Sandia National Laboratories

Exceptional service in the national interest

NEUROMORPHIC COMPUTING

Towards Brain-like Energy Efficiency

Suma George Cardwell, PhD

Principal Member of Technical Staff

Sandia National Laboratories, USA

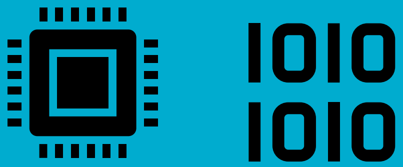
Parallel Processing and Applied Mathematics (PPAM) 2024

Ostrava, Czech Republic

September 9th, 2024

MODERN COMPUTING LANDSCAPE

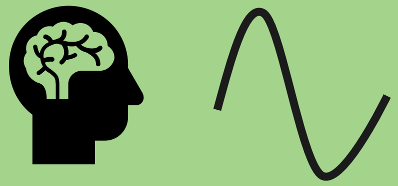
DIGITAL COMPUTING



CPU, GPU, FPGA

- Discrete-valued
- Deterministic
- High-Precision
- Simple(r) building blocks

ANALOG COMPUTING

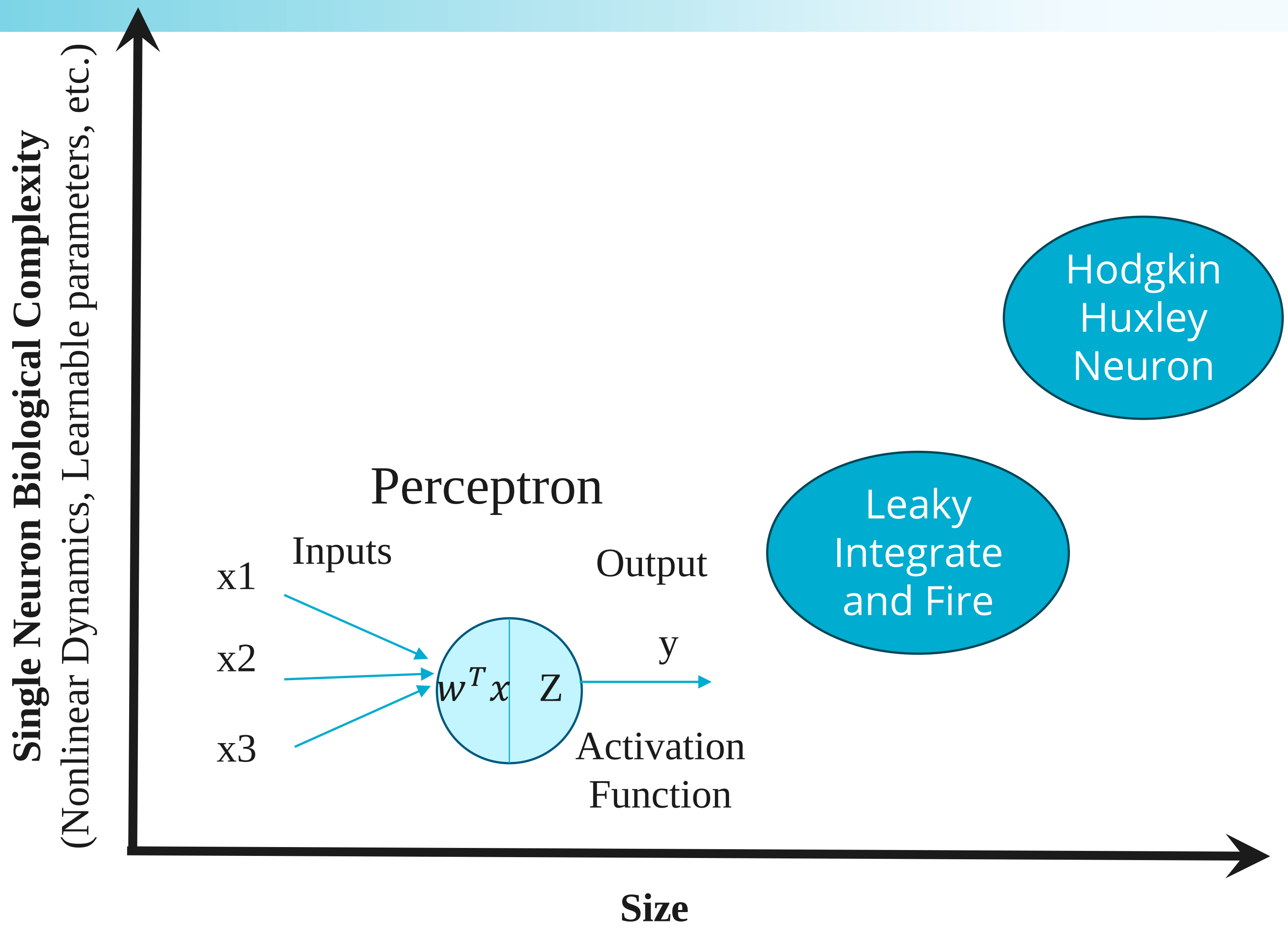


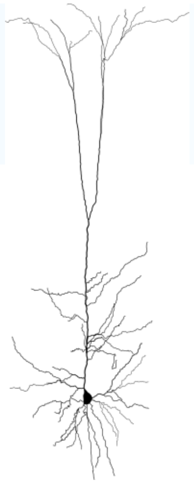
Brain , FPAA

- Continuous-valued
- Stochastic
- Lower-Precision
- Complex building blocks (neurons)

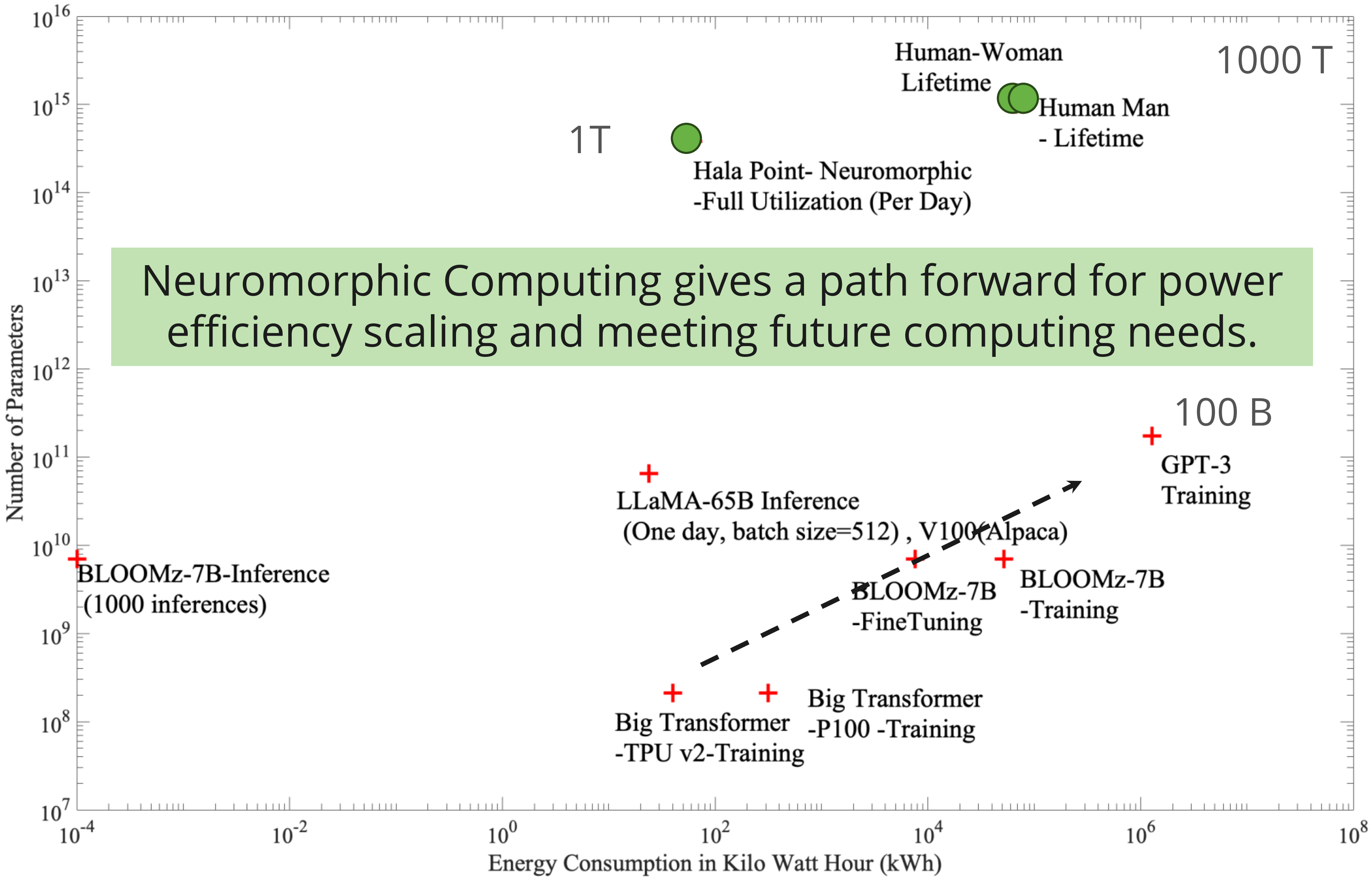
SINGLE NEURON COMPLEXITY

Biological Neurons have rich dynamics





ENERGY CONSUMPTION OF MODERN AI SYSTEMS



NEUROMORPHIC COMPUTING

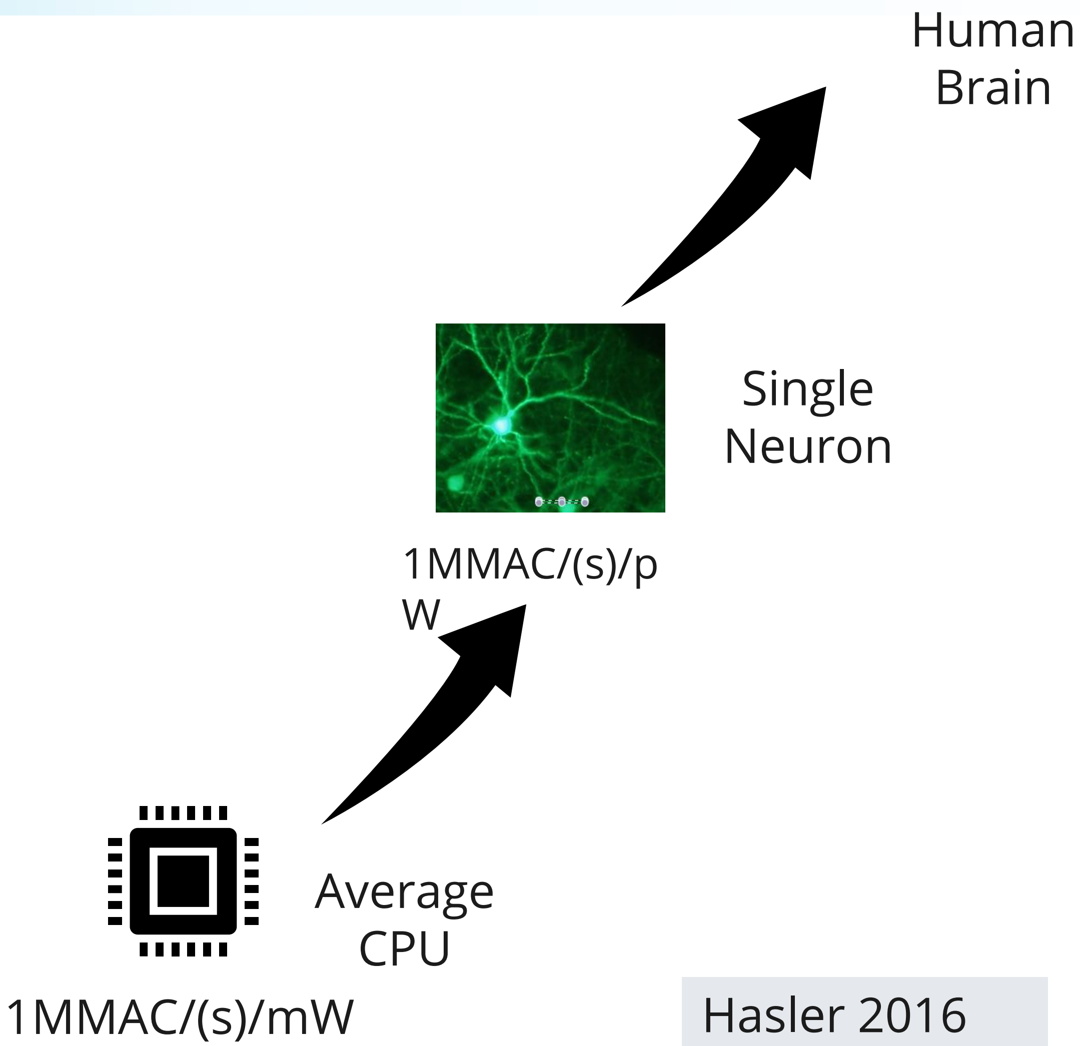
TAKING INSPIRATION FROM BRAINS

Functionality

- Solves ill-structured problems with little training
- Online learning
- Transfer learning
- Continual learning

Attributes

- Computational Memory/In-memory computing
- More complexity and computation/ single unit
- Self-organizing/ Reconfigurable
- Spiking/event-driven communication
- Sub-threshold computation
- Stochasticity as a feature
- Dense local connectivity
- Massively parallel computation

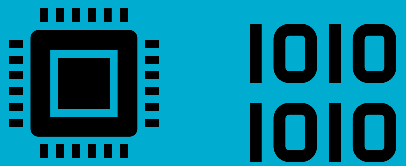


Hasler 2016

NEUROMORPHIC COMPUTING

Varied solutions proposed spanning digital, mixed-signal and beyond-CMOS

DIGITAL CMOS



CPU, GPU, FPGA

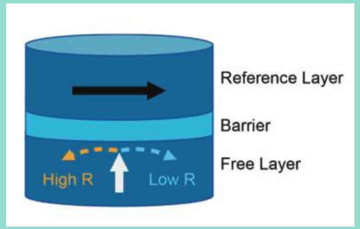
- Scaled to 1.15 B/ 2B neurons
- Deterministic
- High-Precision

ANALOG/ MIXED-SIGNAL

NVM, Analog, Sub-threshold

- Scaled to 1 M
- Analog/ Stochastic
- Low-Precision

BEYOND-CMOS



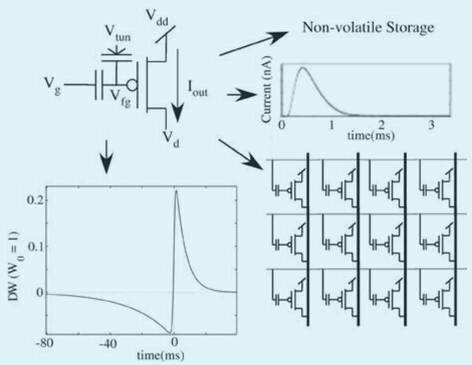
Memristors, FeFETs, MTJs etc.

- Large focus on in-memory computing
- Analog/ Stochastic
- Low-precision
- Integration with CMOS

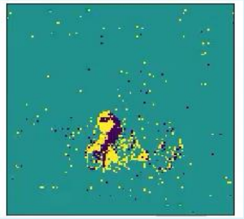
NEUROMORPHIC BUILDING BLOCKS

Neuromorphic offers computational richness

In-memory Computing With Synapses

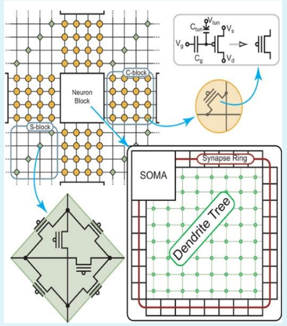


Sensory Processing



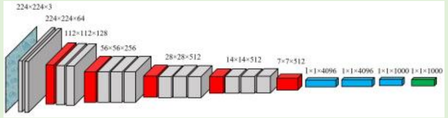
Silicon Retina/ Event Sensor
Silicon Cochlea etc.

Neural Processing



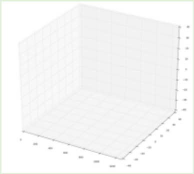
Dendrites, Learning, Multi-modal

APPLICATIONS

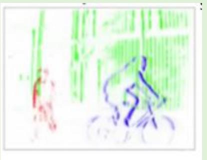


AI/ML
(ANN, SNN)

Brain-Inspired
Algorithms



Scientific
Computing



Edge
Computing

High Performance
Computing

NEUROMORPHIC SUPERCOMPUTER

Neuromorphic systems with billions of neurons

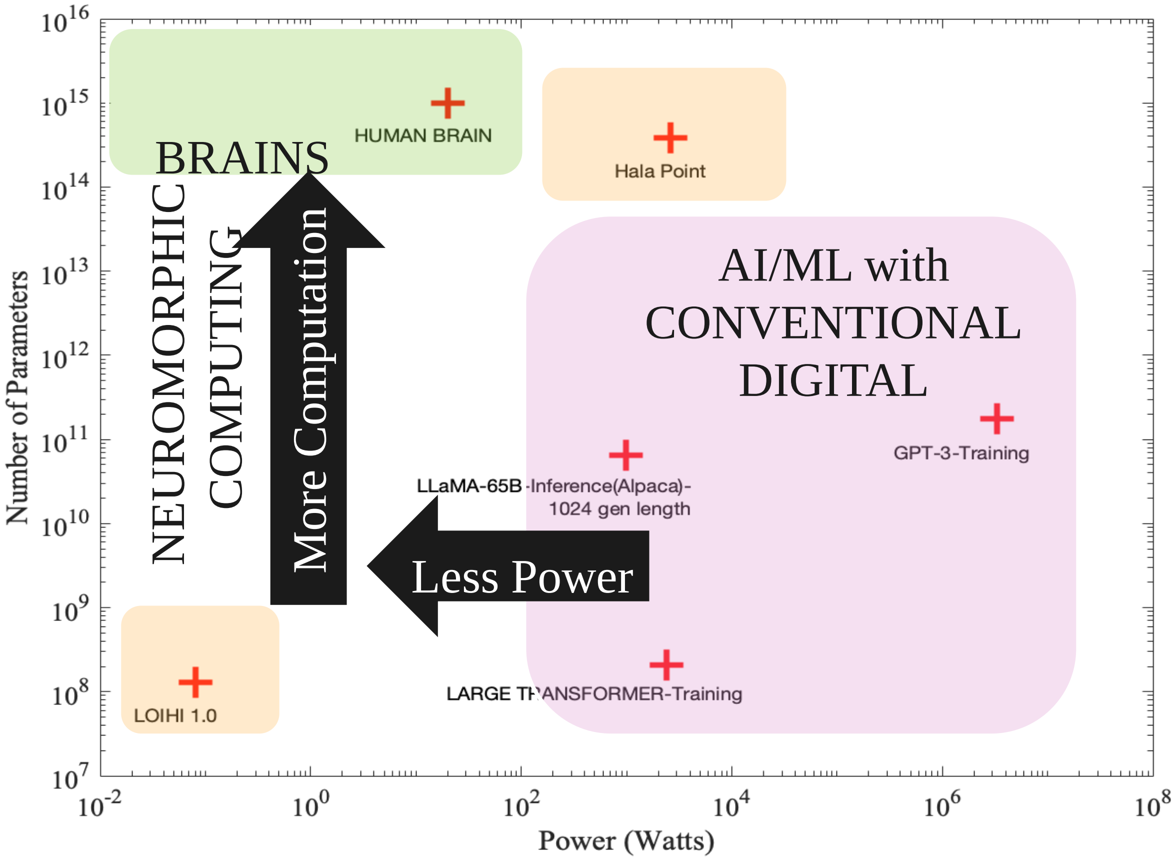
- SNL hosts Intel's Hala Point, which utilizes the Loihi 2 chips to realize one of the largest neuromorphic supercomputers in the world.
- **1.15 Billion neurons** and **trillions of synapses** with a total power consumption of only **2600 W**
- The Hala Point system incorporates 1,152 Loihi 2 processors, each of which can simulate a million neurons.
- Capable of **15 TOPS-per-watt at 8-bit precision** and does not require extensive data-processing or batching in advance.

NEUROMORPHIC COMPUTING FOR PERFORMANCE

Next-generation Neuromorphic Architectures for HPC and Edge

Key attributes we need:

- Increase complexity to get more computational/unit
- Leverage stochasticity as a feature
- Scalability and Complexity
- Codesign across scales
- Software tools for Codesign
- Application specific Solutions
- Design of Heterogeneous Architectures

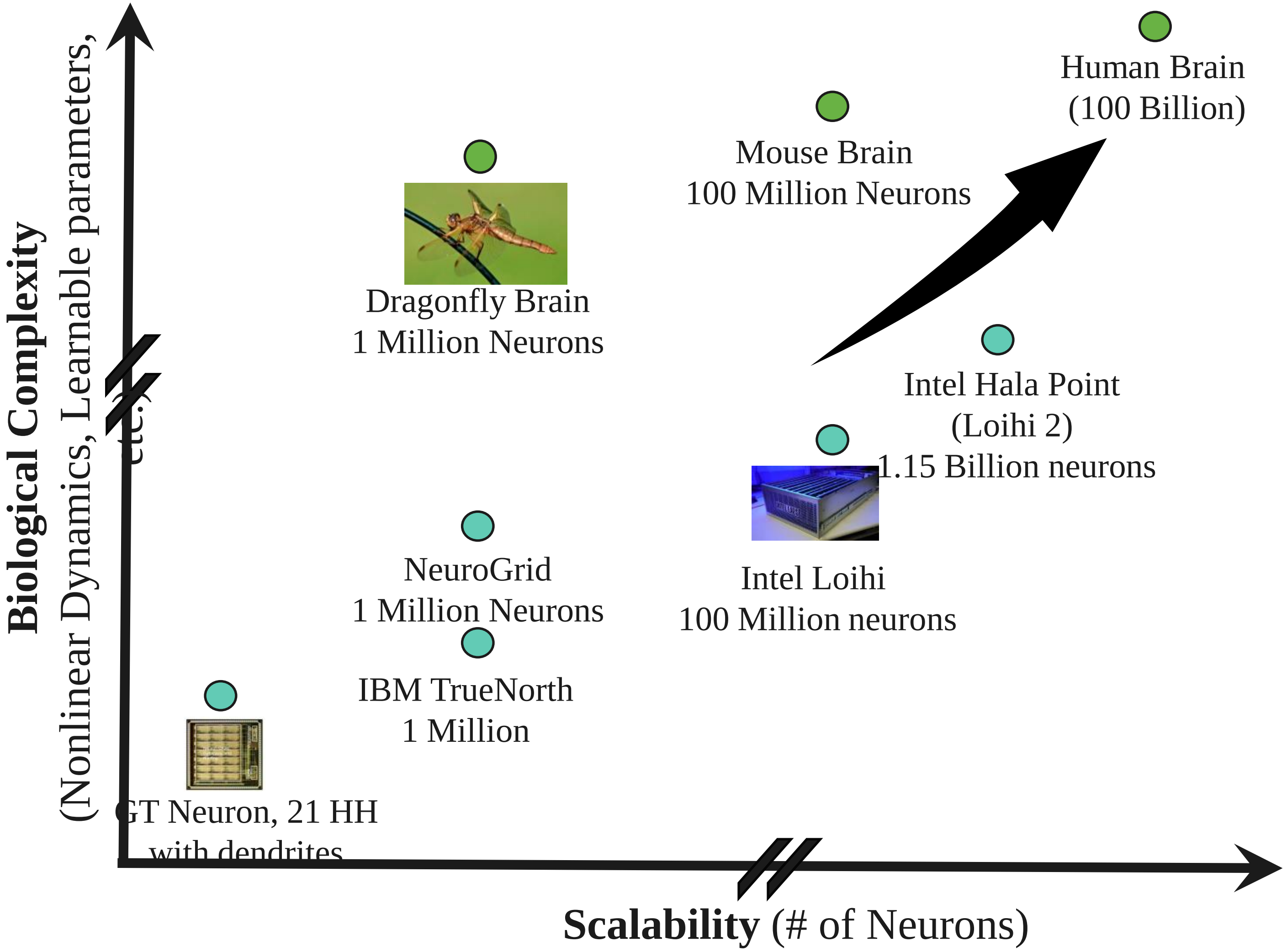


Cardwell et al., Neuromorphic IOP 2024
Under Review

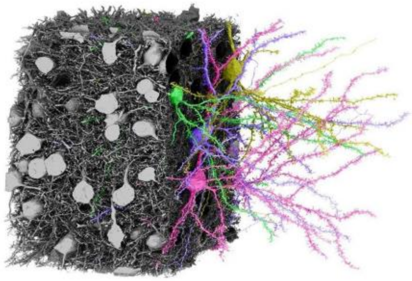
OPPORTUNITIES IN NEUROMORPHIC COMPUTING

SCALABILITY VS. COMPLEXITY

NEUROMORPHIC COMPUTING NEEDS BOTH!



- We need to increase computational efficiency as well as computational density for neuromorphic systems.
- We can improve the complexity of a single neuron.
- Novel devices and materials can help bridge this gap but codesign is a challenge.



Biological neurons have rich dynamics and a lot more computational power.

DENDRITES IMPROVE COMPUTATION WITHIN A SINGLE NEURON

SIMPLE NEURONS

- **Simpler neuron models requires larger neural networks.**
  - Power hungry ANNs
  - Inefficient scaling
  - Inefficient hardware implementations
  - Focus only on synapses and

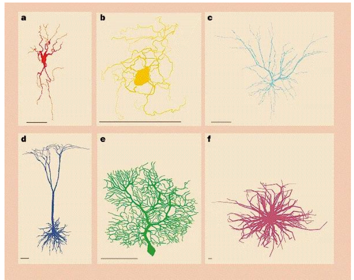
<u>Our Current Research</u>

- Dendrites supporting shunting Inhibition
  - Pattern Recognition (NICE 2023, ICONS 2023)
  - Direction-Selective (ICONS 2024)
- Software library for analog hardware-based dendrites (NeurIPS workshop, NICE 2023)
- DEND-NET: SNN with dendrites (Neuro IOP 2024 - *In Review*)
- Neuromorphic Design Space Exploration with SANA-FE tool in collaboration with UT Austin.

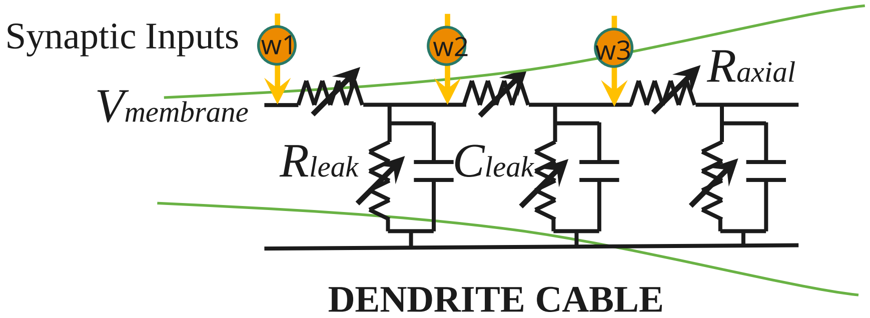
COMPLEX NEURONS

- **Active Dendrites improves performance of Artificial Neural Networks**
  - More energy efficiency ("Neural Network within a neuron"), non-linear filtering
  - Heterogeneity
  - More computations/unit
  - Better Connectivity and fan-in (3D architectures)
  - Scalable with CMOS+X approaches.
- **Plastic Synapses**
- **Learning Mechanisms**

Heterogeneity in Dendrites



Segev, Nature 1998

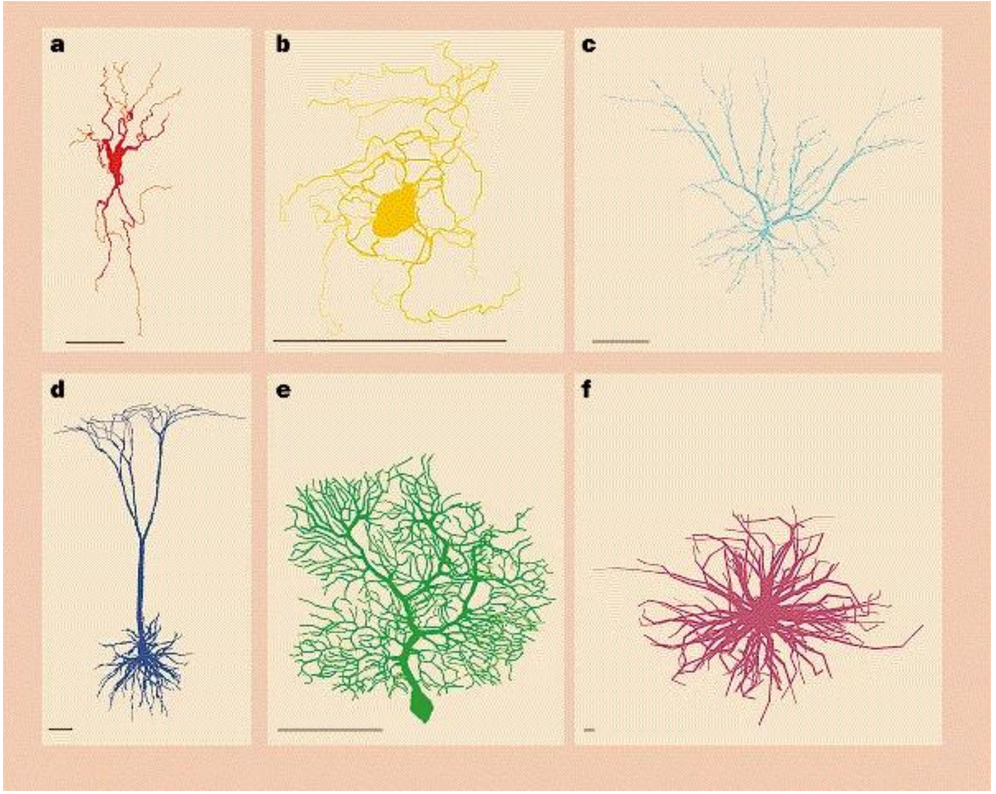


Analog devices and circuits well suited to model dendrites efficiently in hardware.

COMPUTATION USING THE DENDRITES

Dendrites are not just wires!

- Over 70% of the neuron's volume (Stuart 2016)
- Great diversity of dendrites within a single brain and across animal species.
- Insights from neuroscience are foundational to the pursuits of neuromorphic computing.
- Biological dendrites are known for their complex physical structures that incorporate significant fan-in (~10,000 inputs).



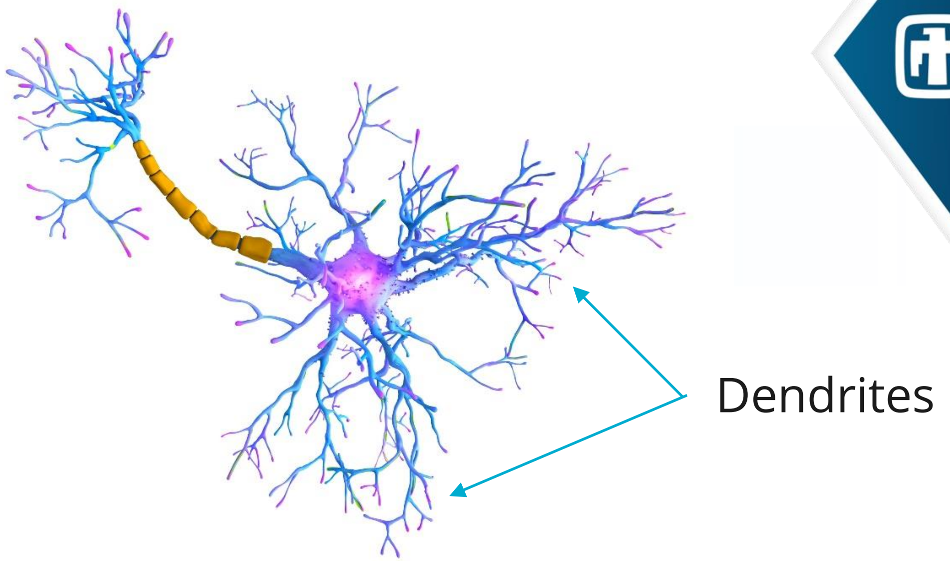
Segev, Nature 1998

DENDRITIC TOOLKIT FOR COMPUTATION

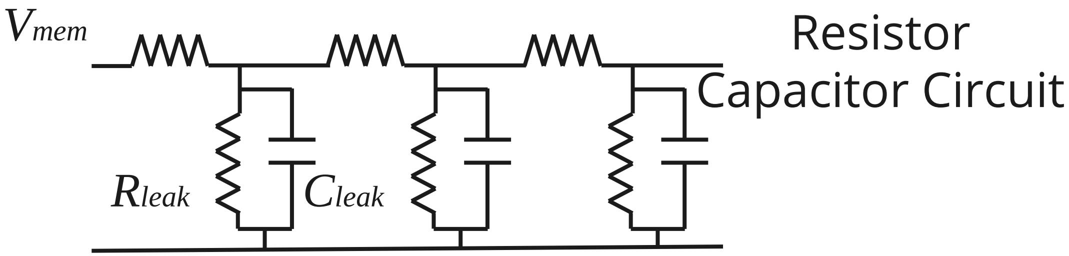
Dendrites are tree-like structures that connect neur

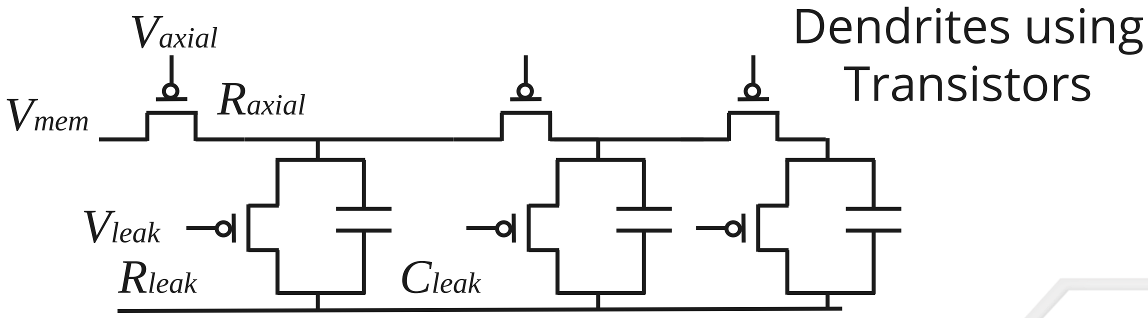
- **Dendrites are not *just* wires!**
- They can perform interesting computation like:
  - Coincidence Detection
  - Current Summation
  - Directional selectivity
  - Non-linear filtering
  - Amplification of Synaptic inputs

London 2005, Poirazi 2020



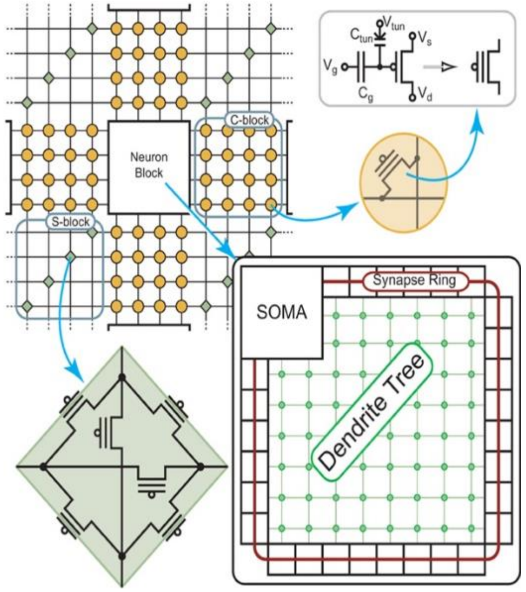
Increased Connectivity and Computation





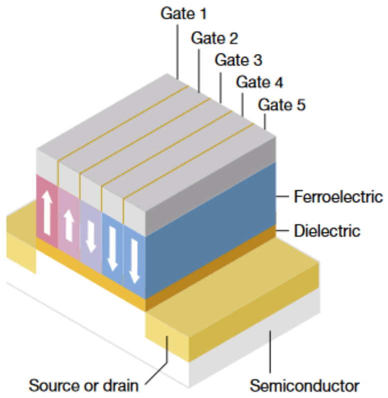
NEUROMORPHIC ARCHITECTURES WITH DENDRITES

Dendrites have been modeled to different degrees in neuromorphic hardware.

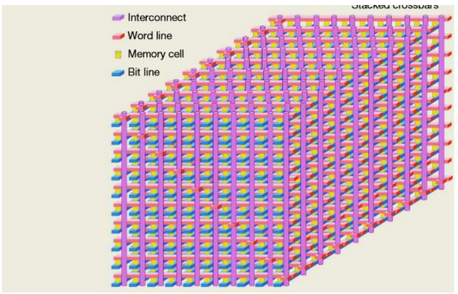


Hodgkin Huxley Neurons with Active Dendrites
Ramakrishnan, 2013

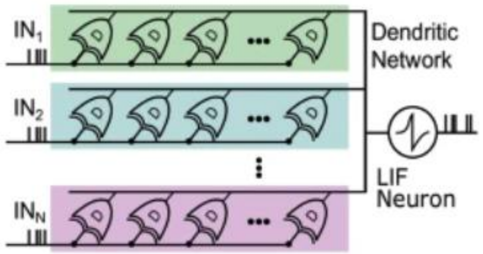
- Active Dendrites with Calcium and NMDA channels: BrainScales
- Floating-gate based active dendrites: Georgia Tech Neuron Chips
- Dendrocentric Learning with mult-gate FeFETS (Boahen 2020)
- DenRAM: Dendrites with Resistive RAM (Payvand 2024)



NanoDendrite Multi-gate FeFET



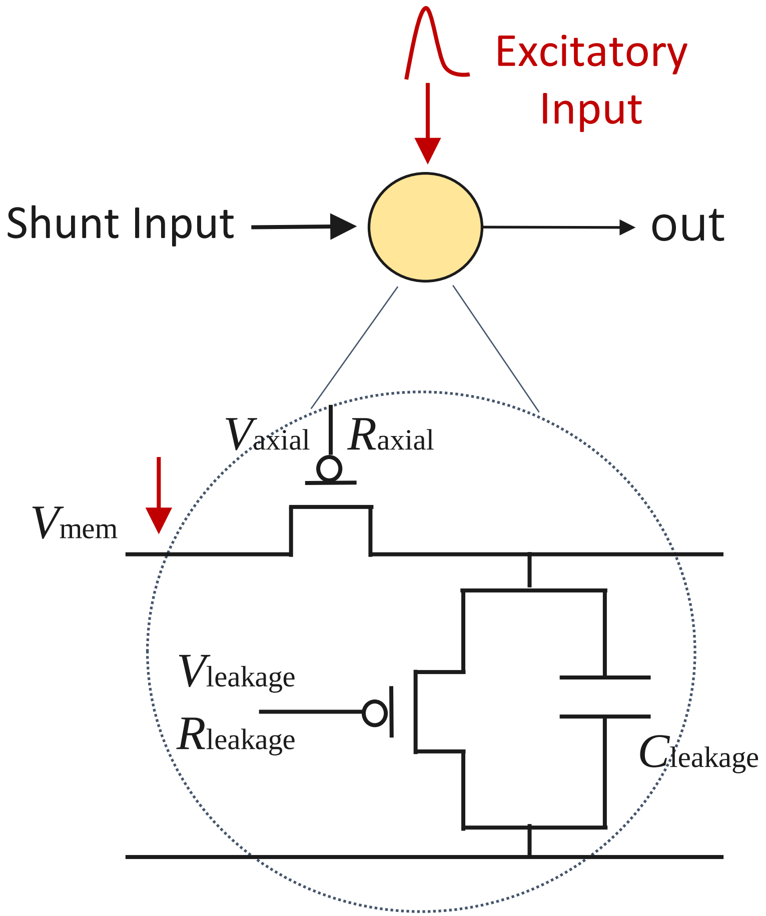
Proposed Dendrite 3D Architecture

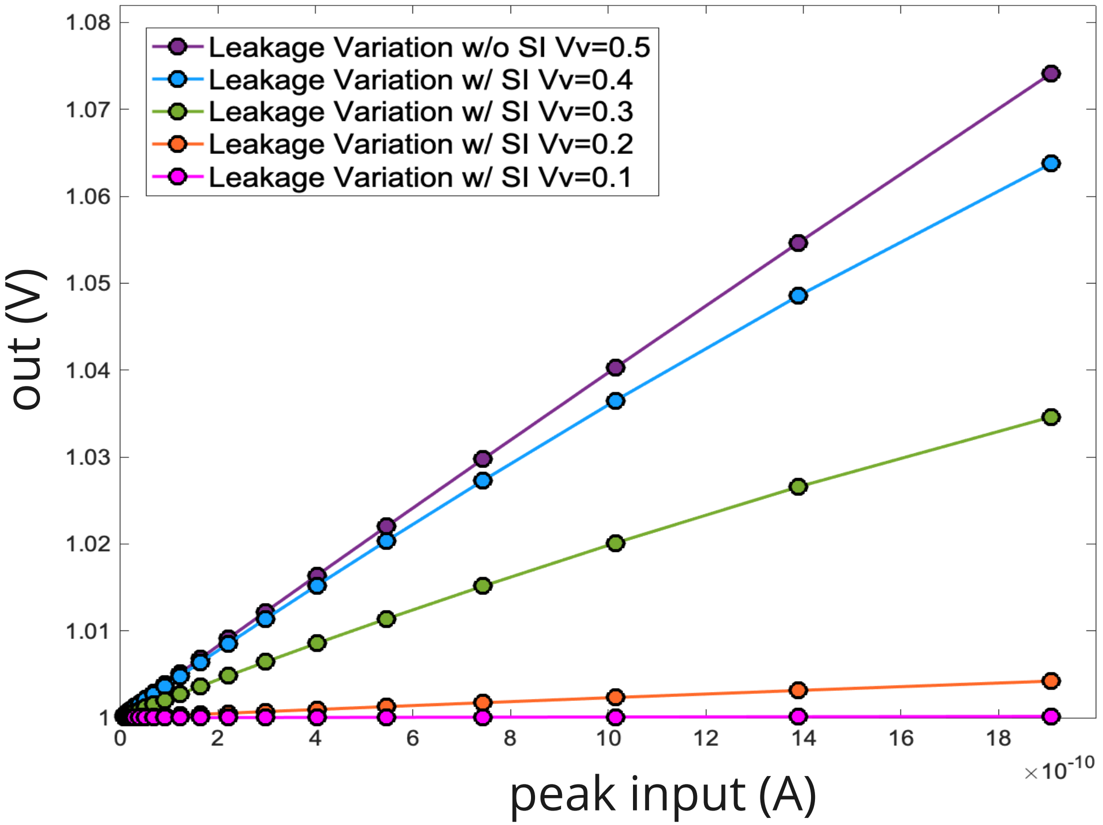


DenRAM with RRAM dendrites

ACTIVE DENDRITES

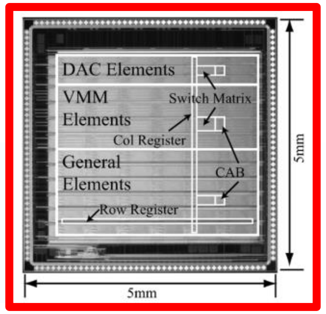
Using shunting inhibition for gain modulation





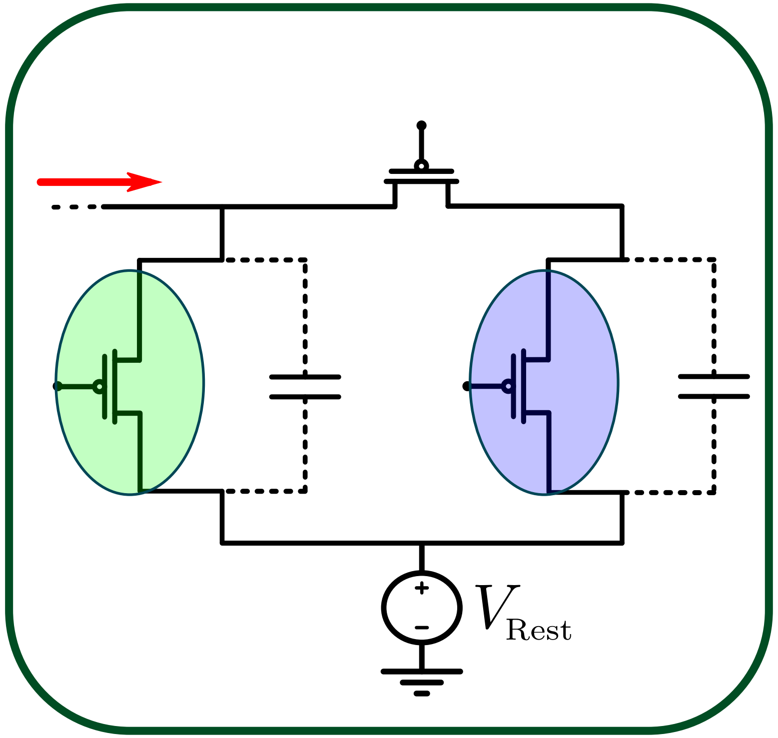
DENDRITES FOR DIRECTION SELECTIVITY

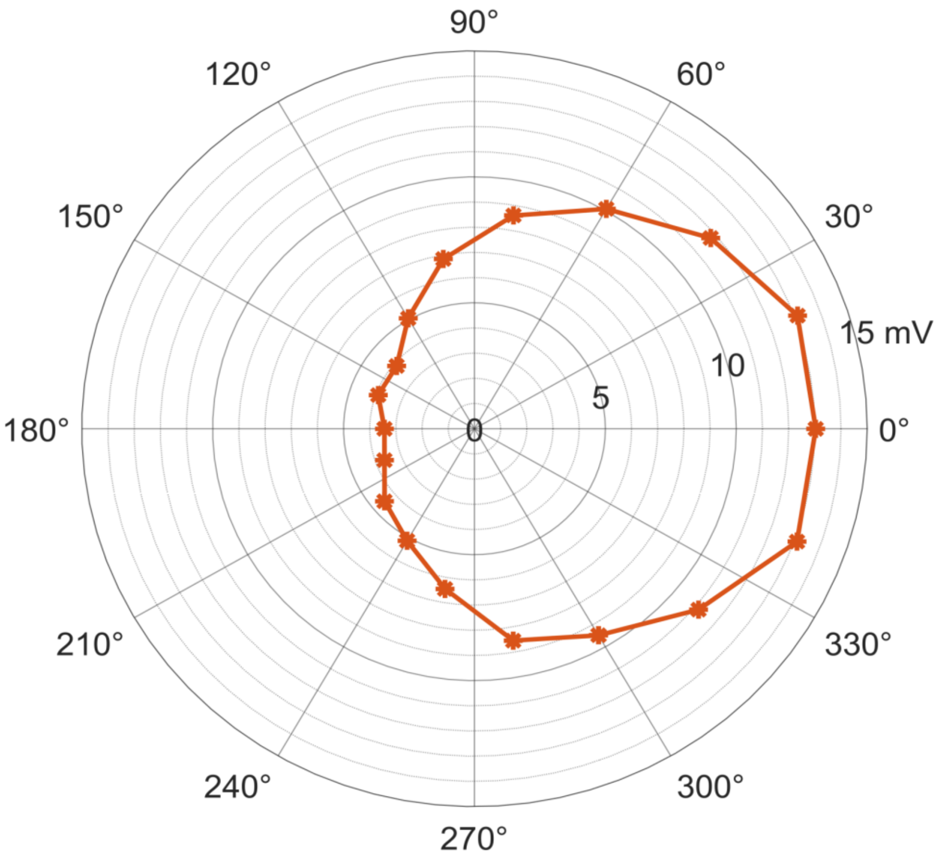
Experimental Demonstration using analog circuits

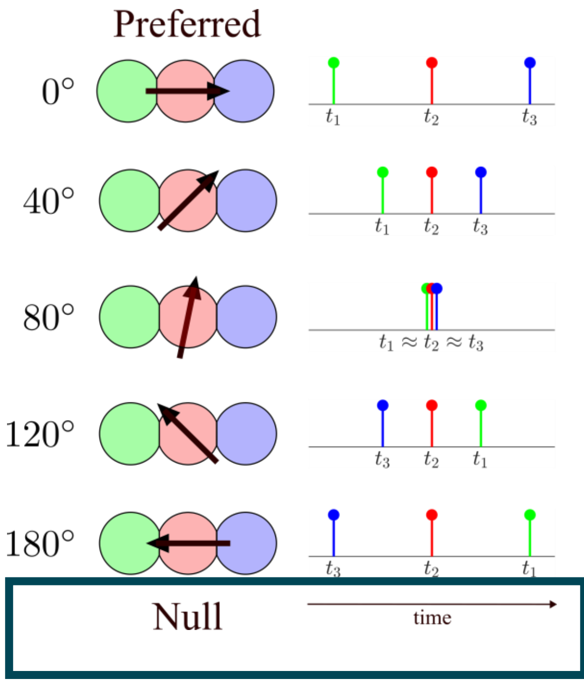


FPAA 2.9V

Dendrite Circuit



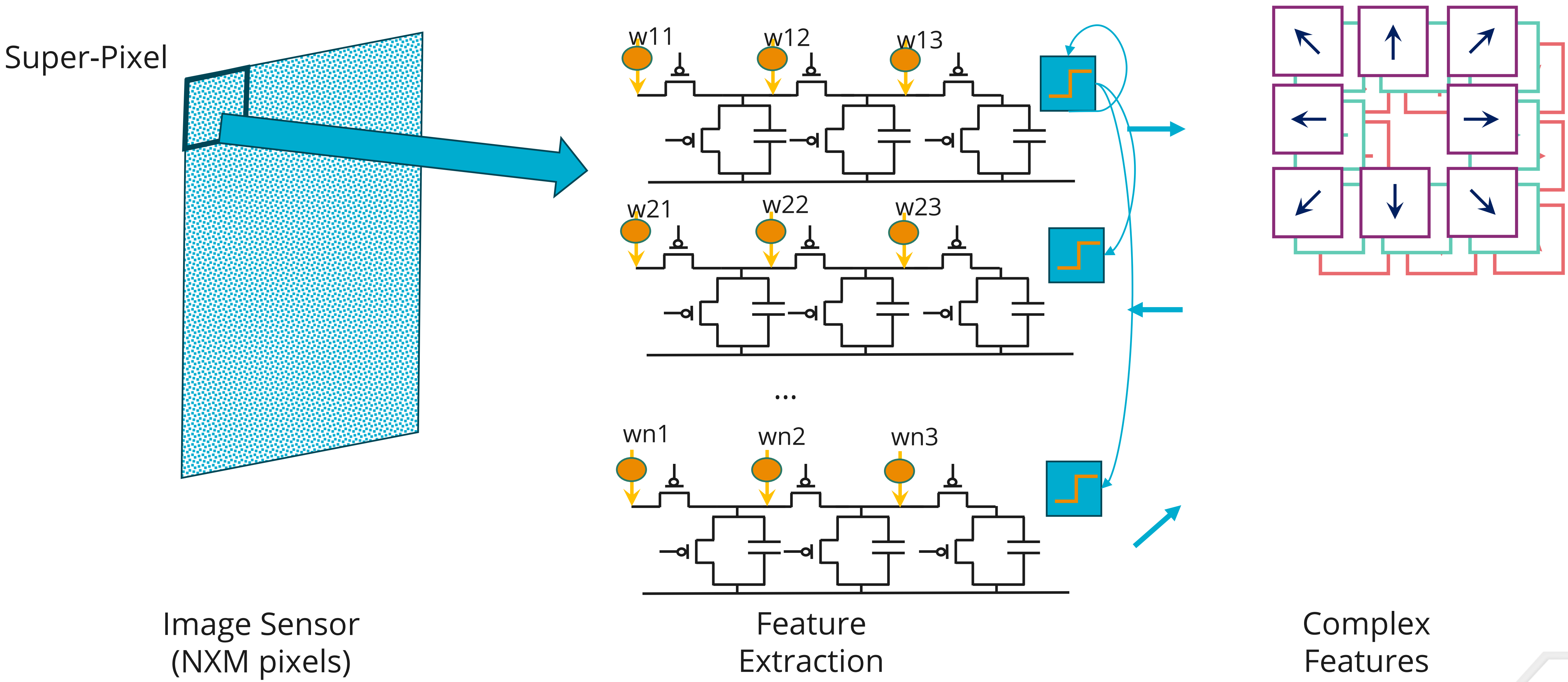




Parker et. al IEEE ICONS 2024

DENDRITES FOR DIRECTION SELECTIVITY

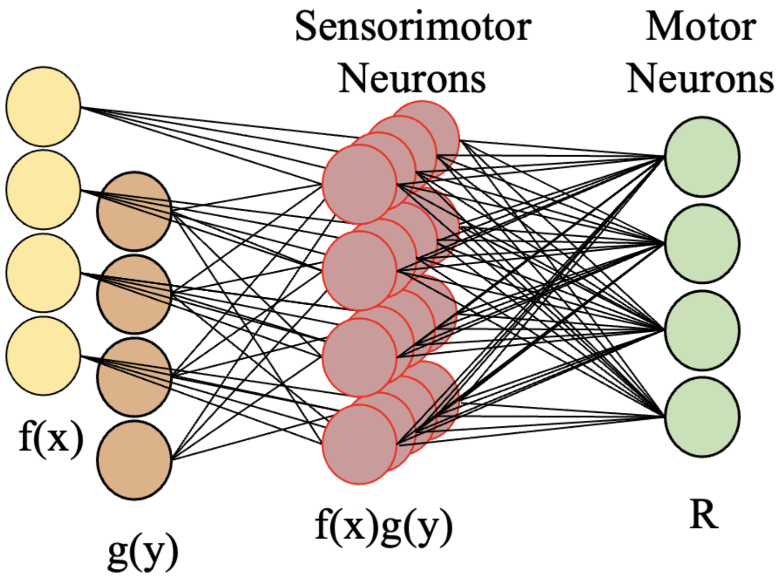
Dendrites for pattern and velocity estimation



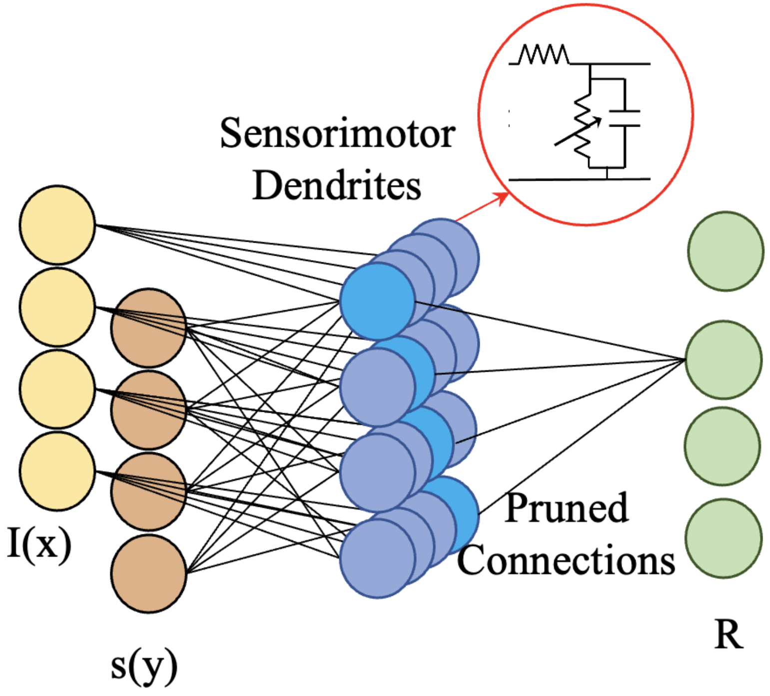
DRAGONFLY WITH DENDRITES

Using shunting inhibition for gain modulation

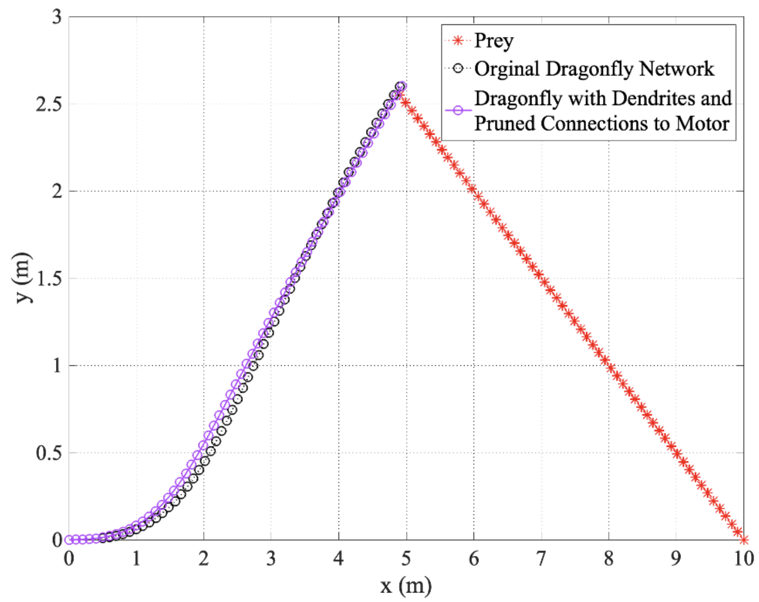
Cardwell et al. *In Review*



Original Dragonfly Interception Circuit (Chance 2020)



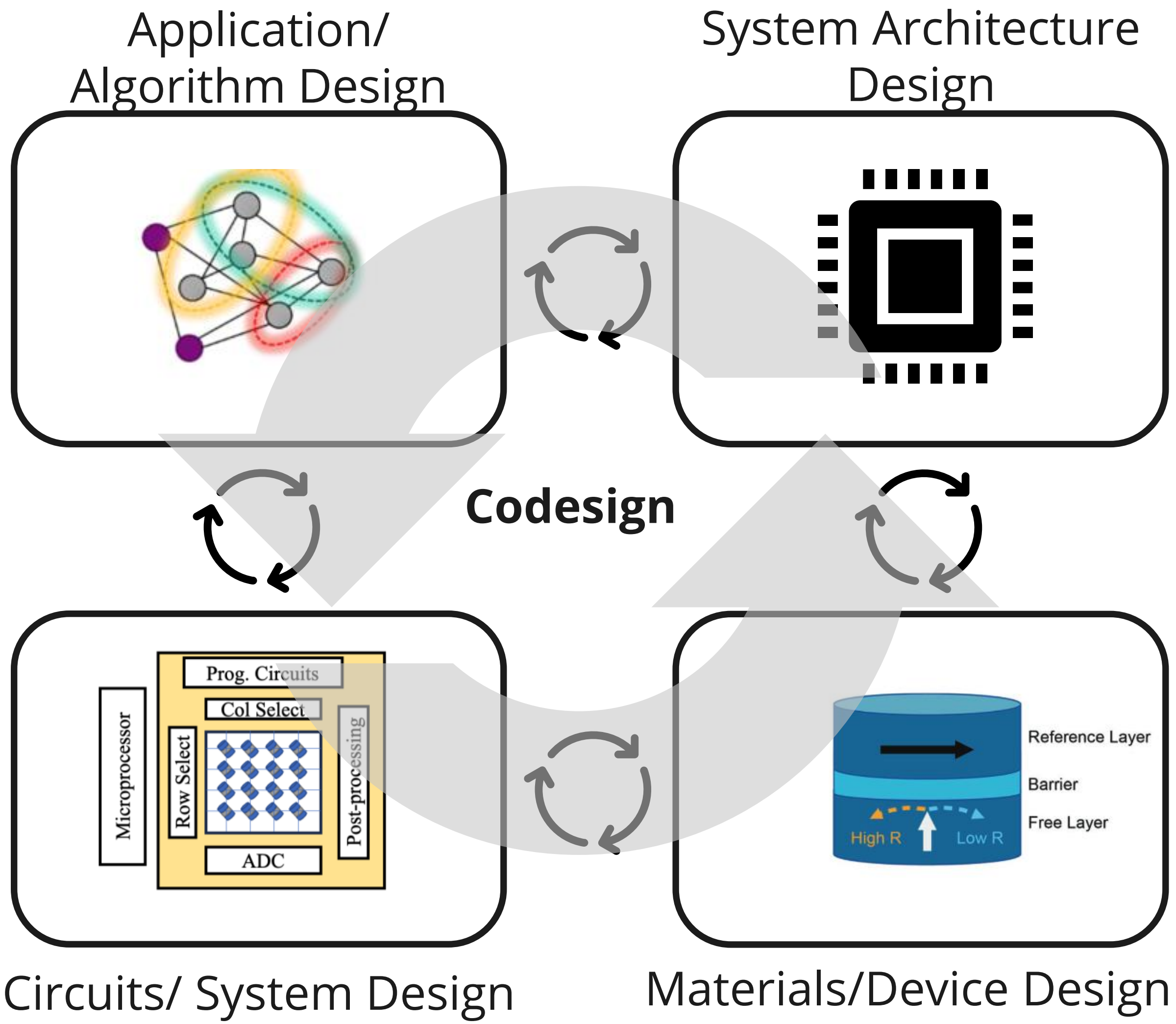
Dragonfly NN with Dendrites



Comparison with original interception circuit

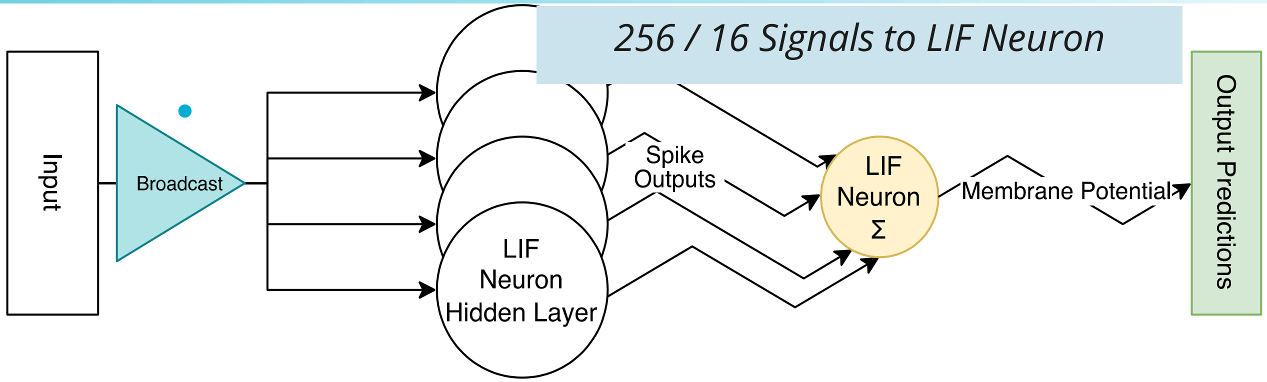
CO-DESIGN IS CHALLENGING

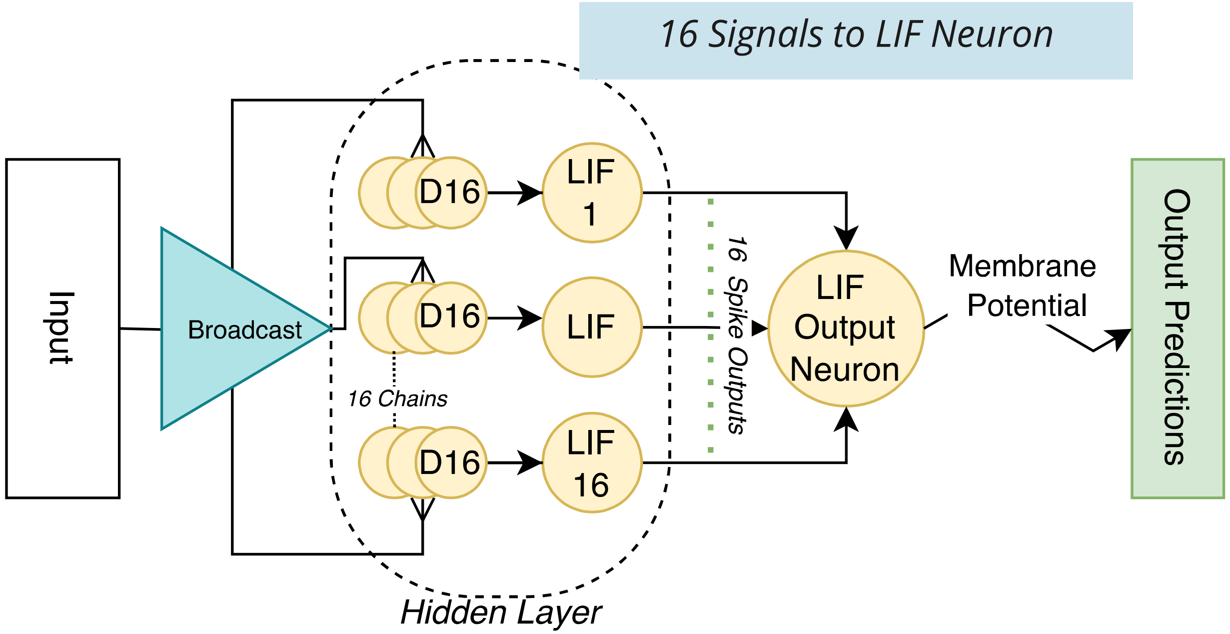
Software tools are critical for design and co-optimization

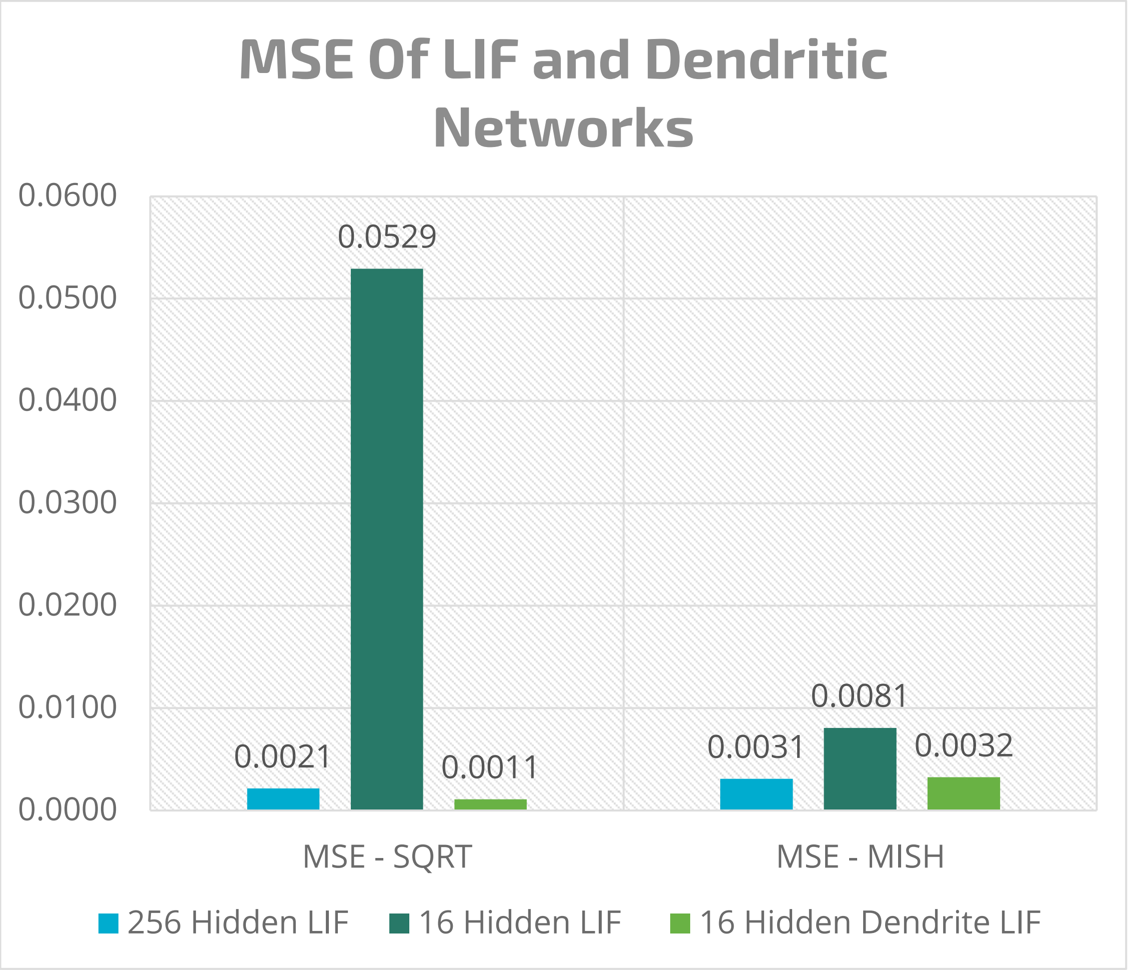


DENDRITES IN SPIKING NEURAL NETWORKS

Comparing networks with LIF and LIF +Dendrites



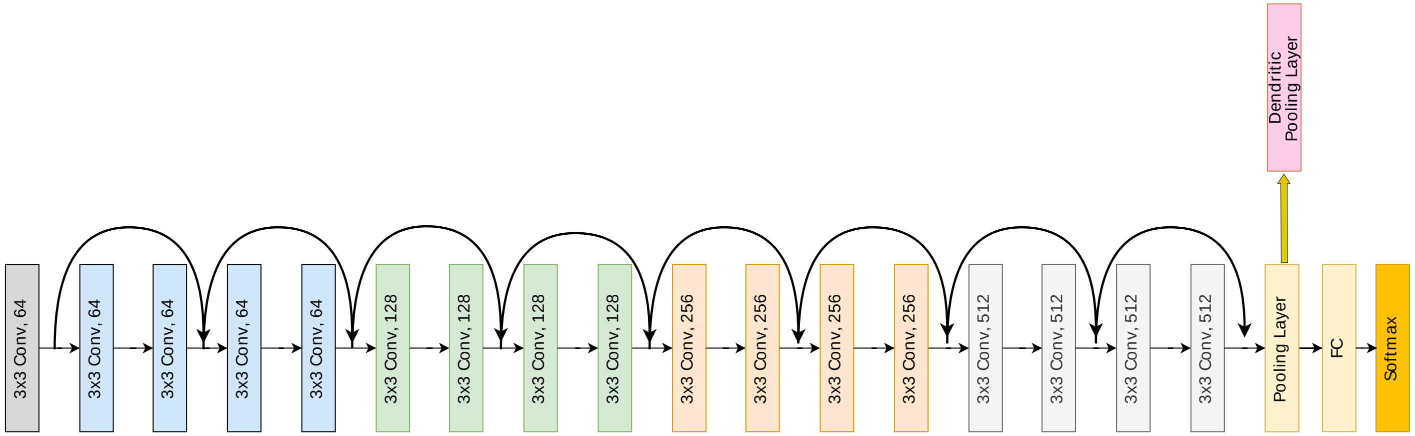




DENDRITE ENABLED SPIKING NEURAL NETWORKS

Leveraging inherent properties of dendrites

- Implemented a “Dendrite Pooling Layer” for use in ANN
- Trained ResNet18 on CIFAR-10 for 300 epochs
  - ResNet + Dendrite layer took significantly longer to train
  - Simplified ODE layer adds state and loops
- Found accuracy to be comparable
- Dendritic pooling has potential in ANNs



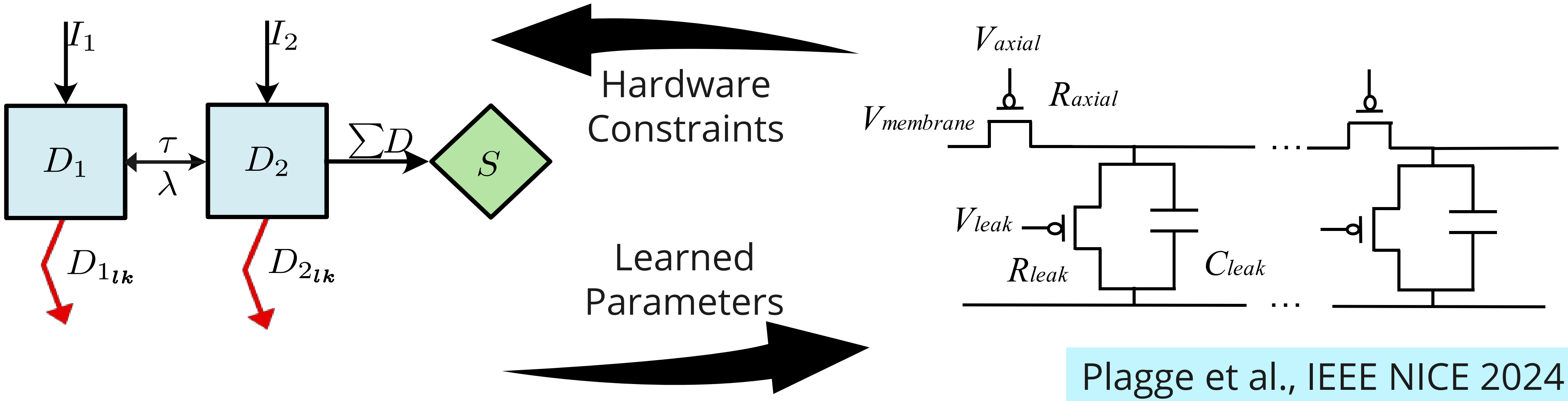
Pooling Layer Hardware	Energy
Digital Nvidia Jetson	504.41 uJ
Dendritic pooling	0.265 uJ

Plagge et al., IEEE ICRC 2024

DENDRITE ENABLED SPIKING NEURAL NETWORKS

- Implemented Torch library with a dendritic chain
  - Simplified version of the complex ODE dendrite solution.
  - Wrapped dynamics into a set of constants and parameters.
- Dendrites support SNNTorch & Non-Spiking Torch.
- Provides a trainable 1-D chain of dendrites

Value	Type
Lambda	”**Spatial**” constant: Represents Distance
Tau	“**Temporal**” constant: Capacitance and Resistance
Leak	Signal loss for each tap
Input Weight	Increases or Decreases signal strength



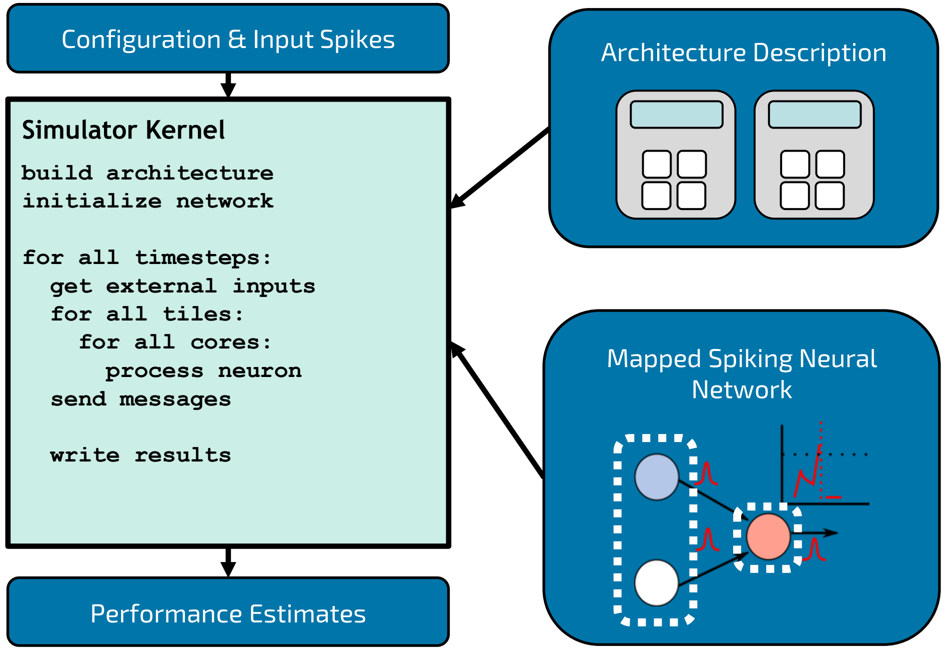
Plagge et al., IEEE NICE 2024

CHALLENGE: NOVEL CODESIGN TOOLS

SANA-FE: Specialized tool to explore novel neuromorphic architectures

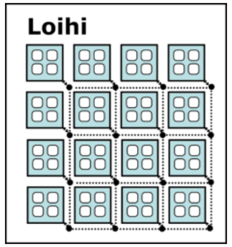
- Rapidly estimate performance of neuromorphic architectures for design-space exploration
- General & extensible spiking H/W simulator
- Model functional behavior & track performance
- Schedule messages & intra-core interactions
- Calibrate simulator to real-world systems
- Accurately predicts latency & energy of gesture categorization spiking neural network (SNN)

SANA-FE: Simulating Advanced Neuromorphic Architectures for Fast Exploration

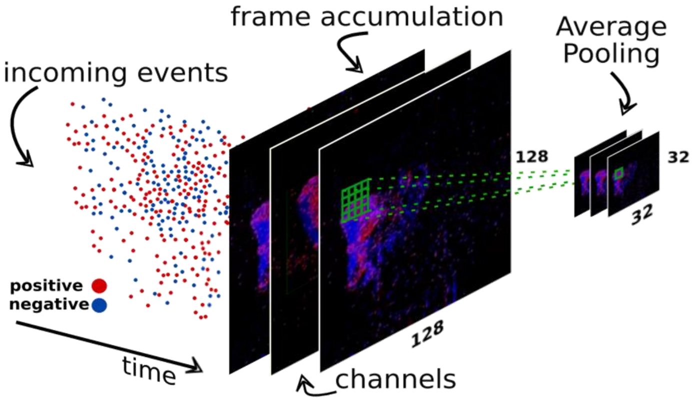


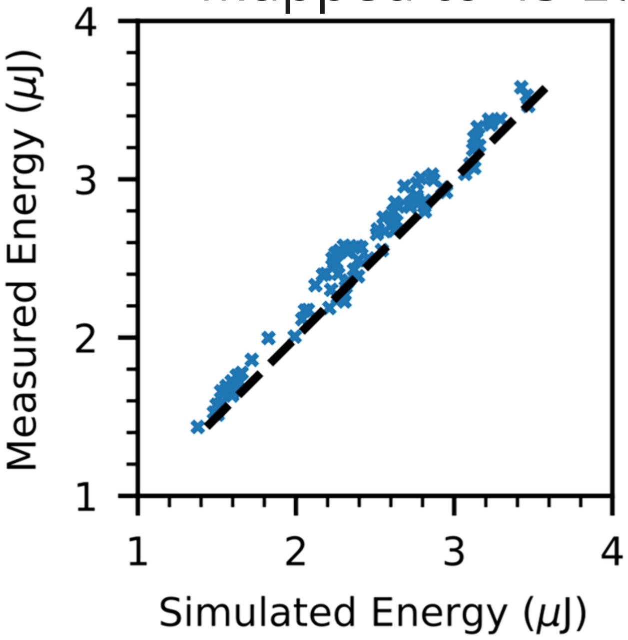
J. Boyle et al. ICONS 2023

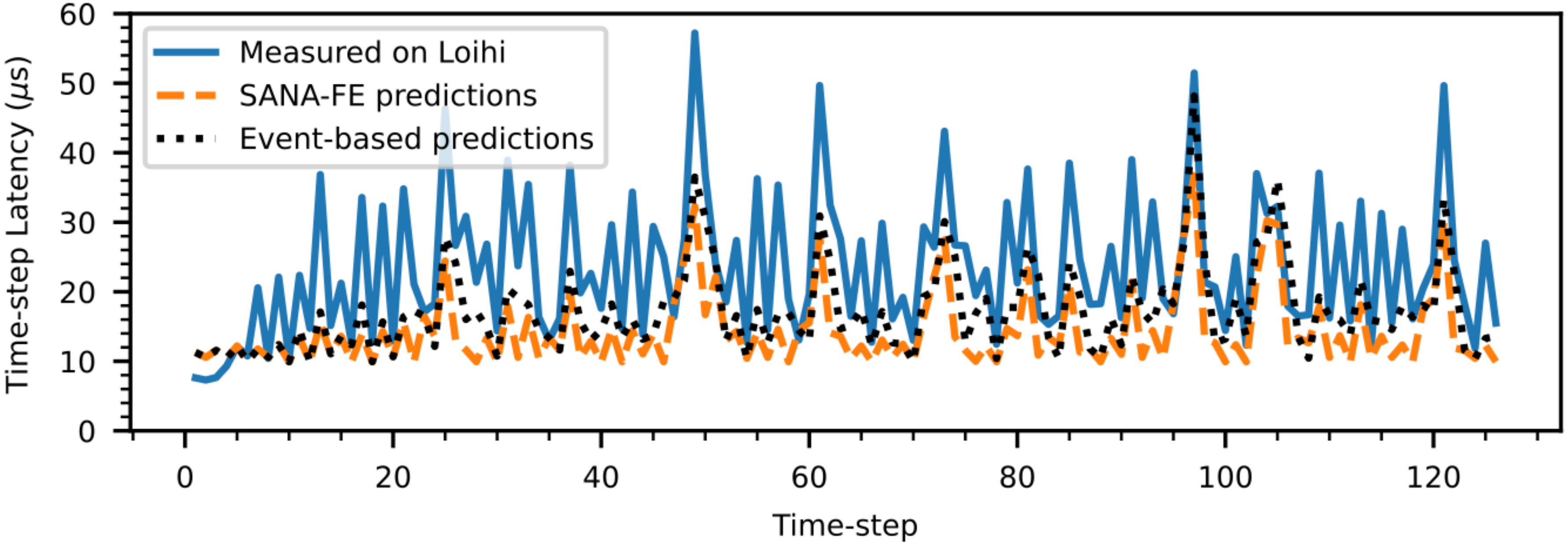
DVS GESTURE RECOGNITION APPLICATION



- Predict energy & performance for larger real-world neuromorphic applications
  - SNN trained on DVS gesture data-set [Massa'20]
  - 18,678 neurons across 6 layers
  - Mapped to 45 Loihi cores out of 128







DVS GESTURE DESIGN SPACE EXPLORATION

- Design-space exploration using DVS gesture application
  - Loihi-based designs, traded-off core count (c) vs neurons per core (n)
  - Optimum design had 170 cores, 21% faster than **Loihi (128 cores)**
  - Design-space sweep took 29 s

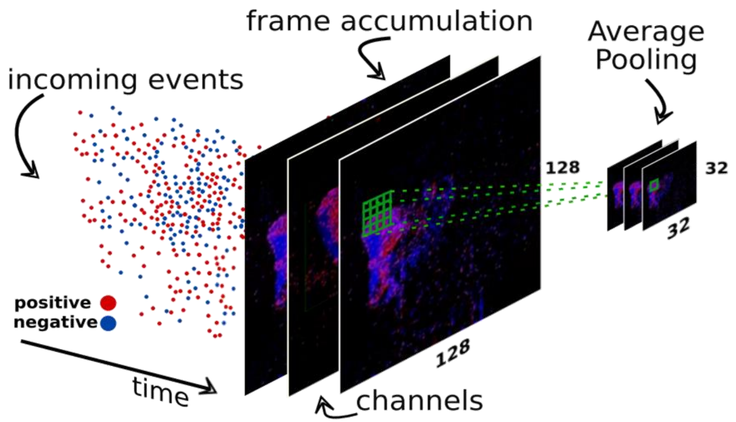
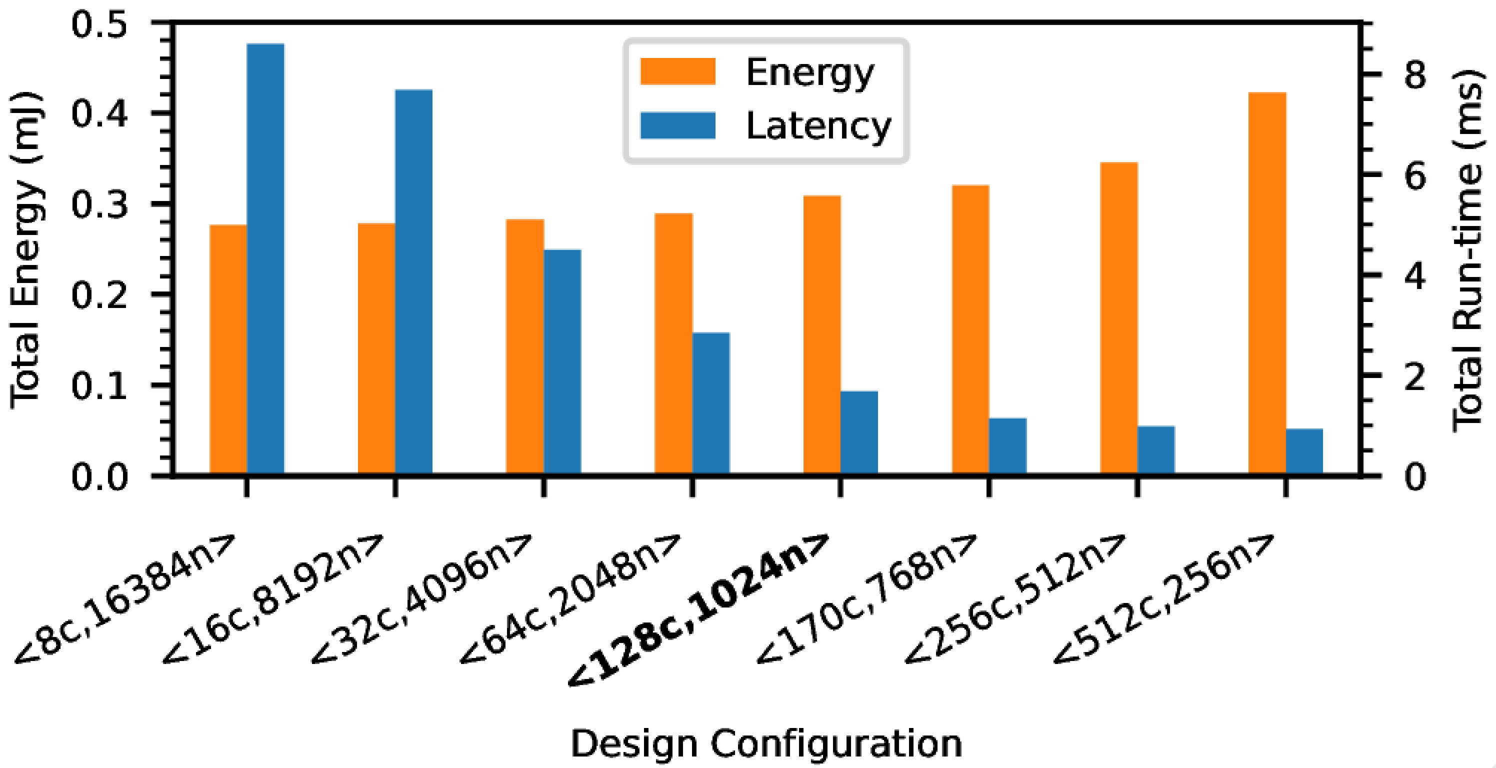
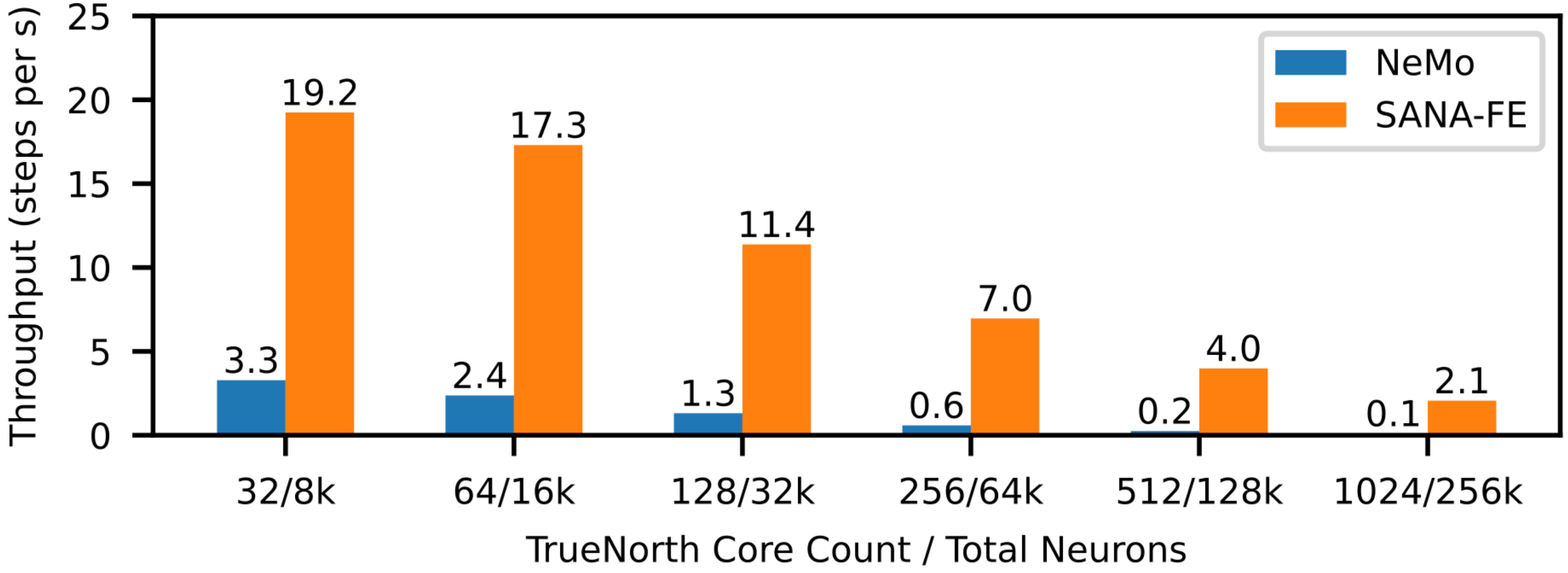


Image reproduced from [Massa,'20]



SIMULATOR SPEED RESULTS

- **Compared to existing discrete-event based spiking simulator (NeMo)**



  - Simulating IBM TrueNorth architecture
  - Randomized SNN with 80% of spikes intra-core, 20% spikes between cores

➢ **Over 20x faster than NeMo for 1024 cores**

DENDRITES IN NEUROMORPHIC ARCHITECTURE

- Further develop links with Dendrite-SNN hardware simulations – SanaFe
- Work on a spiking self-attention network with dendrites:
  - Dendritic attention layer (Temporal coherence and context)
  - Dendritic pooling layers (More efficient summary layer)
- Other compelling network designs
- Release as stand-alone library or as SNNTorch add-on

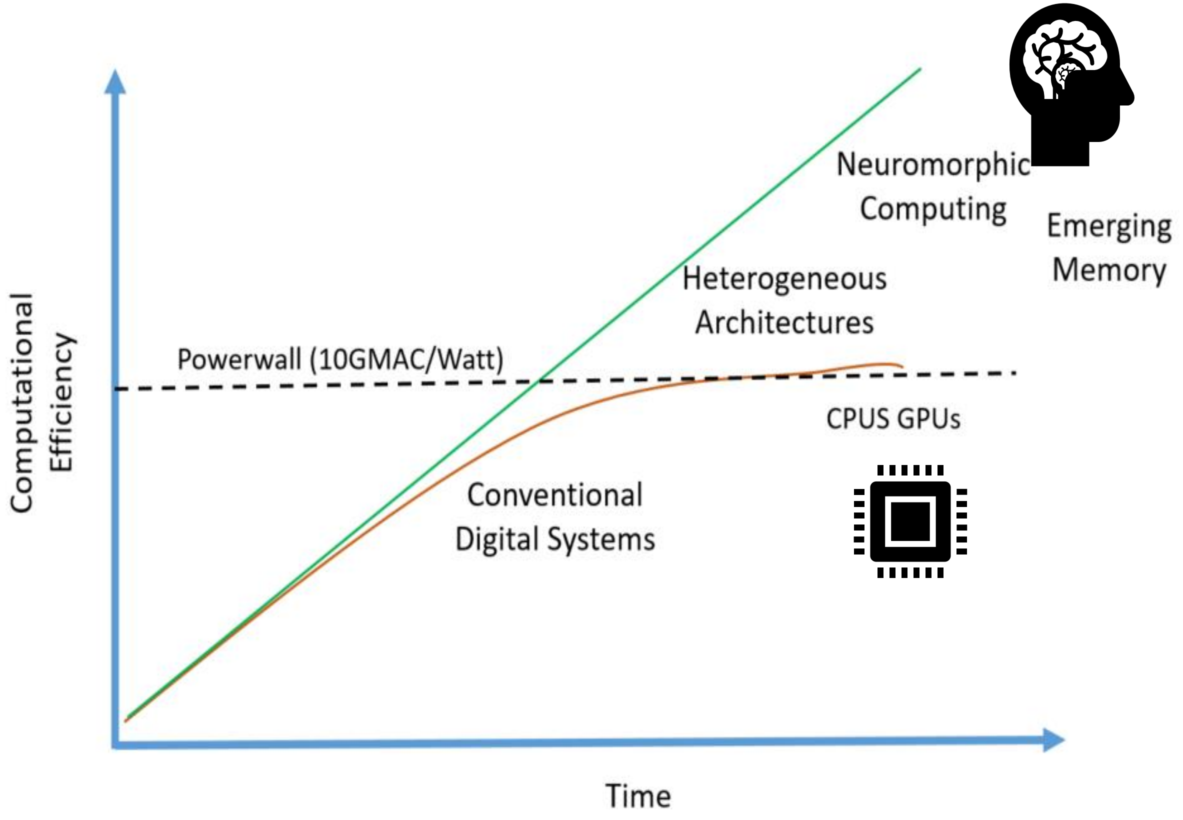
SanaFe – Hardware Simulator

An in-progress tool to estimate timing and energy of neuromorphic systems.
Currently supports Loihi.
Analog components are WIP

HETEROGENEOUS ARCHITECTURES

Embracing Heterogeneity

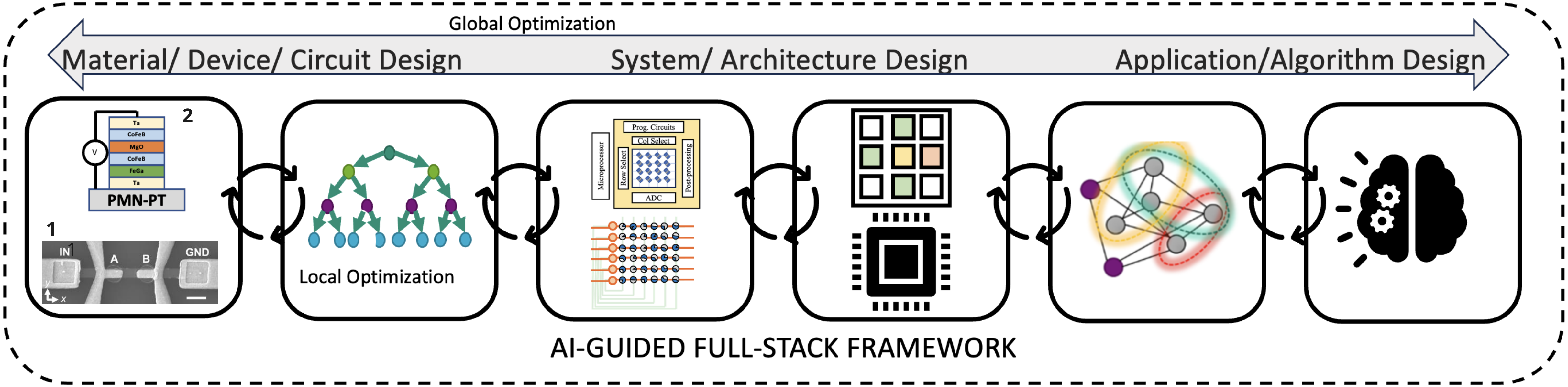
- Move past general-purpose solutions and only use them for prototyping
- Heterogeneous and reconfigurable systems
- Get the best performance based on system needs.
- System designed with application in mind for optimization
- Near-sensor processing



'Truly Heterogeneous Computing', Cardwell et al., SMC 2020

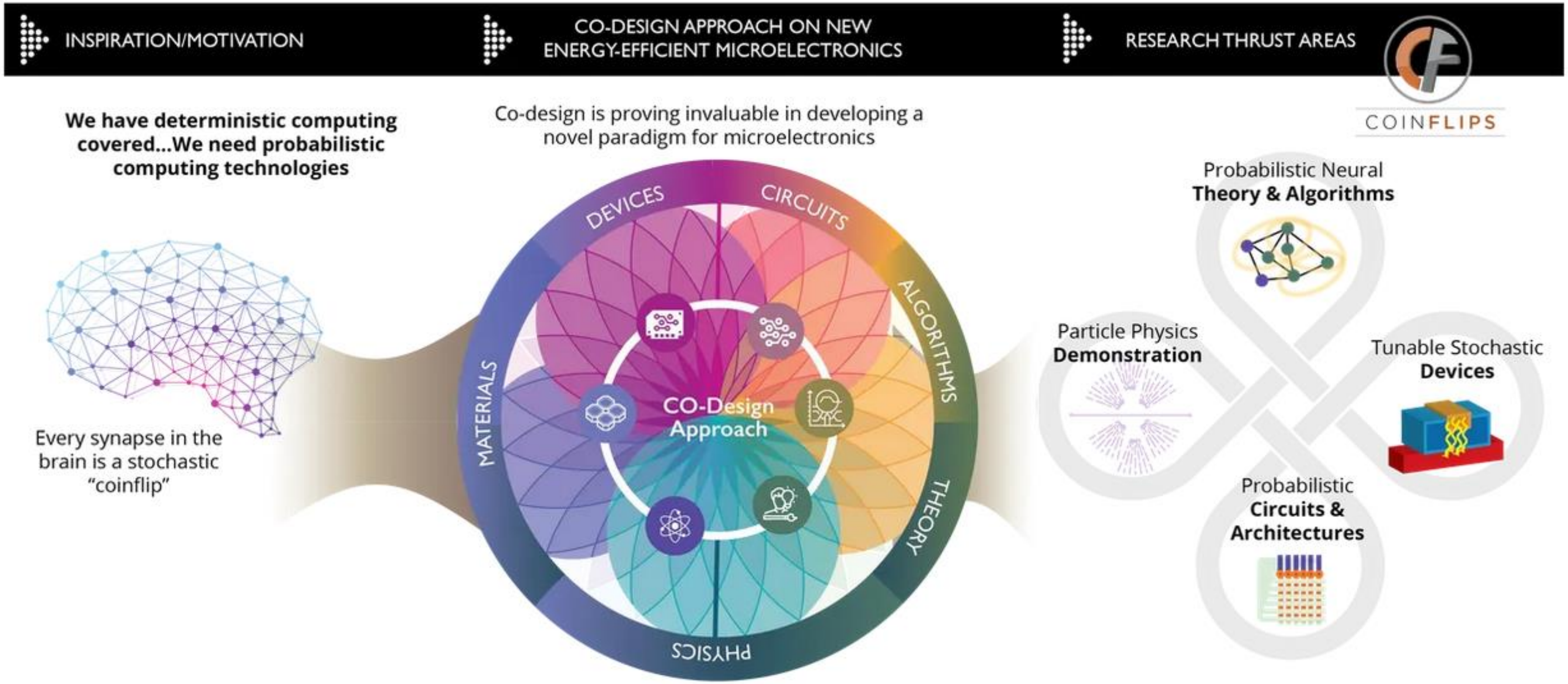
APPLICATION SPECIFIC SOLUTIONS

We are in the "golden-age of computer architectures" - Patterson



1. Cui et al. arXiv 2024
2. Karki et al. Journal of Materials Research 2024

PROBABILISTIC NEURAL SYSTEMS: COINFLIPS



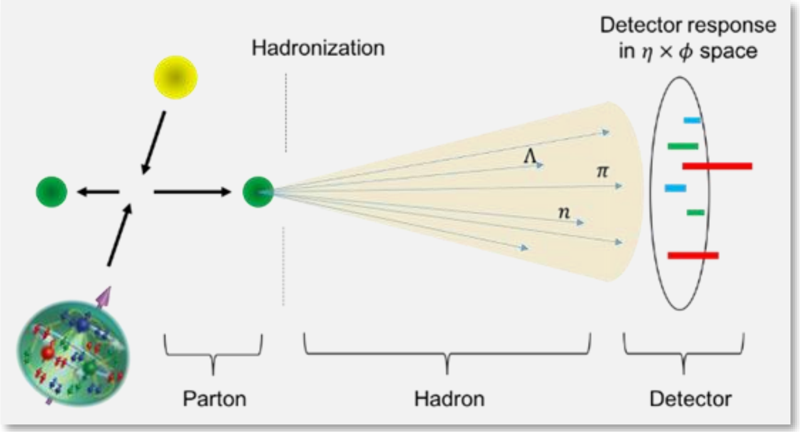
OBJECTIVE: Leverage stochasticity in computing by exploiting the underlying physics of emerging random number generator (RNG) devices to build probabilistic neural architectures.

Collaborators: NYU, ORNL, Temple University, UT-Austin and UT-Knoxville, USA

TRUE RANDOM NUMBER GENERATION (TRNG)

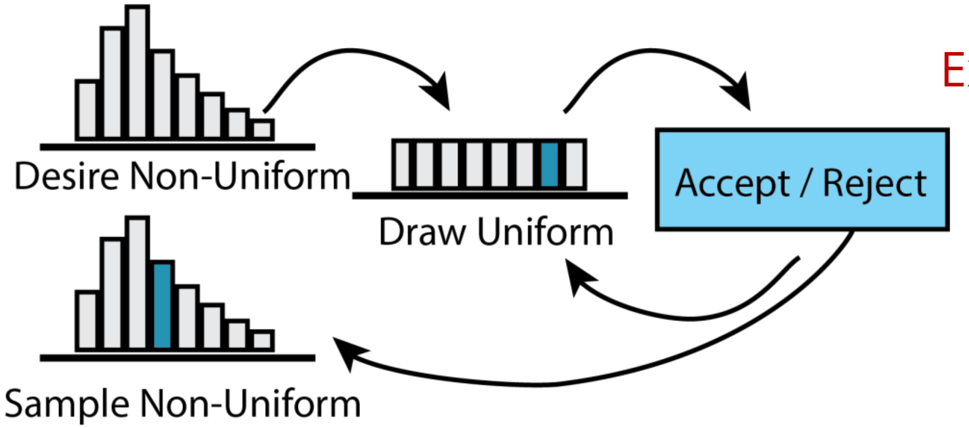
APPLICATION: HIGH ENERGY PHYSICS

- Current PRNG Methods take 40—50% of CPU compute time.
- TRNGs leveraging stochastic devices can lead to significant energy and latency savings.



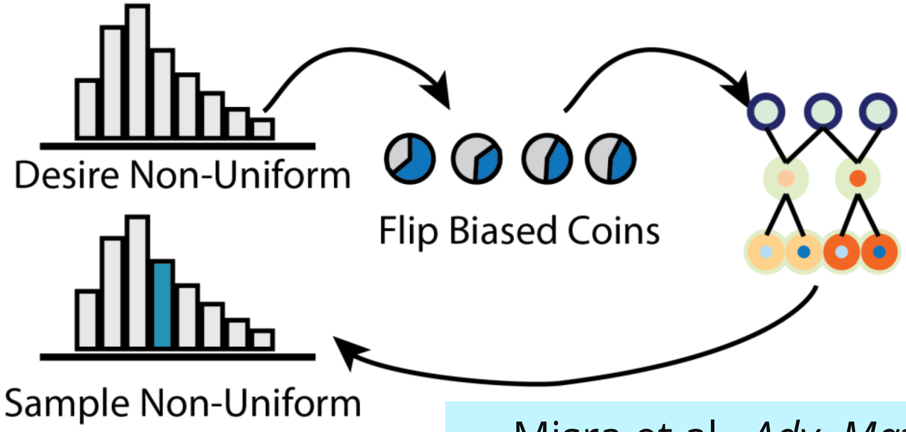
Pierog et al., Phy Rev. 2022

RNG TODAY



Expensive

COINFLIPS APPROACH



Efficient

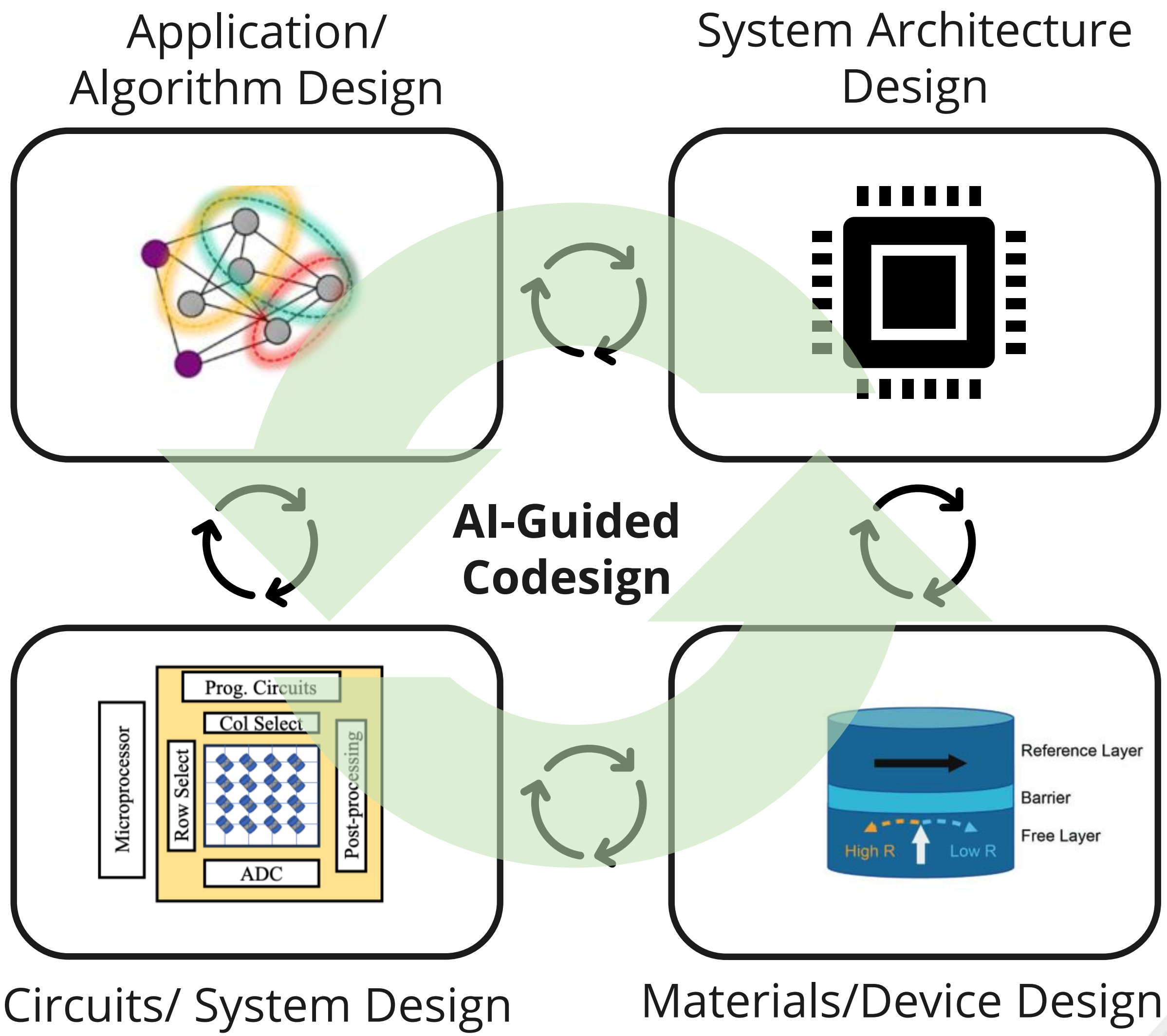
Misra et al., *Adv. Materials 2022*

AI TO ACCELERATE CODESIGN FOR EMERGING COMPUTING

AI-Guided Tools a force multiplier for large design space

- Computing demands are constantly increasing.
- Emerging computing techniques can alleviate these challenges.
- However, the design space is huge and optimization is needed across the stack.
- AI-guided techniques can alleviate these challenges.

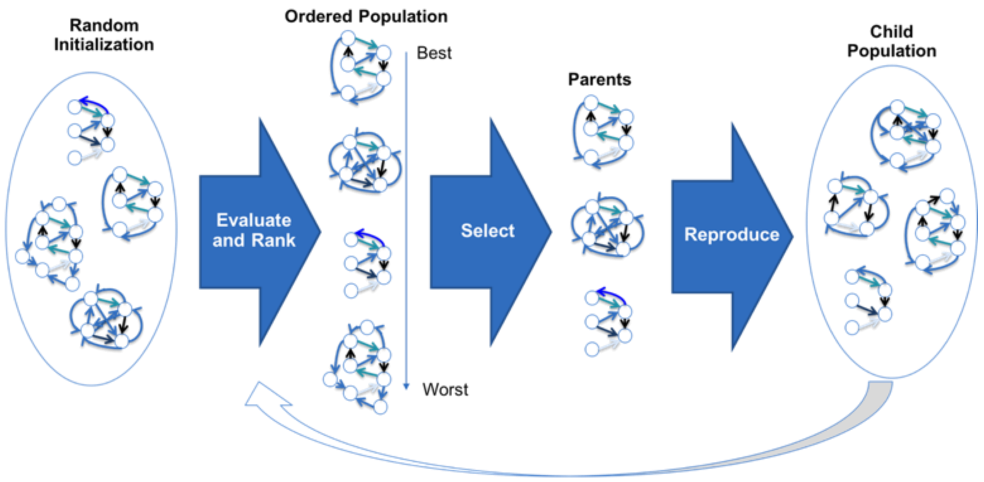
Misra et al., *Adv. Materials 2022*



AI-GUIDED METHODS FOR EMERGING COMPUTING

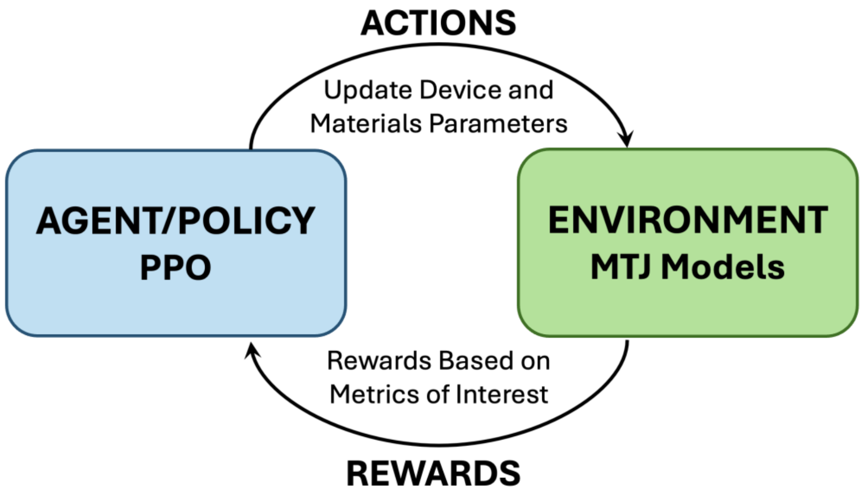
EVOLUTIONARY OPTIMIZATION

- Evolutionary algorithms
  - LEAP: Library of Evolutionary Algorithms in Python
  - EONS: Evolutionary Optimization for Neuromorphic Systems



REINFORCEMENT LEARNING

- Trains agent to make optimal decisions in an environment to maximize rewards.
  - Agent trained on PPO policy
  - Environment: Physics-based Device model



Other AI Approaches: Physics-aware machine learning, Generative AI, Neural ODEs

EA FOR CIRCUITS: MULTI-OBJECTIVE OPTIMIZATION

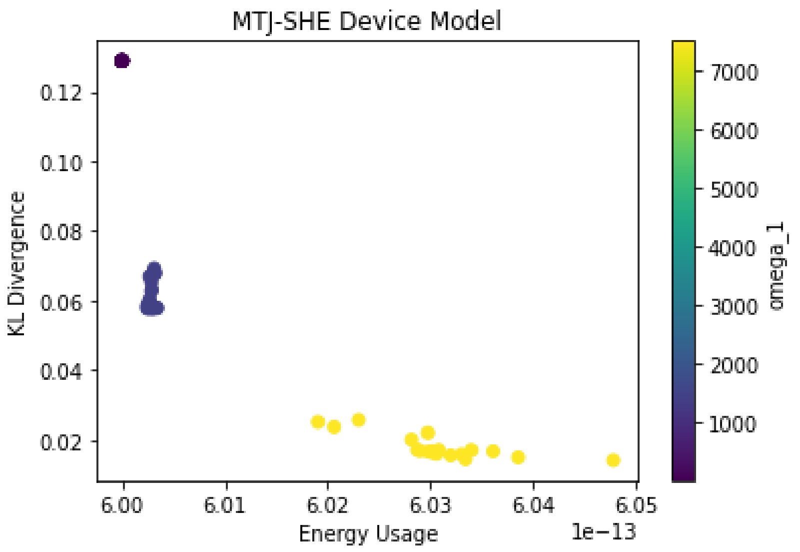
FITNESS FUNCTION

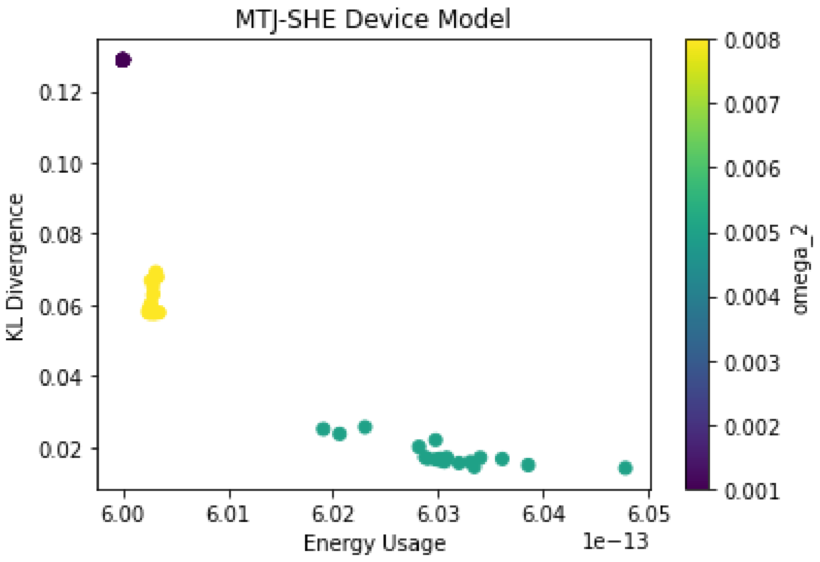
$$f(w, p1, p2, q1, q2) \quad = \boldsymbol{\omega_1} KL(p_1, p_2, q_1, q_2) + \boldsymbol{\omega_2}\left(\sum_{i=1}^{2} |p_i \ - 0.5| + \sum_{i=1}^{2} |q_i \ - 0.5|\right) + \boldsymbol{\omega_3} EN\ (p_1, p_2, q_1, q_2)$$

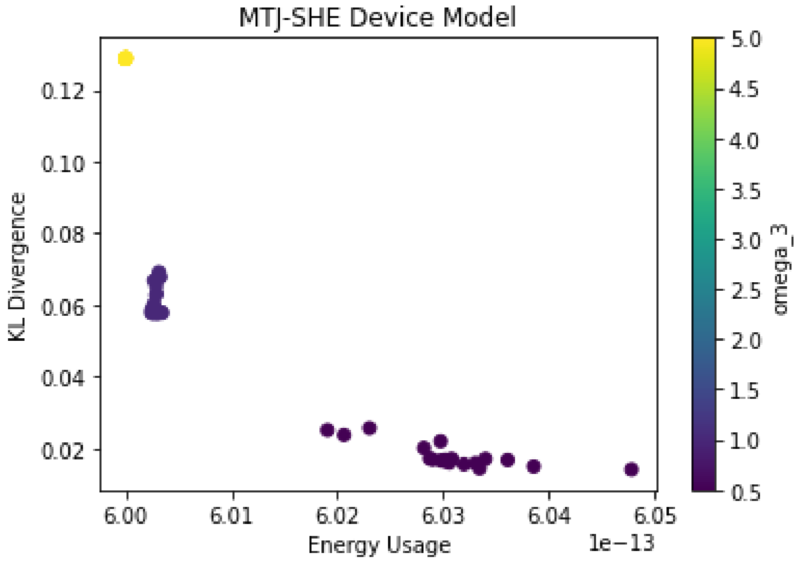
Kullback-Leibler Divergence

Difference of weight from a fair coin

Energy of a coinflip







SHE: Spin Hall Effect

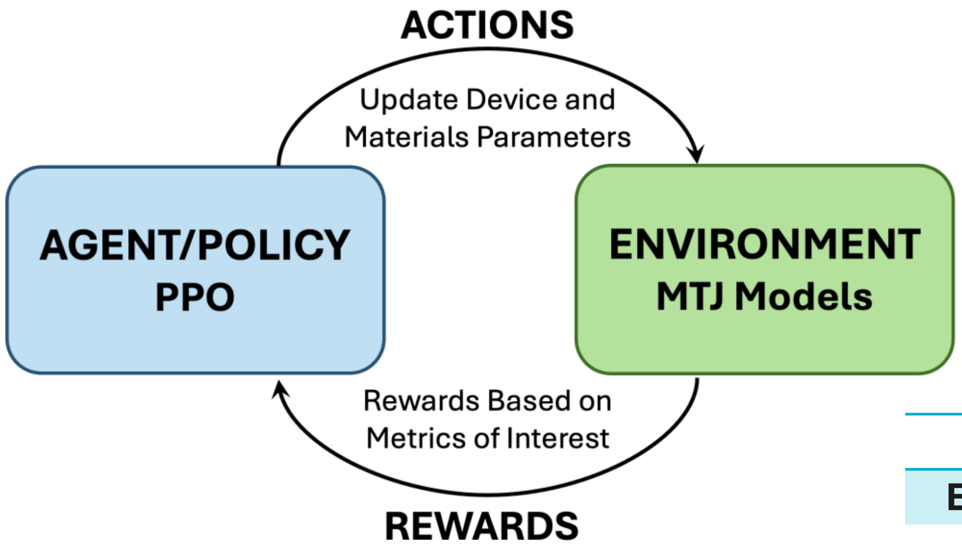
Multi-objective optimization of weights ω1, ω2, ω3 for optimal KL divergence and energy usage of MTJ-SHE devices

Cardwell et al., IEEE ICRC 2022

REINFORCEMENT LEARNING FOR DEVICE CODESIGN

Discover best device and material characteristics for TRNG

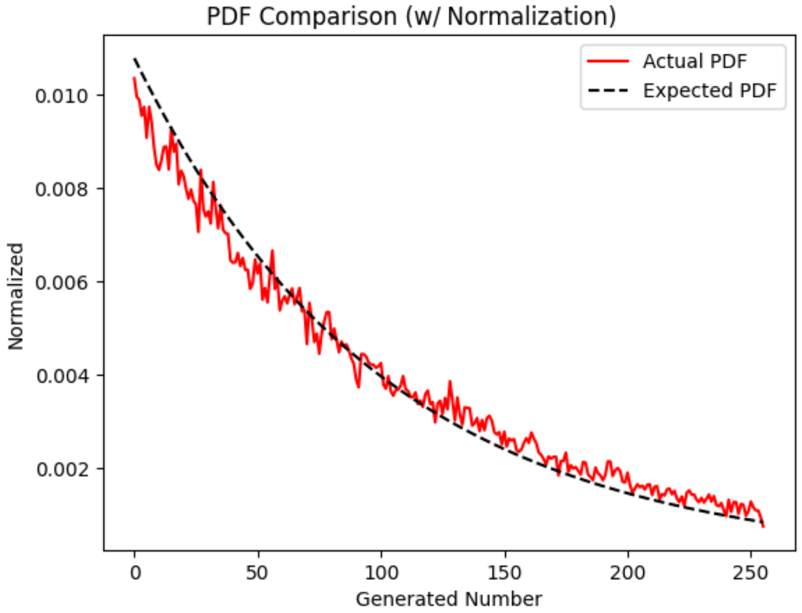
Device Check



Metrics: Energy, KL Divergence

- RL agent trained using PPO policy to find best device and material configuration.
- Device validity checked.
- Best configuration found normalizing for energy and KL-divergence.

Metric	Best Config
Energy (J)	2.568×10^{-14}
KL-Divergence	0.0532

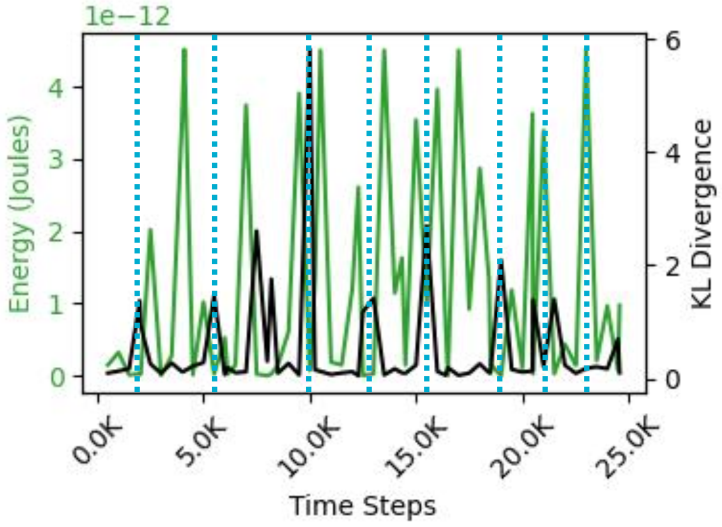


Cardwell et al., *IEEE ISCAS 2024*

DEVICE PERFORMANCE TRADEOFFS

TRAINING RL AGENT

- Agent had to balance both energy and KL-divergence for optimization which seemed to have an inverse relationship.
- Reward schema is extremely important.

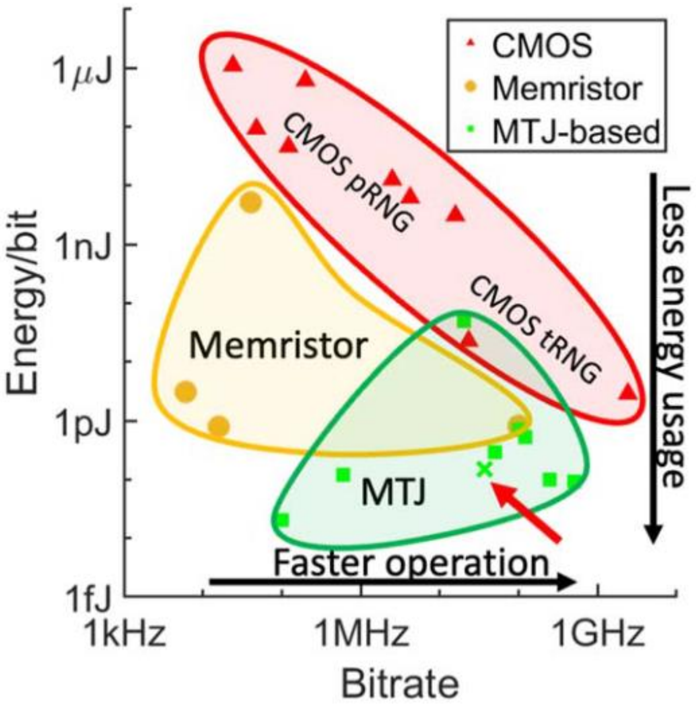


Energy and KL-Divergence

Cardwell et al., *IEEE ISCAS 2024*

BENCHMARKING TRNGS

- Comparing CMOS pRNG, tRNG, memristor tRNG and MTJ tRNG

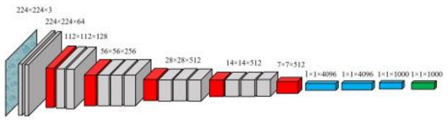


Maicke et al., *IOP Nano 2024*

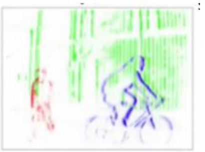
NEXT-GENERATION NEUROMORPHIC ARCHITECTURES

Re-think how we design computer architectures

- We need more dynamics and complexity per computational unit.
- Leverage Stochasticity as a feature not a bug.
- We need systems that do not just process, but can learn, adapt and reconfigure.
- Novel integration: 3D architectures, wafer scale etc. for scaling and dense connectivity
- Many areas where neuromorphic can have impact from HPC to edge.
- Codesign tools can accelerate design space exploration and lead to creative solutions.



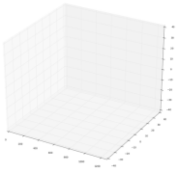
AI/ML
(ANN, SNN)



Edge
Computing

Brain-Inspired
Algorithms

High Performance
Computing



Scientific
Computing

Probabilistic
Computing

Sandia National Laboratories
NEUROMORPHIC

THANK YOU!

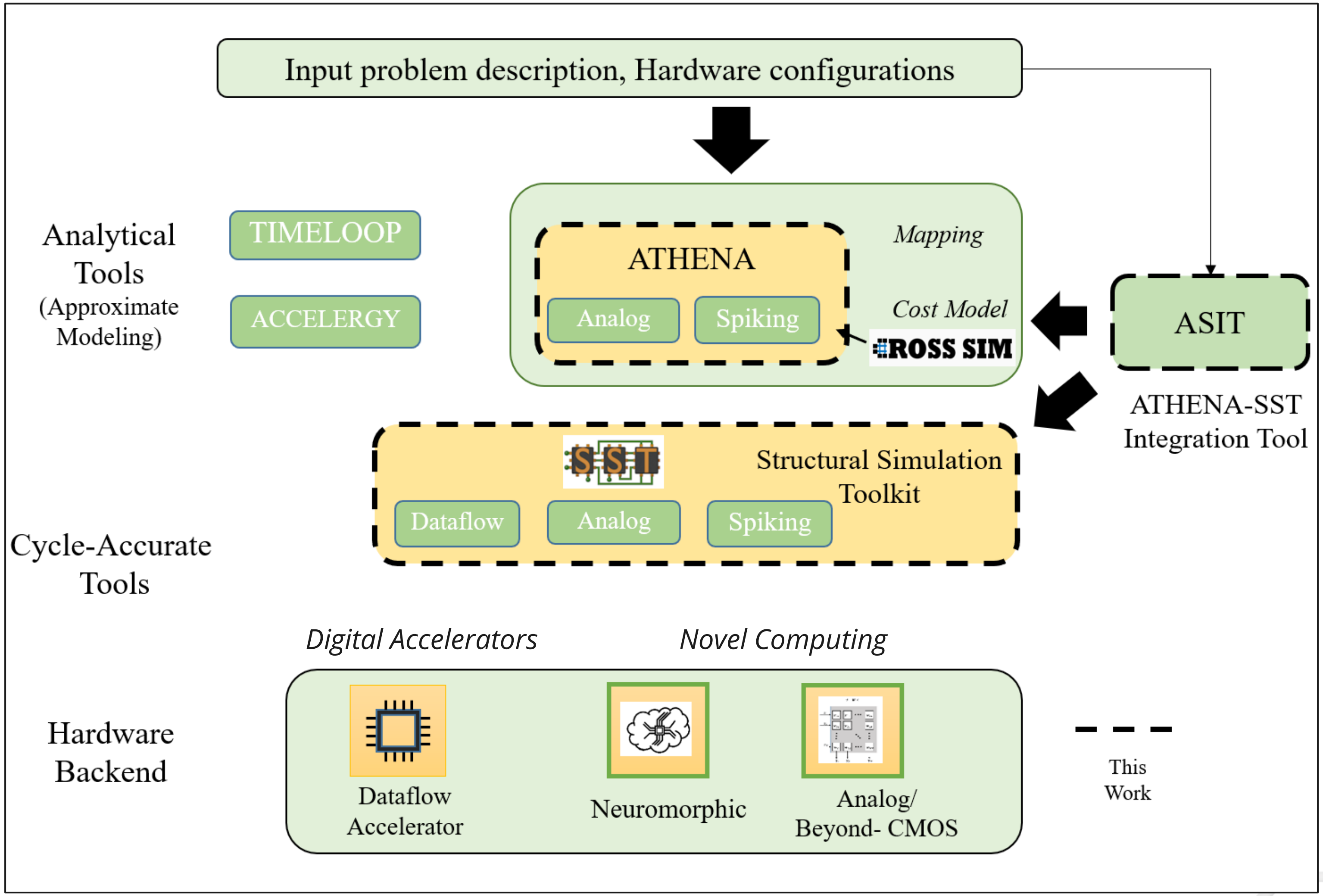
SGCARDW@SANDIA.GOV

BACKUPS

ATHENA
(ANALYTICAL TOOL TO EVALUATE HETEROGENEOUS NEUROMORPHIC ARCHITECTURES)

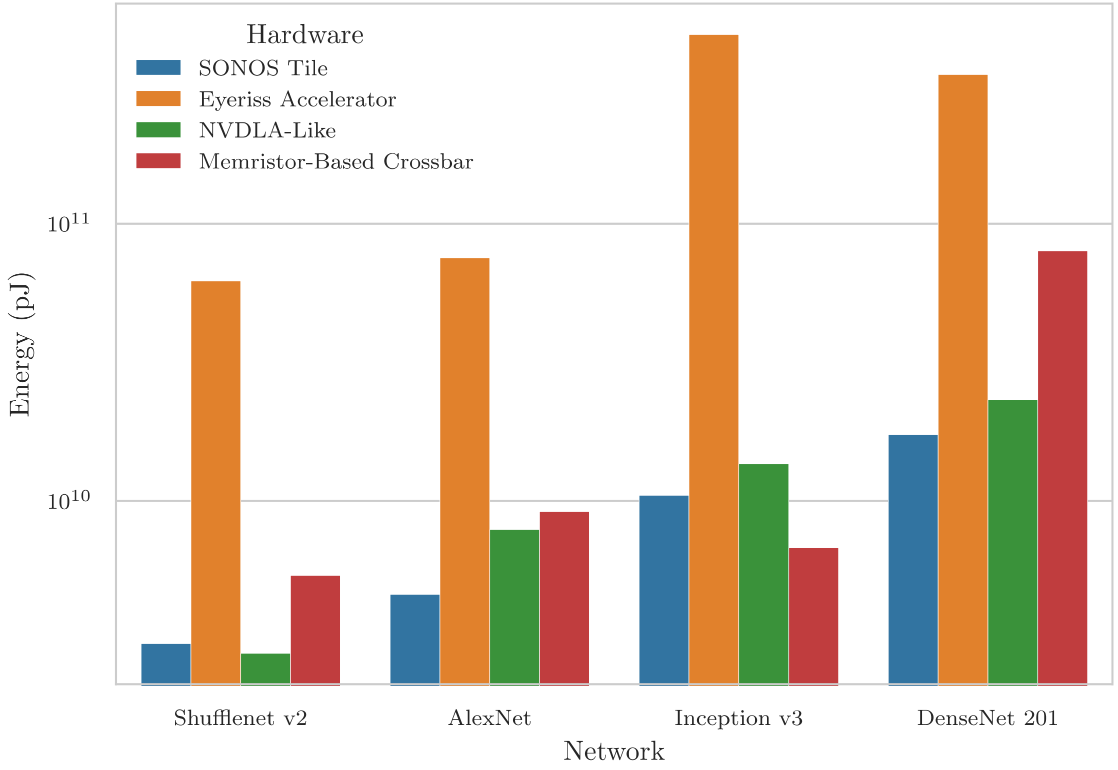
- ATHENA will quickly evaluate performance metrics of analog architectures

- Developed as part of a larger ecosystem
  - Tools to enable next-generation hardware design prototyping



Plagge et al., International Conference on Rebooting Computing (ICRC) 2022

ATHENA – HARDWARE PERFORMANCE

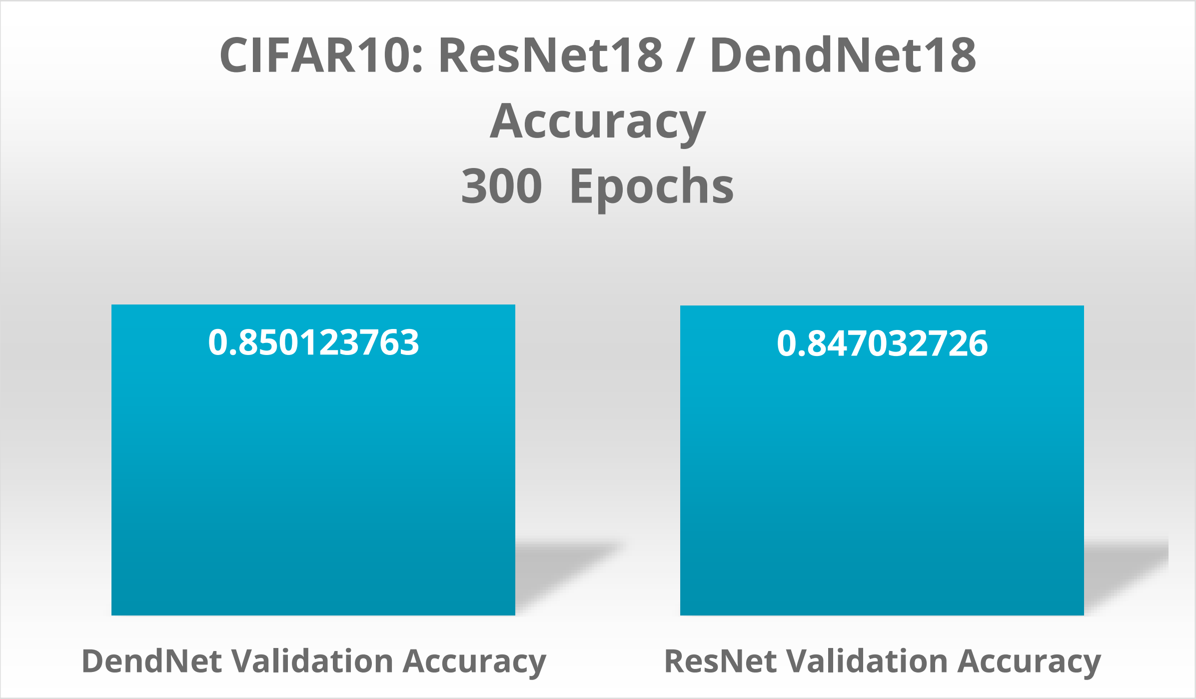


Plagge et al., International Conference on Rebooting Computing (ICRC) 2022

- ATHENA was used to compare the performance of multiple hardware devices against various deep learning networks
- The SONOS tile-based architecture performed well across networks, with one notable exception: the Inception v3 network
- This performance difference could be explored – showing ATHENA's potential for codesign work

ANN NETWORK APPLICATIONS – RESNET18

- Working with a graduate student intern at SNL
- Implemented a "Dendrite Pooling Layer" for use in AI ML
- Replaced traditional pooling layer with Dendrite Layer
- Trained ResNet18 on CIFAR-10 for 300 epochs
  - ResNet + Dendrite layer took significantly longer to train
  - Simplified ODE layer adds state and loops
- Found accuracy to be comparable
  - Dendritic pooling has potential in ANNs



Working with Priyam Mazumdar

ML NETWORK APPLICATIONS – RESNET18

- ResNet18 – Was slower to train with a dendritic layer
  - In hardware however, dendrites will be highly efficient

- Rough estimate of efficiency based on
  - Energy = $C(V_{mem} - Ek)V_{dd} = 500fJ$
    - C = 10pF
    - V_{dd}=2.5V
    - $V_{mem} - Ek$ = 100mV
  - Nvidia

Pooling Layer on Digital Nvidia Jetson	504.41 Micro Joules
Dendritic pooling	0.265 Micro Joules

Rodrigues, Crefeda Faviola, Graham Riley, and Mikel Luján.
"Energy predictive models for convolutional neural networks on mobile platforms."
arXiv preprint arXiv:2004.05137 (2020).