# Energy-Efficient Hardware for AI: Subthreshold Silicon Neurons and Gain-Cell Analog In-Memory Computing

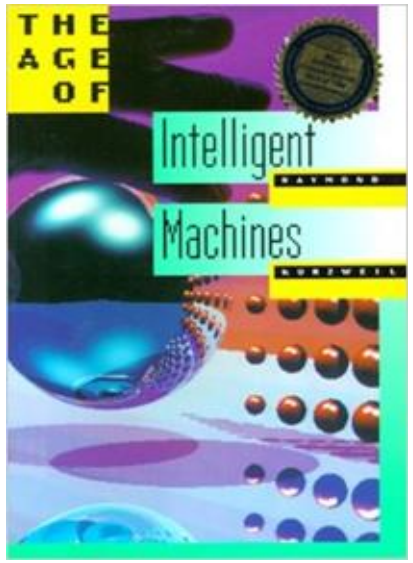

Yiran Chen

*Department of Electrical and Computer Engineering*
*Duke University Center for Computational Evolutionary Intelligence (CEI)*
*NSF IUCRC For Alternative Sustainable and Intelligence Computing (ASIC)*
*NSF AI Institute for Edge Computing Leveraging Next-generation Networks (Athena)*

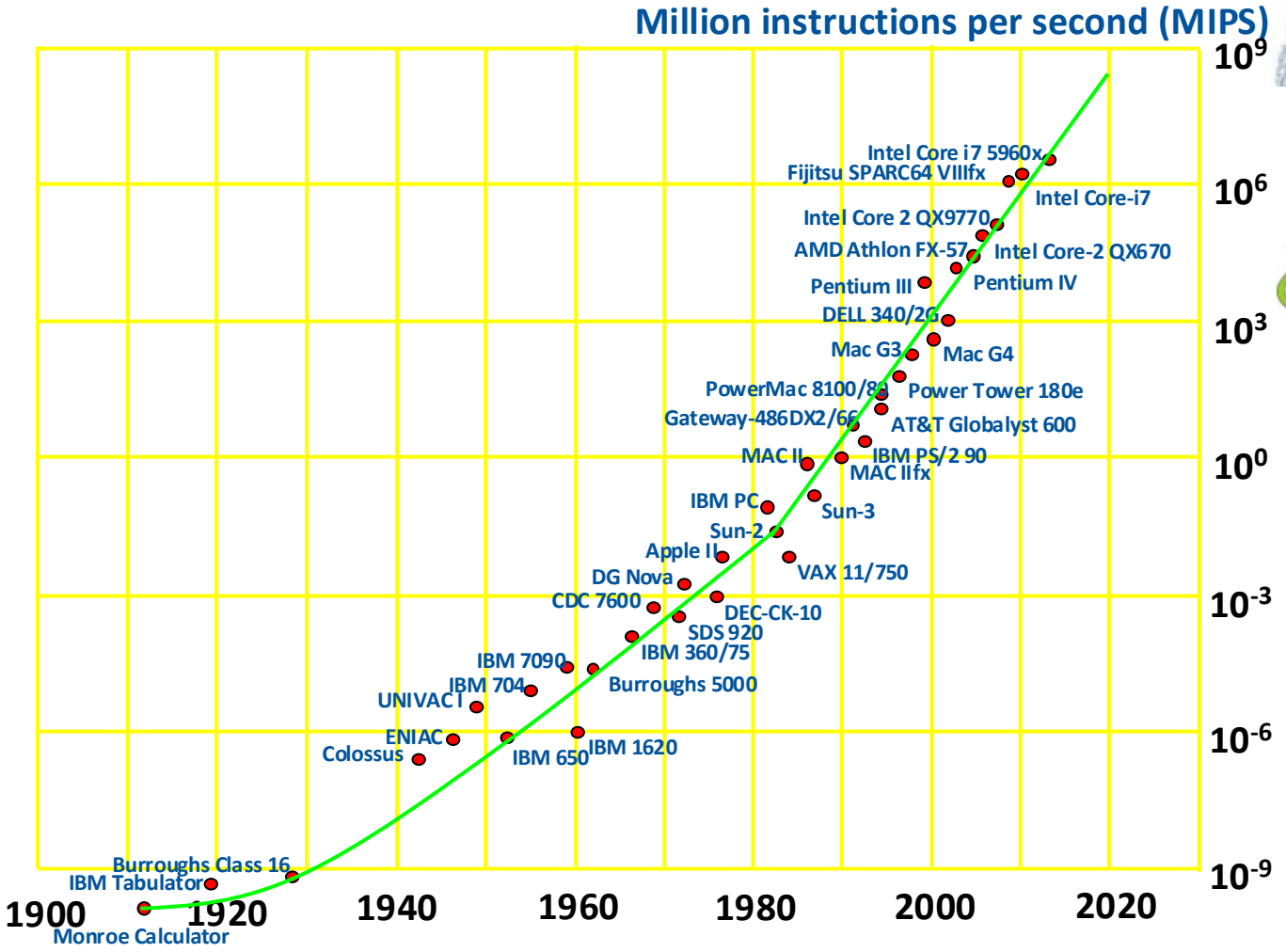
Duke

# Evolution of Computing Power

Dr. Raymond Kurzweil




The Age of Intelligent Machines
MIT Press, 1990 ISBN 0-262-11121-7




**Human**
100G Neurons

**Monkey**
3G Neurons

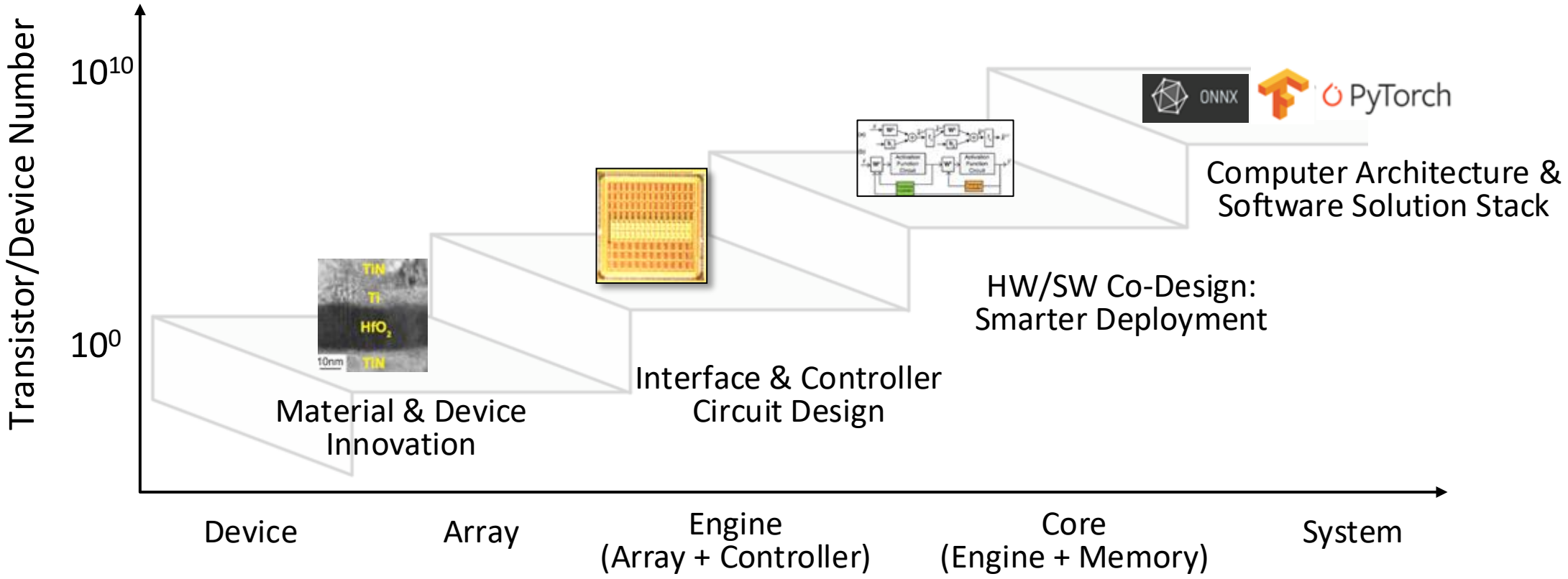
**Lizard**
2M Neurons

**Worm**
300 Neurons

**Bacteria**
1 “Neuron”

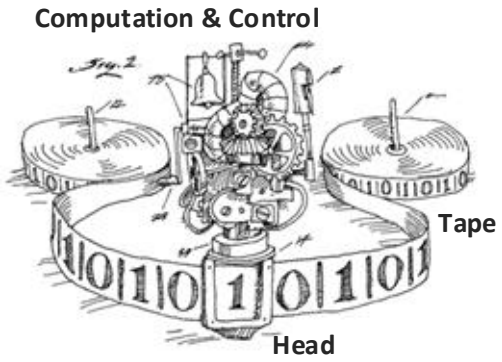
Duke

# AI Systems – An Interdisciplinary Field

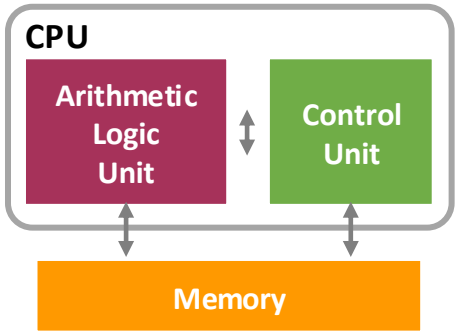



Duke

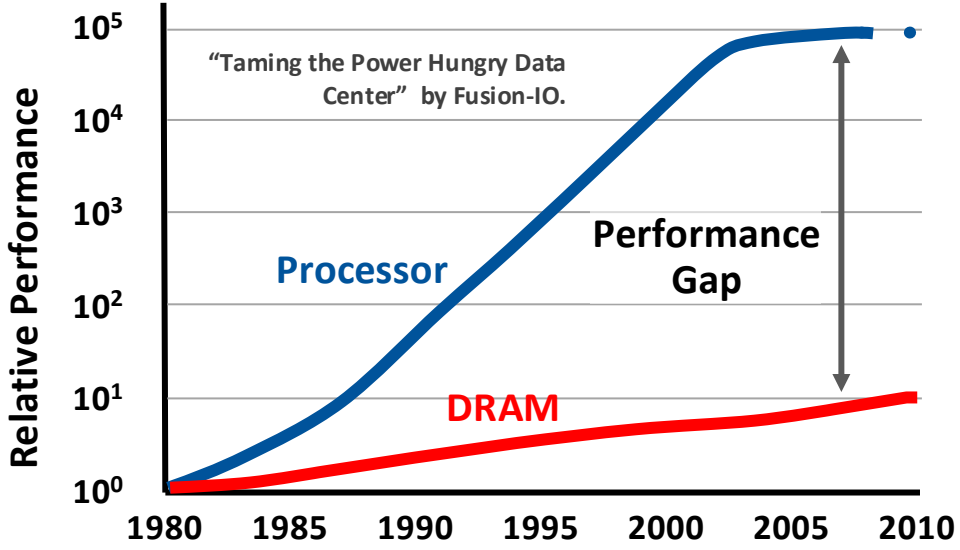
# Von Neumann Bottleneck

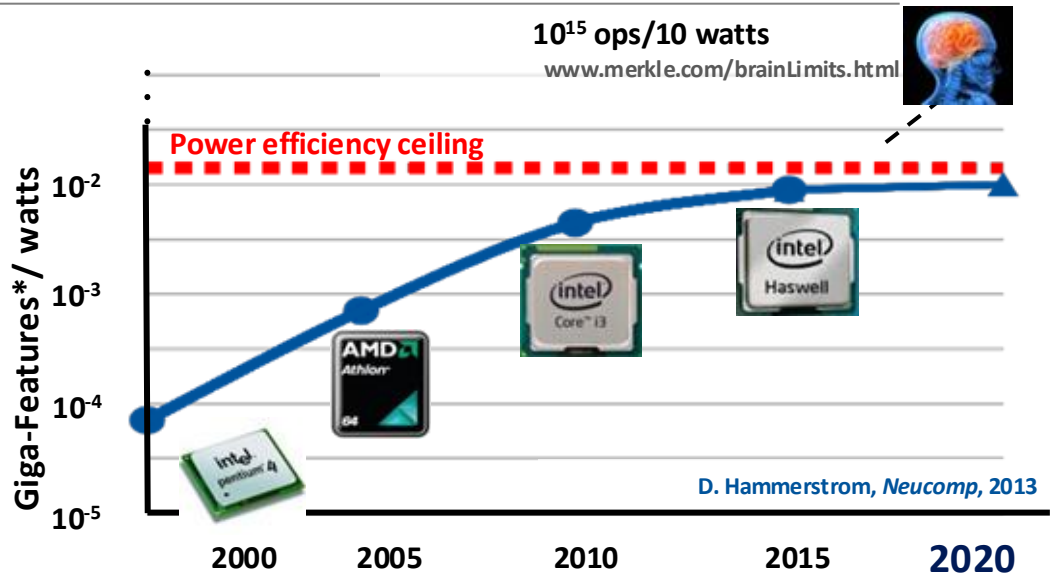



Turing Machine




von Neumann Arch.

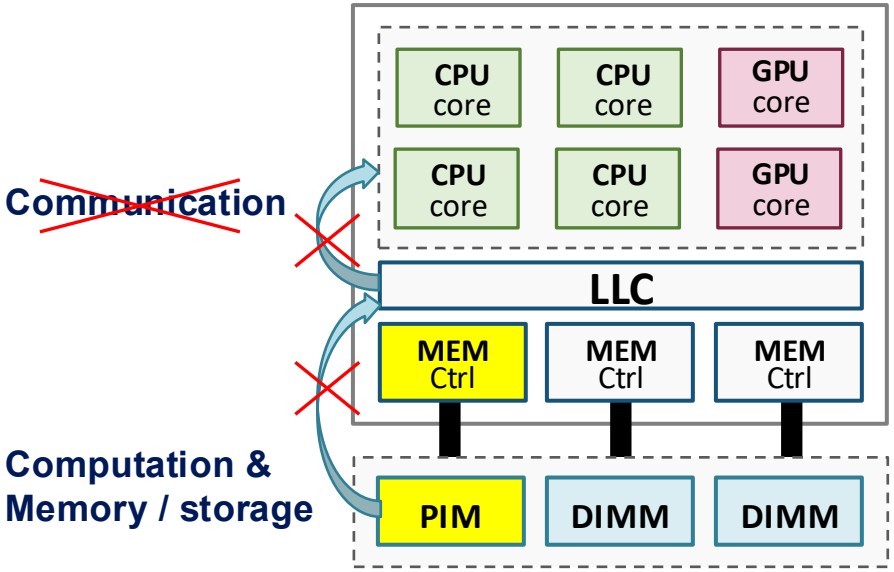



- ✓ Stalled single-thread performance
- ✓ Limited data throughput
- ✓ Constrained power efficiency

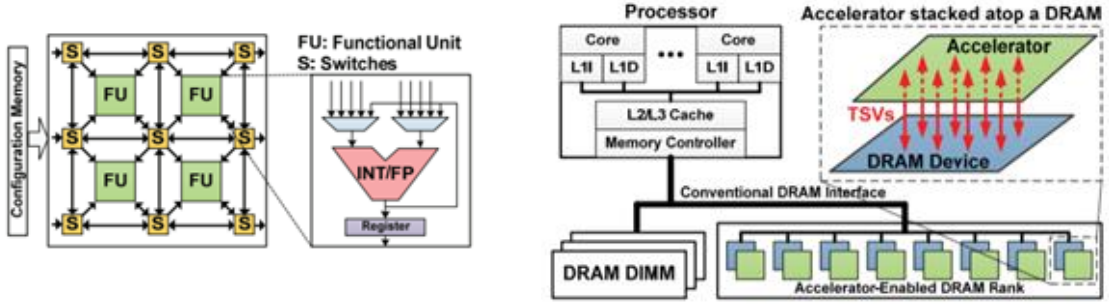
Duke

# Memory Wall & Memory Centric Design

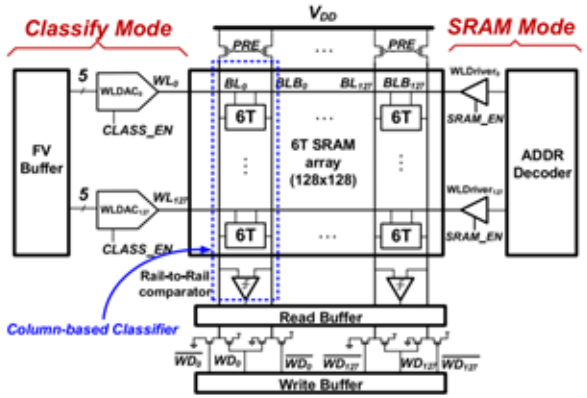



**Energy consumption for 45 nm CMOS process [Han+, ISCA'16].**

| Operation | Energy [pJ] | Relative Cost |
|---|---|---|
| 32 bit int ADD | 0.1 | 1 |
| 32 bit float ADD | 0.9 | 9 |
| 32 bit int MULT | 3.1 | 31 |
| 32 bit float MULT | 3.7 | 37 |
| 32 bit 32KB SRAM | 5 | 50 |
| ~~32 bit DRAM~~ | ~~640~~ | ~~6400~~ |




DRAM-based Near Memory Processing System [ISCA'16 A. F.-Farahani, et al.]




SRAM-based Processing-In-Memory System

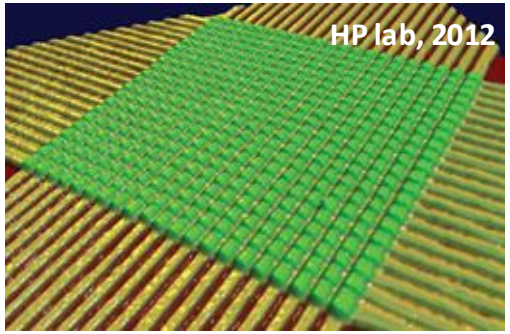
- **Near Memory Processing**
  - Data movement overhead still exists.
- **PIM with Conventional Memory**
  - Large Area/Power Overhead
  - Only support simple classification algorithms with poor accuracy.

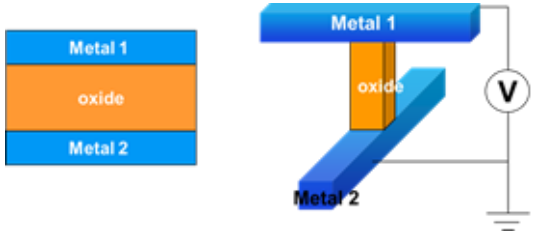
**[JSSC 17, J. Zhang et al.]**

Duke

# eNVM – Computing Inspired by Brain

Synapse Network ⟺ Memristor Crossbar



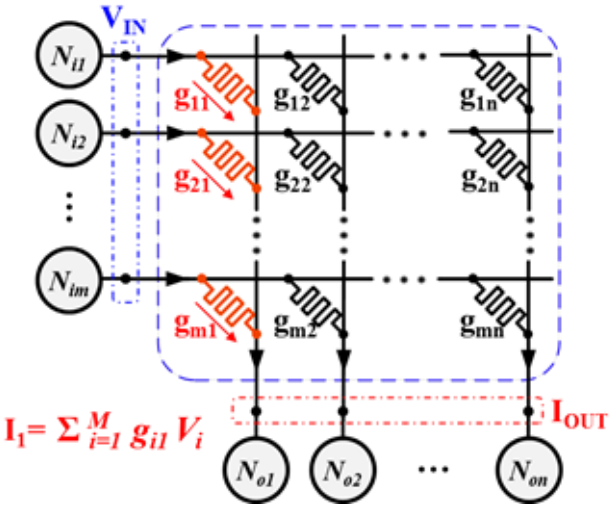




Natural matrix operation

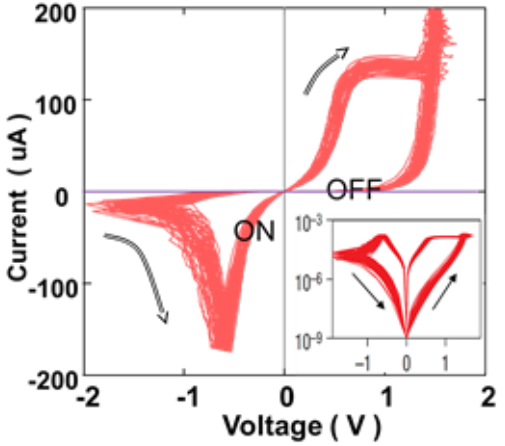
$$[x_1 \; x_2 \; \dots \; x_m] \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix} = [\, y_1 \quad y_2 \quad \cdots \quad y_n \,]$$

$$y_1 = \Sigma \, x_i \cdot g_{i1}$$

CEI DAC'12




High density




[Nature Nanotechnol. 08, J. Joshua Yang et al.]

Duke

# Outline

- Motivation: Neuromorphic hardware testbeds
- Biologically-plausible silicon neurons
- Gain-cell analog in-memory computing
- Conclusions & Takeaways

Duke

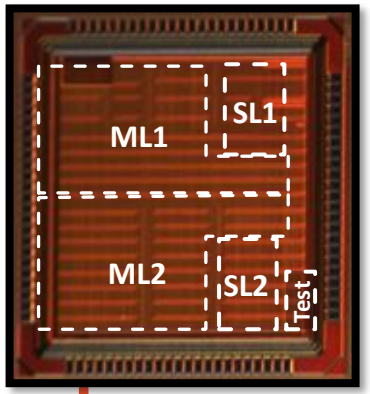
# Outline

- Motivation: Neuromorphic hardware testbeds
- Biologically-plausible silicon neurons
- Gain-cell analog in-memory computing
- Conclusions & Takeaways

Duke

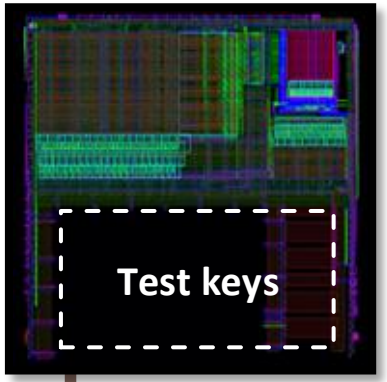
# Memristor-Chip Development at Duke

**Chip 1**
**Spiking**
**2520x2520 µm²**



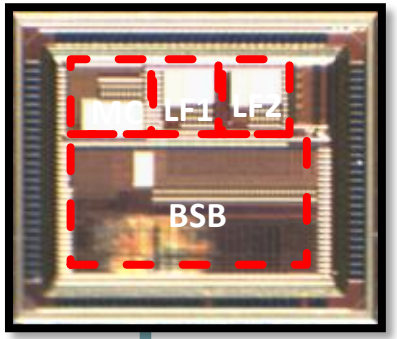

**2015**

**Chip 2**
**Spiking**
**2520x2520µm²**



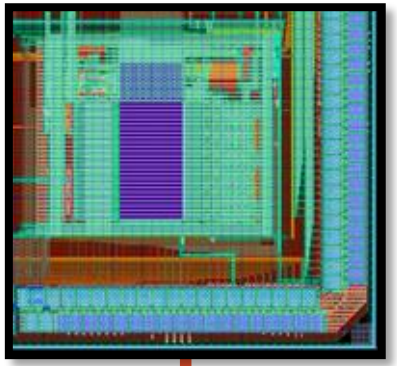

**2016.02**

**Chip 3**
**Level**
**2847x2471 µm²**



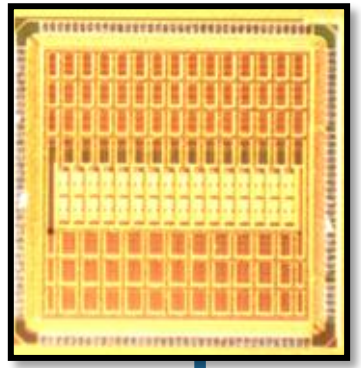

**2016.05**

**Chip 4**
**Spiking**
**700x700 µm²**



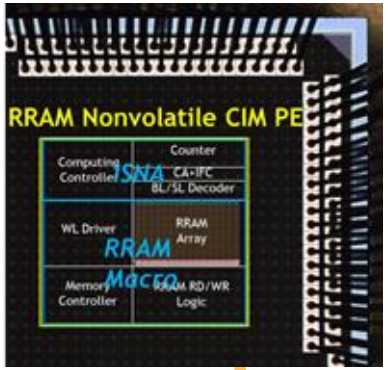
**2016.10**

**Chip 5**
**Spiking**
**3152x3152 µm²**



**2017**

**Chip 6**
**Spiking**
**887x807 µm²**




**2019**

Duke

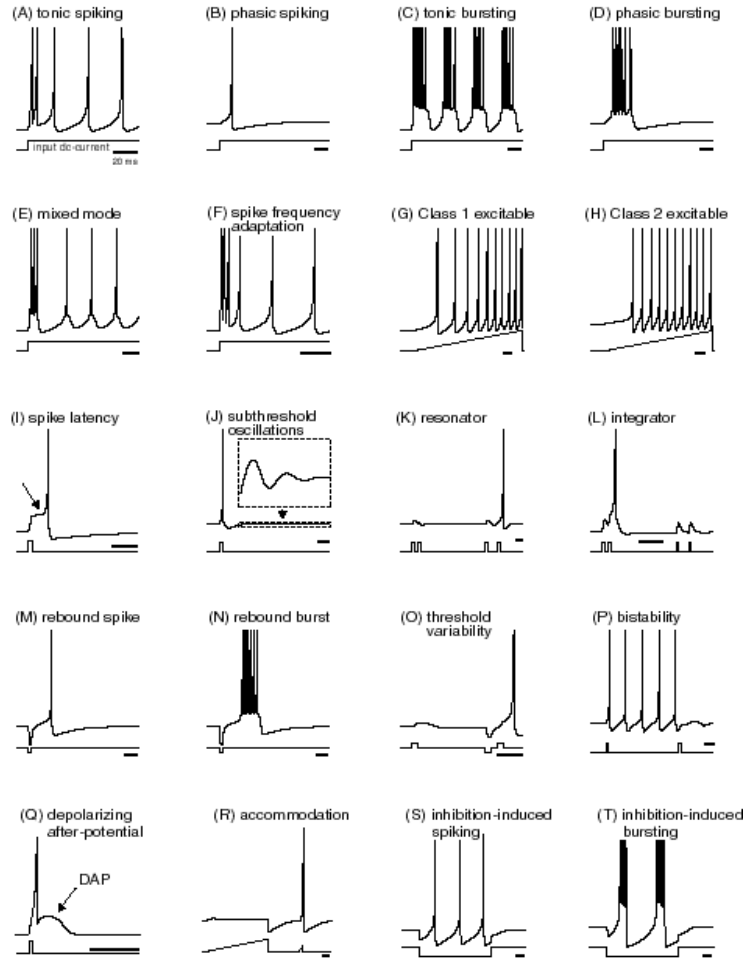
# Outline

- Motivation: Neuromorphic hardware testbeds
- Biologically-plausible silicon neurons
- Gain-cell analog in-memory computing
- Conclusions & Takeaways

Duke

# Goal: A Scalable Neuron

- **Objective:** Robust, low-power, diverse dynamics.
- **Challenge:** Reliability at scale. CMOS for now.
- **Foundation:** Izhikevich neuron model.




1. $\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I$

2. $du/dt = a(bv - u)$

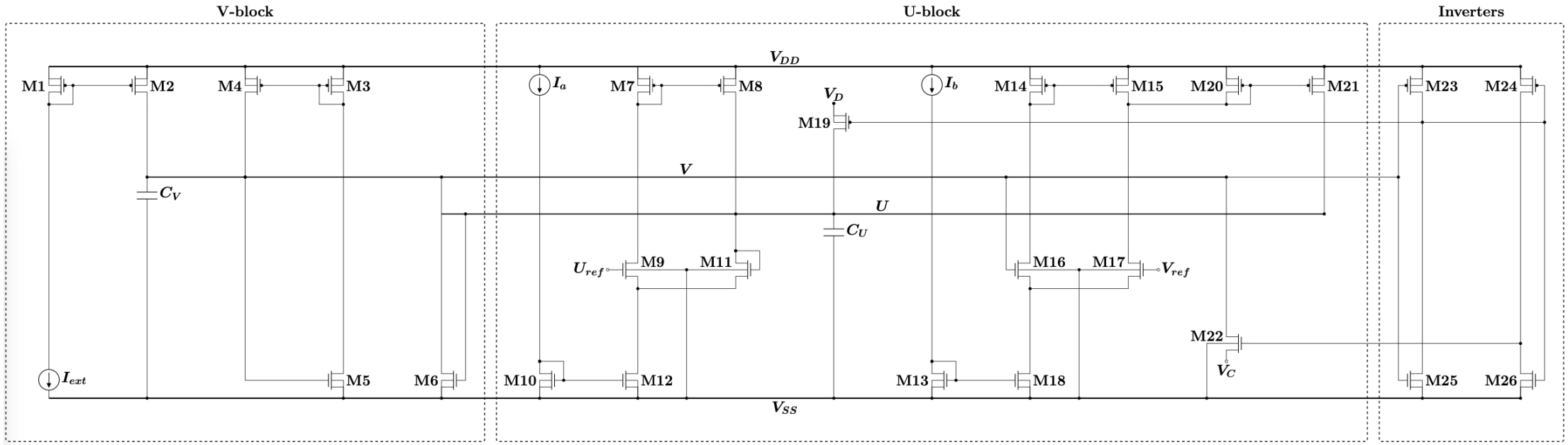
if v ≥ $v_{th}$ :

v ← c , u ← u + d

*a* : recovery time scale of u
*b* : coupling strength from v to u
*c, d* : post-spike reset values of v and u
*v* : membrane potential, the fast state
*u* : recovery variable, the slow state
*I* : input current

Duke

# Schematic




**3-Block Architecture:**

- **V-block**: Membrane potential (M1-M6, $C_V$, $I_{ext}$ )
- **U-block**: Recovery variable (M7-M18, $C_U$,$I_a$, $I_b$)
- **Inverters**: Output stage with threshold ($V_C$)

$$\frac{du}{dt} = \left(\frac{1}{C_u}\right)(I_a + I_b)$$

$$I_a = -g_{ma}(u - U_{ref})$$

$$I_b = g_{mb}(v - V_{ref})$$

$$a = \frac{g_{ma}}{C_u} \qquad b = \frac{g_{mb}}{g_{ma}}$$

*In subthreshold each $g_m$ is set by its OTA tail current, so a and b are tuned by bias, not by W/L.*

**Key Innovations:**

- Orthogonal control of *a* and *b* parameters
- Not hard-coded by W/L ratios
- Fully tunable post-fabrication via bias currents
- Robust to PVT (calibration possible)

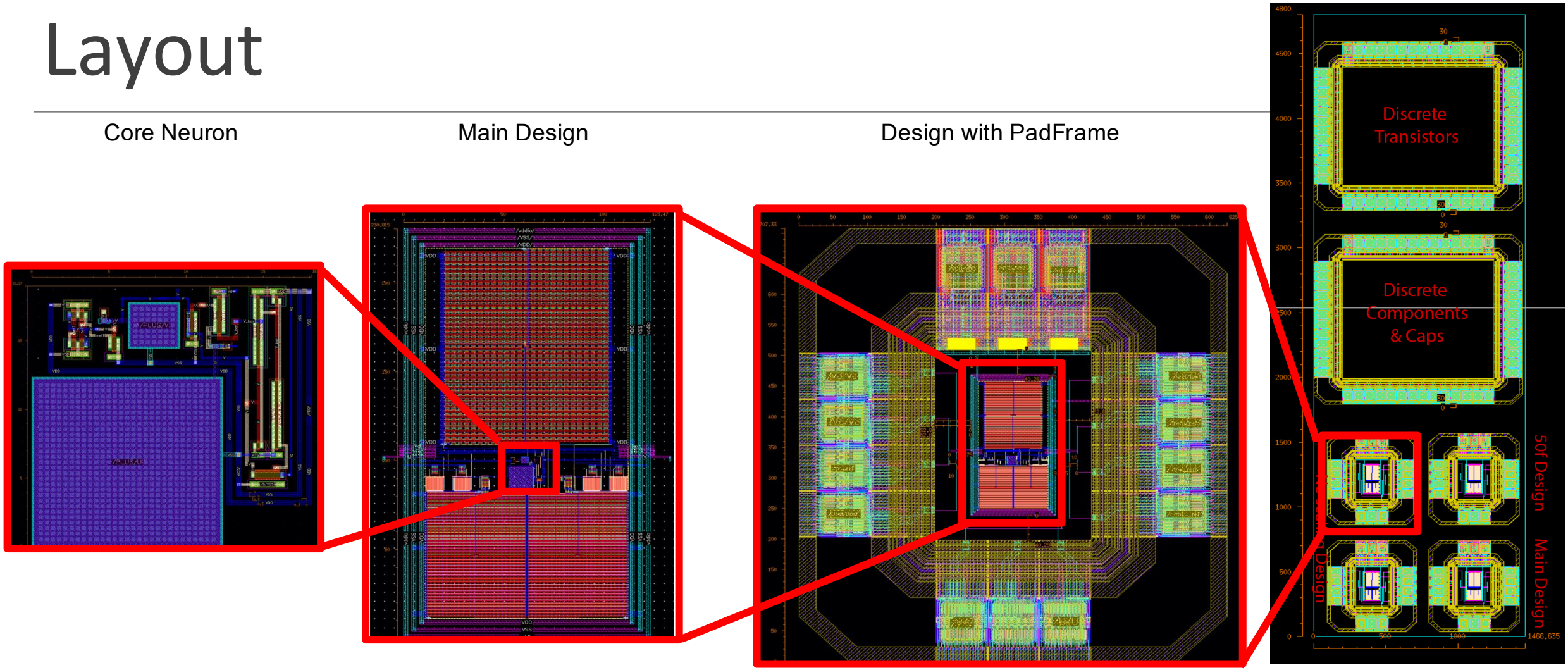
Duke

# Layout

Core Neuron

Main Design

Design with PadFrame




~ 18 µm x 19 µm

~ 123.47 µm x 230.815 µm

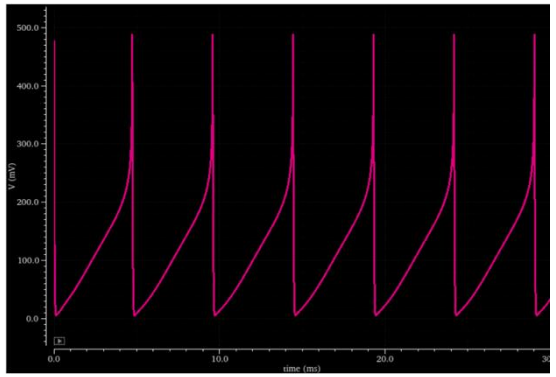
~ 625 µm x 707.33 µm

~ 1466.635 µm x 4800 µm

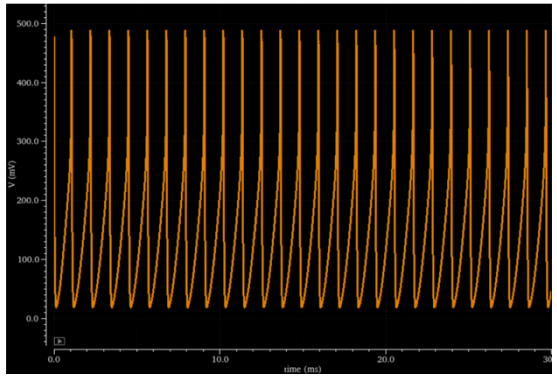
Duke

# Reproduced Spiking Patterns

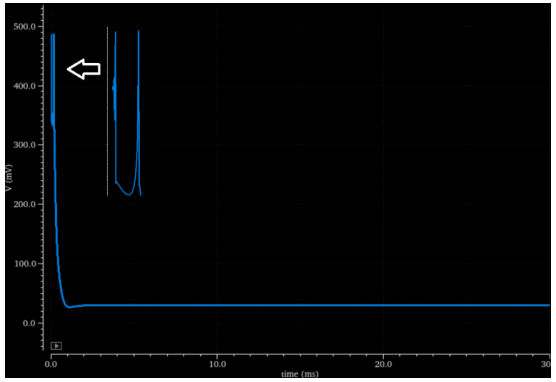
Regular Spiking



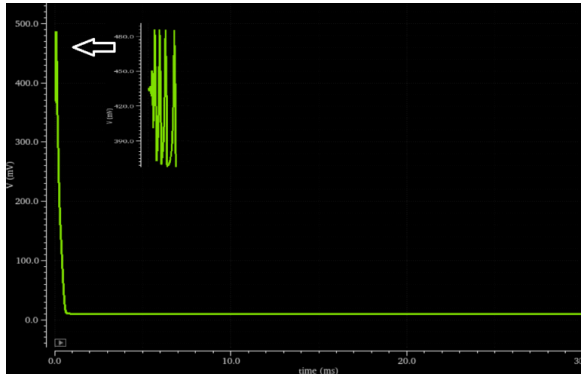

Fast Spiking



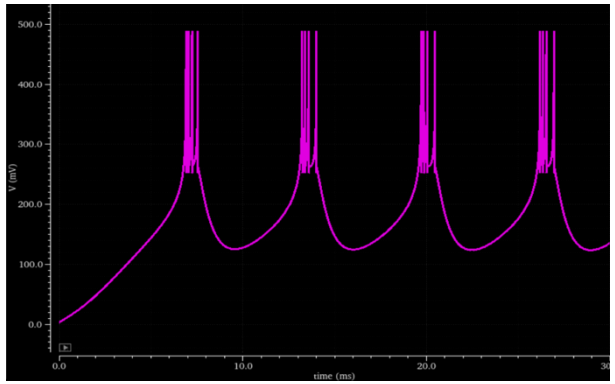

Phasic Spiking



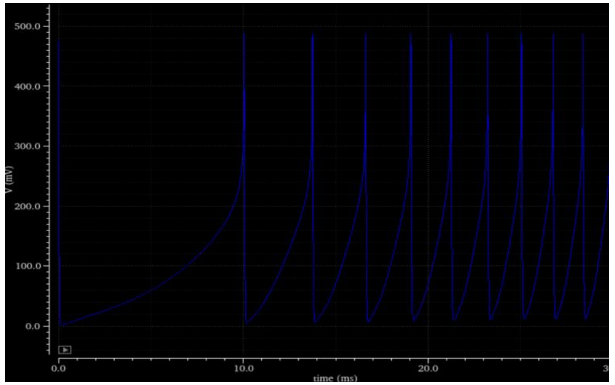

Phasic Bursting



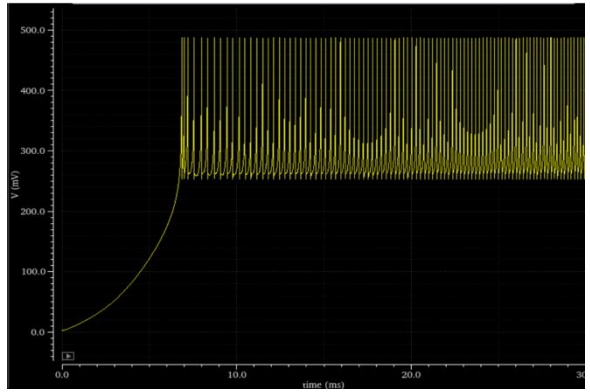

Chattering




Class 1 Excitable




Class 2 Excitable




Duke

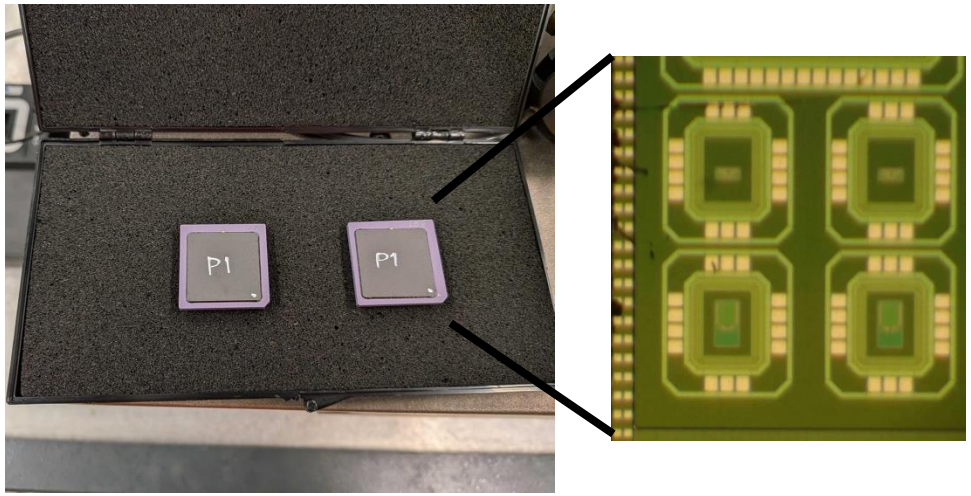
# Comparison with Existing Work

| | Year | Tech (nm) | Patterns | Normalized area | Time scale |
|---|---|---|---|---|---|
| Wijekoon | 2008 | 350 | 4 | 8.19 | Accelerated |
| Schaik | 2010 | 500 | 6 | 58.48 | Biological |
| Rangan | 2010 | 90 | 5 | 8.71 | Biological |
| Grassia | 2011 | 65 | 5 | 2.92 | Biological |
| Ronchini | 2020 | 180 | 3 | 1.38 | Biological |
| **Proposed Nominal** | 2026 | 130 | **7** | **1.0** | Biological |
| **Proposed Scaled** | 2026 | 130 | **7** | **0.53** | Biological |

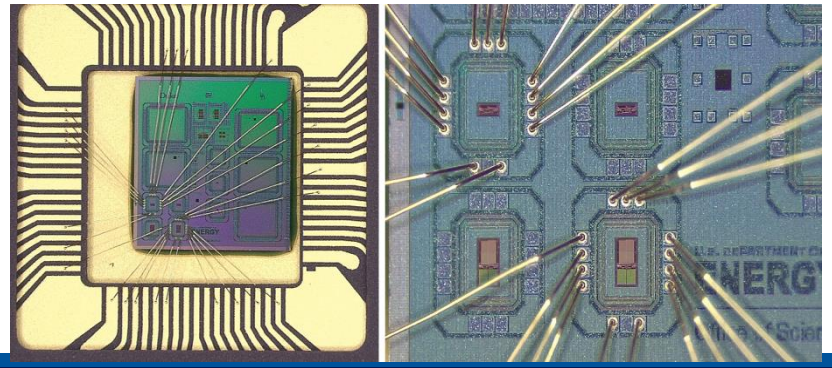
Duke

# Ongoing: Chip Measurement

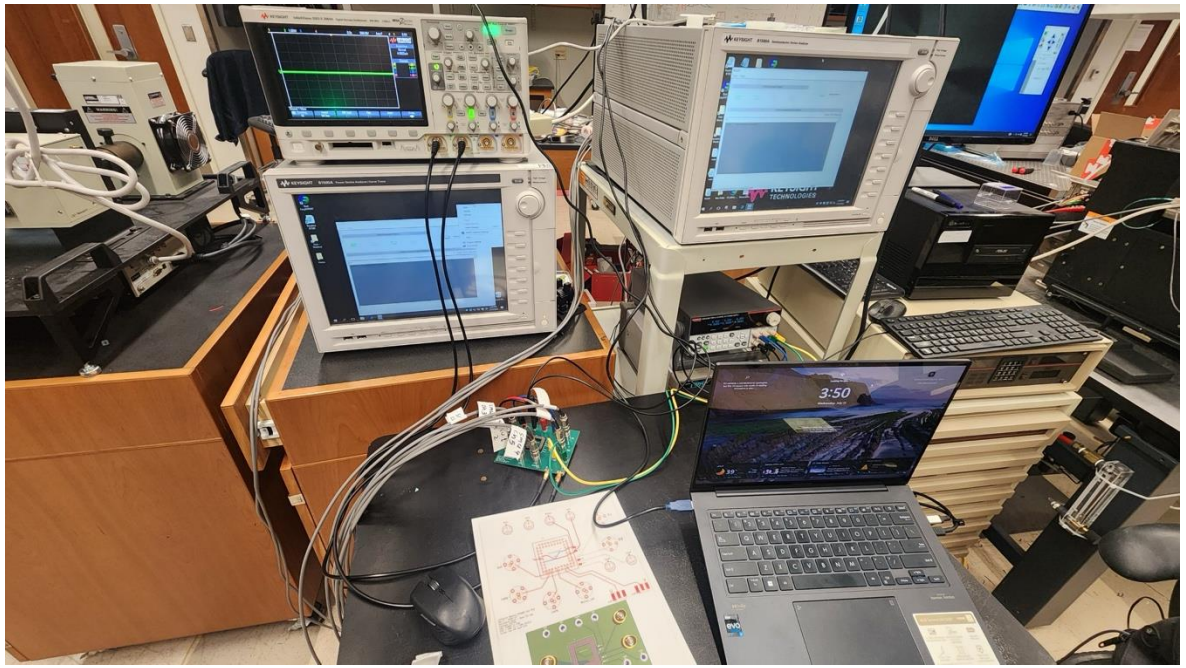
**Fabricated 4 neurons**




**Packaging**



**Experimental setup**

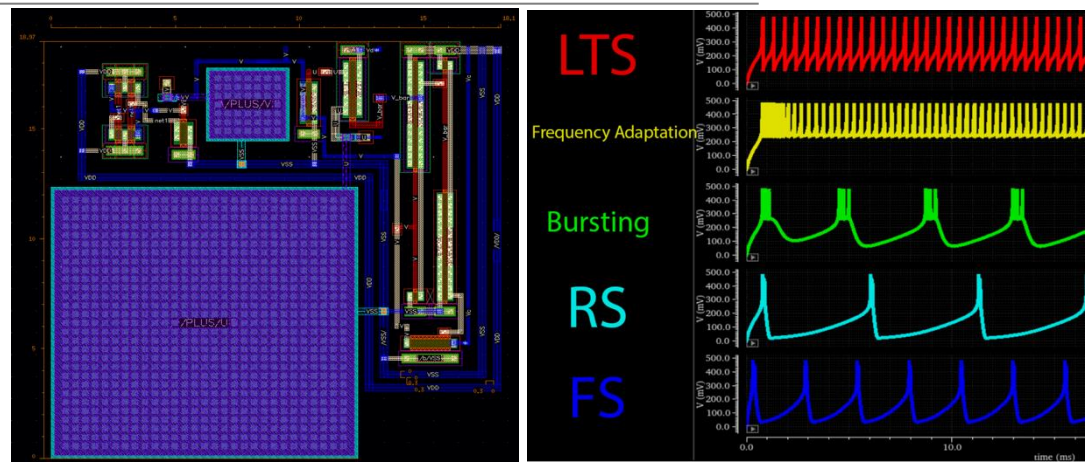



Duke

# Significance and Impact

## Scientific Achievement

- First taped-out silicon neuron reproducing all 7 canonical Izhikevich spiking patterns on SkyWater 130nm.
- Preset bias currents enable post-fabrication process-variation tuning (unlike fixed W/L ratios).
- Core: $\sim 342\ \mu m^2$ (18 × 19 µm), scalable to ~182 µm² at $C_u = 50\ fF$ (smallest reported).




*Fabricated Izhikevich neuron (18 × 19 µm, SkyWater 130nm): 7 firing patterns (RS, FS, bursting, LTS, etc.) from one circuit via bias currents alone. Smallest area, highest fidelity of any taped-out silicon Izhikevich neuron to date.*

## Significance and Impact

- Consumes ~0.1-1 pJ/spike vs. ~10 pJ/spike for biological neurons (10X better), extrapolating to ~1-5 fJ/spike at 2nm (~10,000X).
- Prior silicon Izhikevich neurons reproduce 3–6 patterns with hard-coded parameters; this work achieves 7 distinct modes via bias tuning in one chip design.
- Highest pattern fidelity and smallest area of any taped-out silicon Izhikevich neuron.

*DE-SC0026254 and DE-SC0026382*

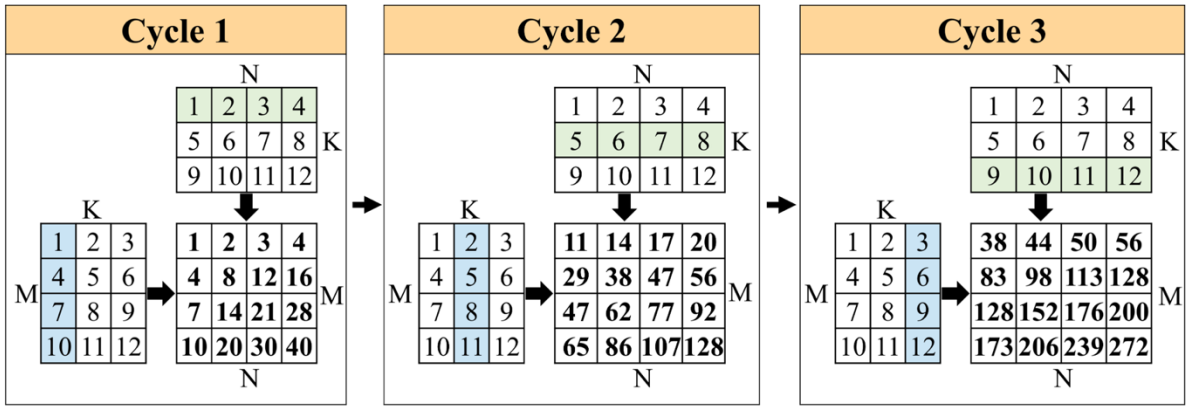
U.S. DEPARTMENT of ENERGY Duke

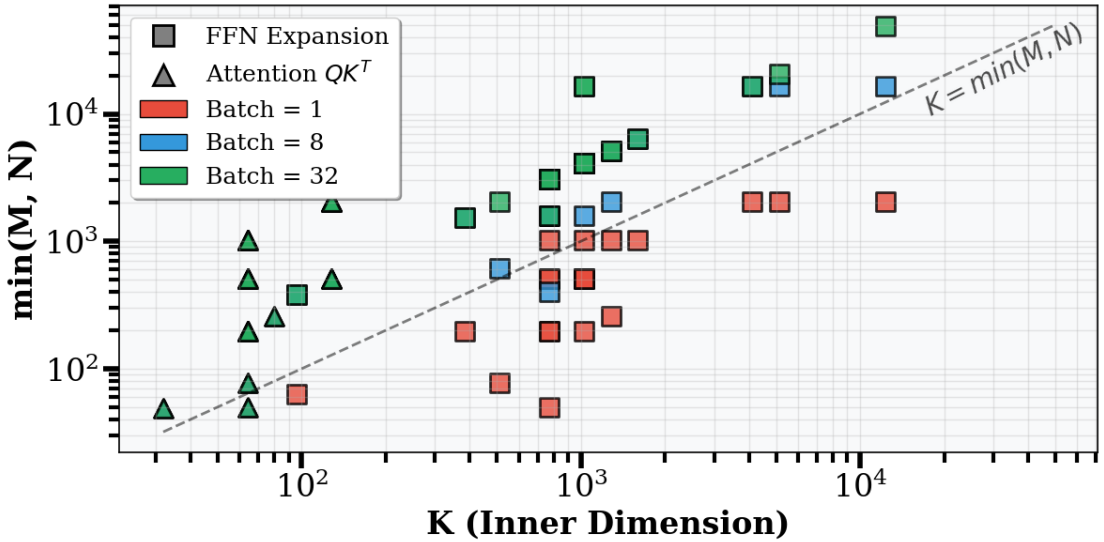
# Outline

- Motivation: Neuromorphic hardware testbeds
- Biologically-plausible silicon neurons
- Gain-cell analog in-memory computing
- Conclusions & Takeaways

Duke

# Analog IMC Has a Dataflow Problem



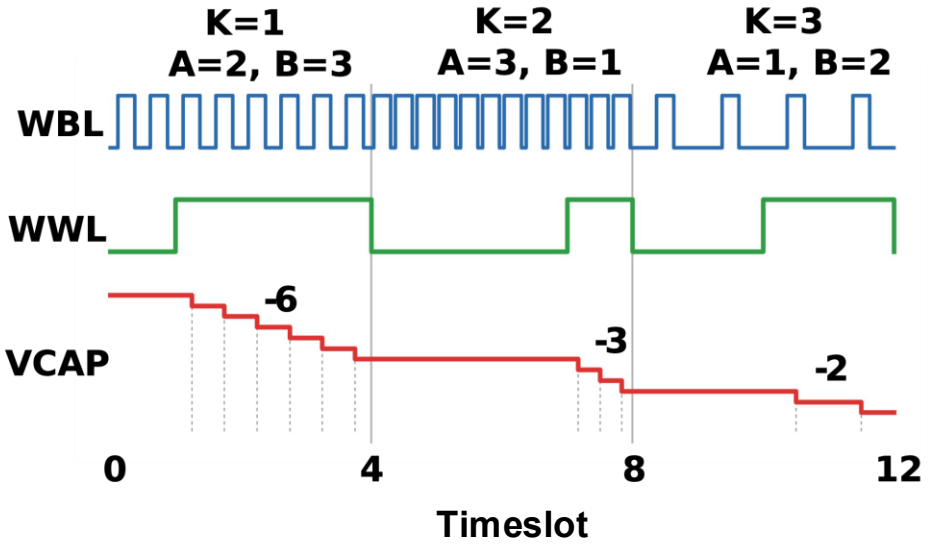



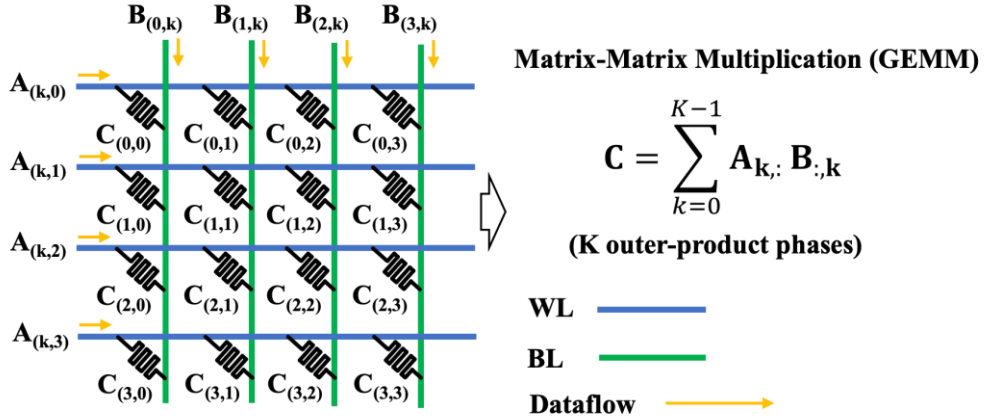

**Weight-stationary IMC spends cycles on the output dimension M, not the inner dimension K, and bitline summing mixes cells so no cell can be calibrated.**

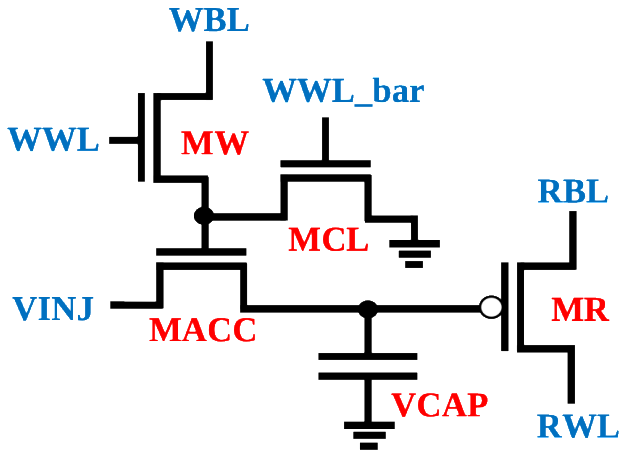
Duke

# Output-Stationary GEMM, and the Cell It Needs

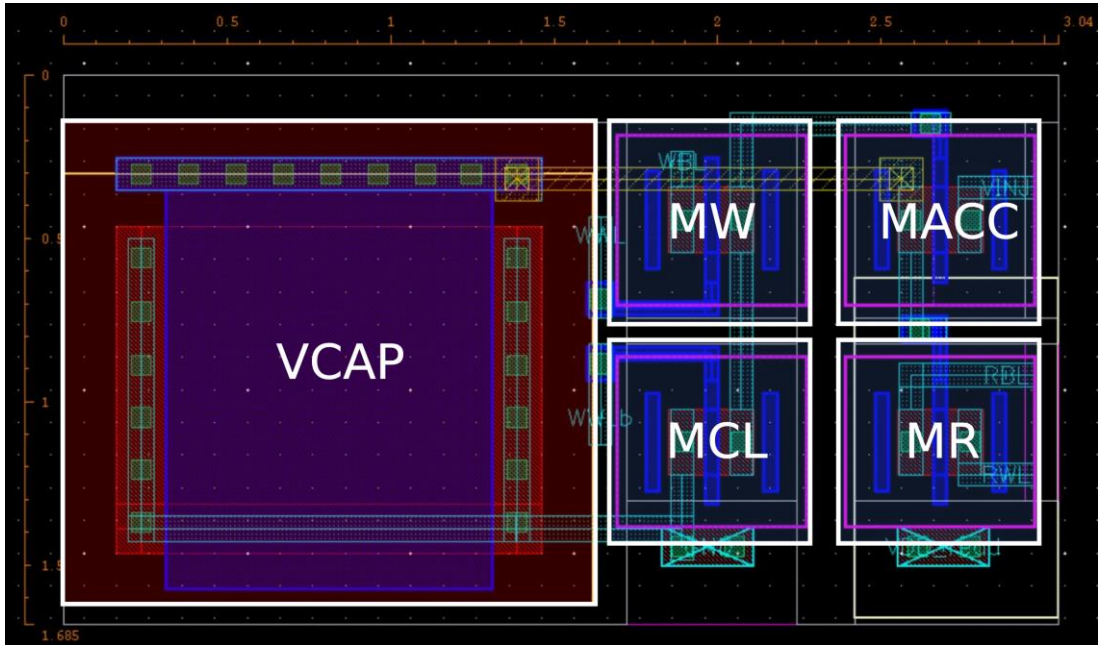
- **Dataflow:** accumulate outer products over K phases, and keep each output in its own cell.
- **Multiply:** input A sets the number of write-bitline pulses, B sets the active word-line window, and each overlap drops $V_{CAP}$ by one step.
- **Accumulate:** $V_{CAP}$ falls by A x B steps per phase, summing over the K phases.
- **Why not nonvolatile:** those devices wear out under repeated writes, so they can't hold an output-stationary accumulation.



# A 4T1C Gain-Cell



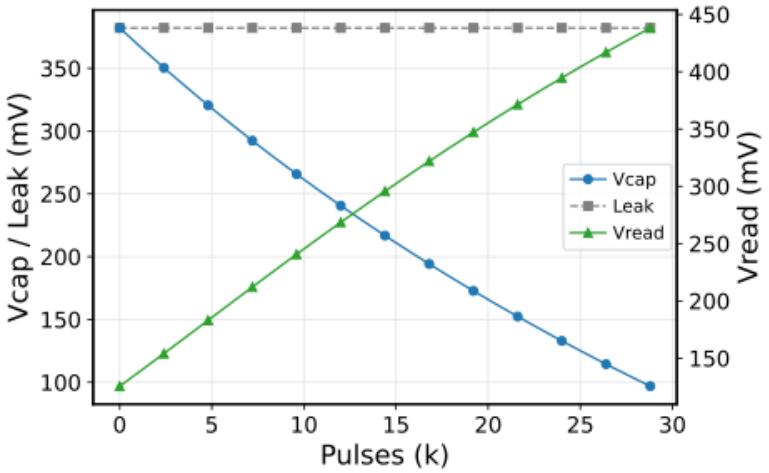

| Device | Cell quality | | | Write cost | |
|---|---|---|---|---|---|
| | States ↑ | Nonlin. ↓ | Var. (%) ↓ | Pulse (V/t) ↓ | Energy ↓ |
| Ag:a-Si [18] | 97 | 2.4 | 3.5 | 3.2 / 300 $\mu$s | 118 pJ |
| TaOx/HfOx [46] | 128 | 0.04 | 3.7 | 1.6 / 50 ns | 1.3 pJ |
| PCMO [34] | 50 | 3.68 | <1 | −2.0 / 1 ms | 174 pJ |
| AlOx/$HfO_2$ [45] | 40 | 1.94 | 5 | 0.9 / 100 $\mu$s | 4.8 nJ |
| GST PCM [23] | 100–120 | 0.105 | 1.5 | 0.7 / 6 $\mu$s | 0.62 nJ |
| EpiRAM [6] | 64 | 0.5 | 2 | 5.0 / 5 $\mu$s | 1.5 nJ |
| HZO FeFET [16] | 32 | 1.75 | <0.5 | 3.65 / 75 ns | —† |
| **Ours** | **29,000** | 0.80* / **<0.01** | **0.46** | 0.27 / **50 ps** | **0.02 fJ** |

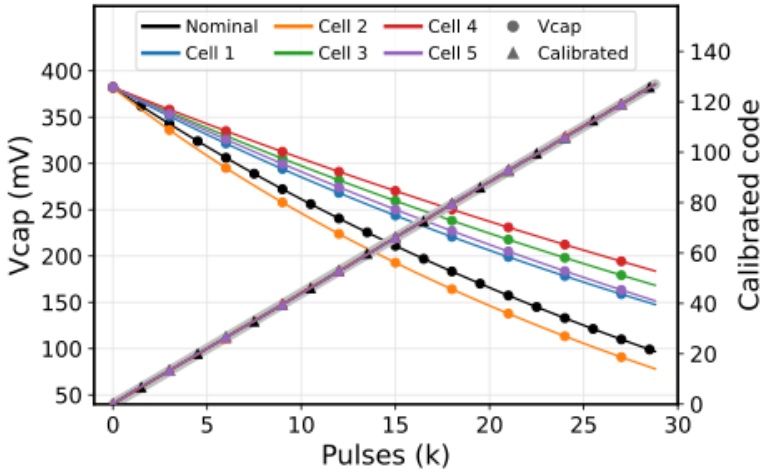
↑/↓: higher/lower is better. Pulse (V/t): write amplitude / duration. *Raw / calibrated nonlinearity; other rows report raw values. †Not reported.

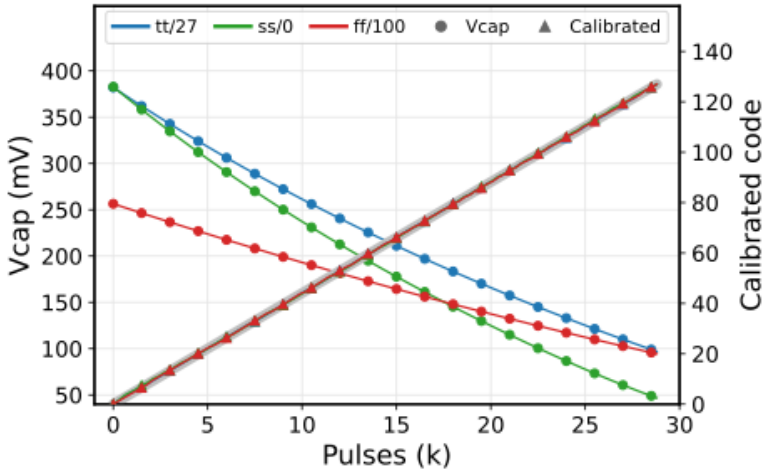
**The cell extends 2T1C so the multiply and the accumulate both happen inside the cell, in 3.02 by 1.54 μm² of TSMC 40 nm.**

Duke

# Per-Cell Calibration Holds the Result



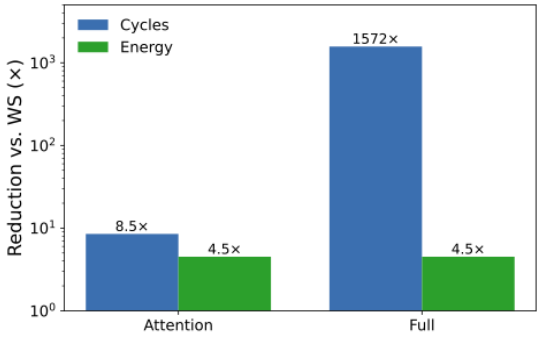

**Per-cell calibration maps a 100 mV spread of Monte Carlo cells onto one code scale, and the GEMM output stays within one 7-bit code of the digital baseline, two at the worst noisy corner.**

Duke

# Significance and Impact

**Scientific Achievement:**

- A 4T1C CMOS gain-cell that does both the multiply and the accumulate inside the cell.
- 29,000 analog states with a 50 ps, 0.02 fJ write, so it takes the repeated writes an output-stationary accumulation needs.
- Per-cell calibration removes mismatch, nonlinearity, and leakage drift, which a bitline-summed array cannot do.




**Significance and Impact:**

- Output-stationary GEMM runs a 1,572x fewer MAC cycles than weight-stationary, and 8.5x fewer for attention.
- Per-MAC energy is 4.5x lower.
- On GPT-2 W4A4, perplexity rises only 0.1% for the attention mapping and 2.5% for the full mapping.
- Establishes the CMOS gain-cell as a practical memory primitive for analog GEMM.

*NSF 2112562, 2233808, 2332744; DARPA W912CG25CA001.*

# Outline

- Motivation: Neuromorphic hardware testbeds
- Biologically-plausible silicon neurons
- Gain-cell analog in-memory computing
- Conclusions & Takeaways

Duke

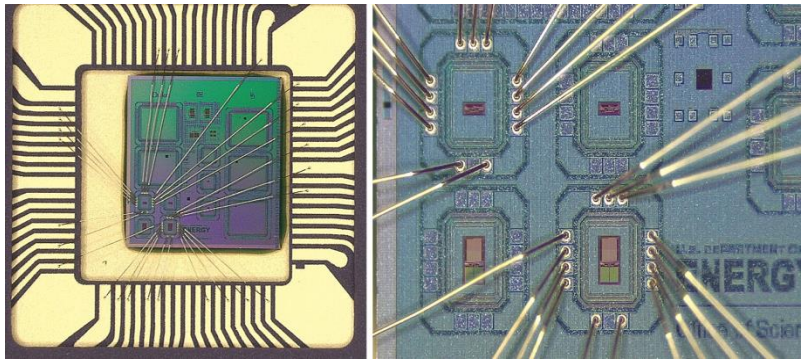
# Conclusion & Takeaways: Analog Primitives You Can Trim

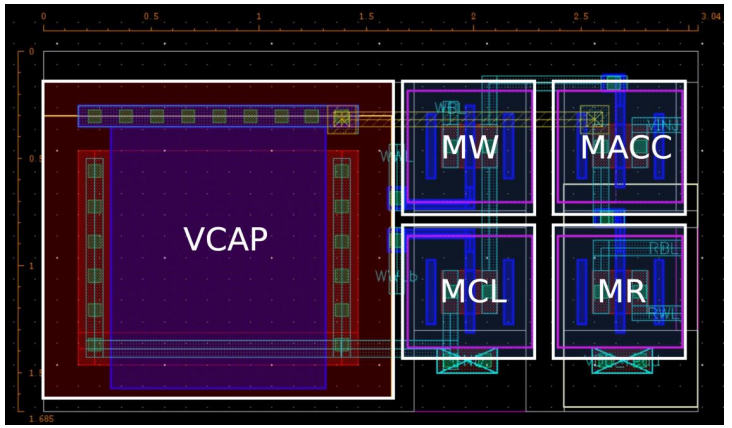
**Silicon primitive**

- A real neuron circuit, taped out in SkyWater 130 nm.
- One circuit produces 7 different firing patterns.
- The core is ~342 µm², scaling to ~182 µm².
- We set the behavior with bias currents, so each neuron can still be tuned after it's built.

**Gain-cell IMC**

- A 4T1C gain cell that multiplies and accumulates inside the cell, in 40 nm CMOS.
- 29,000 analog states, written in 50 ps at 0.02 fJ.
- Per-cell calibration keeps the result within one 7-bit code of digital.
- The dataflow runs 1572x fewer MAC cycles and uses 4.5x less energy.






*Analog only works if you can trim it. The neuron gives us bias currents that set its behavior after fabrication, and the gain cell gives us a per-cell calibration that removes its mismatch.*

# Q&A

Duke