LIGHTELLIGENCE

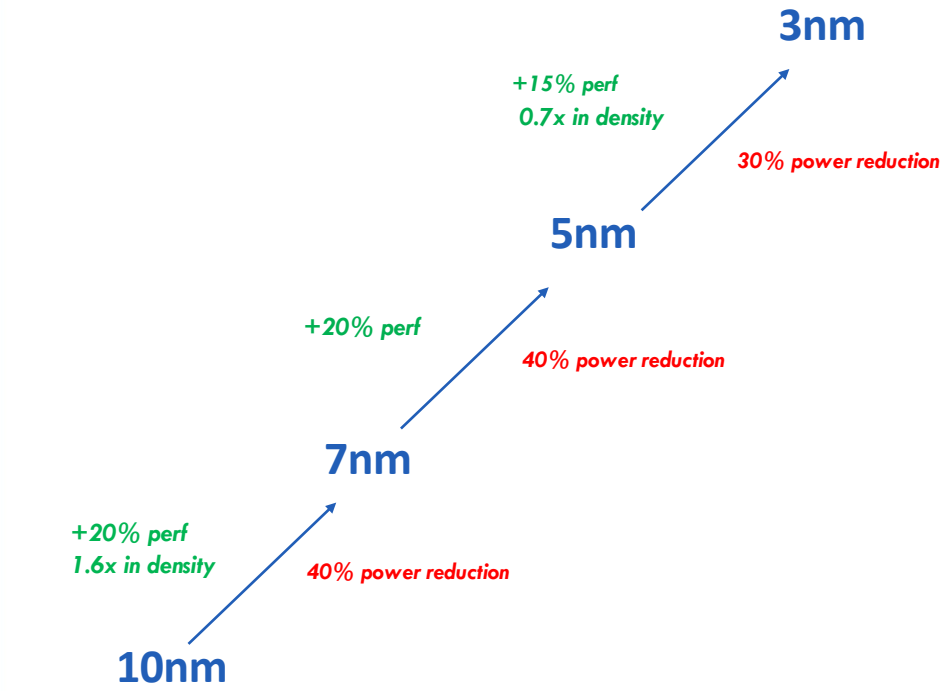
Hummingbird™
Low-latency Computing Engine

Maurice Steinman, Vice President of Engineering

Transistor Scaling Falling behind Demand

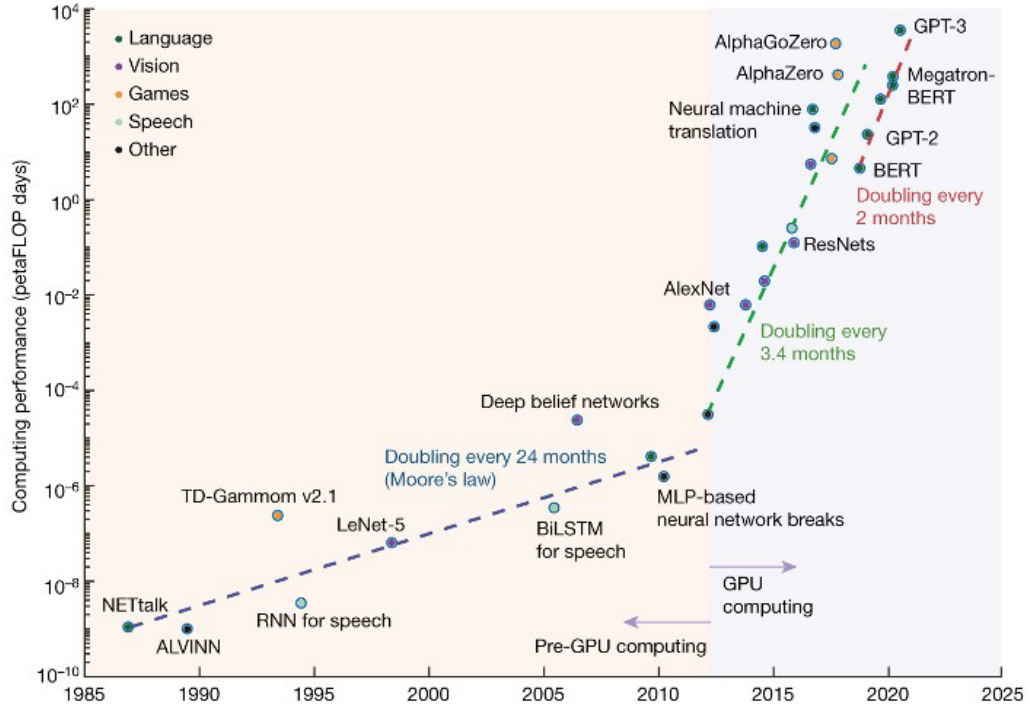
LIGHTELLIGENCE

Single Transistor Level Improvement



Source: TSMC

AI Model Computing Performance Requirement



Source: Nature

- 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 language models cost millions of dollars to train

A New Computing Paradigm

LIGHTELLIGENCE

Photonic Compute

Optical Fabric

Today's focus

Optical MAC

Process Data

Special Purpose Accelerators:

- NP-Complete Problems
- 100X faster than GPUs
- Lower TCO

Optical NOC

Share Data

Flexible Topologies:

- Low Latency Interconnect
- Greater Density/Less Power
- Simplifies SW Development

Optical NET

Transfer Data

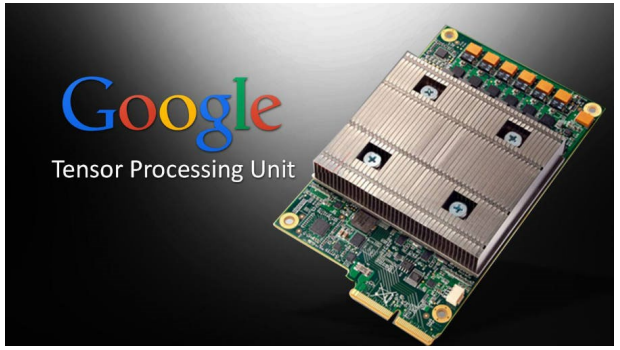
Access to more Memory:

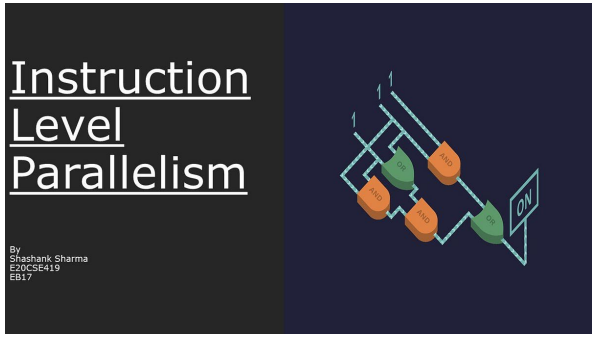
- High Bandwidth Memory Pools
- Rack to Rack Connections
- Lower cost of Model Training

Performance Improvement: Architecture Innovations

Architecture Innovations:

- Domain Specific Architecture (DSA)
  - Tensor Processing Units (TPU)
- Non-Von Neuman, Disruptive Architecture
- Instruction-Level Parallelism
- Transistor Efficiency Improvements
- Increased On-Chip Memory



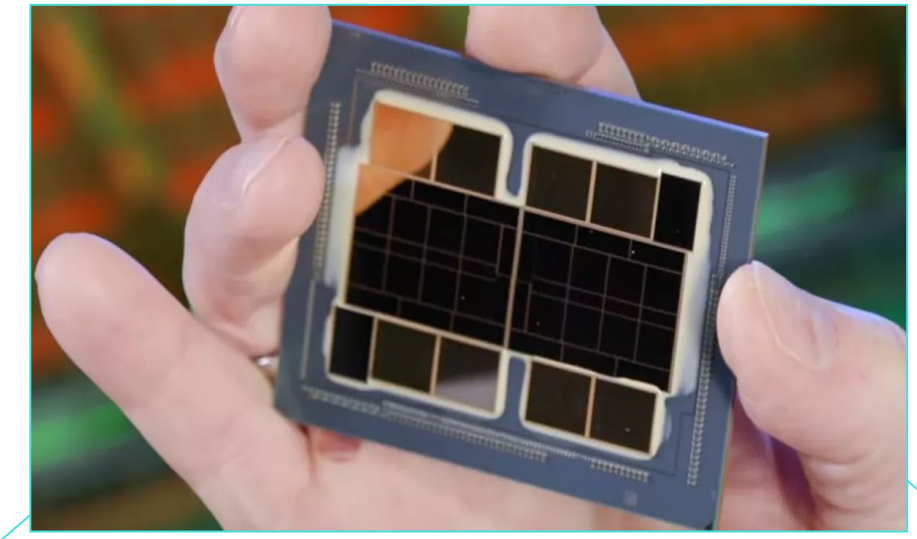


Many sacrifice versatility for performance

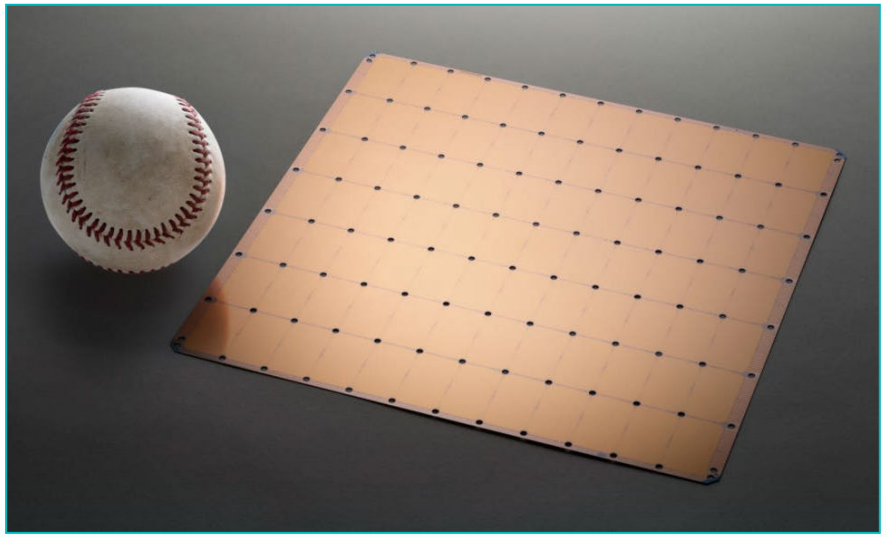
Performance Improvement: Enlarge Silicon Area

LIGHTELLIGENCE

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



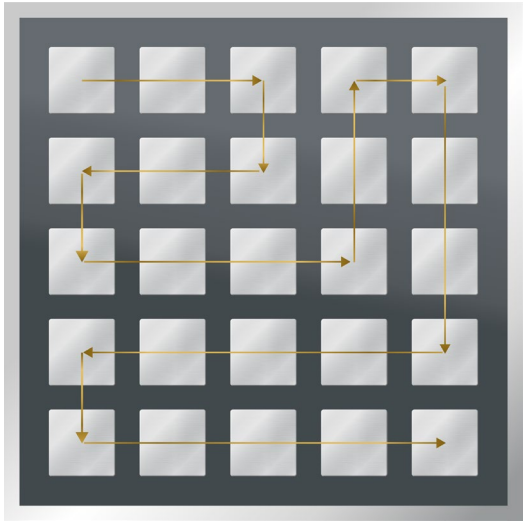
Intel Ponte Vecchio
1,200 mm^2

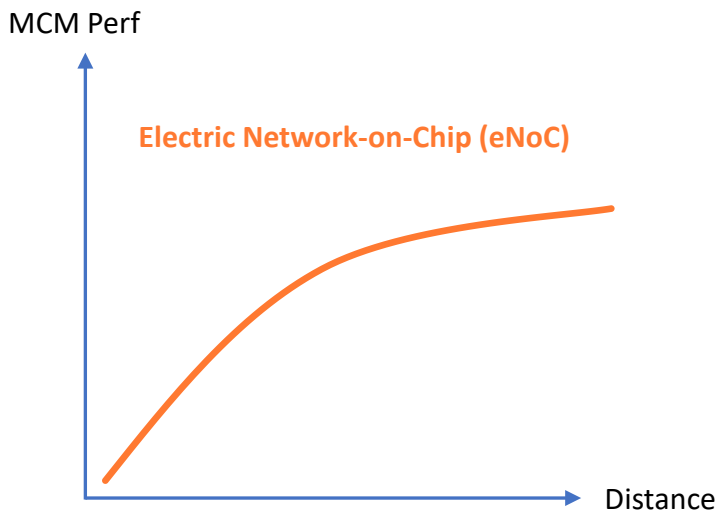


Cerebras Wafer Scale Engine
46,225 mm^2

Inefficient Scaling of Performance

Mapping of workloads to a mesh architecture

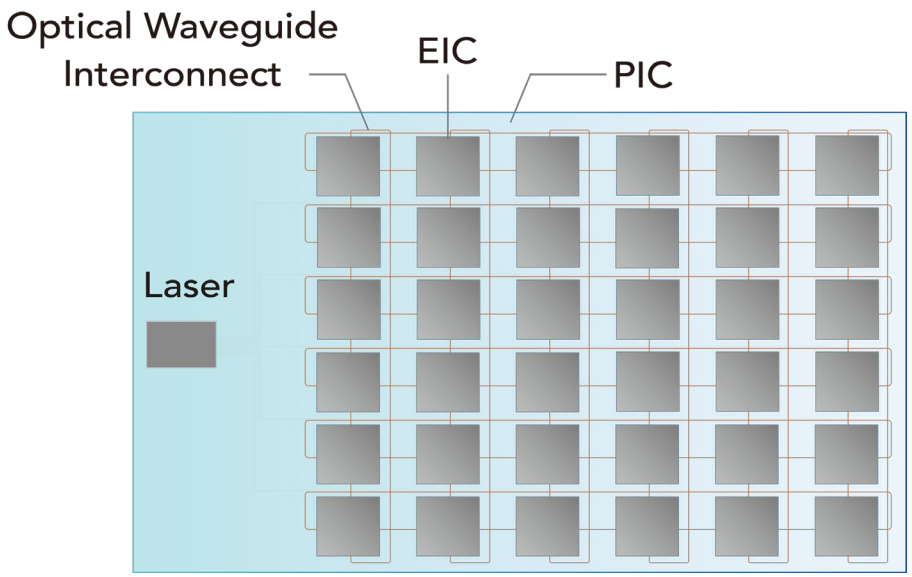




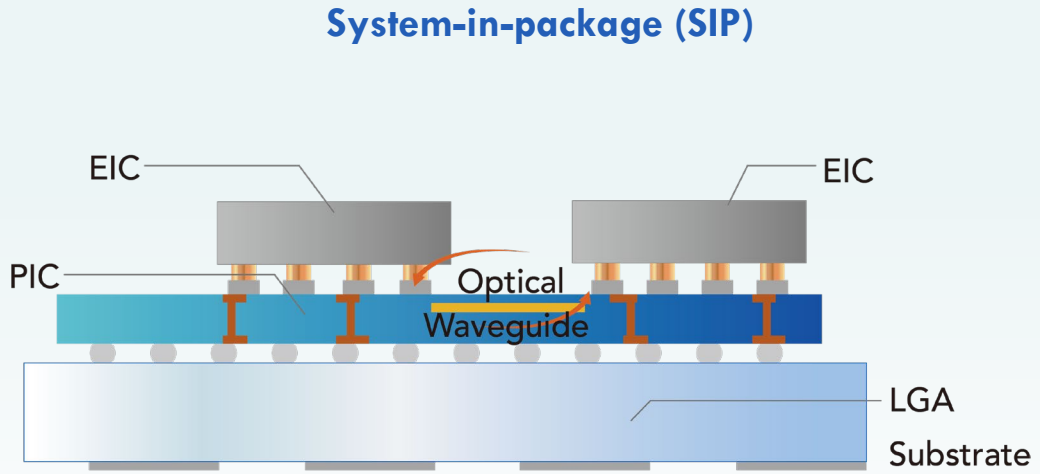
- Electrical signals attenuate over distance
  - Nearest-neighbor only communication
- More hops result in increased latency
- Difficulty in programming
  - Inefficient utilization

A better interconnect solution is needed for large MCMs

Optical Network-on-Chip (oNOC)

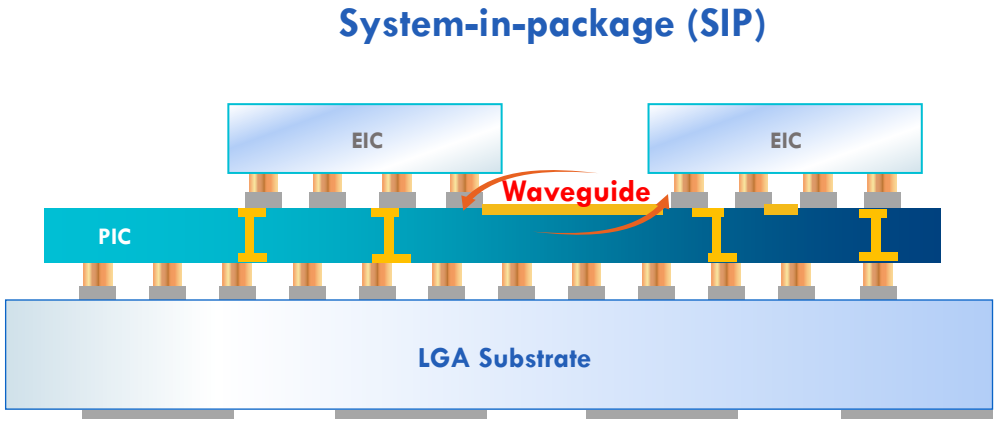


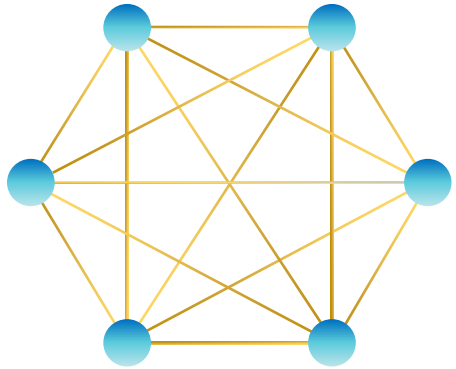
Cross - section



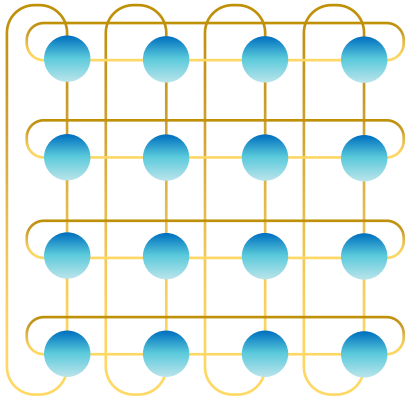
- Optical signal attenuation is small at wafer scale
- Power and latency are independent of distance
- Photonic integrated circuit (PIC) as active interposer to transmit data between Electronic integrated circuits (EICs) using waveguides

Optical Network-on-Chip (oNOC)





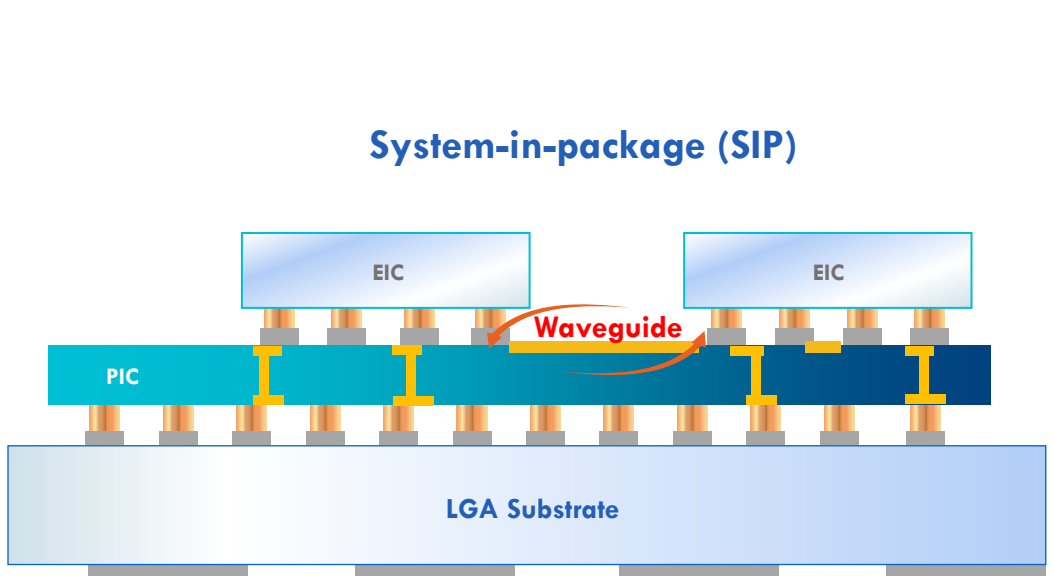
Full mesh

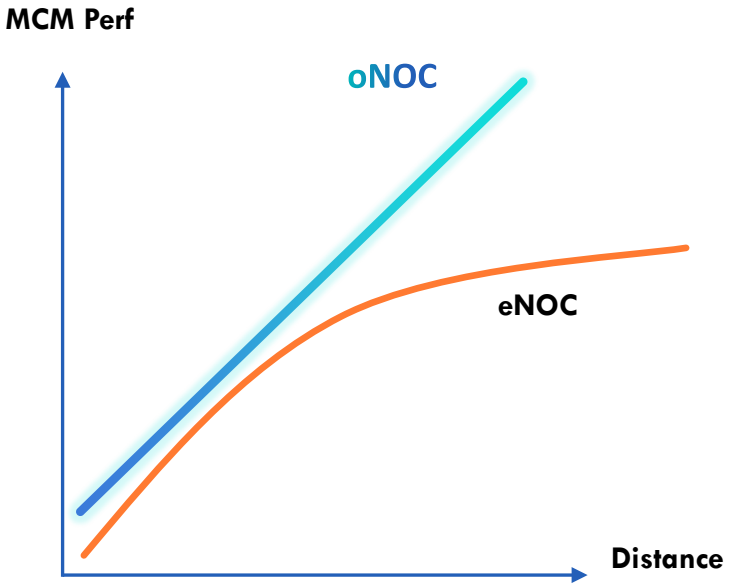


2D Torus

- Optical networking enables diverse topologies
- Inter-chiplet connectivity no longer limited to nearest neighbors

Optical Network-on-Chip (oNOC)

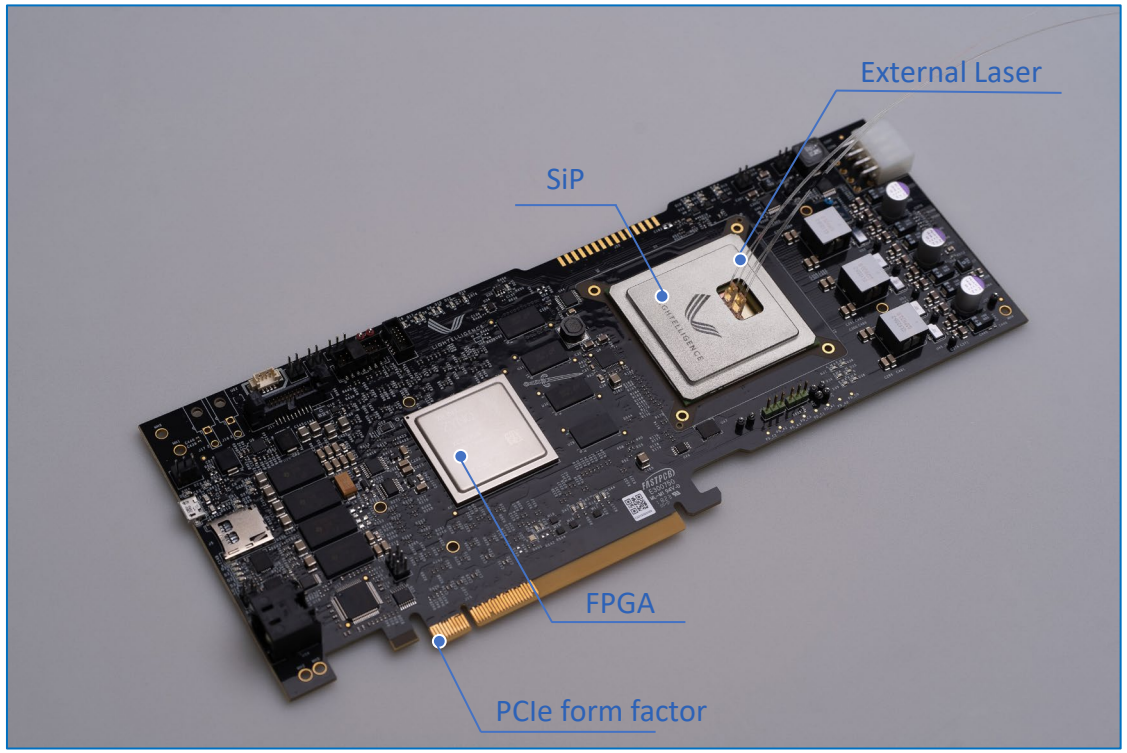




- Mapping workloads to hardware becomes more efficient with flexible network topology
- Close to linear scaling of MCM performance

Hummingbird™: Superior Latency Enabled by oNOC

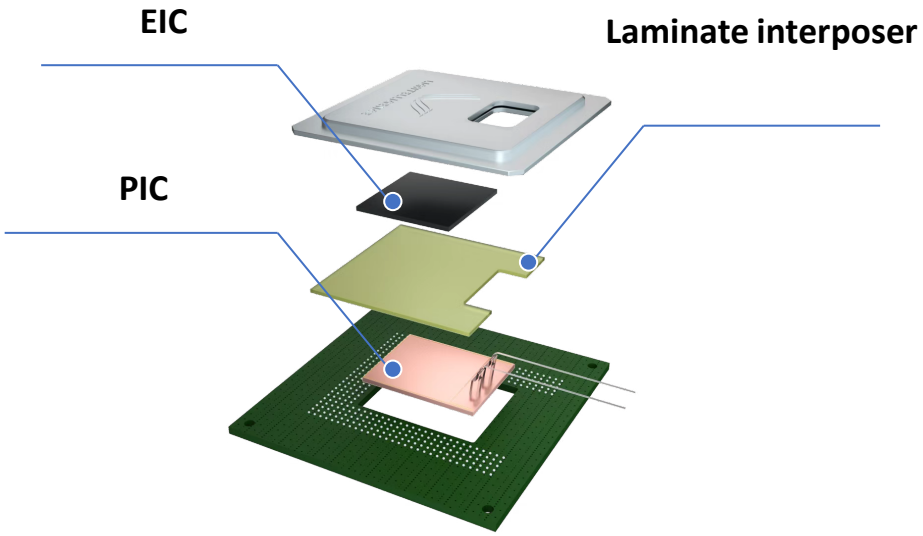
First oNOC-powered system to run commercial workloads



Support AI inference and other applications

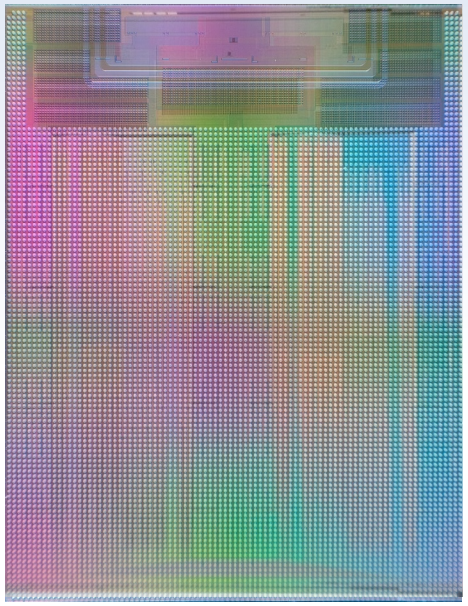
- oNOC All-to-All broadcast
- Ultralow latency data transfer
- Lightelligence SDK

First Computing SiP with oNOC

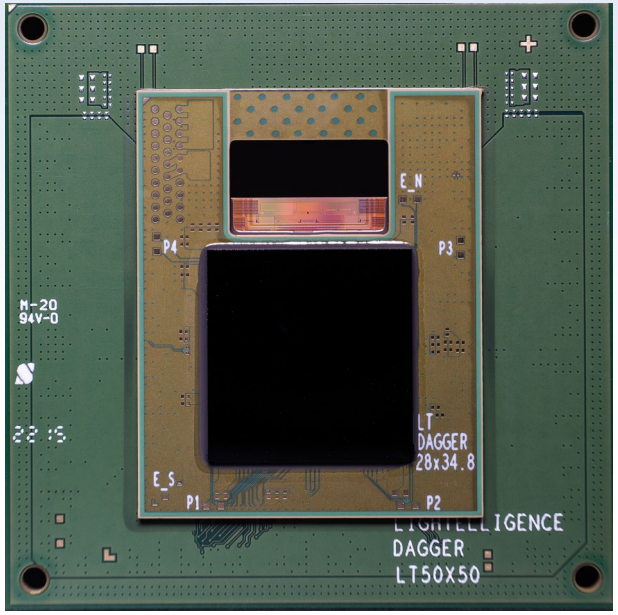


Interposer provides:

- Path from EIC to package substrate for power delivery and external I/O, while maintaining dense connectivity between EIC and PIC

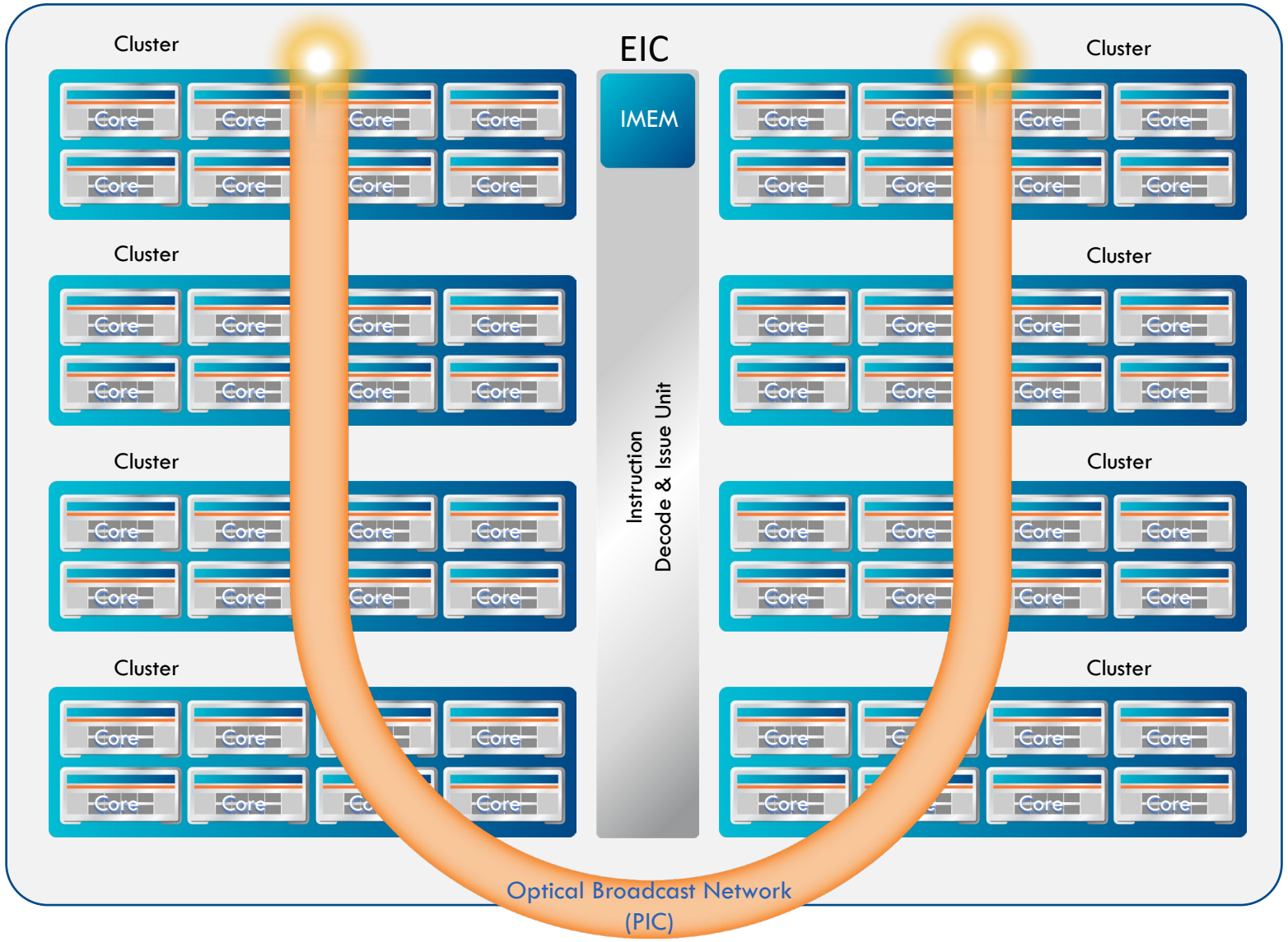


Photonic integrated circuit (PIC)



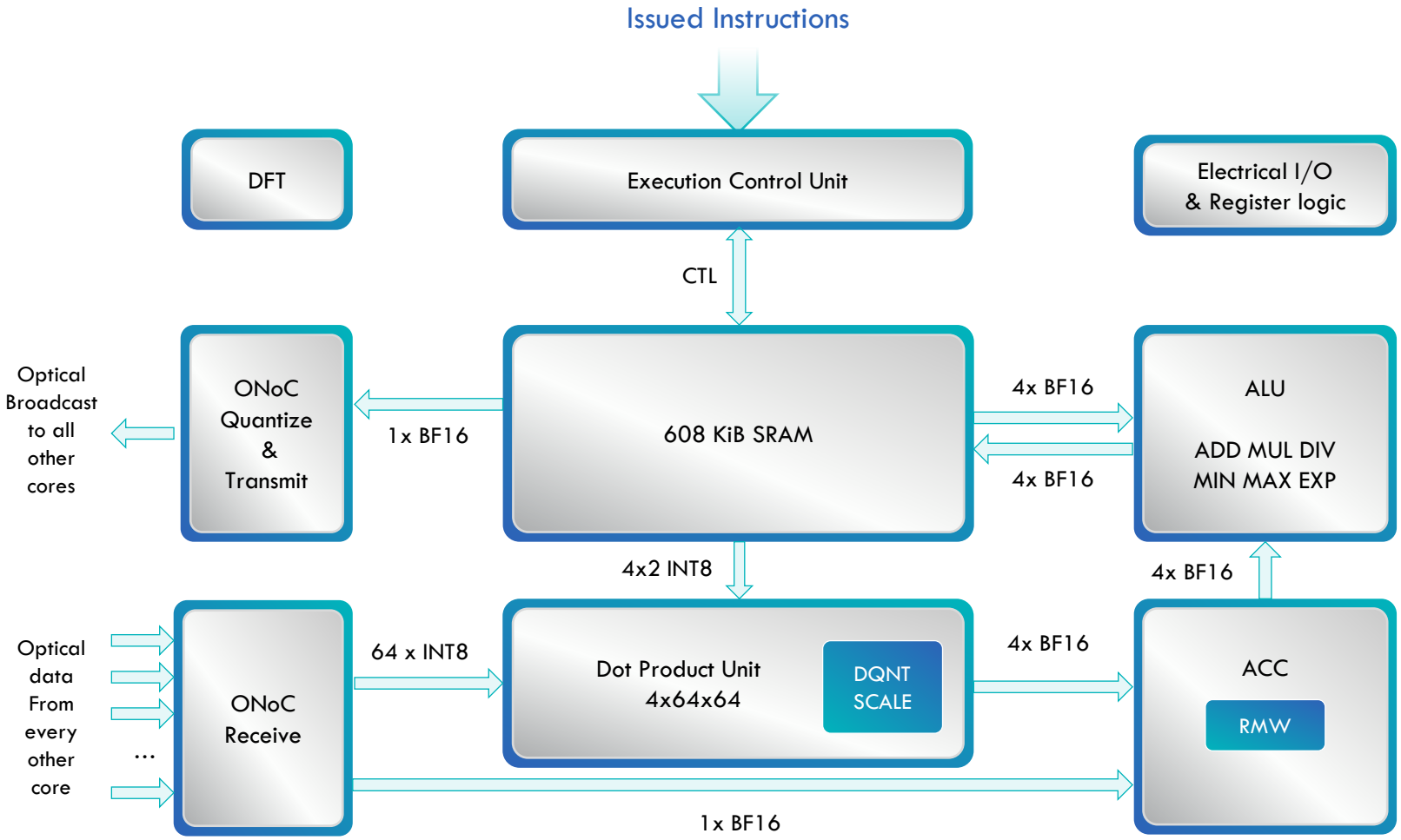
System-in-package

Hummingbird™ SiP Architecture



- SIMD architecture with custom ISA
  - Central instruction unit issues to each of the 64 cores in the EIC in parallel
- All-to-all broadcast
  - oNOC transports data from each core to all other cores in the EIC
  - "U" shape enables all-to-all connectivity without waveguide crossings

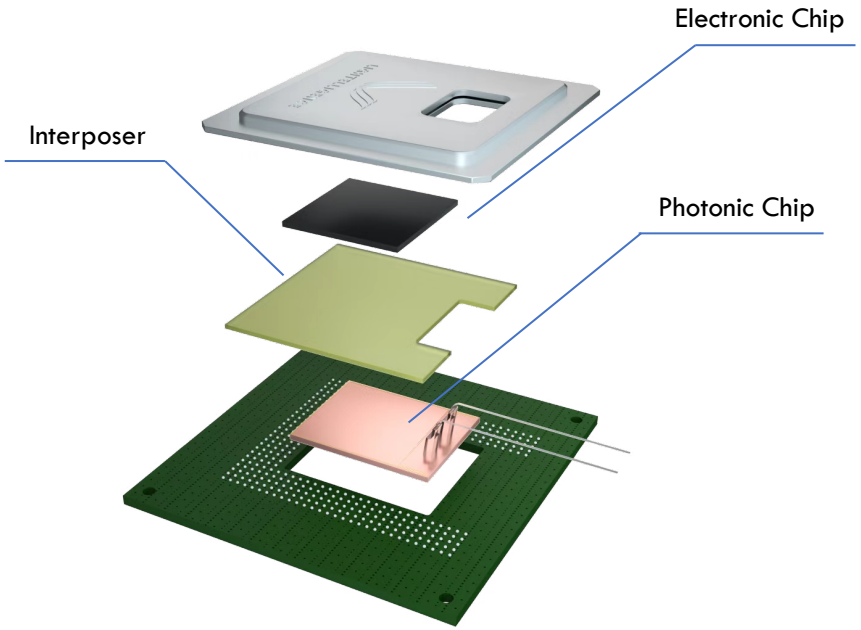
Hummingbird™ Core Microarchitecture



- Bfloat16 data storage format+quantization/ dequantization
- 4-wide INT8 64x64 dot product unit with accumulator
- ALU for scalar and vector functions
- 608 KiB SRAM
- Optical broadcast interface to send and collect results with all other cores

Hummingbird™ Design Metrics

Architecture	Hummingbird™
Compute Cores	64
Precision	INT8
On-Chip Memory	38 MiB SRAM
ECC	SECDED
System Interface	x4 PCIe Gen3
System Memory	2 GB DDR4 SDRAM
Form Factor	Full Length, Dual Slot PCIe
Thermal Solution	Passive
Compute API	LT-SDK
Photonic Transmitters	64
Photonic Receivers	512



Electronic Chip

- TSMC N28 HPC+
- XY: 17mm X 16.5mm
- 500M Transistors
- 0.75 km of wire length

Photonic Chip

- IMEC iSiPP 200
- XY: 21.3mm X 16.5mm
- >20m of silicon waveguides
- 580 Photodiodes
- 64 Data Modulators

Hummingbird™ System Design

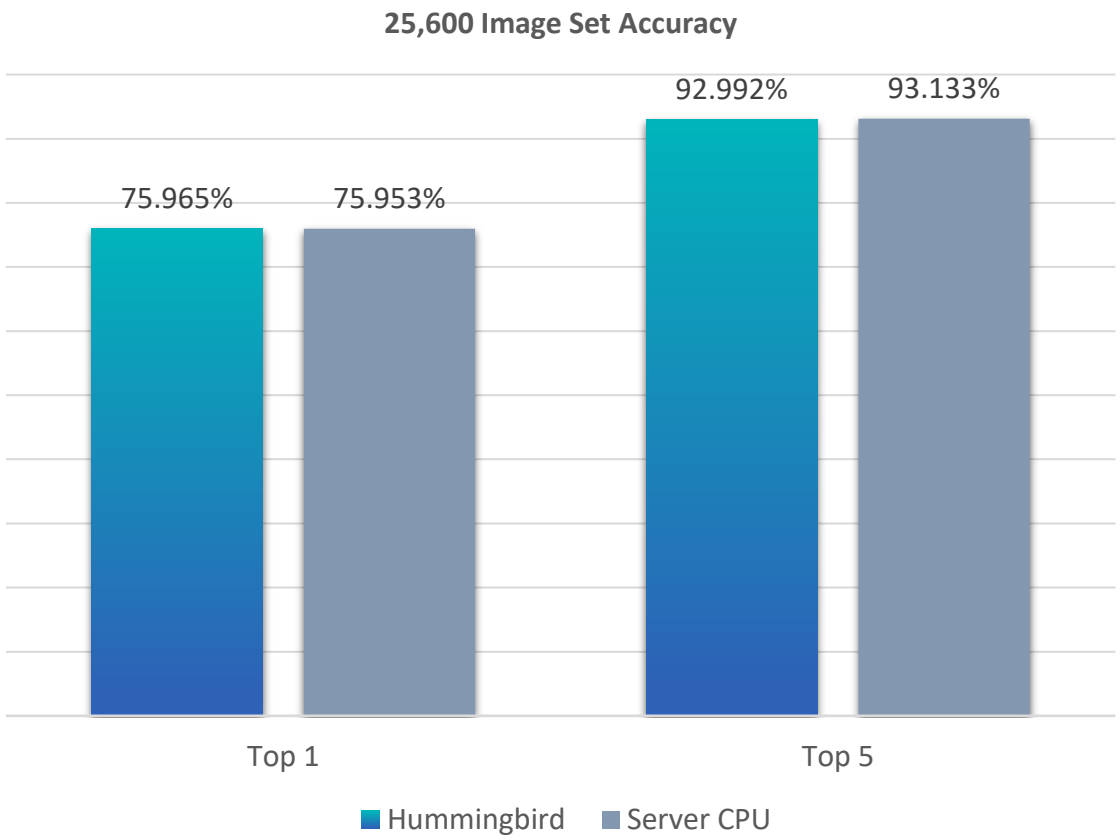
64 Core Distributed Processing Engine
All-to-All Optical Network-on-Chip

Industry Compliant PCIe Form Factor

Ease of adoption with custom SDK and integration ready hardware

Metrics

ResNet50 ML Inference run on Hummingbird using LT-SDK



Power and Performance

- oNOC latency (core-core):
  - $TX_{[Digital+Analog]}$: 3 cycles
  - Transport (pS) $_{[Optical]}$: 52/787/407 (min/max/mean)
  - $RX_{[Analog+Digital]}$: 3 cycles
- oNOC data rate: 1Gbps
- Card max Power (W): 65 (35 SiP) (@Digital F_{MAX}=1GHz)
- ResNet50 Singe-Image Latency (ms): 20 (15.9 SiP)

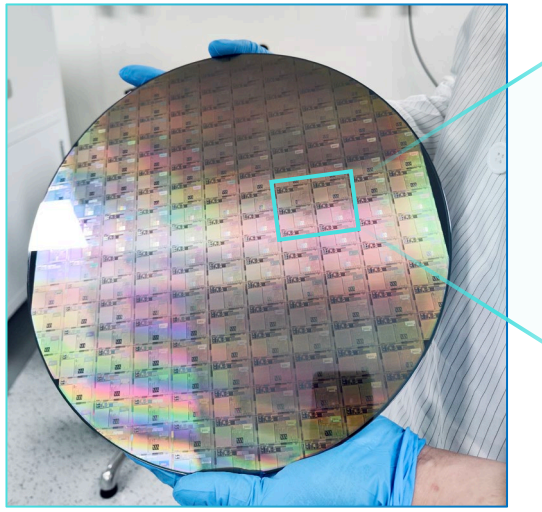
Looking Ahead

- 3D packaging utilizing Through-Silicon Vias (TSV)s
- More advanced node implementation for EIC
- Best in class, customer-specific designs

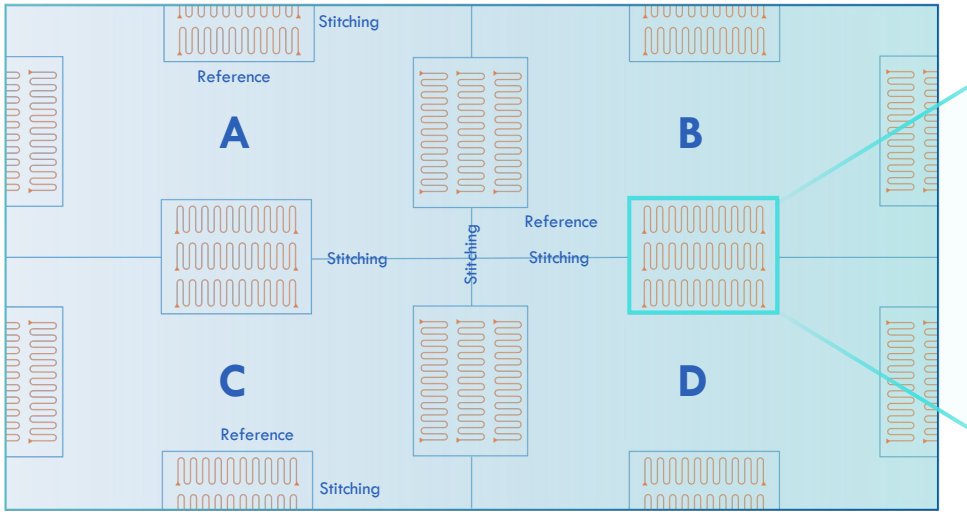
Looking ahead: Reticle Stitching

LIGHTELLIGENCE

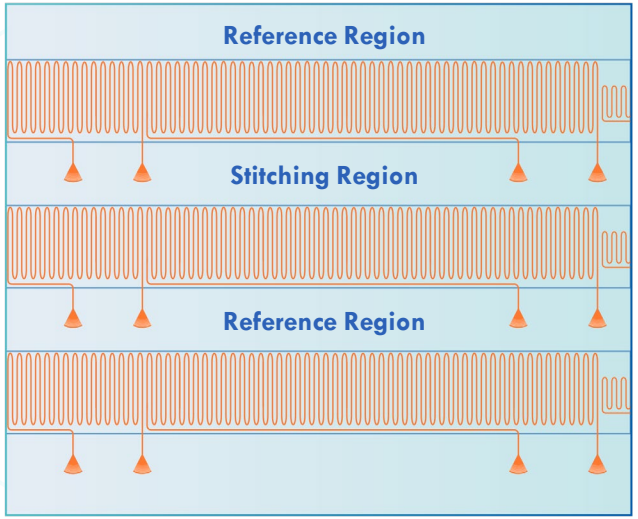
Stitching Wafer



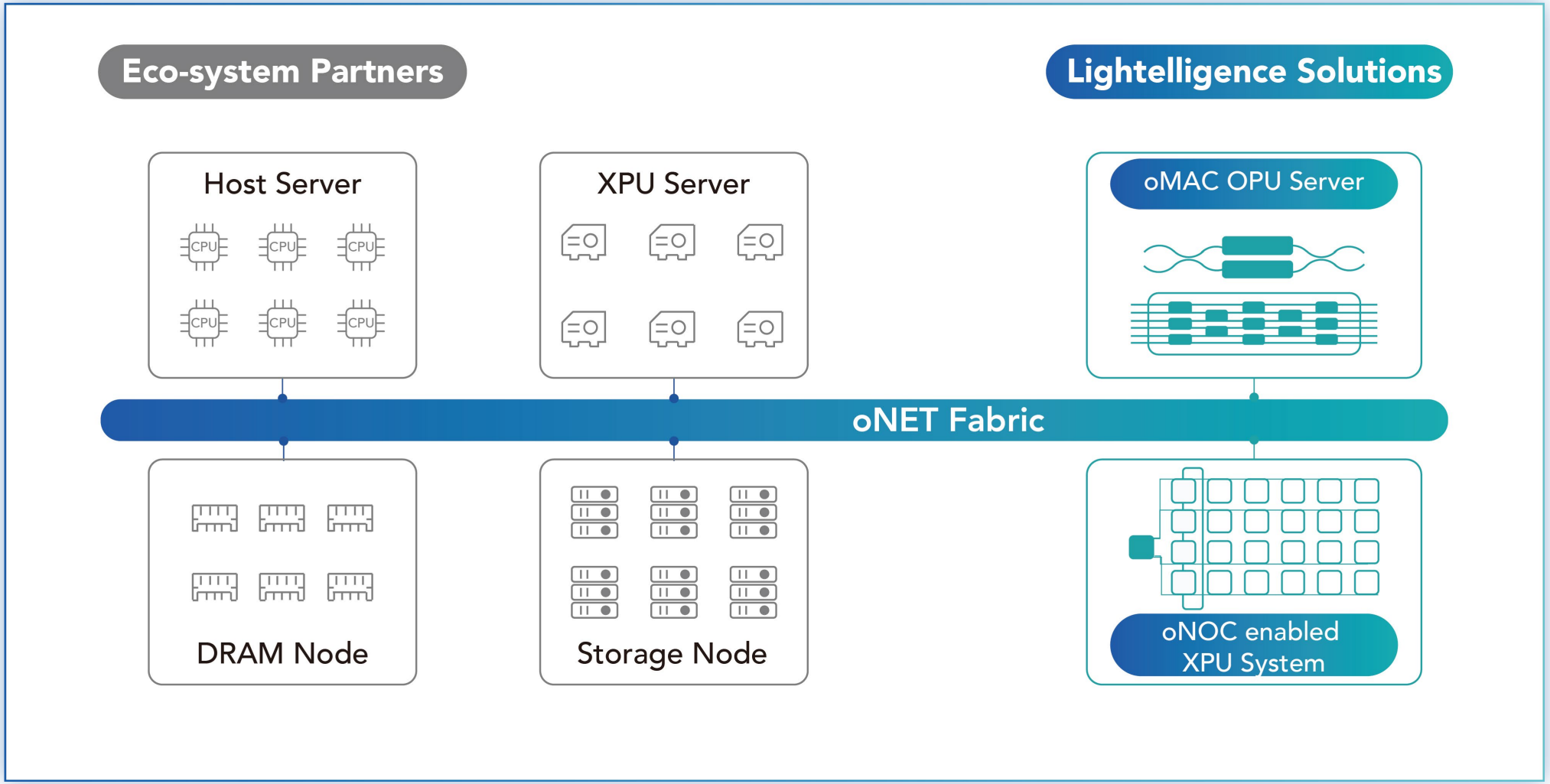
Stitching Design Illustration



Stitching Image



Lightelligence Solutions



oMAC: Optical Multiply Accumulate Operation

oNOC: Optical Network on Chip

oNET: Optical Inter-Chip Networking

Endnotes

LIGHTELLIGENCE

info@lightelligence.ai