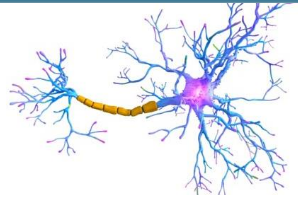
Modeling and Simulation Challenges of Neuromorphic Architectures



Suma George Cardwell

Cognitive and Emerging Computing

Sandia National Laboratories

ModSim 2023

Workshop on Modeling & Simulation of Systems and Applications

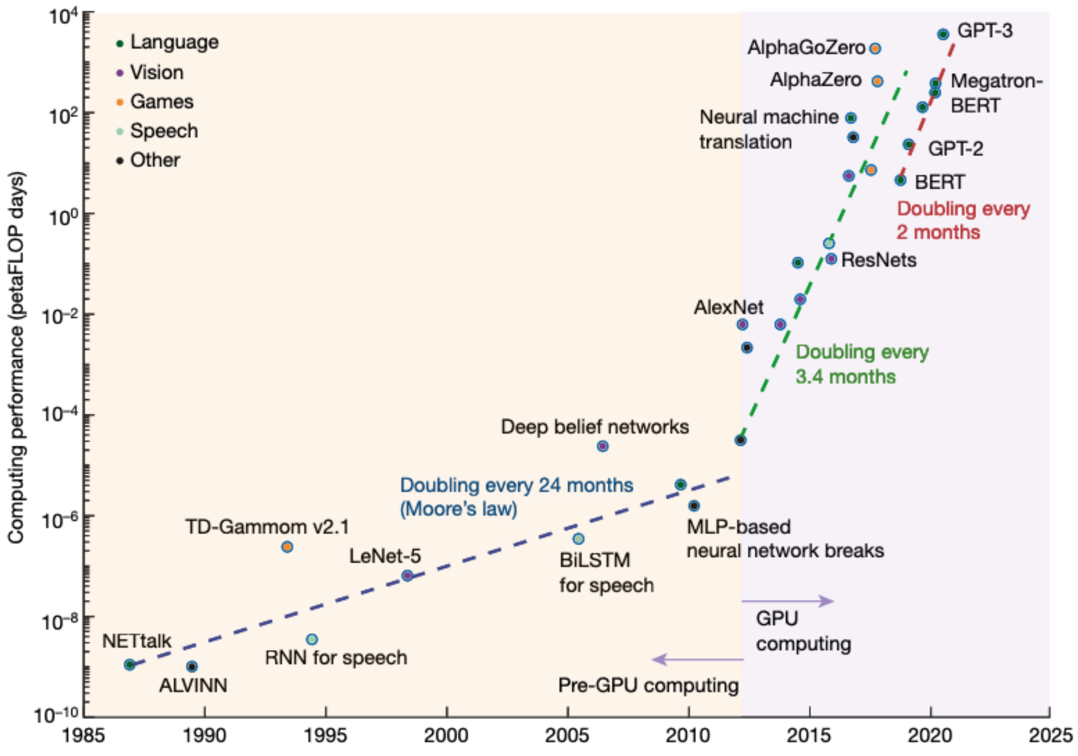
August 9th, 2023

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

SAND2023-07590O

FUNDAMENTAL CHALLENGES IN COMPUTING

- Limits of scaling have ushered in the "Golden Age of Computer Architecture" Hennessy & Patterson 2019
- Inefficiency of generality
- Performance saturation
- Growing demands for HPC and SWaP-constrained edge computing



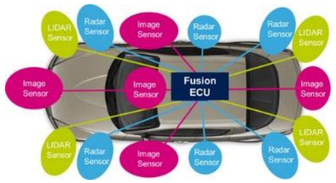
AI compute demands are increasing

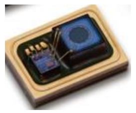
Mehonic & Kenyon 2022, Open AI Research Blog

Neuromorphic Computing gives a path forward for power efficiency scaling and meeting future computing needs.

COMPUTING LANDSCAPE

Sensors

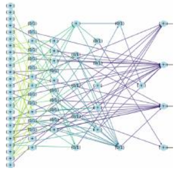




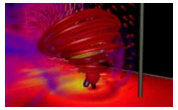

50 billion IoT devices by 2030

Algorithms

- Scientific Computation
- Machine Learning
- Brain-derived algorithms
- Signal Processing



Applications




Modern Computing

Conventional Digital

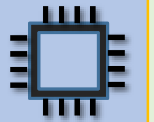
Novel Computing Paradigms

CPUs

GPUs

FPGAs

TPUs



ASICs

Neuromorphic

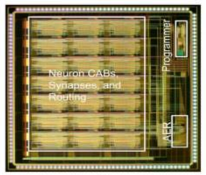
Quantum

Digital Neuromorphic

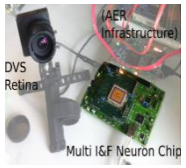
Intel Loihi
Davies 2018

SpiNNaker
Furber 2016

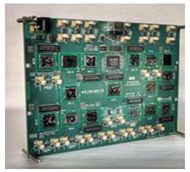
Analog/Mixed-signal Neuromorphic



GT Neuron
Brink 2013

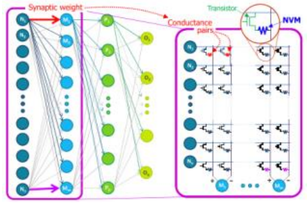


DAVIS 240C, DYNAPSEL

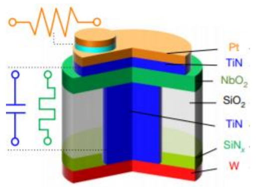


NeuroGrid
Benjamin 2014

Beyond CMOS devices



RERAM
Marinella et al., 2016

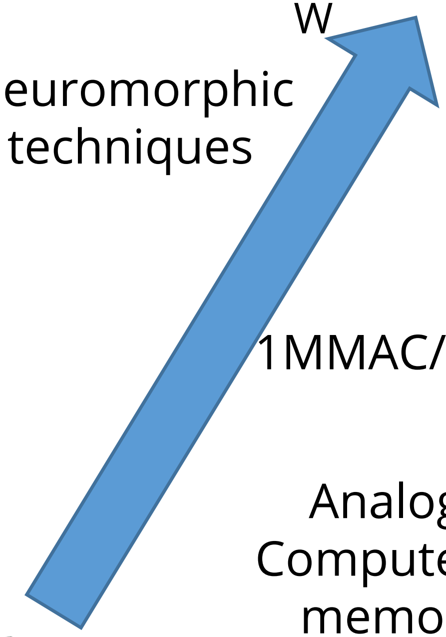


Mott- Memristor
Kumar et al., 2020

NEUROMORPHIC COMPUTING: INSPIRED BY THE BRAIN

Brain and Computing: Why make the connection ?

- High computational efficiency, Single neuron ~1MMAC/pW
- Processing and memory operations performed by the same components
- Self-organizing system
- Online learning
- Solving ill-structured problems
- Transfer learning
- Spiking/event driven communication, subthreshold computation



Hasler 2016

Neuromorphic techniques will be disruptive to how we develop our computing systems

MMAC: Million Multiply Accumulates

NEUROMORPHIC COMPUTING: DIVERSE SOLUTIONS

Digital Neuromorphic

Intel Loihi/ Loihi 2.0

SNL hosts Intel's 50 million neural supercomputer

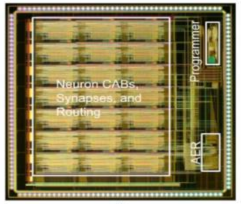
SpiNNaker/ SpiNNaker 2

→Scaled to a billion neurons

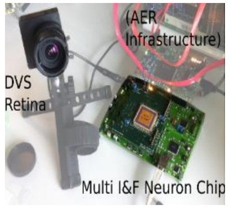
IBM TrueNorth

ODIN (Open-source)

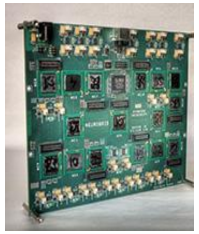
Analog/Mixed-Signal



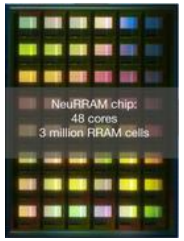
GT Neuron



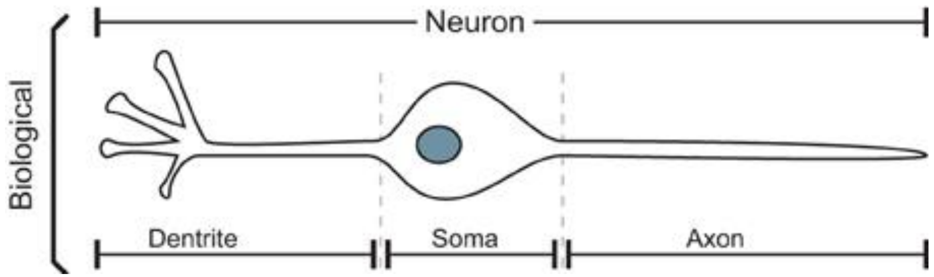
INI, ETH Zurich



Stanford Neurogrid

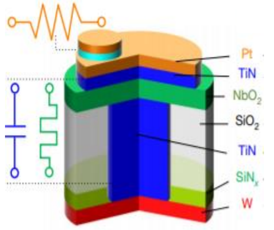


NeuRRAM UCSD/Tsinghua

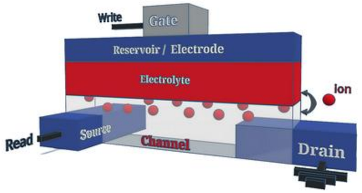


Brink et al., 2013

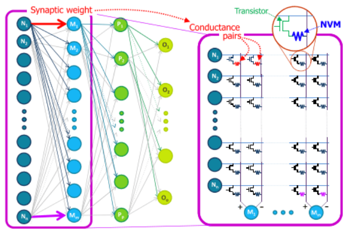
Beyond CMOS Devices



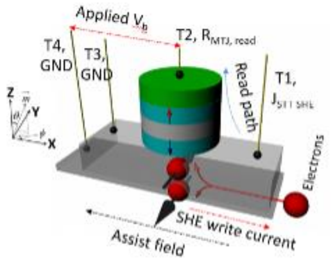
Mott- Memristor



ECRAM



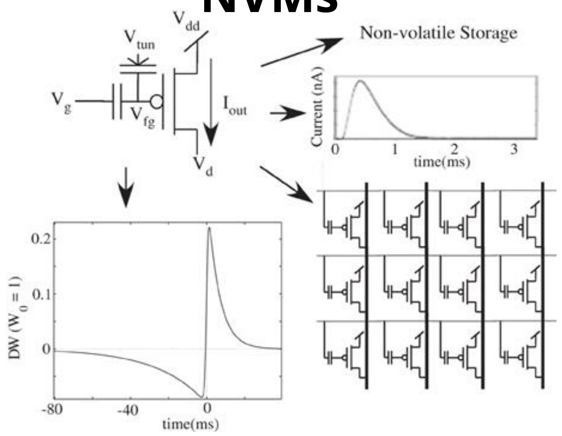
RRAM Crossbar



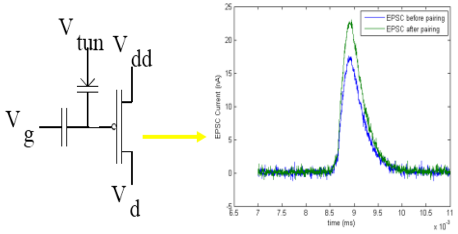
MTJ

NEUROMORPHIC BUILDING BLOCKS

Analog Crossbars using NVMs



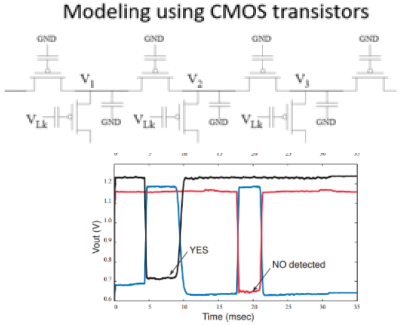
Learning Synapses



Many different models for neurons, synapses, online learning and dendrites.

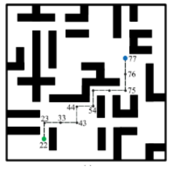
Neuromorphic offers computational richness we can leverage, to move beyond today's computational limitations.

Dendritic Processing



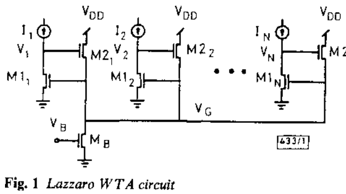
George Cardwell et al. 2013

Neural Path Planning



Koziol et al. 2013

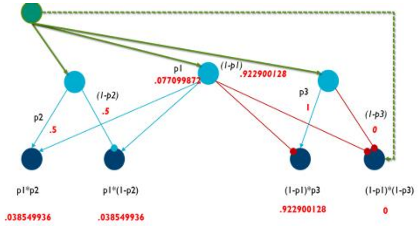
Winner-Take-All



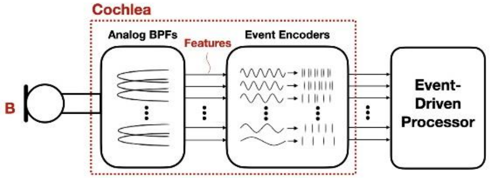
Lazzaro et al. 1988

Random Walks

Smith et al. 2021

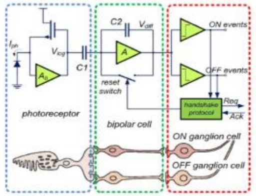


Silicon Cochlea



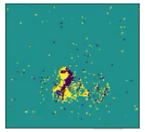
Liu et al. 2020

Silicon Retina/ Event Sensor

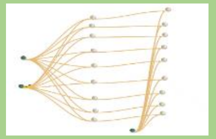


Posch et al. 2014

Delbruck et al. 2020

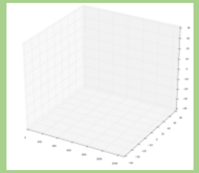


APPLICATIONS



AI/ML (ANN, SNN)

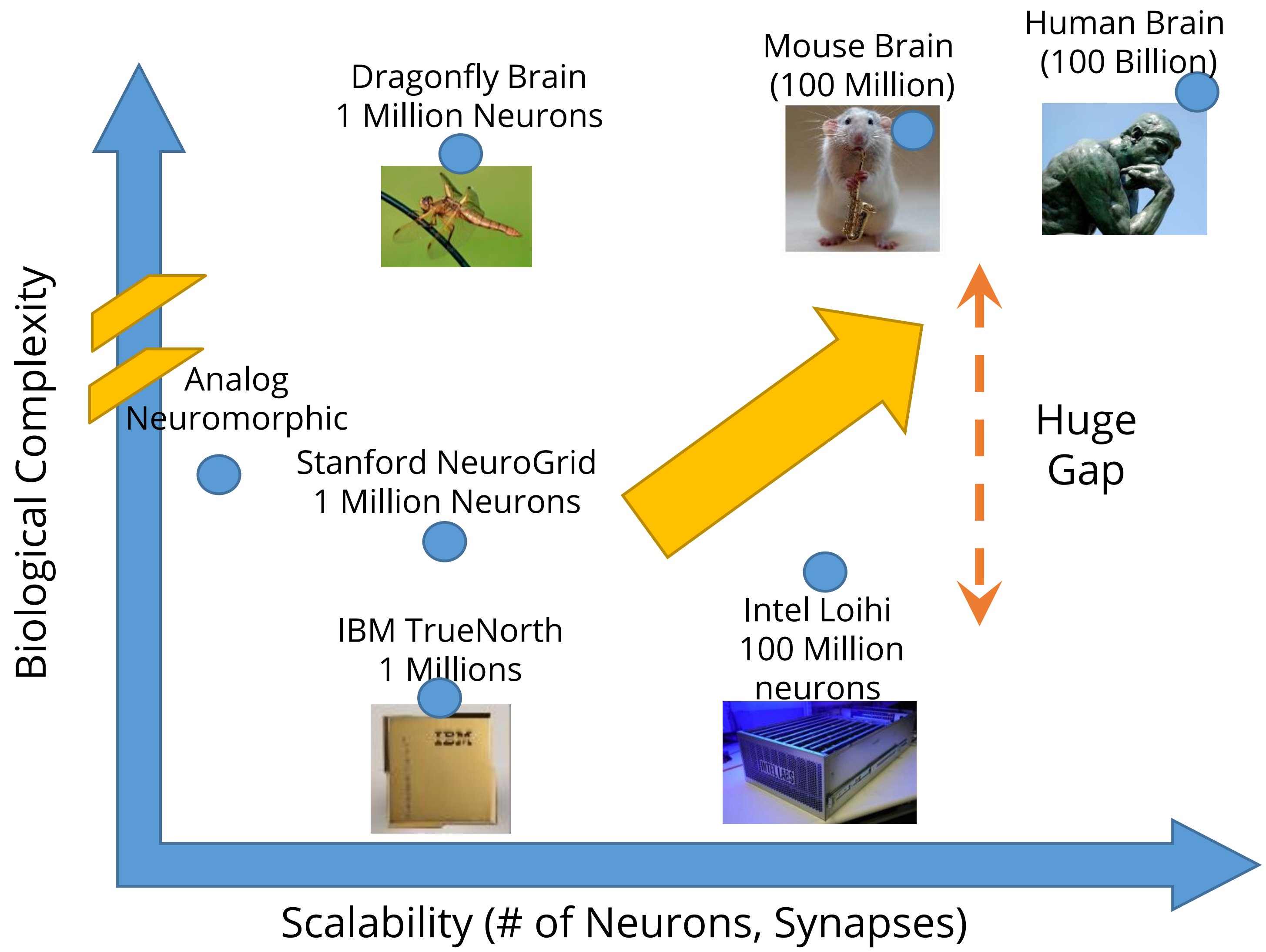
Brain-inspired algorithms



Scientific Computing

Edge Computing

CHALLENGE: SCALABILITY VS. COMPLEXITY



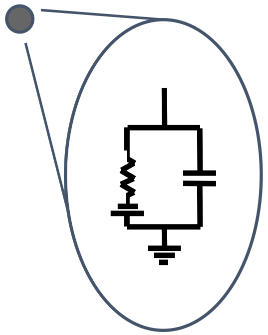
However, to achieve brain-like complexity we need both scaling and rich dynamics.

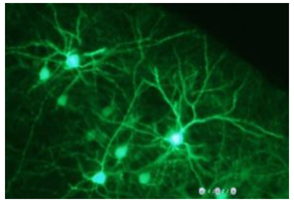
- Solving ill-structured problems
- Online learning
- Transfer learning

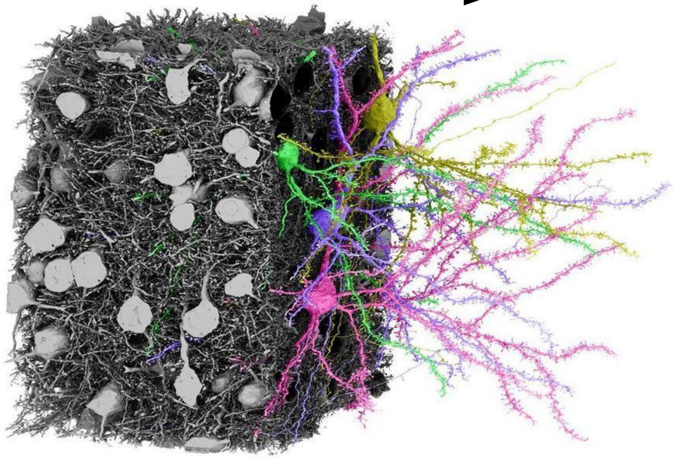
Understanding fundamental mechanisms in neuroscience, translated to algorithms and models will influence next-generation devices, architectures and intelligent computing systems

INCREASING "BIOLOGICAL COMPLEXITY"

Increase computational efficiency
and
Increase computational density







Novel devices and materials
can help bridge this gap.

LIF neuron
- Single passive compartment
- Spikes
- Limited dynamics
- Relatively easy to scale

Biological neuron
- Dendrites = intricate structure and dense connectivity
- Complex pattern of active conductances
- Rich dynamics , multiple patterns of spiking, subthreshold computation
- More computational power, not compact

LIF: leaky Integrate and Fire

DENDRITIC TOOLKIT FOR COMPUTATION

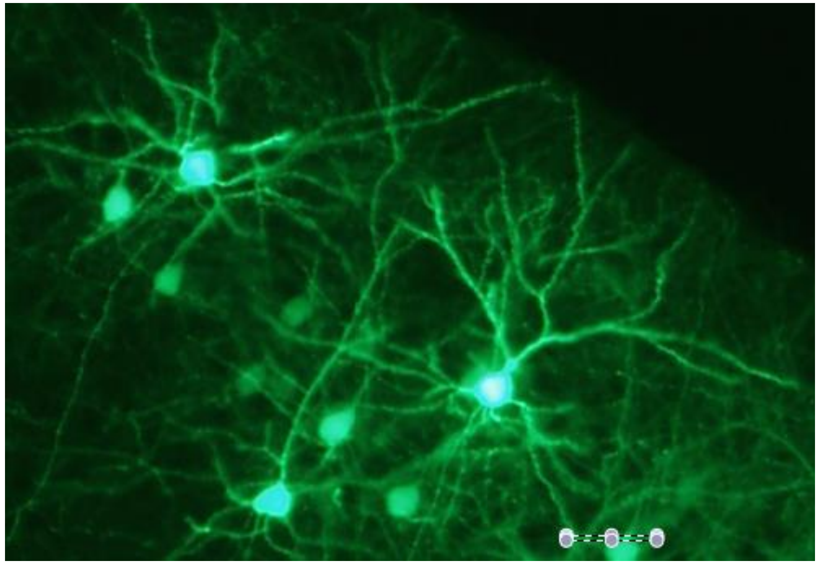
Dendrites are tree-like structures that connect neurons synapses to its soma.

Dendrites are not *just* wires!

They can perform interesting computation like:

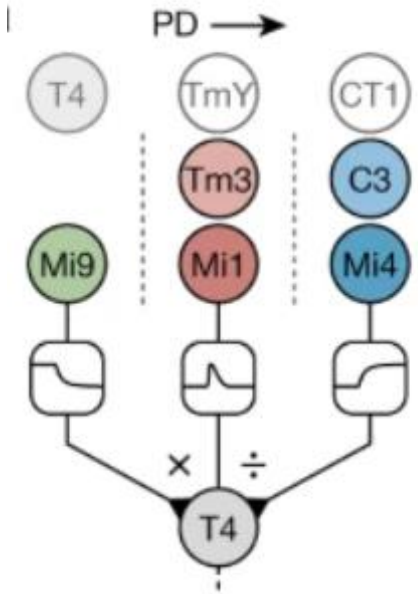
- Coincidence Detection
- Current Summation
- Directional selectivity
- Non-linear filtering
- Amplification of Synaptic inputs

London 2005, Poirazi 2020



Increased Connectivity and Computation

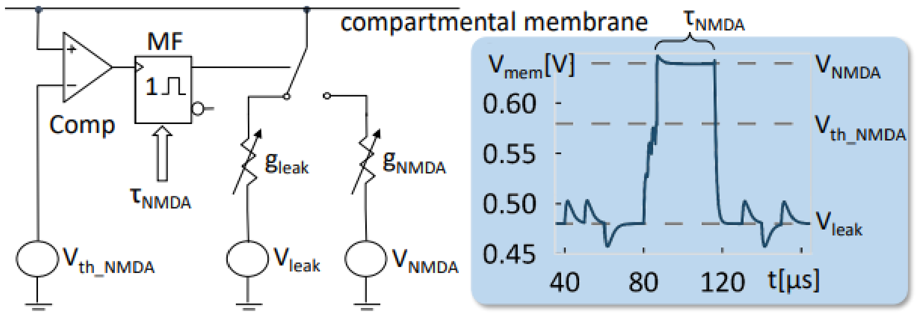
SINGLE NEURON MULTIPLICATION



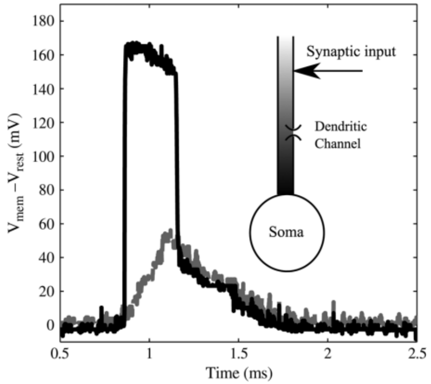
Groschner et al., Nature 2022

Leveraging Inhibition

Shunting Inhibition/ Leveraging Leakage Conductance

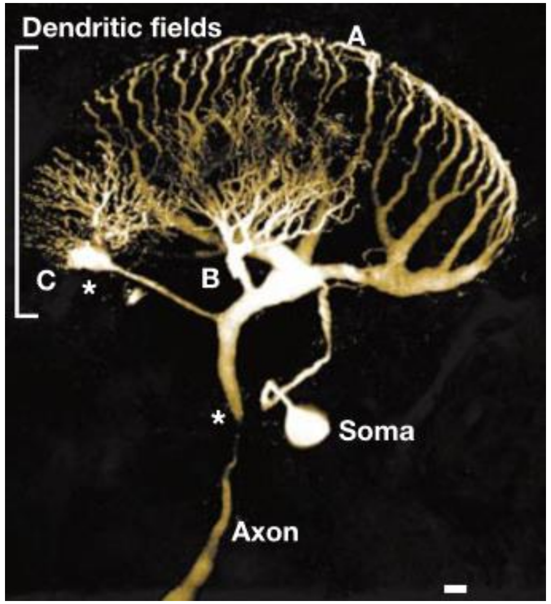


Schemmel, Johannes, et al. , IEEE IJCNN, 2017.



Dendrites with Active Channel, Ramakrishnan et al., IEEE TBIOCAS, 2013.

NMDA /Ca



Lobula giant movement detector (LGMD) of locusts Gabbiani et al., Nature 2002

Multiplication based on dendritic subtraction of two converging inputs encoded logarithmically, followed by exponentiation through active membrane conductances.

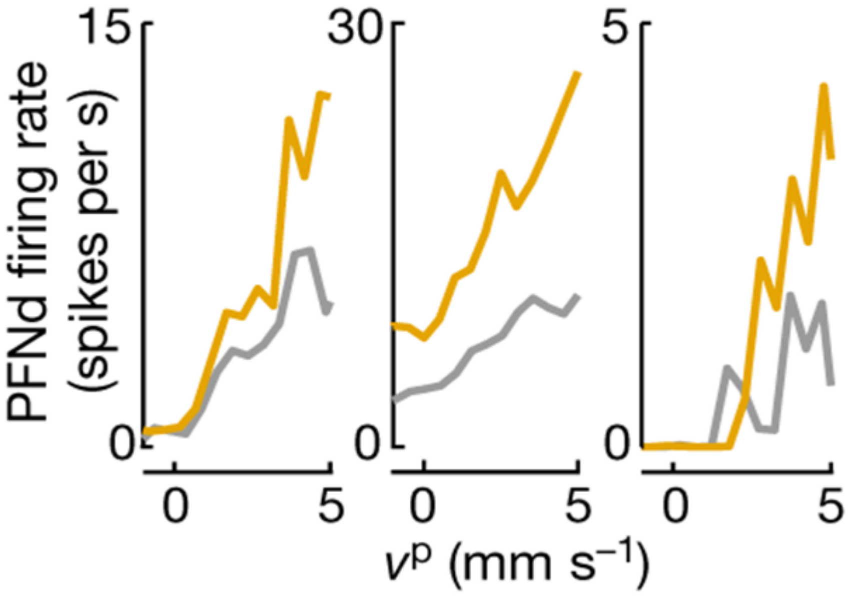
SINGLE NEURON MULTIPLICATION

Algorithms

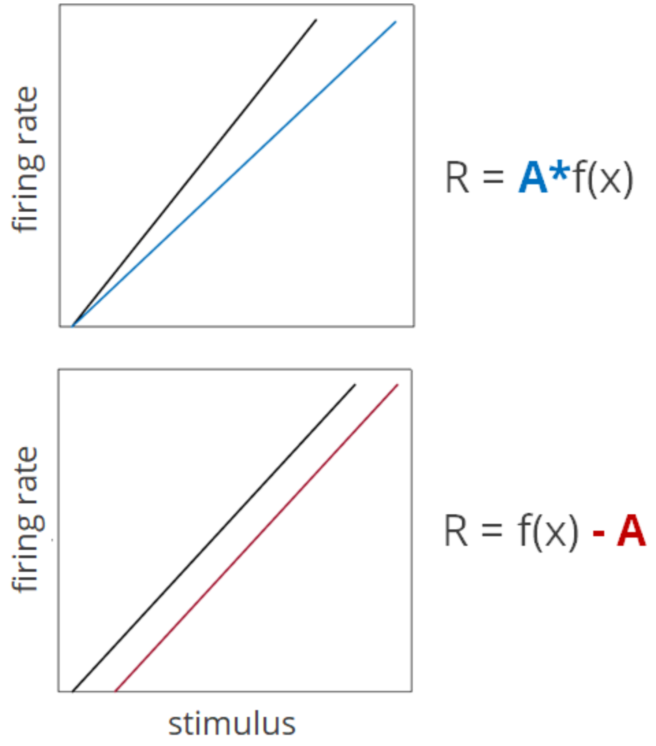
Devices & Circuits

Physics of Computing

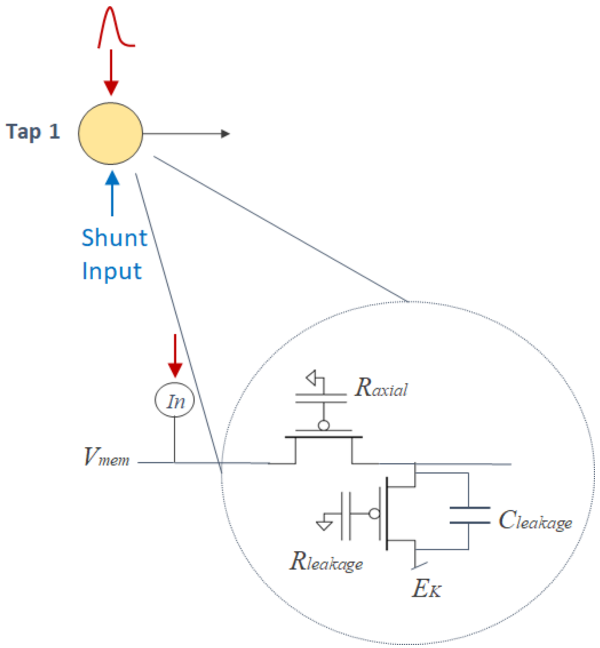
from fan-shaped body of *Drosophila* brain



Lu et al. 2022



Chance & Cardwell
NICE 2023



Shunting Inhibition in
Neuromorphic Dendrite

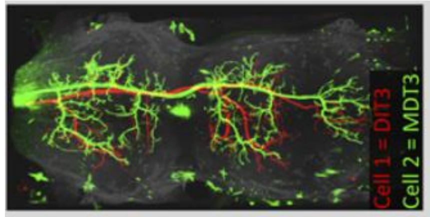
NEUROMORPHIC CODESIGN

Algorithms *Devices & Circuits* *Physics of Computing*

Coordinate transformations from Dragonflies to Neuromorphic Hardware

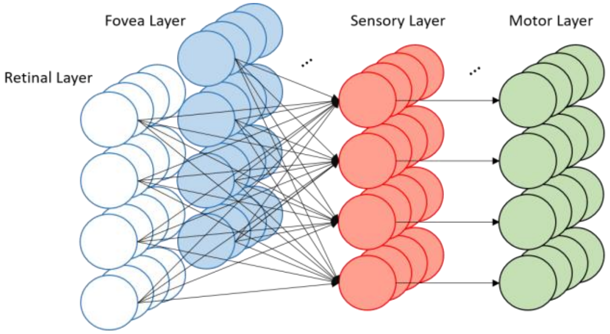
Lead PI: Frances Chance, SNL

October 2021



Gonzalez-Bellido, UMN

DRAGONFLY EXPERIMENTS

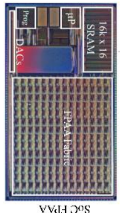


Chance 2020

August 2021

COMPUTATIONAL MODEL

GT FPAA



George Cardwell 2016

Intel's Loihi

Davies 2018

SNL, Baylor

NEUROMORPHIC IMPLEMENTATION

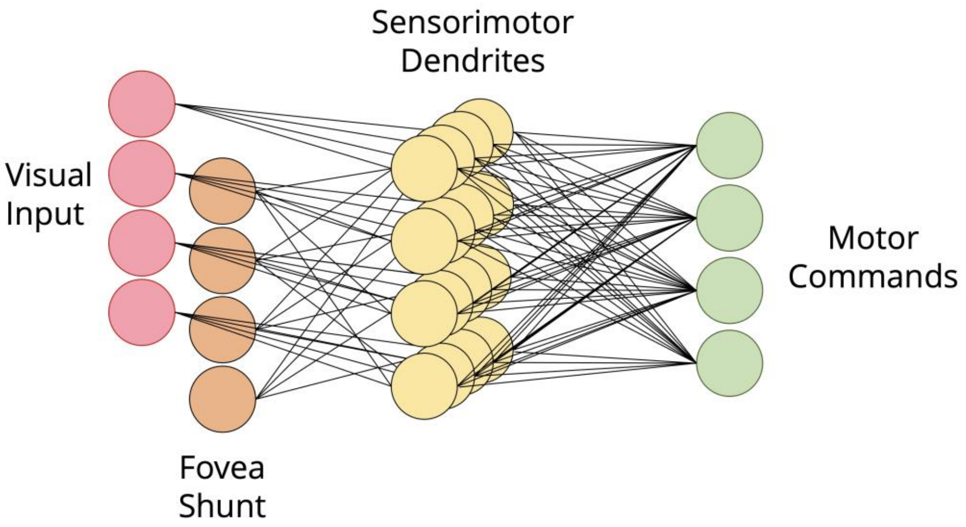
Increased collaboration between neuroscience and neuromorphic engineering will facilitate development of novel neural-inspired architectures.

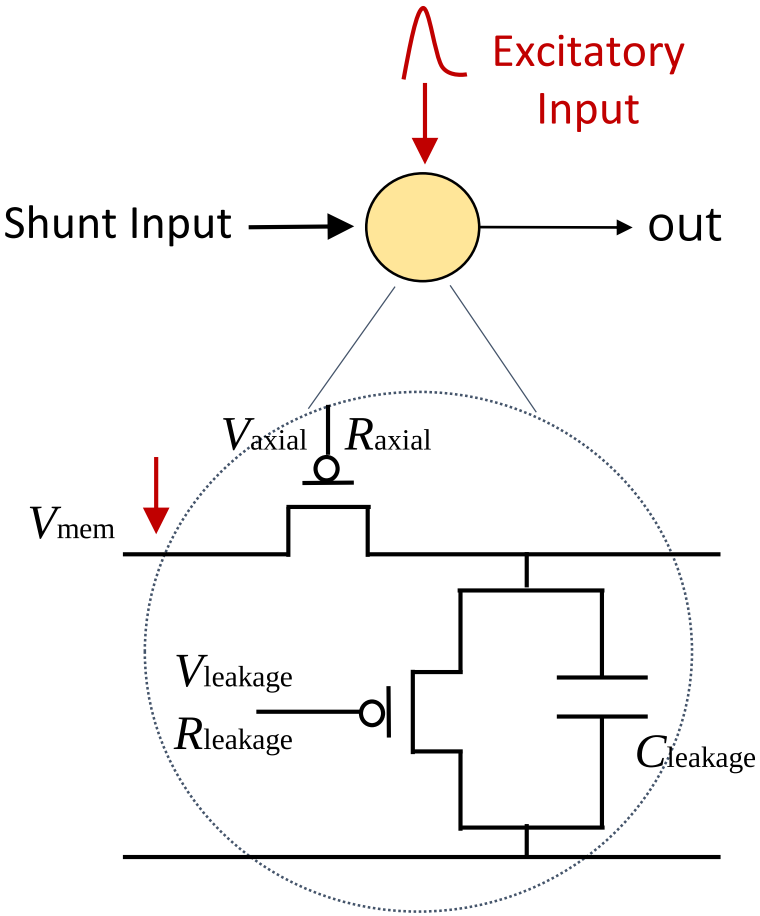
U.S. DEPARTMENT OF ENERGY | Office of Science

DOE ASCR (FY21-24)

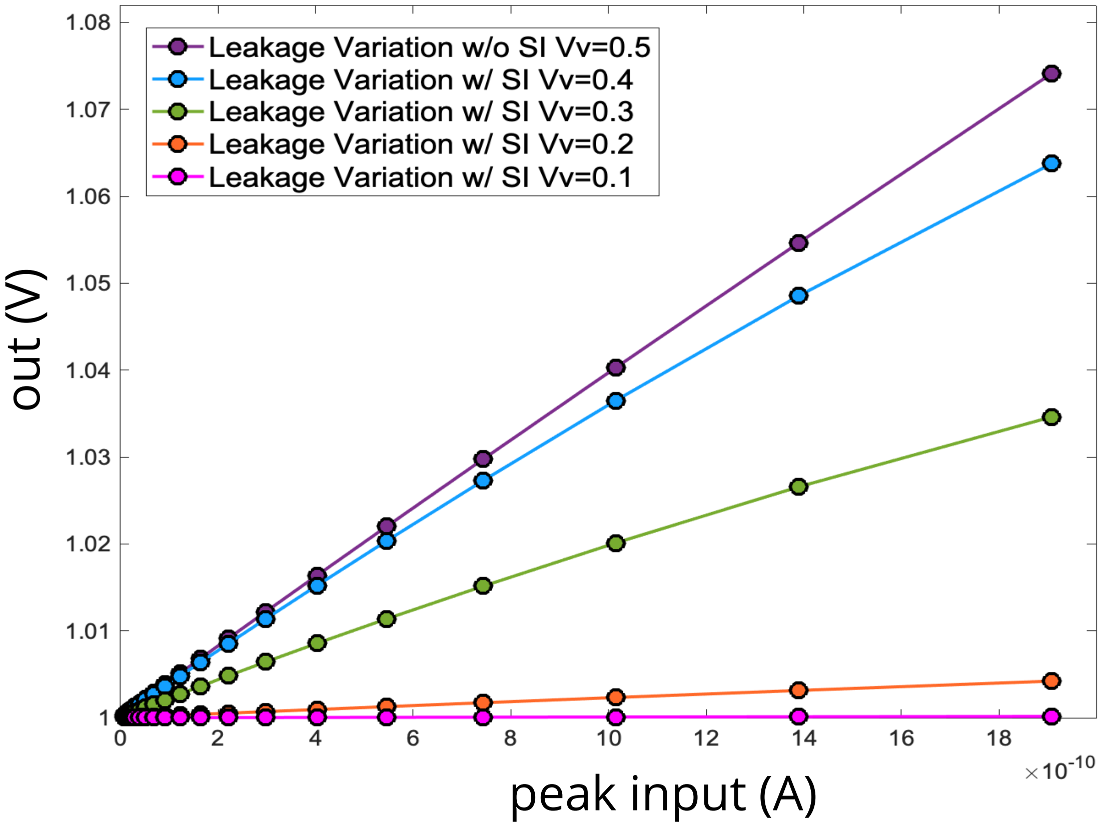
Department of Energy
Advanced Scientific Computing Research

DRAGONFLY WITH DENDRITES

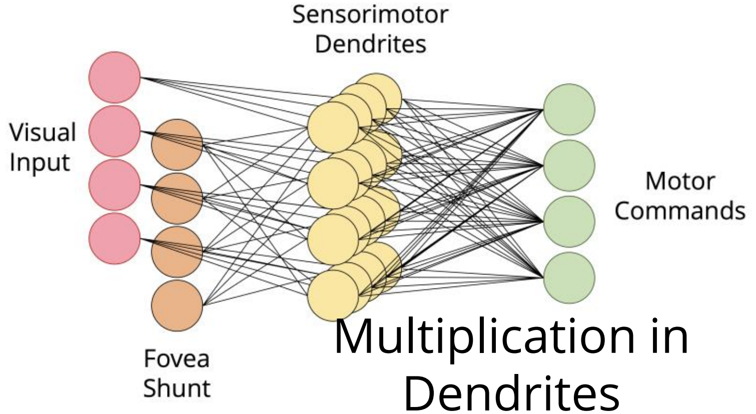




Sub-threshold region of operation



DRAGONFLY INTERCEPTION WITH DENDRITES

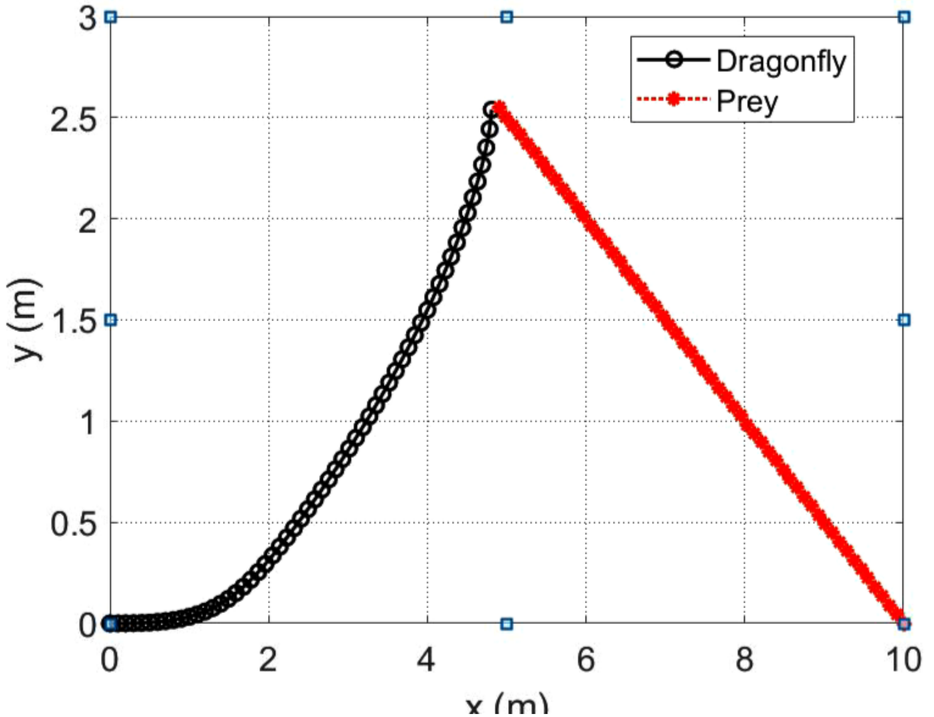


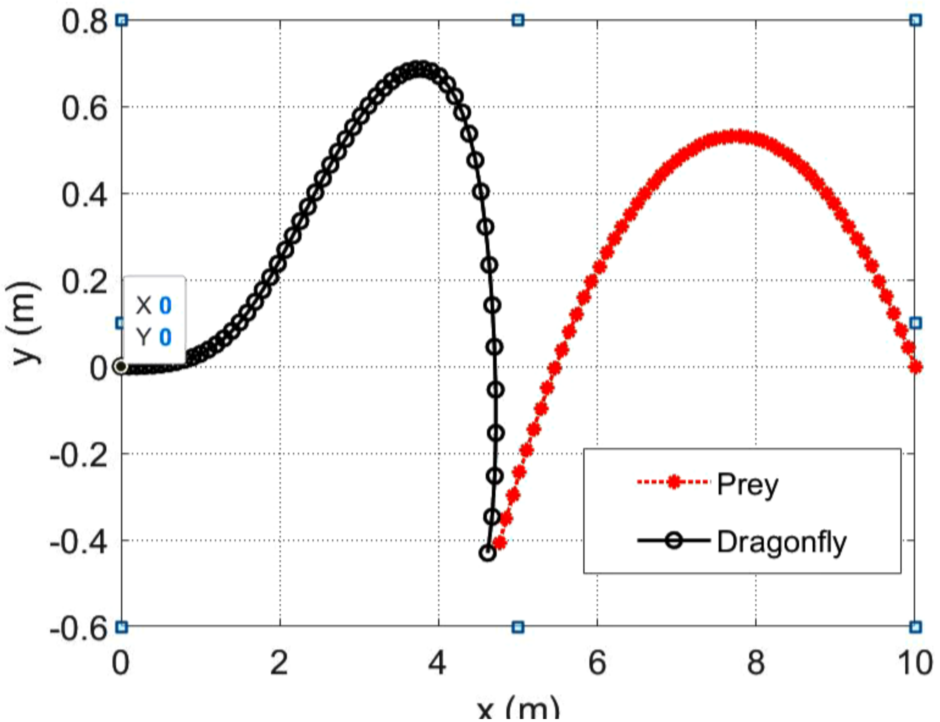
Multiplication in Dendrites

Sensorimotor Dendrites Response

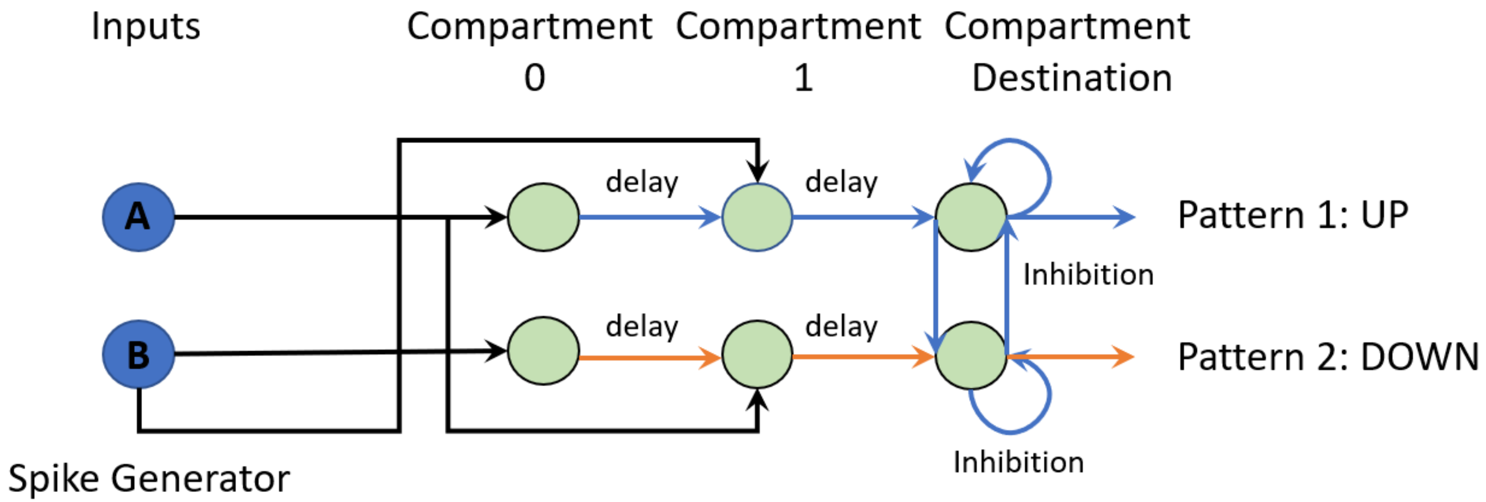
$$S_{ij} = f_i(x) g_j(y)$$

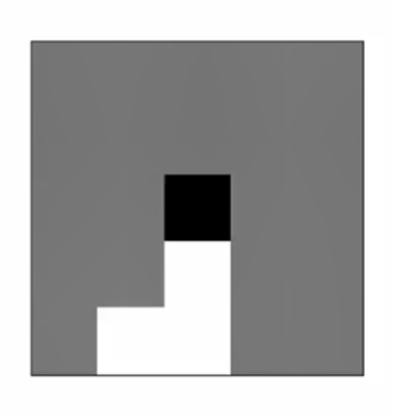
Cardwell & Chance ICONS 23





DIRECTION-SELECTIVE DENDRITES

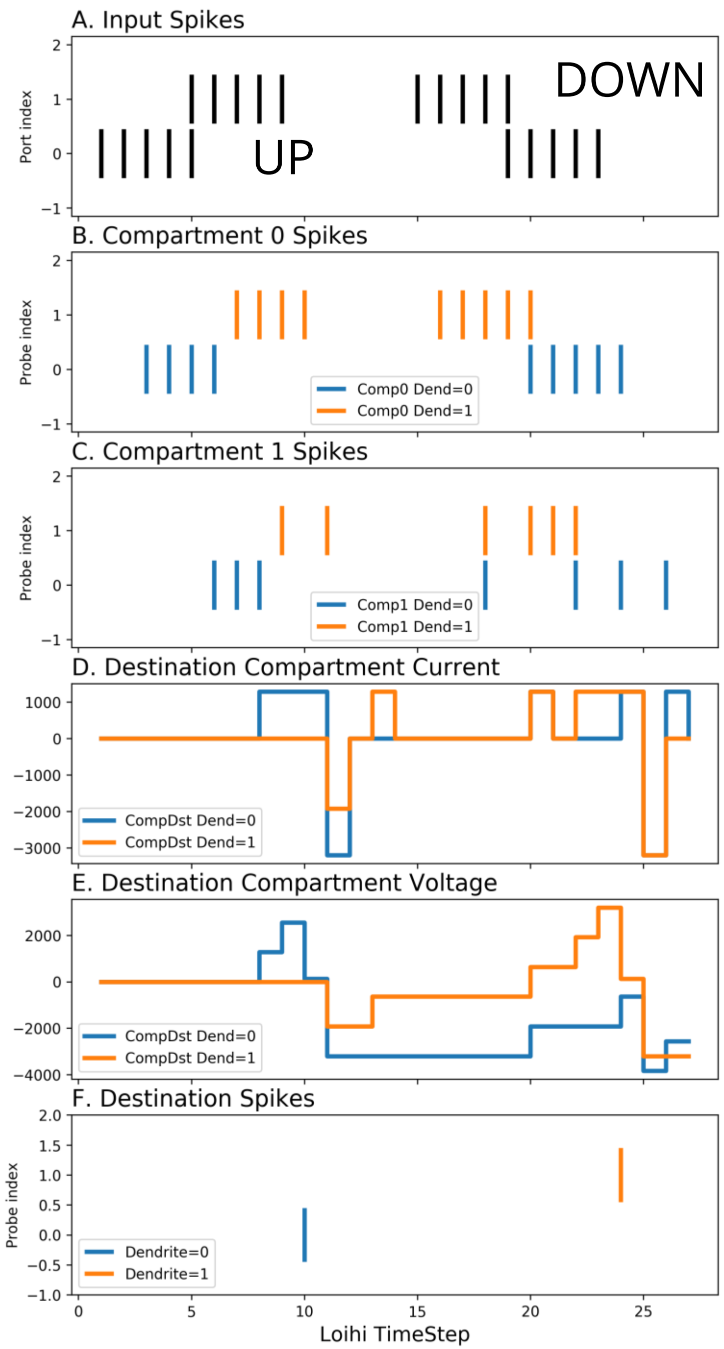




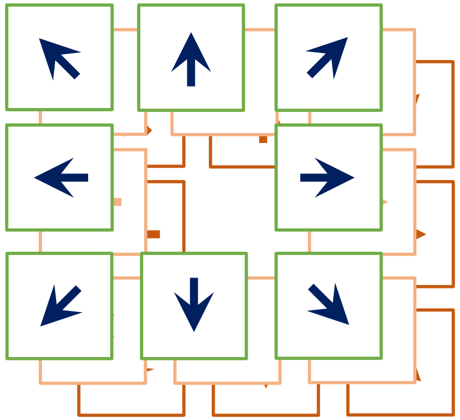
Event Sensor Output

Nahuku (Loihi Chips)
Davies 2018

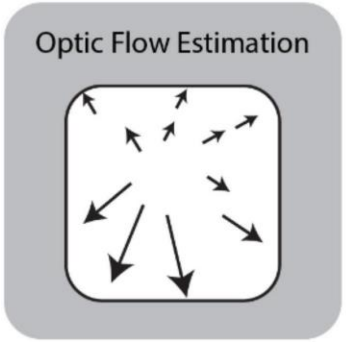
Cardwell & Chance ICONS 23



DIRECTION-SELECTIVE CELLS FOR COMPLEX PATTERNS

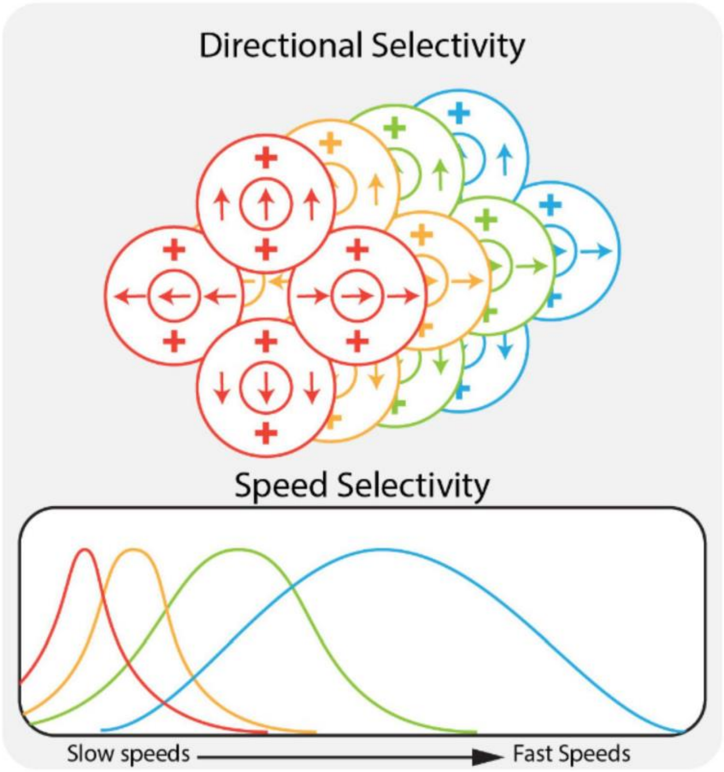


V1



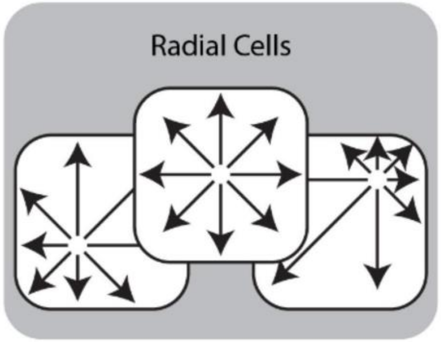
MT (Middle Temporal)

MT+



dorsal medial superior temporal

MSTd

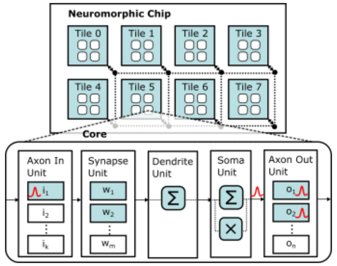


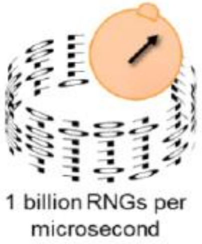
Steinmetz et al. 2022

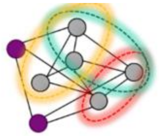
CHALLENGE: CODESIGN TOOLS

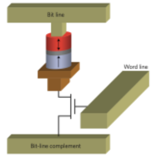
Co-Design Tools for Novel Architectures

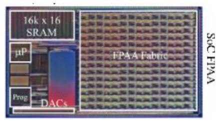
Next-generation Neuromorphic Architectures











ATHENA
Analytical Tool for analog and neuromorphic ML accelerator

ASC-AML (FY20-22)

SANA-FE
Neuromorphic Architecture Exploration

SNL LDRDs (FY21-24)

AI-Enhanced Codesign
Reinforcement Learning/Evolutionary for Circuit and System design

SNL LDRDs (FY21-23)

COINFLIPS
Probabilistic Neural Computing, Leverage stochasticity in beyond-CMOS devices

DOE ASCR/BES (FY21-24)

DRAGONFLY
Dendritic processing, Coordinate transformation from Dragonflies to Neuromorphic hardware,

DOE ASCR (FY21-24)

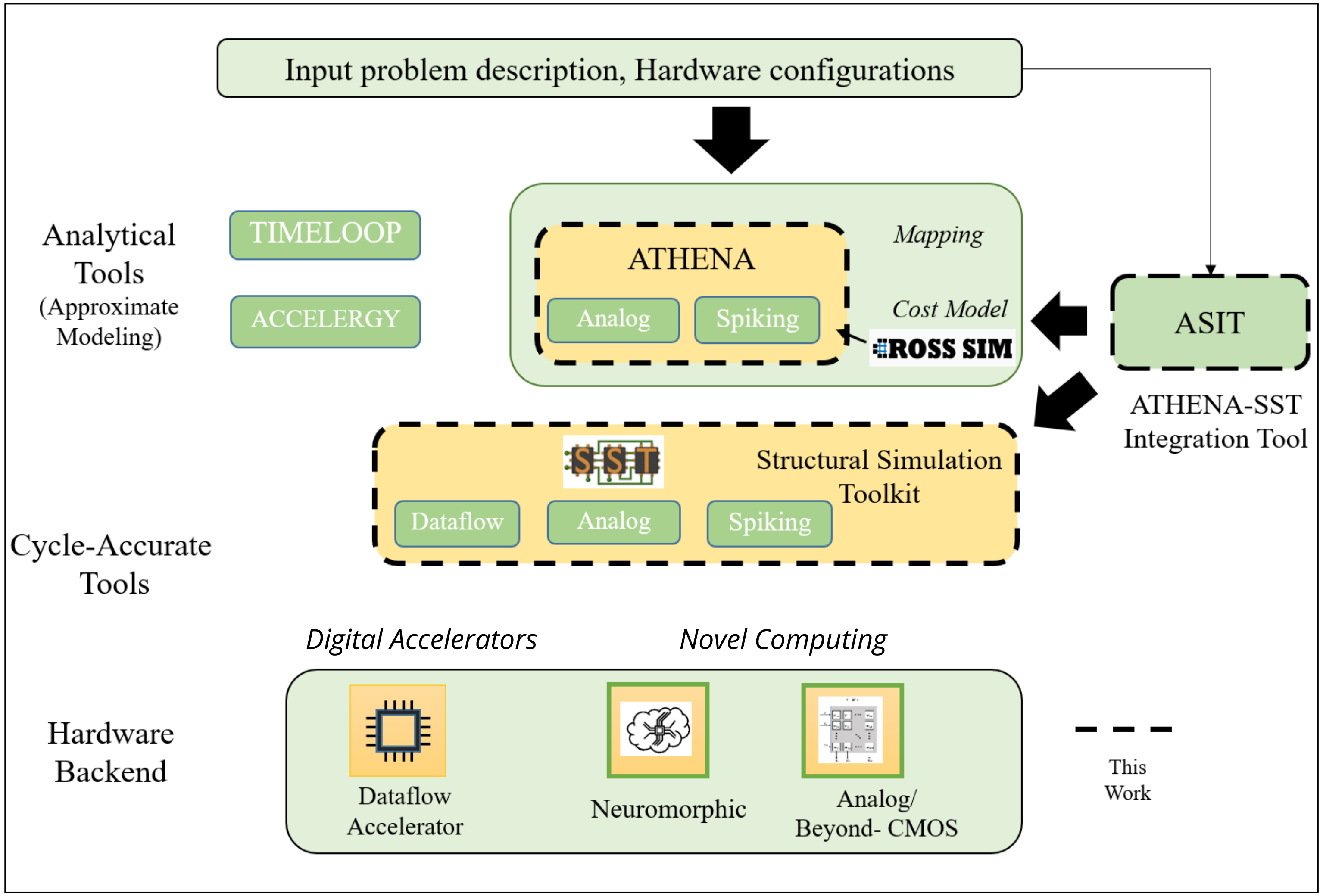
External Collaborators: UT Austin, Intel, NCSU, Infineon Memory Solutions, Georgia Tech, UMN, Baylor University, UT Knoxville, Temple University, NYU, ORNL

ATHENA : ANALYTICAL TOOL TO EVALUATE HETEROGENEOUS NEUROMORPHIC ARCHITECTURES

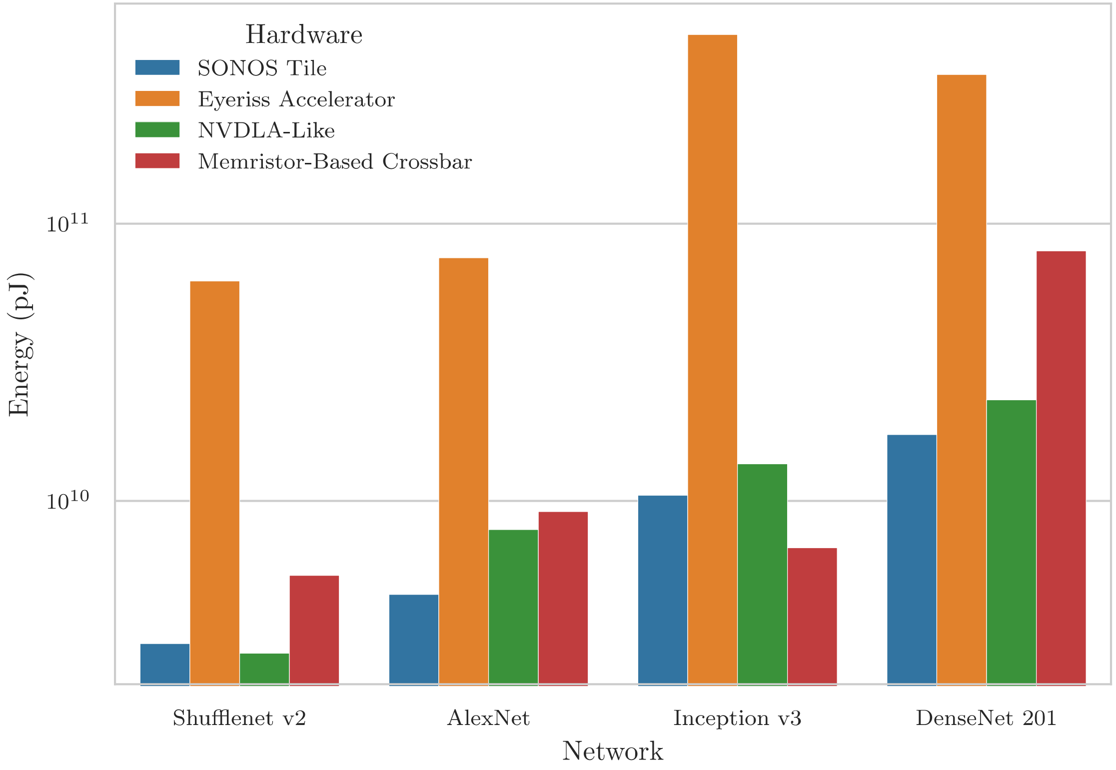
- ATHENA will quickly evaluate performance metrics of analog architectures

- 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



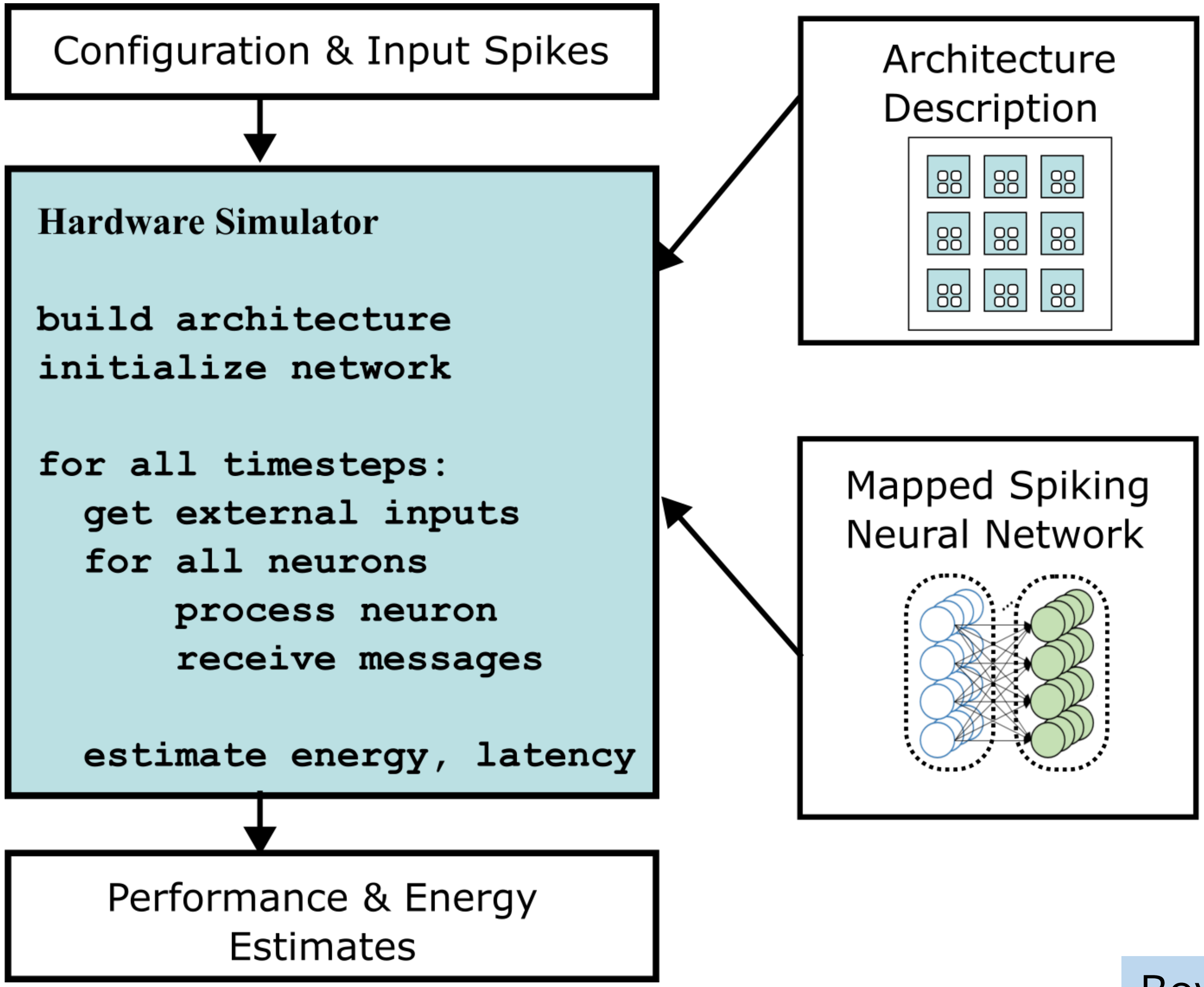
ATHENA – HARDWARE PERFORMANCE



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.
- In the process of making ATHENA open-source.

SANA-FE: Simulating Advanced Neuromorphic Architectures for Fast Exploration



SANA-FE

UT Austin Collaboration

Boyle et al., ICONS 2023

SPIKING ARCHITECTURE TEMPLATE

Tile-based architecture

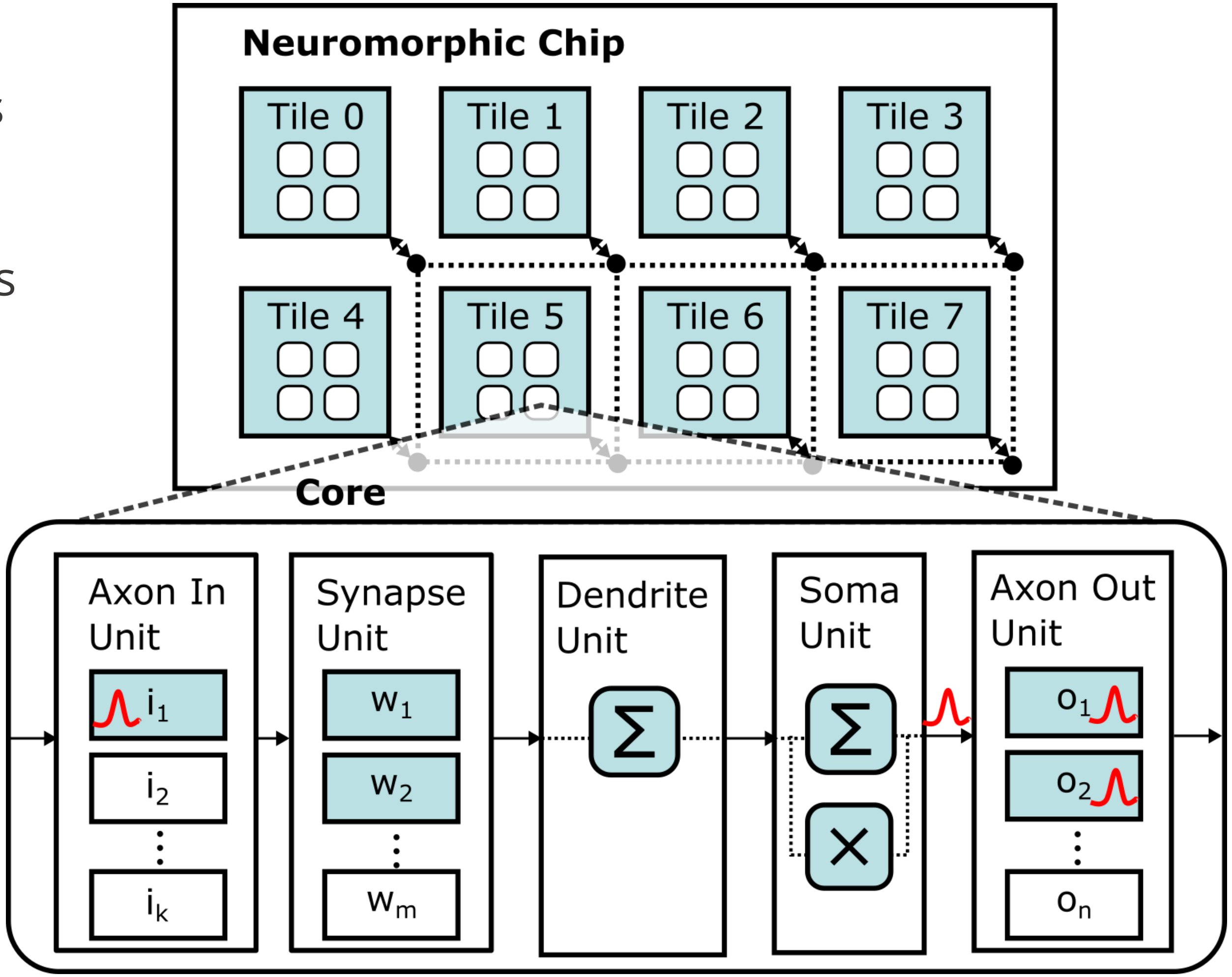
- Network-on-chip connecting neural cores

Many cores per tile

- Cores simulate group of mapped neurons
- Local shared memory

Core pipeline

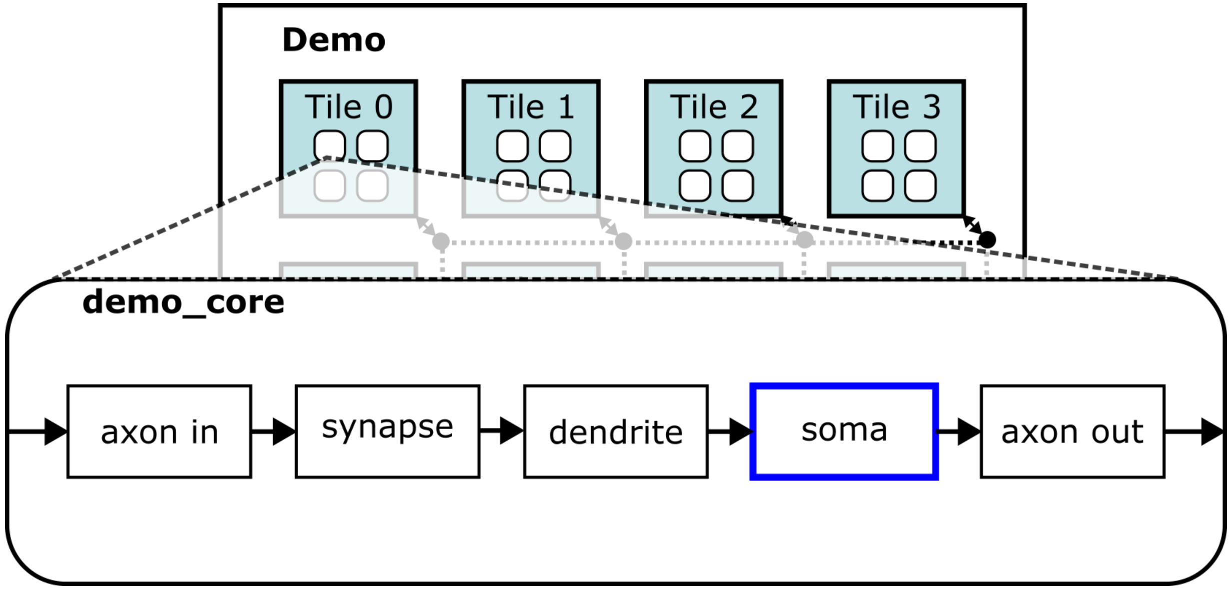
- Axon stage
- Synapse stage
- Dendrite stage
- Soma stage



CHALLENGE: ARCHITECTURE DESCRIPTION

Describes different H/W architectures

- Represents different existing & future spiking designs based on common features
- Defines compute elements of chip
- YAML-based, flexible & extensible

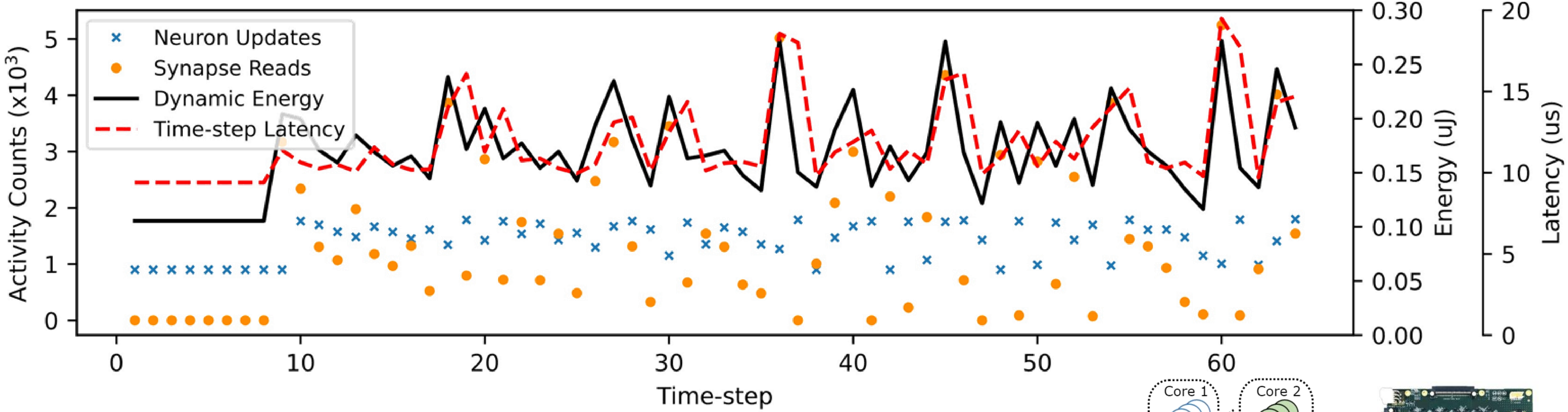


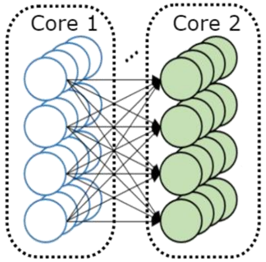

```yaml
architecture:
 name: demo
 tile:
   - name: demo_tile[0..7]
     attributes:
      energy_east_west: 1e-12
      latency_east_west: 2e-9
      ...
     core:
      - name: demo_core[0..3]
        soma:
          - name: core_lif
            attributes:
             energy_spiking: 68e-12
             latency_spiking: 30e-9
...
```

Boyle et al., ICONS 2023

PERFORMANCE MODELING RESULTS

For randomized spiking inputs on the application SNN





- Detailed breakdown of on-chip activity on Loihi
- Captures dynamic energy and latency trends

➢Detailed insight into H/W behavior

Boyle et al., ICONS 2023

RESULTS FOR OTHER NEUROMORPHIC BENCHMARKS

Predict performance & energy for larger real-world neuromorphic applications

- SNN trained on DVS gesture data-set
- 18,678 neurons across 6 layers
- Mapped to 45 Loihi cores out of 128

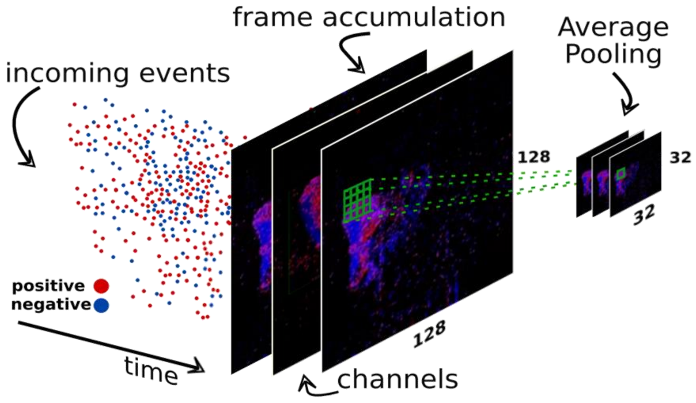
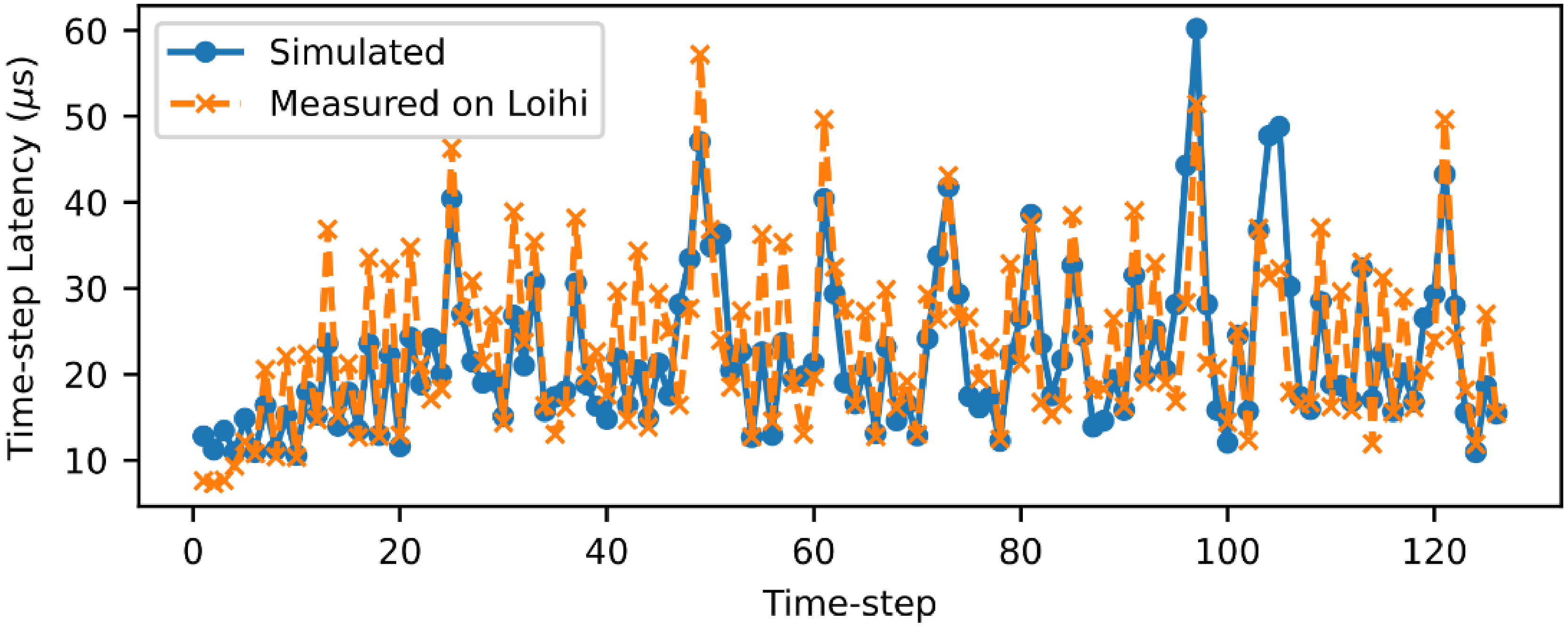
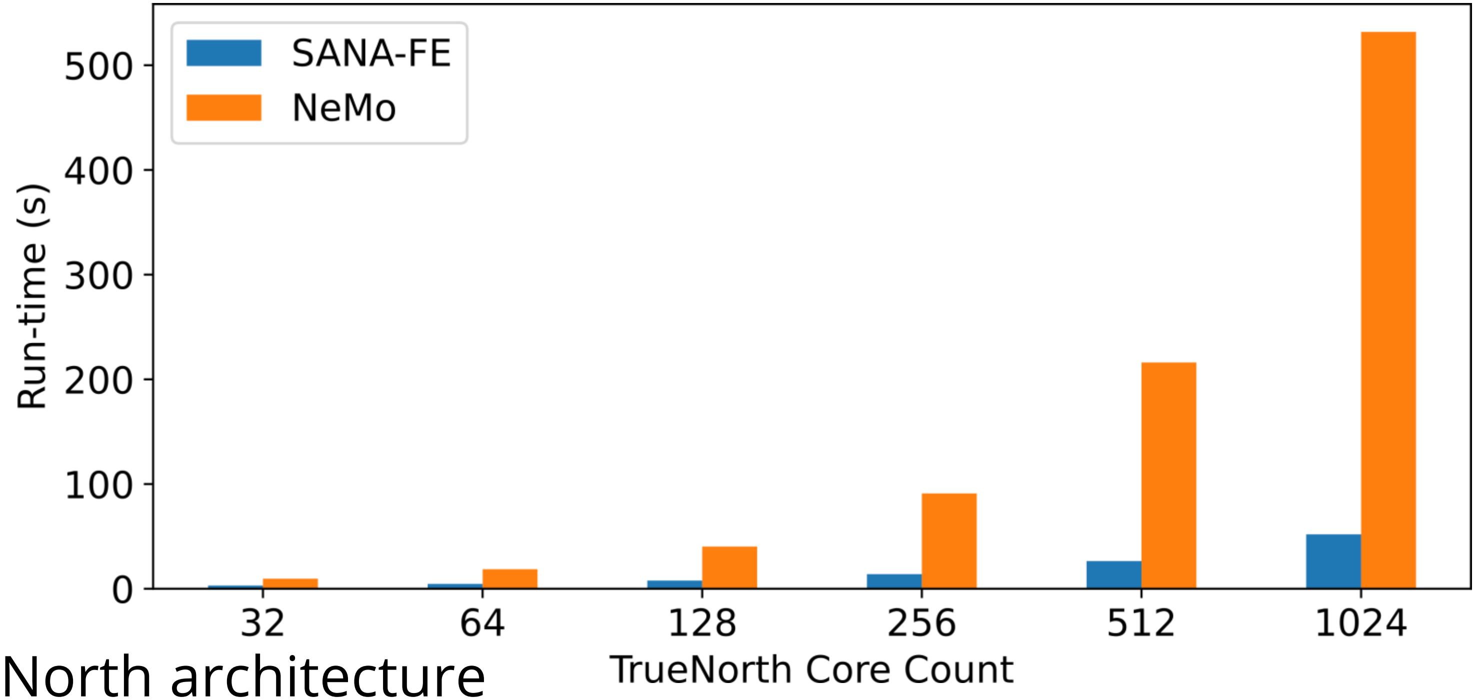


Image reproduced from [Massa,'20]



SIMULATOR SPEED RESULTS

- Compared to existing spiking simulator (NeMo)



  - Simulating IBM TrueNorth architecture
  - Randomized SNN with 80% of spikes intra-core, 20% spikes between cores
- **Over 10x faster than NeMo for 1024 cores**

SANA-FE

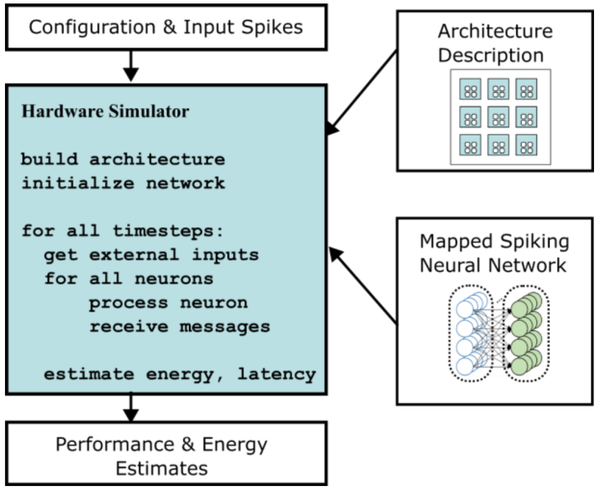
Generic & extensible
- User-defined architecture & SNN
- Supports range of spiking architectures

Fast & accurate
- Time-step based approach
- Detailed hardware activity for each time-step
- Accurately estimates performance & energy

Future work
- Support other existing architectures & scale to larger designs
- Adapt other neuromorphic benchmark applications
- Model analog architectures & novel devices
- Integrate with other frameworks e.g., SST, Fugu & Lava

Access at: https://github.com/SLAM-Lab/sana-fe

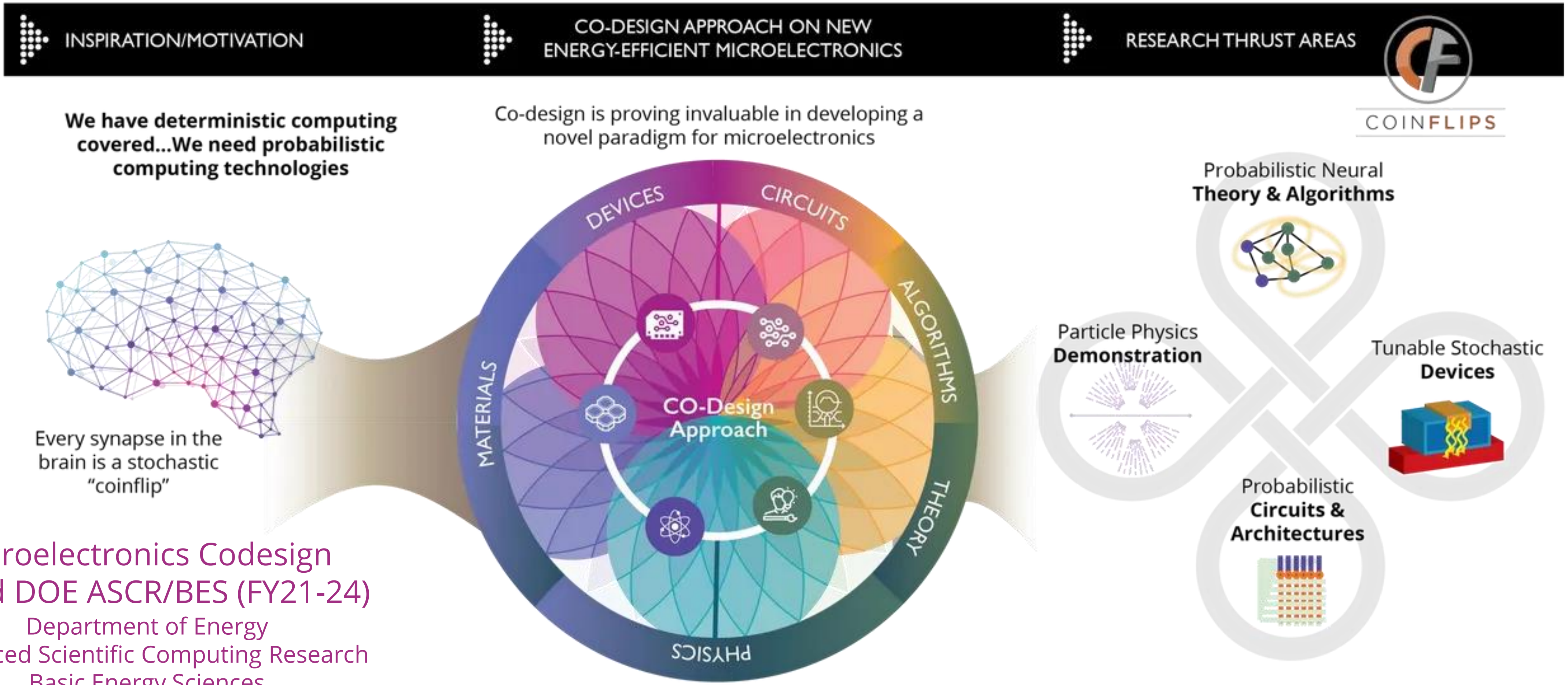
Prof. Andreas Gerstlauer's SLAM Lab @ UT Austin



Boyle et al., ICONS 2023

AI-ENHANCED CODESIGN: COINFLIPS

Lead PI: Brad Aimone



Microelectronics Codesign Award DOE ASCR/BES (FY21-24)

Department of Energy

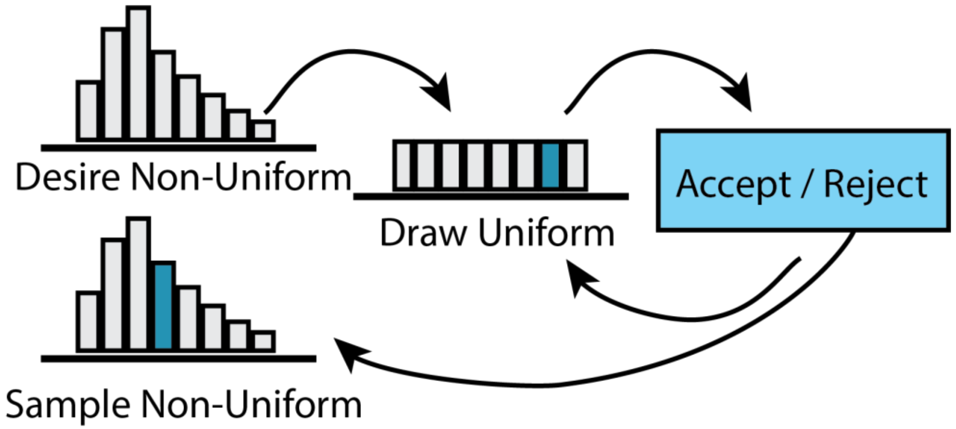
Advanced Scientific Computing Research

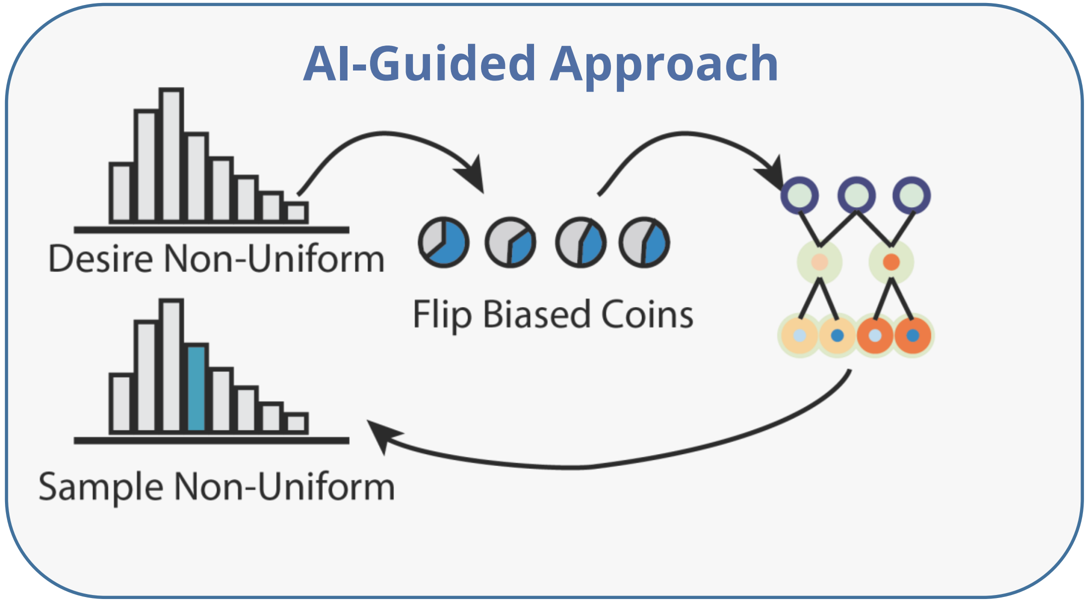
Basic Energy Sciences

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

https://coinflipscomputing.org/

AI-GUIDED CODESIGN OF PROBABILSITIC CIRCUITS





Unfair coins can be combined with AI-designed neural circuits to allow sampling of application desired probability distributions, avoiding accept/reject steps.

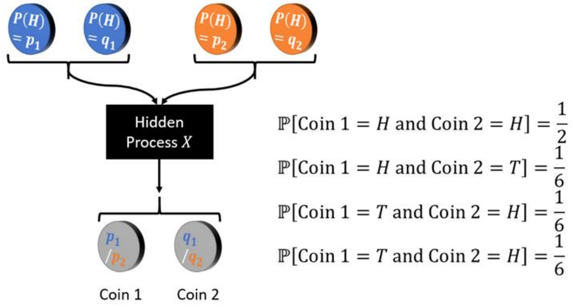
We leveraged evolutionary algorithms for circuit design and optimization

- **LEAP** (Library of Evolutionary Algorithms in Python)
- **EONS** (Evolutionary Optimization for Neuromorphic Systems)- Schuman et al. , 2020

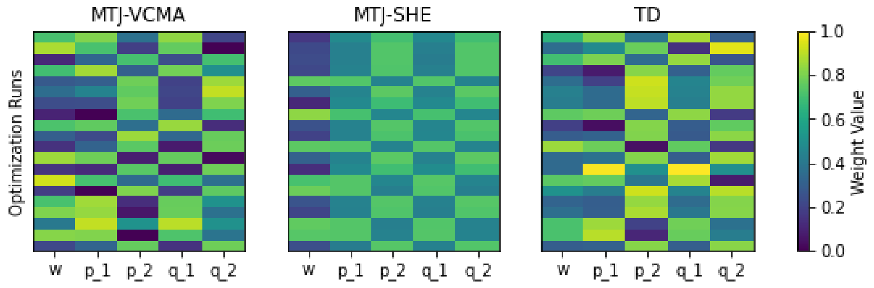
We used abstracted device models for TD and MTJ to capture functionality and energy usage.

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

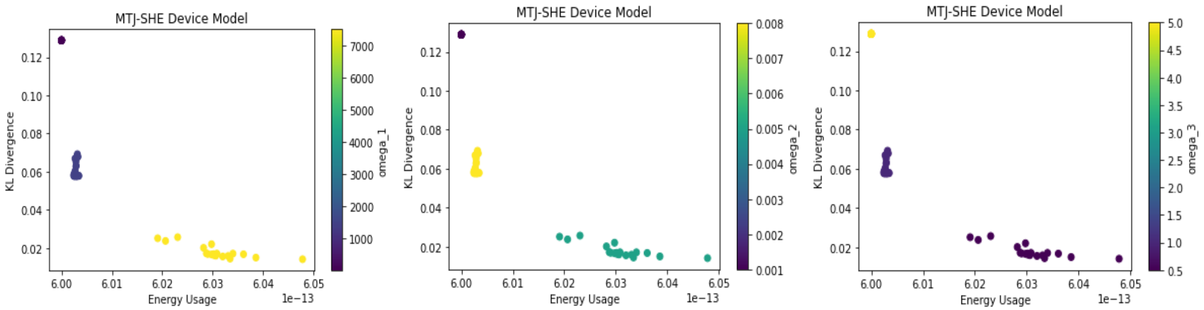
AI-GUIDED CODESIGN OF PROBABILSITIC CIRCUITS



Probabilistic Mixing Algorithm



Optimized weight values for each device



Multi-objective optimization of weights of fitness function for optimal KL divergence, biased weight and energy usage.

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

Weights are customized for the device's behavior to target the best performance in terms of KL divergence and energy usage.

One of the challenges in optimizing for both algorithms and devices was appropriately abstracting the device models and algorithmic constraints.

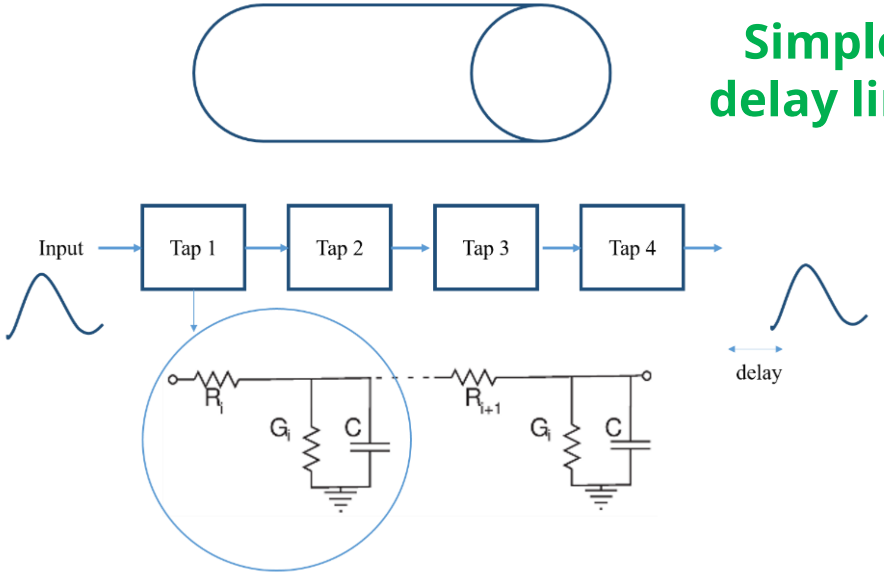
The functional models developed will also be evolved in time as new device data and research emerges.

Our framework can accommodate any emerging device type.

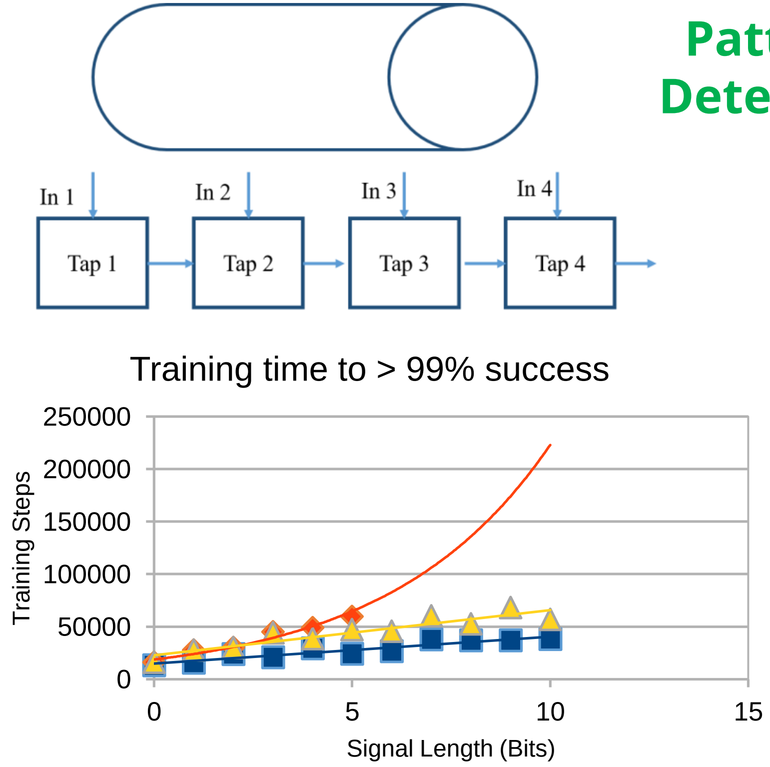
AI-ENHANCED CODESIGN: NEURAL CIRCUITS

We developed an RL algorithm approach which is capable of building very simple circuits.

Simple delay line



Pattern Detection



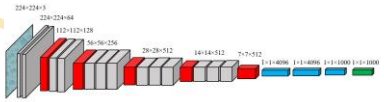
Crowder et al. 2023 (MWCAS 2023)

Algorithms

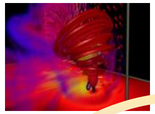
Physics of Computing

Devices & Circuits

NEURONMORPHIC APPLICATIONS

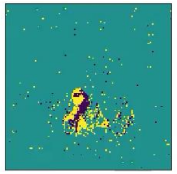


AI/ML Applications
ANNs
SNNs

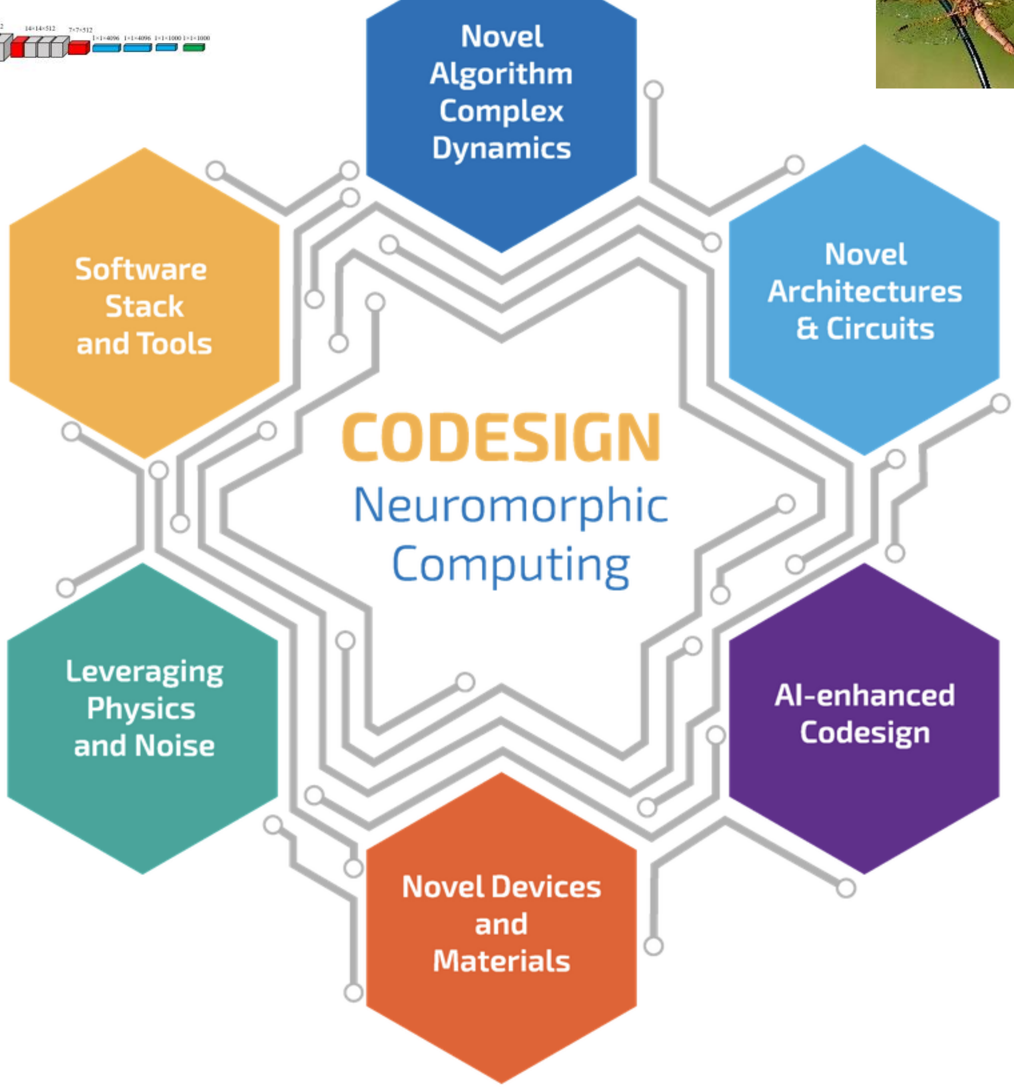


Scientific Computing
- Random Walks

High-fidelity Physics Simulations



Edge Computing
- Event sensors
- Spatio-temporal processing



Brain-inspired Algorithms
Dragonfly
Dendritic Processing

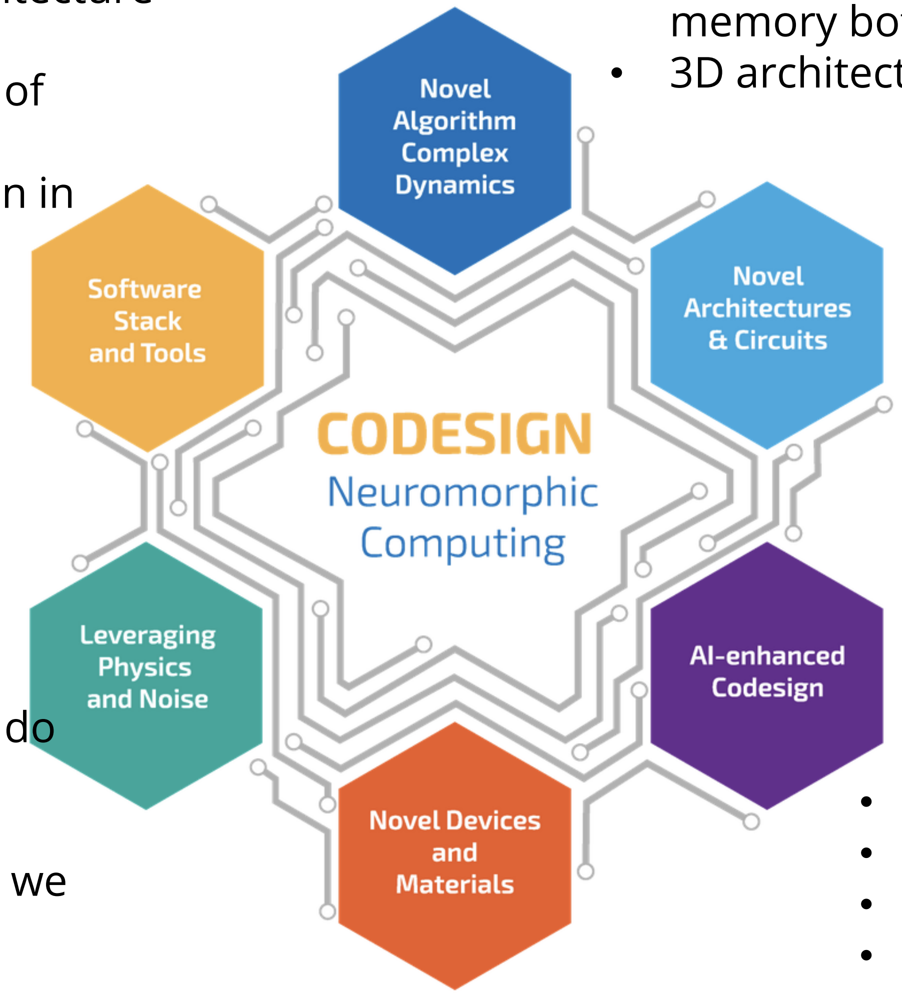
Probabilistic Computing
COINFLIPS

Heterogeneous Computing Applications

CHALLENGES FOR NEXT-GENERATION OF NEUROMORPHIC SYSTEMS

- Algorithms are cognizant of architecture and device constraints.
- Leverage the complex dynamics of devices.
- Bio-inspired techniques, adoption in computing

- Heterogeneous architectures
- CoDesign to optimize communication and memory bottlenecks
- 3D architectures, Photonics



- How can AI-enhanced techniques accelerate scientific discovery?
- Different AI techniques at the device, circuit, system design and architecture level.
- Enable encoding of domain knowledge
- Enable concurrent contribution from researchers

- Software tools to support design and development
- Integration with AI-enhanced techniques?

- Leverage the physics of devices to do computation (analog)
- Embrace stochasticity of devices
- Analog devices are noisy. How can we incorporate this into algorithms?

- Novel devices with complex dynamics
- Radiation-hardened devices
- Reconfigurable devices
- Computational efficiency and computational density

THANK YOU!!

Neural Exploration and Research Lab

WE ARE HIRING!

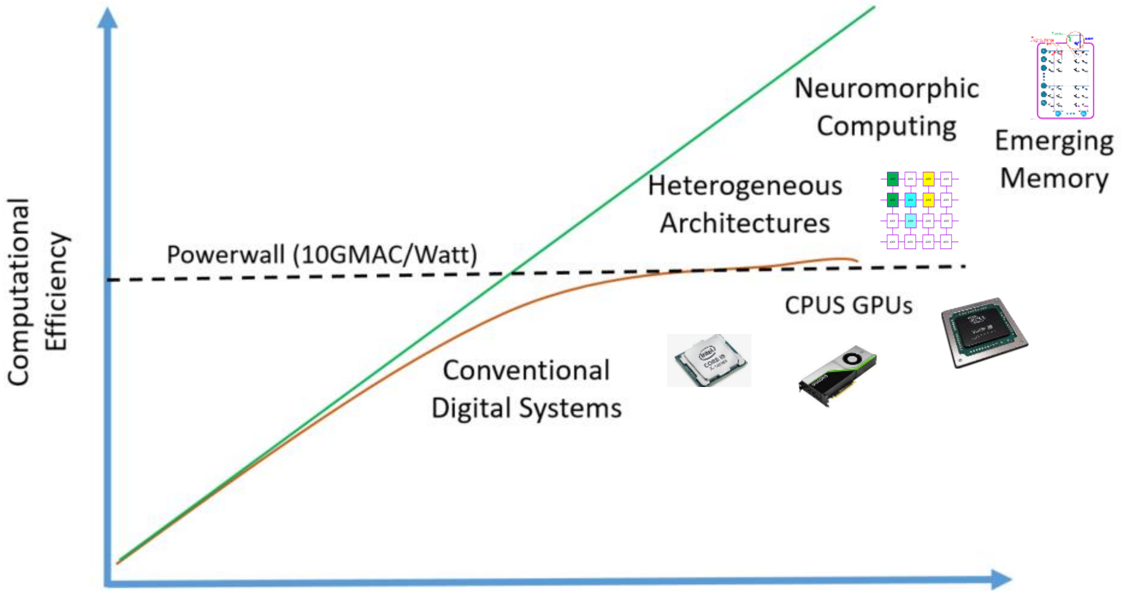
Careers

careers.sandia.gov

https://neuroscience.sandia.gov/

Suma G. Cardwell sgcardw@sandia.gov

FUTURE OF COMPUTING: HETEROGENEOUS ARCHITECTURES



Co-Design is critical to build the next-generation heterogeneous systems

Limits of scaling have ushered in the 'Golden Age of Computer Architecture'

Heterogeneous Architectures

Emerging memory

Neuromorphic Accelerators

Extremely Heterogeneous

Quantum Computing

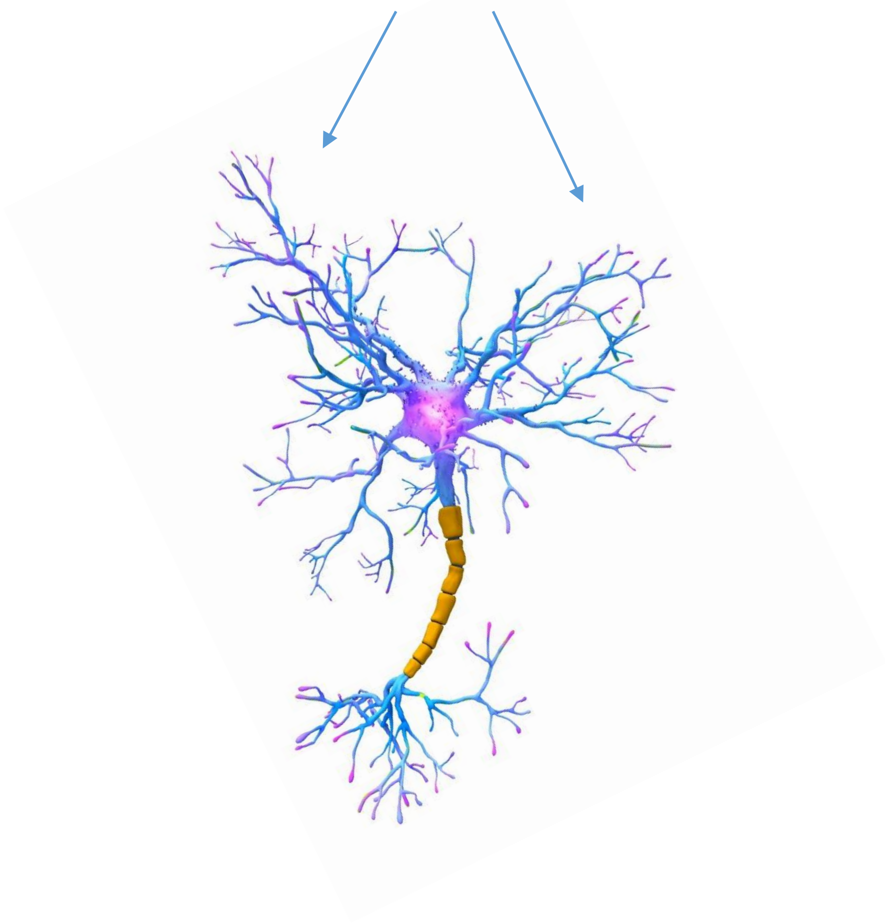
5-10 years

15-20 years

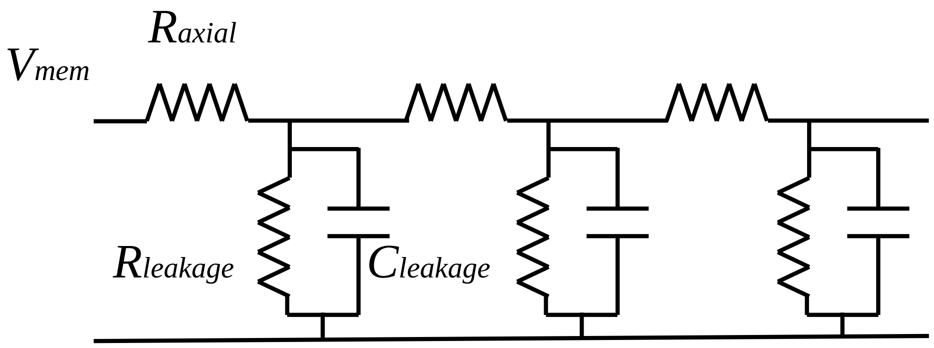
['Truly Heterogeneous Computing', Cardwell et al., SMC 2020]

DENDRITE MODELING

DENDRITES

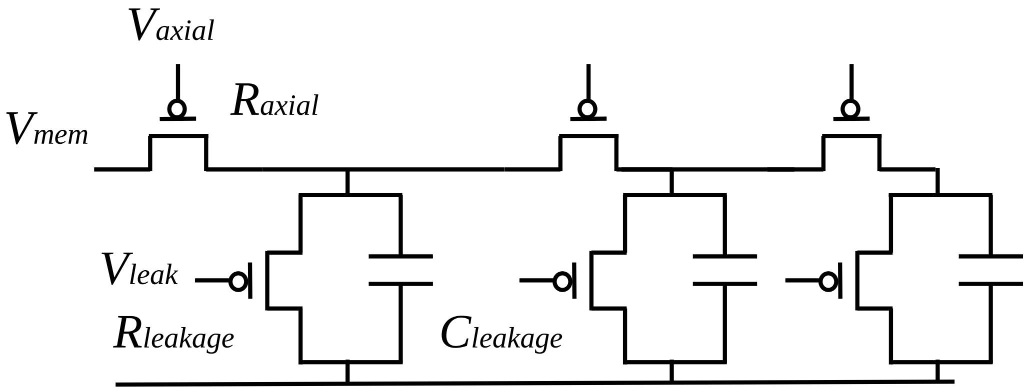


Resistor-Capacitor Circuit



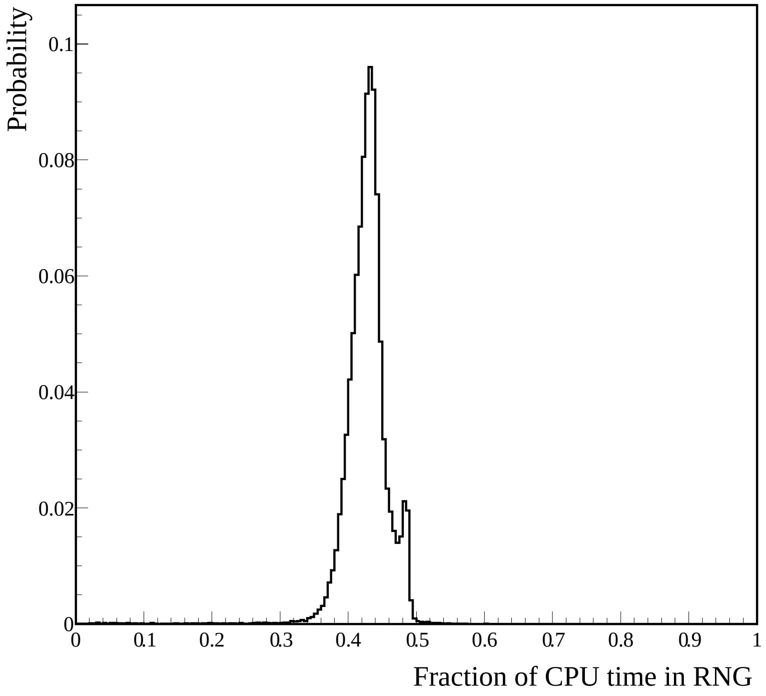
Rall's Cable Model

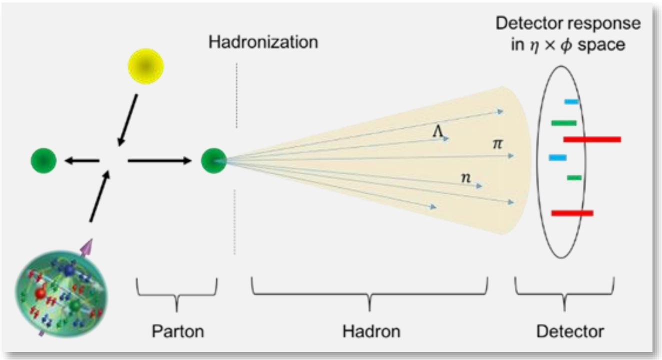
CMOS-transistor based Dendrite



Nease et al. 2011

COINFLIPS APPLICATION: NUCLEAR PHYSICS SIMULATIONS





For a particular collider physics simulation [Pierog et al., Phy Rev. 2022], ~ 270K pseudo- random numbers needed for a single event, with billions of events needing to be simulated.

CPU time is ~ 30-50% of the total compute time

Direct random number generation leveraging stochastic devices can promise significant energy savings for such applications

Misra et al., *Advanced Materials 2022*

Random numbers are a limiting computational cost for some nuclear physics applications