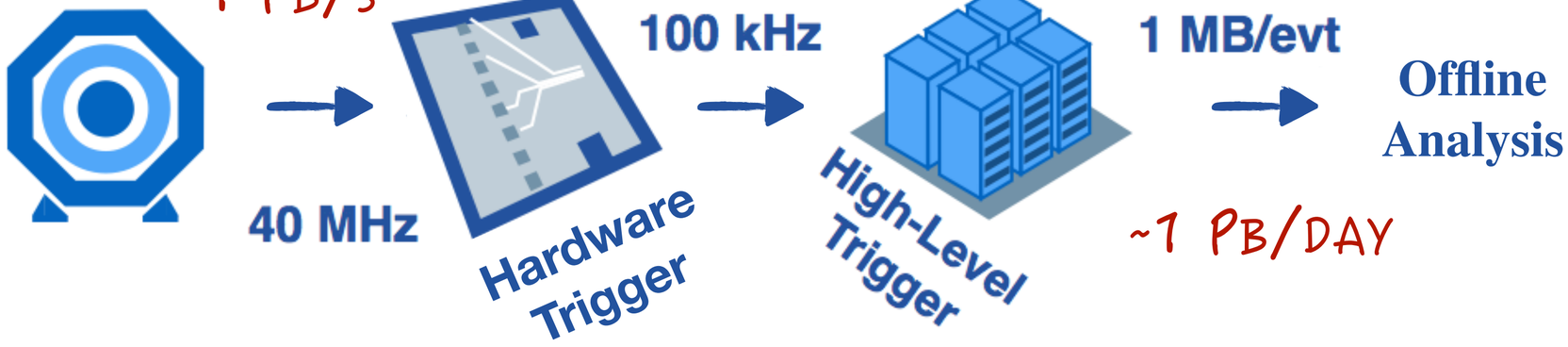


Type B meeting in the fast ML meeting series: 10 AM Fridays (CDT)

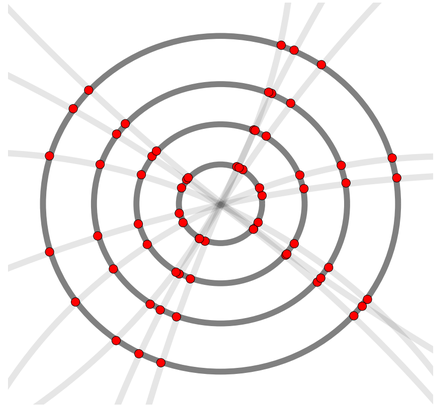
https://indico.cern.ch/category/11844/

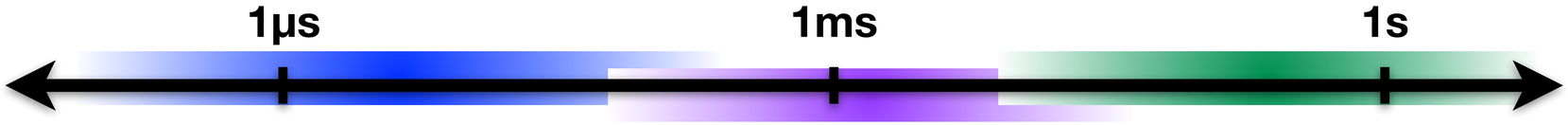
Hope to have continuous discussions on topics in today's workshop

Accelerated discoveries with Real-Time AI



Data

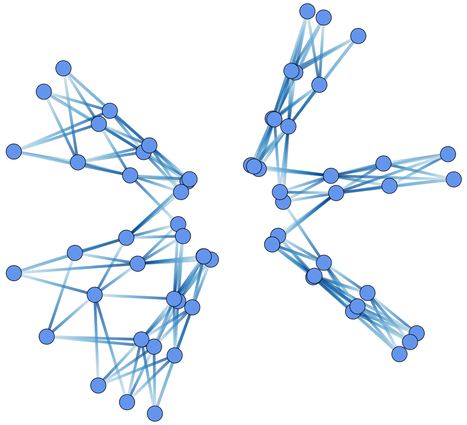




ASIC/FPGA

Heterogenous computing /CPU

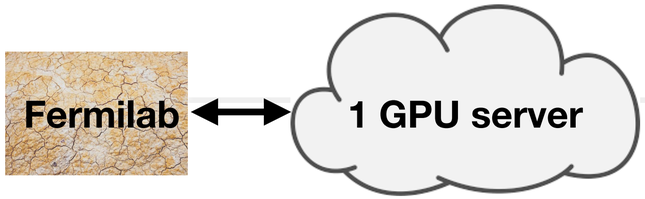
AI Algorithms



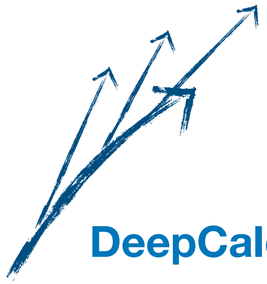
hls4ml

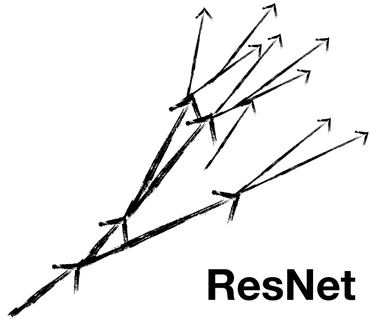
Co-design

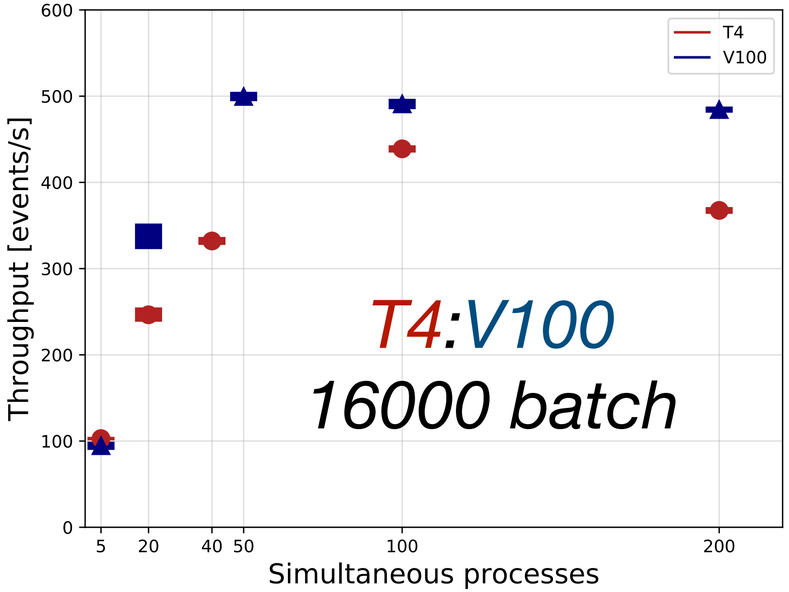
Throughput

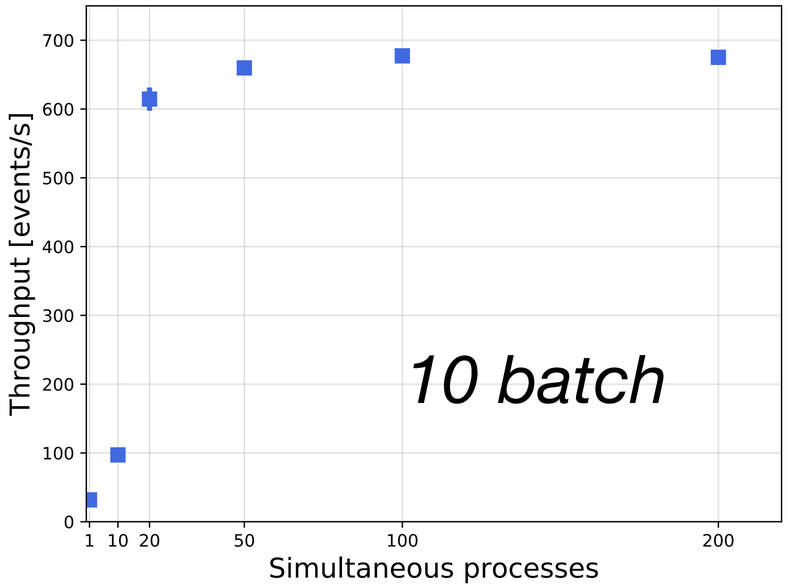

Fermilab
1 GPU server

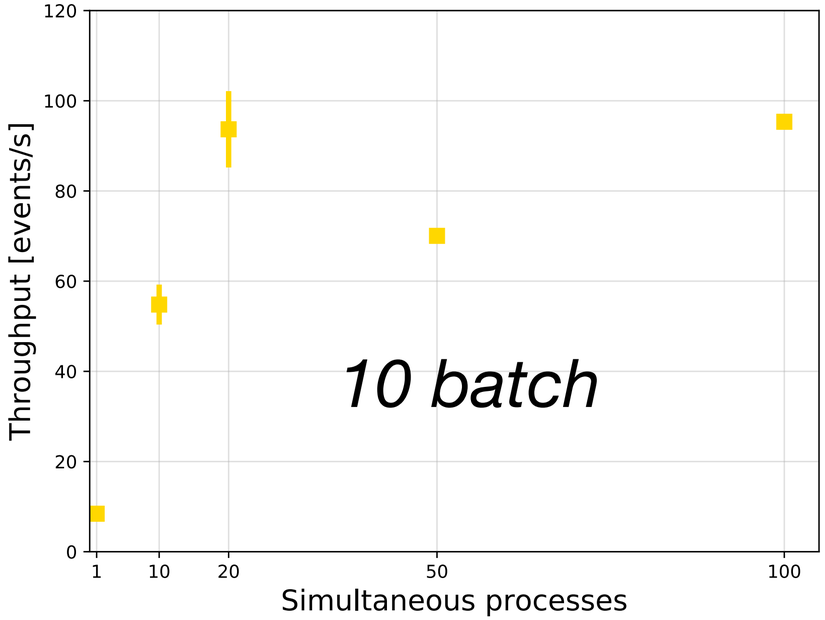

FACILE


DeepCalo


ResNet


T4
V100
T4:V100
16000 batch
Throughput [events/s]
Simultaneous processes
0
100
200
300
400
500
600
5
20
40
50
100
200


10 batch
Throughput [events/s]
Simultaneous processes
0
100
200
300
400
500
600
700
1
10
20
50
100
200


10 batch
Throughput [events/s]
Simultaneous processes
0
20
40
60
80
100
120
1
10
20
50
100

SONIC: Not limited to ML

Given a heterogenous computing hardware:

re-write physics algorithms for new hardware

Language: OpenCL, OpenMP,TBB, HLS, …?
Hardware: FPGA, GPU

Parallelized and Vectorized Tracking Using Kalman Filters

- e.g.*On GPUs*

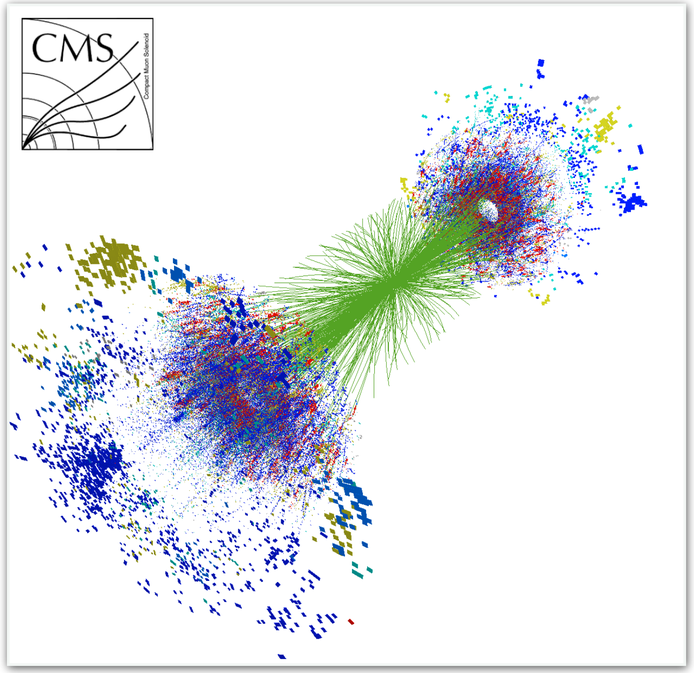
re-cast physics problem as a machine learning problem

Language: C++, Python (TensorFlow, PyTorch,…)

Hardware: FPGA, GPU,ASIC

Tracking with ML

- Algorithms parallelizable
- Solutions with ML e.g., HEP.TrkX.



Charged particle Tracking With graph neural networks