Silicon Photonics for Neuromorphic Computing

Acceleration of Deep Neural Network training

Folkert Horst, IBM Research - Zurich

Overview

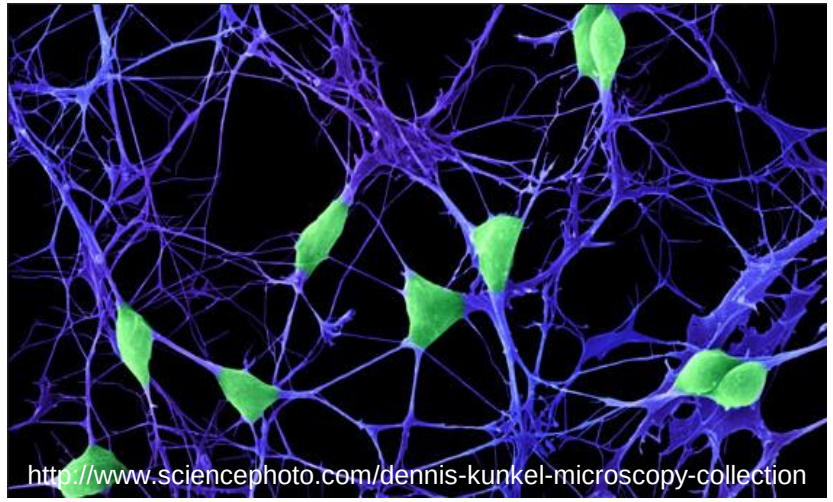
- Neuromorphic computing ➔ Artificial Deep Neural Networks
  - Training of deep neural networks
  - Processing of synaptic weights
  - Need for non Von-Neumann computing architectures
- Analog synaptic weight storage and processing in crossbar arrays
  - Electric crossbar arrays
  - Optical crossbar array using holographic storage and signal processing
- Integrated optical crossbar array in Silicon Photonics
  - Optical components
  - Holographic storage medium
- Summary & Outlook

Neuromorphic computing = Brain inspired computing

Motivation: The outstanding features of the (human) brain:

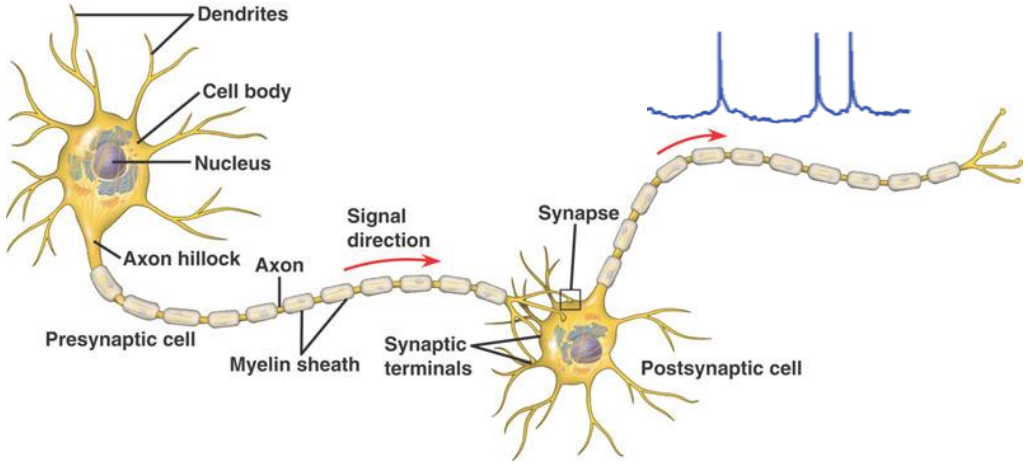
- **Power efficiency** (human brain consumes ~ 20 W)
- **Remarkable pattern recognition performance:** Recognition of (subtle) patterns buried in noise

Brain at neural network level:



- Human brain: ~ 100 billions neurons
- Each neuron is connected to 1'000 – 10'000 other neurons by synapses
- Signal transmitted by a synapse is adjustable: "synaptic weight"

Neuron level:

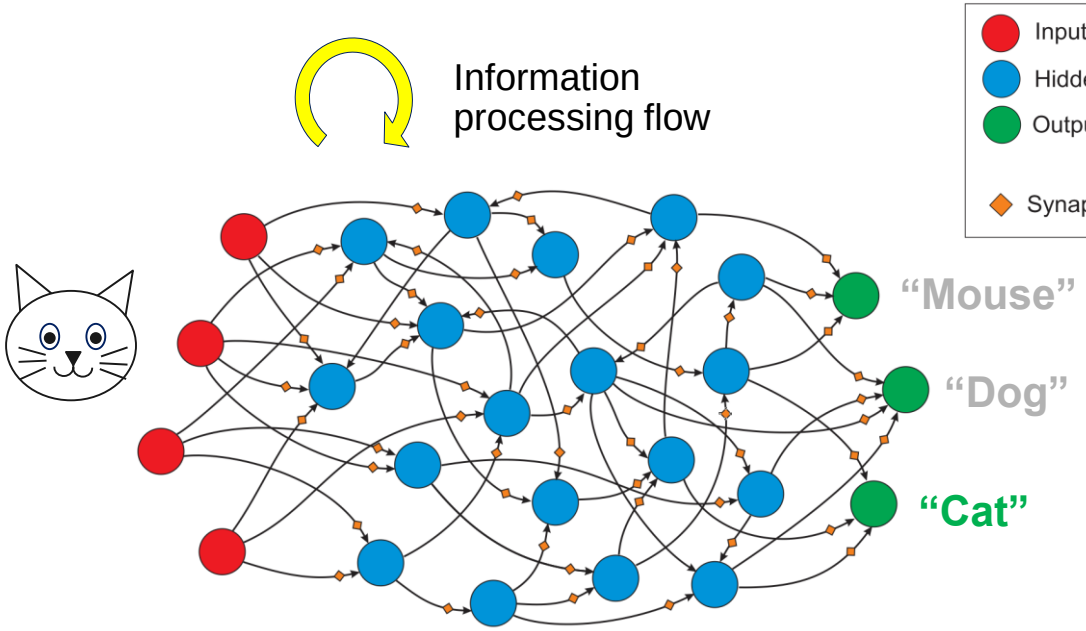


http://biomedicalengineering.yolasite.com/neurons.php

- Signaling between neurons: Spikes, spike trains
- Neuron activation: "Leaky Integrate and Fire"
- Learning: Adjustment of the synaptic weights
  - Spike Timing Dependent Plasticity: "Neurons that fire together wire together"

Brain inspired computing:

Brain-like Neural network:

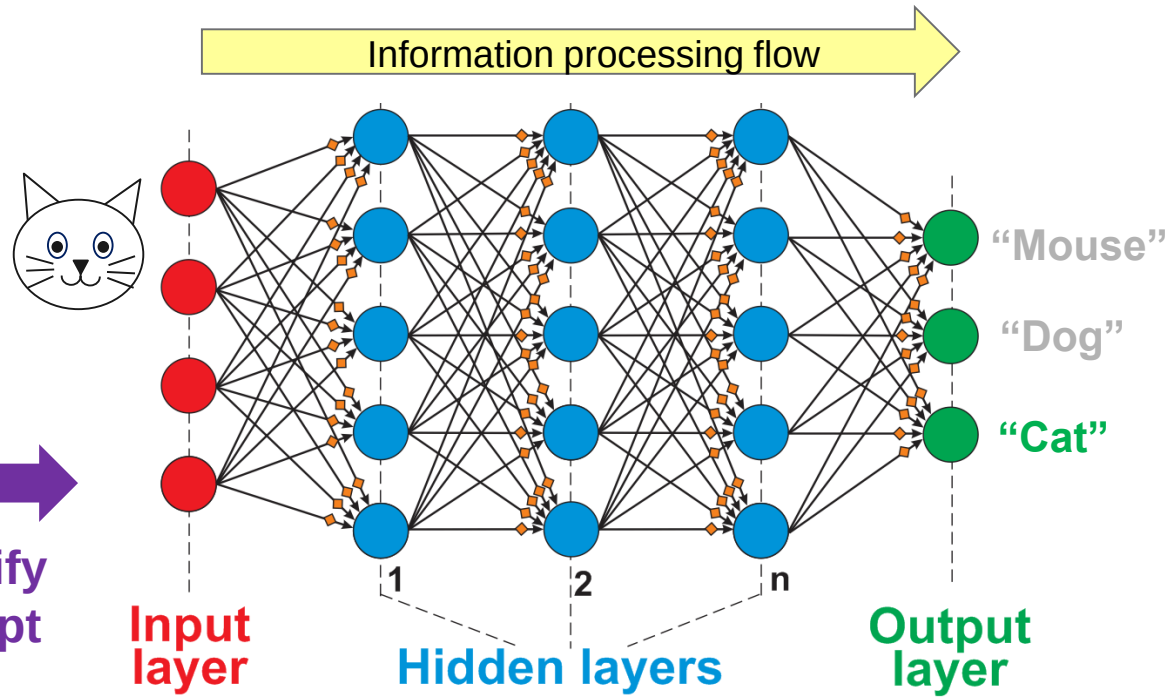


- Omni-directional signal flow
- A-synchronous pulse signals
- Information encoded in signal timing

➔ **Difficult to implement efficiently** on standard computer hardware

Simplify & Adapt

Deep Artificial Neural Network:

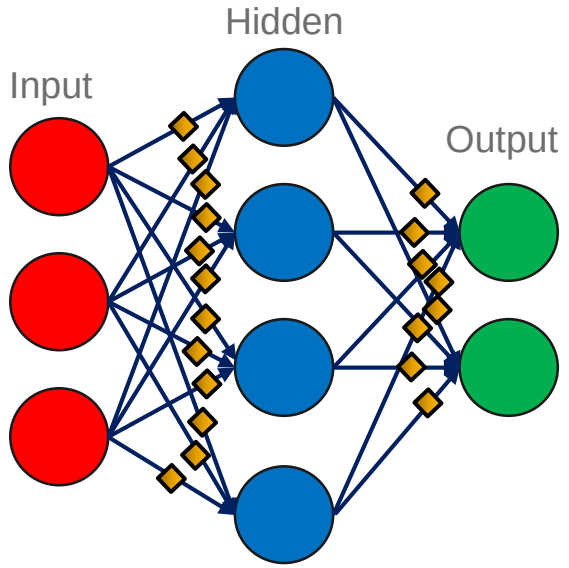


- **Better fit to standard hardware**:
  - Feed-forward sequential processing
  - Information encoded in signal amplitude
  - Multiply and Accumulate for weighted connections
  - Neuron activation: (soft) threshold function
- **Training: Backpropagation Algorithm**

Artificial Neural Network: Computations

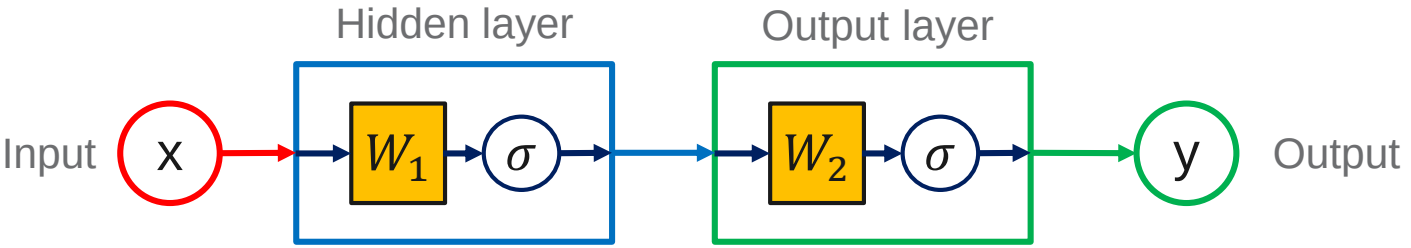
Components:

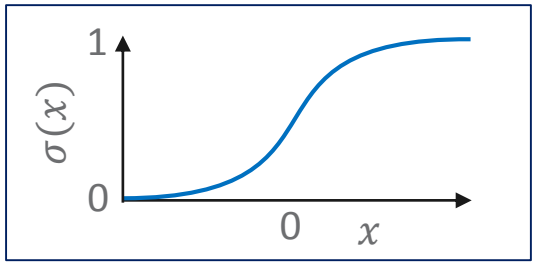
- Layers of neurons
- Synaptic interconnections



Mathematical operations:

- Thick lines: signal vectors
- W: Synaptic weight matrix
- σ: per-element neural activation function (sigmoid)

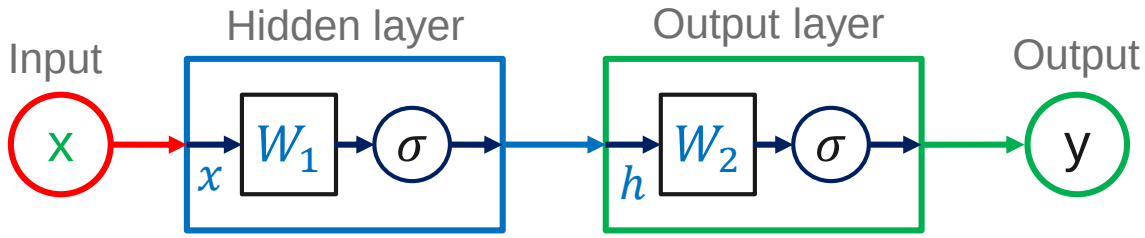




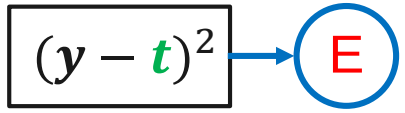
ANN Training: Backpropagation algorithm

Training case x with target response t:

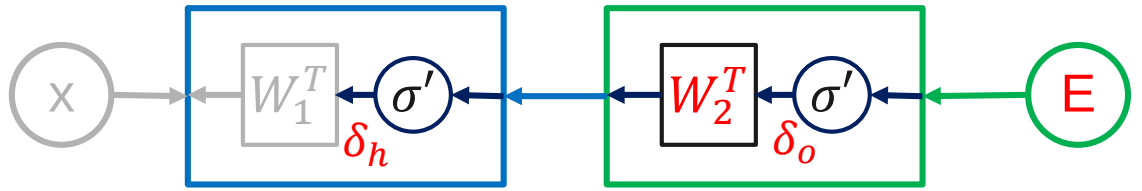
1. Forward Propagate ➔ Response **y**



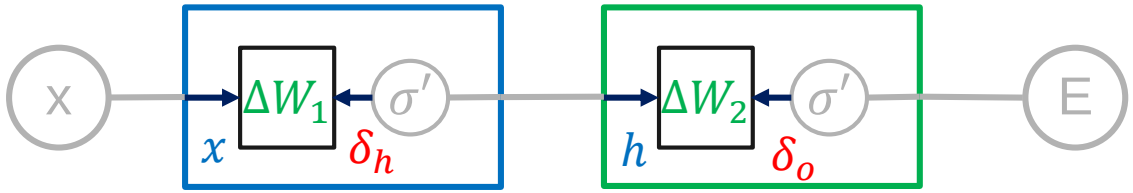
2. Determine output error:



3. Backward Propagate: Find neuron input signals that contributed most to the error



4. Find weights that were active, and that contributed to the error. Adjust weights to reduce error: $\Delta w_{ij} = -\eta \ \delta_i \ x_j$



5. Repeat for many, many different testcases

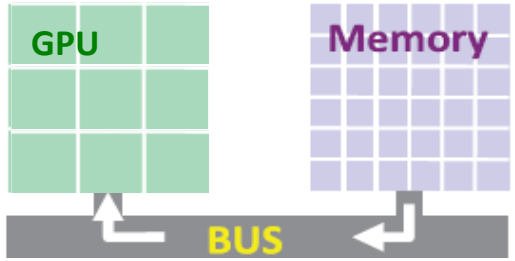
Efficient training of Deep Artificial Neural Networks:

- Training by Backpropagation Method:
  - Forward Propagation: $W_{1,2..}$
  - Backward Propagation: $W^T_{2,3..}$
  - Weight Update: $\Delta W_{1,2..}$

- **Many large matrix operations**
  - Scale $\propto N^2$ (Neurons/layer)

- **Large training datasets**: Thousands of training cases

Weight matrix processing has limited efficiency on standard Von Neumann systems:
- (Mostly) Serial processing
- Low computation to IO ratio → Memory bottleneck



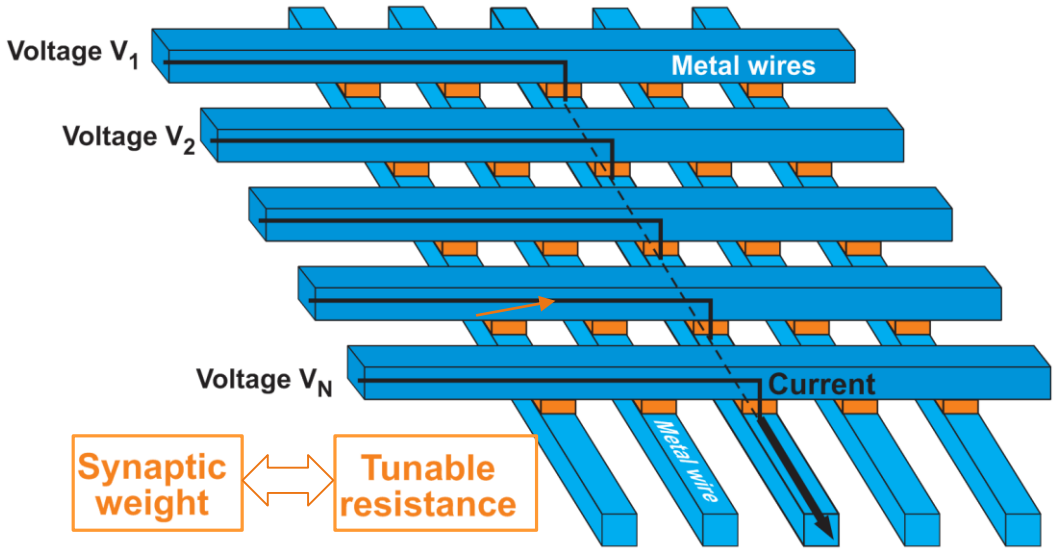
- **To accelerate weight matrix processing: Borrow some concepts from the brain:**
  - Analog signal processing
  - Fully parallel processing
  - Tight integration of processing and memory

Crossbar arrays

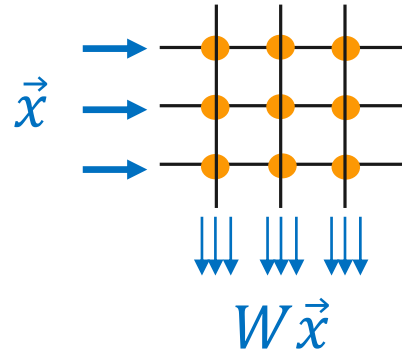
📖 G. W. Burr et al., "Tech. Dig. - Int. Electron Devices Meet. IEDM, vol. 2016–Febru, no. 408, p. 4.4.1-4.4.4, 2016.
📖 T. Gokmen and Y. Vlasov, Front. Neurosci., vol. 10, no. JUL, pp. 1–13, 2016.

Analog crossbar arrays:

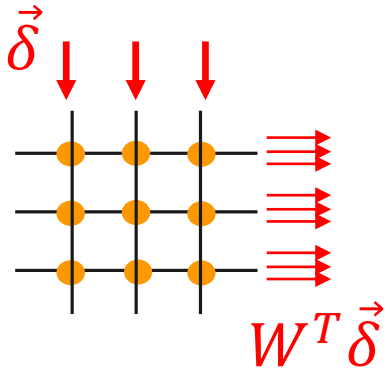
Electrical crossbar array:



Forward propagation:



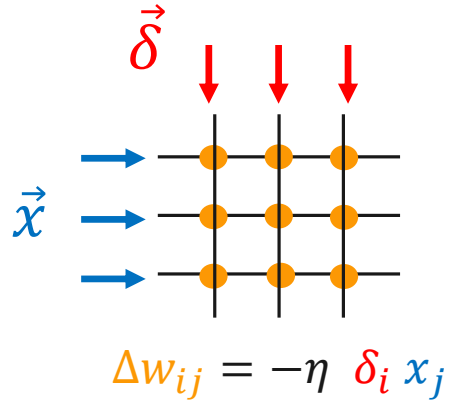
Backward propagation:



Challenge: Tunable weights

- Update must be proportional to signals on row and column
  - Symmetric increase and decrease of weight
  - ~1000 analog levels required
- **Difficult to find material systems** that meet these requirements

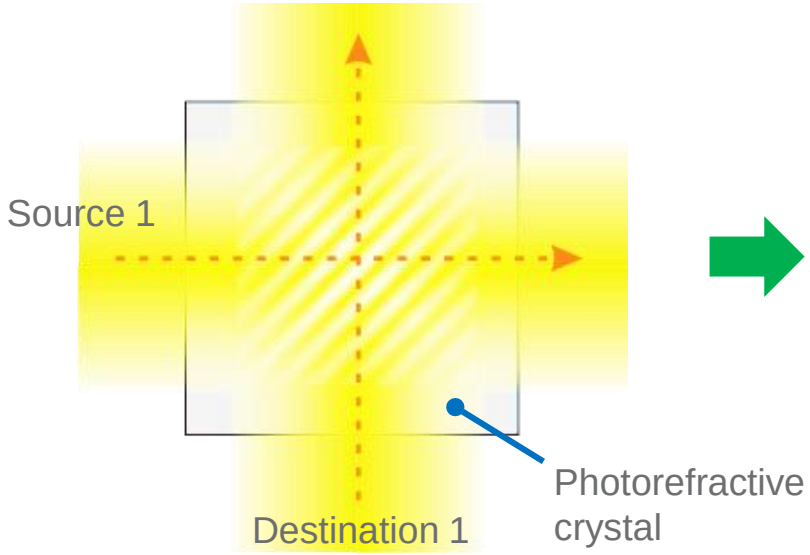
Weight update:



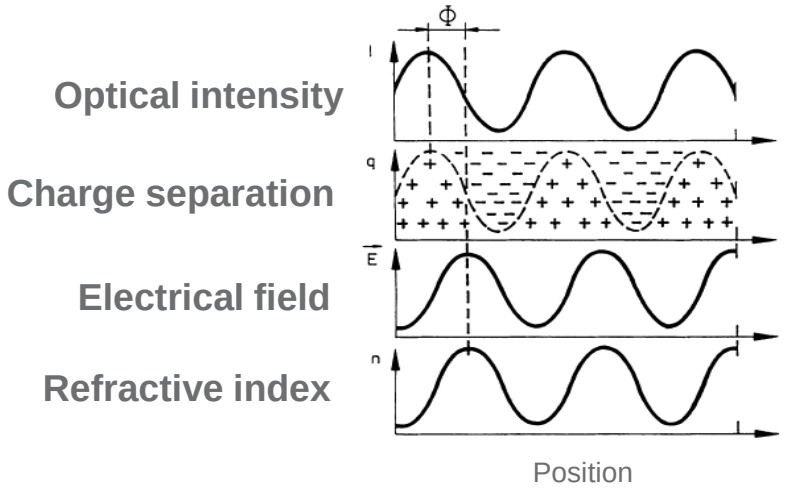
Optical crossbar arrays: Holographic storage and signal processing

Weight Storage:

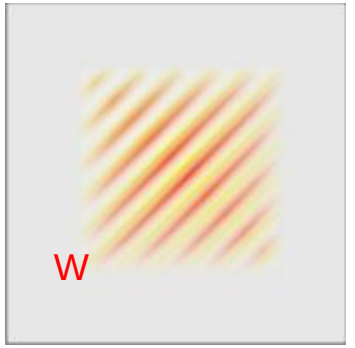
Interference pattern:



Photorefractive effect:



Stored diffraction grating:



📖 Cornelia Denz, Optical Neural Networks, 1998.

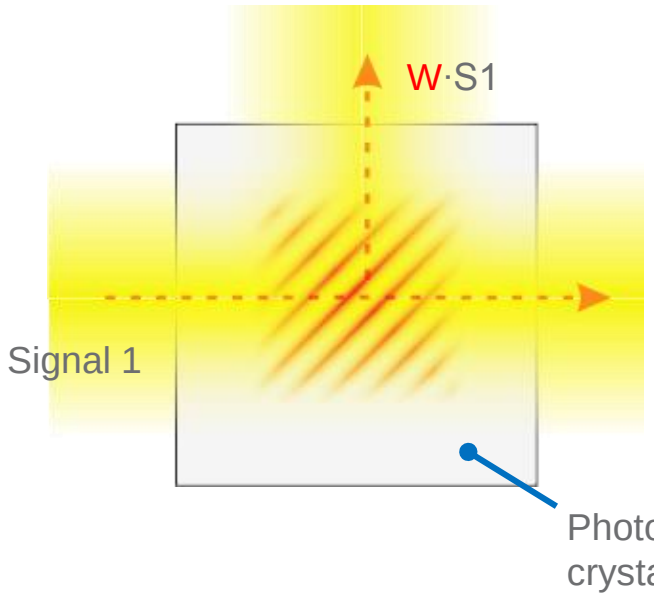
Synaptic weights are stored as refractive index gratings in a photorefractive material:

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

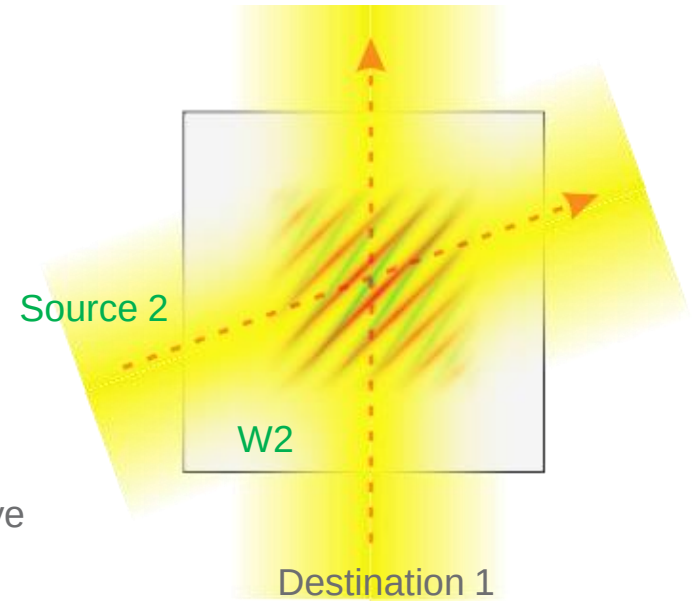
Optical crossbar arrays: Holographic storage and signal processing

Synaptic weight processing:

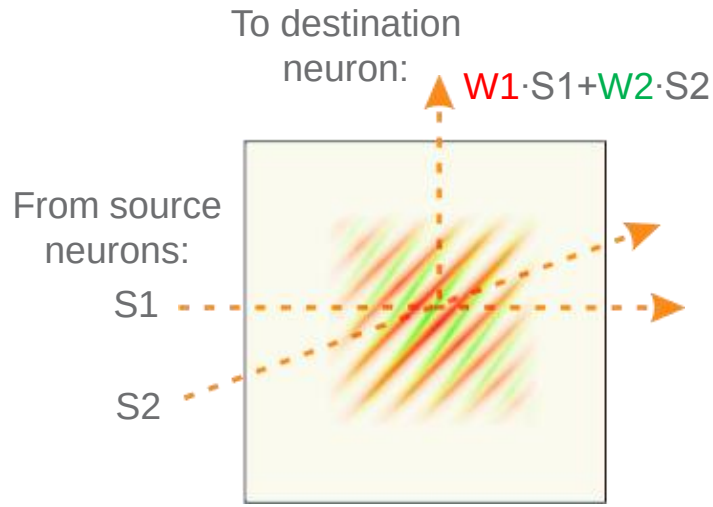
Diffraction grating readout:



Write a second grating:



Multiply & accumulate on two gratings:



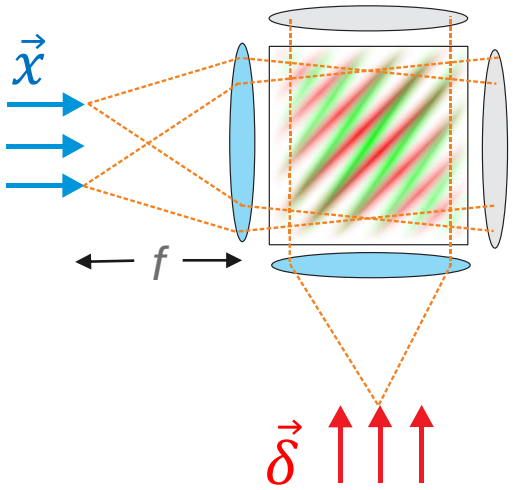
Synaptic weight gratings diffract light from optical input beams to optical output beams

- Different input/output signals are encoded by different beam angles in the crystal
- There is a unique grating for every input-output beam combination
- Optical signaling: amplitude & phase ➔ Bipolar signals and weights

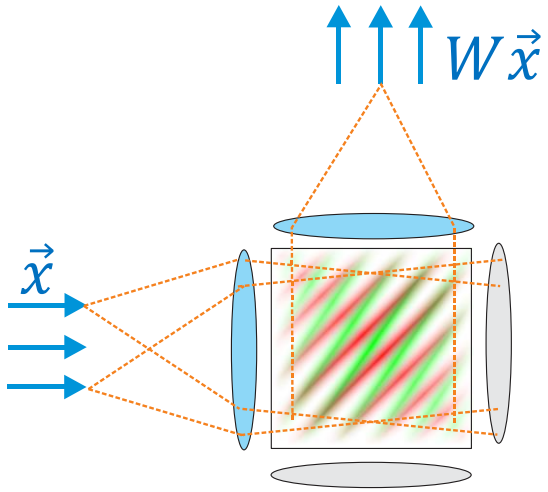
Optical crossbar arrays: Weight processing operations

- Add lenses to shape the optical beams:
  - Arrays of point sources → collimated beams under different angles → arrays of point images

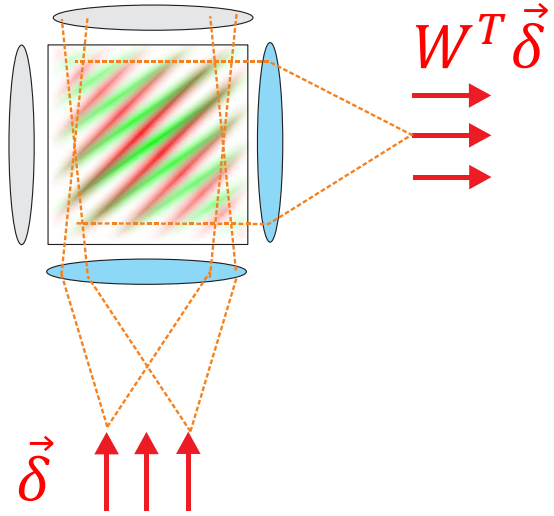
Weight update:



Forward propagation:



Backward propagation:

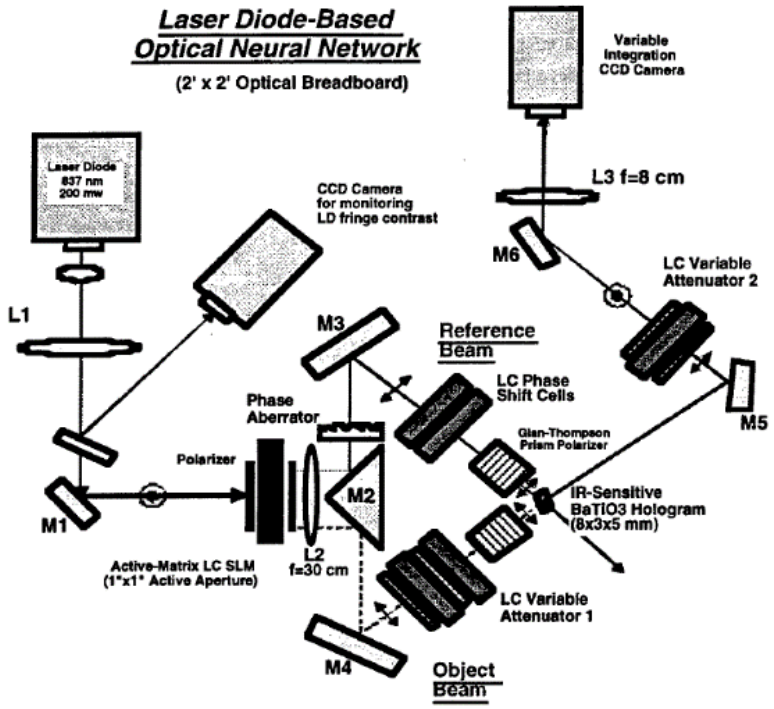


- All weight processing operations for backpropagation supported

Optical crossbar arrays: Integrated Solution

Concept demonstrated in 3D free-space optics

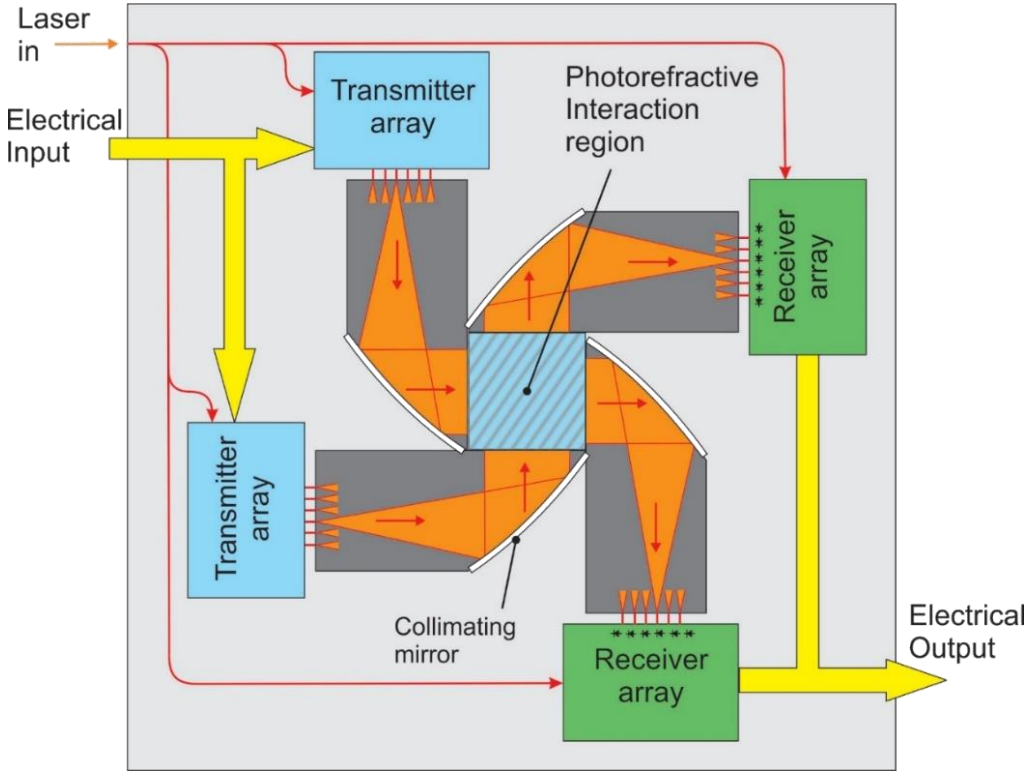
- In the 90s (Hughes Research Laboratories)
- Backpropagation training of ANNs shown
- Large setup, slow electro-optics, stability issues



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

Our approach: Miniaturize using Integrated Optics

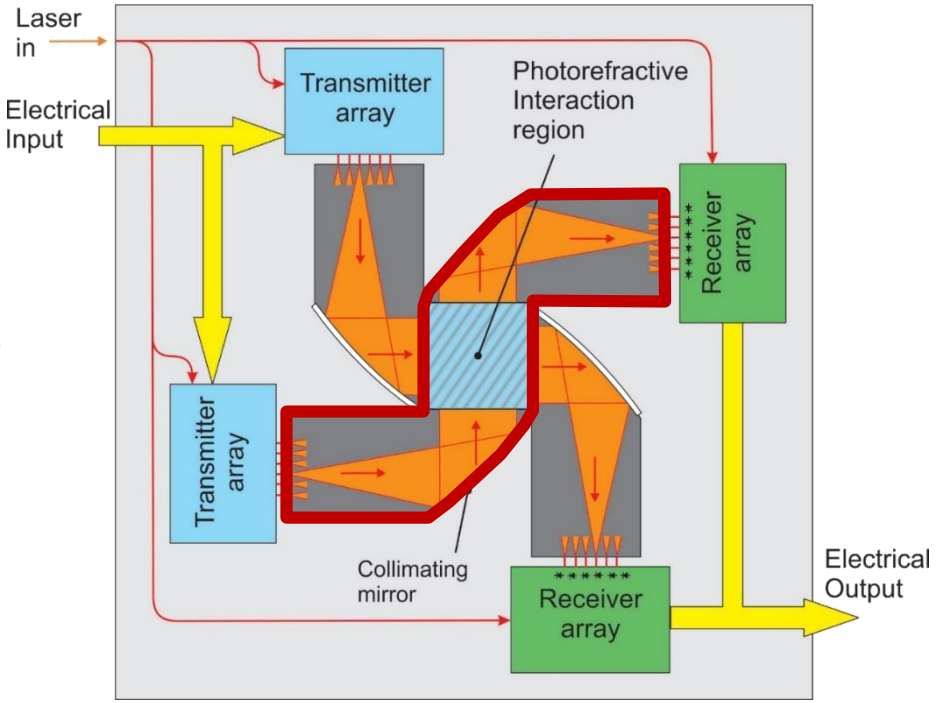
- Electro-optic conversion and beam shaping optics on a Silicon-Photonics chip
- Memory: Photorefractive thin film on silicon



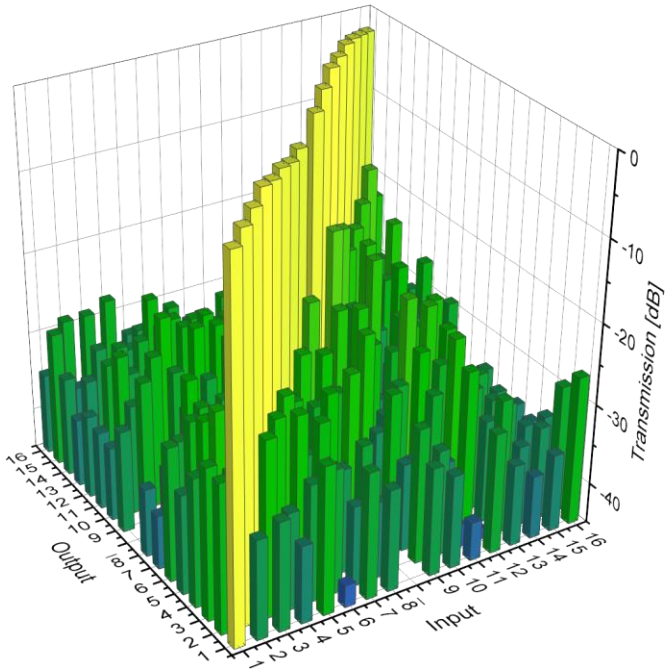
Photonic weight processing unit: Building blocks

Beam shaping optics:

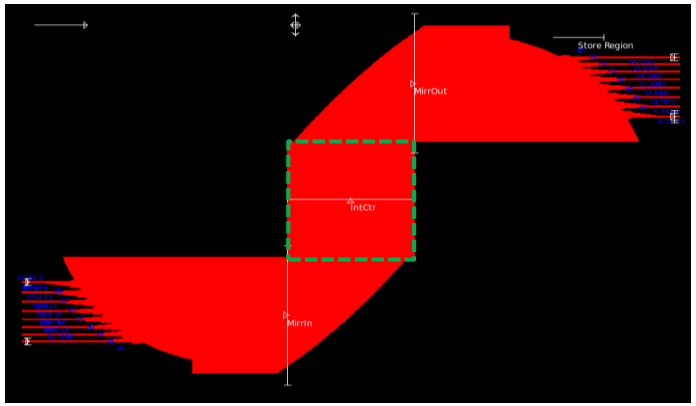
- Converts between point sources and plane waves
  - Parabolic collimating mirrors
  - Curved/tilted focal planes for aberration correction.



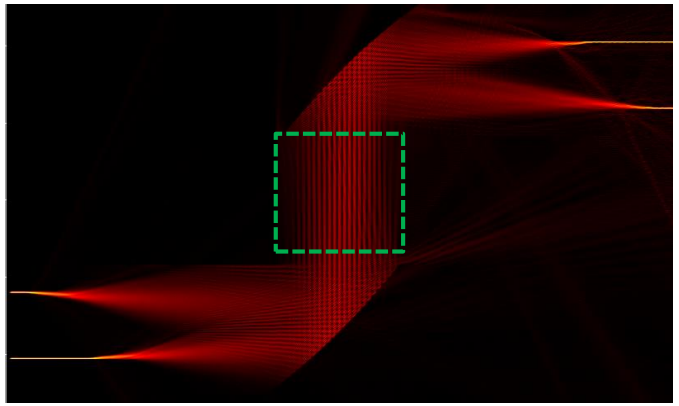
Measured transmission:



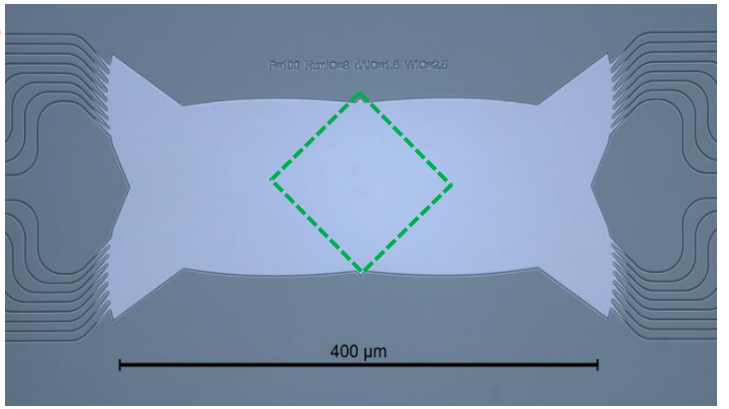
layout



Simulated transmission



Si-Photonics hardware:

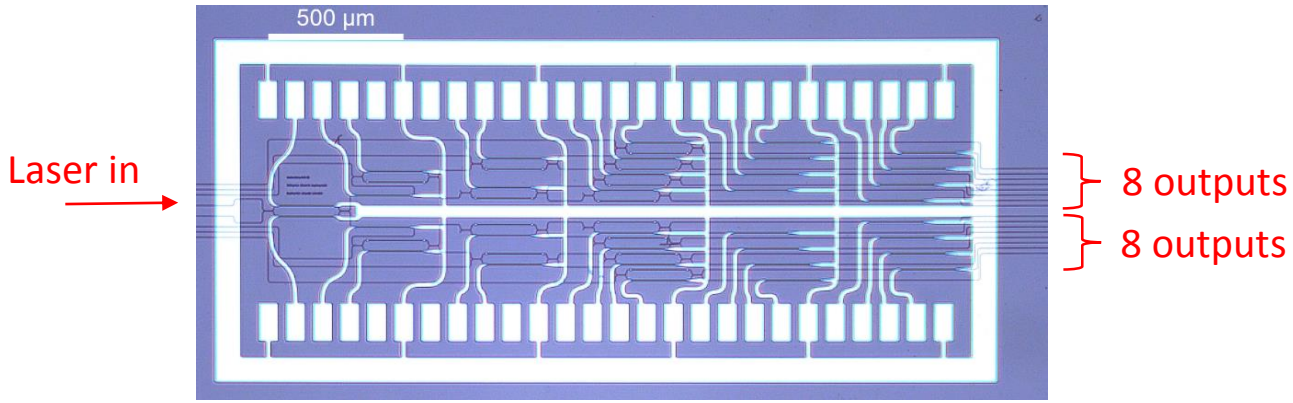


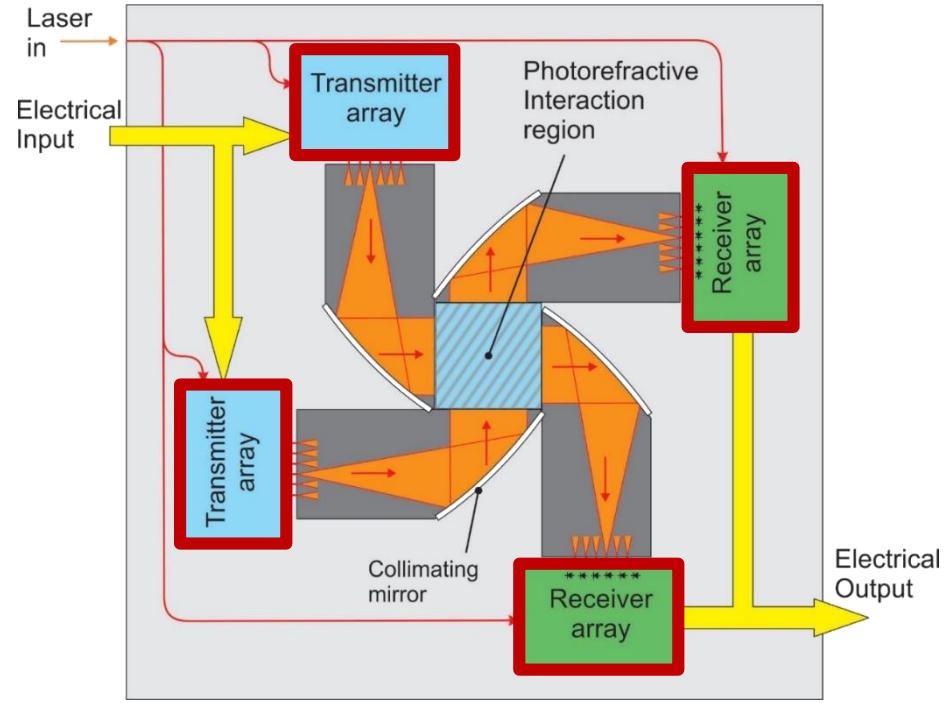
Photonic weight processing unit: Building blocks

Transmitter array:

- Encodes input vectors onto arrays of coherent optical sources.
- Control of amplitude and phase
- Based on standard Si-Photonics components

Si-Photonics hardware:





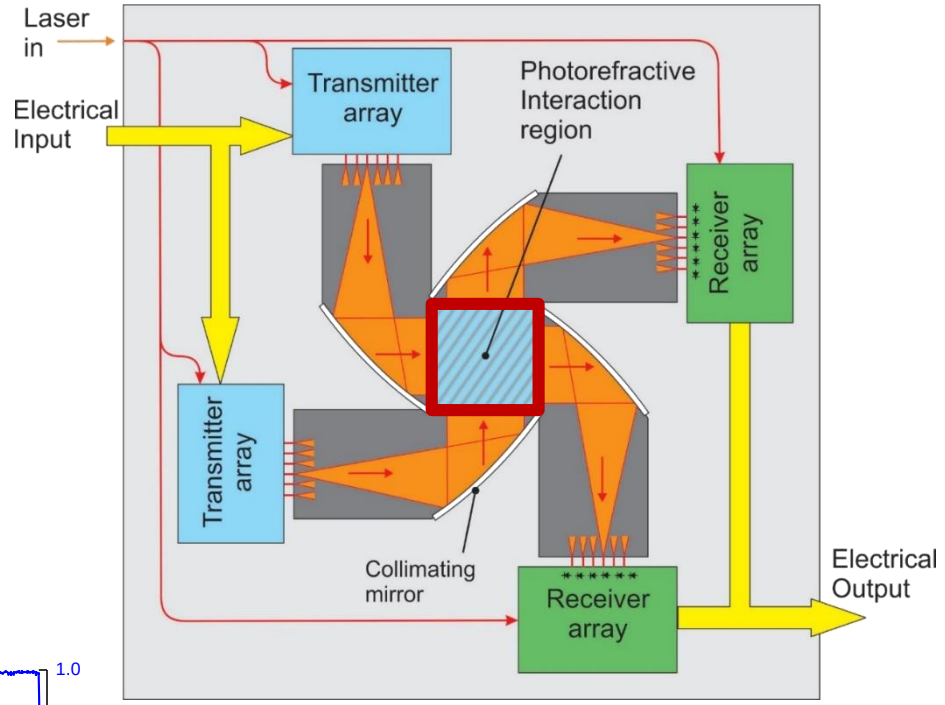
Receiver array:

- Detects amplitude and phase of output signals
- Standard Si-Photonics detector array

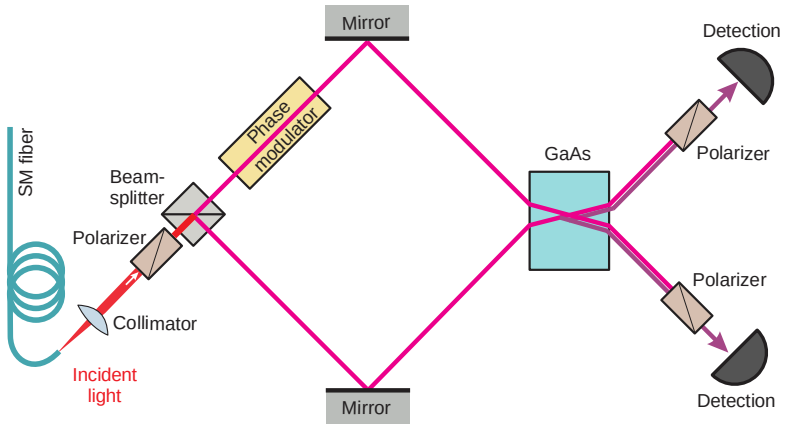
Photonic weight processing unit: Building blocks

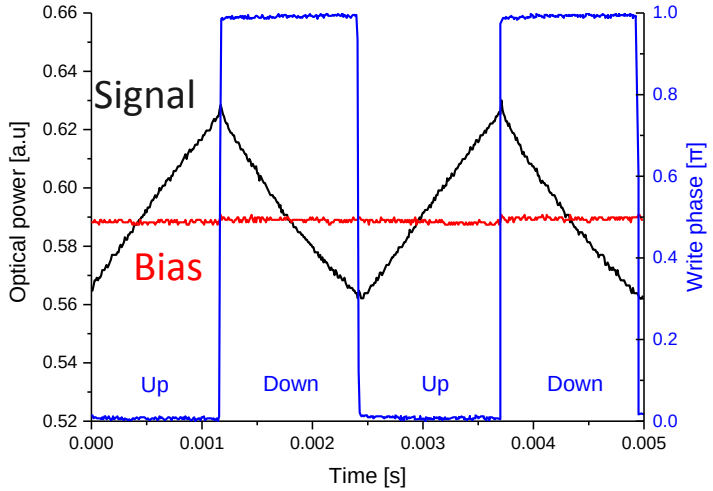
Photorefractive interaction region:

- Stores synaptic weights as refractive index gratings
- Photorefractive material: Semi-Insulating GaAs
  - Matches Si-Photonics wavelength range
  - Compatible with III-V on Silicon processes



Two-wave mixing in bulk GaAs crystal ≈ single synapse:



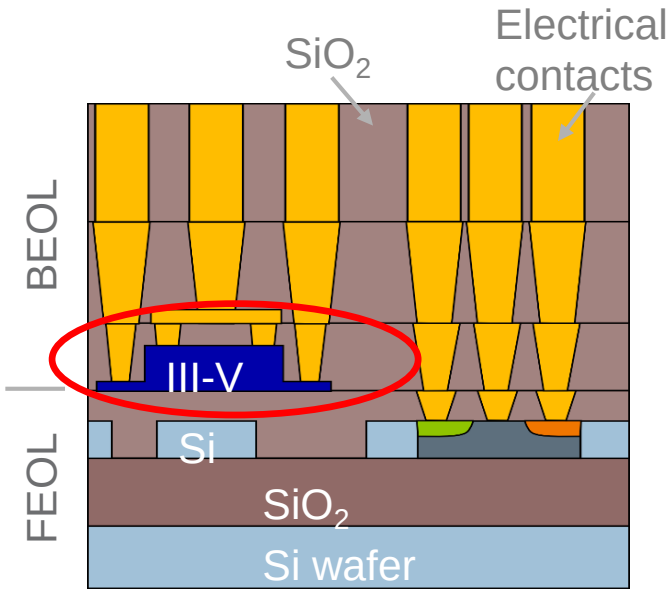


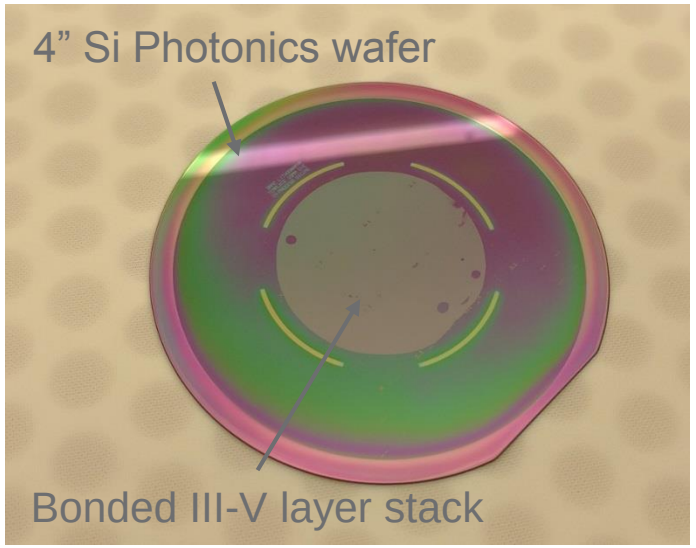
- Low asymmetry
- High dynamic range
- Bipolar weight storage
- To be confirmed in thin film

Photonic weight processing unit: Building blocks

Integration of the photorefractive layer:

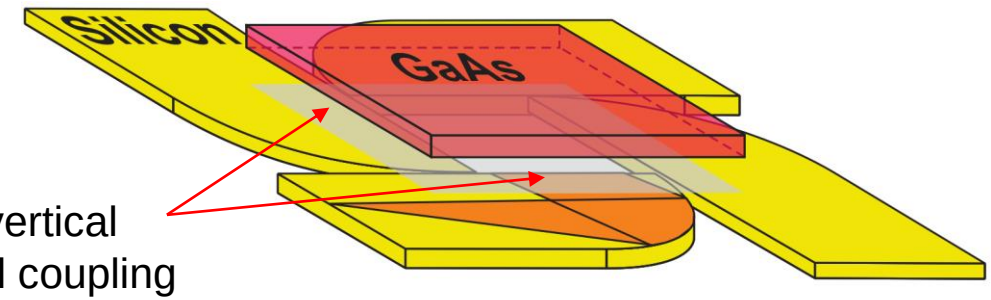
- Bonding technology as demonstrated for other III-V on Si projects:
  - Gain layers for integrated light sources
- Oxide bonding to Silicon-photonics stack



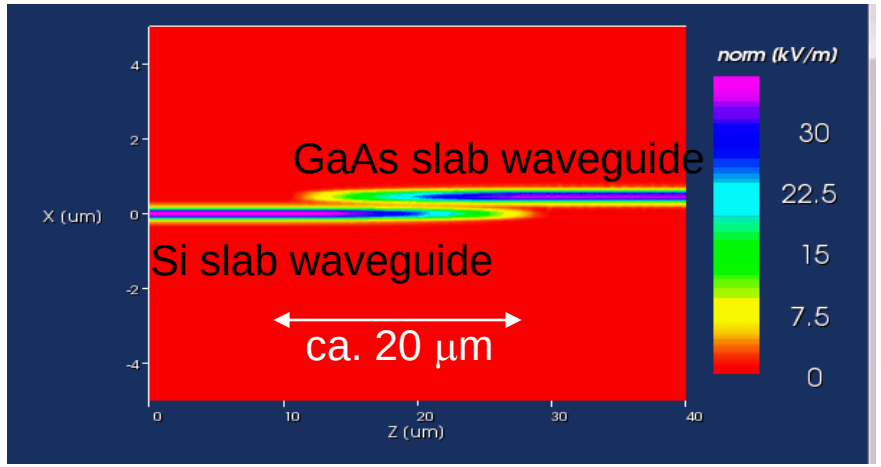


M. Seifried *et al*., "Monolithically Integrated CMOS-Compatible III–V on Silicon Lasers“ doi: 10.1109/JSTQE.2018.2832654.

- Vertical directional coupling for efficient coupling of light between Si-photonic and GaAs layers

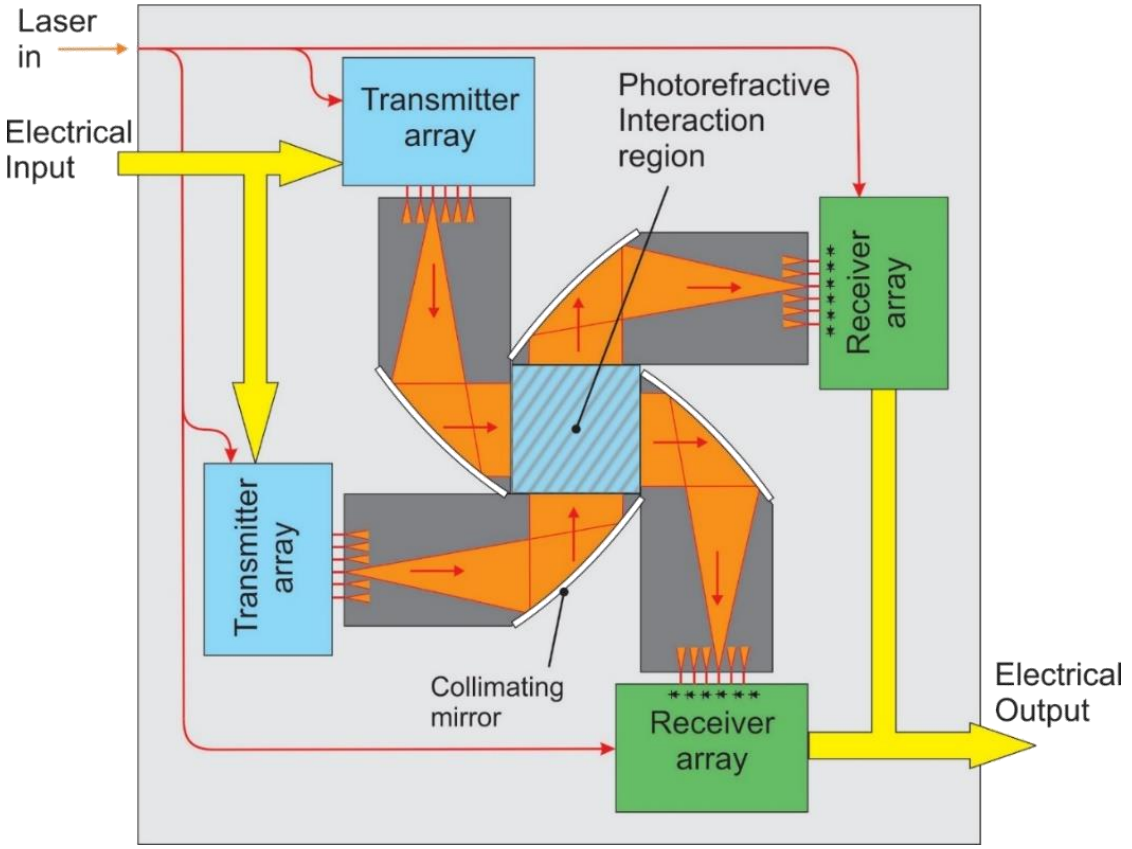


Vertical directional coupling



Summary and Outlook

- **Optical holographic storage and signal processing:**
  - Provides all necessary operations for accelerating training and evaluation of Deep Artificial Neural Networks
- **Integrated photonic synaptic weight processor:**
  - Silicon Photonics for electro-optical conversion and beam shaping
  - GaAs photorefractive layer for holographic weight storage and processing
- **First step: Demonstration of principle**
  - 8 x 8 matrix using in-house facilities
  - BRNC cleanroom @ IBM - Zurich
- **Next: Large scale demonstrator**
  - Si-Photonics foundry
  - III-V integration support required



Acknowledgements

IBM Research – Zurich, Switzerland
Roger Dangel, Yannick Baumgartner, Bert Offrein, Stefan Abel, Marc Seifried, Gustavo Villares, Felix Eltes, Jean Fompeyrine, D. Caimi, L. Czornomaz, M. Sousa, H. Siegwart, C. Caer, D. Jubin, N. Meier, A. La Porta, J. Weiss, U. Drechsler

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