Programming photonic hardware for computing

Francesco Da Ros
fdro@dtu.dk
The computing power challenge

Data source: Epoch (2023)

OurWorldInData.org/artificial-intelligence | CC BY
Scaling of energy efficiency in digital computing

Energy Efficiencies of Top 3 Supercomputers

![Bar chart showing energy efficiencies of top 3 supercomputers over years 2013 to 2023. The chart indicates a doubling time of 2.4 years.]

Data source: Green500 list, data November 2023

End of Moore’s Law?

![Graph showing trend data of transistors, single-thread performance, frequency, typical power, and number of logical cores over 50 years. The data suggests the end of Moore's Law.]

https://github.com/karlrupp/microprocessor-trend-data
Scaling of energy efficiency in digital computing

Energy Efficiencies of Top 3 Supercomputers

Common carbon footprint benchmarks

- Roundtrip flight b/w NY and SF (1 passenger) - 1,984
- Human life (avg. 1 year) - 11,023
- American life (avg. 1 year) - 36,156
- US car including fuel (avg. 1 lifetime) - 126,000
- Transformer (213M parameters) w/ neural architecture search - 626,155

Data source: Green500 list, data November 2023

End of Moore's Law?

Data source: Green500 list, data November 2023

https://github.com/karlrupp/microprocessor-trend-data
A basic neural network

\[ z = f_{act}(w^T x) \]
Vector-Matrix Multipliers

"small" NN but >400 connections/weights
# interconnections $\propto (\# \text{ nodes})^2$

<table>
<thead>
<tr>
<th>Electronics (metal wiring)</th>
<th>Photonics (waveguides)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher energy consumption</td>
<td>Lower energy consumption</td>
</tr>
<tr>
<td>High loss</td>
<td>Low losses</td>
</tr>
<tr>
<td>Narrow freq. bandwidth</td>
<td>Wide freq. bandwidth</td>
</tr>
<tr>
<td>Sensitive to interference</td>
<td>Lower sensitivity to interference</td>
</tr>
</tbody>
</table>

* Analog electronics for computing is also a very active research topic.
Outline of this talk

• Integrated photonics for computing
  – Building blocks and architectures

• A key challenge: programming the circuit
  – Why is accurate programming important?
  – Offline programming
  – Online programming

• Conclusions
Building blocks for integrated photonic neural networks

Weights – Vector-matrix multipliers

Nodes – activation functions

SOA: Semiconductor optical amplifier
Vector-matrix multipliers – MZI meshes

- Several configurations proposed
  - Reck (triangular)
  - Clements (rectangular)
  - Universal generalized (UGMZI)
  - Diamond
  - Hexagonal mesh
  - …

- Coherent vs. incoherent operation

- MZIs can be replaced by MRRs

Tradeoffs between scalability (# couplers/phase shifters), path loss difference, circuit depth, tolerance to errors

M. Reck, et al. PRL 1994
Y. Shen, et al., Nat Phot 2017
W.R. Clements, et al., Optica 2016
A. Cem, JLT 2023
R. Hamerly, PRA 2022

MZI: Mach-Zehnder interferometer
MRR: Microring resonator
MMI: Multimode interferometer
Vector-matrix multipliers – Xbar arrays

- Several configuration proposed for the weighting elements
  - MRRs
  - Phase-change materials
  - MZIs and Phase shifters
  - …

- Coherent vs. incoherent operation

- Potential for matrix-matrix multiplication through WDM (and WDM+FM) up-scaling

Potential scaling issues and very sensitive to phase-mismatch in the optical path.

Vector-matrix multipliers – Weight & Add

- WDM dimension used for multiplexing columns
- Mainly incoherent operation

Potential scaling issues, matrix size is limited by the number of wavelengths.

B. Shastri et al., Nat. Phot. 2021     A.N. Tait et. al., JSTQE 2016
Vector-matrix multipliers – SOA banks

- WDM dimension used for multiplexing columns
- Mainly incoherent operation
- Allows for weights > 1

Single platform for linear and nonlinear operations and inherent signal amplification but higher energy consumption and added noise

SOA: Semiconductor optical amplifier
Alternative photonic NN architectures

Reservoir computing

Nonlinear propagation

Diffractive networks

Spiking networks


B. Rahmani, et al., Nanophotonics 2022
L.G. Wright, et al., Nature 2022

X. Li, et al., Science 2018
Z. Chen, et al., Nat. Phot. 2023

Outline of this talk

• Integrated photonics for computing
  – Building blocks and architectures

• A key challenge: **programming** the circuit
  – Why is accurate programming important?
  – Online programming
  – Offline programming

• Conclusions
Importance of programming accuracy

Bit Precision = \log_2 \left( \frac{w_{max} - w_{min}}{n \cdot \sigma} \right)

Error in weights

PDF

Weight error [dB]

0
0.5
1
1.5

10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}

Bit precision

n=3
n=1

GPU

GPU/TPU

W. Zhang, et al., Optica 2022

Programming photonic hardware for computing
From modelling error to task performance

Training with inaccurate models leads to performance penalty during inference

A. Cem, et al., JLT 2023

W. Zhang, et al., Optica 2022
Training photonic networks

**On-line training (in situ)**
- Performed on the specific PIC
- Generally requires extra hardware
- Iterative procedure (re-train → re-start)

**Off-line training (in silico)**
- Relies on an accurate/fast PIC model
- Allows for faster re-configurability
- Does not capture drifts

---

S. Pai, et al., Science 2023
M. Milanizadeh, et al., JSTQE 2020
H. Zhang, et al., ACS Phot 2021

S. Bandyopadhyay, et al., Optica 2021
A. Cem, et al., JLT 2023
In-situ training (I)

Calibration and hardware correction

- Simple sequential procedure
- Many calibration measurements required and not easily accounting for cross-talk effects

S. Bandyopadhyay, et al., Optica 2021
and many more...

In-situ backpropagation

- Requires monitoring – on-chip or off-chip hardware
- Effective but scalable?

S. Pai, et al., Science 2023
In-situ training (II)

Gradient-approximation algorithms

- Current demonstration/algorithms are circuit-specific
  
  A. Momeni, et al., arXiv 2304.11042 2023

Forward-only algorithms

- Algorithms fine-tuned to the specific circuit
  
  I. Oguz, Opt. Lett. 2023
  E. Martin, et al., iScience 2021

Auxiliary training circuit

- On-chip training
- Challenges in scaling up the circuit size
  
  M. Filipovich, et al., Optica 2022
In silico training

Accurate physical models of the building blocks exist.

Packing MZI/MRR meshes tightly:

- Optical crosstalk – waveguide crossing
- Thermal crosstalk – thermal diffusion
- Electrical crosstalk – voltage delivery network
- Fabrication errors/tolerances

Are simple models accurate enough?
Simple MZI mesh model

\[ P_{out,2} = W_{21}(V_1, V_2, \ldots)P_{in,1} \]

Loss Extinction ratio

\[ W_{ij} = L_{ij} \prod_{k \in K_{ij}} \left( \frac{1}{4} \frac{\sqrt{ER} - 1}{\sqrt{ER} + 1} - \exp \left( -\sqrt{-1} \left( \phi_k^{(0)} + \phi_k^{(2)} V_k^2 \right) \right) \right)^2 \]
MZI mesh model with crosstalk

\[ W_{ij} = L_{ij} \prod_{k \in K_{ij}} \frac{1}{4} \left( \sqrt{ER - 1} + \sqrt{ER + 1} \right) - \exp \left( \sqrt{-1} \left( \phi_{ij,k}^{(0)} + \sum_{m=1}^{M} \phi_{ij,k,m}^{(2)} V_{m}^{2} \right) \right) \]

\[ P_{out,2} = W_{21}(V_1, V_2, \ldots) P_{in,1} \]

A. Cem, et al., OFC 2022  S. Bandyopadhyay, et al., Optica 2021
Grey-box NN MZI mesh model
Experimental setup and 3x3 MZI mesh

Measurements:

1. Individually sweep one voltage \([0, V_{\pi}]\)
2. Randomly chosen voltages

Dataset = \{ Voltages | Weights \}

Including thermal cross-talk improves performance but not as much as a grey-box ML model.

\[
W_{ij} = L_D \prod_{k \in K_{ij}} \left| \frac{\sqrt{ER} - 1}{\sqrt{ER} + 1} - \exp \left( \sqrt{-1} \left( \phi_i^{(k)} + \phi_j^{(k)} \right) \right) \right|^2
\]

\[
W_{ij} = L_D \prod_{k \in K_{ij}} \left| \frac{\sqrt{ER} - 1}{\sqrt{ER} + 1} - \exp \left( \sqrt{-1} \left( \phi_i^{(k)} + \frac{M}{m=1} \phi_j^{(k)} \right) \right) \right|^2
\]

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (w_i - \hat{w}_i)^2}
\]

RMSE = 3.26 dB

RMSE = 1.44 dB

RMSE = 0.53 dB

A. Cem, et al., JLT 2023
Full model with thermal crosstalk

1. FTDT analysis
2. 3D Thermal analysis
3. Comparison w/ measurements

Fabrication errors, electrical and optical crosstalk not modelled

M. Orlandin, et al., NUDOS 2023
Modelling hexagonal MZI meshes

Even for chips designed to minimize the impact of crosstalk, sensitive applications can be affected.

A. Cem, et al., IPC 2023

ASE: amplified spontaneous emission
OSA: optical spectrum analyzer
Fitted thermal diffusion model

\[ y = \sum_{i=1}^{66} \left( a_1 e^{-a_2 d_i} + a_3 d_i + a_4 \right) \phi_i \]

RMSE = 0.42 pm
Cross-talk compensation

[Graph showing cross-talk compensation with wavelength and optical power axes]
Data scarcity – MZI meshes

Simple Analytical Model with Thermal Crosstalk (SAM+XT)  Neural Network Model (NN)

Simpler models are less accurate but more data-efficient to train

A. Cem et al., Opt. Lett. 2023
Transfer learning for data-efficient modelling

1. Train simpler model with experimental data

2. Generate synthetic data

3. Pre-train NN model with synthetic data

4. Re-train NN model with experimental data
Generalizable crosstalk models

Can physics help in building more efficient/generalizable models?

I. Teofilovic JLT 2024 (in preparation)
Generalizable crosstalk models

Use symmetry arguments to extend the model of a small part of the circuit

I. Teofilovic JLT 2024 (in preparation)
Physical knowledge allows models to generalize (e.g. by extrapolating).

I. Teofilovic Frontiers 2024 (in preparation)
Hardware-aware modelling/training

Include a physically-informed description of the photonic NN during training improves inference.

M. Moralis-Pegios, et al., JLT 2022

V. Shah, N. Youngblood, APL Mach. Learn. 2023
Conclusions

• Accurate training of photonic circuit is necessary to guarantee task performance

• In-situ and in-silico approaches provided a plethora of specific methods but with significant trade-offs required by every method

• No one-fits-all solution yet but lots of interesting directions

• General shift towards physics-informed modelling and online algorithms tuned for photonics circuits
Conclusions

• Accurate training of photonic circuit is necessary to guarantee task performance

• In-situ and in-silico approaches provided a plethora of specific methods but with significant trade-offs required by every method

• No one-fits-all solution yet but lots of interesting directions

• General shift towards physics-informed modelling and online algorithms tuned for photonics circuits

Questions: now and fdro@dtu.dk

Acknowledgements: Villum Foundation through the Villum Young Investigator project OPTIC-AI (grant no. VIL29344), the Horizon Europe research and innovation project PROMETHEUS (grant no. 101070195), the independent research fund Denmark project QUARCOM (grant no. 10.46540/2032-00161B) and the Swedish research council project BRAIN (grant no. 2022-04798).