Physics Community Needs, Tools, and Resources for Machine Learning

- Machine Learning (ML) is becoming an increasingly important component of cutting-edge physics research
  - Its computational requirements present significant challenges
- I will discuss the ML needs of the physics community e.g., across latency and throughput regimes
- Some Tools and Resources that can satisfy these needs and how these can be best utilized in the coming years

Accelerated AI Algorithms for Data-Driven Discovery

Elham E Khoda

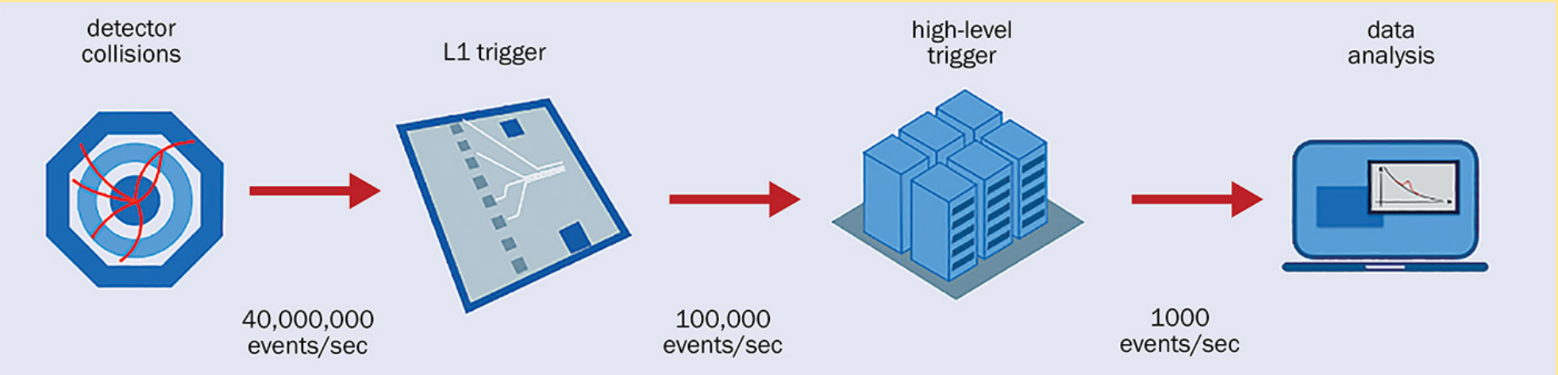
UNIVERSITY of WASHINGTON

Collaborators:
Philip Harris[1], Erik Katsavounidis[1], William Patrick McCormack[1], Dylan Rankin[1], Yongbin Feng[2], Abhijith Gandrakota[2], Christian Herwig[2], Burt Holzman[2], Kevin Pedro[2], Nhan Tran[2], Tingjun Yang[2], Jennifer Ngadiuba[2], Michael Coughlin[3], Scott Hauck[4], Shih-Chieh Hsu[4], Deming Chen[5], Mark Neubauer[5], Javier Duarte[6], Georgia Karagiorgi[7], Mia Liu[8]

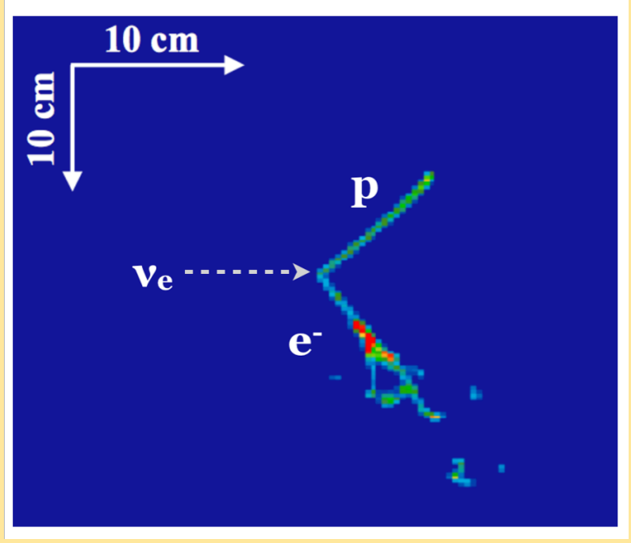
[1] MIT, [2] Fermilab, [3] University of Minnesota, [4] University of Washington, [5] University of Illinois Urbana-Champaign, [6] University of California San Diego, [7] Columbia University, [8] Perdue University

Community Needs

Collider Physics:



- ML algorithms needs to be fast ~ $\mathcal{O}(10\mu s)$ for Level-1 trigger
- Intensive use of ML algorithms in offline reconstruction and data analysis

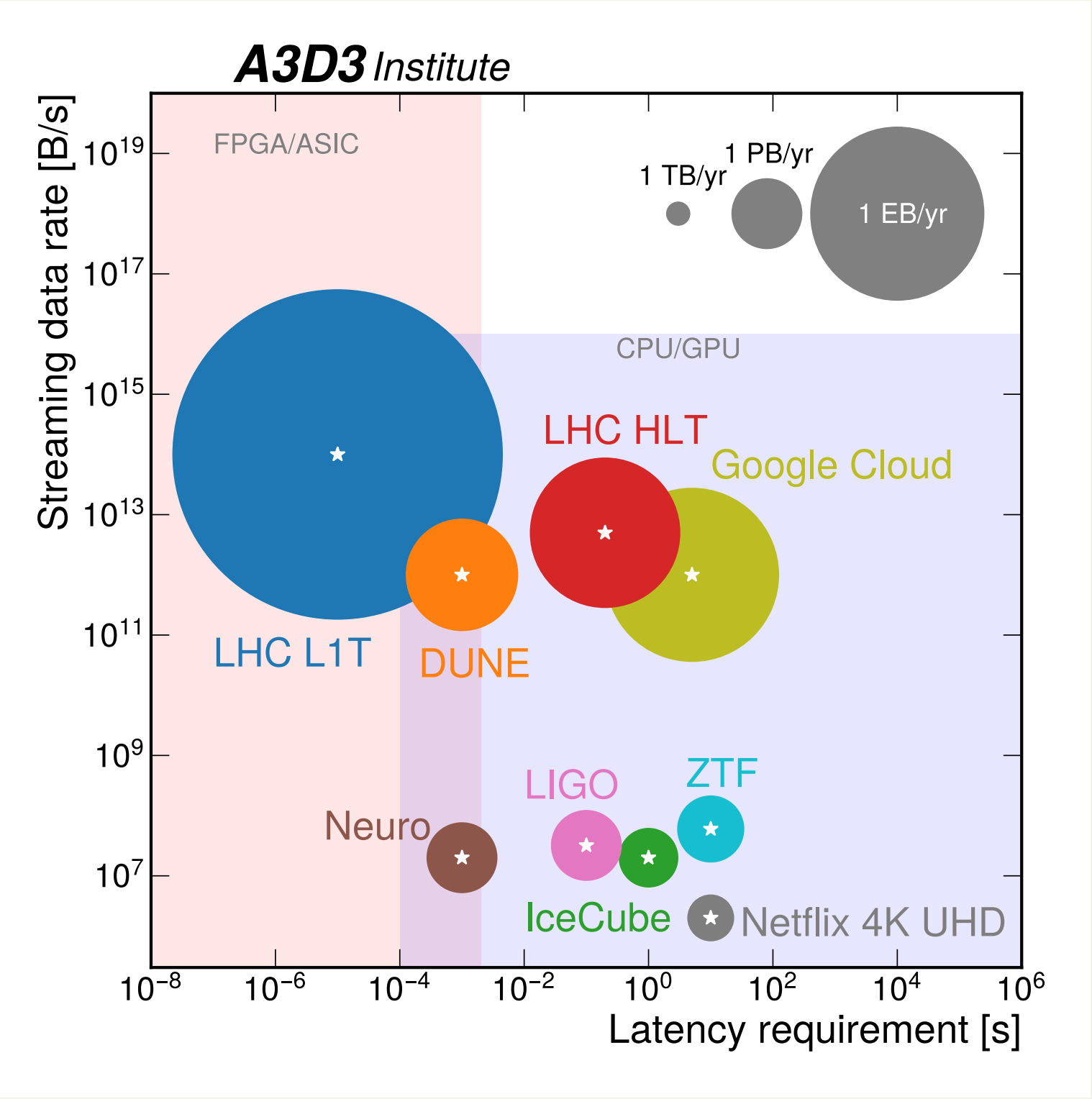


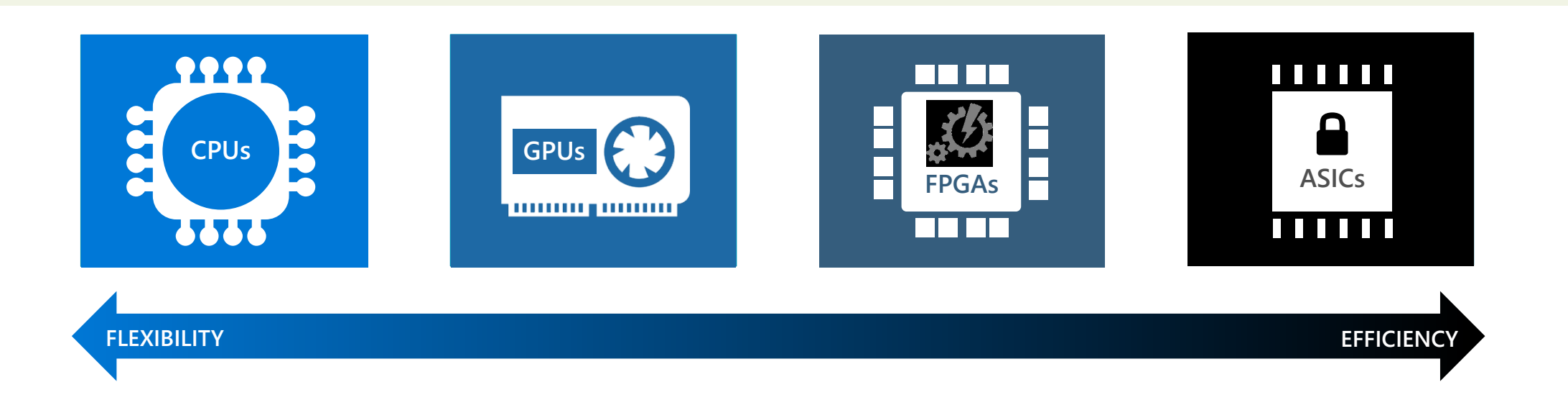
Neutrino Physics:

- High-resolution 2D projection images using new detector technology → ideal for using computer vision algorithms
- 1 TB/sec data expected in future DUNE

Astrophysics:

- Exponential growth of datasets and the interconnections between observations with all messengers
- $\mathcal{O}(1\text{s})$ data processing latency





GPU
Efficient for large matrix operations

Nvidia A100 GPUs $\mathcal{O}(100)$ times faster than Intel Xeon CPUs

FPGAs
Configurable digital circuits
Specialized multiplication units (DSPs)

AI

CPU
Widely used

Not efficient for ML inference

Future
specialized processors
- optical-based ML processors
- Tensor processing units (TPUs)
- Intelligence Processing Units (IPUs)

Software and Resources

- Open source industry tools: PyTorch, TensorFlow, ONNX, Triton e.t.c.
- Other industry tools for inference, automation, orchestration (Kubernetes)

ML-inference on FPGA: hls4ml, FINN, Vitis AI

As-a-service (aaS) computing paradigm:

- Client-server computing model. Cost-effective and performance-efficient.
- Services for Optimized Network Inference on Coprocessors (SONIC)

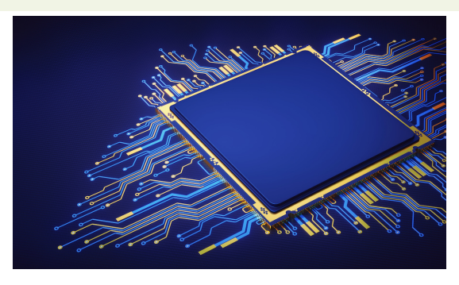
Lessons from Industry: Coordinated efforts among researchers from different domains of science engineering, and hardware systems

Cloud computing

- Flexibility in resource selection
- High cost but good for short-term development

High Performance Computing (HPC)

- Fair-share scheduler
- GPU clusters
- ORNL: 27k V100
- NERSC: 6k A100

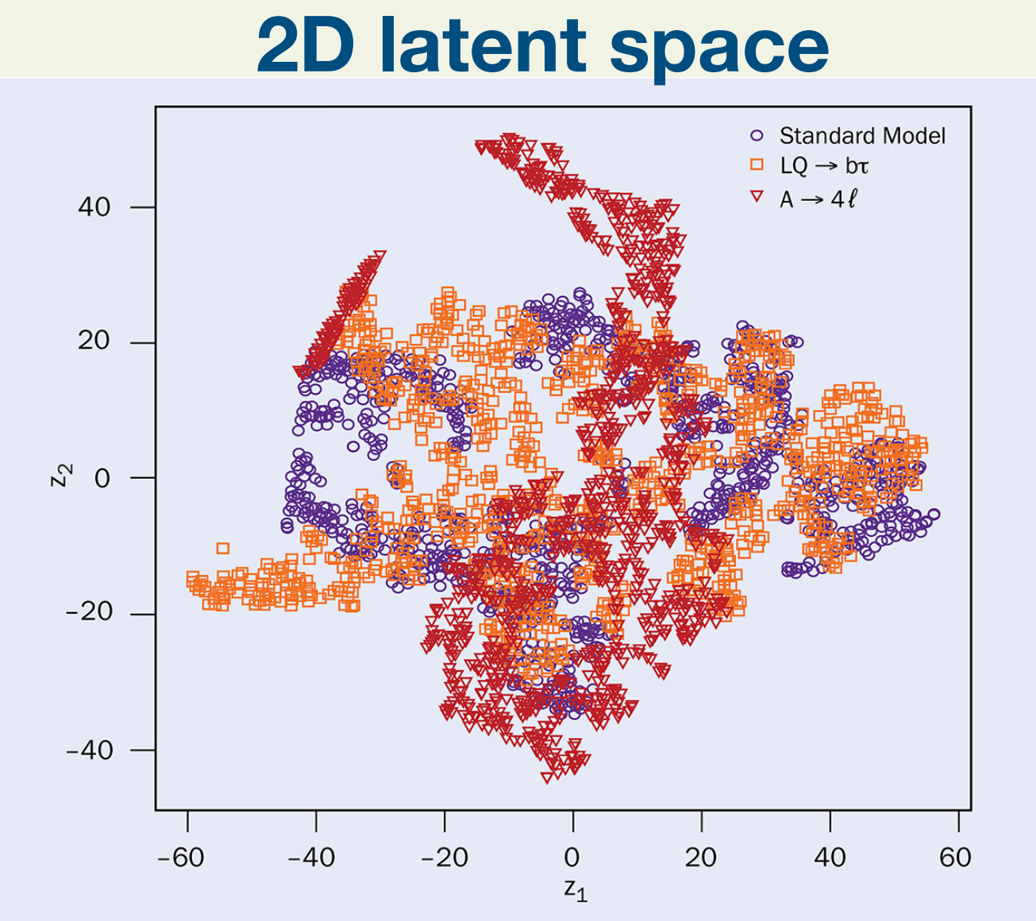


Hardware and Electronic Design Automation (EAD) tools

- Expensive industry tools
- Industry collaboration
- Open source solutions

Applications

2D latent space



Collider:

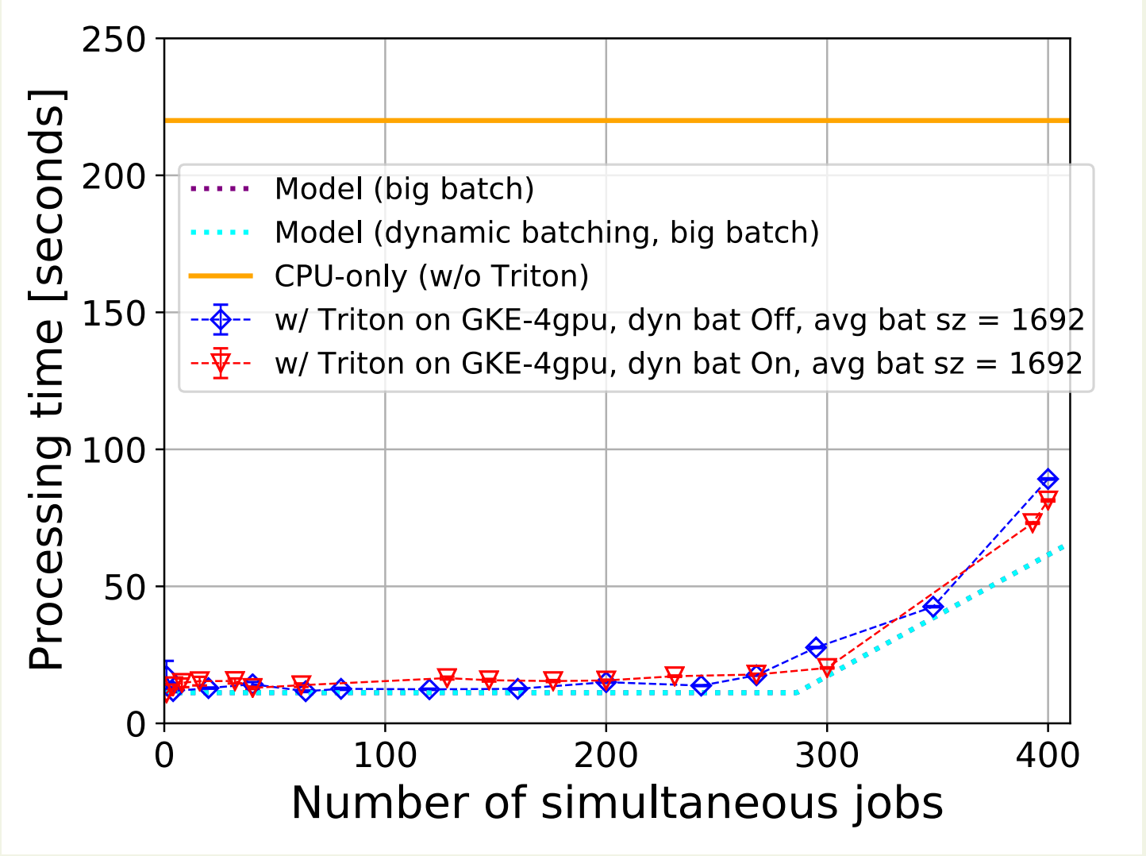
- Level-1 and High-Level Trigger
- Reconstruction and calibration of final objects or lower-level inputs like trajectories, vertices, calorimeter clusters
- Identification of long-lived particles
- Anomaly detection using Autoencoders in Run-3 and High-Luminosity LHC

Neutrinos:

- Event reconstruction with ML. The SONIC integrated framework shows a factor of 3 speed up for ProtoDUNE event reconstruction
- CNN-based event selection with FPGA for DUNE
- Future DUNE Far Detector parallel process: up to several GB information with millisecond latency.

Astrophysics:

- Denoising and astrophysical source identification.
- Gravitational-wave detection and parameter estimation.
- Attain sub-second latencies with hardware accelerators



Summary

- ML will help overcome some of the challenges of physics research in the coming decade. But ML is computationally expensive.
- Potential hardware solution: GPUs, FPGAs, and ASICs .
- Use industry tools and develop open source software for specific needs.
- Continued collaboration with industry and HPC centers will be critical.

References

1. A3D3 Institute, https://a3d3.ai/, 2022
2. P.Harris, *et al.* arXiv:2203.16255
3. J. Ngadiuba and M. Pierini, Hunting anomalies with an AI trigger, CERNCOURIER, 31 Aug 2021
4. M. Wang, *et al.* Front. Big Data 3 (2021) 604083 arXiv:2009.04509
5. MicroBooNE collaboration, Phys. Rev. D 99 (2019) 092001 arXiv 1808.07269
6. A. Gunny, D. Rankin, J. Krupa *et al.* Nat Astron 6, 529–536 (2022)