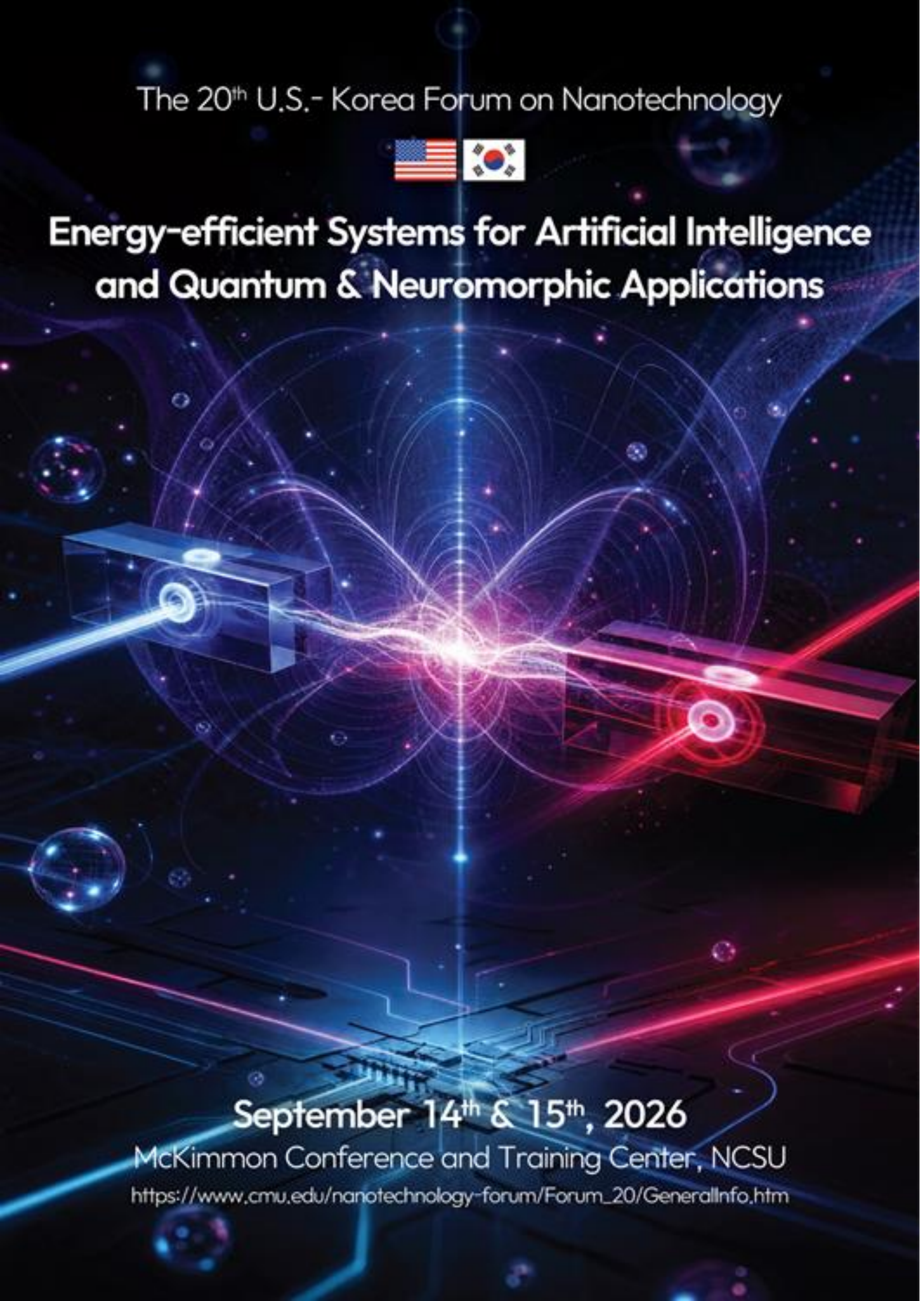


Invited Talk

The Next Horizon of AI Hardware: Silicon That Learns Continually

Dhireesha Kudithipudi, PhD
In collaboration with several NUAI Lab Members

[matrix] The UTSA AI Consortium for Human Well-Being

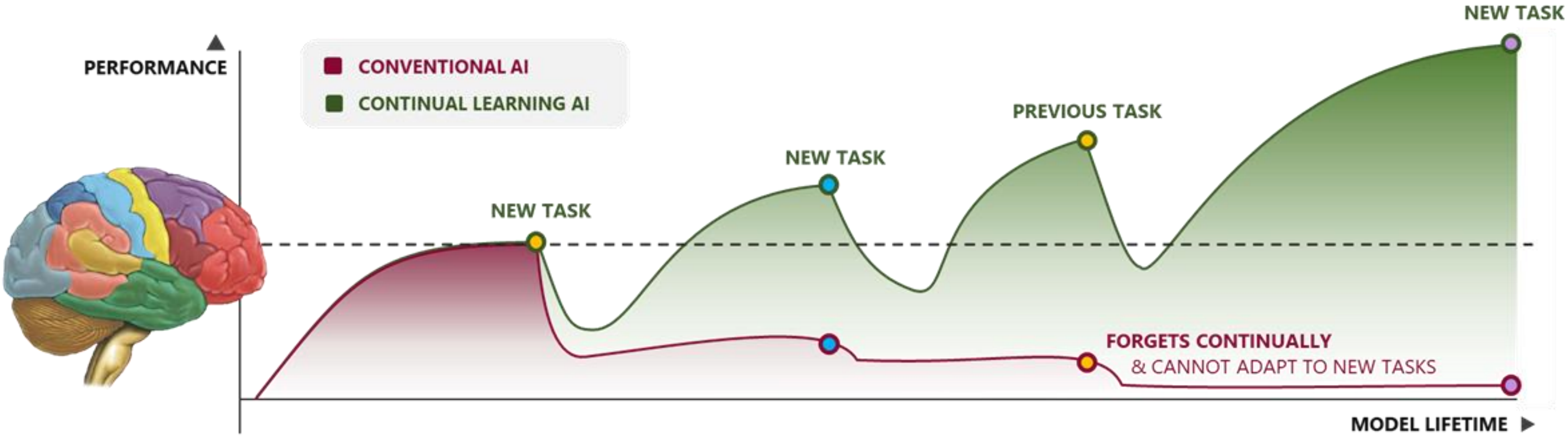
UT San Antonio™

Neuromorphic AI Lab

14th Sep 2026

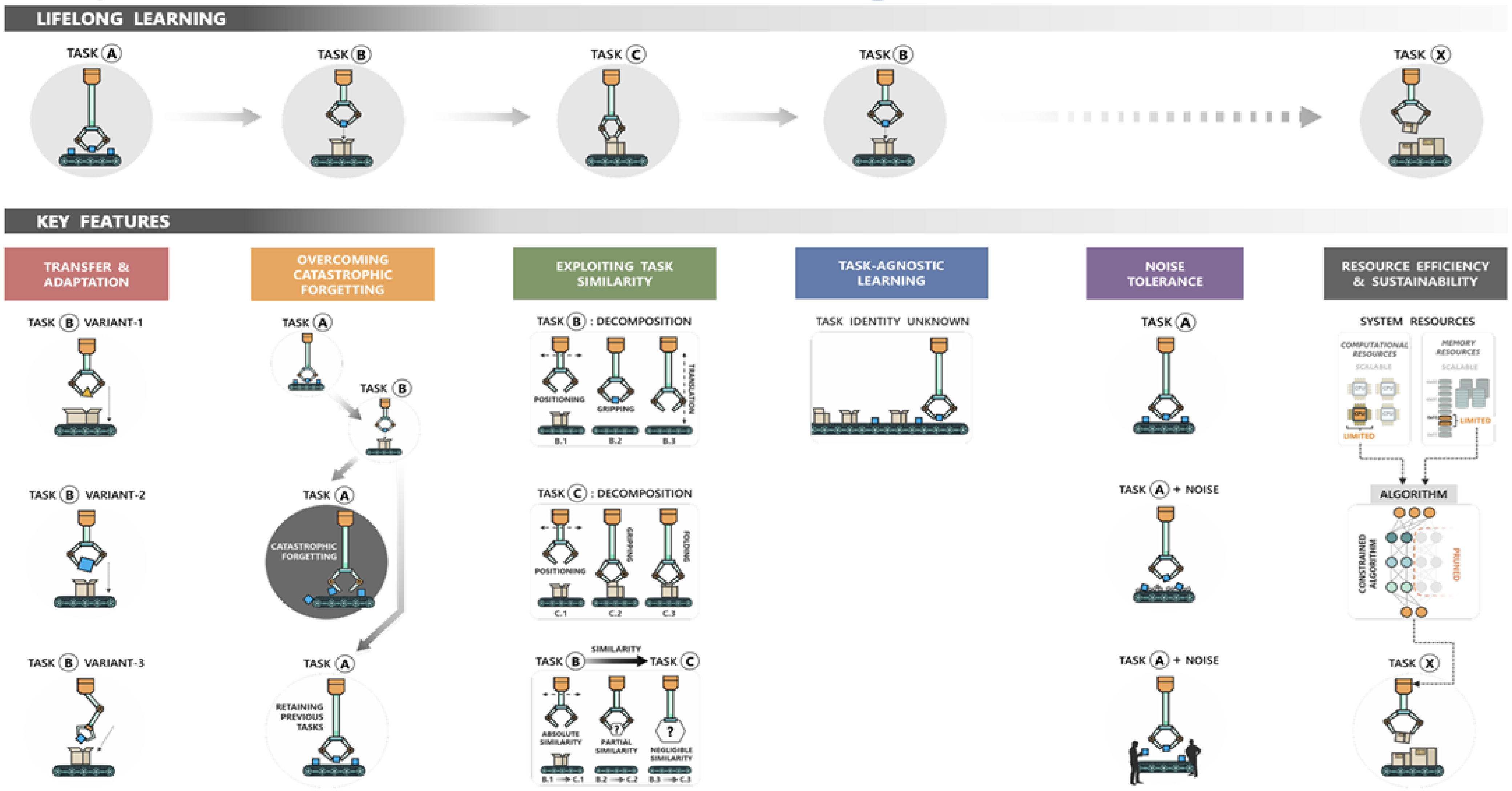
Continual Learning

- Continual learning is an ability to learn incrementally from a non-stationary stream of data
    - **Non-stationary**: distribution of data from what is learned changes over time
    - **Incrementally**: new learning should not overwrite what was learned before, but knowledge should be accumulated



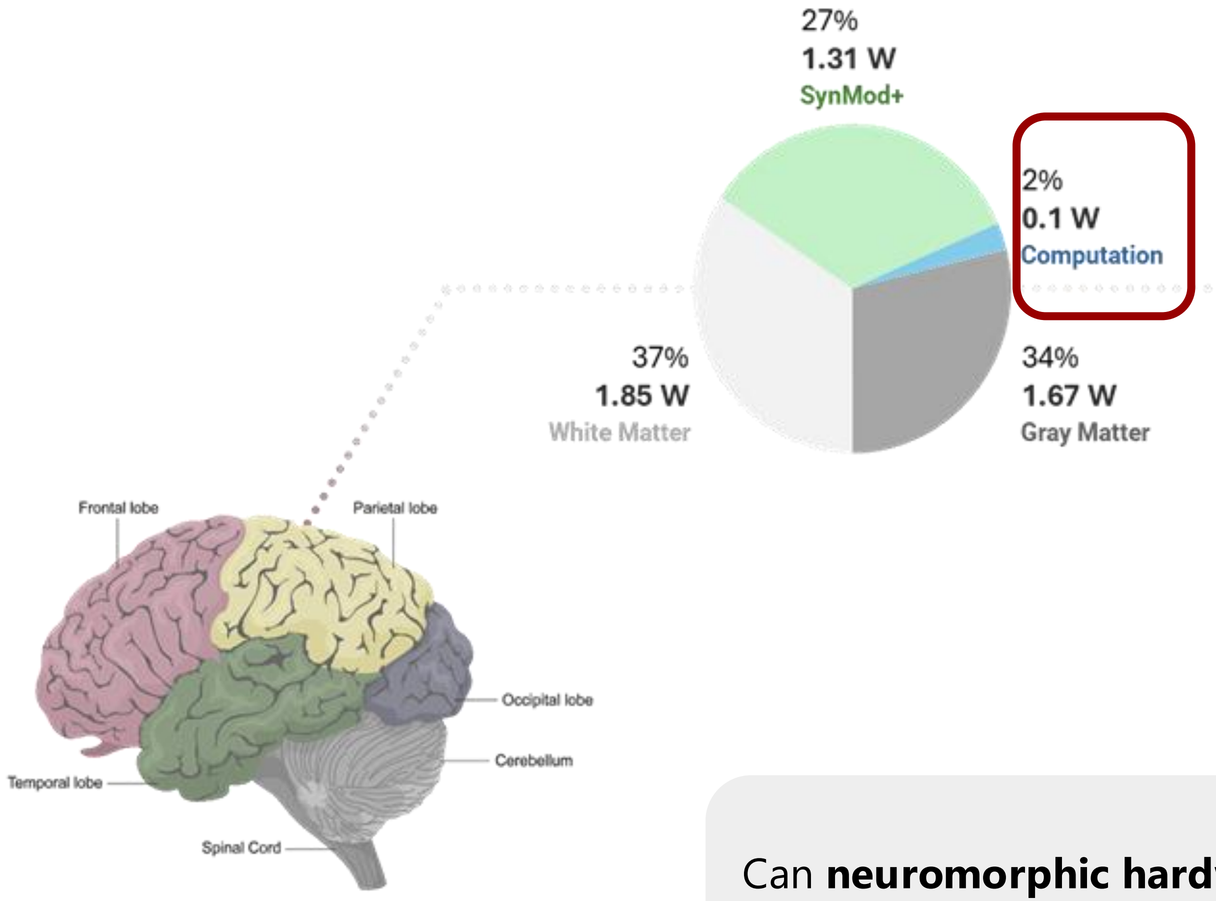
DARPA L2M,
Kudithipudi et.al.Biological underpinnings for lifelong learning machines, Nature Machine Intelligence, 2022
Kudithipudi, NIH NeuroAI 2024; Nature Conference NeuroAI,

Key Features of Continual Learning



[1]Kudithipudi, Dhireesha, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P. Brna et al. "Biological underpinnings for lifelong learning machines." *Nature Machine Intelligence* 4, no. 3 (2022): 196-210.

Cost of Computation: AI & Brains



OpenAI Five (DOTA 2 video game)

~45,000 training **years**

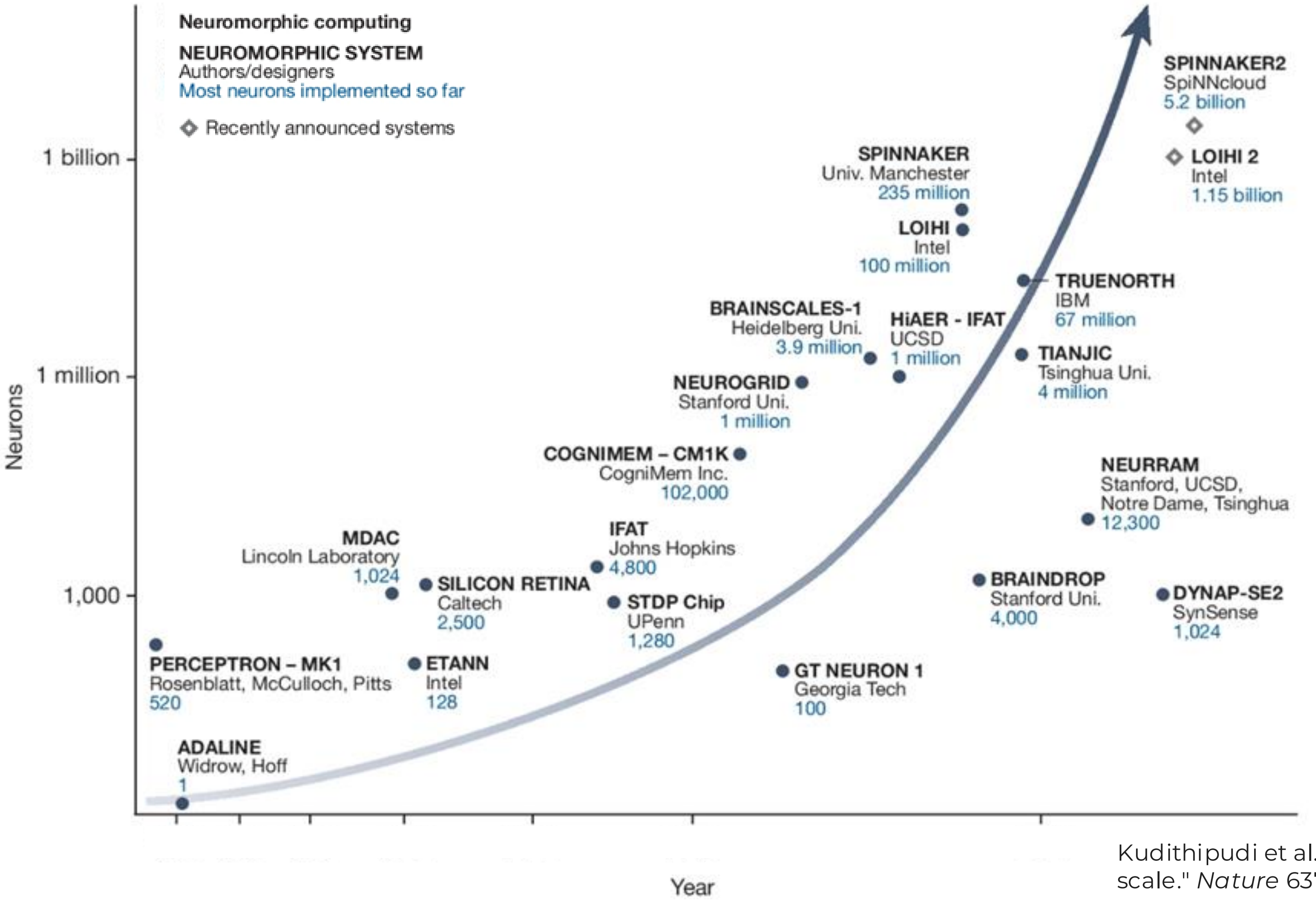
~**100's of MWh** in energy cost

~**$10M** in training cost

Can **neuromorphic hardware** serve as a natural substrate for deploying continual learning AI models ?

Kudithipudi et.al, Design principles for Lifelong LearningAI Accelerators, Nature Electronics, Nov 2023
Kudithipudi, NIH NeuroAI 2024
Levy WB, Calvert VG, PNAS, 2021

Neuromorphic Computing at Scale

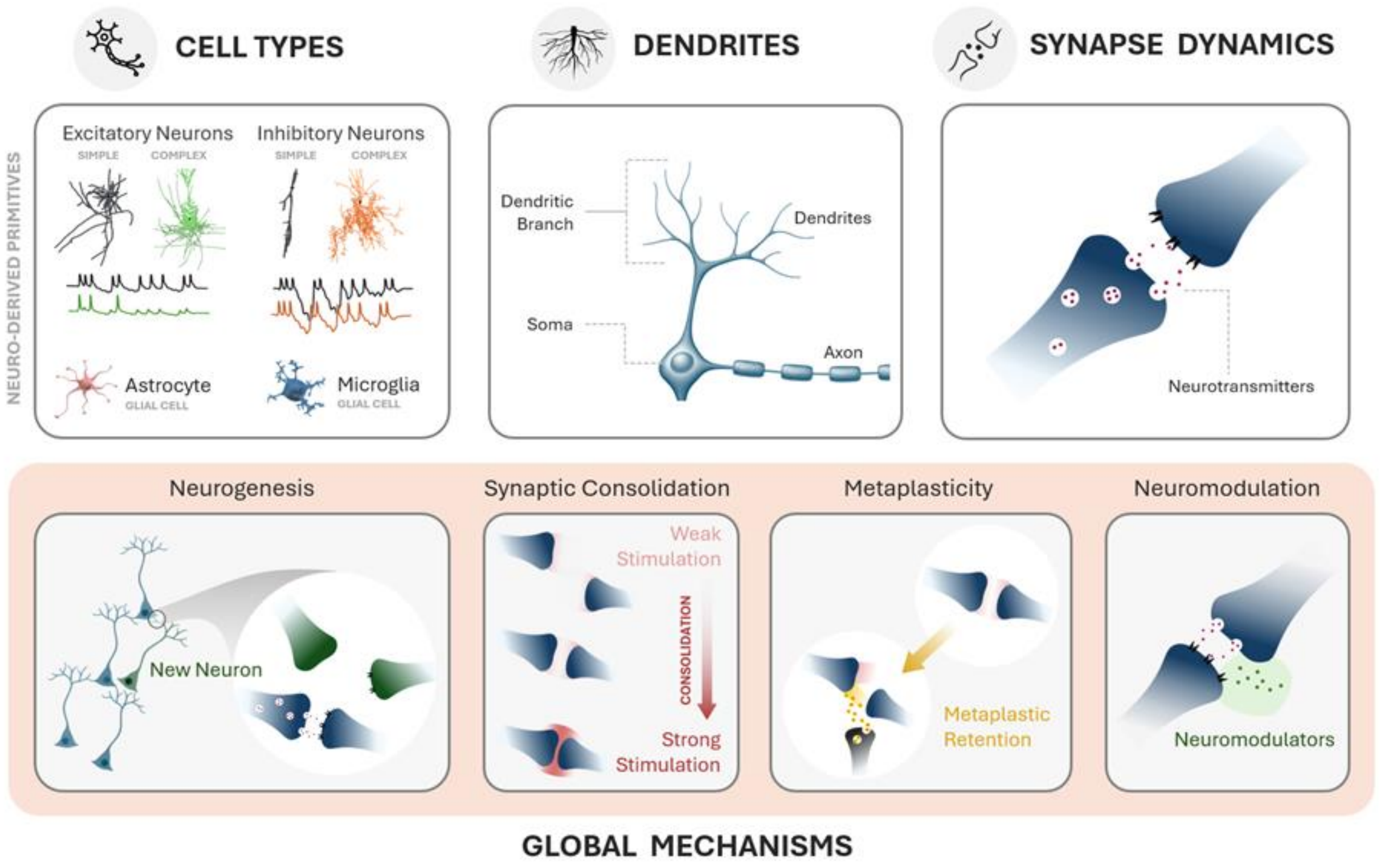


Neuromorphic systems have grown in scale from

a single neuron in the early ***1960s***

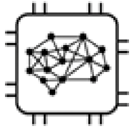
to many billions by ***2025***

Kudithipudi et al. "Neuromorphic computing at scale." *Nature* 637.8047 (2025): 801-812.



Neuro-derived primitives for Continual Learning capable of addressing the complexity and non-linear dynamics of biological brains.

Mismatch Across Abstractions



Computational Principles → Spiking Neural Networks

- Sparse binary representation
- Event-driven computation
- Local learning rule

Memory Consolidation → Continual Learning Mechanisms

- Metaplasticity
- Episodic replay
- Neuromodulation

Structural Mapping → Emerging Devices and Architectures

- Crossbar architectures with emerging memory
- Synaptic features in materials

Current Brain-Inspired CL Approaches

Assumes high-precision, ideal weights [1]

Emerging Memory Devices

Exhibits low precision and high variability

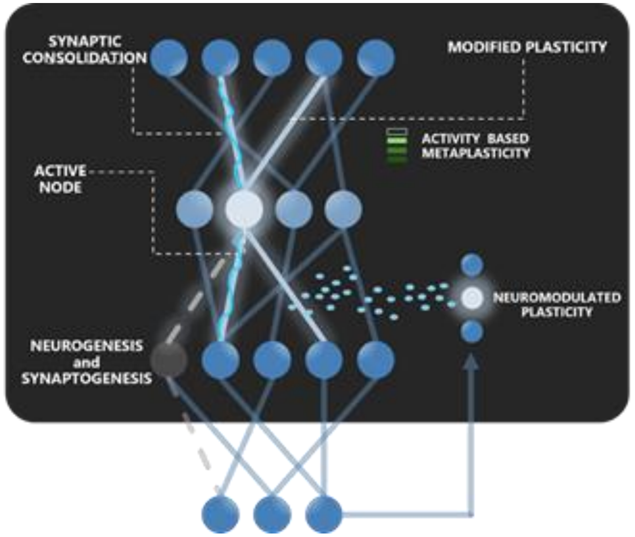
Auxiliary high-precision memory to support continual learning mechanism [2,3]

[1] Soures, Nicholas, et al, “TACOS: Task Agnostic Continual Learning in Spiking Neural Networks” International Conference on Machine Learning workshop 2021 :1–8

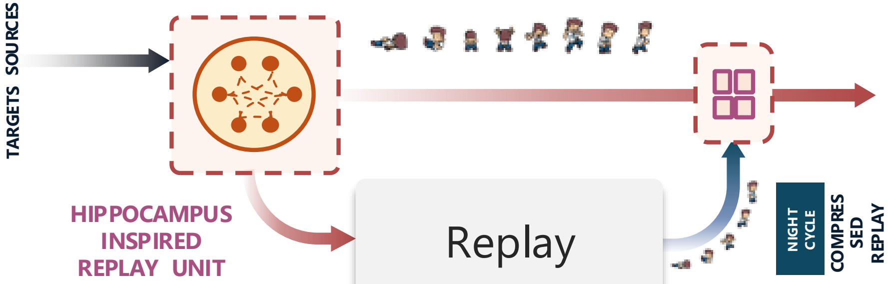
[2] D’Agostino, Simone, et al. "Synaptic metaplasticity with multi-level memristive devices." *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2023.

[3] Drouhin, M., et al. "Leveraging a passive MRAM crossbar for hardware-in-the-loop and continual learning." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* (2026).

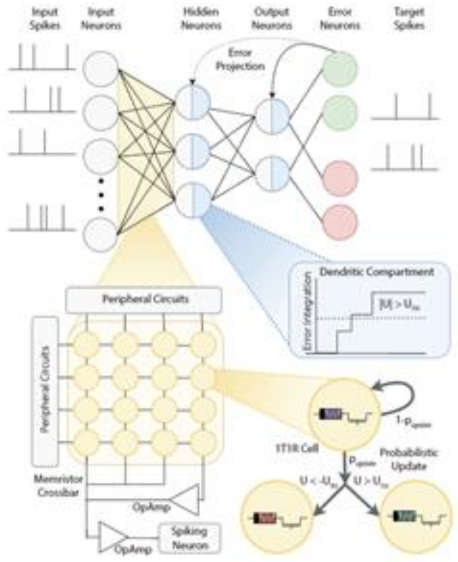
Neuro-Inspired Continual Learning - Across Abstractions



Synaptic Consolidation, Neuromodulation, & Neurogenesis (dynamic architectures)

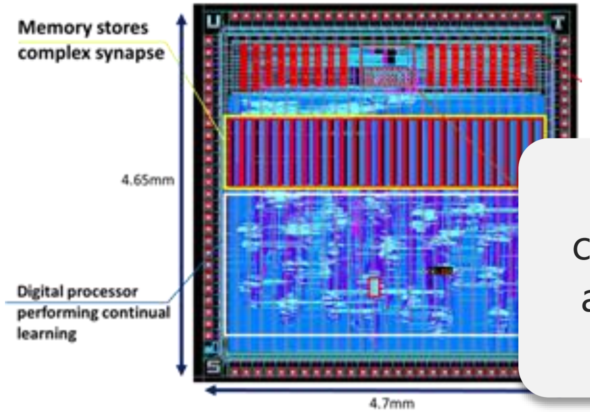


Replay



Metaplasticity

Interplay between **Neuroscience**, **Algorithms** & **Neuromorphic Hardware**

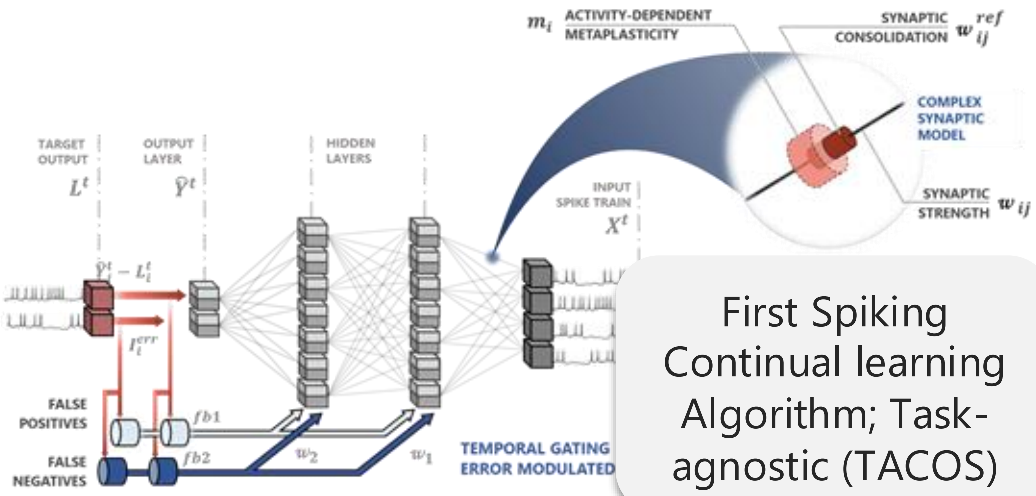


First spiking continual learning accelerator, 2019 Sponsor-AFRL

References

Nature MI'22
Nature Electronics'24
ICML-W'20, ICML'26,
CVPR-W'21, ECAI-W'24,
CCN'26

NUAI LAB



First Spiking Continual learning Algorithm; Task-agnostic (TACOS)

Sponsored by the U.S. Department of Energy
Office of Advanced Scientific Computing Research (ASCR)

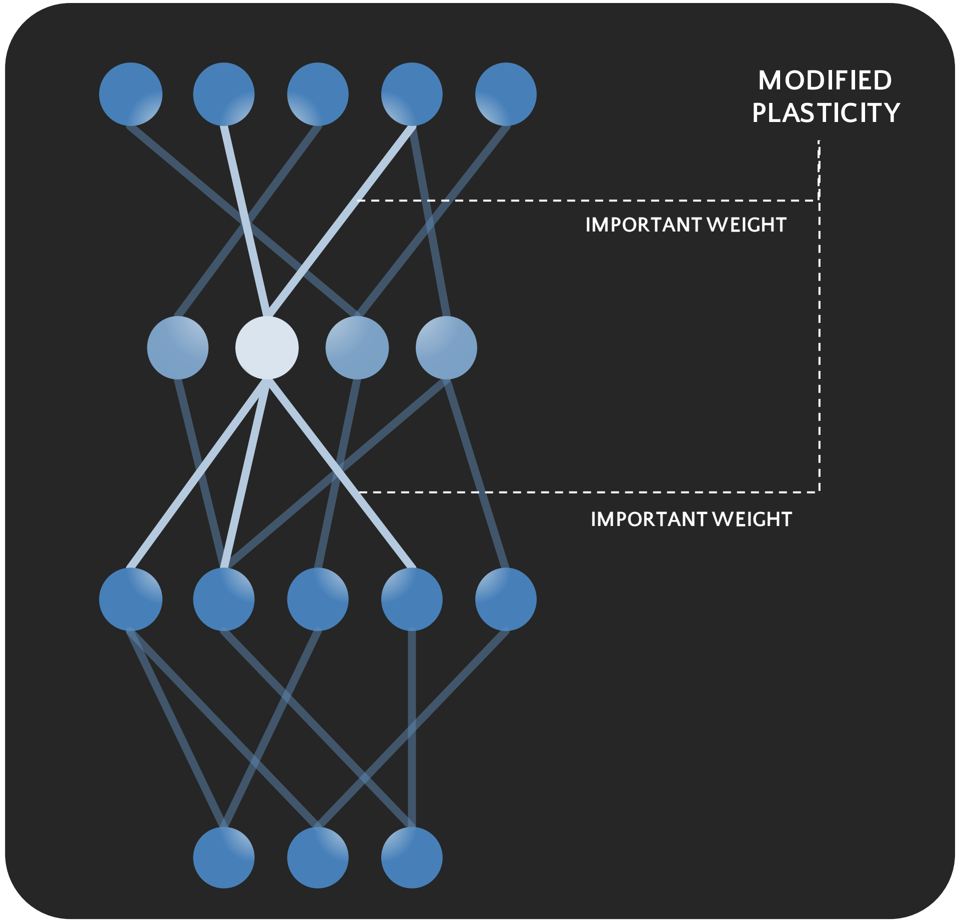
U.S. DEPARTMENT OF ENERGY | Office of Science

Continual Learning with Metaplasticity

- Metaplasticity [1]-inspired methods reduce the plasticity of weights deemed important to retain previously learned information [2]

$$\Delta w_{\text{effective}} = \Delta w \times \text{Metaplasticity\_factor}$$

- Emerging devices exhibit ~3-6 bit conductance resolution [3]
  - Prior solutions rely on auxiliary high-precision memory for gradient accumulation / hidden weights [4,5]



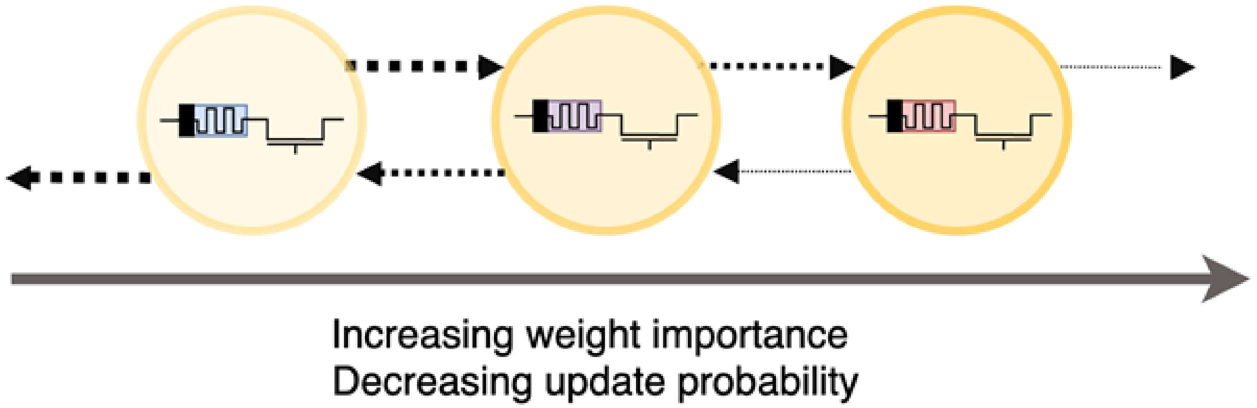
[1] Abraham, Wickliffe C., and Mark F. Bear. "Metaplasticity: the plasticity of synaptic plasticity." *Trends in neurosciences* 19.4 (1996): 126-130.
[2] Kudithipudi, Dhireesha, et al. "Biological underpinnings for lifelong learning machines." *Nature Machine Intelligence* 4.3 (2022): 196-210.
[3] Park, Jaesung, et al. "TiO x-based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing." *IEEE Electron Device Letters* 37.12 (2016): 1559-1562.
[4] D'Agostino, Simone, et al. "Synaptic metaplasticity with multi-level memristive devices." *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)* IEEE, 2023.
[5] Drouhin, M., et al. "Leveraging a passive MRAM crossbar for hardware-in-the-loop and continual learning." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* (2026).

Reformulating Metaplasticity

Computational models of binary metaplastic synapses show that probabilistic transitions can help retain memory. [1]

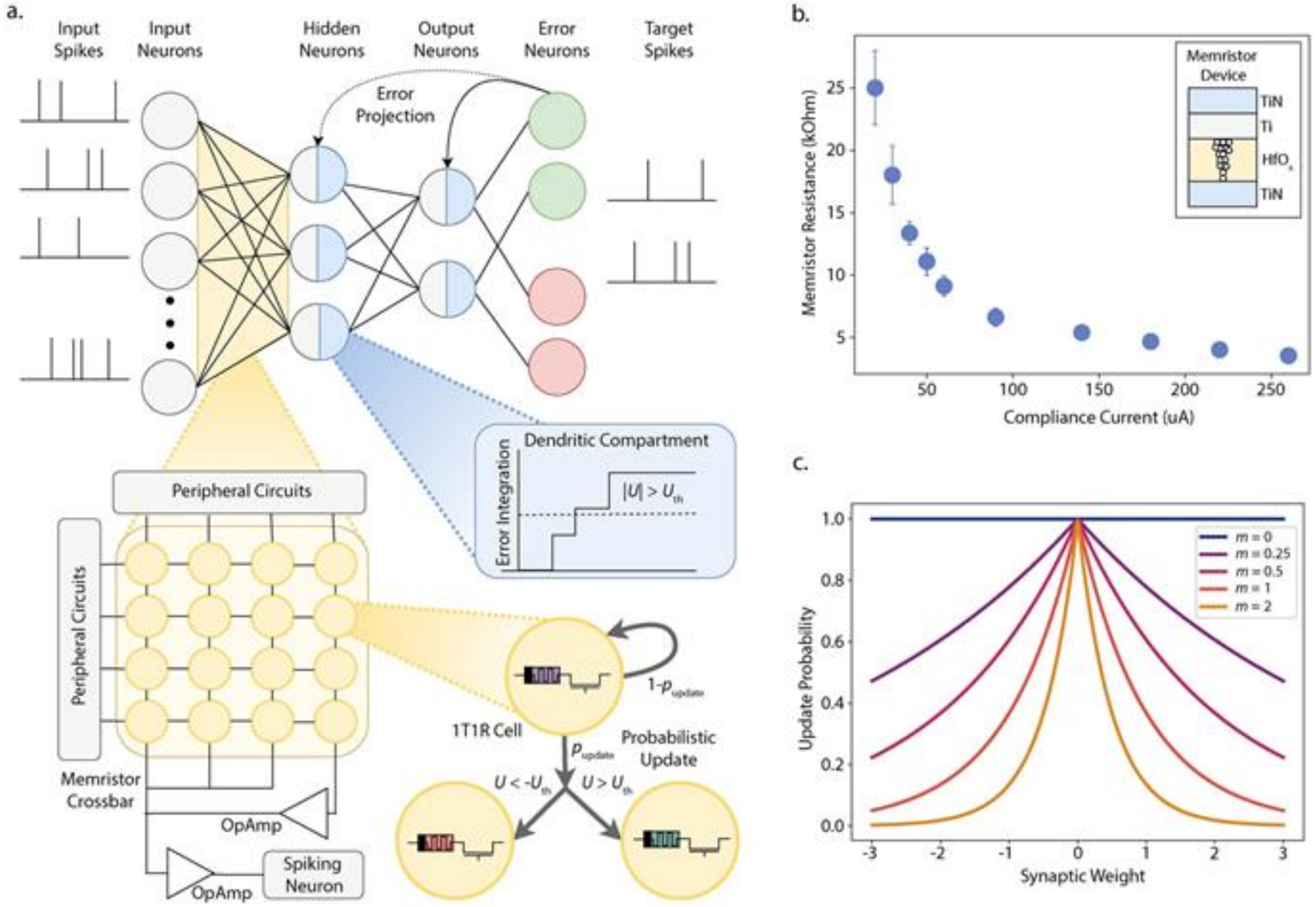
Probabilistic metaplasticity abstracts this principle by interpreting a weight's plasticity as its probability of change instead of the magnitude of change.

Key idea: We consolidate weights by reducing their **likelihood of updating.**



[1] Fusi, Stefano, Patrick J. Drew, and Larry F. Abbott. "Cascade models of synaptically stored memories." *Neuron* 45.4 (2005): 599-611.

Probabilistic Metaplasticity



- Weights are updated probabilistically
- Low update probability assigned to important weights

~60× reduction in energy during the parameter update phase compared to equivalent approach with gradient accumulation.

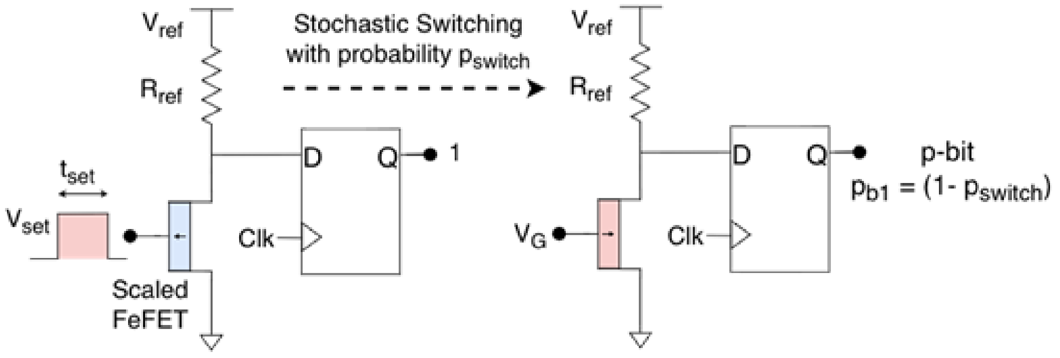
-67% reduction in memory overhead compared to gradient accumulation-based approach

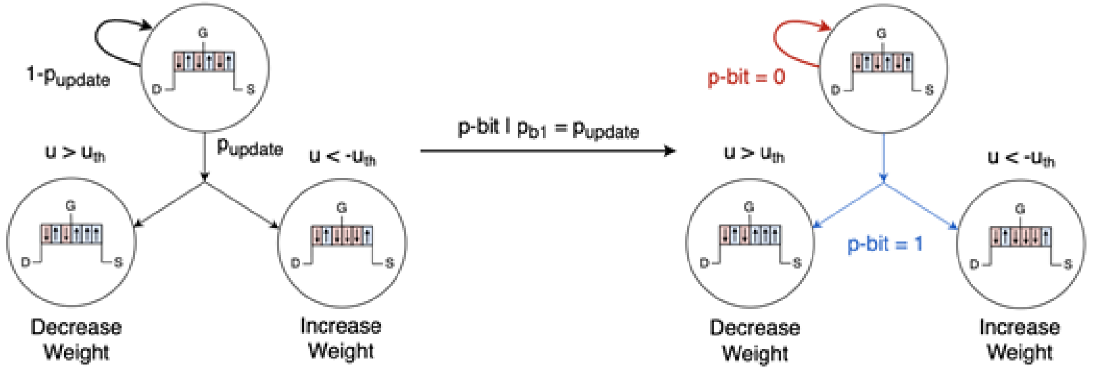
[1] Fatima Tuz Zohora, Vedant Karia, Nicholas Soures, Dhireesha Kudithipudi. “Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks.” Scientific Reports 14(1):1–16, 2024

Probabilistic Metaplasticity with p-bits

- Stochastic polarization switching in scaled FeFETs is used to generate p-bits

- A p-bit is a random binary variable that is 1 with a tunable probability (p_{b1})



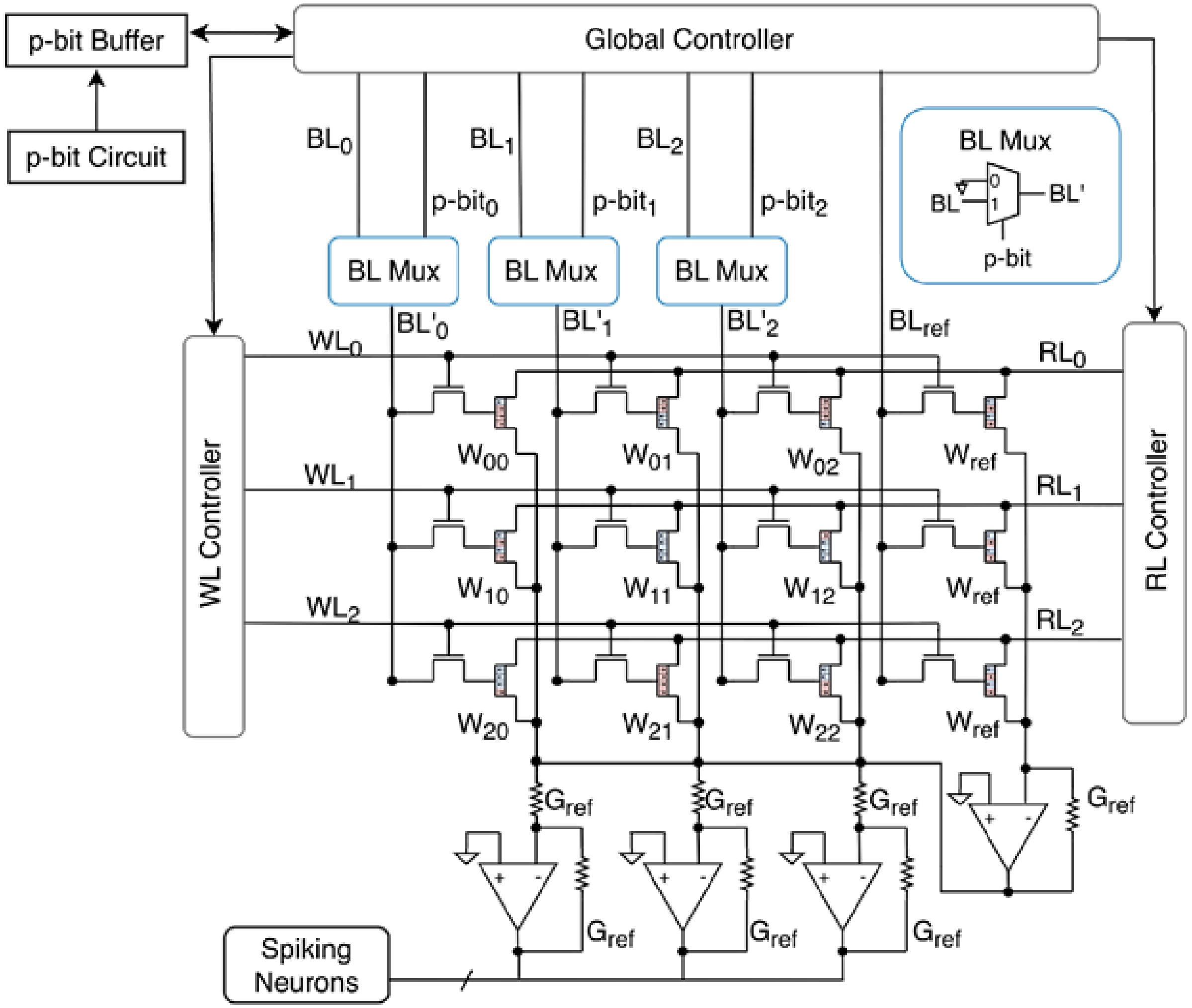


- Stochastic switching in response to a set pulse generates p-bit at the output
  - Output of the circuit, p-bit ~ Bernoulli (1-p_{switch})

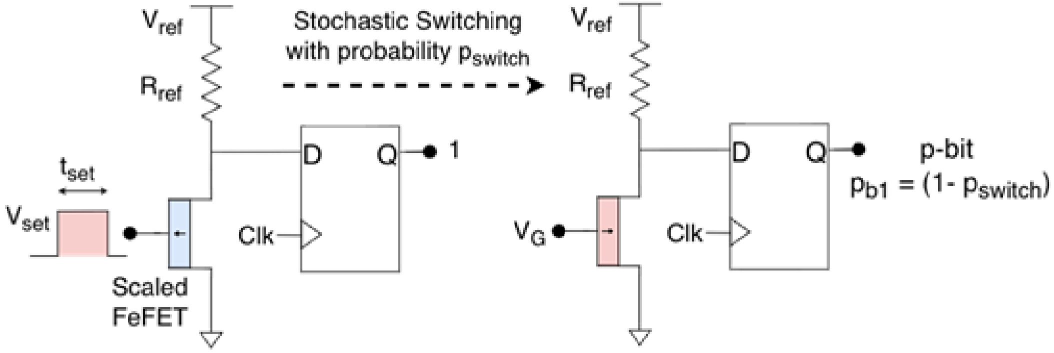
- The set pulse amplitude is calibrated to achieve a target switching probability [1,2]*

[1] Fatima Tuz Zohora Nicolas Ramos, Abinidhi Geethaikrishnan, Hai Li, and Dhireesha Kudithipudi. "Device-Algorithm Co-Design with FeFETs for On-device Continual Learning." *Proceedings of the Great Lakes Symposium on VLSI 2026*. 2026

Crossbar Architecture for On-device Continual Learning



- 1T-1FeFET crossbar architecture interfaces with the p-bit circuit to support parallel probabilistic weight updates
- Larger FeFETs with 16 discrete conductance states are used as synaptic weights



[1] Fatima Tuz Zohora, Vedant Karia, Nicholas Soures, Dhireesha Kudithipudi. "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks." Scientific Reports 14(1):1–16, 2024
[2] Fatima Tuz Zohora and Dhireesha Kudithipudi, "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks," U.S. Non-Provisional Patent Application, Atty. Dkt. No. 6693-17201, filed March, 2026.
[3] Fatima Tuz Zohora Nicolas Ramos, Abinidhi Geethaikrishnan, Hai Li, and Dhireesha Kudithipudi. "Device-Algorithm Co-Design with FeFETs for On-device Continual Learning." *Proceedings of the Great Lakes Symposium on VLSI 2026*. 2026.

Limited Device Endurance

- A finite number of reliable write cycles before the device degrades [3-5]
  - RRAMs can show endurance ~10^5 cycles, much lower compared to SRAMs (~10^{16} cycles)

- Reducing write frequency is critical to extending network lifetime

Endurance of different memory technologies [1,2]

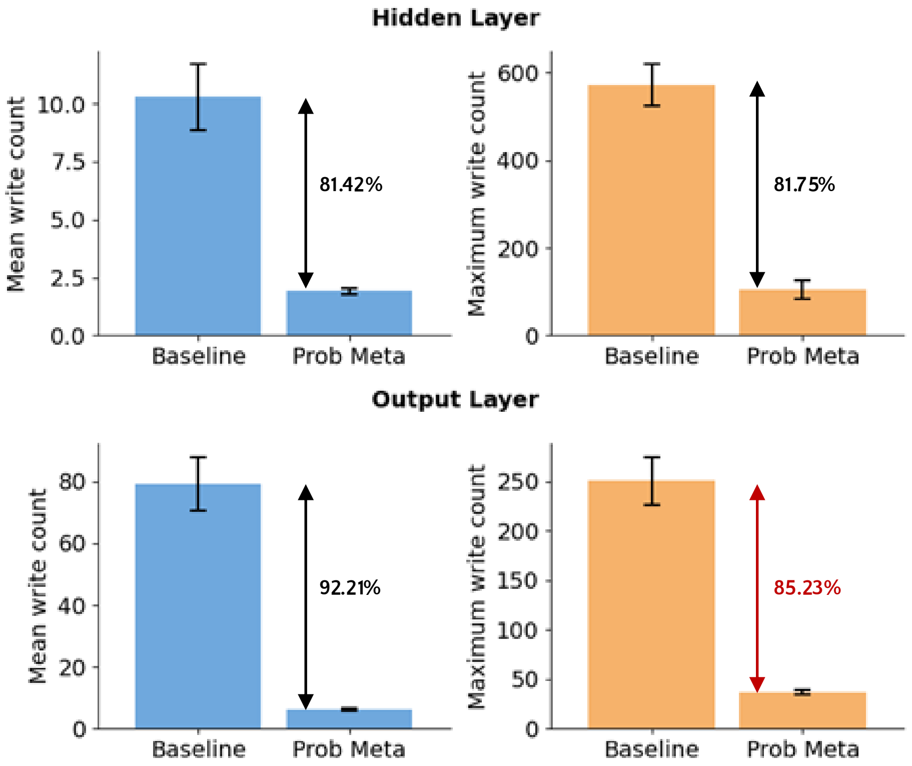
Memory	Endurance
Resistive Random Access Memory	> 10^5
Phase Change Memory	~10^7
Magnetic Random Access Memory	> 10^{12}
Static Random Access Memory	~ 10^{16}

[1] Lanza, Mario, et al. "The growing memristor industry." *Nature* 640.8059 (2025): 613-622.
[2] Lanza, Mario, et al. "Memristive technologies for data storage, computation, encryption, and radio-frequency communication." *Science* 376.6597 (2022): eabj9979.
[3] Zhang, Shuhang, et al. "Lifetime enhancement for rram-based computing-in-memory engine considering aging and thermal effects." *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2020.
[4] Zhang, Shuhang, et al. "Aging-aware lifetime enhancement for memristor-based neuromorphic computing." *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019.
[5] Xia, Lixue, et al. "Fault-tolerant training with on-line fault detection for RRAM-based neural computing systems." *Proceedings of the 54th Annual Design Automation Conference 2017*. 2017

Probabilistic Metaplasticity: Impact on Write Count

Probabilistic metaplasticity reduces likelihood of updating important weights

- >80% reduction in both metrics compared to the baseline model
- The output layer exhibits smaller reduction in maximum writes than in mean writes



Concentrated writes to a subset of weights can cause premature device wear-out.

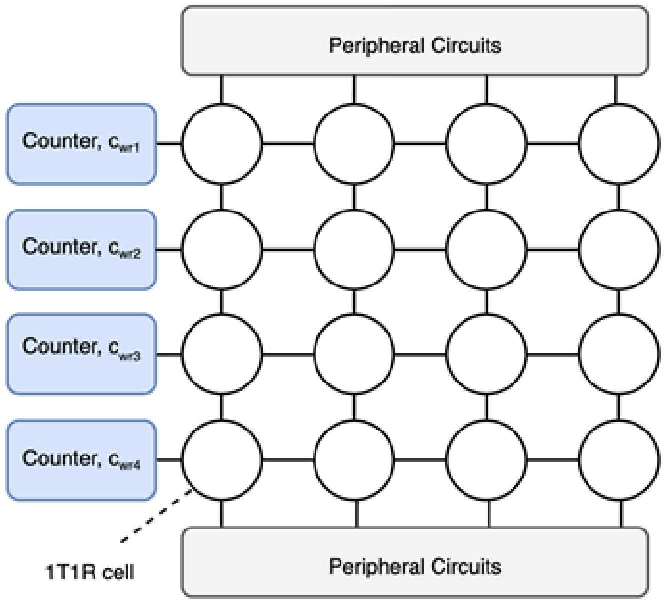
[1] Fatima Tuz Zohora, Vedant Karia, Nicholas Soures, Dhireesha Kudithipudi. “Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks.” Scientific Reports 14(1):1–16, 2024
[2] Fatima Tuz Zohora and Dhireesha Kudithipudi, “Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks,” U.S. Non-Provisional Patent Application, Atty. Dkt. No. 6693-17201, filed March, 2026.
[3] Fatima Tuz Zohora Nicolas Ramos, Abinidhi Geethaikrishnan, Hai Li, and Dhireesha Kudithipudi. "Device-Algorithm Co-Design with FeFETs for On-device Continual Learning." *Proceedings of the Great Lakes Symposium on VLSI 2026*. 2026

Wear-Aware Probabilistic Metaplasticity

- Incorporate device wear effects in the probabilistic metaplasticity model [1]
- The update probability of a weight is given by:

$$p'_{update}(w) = p_{update}(w) * p_{wear}(c_{wr})$$

p_{wear} reduces the likelihood of updating weights in heavily written rows

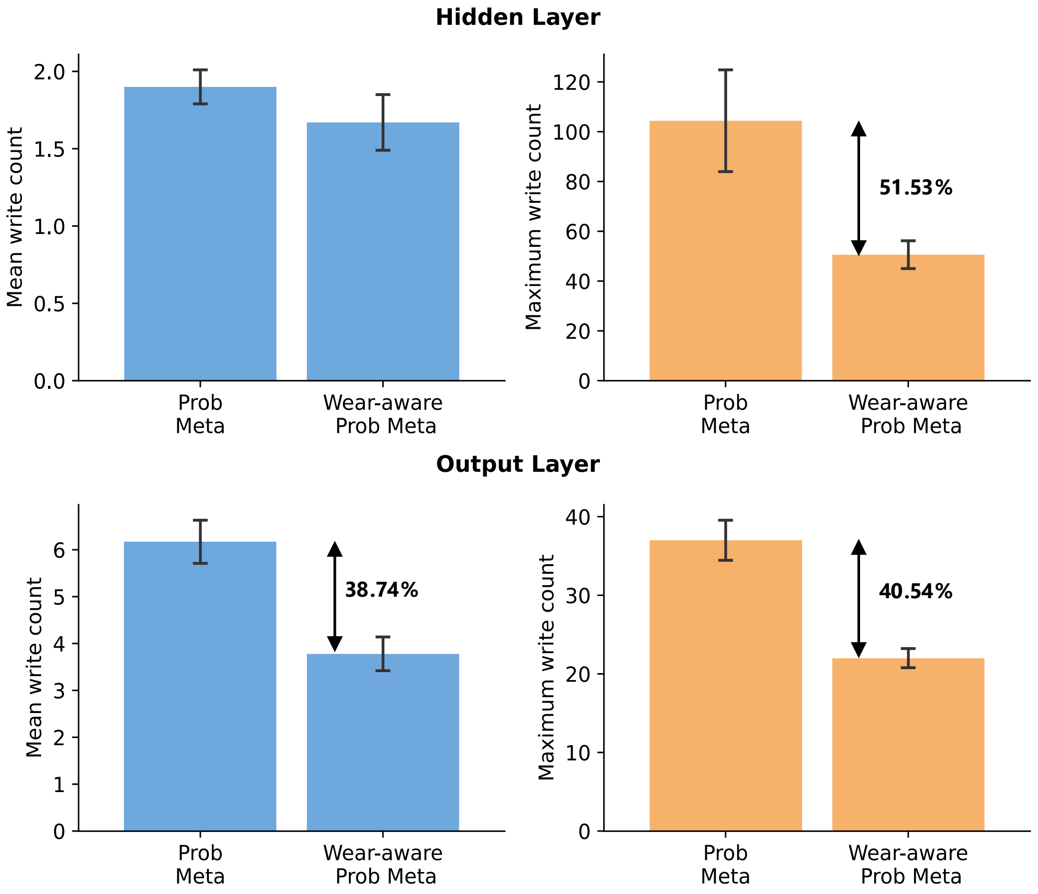


[1] Fatima Tuz Zohora, Vedant Karia, Nicholas Soures, Dhireesha Kudithipudi. "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks." Scientific Reports 14(1):1–16, 2024
[2] Fatima Tuz Zohora and Dhireesha Kudithipudi, "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks," U.S. Non-Provisional Patent Application, Atty. Dkt. No. 6693-17201, filed March, 2026.

Wear-Aware Probabilistic Metaplasticity

- Reduction in maximum write count on both layers
- Maintains comparable mean accuracy across tasks

Method	Mean Accuracy on Split-MNIST
Probabilistic Metaplasticity	82.43 ± 0.89
Wear-aware Probabilistic Metaplasticity	82.34 ± 0.83



[1] Fatima Tuz Zohora, Vedant Karia, Nicholas Soures, Dhireesha Kudithipudi. "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks." Scientific Reports 14(1):1–16, 2024
[2] Fatima Tuz Zohora Nicolas Ramos, Abinidhi Geethaikrishnan, Hai Li, and Dhireesha Kudithipudi. "Device-Algorithm Co-Design with FeFETs for On-device Continual Learning." *Proceedings of the Great Lakes Symposium on VLSI 2026*. 2026.
[3] Fatima Tuz Zohora and Dhireesha Kudithipudi, "Probabilistic Metaplasticity for Continual Learning with Memristors in Spiking Networks," U.S. Non-Provisional Patent Application, Atty. Dkt. No. 6693-17201, filed March, 2026.

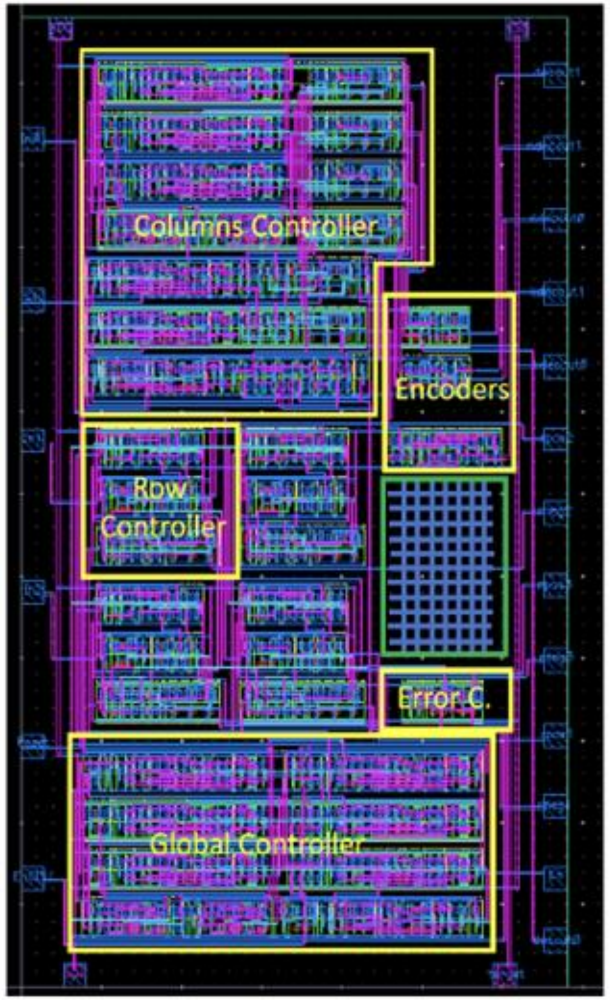
Neuromorphic Continual Learning Accelerators

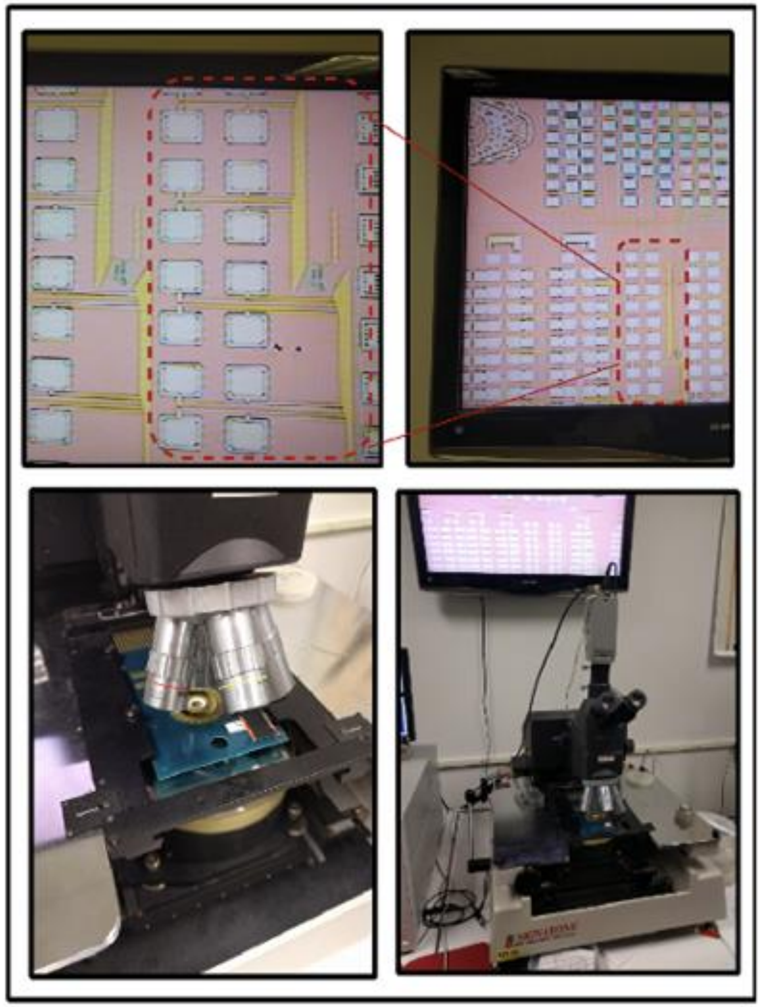
Dr. Vedant Karia

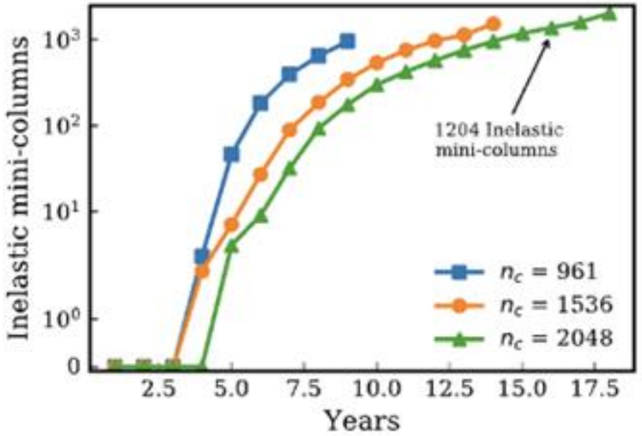
Dr. Fatima Tuz Zohora

Dr. Abdullah Zyarah

Memristor/CMOS Accelerator with Neurogenesis

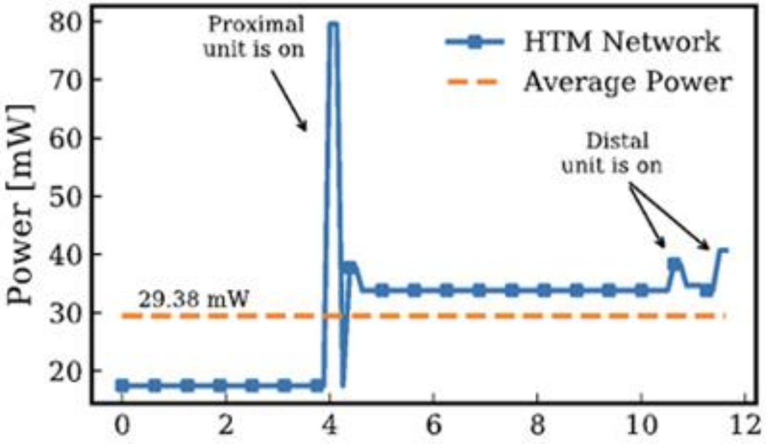






Memristor device endurance 10^6 - 10^{12}

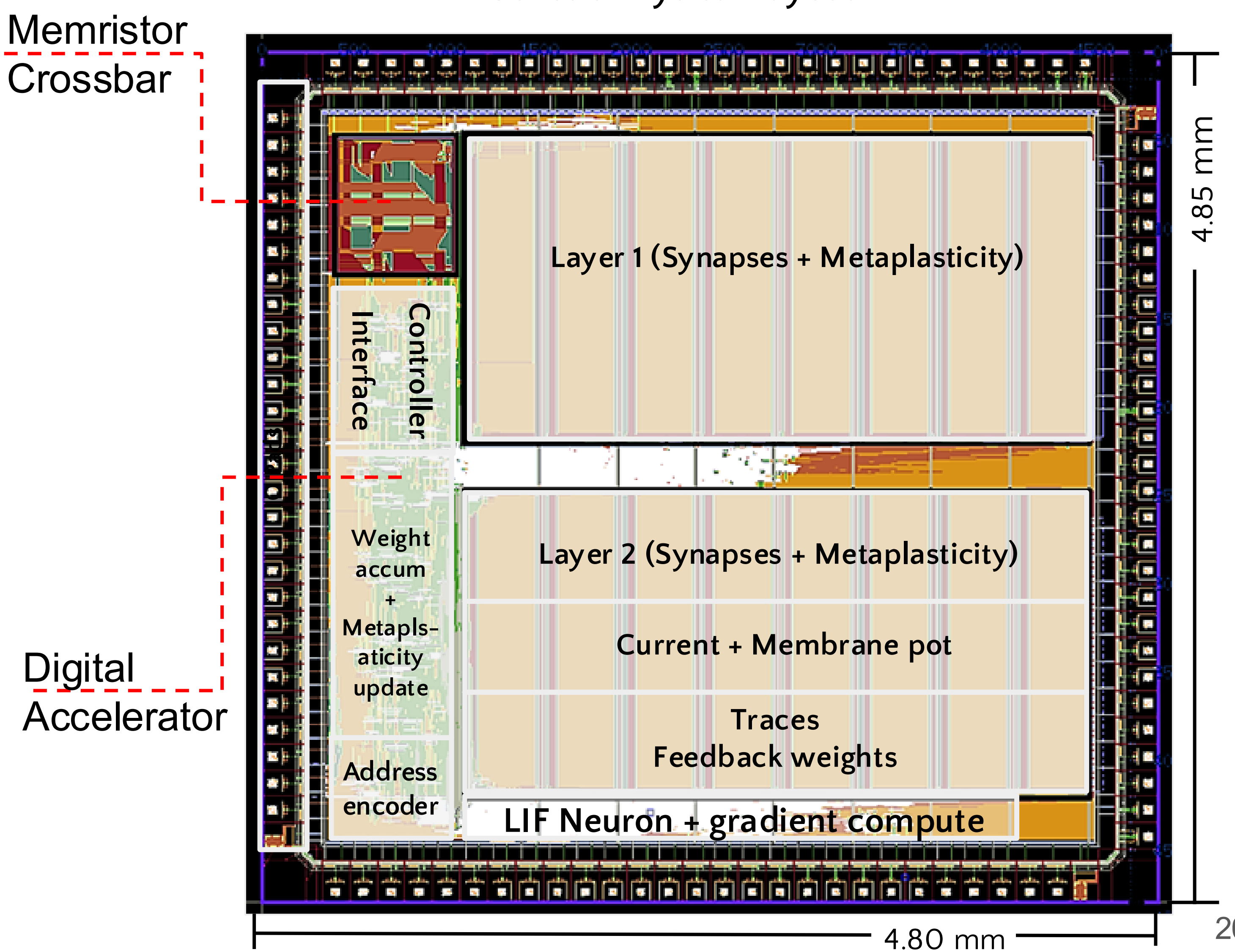
> 8 years of successful learning @ rate of 10 ms



[1]Zyarah, Abdullah M., Kevin Gomez, and Dhireesha Kudithipudi. "Neuromorphic system for spatial and temporal information processing." *IEEE Transactions on Computers* 69.8 (2020): 1099-1112.

Genesis (2.0)

Genesis Physical Layout



[1] Vedant Karia and Dhireesha Kudithipudi, "SPIKING NEUROMORPHIC ACCELERATOR FOR ON-DEVICE CONTINUAL LEARNING," U.S. Non-Provisional Patent Application, Atty. Dkt. No. 6693-17101, filed June, 2026.

US Hub for Neuromorphic Computing – The Neuromorphic Commons

First-of-its-kind, THOR enables richer understanding of computational models, AI/ML algorithms, neuroscience and neuromorphic hardware.

Housed at **UTSA** with **close-knit industry partnerships**

UTSA

[matrix]
The UTSA AI Consortium for Human Well-Being

LEAD INSTITUTION

THE UNIVERSITY OF TENNESSEE KNOXVILLE

UC San Diego

HARVARD UNIVERSITY

COLLABORATING INSTITUTIONS

neuromorphiccommons.com

thor@utsa.edu

Award **# 2346527**

Take Aways

- **Continual learning is a grand challenge**

- **Continued adoption of neural plasticity mechanisms in AI models**
  - Spiking networks are nascent, but offer lot of potential with local plasticity and temporal learning
  - Opportunities for synergy when combining plasticity models, existing and future (compositionality)

- **Enhance capabilities of (Neuromorphic) AI Hardware**
  - Conundrum with AI hardware and models, can be addressed by adopting incremental advances into silicon
  - Emerging technologies offer unique solutions,
    - One should be careful not to fall into the trap of **New devices = New AI systems**

"Coming together is a beginning, staying together is progress, and working together is success." - Henry Ford

Acknowledgements

- Researchers and collaborators in the NUAI Lab.
- External Collaborators on THOR
  - Dr. Catherine Schuman, UTK
  - Dr. Gert Cauwenberghs, UCSD
  - Dr. VJ, Harvard

[matrix] The UTSA AI Consortium for Human Well-Being

Argonne NATIONAL LABORATORY

DARPA

Lifelong Learning Machines (L2M) by DARPA MTO

Sandia National Laboratories

SAN ANTONIO MEDICAL FOUNDATION

NSF

MEERCAT Center

NUAI LAB

NUAI Lab Team

Thank you

dk@utsa.edu