FastML Science Benchmarks: Accelerating Scientific Edge ML

J. Duarte[1], N. Tran[2], B. Hawks[2], **C. Herwig**[2], J. Muhizi[3], S. Prakash[3], V. Janapa Reddi[3]

Introduction and motivation

In pursuit of scientific discovery, experiments constantly evolve to probe physical systems at smaller spatial resolutions and shorter timescales. Order-of-magnitude advancements have lead to an explosion in data volumes and richness, requiring novel methods of **real-time processing on the edge**, where selection and distillation of the complete data increasingly occurs before transmission off-detector.

Machine learning (ML) has emerged as a powerful and flexible framework to process large quantities of information, using algorithms that learn directly from the data. Deep neutral networks in particular have proven capable of solving complex problems across a wide range of scientific domains.

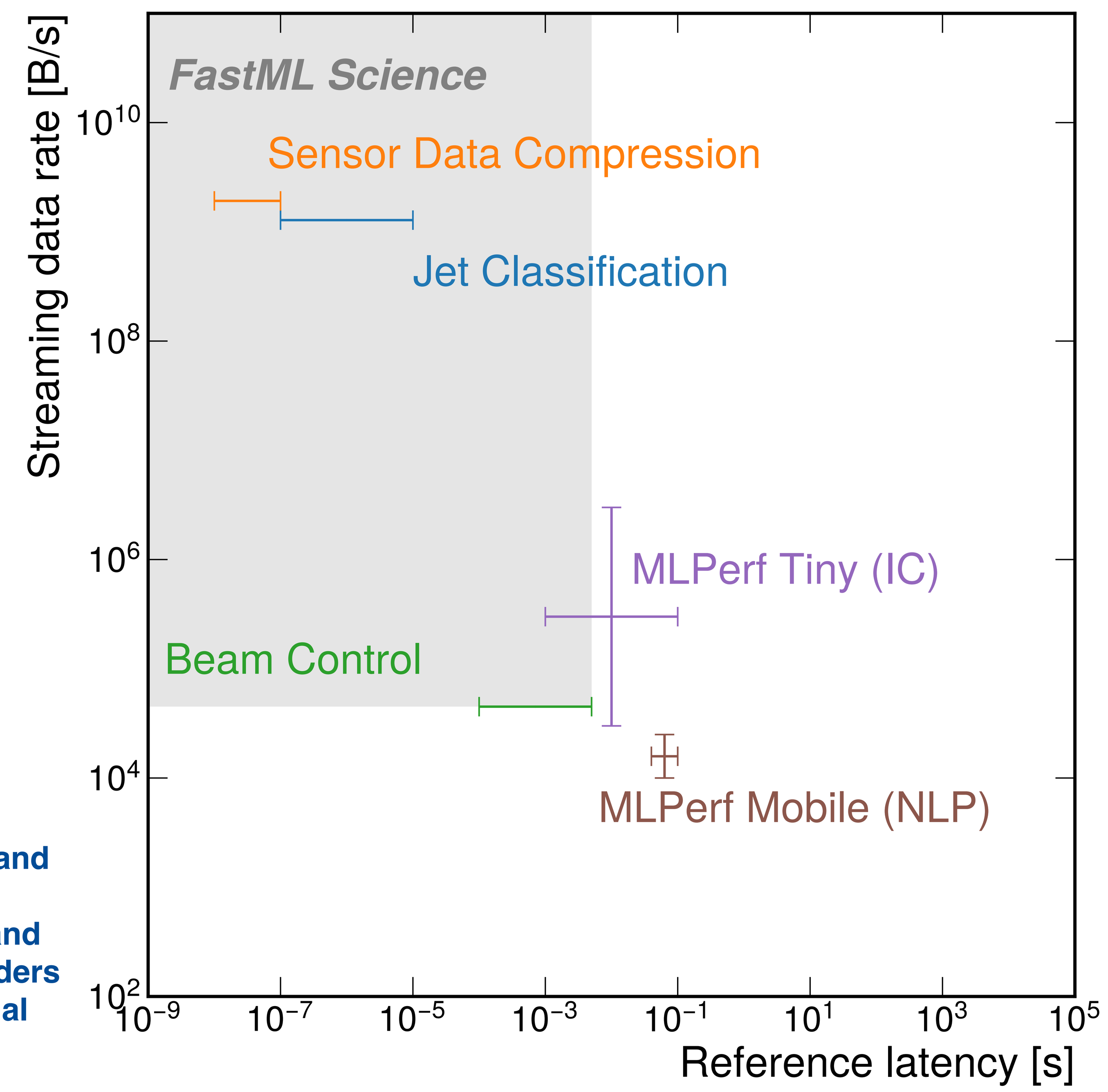


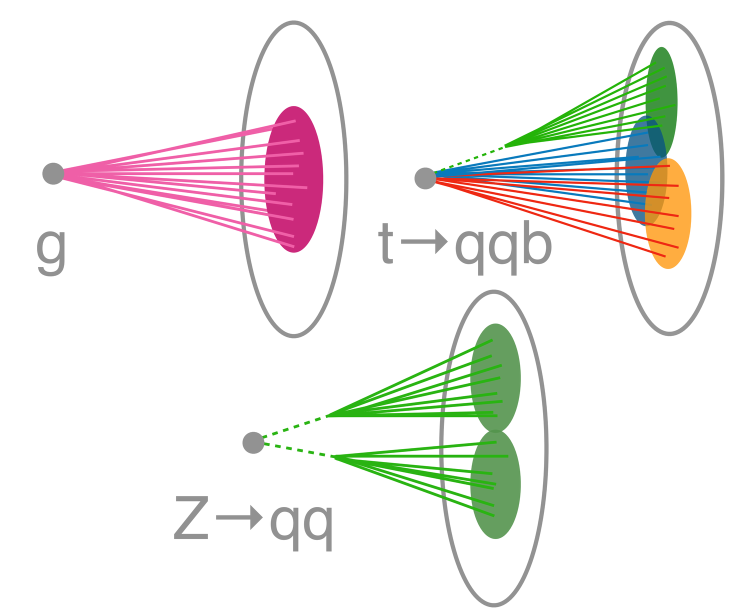
Figure: Reference latencies and streaming input data rates for common benchmarks and those proposed in this work. The FastML Science regime represents data volumes and inference latency requirements that are orders of magnitude more stringent than traditional consumer-facing applications.

We propose new **standardized benchmarks** representing state-of-the-art scientific ML tasks across supervised, unsupervised, and reinforcement learning that eclipse previously-considered levels of throughput and latency.

Table 1 Summary of constraints for the three FastML Science benchmark scenarios.

Type	Benchmark	Input Precision	Pipeline Rate	Real-time Latency
Supervised Learning	Jet Classification	16b	150 ns	1 μs
Unsupervised Learning	Sensor Data Compression	9b	25 ns	100 ns
Reinforcement Learning	Beam Control	32b	5 ms	5 ms

[1]University of California San Diego
[2]Fermi National Accelerator Laboratory
[3]Harvard University



Supervised classification of particle jets

A representative identification task for FPGA-based Large Hadron Collider (LHC) detector trigger systems, which produce 100s of TB/s of data at 40MHz event rates.

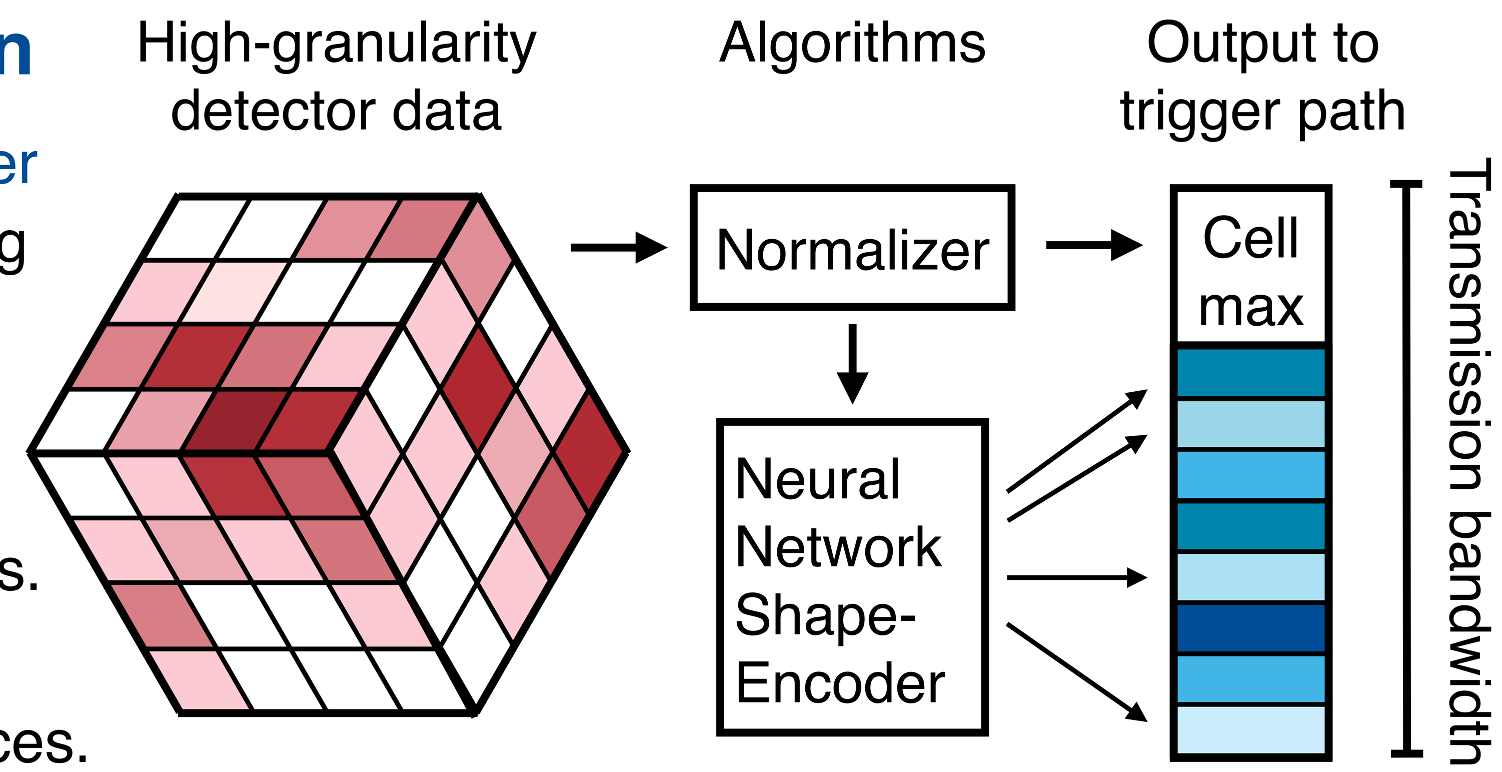
Data: Labeled jets with particle constituents *or* 16 expert features

Metrics: Classification accuracy, and FPR @ 50% TPR.

Baseline: 5-layer MLP; quantization-aware training; Xilinx VU9P target.

Irregular sensor data compression

This next-generation CMS imaging calorimeter will compress data by 400x, without sacrificing the ability to classify and measure particles. This high-radiation, on-chip environment requires an ASIC-friendly design.



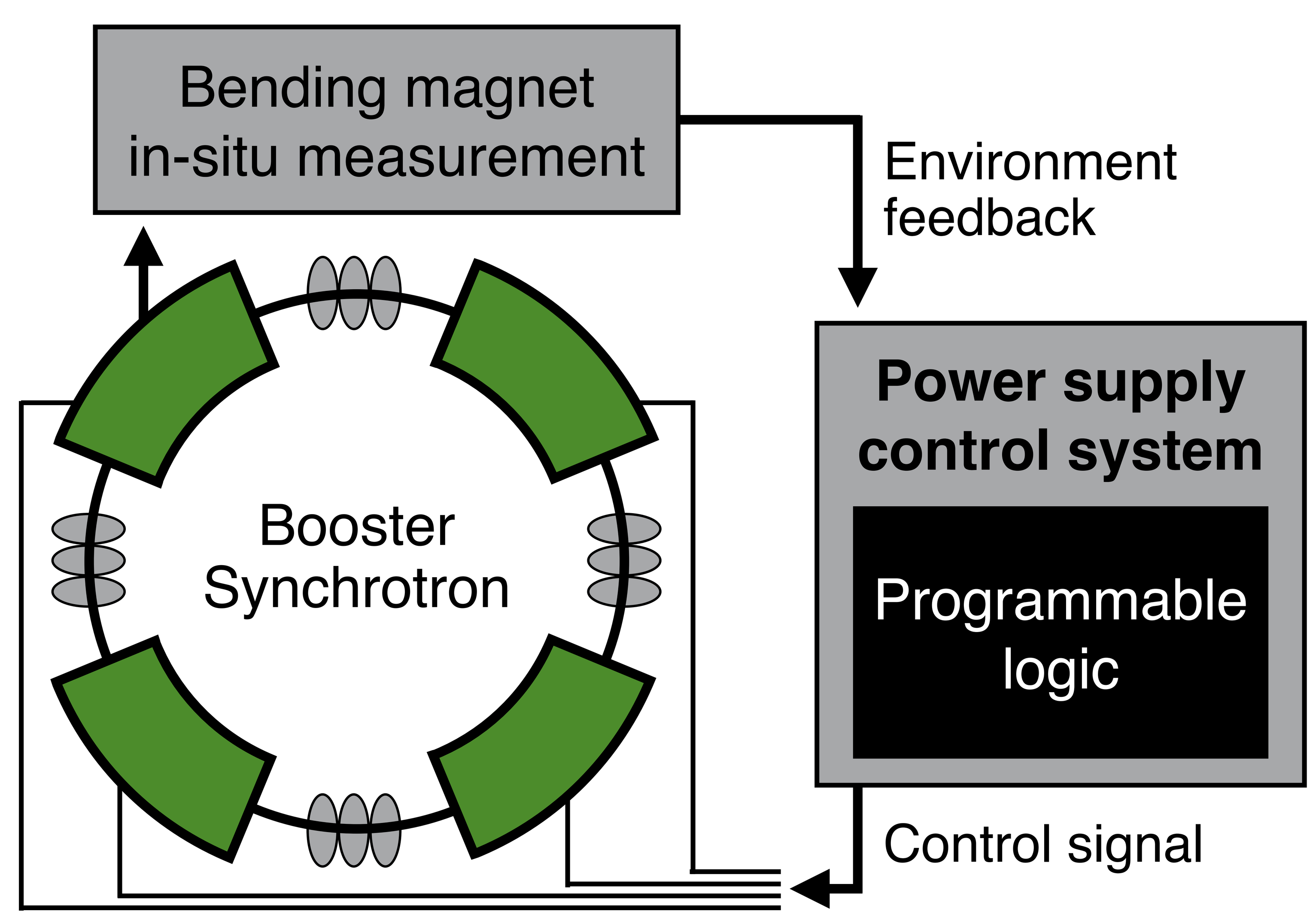
Data: Sensors of 48 normalized trigger cells.

Metrics: Similarity score based on the magnitude and distance of energy differences.

Baseline: Convolutional NN targeting 65nm CMOS process, 3.6mm^2 in area, drawing 60mW.

Accelerator beams control

At Fermilab, the Booster accelerator must guide protons along a precise trajectory in order to achieve maximal intensities. Here an ML agent controls the bending magnets, acting on past trajectories and other external measurements provided by a surrogate model of the accelerator complex.



Dataset: 54 measurement devices, sampled at the 15hz beam repetition rate.

Metrics: Time-averaged difference in the target and measured particle trajectories.

Baseline: Deep-Q network selecting from 7 possible actions (3-layer MLP). Design targets an Intel Arria10 FPGA.