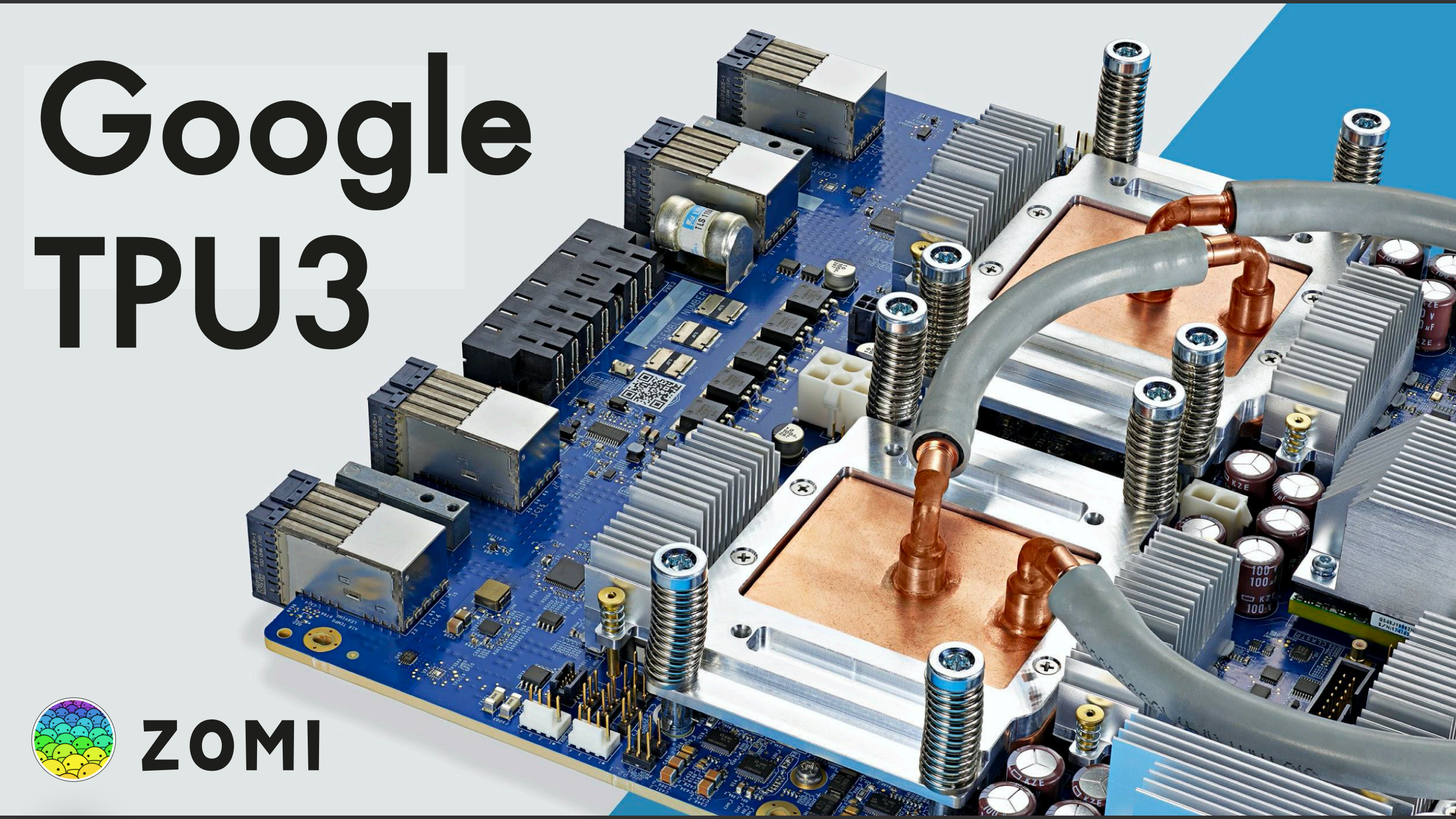
# Google TPU3



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

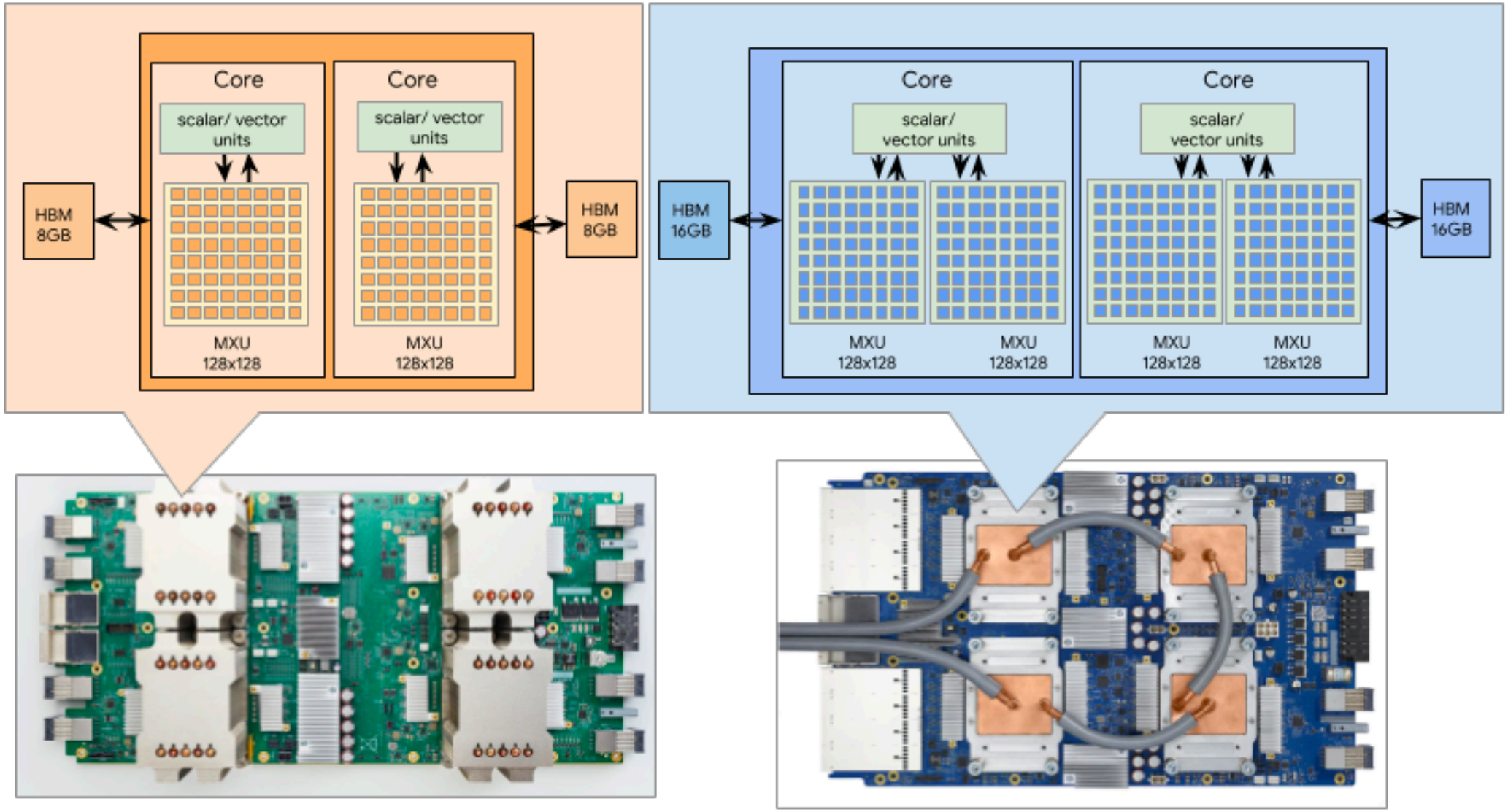
# 1. TPU v3 介绍

# TPU v3 相比 TPU v2

- TPU v3 实际上就是 TPU v2 增强版。TPU v3 相比 TPU v2 有约 1.35 倍的时钟频率、ICI带宽和内存带宽，两倍 MXU 数量，峰值性能提高 2.7 倍。
- 体积只大了不到 10%，MXU 数量翻倍达到 4 个；时钟频率加快了 30%，进一步加快计算速度；内存带宽扩大了 30%，容量翻倍，可使多种应用更加方便；芯片间带宽扩大 30%，可连接的节点数是之前的 4 倍。

# TPU历代芯片

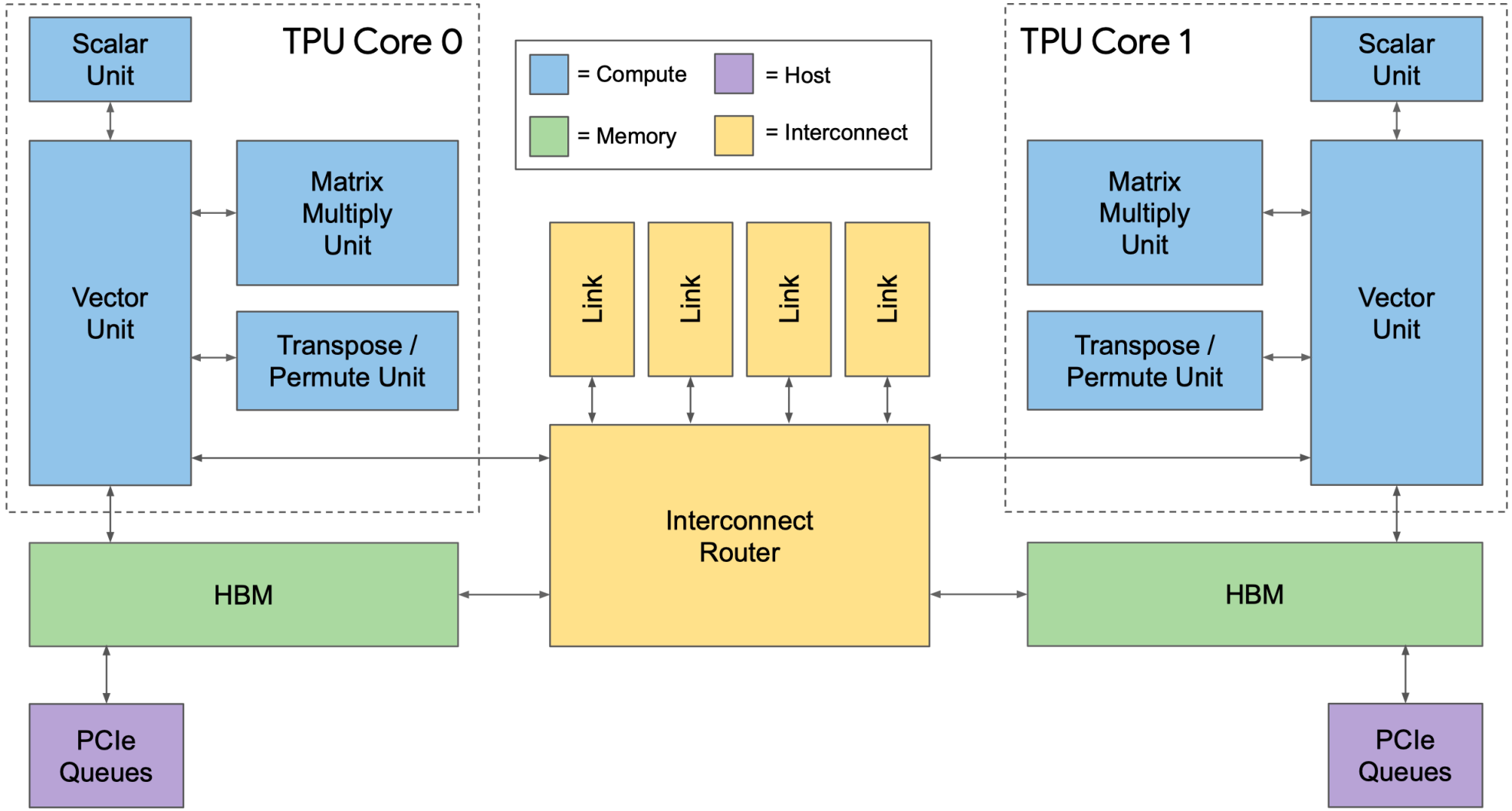
| | TPUv1 | TPUv2 | TPUv3 |
|---|---|---|---|
| Date introduced | 2016 | 2017 | 2018 |
| Process node | 28 nm | 16 nm | 16 nm |
| Die size ($mm^2$) | 330mm | 625mm | 700mm |
| On-chip memory (MB) | 28MB | 64MB | 64MB |
| Clock speed (MHz) | 700MHz | 700MHz | 940MHz |
| Memory | 8 GB DDR3 | 16 GB HBM | 32 GiB HBM |
| Memory bandwidth | 300 GB/s | 700 GB/s | 900 GB/s |
| TDP (W) | 75 | 280 | 450 |
| TOPS (Tera/Second) | 92 | 180 | 360 |
| TOPS/W | 0.31 | 0.16 | 0.56 |



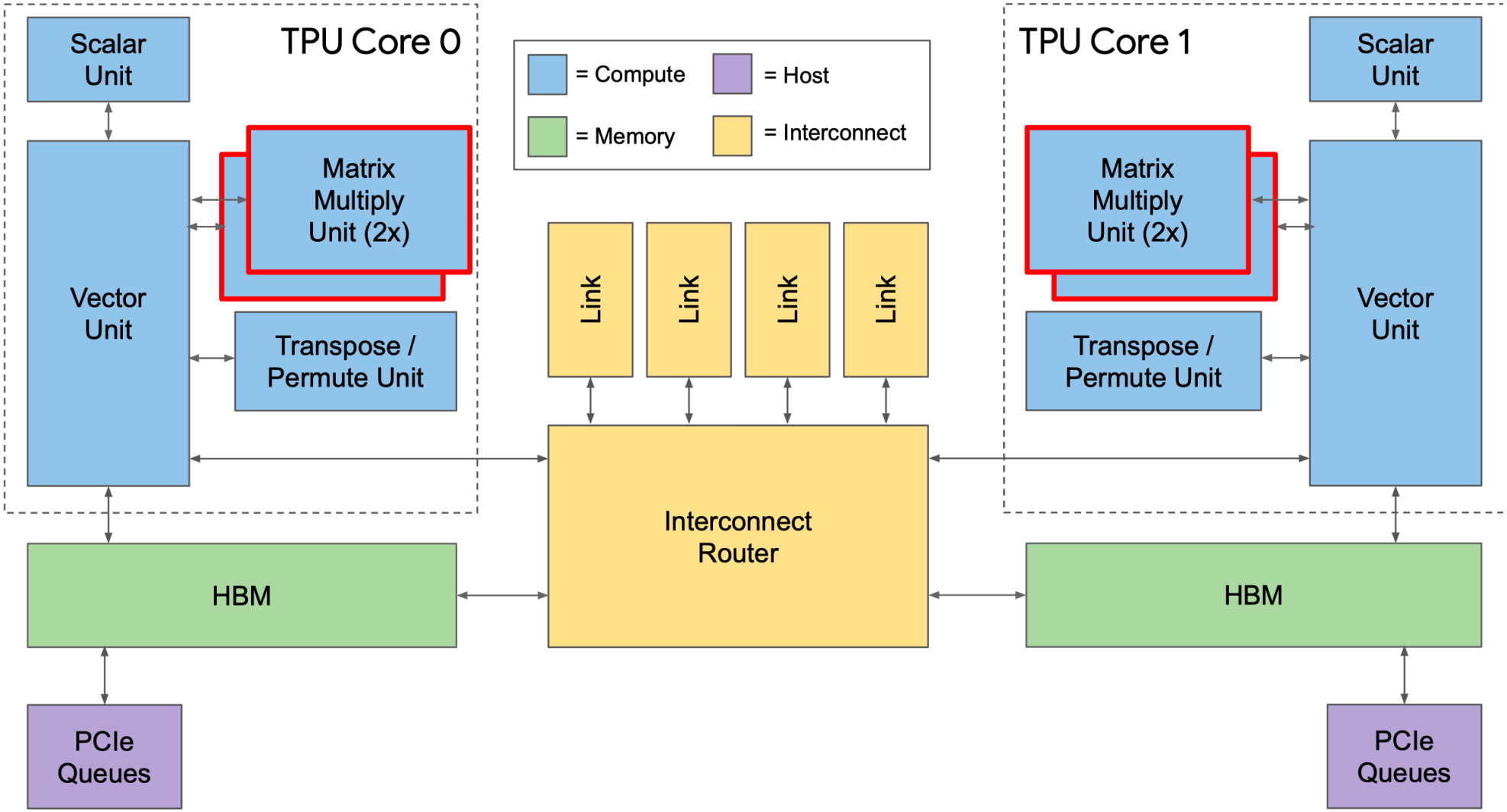

TPU v2 - 4 chips, 2 cores per chip

TPU v3 - 4 chips, 2 cores per chip

# TPU v2 第一款训练卡



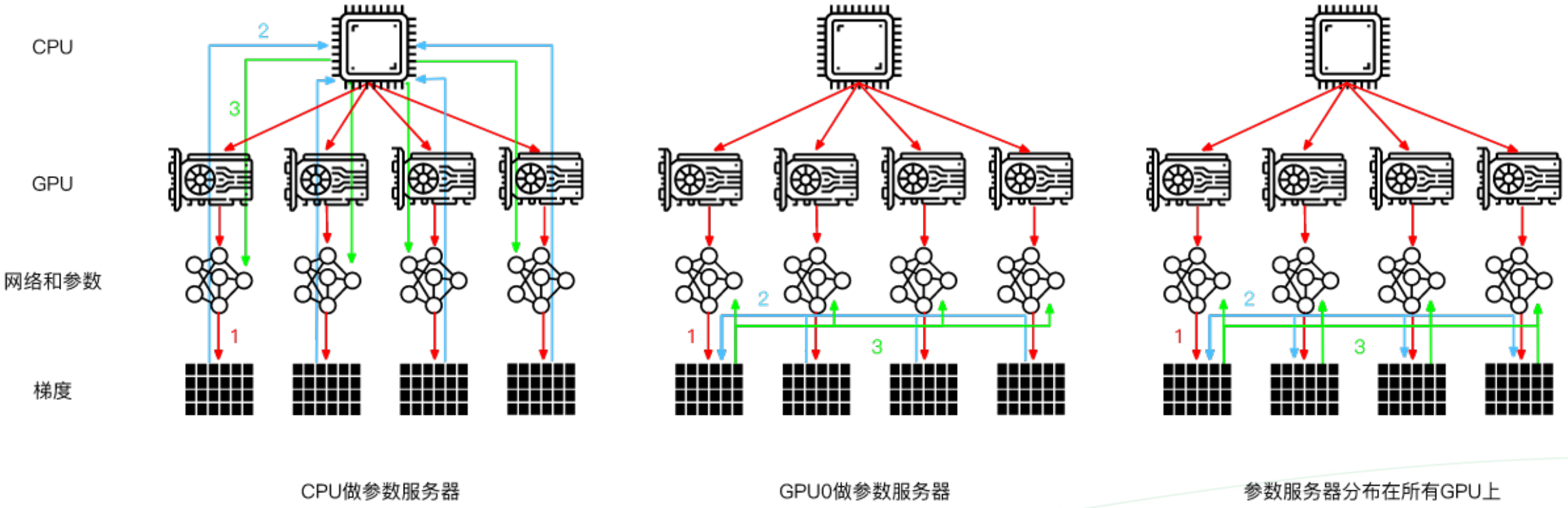
# TPU3 + 30% Freq + 30% b/w + 4x node



# 2. 基本概念澄清

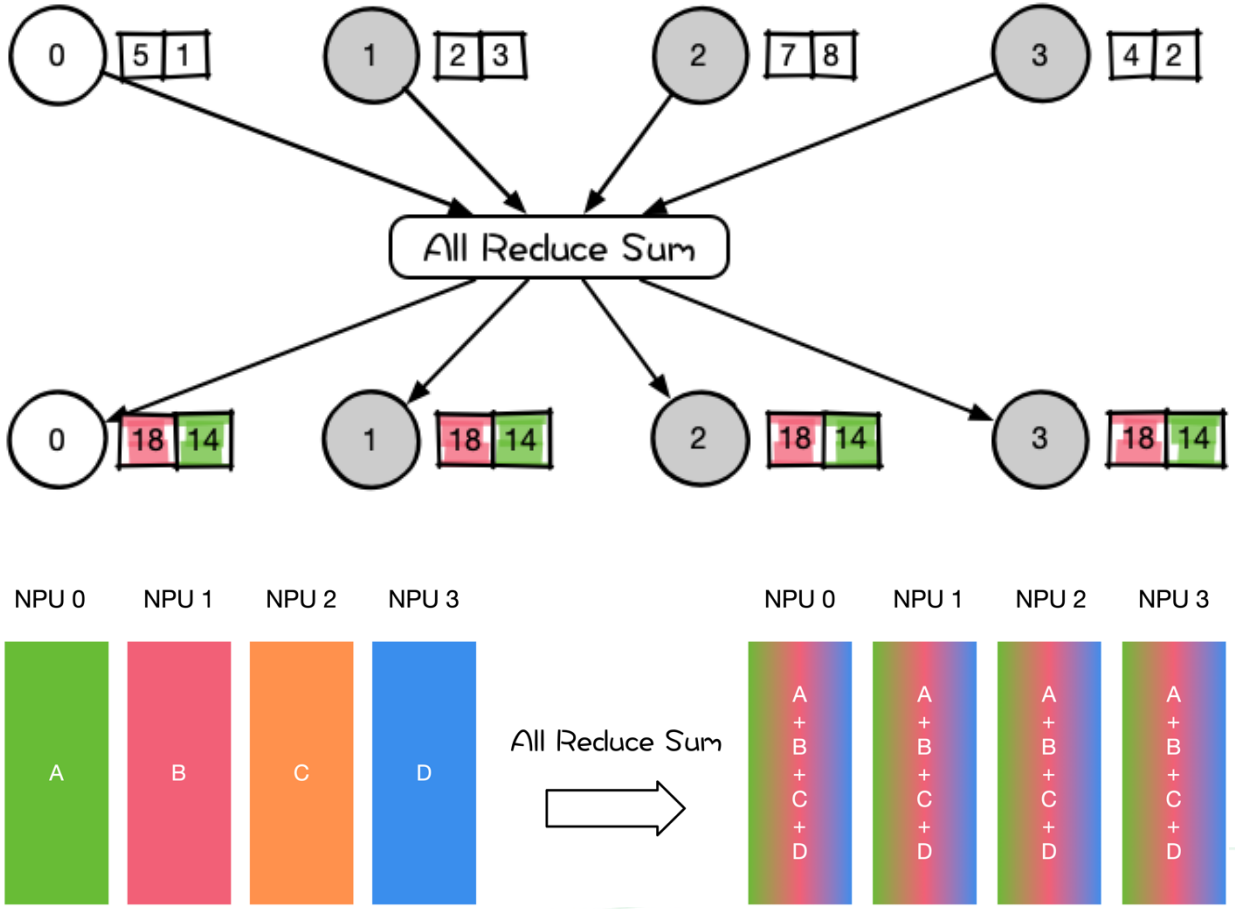
# 概念：分布式架构 - 参数服务器

（1）计算损失和梯度（2）梯度聚合（3）参数更新并参数重新广播



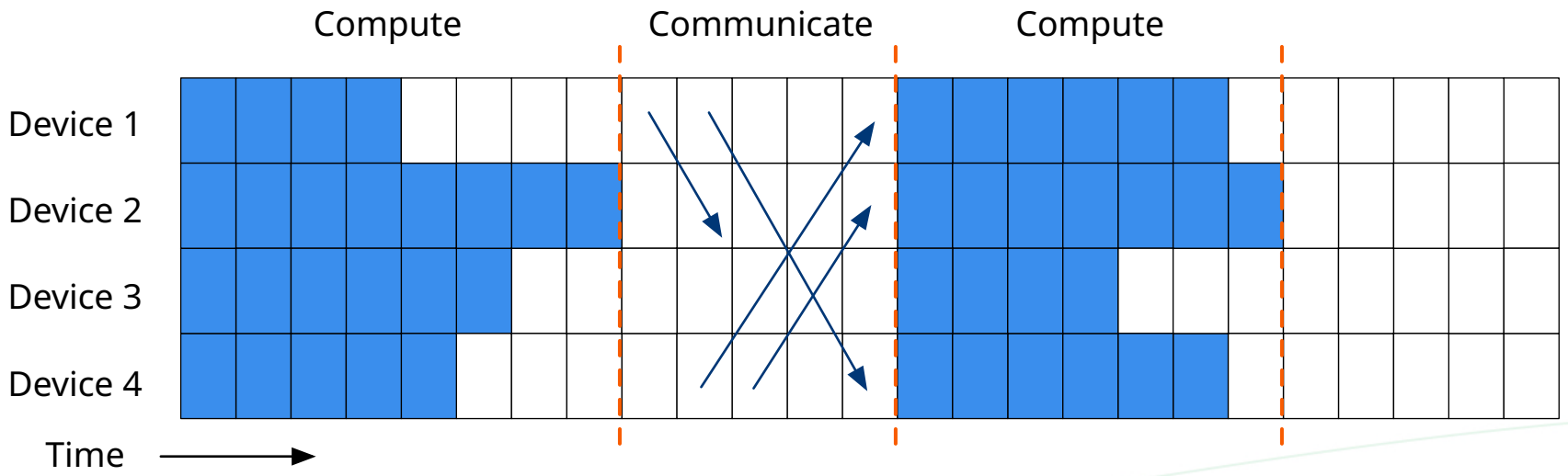
# 概念：POD 中的通信

- 超级计算机中，执行的大部分是神经网络模型的 DP（Data Parallel）计算，大多数流量都是来自所有节点的权重更新时的 all-reduce 通信，所以现在研发说的 Host Bound 和 Device Bound 就来自于这里。AI 的应用在集群环境中属于 Device Bound。



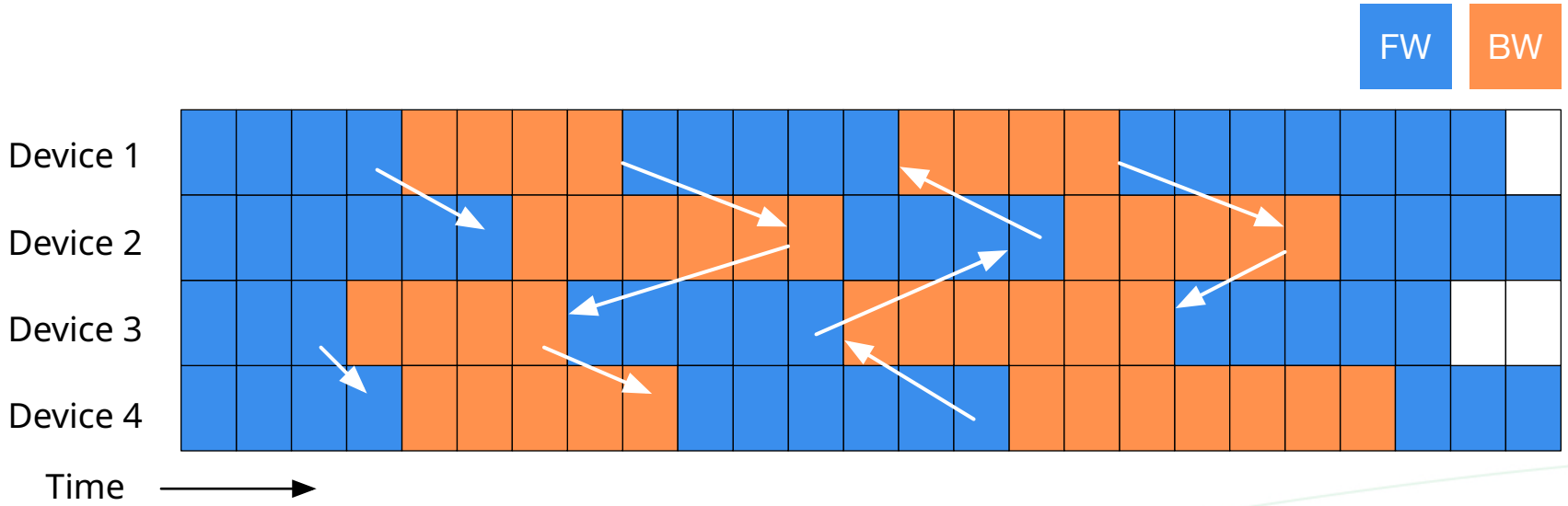
# 概念：分布式 - 同步并行

- 必须等全部工作节点完成了本次通信之后才能继续下一轮本地计算
- **优点**：本地计算和通信同步严格顺序化，能够容易地保证并行的执行逻辑于串行相同
- **缺点**：本地计算更早的工作节点需要等待其它工作节点处理，造成了计算硬件的浪费。

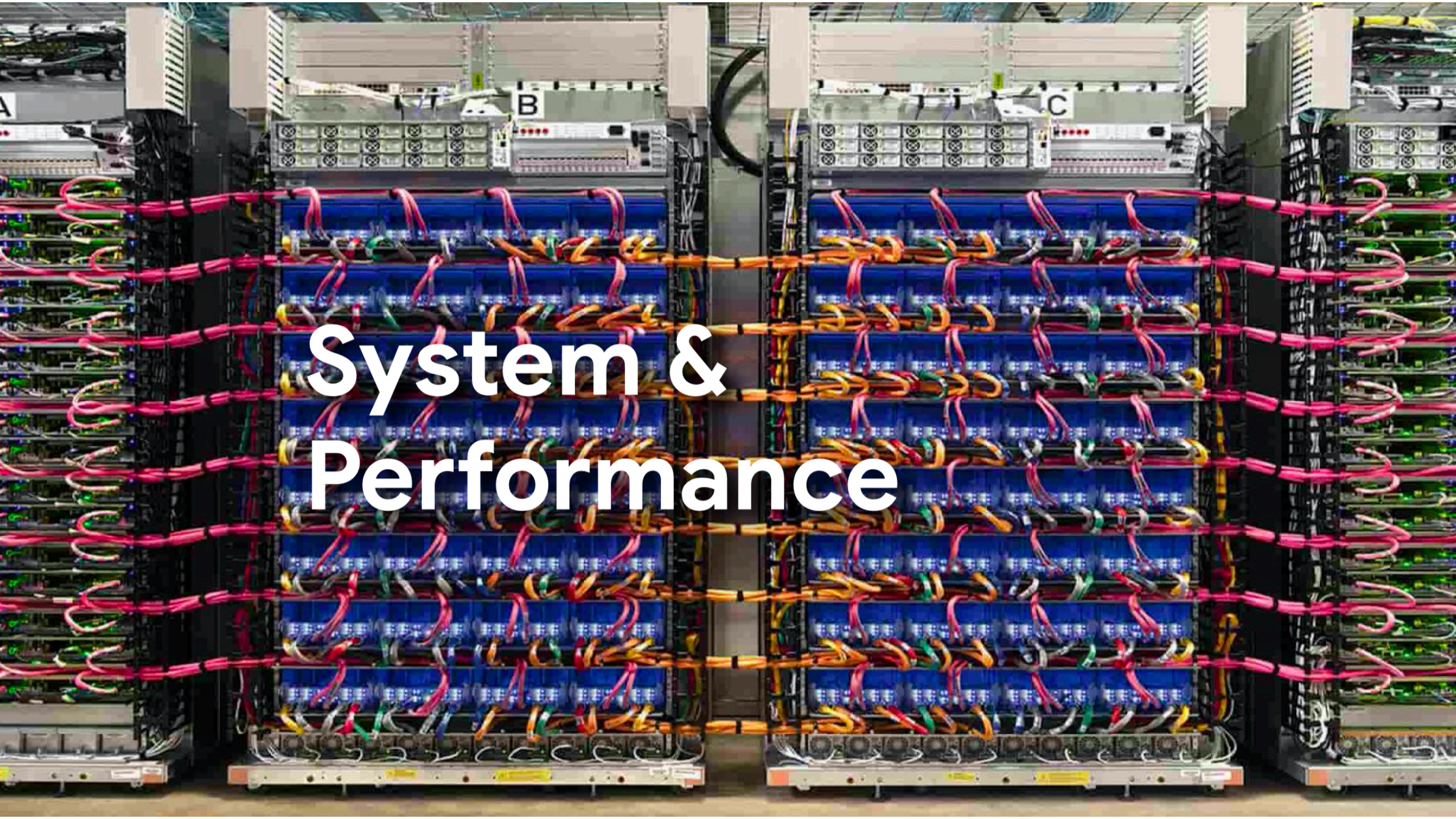


# 概念：分布式 - 异步并行

- 当前batch迭代完后与其他服务器进行通信传输网络模型参数
- **优点**：执行效率高，中间除了单机通信时间以外没有任何通信和执行之间的阻塞等待
- **缺点**：网络模型训练不收敛，训练时间长，模型参数反复使用导致无法工业化



# 3. POD 超算形态



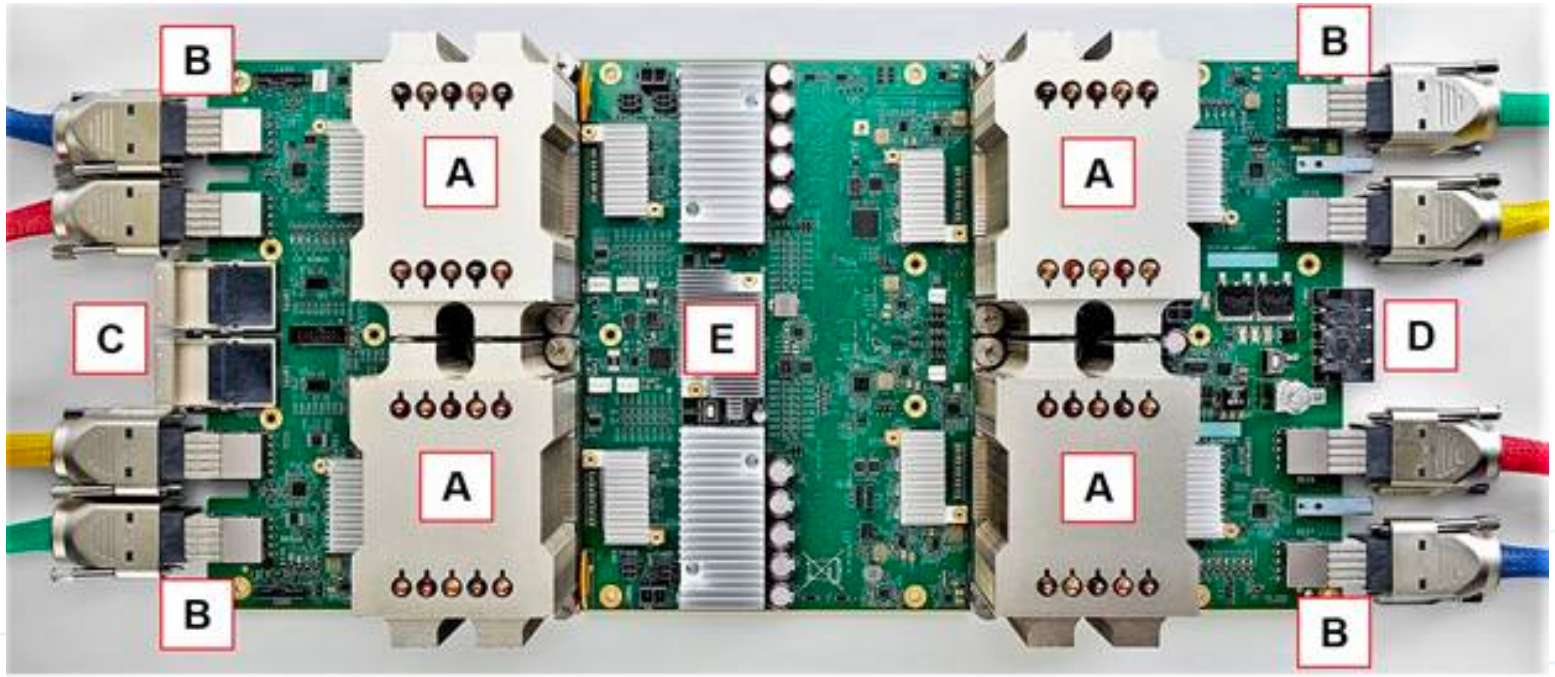

# System & Performance

# 迎来 Supercomputer

- TPUv1: single-chip system -- built as coprocessor to a CPU
  - Works well for inference
- TPUv2, v3: ML Supercomputer – name as POD
  - Multi-chip scaling critical for practical training times
  - Single TPU v2 chip would take 60 - 400 days for production workloads

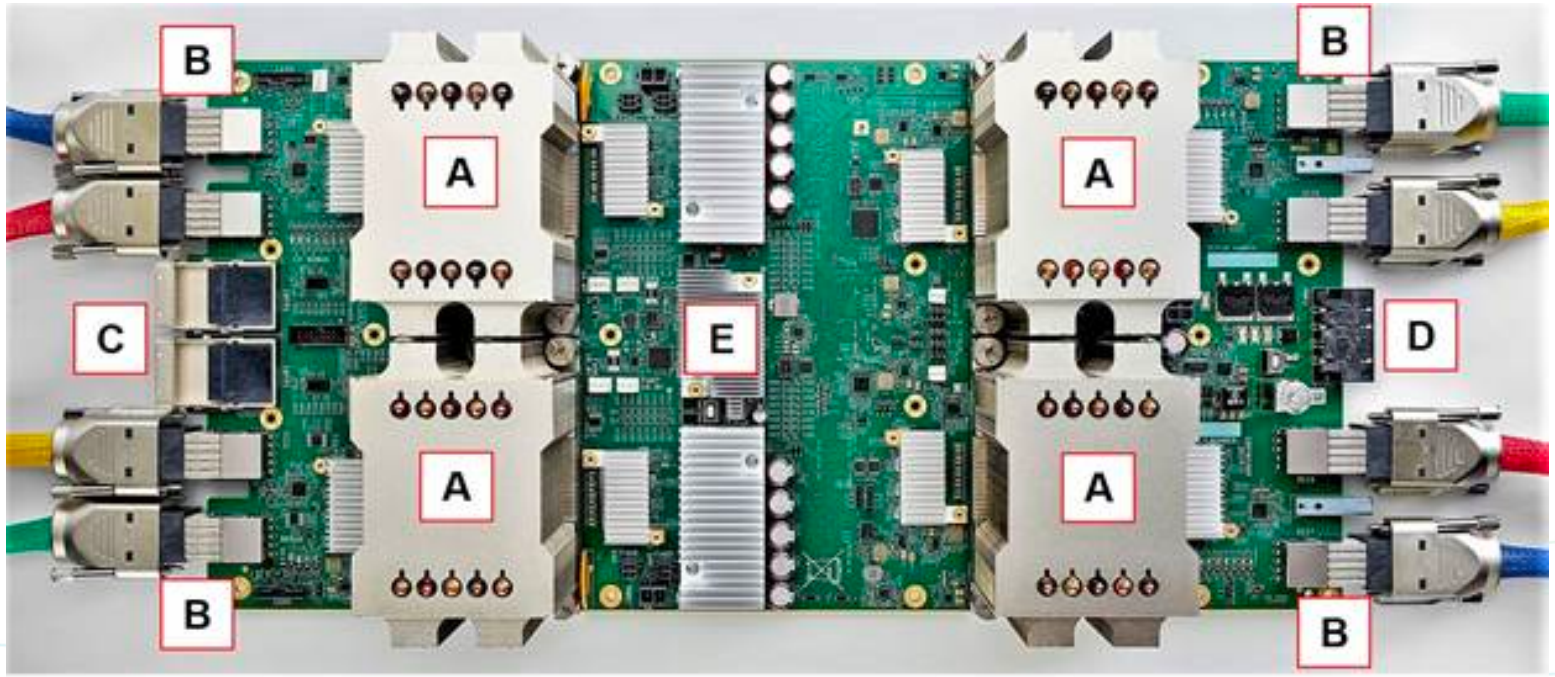
# TPU v2 基板组成

- A：四个TPU v2芯片和散热片；
- B：2 个 BlueLink 25GB/s 电缆接口；
- C：Intel 全路径体系结构（OPA）电缆；
- D：电路板电源连接器；
- 支持两种网络配置，分别为10Gbps 以太网和 100Gbps Intel OPA连接。

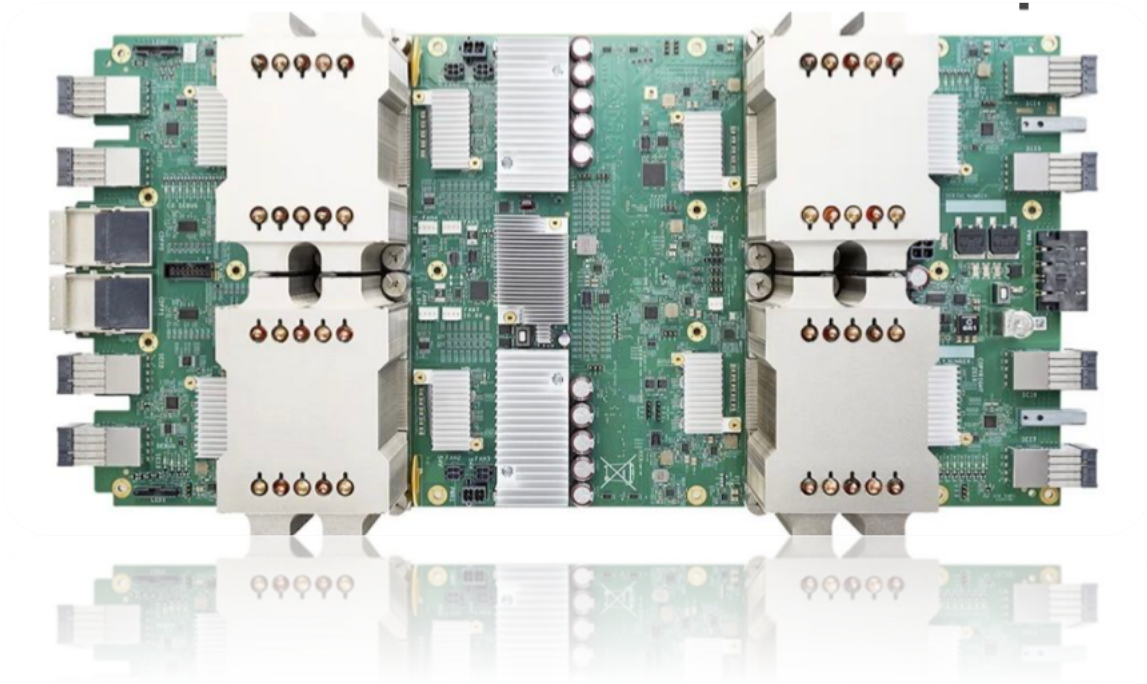


# 概念澄清

- **BlueLink**： IBM BlueLink端口协议，每 Socket 25 Gb/s 带宽。互联节点间进行NUMA通信。
- **OPA**：英特尔 Intel Omni-Path Architecture (OPA) 互联架构，与InfiniBand相似的网络架构。



# Supercomputer with dedicated interconnect

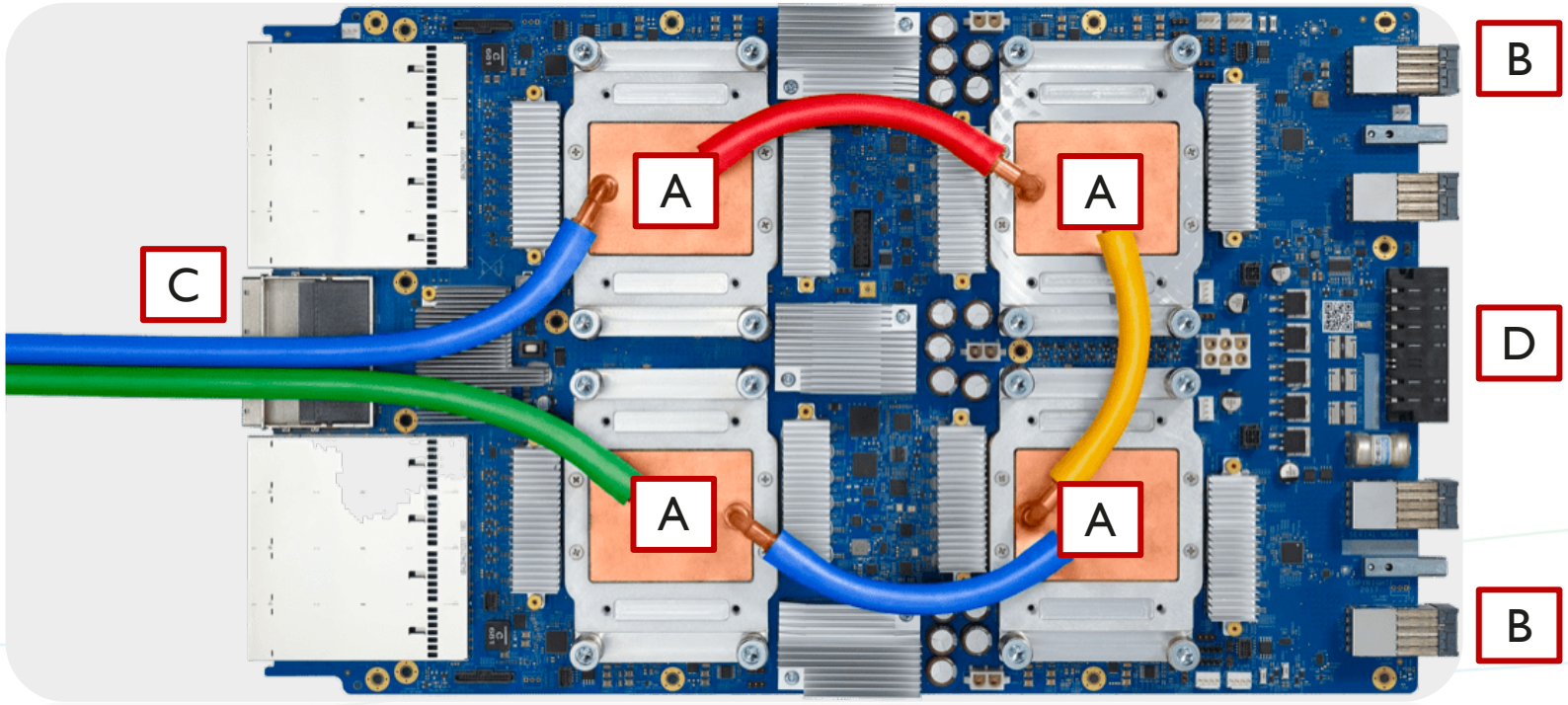


TPUv2 boards = 4 chips

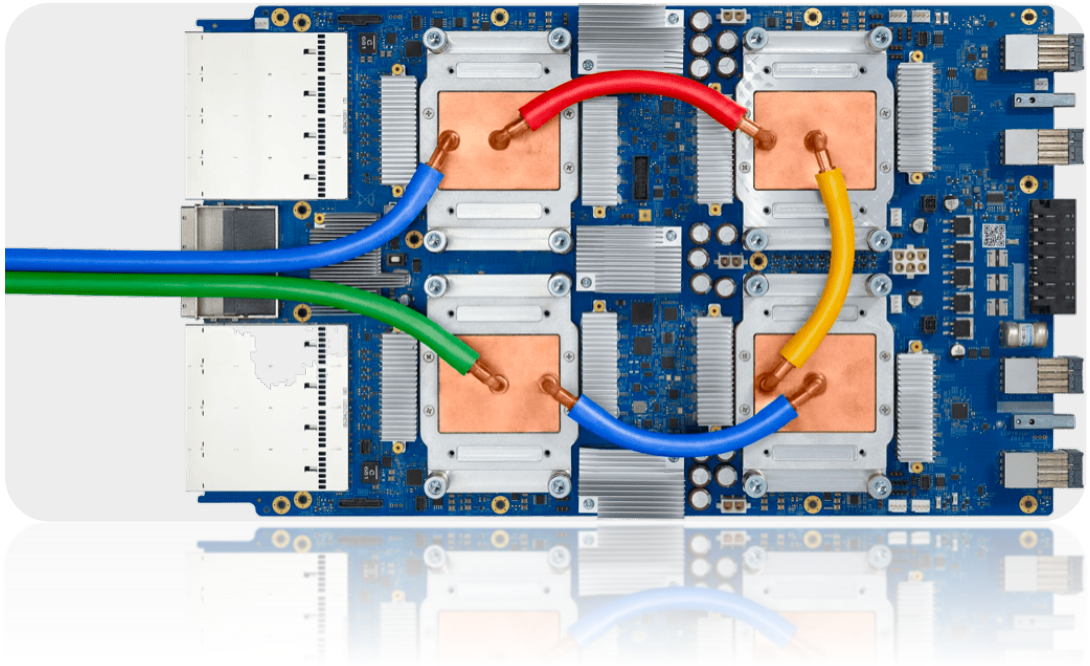

TPUv2 supercomputer (256 chips)

# TPU v3 基板组成

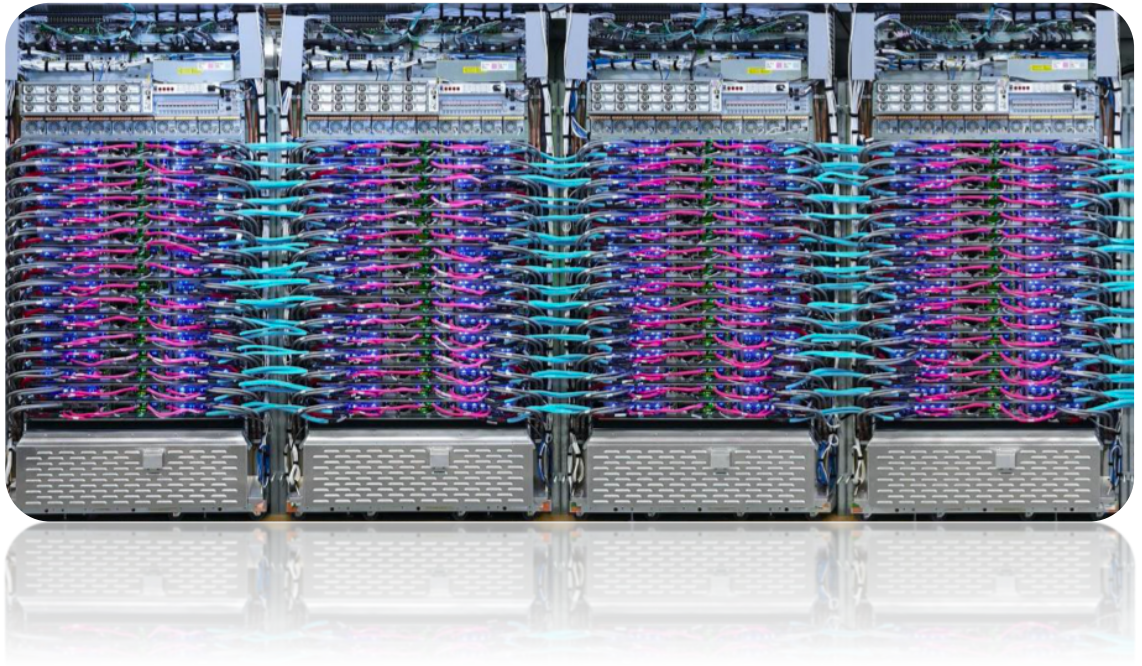
- A：四个TPU v2芯片和液冷散热管；
- B：2 个 BlueLink 25GB/s 电缆接口；
- C：Intel 全路径体系结构（OPA）电缆；
- D：电路板电源连接器；
- 支持两种网络配置，分别为10Gbps 以太网和 100Gbps Intel OPA连接。



# Supercomputer with dedicated interconnect



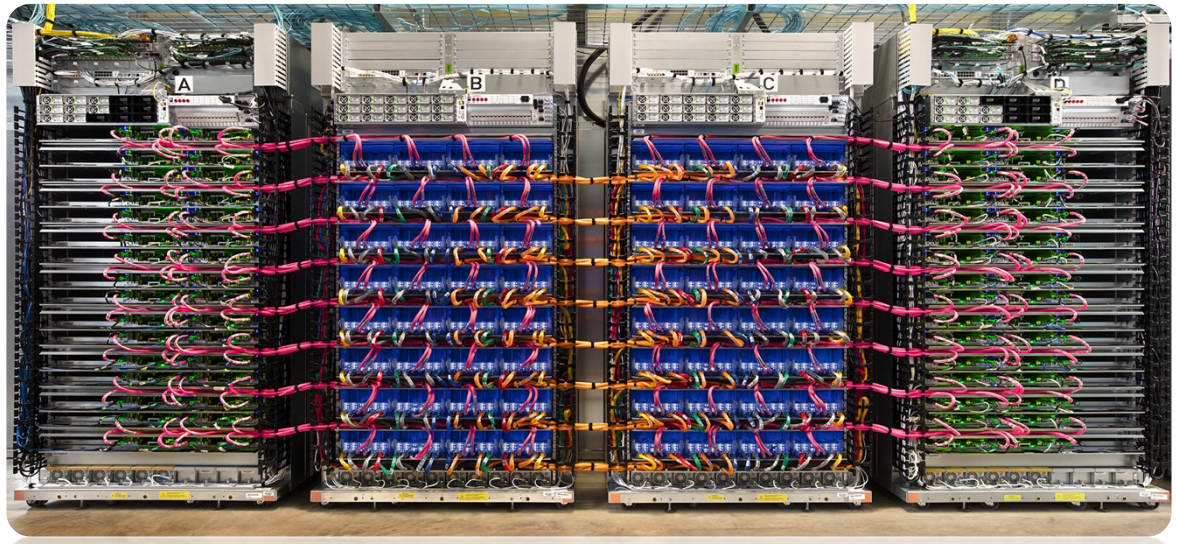
TPUv3 boards = 4 chips



TPUv3 supercomputer (1024 chips)

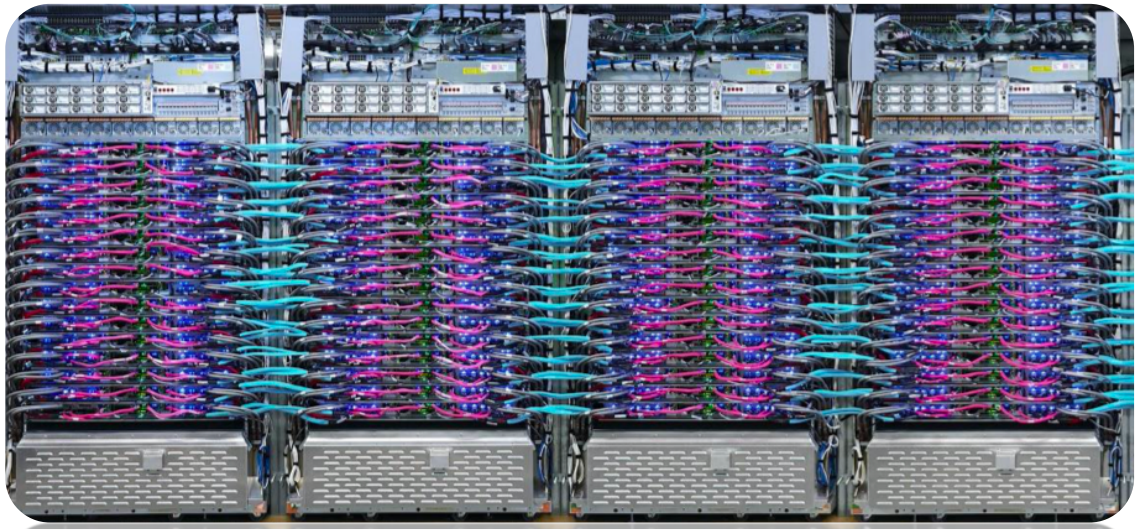
# Supercomputer with dedicated interconnect

TPUv2 supercomputer (256 chips)



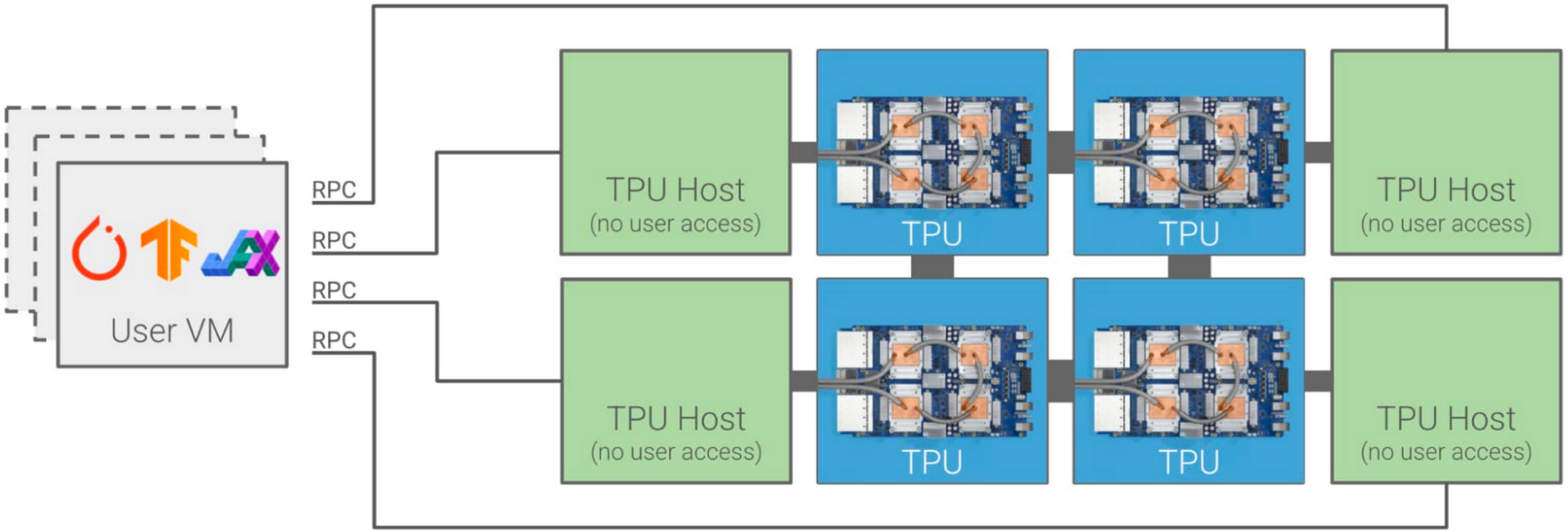

- 11.5 petaflops
- 4 TB HBM
- 2-D torus
- 256 chips

TPUv3 supercomputer (1024 chips)



- 100 petaflops
- 32 TB HBM
- Liquid cooled
- 1024 chips

# 虚拟架构图

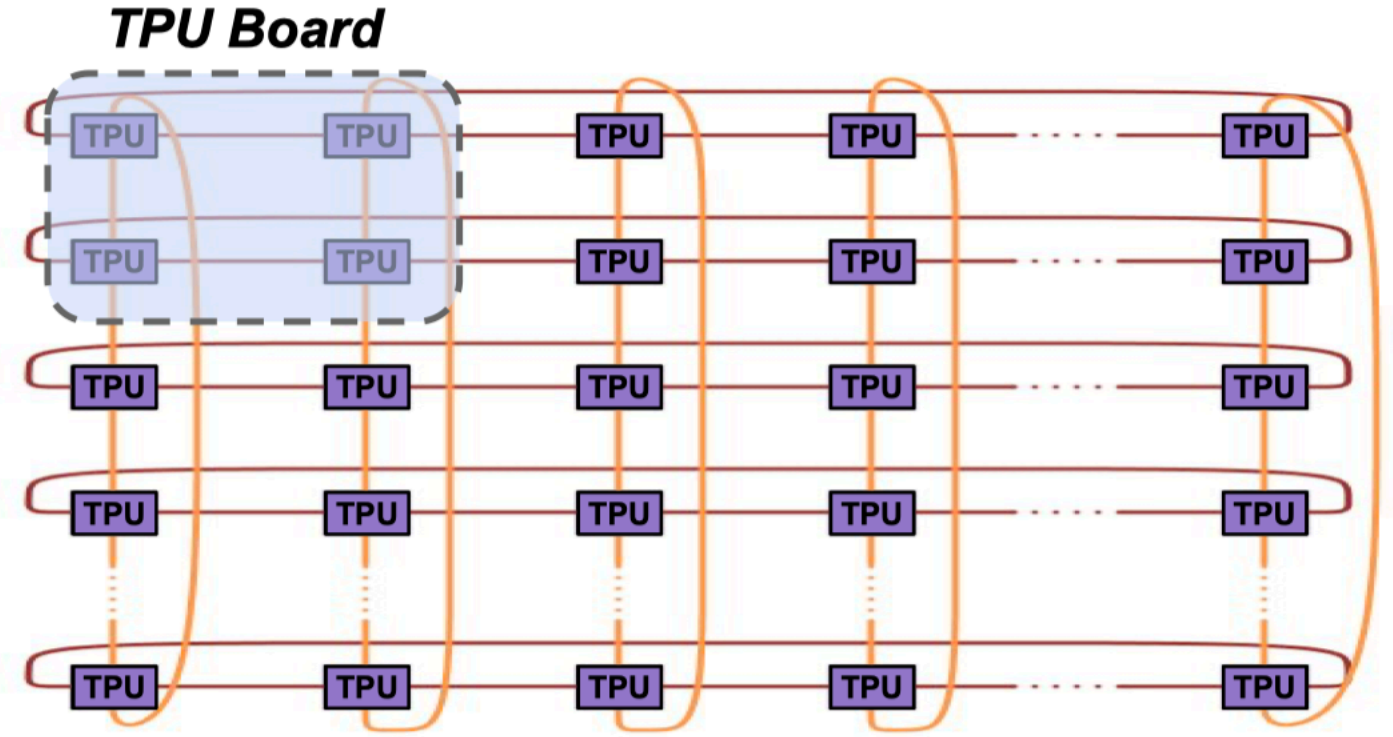

User VM
RPC
RPC
RPC
RPC
TPU Host
(no user access)
TPU
TPU
TPU Host
(no user access)
TPU Host
(no user access)
TPU
TPU
TPU Host
(no user access)

# TPU v2/v3 Supercomputer

1. 模型训练常需要大规模数据才能完成，TPU v2 重要改进是增加了核间互连结构（ICI），使最多 256 个 TPU v2 组成一个超级计算机。Google 基于 TPU v2 开发的超级计算机，会使用多个 TPUv2 组成集群对深度学习网络进行训练。
2. Google 将1024 个 TPU v3 组成超级计算机 TPU POD，服务器形态采用水冷方式，能使功率提升 1.6 倍；TPU v3 模具尺寸只比 TPU v2 大 6%。
3. 通过交换机提供的虚拟电路和无死锁路由功能，再加上 TPUv2 本身存在的核间互连(ICI)结构，便构建出 TPUv2 集群：TPU v2 集群的2D tours，提供了 15.9T/s 的平分带宽，相比传统的集群组网，省去了集群网卡、交换机的成本，以及与集群 CPU 的通信延迟。

# TPU Training Pod Architecture

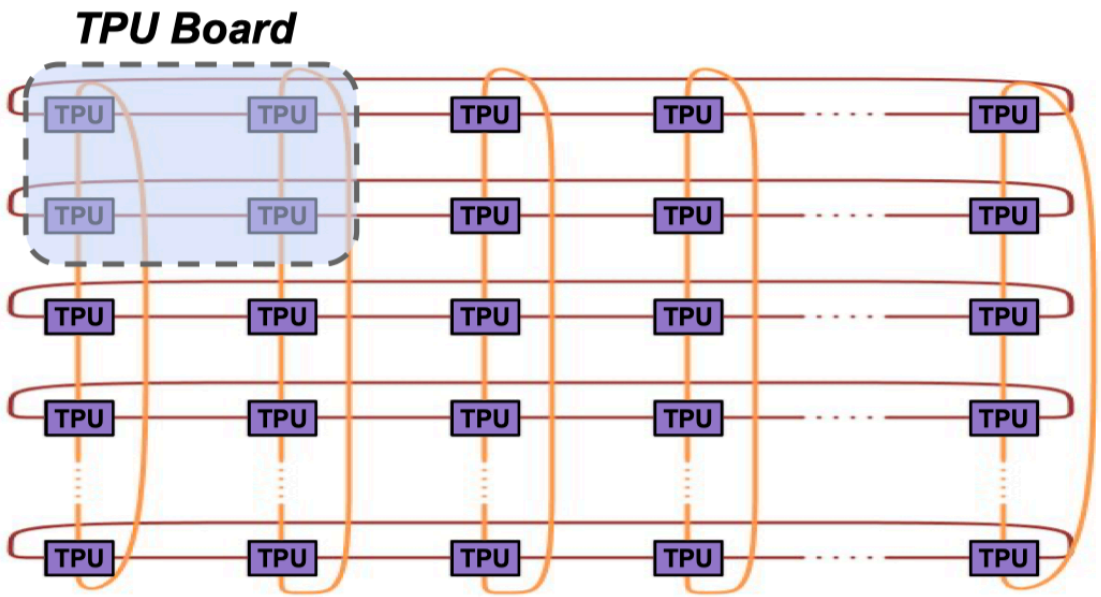
同步训练在同等资源的条件下可以击败异步 SGD 训练，是因为同步训练不需要参数服务器，直接允许节点间的通信，使用 all-reduce 模式就可以获得一致的权重，解决了异步训练时参数服务器与各工作节点带宽不足的问题。

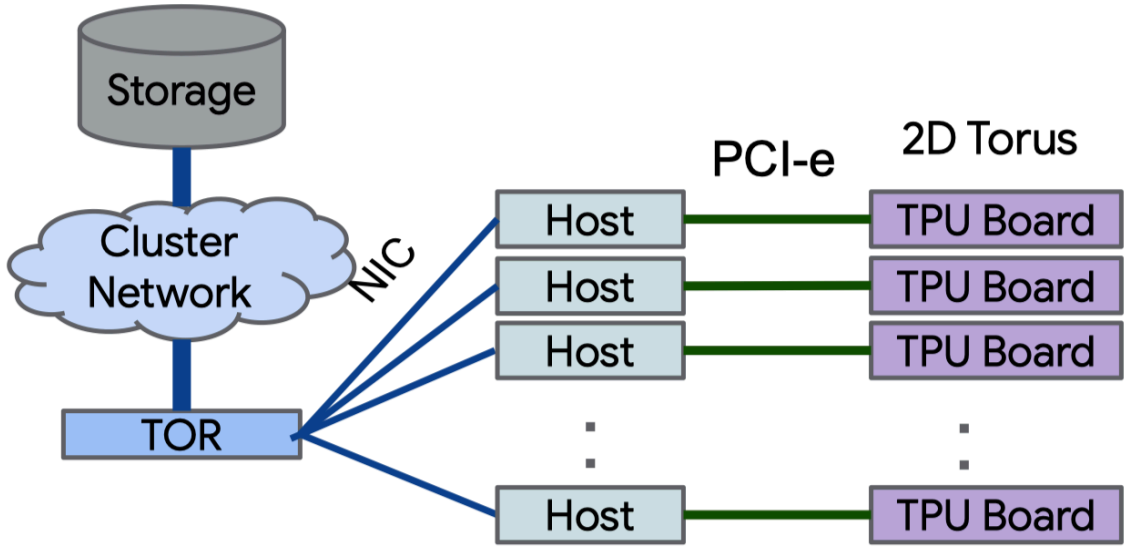



**TPUs interconnected in 2D Torus**

Dedicated network for synchronous parallel training

# TPU Training Pod Architecture



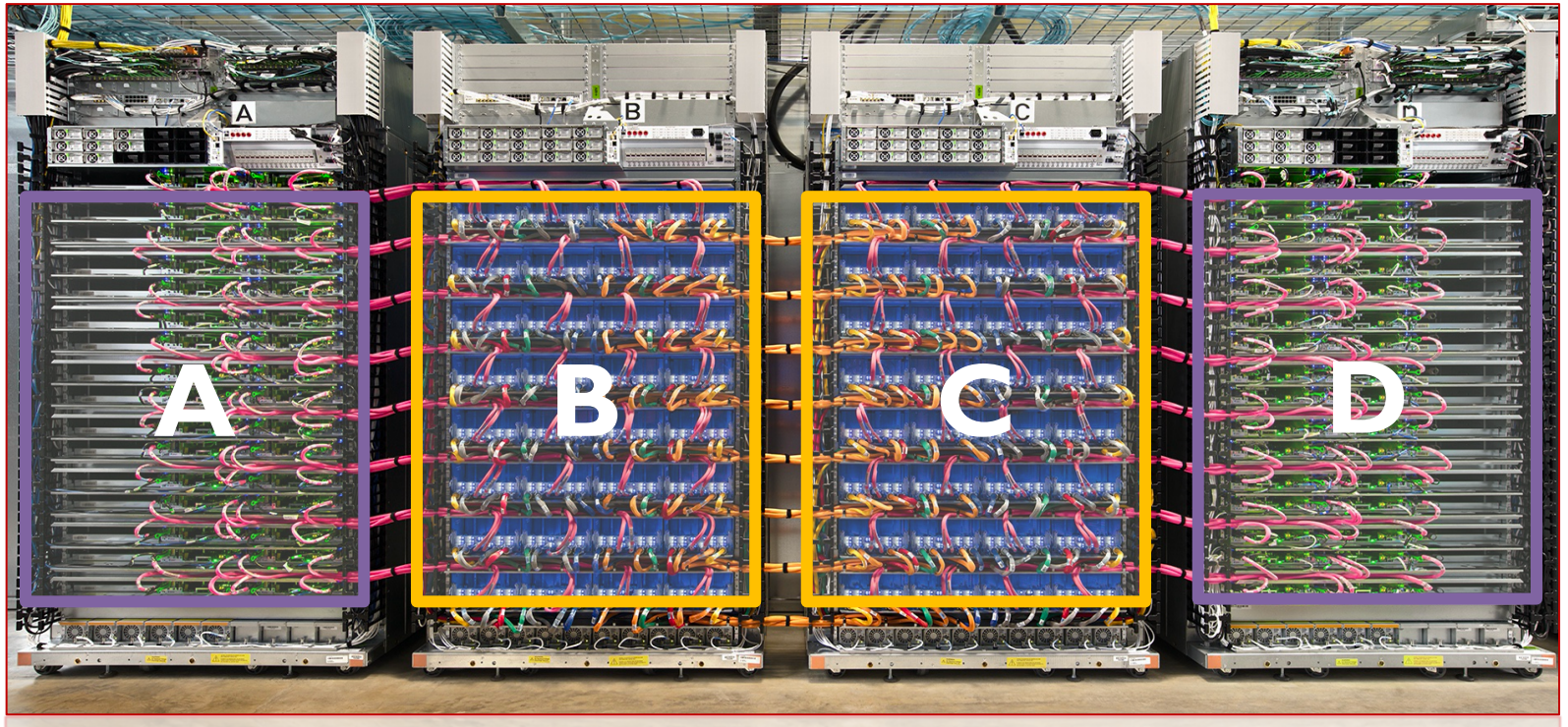

**TPUs interconnected in 2D Torus**

Dedicated network for synchronous parallel training

# 4. POD详情

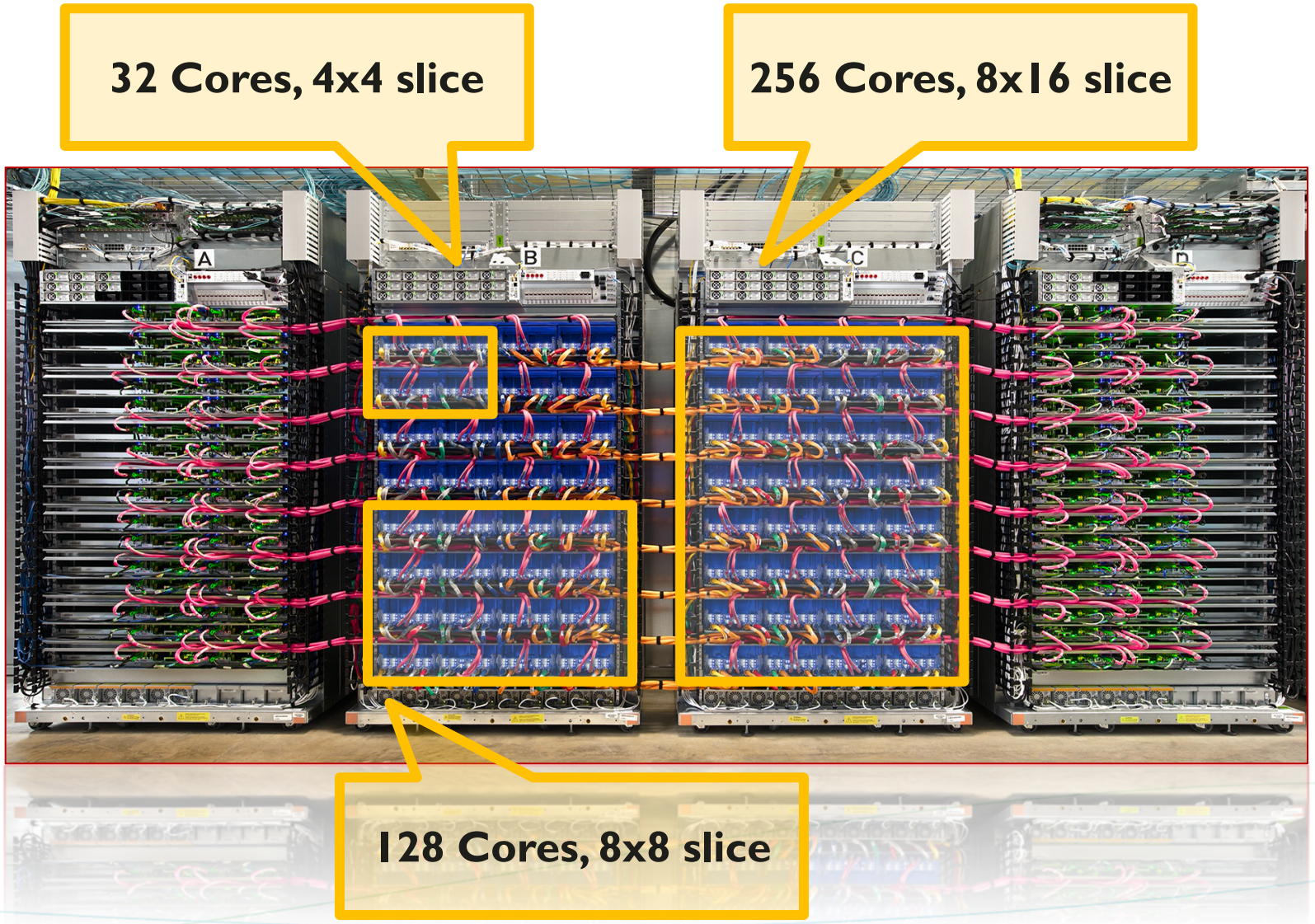
# TPU v2 POD 计算

1. **A 和 D：**CPU 机架，B 和 C 为 TPU v2 机架。
2. **固体箱（蓝色/红色）**：电源管理系统（UPS）& 具体电源接口；
3. **虚线框（绿色）**：机架式网络交换机和机架式交换机顶部；
4. **总体：**每个机柜中有64个CPU板和64个TPU板，共有128个CPU芯片和256个TPU v2芯片。



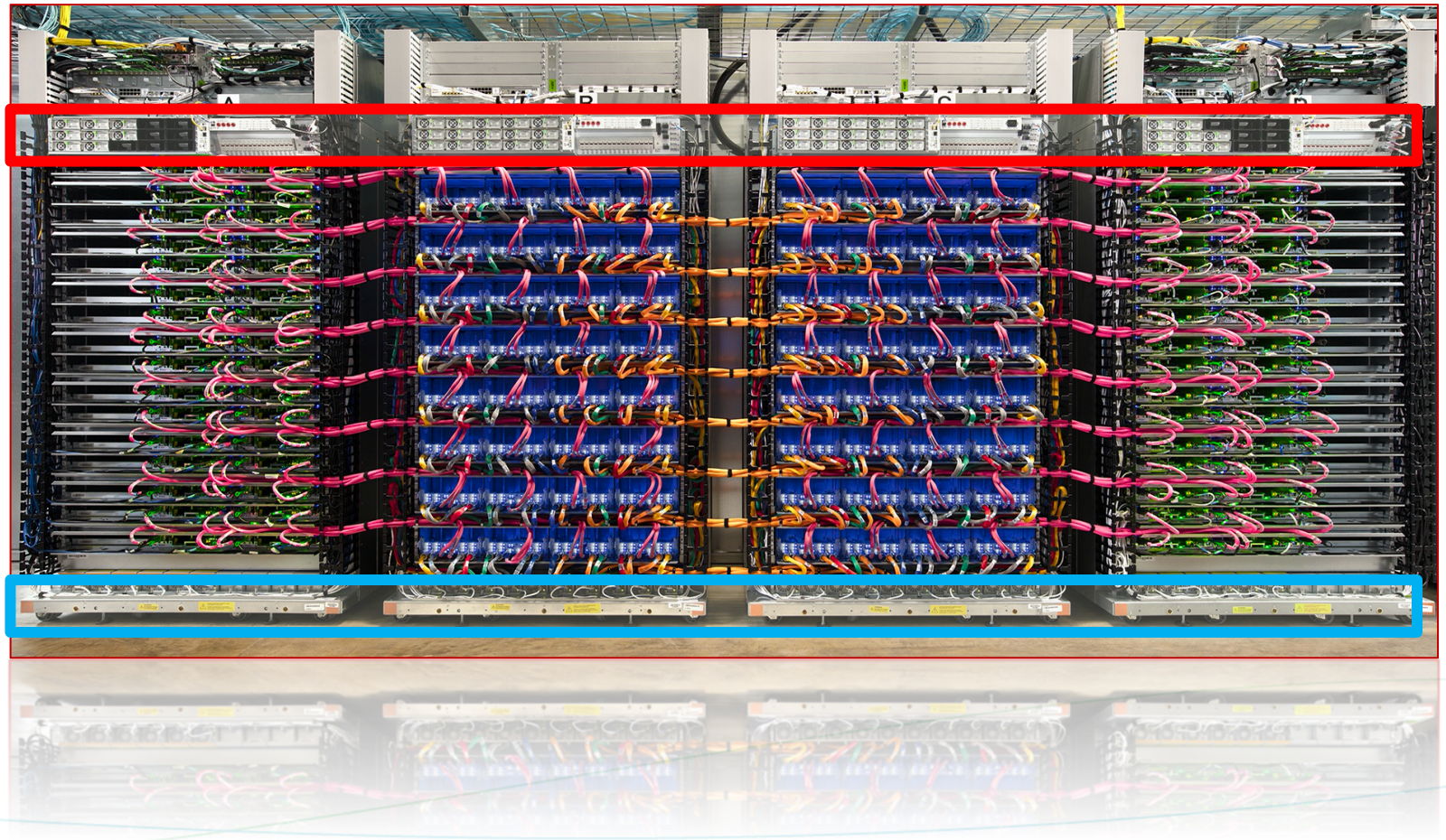
# TPU v2 POD 计算

1. **A 和 D**：CPU 机架，B 和 C 为 TPU v2 机架。
2. **固体箱（蓝色/红色）**：电源管理系统（UPS）& 具体电源接口；
3. **虚线框（绿色）**：机架式网络交换机和机架式交换机顶部；
4. **总体**：每个机柜中有64个CPU板和64个TPU板，共有128个CPU芯片和256个TPU v2芯片。



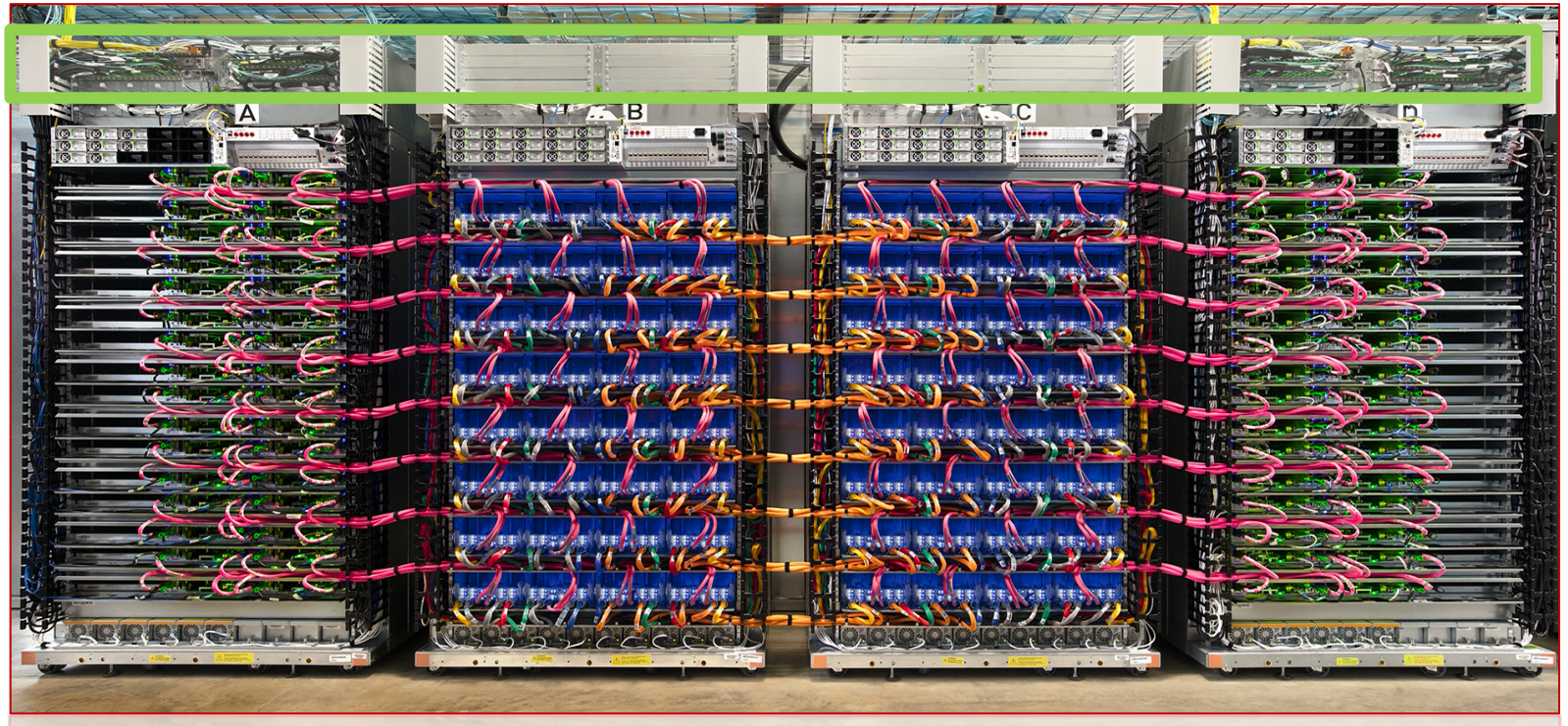
# TPU v2 POD 电源

1. **A 和 D**：CPU 机架，B 和 C 为 TPU v2 机架。
2. **固体箱（蓝色/红色）**：电源管理系统（UPS）& 具体电源接口；
3. **虚线框（绿色）**：机架式网络交换机和机架式交换机顶部；
4. **总体**：每个机柜中有64个CPU板和64个TPU板，共有128个CPU芯片和256个TPU v2芯片。



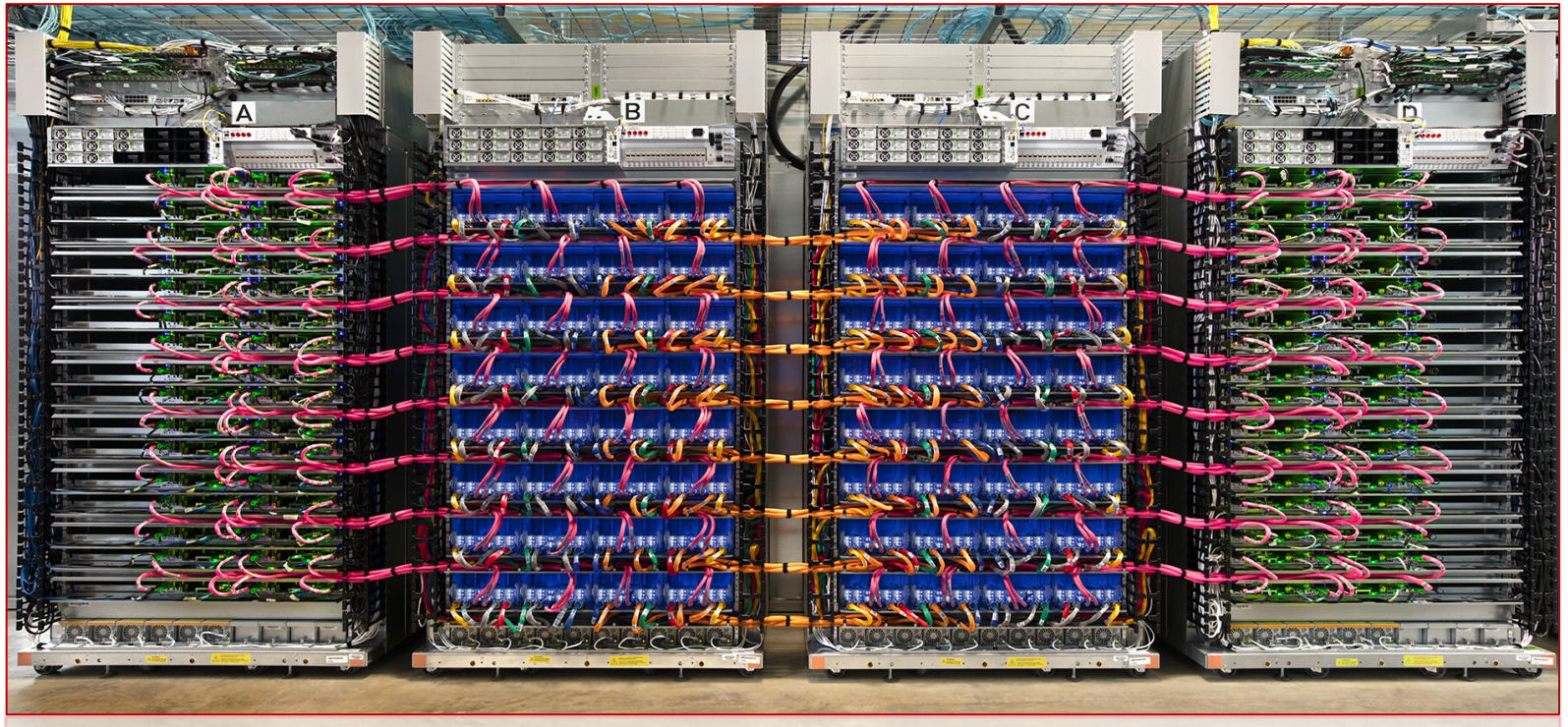
# TPU v2 POD 交换机

1. **A 和 D：**CPU 机架，B 和 C 为 TPU v2 机架。
2. **固体箱（蓝色/红色）**：电源管理系统（UPS）& 具体电源接口；
3. **虚线框（绿色）**：机架式网络交换机和机架式交换机顶部；
4. **总体：**每个机柜中有64个CPU板和64个TPU板，共有128个CPU芯片和256个TPU v2芯片。



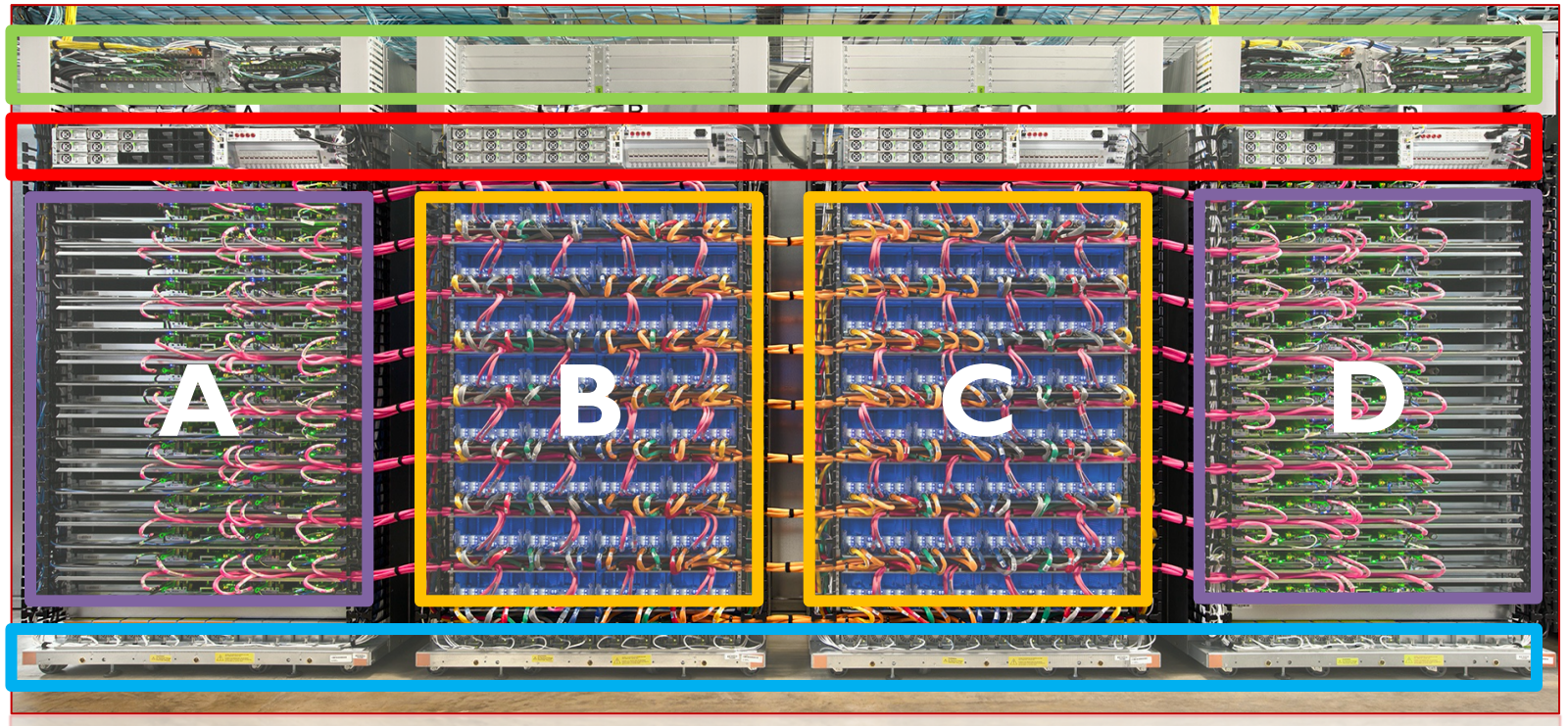
# TPU v2 POD 存储

- 在TPU v2机柜中，看不到任何存储模块。或许这正是下图中机柜上方大量蓝色光纤存在的原因。
- 数据中心网络连接至CPU，同时没有任何光纤连接至机柜B和C的TPU集群，而TPU v2板上也没有任何网络连接。



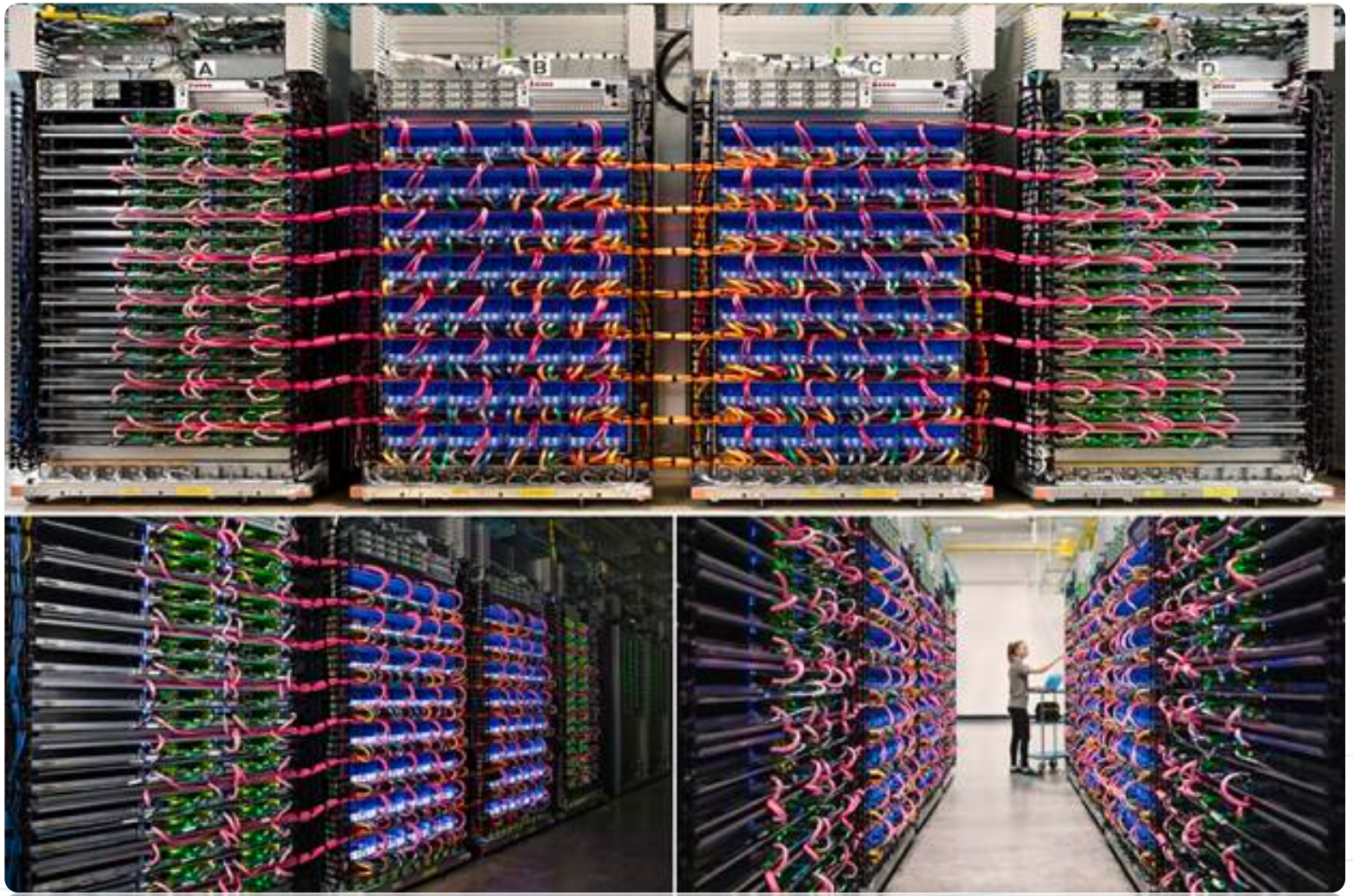
# TPU v2 POD 整体

1. **A 和 D：**CPU 机架，B 和 C 为 TPU v2 机架。
2. **固体箱（蓝色/红色）**：电源管理系统（UPS）& 具体电源接口；
3. **虚线框（绿色）**：机架式网络交换机和机架式交换机顶部；
4. **总体：**每个机柜中有64个CPU板和64个TPU板，共有128个CPU芯片和256个TPU v2芯片。

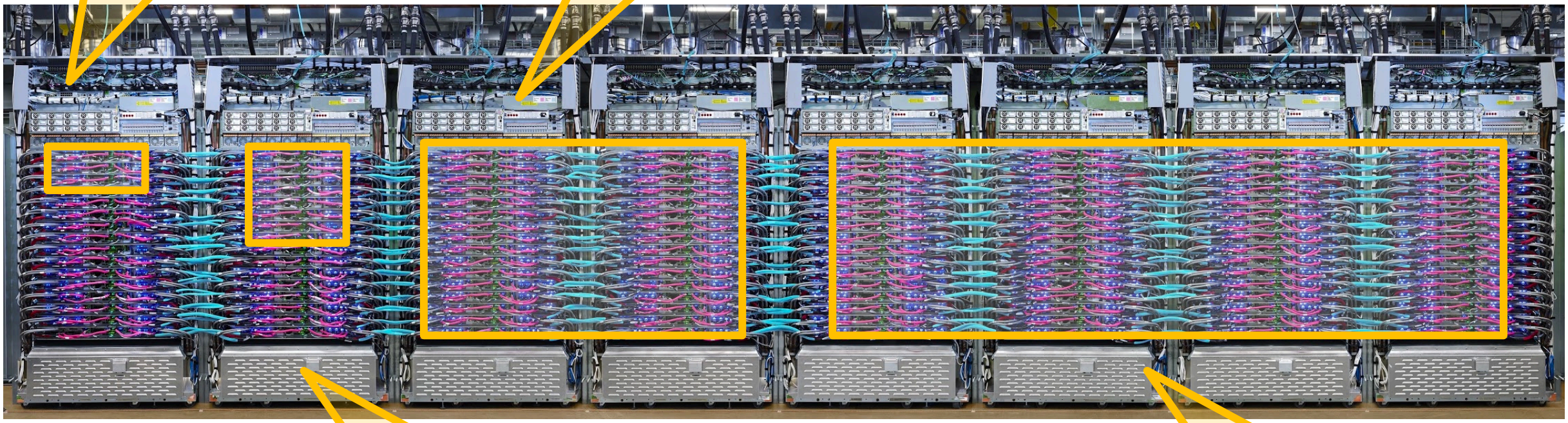


# TPUv2 POD 机柜

- 谷歌展示了 TPUv2 机柜的3张不同照片。在这3张照片中，配置和连线方式一致。
- 为了避免信号衰减带来问题，BlueLink 或 OPA 的铜缆和光纤长度不能太长，因此NPU在中间，CPU在两侧的方式排布。

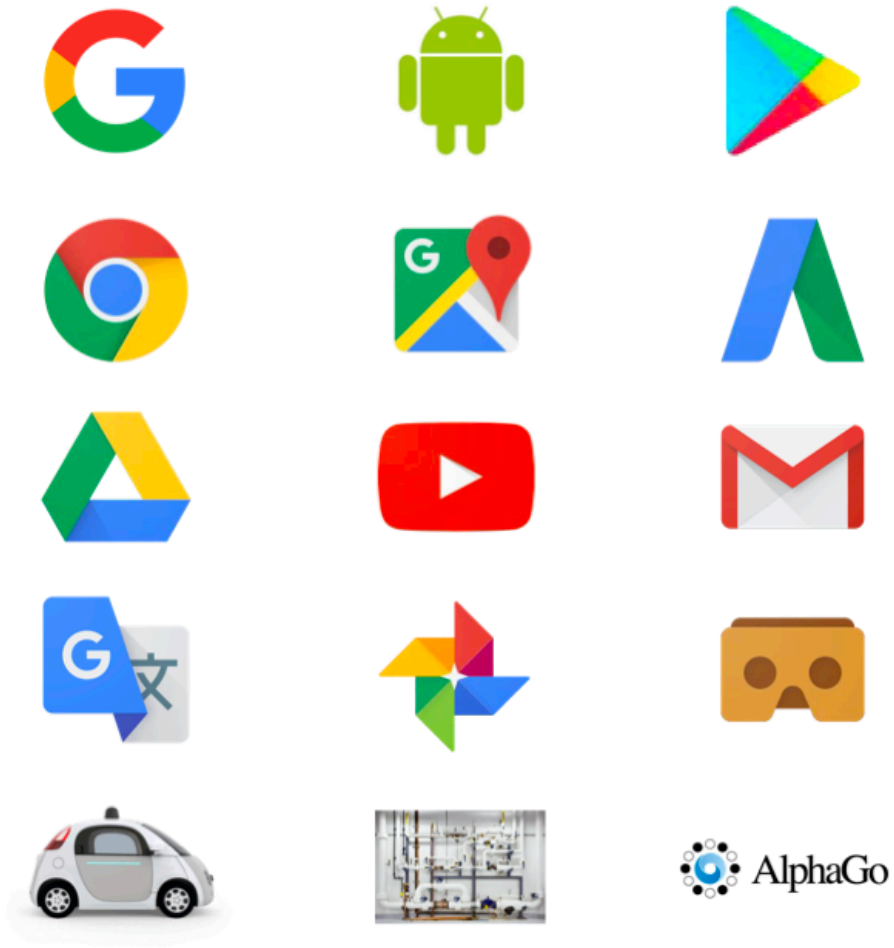


# TPU v2 POD 计算


32 Cores, 4x4 slice
512 Cores, 16x16 slice
128 Cores, 8x8 slice
1024 Cores, 16x32 slice

# Key Takeaways

- TPUv2/v3 supercomputers with 256-1024 chips run production applications at scale, powering many Google products
- TPUs are widely used to accelerate production and research
- Proven results from Model/HW/SW codesign, with more opportunities still available



# 思考

GPU

TPU

1. 英伟达通过 NVLink 迈出 GPU 芯片独立于 CPU，并且将其软件基础架构和工作负载从单一 GPU 扩展到GPU 集群。
2. 机柜中 CPU 和 TPUv2 芯片间如何关联，使得 TPU v2 可以通过超网格中的连接有效地共享数据？
3. TPU v2 芯片可以通过 OPA 上使用远程直接存储器访问（RDMA）从 host 测的内存中加载数据吗？还是要通过CPU 进行调度分发？
4. Google 的 TPU v2 POD 没有纰漏足够的信息能将其与同时代的英伟达 Volta 等商用产品进行比较，POD形态除了硬件指标参数，用什么基准衡量（算力利用率 or MLPerf）？

# 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