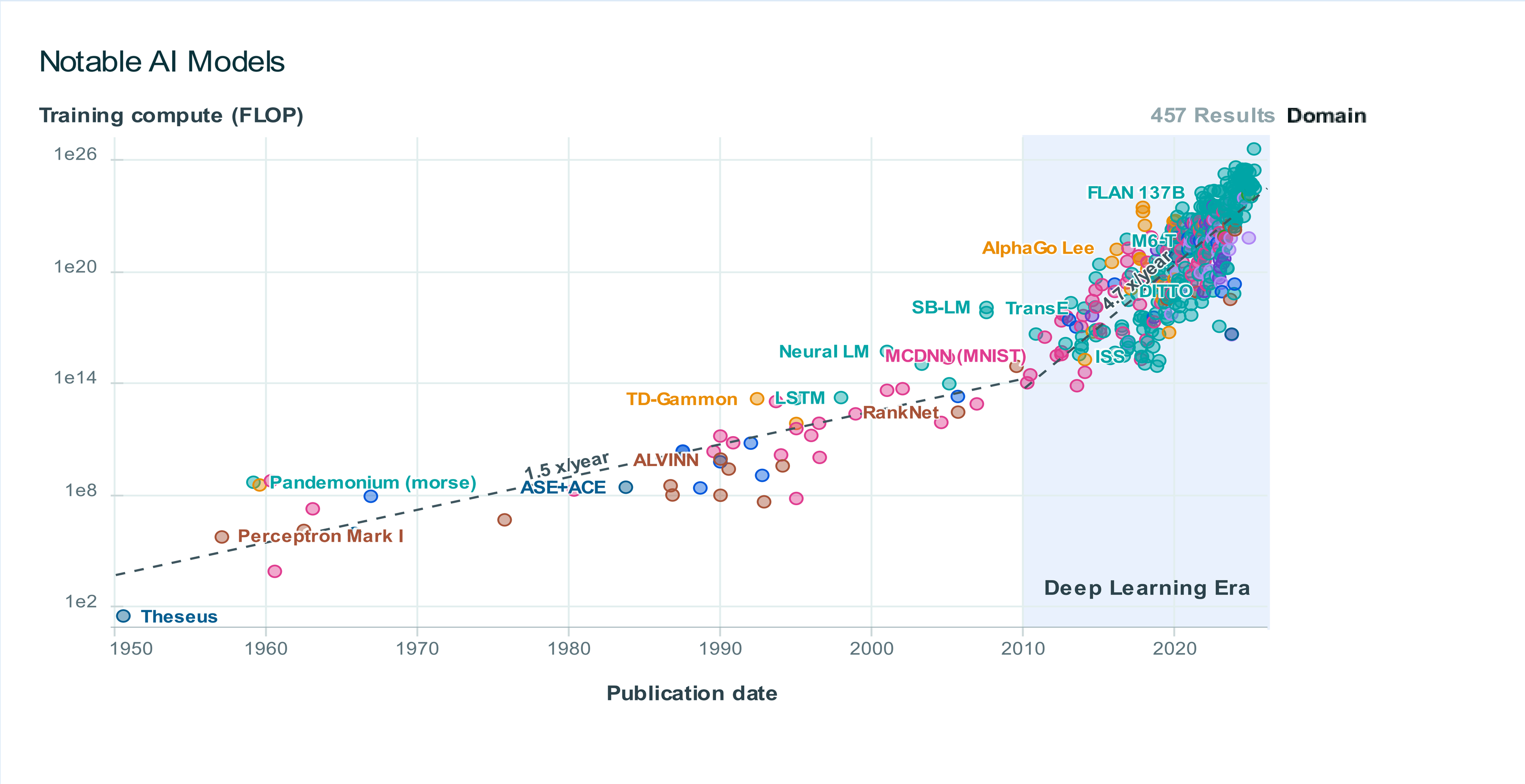
# PHOTONIC ARCHITECTURES FOR ACCELERATING HPC, DL, AND ML OPERATIONS

**Dinah Wobuyaga**
Supervisors: Prof. Philippe Velha and Prof. Flavio Vella
Department of Engineering and Computer Science, University of Trento

ICT DAYS

## INTRODUCTION

- AI computational demands are growing exponentially, with model sizes doubling every 3.4 months
- Modern AI models such as GPT-4 require billions of matrix operations, straining traditional hardware.
- Traditional electronic processors face fundamental limitations in both performance and energy efficiency.




Epoch AI, "AI trends," [Online]. Available: limitations in both performance and energy efficiency. [Accessed: Mar. 13, 2025]. [Accessed: Mar. 13, 2025]

## OBJECTIVES

- To identify and analyze electronic bottlenecks
- To design a photonic architecture that minimizes these bottlenecks
- To develop integration strategies that enable efficient communication between photonic and electronic systems
- To simulate and evaluate photonic accelerator performance in a system environment.

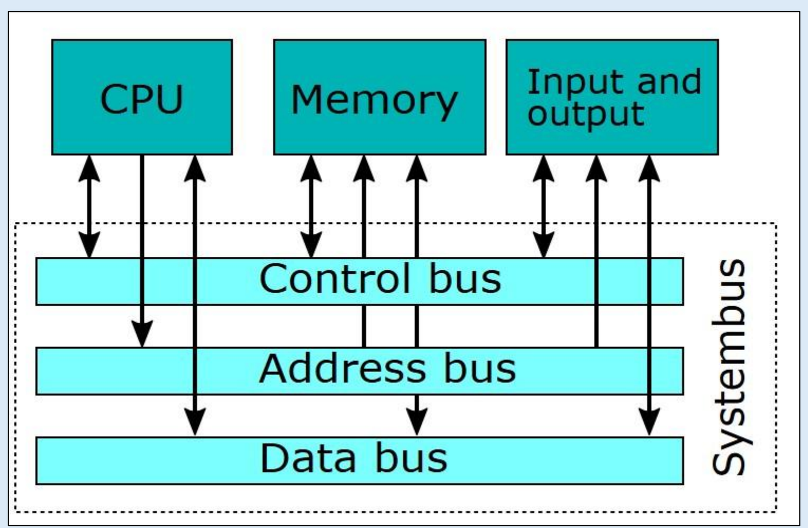
## PROBLEM

### Conventional Computing Bottlenecks

1. Memory access latency creating performance constraints
2. Excessive energy consumption limiting scalability
3. Restricted throughput for matrix operations

### Photonic Computing open Challenges

1. Unexplored memory hierarchy optimizations for photonic accelerators
2. Focus on a single metric
3. Under explored electronic-photonic integration strategies



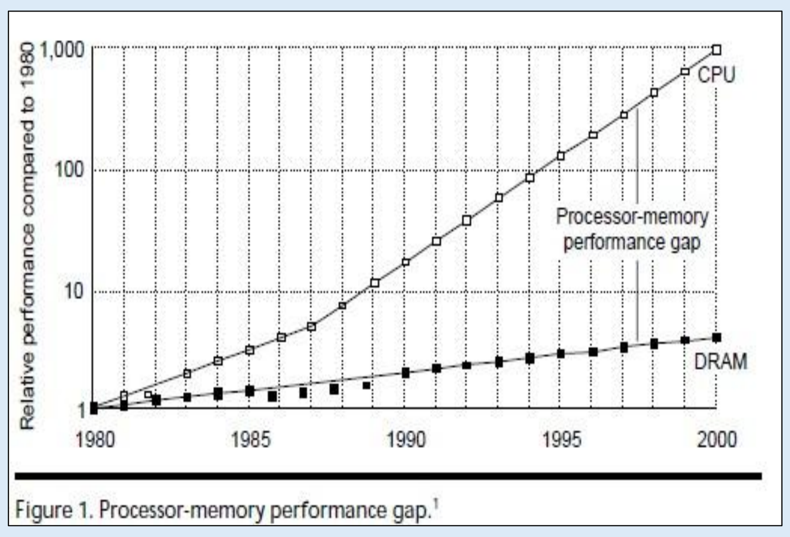




Figure 1. Processor-memory performance gap.[1]

Patterson, David, et al. "A case for intelligent RAM." IEEE micro 17.2 (1997): 34-44.

## SIGNIFICANCE

### Industrial Impact

**Enhancing the adoption of photonic technology**
- Seamless integration with electronics
- Overcoming electronic bottlenecks

**Enabling new architectural innovation**
- Novel Computational Designs
- Optimized for AI and DL operations

**Fueling the next generation of AI**
- Reducing training time

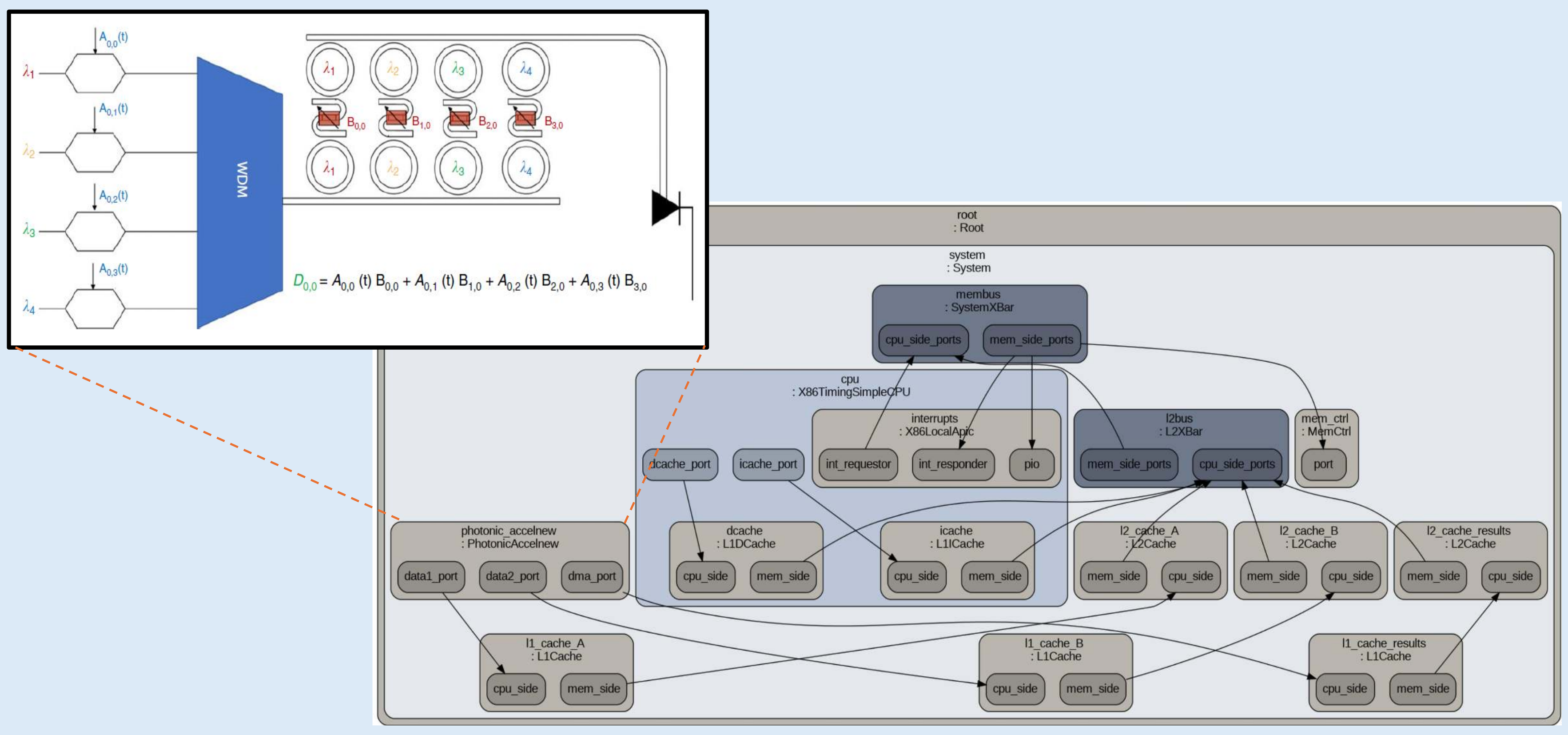
### Meeting future computational Demands

**Speed of light**
- Faster computations
- Parallelism

**Energy efficiency**
- Lower power consumption
- Reduced energy costs

**Low latency**
- Reduced data transmission delays

## PROPOSED APPROACH



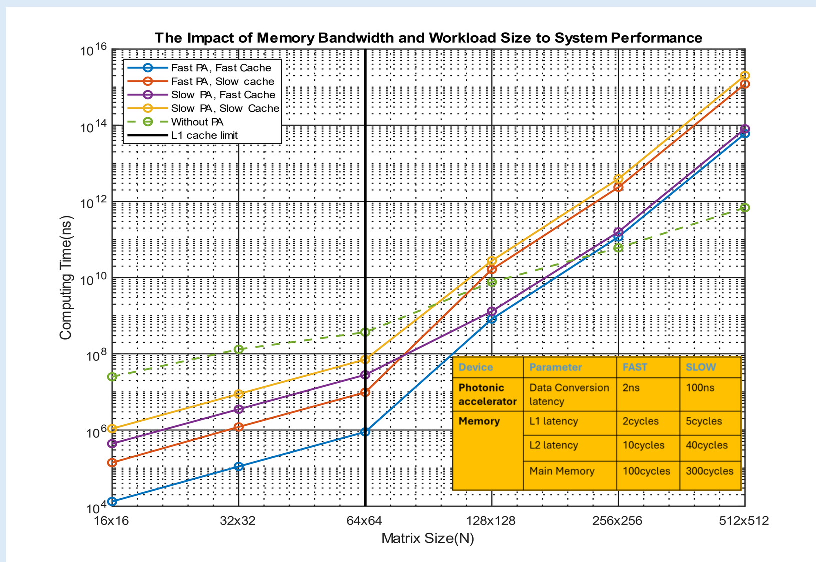

### Simulation tools

- Luceda
- Ansys Lumerical INTERCONNECT
- gem5

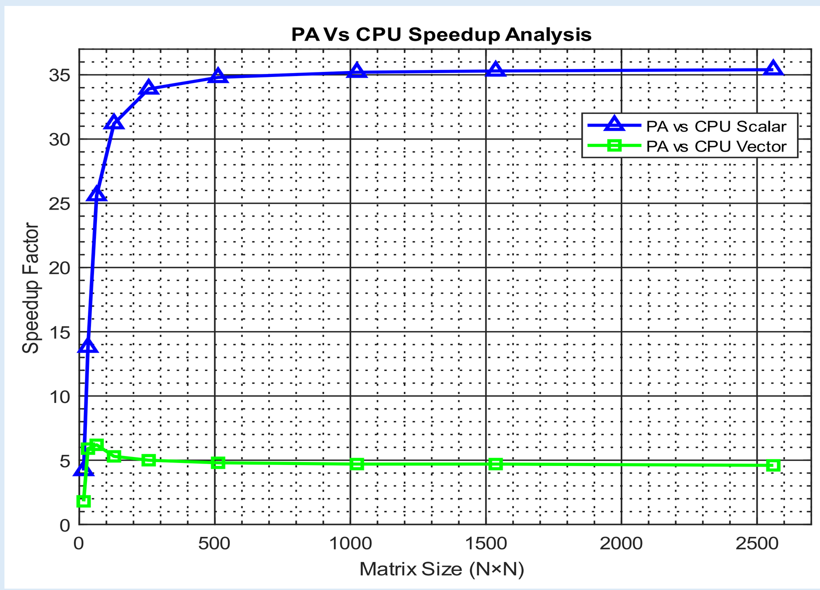
### Core Metric

- Throughput

## PRELIMINARY RESULTS




1. **L1 cache limit:** Performance degrades as matrix sizes exceed the cache capacity
2. Best performance achieved with **fast PA fast cache**
3. Results indicate that memory bandwidth is a bottleneck to the performance of the PA




1. PA outperforms CPU for both scalar and vectorized matrix operations
2. **35x PA** speedup compared to **CPU scalar** processing
3. **~5×** PA speedup to **CPU vectorized** processing

## CONCLUSION

- Photonic architectures enhance **HPC, DL, and ML** performance (GeMM operations).
- Significant **speed** gains observed.
- Cache/memory bottlenecks limit performance, especially for large matrices.
- Photonic accelerators outperform CPU for both **scalar** and **vector execution**.
- Future memory optimization will meet growing computational demands.
- Our photonic architecture achieves peak computational throughput over 10 GFLOPS for matrix operations per channel

## FUTURE WORKS

- Design exploration of photonic system architectures
- Memory hierarchy optimization
- Exploring effective photonic-electronic integration strategies
- Implementing larger model workloads such as CNN, RNN.
- System performance evaluation and analysis using considering other metrics (Latency and Power consumption)

## REFERENCES

[1] Shiflett, Kyle, et al. "Pixel: Photonic neural network accelerator." 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2020.
[2] Demirkiran, Cansu, et al. "An electro-photonic system for accelerating deep neural networks." ACM journal on emerging technologies in computing systems 19.4 (2023): 1-31.
[3] Pasricha, Sudeep. "Optical Computing for Deep Neural Network Acceleration: Foundations, Recent Developments, and Emerging Directions." arXiv preprint arXiv:2407.21184 (2024).
[4] Zhou, Hailong, et al. "Photonic matrix multiplication lights up photonic accelerator and beyond." *Light: Science & Applications* 11.1 (2022): 30.
[5] Lowe-Power, Jason, et al. "The gem5 simulator: Version 20.0+." *arXiv preprint arXiv:2007.03152* (2020).
[6] Epoch AI, "AI trends," [Online]. Available: limitations in both performance and energy efficiency. [Accessed: Mar. 13, 2025]. [Accessed: Mar. 13, 2025]

UNIVERSITÀ DI TRENTO

ICT DAYS