From Beams to Bits:
Real-Time AI Engines for Particle Physics

Shih-Chieh Hsu
University of Washington

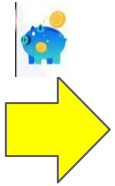
Mar 9 2026
ICEPP, University of Tokyo

PHY-2117997

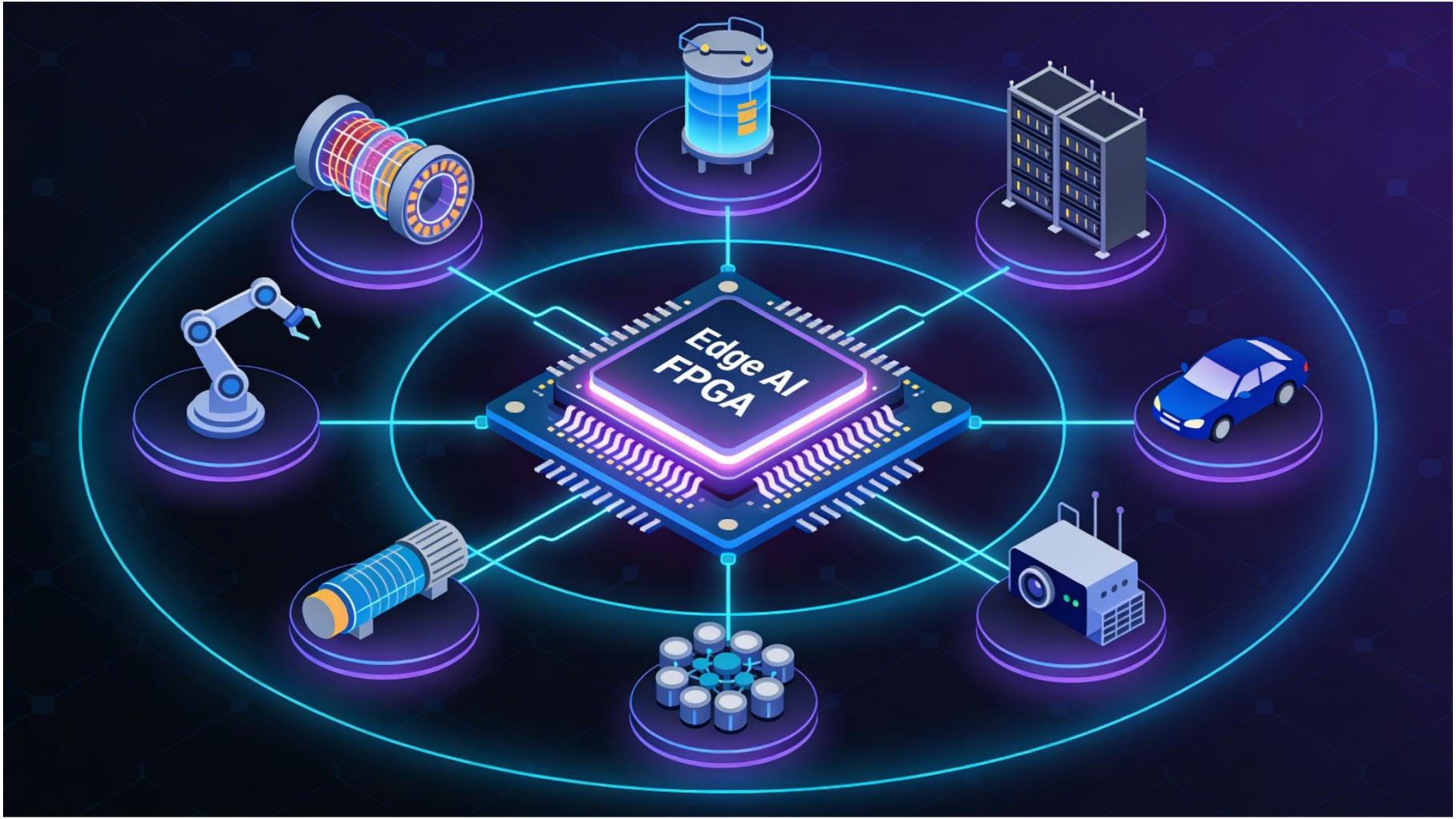
https://a3d3.ai/

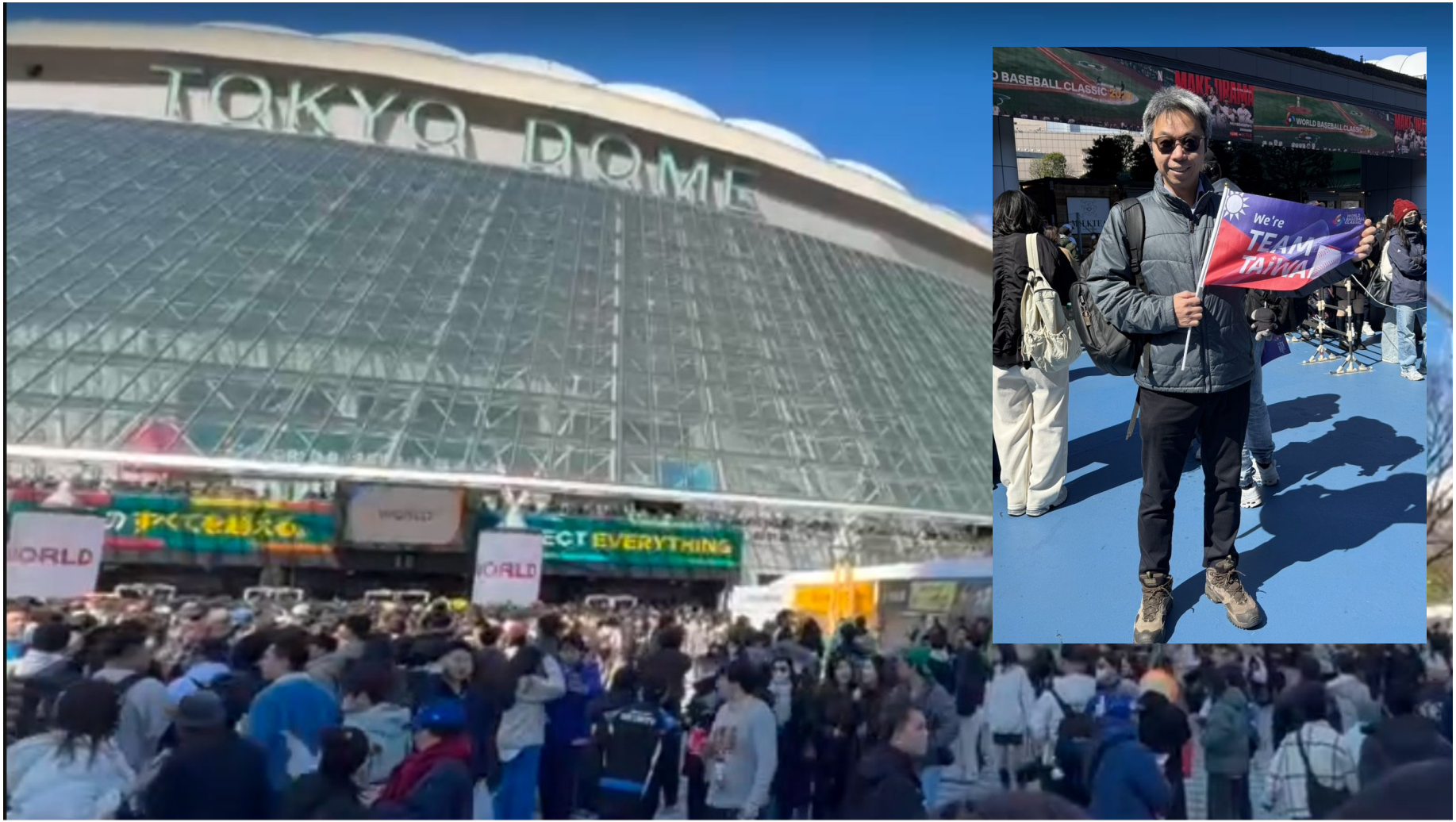
Open-source Python package that converts trained ML models (Keras, PyTorch, ONNX, etc.) into FPGA/ASIC firmware via **high-level synthesis** for low-latency inference

1 FPGA engineer + 0.3 FTE ML engineer.
3 months



0.3 FTE ML engineer (no HDL experiences).
5 day


Edge AI
FPGA


TOKYO DOME
EVERYTHING

We're
TEAM

Shih-Chieh (Jay) Hsu
徐士傑

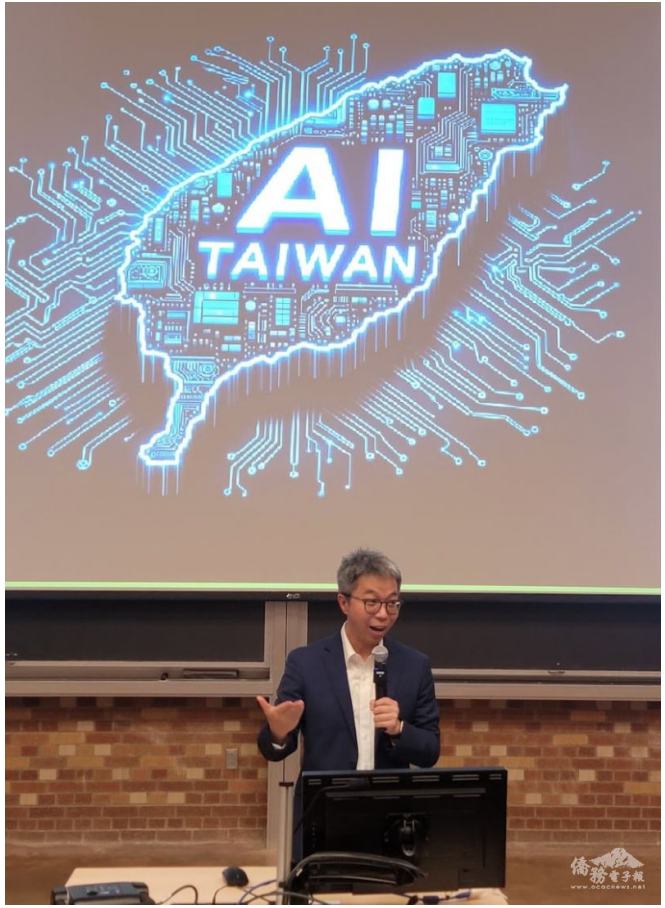
- Professor, University of Washington Physics
- Adjunct Professor, University of Washington Electrical and Computer Engineering
- Adjunct Professor, National Tsing Hua University Physics
- Director, NSF HDR Institute A3D3
- Vice President, National Taiwan University Alumni Association of Northwest America
- Vice President, North America Taiwanese Science and Engineering Association, Seattle Chapter

https://www.linkedin.com/in/shih-chieh-hsu-hep/
https://www.facebook.com/schsuphy

Dr. Hsu studies studies fundamental science using real-time AI.

schsu@uw.edu


2025
U.S. Taiwan Tech Connect
AI Trends & Industry Insights
NATEA SEATTLE
Ethan Tu
Founder & CEO
Taiwan AI Labs
Elaine Lu
VP of Engineering Operations
CoreWeave
Shyam Gollakota
Professor of CSE, UW
Co-founder of Heatvana AI
Kwok Cheung
Founder & CEO
KC-Exousia Consulting
Miriam Teng
Principal Accounting Manager, Amazon
Hubert Tsai
Founder & CTO
Cheehoo
Nov 15th at UW
Scan to register
NSTC 國家科學及技術委員會
National Science and Technology Council
tatwanzonian

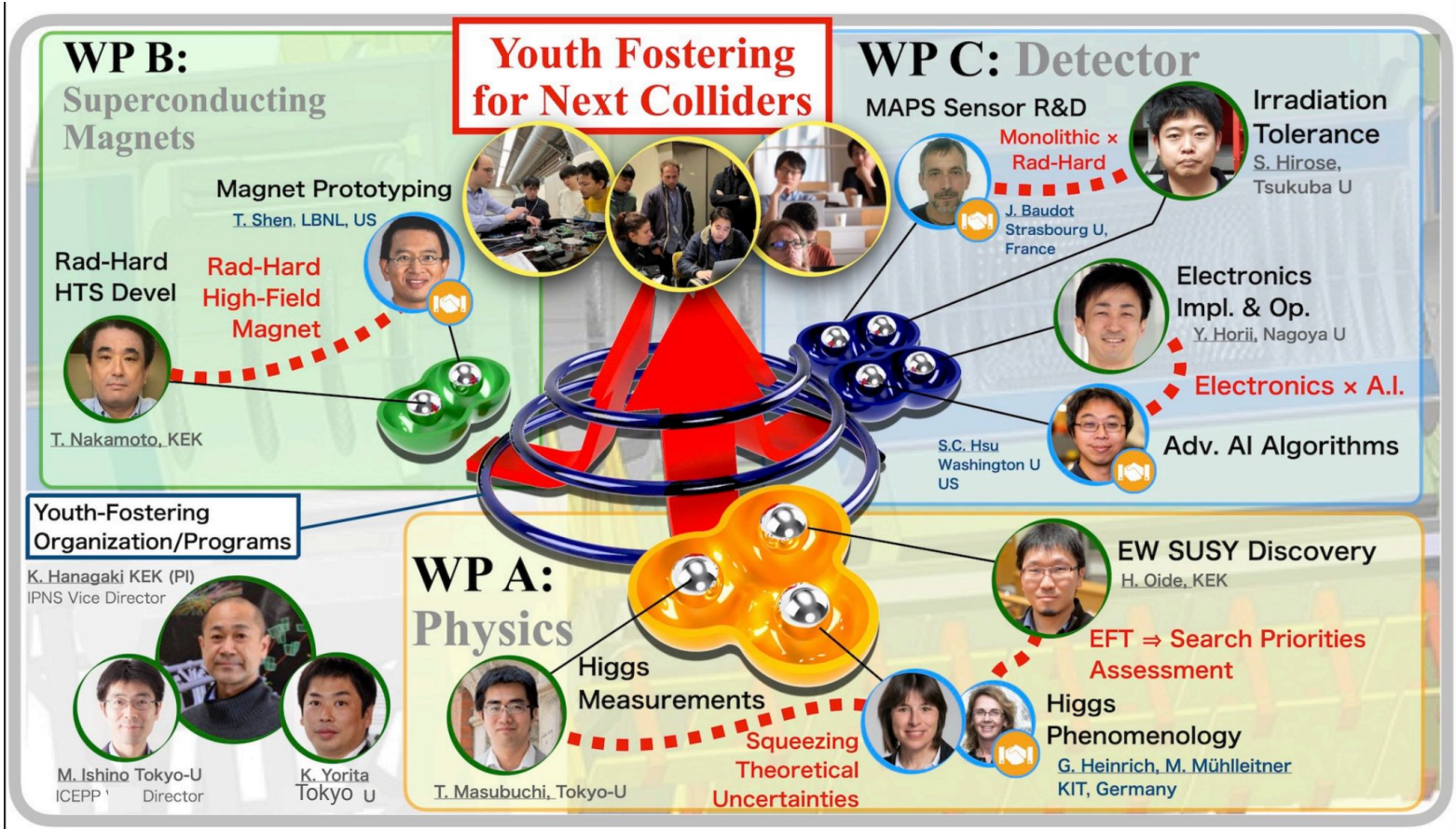

AI
TAIWAN


U.S. Taiwan
Tech Connect


北美台灣工程師協會
North America Taiwanese Engineering & Science Association


Startup Journeys
& Lessons Learned
NATEA SEATTLE
Will Yang
Moderator
Jonathan Chang
Investor
Hubert Tsai
Founder & CTO
Derrick Chiang
Founder & CEO

Introduction/Advertisement of International Leading Research (JSPS grant)
Development of the Energy Frontier and International Human Resources toward Unified Understanding of Interaction



Real-Time Adaptive AI for Sleep Diagnosis, Treatment and Therapy Cross-Pacific Initiative (X-PAI)

Dr. Shih-Chieh Hsu (PI)
Professor of Physics
Director of A3D3

Dr. Scott Hauck
Professor of ECE

Dr. Amy Orsborn
Professor of ECE/BioEng

Dr. Eli Shlizerman
Professor of ECE/AMath

NSF HDR Institute A3D3, University of Washington

Dr. Shun Nakajima (PI)
Associate professor

Dr. Hiroyuki Kitagawa
Professor

Dr. Kei Muroi, MSc. Shio Maeda, MSc. Hiroku Noma
Dr. Azusa Ishii

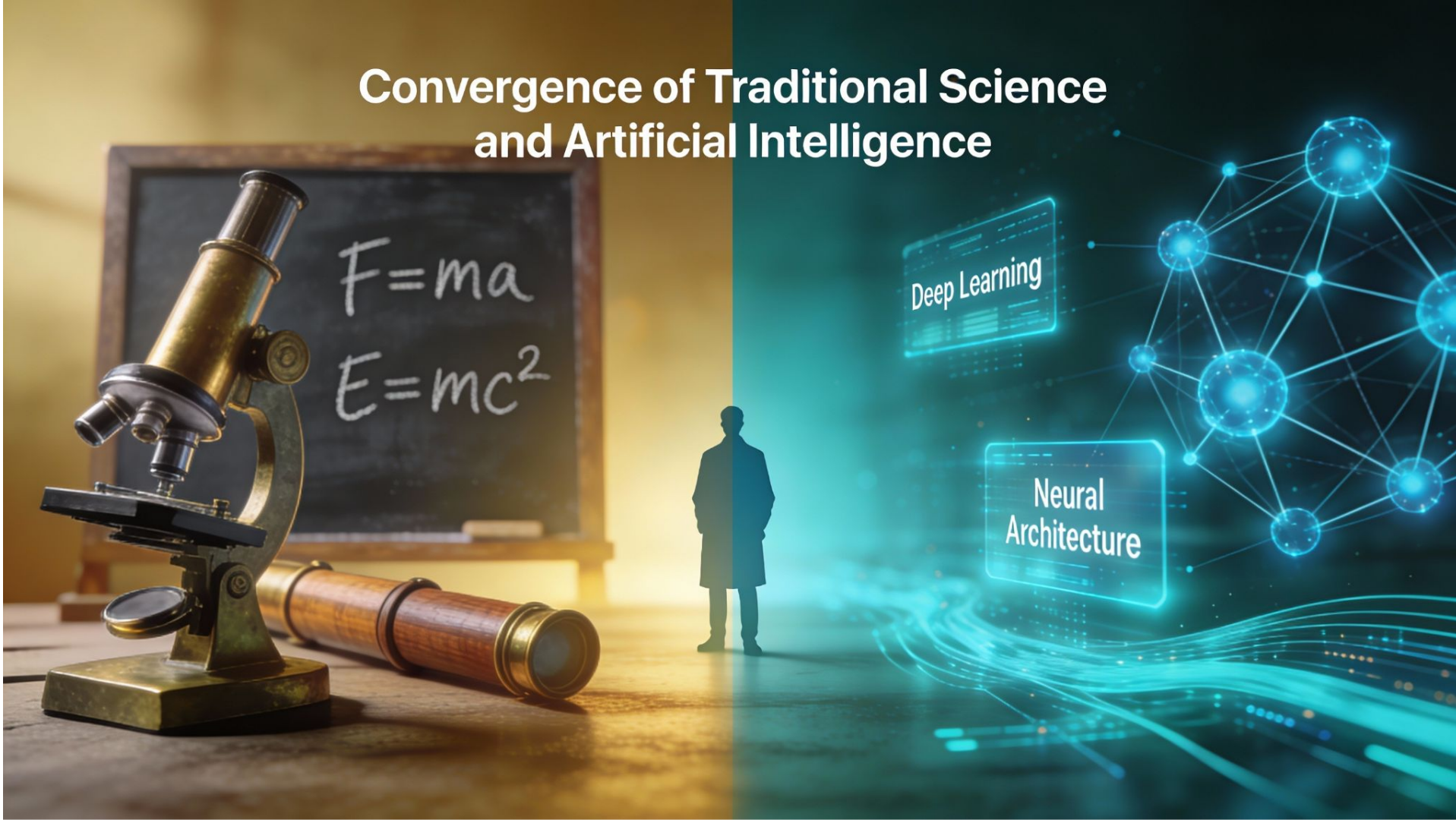
International Institute for Integrative Sleep Medicine
University of Tsukuba

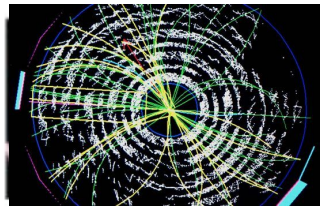
日本睡眠学会専門医療機関A型
岐阜メイツ睡眠クリニック

Gifu Mates Sleep Clinic
Sleep Medicine Specialist

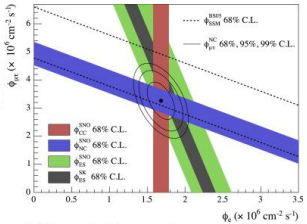
Dr. Haruhito Tanaka (Physician)

Hiroaki Yamamoto
(Registered Polysomnographic Technologist)

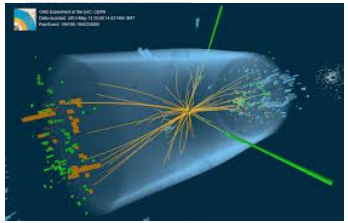

Convergence of Traditional Science and Artificial Intelligence
F=ma
E=mc²
Deep Learning
Neural Architecture



Top discovery
1995



Neutrino Oscillations
2001



Higgs Discovery
2012

?

1995
Support Vector Machine

2001
Gradient Boosting

2012
AexNet

2016
AlphaGo

2020
AlphaFold

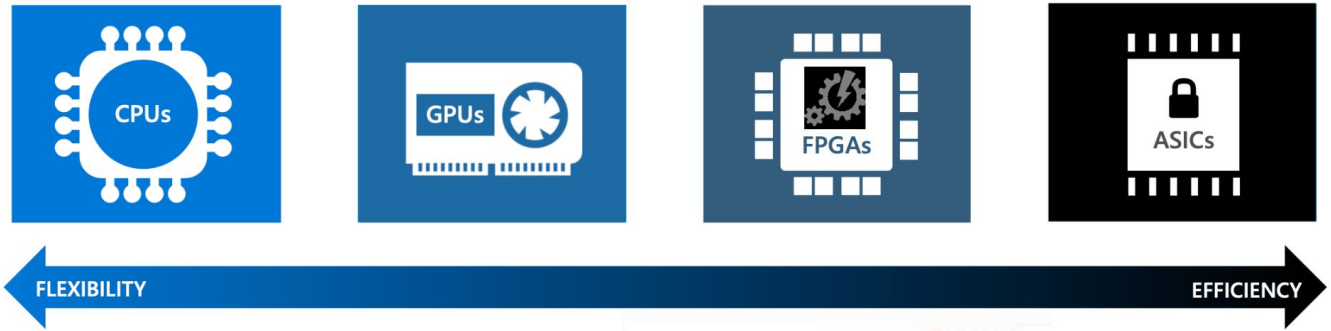
2022
ChatGPT

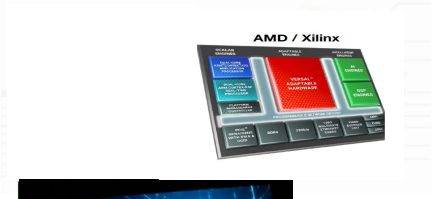
2024
Agentic AI

Deep Learning
revolution

GenAI
revolution

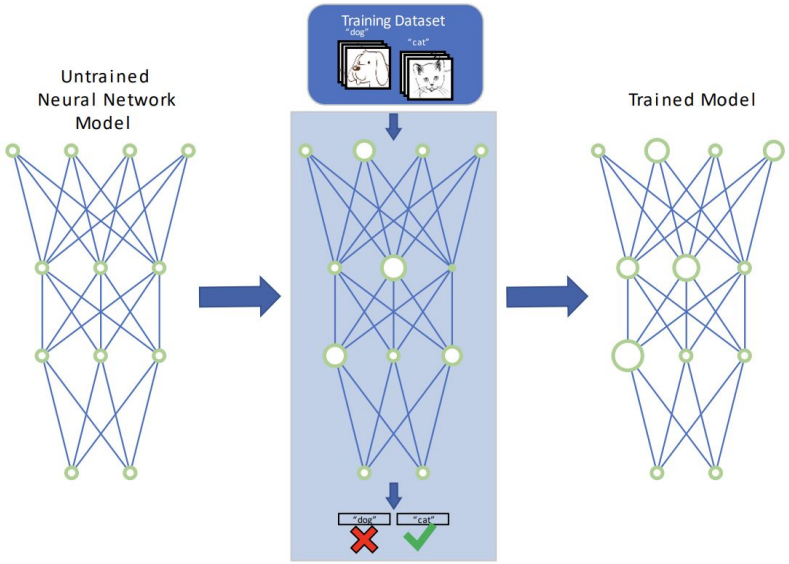
Trending in industry





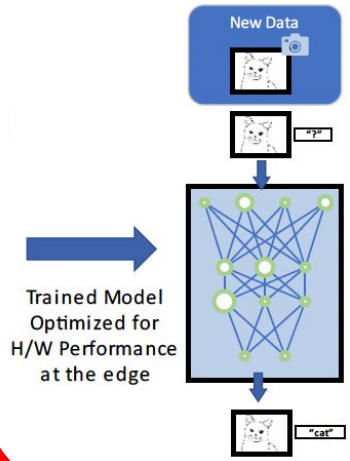
Advances driven by
big data explosion
& machine learning

Training



- Very large datasets & memory, CPU/GPU/TPU farms, floating point required
- Done in an AI/ML Framework (Tensorflow, PyTorch, etc.).

Inference



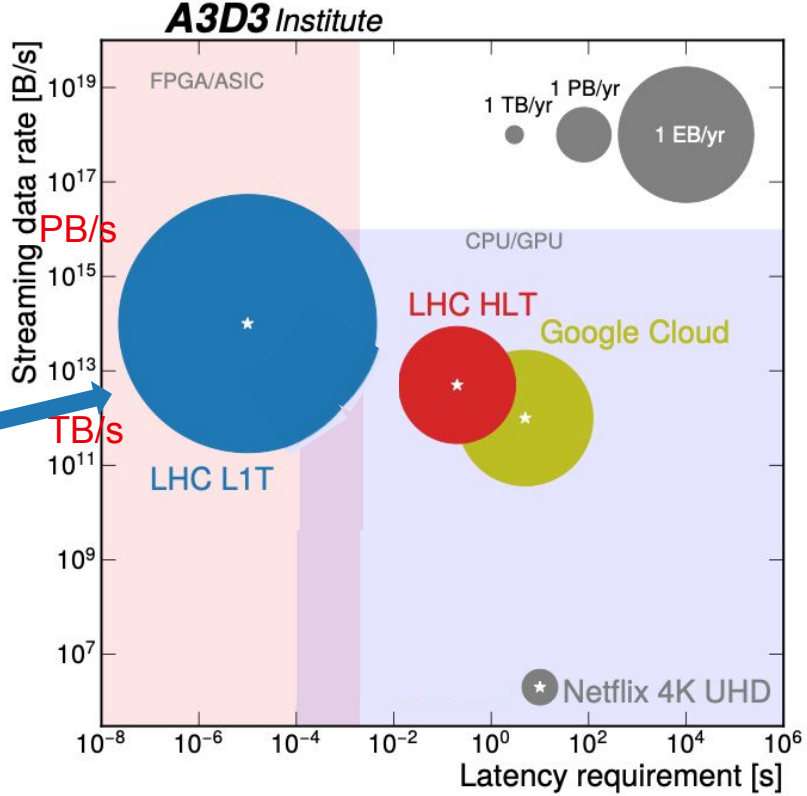
Focus of this talk

- Requires minimal computational resources
- Often has real-time performance/power requirements that require custom hardware, e.g. FPGA/ASIC/Edge devices

Credit: Marzieh Vaez Torshizi

LHC pushes technology limits

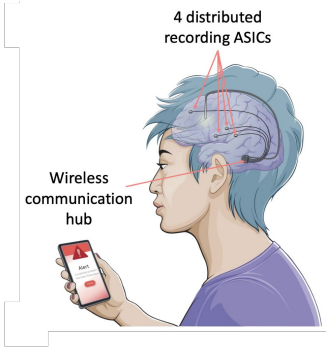
High Energy Physics

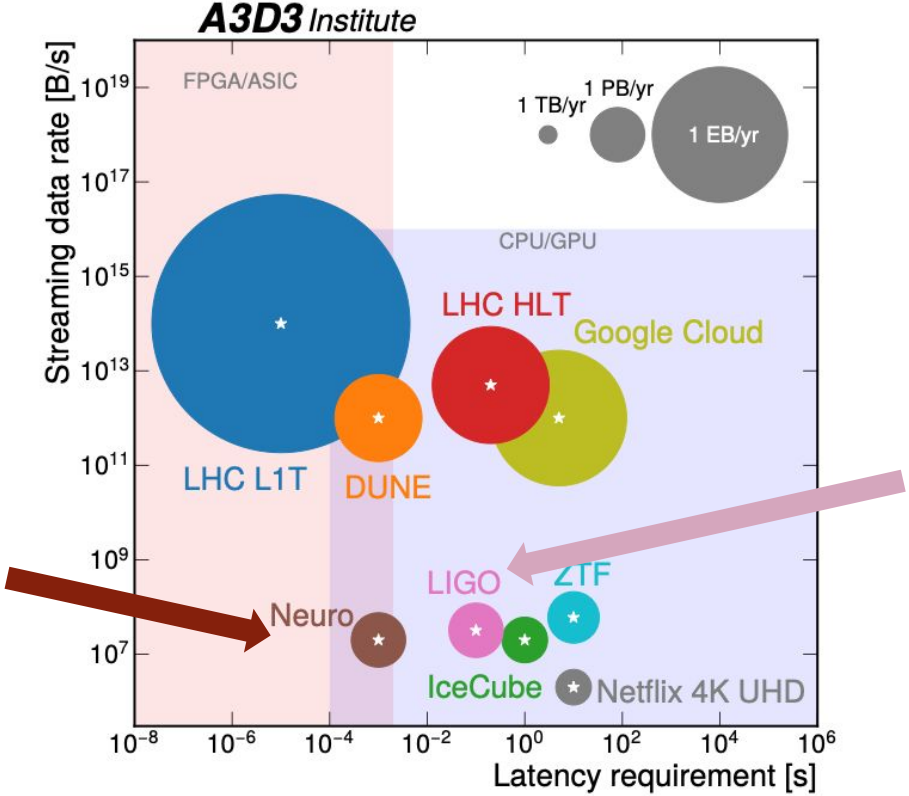


- **Data volumes** surpass industry standards
- **Streaming rates** exceed commercial benchmarks

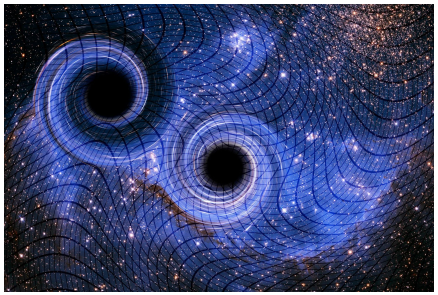
Critical Common challenges

Neuroscience





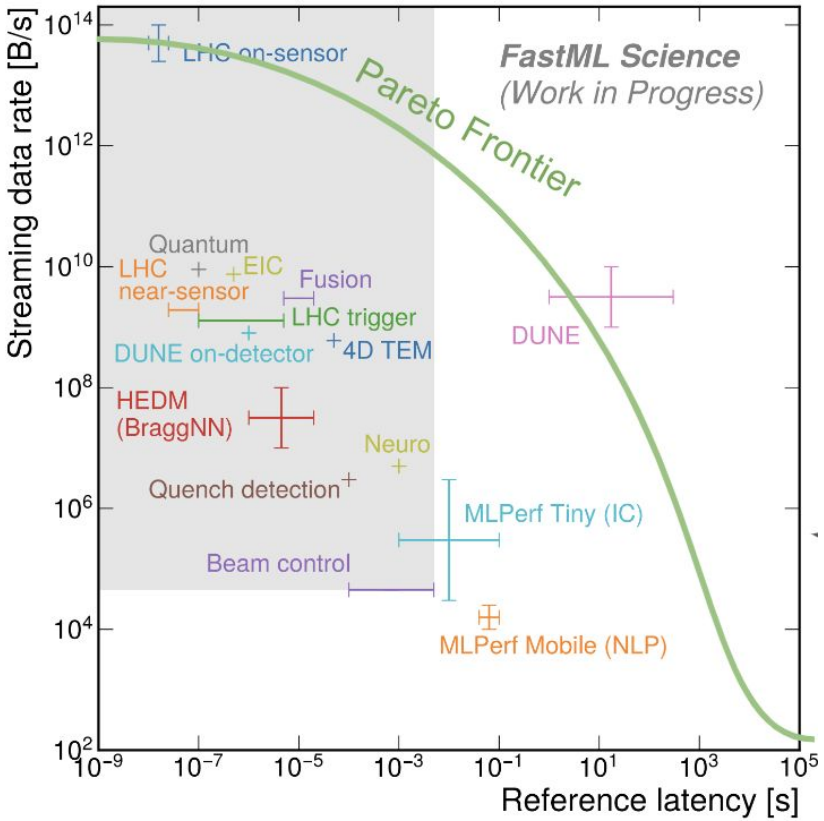
Gravitational Wave



Fast Machine Learning Community

High-performance data systems with

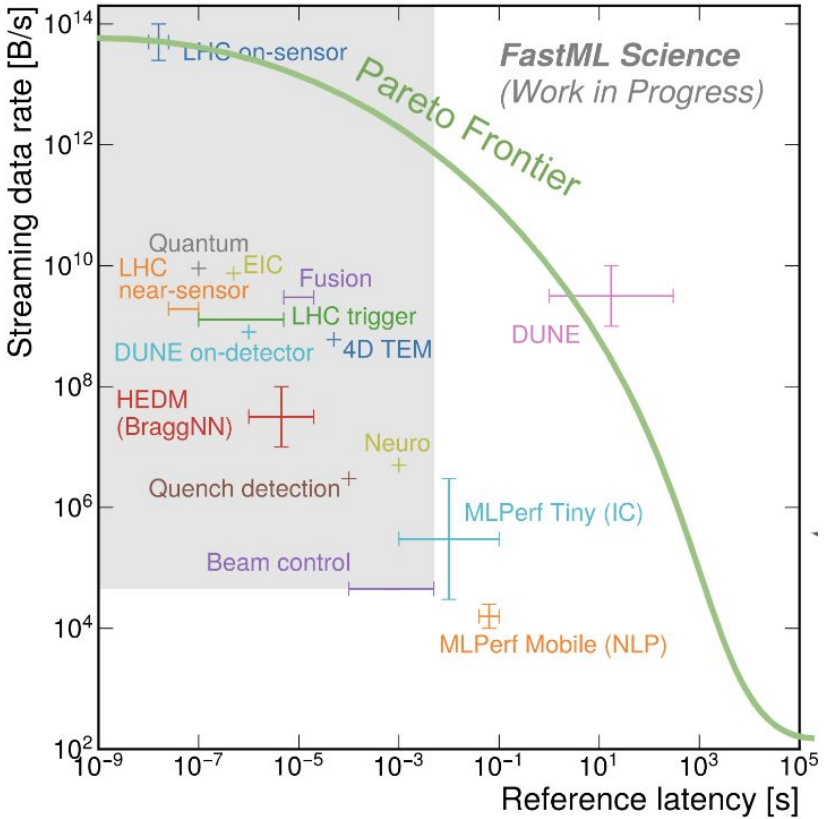
- low latency
- high throughput processing,
- real-time control modules, and
- custom processing elements.



Fast Machine Learning Community

High-performance data systems with

- low latency
- high throughput processing,
- real-time control modules, and
- custom processing elements.



a3d3

pose-hls4ml

EPIGRAPHY

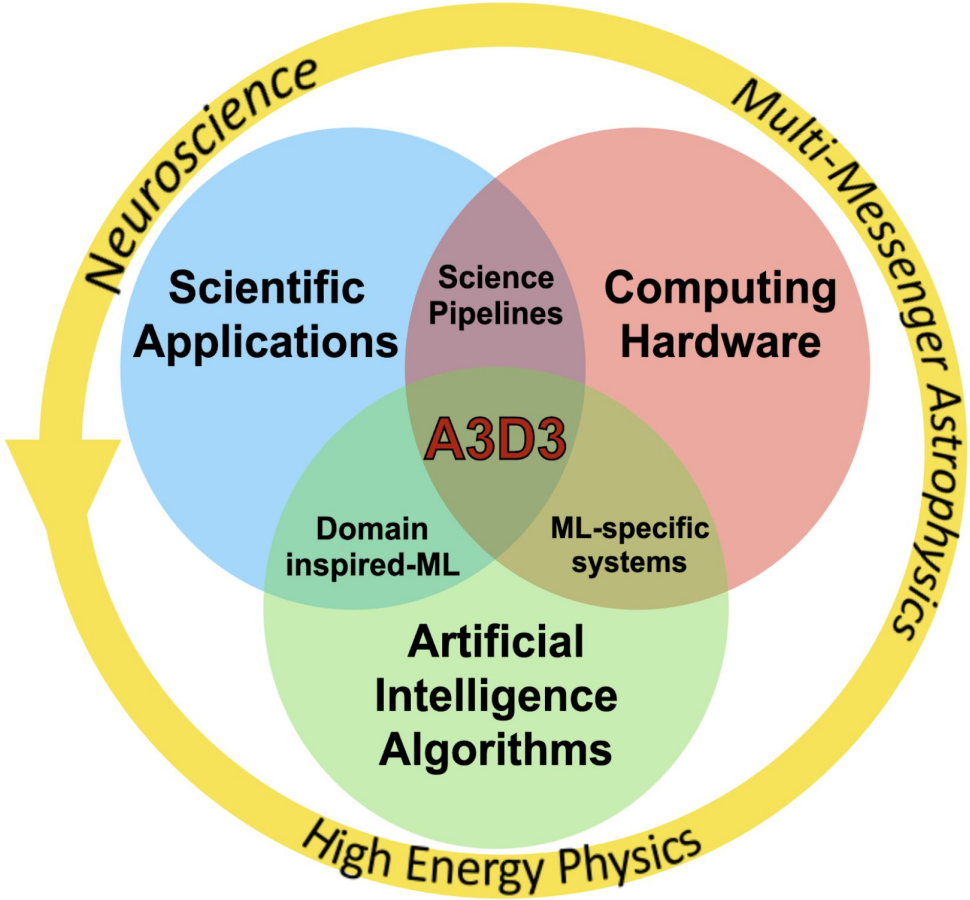
NexTGen
Next Generation Triggers

Accelerated AI Algorithms for Data-Driven Discovery

NSF HDR Institute **A3D3** (since 2021)

***A**ccelerated **A**rtificial Intelligence **A**lgorithms for **D**ata-**D**riven **D**iscovery*

Our Mission is to enable real-time AI techniques for scientific and engineering discovery by uniting three core components: Scientific Applications, Artificial Intelligence Algorithms, and Computing Hardware.

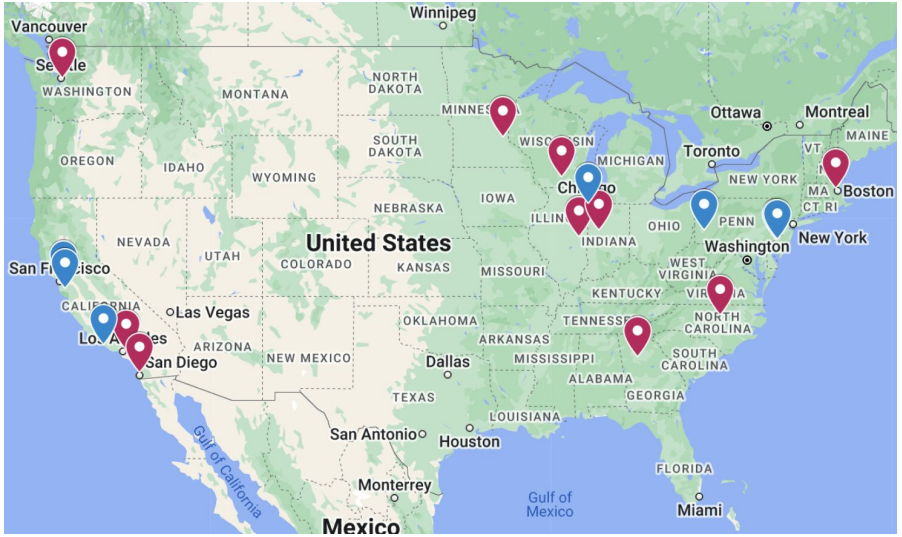

Neuroscience
Multi-Messenger Astrophysics
High Energy Physics
Scientific Applications
Science Pipelines
Computing Hardware
A3D3
Domain inspired-ML
ML-specific systems
Artificial Intelligence Algorithms

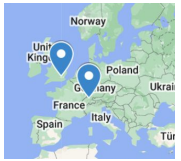
Multi-institution

NSF Harnessing the Data Revolution Institute

21 institutions
170 members
(75% trainees)

($16M/5year)





Multi-disciplines

HEP


Hsu
PI/Director


Harris
co-PI


Neubauer
co-PI


Liu


Duarte


Rankin


Aarestad


Gonski


Carlsen


Cremonesi


Buat

MMA


Coughlin
co-PI


Scholberg
co-PI


Graham


Riedel


Katsavounidis


Li


Sravan


DiPetrillo


Cavanaugh

Neuros


Orsborn


Shlizerman


Dadarlat

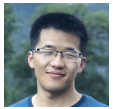

Sun


Yu

35 faculty/staff

CS/EE


Hauck


Li


Chen


Han


Ju

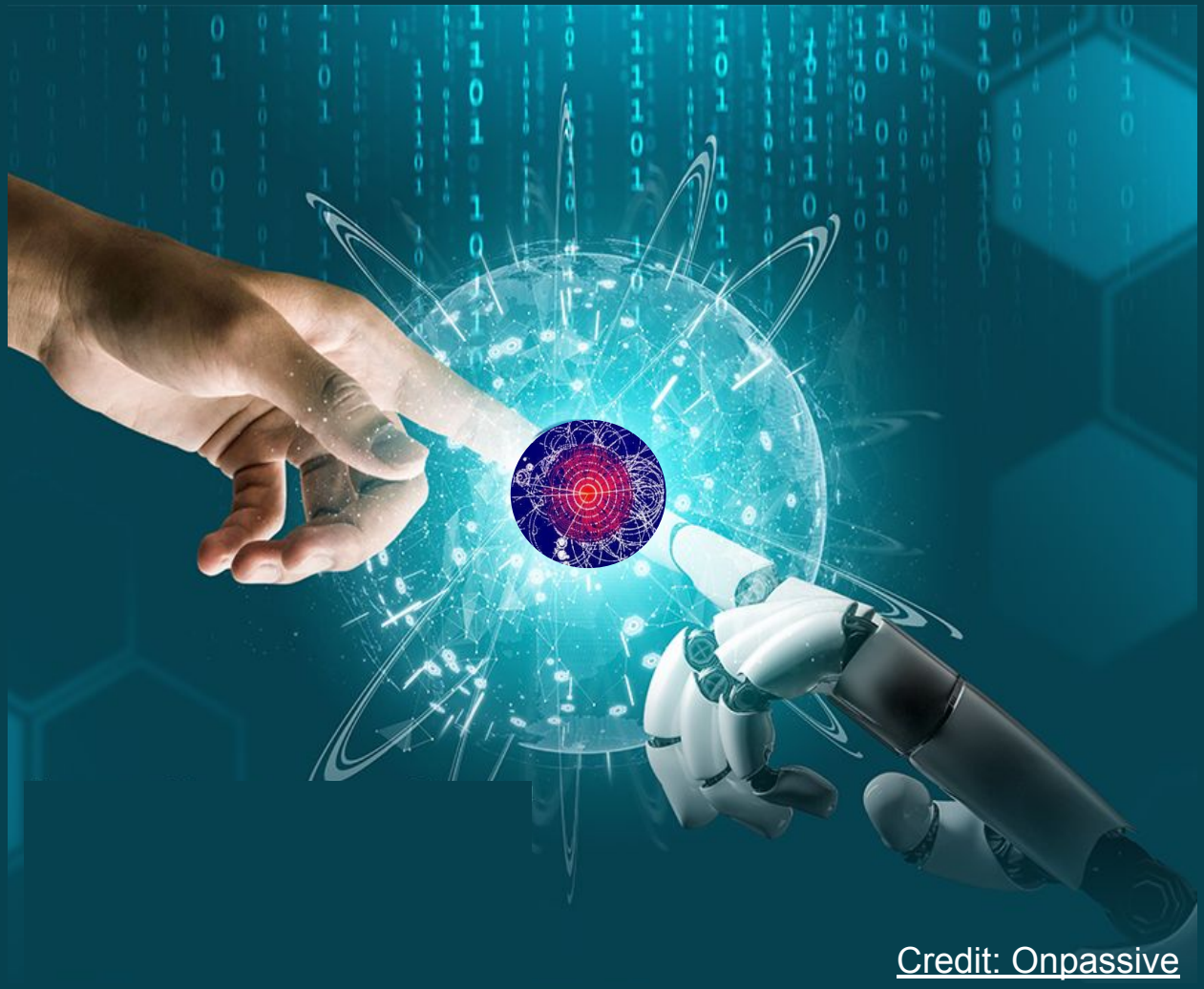

Loncar


Liu

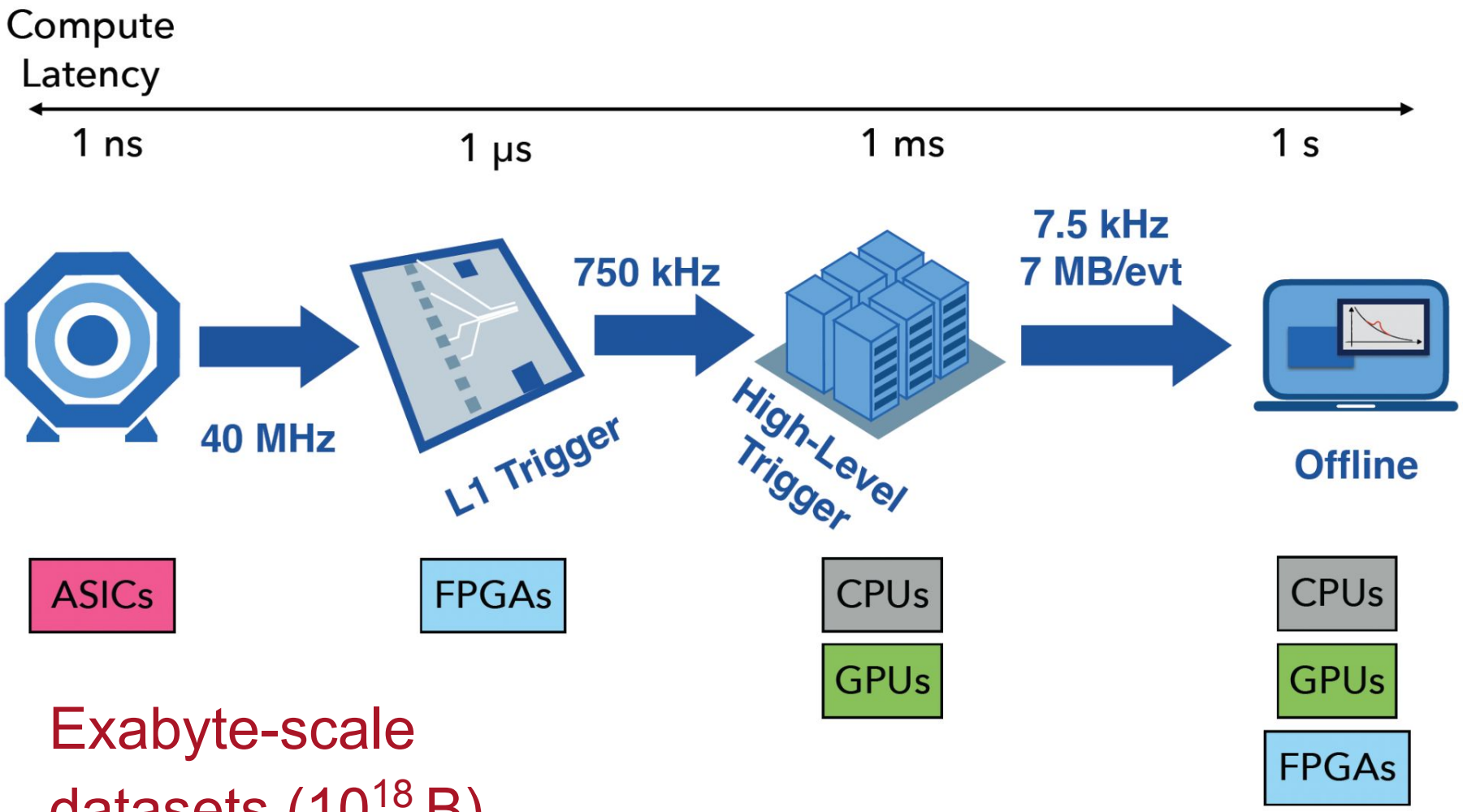

Li


Trivedi


Lai
賴伯承

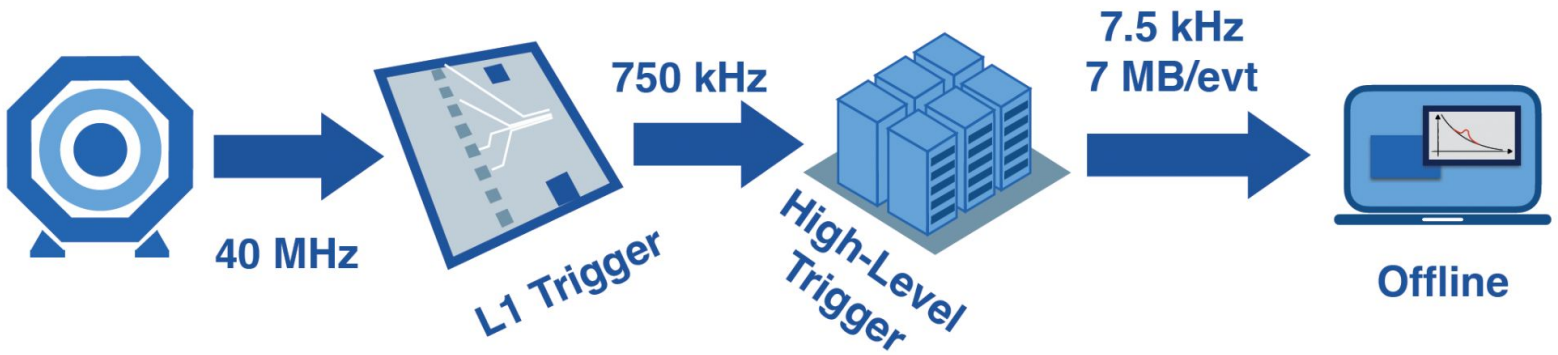


Credit: Onpassive



Exabyte-scale datasets (10^{18} B)

The Need for the FastML



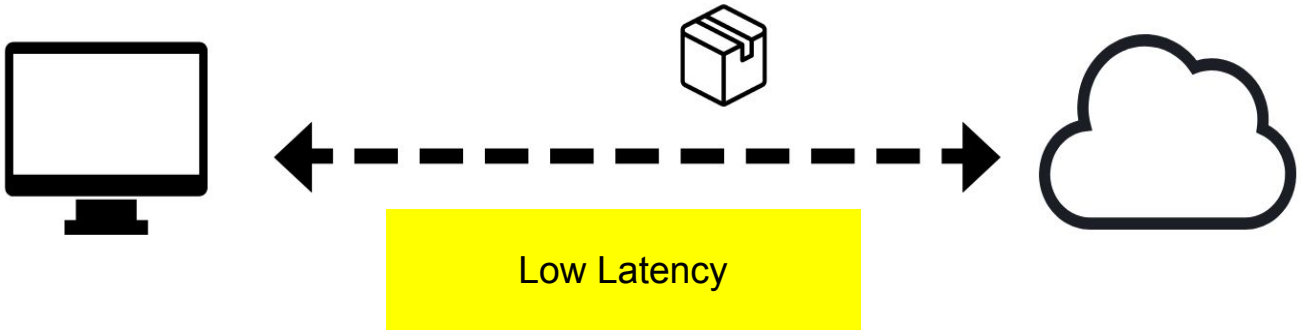
Fast Machine Learning Collaboration

Latency
(HLS4ML)

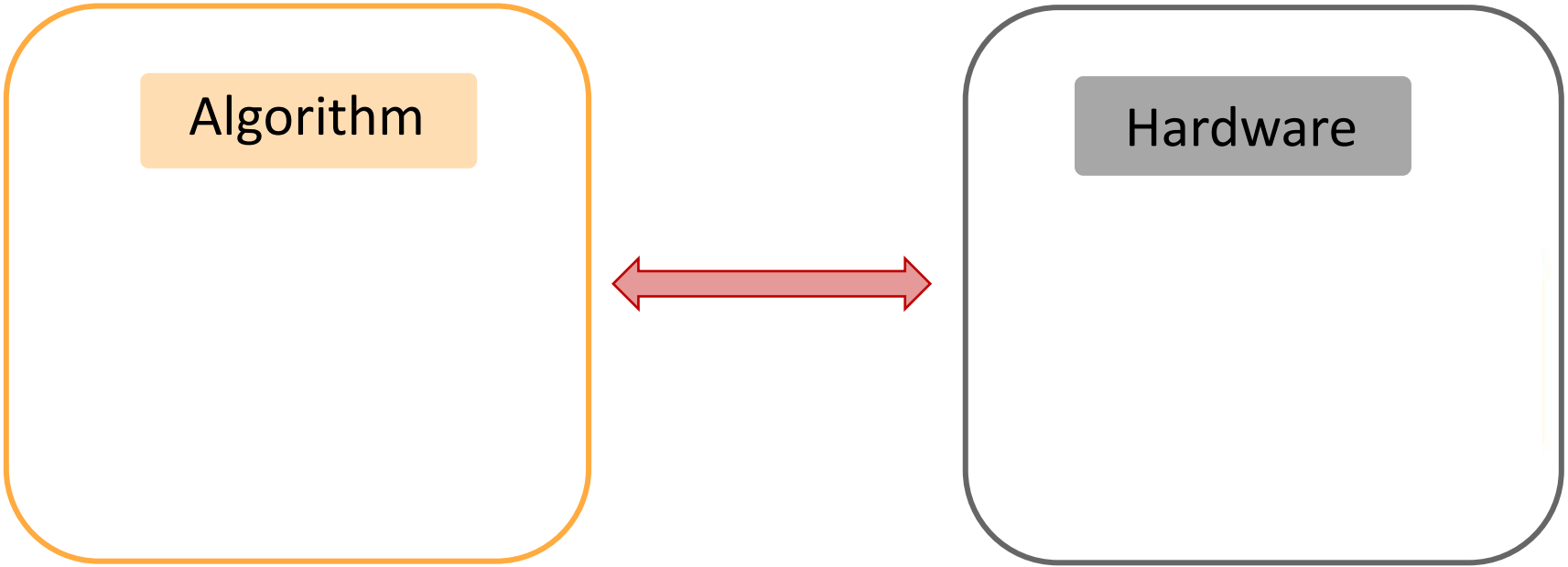
Throughput
(SuperSONIC)

Innovative Algorithms

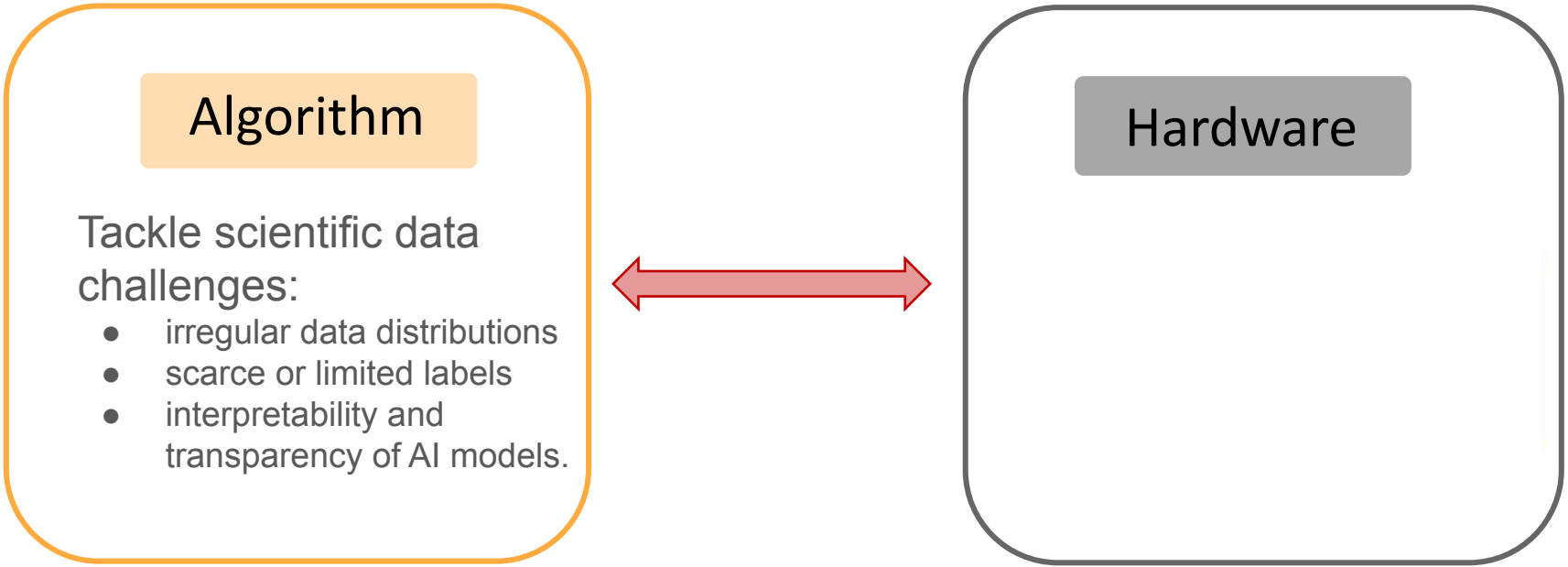
Targeted system (FPGA/ASIC)



Hardware Algorithm Co-design, Automation



Hardware Algorithm Co-design, Automation



Hardware Algorithm Co-design, Automation

Algorithm

Tackle scientific data challenges:

- irregular data distributions
- scarce or limited labels
- interpretability and transparency of AI models.

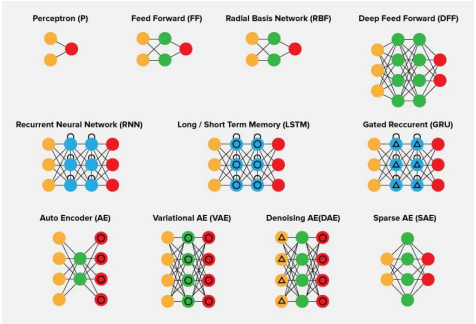
Hardware

Dedicated platforms optimized to

- support these AI algorithms
- addressing the need for low latency, high throughput
- mindful consideration of power and memory constraints.

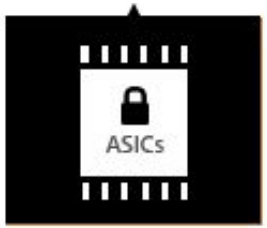
Hardware Algorithm Co-design, Automation

Algorithm



hls 4 ml

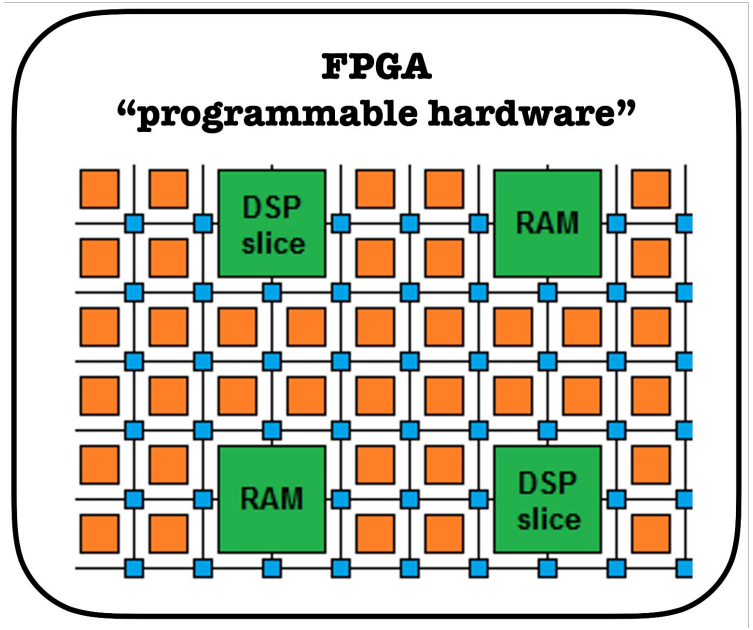
Hardware

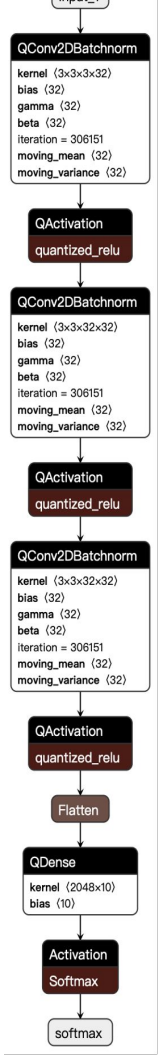


- Enable domain experts to implement and deploy AI algorithms directly on hardware.
- Streamline the conversion from algorithm development to hardware execution.

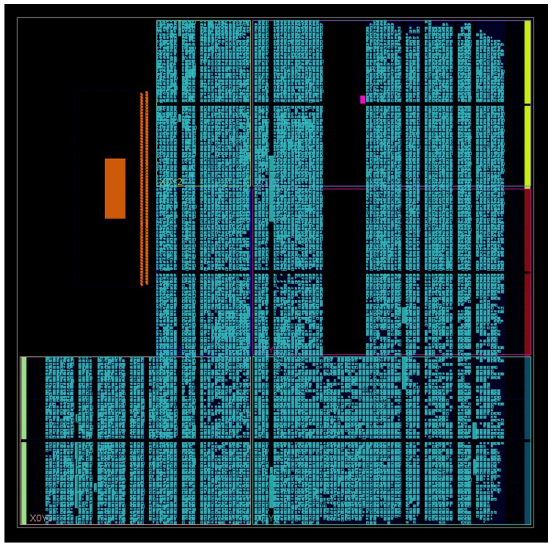
https://github.com/fastmachinelearning/hls4ml

- Pros:
  - Reprogrammable interconnects between embedded components that perform multiplication (DSPs), apply logical functions (LUTs), or store memory (BRAM)
  - High throughput I/O: O(100) optical transceivers running at O(15) Gbps
  - Massively parallel
  - Low power
- Cons:
  - Requires domain knowledge to program (using VHDL/Verilog)





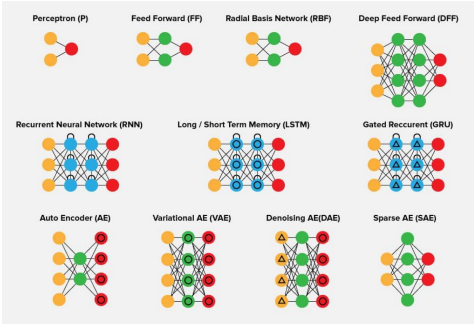
Programming Hardware (FPGAs)



High-Level Synthesis

Hardware Algorithm Co-design, Automation

Algorithm



hls 4 ml

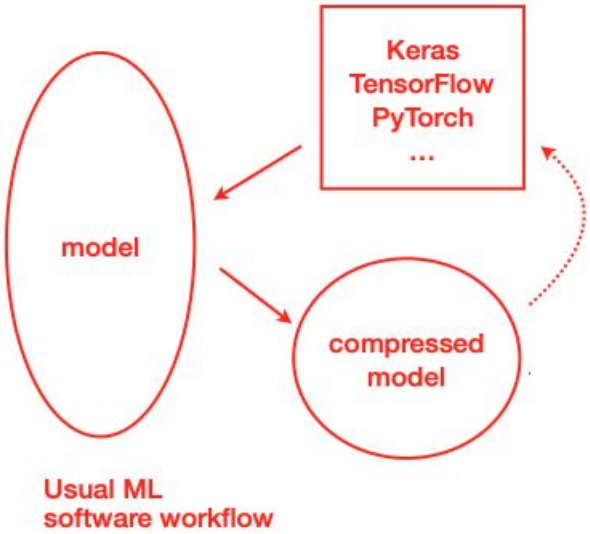
Low latency

Hardware

https://github.com/fastmachinelearning/hls4ml

HLS4ML translating ML into FPGA firmware

HLS4ML translating ML into FPGA firmware



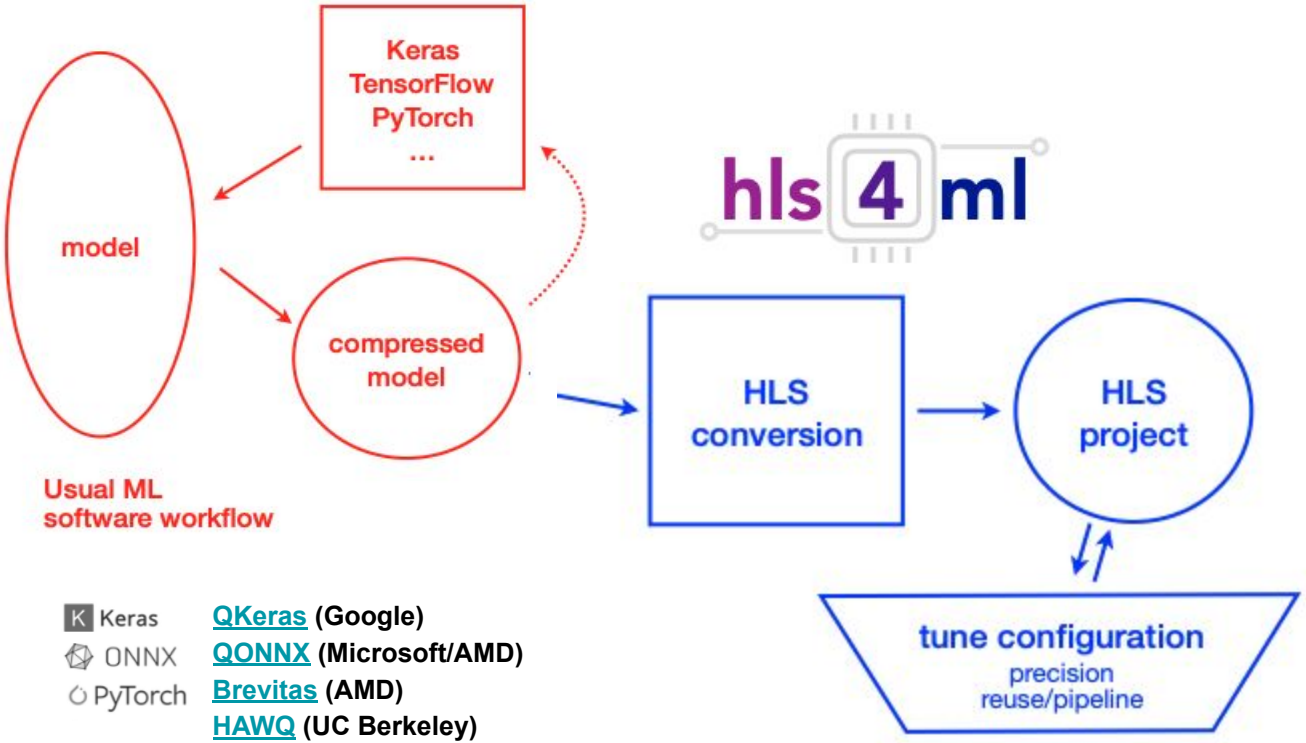
Usual ML
software workflow

hls4ml

Keras ONNX PyTorch

[QKeras](#) (Google)
[QONNX](#) (Microsoft/AMD)
[Brevitas](#) (AMD)
[HAWQ](#) (UC Berkeley)

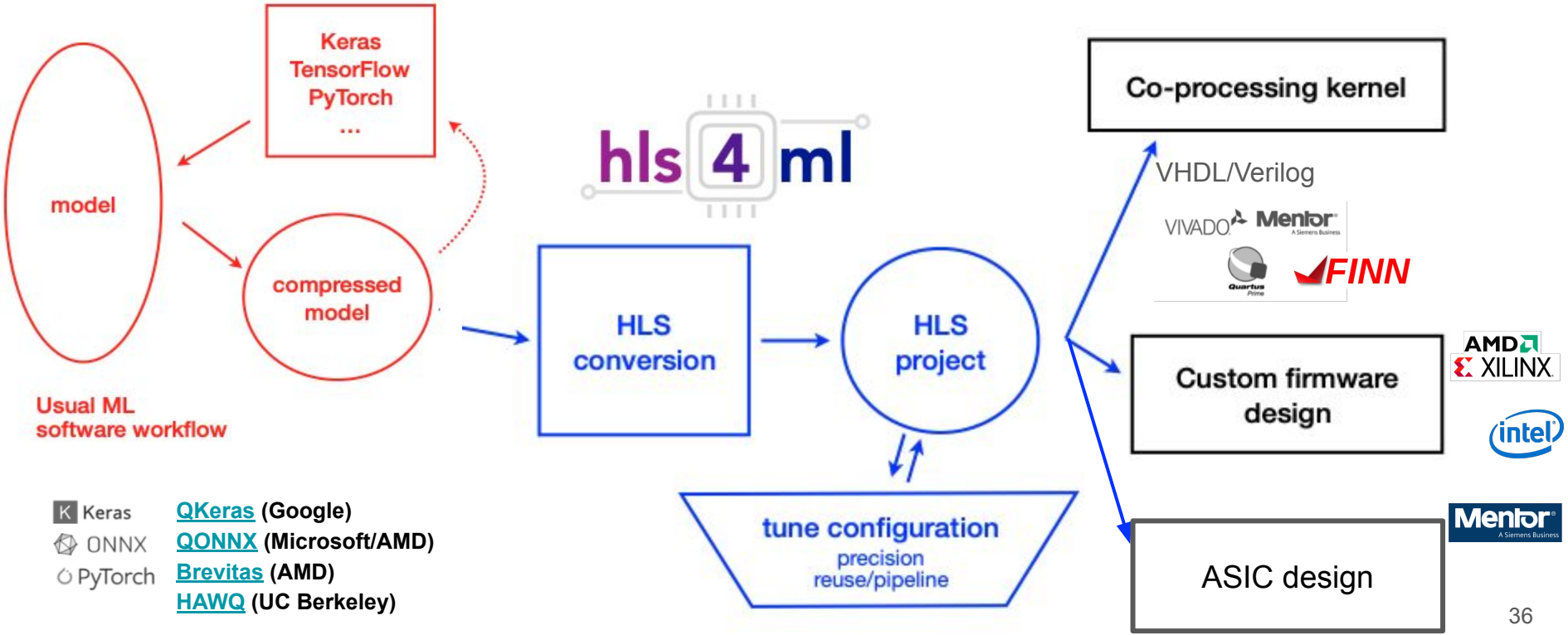
HLS4ML translating ML into FPGA firmware



Keras
ONNX
PyTorch

QKeras (Google)
QONNX (Microsoft/AMD)
Brevitas (AMD)
HAWQ (UC Berkeley)

HLS4ML translating ML into FPGA firmware



How can we accelerate an AI model on FPGA?

Quantization

Xilinx Vivado 2017.2
Clock frequency: 200 MHz
FPGA: Xilinx Kintex Ultrascale
(XCKU115-FLVB2104)

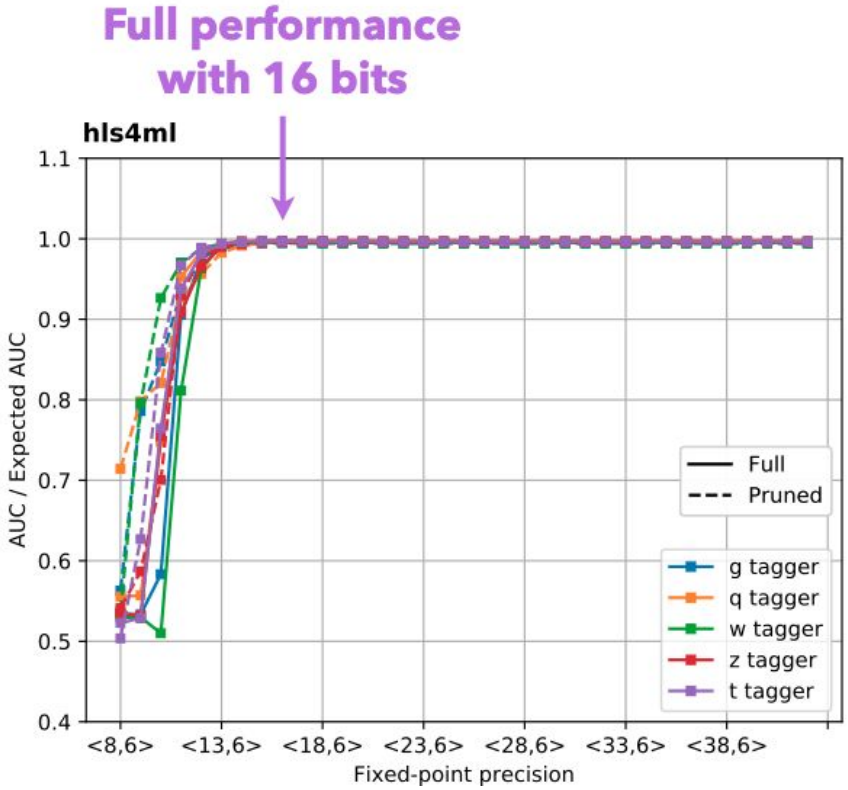
- Scan the bit width until you reach optimal performance

ap_fixed<width,integer>

0101.1011101010

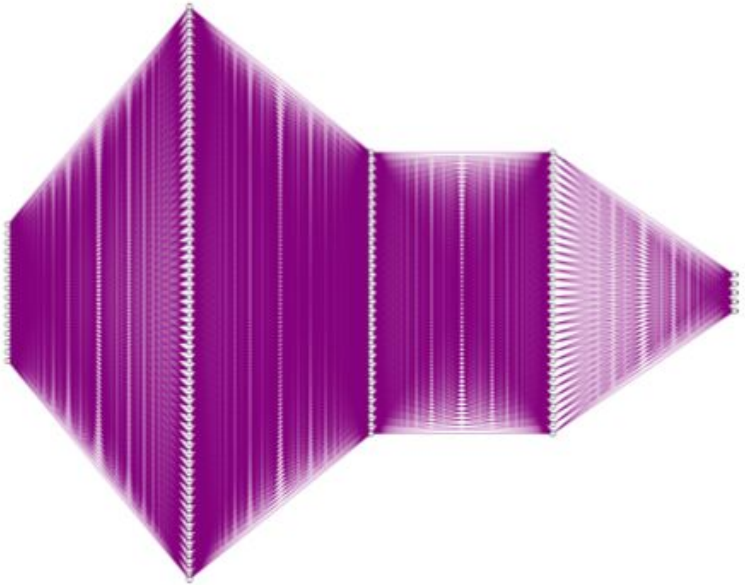
integer | fractional

width



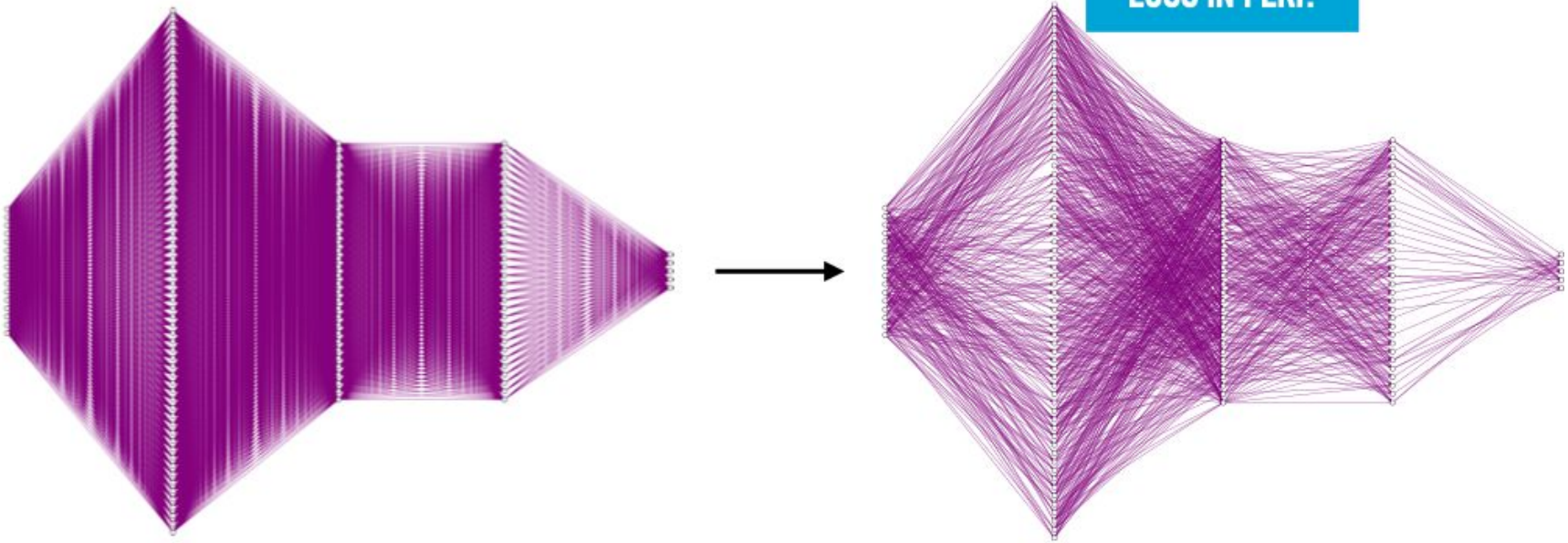
Pruning

- Remove **smallest** weigh
- Iterate

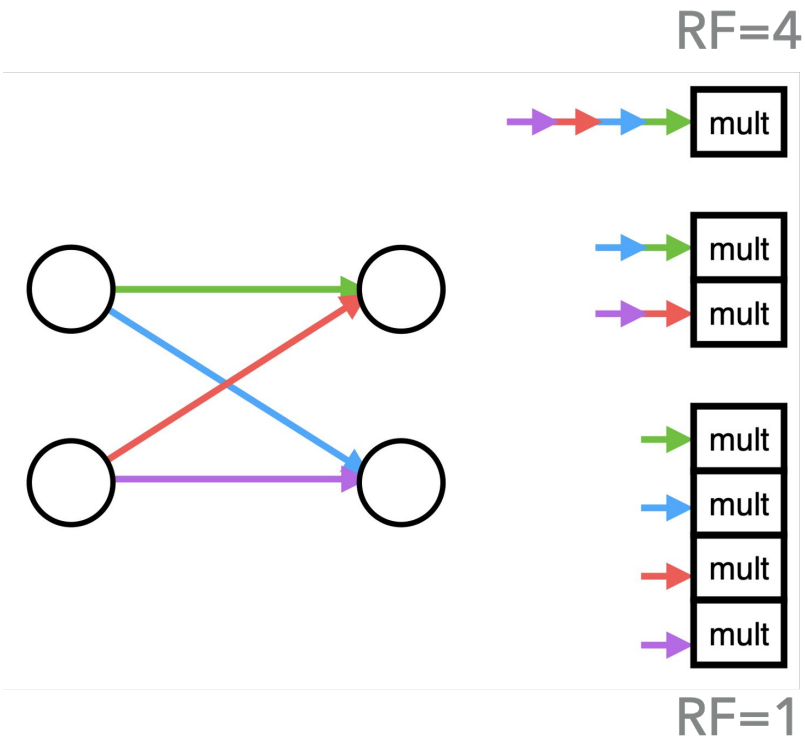


Pruning

- Remove **smallest** weights
- Iterate



Parallelization and Timing

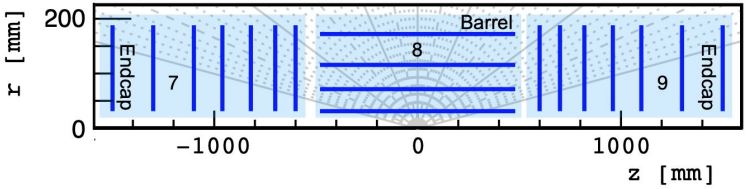


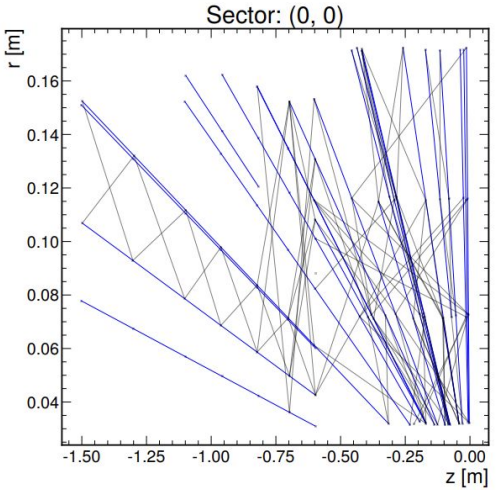
~35 clocks
@ 200 MHz
= 175 ns

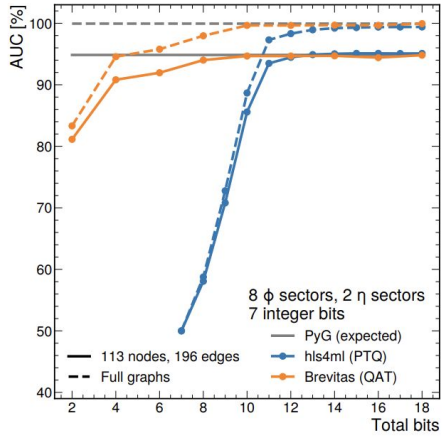
~15 clocks
@ 200 MHz
= 75 ns

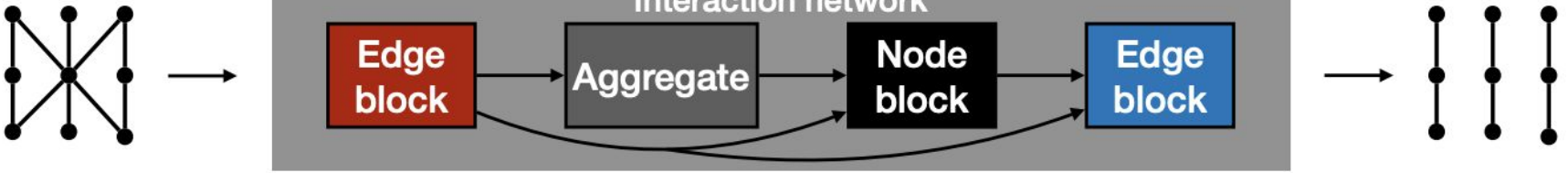
Graph Neural Networks for Tracking on FPGAs

[Front. Big Data, 23 March 2022]



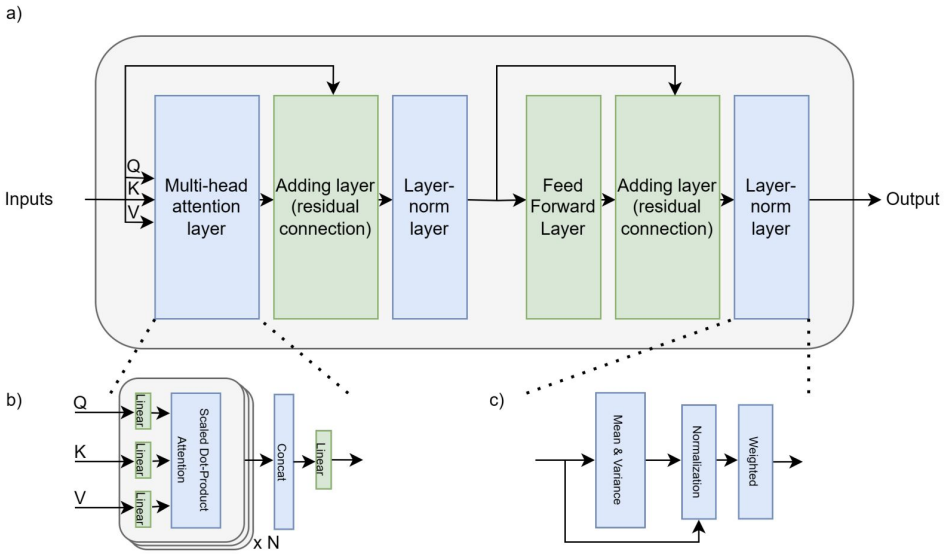


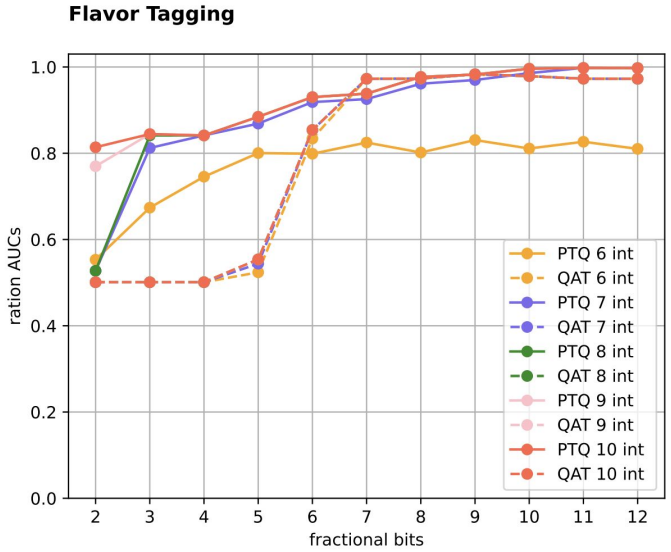




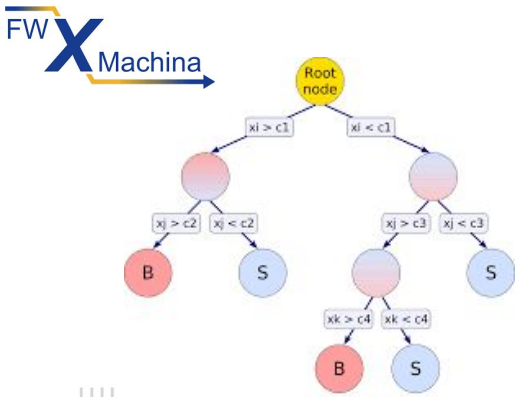
Transformer Inference on FPGAs

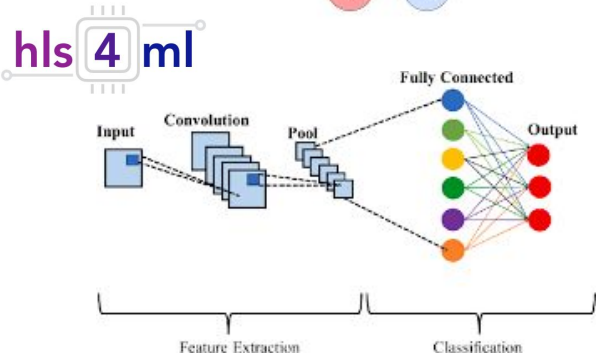
[Zhixing Jiang et al 2025 JINST 20 P04014]



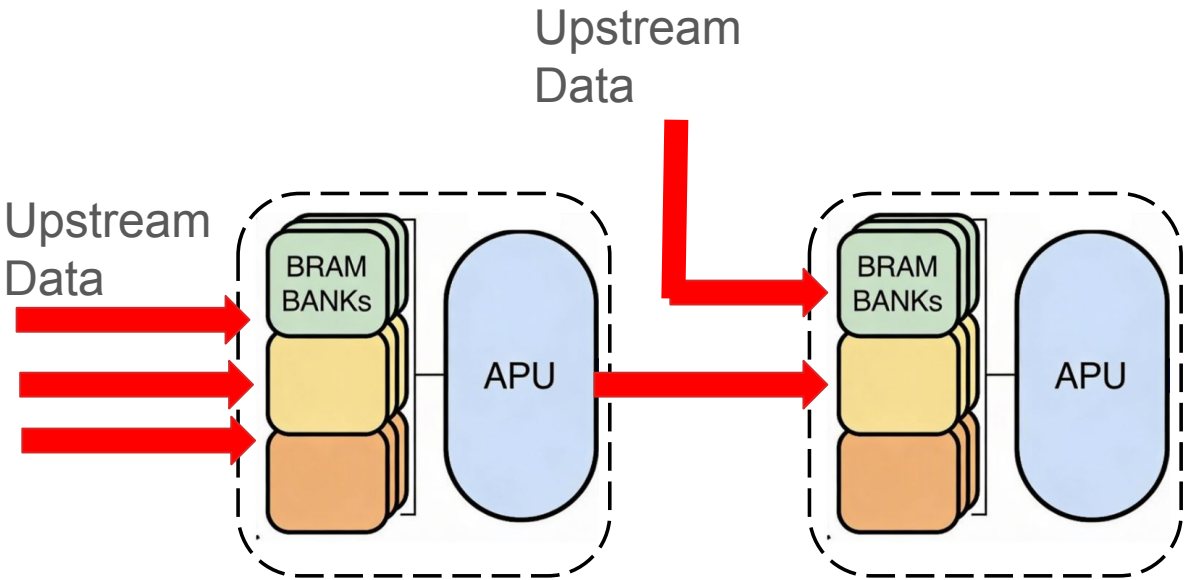


Global Event Processor Evaluation Platform



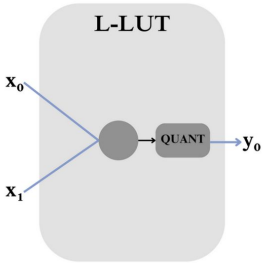


[Zhixing Jiang et al 2024 JINST 19 P05031]



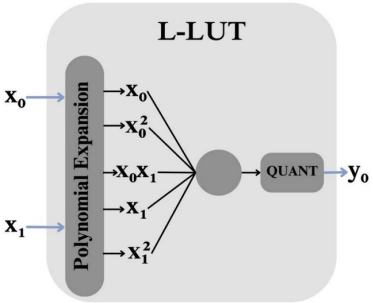
Evolution of LUT-based NNs

Andronic, Cassidy [FastML2025]



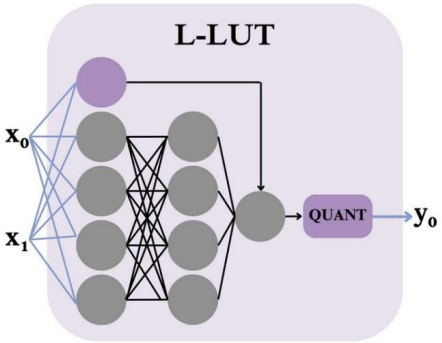
Neuron absorbed in a Logical LUT.

2020



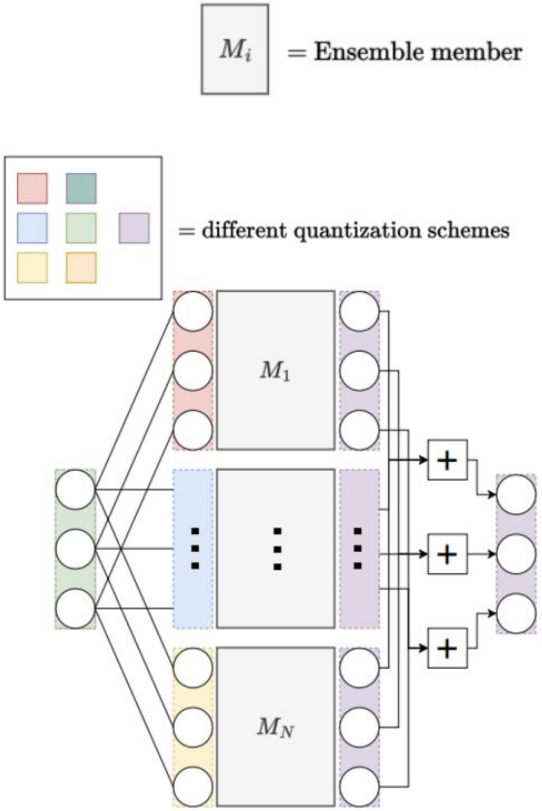
L-LUT hides a multivariate polynomial.

2023



Neuron hides MLP with skip connections.

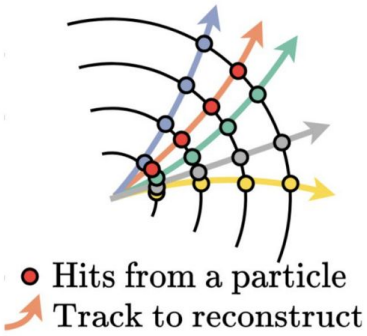
2024

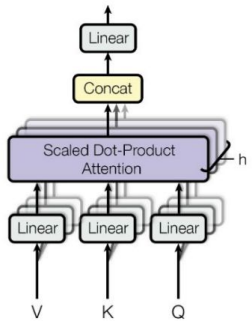


AmigoLUT

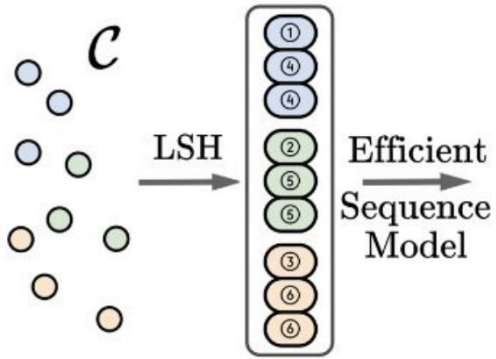
2025

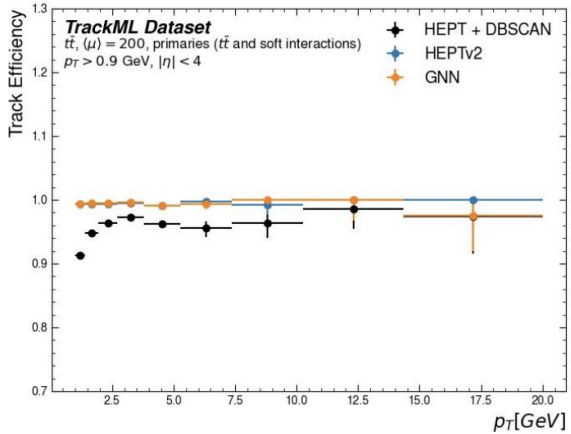
Efficient Transformer Architectures for Point Cloud Data in High Energy Physics





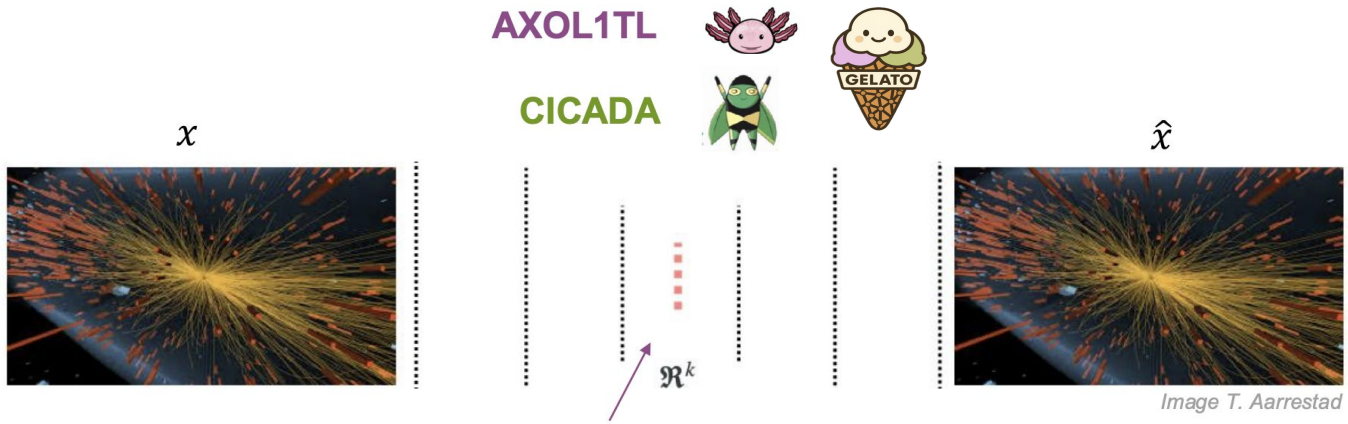
- Charged particles leave hits on multiple detector layers → point cloud, sparse, irregular
- Locality-Sensitive Hashing-Based Efficient Point Transformer (HEPT)





ANOMALY DETECTION AT 40 MHZ

- Anomaly detection algorithm for the trigger based on a variational autoencoder



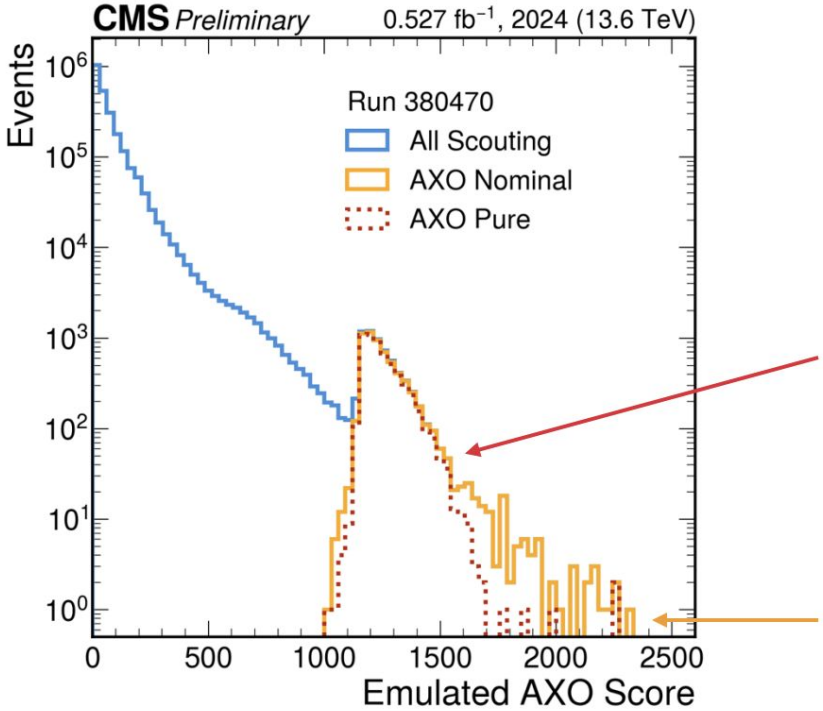
Train on randomly sampled *data*

Bottleneck: autoencoder learns to compress high dimensional inputs into low dimensional latent space

Unsupervised learning: $x - \hat{x}$ represents degree of abnormality

AXOL1TL triggered events

CMS-DP-2024-059

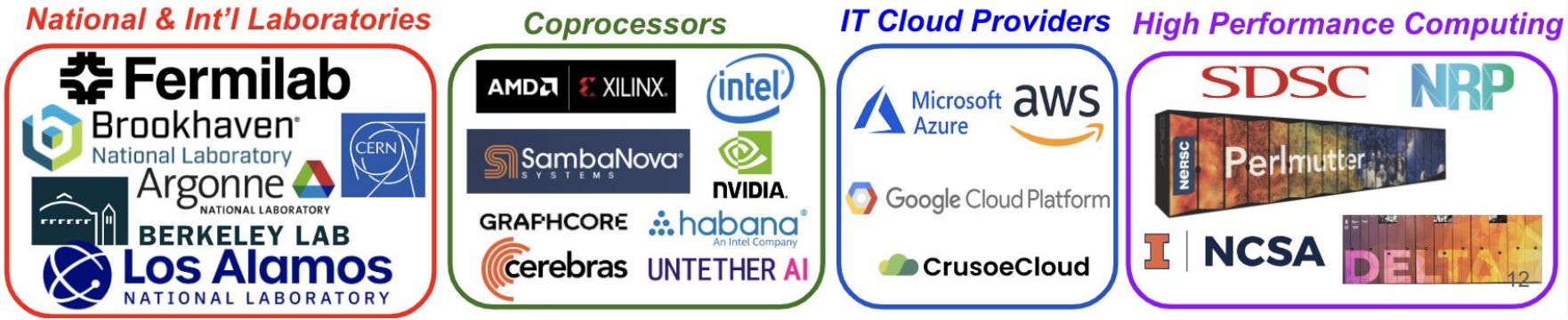


Large fraction of unique events recorded that would otherwise be rejected

High anomaly score events, also triggered by existing L1 trigger

Partnership and FastML Ecosystem

Growing strong industry connections with support through the Fast ML community



A3D3 Postbac

Jada Marshall

Purdue University

Neuroscience

Supervisor: Maria Makin

Kira Nolan

CalTech

MMA

Supervisor: Matthew Graham

Lucie Afko

Duke University

HEP/MMA

Supervisor: Kate Scholberg

Malina Desai

MIT

MMA

Supervisors: Philip Harris, Erik Katsavounidis

Workshop & Conferences

- Summer school

- Workshop

- Conference

Scientific-MOOD FAIR Challenge

2025 FPGA智慧運算與終端節點 創意應用競賽 暨 NSF HDR AI 黑客松競賽

FPGA競賽官網
https://reurl.cc/eMvzAK

NSF HDR AI黑客松官網
https://indico.cern.ch/event/1610056/

Summary

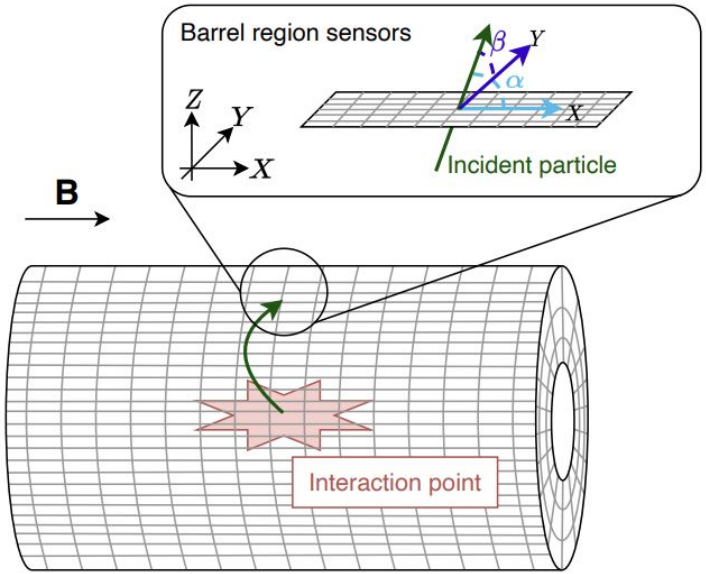
Real-time AI is transforming scientific and engineering discovery.

- **HLS4ML**: Speeds deployment of ML algorithms to FPGA/ASIC, optimizing workflows.
- **Expanding Ecosystem**: A growing community driving innovation across disciplines.

Shih-Chieh Hsu
http://faculty.washington.edu/schsu/
schsu@uw.edu

On-chip probabilistic inference for charged-particle tracking at the sensor edge

Arghya Ranjan Das et al arXiv:2602.15946v2



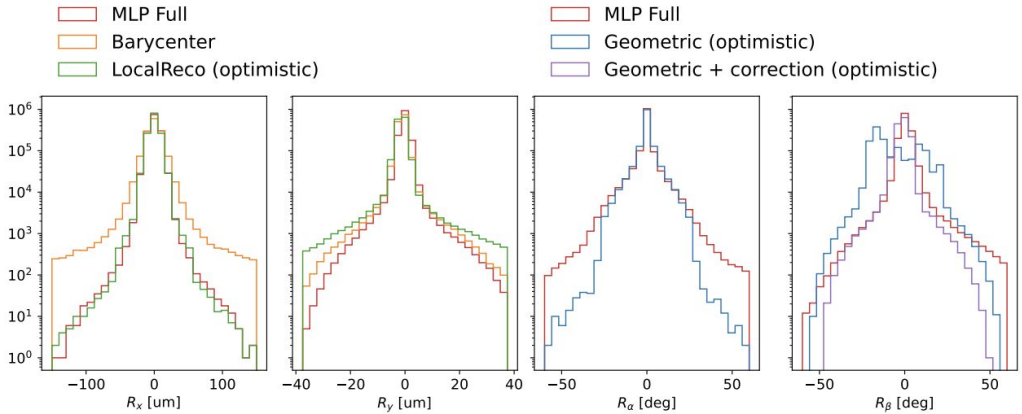
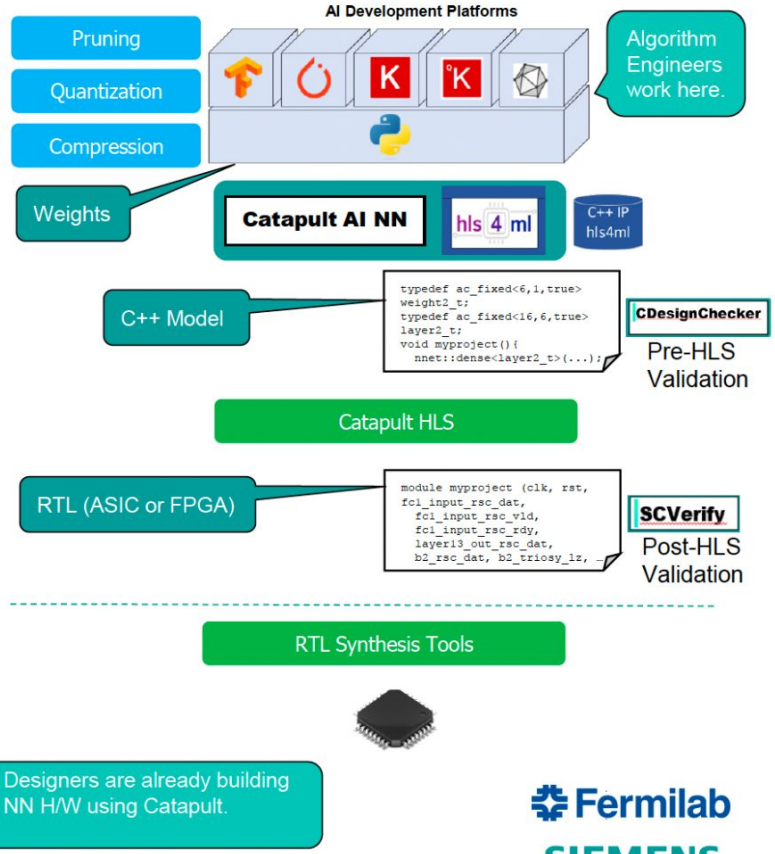


Figure 4: Comparison of the residuals for x, y, α, and β between the Full MLP model and non-ML algorithms. The non-ML algorithms labeled as "optimistic" incorporate information that is not directly available on the ASIC.

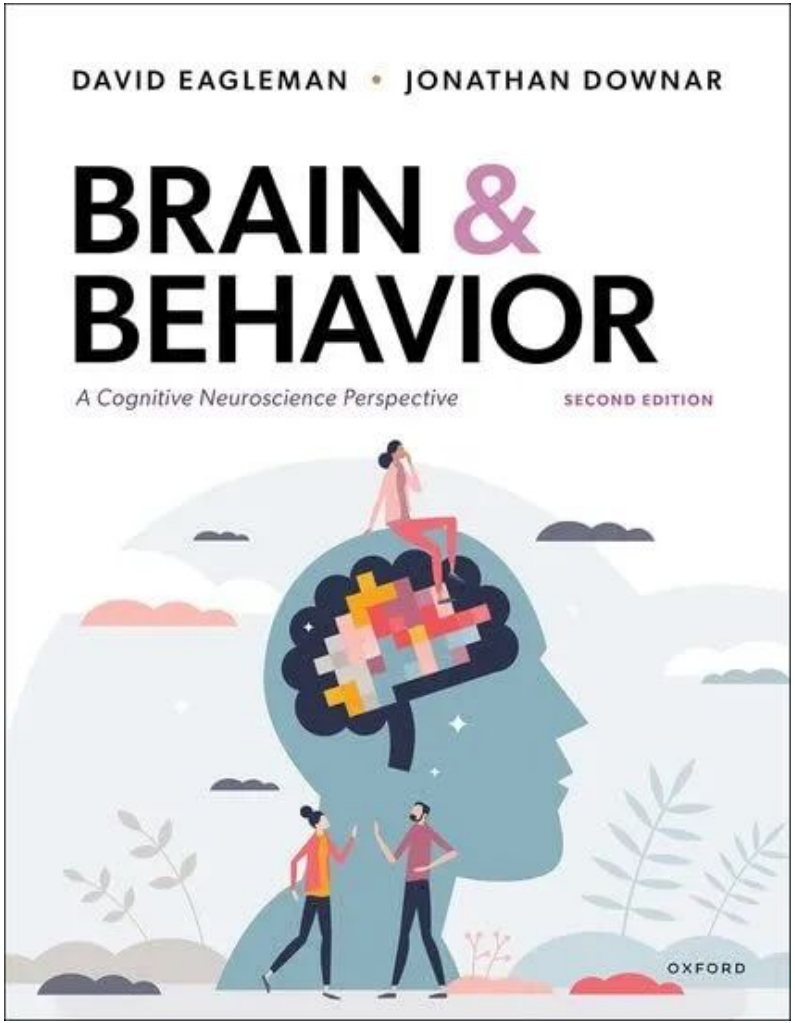
What is HLS4ML?

- A Python package that:
  - Generates a top-level C++ design leveraging a library of C++ machine learning (ML) functions
  - The generated C++ code is a dataflow pipeline of hardened layers
  - Compiles the synthesis-ready C++ design into register-transfer level (RTL) code using High-level synthesis (HLS) tools (FPGA or ASIC)
    - Provides coarse-grained “reuse_factor” for latency/resource control to explore hardware implementations
    - Provides fully parallel or streamed input/output

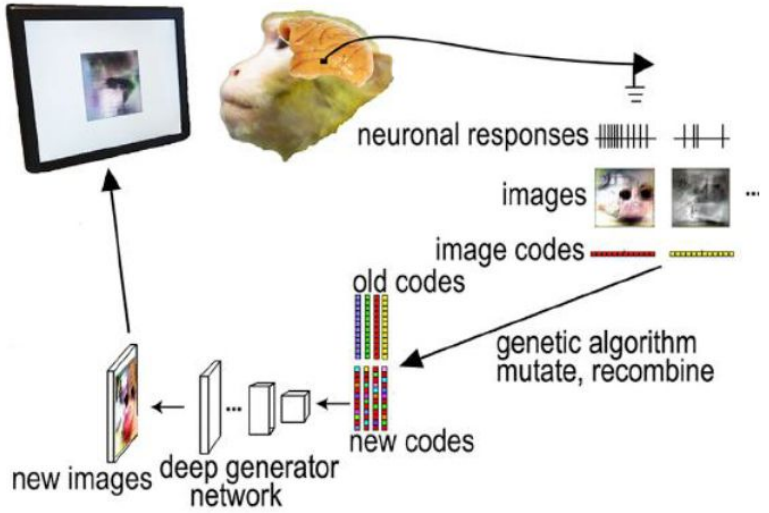


Catapult allows designers to target ASIC platforms

Fermilab
SIEMENS

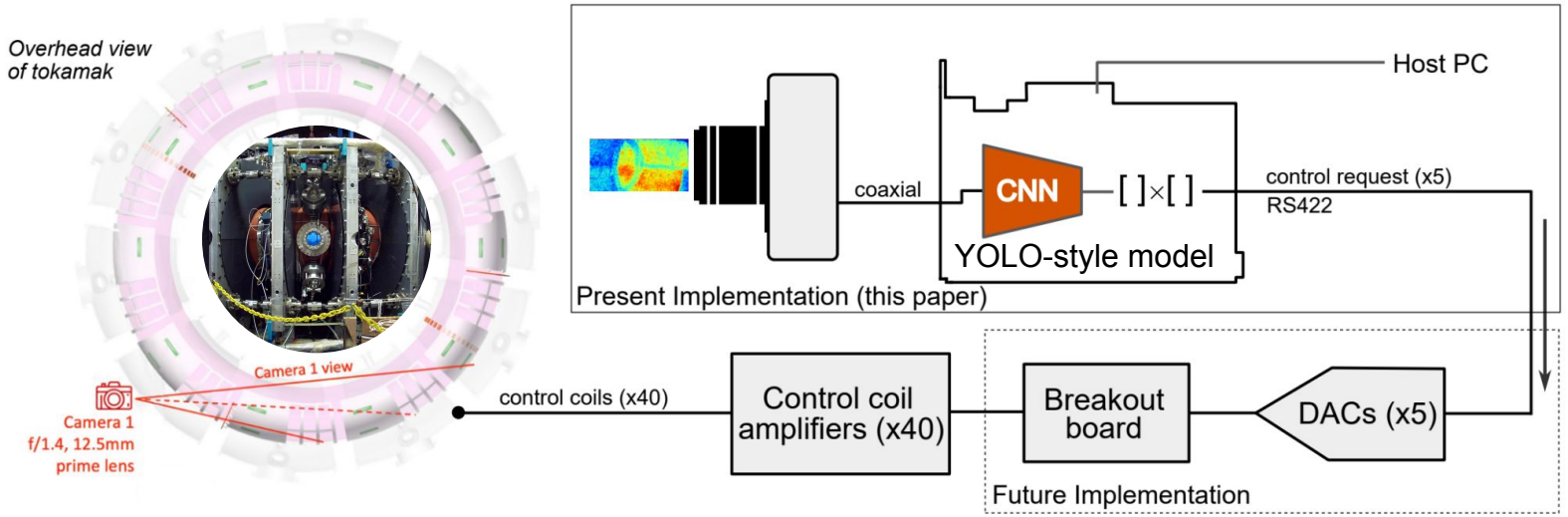


Change stimuli based on brain activity→
map responses quickly



Ponce et al., *Cell* 2019

Fusion: in-situ inference on frame grabber FPGAs in high-speed imaging

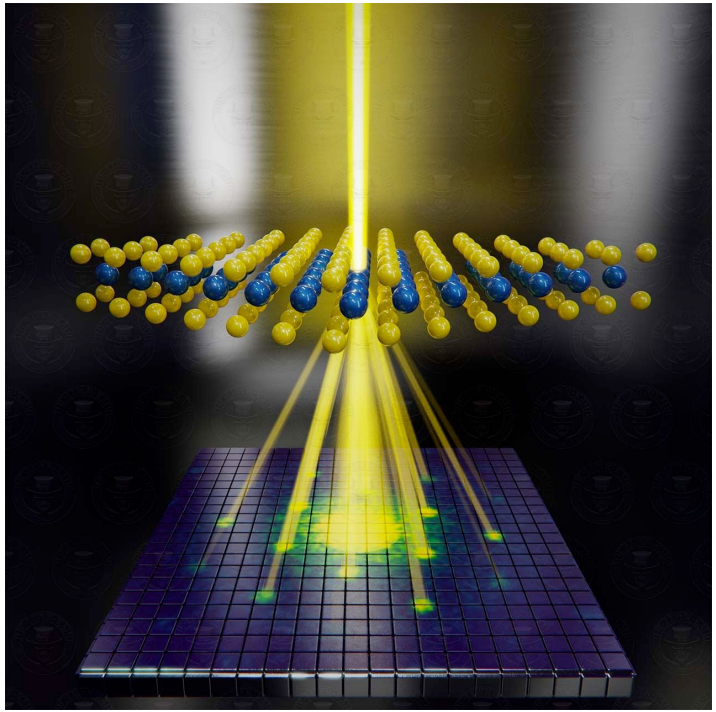


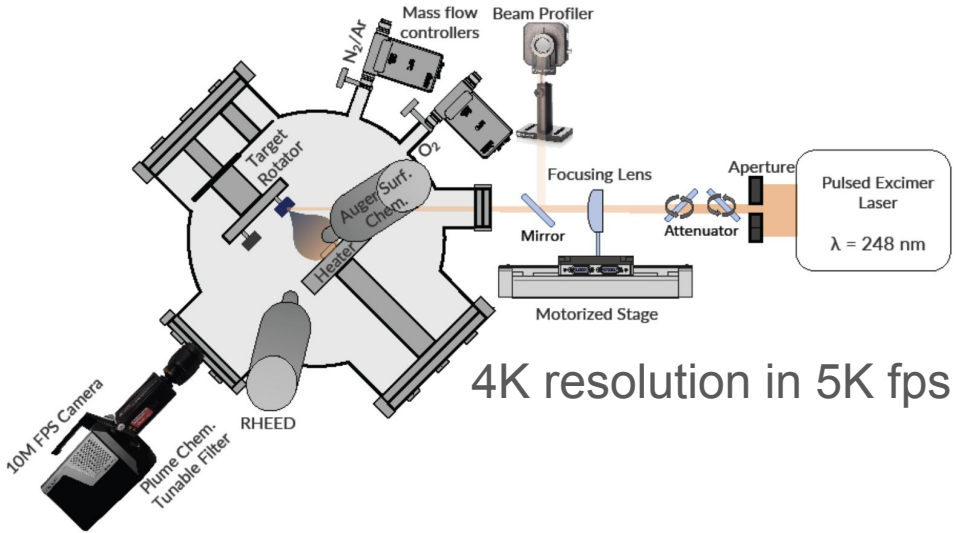
Fusion magneto hydrodynamics and crystal structure detection

Rev. Sci. Instrum. 95, 073509 (2024)

4D Scanning Transmission Electron Microscopy

Agar_FastML_talk

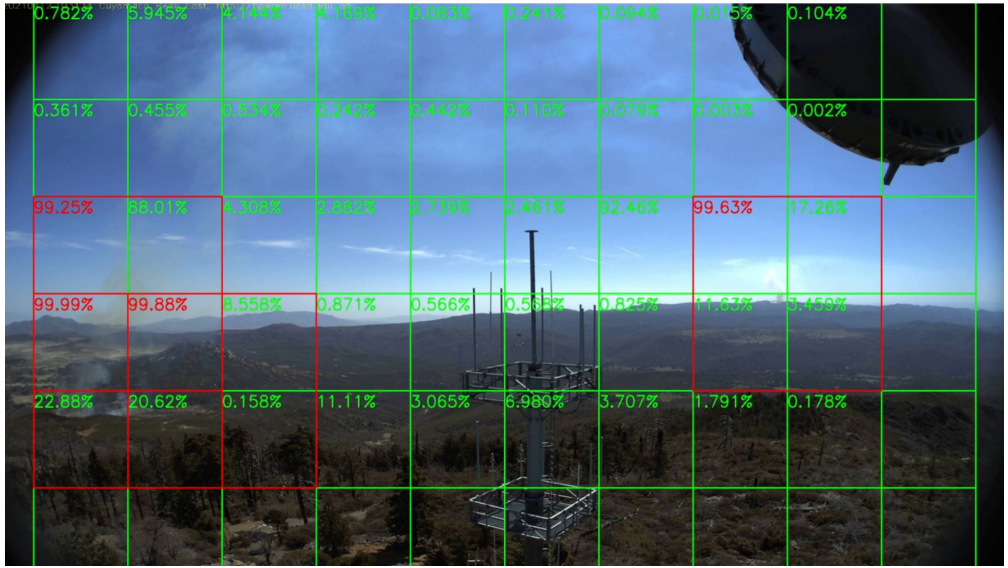




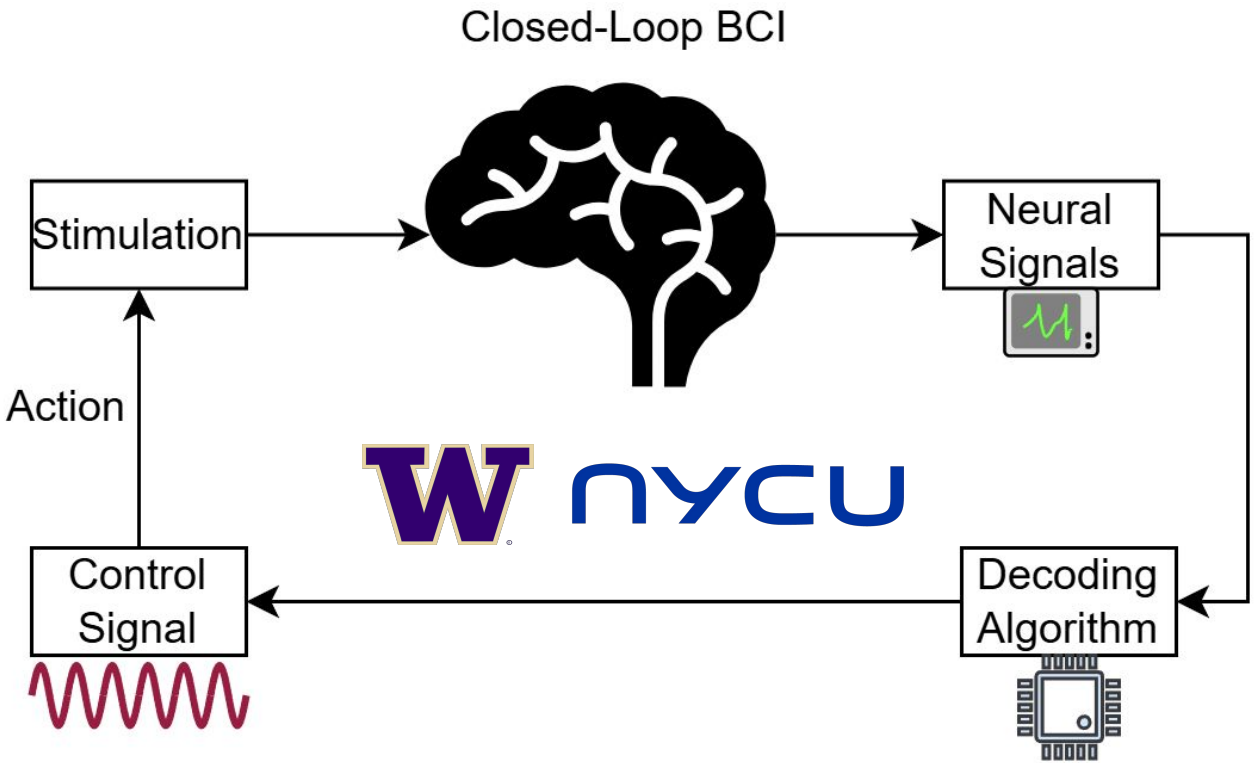
4K resolution in 5K fps

Wildfire: Enabling Embedded Systems and IoT

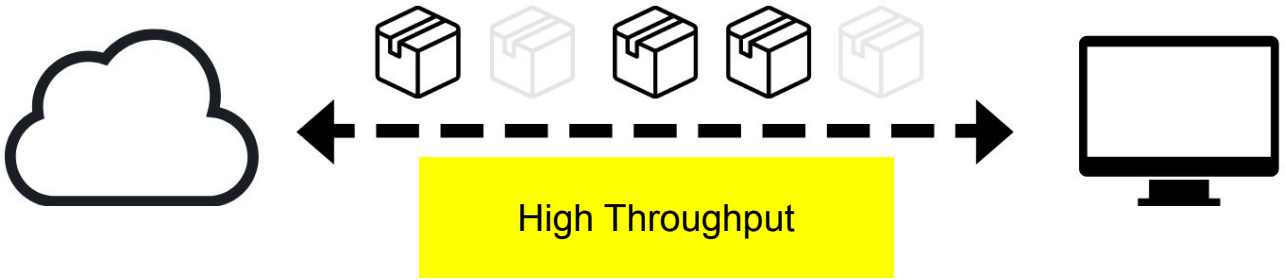
Utilization of hls4ml to integrate AI detection models (trained with PyTorch and TensorFlow-Keras) onto FPGAs, demonstrating the effectiveness of this approach for rapid and efficient wildfire detection and prevention strategies.



Closed-Loop Brian Stimulation

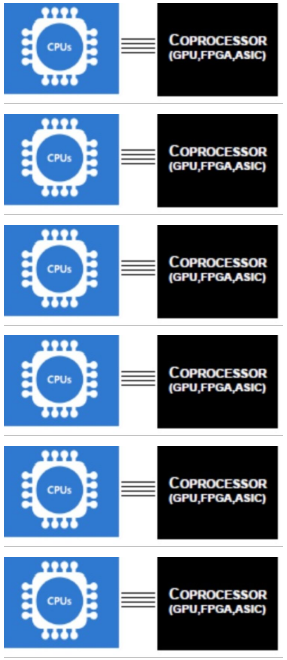


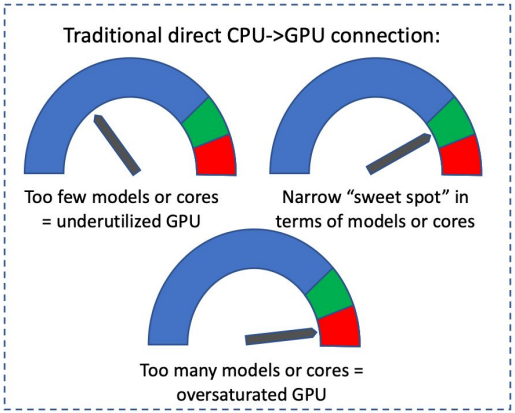
Heterogeneous System



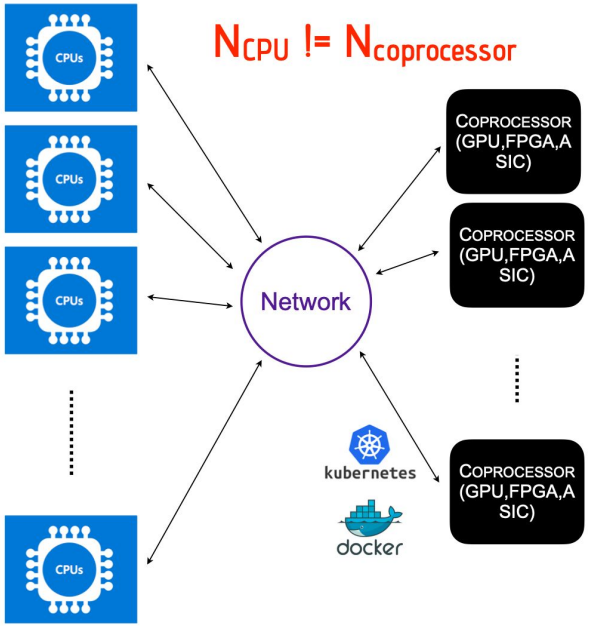
Coprocessor Integration Approaches

1. Direct connection
Inflexible, expensive





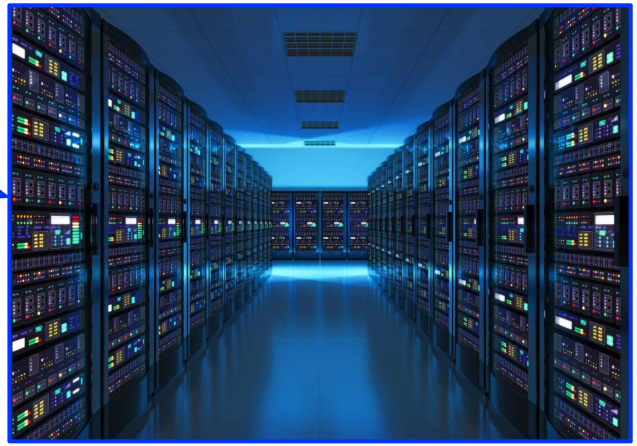
2. "As a service"
Complex, requires R&D



Not so different from ChatGPT!

What to do in Kaohsiung in Summer 2025?

Let’s go dragon boat racing!

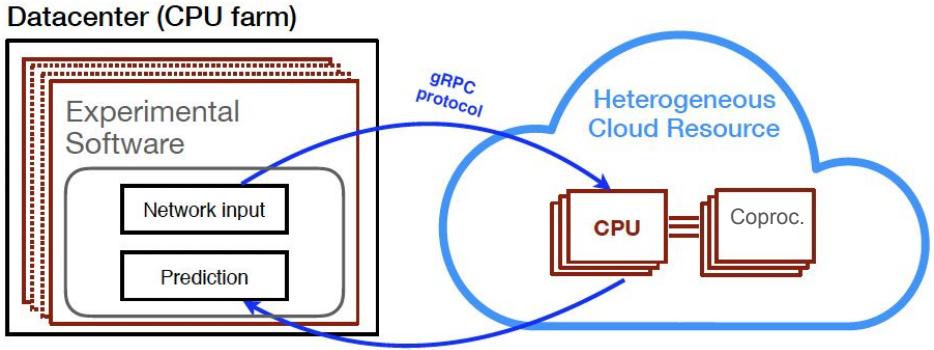


A computing paradigm adaptive to changing hardware landscape

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

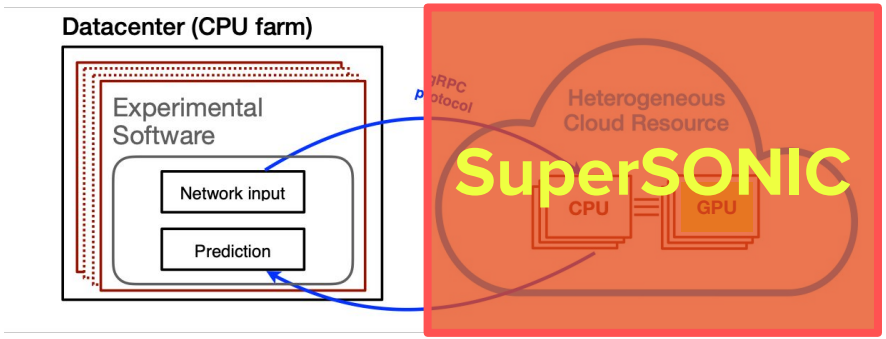
"Inference-as-a-Service":

- Client–server architecture
- Data sent to GPU/FPGA server via gRPC
- Results returned synchronously or asynchronously



SuperSONIC

- SuperSONIC implements SONIC server infrastructure as a portable package based on open-source industry-standard tools.
- GitHub: https://github.com/fastmachinelearning/SuperSONIC



Client workflows vary between experiments, but server feature needs are largely similar.

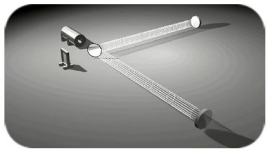
Many applications can benefit from a common server implementation.

Gravitational Wave Data workflow for Alerts

Multi-messenger Astrophysics

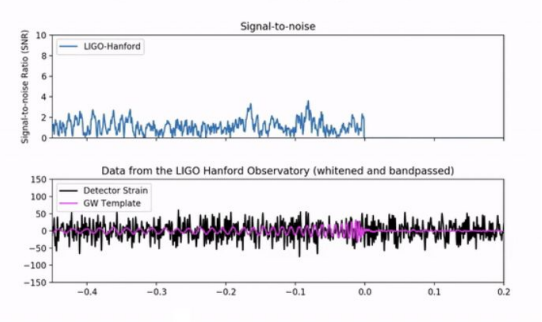
Rapidly identify all signals from multi-messenger events

DETECTOR CHARACTERISATION

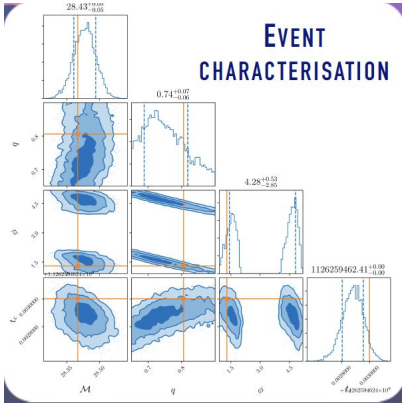


USE INFO FROM WITNESS SENSORS TO PERFORM DATA DE-NOISING

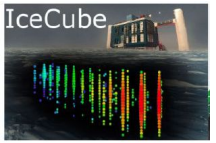
EVENT DETECTION

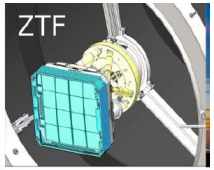


EVENT CHARACTERISATION

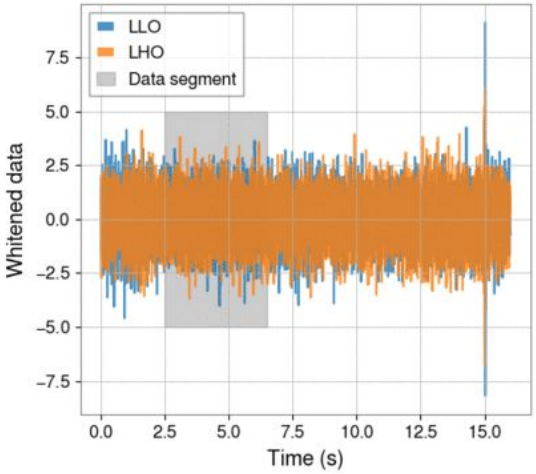


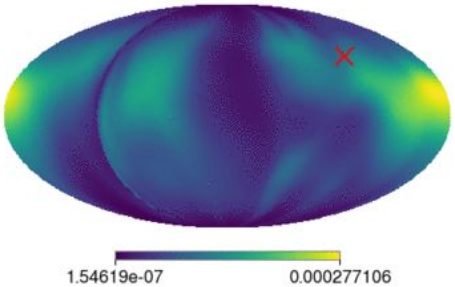
ALERT





MMA: Aframe + AMPLFI deployed Summer 2025!





$$p(\Theta|\mathbf{d}) \propto p(\mathbf{d}|\Theta)\, p(\Theta)$$

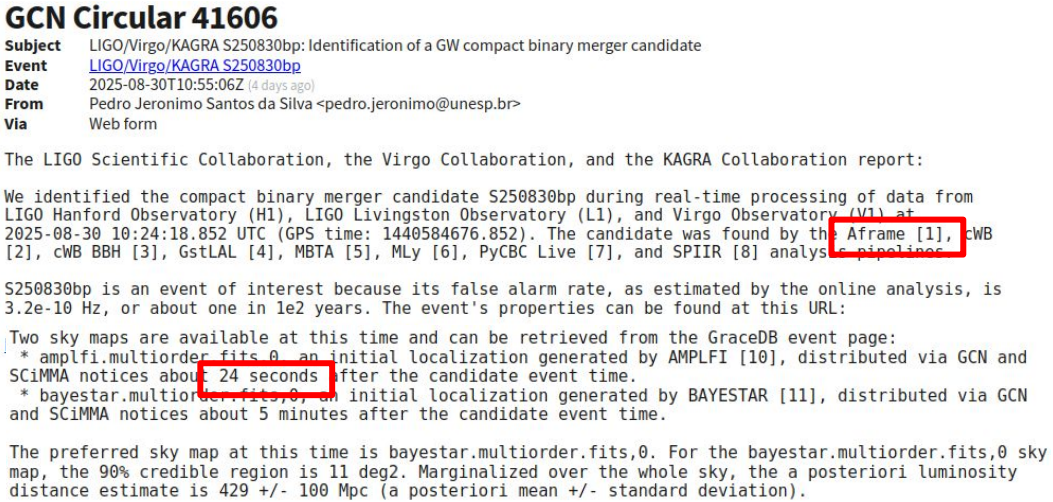
Posterior Likelihood Prior

NN approximator

$$\{\boldsymbol{\Theta}_i, \mathbf{d}_i\}$$

In Likelihood-free Inference:
Learn the distribution from simulations

MMA: Aframe + AMPLFI deployed Summer 2025!



GCN Circular 41606

Subject LIGO/Virgo/KAGRA S250830bp: Identification of a GW compact binary merger candidate
Event LIGO/Virgo/KAGRA S250830bp
Date 2025-08-30T10:55:06Z (4 days ago)
From Pedro Jeronimo Santos da Silva <pedro.jeronimo@unesp.br>
Via Web form

The LIGO Scientific Collaboration, the Virgo Collaboration, and the KAGRA Collaboration report:

We identified the compact binary merger candidate S250830bp during real-time processing of data from LIGO Hanford Observatory (H1), LIGO Livingston Observatory (L1), and Virgo Observatory (V1) at 2025-08-30 10:24:18.852 UTC (GPS time: 1440584676.852). The candidate was found by the Aframe [1], cWB [2], cWB BBH [3], GstLAL [4], MBTA [5], MLy [6], PyCBC Live [7], and SPIIR [8] analysis pipelines.

S250830bp is an event of interest because its false alarm rate, as estimated by the online analysis, is 3.2e-10 Hz, or about one in 1e2 years. The event's properties can be found at this URL:

Two sky maps are available at this time and can be retrieved from the GraceDB event page:
 * amplfi.multiorder.fits,0, an initial localization generated by AMPLFI [10], distributed via GCN and SCiMMA notices about 24 seconds after the candidate event time.
 * bayestar.multiorder.fits,0, an initial localization generated by BAYESTAR [11], distributed via GCN and SCiMMA notices about 5 minutes after the candidate event time.

The preferred sky map at this time is bayestar.multiorder.fits,0. For the bayestar.multiorder.fits,0 sky map, the 90% credible region is 11 deg2. Marginalized over the whole sky, the a posteriori luminosity distance estimate is 429 +/- 100 Mpc (a posteriori mean +/- standard deviation).

- Made the first Compact Binary Coalescence discovery using NNs
  - S250904cv, S250904br, S250904ae, S25
    S250830bp, S250830m
  - First alert from Aframe.