Accelerating Artificial Intelligence for High Energy Physics

Shih-Chieh Hsu (徐士傑)
University of Washington

PHY-2117997

FTCF 2024 (https://indico.pnp.ustc.edu.cn/event/91/)
University of Science and Technology of China
Jan 17 2024

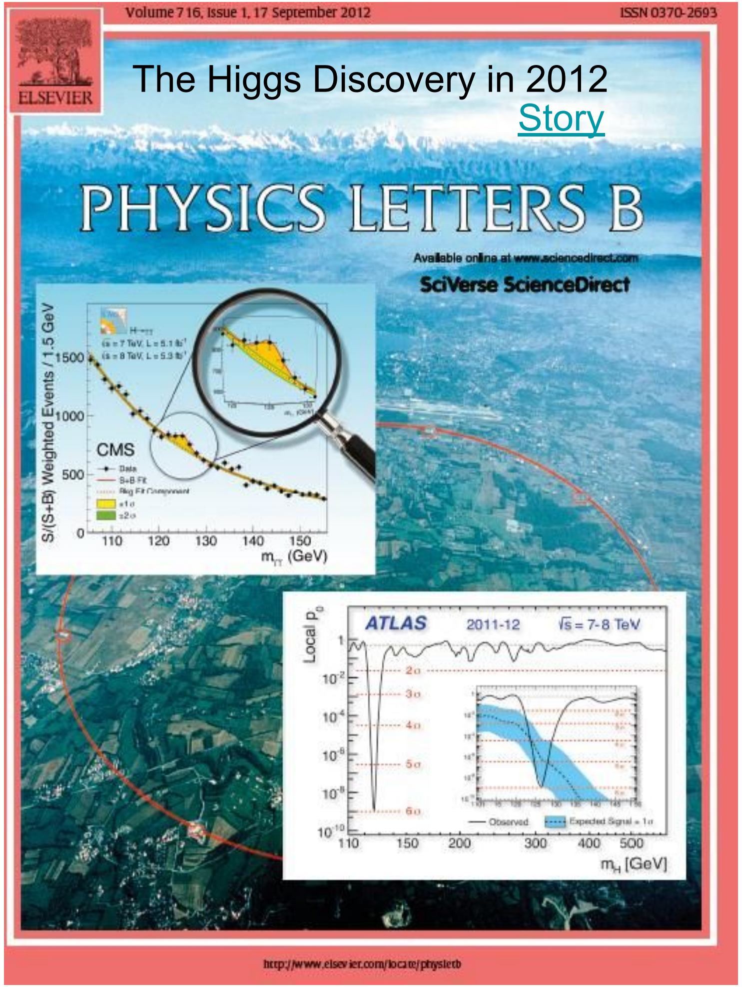
https://a3d3.ai/

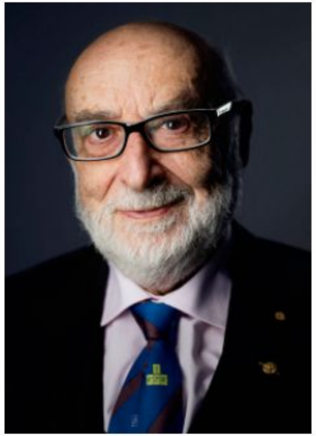
P5 Report (Draft Dec 2023)

https://www.usparticlephysics.org/2023-p5-report/

Investing in the scientific workforce and enhancing computational and technological infrastructure is crucial. To achieve this goal, funding agencies should support programs that foster a supportive, collaborative work environment; help recruit and retain diverse talent; and reinforce professional standards. Targeted increases in support for theory, general accelerator R&D (GARD), instrumentation, and computing will bolster areas where US leadership has begun to erode. These areas align with national initiatives in **artificial intelligence and machine learning (AI/ML)**, quantum information science (QIS), and microelectronics, creating valuable synergies. Such increased support maximizes the return on scientific investments, fosters innovation, and benefits society in domains from medicine to national security.

The Higgs Discovery in 2012
Story





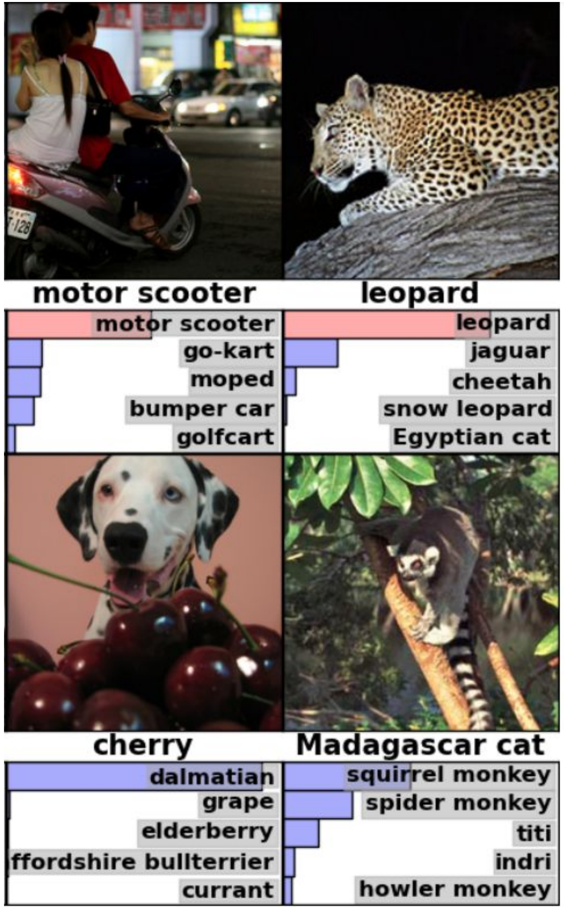
© Nobel Media AB. Photo: A. Mahmoud

François Englert

© Nobel Media AB. Photo: A. Mahmoud

Peter W. Higgs

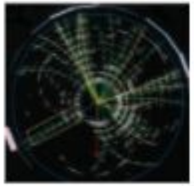
2013



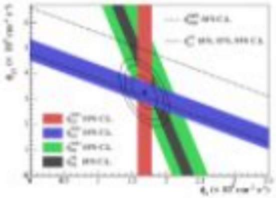
AlexNet Comm. ACM. 60 (6): 84–90

[2012: A Breakthrough Year for Deep Learning]

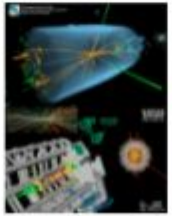
ACM 2018 Turing Award



Top discovery
1995



Neutrino Oscillations
2001



Higgs Discovery
2012

1995
Support Vector Machine

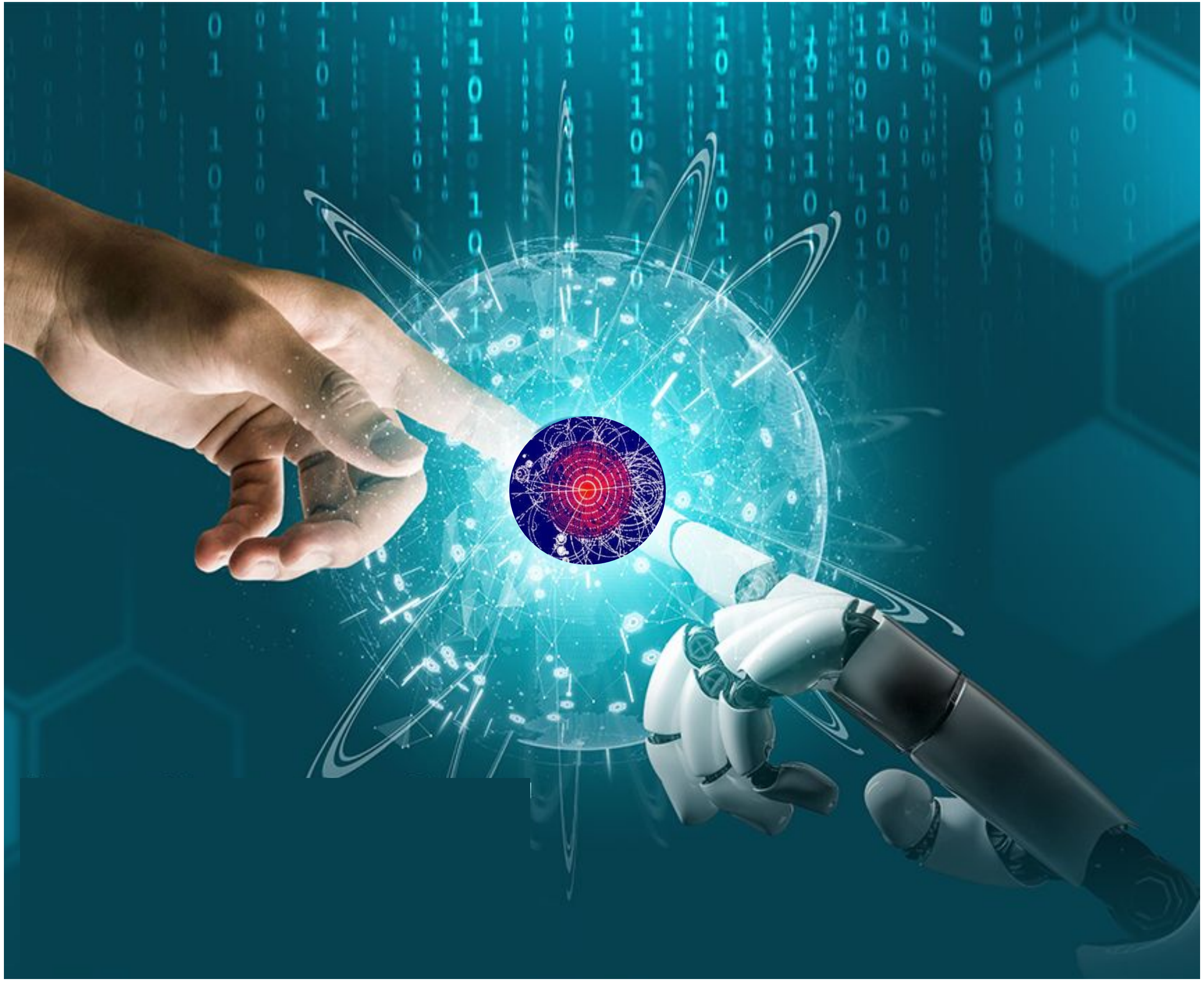
2001
Gradient Boosting

2012
AexNet

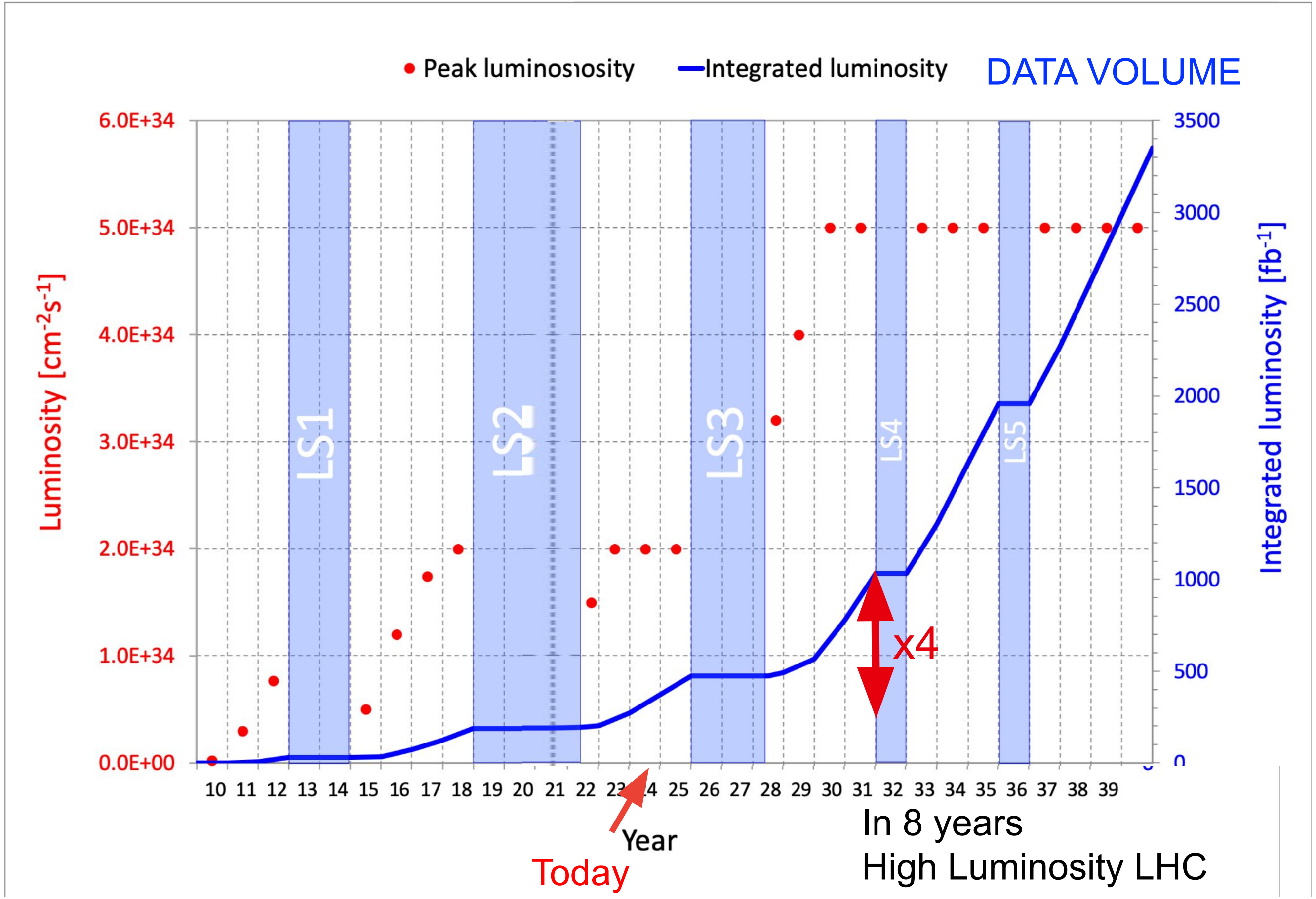
2016
AlphaGo

2022
ChatGPT

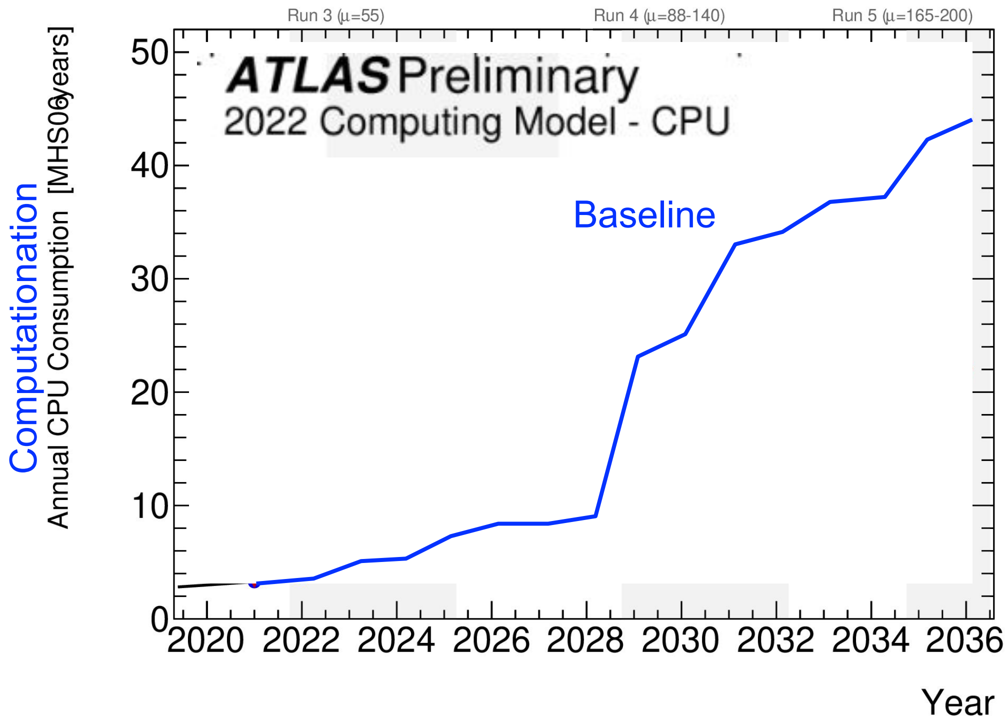
Deep Learning revolution



Credit: Onpassive

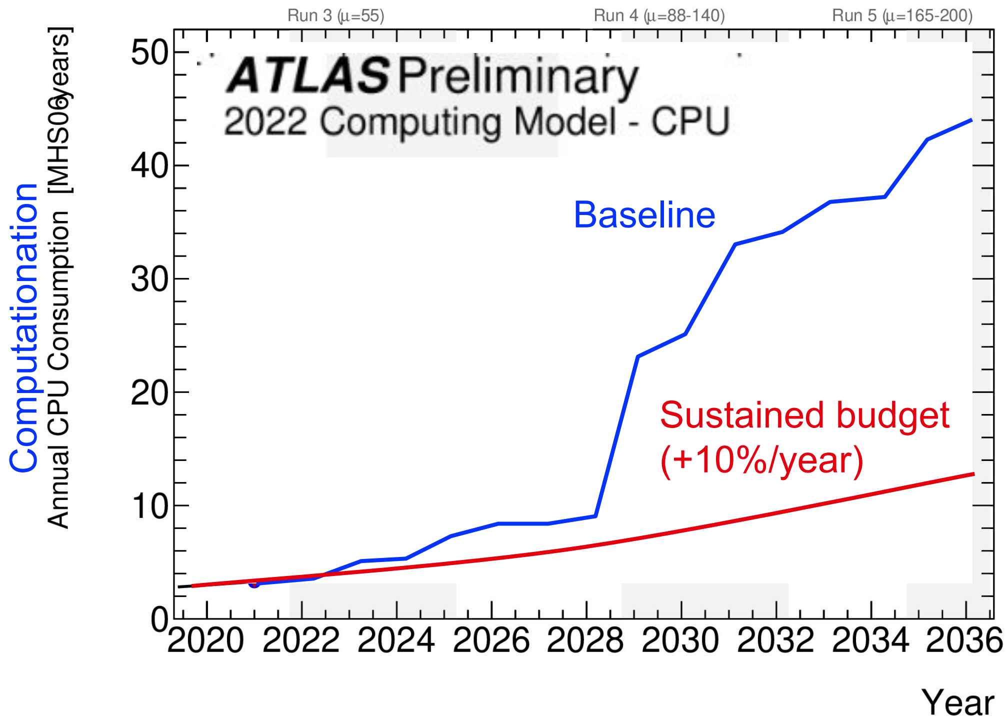

Peak luminosiosity
Integrated luminosity
DATA VOLUME
Luminosity [cm-2s-1]
Integrated luminosity [fb-1]
6.0E+34
5.0E+34
4.0E+34
3.0E+34
2.0E+34
1.0E+34
0.0E+00
3500
3000
2500
2000
1500
1000
500
0
LS1
LS2
LS3
LS4
LS5
x4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Year
Today
In 8 years
High Luminosity LHC

Critical computing challenge



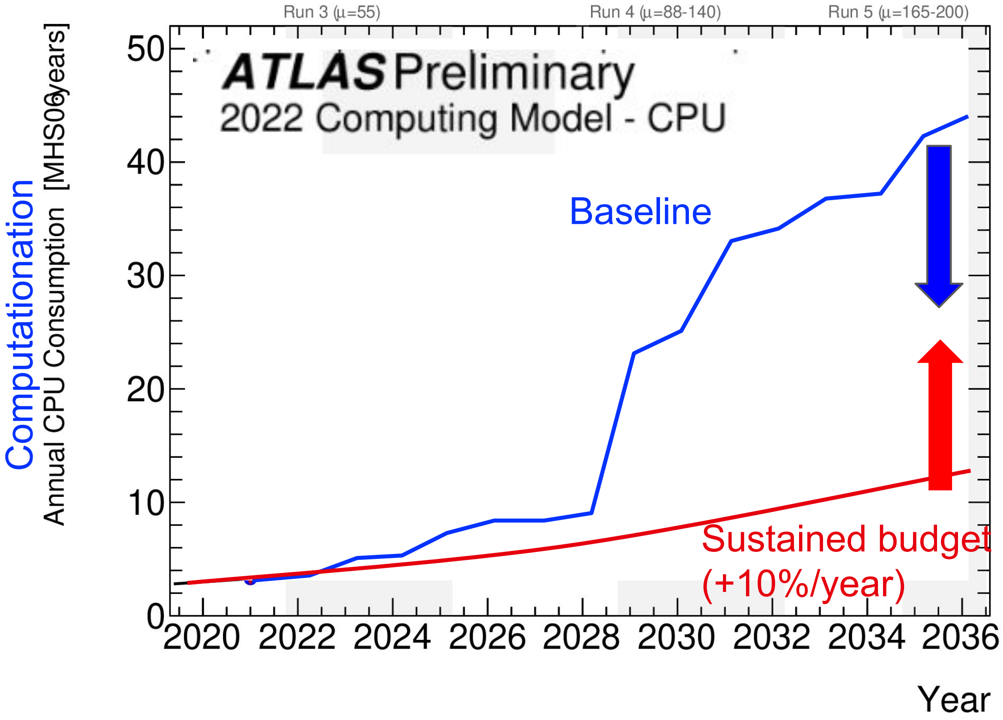
- To preserve current physics we are upgrading the system
  - Our event size will have to be 10x larger
  - We will have to take data at 4 times the current rate

Critical computing challenge



- To preserve current physics we are upgrading the system
  - Our event size will have to be 10x larger
  - We will have to take data at 4 times the current rate

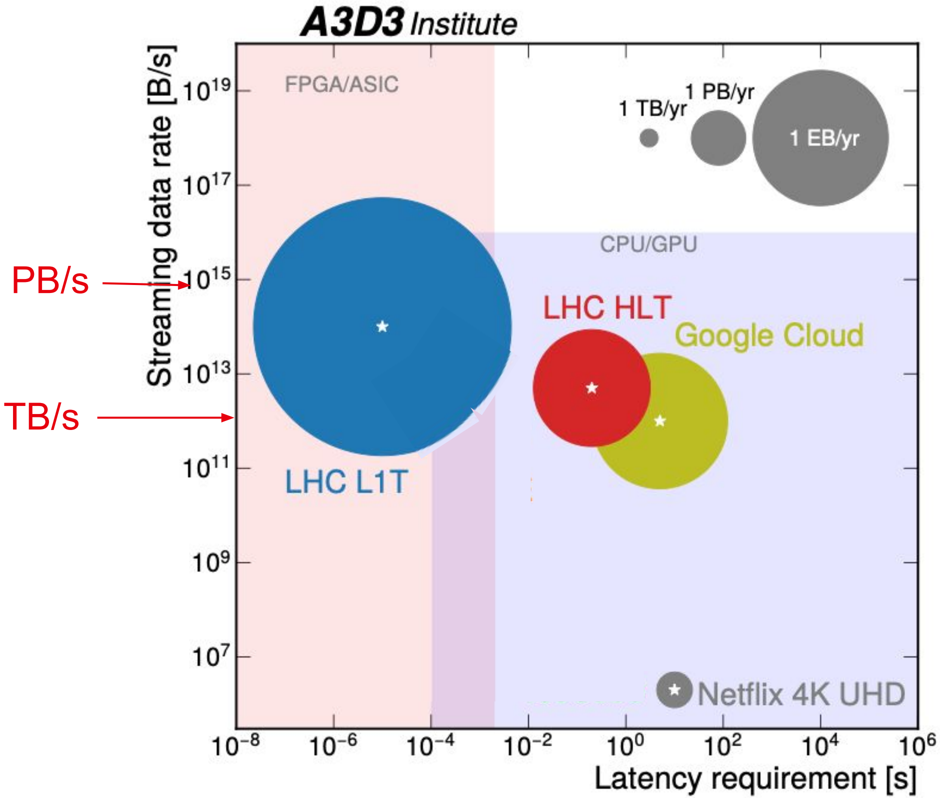
Critical computing challenge



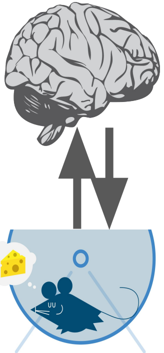
Smarter Algorithms - AI

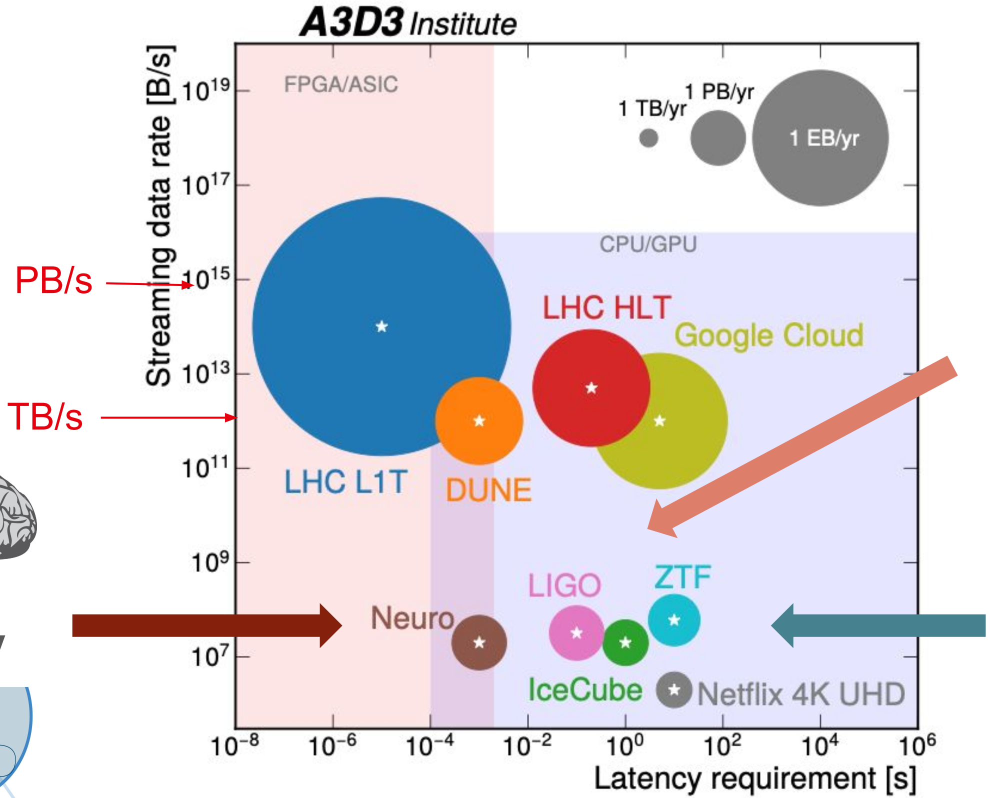
Faster Hardware - Co-processor

Critical computing challenges

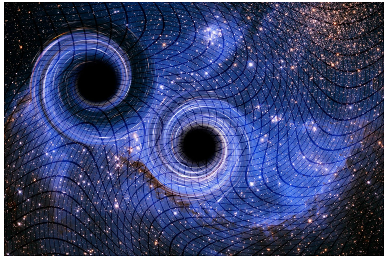


Common big data challenges

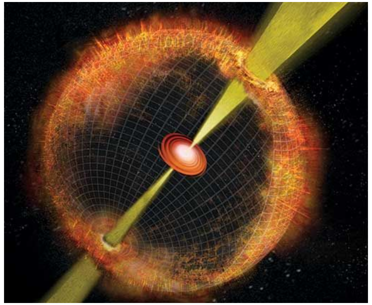




Gravitational Wave



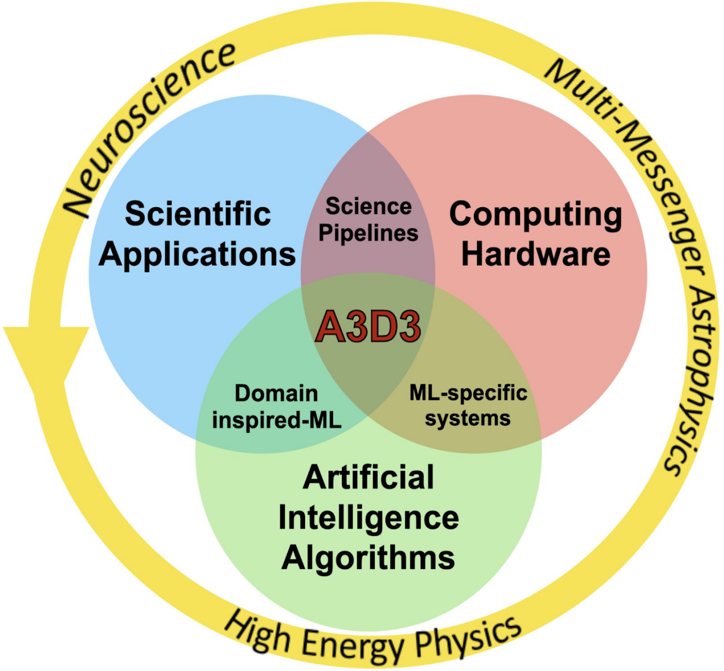
Supernova



NSF HDR Institute A3D3

Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery

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



Cross-institution

Spread across the USA at **16** institutions for **104** members

ETH Zürich

nycu NATIONAL YANG MING CHIAO TUNG UNIVERSITY

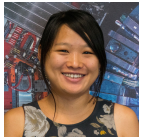
Cross-discipline

HEP


Hsu
PI


Harris
co-PI


Neubauer
co-PI


Liu


Duarte


Rankin


Sravan

MMA


Coughlin
co-PI


Scholberg
co-PI


Graham


Hanson


Katsavounidis

Neuros


Orsborn

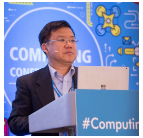

Shlizerman

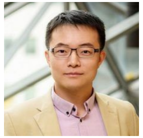

Dadarlat Makin

CS/EE


Hauck


Li


Chen

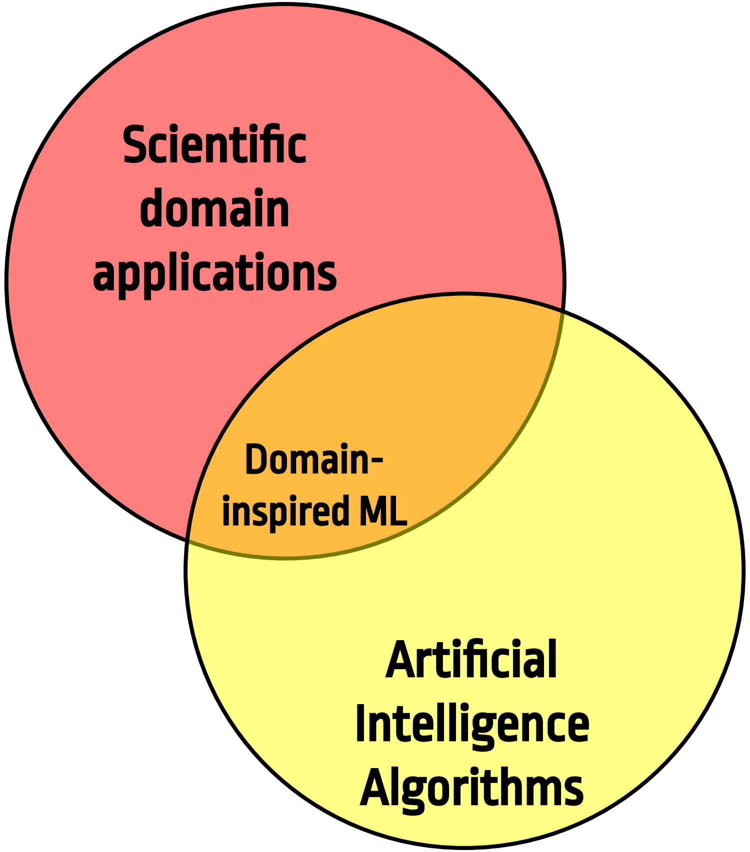

Han


Ju

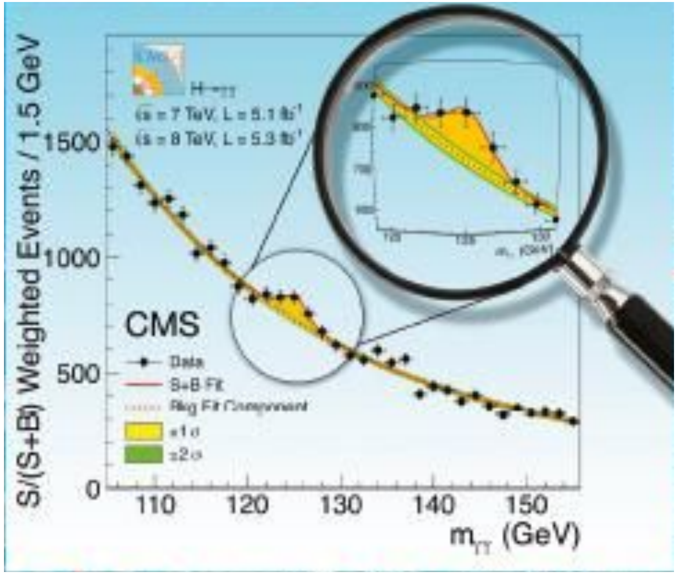

Lai
(NYCU)
賴伯承

17 Senior Personal
5 Affiliates
10 Postdocs
58 Graduates

Smarter Algorithms



AI has made critical contributions to the Higgs Discovery!

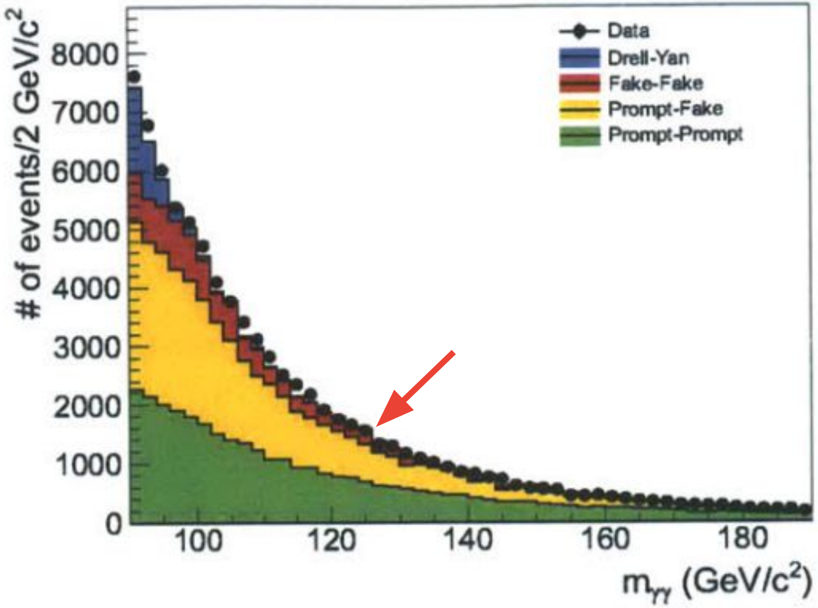


PHYSICS LETTERS B Volume 716, Issue 1, 17 September 2012

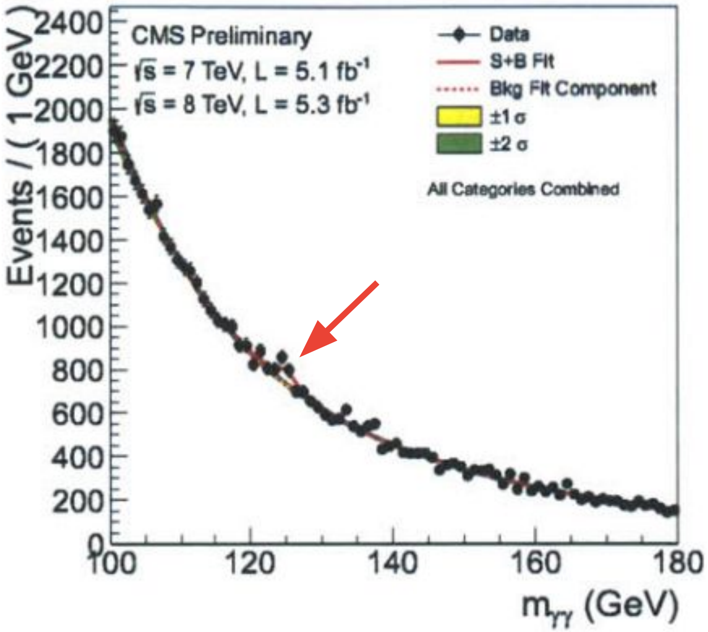
[CERN-THESIS-2013-079]

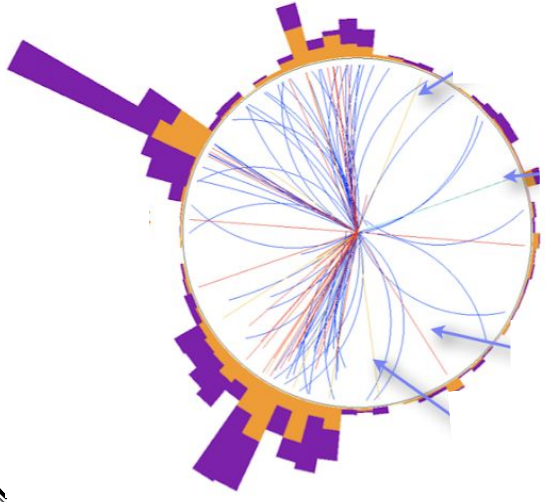
Key for discovery

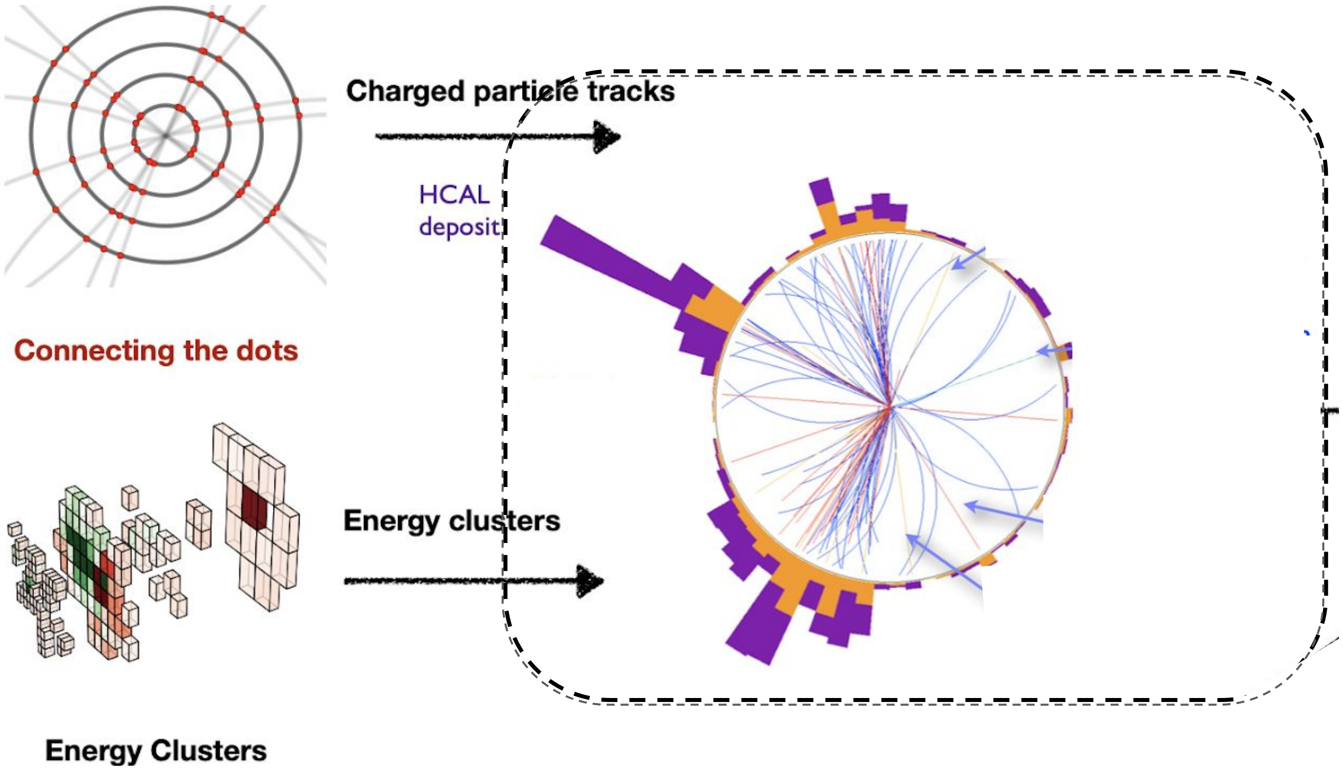
- Optimizing **signal-to-background** ratio

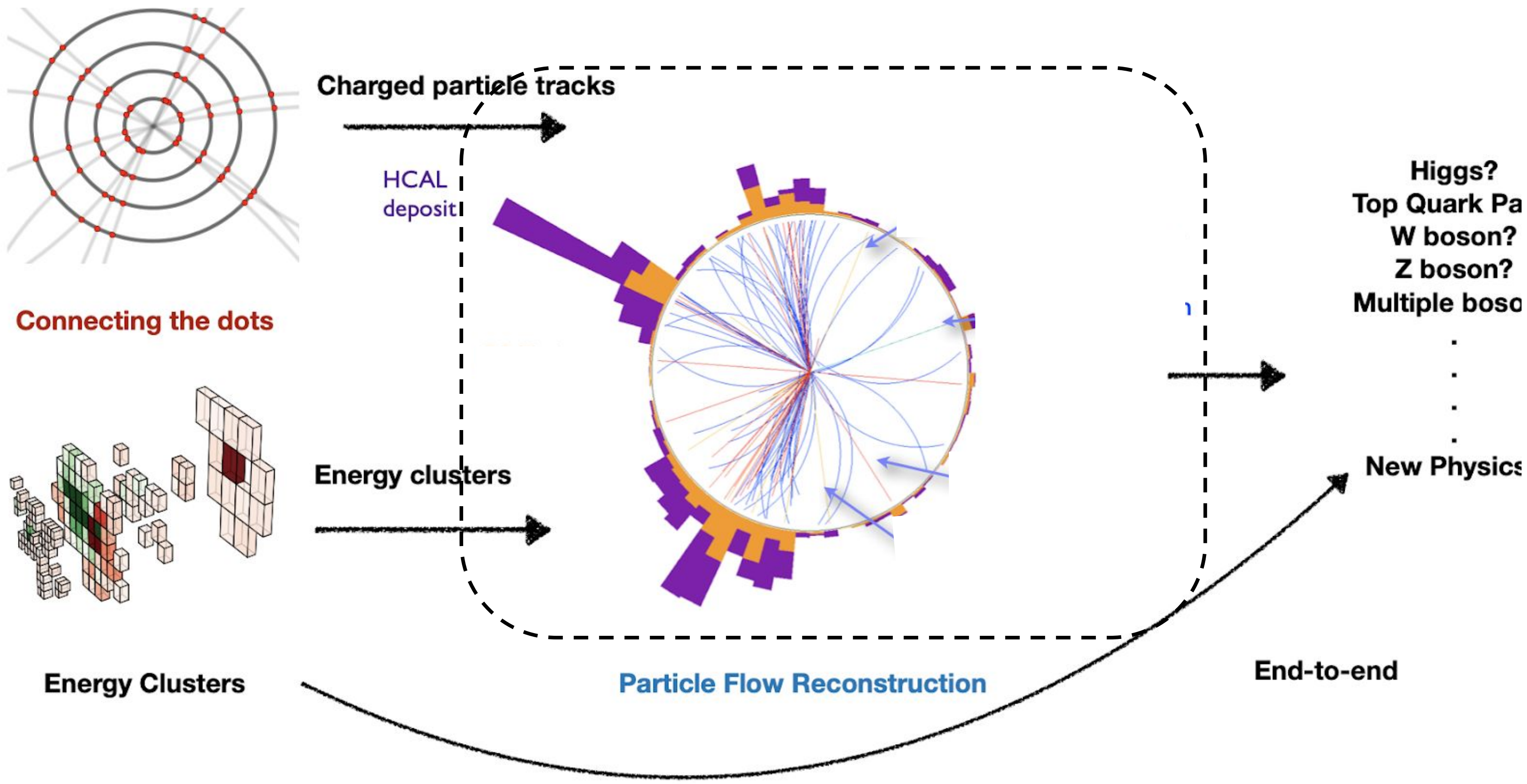


AI



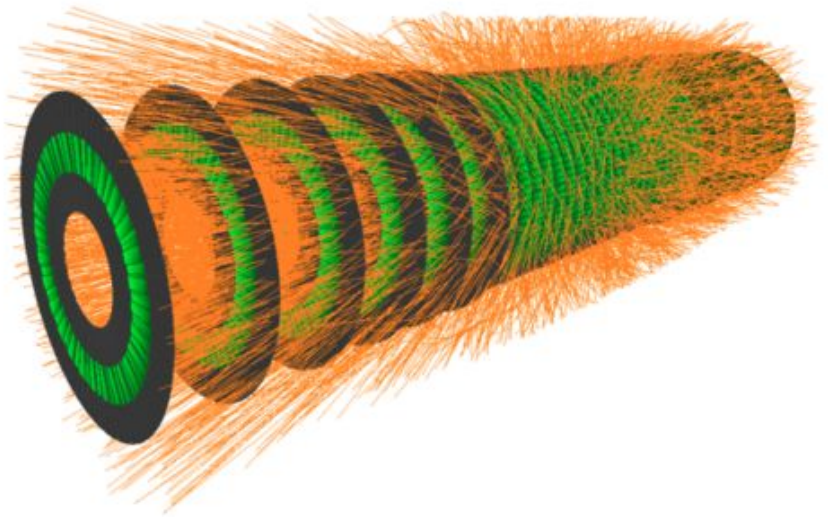


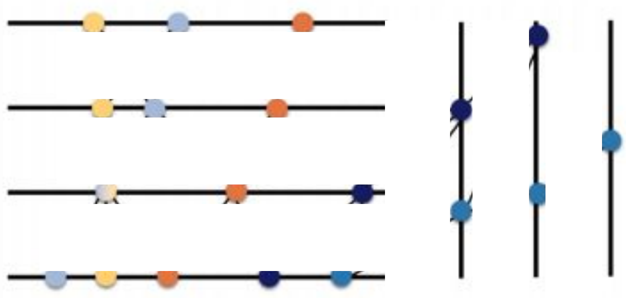

Charged particle tracks
HCAL
deposit
Connecting the dots
Energy clusters
Energy Clusters


Charged particle tracks
HCAL
deposit
Connecting the dots
Energy clusters
Energy Clusters
Particle Flow Reconstruction
Higgs?
Top Quark Pa
W boson?
Z boson?
Multiple boso
New Physics
End-to-end

Track Reconstruction as Graph

arxiv:2103.06995



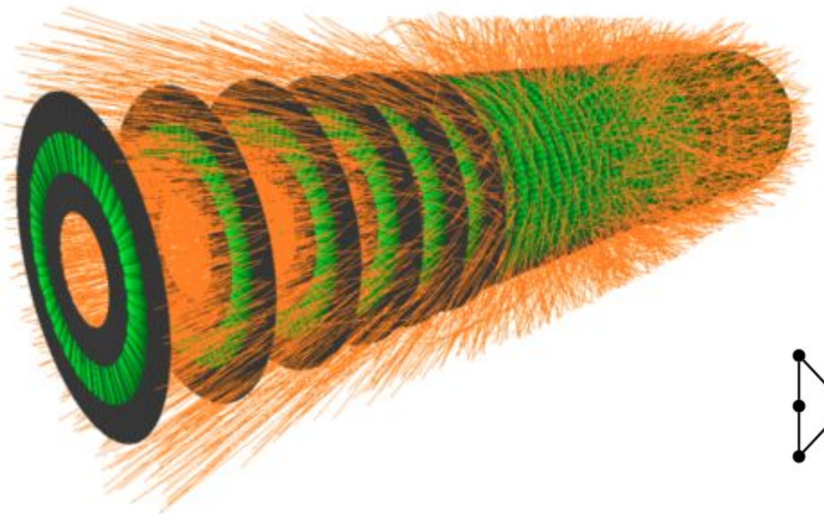


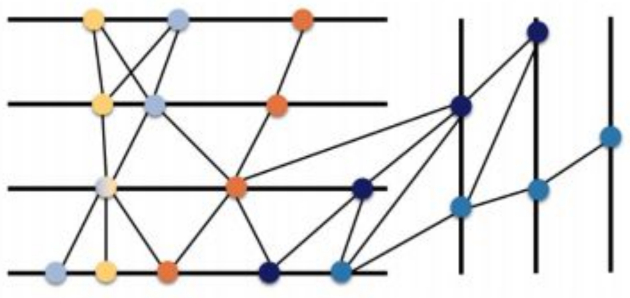
X. Ju, et. al. (Hsu, Thais) EPJC 81, 876 (2021)

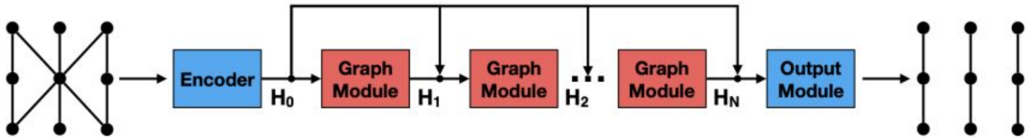
Track Reconstruction as Graph

In FTCF24: J. Zhang BESIII, H. Zhou STCF, A. Salzburger ACTS

Graph Neural Network to identify correct edge connecting adjacent nodes



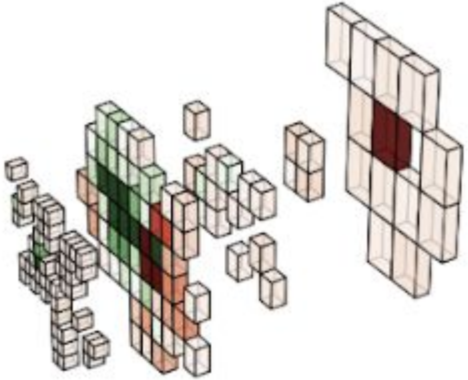




Clustering with Sparse Point Voxel Convolutional Neural Network

J. Krupa FastML'23

- Torchsparse/ Torchsparse++ (Haotian Tang, et al. @ MLSys'22)
  - **2.9X** faster than MinkowskiEngine (NVIDIA)
  - **1.8X** faster than SpConv (TuSimple).



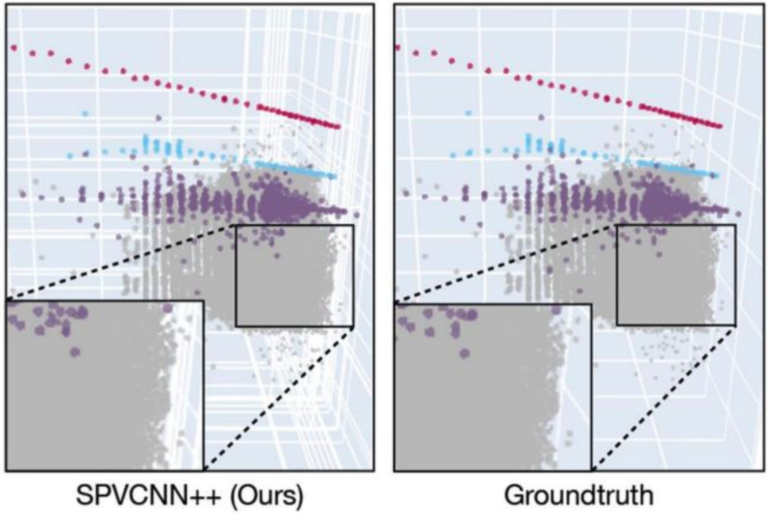
Energy Clusters

Clustering with Sparse Point Voxel Convolutional Neural Network

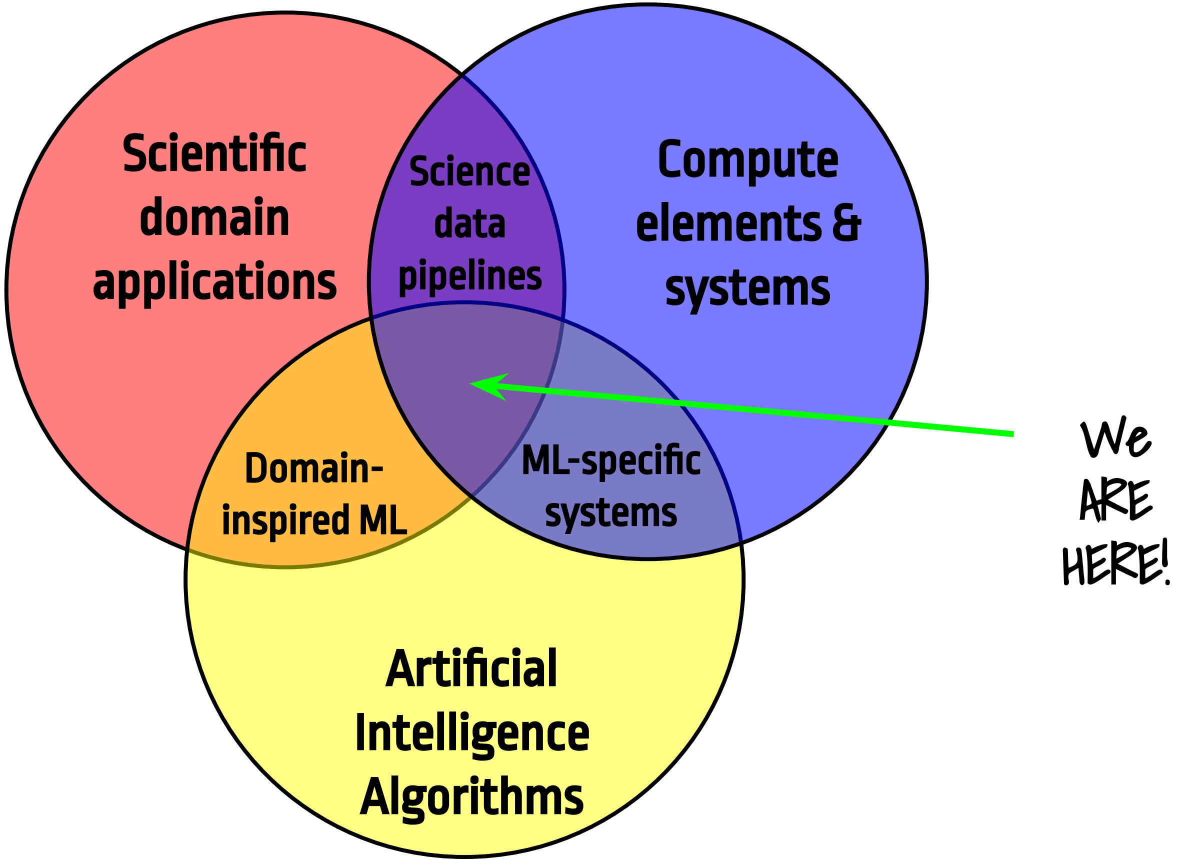
- Torchsparse/ Torchsparse++ (Haotian Tang, et al. @ MLSys'22)
  - **2.9X** faster than MinkowskiEngine (NVIDIA)
  - **1.8X** faster than SpConv (TuSimple).

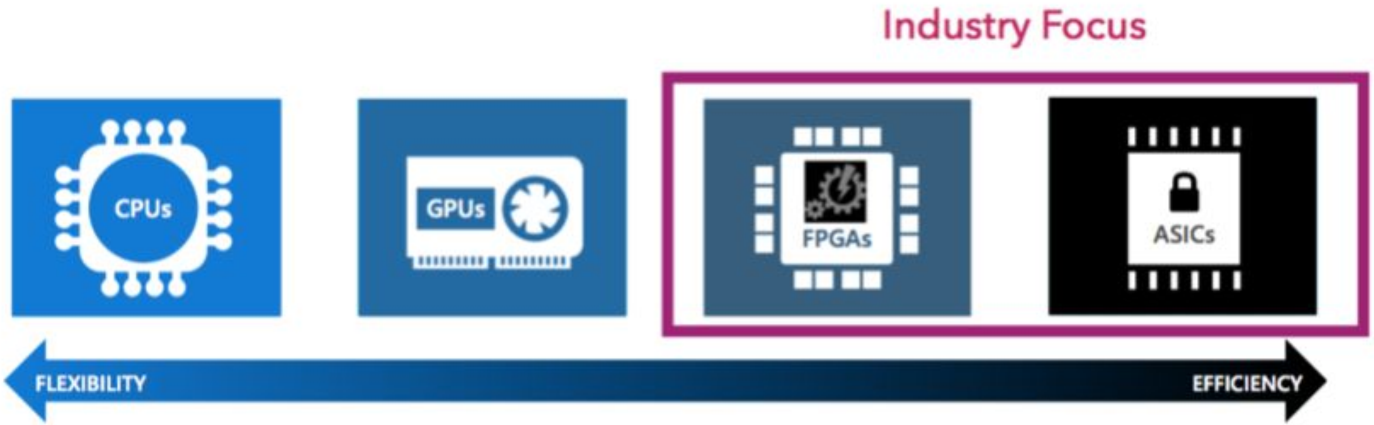
Particles are a set of 3D points and can be processed by our efficient 3D algorithms.

4% higher mIoU and **10+% higher** PQ



Smarter Algorithms and Faster Hardware

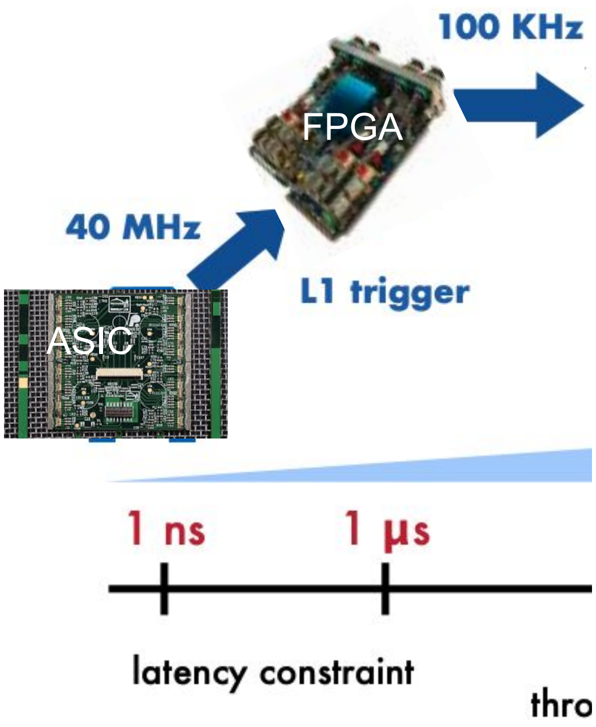



Industry Focus
CPUs
GPUs
FPGAs
ASICs
FLEXIBILITY
EFFICIENCY

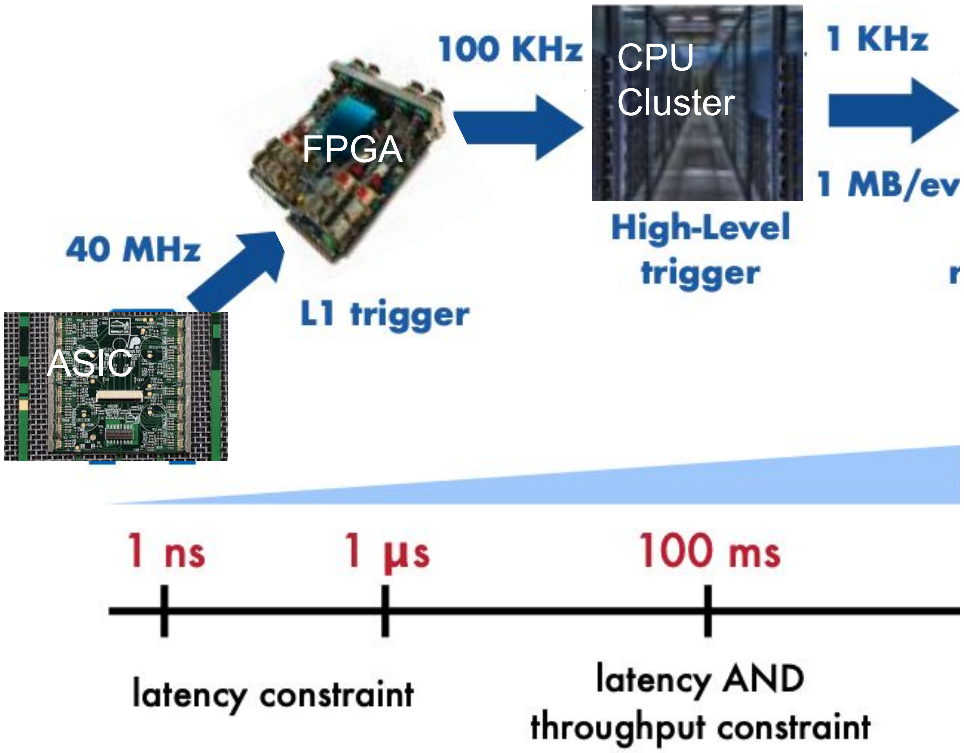
The Need for the FastML

Heterogeneous Computing

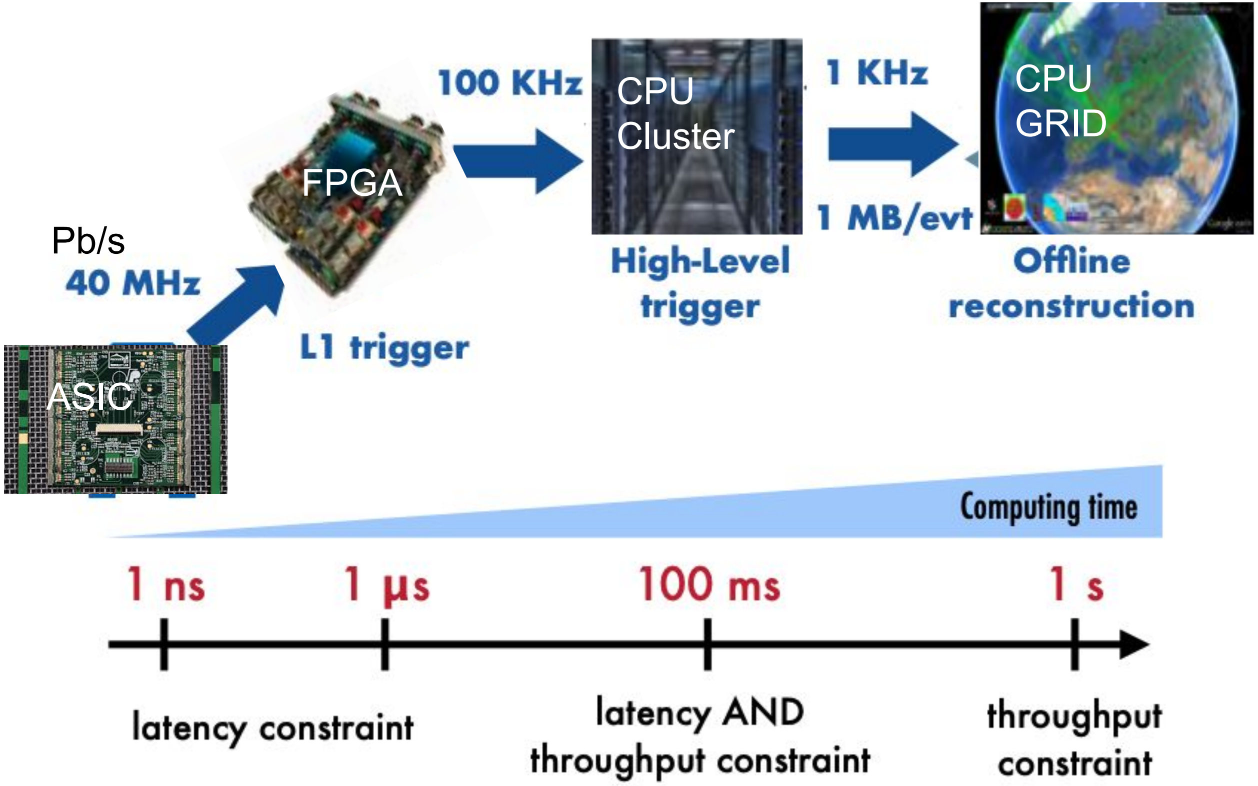
Heterogeneous Computing



Heterogeneous Computing



Heterogeneous Computing



The Need for the FastML

Fast Machine Learning Collaboration

Core
(HLS4ML)

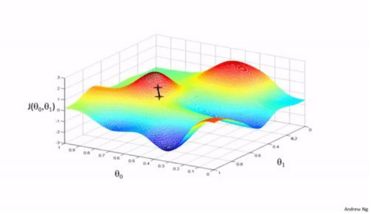
Co-processor
(SONIC)

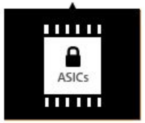
Innovative Algorithms

HAC Research Focus

Co-design, Design Automation

Algorithm ⟷ Hardware





Challenges in Algorithm Design:

- Irregular data (graphs, point clouds)
- Label scarsity
- AI models are hard to be interpreted
- …

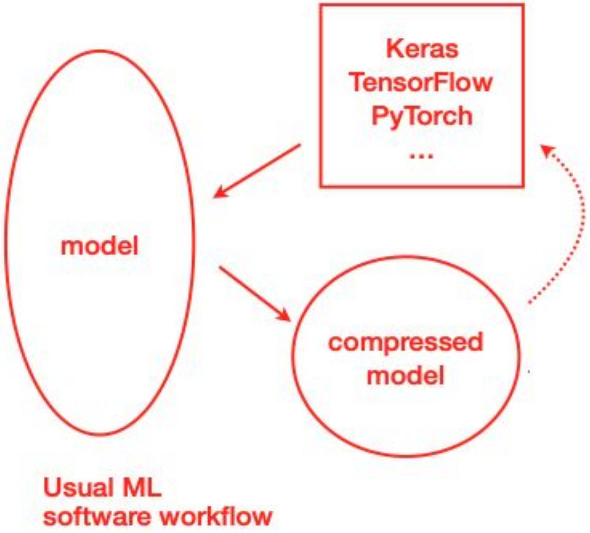
Challenges in Deployment in Hardware:

- Computation efficiency issues
- Power/memory constraints
- Hard to be implemented on FPGA/ASIC

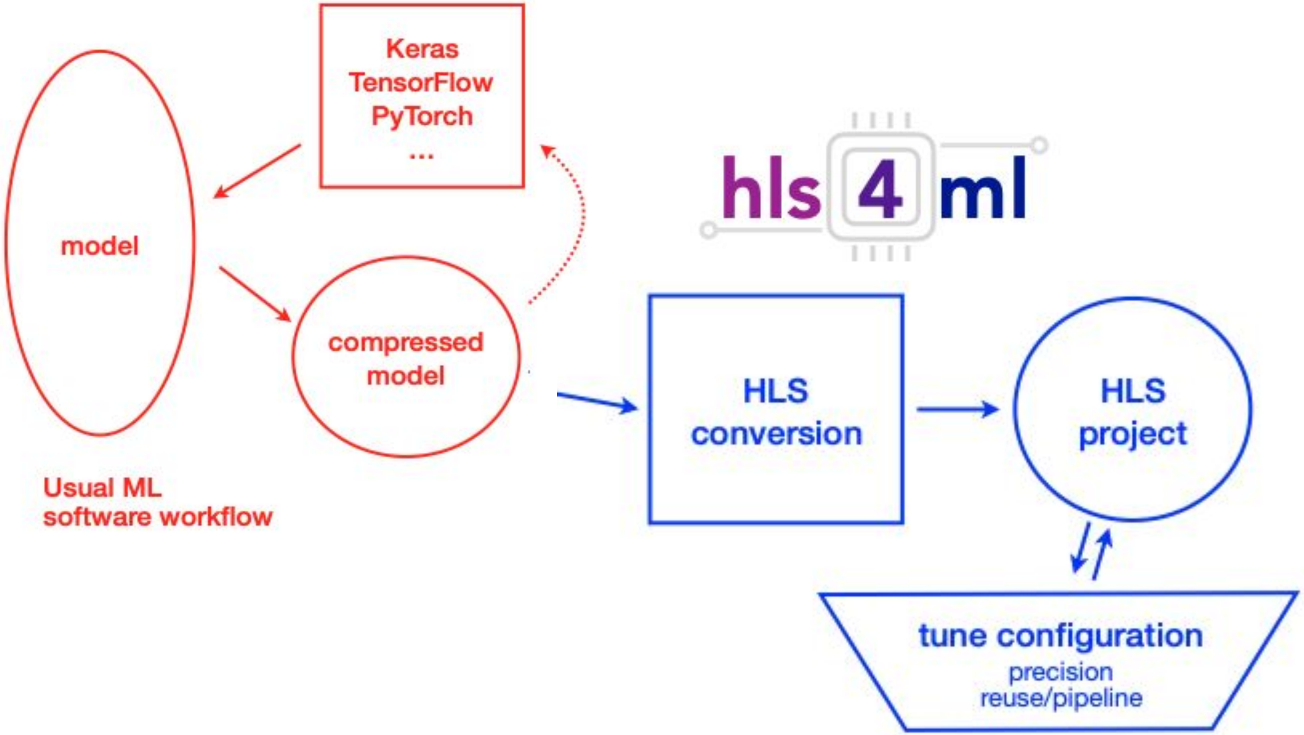
…

--> hardware design automation tools

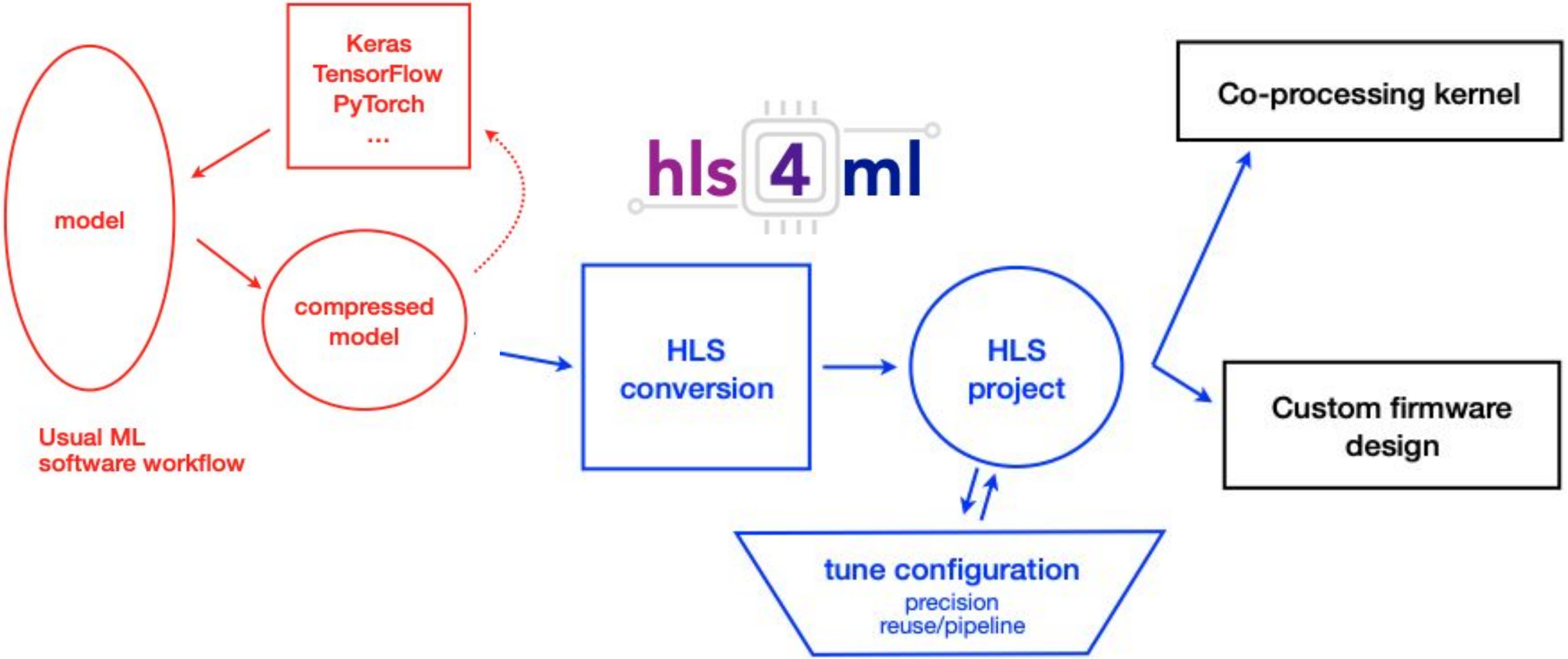
HLS4ML translating ML into FPGA firmware



HLS4ML translating ML into FPGA firmware



HLS4ML translating ML into FPGA firmware



Quantization

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

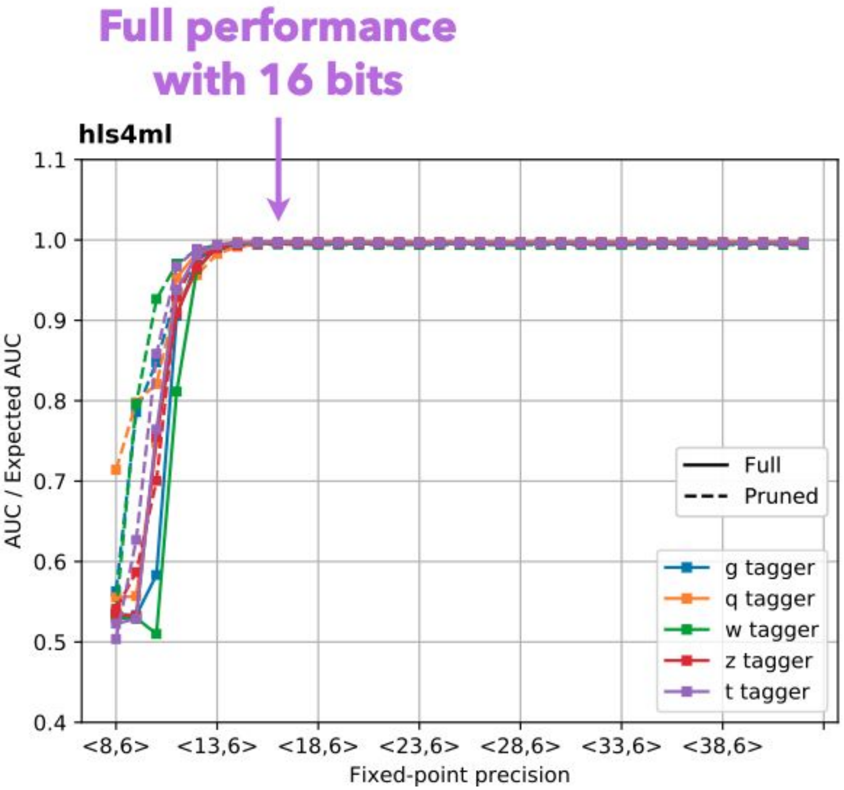
- Scan the bit width until you reach optimal performance

ap_fixed<width,integer>

0101.1011101010

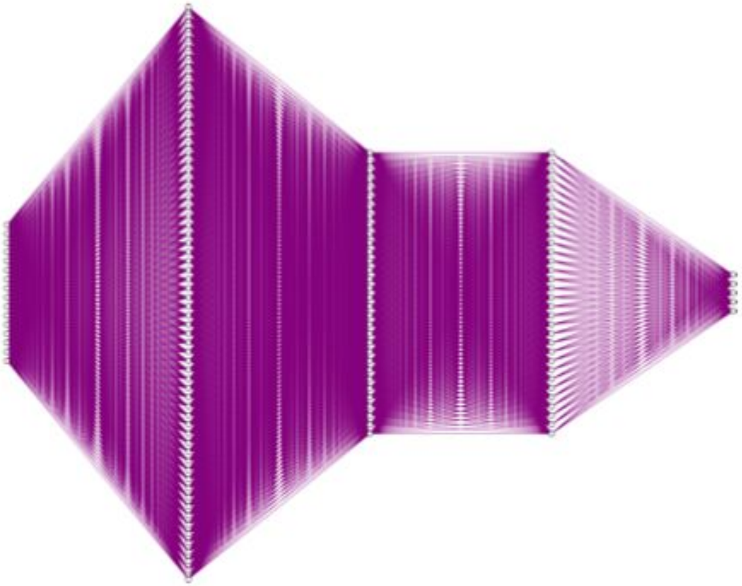
integer | fractional

width



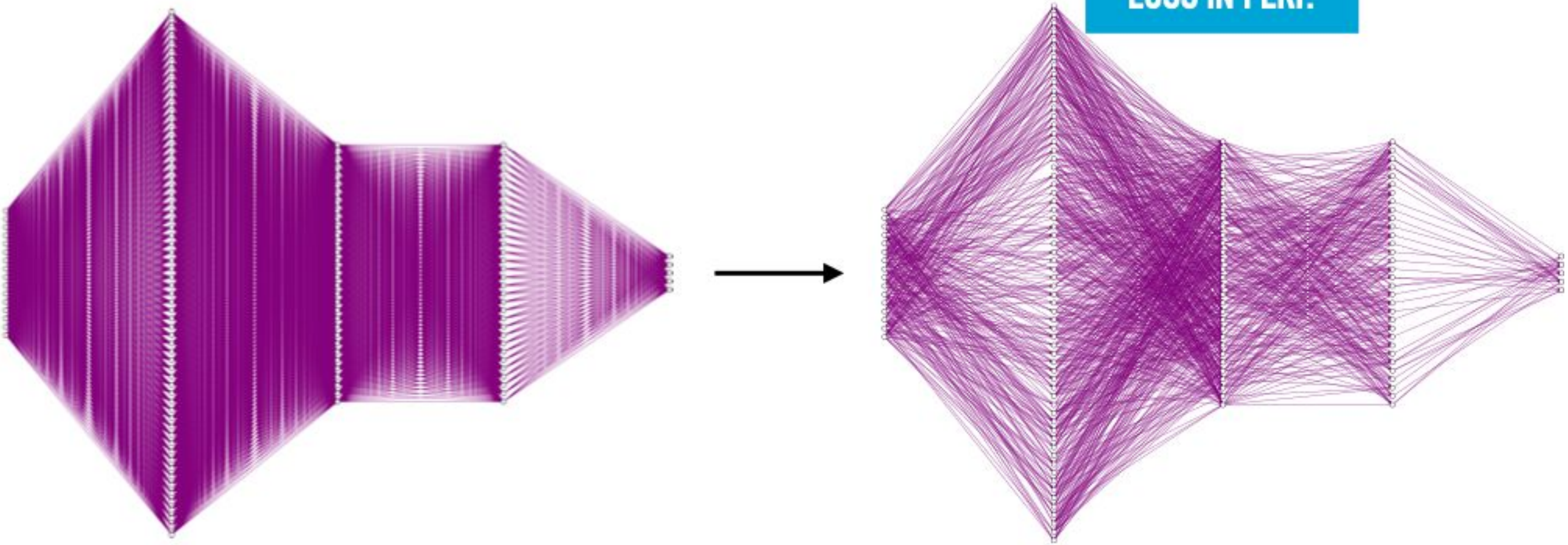
Compression

- Remove **smallest** weigh
- Iterate



Compression

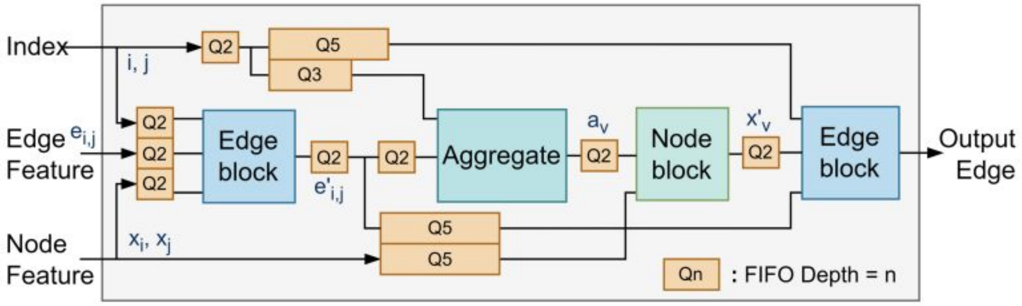
- Remove **smallest** weights
- Iterate

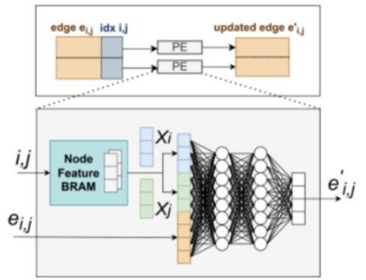


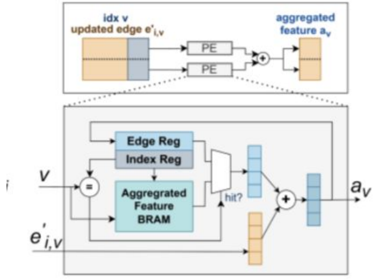
LOW LATENCY EDGE CLASSIFICATION GNN

[Shi-Yu Huang, Yun-Chen Yang, Yu-Ru Si, et. al. FPL 2023]

Modularized parallel architecture for each computational pipelines



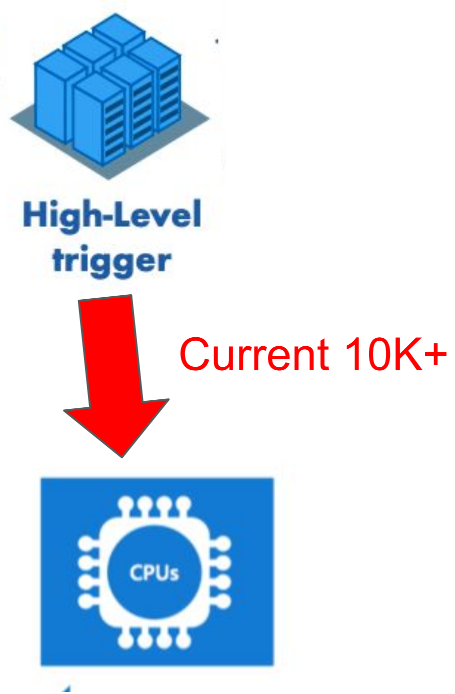




Achieving 2.07 us Latency with 3.225 Throughput (MGPS)

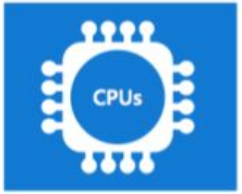
- Xilinx Virtex UltraScale+ VU9P HLS 2019.2

High-Level Trigger (100 KHz, 100 ms latency)



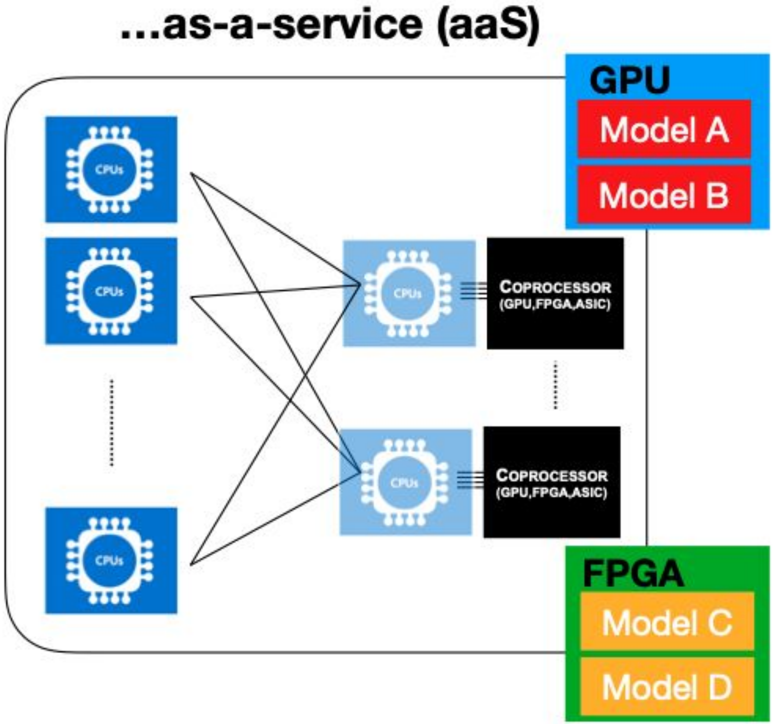
High-Level Trigger (100 KHz, 100 ms latency)

Our proposal



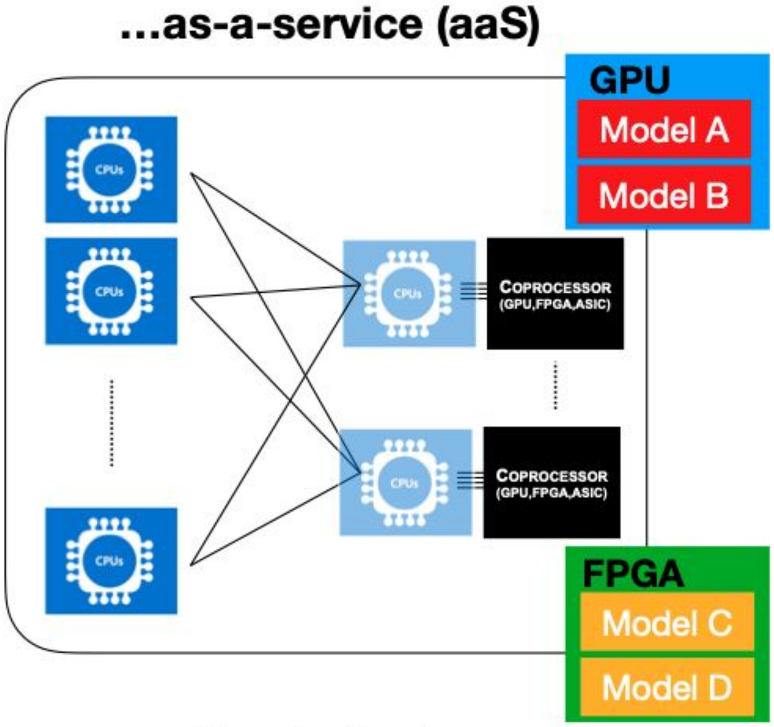
ML-as-a-Service

- Simple support for mixed hardware
- Scaleable
- Throughput optimization for multiple-core
- Simple client-side



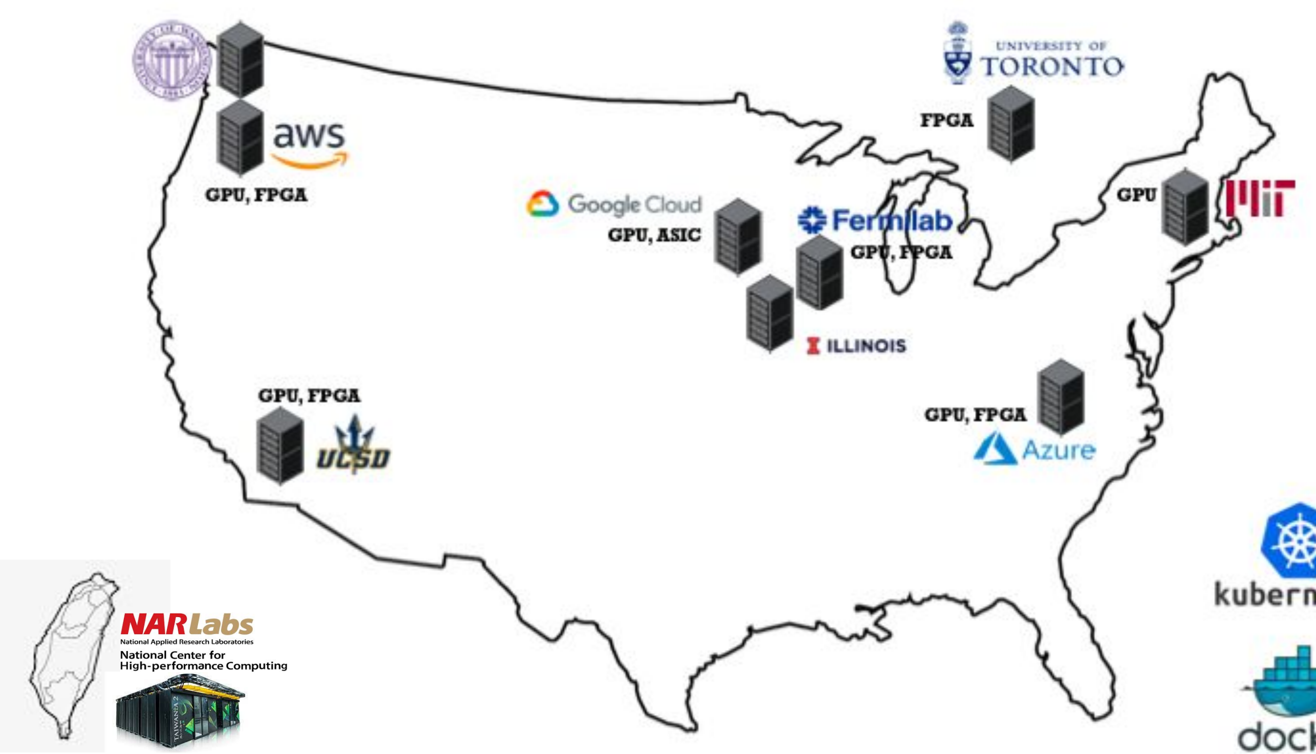
ML-as-a-Service

- Simple support for mixed hardware
- Scaleable
- Throughput optimization for multiple-core
- Simple client-side



NVIDIA
TRITON INFERENCE SERVER
kubernetes
Azure
aws

Scalability Test



kubernetes

docker

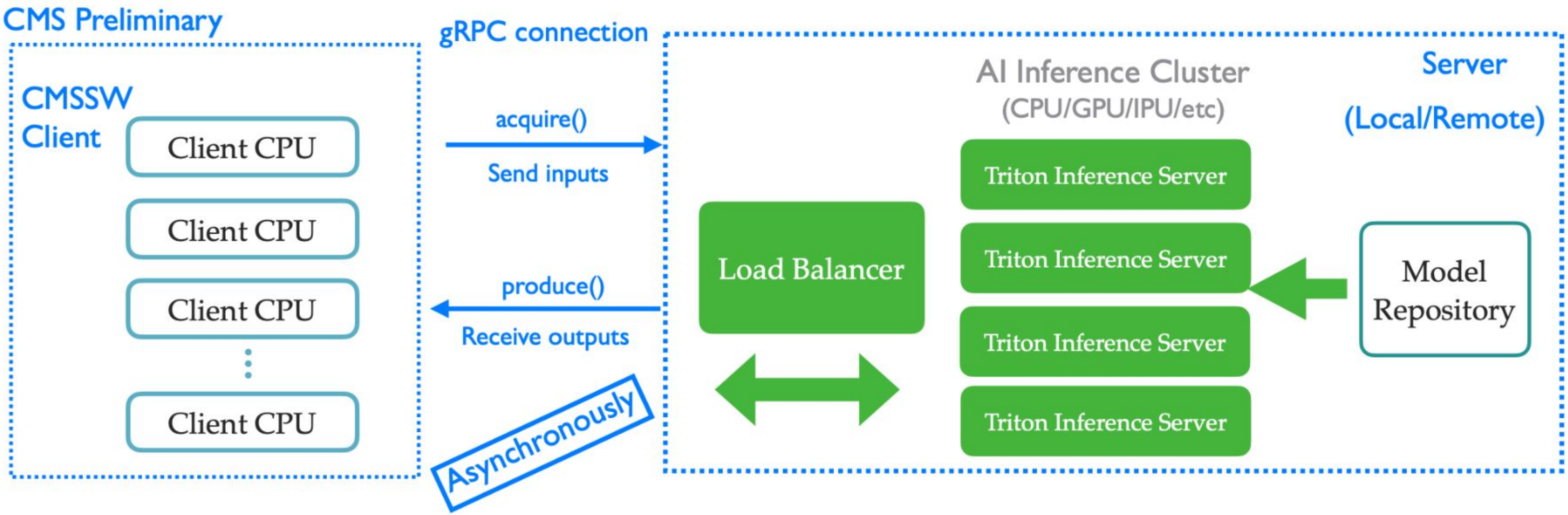
Building a network of heterogeneous resources in the cloud and on-premises

Work-in-progress: how to coordinate and orchestrate distributed heterogeneous resources

SONIC

P. McComack - FastML 2023

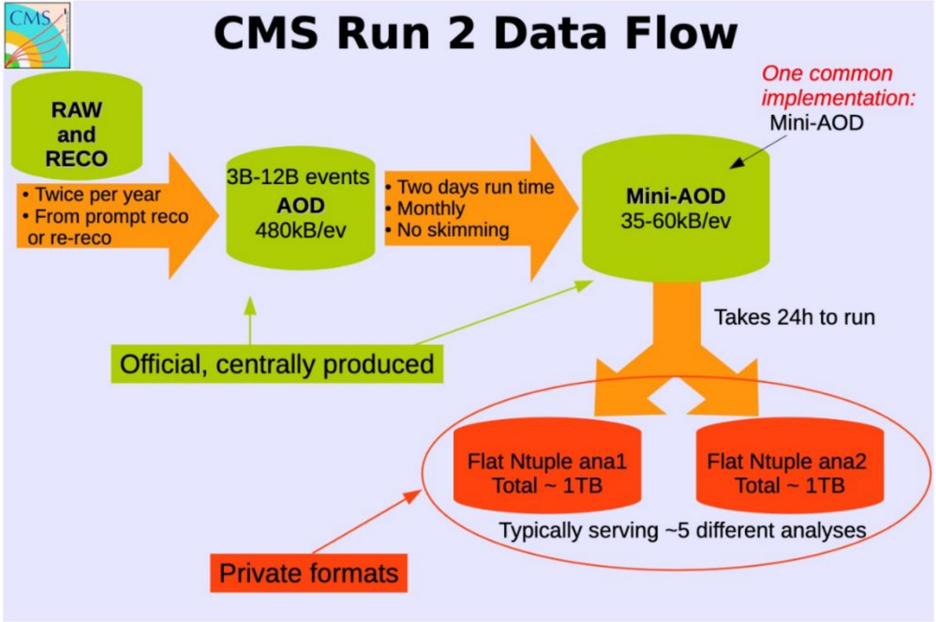
- Within CMS software (CMSSW), the IaaS deployment scheme is called "Services for Optimized Network Inference on Coprocessors" (SONIC)



Studying SONIC at scale

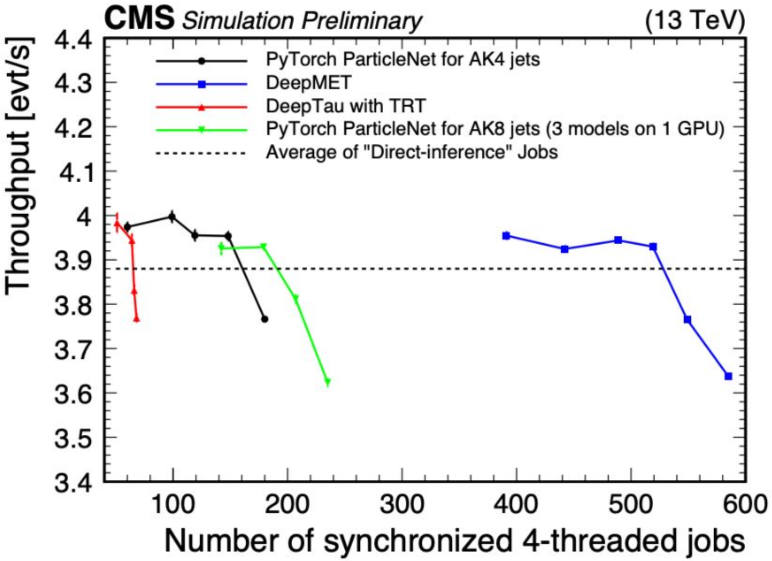
- As a testbed for SONIC-enabled deployment, we created a MiniAOD demonstrator workflow
    - Runs a refinement and slimming step of CMS data processing
    - Full MiniAOD processing workflow typically run ~monthly

[1702.04685]



Mini-AOD production typically takes about 0.5 seconds per event on production grid nodes

Optimizing performance: CPU-to-GPU ratio



- Having explored server parameters, we can test the number of client jobs that a single GPU can handle
- We perform these tests in the cloud, as we need to synchronize jobs running on O(1000) CPU cores

Summary

- Artificial Intelligence heavily applied to Physics Discovery
    - For examples, Higgs discovery!
- HL-LHC confronted Big Data challenge
    - Smart Machine Learning could offer partial solutions
- A3D3 focusing on accelerating AI to solve common challenges through interdisciplinary collaboration
    - Perfect time to join the growing community
    - Upcoming events
        - HDR Ecosystem conference, UIUC (Sep 9 2024)
            - https://indico.cern.ch/event/1364455/
            - Machine Learning Challenge (2024)

Shih-Chieh Hsu

http://faculty.washington.edu/schsu/

schsu@uw.edu

Backup

Studying SONIC at scale

- Inferences for three classes of algorithms were run through SONIC:
  - ONNX-based jet tagger
  - TensorFlow based missing energy calculation
  - TensorFlow based CNN for tau lepton ID
- These algorithms consume about 10% of total workflow latency

Algorithm	Time [ms]	Fraction [%]	Input [MB]
PN-AK4	42.4	4.3	0.04
PN-AK8	11.4	1.1	0.003
DeepMET	13.2	1.3	0.33
DeepTau	21.1	2.1	1.18
ParticleNet+DeepMET+DeepTau	88.1	8.8	1.55
Total	993.3	100.0	—