LIGHTELLIGENCE

Optical Network-on-Chip for Large Scale Chiplet Architectures

Huaiyu Meng, Co-founder and CTO

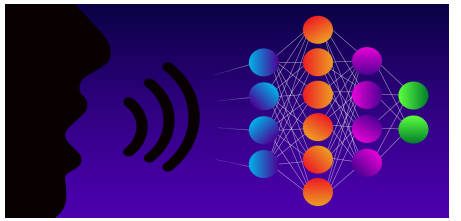
Feb. 26th, 2023

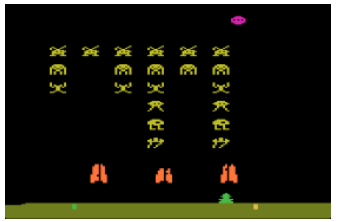
Exponential Growth of Machine Learning Market

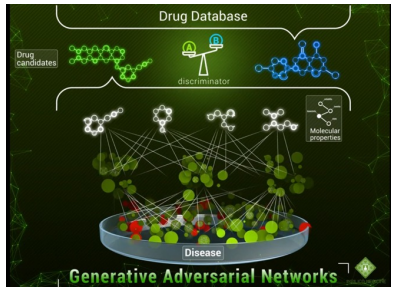
LIGHTELLIGENCE

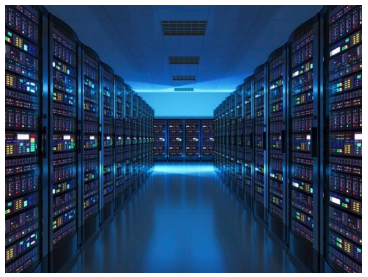
Breakthroughs in deep learning:

- Computer Vision
- Natural Language Processing (NLP)
- Game Playing (Go, Atari)
- Autonomous Vehicles Control
- Advertisement Placement
- Drug or Material Discovery
- ChatGPT
- …





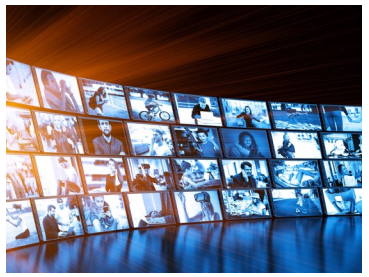




Cloud AI

Finance

Telecom



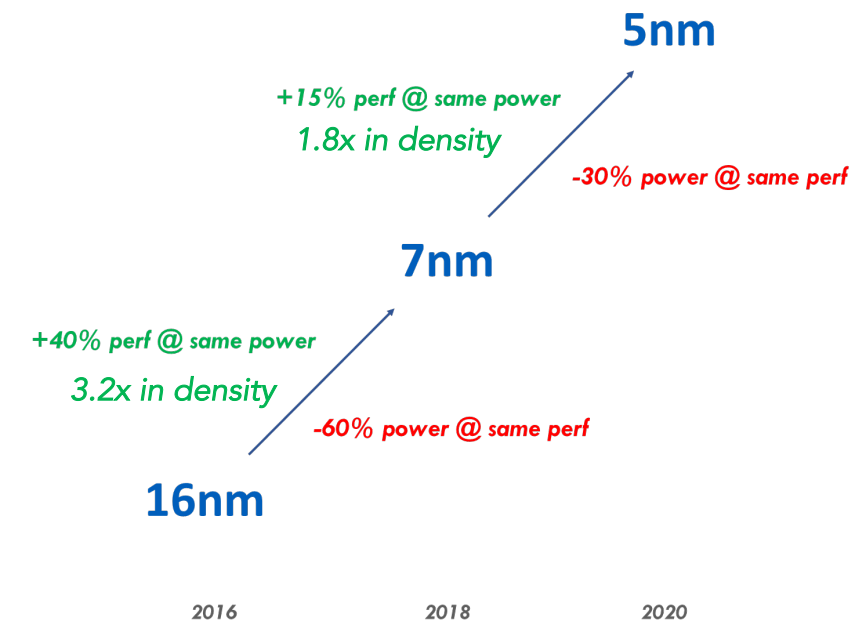
Intelligent Surveillance

Smart Retail

- **Deep learning has extended its application to multiple aspects of our daily life.**
- **Machines are getting better in tasks human used to be good at.**

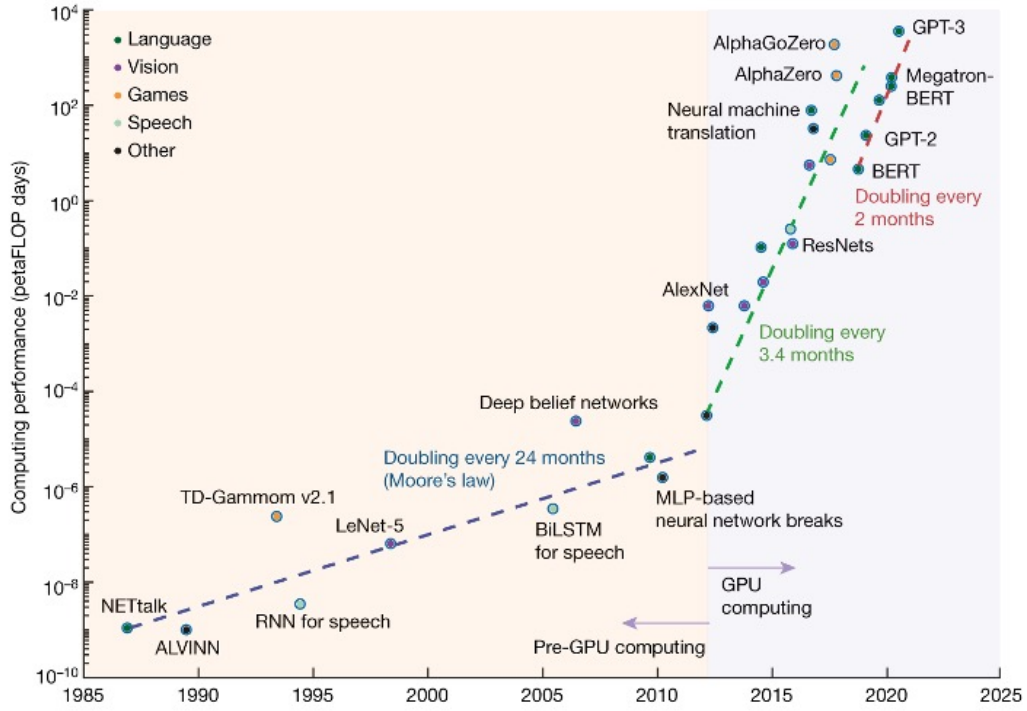
Transistor Scaling Falling behind Demand

Single Transistor Level Improvement



https://www.tsmc.com/english/dedicatedFoundry/technology/logic
https://en.wikipedia.org/wiki/5_nm_process

AI Model Computing Performance Requirement



A. Mehonic and A. J. Kenyon, "Brain-inspired computing needs a master plan," Nature, vol. 604, no. 7905, pp. 255–260, 2022, doi: 10.1038/s41586-021-04362-w.

- Electronics approaching physical limits, hitting walls on power, communication and memory access
- AI model and its computing resource requirement is increasing at a much quicker pace
- Large models cost millions of dollars to train

Performance Improvement: Architecture Innovations

Make transistors work more efficiently: architecture innovations

DSA: trading versatility for performance

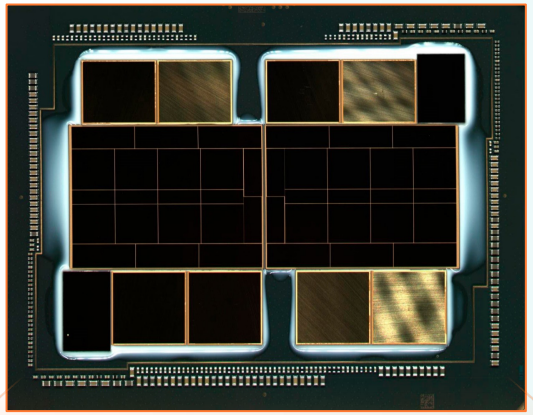
General-purpose innovations

- Instruction-level parallelism (ILP)
- Integration of complex logic function, including more on-chip memory
- Multi-threading, multi-core, and context-based kilo-threading architecture

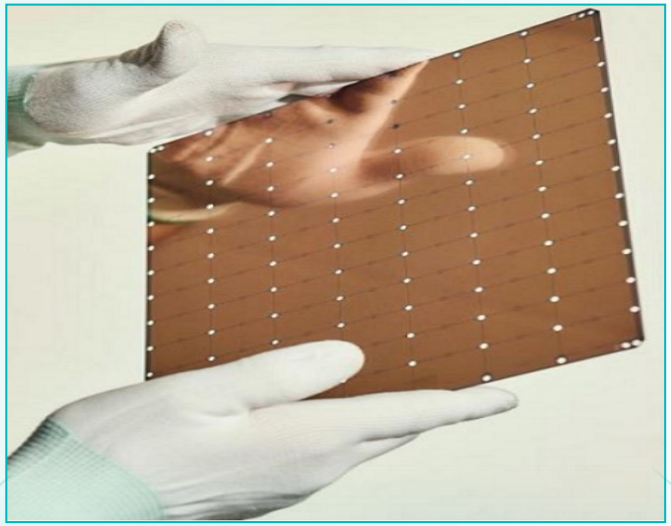
Non-Von Neuman, disruptive architecture

Performance Improvement: Enlarge Silicon Area

Larger area means more transistors: multi-chip module (MCM)

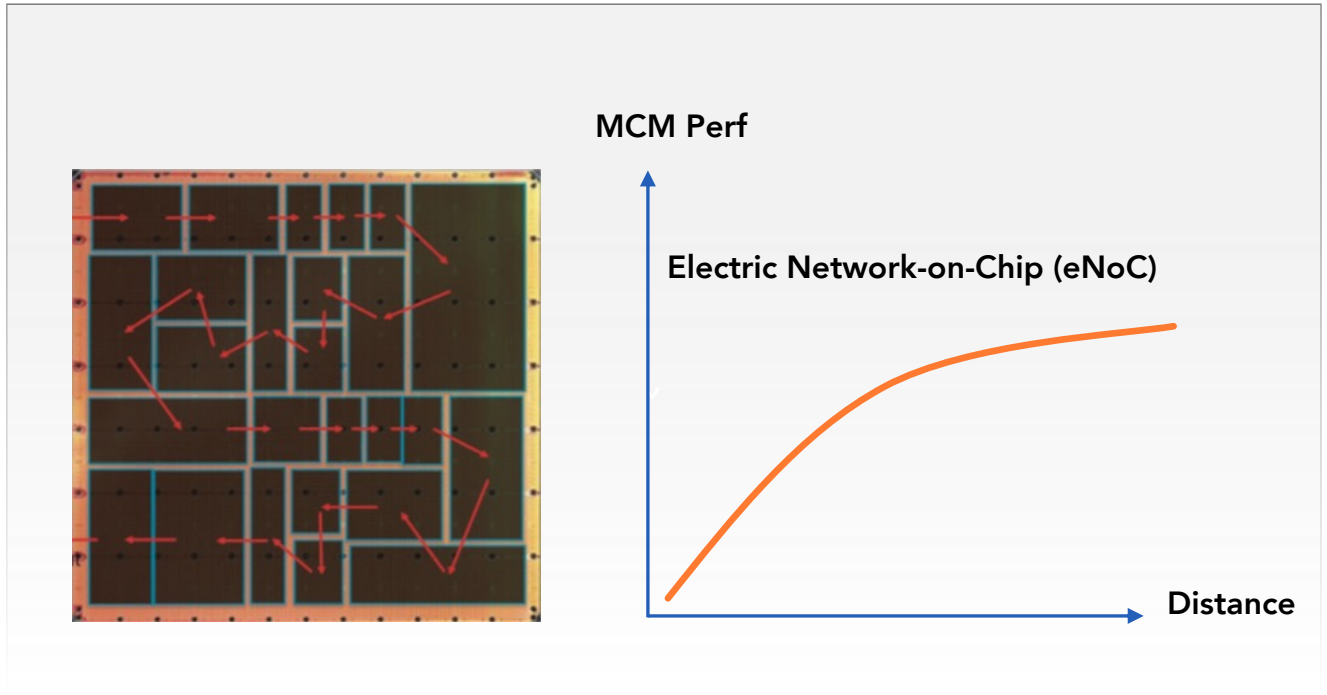


Intel Ponte Vecchio
>1,200 mm^2



Cerebras Wafer Scale Engine
>46,000 mm2

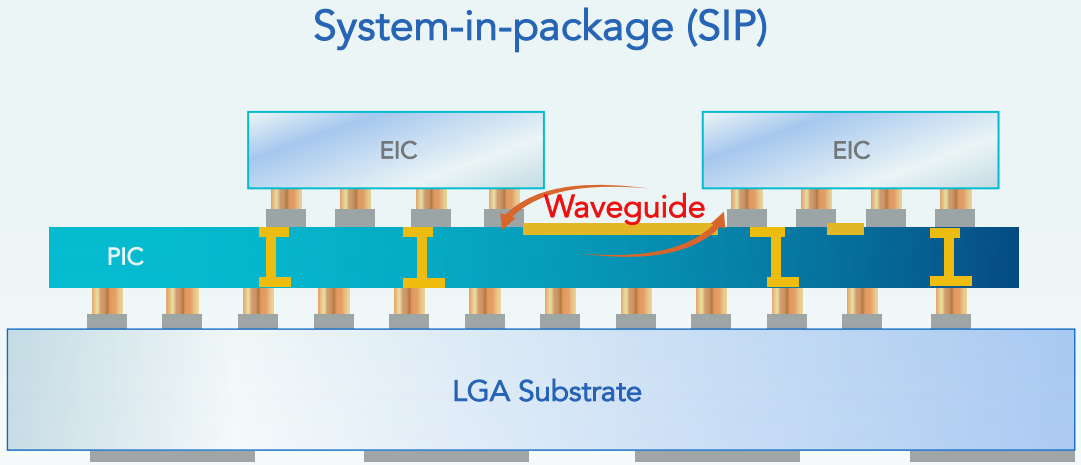
Inefficient Scaling of Performance

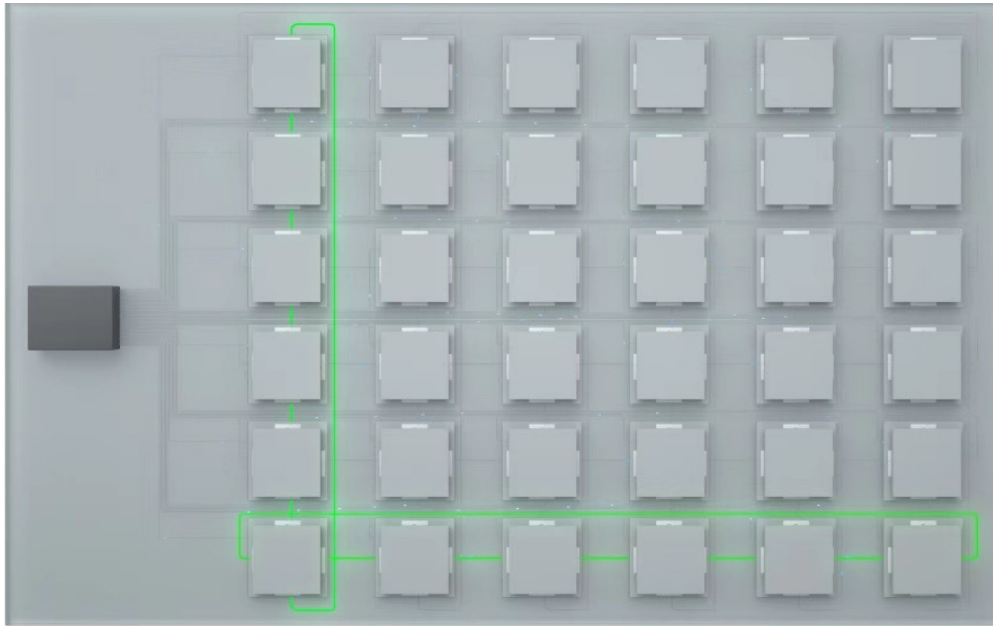


- Electric signal attenuate over distance
  - Nearest neighbor only communication
- More hops equal to more latency
- Difficulty in programming
  - Inefficient utilization

A better interconnect solution is needed for large MCM

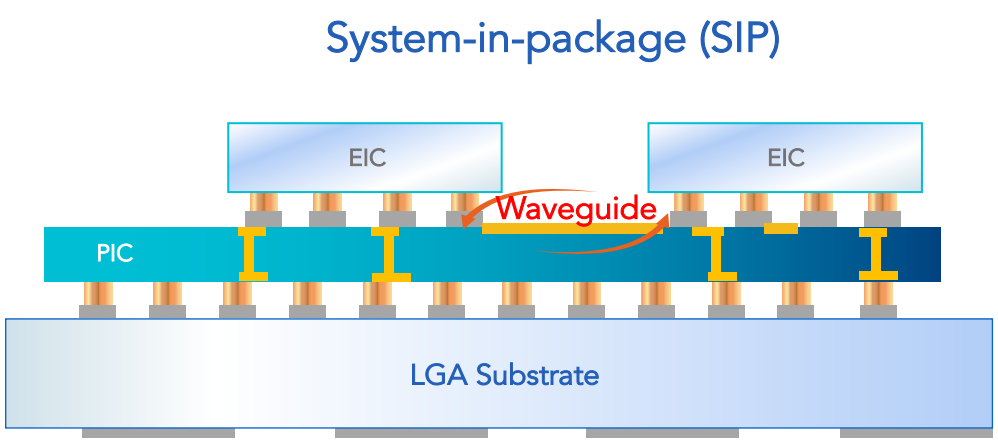
Optical Network-on-Chip (oNOC)

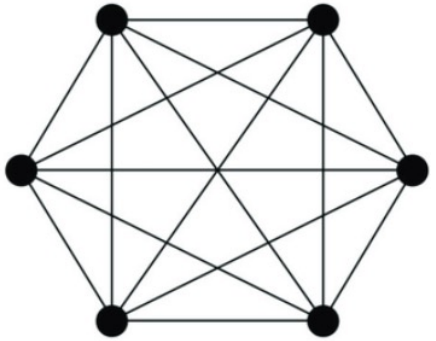




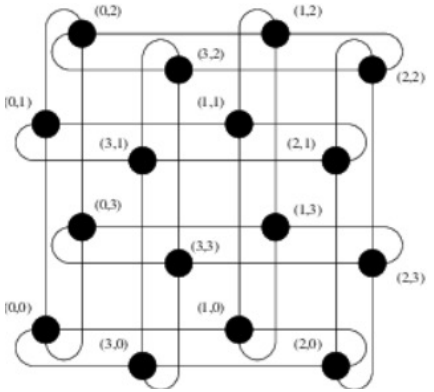
- Optical signal attenuation is small in wafer scale
- Power, latency is independent of distance
- Photonic integrated chip (PIC) as active interposer
- Electrical-optical conversion in PIC

Optical Network-on-Chip (oNOC)





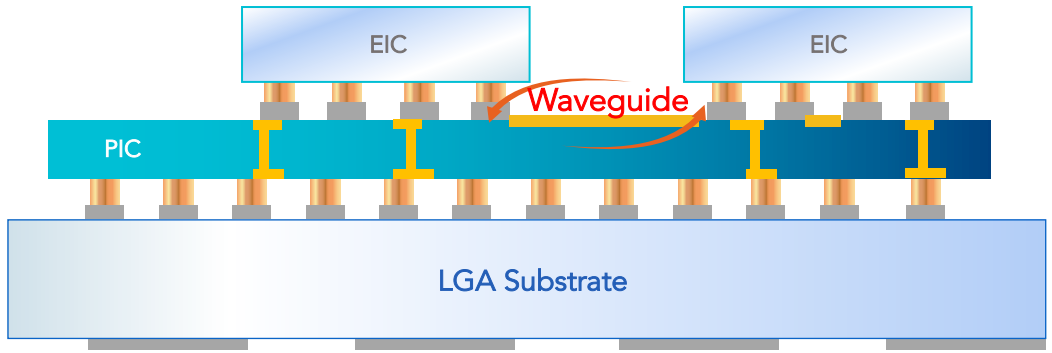
Full mesh

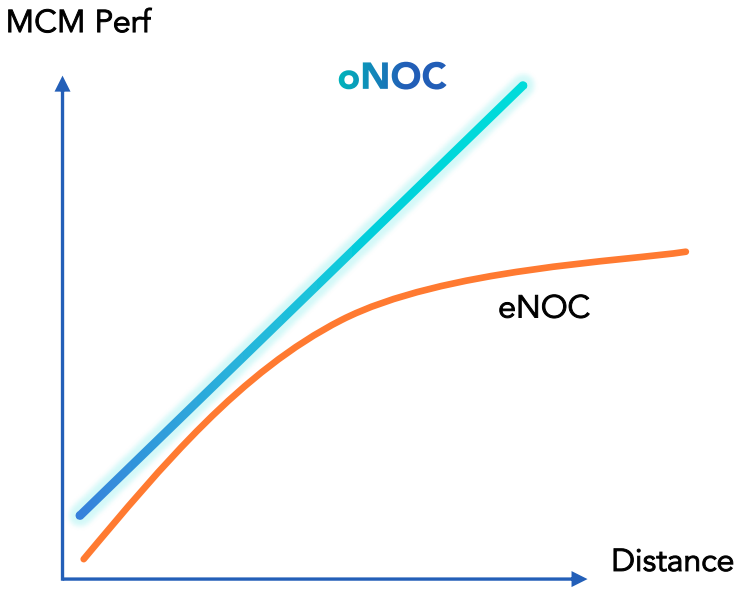


2D Torus

- Inter-chiplet connectivity no longer limited to nearest neighbor
- Various topology is possible

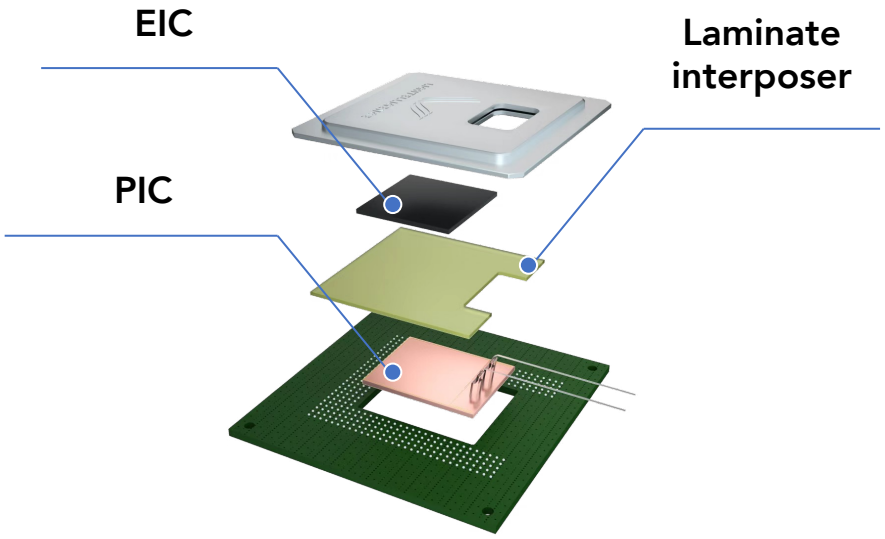
Optical Network-on-Chip (oNOC)

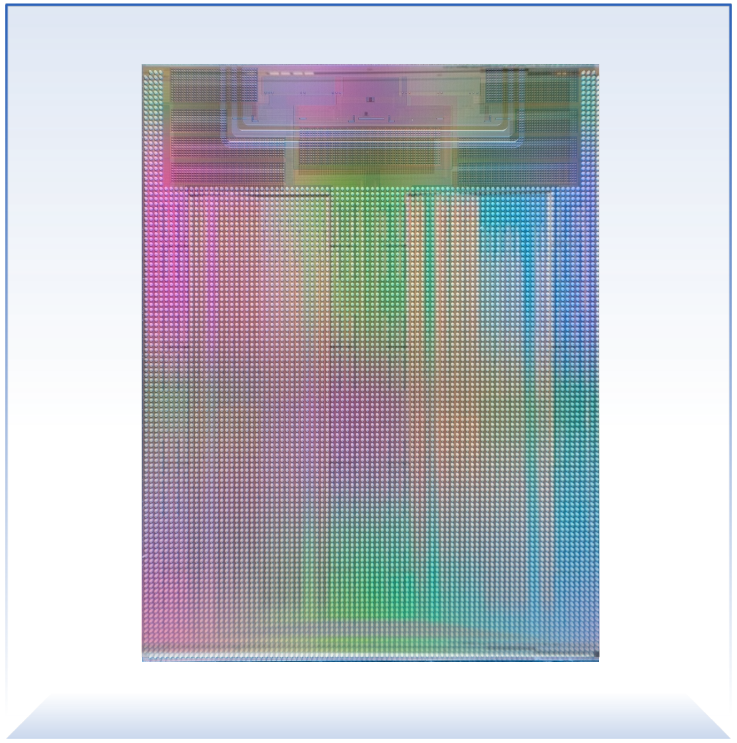




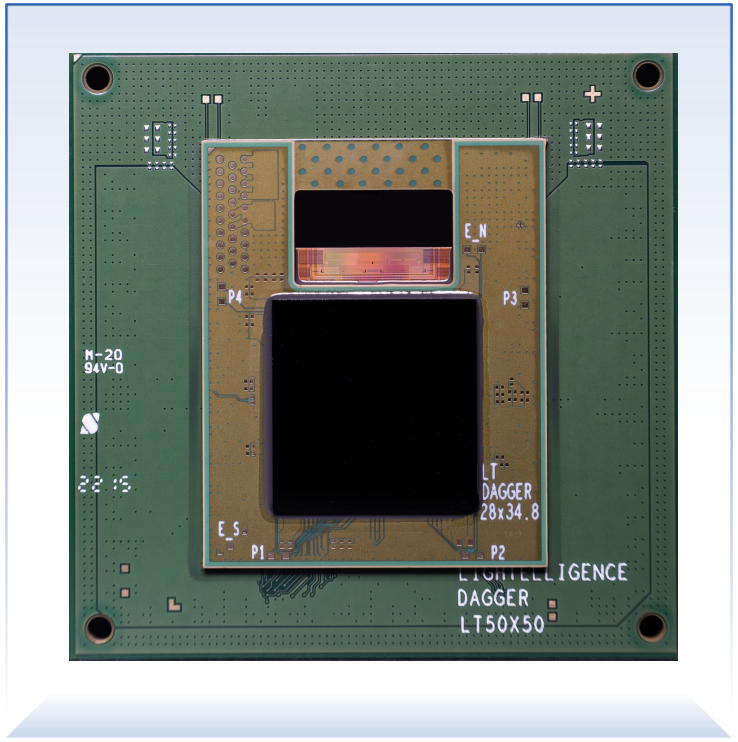
- Mapping workload to hardware become more efficient with flexible network topology
- Close to linear scaling of MCM performance

First Computing SiP with oNOC





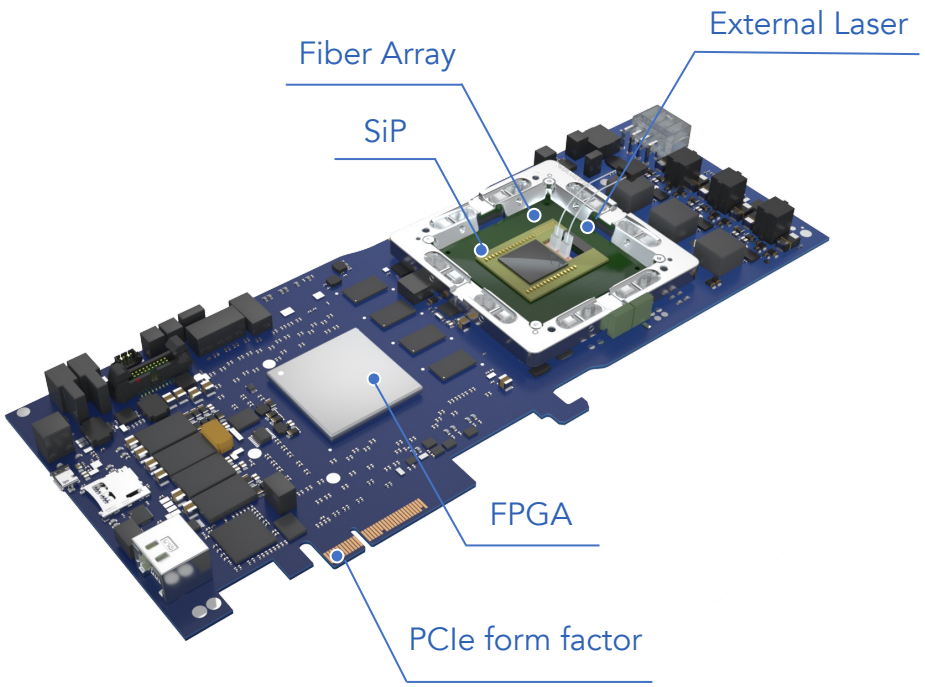
Photonic integrated circuit (PIC)

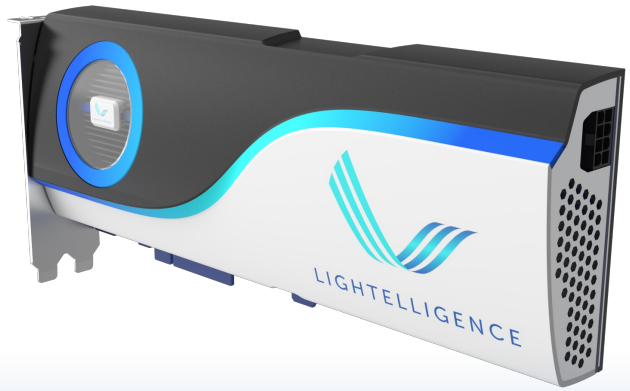


System-in-package

Superior Latency Enabled by oNOC

First commercial grade computing product powered by oNOC





Support AI inference and other applications

- oNoC All-to-All broadcast
- Ultra low latency data transfer
- Lightelligence SDK

Challenges Towards Mass Adoption

LIGHTELLIGENCE

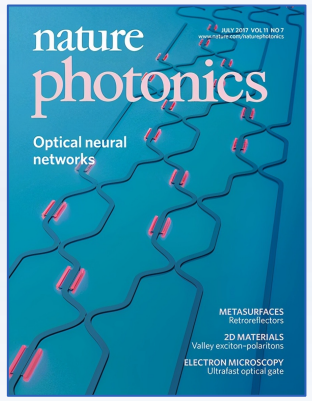
- The maturity of the chiplet ecosystem
- Standard interface between chiplets
  - *UCIE lacks unified inter-chiplet PHY standard*
- Cost of silicon photonic supply chain
  - *Volume production would bring down the cost*

Brief Introduction to Lightelligence

LIGHTELLIGENCE

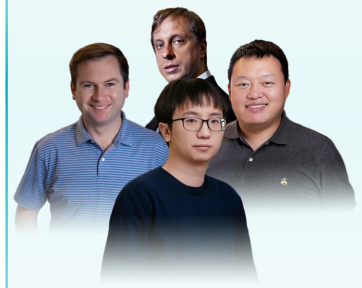
2017

- "Deep Learning with Coherent Nanophotonic Circuits" in Nature



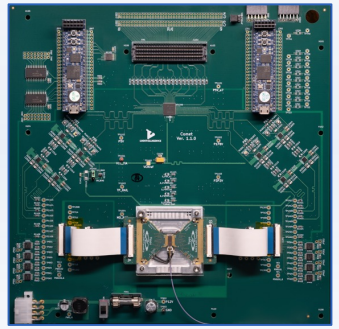
2018

- 1st round of funding
- Team assembled, MIT spinout



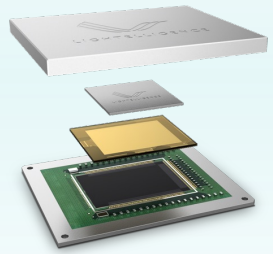
2019

- World 1st photonics AI accelerator demo



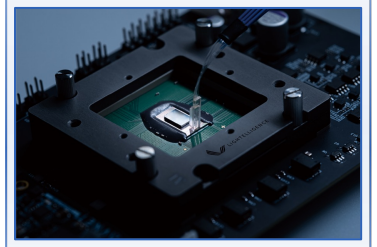
2020

- Successful tape outs
- Scaling optical computing



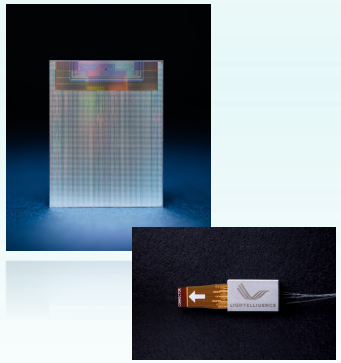
2021

- Announced PACE, demonstrating optical superiority for certain workloads



2022

- Developed Optical Network-on-Chip platform
- Released Moonstone, our on-board optical source



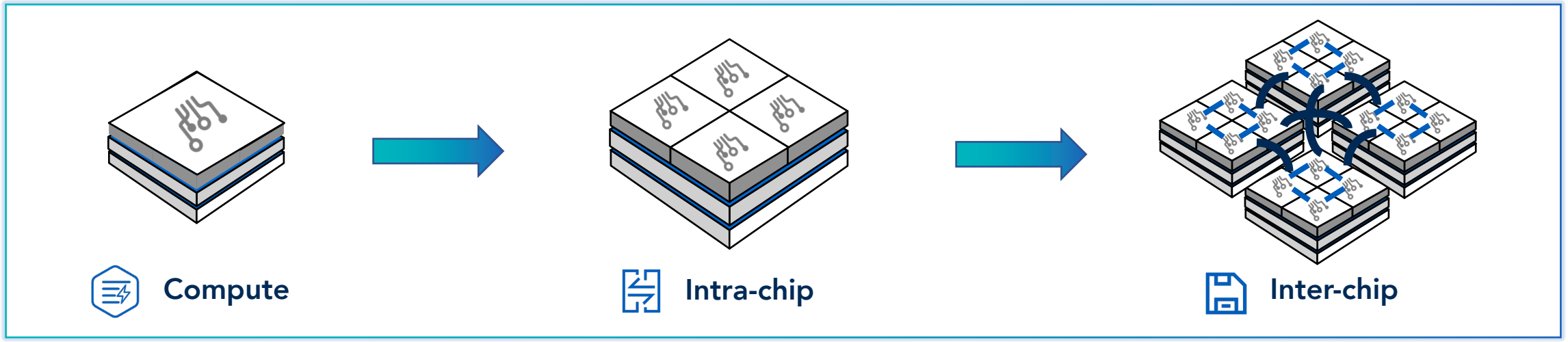
Computing Chip to System

LIGHTELLIGENCE

Single-Compute Engine

Pack More Compute Engine In Single Chip

Connect More Chip to Build Larger System



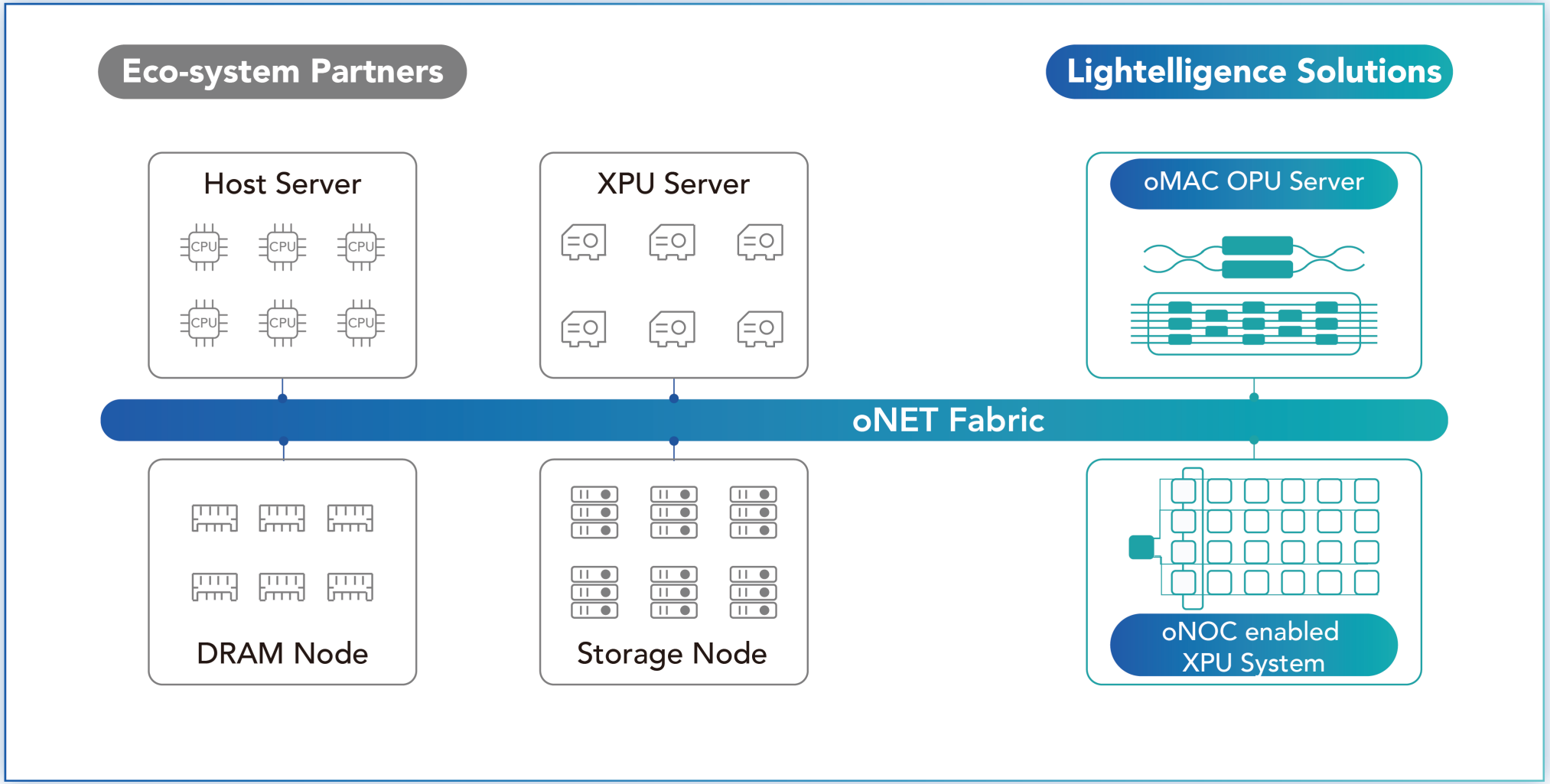
Single compute engine performance improvement

Intra-chip interconnect to connect multiple compute engines

Inter-chip interconnect to create a more disaggregated system

Optics can help on both compute and interconnect side.

Lightelligence Solutions



oMAC: Optical Multiply Accumulate Operation **oNOC:** Optical Network on Chip **oNET:** Optical Inter-Chip Networking

LIGHTELLIGENCE

Thank you