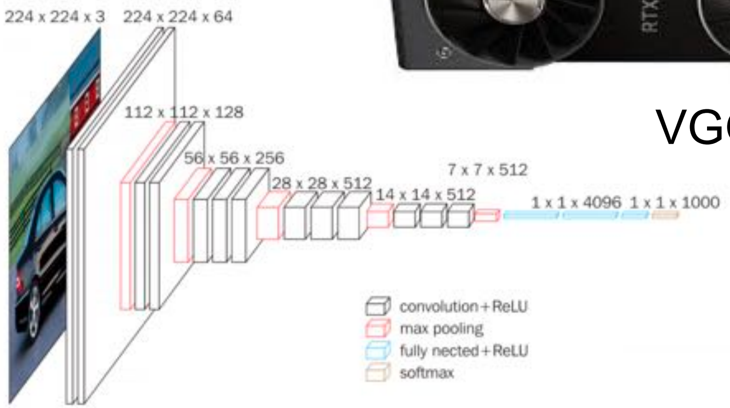
Memory Organization for Energy-Efficient Learning and Inference in Digital Neuromorphic Accelerators

Clemens JS Schaefer[1], Patrick Faley[1], Emre O Neftci[2] and Siddharth Joshi[1]

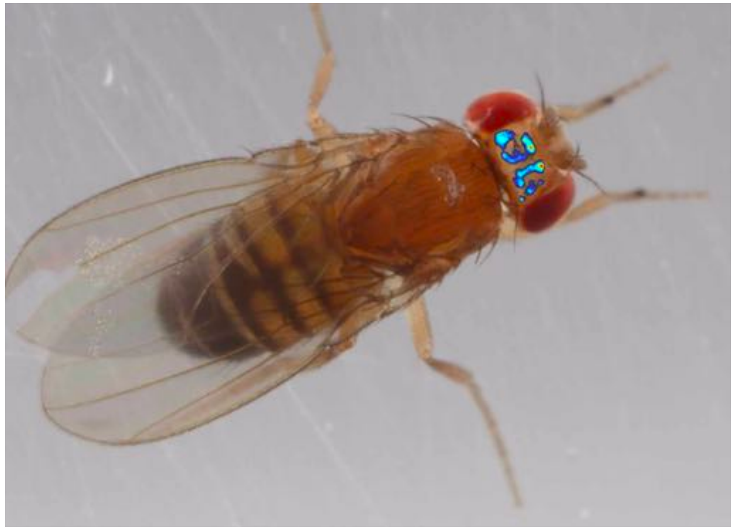
[1] University of Notre Dame du Lac - Intelligent Microsystems Lab
[2]UC Irvine - Neuromorphic Machine Intelligence Lab

Potential of Biological Neural Networks

- Biological systems: continuously learning, unreliable stimuli, noisy environment <u>and still</u> energy efficient



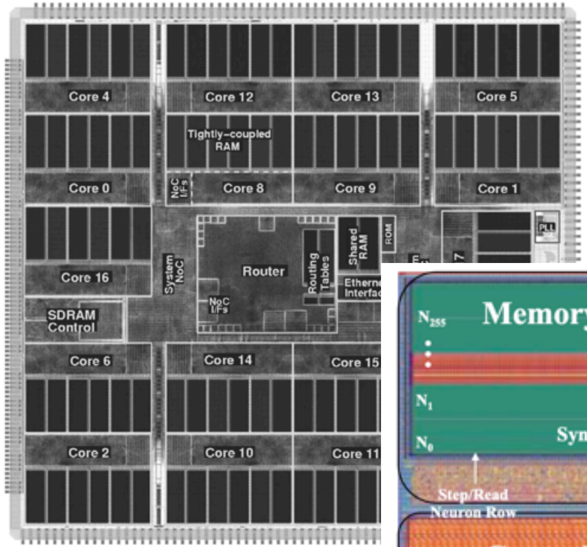
VGG16



Power	180-360W	~0.0000036344W
Volume/Space	267 mm x 116 mm x 35 mm	> 1mm x 1mm x 1mm
Computational Resources (number of neurons)	14,719,656 neurons	~135,000 neurons

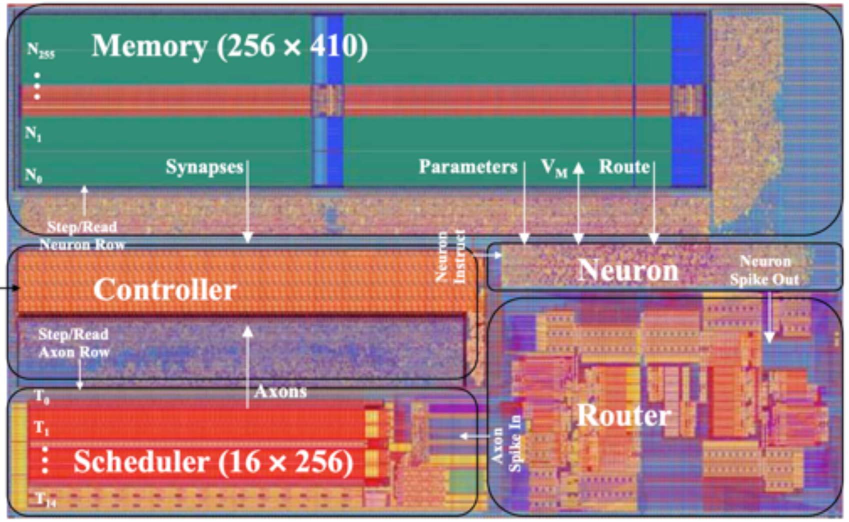
Neuromorphic Systems

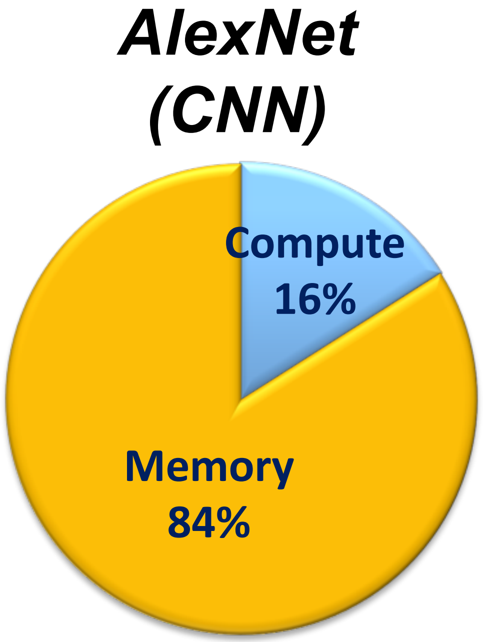
- Co-development of hardware and software to mimic biological systems
- Neurosynaptic core: neuron and synapse subsystem
- Memory access major energy driver

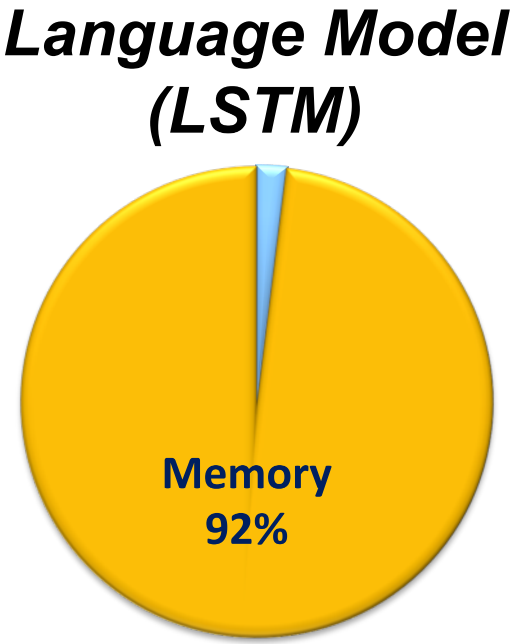


SpiNNaker

TrueNorth (IBM)





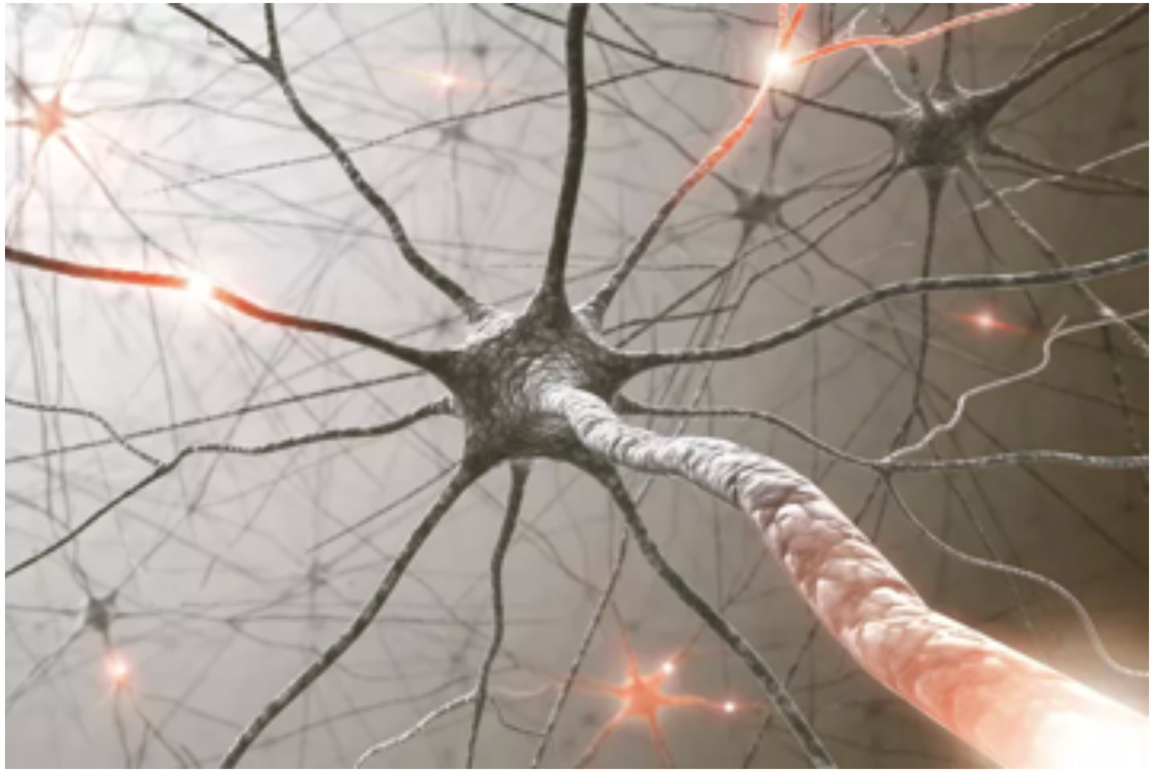


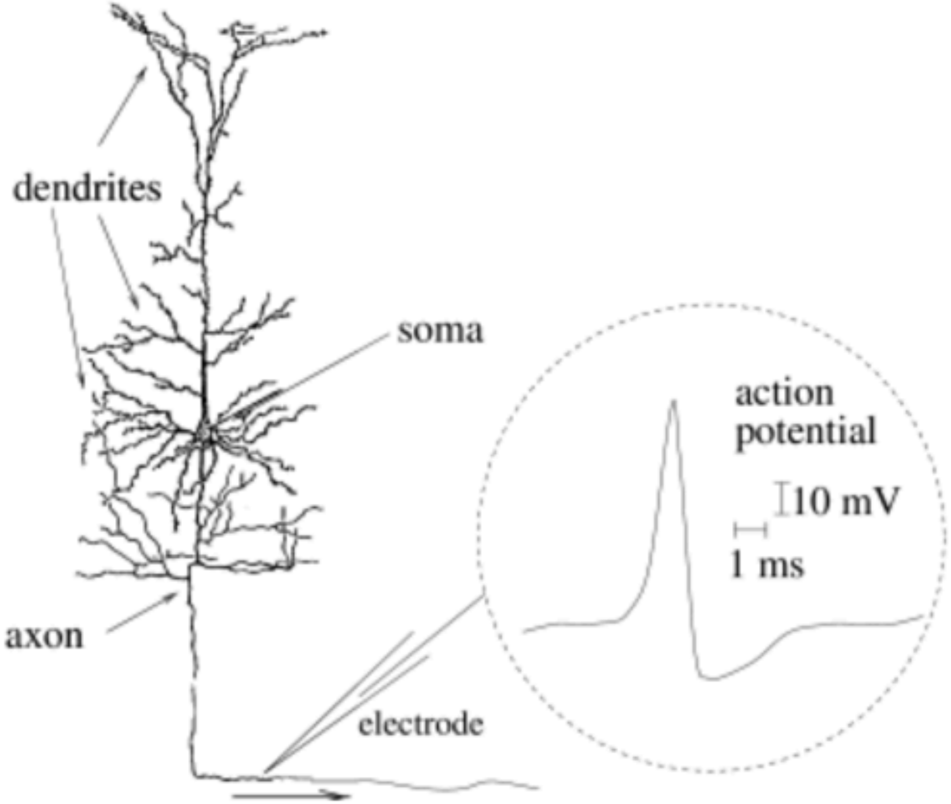
W. Wan et al., "33.1 a 74 tmacs/w cmos-rram neurosynaptic core with dynamically reconfigurable dataflow and in-situ transposable weights for probabilistic graphical models," in 2020 IEEE International Solid- State Circuits Conference - (ISSCC), 2020, pp. 498–500.

Painkras, Eustace, et al. "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation." IEEE Journal of Solid-State Circuits 48.8 (2013): 1943-1953.

Spiking Neural Networks

- Hodgkin and Huxley [7] analyzing biological nervous systems
- Communication and compute through action potential (spikes)
- Various formulations, e.g. Hodgkin-Huxley, Izhikevich, Integrate-and-Fire



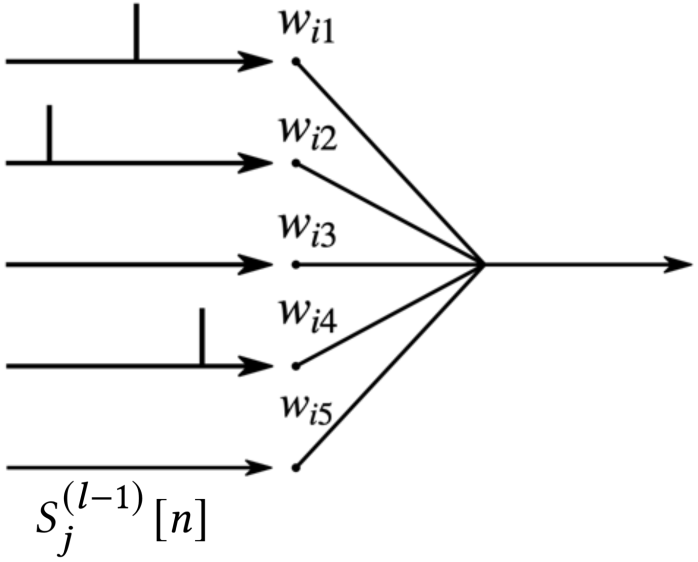


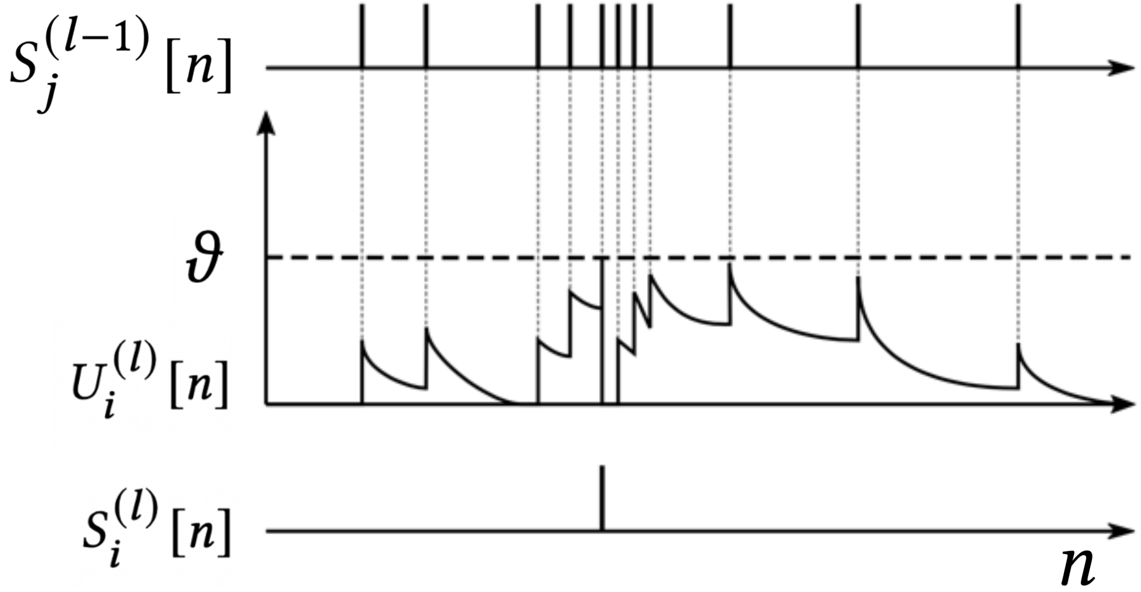
Spiking Neural Networks

- Neuron state is governed by incoming spikes, corresponding weights and reset dynamics [8]
- Spikes are generated by step-function

$$U_i^{(l)}[n] = \sum_j W_{ij}^{(l)} P_j[n] - \delta R_i[n],$$

$$S_i^{(l)}[n] = \Theta(U_i^{(l)}[n] - \vartheta),$$

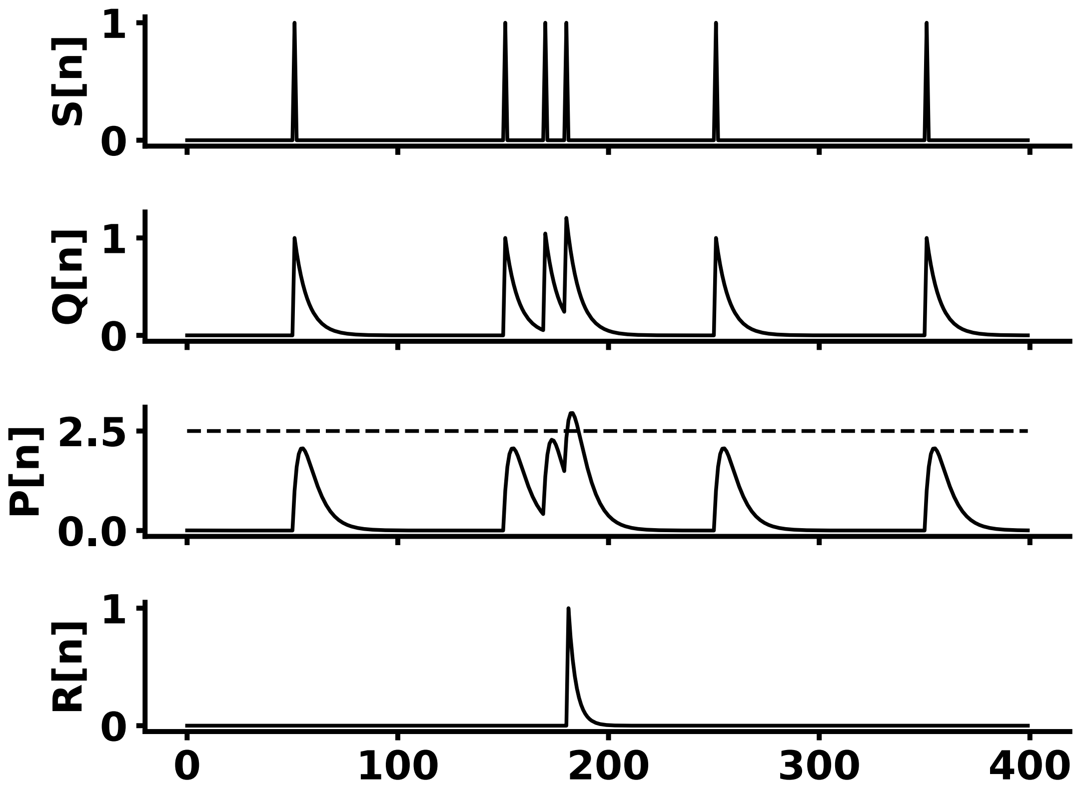




Spiking Neural Networks

- Neural dynamics defined by synapse state (Q), membrane trace (P) and refractory state (R)
- Behavior of neurons defined through time constants (alpha, beta, gamma)

$$Q_j[n+1] = \alpha Q_j[n] + S_j^{(l-1)}[n],$$
$$P_j[n+1] = \beta P_j[n] + Q_j[n],$$
$$R_i[n+1] = \gamma R_i[n] + S_i^{(l)}[n].$$

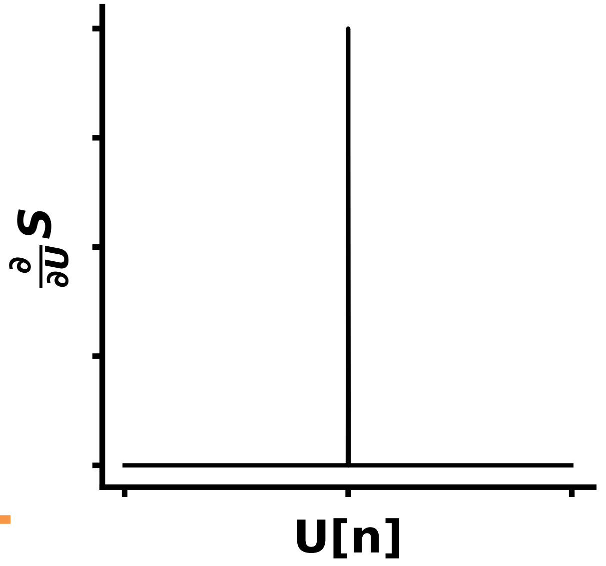


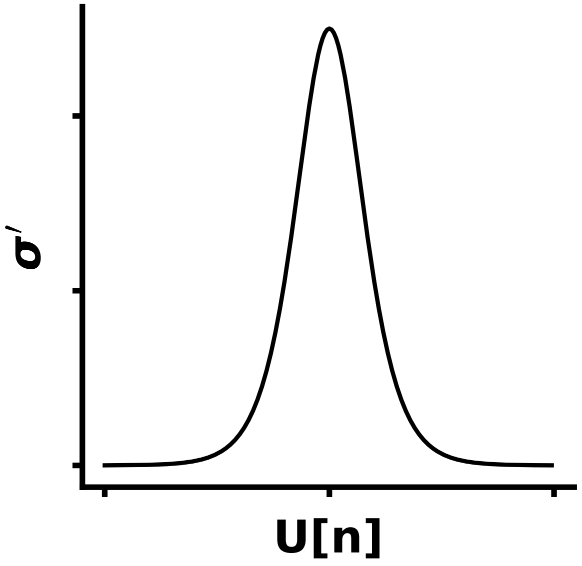
Training Spiking Neural Networks

- Evolutionary learning, e.g. genetic algorithms
- Biologically inspired training algorithm, e.g. hebbian learning (spike-timing-dependent plasticity)
- Traditional machine learning, e.g. backpropagation
- Surrogate Gradient Learning, e.g. SuperSpike [13, 8]

$$\frac{\partial}{\partial W_{ij}}\mathcal{L} = \frac{\partial}{\partial W_{ij}}U_i \frac{\partial}{\partial U_i}S_i \frac{\partial}{\partial S_i}\mathcal{L}$$

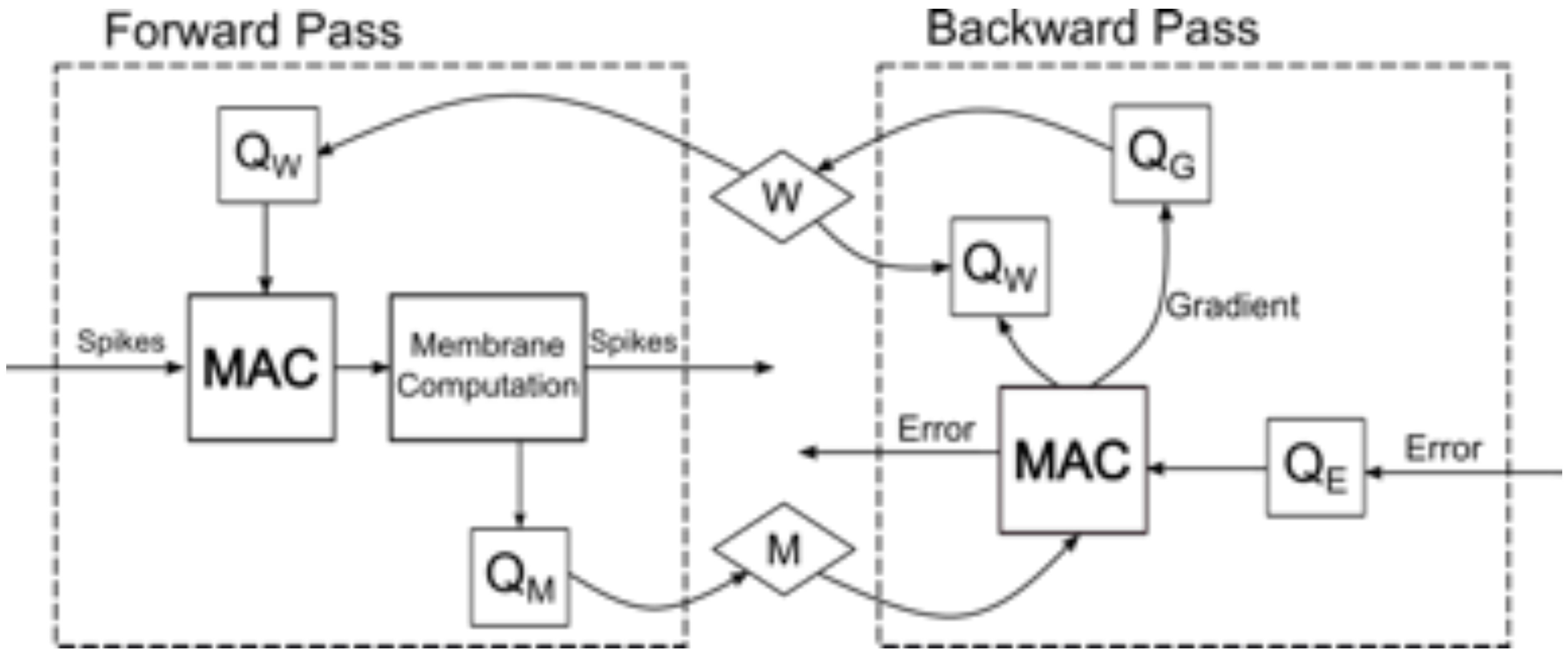
$$-\Delta W_{ij}^{l} \propto \frac{\partial}{\partial W_{ij}}\mathcal{L} = P_j[n]\sigma'(U_i^{l}[n])E_i[n]$$



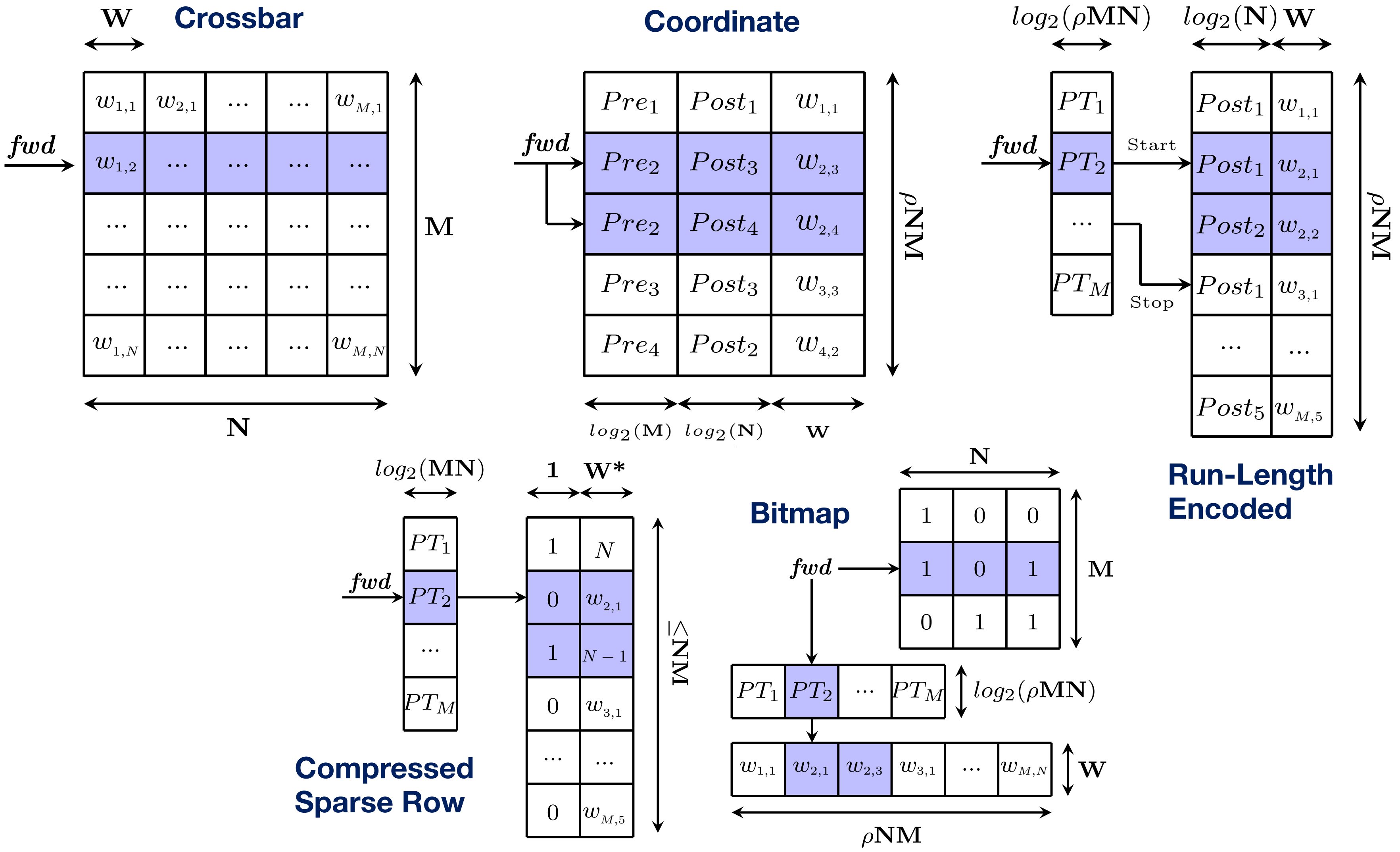


Quantizing SNNs

- Quantization to integers: clipping and rounding of values
- Forward pass: weights and copy of membrane potential
- Backward pass: errors and gradients (with stochastic rounding)



Memory Organization

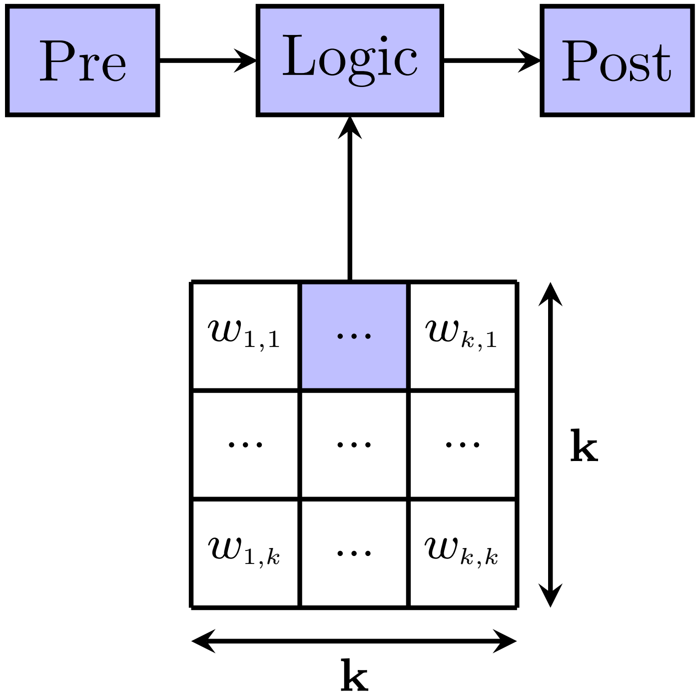


Memory Organization

Functional for 2D Convolution

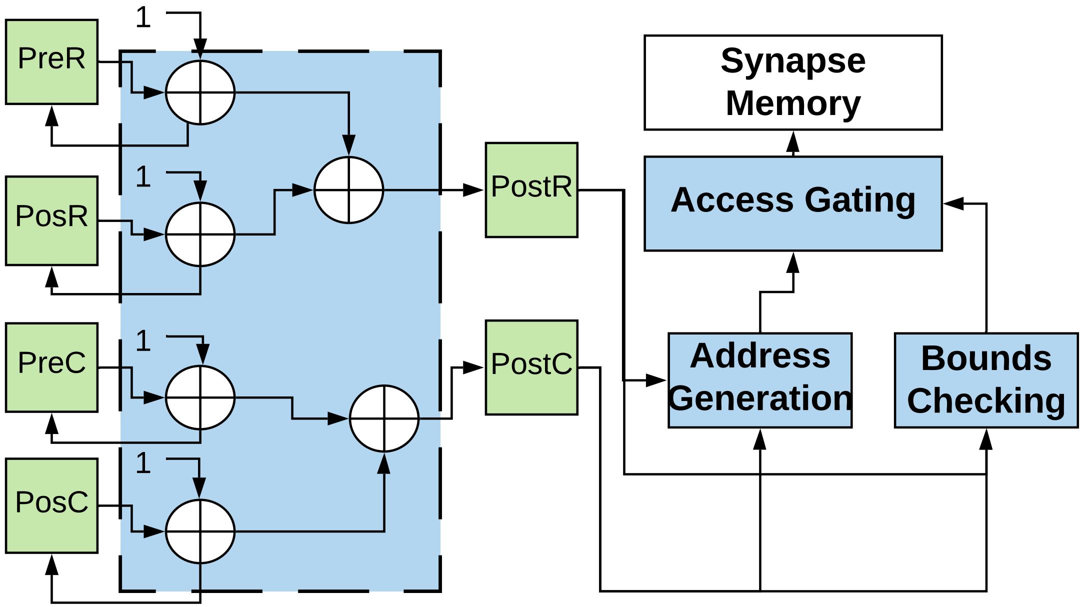
Froward pass

$$f(PreR, PosR) = PreR + PosR$$
$$f(PreC, PosC) = PreC + PosC$$



Backward pass

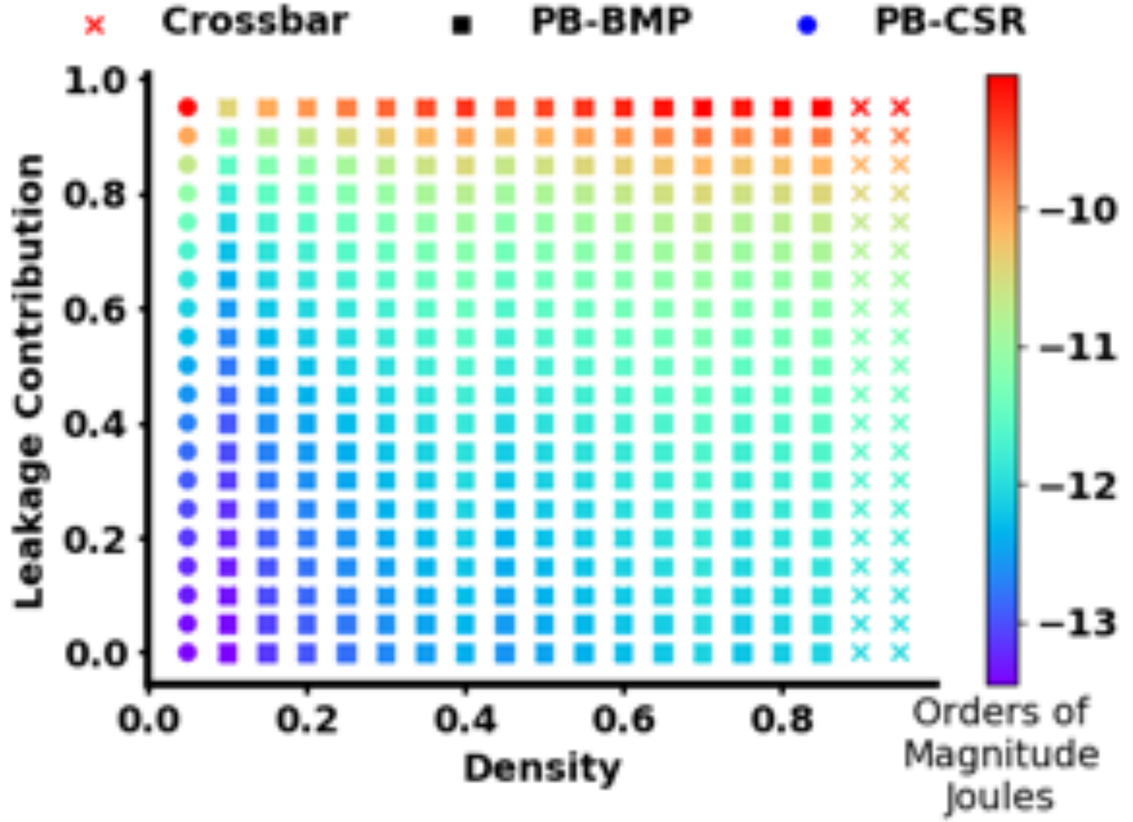
$$f^{-1}(PostR, PosR) = PostR - PosR$$
$$f^{-1}(PostC, PosC) = PostC - PosC.$$

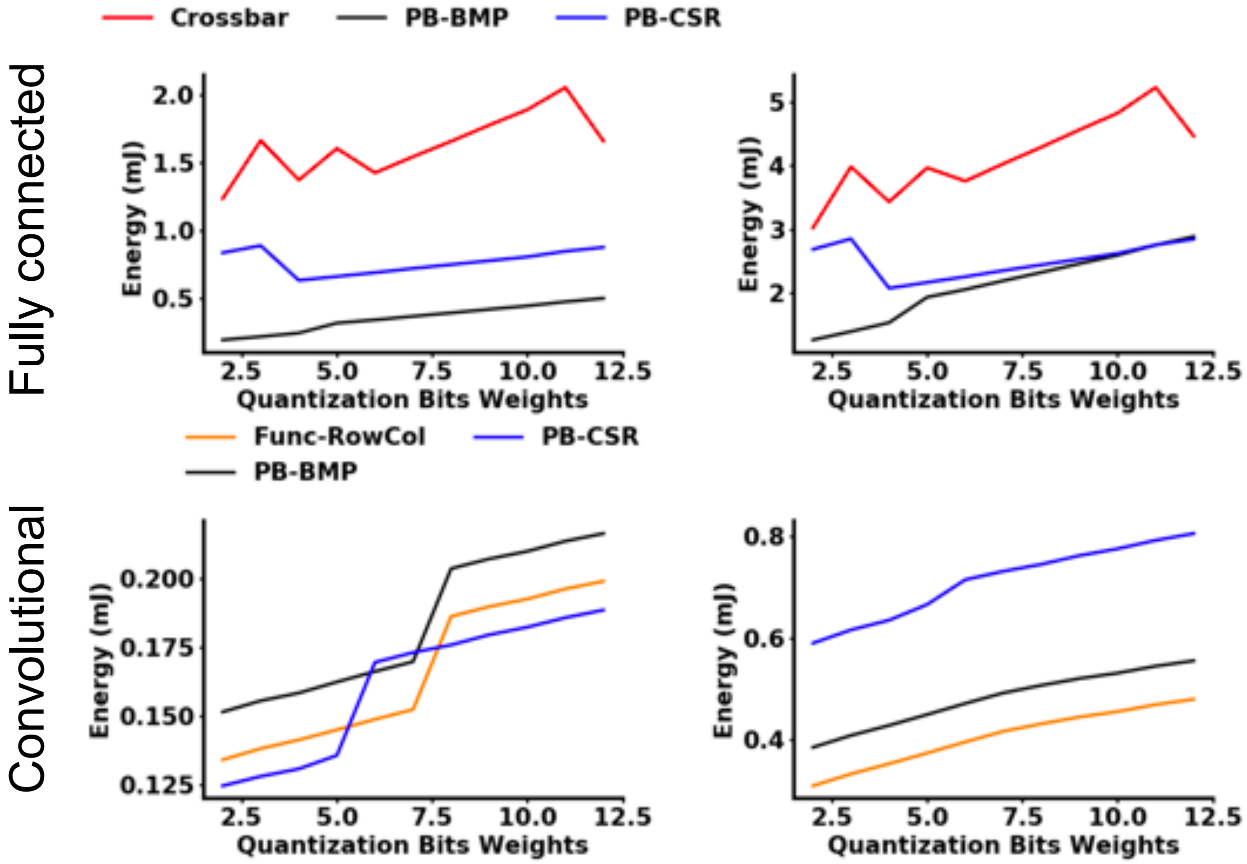


Experiments and Results

- Energy consumption of one fully connected layer (input 728, output 128) and convolutional layer (input 28x28, kernel 3x3, in channels 32, out channels 32)

Fully connected forward + backward pass

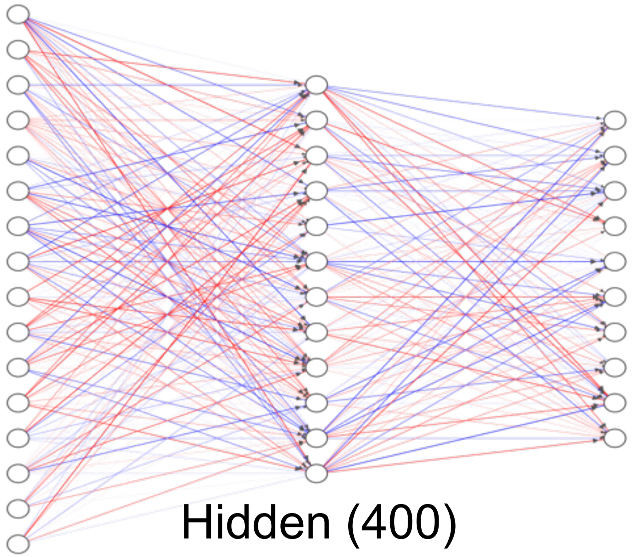




Experiment Setup

- Task: spatial-temporal pattern retention
- Leaky-integrate and fire neurons trained with BPTT and surrogate gradients
- Recording Van Rossum distance over 10000 training epochs

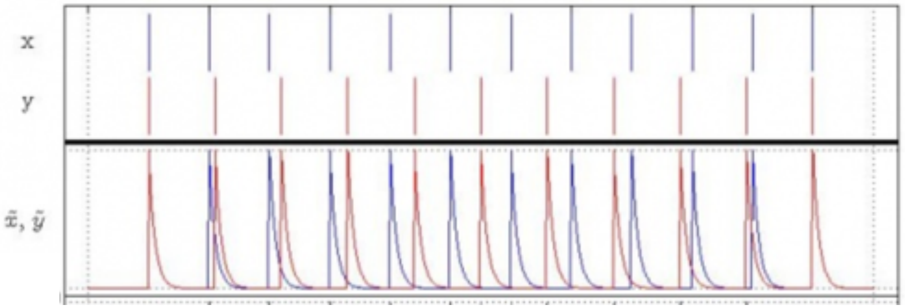
Input
(700)

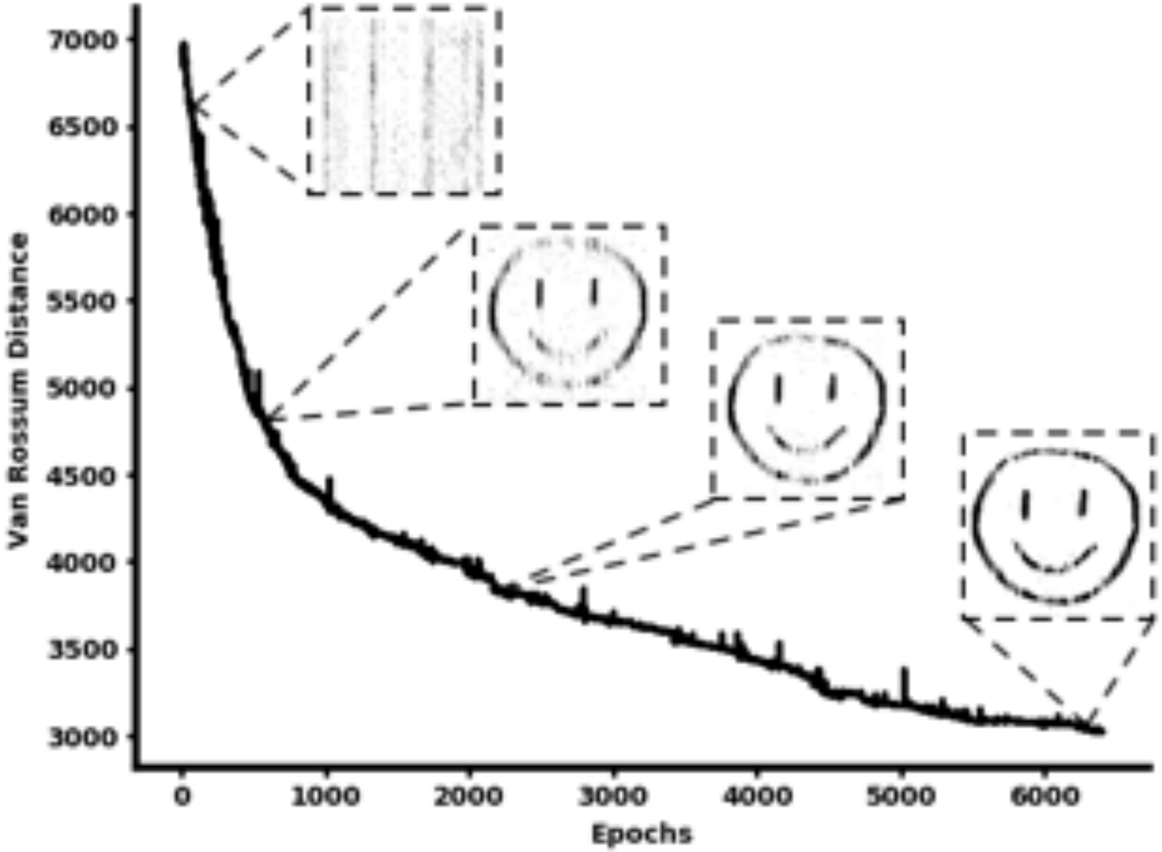


Output
(250)

Hidden (400)

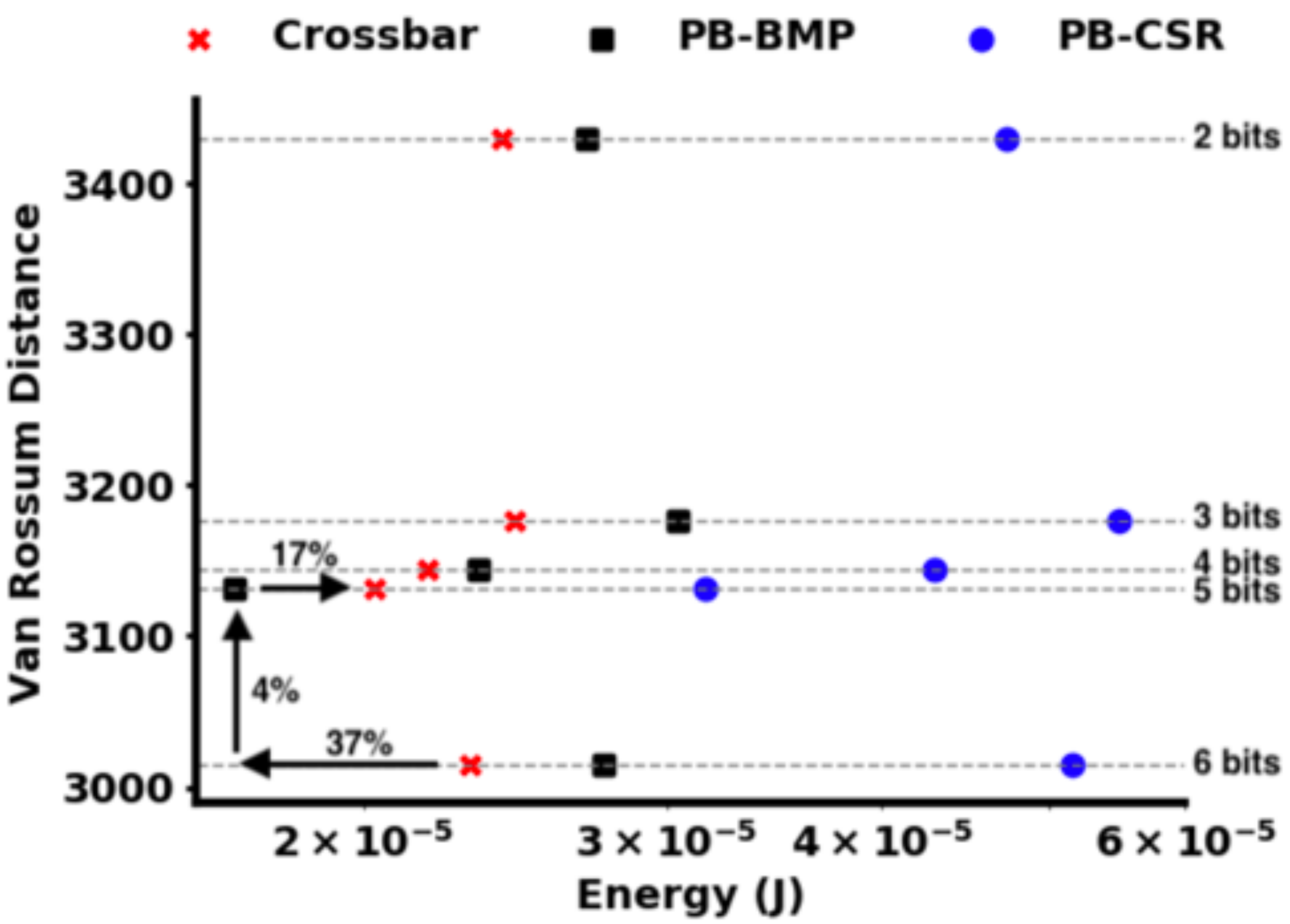
Van Rossum
Distance





Conclusion

- Encoding schemes can exploit weight sparsity for energy-efficiency
- Functional weight encoding for patterned connectivity
- Quantization induced sparsity and accuracy trade-offs translate to accuracy energy trade-offs
- Energy efficiency depends on *access efficiency* **<u>and</u>** *memory size*



References

1. S. Wu, G. Li, F. Chen, and L. Shi, “Training and inference with integers in deep neural networks,” arXiv preprint arXiv:1802.04680, 2018.
2. S. Joshi, B. U. Pedroni, and G. Cauwenberghs, “Neuromorphic event- driven multi-scale synaptic connectivity and plasticity,” in 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 2017, pp. 1–5.
3. M. v. Rossum, “A novel spike distance,” Neural computation, vol. 13, no. 4, pp. 751–763, 2001.
4. M. Davies, N. Srinivasa, T. H. Lin, G. Chinya, P. Joshi, A. Lines, A. Wild, and H. Wang, “Loihi: A neuromorphic manycore processor with on-chip learning,” IEEE Micro, vol. PP, no. 99, pp. 1–1, 2018.
5. E. Neftci, H. Mostafa, and F. Zenke, “Surrogate gradient learning in spiking neural networks,” Signal Processing Magazine, IEEE, Dec 2019, (accepted).
6. F.Zenke and S.Ganguli,“Superspike:Supervised learning in multi-layer spiking neural networks,” arXiv preprint arXiv:1705.11146, 2017.
7. J. Kim, J. Koo, T. Kim, and J.-J. Kim, “Efficient synapse memory structure for reconfigurable digital neuromorphic hardware,” Frontiers in neuroscience, vol. 12, p. 829, 2018.
8. S. Joshi, B. U. Pedroni, and G. Cauwenberghs, “Neuromorphic event- driven multi-scale synaptic connectivity and plasticity,” in 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 2017, pp. 1–5.
9. B. U. Pedroni, S. Sheik, S. Joshi, G. Detorakis, S. Paul, C. Augustine, E. Neftci, and G. Cauwenberghs, “Forward table-based presynaptic event-triggered spike-timing-dependent plasticity,” Oct 2016.
10. P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. Modha, “A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm,” in Custom Integrated Circuits Conference (CICC), 2011 IEEE, Sept. 2011, pp. 1–4.
11. B. U. Pedroni, S. Joshi, S. Deiss, S. Sheik, G. Detorakis, S. Paul, C. Augustine, E. O. Neftci, and G. Cauwenberghs, “Memory-efficient synaptic connectivity for spike-timing-dependent plasticity,” Frontiers in neuroscience, vol. 13, p. 357, 2019.