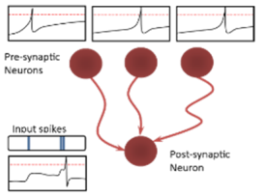
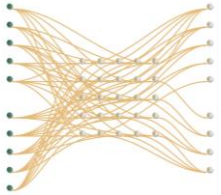
SAND2021-3207C

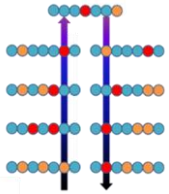
Sandia National Laboratories

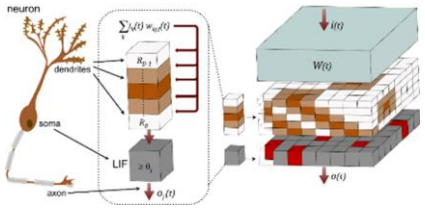
SEEK

Scoping neuromorphic architecture impact enabling advanced sensing capabilities









Presented by Craig M. Vineyard, PhD

Co-PI: Andrew Sornborger, PhD (LANL)

Contact emails: cmviney@sandia.gov, sornborg@lanl.gov

Los Alamos NATIONAL LABORATORY EST. 1943

U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

PROJECT OVERVIEW

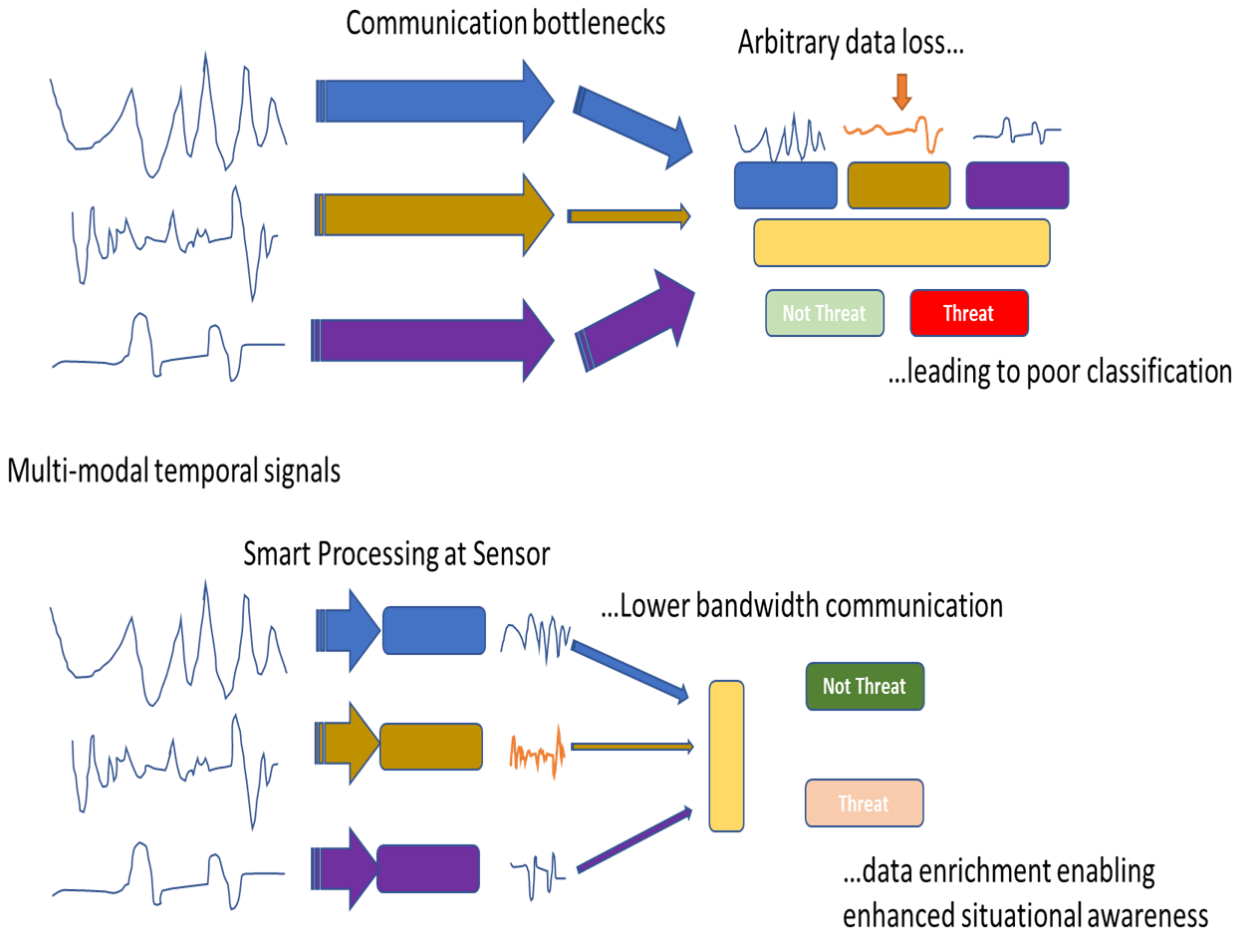
SEEK - Scoping neuromorphic architecture impact enabling advanced sensing capabilities
YEAR 2 of 2

Project Purpose	Goals
As ML approaches are increasingly impactful, revolutionizing signal processing and data science tasks, this effort ***Seeks*** to understand how emerging neural computing approaches can enable NA-22 neural network data analysis needs	Provide an assessment of neural computing approaches yielding insight into the interplay of neural network algorithms and architectures
Approach	**Deliverables**
Scoping study effort integrating empirical studies and analytical developments jointly pursued by SNL & LANL • Architecture Insights • Algorithmic Assessments • Co-Design Analysis	Produce a report providing insight into how existing and emerging neuromorphic computation may benefit NA-22 monitoring activities as well as enable future algorithm development

NA-22 Impact

Neuromorphic computing offers a path to enabling enhanced processing at the sensor

- The amount of data transmitted for subsequent data science processing or analyst interpretation may be reduced
- Initial processing on the raw data at the sensor provides an opportunity to transmit higher order information
- Instead of providing more alerts or signatures to attend to, the neuromorphic pre-processing of remotely sensed data may be filtered providing a data enrichment through machine intelligence processing helping to enable model-driven analytics



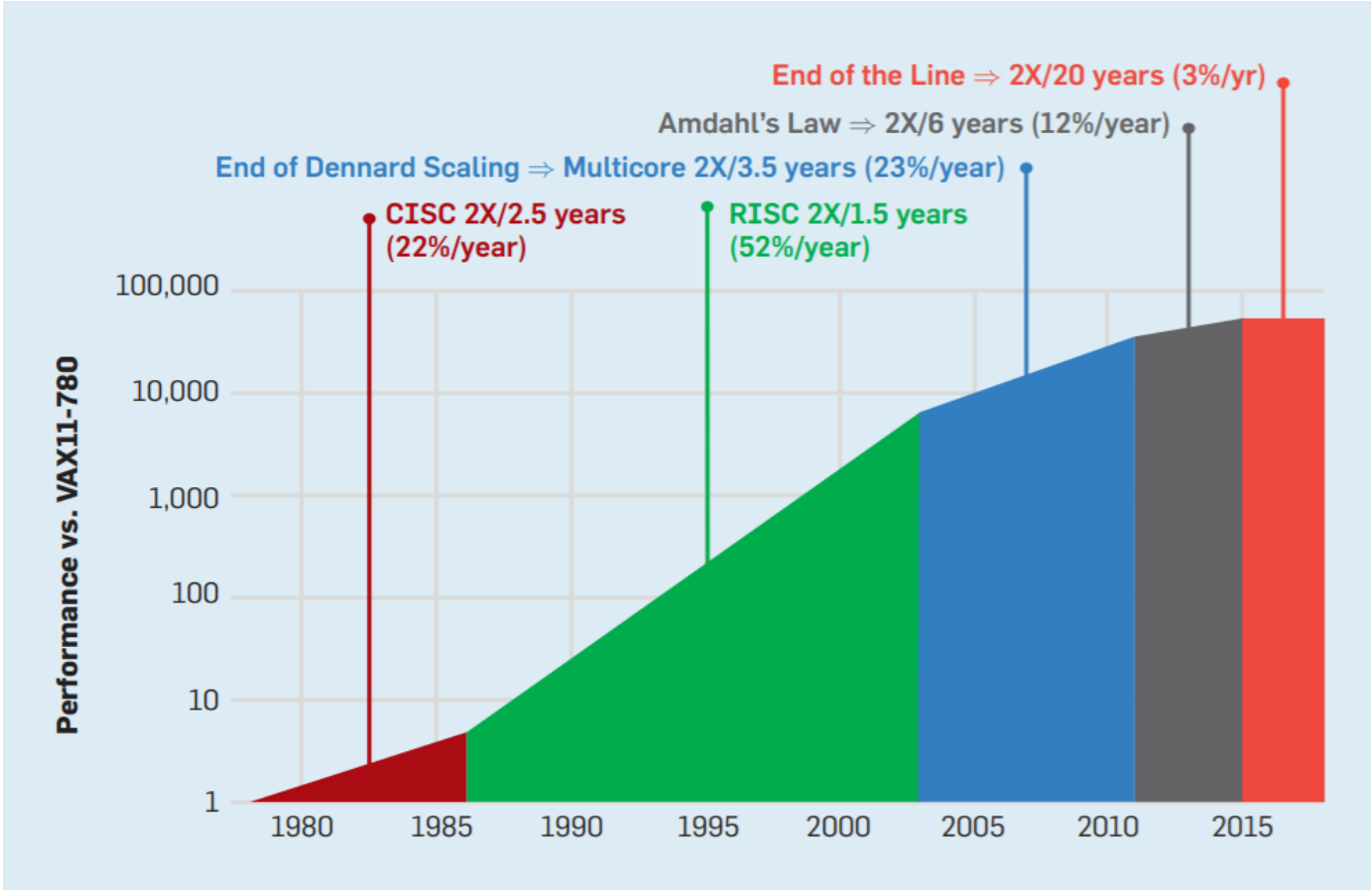
Traditional Computation - "End of the Line"

Driven by immense advances, computer architectures have pursued optimizations of

- Computation complexity
- Communication & storage
- Parallelization
- Scale

But physical limits are reaching an end

- Moore's Law
- Dennard Scaling
- Amdahl's Law

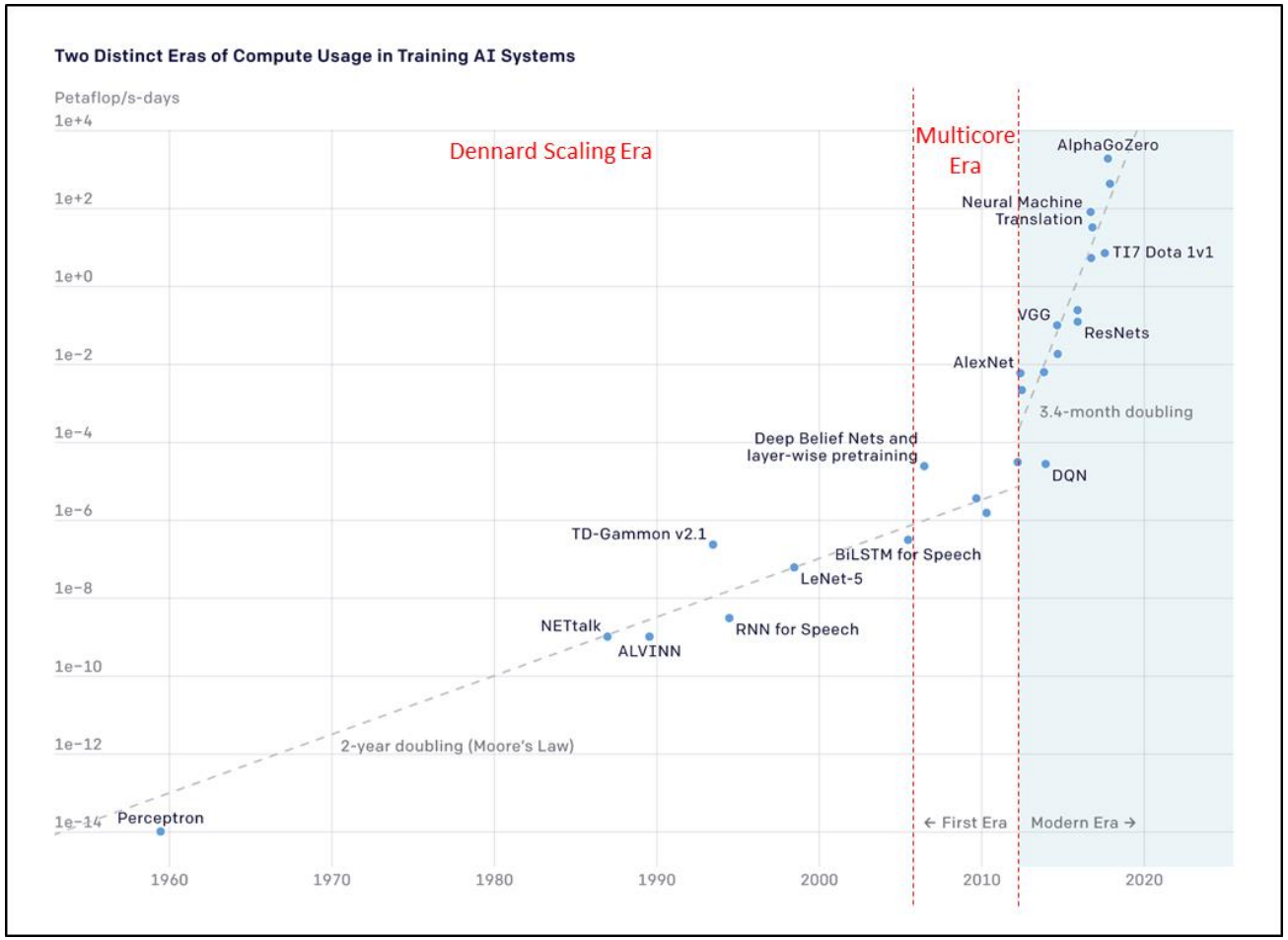


Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. Communications of the ACM, 62(2), 48-60.

Traditional Computation - "End of the Line"

Neural networks have a legacy of taxing the computational capabilities available

- In lieu of relying upon physical scaling, need a new paradigm of computing



AI and Compute (OpenAI May 2018)

Beyond Traditional "End of the Line"

Alternative paradigms – specialization

- GPUs: parallelization in compute density
- FPGAs: programmable hardware adaptivity

What is neural-inspired, neuromorphic, brain-inspired computing?

- Many terms
- Fundamental notion of taking inspiration from how the brain performs computation


THEME ARTICLE: Hot Chips

A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution

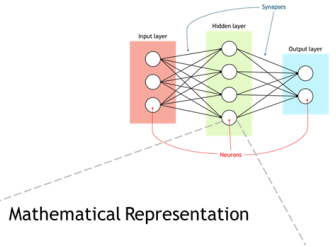
Jeff Dean, David Patterson, and Cliff Young
Google Brain

The end of Moore's law and Dennard scaling has led to the end of rapid improvement in general-purpose program performance. Machine learning (ML), and in particular deep learning, is an attractive alternative for architects to explore. It has recently revolutionized vision, speech, language understanding, and many other fields, and it promises to help with the grand challenges facing our society. The computation at its core is low-precision linear algebra. Thus, ML is both broad enough to apply to many domains and narrow enough to benefit from domain-specific architectures, such as Google's Tensor Processing Unit (TPU). Moreover, the growth in demand for ML computing exceeds Moore's law at its peak, just as it is fading. Hence, ML experts and computer architects must work together to design the computing systems required to deliver on the potential of ML. This article offers motivation, suggestions, and warnings to computer architects on how to best contribute to the ML revolution.

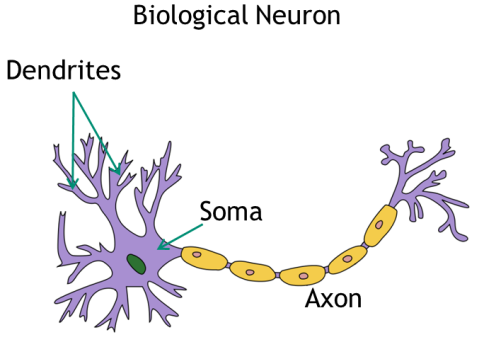
As scientists and engineers, it's our responsibility to address society's biggest problems, such as those in Figure 1. Many of the items on that list might seem to be beyond the skillset of computer scientists and engineers, but happily, that is not the case. We believe advances in ML will lead to significant advances on most of these grand challenges.

The deep-neural-network branch of ML is a transformational technology that, in the last five years, has started revolutions in several fields. In 2012, image recognition didn't work well. AlexNet[1] broke previous accuracy records of the ImageNet competition[2] and has led to a sequence of further improvements such that machines now exceed human accuracy in image-recognition tasks.[3–6] Similar breakthroughs have accelerated fields such as speech recognition,[7] web search,[8] human language translation,[9–10] medical diagnosis,[11–12] and even playing the game

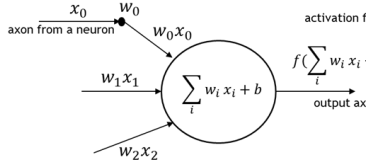
IEEE Micro
March/April 2018
21
Published by the IEEE Computer Society
0272-1732/18/$33.00 ©2018 IEEE



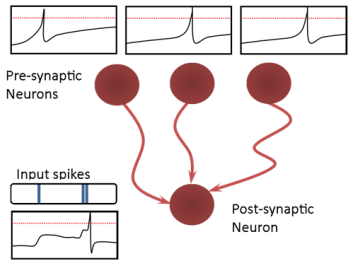
Mathematical Representation



ANN



SNN

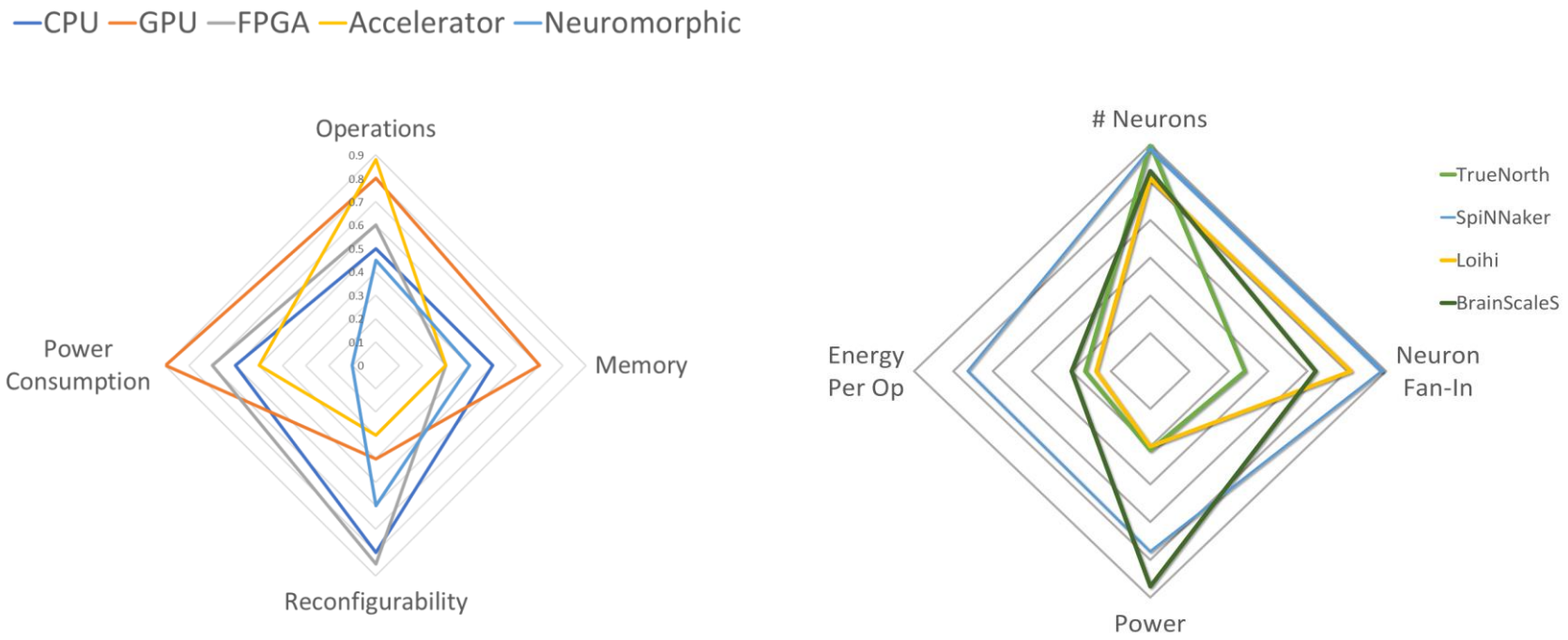


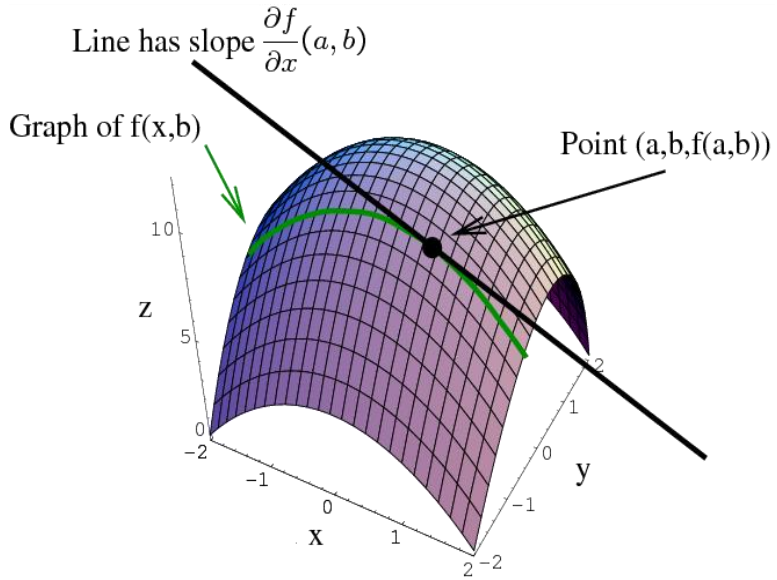
-morphous

Word Origin

1. a combining form with the meaning "having the shape, form, or structure" of the kind or number specified by the initial element, used in the formation of compound words: *polymorphous.*

Which architecture is best?





https://mathinsight.org/image/partial_derivative_as_slope

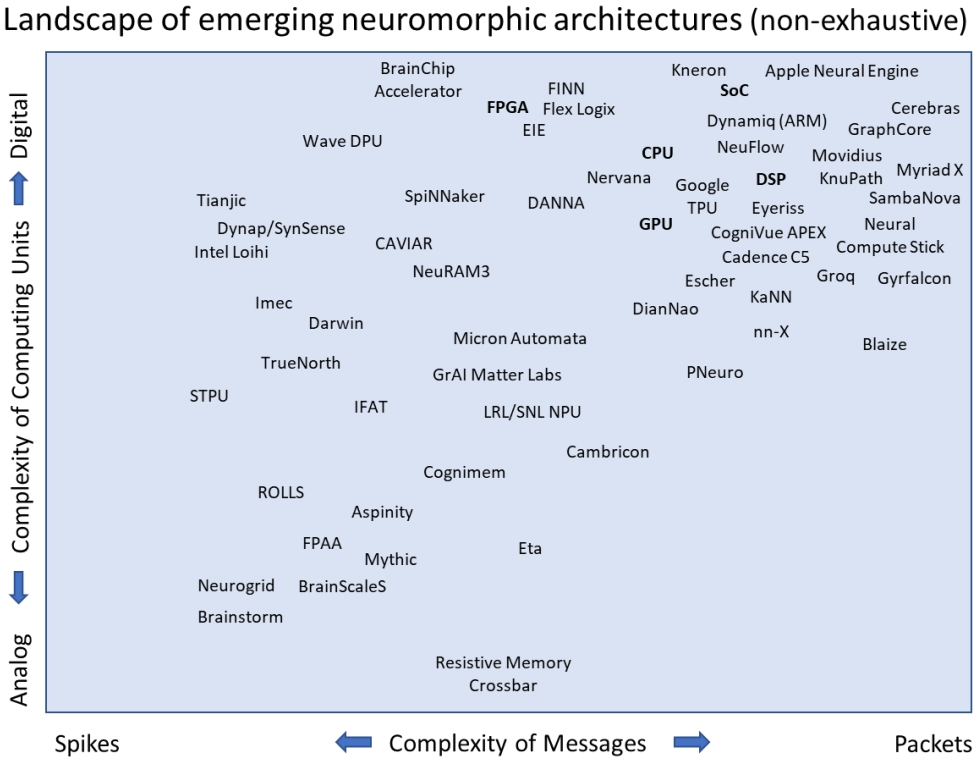
- Broad classes of architectures employ different approaches & objectives
- Neuromorphic is actively exploring architectural tradeoffs
- For a fixed architecture (hardware), the algorithms (software) which are optimal does not mean they are the best overall approach → importance of co-design

…a complex question – but is what this effort is SEEKing to shed light on

Architectural Insight

Architectural explosion

- Worldwide
- Industry as well as academia
- Leading chip vendors as well as startups
- Approaches include
  - Optimizing existing architectures for neural networks
  - Novel materials
  - Analog, digital, optical, asynchronous, event driven
- Scale & Technical maturity
- Emerging software stack



Accordingly, here we are considering insights from architectural analysis, benchmarking, and analytical assessments

Architectural Insight

Emerging trends include

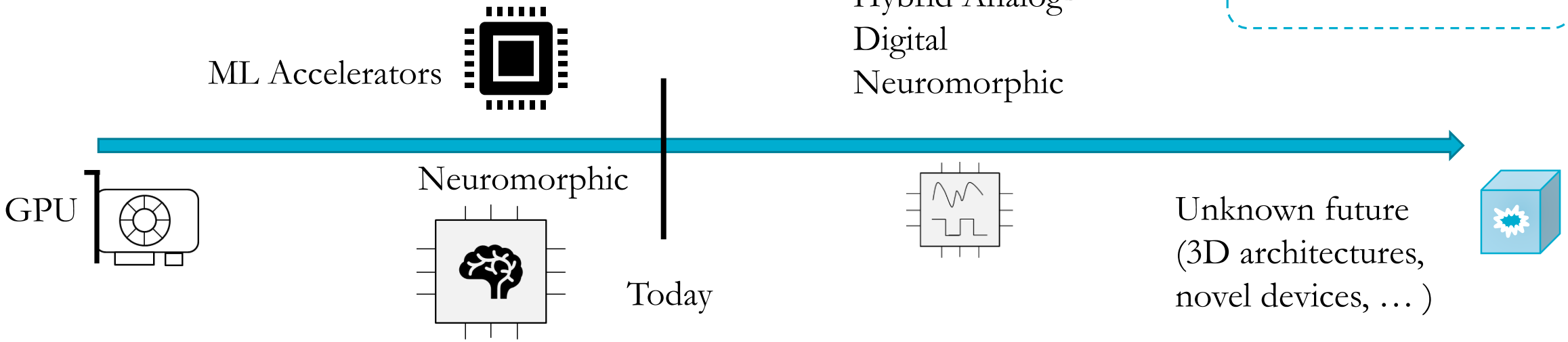
Analog – power & speed efficiencies
Applications include sensor pre-processing directly in analog & time series

Event Driven – computation & sensing
Applications such as video analysis, action recognition, signature analysis

Reconfigurable – data stream optimizations offering versatility in application

Embedded – FPGA & ASIC
Applications include tiny ML for enabling situational awareness

Physics based computing – optical and other paradigms

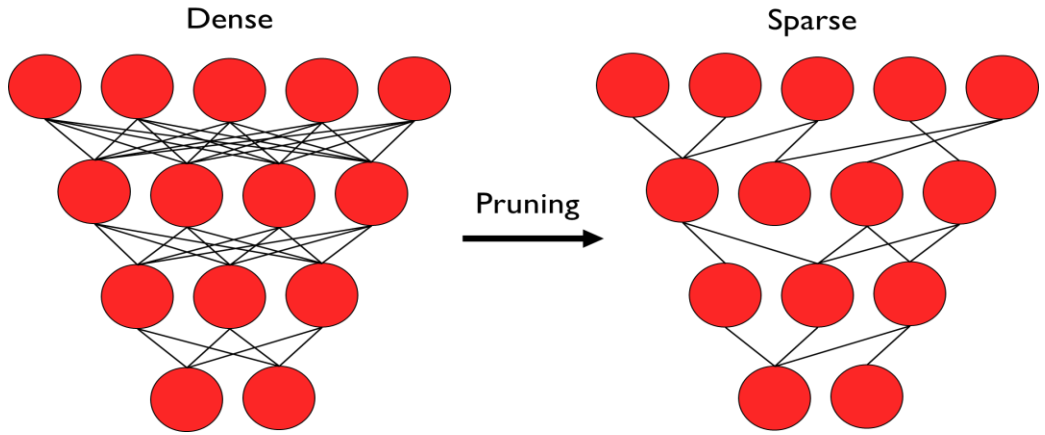


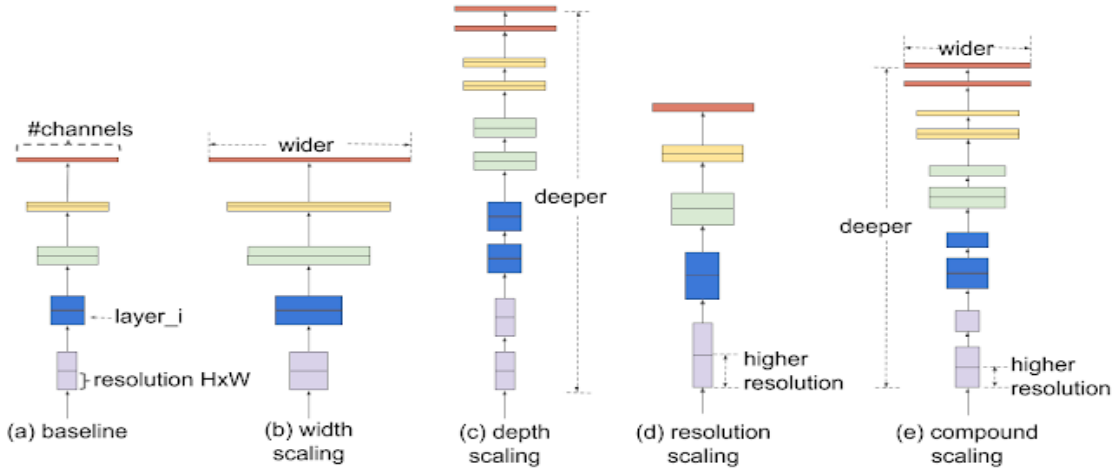
Timeline of neuromorphic impact is both now and emerging

Algorithmic Assessments

Many factors influence the efficiency of neural network computation

- Includes model size & structure, precision, pruning/compression, etc.
- Efforts here include parameter sweeps and weight pruning study of a spiking neural network
  - Highlighted ability to maintain performance while easing computational cost
- Hyperparameter Architecture Search (HAS)
  - Search over models with knobs to adjust their architectural configurations
  - Enables tuning several facets of performance



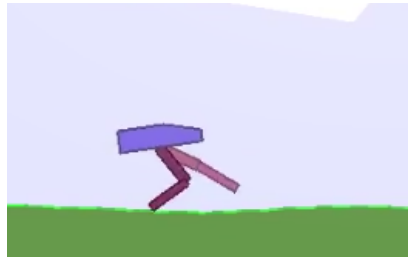


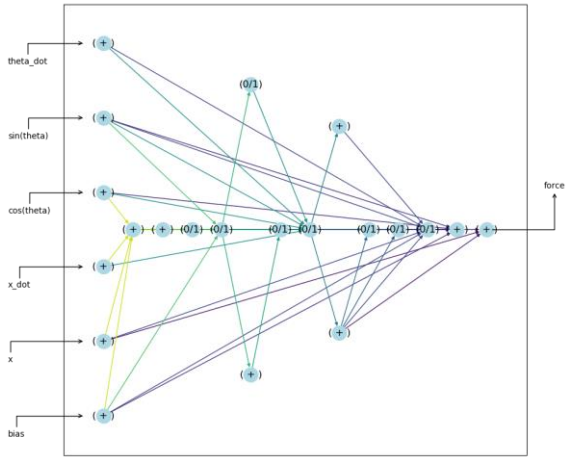
Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *arXiv preprint arXiv:1905.11946* (2019).

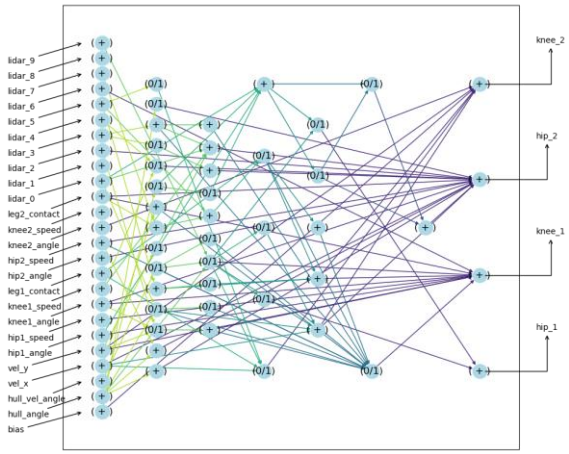
Algorithmic Assessments

Spiking Weight Agnostic Neural Networks (WANN)

- Evolutionary neural architecture search method we have extended to spiking circuits
- Efficient networks which emphasize connectivity more than weights
- Enabling explorations into properties like network size, complexity, noise resilience, multi-sensor fusion







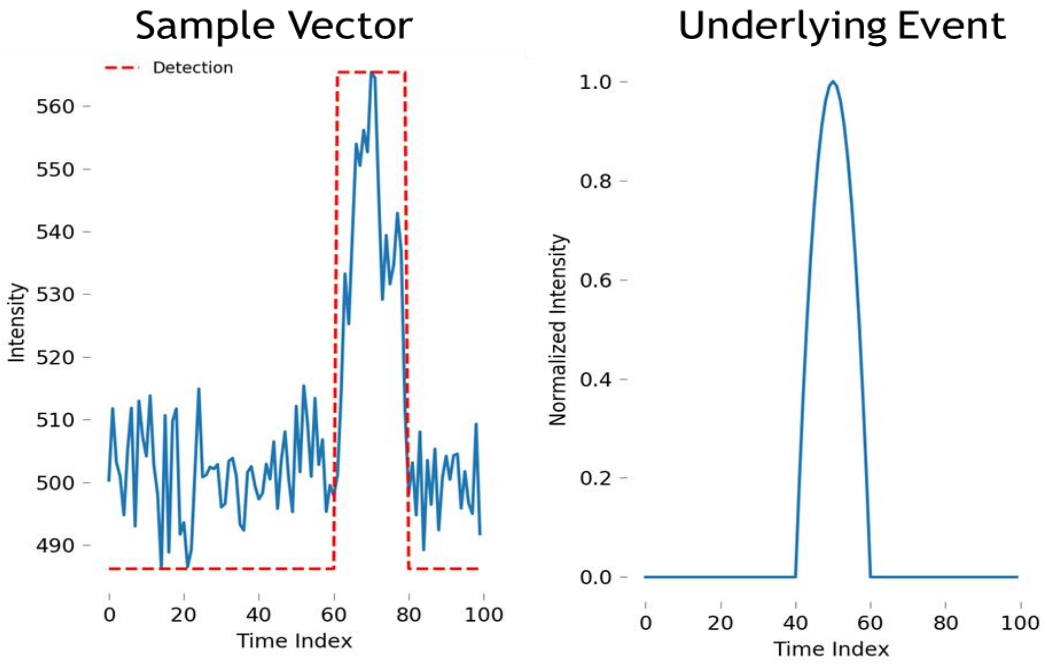
	WANN			Spiking WANN		
	Tuned Shared Weight	Tuned	Network Size	Tuned Shared Weight	Tuned	Network Size
Swingup Cartpole *	723 ± 16	932 ± 6	62	745 ± 11	912 ± 5	56
Bipedal Walker *	261 ± 58	322 ± 7	338	290 ± 22	281 ± 31	210
MNIST	91.9%	94.2%	4553	87.7%	88.2%	3150

*** mean ± std, reward over 100 rollouts**

Algorithmic Assessments

Pursuing WANN robustness

- Combining the convolutional structure of a Whetstone spiking neural network as a preprocessor & feeding extracted features as inputs into the WANN
  - Combining the two methods on a timeseries binary classification task provides a strong boost to performance (85.9% → 98.0%) while striving to improve network efficiency
- Applying noise to WANNs
  - Given emphasis on computation structure over weighted inputs
  - Input noise added with respect to the variance of each input



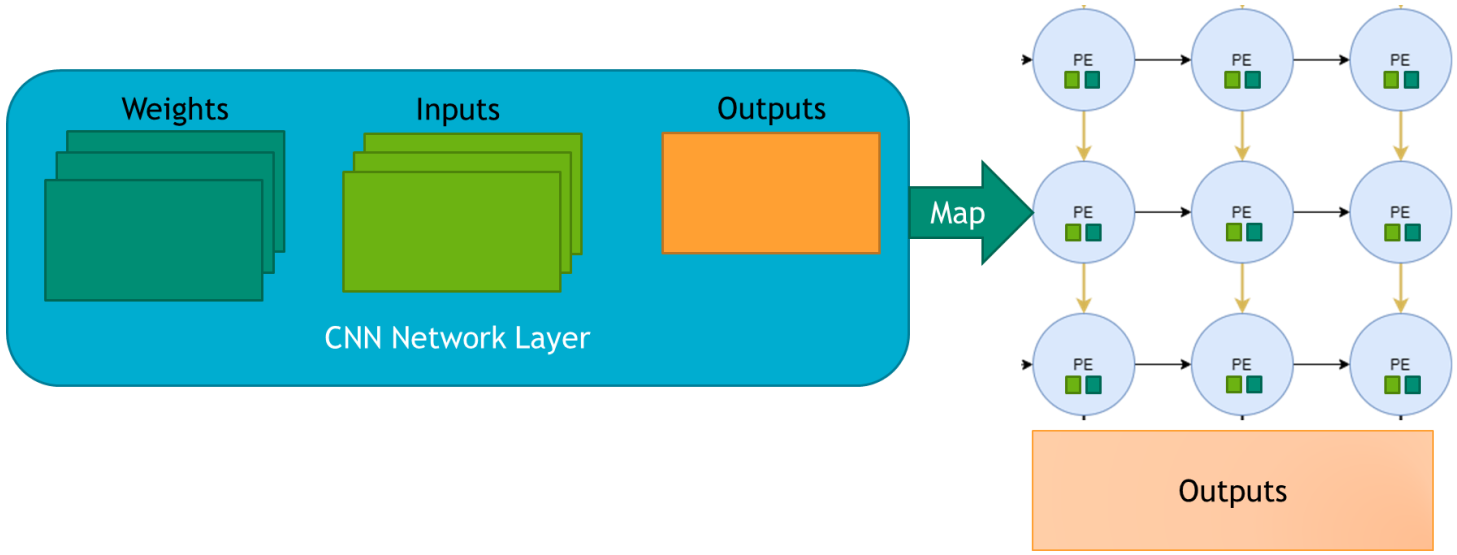
WANN MNIST (16x16) Input Noise Classification Results

Test Noise / Trained Noise	0%	10%	25%	50%	75%	100%	200%	Connections	# of Nodes
0%	91.71%	84.51%	65.76%	40.47%	29.45%	24.12%	16.62%	3403	450
10%	91.04%	89.46%	80.18%	55.96%	40.33%	33.21%	20.07%	3467	486
25%	90.74%	89.65%	84.10%	64.03%	47.61%	37.09%	22.20%	3643	558
50%	85.21%	84.20%	79.48%	65.64%	50.10%	39.42%	20.97%	3499	561
75%	73.66%	72.96%	69.58%	59.05%	46.69%	39.18%	23.35%	3534	595
100%	66.85%	66.62%	64.84%	57.29%	47.70%	39.60%	22.80%	4533	873

Co-Design Analysis

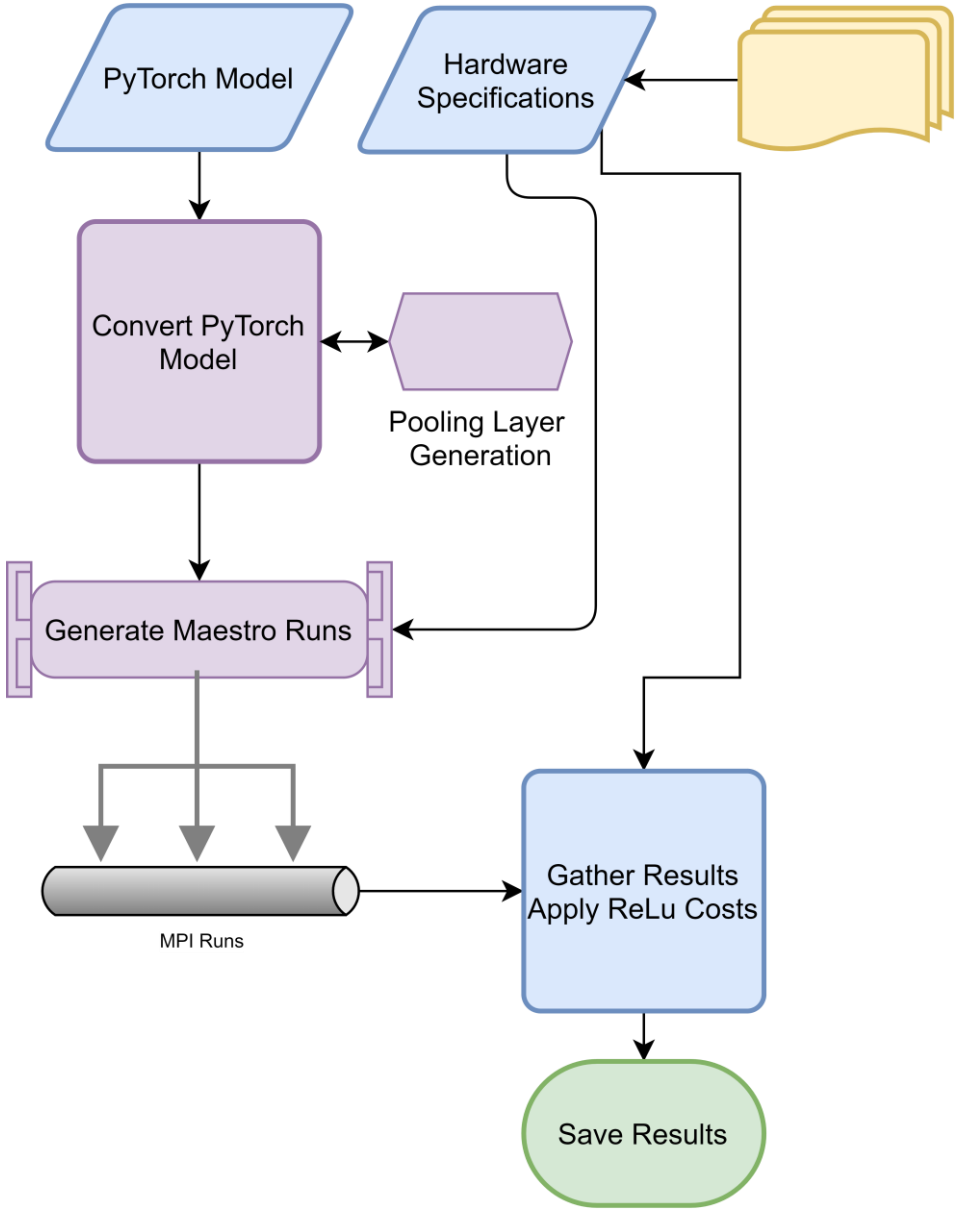
Analytical Architecture Analysis - Assayer

- Tool which given a set of neural network layers (and a dataflow) analyzes systolic array type hardware execution
- Factors in costs associated with computation, memory access, communication (references 30 nm CMOS)
- Produces various metrics (energy, latency, reuse, etc.)
- We've extended base capability to include activation function cost (ReLU) & pooling layers



Co-Design Analysis

- Built a hardware testing harness
  - Input:
    - A set of PyTorch neural networks
    - A selection of hardware parameters
  - Converts:
    - Pooling layers to convolutional layer sizes
    - PyTorch networks to a set of dimensions
  - Runs:
    - Multiple instances of Maestro
    - Controlled through MPI
  - Collects:
    - Result data from Maestro



Co-Design Analysis

Baseline sweeps –

- Applied to 12 CNN models
- 3 dataflow types (output, input/row, & weight stationary)
- Analysis where each dataflow-model combination is over 200k hardware configurations

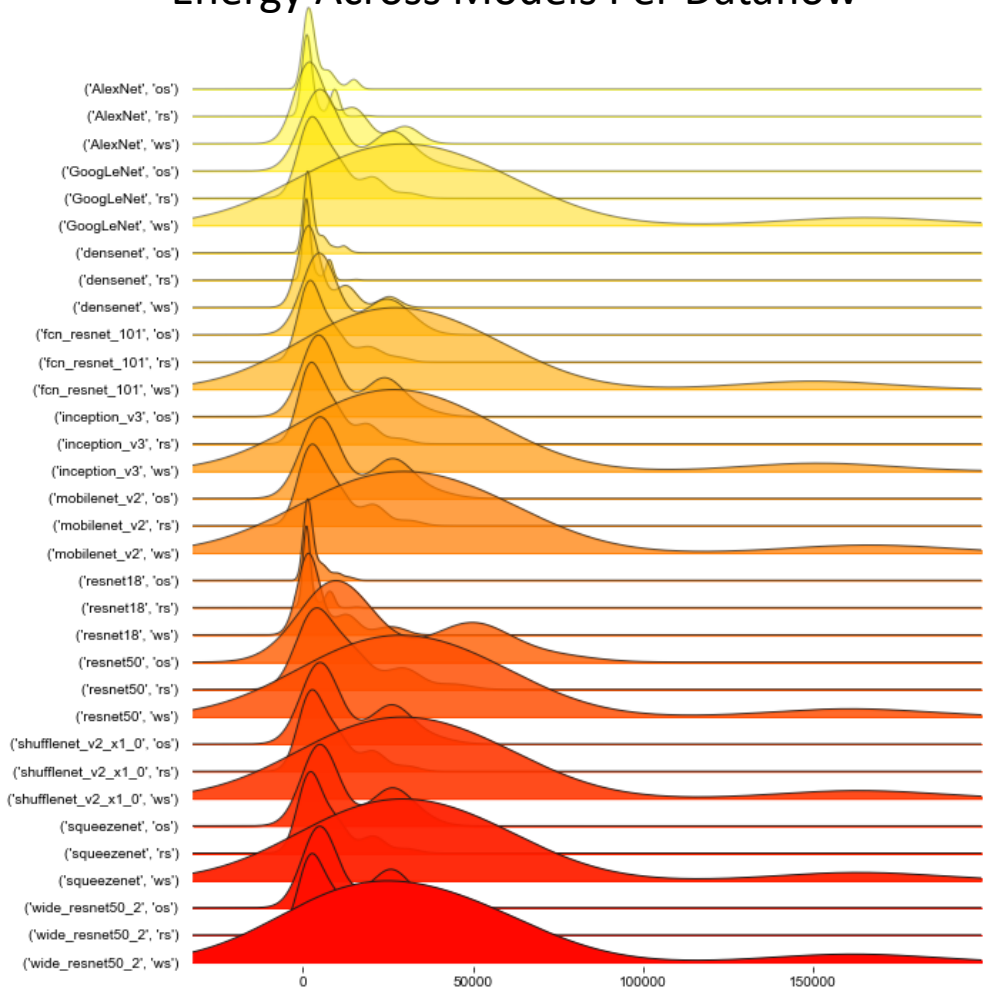
Models Run
Wide Resnet 50
Squeezenet
Shufflenet V2
Resnet 50
Resnet 18
Mobilenet
Inception V3
GooGLeNet
CN Resnet
Densenet / Densenet 201
Alexnet

Hardware Configurations
Number of PEs: 32 – 262144
L1 Sizes: 256 – 262144
L2 Sizes: 2048 – 262144
N.O.C Bandwidth: 16 – 2048
Number of ALUs per PE: 1,2,16,32

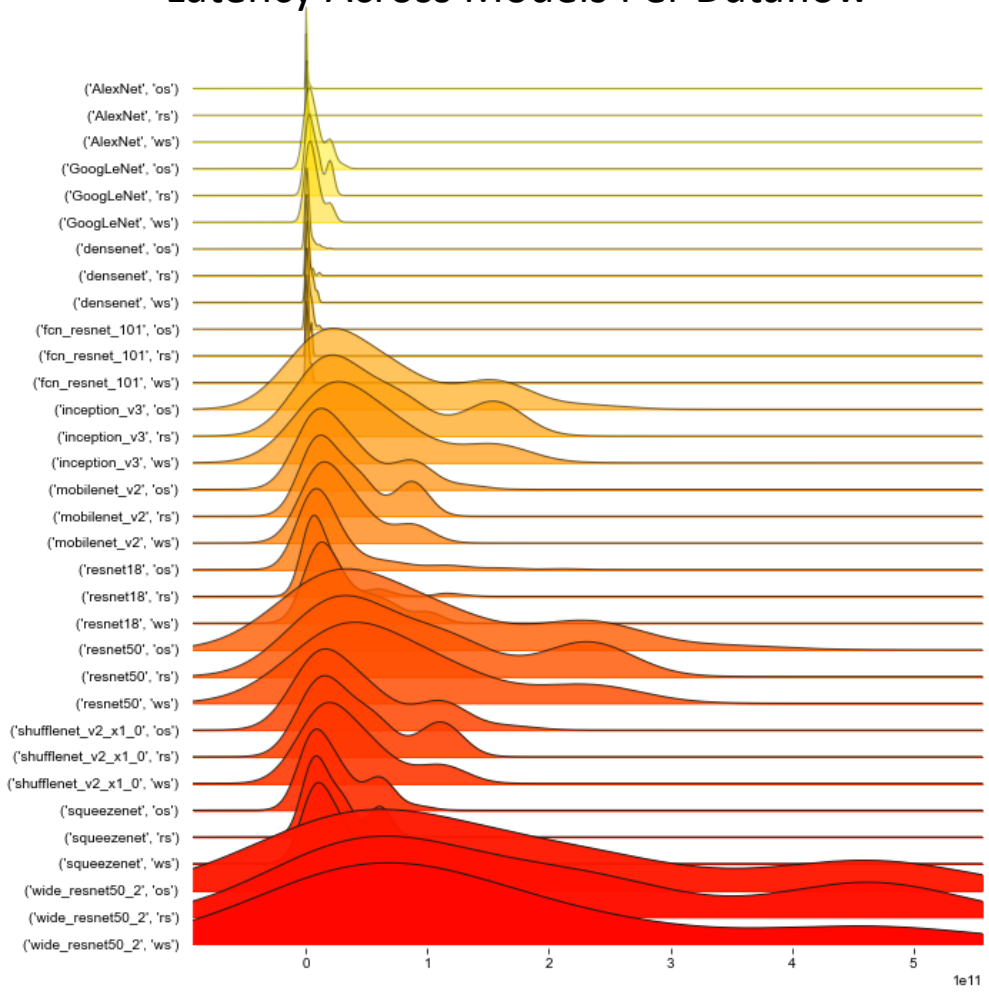
Co-Design Analysis

Per dataflow insights -

Energy Across Models Per Dataflow



Latency Across Models Per Dataflow

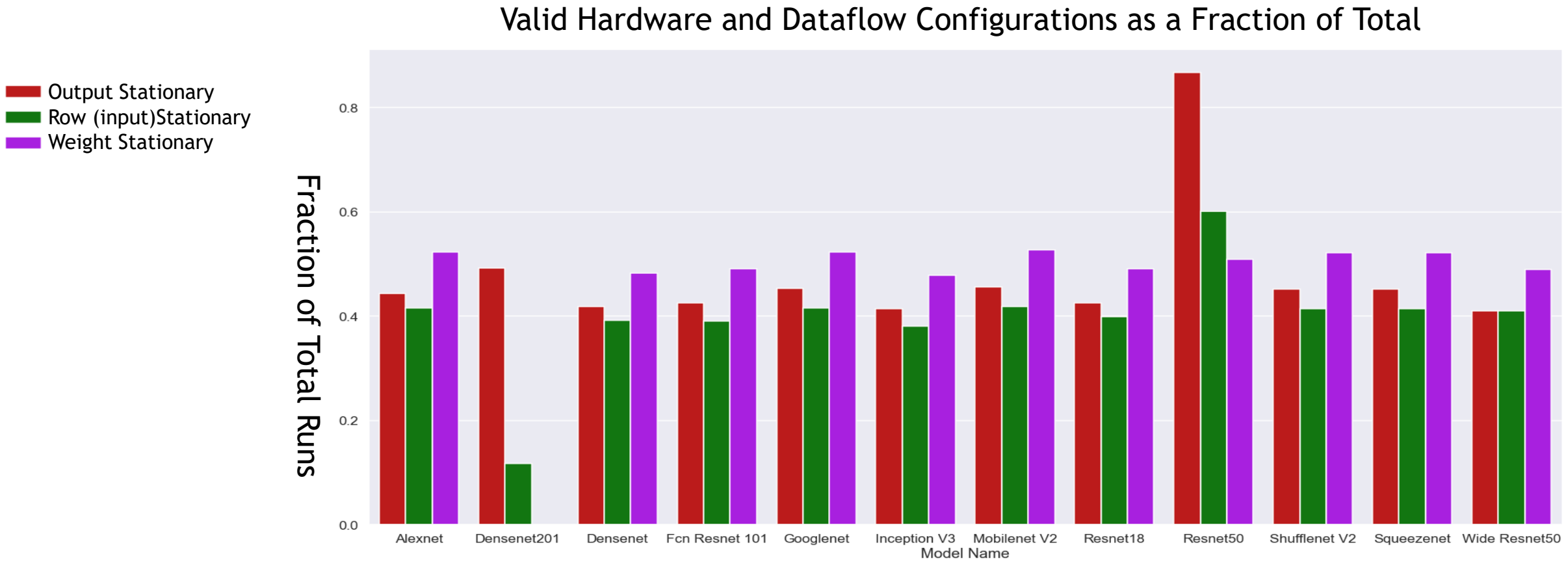


Co-Design Analysis

Finding valid dataflows is critically important

Hardware must have

- Enough L2 buffer to hold data that is reused
- Enough L1 buffer to hold working data
- Some hardware can only work with specific dataflows



Co-Design Analysis

Sweep results – Energy vs Latency mean

Across all valid hardware / dataflow combinations across all models:

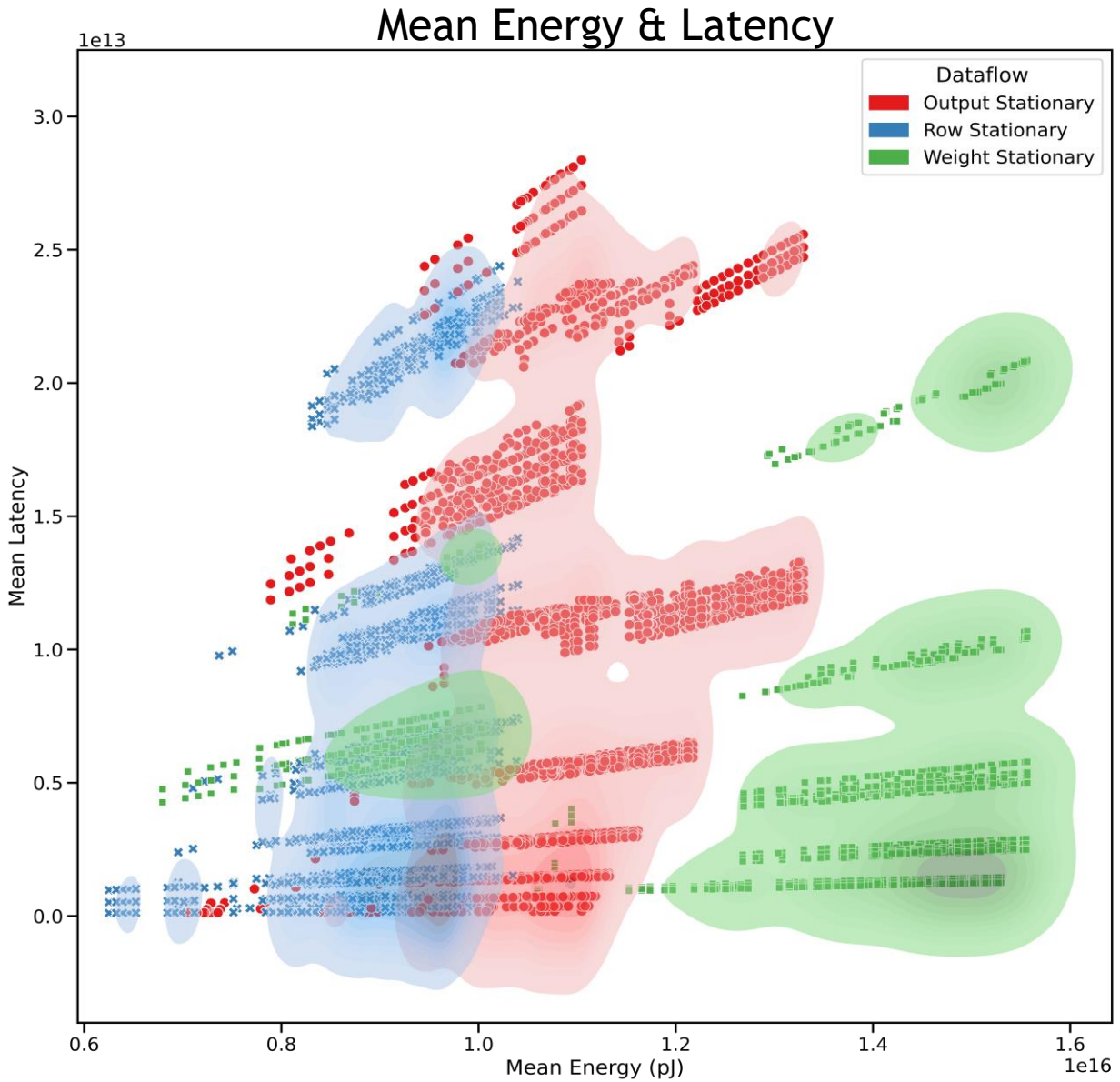
- Mean of energy and latency

Sweeps show a trend where

- Weight stationary can be the worst performing dataflow
- Row stationary is generally well performing
- There is a limit to increasing the compute power and gaining faster results

General trends allow for pathways to further detailed analysis:

- Investigate specific hardware & dataflow combinations



Co-Design Analysis

To further examine hardware configuration performance:

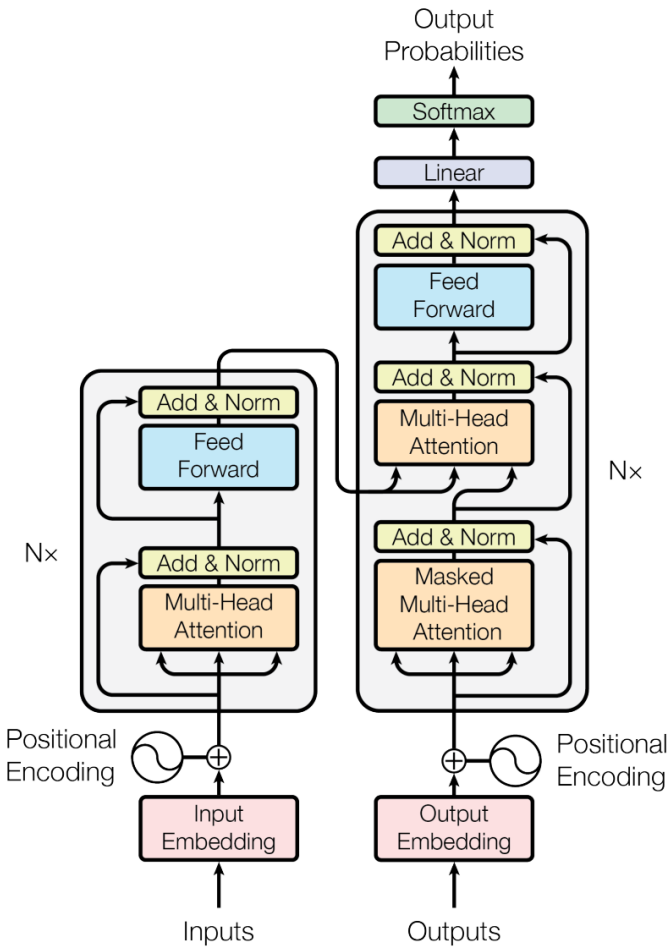
- We are adding new, non Convolutional Neural Network support to Assayer

First example network is a “Transformer” network

- Architecture first described in “Attention is All You Need”
- Uses multiple combinations of matrix multiplication and fully connected networks

The Transformer network posed new problems for mapping to systolic-array hardware:

- The model required significantly large L1 (scratch) memory to run
  - These runs **start** at 32KB of L1 memory
  - Largest CNN sweep was 32KB of L1 memory



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017): 5998-6008.

Co-Design Analysis

Smaller range of energy usage results

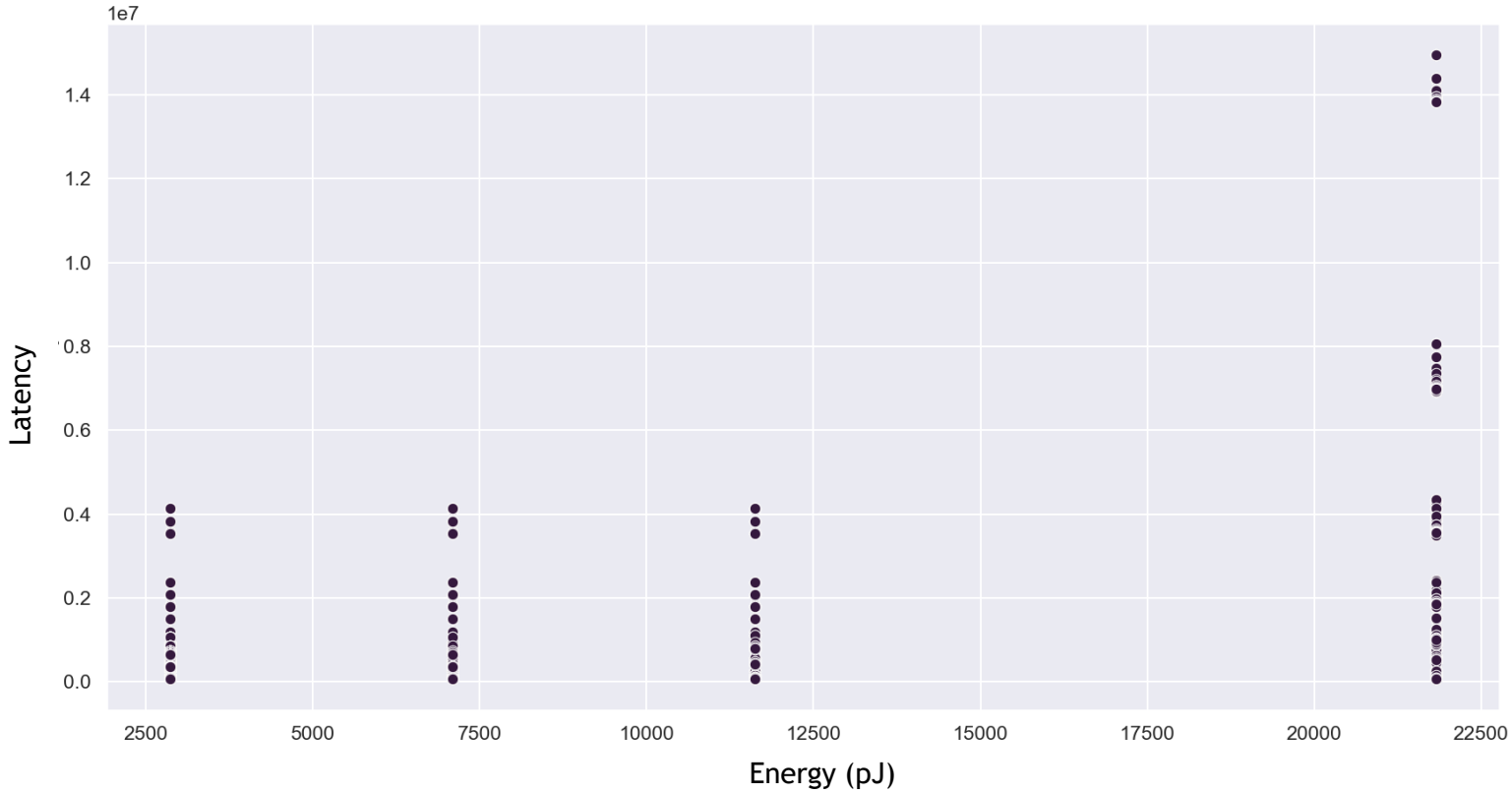
- Lower energy use overall than CNN models

Latency is wildly distributed

- Latency is also lower than many of the CNN networks

Attention is all you Need network is faster and more efficient

- *Assuming* the accelerator hardware has large enough SRAM cache

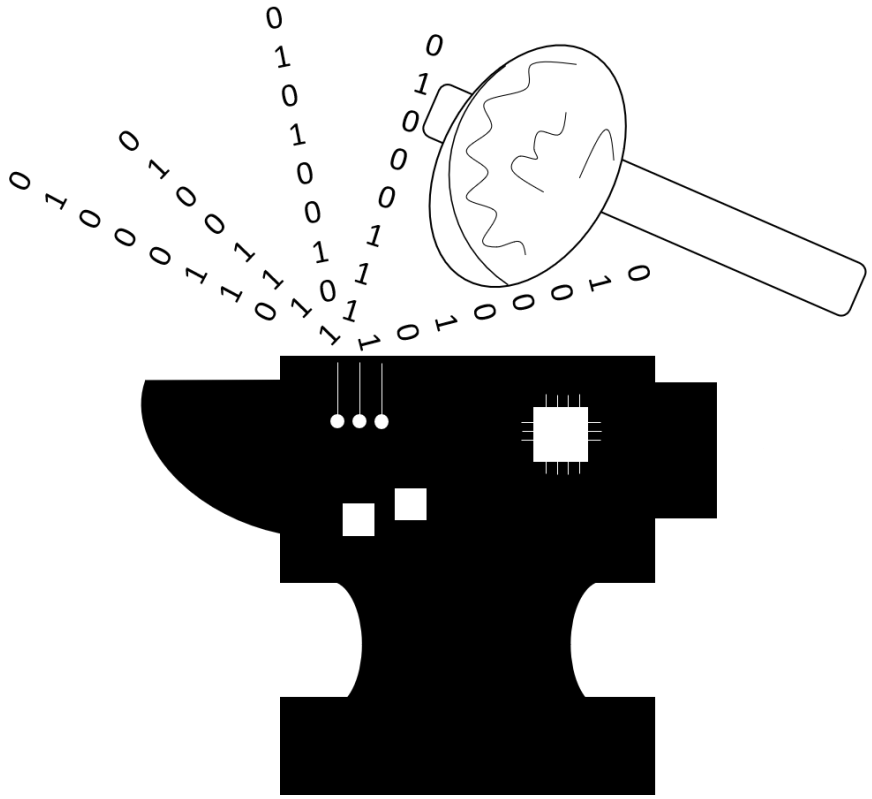


Future Work & Conclusions

- Integrating the presented efforts in architecture insight, algorithmic assessments, & co-design analysis –
  - This project will produce a report providing insight into how existing and emerging neuromorphic computation may benefit NA-22 monitoring activities as well as enable future algorithm development

- Importantly, as neural network algorithms as well as architectures are actively evolving, we believe it is necessary to consider a co-design perspective
  - Not only are there approaches available now for enabling enhanced processing at the sensor, but potential advances are hinting at orders of magnitude improvements
  - For novel computing paradigms to reach this potential, advantageous to consider interplay of algorithms & architectures

Thank you!



Los Alamos NATIONAL LABORATORY EST. 1943

Sandia National Laboratories