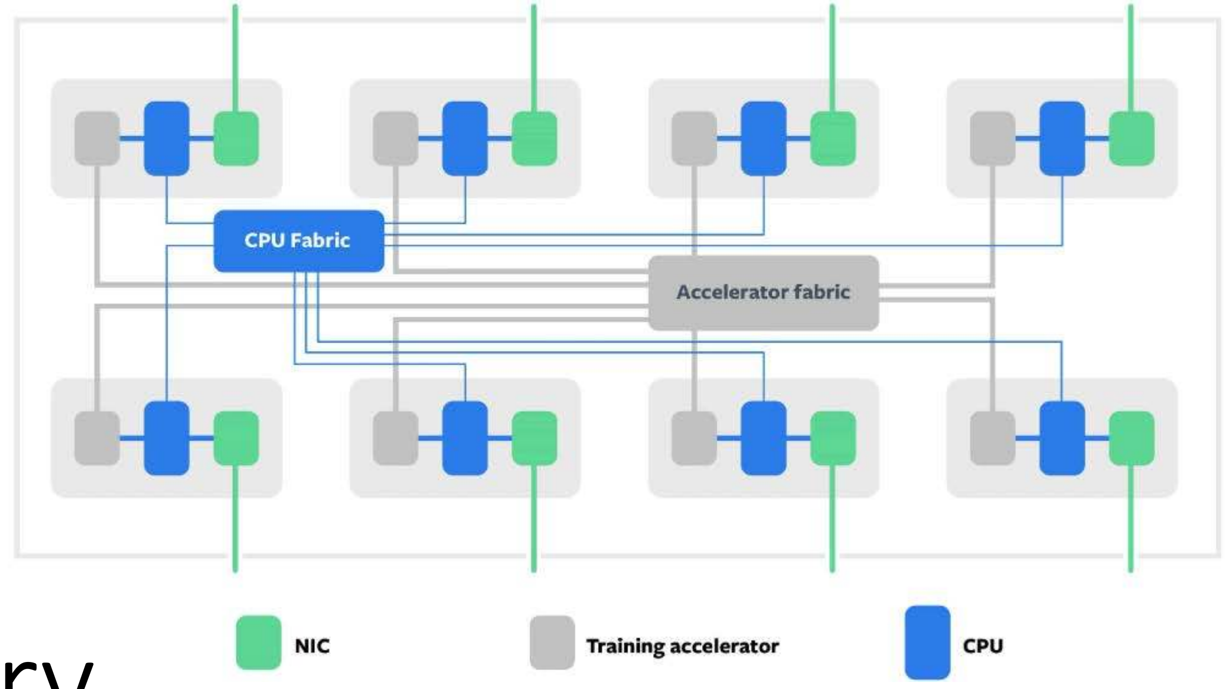
facebook

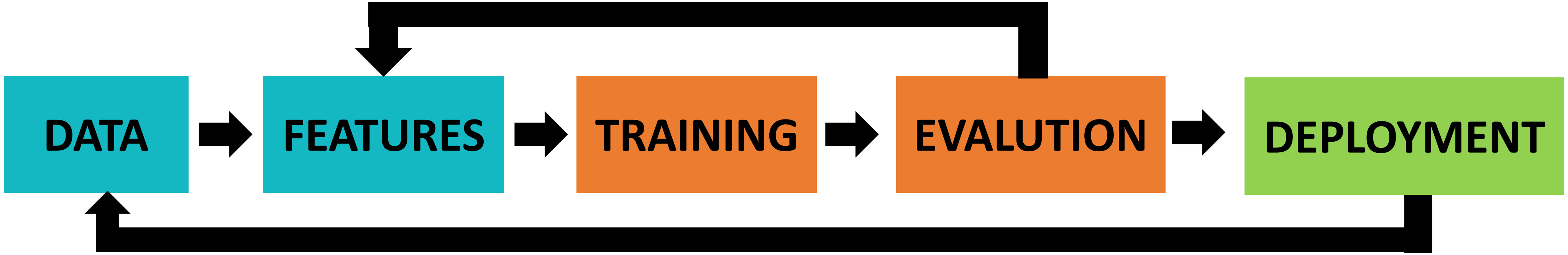
Zion: Facebook Next-Generation Large Memory Training Platform

Misha Smelyanskiy

Hot Chips 31, August 19, 2019



The Growth of ML at Facebook

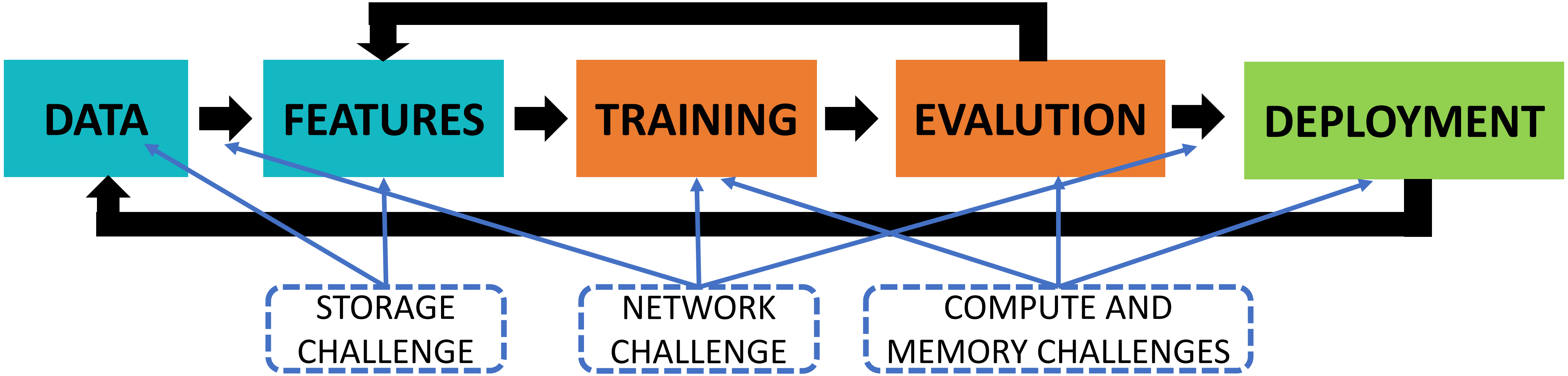


- **ML pipeline data growth**
  - Usage in **2018**: **30%**
  - Usage **today**: **50%**
  - ML data growth in **one year**: **3X**

- **12-month ML Training growth**
  - # of ranking engineers: **2X**
  - Workflows trained: **3X**
  - Compute consumed: **3X**

facebook

Training Infrastructure Challenges



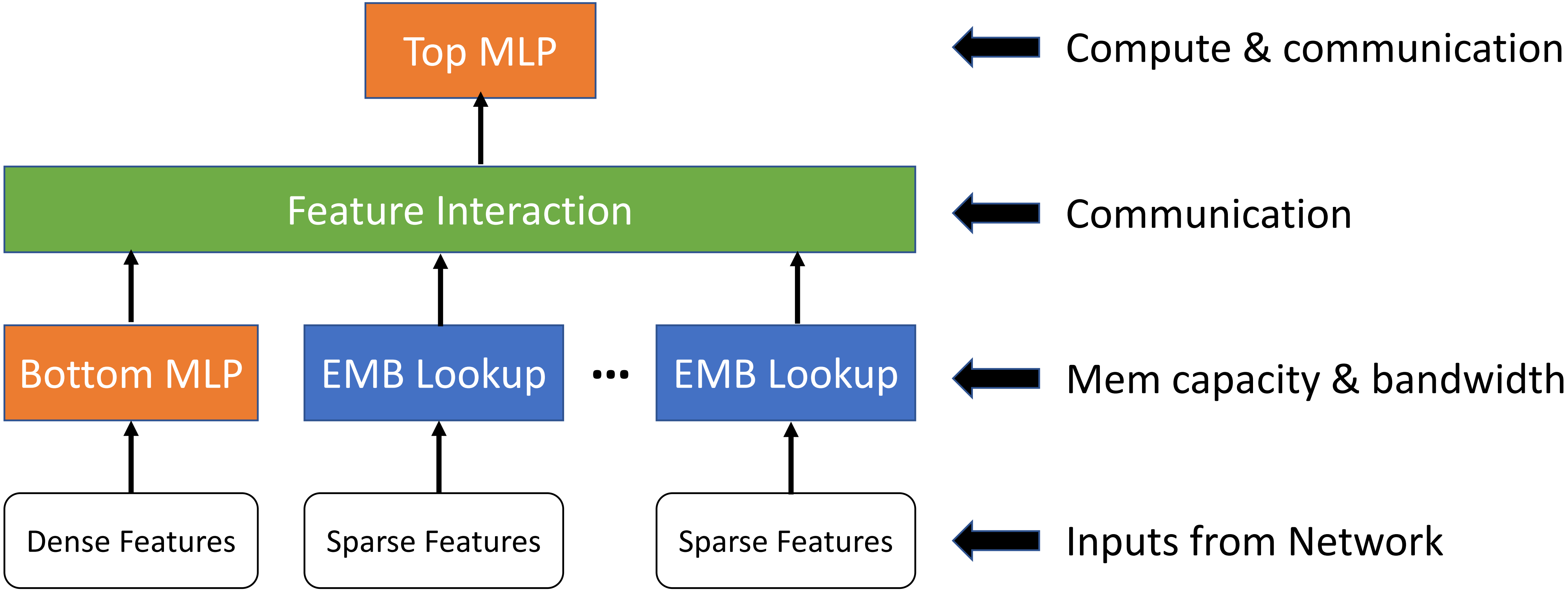
- Strains memory, compute, storage, and network
- ML engineers expect developer efficiency and flexibility
- **Motivated SW/HW co-design of training platform**

Major AI Services @ Facebook

- Ranking and recommendation
  - news feed, and search
- Computer vision
  - image classification, object detection, and video understanding
- Language
  - translation, content understanding

- Recommendation models are among most important models

Deep Learning Recommendation Models



Top MLP

Feature Interaction

Bottom MLP

EMB Lookup

…

EMB Lookup

Dense Features

Sparse Features

Sparse Features

Compute & communication

Communication

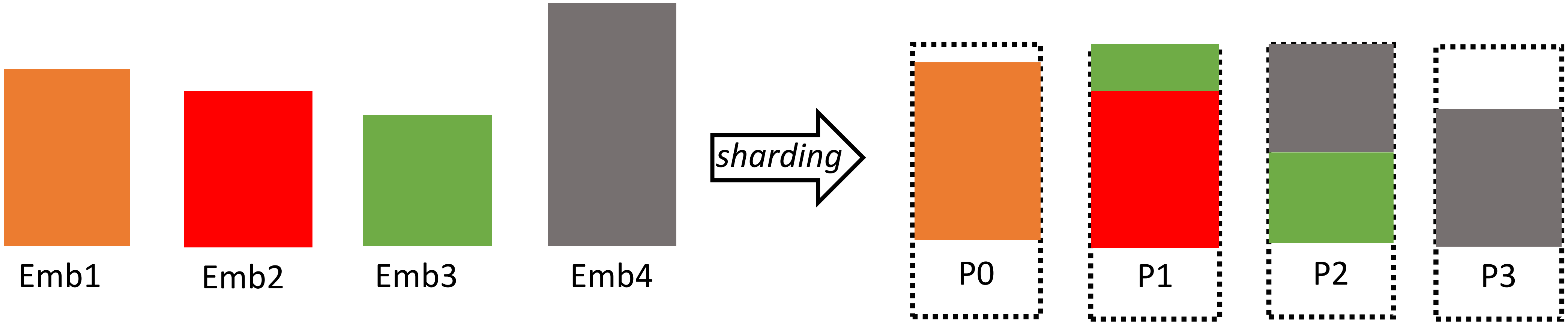
Mem capacity & bandwidth

Inputs from Network

- Open-sourced as a deep learning recommendation model benchmark

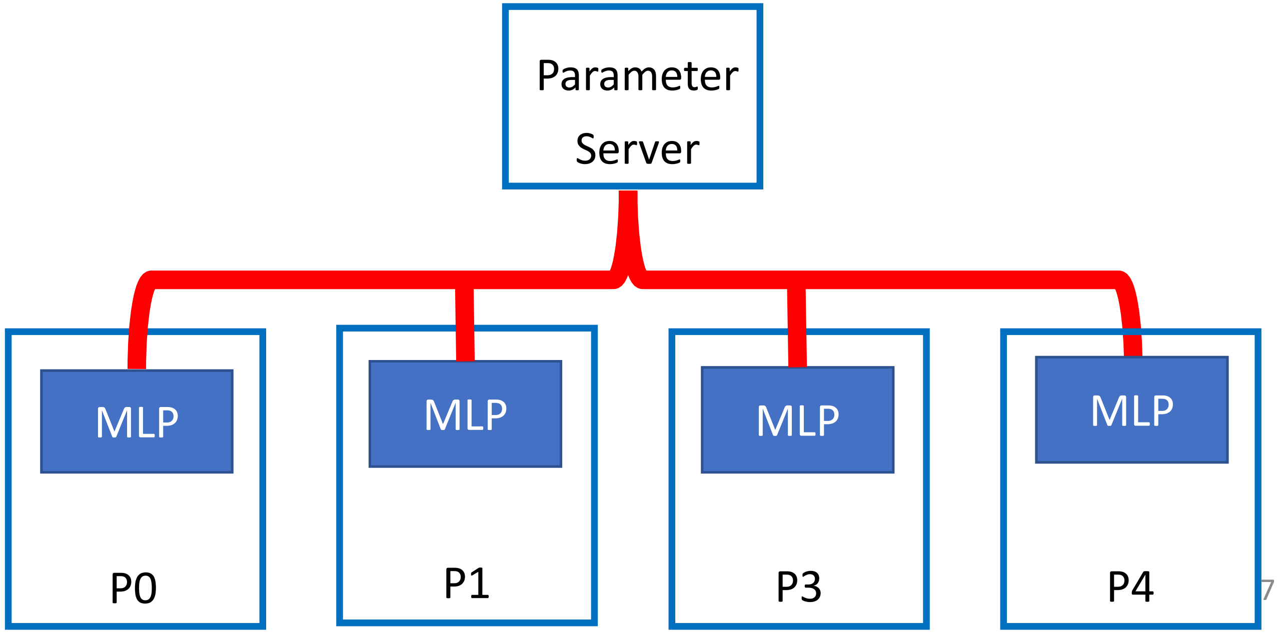
Training Embedding Tables

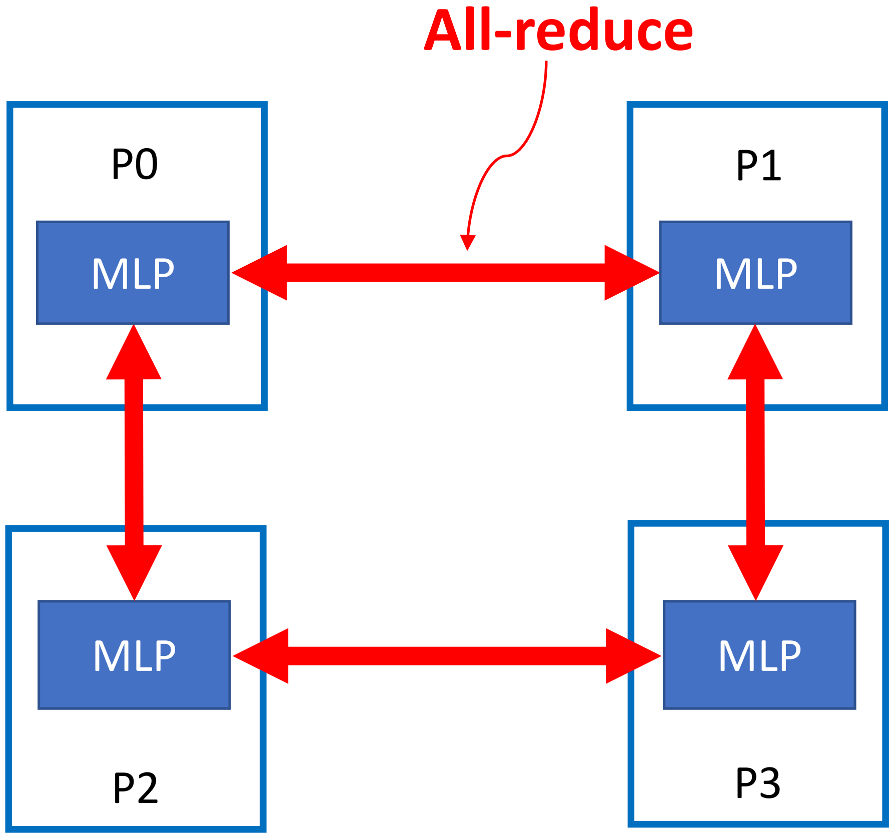
- Very large embedding tables – O(10+) GBYTES
- Low arithmetic intensity, irregular memory accesses
- Model Parallelism
  - Map embedding tables to different compute devices
  - Shard to balance out utilization given memory constraints



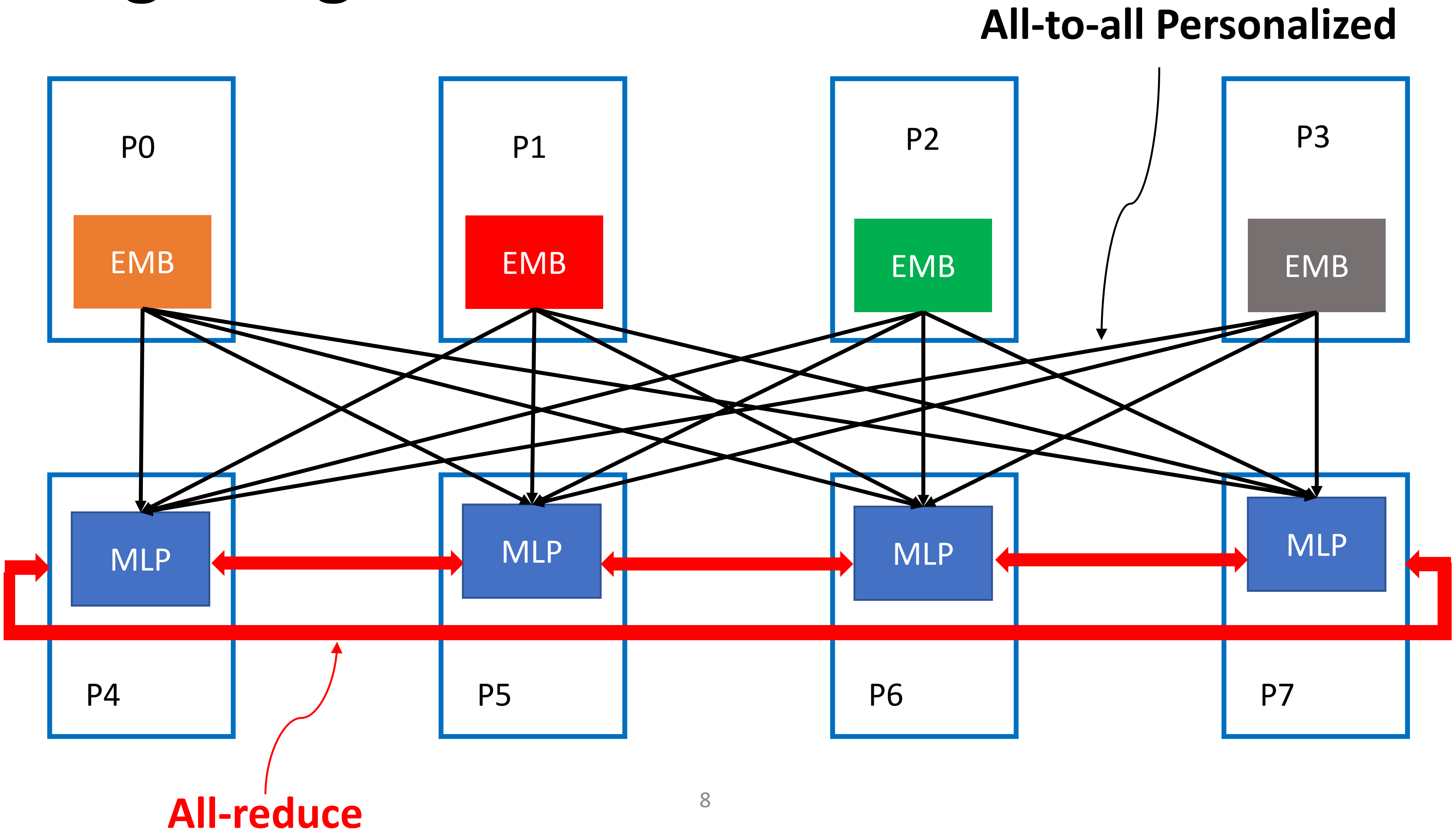
Training MLP

- Parallelism: model or data
- Updates: asynchronous or synchronous (via all-reduce)
- Dense regular compute, tall-skinny GEMMs



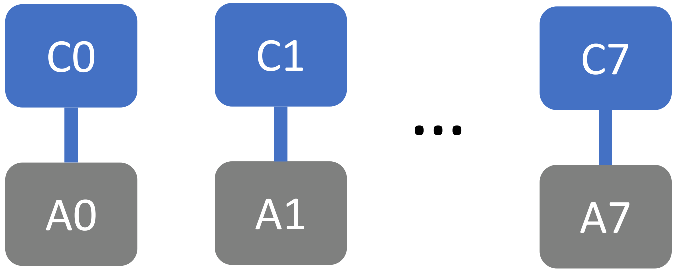


Putting it together



facebook

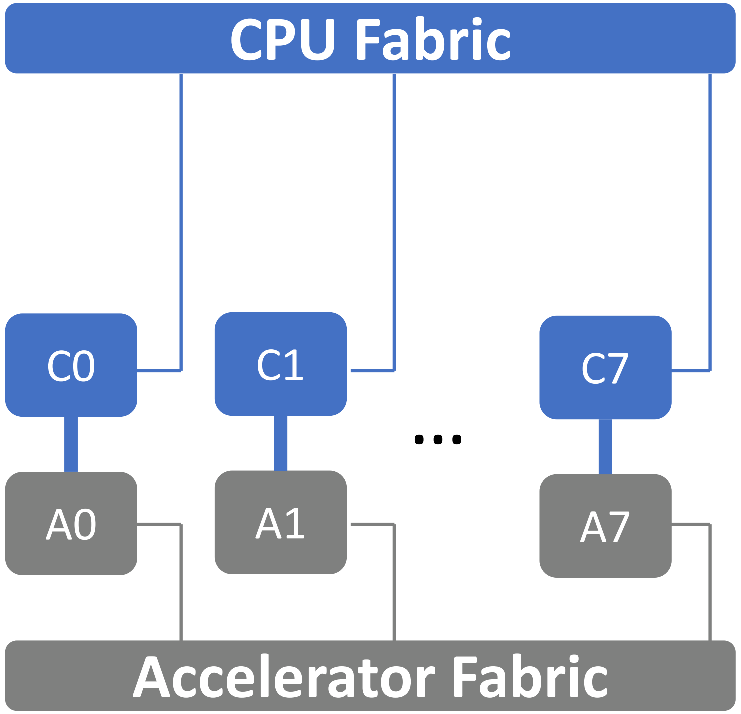
Zion: MLP and Embedding Support



	CPU	**Accelerator**
# of devices	8	8
Total BF16 Compute (TFLOPS)	O(1)	O(10)
Power per device	~100w	~200w
Mem Type	DDR4	HBM2
Total Capacity (GBYTES)	O(1000)	O(100)
Total BW (TB/s)	O(1)	O(10)

- Unified BFLOAT16 format with CPU and accelerators
- High capacity, low BW DDR; low capacity, high BW HBM

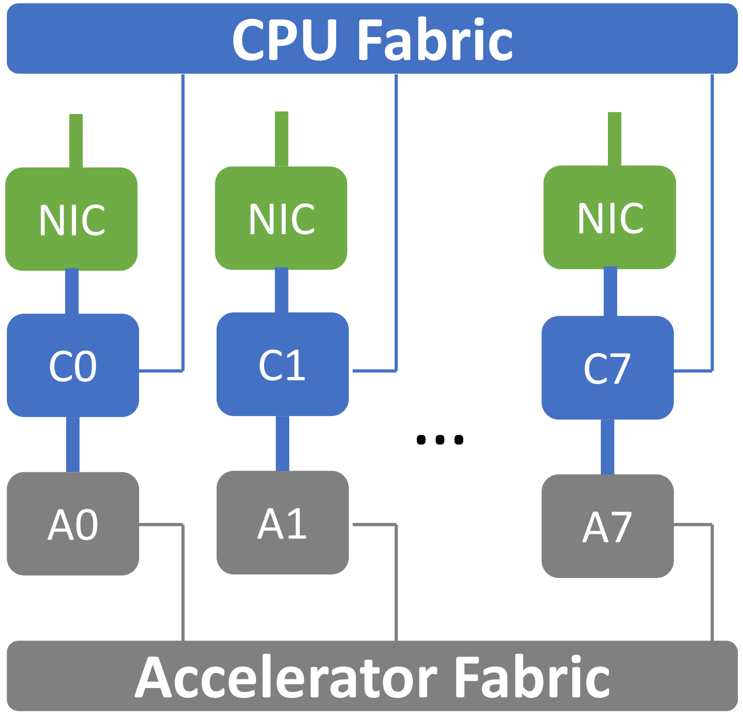
Zion: Communication Support



	CPU	**Accelerator**
Fabric Type	cache-coherent UPI	vendor
Fabric Topology	Twisted Hypercube	varies
Total BW (TB/s)	O(1)	O(1)

- Supports all-reduce and all-to-all
- Twisted hypercube has lower diameter than hypercube
- Use non-temporal stores on CPU to reduce coherent traffic

Zion: Scaling Out



	CPU	Accelerator
NIC (Gbps)	8 x 100	n/a
PCIe (Gen3 or 4)	X16	n/a

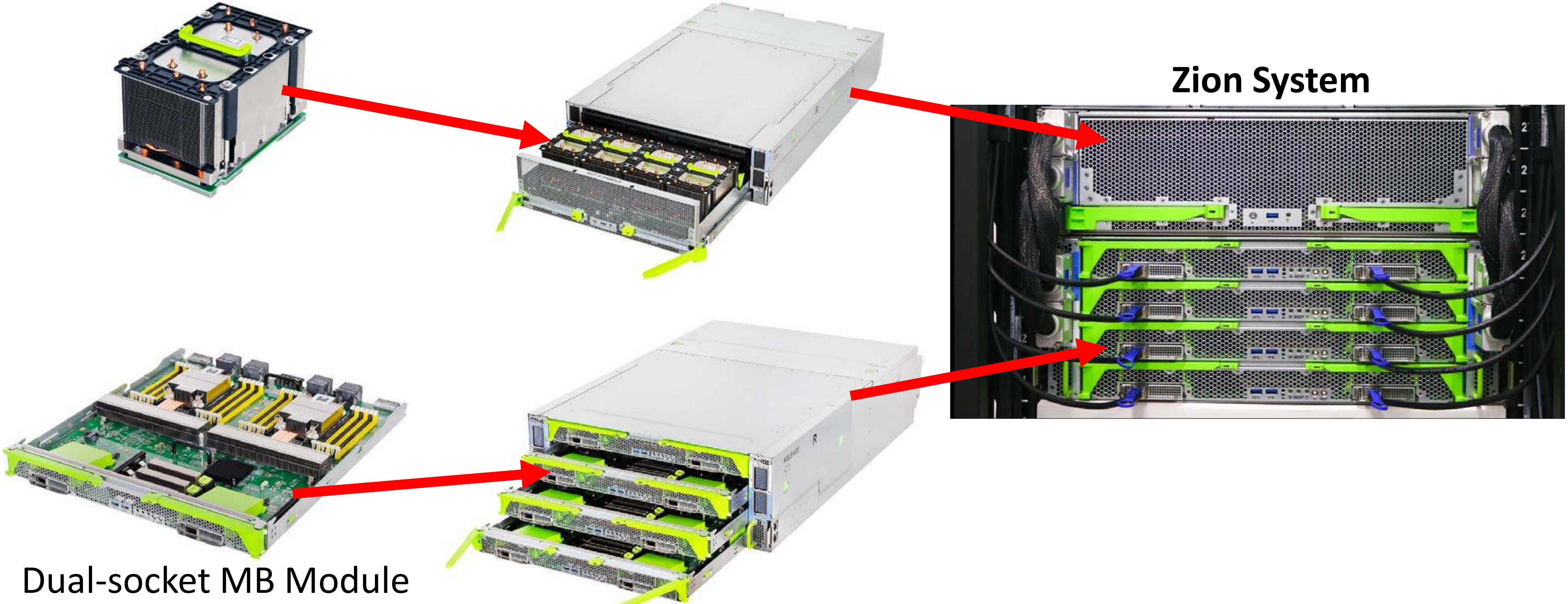
- Via host NIC, P2P, RDMA, PCI-SWITCH

Modular Physical Design

OCP Accelerator Module

8-Accelerator System

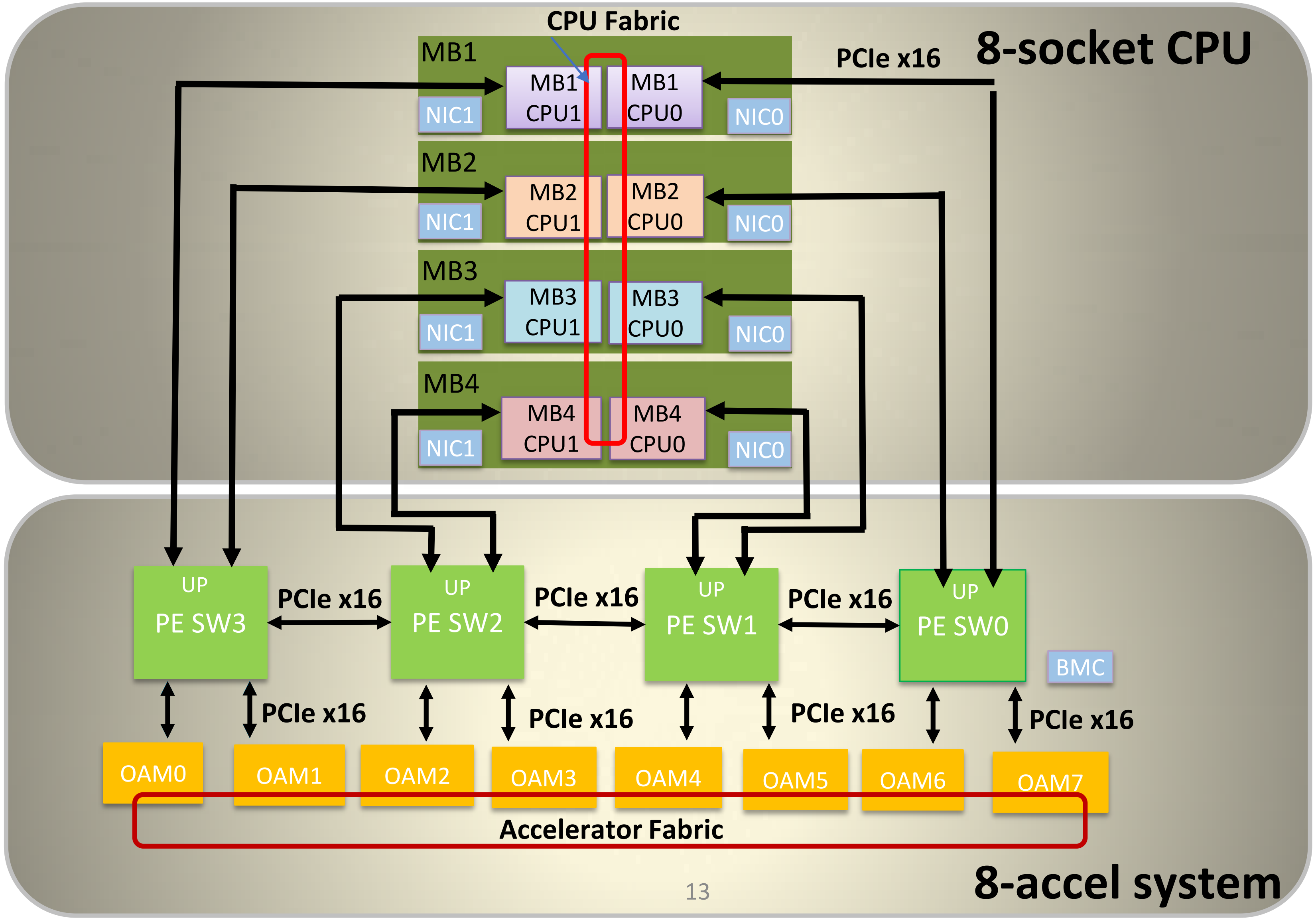
Zion System



Dual-socket MB Module

8-Socket System

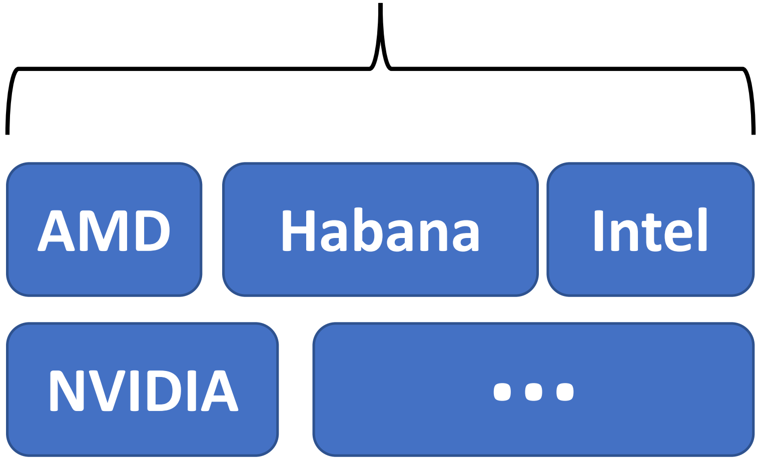
Platform Architecture



OCP Accelerator Module (OAM)

- Challenge: which accelerator do we use?
  - Very large number of accelerators
  - Limited resource to enable multiple systems

- Solution: OCP Accelerator Module(OAM)
  - Facebook led efforts
  - Define vendor-agnostic common form factor



AMD

Habana

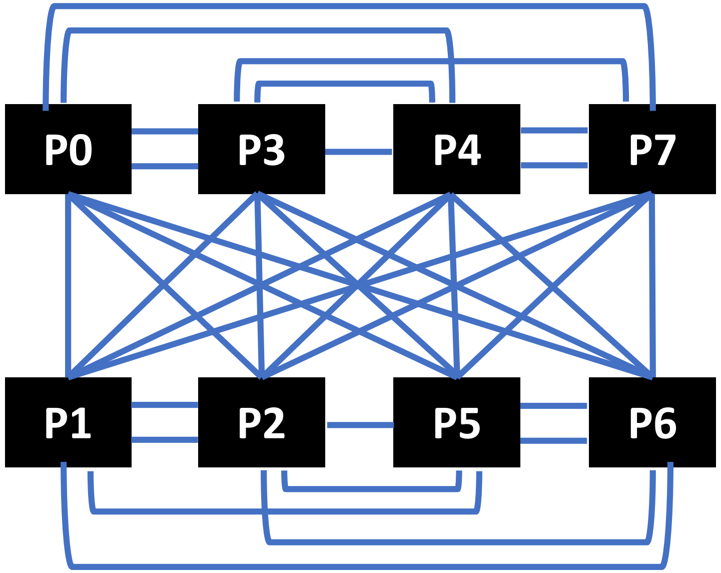
Intel

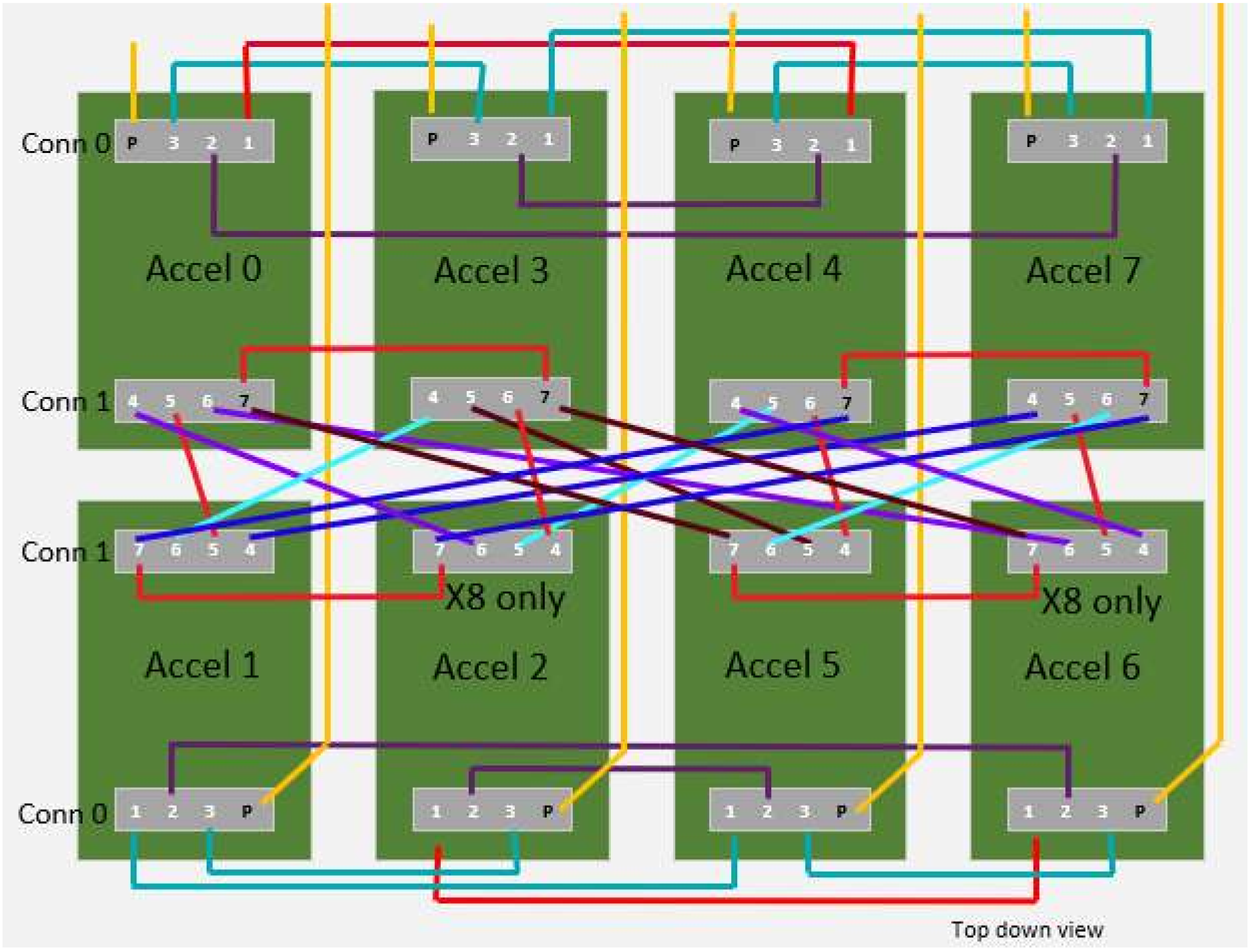
NVIDIA

…

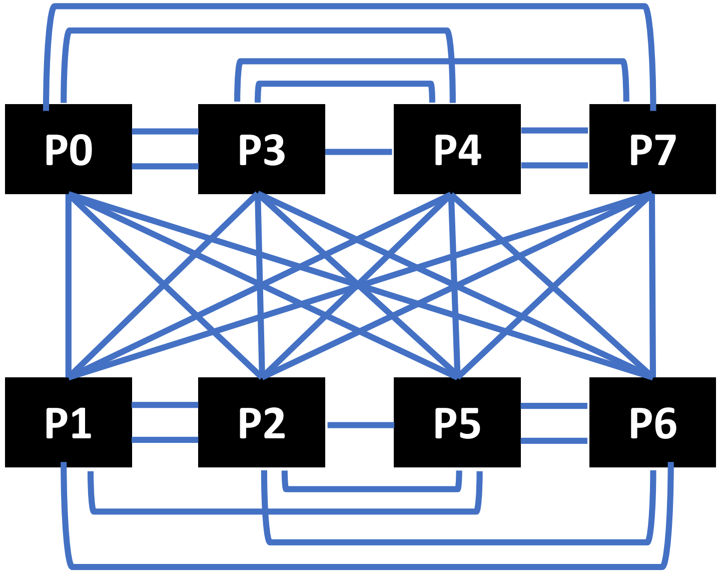
Accelerator Interconnect Topology

- Challenge: vendors support different topologies: FC, AFC, HCM, …
- Solution: superset physical topology

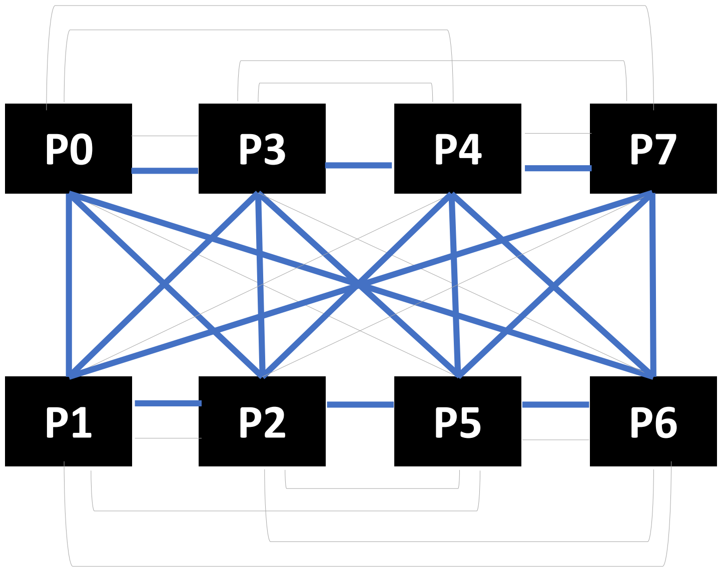




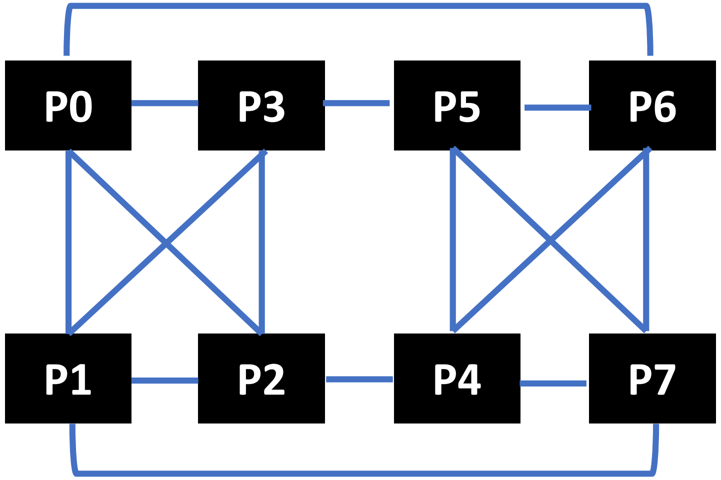
Example: Embedding Hypercube Mesh



Superset topology



Remove unused links



Rotate 4,7,5,6 by 180° → HCM

Software Flexibility

- User can gradually increase SW complexity (and performance)

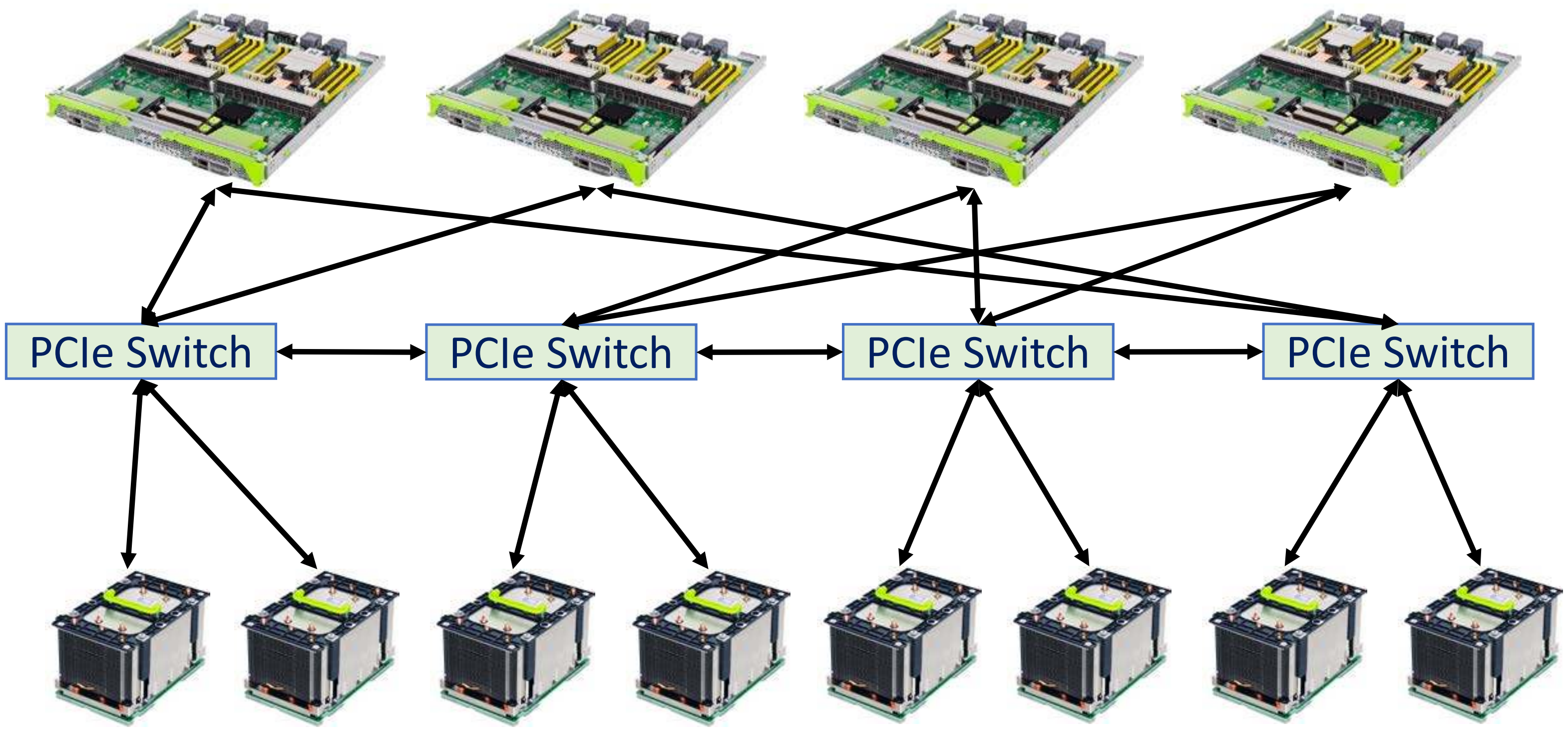
1. CPU-only
2. CPU for embeddings + Accelerators for MLP
3. Use Accelerator HBM for embeddings as well
   - Challenge: table accesses have different frequencies
   - Benefits from run-time profile driven table partitioning
4. Distributed training

- Creates continuum of dev efficiency vs performance tradeoffs

Hardware Flexibility

- Four 2S modules are identical
- Configured based on workloads needs as
  - Up to four 2S systems
  - One or two 4S systems
  - One 8S
- SW -> BMC -> CM -> Configure board IDs to be 2S, 4S or 8S to power on

8 CPU Sockets



PCIe Switch

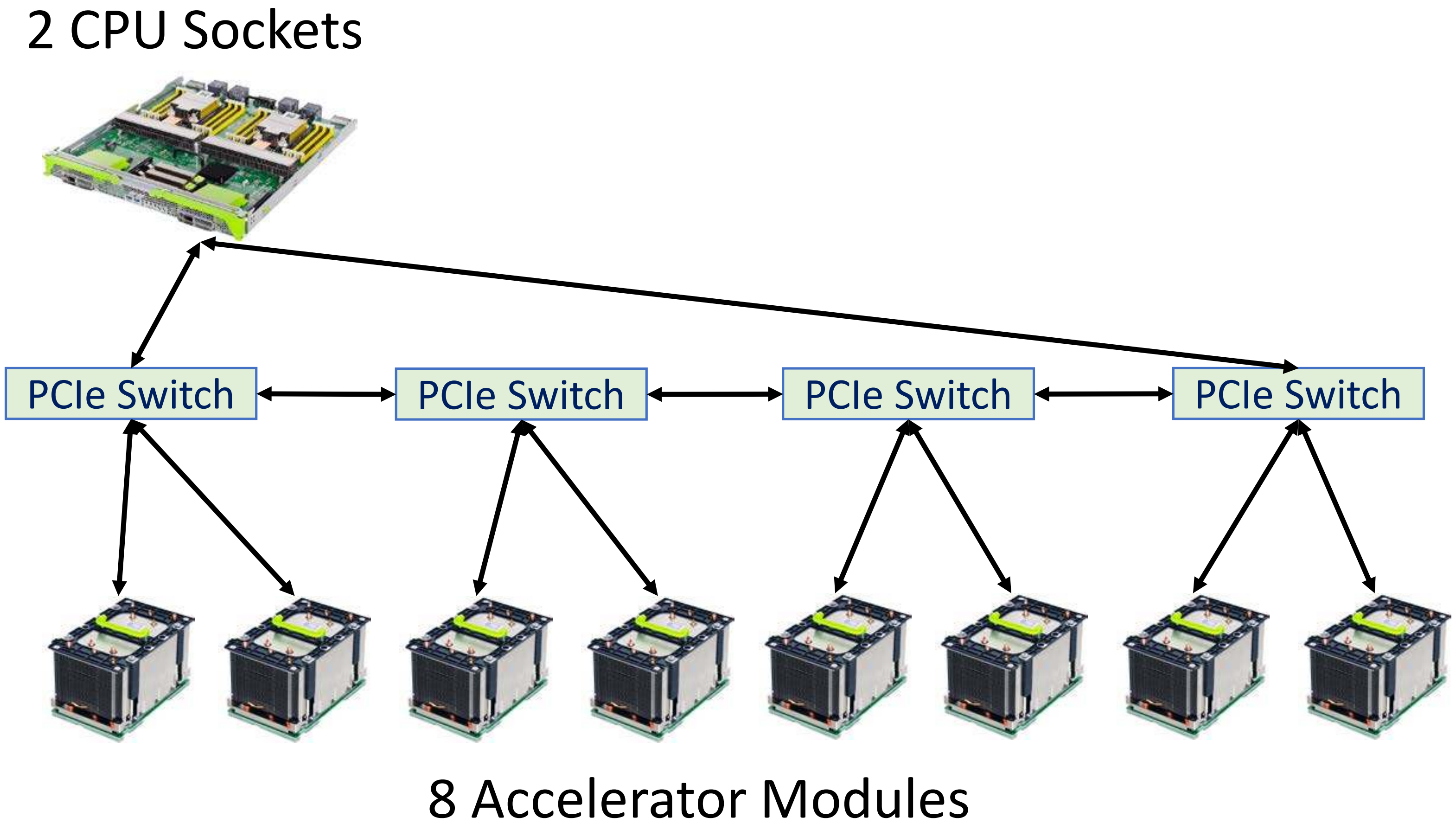
PCIe Switch

PCIe Switch

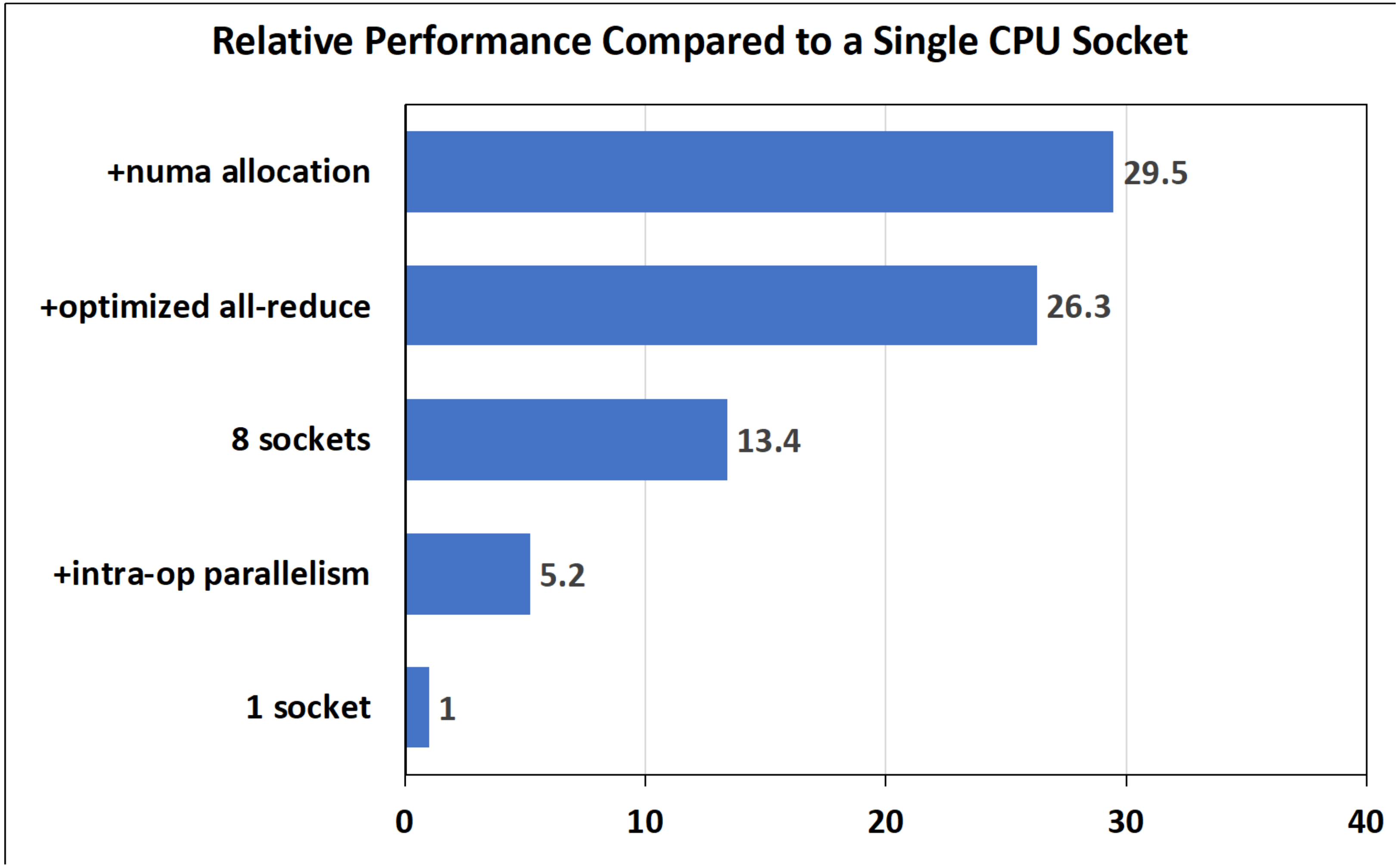
PCIe Switch

8 Accelerator Modules

One 2S Re-configuration Example



Production Performance Results (CPU Only)



Comparison with GPU-based Platform

	Big Basin	Zion
Accelerator	NVIDIA GPU Only	Different accelerators
Interconnect	Hypercube mesh via NVLINK	Richer set of topologies
Memory Capacity	O(100) GB	O(1000) GB
Number of CPUs	Single headnode	Reconfigurable

Conclusions

- Zion is FB next generation flexible training platform
- Co-designed to target demanding recommendation models
- Adopts new vendor-agnostic OCP accelerator module
- Building block that can scale out to a bigger system