Computing goes Light:
Effiziente Datenverarbeitung auf Basis der Photonik

ADV TRENDS Gamechanger IT

Nov 30 2023

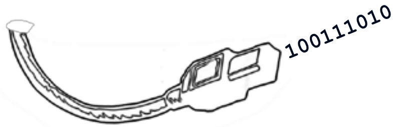
Dr. Bernhard Schrenk

AIT Austrian Institute of Technology

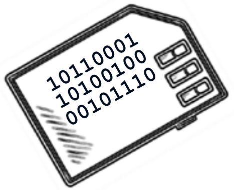
AUSTRIAN INSTITUTE OF TECHNOLOGY

Computing goes Light:

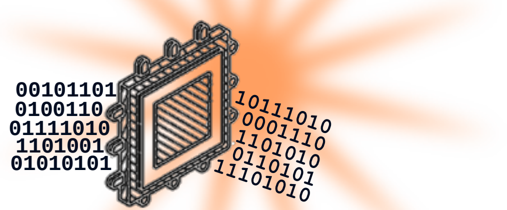
Effiziente Datenverarbeitung auf Basis der Photonik



Communication:
Information in Motion



Storage:
Information at Rest



Processing:
Information Transformation

The Internet Explosion

1987

2 TB

1997

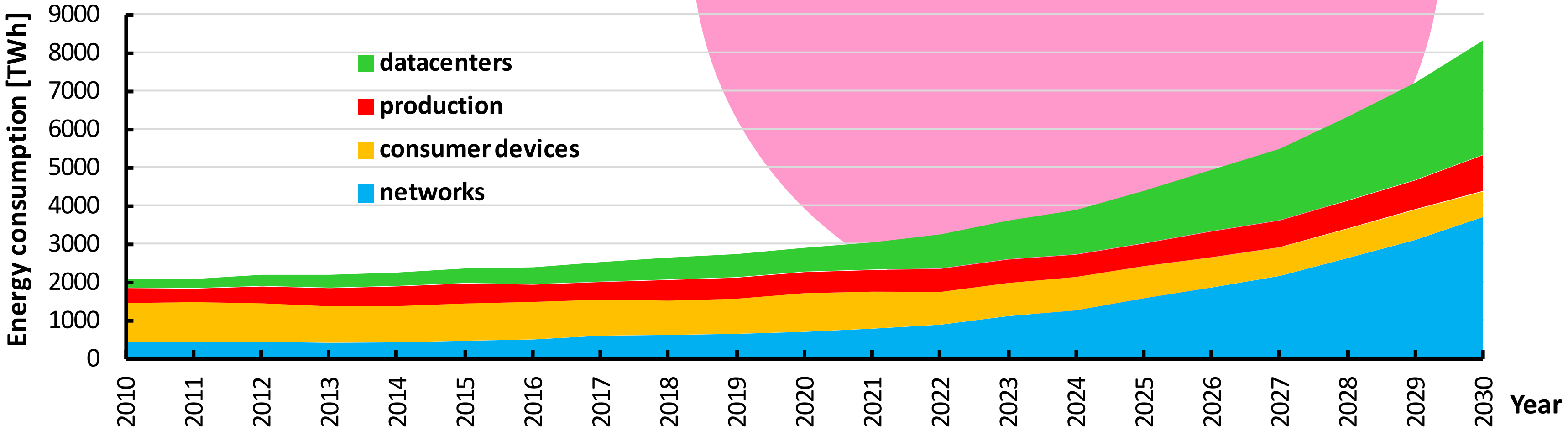
60 PB

2007

50 EB

2017

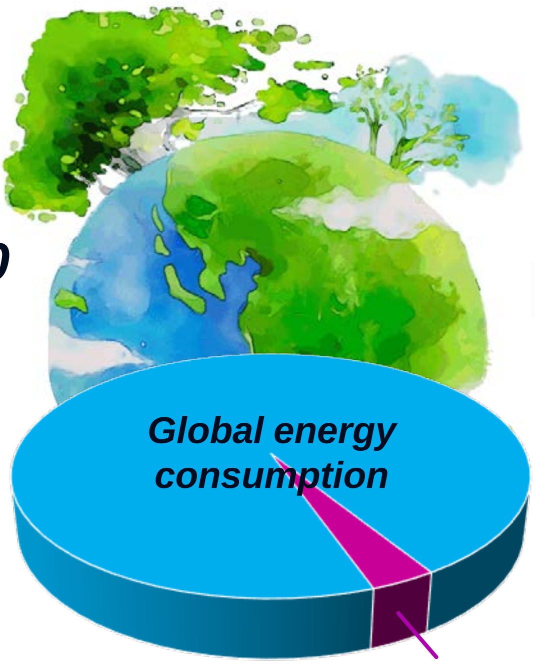
1.1 ZB



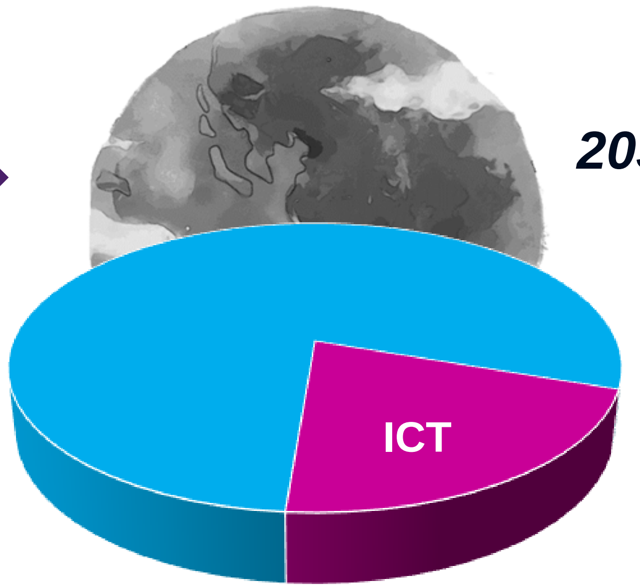
N. Jones, Nature 561, 163 (2018)

Share of ICT in Global Energy Consumption

2020



T + 10 years



2030

ICT: 3 - 4%

~ 24 nuclear reactors

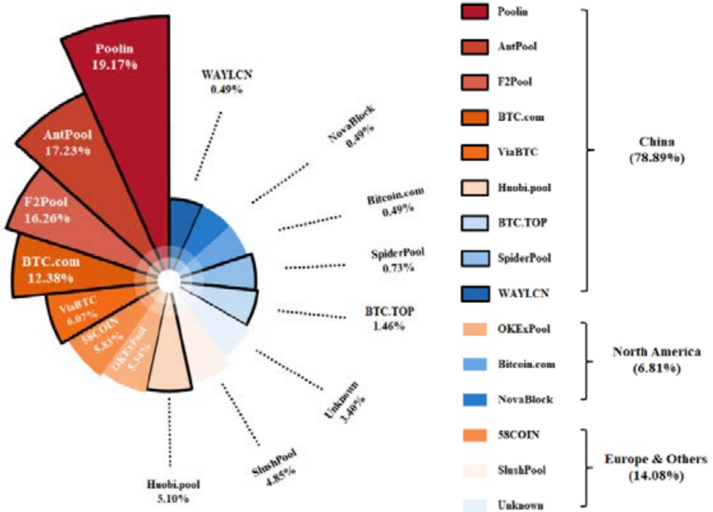
~ aviation industry

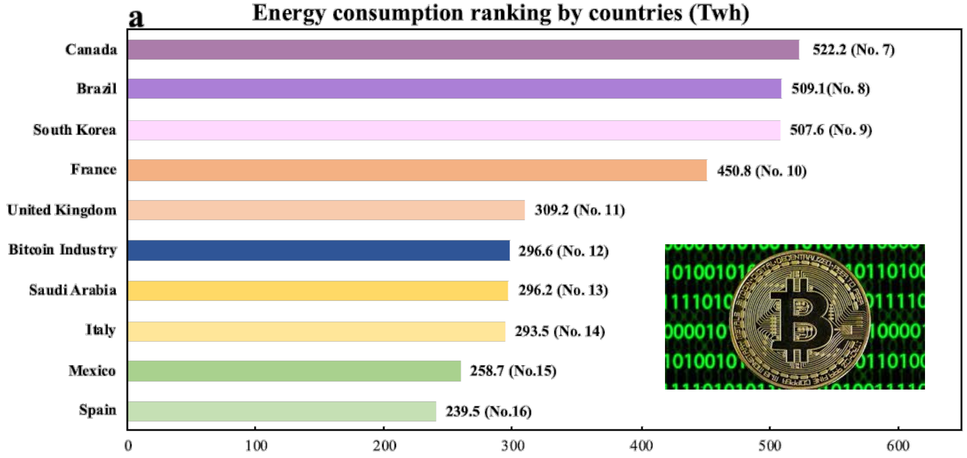
worst case: ~21%
best case: ~8%

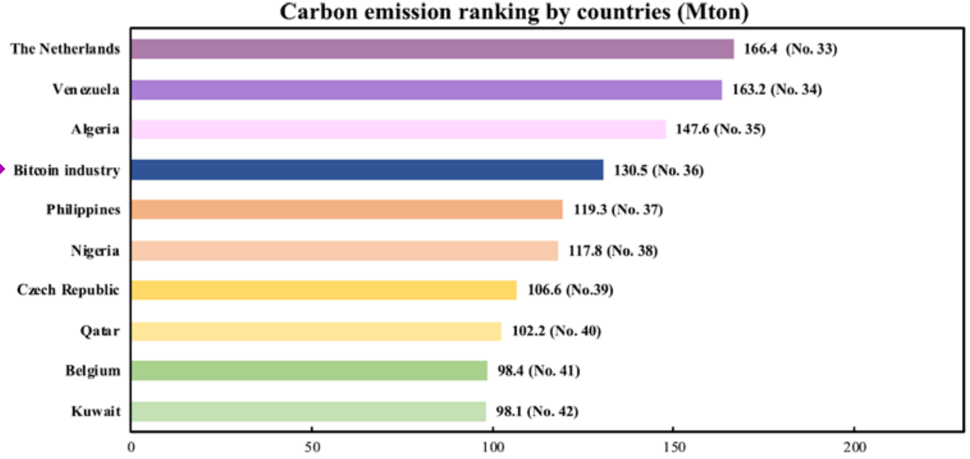
N. Jones, Nature 561, 163 (2018)

Example: Bitcoin Mining

- 80% happening in China
  - 40% thereof fuelled by coal
- equivalent to 0.6% of world's electricity production
- similar footprint as Italy or Saudi Arabia



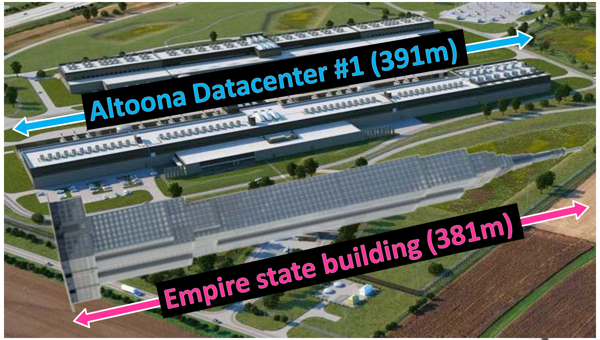




S. Jiang, Nature Comm. 12, 1938 (2021)

Processing: Inside the Information Factory

Cloud Datacenter

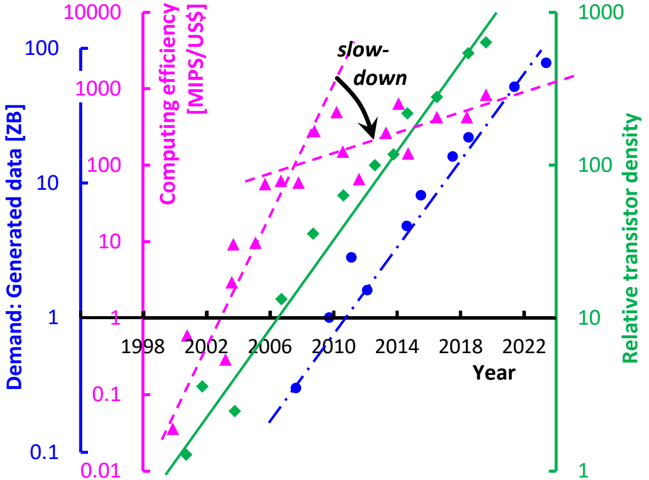


~20 000 servers
~20 MW

Operating a datacenter: installing a server blade in 2030.

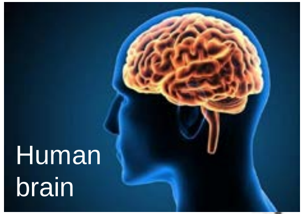
HPC

537 212 TFlop/s
29.9 MW

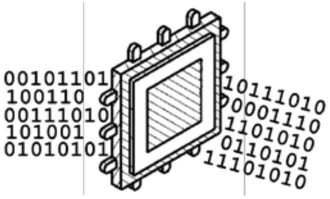


J. Kendall, Appl. Phys. Rev. 7, 011305 (2020)

Human



~2 000 TFlop/s
0.000020 MW



$$\frac{\textbf{Compute ops}}{\textbf{el. Power}}$$

Pattern Recognition

Msot plpeoe wlil hvae no peormbls radneig tihs txet, alothguh the oderr of leterts is rndaom (wtih the epeixoctn of the frist and the lsat leettr).

- There is only 1 correct solution and ~ 121 885 070 000 000 000 000 000 possibilities.
- We compute on-the-fly as we read over the text – a fantastic example of pattern recognition.

AI Hardware

- Multi-layered, deep **neural network**
  - accomodates many **neurons**

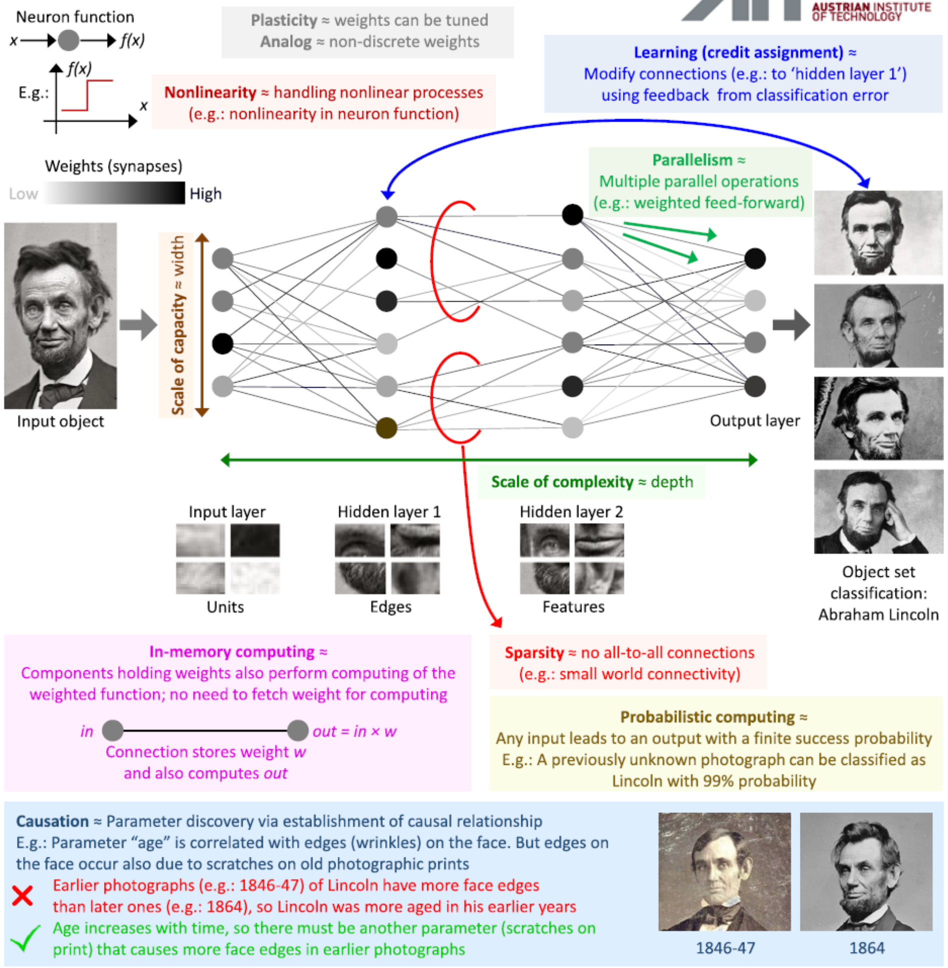
    Human brain: 10^{11}

    Intel Loihi: 130,000

- Weighted synaptic **interconnect**
  - dense vector-matrix multiplications
  - routing becomes challenging when scaling up the data movement

    Human brain: 10^4 inputs/neuron

- Each layer needs to be **trained** …
- … to yield **time-of-flight inference**

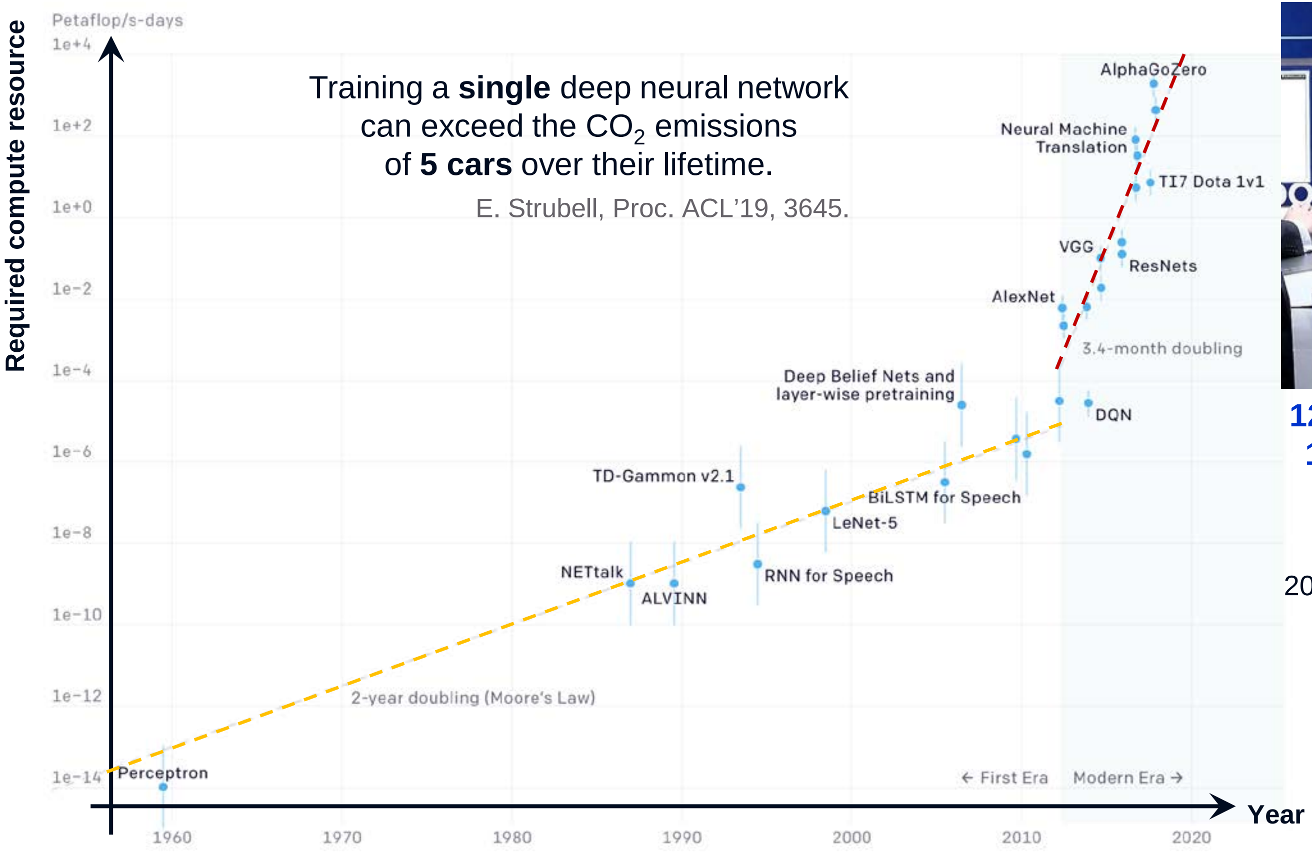


J. Kendall, Appl. Phys. Rev. 7, 011305 (2020)

First: Training the AI

Training a **single** deep neural network can exceed the CO_2 emissions of **5 cars** over their lifetime.

E. Strubell, Proc. ACL'19, 3645.



https://openai.com/blog/ai-and-compute/

2016: AlphaGo defeated Lee Sedol

1202 CPUs
176 GPUs
vs **1 human brain**

1 MW vs **20 W**

2023: Kellin Pelrine defeated Go-playing AI

Second: Enjoy Inference at Low Latency

Speech
and object
recognition

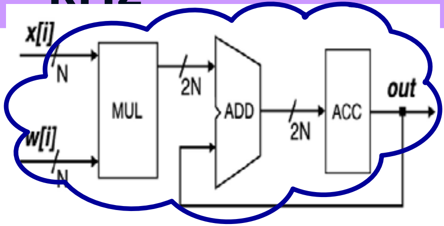
Deep surveillance in real-time

Accelerators / co-processors
for vector-matrix multiplication
and deep learning inference,

ultra-fast **control**,

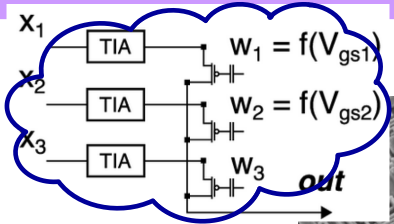
intelligent **signal processing**
(wireless, fiber comms, edge computing)

Hz – kHz

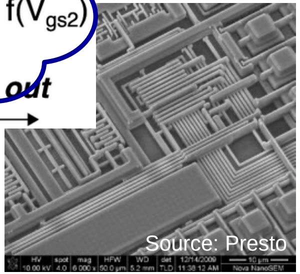


Machine learning
with computers:
AI **software**

kHz – MHz

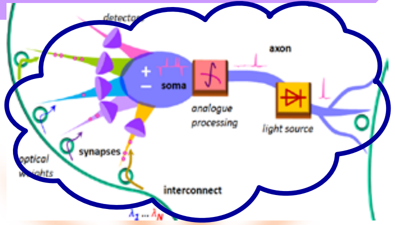


Neuromorphic electronics:
Hardware neural networks



Challenge: **Interconnect**
(capacitive loading, BW, EMI, routing, leakage / energy)

GHz



Neuromorphic photonics:
Hardware-based

Challenge: **scaling (PICs), all-optical NNs**

PhotonICs

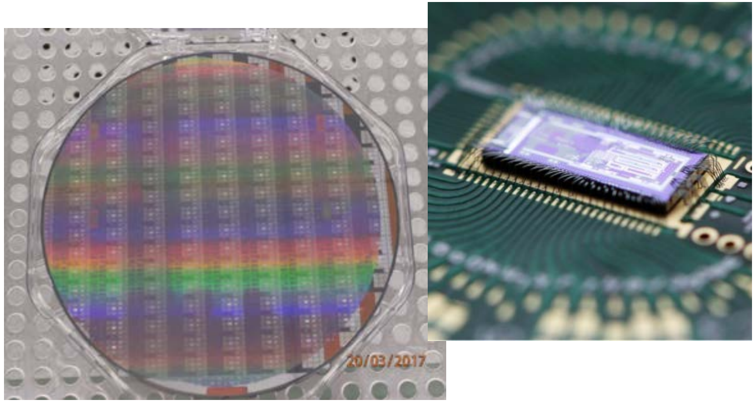
2020 **Silicon Photonics**

1986 ***Principal idea:***
using the silicon manufacturing supply chain to produce photonics

The quest for an 'optical silicon'

Even if some researchers would disagree that gallium arsenide and its cousins are the best way to go, most agree on the need for a practical material with all the required optoelectronic properties—what Tanguay has called an "optical silicon." **This ideal material should be versatile, stable, easy to work with, manufacturable, reproducible, and cheap.**

IEEE Spectrum, 1986



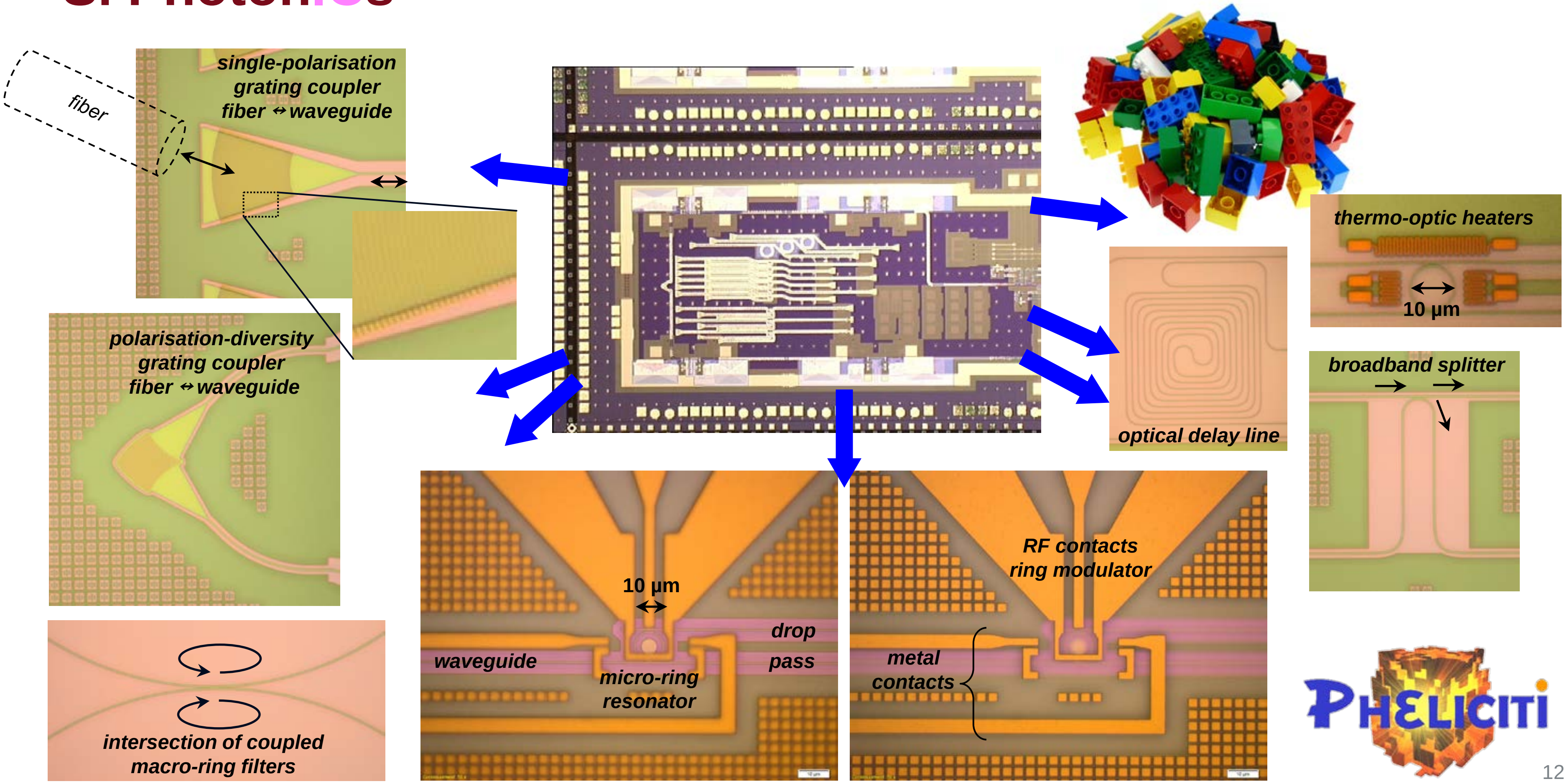
SiPh integrated multi-lane transceivers
on 200-mm wafer scale.

TCO: **< 0.3€ per Gb/s**

0.2€ / mm^2 in shared fab for 10M chips/year
Saturated 200-mm fab: 50M chips/month
50 datacenters equivalent: 1 Gchips or <2 fab years

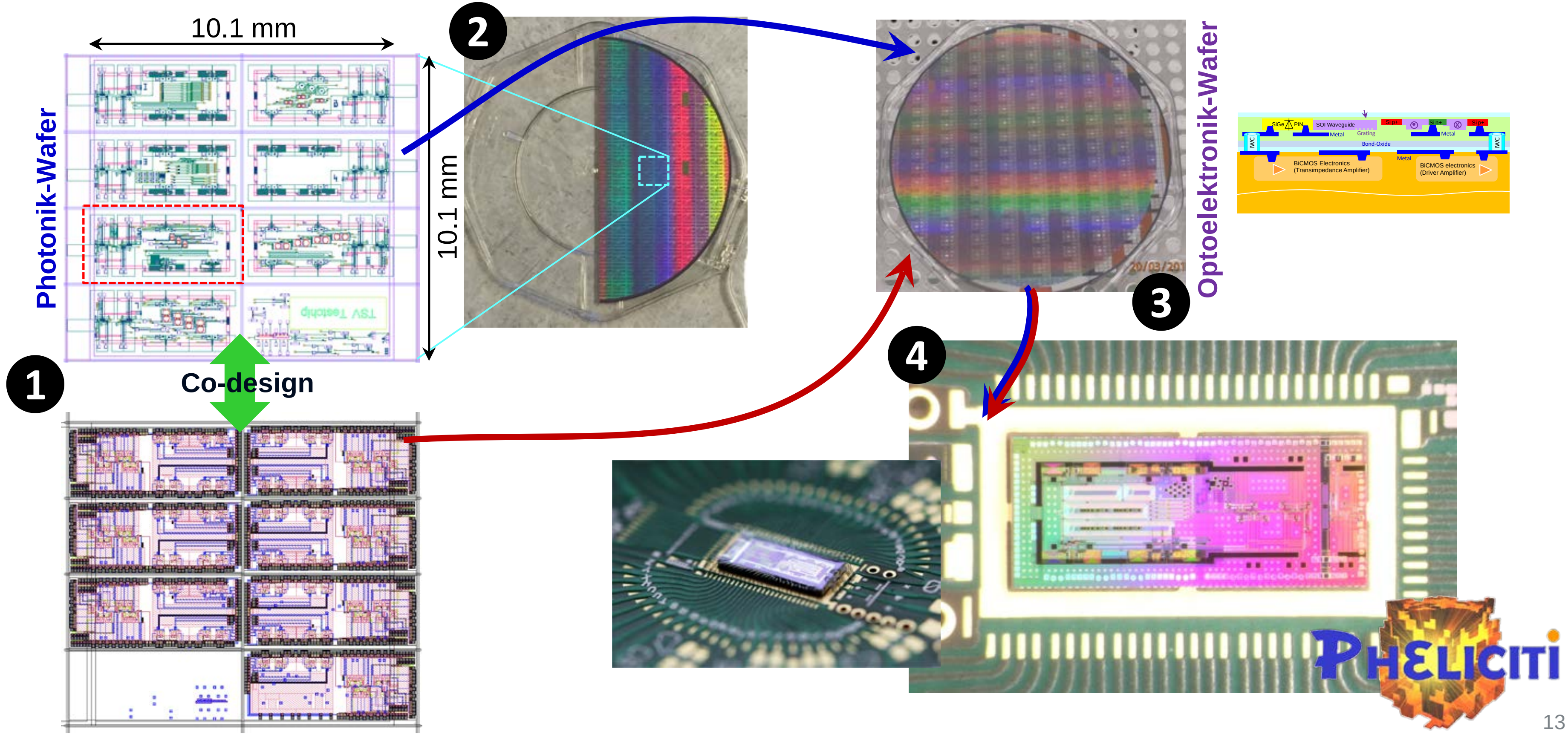
W. Bogaerts, 2017

Si PhotonICs



Si PhotonICs

AUSTRIAN INSTITUTE OF TECHNOLOGY

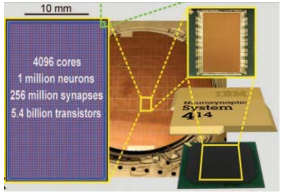


Electronic vs. Photonic Artificial Neural Network

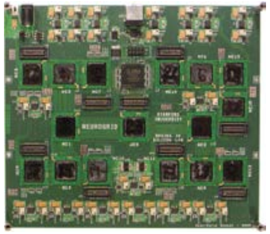
AUSTRIAN INSTITUTE OF TECHNOLOGY

A neuromorphic processor requires a large number of interconnects!

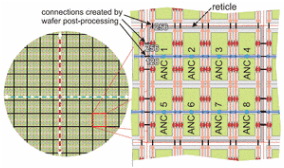
ELECTRONICS



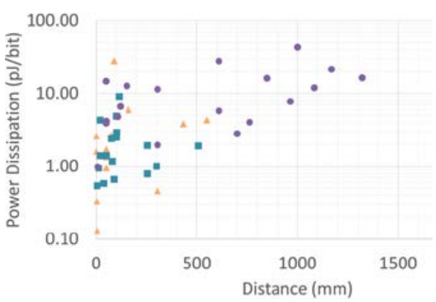
TrueNorth



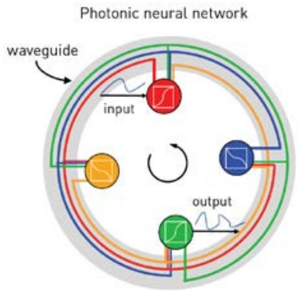
Neugrid



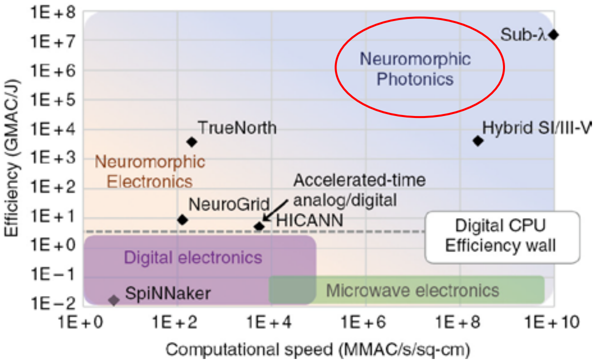
HICANN

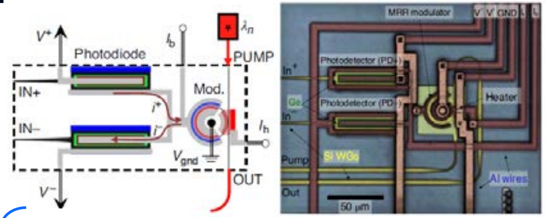


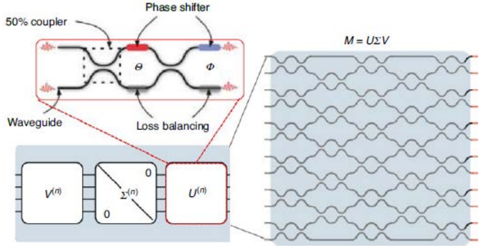
- Bandwidth-distance trade-off
- Huge energy consumption
- kHz-MHz clock rates

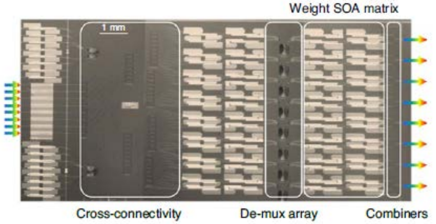


- Optical multiplexing
- Low energy consumption
- GHz information rates







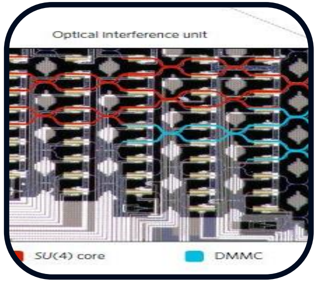


PHOTONICS

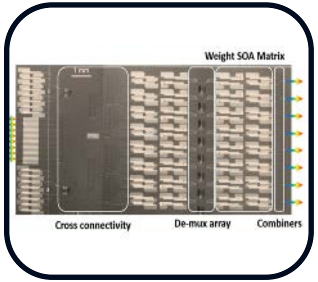
P. Merolla et. al., Science 345, 668 (2014)
J. Schemmel et al., ISCAS'10 (2010)
B. Benjamin et al., Proc. IEEE 102, 1174 (2014)
Y. Shen et. al., JLT 37, 245 (2019)

T. Ferreira de Lima et al., Nanophot. 6 (2017)
B. J. Shastri et al., Springer (2018)
A. Tait et al., Phys. Rev. Appl. 11 (2019)
Y. Shen et al., Nat. Phot. 11 (2017)
B. Shi et. al., JSTQE 26 (2020)

The Neuron Goes "Light"



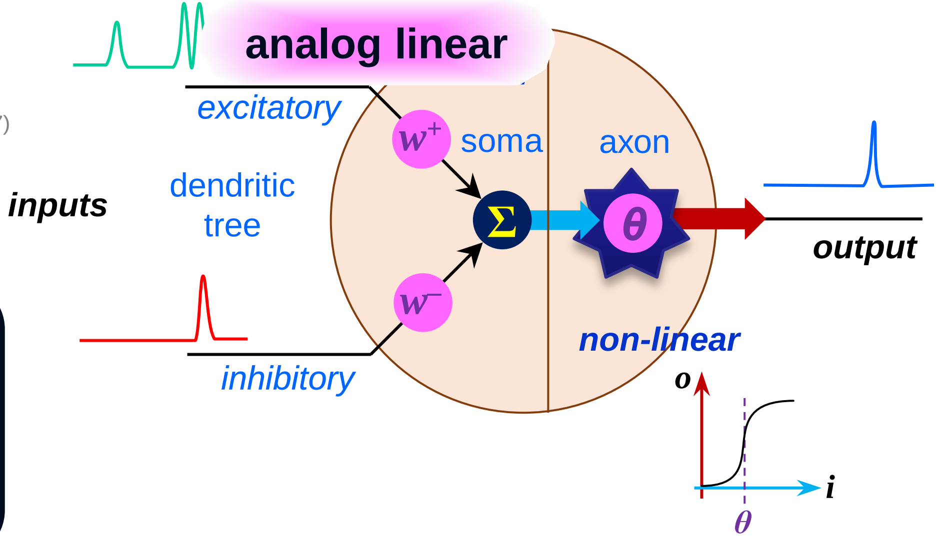
Y. Shen et al., Nat. Phot. 11 (2017)

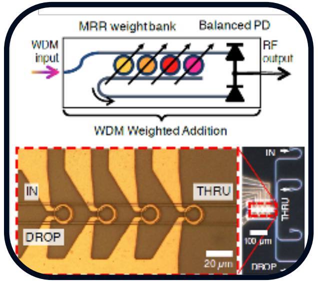


B. Shi et al., JSTQE 26 (2020)

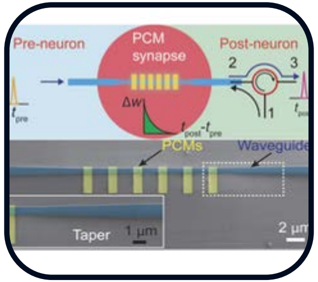
$$\sum_k v^{(k)} w_k$$

$$w_k \in [-1, +1]$$

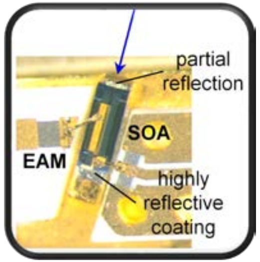




A. Tait et al., JSTQE 22 (2016)

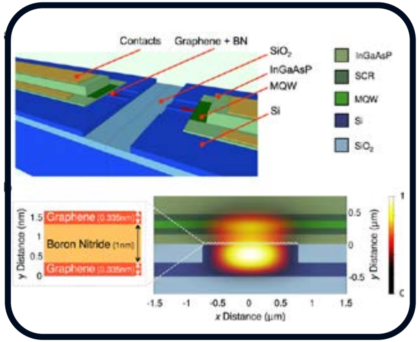


A. Tait et al., JSTQE 22 (2016)

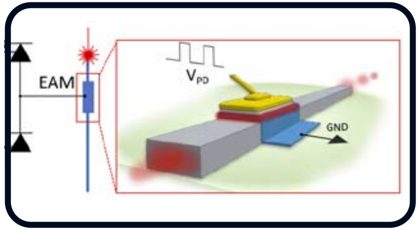


M. Stephanie et al., JLT 41 (2023)

The Neuron Goes "Light"



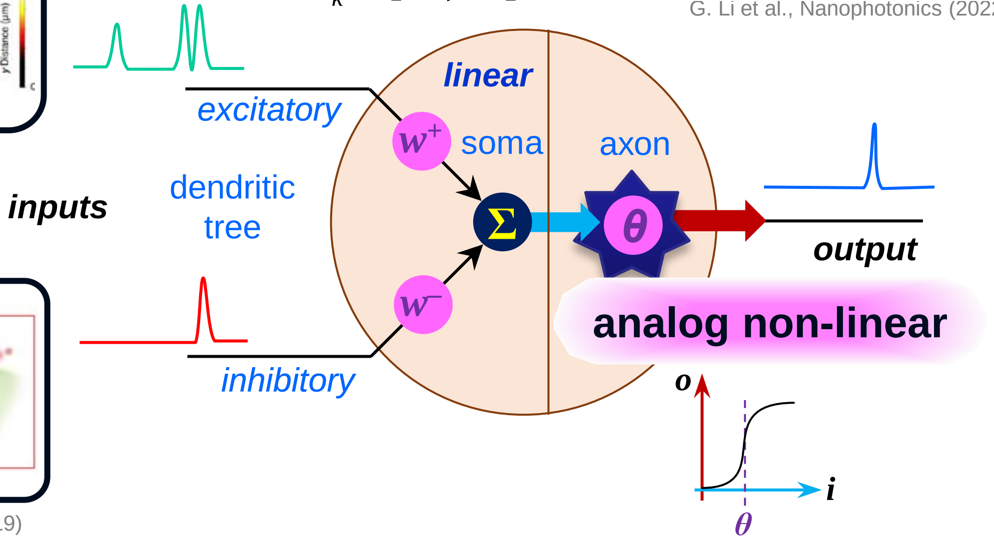
B. Shastri et al., Sci. Rep. **6** (2016)

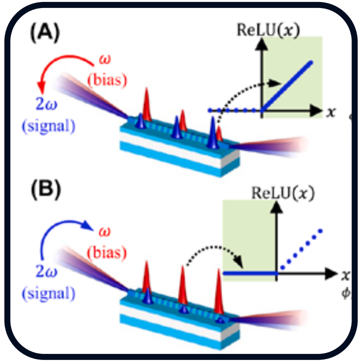


R. Amin et al., APL Materials **7** (2019)

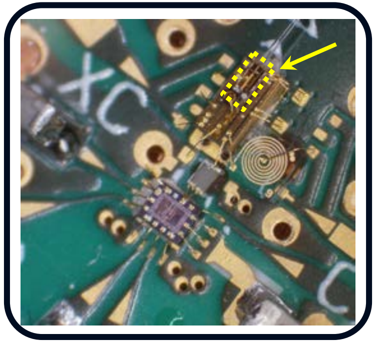
$$\sum_k v^{(k)} w_k$$

$$w_k \in [-1, +1]$$

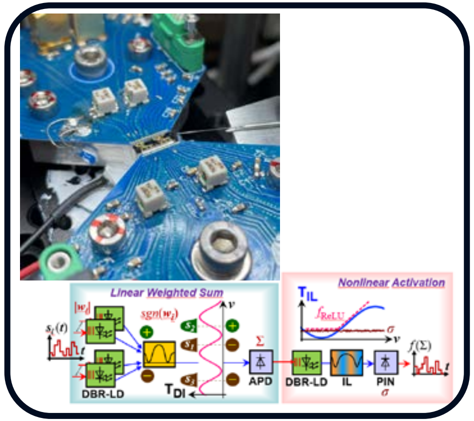




G. Li et al., Nanophotonics (2022)



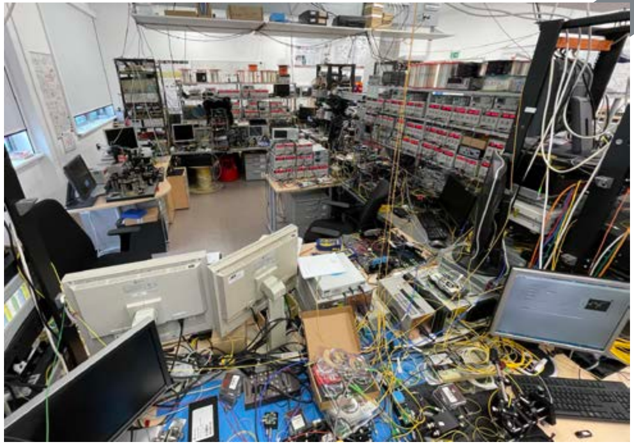
B. Schrenk, ECOC'19, We.P27 (2019)

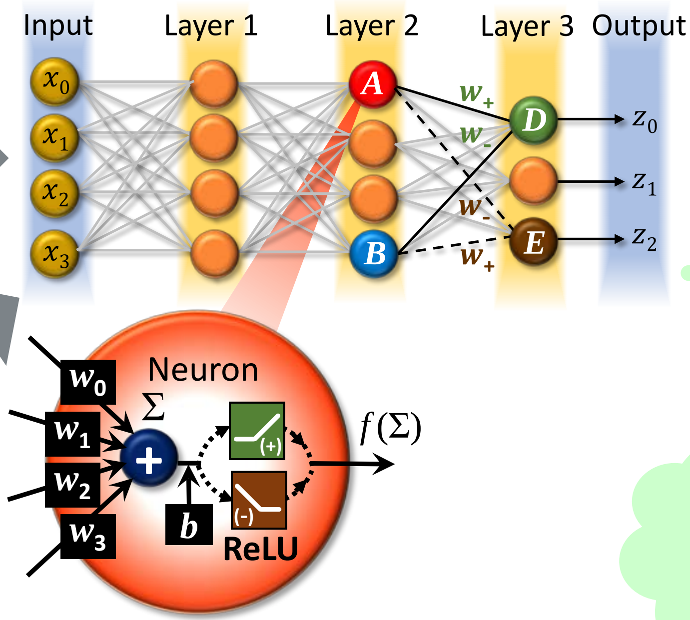


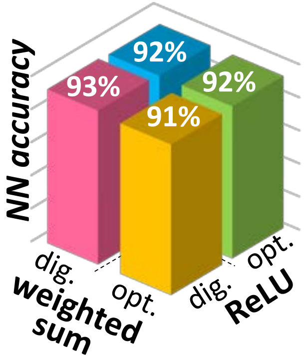
M. Stephanie et al., Proc. SUM, MF4.4 (2023)

Accuracy at Speed

Iris Classification Problem
150 flower samples







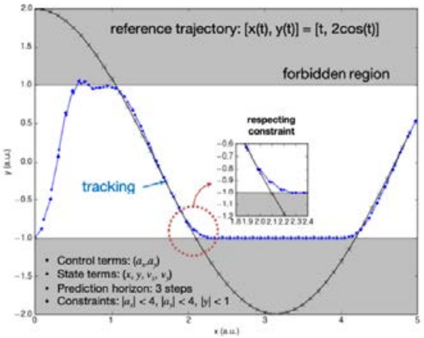
Accuracy

digital NN: 93%

optical NN: 91-92%
but one flower every ns!

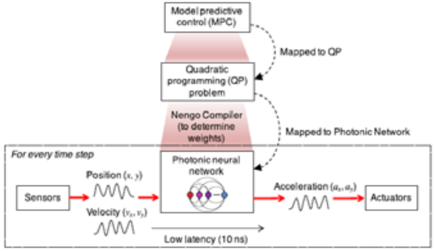
M. Stephanie et al., Proc. SUM, MF4.4 (2023)

What about “real” Applications?

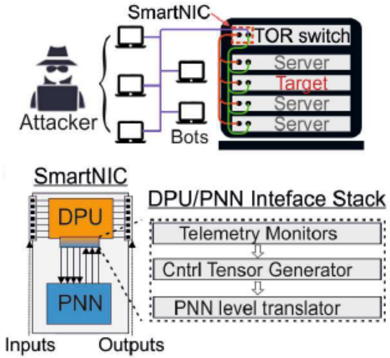


Predictive Control

- for object at flight
- 24 neurons
- convergence time of 10 ns

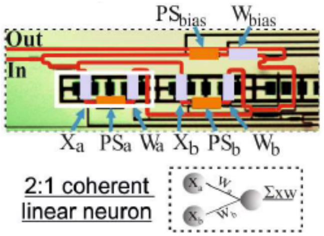


T. Ferreira de Lima et. al., JLT 37 (2019)

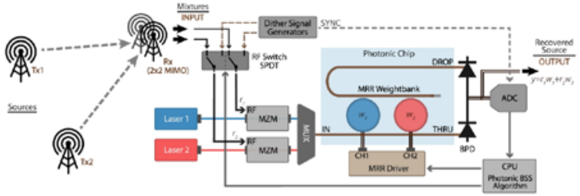


Distributed Denial of Service (DDoS) Attack Identification

- Using silicon photonic processor
- 50 GHz signal rate

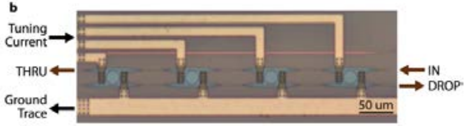


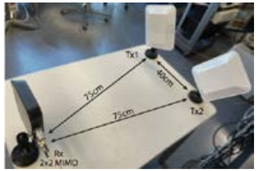
A. Tsakyridis et al., Proc. OFC (2023)



Blind Source Separation

- Separating an unknown mixture of unknown independent signals
- Using microring weight bank
- achieving a processing bandwidth of up to 19.2 GHz



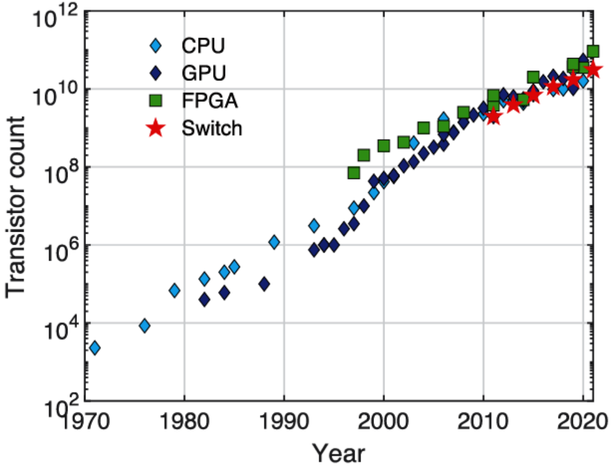


W. Zhang et. al., Nature Comm. 14 (2023)

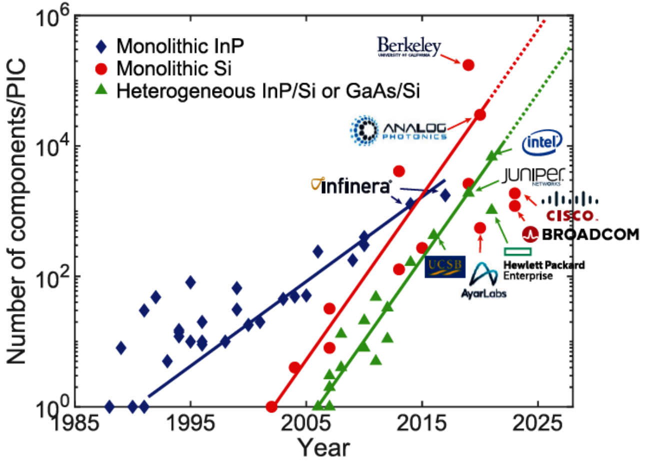
Is there a Moore's Law for PICs?

AUSTRIAN INSTITUTE OF TECHNOLOGY

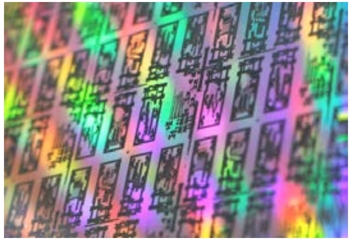
Electronics



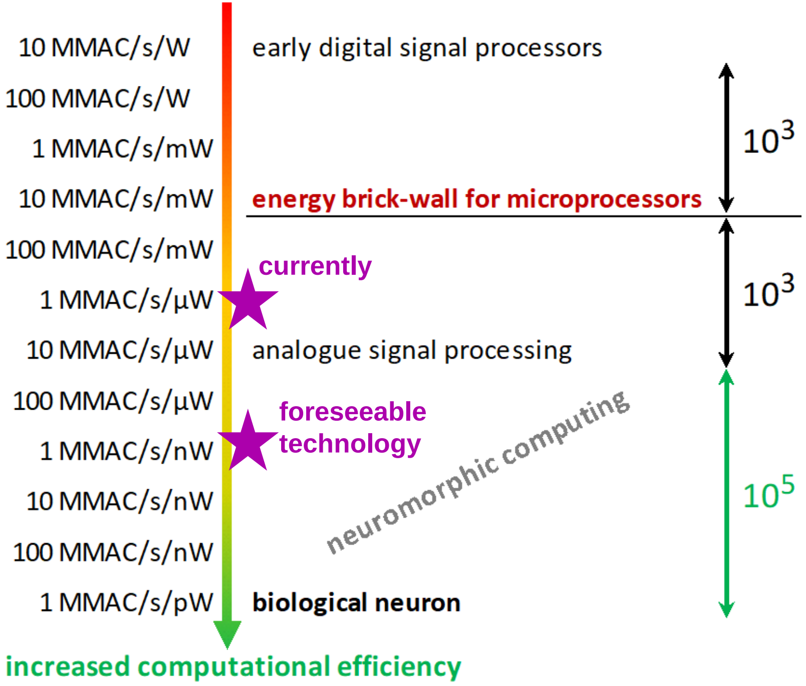
Photonics

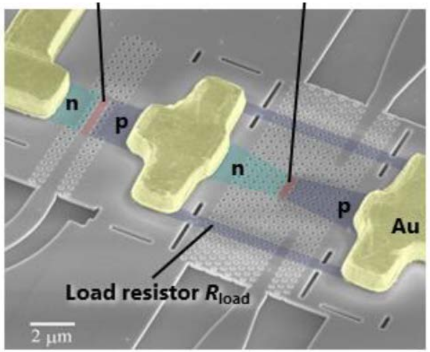


N. Margalit, Appl. Phys. Lett. 118, 220501 (2021)



…and last, what about Energy Efficiency?





K. Nozaki, Nat. Photon. **13**, 454 (2019)

foreseeable:

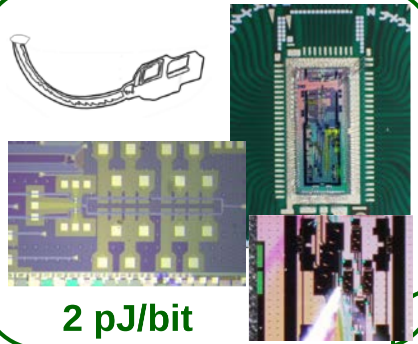
1.1 fJ/MAC

Lao Tzu

Take Awayyy

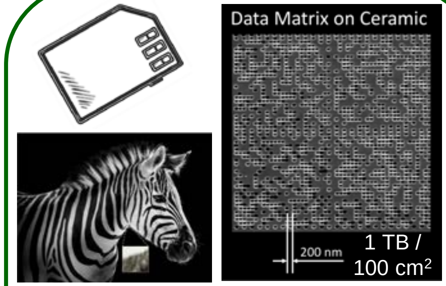
There are promising solutions to keep the raising ICT energy footprint in check.

At the same time, we obtain better performance than the state-of-play.



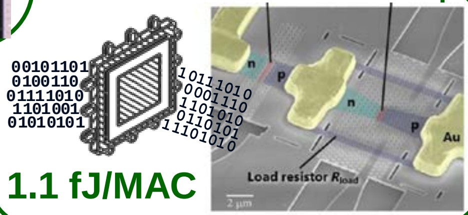
2 pJ/bit

ICT



(c) Ceramic Data Solutions

Gb/s WR/RD speed,
1 nJ/bit
Migration obsolete.



1.1 fJ/MAC

Another Decathlon for Another Decade!

Bernhard Schrenk
bernhard.schrenk@ait.ac.at

Center for Digital Safety & Security
AIT Austrian Institute of Technology

Jollybee

COYOTE
COHERENT OPTICS EVERYWHERE

Part of the work shown was supported by the Austrian Research Promotion Agency through the JOLLYBEE project (grant nº 887467).

Part of the work is also supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 804769).

erc

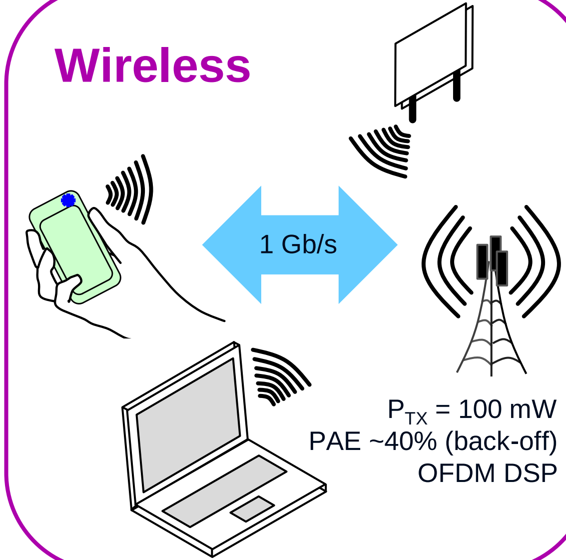
Communication: Information in Motion

- between people
- between machines

Energy / bit = Power consumption / Data rate

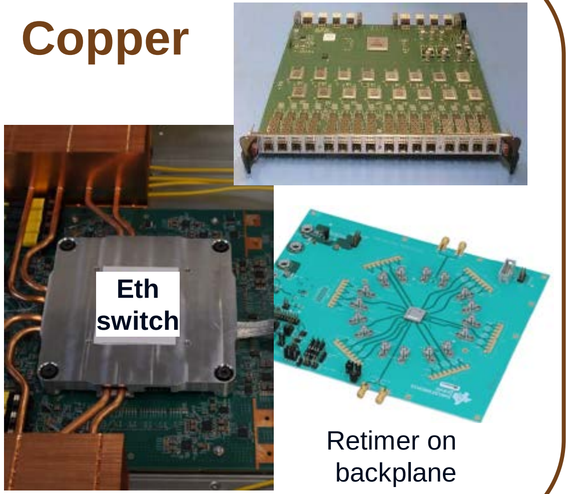
$$\text{Energy / bit} = \frac{\text{Power consumption}}{\text{Data rate}}$$

Wireless



P_{TX} = 100 mW
PAE ~40% (back-off)
OFDM DSP

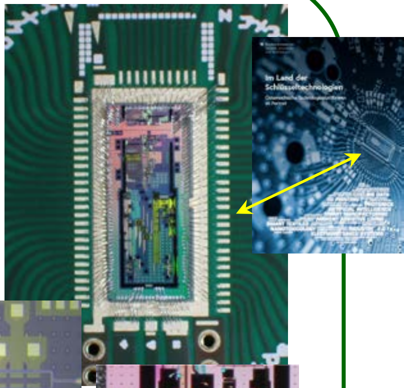
Copper

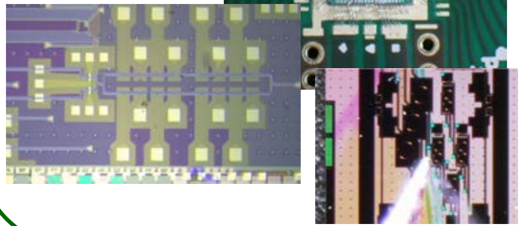


Retimer on backplane

Optical

Opto-el. transceivers (Si, InP)





1000 pJ/bit → LiFi? (10-100 Gb/s)

18 pJ/bit → **2 pJ/bit** Active optical cables, PON, etc

Storage: Information at Rest

- short-term caching
- long-term storage

$$\frac{\textbf{Energy}}{\textbf{Preserved bit}} + \textbf{Hardware Resources}$$

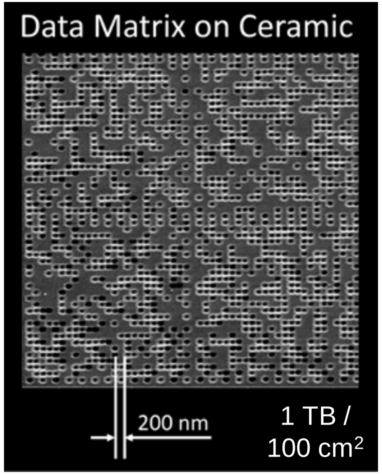
HDD / SSD

30-60 nJ/bit/year

migrate every 2-3 years

Optical memory

Cold storage: archival (30+ years)



(c) Ceramic Data Solutions

write / read at Gb/s speed and 1 nJ/bit

No migration required.