

FastML Science Benchmarks: Accelerating Real-Time Scientific Machine Learning

Javier Duarte³, Nhan Tran¹, Ben Hawks¹, Christian Herwig¹,
Jules Muhizi^{1,2}, Shvetank Prakash², Vijay Janapa Reddi²


ML Challenge
December 15th, 2022

¹Fermi National Research Laboratory

²Harvard University

³UC San Diego





FastML Science Benchmarks: Accelerating Real-Time Scientific Machine Learning

Javier Duarte³, Nhan Tran¹, Ben Hawks¹, Christian Herwig¹,
Jules Muhizi^{1,2}, Shvetank Prakash², Vijay Janapa Reddi²

ML Challenge
December 15th, 2022

¹Fermi National Research Laboratory

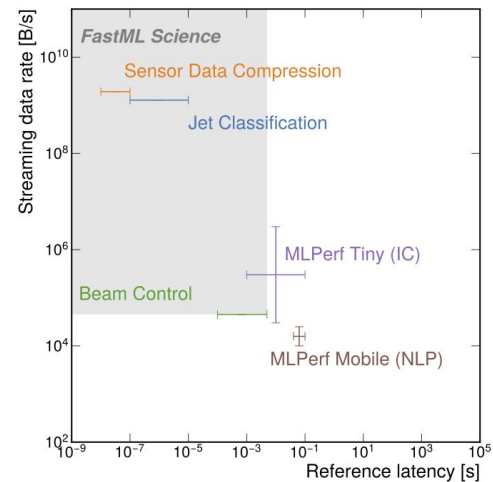
²Harvard University

³UC San Diego



Machine Learning!

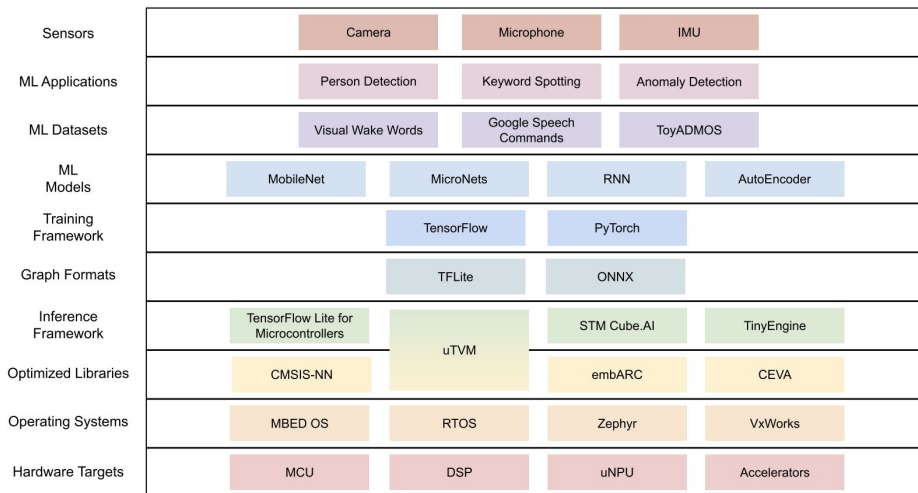
- Rise of ML as a data processing framework for large data
- DNNs have proven to be versatile at complex problems
- Scientific domain latency budget a **much** smaller than industry



ML Commons

06.16.2021 - San Francisco, CA

MLPerf Tiny Inference Benchmark



Use Case	Dataset (Input Size)	Model (TFLite Model Size)	Quality Target (Metric)
Keyword Spotting	Speech Commands (49x10)	DS-CNN (52.5 KB)	90% (Top-1)
Visual Wake Words	VWW Dataset (96x96)	MobileNetV1 (325 KB)	80% (Top-1)
Image Classification	CIFAR10 (32x32)	ResNet (96 KB)	85% (Top-1)
Anomaly Detection	ToyADMOS (5*128)	FC-AutoEncoder (270 KB)	.85 (AUC)

Table 1: MLPerf Tiny v0.5 Inference Benchmarks.

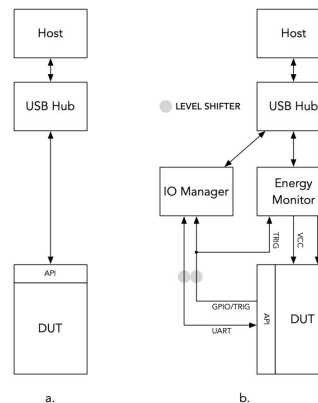


Figure 3: The two configuration modes of the benchmark framework for (a.) latency and accuracy measurement, or (b.) energy measurement.

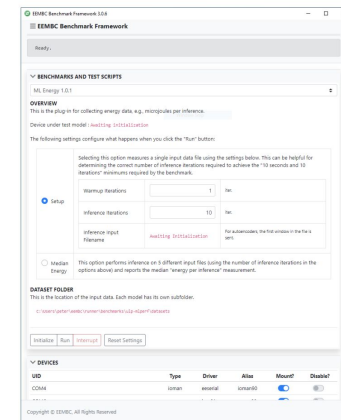


Figure 4: The graphical user interface (GUI) for the benchmark runner.



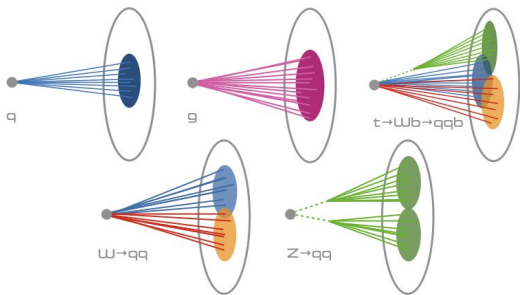
Key Challenges

How do we design a generally applicable ML benchmark using specific **scientific** applications?

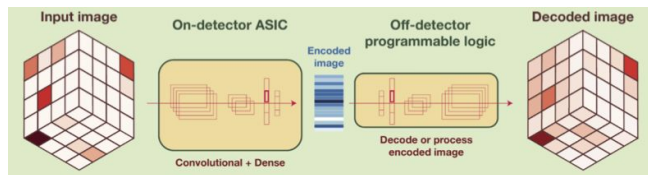
How can we design benchmark tasks to satisfy challenging **system-level** requirements while maintaining commonality?

FastML Science Benchmarks

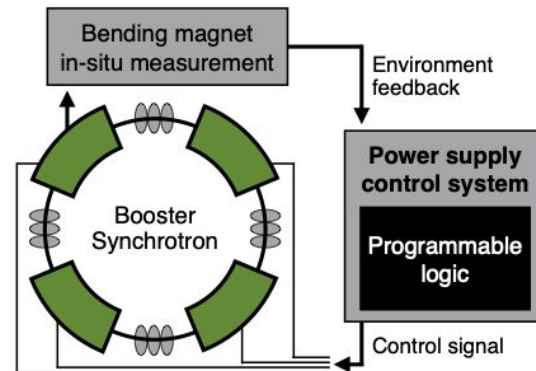
Supervised learning for rare physics event classification



Unsupervised compression of sensor data



Reinforcement learning for accelerator beam control





Agenda

- Existing Works
- Benchmark Design Philosophy
- Benchmark
 - Supervised Learning for Physics event triggering
 - Unsupervised learning for lossy compression of sensor data
 - Reinforcement learning for accelerator beam control

Existing Works

	Formalized Benchmark	Scientific Workload(s)	Edge Computing	Real-Time Constraints
FastML Science Benchmarks (this work)	✓	✓	✓	✓
SciMLBench (Thiyagalingam et al., 2021)	✓	✓	✓	×
LHC New Physics Dataset (Govorkova et al., 2021)	×	✓	✓	✓
MLPerf HPC (Farrell et al., 2021)	✓	✓	×	×
BenchCouncil AIBench HPC (BenchCouncil, 2018)	✓	✓	×	×
MLCommons Science (MLCommons, 2020)	✓	✓	×	×
ITU Modulation Classification (ITU, 2021)	×	×	✓	✓

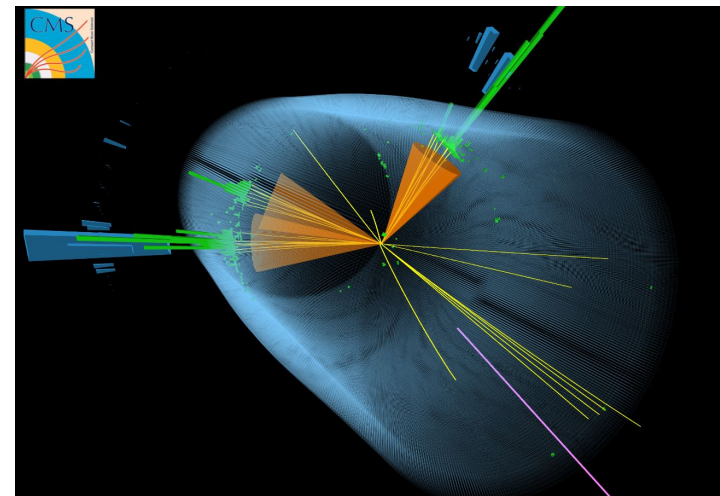


Benchmark Design Philosophy

- Applications are for the extreme edge
- Contrasting features between tasks
 - Quantization specification
 - Task specific performance metric
 - System level constraints on each benchmark
 - Latency
 - Power & Area

Jet Classification

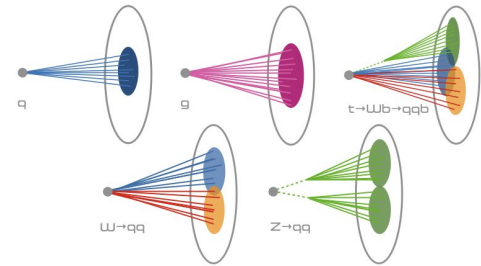
- CMS experiment observes $\sim 40\text{MHz}$ collision rate
- Data rates must be reduced by **triggering***
- Custom FPGA platforms in use as triggers at μs latency



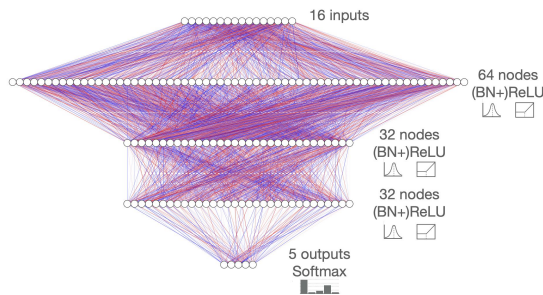
*triggering: real-time filtering
to save only certain events

Supervised Learning: Jet Classification

- Trigger only **interesting** events
- Jet tagging as supervised learning
- Baseline platform: Xilinx FPGAs within custom electronics



Baseline Model



Metrics

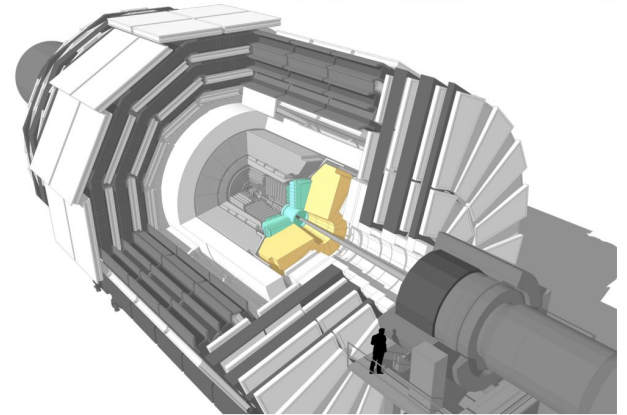
		Actual	
		Neg(0)	Pos(1)
Predicted	Neg(0)		
	Pos(1)		

Constraints

Input precision	Pipeline interval	Real-time latency
16b	150ns	1 μ s

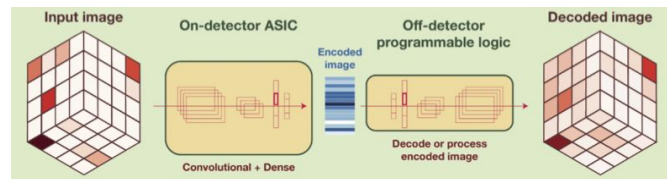
Sensor Data Compression

- High Granularity Calorimeter imaging detector produces large data
- Big data challenge posed by need to compression large for decision making
- Generalizable task to on-detector sensor data compression

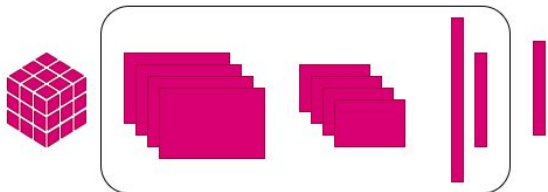


Unsupervised Learning: Irregular Sensor Data Compression

- Compress data for downstream processing
- Unsupervised data compression
- Reference platform: ASIC compresses sensor data



Baseline Model



Metrics

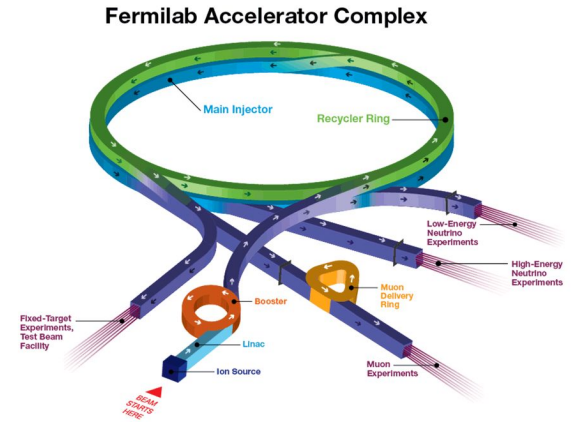
Similarity score using magnitude and distance of sensor data output

Constraints

Input precision	Pipeline interval	Real-time latency
9b	25ns	100 ns

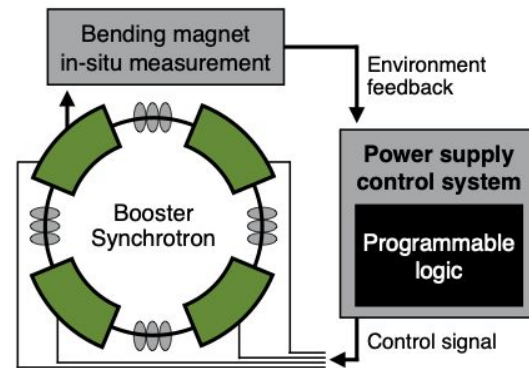
Beam Control

- Intense particle beam control is useful in general scientific work (optics, cancer therapy ...etc)
- Precise control key to operation at DOE facilities
- Control systems problem:
 - Fermilab booster synchrotron: drive particle beam intensity and reduce beam intensity loss

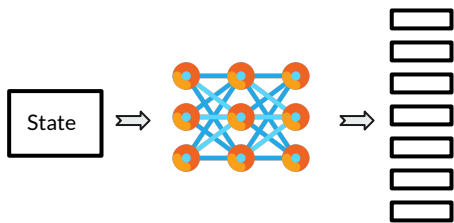


Reinforcement Learning: Beam Control

- Proton beam control critical to physics experiment
- Controls: reduce beam intensity loss
- Benchmark platform: Arria10 SoC



Baseline Model



Metrics

Difference in target and measured beam intensities

Constraints

Input precision	Pipeline interval	Real-time latency
32b	5ms	5ms



Review Key Challenges

- How do we design a generally applicable ML benchmark using specific **scientific** applications?
 - Abstract away scientific complexity where applicable
 - Allow for new additions from scientific domain experts
 -
- How can we design benchmark tasks to satisfy challenging **system-level** scientific while maintaining commonality?
 - Features such as quantization are innate to data at the edge
 - We vary our platforms from ASIC to FPGA
 - Take inspiration from MLPerf Tiny™ to standardize platforms



Outlook

- Dennard scaling and Moore's will become more and more apparent
- Edge computing and processing exceedingly crucial
- Motivate other science domain experts to bring more applications

Visit the repo and checkout the paper!

The screenshot shows the GitHub repository page for 'fastmachinelearning/fastml-science'. The repository is public and has 3 branches and 2 tags. The main branch is selected. The repository contains several folders and files: 'beam-control' (update readme and include quantized MLP, 23 days ago), 'jet-classify' (Update README.md (#4), 8 months ago), 'sensor-data-compression' (sensor data compression repo, 2 months ago), '.gitignore' (Initial commit, 8 months ago), 'LICENSE' (Create LICENSE (#5), 8 months ago), and 'README.md' (Update README.md, 8 months ago). A pull request #9 is also visible. At the bottom, there is a section titled 'Fast Machine Learning Science Benchmarks'.

The screenshot shows the arXiv paper page for 'FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning'. The paper is in the 'Computer Science > Machine Learning' category and was submitted on 16 Jul 2022. The authors are Javier Duarte, Nhan Tran, Ben Hawks, Christian Herwig, Jules Muhizi, Shvetank Prakash, and Vijay Janapa Reddi. The abstract states: 'Applications of machine learning (ML) are growing by the day for many unique and challenging scientific applications. However, a crucial challenge facing these applications is their need for ultra low-latency and on-detector ML capabilities. Given the slowdown in Moore's law and Dennard scaling, coupled with the rapid advances in scientific instrumentation that is resulting in growing data rates, there is a need for ultra-fast ML at the extreme edge. Fast ML at the edge is essential for reducing and filtering scientific data in real-time to accelerate science experimentation and enable more profound insights. To accelerate real-time scientific edge ML hardware and software solutions, we need well-constrained benchmark tasks with enough specifications to be generically applicable and accessible. These benchmarks can guide the design of future edge ML hardware for scientific applications capable of meeting the nanosecond and microsecond level latency requirements. To this end, we present an initial set of scientific ML benchmarks, covering a variety of ML and embedded system techniques.' The paper has 9 pages, 4 figures, and is a contribution to the 3rd Workshop on Benchmarking Machine Learning Workloads on Emerging Hardware (MLBench) at the 5th Conference on Machine Learning and Systems (MLSys). The report number is FERMLAB-CONF-22-514-PPD-SCD. The paper can be cited as: arXiv:2207.07958 [cs.LG] (or arXiv:2207.07958v1 [cs.LG] for this version) with the DOI: https://doi.org/10.48550/arXiv.2207.07958. The submission history shows it was submitted by Javier Duarte on 16 Jul 2022 at 14:30:15 UTC (394 KB).

Correspond with us:
Main contact - Javier Duarte jduarte@physics.ucsd.edu
Presenter - Jules Muhizi - jmuhizi@fnal.gov