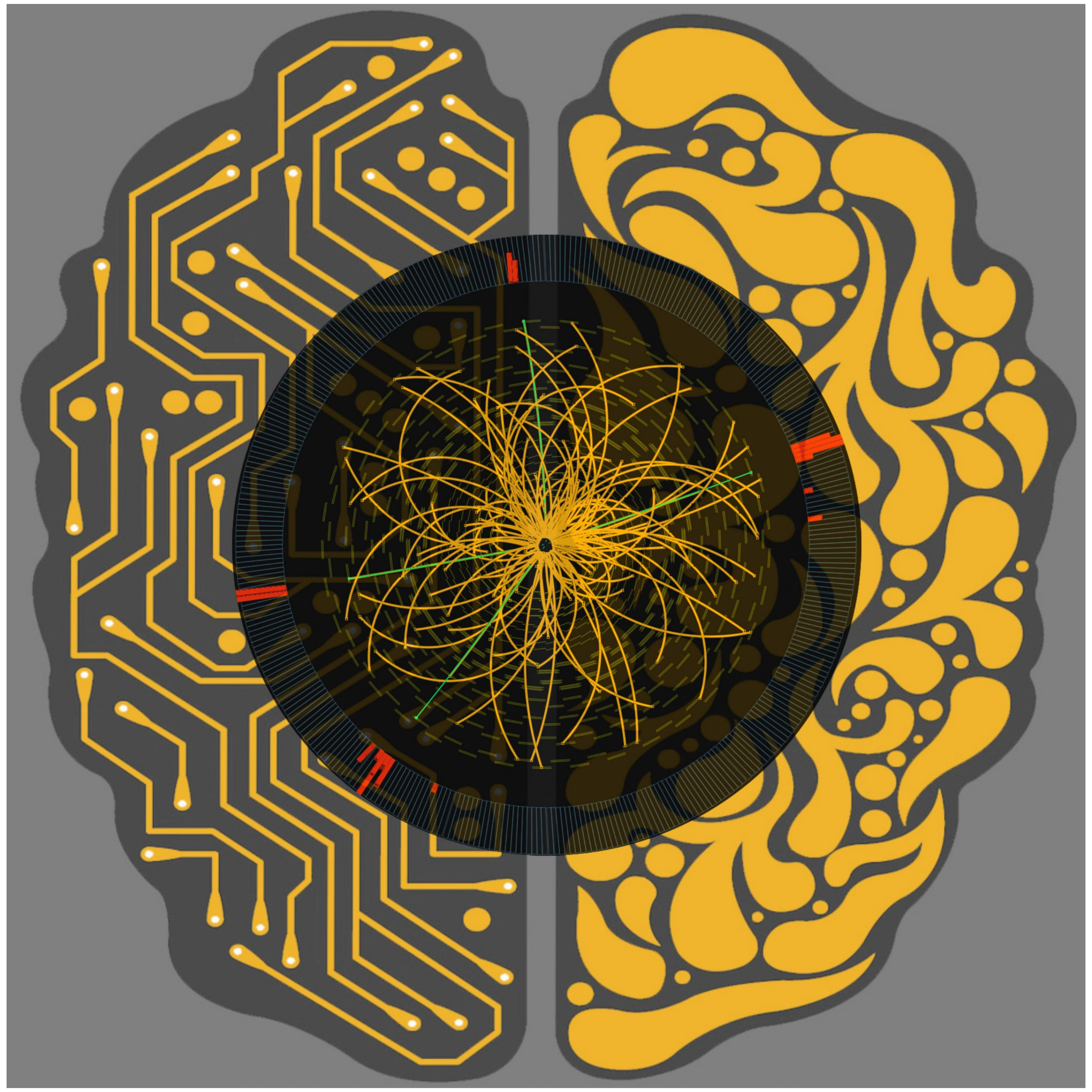
Fermilab

FPGA-accelerated machine learning inference as trigger and computing solutions in particle physics

Mia Liu
FNAL

Fast Machine Learning Workshop

September.11.2019



OUTLINE

Big data challenge in particle physics

an example:

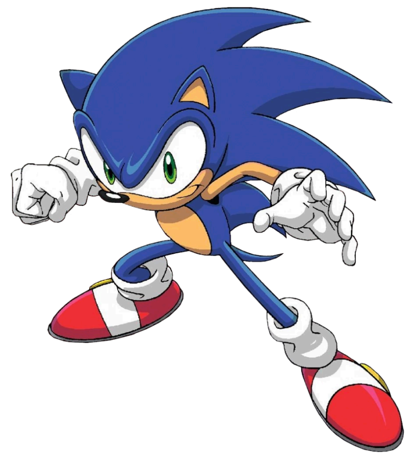
Trigger and computing challenges@LHC: speed, volume, complexity

Potential machine learning solutions

SONIC paper

Proof of concept study with Brainwave:

Heterogenous computing for particle physics

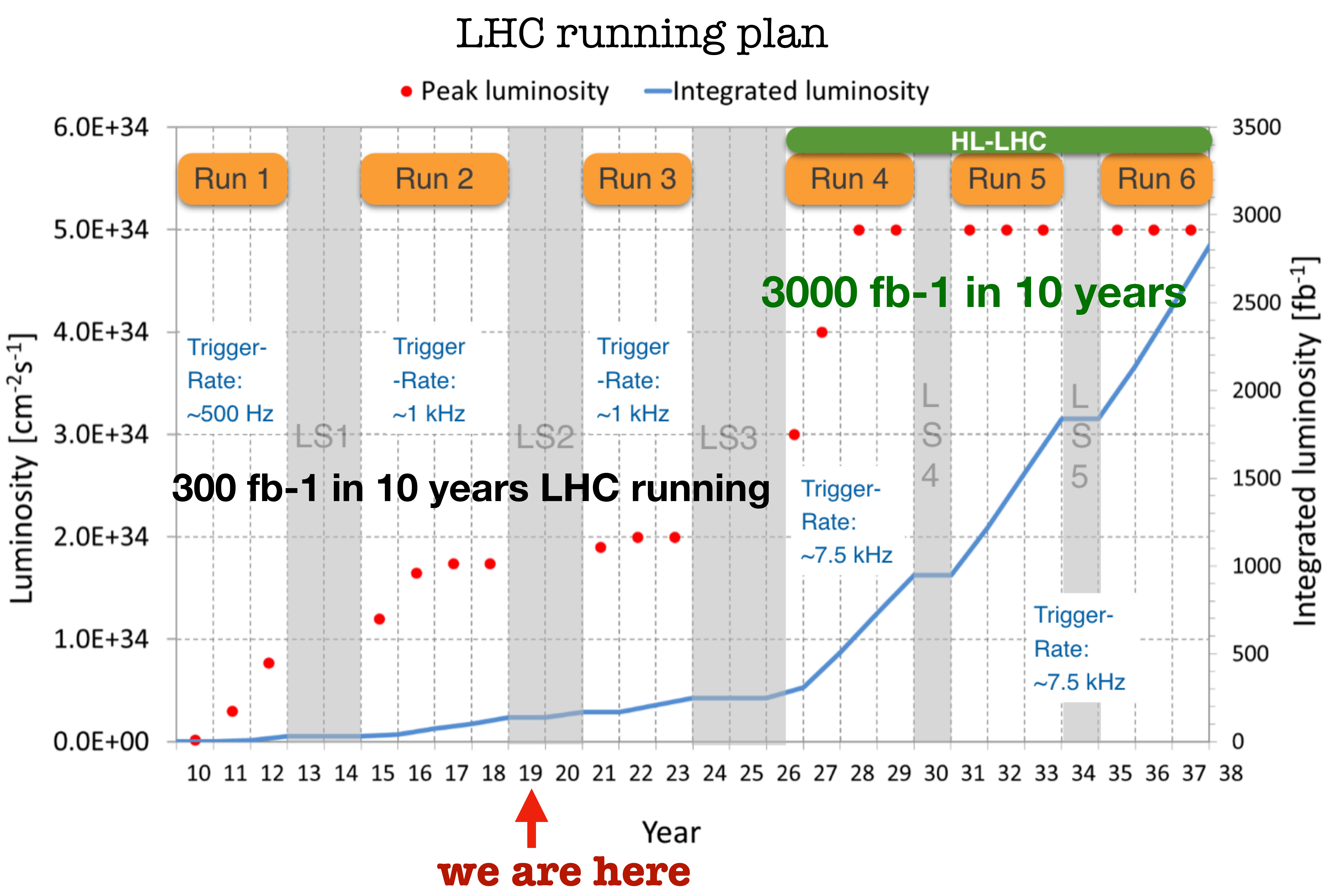


Outlook& takeaways

Massachusetts Institute of Technology

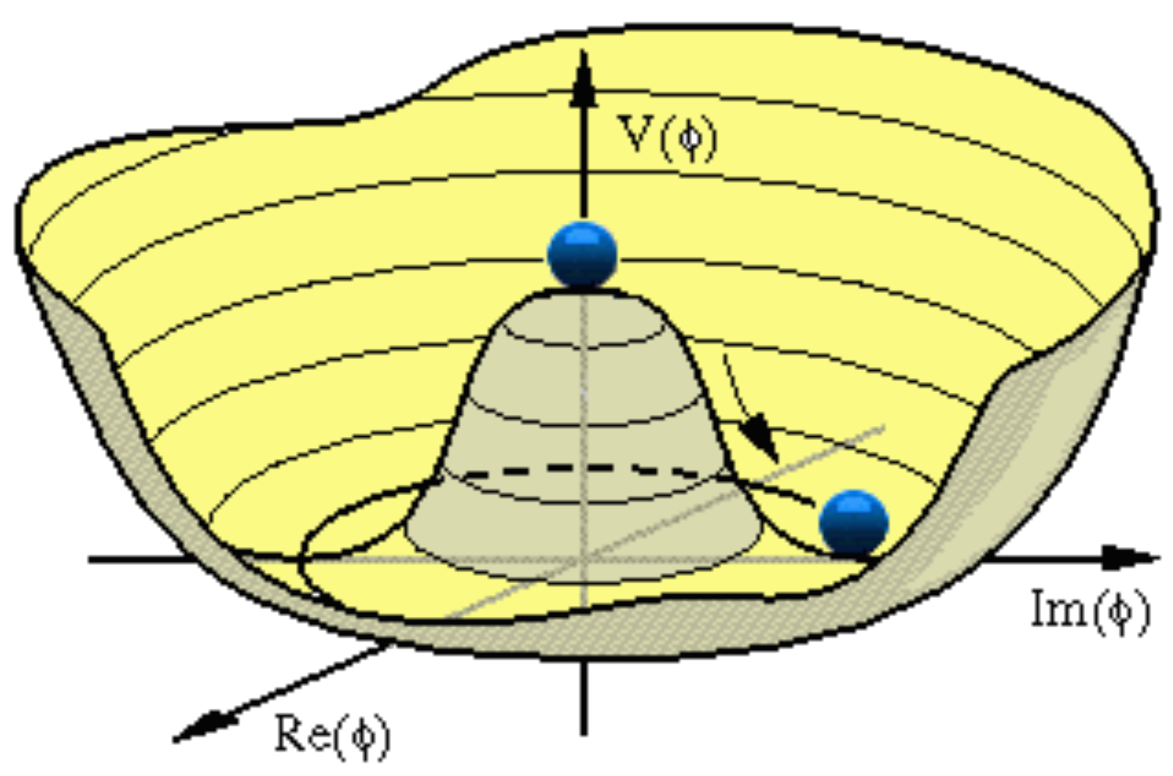
The Large Hadron Collider

95% DATA TO BE COLLECTED

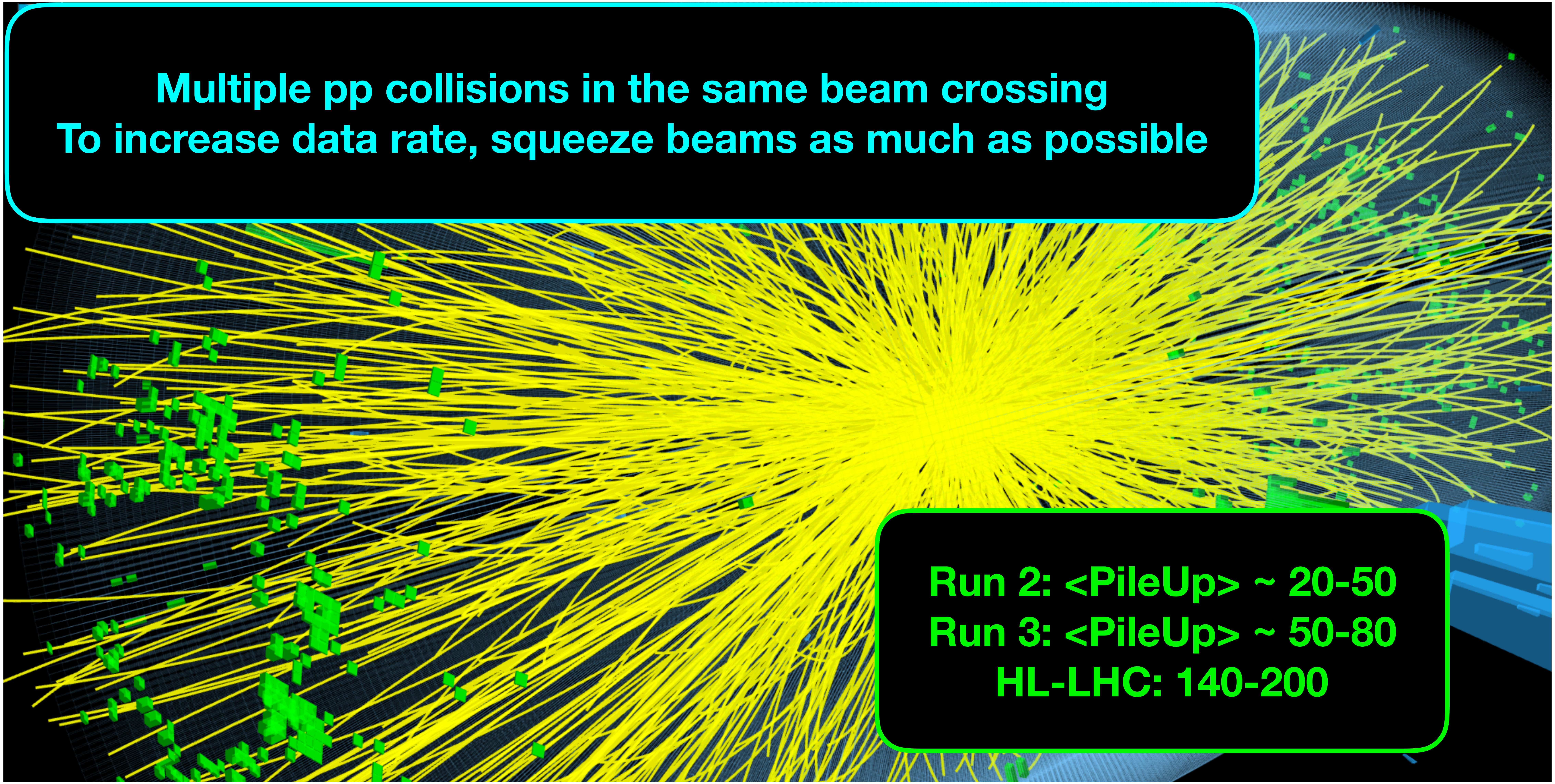


We need the full dataset

example: Study the Higgs boson potential



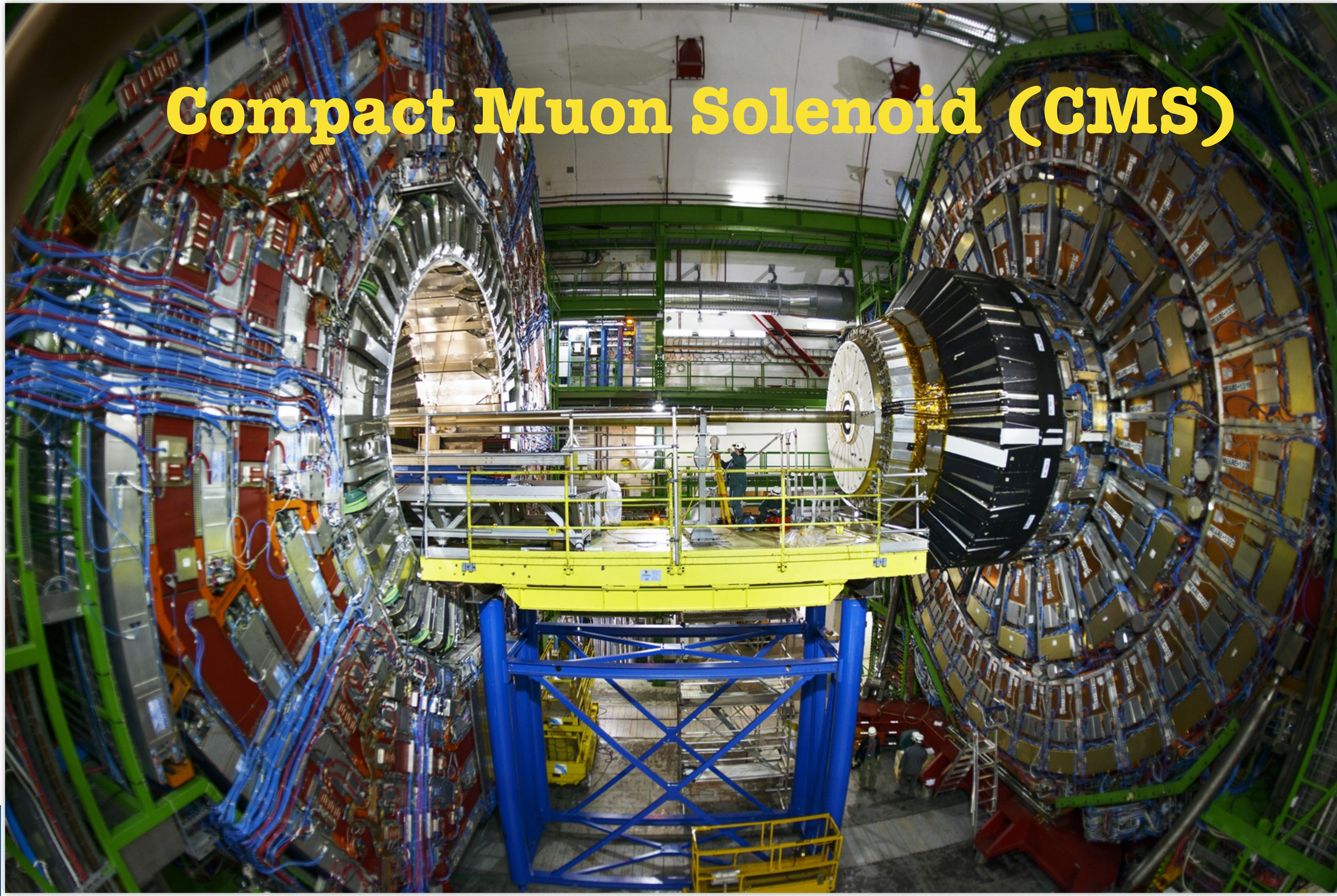
THE 'PILEUP' CHALLENGE



Fermilab

FASTER AND MORE COMPLEX DETECTORS

CMS pixel ('ultra high speed/resolution camera')	Number of channels
'Phase-0'	66 M
'Phase-1'	123 M
'Phase-2'	2B



- Building faster detectors with better resolution
- Trigger& computing challenges@HL-LHC: machine learning solutions?

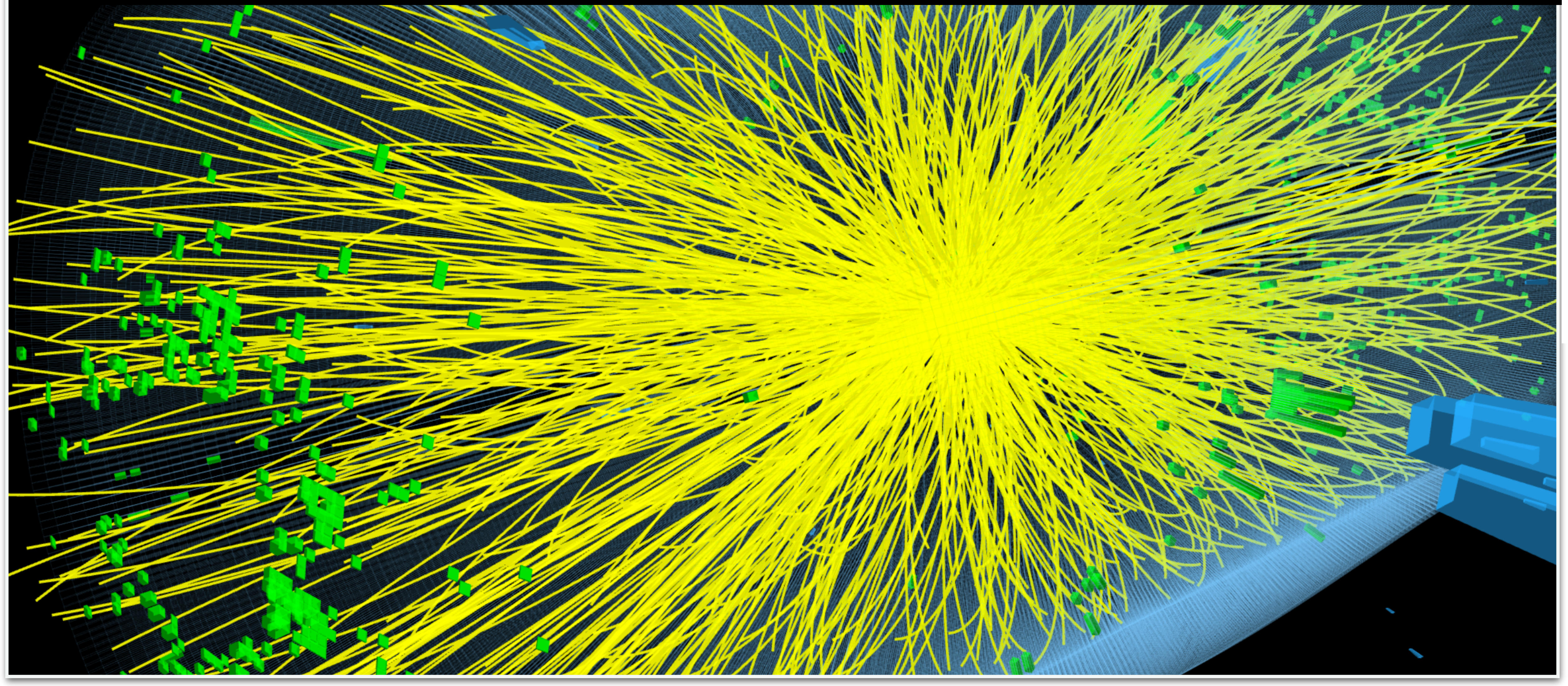
M.LIU

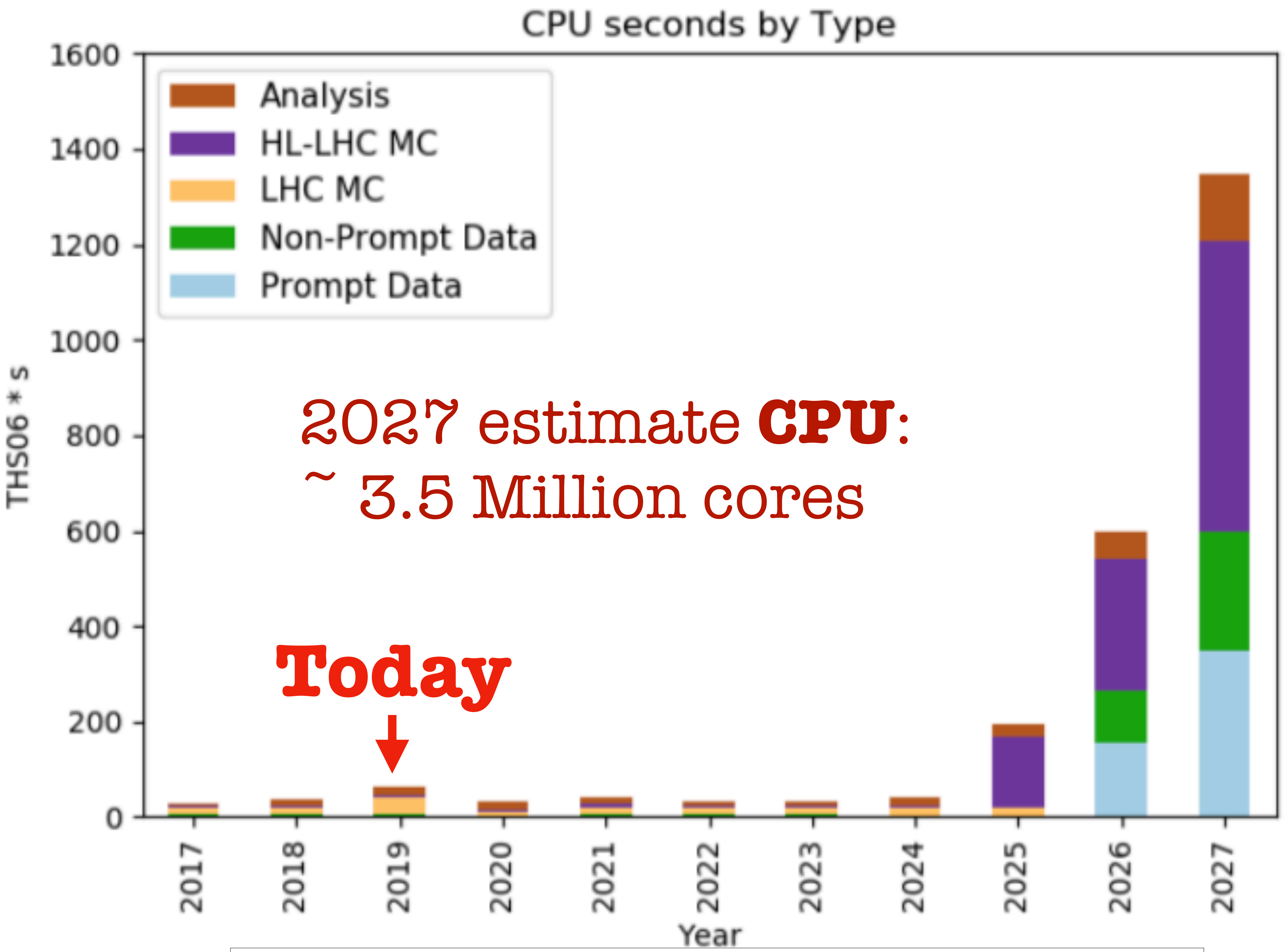
INCREASED DATA VOLUME AND COMPLEXITY

10X more data at the High-luminosity LHC

Event complexity

Current: ~5 minutes per HL-LHC event



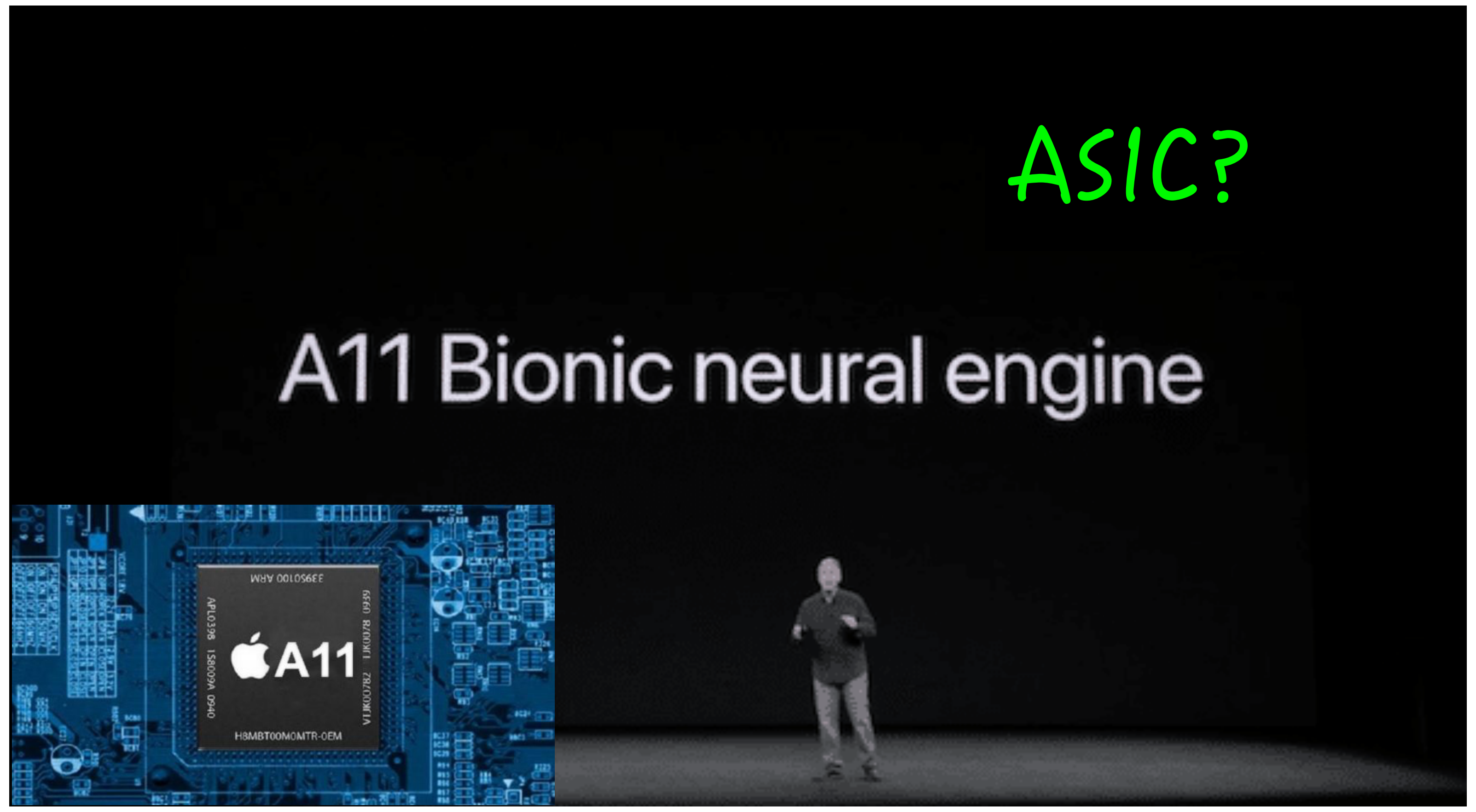


2027 estimate **CPU**: ~ 3.5 Million cores

Today

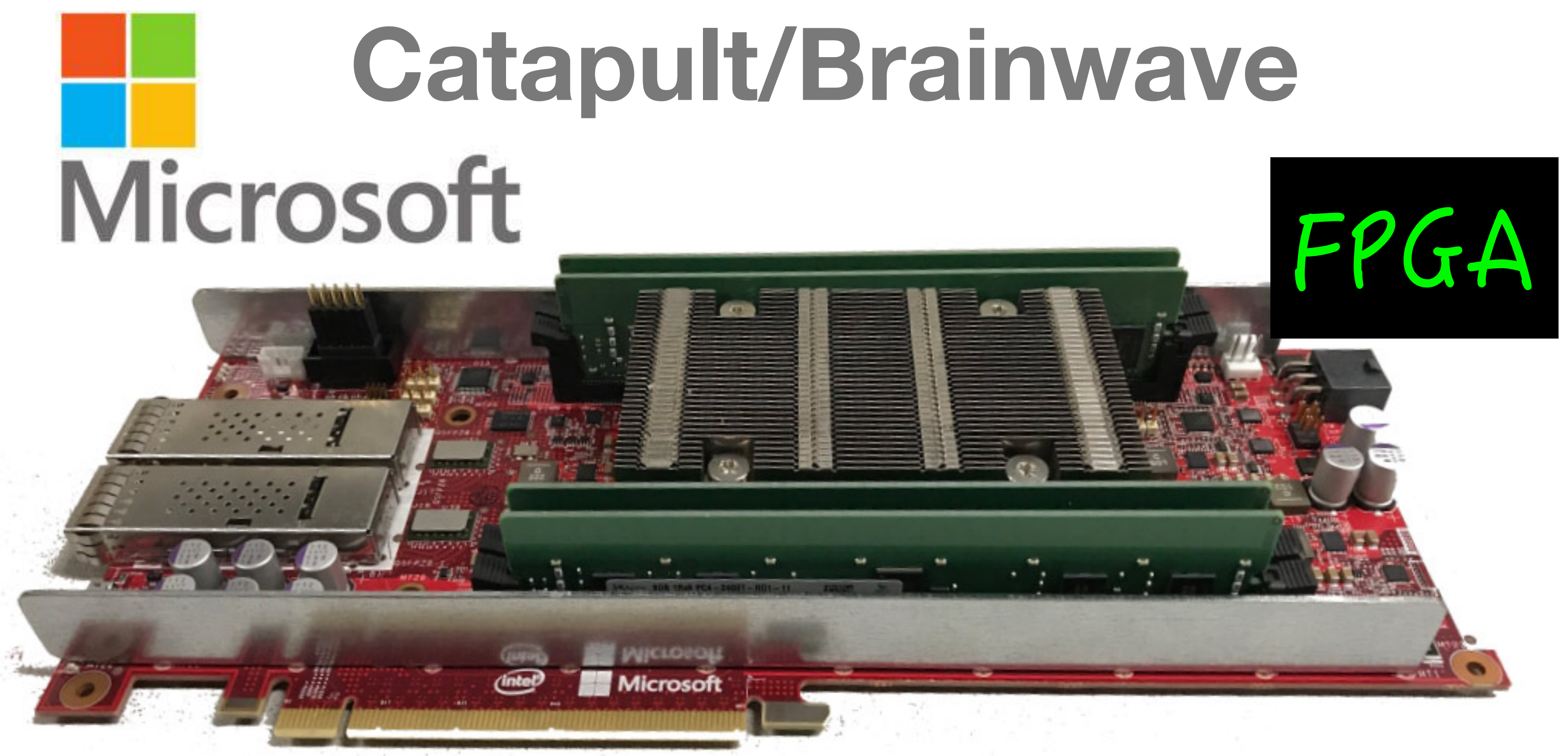
Problem won't get solved by itself: Moore's Law continues ...but Dennard Scaling fails

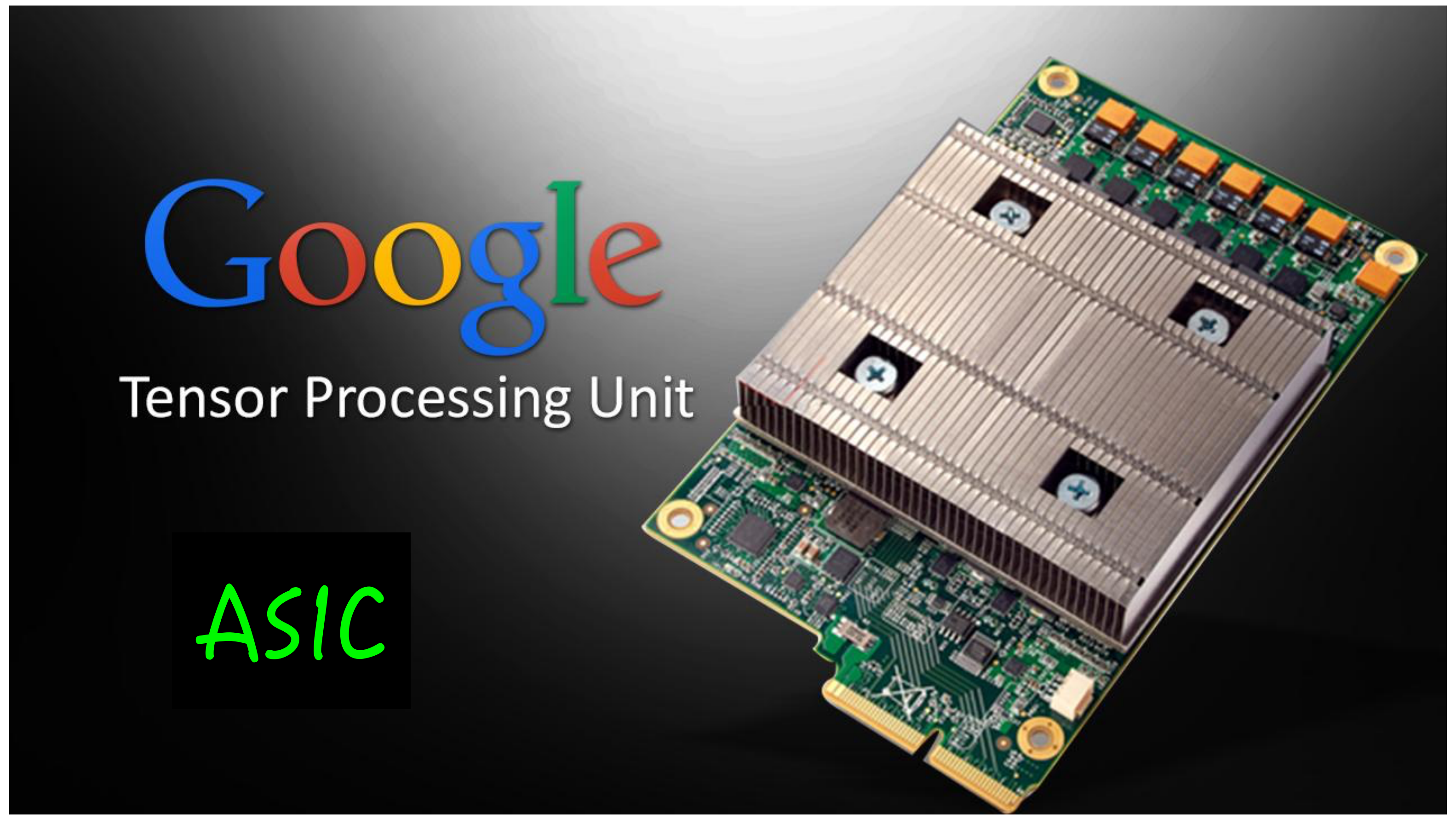
TRENDING: CO-PROCESSORS

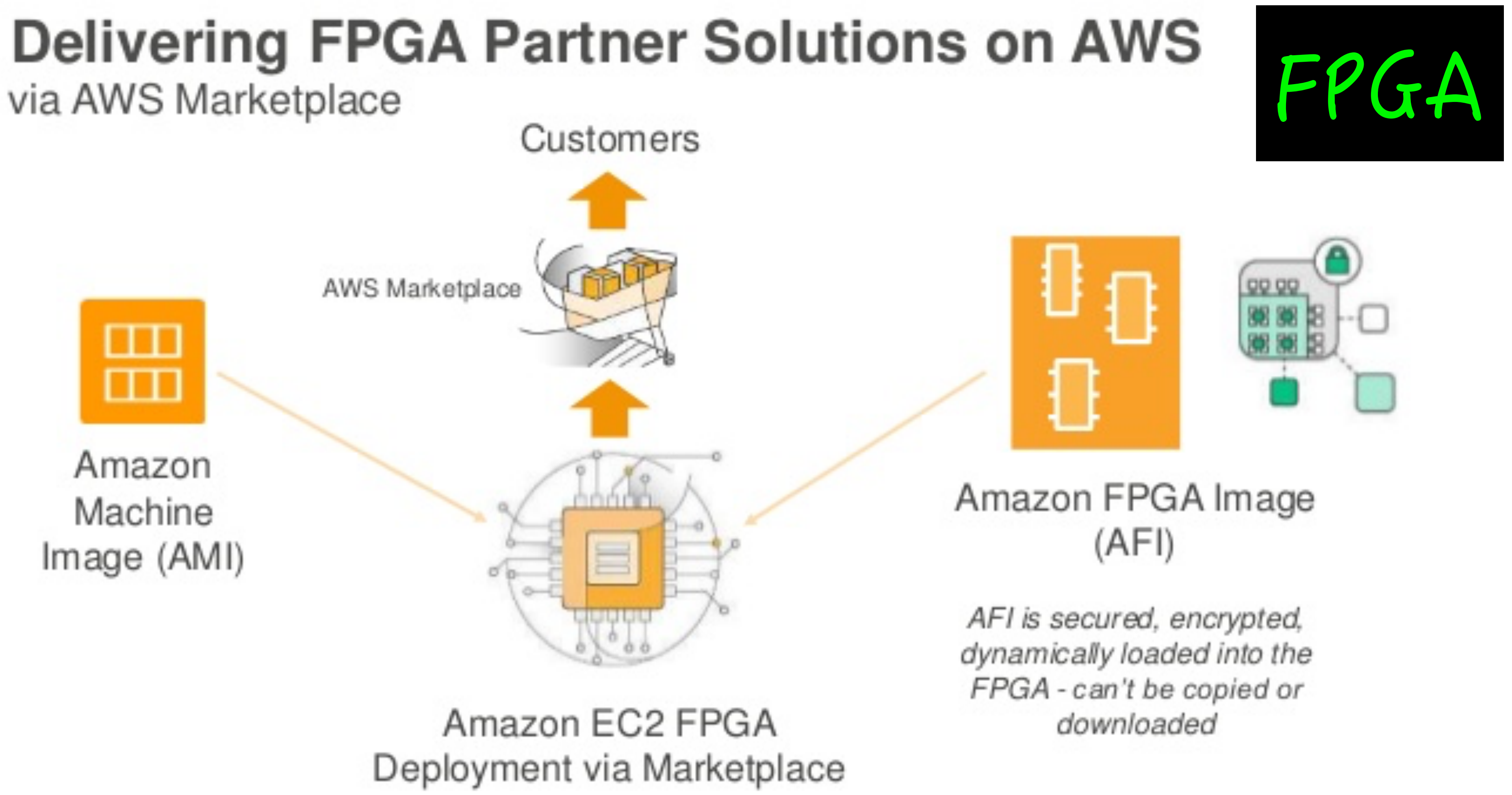


Specialized co-processor hardware for machine learning:

flexibility vs speed (efficiency)







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

SOLUTIONS TO PARTICLE PHYSICS PROBLEMS

Option 1

re-write physics algorithms for new hardware

Language: OpenCL, OpenMP,TBB, HLS, …?

Hardware: FPGA, GPU

Option 2

re-cast physics problem as a machine learning problem

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

Hardware: FPGA, GPU, ASIC

e.g. track reconstruction

Option 1: Parallelized and Vectorized Tracking Using Kalman Filters

Option 2: Tracking using CNN and Graph Networks: e.g. HEP.TrkX project.

Advantages:

- Algorithms expressed as matrix multiplications: intrinsically parallelizable
- Take advantage of co-processors optimized for ML fast inference

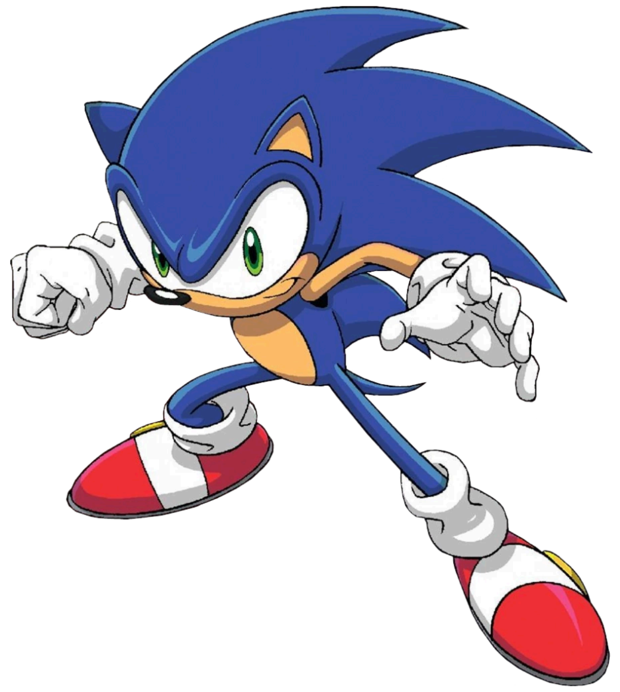
Caveat: challenges of solving our reconstruction problems with NNs.

PROOF OF CONCEPT: SONIC

Services for **O**ptimized **N**etwork **I**nference on **C**o-processors

FPGA-accelerated machine learning inference as a service for particle physics computing

Javier Duarte · Philip Harris · Scott Hauck · Burt Holzman · Shih-Chieh Hsu · Sergo Jindariani · Suffian Khan · Benjamin Kreis · Brian Lee · Mia Liu · Vladimir Lončar · Jennifer Ngadiuba · Kevin Pedro · Brandon Perez · Maurizio Pierini · Dylan Rankin · Nhan Tran · Matthew Trahms · Aristeidis Tsaris · Colin Versteeg · Ted W. Way · Dustin Werran · Zhenbin Wu



https://arxiv.org/abs/1904.08986

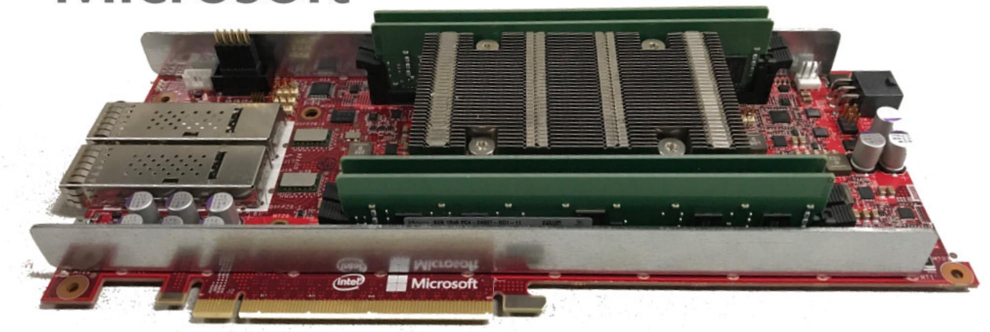
Massachusetts Institute of Technology

Question:
How do we help with physics event data processing model with industry developments in co-processors?

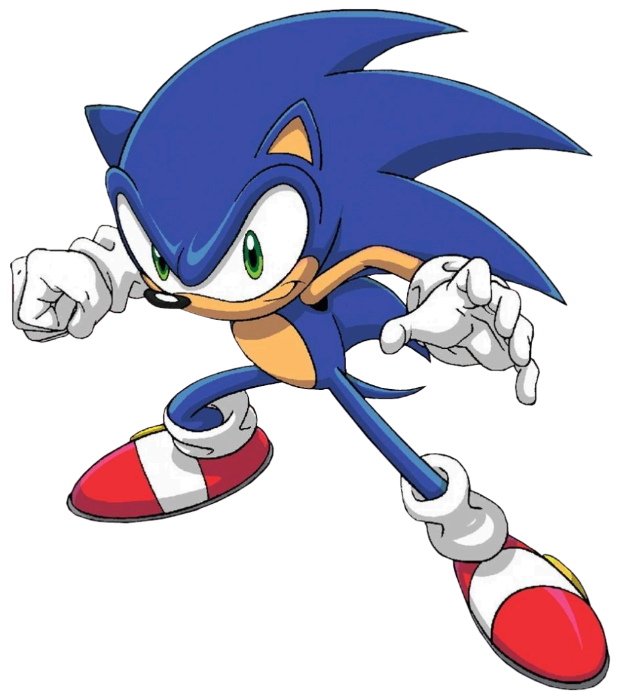
Focus on speeding up the **inference**.

Catapult/Brainwave

Microsoft



SONIC PAPER



CERN

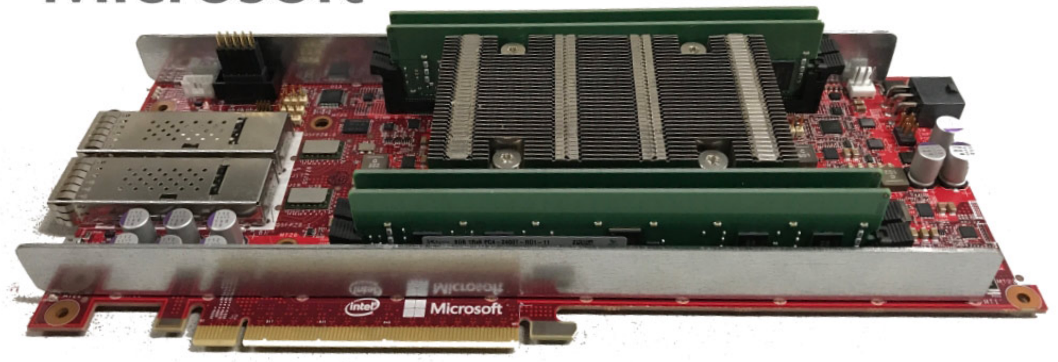
Massachusetts Institute of Technology

UNIVERSITY · OF · WASHINGTON · 1861 ·

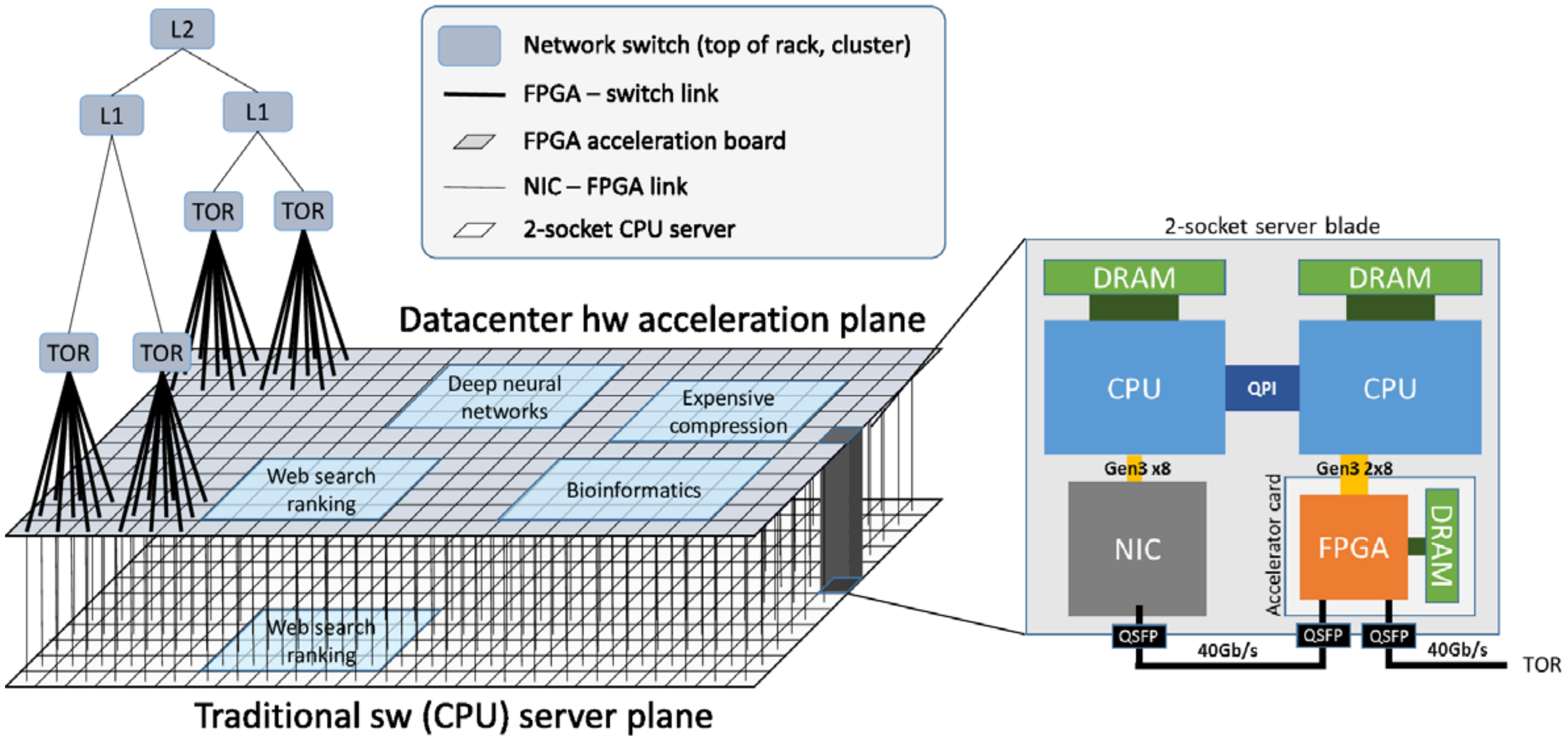
UIC

- Proof of concept: Top tagging/neutrino image classification on Brainwave
- Implementation as service in CMSSW in non-disruptive way.
- Speed and data throughput performance

Microsoft

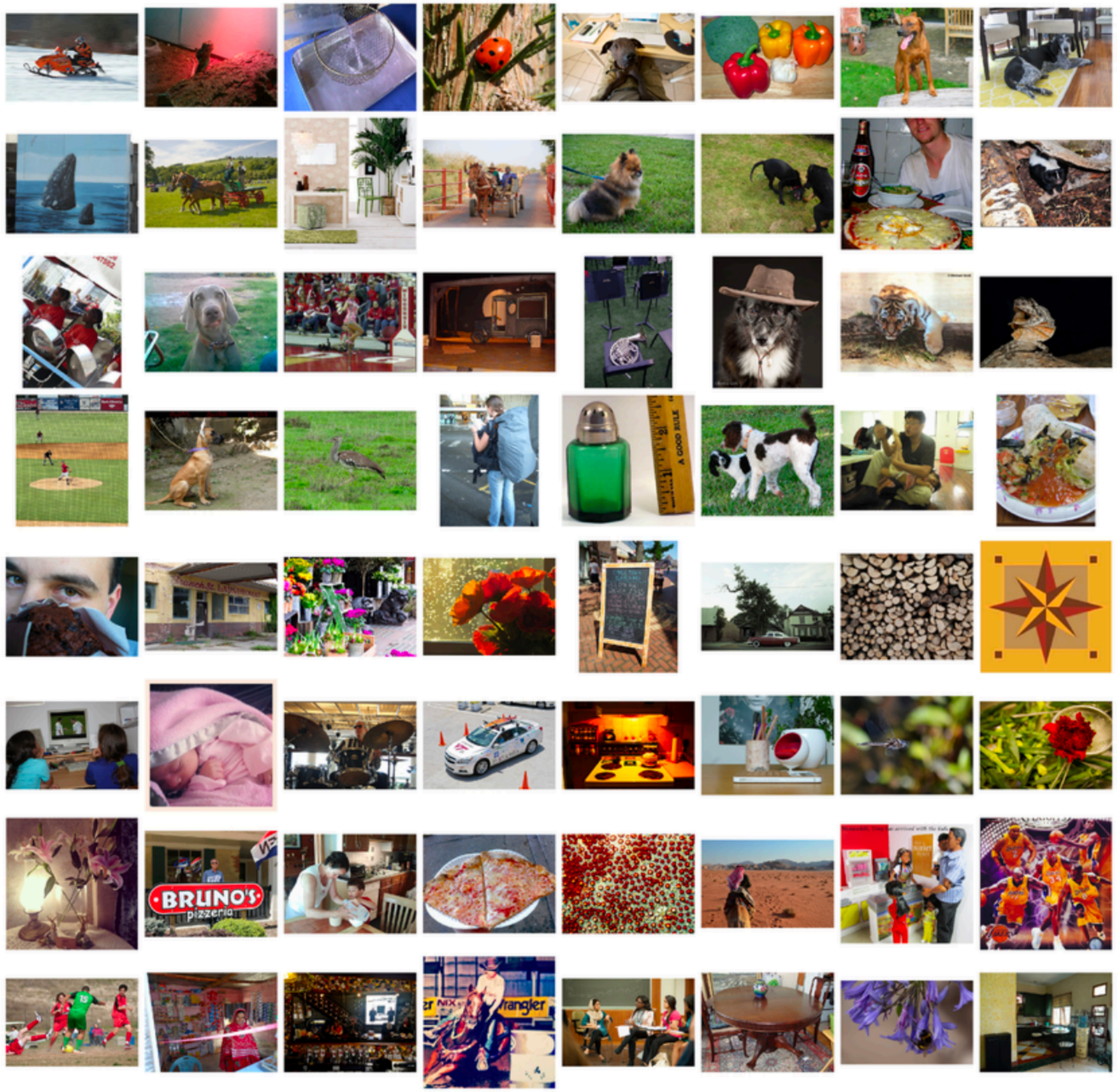


BRAINWAVE

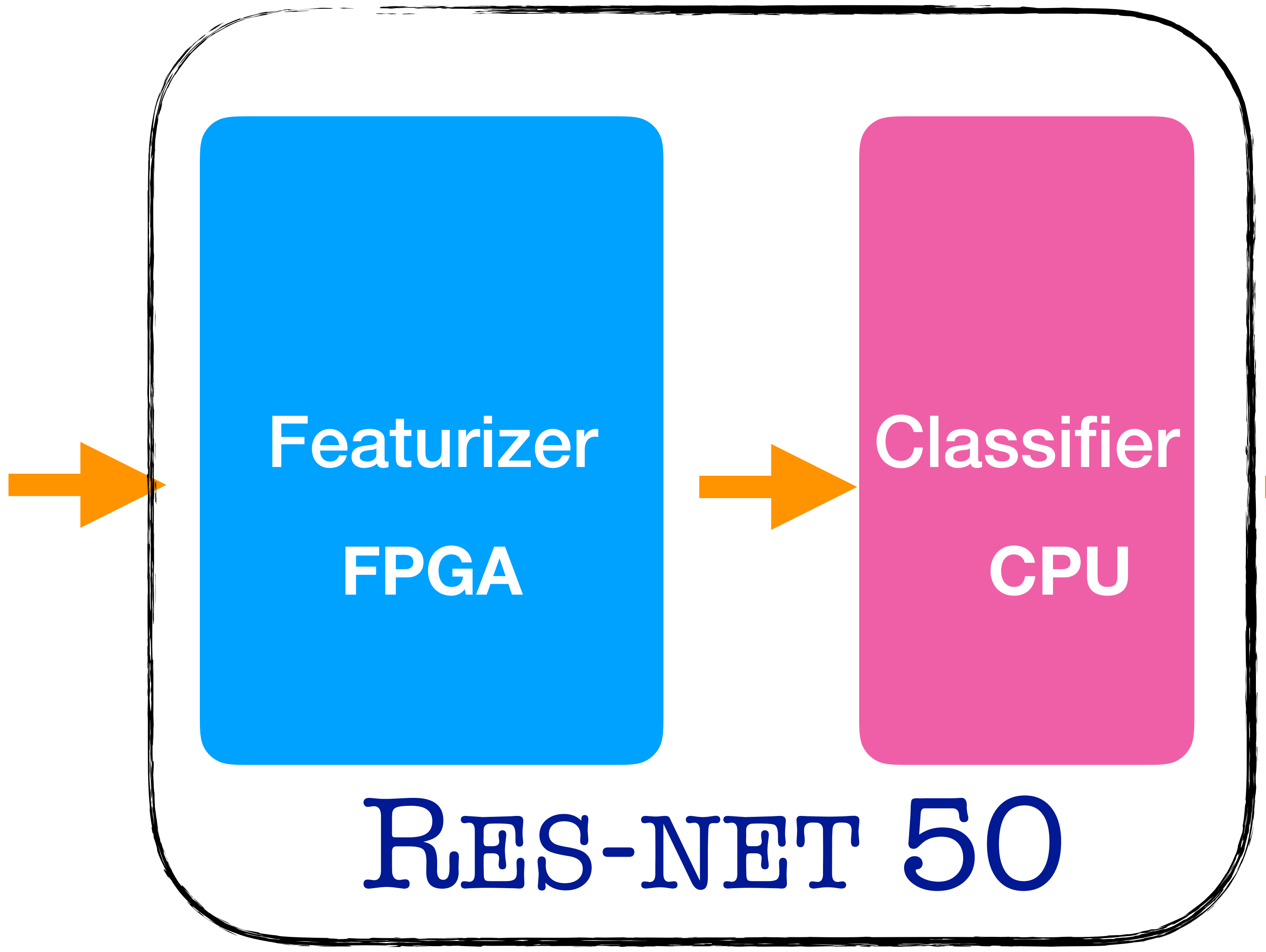


- Provides a full service at scale (more than just a single co-processor)
- Multi-FPGA/CPU fabric accelerates both **computing** and **network**
- Models supported:
  - ResNet50, ResNet152, DenseNet121 ,VGGNet 16
  - Weight retuning available

A PHYSICS CASE: JET TAGGING WITH RES-NET 50



ImageNet

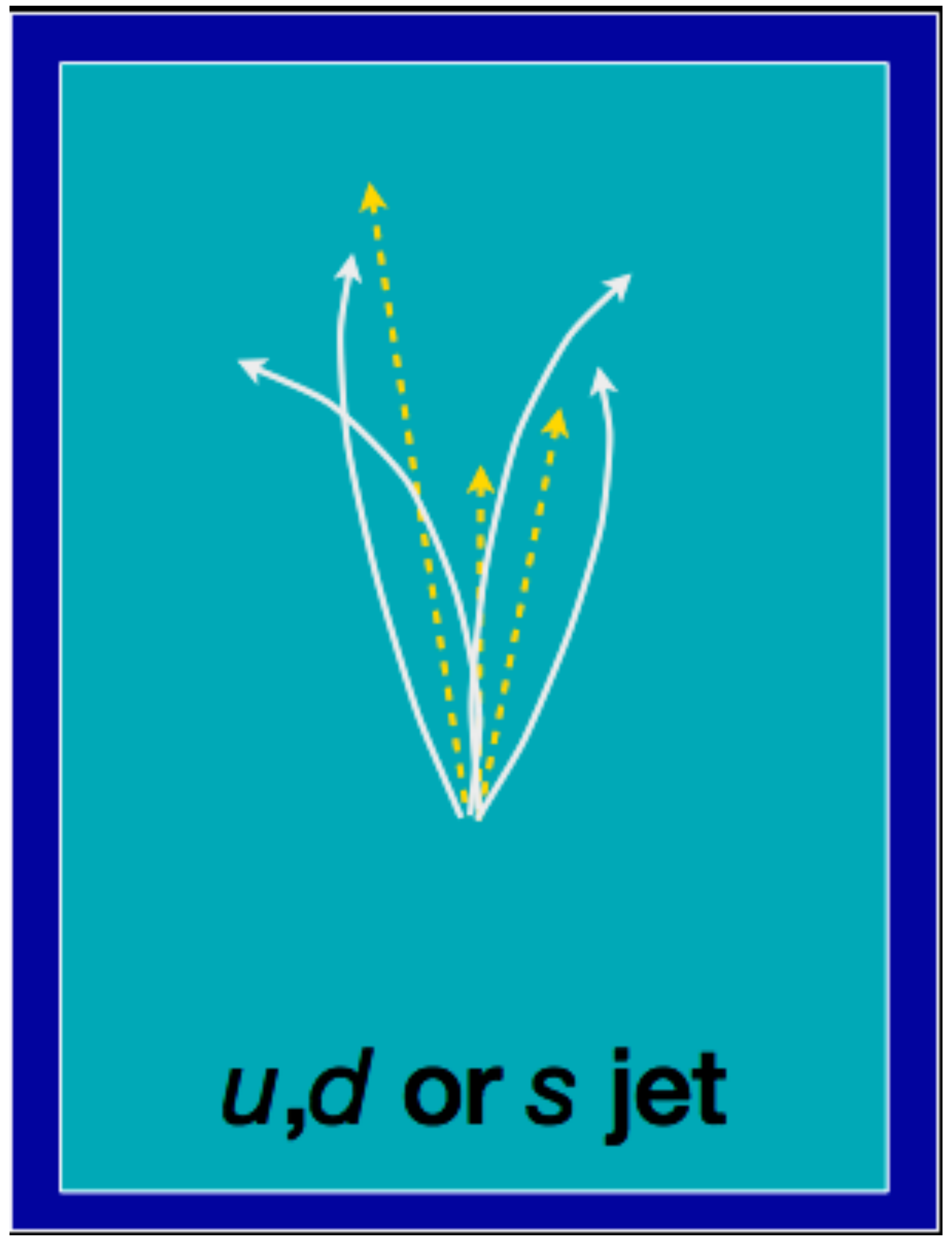


1000 classes
(cats,dogs...)

25M parameters

4 G-ops/inference

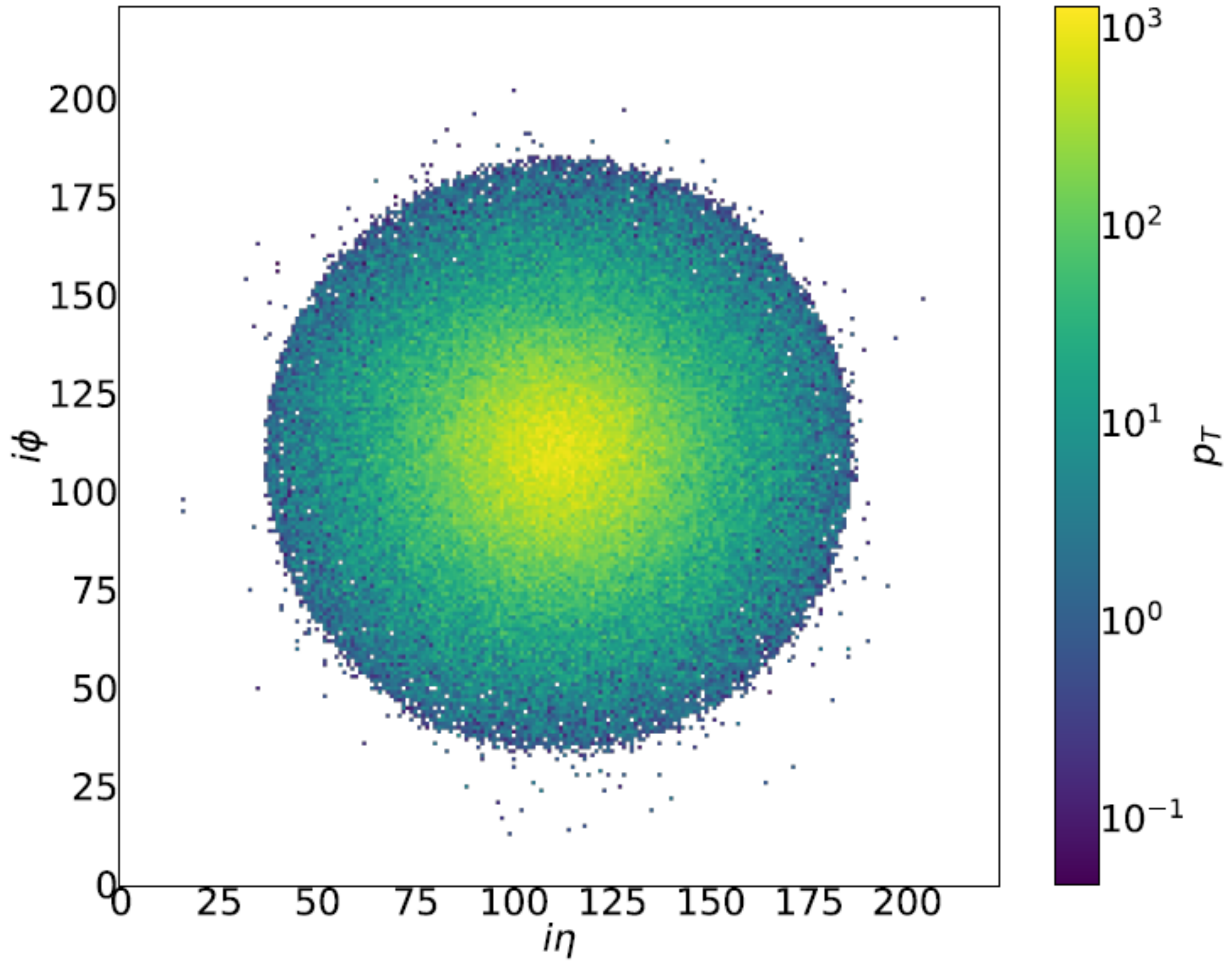
PASSING JET IMAGES TO RE-TRAIN RES-NET 50



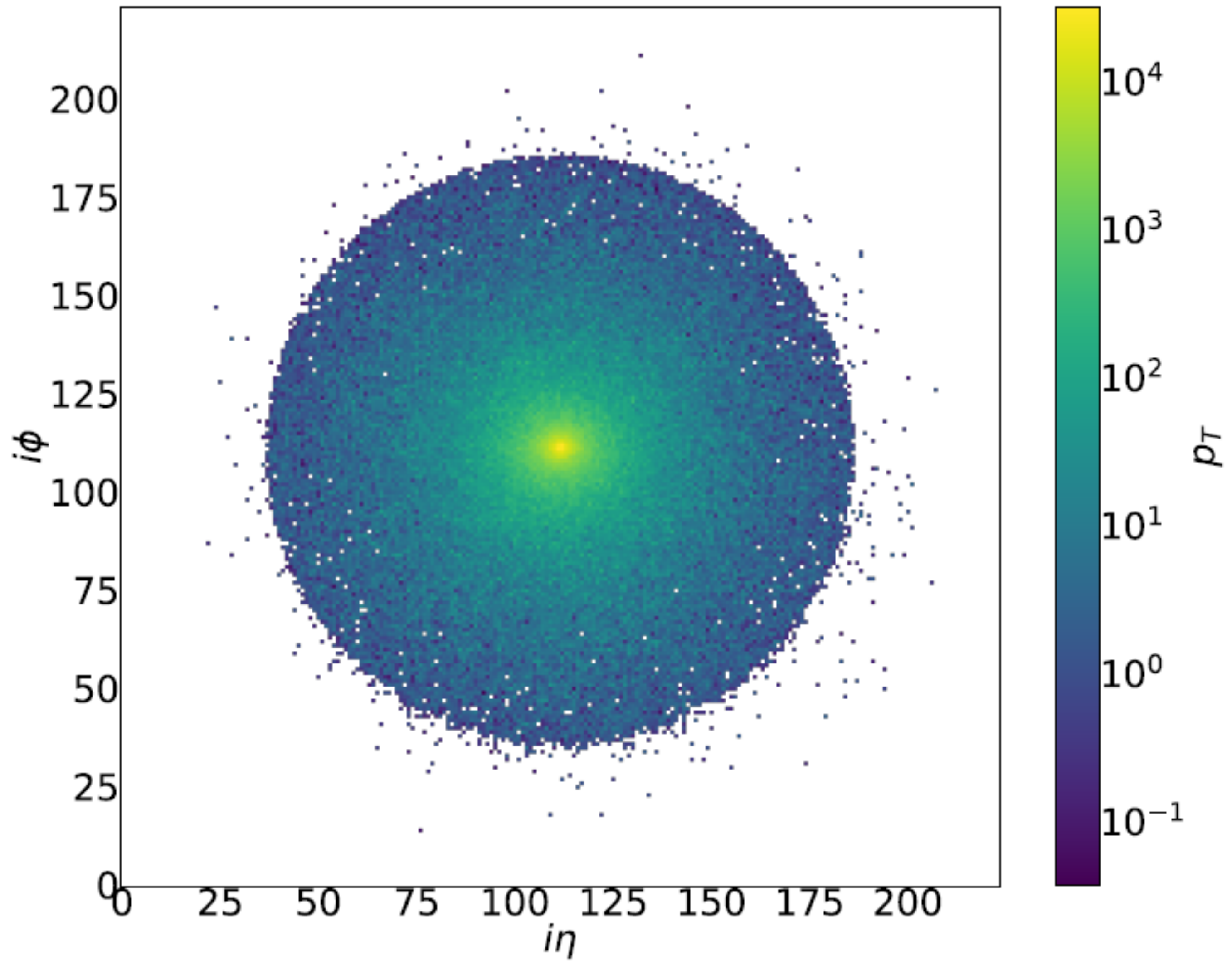
Top tagging benchmark dataset

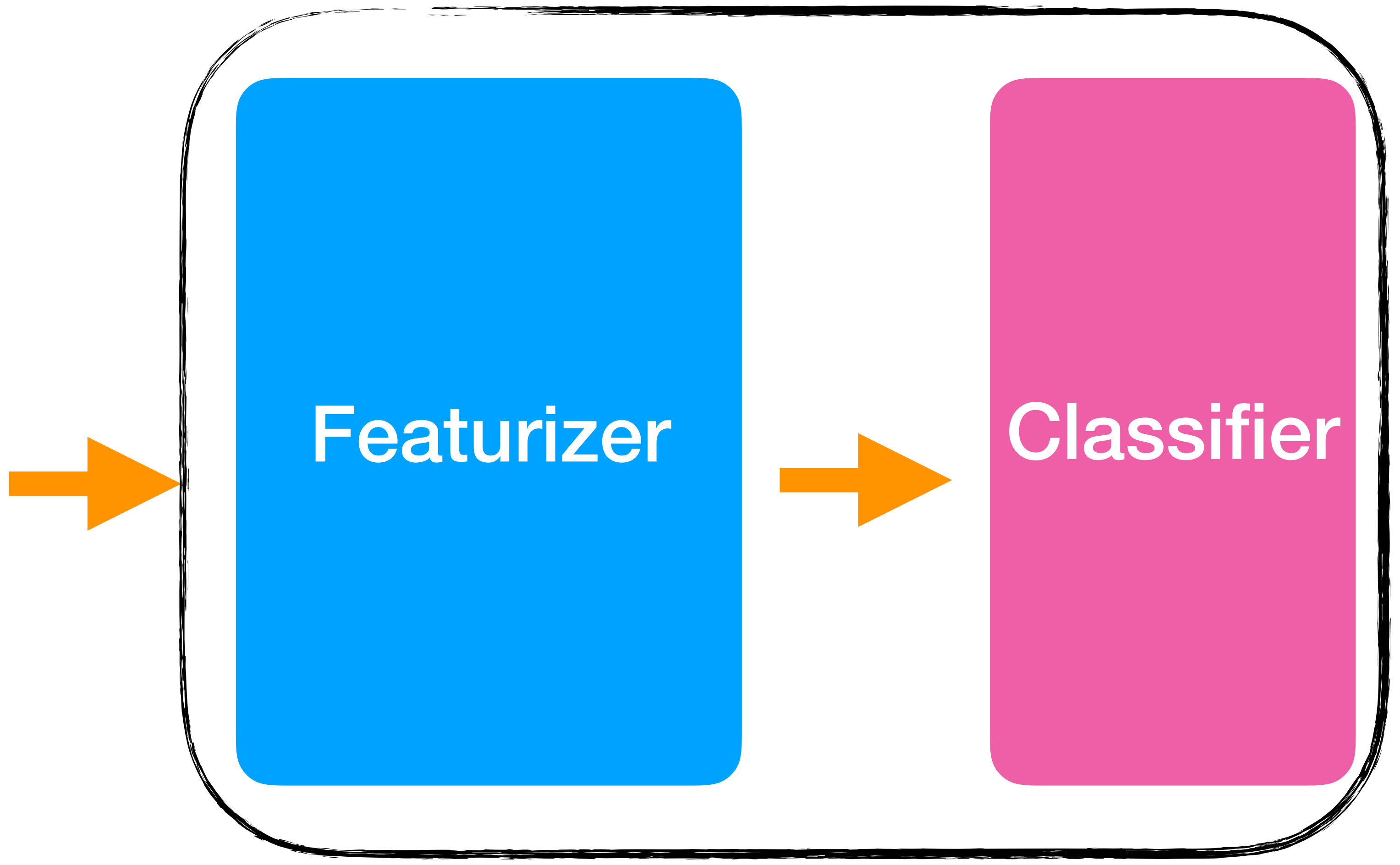
- Images made from density map of the pt of jet constituents in η*φ space.
  - Grey image, duplicated to RGB.

top, averaged over 5k jets



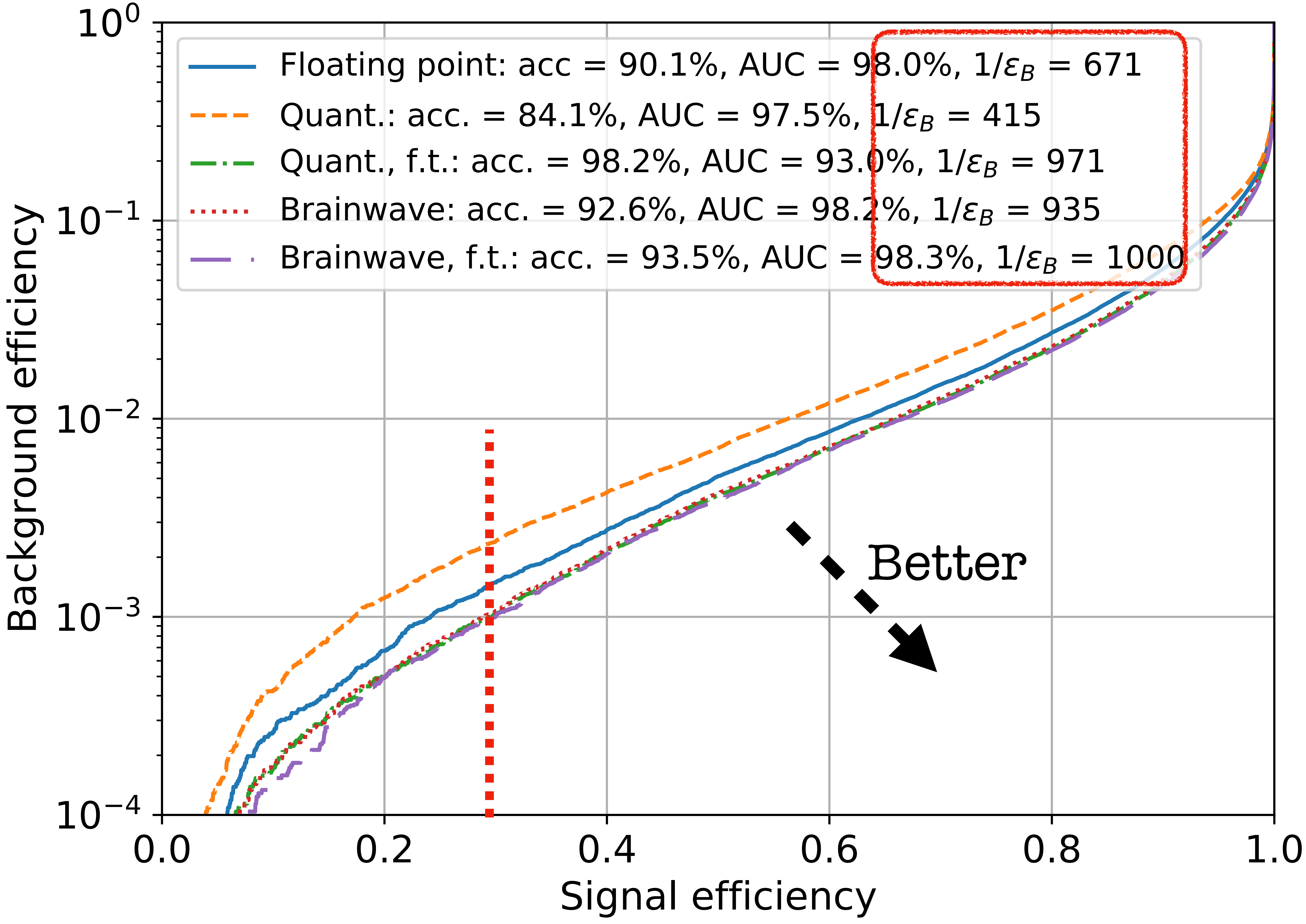
QCD, averaged over 5k jets





RETRAIN RES-NET 50

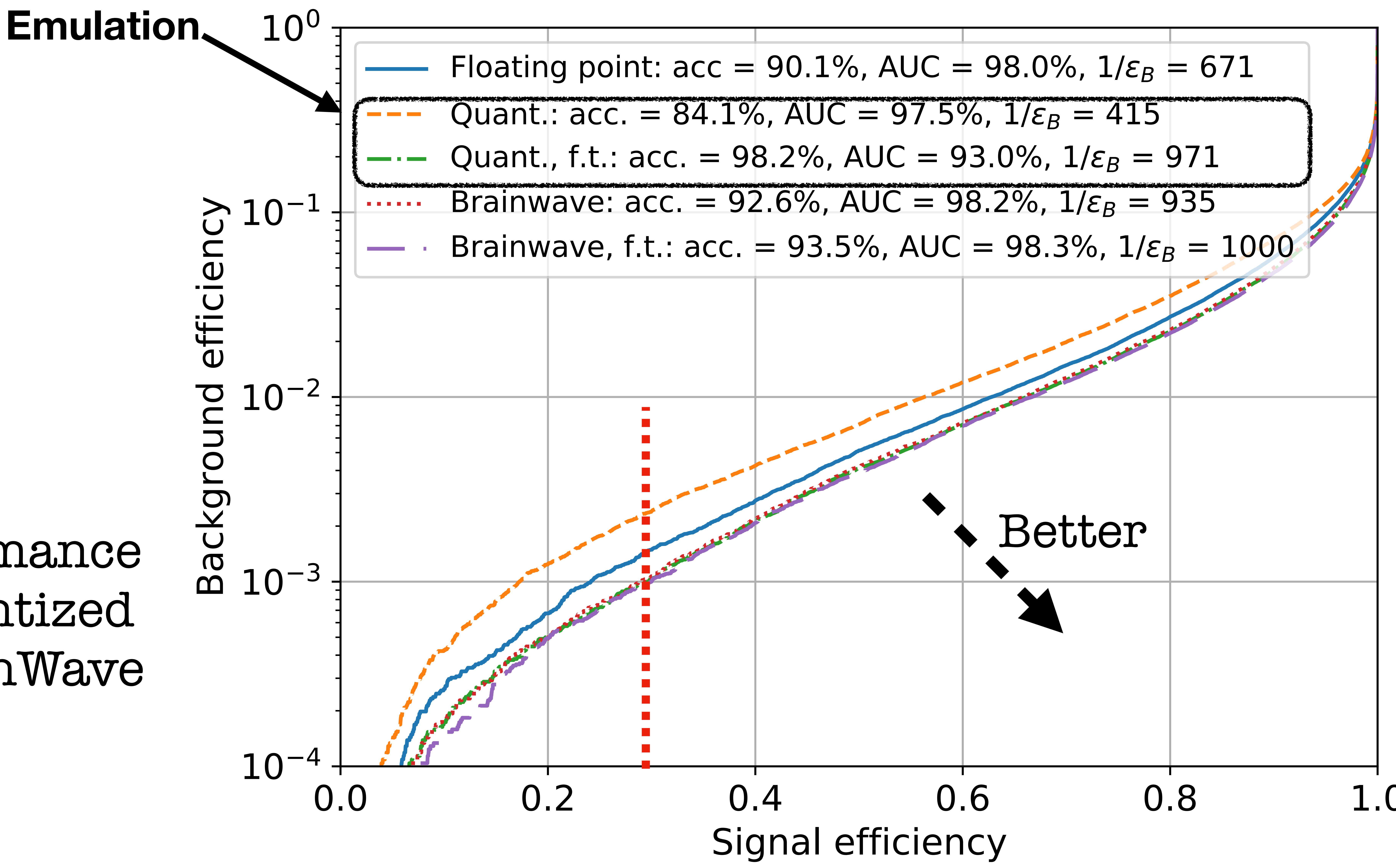
RE-TRAIN RESNET-50 FOR TOP TAGGING



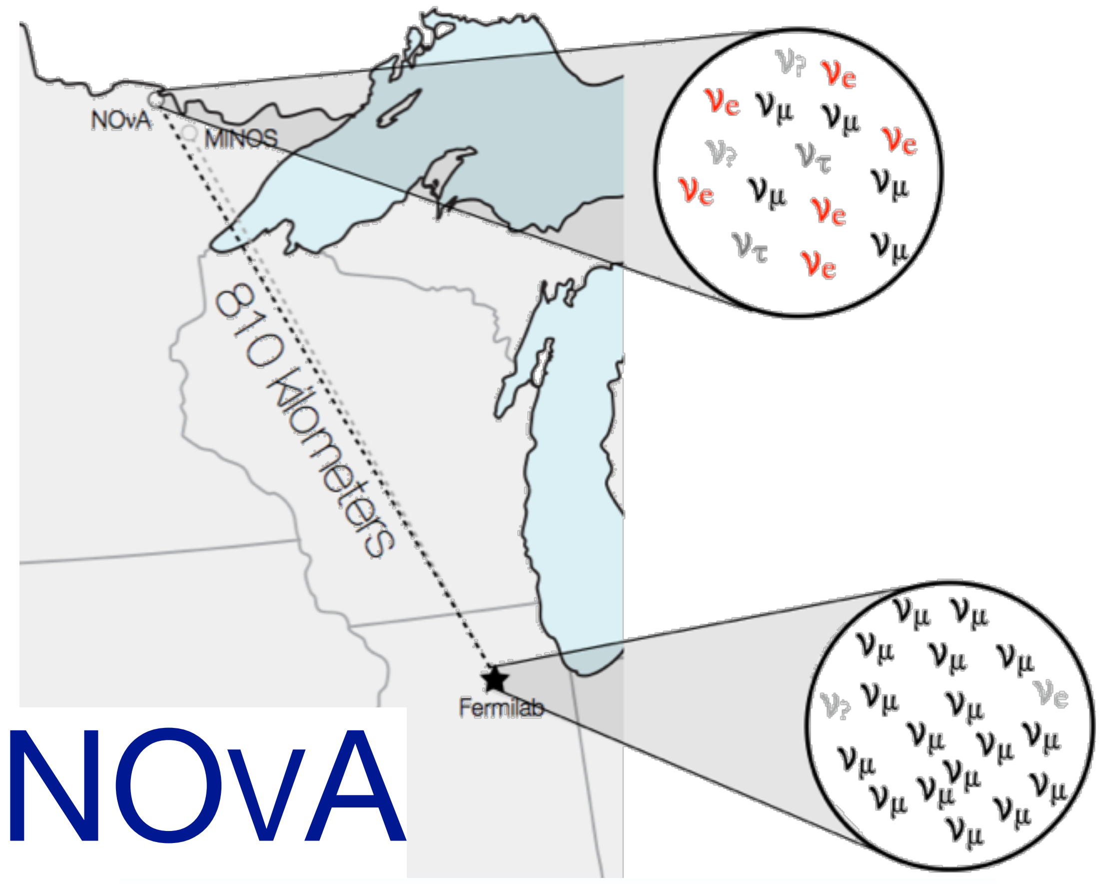
RE-TRAIN RESNET-50 FOR TOP TAGGING

Quantized model:

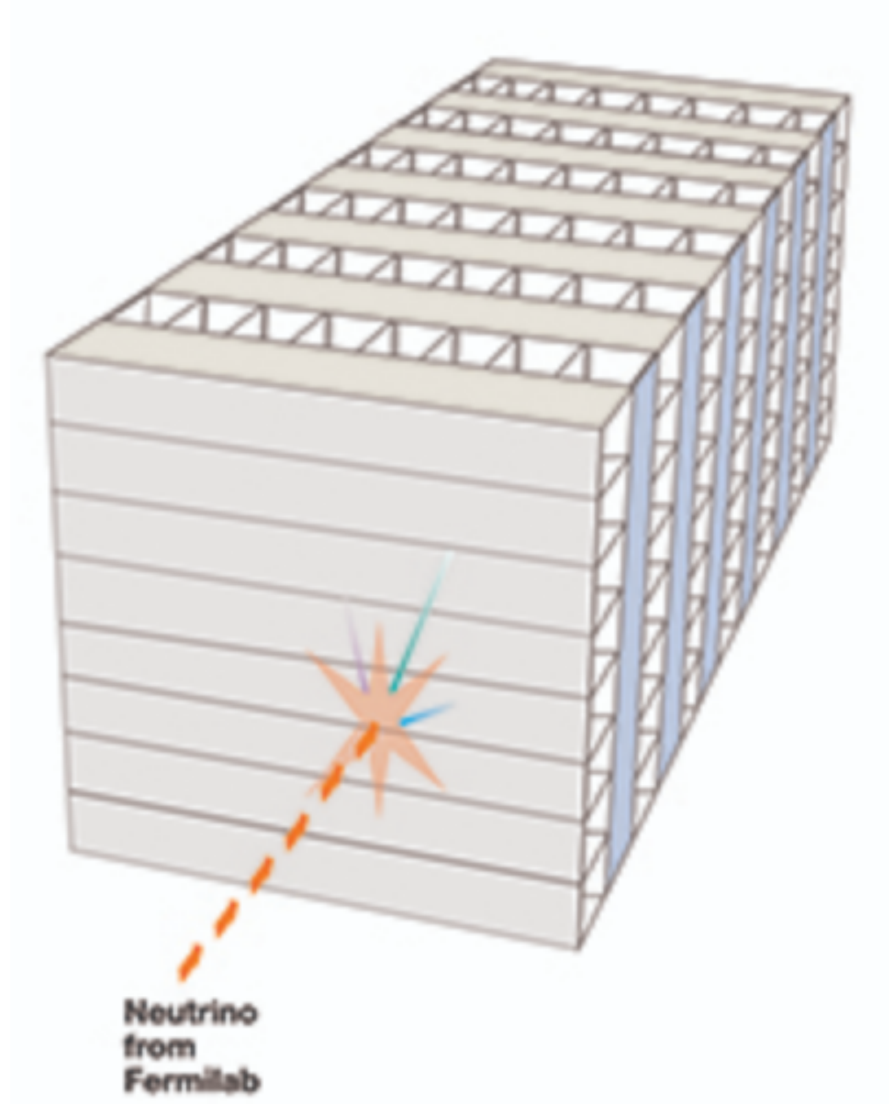
- Brainwave's implementation of ResNet50 on FPGA
- Can tune weights
- State of art performance achieved with quantized ResNet 50 on BrainWave service

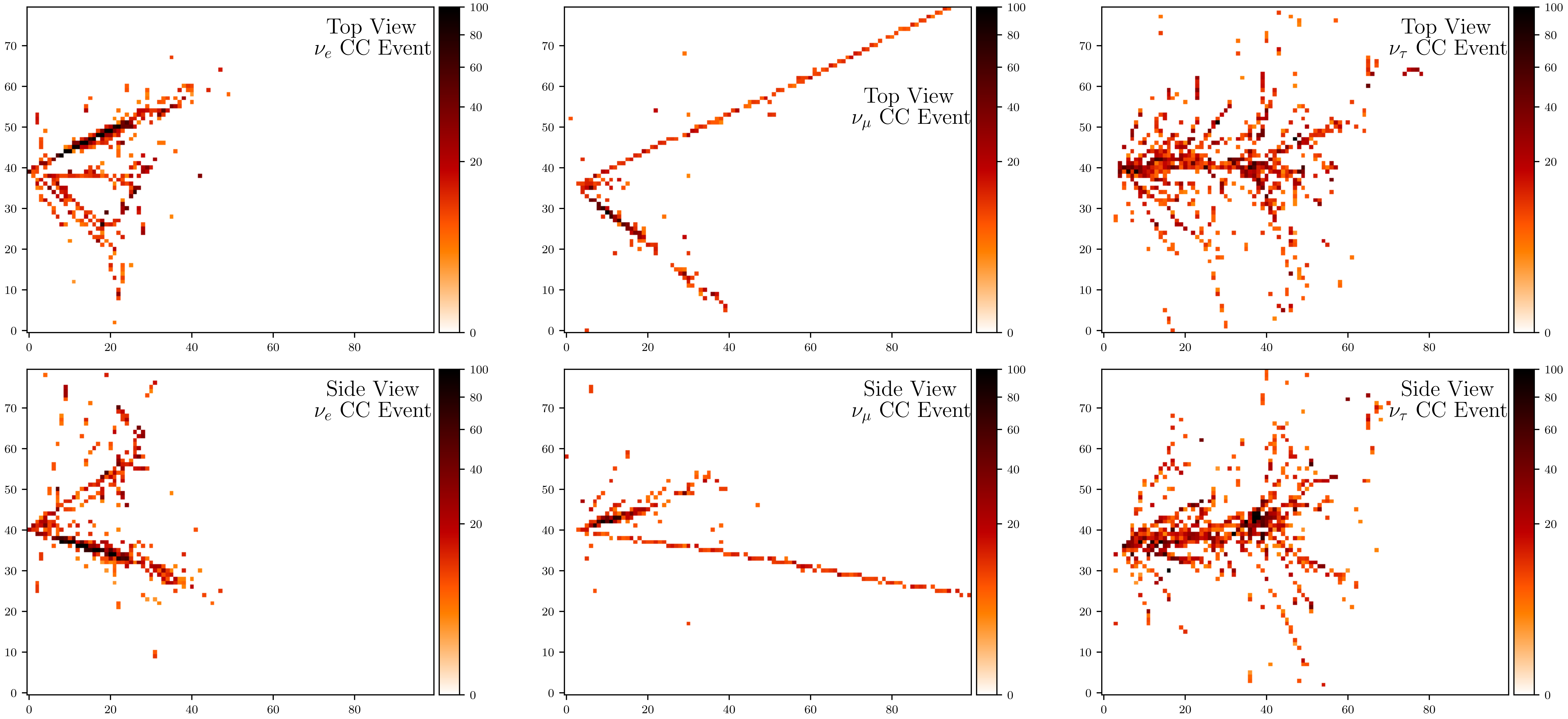


CLASSIFY NEUTRINO IMAGES



NOvA

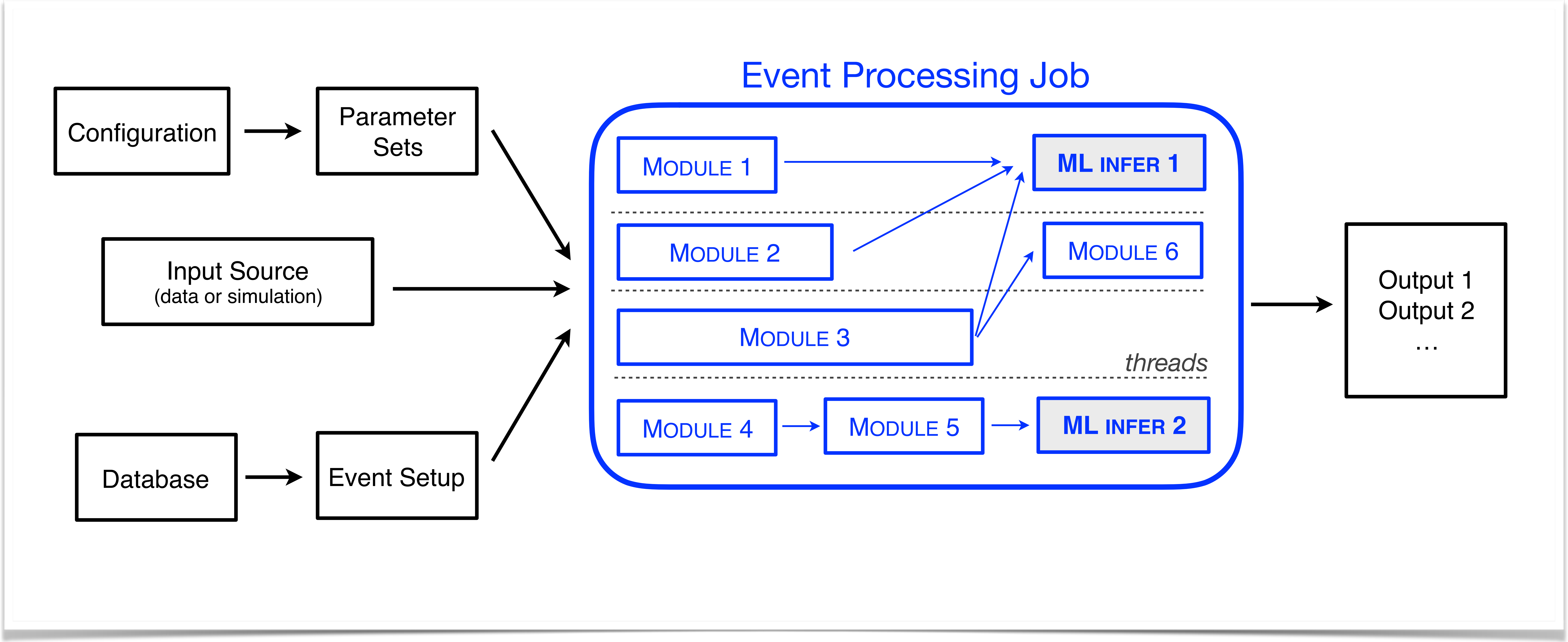


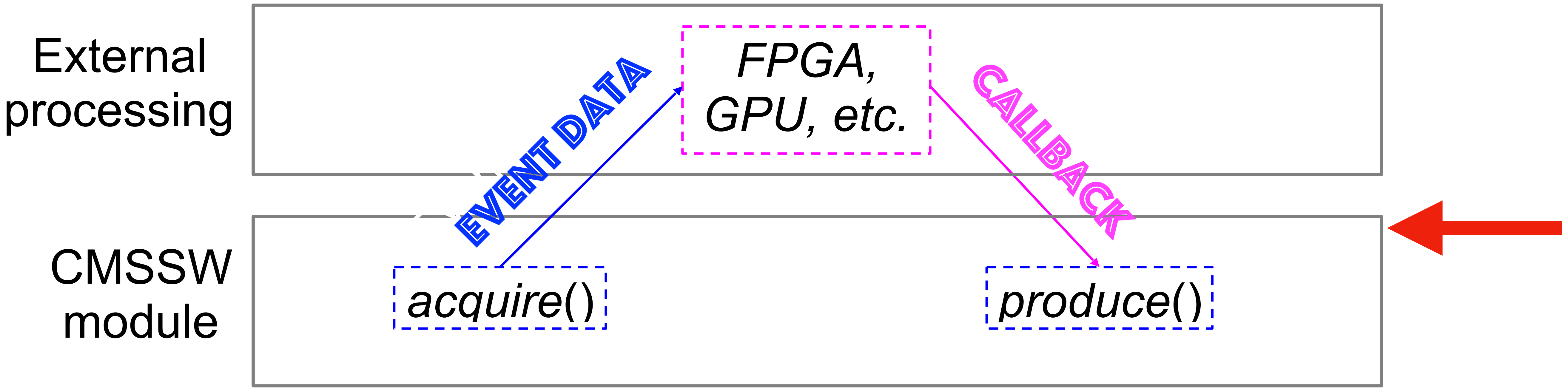


Electron neutrino | Muon neutrino | Tau neutrino

- Primary goal of NOvA: measurement of neutrino oscillations via $\nu\mu \rightarrow \nu e$: Classifying neutrinos with ResNet50 (transfer learning).
- Can be used in NOvA event processing as of today: see Thomas's lightening round talk.

CMS COMPUTING MODEL

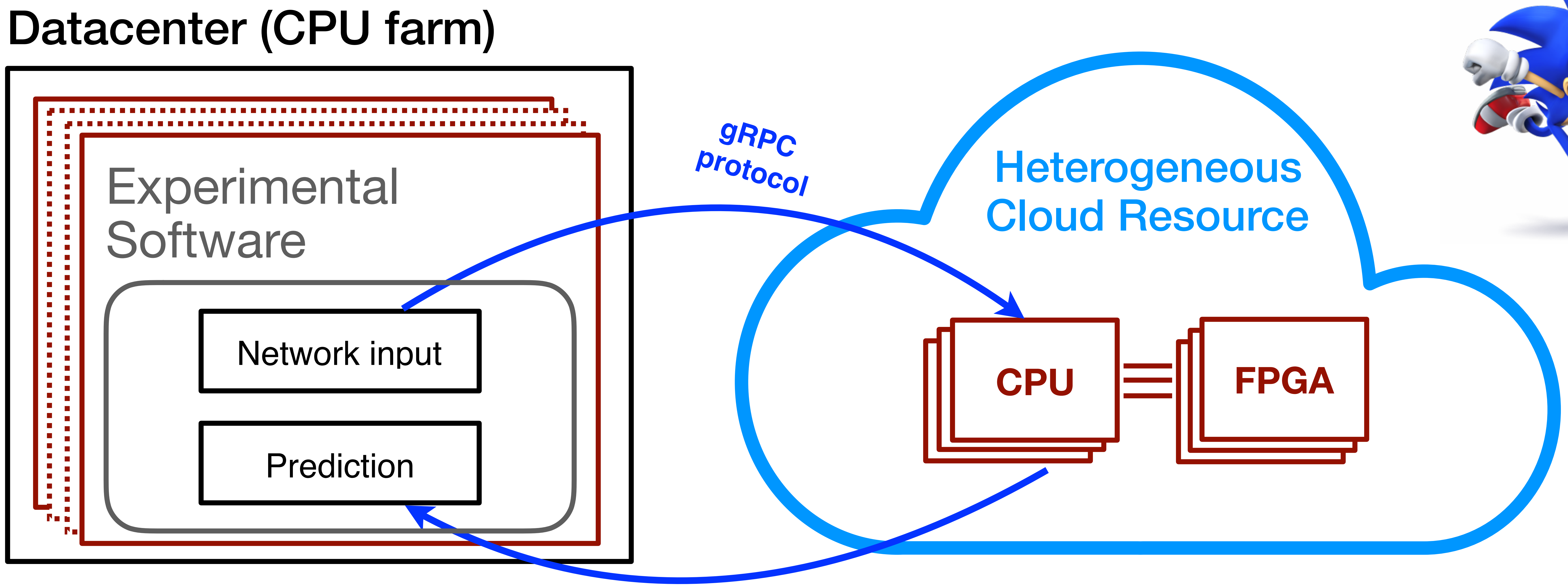


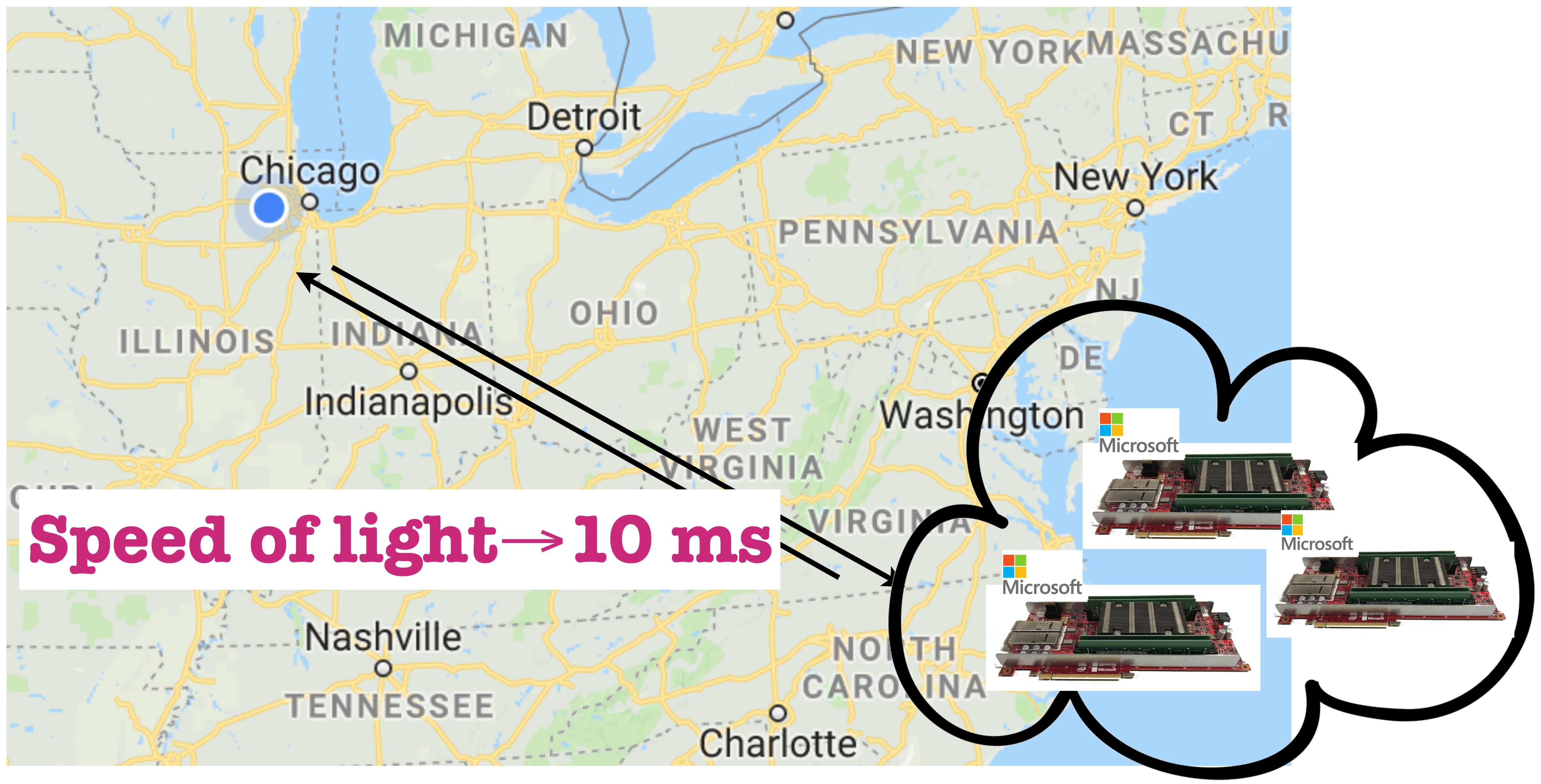


Deploy MS Brainwave as a service:

- Implemented with CMSSW ExternalWork module
- Fits CMS computing model in a non-disruptive way

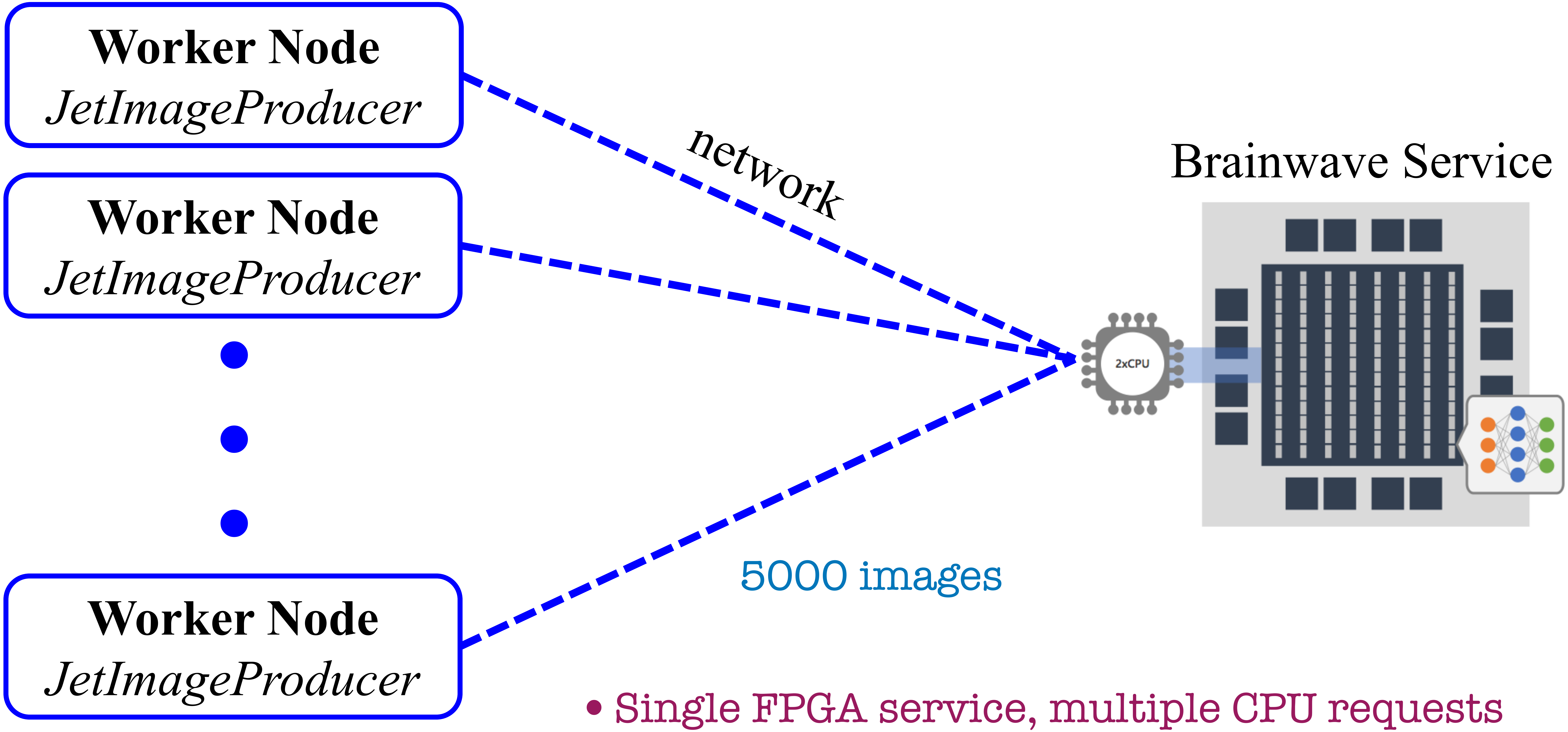
Single inference speed tests





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

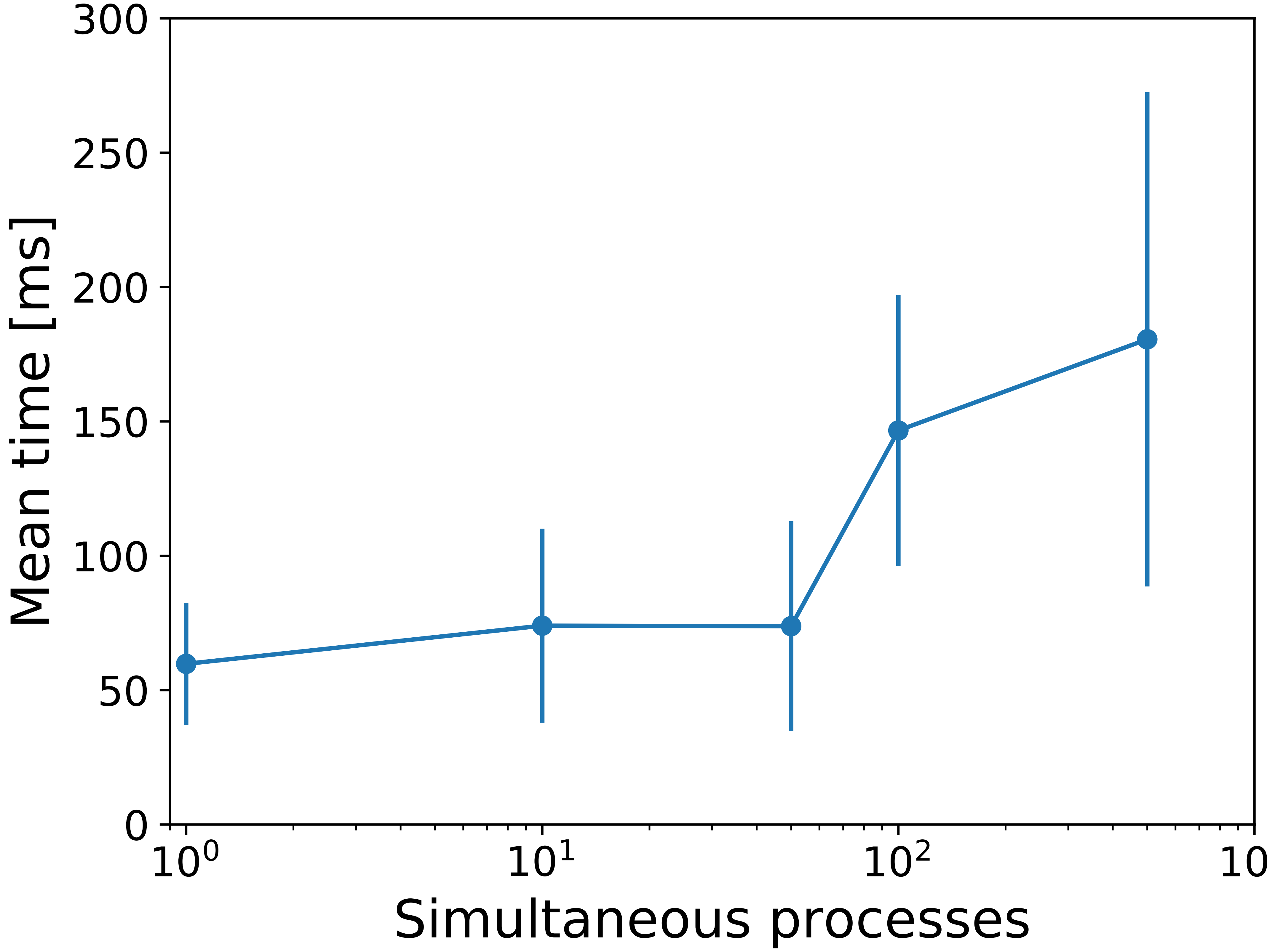
Latency Scaling test

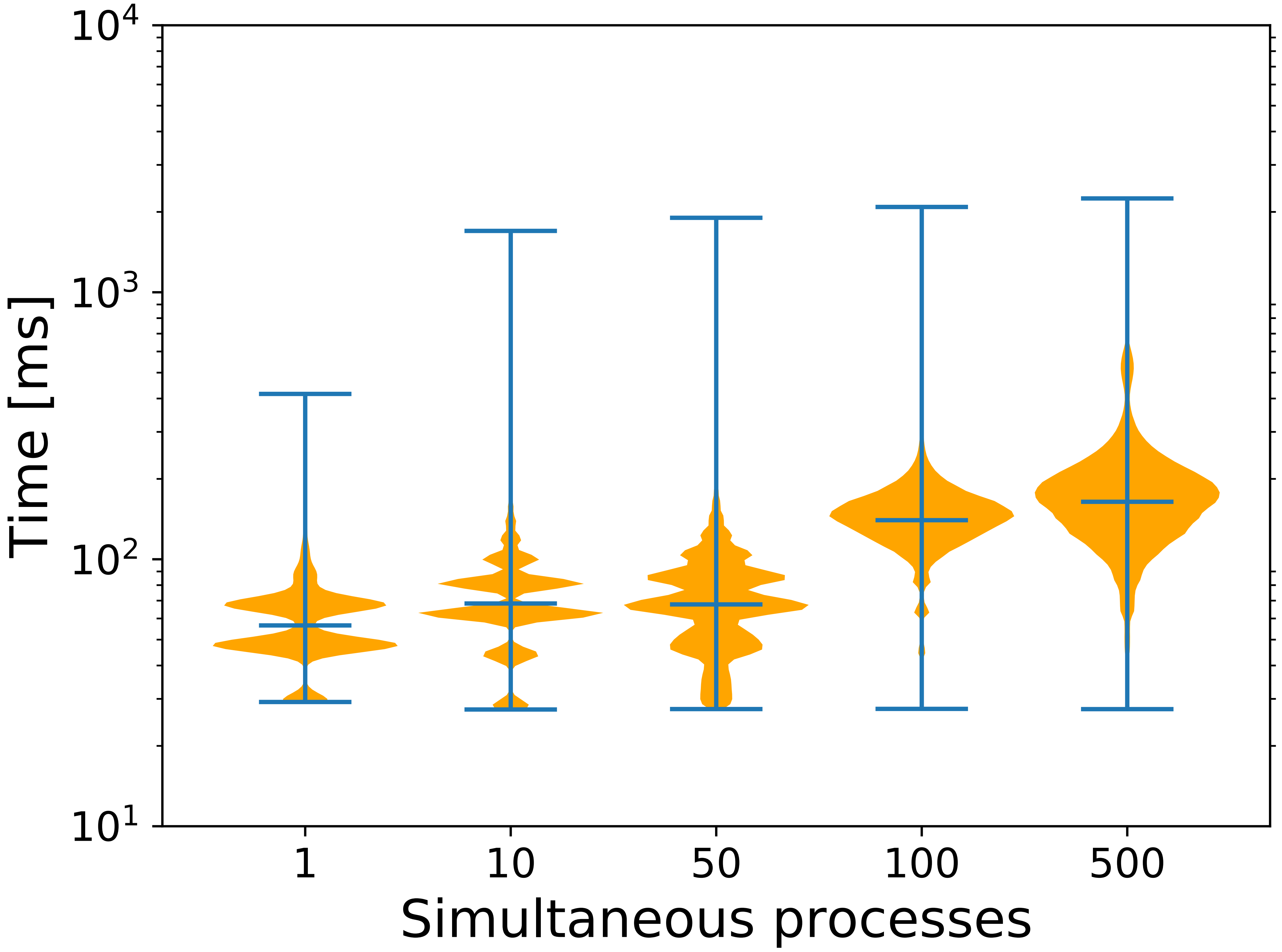


- Single FPGA service, multiple CPU requests
- Each request sends 5000 images

N: simultaneous processes

Latency Scaling test





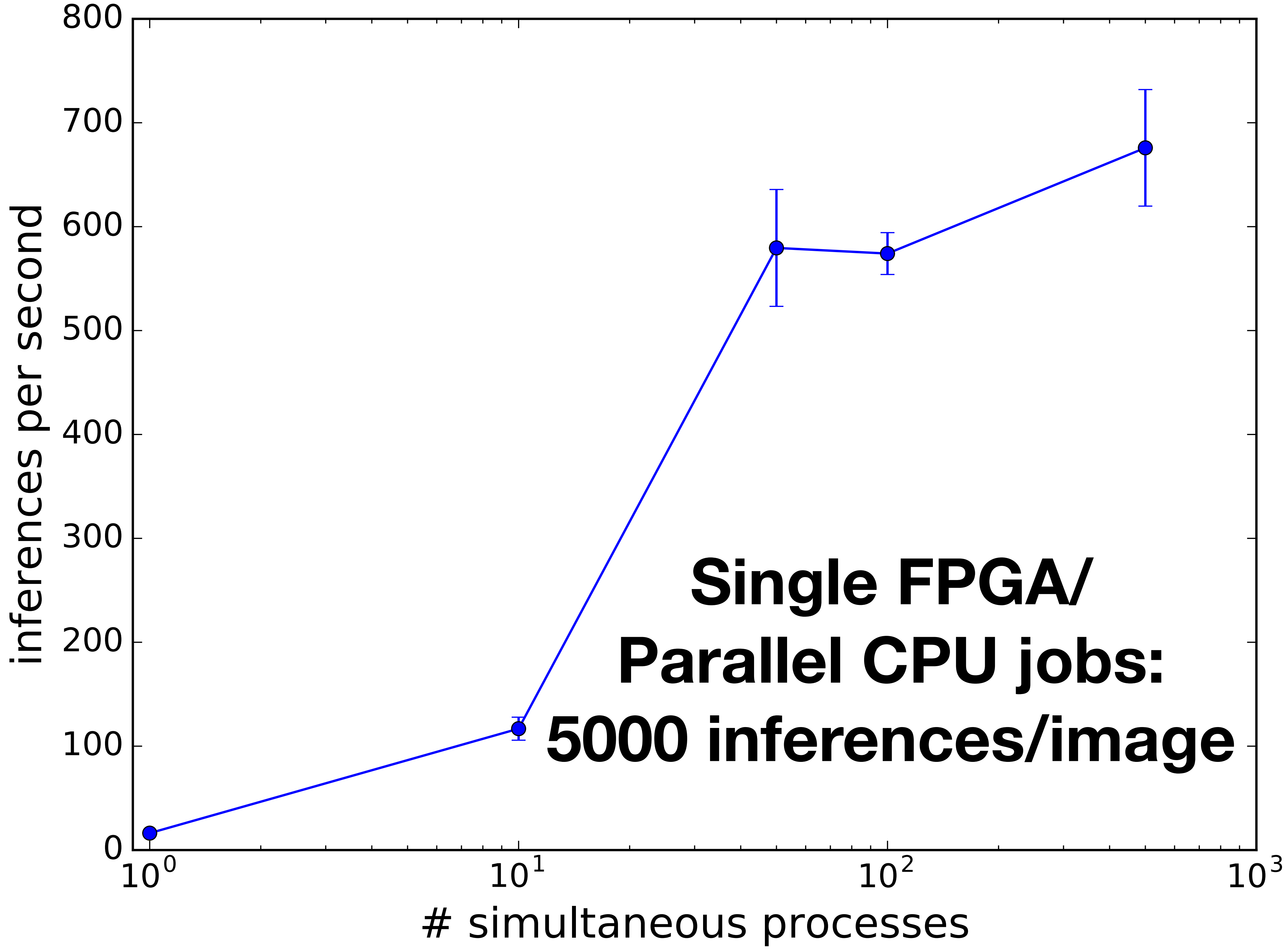
Tests:N=1,10,50,100,500

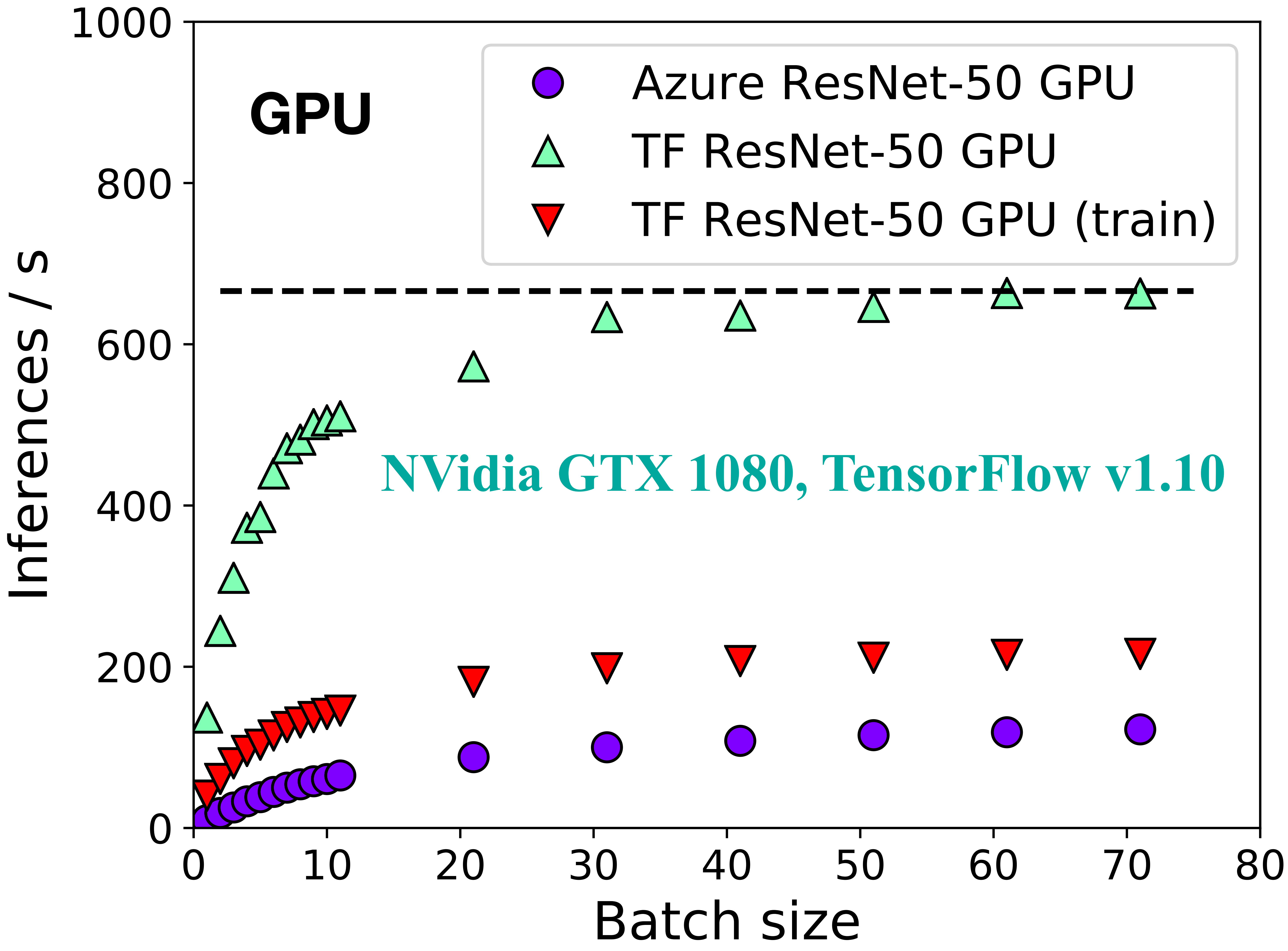
- Moderate increases in mean, standard deviation, and long tail for latency

o Fairly stable up to N = 50

DATA THROUGHOUT COMPARED TO GPUs

Brainwave cloud service

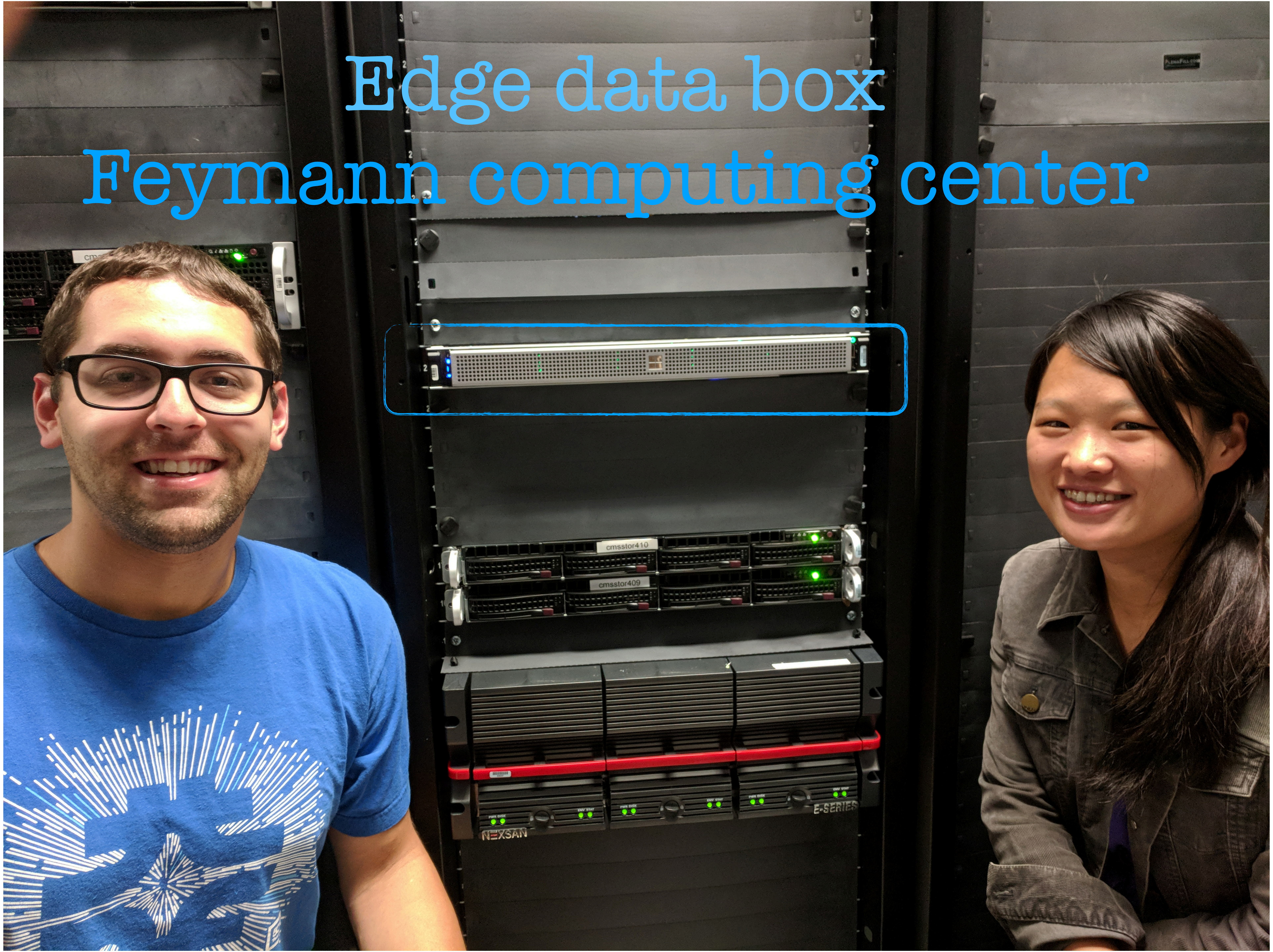




Comparable max data throughout: 600-700 images/

Fermilab

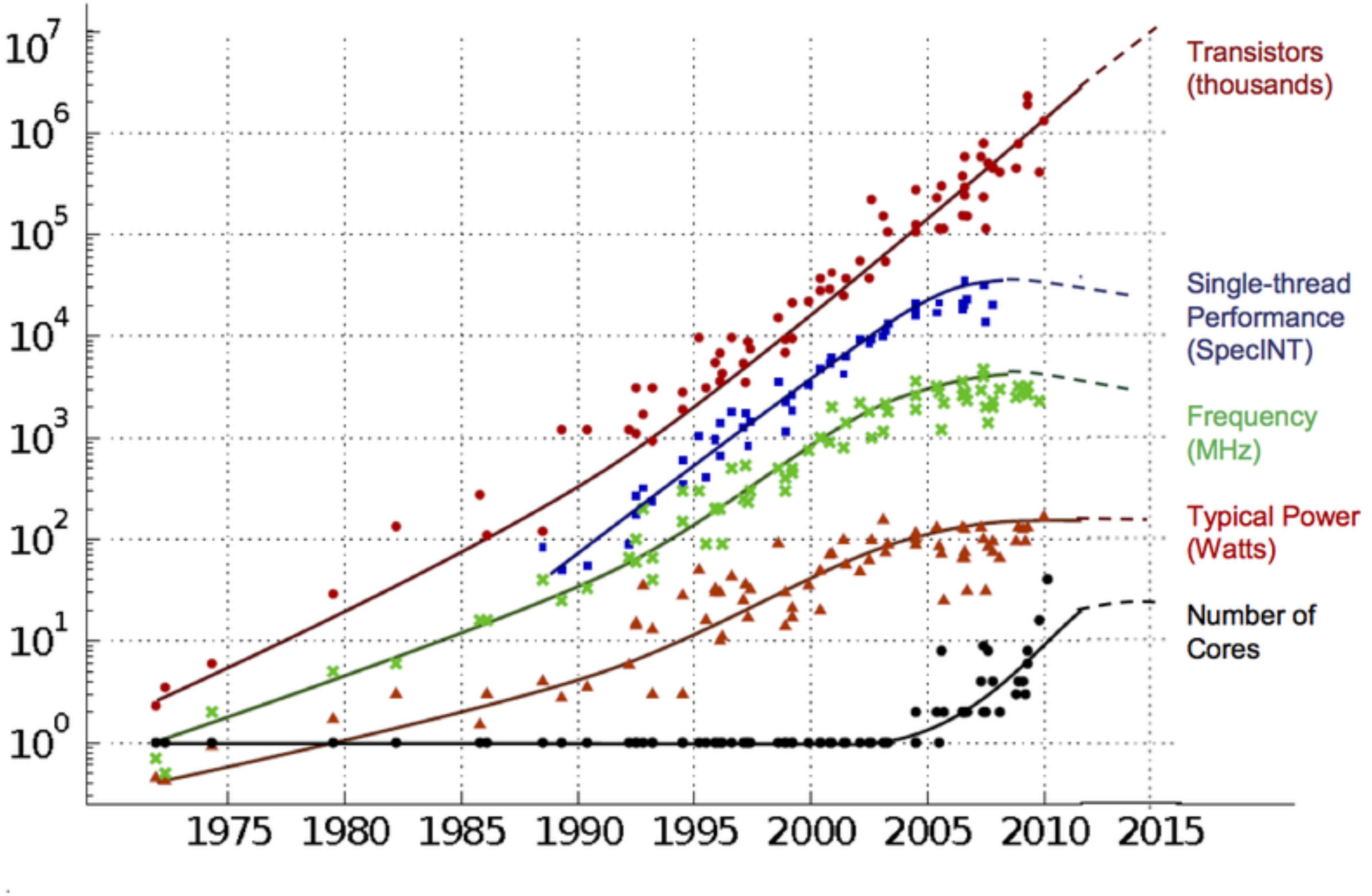
M.LIU



Summary and outlooks

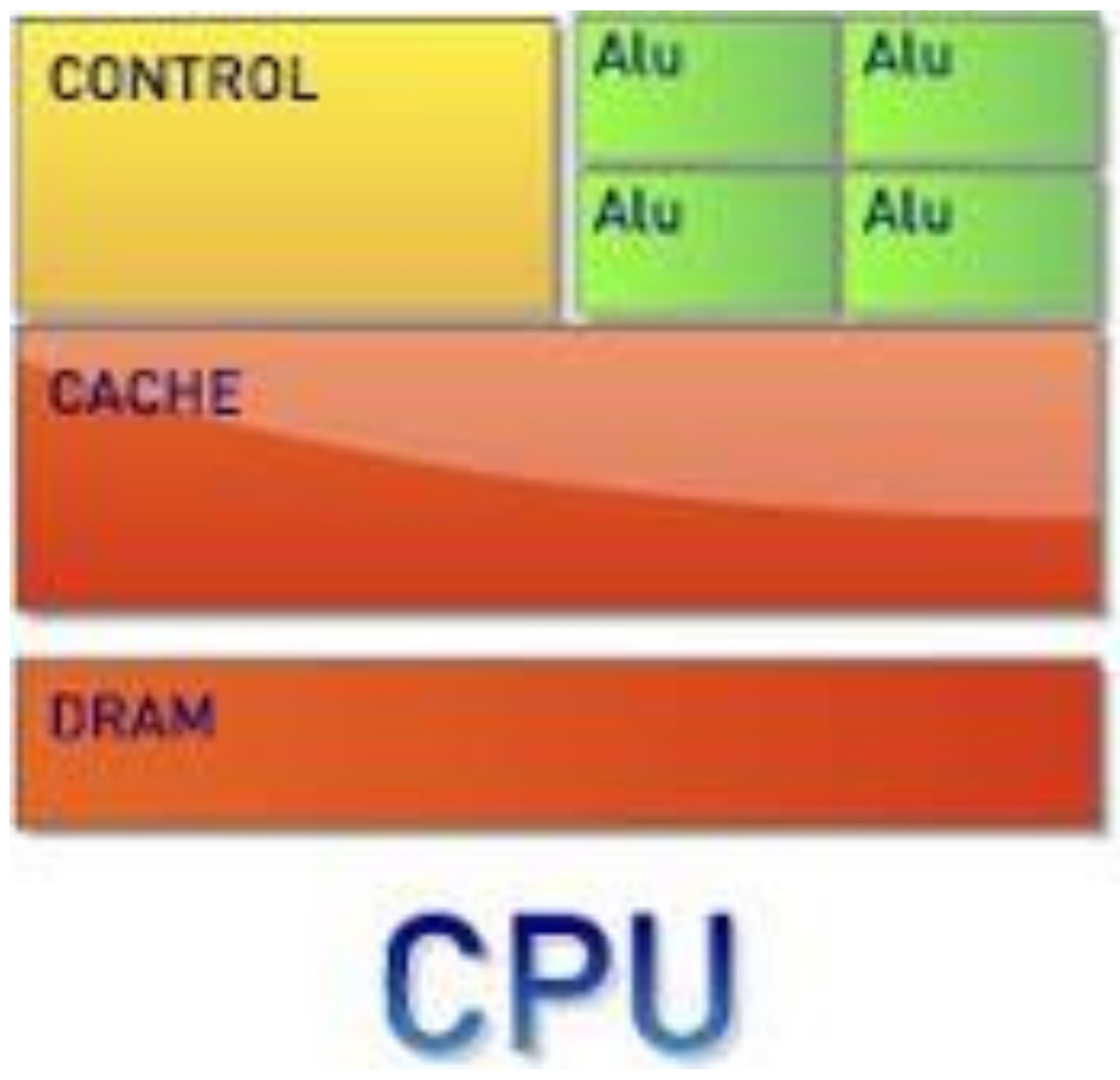
- **Proof of concept study with Brainwave:**
  - Integrate heterogeneous computing in our software framework
- **We are doing studies to benchmark other options (speed and scaling):**
  - Intel Open Vino, AWS, Google TPU…
- **Other (top) user's considerations for a dream Heterogeneous Computing Platform:**
  - Flexibility: Model support
    - e.g. Support for Graph Neural Networks
    - Support for ML framework
  - As a service: Cloud and Edge. Cost model.
- **Only works if we can solve our problems with ML! See showcases in the Lightning round talks.**

Moore's Law and Dennard Scaling



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Moore's Law continues
...but Dennard Scaling fails



Single threaded performance not improving

~2005: "The Era of Multicore"

Computing Platform Pros & Cons:

More Flexible

Homogeneous

Specialized

More Efficient

	Perf/W	
CPUs	1X	Today's standard, most programmable, good for services changing rapidly
Manycore CPUs	3X	Many simple cores (10s to 100s per chip), useful if software can be fine-grain parallel, difficult to maintain.
GPUs	5-30X	Good for data parallelism by merged threads (SIMD), High memory bandwidth, power hungry
FPGAs	5-30X	Most radical fully programmable option. Good for streaming/irregular parallelism. Power efficient but currently need to program in H/W languages.
Structured ASICS	20-100X	Lower-NRE ASICs with lower performance/efficiency. Includes domain-specific (programmable) accelerators.
Custom ASICs	> 100X	Highest efficiency. Highest NRE costs. Requires high volume. Good for functions in very widespread use that are stable for many years.

Conventional programming — C/C++

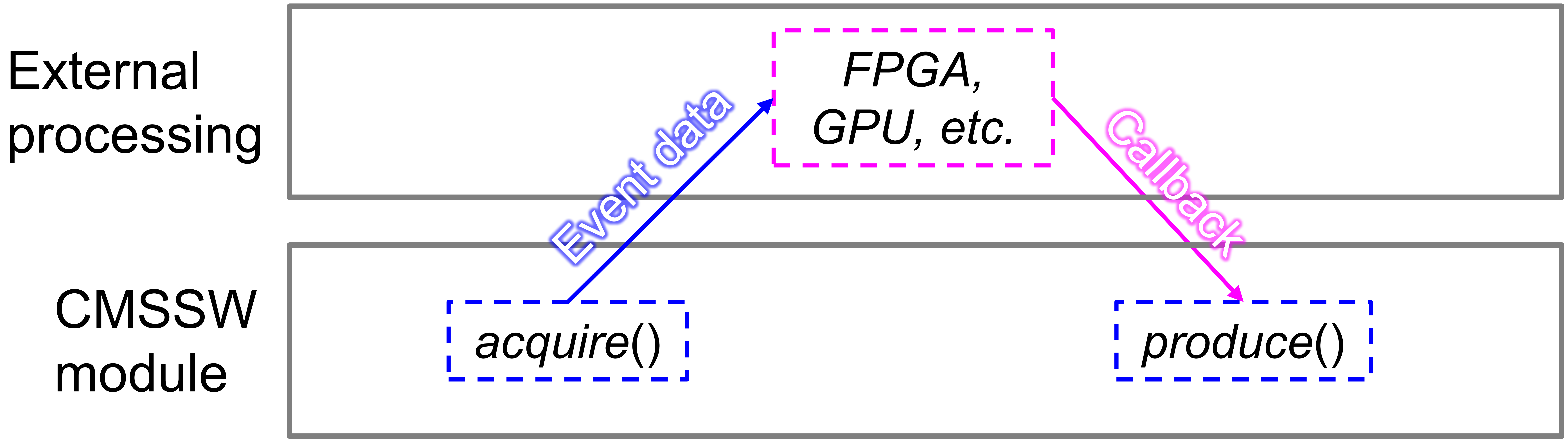
Alternative programming — CUDA, Verilog

Can't change functionality — Verilog

Electricity bill

Software/Electrical engineer hours

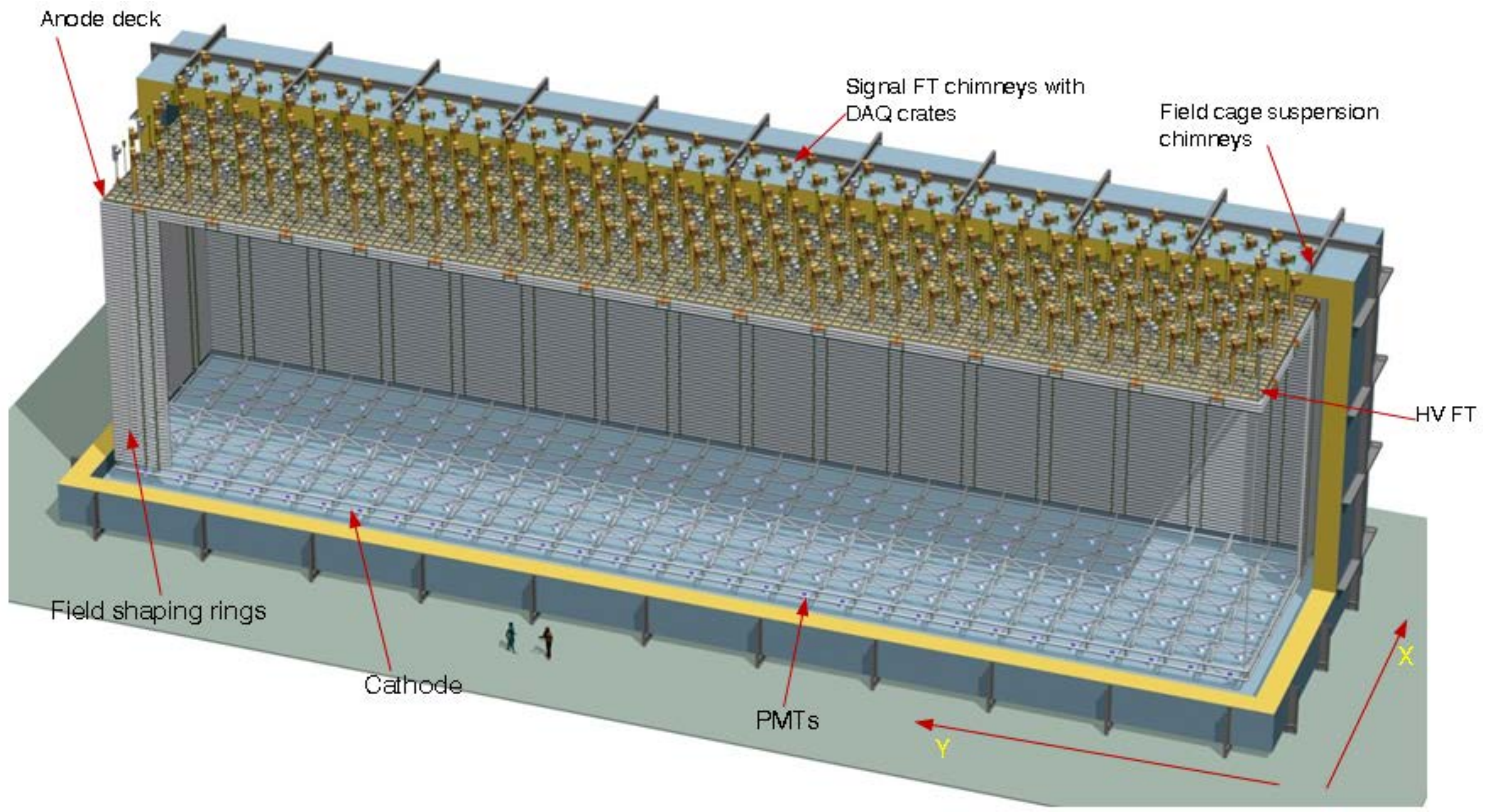
SONIC: IMPLEMENTATION IN CMSSW



Deploy MS Brainwave as a service:

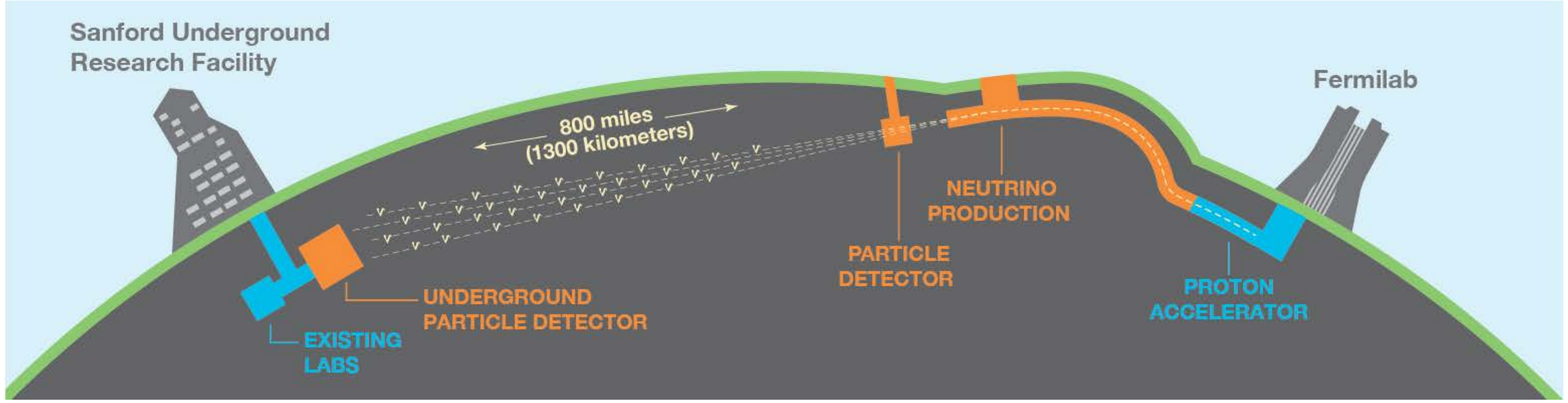
- Implemented with CMSSW ExternalWork module
- Fits CMS computing model in a non-disruptive way

Neutrino Computing Challenges



Intensity frontier: DUNE

- Largest liquid argon detector ever designed
- ~1M channels, 1 ms integration time w/ MHz sampling → 30+ petabytes/year



➢ CPU needs for particle physics will increase by *more than an order of magnitude* in the next decade

Run-2&3 data enables first observations

Example: First observation of ttH using particle Run 2 data.

ttH production