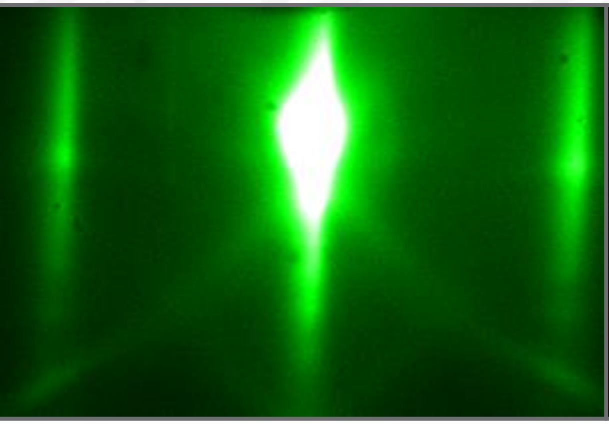
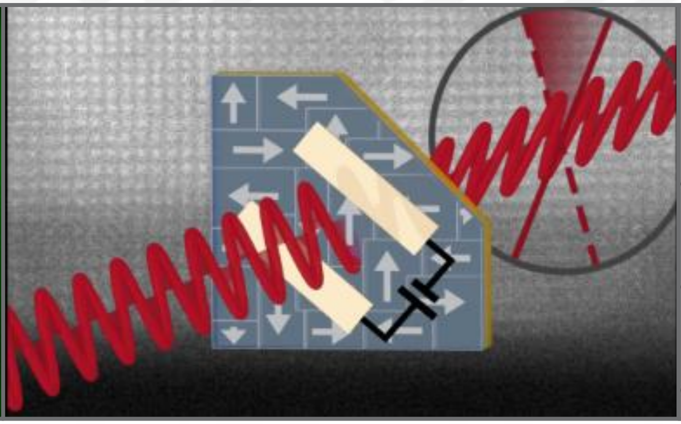
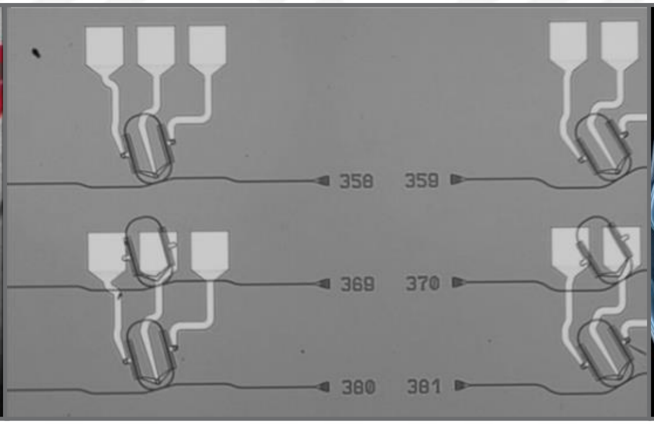
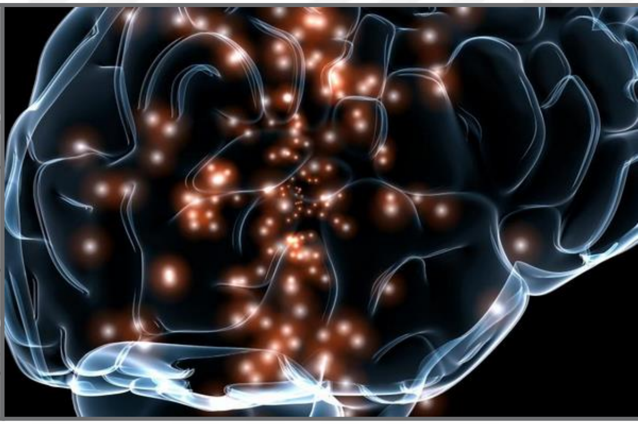
PICs for enabling neuromorphic computing

Folkert Horst, Elger Vlieg, Bert Jan Offrein



IBM Research Europe - Zurich, 8803 Rüschlikon, Switzerland

Neuromorphic Devices and Systems Group

Outline

- Neuromorphic computing
- Integrated-optic neuromorphic computing concepts
- Discussion

Experiment: “Human Brain vs. Computer”

Task 1: Mathematics

$$\sqrt{2} = ?$$

Task 2: Image recognition

Traditional silicon scaling ended
New types of problems gain interest

Explore new functionalities, More than Moore
Explore new computing paradigms

Neuromorphic Computing – ?

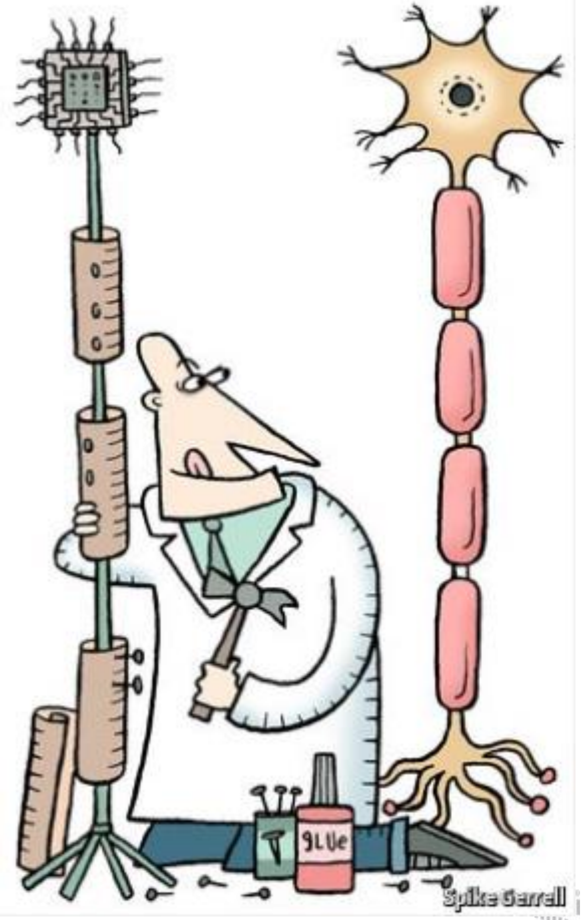
Ethymological: **“neuro”** ⇔ related to nerves or nervous system
“morphic” ⇔ having form or structure of…

Definition: Neuromorphic computing is a **brain-inspired signal processing technology** that tries to **mimic** the **neuro-biological architecture** of the **brain and its functions.**
As interdisciplinary technology, it involves

- biological,
- physical,
- mathematical,
- computer science,
- and electronic engineering concepts

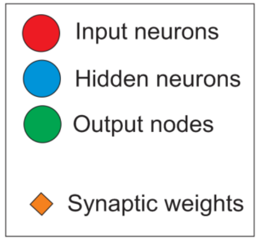
to design and realize new artificial neural network systems.

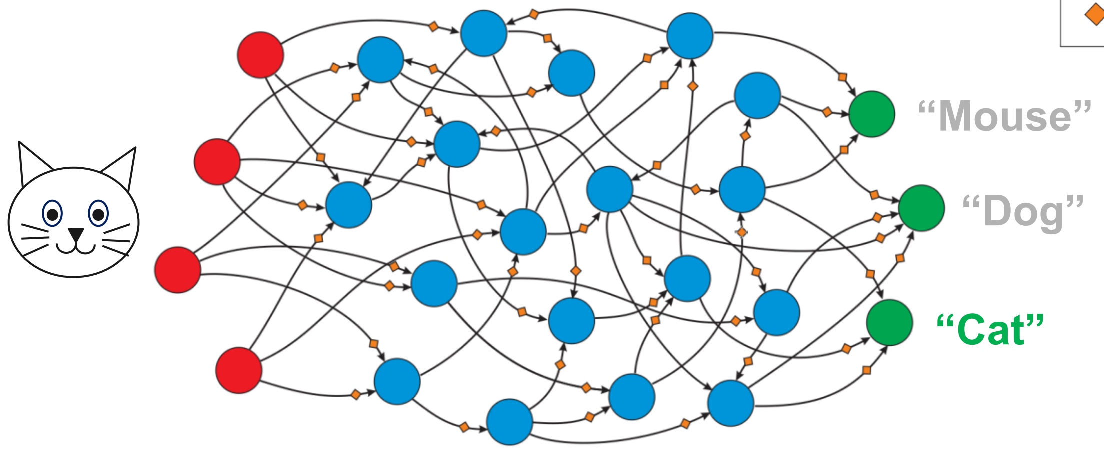
http://www.web3.lu/category/science-philosophy/



Brain inspired computing:

Brain-like Neural network:



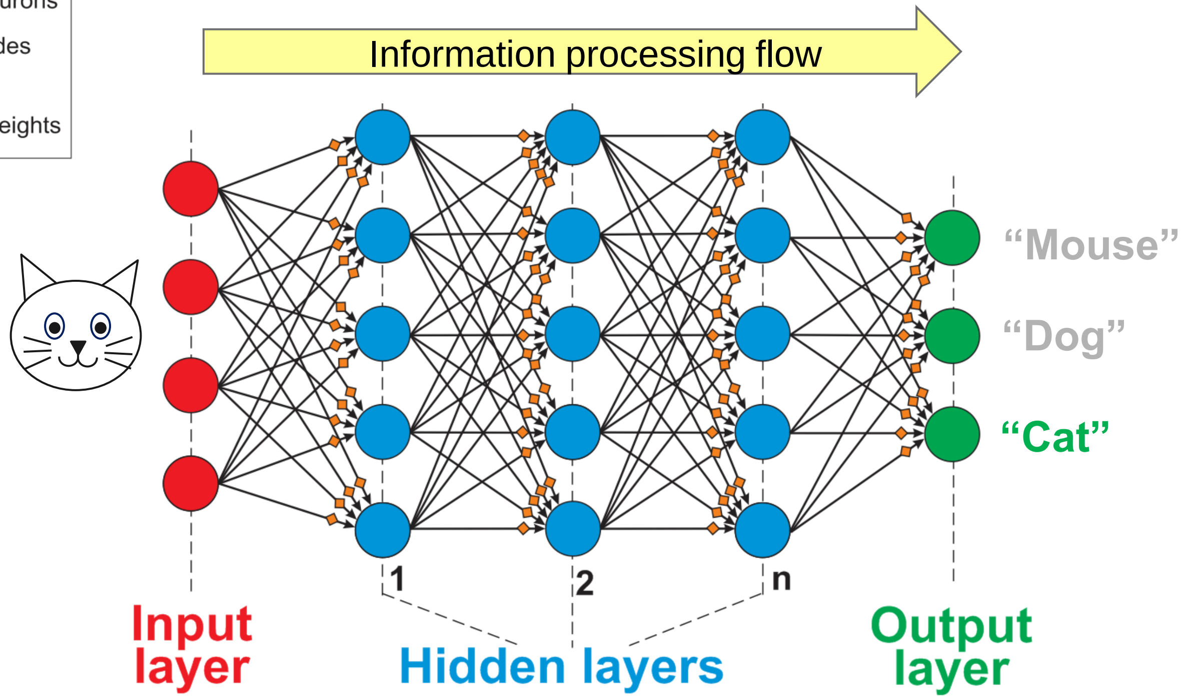


- Omni-directional signal flow
- A-synchronous pulse signals
- Information encoded in signal timing

➔ Difficult to implement efficiently on standard computer hardware

Simplify

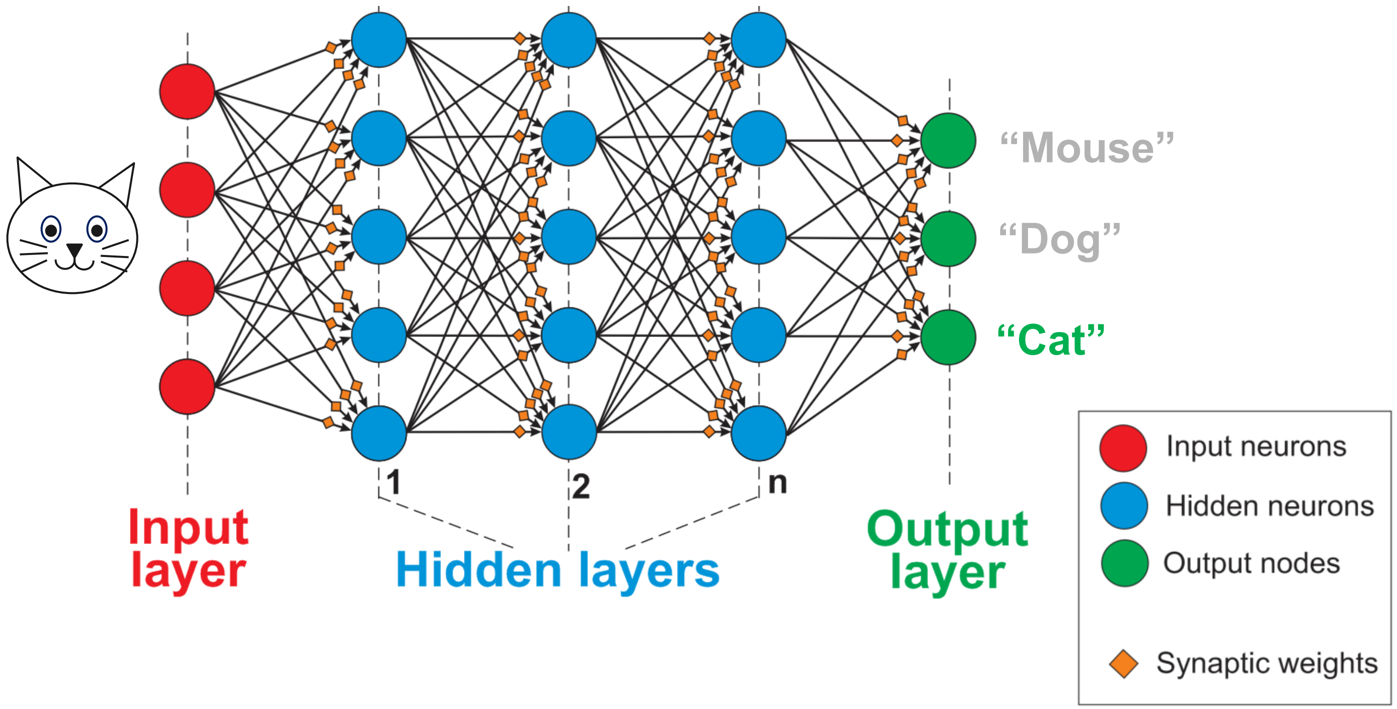
Deep Artificial Neural Network:



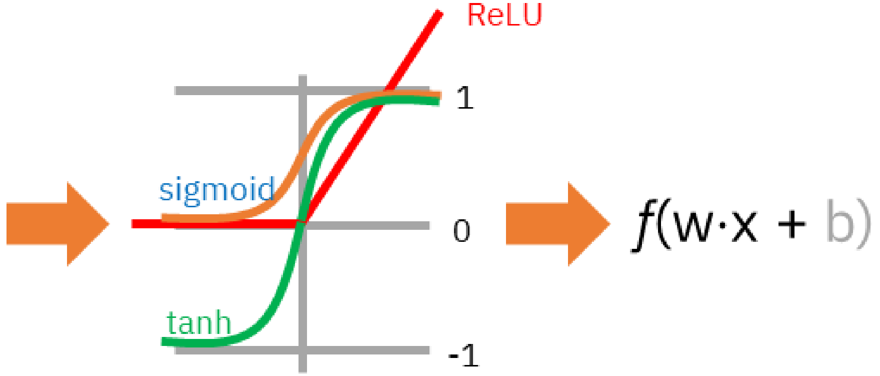
- Feed-forward sequential processing
- Information encoded in signal amplitude
- Neuron activation: Accumulate + Threshold
- **Training: Backpropagation Algorithm**

Signal processing in neuromorphic computing

Artificial deep neural network



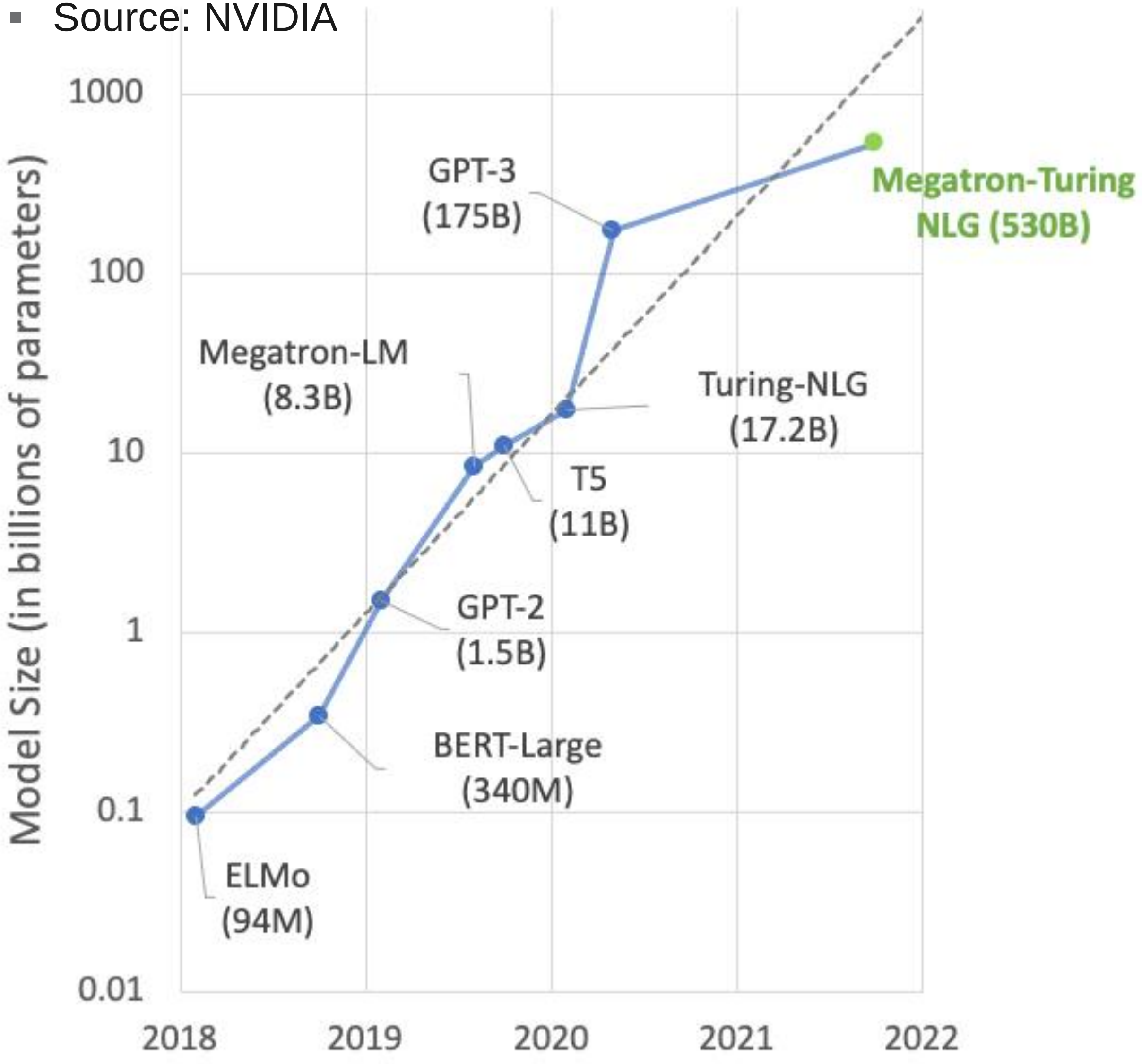
$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ & & \cdots & \\ w_{j1} & w_{j2} & \cdots & w_{jn} \\ & & \cdots & \\ w_{mn} & w_{mn} & \cdots & w_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$



Synaptic function: Multiply accumulate → Vector matrix multiplication → O(N^2)
Neuron: Nonlinear activation → O(N)

The neural network *size* explosion

- Source: NVIDIA



MIT Technology Review

A data center

DEAN MOUHTAROPOULOS | GETTY; EDITED BY MIT TECHNOLOGY REVIEW

Artificial Intelligence / Machine Learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

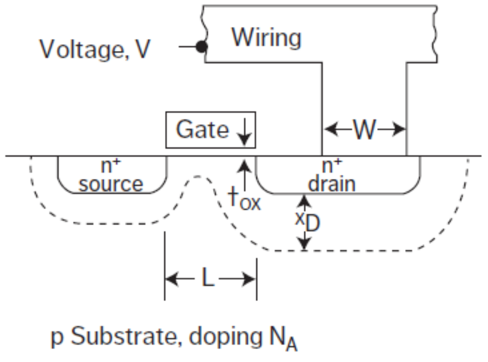
by **Karen Hao** Jun 6, 2019

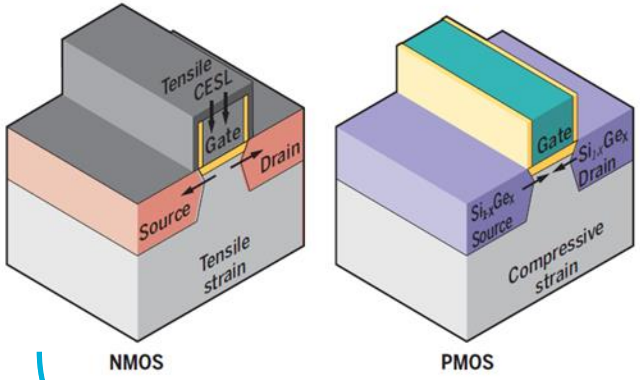
The artificial-intelligence industry is often compared to the oil industry: once mined and refined, data, like oil, can be a highly lucrative commodity. Now it seems the metaphor may extend even further. Like its fossil-fuel counterpart, the process of deep learning has an outsize environmental impact.

E. Strubell et al., arXiv:1906.02243

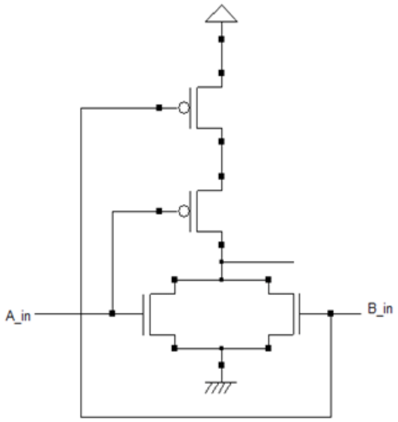
The computing hardware

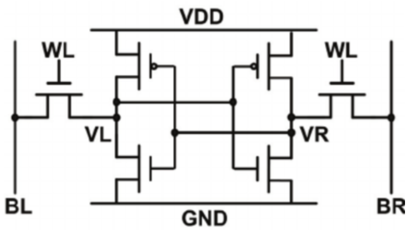
Transistors



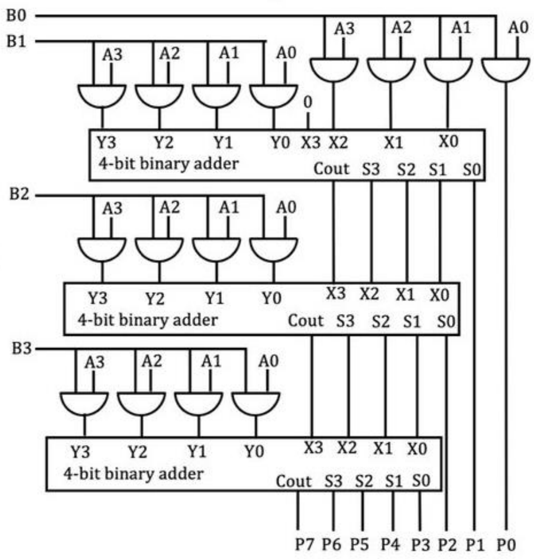


Logic & SRAM

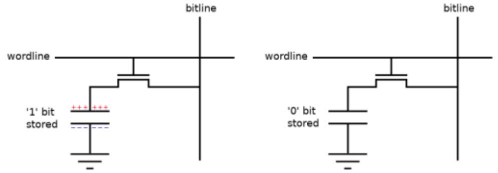


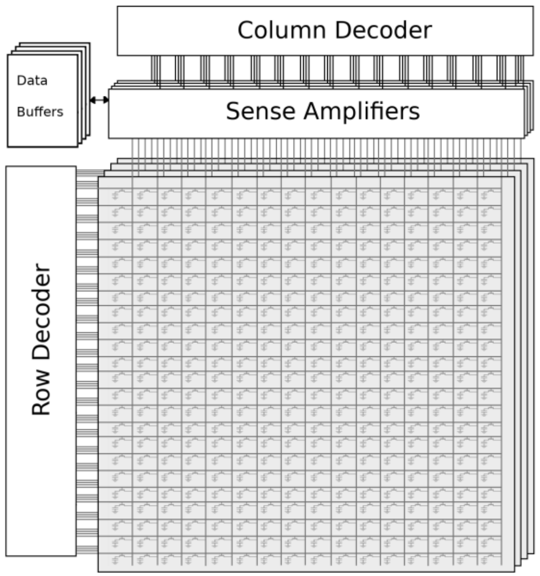


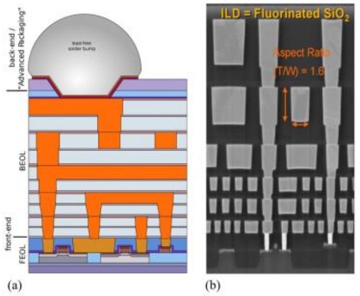
Circuits



Memory & Storage

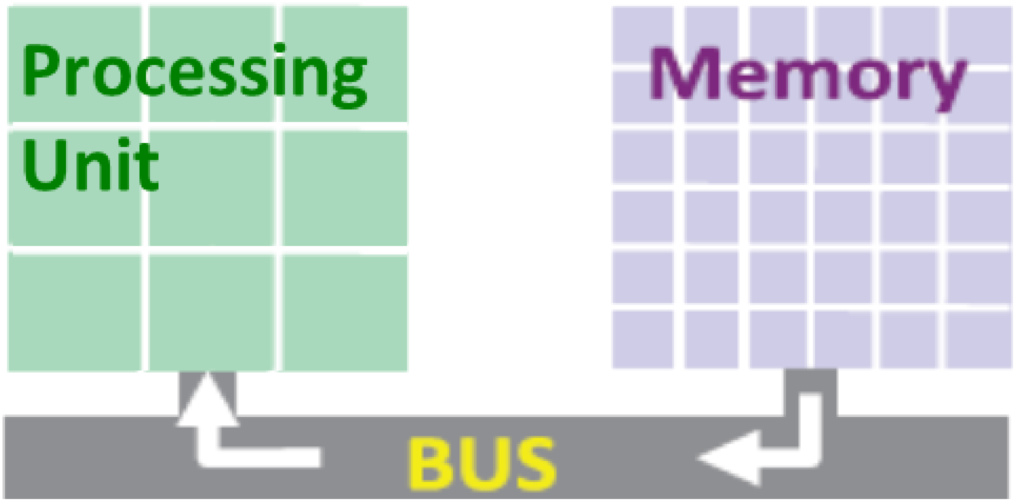




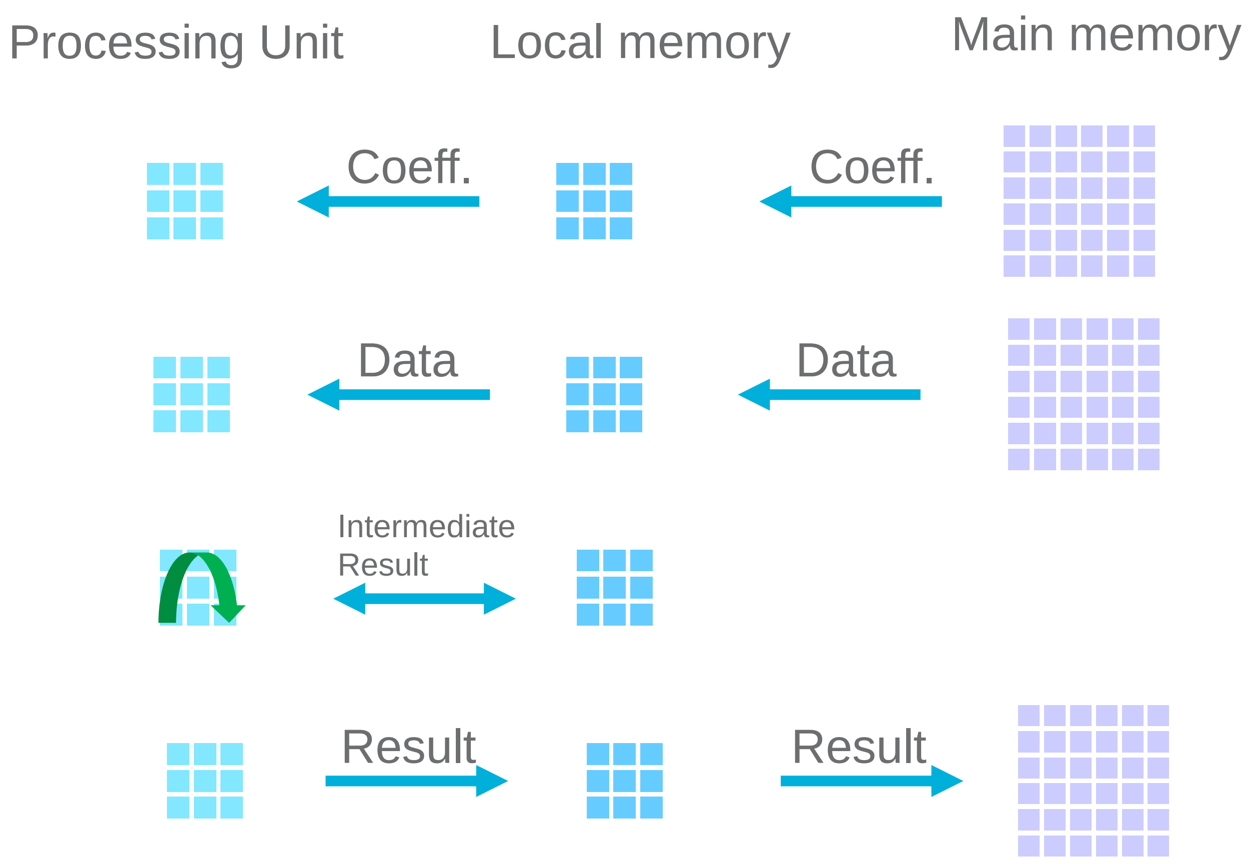


Digital signal processing

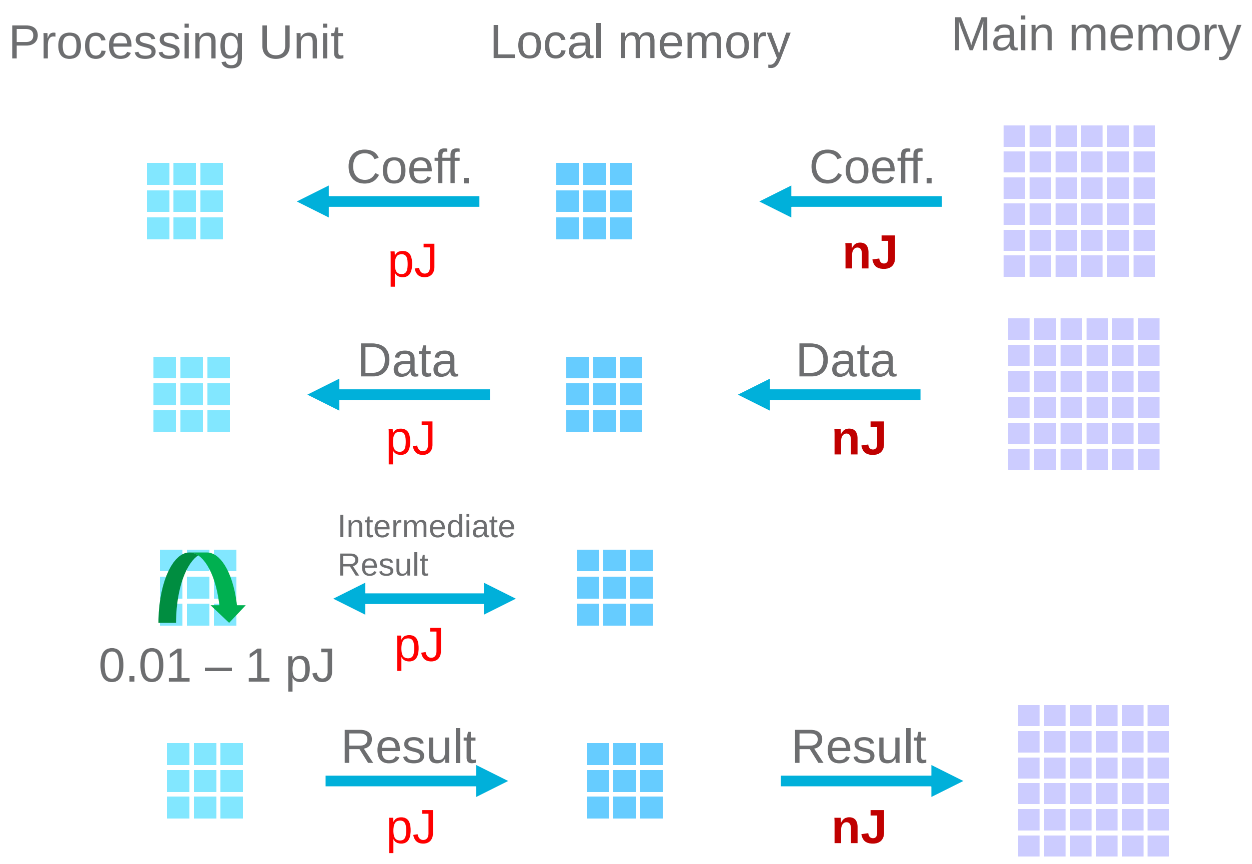
- The Von Neumann architecture
  - Memory for programs and data, a bus for memory access, an arithmetic unit & a program control unit



Let’s have a closer look at the processing steps



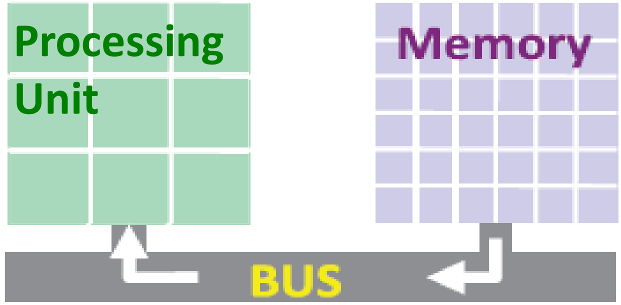
Let’s have a closer look at the processing steps



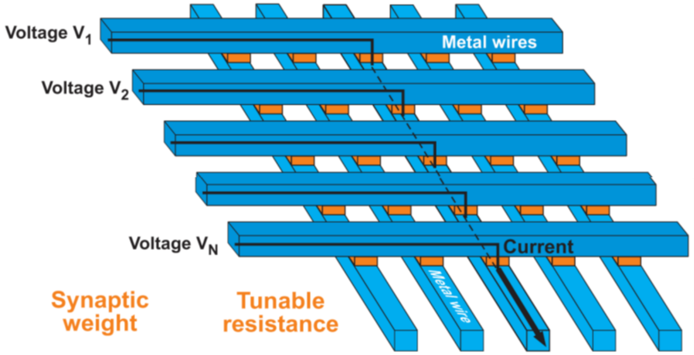
Analog signal processing for scalability

- **Limiting factors**
  - Memory access
  - Sequential operations
  - Digital signal processing

- **Overcome by**
  - In-memory computing
  - Parallel operations
  - Analog signal processing



Compute effort ~O(#Neurons^2)

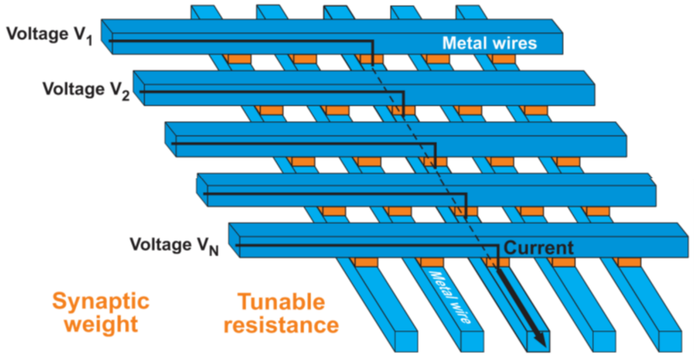


Compute effort ~O(1)

Electrical and optical solutions are viable candidates

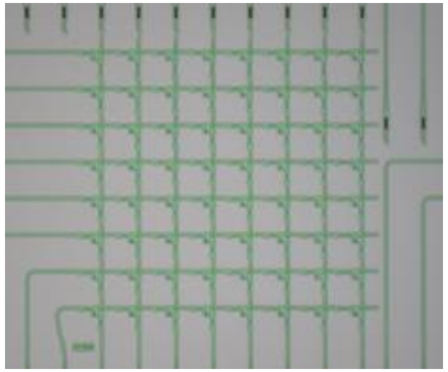
Analog signal processing systems

- Electrical

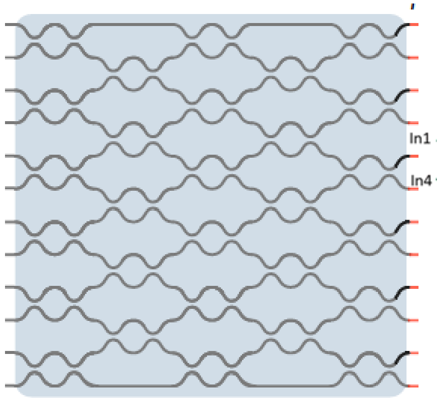


Ohm’s and Kirchhoff’s law

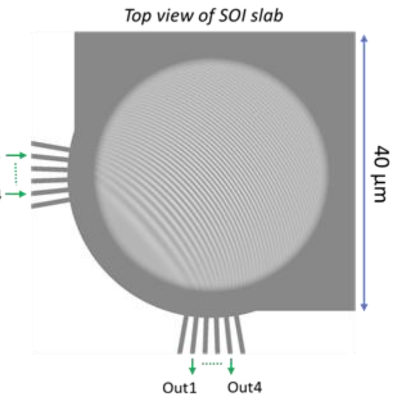
- Integrated photonics



From: EU PHOENICS (U Oxford).



From: Y. Shen et al., doi: 10.1038/nphoton.2017.93.



From: F. Horst, IBM.

Attenuation, interference, diffraction

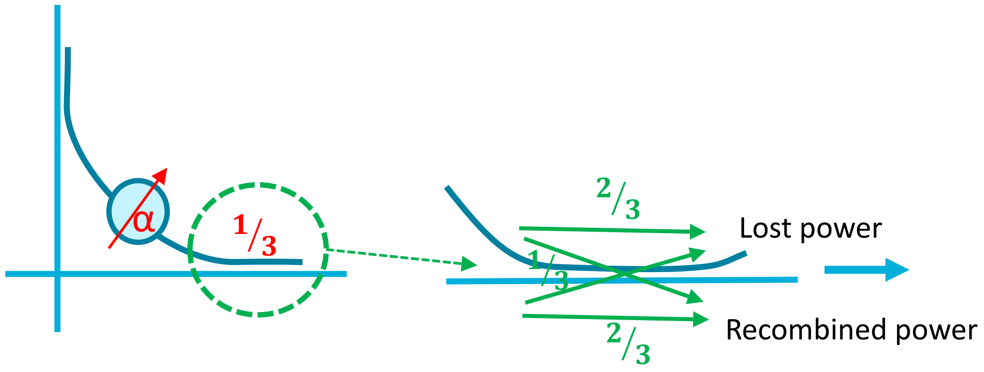
- Memristive devices in a crossbar
  - PCM
  - OxRAM
  - FERAM

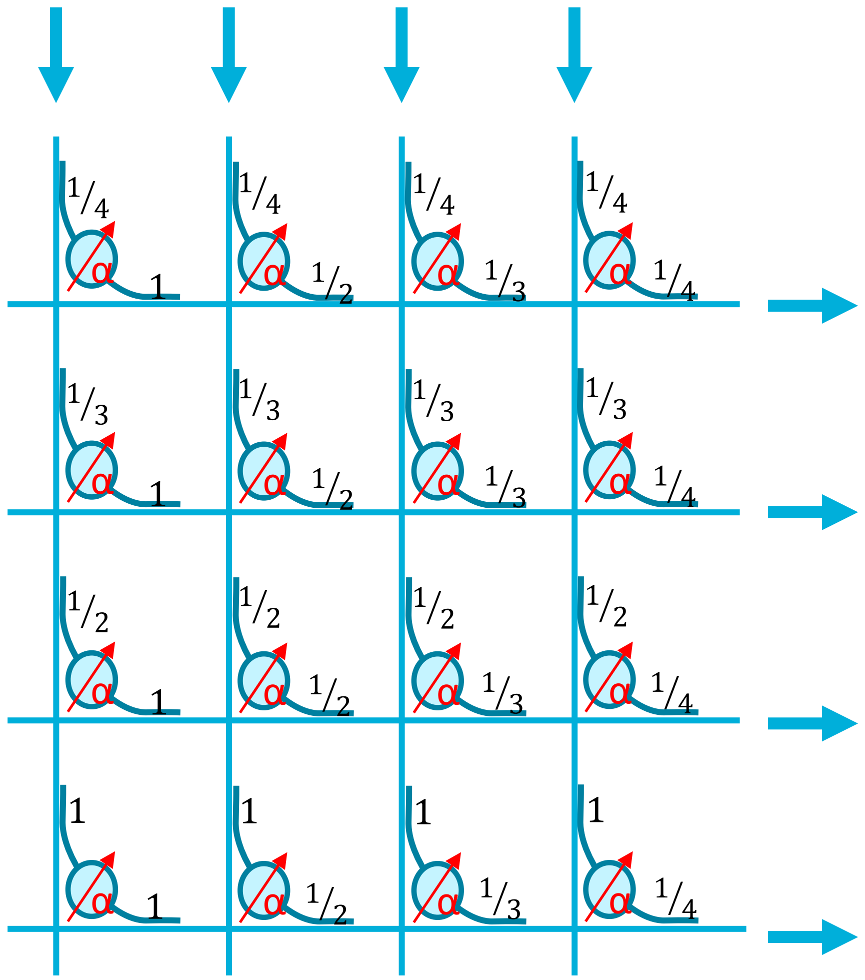
- Various device concepts and materials
  - Crossbar
  - Mach-Zehnder interferometer
  - Diffractive

For inference & training

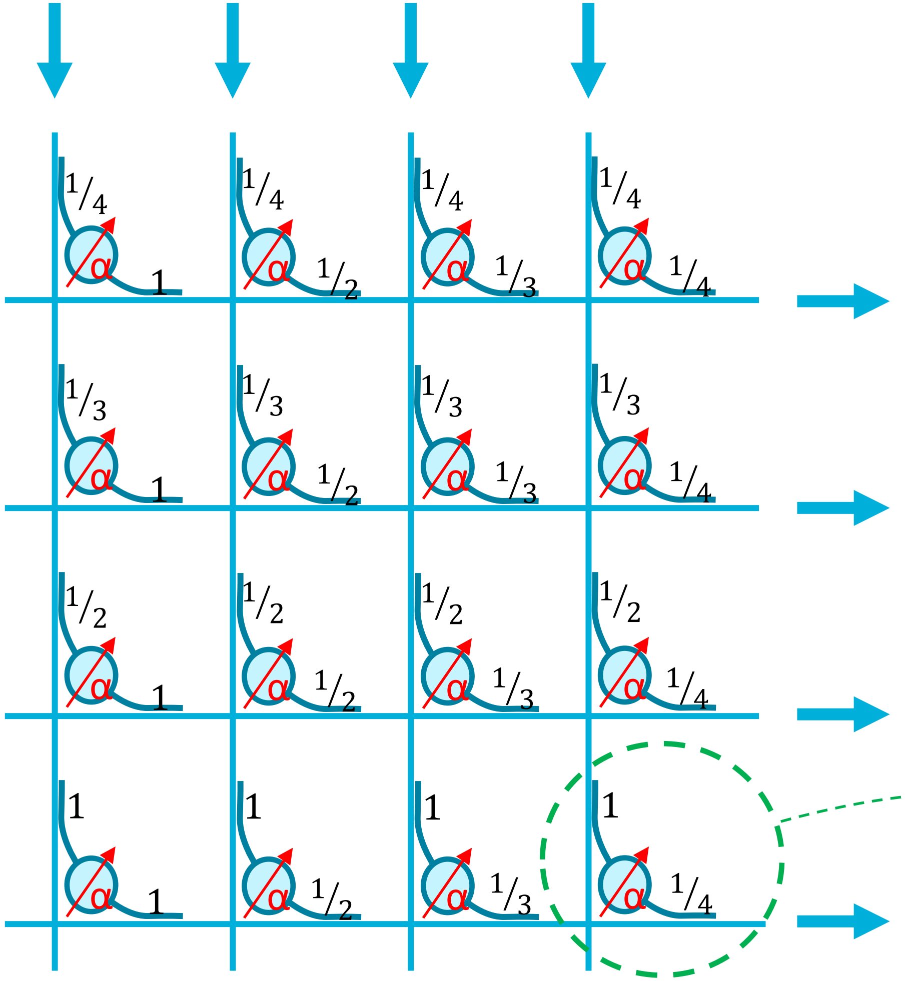
Crossbar with tunable attenuators, incoherent light

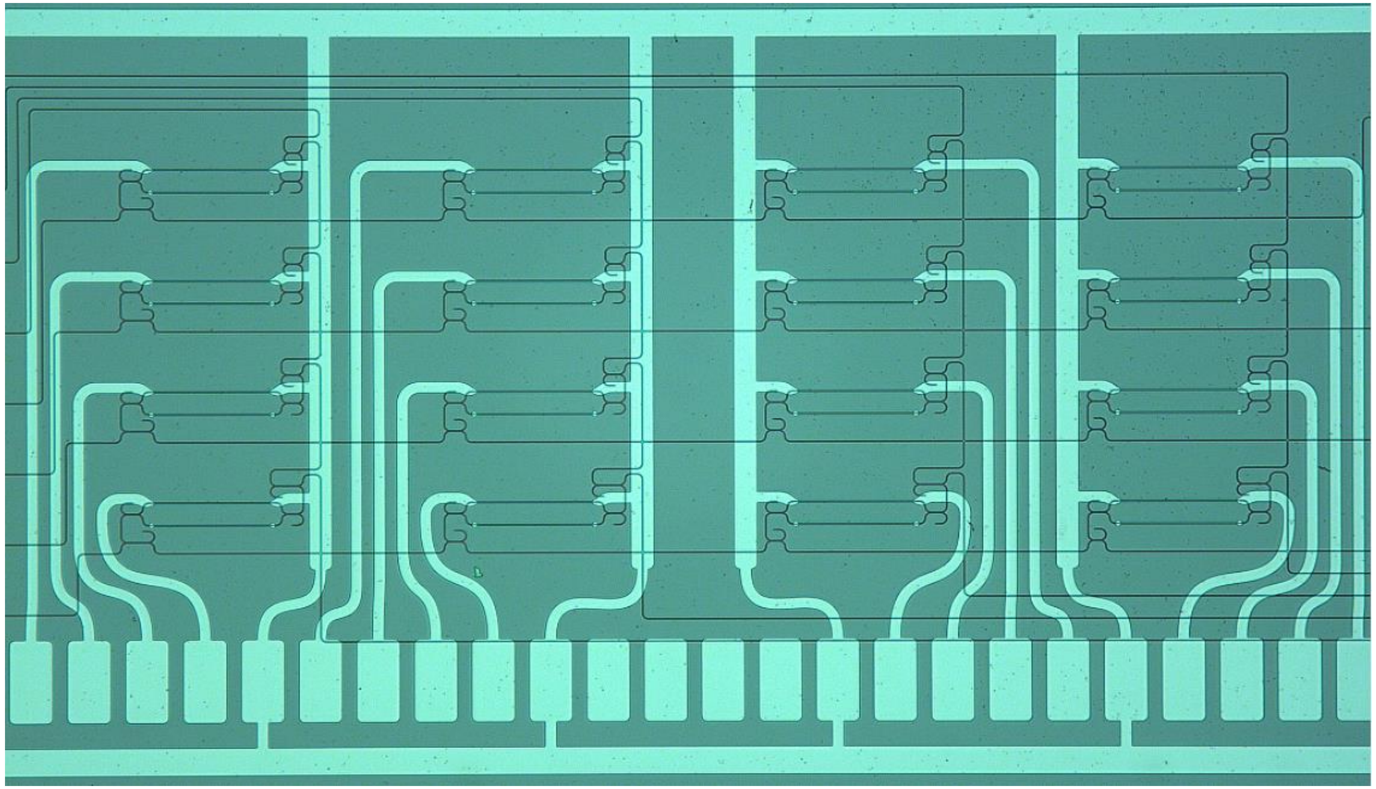
- Equal signal distribution along columns
- Equal signal accumulation along rows
- One tunable attenuator per intersection/coefficient:
  - N^2 heaters
  - Simple control
- **However:** Power loss (factor 1/N) in the directional couplers for signal accumulation along the output rows:

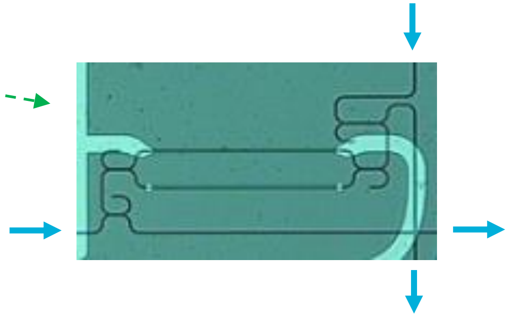




Crossbar with tunable attenuator: Hardware



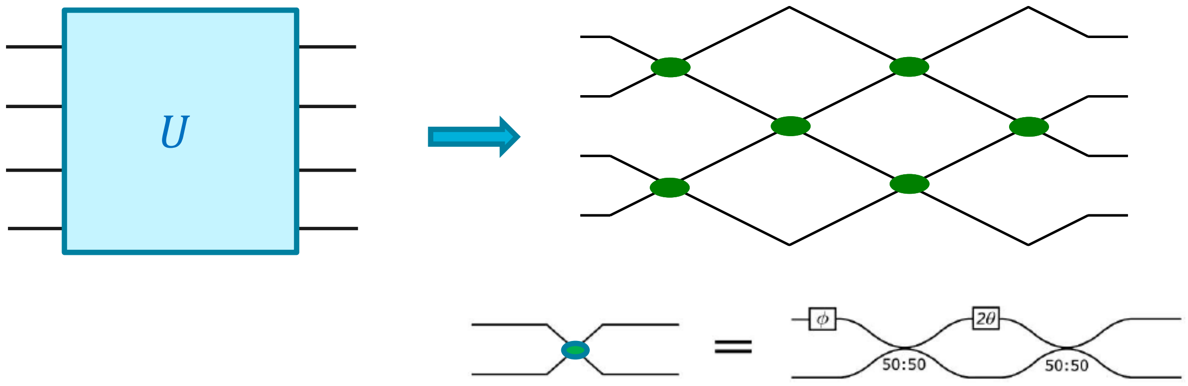




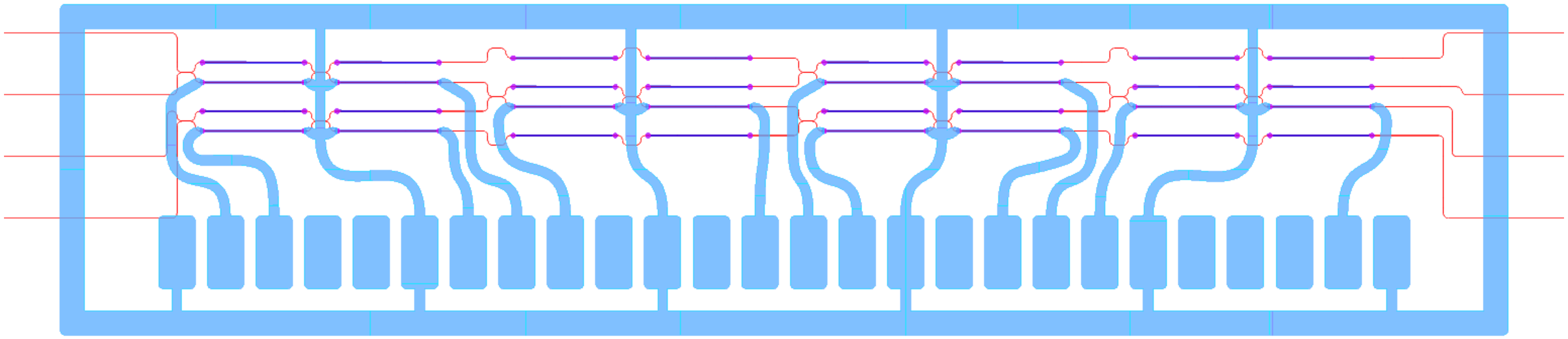
Synaptic interconnect, coherent light

Optical implementation of a unitary matrix multiplier:

- Control requires N*(N-1) heaters
- Complicated (sensitive?) tuning algorithm



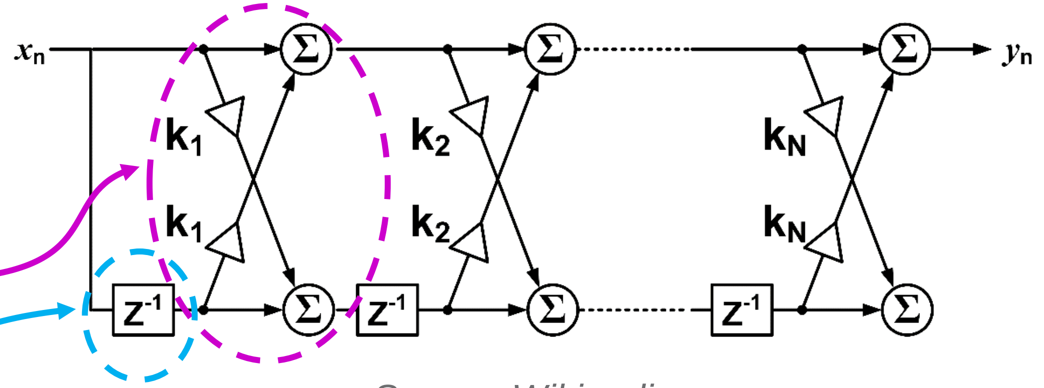
W.R. Clements et.al. , “Optimal design for universal multiport interferometers”
http://dx.doi.org/10.1364/OPTICA.3.001460



Lattice filters for optical convolution processing

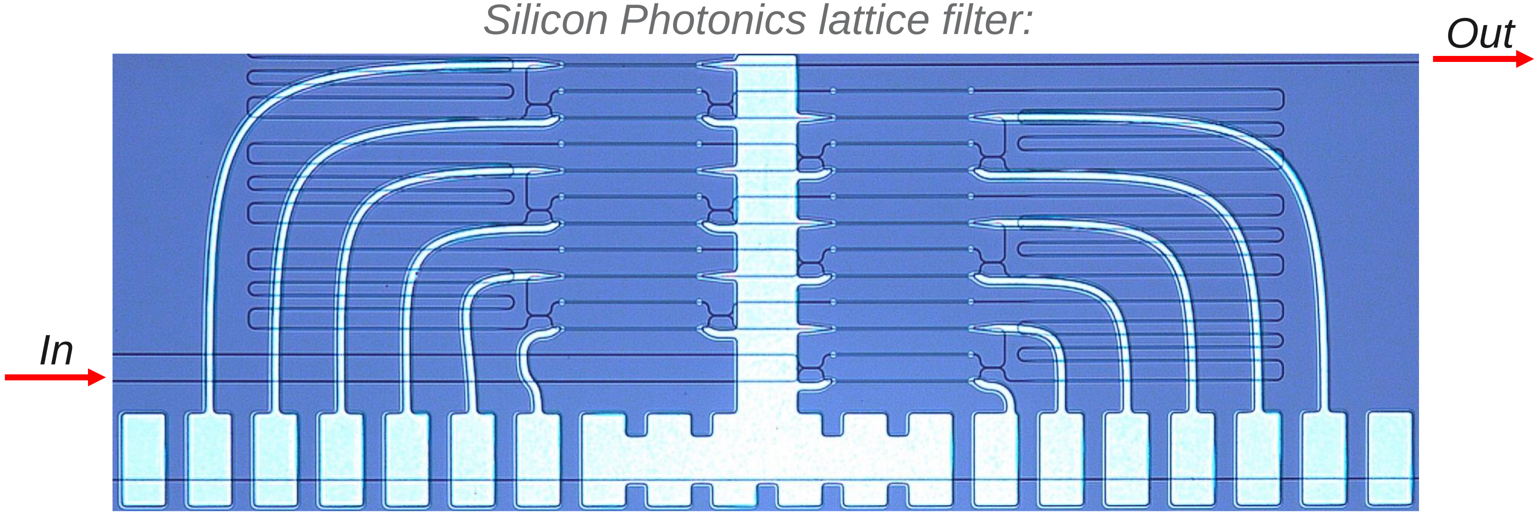
- A **Finite Impulse Response** filter performs a **convolution** on a discrete time series of input data
- Implementation in Silicon photonics:
  - Tunable Mach-Zehnder Interferometers, as power splitter-combiners
  - Folded waveguides as delay lines
  - Thermo-optic phase shifters for control

Lattice form discrete time series FIR filter:

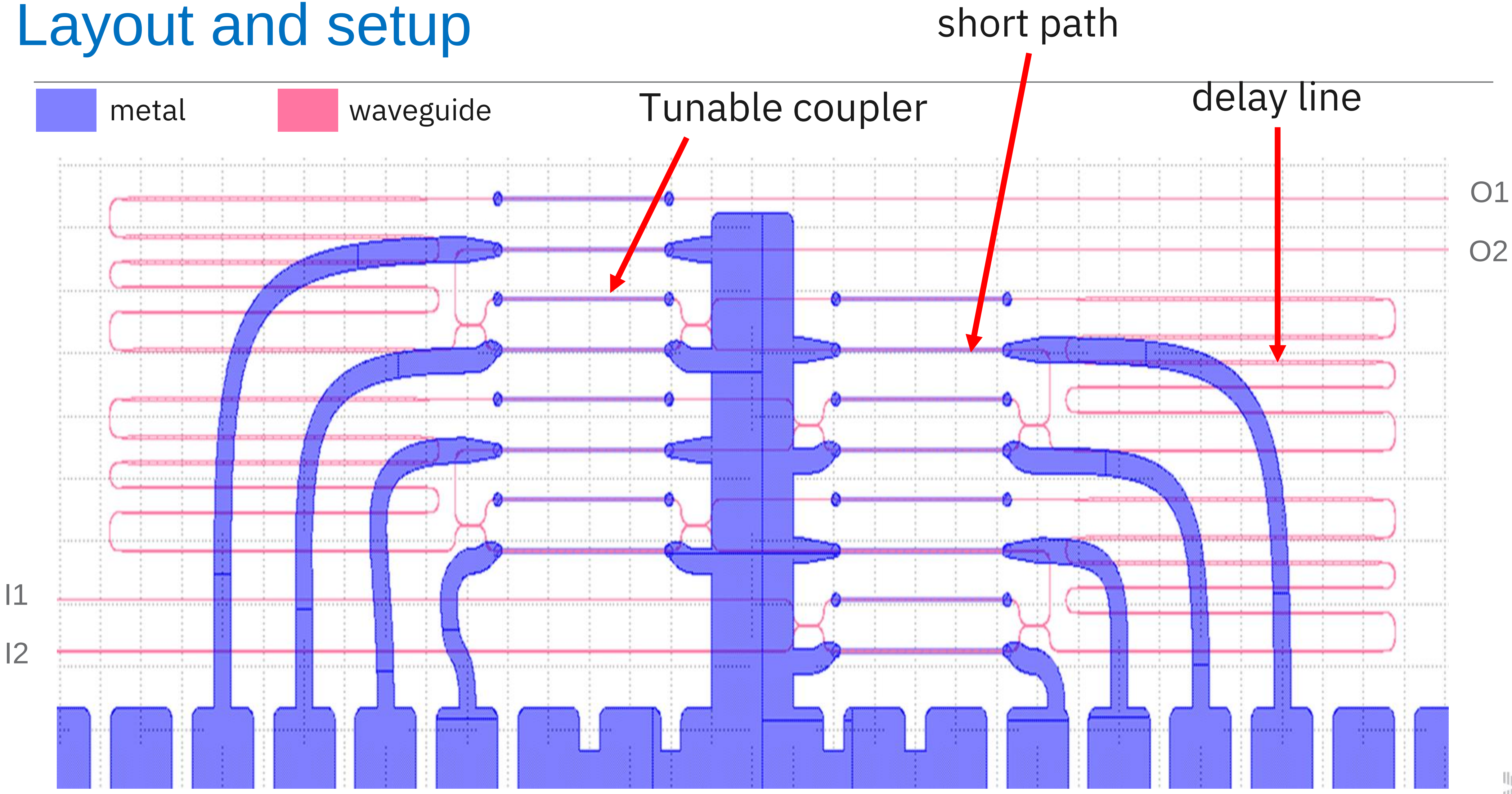


Source: Wikipedia

Silicon Photonics lattice filter:

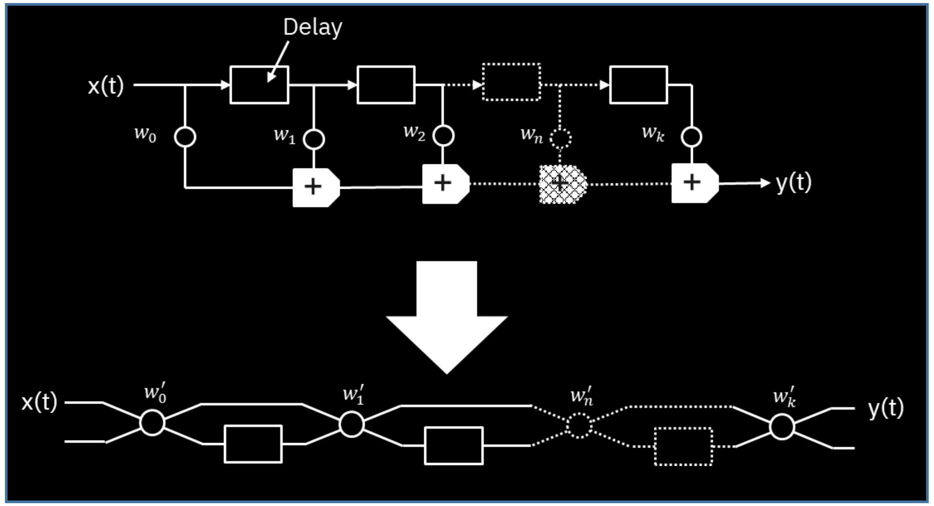


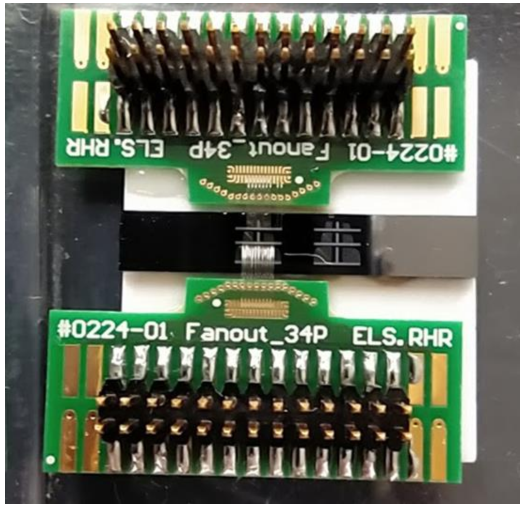
Layout and setup

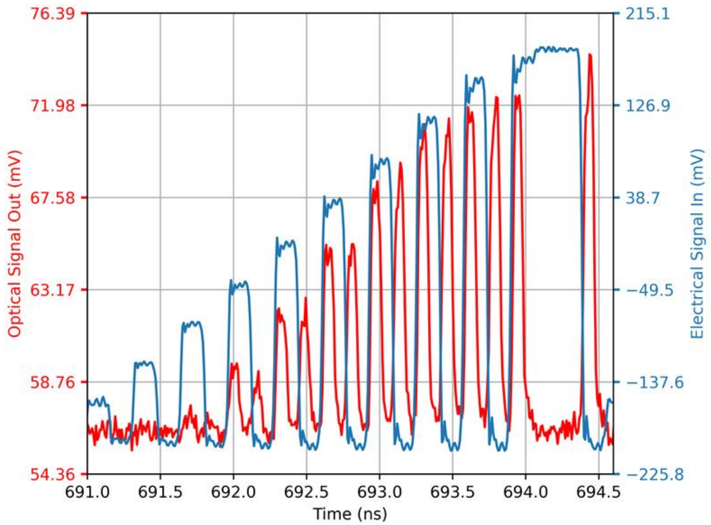


Optical convolutional signal processor

- Photonic implementations, volatile weights but well controlled and fast set







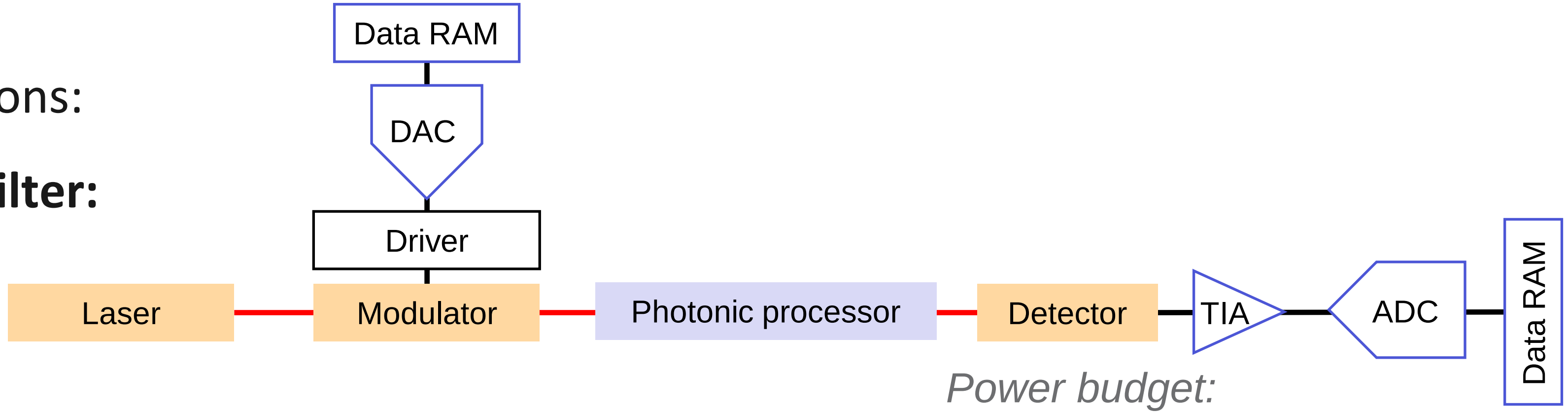
Measurements by Pascal Stark

- Time domain operation
- High-speed signal processing (12.5 GSample/s)
- Fast and efficient reconfiguration (electro-optic modulators)

Lattice filter: Link and Power budget

Link and Power budget calculations:

- **Scaling limits for the lattice filter:**
  - **Stage loss**
  - **Control complexity**



Link budget:

Parameter	Value	Unit
CW laser launch power	**13.0**	**dBm**
Laser to chip coupling	0.2	dB
Modulator insertion loss	3	dB
Lattice filter loss: **9 stages** @ 0.58 dB/stage	5.2	dB
Kernel normalization loss	2	dB
Detector coupling loss	0.2	dB
Optical power at photodetector	**2.6**	**dBm**
Power penalties (jitter, crosstalk, ISI etc.)	1.7	dB
Effective optical power at photodetector	**0.9**	**dBm**
Optical Sensitivity for a resolution of 4 bits, at 32 GSps	**-2.4**	dBm
Available link margin	**3.3**	**dB**

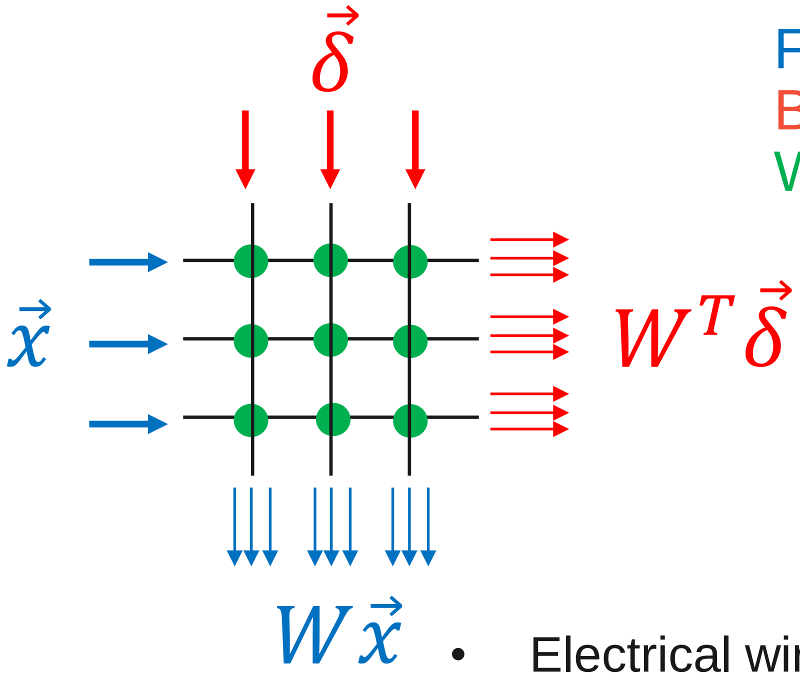
Power budget:

Building Block	Power in mW at 32 GSps:
Data RAM (read)	11
High Speed DAC	67
Driver and Modulator	70
Detector and TIA	6
Output ADC	115
Results RAM (write)	10
CW Laser	200
Sum of Power	476
Efficiency TOps/Watt	**2.49**

- Comparable to existing digital hardware, but
  - High-speed, low latency / Real time
  - Can do complex data and kernels
  - Room for further improvements

Analog signal processing for neural network training

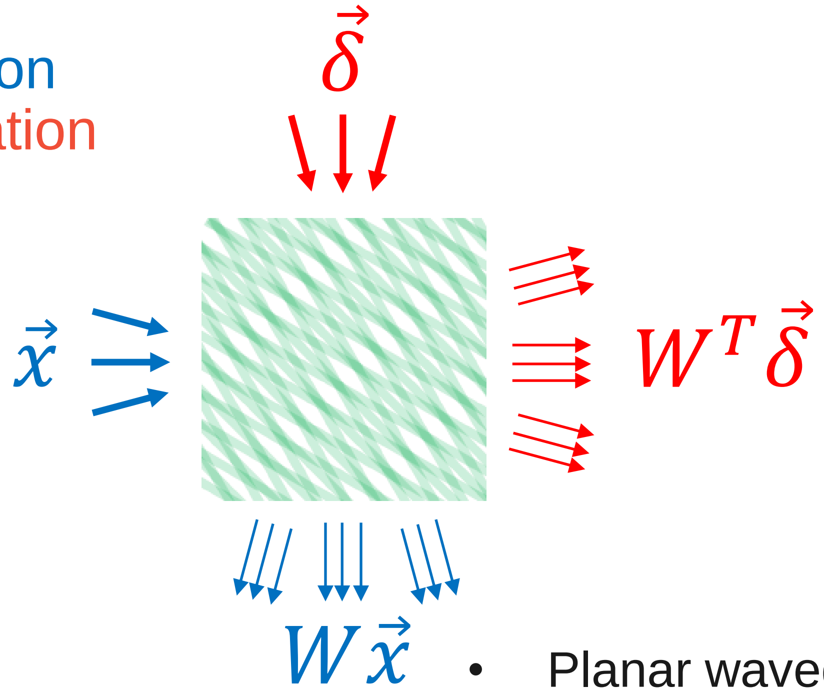
Electrical crossbar



- Electrical wires
- Local weights
- Resistance tuning

Forward propagation
Backward propagation
Weights

Photonic crossbar



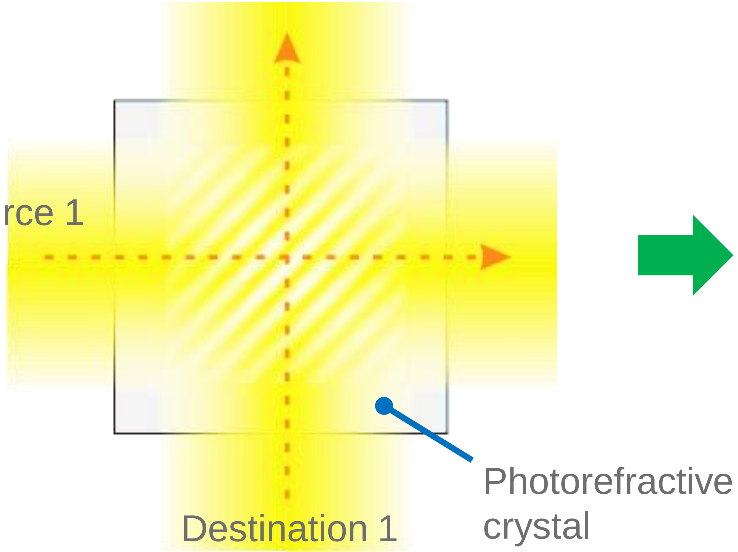
- Planar waveguiding
- Distributed weights
- Refractive index tuning

Writable photorefractive gratings provide the same functionality as the tunable resistive elements in a crossbar unit

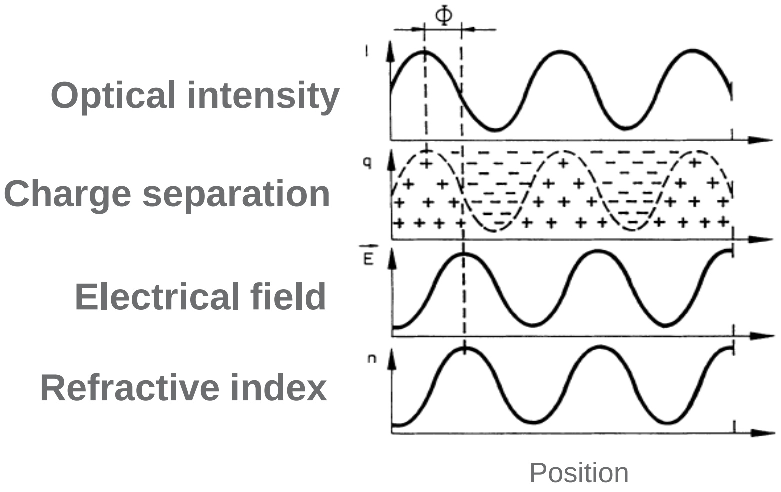
Optical crossbar arrays: Holographic storage and signal processing

Weight Storage:

Interference pattern:

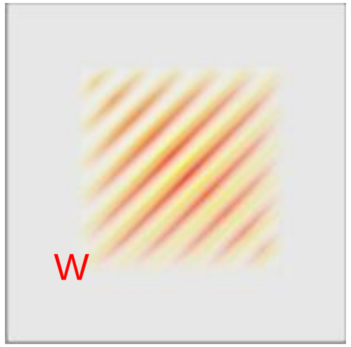


Photorefractive effect:



📖 Cornelia Denz, Optical Neural Networks, 1998.

Stored diffraction grating:



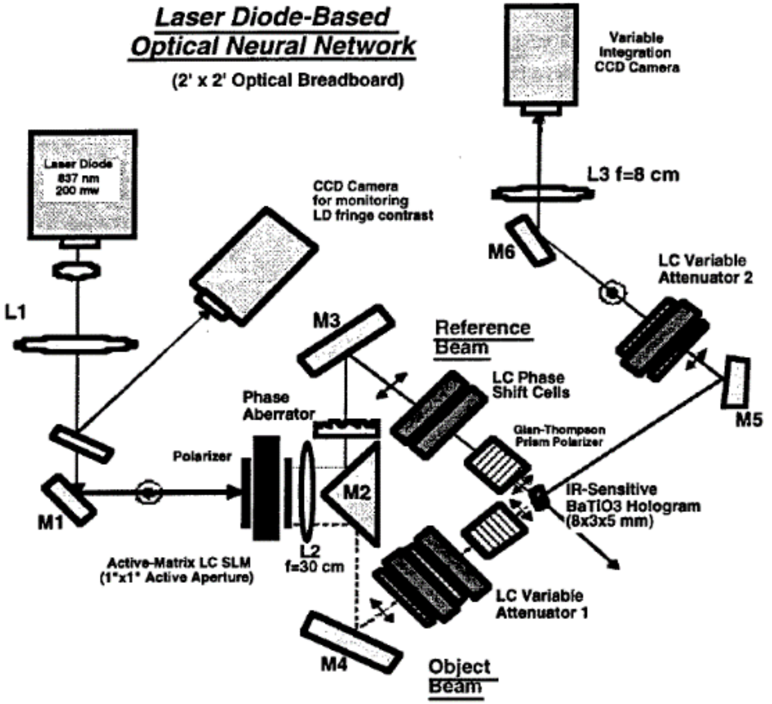
Synaptic weights are stored as refractive index gratings in a photorefractive material:

- Grating are written by two interfering optical beams
- Photorefractive effect: Optically active electron traps + Pockels effect → refractive index grating
- Linear and symmetric process

Optical crossbar arrays: Integrated Solution

Concept demonstrated in bulk optics

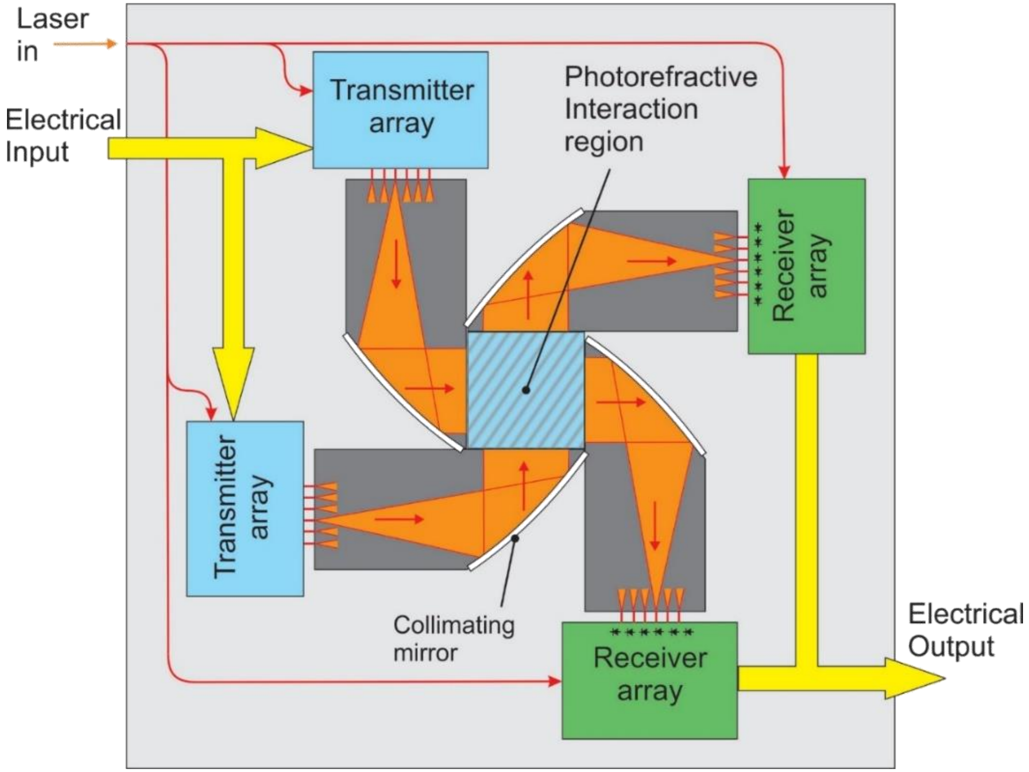
- Backpropagation training of neural networks with hidden layers
- Large setup, slow electro-optics, stability issues



📖 Yuri Owechko and Bernard H. Soffer, "Holographic neurocomputer utilizing laser diode light source“, 1995

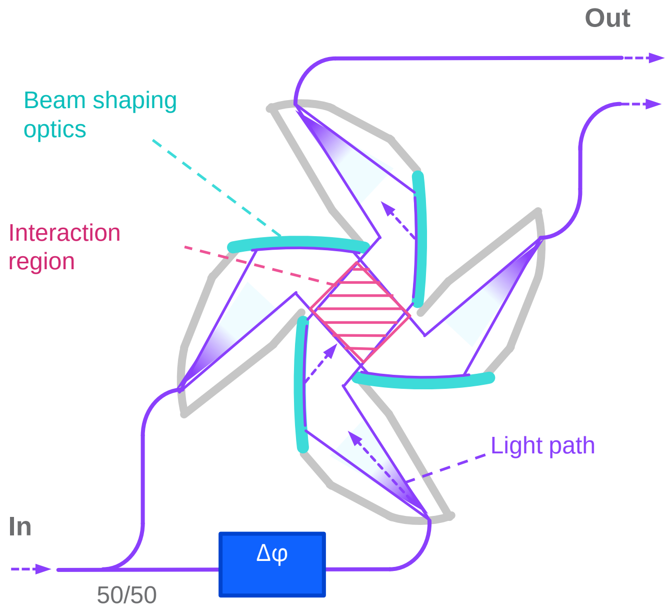
→: Miniaturize using Integrated Optics

- Electro-optic conversion and beam shaping optics on a silicon photonics chip
- Memory: Photorefractive thin film on silicon

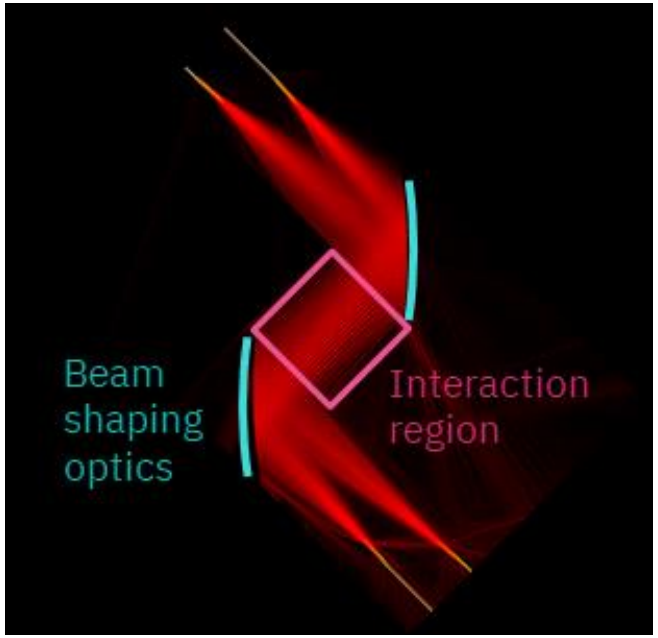


Photorefractive gratings in GaAs – Integrated photonic implementation

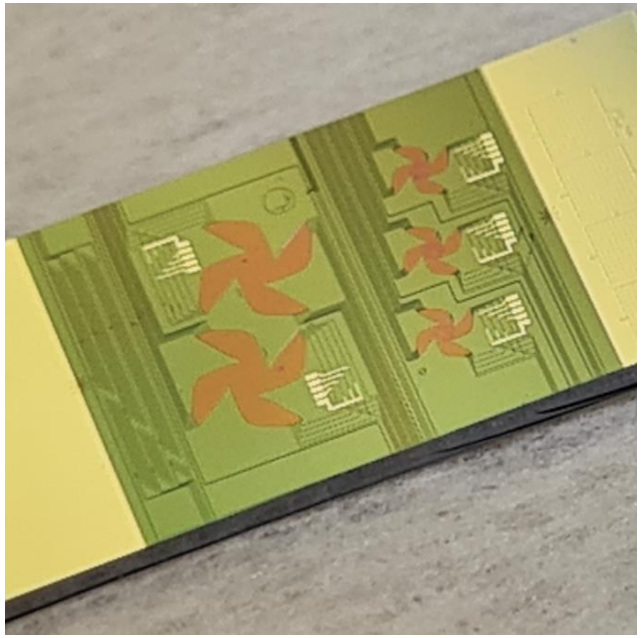
Photorefractive processor



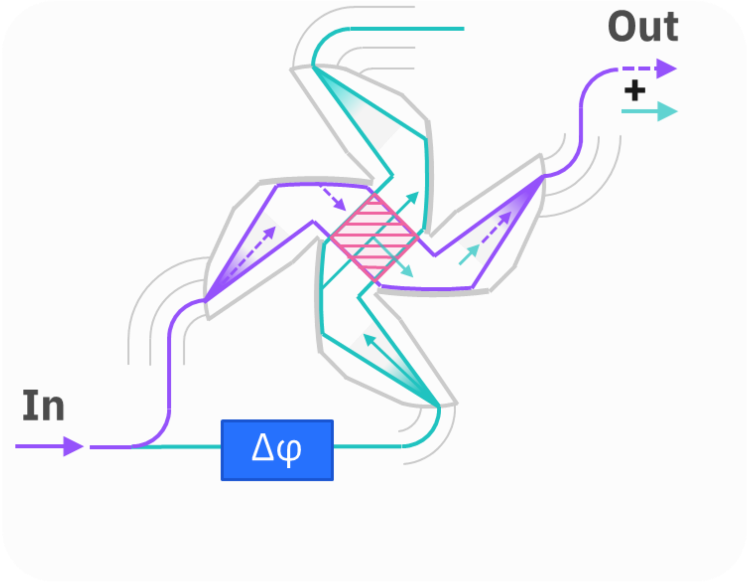
Simulated transmission

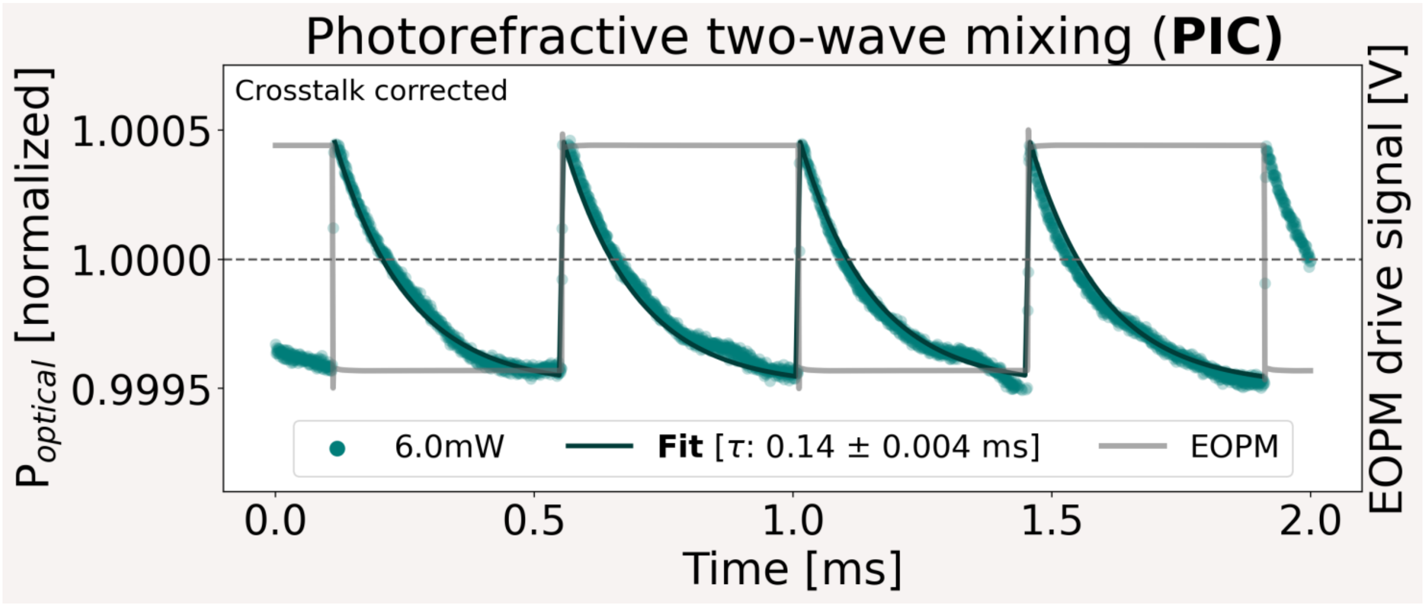


Manufactured chip



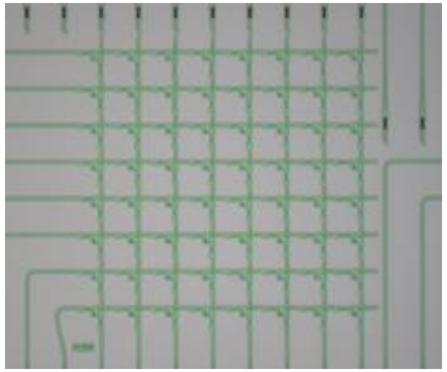
Periodic synapse writing





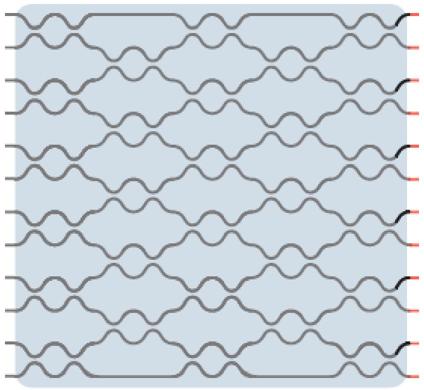
Concept comparison

Photonic crossbar



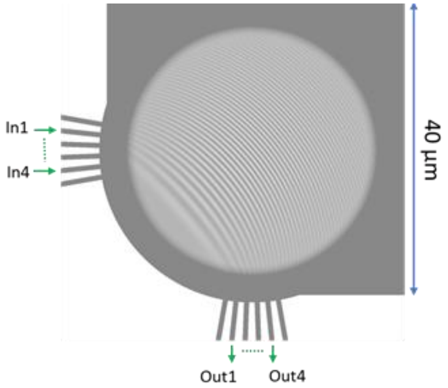
- Inference
- Incoherent
- Available
- ‘Simple’ control
- Inherent loss (1/N)
- Scalable to N≈20
  - Loss limited

Interferometric



- Inference
- Coherent
- Available
- ‘Complex’ control
- Lossless
- Scalable to N≈64
  - Complexity limited

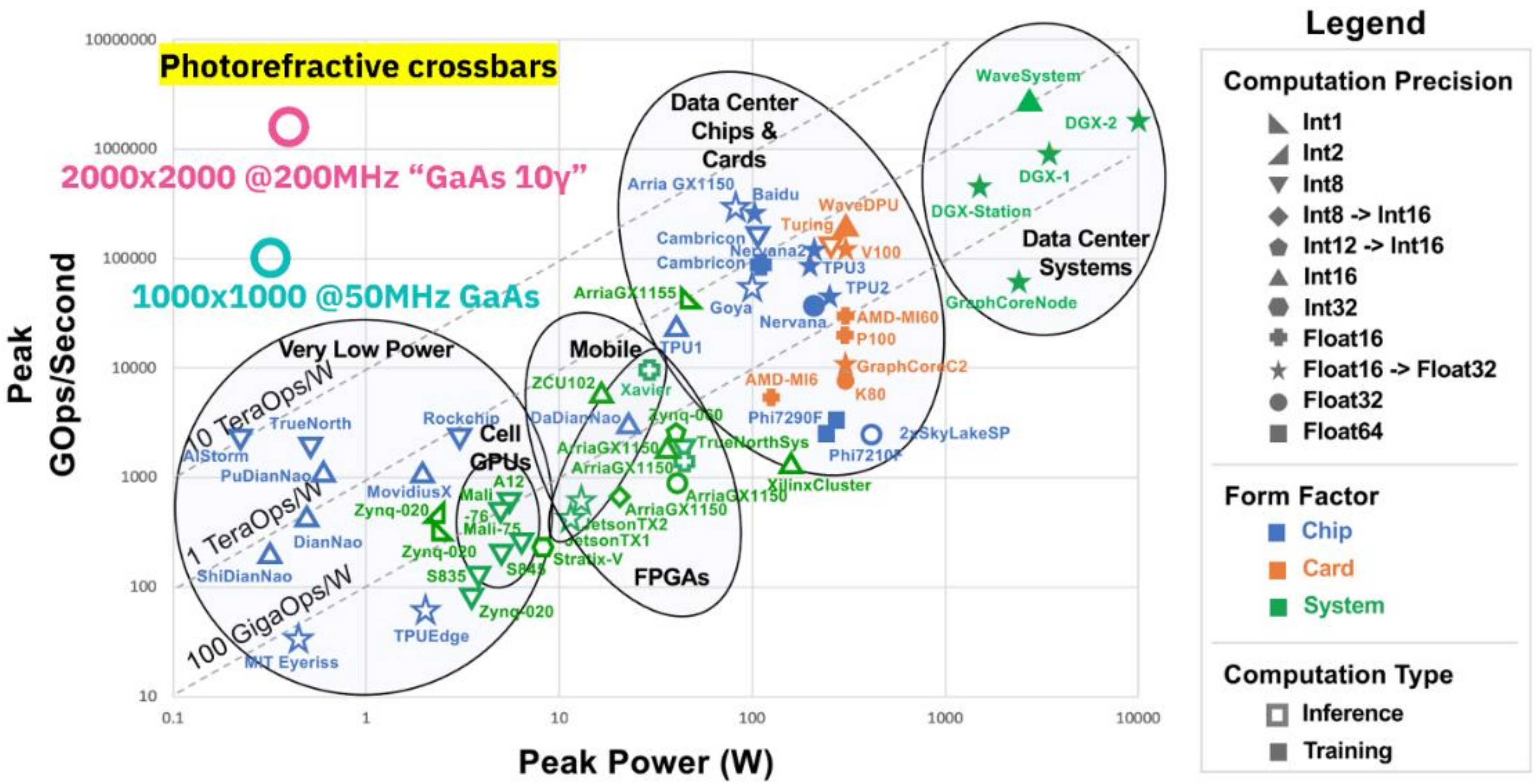
Diffractive



- Inference & training
- Coherent
- Partially available
- ‘Complex’ control
- Low loss
- Scalable to N≈256
  - IO circuit limited

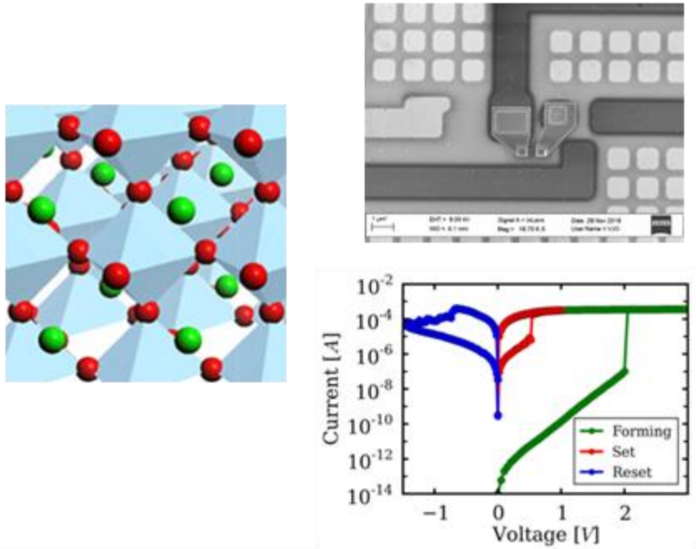
- Power-efficiency improves with N
- S/N determines the resolution and operating speed

Power-efficiency and scalability

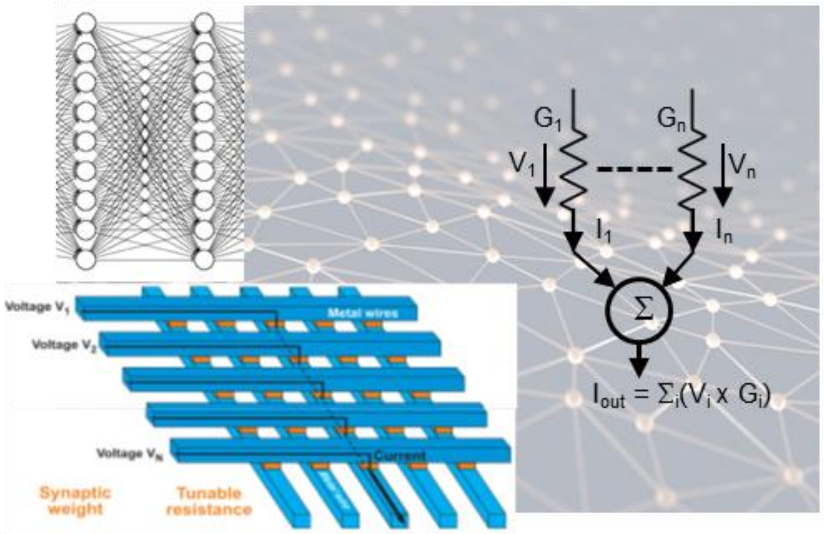


Doi: 10.1109/HPEC43674.2020.9286149

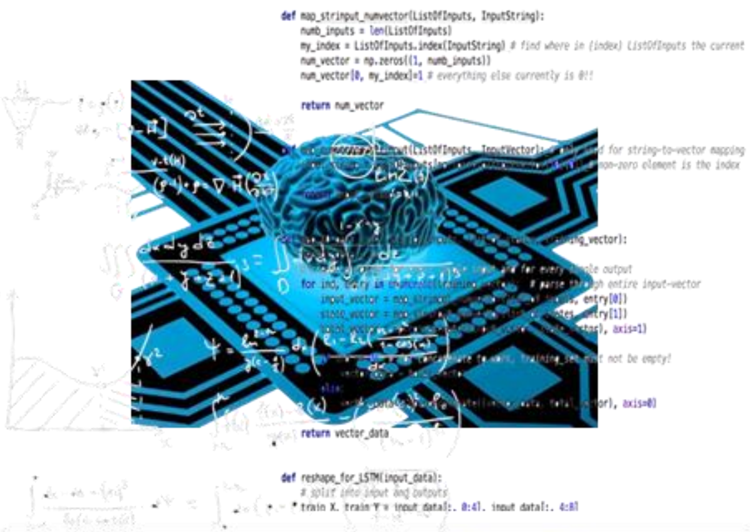
Innovation required at all levels



New Materials and Devices



Non *von Neumann* Architecture



Hardware – Algorithm Interplay

New technologies for Artificial Intelligence - The team

Acknowledgments

IBM Research – Zurich, Switzerland
Neuromorphic Devices and Systems team

The IBM BRNC cleanroom opteam

Co-funded by the European Union Horizon 2020 Programme and the Swiss National Secretariat for Education, Research and Innovation (SERI)

EU & CH-SERI
PHOENICS, PHOENIX, PROMETHEUS

Thank you for your attention!

OFB@zurich.ibm.com