

DOJO

训练系统

The Microarchitecture of Tesla's Exa-Car Computer

Emily Tapp, Douglas Williams, Rob Jitkaas

AI 芯片 - NPU 详解



Talk Overview

I. AI 计算体系

- 深度学习计算模式
- 计算体系与矩阵运算

2. AI 芯片基础

- 通用处理器 CPU
- 从数据看 CPU 计算
- 通用图形处理器 GPU
- AI 专用处理器 NPU/TPU
- 计算体系架构的黄金10年

I. 华为昇腾 NPU

- 达芬奇架构
- 昇腾AI处理器

2. 谷歌 TPU

- TPU 核心脉动阵列
- TPU 系列架构

3. 特斯拉 DOJO

- DOJO 架构

4. 国内外其他AI芯片

- AI芯片的思考

Talk Overview

I. 基本内容

- DOJO 整体架构
- DOJO Core 架构
- DOJO Core 前端处理
- DOJO Core 执行引擎
- SRAM 与内存
- DOJO 训练矩阵
- 问题与思考

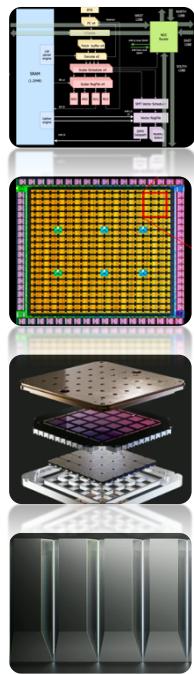
DOJO

整体结构回顾



DOJO 架构设计哲学

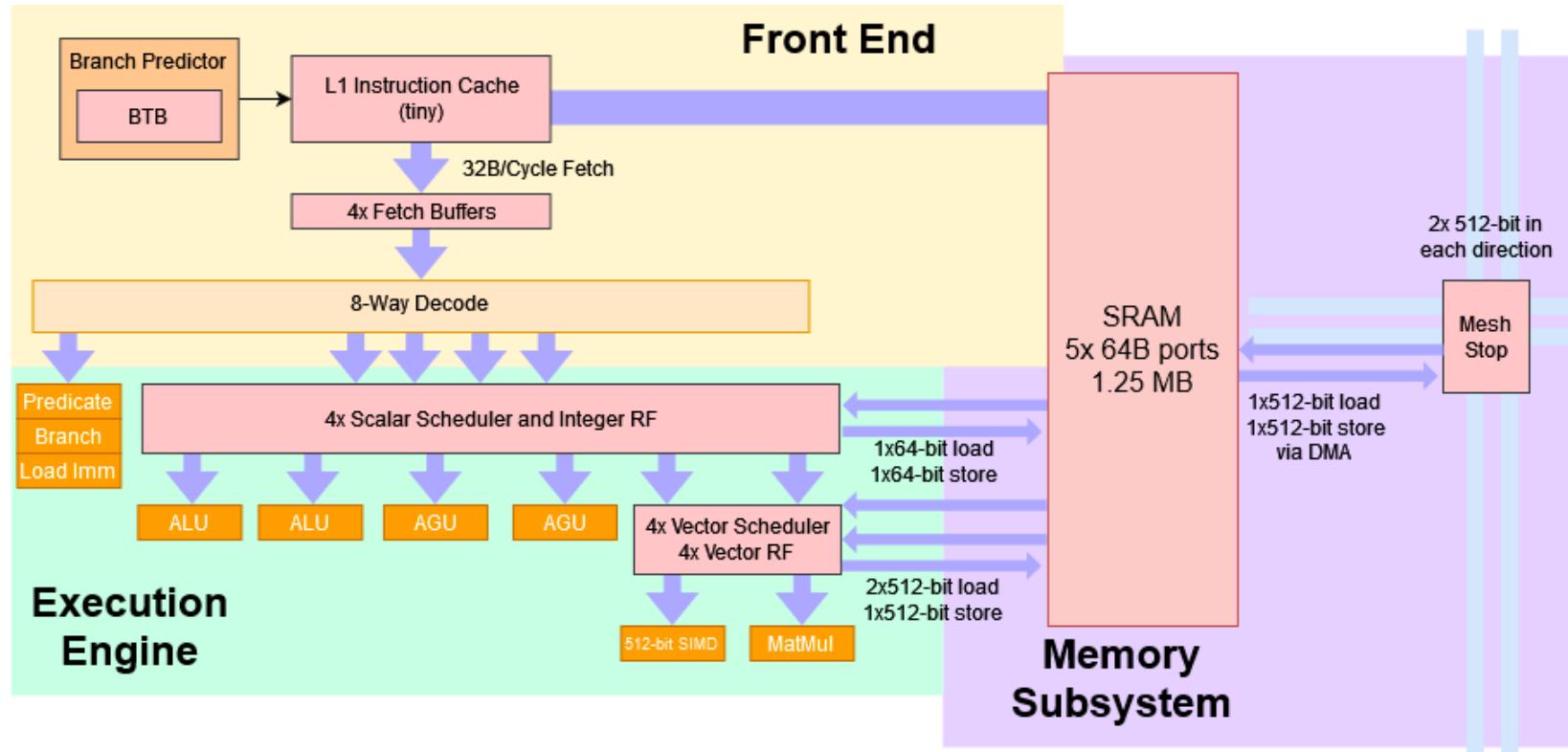
- DOJO 采用存算一体架构（“存内计算”或者“近存计算”），单个可扩展计算平面、全局寻址快速存储器和统一的高带宽+低延迟。



分层	名称	片上SRAM	算力	备注
内核	DOJO Core	1.25MB	1.024 TFLOPS	单个计算核心，64bit，4个8x8x4矩阵计算核心，2GHz主频
芯片	DOJO DI DIE	440MB	362 TFLOPS	单芯片，354核心数，654mm ²
格点	DOJO Tile	11GB	9050 TFLOPS	单个训练模组，每5x5个芯片组成一个训练模组
集群	ExaPOD	1320GB	1.1 EFLOPS	训练集群，每12个训练模组组成一个机柜，每10个机柜组成一个ExaPOD，一共3000个DI芯片

DOJO Core 的结构和特点

- DOJO Core 由前端、执行引擎、SRAM 和 NoC 路由4部分组成，比CPU和GPU的控制部件都更少，具有类似CPU的AGU和类似GPU Tensor core 的矩阵计算单元。



DOJO 训练矩阵



BUILDING A BETTER CONNECTED WORLD

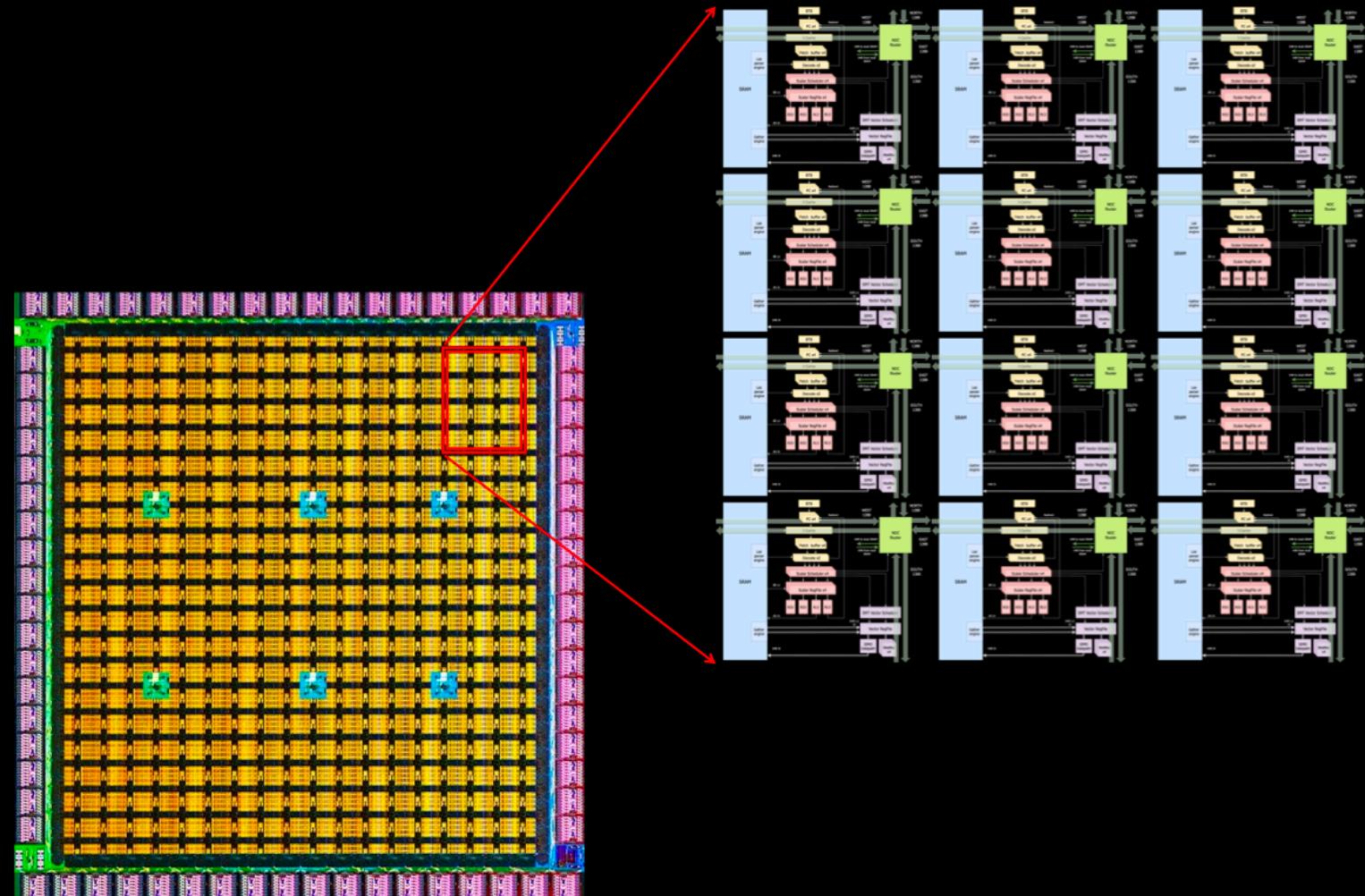
Ascend – ZOMI

7

 www.hiascend.com

D1 Die : 354 个 DOJO Core 和 440MB SRAM

- 片上网络路由：将多个 Core 连接在一起形成一个 2D 网格。
- NOC 路由：可以处理跨节点边界的数据包（boundary），每个方向 64 B；每个时钟周期，在四个方向上一个数据包输入和一个数据包输出到网格中每个核心最近邻居。
- 路由读写：该路由器还可以在每个周期对本地 SRAM 进行一次 64 B 读取和一次 64 B 写入，因此可以在内核之间移动数据。

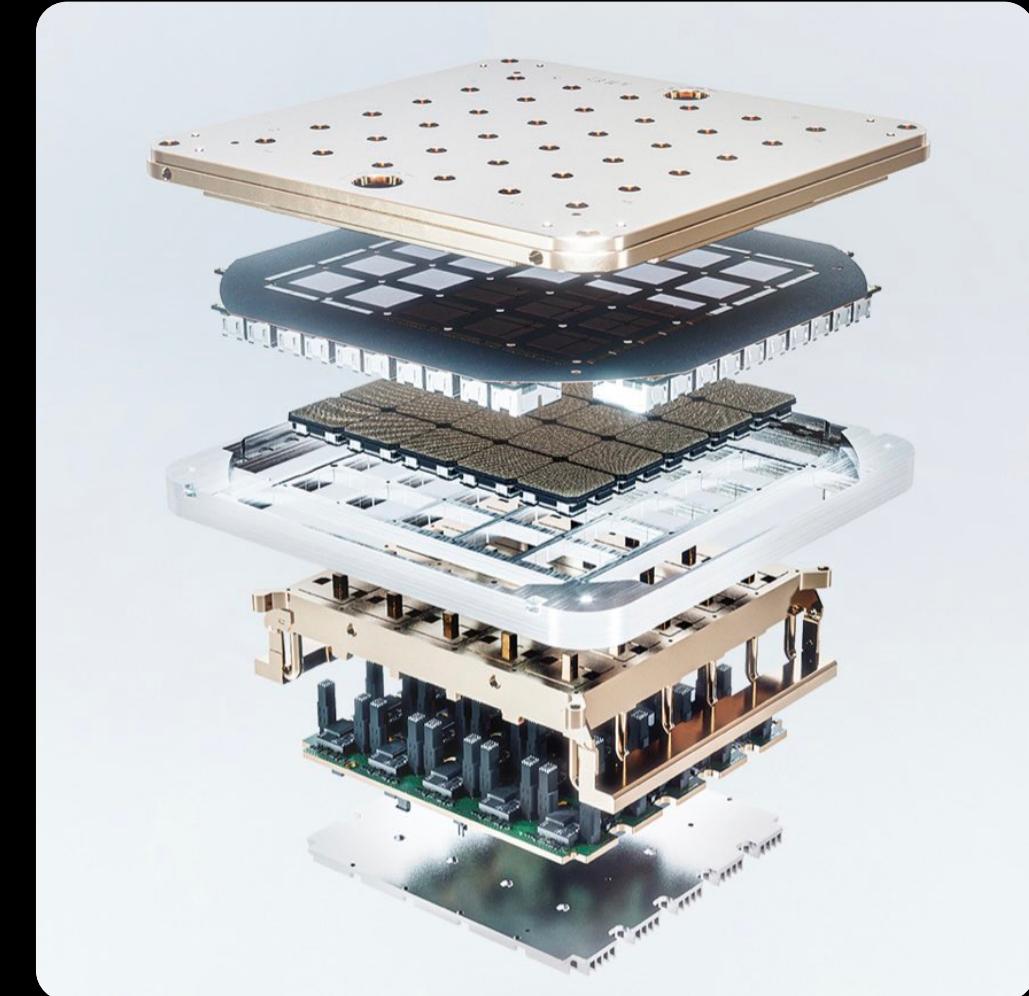


D1 Die

核心数	354(18x20)	工艺节点	7nm
算力	362TFLOPS@BF16/CFP8 22TFLOPS@FP32	主频	2GHz
片上SRAM	440MB	面积	645MM ²

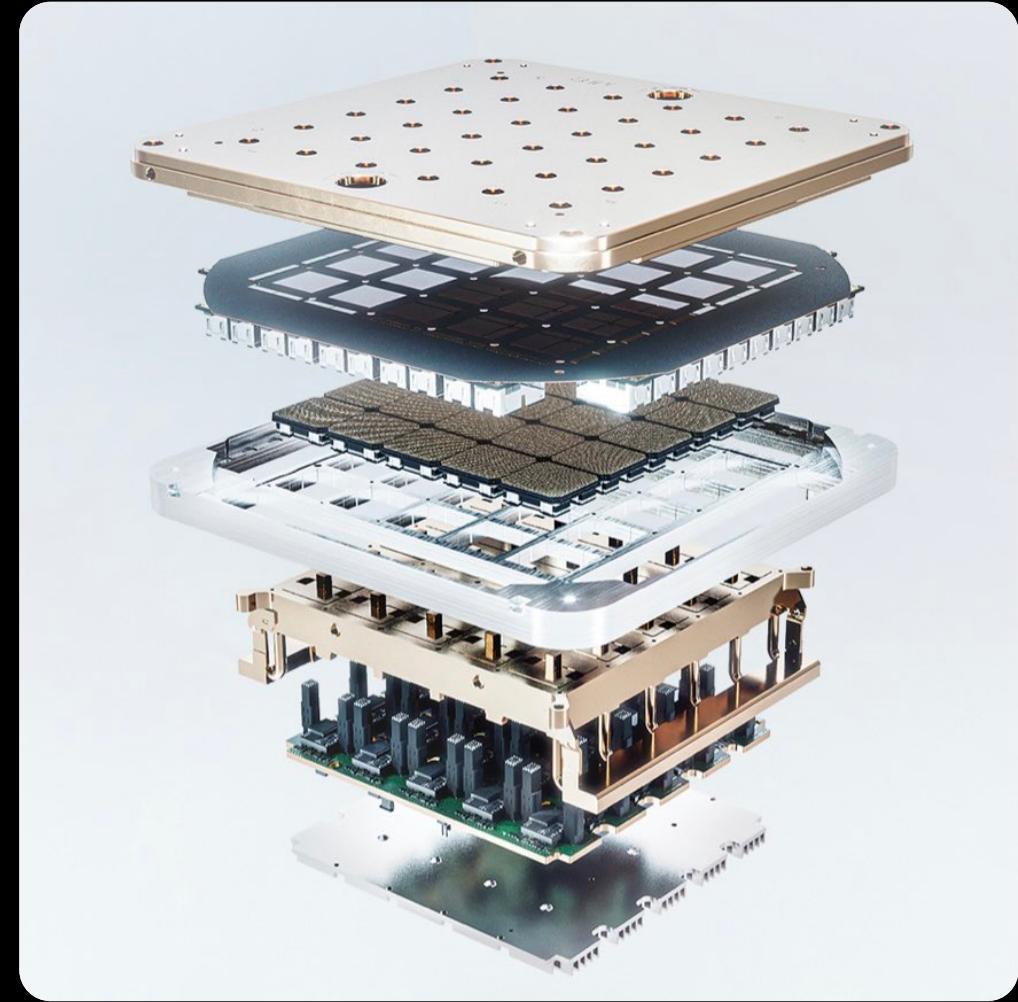
训练 Tile

- DI Die 被组装成一个 5×5 的 Training Tile , 每个边缘有 4.5TB/s 的带宽。DI 芯片以 2 GHz 运行，在这些内核上共有 440 MB 的 SRAM，在 BF16 或 CFP8 上提供 376 teraflops，在 FP32 上提供 22 teraflops。每个训练 tile 总共有 11 GB 的 SRAM 主存。
- 每个Dojo核心都是一个相对完整的带矩阵计算能力的独立单元，数据直接在各个处理核心之间流转，无需回到内存。



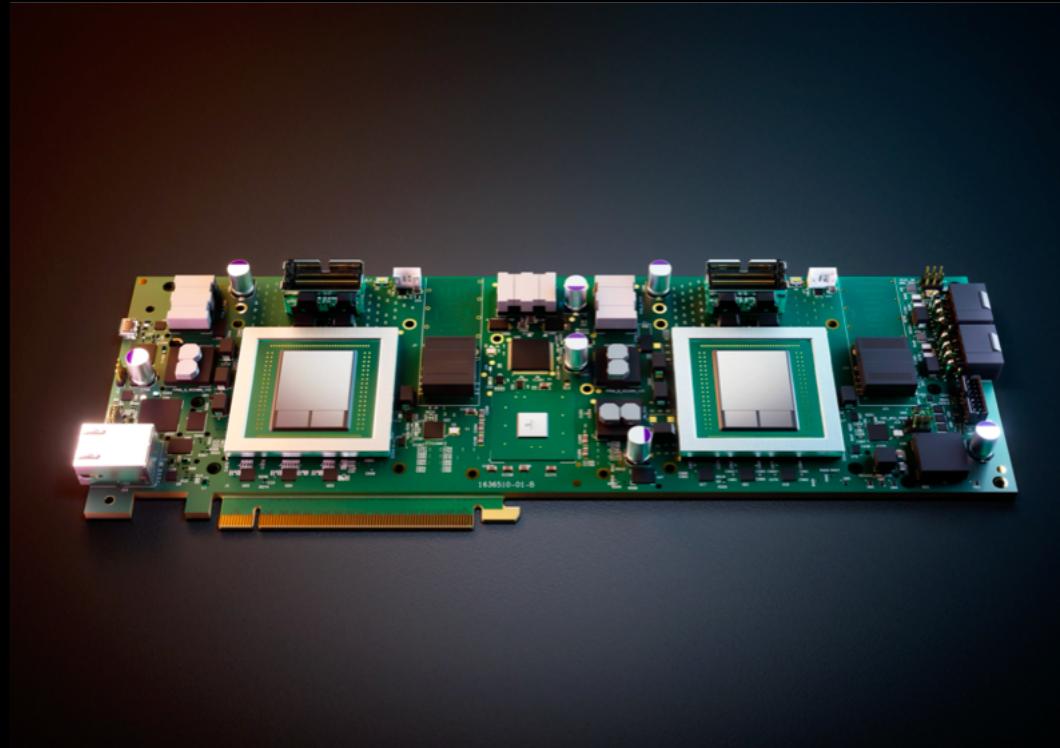
训练 Tile

- SRAM：DI芯片运行在2GHz，拥有巨大的440MB SRAM。特斯拉将设计重心放在计算网格中的分布式SRAM，通过大量更快更近的片上存储和片上存储之间的流转减少对内存的访问频度，来提升整个系统的性能，具有明显的存算一体架构特征。
- 每颗DI 芯片有 576 个双向 SerDes 通道，分布在四周，可连接到其他 DI 芯片，单边带宽为 4 TB/秒。
- 接口处理器：位于2D网格的边缘，每个训练块有11GB的SRAM和160GB的共享DRAM。



DIP : Dojo Interface Processor

- **内存**：有两个 I/O 处理器，每个处理器带两个 HBM 内存组，提供 32 GB HBM 内存和 800 GB/s 带宽。
- **传输**：通过 PCI-Express 实现 Tesla 传输协议 (TTP)，将完整 DRAM 内存带宽提供给 DOJO 训练 Tile。
- **网络**：50 GB/s TTP 协议链路在以太网 NIC 上运行，连接到 400 Gb/s 或成对 200 Gb/s 端口。



DOJO 超级计算机系统

- TTPOE 可将标准以太网转换至 Z 平面拓扑，拥有高 Z 平面拓扑连。最多可以将 5 个 DIP 以 900GB/s 的速度连接到一个训练瓦片上，以达到 4.5TB/s 的总量，每个训练 Tile 共有 160GB 的 HBM。

DISAGGREGATED HIGH SPEED MEMORY

Standard PCIe form factor
20x cards per tray & 32GB HBM per card

HIGH-BANDWIDTH INGEST

PCIe & Ethernet connectivity

HIGH RADIX Z-PLANE CONNECTIVITY

Shortcuts across compute plane

640 GB HIGH-BANDWIDTH DRAM

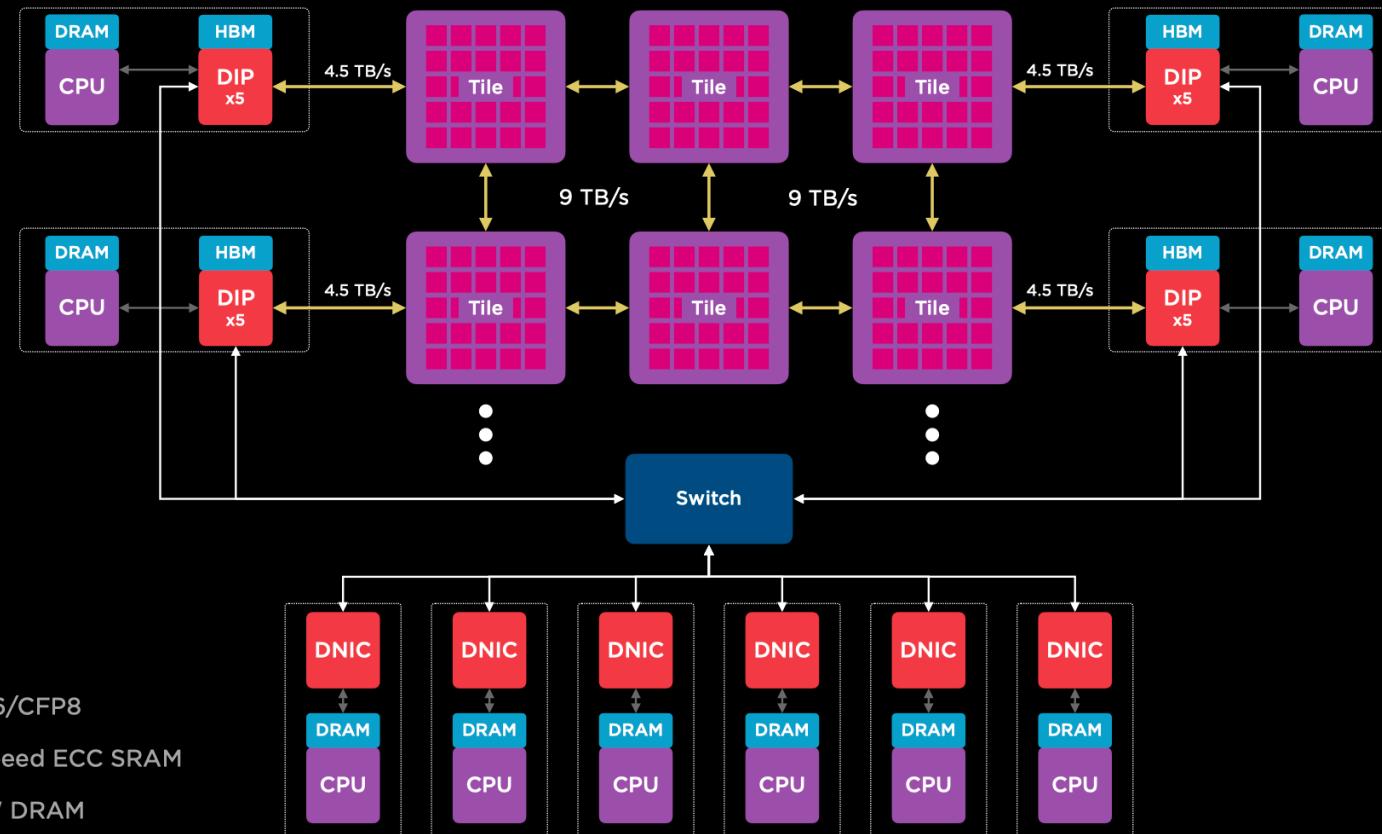
1 TB/S ETHERNET BANDWIDTH

18 TB/S AGGREGATE BANDWIDTH TO TILES



训练矩阵

- Dojo VI 训练矩阵是 Tesla 正在构建的基础系统，它有 6 个训练 tile、4 个主机服务器上的 20 个 DIP，以及一组连接到以太网交换结构的服务器：



DOJO 超级计算机系统

- DOJO 主机接口

INGEST PROCESSING

PCIe connectivity to Interface Processors
Hardware Video Decoder Support

USER APPLICATIONS

x86 Linux Environment
User-Scheduled jobs

512 TOTAL X86 CORES

8 TB TOTAL MEMORY

640 GB/S PCIE BANDWIDTH

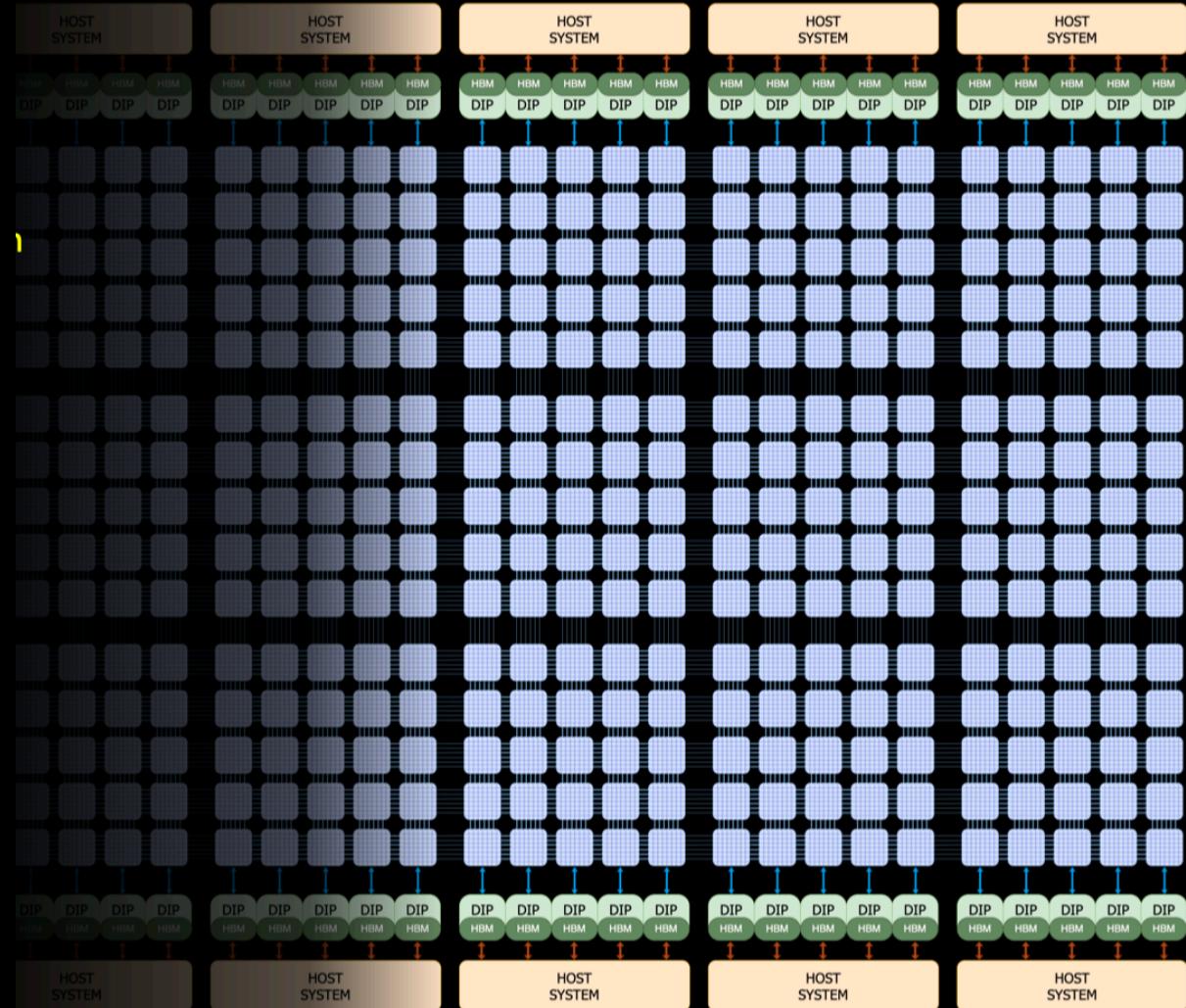


DOJO 通信方式



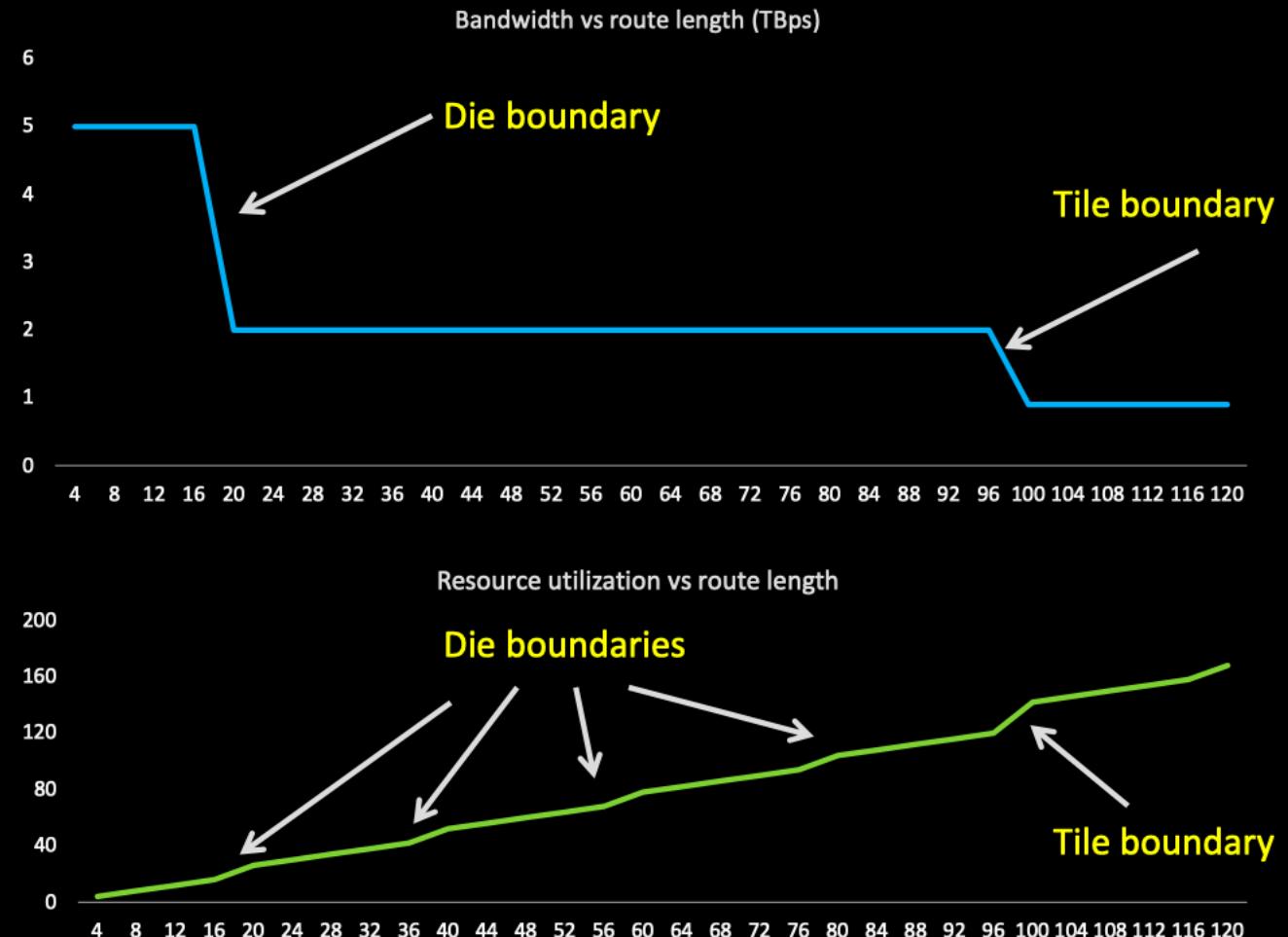
通信方式

- **HBM DIP**：连接到 DI 网格以及为 DIP 供电并执行系统管理的主机系统功能。每个DIP和主机系统提供32GB/s的链接。
- 系统需要某种更大内存时，Tesla 选择创建一个 DIP 内存和 I/O 协处理器，其中包含 32 GB 共享 HBM 内存，以及以太网接口到外部世界以及在tile和核心之间进行比通过这个巨大的网格更直接的跳跃。



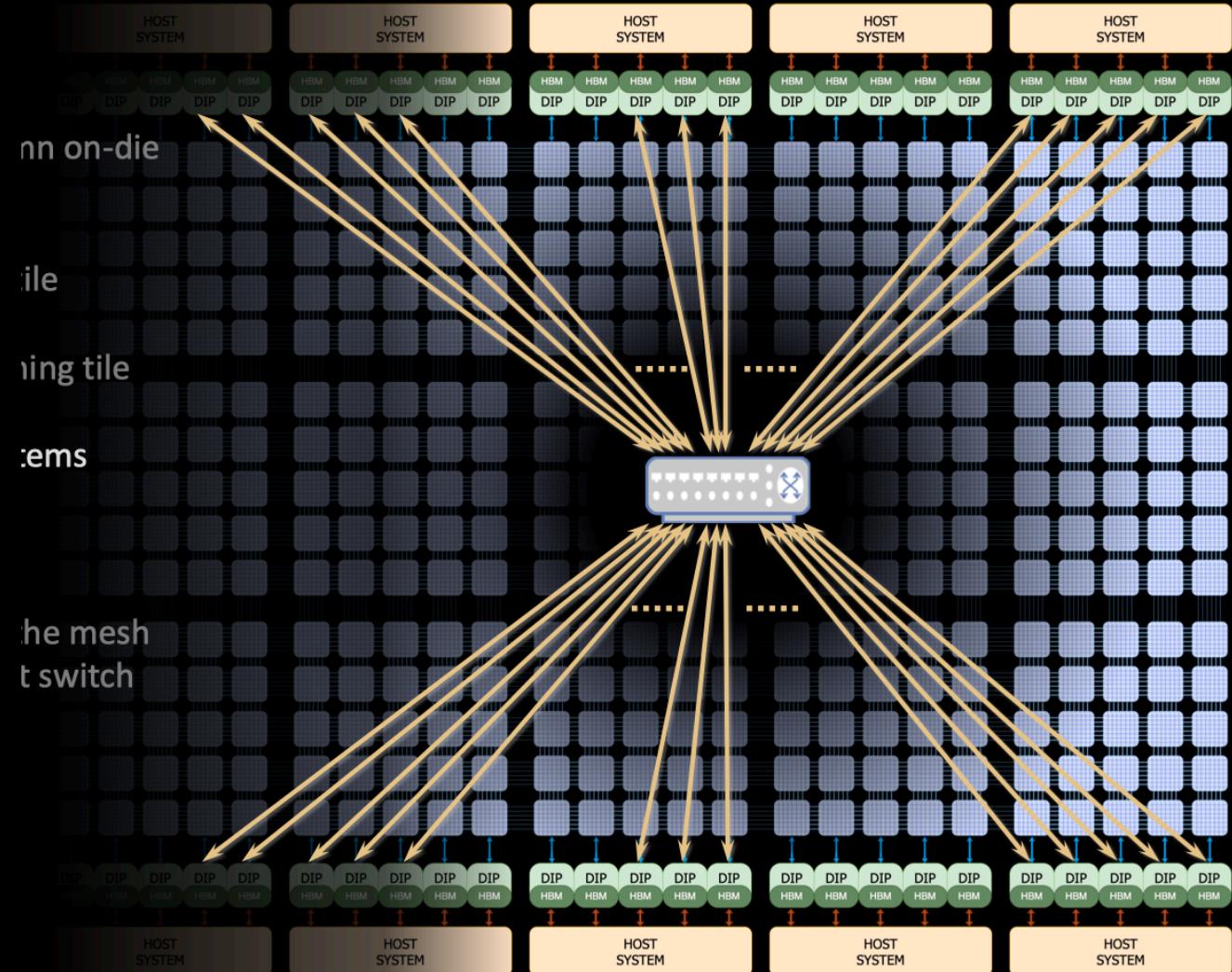
通信方式

- 右图是训练 DIE 和 TILE 延迟边界，传输的链路越长，带宽限制越明显。资源利用率越底。
- 这就是为什么在Dojo中对它们进行不同处理的原因，需要Z平面链路的原因是，长路径很昂贵。



Z平面链路

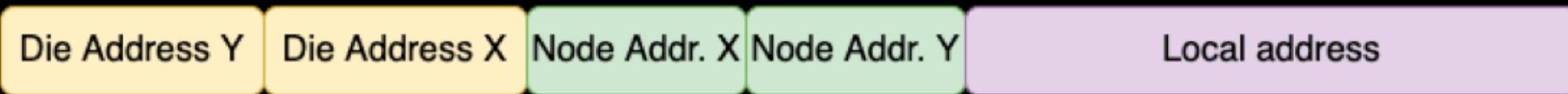
- DIP 不仅将 DRAM 实现为本地存储，而且提供了另一种网络维度（用于更长路线的Z平面链接），可用于绕过 2D 网格。
- 任何处理节点都可以跨系统访问数据。每个节点都可以将数据推送或拉取到SRAM或DRAM。



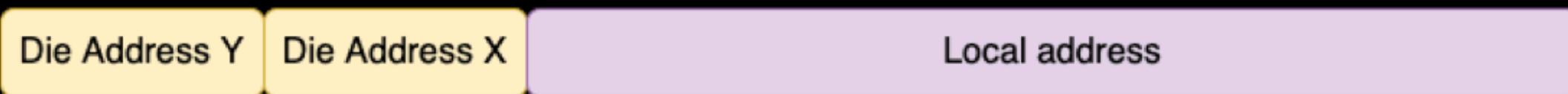
ISA

- Dojo 使用平面寻址方案进行通信。

D1 address format:

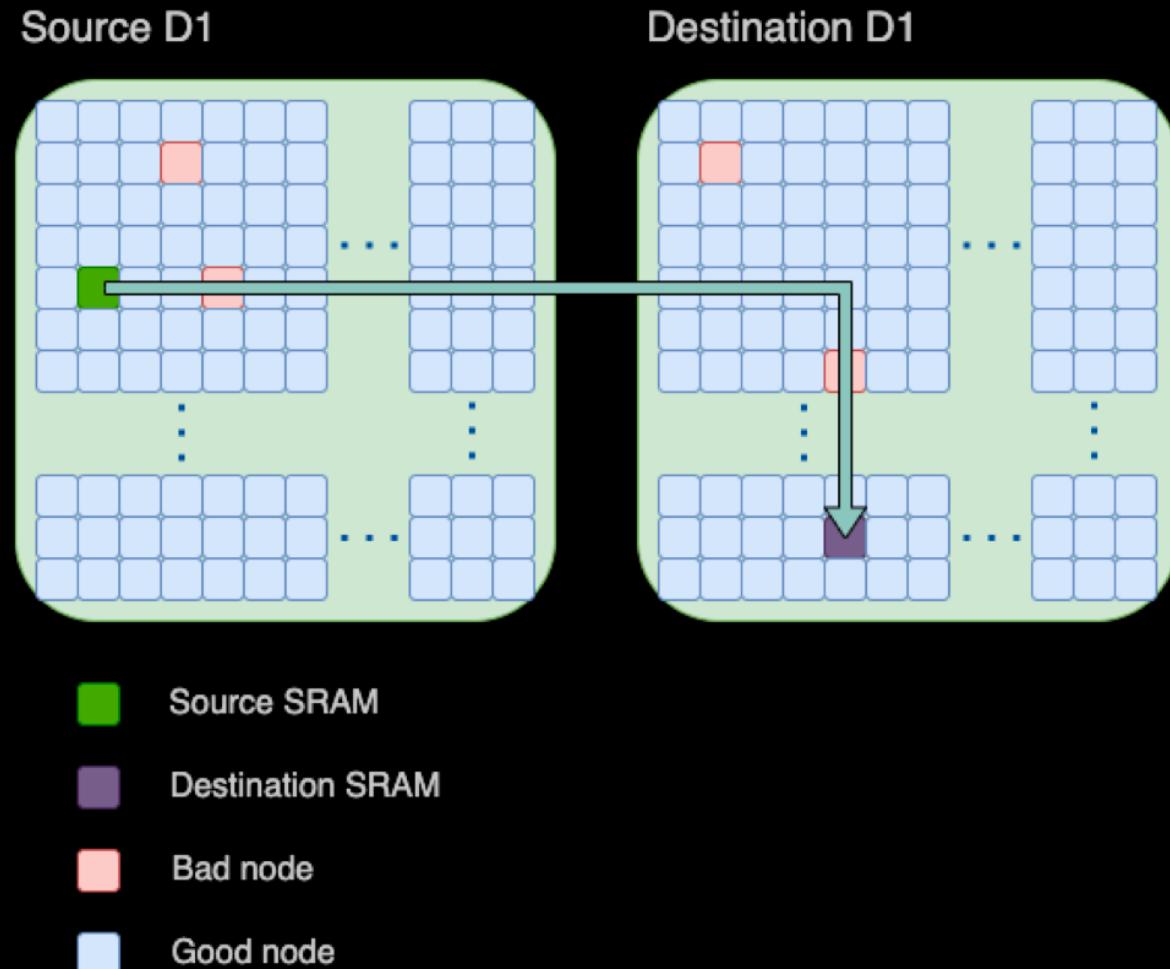


DIP address format:



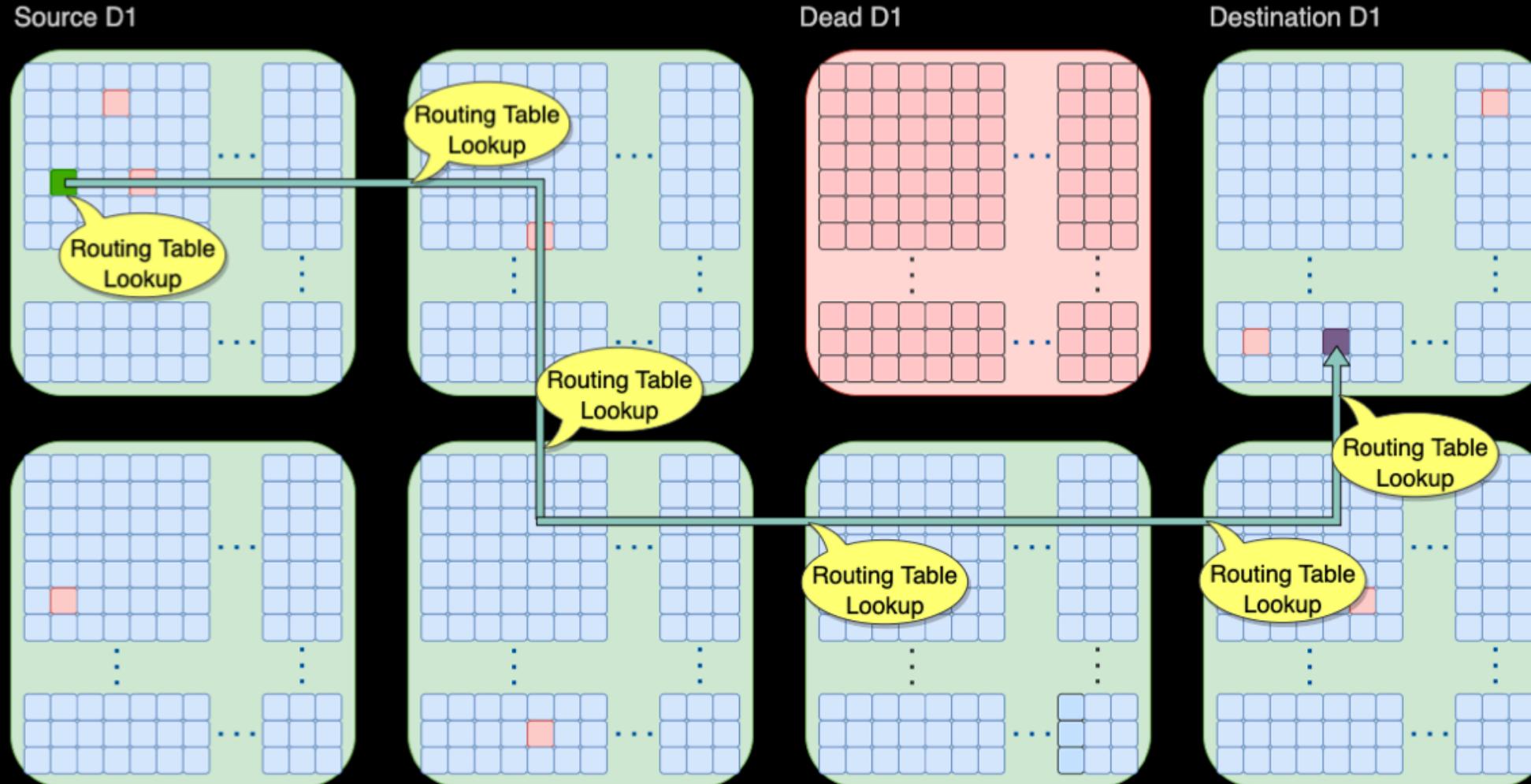
DOJO系统网络通信方式

- 不同的Dojo Core 通过片上网络路由 (NoC Route) 进行连接和数据同步，而不是共享数据缓存。
- NoC 可以处理跨节点边界 4 个方向 8 个数据包，每个方向 64 B/cycl
e. 该 NoC Route 还可以在每个周期对 Core 中 SRAM 进行一次 64 B 双向读写。



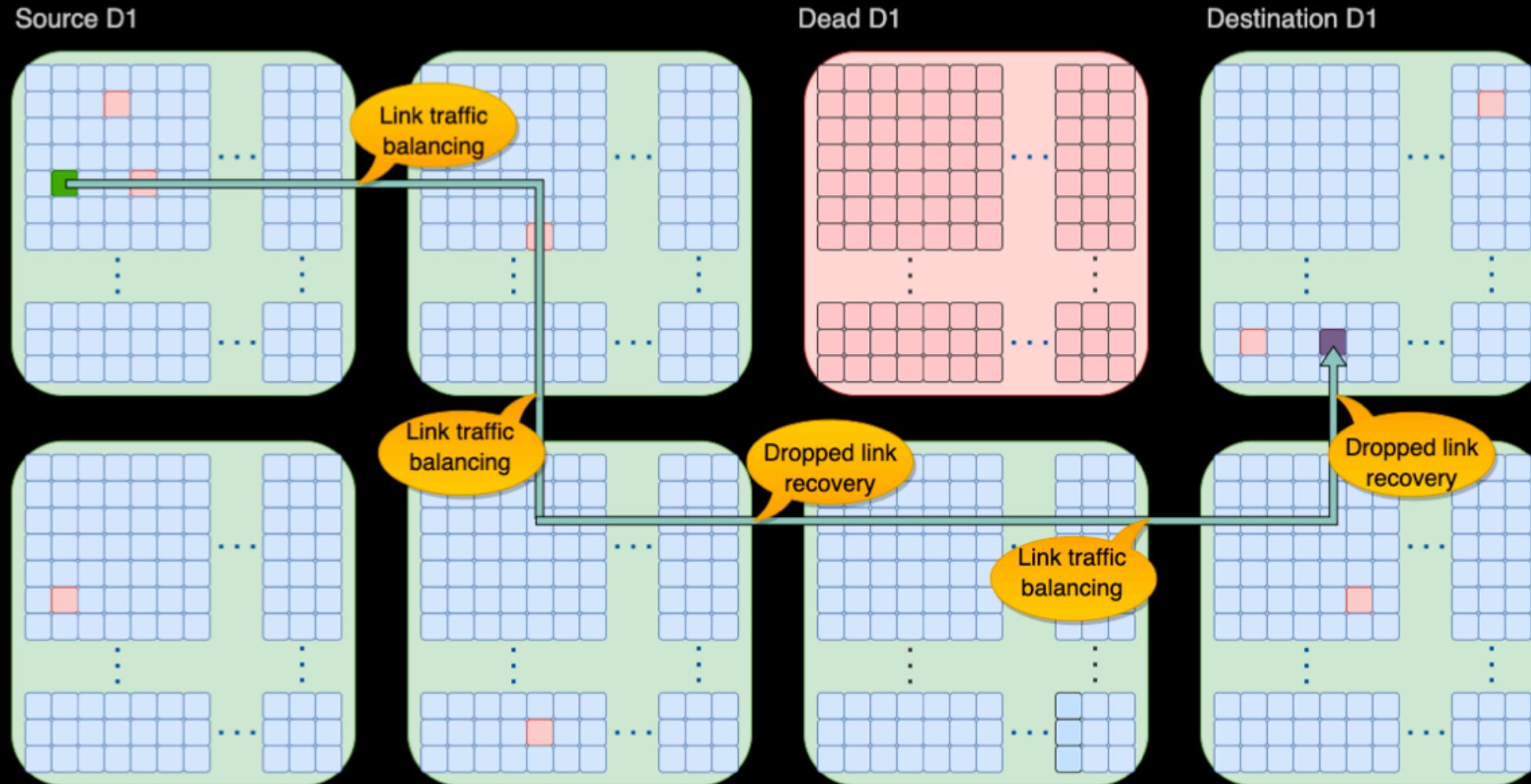
DOJO系统网络通信方式

- Dojo不保证端到端的流量排序，因此需要在目的地对数据包进行计数。



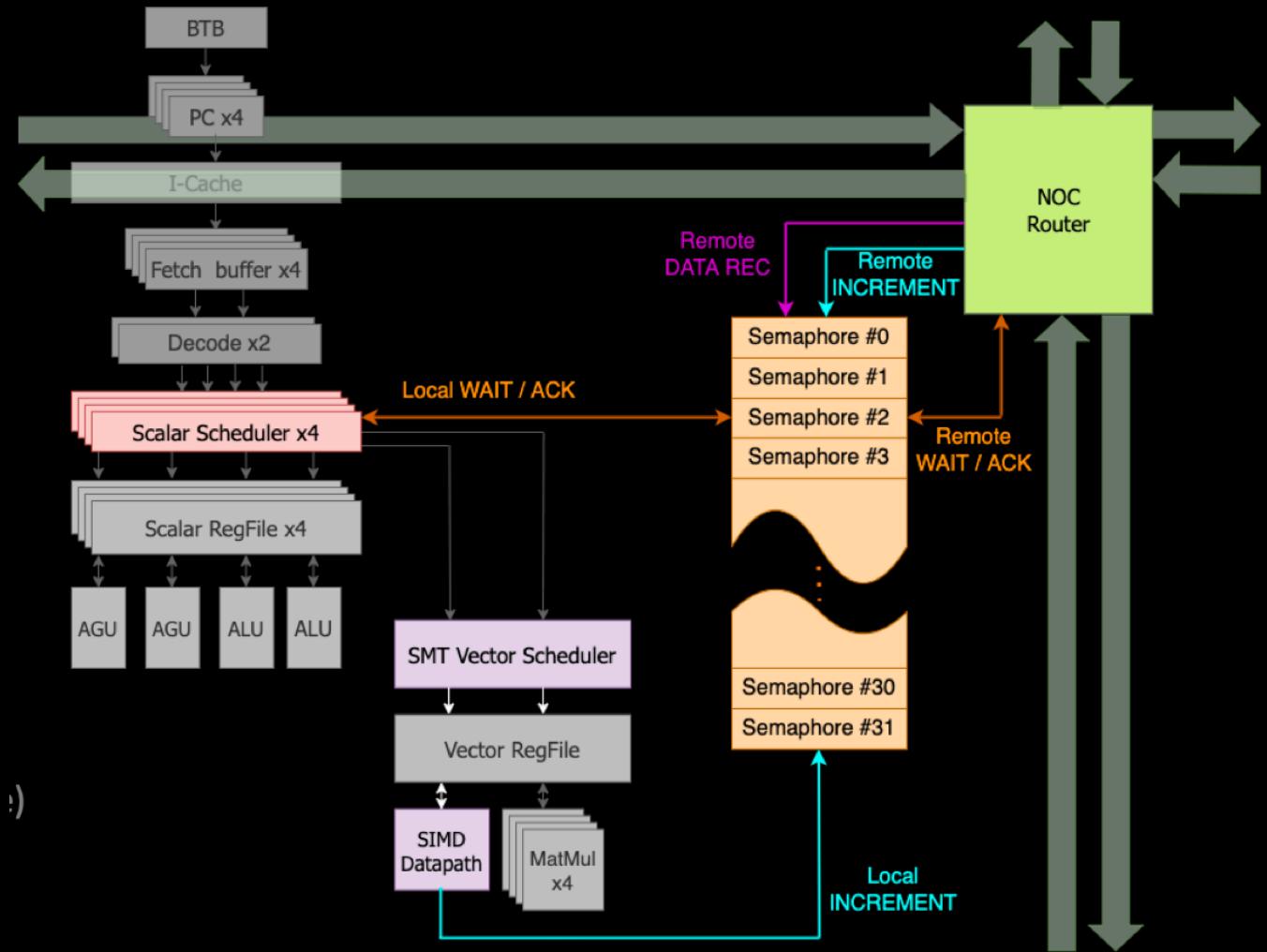
DOJO系统网络通信方式

- Dojo不保证端到端的流量排序，因此需要在目的地对数据包进行计数。



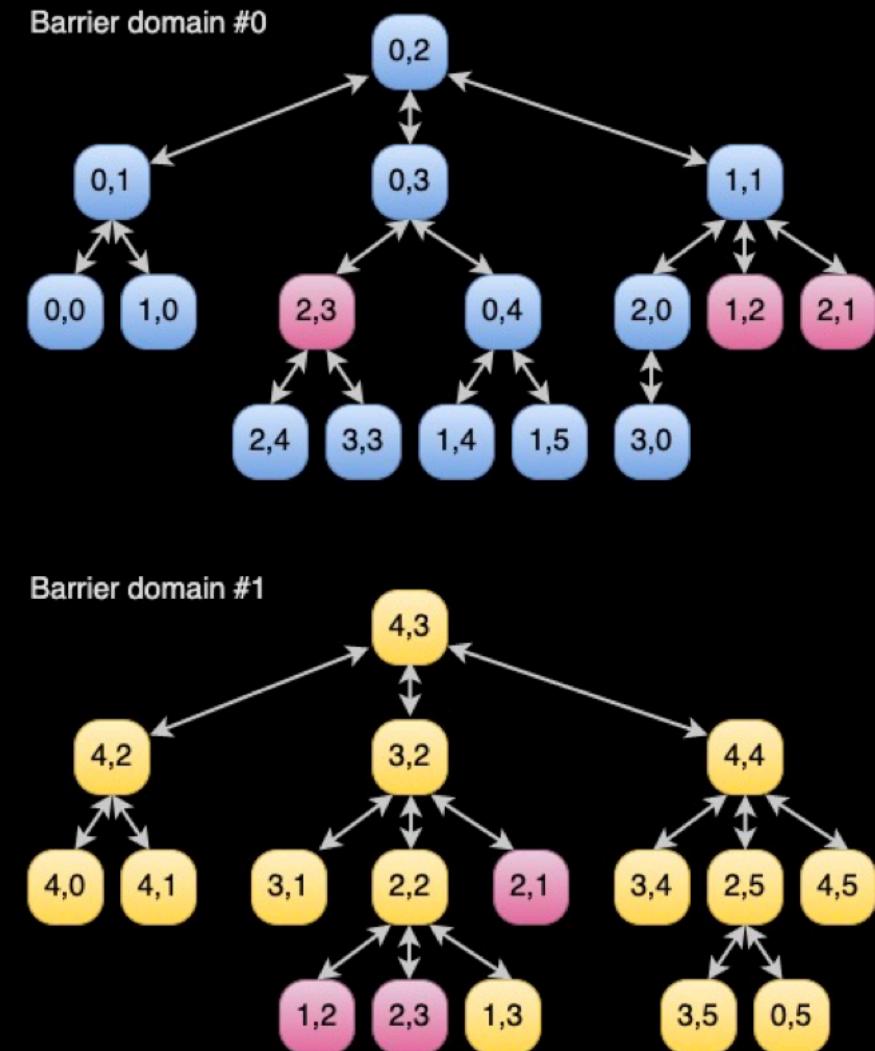
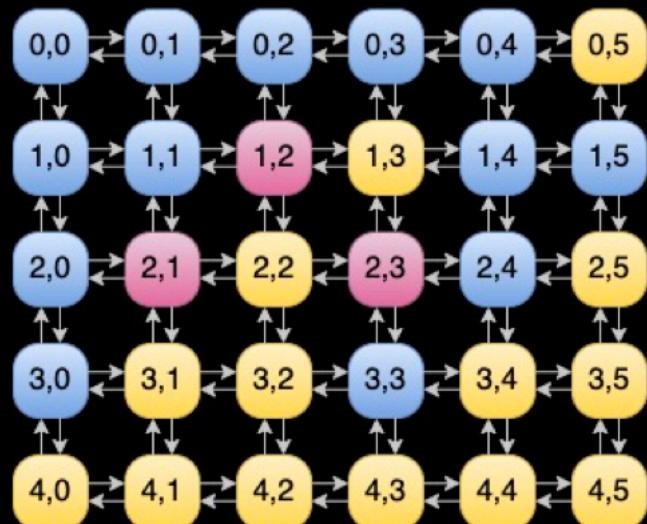
通信方式

- **任务生成**：从远程获取数据包，然后分为远程 Remote 线程执行和本地 Local 线程执行两种任务执行方式。
- **任务监控**：等待线程进行休眠状态，直到远程 Remote 和本地 Local 线程触发。



DOJO 系统同步

- 编译器定义一个带节点的通信树，对计算区域平面进行划分。通过软件层面去检查和触达所有树节点，对系统数据进行同步。



软件生态



BUILDING A BETTER CONNECTED WORLD

Ascend – ZOMI

26

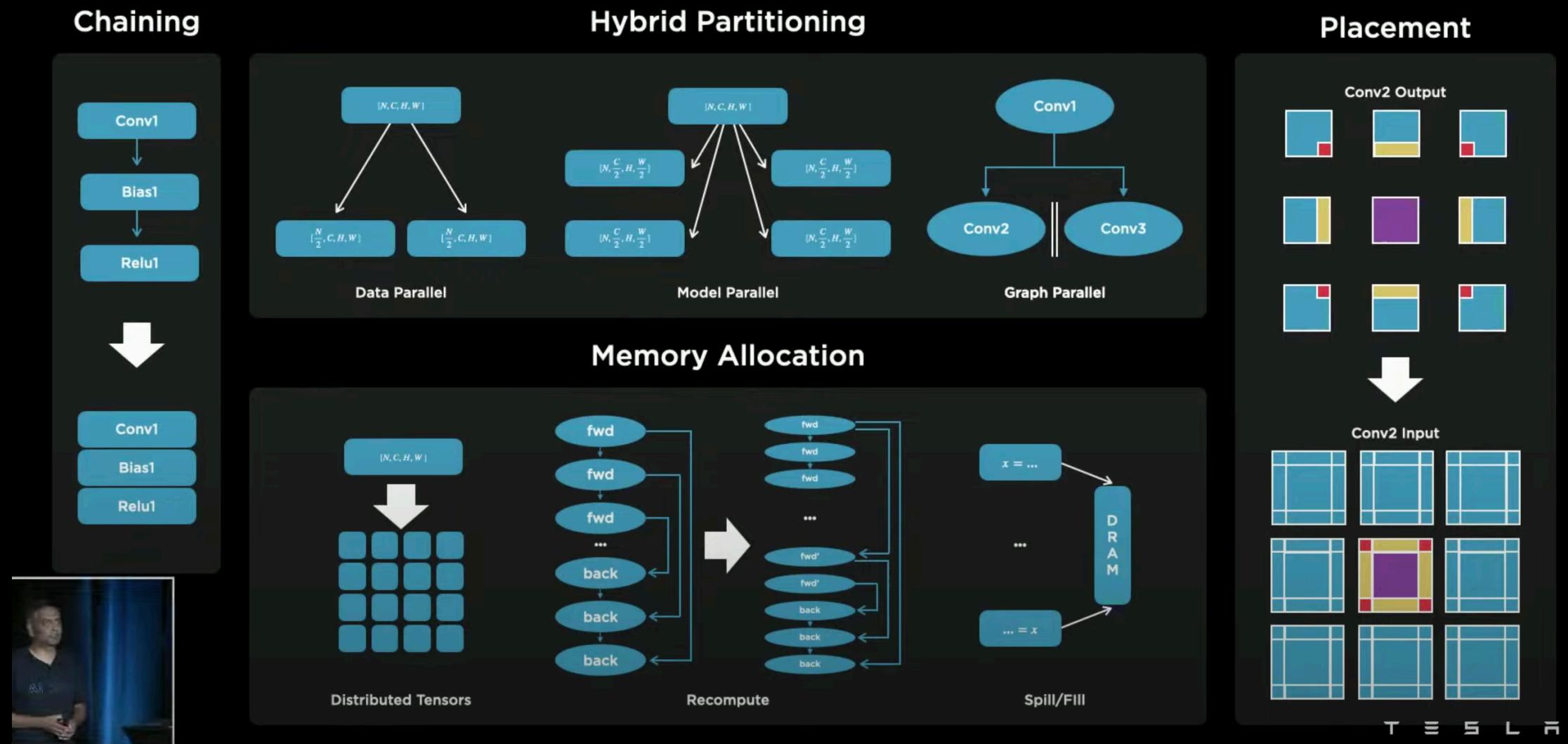


www.hiascend.com

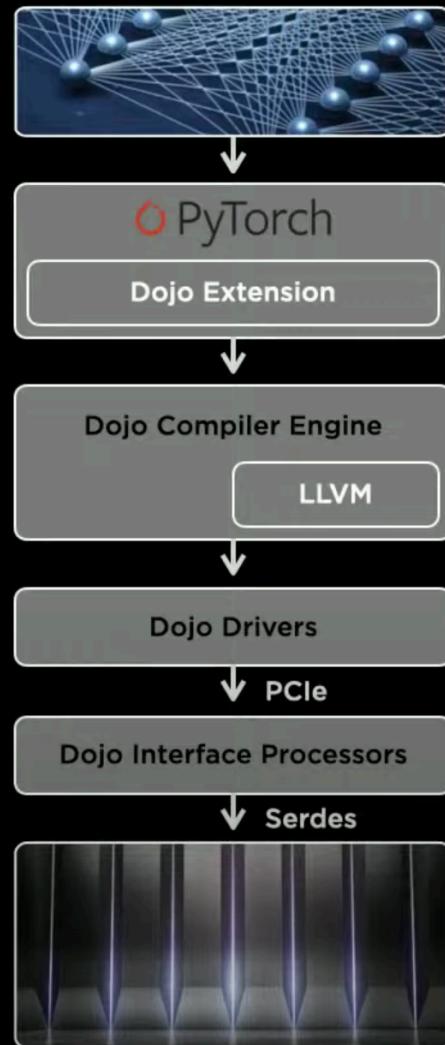
分布式系统



后端编译引擎



软件栈对接PyTorch



Neural Net Models

PyTorch-Extension

JIT NN Compiler

LLVM Backend

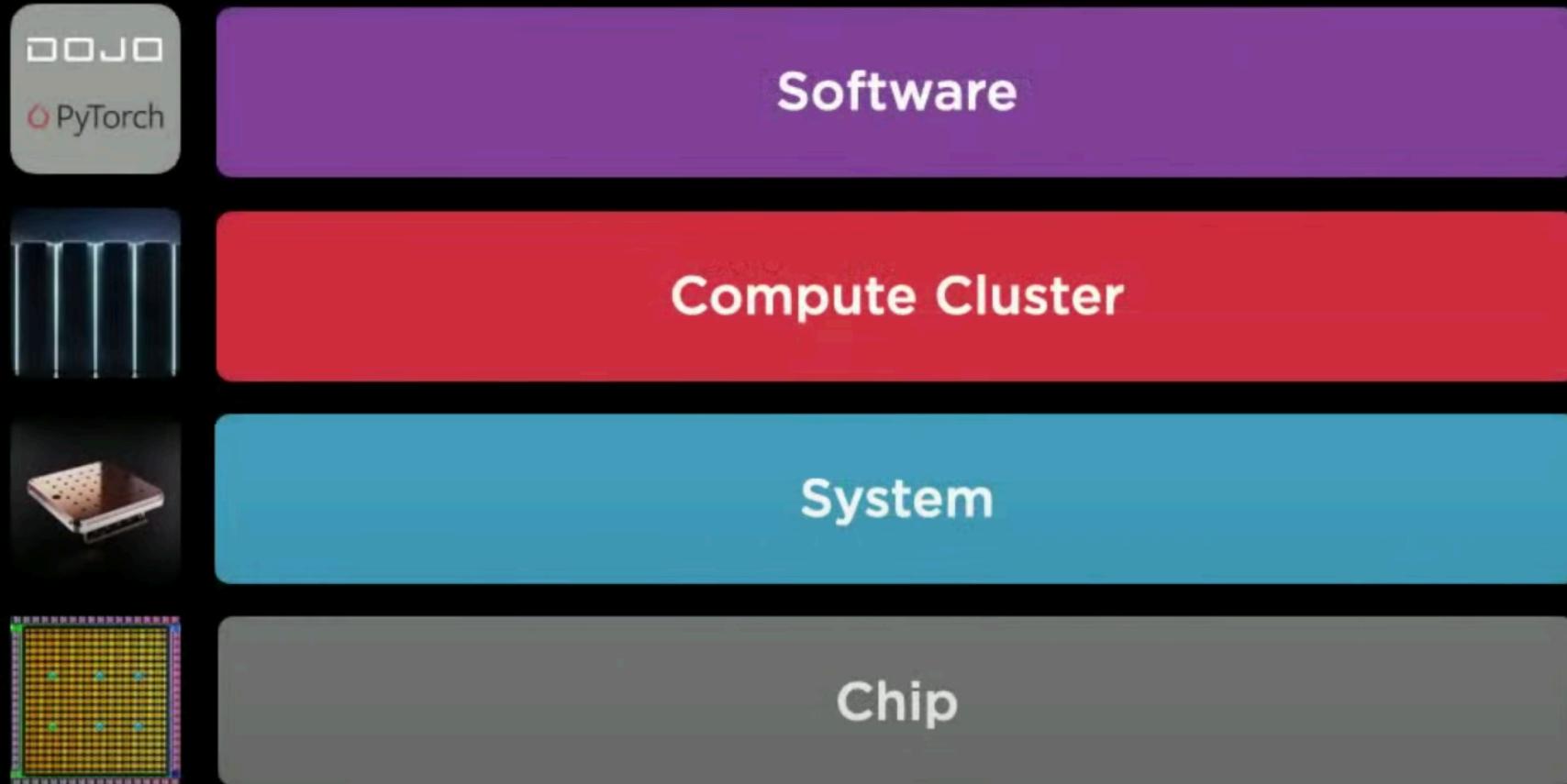
Multi-Host, Multi Partition Management

Ingest & Shared Mem

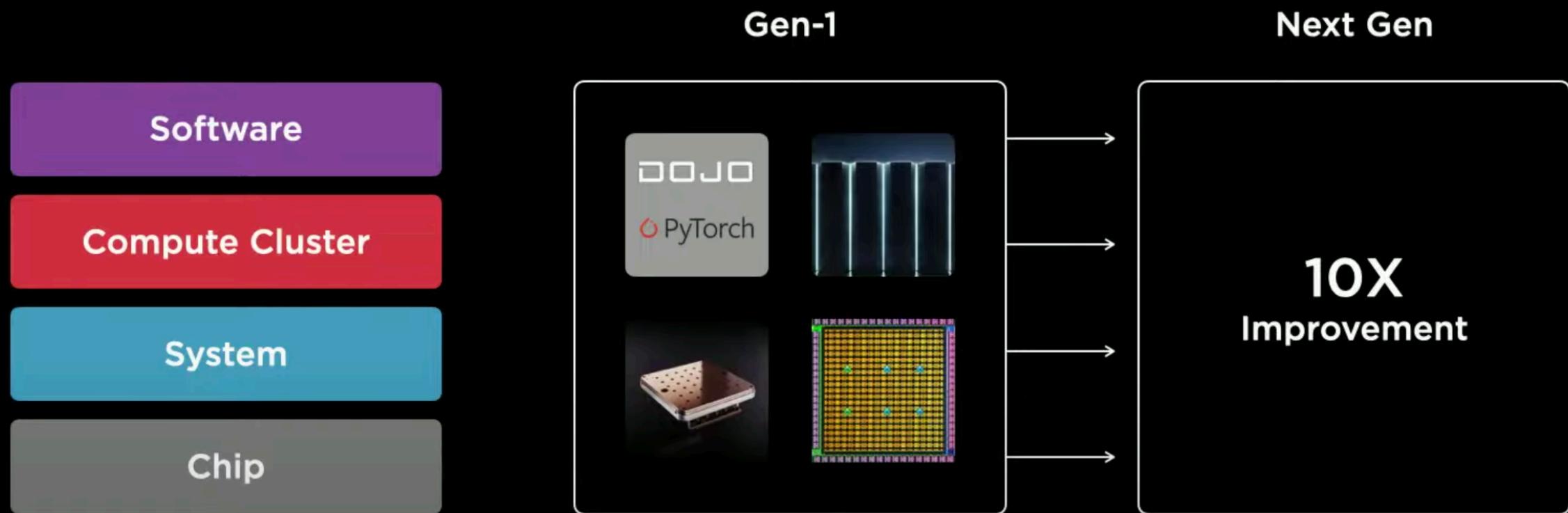
ExaPOD

T E S L A LIVE

各层对应的软硬件栈



RoadMap



问题与思考



Question?

1. 业界关于AI编译器没有形成统一的方案，处于快速演进期；但是 DOJO 需要的编译器跟业界 AI编译器差异大，如何兼容软件易用性？
2. Tesla 不需要 DOJO 内核来同时处理多个正在运行的应用程序，DOJO 只需运行内部的可信代码，并不关心虚拟内存支持。在多任务场景如何合理分配任务？
3. DOJO 上运行机器学习程序将在编写时考虑到该特定系统，为了提供多的芯片面积给计算，去掉了precise exceptions 等模块，在训练场景如何有效调试代码？



Question?

1. 不能同时运行多个应用程序。单个行为不端的应用程序可能会迫使您重新启动系统。如果您没有足够的 RAM 来运行某个程序，您可以完全忘记运行它（无需分页到磁盘）
2. DOJO做了很多硬件的定制化和阉割方案，Extra POD 作为服务器集群，如何考虑云服务，提供给不同团队使用？虚拟化和 Device 业务隔离如何考虑？



Reference 引用&参考

1. <https://www.youtube.com/watch?v=uE2f7kiRhmw>
2. <https://www.youtube.com/watch?v=QurtwJdb5Ew>
3. <https://www.youtube.com/watch?v=DSw3lwsgNnc>
4. <https://chipsandcheese.com/2022/09/01/hot-chips-34-teslas-dojo-microarchitecture/>
5. https://en.wikipedia.org/wiki/Tesla_Dojo
6. <https://www.qbitai.com/2022/08/37209.html>
7. <https://zhidx.com/p/347884.html>
8. <https://www.cnblogs.com/wujianming-110117/p/17115152.html>
9. <https://mp.weixin.qq.com/s/xGytSXTW7-CL-O1V7y7QRw>
10. <https://mp.weixin.qq.com/s/uBL4x4MJz1GiCf2a0gnnUQ>
11. https://mp.weixin.qq.com/s/N_sMmndpyiq0_qbdwq3MuA
12. <https://www.51cto.com/article/717372.html>
13. <https://mp.weixin.qq.com/s/vmNsRVwmI3Azo-bPDanc2g>



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.