EPFL

EMBEDDED SYSTEMS LABORATORY

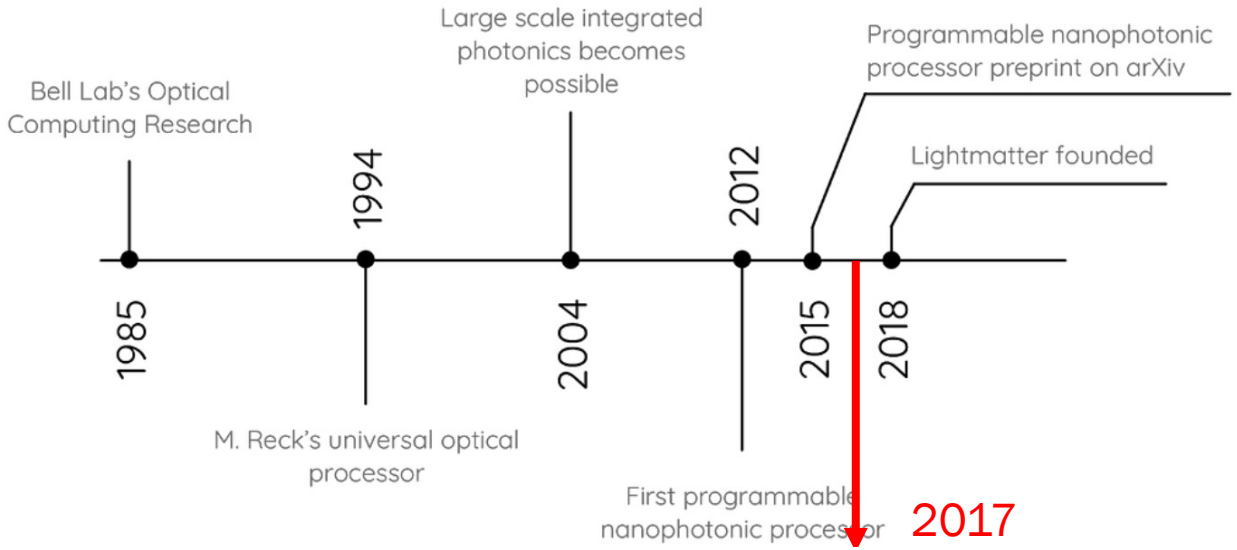
Enterprise of Optical Computing

Lightmatter & Lightelligence

Micro-608 Optical Computing

Yu, Pengbo
EPFL - Embedded Systems Laboratory
Pengbo.yu@epfl.ch
2024.04.15

The History of Lightmatter & Lightelligence



Yichen Shen, second from left, and Nicholas Harris, third from right, were both on the team that won a 2017 MIT pitch competition for their AI-focused chip technology. Now each is helming his own startup to commercialize the tech.

SAMUEL ADELMAN

nature photonics
ARTICLES
PUBLISHED ONLINE: 12 JUNE 2017 | DOI: 10.1038/NPHOTON.2017.93

Deep learning with coherent nanophotonic circuits

Yichen Shen[1*†], Nicholas C. Harris[1*†], Scott Skirlo[1], Mihika Prabhu[1], Tom Baehr-Jones[2], Michael Hochberg[2], Xin Sun[3], Shijie Zhao[4], Hugo Larochelle[5], Dirk Englund[1] and Marin Soljačić[1]

Nanalyze

A Race Between Lightmatter and Lightelligence

An article on Medium by the lead investor in Lightmatter, Matrix Partners, talks about how matrix multiplication is actually the bottleneck for...

Co-first authors

Yichen Shen, MIT, Advised by Prof Marin Soljačić. → Founder of Lightelligence (2017)

Nicholas Christopher Harris, MIT, Advised by Prof Dirk Robert Englund. → Founder of Lightmatter (2018)

Prof Marin Soljačić, MIT, Physics → Co-Founder of Lightelligence (2017)

Prof Dirk Robert Englund, MIT, EECS → Advisor of Lightmatter (2018)

[1] https://medium.com/lightmatter/the-story-behind-lightmatters-tech-e9fa0facca30

nature photonics

ARTICLES

PUBLISHED ONLINE: 12 JUNE 2017 | DOI: 10.1038/NPHOTON.2017.93

Deep learning with coherent nanophotonic circuits

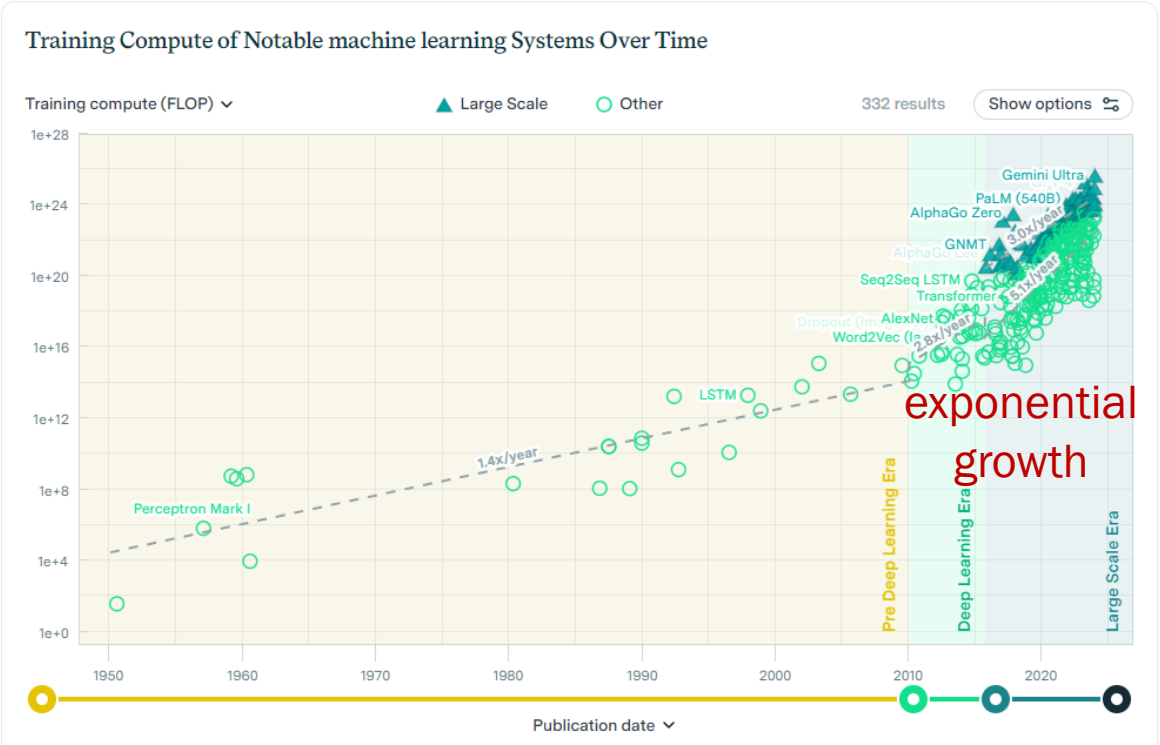
Yichen Shen[1*†], Nicholas C. Harris[1*†], Scott Skirlo[1], Mihika Prabhu[1], Tom Baehr-Jones[2], Michael Hochberg[2], Xin Sun[3], Shijie Zhao[4], Hugo Larochelle[5], Dirk Englund[1] and Marin Soljačić[1]

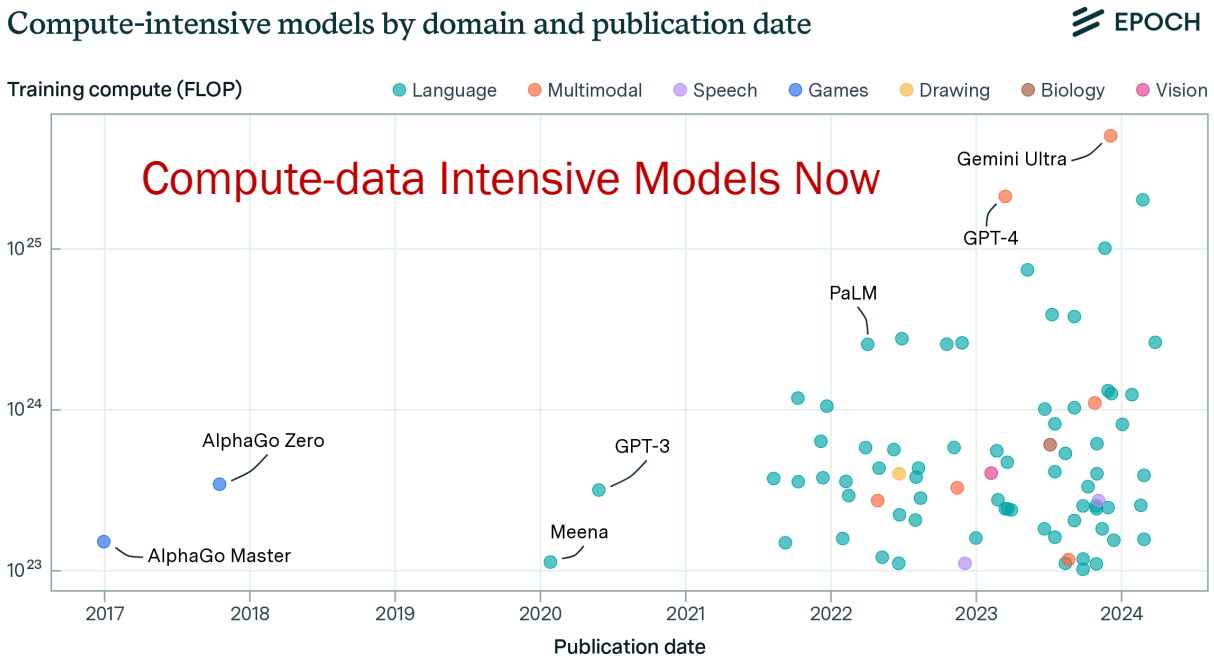
Deep learning with coherent nanophotonic circuits

Shen, Y., Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D. and Soljačić, M., 2017. Deep learning with coherent nanophotonic circuits. Nature photonics, 11(7), pp.441-446.

Background: Why Optical Computing is Attractive for AI?

Computation of Artificial Intelligence





Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703

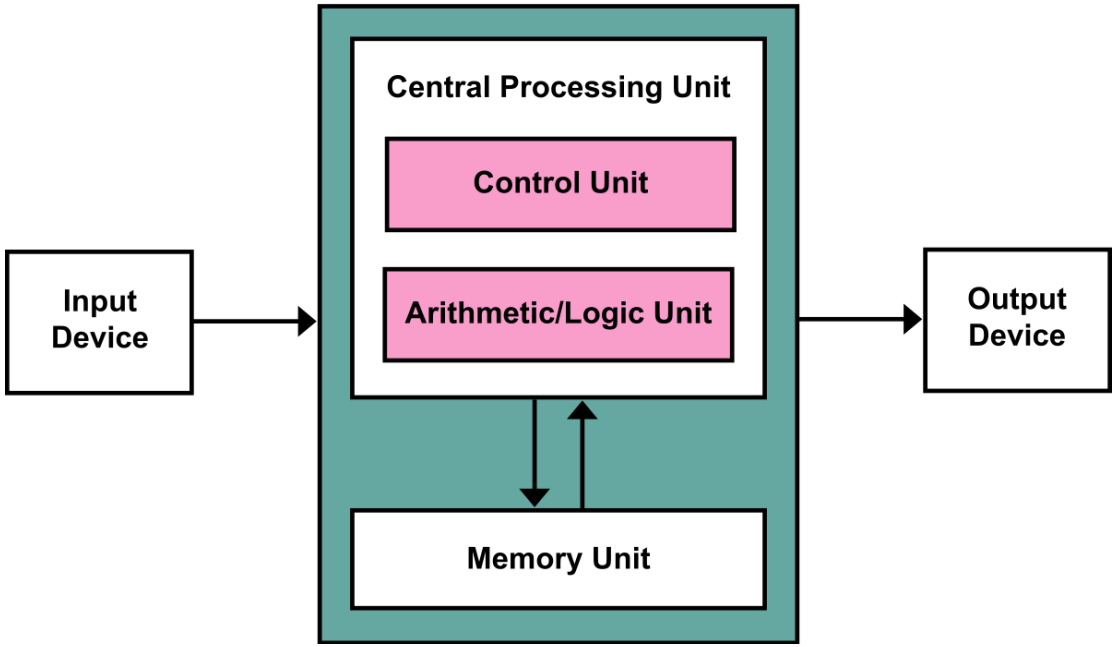
FLOP: Floating-point operations per second

1P = 10^15

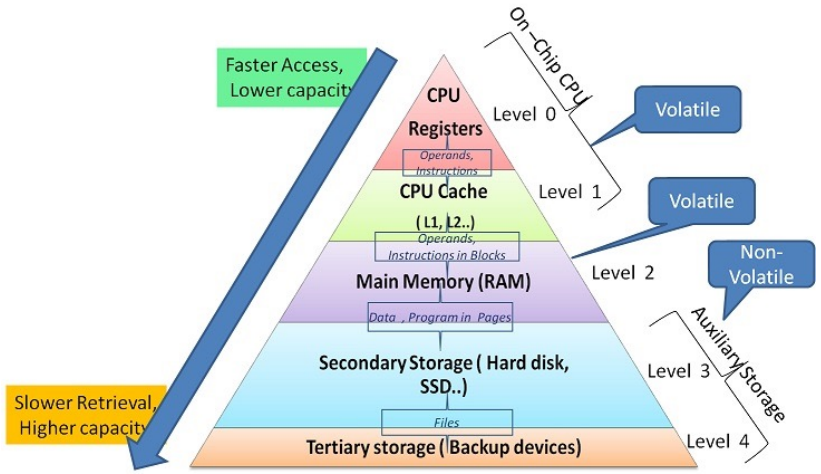
10^23 means 60,000 times of the 1st Supercomputer peak performance

[1] https://epochai.org/blog/tracking-compute-intensive-ai-models
[2] https://epochai.org/blog/compute-trends
[3] https://www.top500.org/lists/top500/2023/11/

The Traditional Computing Architecture



Von neumann architecture



Memory hierarchy

The Need for High Efficient Computing

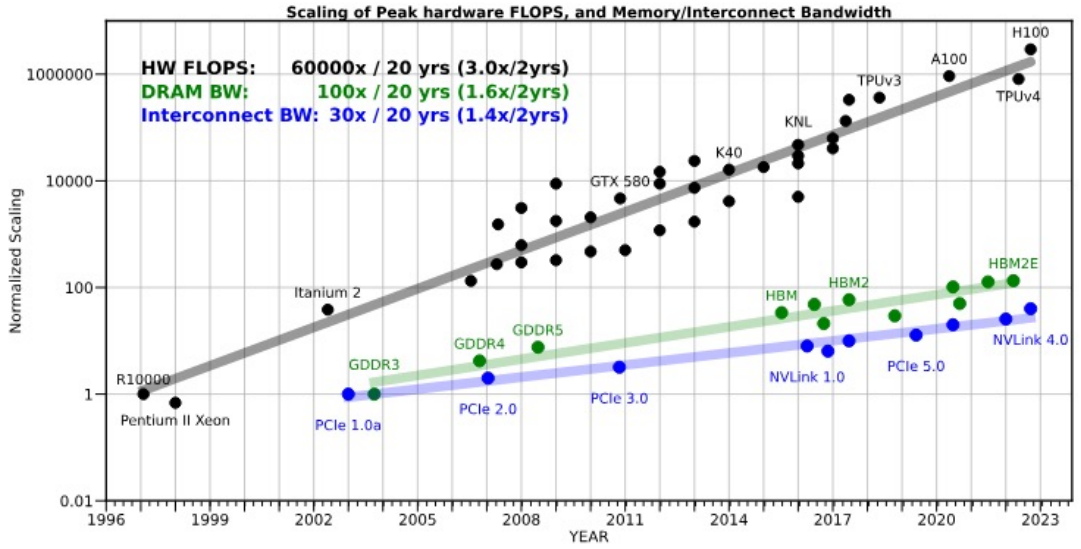
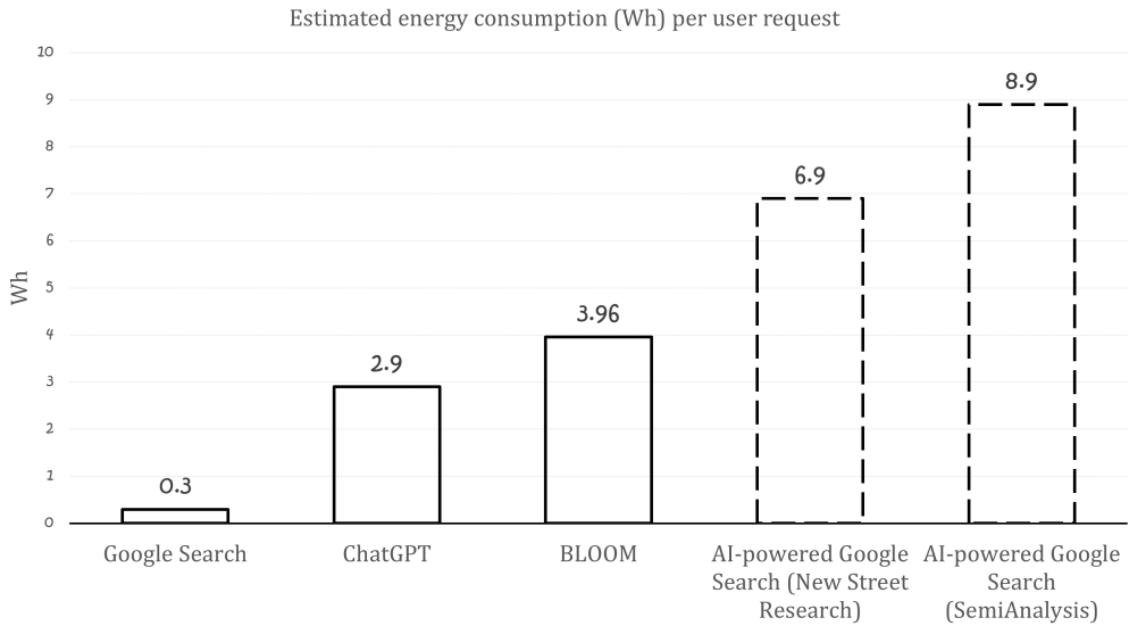


Fig. 1: *The scaling of the bandwidth of different generations of interconnections and memory, as well as the Peak FLOPS. As can be seen, the bandwidth is increasing very slowly. We are normalizing hardware peak FLOPS with the R10000 system, as it was used to report the cost of training LeNet-5.*



Trainning +billions of inference task
Energy efficient is needed

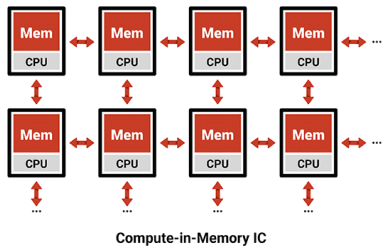
Memory bandwidth restricts computing power

[1] Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M.W. and Keutzer, K., 2024. Ai and memory wall. IEEE Micro.
[2] https://www.bellwether.works/ai-is-huge-and-so-is-its-energy-consumption/

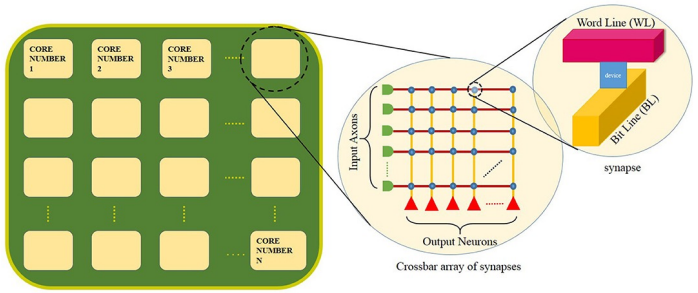
Possible solutions

Traditional electronics

- GPU / NPU / TPU / ASIC / DSIP
  (Optimal design for AI Hardware)
- In/Near Memory Computing
  (SRAM, DRAM, FLASH, RRAM, PCRAM, MRAM)



Compute-in-Memory IC



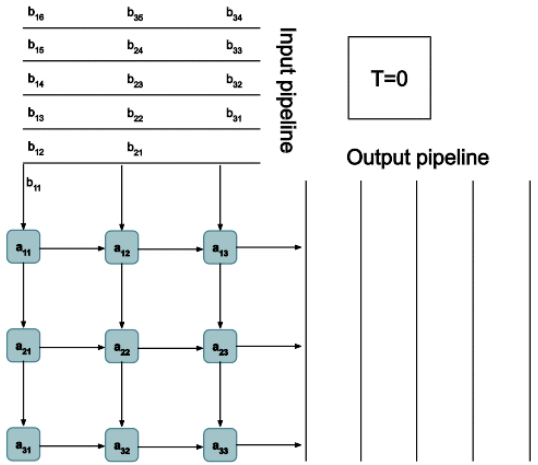
- The majority of AI algorithms can be decomposed to basic Multiply-accumulate (MAC) operations
  (MAC operation or Matrix Multiplication)
- MAC operations can be accelerated by many ways including Optical Computing

Emerging electronics

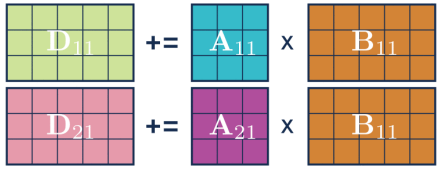
- Optical computing

① High bandwidth
② Low power
③ Feasible process

- Quantum computing



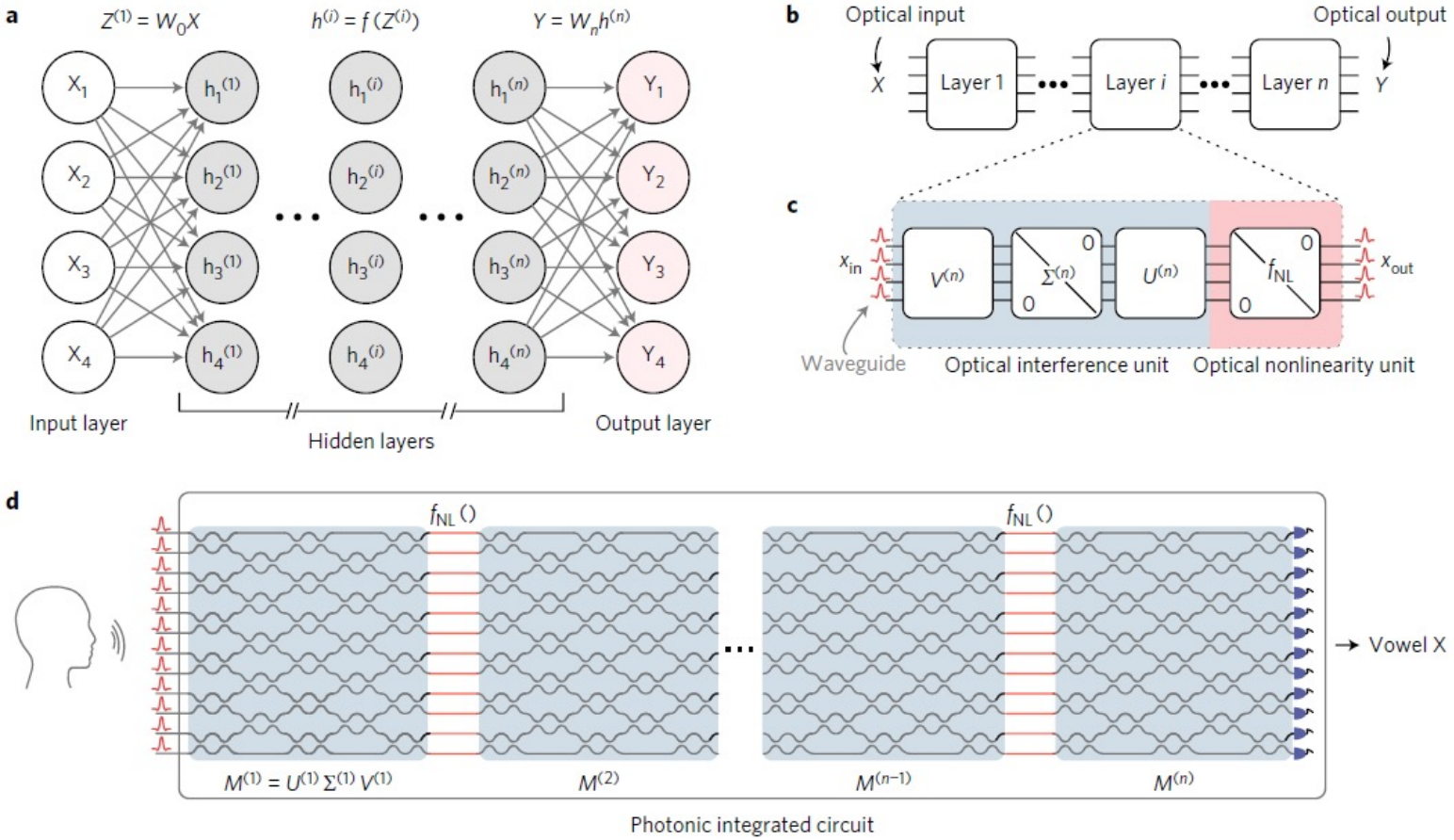
a spatial accumulation systolic array.



MAC operation

[1] Gopalakrishnan, R.,et.al. Hfnet: A cnn architecture co-designed for neuromorphic hardware with a crossbar array of synapses.

Fully Optical Neural Networks (ONNs)



① Input data is <u>preprocessed to a high-dimensional vector</u>.

② The preprocessed signals are then <u>encoded in the amplitude of optical pulses</u> propagating in the photonic integrated circuit.

③ Each layer of ONN has an <u>optical interference unit (OIU) for optical matrix multiplication</u> and <u>an optical nonlinearity unit (ONU) for the nonlinear activation.</u>

④ <u>Cascade for high depth</u>.

Optical Matrix Multiplication

- singular value decomposition (SVD)

$M = U \Sigma V^{\dagger}$

M is $m \times n$,

U is $m \times m$ (unitary matrix),

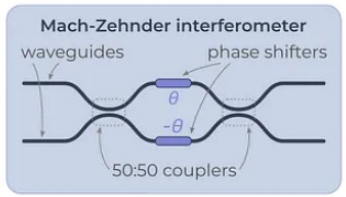
Σ is $m \times n$ (diagonal matrix),

$V^{\dagger}$ is $n \times n$ (unitary matrix),

$UU^{\dagger}=U^{\dagger}U= VV^{\dagger}=V^{\dagger}V = I$ (Identity matrix)

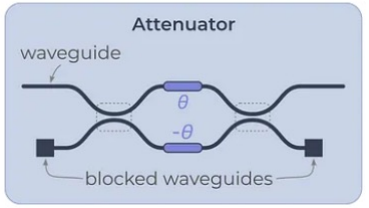
- SVD in Optics

$M = U \Sigma V^{\dagger}$



$$\Rightarrow \quad \mathbf{U}_{\mathrm{MZI}}\,\mathbf{x} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

- The elements in unitary matrix U, $V^{\dagger}$ means the light phase (hence the interference of light, phase/amplitude).
  U, $V^{\dagger}$ can be implemented with optical beamsplitters and phase shifters (2D rotation and phase transformation)



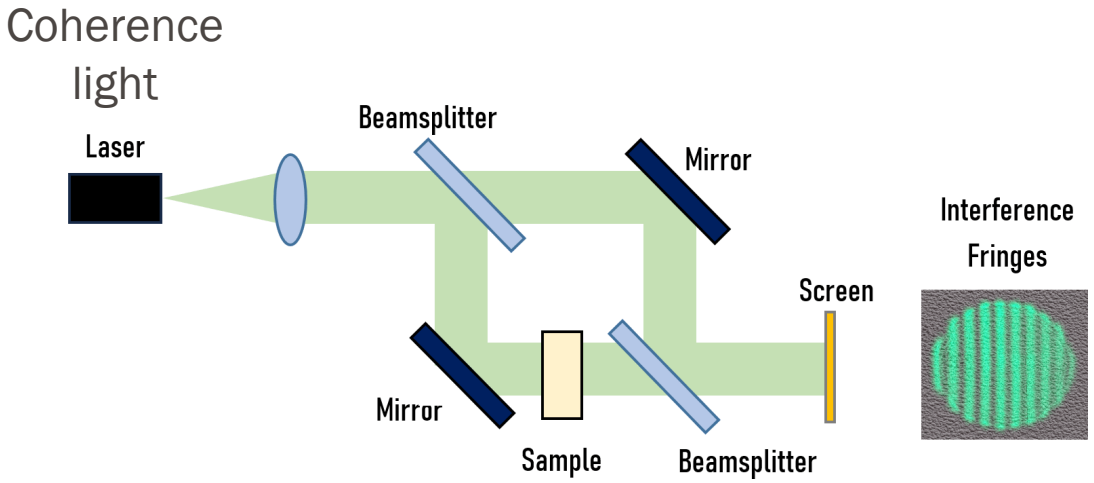
$$\Rightarrow \quad \mathbf{D}_{\mathrm{attenuator}}\,\mathbf{x} = x_1 \cos\theta$$

- The elements in the diagonal matrix Σ represent the scaling factors of different signal channels (light intensity).
  Σ can be implemented using optical attenuators—optical amplification materials

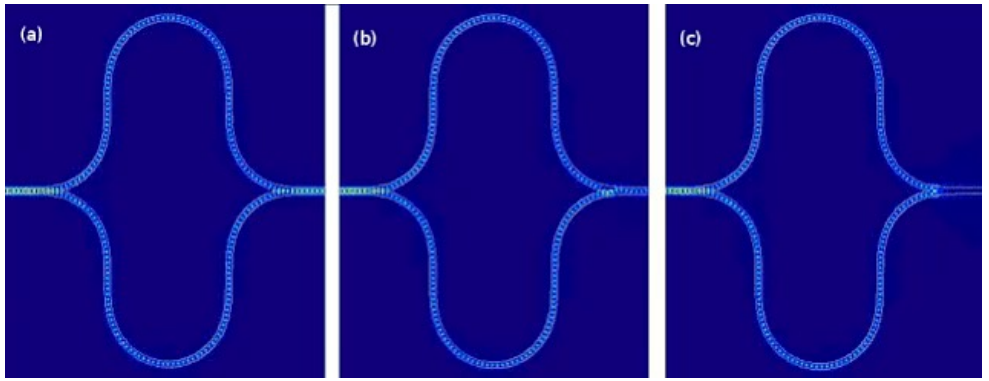
$f(O) = f(A * W) = f(U \Sigma V^{\dagger})$, f means non-linear operation

- The matrix multiplication is equal to the adjustment of light phase and light intensity.
- The non-linear is performed by traditional computer in this paper.
- Matrix multiplication has no energy cost in theory.

Mach–Zehnder Interferometers

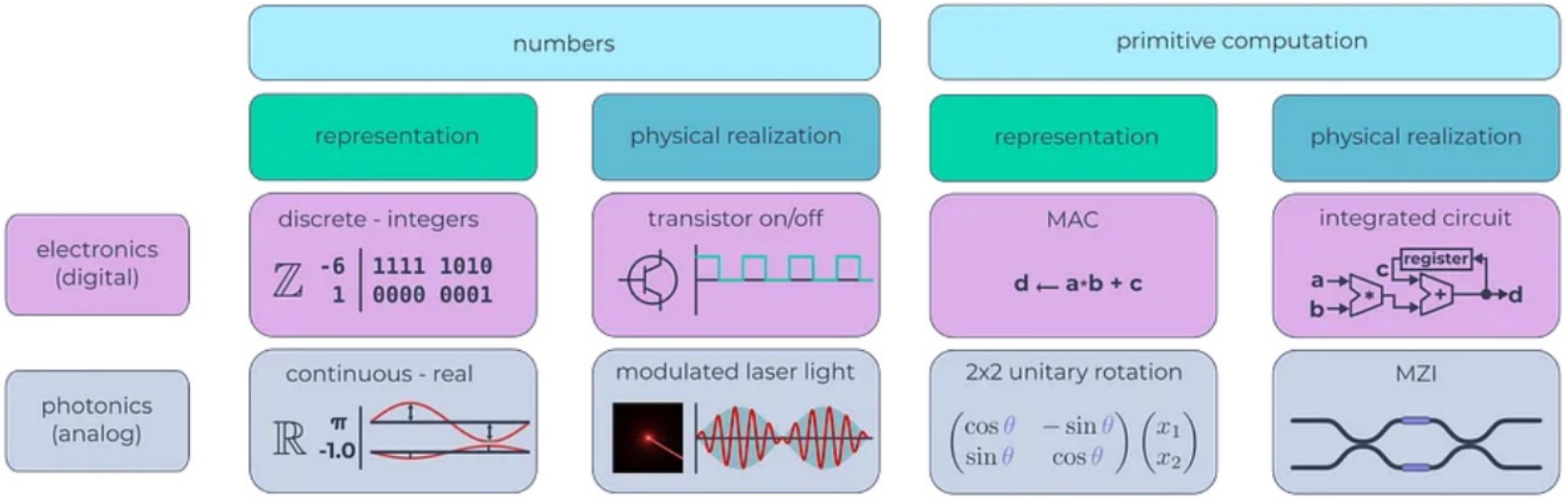


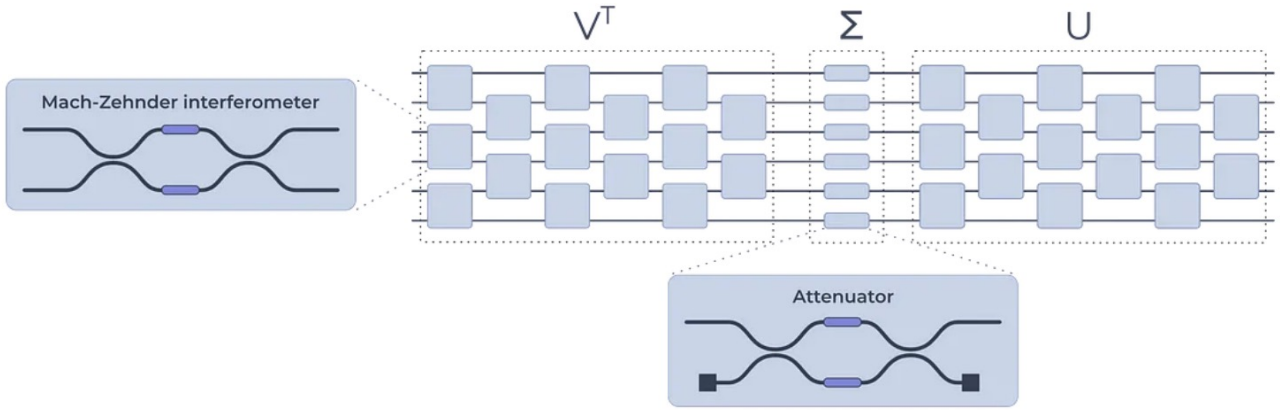
- 2D rotation and phase transformation
- Cascade for complex behavior



Optical Matrix Multiplication

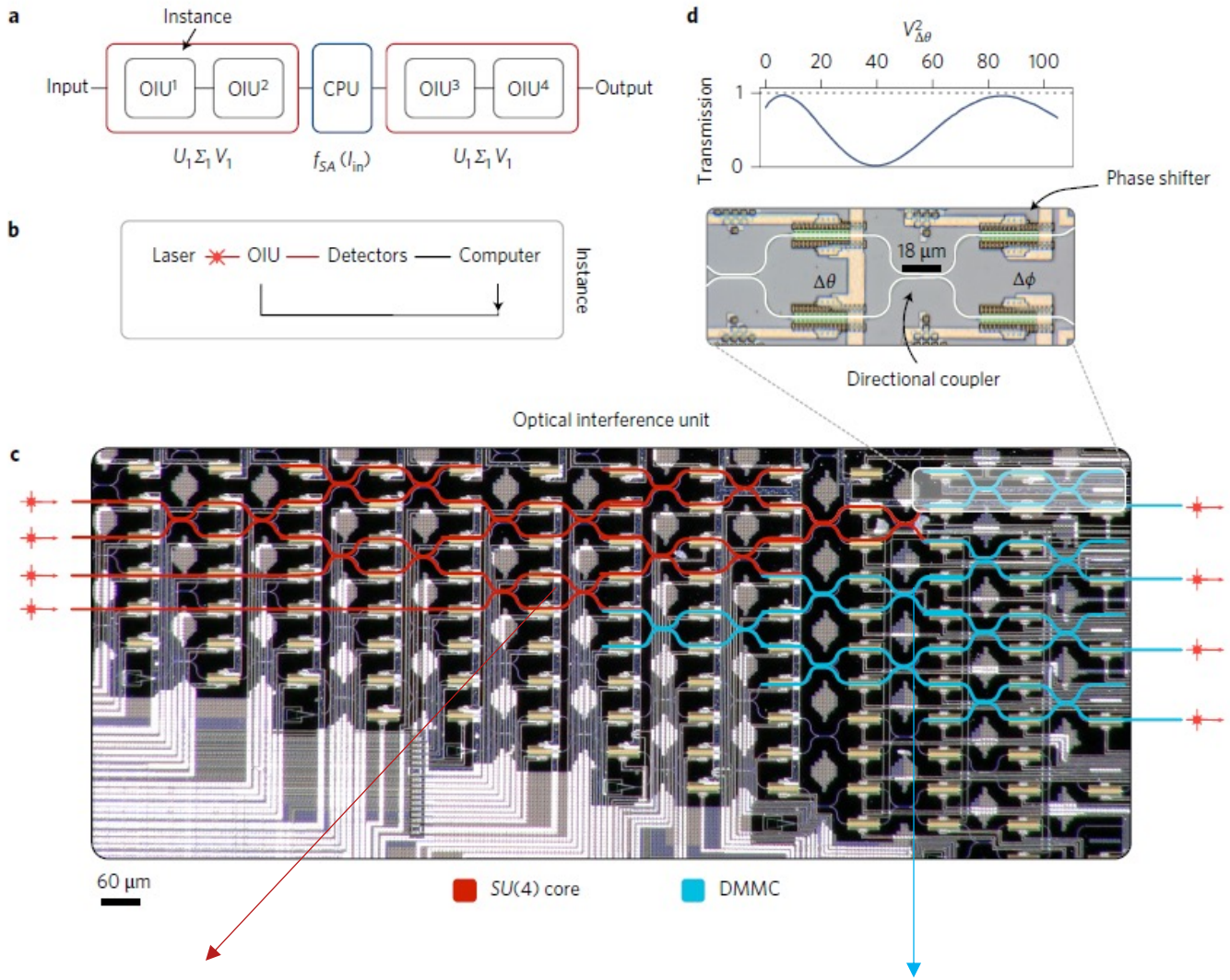
Fully Optical Neural Networks (ONNs)





[1] Martin Forsythe, Machine Learning Scientist @ Lightmatter

ONN Experiment



matrix multiplication

attenuation

Benchmark:
360 data points were generated by 90 different people speaking four different vowel phonemes

Hardware:
silicon photonic integrated circuit fabricated in the OPSIS foundry.

- 56 programmable MZIs,
- each has a thermo-optic phase shifter (θ) between two 50% directional couplers, followed by another phase shifter (φ).
- The MZI splitting ratio was controlled with an internal phase shifter and the differential output phase was controlled with the external phase shifter.

ONN Experiment

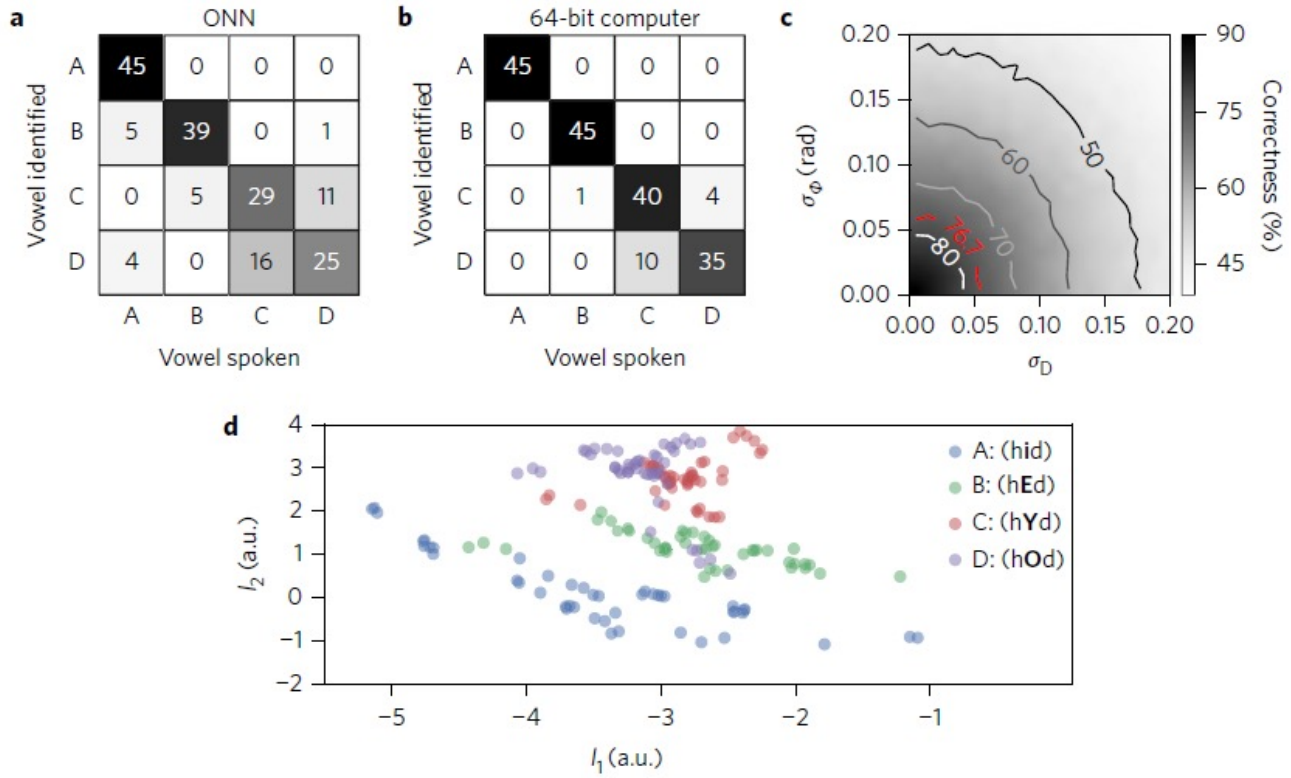


Figure 3 | Vowel recognition. a,b, Correlation matrices for the ONN and a 64-bit electronic computer, respectively, implementing two-layer neural networks for vowel recognition. Each row of the correlation matrices is a histogram of the number of times the ONN or 64-bit computer identified vowel X when presented with vowel Y. Perfect performance for the vowel recognition task would result in a diagonal correlation matrix. **c**, Correct identification ratio in percent for the vowel recognition problem with phase-encoding (σ_Φ) and photodetection error (σ_D). The definitions of these two variables are provided in the Methods. Solid lines are contours for different correctness ratios. In our experiment, $\sigma_D \simeq 0.1\%$. The contour line shown in red marks an isoline corresponding to the correct identification ratio for our experiment. **d**, Two-dimensional projection (log area ratio coefficient 1 on the *x* axis and 2 on the *y* axis) of the testing data set, which shows the large overlap between spoken vowel C and D. This large overlap leads to lower classification accuracy for both a 64-bit computer and the experimental ONN.

- Training is performed by conventional computer
- ONN adapts the trained parameter and shows close accuracy for inference
- It is a very simple demo, but shows the feasibility for larger ONN applications.
- ONN is sensitive to phase-encoding and device vibration (error).

ONN Discussion and Expectation

1. **Resolution (Analog Computing)**
   - The finite precision of optical phase (16-bit)
   - Cascade crosstalk
   - Device vibration
   - Nosise

2. **Computation speed and energy efficiency**
   - Less even zero energy cost (now ~10 mW per phase modulator)
   - Low latency and larger throughput
   - Only limited by hardware optical system

3. **On-chip training**
   - Back propagation could be replaced by forward propagation.

Follow-up Progress

An Electro-Photonic System for Accelerating Deep Neural Networks ------ From system level evaluation instead of sing ONN

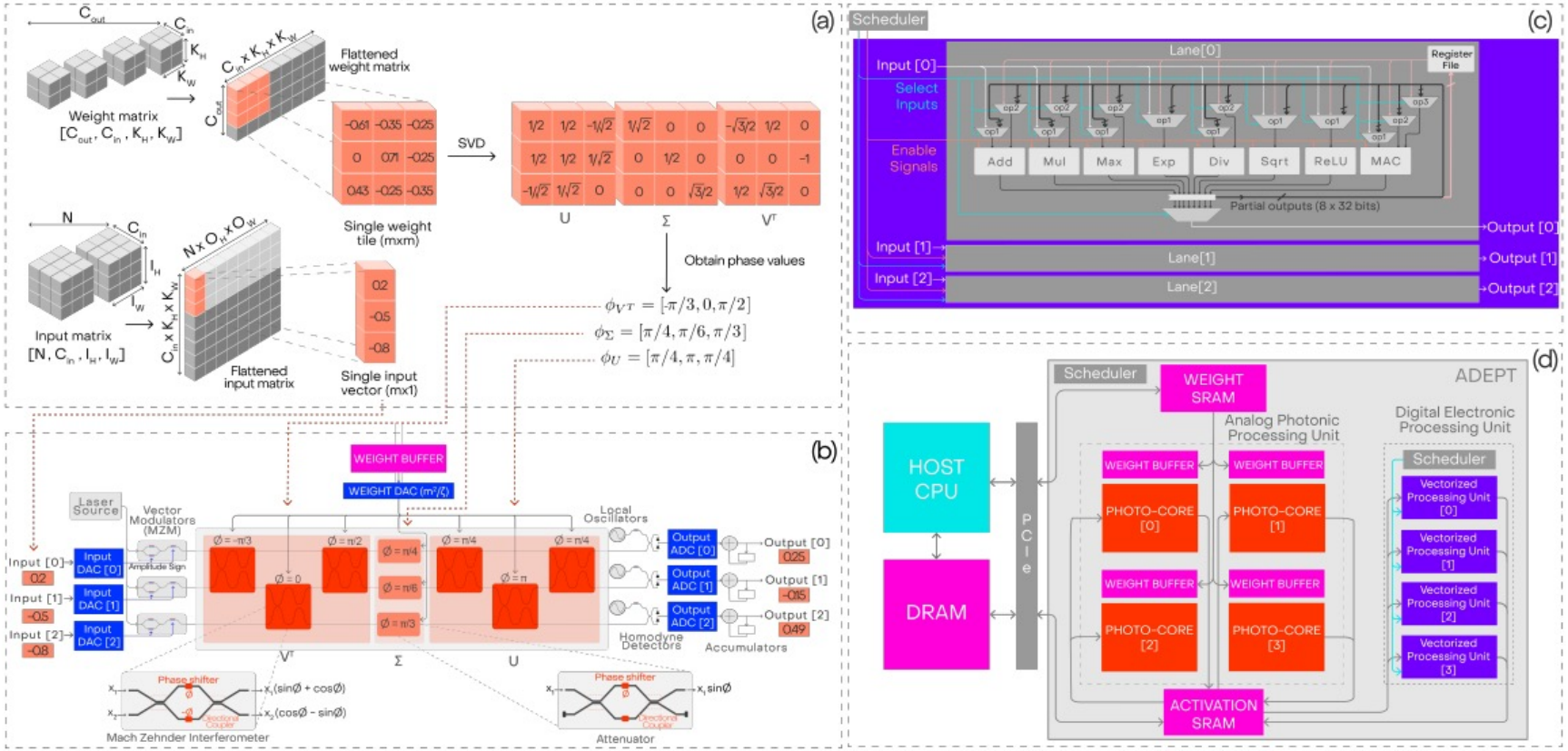


Figure 1: **Diagram showing different components of ADEPT and how operations are performed.** (a) Example GEMM operation in the photo-core. (b) Programming input and weight matrices into the photo-core. The $m \times m$ (here $m = 3$ as an example) photo-core consists of $2 \times m(m-1)/2 = 6$ MZIs (for U and V^T) and 3 attenuators (for Σ). (c) Microarchitecture for a single digital electronic vectorized processing unit. The unit comprises $m = 3$ digital lanes, each consisting of arithmetic units to perform non-GEMM operations. (d) Full system architecture including the host CPU, the DRAM, and ADEPT—interconnected using a PCI-e interface. As an example, we show four photo-cores and four vectorized processing units.

- Host CPU
  top level schedule and control
- DRAM
  external memory, low speed
  but larger capacity size
- SRAM
  internal memory, high speed
  and medium capacity size
  store photonics value
- Photonics Core
  MAC operation acceleration
- ASCI
  input pre[rocessing
  non-linear operation (ReLU, etc.)
  ADC, DAC (8-bit)

[1] Demirkiran, et.al. An electro-photonic system for accelerating deep neural networks. ACM Journal on Emerging Technologies in Computing Systems, 19(4), pp.1-31.

An Electro-Photonic System for Accelerating Deep Neural Networks ------ From system level evaluation instead of sing ONN

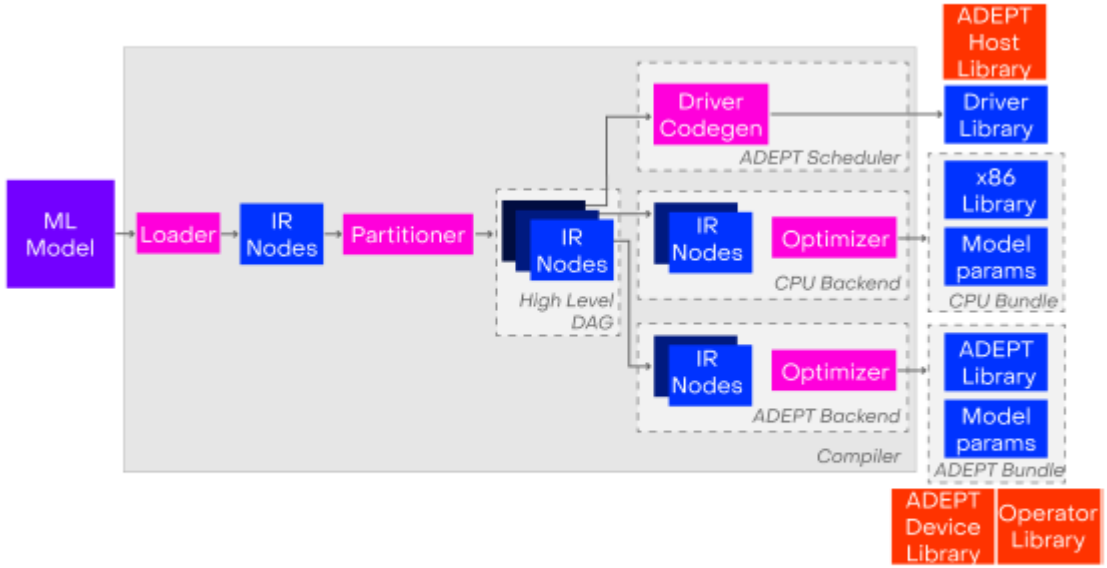


Figure 2: **Execution model.** Compilation process of an ML model for ADEPT.

Software-Hardware co-optimization for the mapping and scheduling

[1] Demirkiran, et.al. An electro-photonic system for accelerating deep neural networks. ACM Journal on Emerging Technologies in Computing Systems, 19(4), pp.1-31.

An Electro-Photonic System for Accelerating Deep Neural Networks ------ From system level evaluation instead of sing ONN

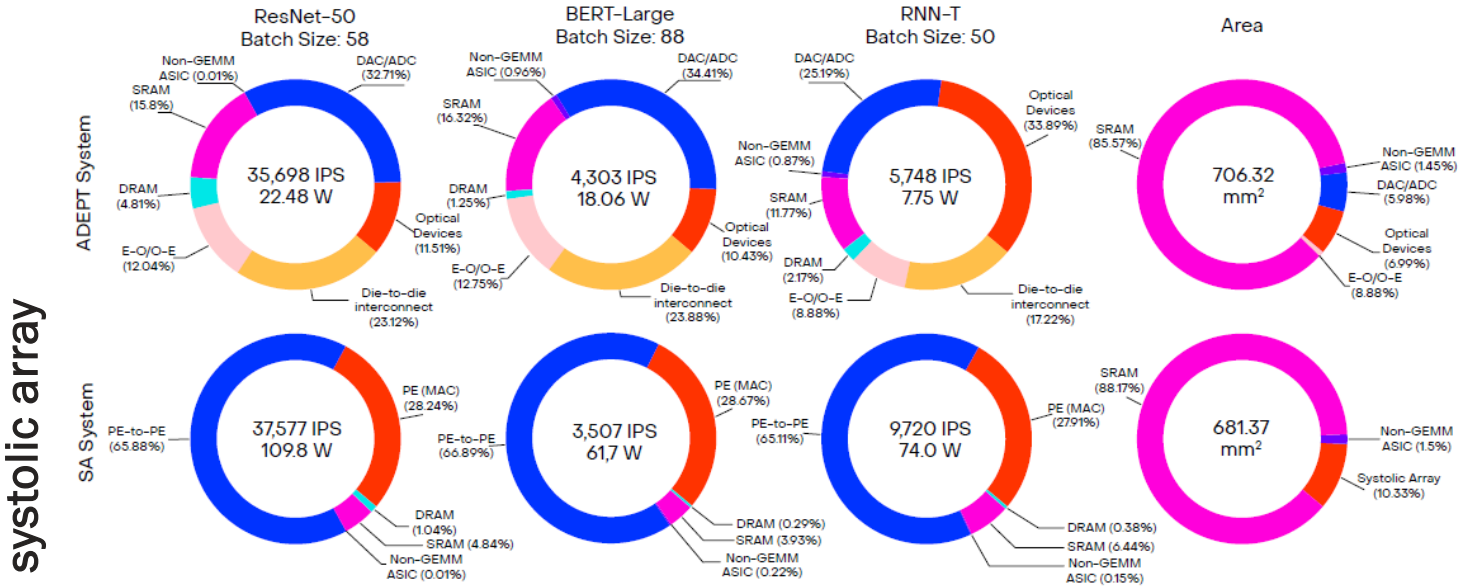


Figure 8: *Average total (static and dynamic) power distribution and area distribution of ADEPT* (128 × 128, *10 GHz photo-core) and the SA system* (128 × 128, *10×1 GHz array, OS dataflow).*

- SRAM takes most area
- The energy cost by DRAM, ADC/DAC, electrical-to-optical (O-E/E-O) and die2die interconnect can not be ignored.

➢ Still, optical computing has great benefit.

Table 2: Comparison against state-of-the-art electronic and photonic accelerators.

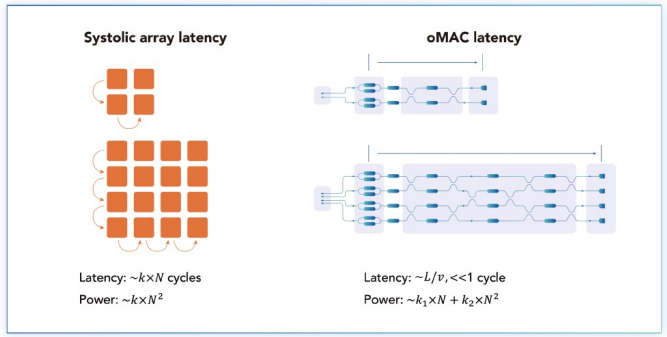
	ADEPT (This work)		Eyeriss [15]	Eyeriss v2 [18]	UNPU [46]	TPU v3 [42]
Tech Node	90 nm photonics + 22 nm CMOS		65 nm	65 nm	65 nm	16 nm
Clock rate	10 GHz		200 MHz	200 MHz	200 MHz	940 MHz
Benchmark	AlexNet	ResNet-50	AlexNet	AlexNet	AlexNet	ResNet-50
Batch size	192	58	4	1	15	N/A
IPS	217, 201	35,698	35	102	346	32,716
IPS/W	7,476.78	1,587.99	124.80	174.80	1,097.50	18.18
IPS/W/mm²	10.59	2.25	10.18	N/A	68.59	0.01

➢ The benefit of optical computing is limited by peripheral devices, and it is difficult to go to THz inference as expected in theory.

[1] Demirkiran, et.al. An electro-photonic system for accelerating deep neural networks. ACM Journal on Emerging Technologies in Computing Systems, 19(4), pp.1-31.

Lightelligence Whitepaper

Optical Computing



- The key advantage is low latency
- Relative less energy
- Relative high bandwidth

- ☐ Precision limit (8bit)
- ☐ Computing noise

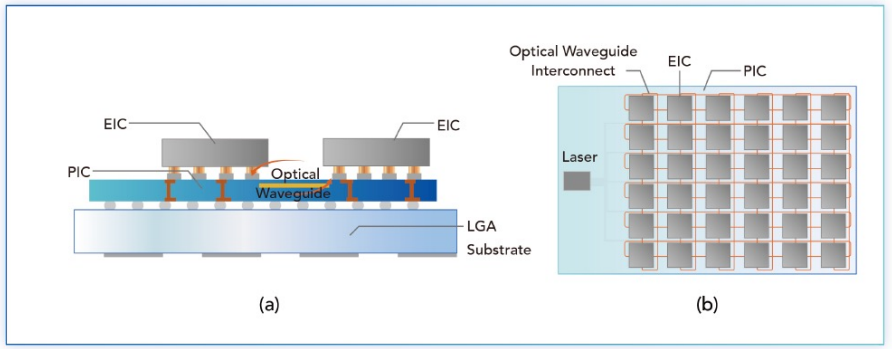


Figure 5 Cross-sectional view (a) and top view (b) of oNOC system where electronic chips are interconnected by optical waveguide based links

- Large scale integration
- Chiplet by packging

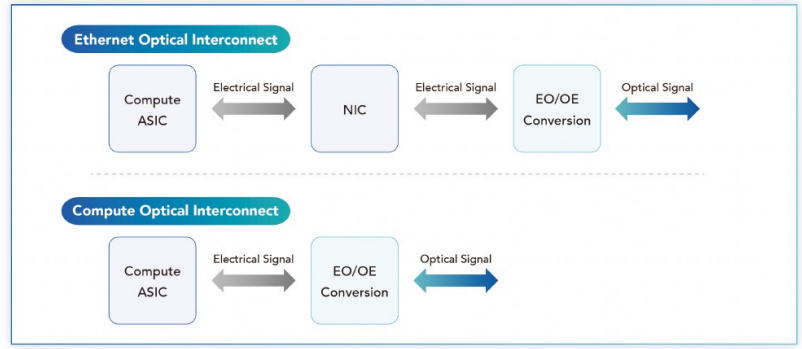


Figure 6 Ethernet optical interconnect and compute optical interconnect

- Optical inter-chip Networking

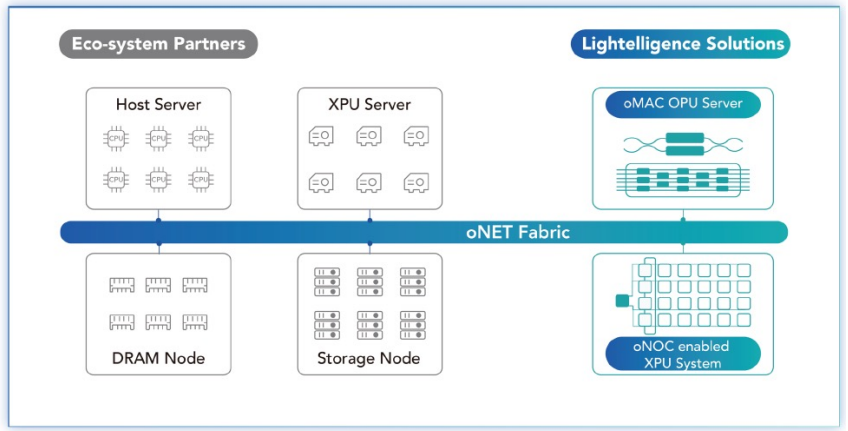
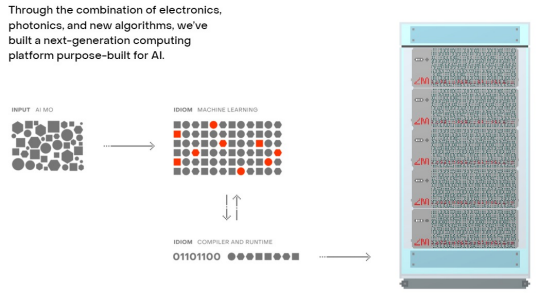
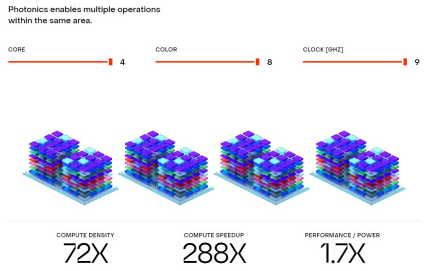
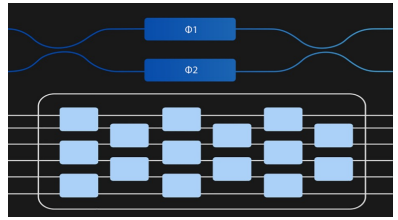
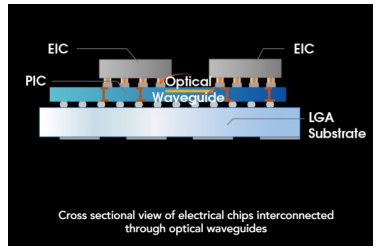
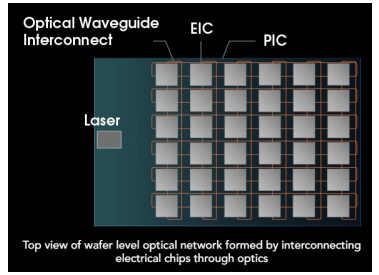


Figure 9 Schematics of a new data center architecture with integrated silicon photonics technology

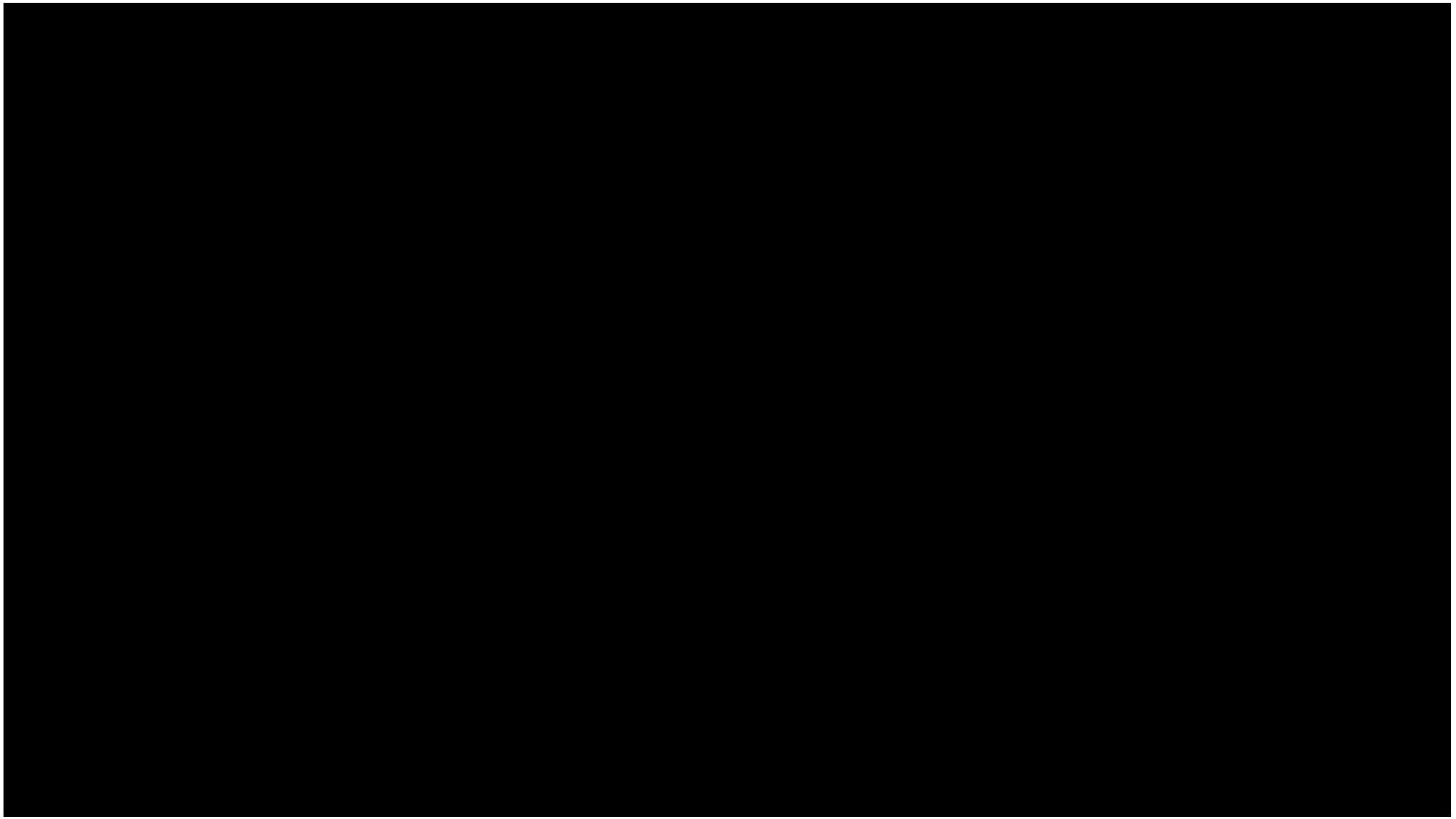
Heterogeneous computing with traditional electronics platforms and emerging optical computing

Commercial Products

Lightmatter & Lightelligence Products

	Lightmatter	Lightelligence
Optical Computing	  **Envise™** features 16 Envise Chips in a server configuration with only 3kW power consumption. It has **3 times higher IPS than the Nvidia DGX-A100 and 8 times the IPS/W on BERT-Base SQuAD (a benchmark).** 500MB Massive on-chip activation and weight storage enabling state-of-the-art neural network execution without leaving the processor. PCI-E 4.0 Standards-based host and interconnect interface. Revolutionary compute, standard communications. 256 RISC cores per Envise™ processor. Generic off-load capabilities. Ultra-high performance out-of-order super-scalar processing architecture. RAS Deployment-grade reliability, availability, and serviceability features. Next generation compute with the reliability of standard electronics. 400Gbps Lightmatter® interconnect fabric per Envise™ chip — enabling large model scale-out. Running the most advanced neural networks on the planet. ➢ 16xLightmatter® Envise™ ➢ 2xAMD EPYC 7002 host processors ➢ 3TB NVMe SSD ➢ 6.4Tbps LM-Fabric for scale-out ➢ 2x200G Ethernet Smart NIC ➢ Gigabit ethernet for IPMI management ➢ 3kW TDP ➢ 4U form factor	**Photonic Arithmetic Computing Engine (PACE)** A fully integrated photonic computing platform. **It has a 64x64 optical matrix multiplier in an integrated silicon photonic chip (150ps delay) and a CMOS microelectronic chip.** It also contains **over 12,000 discrete photonic devices and has a system clock of 1GHz.**   **HUMMINGBIRD** Hummingbird serves as **the communications network for data centers and other high-performance applications.** It has **64 transmitters and 512 receivers.** **PCIe bus and Lightelligence Software Development Kit (SDK).**  Cross sectional view of electrical chips interconnected through optical waveguides  Top view of wafer level optical network formed by interconnecting electrical chips through optics

Lightmatter ------- ENVISE™



Lightelligence ------ Optical Multiply Accumulate

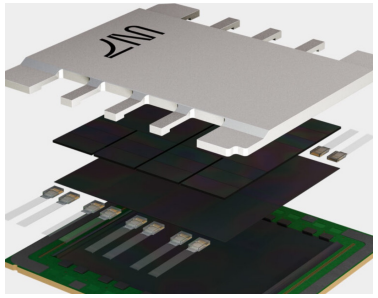
Lightmatter & Lightelligence Products

Lightmatter

Photonics Interconnect

PASSAGE™:

Designed for **interconnection among optical computing chips** for high bandwidth.



Fully integrated chiplet, interconnect solution with direct fiber attach all in a single assembly.

Uniform architecture for flexible dicing (e.g. 2×2, 2×4, 2×8).

CPU GPU/ ACCELERATOR MEMORY

Wafer-scale processing with heterogeneous tiles of CPUs, GPUs, FPGAs, DRAM, and ASICs.

Transistors and photonics integrated side-by-side. SerDes signals from chiplets directly modulated onto waveguides. Standards-based D2D interfaces supported including UCIe, AIB, and others.

Any Topology Anytime with 1-Hop-Everywhere. Dynamically reconfigure network configurations in microseconds.

Thousands of waveguides with cross-reticle stitching. Every chiplet directly connected to every other at multi-terabit speed and near-zero latency.

40X

Dramatic interconnect density improvement. 40 waveguides in the space of one optical fiber.

800+ TBPS

Input/output bandwidth from each chiplet site for full reticle. And up to 250+ Tbps per chiplet site edge.

100X

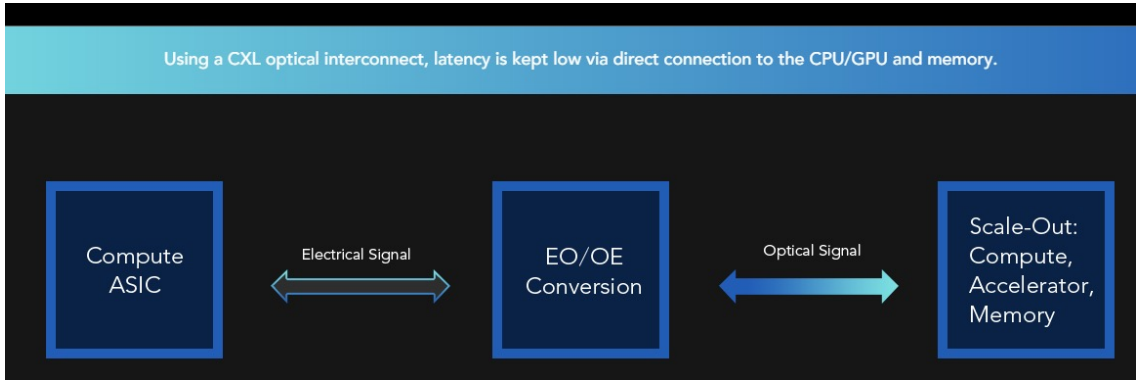
More bandwidth than existing chip-to-chip interconnect solutions.

<2NS

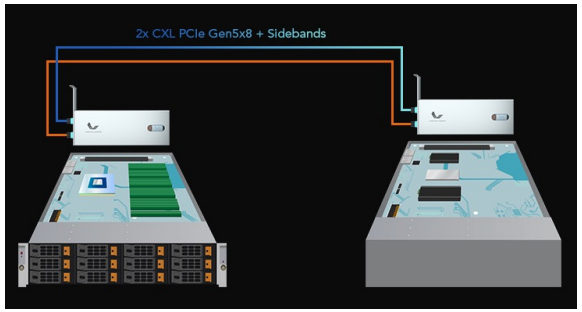
Chiplet to chiplet latency, single hop connectivity between every site.

Lightelligence

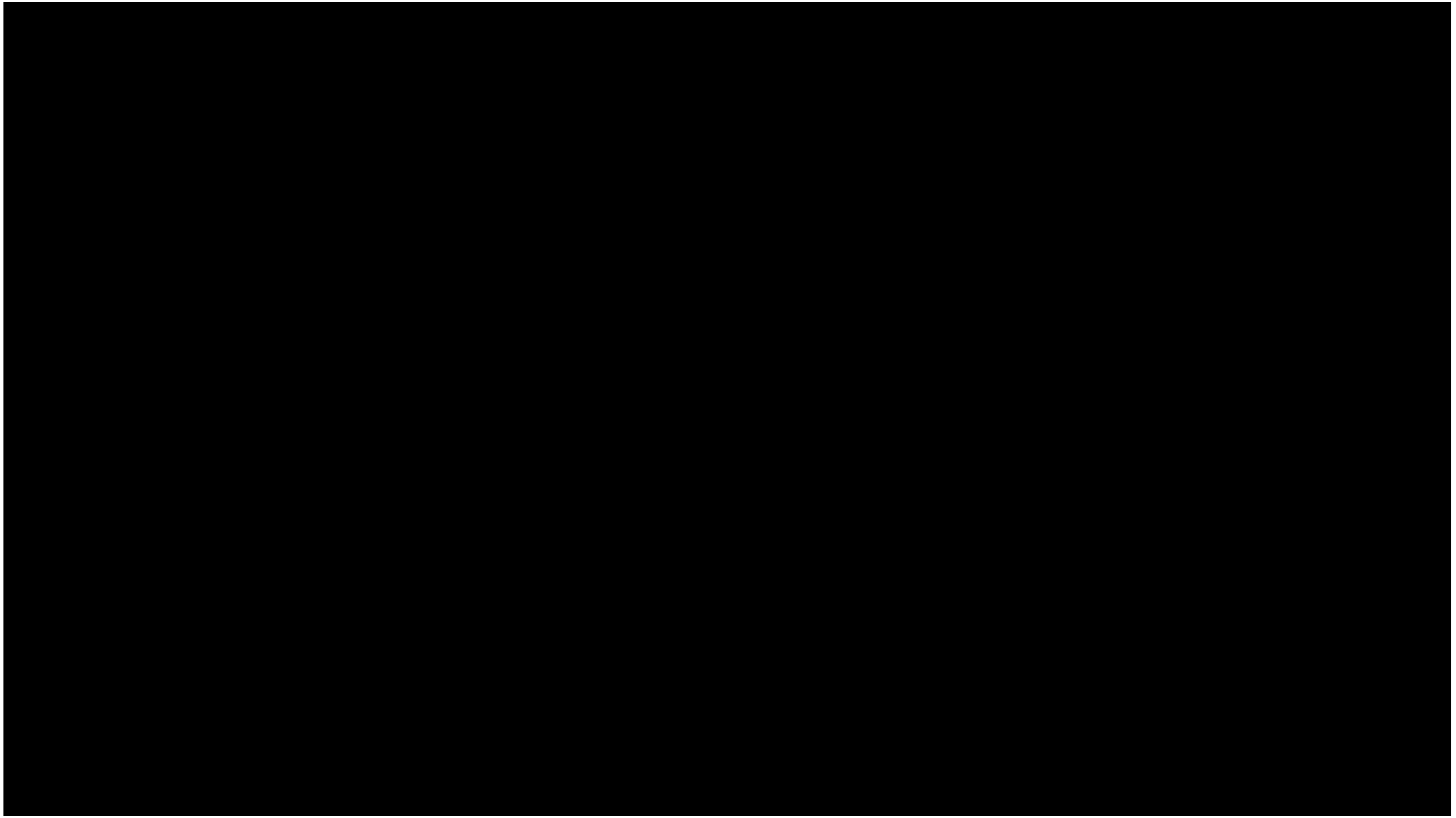
Photowave™



optical communications hardware that is kind designed for PCIe and Compute Express Link (CXL) connectivity. Leveraging the significant latency and energy efficiency benefits of photonics, it enables data center managers to scale resources within or across server racks.



Lightmatter ------- PASSAGE™

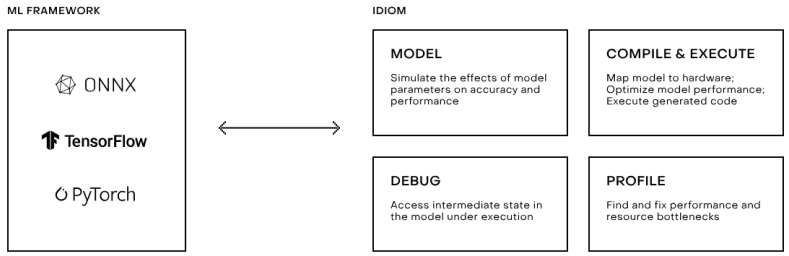


Lightelligence ------ Optical Network on Chip

Lightelligence ------ Optical Networking

Lightmatter & Lightelligence Products

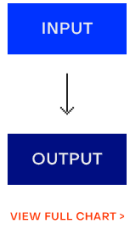
	Lightmatter	Lightelligence
Others	**IDIOMP® SOFTWARE TOOLS FOR AI DEPLOYMENT**	**Moonstone** **Single and Multi-wavelength Optical Sources**



GRAPH COMPILER

idCompile automates the programming by partitioning (large) neural networks for parallel programming within and between Envise blades

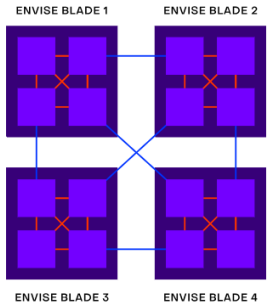
- Automatic conversion from floating-point numbers for mixed-precision inference
- Automatic generation of optimized execution schedule
- Supports multiple parallelism strategies: data parallelism, model parallelism, and pipelining



MULTI-BLADE ENVISE PARTITIONING

Idiom® automatically performs the partitioning between multiple Envise™ blades.

- Proprietary Lightmatter® fiber optical communication links Envise™ blades, while Idiom® synchronizes the Envise chips together in a single runtime
- Automatic partitioning chooses the best parallelism model for performance
- Virtualizes each Envise™ blade automatically and multiple users can apportion the number of chips used



≥18dBm/ch single wavelength output optical power
≥14dBm/λ/ch multi-wavelength output optical power

- Moonstone™ is a **high power, multi-channel, single or multi-wavelength DFB laser source.**
- It has **a smaller footprint, better operating temperature ranges** and is field replaceable with advanced packaging at a much lower price point.
- **Ideal for telecommunications, LIDAR, ethernet switching**, along with a broad range of test and sensor equipment.

Venture Capital of Lightmatter & Lightelligence

Lightmatter

Financing process 6

Export 企查查

serial number	date	Financing rounds	Valuation amount	Financing Amount	Investment agency	News link
1	2024-01-09	C+ round	US$1.2 billion	not disclosed	SIP Global Partners	news source
2	2023-12-18	C+ round	US$1.2 billion	$155 million	Viking Global Investors, Google Ventures	news source
3	2023-05-30	C round	not disclosed	$154 million	Fidelity Management Research , Viking Global Investors , Google Ventures , SIP Global Partners , HPE Pathfinder	news source
4	2021-05-05	Round B	not disclosed	US$80 million	Viking Global Investors, Lockheed Martin Ventures, Spark Capital, Google Ventures, Hewlett Packard Enterprise, 经纬创投, SIP Global Partners	news source
5	2019-02-25	A+ round	not disclosed	$22 million	Spark Capital , Google Ventures , Matrix Partners	-
6	2018-02-06	Series A	not disclosed	$11 million	Spark Capital , Matrix Partners	-

Round C, more than 400 million USD

Lightelligence

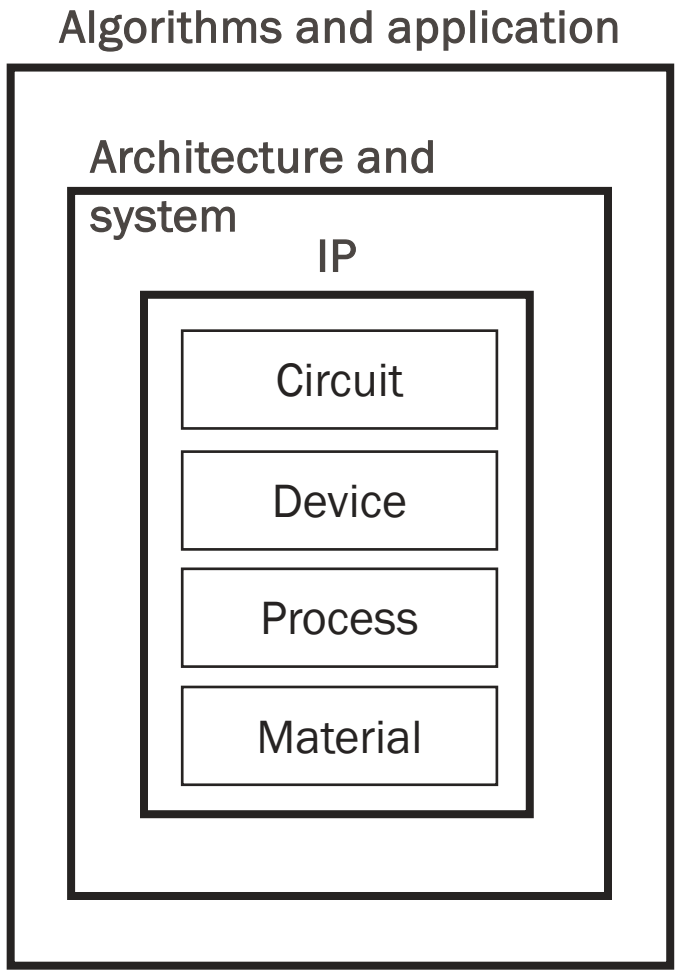
Financing process 3

天眼查

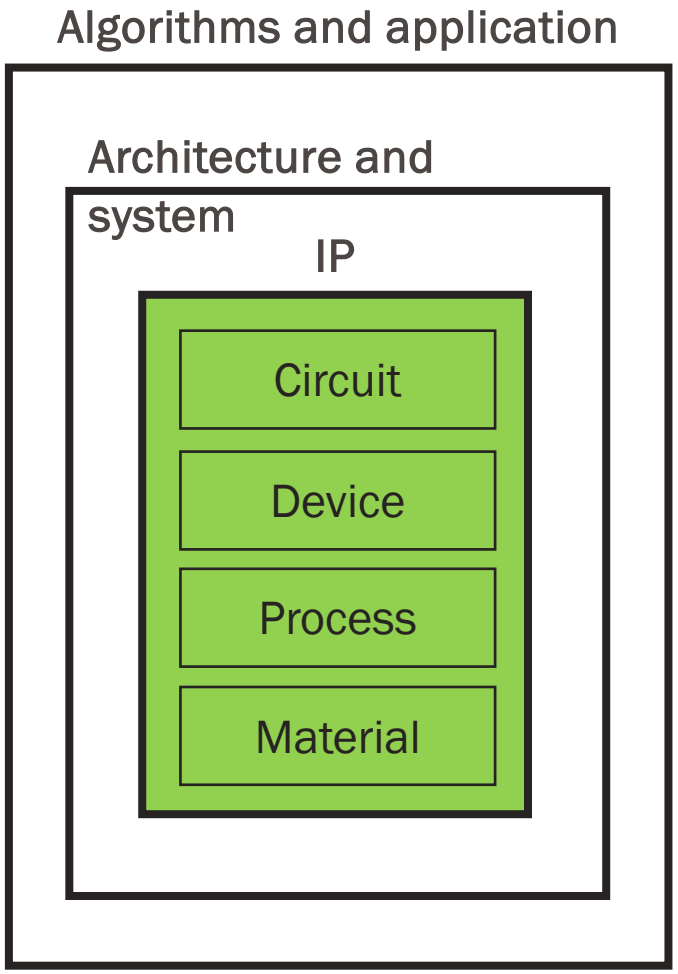
serial number	Disclosure date	Amount of the transaction	Financing rounds	Valuation	Proportion	investor	news source
1	2020-07-06	tens of millions of dollars	A+ round	-	-	Heli Capital Henry Yuan	Photonic AI chip company Xizhi Technology completed tens of millions of dollars in Series A+ financing
2	2020-04-08	$26 million	Series A	-	-	Matrix Partners CICC Capital Zhongke Chuangxing Fengrui Capital Vertex Investment China Fund BV Baidu Venture Capital China Merchants Venture Capital	The world's first photonic chip company completed US$26 million in financing, with Basalt serving as financial advisor
3	2018-02-04	US$10 million	angel wheel	-	-	BV Baidu Venture Capital ZhenFund Dexun Investment	Baidu invests in optical AI chip startup Lightelligence

Round A+, more than 40 million USD

In Summary



Tradition electronics computing



Optical Computing

Thank you!

Yu, Pengbo
EPFL - Embedded Systems Laboratory
Pengbo.yu@epfl.ch