Data Reduction codesign at the extreme edge (XDR)

FERMILAB-SLIDES-24-0132-CSAID

PIs: Josh Agar[1], Javier Duarte[2], Amir Gholami[3], Phil Harris[4], Ryan Kastner[2], Michael Mahoney[3], Jennifer Ngadiuba[5], ***Nhan Tran***[5]

[1]Drexel University, [2]UCSD, [3]ICSI/Berkeley, [4]MIT, [5]Fermilab,

Abstract

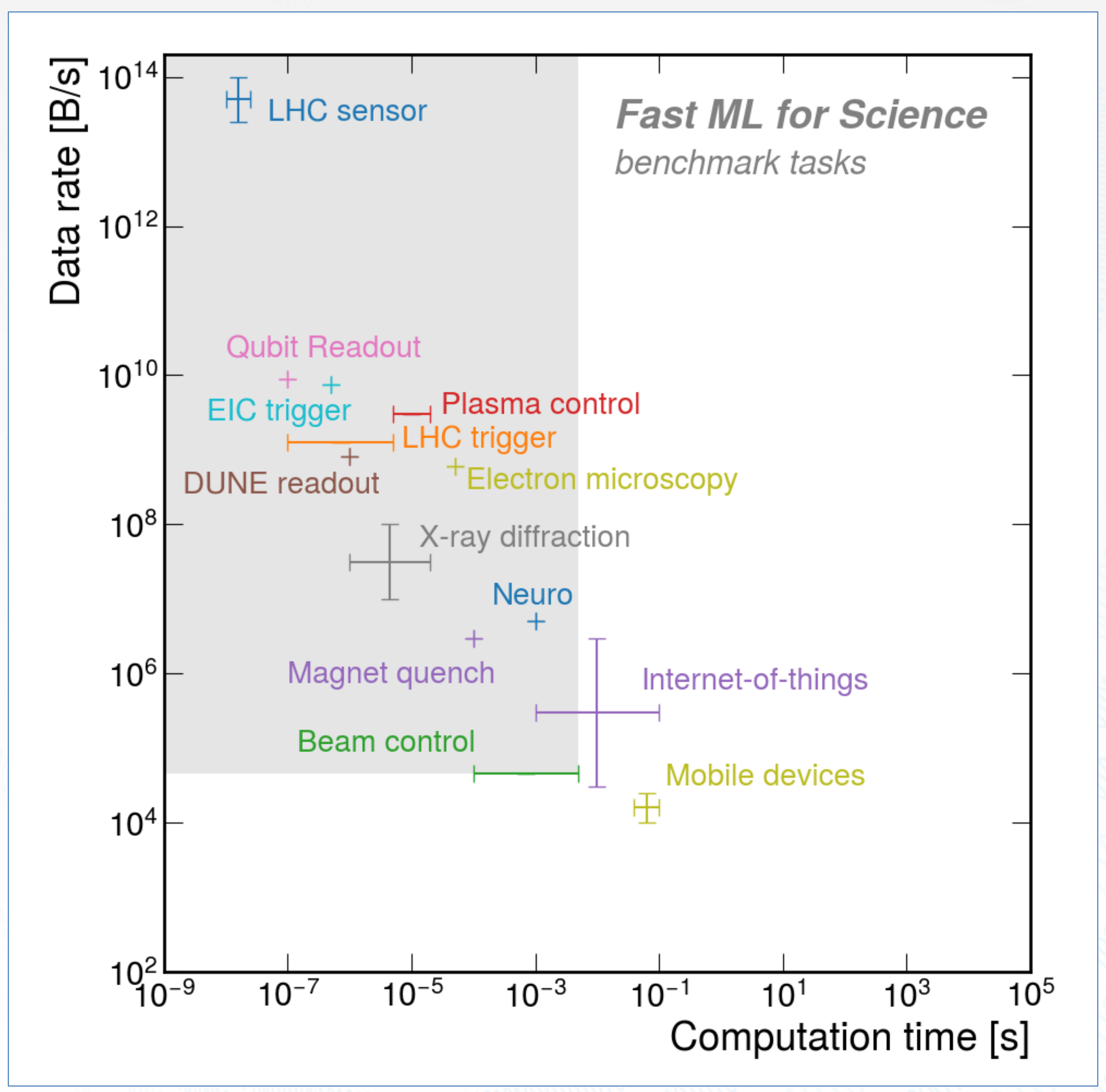
Intelligent ML-based data reduction **as close as possible to the data source promises to vastly accelerate scientific discovery potential**. Per sensor compression and efficient aggregation of information while preserving scientific fidelity can have a huge impact on experiment data flow, analysis, ,control, and operation; and ultimately how quickly experiments can be performed and hypotheses explored.

We concentrate on powerful, specialized compute hardware at the extreme edge such as FPGAs, ASICs, and systems-on-chip — on platforms common to many scientific experiments. We aim to:

- develop **performant and reliable AI algorithms** for science at the edge
- develop **codesign tools to build efficient implementations** of those algorithms in hardware;
- enable rapid **exploration for domain scientists and system designers** with an **accessible tool flow**.

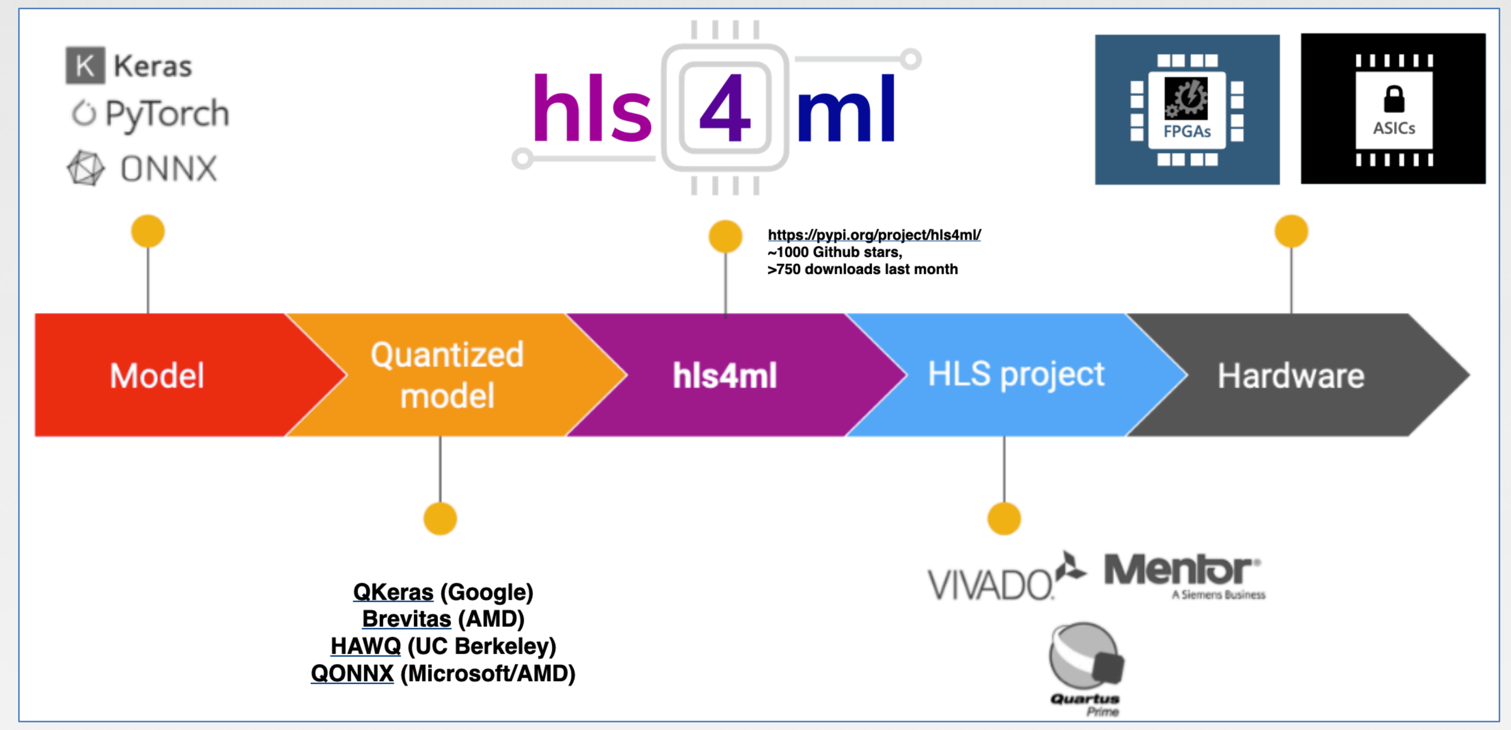
Motivation

Grand challenges spark imaginations, benchmarks bring innovation [3]



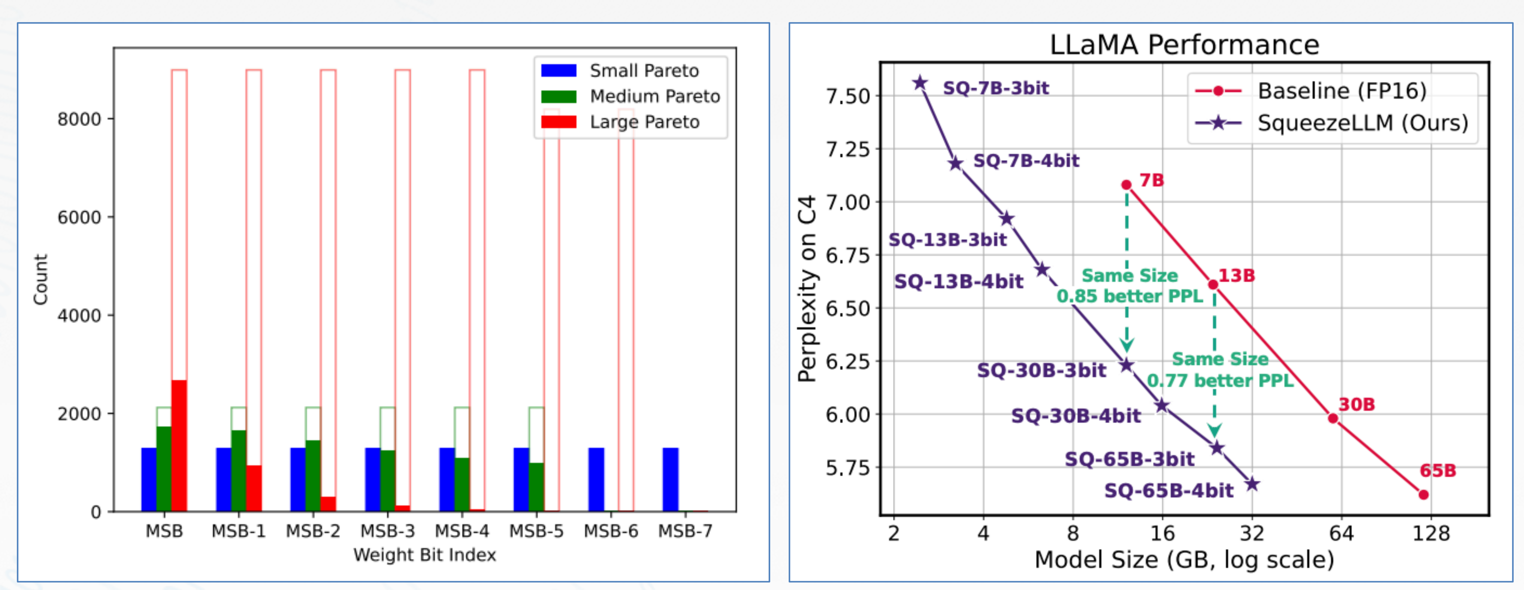
Approach

Robust and efficient AI through accessible workflows with hls4ml [1]

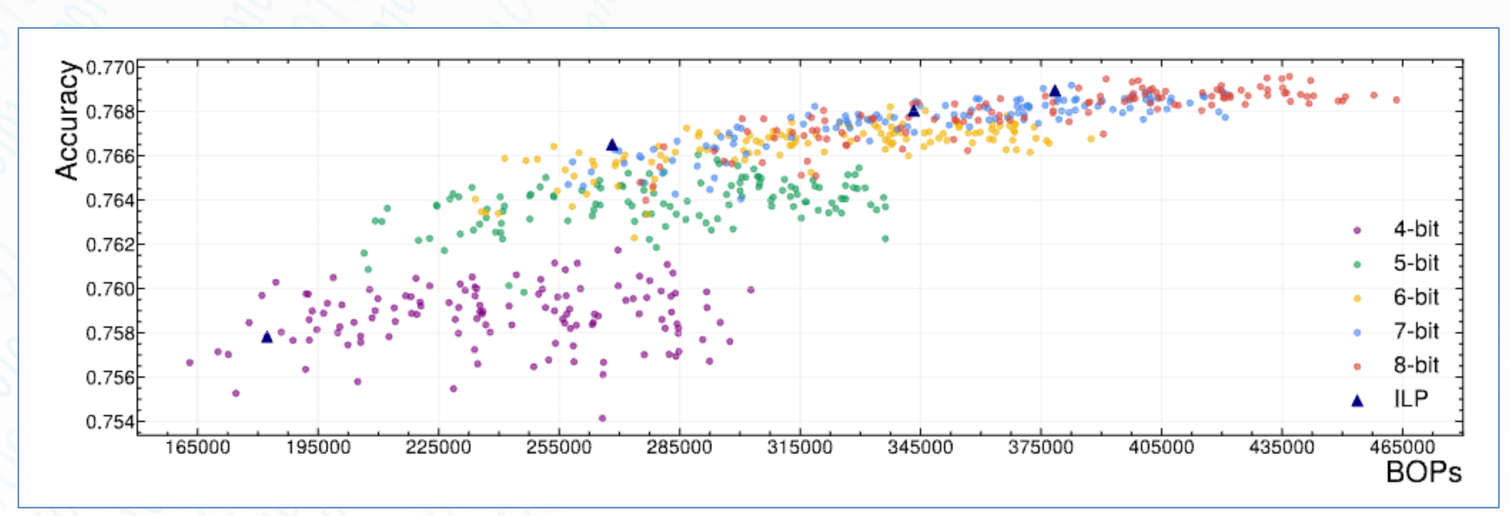


Methods and Highlights

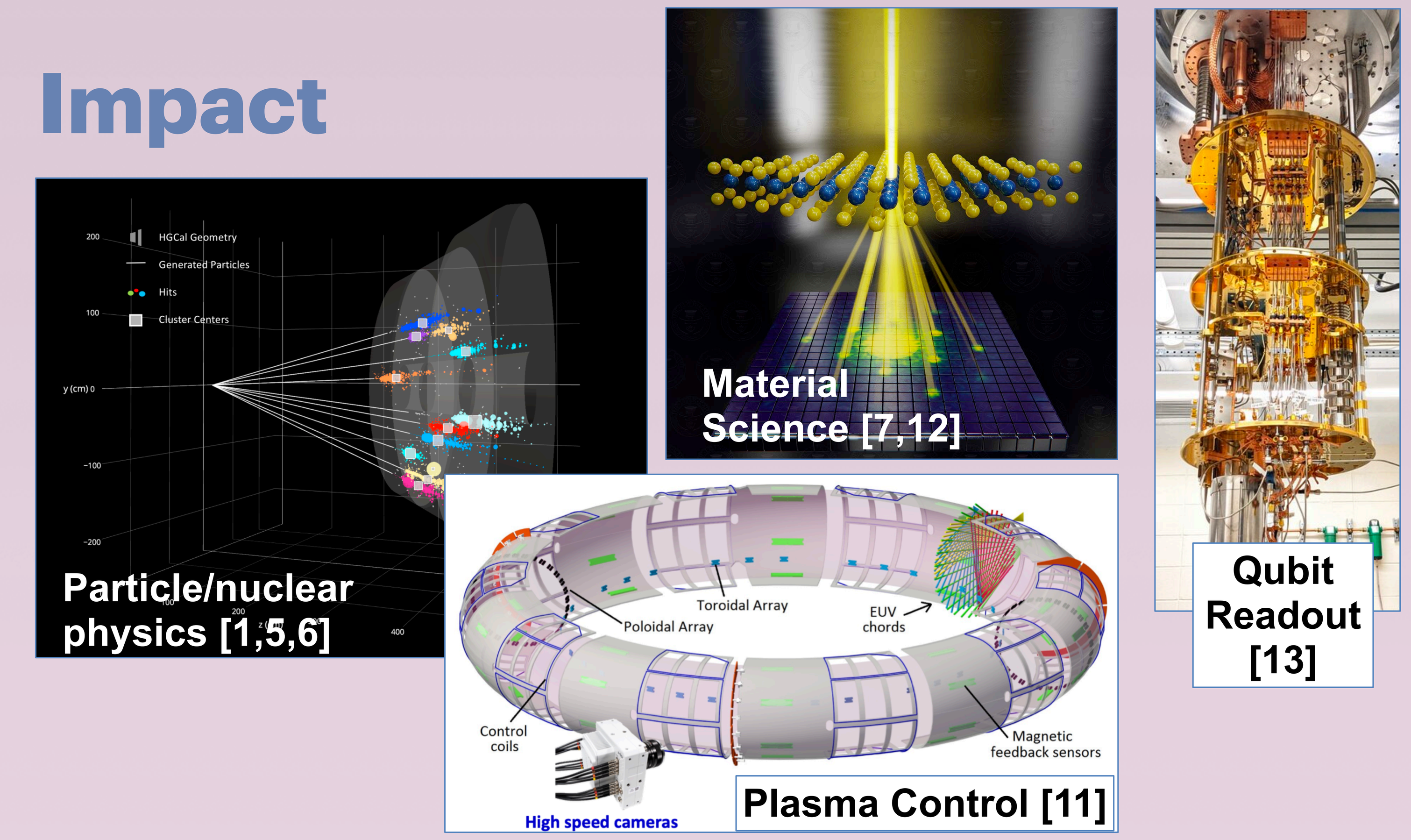
Energy-efficient algorithms with fine-grained Hessian quantization-aware training and sparsification [2,4,9,10,15]



Programming models for novel AI hardware and architectures with user-driven accessible codesign tool flows [1,6,9]

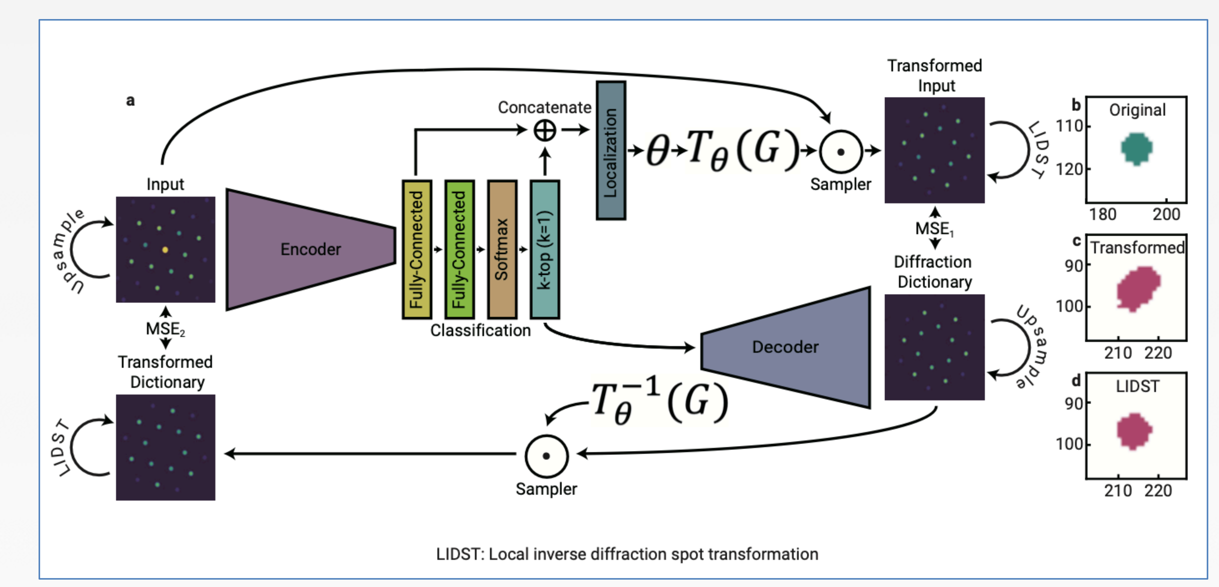


Impact



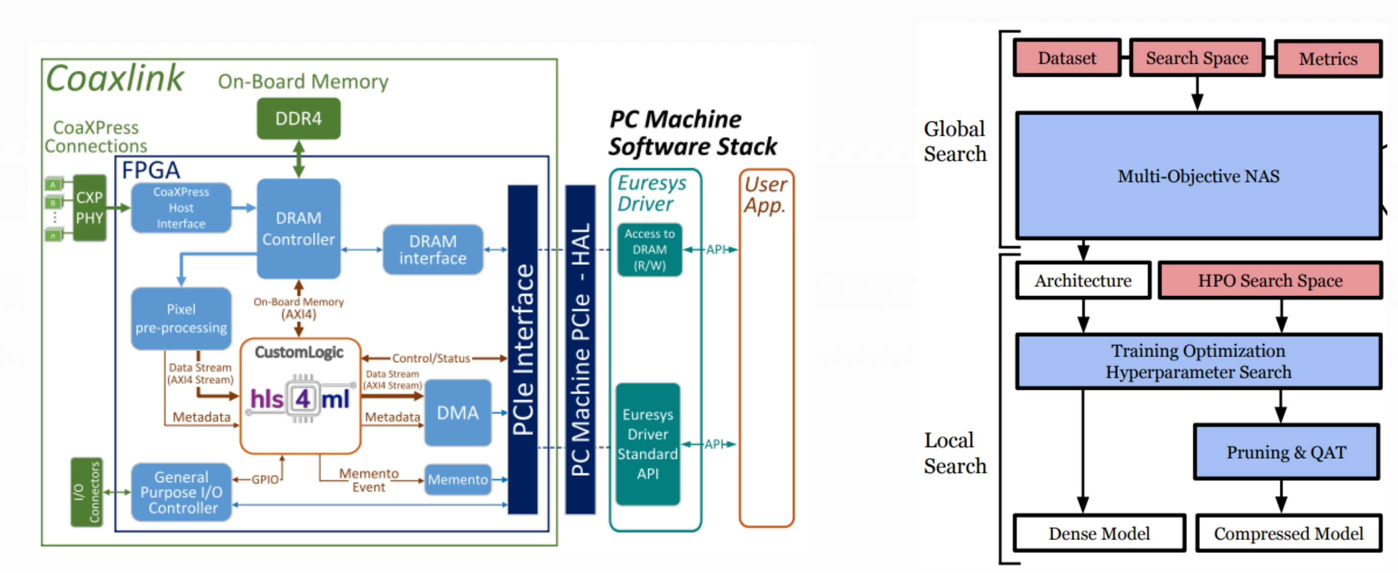
Bring new experimental paradigms and capabilities to many DOE SC applications and beyond!

Physics-aware architectures for robust parameter extraction [5,7]



Methods for scientific model robustness: fault tolerance, noise, and loss landscape [8,10,14]

Neural architecture codesign tools to build custom hardware implementations for scientific applications [1,11,12,13]



Collaboration Opportunities

Methods for robust, reliable, efficient ML codesign;
Programming models for embedded hardware architectures;
Low-latency scientific applications

References

1. Fast inference of deep neural networks in FPGAs for particle physics
2. HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision
3. FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning
4. Tailor: Altering Skip Connections for Resource-Efficient Inference
5. Differentiable Earth Mover's Distance for Data Compression at the High-Luminosity LHC
6. End-to-end codesign of Hessian-aware quantized neural networks for FPGAs
7. Extremely Noisy 4D-TEM Strain Mapping Using Cycle Consistent Spatial Transforming Autoencoders
8. Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior
9. SqueezeLLM: Dense-and-Sparse Quantization
10. FKeras: A Sensitivity Analysis Tool for Edge Neural Networks
11. Low latency optical-based mode tracking with machine learning deployed on FPGAs on a tokamak
12. Neural Architecture Codesign for Fast Bragg Peak Analysis
13. Implementing Machine Learning Methods on QICK hardware for Quantum Readout
14. Taxonomizing local versus global structure in neural network loss landscapes
15. Full Stack Optimization of Transformer Inference: a Survey

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Data Reduction codesign at the extreme edge (XDR)

PIs: Josh Agar[1], Javier Duarte[2], Amir Gholami[3], Phil Harris[4], Ryan Kastner[2], Michael Mahoney[3], Jennifer Ngadiuba[5], ***Nhan Tran***[5]

[1]Drexel University, [2]UCSD, [3]ICSI/Berkeley, [4]MIT, [5]Fermilab,

Abstract

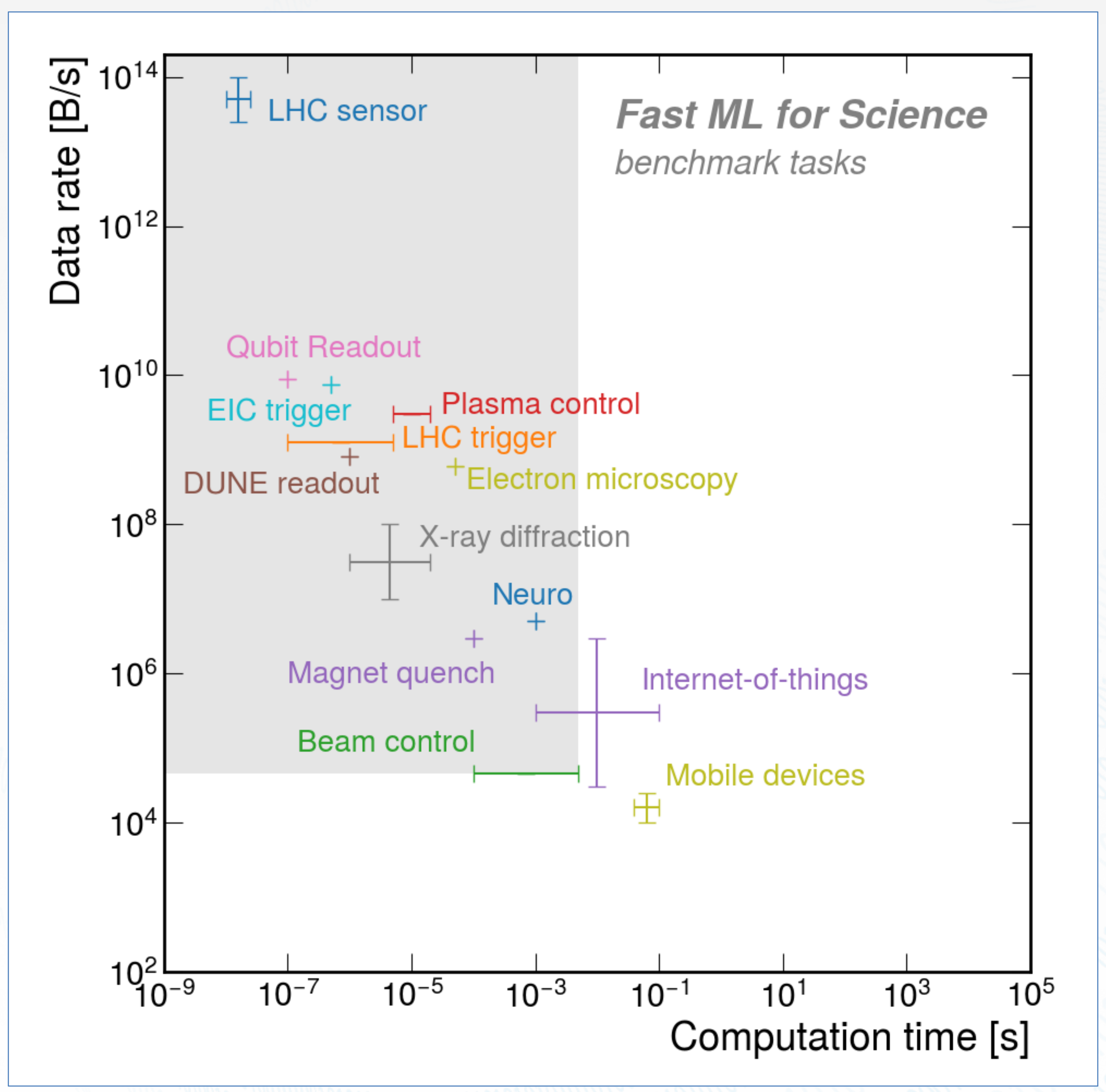
Intelligent ML-based data reduction **as close as possible to the data source promises to vastly accelerate scientific discovery potential**. Per sensor compression and efficient aggregation of information while preserving scientific fidelity can have a huge impact on experiment data flow, analysis, ,control, and operation; and ultimately how quickly experiments can be performed and hypotheses explored.

We concentrate on powerful, specialized compute hardware at the extreme edge such as FPGAs, ASICs, and systems-on-chip — on platforms common to many scientific experiments. We aim to:

- develop **performant and reliable AI algorithms** for science at the edge
- develop **codesign tools to build efficient implementations** of those algorithms in hardware;
- enable rapid **exploration for domain scientists and system designers** with an **accessible tool flow**.

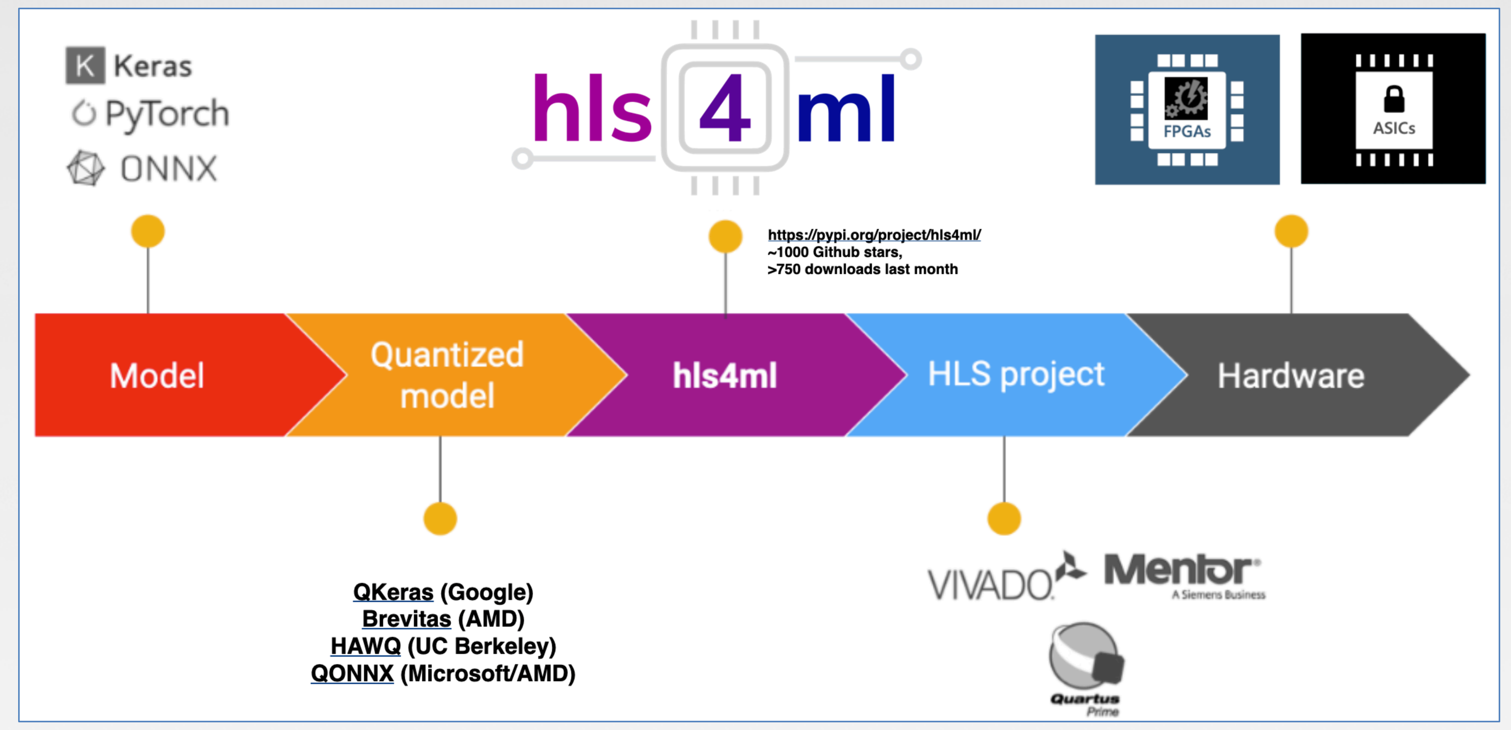
Motivation

Grand challenges spark imaginations, benchmarks bring innovation [3]



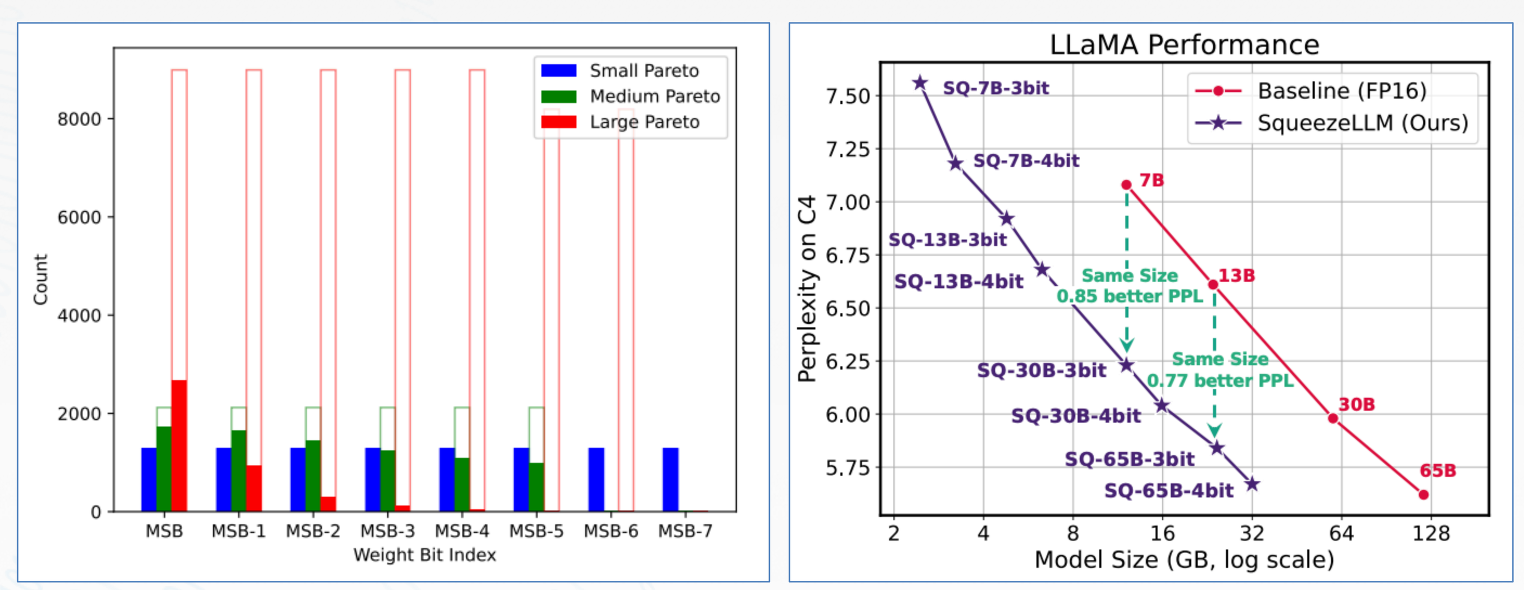
Approach

Robust and efficient AI through accessible workflows with hls4ml [1]

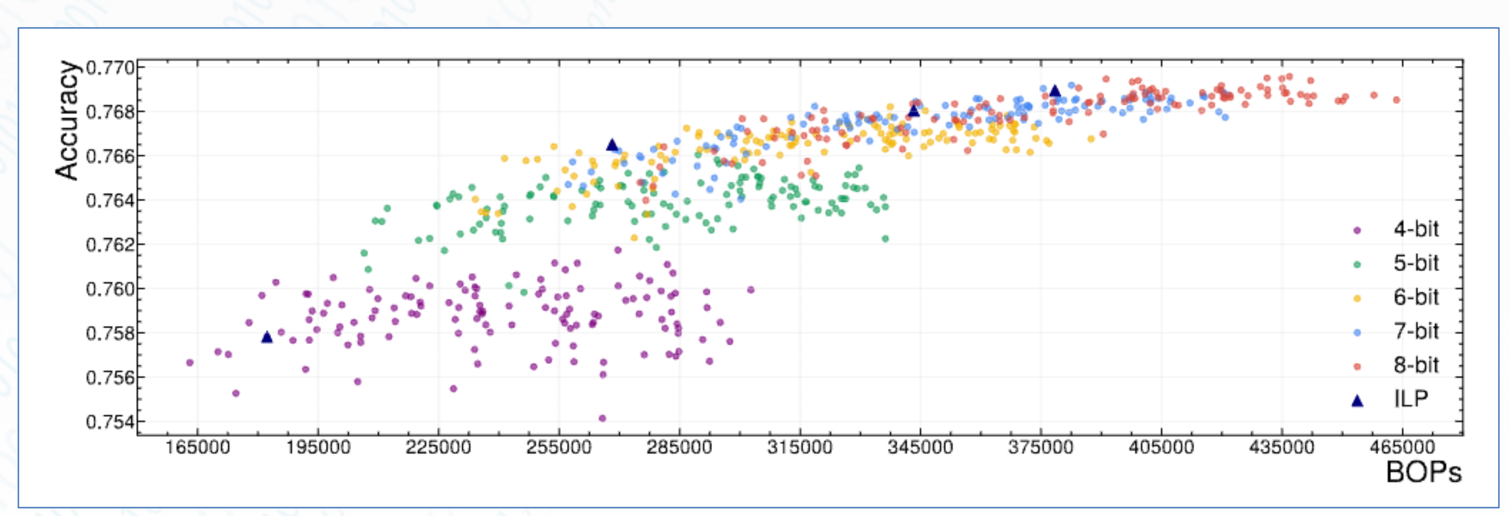


Methods and Highlights

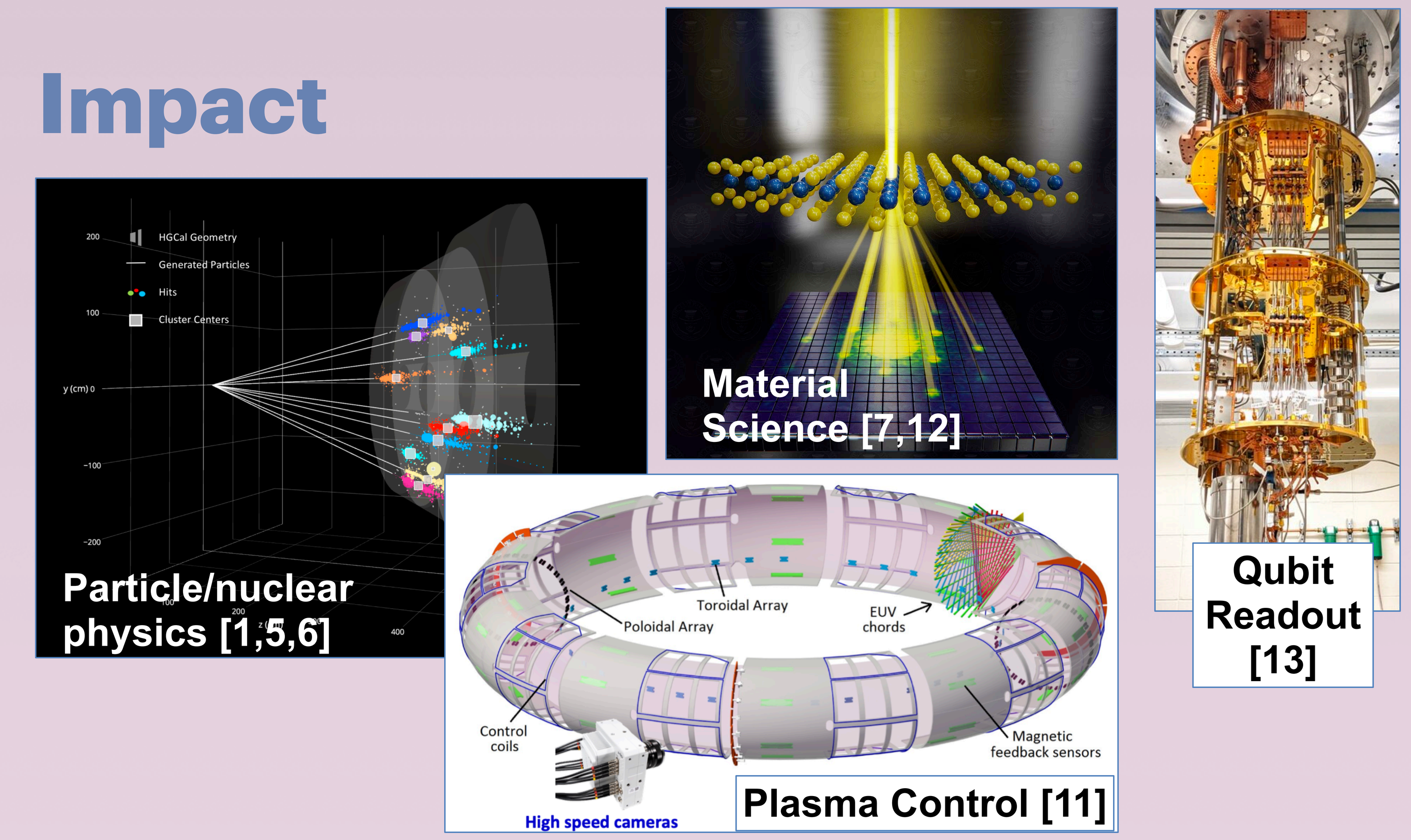
Energy-efficient algorithms with fine-grained Hessian quantization-aware training and sparsification [2,4,9,10,15]



Programming models for novel AI hardware and architectures with user-driven accessible codesign tool flows [1,6,9]

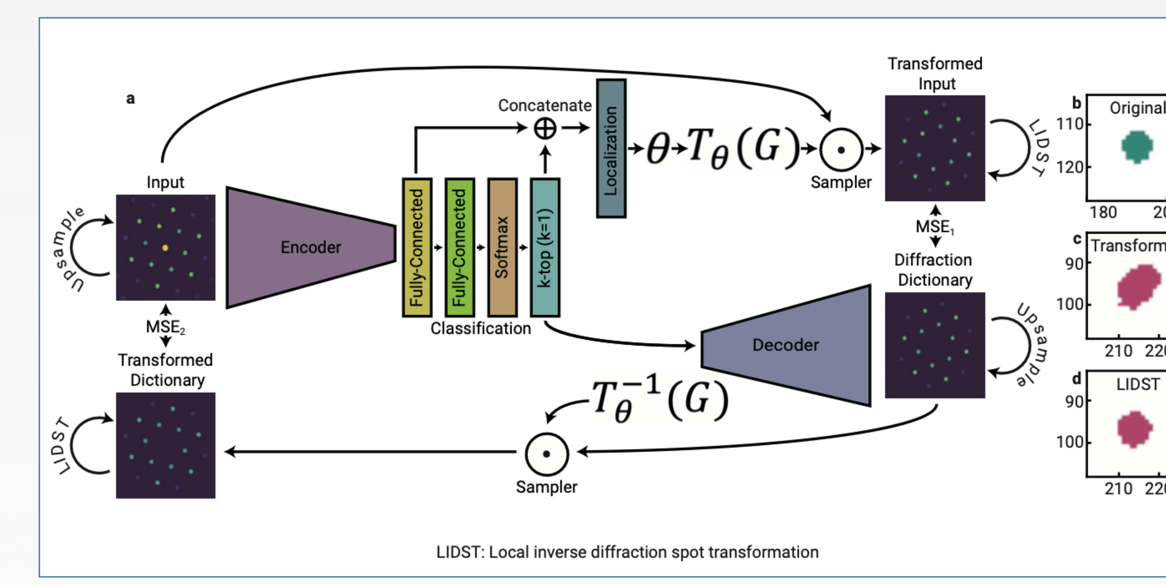


Impact



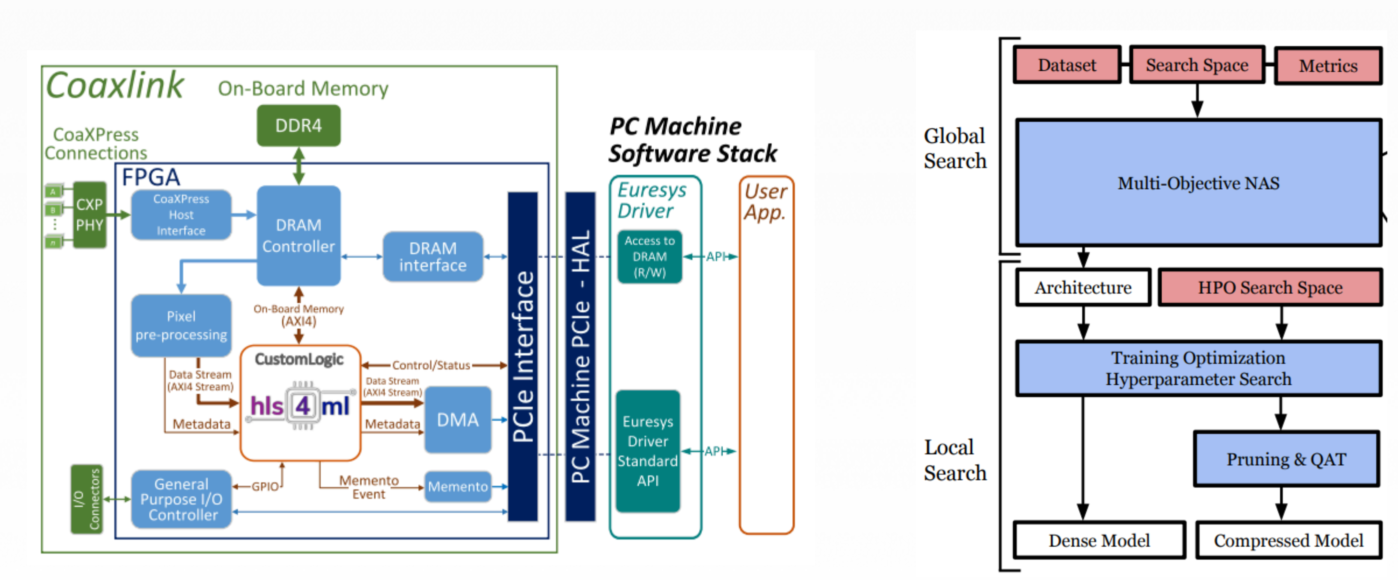
Bring new experimental paradigms and capabilities to many DOE SC applications and beyond!

Physics-aware architectures for robust parameter extraction [5,7]



Methods for scientific model robustness: fault tolerance, noise, and loss landscape [8,10,14]

Neural architecture codesign tools to build custom hardware implementations for scientific applications [1,11,12,13]



Collaboration Opportunities

Methods for robust, reliable, efficient ML codesign;
Programming models for embedded hardware architectures;
Low-latency scientific applications

References

1. Fast inference of deep neural networks in FPGAs for particle physics
2. HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision
3. FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning
4. Tailor: Altering Skip Connections for Resource-Efficient Inference
5. Differentiable Earth Mover's Distance for Data Compression at the High-Luminosity LHC
6. End-to-end codesign of Hessian-aware quantized neural networks for FPGAs
7. Extremely Noisy 4D-TEM Strain Mapping Using Cycle Consistent Spatial Transforming Autoencoders
8. Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior
9. SqueezeLLM: Dense-and-Sparse Quantization
10. FKeras: A Sensitivity Analysis Tool for Edge Neural Networks
11. Low latency optical-based mode tracking with machine learning deployed on FPGAs on a tokamak
12. Neural Architecture Codesign for Fast Bragg Peak Analysis
13. Implementing Machine Learning Methods on QICK hardware for Quantum Readout
14. Taxonomizing local versus global structure in neural network loss landscapes
15. Full Stack Optimization of Transformer Inference: a Survey

FERMILAB-POSTER-24-0005-CSAID
This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

DOE ASCR Computer Science Principal Investigators (PI) Meeting, Atlanta, GA February 5-7, 2024