UNIVERSITY of York

Ulster University

UKRI Engineering and Physical Sciences Research Council

NERVOUS SYSTEMS

Neuromorphic Hardware Overview

MARTIN TREFZER, ANDY TYRRELL, ANDREW WALTER & SHIMENG WU

SCHOOL OF PHYSICS, ENGINEERING & TECHNOLOGY

UNIVERSITY OF YORK

JIM HARKIN, LIAM MCDAID, MALACHY MCELHOLM & THANDASSERY NIDHIN

SCHOOL OF COMPUTING, ENGINEERING & INTELLIGENT SYSTEMS

ULSTER UNIVERSITY

XILINX

ThalesAlenia Space

a Thales / Leonardo company

arm

What is “Neuromorphic” Hardware?

NEURO: “RELATING TO NERVES OR THE NERVOUS SYSTEM”
MORPHIC: “OF OR PERTAINING TO FORM”

- **Term coined by Carver Mead in 1990[1], referring to analogue VLSI mimicking biological neural systems.**
- **The original idea of Brain-like computing is much older, discussed by Alan Touring[2] and John von Neumann[3] in the 1950’s.**
- **Neuromorphic systems are typically based on some combination of:**
  - **analogue data processing**
  - **asynchronous communication**
  - **massively parallel information processing**
  - **spiking-based information representation**
  - **collocating memory and processing**
  - **efficient, robust, low-power**

MARTIN.TREFZER@YORK.AC.UK

Neuromorphic vs von Neumann

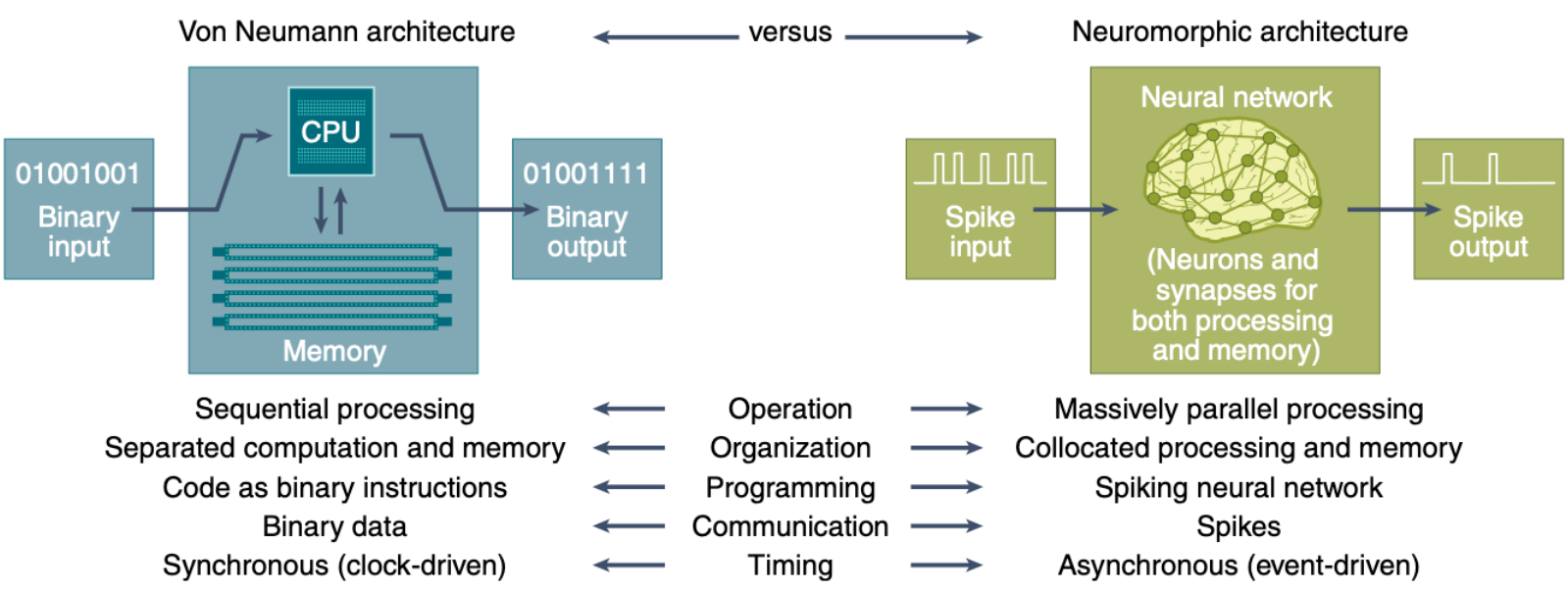


Image from [16] Schuman, Catherine D., Shruti R. Kulkarni, Maryam Parsa, J. Parker Mitchell, Prasanna Date, and Bill Kay. 2022. "Opportunities for Neuromorphic Computing Algorithms and Applications." Nature Computational Science 2 (1): 10–19.

Why Neuromorphic Hardware?

1980's Parallelism[4-6]
Simple processing elements, densely connected (Brain inspired).

von Neumann bottleneck

1990's Real-time Comp[10-11]
(Autonomous) robot control, image reconstruction, parallelism for application performance.

End of Dennard scaling

2010's Low-power[13]
20W Brain power, no A2D/D2A, AI/ML for edge applications.

1990's Computation Speed[7-9]
Inherent parallelism, machine learning acceleration with custom HW.

New devices and materials

late 1990's Robustness[4,12]
Inherent fault-tolerance, capability for HW to adapt and self-heal, advantage when dealing with intrinsic variability of novel devices.

Deep learning "AI"

All-time goals[14-16]**:**
- **Scalability**
- **von Neumann bottleneck (in-memory computing)**
- **Autonomous – online learning**
- **Neuroscience**

MARTIN.TREFZER@YORK.AC.UK

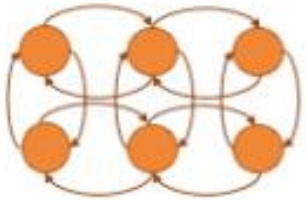
Designing Neuromorphic HW Systems

NETWORK TOPOLOGY AND HARDWARE MAPPING

Feed-forward

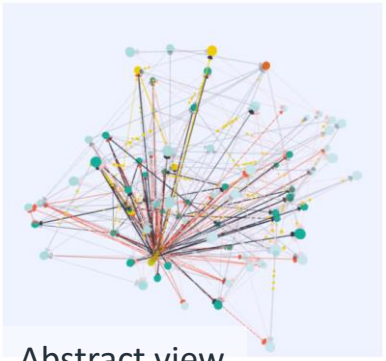
Feed-forward
with some recurrence

Sparsely-connected
Recurrent



Locally-connected
Recurrent

Fully-connected
Recurrent



Abstract view

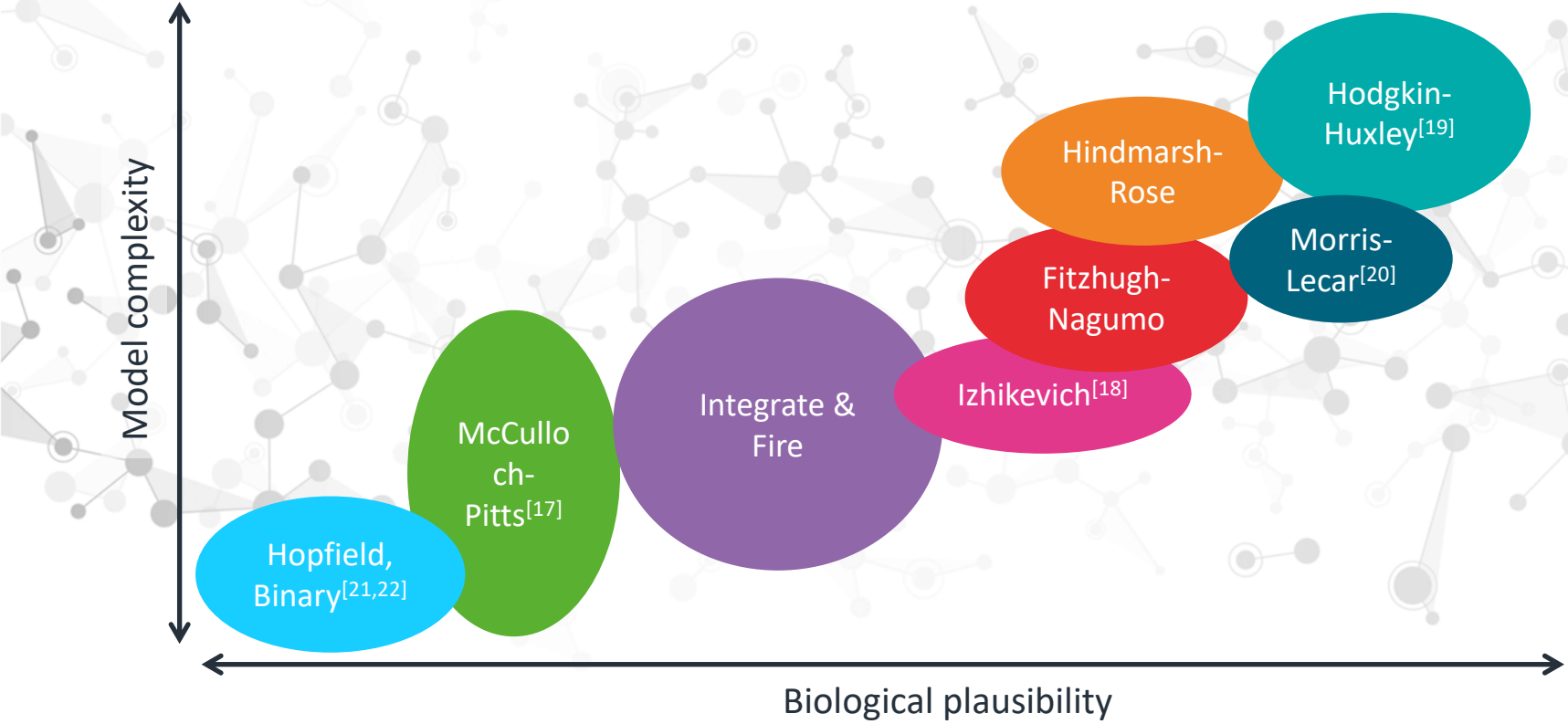
Hardware mapping

Illustrations from [15] Schuman, Catherine D., Thomas E. Potok, Robert M. Patton, J. Douglas Birdwell, Mark E. Dean, Garrett S. Rose, and James S. Plank. 2017. “A Survey of Neuromorphic Computing and Neural Networks in Hardware.” arXiv [cs.NE]. arXiv. http://arxiv.org/abs/1705.06963.

MARTIN.TREFZER@YORK.AC.UK

Neuron (& Synapse) Models

BIO-INSPIRED COMPUTATION MODEL ABSTRACTIONS

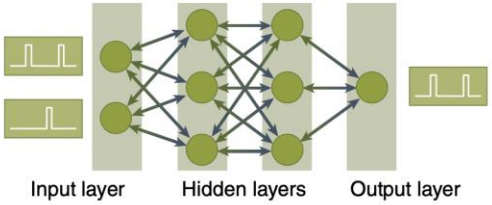


MARTIN.TREFZER@YORK.AC.UK

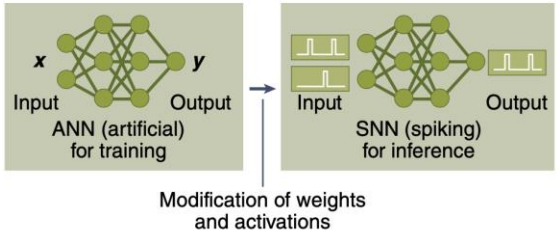
Learning Rules and Training Methods

- Supervised
- Unsupervised
- L2L

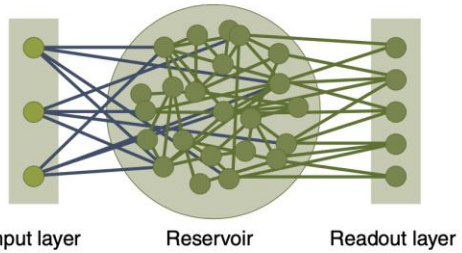
a (Quasi) Backpropagation



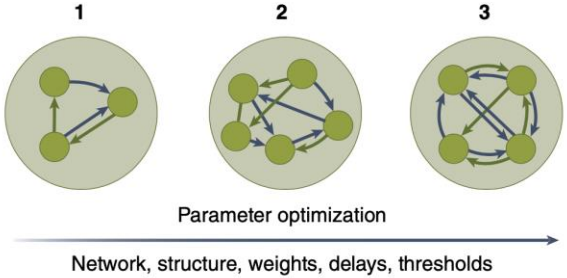
b Mapping Post-training



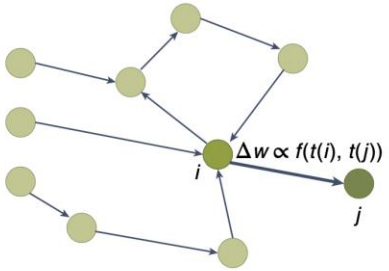
c Reservoir Computing



d Evolutionary Optimisation



e Hebbian Rules, STDP



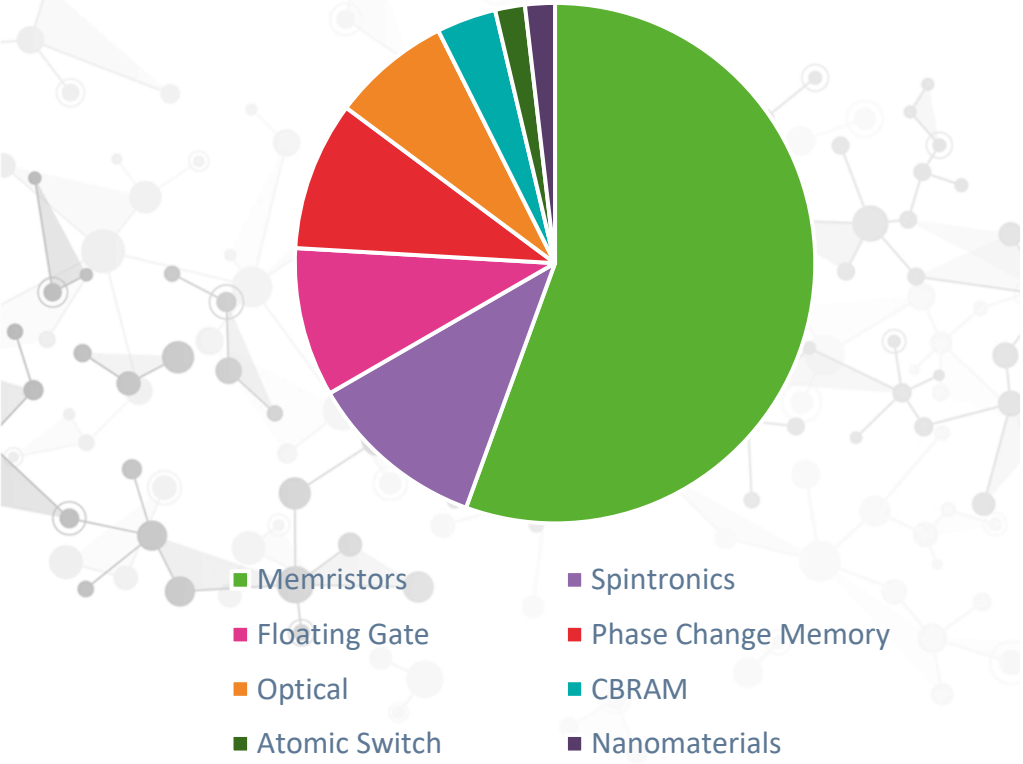
ALGORITHMS PROS AND CONS

Algorithm Class	Any Model	Device Quirks	Complex to Implement	On-Line	Fast Time to Solution	Demonstrated Broad Applicability	Biologically-Inspired or Plausible
Back-Propagation	No	No	Yes	No	Yes	Yes	No
Evolutionary	Yes	Yes	No	No	No	Yes	Maybe
Hebbian	No	Yes	No	Yes	Maybe	No	Yes
STDP	No	Yes	Maybe	Yes	Maybe	No	Yes

Image sources [15,16]

MARTIN.TREFZER@YORK.AC.UK

Materials and Technologies



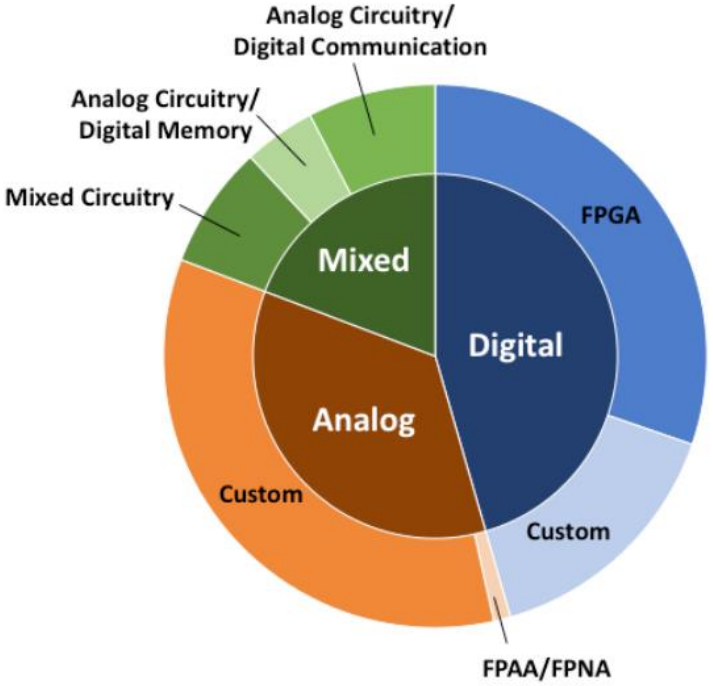
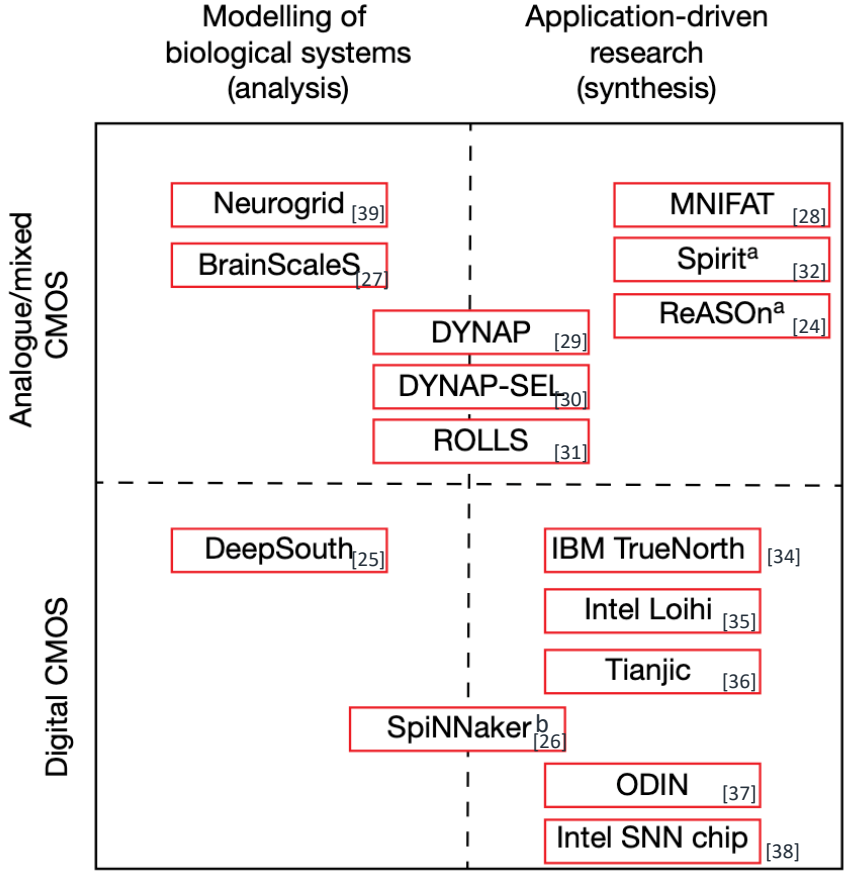


Image source [15], VLSI [40]

MARTIN.TREFZER@YORK.AC.UK

Landscape of Neuromorphic Hardware



Including properties such as:

- In-memory computing
- Fine-grained parallelism
- Learning in hardware
- Event-based and asynchronous communication
- Reduced precision
- Spike-based processing
- Adaptability
- Leveraging noise and stochasticity
- Brain-inspired

a. memristor technology
b. SpiNNaker2 with FP+other features

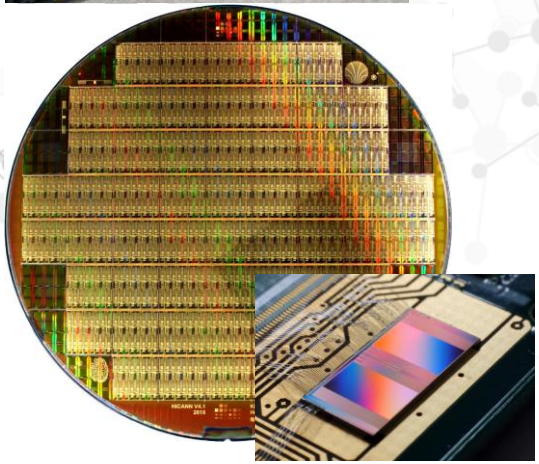
Image source Mehonic et al. [23]

MARTIN.TREFZER@YORK.AC.UK

BrainScaleS vs SpiNNaker (HBP)

BrainScaleS

- physical VLSI (analogue or mixed-signal)
- emulations of neuron, synapse and plasticity models
- digital connectivity
- 10,000X times faster than real time



SpiNNaker

- digital multicore chips
- numerical models of neuron, synapse and plasticity on ARM
- custom optimised connectivity
- 10X real time on 1-Mio cores

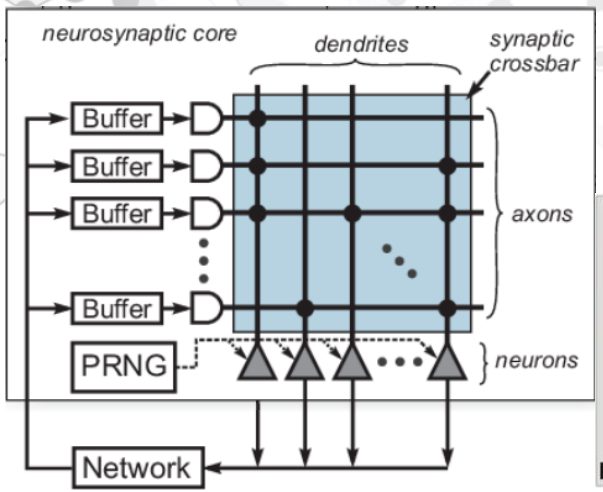
https://wiki.ebrains.eu/bin/view/Collabs/neuromorphic/

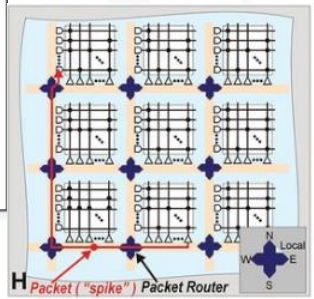
MARTIN.TREFZER@YORK.AC.UK

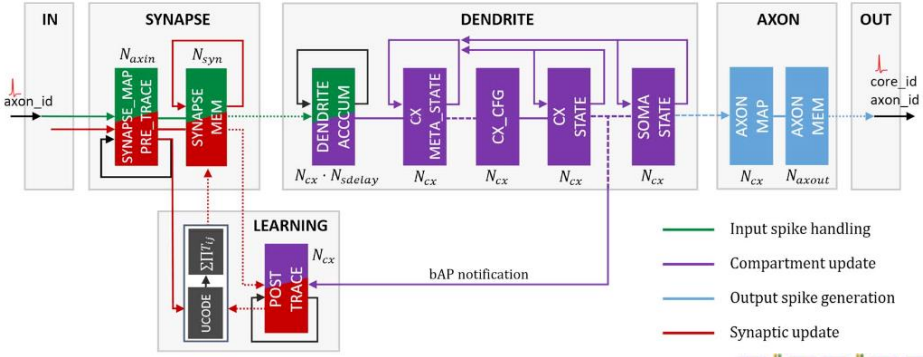
TrueNorth & Loihi

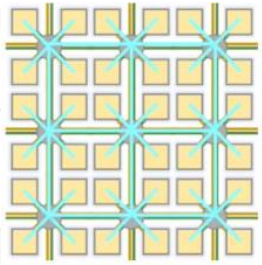
IBM TrueNorth [34,41]

- 4096 Neuro Cores
- Each 256 fully-connected neurons
- 256x256 crossbar (locally synchronous)
- mesh event routing between Neuro Cores (globally asynchronous)
- LIF neurons, multi-valued synapses







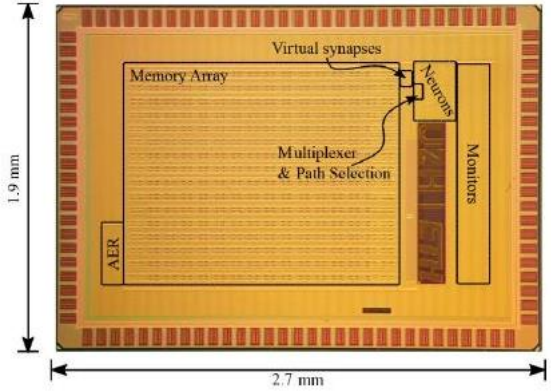


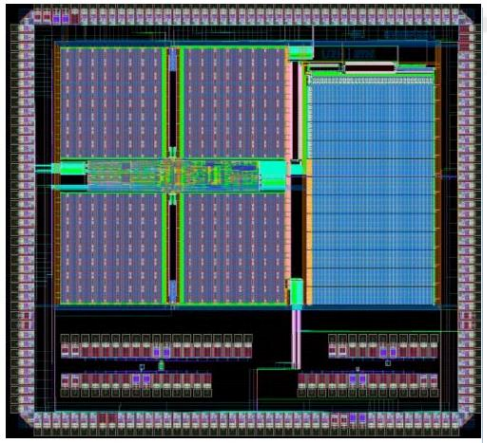
Intel Loihi [35]

- 128 Neuromorphic Cores
- Each 1024 primitive neurons
- scalable mesh routing (locally synchronous, globally asynchronous)
- LIF neurons, precision synapses
- Programmable on-chip synaptic learning rules (STDP)
- Programmable synaptic delays

MARTIN.TREFZER@YORK.AC.UK

ReASOn / Dynap-sel (28nm FD-SOI) [24]

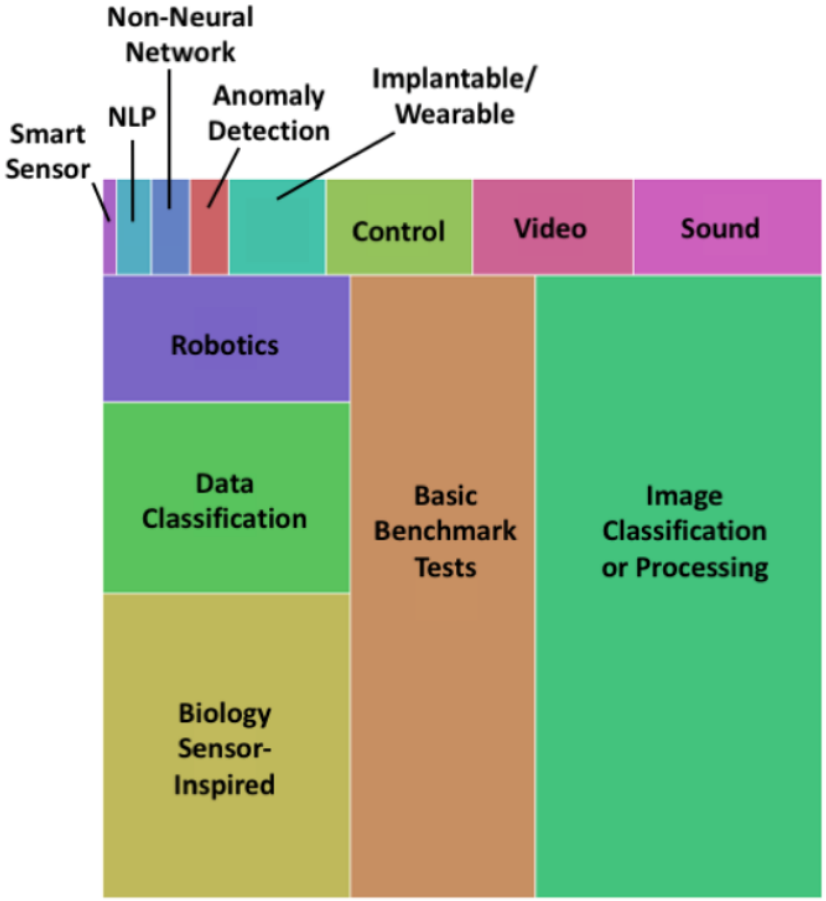




Block	Number	Detail
Non-Plastic Core	X4	256 X 4 analog AExp Leaky I&F neurons, 16k X4 TCAM-based synapses
Plastic Core	X1	64 analog AExp Leaky I&F neurons, 8k digital plastic synapses 4k digital non-plastic synapses 256 virtual synapses
Neuron		Computation: Pure current based Biological features: NMDA, frequency adaptation, tunable refractory period, firing threshold, fast/slow leaky time constants. Tunable Parameter: 12 biases for Ith, Itau1/2, Idc, Irefr, Imemthr... Local latches for configuration: NMDA, Monitor, tau1/2
Non-plastic Synapse		11-bit TCAM(synapse addresses)+5-bit SRAM (Exc/Inh, 4-bit synaptic weight)
Plastic Synapse		4-bit up/down counter-based digital with Fusi stop learning rule Configurable feature: Exc/Inh, Set weight from 0 to 15, weight monitor enable, Learning Enable/disable, Broadcast event enable/disable

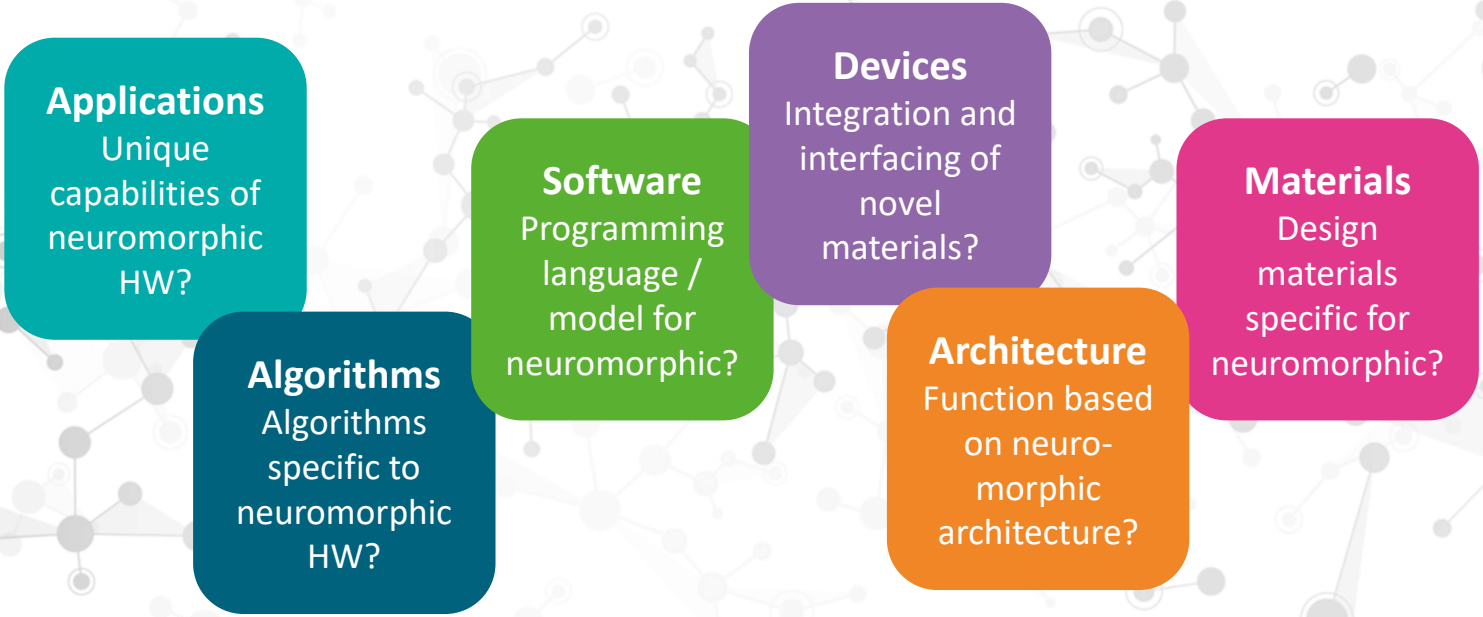
MARTIN.TREFZER@YORK.AC.UK

Overview of Applications – Next Talk!



- Data from 2017[15]
- In 2023, probably larger portion of smart sensors and edge AI

Research Areas and Challenges



MARTIN.TREFZER@YORK.AC.UK

References

[1] C. Mead, “Neuromorphic electronic systems,” Proceedings of the IEEE, vol. 78, no. 10, pp. 1629–1636, Oct 1990.
[2] A.M.Turing,“Computingmachineryandintelligence,”Mind,vol.59, no. 236, pp. 433–460, 1950.
[3] J. Von Neumann and R. Kurzweil, The computer and the brain. Yale University Press, 2012.
[4] A. F. Murray and A. V. Smith, “Asynchronous VLSI neural networks using pulse-stream arithmetic,” Solid-State Circuits, IEEE Journal of, vol. 23, no. 3, pp. 688–697, 1988.
[5] F. Blayo and P. Hurat, “A vlsi systolic array dedicated to Hopfield neural network,” in VLSI for Artificial Intelligence. Springer, 1989, pp. 255–264.
[6] F. Distante, M. Sami, and G. S. Gajani, “A general configurable architecture for VLSI implementation for neural nets,” in Wafer Scale Integration, 1990. Proceedings.,[2nd] International Conference on. IEEE, 1990, pp. 116–123.
[7] J. B. Burr, “Digital neural network implementations,” Neural networks, concepts, applications, and implementations, vol. 3, pp. 237– 285, 1991.
[8] M. Chiang, T. Lu, and J. Kuo, “Analogue adaptive neural network circuit,” IEEE Proceedings G (Circuits, Devices and Systems), vol. 138, no. 6, pp. 717–723, 1991.
[9] A. F. Murray, D. Del Corso, and L. Tarassenko, “Pulse-stream vlsi neural networks mixing analog and digital techniques,” Neural Networks, IEEE Transactions on, vol. 2, no. 2, pp. 193–204, 1991.
[10] J.-C. Lee and B. J. Sheu, “Parallel digital image restoration using adaptive vlsi neural chips,” in Computer Design: VLSI in Computers and Processors, 1990. ICCD’90. Proceedings, 1990 IEEE International Conference on. IEEE, 1990, pp. 126–129.
[11] L. Tarassenko, M. Brownlow, G. Marshall, J. Tombs, and A. Murray, “Real-time autonomous robot navigation using vlsi neural networks,” in Advances in neural information processing systems, 1991, pp. 422– 428.
[12] L. Akers, M. Walker, D. Ferry, and R. Grondin, “A limited- interconnect, highly layered synthetic neural architecture,” in *VLSI for artificial intelligence*. Springer, 1989, pp. 218–226.
[13] J. Schemmel, D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, “A wafer-scale neuromorphic hardware system for large- scale neural modeling,” in Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010, pp. 1947– 1950.
[14] H. Markram, “The human brain project,” Scientific American, vol. 306, no. 6, pp. 50–55, 2012.
[15] Schuman, Catherine D., Thomas E. Potok, Robert M. Patton, J. Douglas Birdwell, Mark E. Dean, Garrett S. Rose, and James S. Plank. 2017. “A Survey of Neuromorphic Computing and Neural Networks in Hardware.” arXiv [cs.NE]. arXiv. http://arxiv.org/abs/1705.06963.
[16] Schuman, Catherine D., Shruti R. Kulkarni, Maryam Parsa, J. Parker Mitchell, Prasanna Date, and Bill Kay. 2022. “Opportunities for Neuromorphic Computing Algorithms and Applications.” Nature Computational Science 2 (1): 10–19.

MARTIN.TREFZER@YORK.AC.UK

References

[17] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas im- manent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
[18] E. M. Izhikevich, “Which model to use for cortical spiking neurons?” *IEEE transactions on neural networks*, vol. 15, no. 5, pp. 1063–1070, 2004.
[19] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.
[20] A. Borisyuk, “Morris–lecar model,” in *Encyclopedia of Computa- tional Neuroscience*. Springer, 2015, pp. 1758–1764.
[21] A. Deshmukh, J. Morghade, A. Khera, and P. Bajaj, “Binary neural networks–a cmos design approach,” in Knowledge-Based Intelligent Information and Engineering Systems. Springer, 2005, pp. 1291– 1296.
[22] P. W. Hollis and J. J. Paulos, “An analog bicmos hopfield neuron,” in *Analog VLSI Neural Networks*. Springer, 1993, pp. 11–17.
[23] Mehonic, A., and A. J. Kenyon. 2022. “Brain-Inspired Computing Needs a Master Plan.” *Nature* 604 (7905): 255–60.
[24] Resistive Array of Synapses with ONline Learning (ReASOn) Developed by NeuRAM3 Project https://cordis.europa.eu/project/id/687299/reporting (2021).
[25] Wang, R. et al. Neuromorphic hardware architecture using the neural engineering framework for pattern recognition. IEEE Trans. Biomed. Circuits Syst. 11, 574–584 (2017).
[26] Furber, S. B., Galluppi, F., Temple, S. & Plana, L. A. The SpiNNaker Project. Proc. IEEE 102, 652–665 (2014).
[27] Schmitt, S. et al. Neuromorphic hardware in the loop: training a deep spiking network on the BrainScaleS wafer-scale system. In 2017 Intl Joint Conf. Neural Networks (IJCNN) https://doi.org/10.1109/ijcnn.2017.7966125 (IEEE, 2017).
[28] Lichtsteiner, P., Posch, C. & Delbruck, T. A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J. Solid-State Circuits 43, 566–576 (2008).
[29] Moradi, S., Qiao, N., Stefanini, F. & Indiveri, G. A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs). IEEE Trans. Biomed. Circuits Syst. 12, 106–122 (2018).
[30] Thakur, C. S. et al. Large-scale neuromorphic spiking array processors: a quest to mimic the brain. Front. Neurosci. 12, 891 (2018).
[31] Qiao, N. et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. Front. Neurosci. 9, 141 (2015).
[32] Valentian, A. et al. in 2019 IEEE Intl Electron Devices Meeting (IEDM) 14.3.1–14.3.4 https://doi.org/10.1109/IEDM19573.2019.8993431 (IEEE, 2019).
[33] Mehonic, A. et al. Memristors—from in-memory computing, deep learning acceleration, and spiking neural networks to the future of neuromorphic and bio-inspired computing. Adv. Intell Syst. 2, 2000085 (2020).
[34] Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).

MARTIN.TREFZER@YORK.AC.UK

References

[35] Davies, M. et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).
[36] Pei, J. et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019). 2
[37] Frenkel, C., Lefebvre, M., Legat, J.-D. & Bol, D. A 0.086-mm 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS. *IEEE Trans. Biomed. Circuits Syst.* **13**, 145–158 (2018).
[38] Chen, G. K., Kumar, R., Sumbul, H. E., Knag, P. C. & Krishnamurthy, R. K. A 4096-neuron 1M-synapse 3.8-pJ/SOP spiking neural network with on-chip STDP learning and sparse weights in 10-nm FinFET CMOS. *IEEE J. Solid-State Circuits* **54**, 992–1002 (2019).
[39] Benjamin, B. V. et al. Neurogrid: a mixed-analog–digital multichip system for large-scale neural simulations. *Proc. IEEE* **102**, 699–716 (2014).
[40] Indiveri, G. et al. Neuromorphic silicon neuron circuits. *Front. Neurosci.* **5**, 73 (2011).
[41] Cassidy, Andrew, Merolla, Paul et al. (2013). Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores. Proceedings of the International Joint Conference on Neural Networks. 10.1109/IJCNN.2013.6707077.

MARTIN.TREFZER@YORK.AC.UK