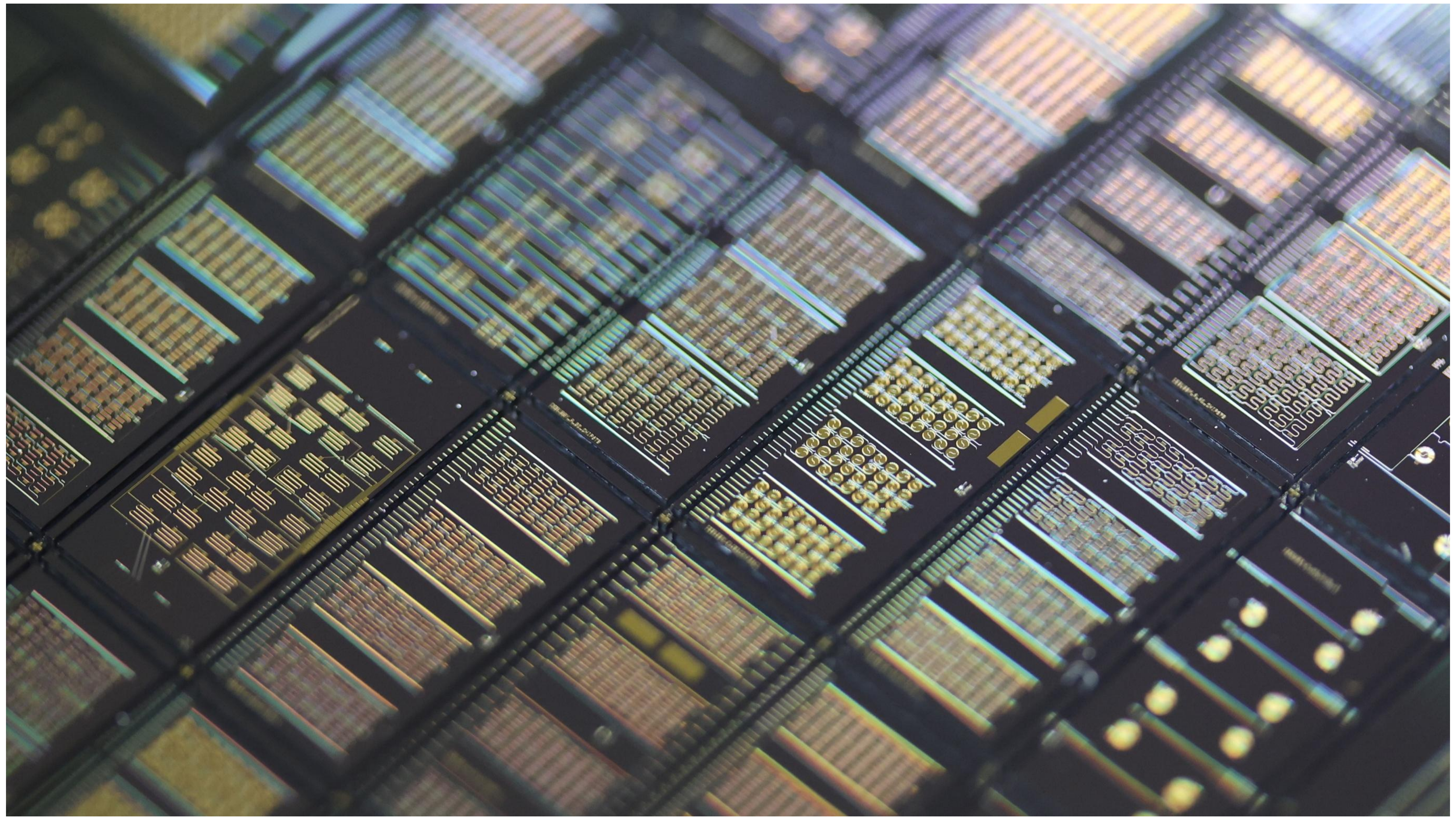


Photonic computing with Lithium Niobate integrated Photonics

from Concepts to Products

Dr. Mathias Metsch

Q.ANT at a glance: key company figures

2018
founded

2
Business Units

2.300
sqm Workspace

> 100
Q.ANTies

21
Nationalities

6
Publicly Funded Projects

>110
Patent Families

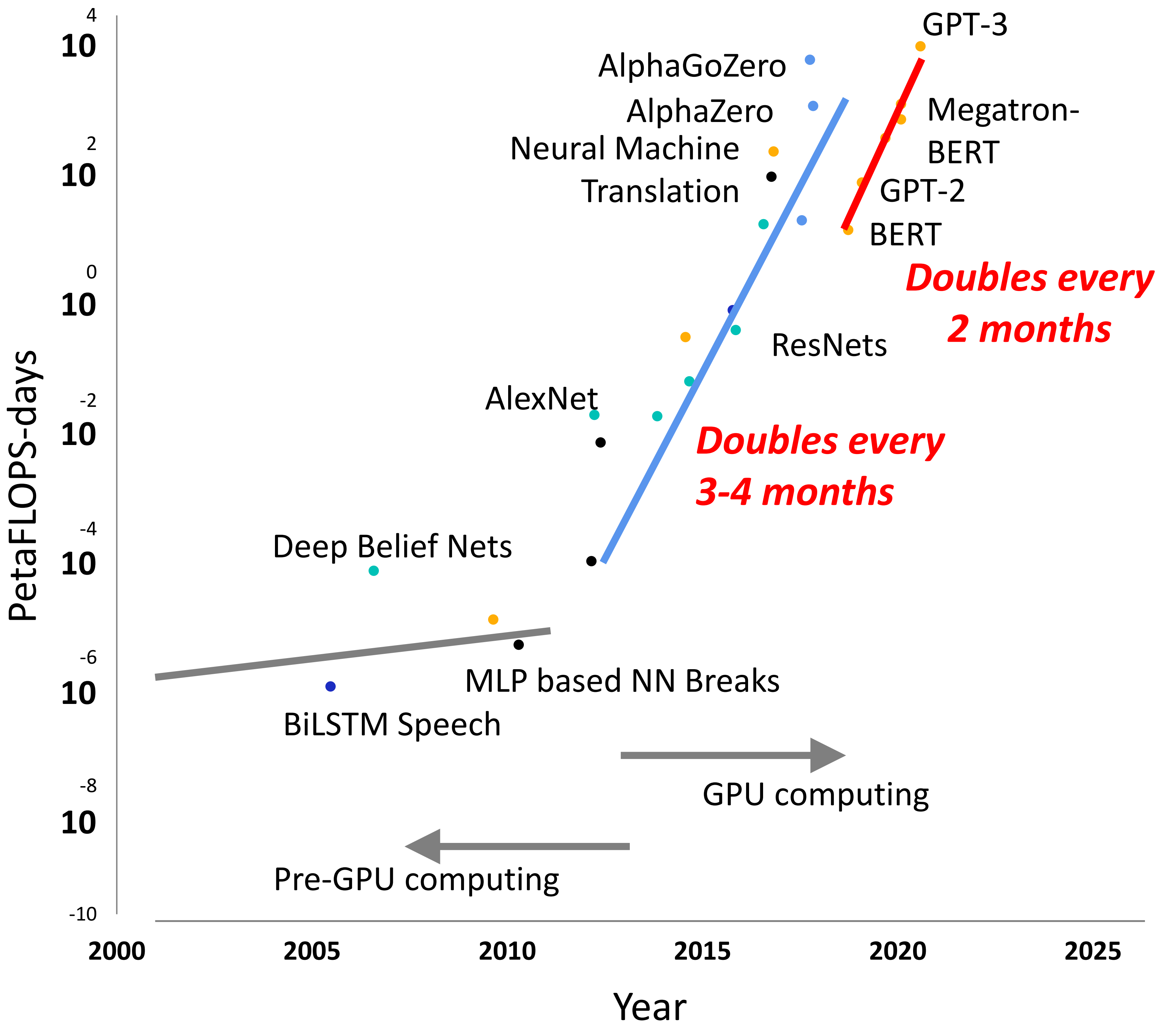
3
World Premiers

7
Coffee Machines

The surge in data centre power consumption continues, mainly driven by sharply rising utilization

Moore's Law for CMOS is long gone.

The demand for compute power doubles every 3-4 months since AI usage on GPUs is rising



Moore's law
Dennard Scaling
Koomey's law
Amdahl's law

Memory wall
Communication bottleneck

Today's Computer ecosystem is self-limiting by three fundamental bottlenecks - A unique opportunity for photonic computing

Several bottlenecks are major obstacles

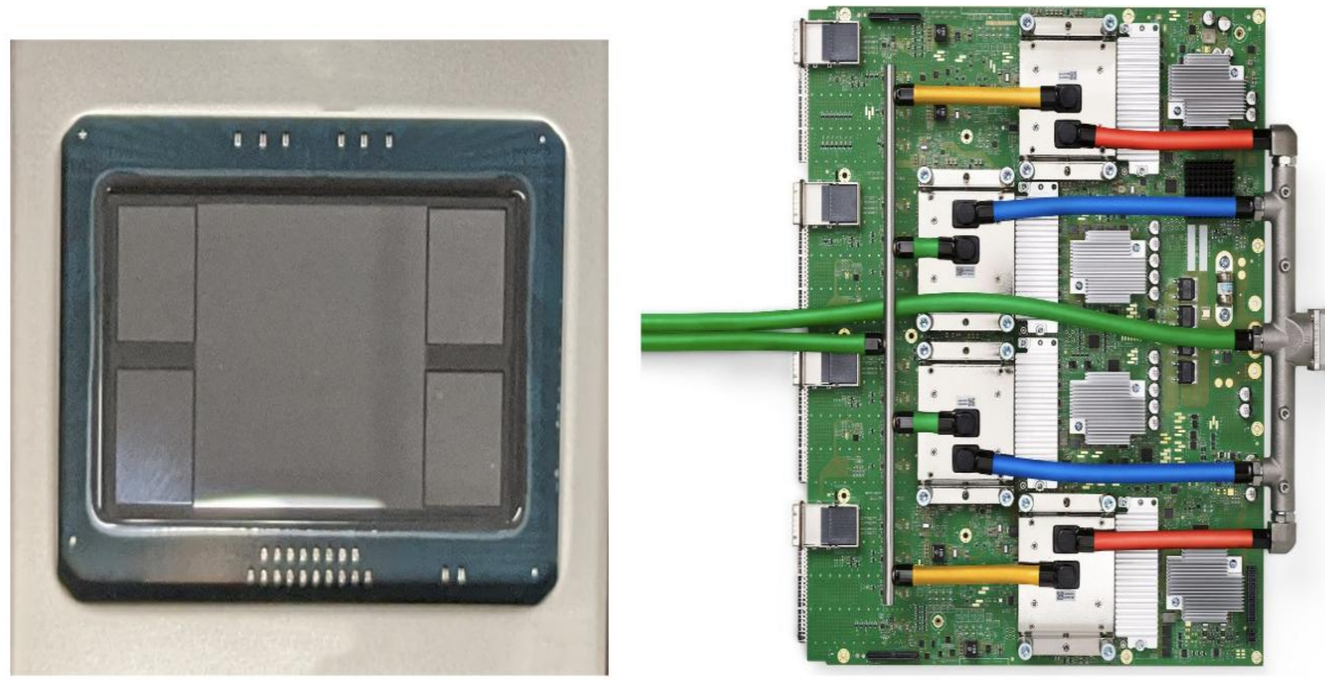
1. Digital implementations are hitting scaling limits and hinder innovations in algorithms
2. The energy demand of data centers will bring existing power infrastructure to its limits
3. Limited semiconductor production capacity gives a handful companies oligopoly power

A window of opportunity for alternative technologies opened

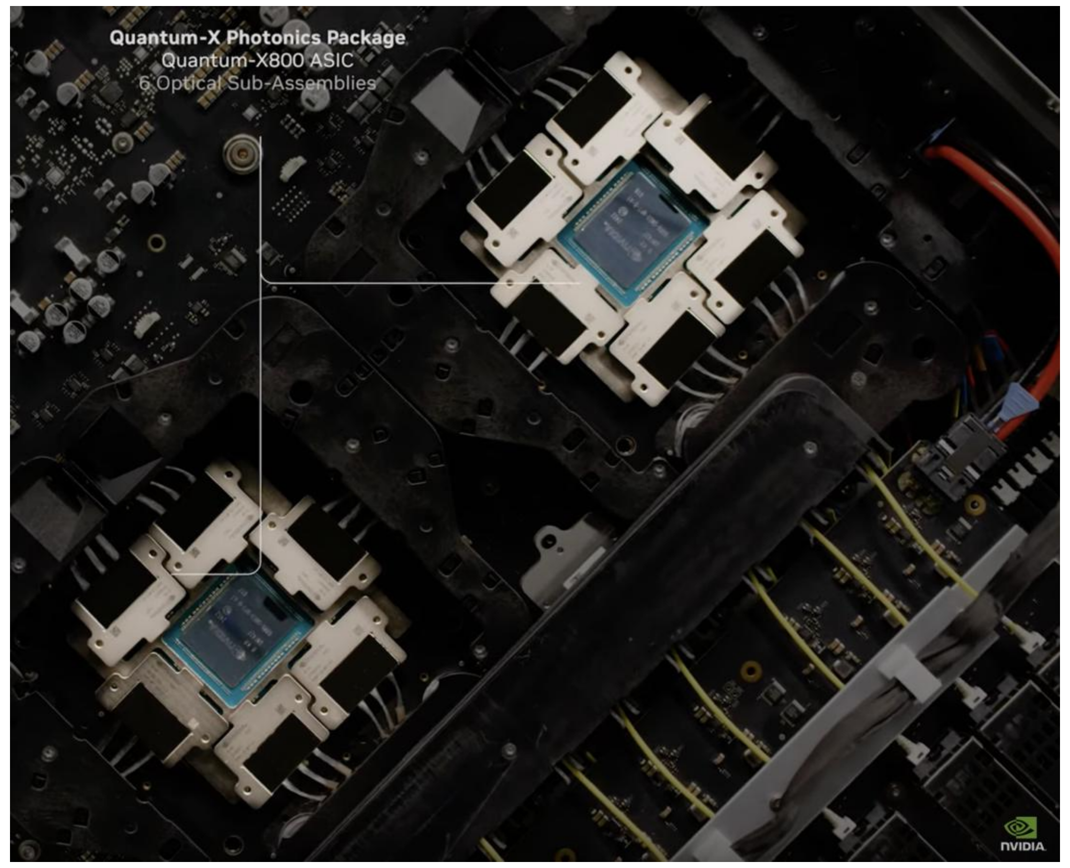
But don't bet against silicon or CMOS!

Addition yes, replacement no

- CMOS has evolved over the last years
  - Multi-die / chiplet approach
  - 2.5D packaging with stacked memory
- Even completely different materials are common where power is key
  - GaN & SiC in power electronics
- Photonics plays its role in communication and HPC



https://arxiv.org/pdf/2304.01433

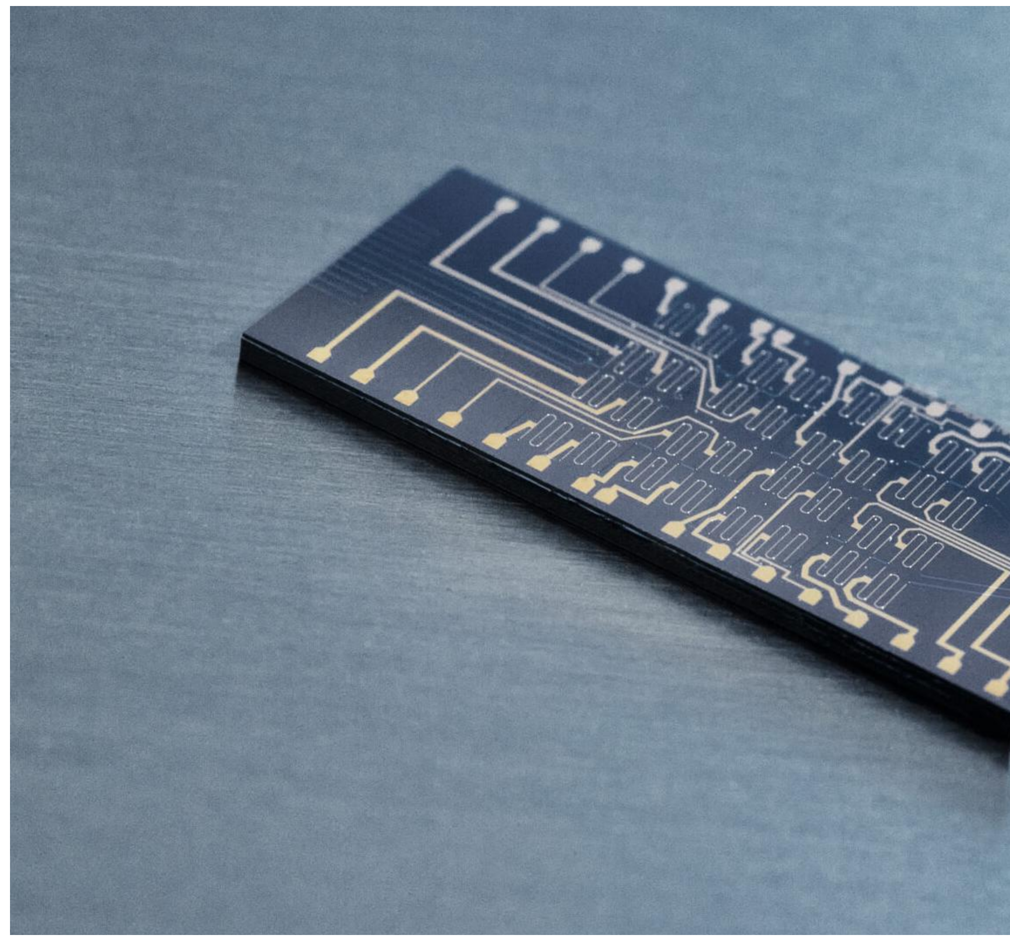


Nvidia, https://www.youtube.com/watch?v=kS8r7UcexJU

1988: First optical transatlantic cable → 2001: SFP standard for pluggable modules → 2025: Co-packaged optics enables AI scale out → Photonic computing

Electronics vs Photonics

	Electronics	Photonics
Carrier of information & energy	Electrons	Photons
Propagation in	Conductive tracks	Dielectric waveguides
Typical geometry	Manhattan	Curvi-linear
Minimal structure size	Few tens of nm	Few hundreds of nm
Complexity	Billions of units	Hundreds of units
Typical chip size	Well sub mm^2 to cm^2	Few mm^2 to cm^2
Main material	Si by far	Si, but more diverse
Nonlinearity	PN-junctions	Gain and $\chi^{(2)}$ (and $\chi^{(3)}$)
Typical wavelength		1 to 1.5 um, mostly 1550nm
Development state	Extremely mature and reaching peak maturity?	Maturing and developing



Photonic computing scales not just physically but also through higher density math and higher density transformations

Scaling lever	Variables	Example
1 Physical scaling	▪ Clock rate ▪ Modes ▪ Wavelength ▪ # of photonic chips	▪ 32 Modes x 1 GHz x 32 wavelengths x 8 chips = 262 TOPS
2 Arithmetical scaling	▪ Use of mathematical functions that are simple to implement in optics	▪ 1024 DFT = 3072 OPS ▪ DFT at 1 GHZ = 3 TOPS
3 Algorithmic scaling	▪ Use of algorithms that need less cycles to achieve the same detection rates	▪ Some transforms are better suited than others for predicting outcomes

Photonics is a great platform for computing

1 Distributed computing is native to photonics

Enabling an essential ingredient to any modern HPC

2 The native bandwidth of photonics is 1000 times higher than in electronics

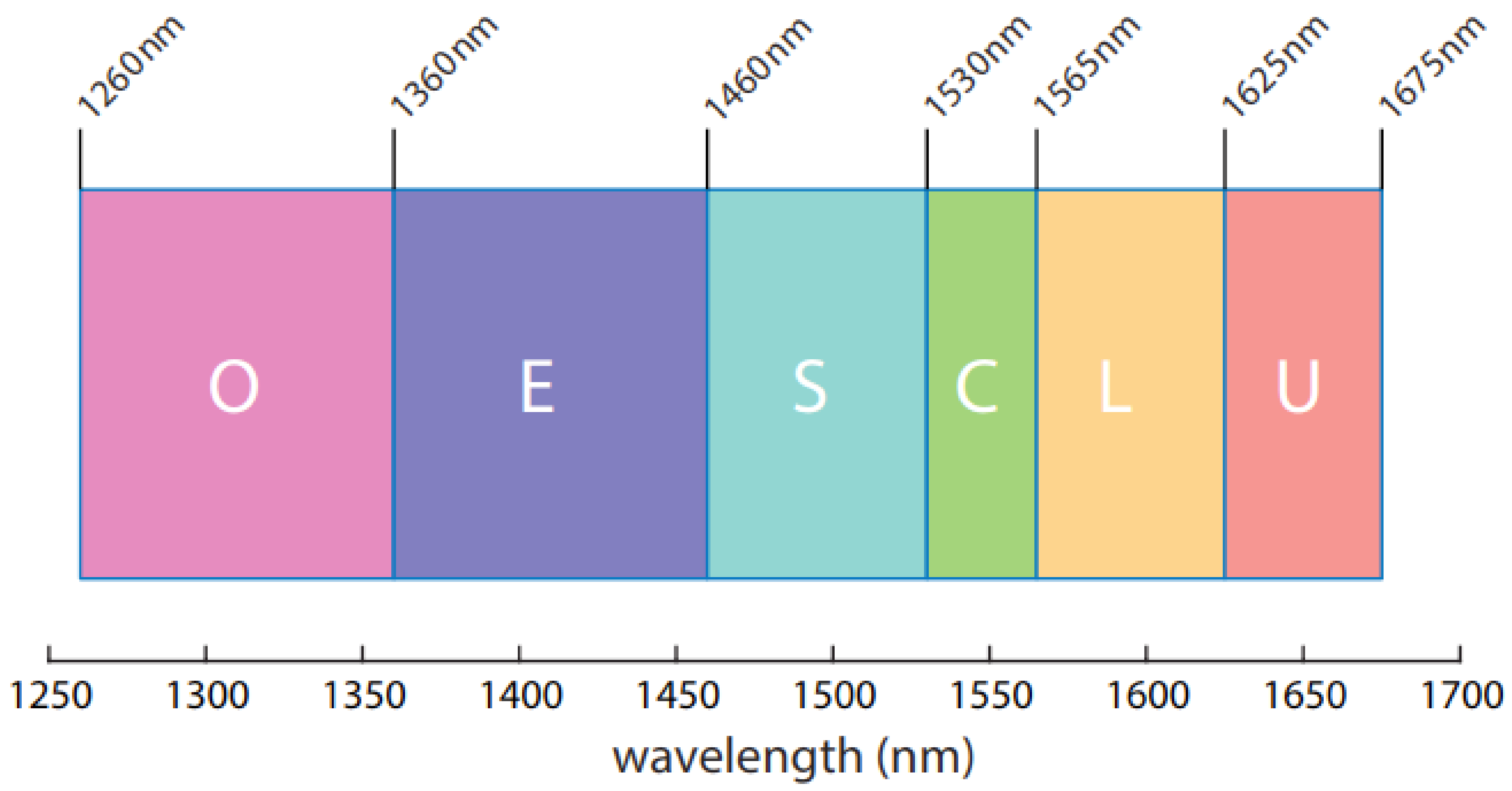
Enabling massive frequency multiplexing

3 Complex operations are native to photonics

Enabling their direct and energy efficient computations in the analog domain

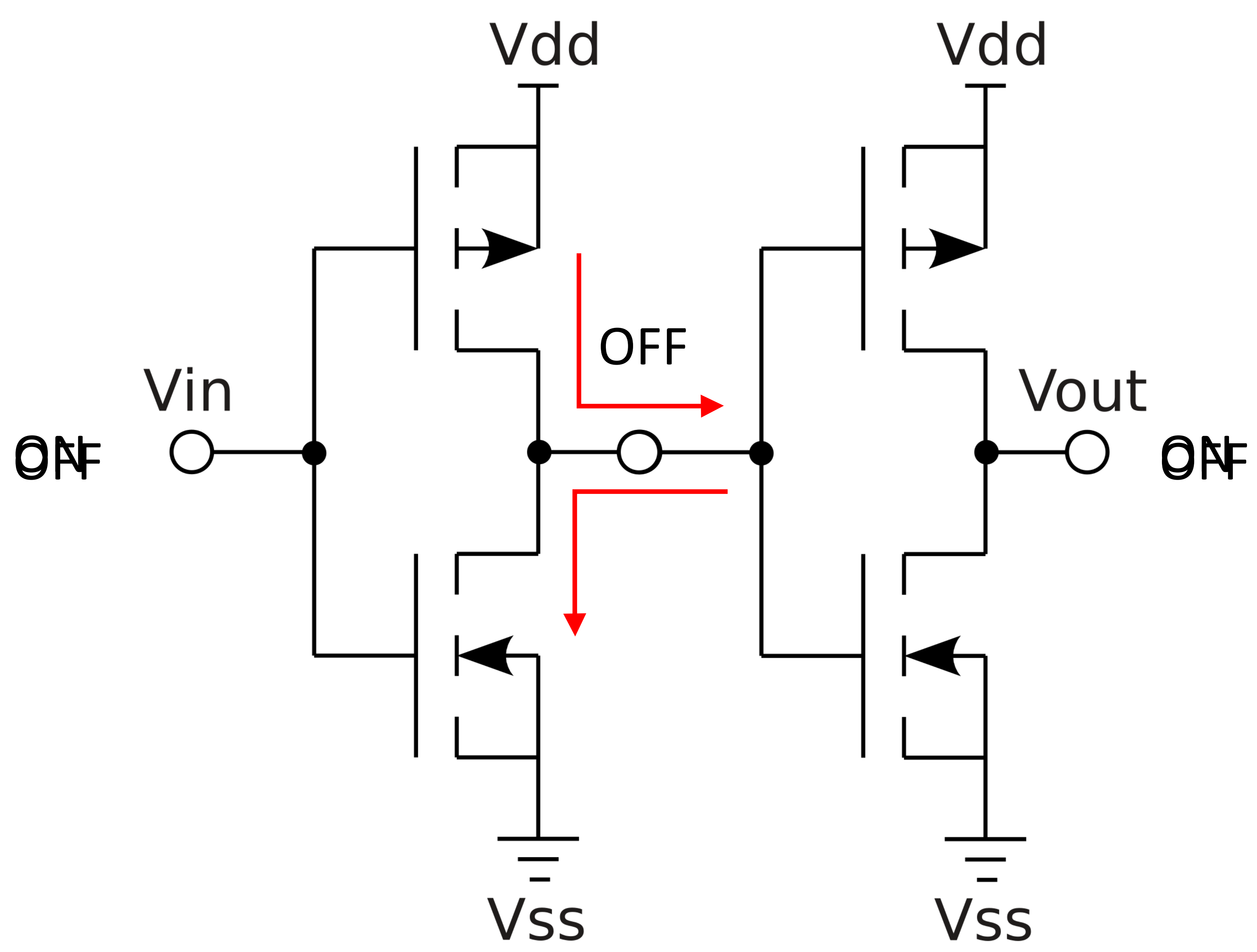
4 Data flow processing architecture can be implemented with light

Near-deterministic processing



Photonic chips for data processing can solve this conundrum, running at the speed of light

Digital Computing



- In CMOS, electrons are moving
- Switching uses electrical currents

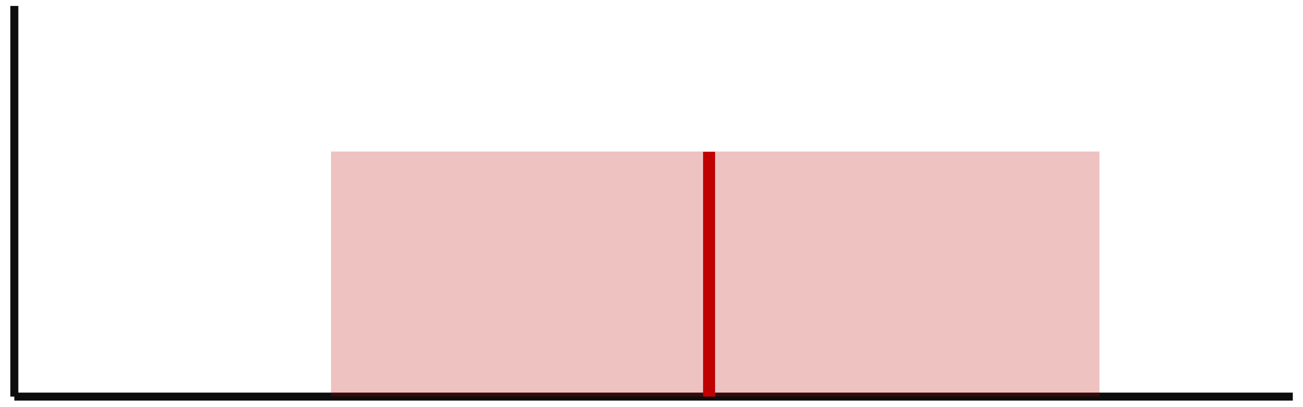
Photonic Computing

- In photonics, photons are flowing
- Control uses electrical fields

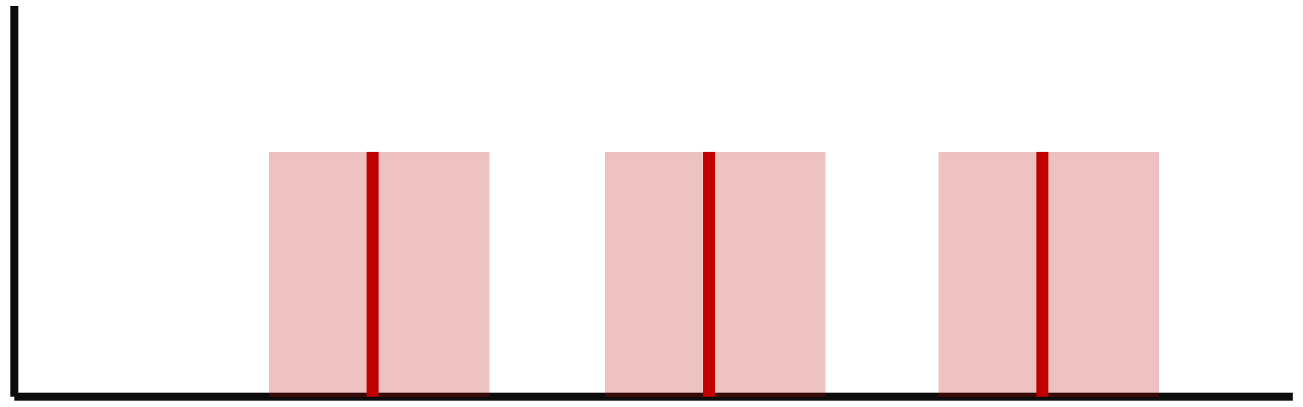
How we make full use of the large optical bandwidth

- The large optical frequency bandwidth opens a new window of parallelism not available in the CMOS world – frequency parallelism (or wavelength parallelism/WDM)

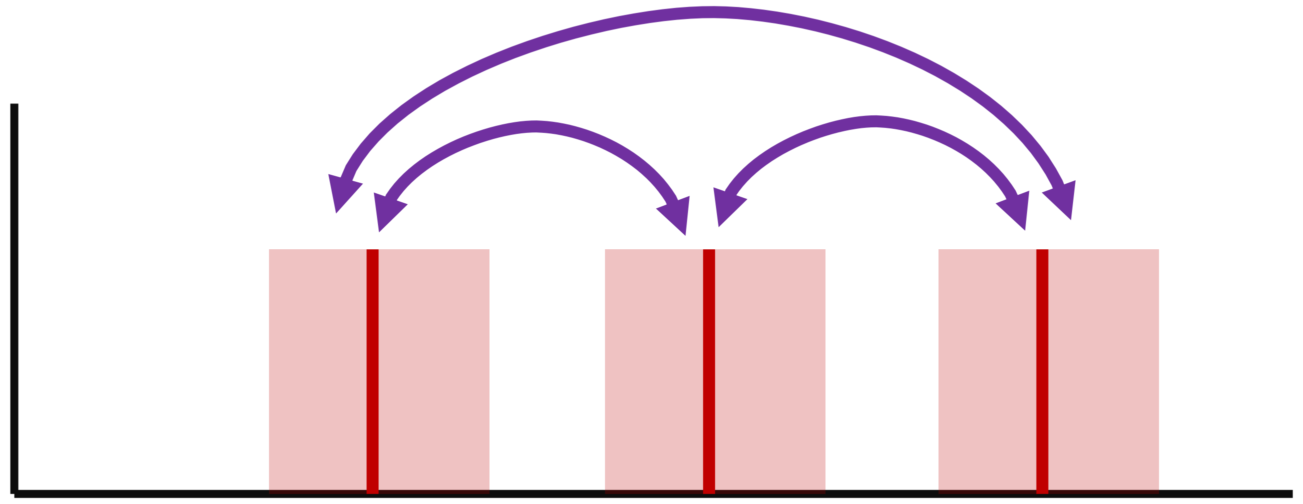
- Modulation with high frequency but in-sync with electronics
  - Regular computing CMOS work with up to 5 GHz



- Implementing WDM
  - Parallel, independent processing

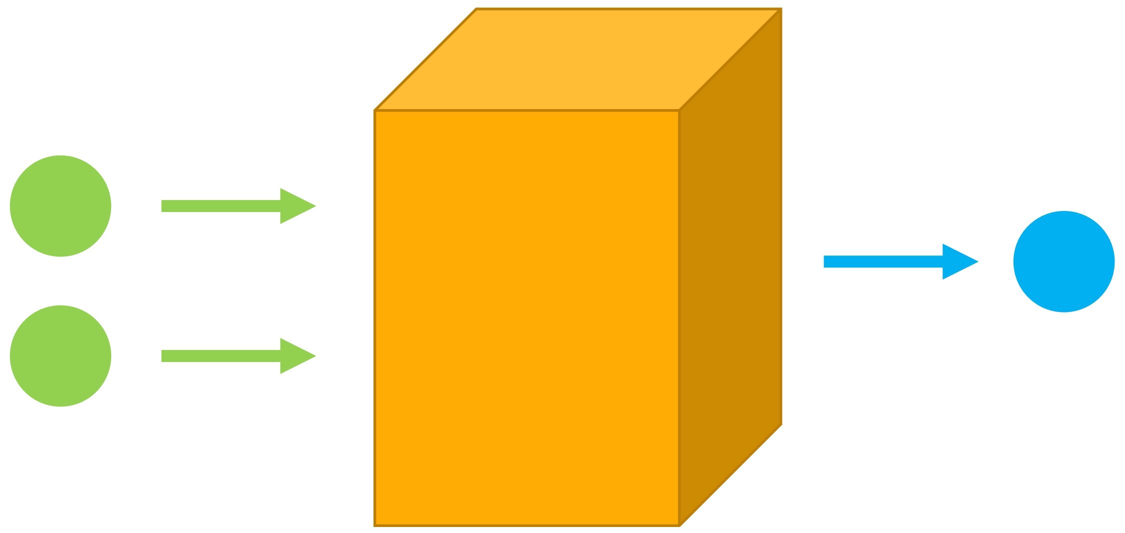


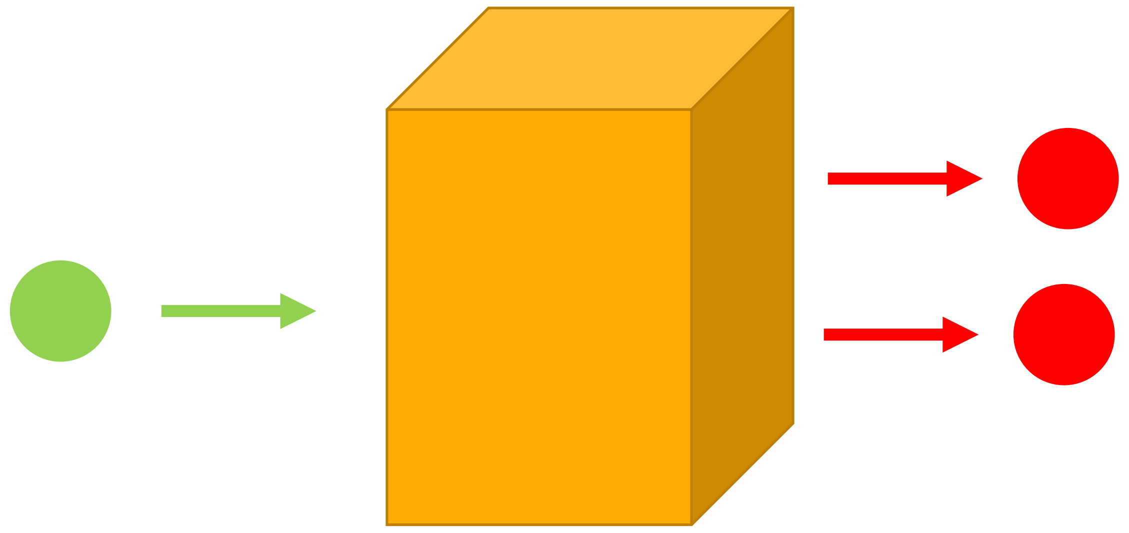
- Interaction in between light
  - Optical nonlinearities (multiplication in amplitude, addition in frequency)
  - Photodetection (summation in power)



What nonlinear $\chi^{(2)}$ optics can do

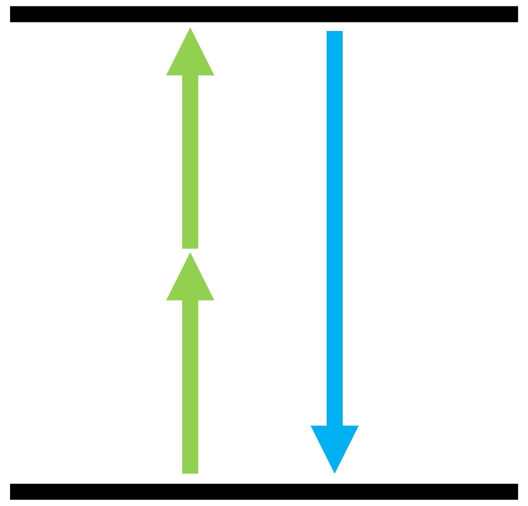
Interactions of photons with photons via matter





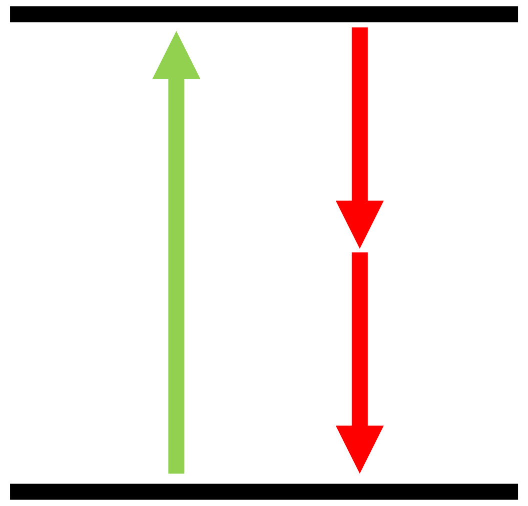
Second Harmonic Generation

$E_{out} = E_{in} + E_{in}$

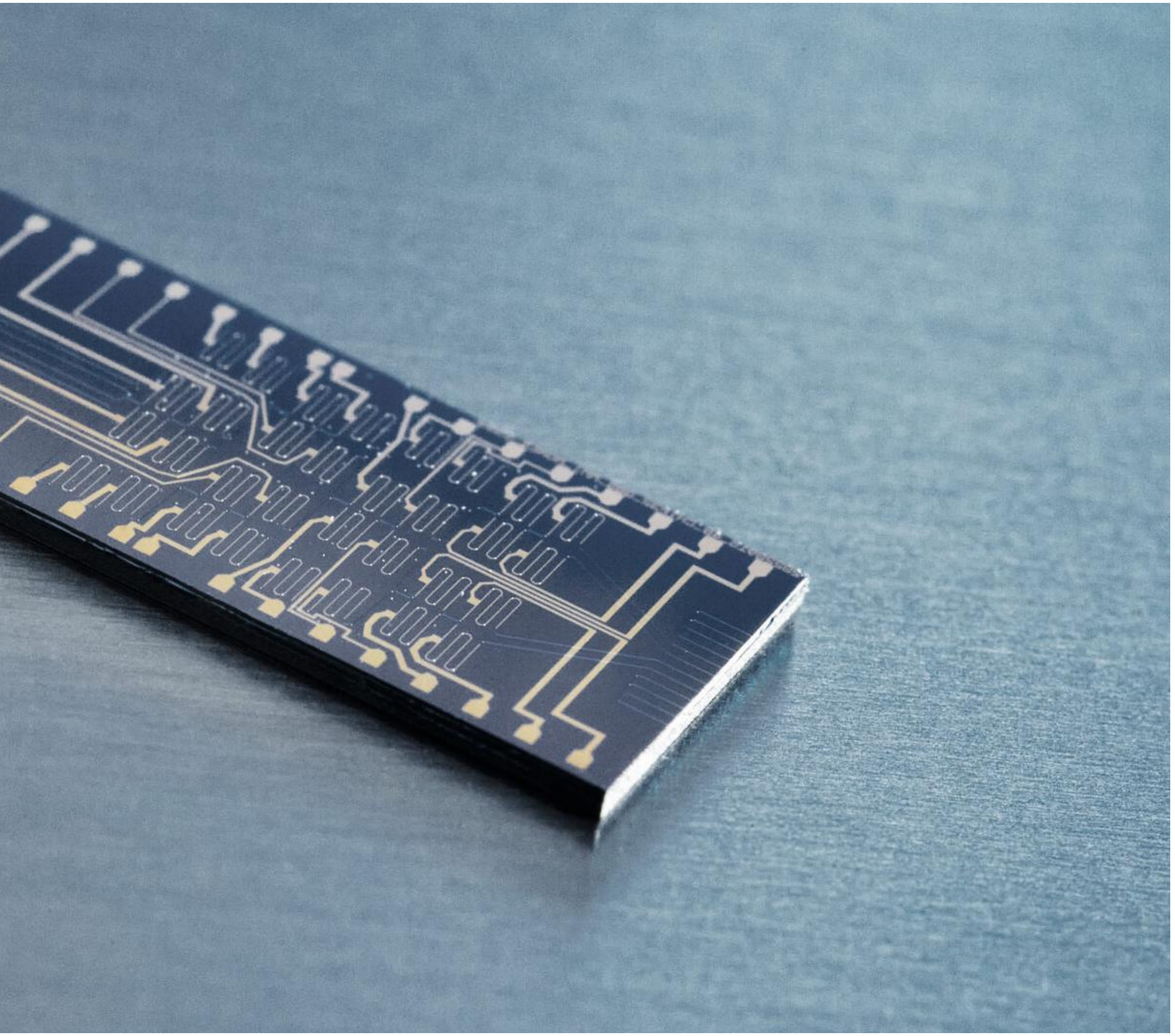


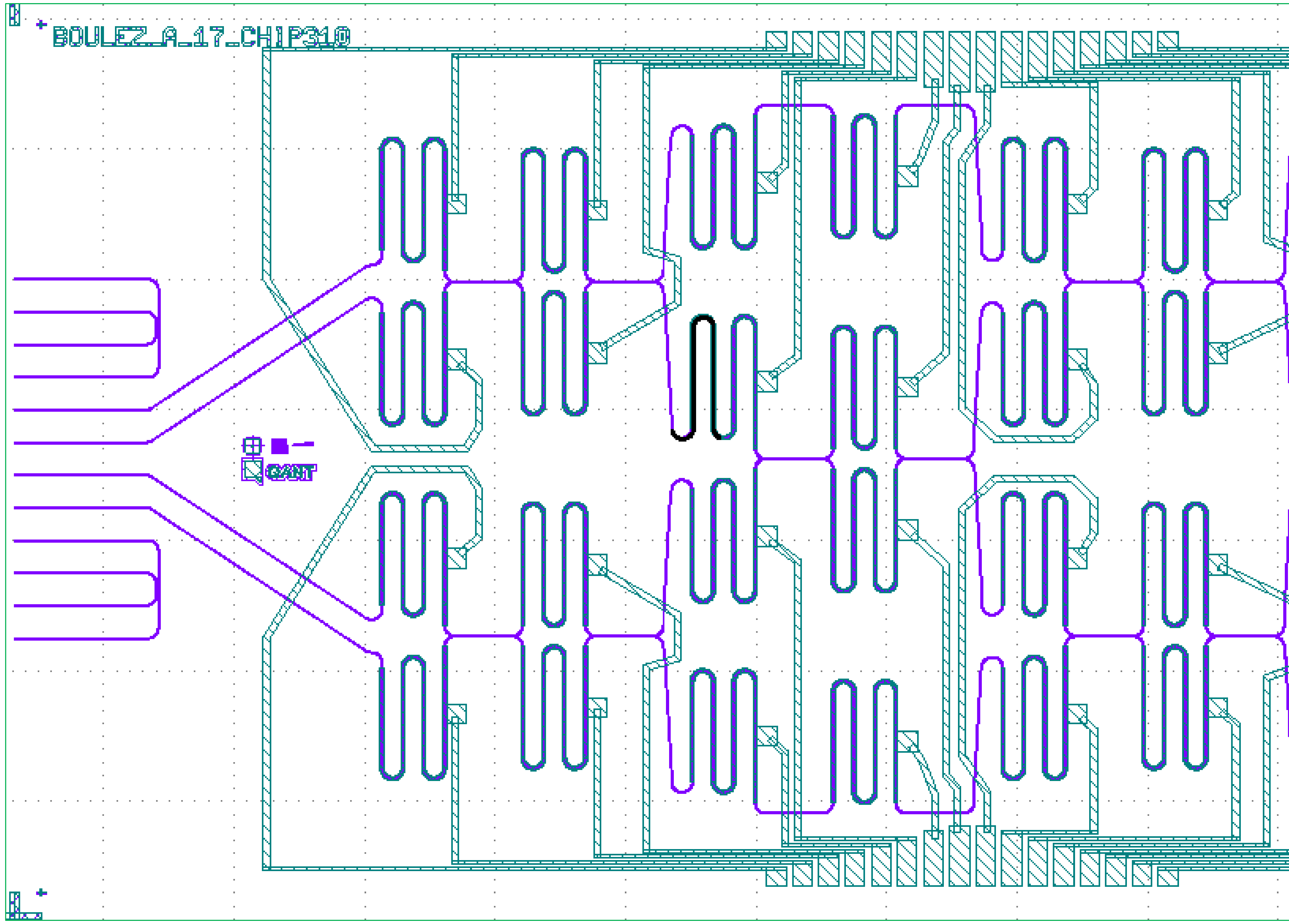
Parametric down conversion

$E_{out1} + E_{out2} = E_{in}$



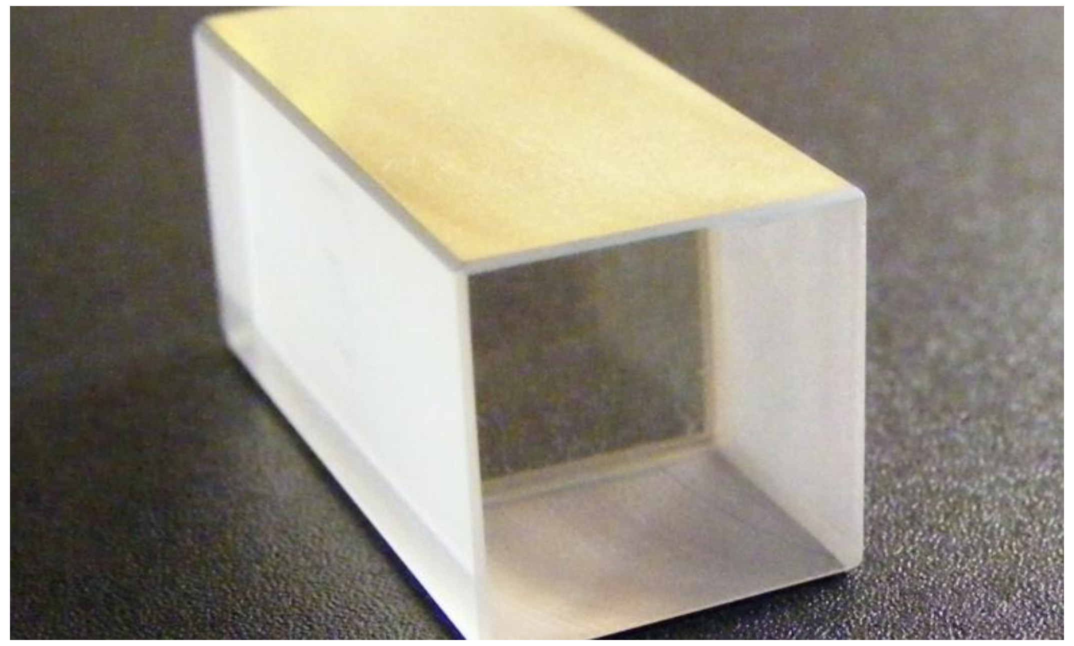
What is integrated photonics?





Lithium niobate is a great material for photonic compute

Transparent



- Large transparency window spanning visible and IR
- Refractive index around 2 – comparable to SiN

Nonlinear

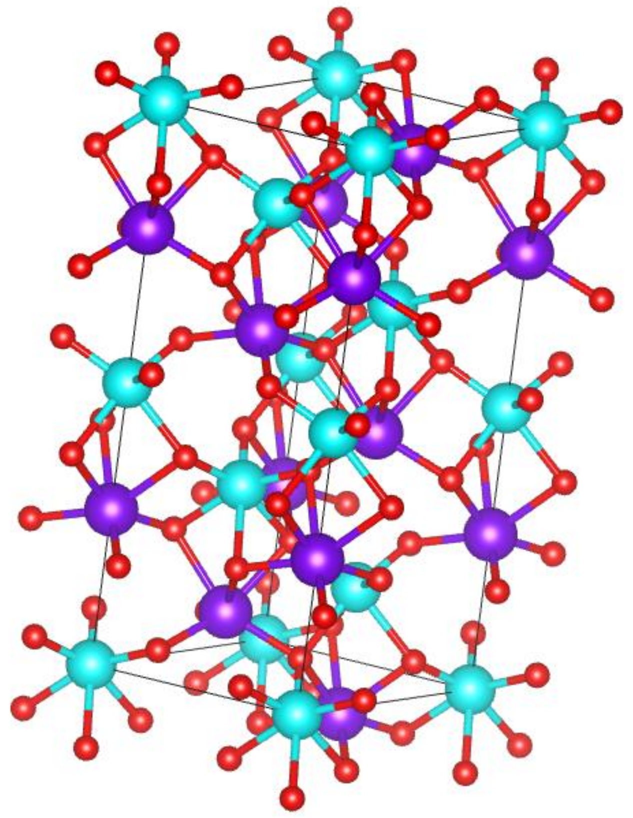
$$\chi^{(2)}\mathbf{E}^2(t) +$$

- Very high $\chi^{(2)}$ nonlinearity
- Can be poled
- High $\chi^{(3)}$ nonlinearity

Electro-optic

- Very high electro-optic coefficient
- Used in the fastest modulators today

Known

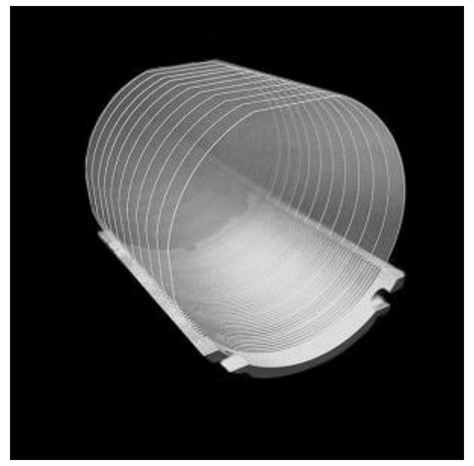


- It has been around for a while
- Commercially availabe as wafers of different sizes
- Commercially available as thin film on insulator (LNOI/TFLN)

Pictures from: deltroniccrystalindustries.com, Wikipedia, ixblue.com

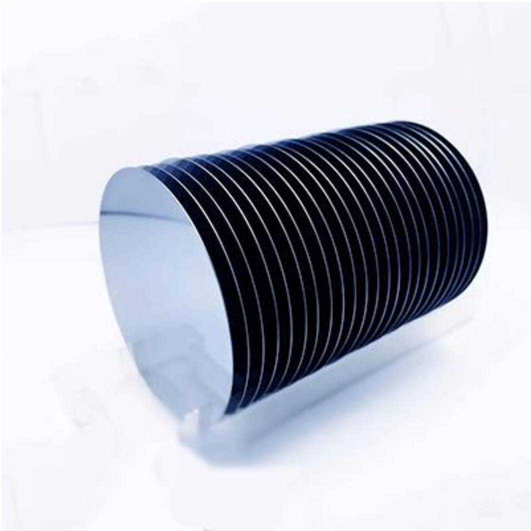
Q.ANT controls the entire value chain of our Native Processing Server (NPS)

$LiNbO_3$ Wafer



+

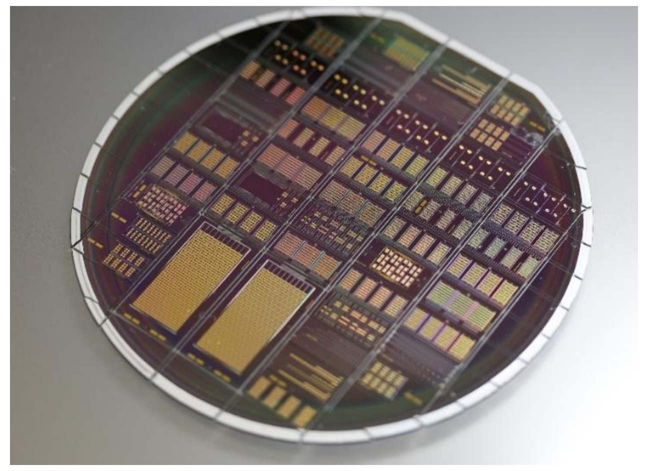
Silicon Wafer



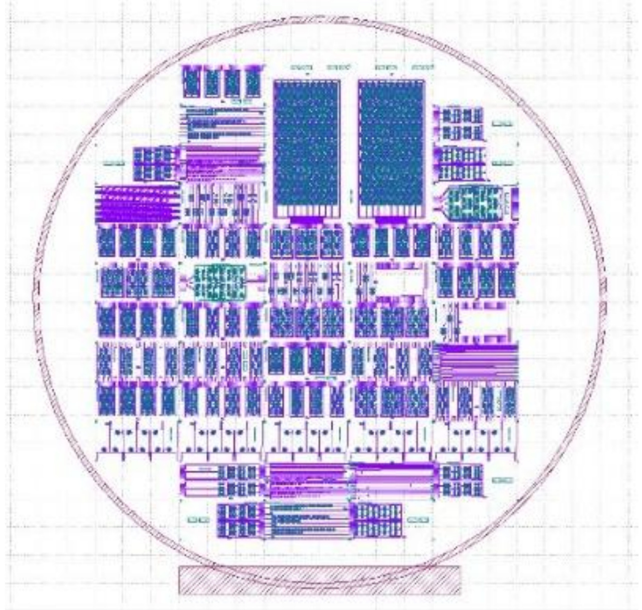
externally sourced

Thin-Film Lithium Niobate Wafer

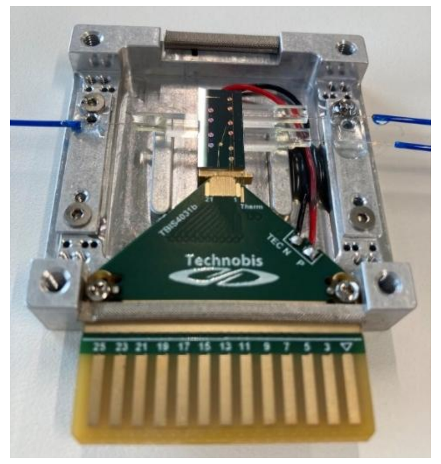
Processed Wafer



PIC Design



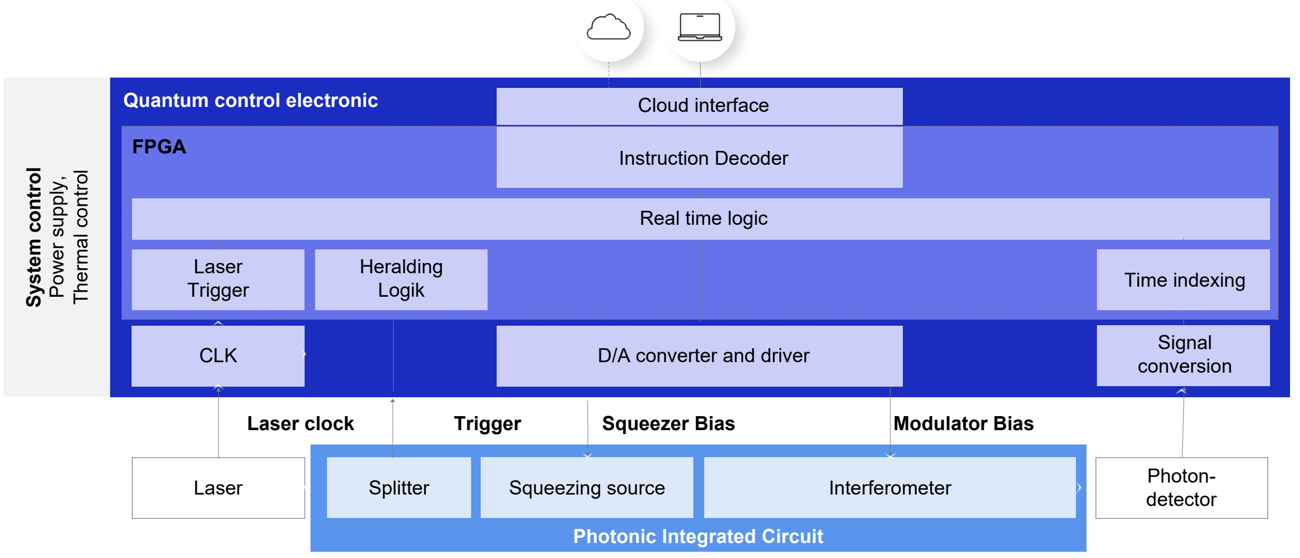
Packaged Chip



NPS

Computing System

System Design



In a “world-first”, Q.ANT demonstrated the Native Processing Unit (NPU) at the end of 2024

NPU for accelerating matrix multiplication

Specifications

- System: PCIe card, double slot
- Throughput: 100 MOPS
- Power consumption: 30 W
- Host interface: PCIe Gen. 3 x8
- Linux device driver
- FPGA based control electronics

NPUs come as turn-key system in Q.ANT's Native Processing Server (NPS)

Native Processing Server

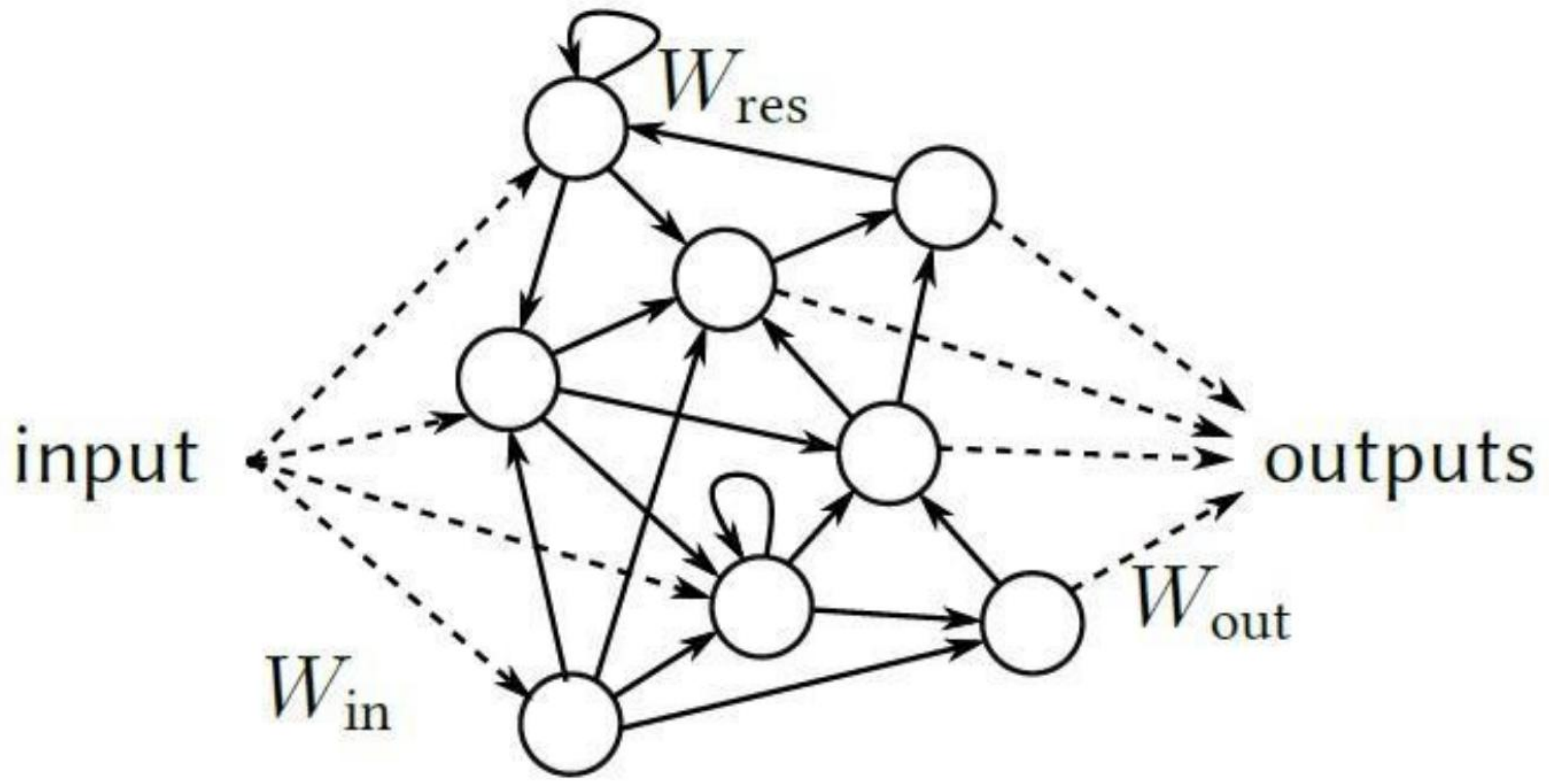
Specifications

- System: host system based on x86 19''
  4U rack mount
- At least one NPU; upgradeable to the latest versions
- Power consumption: 500 W
- Network Interface: 10 Gbit Ethernet
- Linux operating system

There is not just the one way to do photonic computing

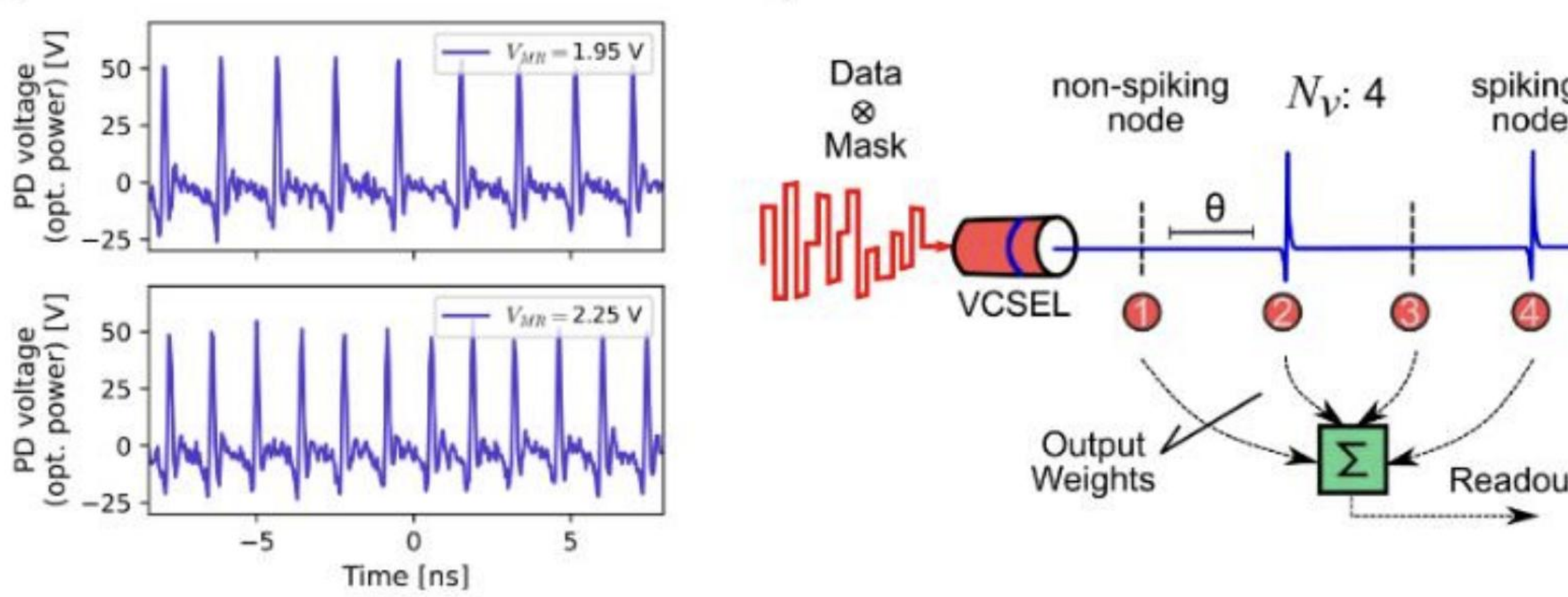
Nearly all approaches are analog but there are substantial differences:

- Just communication
- Reservoir computing
- Spiking neural networks
- Pure-GEMM
- Hybrid



And most of these still rely on substantial electronic hardware:

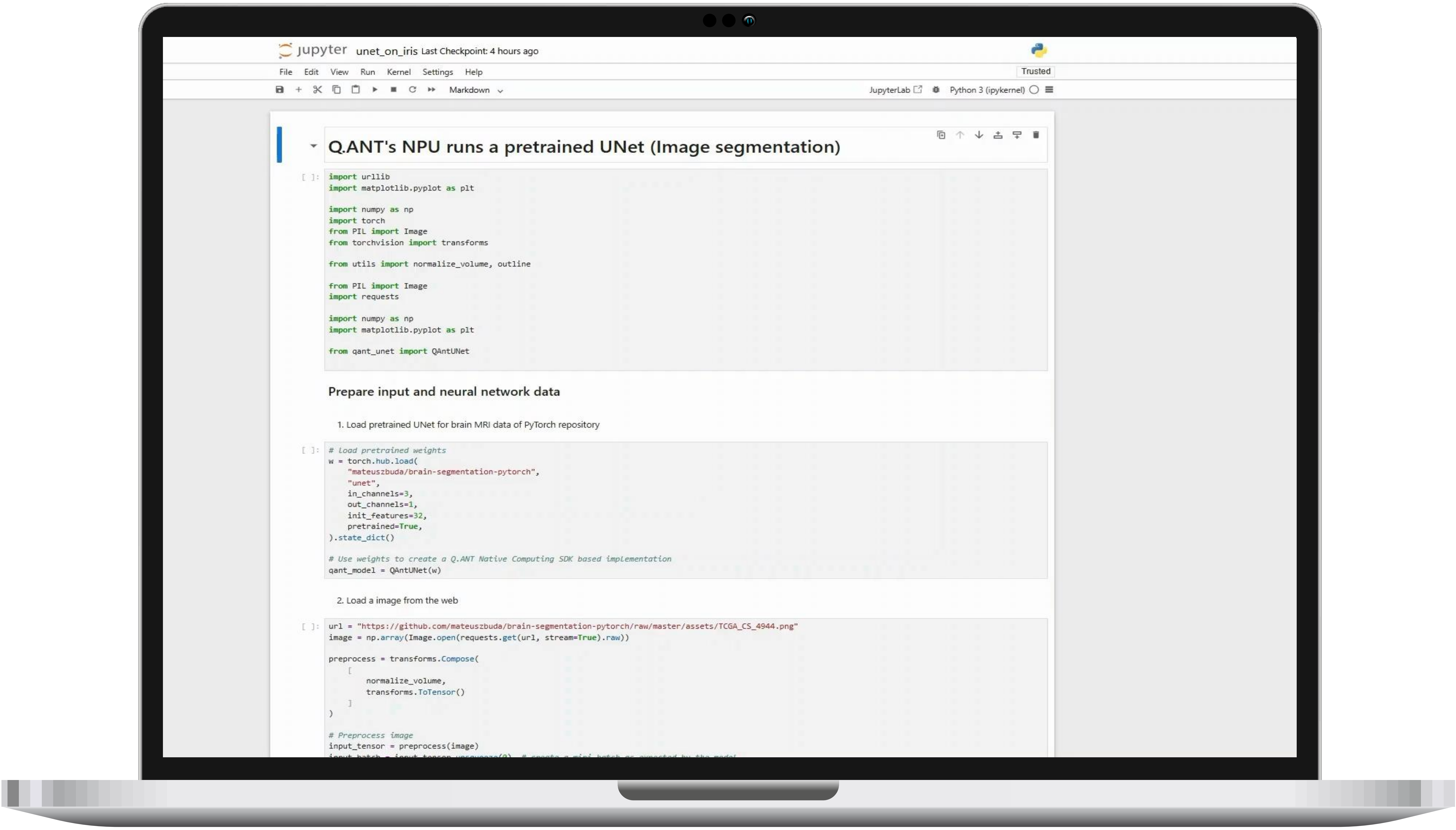
- Input and output
- Memory
- Clocking
- Control routines and programming



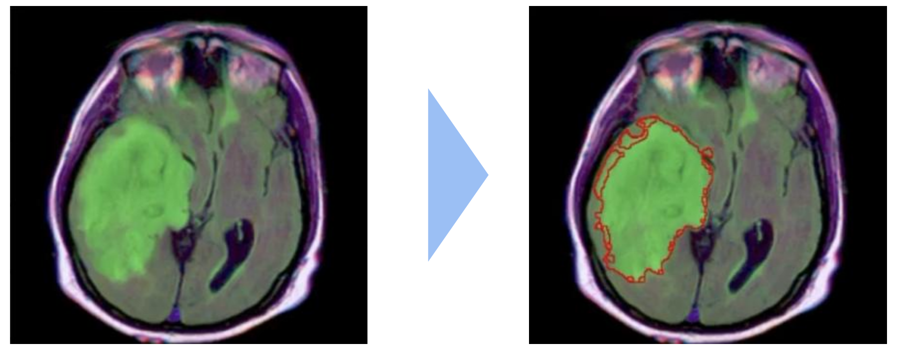
https://arxiv.org/pdf/2501.07917

Image segmentation demonstrates the capabilities of the NPS

Image segmentation on NPU Gen. 1 (100 MOPS) using U-Net:

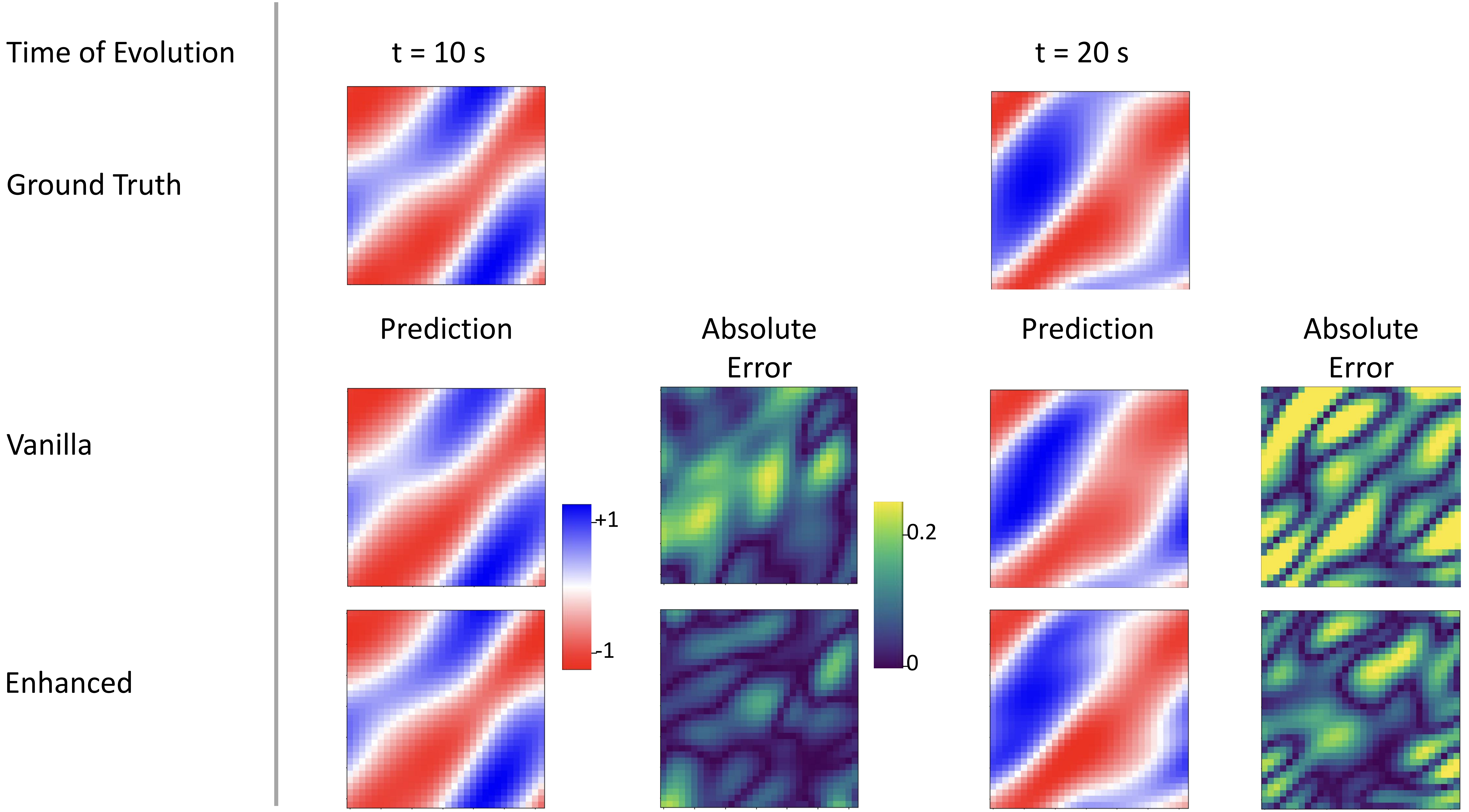


- Image segmentation of a tumor in a brain MRI scan



- Using the 'U-Net' convolutional neural network architecture, with 5.4×10^{10} MAC operations performed on Q.ANT NPU Gen. 1
- Error propagation of analog compute is not relevant
- AI based on photonic compute is feasible, with same performance at lower energy and less footprint

More accurate prediction of Navier-Stokes time evolution



Beyond technical issues there is the market

- There is a strong dominance in the GPU/accelerator market today
- Software stack and Open Source
  - Potential for integration of different solutions into one stack
  - Libaries enable integration without a full blown approach like CUDA

(Image: Julian Holzwarth / HLRS)

Outlook for the future of photonics computing

- Photonics does play a major supporting role for computing already
- Q.ANT works on making photonics one of the main actors for efficient computing
- We have first products aimed at R&D and early adopters
- We will release new hardware in the next year
- We are working hard to increase the depth and complexity of computing possible in photonics

Q.ANT