NUAI LAB

Energy-Efficient Continual Learning Systems

Dhireesha Kudithipudi, PhD

Semiconductor Industry Energy Efficiency Scaling (EES2) Technical Workshop

September 14, 2022

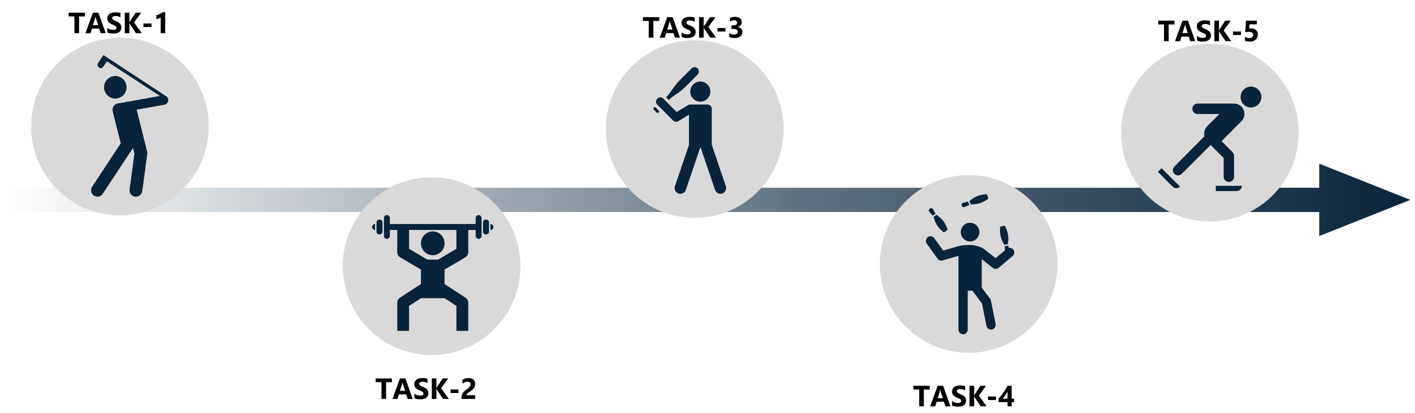
UTSA

The University of Texas at San Antonio™

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

Continual Learning/Lifelong Learning

- **Rapid adaptation** to new conditions without forgetting previous knowledge
- **Continuous improvement** in performance while executing previously learnt tasks and novel tasks



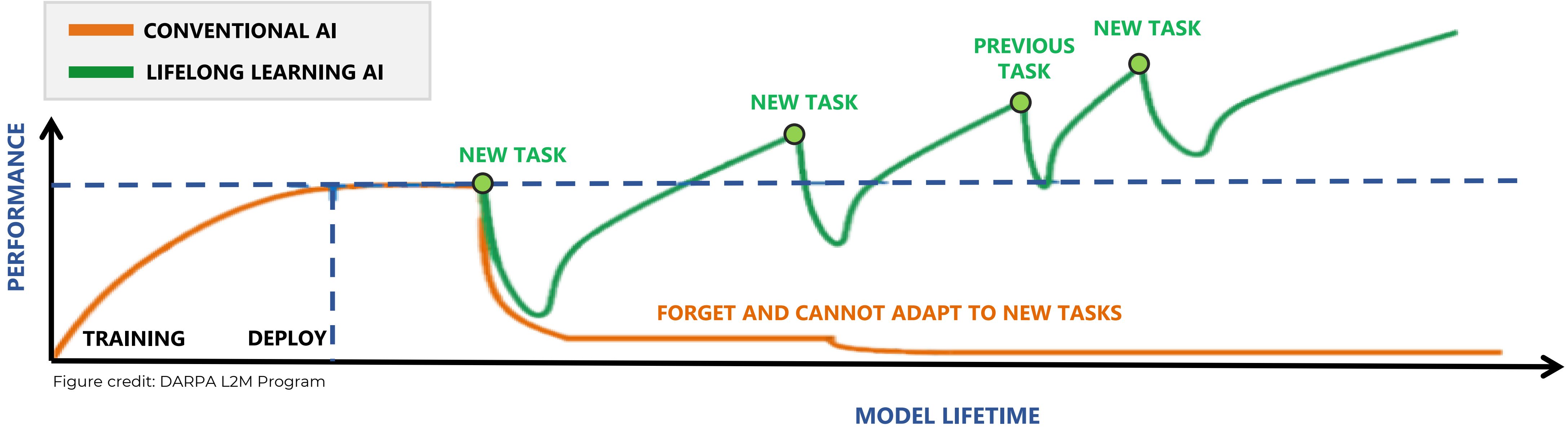
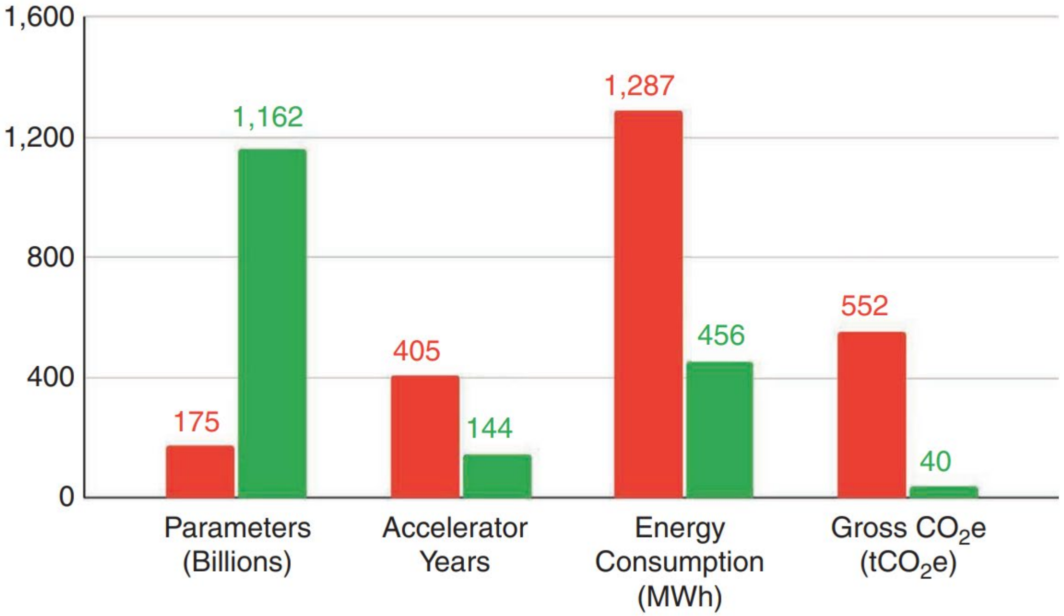


Figure credit: DARPA L2M Program

To Learn More: Kudithipudi, D., Aguilar-Simon, M., Babb, J. et al. Biological underpinnings for lifelong learning machines. Nature Mach Intell 4, 196–210 (2022).

Current State of **AI Compute Hardware**

A paradigm shift is needed in compute architectures to sustain the growing AI compute demands



OpenAI Five (DOTA 2 video game)

- Training equivalent ~ 45,000 years
- Operations ~ 770 PFLOP/s-days

Dactyl (Deep RL)

- 2.8GWh of electricity
- 100 years of training

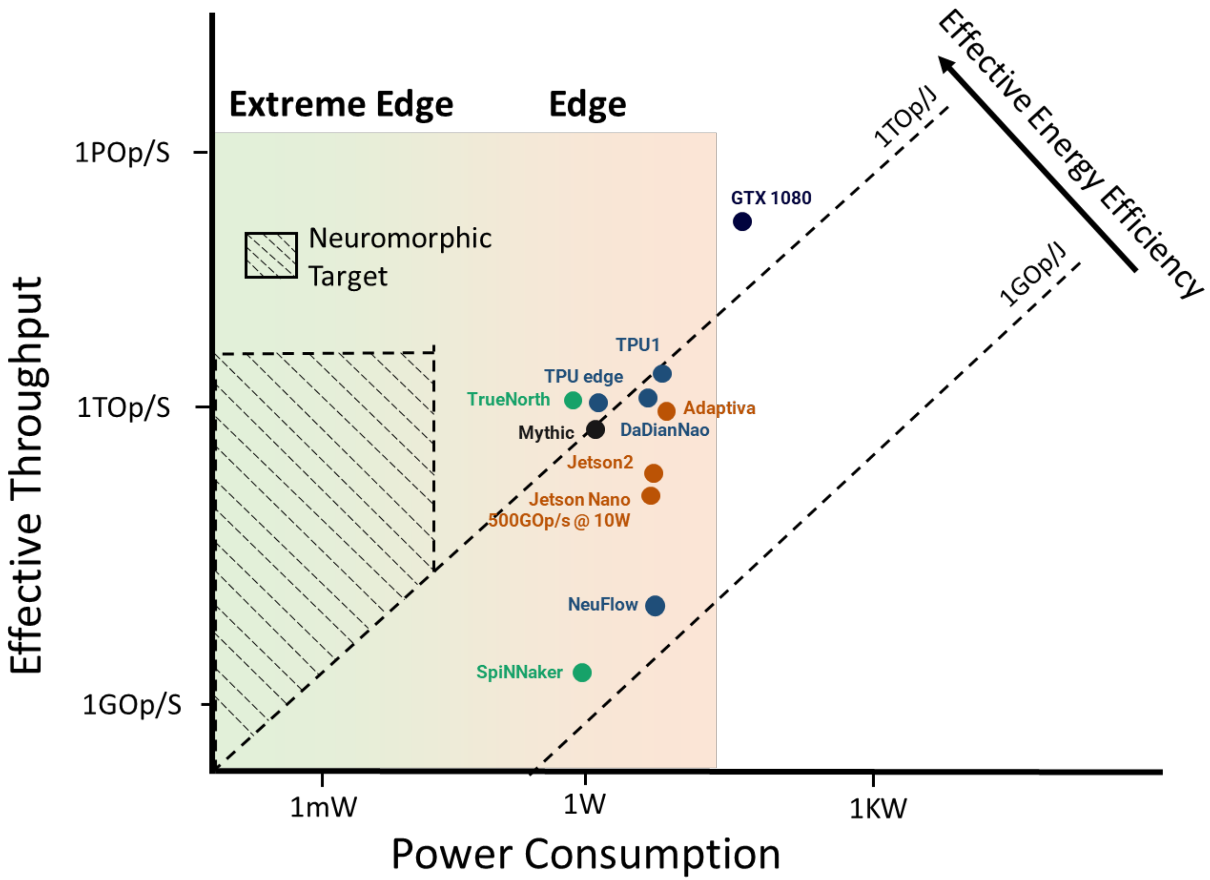
- **AI compute is doubling ~3.5 months**
- **Compute cost decreasing by 1 order ~4-12 years**
- **Training cost of top models ~$10M**

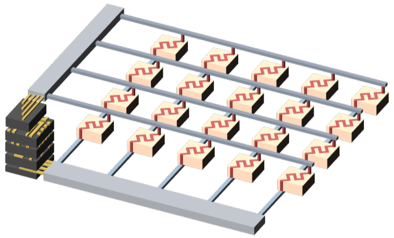
D. Patterson et al., "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," in Computer, vol. 55, no. 7, pp. 18-28, July 2022

OpenAI.com

Kudithipudi, Nature Conference on AI, Neuroscience & Hardware,2022

Current State of **AI Compute Hardware**





Neuromorphic Accelerators

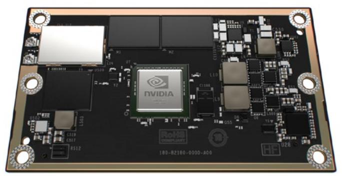
20.7x more energy efficient

45.2x less power consumption

1.41x faster inference

Co-located memory and computation

NVIDIA Jetson TX1



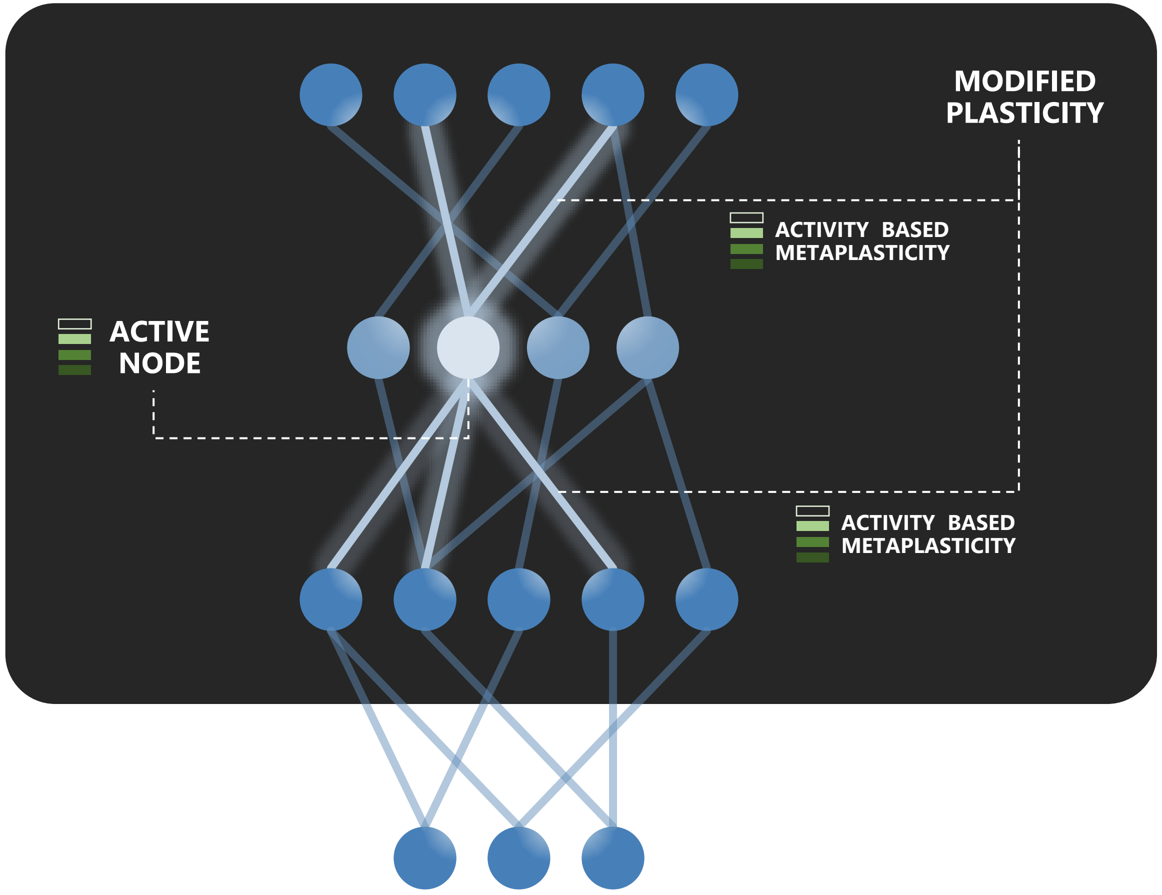
A. Reuther,et al., "Survey of Machine Learning Accelerators," *2020 IEEE HPEC*

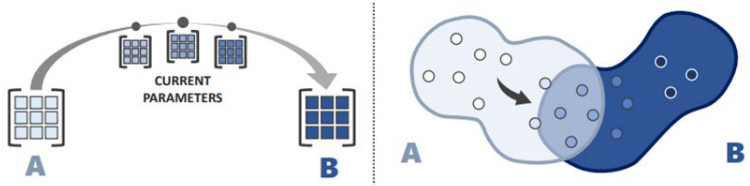
Shih-Chii Liu, “Edge AI with Neuromorphic Spiking Sensors”

Blouw, Peter, et al. "Benchmarking keyword spotting efficiency on neuromorphic hardware." NICE. 2019.

Example Plasticity Mechanism for Lifelong Learning

Metaplasticity: Protects previous knowledge encoded in important synapses to reduce catastrophic forgetting





- **Metaplasticity** can be based on
  - Activity
  - Weight Change
  - Temporal Correlation
  - STDP

- Plasticity of previously active neurons is reduced to prevent them from being overwritten.

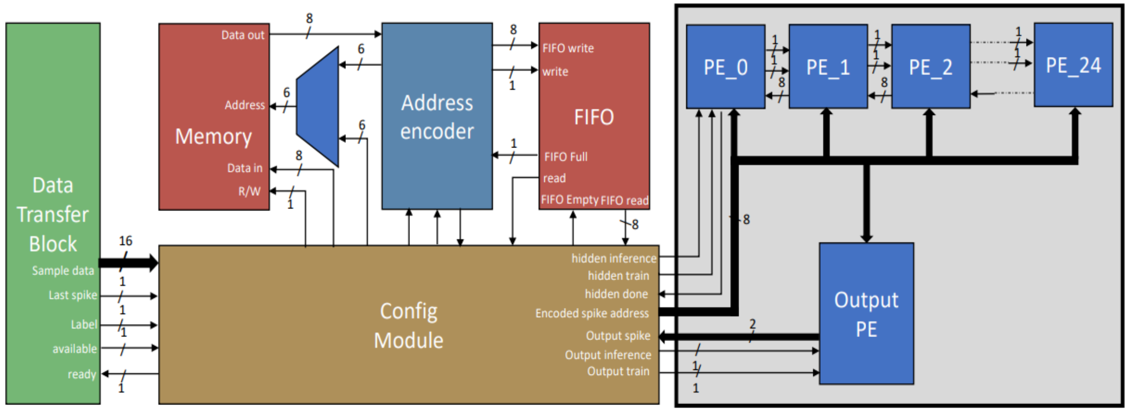
- Plasticity of inactive neurons is increased to allow them to respond to multiple stimuli.

$$Plasticity = \exp(-abs(m * w))$$

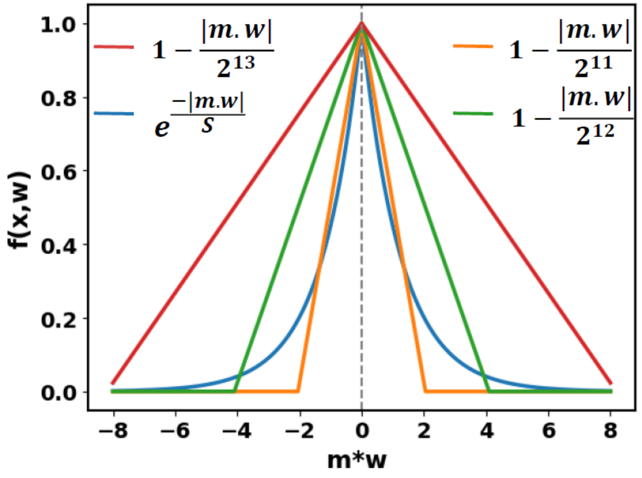
N. Soures, P. Helfer, A. Daram, T. Pandit, and D. Kudithipudi, “TACOS: Task agnostic continual learning in spiking neural networks,” ICML Workshop on theory and foundations in Continual Learning, 2021.

N. Soures and D. Kudithipudi. “Exploring the Plasticity-Stability Trade-Off in Spiking Neural Networks,” in CCN, 2022.

Example **Accelerator** with **Metaplasticity**

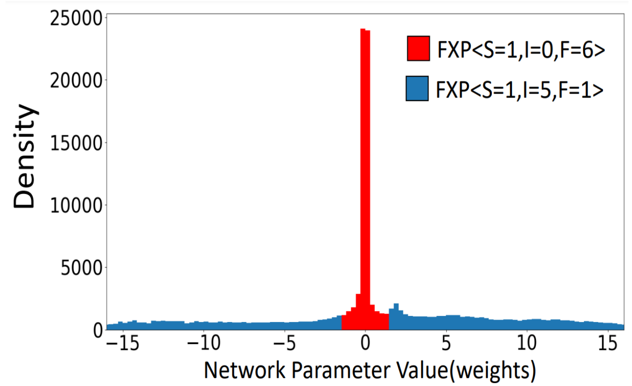


Architecture of the digital accelerator



Options for hardware friendly bi-linear metaplasticity functions

- **Systolic array p**rocessing elements(PEs) for parallel processing and efficient data movement
- Each PE incorporates bi-linear metaplasticity function to enable lifelong learning, reducing the computational overhead
- Model parameters represented with 8-bit dual fixed-point to reduce quantization error and memory footprint

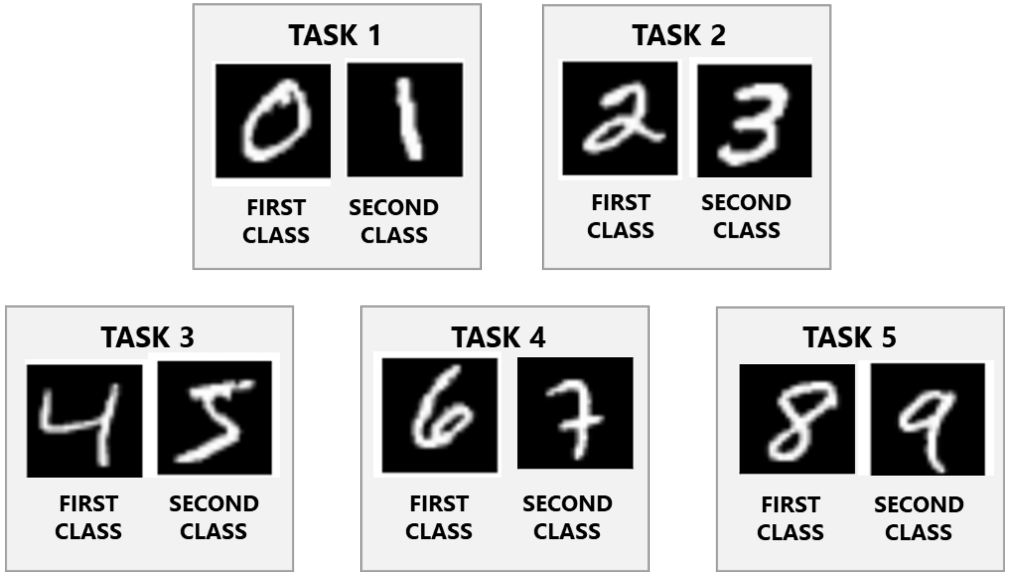


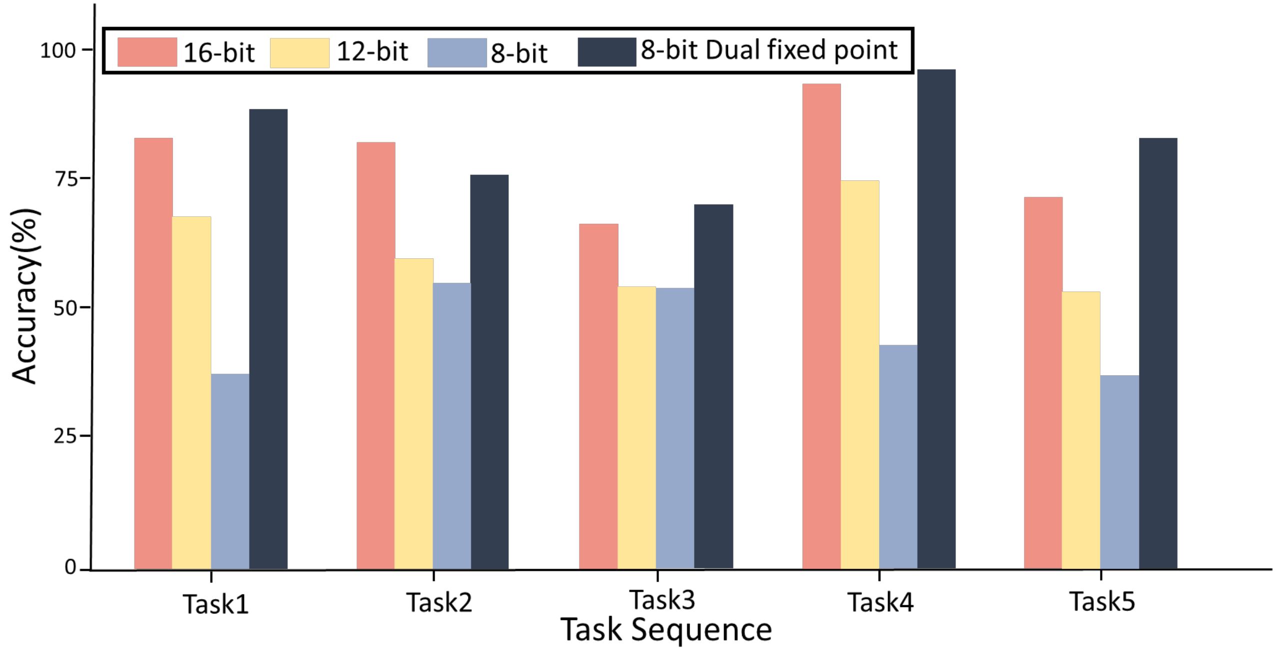
Model parameter distribution split matching dual fixed-point representation

V. Karia, F. T. Zohora, N. Soures, and D. Kudithipudi, “SCOLAR: A spiking digital accelerator with dual fixed point for continual learning.” 2022 IEEE ISCAS

Accelerator Performance

SPLIT MNIST DATASET





Lifelong Learning Performance on Split MNIST benchmark in **Domain IL scenario**

Realtime training at 30 images/sec

1.573W power consumption on Xilinx ZYNQ FPGA

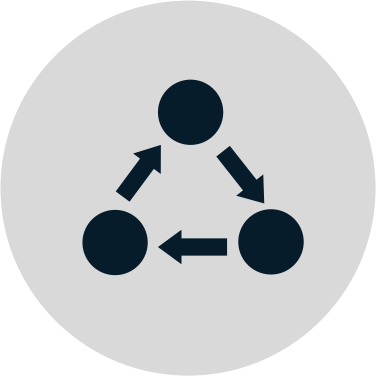
2MB on-chip SRAM memory

Near SOTA accuracy on **Split MNIST** dataset with dual fixed-point representation

V. Karia, F. T. Zohora, N. Soures, and D. Kudithipudi, “SCOLAR: A spiking digital accelerator with dual fixed point for continual learning.” 2022 IEEE ISCAS

Research in **Neuromorphic AI Lab**

LIFELONG LEARNING



- Neural Plasticity
- Dynamic Algorithms

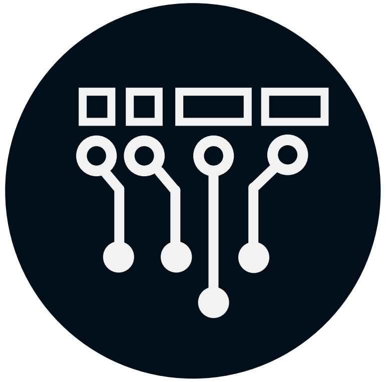
NEURALLY INSPIRED ALGORITHMS

- Spiking Neural Networks
- Rate Based Neural Networks

NEUROMORPHIC SYSTEMS

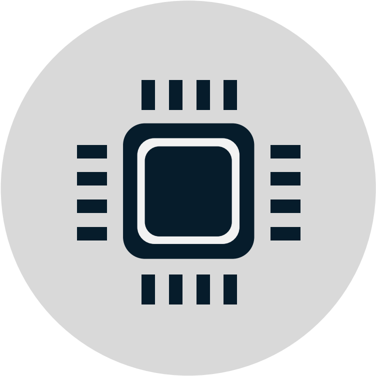
- Memristors
- Digital & Mixed Signal Design

ENERGY EFFICIENT ML



- Model Compression
- Low Precision formats
- Circuit & Architecture Optimization

EMERGING TECHNOLOGIES



- 3D
- FeFETs

Acknowledgements

AIR FORCE RESEARCH LABORATORY

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

SEAGATE

Argonne NATIONAL LABORATORY

DARPA

Lifelong Learning Machines **(L2M)** by **DARPA MTO**

Sandia National Laboratories

SAN ANTONIO MEDICAL FOUNDATION

NSF

Thank You

e-mail: dk@utsa.edu

www.nuailab.com