# Fast ML for Science benchmarks and architectural implications

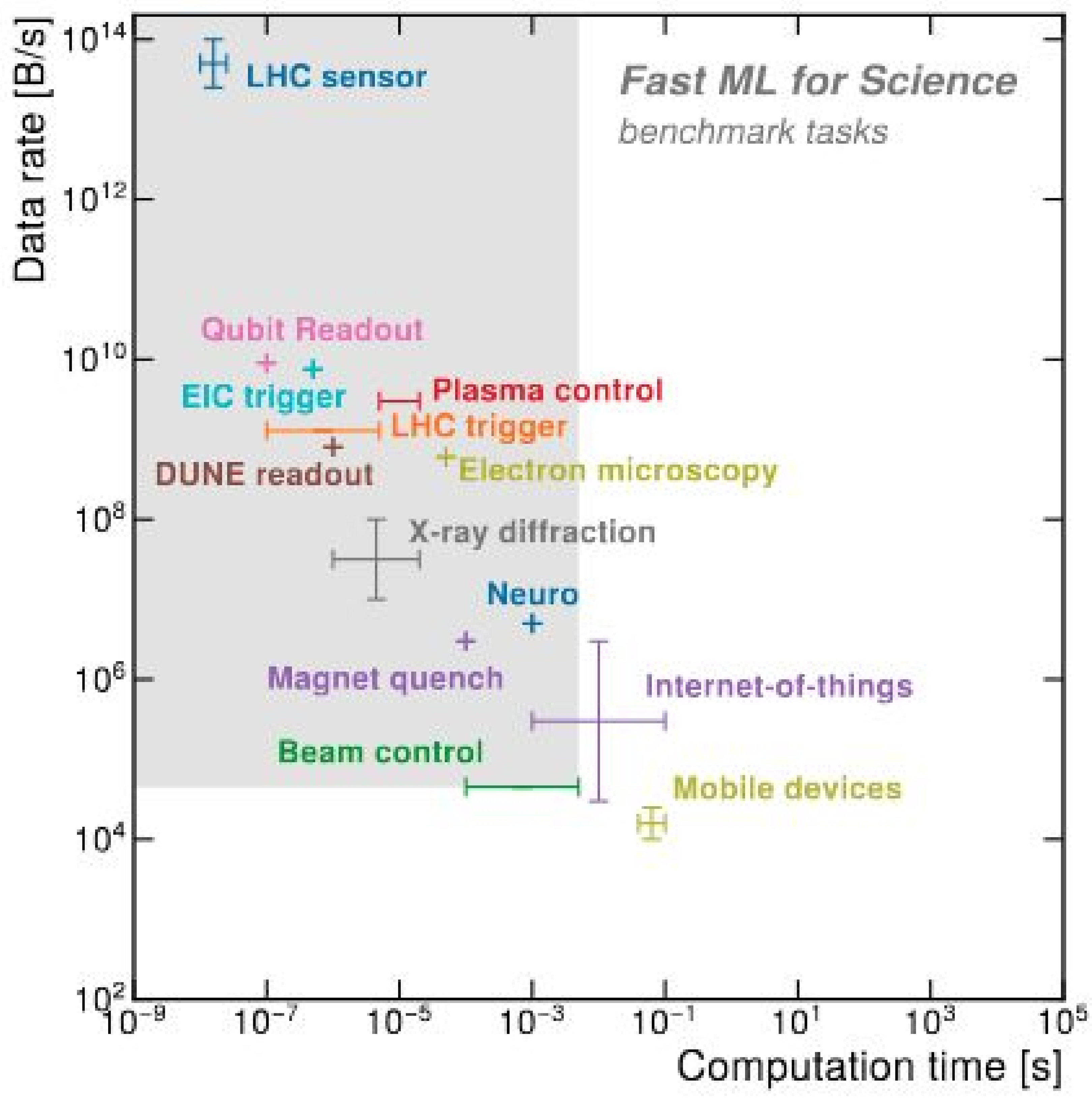
Ben Hawks, Nhan Tran

Olivia Weng, Javier Duarte, Ryan Kastner, Alexander Redding

***Grand vision*: autonomous experiments operating at the timescales of nature to solve mysteries of the universe, discover new materials, sources of energy, and beyond!**

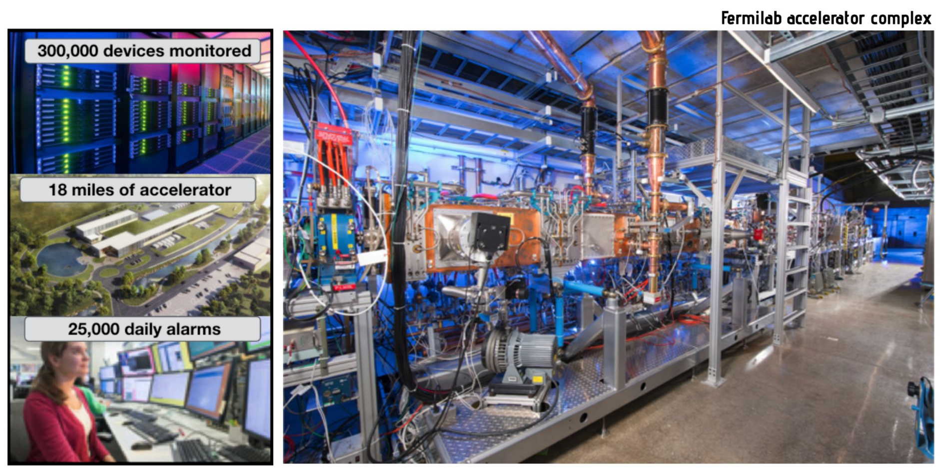
"Scientific discoveries come from groundbreaking ideas and the capability to validate those ideas by **testing nature at new scales-finer and more precise temporal and spatial resolution**. This is leading to an explosion of data that must be interpreted, and ML is proving a powerful approach. The more efficiently we can test our hypotheses, the faster we can achieve discovery. To fully unleash the power of ML and accelerate discoveries, it is necessary **to embed it into our scientific process, into our instruments and detectors.**"

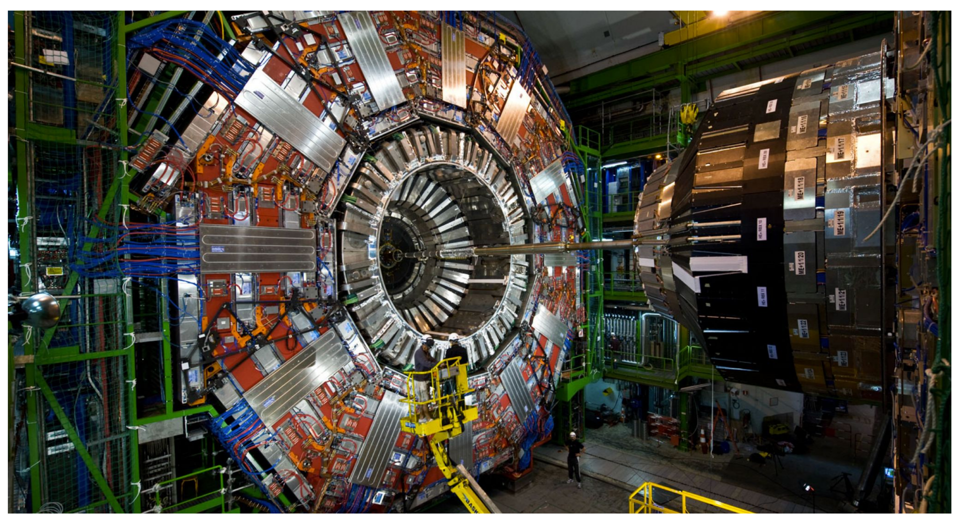
Based on "FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning"
Javier Duarte, Nhan Tran, Ben Hawks, Christian Herwig, Jules Muhizi, Shvetank Prakash, Vijay Janapa Reddi
https://doi.org/10.48550/arXiv.2207.07958

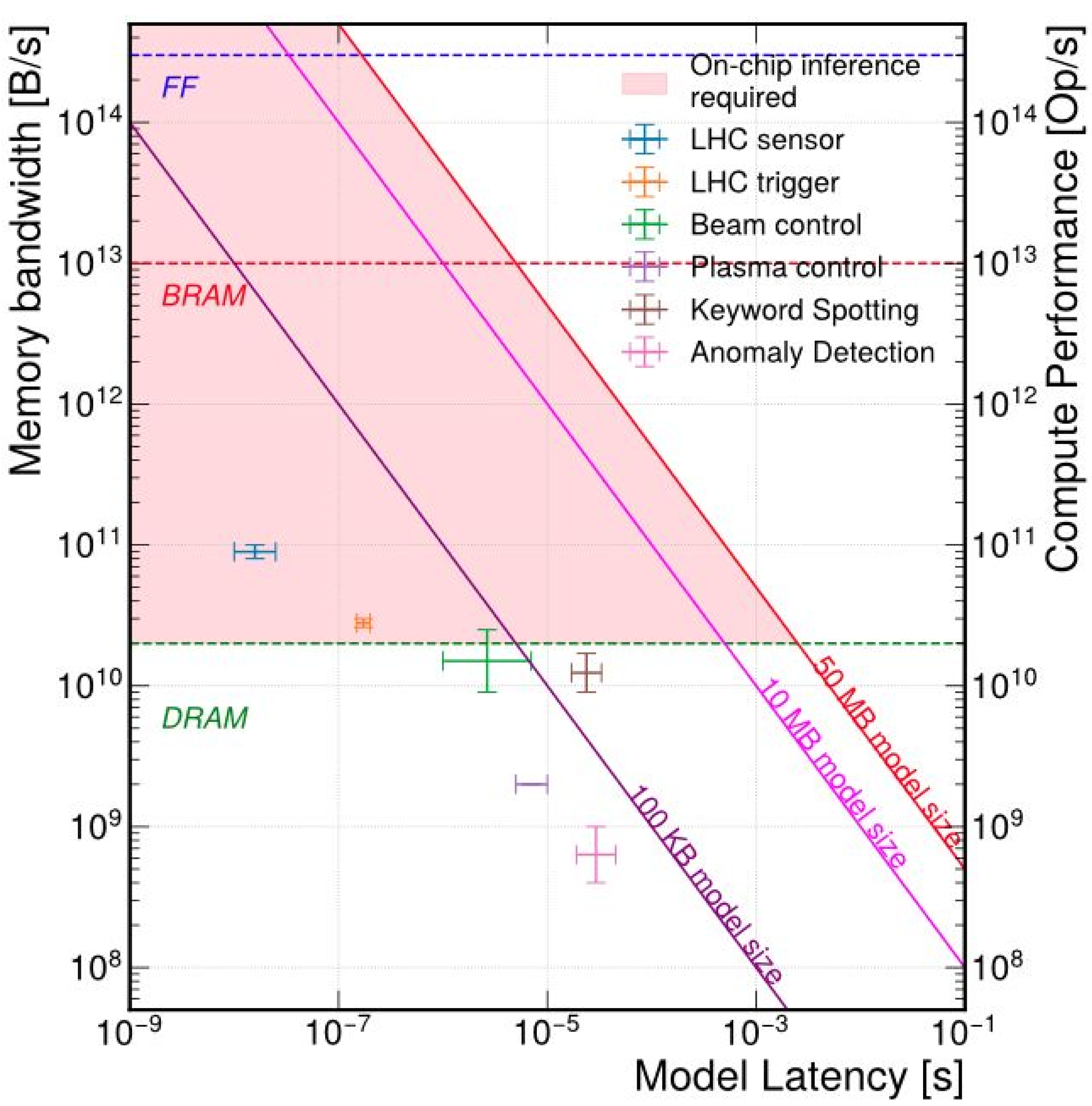



## Real-time SciML Benchmarks

*particle physics, nuclear physics, neuroscience, material science, fusion, particle accelerators, superconducting magnets, etc.*






"Architectural Implications of Neural Network Inference for High Data-Rate, Low-Latency Scientific Applications"
Olivia Weng, Alexander Redding, Nhan Tran, Javier Mauricio Duarte, Ryan Kastner
https://doi.org/10.48550/arXiv.2403.08980

## Unique architectural challenges for science:

- ultra-fast all-on-chip inference
- highly customizable models for custom applications including multiple platforms (SoC, AIE, PL, …)
- Adaptive workflows for extreme environments and changing conditions, e.g. digital twin interfaces

## Benchmarks lead to innovations and improvements

particle physics jet substructure task –
~500x improvement in LUT x ns from original works!

| Dataset | Model | Accuracy /EMD | LUT | FF | DSP | BRAM | Latency (ns) | FMax (MHz) | Area × Delay (LUT × ns) |
|---|---|---|---|---|---|---|---|---|---|
| MNIST | **AmigoLUT-LogicNet-XS** (2 models) | 94.7 | 9 711 | 9 047 | 0 | 0 | 12.3 | 569 | 119 445 |
| | **AmigoLUT-NeuraLUT** (4 models) | 95.5 | 16 081 | 13 292 | 0 | 0 | **7.6** | **925** | 122 216 |
| | PolyLUT [3] | 96 | 70 673 | 4 681 | 0 | 0 | 16 | 378 | 1 130 768 |
| | NeuraLUT [4] | 96 | 54 798 | 3 757 | 0 | 0 | 12 | 431 | 657 576 |
| | PolyLUT-Add [16] | 96 | 14 810 | 2 609 | 0 | 0 | 10 | 625 | 148 100 |
| | DWN [5] | **97.8** | **2 092** | **1 757** | 0 | 0 | 9.2 | 873 | **19 246** |
| JSC | **AmigoLUT-NeuraLUT-XS** (4 models) | 71.1 | 320 | 482 | 0 | 0 | 3.5 | **1 445** | 1 120 |
| | **AmigoLUT-NeuraLUT-XS** (16 models) | 72.9 | 1 243 | 1 240 | 0 | 0 | 5.0 | 1 008 | 6 215 |
| | **AmigoLUT-NeuraLUT-S** (32 models) | 74.4 | 42 742 | 4 717 | 0 | 0 | 9.6 | 520 | 410 323 |
| | LogicNet-L | 73.1 | 36 415 | 2 790 | 0 | 0 | 6 | 390 | 218 490 |
| | PolyLUT | 72 | 12 436 | 773 | 0 | 0 | 5 | 646 | 62 180 |
| | PolyLUT | 75 | 236 541 | 2 775 | 0 | 0 | 21 | 235 | 4 967 361 |
| | NeuraLUT | 72 | 4 684 | 341 | 0 | 0 | **3** | 727 | 14 052 |
| | NeuraLUT | 75 | 92 357 | 4 885 | 0 | 0 | 14 | 368 | 1 292 998 |
| | PolyLUT-Add | 75 | 36 484 | 1 209 | 0 | 0 | 16 | 315 | 583 744 |
| | PolyLUT-Add | 72 | 895 | 1 649 | 0 | 0 | 4 | 750 | 3 580 |
| | DWN | 73.7 | **134** | **106** | 0 | 0 | 3.7 | 1 361 | **496** |
| | DWN | **76.3** | 6 302 | 4 128 | 0 | 0 | 14.4 | 695 | 90 749 |
| HGCal | **AmigoLUT-LogicNet-S** (2 models) | 1.386 | **26 499** | 4 040 | 0 | 0 | 15.6 | **512** | 411 840 |
| | **AmigoLUT-LogicNet-S** (16 models) | **1.270** | 195 724 | 23 515 | 0 | 0 | 24.0 | 334 | 4 697 376 |
| | LogicNet-L | 1.407 | 32 529 | **2 340** | 0 | 0 | **12.2** | 323 | **396 854** |

**Table 4: Comparing model resource utilization and performance metrics across the three datasets/tasks with prior work.**

*See presentation from Olivia Weng et al from Thursday*

Greater than the Sum of its LUTs: Scaling Up LUT-based Neural Networks with AmigoLUT.
Olivia Weng, Marta Andronic, Danial Zuberi, Jiaqing Chen, Caleb Geniesse, George A. Constantinides, Nhan Tran, Nicholas Fraser, Javier Mauricio Duarte, Ryan Kastner.
https://doi.org/10.1145/3706628.3708874

Fermi National Accelerator Laboratory

Fermilab U.S. DEPARTMENT OF ENERGY

BH and NT are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy (DOE), Office of Science, Office of High Energy Physics. BH, JD, OW, RK, AR, and NT are supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research under the "Real-time Data Reduction Codesign at the Extreme Edge for Science" Project(DE-FOA-0002501).

FERMILAB-POSTER–25-0008-CSAID