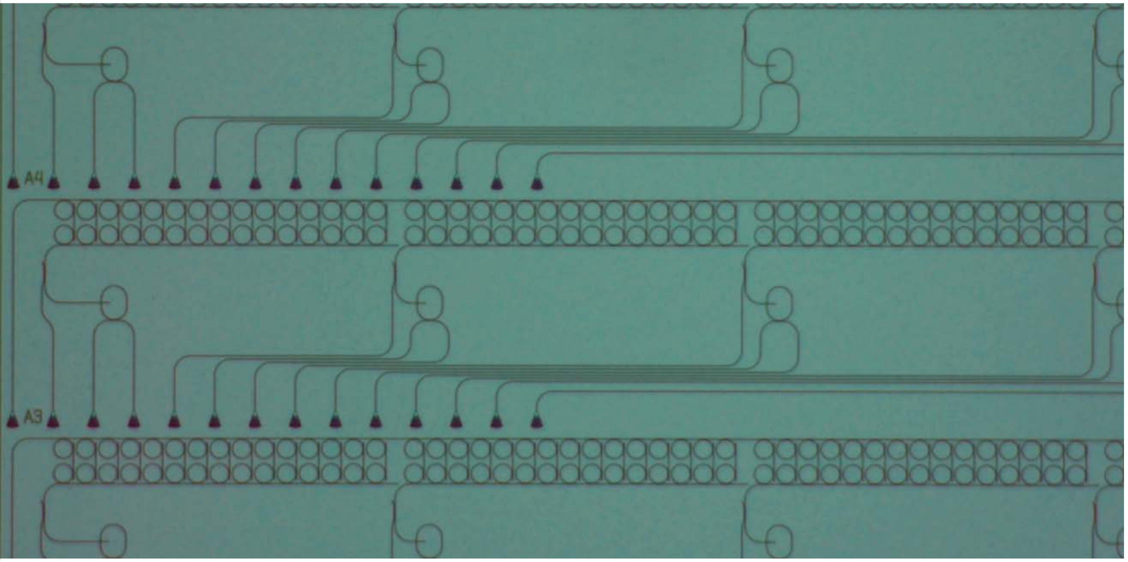
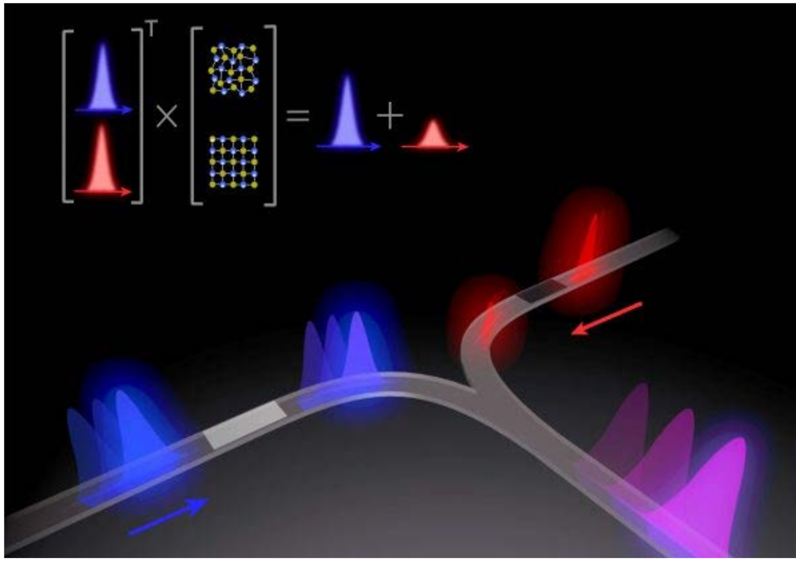
Ultrafast photonic convolution processing

Wolfram Pernice
http://www.uni-muenster.de/Physik.PI/Pernice/

Universität Münster (WWU), Physikalisches Institut

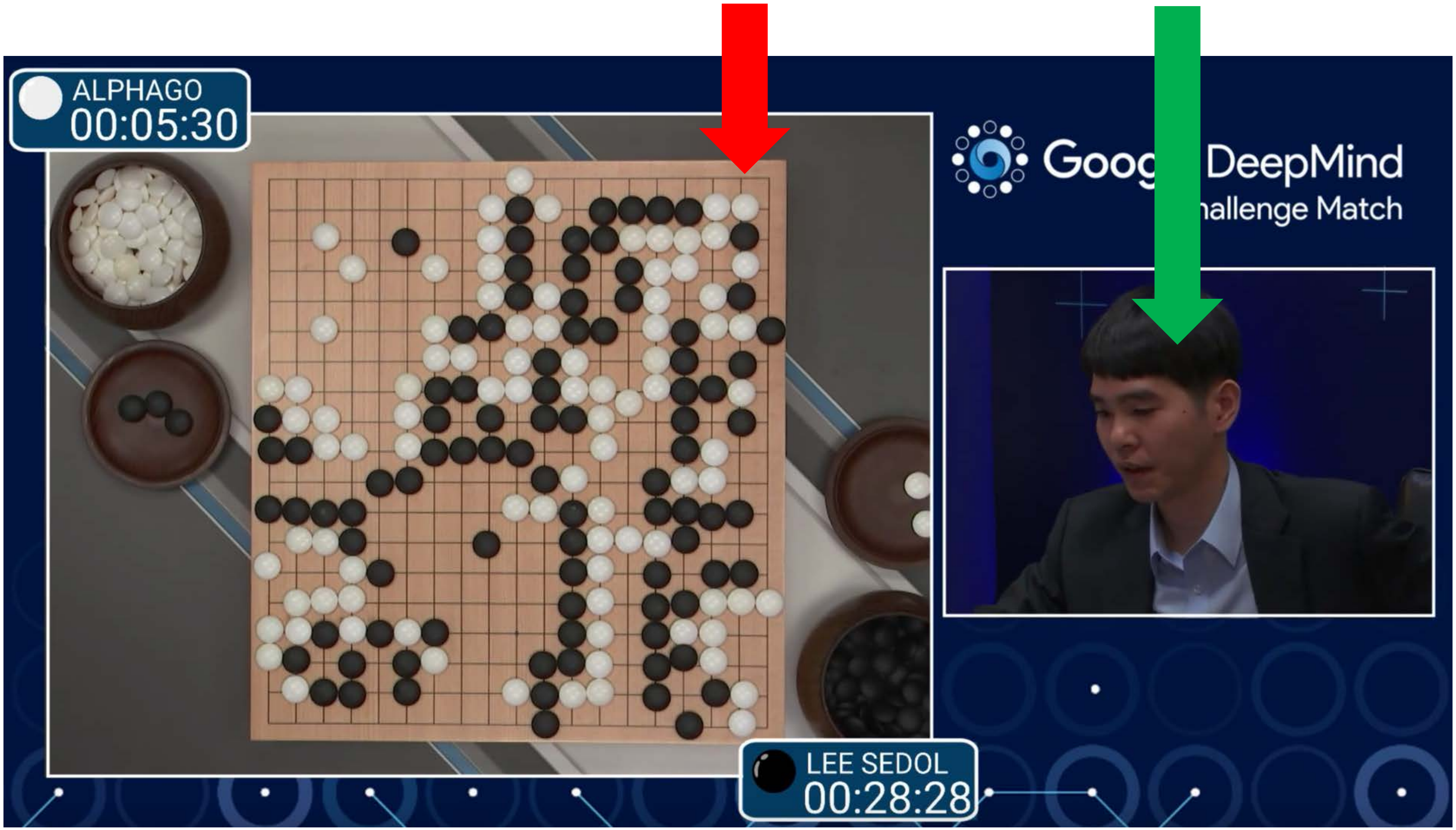




AlphaGo (2016)

~1,000,000 W

~20 W

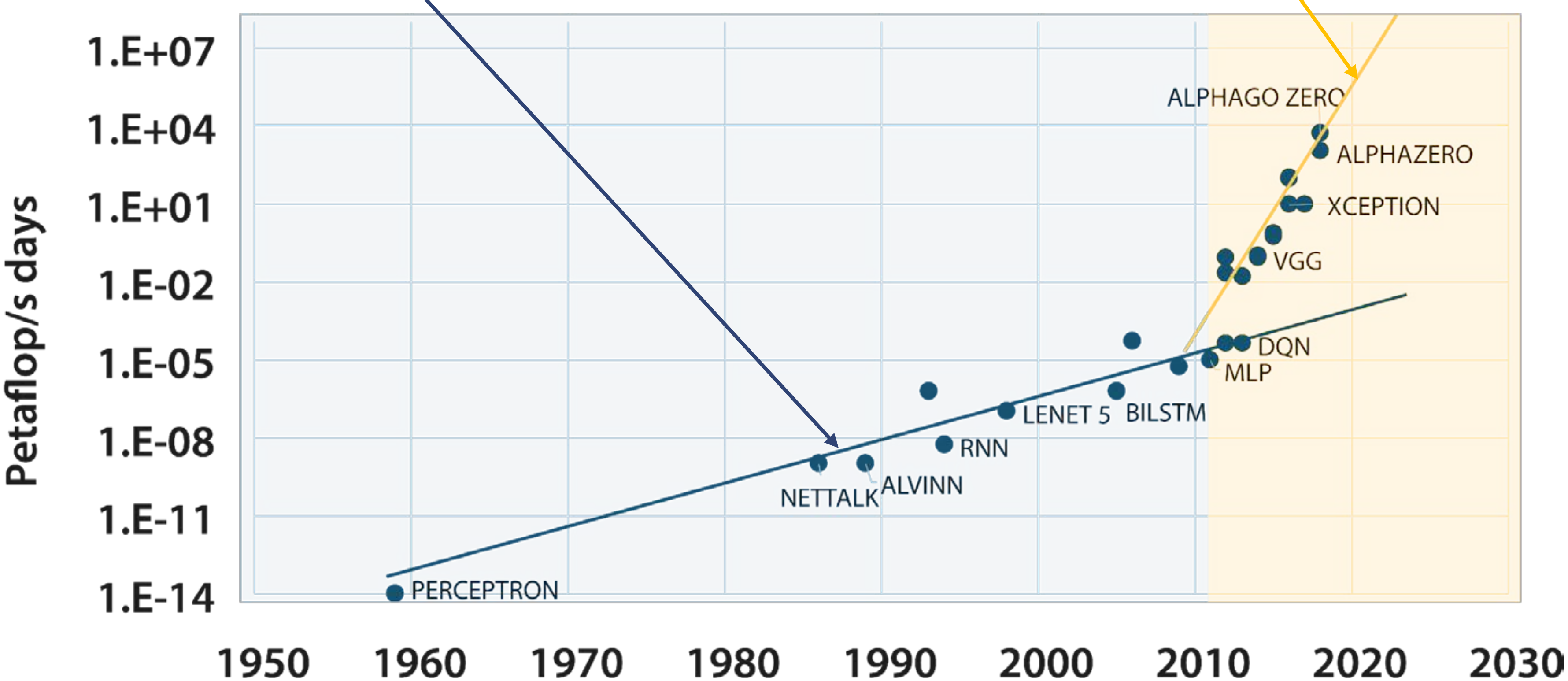


- 1202 Central Processing Units (CPUs)
- 176 Graphics Processing Units (GPUs)

Moore‘s law – revisited

Before: Processing power doubles every 2 years

Now: Processing power doubles every 3.5 months

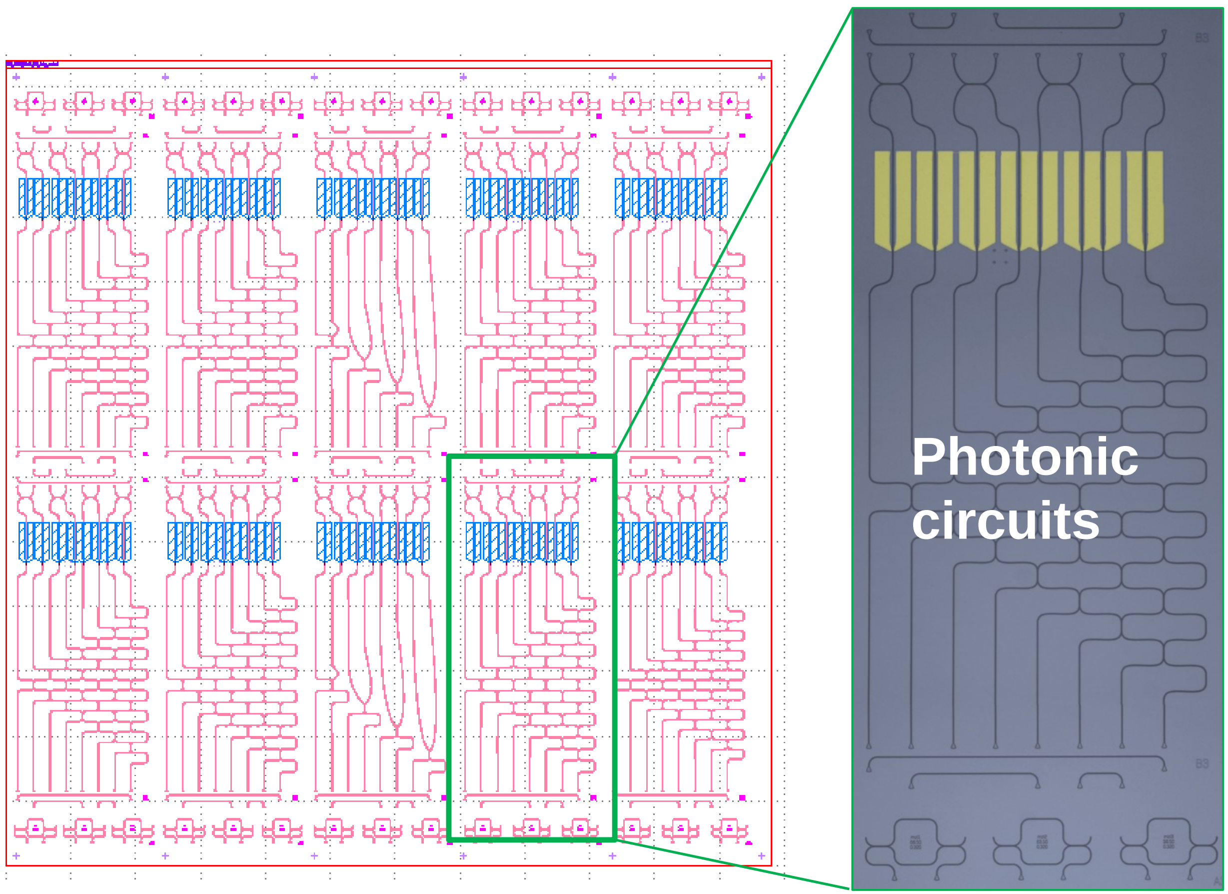


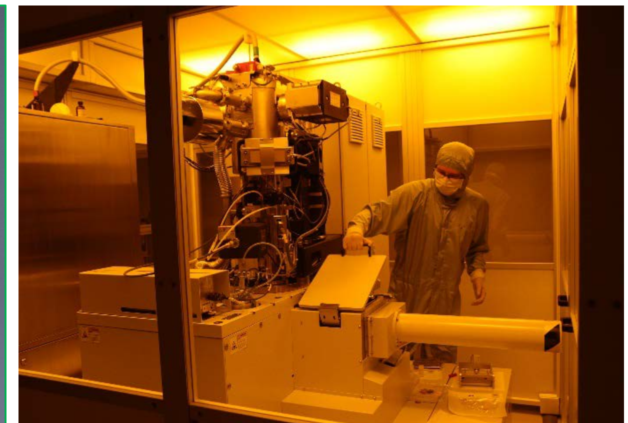
Nanophotonic circuits @ WWU

- Integrated photonic components
- Multiple layers of lithography using alignment
- Photonic CAD with Python framework

MÜNSTER NANOFABRICATION FACILITY

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

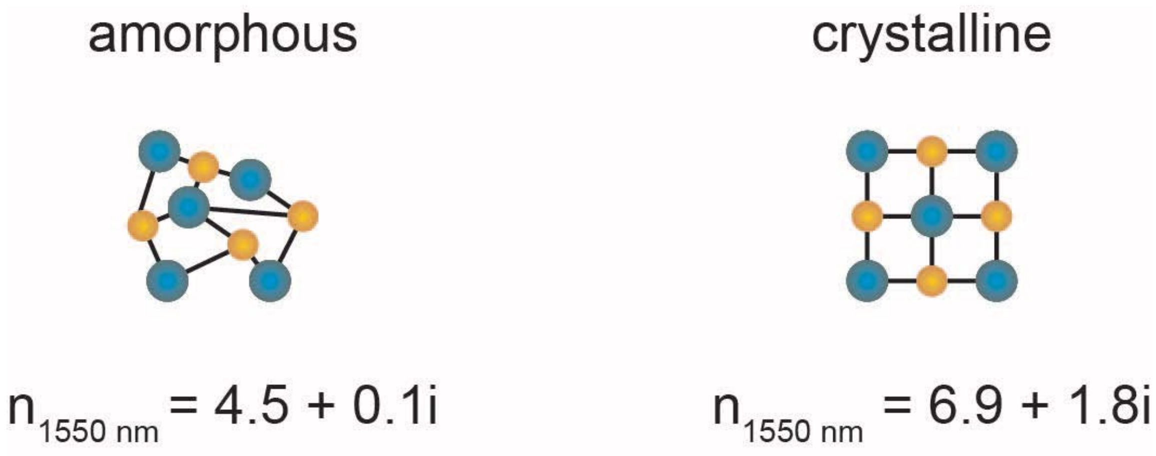




Gehring, et al., OSA Contin. 2, 3091 (2019)

Phase-change photonics

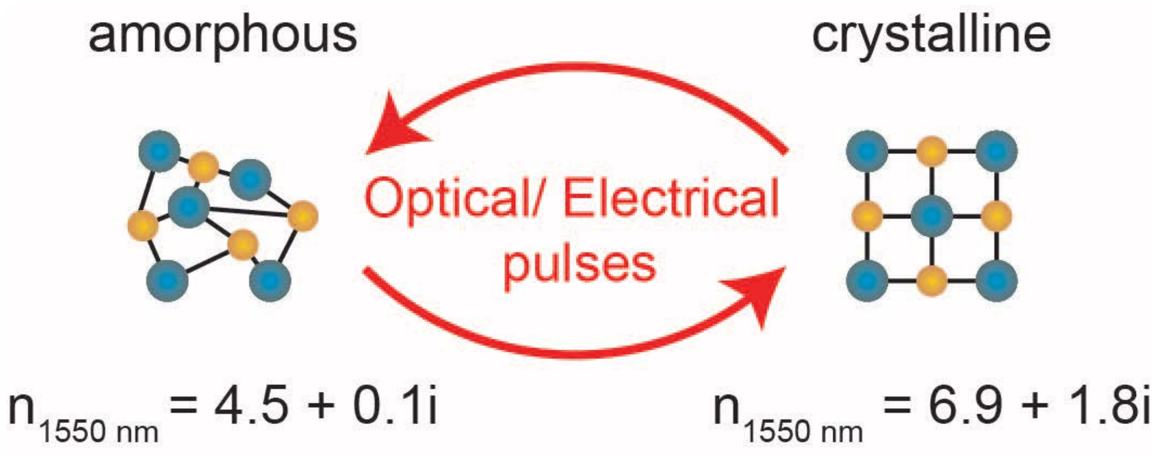
- Add active elements to passive waveguides
- Implement synapses and neuron soma with phase change materials (PCMs)



$Ge_2Sb_2Te_5$ (GST)

Phase-change photonics

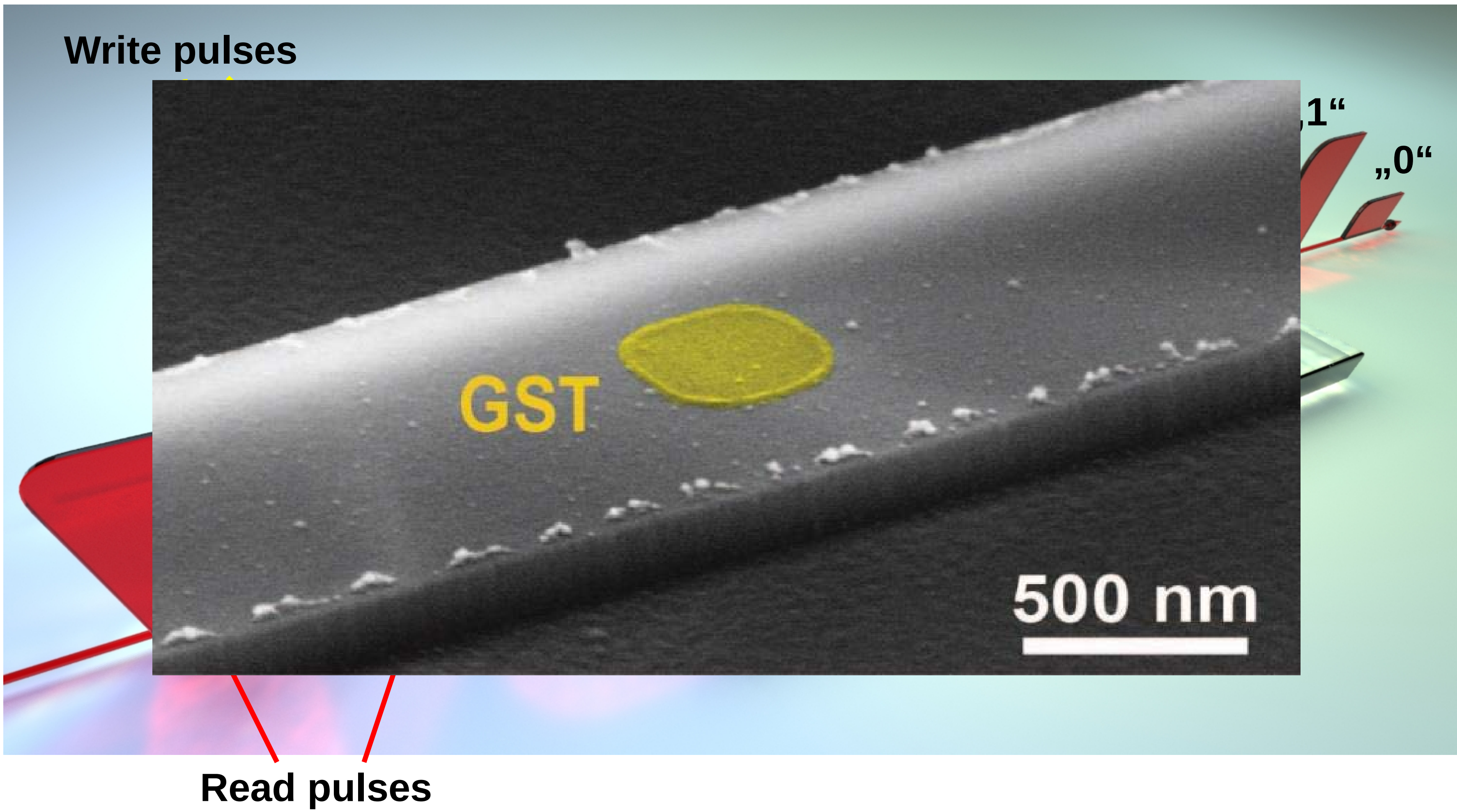
- Add active elements to passive waveguides
- Implement synapses and neuron soma with phase change materials (PCMs)
- All-optical reconfiguration within sub-nanoseconds



$Ge_2Sb_2Te_5$ (GST)

PCM nanophotonic devices

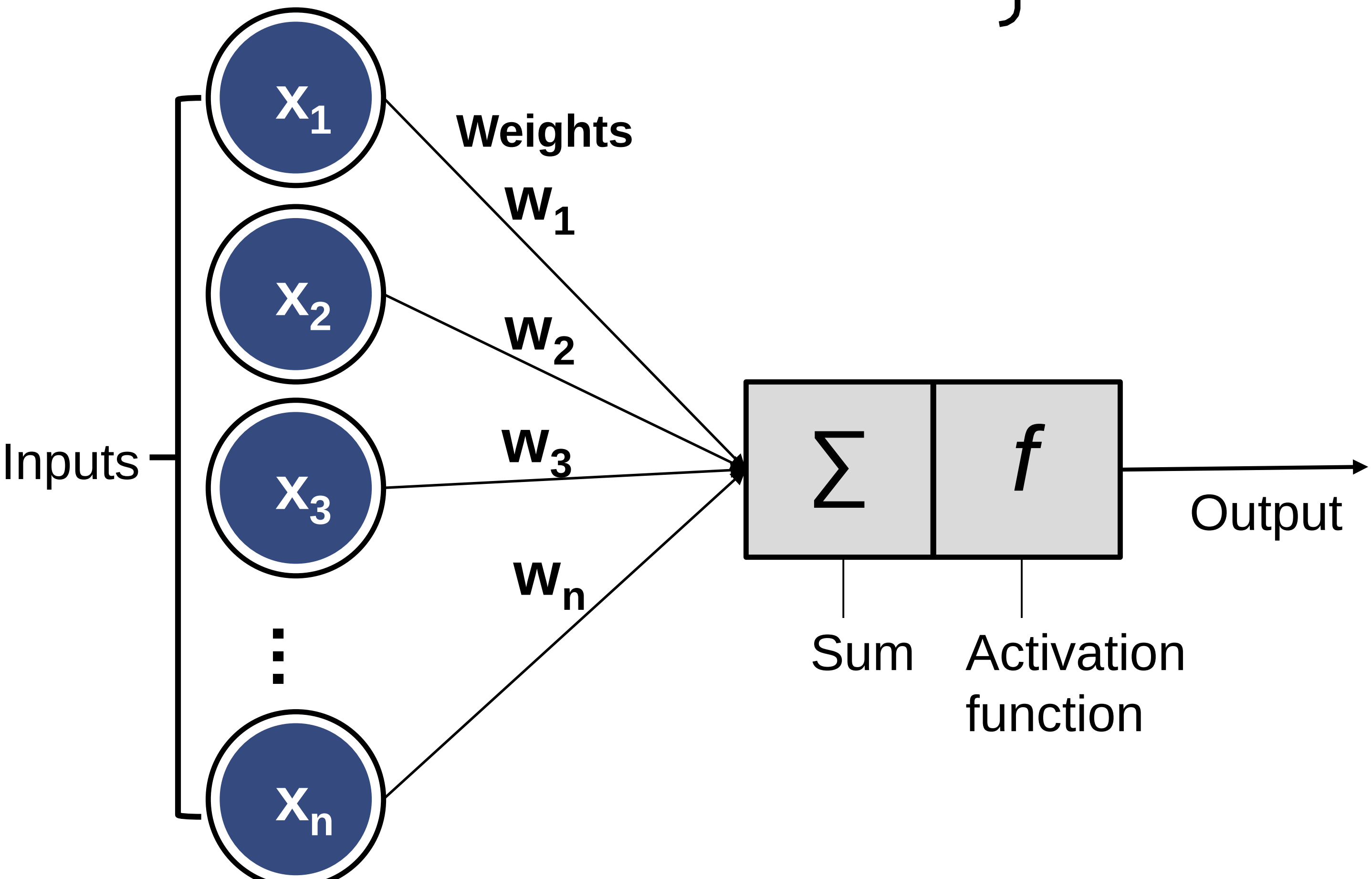
- Place PCM in near-field of optical waveguide
- Data is encoded in the amount of transmitted power



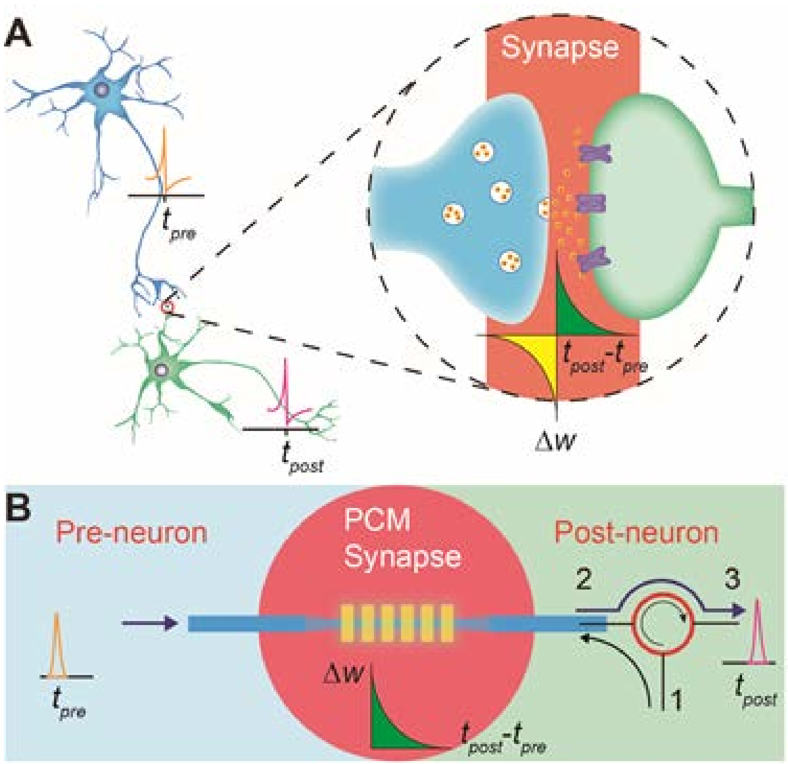
Photonic neurons

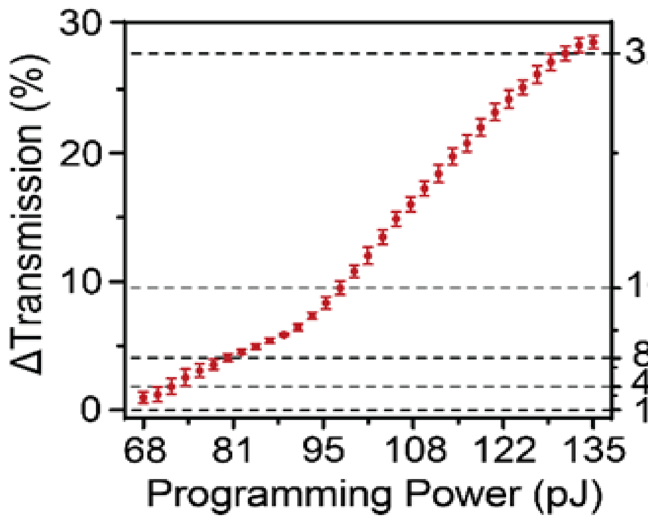
- Use for matrix multiplication:
  - Multiplication
  - Addition

  } Multiply-accumulate (MAC)



A photonic synapse (the weights)

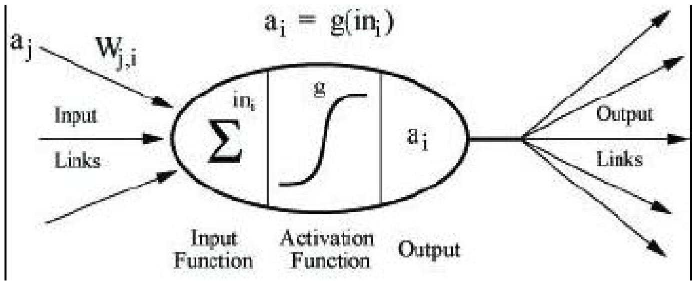


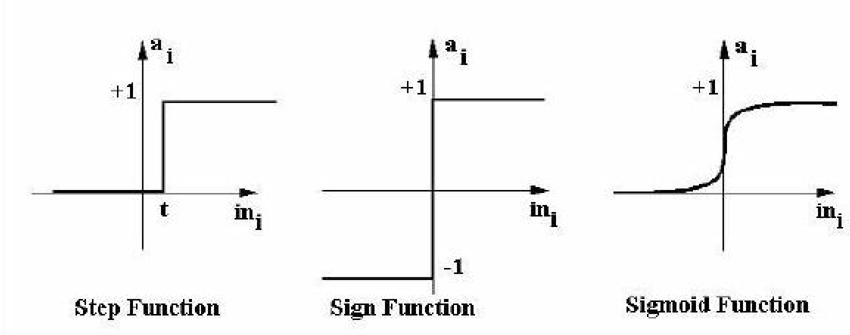


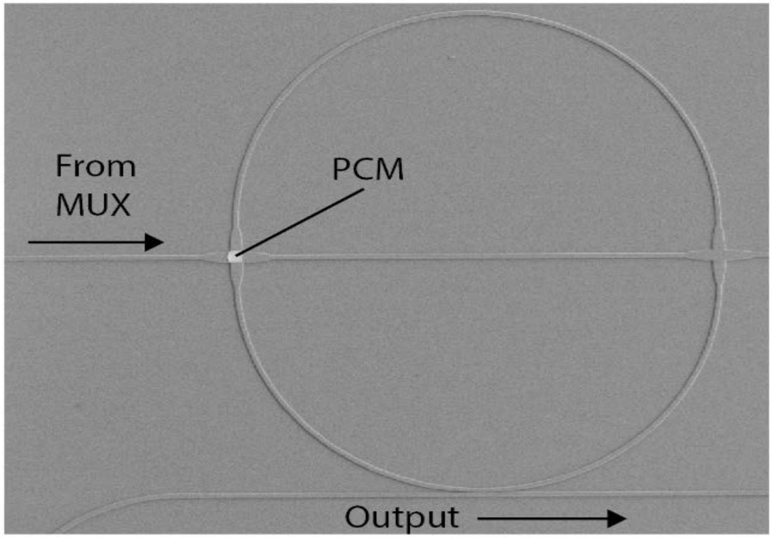
- Partial crystallization allows storage of multiple bits per cell
- Number of levels depends on optical contrast and noise performance
- Wheights are stored permanently in crystal state of PCM

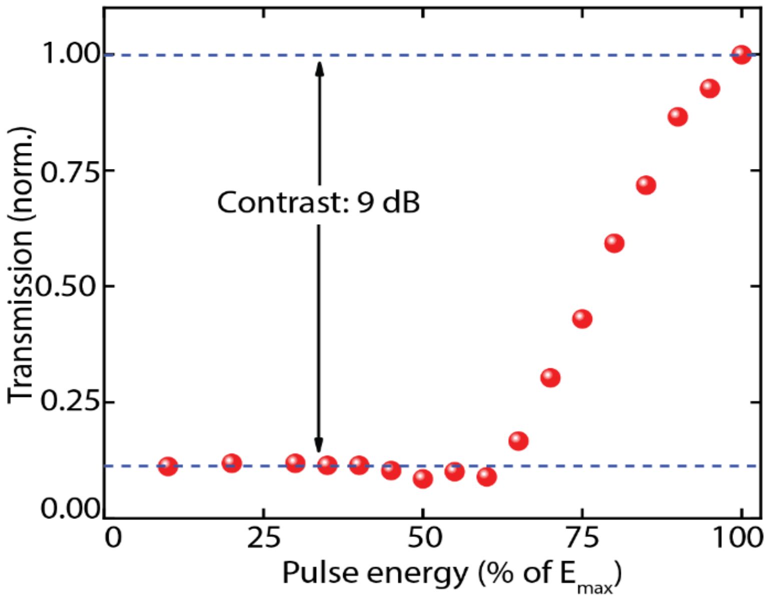
Rios et al., Science Advances 5, eaau5759 (2019)
Li, et al., Optica 6, 1, (2018)
Cheng, et al., Science Advances 3 e1700160 (2017)

Implementation of threshold function (the „soma“)



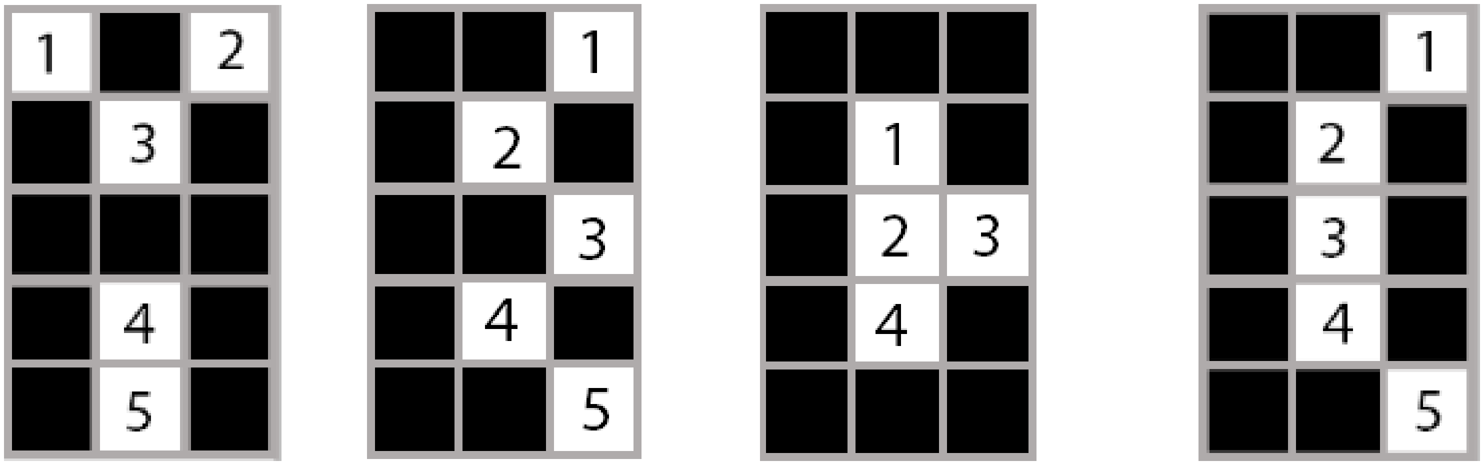


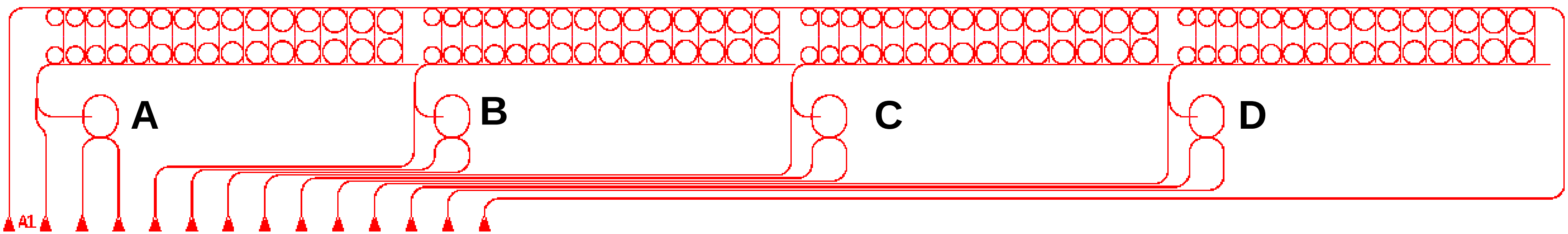




- Threshold function is provide by ring resonator
- Resonance tuning with embedded PCM element

A small-scale ANN

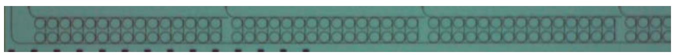


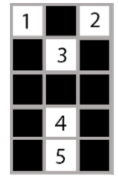


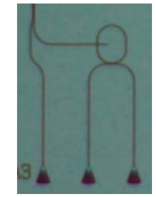
- 15 input neurons and 4 output neurons
- Each letter is pixelized into 15 digital elements

Feldmann et al., Nature 569, 208 (2019)

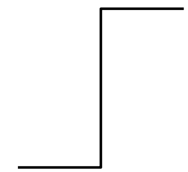
A closer look at the phontonic ANN







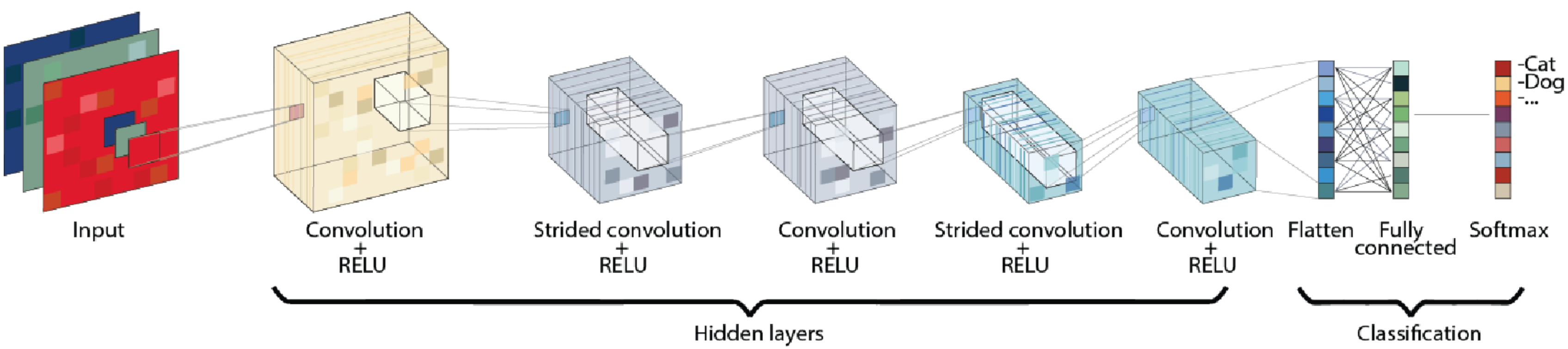
$$\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \times \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} \times$$



Synaptic weights

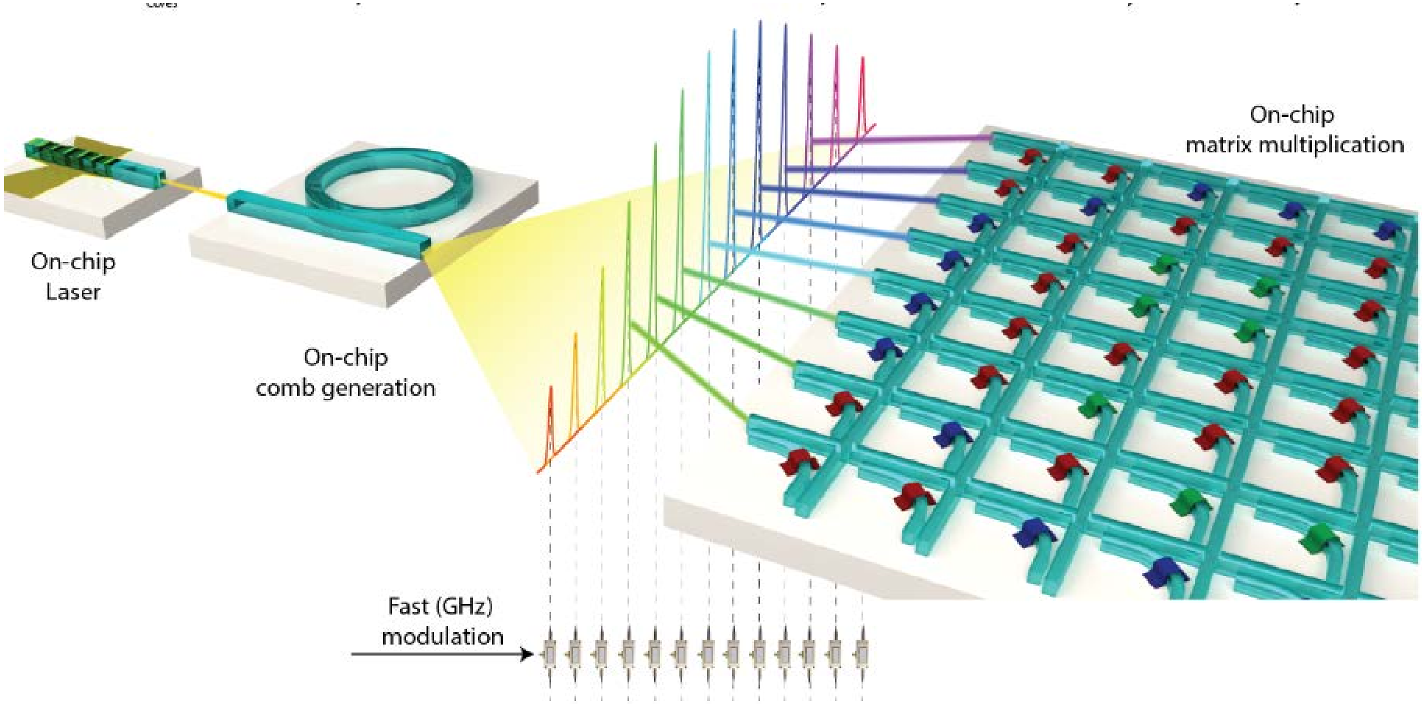
Input-Vector

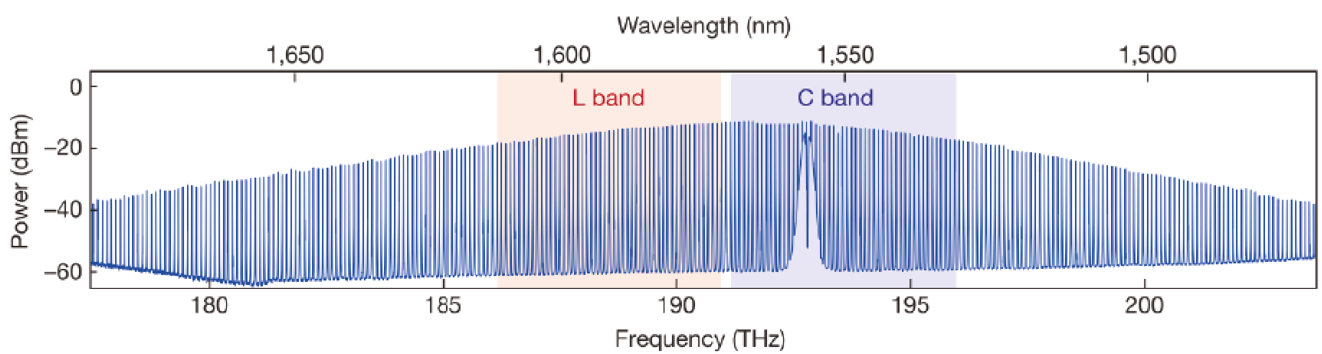
Rectification



Convolutional neural networks

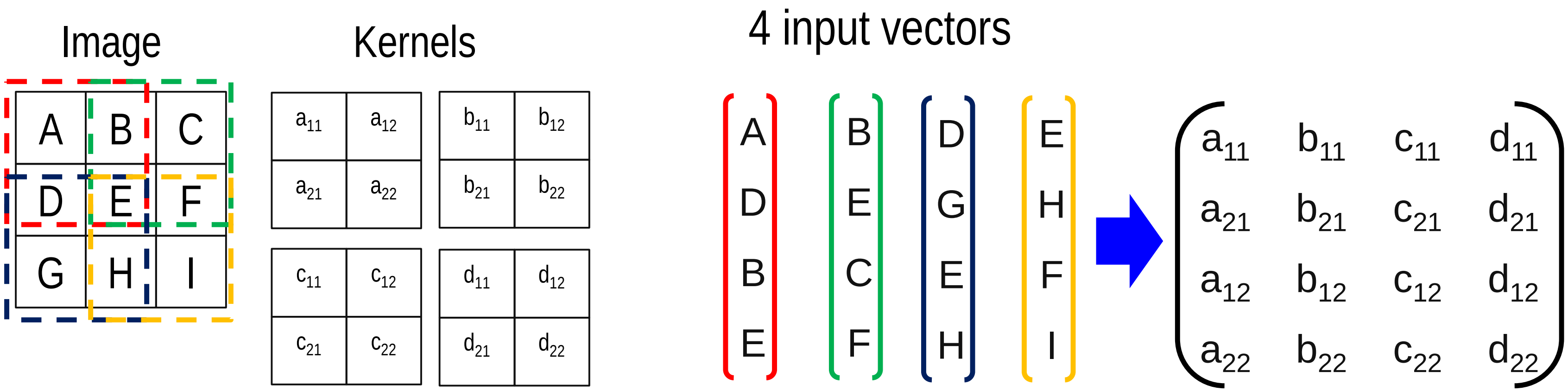
Ultrafast convolution processing





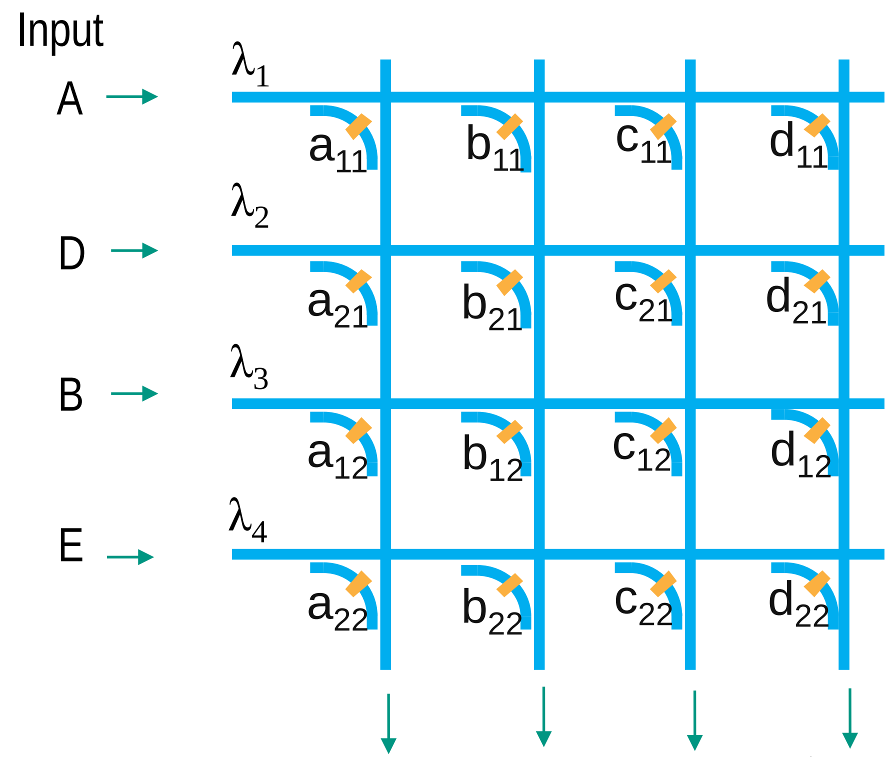
Frequency comb, Kippenberg group (EPFL)

Example: 3X3 pixel image and 4 kernels of 2X2 dimension



Example: 3X3 pixel image and 4 kernels of 2X2 dimension

- Using waveguide crossing matrix on the right
- Splitting ratios (directional couplers) adjusted for equal power distribution
- Each input on different wavelength (avoid interference)
- Each column gives convolution output for one of the kernels

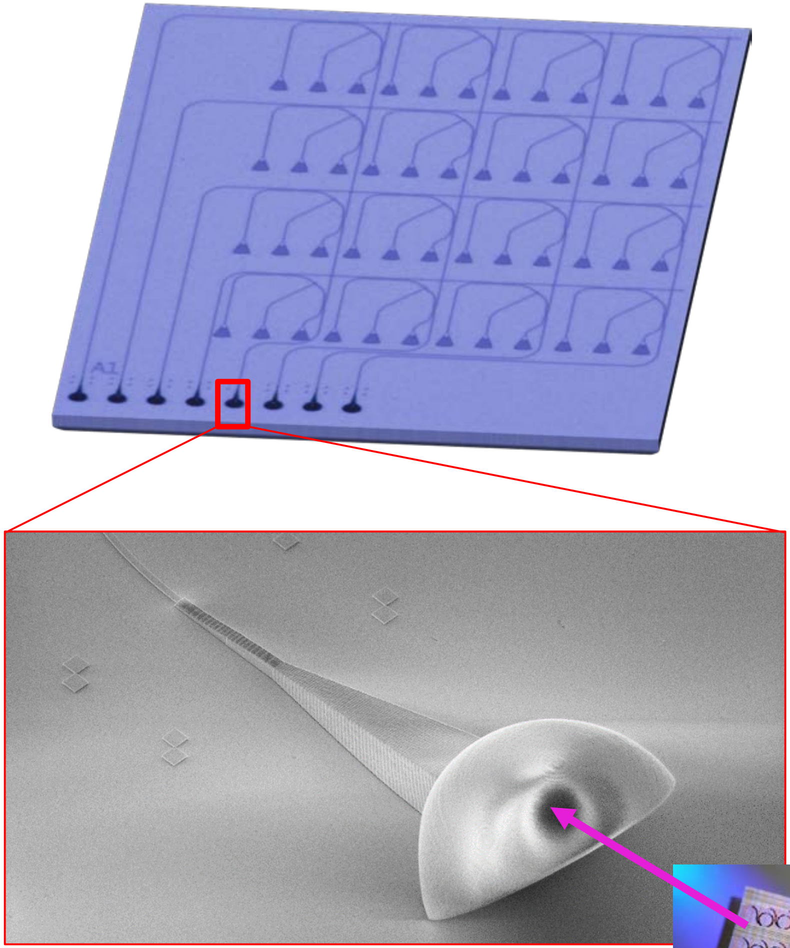


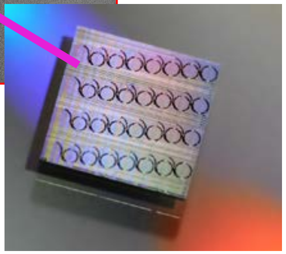
$\frac{1}{16}(Aa_{11}+Da_{21}+Ba_{12}+Ea_{22}) = \frac{1}{16}X_1$ $\quad \frac{1}{16}X_2$ $\quad \frac{1}{16}X_3$ $\quad \frac{1}{16}X_4$

Use WDM to perform multiple convolutions at the same time

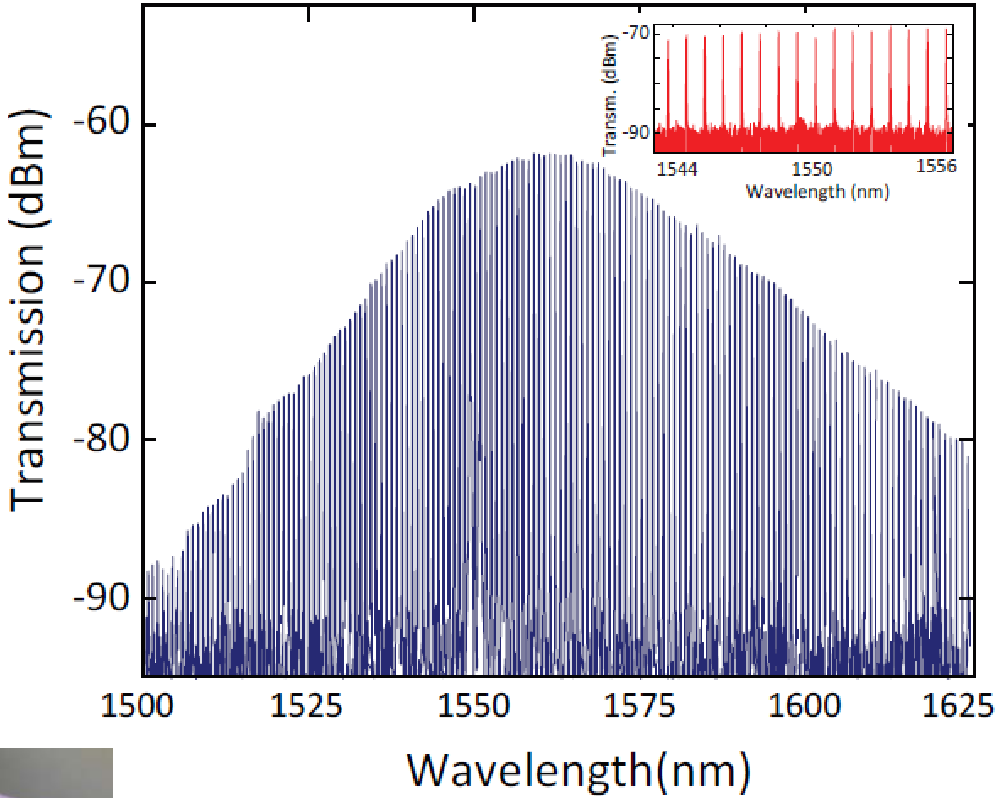
Ultrafast convolution processing

PCM Matrix chip



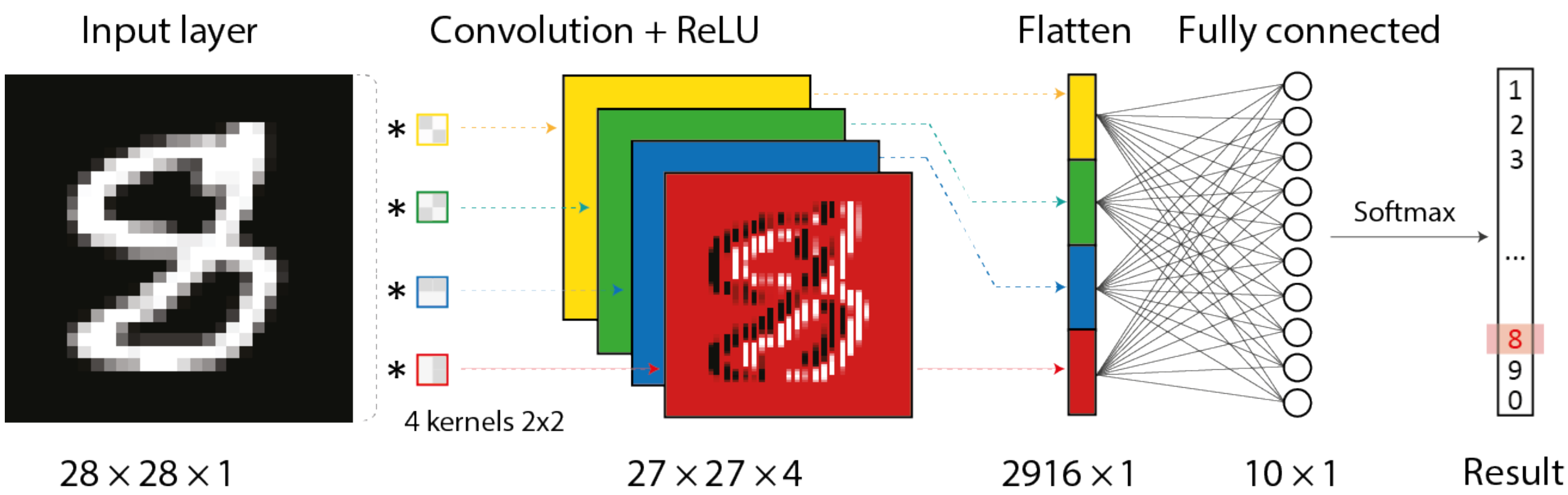


Comb input, EPFL



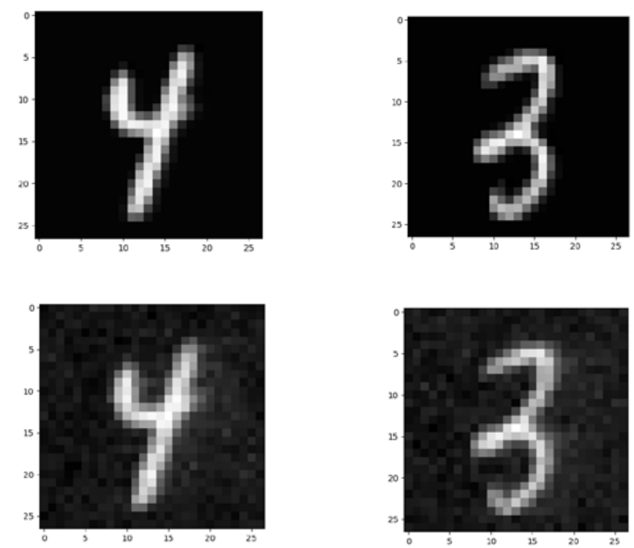
Feldmann et al., Nature 589, 52 (2021)

Full digit recognition with photonic NNs



~95% accurate

Prediction \ Label	0	1	2	3	4	5	6	7	8	9
0	97.6		0.6		0.6		2.3			
1		99.4	0.6					0.7		
2			93.0	0.6				4.0	1.4	
3		0.6	0.6	98.7		2.2		1.3	0.7	
4					95.8			0.7	1.4	2.1
5						93.4	0.8		1.4	1.4
6	1.6		1.7		1.2	0.7	95.4			
7			0.6					89.3	0.7	3.5
8	0.8		2.3	0.6	2.4	3.6	1.5	1.3	94.3	4.2
9			0.6					2.7		88.9

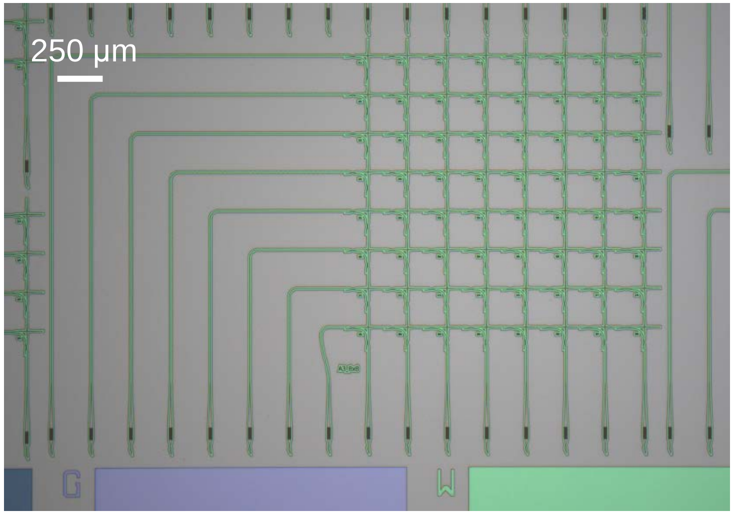


Feldmann et al., Nature 589, 52 (2021)

Why this is exciting

NVIDIA flagsip tensor processor (Tesla V100)

- 195 GigaFLOPS/core
- 1 FLOP ~ 0.5 multiply-accumulate (MAC) operations



PCM matrix processor

- **2 TeraMAC/s**

More than 1000 HDTV video streams in parallel

The people who really do the work:

<u>At WWU:</u>
C. Schuck and team

R. Bratschitsch, J. Kern, P. Tonndorf

J. Feldmann, N. Gruhler, A. Ovvyan, S. Ferrari, W. Hartmann, N. Walter, F. Beutel, M. Stappers, H. Gehring, C. Kaspar, F. Lenzini, T. Grottke, J. Lin, J, Schütte, E. Lomonte, R. Stegmüller, Y Liu

<u>At Oxford:</u>
N. Youngblood
H. Bhaskaran
X. Li

<u>At Exeter:</u>
D. Wright
E. Gemo
S. Garcia-Cuevas Carrillo

<u>At EPFL:</u>
T. Kippenberg
M. Karpov

<u>At IBM:</u>
A. Sebastian

DFG erc VolkswagenStiftung QUANTERA European Commission HORIZON 2020