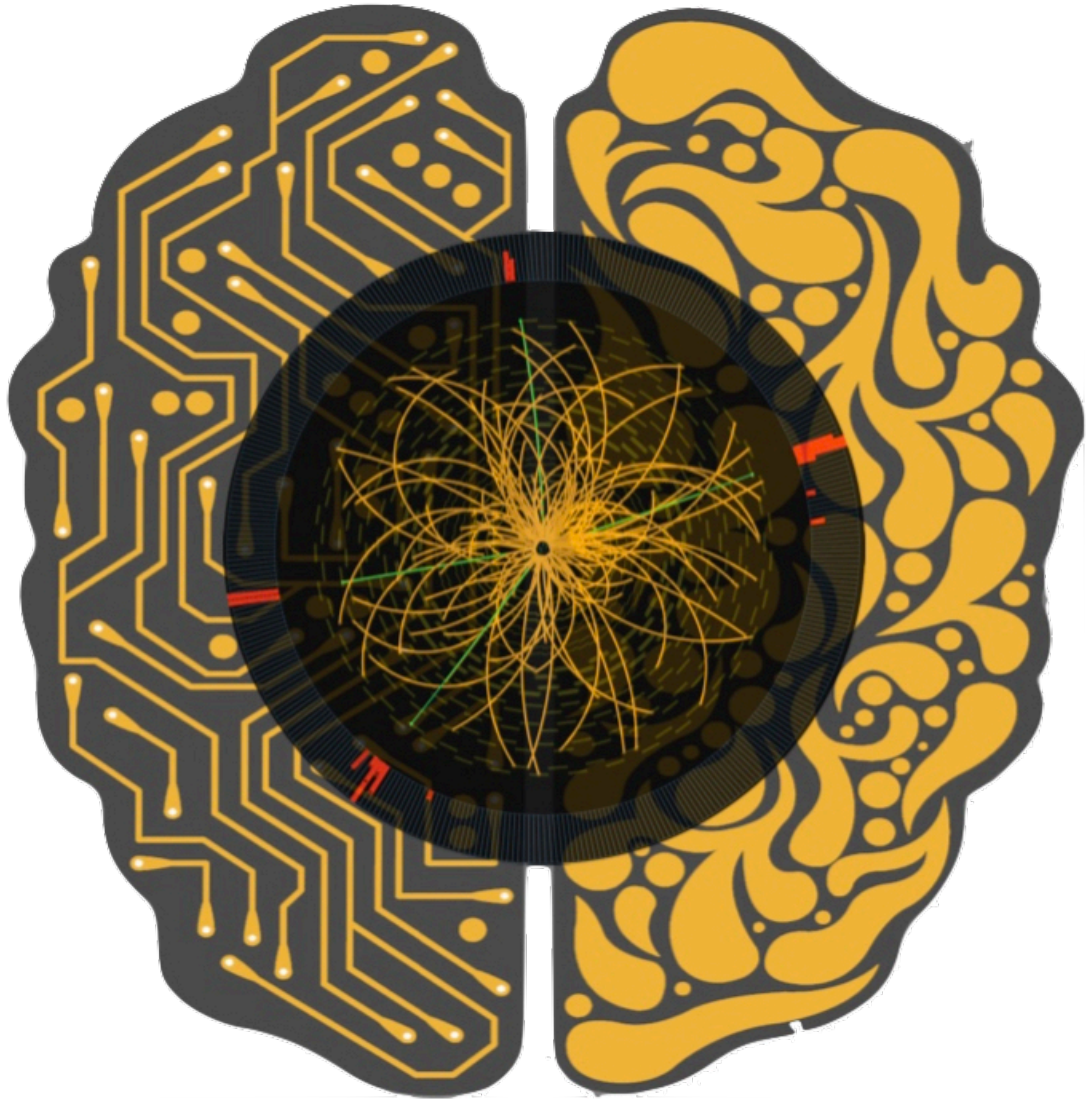
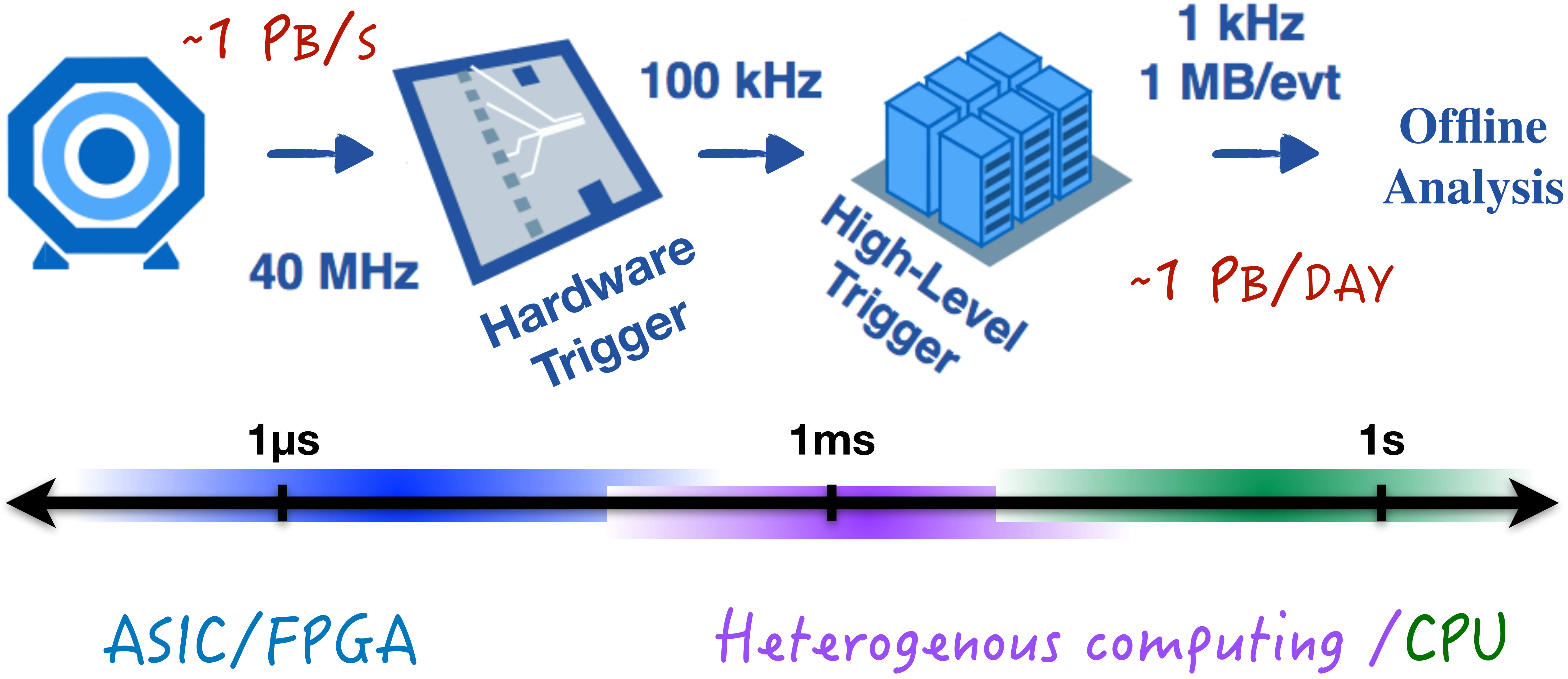
Real-time Artificial Intelligence
— *with heterogeneous compute*

Mia Liu
Purdue University

Oct. 13. 2020
IEEE Real-Time Conference

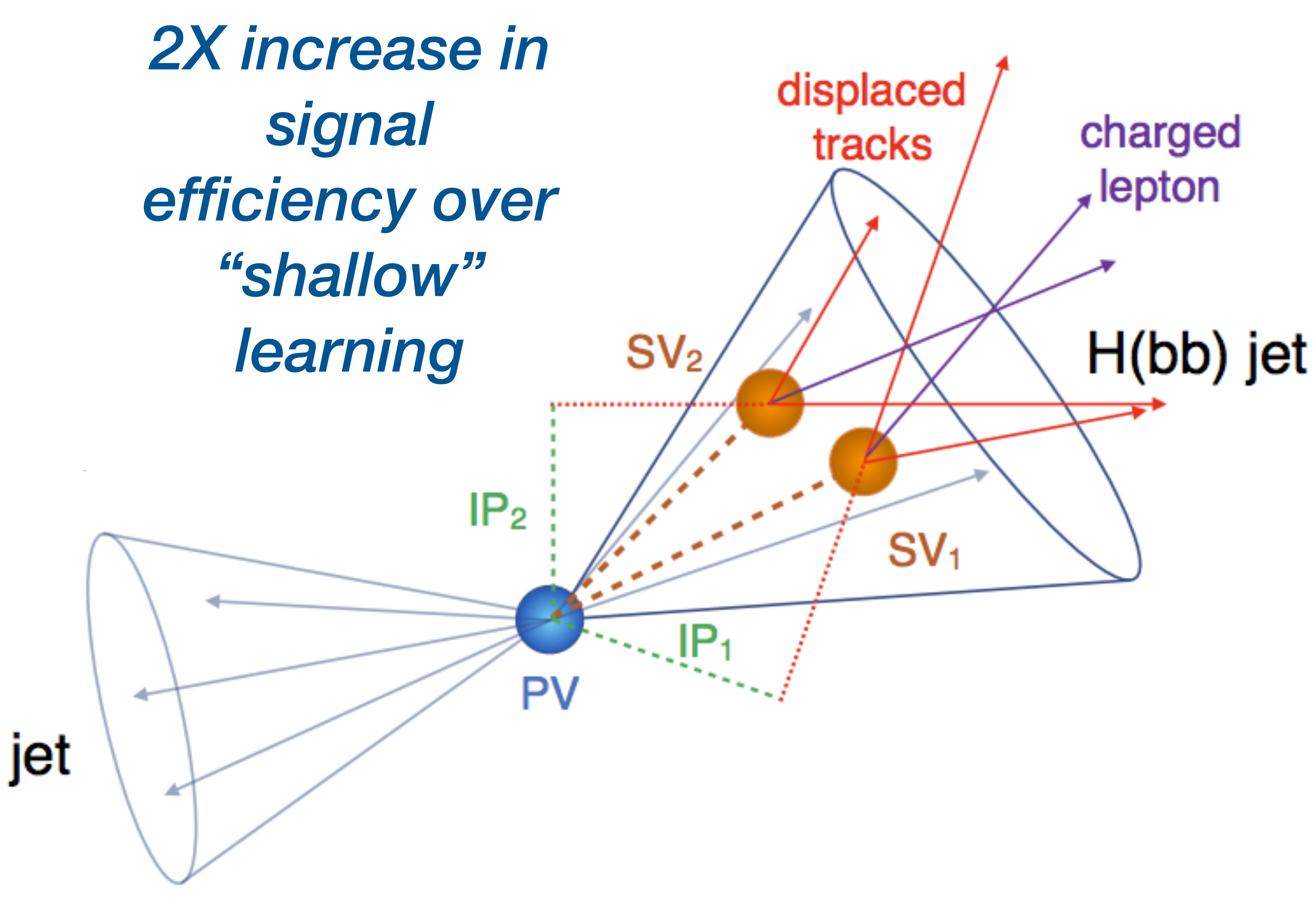


Data processing in Particle Physics

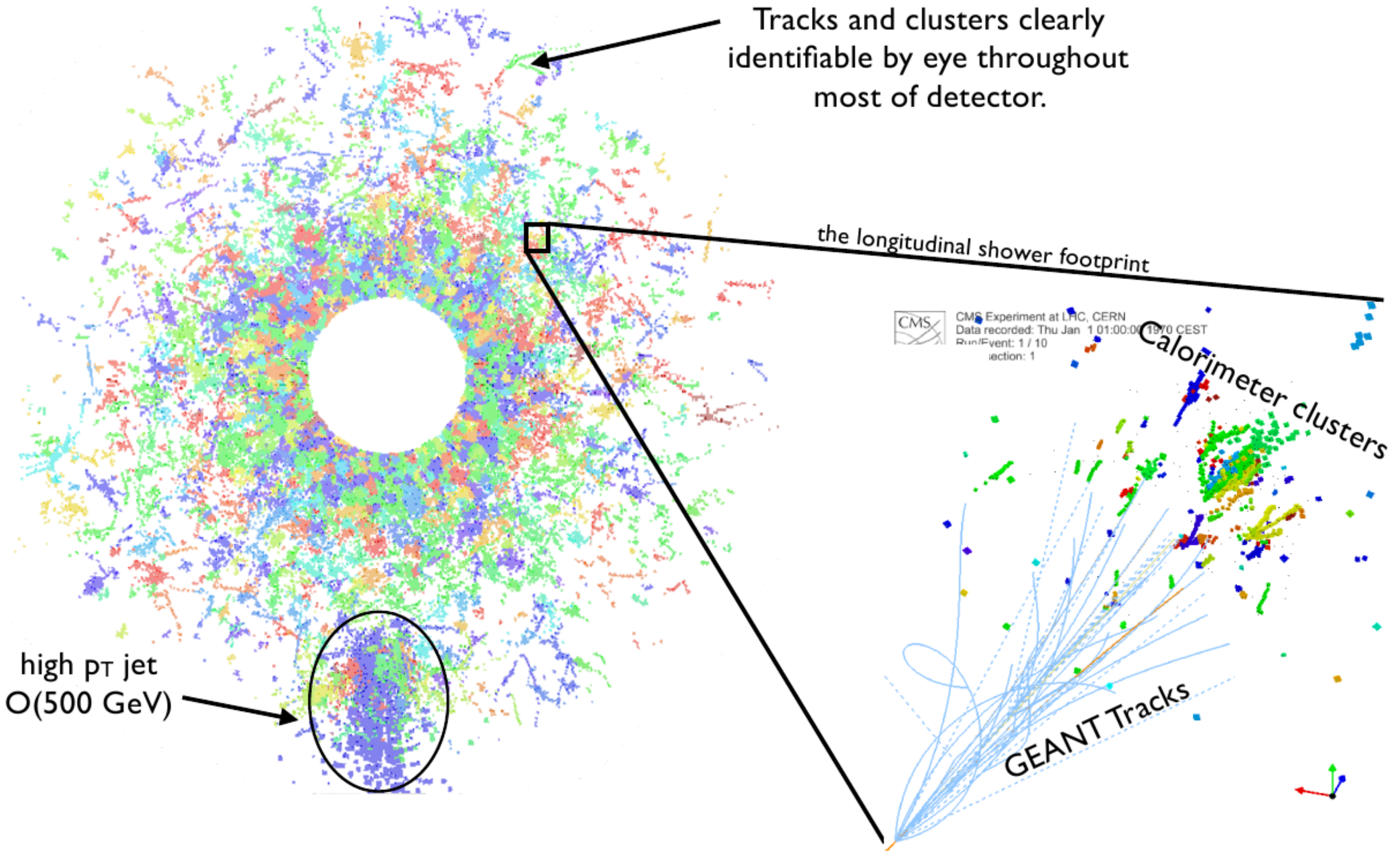


CMS as an example

A quest for accelerated Machine Learning inference



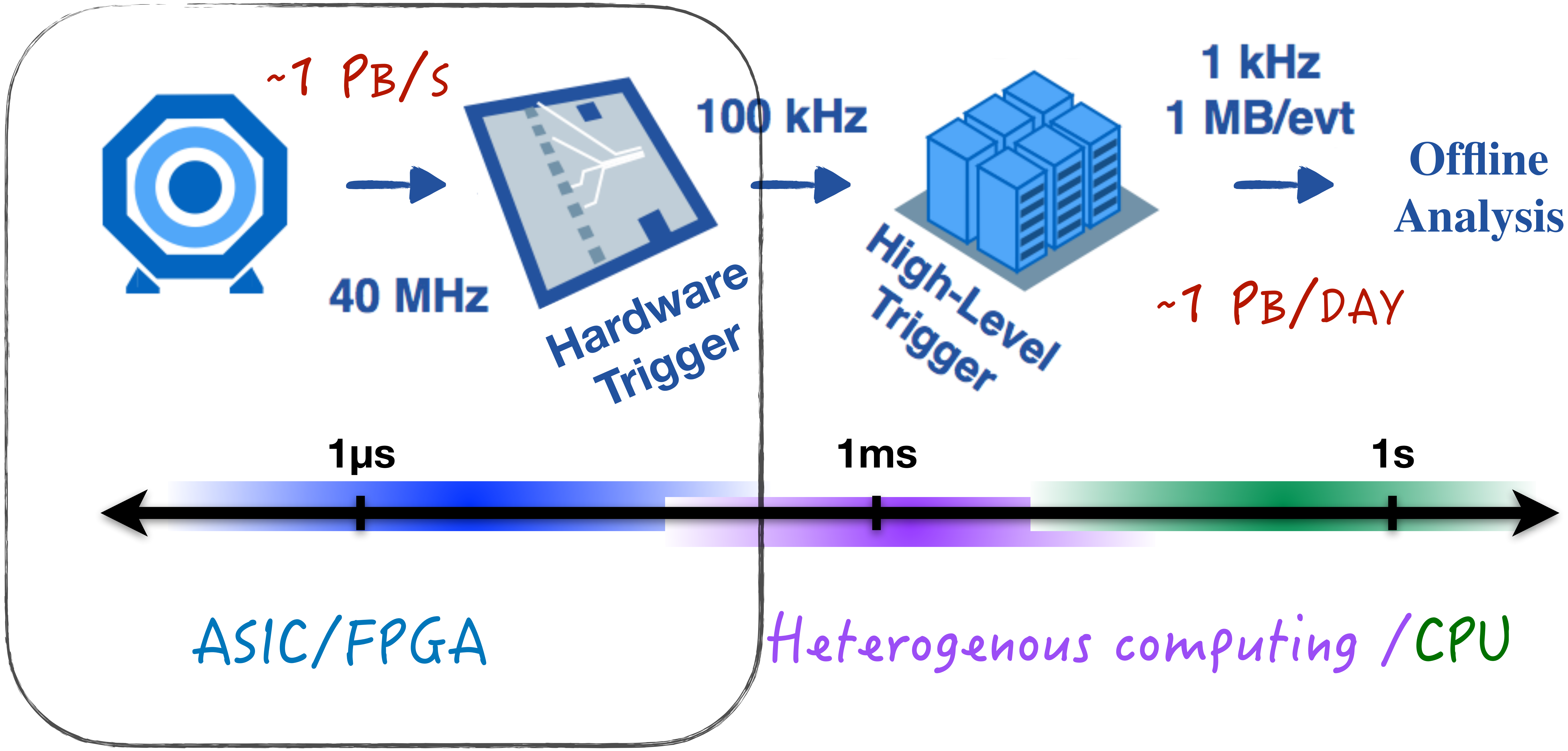
Heavy flavor jet tagging



CMS High Granularity Calorimeter Reconstruction

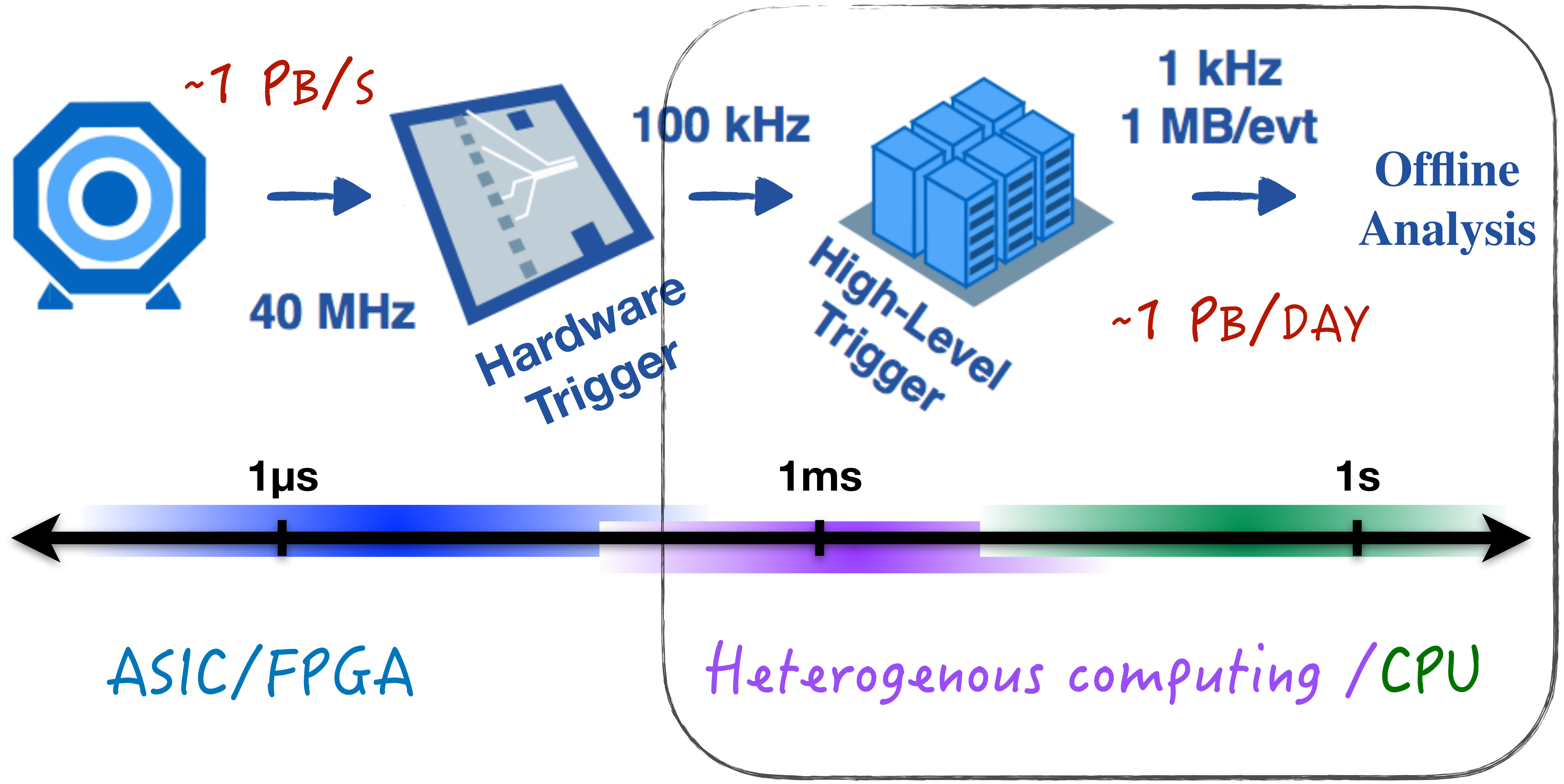
Accelerated machine learning opens up AI application domain in real-time system and offers novel solutions to computing challenges. See our white paper.

Accelerated ML in embedded systems



See Nhan's talk on Monday:
Real-time machine learning in embedded systems for particle physics

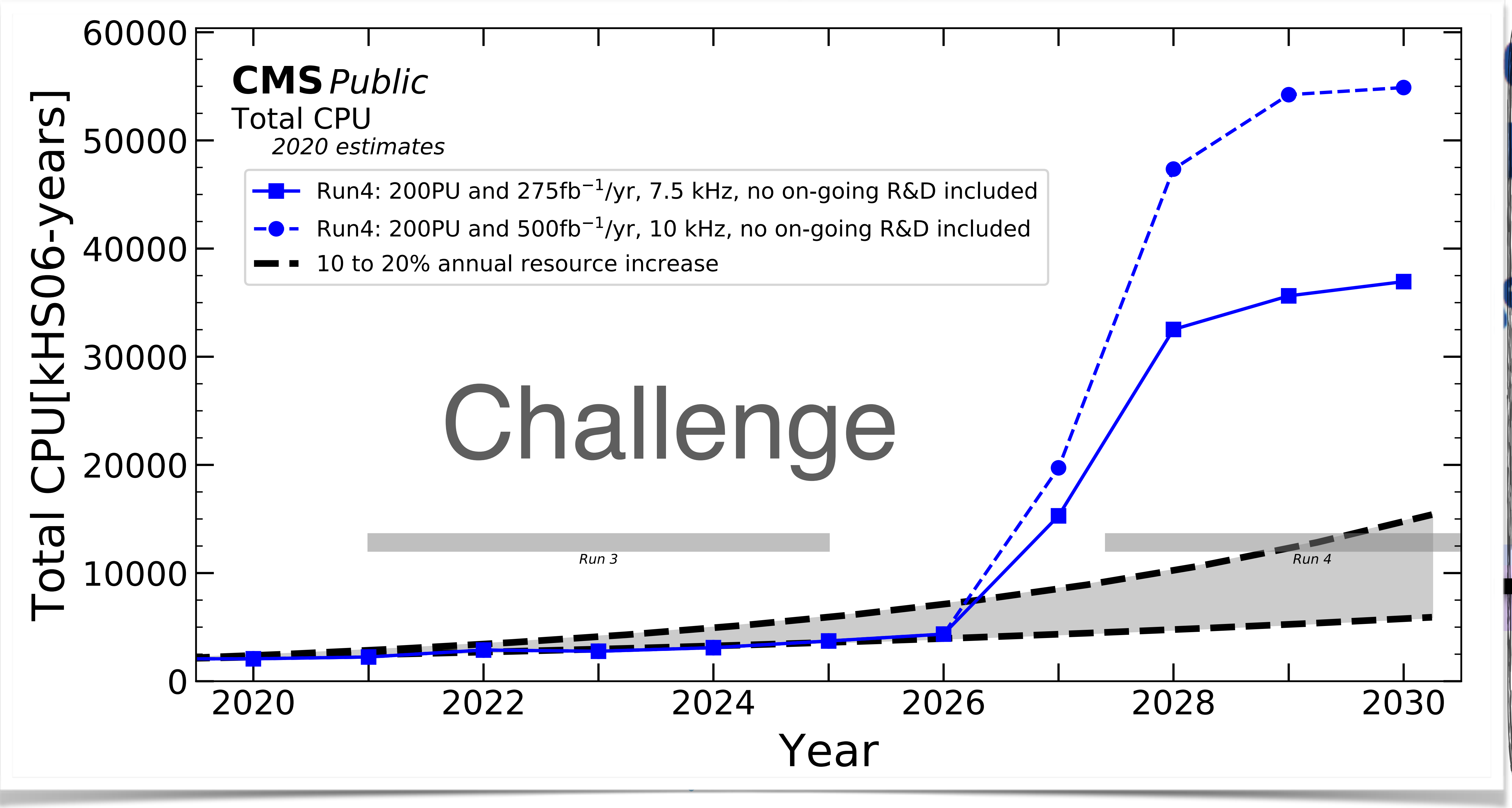
Accelerated ML for HLT/offline

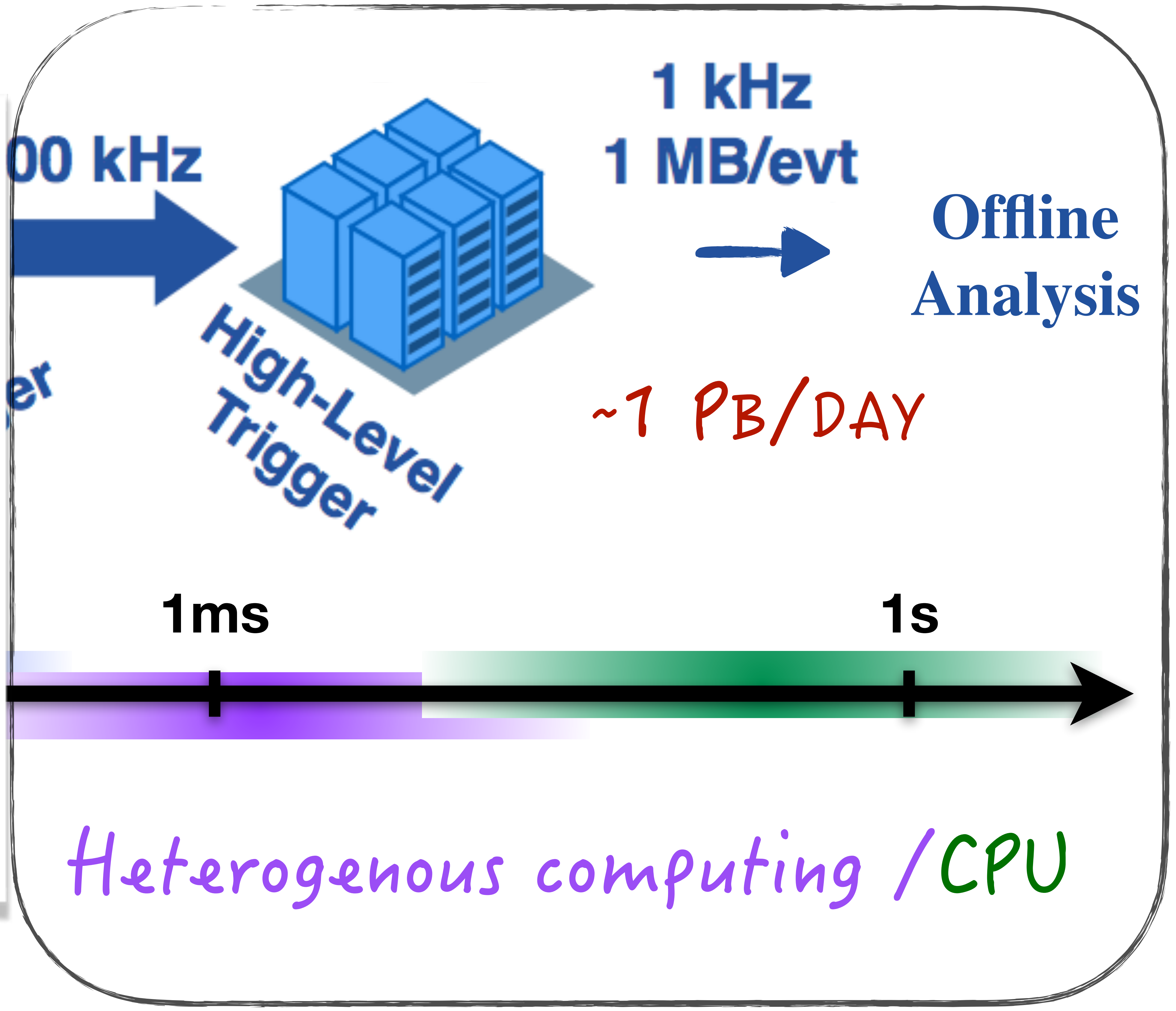


I will focus on this

No faster CPUs for free

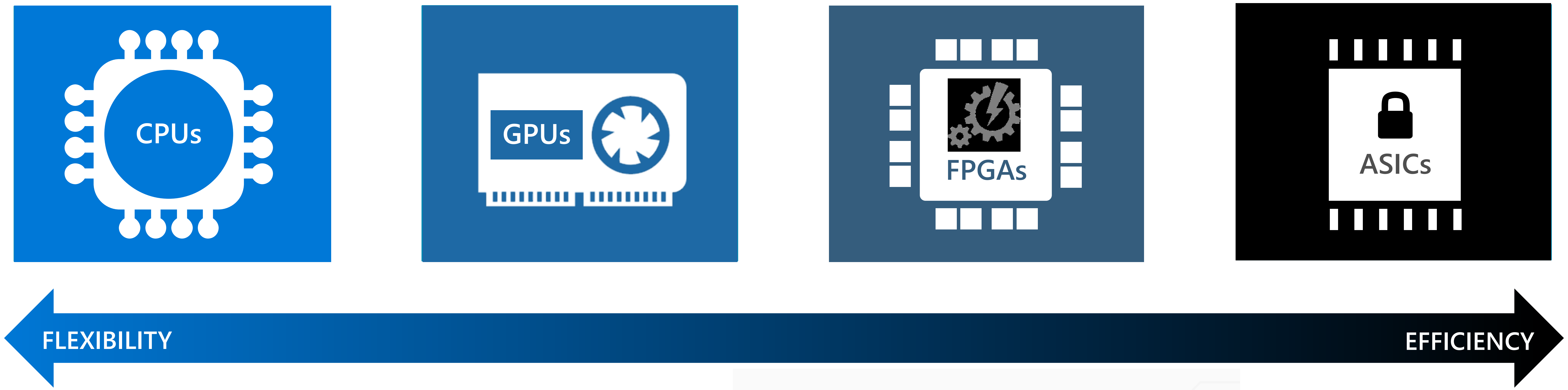
M.LIU

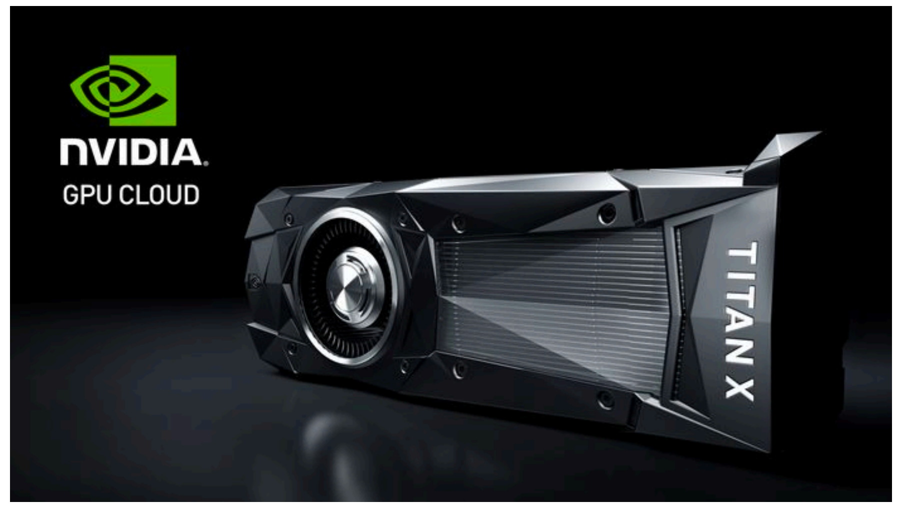


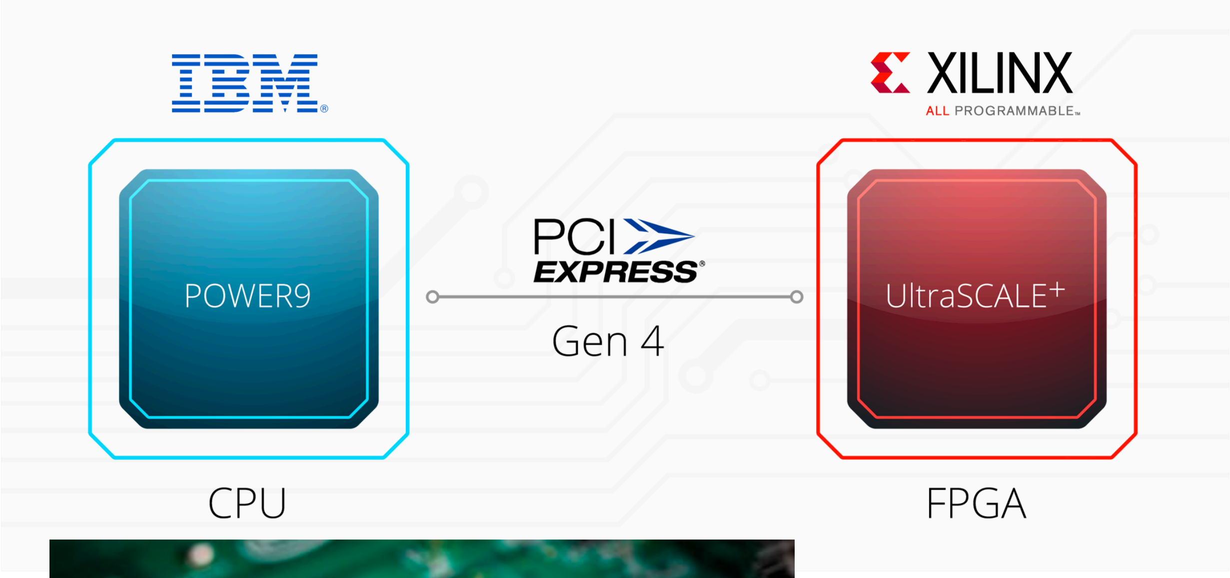


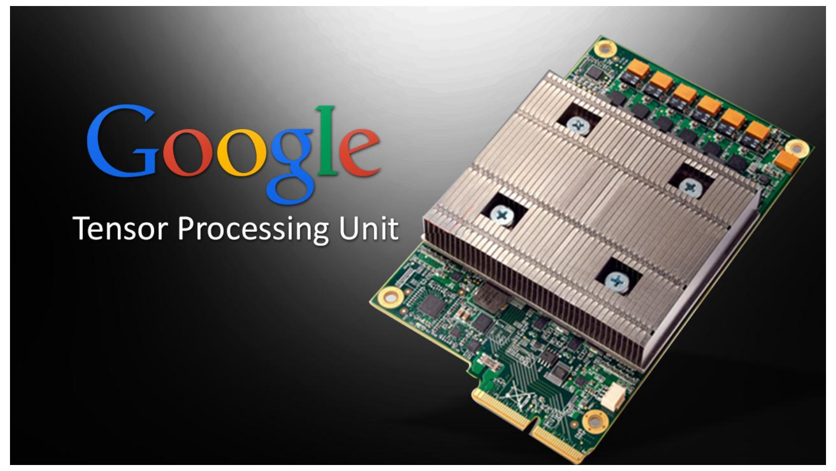
Highlight the opportunities and challenges.
Review current developments… with a bias

#Trending in Industry: Heterogeneous Computing

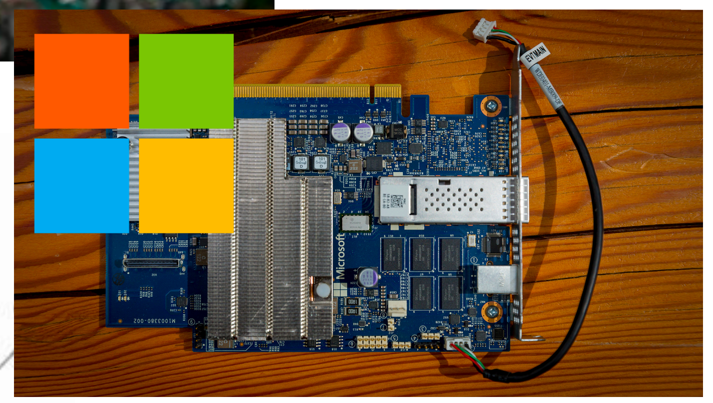




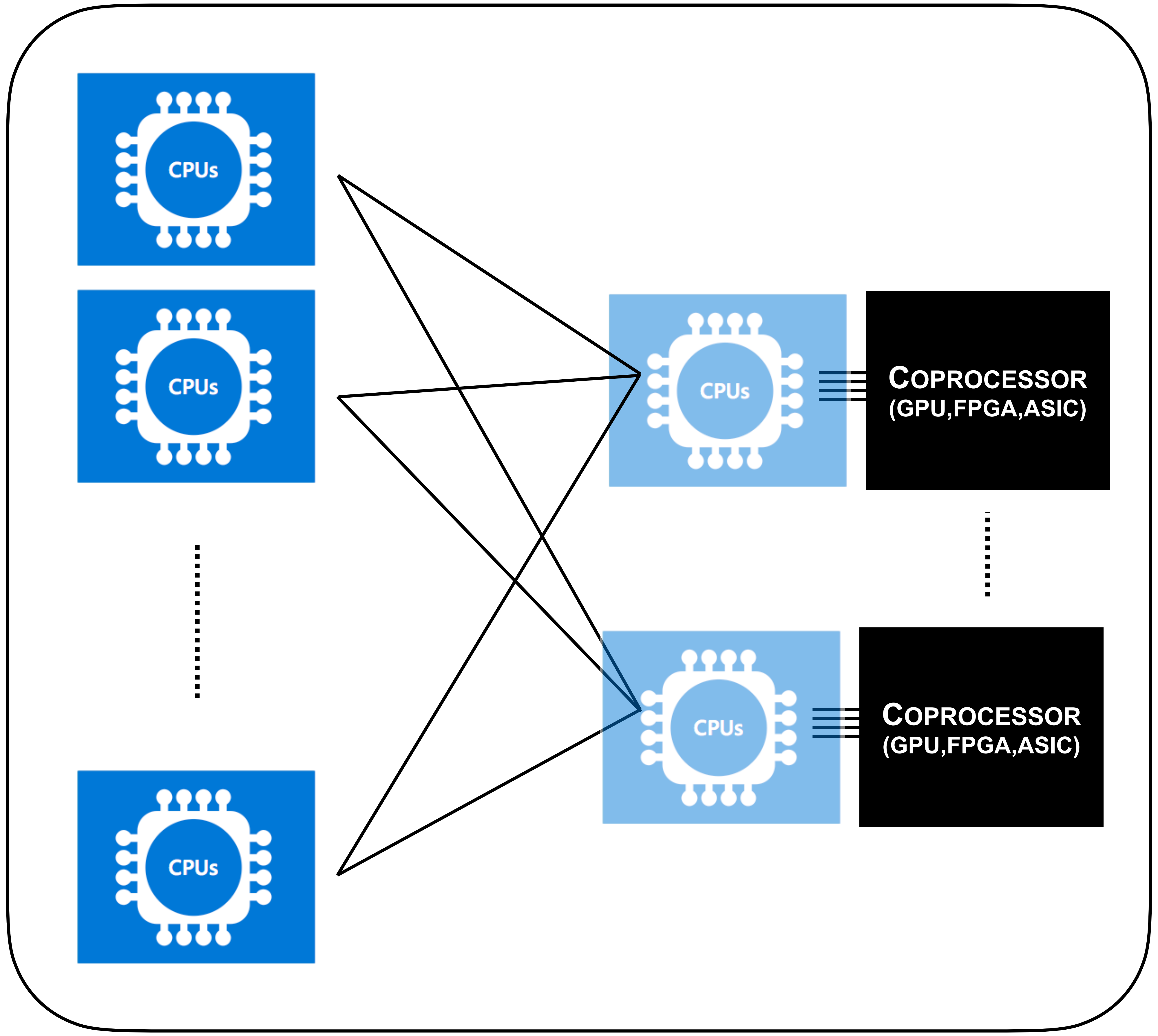




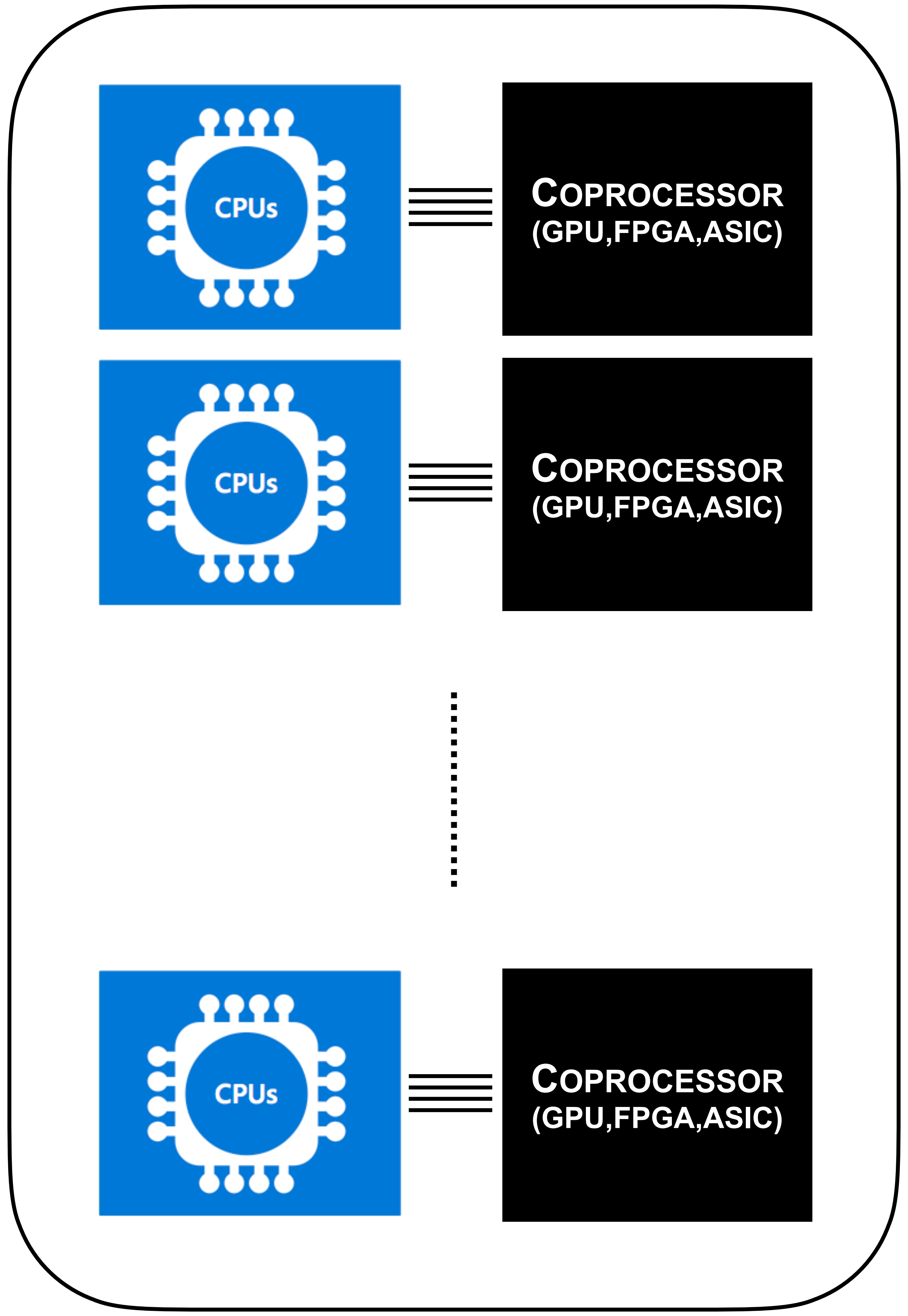
Advances driven by
big data explosion
& machine learning



#Heterogeneous Computing Paradigm

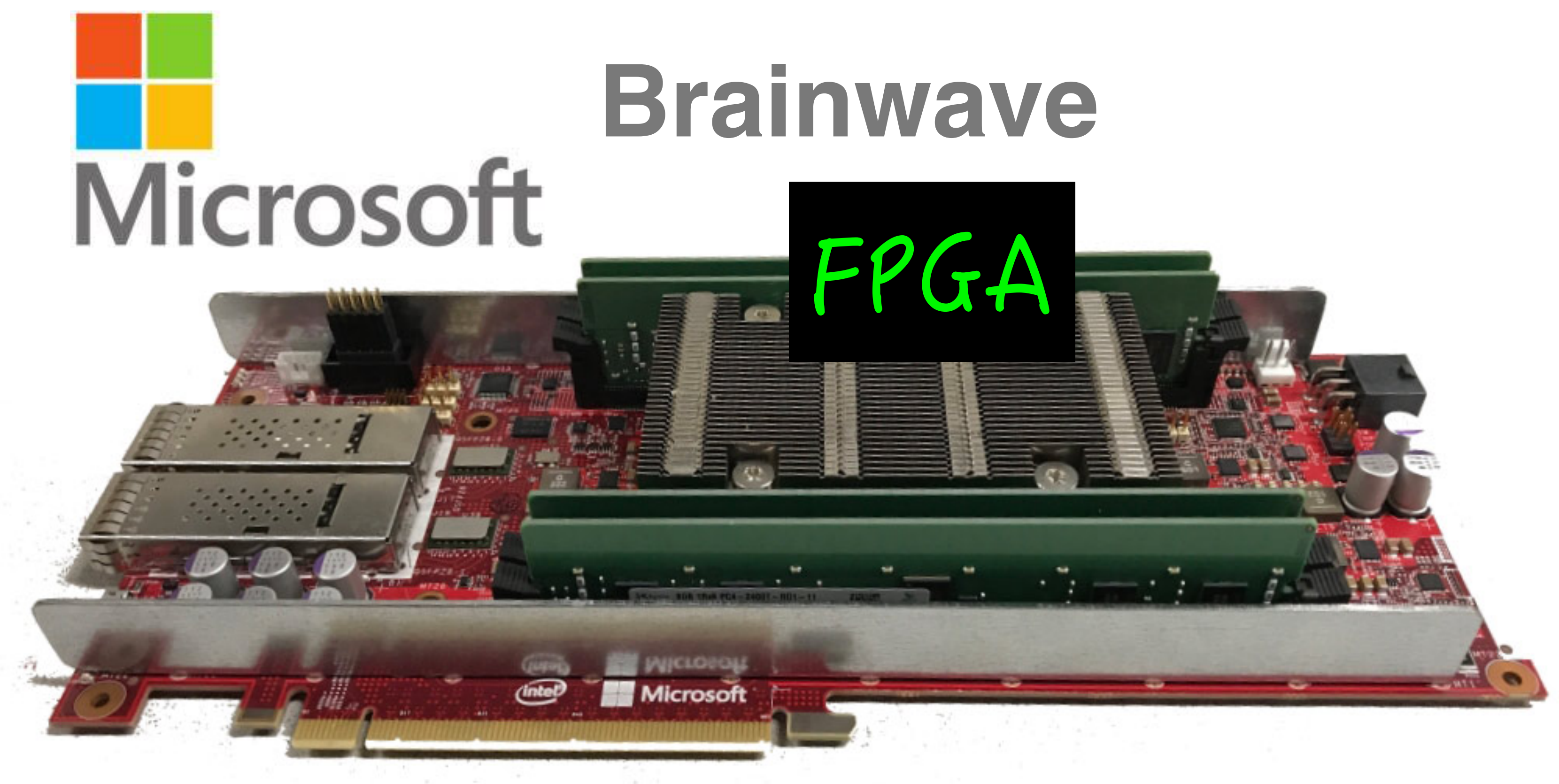


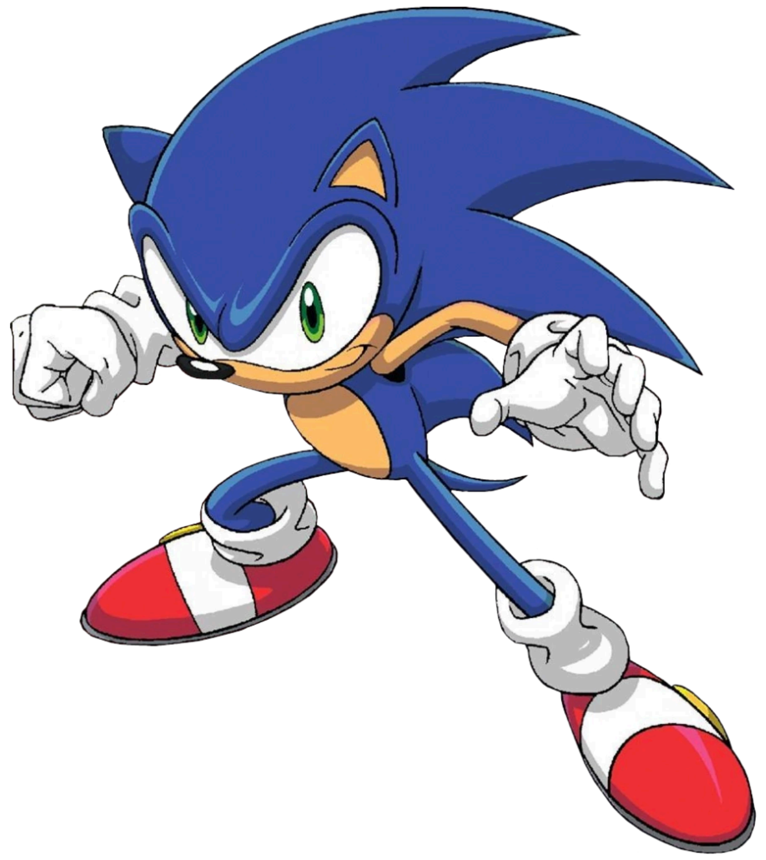
Pros:
scalable algorithms
scalable to the grid/cloud
Heterogeneous heterogeneity (mixed hardwares)



Pros:
less system complexity
no network latency

Services for Optimized Network Inference on Co-processors



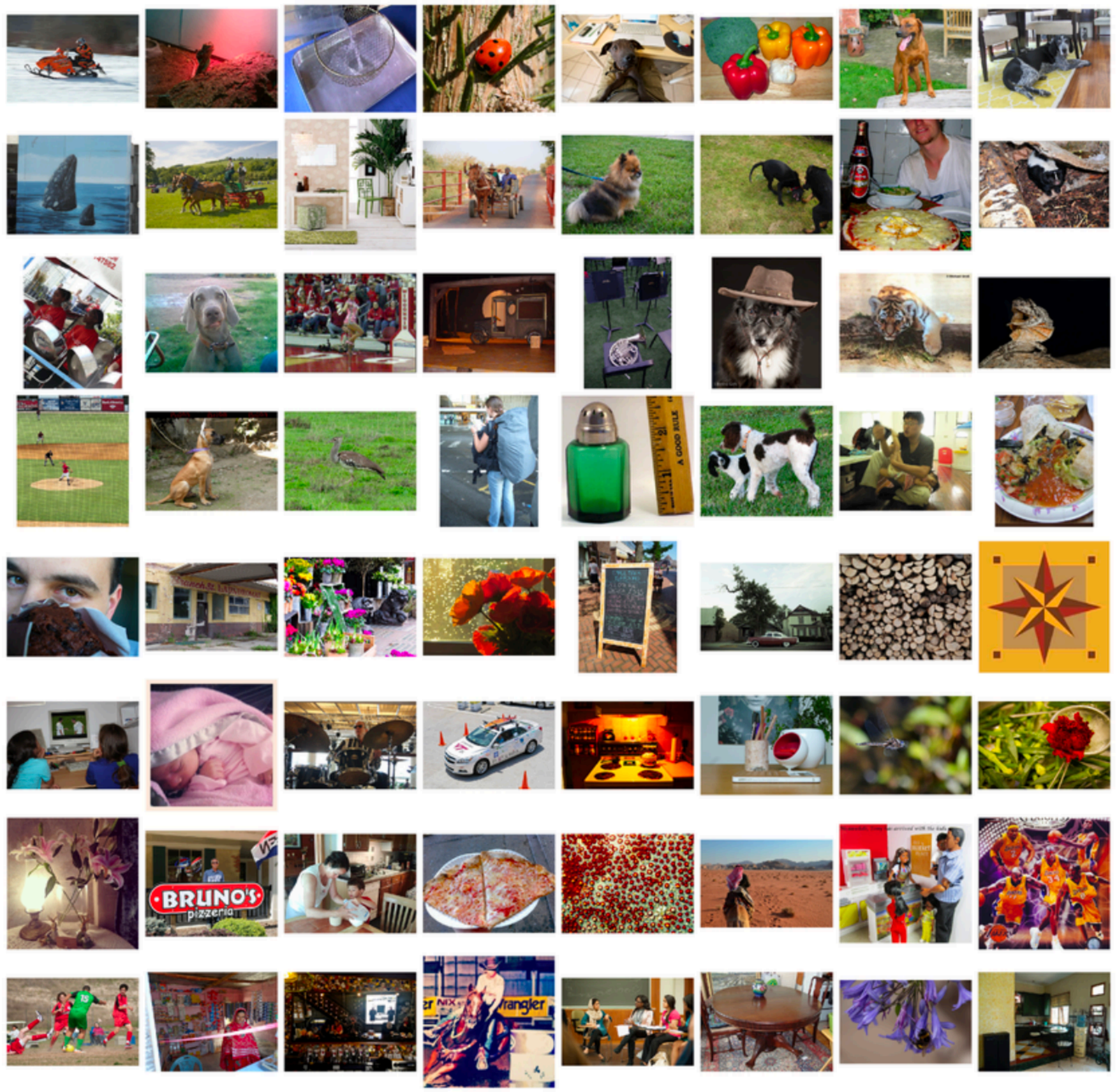


SONIC

Published in CSBS

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

‘Teach’ Res-Net 50 about particle physics



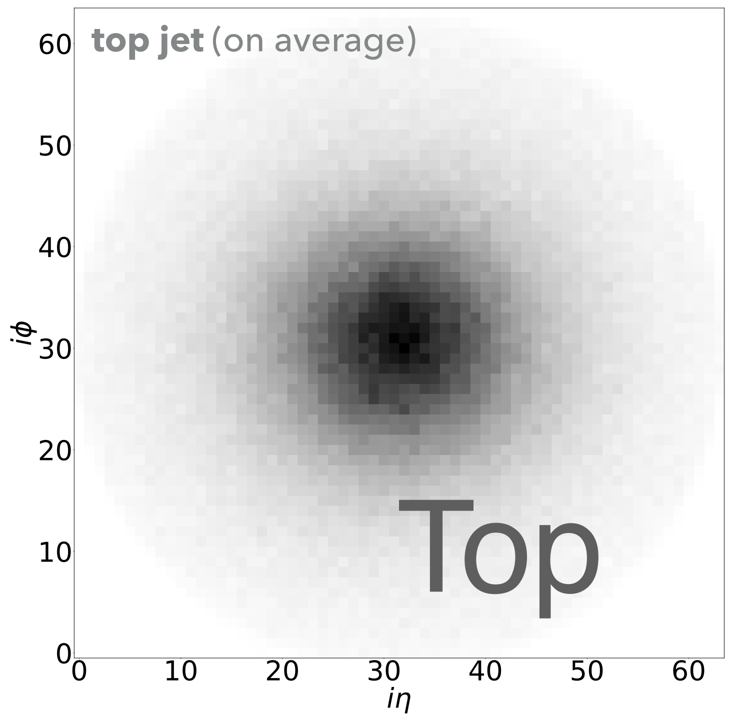
1000 classes
(cats, dogs…)

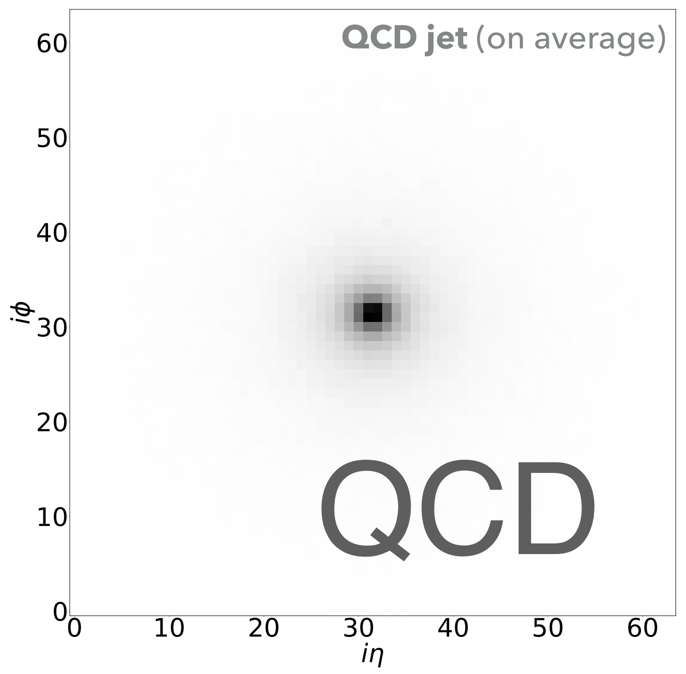
Res-Net 50 (25M parameters)

Featurizer
FPGA

Classifier
CPU

Train with top vs QCD labels

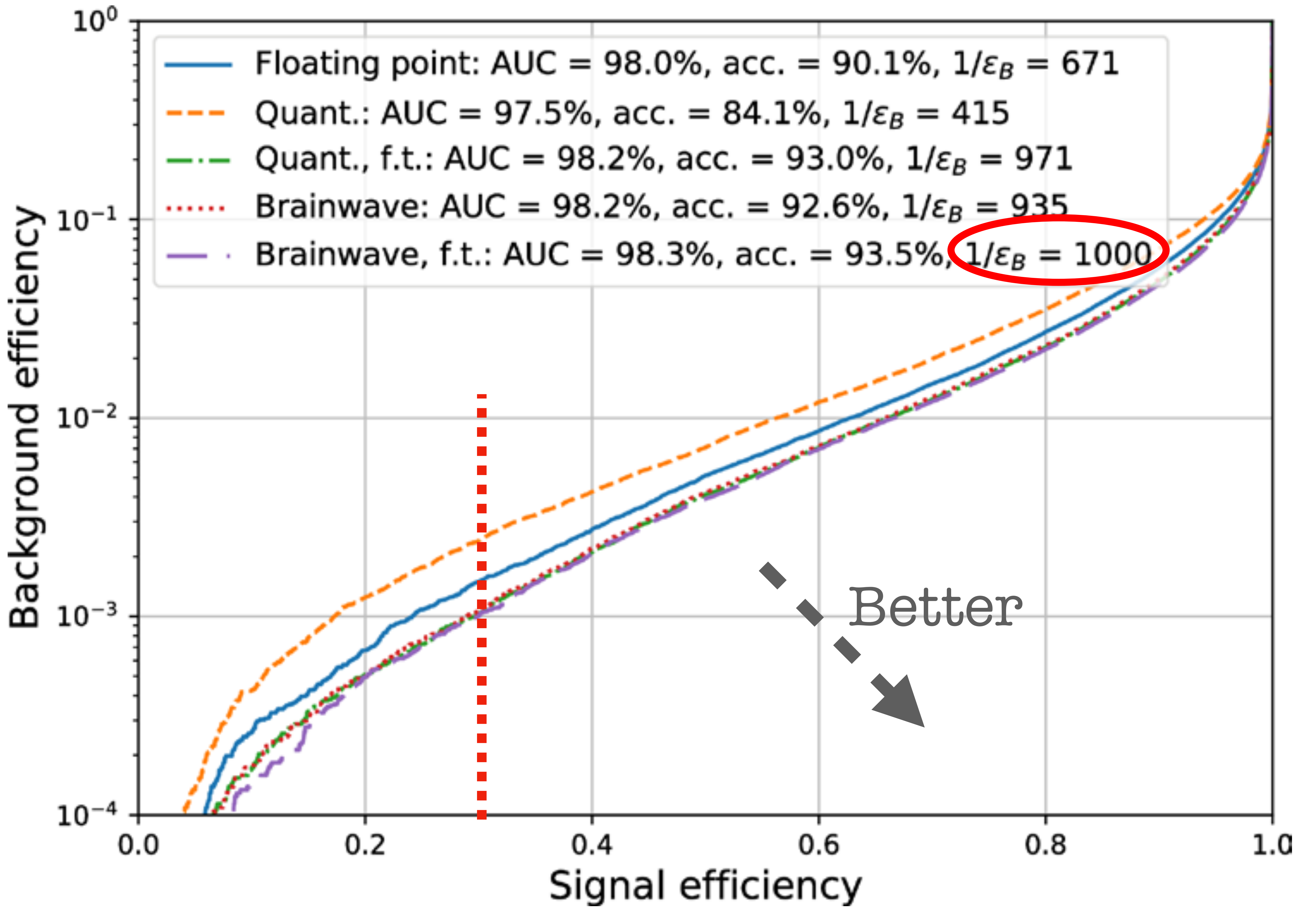




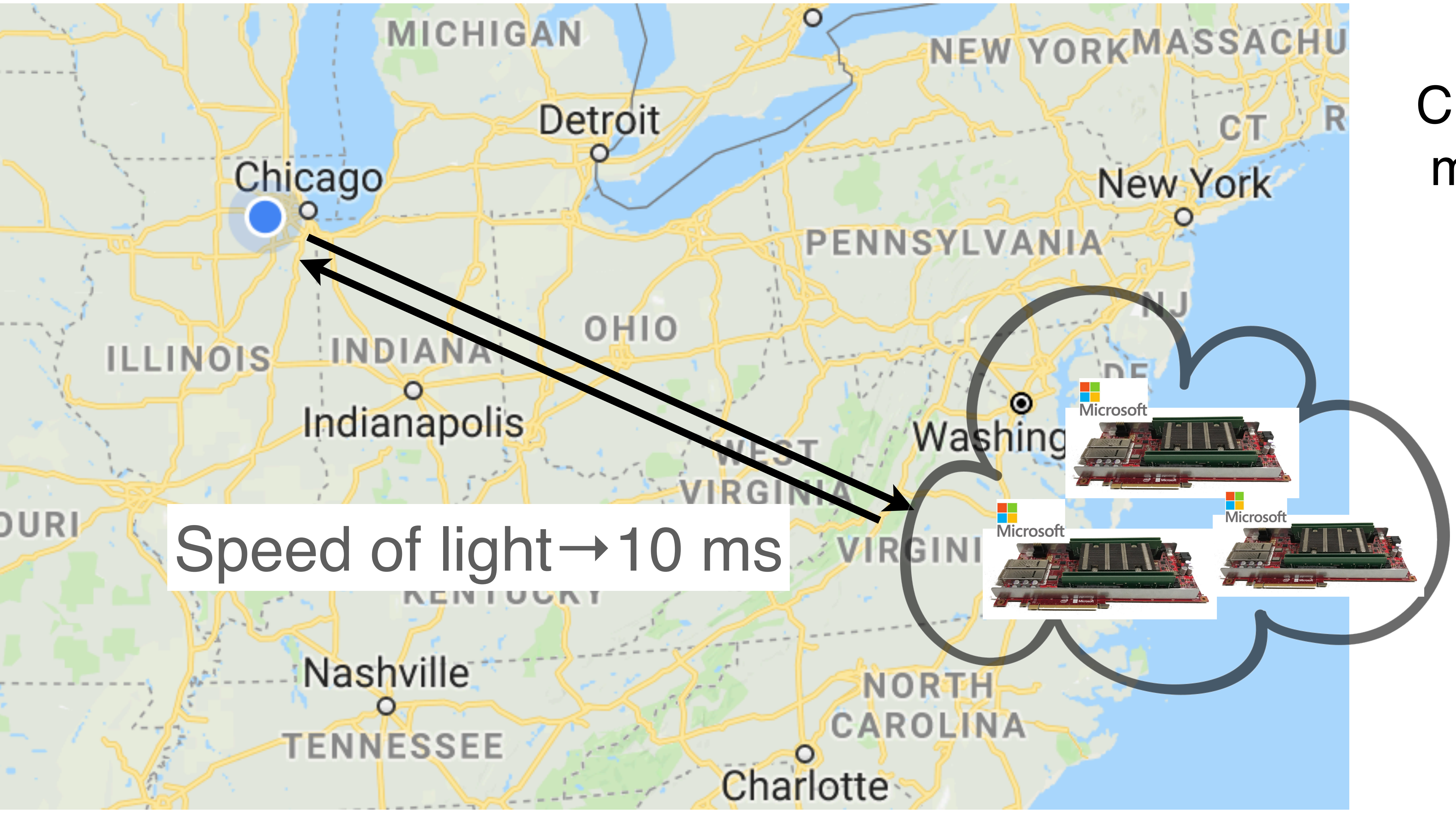
Quantized Res-Net 50 performance

Quantization matters:

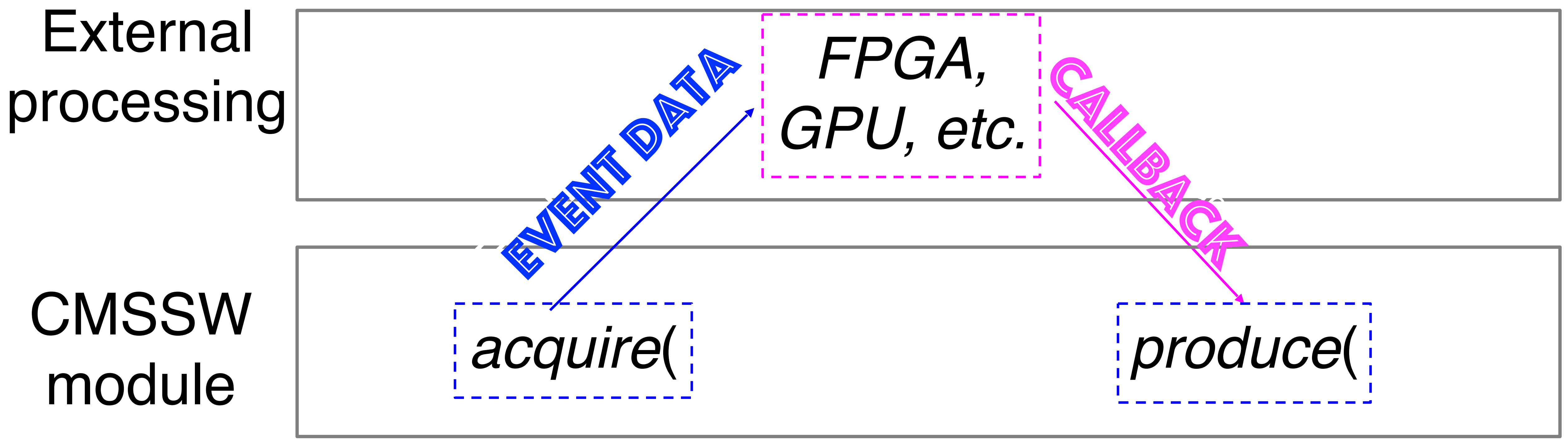
- Floating point —> Quantized model brainwave's implementation of ResNet50 on FPGA
  - Loss in performance
- Re-train the model with fixed precision regains the performance



Is it faster? Inference speed

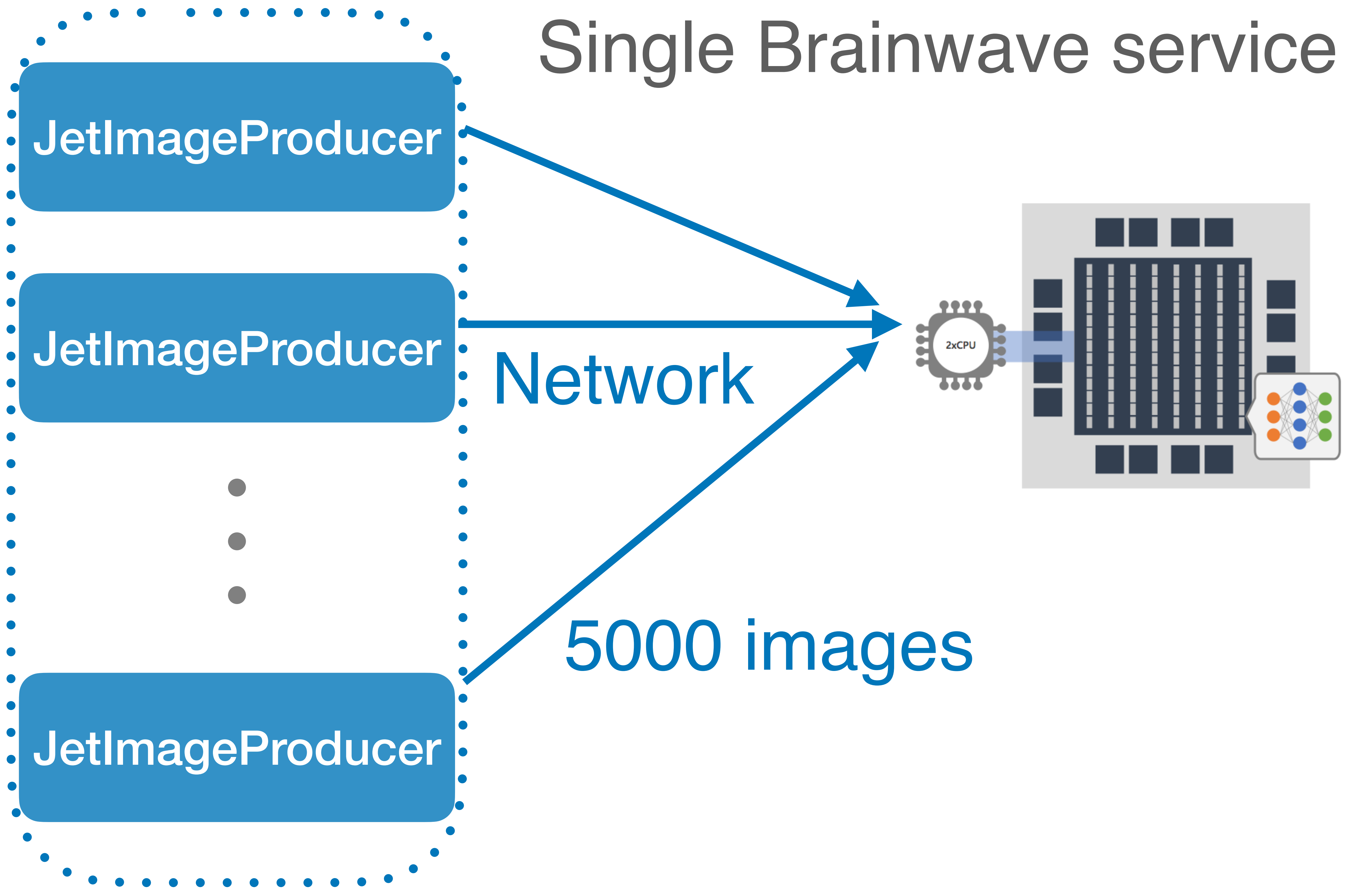


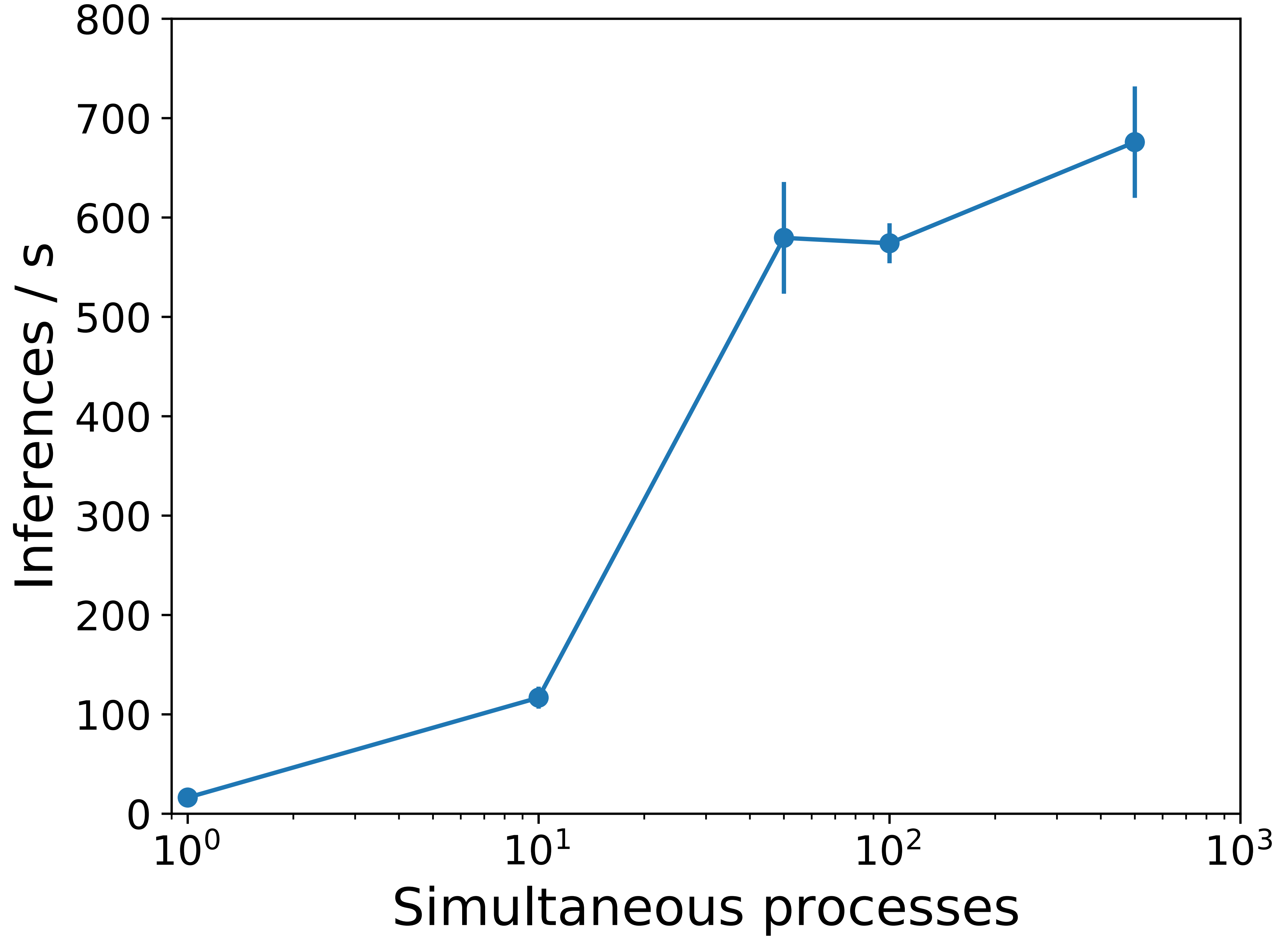
Test integrated in CMS software stack



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

Computing: data throughput





Max data throughout: 600-700 images/sec, Comparable with V100 GPU (with large batch sizes).

SONIC: latest explorations

GPU-as-a-service at the LHC

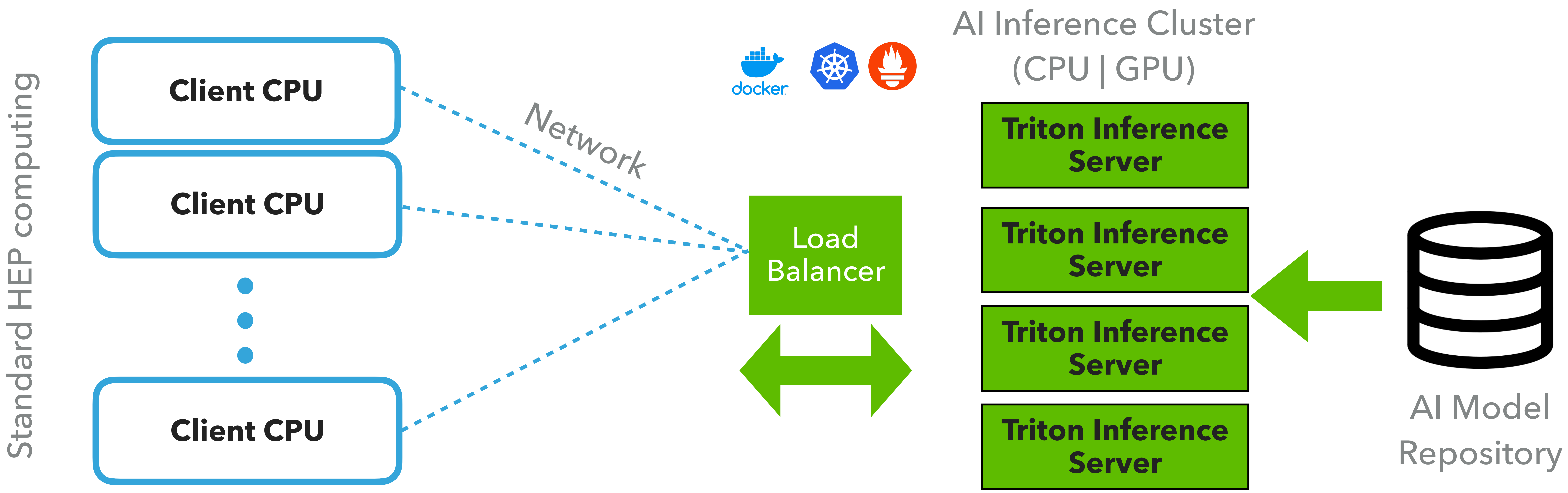
https://arxiv.org/abs/2007.10359

Hardware platforms

GPU-as-a-service for DUNE

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

GPU as-a-service with Triton



Example in neutrino: speedup, saturate GPUs

Accelerating Proto-DUNE reconstruction

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

Proto-DUNE

Largest LArTPC ever built

Busy environment: cosmic ray muons & beam

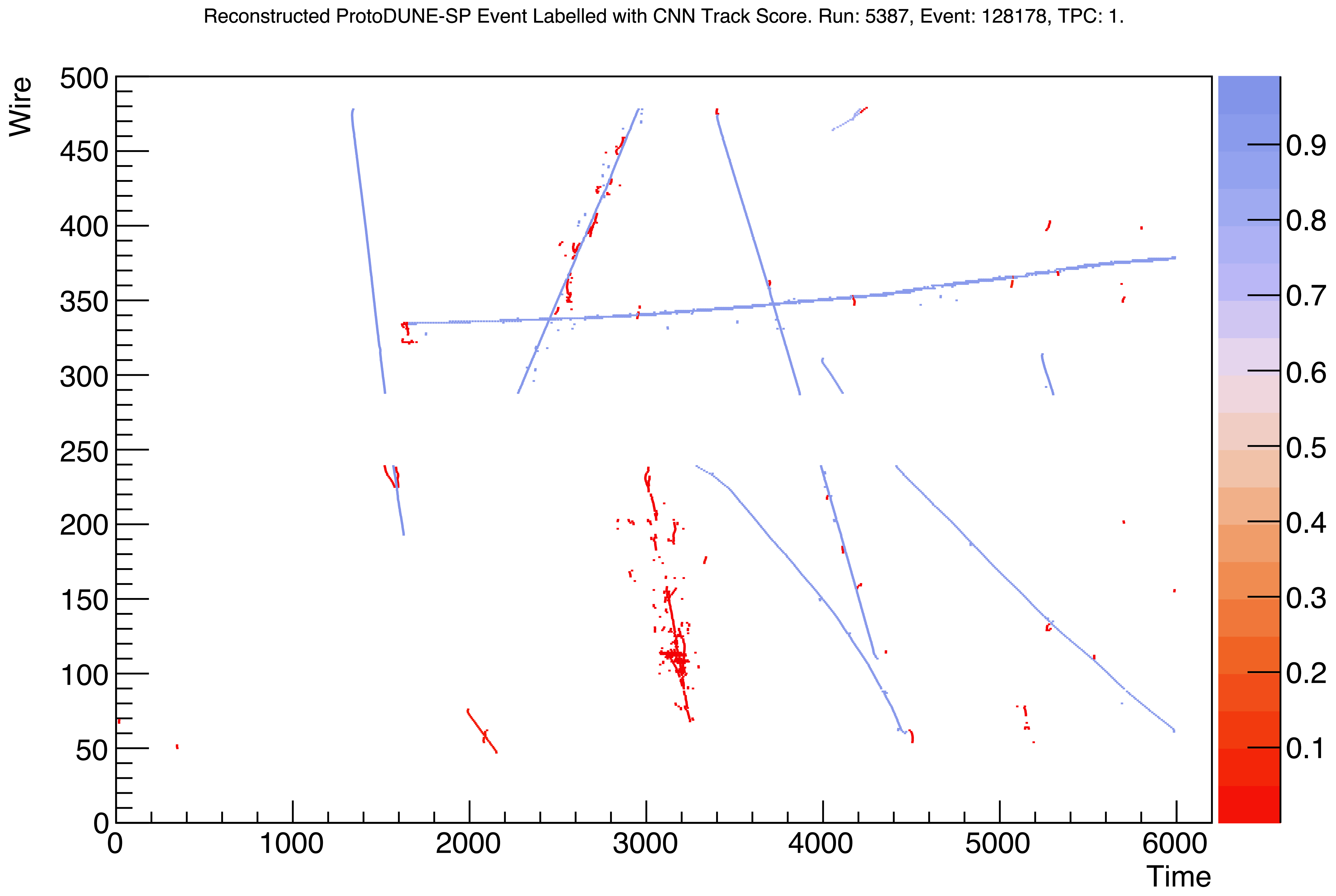
Reconstruction chain:

Noise mitigation,

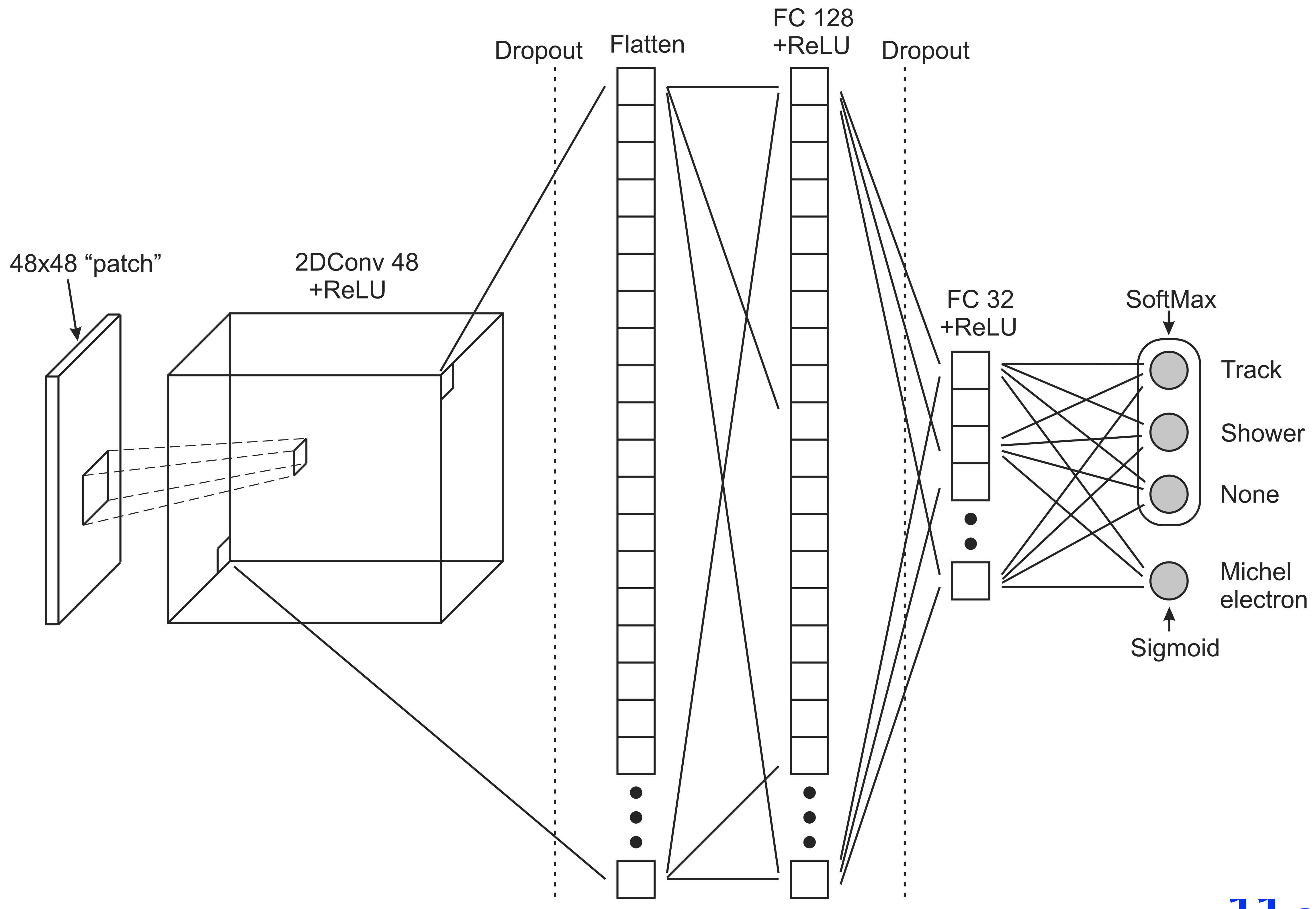
hit finding,

pandora pattern recognition,

-> CNN EmTrkMichelId



Speed up with GPU server



CNN EmTrkMichelId

~20x speedup of EMMichelTrackID module

CPU

Wall time (s)		
ML module	non-ML modules	Total
220	110	330

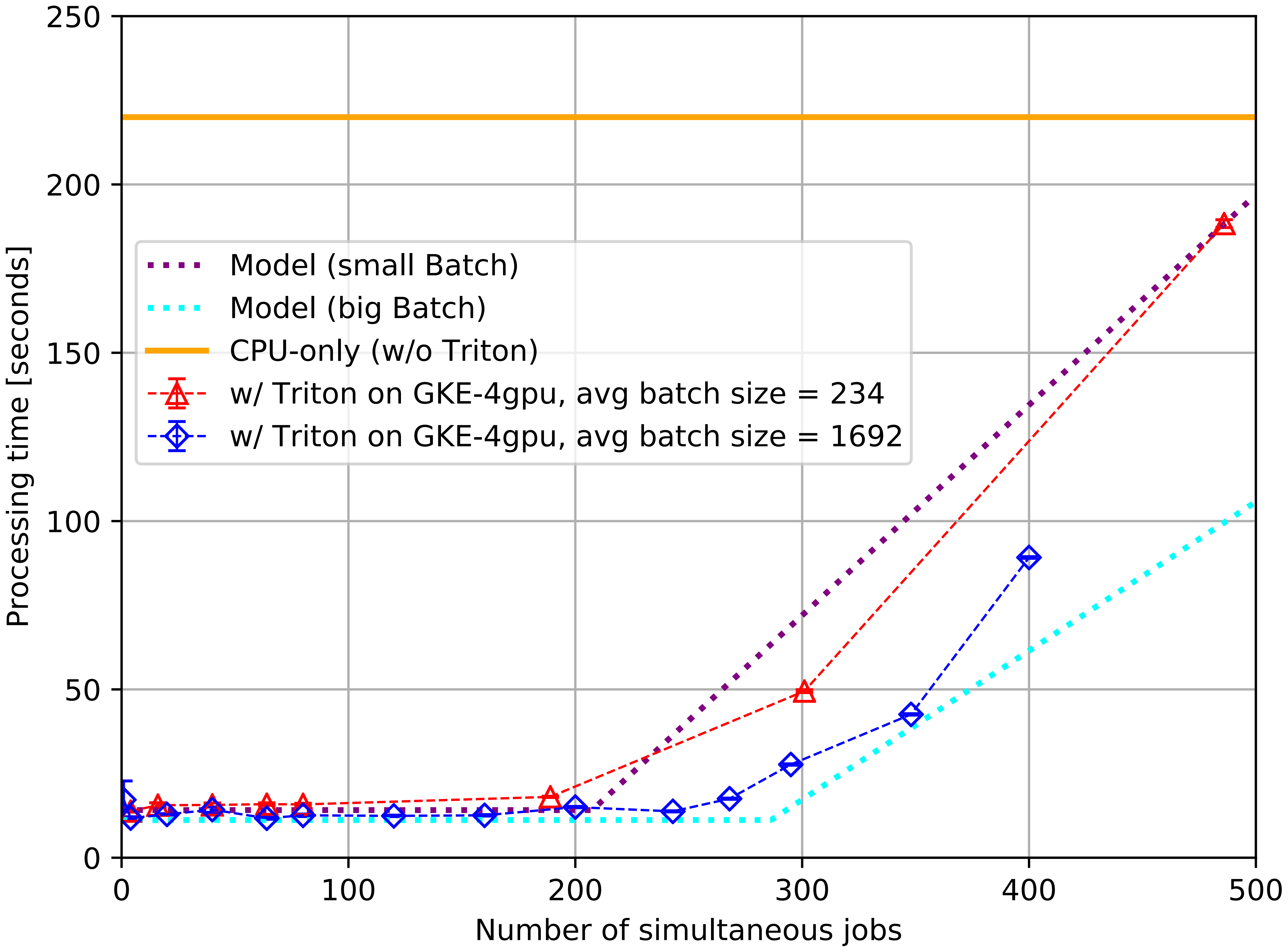
Single GPU server (NVIDIA T4)

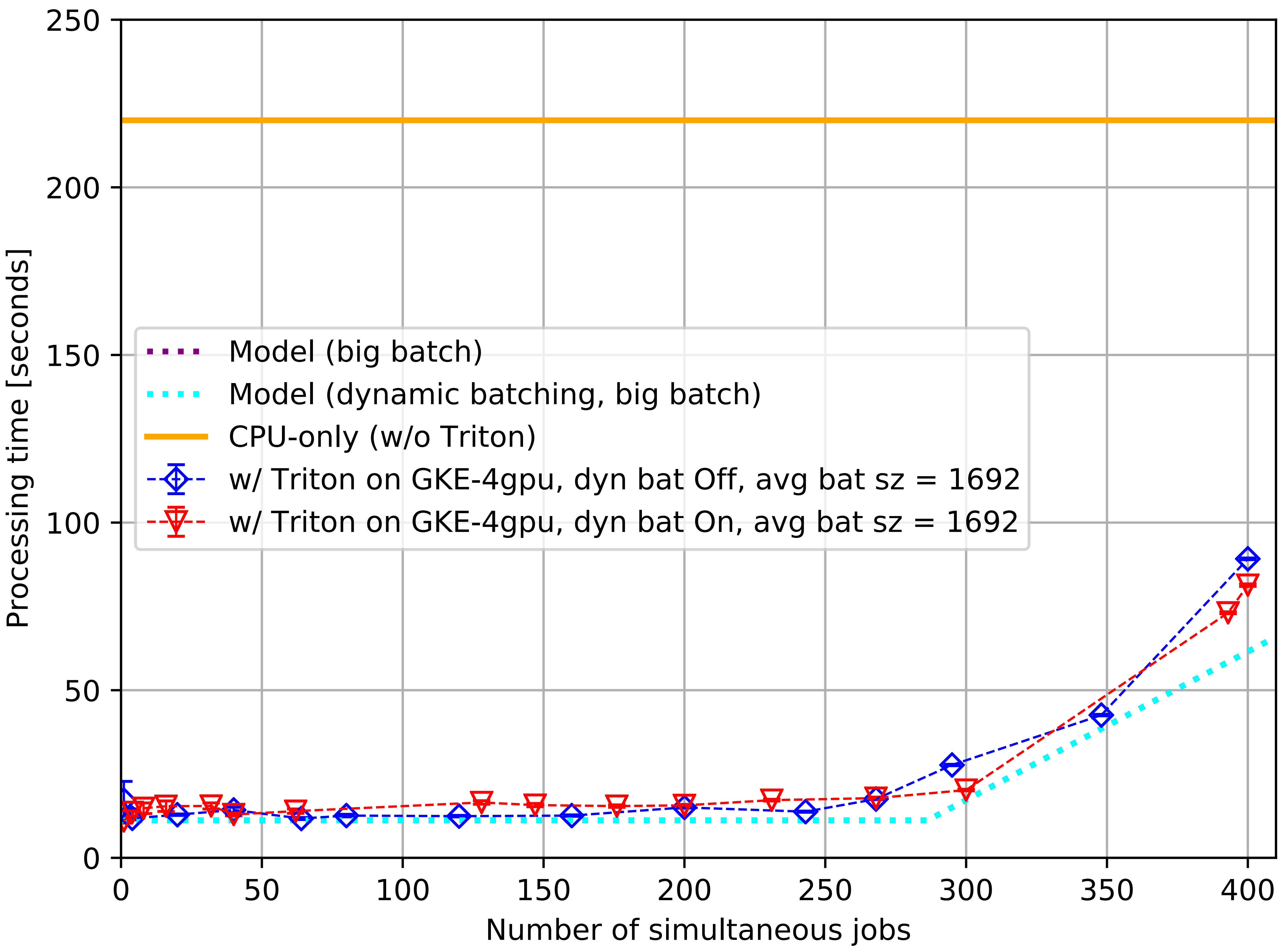
Wall time (s)		
ML module	non-ML modules	Total
~~220~~ ~11s	110	330

$11s \sim t_{preprocess} + t_{transmit} + t_{travel} + t_{GPU}$

$t_{preprocess}$	$t_{transmit}$	t_{travel}	t_{GPU}
7s	2s	0.4s	1.8s
On CPU, preparing NN inputs	Based on 2Gbps ethernet bandwidth	Ping latency between Iowa and FNAL	Time on the GPU

Saturating the GPUs





$$t_{\mathrm{SONIC}} = (1-p)\times t_{\mathrm{CPU}} + t_{\mathrm{GPU}}\left[1+\max\boxed{\left(0, \frac{N_{\mathrm{CPU}}}{N_{\mathrm{GPU}}} - \frac{t_{\mathrm{ideal}}}{t_{\mathrm{GPU}}}\right)}\right] + t_{\mathrm{latency}}$$

GPU saturates

2.7x speed up of the full ProtoDUNE-SP processing chain

1 GPU can handle 68 CPU processes simultaneously

SONIC: latest explorations

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

Hardware platforms

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

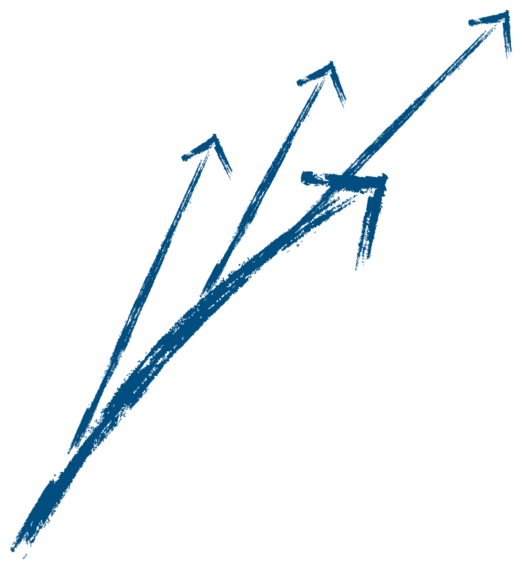
Algorithm complexity… at the LHC

FACILE

CMS Hadronic Calorimeter channel regression

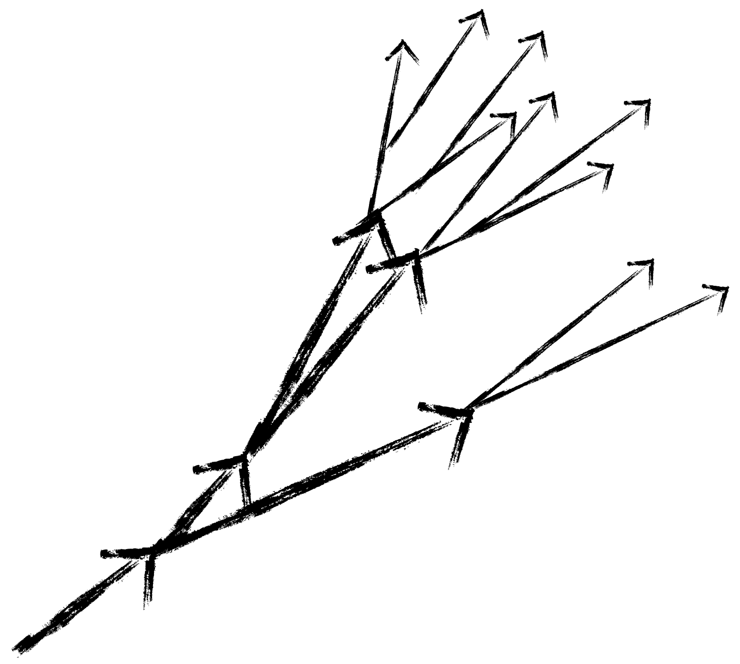
DeepCalo*

ECAL cluster regression



ResNet

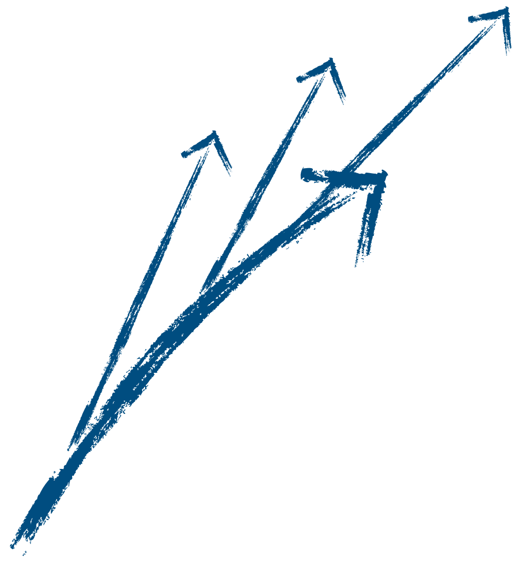
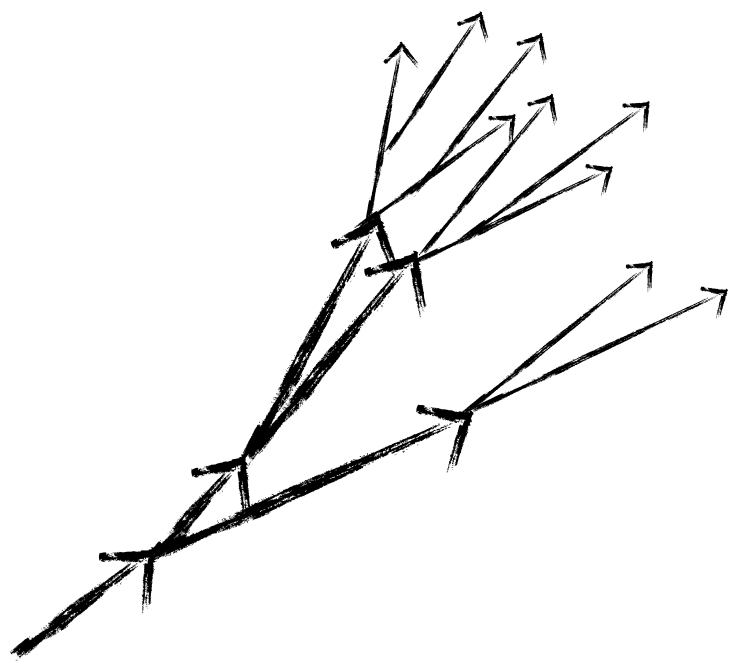
top quark image classification



2k parameters

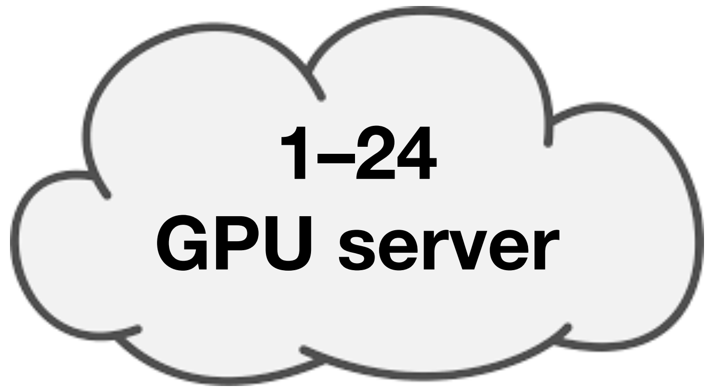
10 M parameters

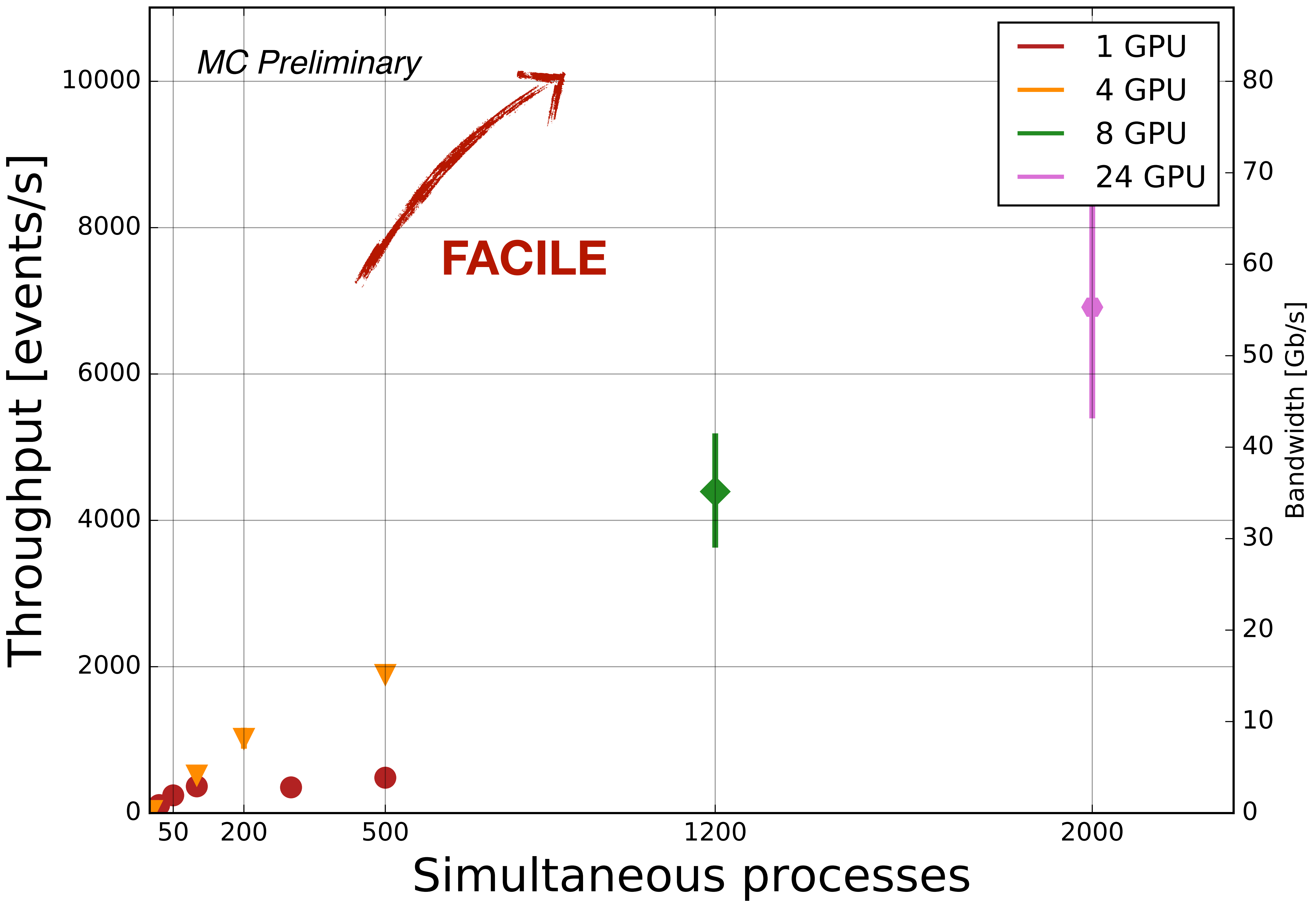
Single GPU server inference speedup

	FACILE	DeepCalo*	ResNet
	HCAL channel regression	ECAL cluster regression	top quark image classification
			
CPU	16 ms	75 ms	~1 s◆
GPU as-a-service	2 ms (GPU)	0.1 ms	1-2 ms (GPU/FPGA)
Gain	**8x (GPU)**	**750x**	**500x**

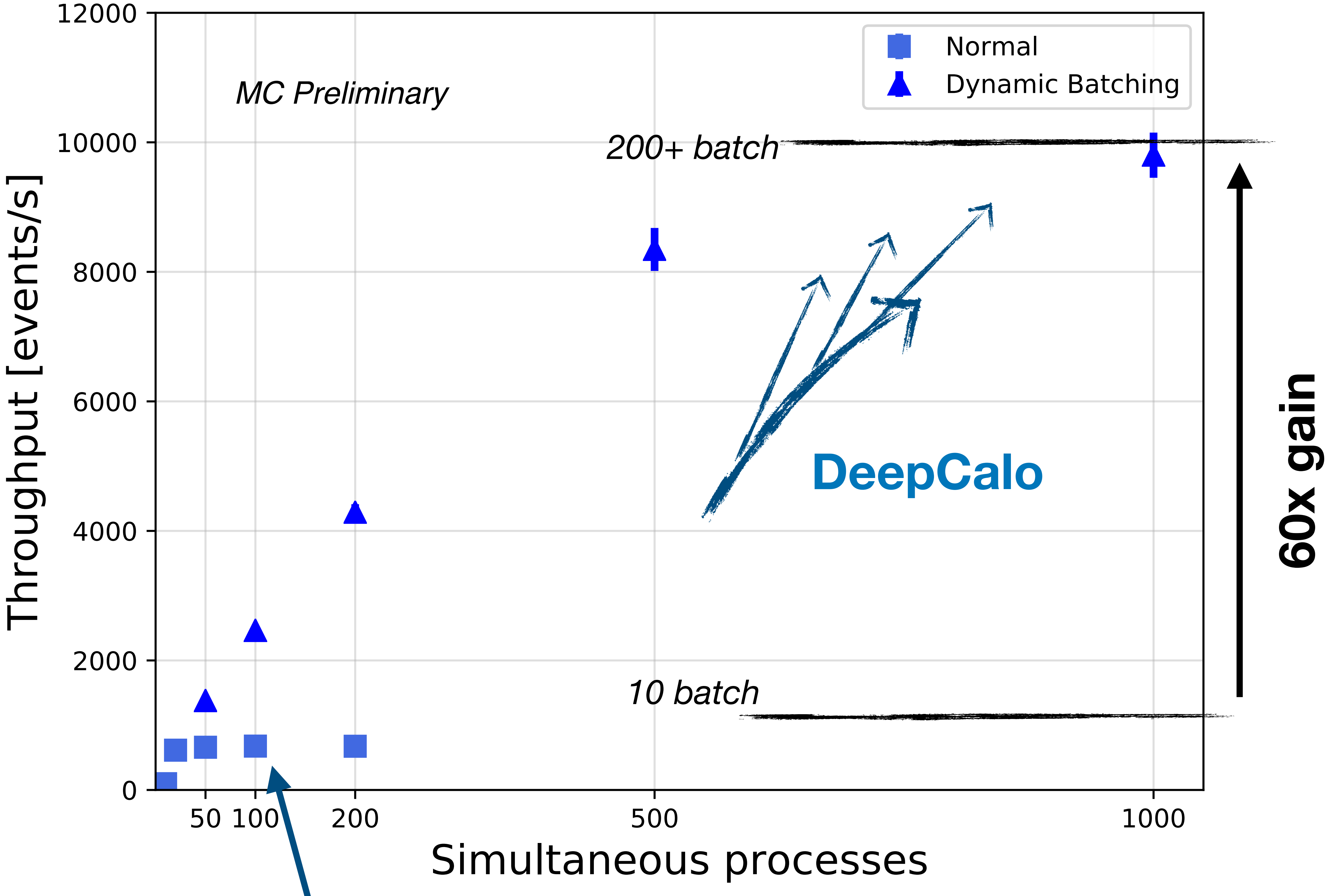
GPU scaling and Dynamic batching

M.LIU



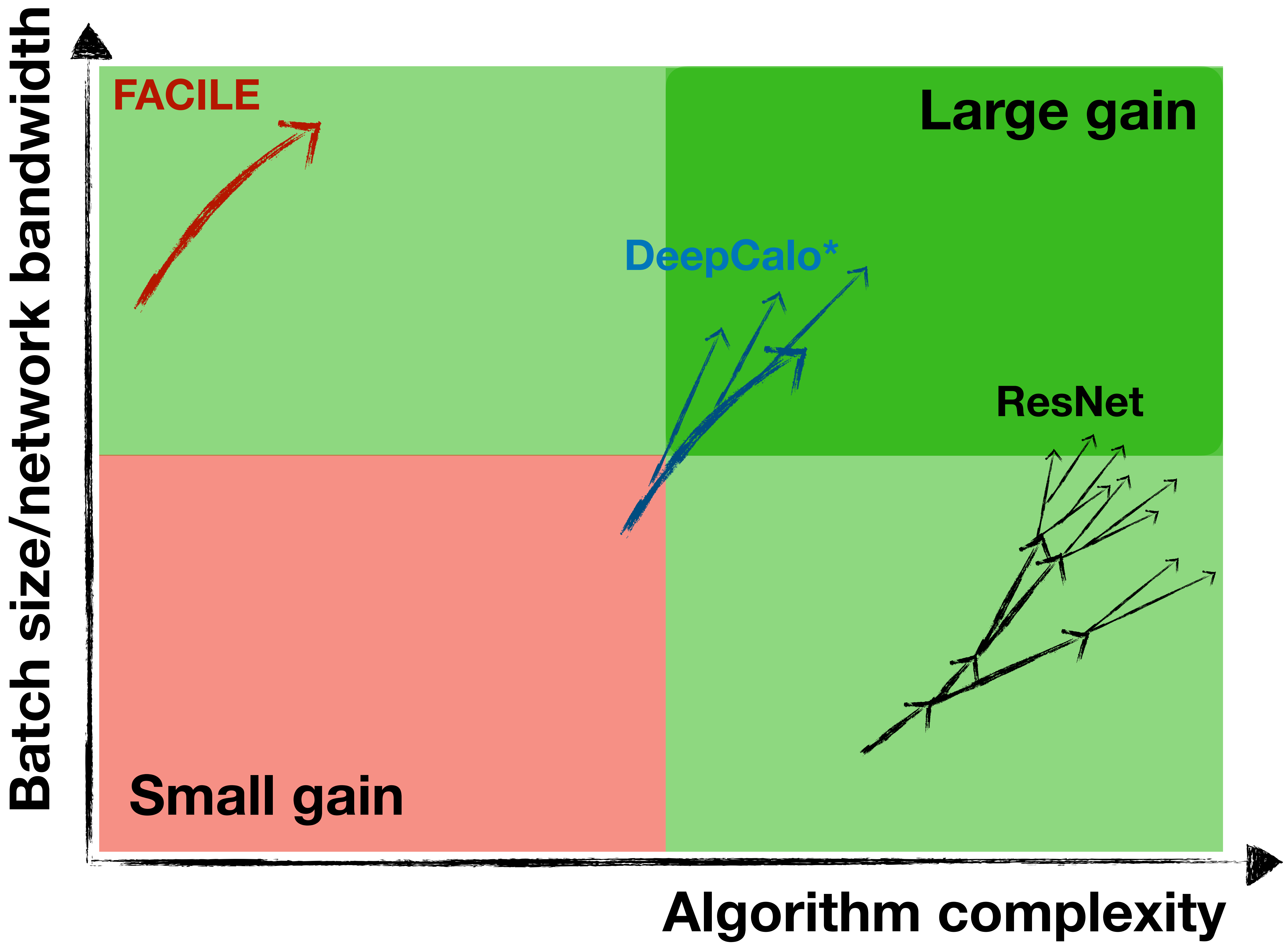


Significant gain from dynamic batching



GPU usage: 45%

Where do we gain?



SONIC: latest explorations

FPGA-as-a-service Toolkit

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

Hardware platforms

Open source tools: flexibility

hls4ml

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

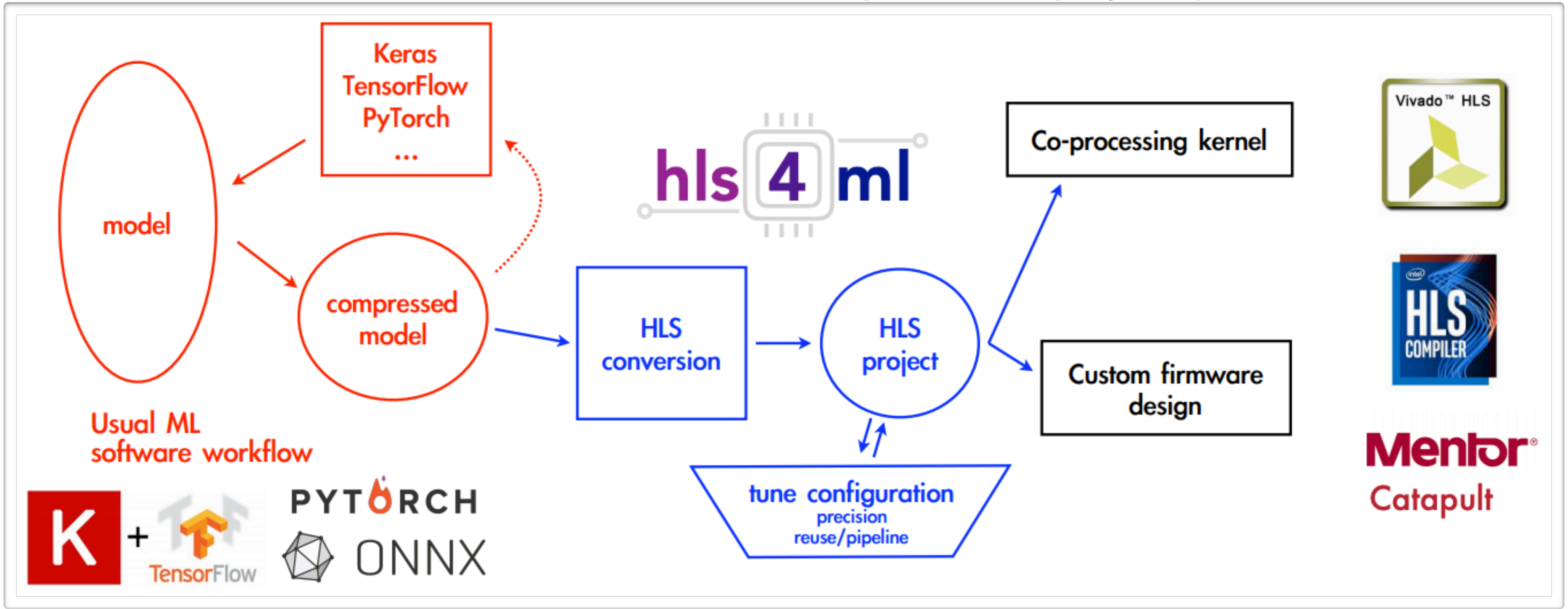
Algorithm complexity

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

hls4ml: accelerating ML on hardware

fastmachinelearning.org/hls4ml

An open source project - join the conversation!



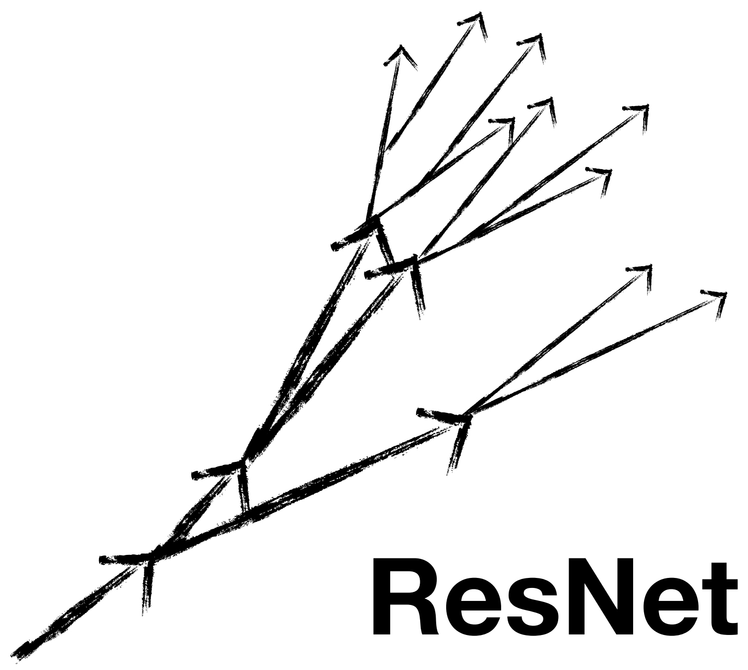
Originally designed for LHC triggers applications but broad and growing user base

FPGAs-as-a-Service Toolkit (FaaST)

VITIS + hls4ml

FACILE

Algorithm	Platform	Number of Devices	Batch Size	Inf./s [Hz]	Bandwidth [Gbps]
FACILE	AWS EC2 F1	1	16,000	36 M	23
FACILE	Alveo U250	1	16,000	86 M	55
FACILE	T4 GPU	1	16,000	8 M	5.1
ResNet-50	AWS EC2 F1	8	10	1400	6.7
ResNet-50	V100 GPU	8	10	1,700	8.1
ResNet-50	ASE	1	1	460	2.2
ResNet-50	T4 GPU	1	10	250	1.2



ResNet

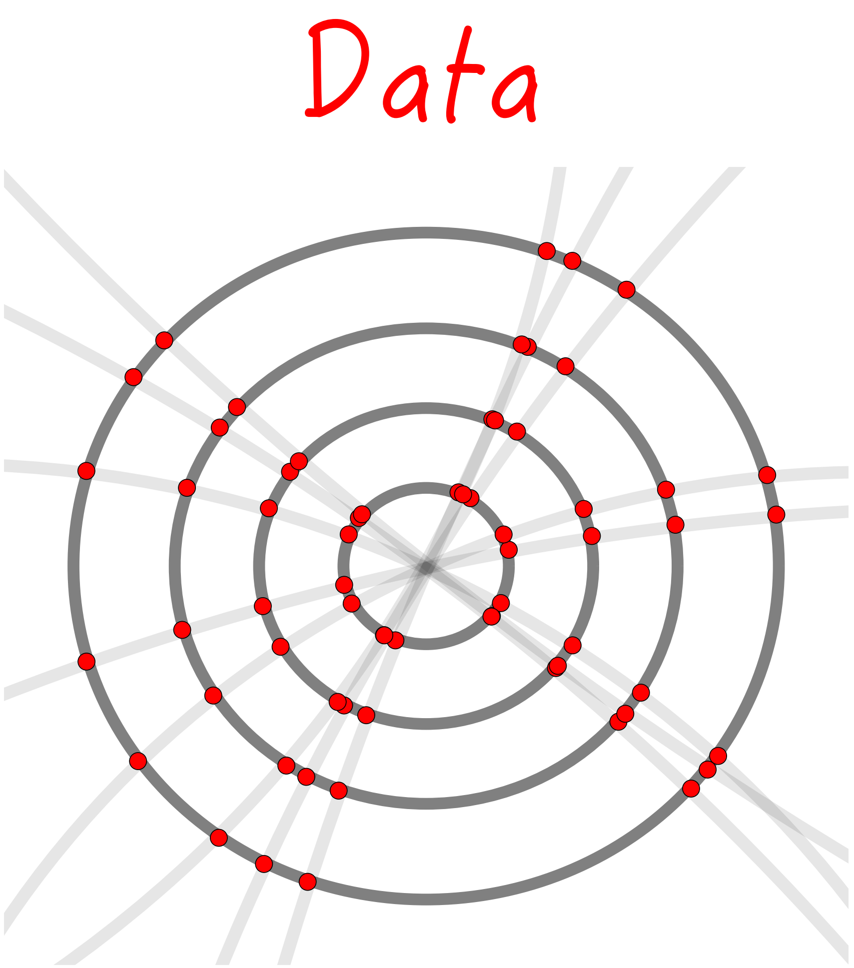
Xilinx Machine Learning Suite

FPGA-as-a-service Toolkit

Future prospects

Integration in full scale production in experiments

: processing for full-scale protoDune-SP reconstruction, FACILE@HLT in CMS…scaling with multiple models.

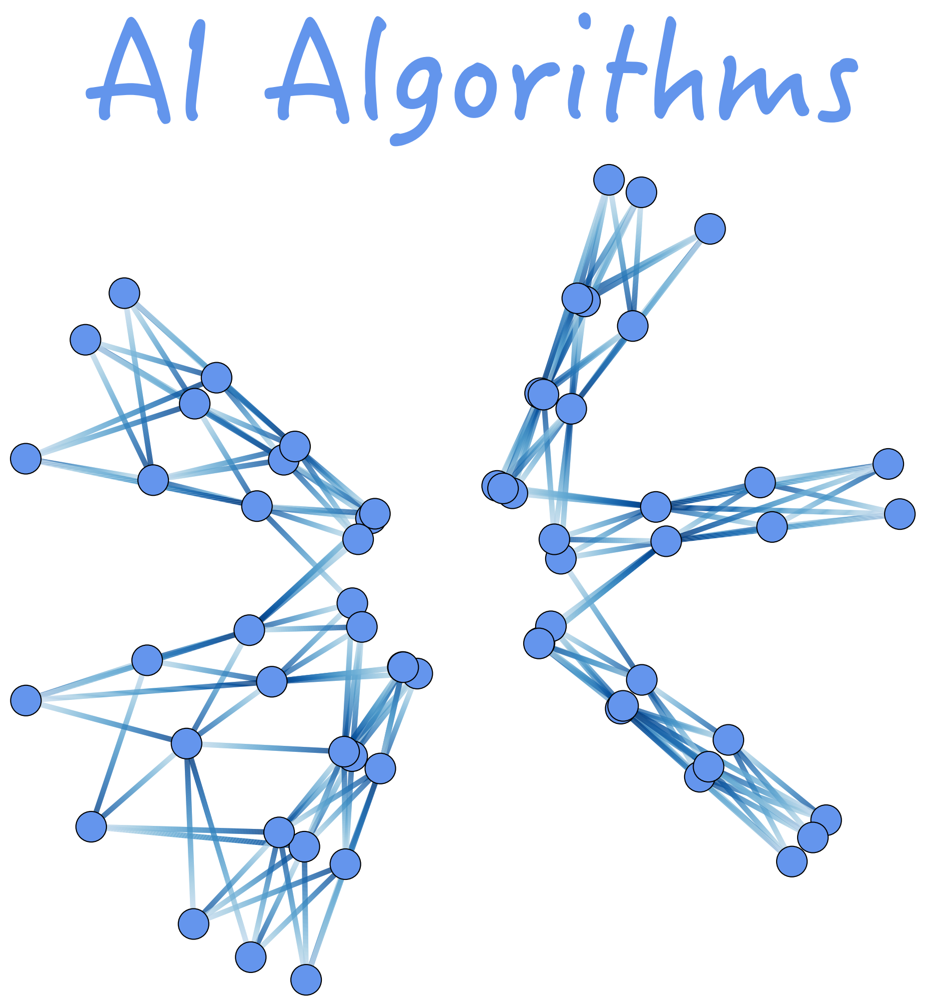


AI algorithms suitable for physics data/with domain knowledge embedded

Graph neural networks, Energy flow networks, see Nhan's talk.

Hardware awareness in training for co-processor workflow

e.g. Brainwave studies explore re-training with quantized version to achieve the best performance in precision.



Most studies performed using Cloud services/on-premises clusters

SONIC in High Performance Computers (HPCs)

: accelerate ML-based simulation/reconstruction etc…

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

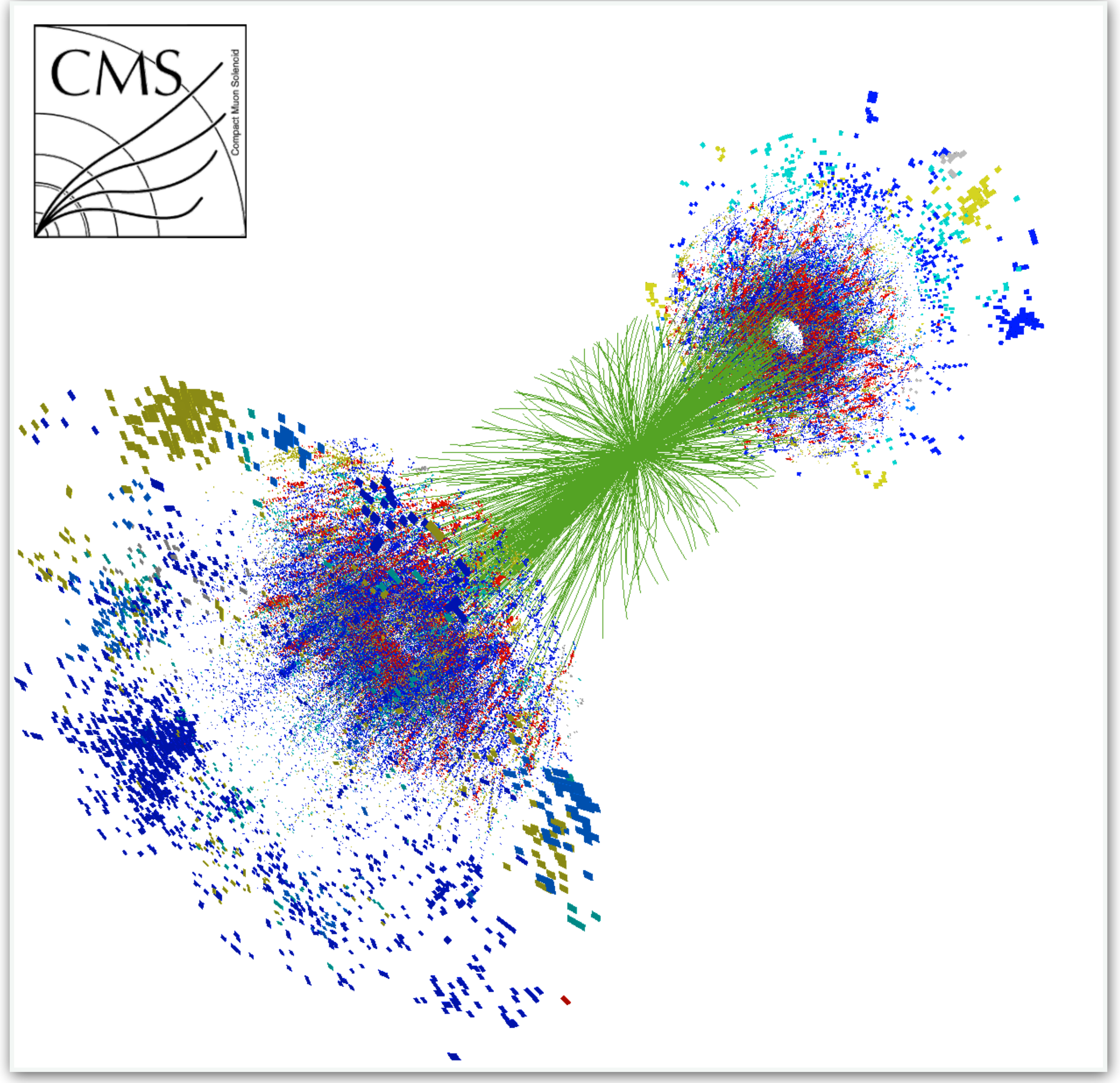
- e.g.*On GPUs*

re-cast physics problem as a machine learning problem
Language: C++, Python (TensorFlow, PyTorch,…)

Hardware: FPGA, GPU,ASIC

Tracking with ML

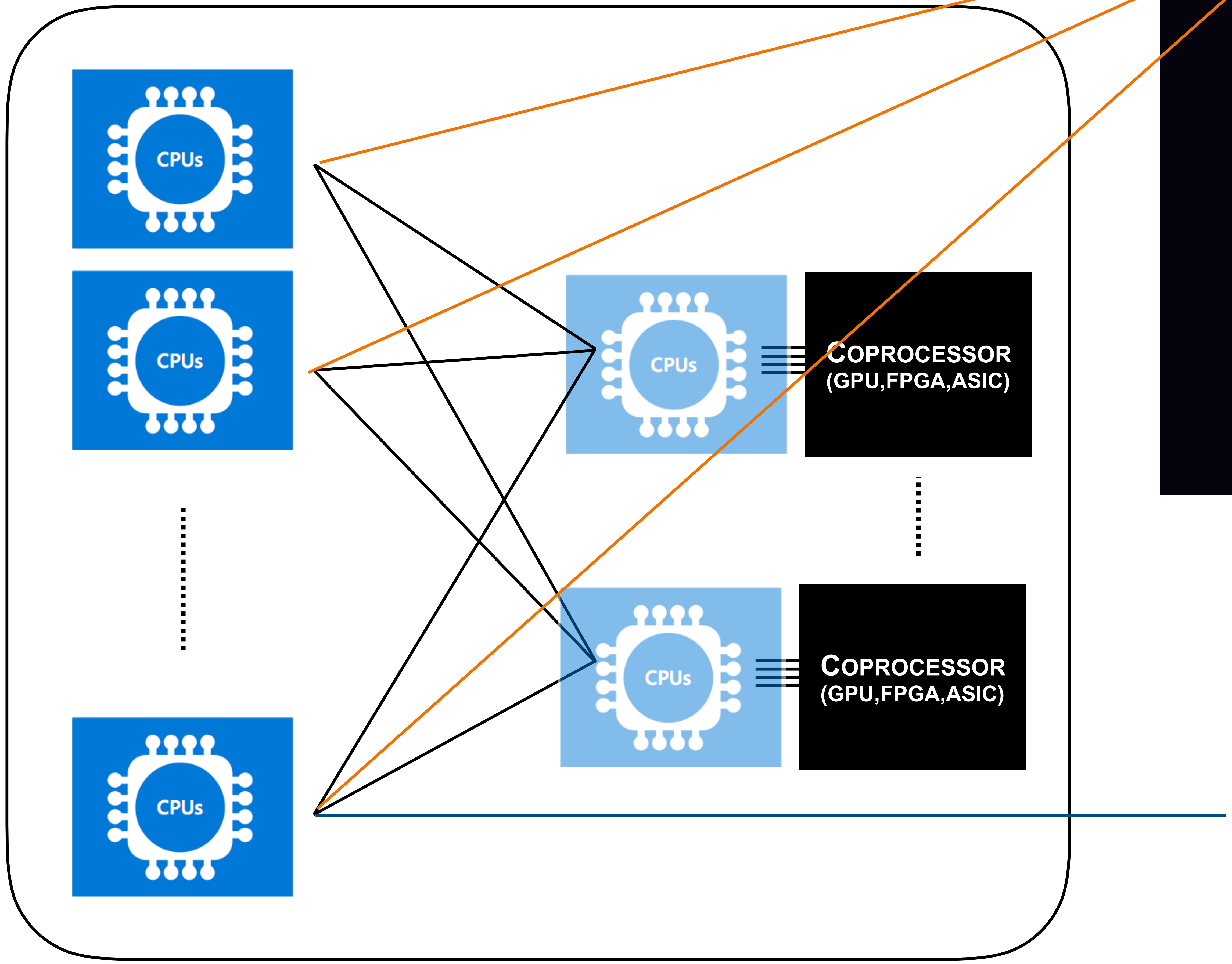
- Algorithms parallelizable
- Solutions with ML e.g., HEP.TrkX.

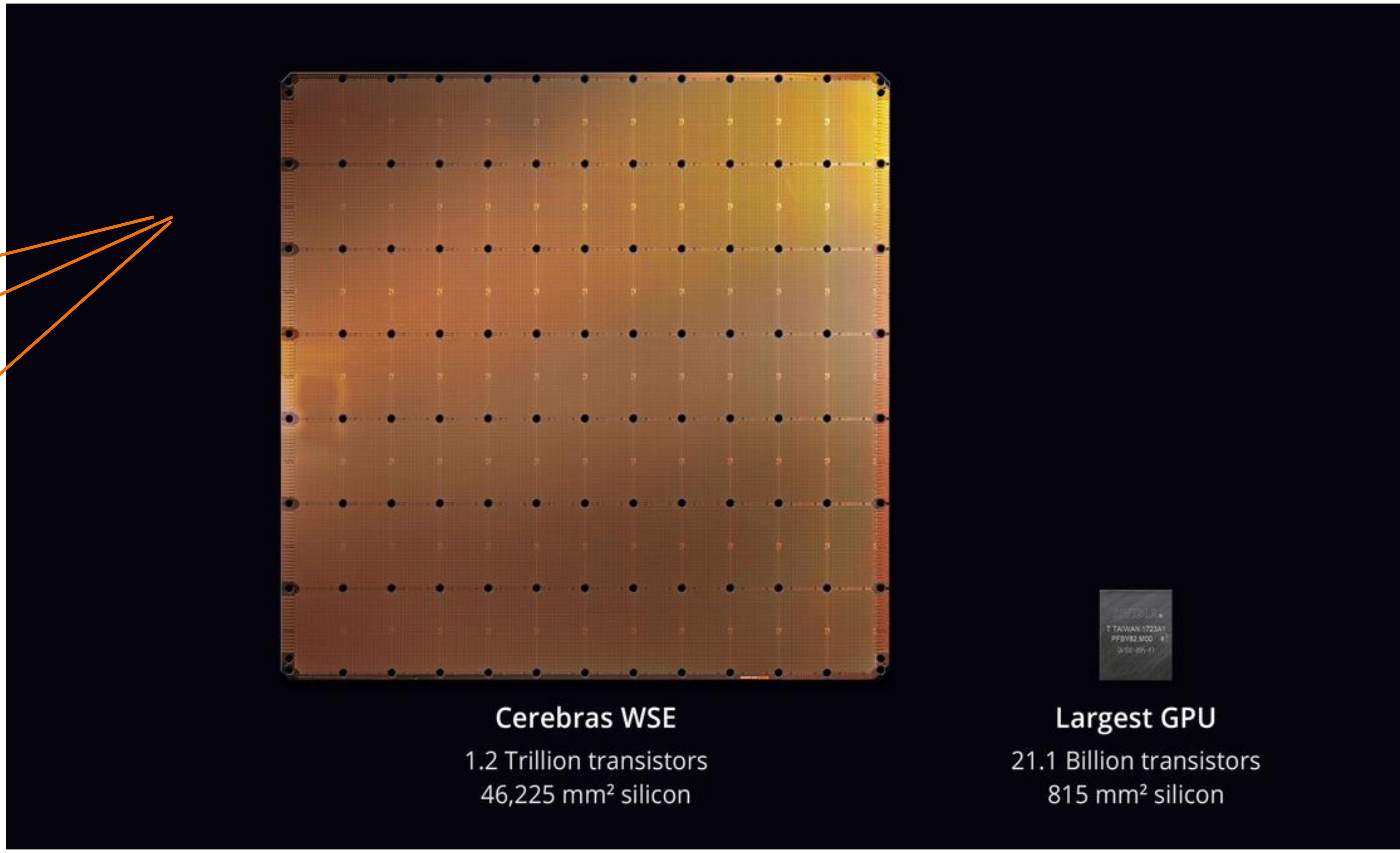


Charged particle Tracking With graph neural networks

The now/future computing paradigm?

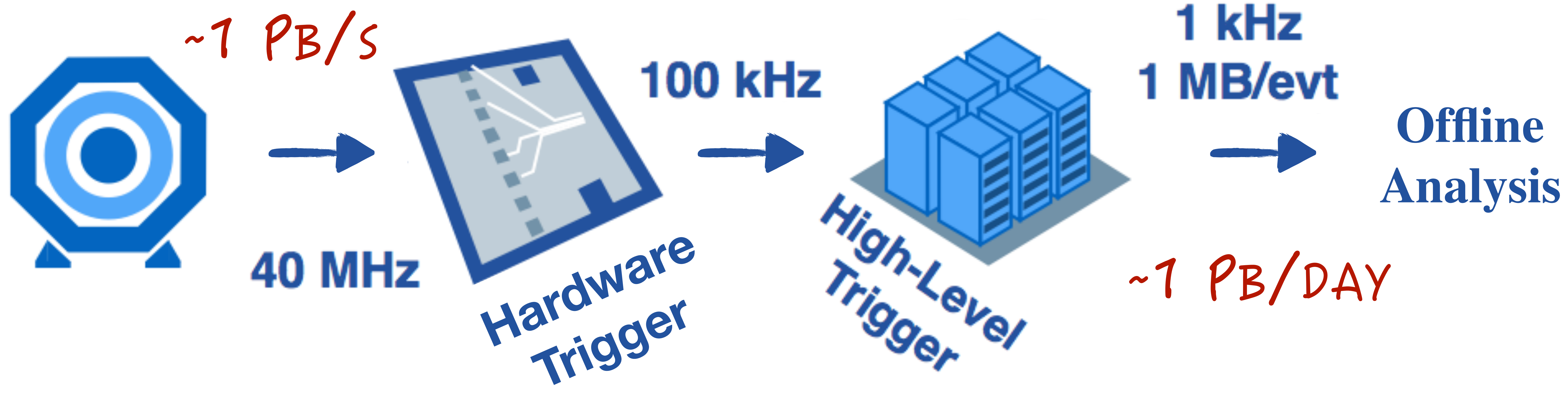
Heterogeneous computing as-a-service



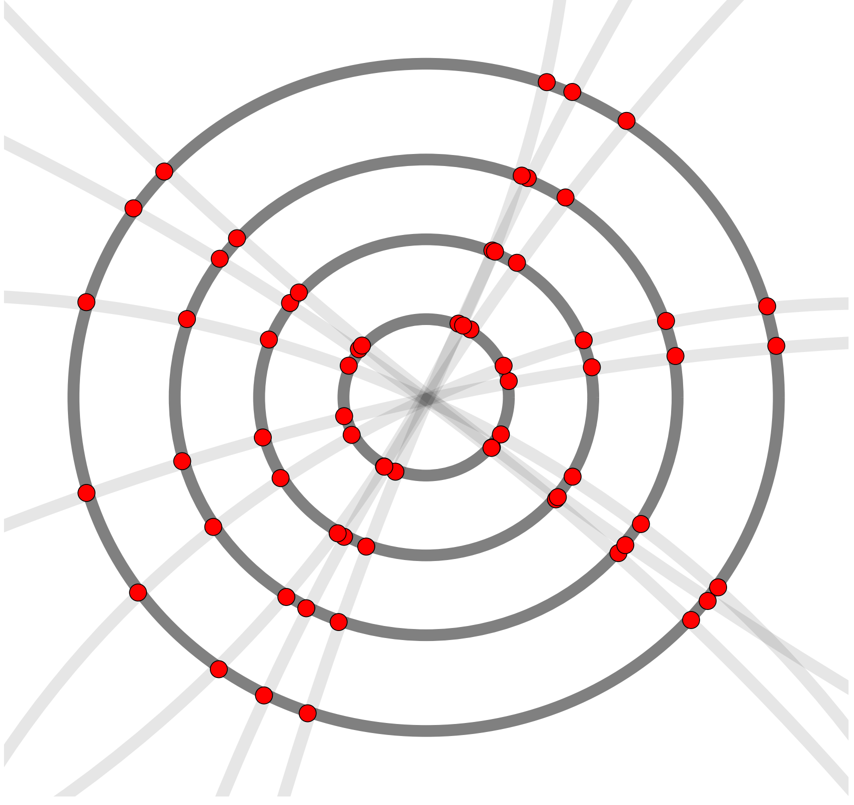


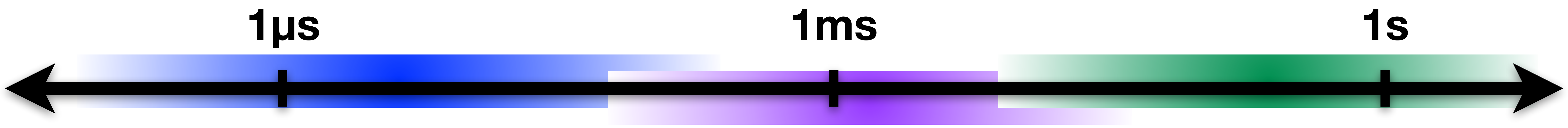
Emerging technologies…

Accelerated discoveries with Real-Time AI



Data

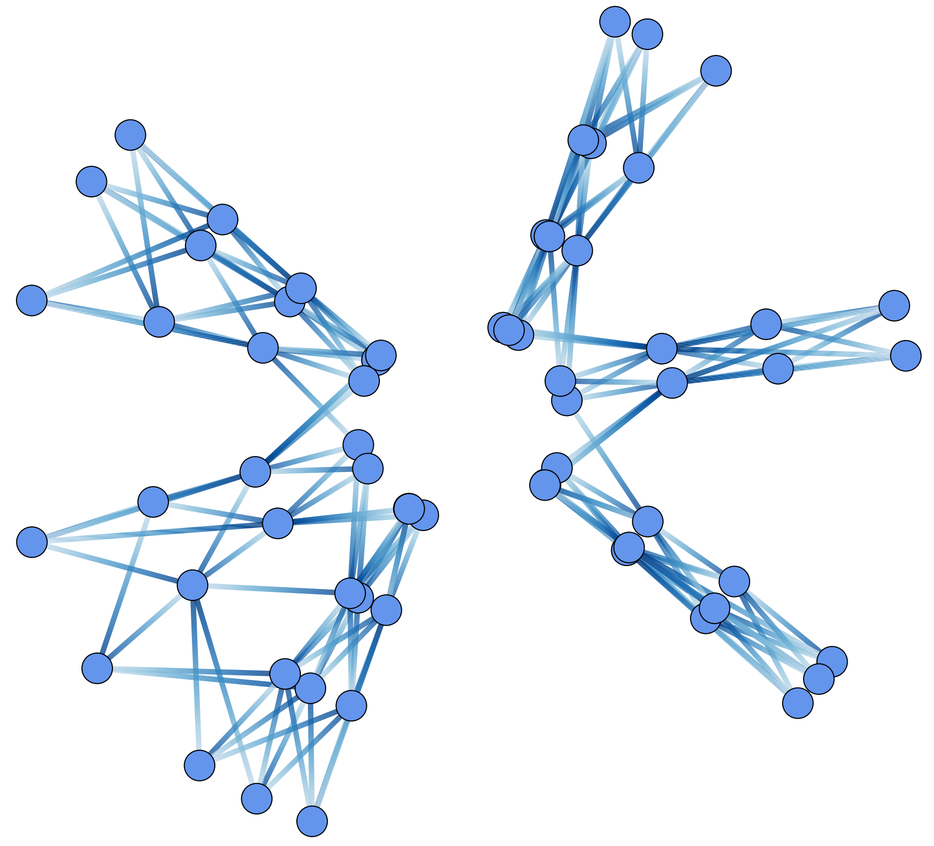




ASIC/FPGA

Heterogenous computing /CPU

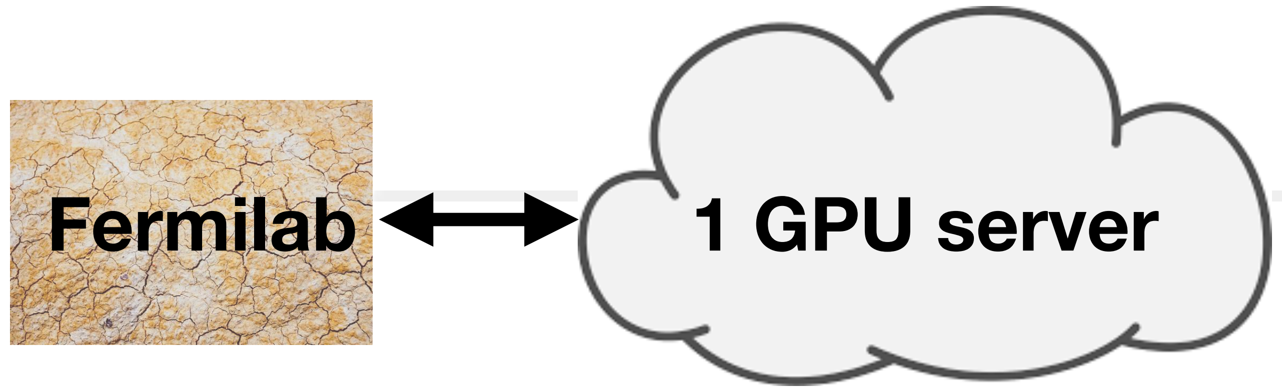
AI Algorithms



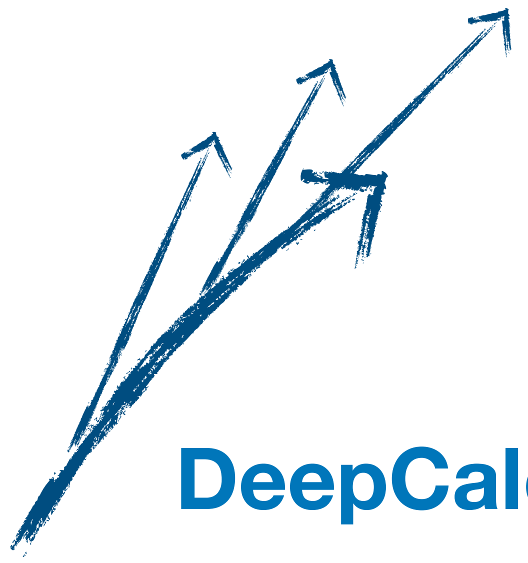
hls4ml

Co-design

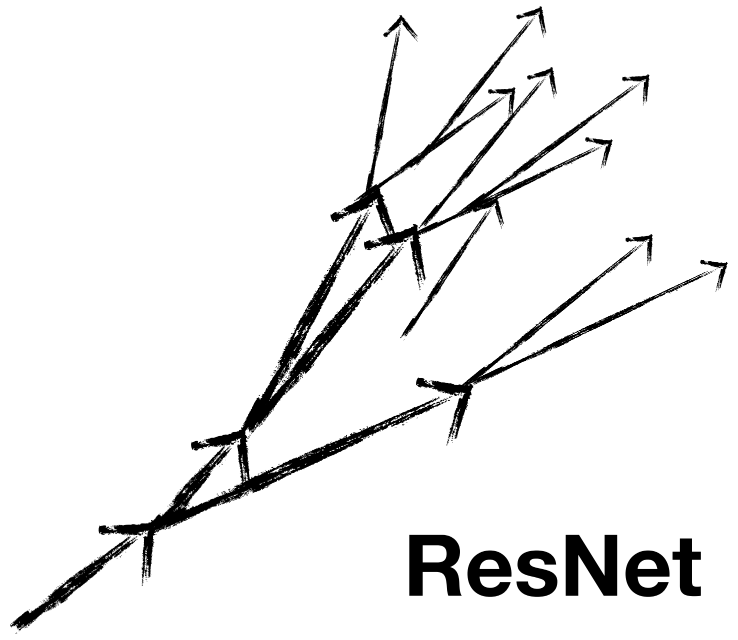
Throughput



FACILE



DeepCalo



ResNet