**ETLTC-ICETM-ICES2025**
**Aizu-Wakamatsu, Japan**

# Neuromorphic Computing: A Paradigm for Low-Power Intelligent Systems

Khanh N. Dang

University of Aizu


January 21, 2025

# Table of Contents

# Enviromental Impact of AI computing

AI (Artificial Intelligence) applications (e.g., deep learning, data analysis, large language models) require **high-end devices, massive computational power, resulting in high energy consumption**:

- Embodied carbon emission * and operational carbon emission is challenging for the environment. AI's carbon footprint is projected to range from 2.1% to 3.9% of the total share of greenhouse gas (GHG) emissions[1].
- Llama 3.1 [2] has 405 billion parameters, and requires approximately 3.8 x $10^2$5 floating-point operations to train which translates into **energy consumption of around hundreds of MWh**.
- NVIDIA and Amazon estimate that over **80% of their energy consumption is for AI inference** [3], while Google estimates that **inference accounts for around 60%**[4].

**A solution for green and sustainable AI** is needed.

*raw material extraction, manufacturing, transportation, installation, maintenance and end-of-life

# Complexity of AI models

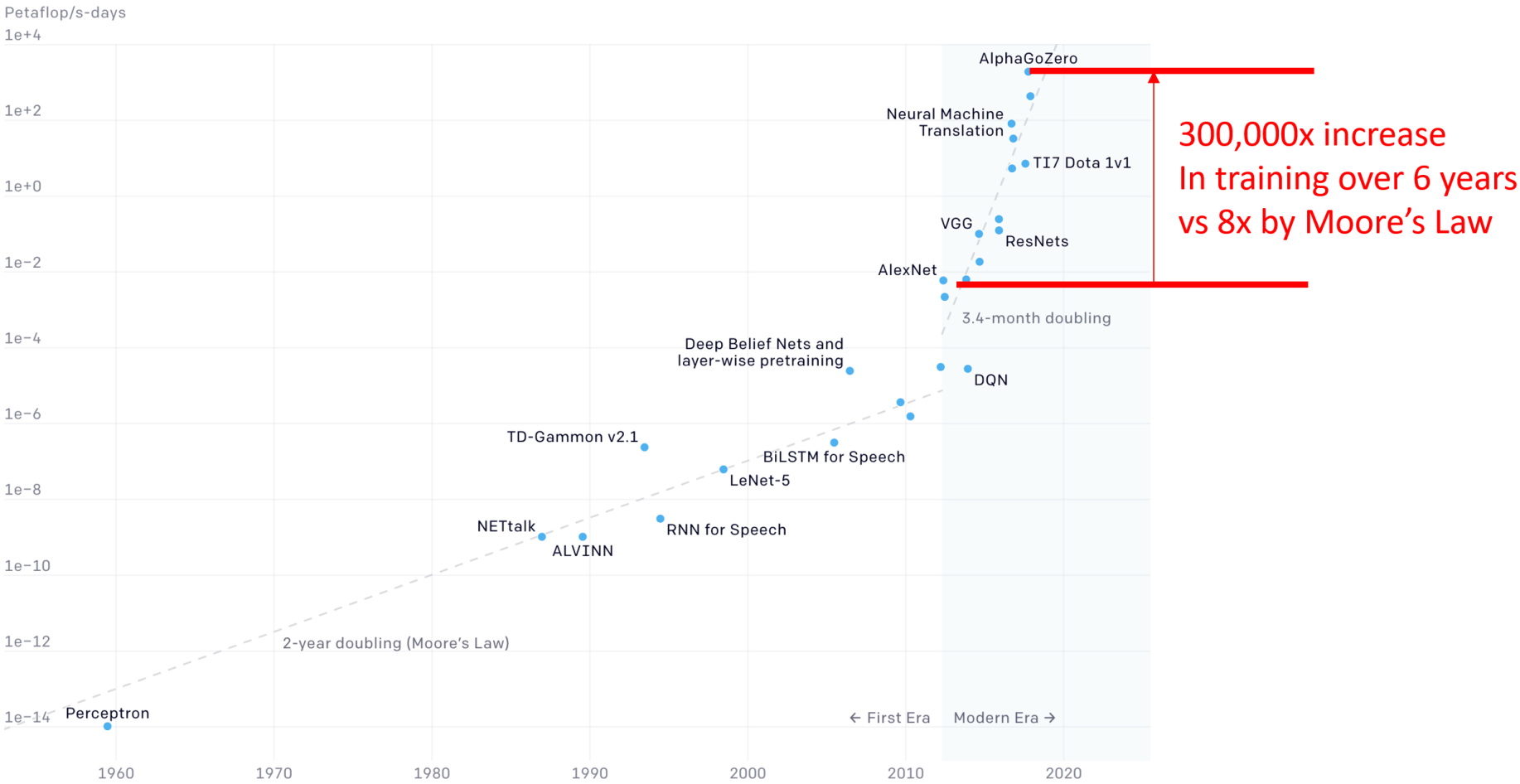



Figure 1: Rapid power hungry of training for deep learnings. Source: Intel Labs[5].

# Nature-Inspiration Design

Parrot

Brain
Power: 50mW
Weight: 2.2 grams

Can learn to speak words

Navigates and learns unknown environments at 35km/h

Can learn to manipulate cups to drink

Autonomous Drone

CPU/GPU
Power: 18,000mW
Weight: 40 grams

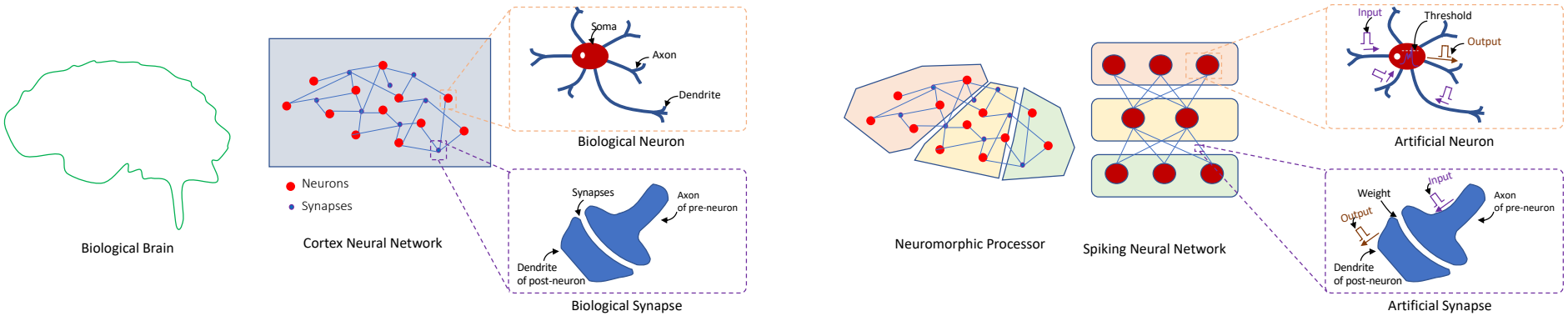
Cannot learn anything online

Pretrained to flight between known gates at walking pace

Figure 2: A brief comparison of Neuromorphic System and Conventional AI. Source: Intel Labs[5].

# Neuromorphic Computing as a Solution

- Neuromorphic computing **mimics the structure and functionality of biological neural systems**.
- Neuromorphic hardware uses **event-driven processing**, activating components only when necessary, unlike traditional systems which are always running.




(a) Biological Brain

(b) Neuromorphic System

Figure 3: Biological Brain and Neuromorphic System.

# Neuromorphic vs von Neumann

- von Neumann architectures require **back-and-forth data transferring between memory and CPU/GPU**: high latency and energy consumption
- Neuromorphic architectures **integrate memory and processing units**: minimal latency and energy.

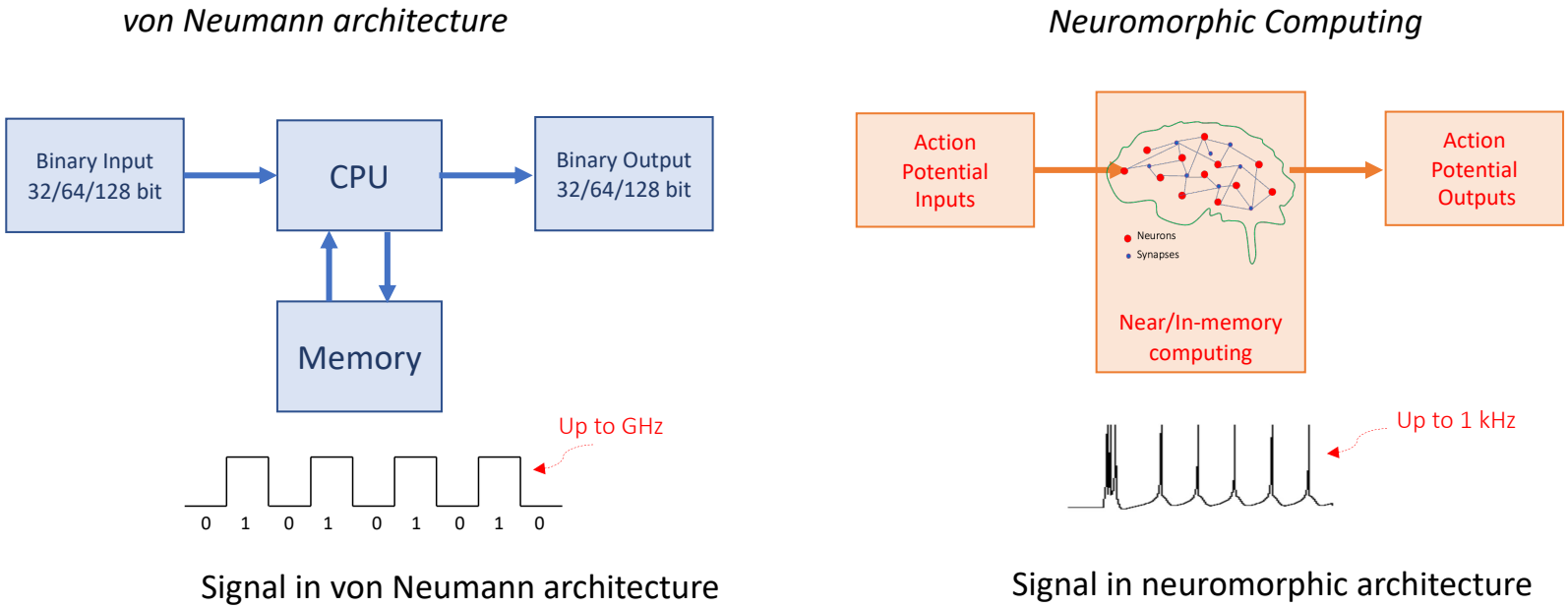



Figure 4: von Neumann vs Neuromorphic.

# Table of Contents

# Key Concepts of Neuromorphic Computing

- Input and output: events (or spikes).
- Communication: address event representation.
- Computation:
  1. incoming spikes are multiplied with synaptic weights → weighted inputs
  2. weighted inputs are accumulated to the membrane potential
  3. once the membrane potential crosses the threshold, the neuron issues outgoing spikes and reset the membrane potential.

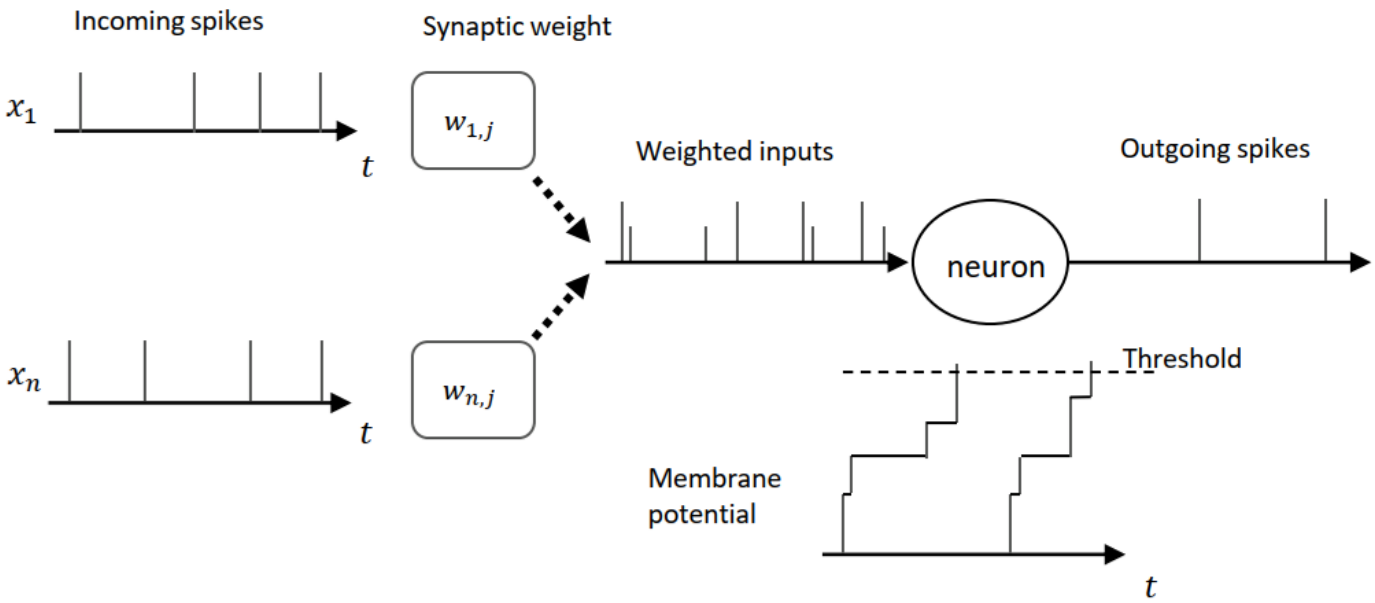



Figure 5: Spiking Neuron Model.

# Neuromorphic Computing Platforms (1/2)

- **Software Platform:** compute neuromorphic system in von Neumann machine using Machine Learning frameworks.
  - Examples: BindsNet[6], NEST[7], Brian[8].
  - Main objectives: neuroscience study and fast development of neuromorphic algorithms.
- **Hardware Platform:** Compute neuromorphic system with application-specific chips for neuromorphic.
  - Examples: IBM TrueNorth[9], Intel Loihi[10], SpiNNaker[9] (SpiNNaker use ARM processors), BrainScaleS[11], FPGA-based[12].
  - Main objectives: low-power & low-latency inferences.

# Neuromorphic Computing Platforms (2/2)

| | **Human Brain** | **SpiNNaker**[9] | **HiCANN/BrainScaleS**[11] | **TrueNorth**[9] | **Loihi**[10] |
|---|---|---|---|---|---|
| Neurons | 100 billion | 1 billion | $\approx$ 4 million | 1 million | 131,072 |
| - using | 1.4 kg | 10 racks | 20 wafers | 1 chip | 1 chip |
| Mean synapses per neuron | 7,000 | Prog. | 224 | 256 | 1,000 |
| Max synapses per neuron | $\approx$ 15,000 | Prog. | 14,336 | 256 | 1,000 |
| Energy per spike | 8 fJ | 4 nJ | 0.1–10 nJ | 26 pJ | $>$23.6 pJ |
| - compared to brain | 1 | 500,000 | 12,500–1,250,000 | 3,250 | $>$2,950 |
| Speed up | 1 | 1 | $10^3 - 10^5$ | 1 | 1 |
| Run time plasticity | Yes | Prog. | STDP | No | STDP |
| Neuron model | Diverse | Prog. | Adaptive exponential | LIF | LIF |

Table 1: Comparison of Neuromorphic Architectures

- The Human Brain has a massive scale in terms of neurons and synapses.
- The Human Brain has exceptional energy consumption.

# Table of Contents

# Challenges and Research Directions

Although neuromorphic computing can offer low-power solutions, there are some existing challenges to be addressed

- **Training for Neuromorphic Systems**: Designing effective training algorithms for neuromorphic systems is challenging due to their non-differentiable nature and reliance on event-driven spiking activity.
- **Reaching Carbon-Neutrality or Net-Zero Computing**: Achieving carbon-neutrality in computing requires minimizing energy consumption and adopting sustainable practices throughout the hardware lifecycle, from design to operation and recycling.

# Toward Net-Zero Neuromorphic Design (1/2)

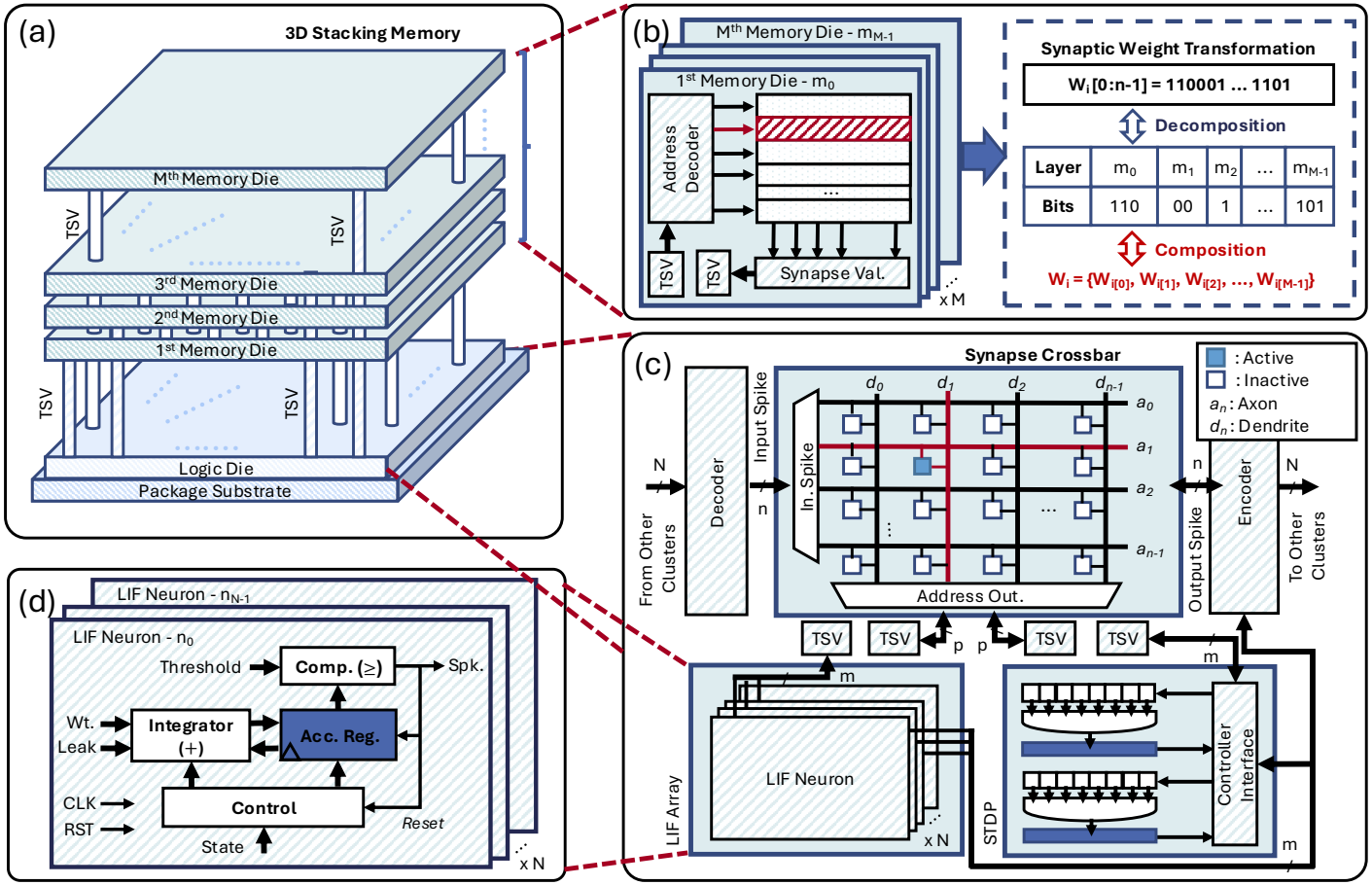



Figure 6: Overview of our system: (a) 3D stacking memory with $M$ layers; (b) Approximate Stack Memory; (c) Computing Core; (d) LIF neuron.

# Toward Net-Zero Neuromorphic Design (2/2)

THE UNIVERSITY OF AIZU

Our Neuromorphic Design[†]

- **3D-Integrated Circuit-based Stacking Memory**: support weight decomposition for approximation.
- **Approximation Circuits for Neuron**: support inaccurate adders with low-power consumption.
- **On-Chip Learning**: Spike-timing-dependent plasticity.
- **Neural Searching Platform:** Evolutionary Algorithm to search for approximate adders and approximation level in memory.

[†] "Energy-Efficient Spiking Neural Networks Using Approximate Neuron Circuits and 3D Stacking Memory", https://ieeexplore.ieee.org/document/10819545
"Power-aware Neuromorphic Architecture with Partial Voltage Scaling 3D Stacking Synaptic Memory" https://ieeexplore.ieee.org/document/10269541

# Evaluation Results

| Model | MNIST Acc.(%) | Arch. | Tech. | Energy per SOP (pJ) | Energy per SOP (pJ) (in 14nm) |
|---|---|---|---|---|---|
| TrueNorth [13] | 91.94 | 2D | 28nm | 26 (0.775V) | 4.902 |
| Loihi [10] | 96 | 2D | 14nm FinFET | 23.6 (0.75V) | 23.6 |
| ODIN [14] | 84.5 | 2D | 28nm FD-SOI | 8.4 | 1.078 |
| NASH [15] | 79.4 | 3D | 45nm | 11.3 (1.1V) | 0.648 |
| [16] | 95.35<br>94.84<br>88.77 | 3D | 45nm | 244.28<br>191.46<br>81.16 | 14.02<br>10.98<br>4.65 |
| [17] | 94.8<br>93.9<br>77.6 | 3D | 45nm | 20.33<br>13.28<br>8.374 | 1.167<br>0.762<br>0.48 |
| Ours | $97.74^{1}$<br>$97.11^{2}$<br>$94.57^{3}$<br>$90.30^{4}$<br>$86.38^{5}$ | 3D | 45nm | $8.797^{1}$<br>$5.163^{2}$<br>$3.057^{3}$<br>$5.900^{4}$<br>$3.898^{5}$ | $0.504^{1}$<br>$0.296^{2}$<br>$0.175^{3}$<br>$0.338^{4}$<br>$0.223^{5}$ |

[1] Case 1: Accurate implementation (four-layer model).
[2] Case 2: $V_{DD} = 0.8V$ in UV3 mode using the configuration $X_1$.
[3] Case 3: $V_{DD} = 0.8V$ in UV-PG3 mode using the configuration $X_1$.
[4] Case 4: $V_{DD} = 0.7V$ in UV-PG1 mode using the configuration $Y_3$.
[5] Case 5: $V_{DD} = 0.8V$ in UV-PG3 mode using the configuration $Y_3$.

# Evaluation Results

| Model | MNIST Acc.(%) | Arch. | Tech. | Energy per SOP (pJ) | Energy per SOP (pJ) (in 14nm) |
|---|---|---|---|---|---|
| TrueNorth [13] | 91.94 | 2D | 28nm | 26 (0.775V) | 4.902 |
| Loihi [10] | 96 | 2D | 14nm FinFET | 23.6 (0.75V) | 23.6 |
| ODIN [14] | 84.5 | 2D | 28nm FD-SOI | 8.4 | 1.078 |
| NASH [15] | 79.4 | 3D | 45nm | 11.3 (1.1V) | 0.648 |
| [16] | 94.84 | 3D | 45nm | 191.46 | 10.98 |
| [17] | 94.8<br>93.9<br>77.6 | 3D | 45nm | 20.33<br>13.28<br>8.374 | 1.167<br>0.762<br>0.48 |
| Ours | 97.74[1]<br>97.11[2]<br>94.57[3]<br>90.30[4]<br>86.38[5] | 3D | 45nm | 8.797[1]<br>5.163[2]<br>3.057[3]<br>5.900[4]<br>3.898[5] | 0.504[1]<br>0.296[2]<br>0.175[3]<br>0.338[4]<br>0.223[5] |

**28.06% energy saving with similar accuracy**
**65.28% energy saving with 3.18% accuracy loss**

[1] Case 1: Accurate implementation (four-layer model).
[2] Case 2: $V_{DD} = 0.8V$ in UV3 mode using the configuration $X_1$.
[3] Case 3: $V_{DD} = 0.8V$ in UV-PG3 mode using the configuration $X_1$.
[4] Case 4: $V_{DD} = 0.7V$ in UV-PG1 mode using the configuration $Y_3$.
[5] Case 5: $V_{DD} = 0.8V$ in UV-PG3 mode using the configuration $Y_3$.

# Distributed On-Chip Learning

- Train on-chip learning in remote devices.
- Upload and synthesize the sub-models into a single models.

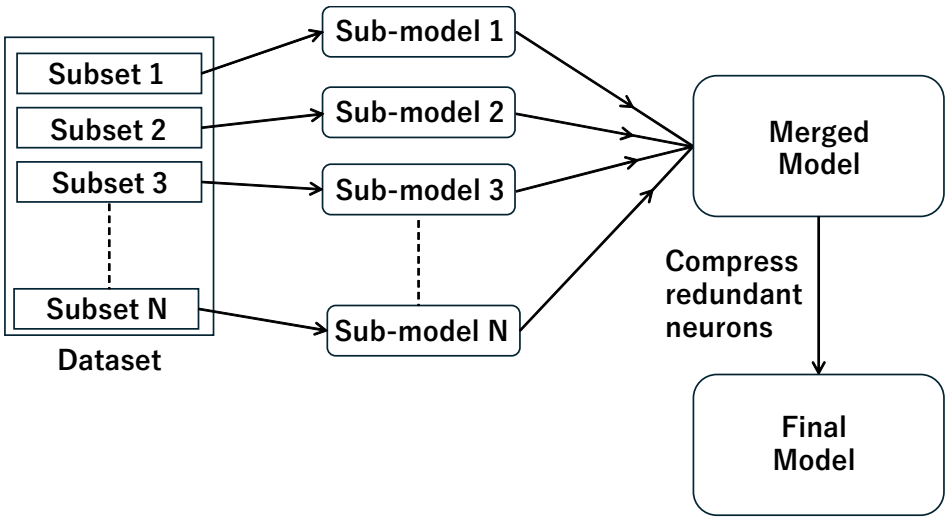



Figure 7: Ensemble STDP learning.[c]

Table 2: 300 neurons model and merging 5×100 neurons sub-models.

| Model | [18] | Ours |
|---|---|---|
| #neurons | 300 | 300 (5×100-200) |
| Training Time (minutes) | 53.13 | 10.58 |
| Classification Accuracy | 88.87% | 85.42% |

Table 3: 300 neurons model and merging 2×250 neurons sub-models.

| Model | [18] | Ours |
|---|---|---|
| #neurons | 300 | 300 (2×250-200) |
| Training Time (minutes) | 53.13 | 26.8 |
| Classification Accuracy | 88.87% | 89.28% |

[c] "EnsembleSTDP: Distributed in-situ Spike Timing Dependent Plasticity Learning in Spiking Neural Networks", https://ieeexplore.ieee.org/document/10819516

# Energy Consumption

- Train with 60K images, 1000 neurons.
- Local Node: Low-power Intel Chip. 10 Nodes.
- Server Node: GPU 4070 GPU + Ryzen 7.

Table 4: Energy Consumption for Distributed STDP learning.

| **Model** | [18] on Server | **Ours: Local Node + Server** |
|---|---|---|
| **#neurons** | 1000 | 1000 (2×100) |
| **Training Energy (Jules)** | 111,320 | 85,410 |
| **Data Transfer Energy (Jules)** | 17.82 | 2.4552 |
| **Merging Energy (Jules)** | 0 | 52.52 |
| **Total Energy (Jules)** | 111337.82 | 85464.97 (-23.24%) |

- Local Node with an average power consumption of 2.19 Watt
- Can be offset by power harvesting (solar power). Estimate with Quartz Solar Forecast → Carbon-neutrality in computing (training and inferencing).

# Sustainable AI computing (1/2)

Besides the energy challenges, one of the critical issues is the hardware lifecycle.

- Defective devices after manufacturing lead to wasted energy and carbon emissions. According to Apple[19], 75% of carbon emissions belong to manufacturing while 19% belong to operation.
- Aging and wear-out are also major concerns on environmental impact. **e-waste is a huge problem for the environement.**

Our solutions:

- Reducing the embodied carbon footprint in manufacturing with **yield improvement**.
- Extending lifetime expectancy with **reliability improvement approaches**.

# Sustainable AI computing (2/2)

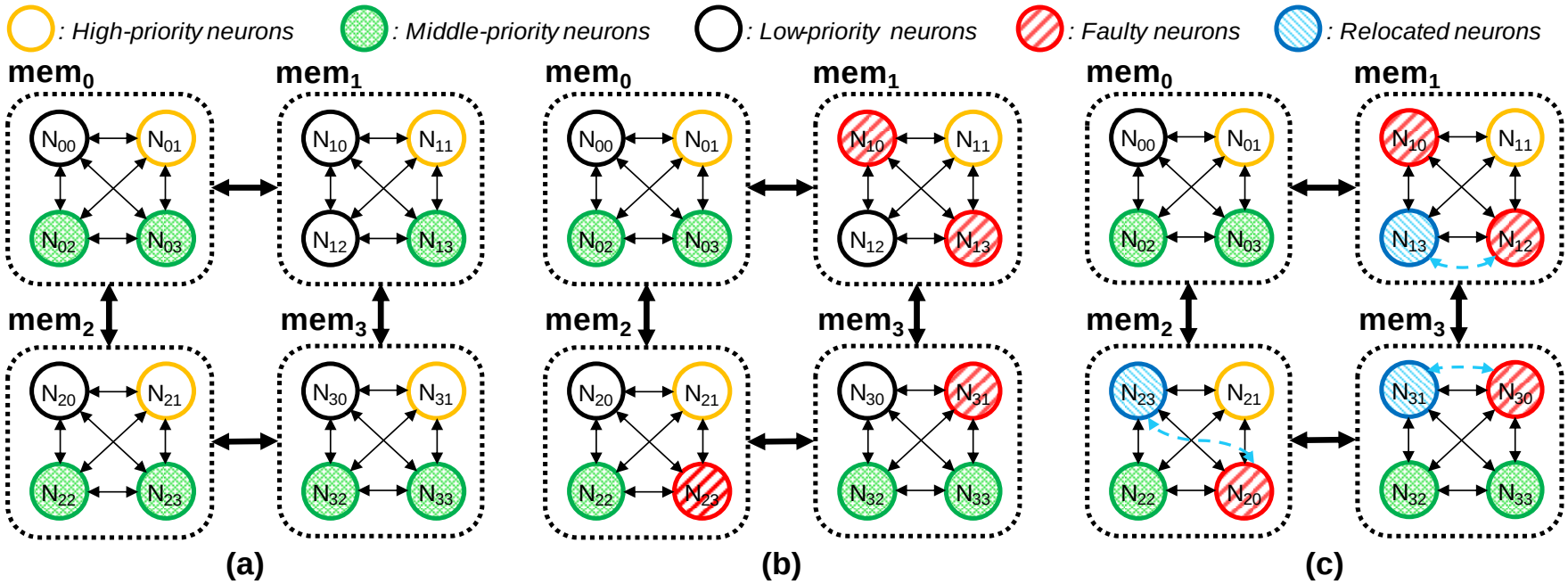



Figure 8: (a) Original neuron allocation. (b) Faults occurrence. (c) Recovery solution.

Junkyard Computing (compute with inferior hardware)[§]:

- Study AI models to find **critical and non-critical neurons and memory blocks**.
- **Reallocate** faulty neurons/memory blocks for non-critical ones.

[§] "NOMA: A Novel Reliability Improvement Methodology for 3-D IC-based Neuromorphic Systems", IEEE Transactions on Components, Packaging and Manufacturing Technology, 2024. https://ieeexplore.ieee.org/document/10738829/

# Results

Table 5: Comparison Results to Existing Works.

| | **Our work** | ReSpawn [20] | SoftSNN [21] |
|---|---|---|---|
| Network Size | 784:256:256:10 | 784:400 | 784:400 |
| Hardware Architecture | 3-D SNN | 2-D SNN | 2-D SNN |
| Benchmark | MNIST | MNIST | MNIST |
| Tolerance Technique | Swapping Weights | Fault-Aware Mapping | Bound-and-Protect |
| Bit Error Rate | 0.10 | 0.10 | 0.10 |
| Baseline Accuracy | 97.78% | $\sim$ 86%[1] | $\sim$ 86%[1] |
| Accuracy Loss | 0.01-0.24% | $\sim$ 10%[1] | $\sim$ 12%[1] |

[1] We calculated the accuracy loss based on the provided images.

# Results

Table 5: Comparison Results to Existing Works.

| | **Our work** | ReSpawn [20] | SoftSNN [21] |
|---|---|---|---|
| Network Size | 784:256:256:10 | 784:400 | 784:400 |
| Hardware Architecture | 3-D SNN | 2-D SNN | 2-D SNN |
| Benchmark | MNIST | MNIST | MNIST |
| Tolerance Technique | Swapping Weights | Fault-Aware Mapping | Bound-and-Protect |
| Bit Error Rate | 0.10 | 0.10 | 0.10 |
| Baseline Accuracy | 97.78% | $\sim$ 86%[1] | $\sim$ 86%[1] |
| Accuracy Loss | 0.01-0.24% | $\sim$ 10%[1] | $\sim$ 12%[1] |

**Maintain at most 0.24% accuracy loss at 10% error rate.**

[1] We calculated the accuracy loss based on the provided images.

# Table of Contents

# Conclusion and Future Outlook

**Conclusion:**

- Neuromorphic computing provides a **low-power, efficient solution** to the increasing energy demands of AI.
- Innovative designs, such as event-driven spiking neural networks and approximate 3D-stacked memory, promise **significant energy savings**.
- Towarding **sustainable AI computing** not just including energy consumption but also managing device lifecycle.

**Future Directions:**

- Integrate sustainability goals into hardware lifecycle management.
- Tailor the approach on scheduling to efficiently off-set the carbon emission.

# References (1/2)


[1] K. Kirkpatrick, "The carbon footprint of artificial intelligence," *Commun. ACM*, vol. 66, p. 17–19, July 2023.

[2] A. . M. Llama Team, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[3] J. McDonald *et al.*, "Great power, great responsibility: Recommendations for reducing energy for training language models," *arXiv preprint arXiv:2205.09646*, 2022.

[4] D. Patterson, "Good news about the carbon footprint of machine learning training." `https://blog.research.google/2022/02/goodnews-about-carbon-footprint-of.html`, 2022. Accessed: 2025-01-14.

[5] Intel Corporation, "Neuromorphic computing." `https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html`, 2025. Accessed: 2025-01-14.

[6] H. Hazan, D. J. Saunders, H. Khan, D. Patel, D. T. Sanghavi, H. T. Siegelmann, and R. Kozma, "Bindsnet: A machine learning-oriented spiking neural networks library in python," *Frontiers in neuroinformatics*, vol. 12, p. 89, 2018.

[7] J. M. Eppler, M. Helias, E. Muller, M. Diesmann, and M.-O. Gewaltig, "Pynest: a convenient interface to the nest simulator," *Frontiers in neuroinformatics*, vol. 2, p. 363, 2009.

[8] D. F. Goodman and R. Brette, "The brian simulator," *Frontiers in neuroscience*, vol. 3, p. 643, 2009.

[9] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber, "Spinnaker: A 1-w 18-core system-on-chip for massively-parallel neural network simulation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, 2013.

[10] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[11] C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, and J. Schemmel, "The brainscales-2 accelerated neuromorphic system with hybrid plasticity," *Frontiers in Neuroscience*, vol. 16, p. 795876, 2022.

# References (2/2)


[12] Y. Liu, Y. Chen, W. Ye, and Y. Gui, "Fpga-nhap: A general fpga-based neuromorphic hardware acceleration platform with high speed and low power," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2553–2566, 2022.

[13] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 34, no. 10, pp. 1537–1557, 2015.

[14] C. Frenkel *et al.*, "A 0.086-mm$^2$ 12.7-pJ/SOP 64k-Synapse 256-Neuron Online-Learning Digital Spiking Neuromorphic Processor in 28-nm CMOS," vol. 13, no. 1, pp. 145–158, 2018.

[15] O. M. Ikechukwu *et al.*, "On the Design of a Fault-Tolerant Scalable Three Dimensional NoC-Based Digital Neuromorphic System With On-Chip Learning," *IEEE Access*, vol. 9, pp. 64331–64345, 2021.

[16] N.-D. Nguyen *et al.*, "Power-Aware Neuromorphic Architecture With Partial Voltage Scaling 3-D Stacking Synaptic Memory," vol. 31, no. 12, pp. 2016–2029, 2023.

[17] R. Kobayashi *et al.*, "Energy-Efficient Spiking Neural Networks Using Approximate Neuron Circuits and 3D Stacking Memory," in *17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip*, IEEE, 2024.

[18] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in computational neuroscience*, vol. 9, p. 99, 2015.

[19] A. Inc., "Environment - apple," 2024.
Accessed: 2024-01-15.

[20] R. V. W. Putra *et al.*, "ReSpawn: Energy-Efficient Fault-Tolerance for Spiking Neural Networks considering Unreliable Memories," in *2021 IEEE/ACM ICCAD*, pp. 1–9, 2021.

[21] R. V. W. Putra *et al.*, "SoftSNN: low-cost fault tolerance for spiking neural network accelerators under soft errors," in *Proceedings of the 59th ACM/IEEE DAC*, (New York, USA), p. 151–156, 2022.

Thank you for your attention!