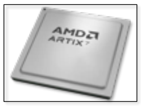
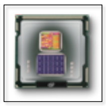
Christof Teuscher

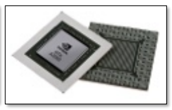
ECE 410/510: Hardware for AI and ML

Emerging devices and technologies

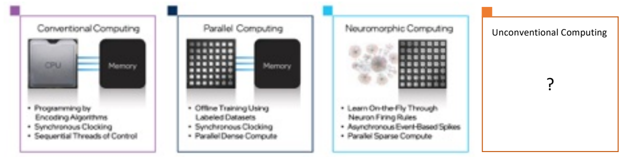
Portland State University
Department of Electrical and Computer Engineering (ECE)
www.teuscher-lab.com
teuscher@pdx.edu





Unconventional computing

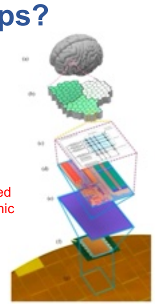


Technology Brief | Taking Neuromorphic Computing to the Next Level with Loihi 2

What are neuromorphic chips?

- Neuromorphic chips are specialized hardware designed to mimic the structure and function of the human brain. **Unlike traditional computing architectures, these chips are built to process information in ways similar to biological neural networks.**
  - Cerebras WSE-3
  - 300-millimeter wafer
  - 4 trillion transistors
  - TSMC 5nm
  - 57x larger than an NVIDIA H100
  - Accelerate AI workloads, particularly training and inference of large language models.

Not considered a neuromorphic chip



The usual trade-offs…!

Why not do everything in software?

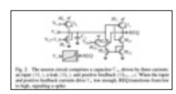
SpiNNaker

Akida
Loihi
Loihi/2
BrainScaleS/2
T1

IBM's TrueNorth
IBM's NorthPole

Flexibility
More digital circuits

Other trade-offs:
- Power
- Reliability
- Programmability
- On-chip learning
- Cost
- Scalability



LIF neuron
DOI: 10.1109/TCSI.2010.2089556

Performance
More analog circuits

But:
- Digital backend needed
- DAC/ADC

TABLE 3 Comparison of neuromorphic chips.

Chip/neural computer	In-memory computation	Signal	Size neurons/synapses	On-device learning	Analog	Event-based	nm	Features
CPU/GPU/TPU	No	Real numbers, spikes	-	Backprop/STDP	No	No	5	High popularity, rich ecosystem, advanced engineering technologies
TrueNorth	Near-memory	Spikes	1M/256M	No	No	Yes	28	First industrial neuromorphic chip without training (IBM)
Loihi	Near-memory	Spikes	128K/128M	STDP	No	Yes	14	First neuromorphic chip with training (Intel)
Loihi2	Near-memory	Real numbers, spikes	120K/1M	STDP, surrogate backprop	No	Yes	7	Development of Loihi ideas, non-binary spikes, neurons can be programmed
Tianjic	Near-memory	Real numbers, spikes	40K/10M	No	No	Yes	28	Hybrid chip with effective support of both SNN and ANN, energy efficiency
SpiNNaker	Near-memory	Real numbers, spikes	-	STDP	No	No	22	Scalable computer for SNN simulation
Brain-ScaleS	Yes	Real numbers, spikes	512/130K	STDP, Surrogate gradient	Yes, membrane	Yes	65	Analog neurons at RC circuits, large size
GrAIOne (Neuron- Flow)	Near-memory	Real numbers, Spikes	200K/	No	No	Yes	28	NeuronFlow architecture, effective support of sparse computations, support of ANN and SNN
DYNAP SE2, SEL, CNN	Near-memory	Spikes	1K/65K 1K/80K 1M/4M	STDP (SEL)	SE2, SEL	Yes	22	Proprietary communication protocol
Akida	Near-memory	Spikes	1.2M/10B	STDP (last layer)	No	Yes	28	First commercial neuromorphic processor with incremental, one-shot, and continuous learning for CNN
Mythic	In-memory	Real numbers	/80M	-	Yes	Yes	40	-
Memristor (Tsinghua University)	Yes	Real numbers	192/ 2048	No	Yes (15 signal levels)	Yes	500	CNN-optimized memristor chip, one chip contains 2048 1T1R elements
Memristor (Univ. of Massachusetts)	Yes	Spikes	192/ 2048	No	Yes	Yes	2 μm	128 × 64 memristor array according to 1T1R circuit
Memristor (IBM)	Yes	Spike	512/ 64k	Yes	Yes	Yes	50	2T1R design allows each synaptic cell to operate asynchronously in either LIF or STDP mode

https://doi.org/10.3389/fnins.2022.959626

Feature	Loihi	Loihi 2
Release Date	2017	2021
Manufacturing Process	14nm	Intel 4 (pre-production)
Die Size	60mm²	31mm² (about half the size)
Processor Cores	128 neuromorphic cores + 3 x86 cores	128 neuromorphic cores + 6 Lakemont x86 cores
Neuron Capacity	~130,000 neurons	Up to 1 million neurons (approximately 8x more)
Synapse Capacity	~130 million synapses	Up to 120 million synapses
On-chip Memory	Over 33MB SRAM	~25MB SRAM
Neuron Model	Fixed leaky integrate-and-fire	Programmable neuron models with custom microcode
Spike Processing Speed	Baseline	Up to 10x faster than Loihi
Spike Events	1-bit (binary)	Up to 32-bit graded spike events
Learning Capabilities	Basic learning rules	Enhanced programmable learning with support for more algorithms
Connectivity	FPGA mediated	Direct support for 10G Ethernet, GPIO, and SPI
Software Framework	None at launch	Lava (open-source software framework)
Power Efficiency	High (compared to conventional solutions)	Further improved
Application Focus	Basic neuromorphic algorithms	Expanded for multiple AI methods and applications
Backpropagation Support	Limited	Improved approximation of backpropagation algorithm
Programmability	Specialized for specific SNN model	Fully programmable with support for arithmetic, comparison, and control flow operations

Key Improvements in Loihi 2

1. **Increased Neural Density**: 8x more neurons in half the die area
2. **Programmable Neuron Models**: Support for custom neuron behaviors beyond fixed models
3. **Faster Processing**: Up to 10x faster spike processing
4. **Better Connectivity**: Improved chip-to-chip communication with 10G Ethernet support
5. **Enhanced Learning**: Support for more advanced learning algorithms
6. **Lava Software Framework**: Open-source tooling for development
7. **Expanded Application Range**: Support for broader workloads and use cases
8. **Improved Integration**: Easier connection to conventional computing systems

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

Loihi 2

Loihi implements the discrete Leaky Integrate and Fire (LIF) neuron model

$$a_i[t] = \sum_j w_{ij} s_j[t-1] \quad (1)$$
$$u_i[t] = \lambda_u u_i[t-1] + a_i[t] \quad (2)$$
$$v_i[t] = \lambda_v v_i[t-1] + u_i[t] \quad (3)$$

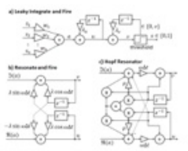


Fig. 2: Block diagrams of the discrete computation for the LIF neuron, and RF and Hopf resonators. The Hopf Resonator shows an implementation of the Euler method, but in practice we use a 4th order Runge-Kutta.

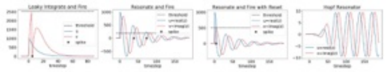


Fig. 1: Response of neuron models used in this paper to an impulse (spike) at time 0. a) A Leaky Integrate and Fire model which spikes whenever voltage exceeds threshold. b) A complex valued Resonate and Fire model used for spatiotemporal filtering, which spikes whenever z crosses the real axis and $real(z)$ is greater than threshold. c) A complex valued Resonate and Fire model with reset used in deep networks. d) A complex valued Hopf Resonator with a stable limit cycle.

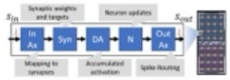


Fig. 3: The Loihi 2 chip plot (right) and processing flow for a single core (left). Incoming spikes are mapped to lists of synapse weights which are accumulated for consumption in the next timestep. Meanwhile, neurons update using the previous timestep's accumulated activation and generate spikes which are routed to other cores by the Output Axon stage.

teuscher Lab teuscher-lab.com   DOI 10.1109/SiPS52927.2021.00053   Portland State University

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

LLM on Loihi 2

Published at ICLR 2025 Workshop (SCOPE)

v2 [cs.NE] 25 Mar 2025

NEUROMORPHIC PRINCIPLES FOR EFFICIENT LARGE LANGUAGE MODELS ON INTEL LOIHI 2

Steven Abreu, University of Groningen & Intel Labs
Sumit Bam Shrestha, Intel Labs
Rui-Jie Zhu, University of California, Santa Cruz
Jason Eshraghian, University of California, Santa Cruz

ABSTRACT

Large language models (LLMs) deliver impressive performance but require large amounts of energy. In this work, we present a MatMul-free LLM architecture adapted for Intel's neuromorphic processor, Loihi 2. Our approach leverages Loihi 2's support for low-precision, event-driven computation and stateful processing. Our hardware-aware quantized model on GPU demonstrates that a 370M-parameter MatMul-free model can be quantized with no accuracy loss. Based on preliminary results, we report up to 3× higher throughput with 2× less energy, compared to transformer-based LLMs on an edge GPU, with significantly better scaling. Further hardware optimizations will increase throughput and decrease energy consumption. These results show the potential of neuromorphic hardware for efficient inference and pave the way for efficient reasoning models capable of generating complex, long-form text rapidly and cost-effectively.

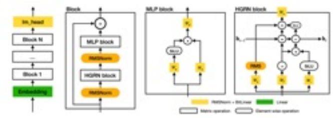


Figure 1: Model architecture of the MatMul-free language model from Zhu et al. (2024).

Why is that gain performance not impressive?

teuscher Lab teuscher-lab.com   arXiv:2503.18002v2   Portland State University

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

Akida

- Akida (Vanarse et al., 2019) is the first commercial neuromorphic processor, commercially available since August 2021.
- BrainChip's Akida is an ultra-low-power neuromorphic processor inspired by the brain's neural architecture. It accelerates complex AI at the edge through event-based processing, on-chip learning abilities, and support form advanced neural networks like CNNs, RNNs & custom Temporal Event-based Nets.
- **Temporal Event-based Neural Networks (TENNs), which combine spatial and temporal convolutions to efficiently process sequential spatiotemporal data.**
- Akida's ecosystem provides a free chip emulator, TensorFlow compatible framework MetaTF for the transformation of convolutional and fully connected neural networks into SNN, and a set of pre-trained models.

teuscher Lab teuscher-lab.com   Portland State University

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

Do we need brain-like HW?

"Although efforts to address neuromorphic solutions through hardware based on top-down CMOS-based technologies have obtained interesting results in terms of energetic efficiency improvement, the replication of brain's **self-assembled** and **redundant architectures** is not considered in the roadmaps of data processing electronics."

teuscher Lab teuscher-lab.com   DOI: 10.3389/fnins.2024.1465789   Portland State University

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

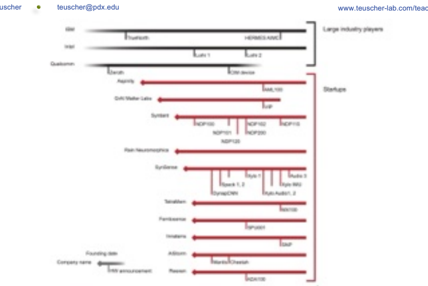


Fig. 3 | Timeline of recent commercial NM compute firms and new hardware.

teuscher Lab teuscher-lab.com   https://doi.org/10.1038/s41467-025-57352-1   Portland State University

Christof Teuscher • teuscher@pdx.edu   www.teuscher-lab.com/teaching

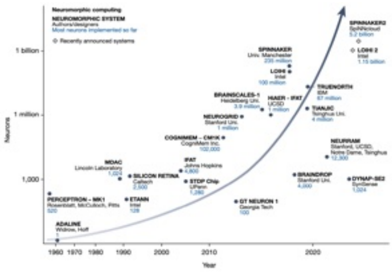
Review

Neuromorphic computing at scale

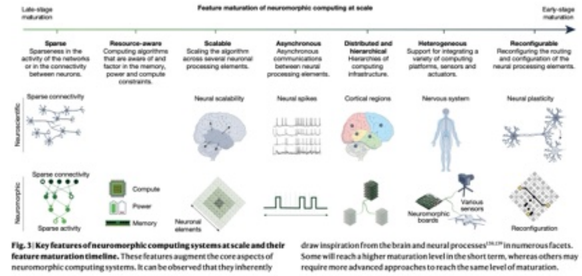
AlexNet-like moment for neuromorphic computing

teuscher Lab teuscher-lab.com   https://doi.org/10.1038/s41586-024-08253-8   Portland State University

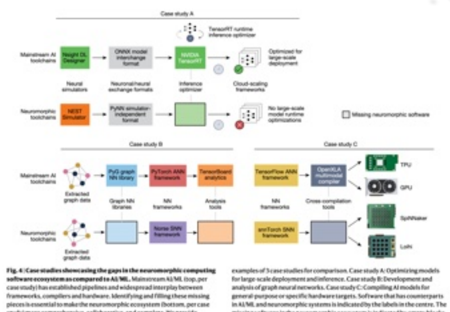
Progression of neuromorphic computing systems



Feature maturation of neuromorphic computing

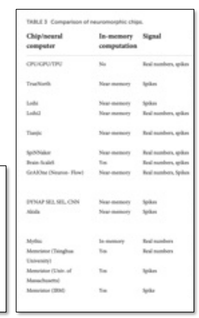


Gaps in neuromorphic computing



Are spikes the future for LLMs?

- SNN can't be trained with backpropagation. STDP is used instead.
  - Local learning only
  - Credit assignment problem
  - Lack of scalability and efficiency
  - Lack of global optimization objective
  - No mathematical convergence guarantee
- Transformers are inherently un-brain–like.
- The attention mechanism requires matrix-matrix multiplication. Hard to implement in neuromorphic architectures.



Recap: neuromorphic chip design choices

- Spikes vs analog
- Stochastic vs reduced precision vs full precision
- In-memory vs near-memory vs von Neumann
- NoC vs shared bus vs wires
- Inference and/or learning
- Crossbar vs non-crossbar
- Neuron circuitry vs configurable neuron vs programmable neuron
- STDP vs backprop vs other learning rules
- Embedded vs datacenter
- Number of neurons vs complexity of neuron and/or learning rate
- Amount of memory

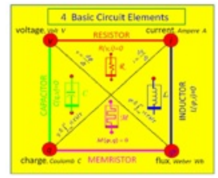
What other options do we have to accelerate AI/ML algorithms?

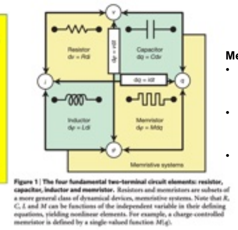
Mem-elements

- Compute closer to the physics.
- Compute with the physics.

Mem-elements

- 1971: Chua predicted missing circuit elements



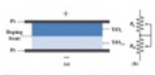


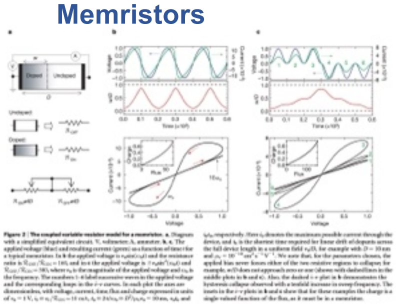
Memristor:
- Variable resistance that depends on the history of current that has flowed through the device.
- Ability to "remember" its previous resistance state even when power is turned off.
- Non-volatile memory properties (retains information without power)

HP: TiO_2 devices

https://doi.org/10.1038/nature06932

Memristors

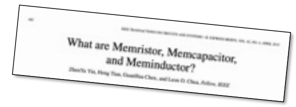




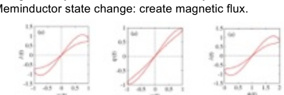
https://doi.org/10.1038/nature06932

Memcapacitor and meminductor

Feature	Memristor	Memcapacitor	Meminductor
Basic Definition	Memory resistor	Memory capacitor	Memory inductor
Conventional Counterpart	Resistor	Capacitor	Inductor
Controls/Responds To	Current history	Voltage history	Flux history
State Variable	Resistance (memristance)	Capacitance (memcapacitance)	Inductance (meminductance)
Fundamental Relationship	Links charge (q) and magnetic flux (φ)	Links charge (q) and voltage (v)	Links current (i) and magnetic flux (φ)
Energy Storage	Minimal/none	Electric field	Magnetic field
Primary Physical Mechanisms	Ion migration, filament formation, phase changes	Polarization changes, ionic redistribution	Magnetic domain reorientation
State Change Method	Apply current	Apply voltage	Apply magnetic flux
First Theoretical Proposal	1971 (Leon Chua)	2009 (Di Ventra, Pershin, Chua)	2009 (Di Ventra, Pershin, Chua)
First Physical Implementation	2008 (HP Labs)	Limited experimental demonstrations	Limited experimental demonstrations
Development Status	Multiple working prototypes	Mostly theoretical, some prototypes	Mostly theoretical, fewer prototypes
Key Applications	Non-volatile memory, neuromorphic computing	Signal processing, tunable circuits	Power systems, signal filters
Characteristic Property	Non-linear resistance with memory	Non-linear capacitance with memory	Non-linear inductance with memory



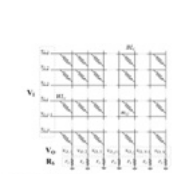
- Memristor state change: apply an electric current through the device.
- Memcapacitors store energy in electric fields and respond to voltage history rather than current history.
- Meminductor state change: create magnetic flux.

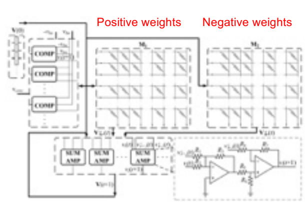


DOI: 10.1109/TCSII.2014.2387653

Memristors for neuromorphic chips

What to do with negative weights?

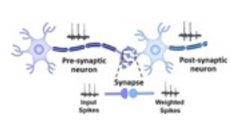


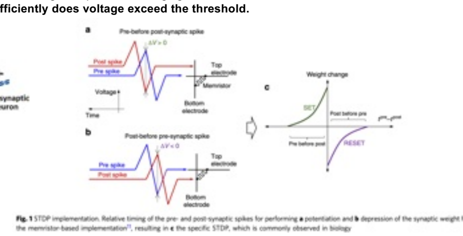


DOI: 10.1109/TNNLS.2013.2296777

Memristors can implement STDP

- STDP emerges naturally from device physics without complex circuits.
- STDP: The time difference between spikes determines magnitude of change (closer timing = stronger effect)
- STDP refines Hebbian learning by adding precise timing relationships. Unsupervised learning!
- Memristors typically have threshold voltages required for changing state.
- **Only when spikes overlap sufficiently does voltage exceed the threshold.**





https://doi.org/10.1038/s41467-018-07757-y

VMM using memristive crossbars

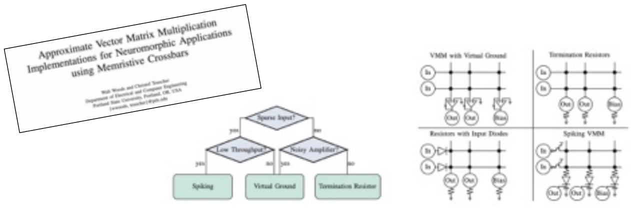


Fig. 1: Decision tree for application-ideal VMM based on the results of this work; architectures are shown in Fig. 3. With 1% input activity, spiking architectures consumed 99.9% less power than a virtual ground architecture, but were several thousand times slower. Without a noisy amplifier, termination resistor-based architectures consumed 75% less power than a virtual ground approach and ran just as fast.

Fig. 3: VMM architectures surveyed. All crossbar junctions are memristive devices. A full implementation of these VMMs also requires a differential amplifier between each output measurement and the bias measurement, with an architecture-specific gain. A relation between each architecture and its ideal application is shown in Fig. 1.

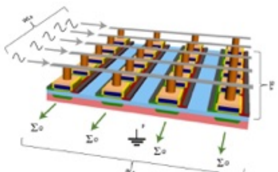
DOI: 10.1109/NANOARCH.2017.8053729

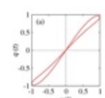
What is one problem of memristors?

https://www.nature.com/articles/s41928-021-00649-y

Memcapacitors

- Memcapacitors offer several unique benefits, but one of their most significant advantages is their exceptional energy efficiency in dynamic computing applications.
- Memcapacitors can implement STDP too (through charge).





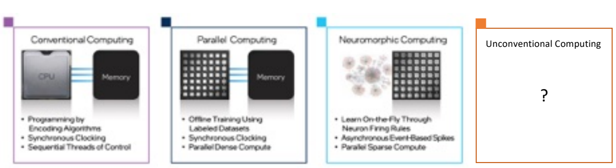
https://doi.org/10.1038/s41928-021-00649-y

Issues with mem-devices

- Device-to-device variation
- Cycle-to-cycle variation
- Limited write cycles (10^6-10^9 compared to CMOS >10^{15})
- Retention issues / state drift
- Programming complexity (requires precise pulses)
- Limited switching speeds
- Lack of design tools

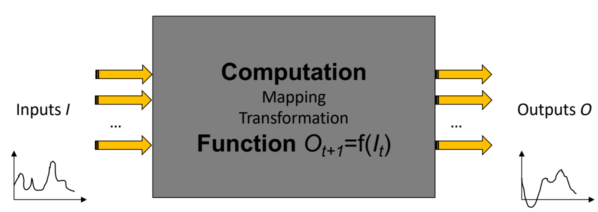
https://doi.org/10.1038/s41928-021-00649-y

Unconventional computing



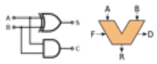
Technology Brief | Taking Neuromorphic Computing to the Next Level with Loihi 2

What is computation?



How can we "realize" this mapping?

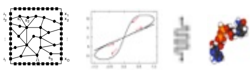
Intrinsic vs designed computation



Desired computation — Computing architecture

Intrinsic dynamics — Device network

How to obtain a desired computation?



What is *Unconventional Computing* (UCOMP)?

There is no precise definition…

"[…] a computing scheme that today viewed as unconventional may well be so because its time hasn't come yet – or is already gone."
— Tommaso Toffoli

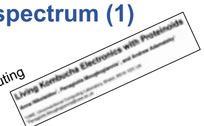
"Today, UCOMP is broad but (relatively) shallow, whilst CCOMP is narrow, but incredibly deep. What would computation look like if UCOMP were as deep as CCOMP, and there were an integrated theory combining all its aspects?" — Susan Stepney

Intel Unveils Oversized Novelty Processor

Today's unconventional computing spectrum (1)

Most definitely incomplete…

Multi-core
In-memory computing
Quantum computing
Stochastic computing
Ising machines
Mem-element computing
Chaos computing
Membrane computing



More classical ⟷ More unconventional

GPU
FPGA
Neuromorphic
Reservoir computing
Analog computing
Thermodynamic computing
Fluidics
Cellular automata
DNA computing
Hypercomputation
Bacterial computation
"Fringe"

What should I "invest" in to reboot computing?

Materials – Devices & Circuits – Architectures – Software – Algorithms – Applications

Today's unconventional computing topics spectrum (2)

Aims and Scope

The International Journal of Unconventional Computing offers the opportunity for rapid publication of theoretical and experimental results in non-classical computing. Specific topics include but are not limited to:

- physics of computation (e.g. conservative logic, thermodynamics of computation, reversible computing, quantum computing, collision-based computing with solitons, optical logic)
- chemical computing (e.g. implementation of logical functions in chemical systems, image processing and pattern recognition in reaction-diffusion chemical systems and networks of chemical reactors)
- bio-molecular computing (e.g. conformation based, information processing in molecular arrays, molecular memory)
- cellular automata as models of massively parallel computing
- complexity (e.g. computational complexity of non-standard computer architectures; theory of amorphous computing; artificial chemistry)
- logics of unconventional computing (e.g. logical systems derived from space-time behavior of natural systems; non-classical logics; logical reasoning in physical, chemical and biological systems)
- smart actuators (e.g. molecular machines incorporating information processing, intelligent arrays of actuators)
- novel hardware systems (e.g. cellular automata VLSIs, functional neural chips)
- mechanical computing (e.g. micromechanical encryption, computing in nanomachines, physical limits to mechanical computation).

Aims & Scope

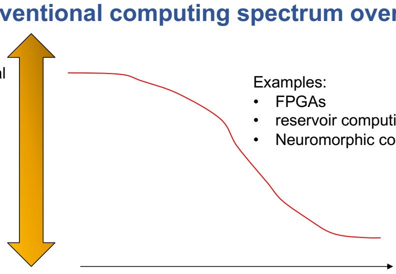
npj Unconventional Computing considers all aspects of research applying unconventional approaches to improve computing efficiency and performance.

The journal covers a broad range of topics including but not limited to:

- Brain-inspired algorithms
- Quantum computing and quantum-inspired algorithms
- Molecular, and DNA computing
- Membrane computing and P systems
- Neuromorphic computing and brain-inspired architectures
- Amorphous computing and swarm robotics
- Cellular automata and complex systems
- Artificial life and evolutionary computation
- Chaos computing and dynamical systems
- Reversible and adiabatic computing
- Thermodynamic computing
- Probabilistic computing (p-bit computing)
- Superconducting computing
- Application of physical phenomena to computing
- Design, fabrication, and testing of unconventional hardware, physical substrates, and emerging technologies for computing

npj | unconventional computing

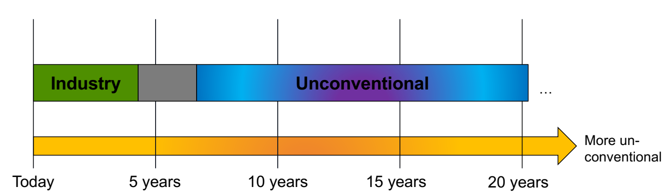
The unconventional computing spectrum over time



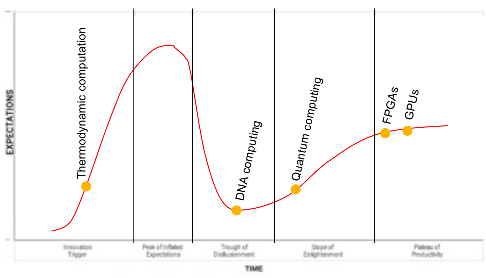
Examples:

- FPGAs
- reservoir computing
- Neuromorphic computing

The unconventional computing time horizon



Hype or computation?



Gartner's hype cycle

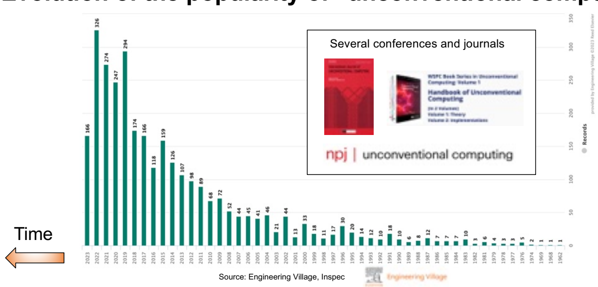
UCOMP supporters

- New technologies need time (and money) to mature.
- Current comparison with state-of-the-art are not fair comparisons.
- We need to start somewhere.
- Often on shoestring budgets only.

UCOMP critics

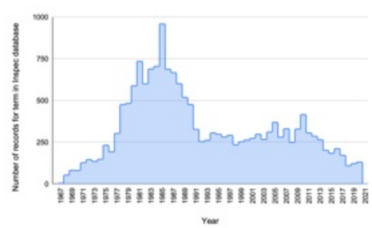
- No useful paradigms that can compete with conventional approaches.
- The real, practical challenges to be solved are always kept comfortably far away.
- Proving that we can compute a NAND function with a self-assembled nanowire network, for example, doesn't move the needle.

Evolution of the popularity of "unconventional computing"



Source: Engineering Village, Inspec

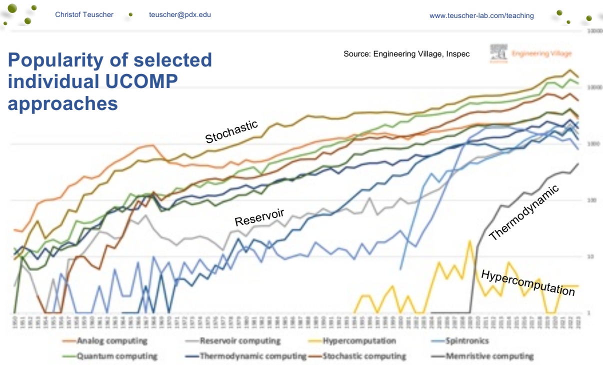
The decline of traditional computer architecture



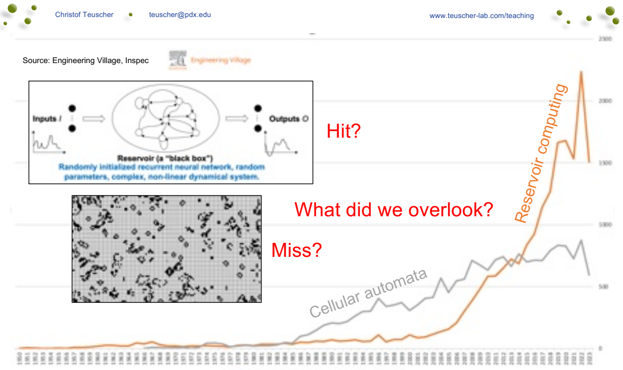
The number of records over the years that were found for the "**computer architecture**" thesaurus term. The field had its peak around 1985 and has since been in decline. Data source: Elsevier Engineering Village thesaurus.

Popularity of selected individual UCOMP approaches

Source: Engineering Village, Inspec

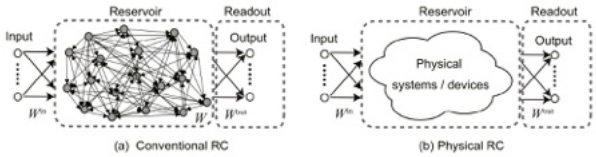


Source: Engineering Village, Inspec



What is material and Physical Reservoir Computing (PRC)?

A physical reservoir computer is a system in which the reservoir is realized with a physical system.



Source: Tanaka et al., Recent advances in physical reservoir computing: A review, 2019, https://doi.org/10.1016/j.neunet.2019.03.005

Questions of interest for Physical Reservoir Computing (PRC)

- How much control does on have over the fabrication process?
- Reproducibility of input-output mapping?
- How to interface with a material/device?
- Memory capacity?
- Signal attenuation/amplification?
- Cycle-to-cycle variation?
- Device-to-device variation?
- In-situ or ex-situ training? The challenging part tends to be the *in materio* implementation of the readout and training.
- What pre- and post-processing is necessary? What will it "cost?"
- …

Pattern recognition in a bucket (1)

4 electric motors perturb water surface

320x240 pixels, 5 frames webcam films inference patters

Sobel edge detection mask that was suitably thresholded to remove noise and averaged to produce a 32x24 grid. The cells of this matrix were used as the input values for 50 perceptrons in parallel

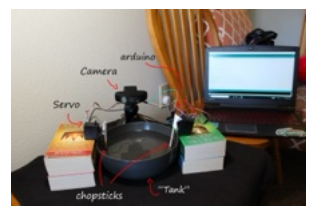
Overhead projector

Fig. 1. The Liquid Brain.

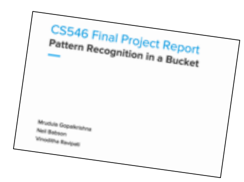
- XOR task
- Speech recognition ("zero"/"one" task)
  - 25% mistakes with single perceptron
  - 1.5% mistakes when passed through water.

DOI: 10.1007/978-3-540-39432-7_63

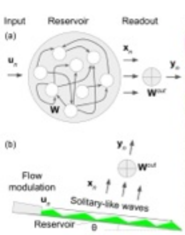
Pattern recognition in a bucket (2)



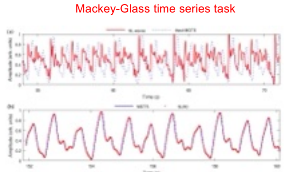
Fernando and Sajakka experiment replication.



RC based on solitary-like waves dynamics of film flows



Mackey-Glass time series task

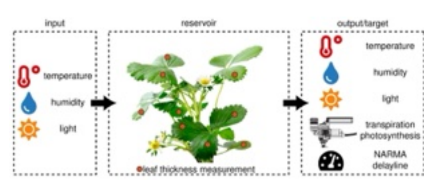


An electric signal modulates the pump flow: '1' states are implemented as a brief change in the flow rate that results in the generation of one.

Camera used for outputs. Output layer in software.

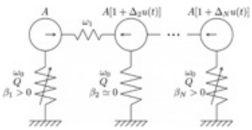
Maksymov & Pototsky, Reservoir computing based on solitary-like waves dynamics of film flows: a proof of concept, 2023, https://arxiv.org/abs/2303.01801

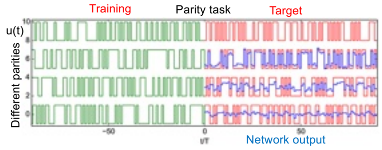
Plant reservoir computing

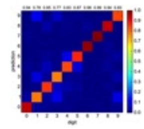


Computing with networks of nonlinear mechanical oscillators

Anharmonic = not simply harmonic
Springs with damping. Random couplings.





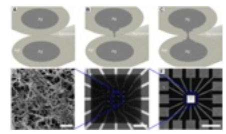


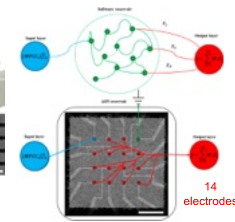
N= 400 oscillators
Parity + spoken words tasks
Numerical simulations only.

Coulombe, Computing with networks of nonlinear mechanical oscillators, 2017, https://doi.org/10.1371/journal.pone.0178663

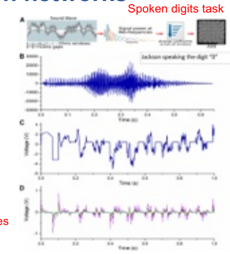
Reservoirs with atomic switch networks

Silver iodide (AgI) junctions





Spoken digits task



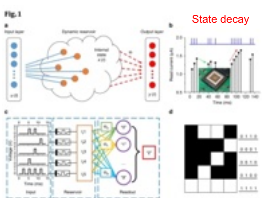
Self-organizing systems

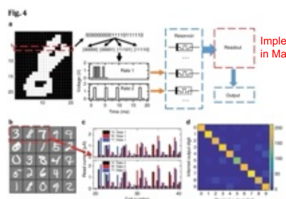
14 electrodes

Lilak, Woods, Scharnhorst K, Dunham, Teuscher, Stieg, Gimzewski, Spoken Digit Classification by In-Materio Reservoir Computing With Neuromorphic Atomic Switch Networks, 2021, http://doi.org/10.3389/fnano.2021.675792

Dynamic memristors for temporal information processing

Experimental demonstration of a memristor-based RC system using dynamic memristor devices that offer internal, short-term memory effects. 88 memristors.

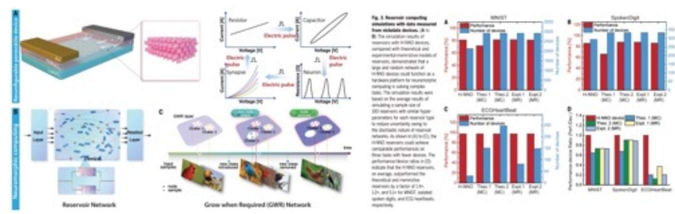




Source: Du et al. Reservoir computing using dynamic memristors for temporal information processing. Nat Commun 8, 2204 (2017). https://doi.org/10.1038/s41467-017-02337-y

Reconfigurable perovskite nickelate electronics

Perovskite nickelate devices act as memristors and memcapacitors for RC.



Source: Zhang et al., Reconfigurable perovskite nickelate electronics for artificial intelligence, 2022, https://doi.org/10.1126/science.abj7943

First Josephson Junction (JJ) reservoir (1)

Josephson transmission line (superconducting oscillators) >100GHz
Inputs to selected JJs only

Wave-like reservoir response

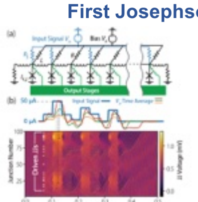


FIG. 1. (a) The superconducting reservoir formed by a Josephson transmission line with the input signal distributed to some fraction f of the junctions across individual resistances R_i. The junctions are connected to decimating output stages. (b) The reservoir response to the displayed 25-ps sample-and-hold input waveform, showing the voltages across each JJ in a chain of 100. Only twenty percent of JJs (white dots) are driven with the input signal shown above. The time-averaged voltages $V_{JJ}(t)$ are shown, for a few indicated junctions, superimposed on the input signal.

Toggle FF
Signal reconstruction

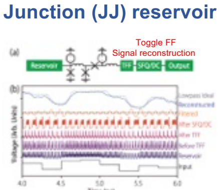
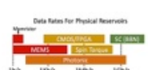


FIG. 2. (a) Detailed view of the output stage, which consists of a buffered output junction, a toggle flip-flop, and an SFQ-to-DC converter. (b) The signal as it evolves, from bottom to top, while passing through the output stage in response to the reservoir stimulus shown at the bottom. The threshold for signal reconstruction is shown as the dotted line at the mean of the "filtered" response and the resulting reconstruction is compared to the "ideal" time-averaged reservoir response.



https://www.youtube.com/watch?v=1bWjyQ1326g

The outputs are the oscillation rates of some subset of the junctions in the reservoir. This is more in line with the liquid state machine approach.

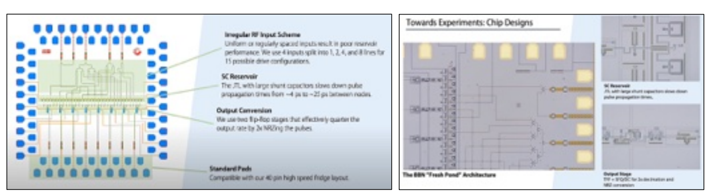
In **simulations** we treat two cases with similar performance:

1. **Full rate reservoir**, assuming on-board readout logic. Sample and hold times as short as 15 ps.
2. **Slower reservoir** with additional capacitance and realistic microwave chain for room temperature control and readout. Sample and hold times 250 ps.

Task: speech recognition. ~70% accuracy. 5 JJs.

Rowlands et al., Reservoir Computing with Superconducting Electronics, 2021, https://arxiv.org/abs/2103.02522

First Josephson junction reservoir (2)



Actual JJ chip

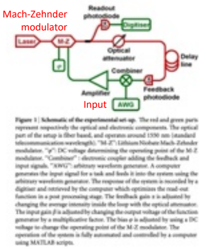
Screenshots from https://www.youtube.com/watch?v=1bWjyQ1326g

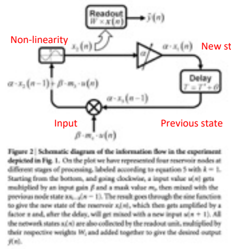
Article: Rowlands et al., Reservoir Computing with Superconducting Electronics, 2021, https://arxiv.org/abs/2103.02522

Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

Optoelectronic RC (general idea)

Photonic reservoirs are all about high speed. The challenge is to make them small.





Paquot et al., 2012, http://www.doi.org/10.1038/srep00287

teuscher Lab teuscher-lab.com Portland State University

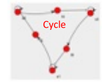
Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

DNA oscillators as reservoirs

$$\frac{d[P_1]}{dt} = k\beta[S_1]([G_1]-[P_3]) - \delta[P_1], \quad \frac{d[S_1]}{dt} = \frac{s}{V} - k\beta[S_1]([G_1]-[P_3]) - \delta[S_1],$$
$$\frac{d[P_2]}{dt} = k\beta[S_2]([G_2]-[P_1]) - \delta[P_2], \quad \frac{d[S_2]}{dt} = \frac{s}{V} - k\beta[S_2]([G_2]-[P_1]) - \delta[S_2],$$
$$\frac{d[P_3]}{dt} = k\beta[S_3]([G_3]-[P_2]) - \delta[P_3], \quad \frac{d[S_3]}{dt} = \frac{s}{V} - k\beta[S_3]([G_3]-[P_2]) - \delta[S_3].$$

Goudarzi, Lakin, and Stefanovic, DNA reservoir computing: a novel molecular computing approach, 2013, https://doi.org/10.1007/978-3-319-01928-4_6

- Yahiro et al., Alife, 2018, https://doi.org/10.1162/isal_a_00013
- Used same oscillator as Goudarzi et al.
- Simulation only. Similar performance as Goudarzi.



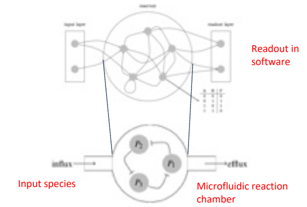


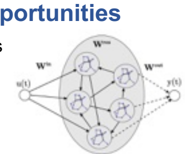
Fig. 4: Three products form an inhibitory cycle that leads to oscillatory behavior in the reservoir. Each product P_i inhibits the production of P_{i+1} by the corresponding deoxyribozyme (cf. Equation 4).

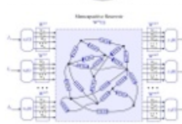
teuscher Lab teuscher-lab.com Portland State University

Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

Open problems, challenges, and opportunities

- Scale up PRC systems to **massive** numbers of neurons
- Train **modular/hierarchical** reservoirs efficiently.
- **Multitasking** reservoirs.
- **Lifelong/incremental** learning with growing reservoirs (developmental systems).
- Implement readout/learning directly *in materio*.
- Find **sweet spots** for high-speed, low-power, high performance, and high robustness.
- Reservoirs that adapt, are tuned, or reconfigured (*in-situ*).
- Material reservoir (deep) **co-evolution**.





teuscher Lab teuscher-lab.com Portland State University

Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

Quo vadis?

- **Material computing** has attracted increasing attention: it allows for solving computational problems for which traditional, general-purpose architectures are not efficient.
- PRC has great potential for physically "embedding" computing in a substrate by using the substrate itself to perform the computation.
- A broad(er) success of PRC will be contingent on finding solutions to the open questions and challenges: identify physical substrates, scaling up hardware, scaling up programmability → **(deep) co-design**.

teuscher Lab teuscher-lab.com Portland State University

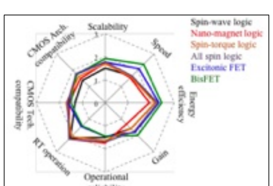
Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

What are good metrics to evaluate UCOMP approaches?

1. Gain
2. Signal to noise ratio
3. Scalability

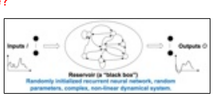
S. Borkar, "Electronics beyond nano-scale CMOS," in Proceedings of the Design Automation Conference, New York, NY, 2006, pp. 807–808.

"[...] there is nothing on the horizon that has promise to replace CMOS at least in the next ten to fifteen years." Is that still true?



(c) Non-charge-based beyond-CMOS devices

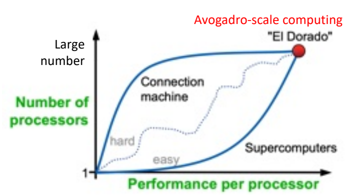
IRDS 2022, BC

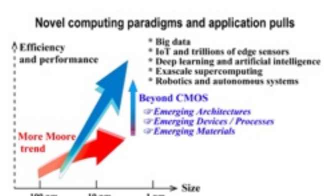


teuscher Lab teuscher-lab.com Portland State University

Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

The goals and drivers have changed and keep changing





IDRD, 2022, BC

teuscher Lab teuscher-lab.com Portland State University

Deep co-design: our best path forward?

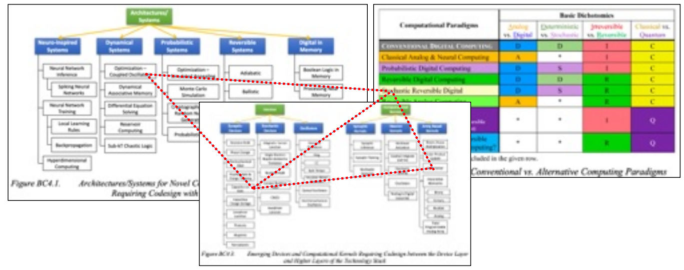
Given a proposed unconventional computing substrate, we can ask:
- Does it actually compute?
- If so, how well does it compute?
- Can it be made to compute better?

Given a proposed unconventional computational model, we can ask:
- How powerful is the model?
- Can it be implemented in a substrate?
- How faithfully or efficiently can it be implemented?

Given complete freedom in the choice of model and substrate, we can ask:
- Can we co-design a model and substrate to work well together?

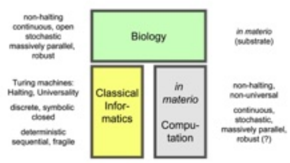
Stepney, S., Co-designing the computational model and the computing substrate. In Proceedings of the 18th International Unconventional Computation and Natural Computation Conference, pp 5-14, 2019.

Deep co-design across the entire stack



Deep co-design: what we need to make this happen…

- A formal framework for *in-materio* computing:
  - continuous
  - open
  - stochastic
  - non-equilibrium
  - non-linear
  - …
- Revisit the stack (or get rid of it?)
- More comprehensive design space exploration at all levels.



Stepney, 2018

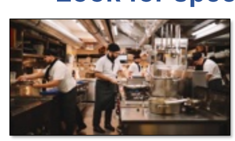
Computing Paradigm	Key Promises	Major Drawbacks
Quantum Computing	• Exponential speedup for specific problems (cryptography, optimization, simulation) • Superior performance in drug discovery and materials science • Potential to solve previously intractable problems • Parallel processing of quantum states	• Extreme environmental requirements (near absolute zero) • High error rates requiring complex error correction • Limited to specific problem types • Extremely expensive infrastructure • Decoherence issues limiting computation time
Neuromorphic Computing	• Ultra-low power consumption • Real-time processing capabilities • Excellent for pattern recognition and AI tasks • Inherent fault tolerance • Adaptive learning capabilities	• Difficult to program with traditional methods • Limited general-purpose computing ability • Lack of standardized architectures • Manufacturing complexity • Limited software ecosystem
Photonic Computing	• Speed of light data transmission • Minimal heat generation • High bandwidth capabilities • Low power consumption for data movement • Immune to electromagnetic interference	• Difficulty in creating optical transistors • Large physical footprint currently • Integration challenges with electronic systems • Limited ability for data storage • High initial development costs
DNA Computing	• Massive parallel processing potential • Extremely high data density • Energy efficient at molecular scale • Biocompatibility for medical applications • Self-assembly capabilities	• Extremely slow compared to electronic computers • High error rates in biological processes • Requires wet lab conditions • Limited to specific computational problems • Difficult to interface with traditional systems
Reversible Computing	• Theoretically zero energy dissipation • Potential for ultra-low power devices • Quantum computing compatibility • Reduced heat generation • Environmental sustainability	• Requires complete redesign of algorithms • Memory overhead for storing computation history • Currently slower than conventional computing • Limited practical implementations • Complex circuit designs
Memristive Computing	• Non-volatile memory and computation fusion • High density storage and processing • Low power consumption • Neuromorphic computing applications • Fast switching speeds	• Manufacturing challenges at scale • Reliability and endurance concerns • Limited commercial availability • Integration complexity with CMOS • Variability in device characteristics
Spintronics	• Non-volatile operation • Low power consumption • High-speed operation potential • Radiation hardness • Compatible with existing semiconductor processes	• Complex physics requiring precise control • Temperature sensitivity • Limited commercial applications currently • High research and development costs • Material engineering challenges

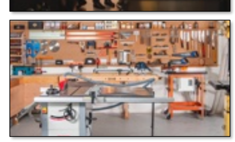
Each paradigm has unique **strengths** and faces **distinct challenges**. The most practical near-term applications are likely in specialized neuromorphic chips while quantum and DNA computing represent longer-term possibilities for specific problem domains.
The reality is that future computing systems will likely be hybrid, combining multiple paradigms to leverage their respective strengths - such as using photonic interconnects with electronic processors, or quantum-classical hybrid systems for optimization problems.

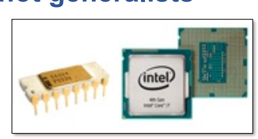
Where should we look for the next big thing?

ICRC 2023 poll

Look for specialists, not generalists







Action items for the community

- Look for specialists
- Formal frameworks
- (Unified?) evaluation metrics
- Roadmap(s)
- Timelines to solutions

Scott Holmes