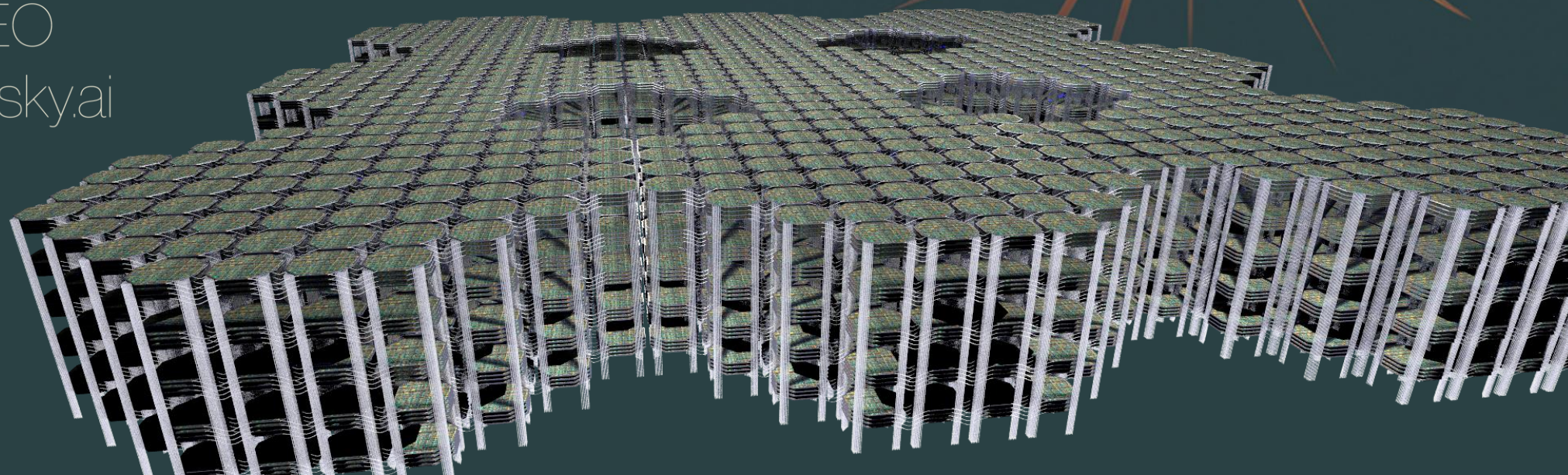
Great Sky

Rebuilding AI from first principles

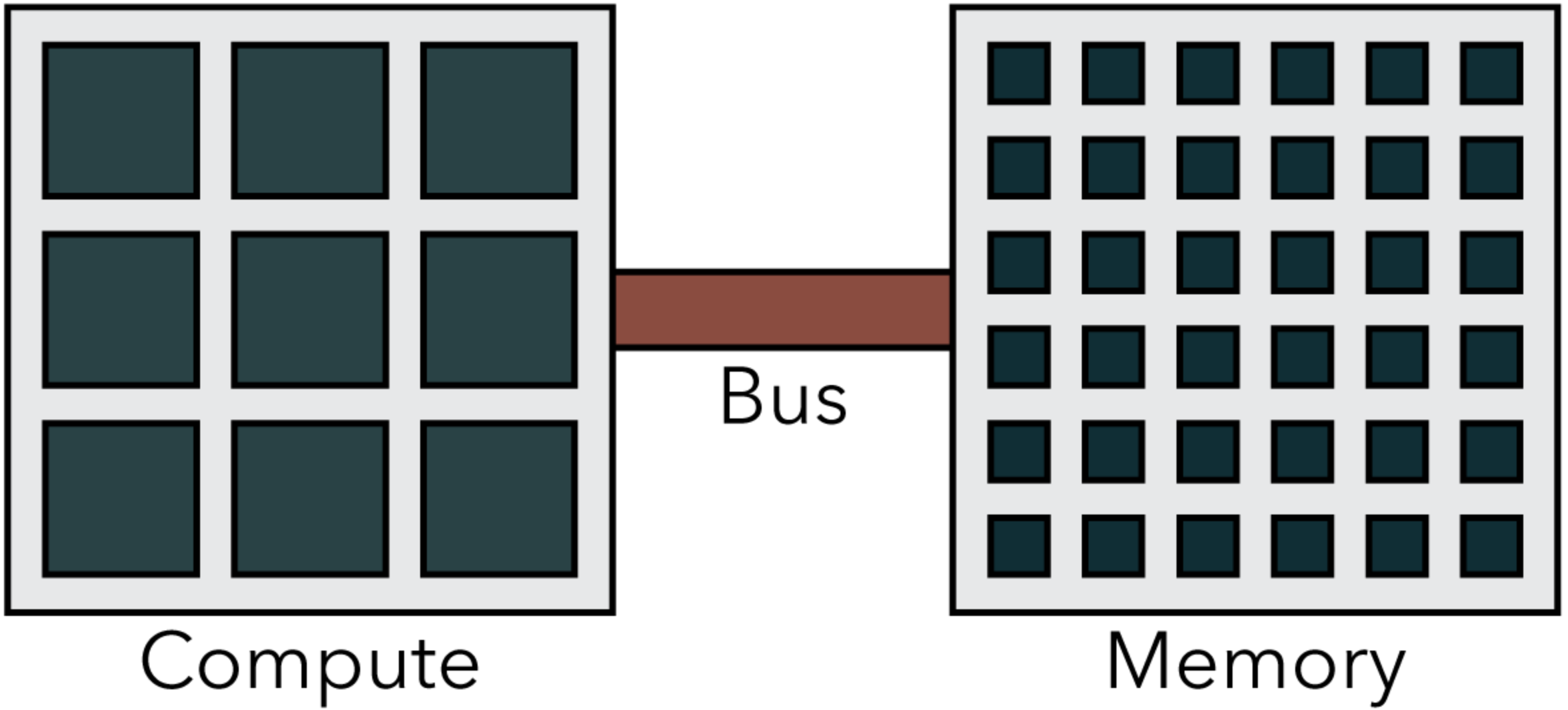
Jeff Shainline
Founder and CEO
shainline@greatsky.ai



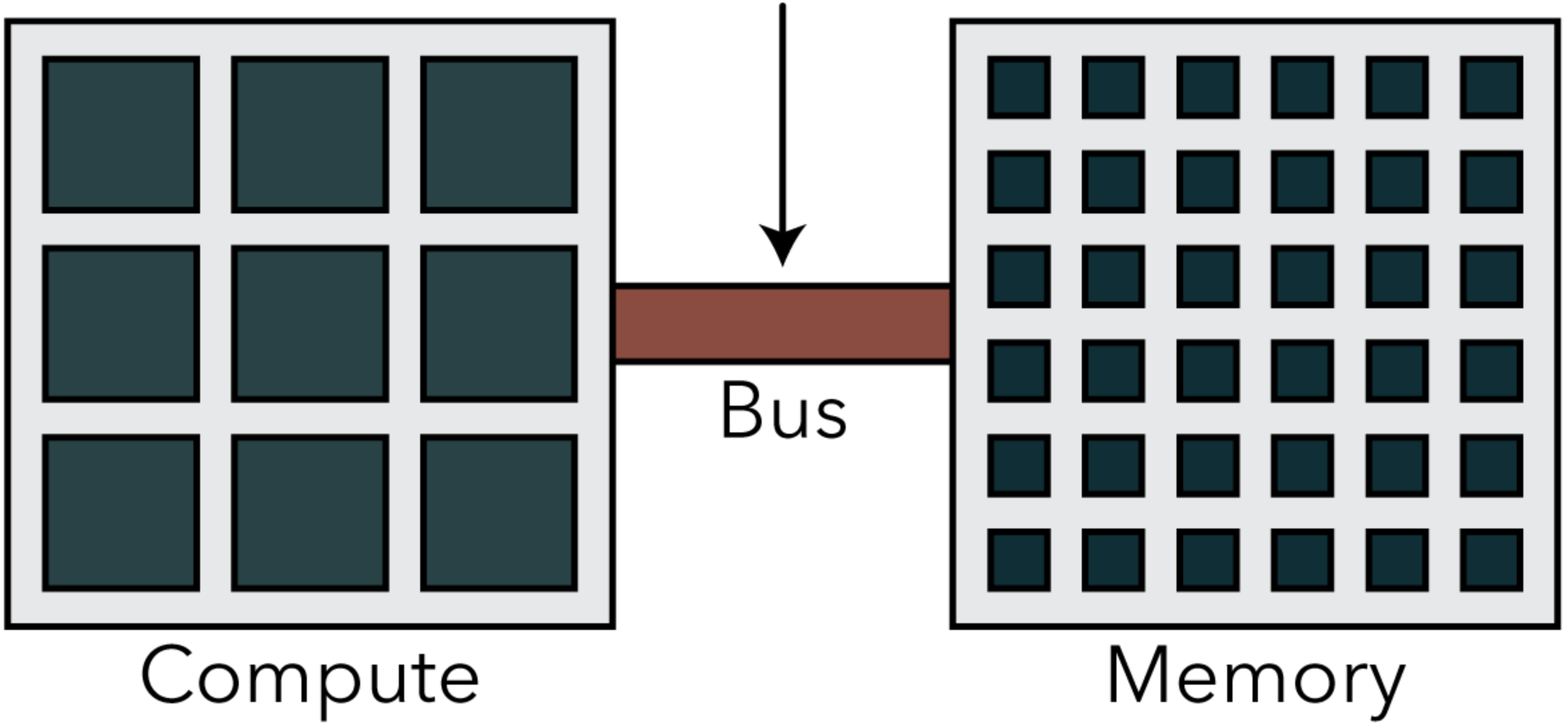
Today's AI is only scratching the surface.

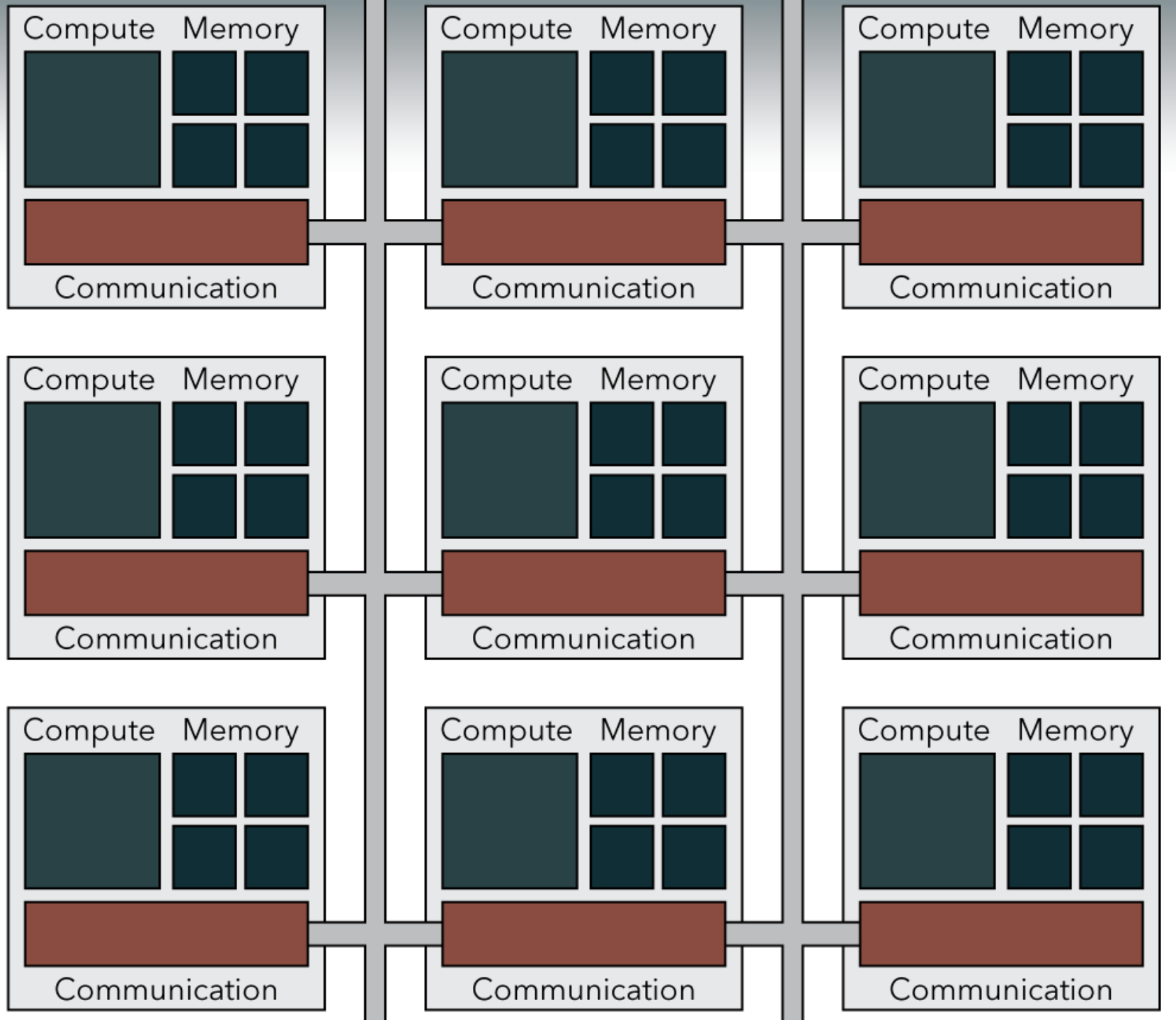
Neuroscience provides a path, but hardware is unfit to implement key neural principles.

We address the root cause of this shortcoming.

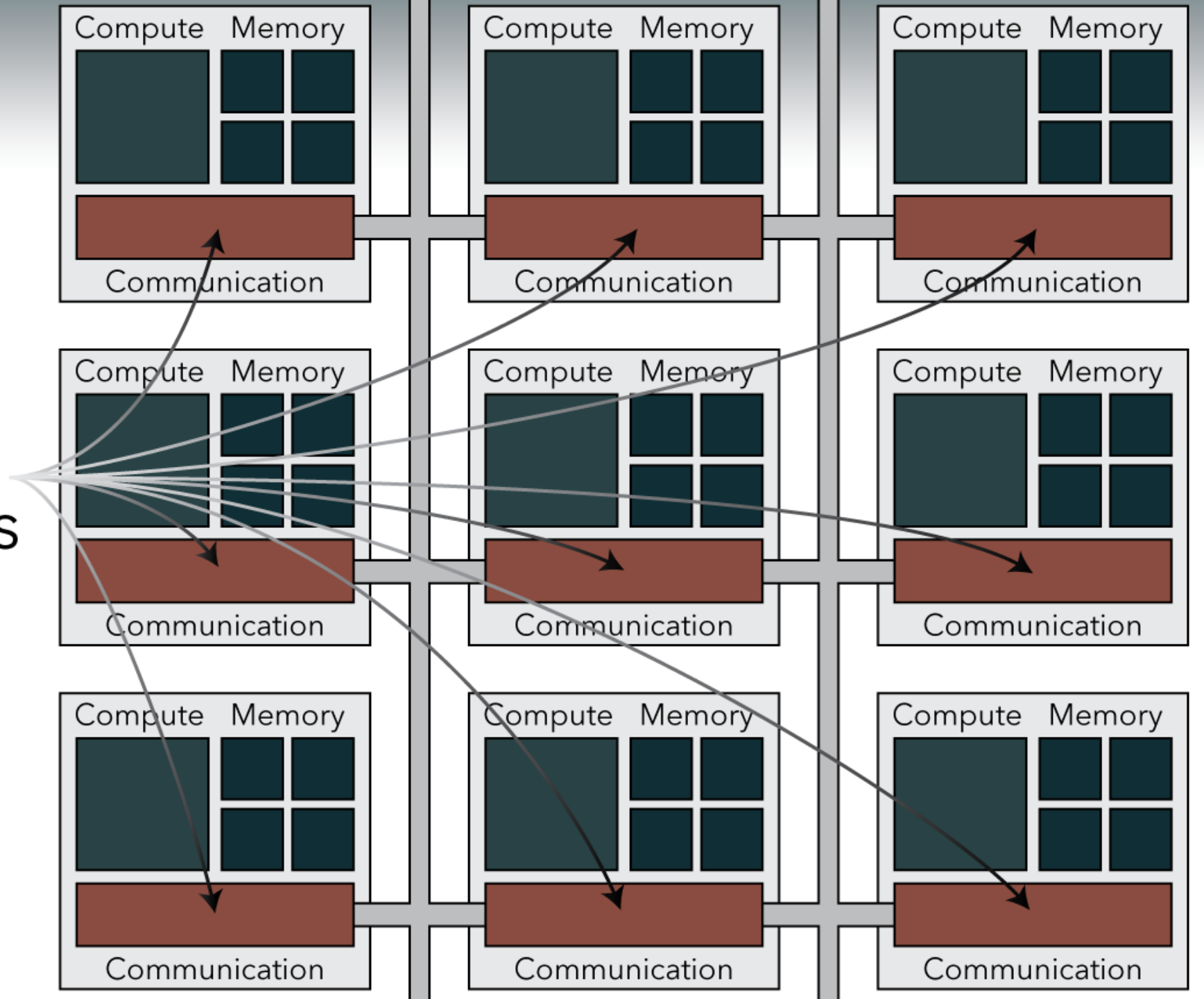

Bus
Compute
Memory

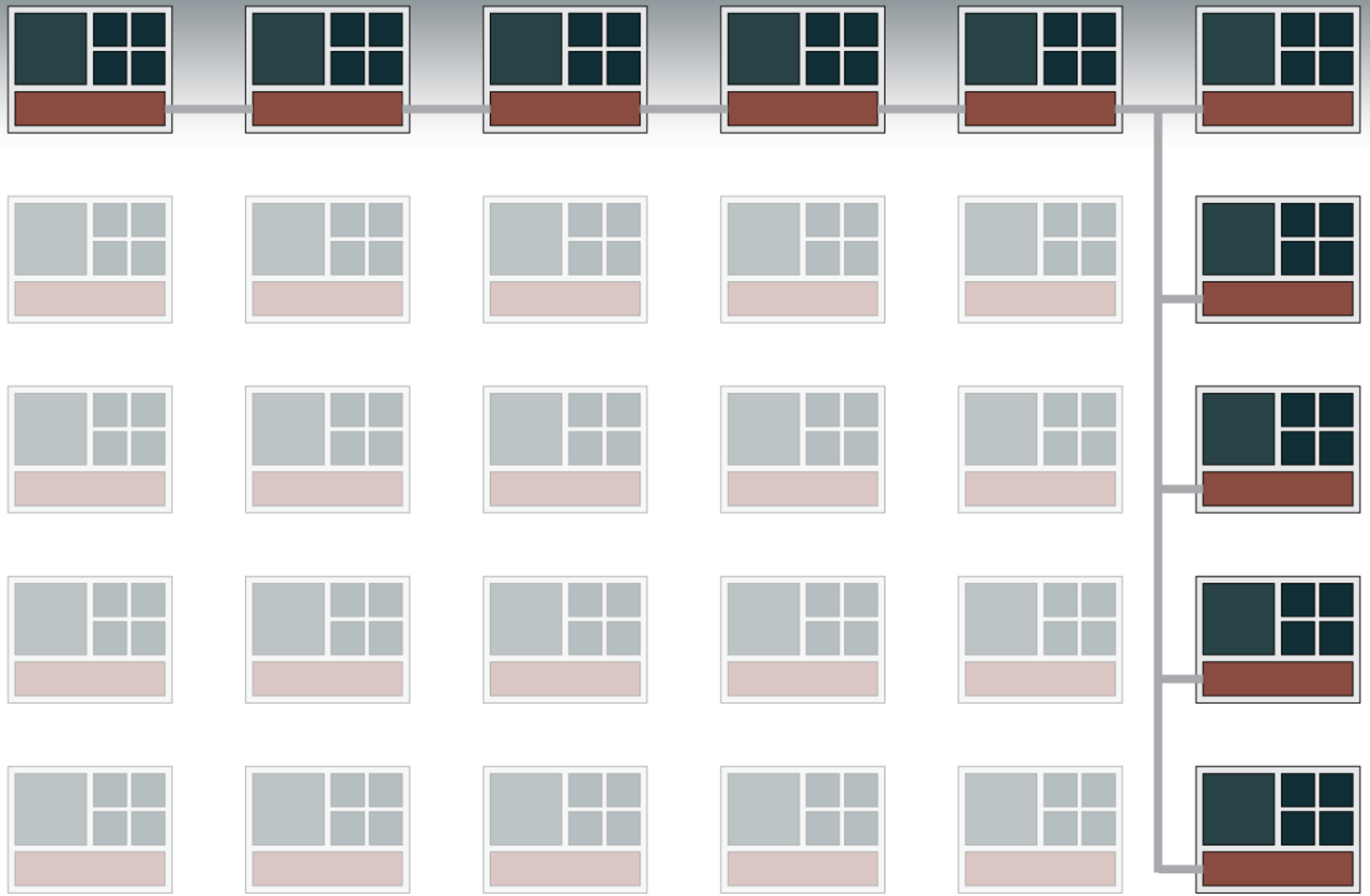
Memory Wall / Bottleneck




Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication
Compute
Memory
Communication

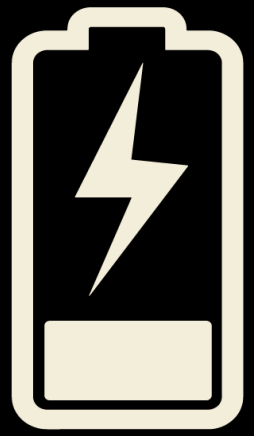
Many Mini Bottlenecks



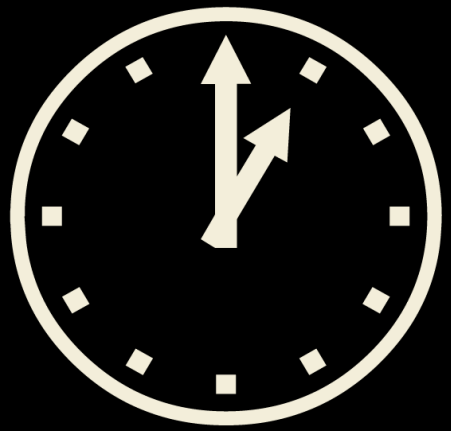


By addressing the deeper problem, we eliminate the symptoms

Energy

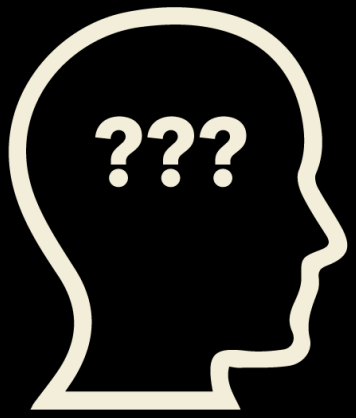


Slow

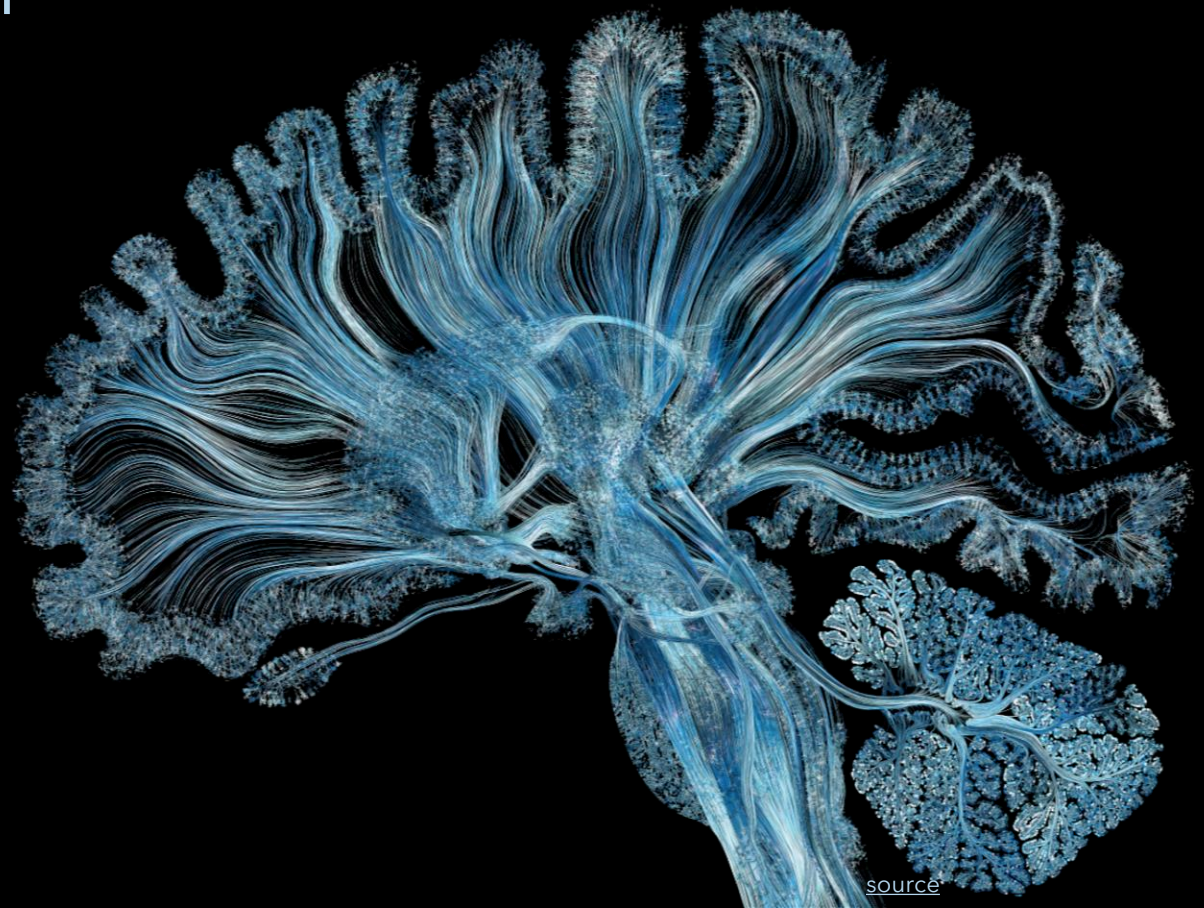


Data

Intelligence

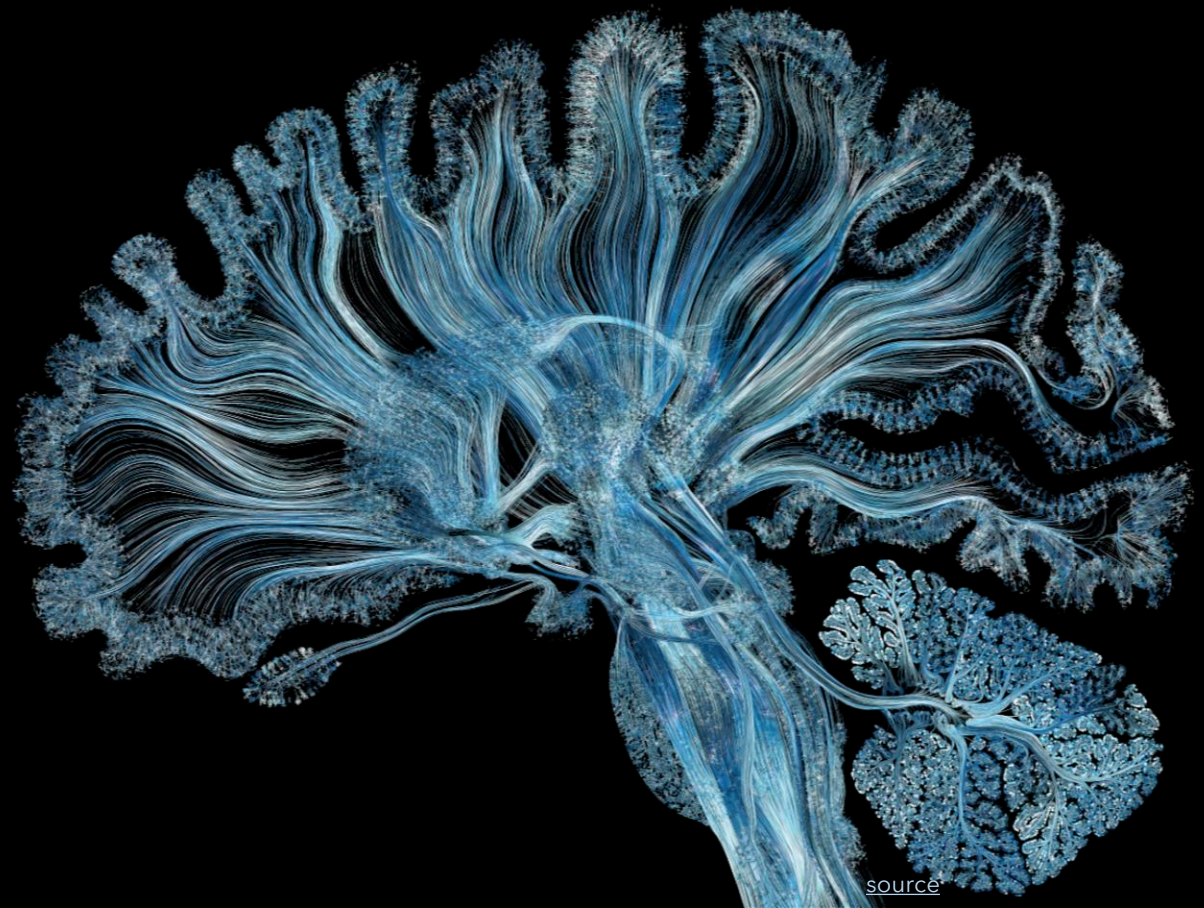


Intelligence requires efficient information integration across vast networks



source

Efficient information integration requires radically new hardware

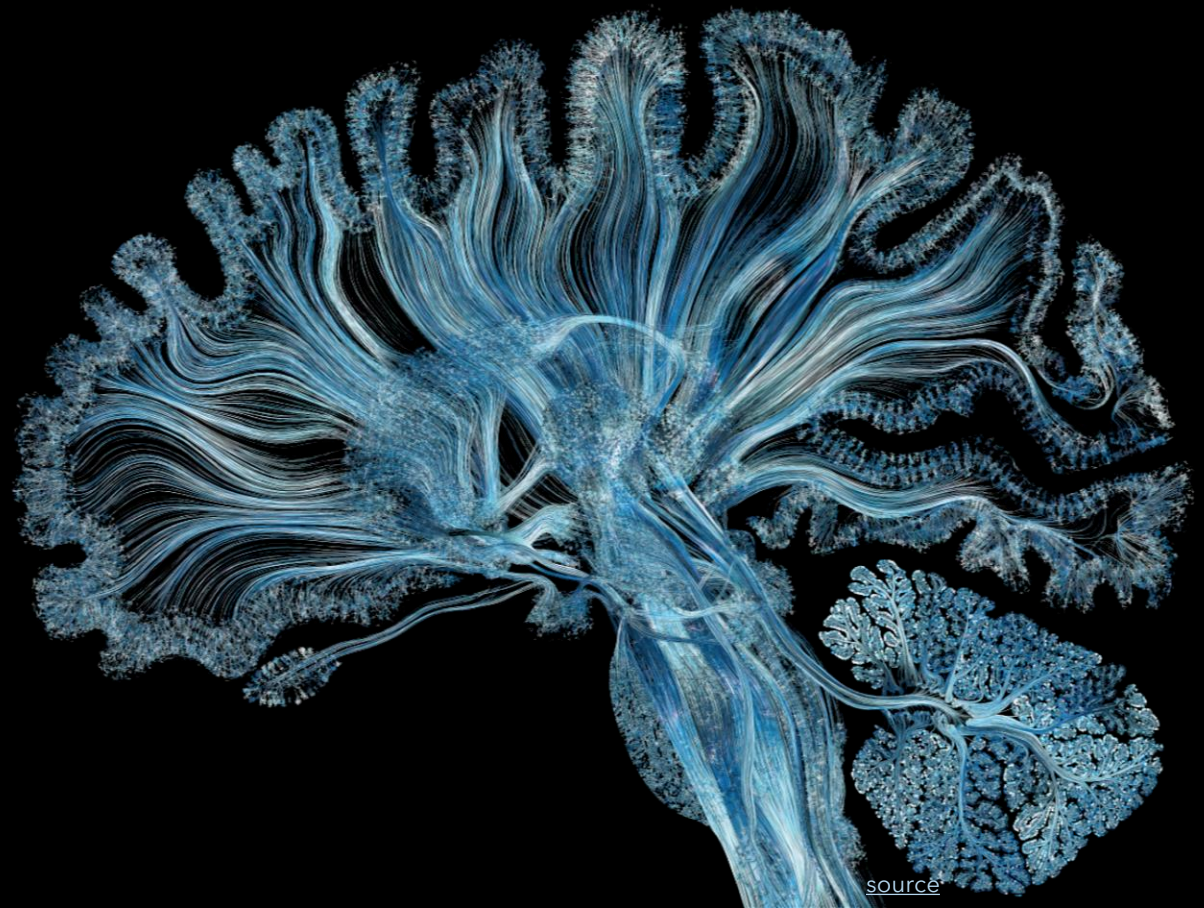


source

Analog processing

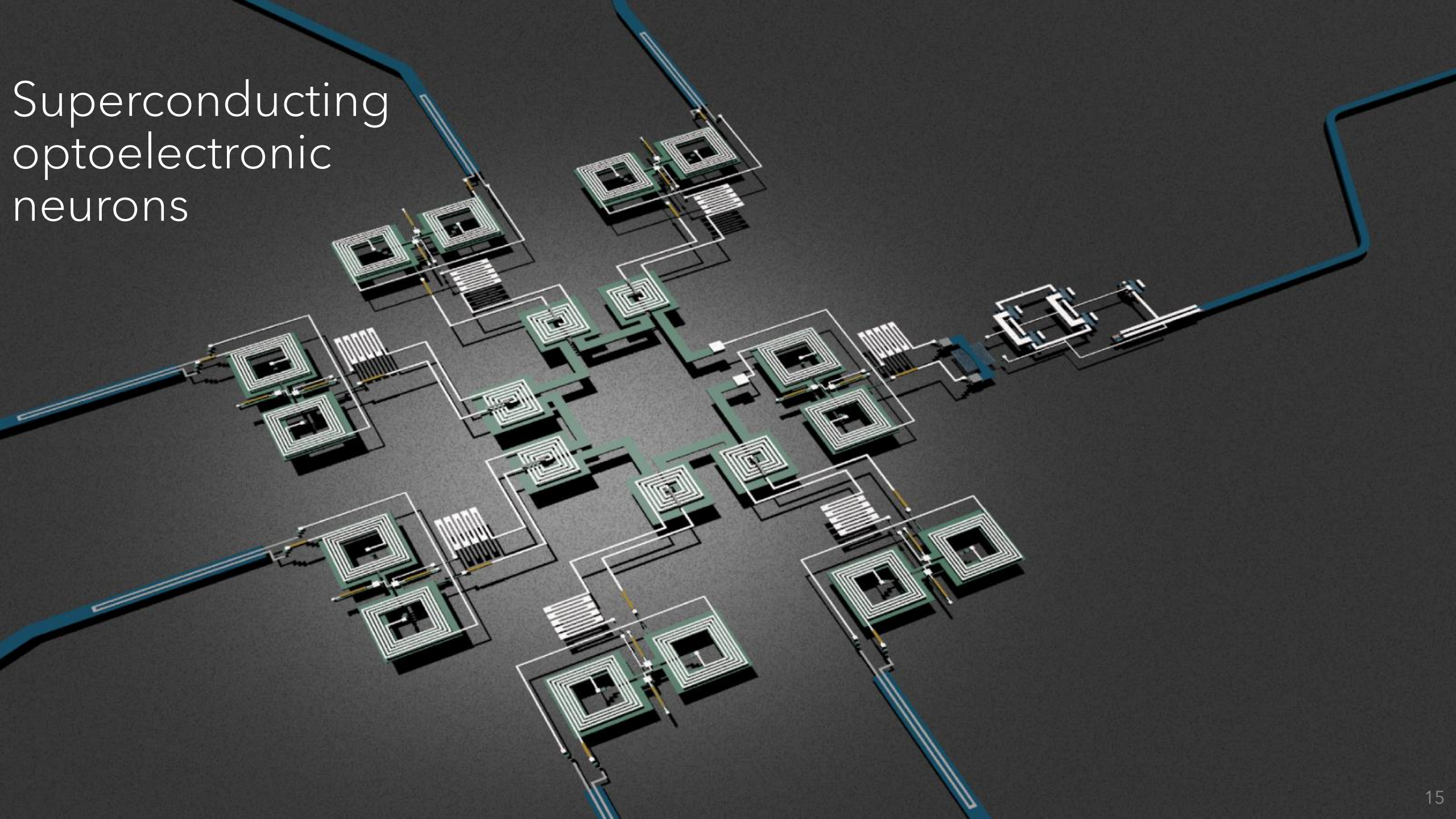
Distributed memory

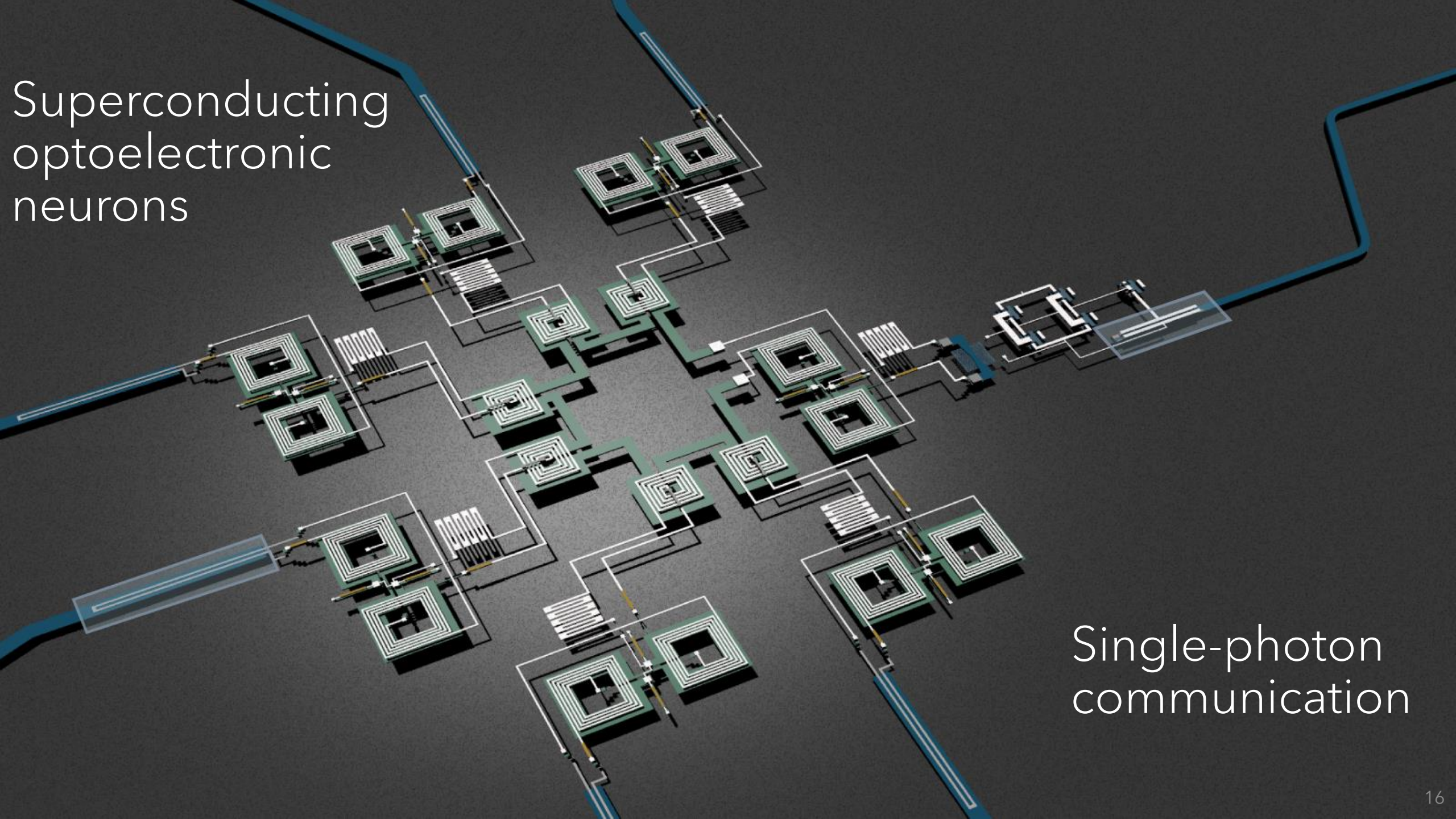
Direct connections

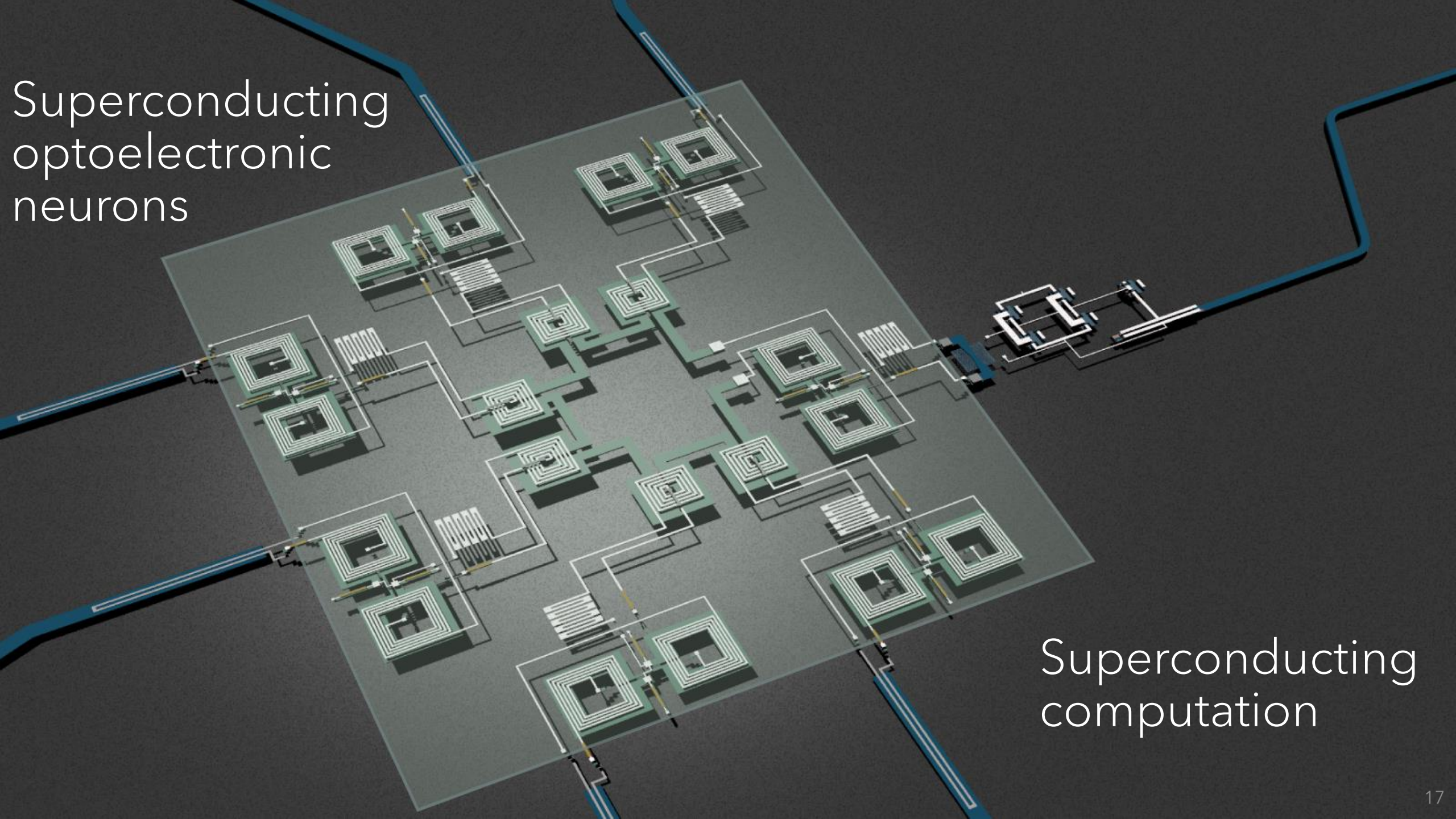


source

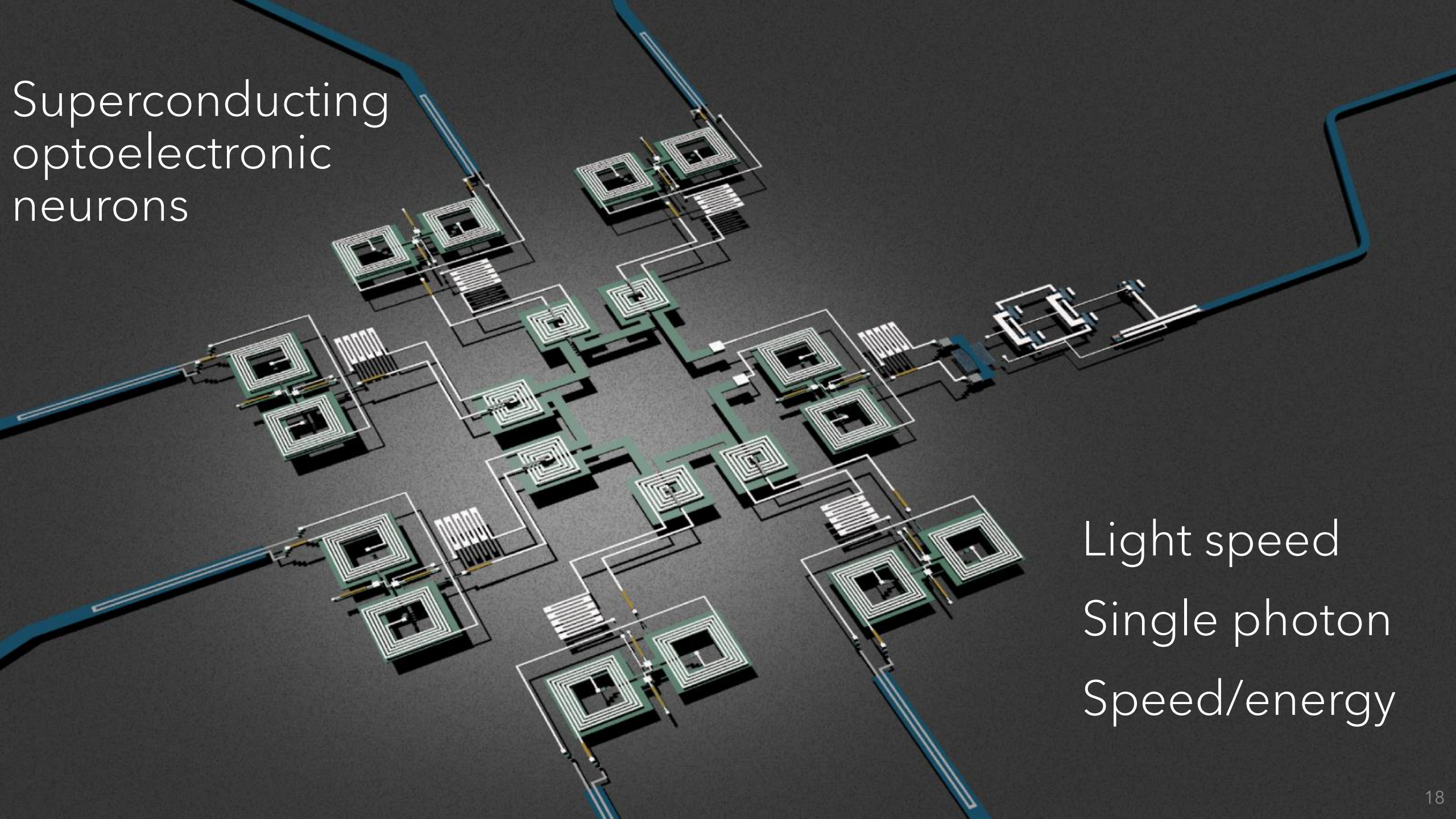
Superconducting optoelectronic neurons




Superconducting optoelectronic neurons
Single-photon communication


Superconducting optoelectronic neurons
Superconducting computation

Superconducting optoelectronic neurons



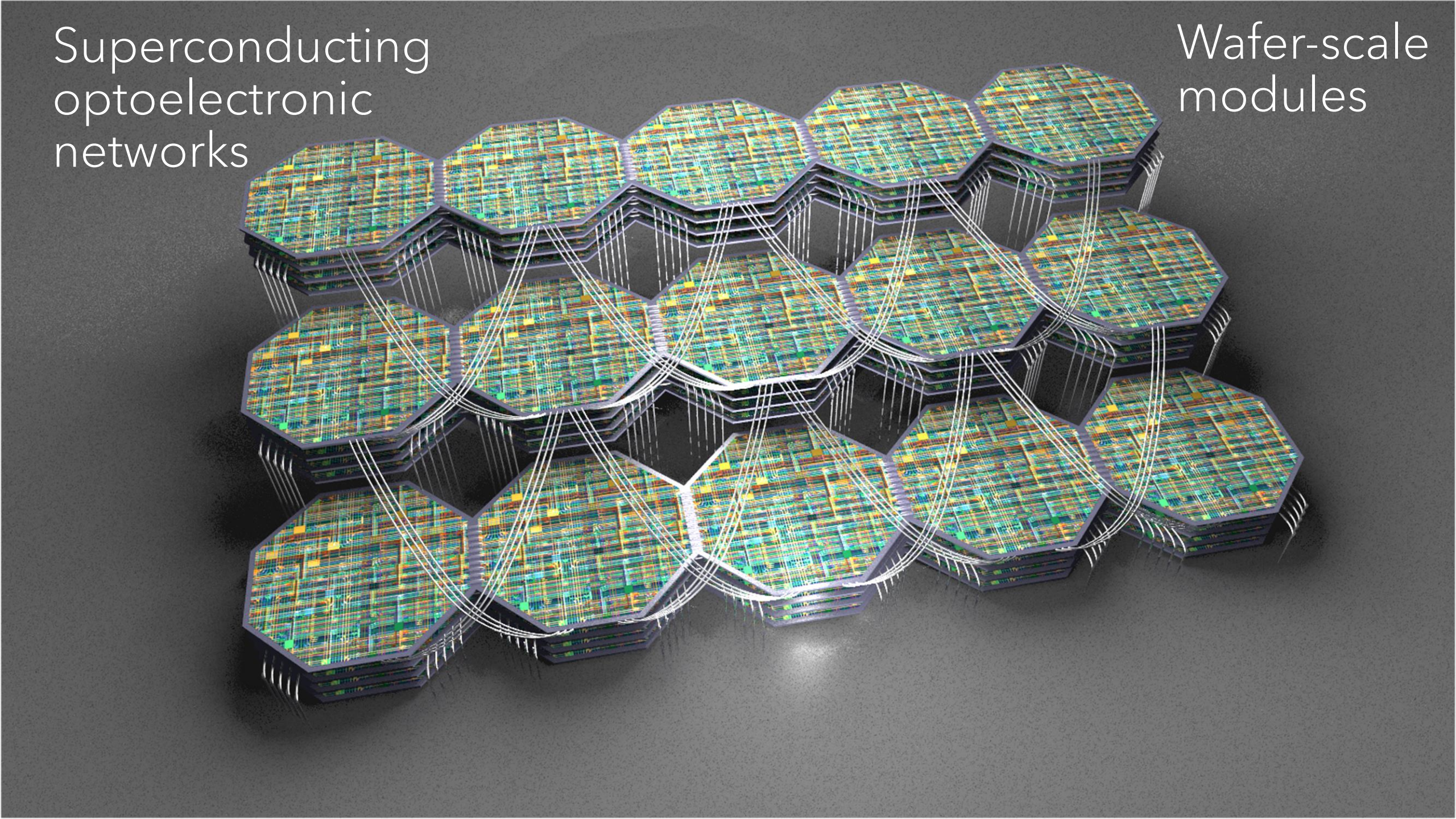
Light speed

Single photon

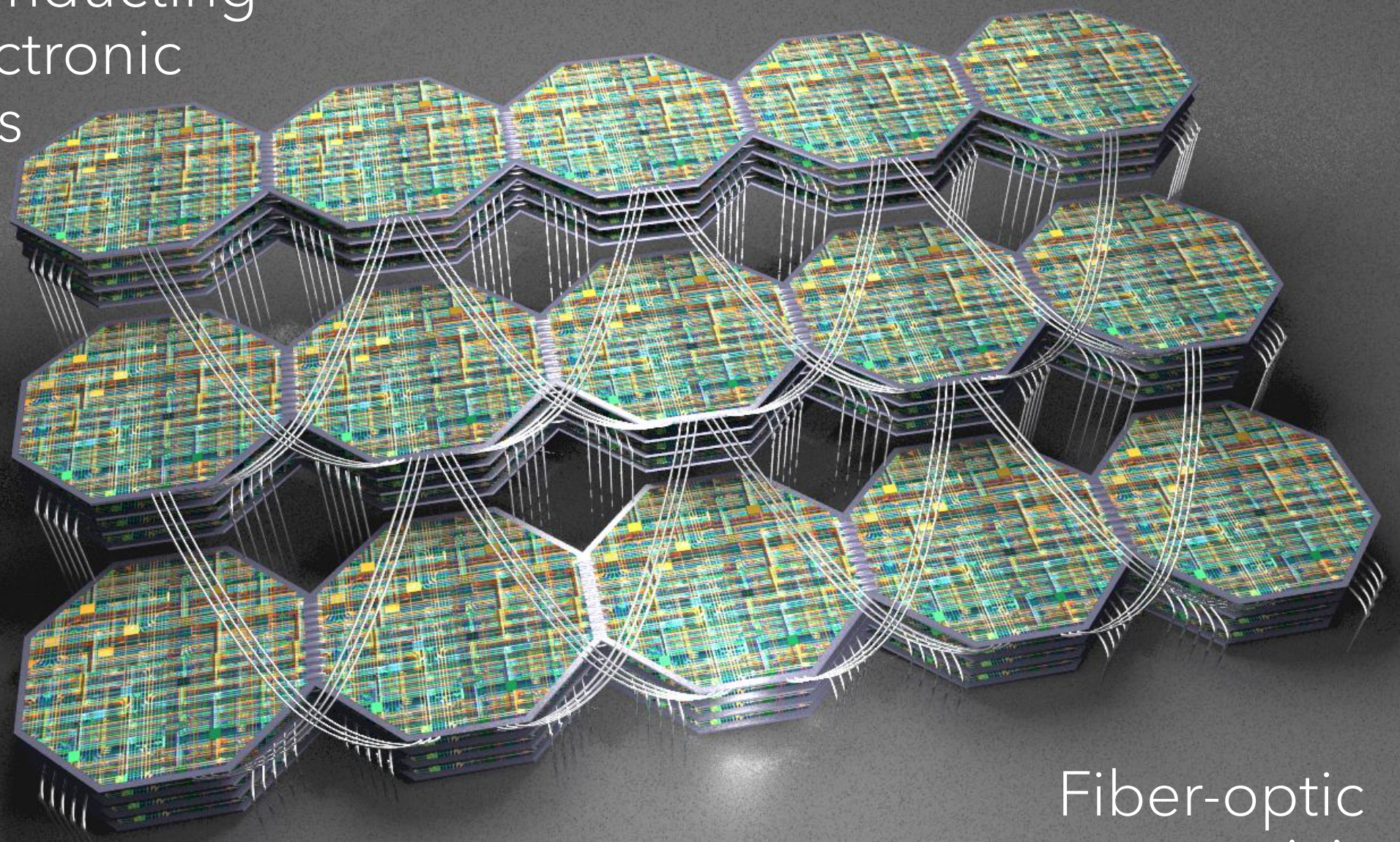
Speed/energy

Superconducting optoelectronic networks

Wafer-scale modules

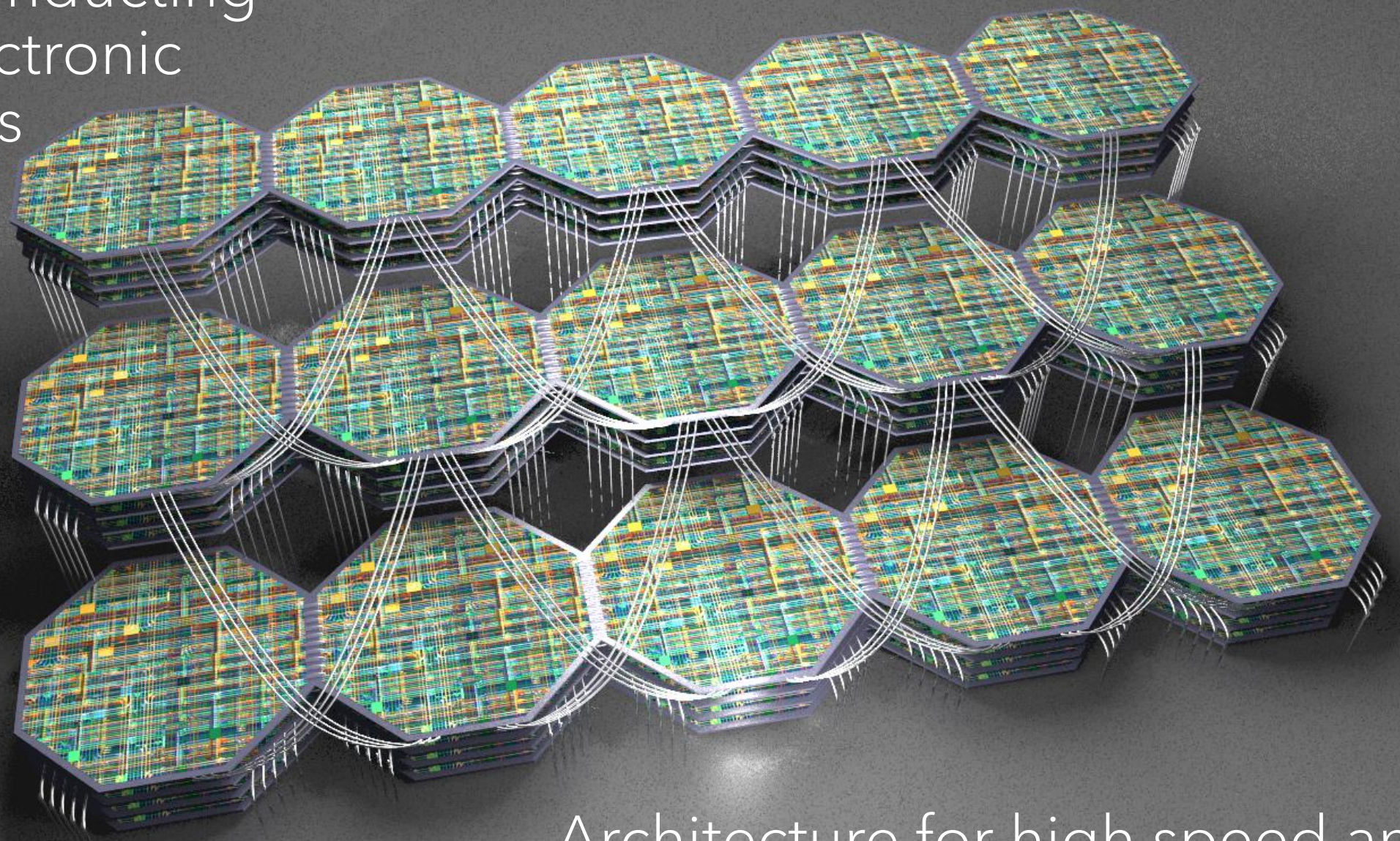


Superconducting optoelectronic networks



Fiber-optic connectivity

Superconducting optoelectronic networks



Architecture for high speed and low power at very large scale

Compared to GPUs:

ImageNet classification:

100x less energy

200x faster

vs *GPUs*

H200

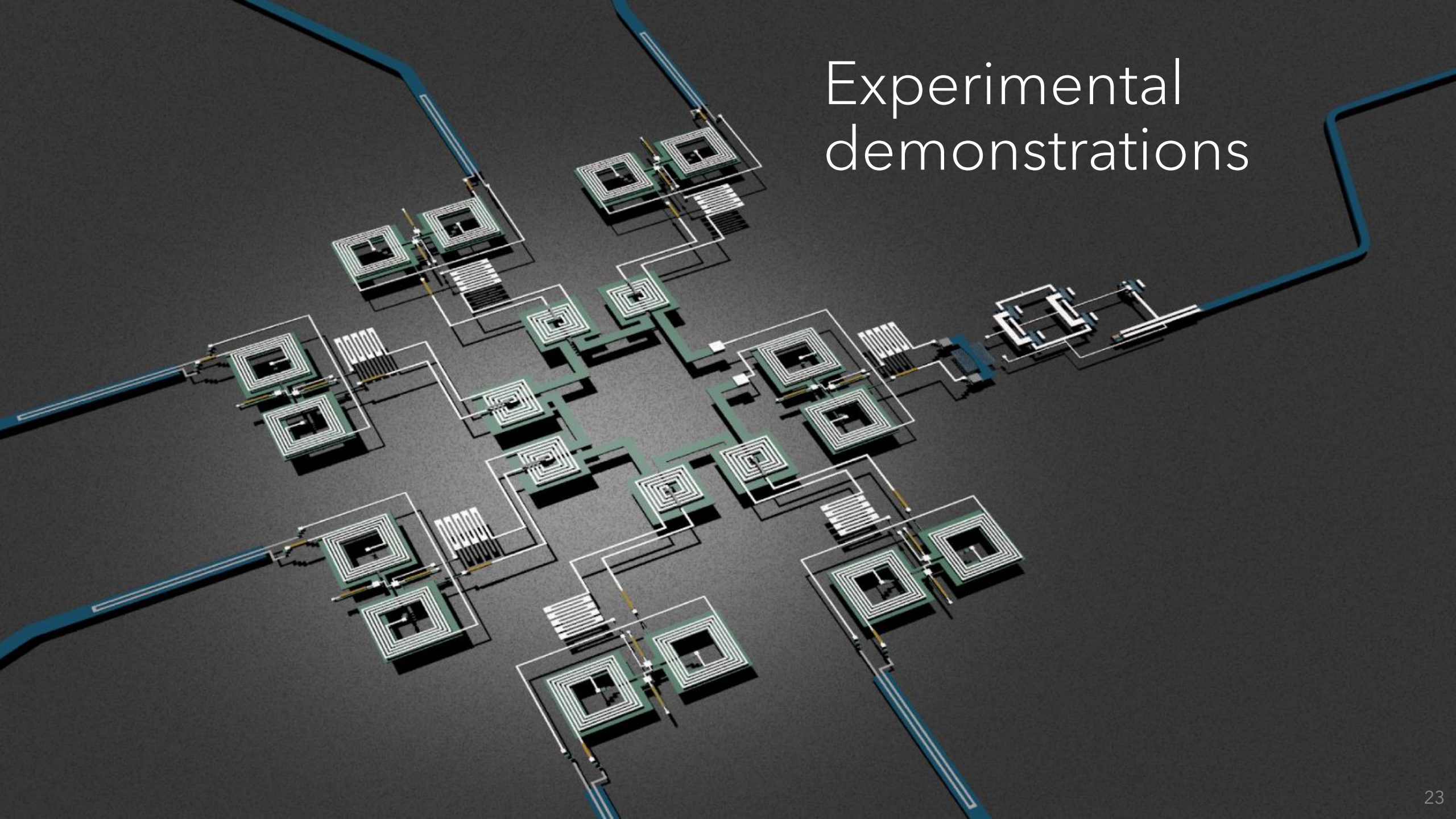
Per Op.

Operation	SOENs	GPU	Improvement
8-bit MAC / Int-8	7.0×10^{-15} joules	5.0×10^{-13} joules	71x
8-bit communication	1.5×10^{-14} joules	6.4×10^{-11} joules	4300 x

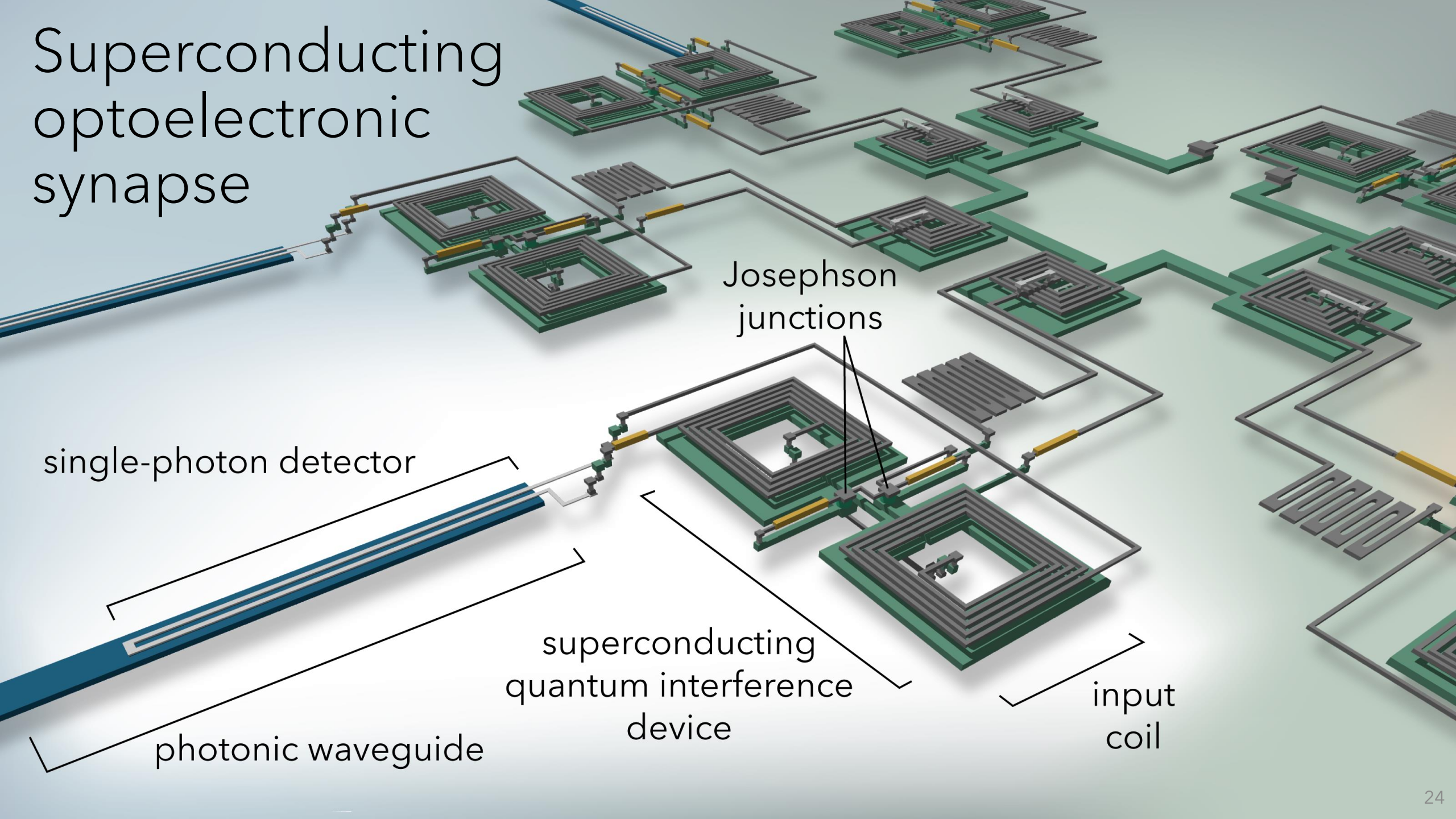
Time-to-first-spike algorithm

Comm = DRAM access
Batch size of 128

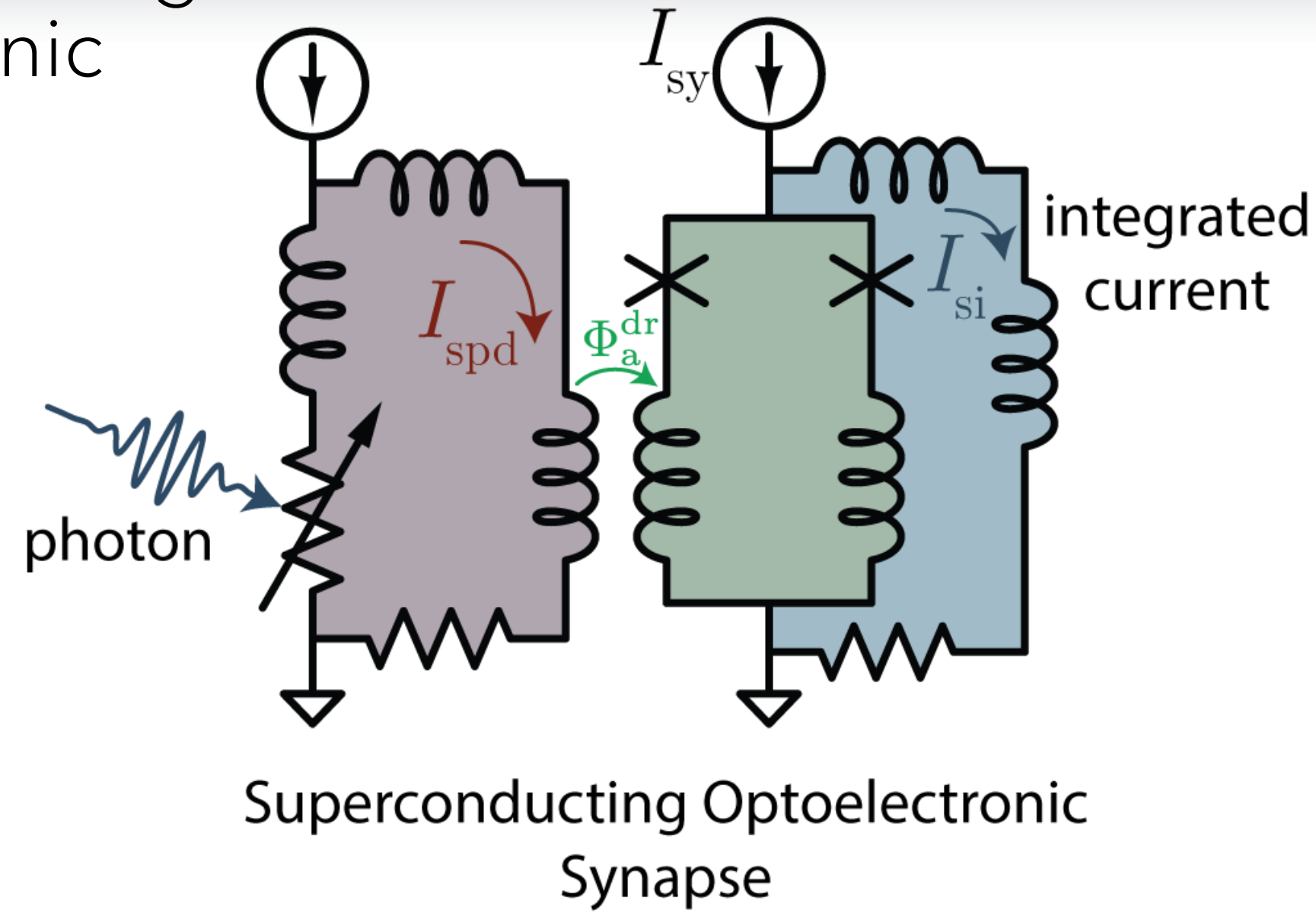
Experimental demonstrations



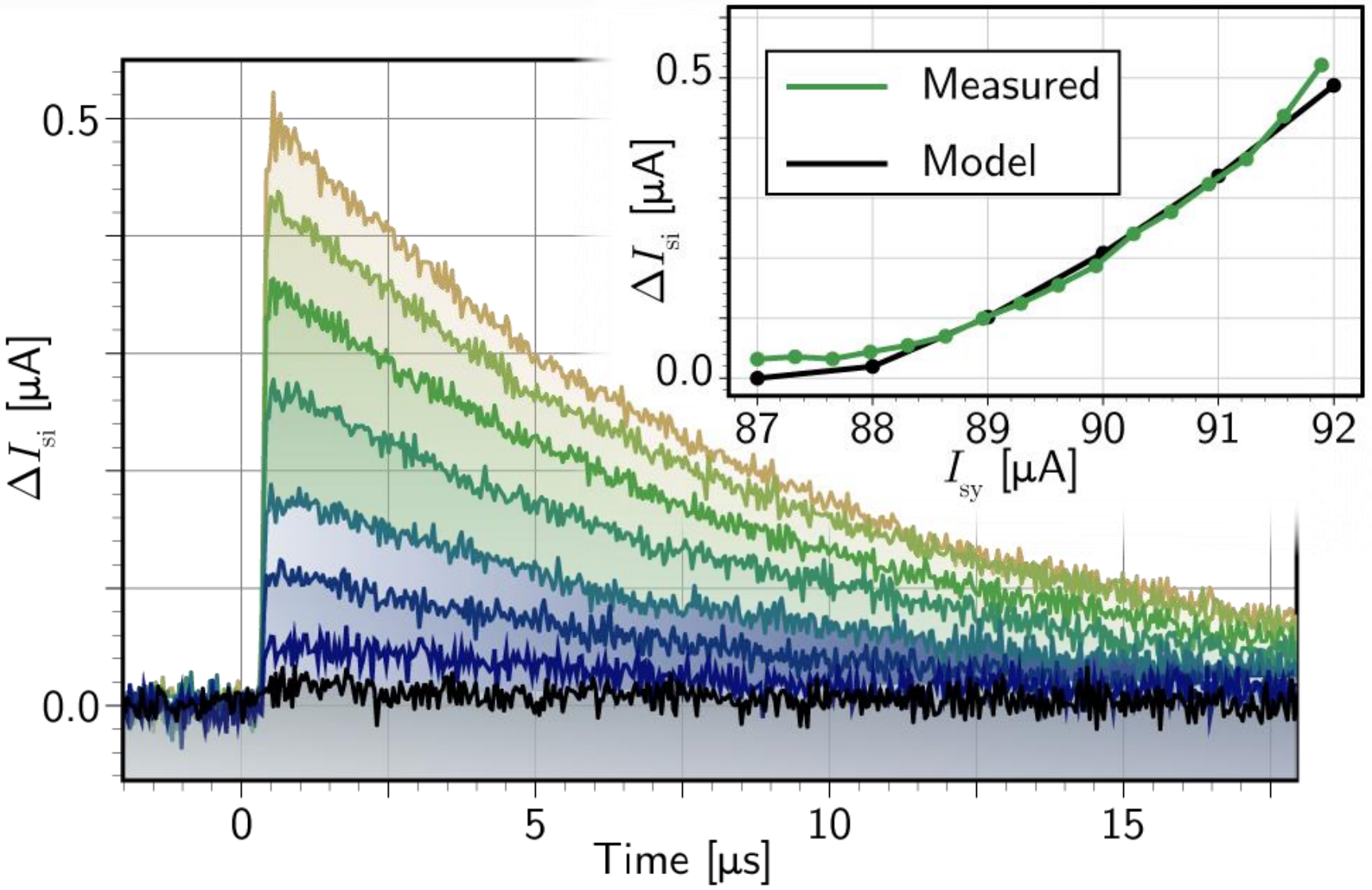
Superconducting optoelectronic synapse

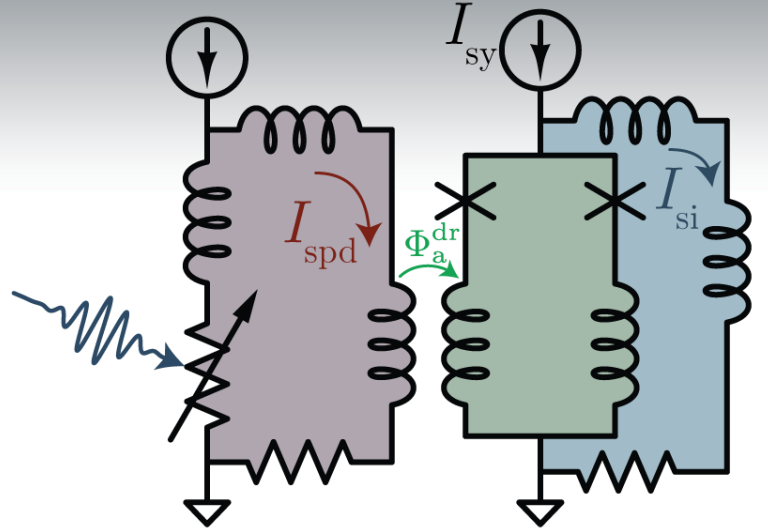


Superconducting optoelectronic synapse



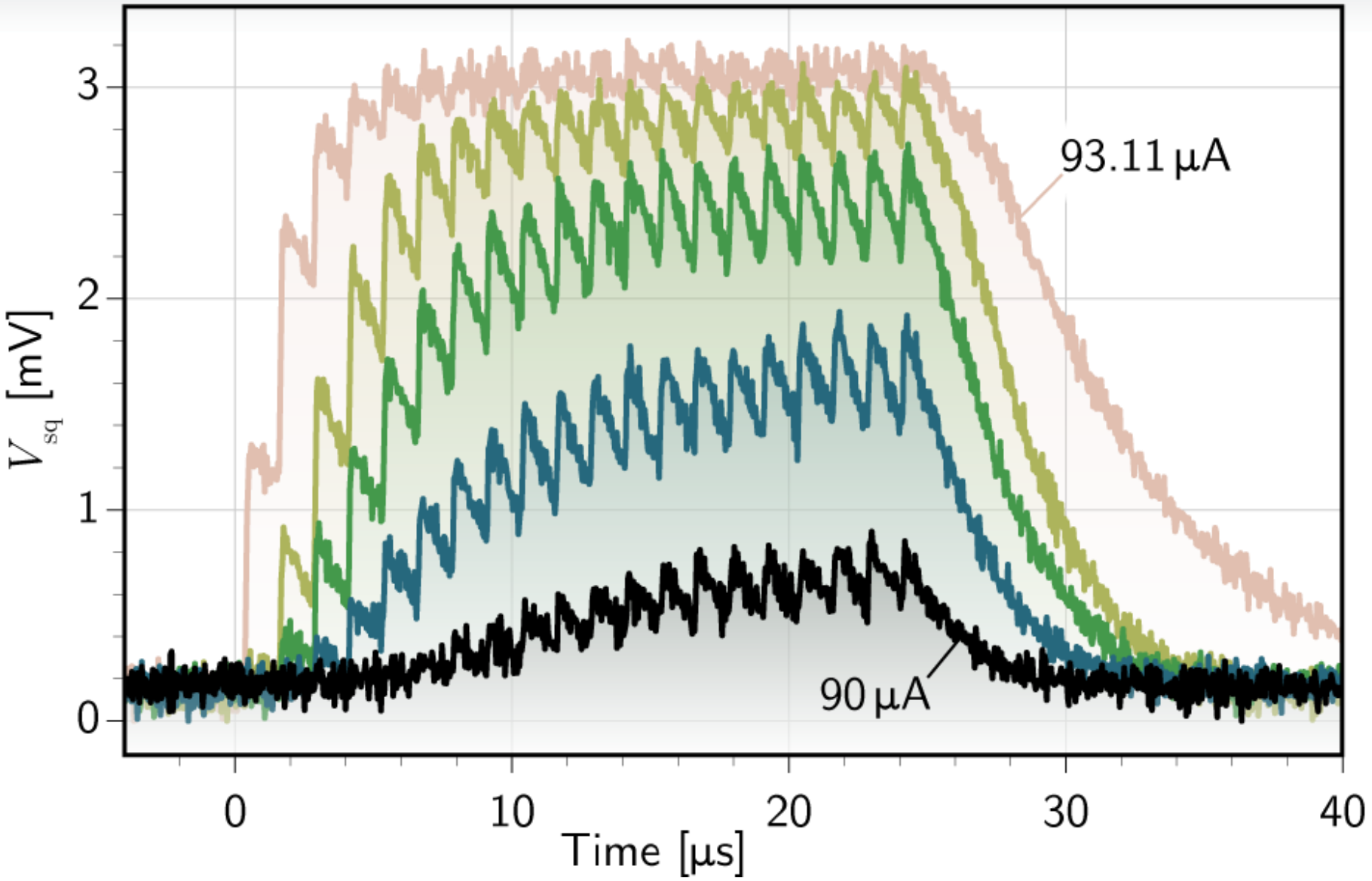
Synaptic weighting





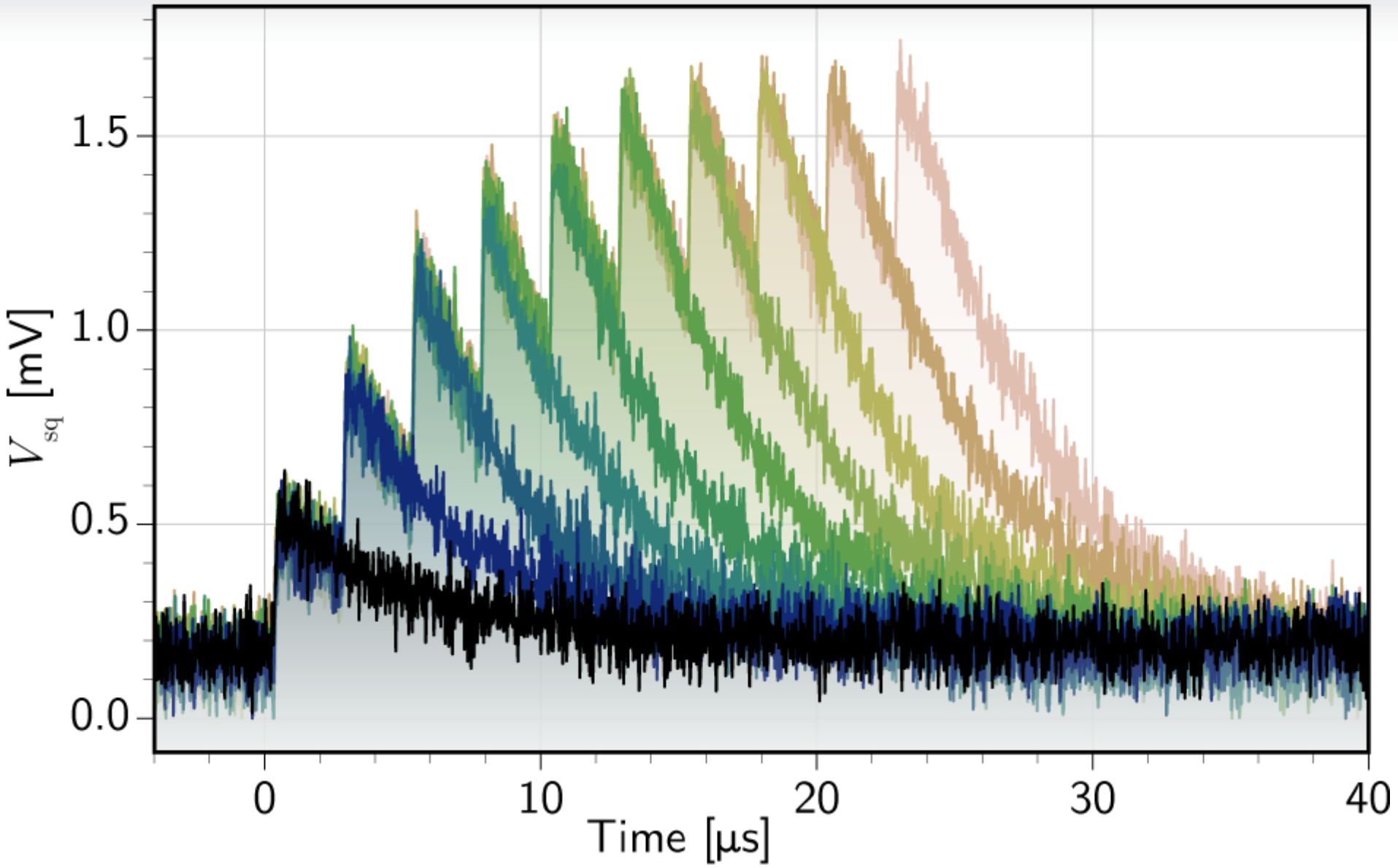
Khan, Primavera et al., Nature Electronics (2022)

Weighting and integration



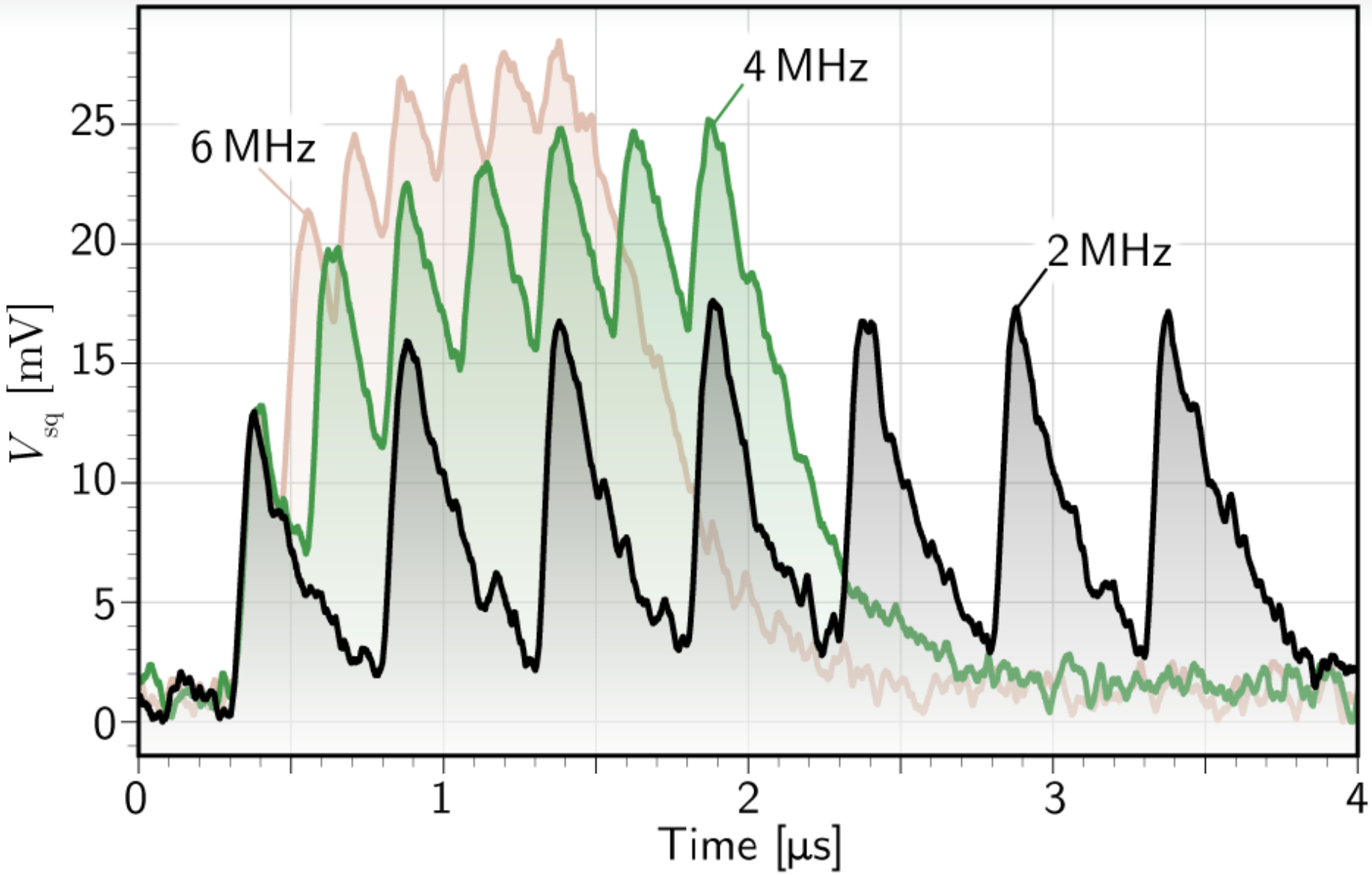
Khan, Primavera et al., Nature Electronics (2022)

Burst coding



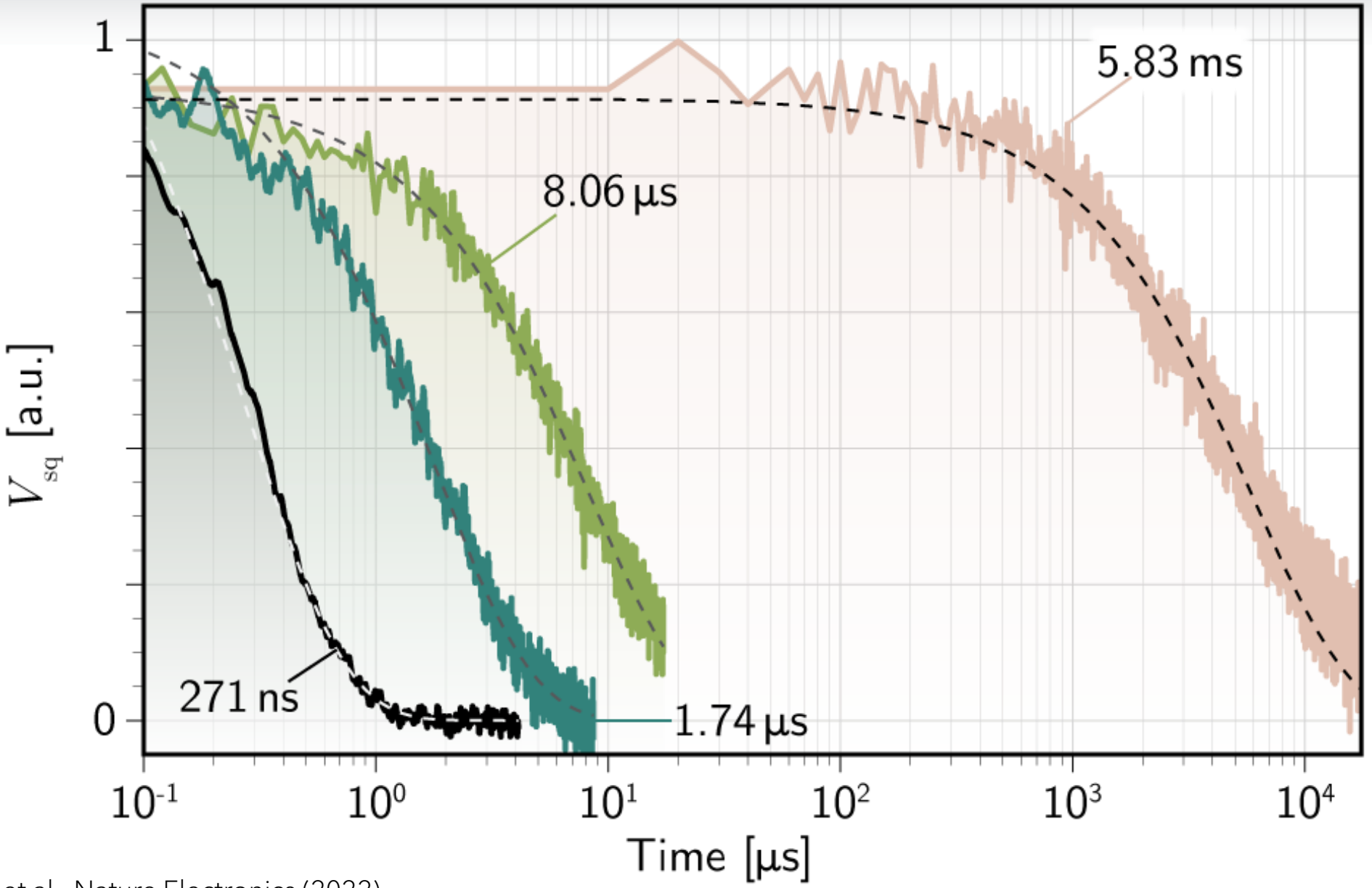
Khan, Primavera et al., Nature Electronics (2022)

Frequency coding



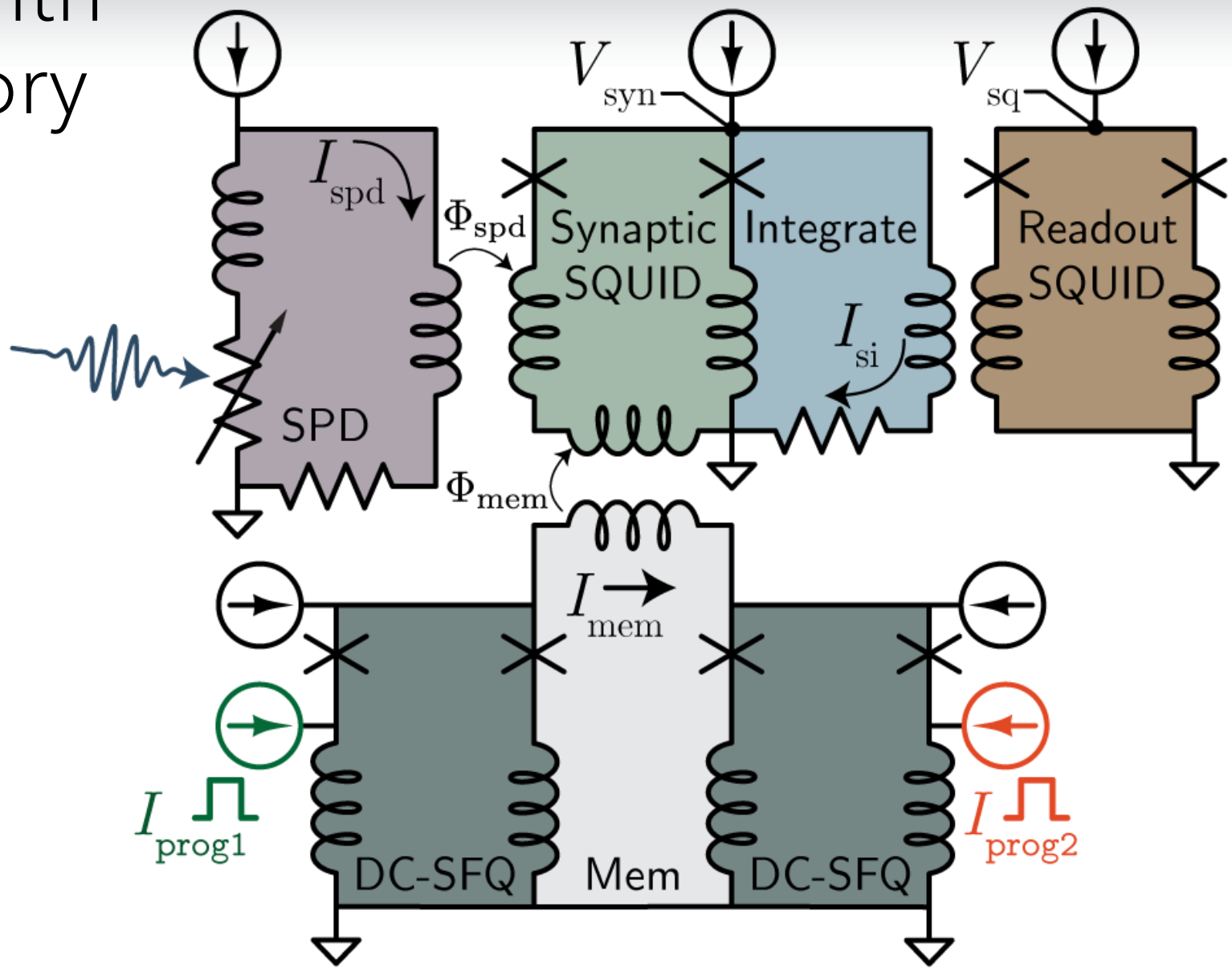
Khan, Primavera et al., Nature Electronics (2022)

Broad temporal scales

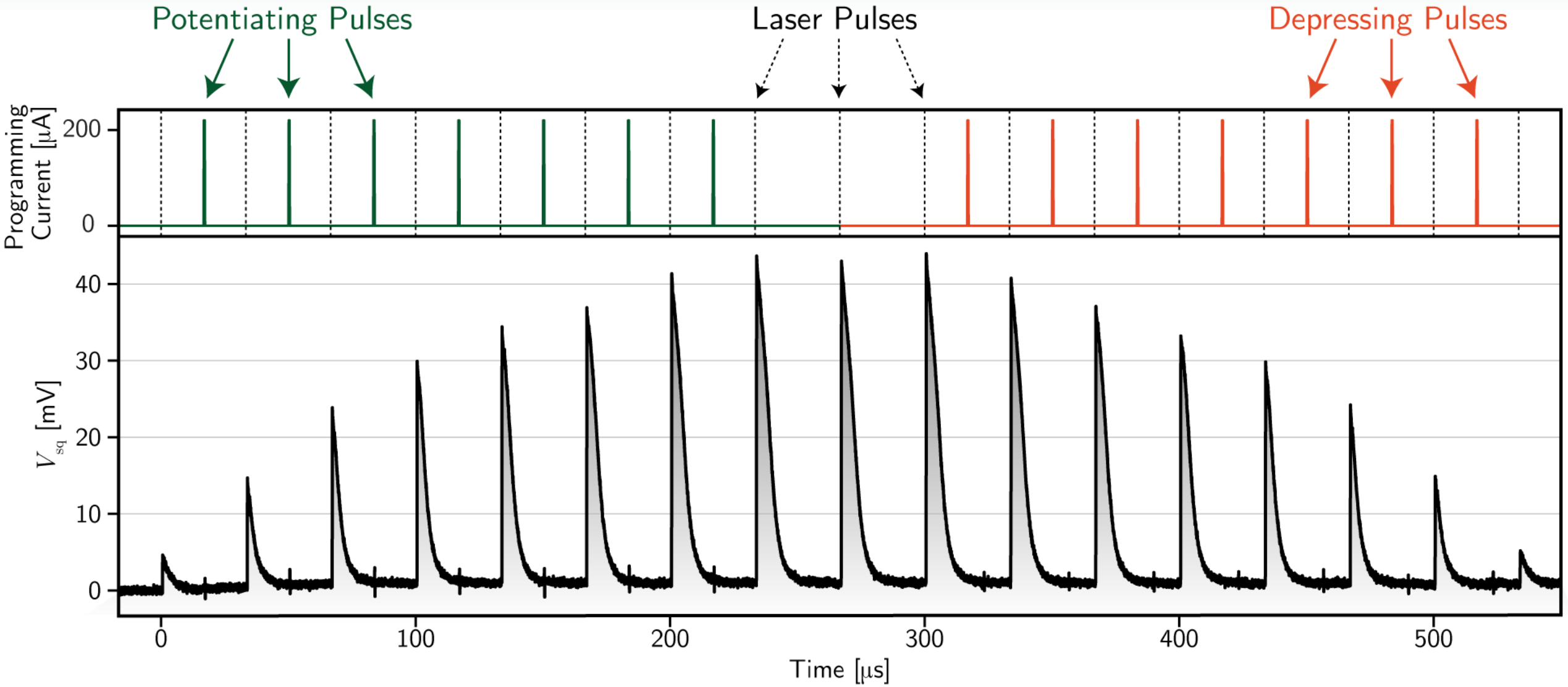


Khan, Primavera et al., Nature Electronics (2022)

Synapses with local memory

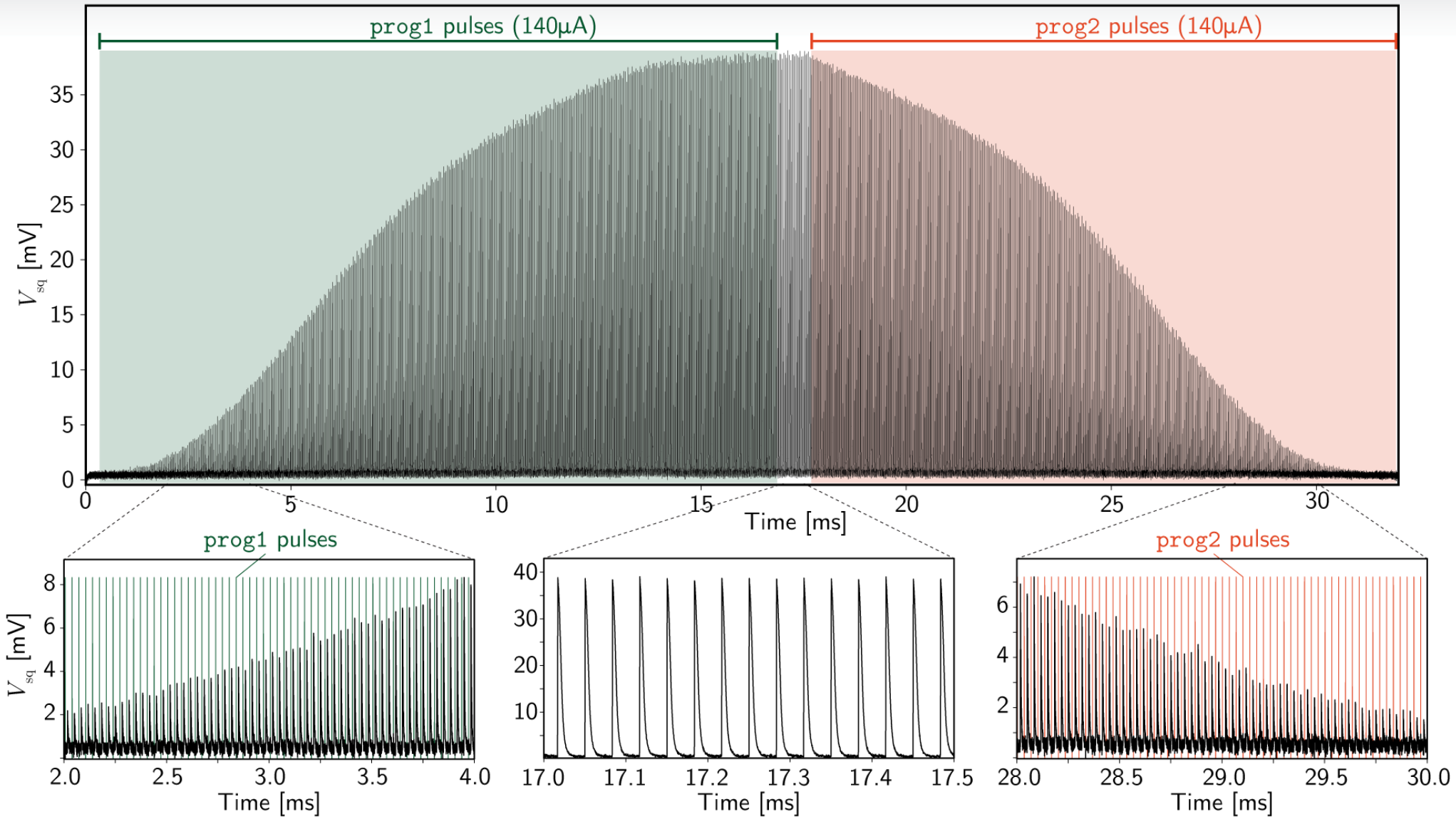


Synapses with local memory: 3bit



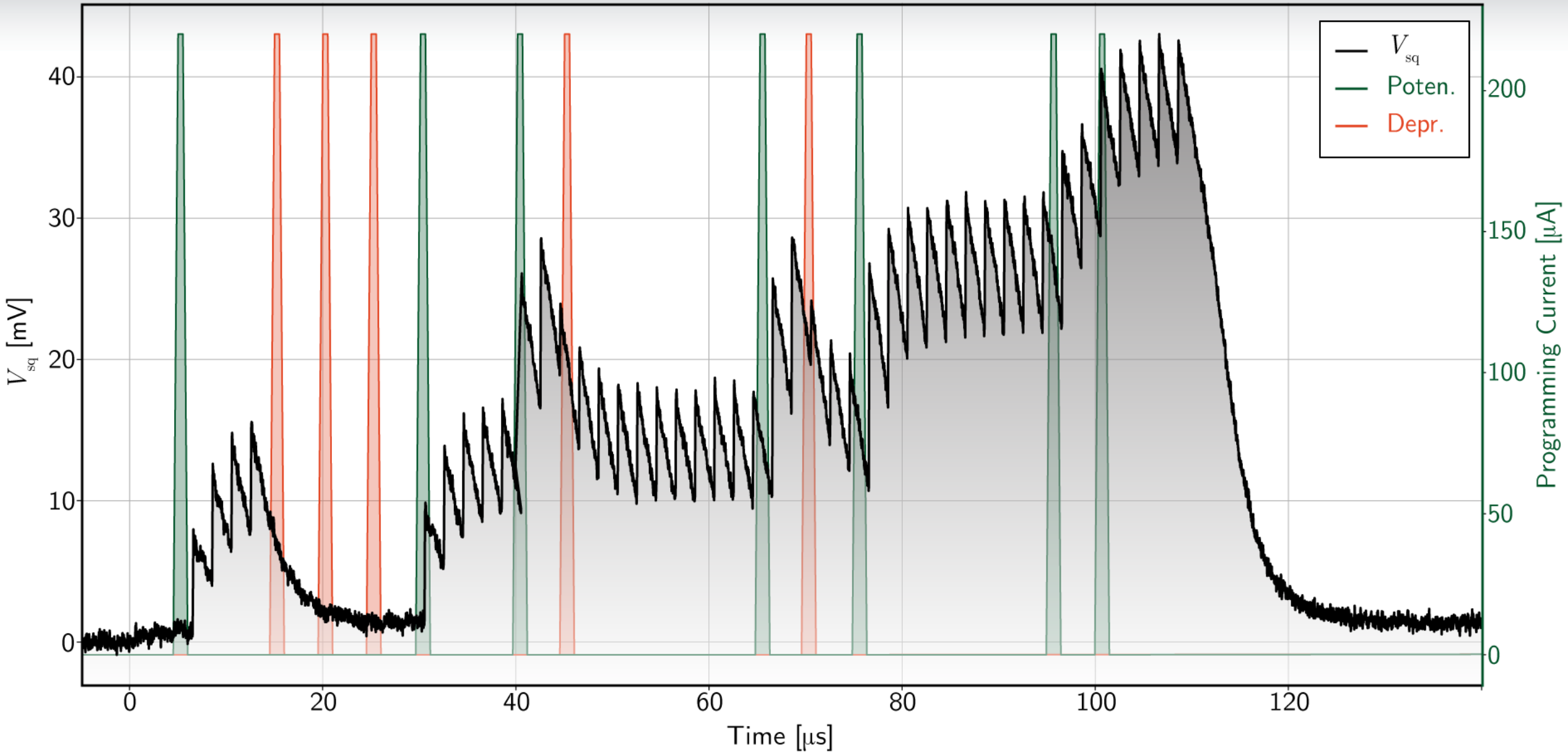
Primavera et al., APL Machine Learning, 2, 026122 (2024)

Synapses with local memory: 8bit



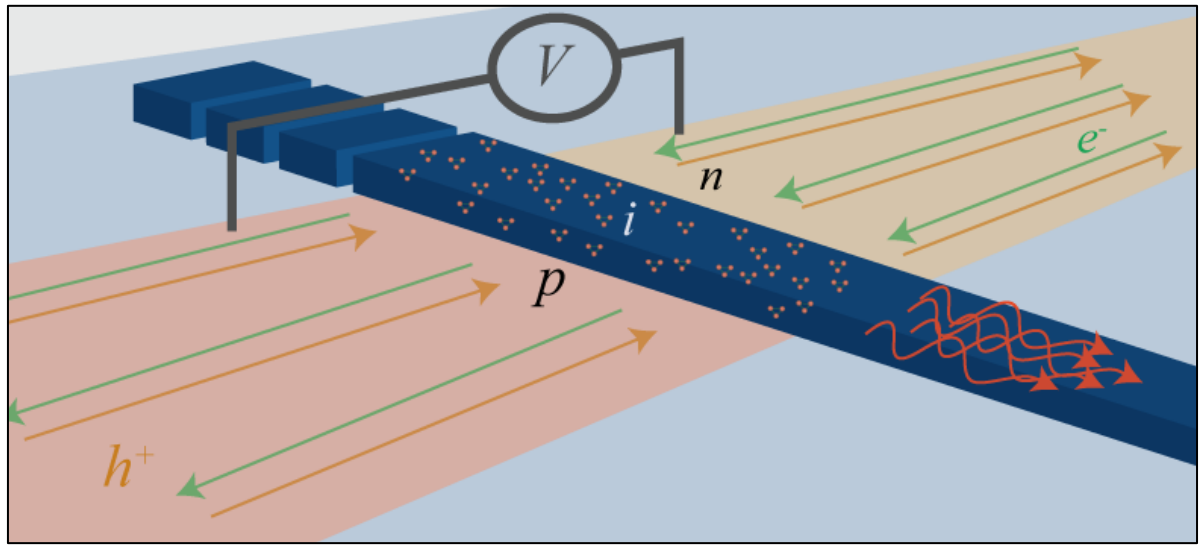
Primavera et al., APL Machine Learning, 2, 026122 (2024)

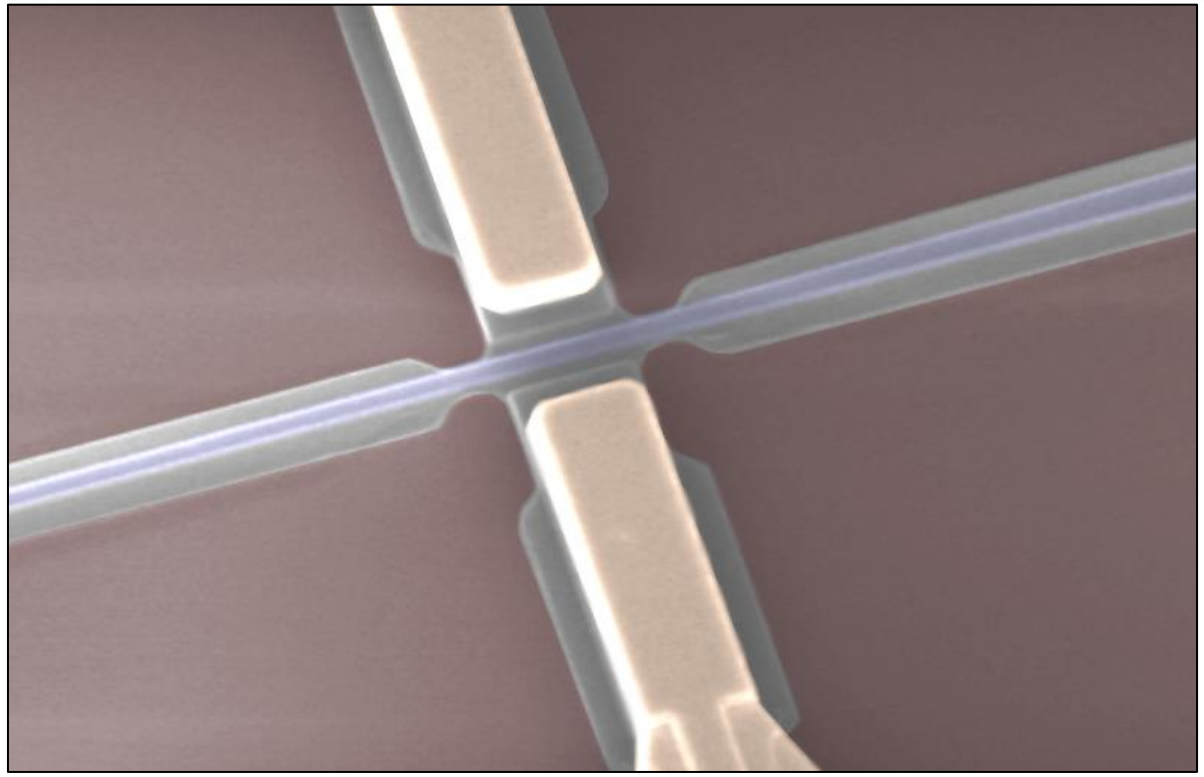
Programming during spike trains



Primavera et al., APL Machine Learning, 2, 026122 (2024)

Silicon light sources

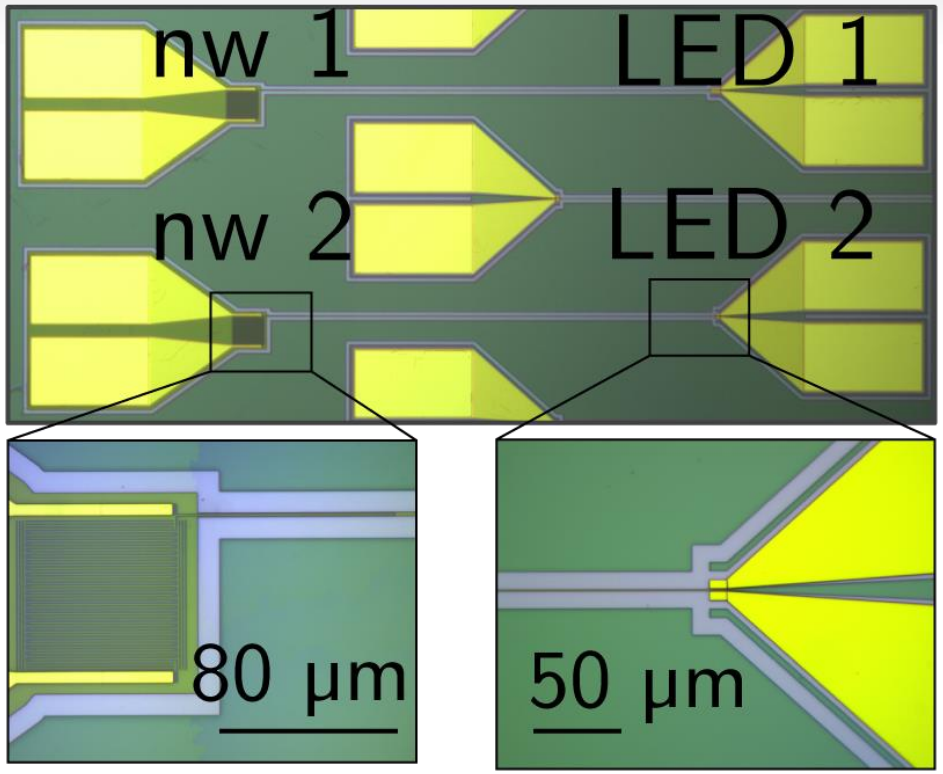


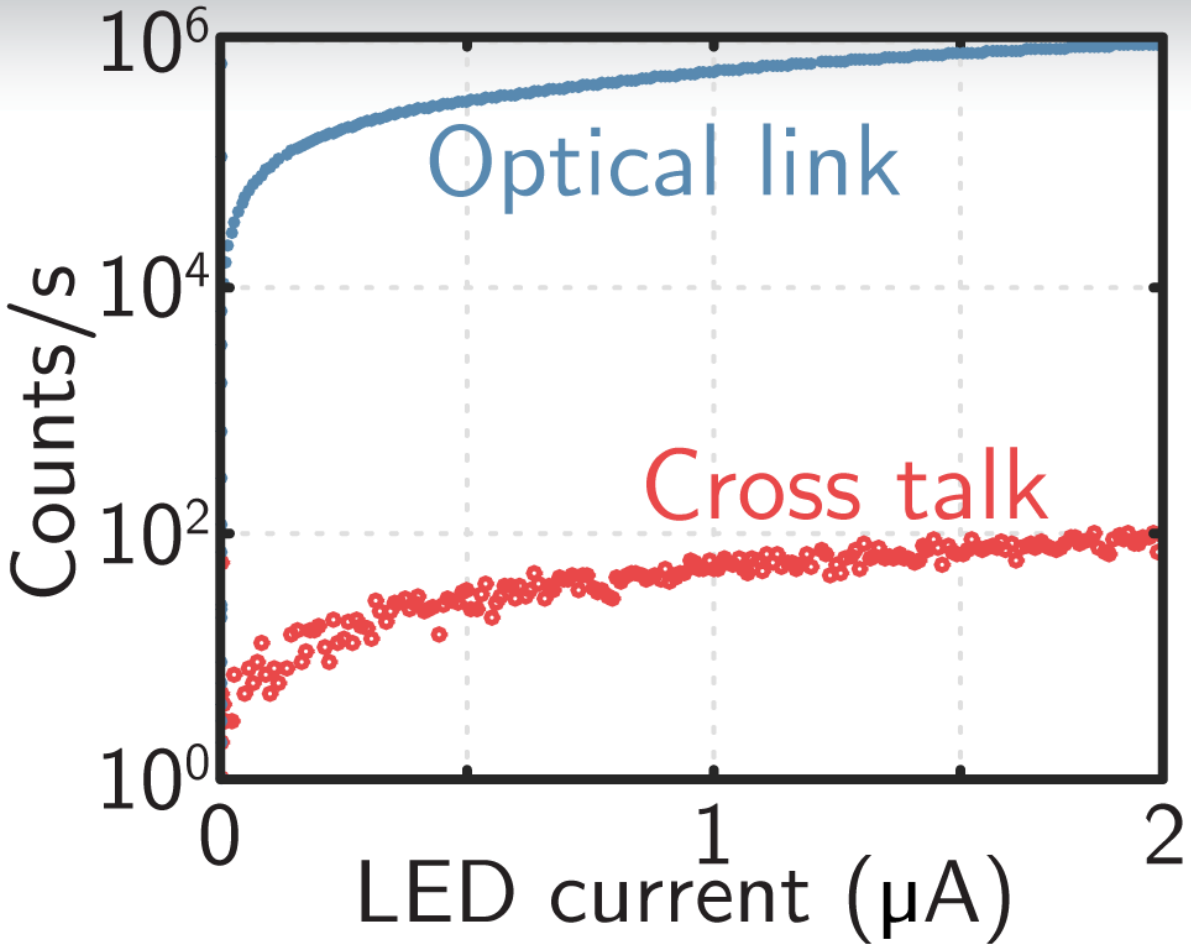


Light sources as simple as transistors

Buckley et al., Appl. Phys. Lett., 111, 141101 (2017)

Silicon light sources



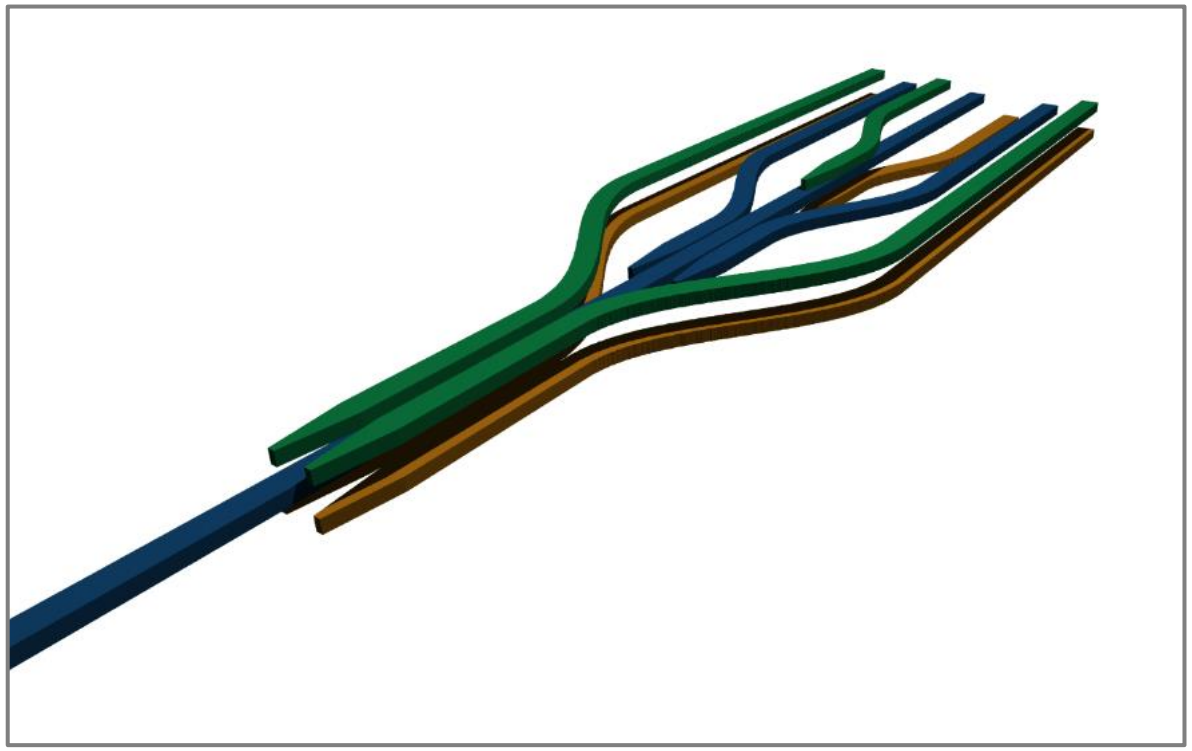


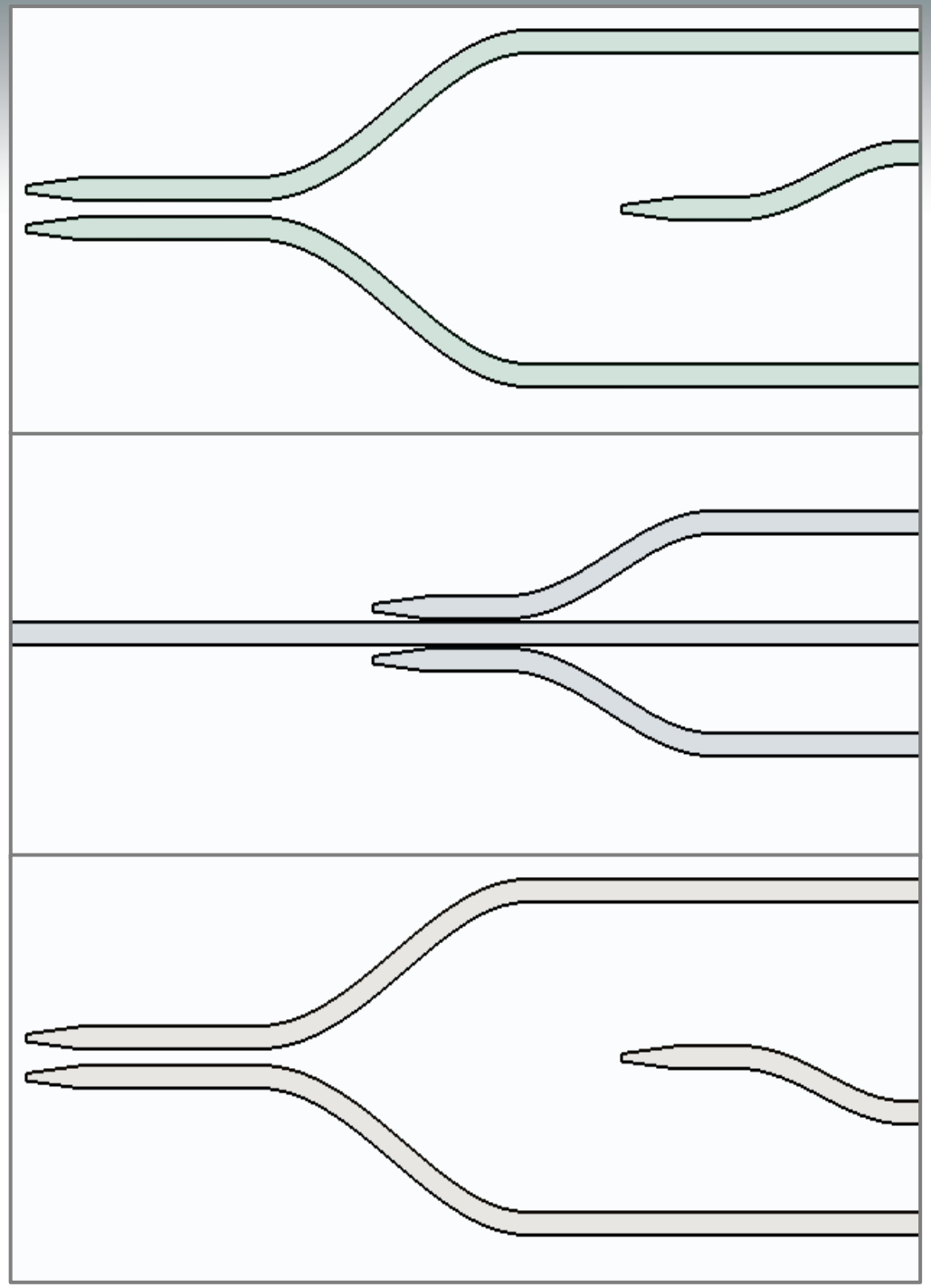
All-silicon photonic links at the single-photon level

Only possible at low temperature

Buckley et al., Appl. Phys. Lett., 111, 141101 (2017)

Photonic interconnects





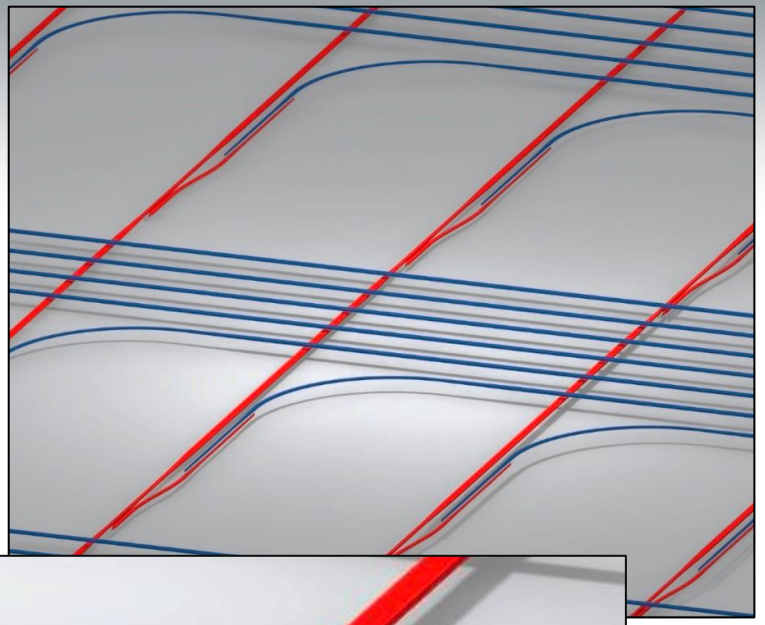
Chiles et al., APL Photonics, 2, 116101 (2017)

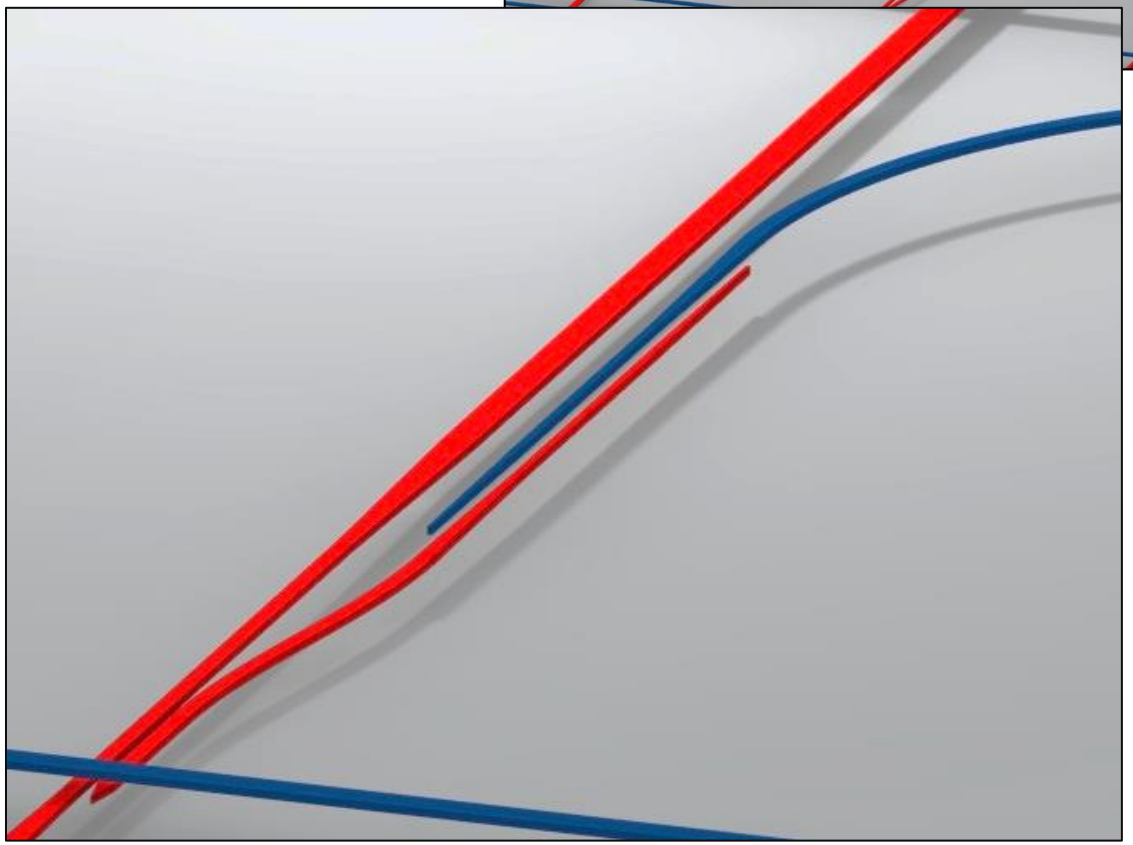
Photonic interconnects

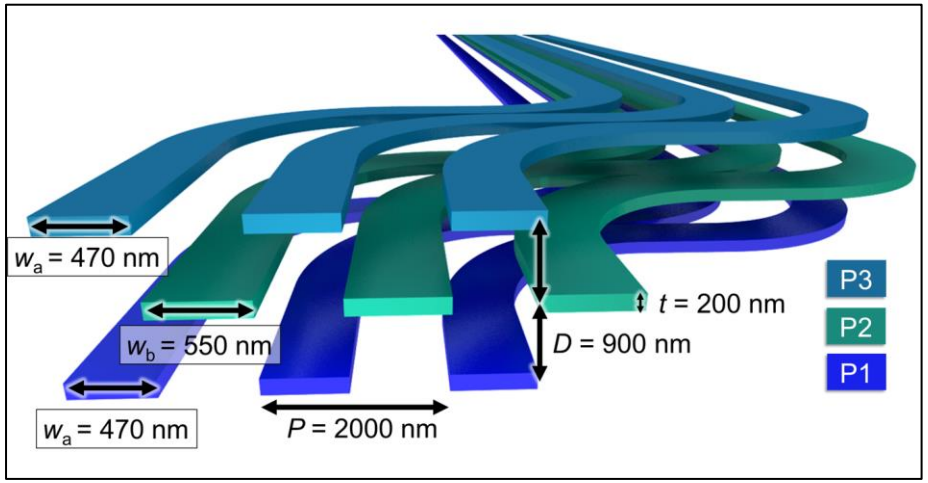
Multiple planes of low-loss routing waveguides

Built-in crosstalk / interference mitigation

3D integration also possible with superconductors



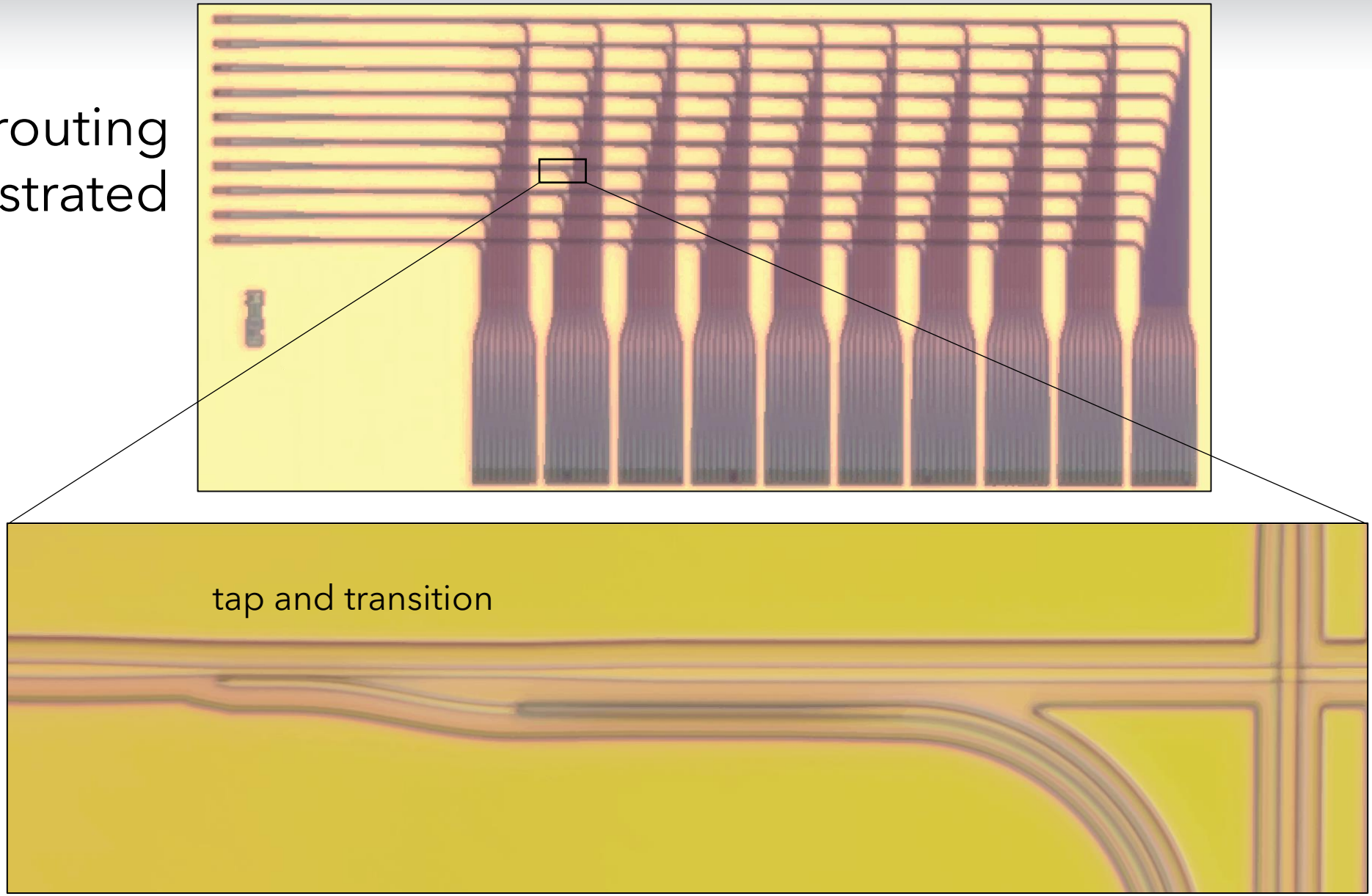




Chiles et al., APL Photonics, 2, 116101 (2017)
Chiles et al., APL Photonics, 3, 106101 (2018)

Photonic interconnects

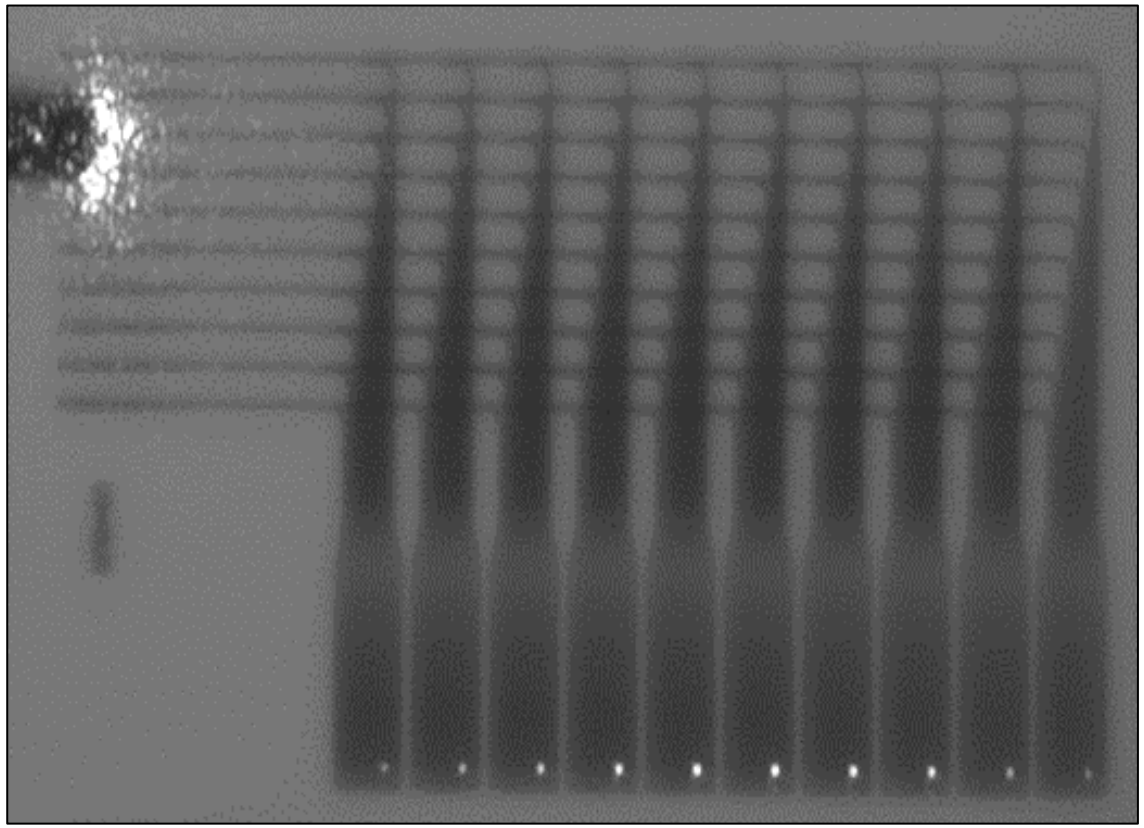
10x100 routing manifold demonstrated

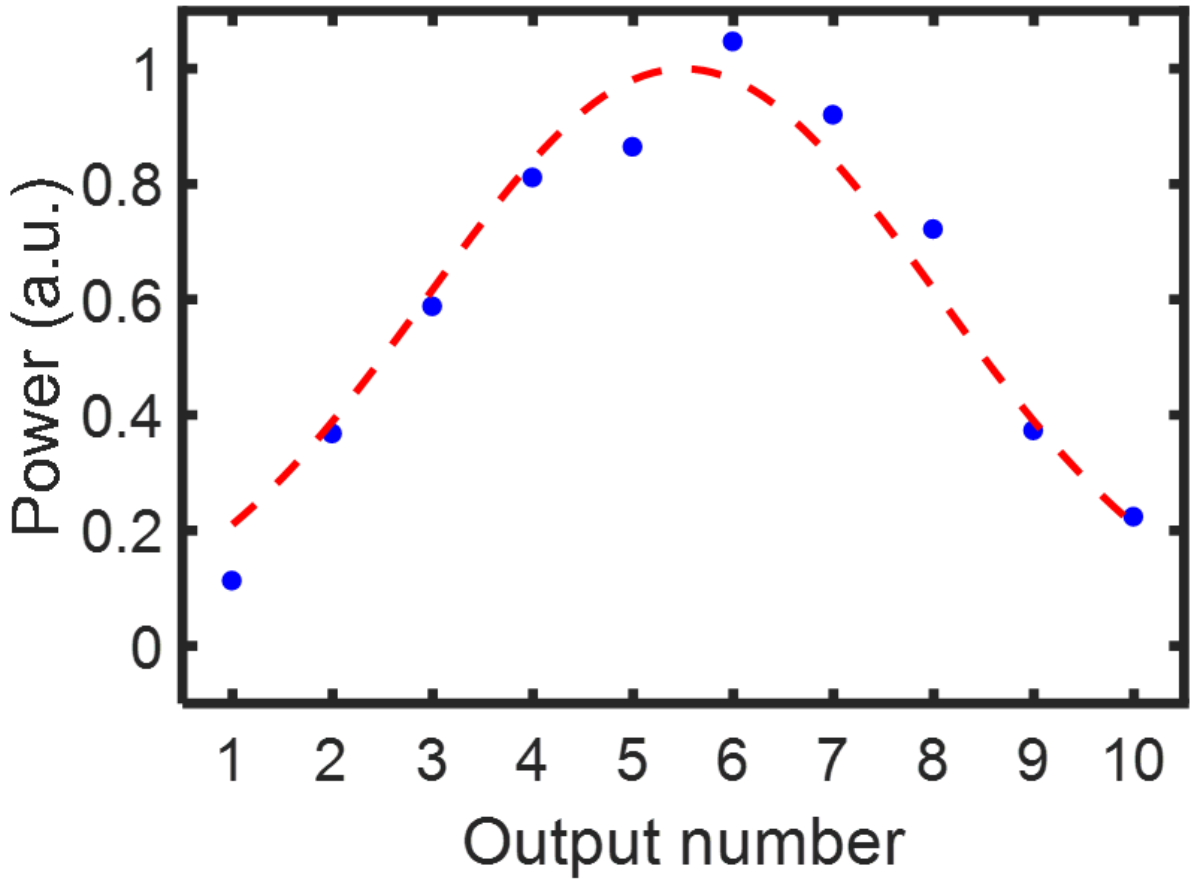


Chiles et al., APL Photonics, 3, 106101 (2018)

Photonic interconnects

10x100 routing manifold demonstrated

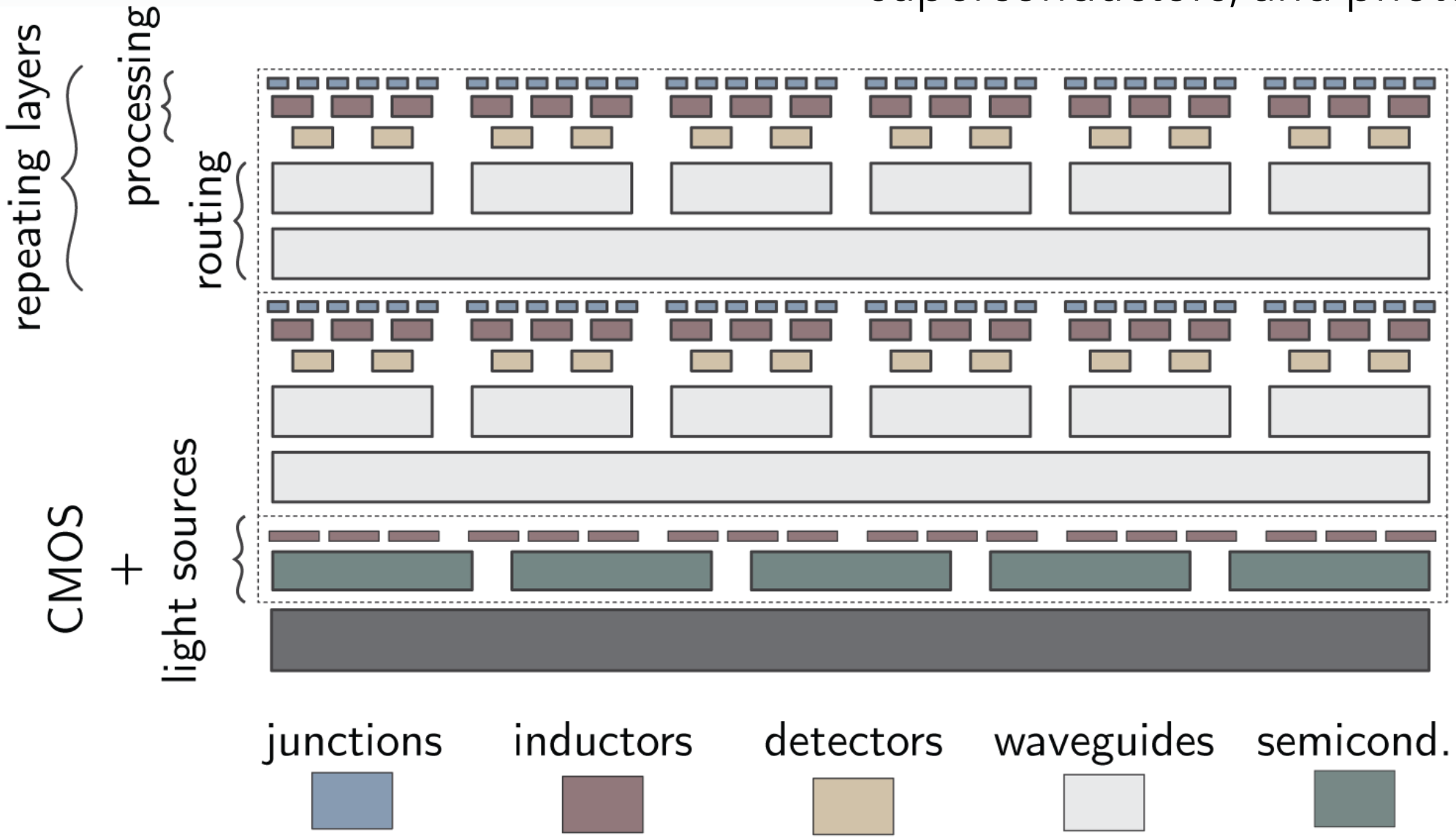




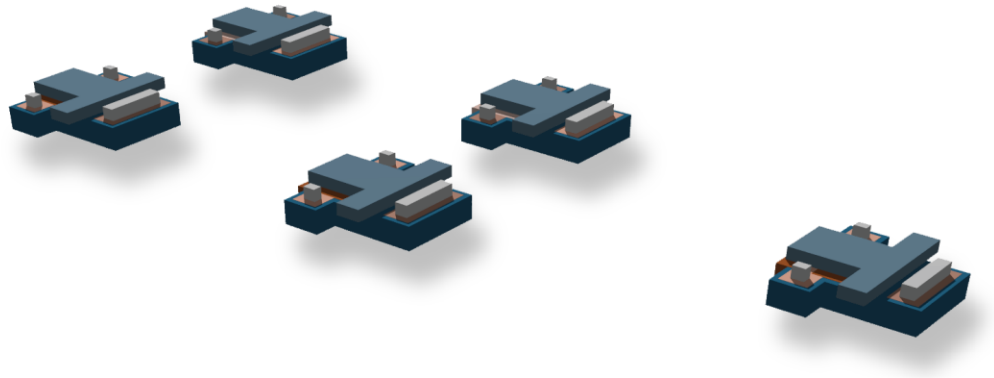
Chiles et al., APL Photonics, 3, 106101 (2018)

The fabrication stack

Combining semiconductors, superconductors, and photonics



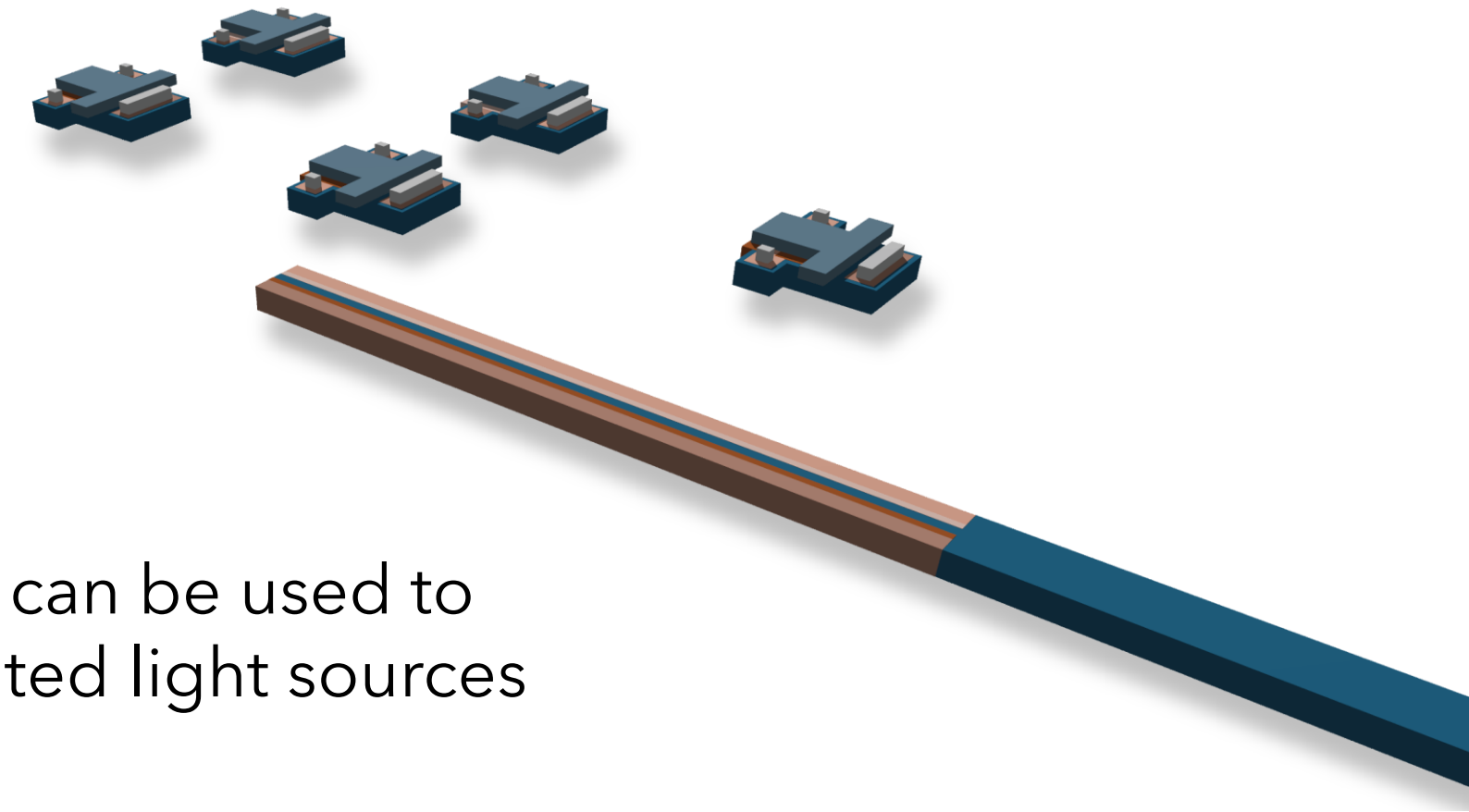
Fabrication begins with semiconductors



Conventional CMOS for transmitter driver circuits and control functions

45nm process will suffice for long term

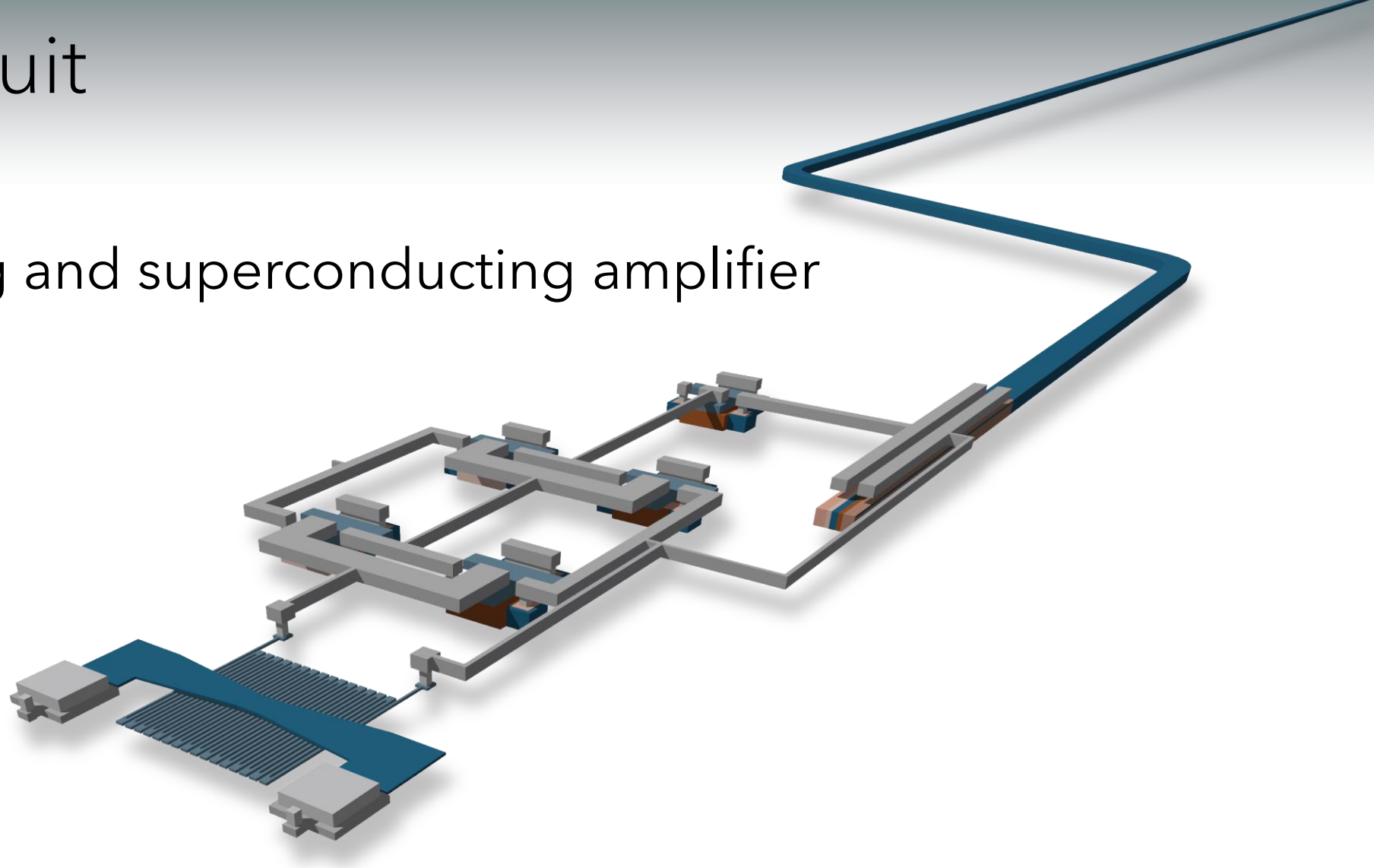
Semiconductor components



Same body silicon layer can be used to form waveguide-integrated light sources

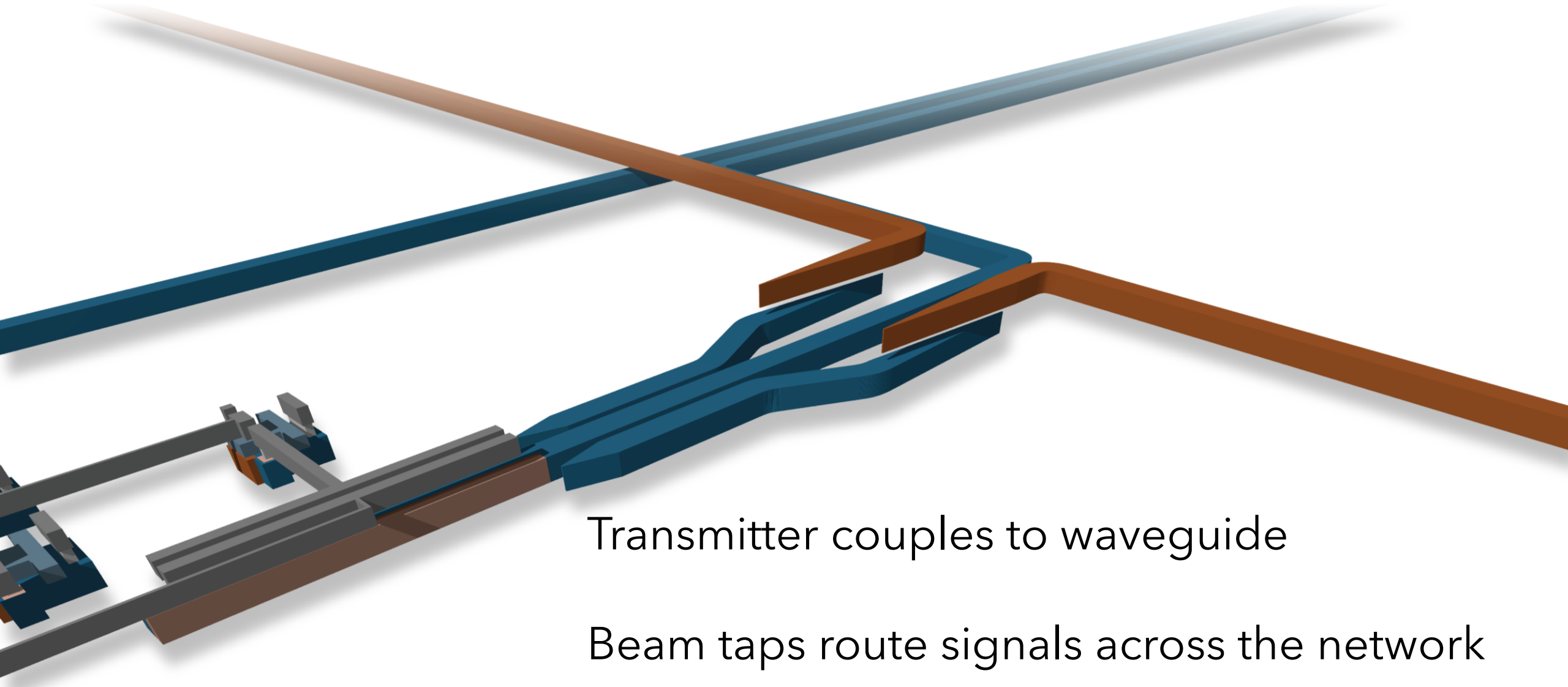
Transmitter circuit

Interconnect wiring and superconducting amplifier



Amplifier that allows superconducting circuits (mV signal)
to drive the semiconductor transmitter (1V signal)

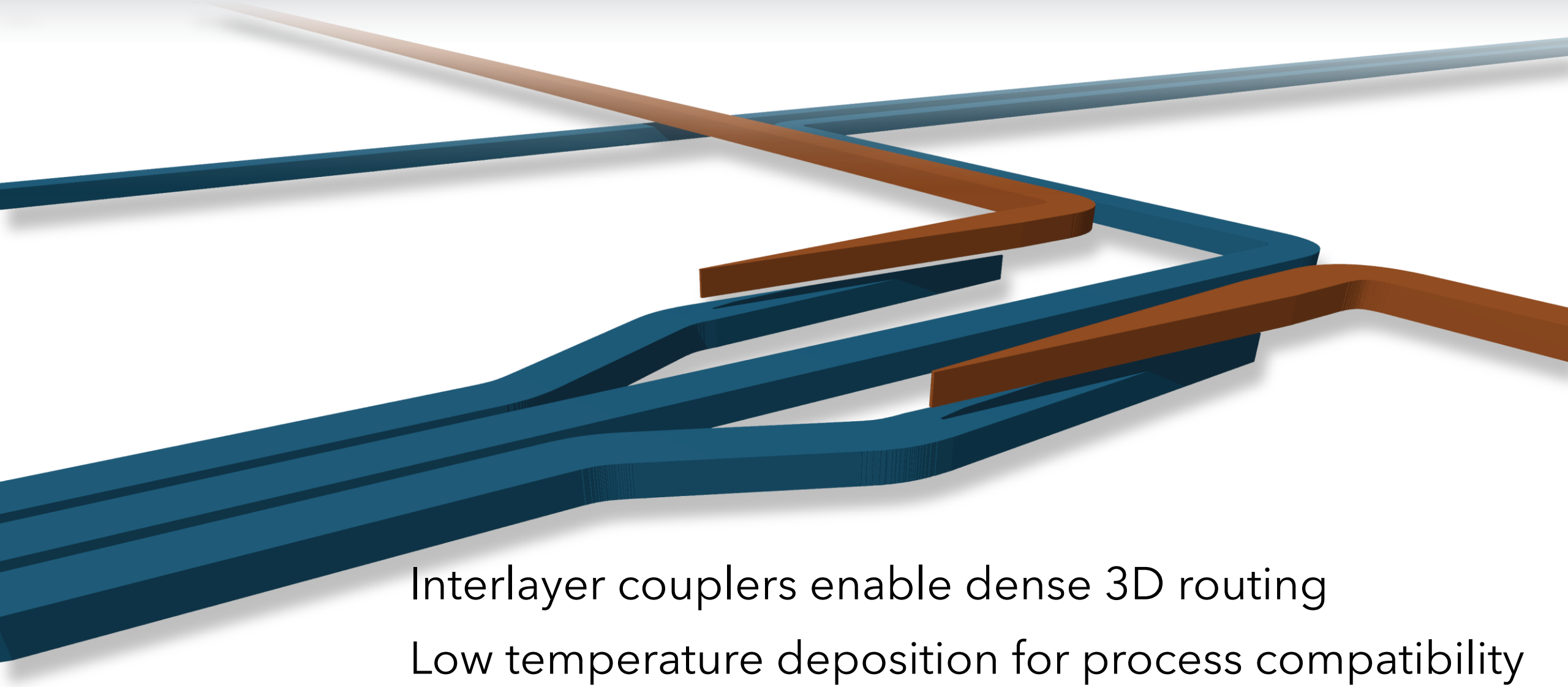
Waveguide coupling and fan-out



Transmitter couples to waveguide

Beam taps route signals across the network

Multiplanar passive photonics



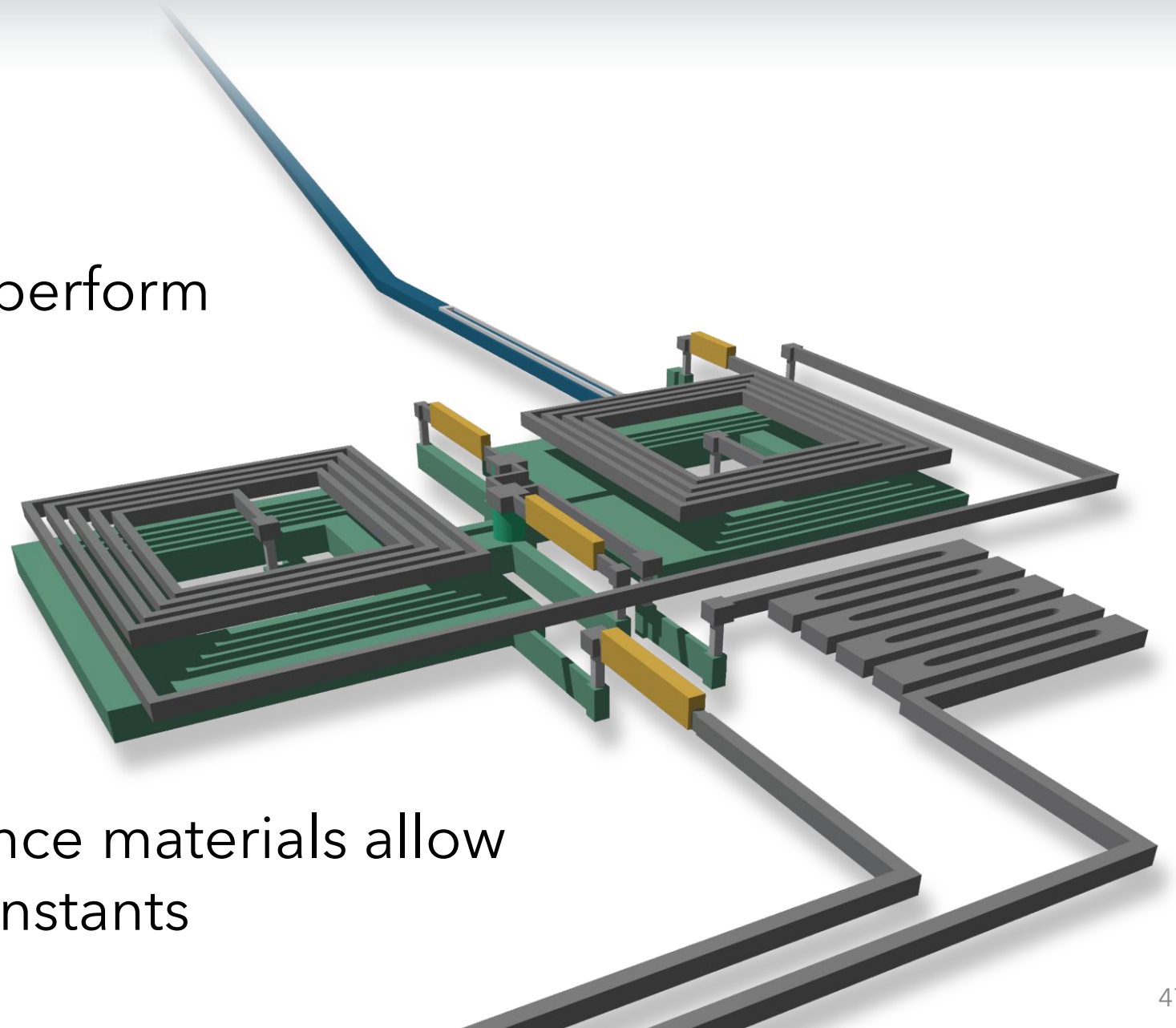
Interlayer couplers enable dense 3D routing

Low temperature deposition for process compatibility

Low propagation loss for cross-wafer communication

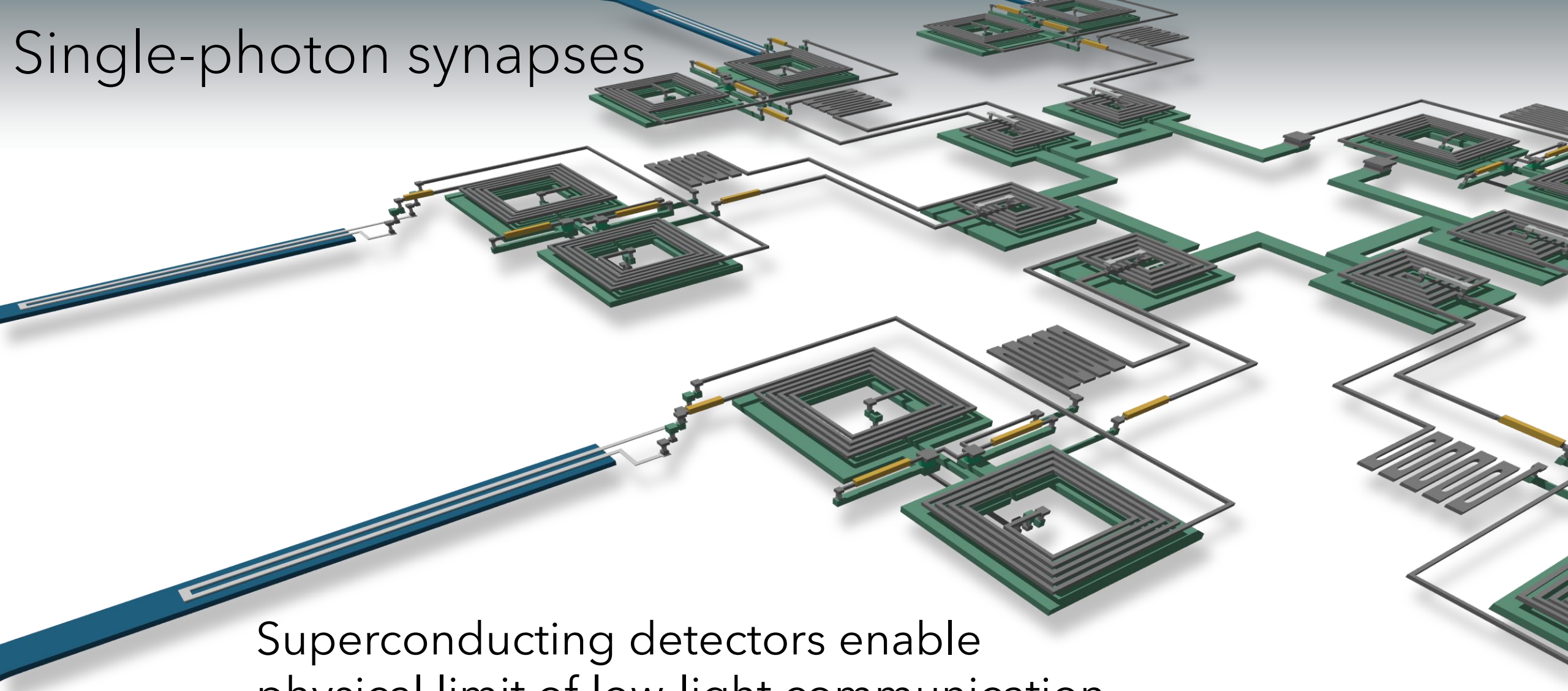
SQUIDs for computation

Superconducting circuits perform neural computations



High-kinetic-inductance materials allow a diversity of time constants

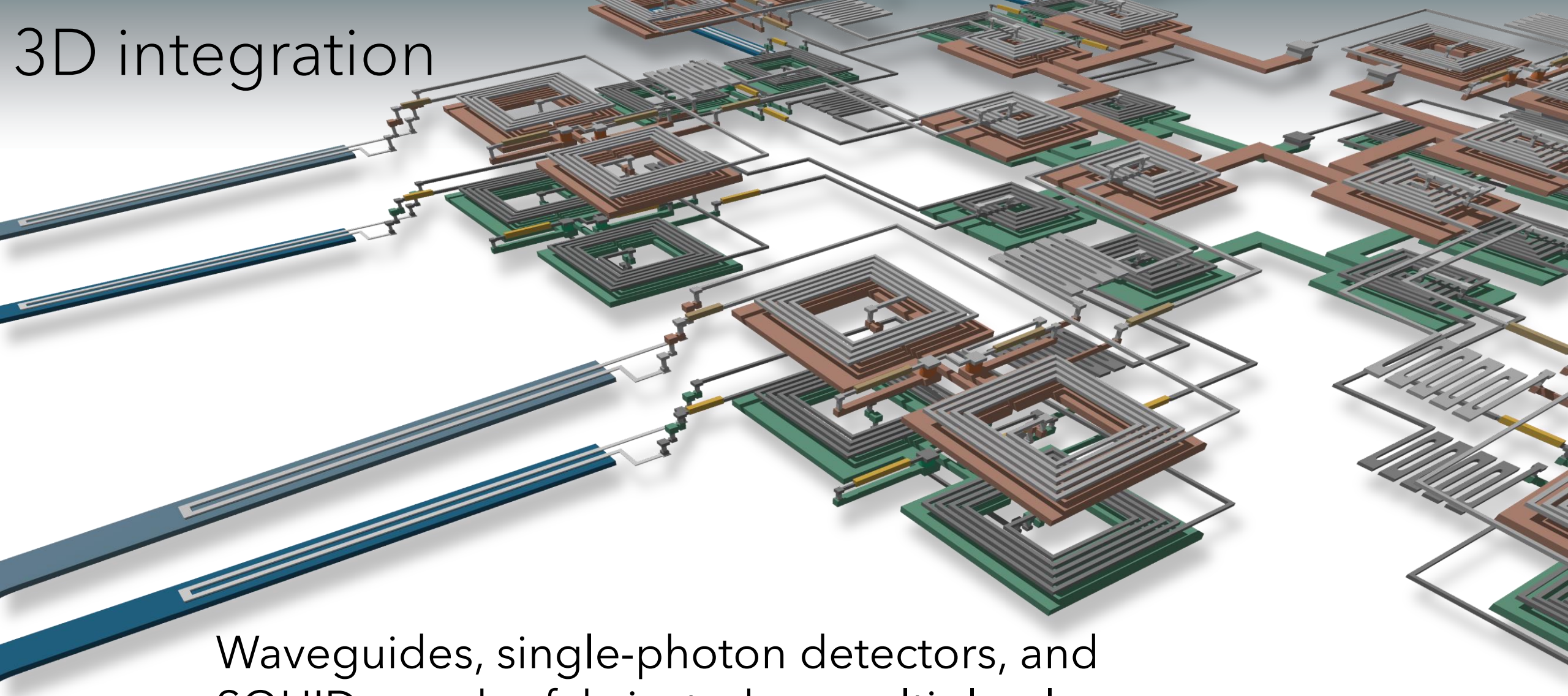
Single-photon synapses



Superconducting detectors enable physical limit of low-light communication

Transduction to electrical signals at SQUIDs enables neural computation

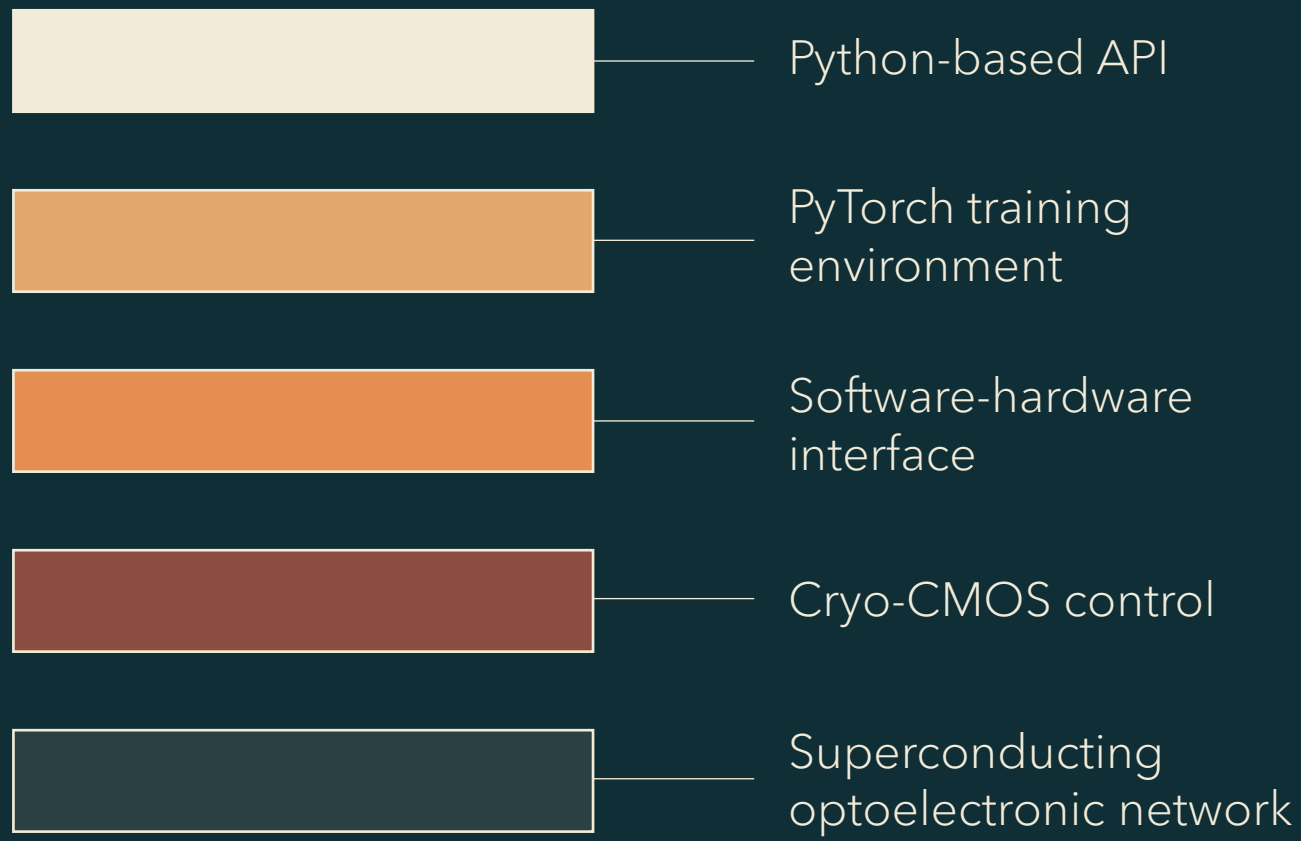
3D integration



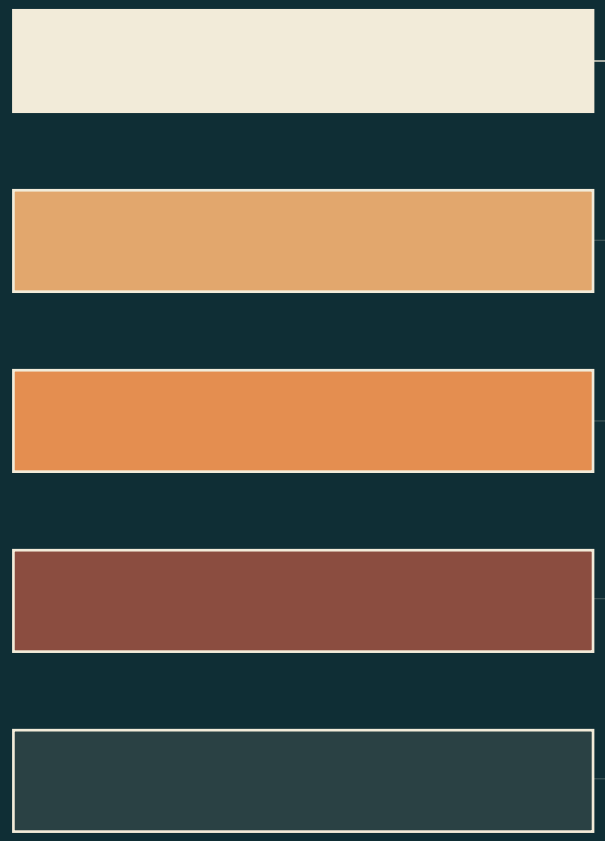
Waveguides, single-photon detectors, and SQUIDs can be fabricated on multiple planes

3D integration of active superconducting devices much easier than semiconductor logic

The Stack


Python-based API
PyTorch training environment
Software-hardware interface
Cryo-CMOS control
Superconducting optoelectronic network

The Stack



Python-based API

PyTorch training environment

Software-hardware interface

Cryo-CMOS control

Superconducting optoelectronic network

Instantiate network architectures

Utilize training data

Analyze outputs from hardware

The Stack

Python-based API

PyTorch training environment

Software-hardware interface

Cryo-CMOS control

Superconducting optoelectronic network

Develop algorithms/architectures

Train models in software

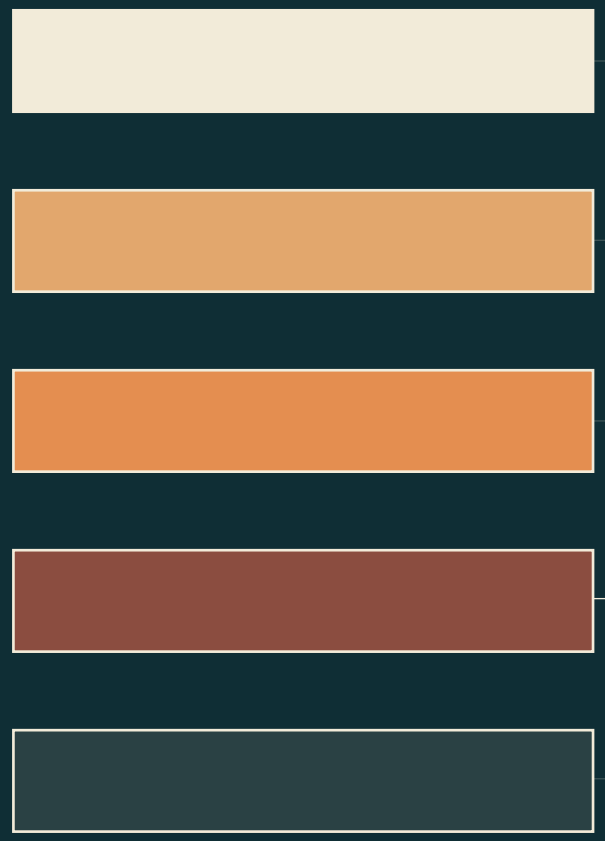
Design networks to be mapped to hardware

The Stack

Translate model parameters to electrical signals that configure hardware

Room-temperature hardware infrastructure for data input, algorithm implementation, and signal readout

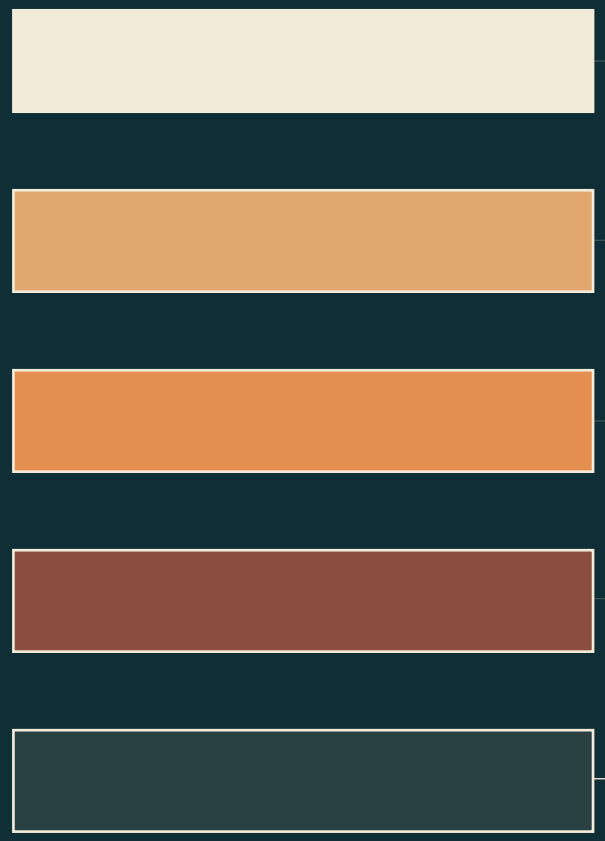
The Stack



Cryo-CMOS control circuits at temperature stages from 77K to 4K

Establish biases, program weights, read outputs, communicate to room temperature

The Stack



Python-based API

PyTorch training environment

Software-hardware interface

Cryo-CMOS control

Superconducting optoelectronic network

Synapses, dendrites, neurons

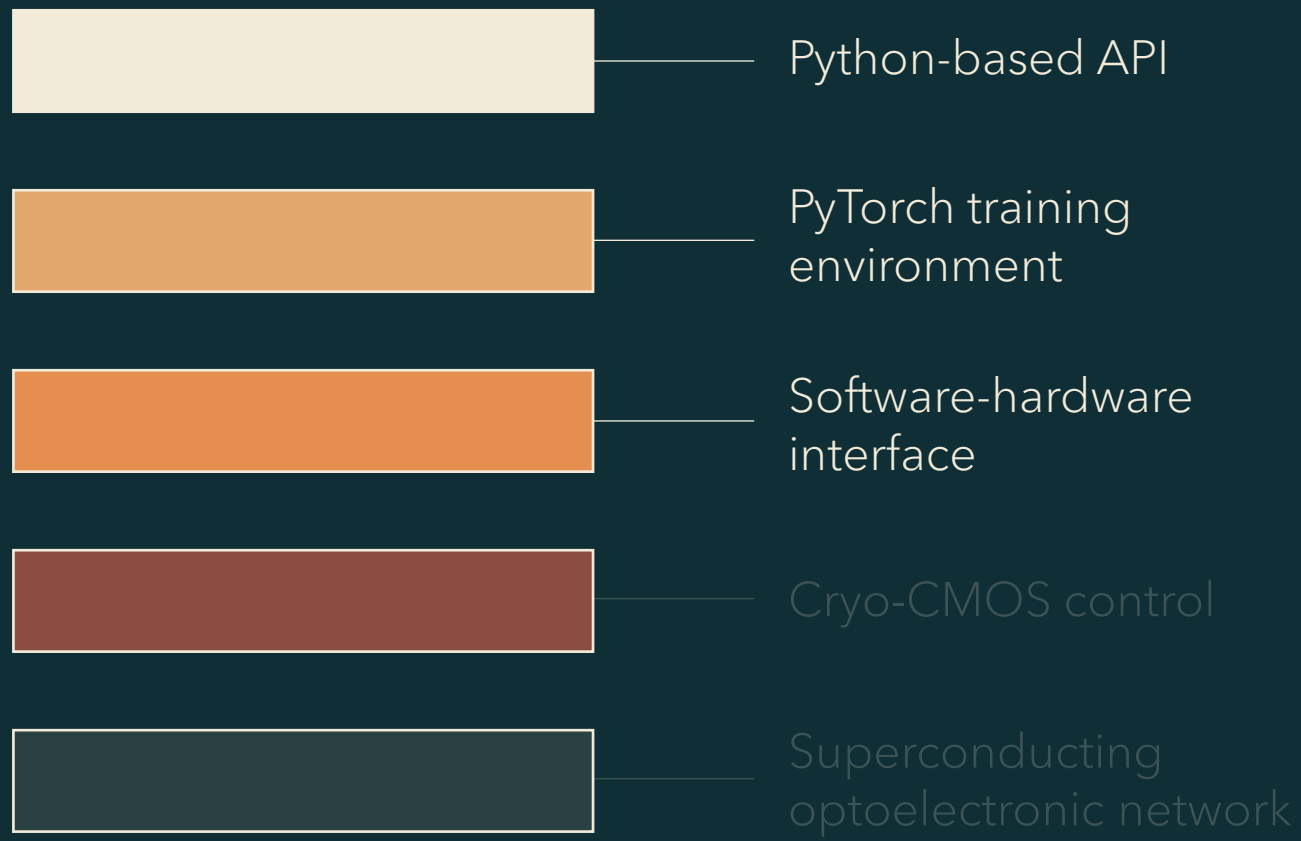
Optical interconnection network

All hardware to execute models

The *Software* Stack

"What's your CUDA?"

The *Software* Stack



The *Software* Stack

Standard neural networks map directly to our hardware

Design and train our networks with PyTorch

Parameters learned with backprop

Rapid training of models with parallel scan on GPUs

Two use cases

Inference engine:

Design network in simulation, determine parameters with backprop run on GPUs

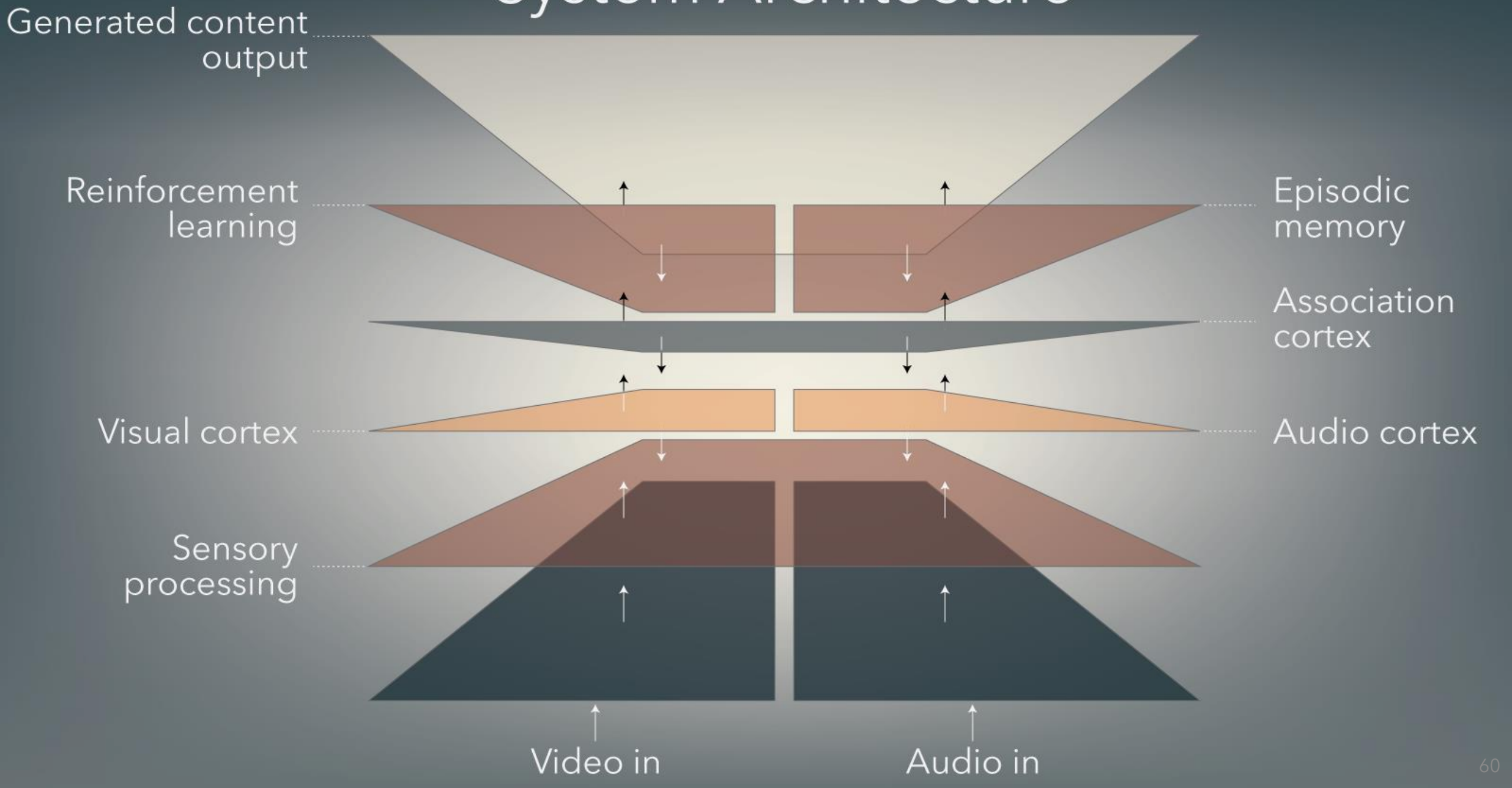
Cryo-CMOS control circuitry to program parameters learned offline

Learning in Hardware:

Implement neural learning algorithms on-chip for faster learning with lower power

Enables lifelong learning with continually evolving data and context

System Architecture



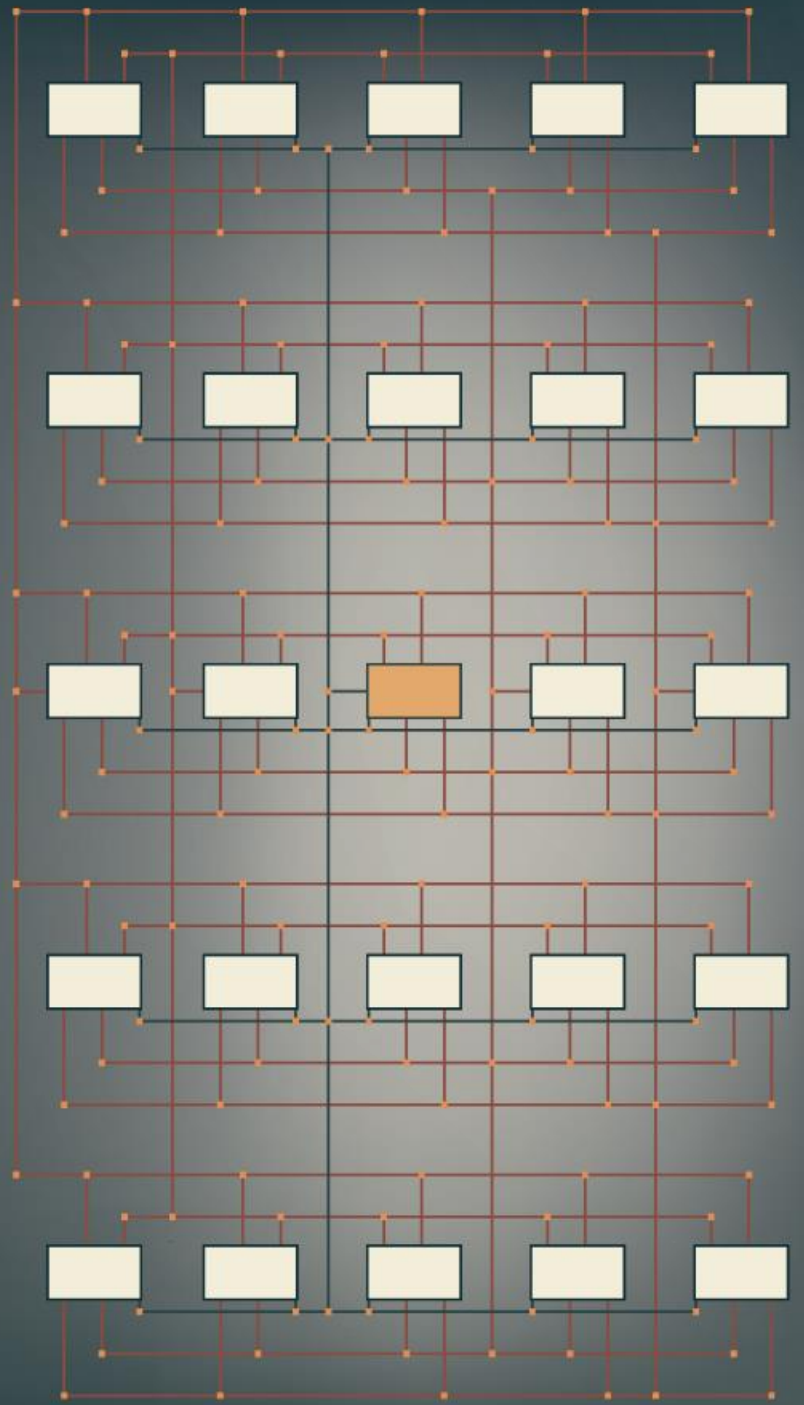
Unique Advantages

Natively implement recurrent neural networks

Infinite context windows

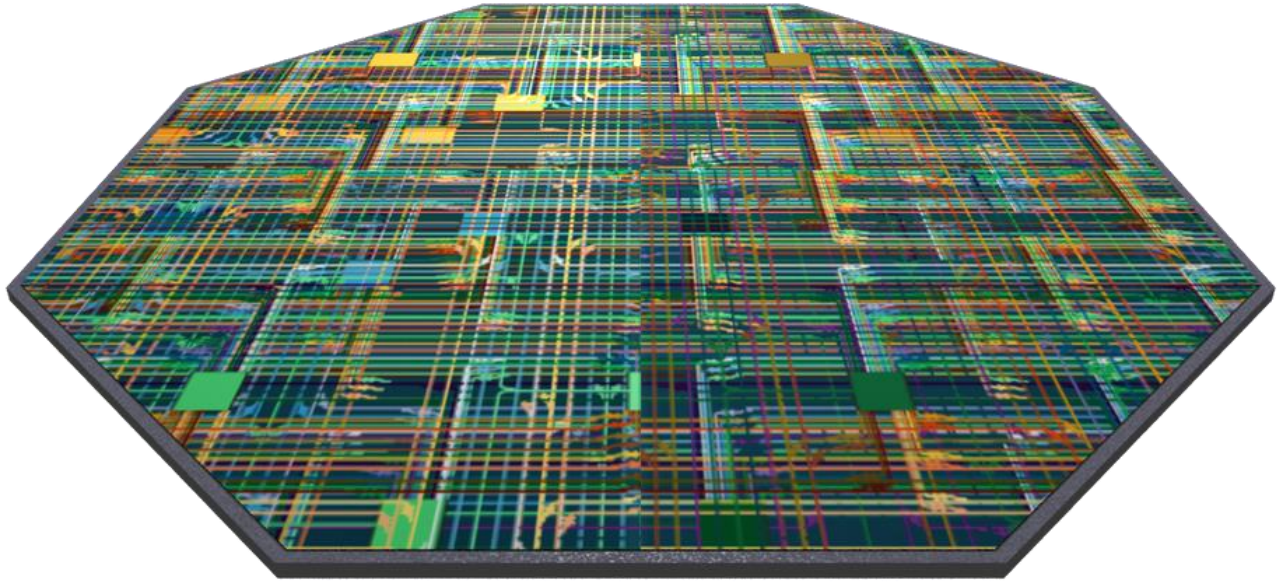
Computational resources grow linearly with sequence length

AI continues to move toward neuroscience



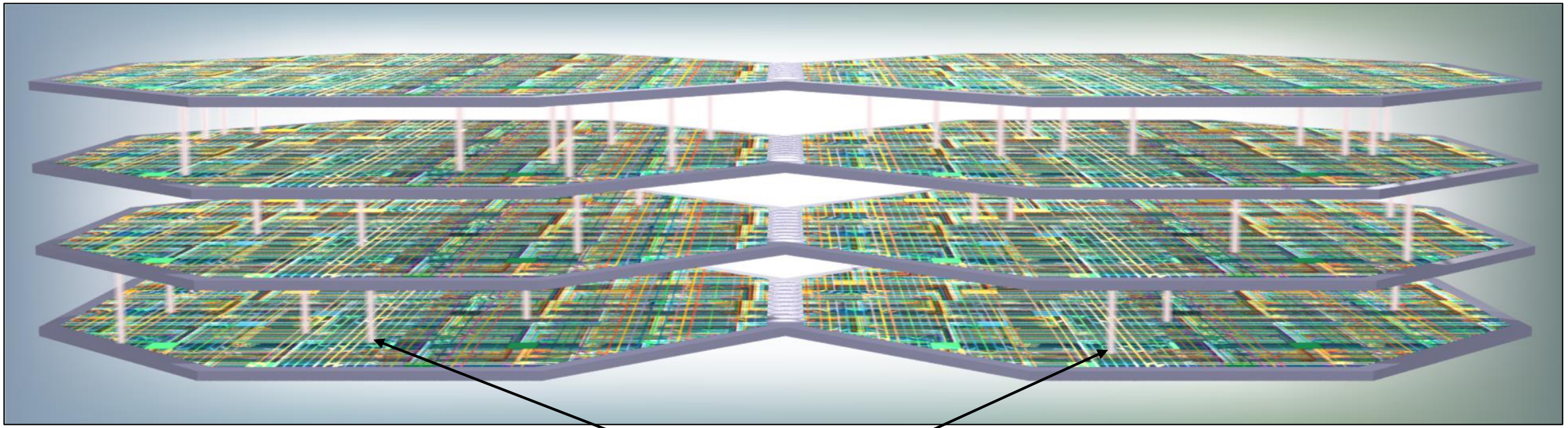
Wafer-scale modules

10^6 neurons per wafer
10^9 synapses per wafer



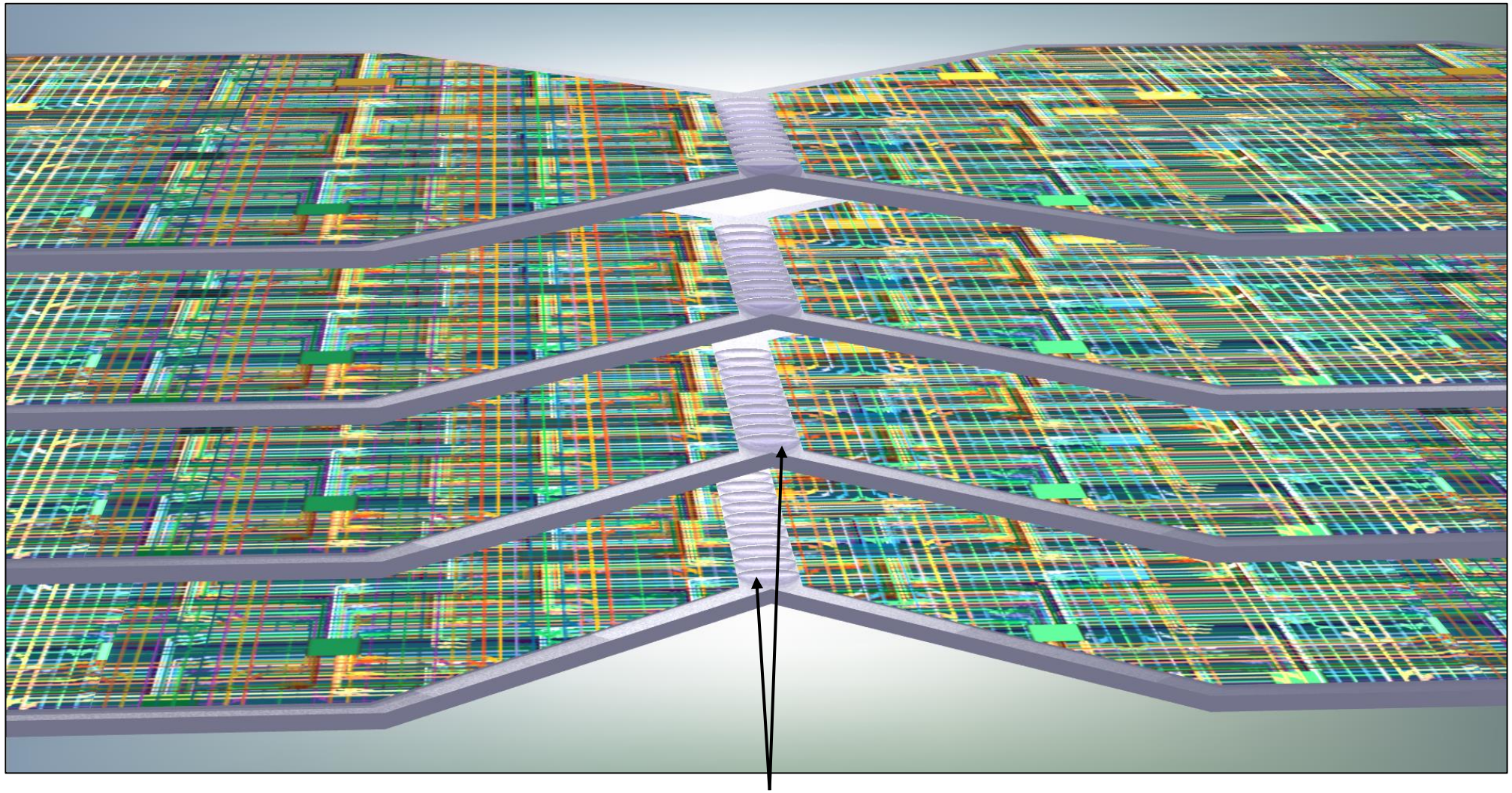
300 mm

Wafers stacked and tiled in cortical columns



Free-space optics for inter-wafer coupling

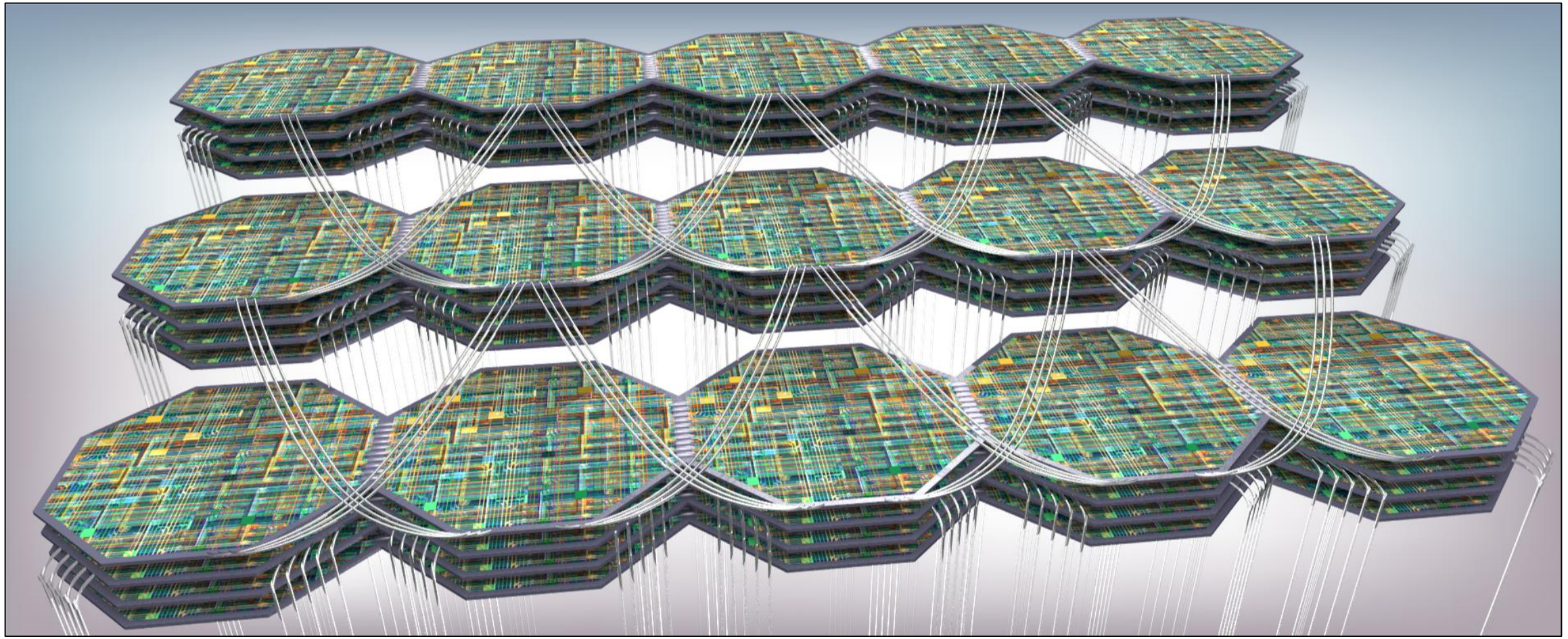
Wafers stacked and tiled in cortical columns



In-plane inter-wafer lateral connections

Superconducting optoelectronic networks (SOENs)

Optical communication crucial at all scales

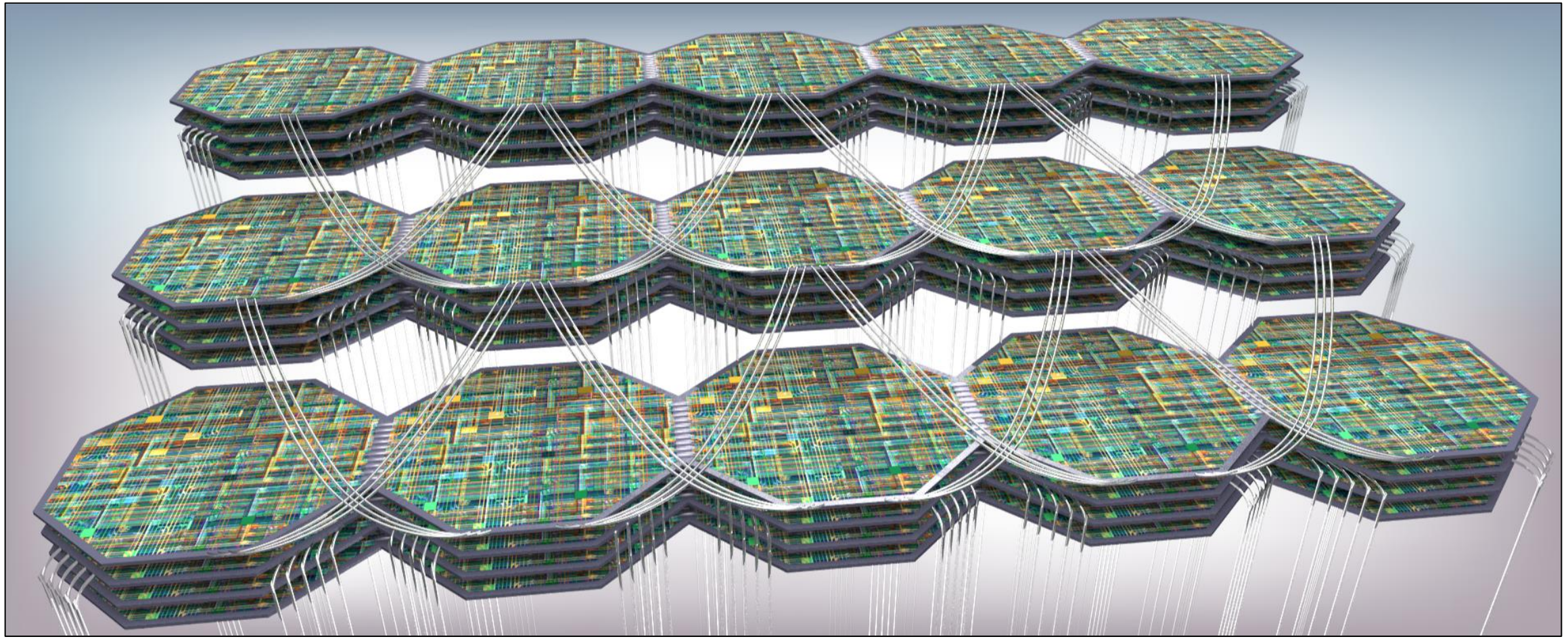


Superconducting optoelectronic networks (SOENs)

10 billion neurons
10 trillion synapses
2-meter cube

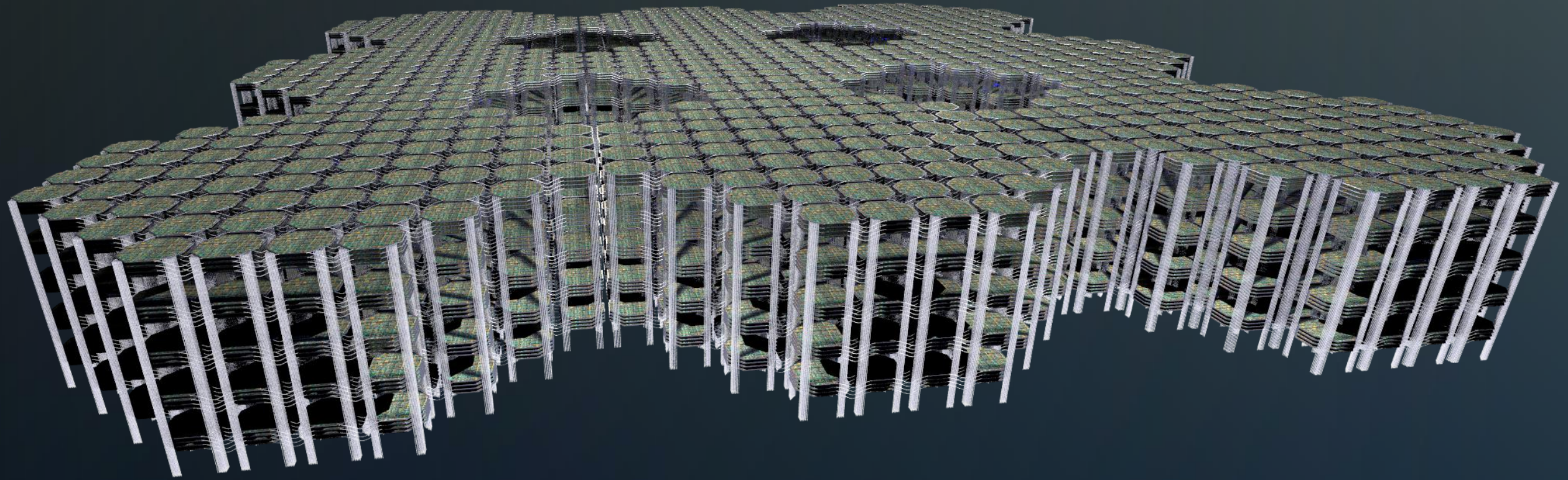
10kW - 1MW

η_{LED} = 10% - 0.1%



Neuromorphic supercomputing

250,000 times faster than the human brain



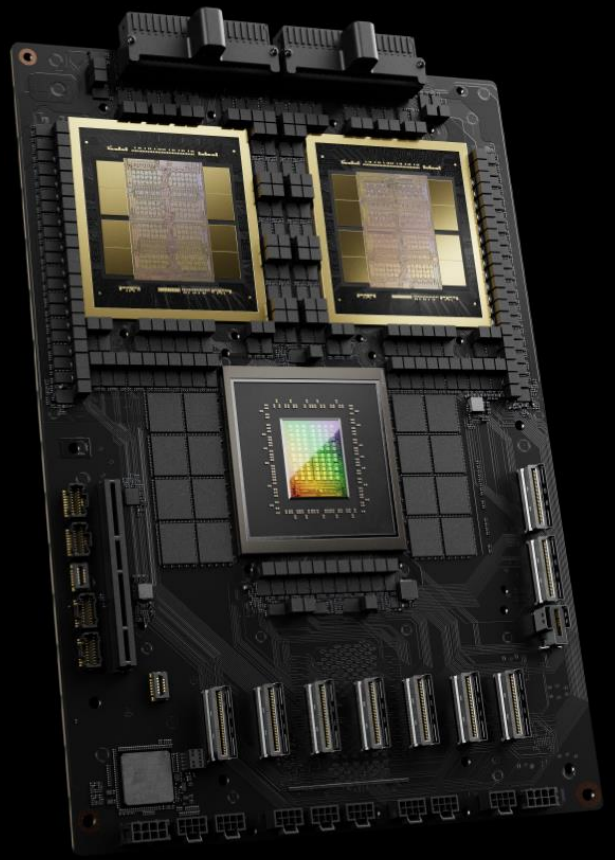
Market

Nvidia’s revenue in 2024: $61B

Major tech companies each spend $10B - $50B annually on compute hardware

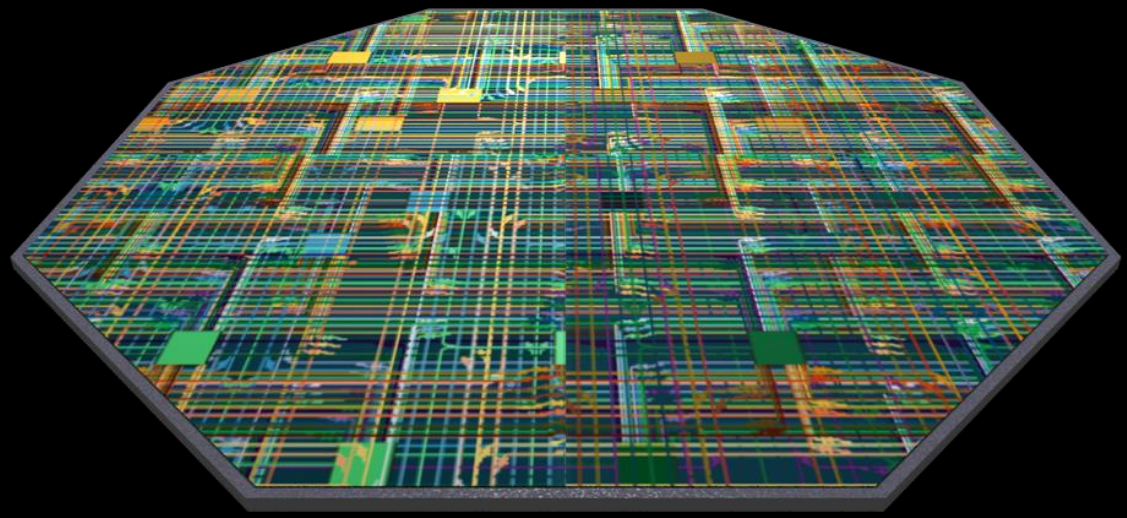
Total AI market: $3T by 2028

Economic Unit



Nvidia Grace Blackwell

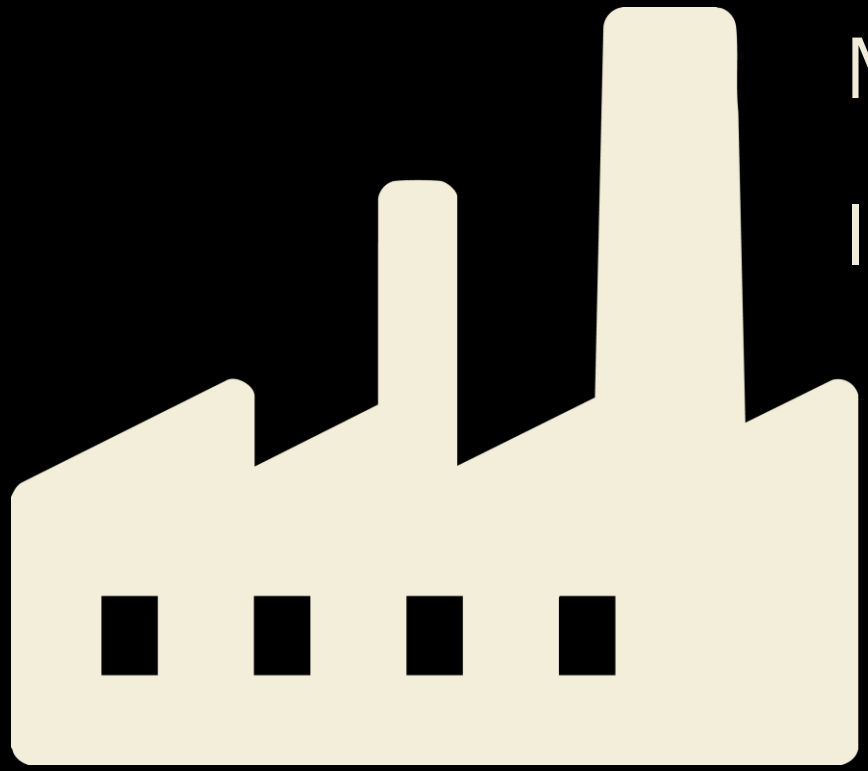
$30k



Great Sky Nova1

$30k

Manufacturing



Manufactured with existing 45nm nodes

Inexpensive: no EUV litho

Revenue

One foundry:

10,000 wafers per day

$30k revenue per wafer

$10k profit per wafer

One 45nm node, $100B annual revenue

First principles of physics and neuroscience

Integrated optoelectronic technology

Physical limits of cognition

shainline@greatsky.ai

Vastly faster than semiconductor neuromorphic

Latency doesn't increase with scale until it is ***limited by the speed of light***

Decades spent incorporating neuroscience in silicon.

The physics doesn't match.

At 10 billion neurons:

Five orders of magnitude faster than the brain

Nine orders of magnitude faster than semiconductors