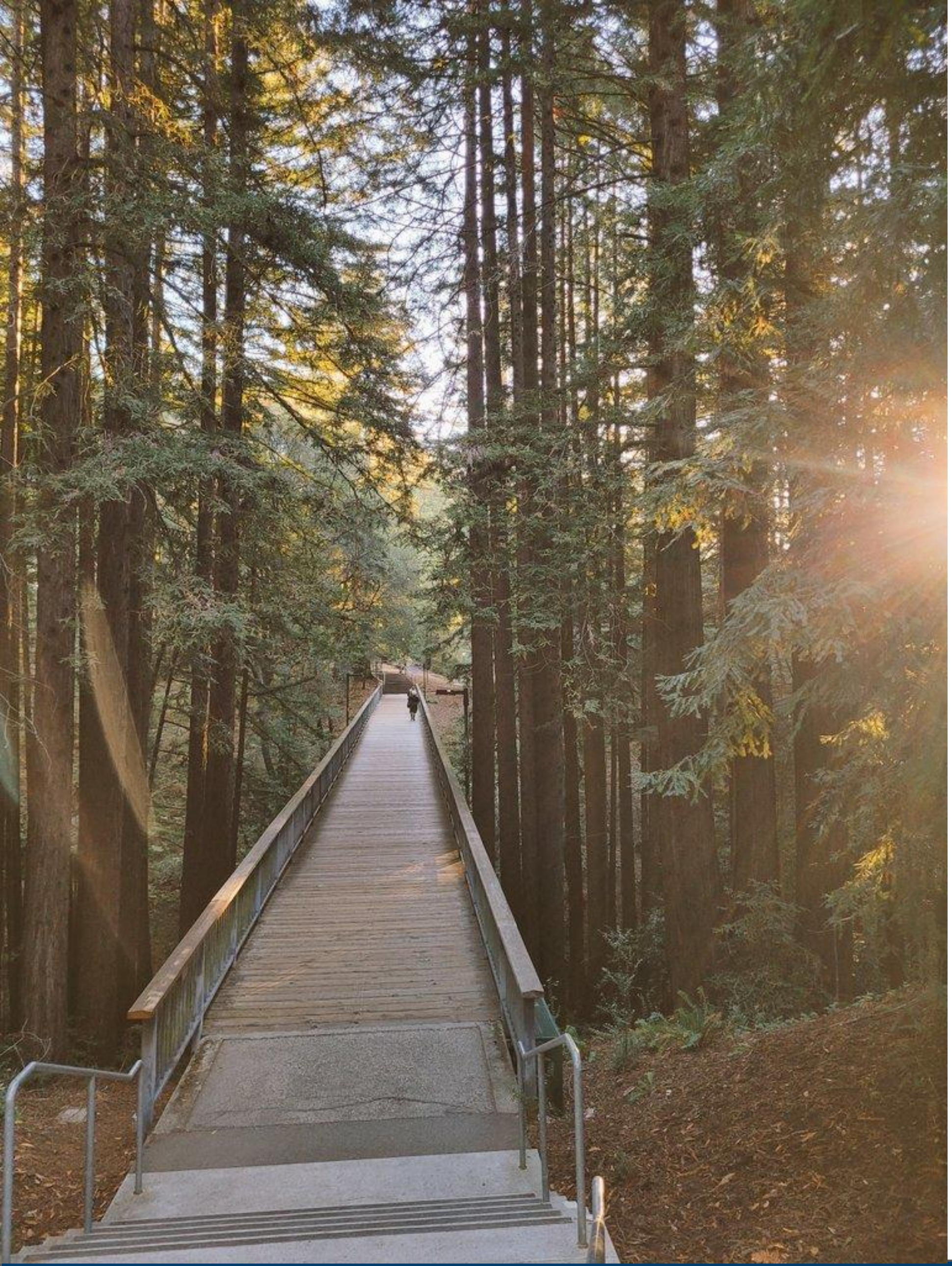


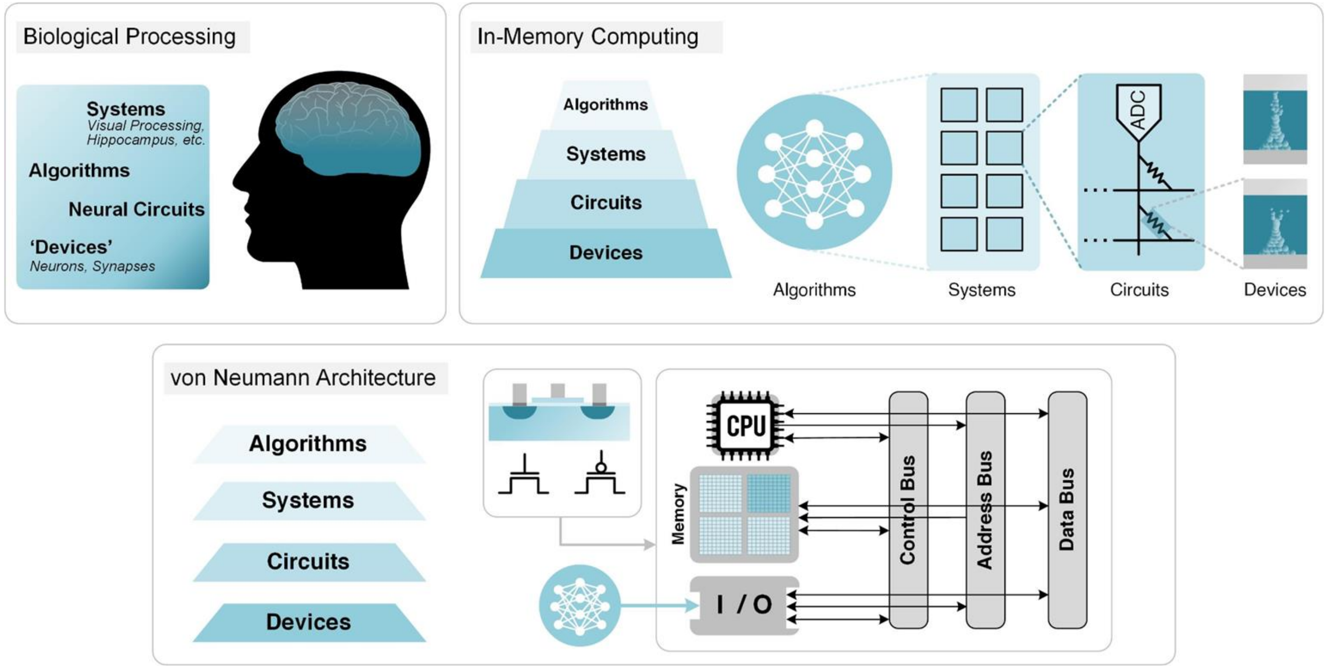
Open-Source Neuromorphic Computing

Jason K. Eshraghian
Assistant Professor, ECE, UC Santa Cruz

13 December 2023

UCSC Neuromorphic Computing Group

Lab Logo Generated by Stable Diffusion



Ph.D. Students

Binh Nguyen
Ruijie Zhu
Juan Lu
Coen Arrow
Assel Kembay

Power bill of training AI in data centers: **>$1 million**

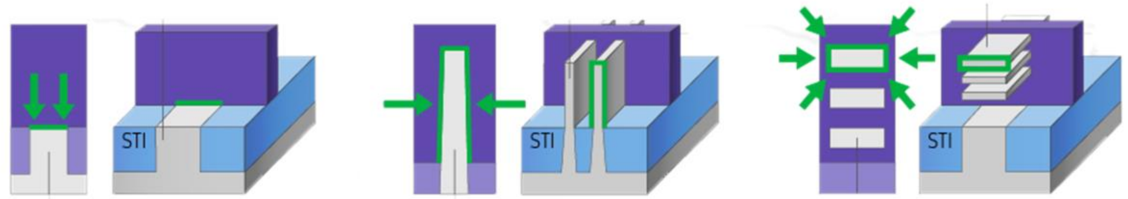
Power budget of brains: **~12 watts**

At the UCSC NCG, our goal is to bridge this gap by applying principles from neuroscience to build better technology.

Undergraduate Students

Farhad Modaresi
Sreyes Venkatesh
Skye Gunasekaran
Hannah Cohen-Sandler
Ruhai Lin
Dylan Louie
Sahil Konjarla

Circuits from 1D→ 2D → 3D → ...4D?



Planar FET
1 gate on channel

FinFET
3 gates on channel

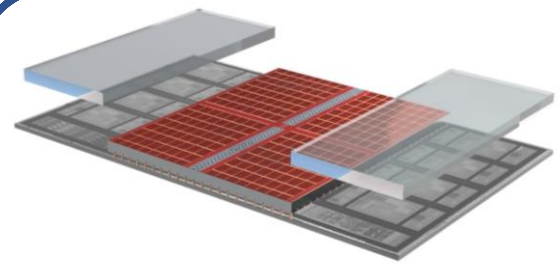
Gate-All-Around
4 gates on channel

Process Technology

Samsung Newsroom, 2019.

Memory Circuits

Western Digital, 2018.



AMD 3D V-Cache
64MB L3 Cache Extension

Microprocessors

AMD, ISSCC 2022.

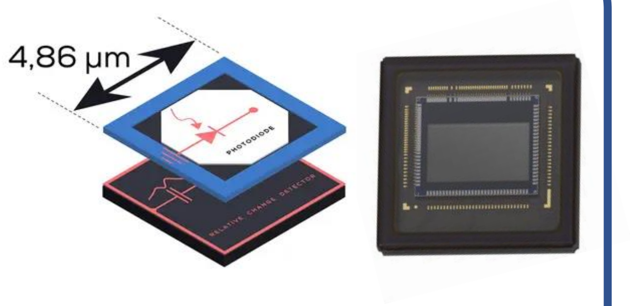


Image Sensors

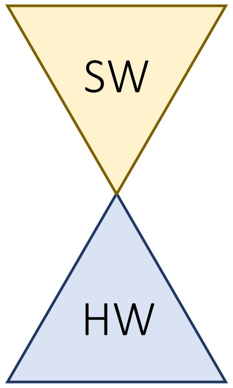
Sony & Prophesee, 2021.

Take the computational principles that underpin the brain and integrate them with silicon

- TinyML
- Compressed Models
- Quantization

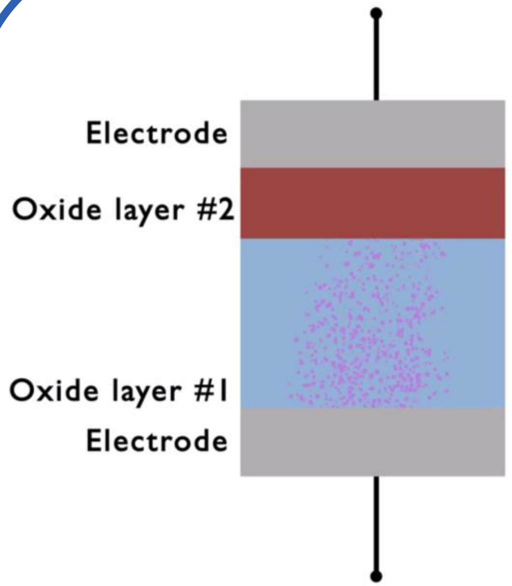
- Spike-based processing
- Temporal coding

- In-memory computing
- Dataflow

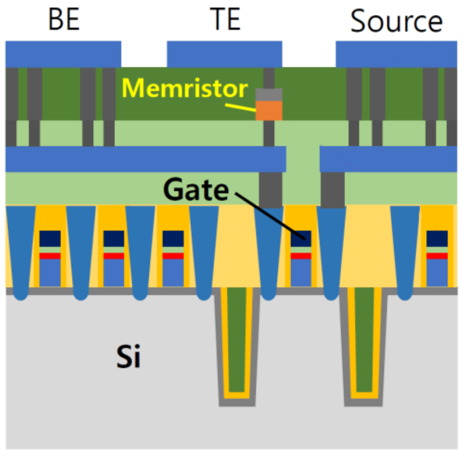


Neuromorphic Computing

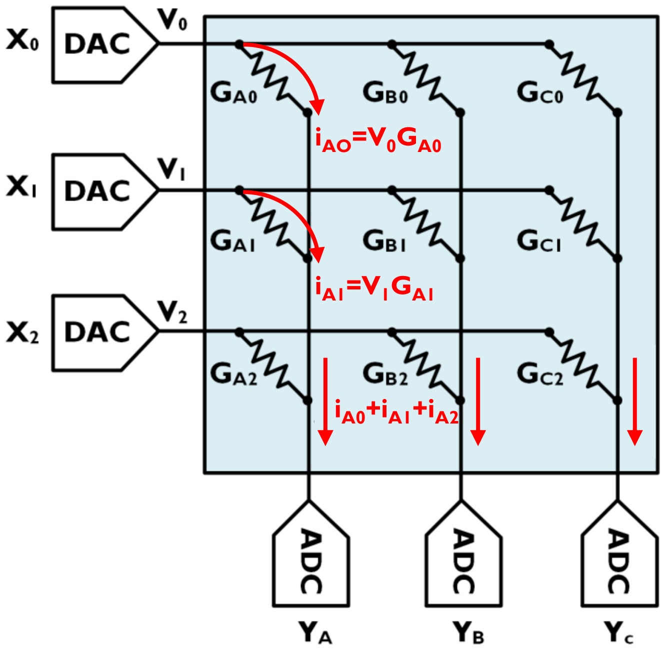
In-Memory Processing Using RRAM Crossbars



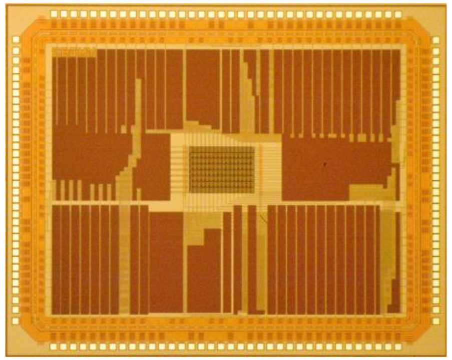
Valence Change Mechanism



BEOL metal-insulator-metal structure



Bit-line current summation



Chip micrograph: 65-nm mixed-signal 16x16 RRAM Crossbar

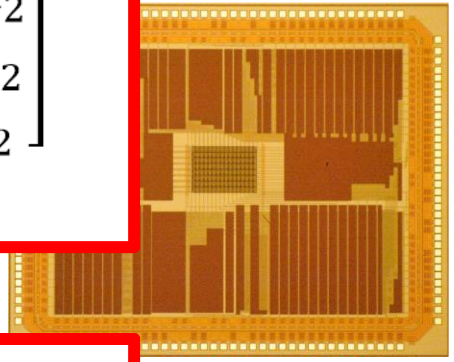
In-Memory Processing Using RRAM Crossbars

Electrode

Oxide layer #2

Oxide layer #1

Electrode



Input Data

Neuron Weights

Output Equations

$$[x_0 \quad x_1 \quad \cdots \quad x_n] * \begin{bmatrix} a_0 & b_0 & c_0 \\ a_1 & b_1 & c_1 \\ \vdots & \vdots & \vdots \\ a_n & b_n & c_n \end{bmatrix} = \begin{bmatrix} y_a = x_0 a_0 + x_1 a_1 + x_2 a_2 \\ y_b = x_0 b_0 + x_1 b_1 + x_2 b_2 \\ y_c = x_0 c_0 + x_1 c_1 + x_2 c_2 \end{bmatrix}$$

Voltage Input

RRAM Conductance

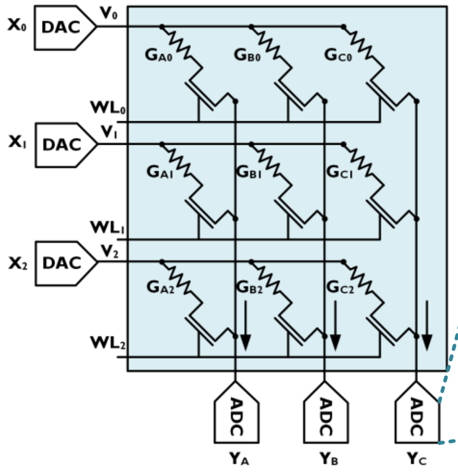
Column Current

$$[\mathrm{V}_0 \quad \mathrm{V}_1 \quad \cdots \quad \mathrm{V}_n] * \begin{bmatrix} \mathrm{G}_{A0} & \mathrm{G}_{B0} & \mathrm{G}_{C0} \\ \mathrm{G}_{A1} & \mathrm{G}_{B1} & \mathrm{G}_{C1} \\ \vdots & \vdots & \vdots \\ \mathrm{G}_{An} & \mathrm{G}_{Bn} & \mathrm{G}_{Cn} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_A = \mathrm{V}_0\mathrm{G}_{A0} + \mathrm{V}_1\mathrm{G}_{A1} + \mathrm{V}_2\mathrm{G}_{A2} \\ \mathrm{I}_B = \mathrm{V}_0\mathrm{G}_{B0} + \mathrm{V}_1\mathrm{G}_{B1} + \mathrm{V}_2\mathrm{G}_{B2} \\ \mathrm{I}_C = \mathrm{V}_0\mathrm{G}_{C0} + \mathrm{V}_1\mathrm{G}_{C1} + \mathrm{V}_2\mathrm{G}_{C2} \end{bmatrix}$$

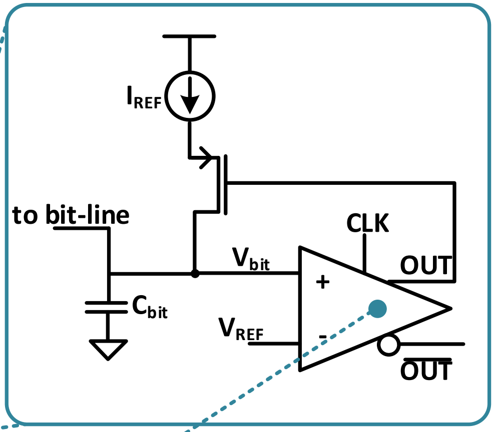
Valer
Me

ip micrograph:
mixed-signal 16x16
RAM Crossbar

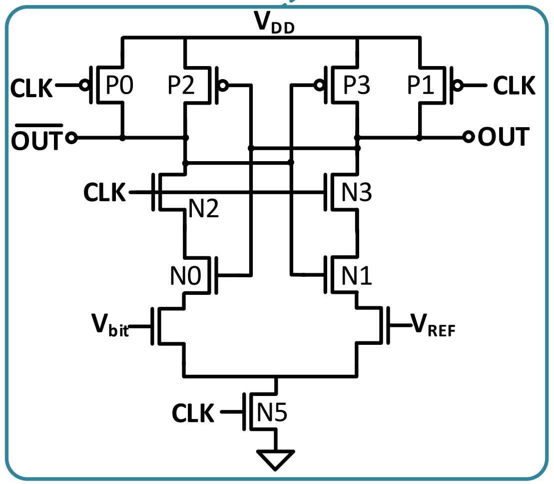
In-Memory Processing Using RRAM Crossbars



1T1R Array

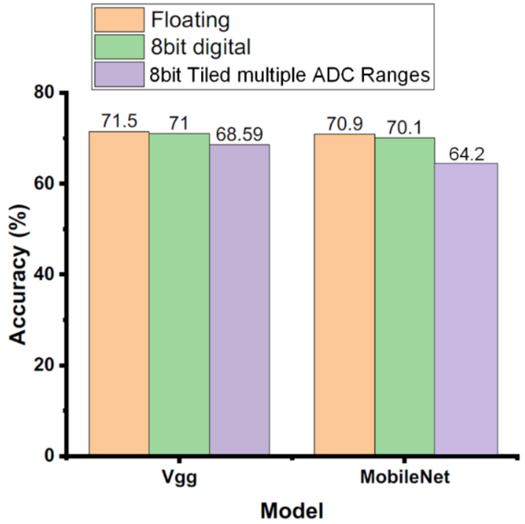


Delta-Sigma ADC
(simplified)



Pre-Charged
Sense-Amplifier

	Our Work	Mythic
Technology	65 nm	40 nm
CLK Freq (MHz)	40 - ADC 100 - RRAM	170
Power (W)	1.89	5
TOPS/W	5.90	4
Quantization	Reconfigurable	8-bit



Top-5 ImageNet classification accuracy

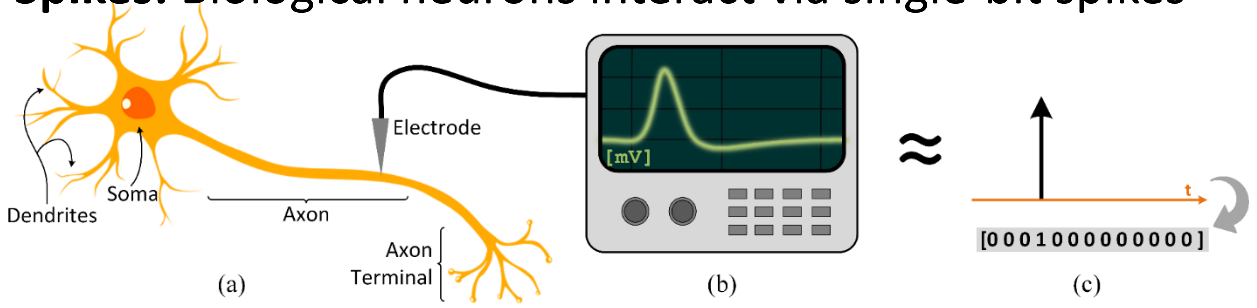
Challenges

- 1.5-3x improvements are insufficient to justify startup costs for a new process
- Scaling challenges: advanced processes are not optimized for analog operation: thermal noise
  - ADCs still <u>dominate latency</u>, and <u>~50% energy consumption</u>
    - Removal of ADCs requires algorithm-level optimization

spike-based processing?

Toward Biological Networks: The Three S's

Spikes: Biological neurons interact via single-bit spikes

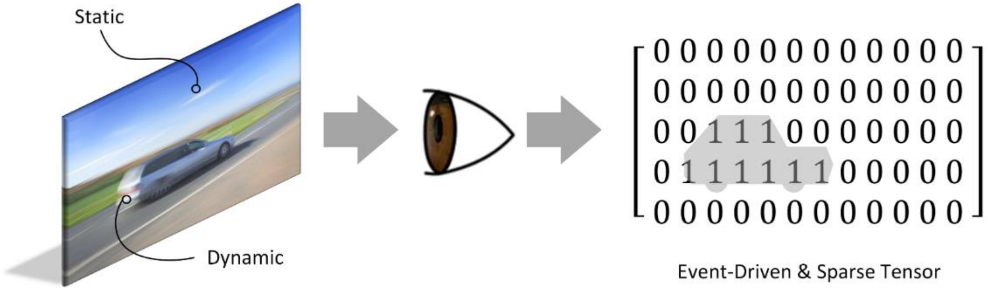


Sparsity: Biological neurons spend most of their time at rest, setting most activations to *zero* at any given time

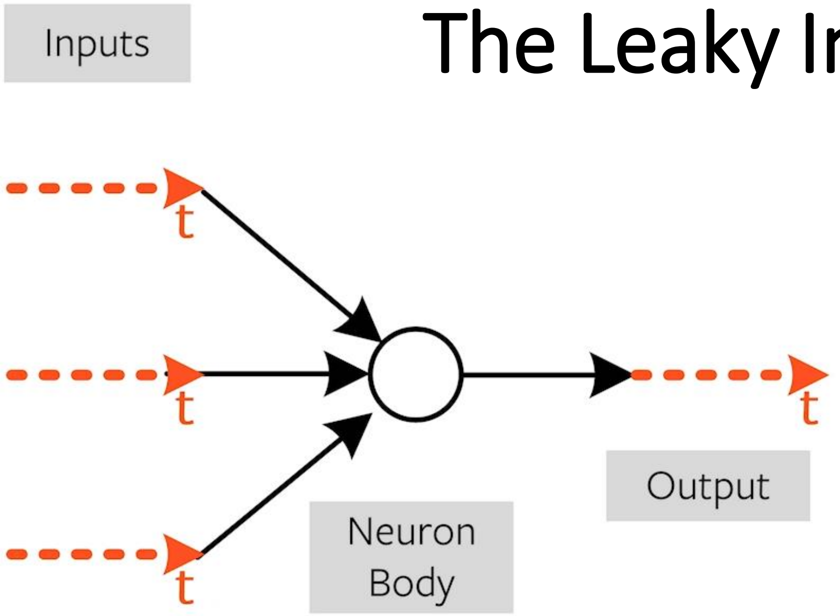
$[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5]$

"7 at position 10; 5 at position 20"

Static Suppression (aka Event-driven Processing): The sensory periphery only processes information when there is *new* information to process

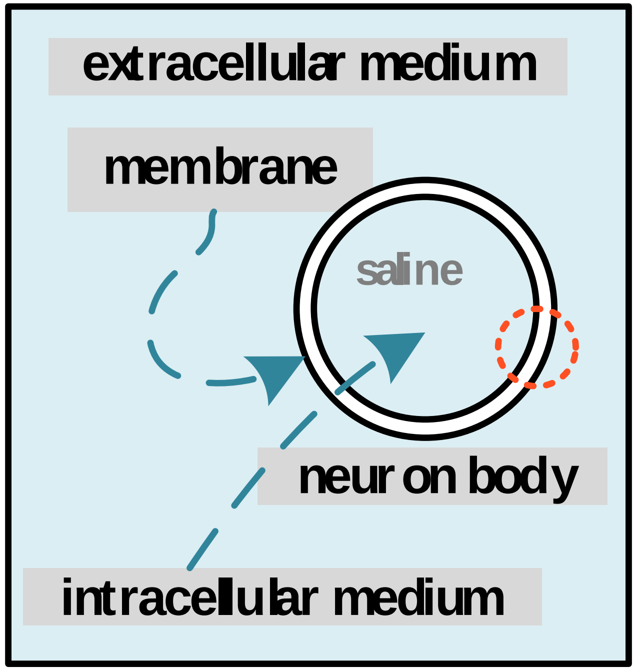


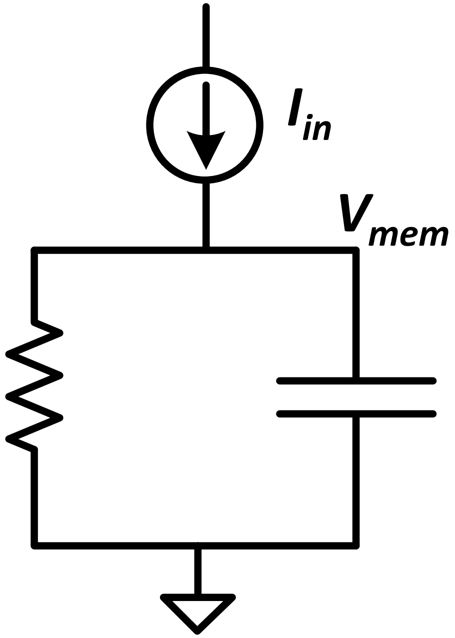
The Leaky Integrate-and-Fire Neuron

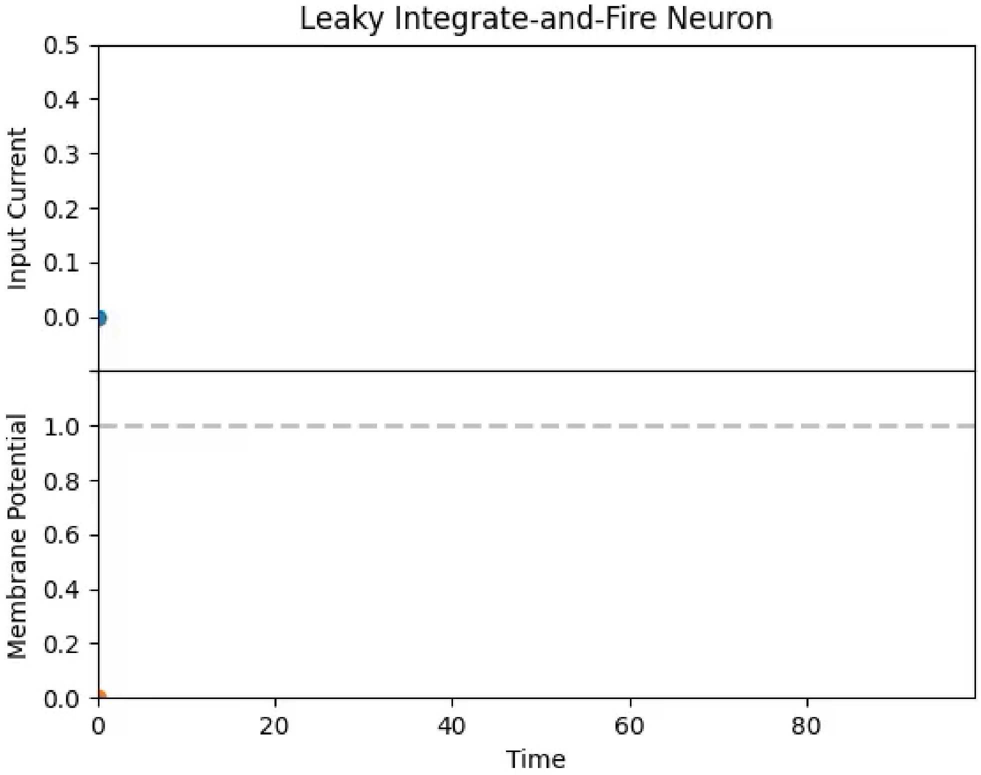


- Bilipid thin-film membrane surrounded by ions: **capacitive**
- Ion-leakage/transfer: **resistive**
- The leaky integrate-and-fire neuron is just a 1st-order low-pass filter, i.e., an **RC circuit**

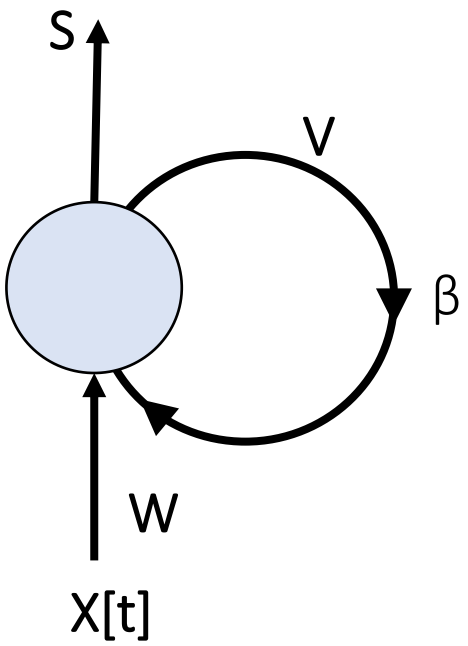
The neuroscientists stole from electrical engineers so it's time to steal back from them

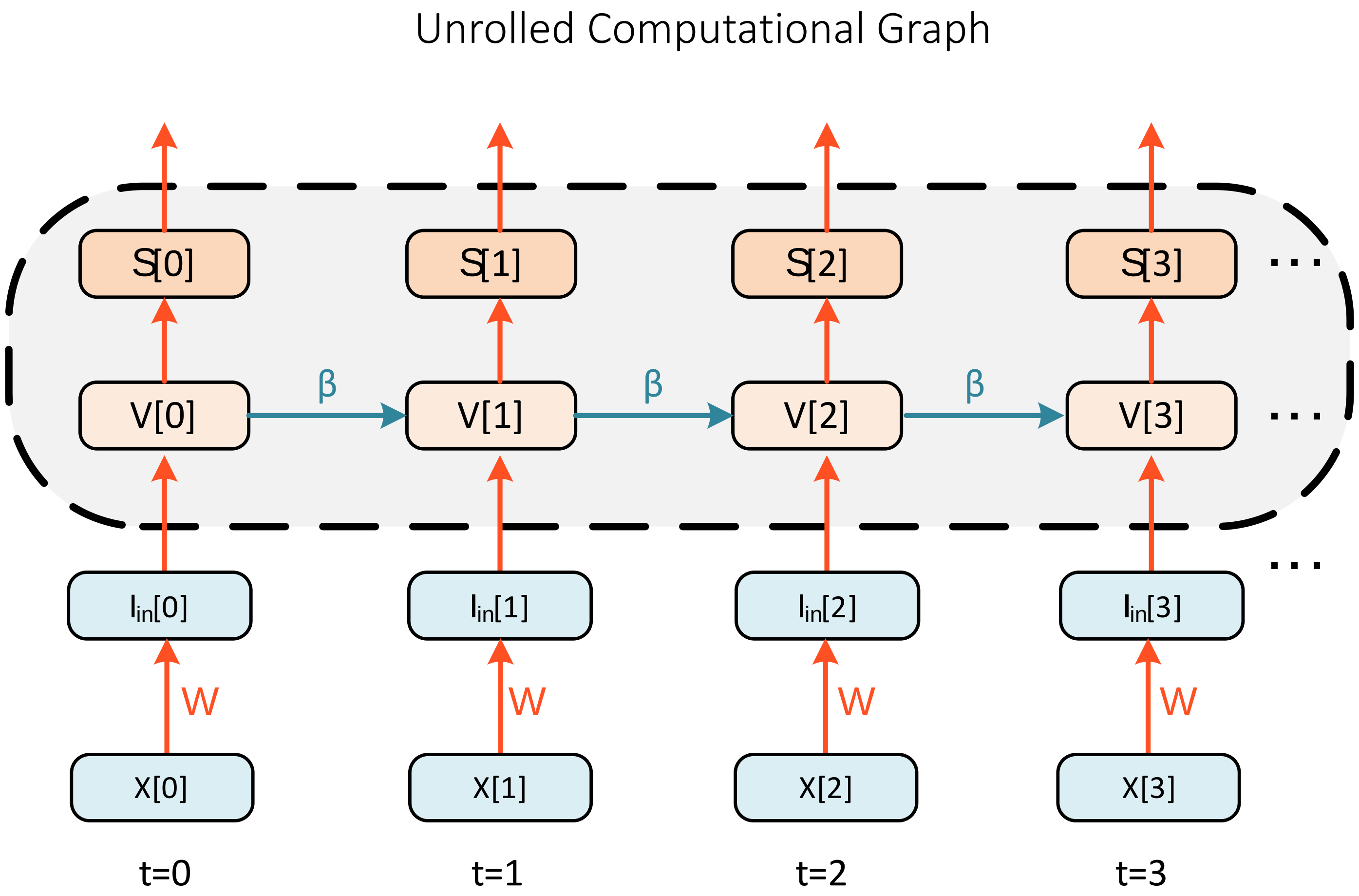






A Recurrent Representation





Spiking Dynamics

$$S\,[t+1] = H(V[t+1] - V_{thr})$$

Membrane Potential Dynamics

$$V\,[t+1] = \beta V[t] + (1-\beta) I_{in}[t]$$

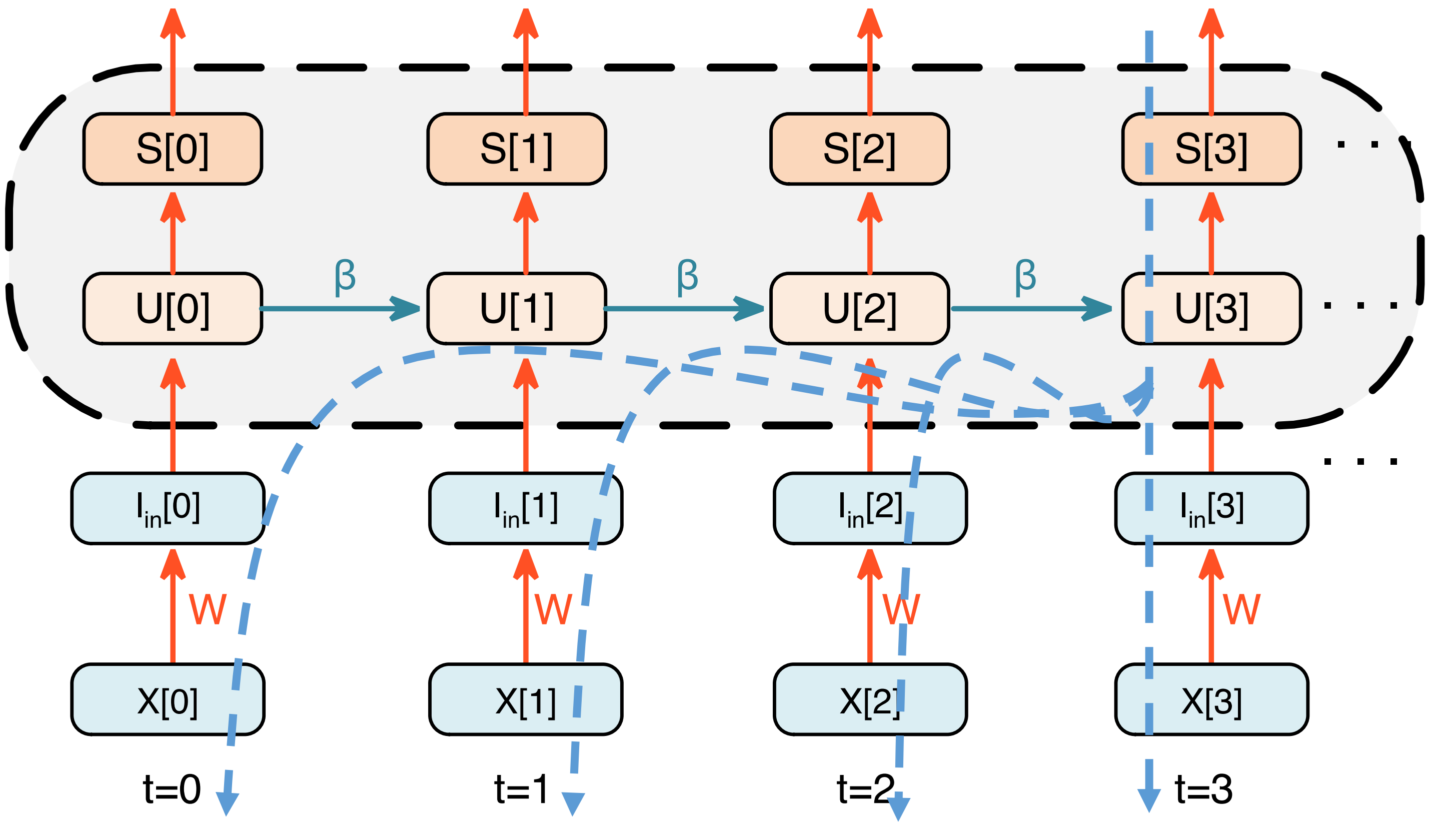
The leaky integrate-and-fire neuron is now compatible with all the tricks and hacks that go with training deep learning models.

Note:
$(1-\beta)$ is removed, and
$I_{in}[t] = WX[t]$

How do we train models that change over time?

Error Backpropagation ...Through Time

Unrolled Computational Graph



Spatial Credit Assignment
How does a loss assign 'blame' to a weight that is spatially far?

Temporal Credit Assignment
All states throughout history must be stored to run the BPTT algorithm
i.e., the entire graph must be stored
Memory complexity: **O(nT)**

snnTorch

Gradient-based Learning with Spiking Neural Networks

Python package for gradient-based optimization of SNNs

real-time online learning

seamless integration with PyTorch

CUDA + IPU accelerated

neuromorphic HW compatible

github.com/jeshraghian/snntorch

Tutorial	Title	Colab Link
Tutorial 1	Spike Encoding with snnTorch	Open in Colab
Tutorial 2	The Leaky Integrate and Fire Neuron	Open in Colab
Tutorial 3	A Feedforward Spiking Neural Network	Open in Colab
Tutorial 4	2nd Order Spiking Neuron Models (Optional)	Open in Colab
Tutorial 5	Training Spiking Neural Networks with snnTorch	Open in Colab
Tutorial 6	Surrogate Gradient Descent in a Convolutional SNN	Open in Colab
Tutorial 7	Neuromorphic Datasets with Tonic + snnTorch	Open in Colab

Advanced Tutorials	Colab Link
Population Coding	Open in Colab
Regression: Part I - Membrane Potential Learning with LIF Neurons	Open in Colab
Regression: Part II - Regression-based Classification with Recurrent LIF Neurons	Open in Colab
Accelerating snnTorch on IPUs	–

Python package for gradient-based optimization of SNNs

real-time online learning

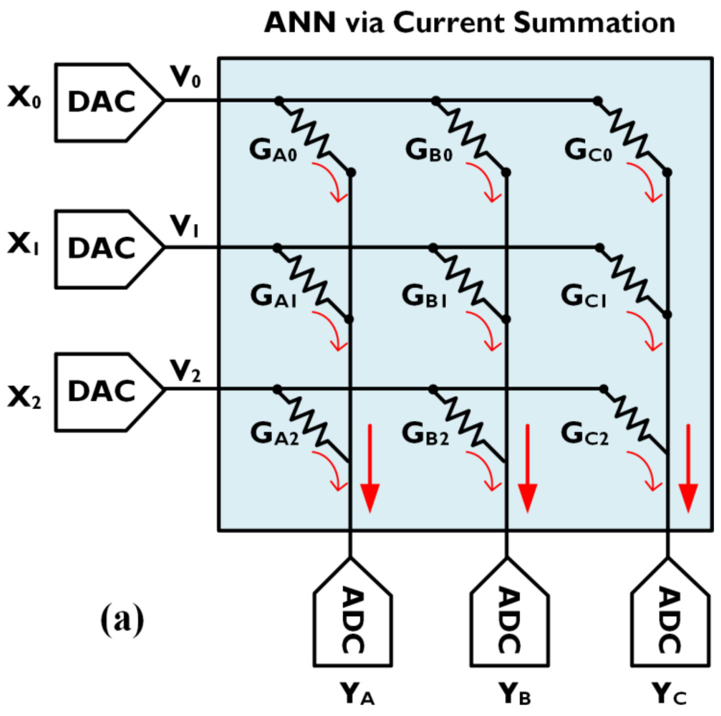
seamless integration with PyTorch

CUDA + IPU accelerated

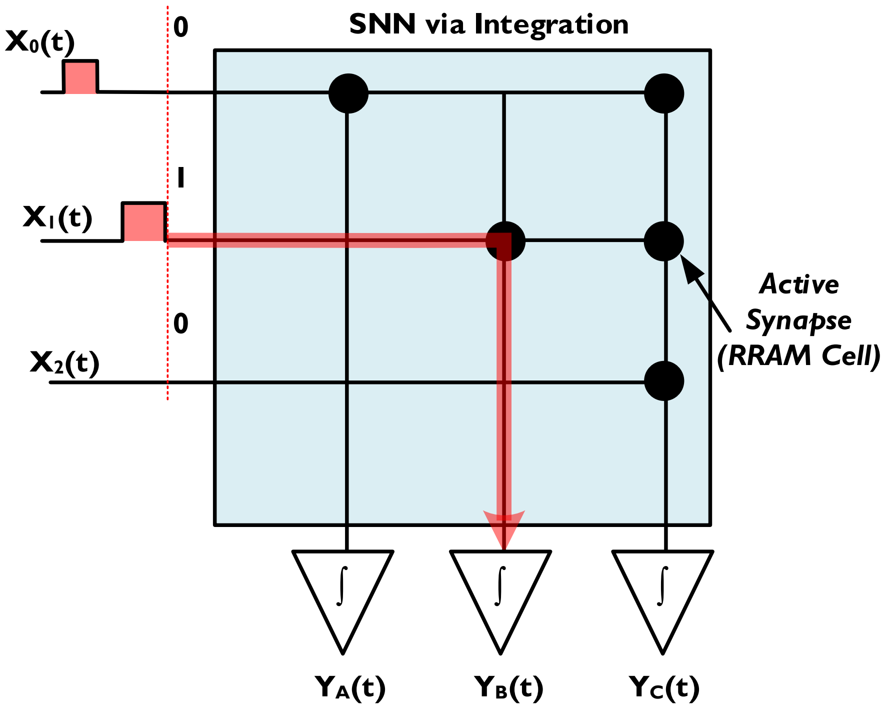
neuromorphic HW compatible

github.com/jeshraghian/snntorch

RRAM-based SNN Processing



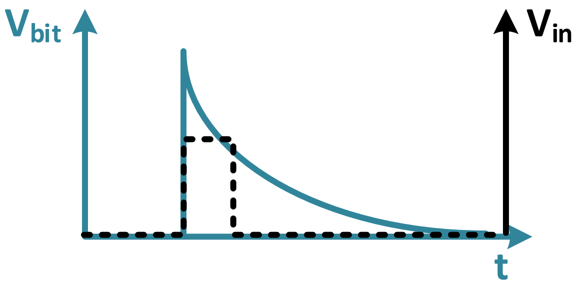
Communicating noise-prone analog signals down long lengths of wire: noise-prone



Communicating noise-robust digital signals down long lengths of wire: noise-tolerant

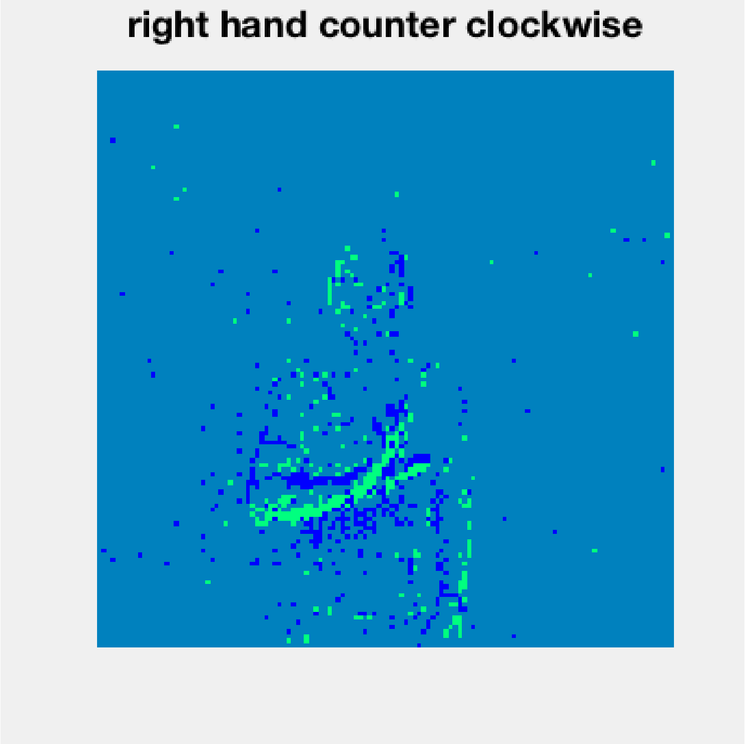
An RRAM Approach to Spiking Neuron Dynamics

- Charge-based integration on the bit-line capacitance
- State decay using heterogeneous RRAM time constants

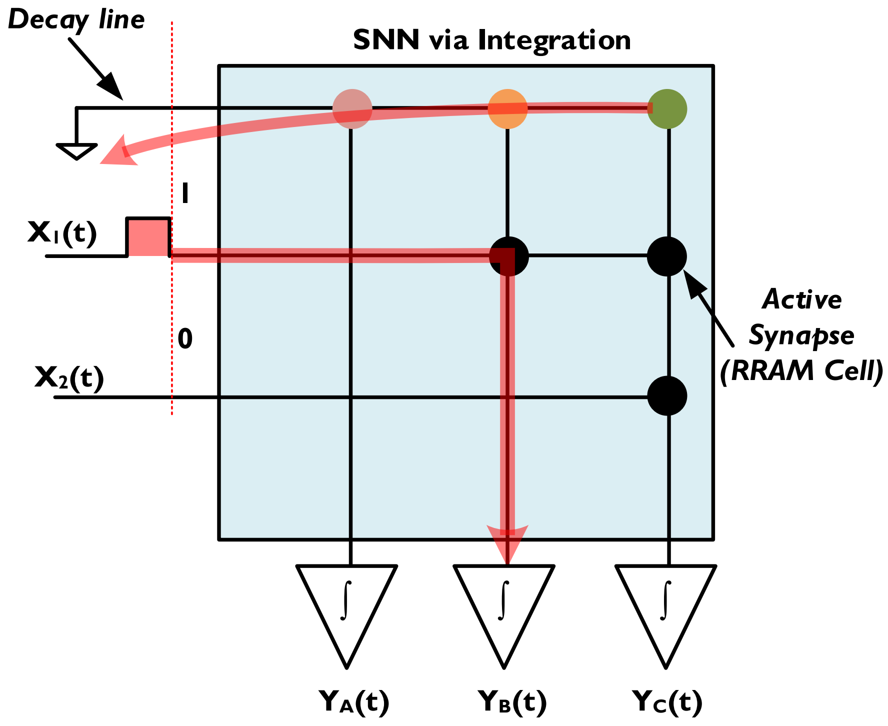


J.K. Eshraghian et al., "Memristor-based Binarized SNNs" IEEE Nanotech. Mag. 2022

RRAM-based SNN Processing



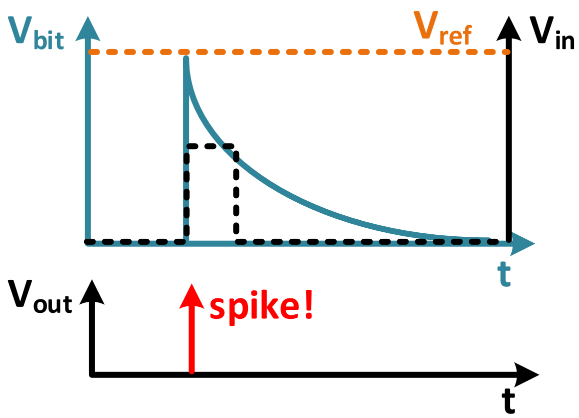
Communicating noise-prone analog signals down long lengths of wire: noise-prone



Communicating noise-robust digital signals down long lengths of wire: noise-tolerant

An RRAM Approach to Spiking Neuron Dynamics

- Charge-based integration on the bit-line capacitance
- State decay using heterogeneous RRAM time constants
- Pre-charged sense amplifier for spiking dynamics (replace ADC)



J.K. Eshraghian et al., "Memristor-based Binarized SNNs" IEEE Nanotech. Mag. 2022

RRAM-based SNN Processing

Classical pain points of IMC RRAM are largely addressed by spikes

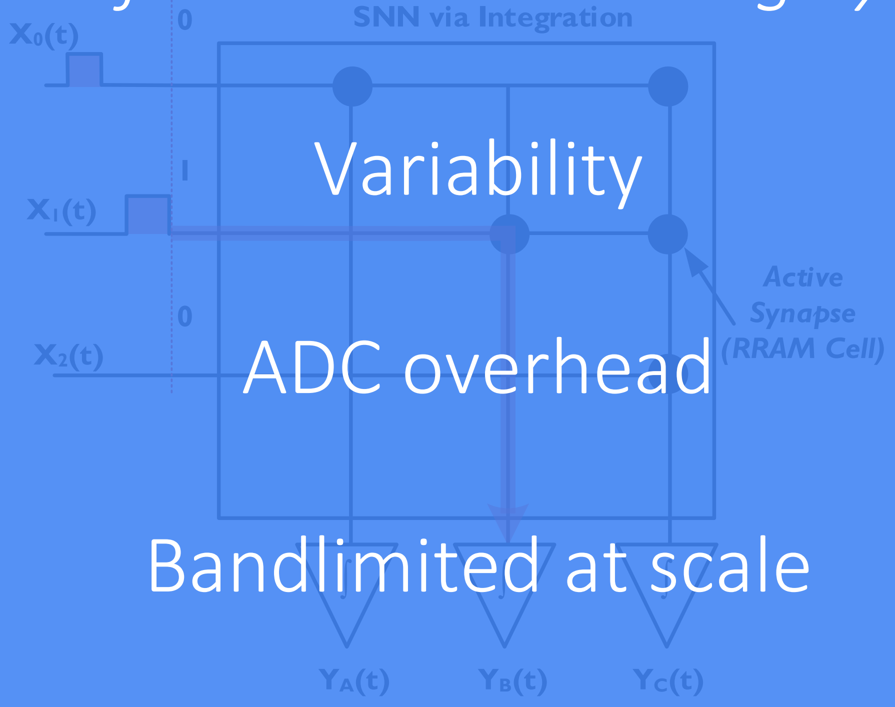
Variability

ADC overhead

Bandlimited at scale

Endurance

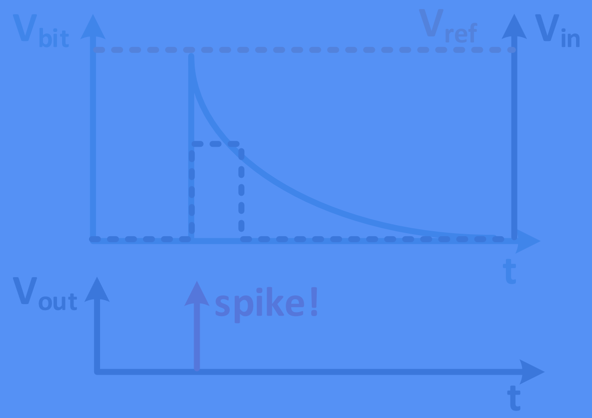
Communicating noise-prone analog signals down long lengths of wire: noise-prone



Communicating noise-robust digital signals down long lengths of wire: noise-tolerant

An RRAM Approach to Spiking Neuron Dynamics

- Charge-based integration on the bit-line capacitance
- State decay using heterogeneous RRAM time constants
- Pre-charged sense amplifier for spiking dynamics (replace ADC)



J.K. Eshraghian et al., "Memristor-based Binarized SNNs" IEEE Nanotech. Mag. 2022

RRAM-based SNN Processing

Classical pain points of IMC RRAM are largely addressed by spikes

~~Variability~~	Heterogeneity
~~ADC overhead~~	Single-bit Spikes
~~Bandlimited at scale~~	Sparse data movement
~~Endurance~~	Weight updates only occur at spike times

Communicating noise-prone analog signals down long lengths of wire: noise-prone

Communicating noise-robust digital signals down long lengths of wire: noise-tolerant

An RRAM Approach to Spiking Neuron Dynamics

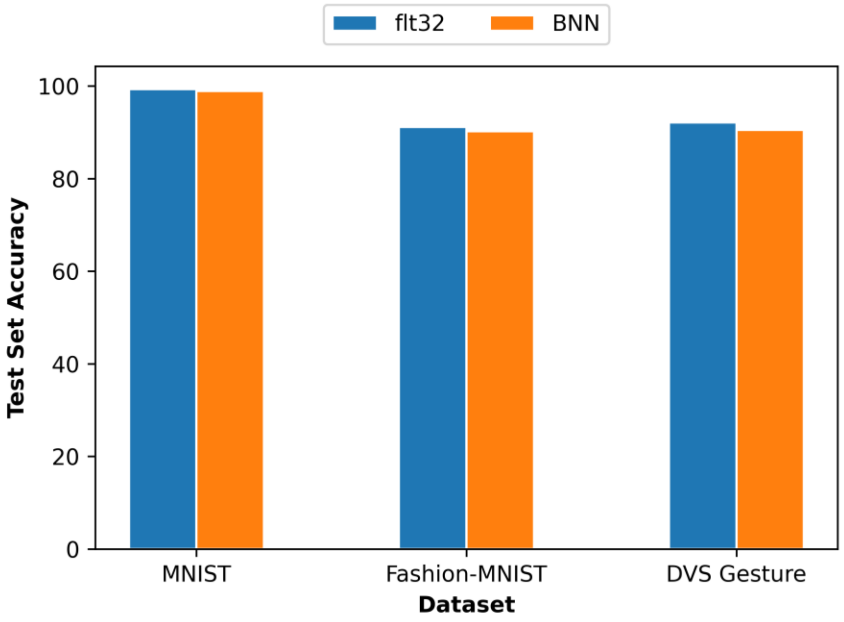
- Charge-based integration on the bitline capacitance
- State decay using heterogeneous RRAM time constants
- Pre-charge sense amplifier for spiking dynamics (replace ADC)

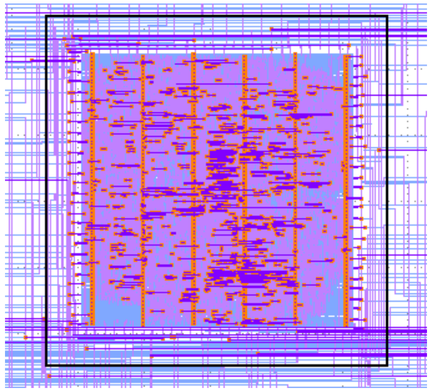
J.K. Eshraghian et al., "Memristor-based Binarized SNNs" IEEE Nanotech. Mag. 2022

SkyWater 130 Neuromorphic Accelerator

Open-Source Neuromorphic IP

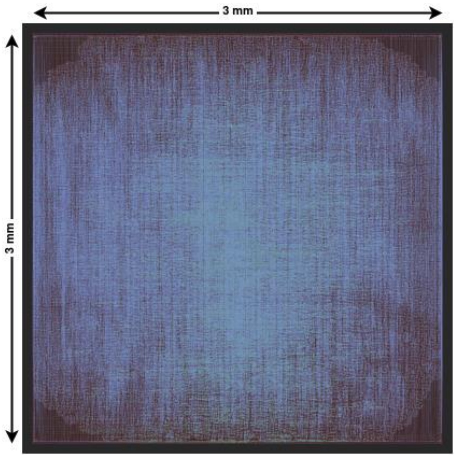
- Google-sponsored tape-outs using SkyWater 130nm process
- Deep learning success was in part due to open-source; let’s port this hype train over to silicon
- 2x successful tape-outs + 1x tape-in…





MPW6 BSNN Streaming Accelerator

- Online event streaming accelerator
- HD Event Cameras: ~10 MHz spike rate
- Die area: 0.09 mm^2
- Average Power: 11 nW
- GPIO: 50 MHz peak, Core: 210 MHz



MPW8 SNN Accelerator
Lead Designer: Farhad Modaresi

- 154 kB of dual-power 2 kB SRAM macros
- 9 mm^2 core, 21.89 mm^2 memory
- SRAM: 40 MHz, Core: 20 MHz
- Custom precision (up to 8-bits)

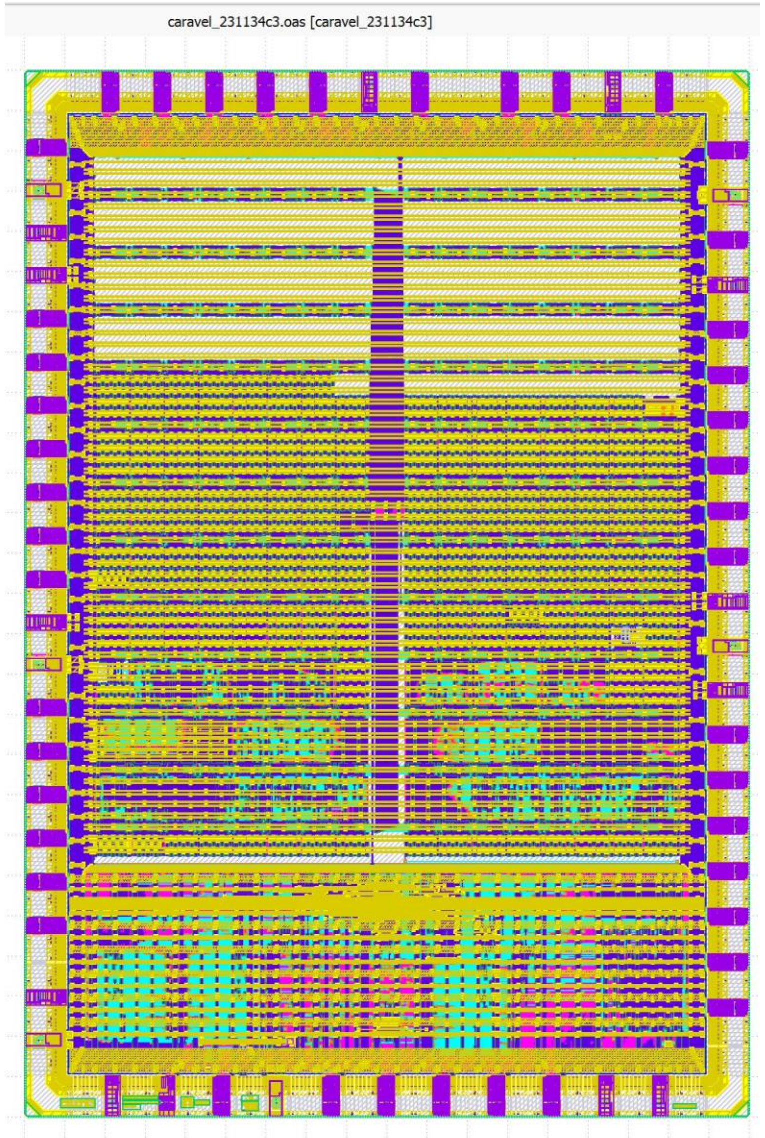
F. Modaresi, M. Guthaus, J. K. Eshraghian, ISCAS 2023.

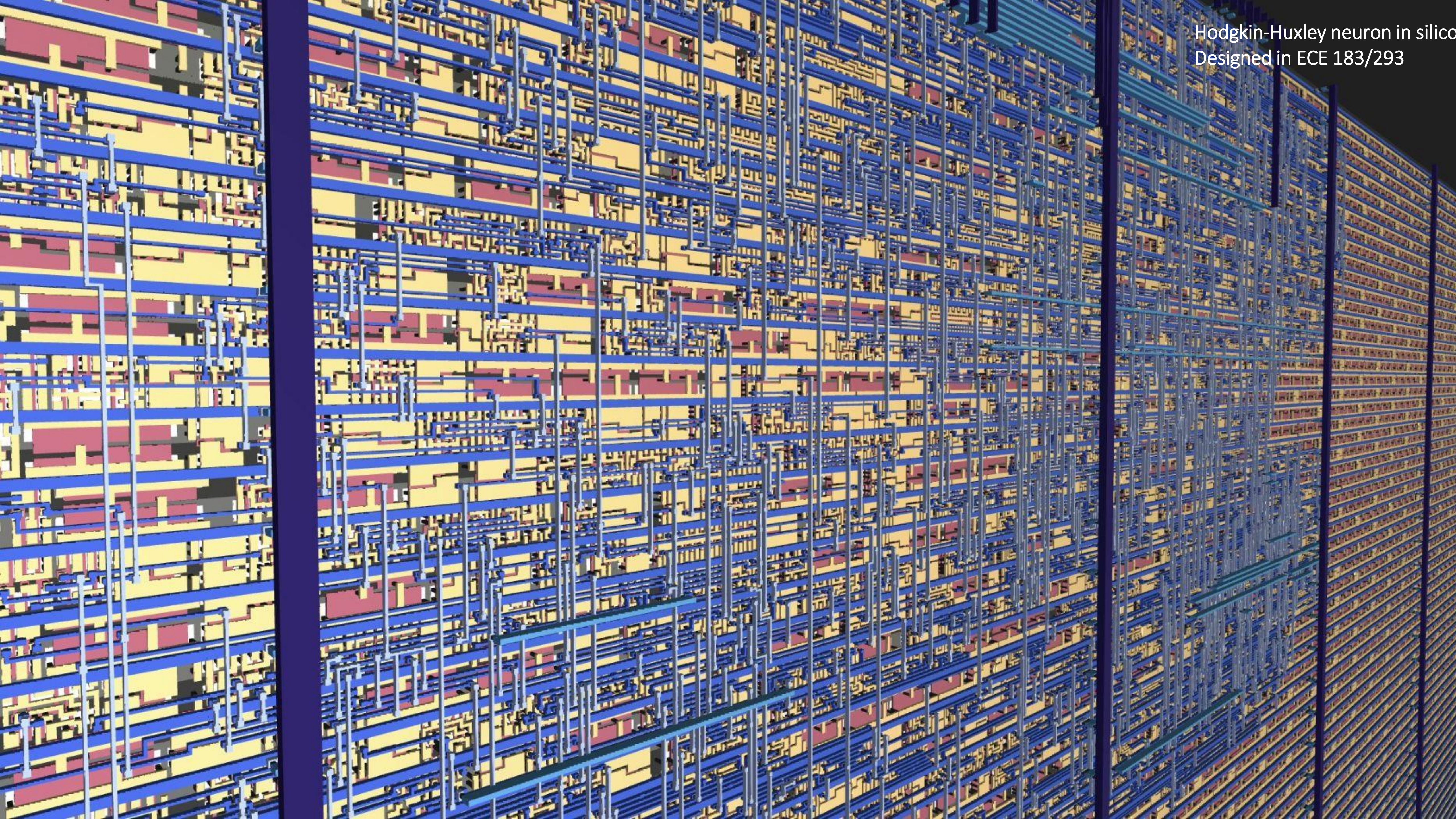
The next generation of neuromorphic engineers

31 first-time chip designers at UCSC taking the "Brain-Inspired Machine Learning" class defined their own design and submitted a chip.

It is currently being manufactured and will ship back in April 2023.

www.tinytapeout.com





Hodgkin-Huxley neuron in silico
Designed in ECE 183/293

How many of you have converted your garage to a semiconductor manufacturing plant?

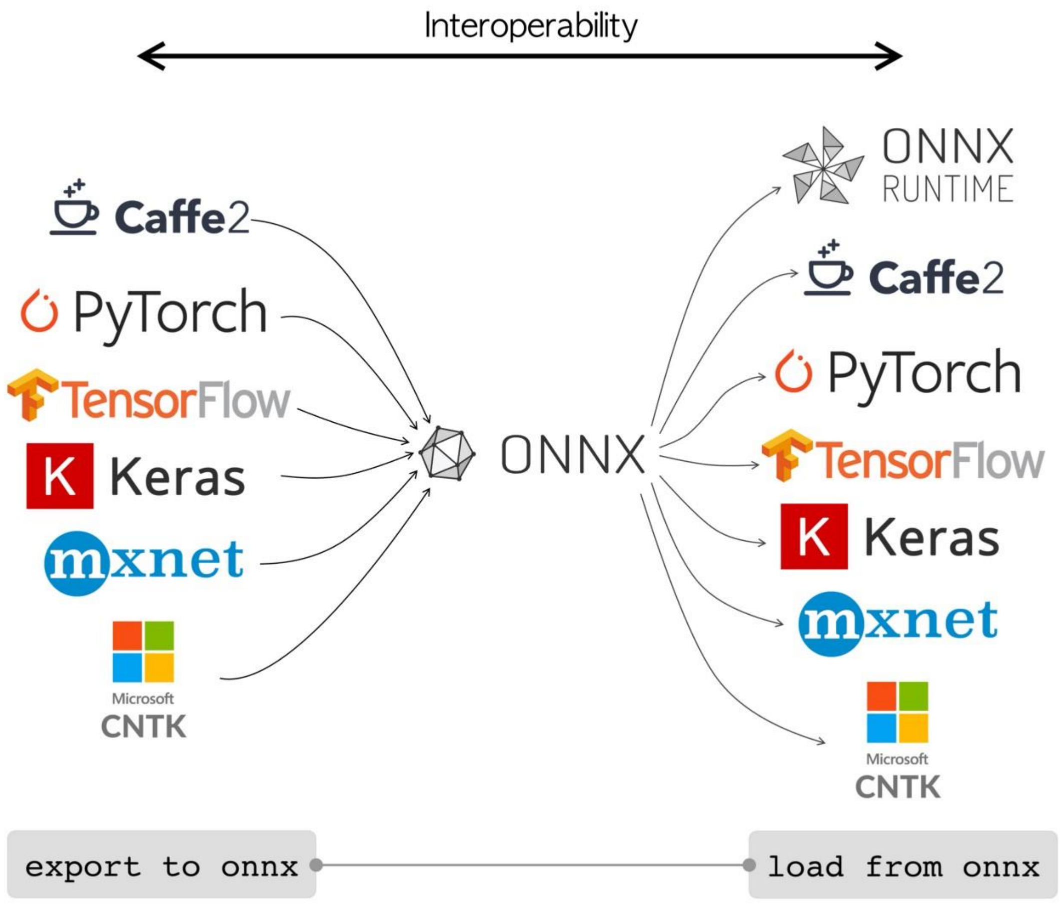
I didn’t think so.

How can we port dynamical, time-varying neural networks across different systems?

Solutions already exist for vanilla deep learning.

Nothing but pain exists for spiking neural networks.

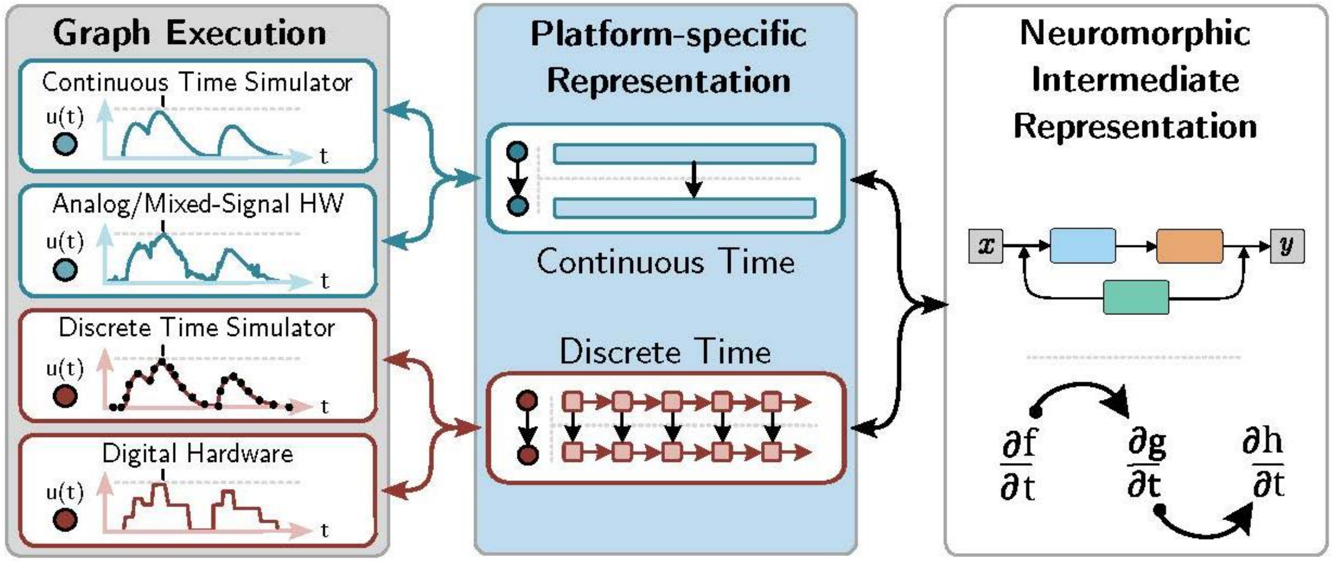
… until now.



NIR

Neuromorphic Intermediate Representation

A Unified Instruction Set for Interoperable Brain-Inspired Computing



- Provides a common standard to enable continuous-time, dynamical neural networks to interface with each other

- Allows the broader community to tap into commercial & exotic hardware with their software of choice – e.g., Intel Loihi

```
>> model.to_nir()
```

We wrote 1000s of lines of code so you only have to write 1.

Pederesen, Abreu, Eshraghian et al., arXiv, Nov. 2023.

Neuromorphic Intermediate Representation

A Unified Instruction Set for Interoperable Brain-Inspired Computing

snnTorch

Server-scale neuromorphic HW:
Loihi, Intel Labs

Edge neuromorphic hardware:
SynSense, BrainChip

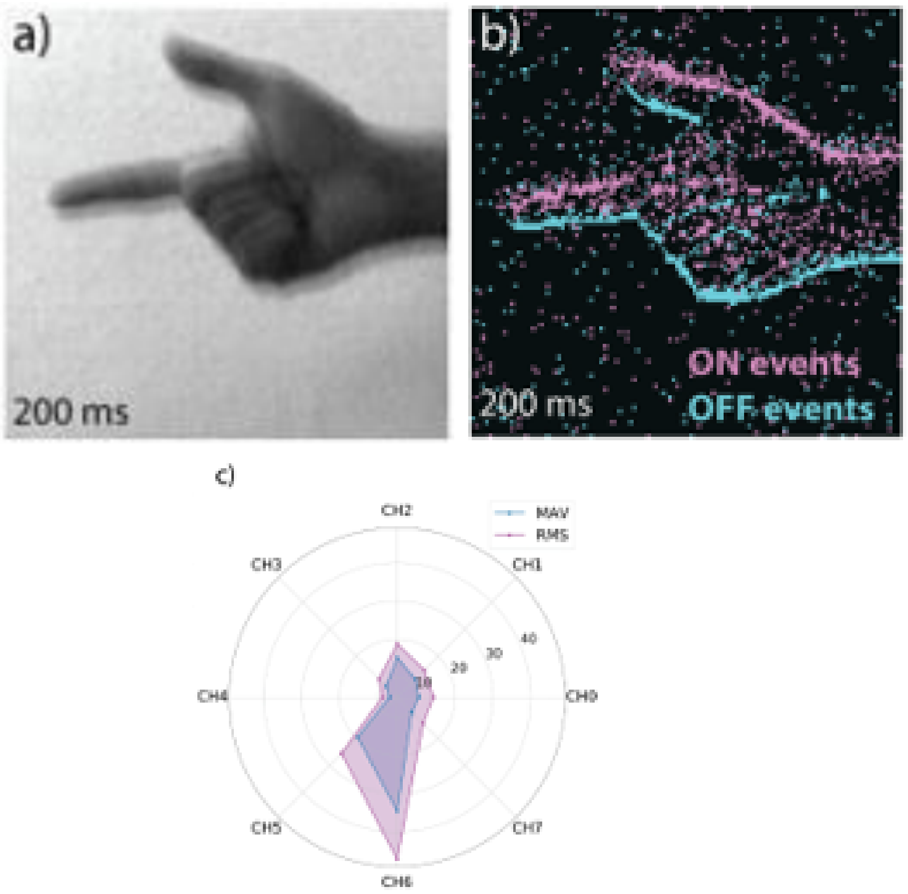
- Provides a common standard to enable continuous-time, dynamical neural networks to interface with each other
- Allows the broader community to tap into commercial & exotic hardware with their software of choice – e.g., Intel Loihi

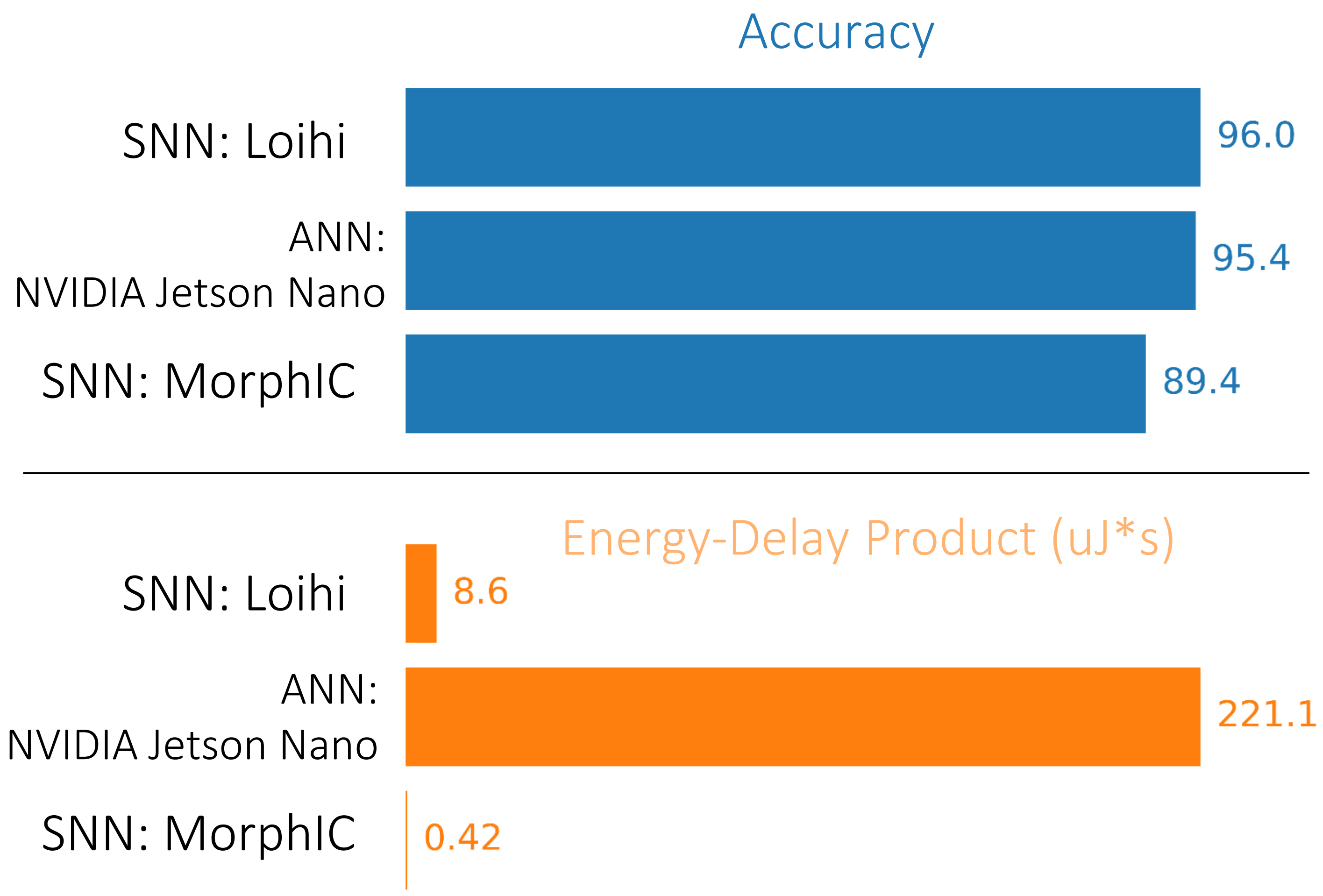
```
>> model.to_nir()
```

We wrote 1000s of lines of code so you only have to write 1.

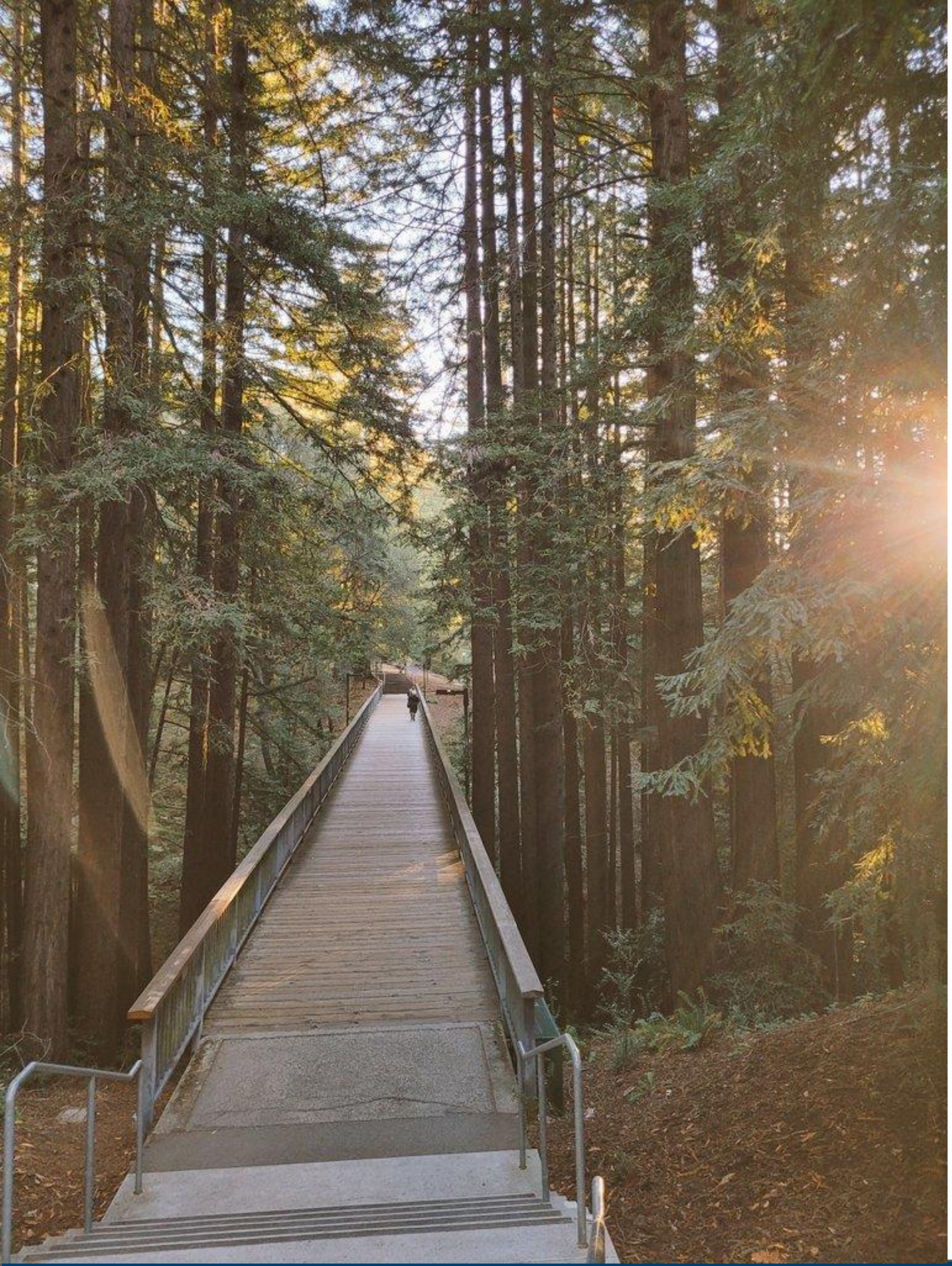
Pedersen, Abreu, Eshraghian et al., arXiv, Nov. 2023.

SNN Evaluation – Multimodal Data





M. Rahimi-Azghadi, J.K. Eshraghian et al., IEEE Trans. Biomedical CAS, Dec. 2020



Open-Source Neuromorphic Computing

Jason K. Eshraghian
Assistant Professor, ECE, UC Santa Cruz

13 December 2023

UC SANTA CRUZ | Baskin Engineering