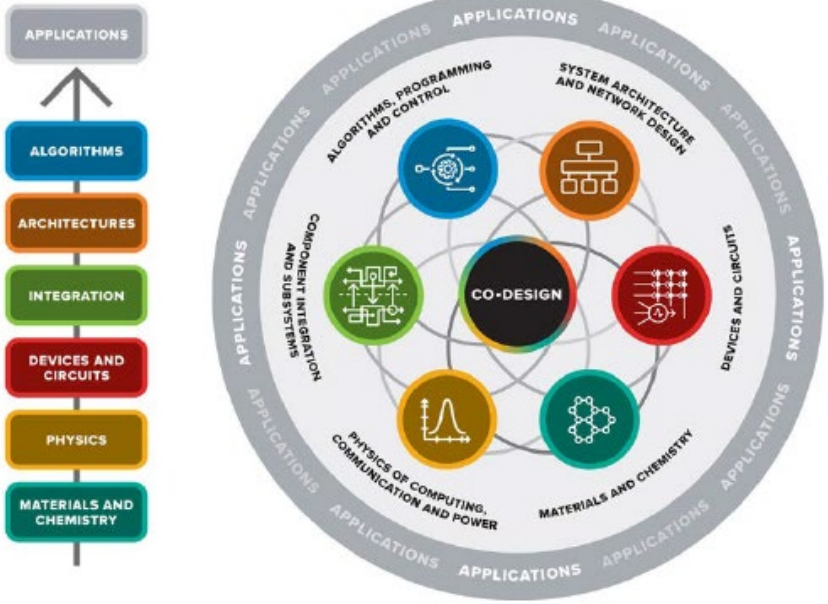
Importance of Co-Design: Long-term and Near-term Impacts

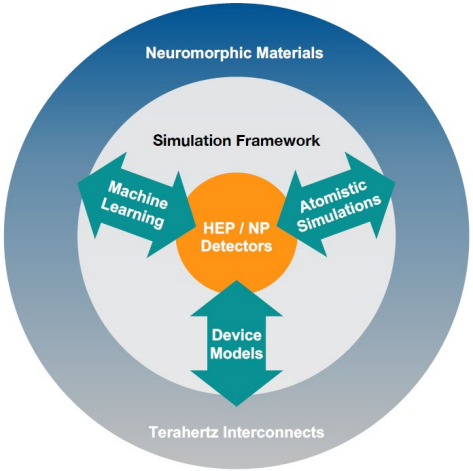
Valerie Taylor
Argonne National Laboratory

Importance of Co-Design

Long-term Impact



Report from DOE SC Workshop on Basic Research Needs for Microelectronics, 2018



Threadwork Project

- Unique devices for tunable kernels within single device for SVM
- 100-fold reduction in device count and power compared to conventional circuits
- Used for classification with HEP detectors

https://doi.org/10.1038/s419280023-01042-7

Hardware-Software Co-Design: Hardware Trends

Near-term Impact

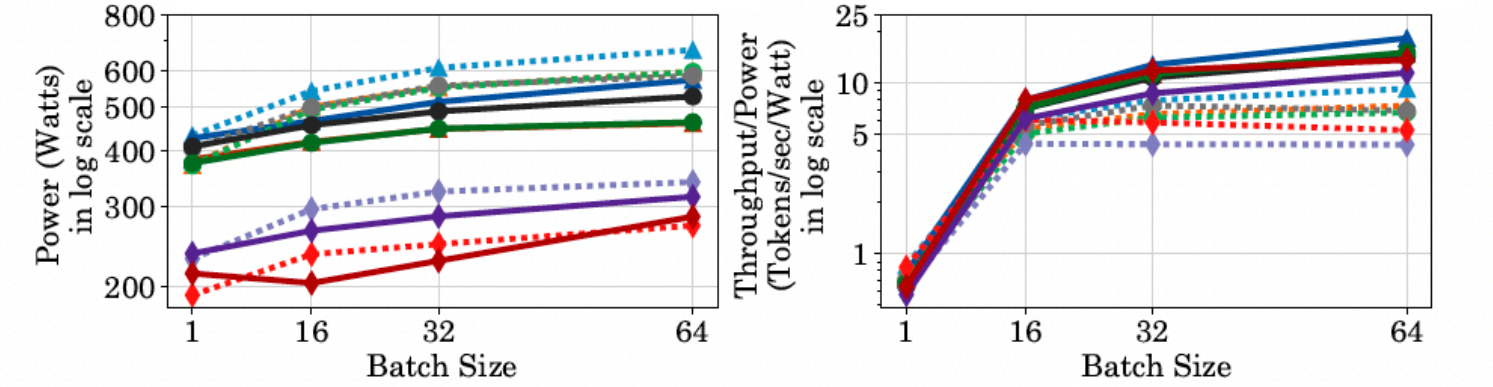
Operation	Energy per Op (pJ) 7nm	Ratio to Int8 ADD
Int8 ADD	0.007	
Int32 ADD	0.03	4.3
BFloat16 ADD	0.11	15.7
IEEE FP 16 ADD	0.16	22.9
IEEE FP 32 ADD	0.38	54.3
Int8 MULT	0.07	10.0
Int32 MULT	1.48	211.4
Bfloat16 MULT	0.21	30.0
IEEE FP 16 MULT	0.34	48.6
IEEE FP 32 MULT	1.31	187.1
8 kB SRAM access	7.5	1071.4
32 kB SRAM access	8.5	1214.3
1 MB SRAM access	14	2000.0
DDR3/4 DRAM access	1300	185,714.3
HBM 2 DRAM access	250-450	35,714.3 – 64,285.7
GDDR6 DRAM access	350-480	50,000.0 – 68,571.4

- Important to reduce data movement as much as possible
- Take advantage of mixed or low precision computations when possible
  - Ozaki scheme for large-scale matrix operations
- Take advantage of Compute-in-Memory

Jouppi, Norman P., et al. 2021, "Ten Lessons from Three Generations Shaped Google's TPUv4i: Industrial Product, ISCA.

Exploring Hardware and Software Options

- *LLM-Inference-Bench to explore:*
  - Inference performance of open source LLMs
  - Different AI accelerators
  - Widely available LLM inference frameworks



https://github.com/argonne-lcf/LLM-Inference-Bench