A Giant Leap

Photonic computing is here
to take us *there*

01 Motivation: Challenges for electronics

02 Opportunity: AI acceleration with GEMM

03 Accelerating AI with photonics

04 Looking ahead

Challenges for electronics

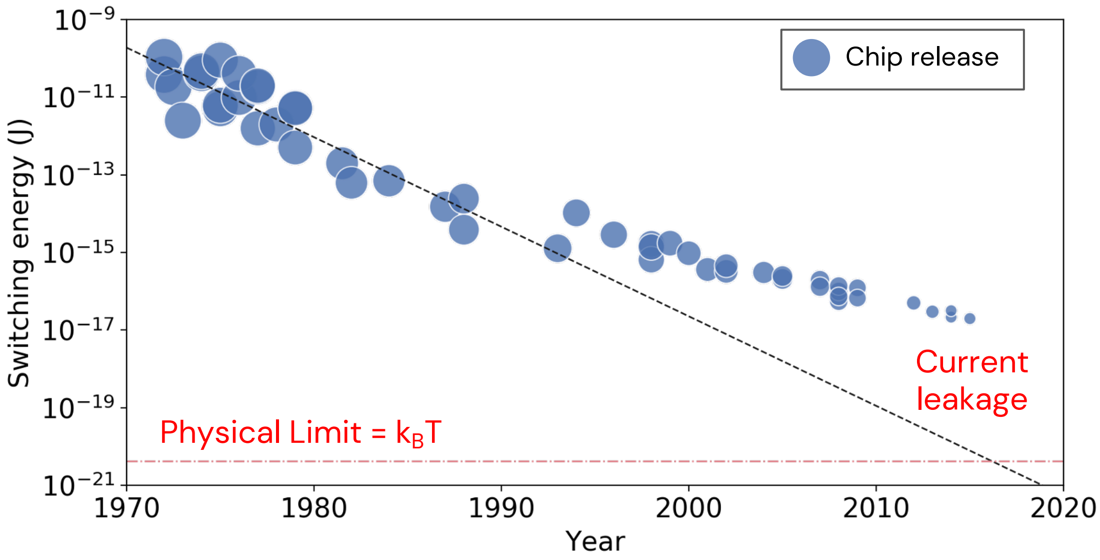
"

Moore's Law lasted for half a century and changed the world as Moore predicted in his 1965 article, but it has ended.

David Patterson
Google, UC Berkeley
Turing Award Laureate

Heat

How transistor scaling and our universe ends



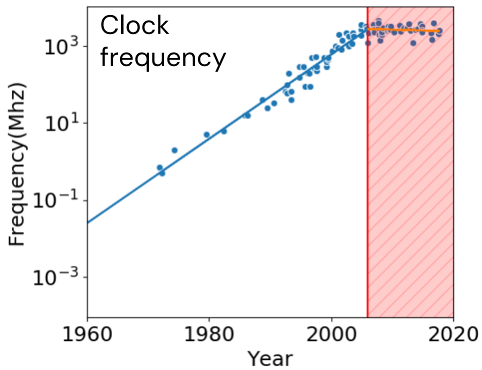
Source: https://github.com/karlrupp/microprocessor-trend-data

Implications

For future electronics

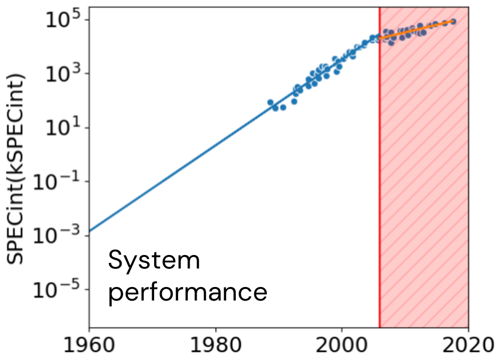
Clock saturation:

Your processors aren't getting much faster.



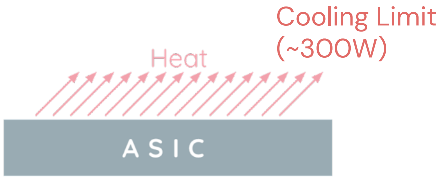
Performance saturation:

Your computers aren't performing better over time.



Dark silicon:

You can't use the whole processor at once (without burning it).

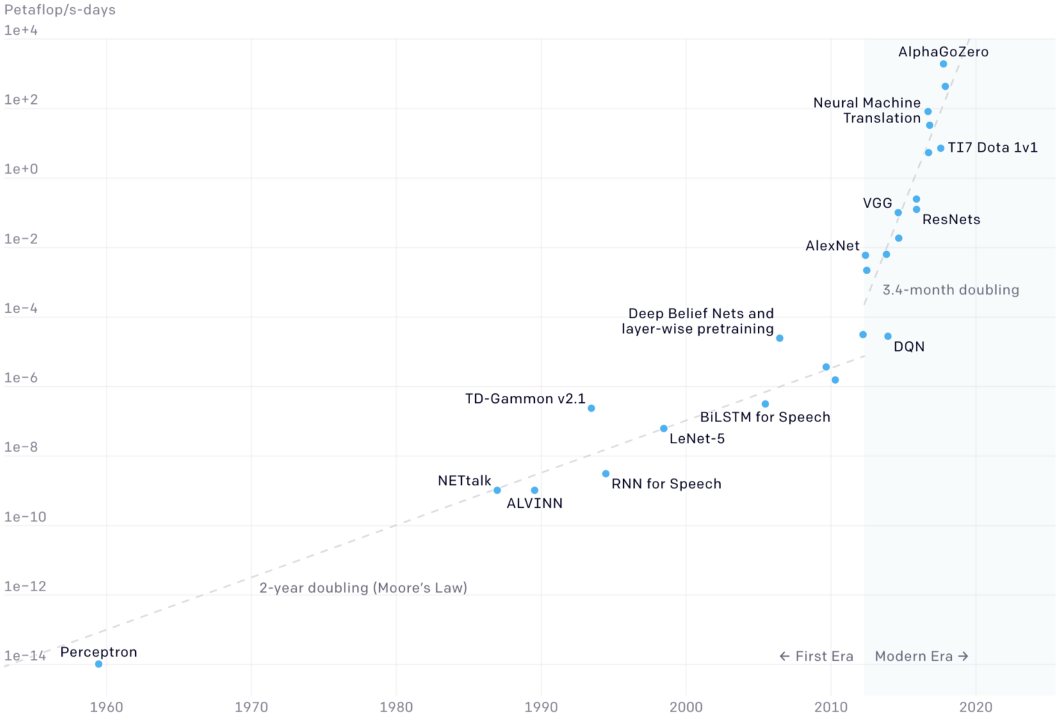


Thermal limit:

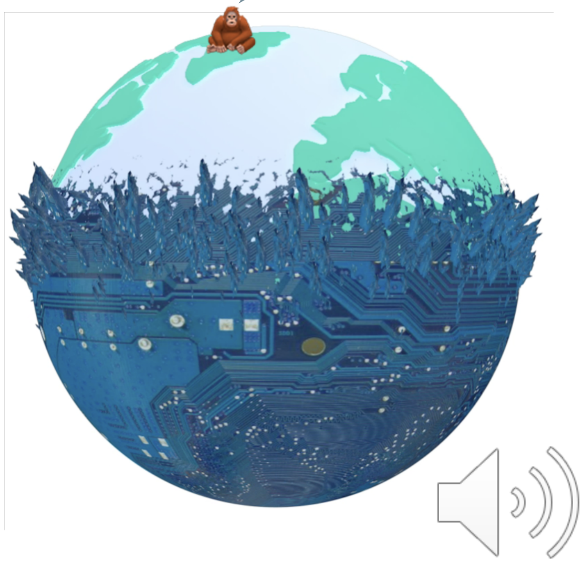
Stacking multiple chips is prohibitive because the processors are too hot.

AI compute requirements

5x the doubling rate of Moore's law



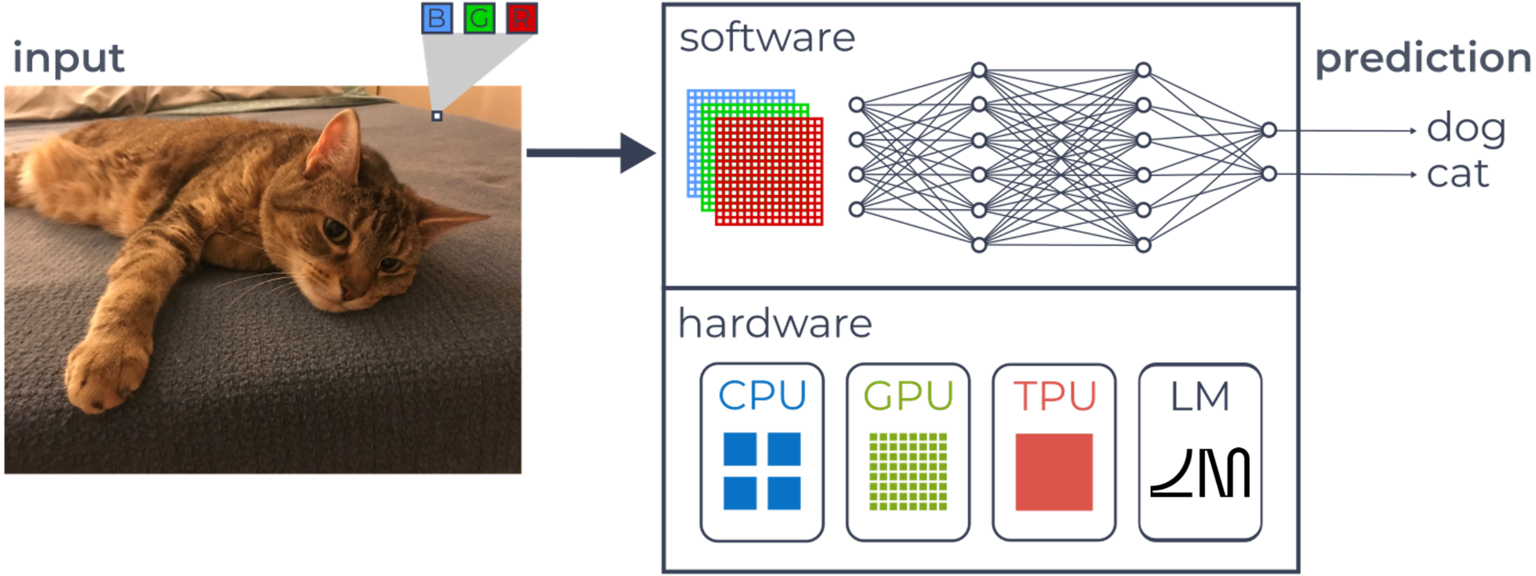
"Cover the planet in datacenters to power AI?"



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

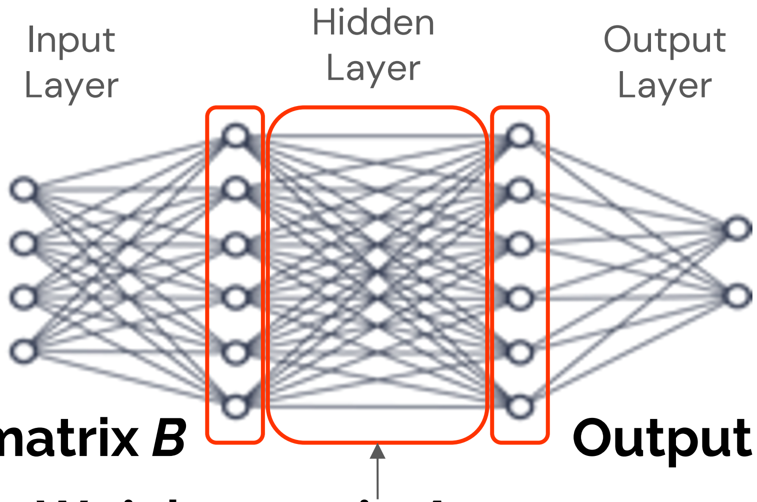
AI acceleration with GEMM

Deep neural networks



General matrix multiply (GEMM)

in deep learning



Input matrix *B*

Output matrix *D*

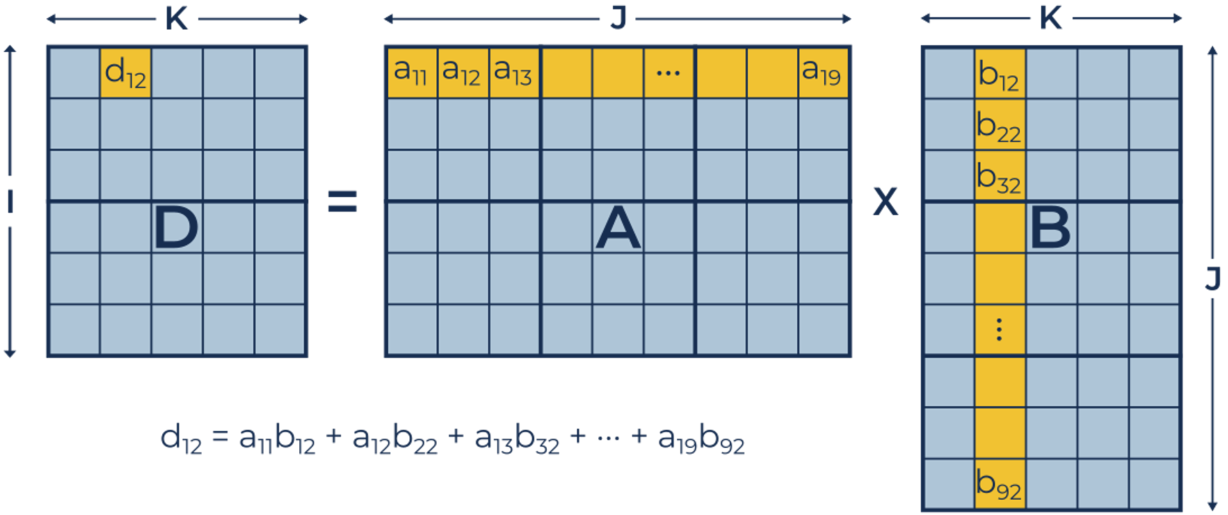
Weight matrix *A*

Representing the weights of the connection in a fully connected layer.

$$D \leftarrow A \times B + C$$

General matrix multiply (GEMM)

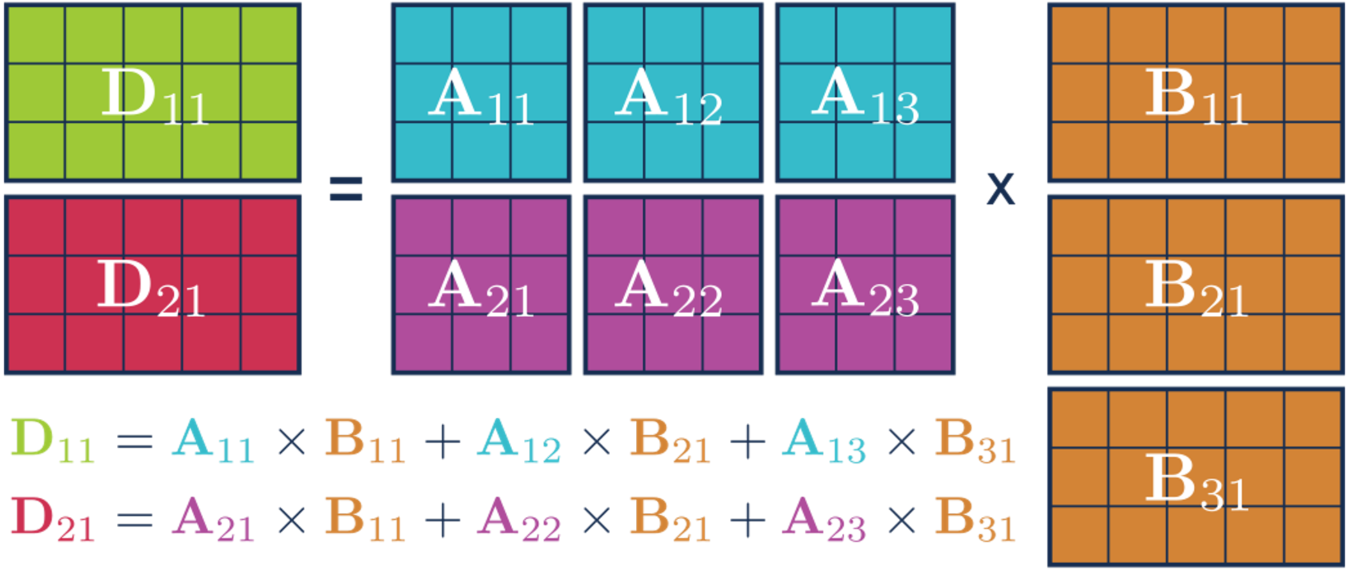
element-by-element



$$d_{12} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} + \cdots + a_{19}b_{92}$$

General matrix multiply (GEMM)

tile-by-tile

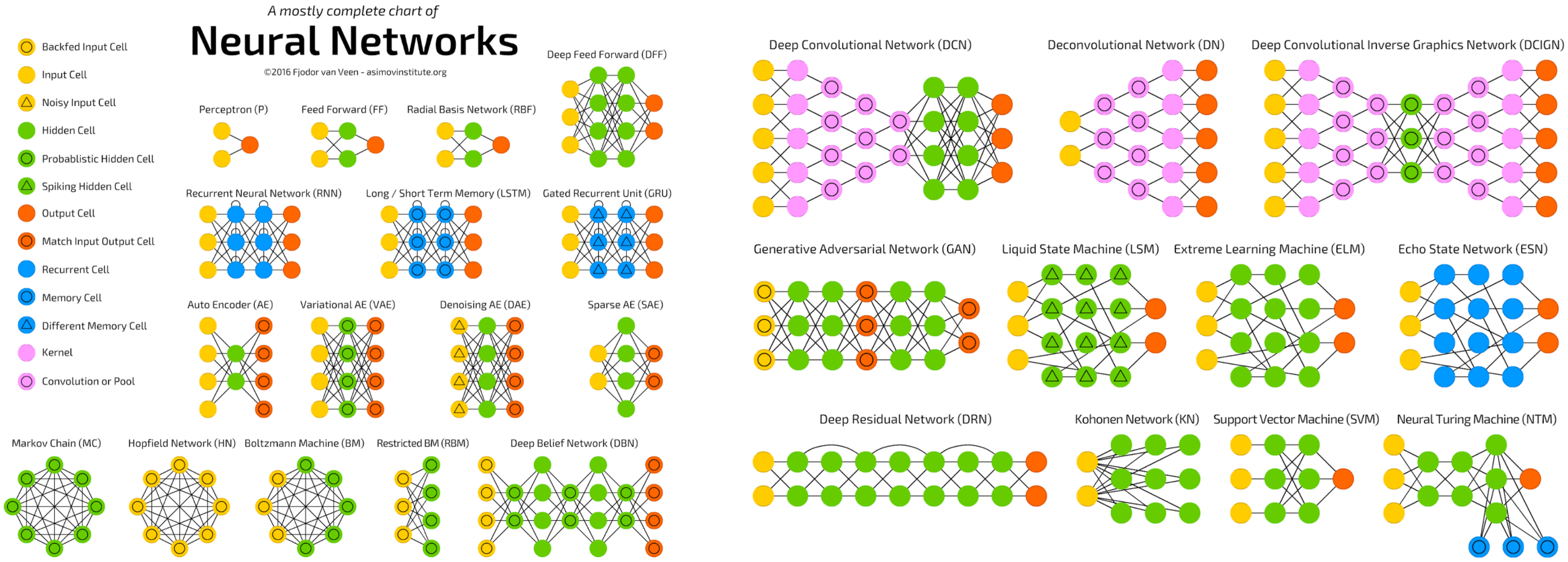


$$D_{11} = A_{11} \times B_{11} + A_{12} \times B_{21} + A_{13} \times B_{31}$$

$$D_{21} = A_{21} \times B_{11} + A_{22} \times B_{21} + A_{23} \times B_{31}$$

GEMM is universal

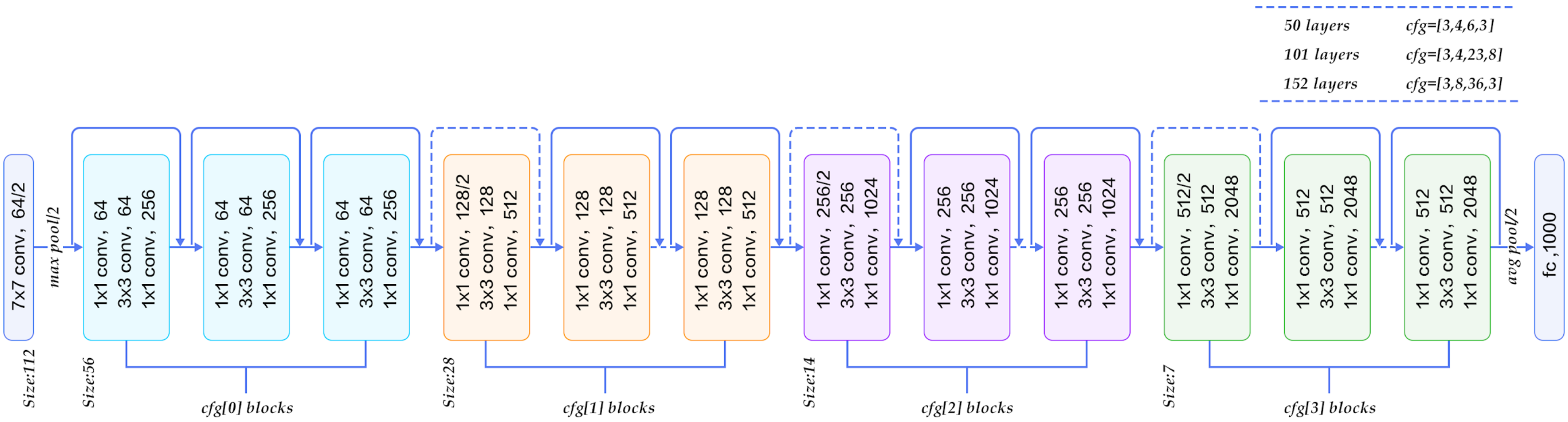
in neural networks



Source: http://www.asimovinstitute.org/neural-network-zoo/

ResNet

A state-of-the-art convolutional neural network (CNN)



Convolution and fully connected layers are linear operations that can be processed with GEMM.

K. He, et al, arXiv:1512.03385

Google Tensor Processing Unit (TPU)

A systolic array for GEMM

In-Datacenter Performance Analysis of a Tensor Processing Unit

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon
Google, Inc., Mountain View, CA USA
jouppi@google.com

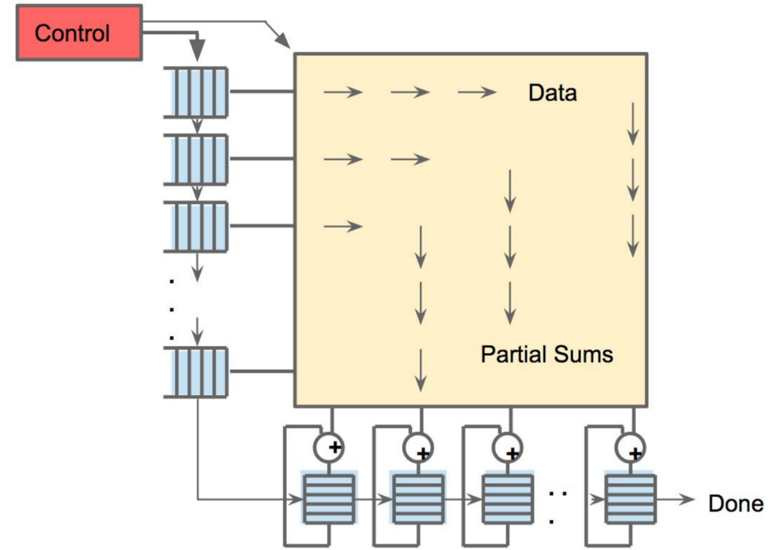


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

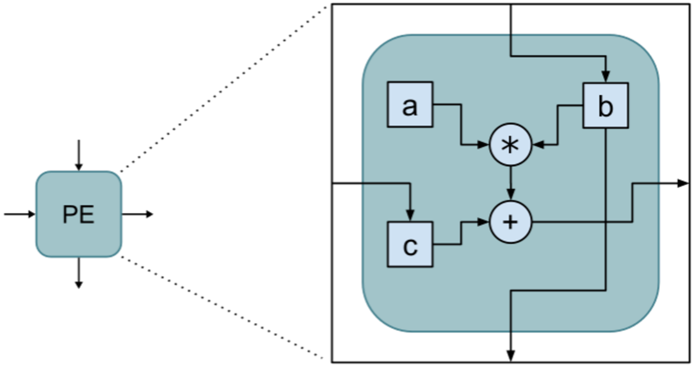
Source: NP Jouppi et al, ISCA, 1-12 (2017).

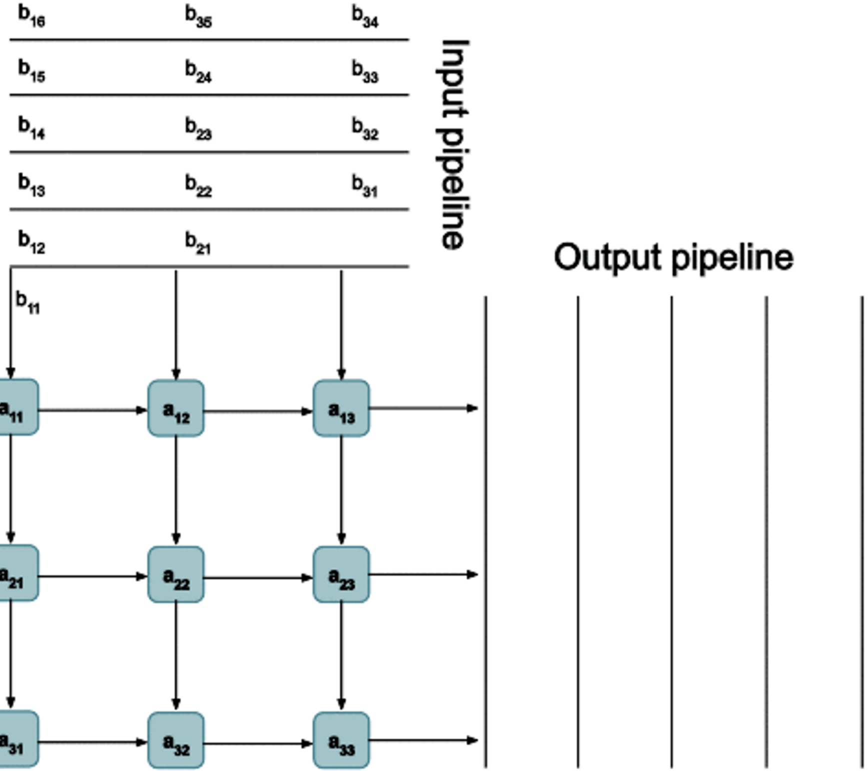
Systolic array multiplier

A small example

Multiply and accumulate (MAC) unit.

$$d \leftarrow a \cdot b + c$$





Matrix A is loaded into the MAC units, and matrix B is streamed through.

Example:

shape(A) = [3 x 3]

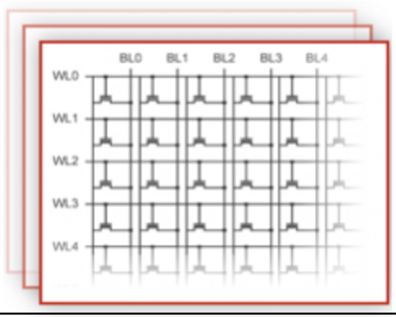
shape(B) = [3 x 9]

Accelerating AI with photonics

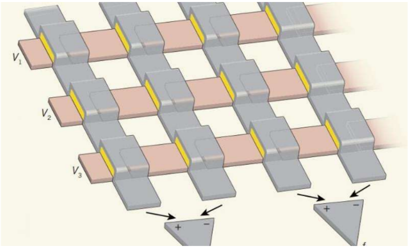
Analog computation

A zoo of GEMM accelerators

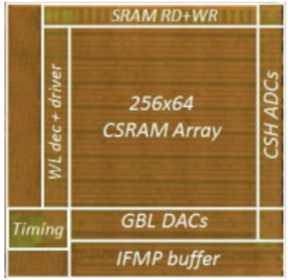
Flash



Memristor



SRAM

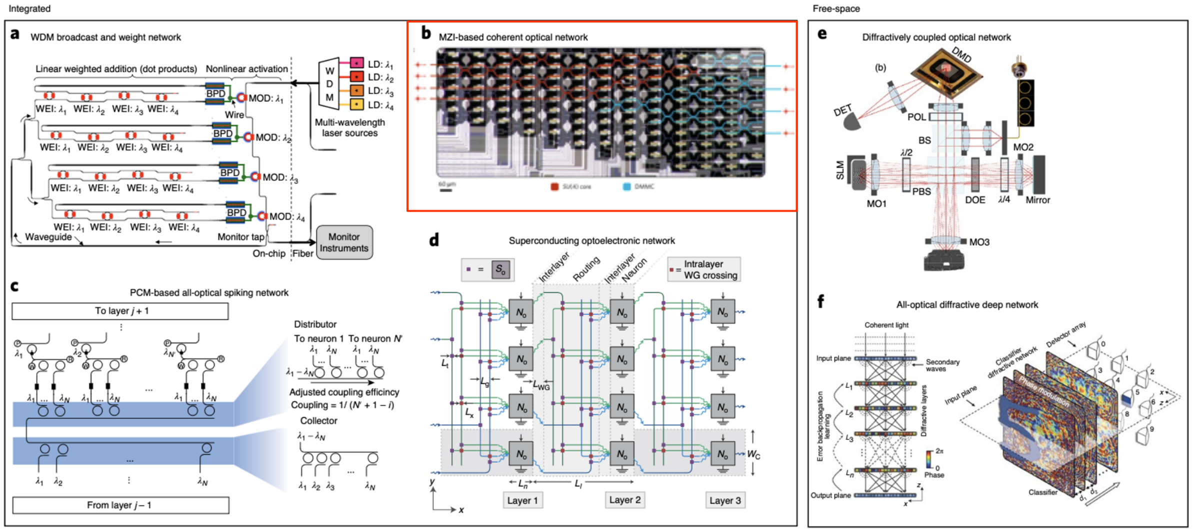


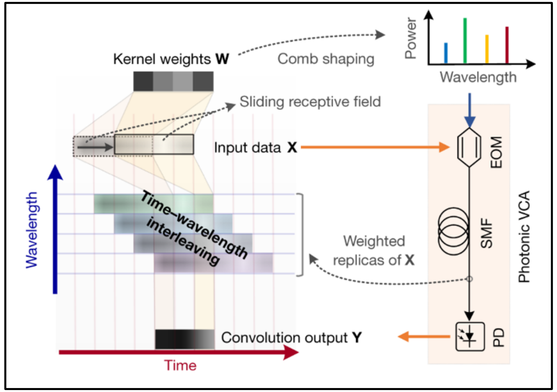
Photonics

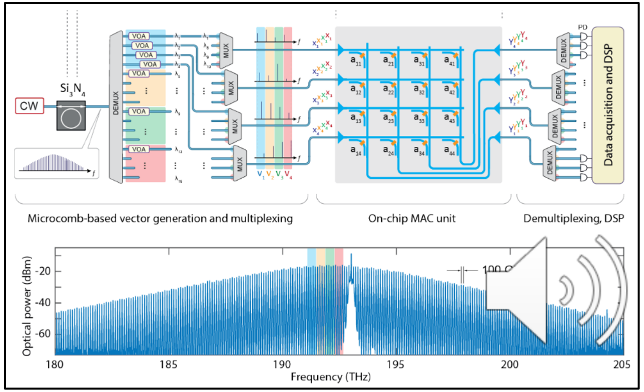
A. Biswas et al., JSSC 54, 217-230 (2018).
S. Ambrogio et al., Nature 558, 60–67 (2018).
D. Fick et al., Hot Chips (2018).

Computation with photonics

A zoo of photonic GEMM accelerators







Reviews:

- B.J. Shastri, et al. Nat. Photonics 15, 102–114 (2021)
- G. Wetzstein, et al. Nature 588, 39–47 (2020)

Recent works:

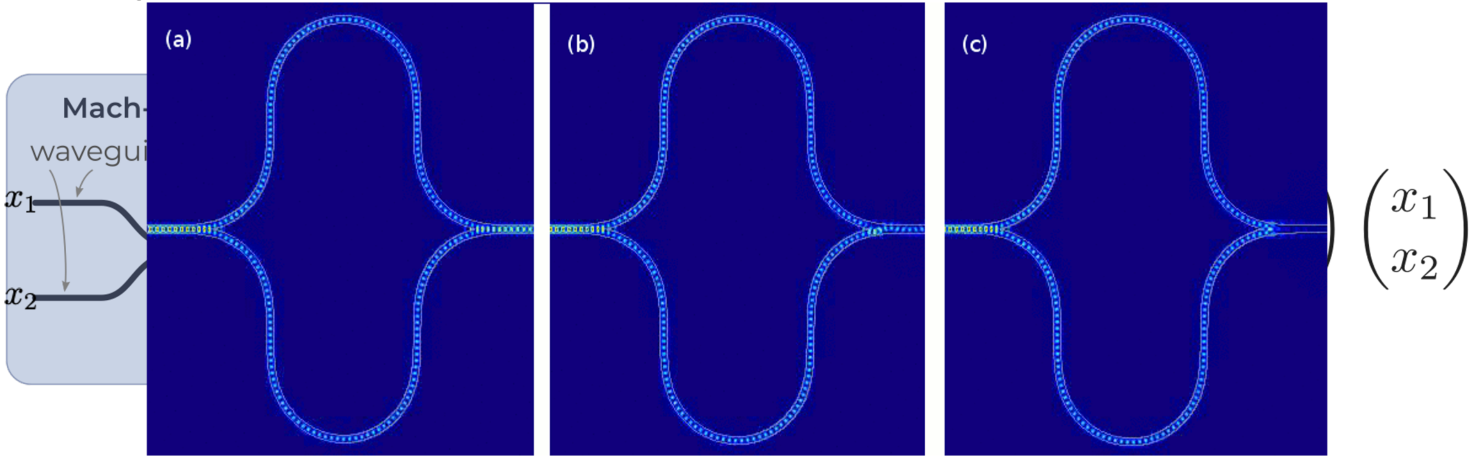
- X. Xu, et al. Nature 589, 44–51 (2021)
- J. Feldmann, et al. Nature 589, 52–58 (2021)

Mach-Zehnder Interferometer (MZI)

Basic building blocks for silicon photonics GEMM accelerators

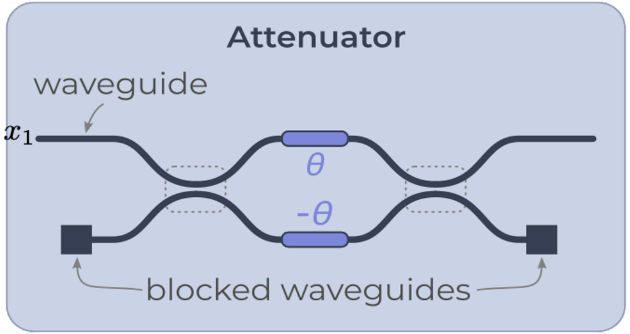
MZI

Interferes light between two waveguides with a relative phase-shift control



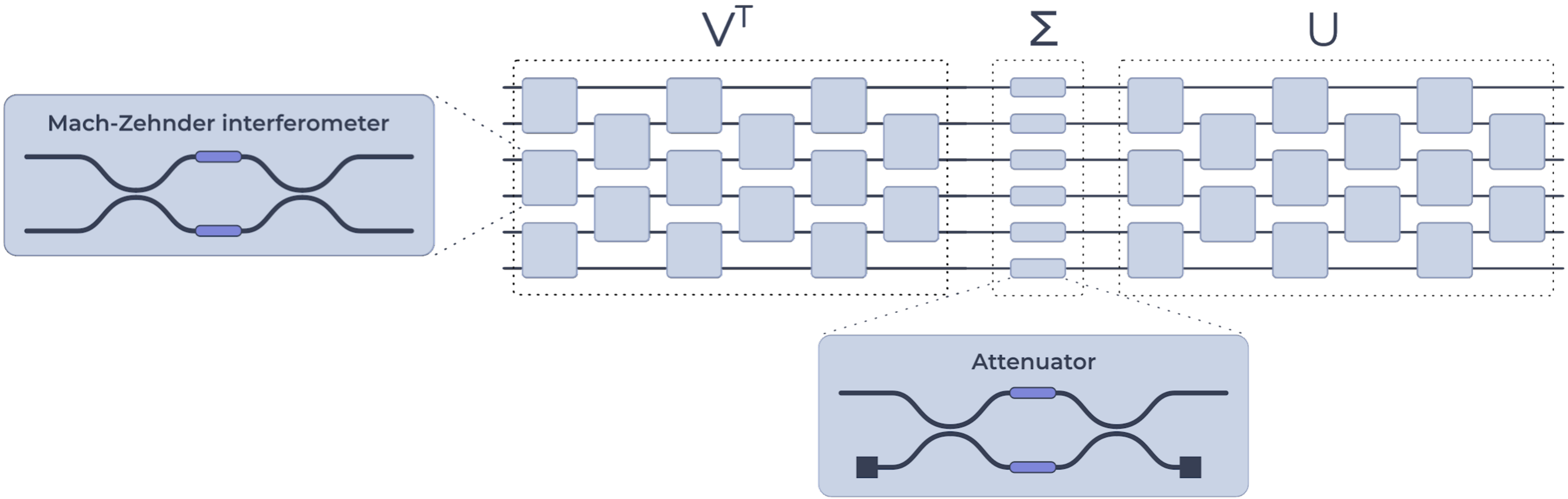
Attenuator

Uses MZI to attenuate the signals



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

Photonic GEMM accelerator



Singular Value Decomposition

$$\mathbf{M} = \mathbf{U\Sigma V}^T$$

N Harris et al, Optica 5, 1623-1631 (2018).
WR Clements et al, Optica 3, 1460-1465 (2016).
DAB Miller, Photon. Res. 1, 1-15 (2013).

Performance scaling

A photonic GEMM multiplier has the same maximum throughput as systolic arrays (e.g., Google TPU)

$$\mathrm{OP/s} \approx 2N^2 f_c$$

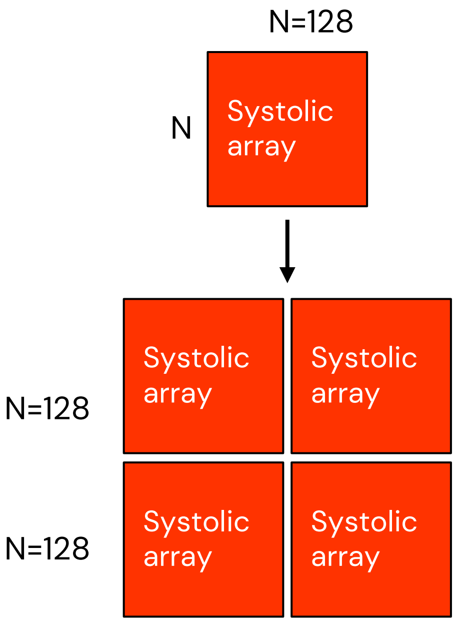
Clock frequency

Energy scaling

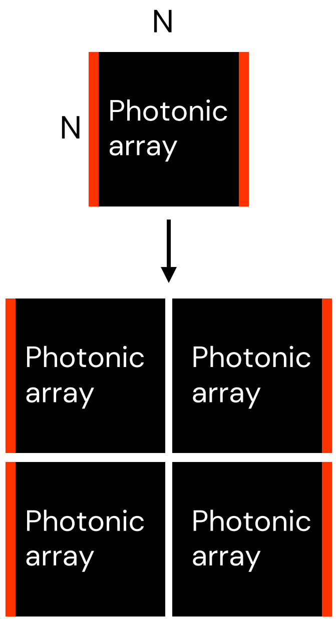
Solving the power problem

Each MAC unit in a systolic array burns power throughout the computation.

Power consumption scales by $O(N^2)$.



4x array →~4x energy consumption



4x array →~2x energy consumption

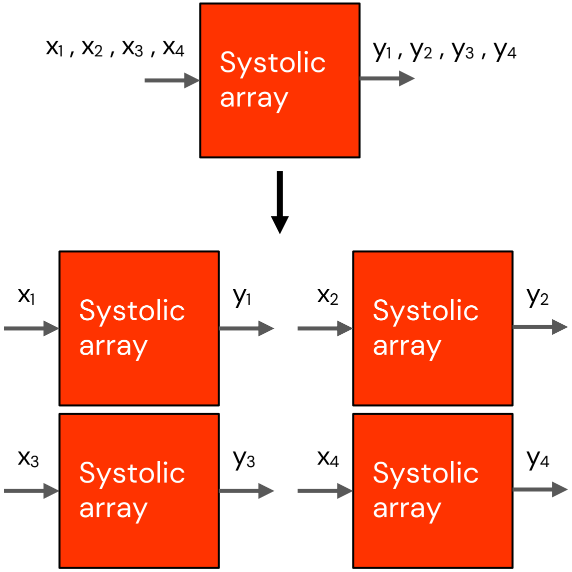
A photonic array mainly only uses input DACs and output ADCs at the edges for its computation.

Power consumption scales by $O(N)$.

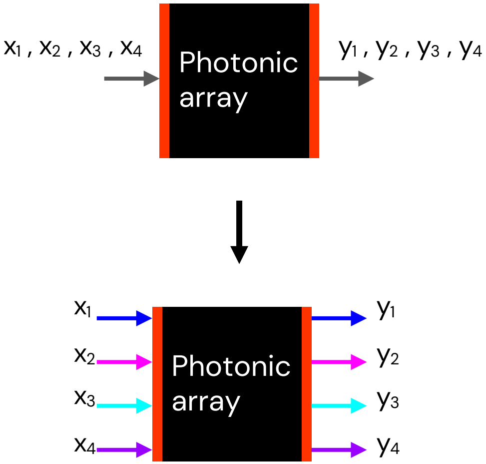
Wavelength-division multiplexing

Solving the area and weight reuse problem

Increasing the throughput of a systolic array (at the same f_c) requires increasing the number of systolic arrays.



4x throughput→~4x area, no weight reuse



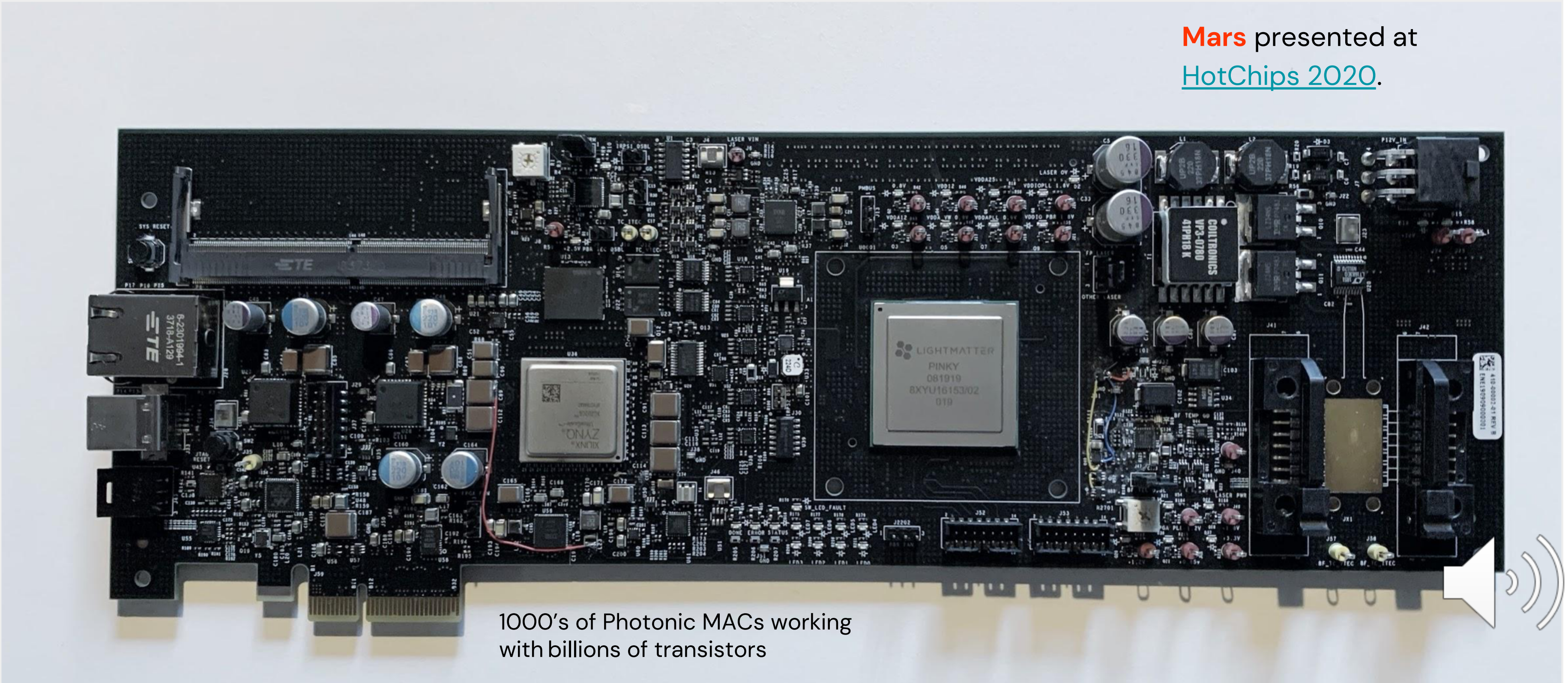
4x throughput →~same area, 4x weight reuse

Wavelength-division multiplexing allows for the simultaneous processing of vectors through the array.

Ideal for machine learning models as multiple operations are typically performed across several cycles of stationary weights.

Photonic Computing is Here

Faster, lower energy, and decoupled from Moore's Law.



Mars presented at HotChips 2020.

1000's of Photonic MACs working
with billions of transistors

Looking ahead

Envise

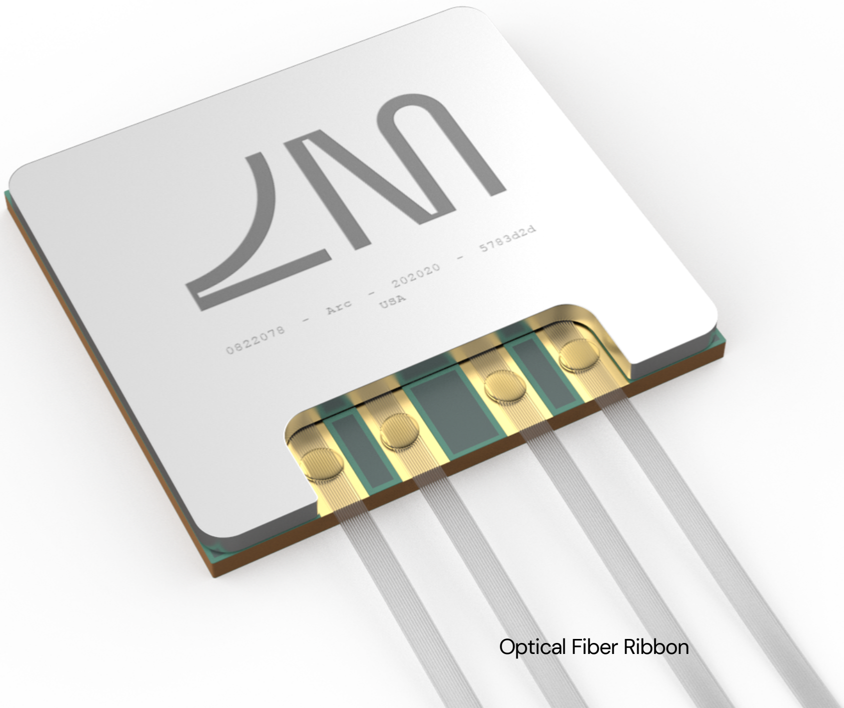
Combining photonics and electronics in a single, compact package

Features

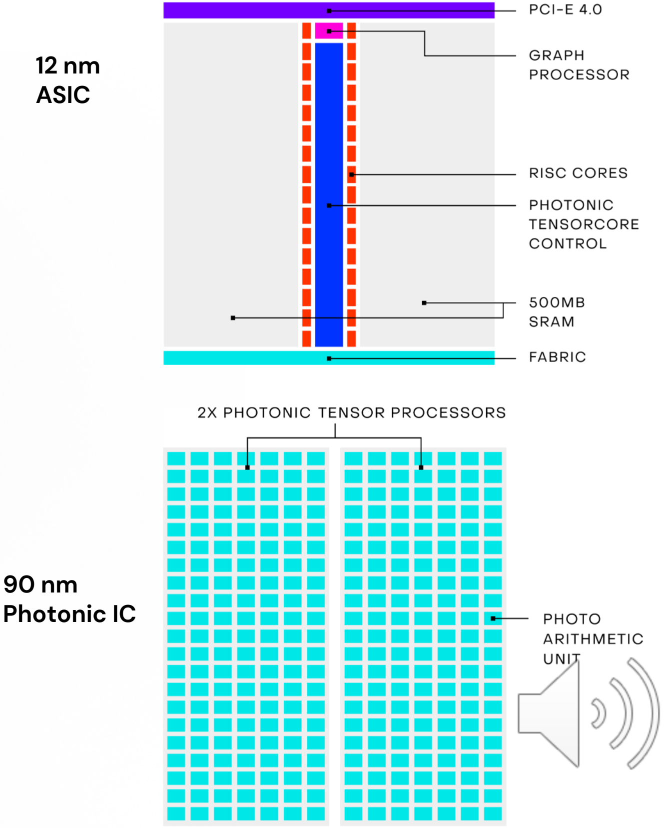
Dual photonic tensor cores
3 GHz Clock
256 RISC cores
500MB SRAM
80W Typ., 130W TDP
3D LGA Package
Reliability, availability, and serviceability (RAS) features

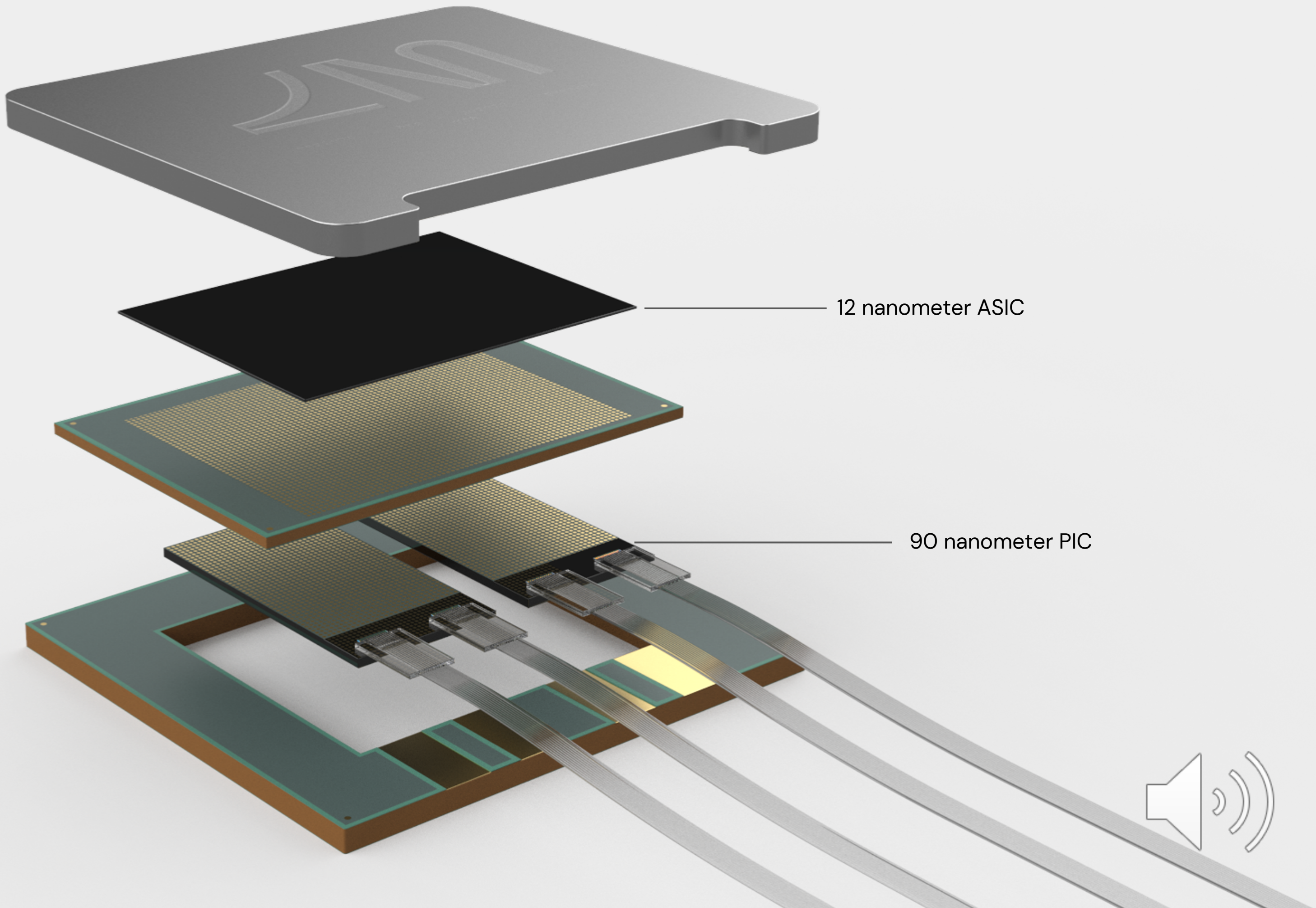
Interfaces

PCI-e 4.0
Local Optical Interconnect (400Gbps)



Optical Fiber Ribbon



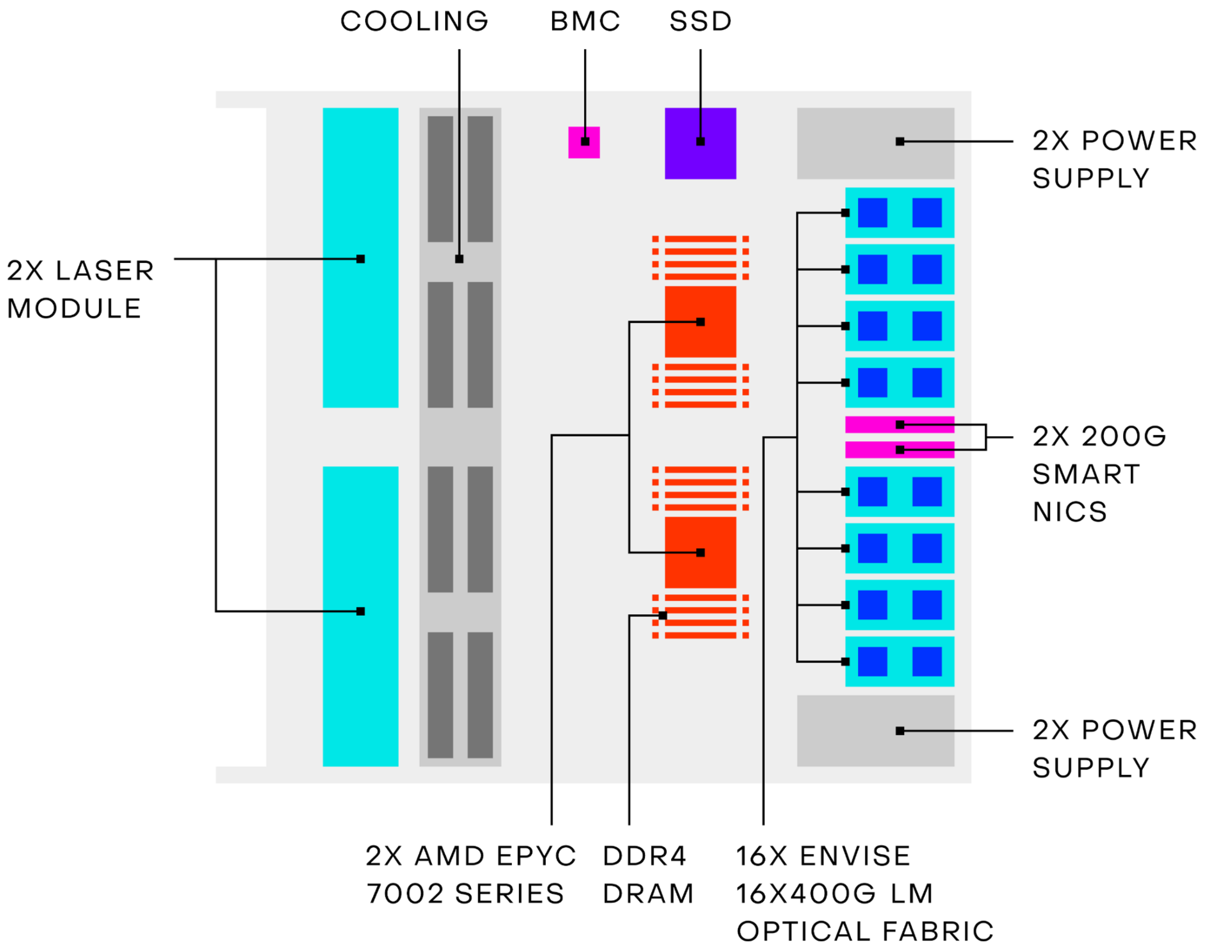

12 nanometer ASIC
90 nanometer PIC

Envise Blade

A general-purpose AI accelerator system

Features

- 16x Lightmatter Envise
- 4U form factor
- 3kW TDP



Envise Blade

- 16x Lightmatter Envise
- 2x AMD EPYC host processors
- 3TB NVMe SSD
- 6.4 Tbps optical fabric interconnect
- 2x 200Gbps Ethernet Smart NIC
- 2x laser modules

IDIOM

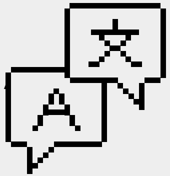
FEATURES

PyTorch TensorFlow ONNX

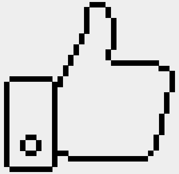
↕

Model	Compile & Execute	Debug	Profile
Simulate the effects of model parameters on accuracy and performance	Map model to hardware Optimize model performance Execute generated code	Access intermediate state in the model under execution	Find and fix performance and resource bottlenecks

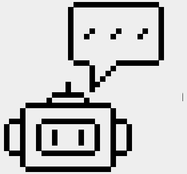
General-Purpose AI Inference Compute



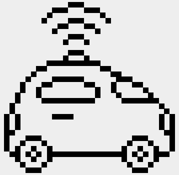
Machine Translation



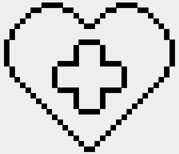
Recommendation



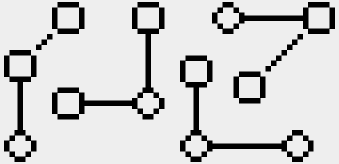
Conversation



Autonomous Driving



Healthcare



Sentiment Analysis