Photonic AI Accelerators

Nikos Pleros

Department of Informatics, Aristotle University of Thessaloniki, Greece
Center for Interdisciplinary Research and Innovation, AUTh, Greece

AI computing and energy consumption

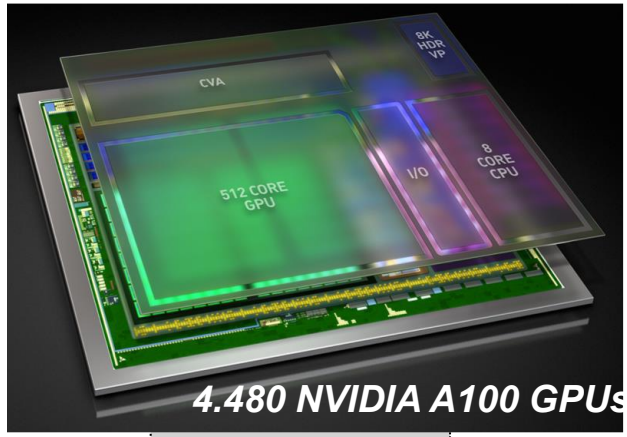
MEGATRON-TURING NLG

Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model

Shaden Smith[§,†], Mostofa Patwary[§,‡], Brandon Norick[†], Patrick LeGresley[‡], Samyam Rajbhandari[†], Jared Casper[‡], Zhun Liu[†], Shrimai Prabhumoye[‡], George Zerveas[*†], Vijay Korthikanti[‡], Elton Zhang[†], Rewon Child[‡], Reza Yazdani Aminabadi[†], Julie Bernauer[‡], Xia Song[†], Mohammad Shoeybi[‡], Yuxiong He[†], Michael Houston[‡], Saurabh Tiwary[†], and Bryan Catanzaro[‡]

NVIDIA Microsoft

arXiv, 2022



530B parameters & **10^{24} FLOPs** compute power

…for 11 days

400MWh

consumes energy that **powers a small city** (16k people) **for 1 year**

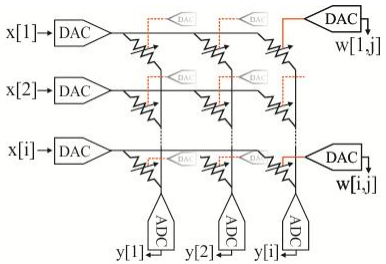
Escape through analog photonic MVMs

wavelength

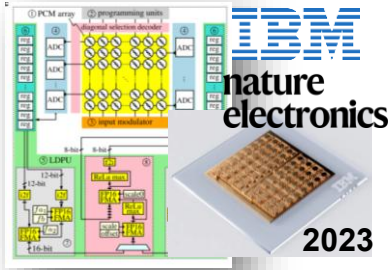
space

time

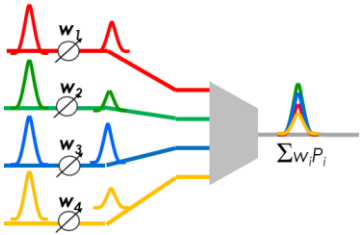
Neuromorphic electronics

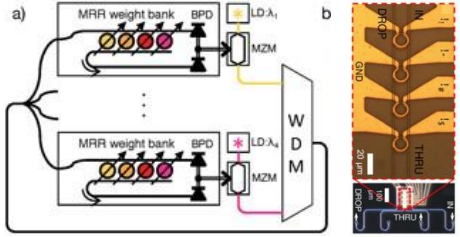


>5x more efficient: ~200 fJ/MAC

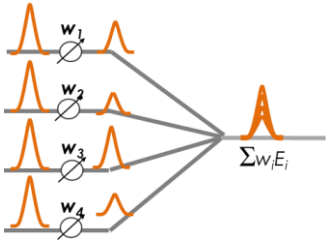


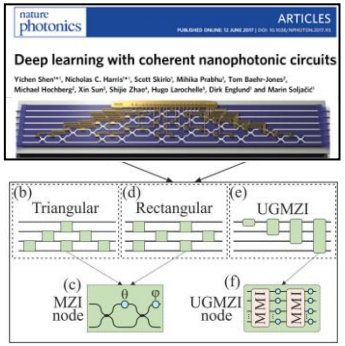
Incoherent photonic accelerators



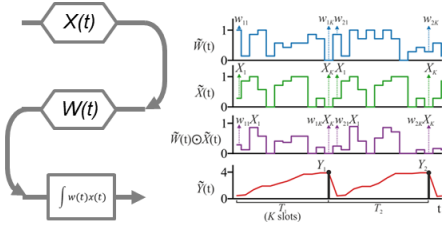


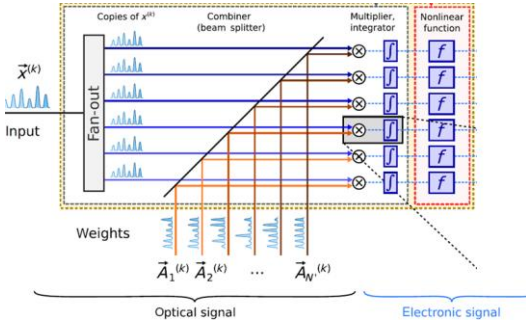
Coherent photonic accelerators



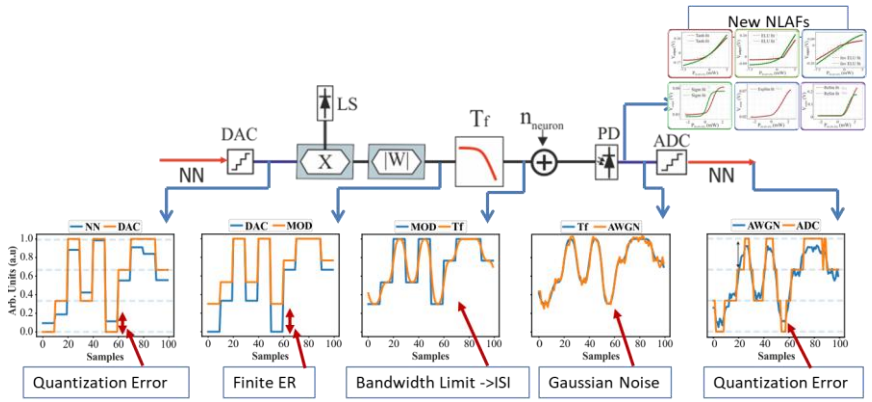


Time-multiplexed photonic accelerators





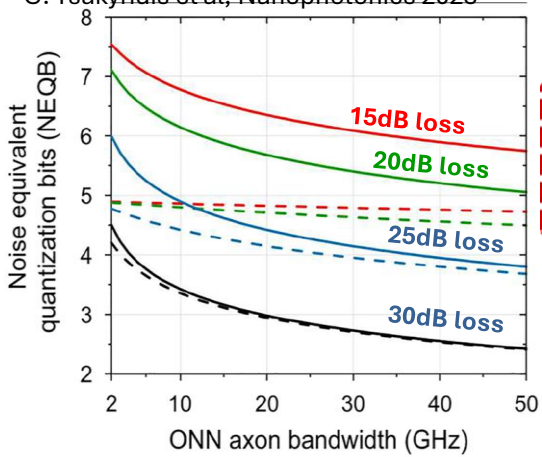
Lessons Learned

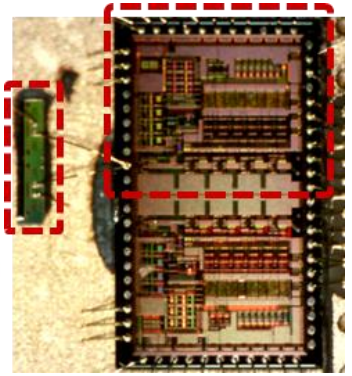


A. Tsakyridis et al, "Photonic Neural Network and Optics-informed Deep Learning Fundamentals", (Invited Tutorial) APL Photonics 2024

✓ **Optics-Informed Deep Learning**: embeds ER, BR, noise, BW etc

G. Tsakyridis et al, Nanophotonics 2023





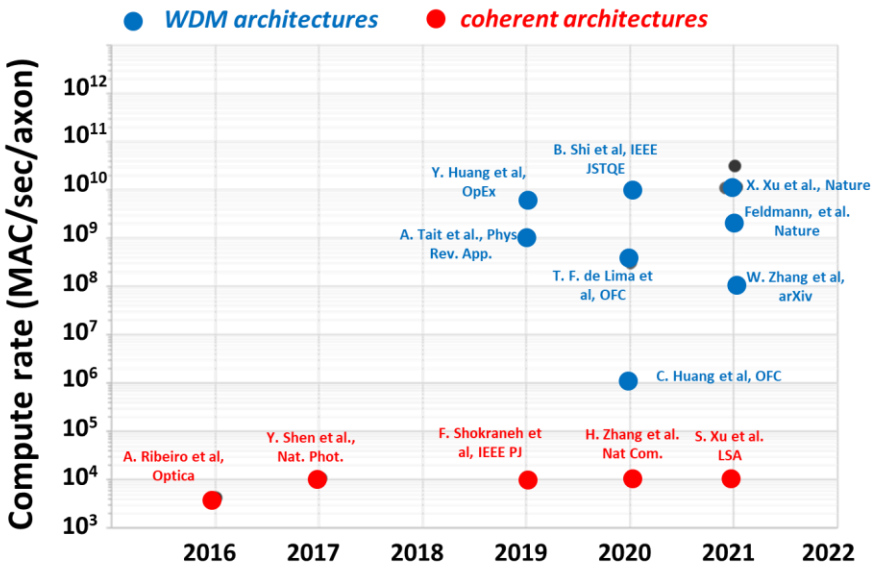
C. Pappas et al, JSTQE 2023
S. Kovaios et al, JLT 2024

✓ High-speed ADCs reduce bit-resolution and energy efficiency

Lesson #1:
Hardware/Software co-design

Lesson #2:
<GHz ADC via TDM PNNs

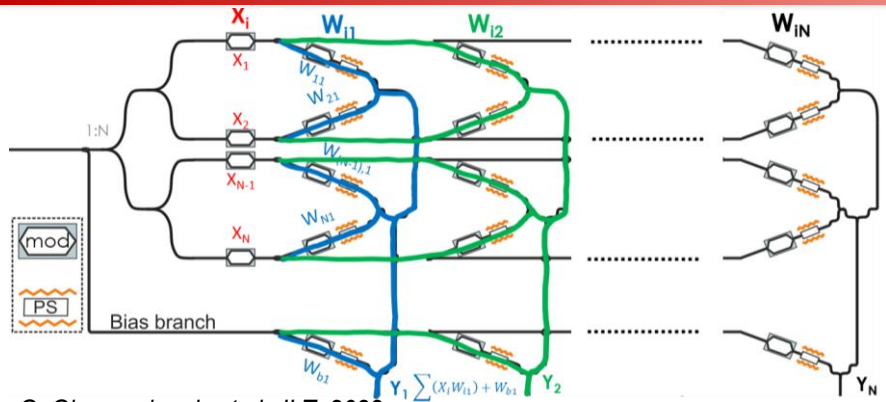
Lessons Learned



- ✓ **Incoherent architectures** support 10's GHz clock rates, but:
  - require one wavelength per axon: **no WDM capability**
- ✓ **SVD coherent architectures** require a single wavelength, but:
  - hardly exceed **GHz range**
  - slow weight update – **small NN sizes**
  - **degrade fidelity**

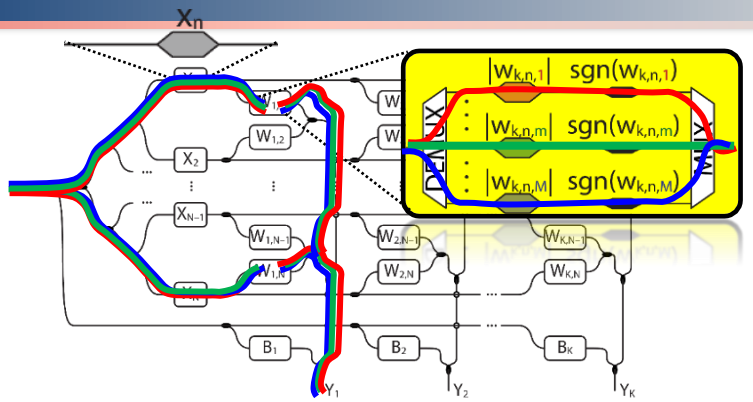
Lesson #3: need for a new MVM architecture (that supports TDM)

The photonic Xbar processor



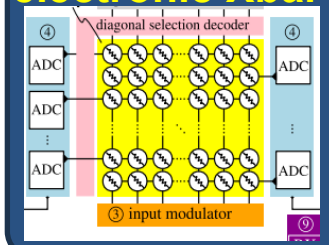
G. Giamougiannis et al, JLT, 2023

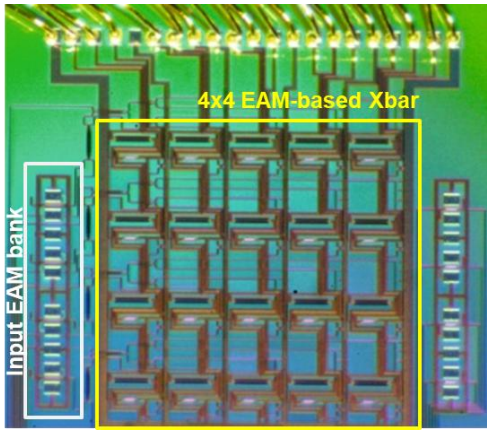
G. Giamougiannis et al, JSTQE 2023



M. Moralis-Pegios et al, Nat. Comms, 2024

the equivalent to the analog electronic Xbar

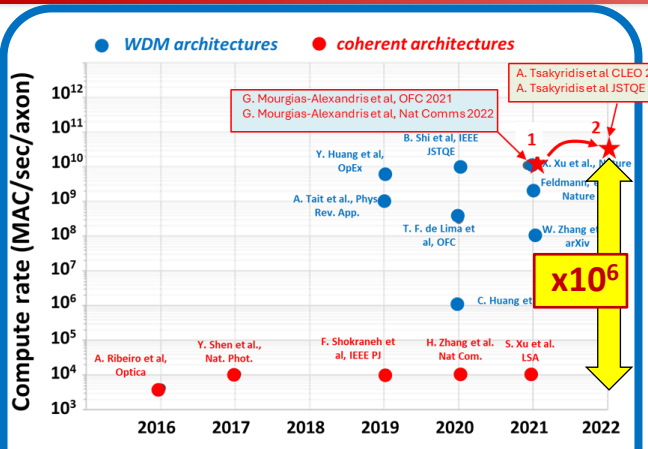




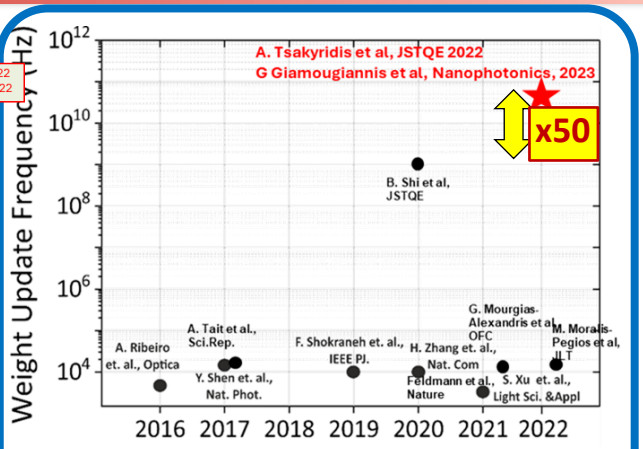
S. Kovaios et al, JLT 2025

A. Totovic et. al., Neuromorph. Comput. Eng. (2022)

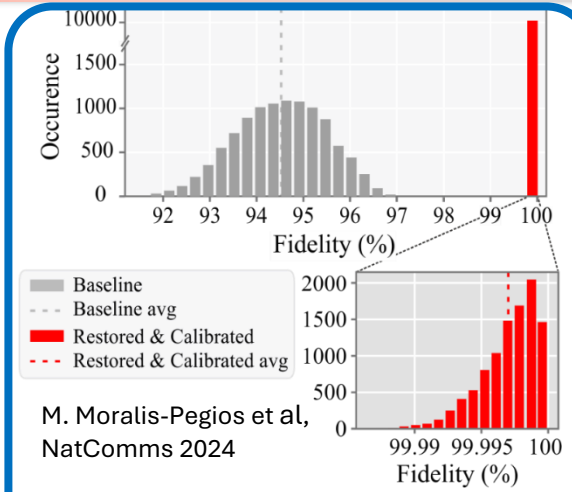
The photonic Xbar processor



- ✓ Record-high **50GHz compute rate**
- ✓ **x10⁶ improvement** (now 60GHz, Z. Lin et al, NatComms 2024)

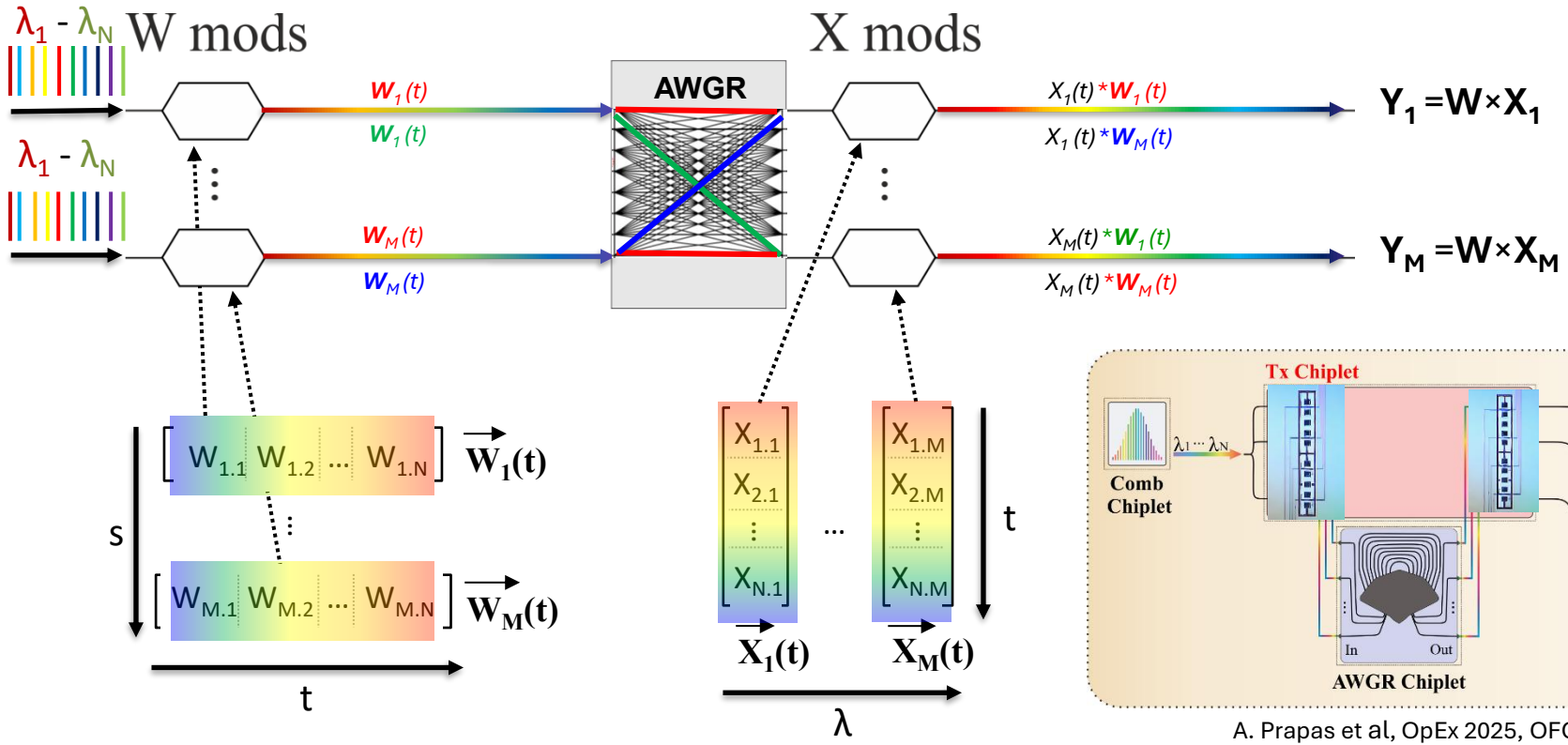


- ✓ **50Gs/sec photonic GeMM** via 50x faster weight update rates
- ✓ **Supports TDM**, i.e. NN size >> PNN



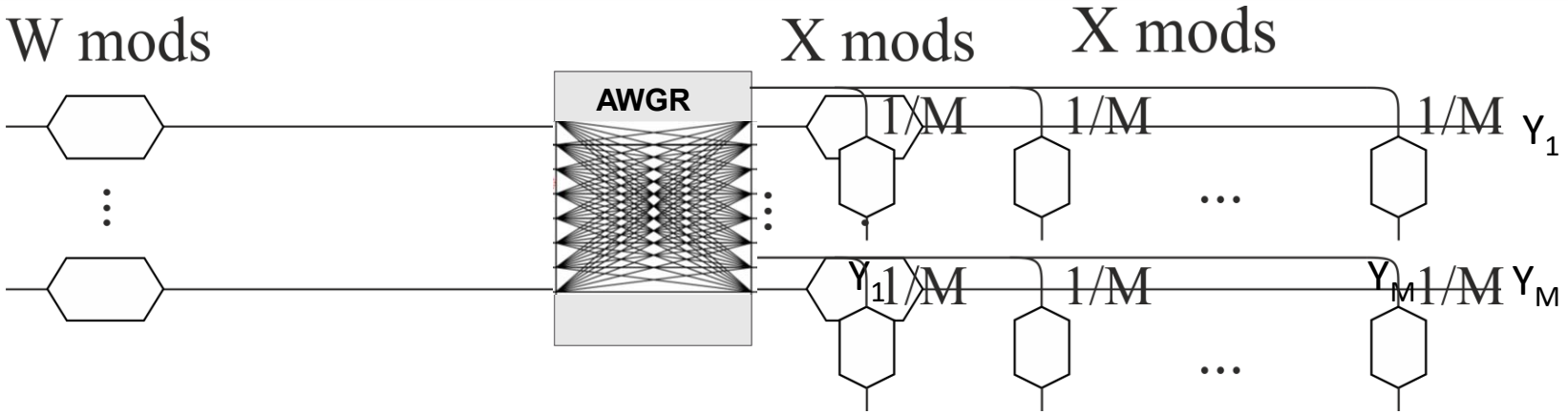
- ✓ Fidelity restoration &..
- ✓ **Record at 99.997 %**

…and a recent AWGR-based TSWDM design



A. Prapas et al, OpEx 2025, OFC 2025

262 TOPs Photonic Tensor Core

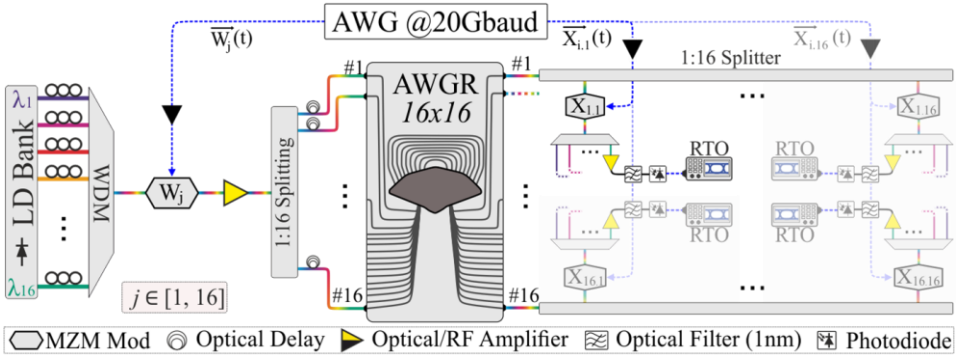


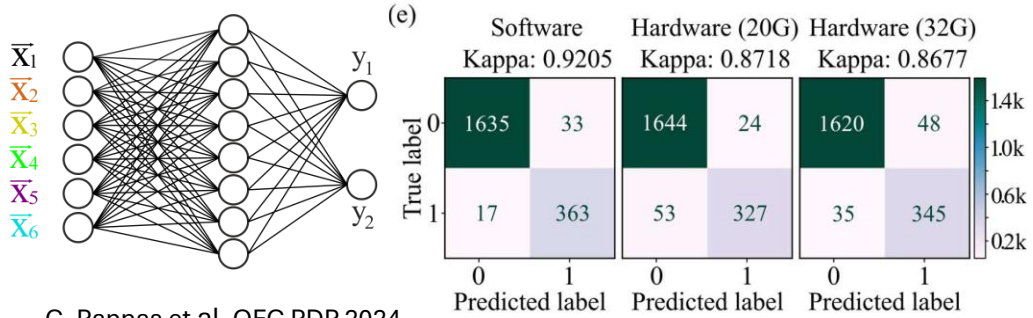
$$\begin{bmatrix} W_{1.1} & W_{1.2} & \cdots & W_{1.N} \\ W_{2.1} & W_{2.2} & \cdots & W_{2.N} \\ \vdots & & \vdots & \\ W_{M.1} & W_{M.2} & \cdots & W_{M.N} \end{bmatrix} \times \begin{bmatrix} X_{1.11} & X_{1.12} & \cdots & X_{1.1M} \\ X_{1.21} & X_{1.22} & \cdots & X_{1.2M} \\ \vdots & & \vdots & \\ X_{1.N1} & X_{1.N2} & \cdots & X_{1.NM} \end{bmatrix} \cdots \begin{bmatrix} X_{M.11} & X_{M.12} & \cdots & X_{M.1M} \\ & & & X_{M.2M} \\ & & & \\ & & & X_{M.NM} \end{bmatrix} \cdots$$

$$Y_1 \cdots \begin{bmatrix} Y_{1.11} & Y_{1.12} & \cdots & Y_{1.1M} \\ Y_{1.21} & Y_{1.22} & \cdots & Y_{1.2M} \\ \vdots & & \vdots & \\ Y_{1.M1} & Y_{1.M2} & \cdots & Y_{1.MM} \end{bmatrix} \begin{matrix} \\ \cdots Y_{2.M} \\ \vdots \\ \cdots Y_{M.M} \end{matrix} M \cdots$$

Y_M

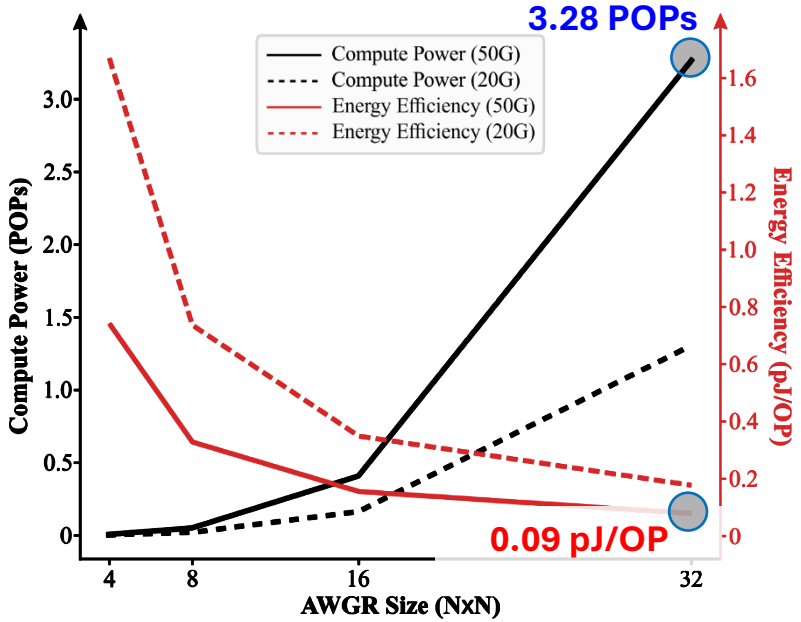
DDoS Attack Detection



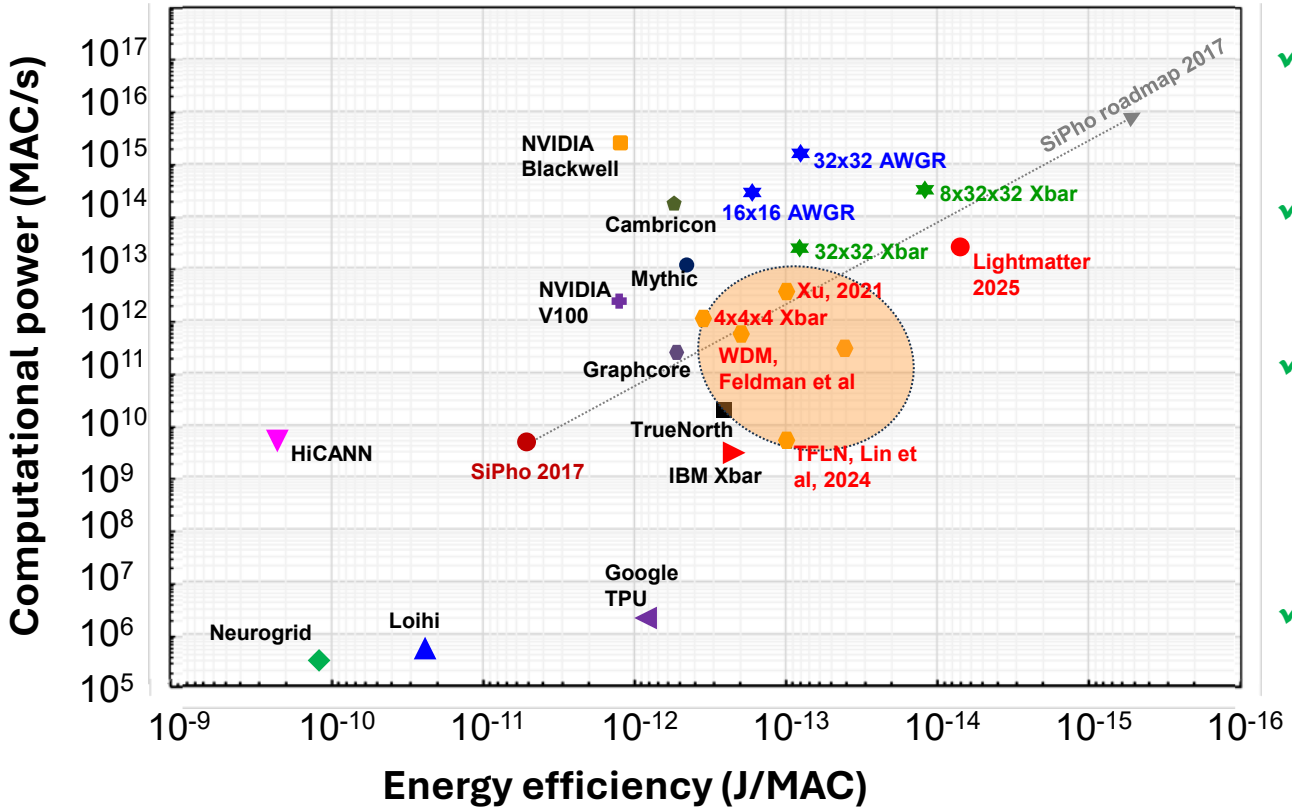


C. Pappas et al, OFC PDP 2024
C. Pappas et al, JLT 2025

- ✓ **Compute** scales with $O(N^3)$
- ✓ **hardware** scales with $O(N^2)$

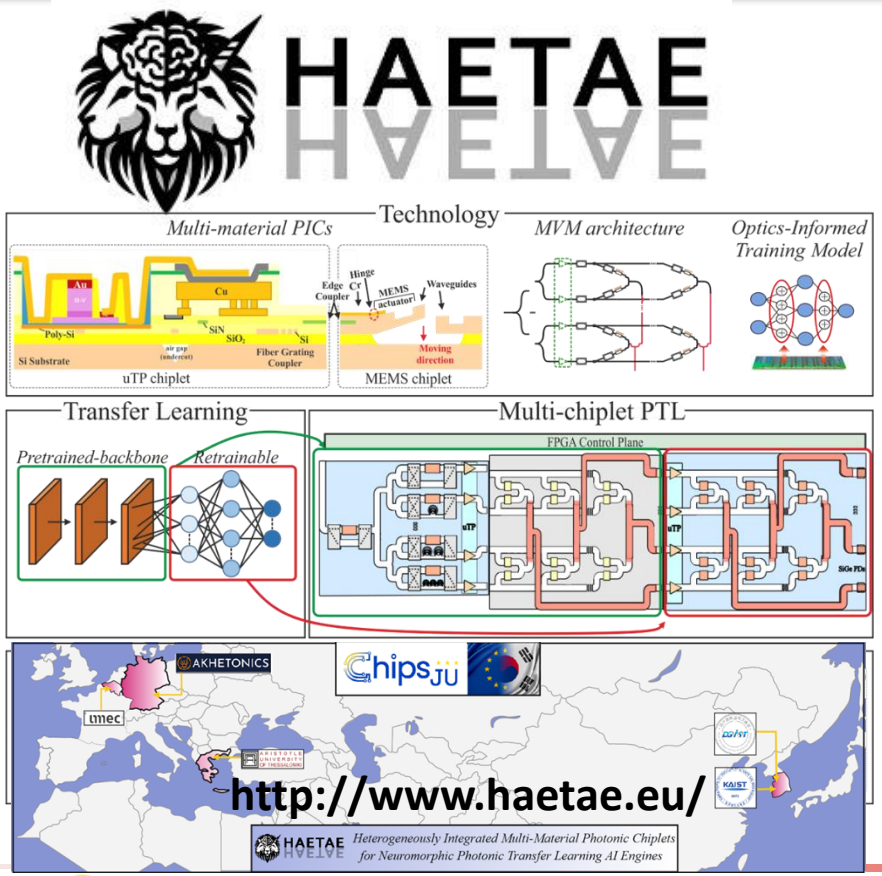


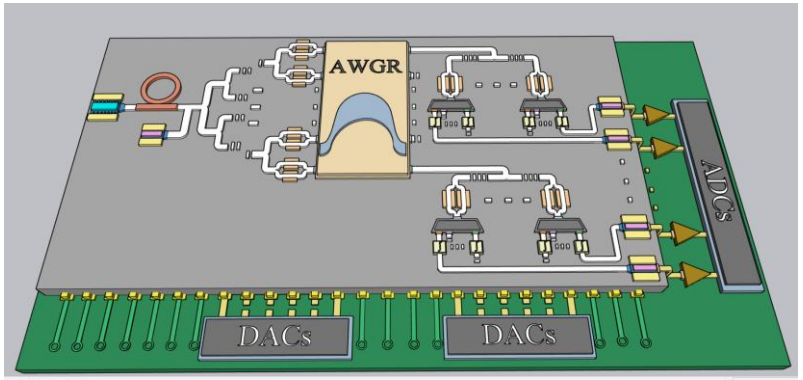
Break the PetaMAC/sec barrier



- ✓ On-chip reconfigurable PNNs up to 10 TOPS
- ✓ Lightmatter Photonic Tensor Core: 65 TOPS
- ✓ 4x4x4 Xbar 1 TOPS on-chip: 32x32x8 Xbar projected to reach 400 TOPS
- ✓ 262 TOPS demonstrated – 32x32@50G AWGR projected to >3 POPS !

On-chip in 3 years from now





[C. Pappas et al, "A 262 TOPs Hyperdimensional Photonic AI Accelerator", submitted at APL Photonics, arXiV pre-print]

Acknowledgements

SiPho-G
Advanced GeSi components for next-generation silicon photonics applications
https://sipho-g.highlite-h2020.eu/

GATEPOST PROJECT

celestial AI

HAETAE

STARLight

Thank you!

Visit us at http://winphos.web.auth.gr/