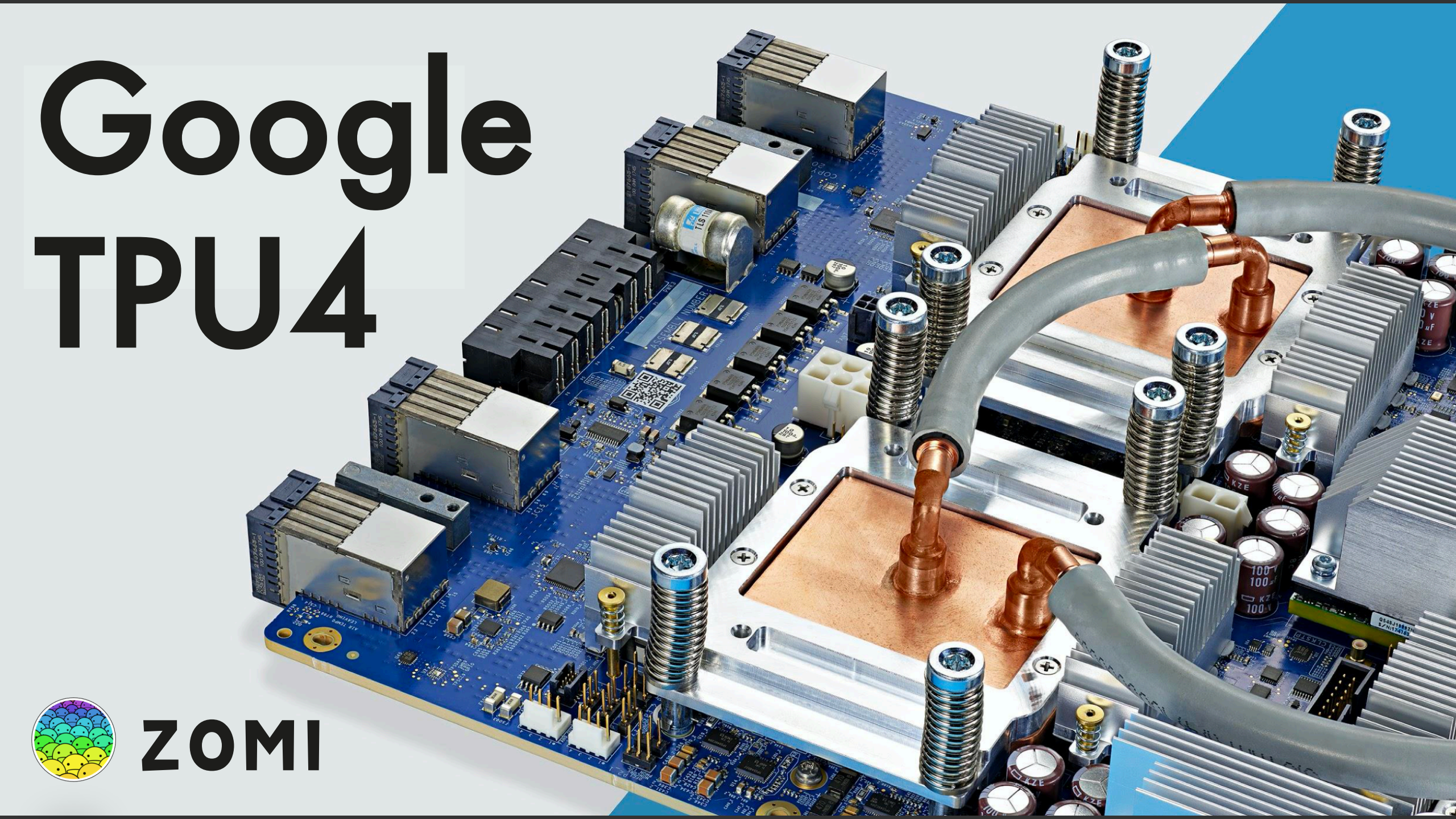
Google TPU4



ZOMI

Talk Overview

1. **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
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
3. **GPU详解**
   - 英伟达GPU架构发展
   - Tensor Core和NVLink
4. **国外 AI 芯片**
   - 特斯拉 DOJO 系列
   - 谷歌 TPU 系列
5. **国内 AI 芯片**
   - 壁仞科技芯片架构
   - 寒武纪科技芯片架构
6. **AI芯片的思考**
   - SIMD&SIMT与编程体系
   - AI芯片的架构思路与思考

Talk Overview

I. 国外 AI 芯片

- 英伟达 GPU 芯片架构剖析
- 特斯拉 DOJO 芯片架构剖析
- 谷歌 TPU 芯片架构剖析

Talk Overview

I. 国外 AI 芯片

- 英伟达 GPU 芯片架构剖析
- 特斯拉 DOJO 芯片架构剖析
- 谷歌 TPU 芯片架构剖析
  - TPU 历史发展
  - TPU1 脉动阵列细节
  - TPU2 第一款训练卡
  - TPU3 性能 POD 超算
  - TPU4 超级互联

TPU历代芯片

	TPUv1	TPUv2	TPUv3	Edge v1	Pixel Neural Core	TPUv4i	TPUv4	Google Tensor
Date introduced	2016	2017	2018	2018	2019	2020	2021	2021
Process node	28 nm	16 nm	16 nm			7nm	7 nm	
Die size (mm^2)	330mm	625mm	700mm			400mm	780mm	
On-chip memory (MB)	28MB	32MB	32MB			144MB	288MB	
Clock speed (MHz)	700MHz	700MHz	940MHz			1050MHz	1050MHz	
Memory	8 GB DDR3	16 GB HBM	32 GiB HBM			8GB DDR	32 GB HBM	
Memory bandwidth	300 GB/s	700 GB/s	900 GB/s			300GB/s	1200 GB/s	
TDP (W)	75	280	450			175	300	
TOPS (Tera/Second)		45	123	4			275	
TOPS/W	0.31	0.16	0.56	2			1.62	

After 4 Years, TPU Be back

- **看时间**：TPUv3 (2018) vs TPUv4(2022)，中间隔了 4 年
- **看竞品**：NV 发布了 Volta、Amber、Hopper一共3代架构，越来越 AI
- **看框架**：TensorFlow已经沉寂，PyTorch成为了AI框架的王者
- **看技术**：大模型涌现对大模型算力的需求和消耗

rack contains 16 tray
cables create a 4×4×4 3D mesh in a rack.
-optical conversions happen at the fiber connector to the
TPU trays. There are no other conversions until the

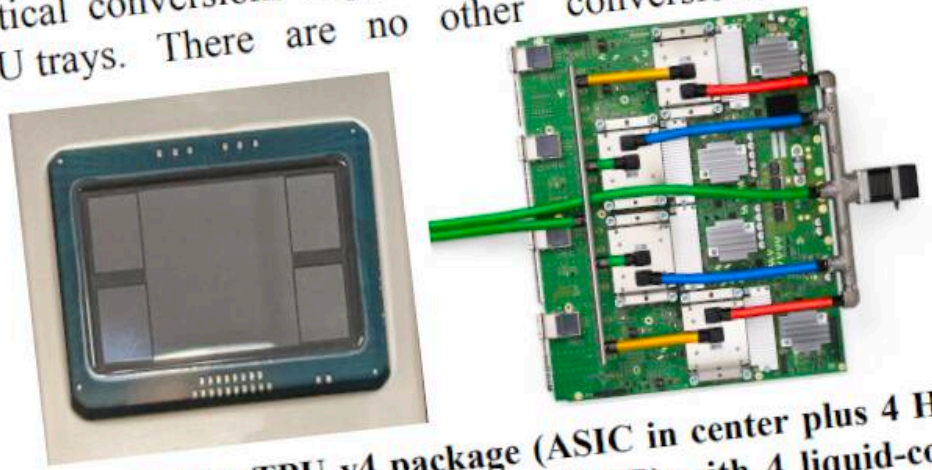


Figure 2: The TPU v4 package (ASIC in center plus 4 HBM stacks) and printed circuit board (PCB) with 4 liquid-cooled packages. The board's front panel has 4 top-side PCIe connectors and 16 bottom-side OSFP connectors for inter-tray ICI links.

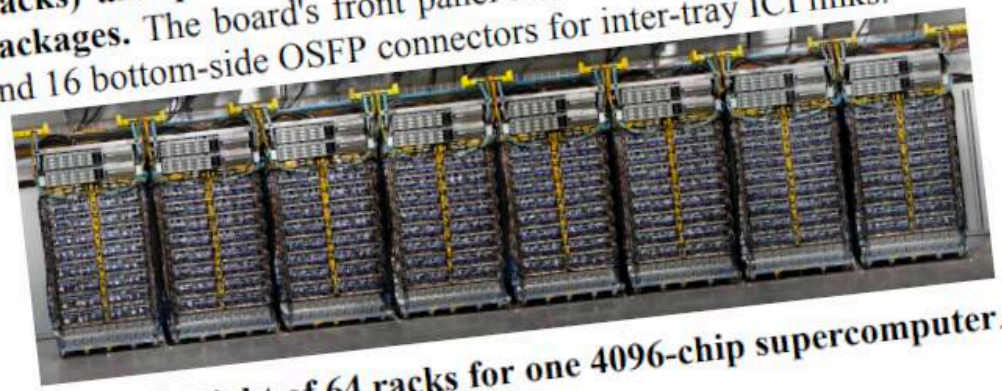


Figure 3: Eight of 64 racks for one 4096-chip supercomputer.

CPU Host Availability with/without OCS

99.9% w OCS

goodput. OCSes
99.5% for most slice sizes. Figure 4 assum
requests are equal, but workloads have many sizes (Table 2).

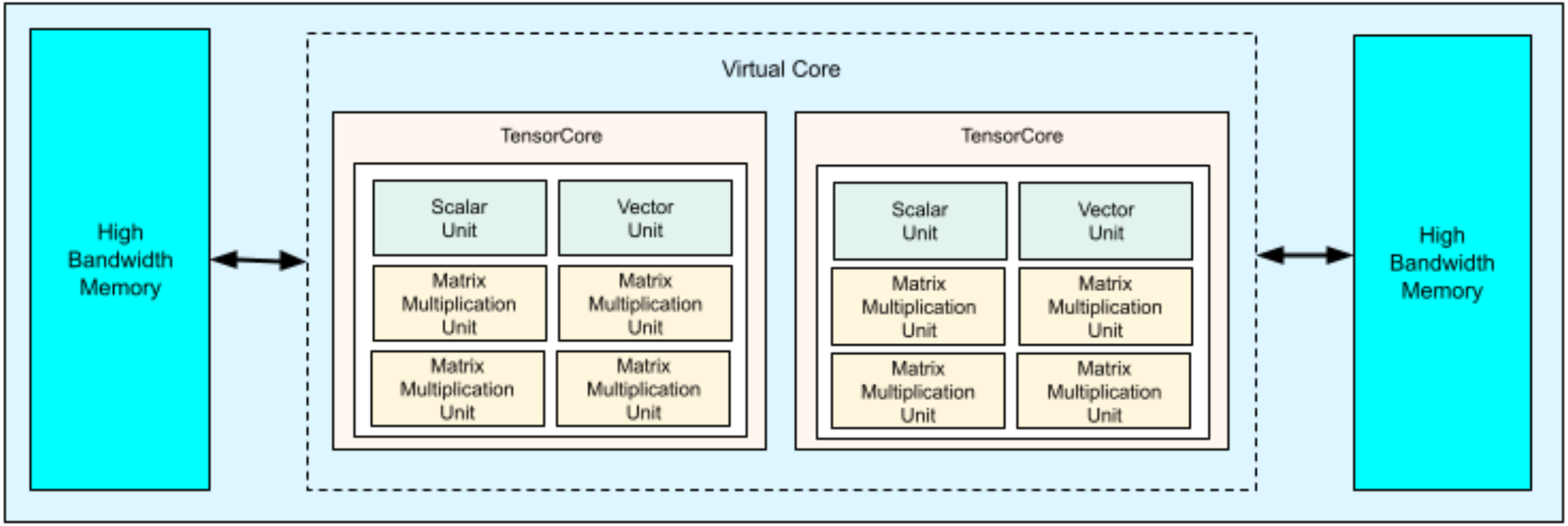
Table 2: Sampling of popularity of TPU v4 slices for a day in Nov. 2022. This table includes all slices used ≥ 0.1%. Twistable (Sec. 2.8) but not twisted means the slice geometry allows twisting (n×n×2n or n×2n×2n), but the user picks the regular topology. The software scheduler requires that slices have dimensions x ≤ y ≤ z. Half of the slices have x, y, and z as either 4 or 8.

Chips	<64		64	
Regular Tori	1×1×1 (1)	2.1%	4×4×4 (64)	13.9%
	1×1×2 (2)	0.4%		
	1×2×2 (4)	6.7%		
	2×2×2 (8)	4.7%		
	2×2×4 (16)	6.4%		
	2×4×4 (32)	8.9%		14%
Total %		29%		
Chips	128-192		256-384	
Twisted Tori	4×4×8_T (128)	16.0%	4×8×8_T (256)	9.2%
Twistable, not twisted Tori	4×4×8_NT (128)	1.5%	4×8×8_NT (256)	1.5%
			4×4×16 (256)	1.0%
Regular Tori	4×4×12n (192)	0.7%	4×8×12 (384)	0.1%
		18%		12%
			1024-1536	

1. TPU v4 介绍

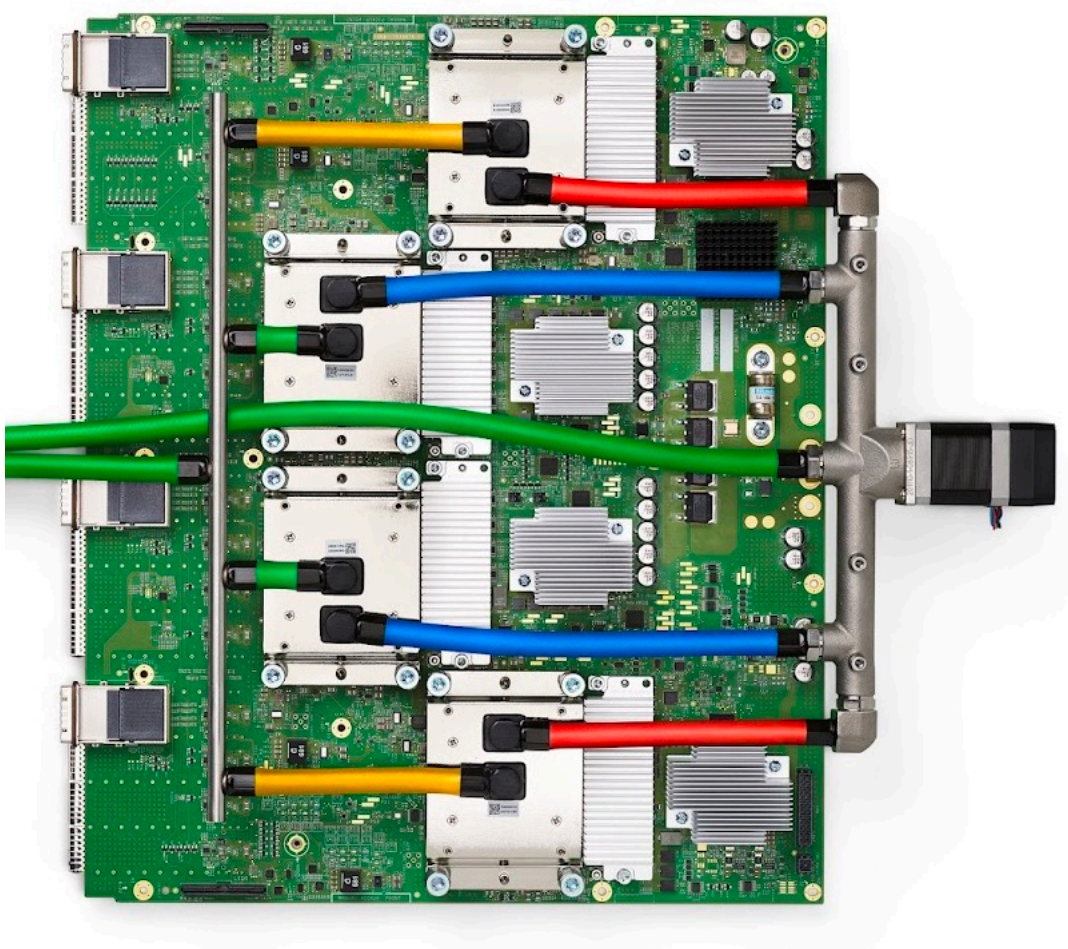
TPU v4 架构图

- 每个 v4 TPU 芯片包含两个 TensorCore。每个 TensorCore 都有四个 MXU、一个矢量单元和一个标量单位。



TPU v4 产品

- TPU v4 是 Google TPU 系列计算引擎的真正升级，工艺从 16 纳米缩小到 7 纳米。MXU 的数量翻了一番，缓存内存增加了 9 倍至 244 MB，HBM2 内存带宽增加了 33% 至 1.2 TB/s，可惜 HBM2 内存容量保持不变 32 GB。
- TPUv4 首次亮相的新 3D torus 互联方式，紧密耦合 4,096 个 TPUv4 引擎，TPU v4 POD 总计提供 1.126 exaflops 的 BF16 峰值算力。
- 稀疏结构硬件专门通过 Sparse Core 支持，基于TPUv4改良的自有 Transformer 模型结构。



Sparse Core：Embedding 层理解

- **Embedding 处理离散型分类特征（Categorical Features），是稀疏化的典型计算范式。**NLP/搜推算法仅支持字符串形式输入（单词短语 Prompt/instruct），表示为离散的稀疏向量特征，稀疏特征不适合映射到硬件上的矩阵乘法单元进行Tensor计算，更像是哈希表。
- 深度学习中由于神经网络通常在稠密 Tensor上计算性能更优，因此会使用 Embeddings 将离散的稀疏分类特征转换成空间更小的稠密 Tensor，作为 NLP/搜推算法模型的第一层。

Sparse Core：Embedding 层理解 (搜推)

- 深度学习中由于神经网络通常在稠密 Tensor上计算性能更优，因此会使用 Embeddings 将离散的稀疏分类特征转换成空间更小的稠密 Tensor，作为 NLP/搜推算法模型的第一层。

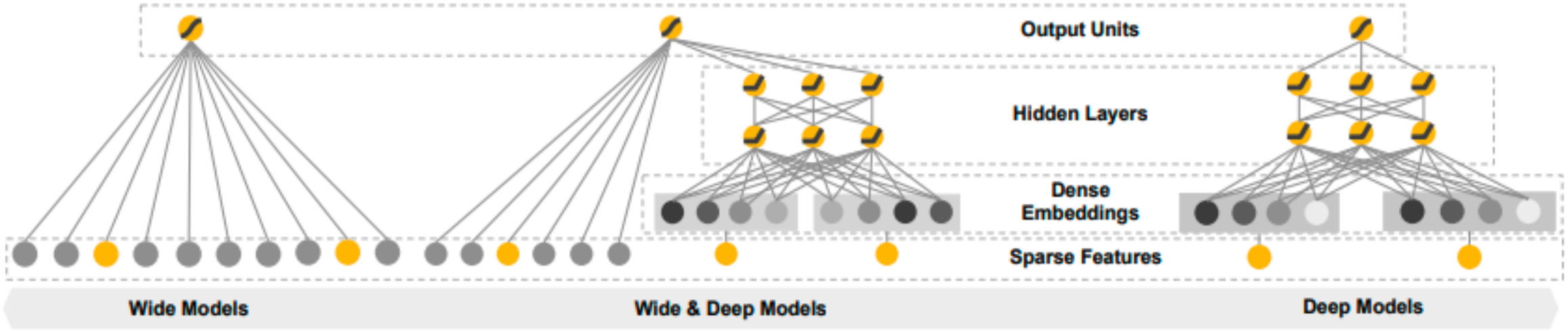
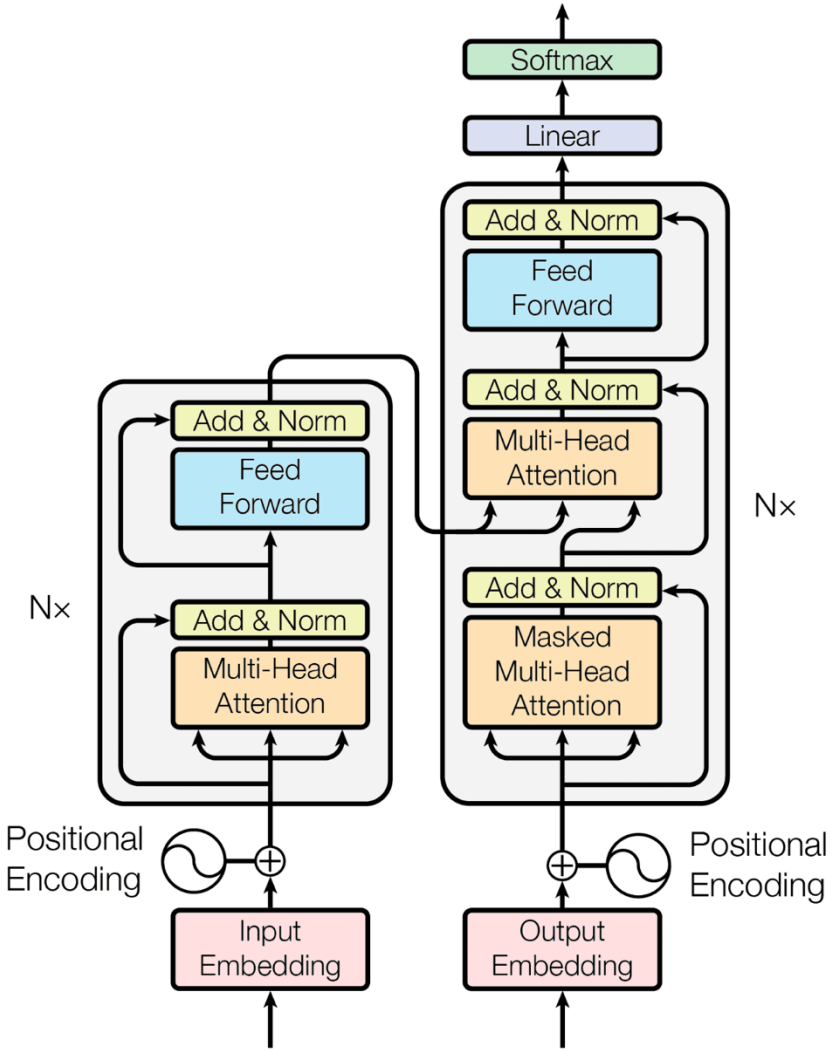
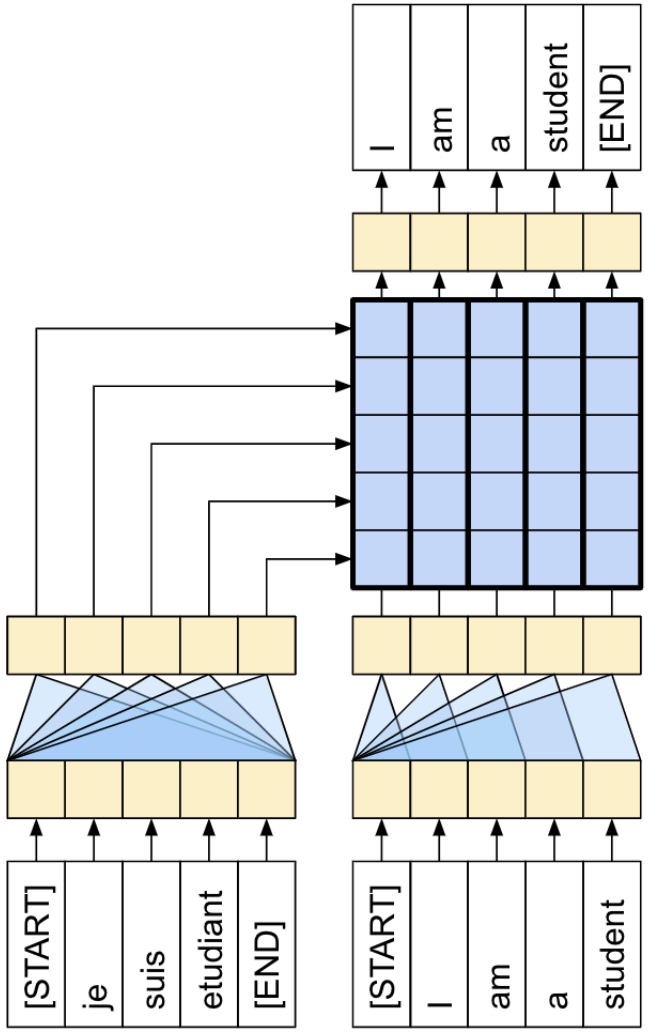


Figure 1: The spectrum of Wide & Deep models.

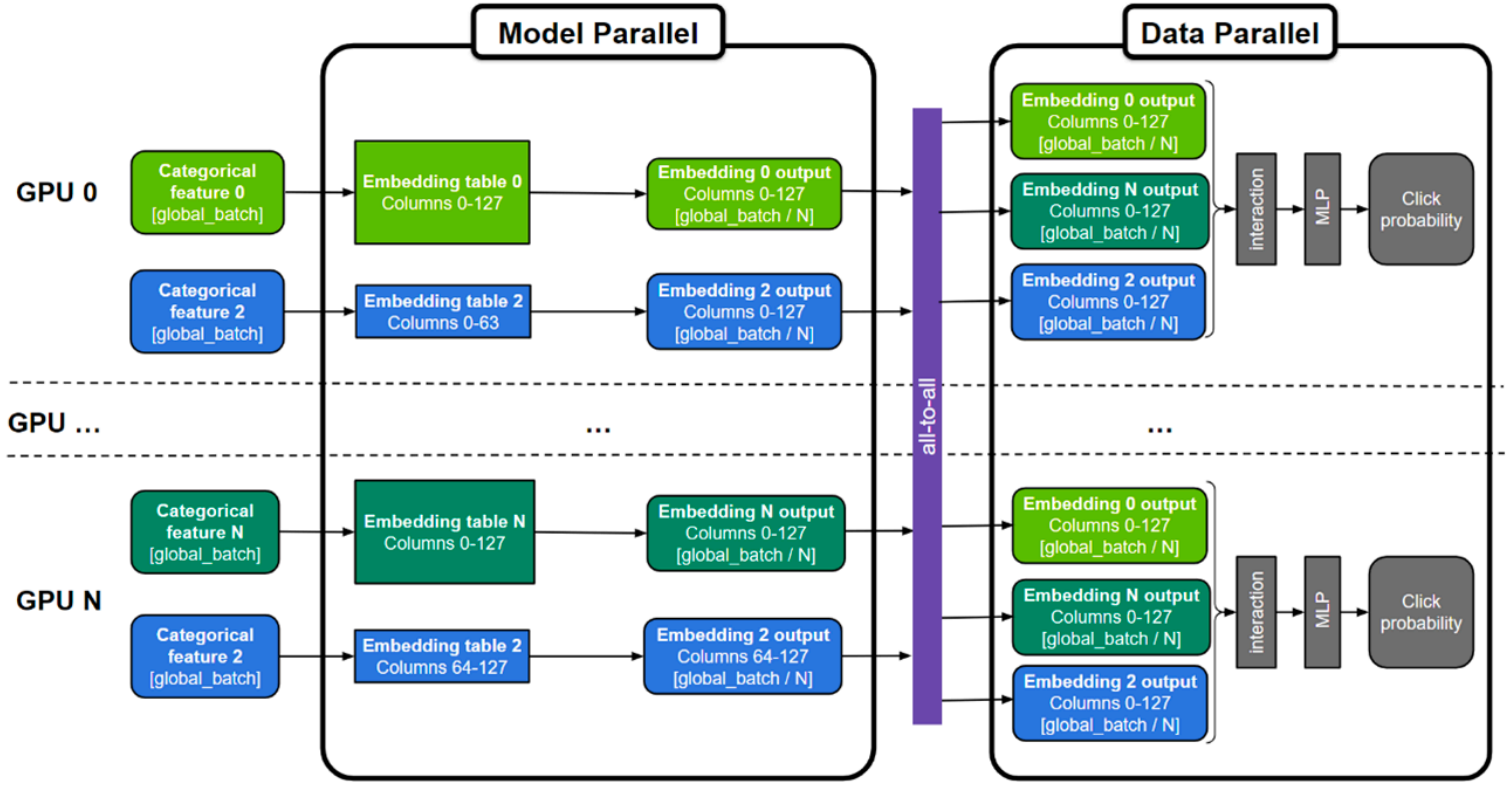
Sparse Core：Embedding 层理解 (NLP)





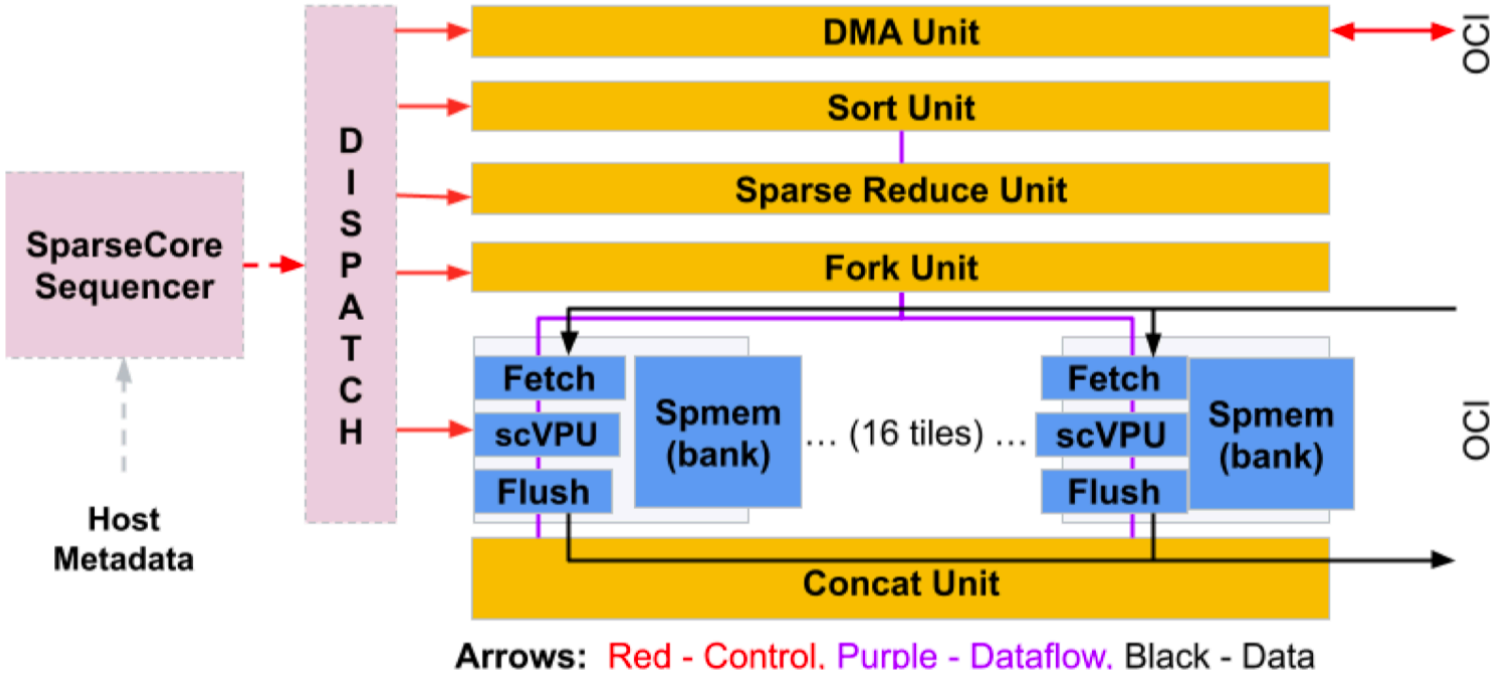
Sparse Core：支持 Embedding 并行

- TPU v4 搭配了一组独立SC（Sparse Core）稀疏核，8bit SIMD，提供了极大的并行化灵活度。可以让超大的 Transformer embedding layer 分布展开在大集群中计算。



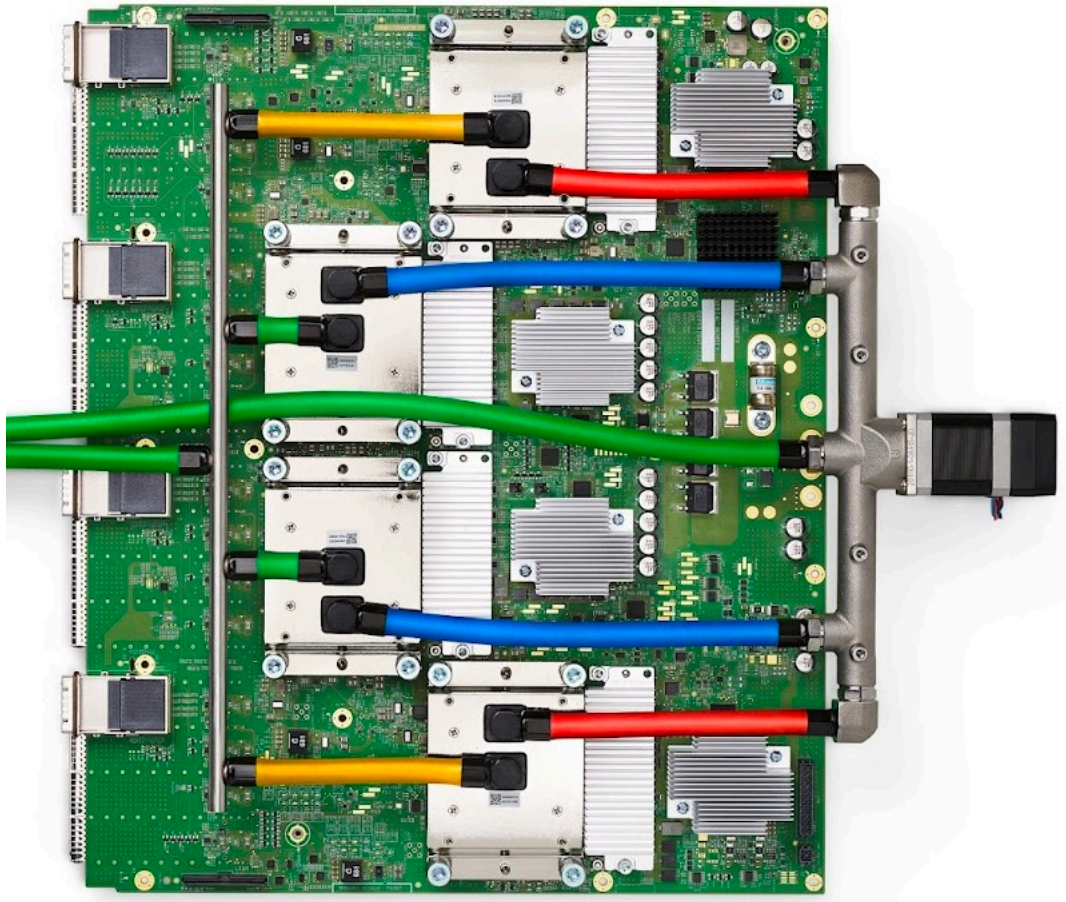
Sparse Core：核心架构

- SC 能够快速访问 HBM（类似于 GPUDirect），增加了独立 fetch, scVPU, flush 等处理单元，以便让数据高效传送到稀疏特征缓存2.5MB Spmem（Sparse Mem），搭配可编程 8bit SIMD 单元（scVPU）可以快速计算稀疏数据。每块 TPU v4 芯片有 4 个 SC 核心，每个 SC 构成16 个 tiles。额外还有一些支持DMA、Sort、SR、Fork等操作的处理单元。



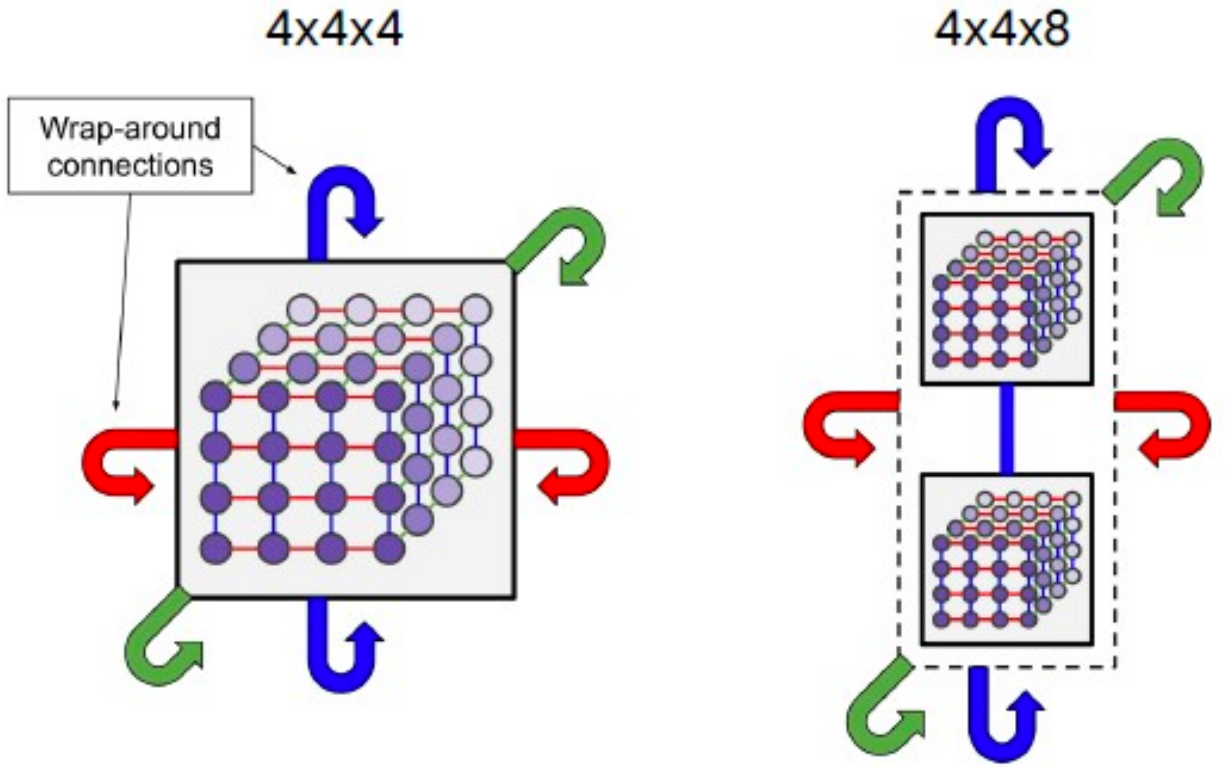
3D torus 互联方式

- 3D torus互联，它具有更多的带宽和更高的基数，它可以紧密耦合4,096个 TPUv4 引擎，总计 1.126 exaflops 的 BF16 计算。端口连接交换机 6 Tb/sec ，用作网络接口卡和 3D 环面网络的基础。



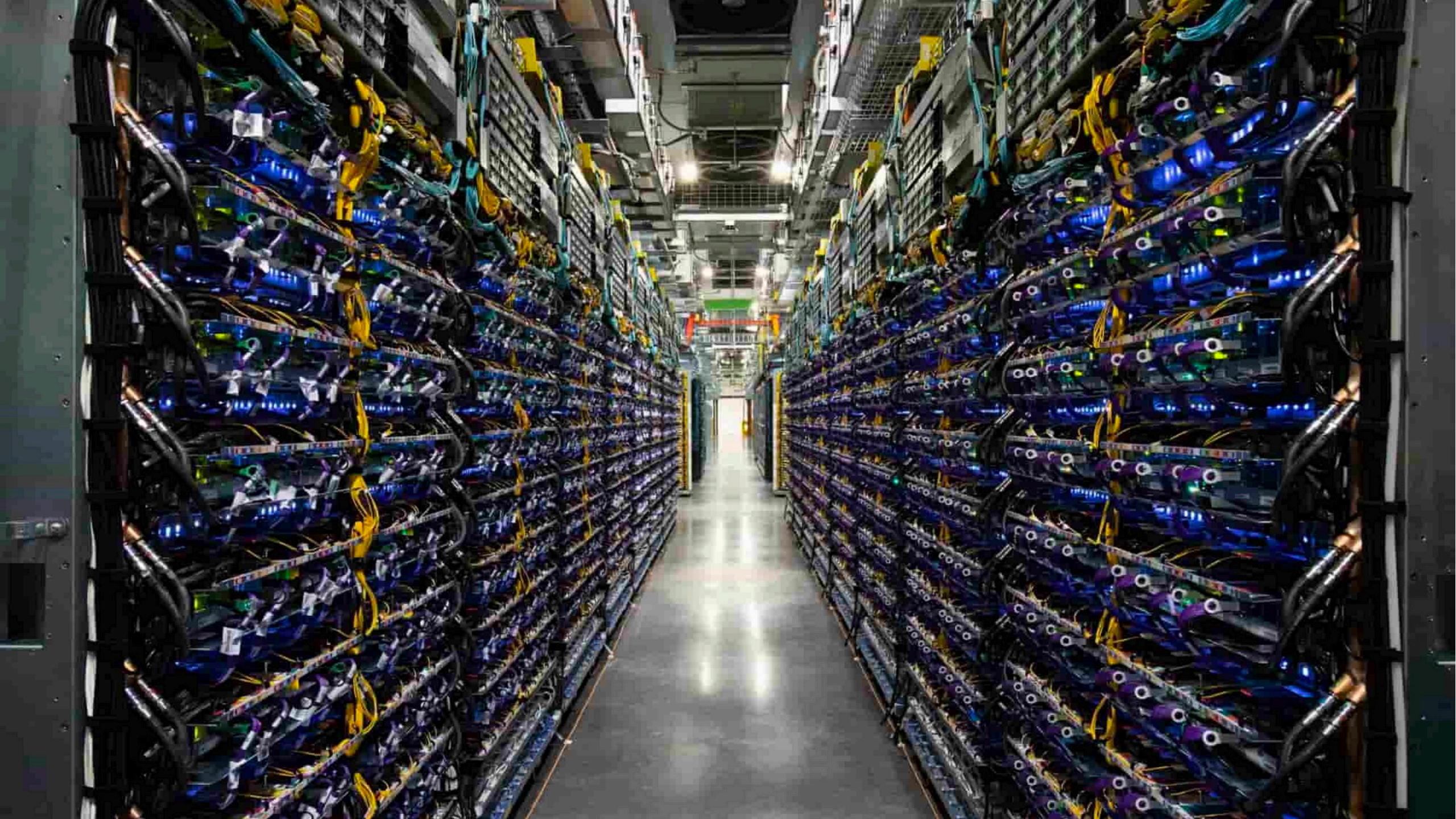
3D torus 互联方式

- Slice now with 3D
- Some Specific slices:
  - 2x2x1(v4-8， One TPU v4 VM)
  - 4x4x4(v4-128， 4-cube)
  - 4x4x8(v4-256)
  - 4x8x8(v4-512)
  - 8x8x8(v4-1024)
  - 8x8x16(v4-2048)
  - 8x16x16(v4-4096， half pod)



立方体切片大小是计算核心数，而不是芯片数

2. POD 形态

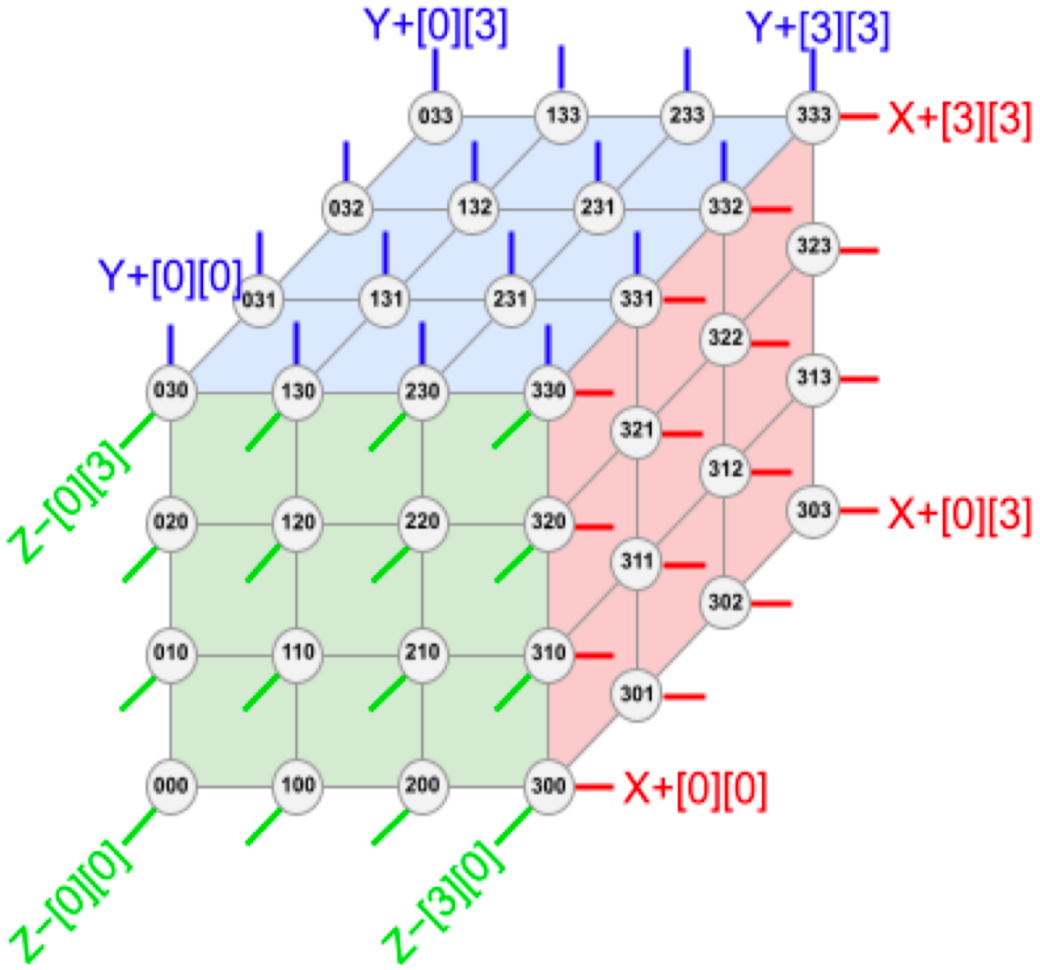


TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings

- 通过光互联可重配置（Optically Reconfigurable）机器学习超级计算机 POD，由 4096 个 TPU v4 单芯片组成的AI计算集群，可释放高达1exaflop（每秒10的18次方浮点运算）的算力，超过了目前全球运算速度最快的超级计算机（富岳）。
- **关键要素**：通过光电路交换机（Optical Circuit Switching，OCS），POD可以动态重新配置芯片之间的连接，避免出现问题并实时调整以提高性能。
- **成功案例**：5400亿参数 PaLM / PaLM2 使用两个 TPUv4 POD 训练 64 天。

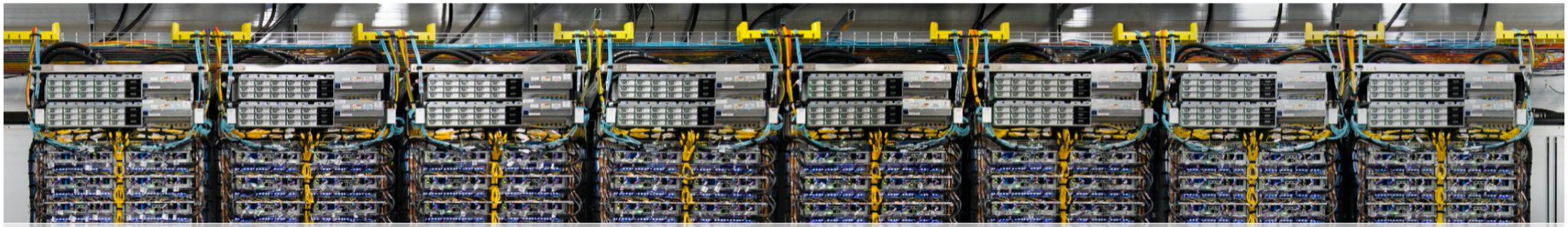
TPU v4 POD 拓扑结构：基本组成

- **组成方式**：将4x4x4（64）个 TPU v4 芯片互联在一起，形成一个立方体结构（Cube）。再把 4x4x4 Cube 用 OCS 连在一起形成一个总共有 4096 个 TPU v4 超级计算机。
- **拓扑结构**：每节点连接到网格中的六个相邻节点（上-下-左-右-前-后），在X-Y-Z三个维度中形成一个闭环。高度互联结构，节点会在三个维度上形成一个连续循环。因此称为 3D Tours。



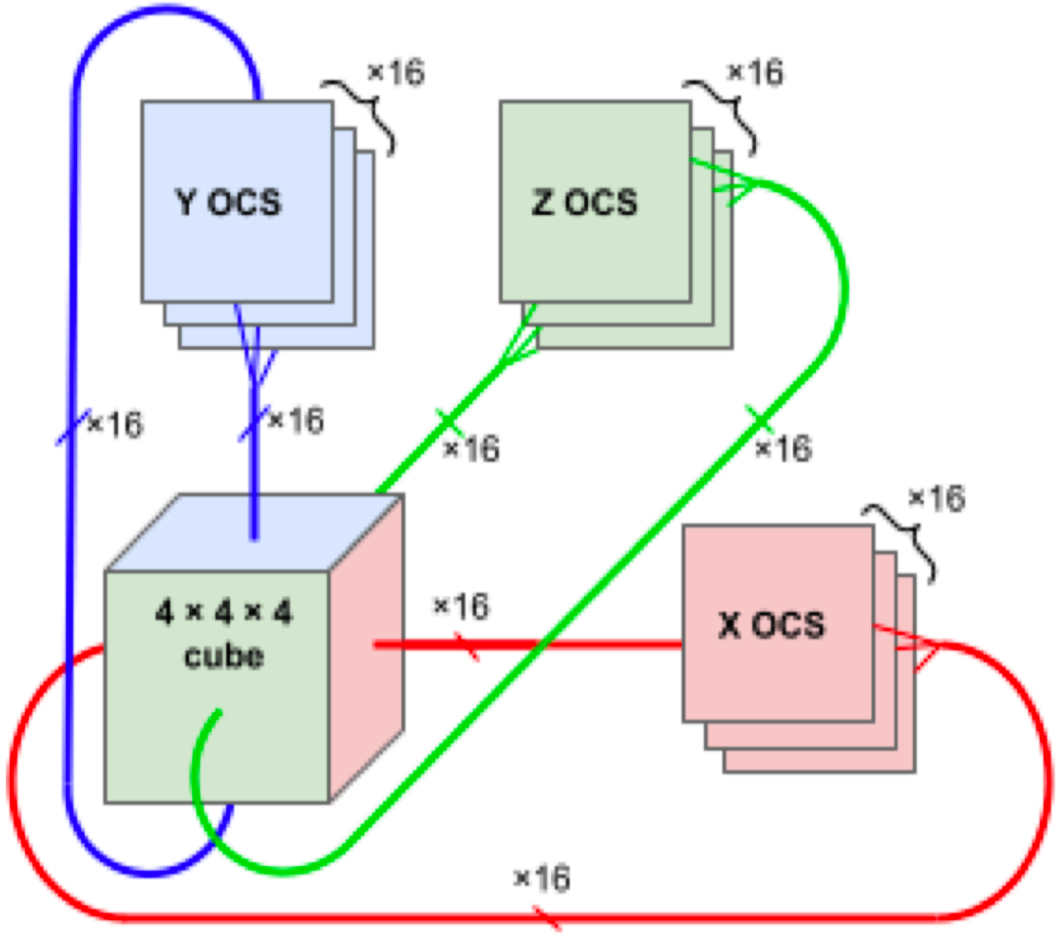
TPU v4 POD 拓扑结构：迎来光互联

- 物理距离较近 Cube 内可以用常规的电互联（ICI）方法连接，距离较远 Cube 间就必须使用光互连。原因就在于在如此大规模的超级计算机中，芯片间互联很大程度上决定整体计算效率；如果数据互联效率不够高的话，很多时候芯片都在等待来自其他芯片数据到达以开始计算。
- 为了避免计算等通信，必须确保芯片之间互联高带宽，低延迟。而光互连对于物理距离较远的芯片就成为了首选。OCS 由 64 颗 TPU 构成一组 Slice 间互连，实现了 Pod 内 Slice 间全光互连（4096 TPUs）；当然也可用于Pod之间互连。



TPU v4 POD 拓扑结构：基本组成

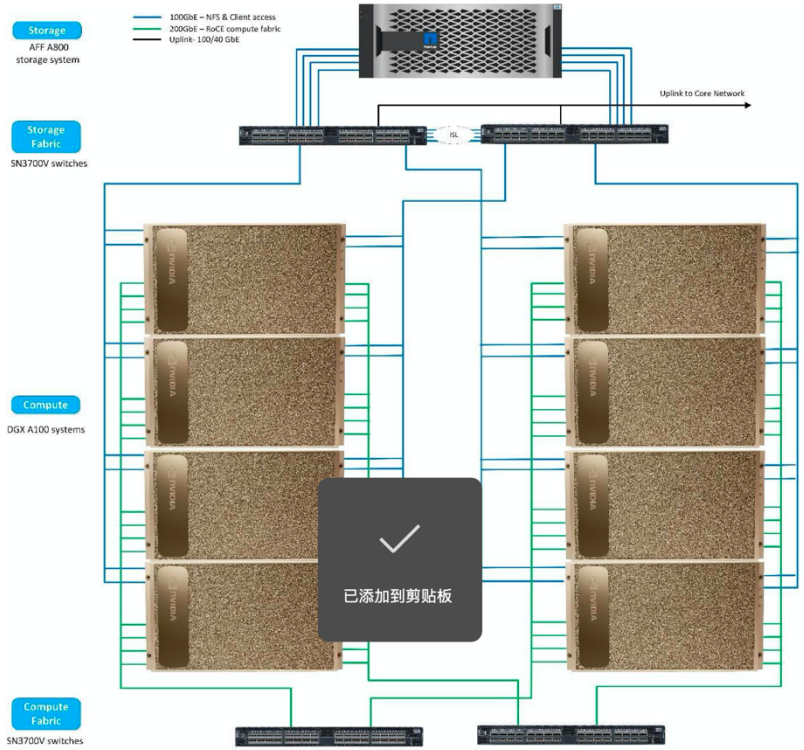
- Cube 要实现 6 面连接，每个面需要 16 条链路，每个块总共有 96 条光链路连接到 OCS 上。
- 要提供 3D 环面链接，相对侧的连接必须连接到相同的 OCS。因此，每个 Cube 连接到 6 × 16 ÷ 2 = 48 个 OCS 上。
- 48 个 OCS 连接来自 64 个 Cube 的 48 对光缆，总共并联 4096 个 TPU v4 芯片。



比的，就是钞能力

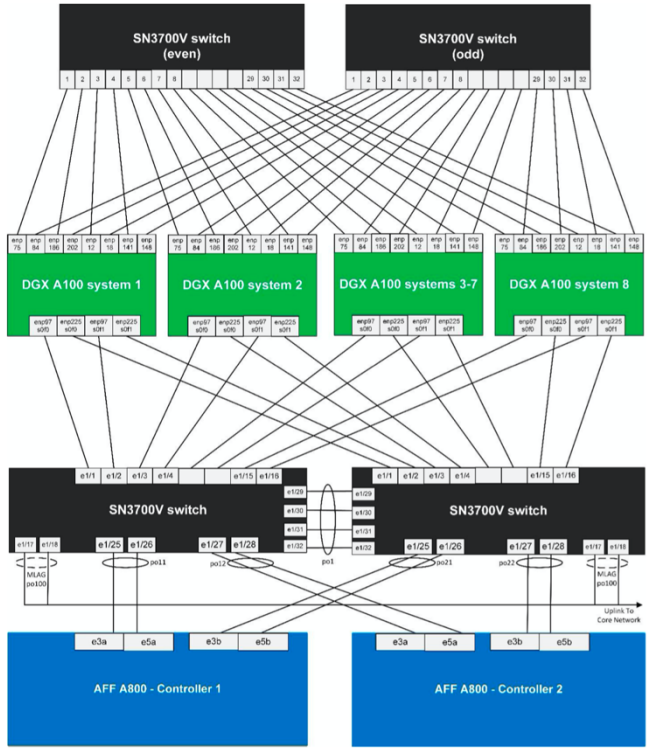
对标NV计算集群

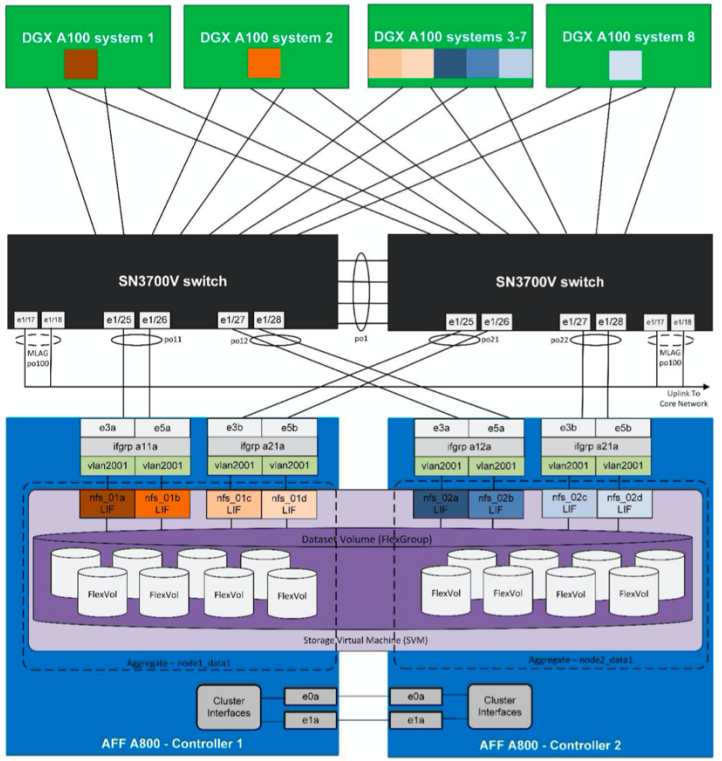
- DGX SuperPod 搭配第四代 NVLink/NVSwitch 最多可以连接 32 个 node 总 256 颗 H100 芯片，并实现每颗 GPU 900G/s的互连带宽。NV 每机架 4台 DGX（共 32 颗 H100 GPU），机架内/外需要光纤连接。**NV的每机架算力密度相对更小/更窄，且需要更多的收发激光器和光纤线材，网络成本高。**



对标NV计算集群

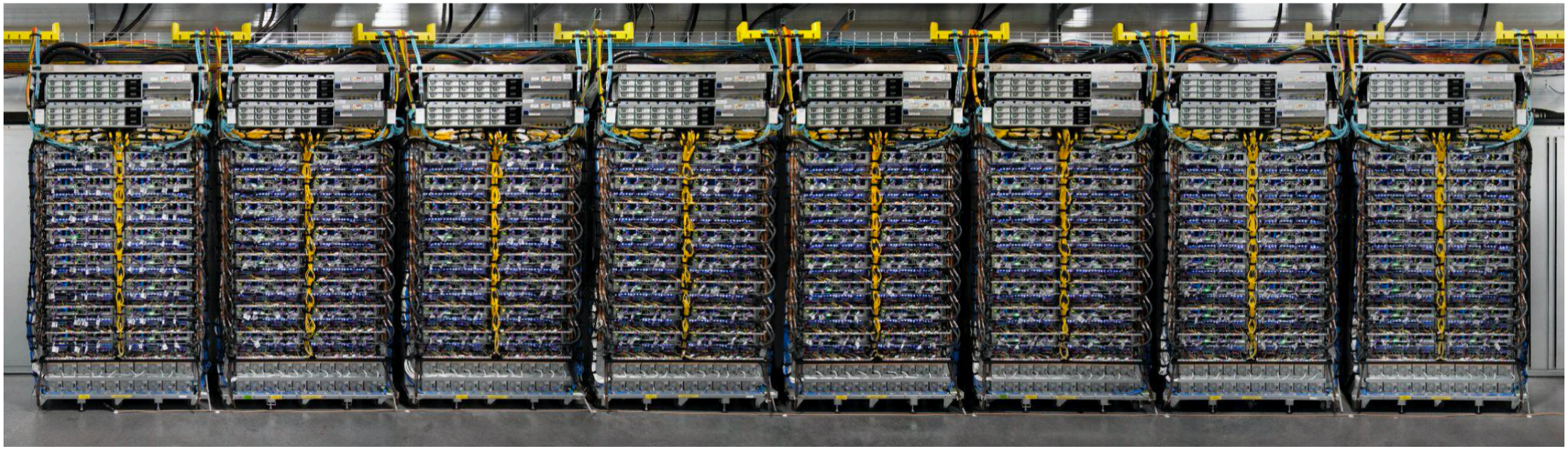
- NV 部署4096颗GPU集群，必须切分成更多个 SuperPod 并独立规划互连网络层，中间完成多层交换，集群内总计需要采购大约568个 Infiniband Switch。





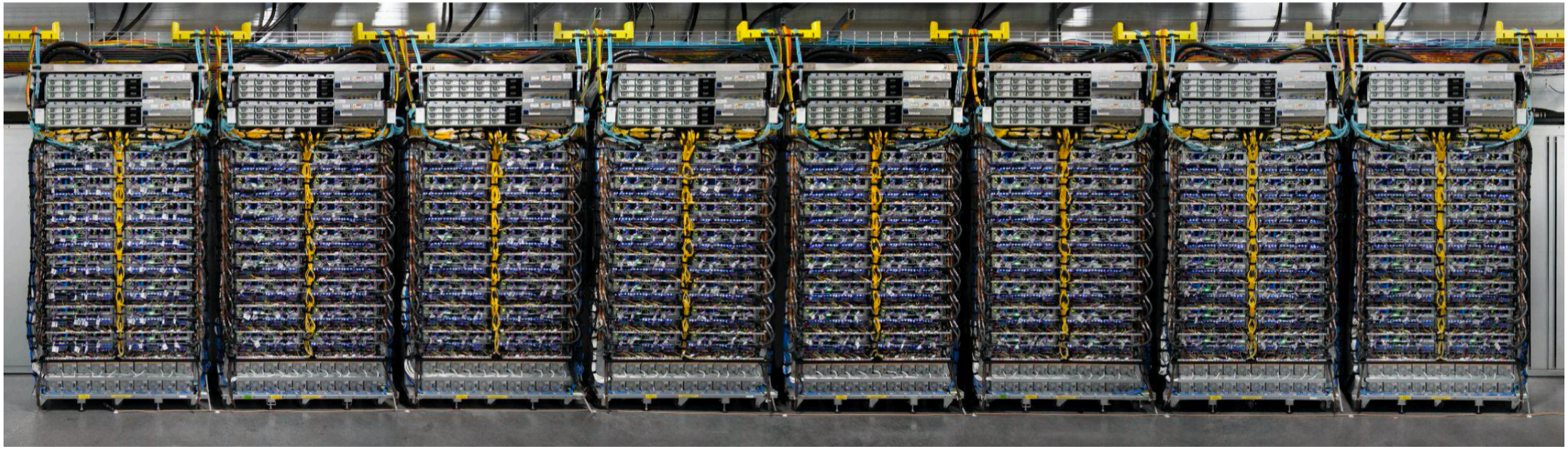
TPU v4 POD

- TPU v4 POD 在ICI网络内部 64 颗 TPU+16 颗 CPU 为一组（TPU Slice），通过直连铜缆连接在4^3 Cube 里面，ICI网络之外为OCS光学背板互连；TPU 集群48 个 OCS Switch 可在单 SuperPod内部署 4096 颗 TPU。**算力密度、同等级带宽下的网络复杂度对比，以及互连设备成本开销的对比较低。**



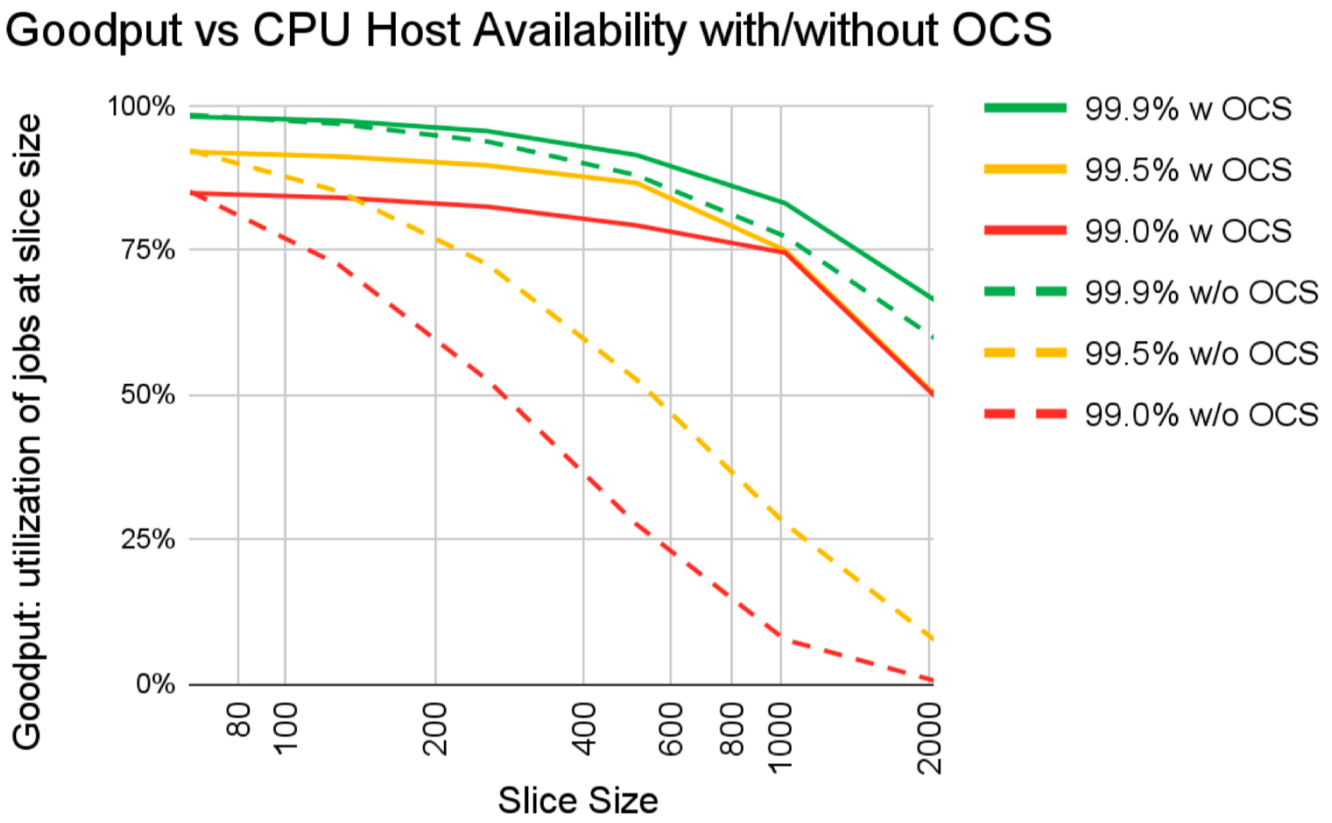
TPU v4 POD

- 与超级计算机一样，工作负载由不同规模的算力承担，称为切片：64 芯片、128 芯片、256 芯片等。与 Infiniband 相比，OCS 的成本更低、功耗更低、速度更快，成本不到系统成本的 5%，功率不到系统功率的 3%。每个 TPU v4 都包含 SparseCores 数据流处理器，可将依赖嵌入的模型加速 5 至 7 倍，但仅使用 5% 的裸片面积和功耗。



TPU v4 POD 测评

- 在使用可配置光互连（以及光路开关）时，假设芯片可靠率在99%的情况下，其整体系统的平均性能提升比不使OCS可高达6倍，可见光互连开关的重要性。

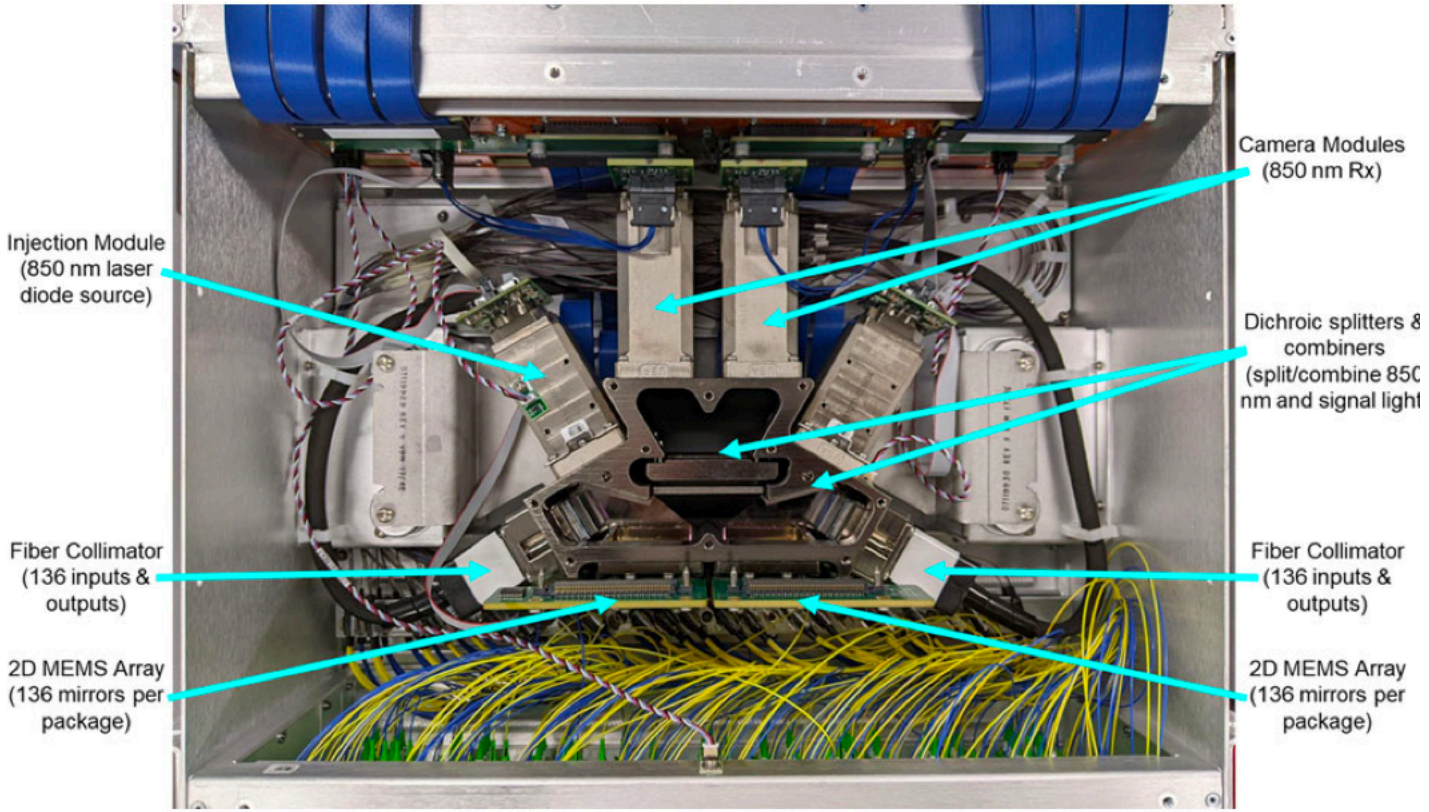


3. 光电路由交换机

Optical Circuit Switching

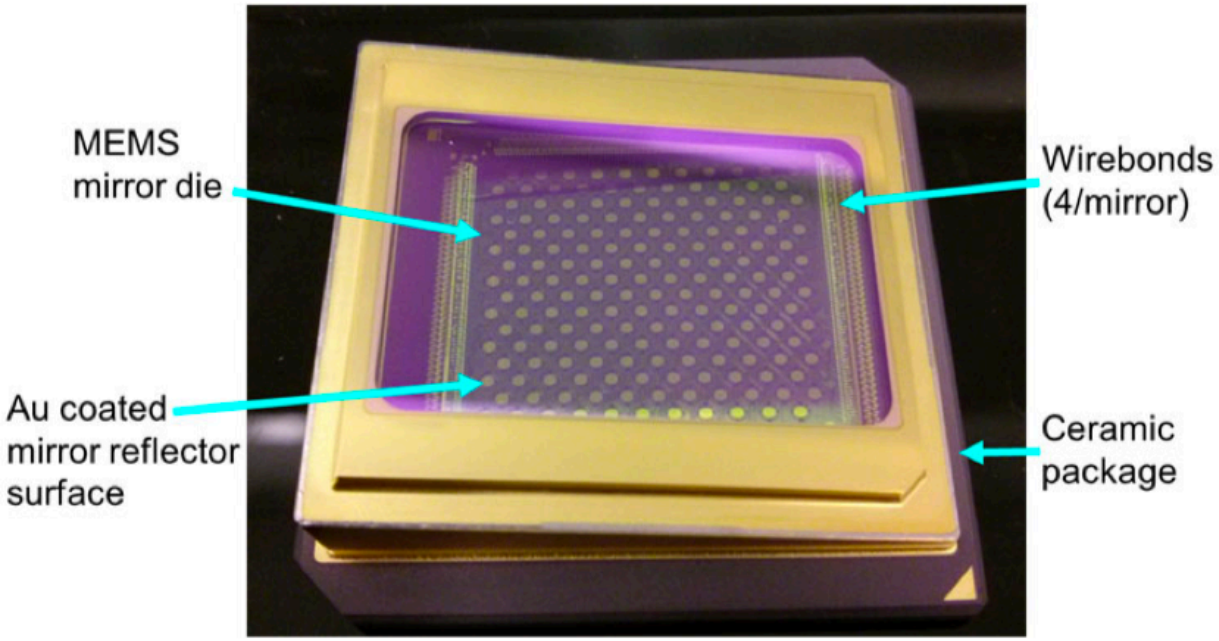
光路开关芯片 Palomar

- 使用的是基于 MEMS 反射镜阵列的技术，具体原理是使用一个 2D MEMS 反射镜阵列，通过控制反射镜的位置来调整光路，从而实现光路切换。使用 MEMS 光路开关芯片可以实现低损耗，低切换延迟（毫秒级别）、低功耗、低成本。

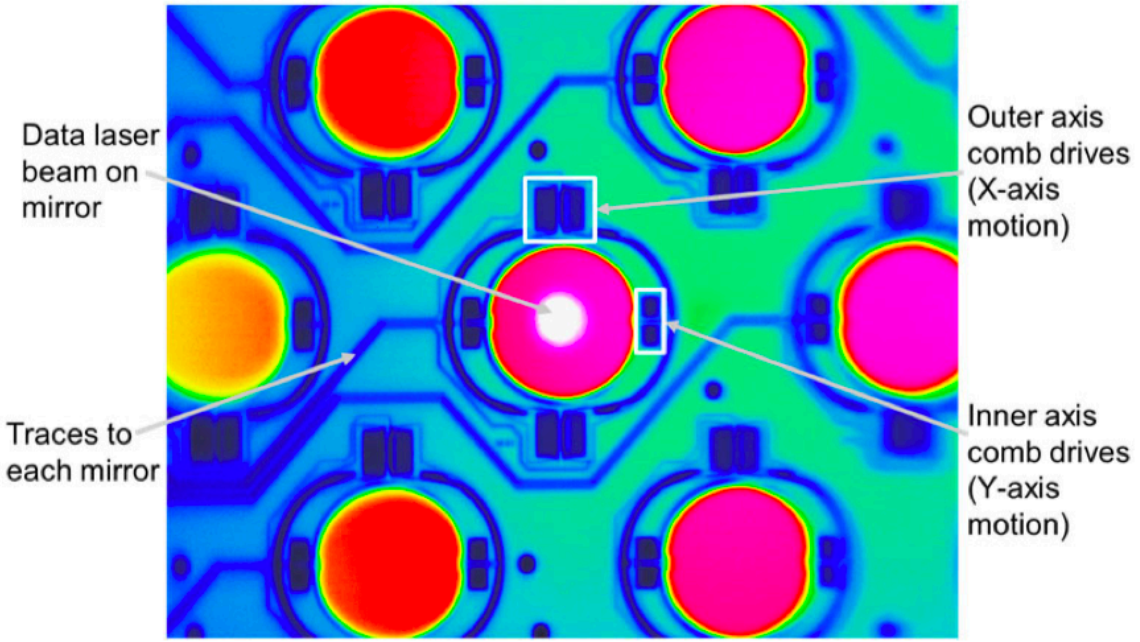


2D MEMS 阵列

- Palomar MEMS反射镜封装的照片。在每个陶瓷封装内部是单个大型芯粒，芯粒里面有176个可单独控制的微反射镜。

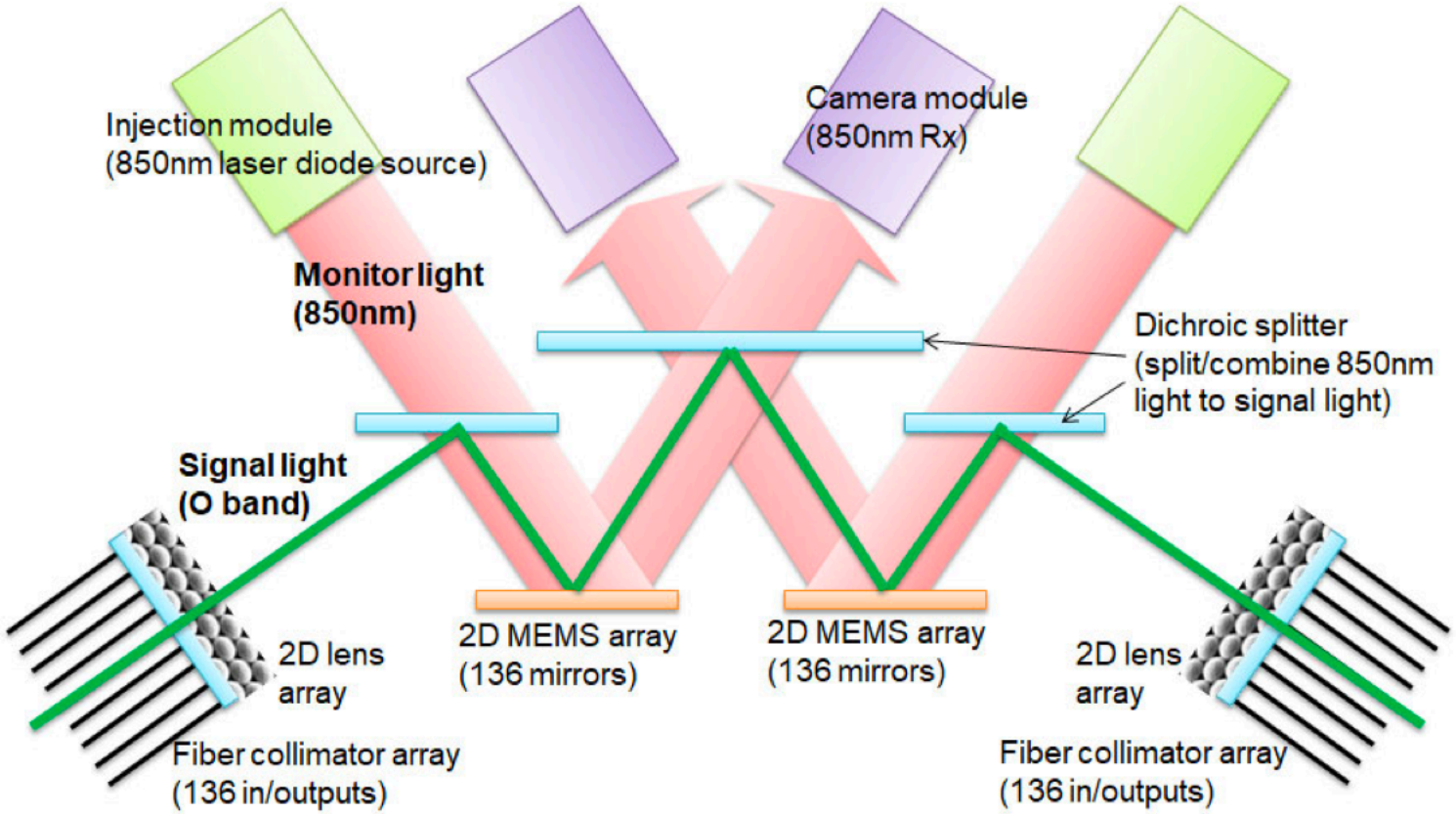


- MEMS反射镜热成像：每个反射镜具有四个梳状驱动区域，用于在两个方向上旋转反射镜。



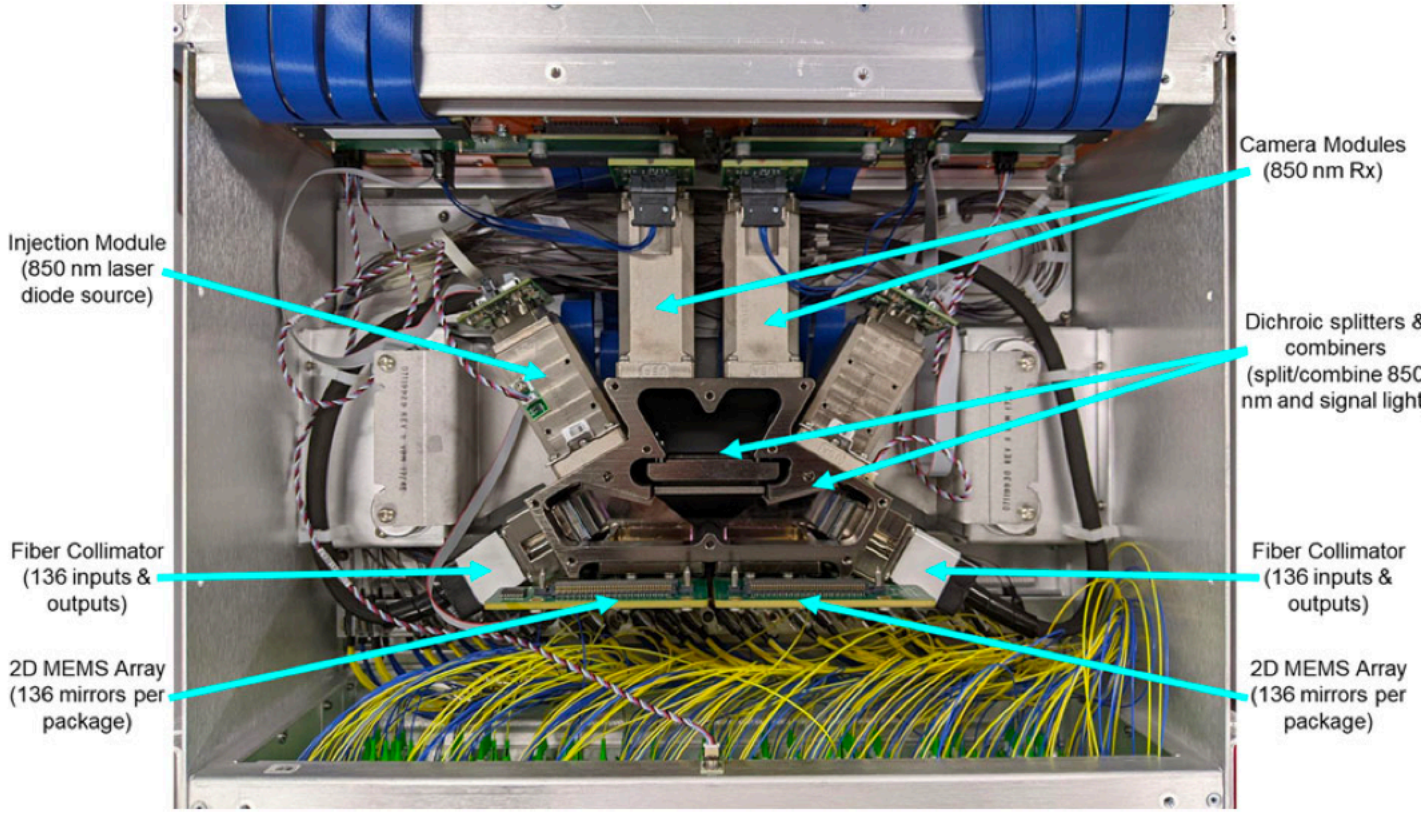
Google 自研光路开关芯片 Palomar 工作原理

- Palomar OCS 光芯设计和光路示意图，使用两个 MEMS 反射镜阵列工作。由绿线指示的带内光信号路径，与带内信号路径叠加，850nm波长通道（红色）用于调节反射镜。
- 不需要光到电到光的转换或耗电的网络分组交换机，从而节省了电力。



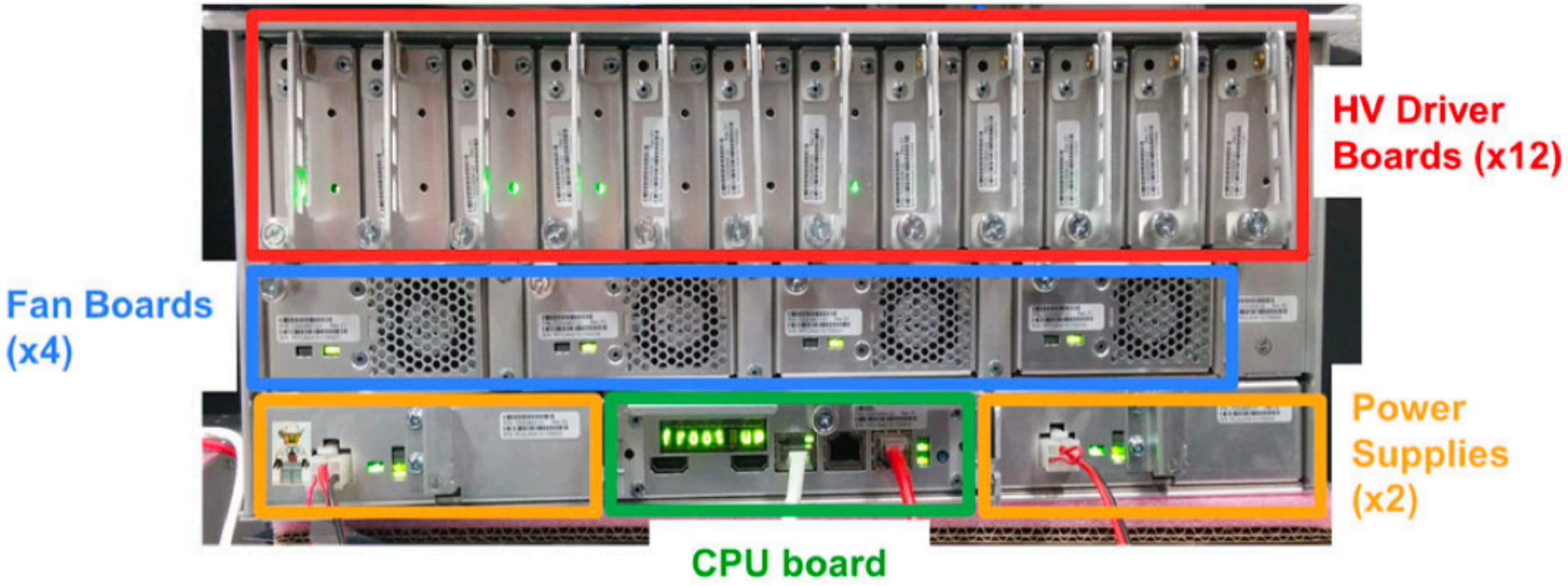
Google 自研光路开关芯片 Palomar 实物图

- Palomar OCS光芯照片，以及相应的关键部件：a）光纤准直器，b）相机模块，c）封装MEMS阵列，d）注入模块，e）二向色分离器和组合器。



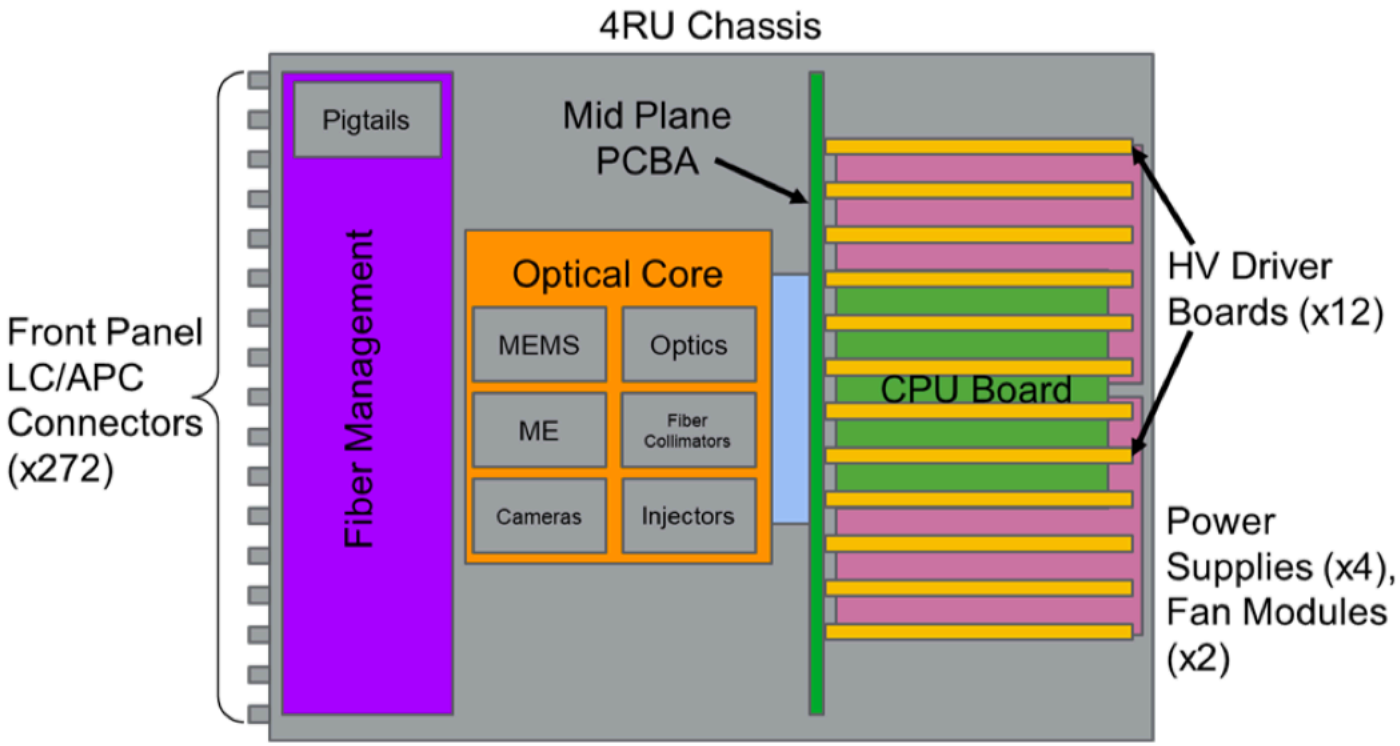
Palomar OCS

- Palomar OCS实物图。带有CPU板、电源、风扇和高压驱动器板的后机箱。



Palomar OCS

- Palomar OCS后机箱的后视图，显示FRU。



4. 思考与优缺点

优点

- **低时延**：3D Tours 因其相邻节点之间的短而直接的连线，可以换来更低的延迟；尤其当节点间需要运行那种密集I/O的、紧耦合的并行任务时特别有用。
- **低网络代价**：对于相同数量的节点， 3D Torus 拓扑网络直径低于 Clos 拓扑，两者相比之下，前者的交换机/线材/连接器的保有量更低，网络层次更少，节省硬件成本。
- **路由可重配**：Google OCS网络支持动态可重配路由，Silic集群在部署之后可以立即投入生产，无需等待整个网络收敛；并且这种特性更容易隔离/下线故障节点。
- **更好集群布局**：集群布局让物理连接上相临的节点间在逻辑上临近，让密集I/O通信、data-flow发生在局部流域，换来更低的通信开销；同时优化了延迟和功耗。这就是3D Torus 将大集群逻辑切割成紧耦合的局部域，局部互连并共享作业。

缺点

- **系统成熟度低：**Clos拓扑本身具备非阻塞特点，性能能够始终保持一致且可以预测，其所有输入/输出都是全带宽同时连接，无冲突无阻塞，这在3D Tours 拓扑中无法保证。
- **拓扑僵硬**：在Clos这种Spine-Leaf脊叶拓扑中，扩容新的叶交换机相对简单，无需更改当前架构；相比之下，扩缩3D Tours 结构比较复杂和耗时，可能需要重新配置整个拓扑。
- **负载均衡问题**：Clos网络在任意两个节点之间提供更多路径，从而实现负载均衡和冗余；虽然3D Tours结构也提供多路径冗余，但显而易见Clos的替代路径数量更多，具体取决于网络的配置。

思考

1. 模型的演进总是要比芯片的设计迭代更快，随着大模型涌现，支持主要依赖AI计算集群，而非提高单芯片能力（单芯片FLOPS没那么香），TPU v4 对于高效互联和规模化的支持。
2. 算法-芯片协同设计是TPU v4的灵魂，人工智能芯片与算法之间的结合继续保持紧密关系。包括对于新数据格式（HF32/BF16）支持，对于稀疏计算支持，对于模型关键算法专用加速器。
3. Benchmark 之外很多生产环境的性能指标都无法直观评测，XLA 编译器的优化效益也难以直观评测。到底TPU v4真的那么香吗？

Reference 引用&参考

1. Jouppi, Norm, et al. "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings." Proceedings of the 50th Annual International Symposium on Computer Architecture. 2023.
2. https://mp.weixin.qq.com/s/xq4G52YA4xo20X0Jo1KRmg
3. https://mp.weixin.qq.com/s/XK6-p6ocjq5XXx5AvGD9Lg
4. https://www.nextplatform.com/2022/10/11/deep-dive-on-googles-exascale-tpuv4-ai-systems/
5. https://unwire.pro/2018/01/06/google-ai-and-tpu/news/
6. https://www.zhihu.com/question/594797182
7. https://blog.csdn.net/df12138/article/details/127045083
8. https://www.youtube.com/watch?v=VQoyypYTz2U
9. https://zhuanlan.zhihu.com/p/564158324
10. https://www.cnblogs.com/sea-wind/p/10993958.html
11. https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext?mobile=false
12. https://www.top500.org/news/google-reveals-major-upgrade-and-expanded-role-for-tpu/
13. https://baijiahao.baidu.com/s?id=1762327109978922365&wfr=spider&for=pc
14. https://baijiahao.baidu.com/s?id=1762787905708085293&wfr=spider&for=pc

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem