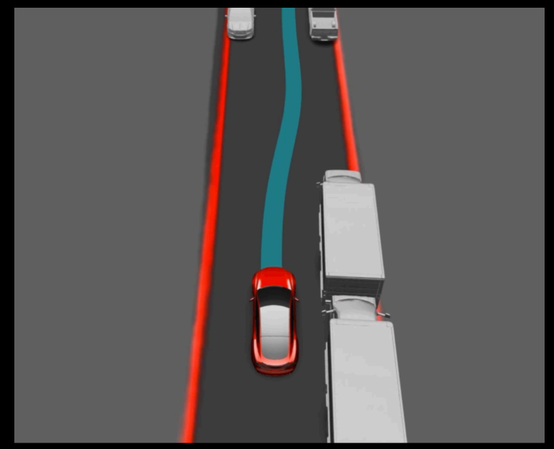
Super-Compute System Scaling for ML Training

Bill Chang, Rajiv Kurian, Doug Williams, Eric Quinnell

Path to General Autonomy

Model Architecture

Vision, Path Planning, Auto-Labeling
New Models Architectures
Parameter Sizes Increasing Exponentially



Training Data

Video Training Data With 4D Labels
Ground Truth Generation

Training Infrastructure

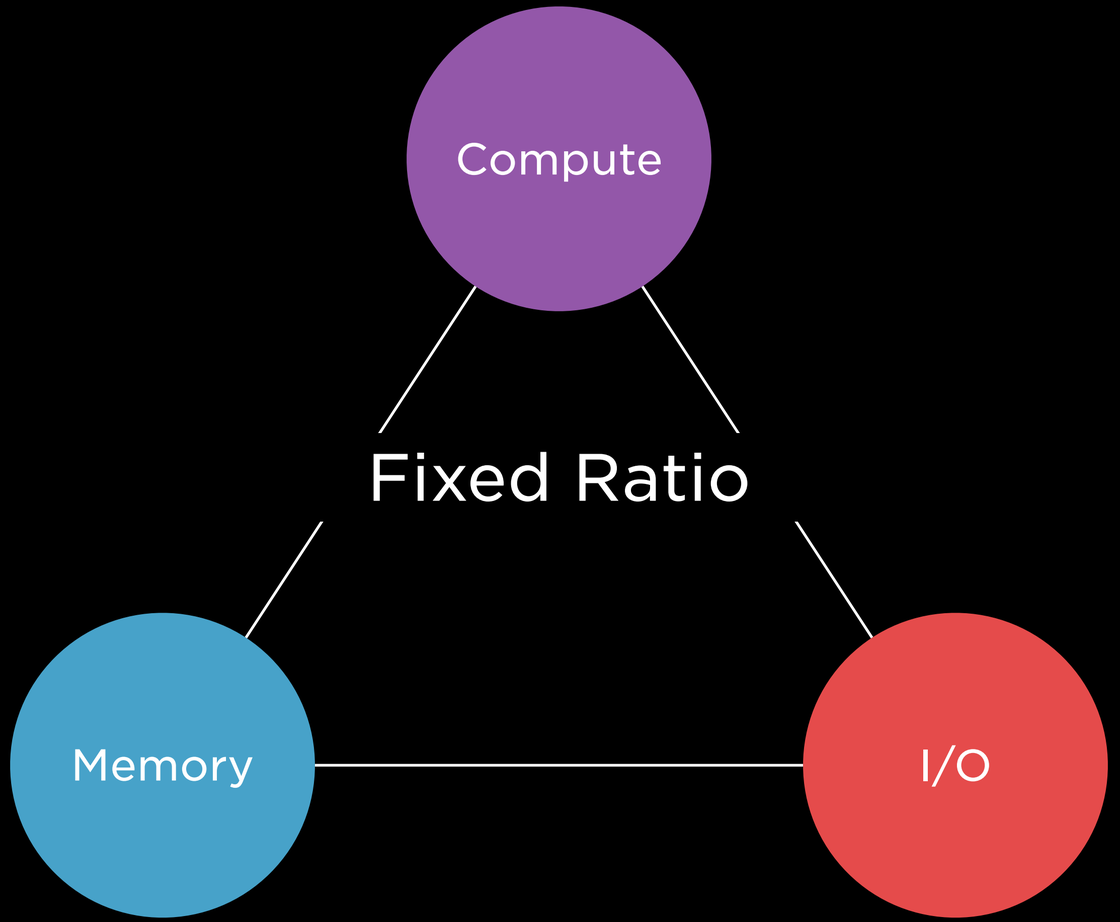
Training and Evaluation Pipeline

Accelerated ML Training System

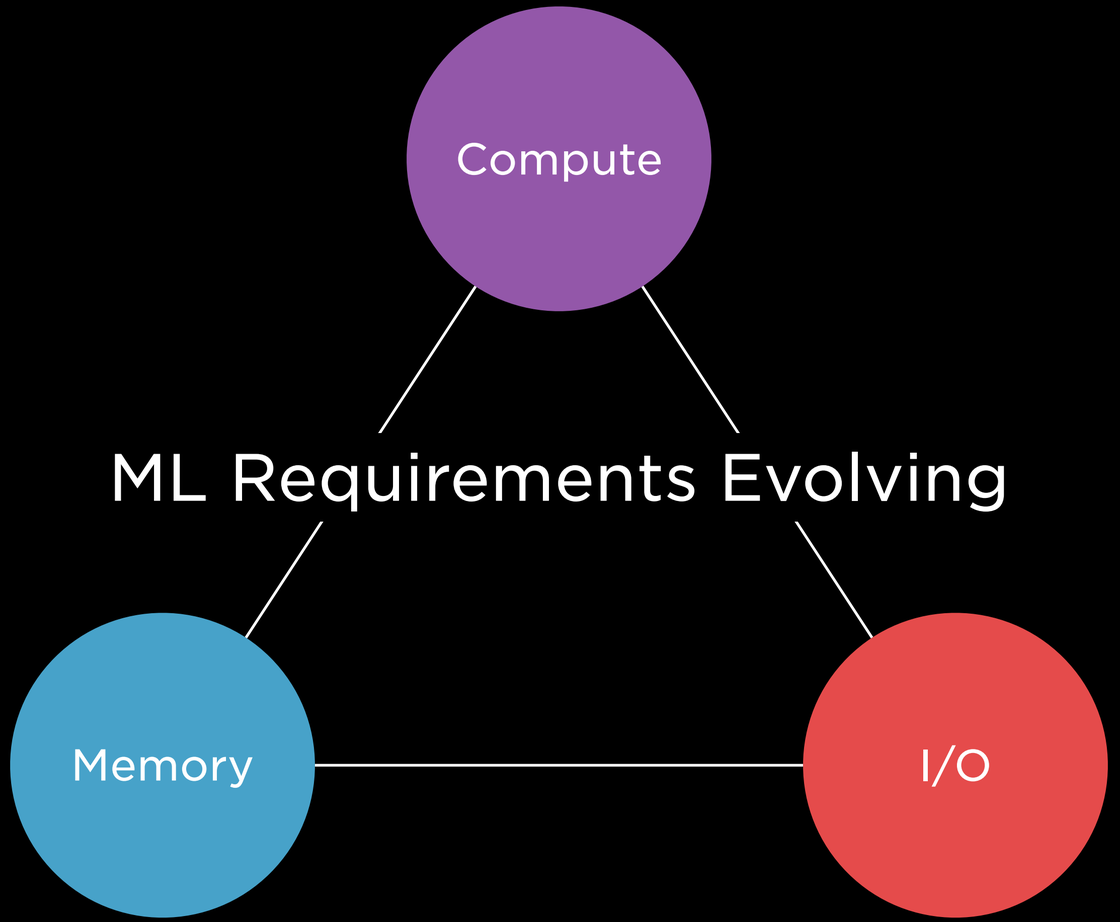
Flexible System Architecture

Software at Scale

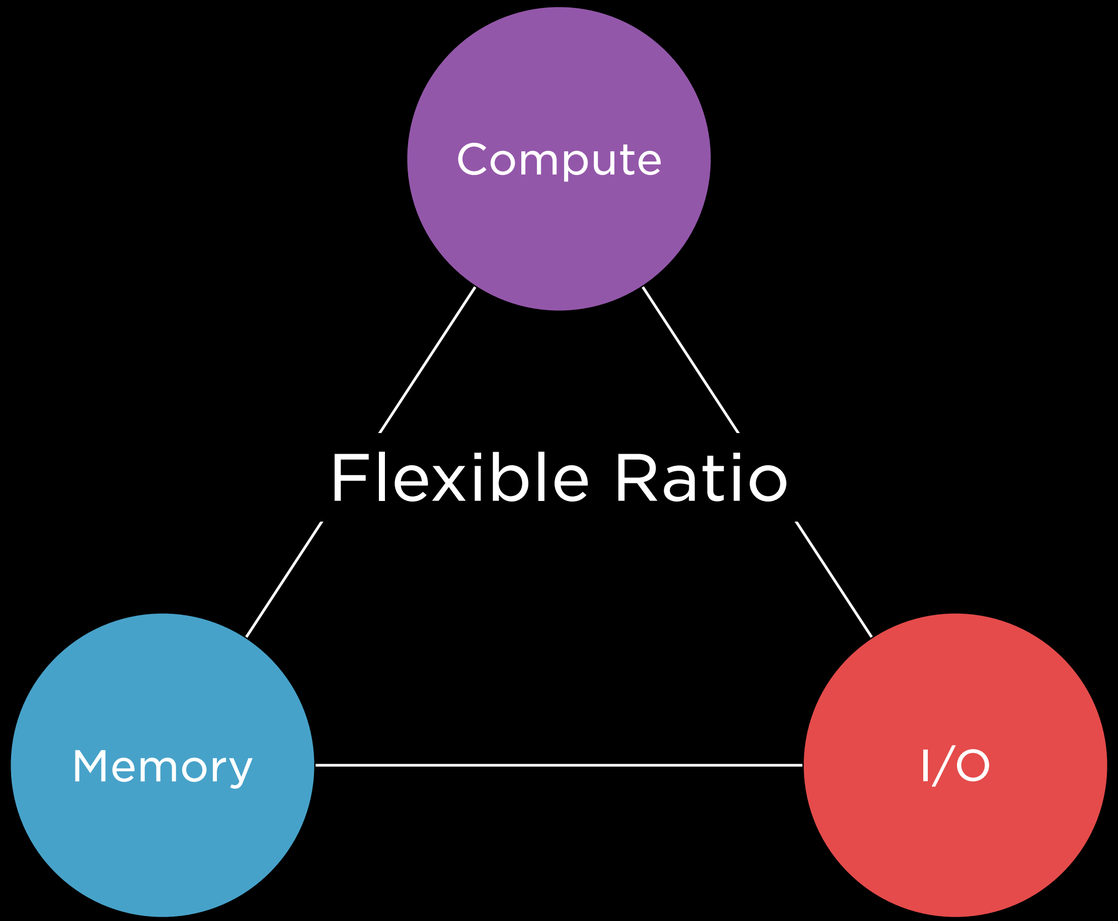
Typical System



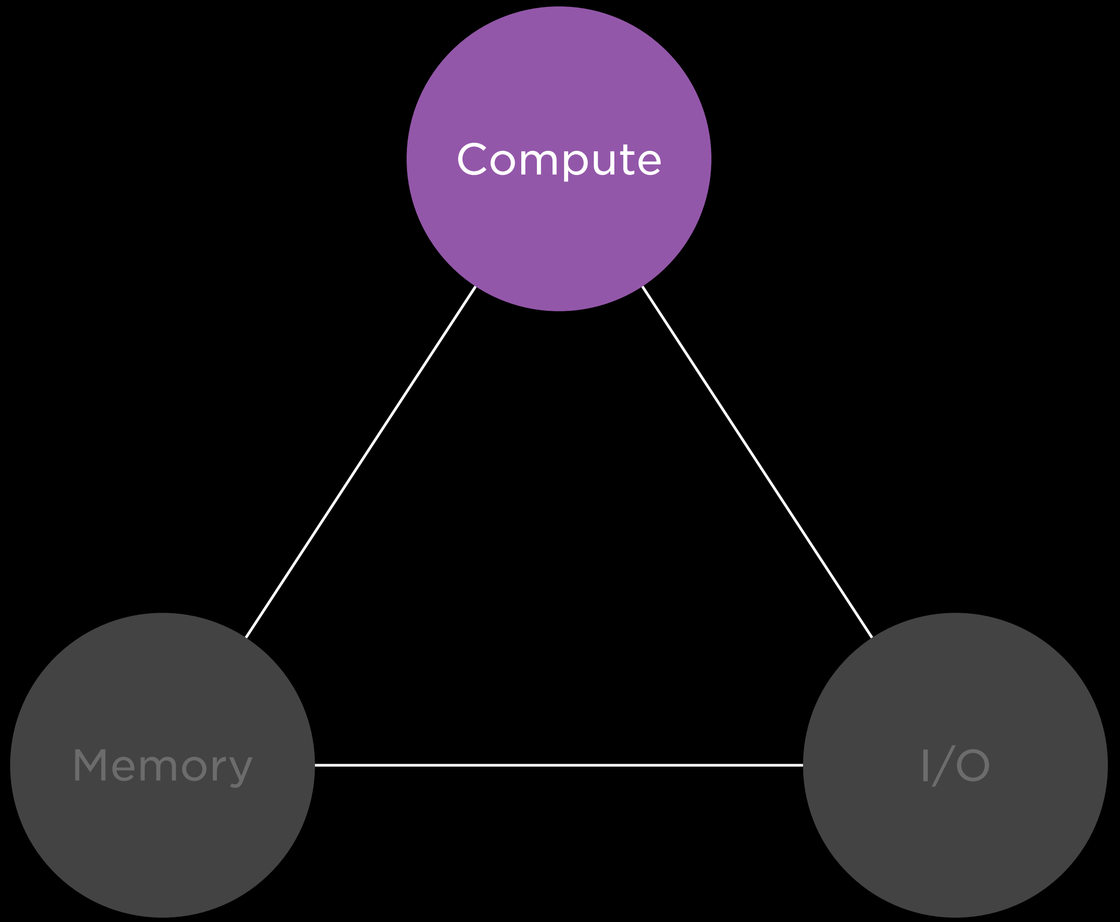
Optimized ML Training System



Disaggregated System Architecture



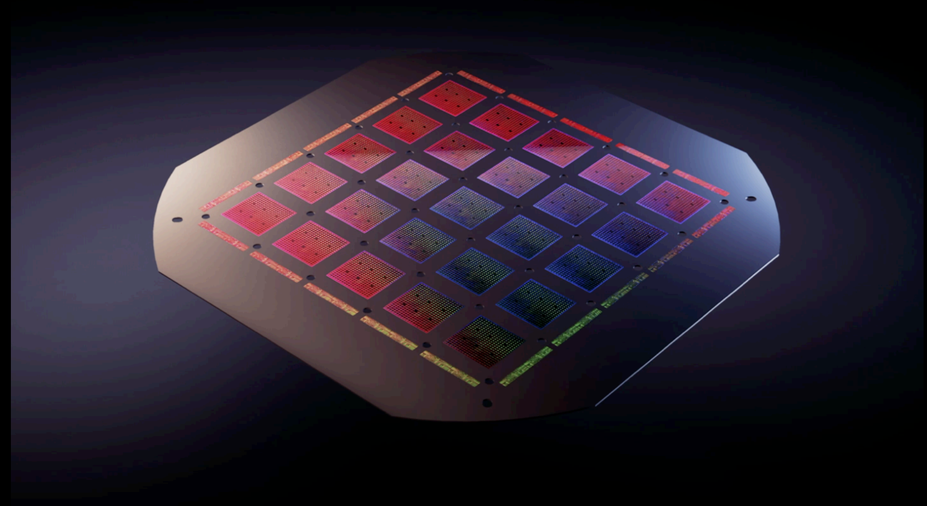
Optimized Compute



Technology-Enabled Scaling

System-On-Wafer Technology

- 25 D1 Compute Dies + 40 I/O Dies
- Compute and I/O Dies Optimize Efficiency and Reach
- Heterogenous RDL Optimized for High-Density and High-Power Layout



Maximize Performance and Yield

- Known Good Die and Fault Tolerant Designs
- Each Tile Assembled With Fully Functional Dies
- Harvesting and Fully Configurable Routing for Yield

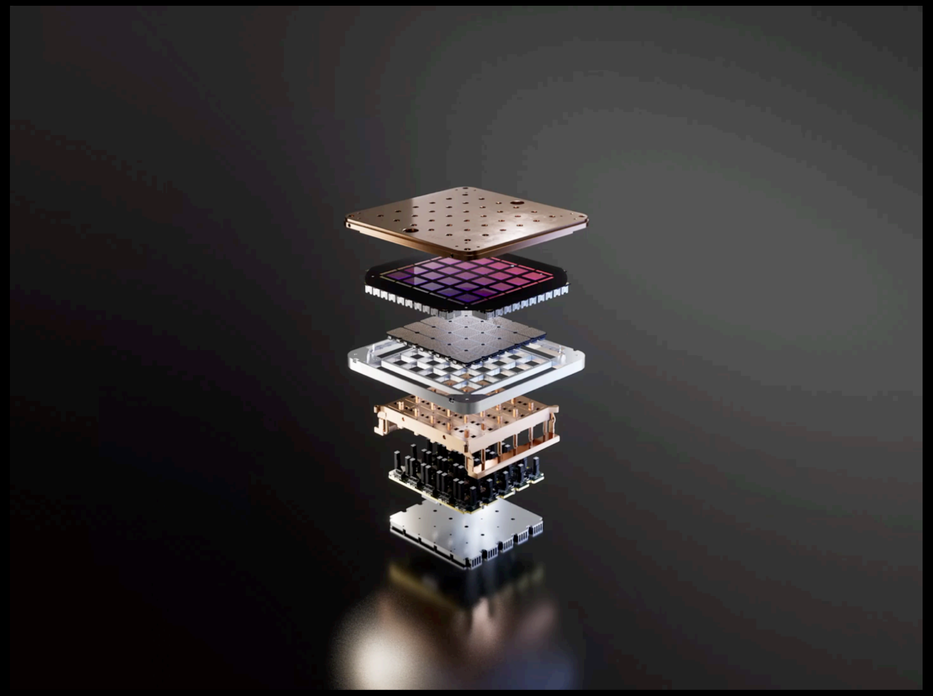
Training Tile

Unit of Scale

- Large Compute With Optimized I/O
- Fully Integrated System Module (Power/Cooling)

Uniform High-Bandwidth

- 10 TB/s on-tile bisection bandwidth
- 36 TB/s off-tile aggregate bandwidth

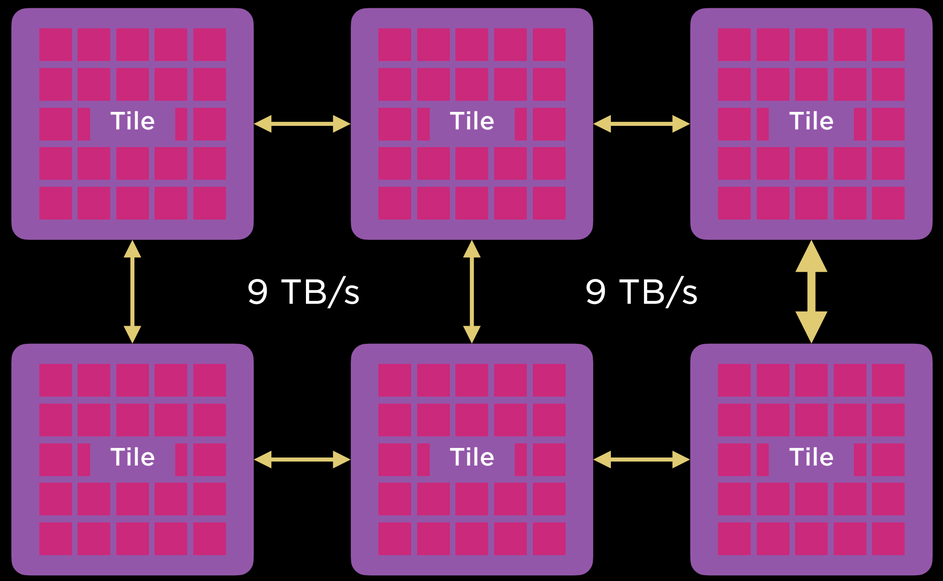


9 PFLOPS BF16/CFP8

11 GB High-Speed ECC SRAM

36 TB/s Aggregate I/O BW

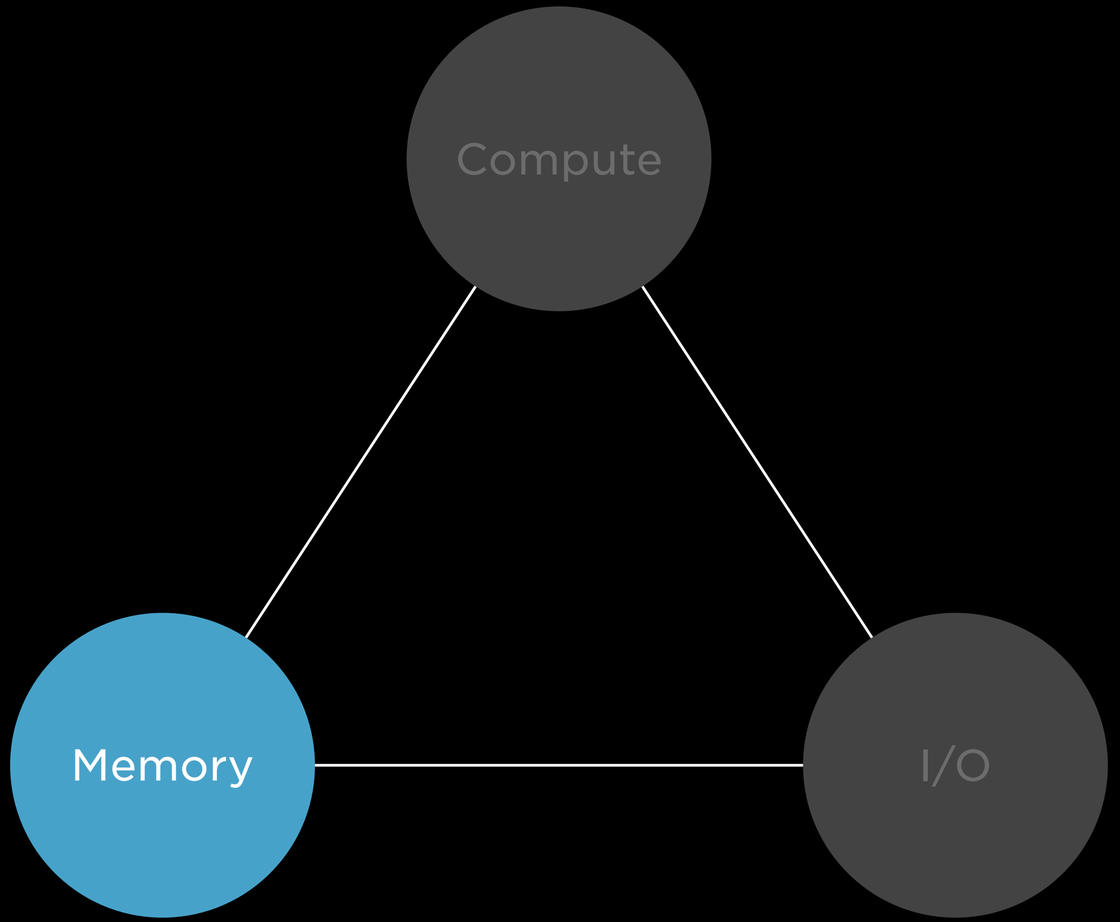
Flexible Building Block



Scale With Multiple Tiles

No Additional Power/Cooling Design Needed

Disaggregated Memory



V1 Dojo Interface Processor

32GB High-Bandwidth Memory

- 800 GB/s Total Memory Bandwidth

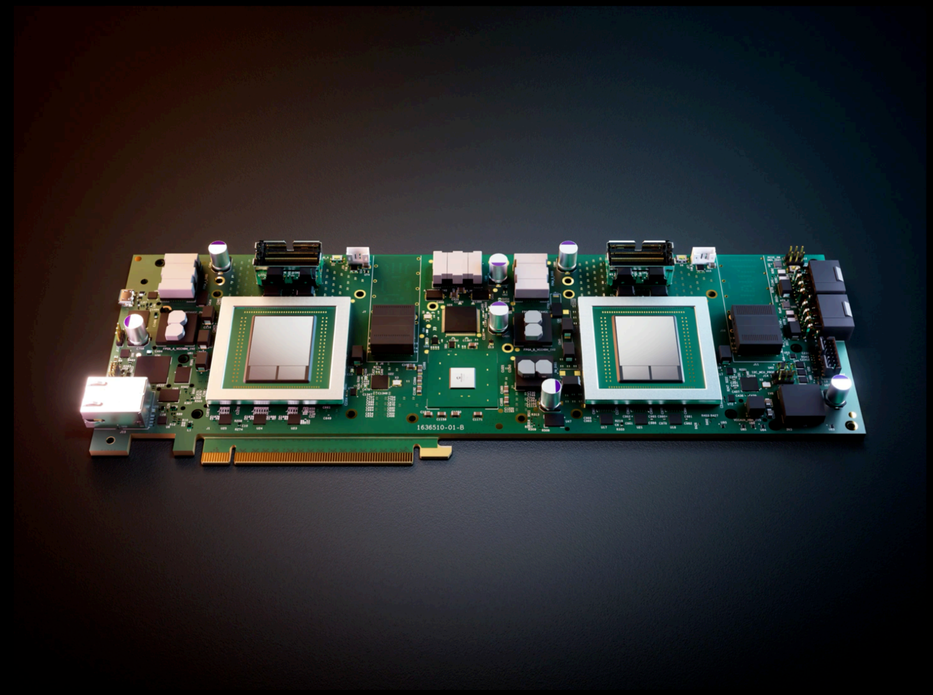
900 GB/s TTP Interface

- Tesla Transport Protocol (TTP) - Full custom protocol
- Provides full DRAM bandwidth to Training Tile

50 GB/s TTP over Ethernet (TTPoE)

- Enables extending communication over standard Ethernet
- Native hardware support

32 GB/s Gen4 PCIe Interface



Dojo Interface Processor - PCIe Topology

160GB Total DRAM per Tile edge

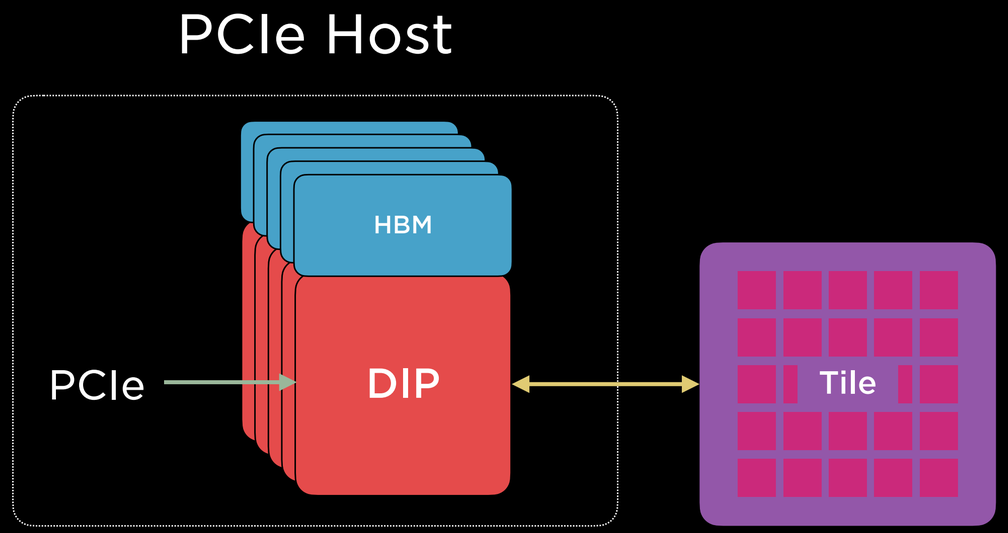
- Shared memory for training tiles

5 DIP Cards Provide Max Bandwidth

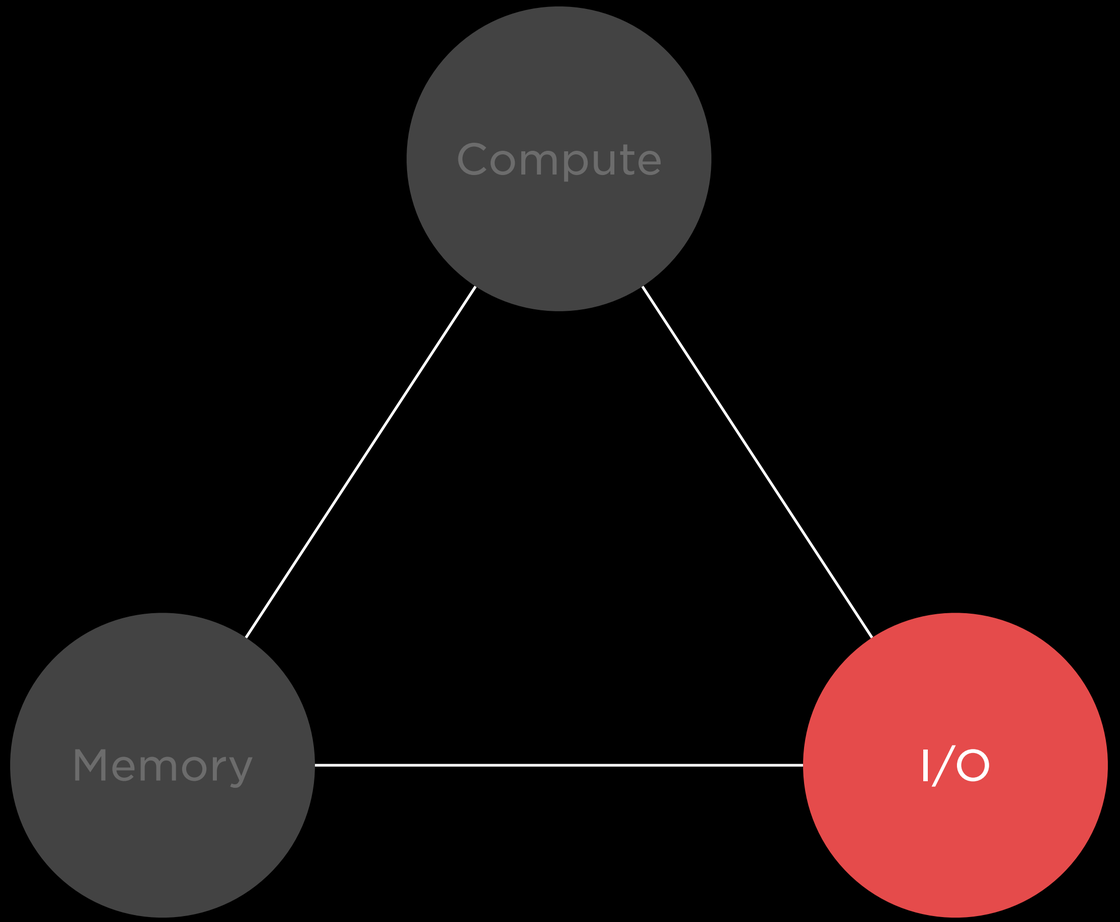
- 4.5 TB/s aggregate bandwidth to DRAM over TTP

80 Lanes PCIe Gen4 Interface

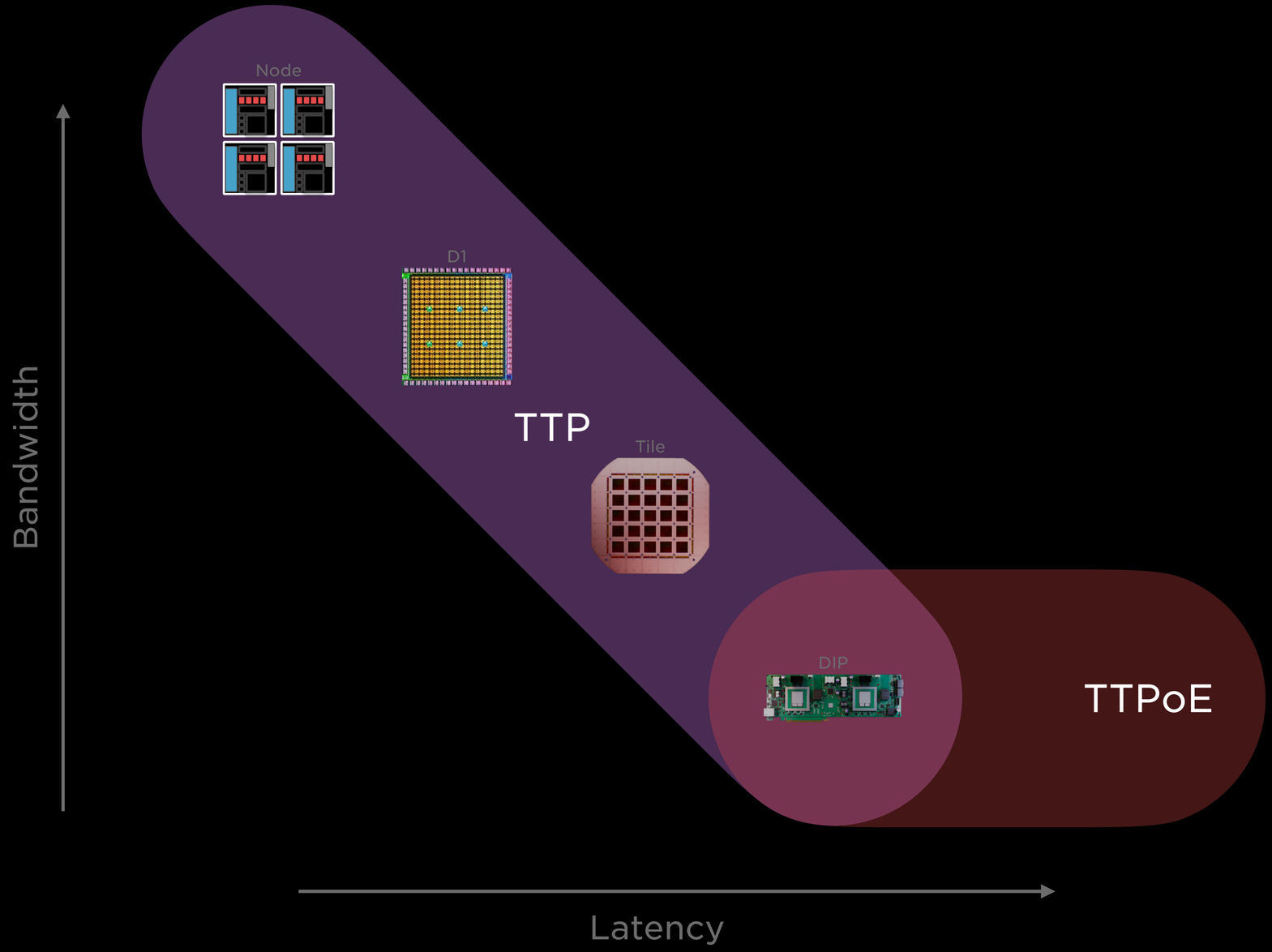
- Provide standard connectivity to hosts



Scalable Communication



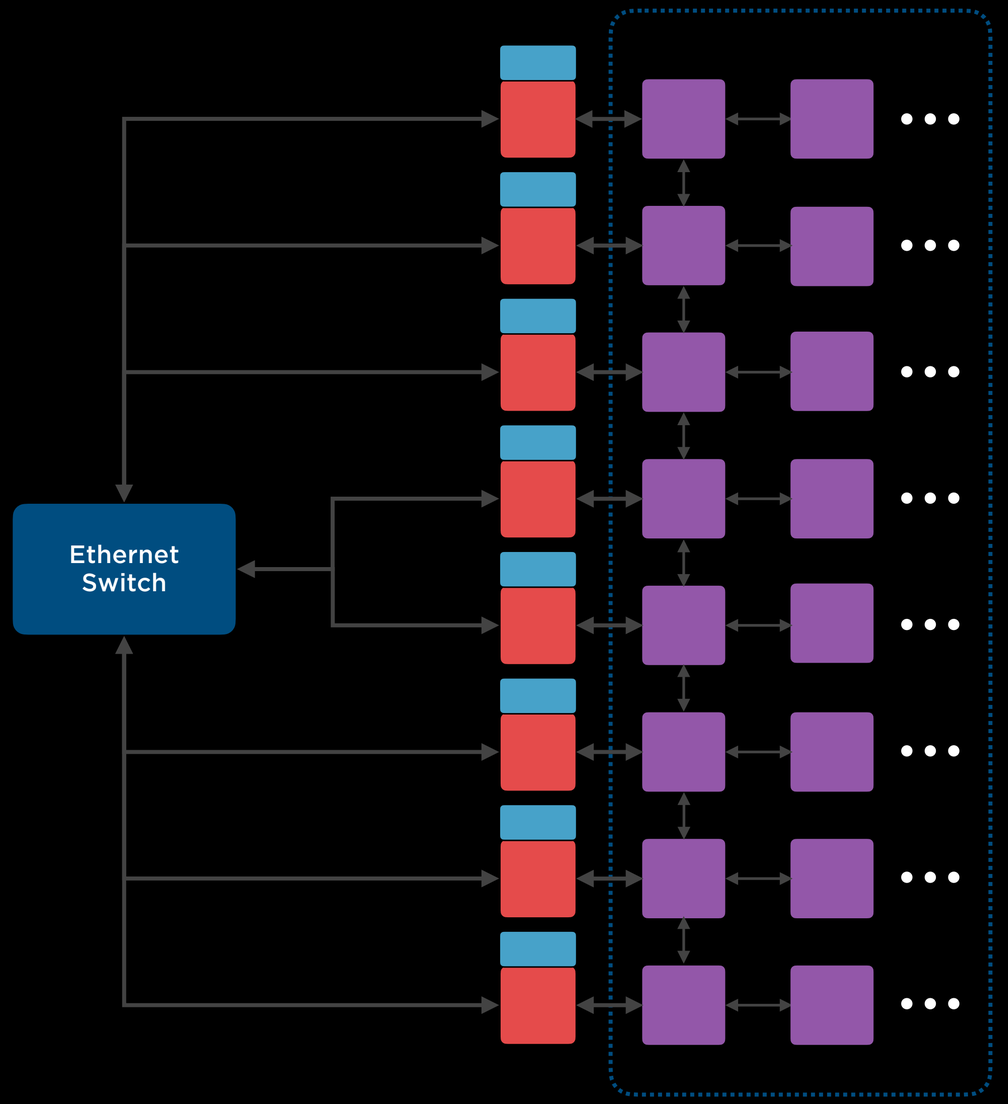
Tesla Transport Protocol



Dojo Interface Processor - Z-Plane Topology

TTPoE - Point-to-Point over Ethernet

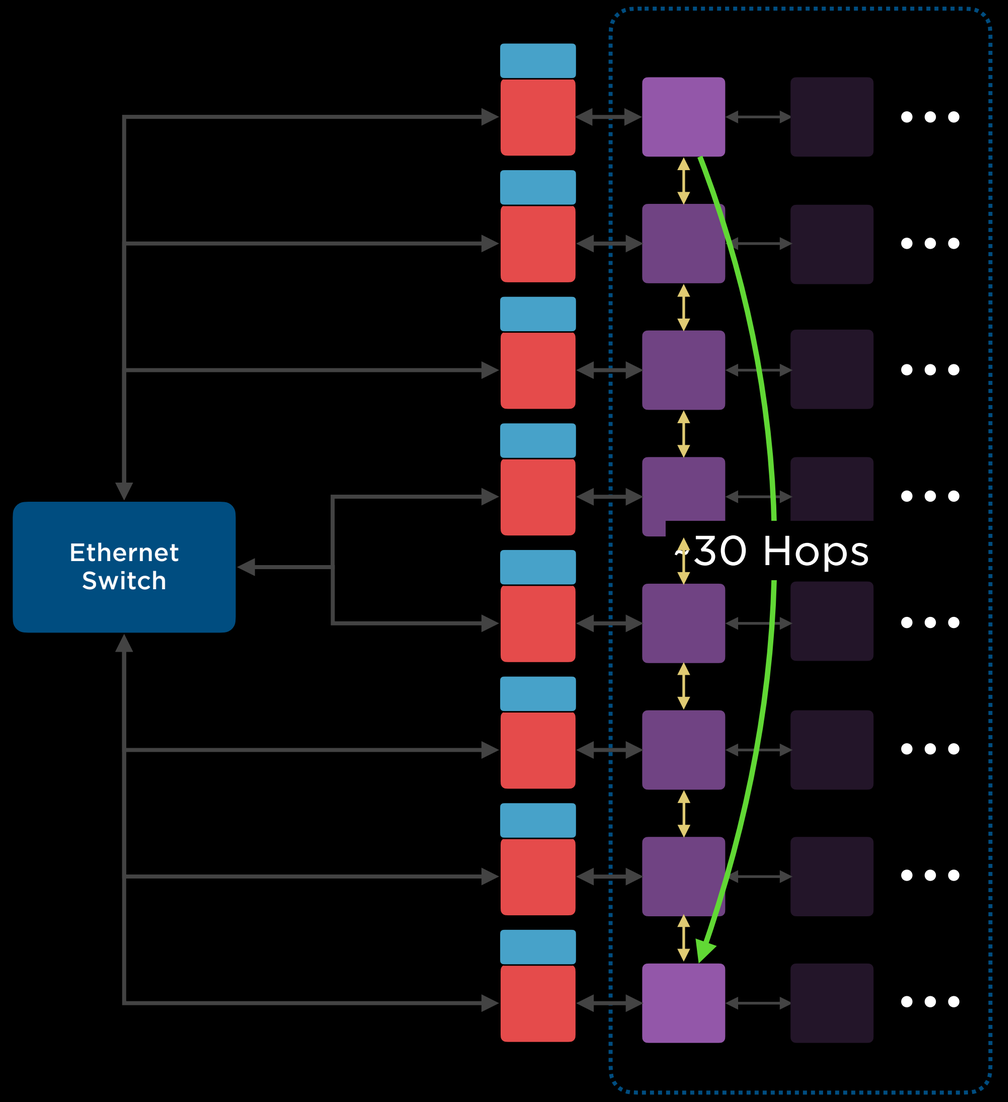
- Provides high-radix connectivity in Z-plane TTP network
- Enables “shortcuts” across the network
- Manage latency for sync and control across compute plane



Dojo Interface Processor - Z-Plane Topology

TTPoE - Point-to-Point over Ethernet

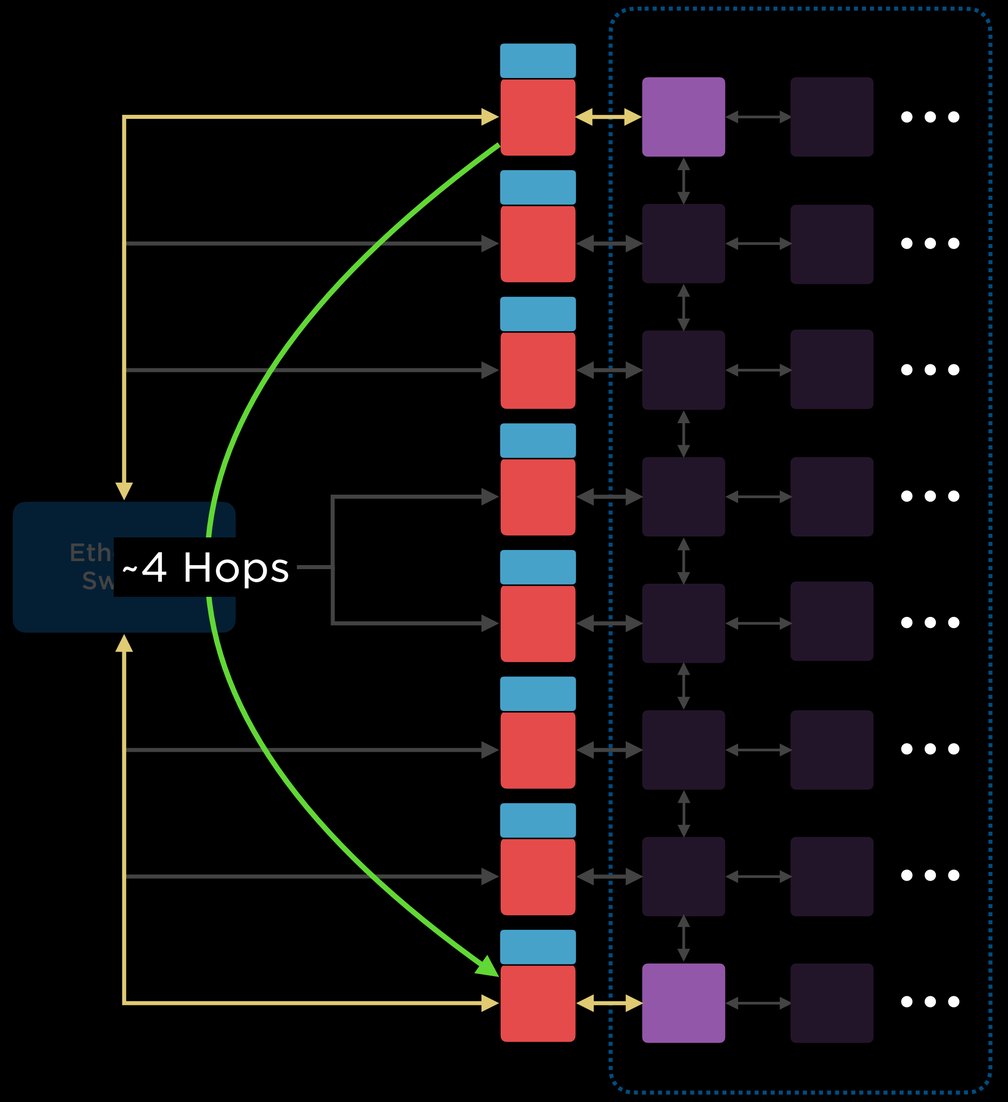
- Provides high-radix connectivity in Z-plane TTP network
- Enables “shortcuts” across the network
- Manage latency for sync and control across compute plane



Dojo Interface Processor - Z-Plane Topology

TTPoE - Point-to-Point over Ethernet

- Provides high-radix connectivity in Z-plane TTP network
- Enables “shortcuts” across the network
- Manage latency for sync and control across compute plane

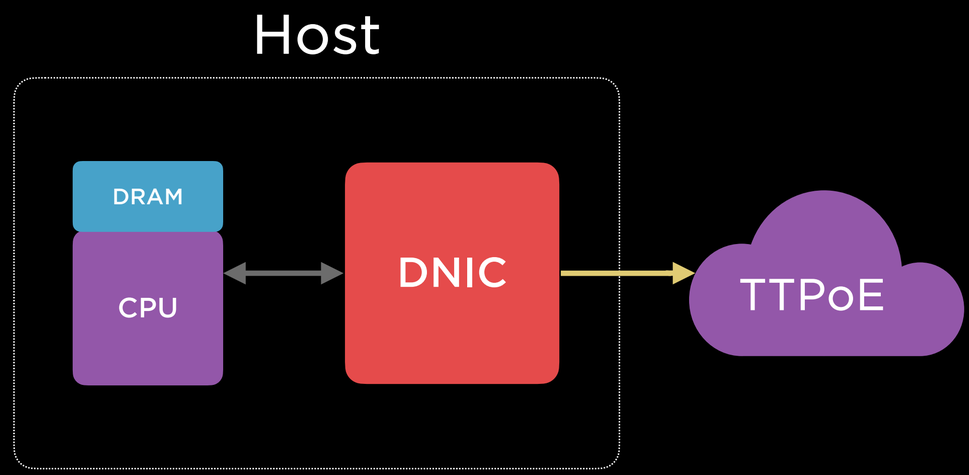


Dojo Network Interface Card

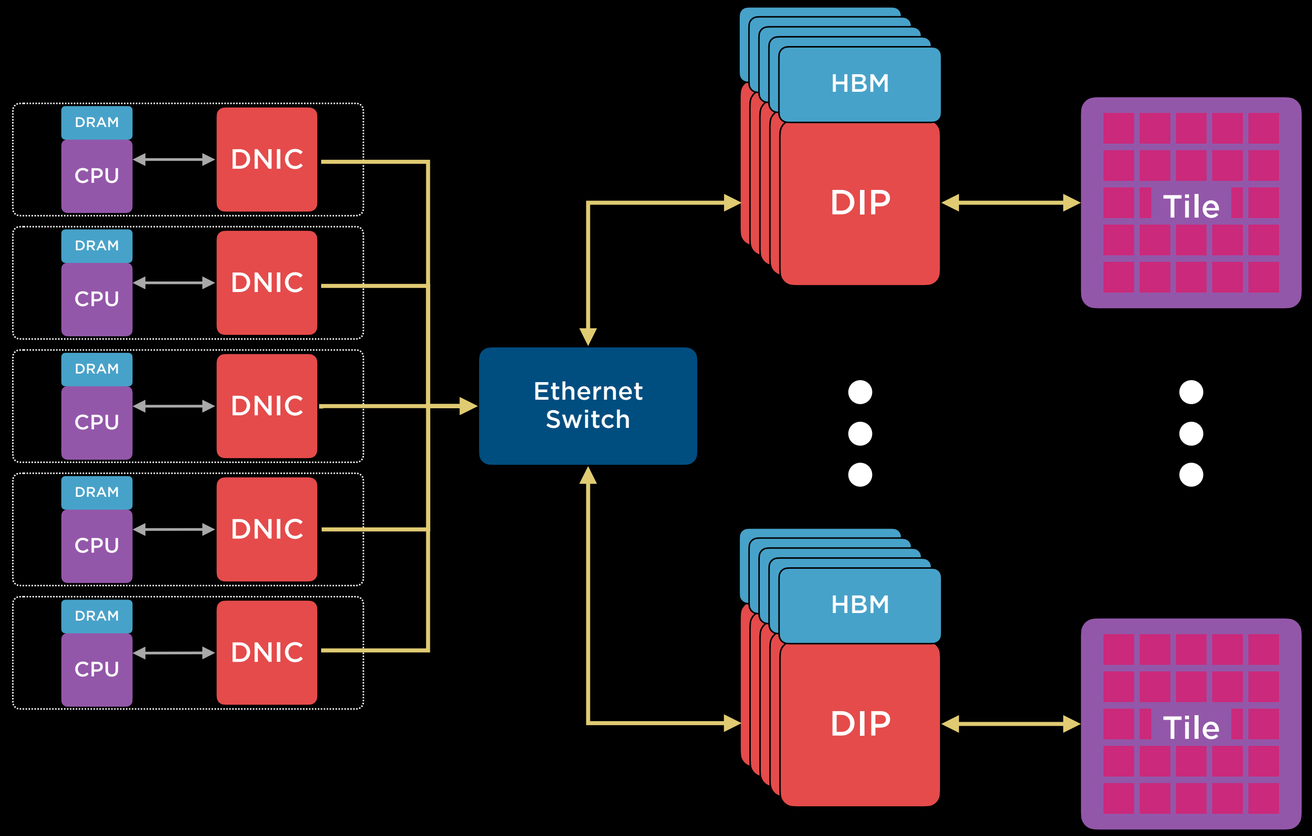
Remote DMA over TTPoE

- DMA to/from any TTP endpoint (compute SRAM, DRAM)
- Leverage switched Ethernet networks

Enables Remote Compute for Pre/post-processing

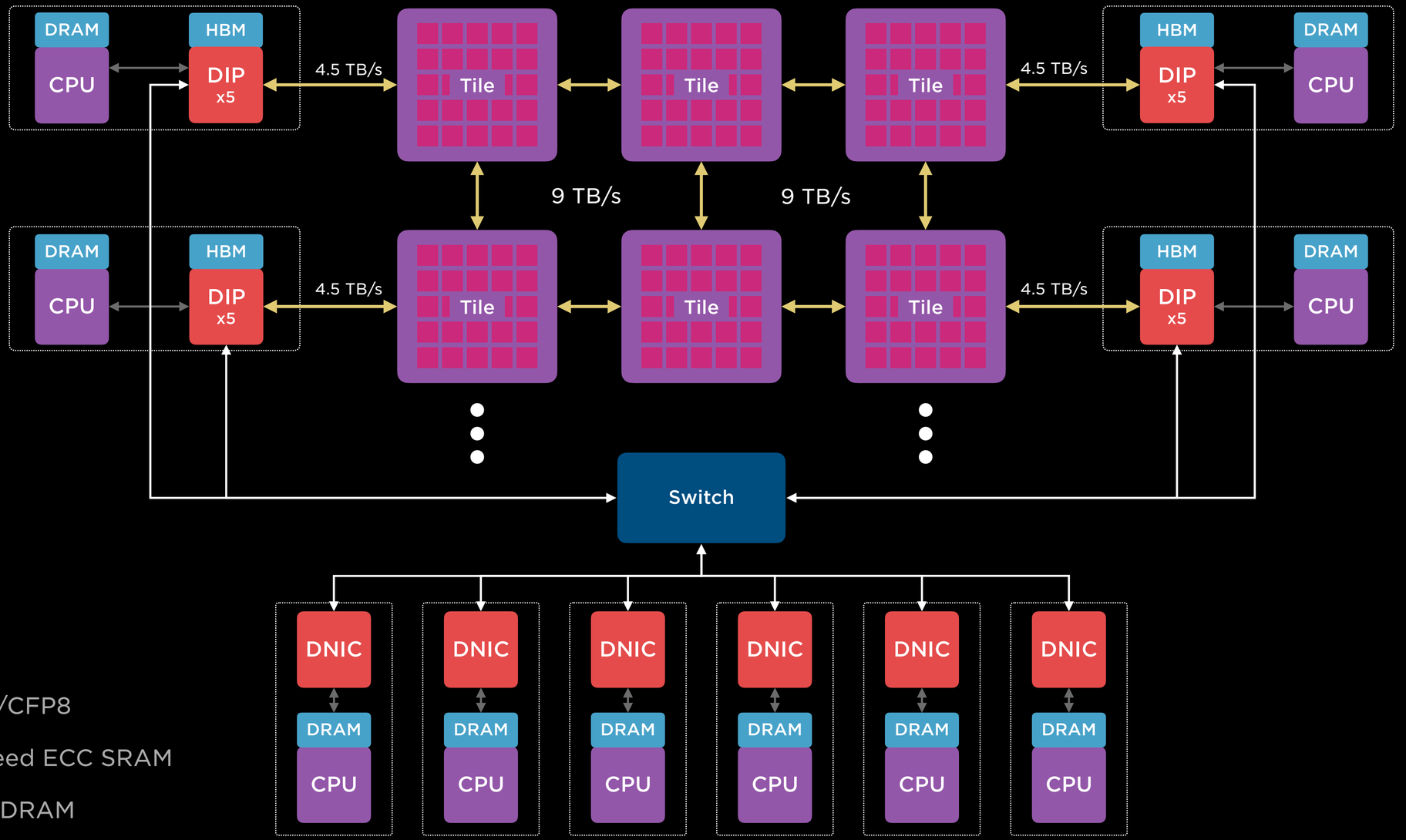


Remote DMA Topology



Scale-Out for CPU/Memory Bound
Pre-Processing Workloads

V1 Dojo Training Matrix



1 EFLOP BF16/CFP8

1.3 TB High-Speed ECC SRAM

13 TB High-BW DRAM

Disaggregated Scalable System

Tile

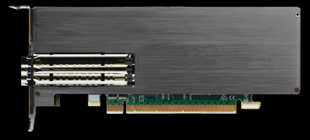
Compute

Memory

I/O

Interface Processor

Network Interface



Software at Scale

Model Execution

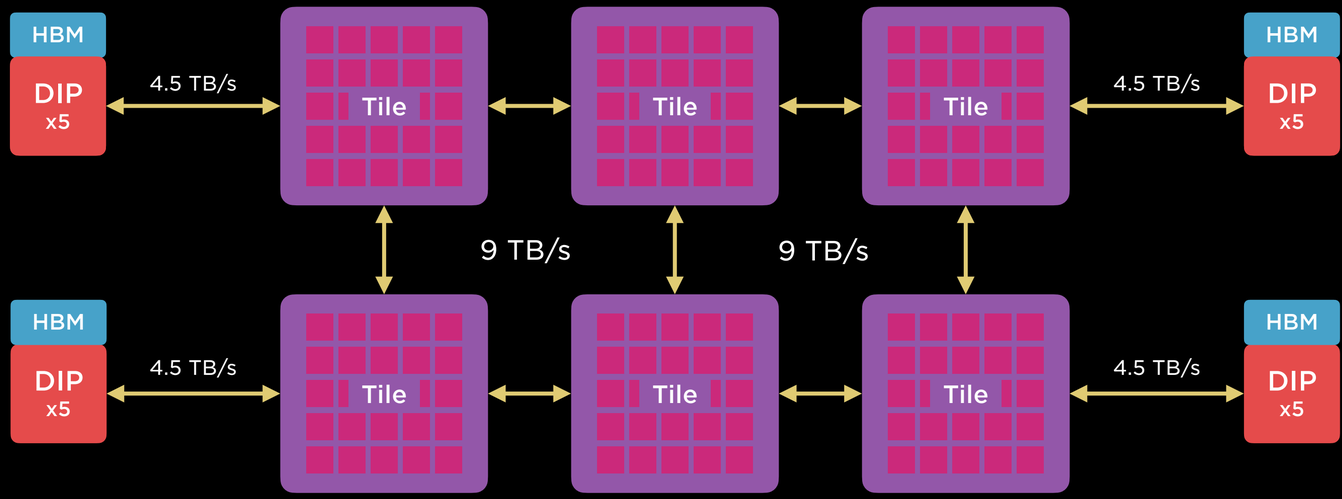
Workloads operate almost entirely out of SRAM

Single copy of parameters - replicated just in time

High utilization

Unlike typical accelerators, all forms of parallelism may cross die boundaries

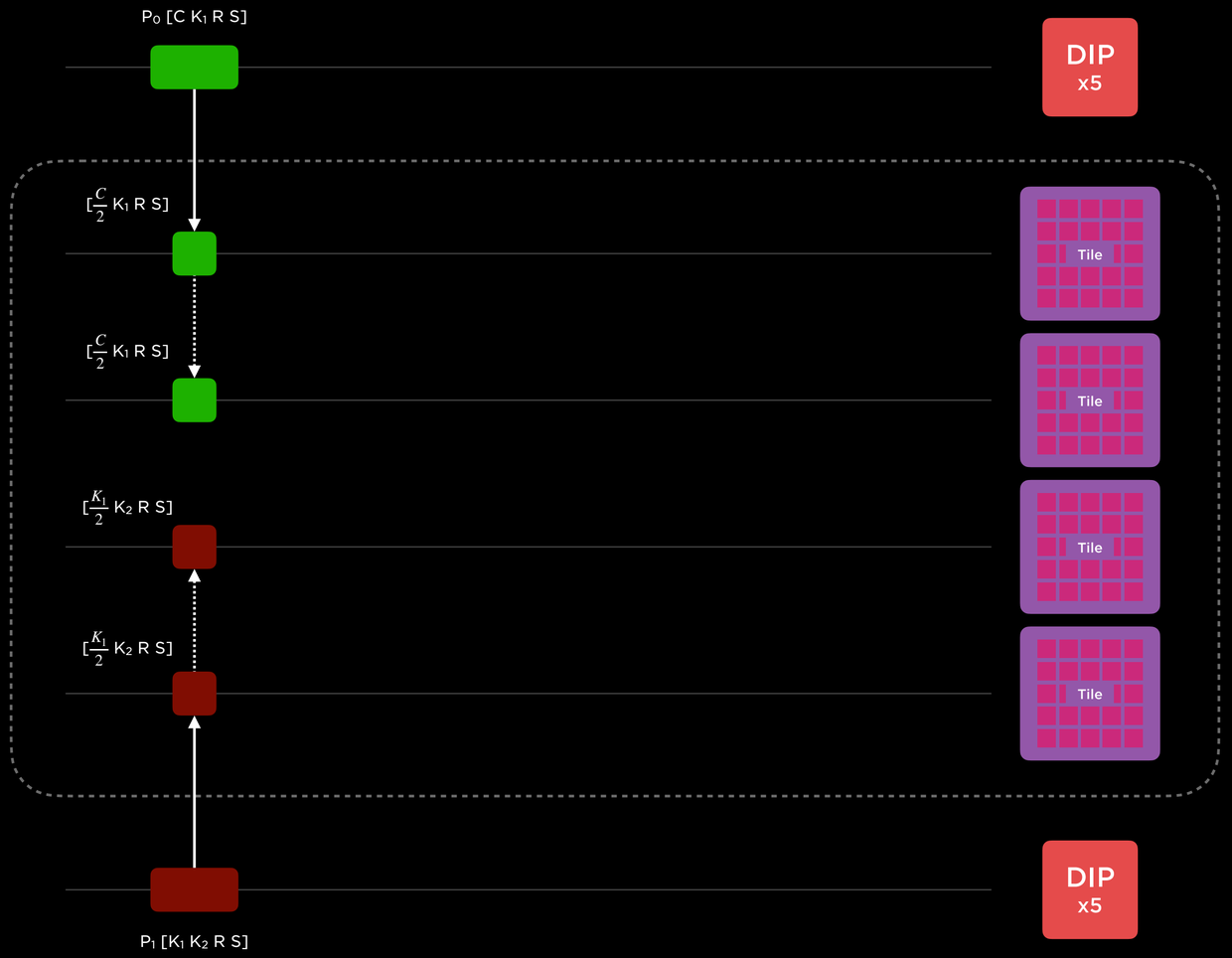
Thanks to High TTP Bandwidth



Model Execution

Parameters Are Distributed Across the DIPs

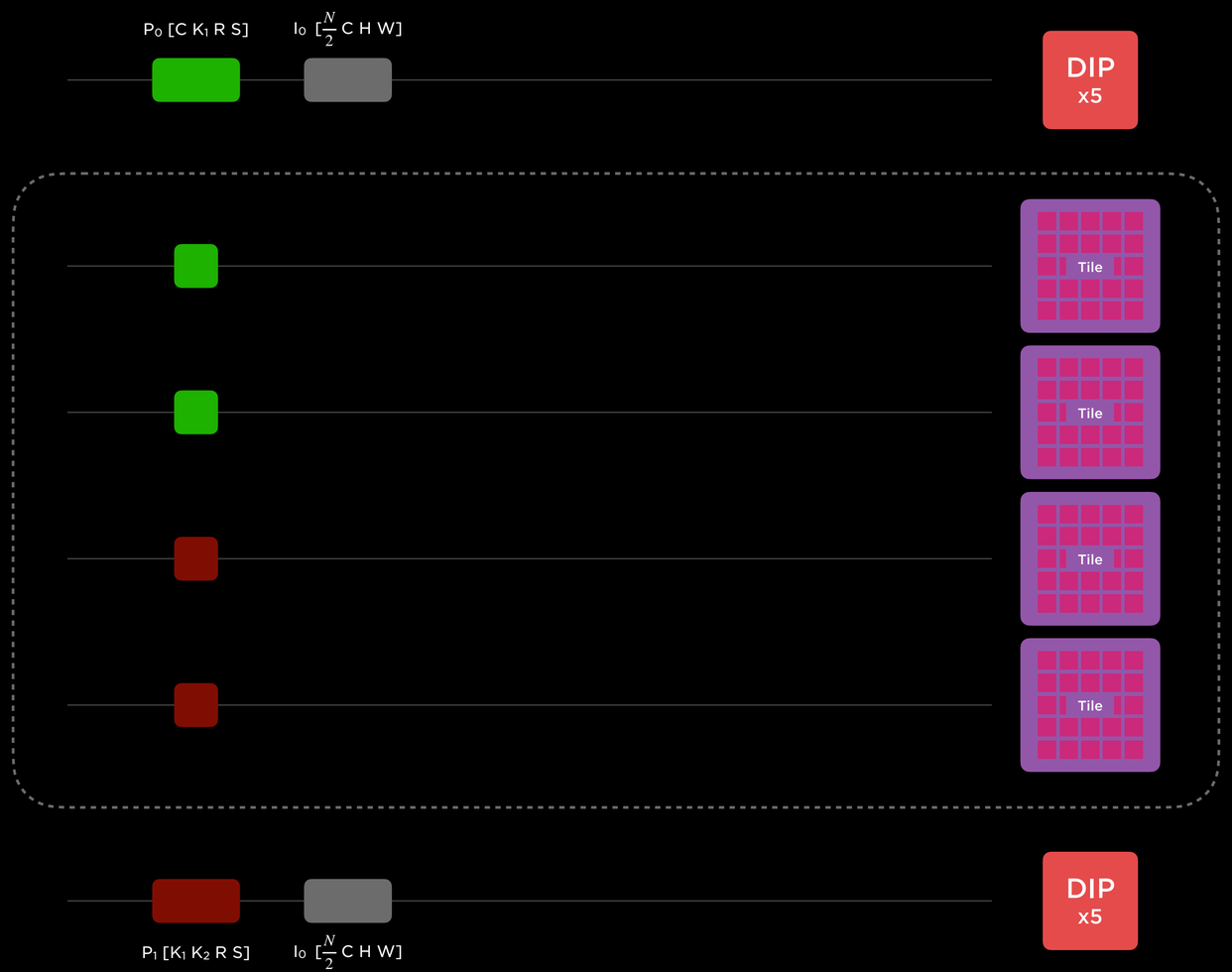
Model Execution



Parameters Are Sharded Across the Tiles at Load Time

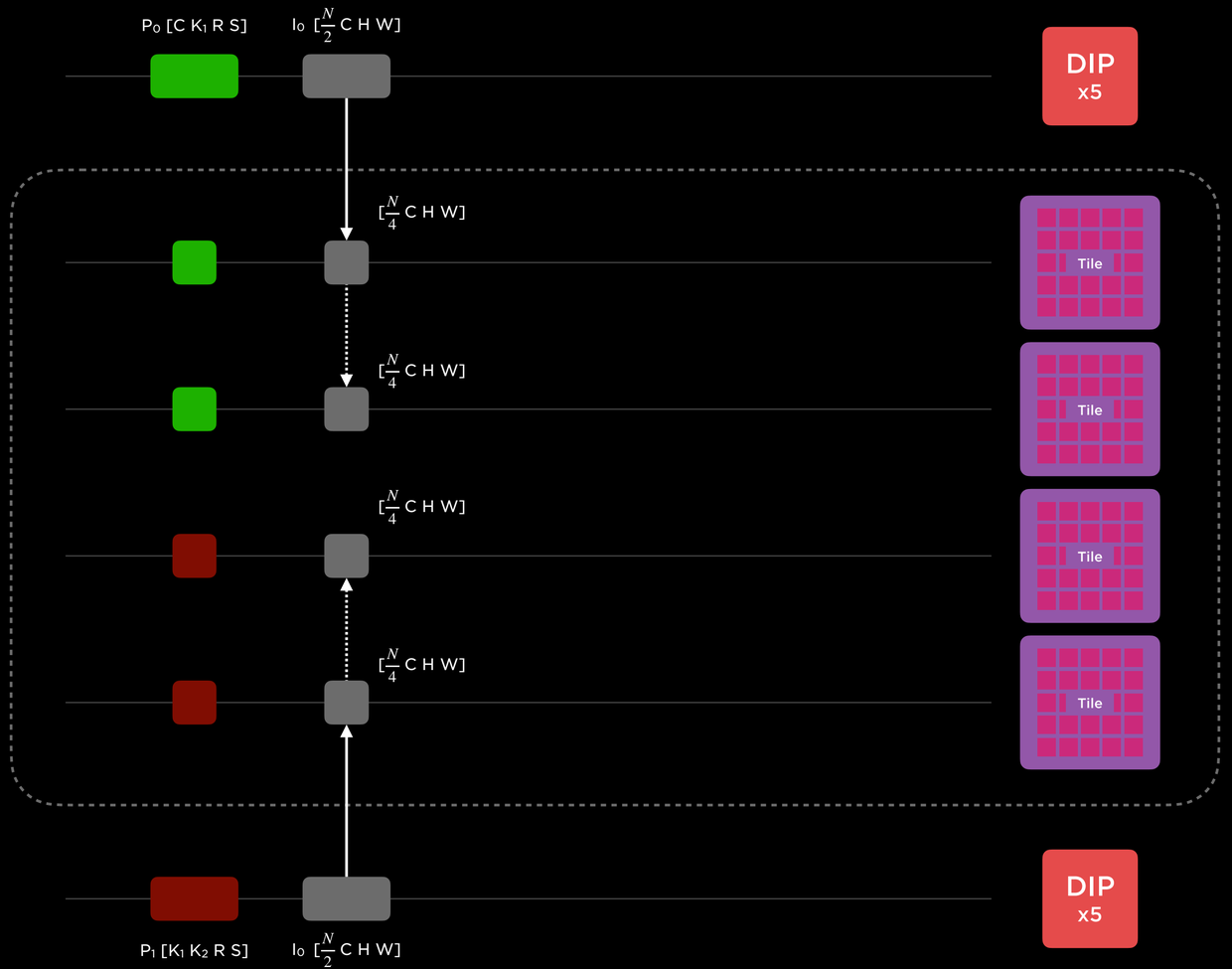
Once per training run

Model Execution



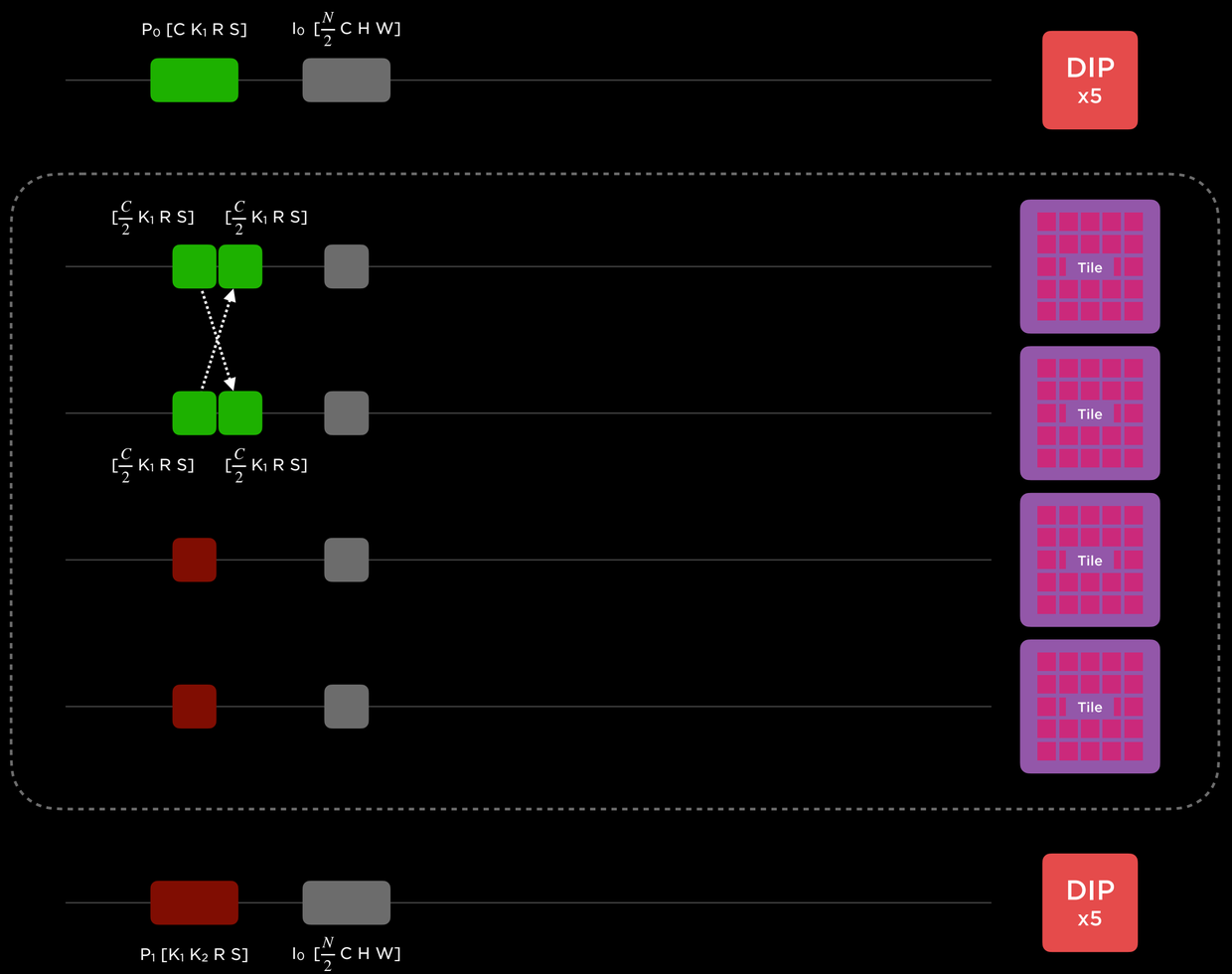
Inputs Sharded Across the DIPs in the Batch Dimension

Model Execution



Inputs Are Also Sharded (by Batch) Across the Tiles

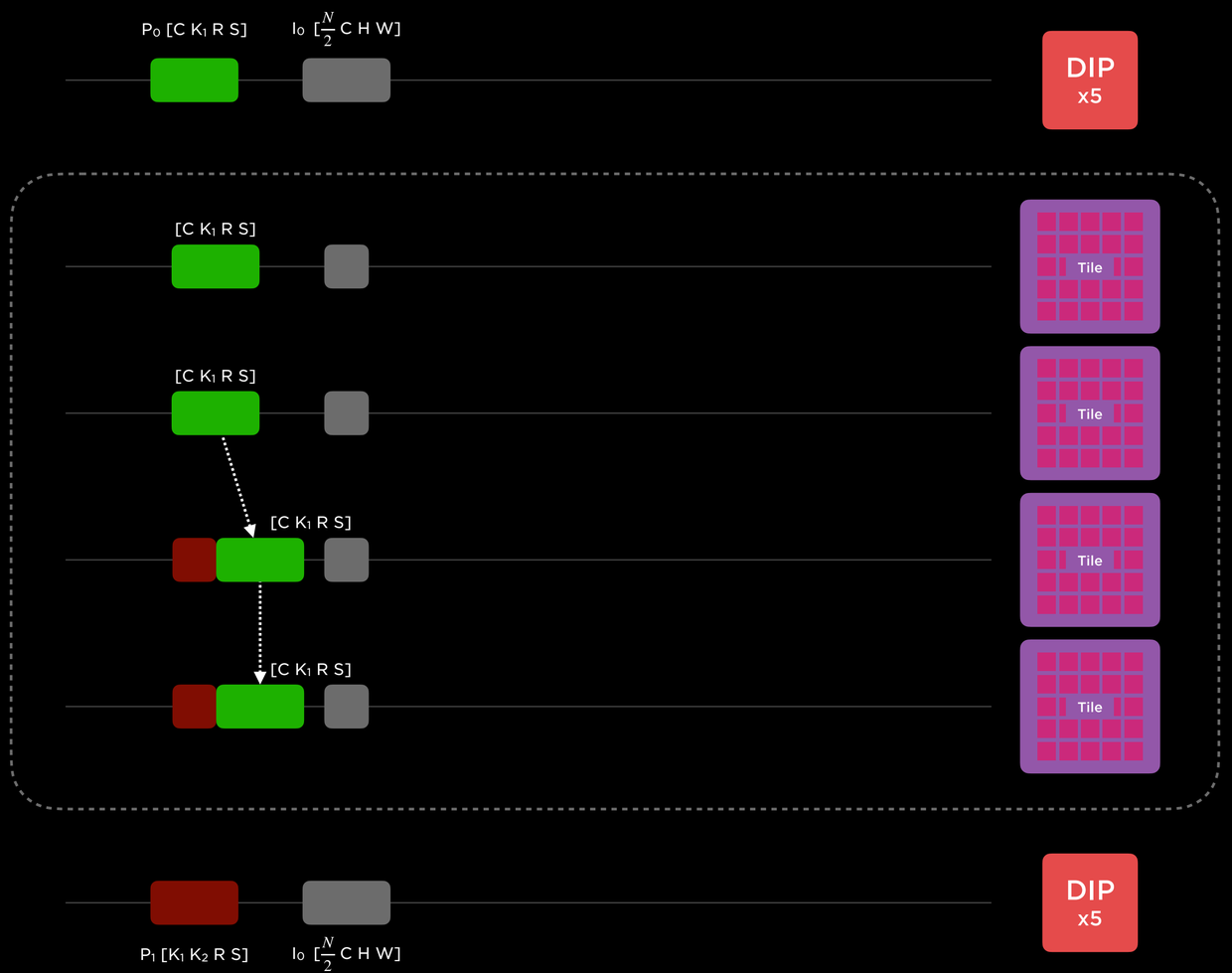
Model Execution



Parameters Are Replicated Across the Tiles Just in Time

A single copy of parameter in the entire system - use the high BW to replicate parameters just in time

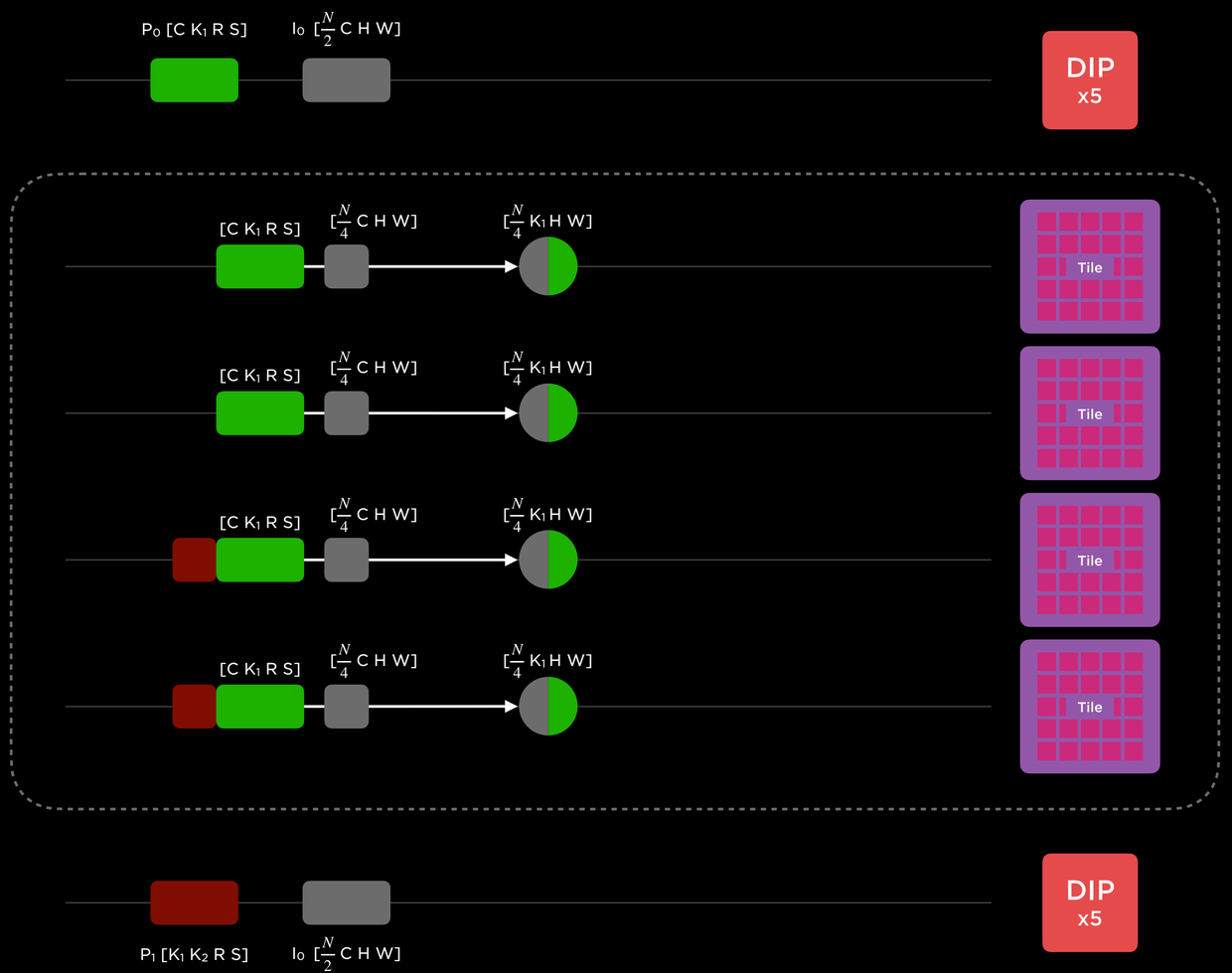
Model Execution



Parameters Are Replicated Across the Tiles Just in Time

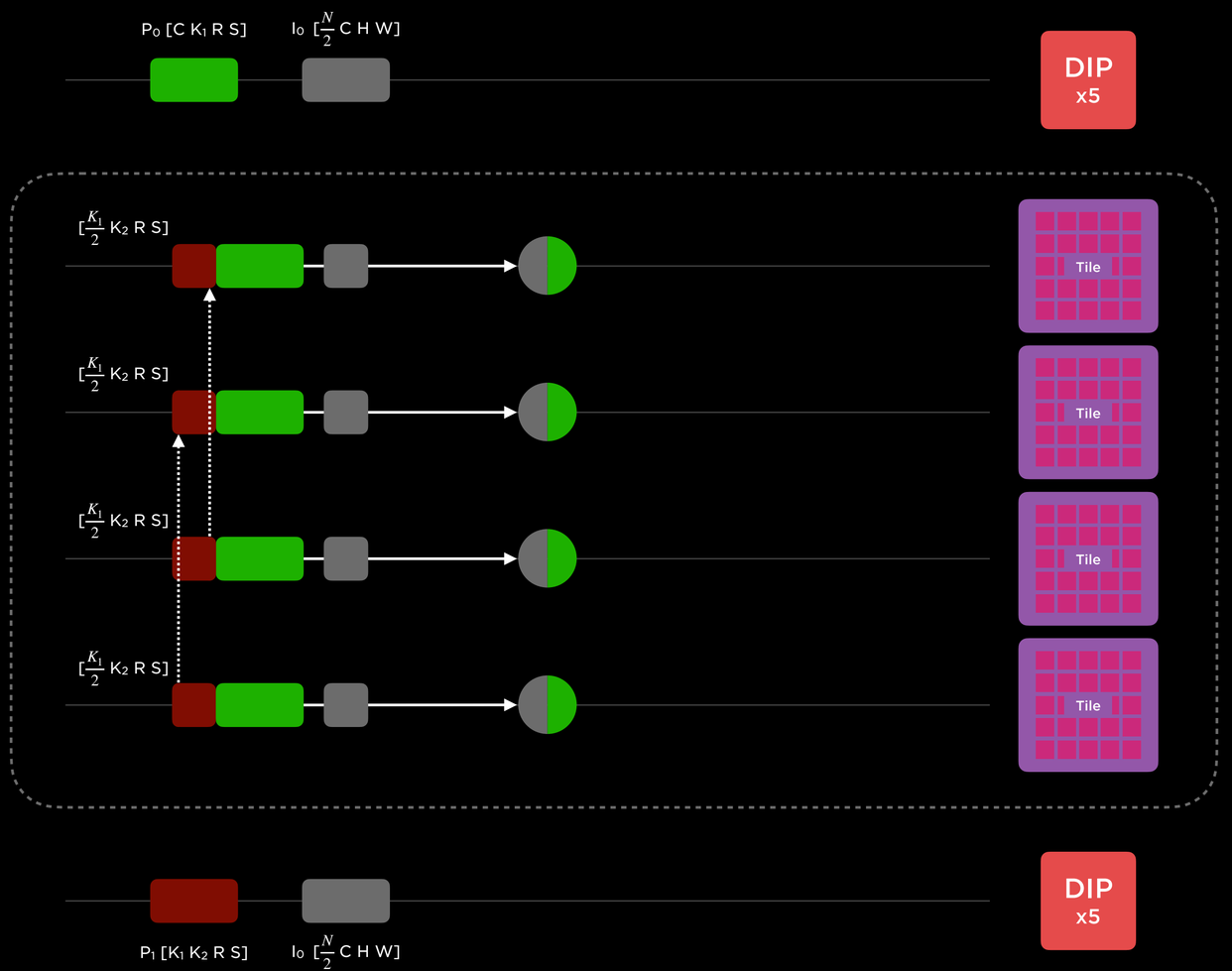
A single copy of parameter in the entire system - use the high BW to replicate parameters just in time

Model Execution



The First Layer Is Run in a Data Parallel Manner

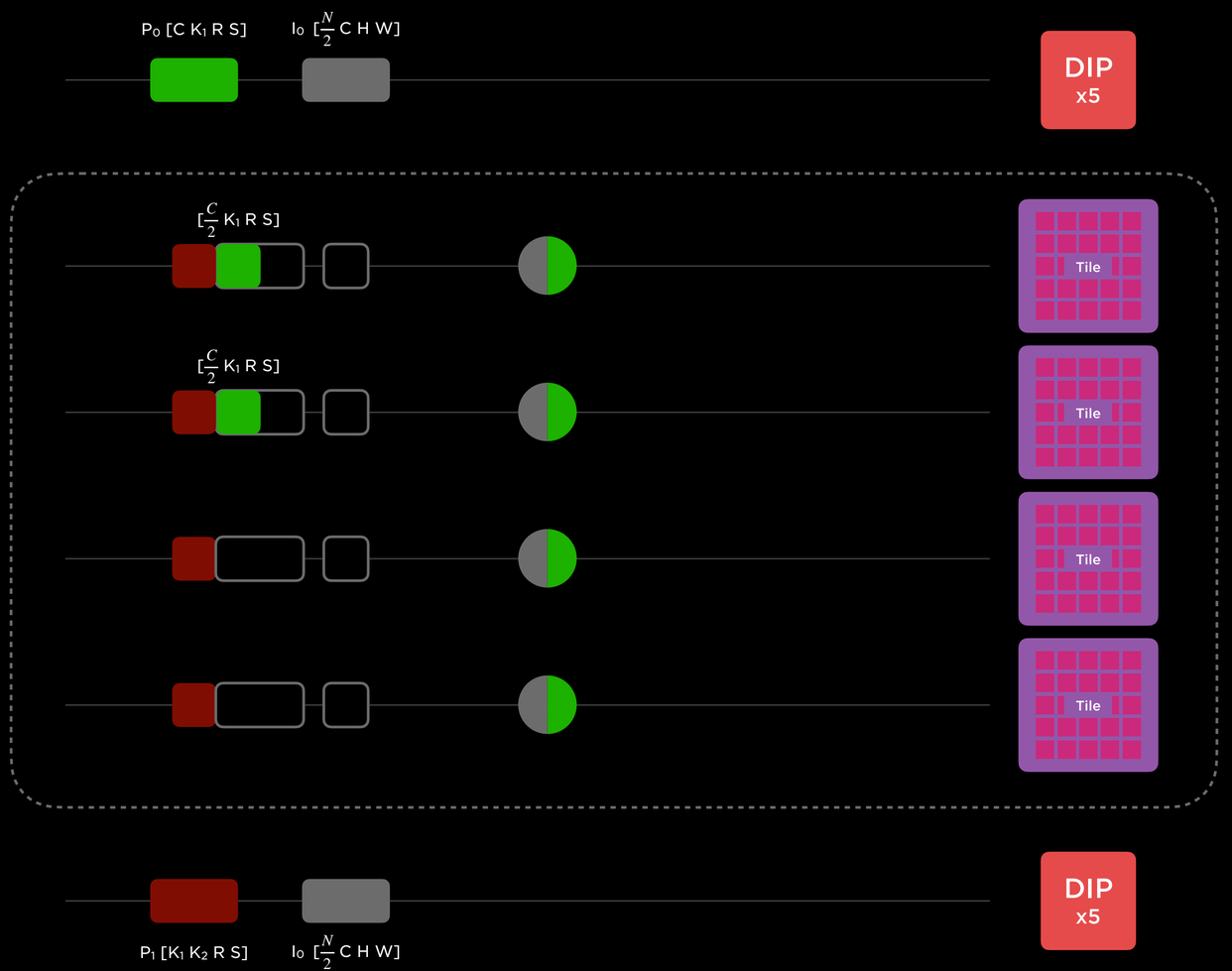
Model Execution



Parameters For the Next Layer Are Replicated Concurrently

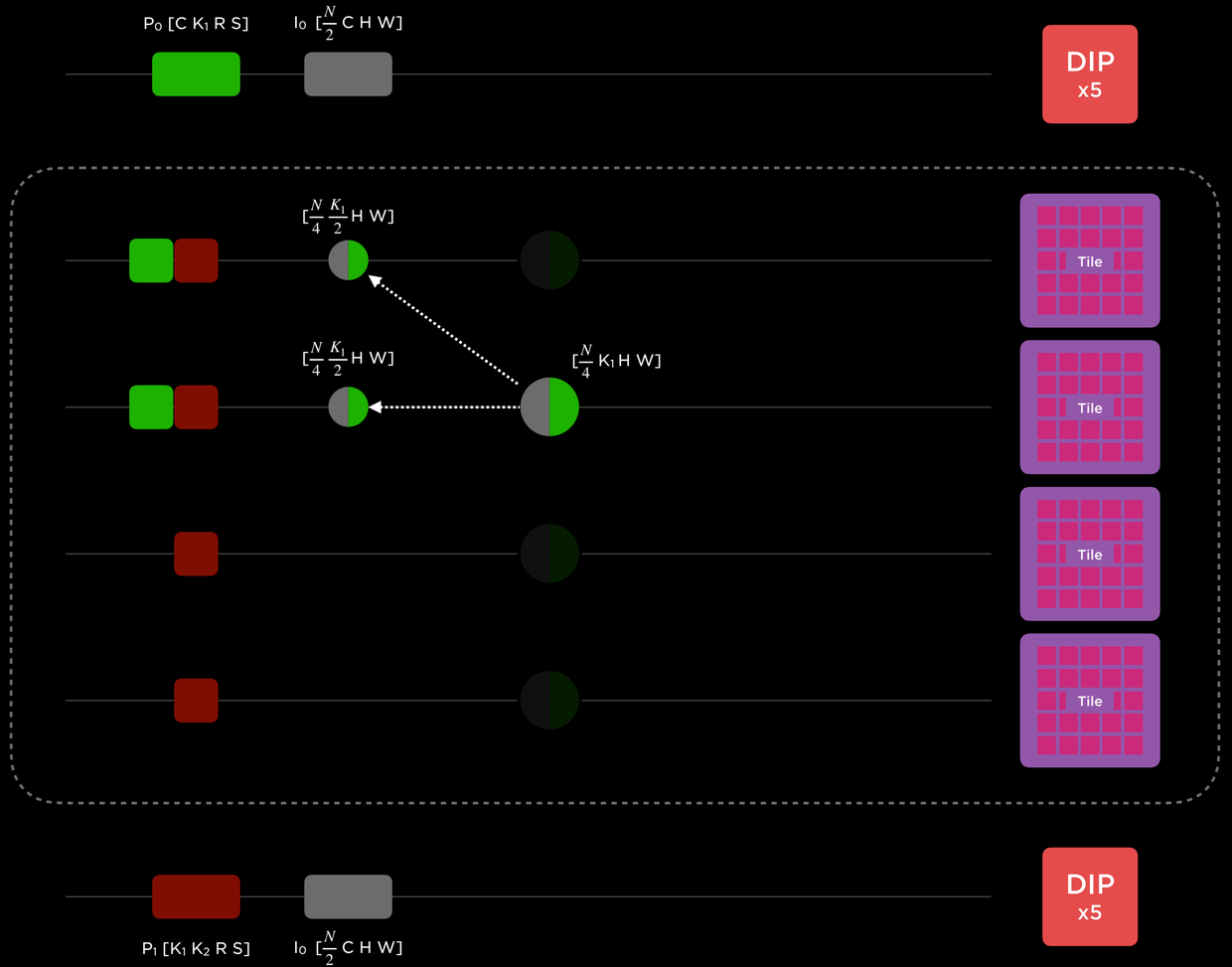
1 copy per 2 tiles. The next layer is better executed in a model parallel manner

Model Execution



Discard Replicated Parameters and Input for Minimal SRAM Footprint

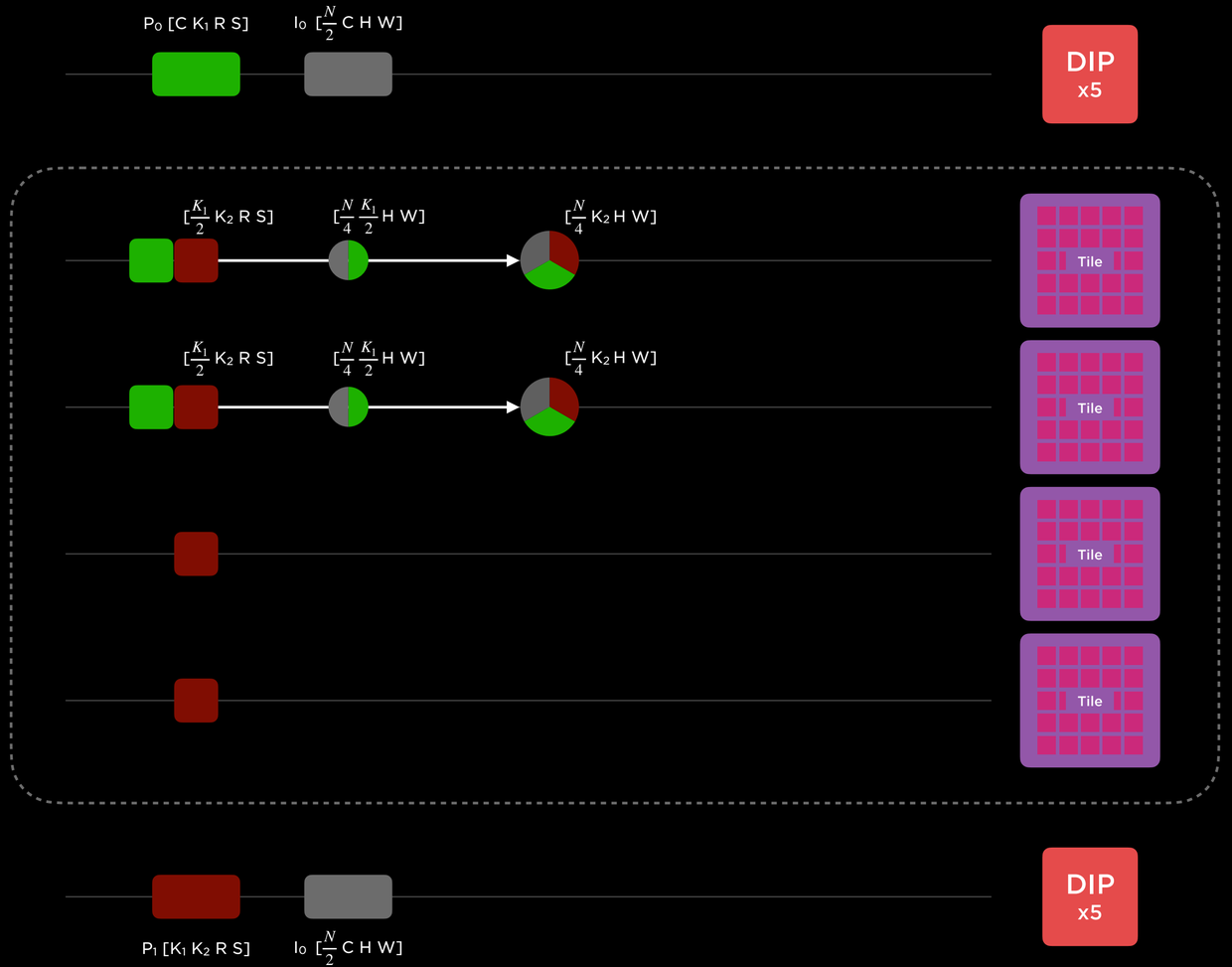
Model Execution



Replicate Input Activation for the Next Layer - Split Across Channels

Only 1 N/4 batch shown

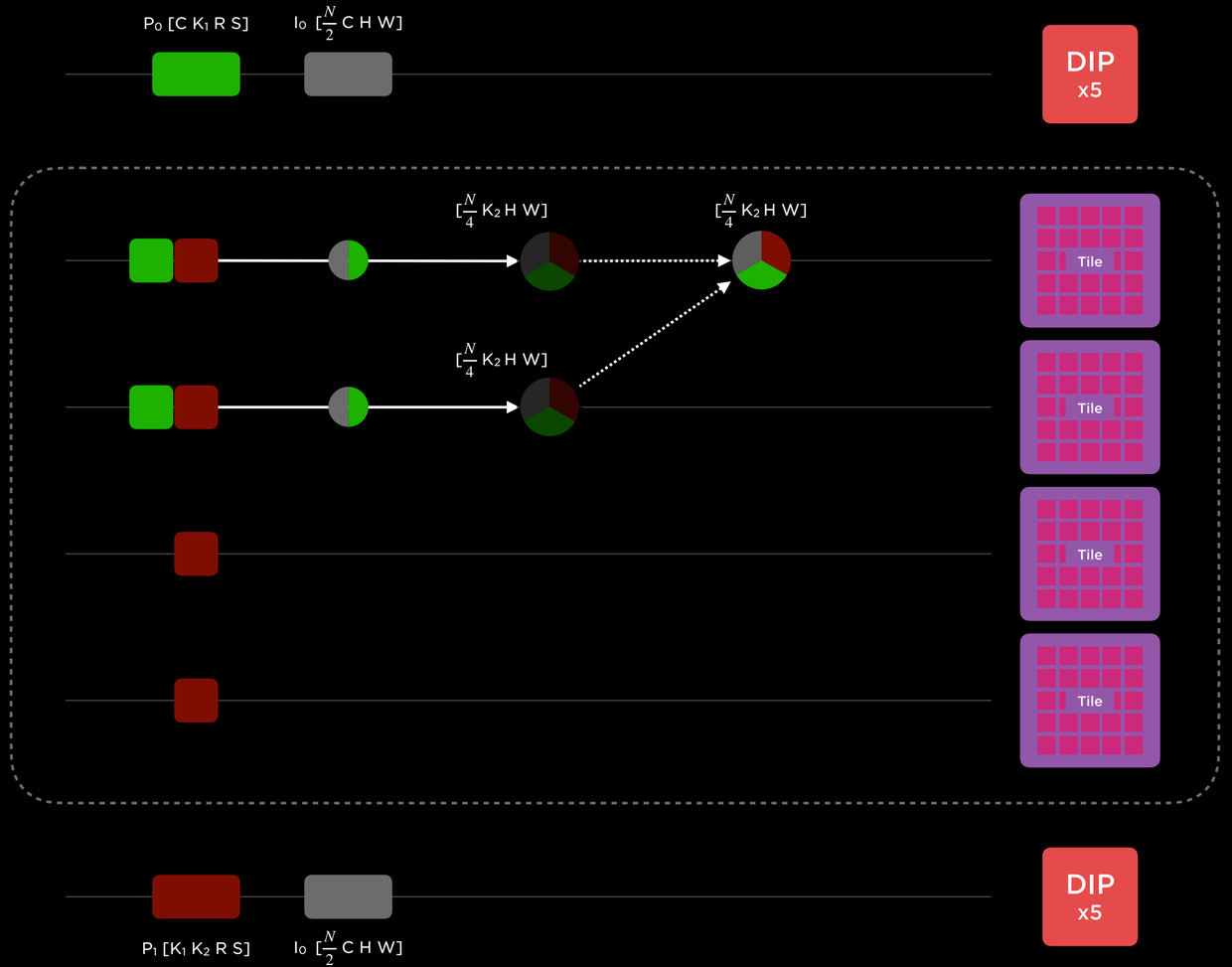
Model Execution



Compute Partial Sum for Each N/4 Batch on Each Tile

Only 1 N/4 batch shown

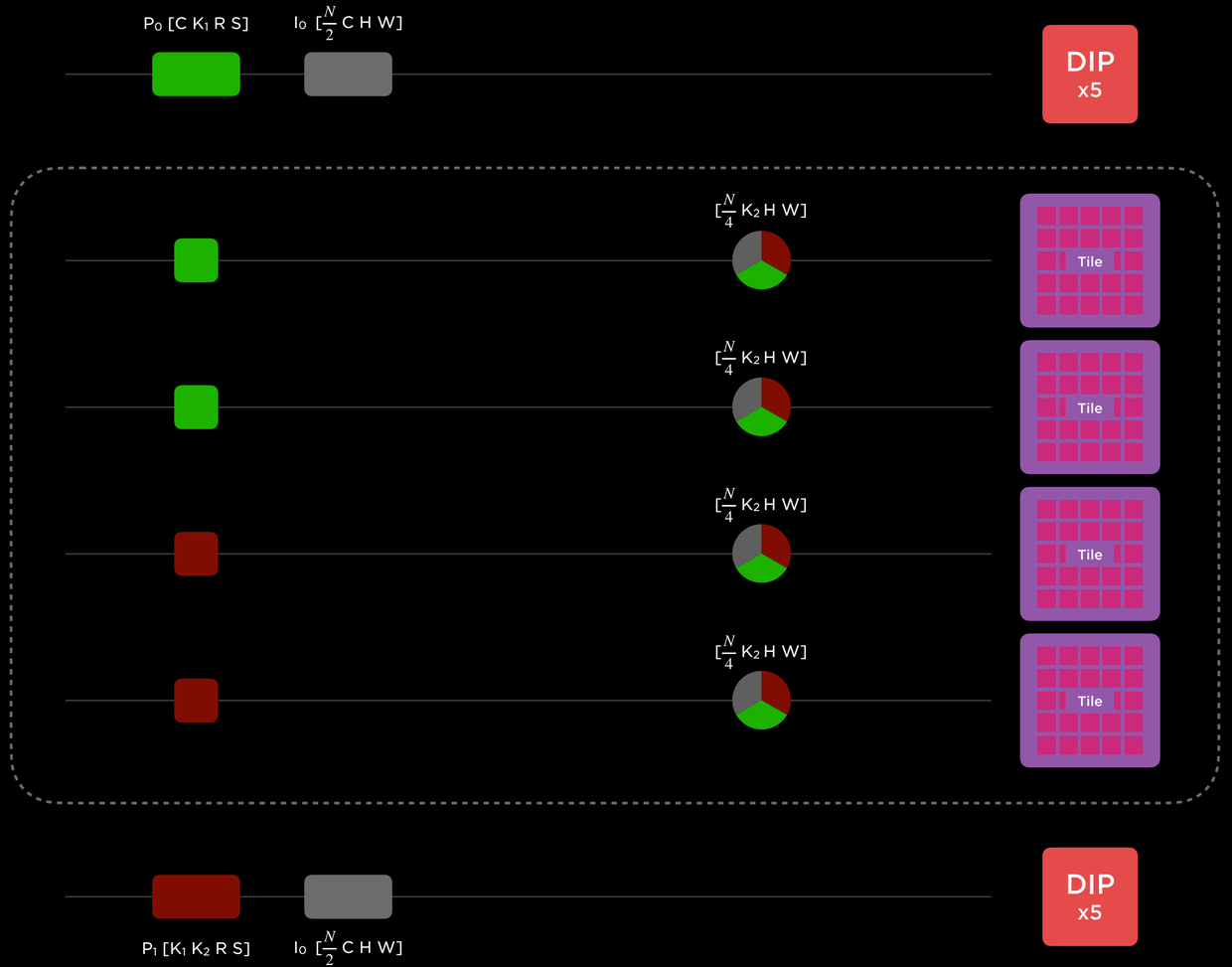
Model Execution



Reduce Partial Sum for Each N/4 Batch Across Tiles

Small packet size, fine-grained synchronization and low-latency network makes pipelined partial sums work

Model Execution



Same Computation Runs on Every Other N/4 Batch

Combination of data and model parallel

End-To-End Training Workflow

File Loading
Decode
Augmentation
Ground Truth Generation

Output
Compression
File Write

Video-Based Training

Data Loading

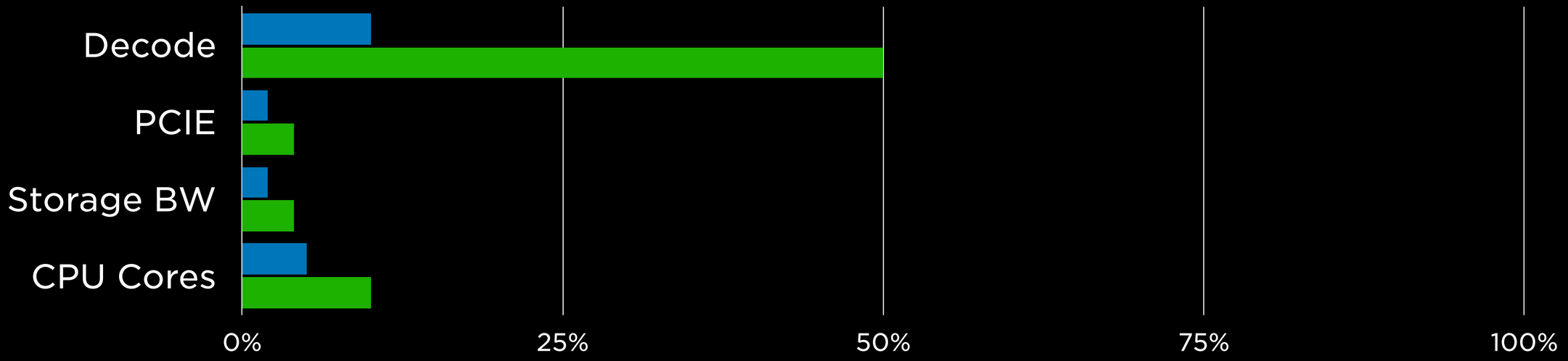
Flexible compute required for:

- Augmentation
- Image rectification
- Ground truth generation

Multi-camera, multi-frame models

- Requires decoding GOP_SIZE/2 frames for first per-camera frame and 1 decode for every frame after

Data Loading Needs of Different Model

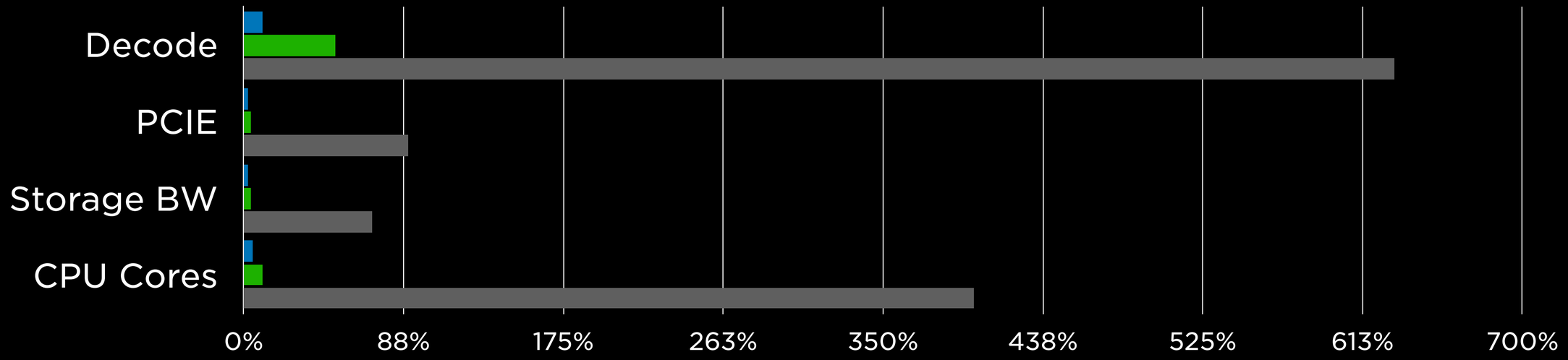


Requirements as % of a Single Host’s Capacity

Model 1

Model 2

Data Loading Needs of Different Models



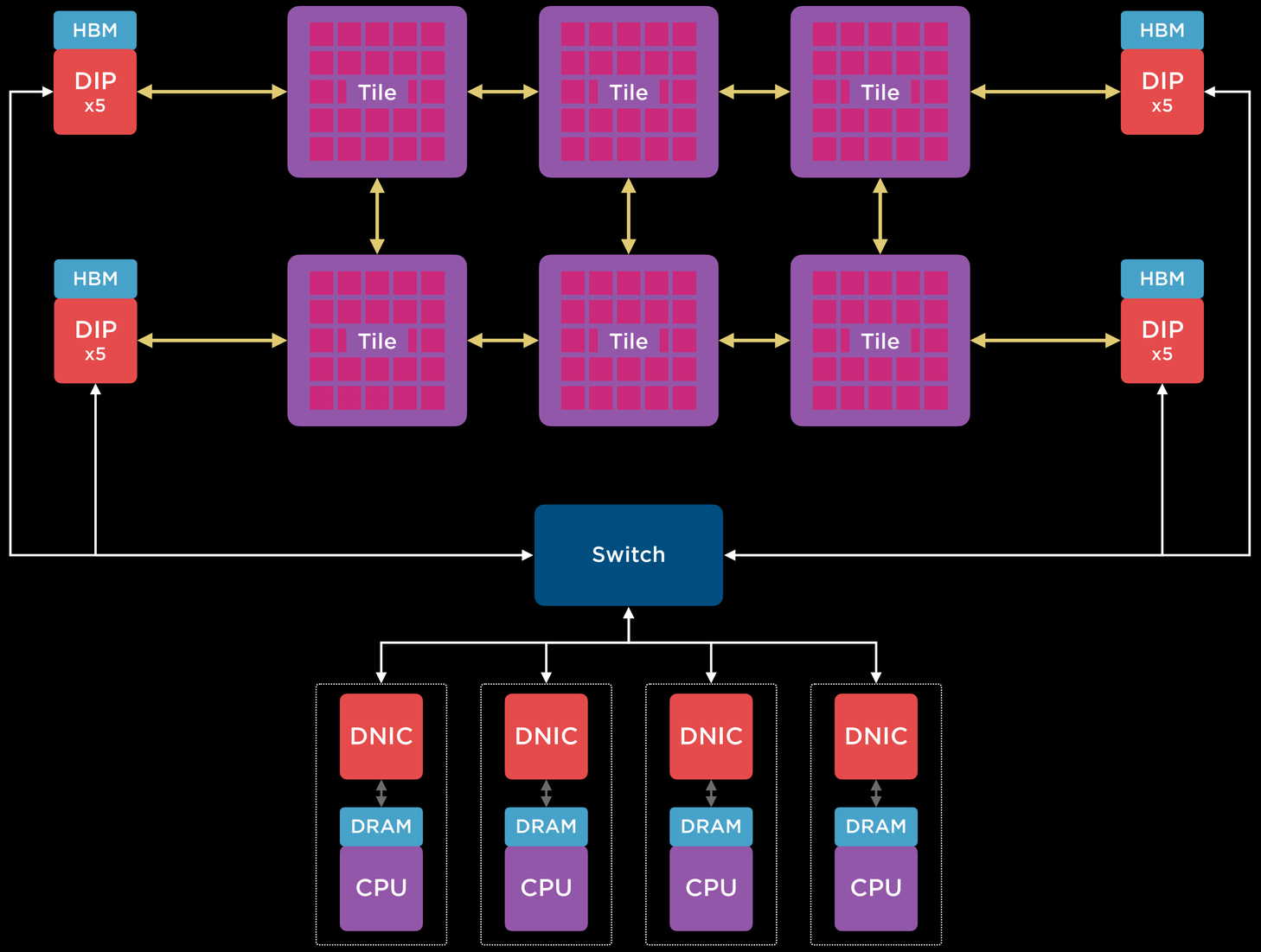
Requirements as % of a Single Host's Capacity

Model 1

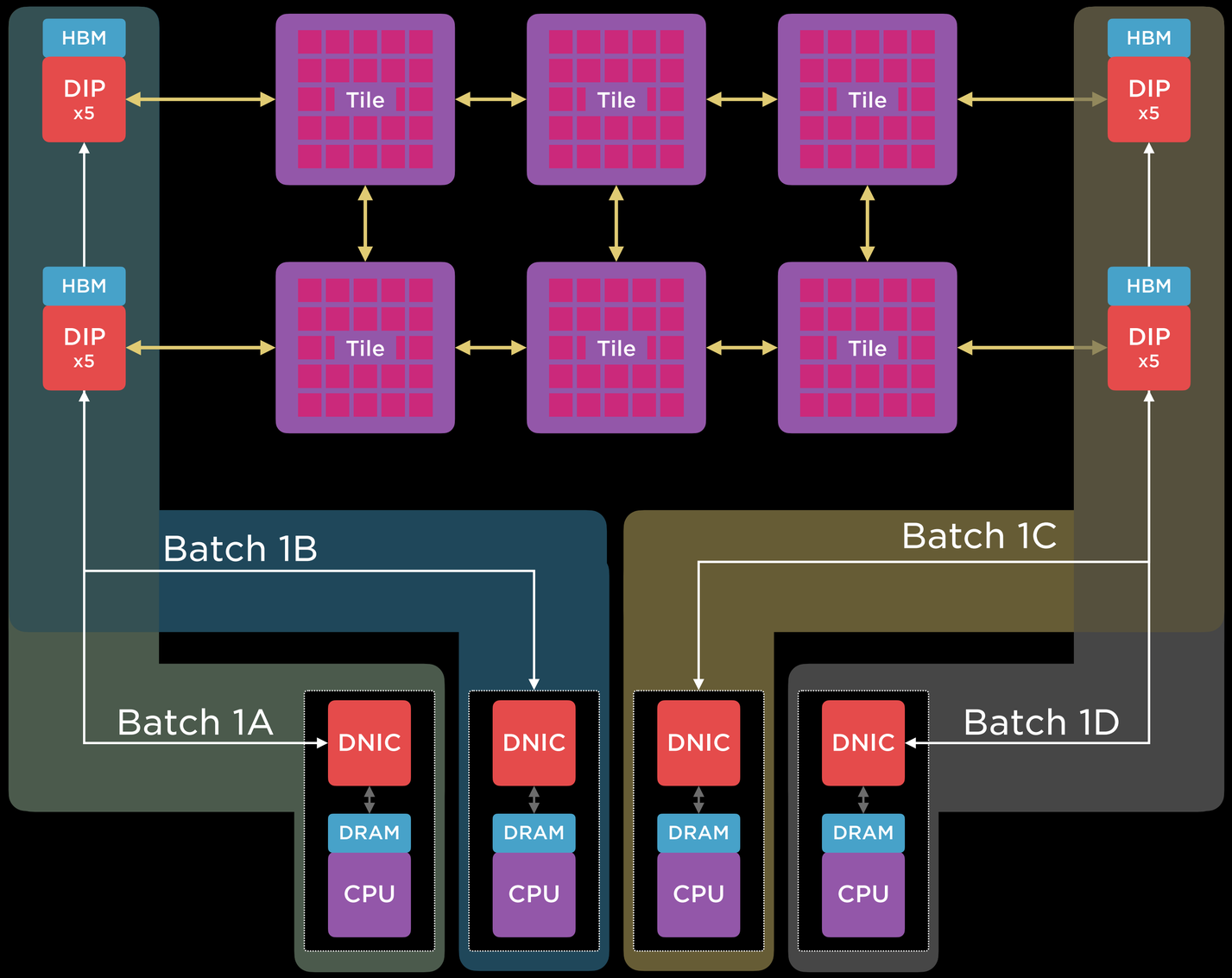
Model 2

Model 3

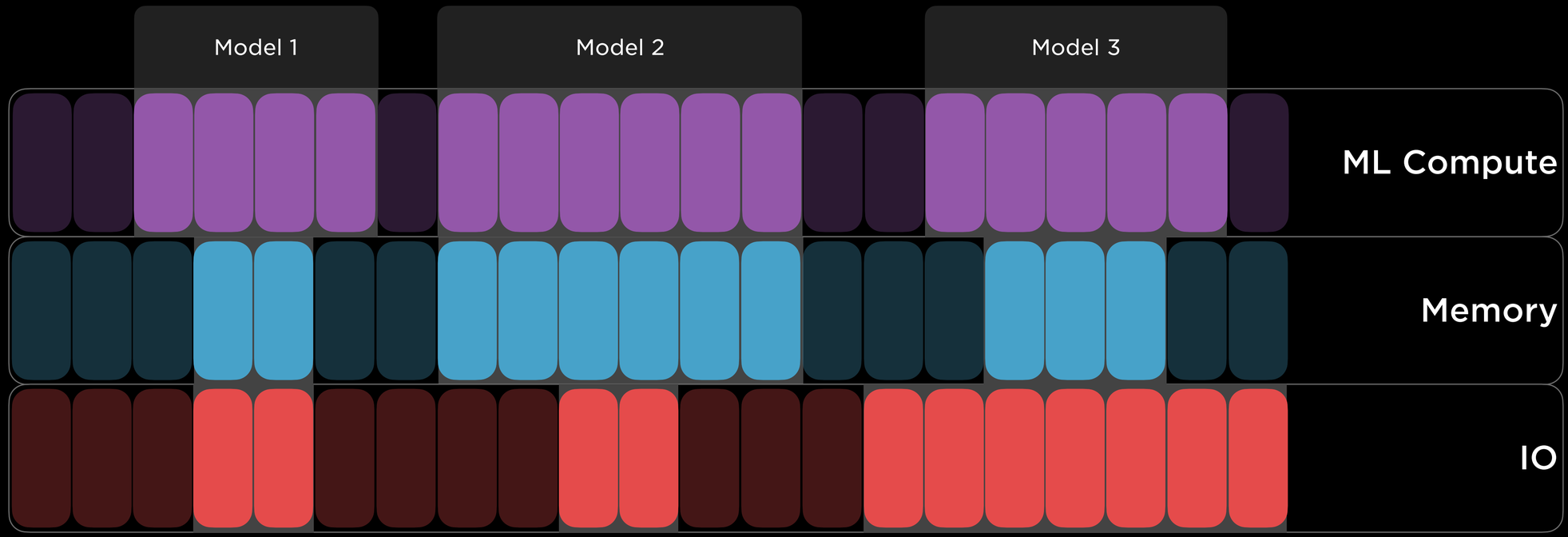
Disaggregated Data Loading Tier



Disaggregated Data Loading Tier



Disaggregated Resources



Resources Can Be Partitioned per Job

Dojo Supercomputer for ML Training

New integration enable high-bandwidth and performance

Uniform high-bandwidth enables full exploitation of parallelism by software

Vertically integrated I/O addresses all workload bottlenecks including data loading