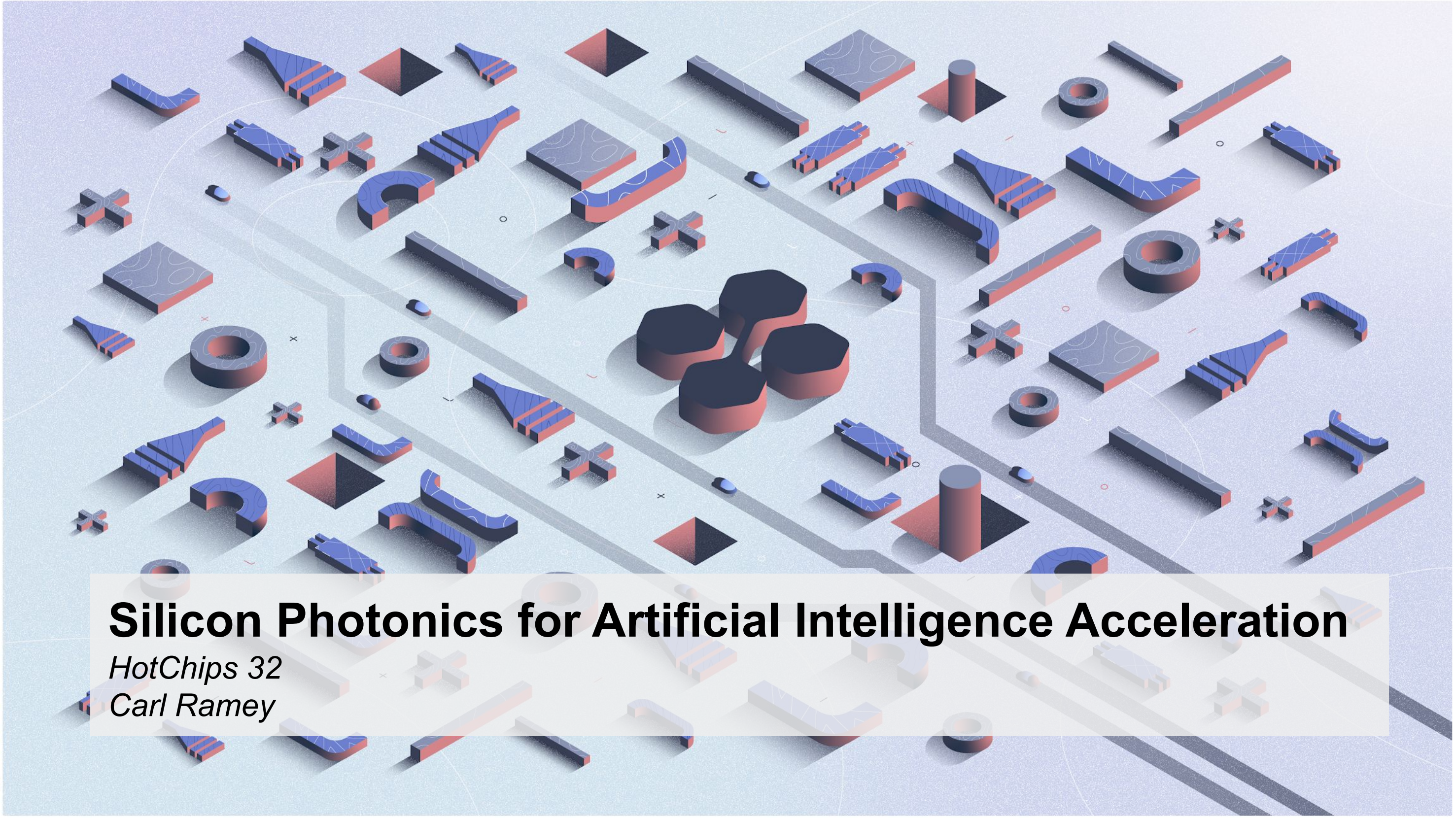
LIGHTMATTER



# Silicon Photonics for Artificial Intelligence Acceleration

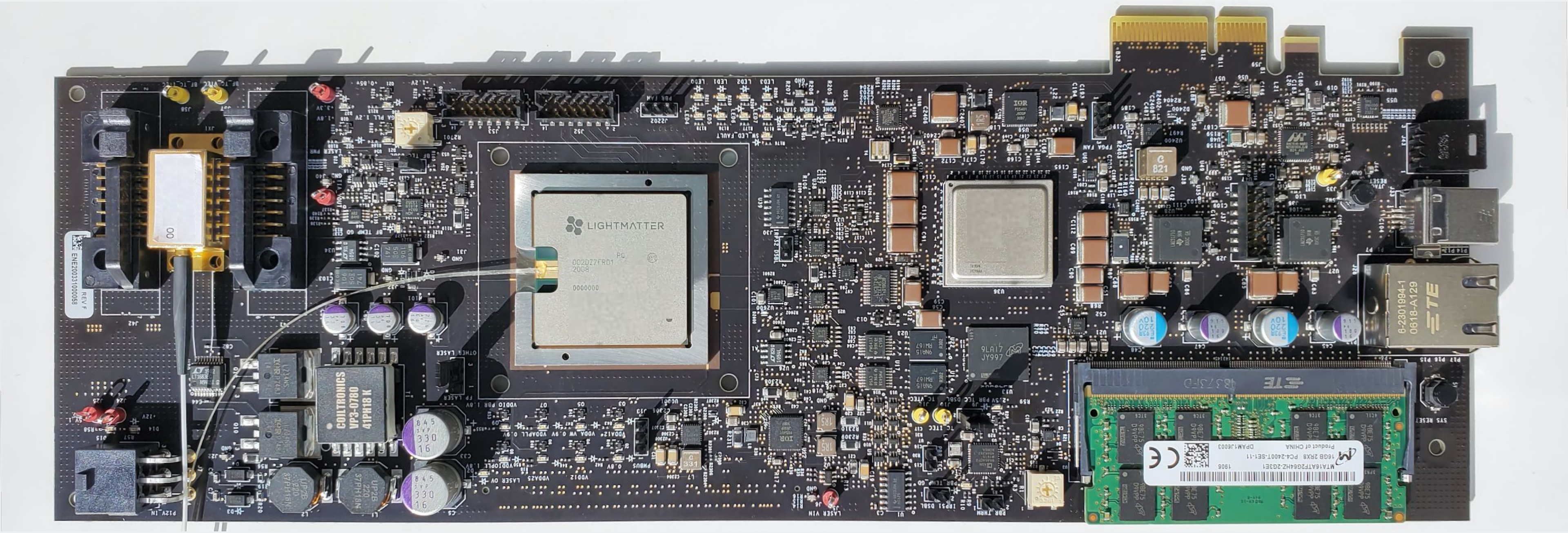
*HotChips 32*

*Carl Ramey*

lightmatter.co

Accelerating AI With Light

# A photonics compute platform to address AI demands

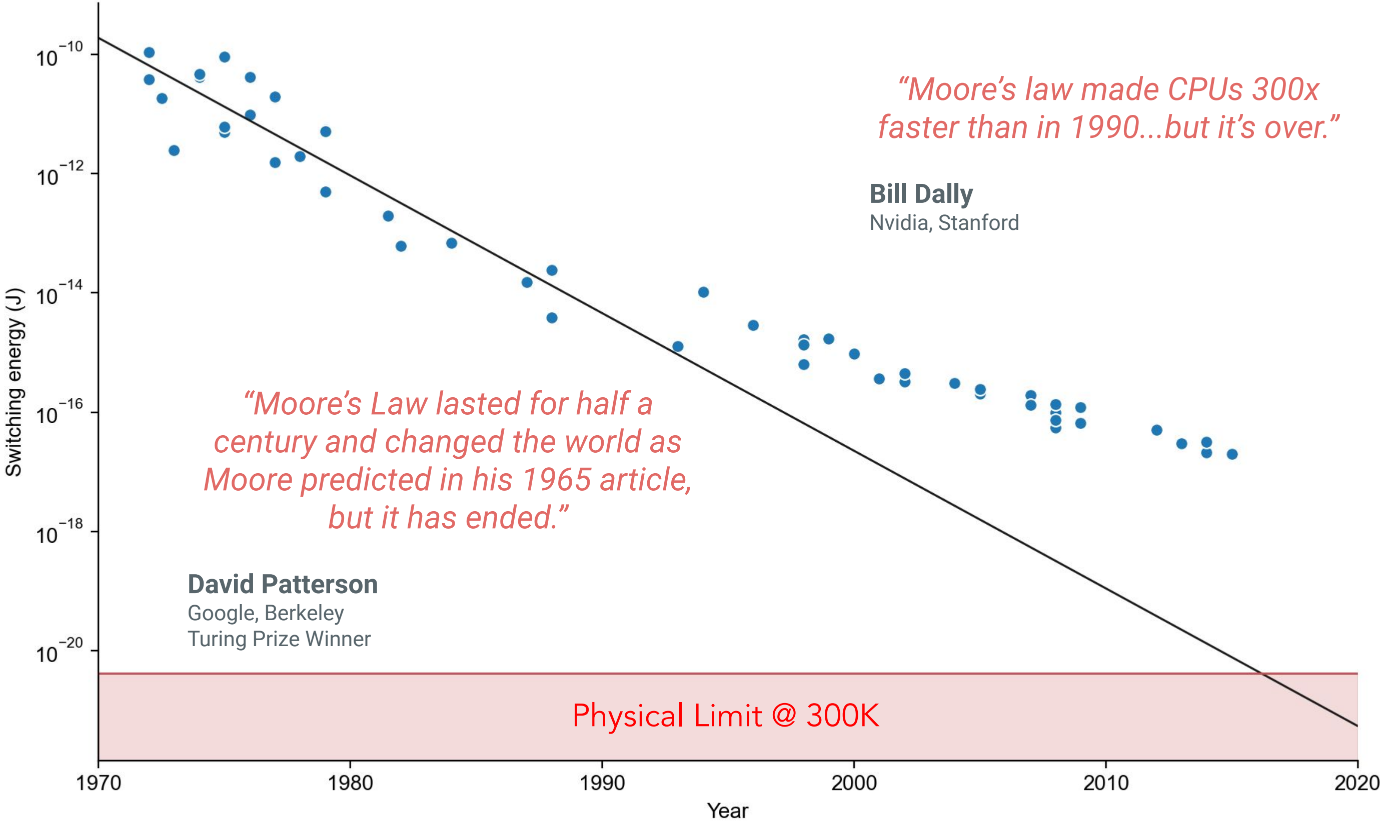



- Lightmatter's *Mars* device accelerates AI workloads
- Core computations performed optically using silicon photonics
- Multi-chip solution to get the best of transistor and photonics technology
- Photonics provides unprecedented opportunities for performance and efficiency

# Section 1

# Motivation

# Transistors aren’t getting more efficient




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

# ⇒Chips are getting hotter

400 Watts <u>per chip package</u>
is a practical limit

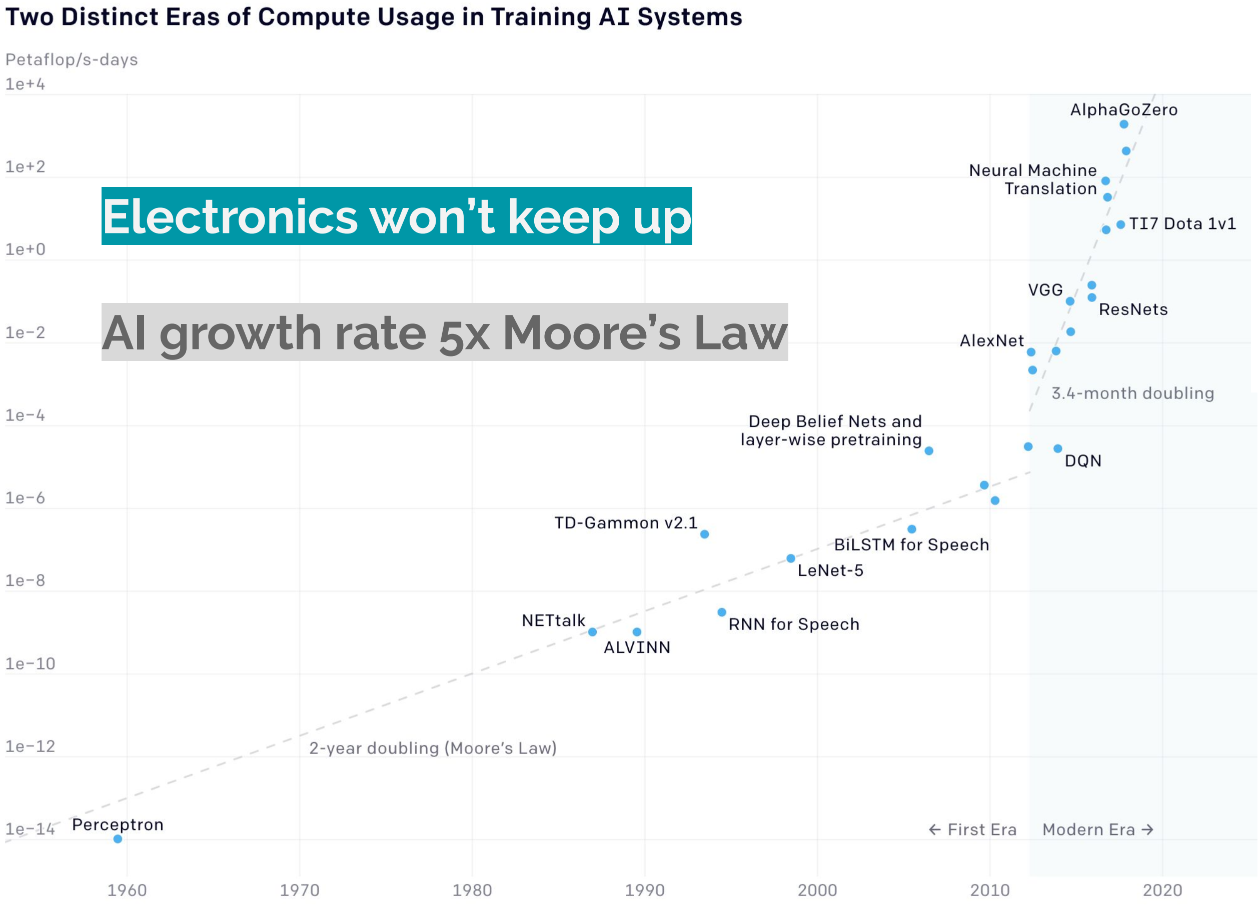
**Dark Silicon**
You can't use the whole chip at once

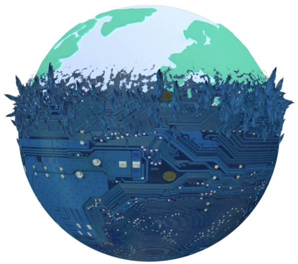
**Clock Frequency Saturation**
Performance per core is limited

**Exotic Cooling Solutions**
Expensive

# Electronics won't keep up

## AI growth rate 5x Moore's Law



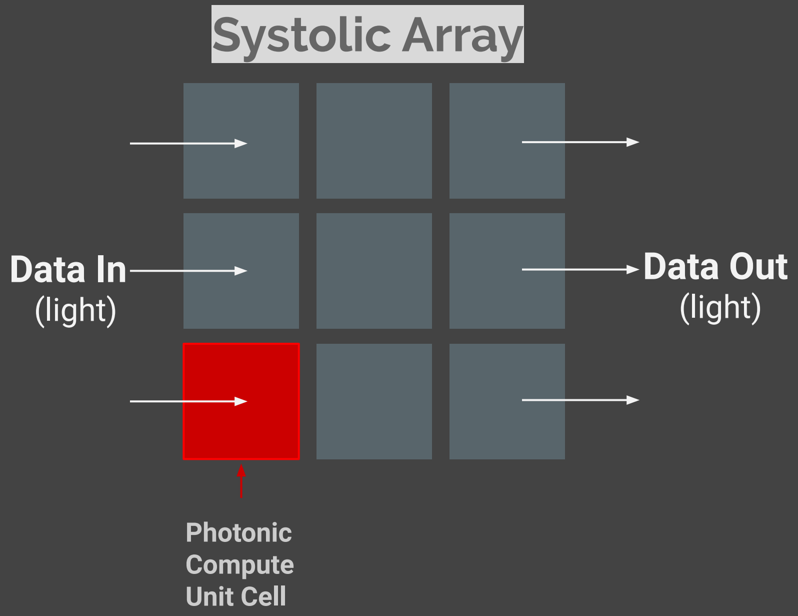



Source: OpenAI

# Section 2
# Photonics Core

# Why photonics?

## Fundamental benefits independent of process node




### Optical data transport

Less energy spent moving data

### Optical tensor core

Higher clock frequency, less energy, same size, lower latency

### Parallel Processing

Wavelength- and polarization-division multiplexing. N processors, 1 physical resource.

| | Photonics | Electronics | Improvement |
|---|---|---|---|
| Latency | 100 ps | 100 ns | $10^3$ |
| Bandwidth | 20 GHz | 2 GHz | 10 |
| Power | 1 uW | 1 mW | $10^3$ |
| Area | 2500um$^2$ | 2500um$^2$ | 1 |

# Optical data transport

Less energy spent moving data

$L = 3cm$

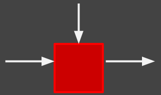

$$P \geq \frac{1}{2} L \cdot C_{wire} v^2 f \cdot N_{lanes}$$

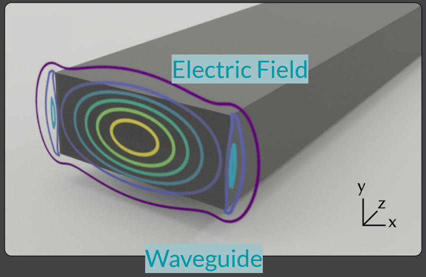
10s of Watts just for passive data transport with electronics.

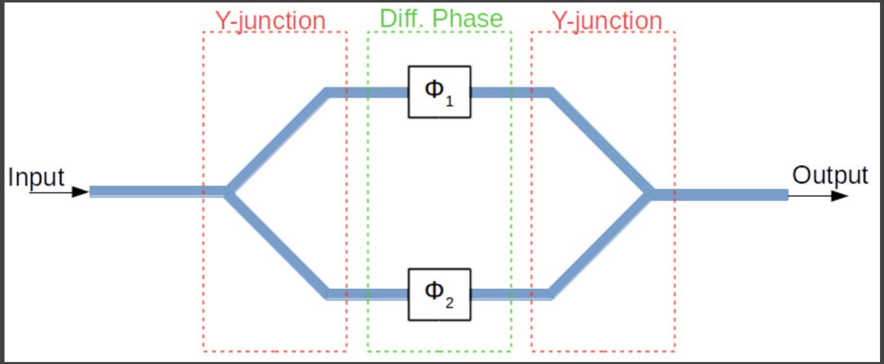
~Free with optics...and no length-dependent RC time constant!

# Mach Zehnder Interferometer (MZI)

Compute tile

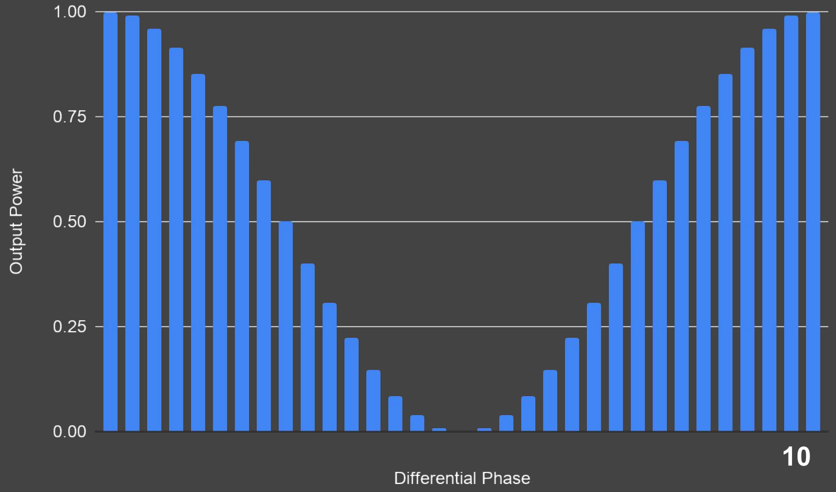




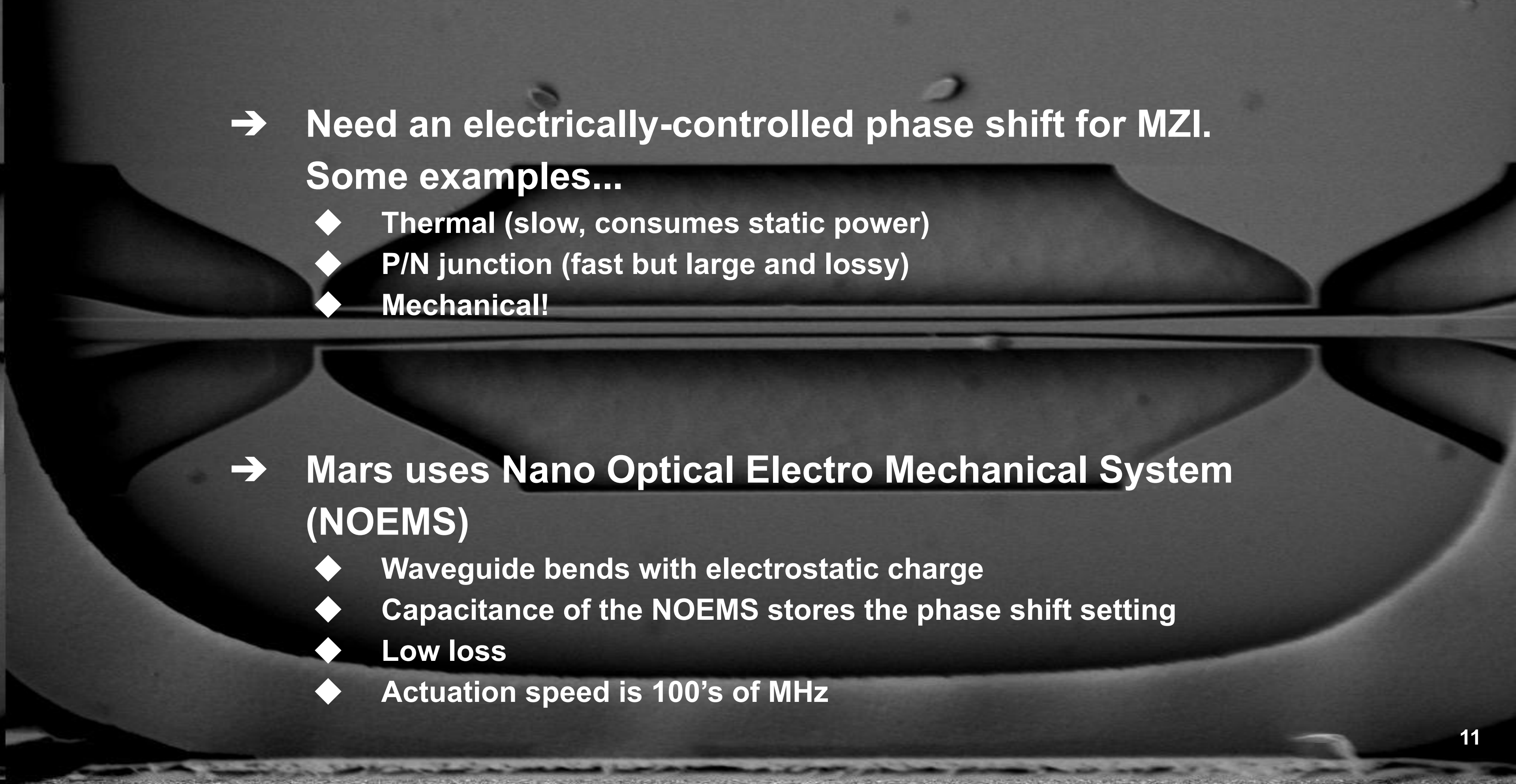




MZIs provide observation of phase shift through interference

These become useful when you modulate the phase on $\Phi_1$ and $\Phi_2$



# Programmable Phase Shifters

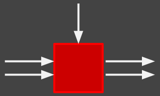


- ➔ Need an electrically-controlled phase shift for MZI. Some examples...
  - Thermal (slow, consumes static power)
  - P/N junction (fast but large and lossy)
  - Mechanical!

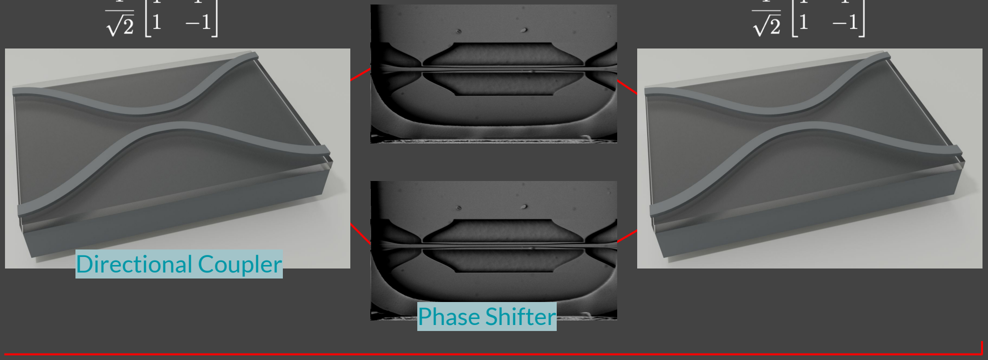
- ➔ Mars uses Nano Optical Electro Mechanical System (NOEMS)
  - Waveguide bends with electrostatic charge
  - Capacitance of the NOEMS stores the phase shift setting
  - Low loss
  - Actuation speed is 100’s of MHz

# Optical Vector MAC

Compute tile



$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix} \quad \begin{bmatrix}e^{i\phi} & 0\\ 0 & e^{-i\phi}\end{bmatrix} \quad \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$




$$\begin{bmatrix}i\sin\phi & -\cos\phi\\ \cos\phi & -i\sin\phi\end{bmatrix}$$

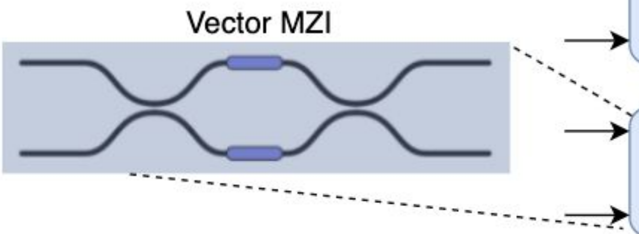
Directional couplers combine input signals in pairs

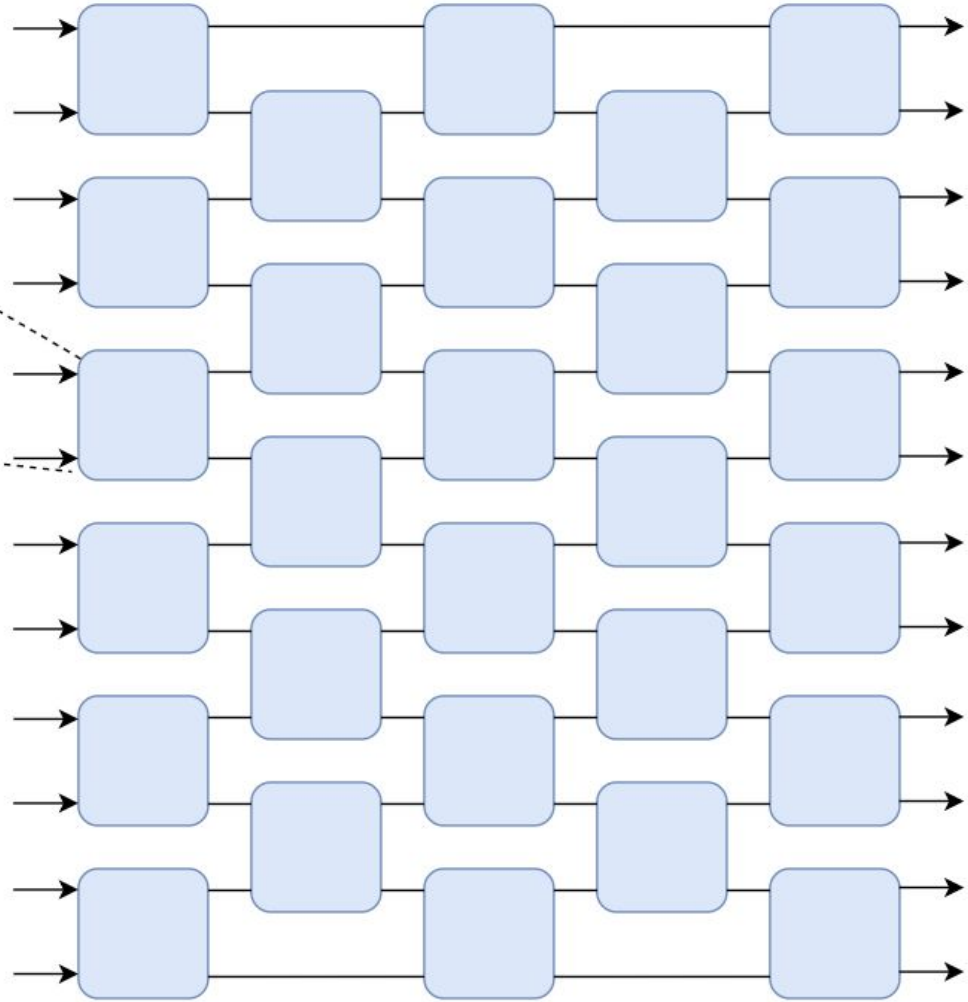
Phase shifters provide programmability

OV MAC computes a 2x2 matrix multiplied by a 1x2 vector

**We just did a 2x2 MVP at the speed of light with near zero power!**
**No RC delays, no dynamic power.**

# Arrays of MZIs



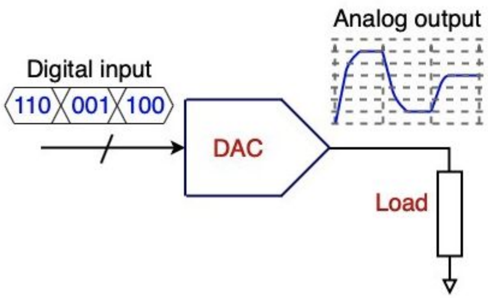



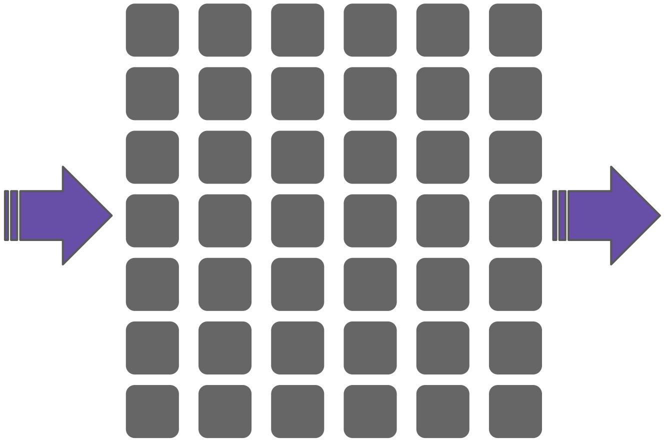
Arrays of 2x2 MZIs provide matrix • vector products

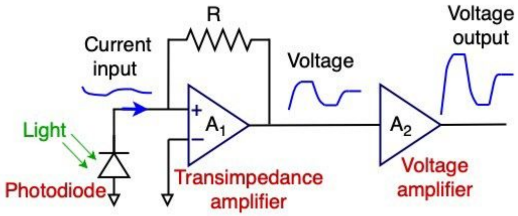
$$\begin{pmatrix} M_{0,0} & \dots \\ \dots & M_{N,N} \end{pmatrix} \begin{pmatrix} Vin_0 \\ \dots \\ Vin_N \end{pmatrix} = \begin{pmatrix} Vout_0 \\ \dots \\ Vout_N \end{pmatrix}$$

# Data Conversion at Edges of Square

DACs encode input vectors

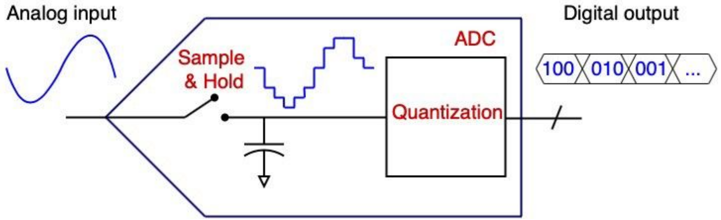




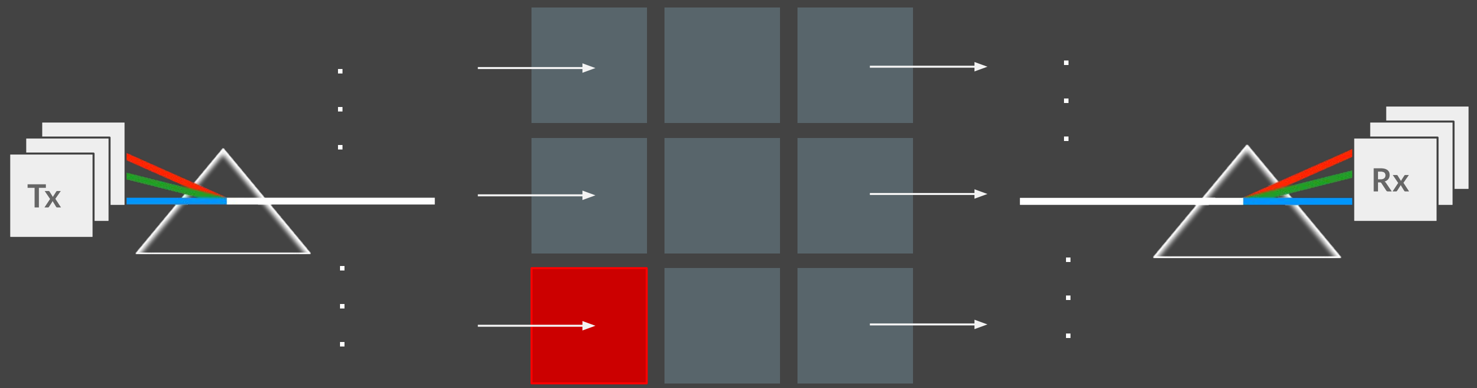

Detectors and ADC convert output vector to digital

- ❏ Mars: 64 DACs and 64 ADCs for 4096 MAC operations
- ❏ Performance scales with area ← like electronics
- ❏ Power scales with sqrt(area) ← this is interesting!



# Parallel Processing

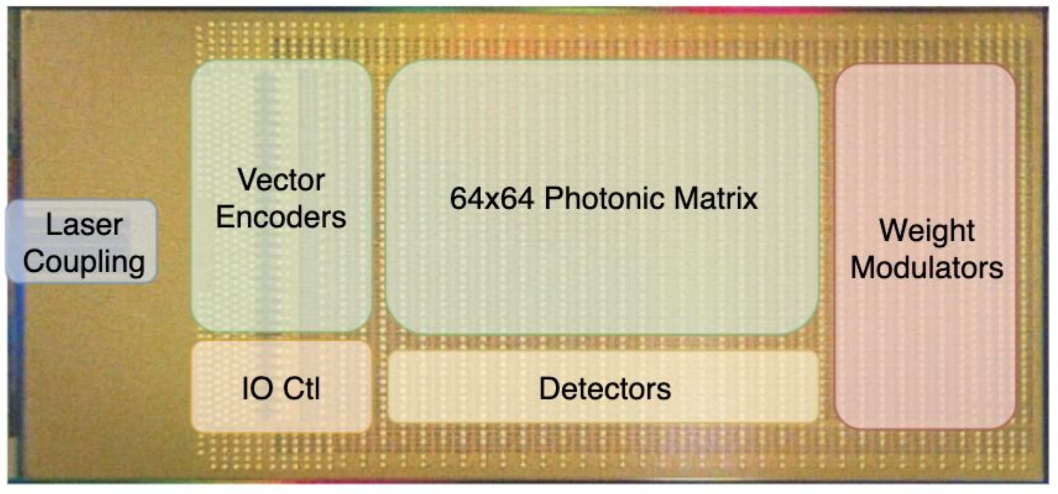
Wavelength- and polarization-division multiplexing. N processors, 1 physical resource.




## 1 Processor ⇒ ‘N’ Parallel Instances

# Mars Photonics Core

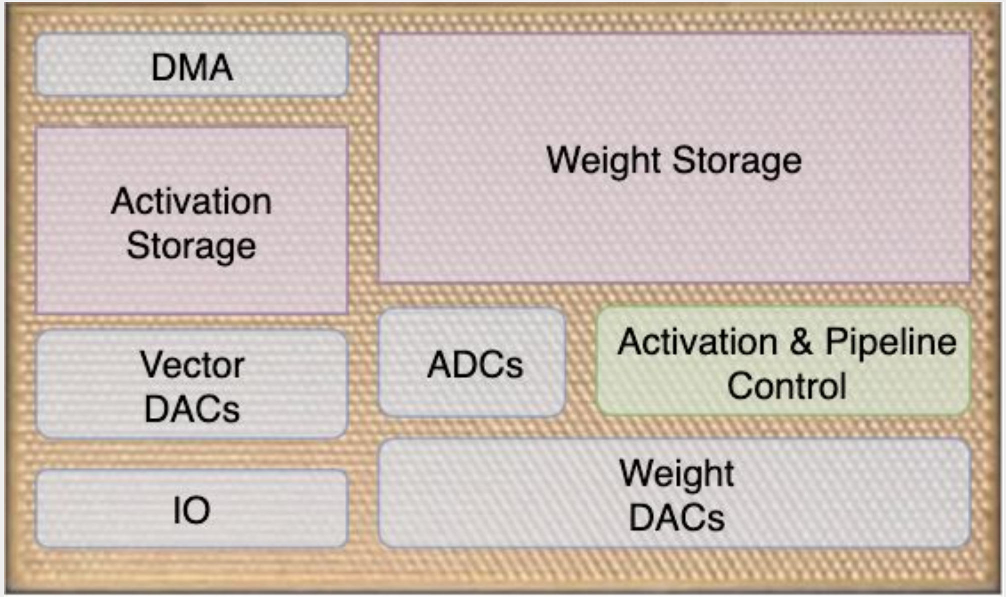
- ❏ 64x64 Matrix * 64 element vector
  - ❏ 8K ops per “cycle”
- ❏ 1GHz vector rate
- ❏ 50mW laser
- ❏ 8-bit signed operands
- ❏ 200ps latency
- ❏ 90nm standard photonics process
- ❏ 150mm$^2$



# Section 3
# Digital System

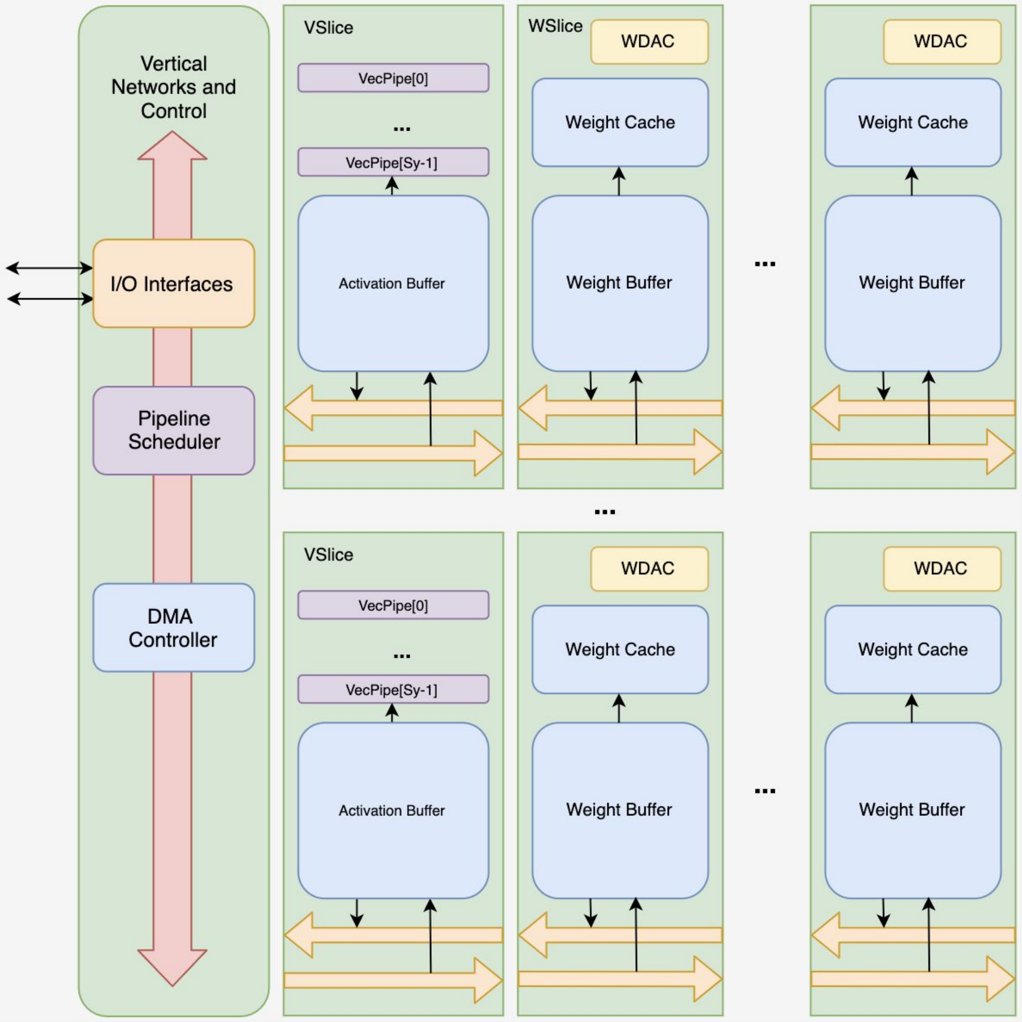
# Mars SoC

- 14nm custom ASIC
- $50mm^2$
- Analog interfaces to Photonics Core
- SRAM for weights and activations
- I/O interfaces
- Digital offloads:
  - Non-linearities (GELU, sigmoid etc.)
  - Scaling and accumulation



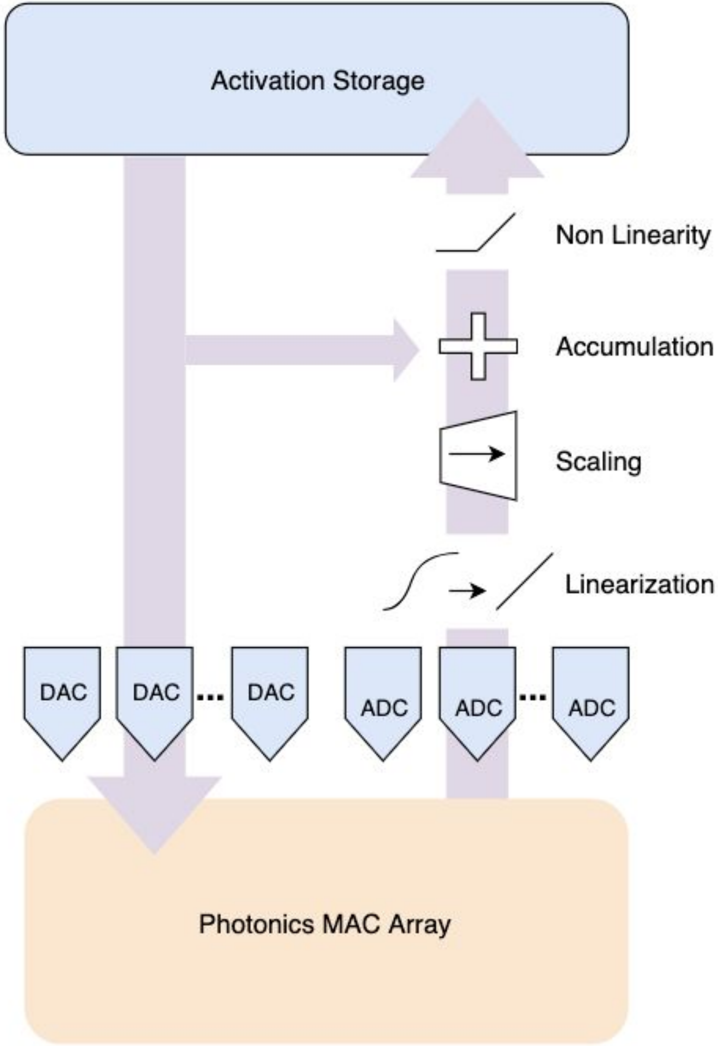
# Digital Architecture

- ❑ **Buffer organization minimizes data movement**
- ❑ **Heterogeneous on chip networks**
- ❑ **Single fully synchronous pipeline scheduler**

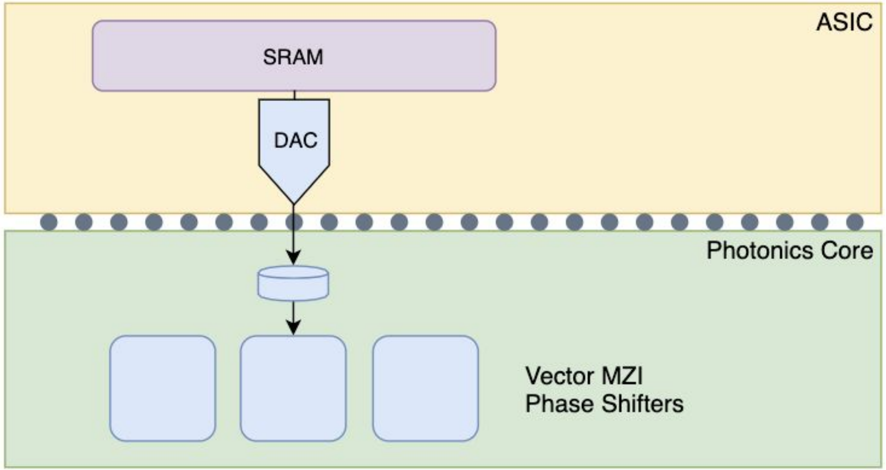


# Activation Pipeline

- ❑ **Digital pipeline maintains operand locality**
- ❑ **SRAM access minimized**
- ❑ **True pipelining for small batch performance**
- ❑ **Photonics outputs physically close to inputs**



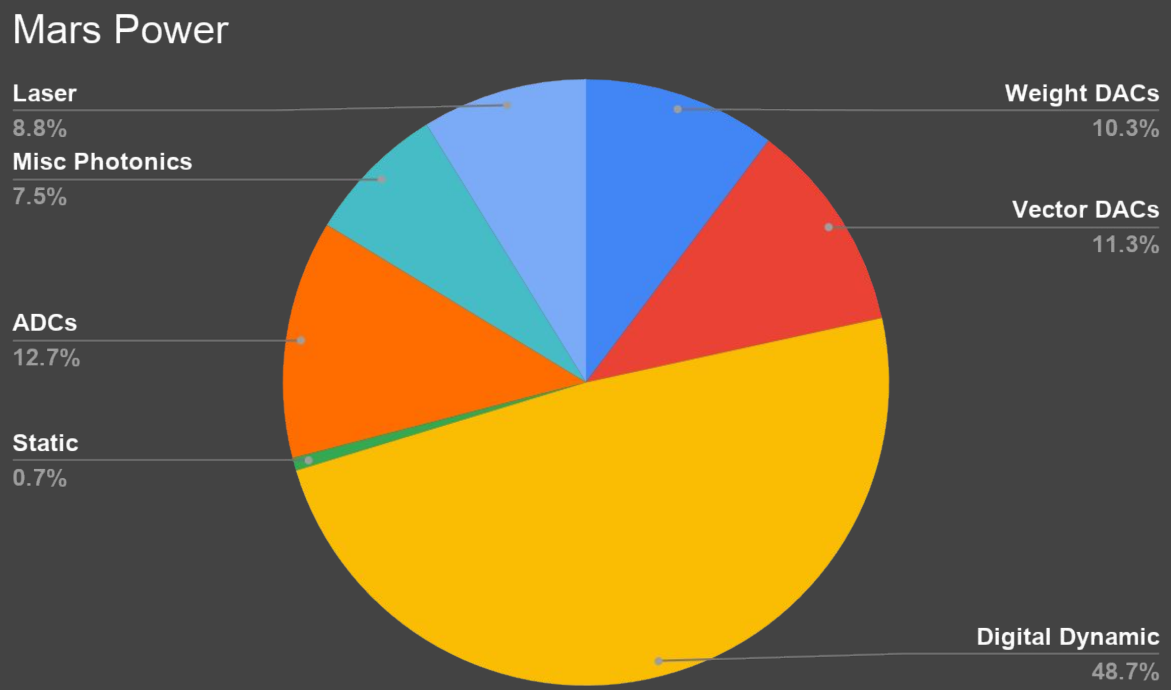
# Weight Updates




- Weights stored near MZI as charge
- Static power near zero
- Computational dynamic power near zero
- Dynamic power for weight update comparable to electronics
- Batching saves SRAM/transport power

**Only pay for data conversion power, not compute**

# Power and Performance




**Under 3W TDP**

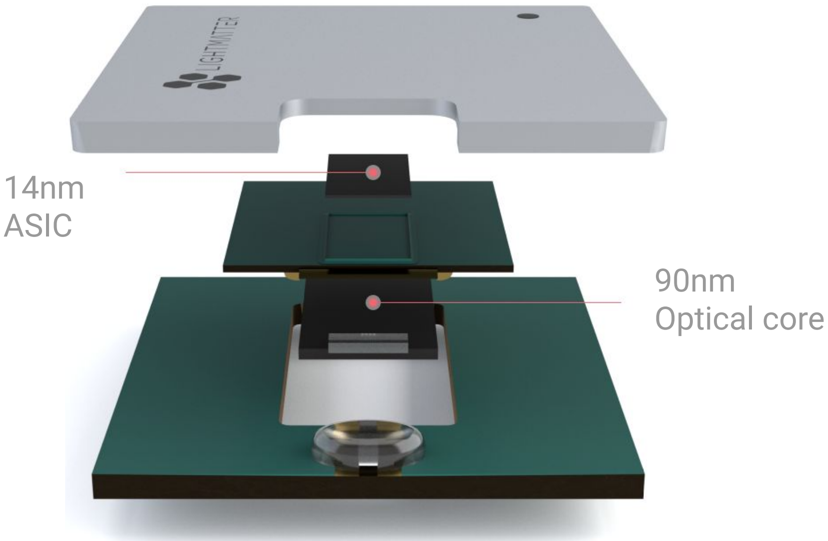
**Most power is data movement**

**ResNet-50 @ 99% accuracy**

*(ImageNet, compared to FP32)*

# 3D Integration

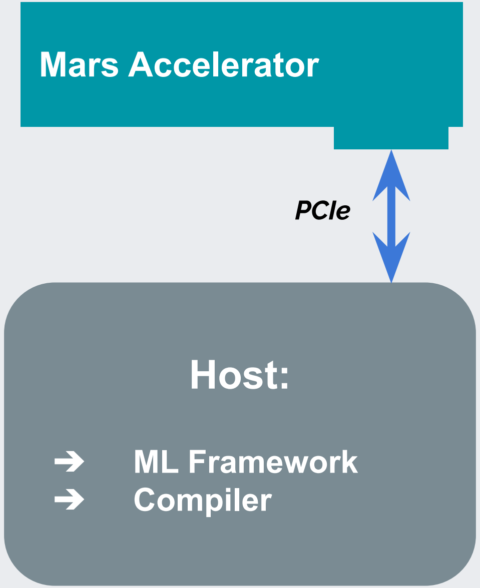
- ❑ **Low power compute allows stacking**
- ❑ **Stacking reduces data movement**
- ❑ **SRAM is closer to compute**




# Putting it all together → Software

## Neural network model optimization and deployment

Interface with standard deep learning frameworks, compilers, and model exchange formats.




ML FRAMEWORKS

PyTorch | TensorFlow | ONNX

Provides Services for...

### Model

Simulate the effects of model parameters on accuracy and performance

### Compile & Execute

Map model to hardware
Optimize model performance
Execute generated code

### Debug

Access intermediate state in the model under execution

### Profile

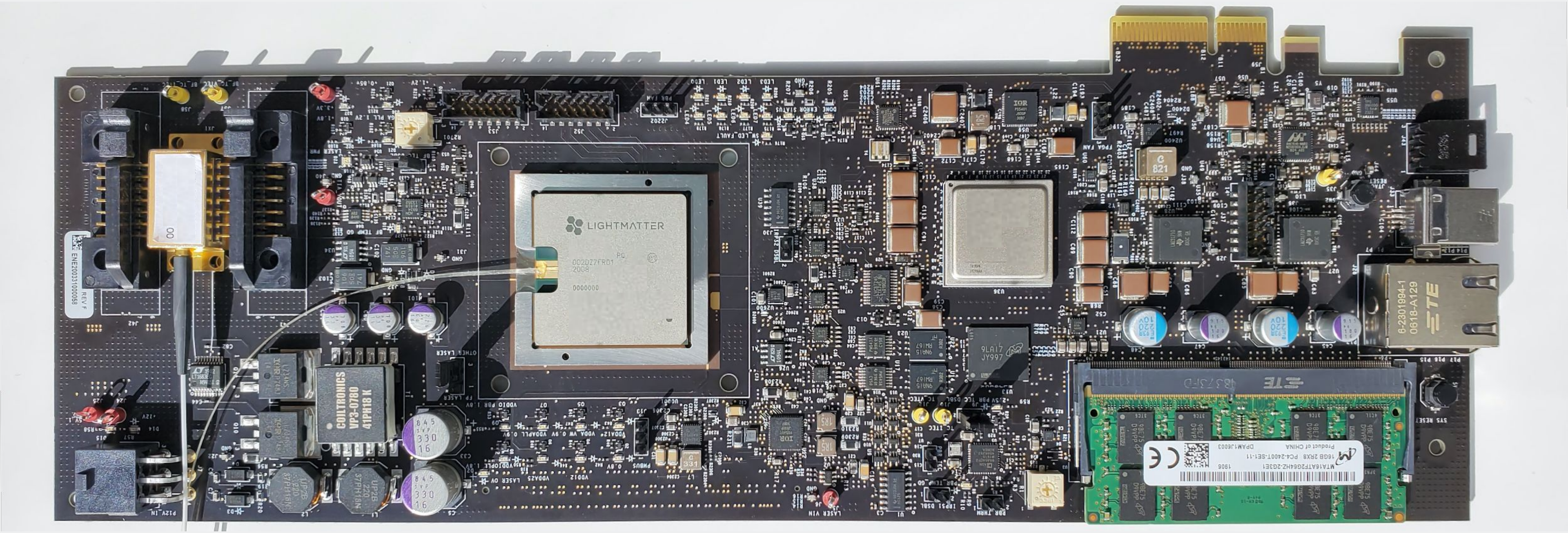
Find and fix performance and resource bottlenecks

# General-purpose AI inference acceleration. Photonics provides core compute.

Mars Accelerator


- Image Recognition
- Sentiment Analysis
- Machine Translation
- Recommendation

# Summary




- Optical computing is here - and focused on AI
- Lightmatter’s Mars chip leverages photonics for compute, electronics for activation and I/O
- 3D stacking brings weights and activations closer to the compute core
- Freedom in power budget allows larger devices and more SRAM