

Accelerated Photonic Deep Learning: Decentralized and Single-Chip Solutions

Ryan Hamerly and Dirk Englund



MIT AI Hardware Program



School of
Engineering



MIT Schwarzman
College of Computing

"Tell me about the problem of the increasing energy consumption of AI..."



Increasing AI Energy Consumpt x

chat.openai.com/chat/fe6164b2-3c8d-443c-80f7-8ee580713aa0

+ New chat

Increasing AI Energy

Clear conversations

Dark mode

OpenAI Discord

Updates & FAQ

Log out

RH Tell me about the problem of the increasing energy consumption of AI

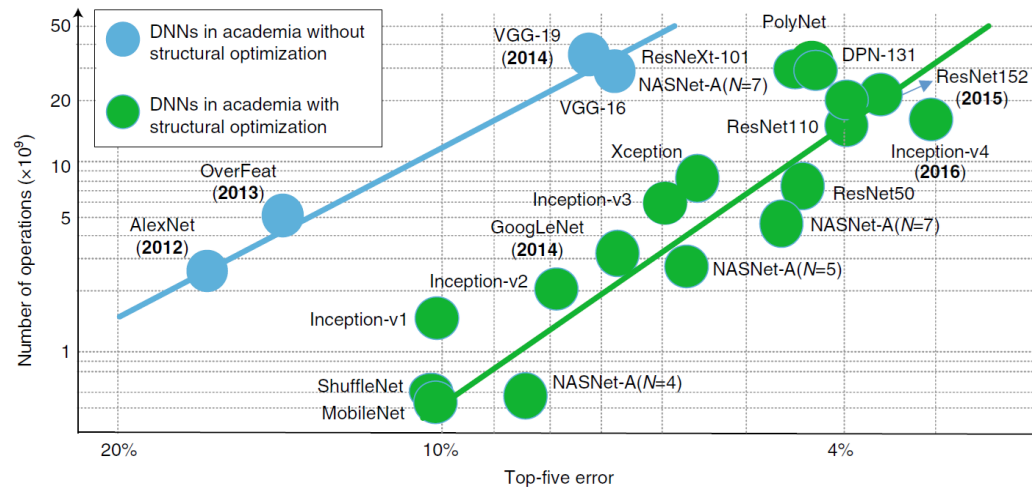
The increasing energy consumption of AI is a growing concern as the use and deployment of AI systems increases. Training large AI models requires significant computational resources, which in turn requires a lot of energy. This energy consumption can have a significant environmental impact, as well as contribute to increasing costs for businesses and organizations that rely on AI. Additionally, the energy consumption of AI systems in operation, particularly those that are constantly running such as in data centers, also add to the overall energy consumption. Efforts to address this issue include developing more efficient algorithms and hardware for AI, as well as utilizing techniques such as model compression and quantization to reduce the computational resources required for AI systems.

Regenerate response

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.



Photonics for AI



From edge sensors to the centralized cloud

The edge computing ecosystem is comprised of four primary areas

Centralized Cloud
Centralized data centers are farthest from the network edge. However, they offer a much greater density of compute, storage, and networking resources.

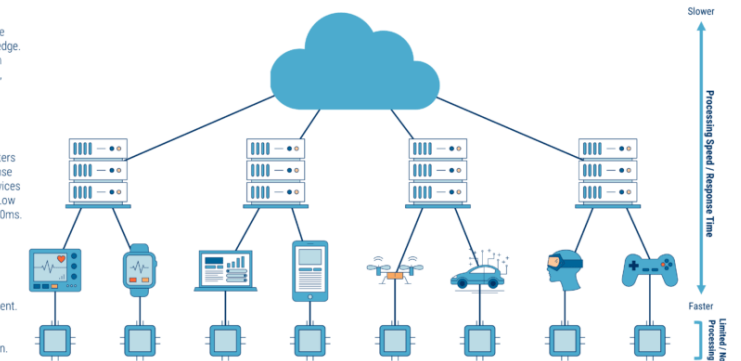
Edge Infrastructure
Small, distributed data centers that provide a resource-dense midpoint between edge devices and the centralized cloud. Low roundtrip latencies of 5 – 10ms.

Edge Devices
Real-time data processing within devices based on application needs. Processing limitations present.

Edge Sensors & Chips
Data collection & origination.

CBINSIGHTS

Source: WinSystems



Exponentially Increasing **compute demand** for AI tasks

Limited gains from Moore's Law due to energy consumption and end of Dennard scaling

Proliferation of **edge computing** applications with severe size, weight and power (**SWaP**) constraints

>> Can photonics provide a way out of this problem? <<



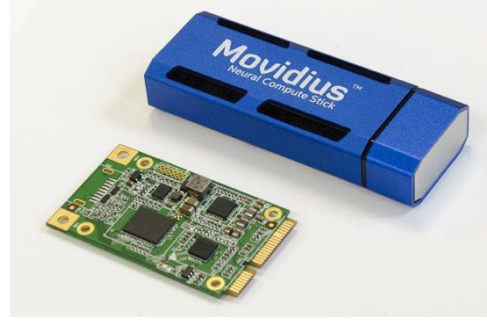
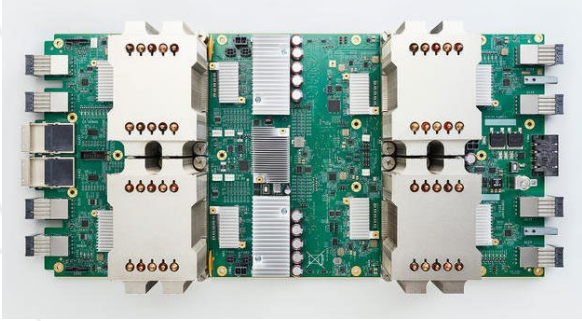
MIT AI Hardware Program



School of Engineering

MIT Schwarzman College of Computing

Hardware “accelerators” for AI: numerous commercial solutions (Google, Intel, Xilinx, Nvidia, etc.)



Three important facts guide design of accelerators:

1. Key metric is **power consumption**.
 - SoA ~ 0.1 -1 pJ/MAC (multiply-and-accumulate)
2. Linear **matrix operations** (synapses) dominate.
 - $O(N^2)$ vs. $O(N)$ scaling
3. **Data movement** & memory access dominates.

[Keckler, *IEEE Micro* 2011; Horowitz, *ISSCC* 2014; Sze, *Proc. IEEE* 2017]

>> Can photonics provide a way out of this problem? <<

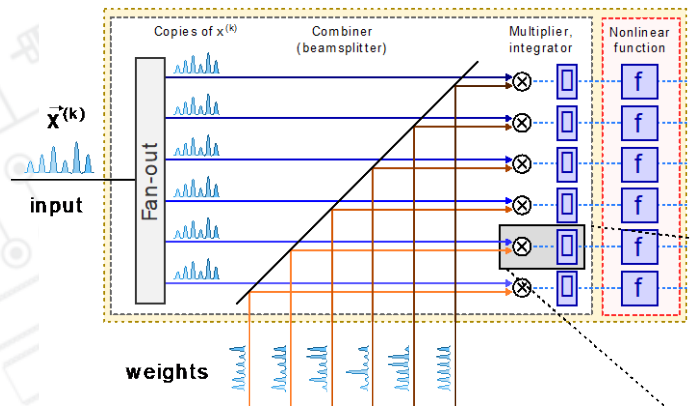
Optical *computing* is hard
Optical *interconnects* are the future

Optical *neural networks* are also promising.

Reasons: (a) Most of deep learning is linear algebra
(b) Linear algebra $\rightarrow$ passive linear optics
(c) High speed, low power, WDM compatible
(d) NN's are robust against analog errors

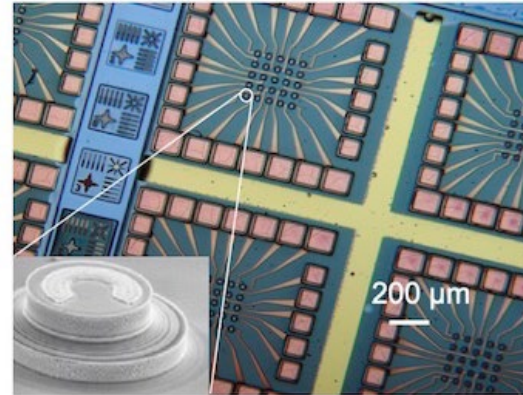
Perspective: MIT QPG / NTT Projects

Coherent Detection



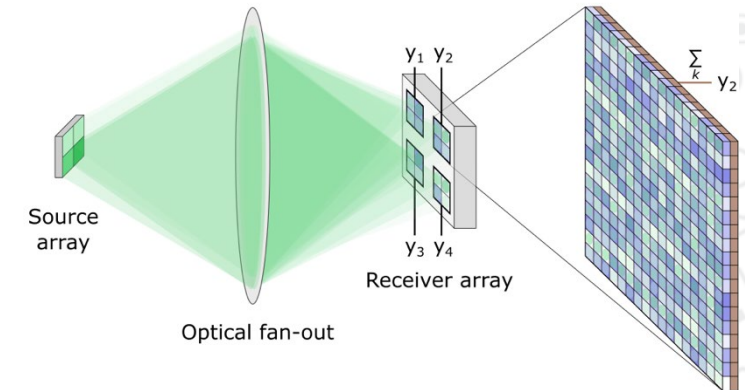
R. Hamerly, *PRX* **9**, 023012 (2019)

VCSEL Arrays



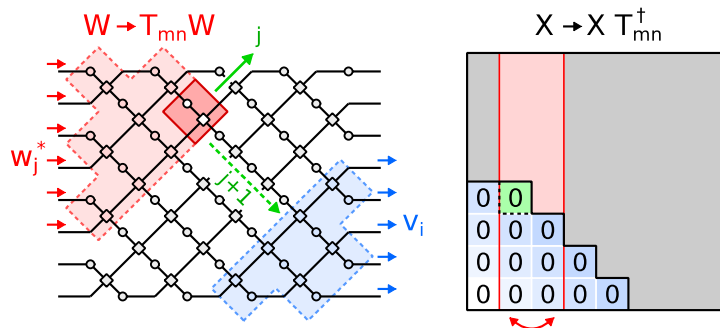
Z. Chen, arXiv:2207.05329 **(12438-18, Tue 9:35)**

Programmable Fan-out



L. Bernstein, arXiv:2205.09103

Photonic Error Correction

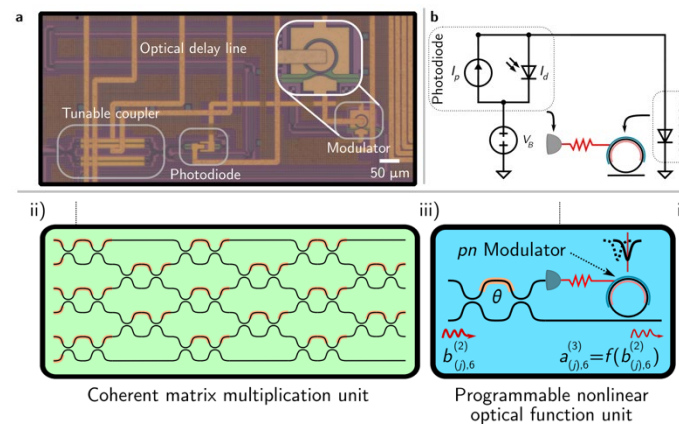


S. Bandyopadhyay, *Optica* **8**, 1247 (2021)

R. Hamerly, *Nat. Comm.* **13**, 6831 (2022)

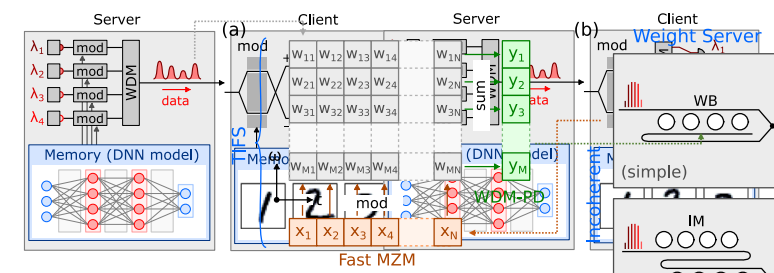
J. Basani, *Nanoph.* 2023 [2209.03335]

3-Layer PNP w/OEO Nonlinearity



S. Bandyopadhyay, arXiv:2208.01623

Netcast: WDM Edge Computing

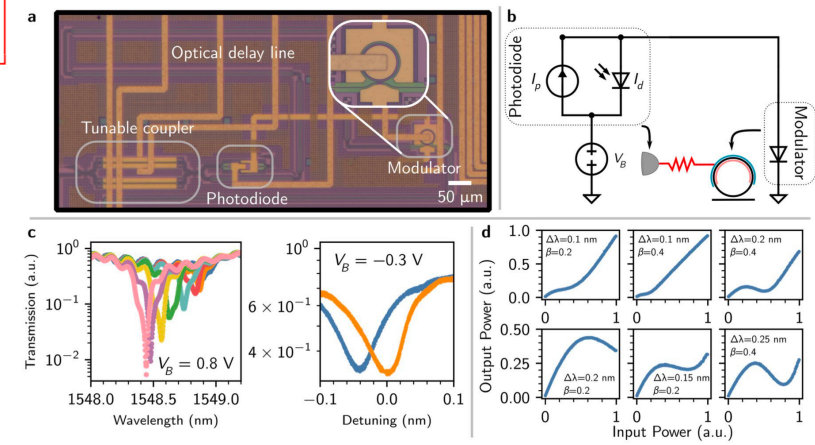
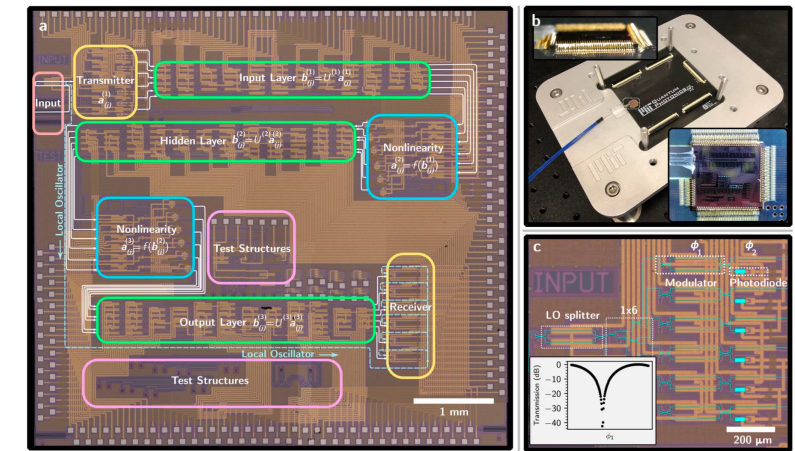
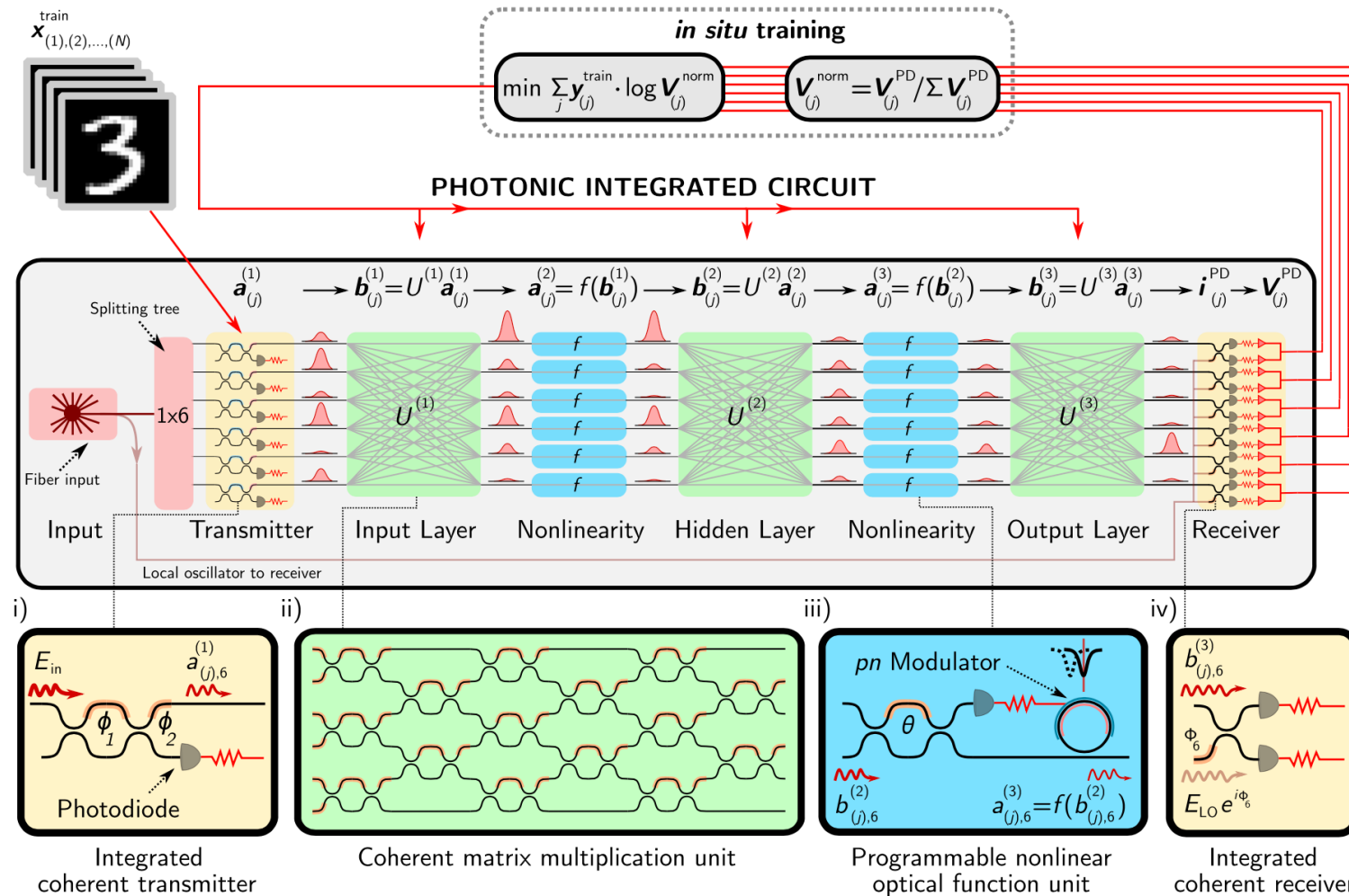


A. Sludds, *Science* **378**, 270 (2022)

R. Hamerly, arXiv:2207.01777 (in review)



Single-Chip Photonic *Deep* NN



“But it’s not scalable!”

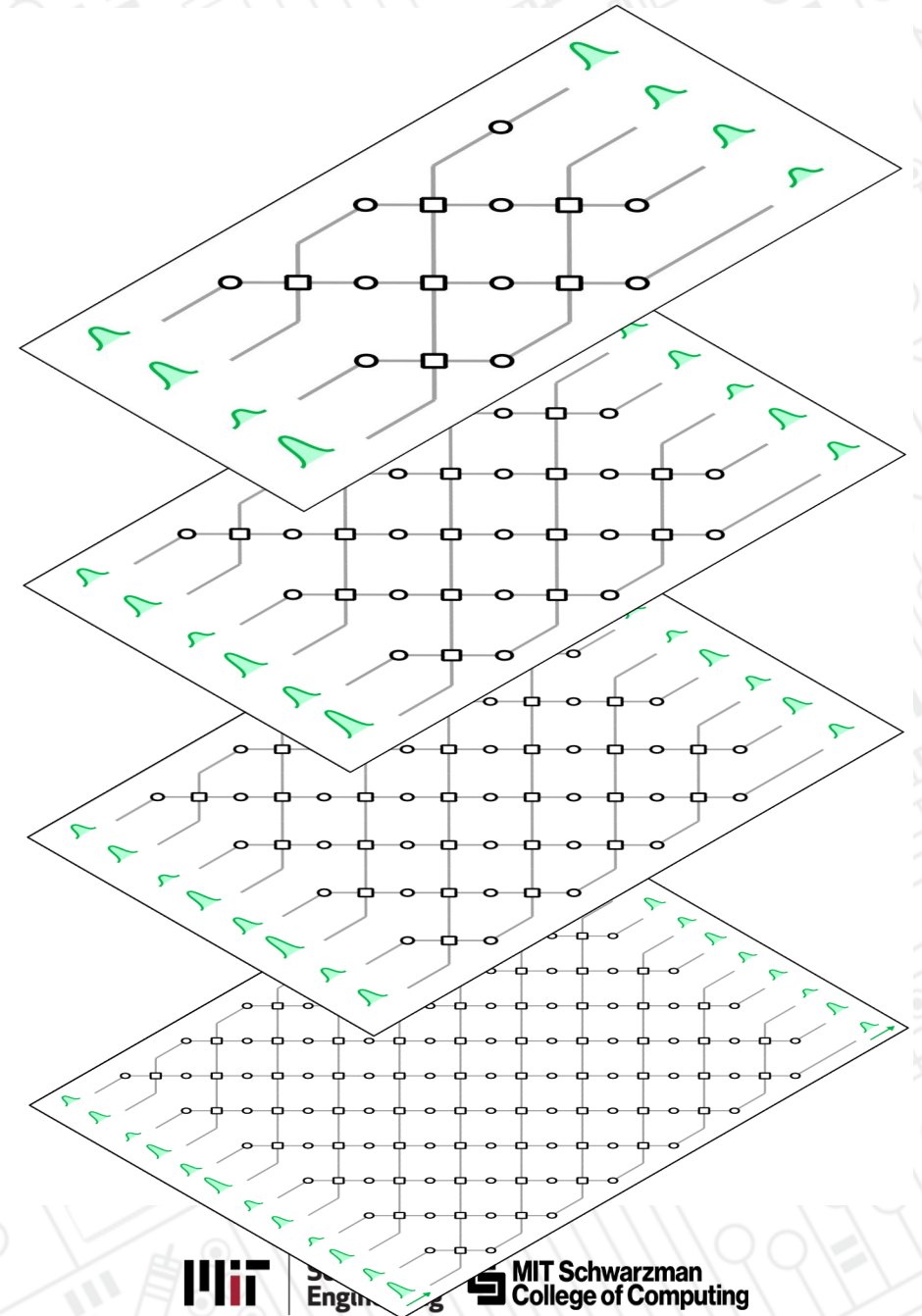
1. Too much chip area / loss / heater power / etc.

Counterpoint: 64x64 is possible with existing technology as of 2020

N. Harris, *OFC 2020*, p.W3A.3



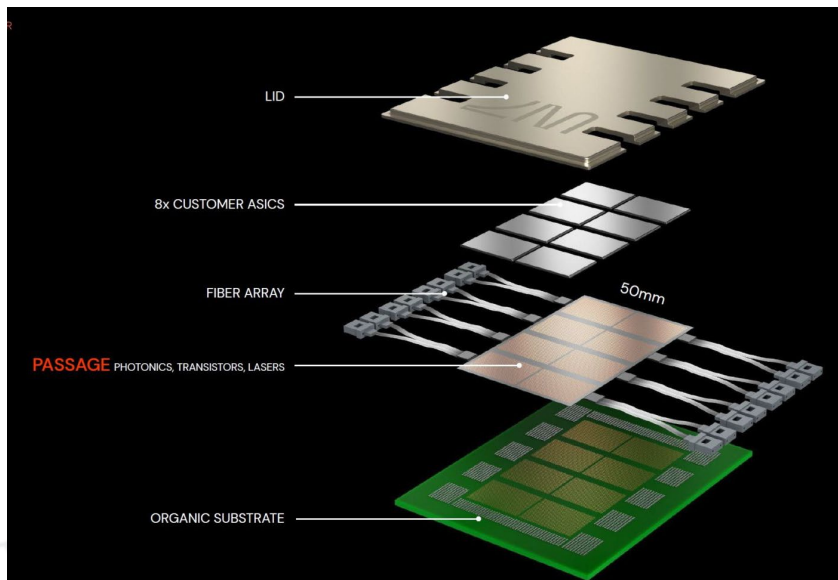
2. It’s analog, and ***analog computing doesn’t scale due to errors.***



Photonics for AI: compute vs. wires

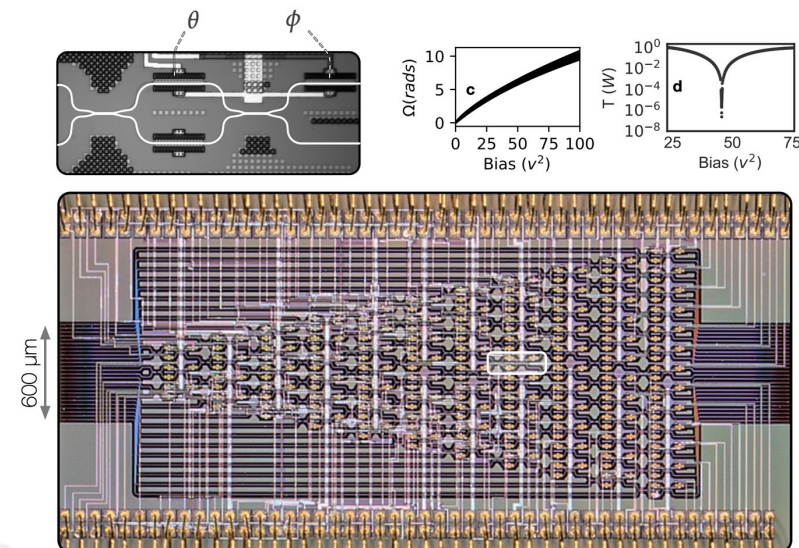
Interconnects

Digital (OOK, PAM4, BPSK, QAM)
Near-Term (i.e. *now*)
Impact: big
Error sensitivity: moderate



Computing

Analog (4-10 bit ENOB)
Long-term (5-10y)
Impact: **bigger** (if it works)
Error sensitivity: **severe**



Error Correction Techniques

Numerical Optimization

Throw everything into a nonlinear optimizer. Minimize norm $|U - U_0|$ (say).

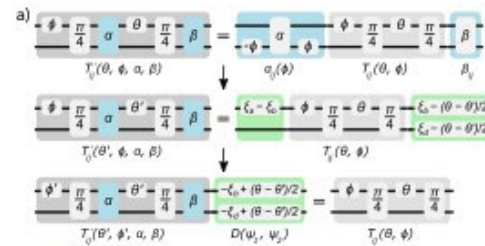
Burgwal et al., OE 25, 28236 (2017)
Mower et al., PRA 92, 032322 (2015)
Pai et al., PRA 11, 064044 (2019)

- (+) Works for arbitrary meshes
- (-) Slow
- (-) Needs pre-calibration (knowledge of MZI errors)

“Local” EC

Given errors, find (θ, ϕ) that fix them per-MZI.

Bandyopadhyay et al., *Optica* 8, 1247 (2021)

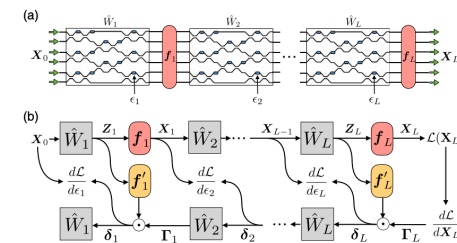


- (+) Arbitrary meshes
- (+) Fast
- (-) Needs pre-calibration

In-situ Training

Adjoint-related method. Send input & weight gradient, get gradients in internal detectors.

Hughes et al., *Optica* 5, 864 (2018)

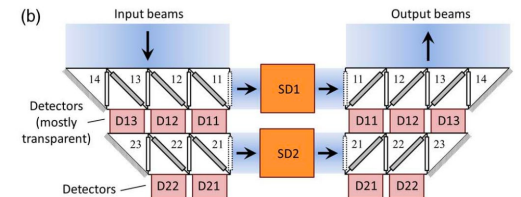


- (+) Arbitrary meshes
- (-) Slow, per-device training
- (+) No pre-calibration
- (-) Additional $O(N^2)$ detectors

Progressive Methods

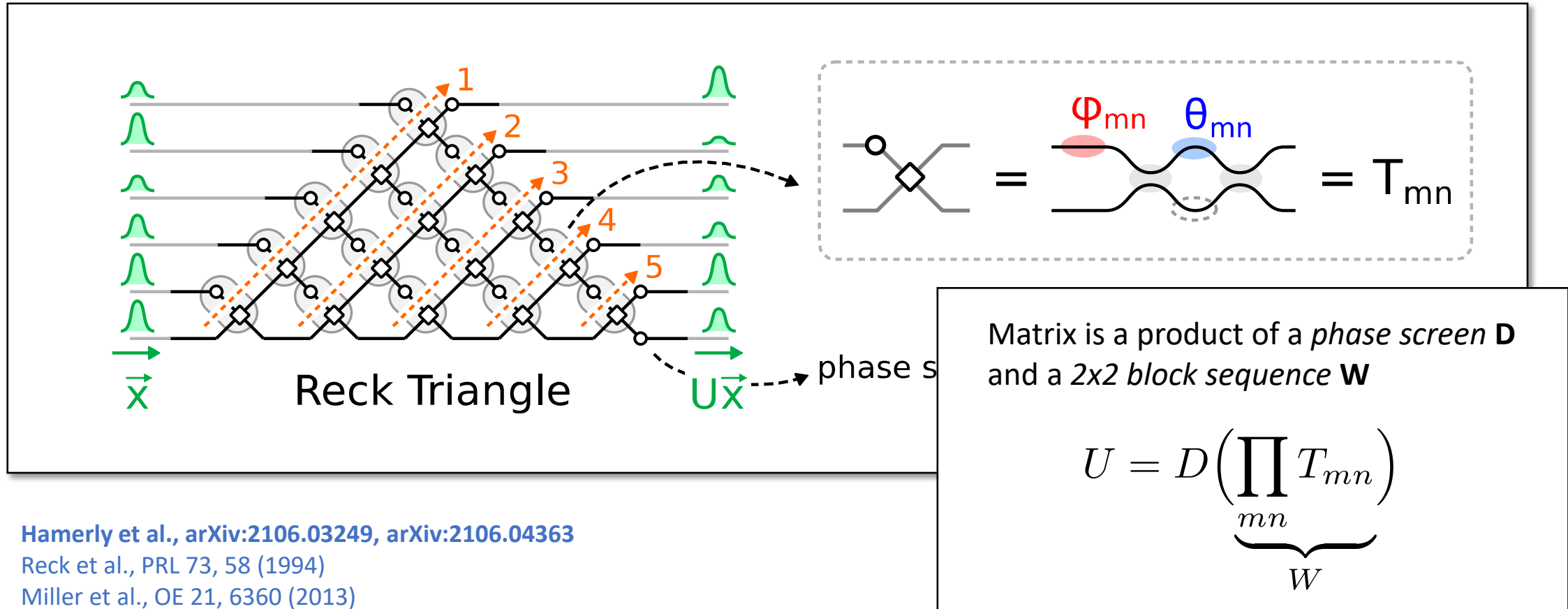
Adjust MZIs one-by-one to match certain I/O modes.

Miller et al., OE 21, 6360 (2013)
Miller et al., Phot Res 1, 1 (2013)
Miller et al., OE 25, 29233 (2017)



- (-) Only triangular meshes
- (+) Fast
- (+) No pre-calibration
- (-) Additional $O(N^2)$ detectors

Error Correction by Self-Configuration



Hamerly et al., arXiv:2106.03249, arXiv:2106.04363

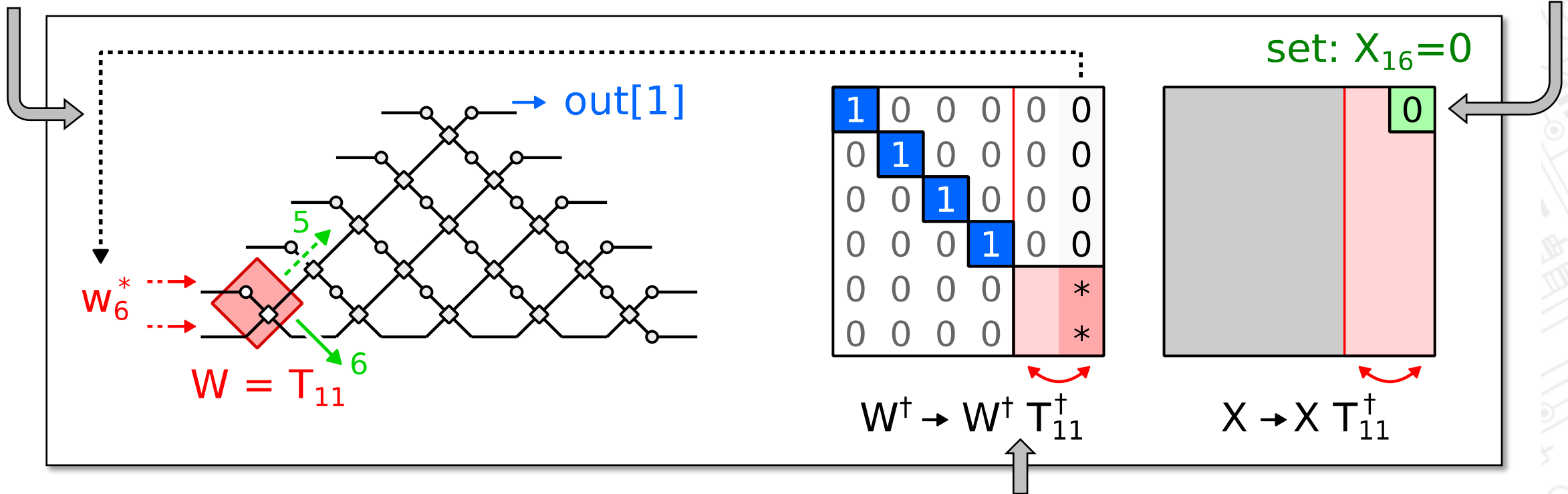
Reck et al., PRL 73, 58 (1994)

Miller et al., OE 21, 6360 (2013)

Error Correction by Self-Configuration

3. Inject w_6^* (column of $W^\dagger$) into device.
Adjust (θ, ϕ) to null power at $\text{out}[1]$

1. Find 2×2 rotation T_{11} that zeroes corner element X_{16} (in computer)



Hamerly et al., arXiv:2106.03249, arXiv:2106.04363

Reck et al., PRL 73, 58 (1994)

MIT AI Hardware Program

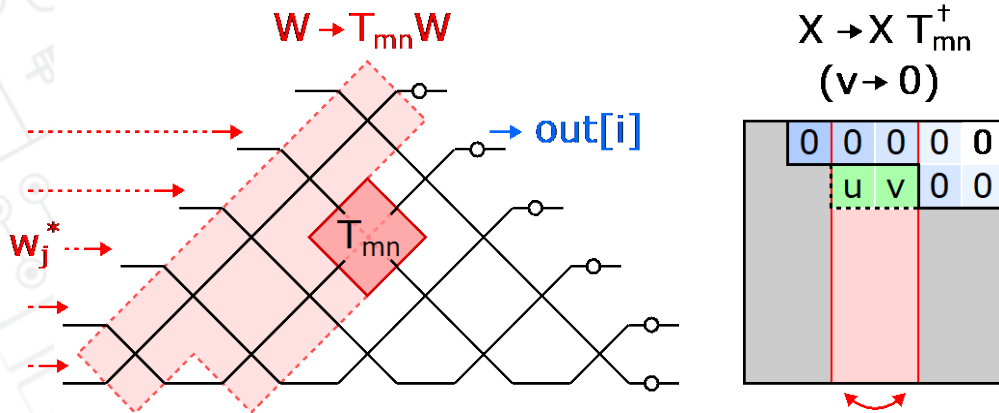
2. Update matrix W
(in computer)



School of Engineering

MIT Schwarzman College of Computing

EC w/3MZI



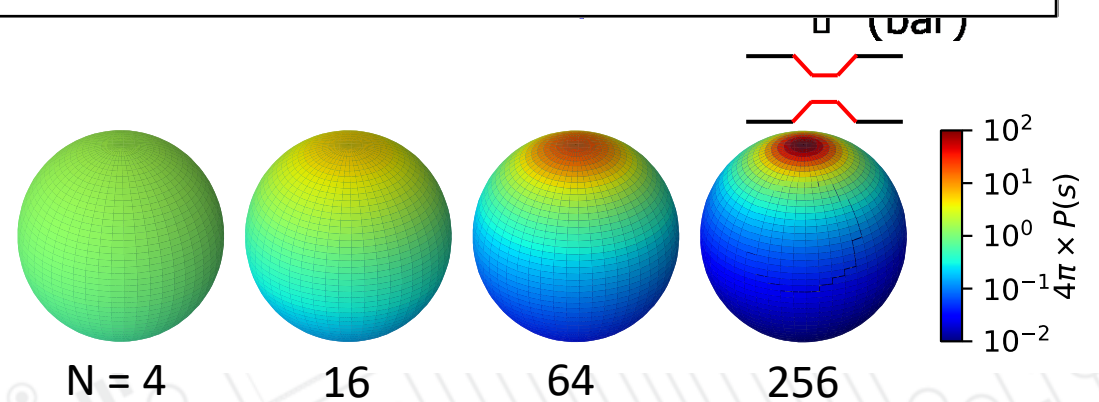
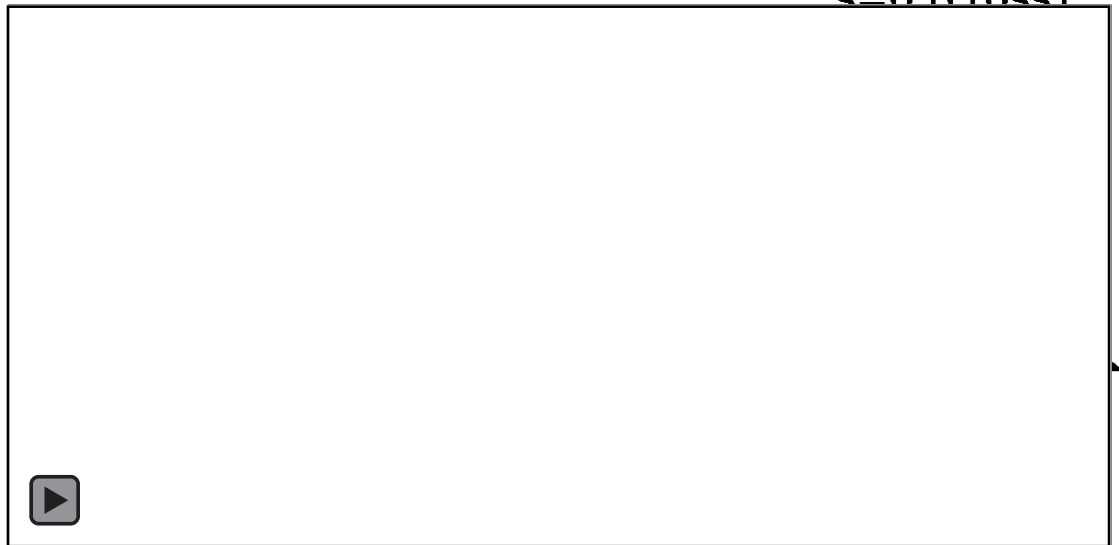
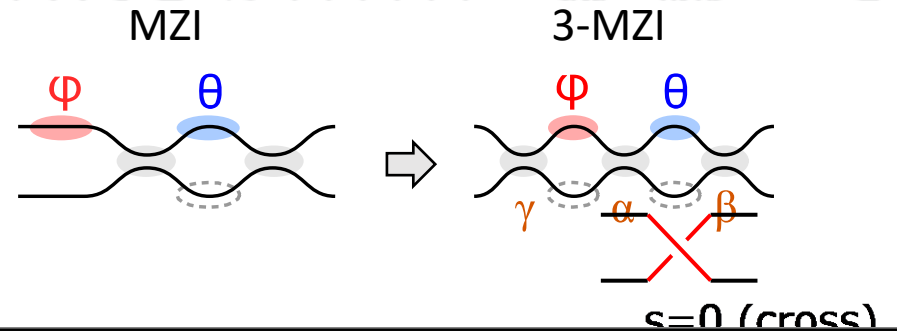
$$[u \quad v] T_{mn}^+ = \begin{bmatrix} * & 0 \end{bmatrix}$$

Below the equation, two small circuit diagrams illustrate the components u^* and v^* and their corresponding matrix elements $*$ and 0 .

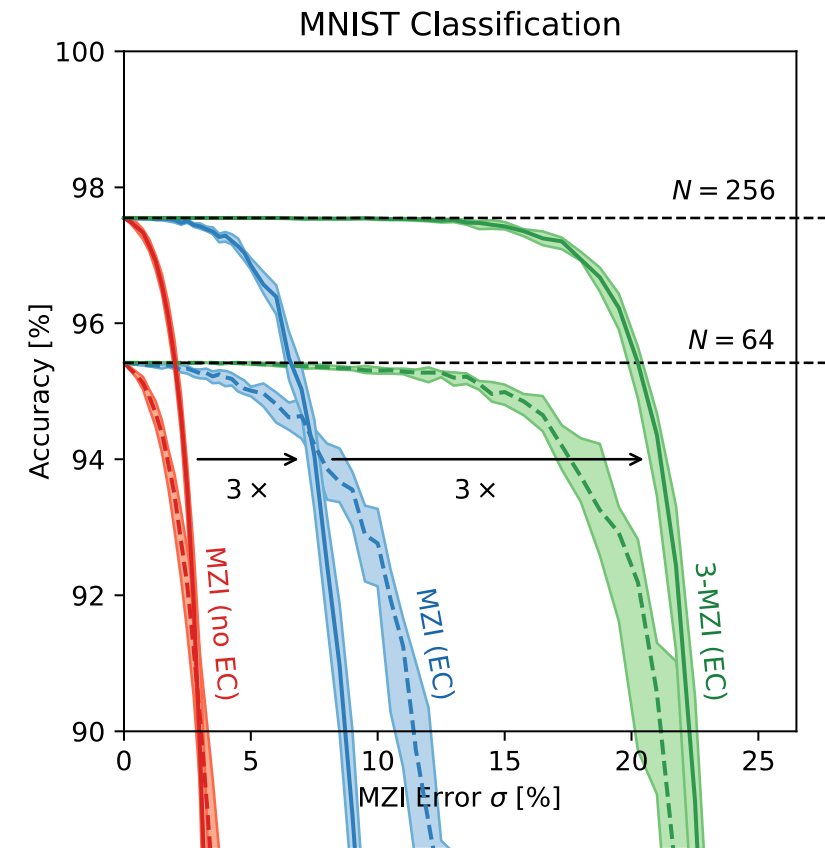
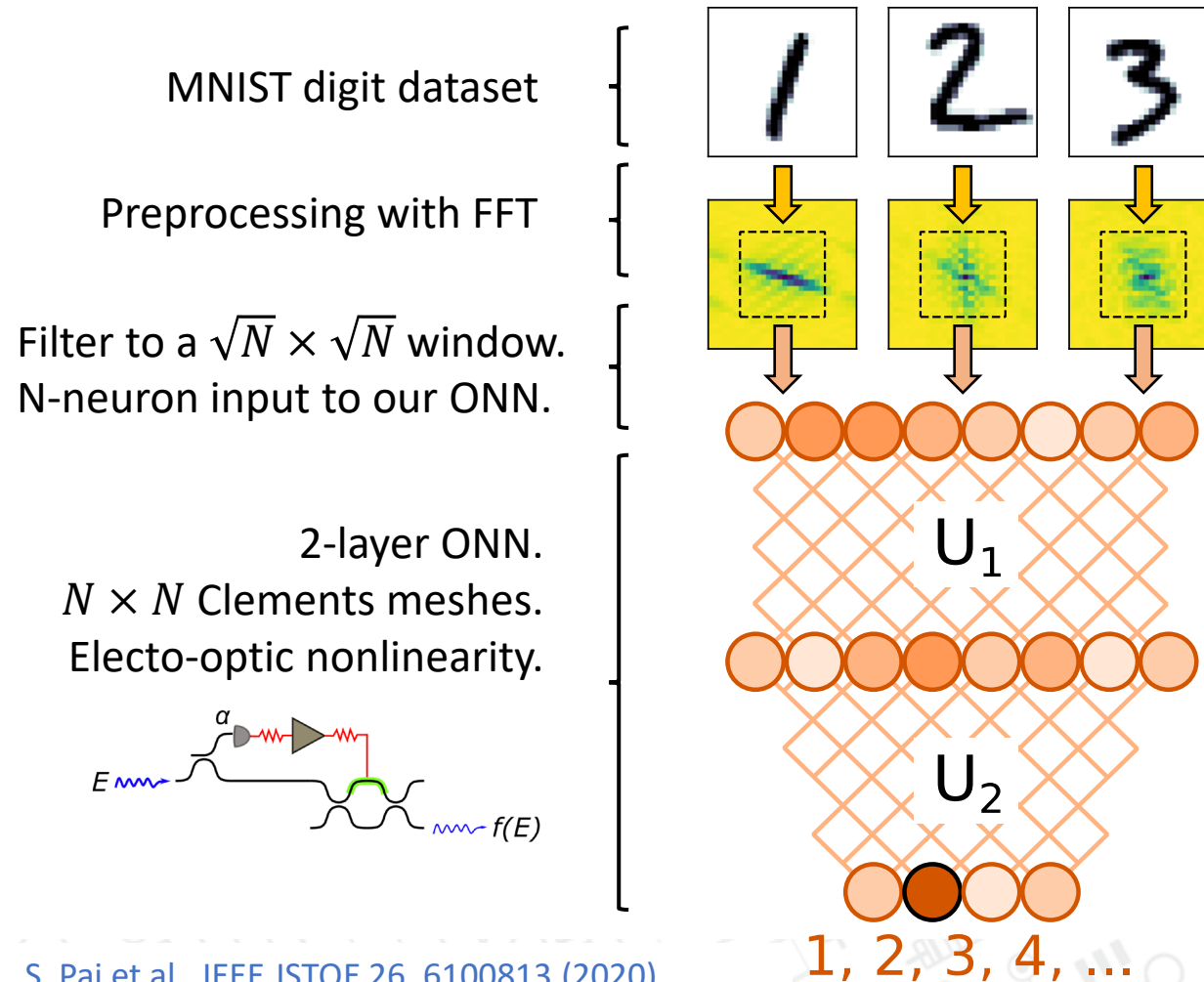
Perfect diagonalization requires correctly setting MZI splitting ratio s :

$$s \equiv \frac{T_{11}}{T_{12}} = \frac{u}{v} \equiv \hat{s}$$

Incorrect $s \rightarrow$ matrix X not diagonalized $\rightarrow$ errors in realized U



Toy Example: ONN Inference

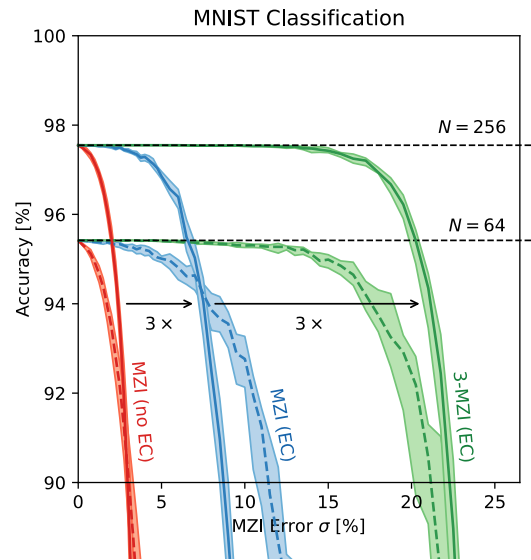
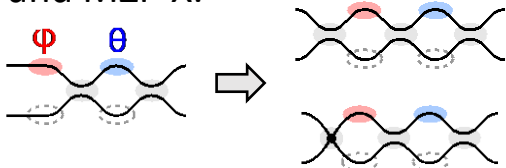


Based on: [S. Pai et al., IEEE JSTQE 26, 6100813 \(2020\)](#)
[I. Williamson et al., IEEE JSTQE 26, 7700412 \(2020\)](#)

Scalable Photonics $\Leftrightarrow$ Robustness

Results

- *Hardware Errors* limit the scalability of photonics
- Self-configuration: an effective *error correction* strategy.
- Introduced *asymptotically perfect* mesh designs: 3-MZI and MZI+X.



References / Code

S. Bandyopadhyay et al., *Optica* **8**, 1247 (2021)
R. Hamerly et al., *PRApp* **18**, 024018 & 024019 (2022)
R. Hamerly et al., *Nat. Comm.* **13**, 6831 (2022)
www.github.com/QPG-MIT/meshes

nature communications



Article

<https://doi.org/10.1038/s41467-022-34308-3>

Asymptotically fault-tolerant programmable photonics

Received: 24 May 2022

Accepted: 20 October 2022

Published online: 29 November 2022

Check for updates

Ryan Hamerly^{1,2}✉, Saumil Bandyopadhyay¹ & Dirk Englund¹

Component errors limit the scaling of programmable coherent photonic circuits. These errors arise because the standard tunable photonic coupler—the Mach-Zehnder interferometer (MZI)—cannot be perfectly programmed to the cross state. Here, we introduce two modified circuit architectures that overcome this limitation: (1) a 3-splitter MZI mesh for generic errors, and (2) a broadband MZI+Crossing design for correlated errors. Because these designs allow for perfect realization of the cross state, the matrix fidelity no longer degrades with increased mesh size, allowing scaling to arbitrarily large meshes. The proposed architectures support progressive self-configuration, are more compact than previous MZI-doubling schemes, and do not require additional phase shifters. This removes a key limitation to the development of very-large-scale programmable photonic circuits.

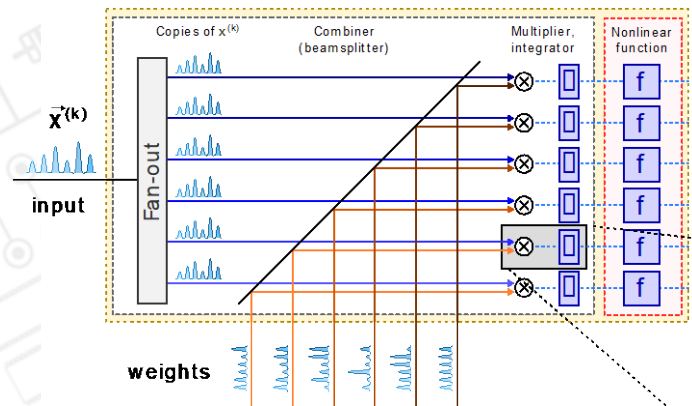
Large-scale programmable photonic circuits are opening up radical new possibilities for optics. Of key importance in many devices is the universal multiport interferometer, which functions as an $N \times N$ reconfigurable feedforward linear circuit. This device, typically constructed with a compact mesh of Mach-Zehnder interferometers (MZIs, Fig. 1a, b)^{1,2}, is widely employed in applications ranging from spatially multiplexed optical communications to machine learning and

resulting effects on chip area (particularly on emerging high-speed platforms where phase shifters have a large footprint^{26,27}), waveguide length (which affects insertion loss and latency²⁸), and electronic complexity (number of pads, traces, DACs/drivers, etc.) make this option unappealing.

In this paper, we propose two mesh architectures that achieve the same perfect scaling without significant added complexity: a 3-splitter

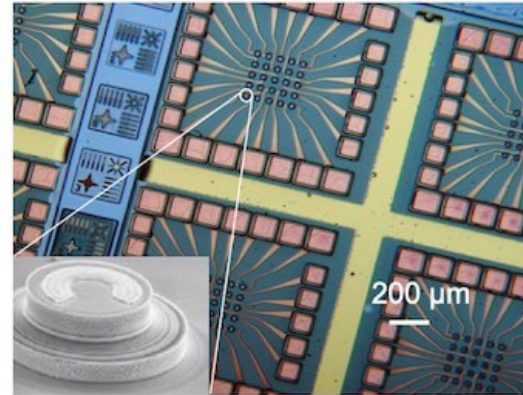
Perspective: MIT QPG / NTT Projects

Coherent Detection



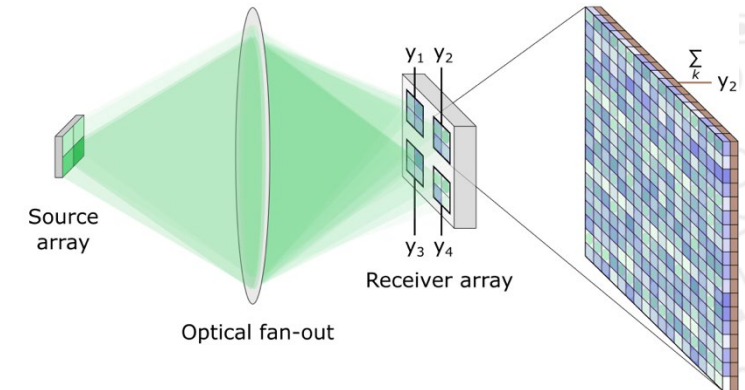
R. Hamerly, *PRX* **9**, 023012 (2019)

VCSEL Arrays



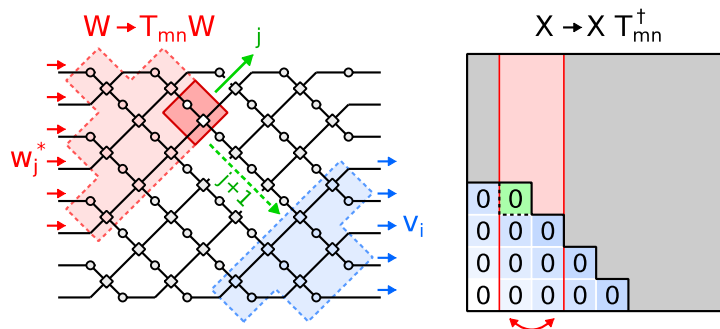
Z. Chen, arXiv:2207.05329 **(12438-18, Tue 9:35)**

Programmable Fan-out



L. Bernstein, arXiv:2205.09103

Photonic Error Correction

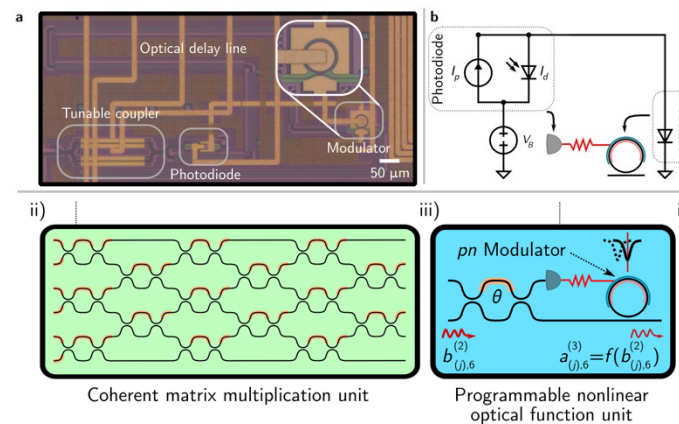


S. Bandyopadhyay, *Optica* **8**, 1247 (2021)

R. Hamerly, *Nat. Comm.* **13**, 6831 (2022)

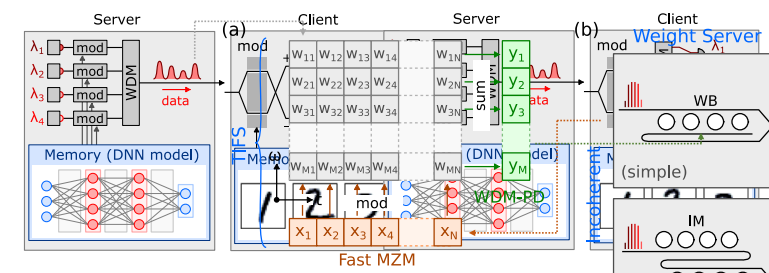
J. Basani, *Nanoph.* 2023 [2209.03335]

3-Layer PNP w/OEO Nonlinearity



S. Bandyopadhyay, arXiv:2208.01623

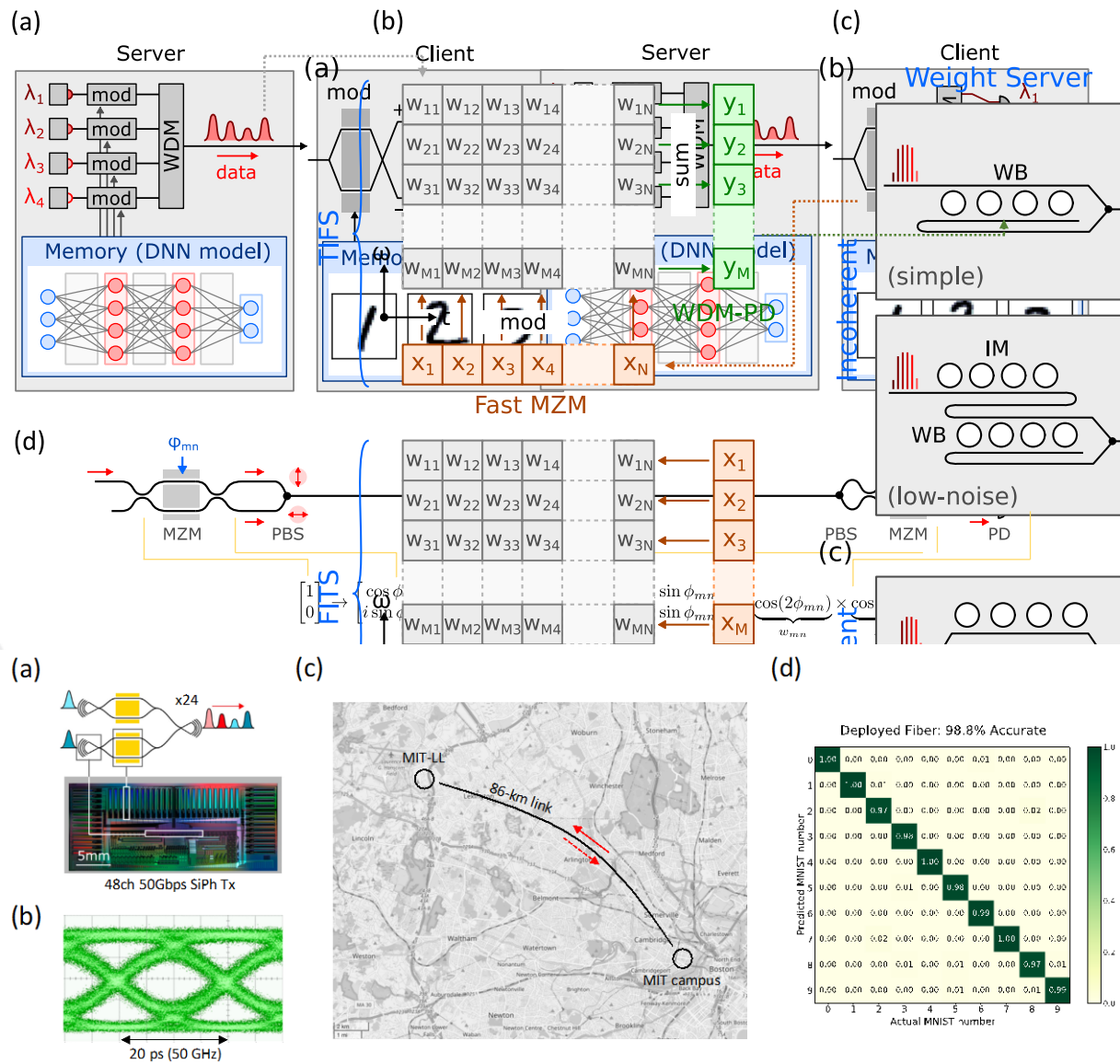
Netcast: WDM Edge Computing



A. Sludds, *Science* **378**, 270 (2022)

R. Hamerly, arXiv:2207.01777 (in review)

Delocalized Photonic Edge ML



RESEARCH

COMPUTER NETWORKS

Delocalized photonic deep learning on the internet's edge

Alexander Sludds^{1*}, Saumil Bandyopadhyay¹, Zaijun Chen^{1†}, Zhizhen Zhong², Jared Cochrane^{1,3}, Liane Bernstein¹, Darius Bunandar^{1,†}, P. Ben Dixon³, Scott A. Hamilton³, Matthew Streshinsky^{4,§}, Ari Novack^{4,§}, Tom Baehr-Jones^{4,§}, Michael Hochberg^{4,§}, Manya Ghobadi², Ryan Hamerly^{1,5*}, Dirk Englund^{1*}

Advanced machine learning models are currently impossible to run on edge devices such as smart sensors and unmanned aerial vehicles owing to constraints on power, processing, and memory. We introduce an approach to machine learning inference based on delocalized analog processing across networks. In this approach, named Netcast, cloud-based “smart transceivers” stream weight data to edge devices, enabling ultraefficient photonic inference. We demonstrate image recognition at ultralow optical energy of 40 attojoules per multiply (<1 photon per multiply) at 98.8% (93%) classification accuracy. We reproduce this performance in a Boston-area field trial over 86 kilometers of deployed optical fiber, wavelength multiplexed over 3 terahertz of optical bandwidth. Netcast allows milliwatt-class edge devices with minimal memory and processing to compute at teraFLOPS rates reserved for high-power (>100 watts) cloud computers.

Advances in deep neural networks (DNNs) are transforming science and technology (1–4). However, the increasing computational demands of the most powerful DNNs limit deployment on low-power devices, such as smartphones and sensors—and this trend is accelerated by the simultaneous move toward Internet of Things (IoT) devices. Numerous efforts are underway to

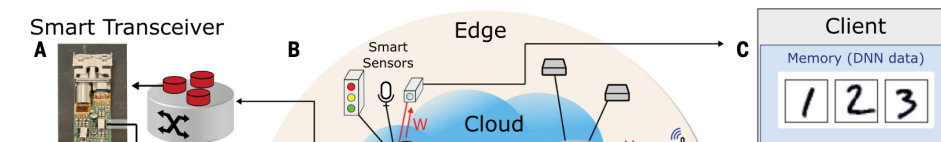
lower power consumption, but a fundamental bottleneck remains because of energy consumption in matrix algebra (5), even for analog approaches including neuromorphic (6), analog memory (7), and photonic meshes (8). In all these approaches, memory access and multiply-accumulate (MAC) functions remain a stubborn bottleneck near 1 pJ per MAC (5, 9–12). Edge devices typically use chip-scale sensors, occupy

millimeter-scale footprints, and consume milliwatts of power. Their small footprint and low power budget mean that performance is limited by the size, weight, and power (SWaP) of computing systems integrated on the device.

To make advanced DNNs at all feasible on low-power devices, industry has resorted to offloading computationally heavy DNN inference to cloud servers. For instance, a smart home device may send a voice query as a vector U to a cloud server, which returns the inference result V to the client (Fig. 1). This offloading architecture adds a ~200-ms latency to voice commands (13), which makes services such as self-driving impossible. Moreover, offloading poses security risks in both the edge and the cloud: Hacking of the communication

¹Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ²Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ³Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02421, USA. ⁴Nokia Corporation, New York, NY 10016, USA. ⁵Physics and Informatics Laboratories, NTT Research Inc., Sunnyvale, CA 94085, USA.

*Corresponding author. Email: asludds@mit.edu (A.S.); rhamerly@mit.edu (R.H.); englund@mit.edu (D.E.)
†Present address: Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA. §Present address: Lightmatter Inc., Boston, MA 02110, USA. §Present address: Luminous Computing Inc., Mountain View, CA 94041, USA.



A. Sludds, *Science* **378**, 270 (2022)

R. Hamerly, *SPIE Proc. ETAL*, p. 118041R (2021)

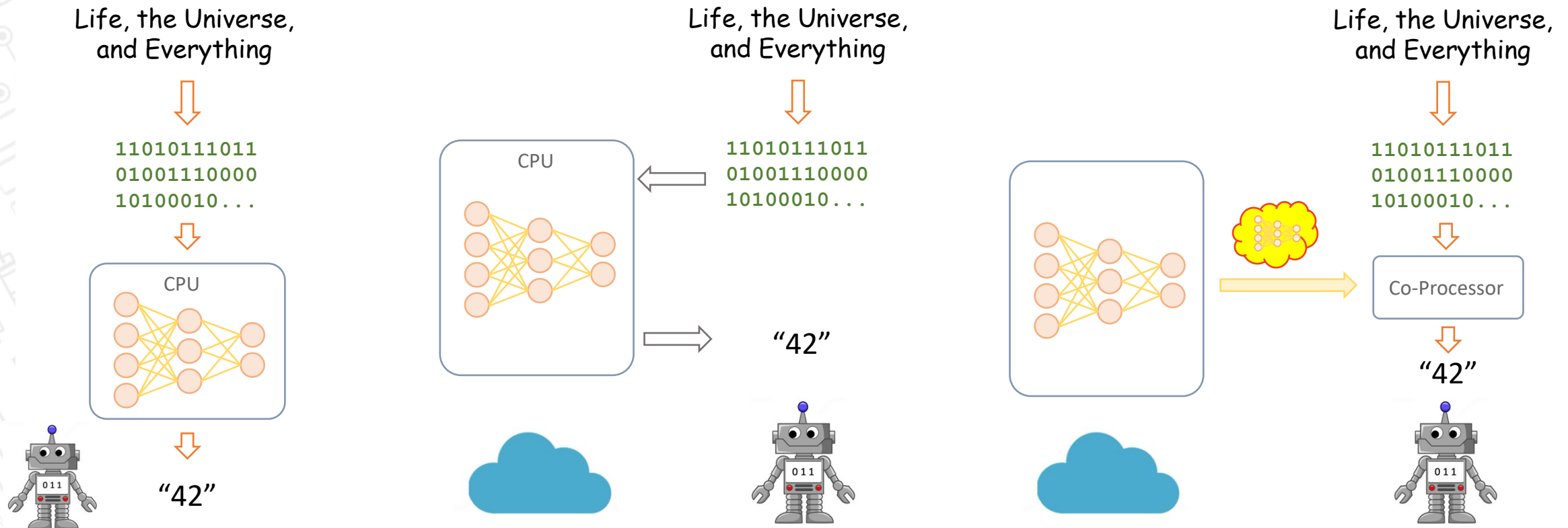
R. Hamerly, *arXiv:2207.01777 (in review)*

Netcast: Delocalized Computing

Desktop Computing

Cloud Computing

Magic-State Computing



☺ Latency, Privacy
☹ High-Power (at edge)

 MIT AI Hardware Program

☺ Low-Power (at edge)
☹ Latency, Privacy
☹ Network

☺ Low-Power (at edge)
☺ Latency, Privacy
☺ And more...



an
puting

Magic States & Teleportation (QIP)

Types of Teleportation

State TP: **transmit qubit** from A to B

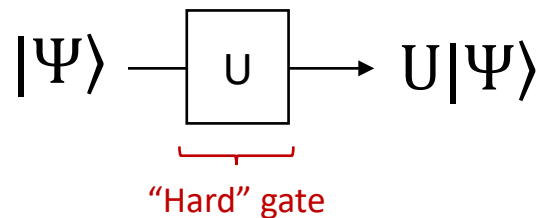
using a **Bell pair**

and **classical feedforward**

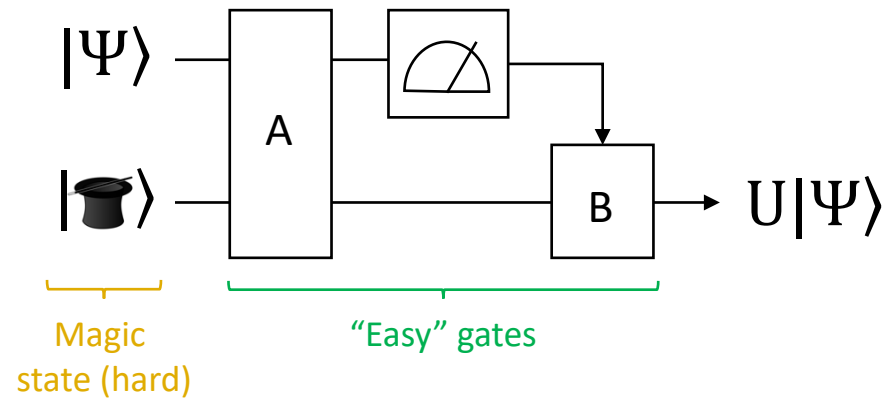
Gate TP: **perform “hard” gate** $|\Psi\rangle \rightarrow U|\Psi\rangle$

using a **magic state**

and **“easy” gates plus feedforward**.



=



Notes:

- Still need to do the “hard” task (compute $|\Psi\rangle$ or get magic state)
- But teleportation **reorders computation**, shifts burden to state generation, making computer more flexible:
 - Allows non-deterministic protocols (KLM, magic state distillation)
 - Facilitates fault-tolerance w/high QEC threshold
 - Opens up new paradigms, e.g. cluster-state MBQC

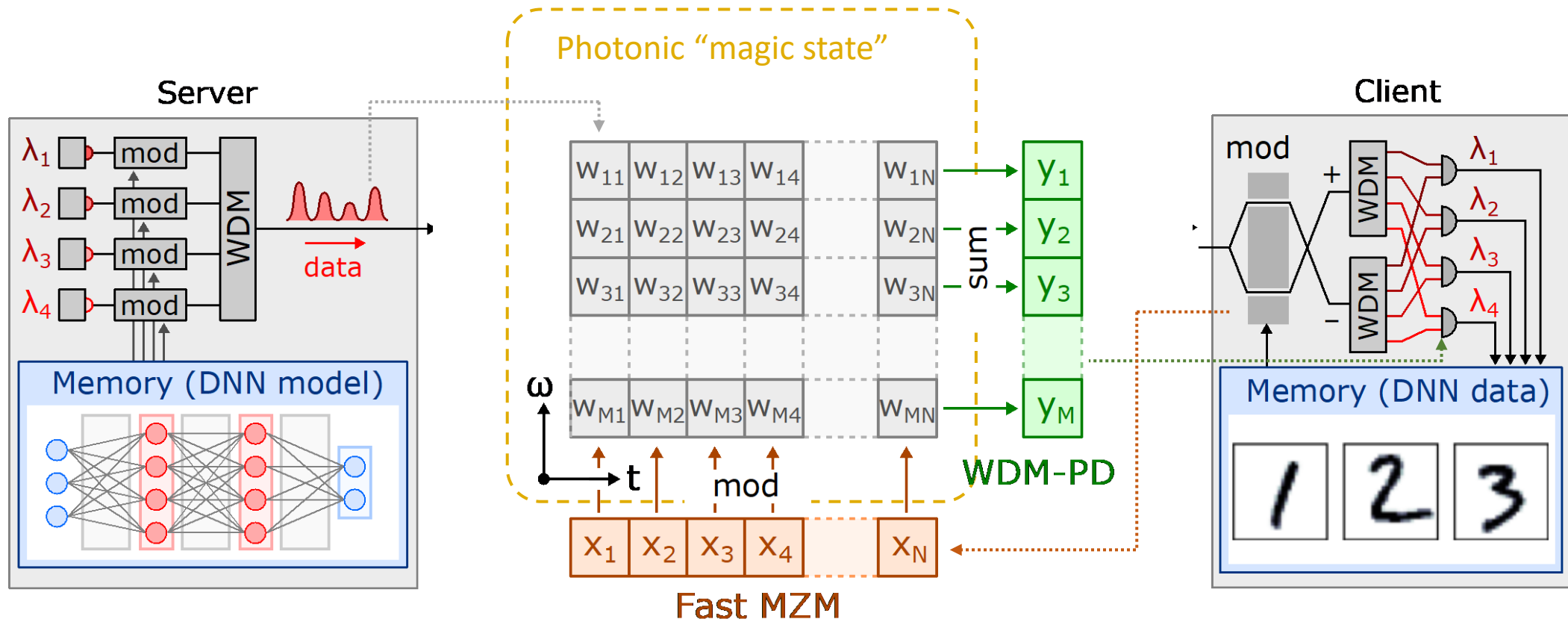
Examples:

- | | | |
|--------------------|---------------|------------------|
| • T gates | → magic state | + Clifford gates |
| • Cubic phase gate | → magic state | + Gaussian gates |
| • GKP-basis gates | → GKP state | + Gaussian gates |



Netcast: “Classical” Magic State DL

An ONN architecture based on (1) weight server + client, (2) output-stationary dataflow, (3) WDM + time multiplexing, (4) broadband modulation & integration

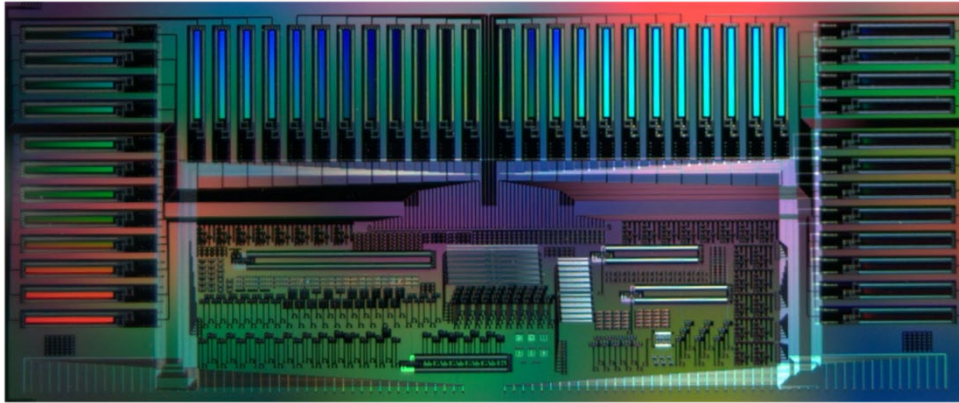


R. Hamerly, *Proc. SPIE OP21*, p.118041R (2021); *Proc. OECC+PSC* (2022, Best Paper)

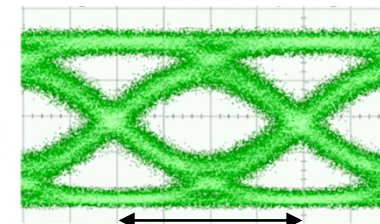
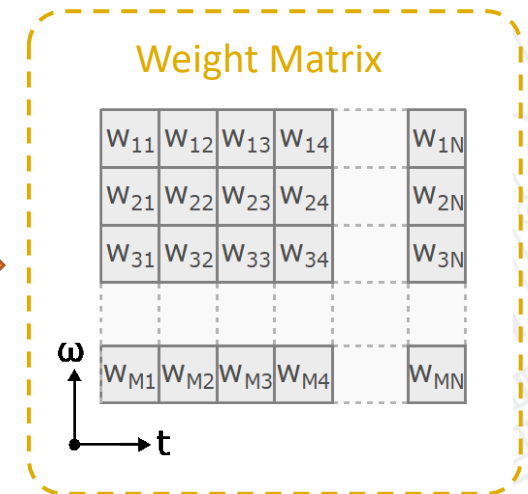
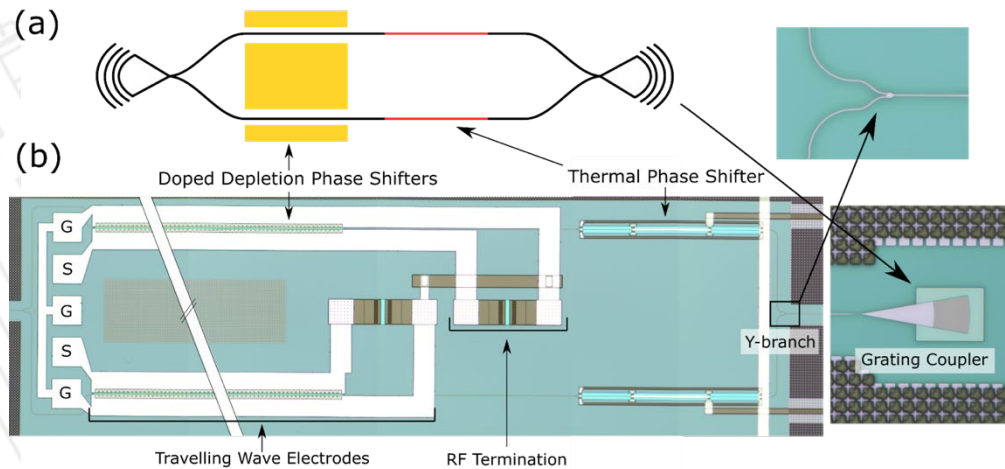
A. Sludds, *Science* **378**, 6617 (2022)

R. Hamerly, arXiv:2207.01777 (in review)

Netcast Experiment: Weight Server



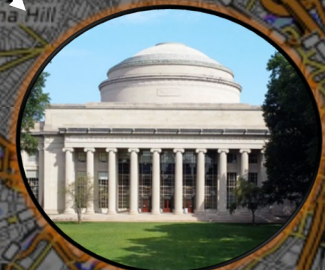
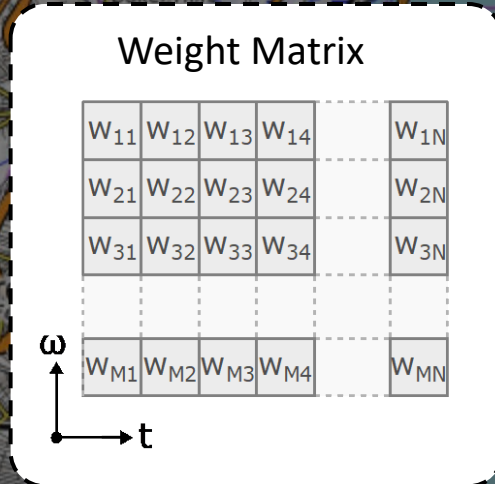
422mm² (half reticle)
48 carrier depletion modulators
50 Gbps each
2.4 Tbps total bandwidth
Foundry chip (OpSIS / IME)



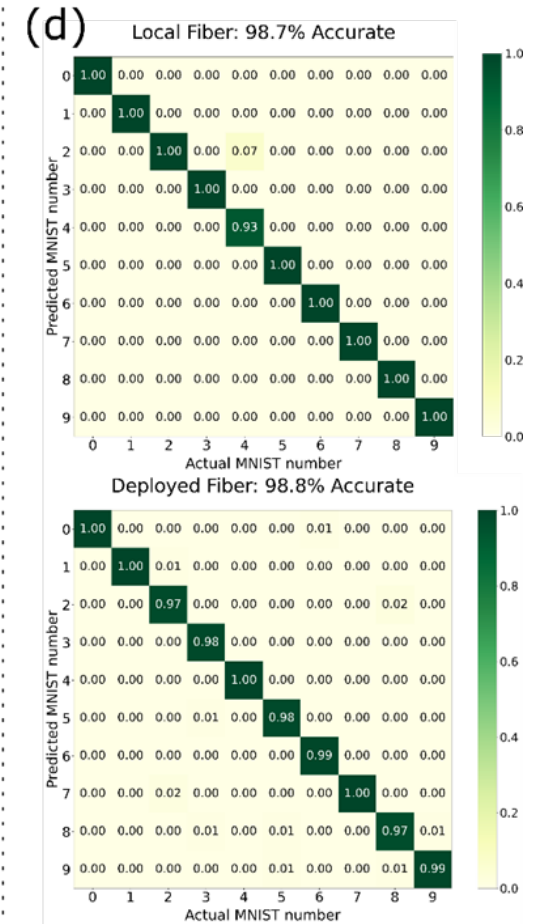
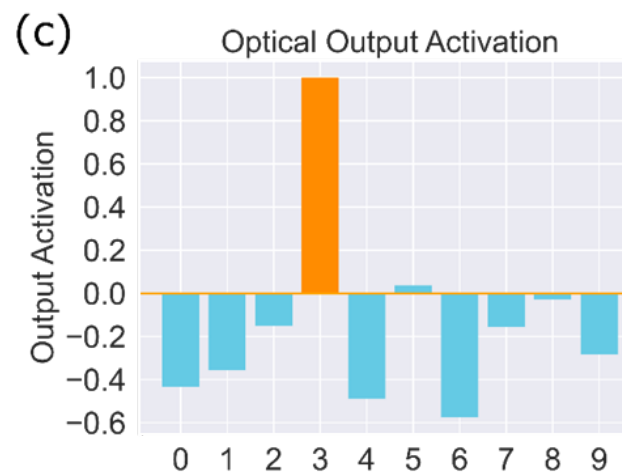
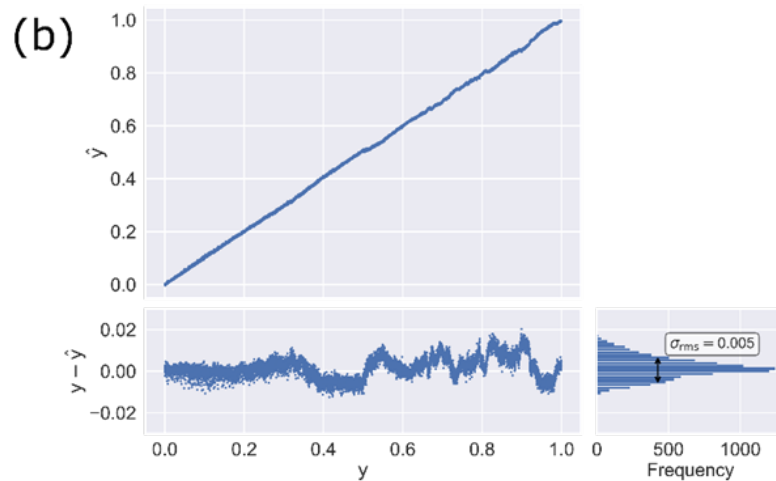
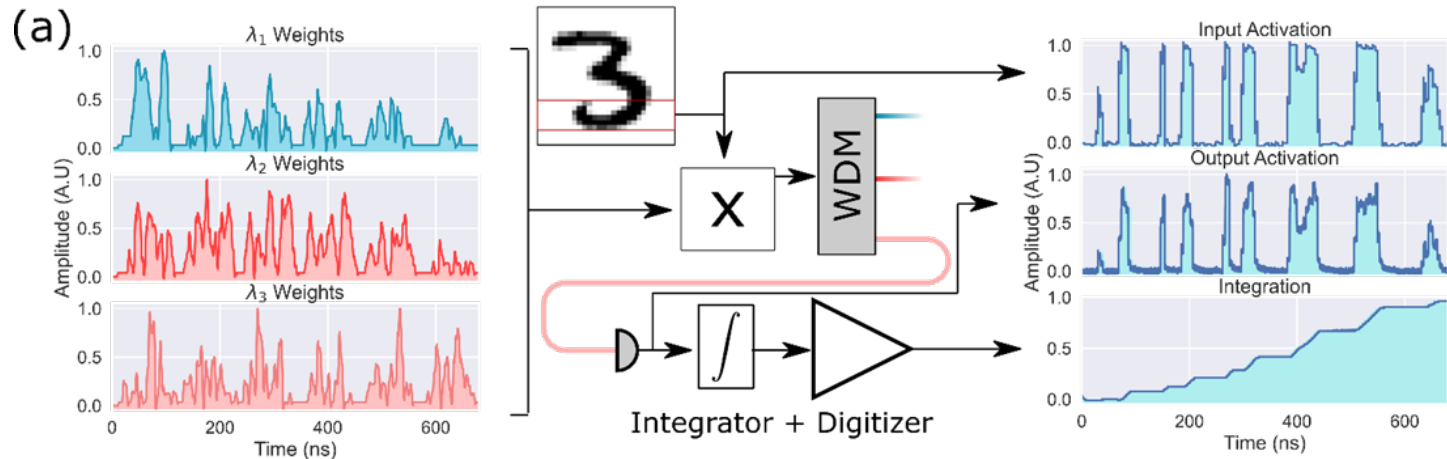
20 ps (50 GHz)



86 km deployed fiber
(round trip), ~30 dB loss



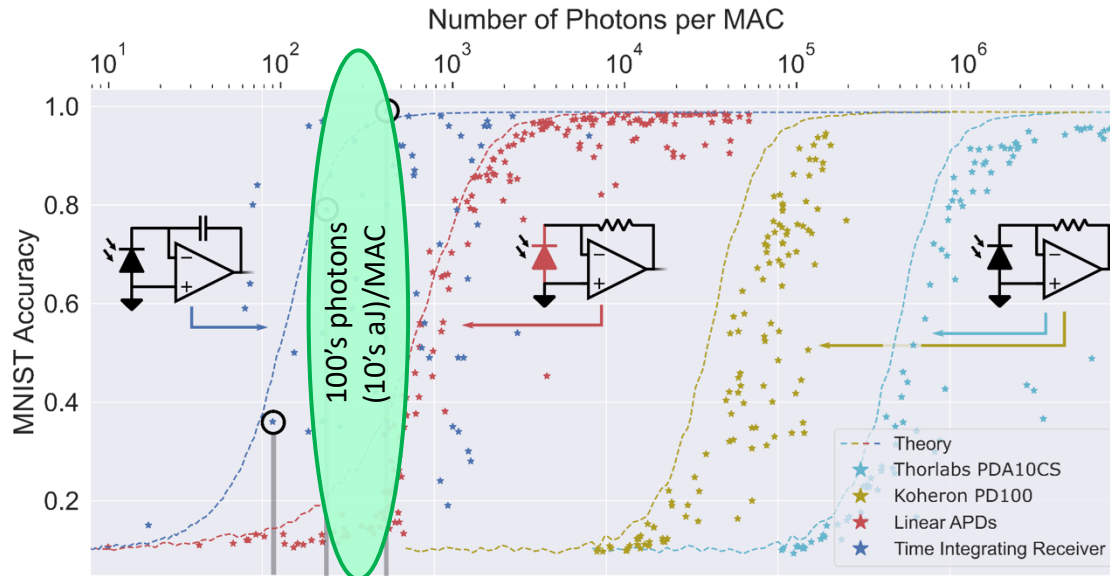
Dataflow & Accuracy



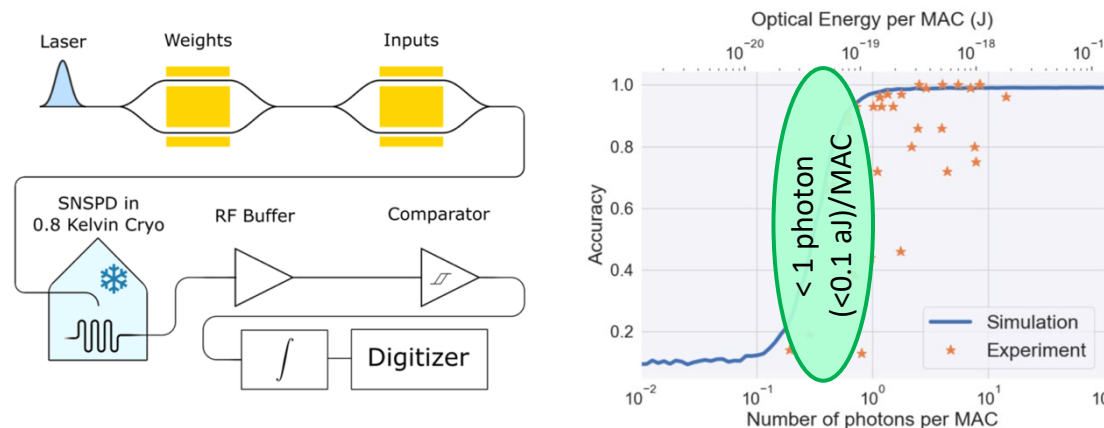
Very robust and accurate analog computation, 0.5% r.m.s. error (ENOB ~ 7 bits)
98.8% MNIST accuracy with 86 km deployed fiber with 3 THz bandwidth

Noise-Limited Energy Consumption

Thermal Noise-Limited

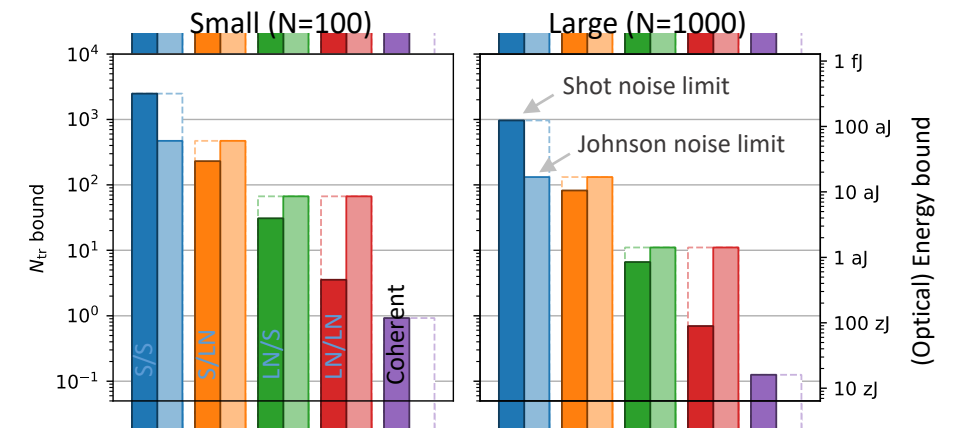


Shot Noise-Limited



Theoretical Limit

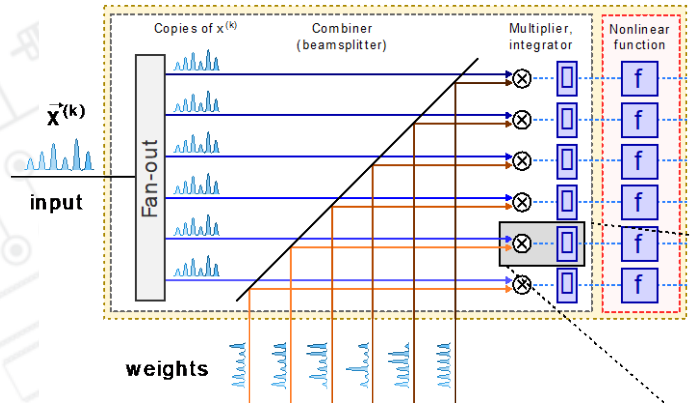
- Consistent with broad trend of experimental data
- Suggests coherent-detection version of NetCast could be far more (optical) energy-efficient
- Optical energy consumption most relevant in high-loss situations (e.g. free-space links)



R. Hamerly et al., arXiv:2207.01777 (submitted)

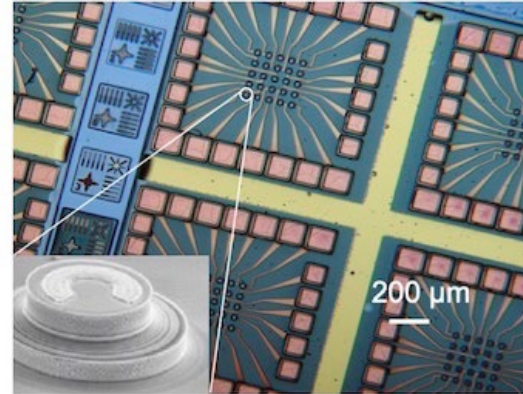
Perspective: MIT QPG / NTT Projects

Coherent Detection



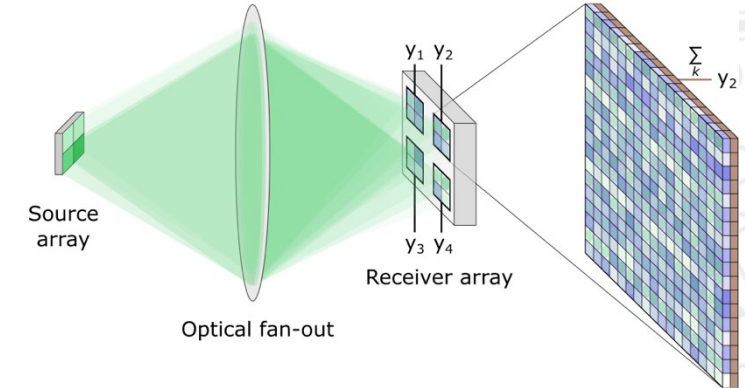
R. Hamerly, *PRX* **9**, 023012 (2019)

VCSEL Arrays



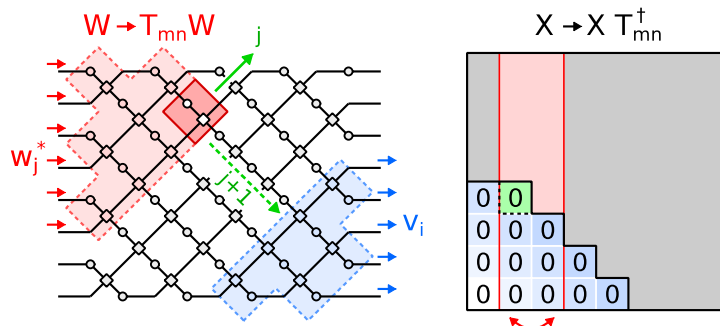
Z. Chen, arXiv:2207.05329 **(12438-18, Tue 9:35)**

Programmable Fan-out



L. Bernstein, arXiv:2205.09103

Photonic Error Correction

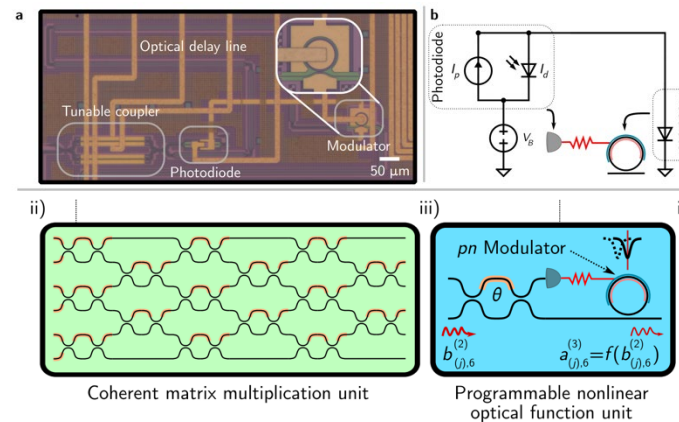


S. Bandyopadhyay, *Optica* **8**, 1247 (2021)

R. Hamerly, *Nat. Comm.* **13**, 6831 (2022)

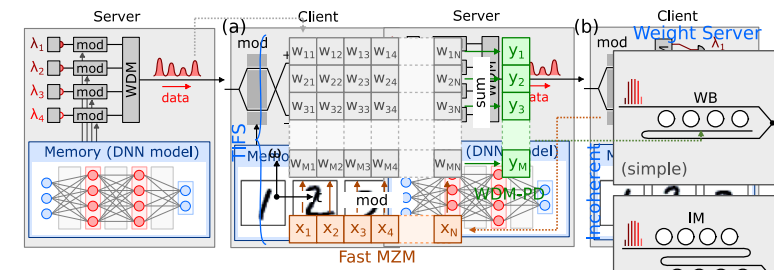
J. Basani, *Nanoph.* 2023 [2209.03335]

3-Layer PNP w/OEO Nonlinearity



S. Bandyopadhyay, arXiv:2208.01623

Netcast: WDM Edge Computing



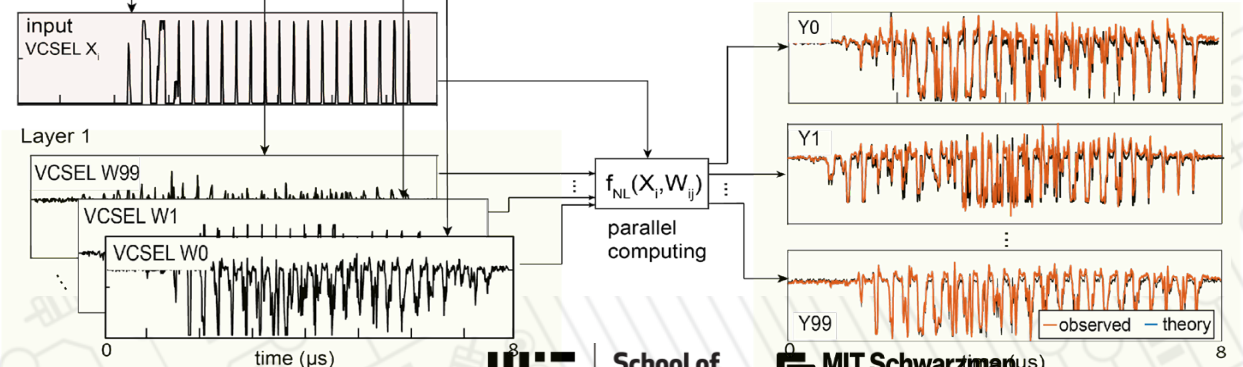
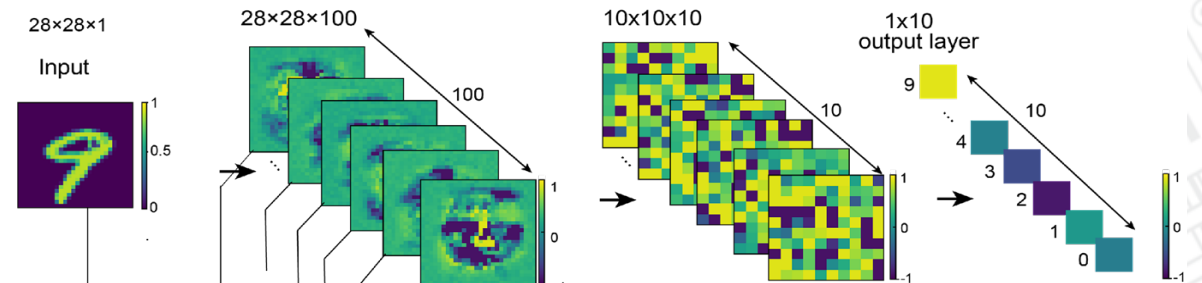
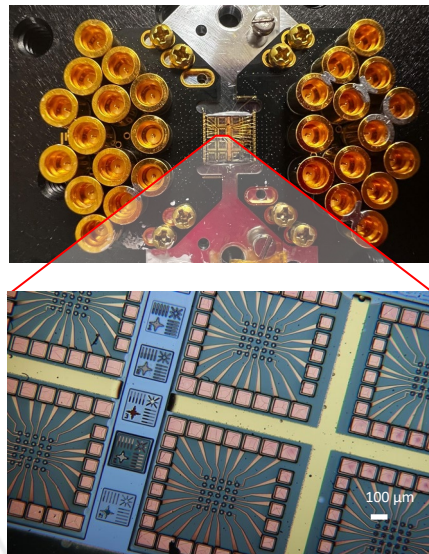
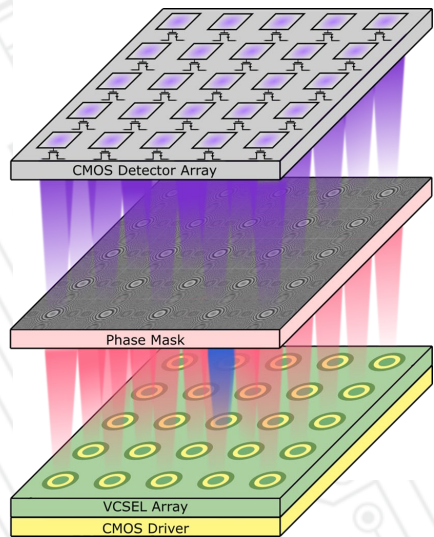
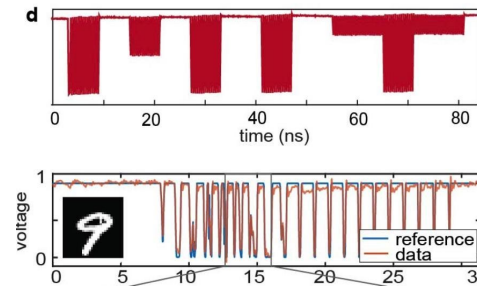
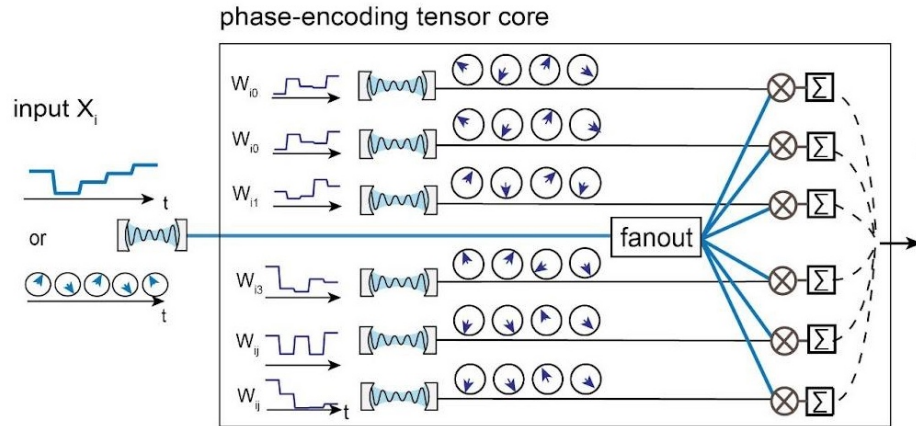
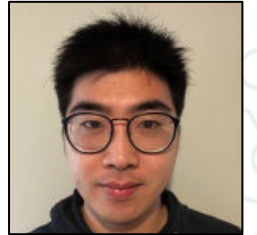
A. Sludds, *Science* **378**, 270 (2022)

R. Hamerly, arXiv:2207.01777 (in review)

Coherent VCSEL Network Computing

Z. Chen, arXiv:2207.05329,
Nat. Photonics (to appear)
 PW2023 (Best Paper)

→ Faculty at USC

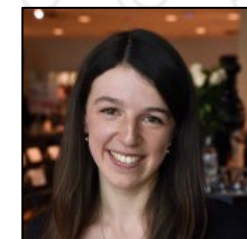


School of Engineering

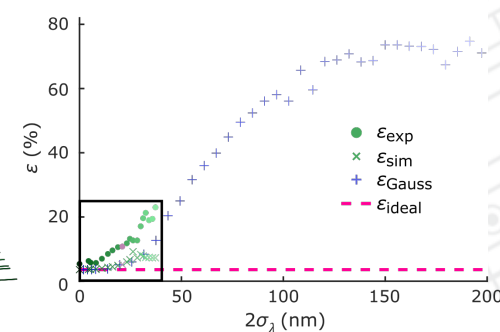
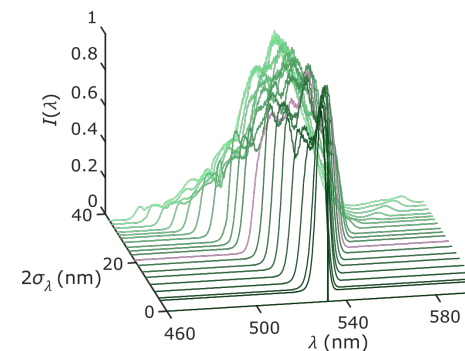
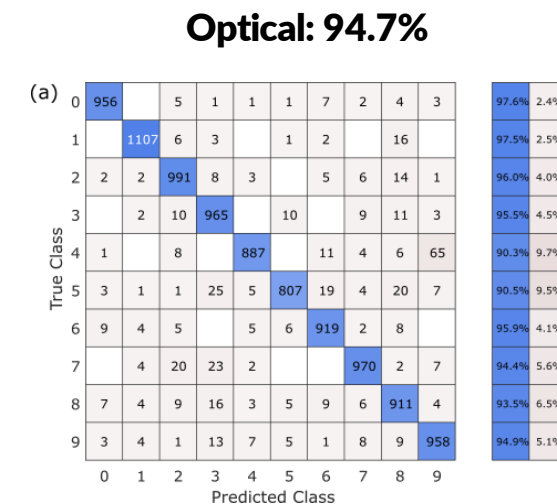
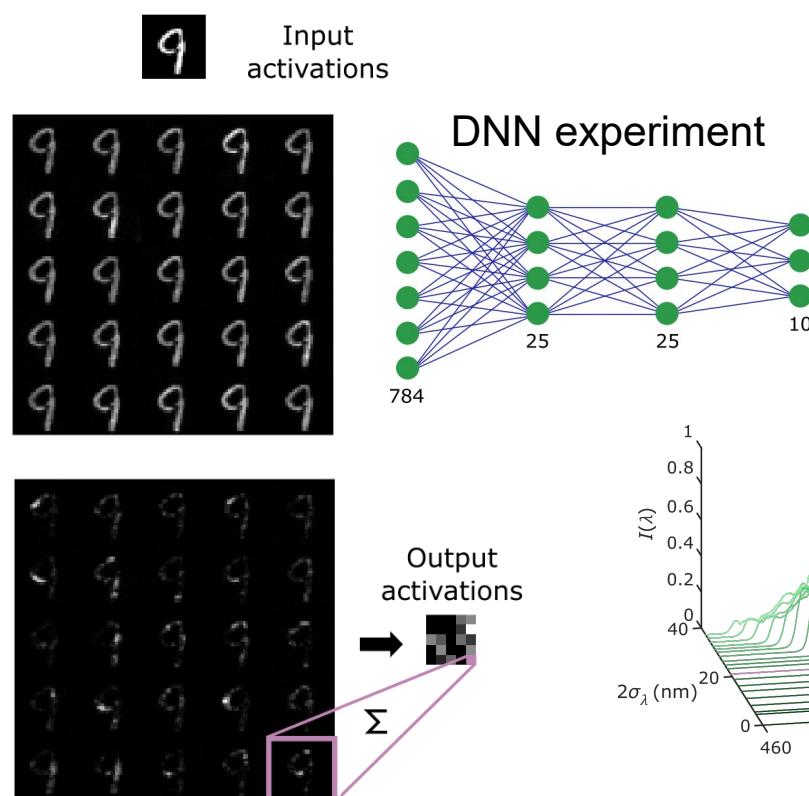
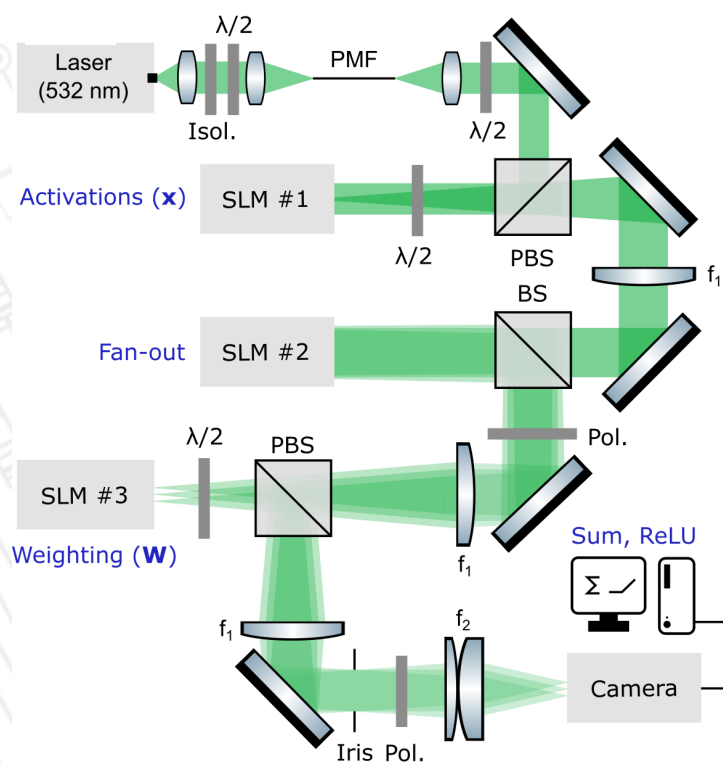
MIT Schwarzman College of Computing



MIT AI Hardware Program



Single-Shot Free-Space ONN



~20,000 MACs/0.02ps

~Exascale throughput!





Acknowledgements



Physics & Informatics Laboratories

Y. Yamamoto

Devices & Hardware

R. Hamerly, M. Jankowski, M-G., Suh



Algorithms & Applications

S. Kako, F. Khoyratee, E. Dormer

Neural Network

H. Tanaka, L.W., T.O., E.N.

Quantum Simulations

M. Fraser, Y. Inui, T. Chervy

Foundations & Theory

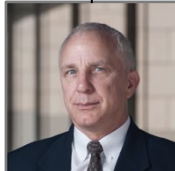
Y. Yamamoto, J.R., S.S., A.G.



D. Englund
(MIT)



W. Oliver
(MIT)



M. M. Fejer
(Stanford)



A. Safavi-Naeini
(Stanford)



K. Vahala
(Caltech)

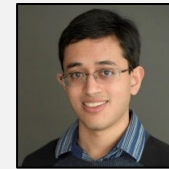


D. Englund

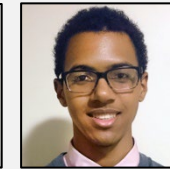
Quantum Photonics Lab (MIT)



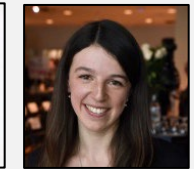
Alex -> LM



Saumil



Ronald



Liane



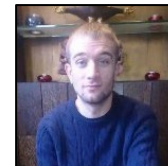
Zaijun -> USC

Chao Luan
Mustafa Yucel

Positions available:

PhD and postdoc at MIT

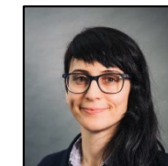
Summer intern with NTT PHI



E. Allen
(Bristol)



D. Miller
(Stanford)



M. Ghobadi
(MIT)



K. Jamshidi
(TU-Dresden)



V. Sze
(MIT)

...



MIT AI Hardware Program



School of
Engineering

MIT Schwarzman
College of Computing