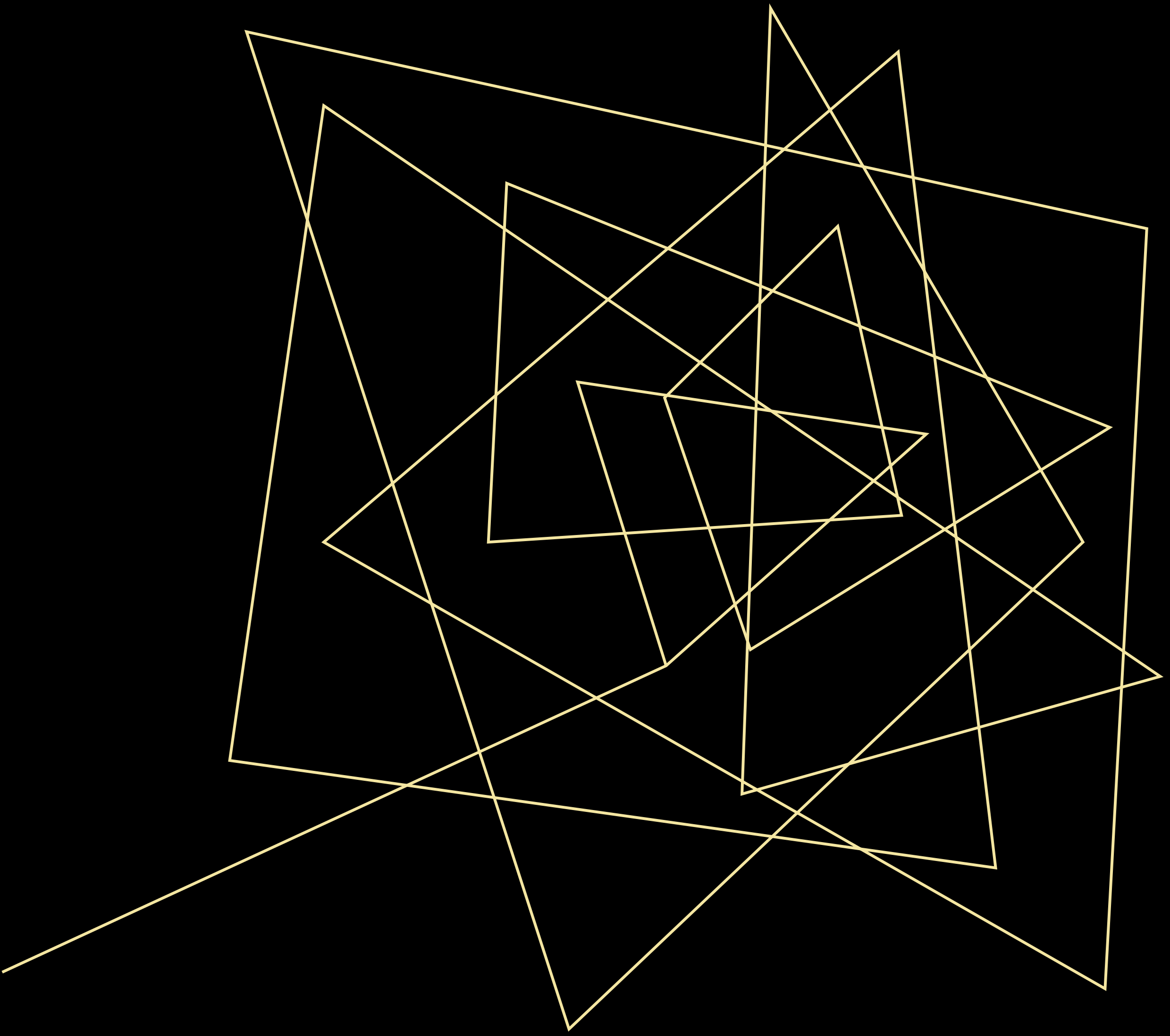
SALIENCE LABS

ENABLING ULTRA-LOW LATENCY EDGE PROCESSING WITH SILICON PHOTONICS

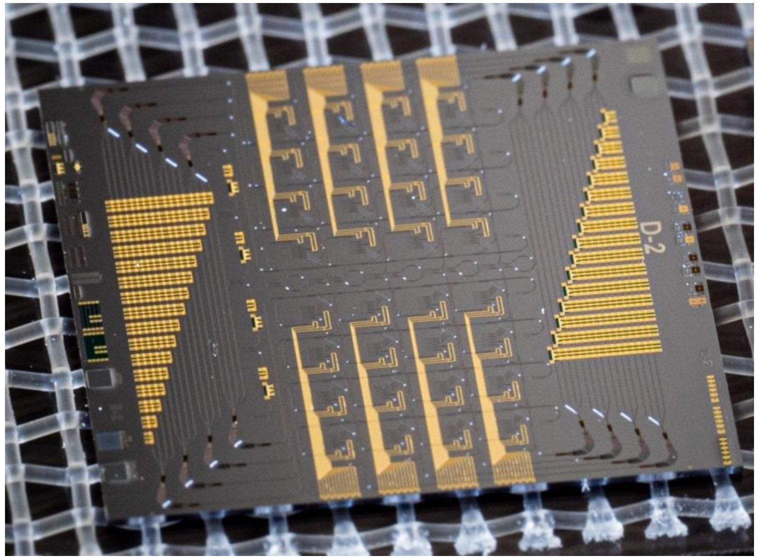
Johannes Feldmann – Salience Labs



WHO, WHAT & WHY?

WHO WE ARE

- Photonic computing and signal processing
- Founded in 2021
- Research from Oxford and Münster University
- Team of 18
- Based in Oxford, UK
- Ultra-low latency AI inference



THE TEAM

Vaysh, CEO

McKinsey & Company

Johannes, CTO

UNIVERSITY OF OXFORD

Chris, VP of Eng

GRAPHCORE NVIDIA

Enzo, Chief Architect

HUAWEI Imagination

Andy, SW Architect

SONY

Joanna, CFO

Mark, Senior SW Eng

Nat, ML Eng

UNIVERSITY OF CAMBRIDGE

Nick, ML Eng

SAMSUNG

Yi-Ling, ML Eng

Imperial College London

Vasileios, Photonics Eng

UNIVERSITY OF Southampton

Gary, Photonics Eng

Optalysys

Jaganath, Principal Analog Eng

intel

Andres, Principal Hardware Eng

INSPACE

Rob, Ops Manager

ULTROMICS

Lakshmi, Lead Verification Eng

cadence

Olufemi, Lead RTL Eng

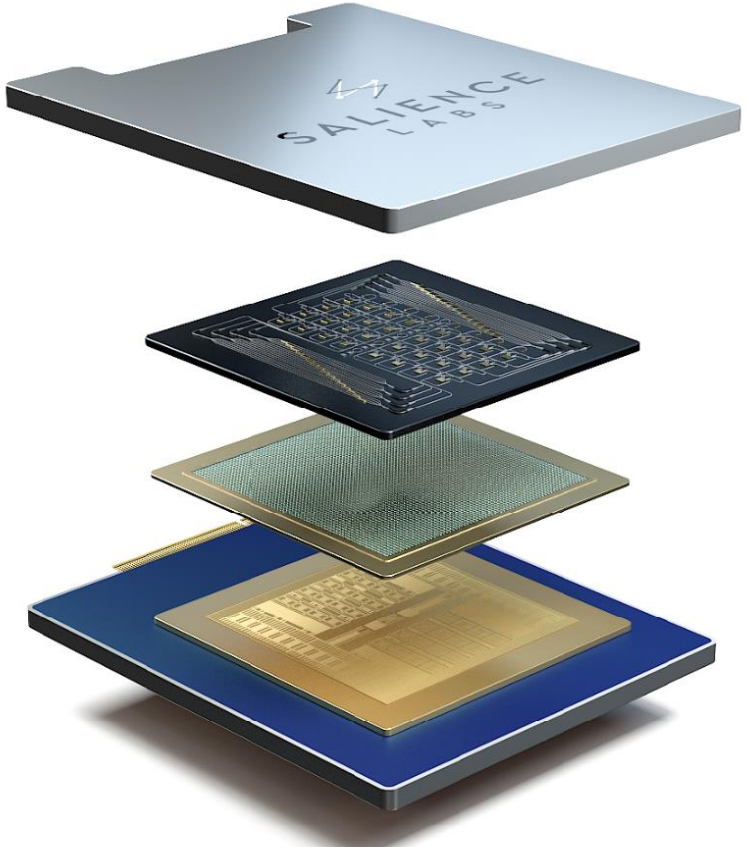
XILINX

Javaid, FPGA Eng

Qualcomm

WHAT WE DO

- AI inference at low latency and low power
- Signal processing: fourier transformation, matrix inversion
- Pattern recognition
- Optical data transfer / interconnects



Ultra-low latency optical compute

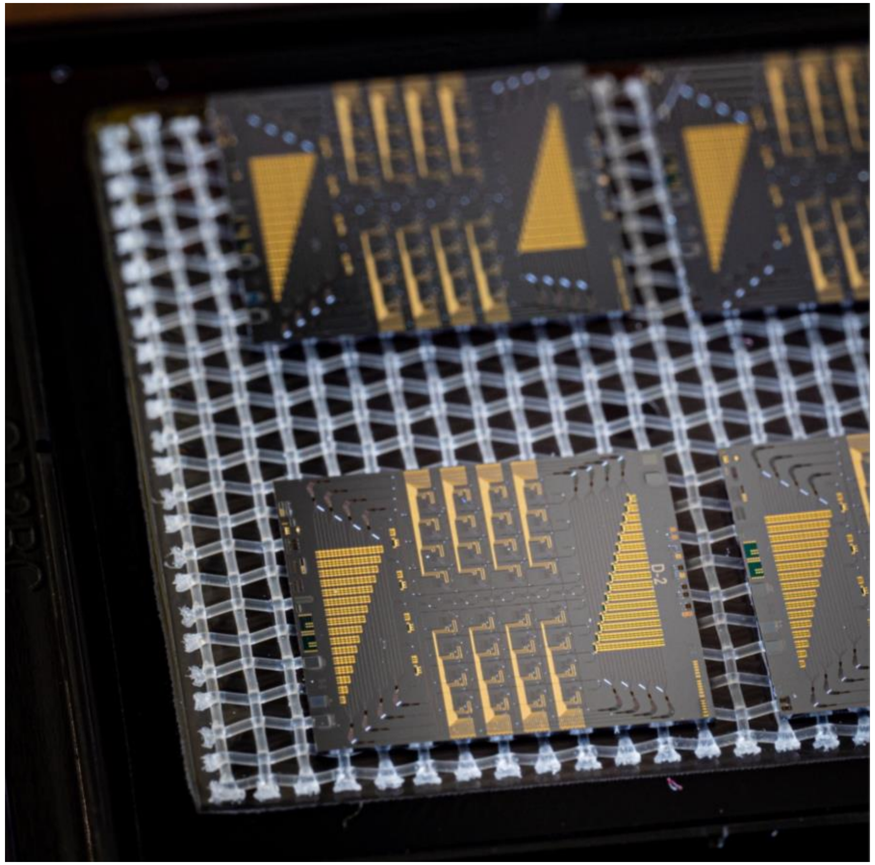
WHY PHOTONIC COMPUTING

SPEED — Run at 10-100 GHz, full matrix vector multiplication in a single clock step

PARALLEL PROCESSING — Many vectors in a single shot on different wavelengths of light

SIMPLICITY — 1-1 mapping of a matrix, no bandwidth limiting components

Efficient low latency compute

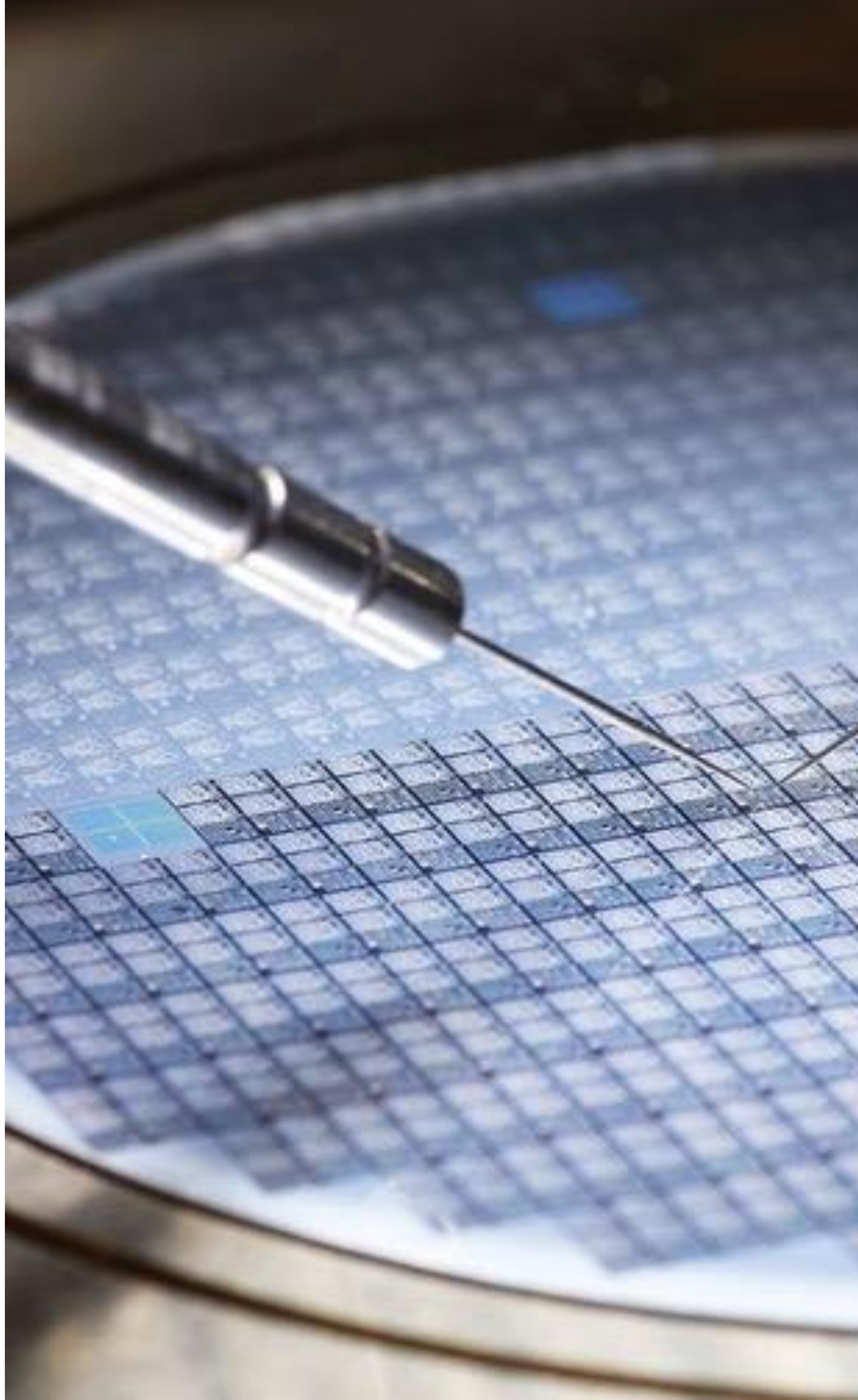


PHOTONICS VS ELECTRONICS

Parameter	Photonics	Electronics
Speed	Up to 100 GHz	Up to a few GHz
Power	Linear operations „for free“ (BUT: EO conversions)	Cost of switching capacitors and leakage currents
Parallelization	Wavelength multiplexing Mode multiplexing Polarization multiplexing Duplication of cores	Via duplication of cores
Footprint	µm scales High compute densities via speed and parallelization	nm scales
Scalability	Modulation speed More wavelengths MAC unit size No need for 3 nm technology	Going to smaller technology nodes expensive reaching physical limits

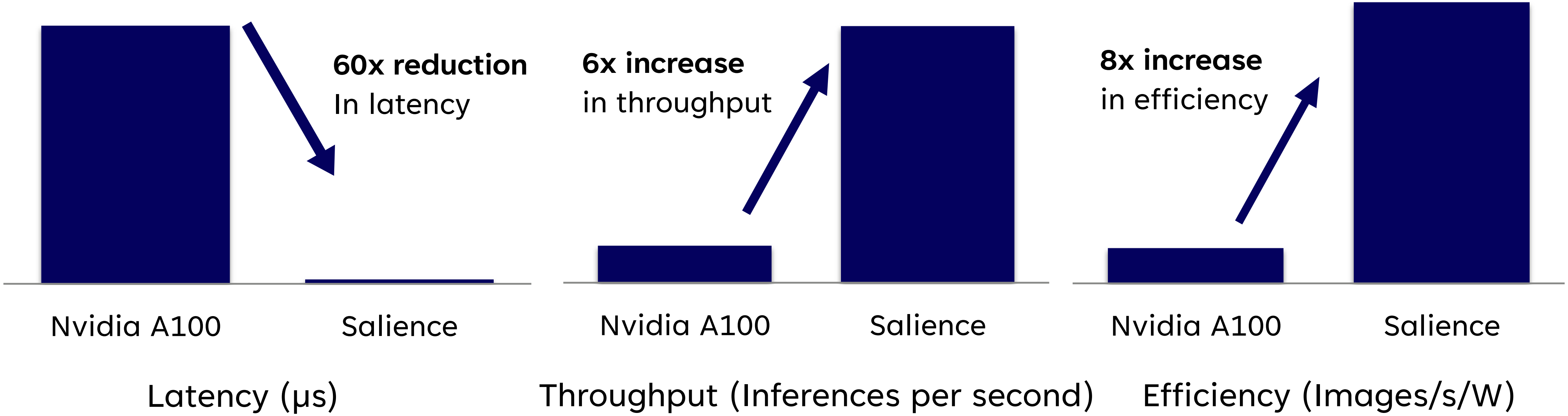
WHY NOW?

- Standard CMOS processes
- Volume manufacturing available now
- Full integration of lightsource, modulators and detectors
- Application specific, not general compute



LOW LATENCY, HIGH THROUGHPUT & EFFICIENCY

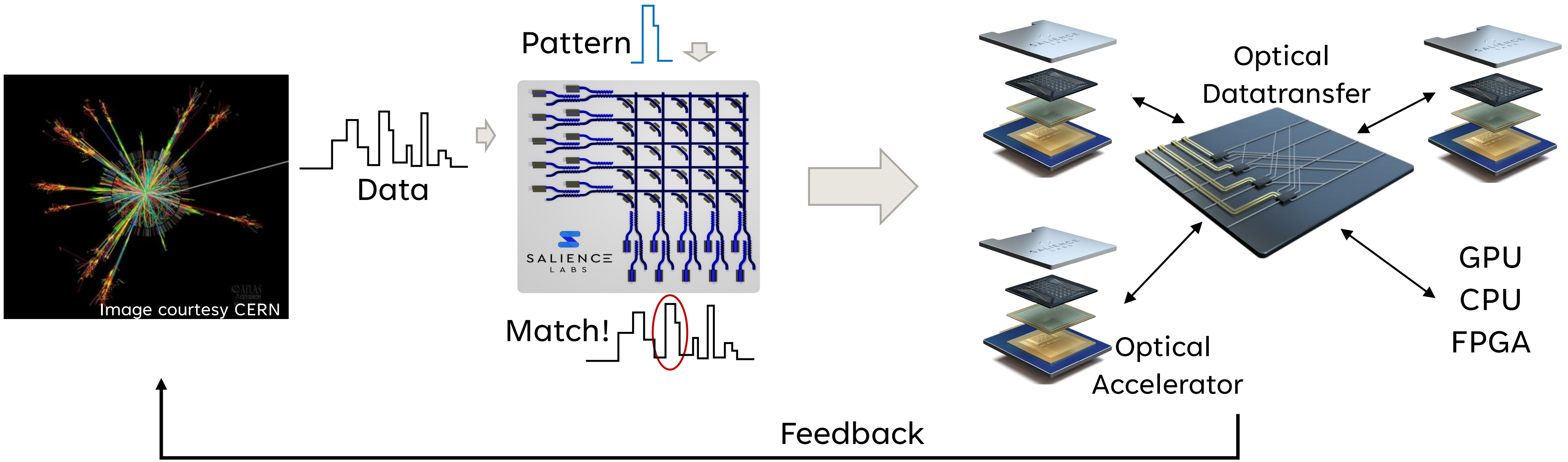
Estimated performance on Salience chip for ResNet 50 on ImageNet database



APPLICATION AREAS

- Ultra-low latency image recognition
- Error corrections / signal cleaning
- Ultra-low latency signal processing
- Nanosecond pattern recognition & correlation detection
- Low precision matrix math

EXAMPLE: DETECTION SYSTEM



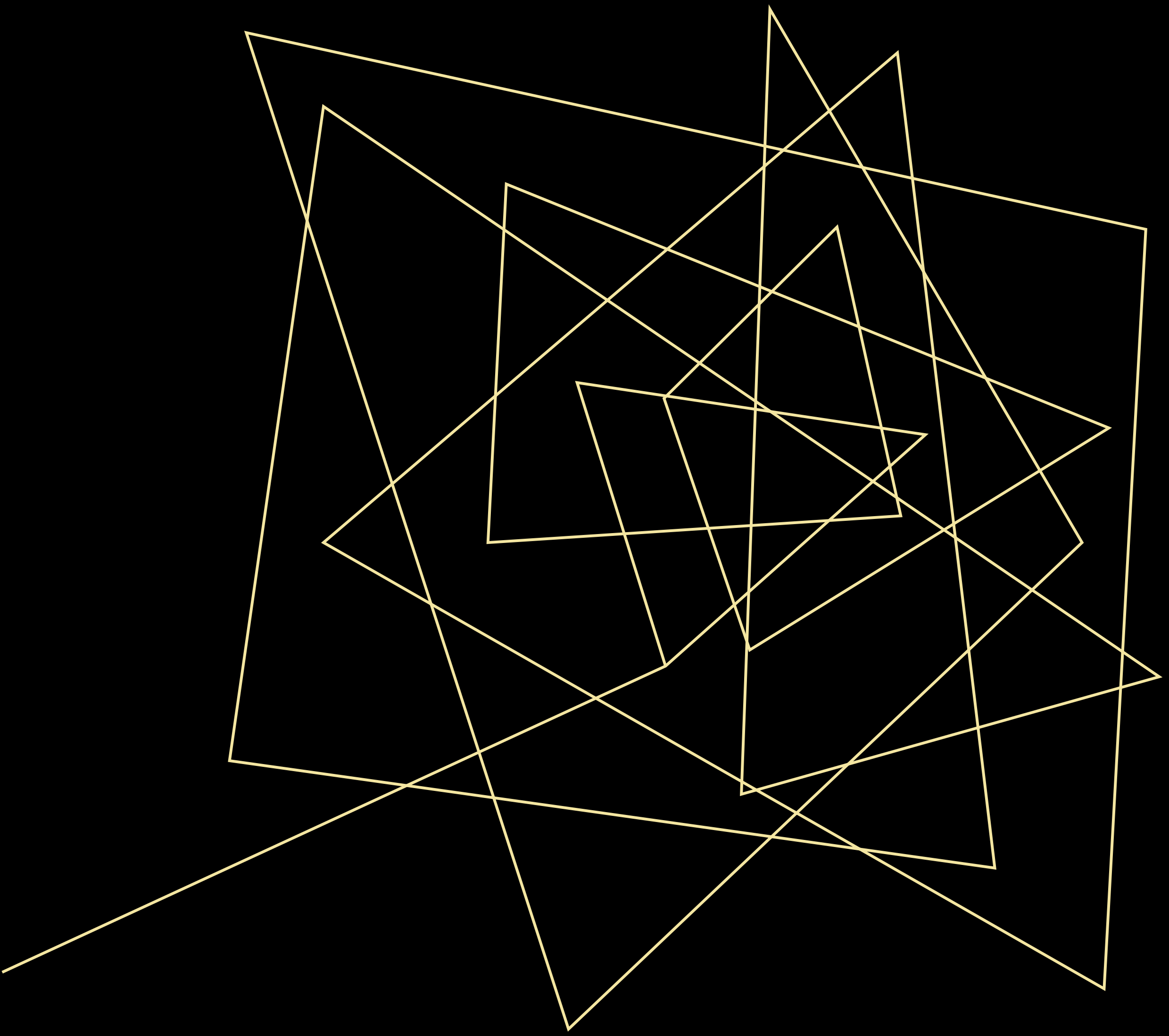
Collision

Pattern recognition

- Detect trigger signal
- Low latency (<1 ns)
- Start processing
- Start storing data

AI Inference cluster

- Multiple compute cores
- Optical interconnect for high bandwidth
- Low latency analysis
- Noise reduction
- Feedback loop

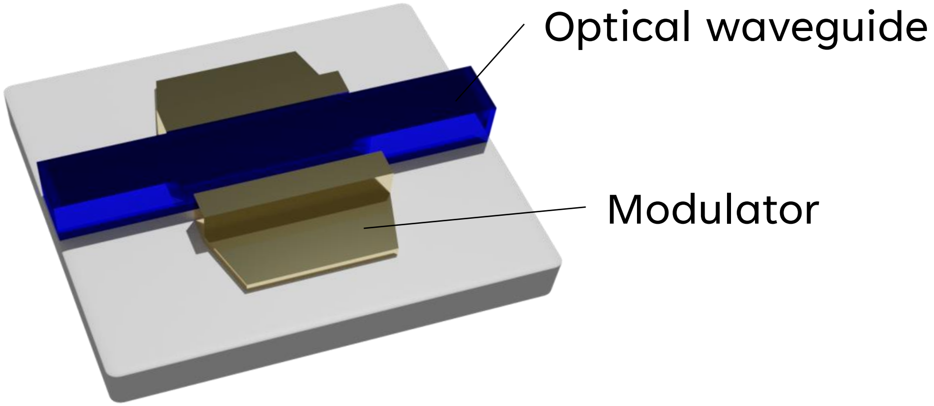


HOW DOES IT WORK?

SALIENCE LABS

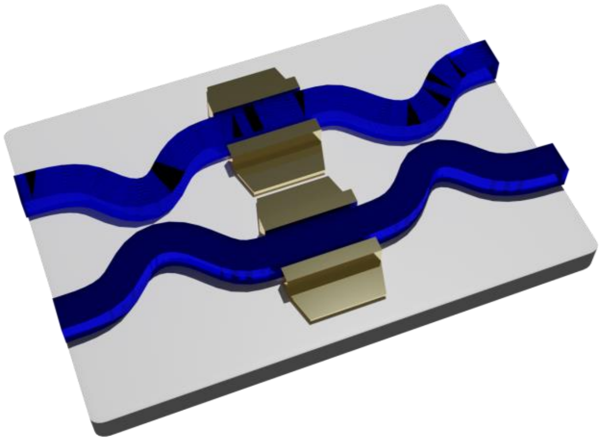
AMPLITUDE VS PHASE

Amplitude modulation



- Multiplication via attenuation
- Not phase sensitive
- No need for coherent light
- Robust

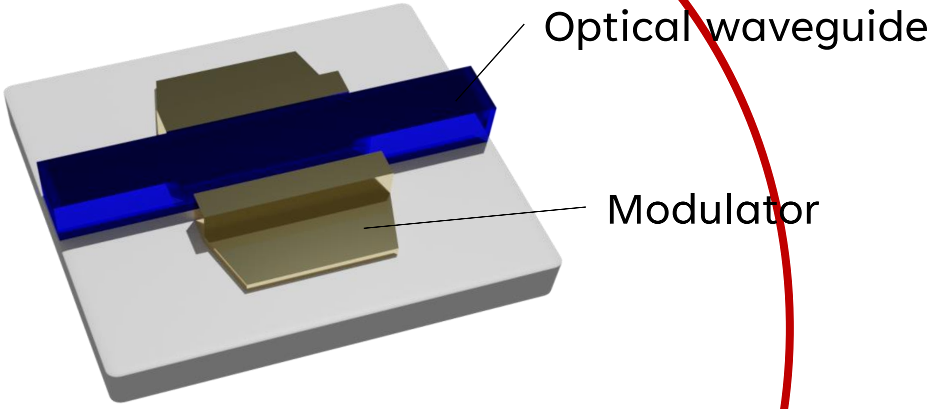
Phase modulation



- Interferometer performs rotations
- Exploits optical interference
- Needs coherent light
- Sensitive to variations

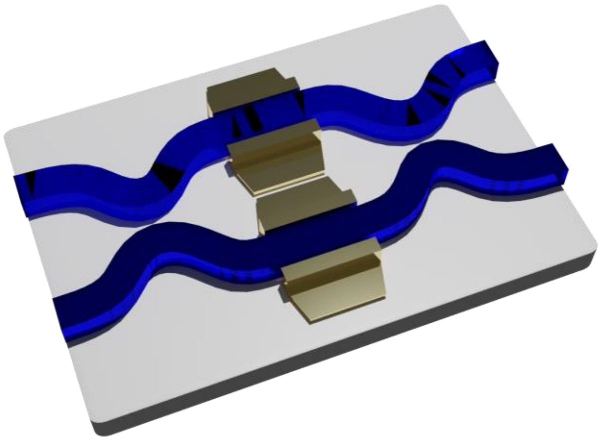
AMPLITUDE VS PHASE

Amplitude modulation



- Multiplication via attenuation
- Not phase sensitive
- No need for coherent light
- Robust

Phase modulation



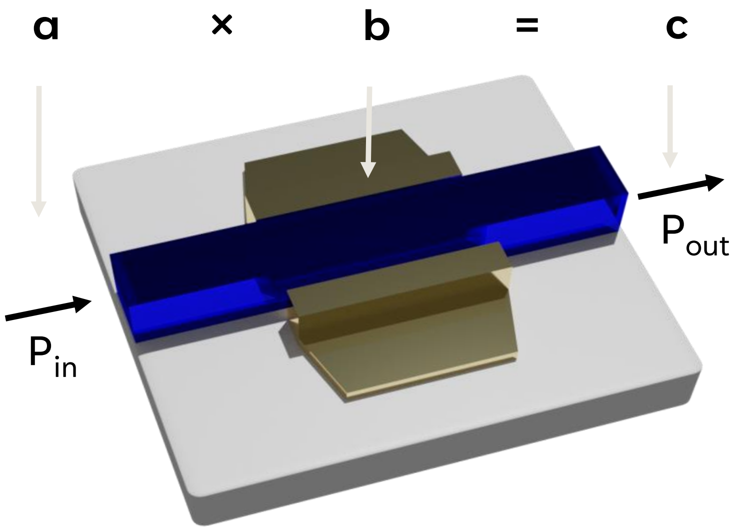
- Interferometer performs rotations
- Exploits optical interference
- Needs coherent light
- Sensitive to variations

DIFFERENT CONCEPTS AND APROACHES

Parameter	Salience, Amplitude	Phase (MZI arrays)	Free space	Ring resonators
Speed	Up to 100 GHz	Few GHz	Typcially kHz- MHz	Up to 100 GHz
Power	Low	Medium	Very low	Medium (tuning)
Parallelization (WDM)	Easy!	Possible	Unlikely	Unlikely
Footprint	Small	Large	Very large	Small
Scalability	Very good: speed, parallelization and MAC unit size	Limited, due to difficult phase control and parallelization	Limited	Limited
Stability	Every component is broadband: high tolerance to variations	Optical phase is very temperature sensitive	Difficult due to alignment and stability of free space parts	Every resonator needs thermal tuning of the sharp resonances to the correct wavelengths

PHOTONIC MULTIPLY ACCUMULATE

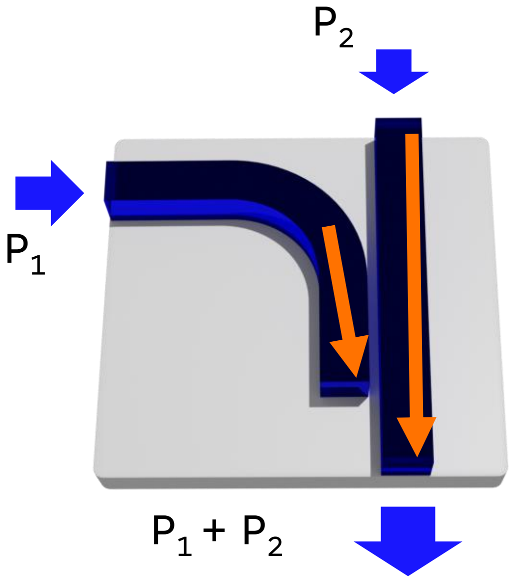
Multiplication



a: Amplitude of input light
b: State of modulator
c: Amplitude of output light

- Multiplication in (passive) transmission measurement
- Amplitude of light weighted by modulating element

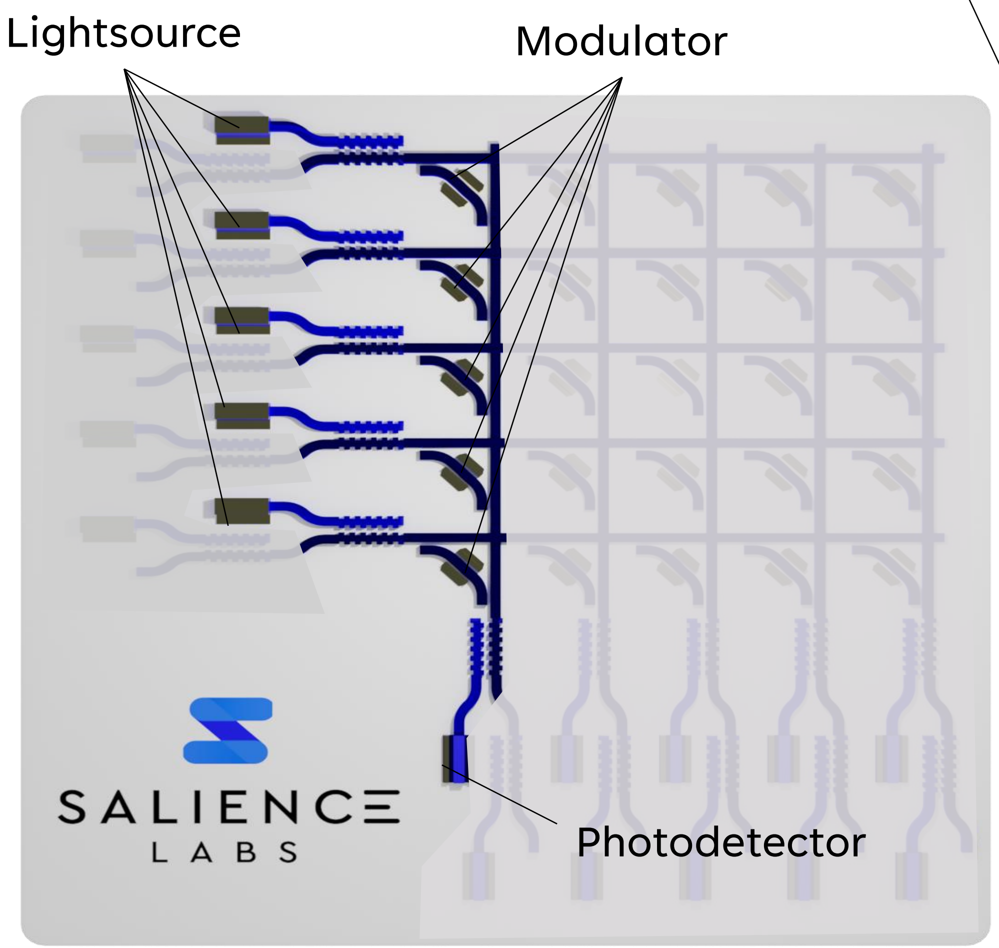
Addition



- Combine power from two waveguides into one
- Avoid interference by use of different wavelengths

PHOTONIC MATRIX MULTIPLICATION

- Combined MAC units calculate dot product: ab+cd+ef+…

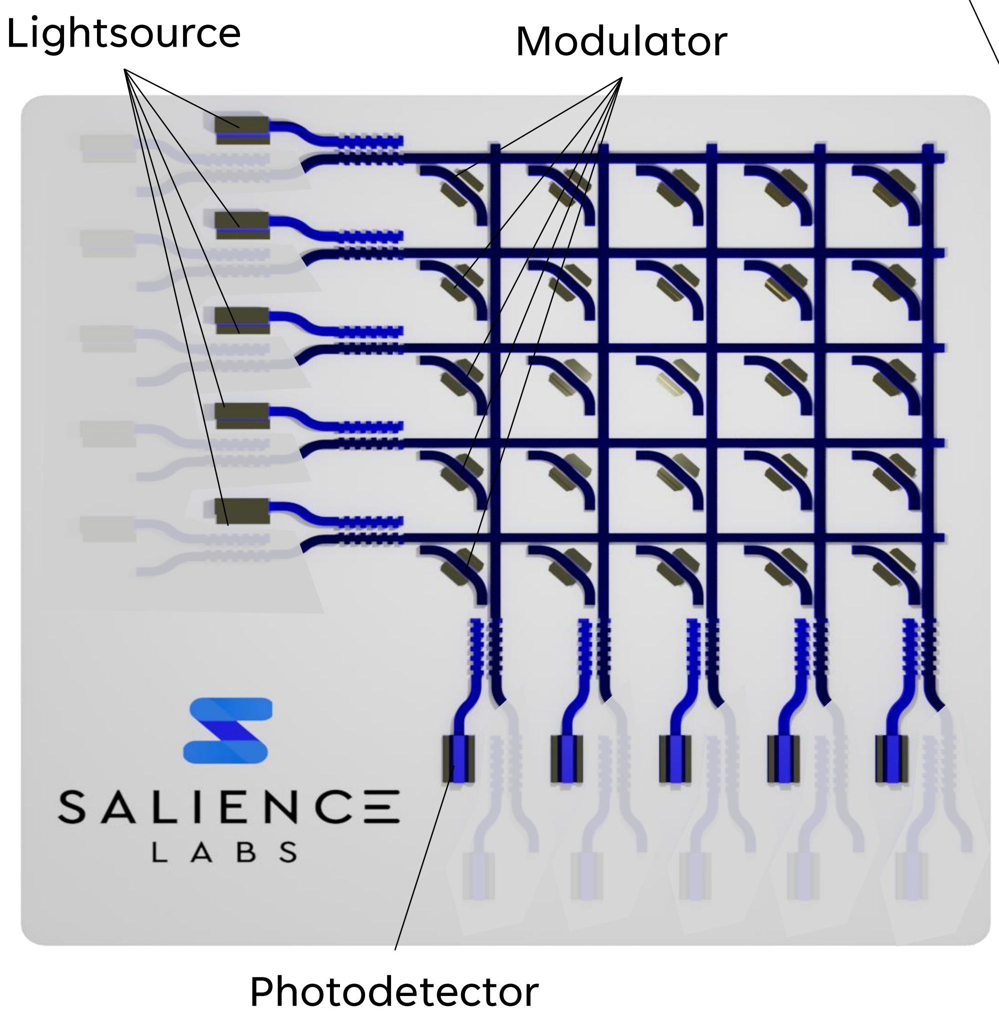


PHOTONIC MATRIX MULTIPLICATION

- Combined MAC units calculate dot product: ab+cd+ef+…
- Multiple columns

⇨ Full matrix vector multiplication

- Single time step (time of flight of the light)

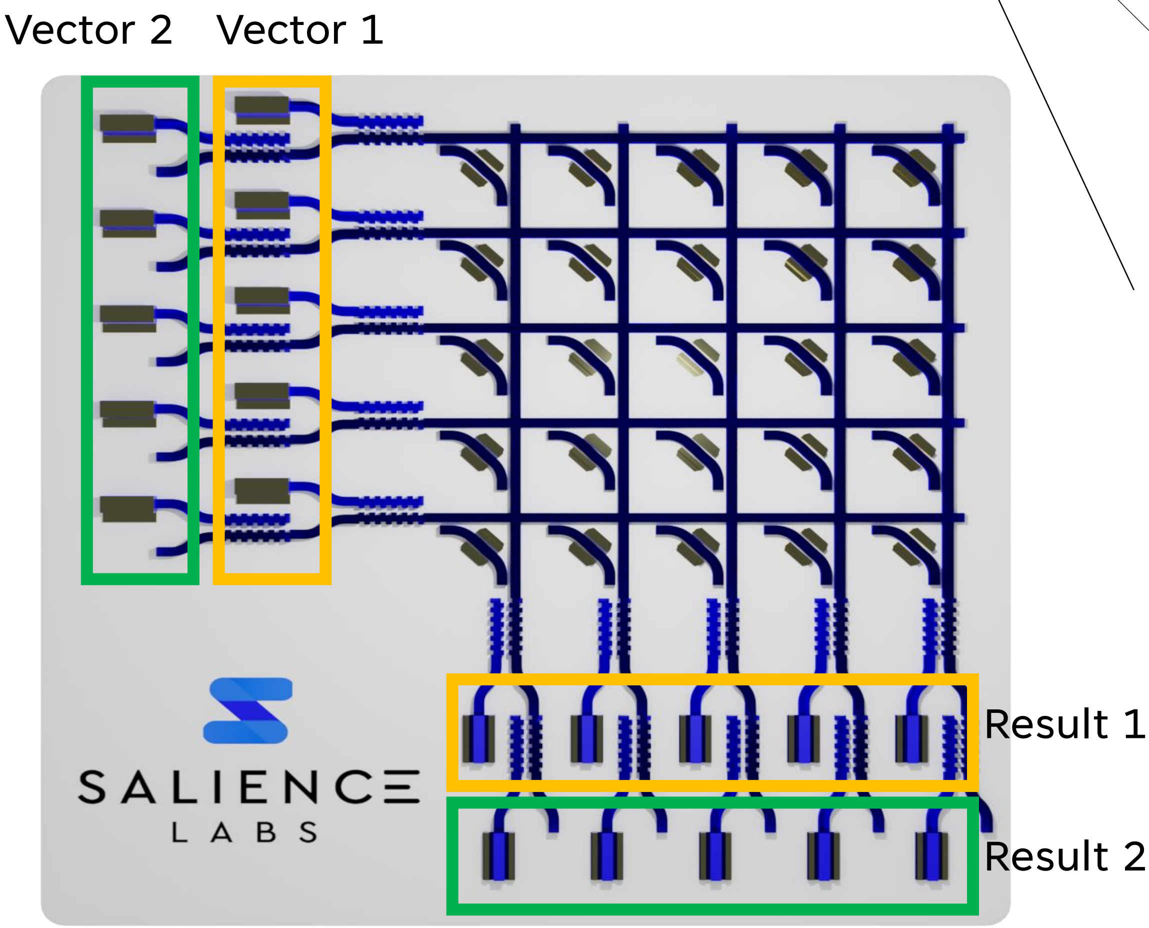


PHOTONIC MATRIX MULTIPLICATION

- Combined MAC units calculate dot product: ab+cd+ef+…
- Multiple columns

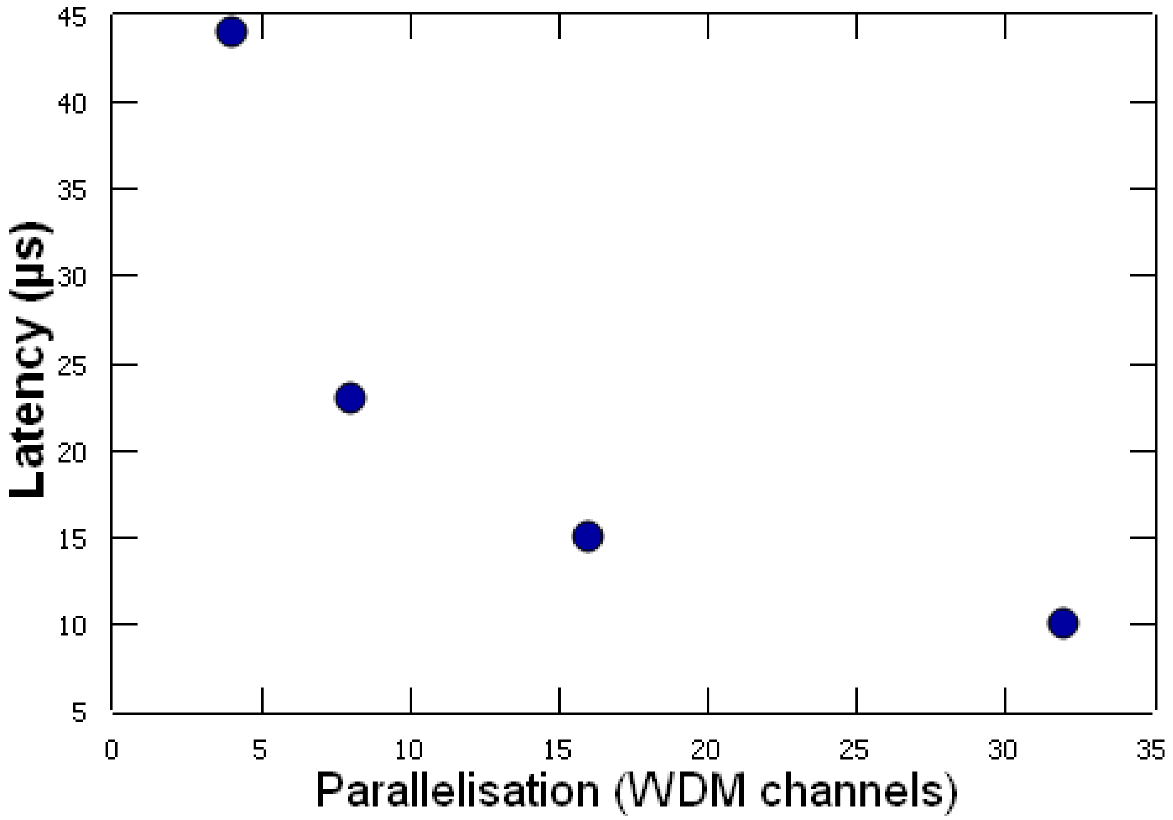
⇨ Full matrix vector multiplication

- Single time step (time of flight of the light)
- Increased compute density via wavelength division multiplexing



IMPACT OF PARALLELISATION

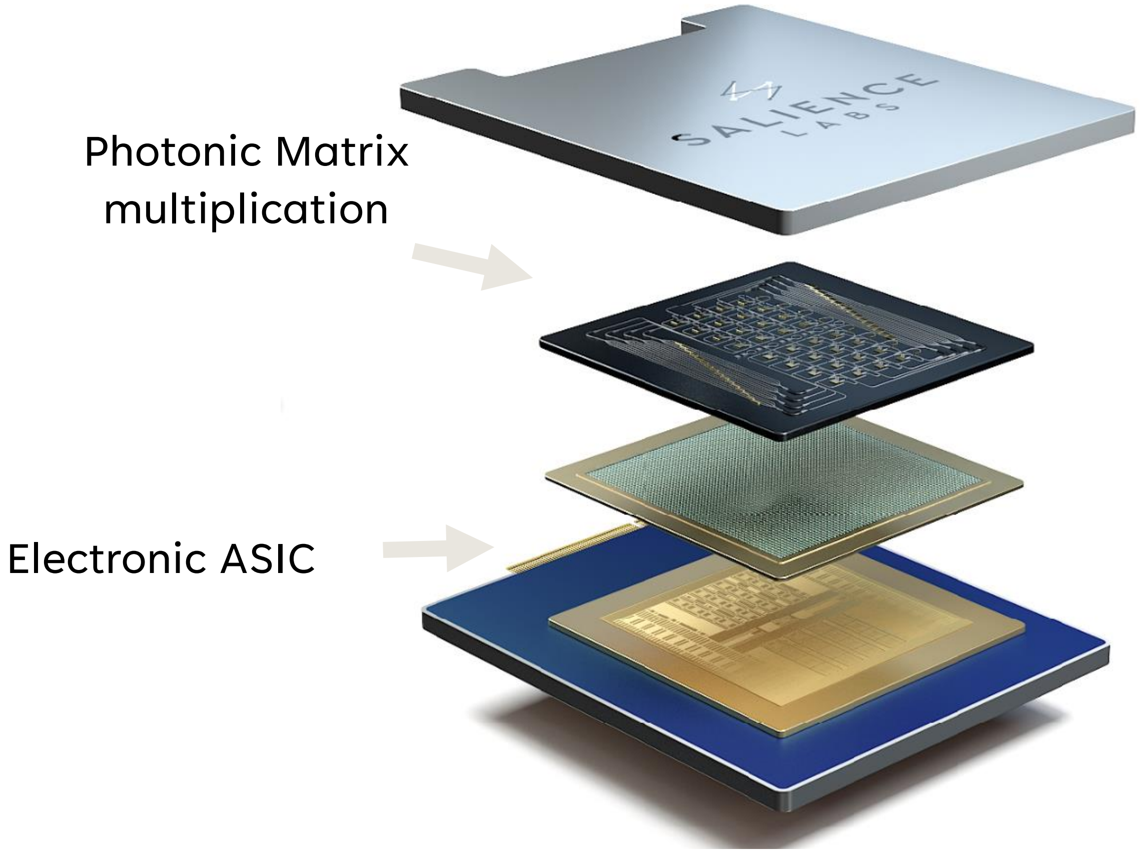
- Data given for Resnet50
- 45 µs is already fastest!
  GPU: ca. 600 µs
- Latency reduction by using multiple input vectors
- Unique to photonics
- Extra boost from extra colours!



THE FULL SYSTEM

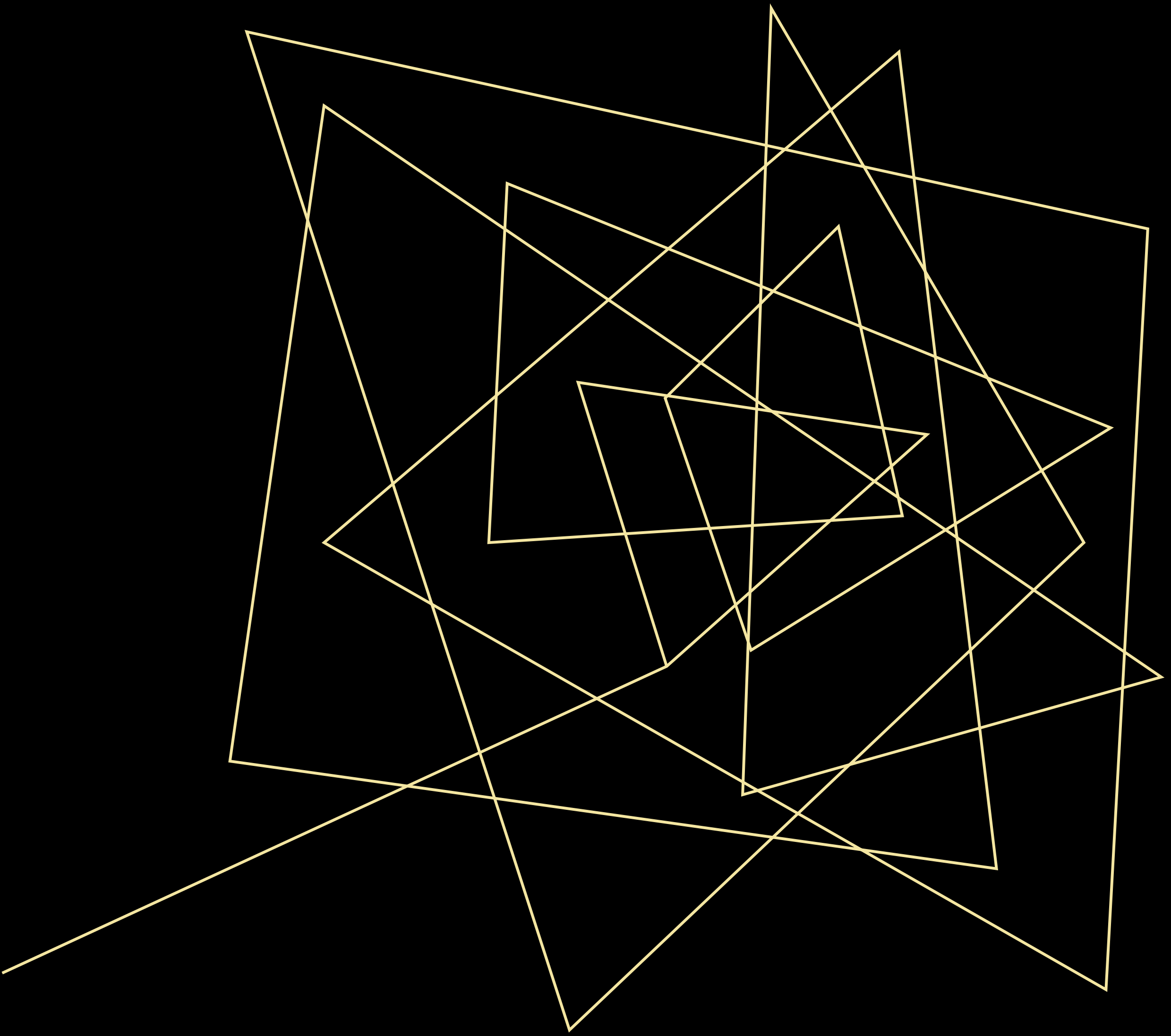
- Photonics: Fast matrix math
- Electronics: Control and nonlinearities
- Chiplet approach
- Standard interfaces:
  Digital electronic in & out

- Photonic interfaces possible



PRIME – SOFTWARE MODEL

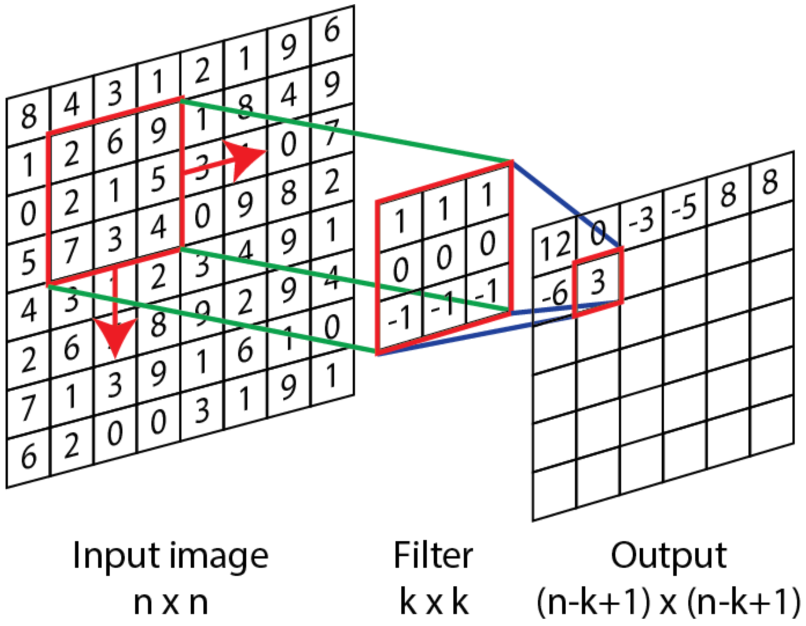
- Software model of the hardware
- Tool to evaluate larger processors before availability of silicon
- Benchmarks for different workloads
- 8 models implemented:
  Resnet50, 3D-Unet, RetinaNet, Beit-L,...
- Software interface: Tensorflow

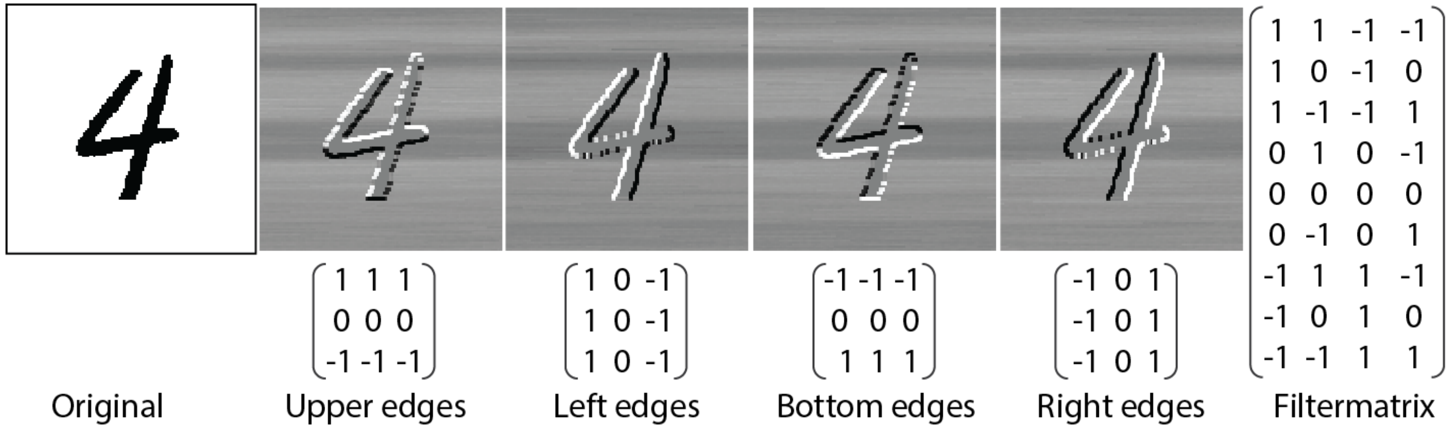


EXPERIMENTAL RESULTS

CONVOLUTIONS

- Scan filter across image
- Image filtering: edge detection, sharpening, smoothing
- Multiple filter kernels can be applied at the same time
- Convolutional neural networks: vision applications, image classification, pattern recognition
- High speed and high efficiency





$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$ Upper edges

$\begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$ Left edges

$\begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$ Bottom edges

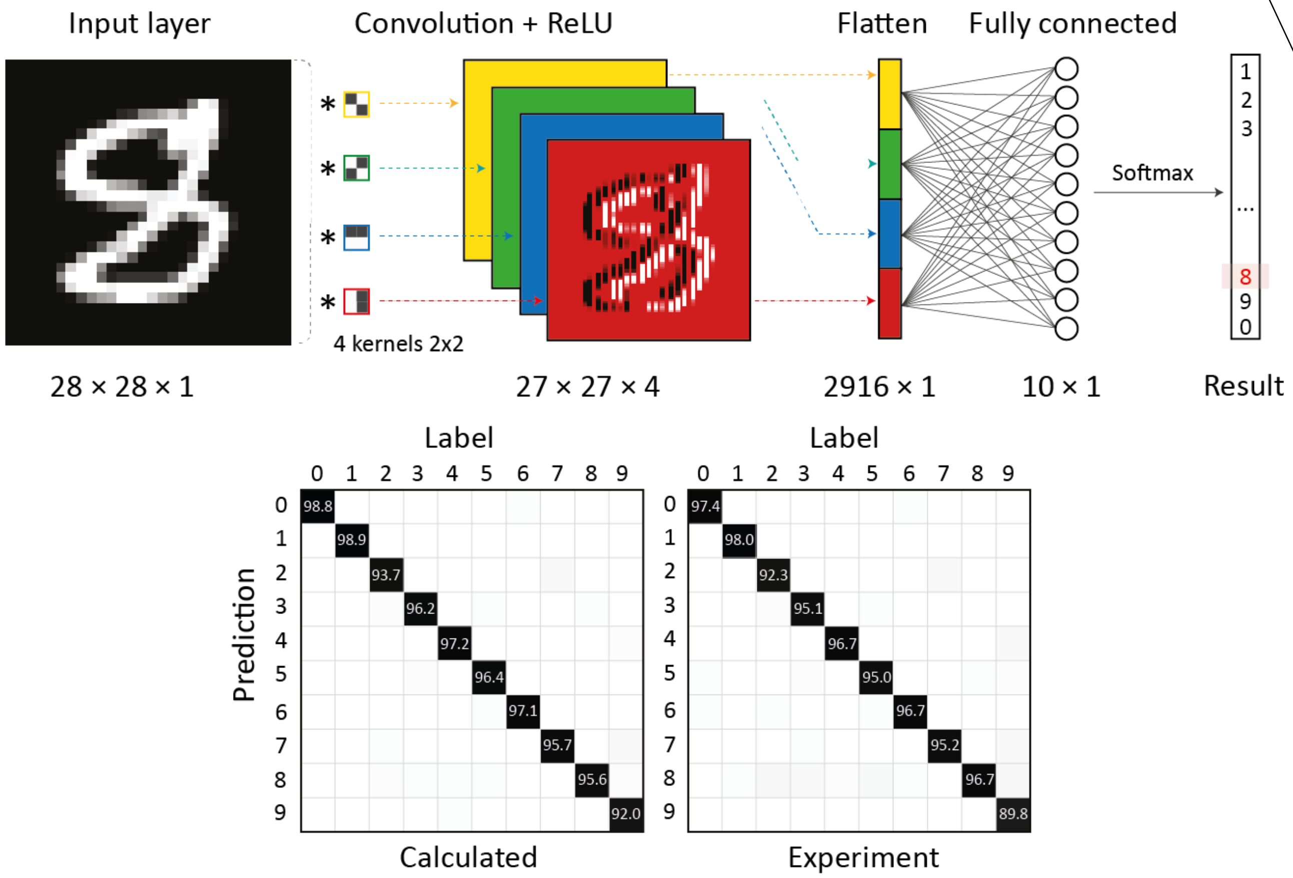
$\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}$ Right edges

$\begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 0 & 1 & 0 \\ -1 & -1 & 1 & 1 \end{pmatrix}$ Filtermatrix

Feldmann, Youngblood, Karpov et al. , Nature 2021, 589, 52-58.

NEURAL NETWORKS

- Handwritten digit recognition
- Simple CNN tested on MNIST database
- Experimental accuracy: 95.3 %
- Theoretical accuracy: 96.1%



Label

	0	1	2	3	4	5	6	7	8	9
0	98.8									
1		98.9								
2			93.7							
3				96.2						
4					97.2					
5						96.4				
6							97.1			
7								95.7		
8									95.6	
9										92.0

Calculated

Label

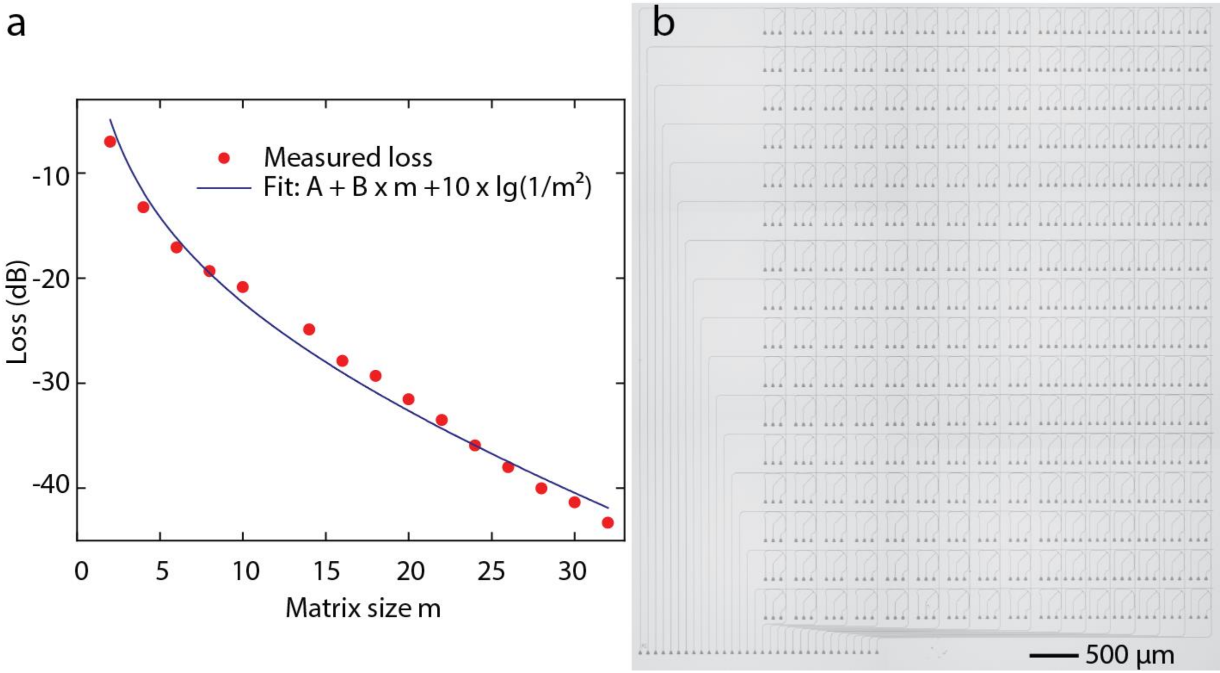
	0	1	2	3	4	5	6	7	8	9
0	97.4									
1		98.0								
2			92.3							
3				95.1						
4					96.7					
5						95.0				
6							96.7			
7								95.2		
8									96.7	
9										89.8

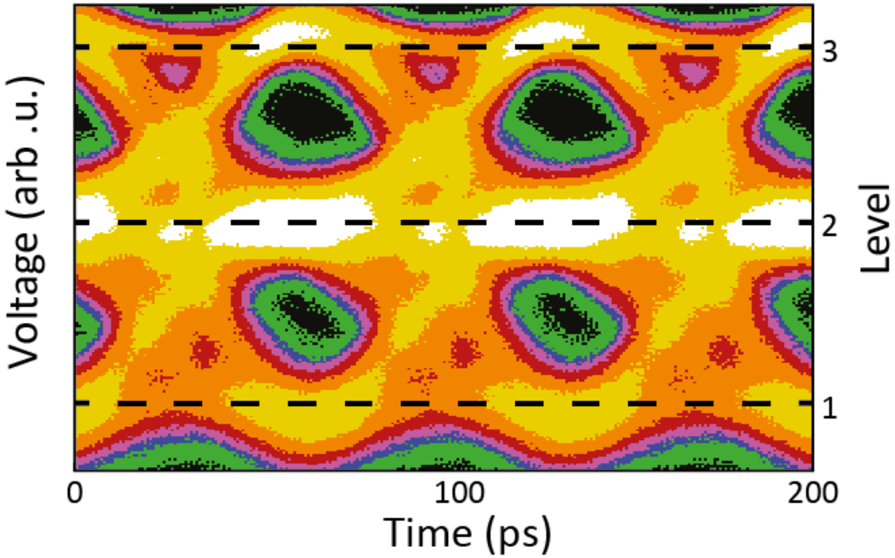
Experiment

Feldmann, Youngblood, Karpov et al. , Nature 2021, 589, 52-58.

SCALABILITY

- Photonics scales in different ways compared to electronics: MAC unit size, modulator speed, parallelisation
- No need for newest technology node!

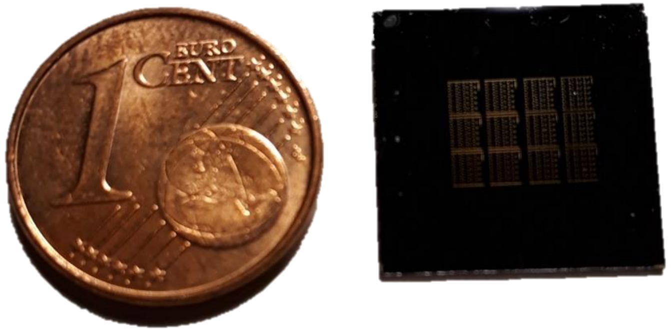




High speed matrix multiplication >13 GHz

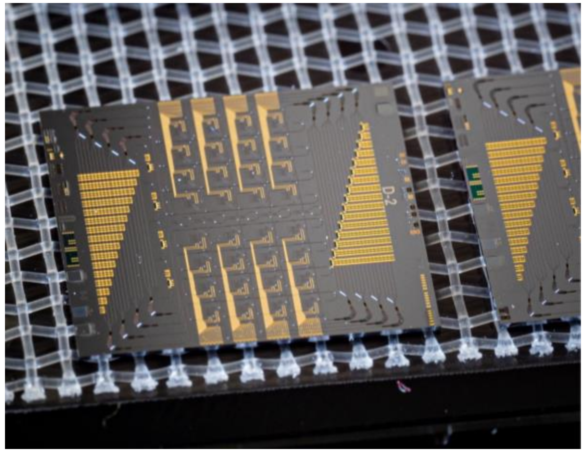
Feldmann et al., Nature, Vol 589, 7 January 2021.

SCALING PERFORMANCE

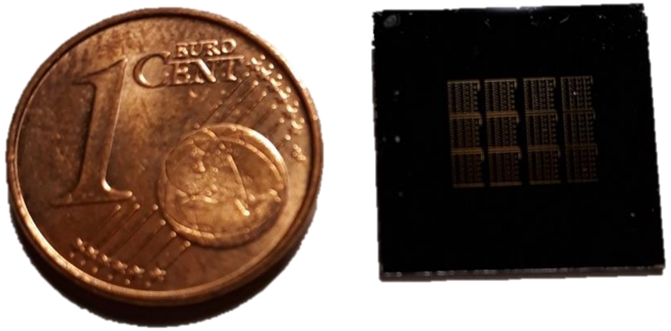


PROTOTYPE CHIPS

9x4 photonic matrix with FPGA
Multiplexing 4 vectors
Up to 14 GHz
Up to 32x32
Foundry compatibility



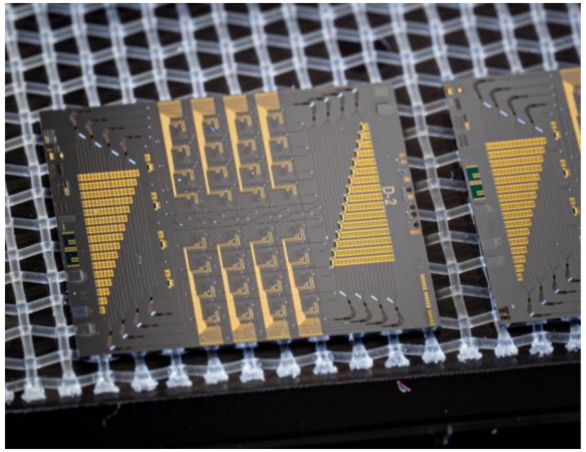
SCALING PERFORMANCE



PROTOTYPE CHIPS

9x4 photonic matrix with FPGA
Multiplexing 4 vectors
Up to 14 GHz
Up to 32x32
Foundry compatibility

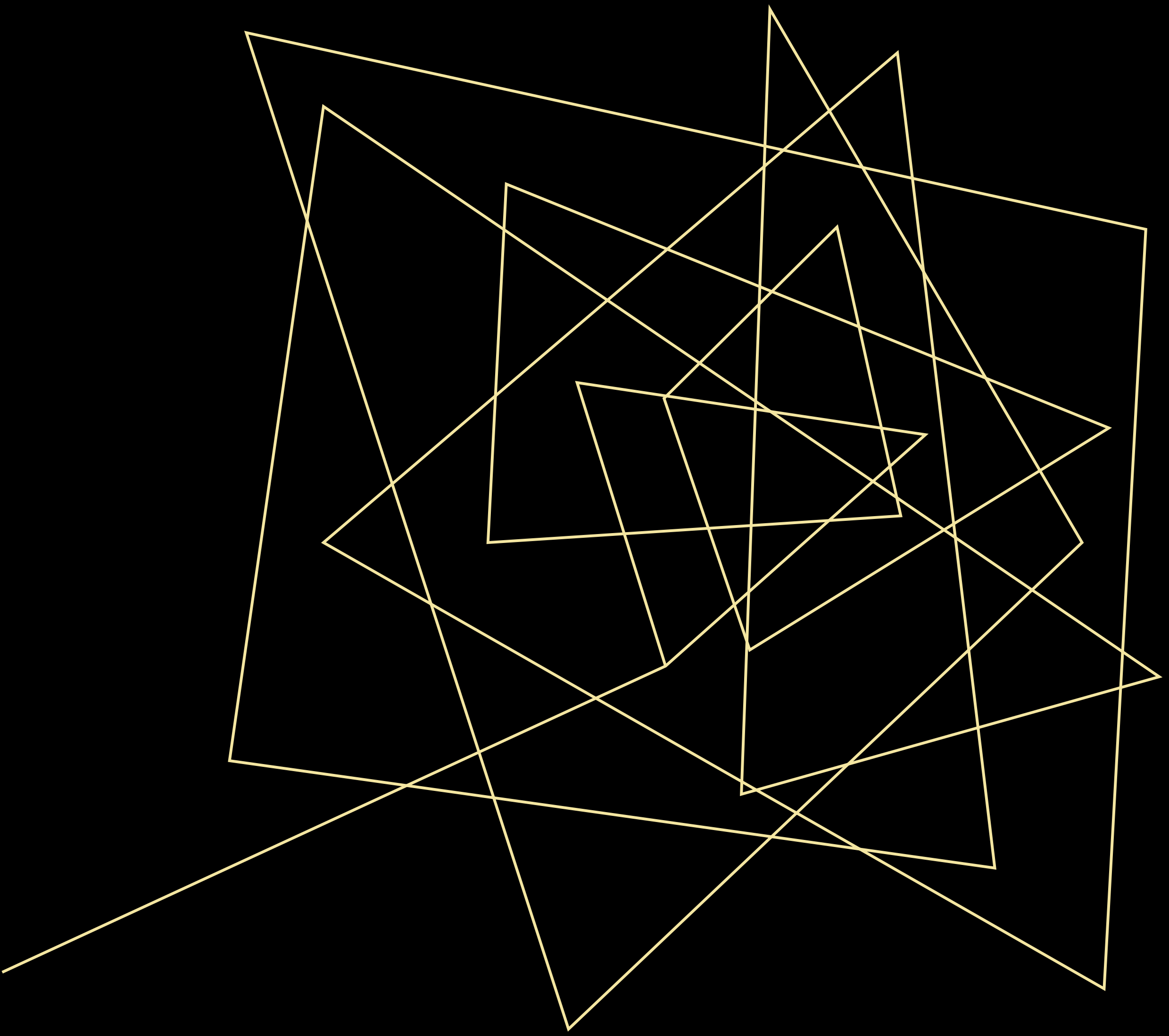
0.5 TOPs



SCALING

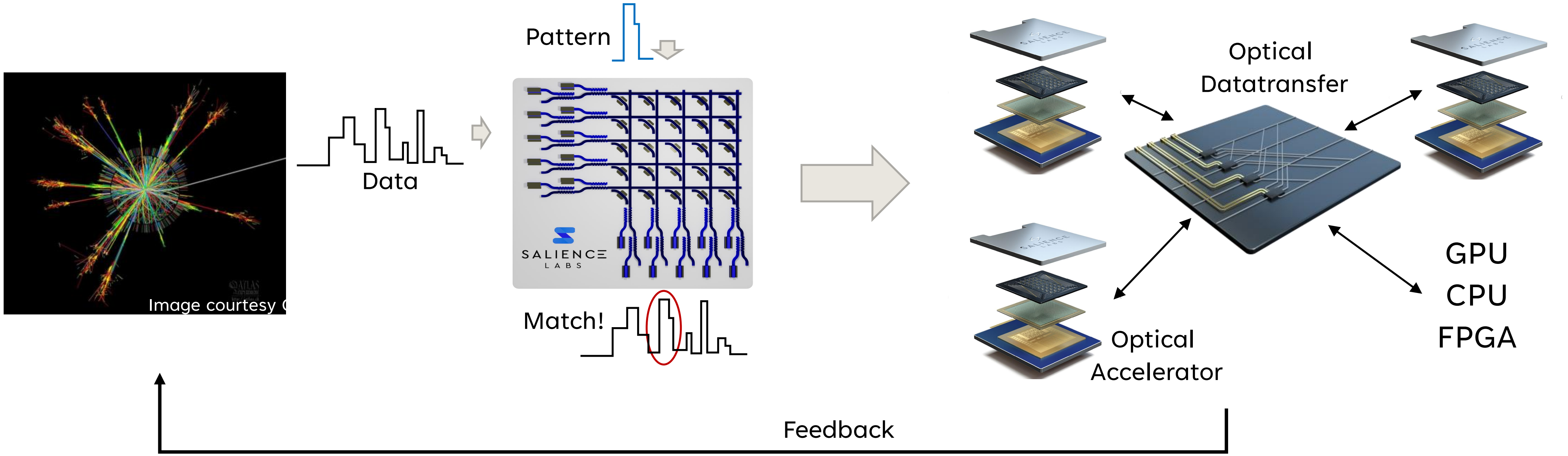
64x64
10 GHz
10 Vectors

1000 TOPs



APPLICATION EXAMPLES

EXAMPLE: DETECTION SYSTEM



Collision

Pattern recognition

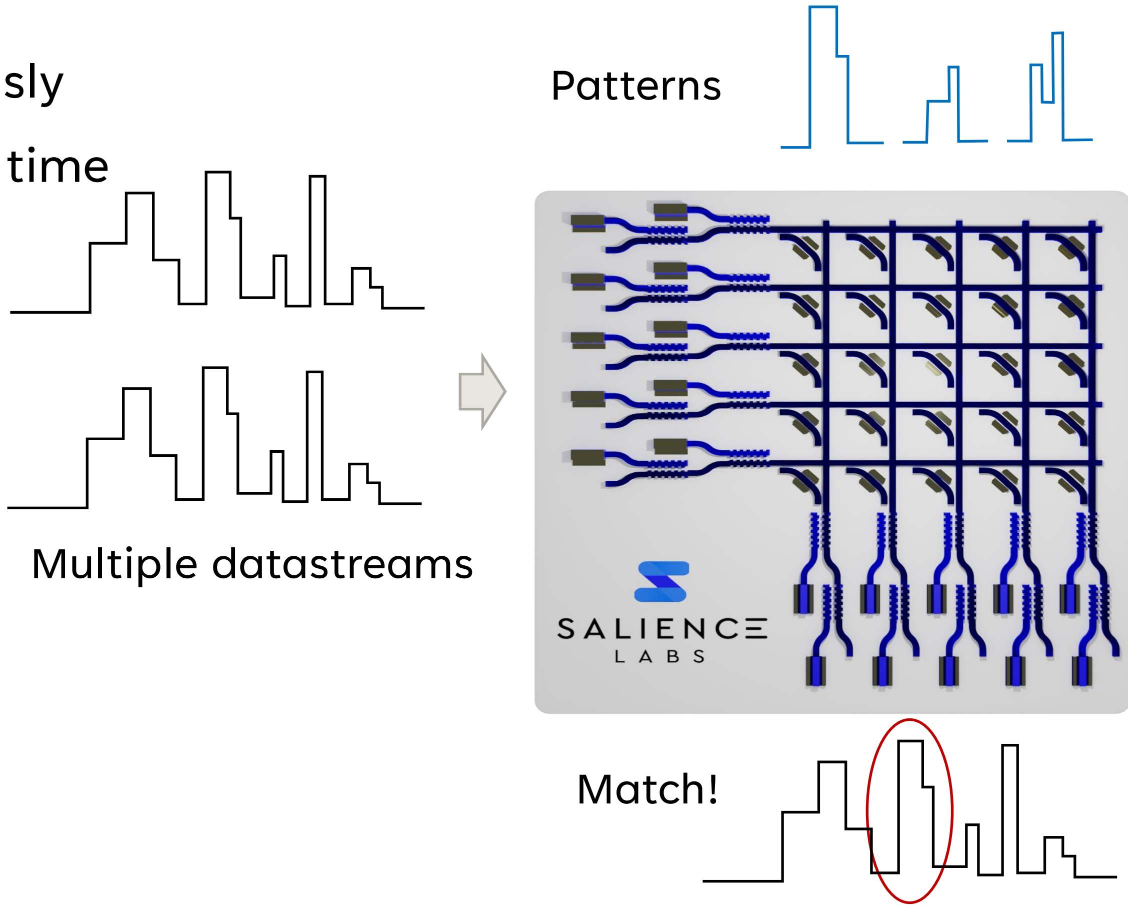
- Detect trigger signal
- Low latency (<1 ns)
- Start processing
- Start storing data

AI Inference cluster

- Multiple compute cores
- Optical interconnect for high bandwidth
- Low latency analysis
- Noise reduction
- Feedback loop

PATTERN RECOGNITION

- Check for multiple patterns simultaneously
- Down to tens of picoseconds evaluation time
- Noise tolerant evaluation



Patterns

Multiple datastreams

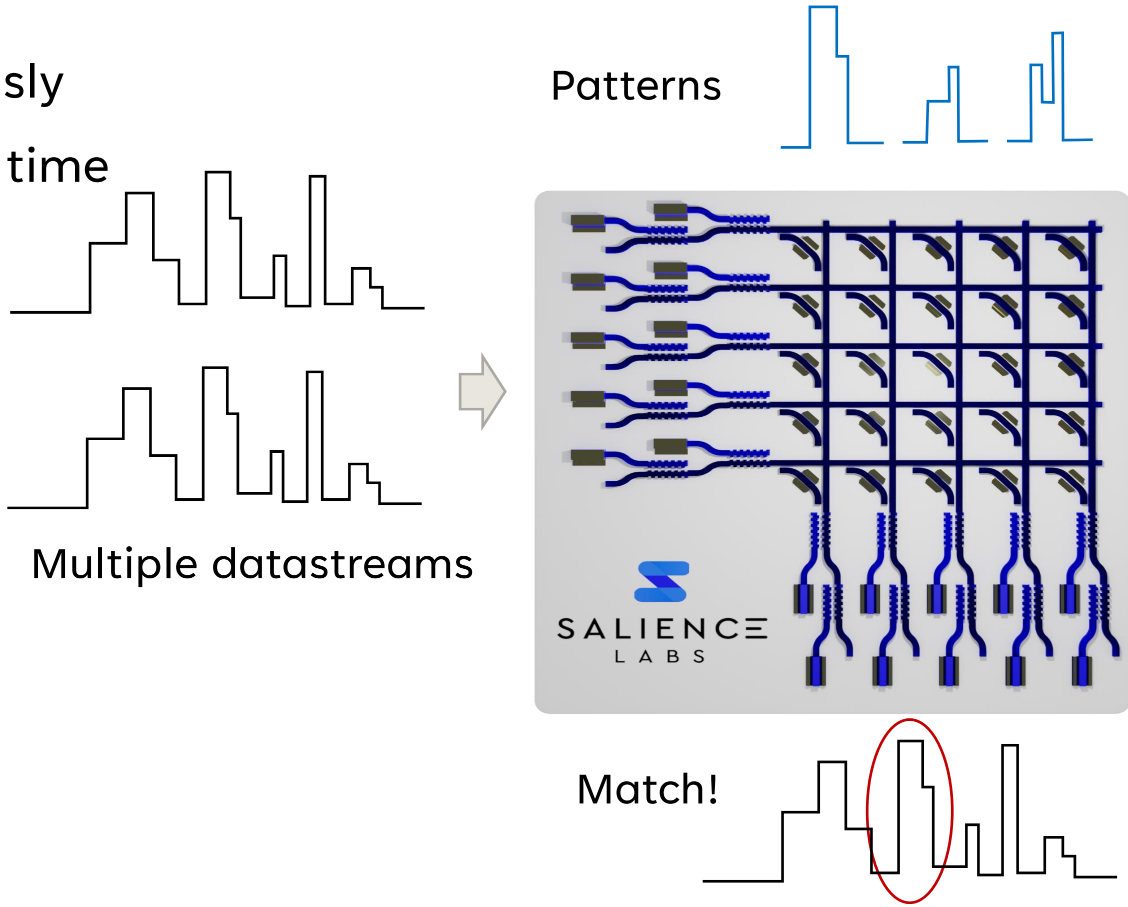
Match!

PATTERN RECOGNITION

- Check for multiple patterns simultaneously
- Down to tens of picoseconds evaluation time
- Noise tolerant evaluation

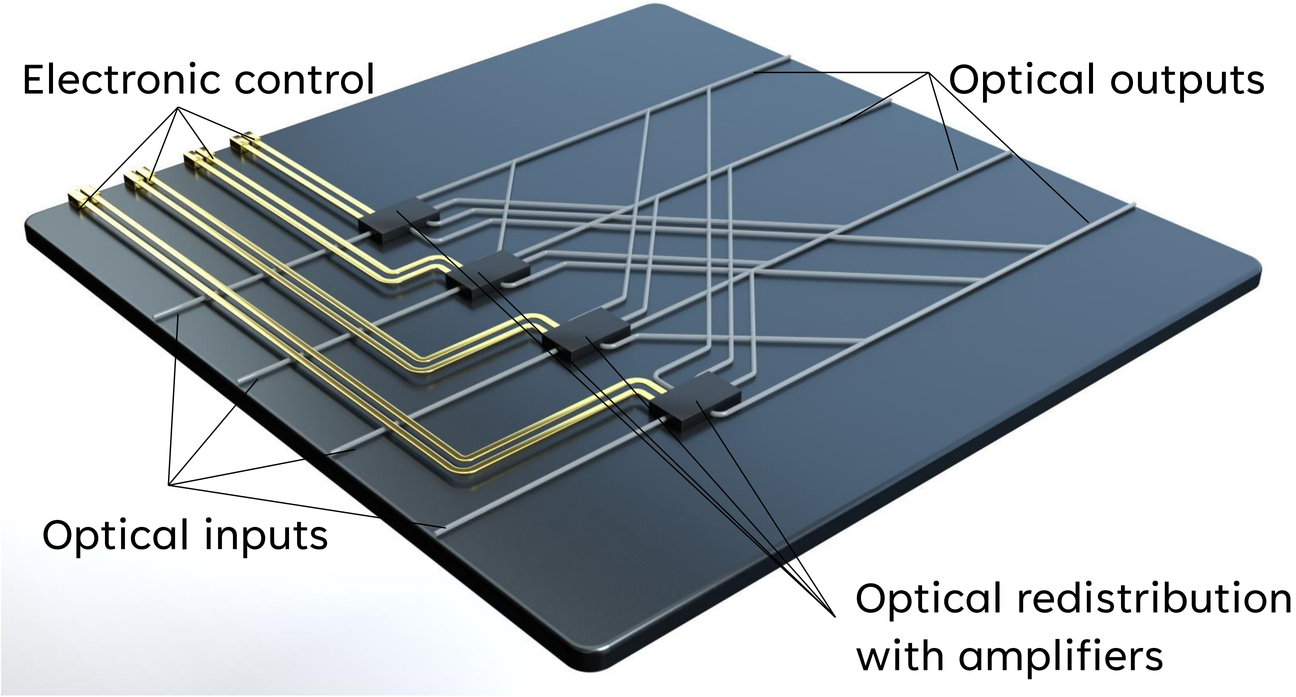
Note:

- Single clock step (up to 100 GHz) **Fourier Transforms**
- If data points match the MAC size, a fourier transformation can be carried out in a single clock step
- Larger transforms possible via decomposition

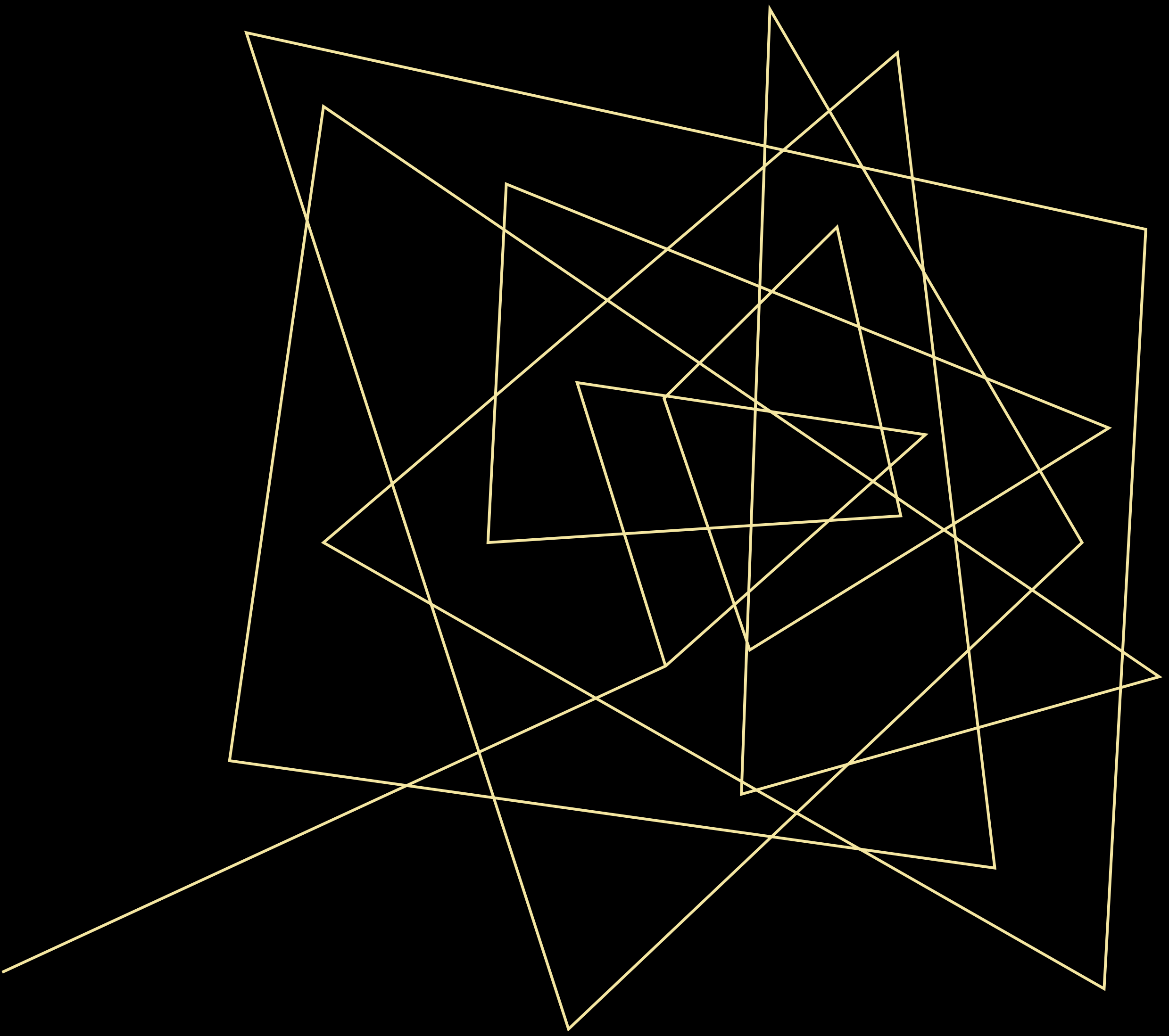


SALIENCE LABS

OPTICAL DATA TRANSFER

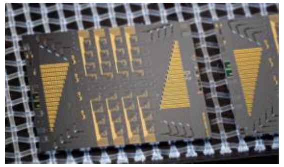
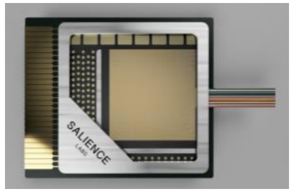


- Optical in, optical out, NxN reconfigurable
- Ultra-low latency
- Signal replication
- Networking in latency critical environments
- High bandwidth, multiple wavelength channels



OUTLOOK

DEVELOPMENT TIMELINES

	Today	Current prototype demonstration: photonic chip driven by FPGA in lab
	Aug 2023	Evaluation board: photonic chip with on chip light source, fabricated at production foundry, with driving electronics
	Oct/Nov 2023	Test chip: prototype with photonic chip packaged to a dedicated ASIC
	2024	Commercial prototype: high performance prototype with photonic chip packaged to a dedicated ASIC

GET IN TOUCH!

Ask: We are looking for collaboration partners who can benefit from our ultra-low latency processing!

Johannes Feldmann:
johannes@saliencelabs.ai

Vaysh Kewada:
vaysh@saliencelabs.ai

www.saliencelabs.ai