Accelerating Artificial Intelligence for High Energy Physics

Shih-Chieh Hsu (徐士傑）
University of Washington

PHY-2117997

IAS HEP 2024 (https://indico.cern.ch/event/1335278/)
HKUST Jockey Club Institute for Advanced Study
Jan 24 2024

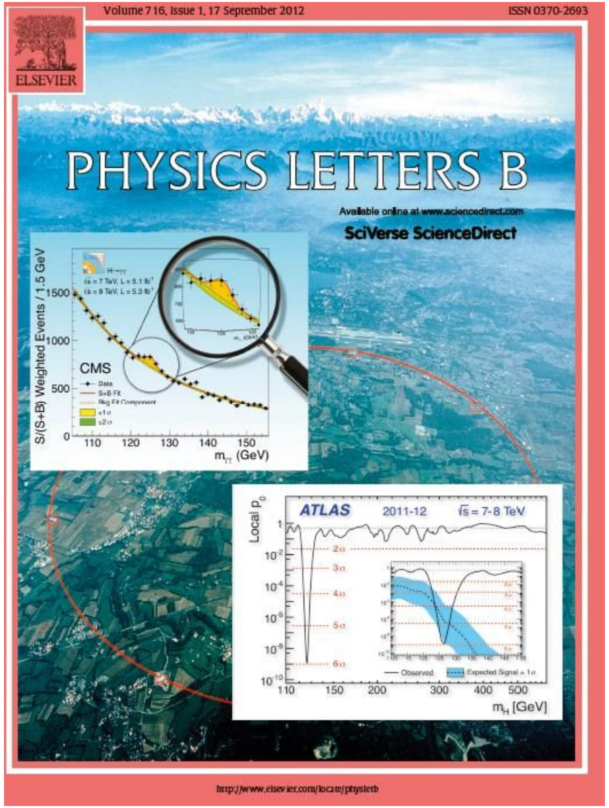
https://a3d3.ai/

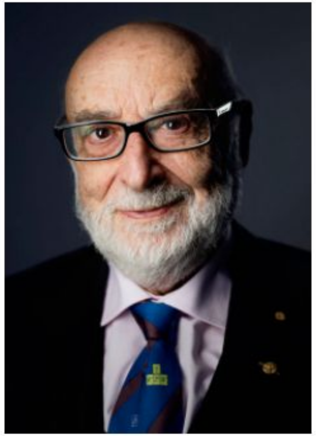
P5 Report (Draft Dec 2023)

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.

AI/ML has made critical contributions to the Higgs Discovery!





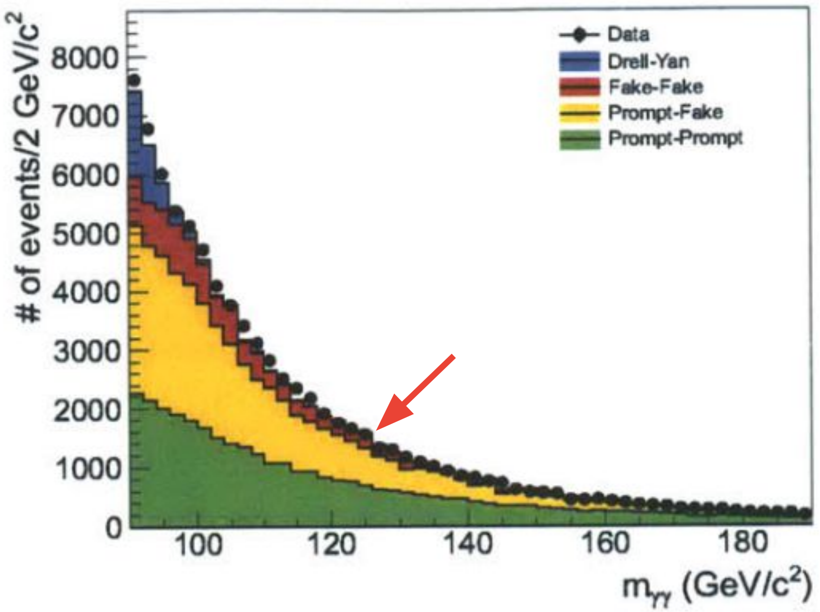
© Nobel Media AB. Photo: A. Mahmoud

François Englert

© Nobel Media AB. Photo: A. Mahmoud

Peter W. Higgs

2013

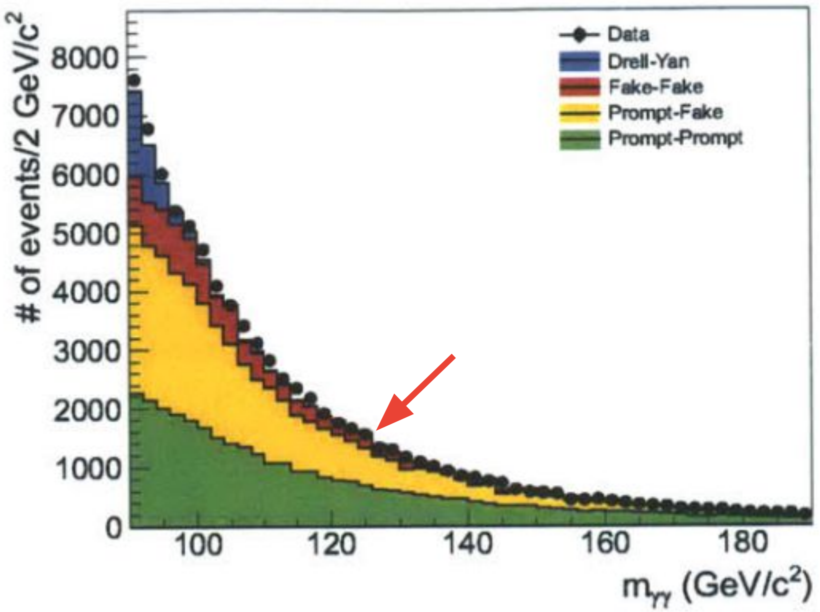


# of events/2 GeV/c^2
8000
7000
6000
5000
4000
3000
2000
1000
0
100
120
140
160
180
$m_{\gamma\gamma}$ (GeV/c^2)
Data
Drell-Yan
Fake-Fake
Prompt-Fake
Prompt-Prompt

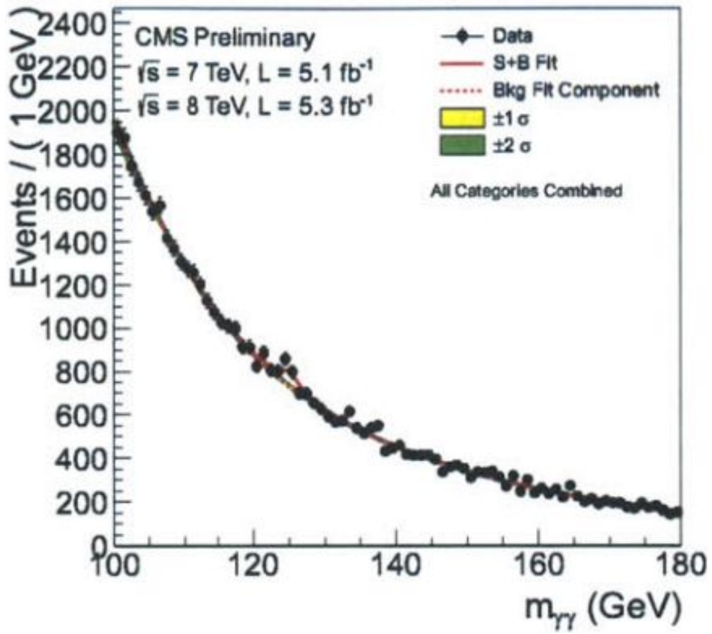
J.L. Bendavid, THESIS-2013-079

Key for discovery

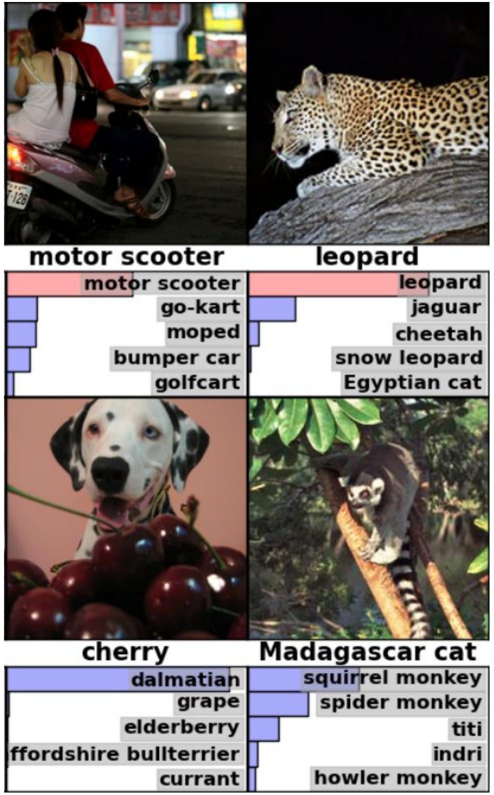
- Optimizing **signal-to-background** ratio



BDT



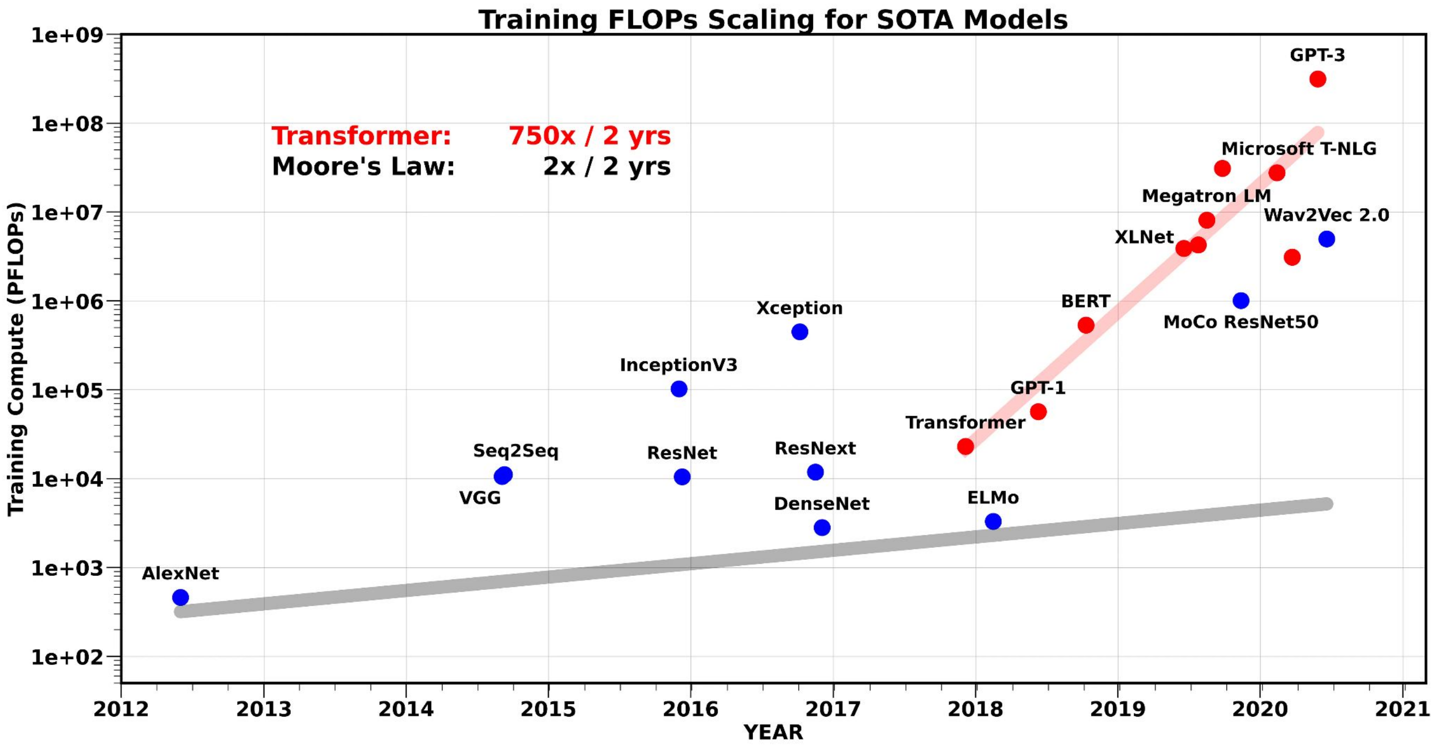
2012: A Breakthrough Year for Deep Learning



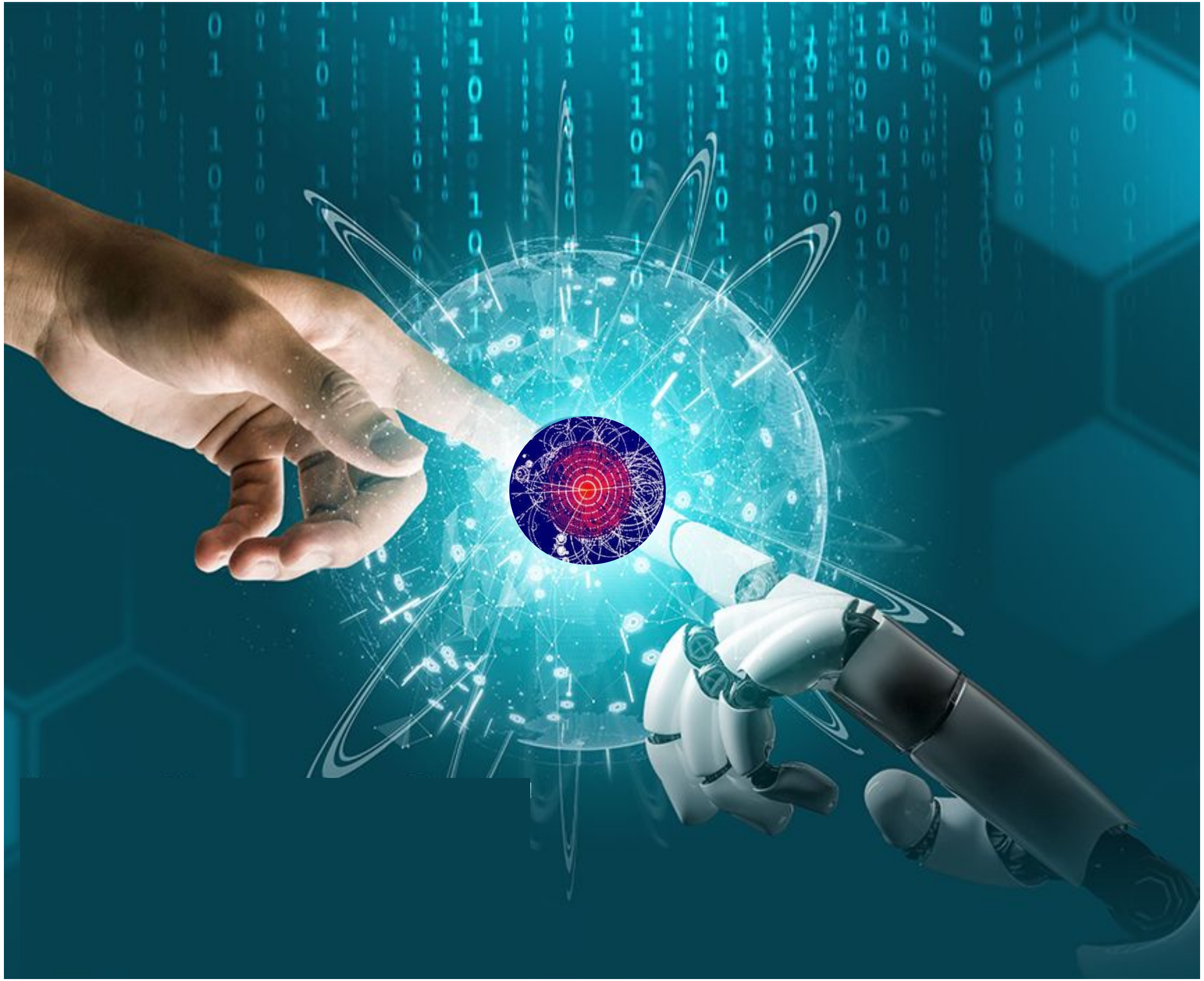
AlexNet Comm. ACM. 60 (6): 84–90

ACM 2018 Turing Award

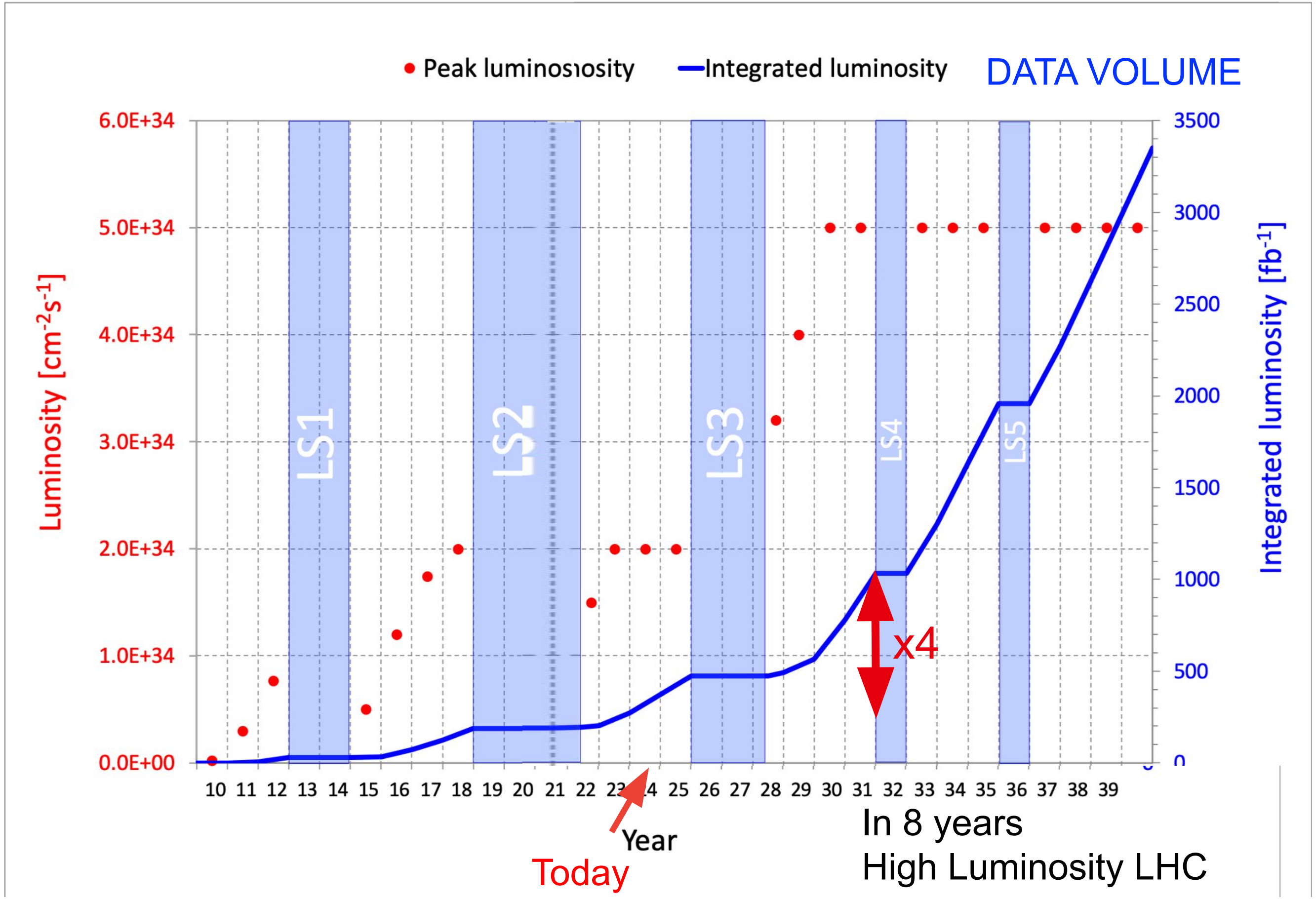
Exponential trend of computation need for AI



A. Gholami

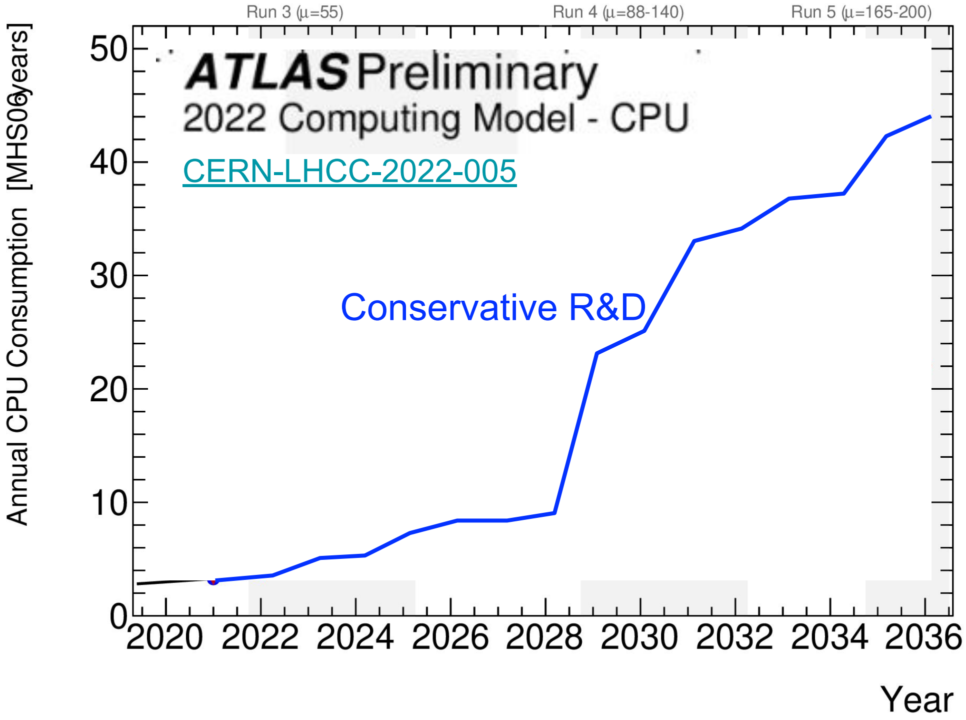


Credit: Onpassive



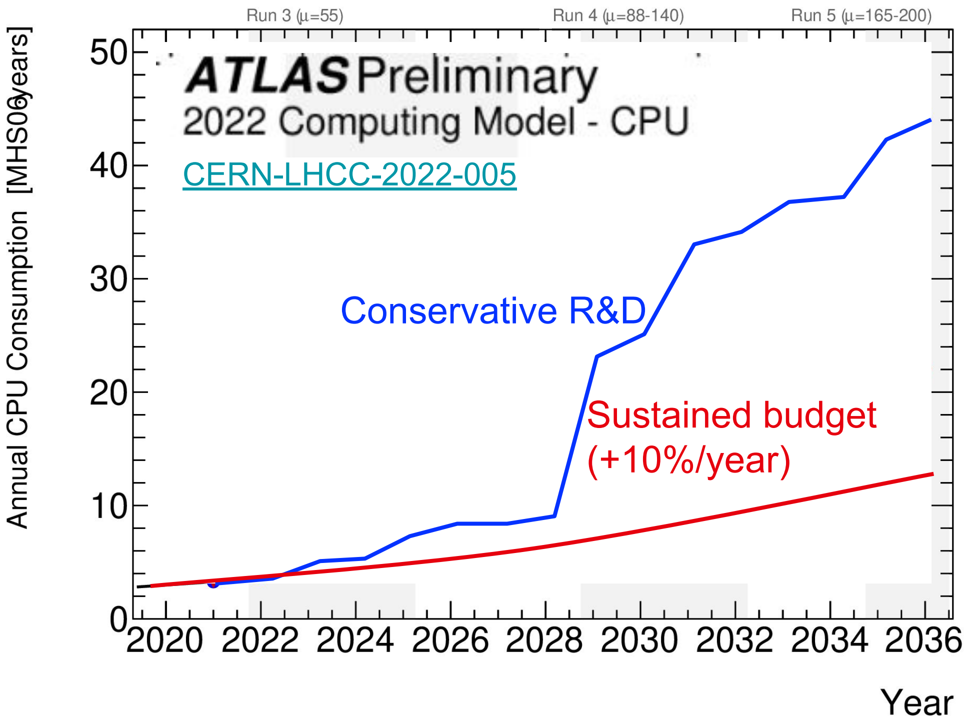
Peak luminosiosity
Integrated luminosity
DATA VOLUME
Luminosity [cm-2s-1]
Integrated luminosity [fb-1]
6.0E+34
5.0E+34
4.0E+34
3.0E+34
2.0E+34
1.0E+34
0.0E+00
3500
3000
2500
2000
1500
1000
500
0
LS1
LS2
LS3
LS4
LS5
x4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37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

Critical computing challenge



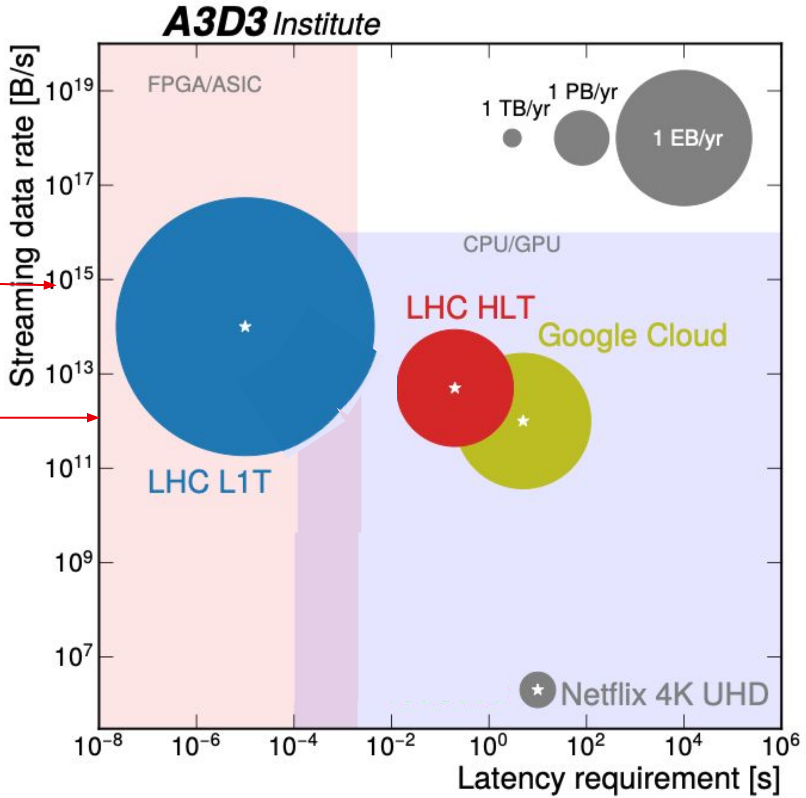
- To preserve current physics we are upgrading the system
  - We will have to take data at 4 times the current rate
  - Our event size will have to be 10x larger

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
- However, we are lacking of sufficient budget to sustain required computing

Critical computing challenges



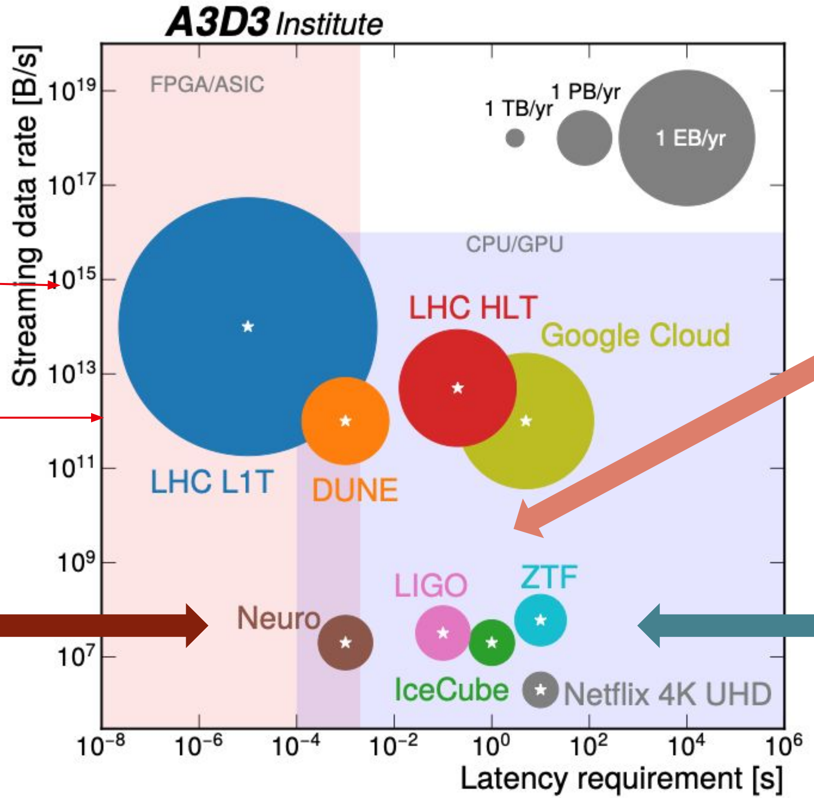
PB/s

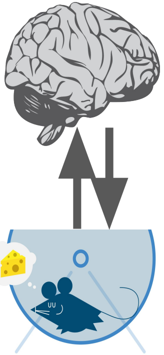
TB/s

Common big data challenges

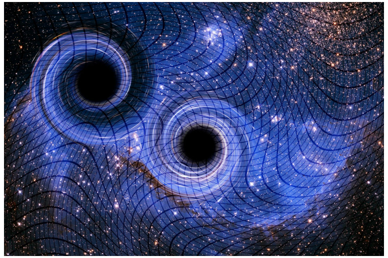
PB/s

TB/s

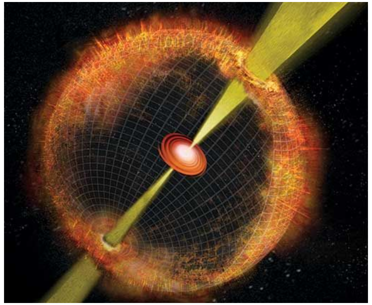




Gravitational Wave



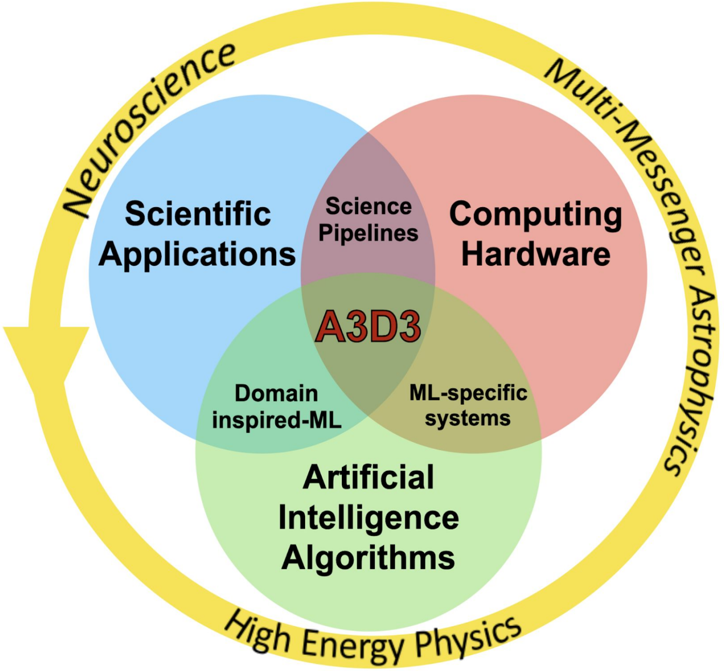
Supernova



NSF HDR Institute A3D3

Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery

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.



Cross-institution

16 institutions
104 members

ETH Zürich

NYCU NATIONAL YANG MING CHIAO TUNG UNIVERSITY

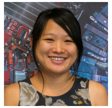
Cross-discipline

HEP


Hsu
PI/Director


Harris
co-PI


Neubauer
co-PI


Liu


Duarte


Rankin


Aarastad


Gonski


Carlsen

MMA


Coughlin
co-PI


Scholberg
co-PI


Graham


Riedel


Katsavounidis


Li


Sravan

Neuros


Orsborn


Shlizerman

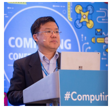

Dadarlat Makin

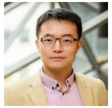

Sun

CS/EE


Hauck


Li


Chen

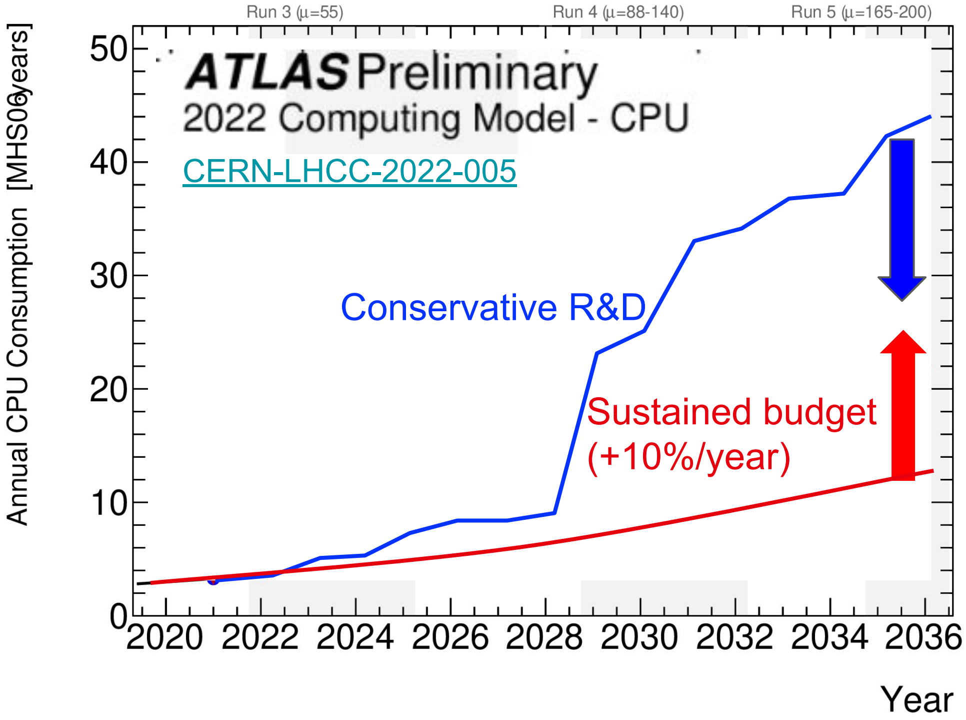

Han


Ju


Lai

17 Senior Personal
9 Affiliates
10 Postdocs
58 Graduates
10 Undergraduates

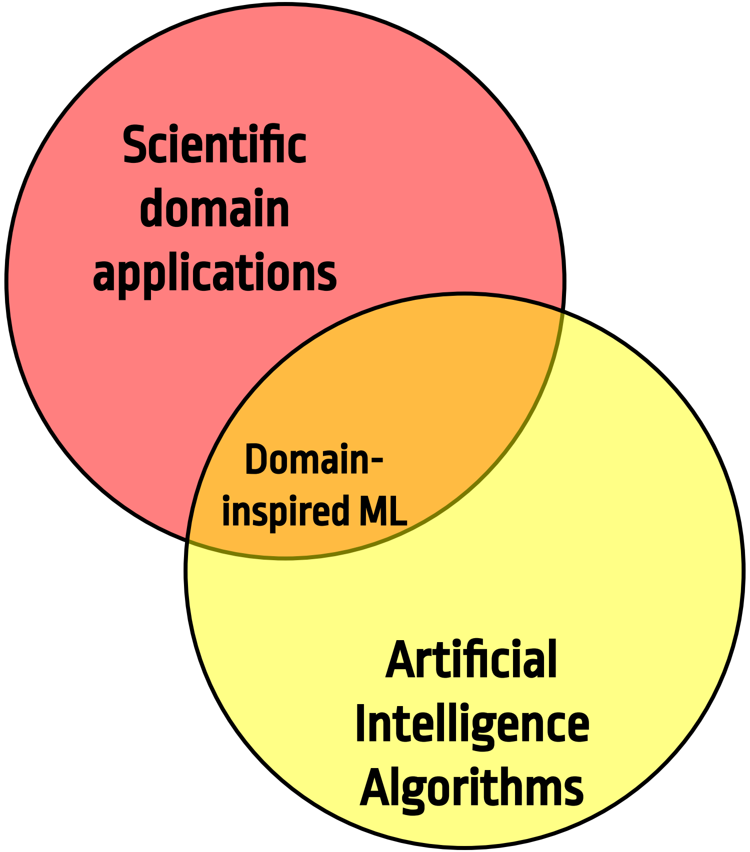
Closing the gap

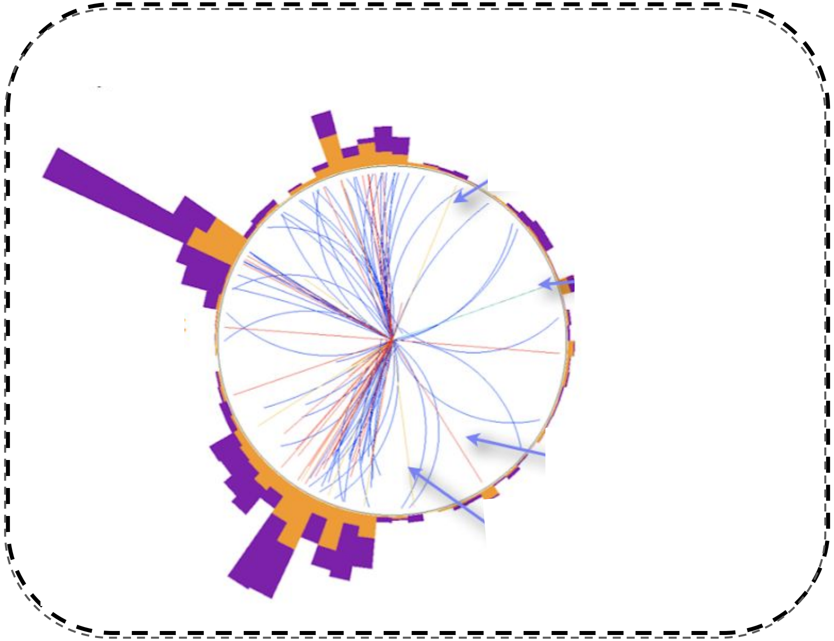


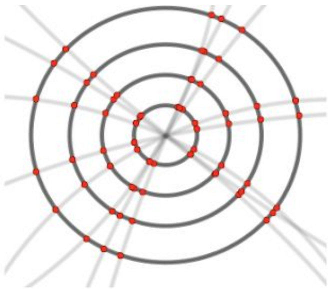
Smarter Algorithms - AI

Faster Hardware - Co-processor

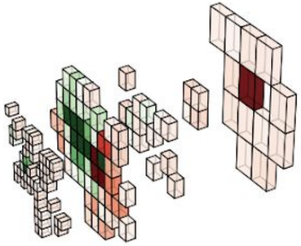
Smarter Algorithms



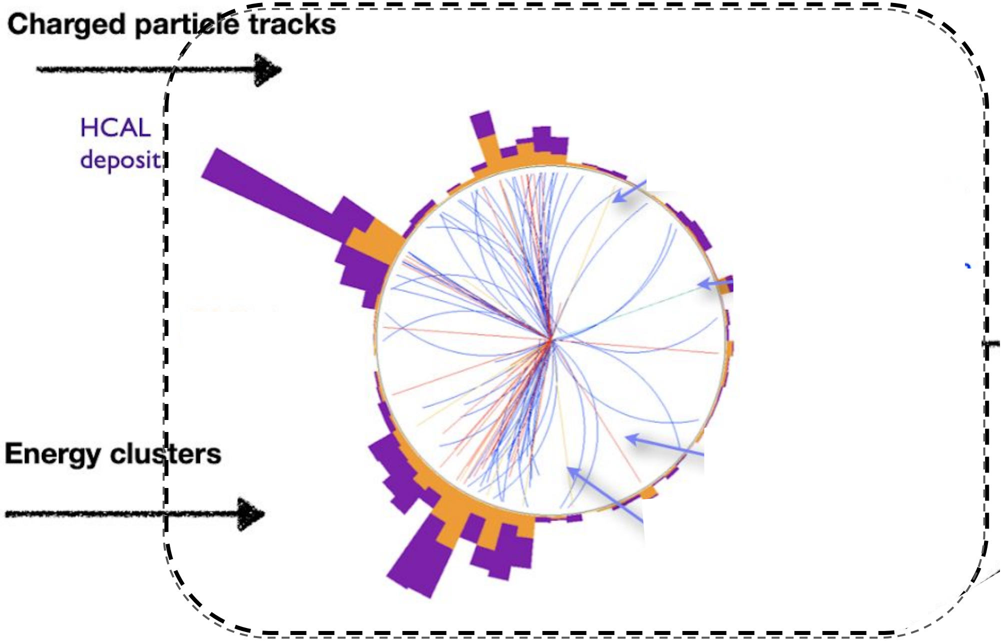


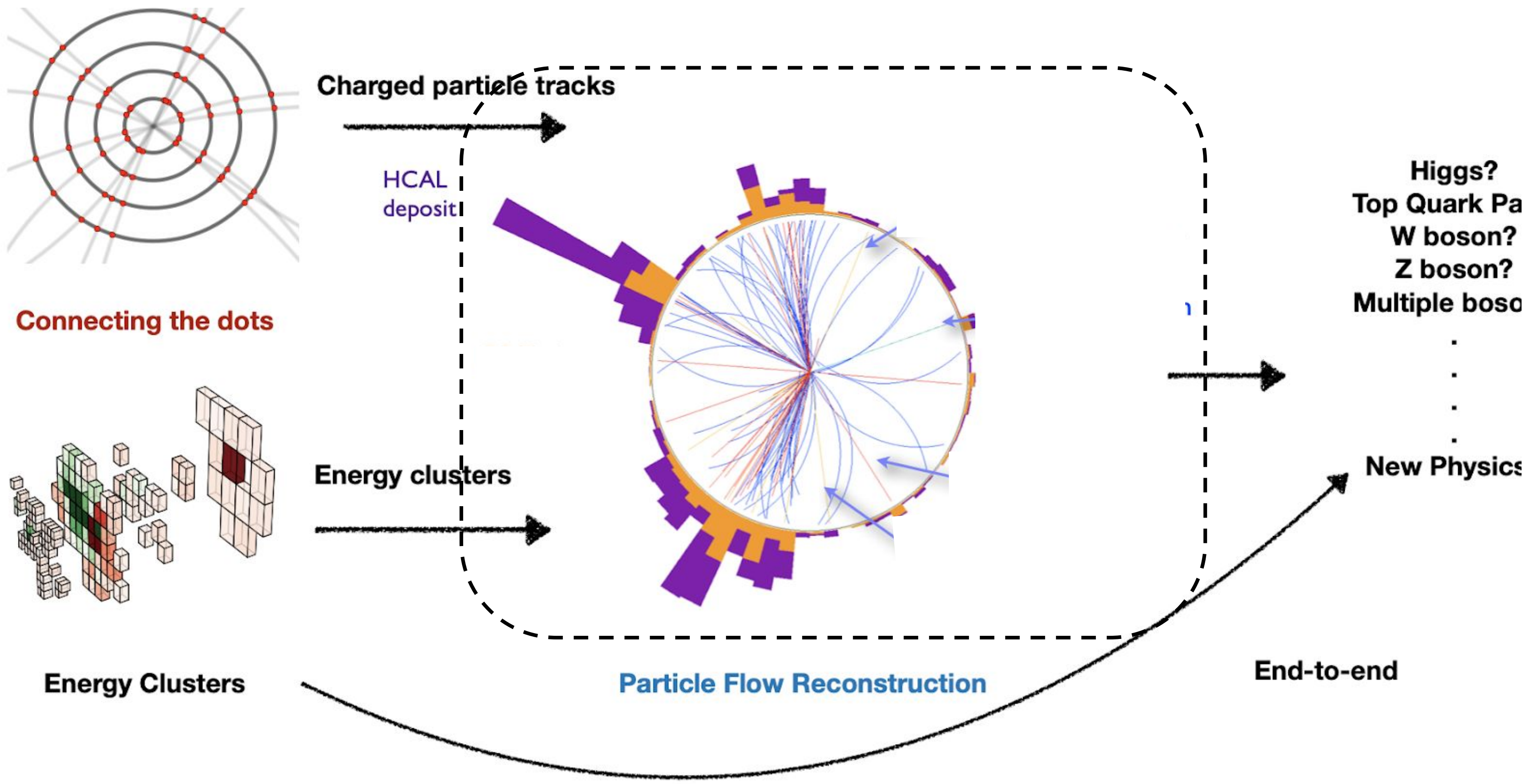


Connecting the dots

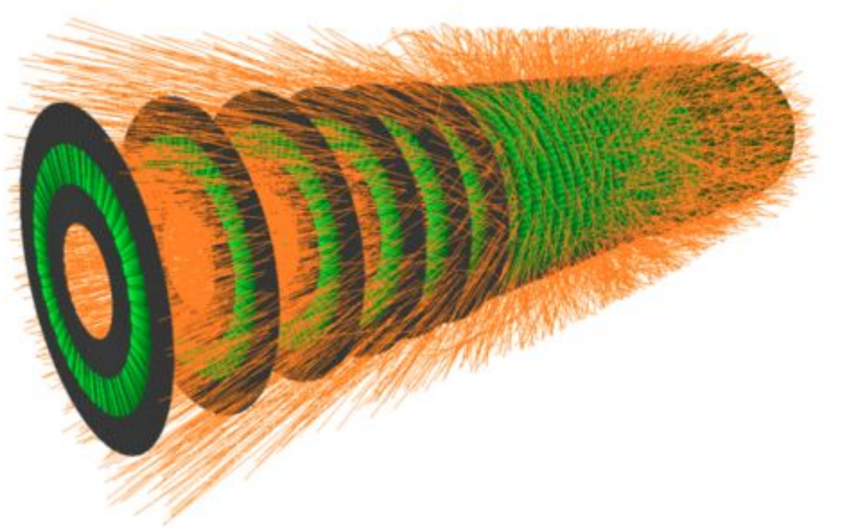


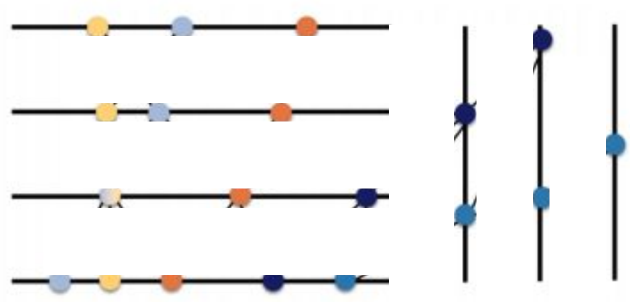
Energy Clusters




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

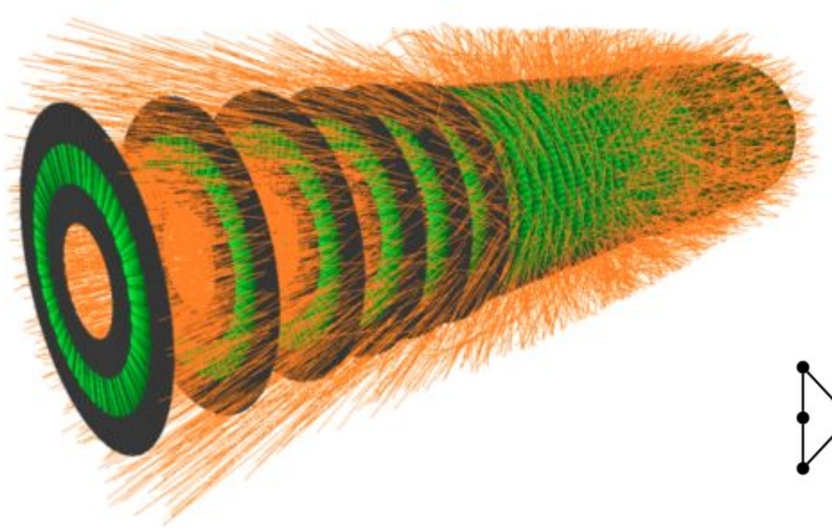




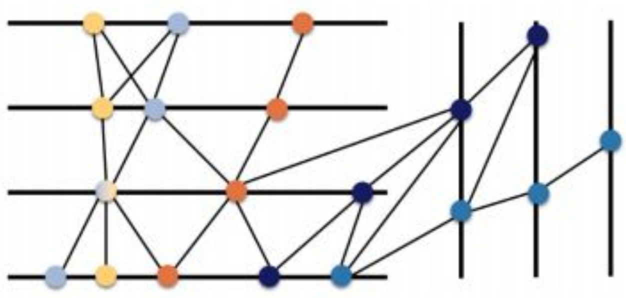
IASHEP: GNN tracking

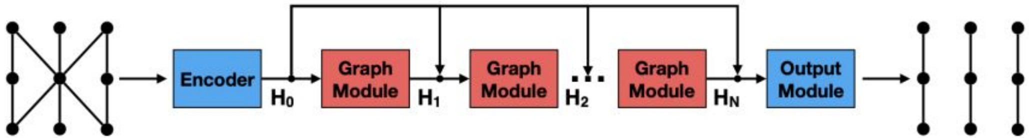
X. Ju, et. al. EPJC 81, 876 (2021)

Track Reconstruction as Graph



Graph Neural Network to identify correct edge connecting adjacent nodes



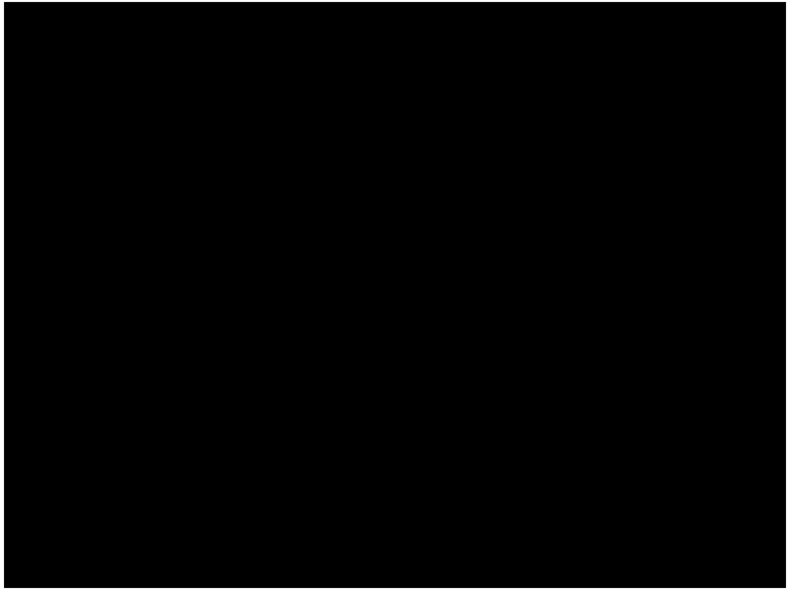


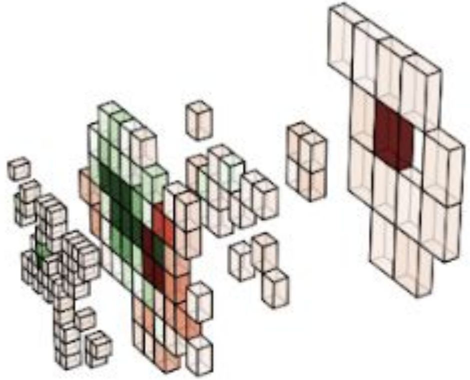
Clustering with Sparse Point Voxel Convolutional Neural Network

J. Krupa FastML'23

- Torchsparse/ Torchsparse++ (Haotian Tang, et al. @ MLSys'22)

  2.9X faster than MinkowskiEngine (NVIDIA)
  1.8X faster than SpConv (TuSimple).





Clustering with Sparse Point Voxel Convolutional Neural Network

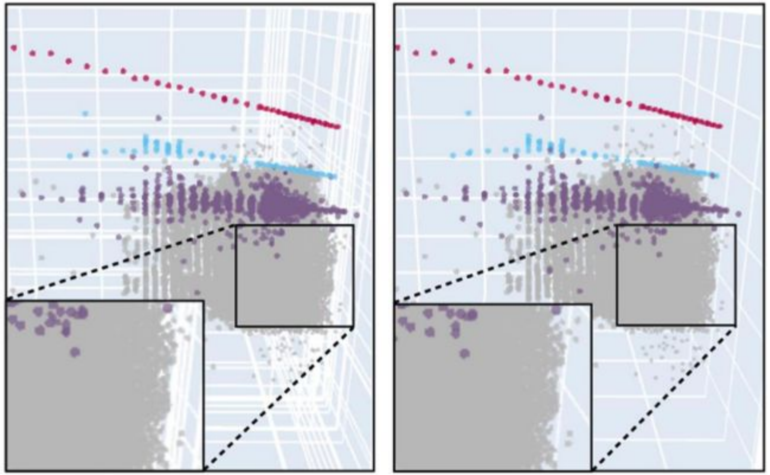
J. Krupa FastML'23

- Torchsparse/ Torchsparse++ (Haotian Tang, et al. @ MLSys'22)

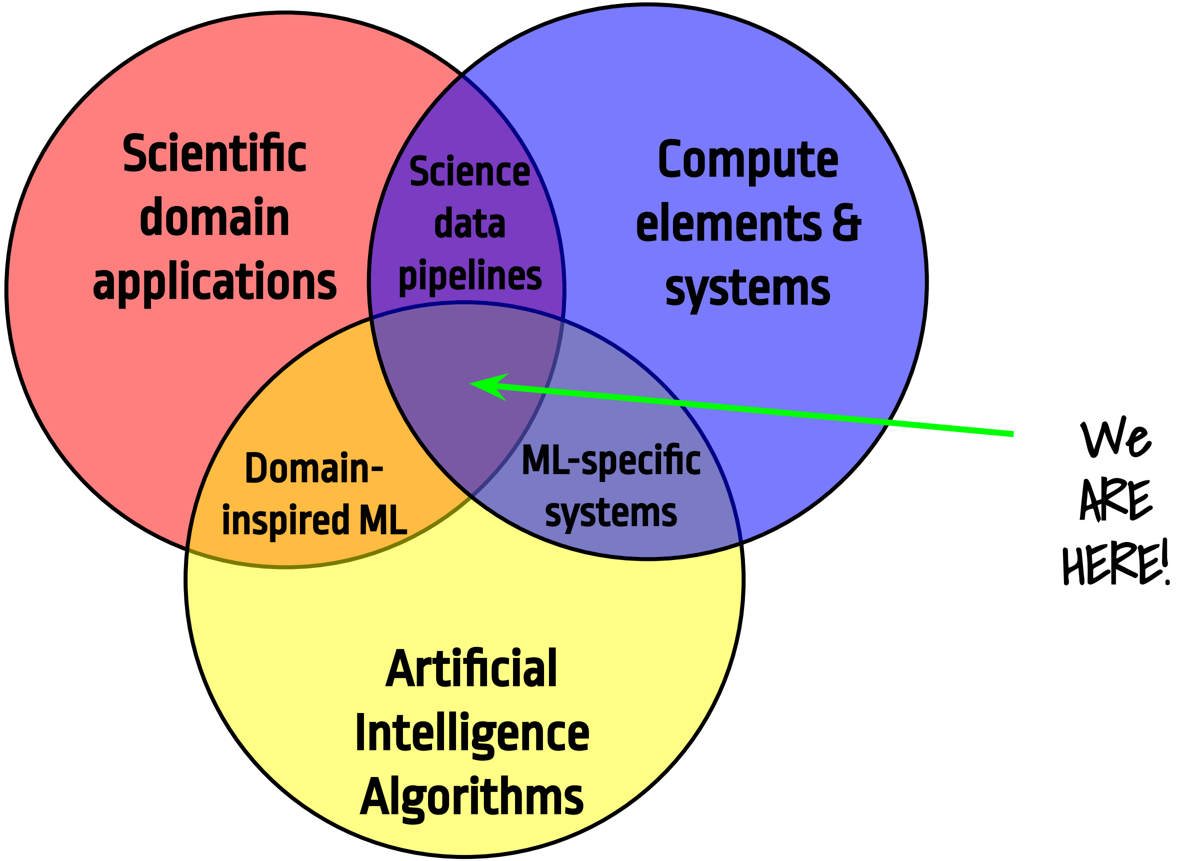
  2.9X faster than MinkowskiEngine (NVIDIA)
  1.8X faster than SpConv (TuSimple).

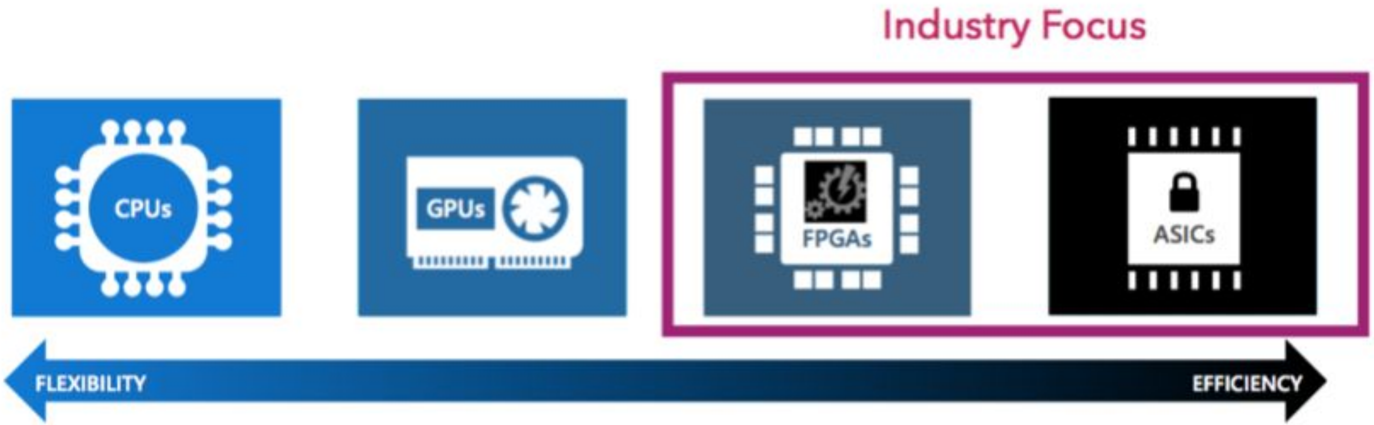
Particles are a set of 3D points and can be processed by our efficient 3D algorithms.

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



Smarter Algorithms and Faster Hardware

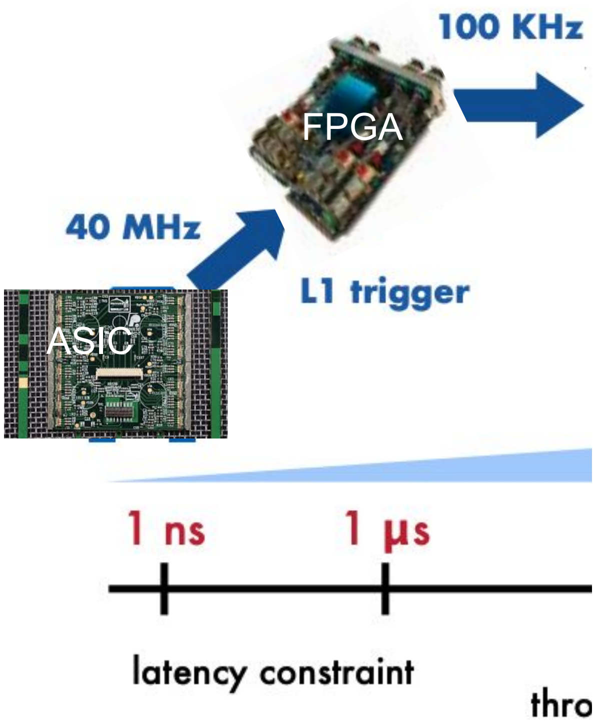



Industry Focus
CPUs
GPUs
FPGAs
ASICs
FLEXIBILITY
EFFICIENCY

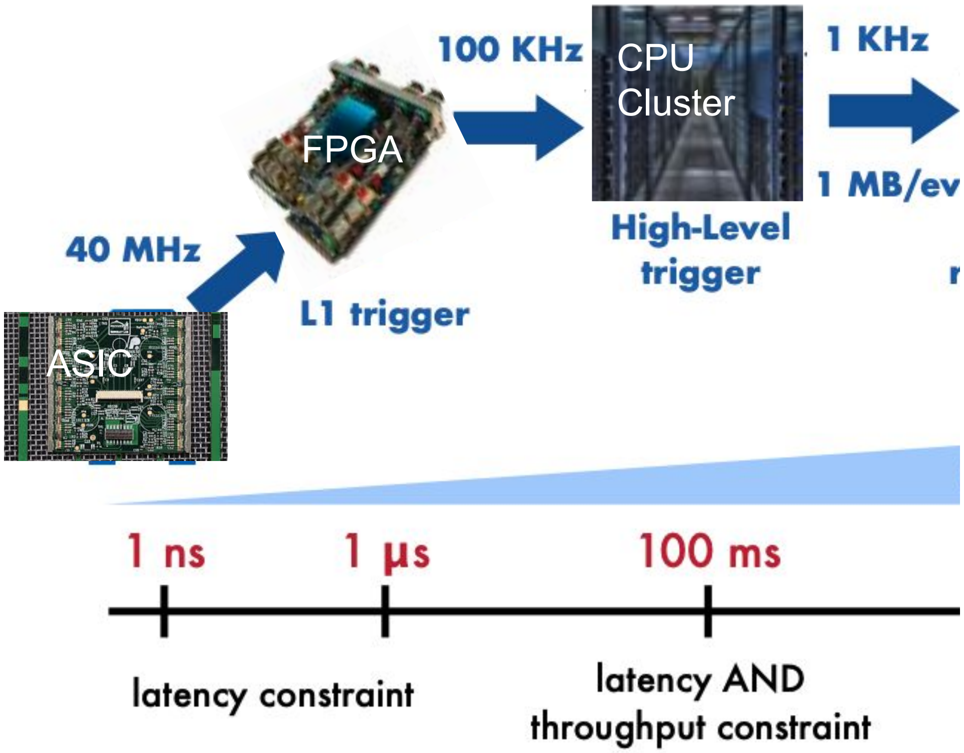
The Need for the FastML

Heterogeneous Computing

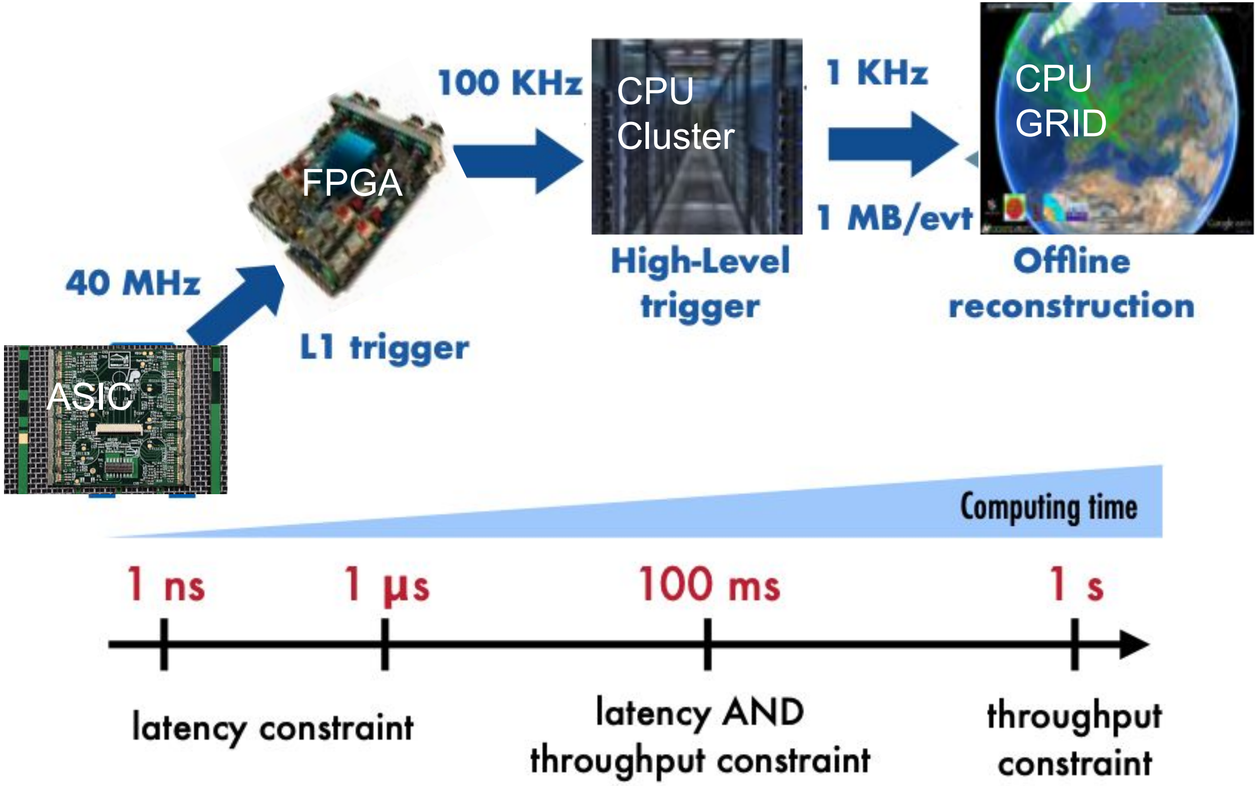
The Need for the FastML



The Need for the FastML



The Need for the FastML



The Need for the FastML

Fast Machine Learning Collaboration

Core
(HLS4ML)

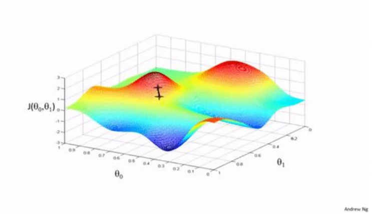
Co-processor
(SONIC)

Innovative Algorithms

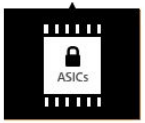
HAC Research Focus

Co-design, Design Automation

Algorithm ⟷ Hardware



hls4ml



Challenges in Algorithm Design:

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

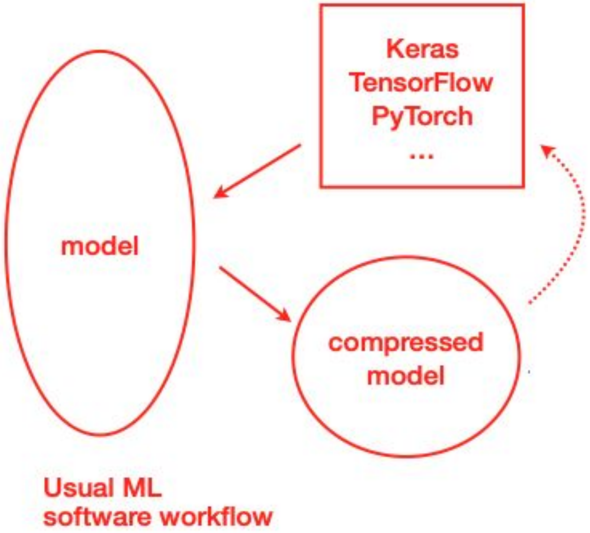
Challenges in Deployment in Hardware:

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

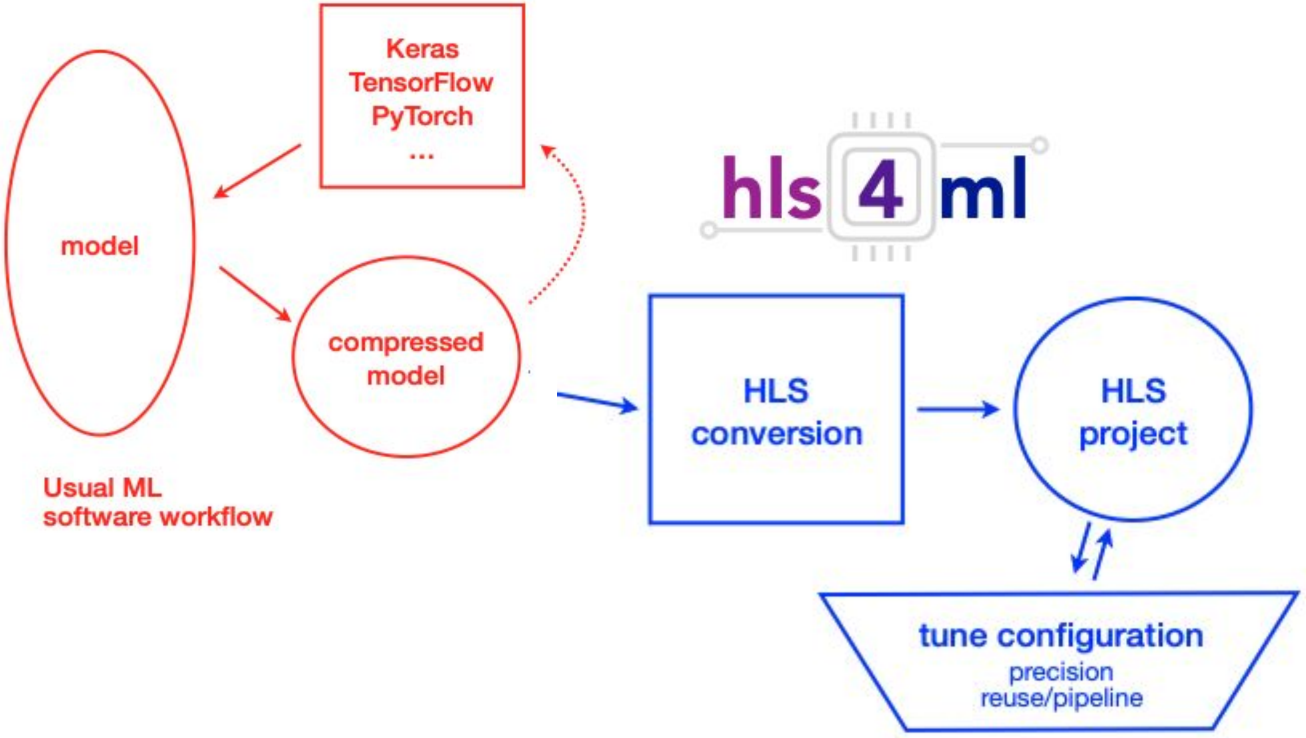
…

--> hardware design automation tools

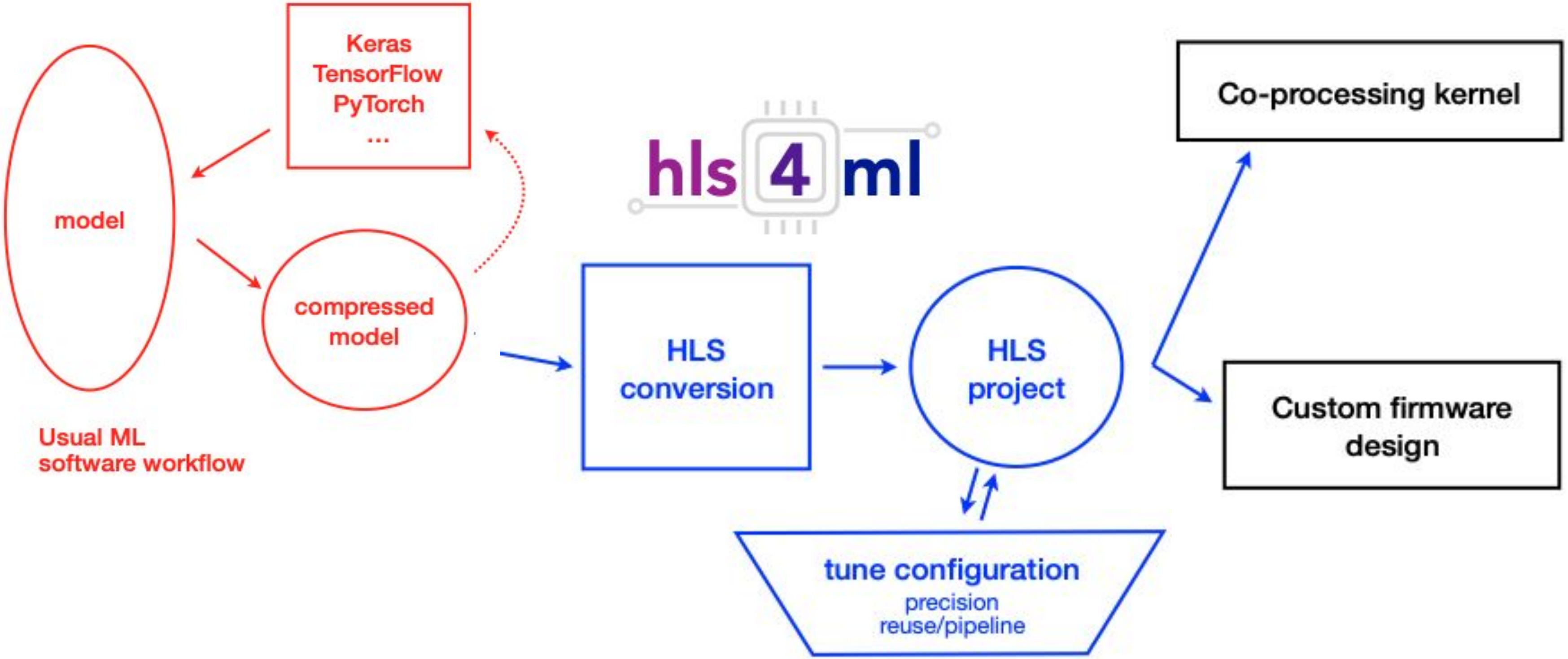
HLS4ML translating ML into FPGA firmware



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)

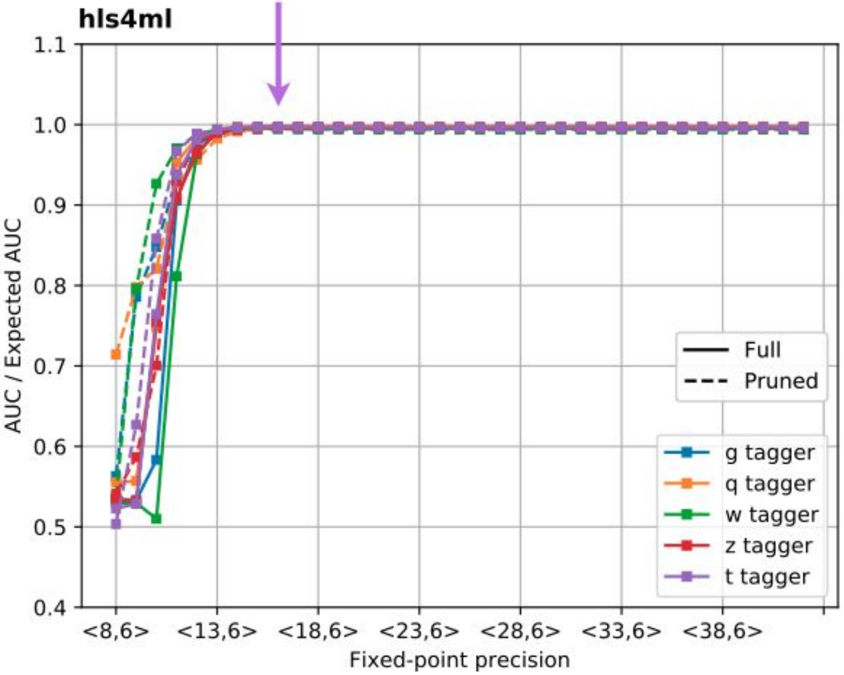
- Scan the bit width until you reach optimal performance

ap_fixed<width,integer>

0101.1011101010

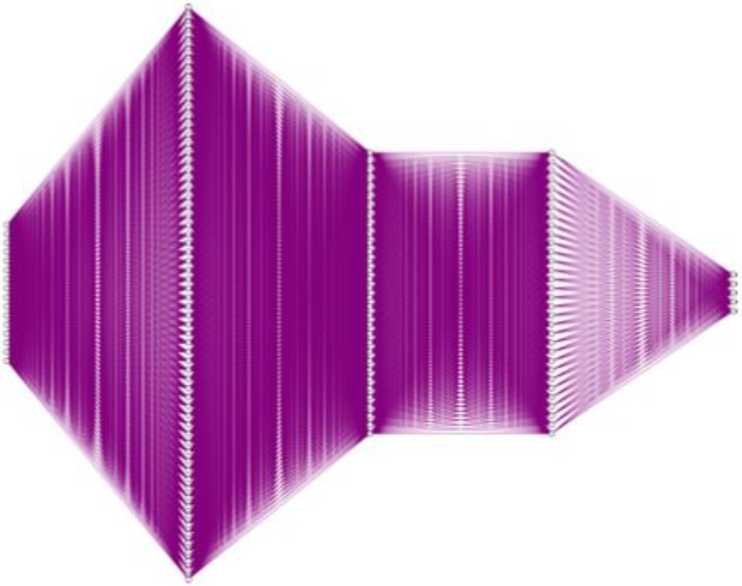
integer | fractional
width

Full performance with 16 bits



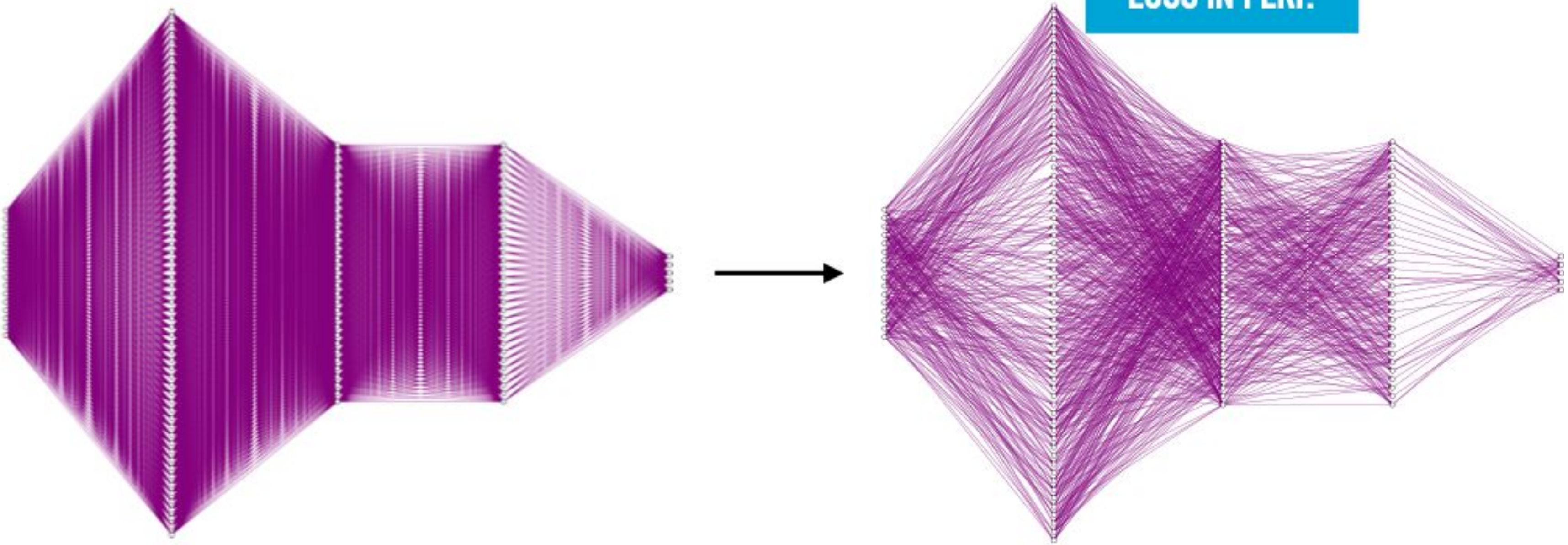
Compression

- Remove **smallest** weigh
- Iterate



Compression

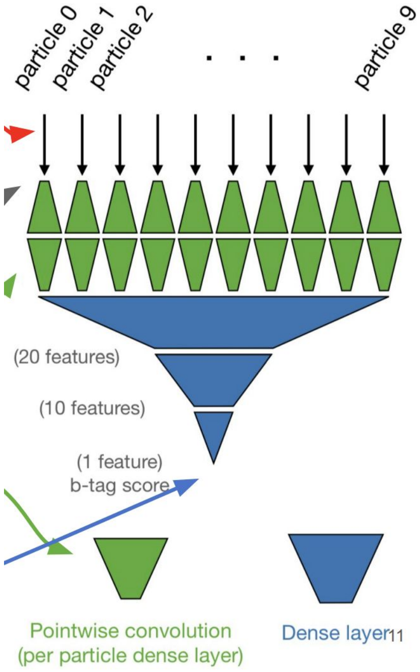
- Remove **smallest** weights
- Iterate



A.V. Aportela, FastML'23

CMS Level-1 trigger

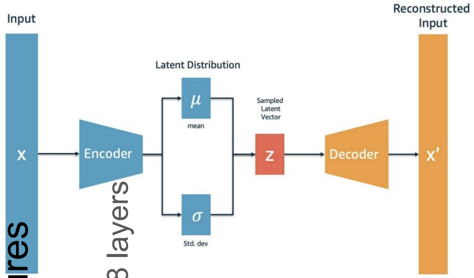
B-tagging



Autoencoder for Anomaly Detection

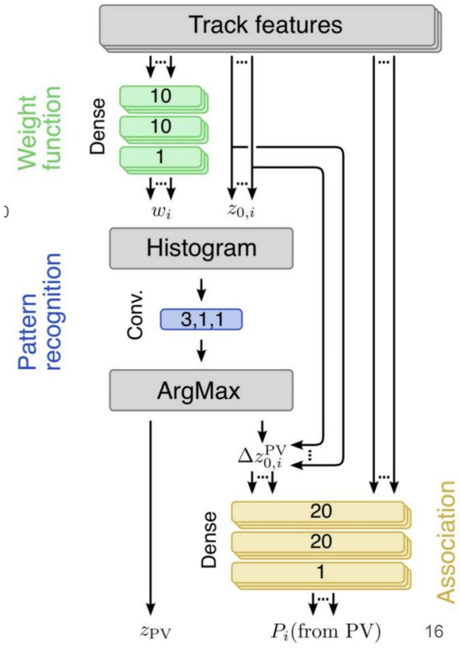
5 PUPPI jets, 18 features

DNN, 8 layers

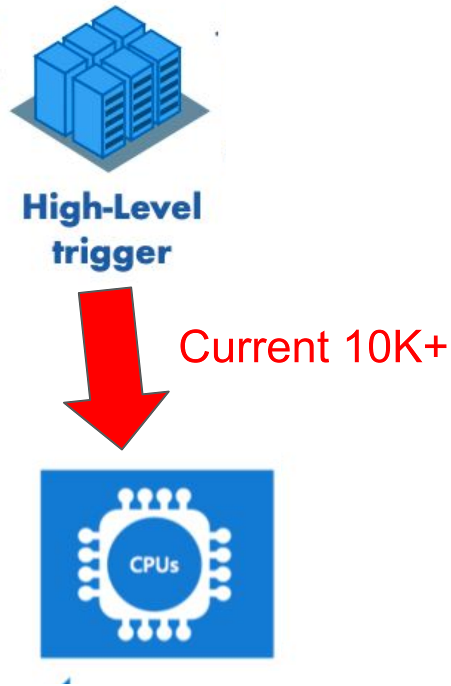


hls 4 ml

End-to-End Vertexing NN

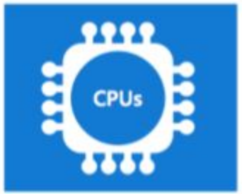


High-Level Trigger (100 KHz, 100 ms latency)



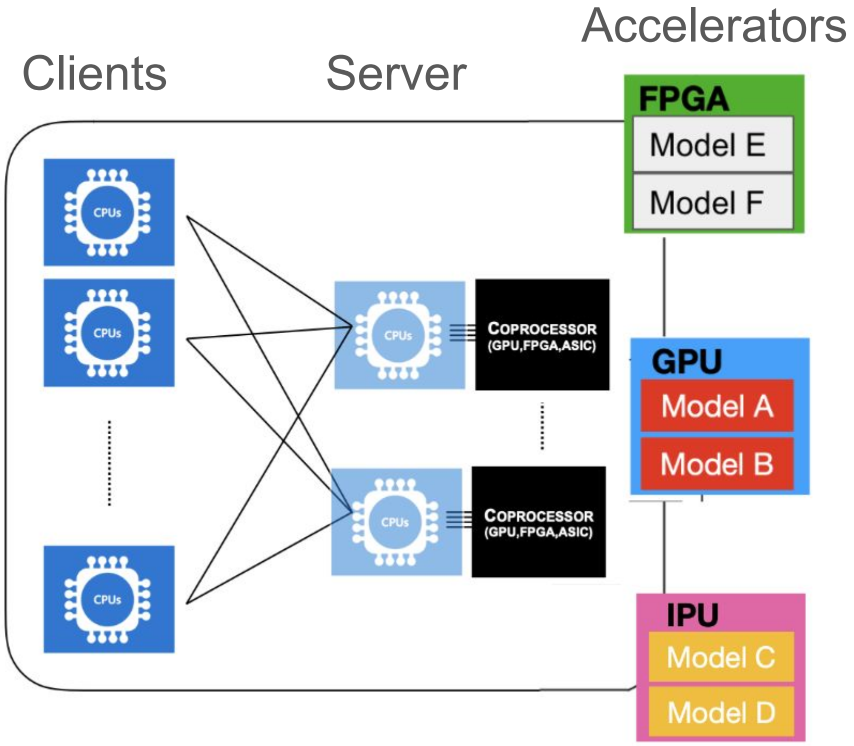
High-Level Trigger (100 KHz, 100 ms latency)

Our proposal



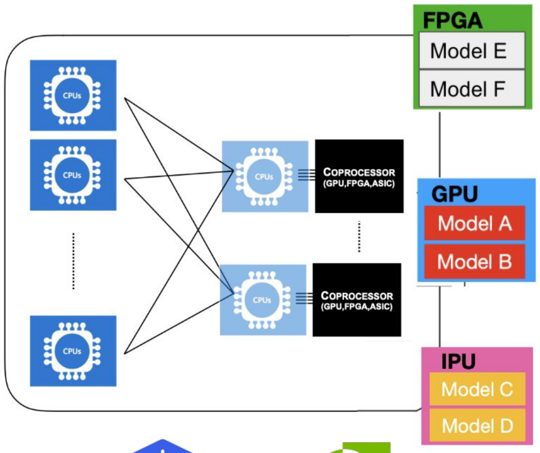
ML-as-a-Service

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



Heterogeneous system for high throughput

- A3D3 develops workflow platforms (SONIC, hermes) using standard industry tools and collaborates with IT Cloud providers & HPCs to evaluate performance



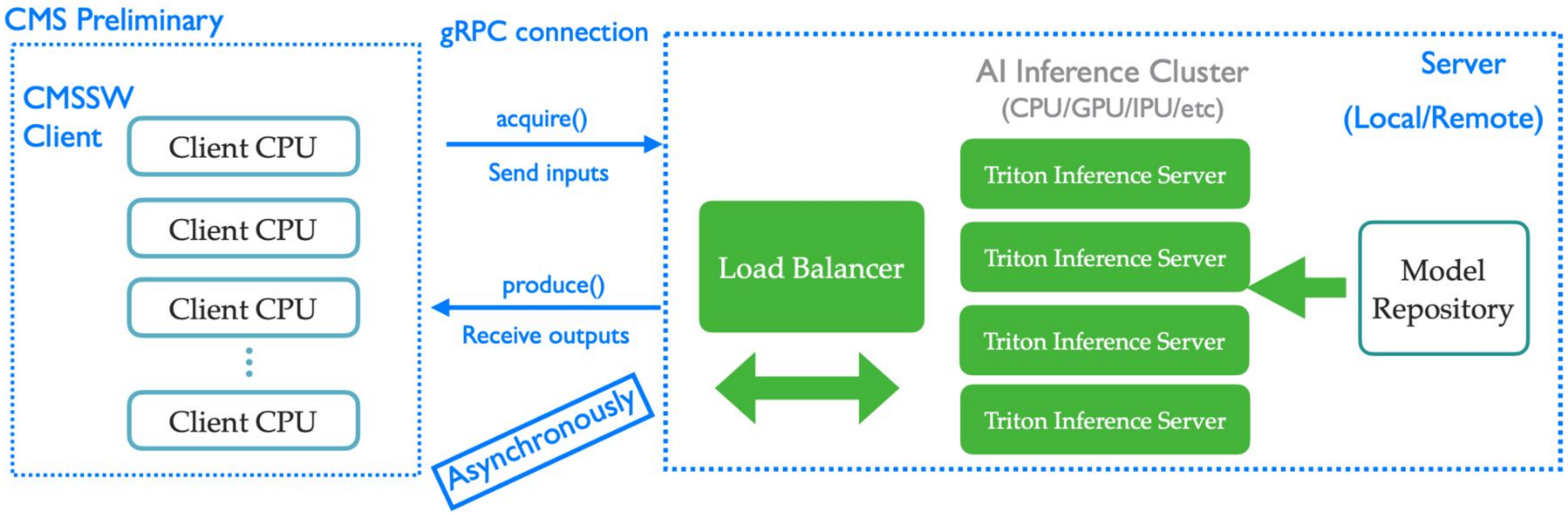
IT Cloud Providers

High Performance Computing

SONIC

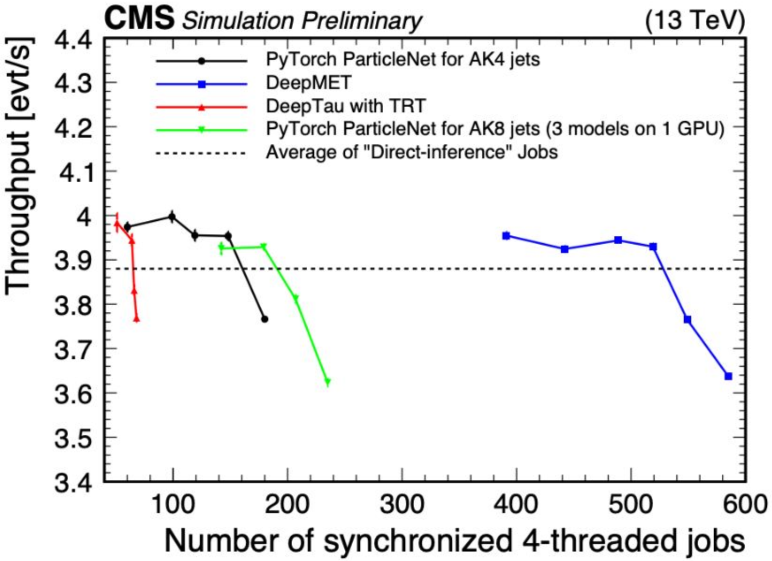
P. McComack - FastML 2023

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



Optimizing performance: CPU-to-GPU ratio

CMS mini-AOD production



- 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

Shih-Chieh Hsu

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

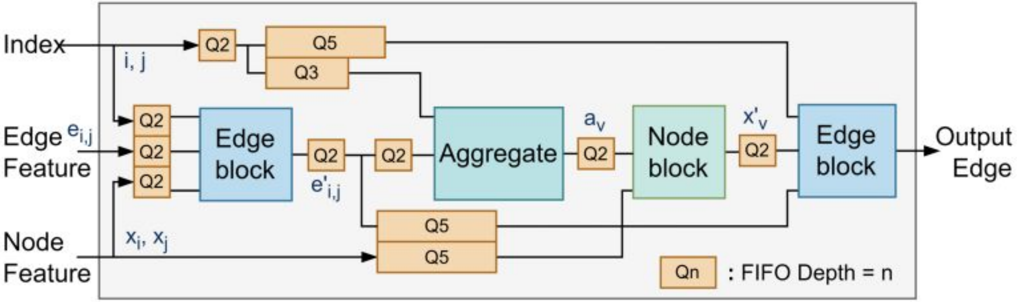
schsu@uw.edu

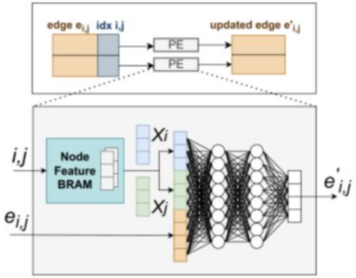
Backup

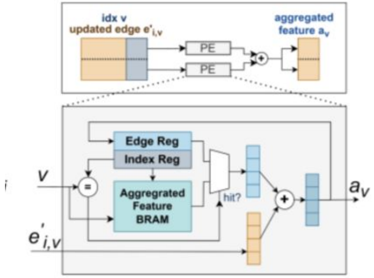
LOW LATENCY EDGE CLASSIFICATION GNN

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

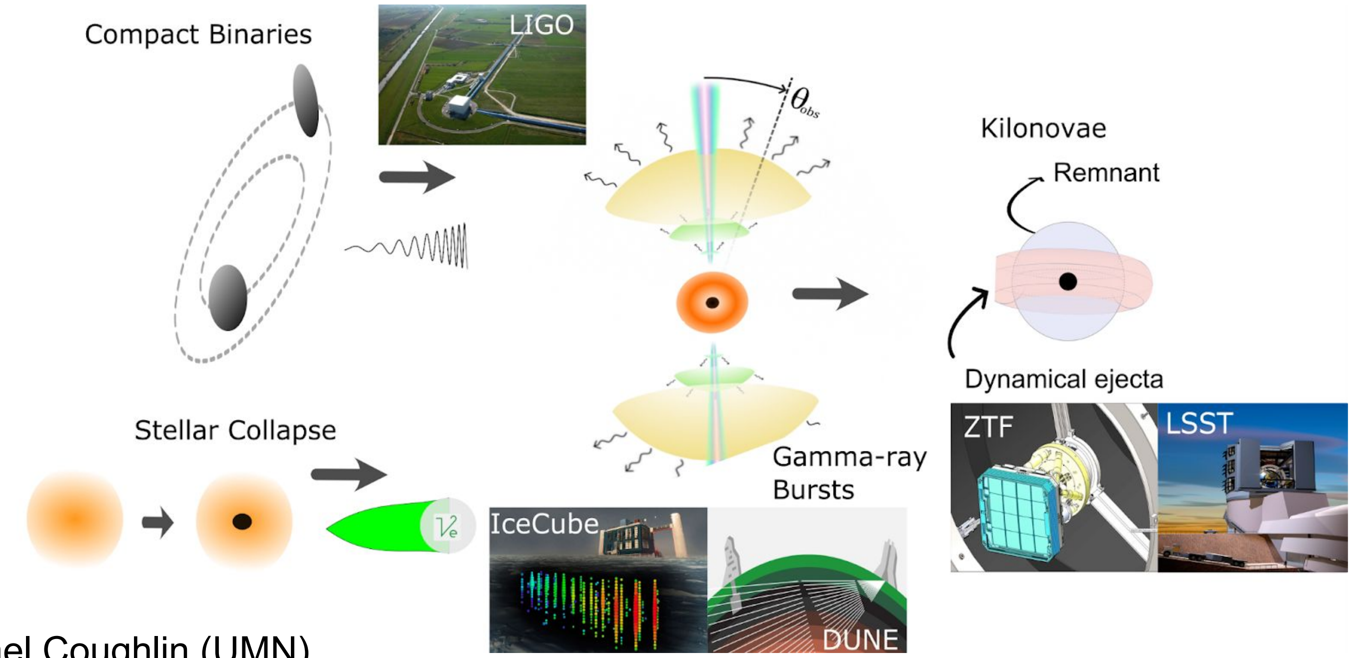
Studying SONIC at scale

- 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	—

Multi-messenger Astrophysics

- Develop and deploy software within astronomical facilities to enable discovery



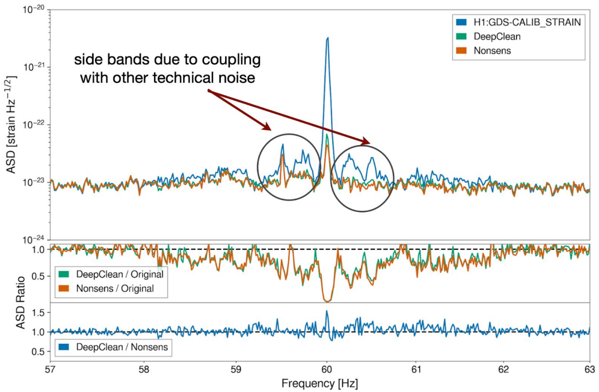
Credit: Michael Coughlin (UMN)

Gravitational Waves (LVK)

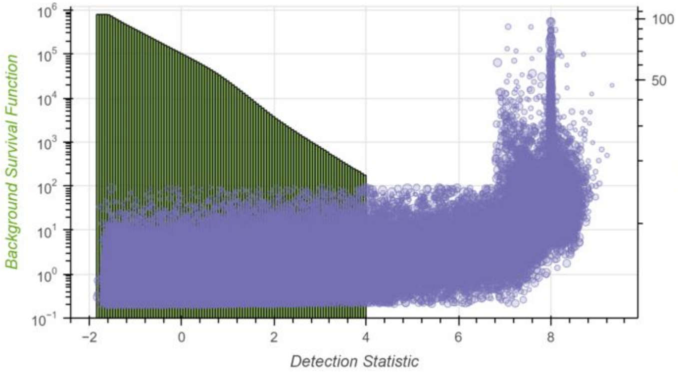
Github: ML4GW

All algorithms use our inference-as-a-service (IaaS) prototype to implement a real-time noise subtraction pipeline (DeepClean), detection (aframe/GWAK), and parameter estimation for use during the fourth observing run (O4) of LIGO-Virgo-KAGRA on dedicated hardware at the detector sites.

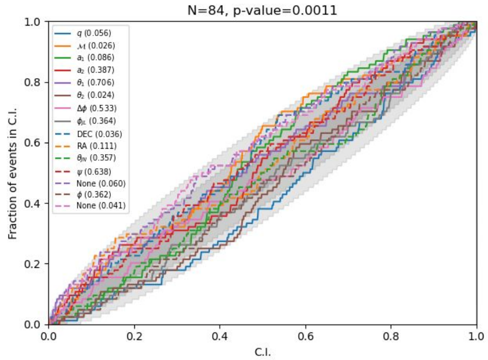
Clean the Data: DeepClean (CNN)



Detect the GWs: aframe (CNN)/GWAK (autoencoders)

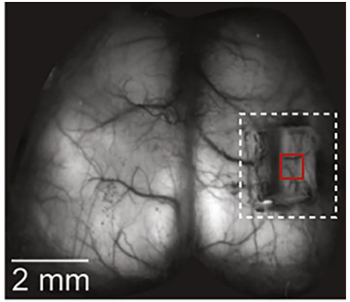


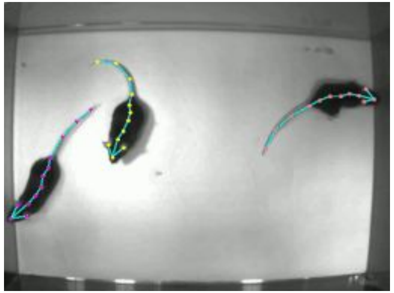
Characterize the GWs: (MAF*)



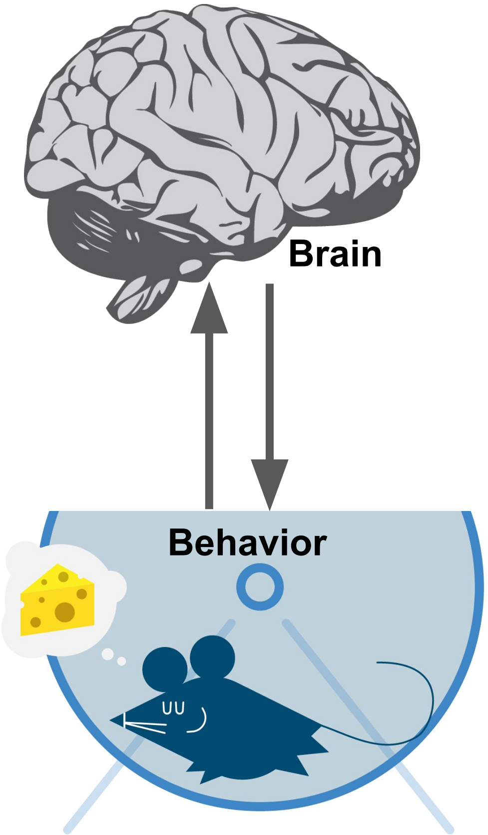
Neuroscience needs high-throughput & real-time AI

Rapid increase in number, type of measurements

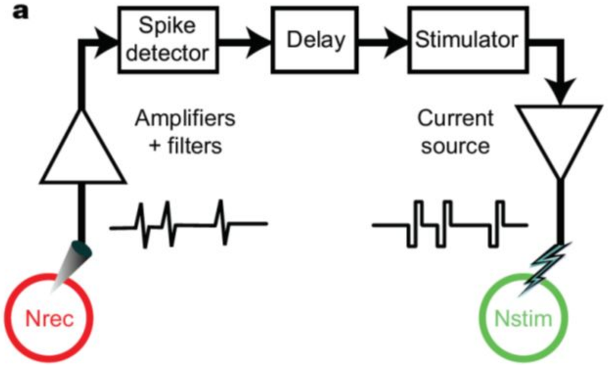




Need: data-driven discovery of relevant features, structure in data



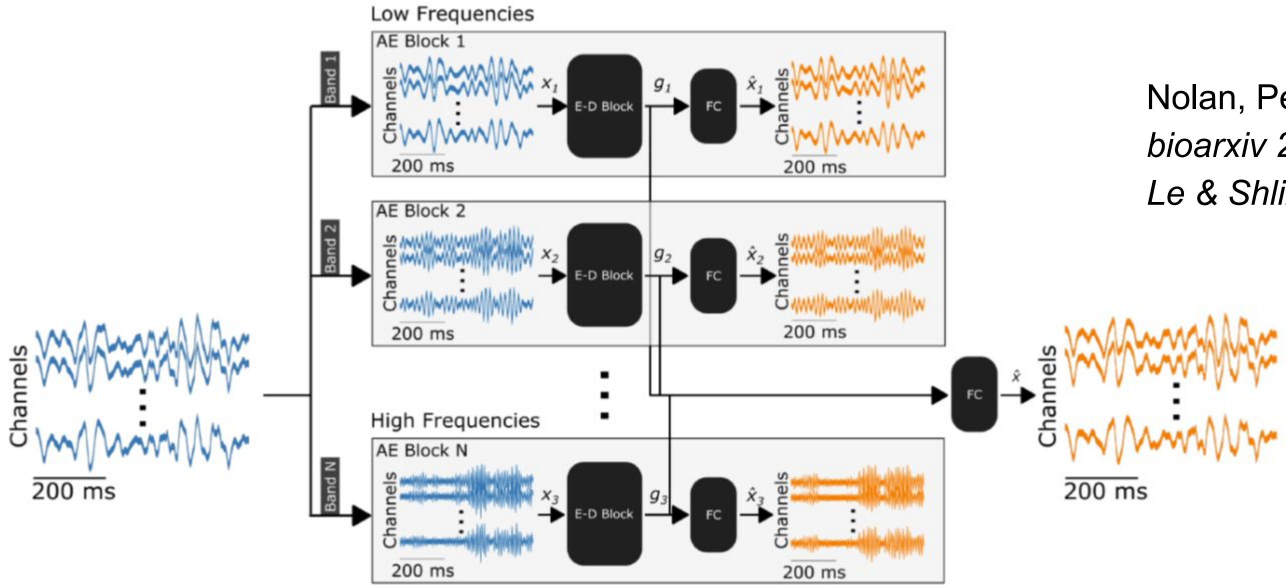
Must *perturb* the system to disentangle causality, treat disorders.



Need: low-latency algorithms (<1ms)

Improved time-series reconstruction methods

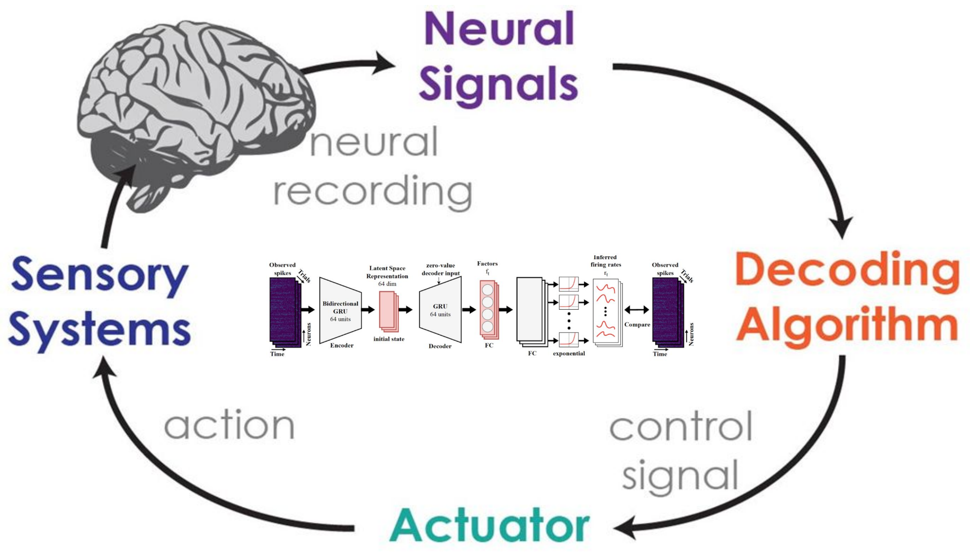
- Developed new Multi-block Recurrent Auto-Encoder (MRAE) to increase bandwidth more efficiently
- Developed Spatio-Temporal Transformer for Spiking Neural Data



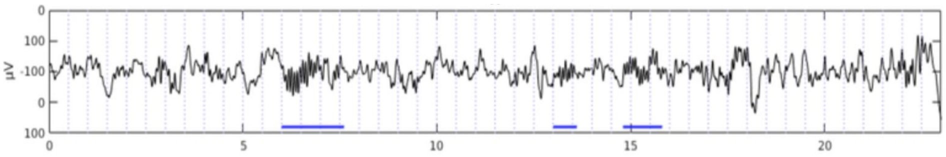
Nolan, Pesaran, Shlizerman & Orsborn, *bioarxiv 2022*
Le & Shlizerman, NeurIPS 2022

NeuroAI Integration

- A popular autoencoder model used on neural data (LFADS) in FPGA, Elham Khoda's talk



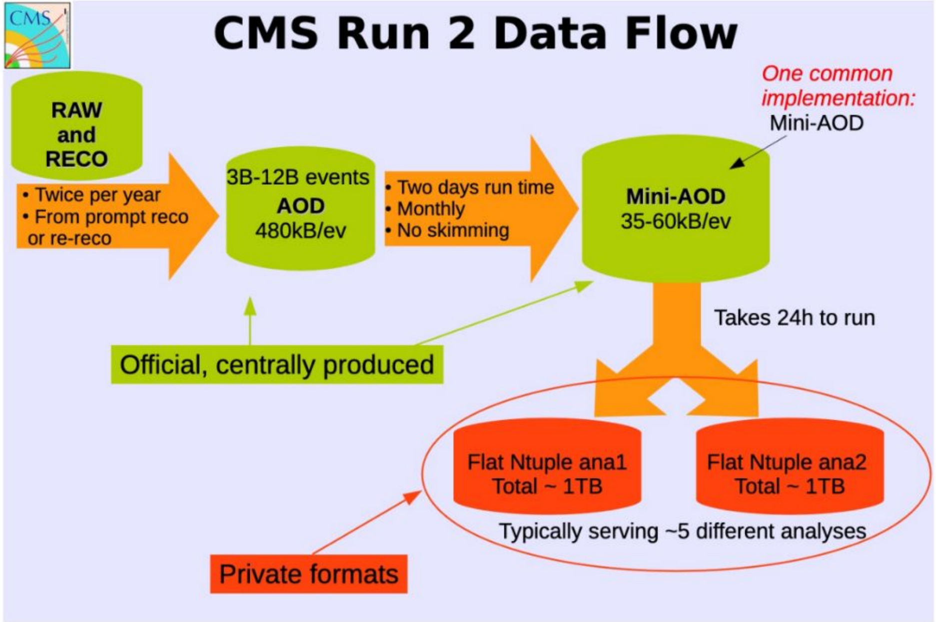
- Neuro A3D3 develops methods for reconstruction, forecasting and clustering of time-series
- Potential applications/uses:
    - Detect noise and artifacts
    - Detect rare neural events of interest (e.g., seizures, spindles, etc)



Studying SONIC at scale

[1702.04685]

- 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



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