OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

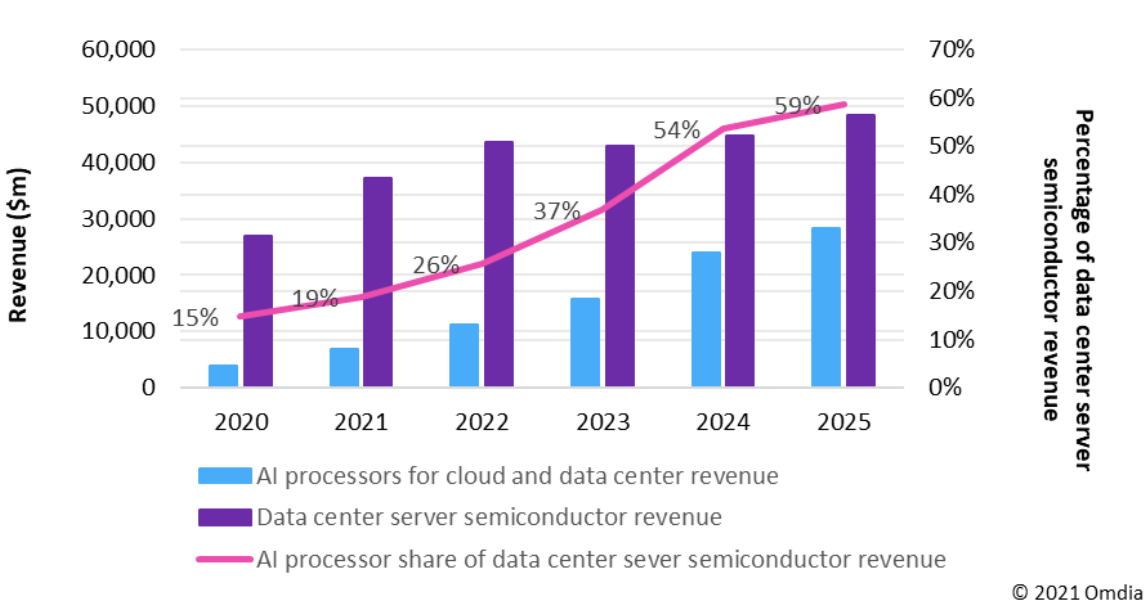
November 8, 2021 | San Jose, CA

AI Co-Design Opportunities in OCP

Matt Bergeron
Yang Seok Ki
Matt Short

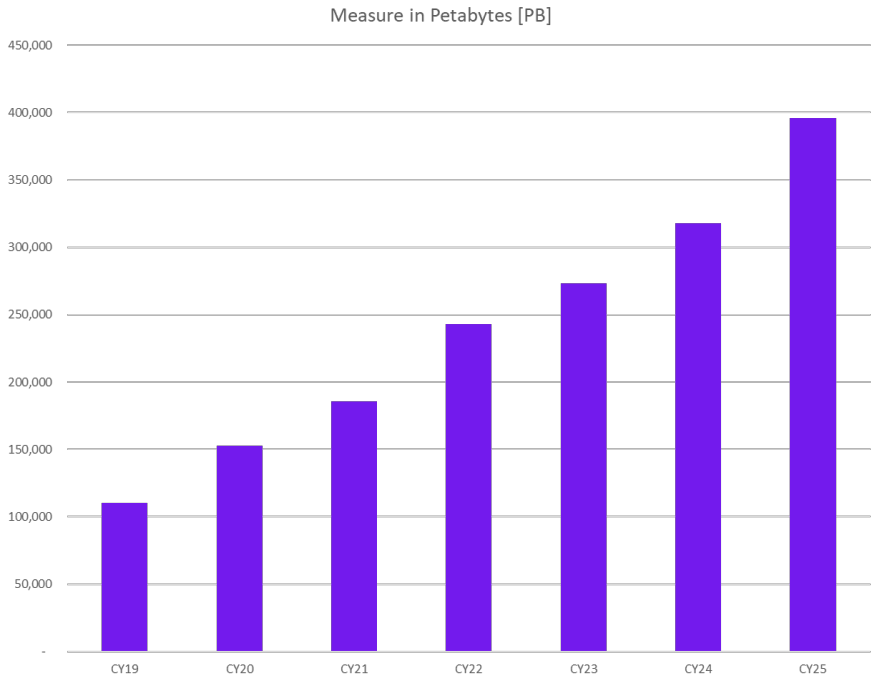
Compute Spend is Changing Dramatically

- Let's start by setting the stage
  - Why is this so important?
  - Why do I need to take action?
- If "AI" was a country
  - Today top #100, between Serbia and Lithuania
  - By 2025 top #50, comparable to Greece
- By 2025, the majority of data center compute spend will be for AI
- Great! But…why is there is always a but?



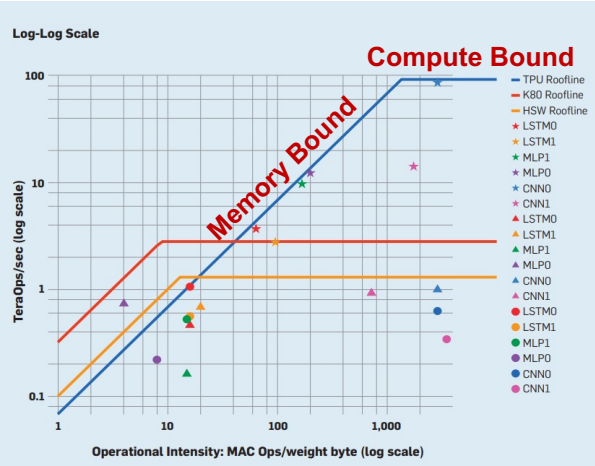
Data Drivers - AI and ML Workflows

- How we segment the problem tells us a lot
  - Spans 28 Major industries
  - 14 Major Applications
  - 208 Use case / technology types
  - 5 major regions
  - 1000+ vendors (HW, SW, OEM, Sis, SPs, etc)
- The need for Industry catalysts
  - APIs, Digital Twins, Tools of all sorts
  - Most important – Industry Collaboration
  - OCP is here to play our part!

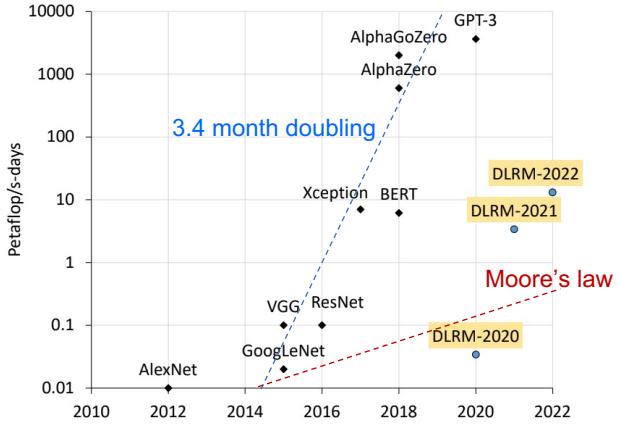


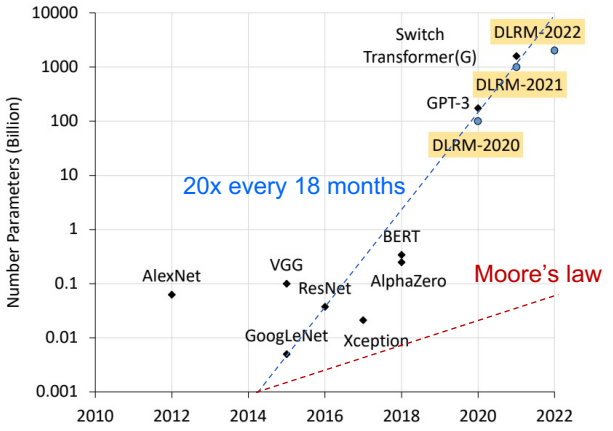
AI Resource Requirements & Trend

- AI models grow faster than technology advances



Source: N. P. Jouppi, COMMUNICATIONS OF THE ACM, Sep. 2018

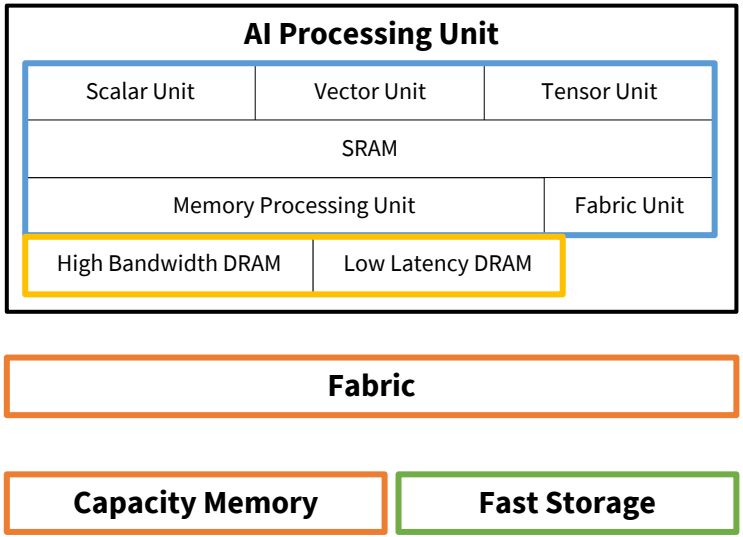




Source: D. Mudigere, https://arxiv.org/abs/2104.05158

Inference System

Challenges



Potentials

AI/Watt High throughput

- AI accelerator
- Analog/Photonic chip
- **FTS presentation**: Accelerating DLRM on an OCP M.2 Accelerator

Capacity Memory sufficient for huge models

- Tiered memory, memory pooling
- SSD
- **FTI SDM Work Stream**

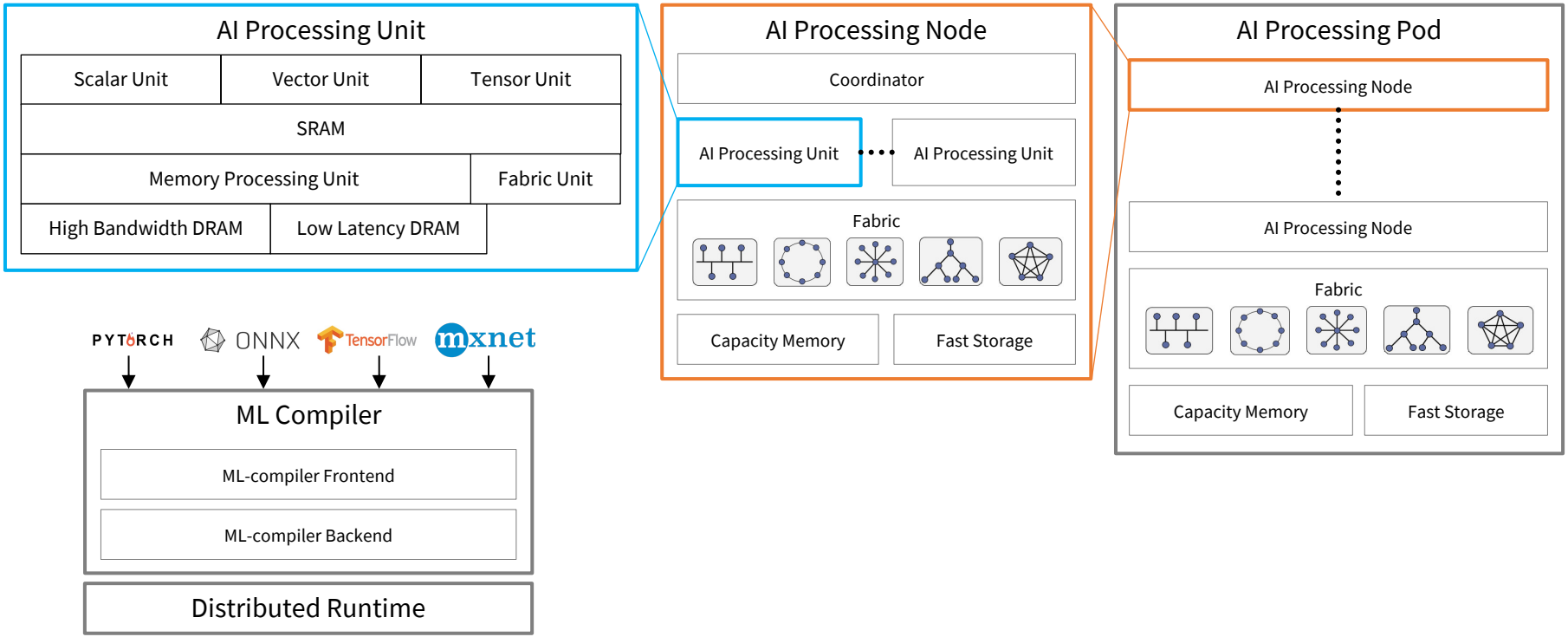
Bandwidth High effective bandwidth

- High bandwidth IO
- Configurable network

Latency Real-time SLA (Service Level Agreement)

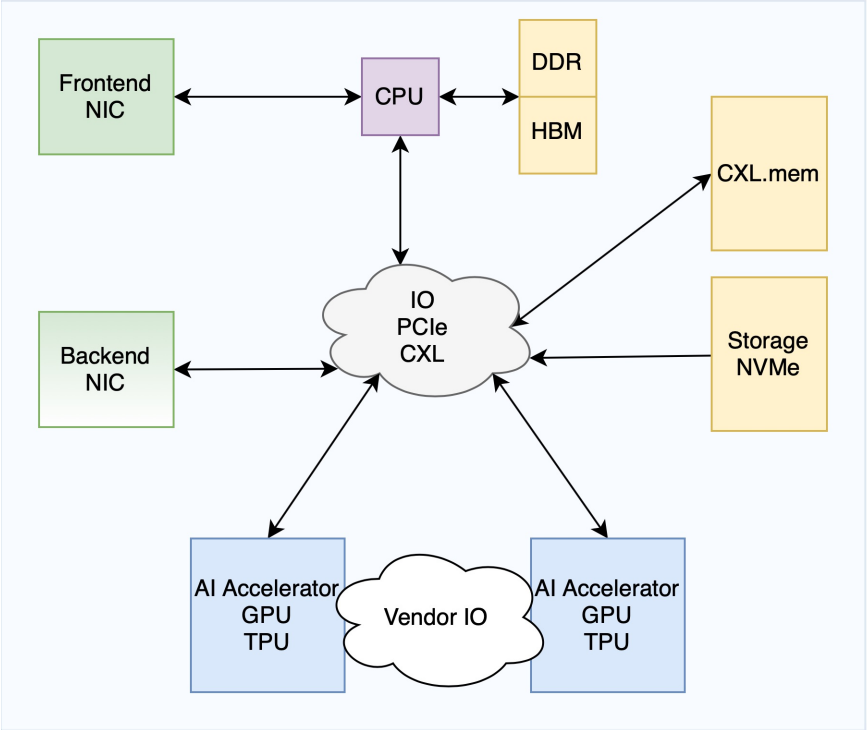
- High IOPS SSD with short tail latency
- **FTS poster**: Can NAND Flash SSDs Meet the Needs of Recommendation Models for Inference?

Inference HW and SW Architecture



Training Systems

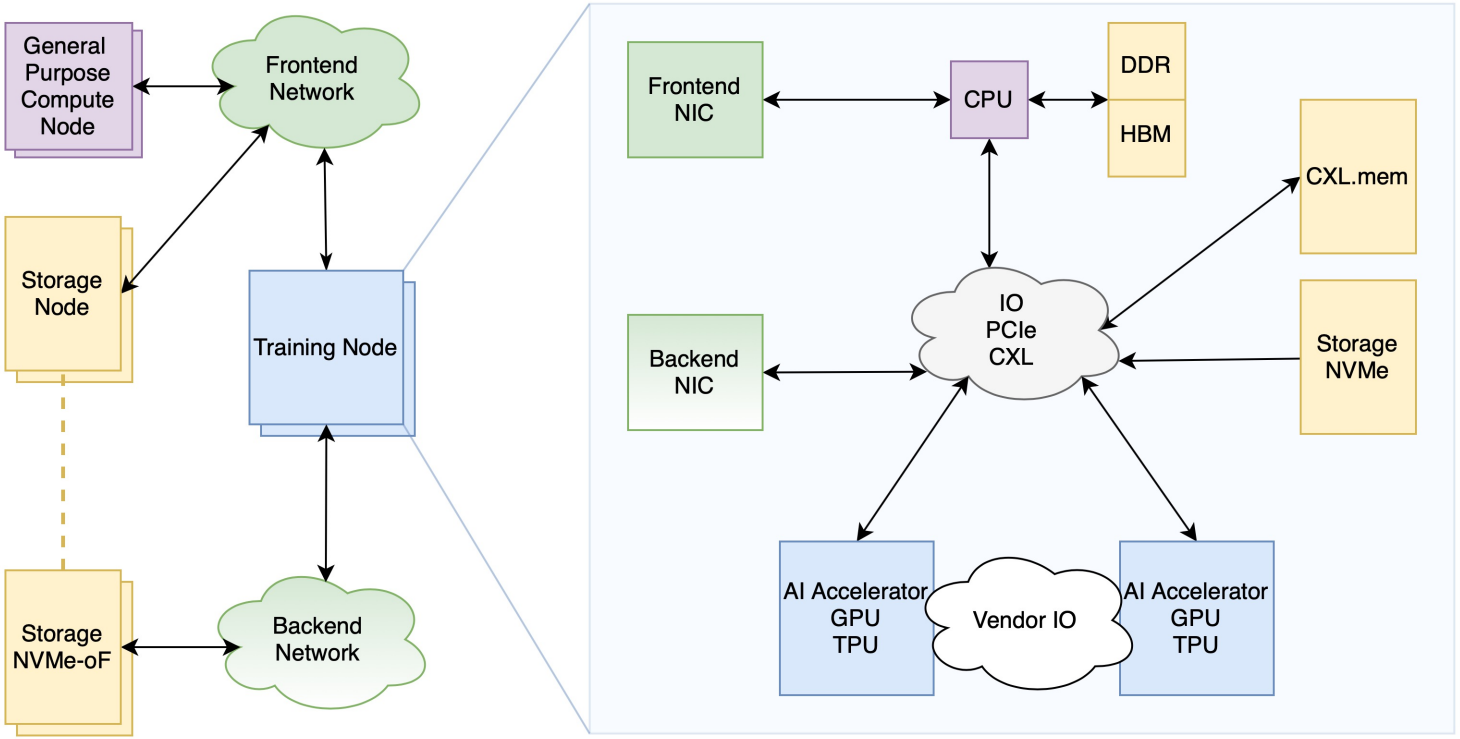
- Compute elements
- IO networks
- Storage types
- Inter-node networks



Inter-node Networks

Ethernet

HPC,
Ethernet,
Infiniband



Collaboration

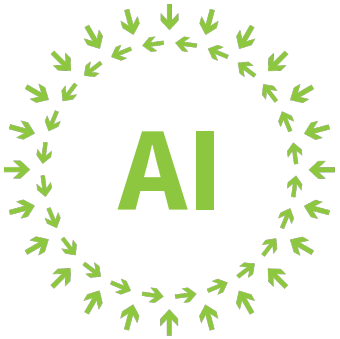
- Modelling
  - Simulation
  - Abstraction (Glow)
- Democratization
  - Open vendor IO bus
  - Physical partitioning

Actions

- Contributors
- Software
- University collaboration
- Industry collaboration

- **Afternoon session**

Goal



Formal OCP project group in 2022


OCP
TOTAL SOLUTIONS

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA