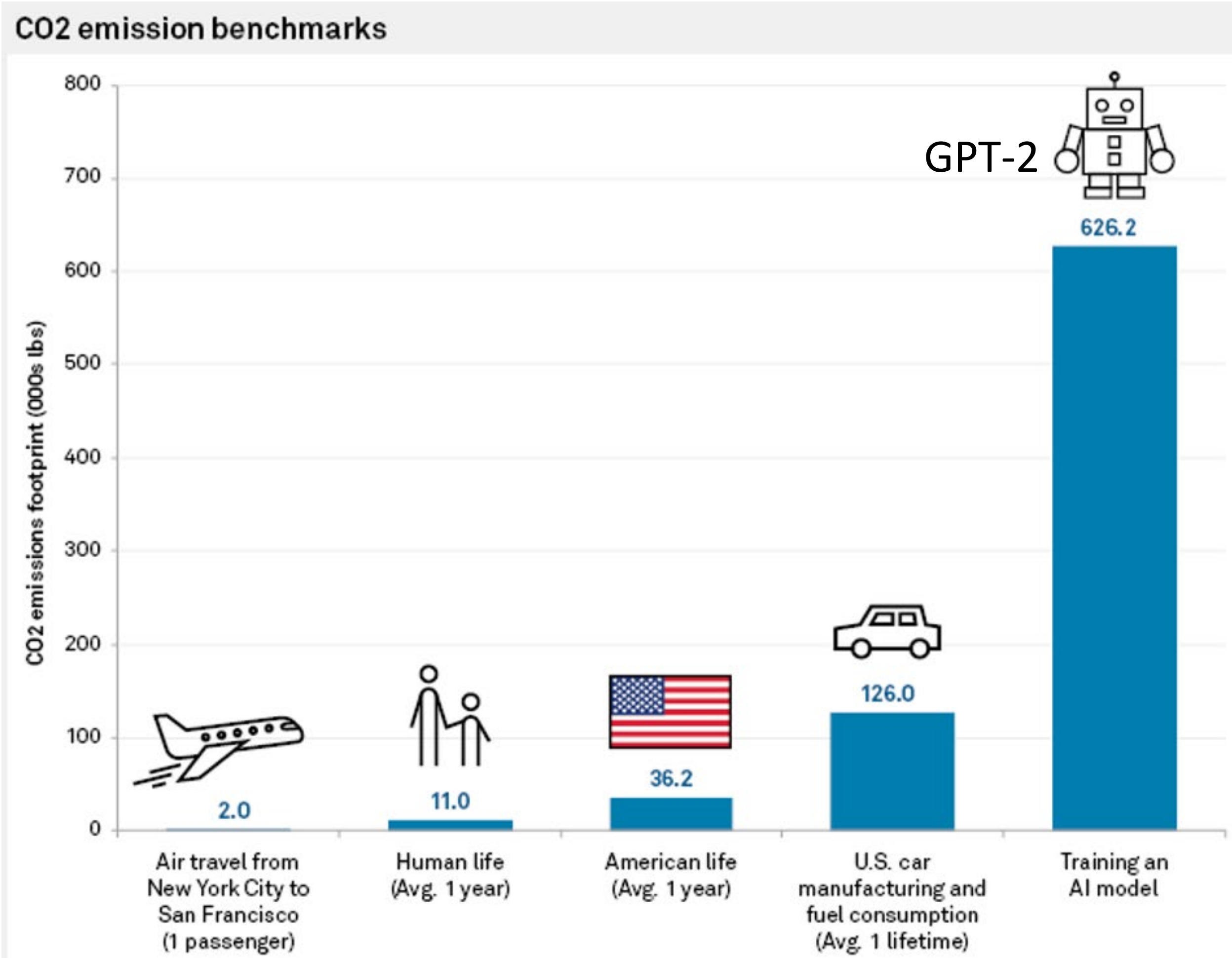
SEA-CROGS: THRUST III SPIKING NEURAL NETWORKS

- **Priya Panda, Yale University**
- Panos Stinis (PNNL)
- George Karniadakis (Brown/ PNNL)
- Thomas Serre, Jerome Darbon (Brown)
- Eric Cyr (SNL)

AI & Environmental Sustainability

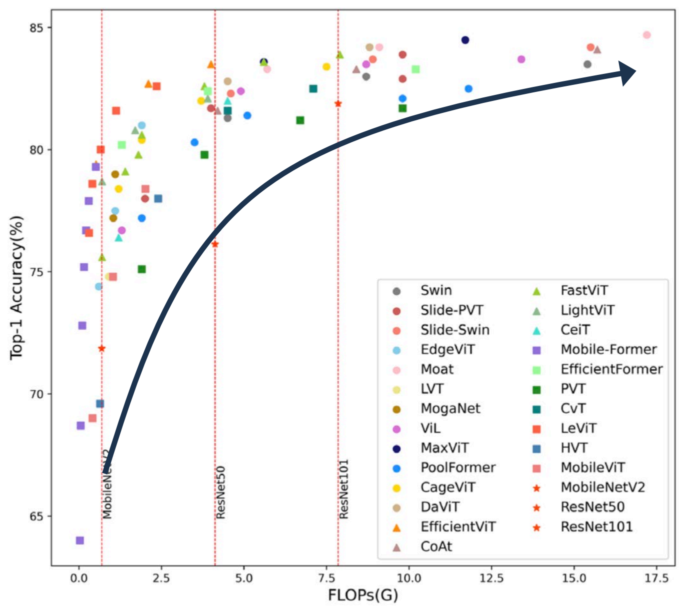


Source: Forbes article, Nov. 2022

Scaling Laws with No Bounds

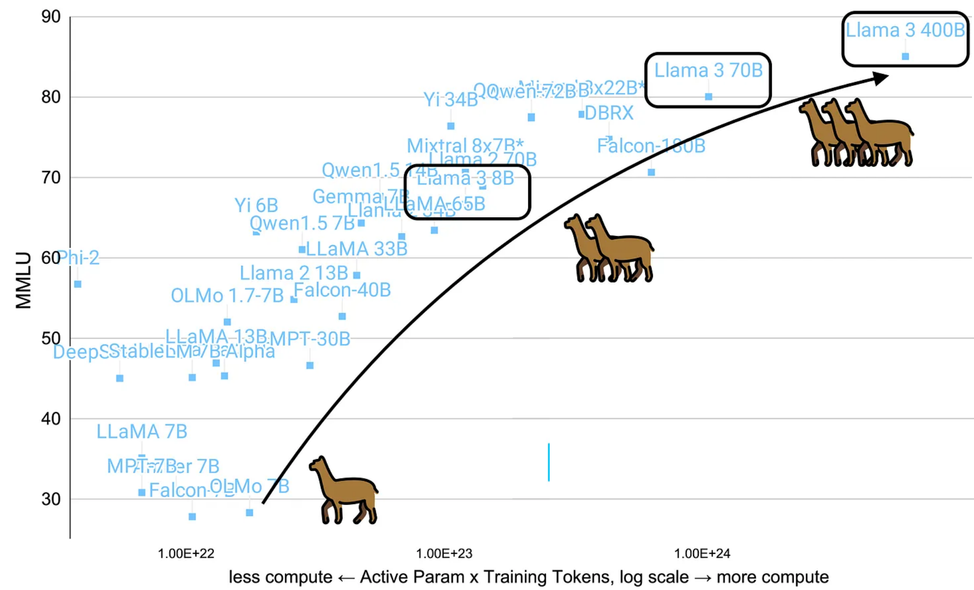
- Scale Argument– More FLOPs, Higher performance

Vision: ImageNet-1k Benchmark Accuracy vs. FLOPs



Lee et al., Vision transformer models for mobile/edge devices: a survey, Multimedia Systems (2024)

NLP: MMLU Benchmark Performance vs. Computing cost

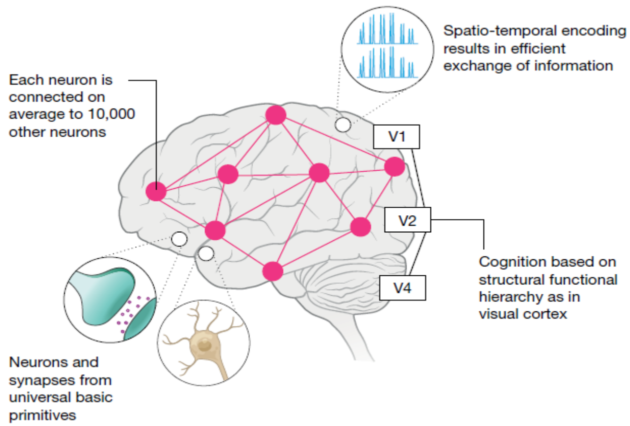


https://www.interconnects.ai/p/llama-3-and-scaling-open-llms

SEA-CROGS

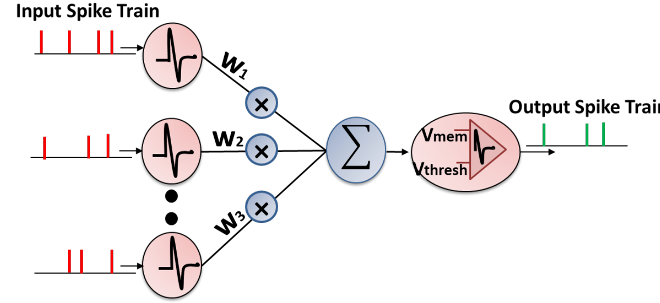
Approaching Sustainability with Spike-based Machine Intelligence

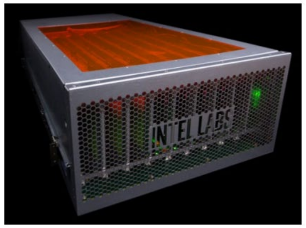
Human Brain

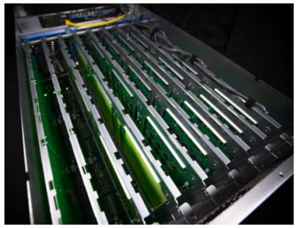


- Performing impressive feats with a power budget of nearly **20 W**
- Spike-driven communication
- Co-located neurons and synapses

Spiking Neural Network (SNN)







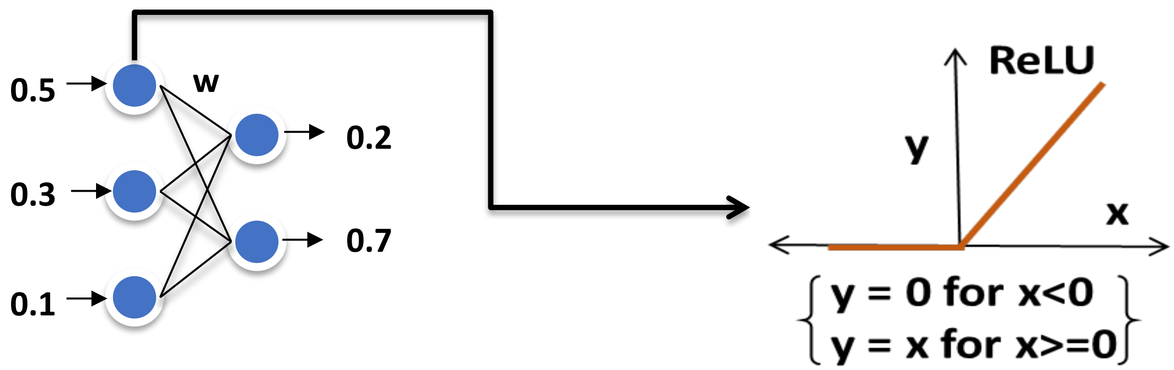
Neuromorphic Hardware

- **Use Spiking Neural Networks as a means to integrate brain-inspired cues to harness energy-efficiency as well as improved learning capability**
- **Use Neuromorphic Hardware (Intel Loihi, SpiNNaker) for more efficient computations**

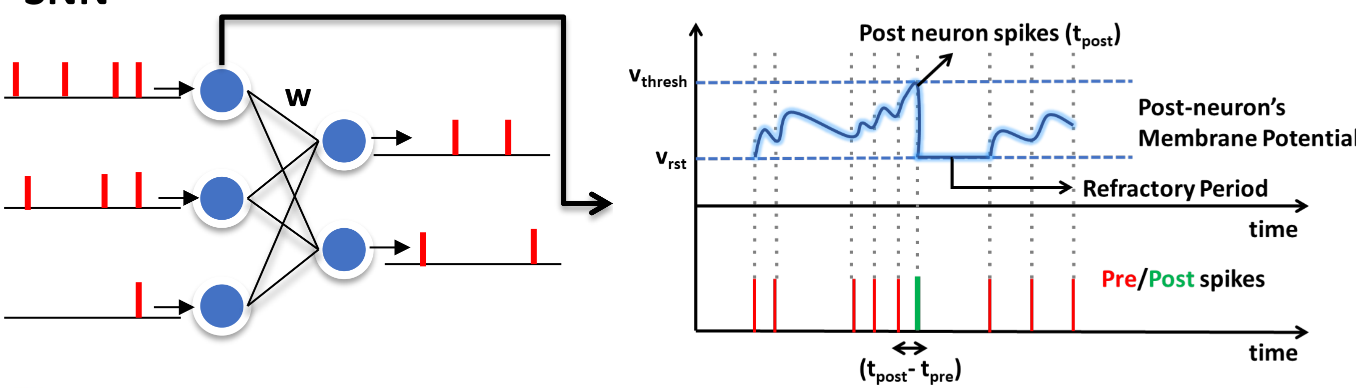
SEA-CROGS

SNN vs. ANN: Fundamental Differences

ANN



SNN



Features

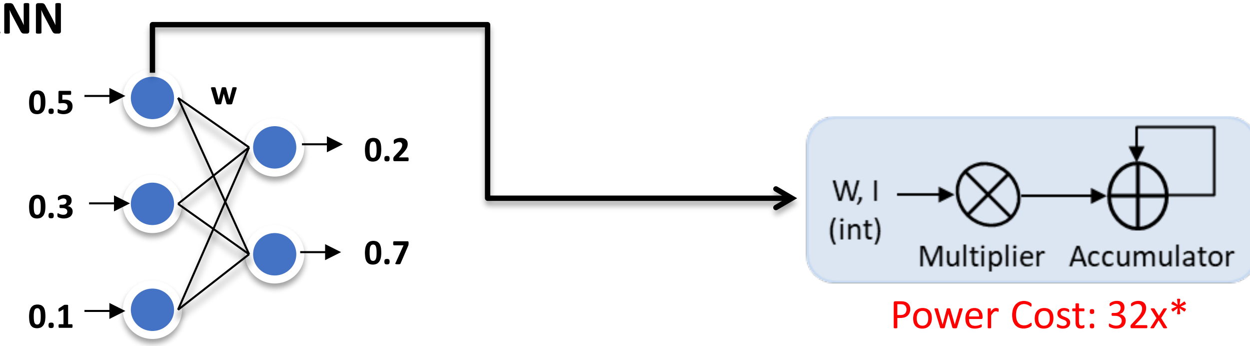
(+) High Performance
(+) Various Applications
(+) Easy to train with gradient descent and platform support

(+) Spatio-Temporal Encoding
(+) Activation Sparsity (>90%)

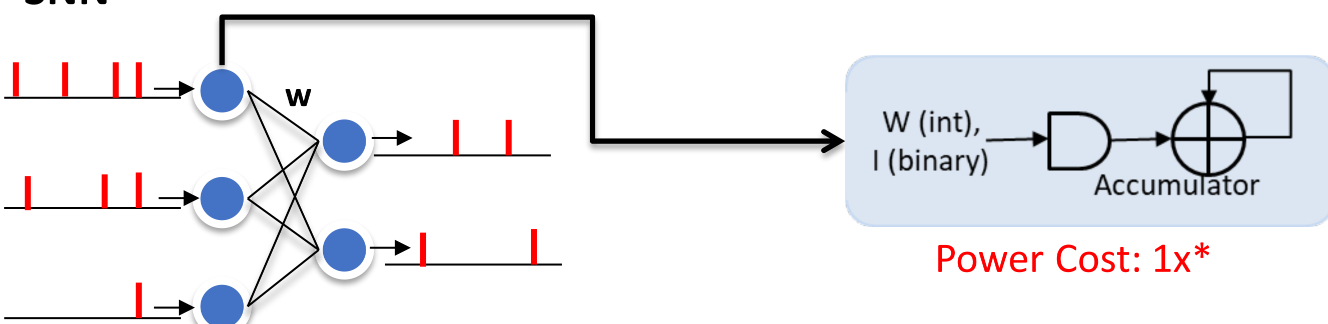
(-) Training with surrogate gradients and customized function writing on Pytorch, JAX

SNN vs. ANN: Fundamental Differences

ANN



SNN



Features

(+) High Performance
(+) Various Applications

(-) High computational cost with FP/INT Multiplier

(+) Low computational cost – Multiplier-less

→ **Energy-efficient and suitable for Edge AI**

SEA-CROGS

* TSMC 65nm technology

Challenge I: Can we efficiently train deep SNNs?

SNN Training

STDP Learning

Pros: Unsupervised local learning

Cons: Limited accuracy and shallow networks

SOTA	MNIST Accuracy
Lee *et al.* TCDS 2018	91.10%

ANN-SNN Conv

Pros: Takes advantage of standard ANN training

Cons: Long time-steps

SOTA	ImageNet Accuracy
Li *et al.* ICML 2021 (UESTC)	75.45%

Backprop in SNN

Pros: Lower Time-steps

Cons: Limited scalability, Discontinuous spike activities

SOTA	ImageNet Accuracy
Guo *et al.* NeurIPS 2022 (Peking)	70.65%

Forwardprop in SNN

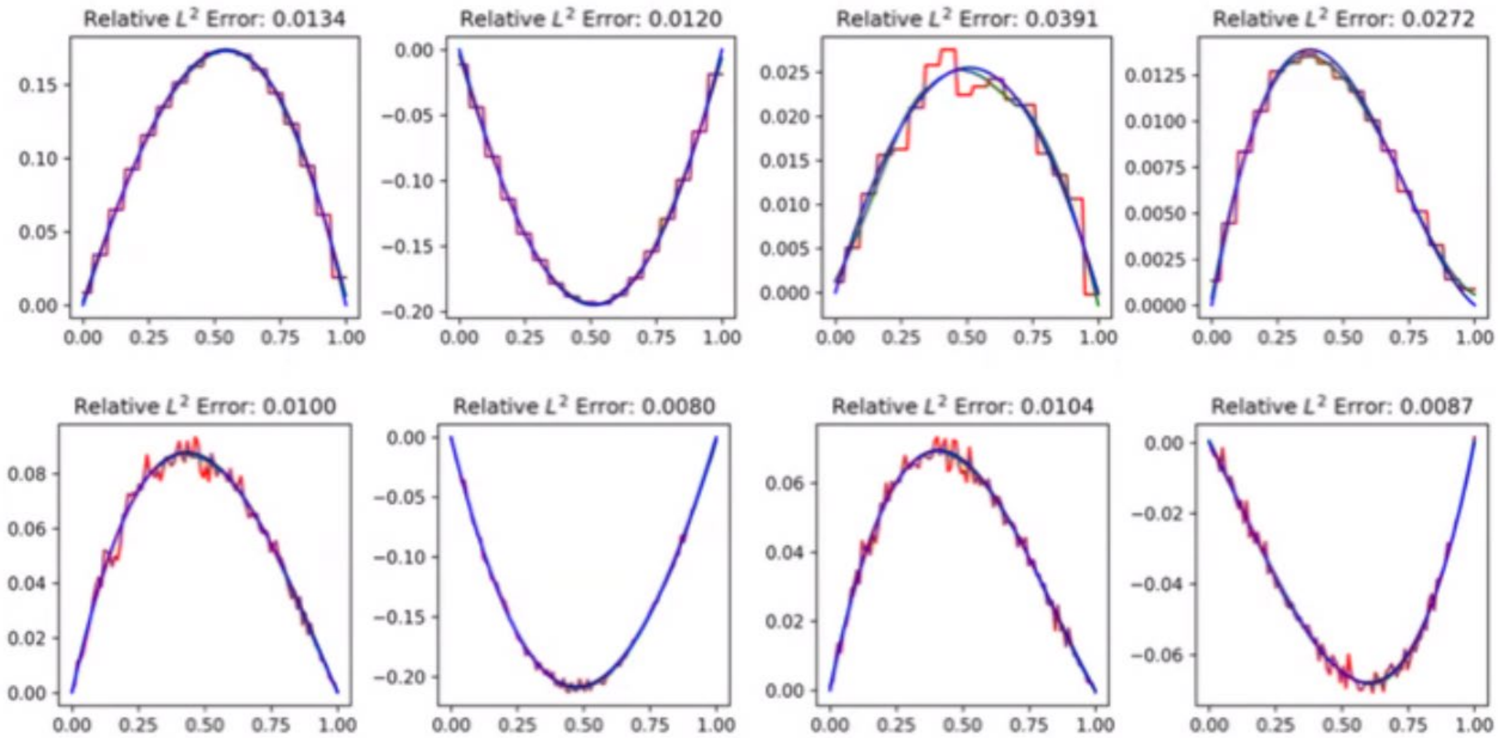
Pros: Lower Time-steps, No Backprop!!

Cons: Limited Scaling

SOTA	DeepONet
Zhang et al (Brown)	Regression Tasks

Training from scratch with Backprop Through Time or Forwardprop leverages time statistics – **Efficiency, Accuracy, Robustness**

Challenge II: Loss of Accuracy on Regression Tasks

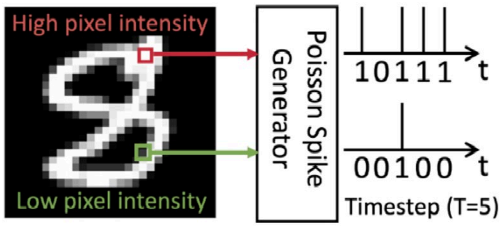


Spiking Neural Operators on 1D Poisson Equation

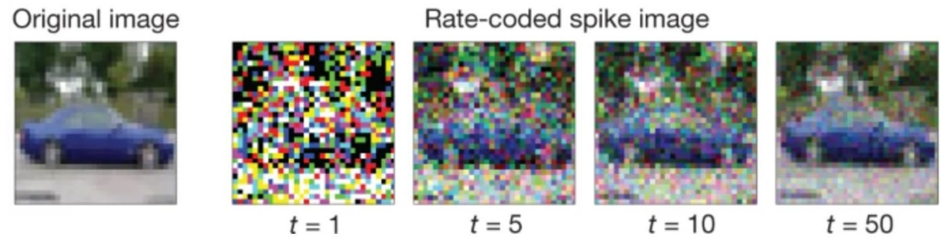
SEA-CROGS

Challenge III: Input Coding Techniques

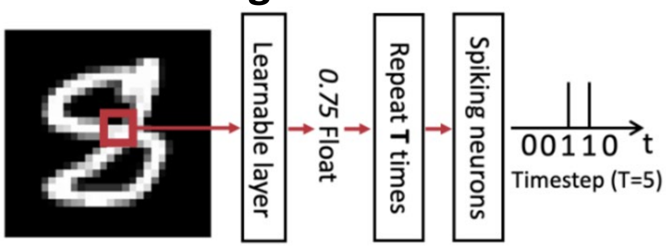
Rate coding



Spike frequency represents information

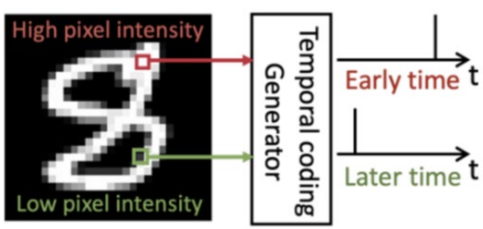


Direct coding



Learnable encoding layer before LIF neurons

Temporal coding



Early spike – important information
Late spike – less important information

SEA-CROGS

Thrust III: SNN Research Overview

Research subThrust 1

Fundamental training algorithms for improving SNNs

Spatio-Temporal Spiking Transformer

Query

1	0	1
0	0	0
1	1	0

Key

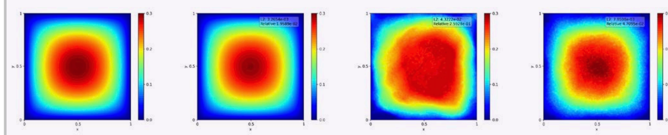
0	1	1
0	1	0
0	0	1

Value

1	0	0
0	0	1
1	1	1

Attention across space and time for improved performance at lower costs

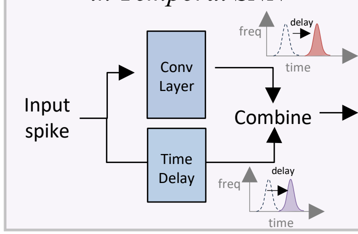
SNN conversion for PINN



(a) Reference solution (b) PINN result (c) Conversion w/o calibration (d) Conversion w/ calibration

Convert Pre-trained PINN with spiking neurons

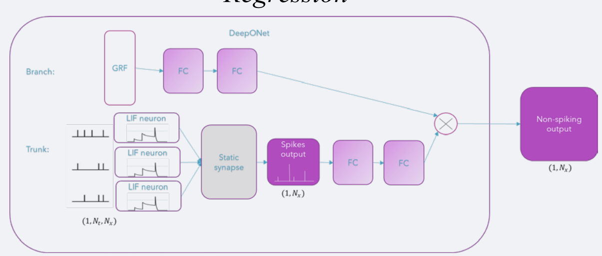
Skip Connection Architecture in Temporal SNN



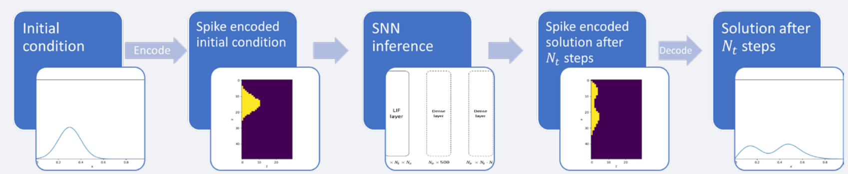
Research subThrust 2

Sci-ML application with SNNs

Regression



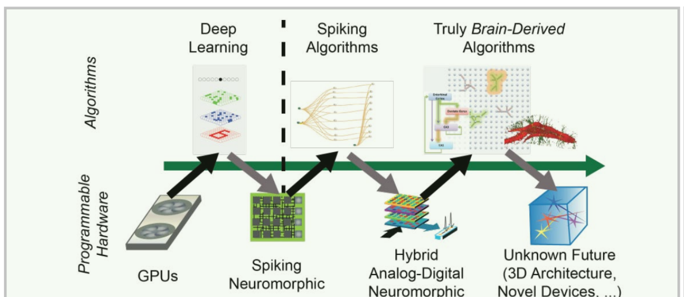
PDE solver



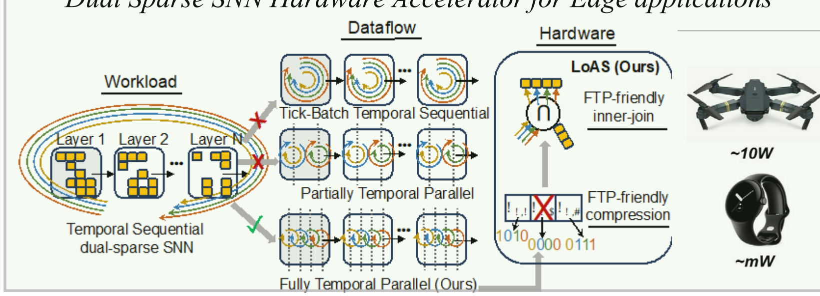
Physics informed machine intelligence

Research subThrust 3

Hardware-Algorithm Co-design based on Loihi Platform



Dual Sparse SNN Hardware Accelerator for Edge applications



Application Benefits and Cost analysis on Real Hardware

Algorithm driven applications & establishing new SOTA

Till Year 2 Goals & Accomplishments

Exploring fundamental SNN algorithm for spike-based SciML

- Developing generic neuromorphic resource model and mapping onto hardware (Sandia, Brown)
  - Energy scaling with spike count / neuron connectivity
  - Loihi/Lava SNN framework for benchmarking regression tasks
  - Quantized Spiking PINN implementation on Loihi

- Developing new spiking graph neural network (Sandia, UPenn, Brown)
  - Connecting spikes to graphs

- Exploring optimal SNN architectures for stable & efficient learning (Yale, Brown, PNNL)
  - Spatio-temporal Attention in Spike Transformers
  - Sparse Matrix-Sparse Matrix Hardware Accelerator for SNNs
  - Temporally Coded & Dynamic Timestep SNNs
  - Quantized SNNs

Sparser SNN and
Low Latency

- Applying various coding schemes, neuron models to SNN regression tasks (Brown, Sandia, Yale)

- Converting PINN into spiking PINN (Brown, Yale, Sandia)

SEA-CROGS

Till Year 2 Goals & Accomplishments

Exploring fundamental SNN algorithm for spike-based SciML

- Developing generic neuromorphic resource model and mapping onto hardware (Sandia, Brown)
  - Spiking PINN and Spiking DeepONet implementation on neuromorphic hardware (Loihi 2)
  - Energy and throughput profiling on Loihi 2
  - Improved hardware-aware quantization for spiking DeepONets.

→ Brad, Sandia

- Exploring optimal SNN architectures for stable & efficient learning (Yale, Brown, PNNL)
  - **Spatio-temporal attention in spiking transformers** ✓
  - **Sparse Matrix-Sparse Matrix Hardware Accelerator for SNNs** ✓

Sparser SNN and Low Latency

- Forward model training for spiking DeepONets (Brown)
- Converting PINN into spiking PINN (Brown, Yale)

→ Qian, Brown

SEA-CROGS

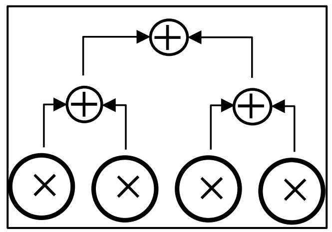
Spiking Transformer

- **Bottlenecks of Self-Attention in Standard Transformer**

ViT

High MAC operations

- QKV Generation
- Projection
- QK^T and SM·V

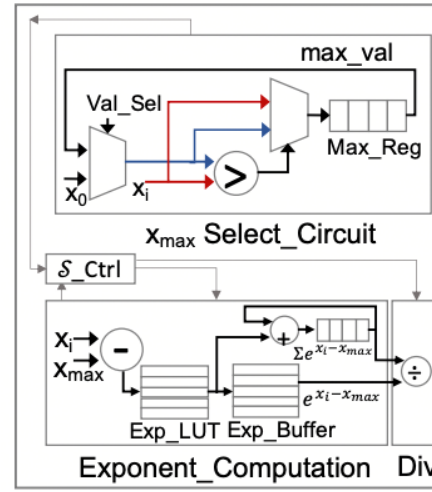


Costly Multipliers

Softmax

$$\frac{e^{x-x_{max}}}{\sum e^{x-x_{max}}}$$

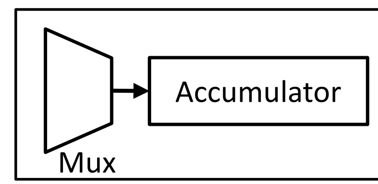
- Sequential Ops
- Complex Ops



Spiking Transformer

$$LIF\left(\overset{Query}{\begin{bmatrix}1&0&1\\0&0&0\\1&1&0\\0&0&1\end{bmatrix}} \times \overset{Key}{\begin{bmatrix}0&1&0&0\\0&0&0&0\\1&1&0&1\end{bmatrix}} \times \overset{Value}{\begin{bmatrix}0&0&1\\0&0&0\\1&1&0\\0&0&1\end{bmatrix}}\right)$$

$$Attn = LIF\{(Q \odot K^{\mathsf{T}}) \odot V\}$$



(+) MAC Operations converted to Mux & Accumulations

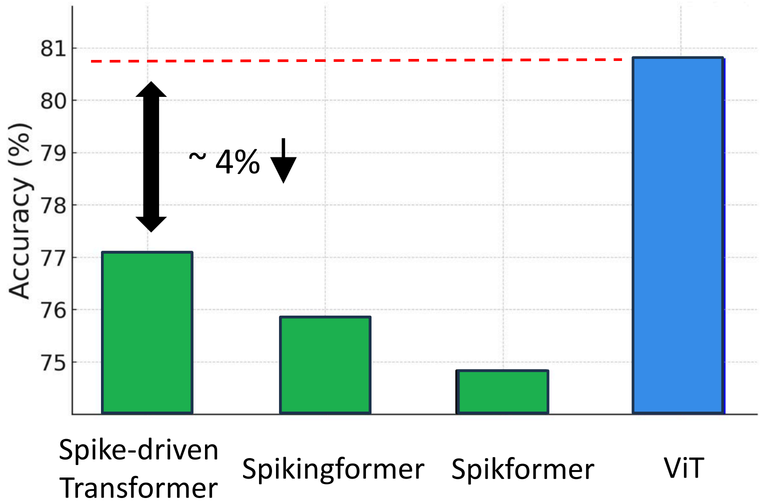
(+) No Softmax Required

SEA-CROGS

Moitra, Abhishek, et al. "TReX-Reusing Vision Transformer's Attention for Efficient Xbar-based Computing." *IEEE TETC* 2024.

Accuracy Drop with Spiking Transformer

- **Accuracy on ImageNet**



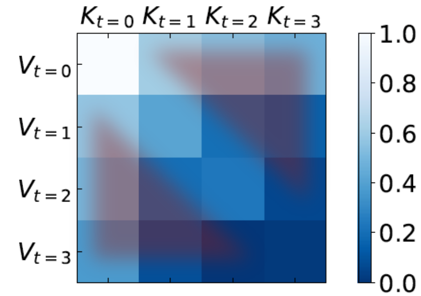
Zhou, Zhaokun, et al. "Spikformer: When spiking neural network meets transformer." *arXiv preprint arXiv:2209.15425* (2022).

Zhou, Chenlin, et al. "Spikingformer: Spike-driven residual learning for transformer-based spiking neural network." *arXiv preprint arXiv:2304.11954* (2023).

Yao, Man, et al. "Spike-driven transformer." *Advances in neural information processing systems* 36 (2024).

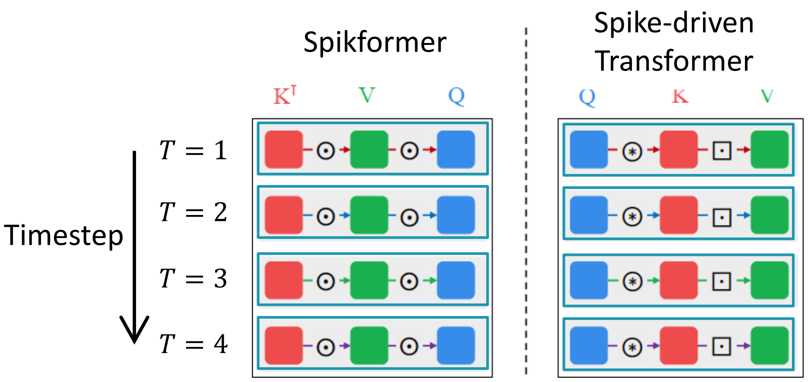
- **Spike Patterns in Self-Attention**

✓ Spike features are various across the timestep



White: high similarity
Blue: low similarity

→ **Q, K, V information are different across the timesteps**



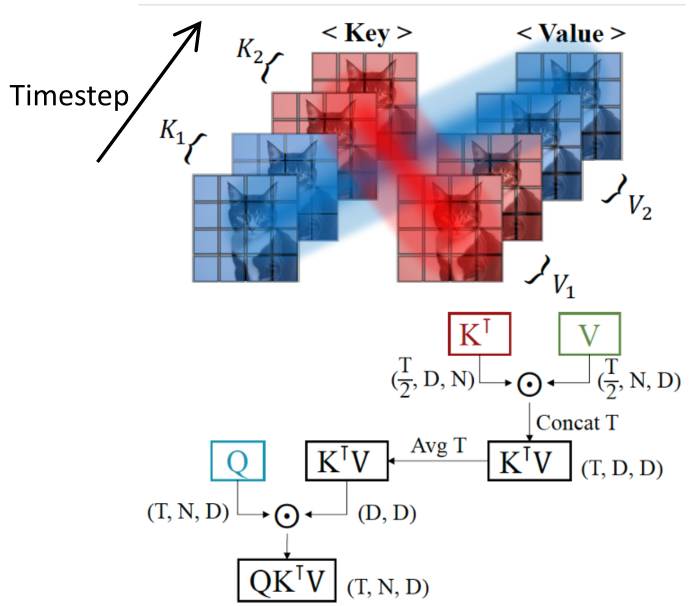
→ **Only spatial correlation**

SEA-CROGS

Spatio-Temporal Attention (STAtten)

- **Spatio-Temporal Attention (STAtten)**

1) Divide K and V into two groups (K_1, K_2, V_1, V_2)
2) Cross-attention between different timestep



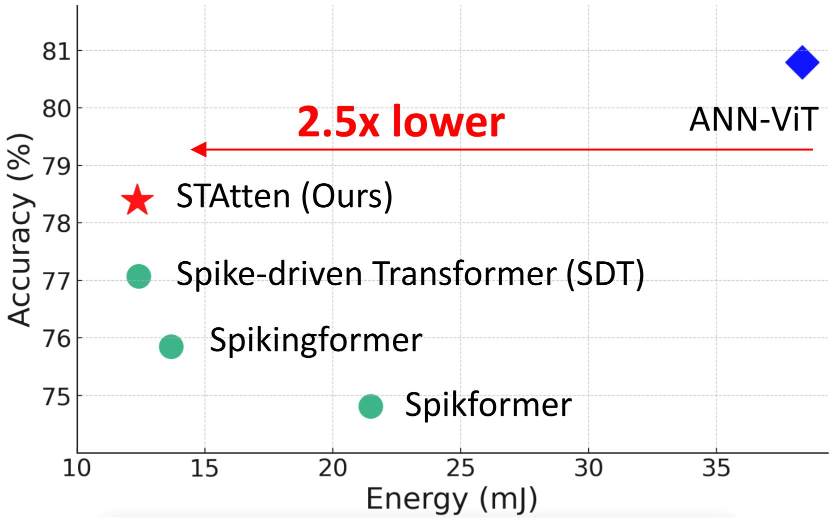
Complexity: $\mathcal{O}(TND^2)$ - Ours (Linear in N)

$\mathcal{O}(T^2N^2D)$ - Conventional (Quadratic in N)

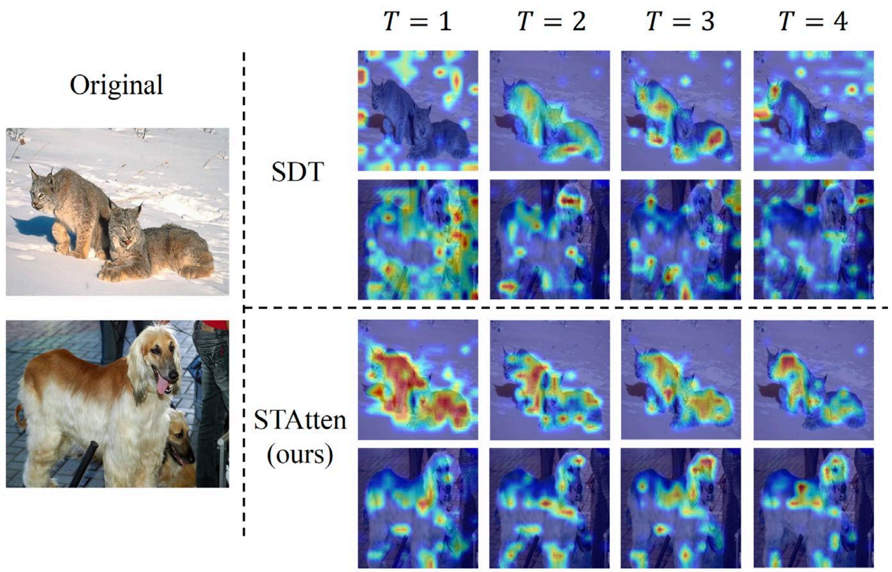
Lee, Donghyun, et al. "Spiking Transformer with Spatial-Temporal Attention." arXiv:2409.19764 (2024).

Spatio-Temporal Attention (STAtten)

- **Experiments on ImageNet**



- **Grad-Cam**



→ Consistent information capture across the timestep

SEA-CROGS

Lee, Donghyun, et al. "Spiking Transformer with Spatial-Temporal Attention." arXiv:2409.19764 (2024).

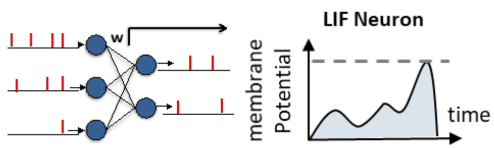
Summary

Spiking Transformer with Spatial-Temporal Attention

Spiking Activation based Efficient Transformer

① Memory Bottleneck ↓

✓ 1-bit Precision QKV by Leaky-Integrated Fire (LIF) neuron



Query

1	0	1
0	0	0
1	1	0

Key

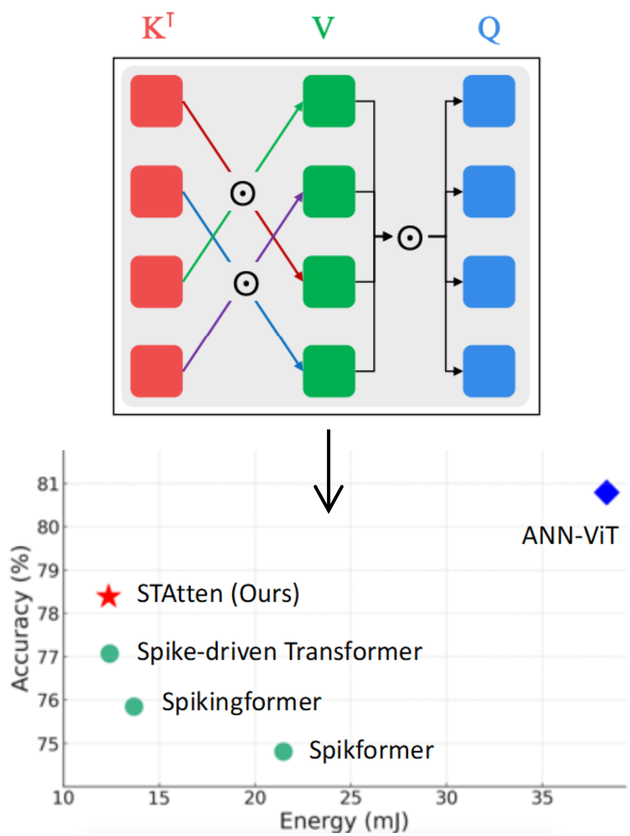
0	1	1
0	1	0
0	0	1

Value

1	0	0
0	0	1
1	1	1

② No Softmax

$$Attn = LIF\{(Q \odot K^{\top}) \odot V\}$$



Future Work

- Implementation on FPGA

https://www.xilinx.com/products/boards-and-kits.html#resources

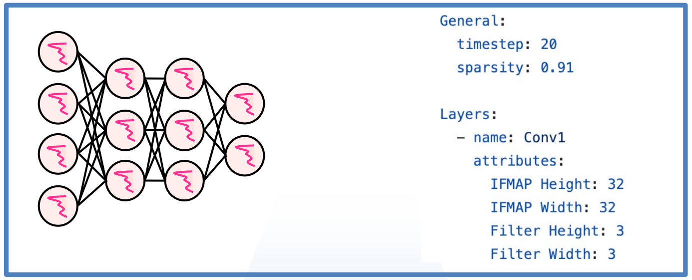
SEA-CROGS

https://www.arxiv.org/abs/2409.19764

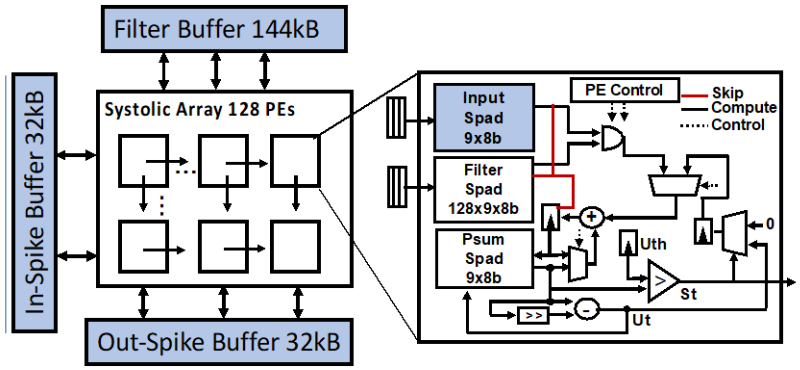
SNNs are Energy-efficient at Inference

With our in-house hardware benchmarking tools for SNNs (e.g., **SATA** [1]), we compare the energy-efficiency of various SNN workloads vs. their ANN counterparts.

On average, **2.3x ~ 6.8x** of energy efficiency improvements can be observed **[2]**.



SATA workloads

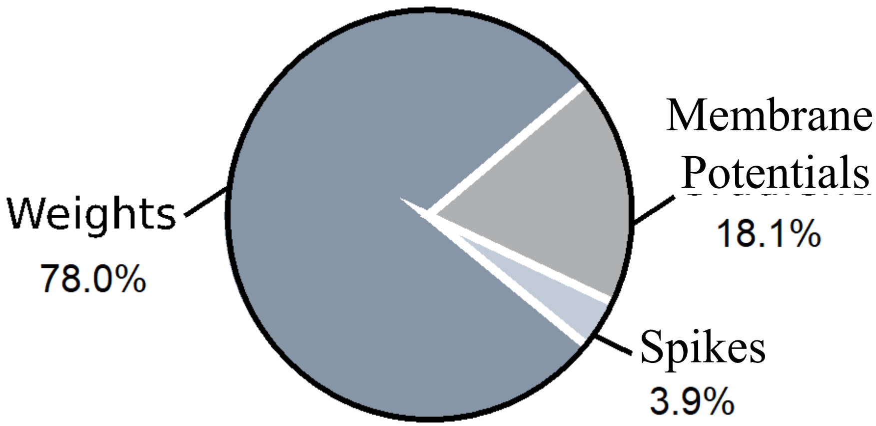


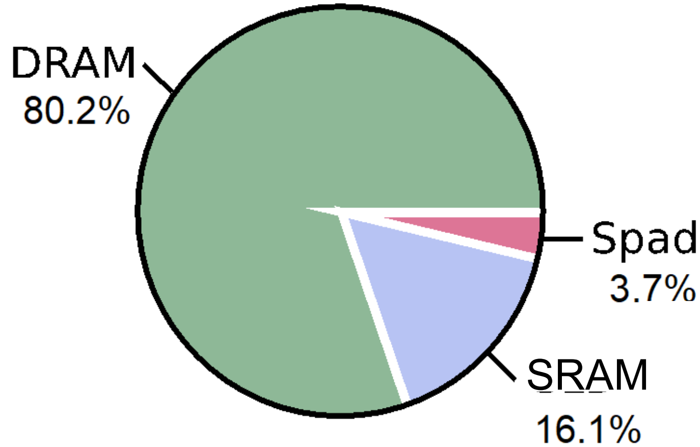
SATA hardware simulation

SEA-CROGS

[1] Yin, Ruokai et al. "SATA: Sparsity-Aware Training Accelerator for Spiking Neural Networks." *IEEE TCAD* (2022).
[2] Bhattacharjee, Abhiroop, et al. "Are SNNs Truly Energy-efficient?—A Hardware Perspective." *IEEE ICASSP* (2024).

Further Improvements?





Memory accesses to DRAM & SRAM for weights & membrane potentials are expensive!

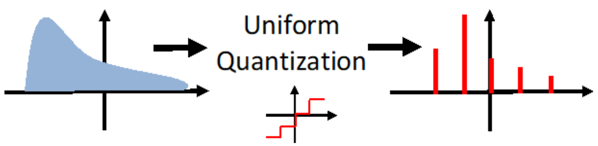
It is possible for the memory operations to dilute the computation efficiency brought by unary spikes.

SEA-CROGS

Further Improvements?

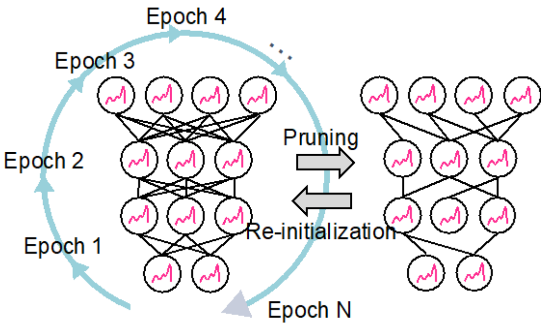
To mitigate the data movement overheads, we can compress the size of the data. Two popular algorithm solutions are there:

Quantization



Year 1:Yin et al., ASP-DAC 2024
(Best Paper Award Nominee)

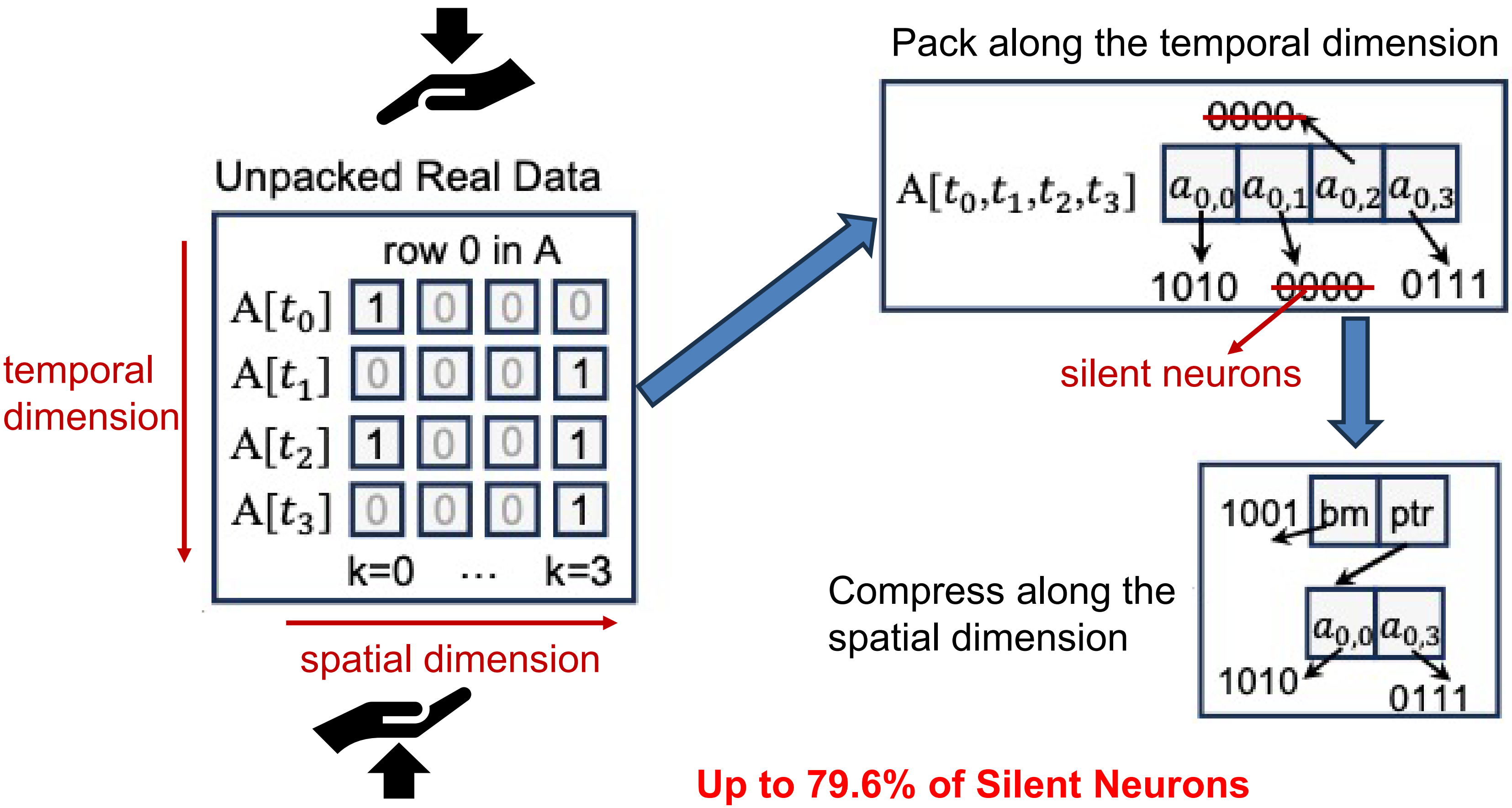
Pruning



SEA-CROGS

Ruokai, Yin, et al. "MINT: Multiplier-less INTeger Quantization for Energy Efficient Spiking Neural Networks." *ASP-DAC* (2024). **Best paper nomination.**
Code: https://github.com/Intelligent-Computing-Lab-Yale/MINT-Quantization

Packing spikes along temporal dimension: Silent Neuron Sparsity based Acceleration



Up to 79.6% of Silent Neurons

Only store the compressed non-silent neurons

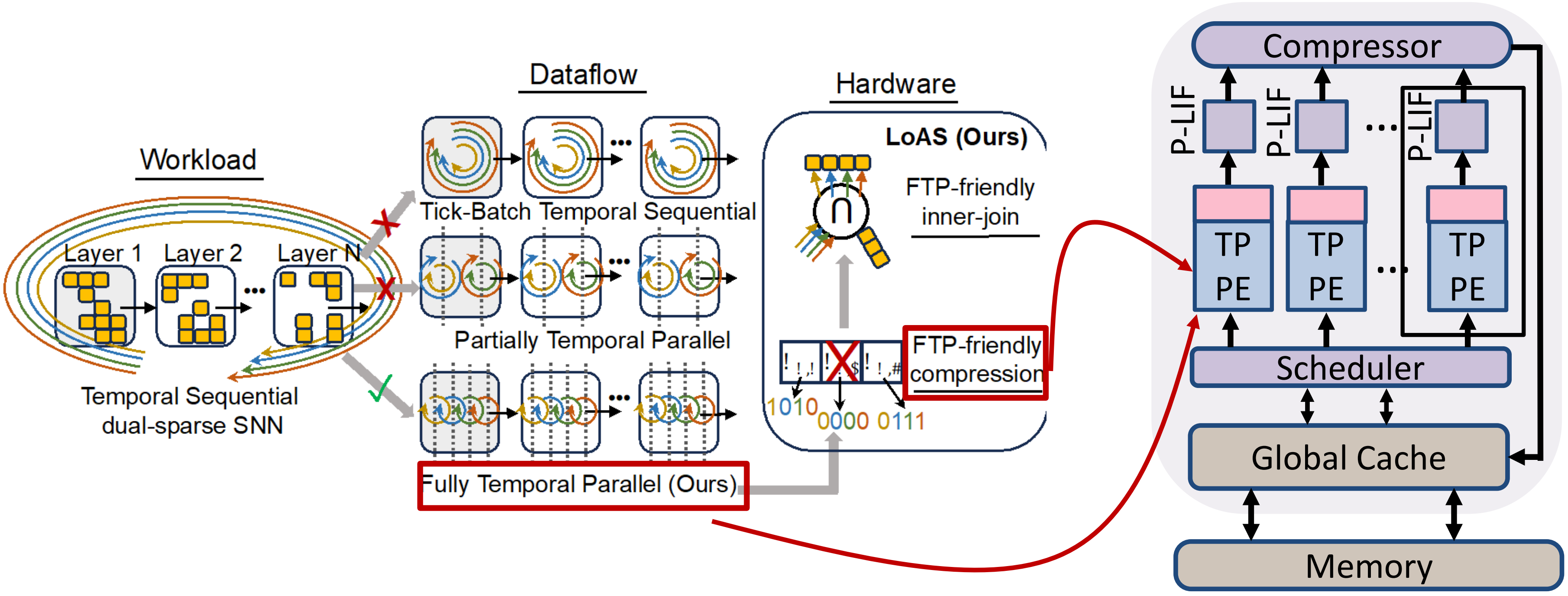
Compression efficiency is now (8/4=2)

SEA-CROGS

Ruokai, Yin, et al. " LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks." *MICRO* (2024)

LoAS: A Dual Sparse SNN Accelerator

We propose LoAS, a dual-sparse SNN accelerator that employs our compression method together with a fully temporal parallel dataflow.



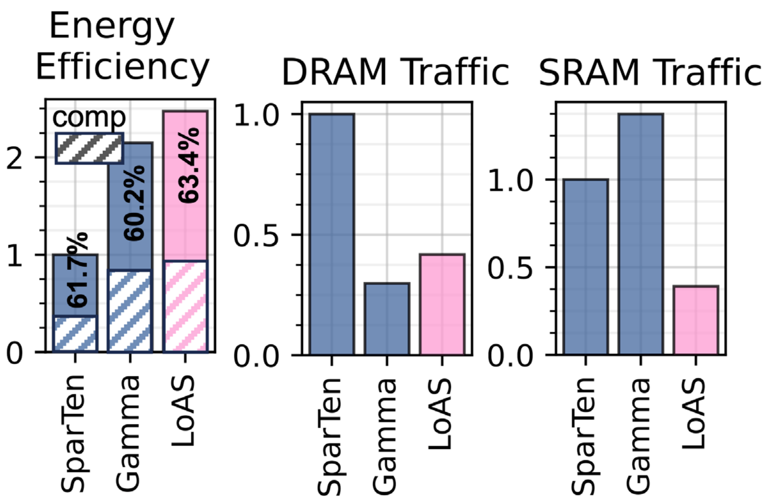
SEA-CROGS

Ruokai, Yin, et al. " LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks." *MICRO* (2024)

Experimental Results

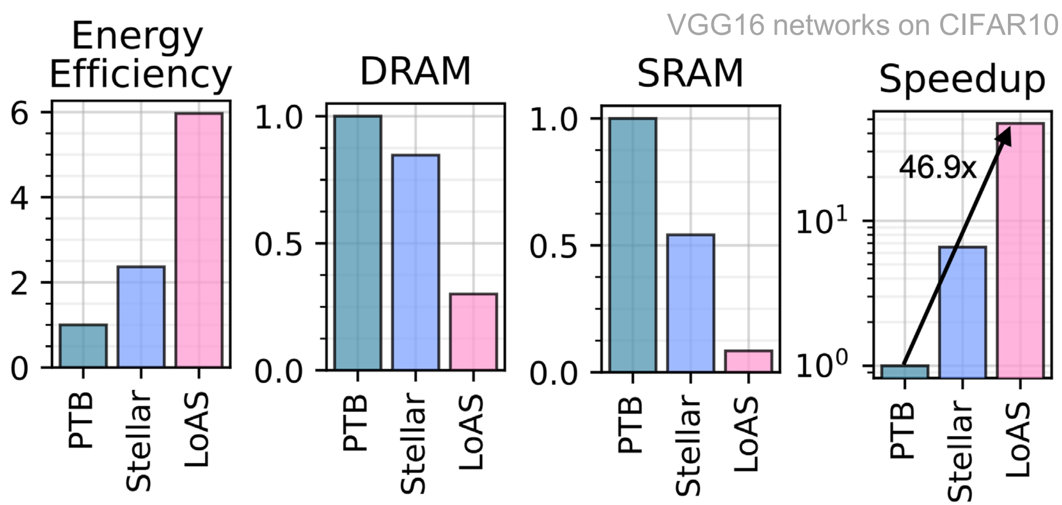
With the help of FTP dataflow and compression, LoAS is more efficient

vs. dual-sparse ANNs



On average 1.9x more energy efficient

vs. existing dense SNN accelerators

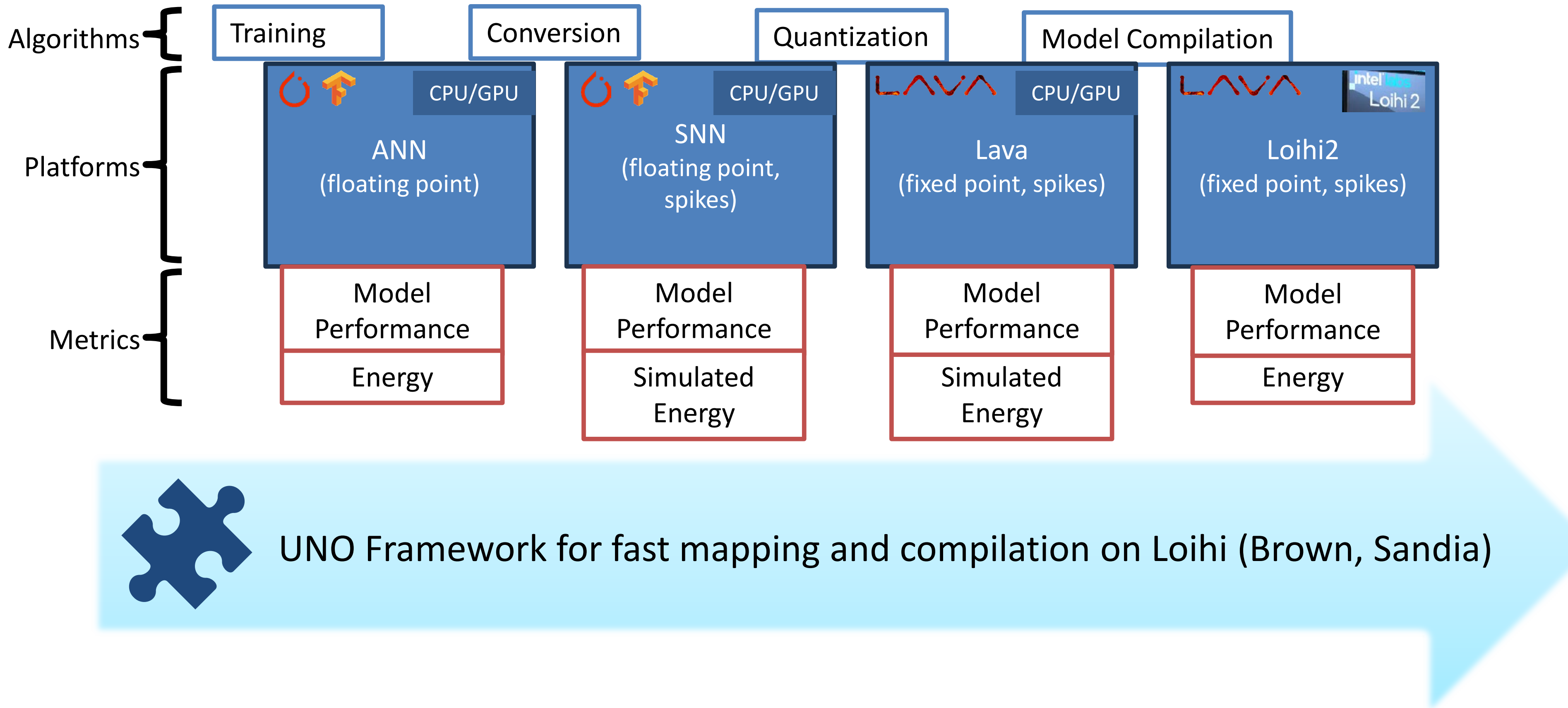


On average 2.8x more energy efficient, >45x speedup

SEA-CROGS

Ruokai, Yin, et al. " LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks." *MICRO* (2024) Code: https://github.com/Intelligent-Computing-Lab-Yale/LoAS

Spiking SciML: Developing Spiking Workflow
(Brad, Sandia)

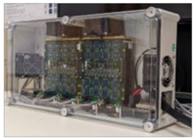


SEA-CROGS

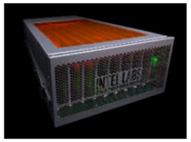
Key Takeaway from Thrust III

- We are the first to comprehensively explore SPIKES for scientific computations
- We are the first to demonstrate Spiking PINNs on Intel Loihi

Intel Loihi2



Intel Loihi1



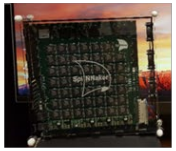
Intel Loihi
Kapoho Bay USB

Inilabs DAVIS
240C DVS

GraphCore

SpiNNaker2

SpiNNaker1



IBM TrueNorth*

Prophesee
Event-Sensor

Groq

Benefits:

- Unique Access to Neuromorphic Hardware (Intel Loihi1, Loihi2) through partnership with Sandia National Labs and Intel Labs
- Opportunities to lead and explore SciML in future platforms (SpiNNaker and others)

5 Year Trajectory & Interaction

SEA-CROGS

5 Year Trajectory
Team: Yale, Sandia, Brown, PNNL

Objective: Advance the utility of Neuromorphic architectures in SciML

1

- **Develop generic neuromorphic resource model**
- **Develop spiking graph neural network**
- **Explore optimal SNN architectures for stable learning**
- **Exploring efficient SNNs**
- **Apply various coding schemes to regression tasks**
- **Convert PINN into spiking PINN**

2

- Apply local learning ideas to spiking GNNs
- Spiking marching scheme
- **Hybrid training (conversion and fine-tuning)**

3

- Develop loss functions from formal resource model
- Apply SNN inference in a distributed model
- **Optimization beyond backprop (Hebbian Learning)**

4

- **Transformer-based embeddings for spiking representations**
- **Simulation of advanced backpropless training algorithm on hardware**
- **Direct training algorithm for spiking PINN**

5

- **Run Graph SNNs on neuromorphic architectures**
- Demonstrate ability to minimize power needs
- Simulation of distributed training and inference on neuromorphic

SEA-CROGS

Thank You!!

Questions??