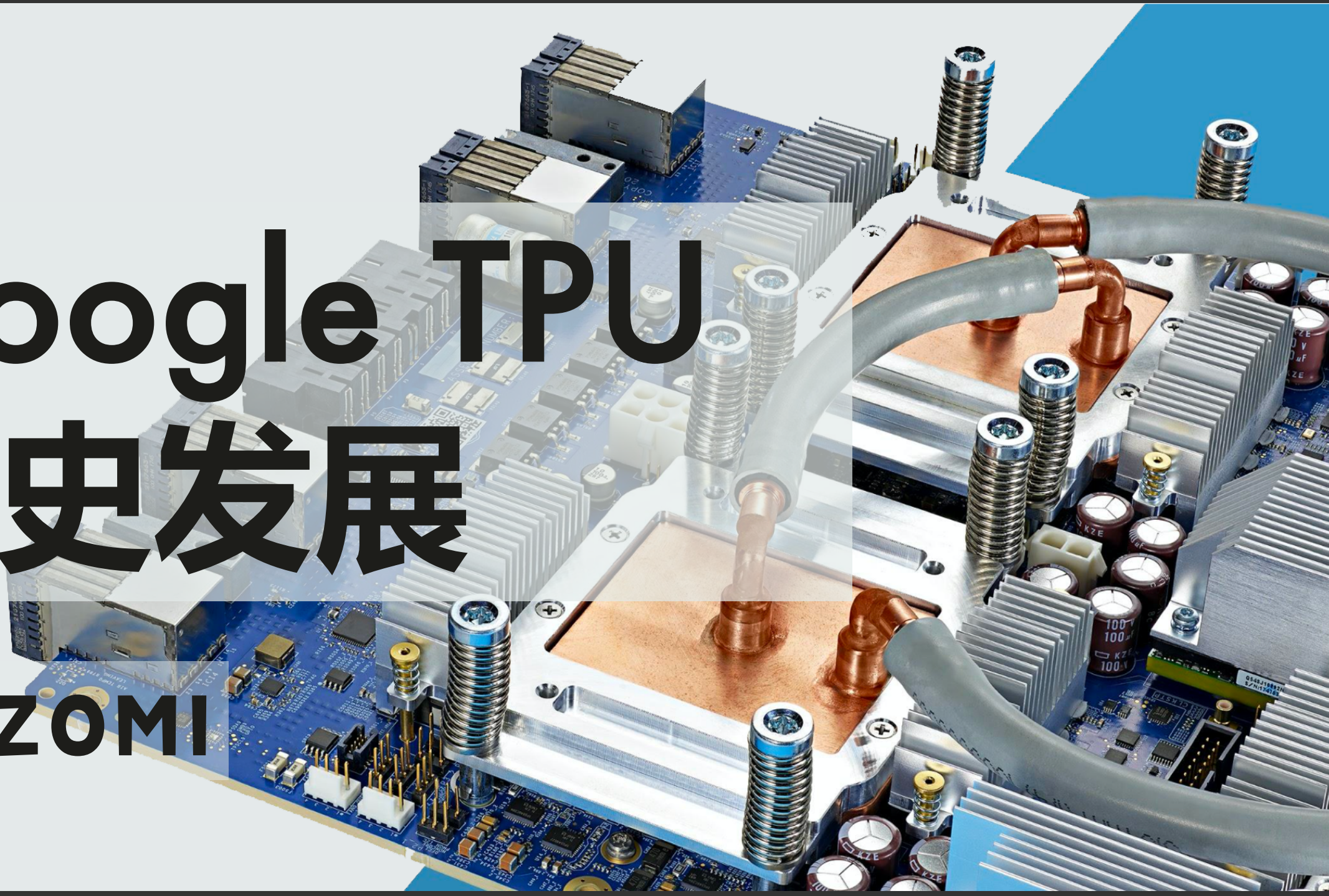
Google TPU 历史发展



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
   - 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

1. **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. 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
- AI 芯片架构的思考

Talk Overview

I. 国外 AI 芯片

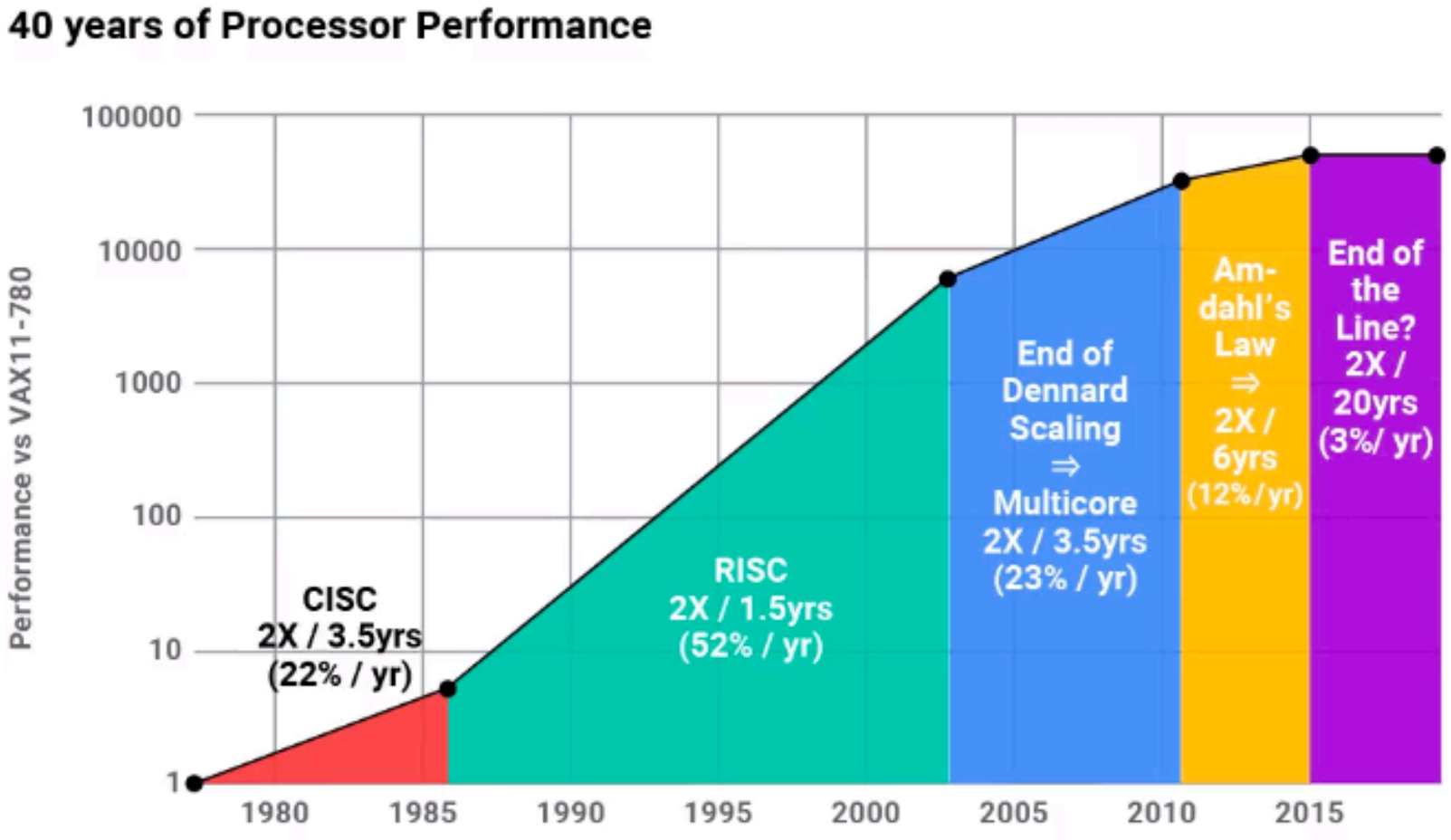
- 英伟达 GPU 芯片架构剖析
- 特斯拉 DOJO 芯片架构剖析
- 谷歌 TPU 芯片架构剖析
  - TPU 历史发展
  - TPU1 脉动阵列细节
  - TPU2 第一款训练卡
  - TPU3 性能 POD 超算
  - TPU4 超级互联
- AI 芯片架构的思考

1. TPU的诞生

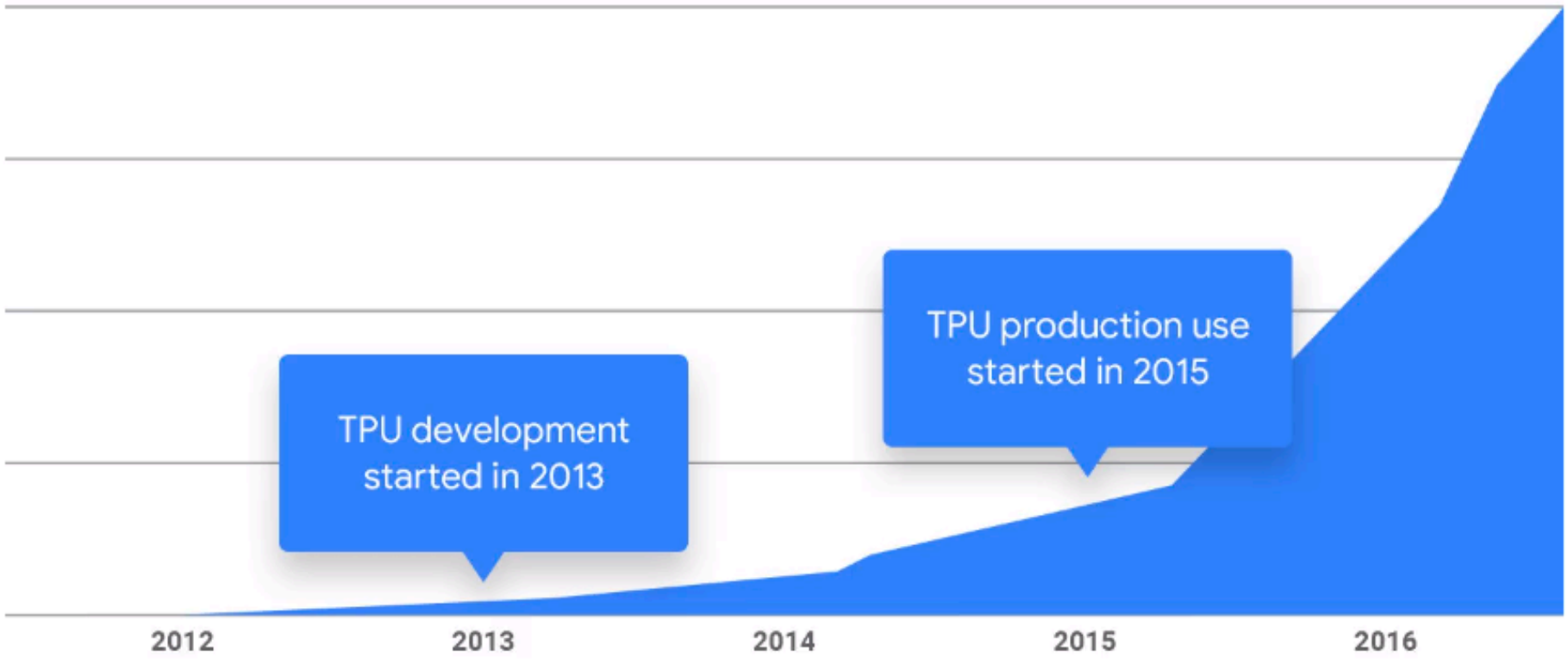
算力消耗越来越大

- 2013年，Google AI 负责人 Jeff Dean 经过计算后发现，如果有 1 亿安卓用户每天使用手机语音转文字服务 3 分钟，消耗的算力就已是 Google 所有数据中心总算力的两倍，何况全球安卓用户远不止 1 亿。

处理器的性能提升


40 years of Processor Performance
Performance vs VAX11-780
100000
10000
1000
100
10
1
1980
1985
1990
1995
2000
2005
2010
2015
CISC
2X / 3.5yrs
(22% / yr)
RISC
2X / 1.5yrs
(52% / yr)
End of
Dennard
Scaling
⇒
Multicore
2X / 3.5yrs
(23% / yr)
Am-
dahl's
Law
⇒
2X /
6yrs
(12%/yr)
End of
the
Line?
2X /
20yrs
(3%/ yr)

Google 搜索中 AI 算力的增长



2. 历代TPU

参数与产品

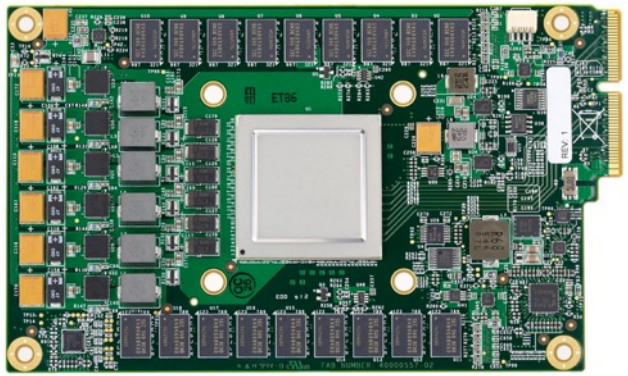
TPU历代芯片

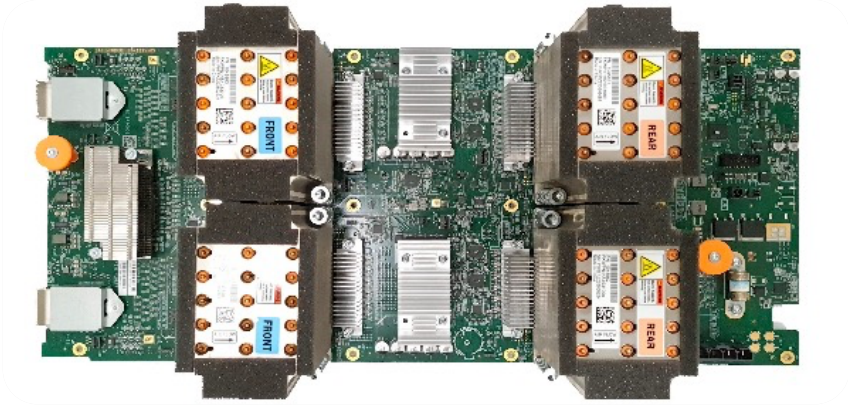
	TPUv1	TPUv2	TPUv3	Edge v1	Pixel Neural Core	TPUv4i	TPUv4	Google Tensor
Date introduced	2016	2017	2018	2018	2019	2020	2021	2021
Process node	28 nm	16 nm	16 nm			7nm	7 nm	
Die size (mm^2)	330mm	625mm	700mm			400mm	780mm	
On-chip memory (MB)	28MB	32MB	32MB			144MB	288MB	
Clock speed (MHz)	700MHz	700MHz	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	

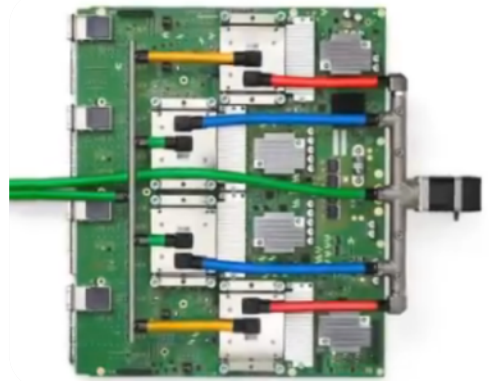
TPU历代产品

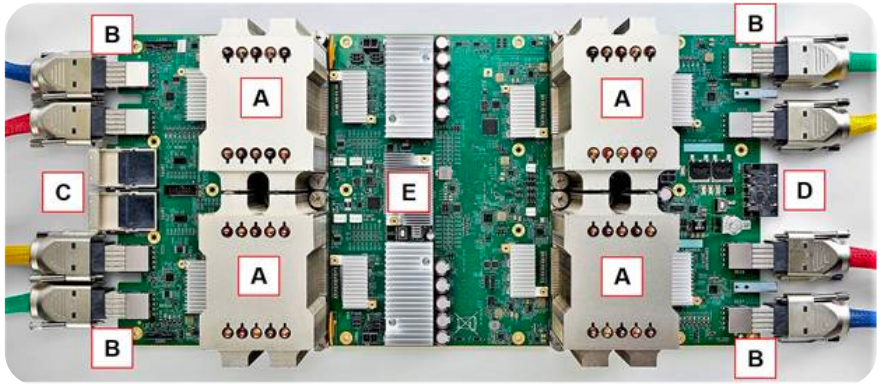
名称	时间	性能	应用
TPUv1	2016年	92Tops + 8GB DDR3	数据中心推理
TPUv2	2017年	180TFlops(集成4块芯片) + 64GB(HBM)	数据中心训练和推理
TPUv3	2018年	420TFlops + 128GB(HBM)	数据中心训练和推理
Edge TPU	2018年	可处理高吞吐量的流式数据	IoT 设备
TPUv2 pod	2019年	11.5千万亿次浮点运算/s，4TB（HBM），二维环面网状网络	数据中心训练和推理
TPUv3 pod	2019年	>100千万亿次浮点运算/s，32TB（HBM），二维环面网状网络	数据中心训练和推理
TPUv4	2021年		数据中心训练和推理
TPUv4 pod	2022年		数据中心训练和推理

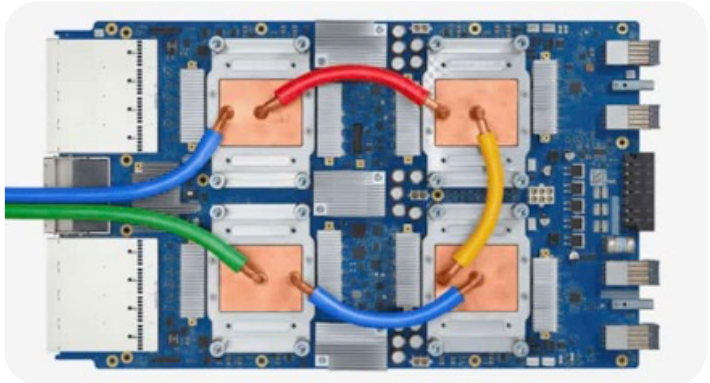
TPU历代芯片



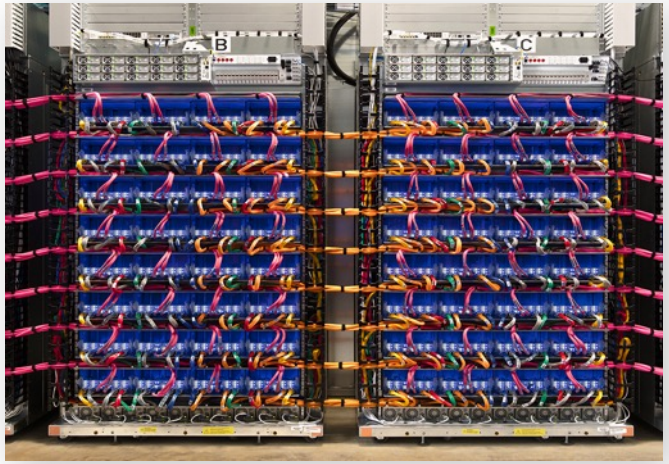


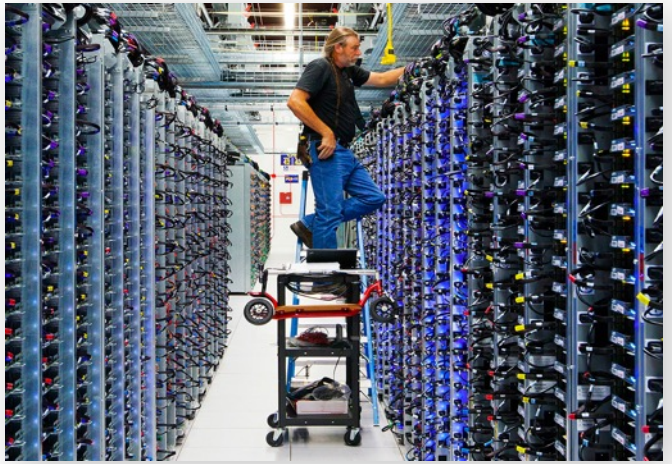


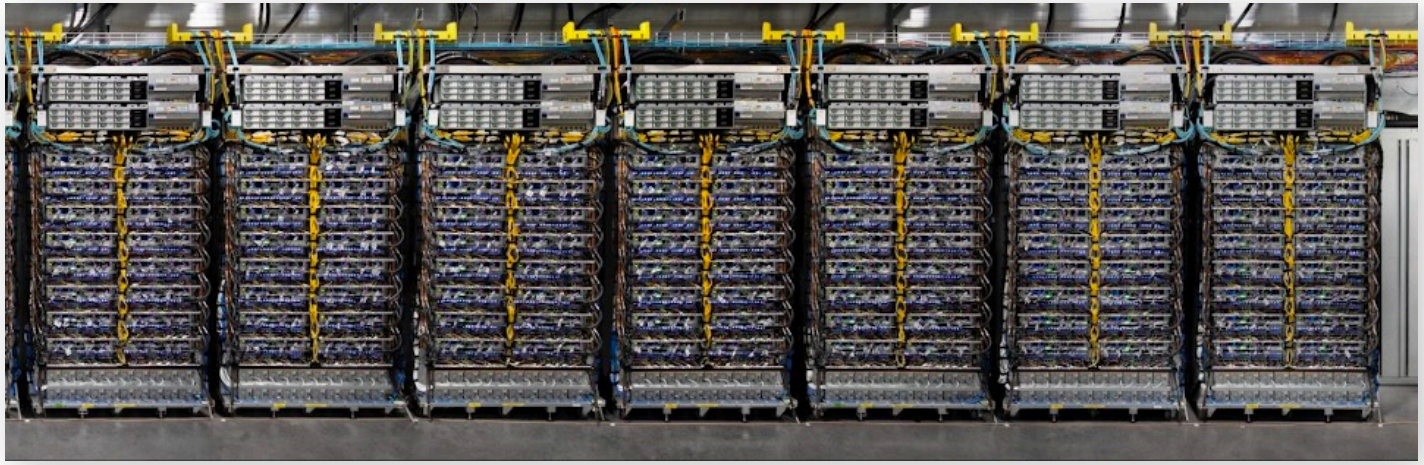




TPU历代芯片服务器







TPU历代芯片产品

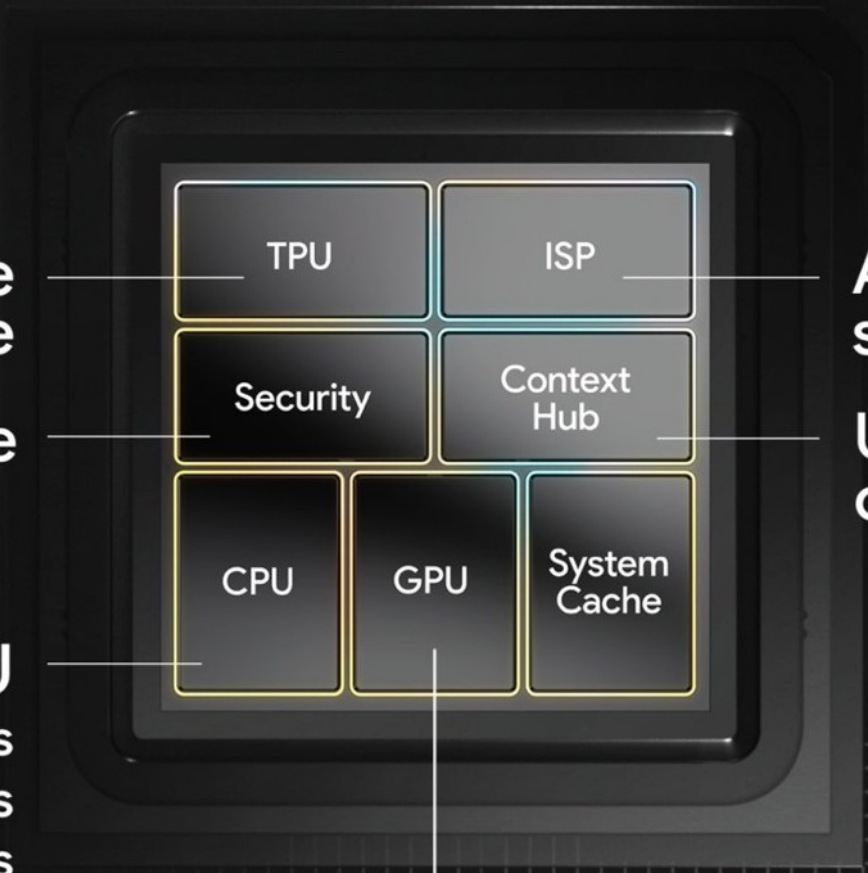
Pixel 5 5G

Google Tensor

Google's machine learning engine

Tensor security core

Powerful CPU
2 high-performance cores
2 mid cores
4 high-efficiency cores



Advanced image signal processor

Ultra-low power context engine

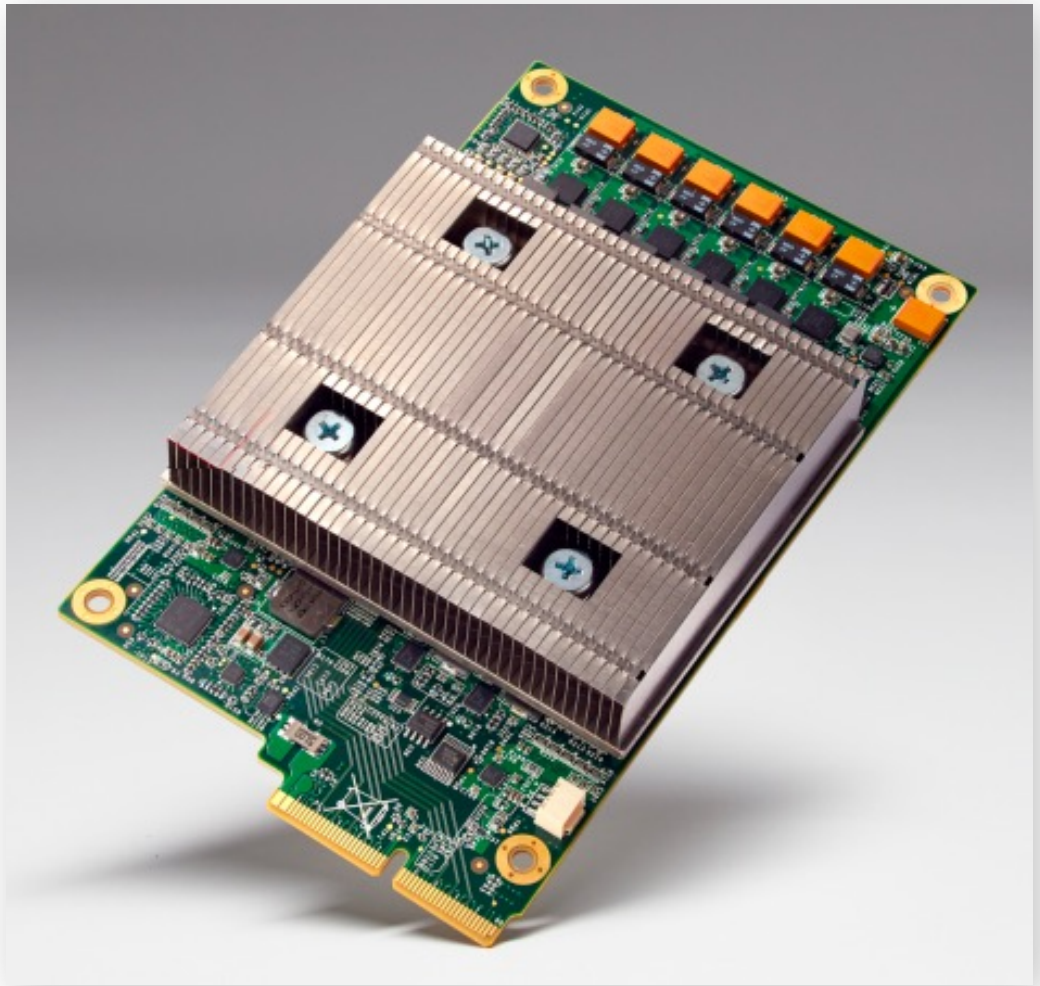
20-core GPU

基于Google Tensor 的 pixel 系列

3. TPU的演进

TPU v1 概览：Deterministic Execution Model

- TPU1 ASIC 采用 28nm 工艺制造；
- 主频700MHz，功耗40W；
- 为了尽快把 TPU 部署到现有服务器中，Google 选择把 TPU1 做成外部扩展加速器，通过 PCIe Gen3 x16 总线与 CPU 主机相连，提供 12.5GB/s 有效带宽；



TPU v1 性能：推理场景受限

受制于时代限制，TPU v1 主要针对2015年最火神经网络进行优化，主要可以分为以下三类：

- MLP 多层感知机
- CNN 卷积神经网络
- RNN & LSTM 递归神经网络 & 长短期记忆

TPU v1 性能：推理性能同期产品炸天

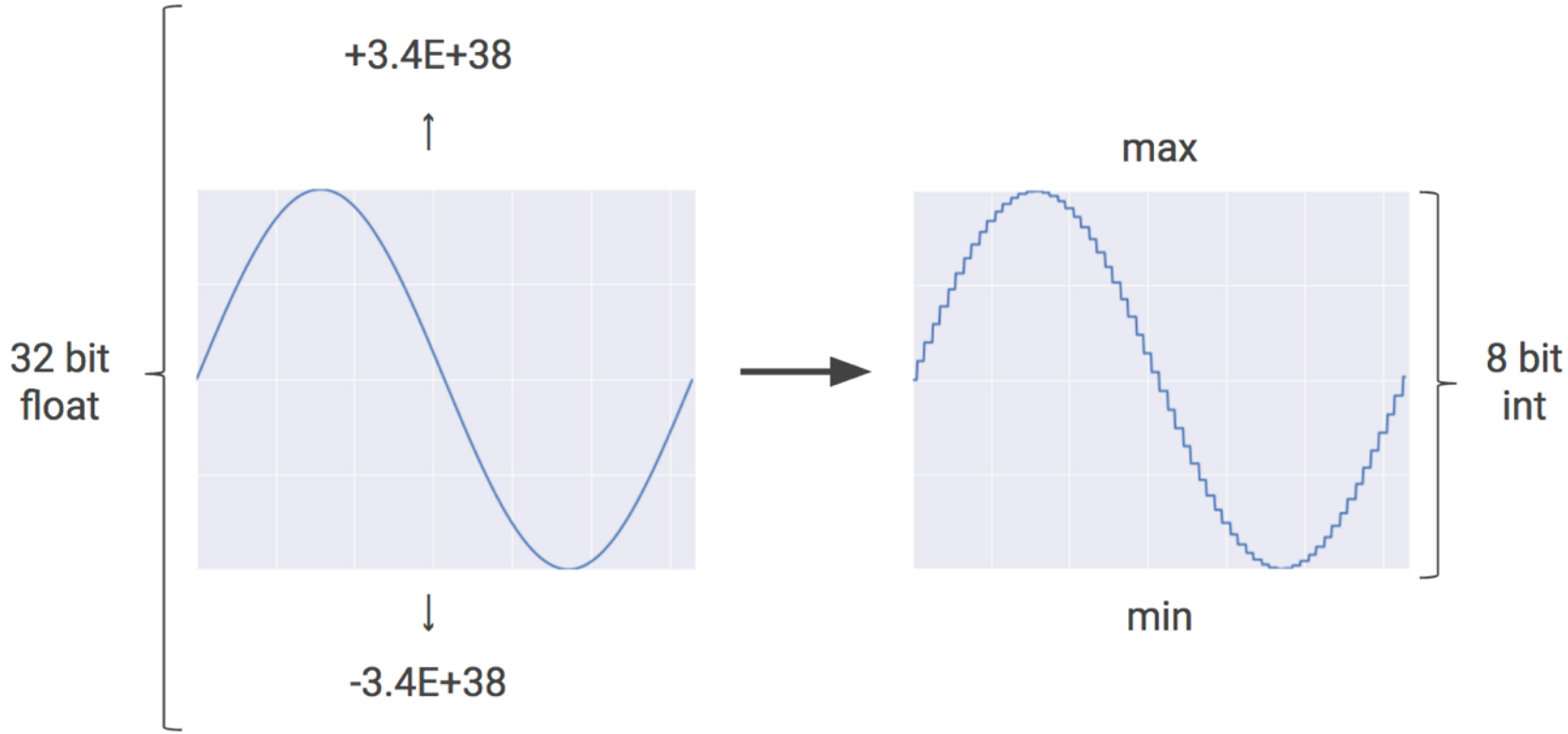
- 深度学习专用的 DSA （ Domain Specific Architecture ）架构硬件 —— 脉动阵列。适合用于非常规整简单的运算，但正巧矩阵乘和卷积就是这种规整又简单的运算，来自两个方向的数据以一定的间隔到达阵列中的 MXU，并在那里进行运算。不过，MXU 中的权值阵列专门为矩阵乘法运算进行了优化，并不适用于通用的逻辑计算。

TPU v1 性能：推理性能同期产品炸天

- Reduced Precision 低精度的数据格式
- Matrix Processor 矩阵乘加专用处理器
- Minimal and deterministic design 专用硬件

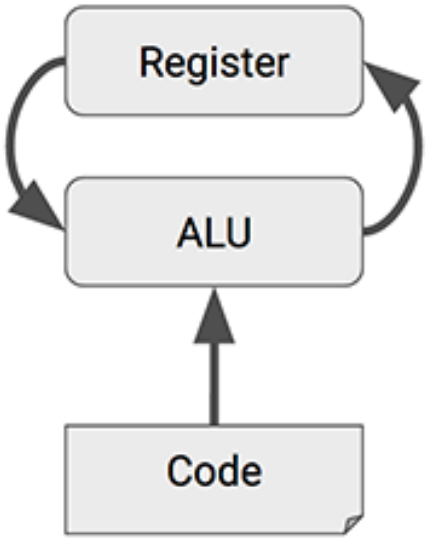
特性1：低精度

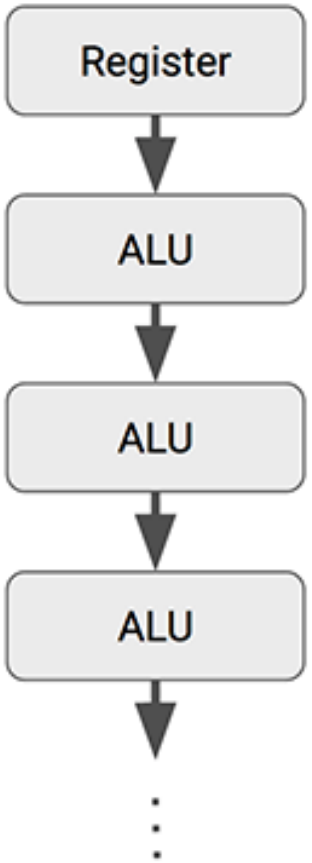
- 神经网络推理不需要FP32/FP16计算精度，通过量化压缩，可以用 Int8 对神经网络进行预测，并保持适当的准确度。这称为量化，使用 Int8 来近似预设最小值和最大值之间任意数值的优化技术。



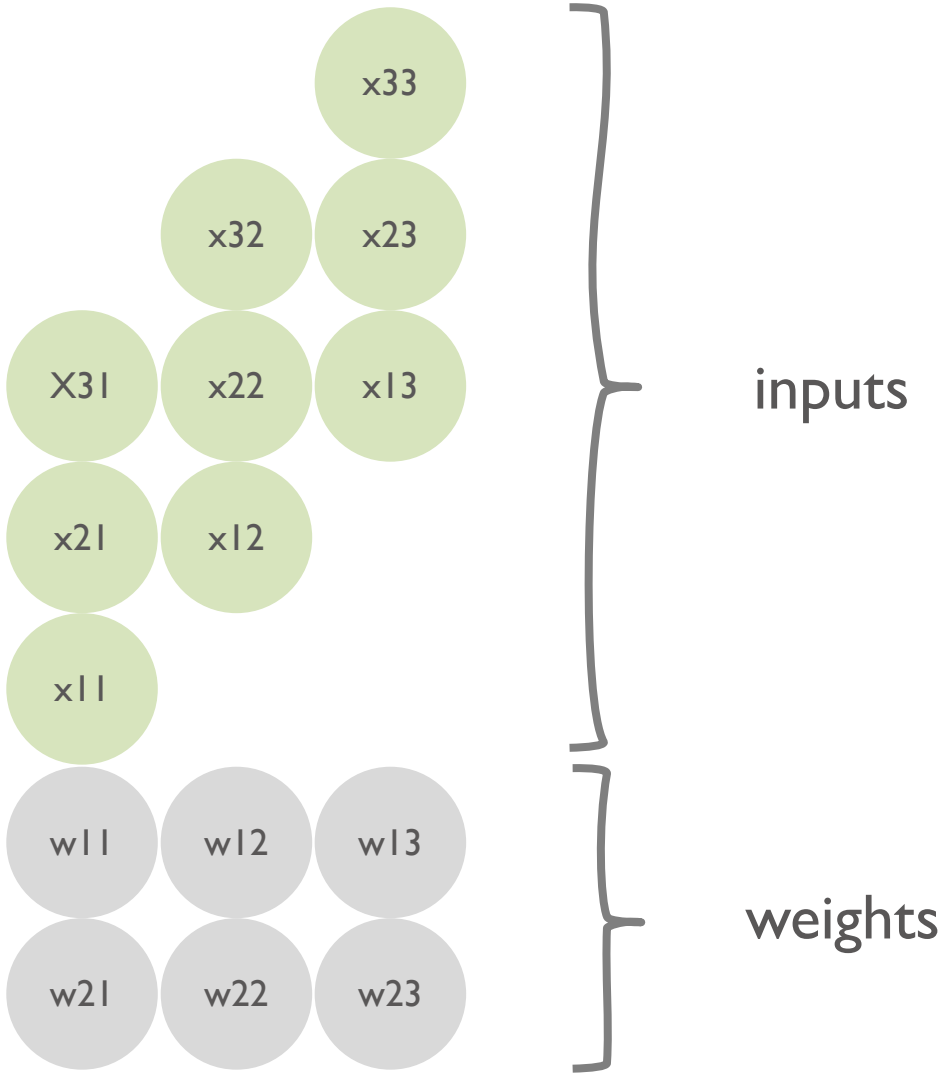
特性1：脉动阵列

- CPU 和 GPU 在每次运算中都需要从多个寄存器（register）中进行存取；
- TPU 脉动阵列将多个运算逻辑单元（ALU）串联在一起，复用从一个寄存器中读取的结果。



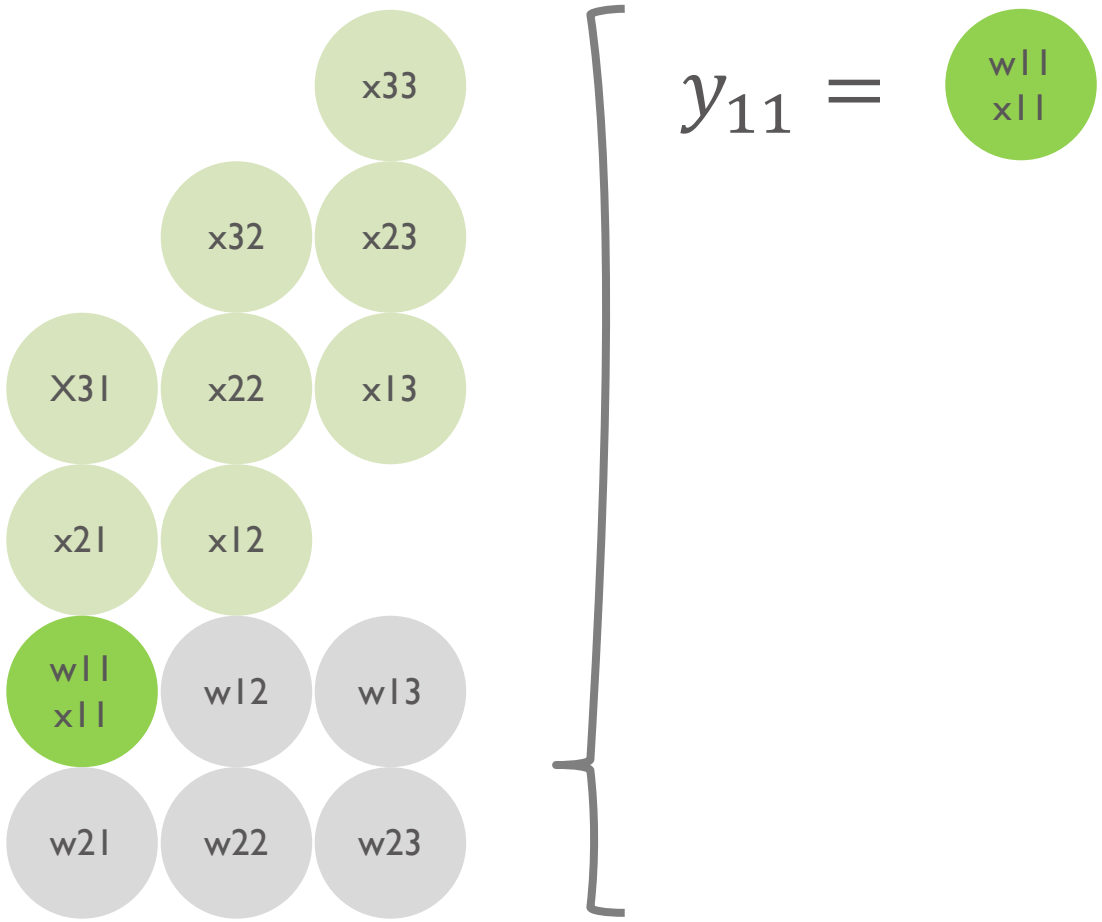


The Core of TPU: Systolic Array



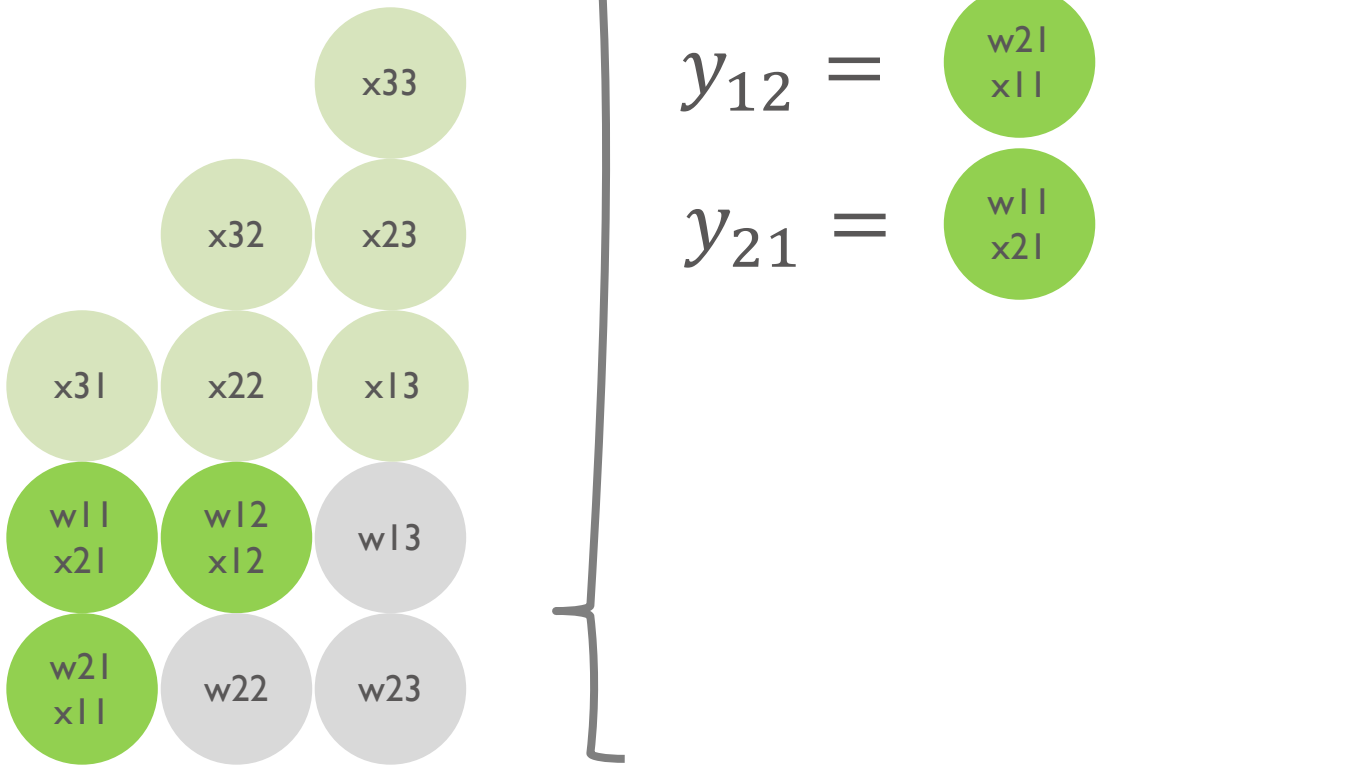
The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access



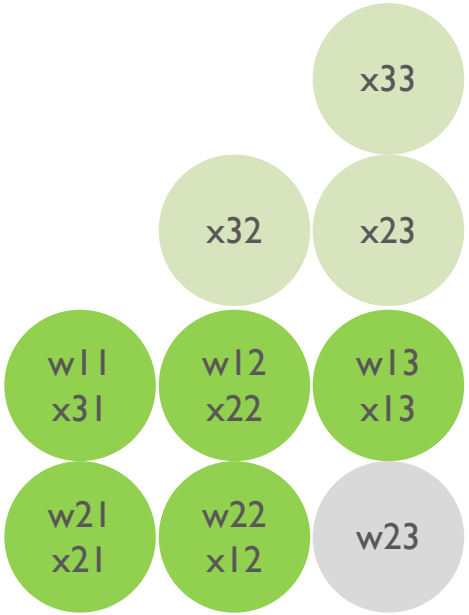
The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access



The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access



$$
\begin{cases}
y_{11} = w11\,x11 + w12\,x12 + w13\,x13 \\
y_{12} = w21\,x11 + w22\,x12 \\
y_{21} = w11\,x21 + w12\,x22 \\
y_{22} = w21\,x21 \\
y_{31} = w11\,x31
\end{cases}
$$

The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access

$$
\begin{cases}
y_{11} = & w11\ x11 + w12\ x12 + w13\ x13 \\
y_{12} = & w21\ x11 + w22\ x12 + w23\ x13 \\
y_{21} = & w11\ x21 + w12\ x22 + w13\ x23 \\
y_{22} = & w21\ x21 + w22\ x22 \\
y_{31} = & w11\ x31 + w12\ x32 \\
y_{32} = & w21\ x31
\end{cases}
$$

The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access

$$
\begin{cases}
y_{11} = w11\,x11 + w12\,x12 + w13\,x13 \\
y_{12} = w21\,x11 + w22\,x12 + w23\,x13 \\
y_{21} = w11\,x21 + w12\,x22 + w13\,x23 \\
y_{22} = w21\,x21 + w22\,x22 + w23\,x23 \\
y_{31} = w11\,x31 + w12\,x32 + w13\,x33 \\
y_{32} = w21\,x31 + w22\,x32
\end{cases}
$$

The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access

$$
\begin{cases}
y_{11} = w11\,x11 + w12\,x12 + w13\,x13 \\
y_{12} = w21\,x11 + w22\,x12 + w23\,x13 \\
y_{21} = w11\,x21 + w12\,x22 + w13\,x23 \\
y_{22} = w21\,x21 + w22\,x22 + w23\,x23 \\
y_{31} = w11\,x31 + w12\,x32 + w13\,x33 \\
y_{32} = w21\,x31 + w22\,x32 + w23\,x33
\end{cases}
$$

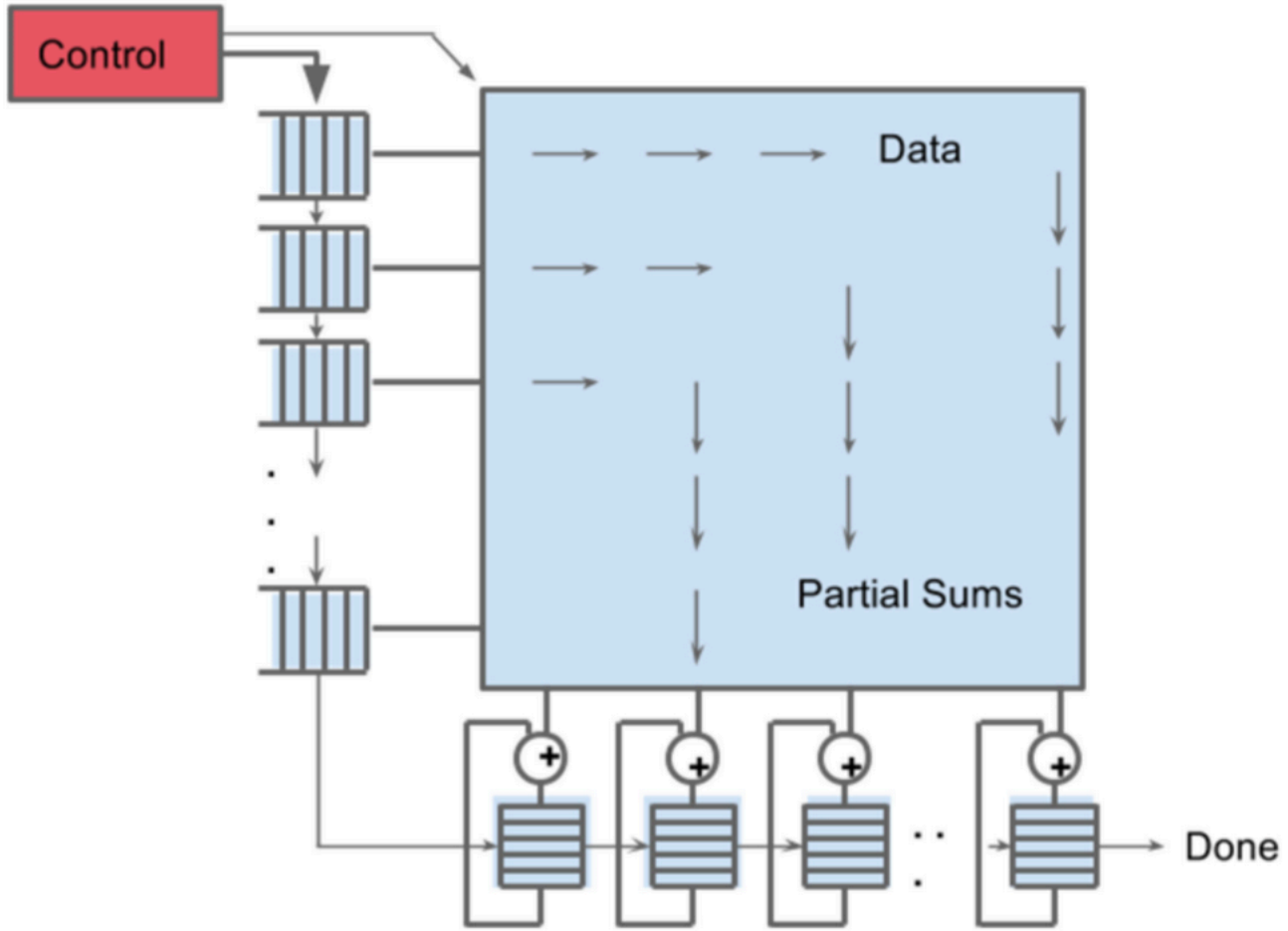
The Core of TPU: Systolic Array

- Large hard-wired matrix calculation without memory access

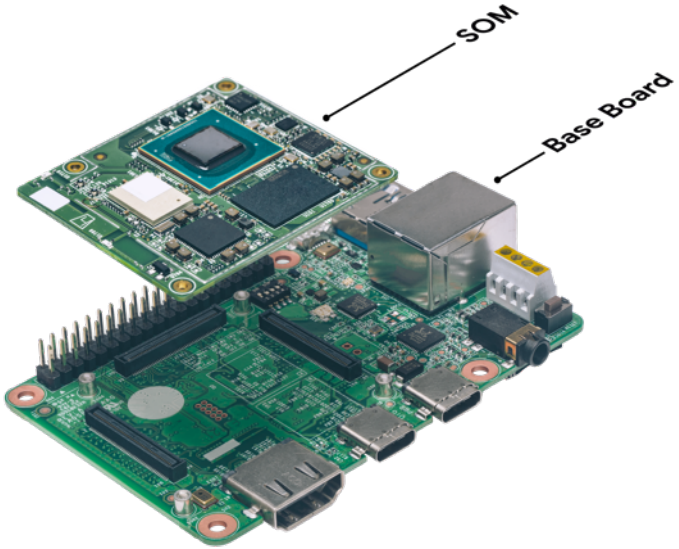
$$
\begin{cases}
y_{11} = w11\,x11 + w12\,x12 + w13\,x13 \\
y_{12} = w21\,x11 + w22\,x12 + w23\,x13 \\
y_{21} = w11\,x21 + w12\,x22 + w13\,x23 \\
y_{22} = w21\,x21 + w22\,x22 + w23\,x23 \\
y_{31} = w11\,x31 + w12\,x32 + w13\,x33 \\
y_{32} = w21\,x31 + w22\,x32 + w23\,x33
\end{cases}
$$

Matrix Multiply Unit(MXU): a BIG systolic array

- Up to 256K ops / cycle
- Up to 256M ops / instruction



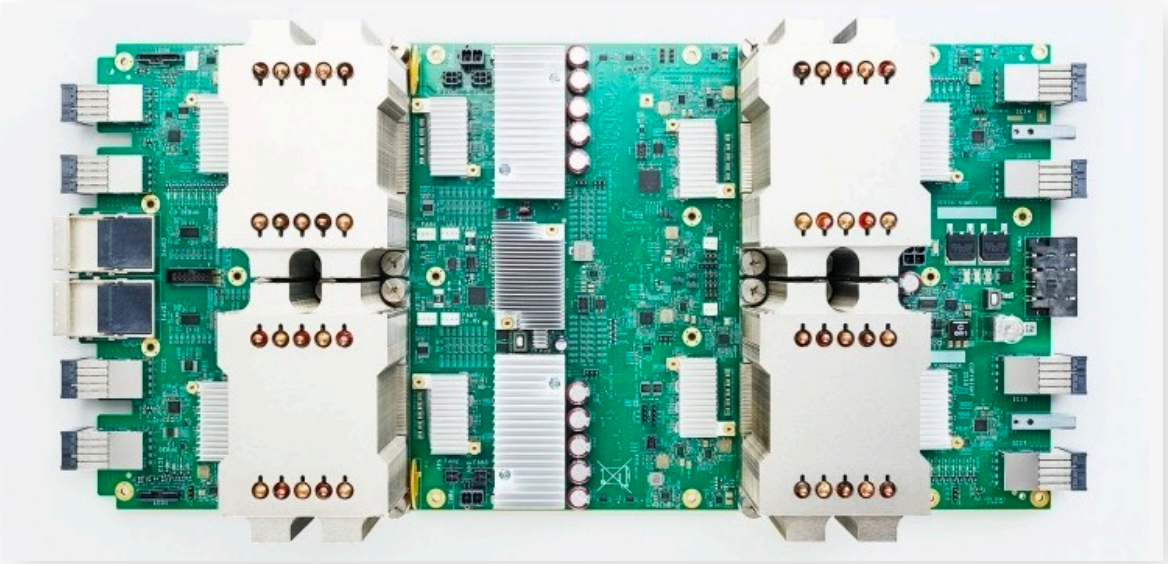
Google Application

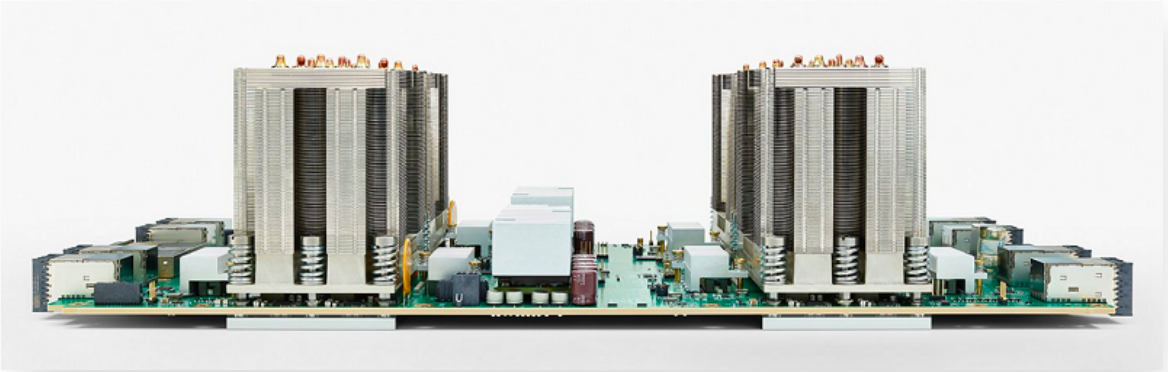


4. TPU2

TPU v2

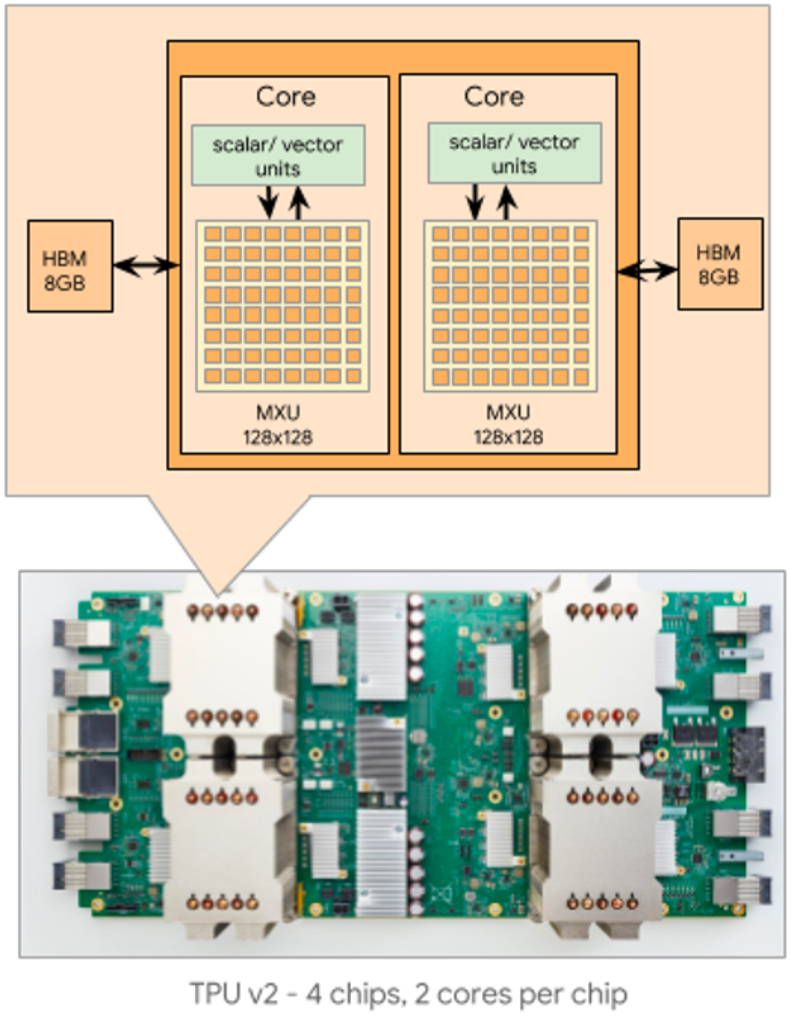
- TPU v2 于 2017 年 5 月发布。使用 16 GB 高带宽内存 HBM，可将带宽提升到 600 GB/s，峰值算力达到 45 TFLOPS。
- TPUv2 被排列成性能为 180 TFLOPS 的四芯片模块，并将其中 64 个模块组装成 256 芯片的 Pod 超算，峰值算力达到 11.5 PFLOPS。
- 从第二代 TPUv2 还可以进行浮点运算，引入 BF16。





TPU v2 芯片架构拓扑

- 只要将 TPUv1 架构设计稍作更改，就可以得到 TPUv2；
- 而对 TPUv2 架构的修改，就是训练与推理的差别。



TPU v2 - 4 chips, 2 cores per chip

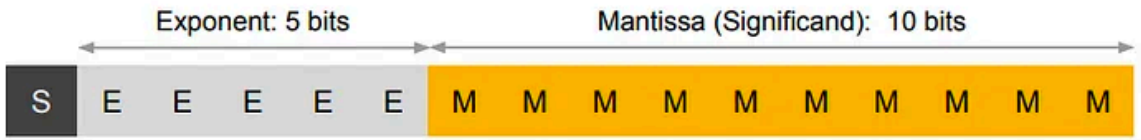
BFLoat16

float32: Single-precision IEEE Floating Point Format

Range: ~1e^{-38} to ~3e^{38}

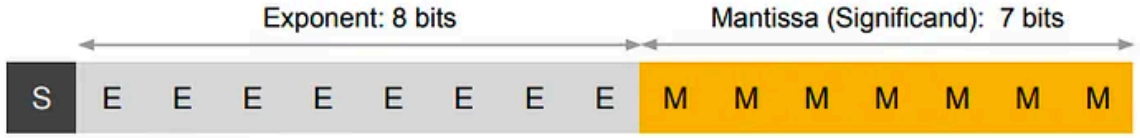
float16: Half-precision IEEE Floating Point Format

Range: ~5.96e^{-8} to 65504



bfloat16: Brain Floating Point Format

Range: ~1e^{-38} to ~3e^{38}



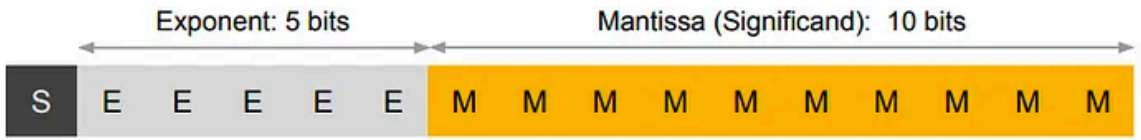
如何计算指数位和小数位？

float32: Single-precision IEEE Floating Point Format

Range: ~1e^{-38} to ~3e^{38}

float16: Half-precision IEEE Floating Point Format

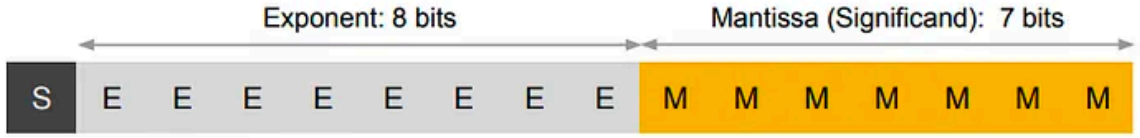
Range: ~5.96e^{-8} to 65504



bfloat1(Format

8 位指数位，7 位小数位

Range: ~1e^{-38} to ~3e^{38}



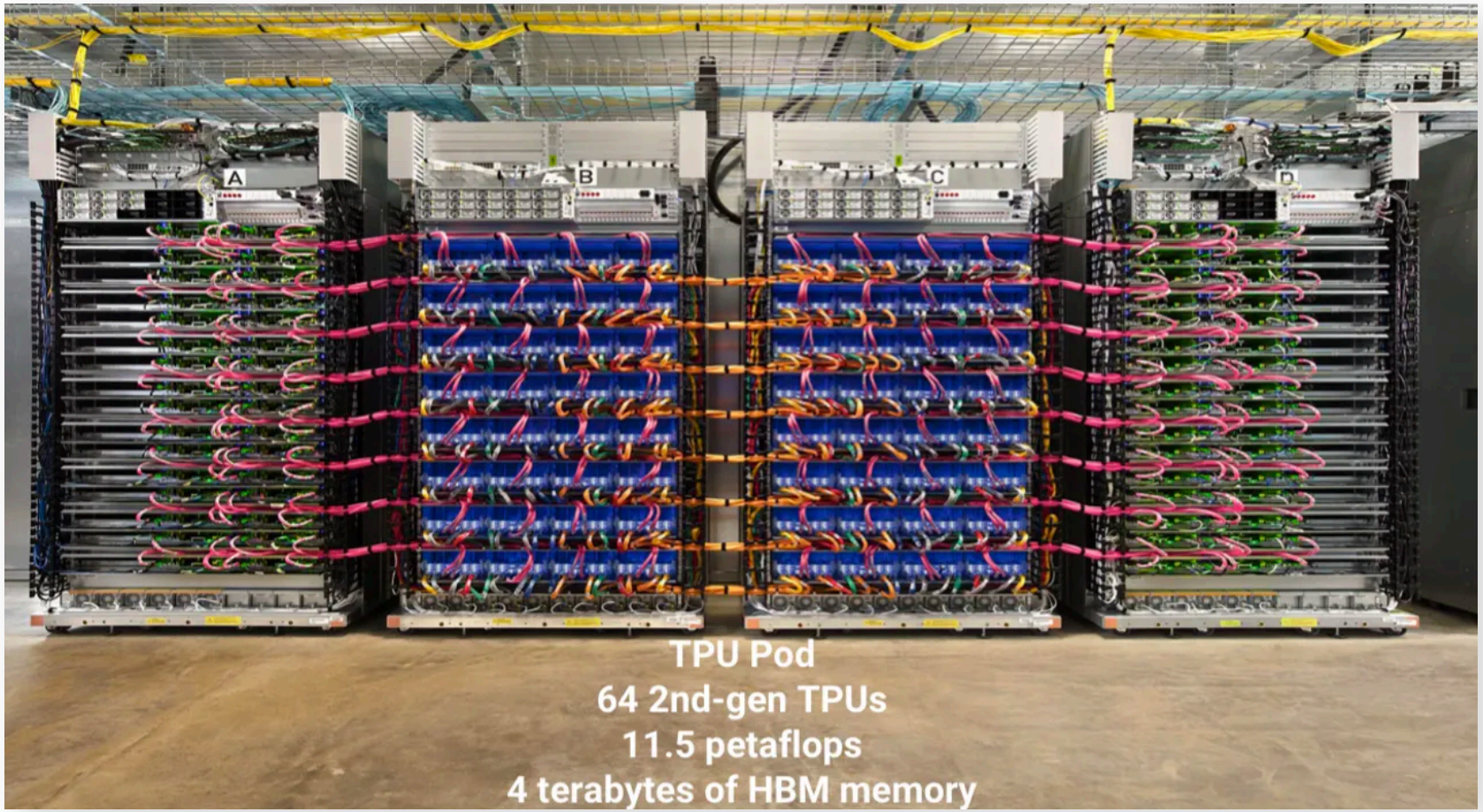
BFLoat16

将值明确类型转换为 bfloat16 有两个原因：

1. 以 bfloat16 格式存储值可节省片上内存，使TPU 能够训练更大的模型或使用更大的批量大小。
2. 某些操作受内存带宽限制，这意味着从内存加载数据所需的时间会减慢执行计算的总体时间。
3. 以 bfloat16 格式存储这些运算的操作数和输出可减少必须传输的数据量，从而提高整体速度。

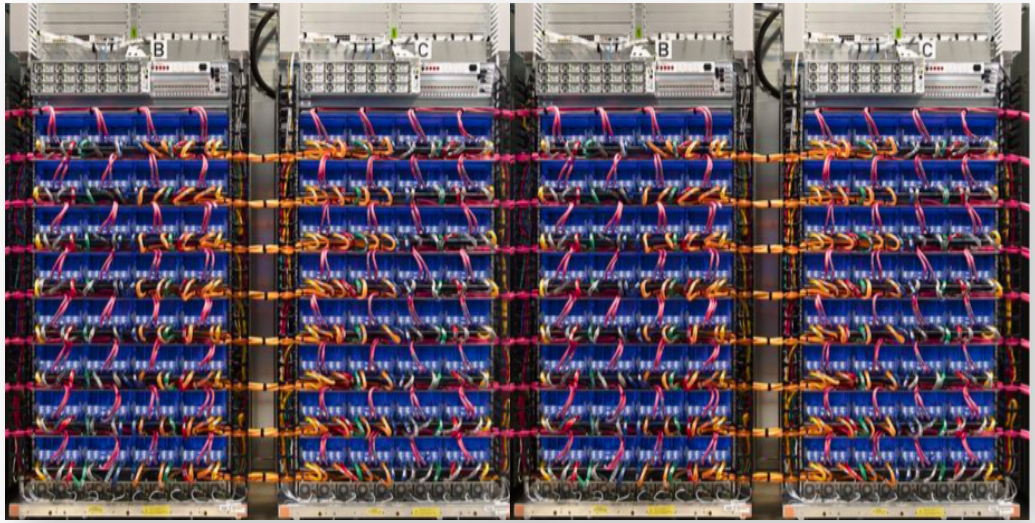
TPU v2 POD

- Google's HPC cluster for ML 11.6 PFLOPS with 64 Cloud TPUs.



TPU v2 POD

TPUv2 supercomputer (256 chips)



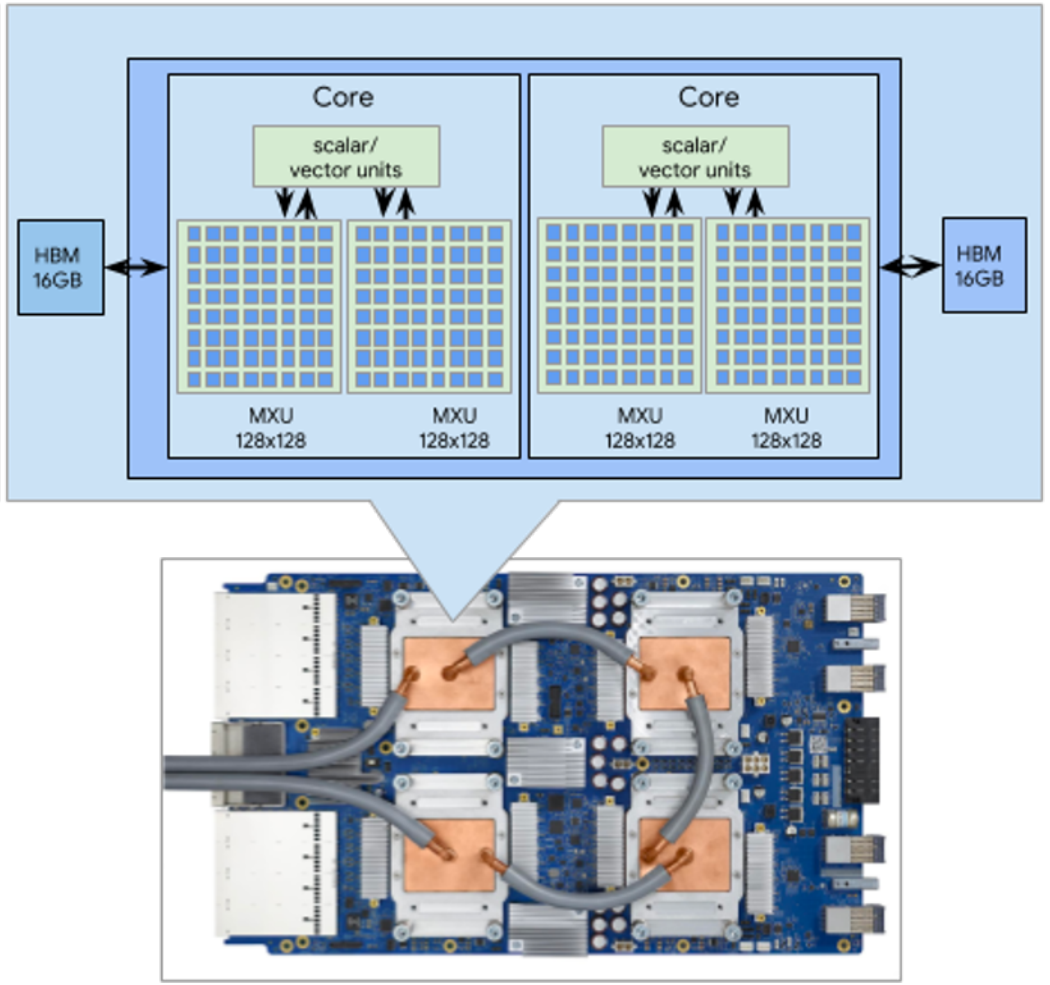
- 11.5 petaflops
- 4 TB HBM
- 2-D torus
- 256 chips

- Real Data: **77,392** images/sec
- Final Accuracy: **93%**
- Training time: **30 min**

5. TPU3

TPU v3

- TPU v3 是对 TPU v2 增量升级，相同工艺下晶体管数量仅增加了 11%，在带宽和计算能力都有 30% 左右的提升，裸片尺寸增加了 12%（16 纳米），芯片性能提高 2.67 倍，HBM 主存储器容量提高 2 倍，能够处理更大的数据集。
- 与 TPU v3 最大的不同在于，互联方式 2D torus 互连从 TPU v2 中 256 个芯片扩展到 TPU v3 中 1,024 个芯片，这让 Pod 超算型号处理能力增加了 10.7 倍，计算理论峰值从12 petaflops 到 126 petaflops（BF16）。



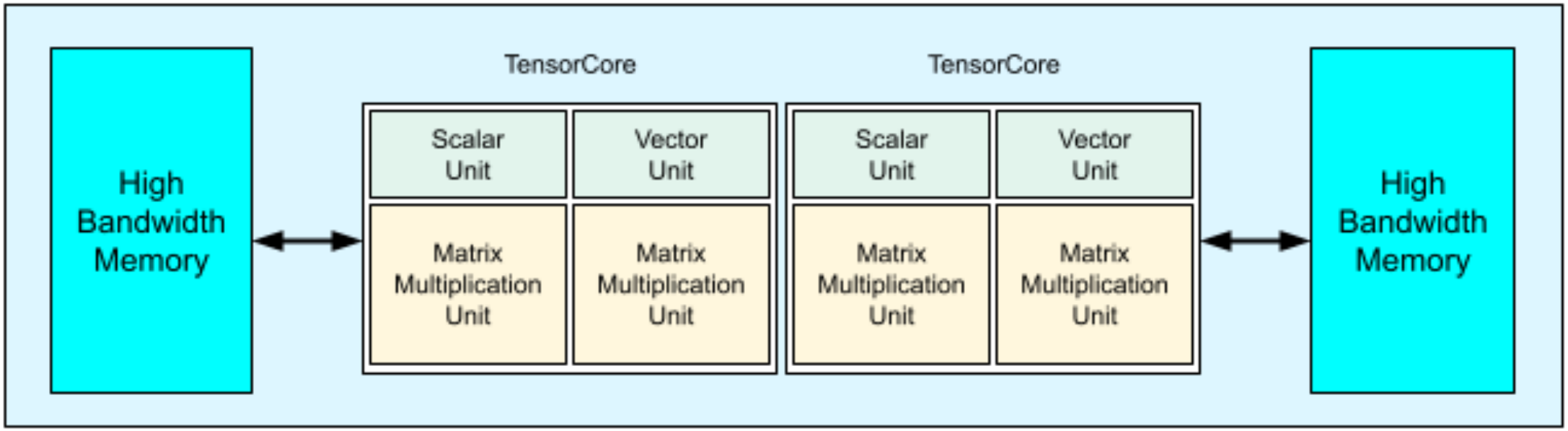
TPU v3 - 4 chips, 2 cores per chip

TPU v3

关键规范	v3 Pod 值
每个芯片的峰值计算次数	123 万亿次浮点数 (bf16)
HBM2 容量和带宽	32 GiB、900 GBps
测量的最小值/平均值/最大值	123/220/262 瓦
TPU Pod 大小	1024 条状标签
互连拓扑	2D 环形
每个 Pod 的峰值计算次数	126 拍拍 (bf16)
每个 Pod 的全宽带宽	340 TB/秒
每个 Pod 的对分带宽	6.4 TB/秒

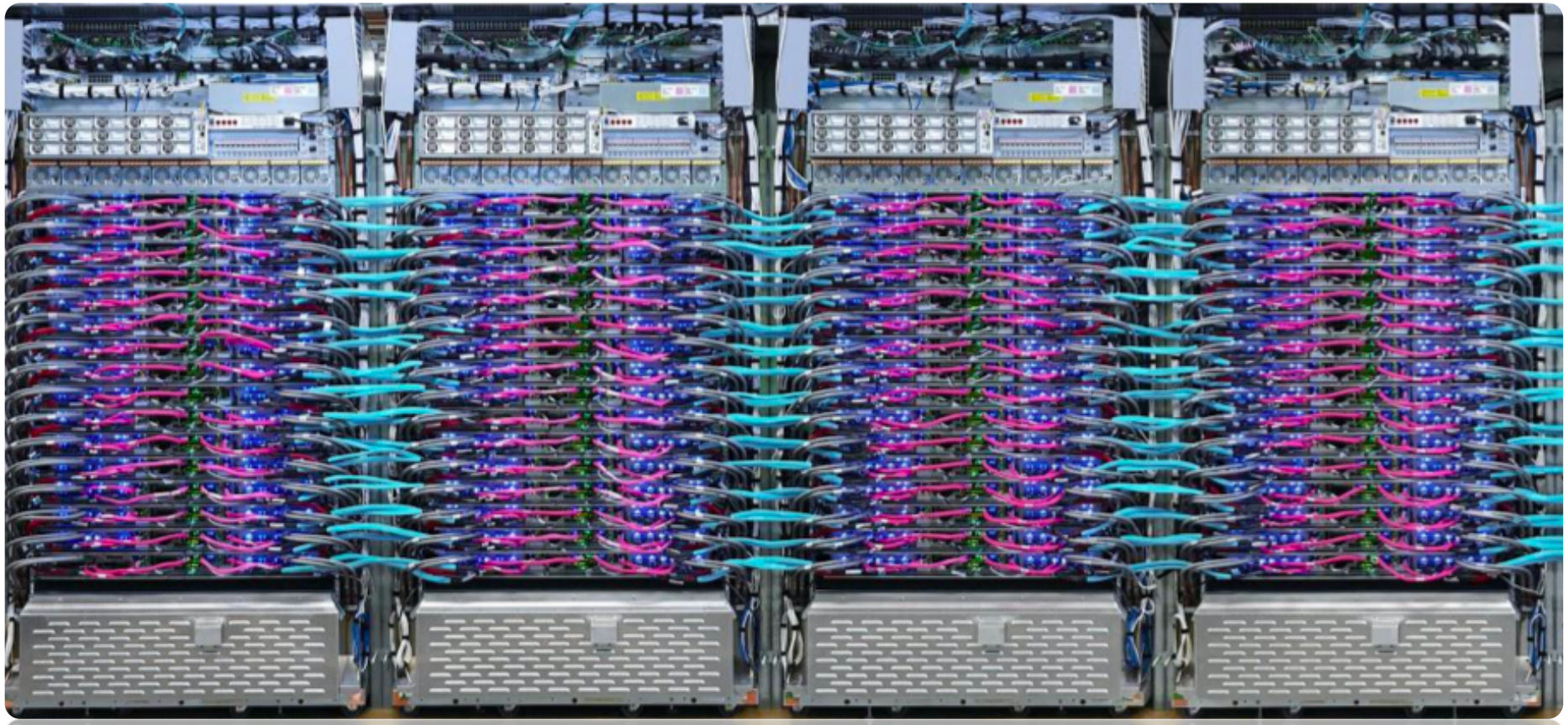
TPU v3

- 每个 v3 TPU 芯片包含两个 TensorCore。每个 TensorCore 都有两个 MXU、一个矢量单元和标量单位。



TPU v3 POD

- Google's HPC cluster TPU 3.0 Pod: 100 PFLOPS(8X faster than v2)



6. TPU4

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

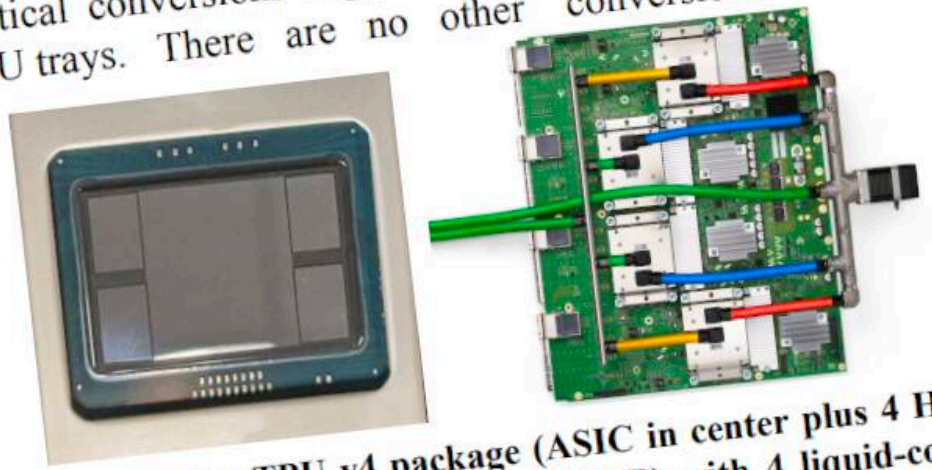


Figure 2: The TPU v4 package (ASIC in center plus 4 HBM stacks) and printed circuit board (PCB) with 4 liquid-cooled packages. The board's front panel has 4 top-side PCIe connectors and 16 bottom-side OSFP connectors for inter-tray ICI links.

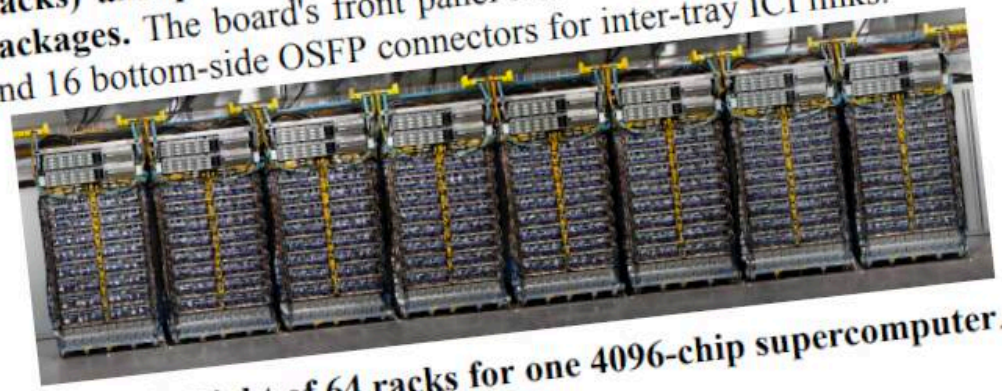


Figure 3: Eight of 64 racks for one 4096-chip supercomputer.

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).

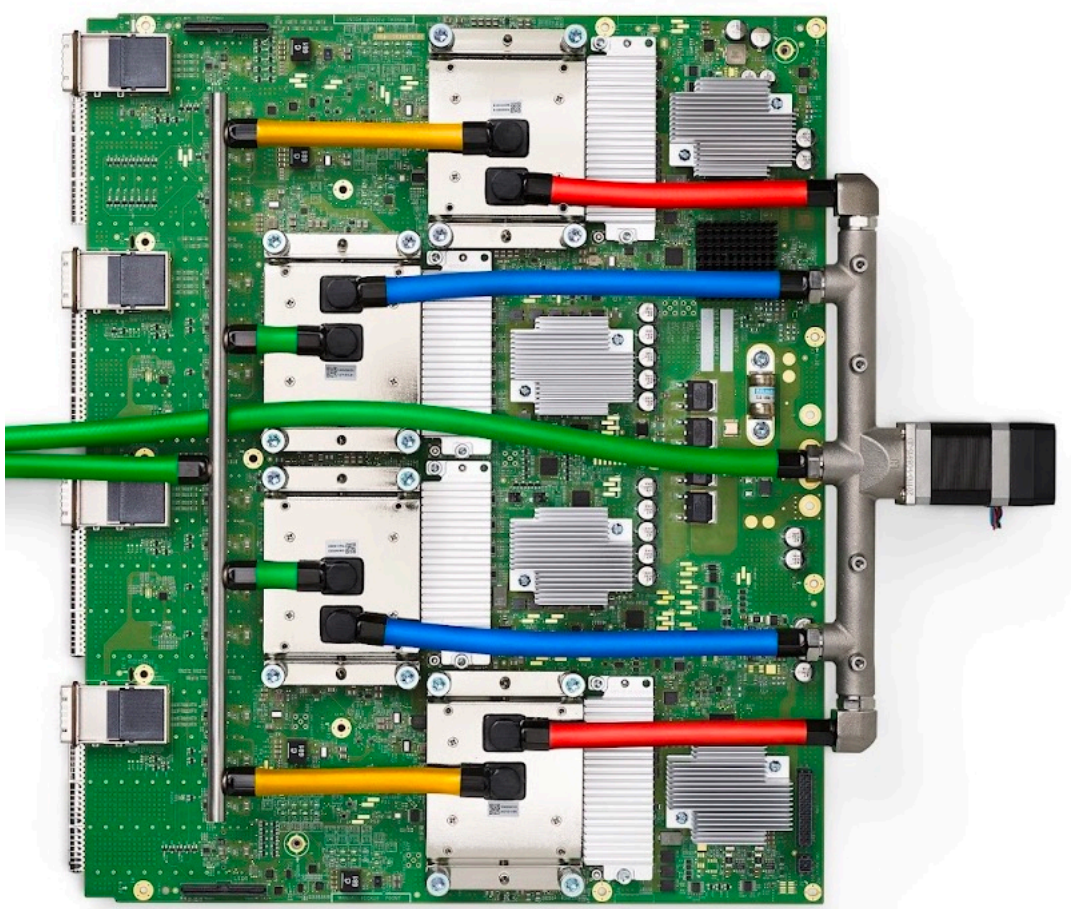
Table 2: Sampling of popularity of TPU v4 slices for a day in Nov. 2022. This table includes all slices used ≥ 0.1%. Twistable (Sec. 2.8) but not twisted means the slice geometry allows twisting (n×n×2n or n×2n×2n), but the user picks the regular topology. The software scheduler requires that slices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

TPU v4

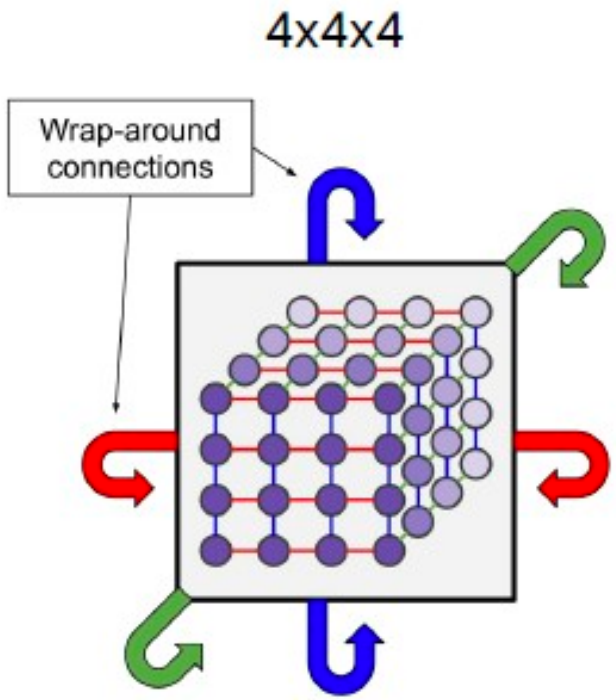
- TPU v4 是 Google TPU 系列计算引擎的真正升级，工艺从 16 纳米缩小到 7 纳米。MXU 的数量翻了一番，缓存内存增加了 9 倍至 244 MB，HBM2 内存带宽增加了 33% 至 1.2 TB/s，可惜 HBM2 内存容量保持不变 32 GB。

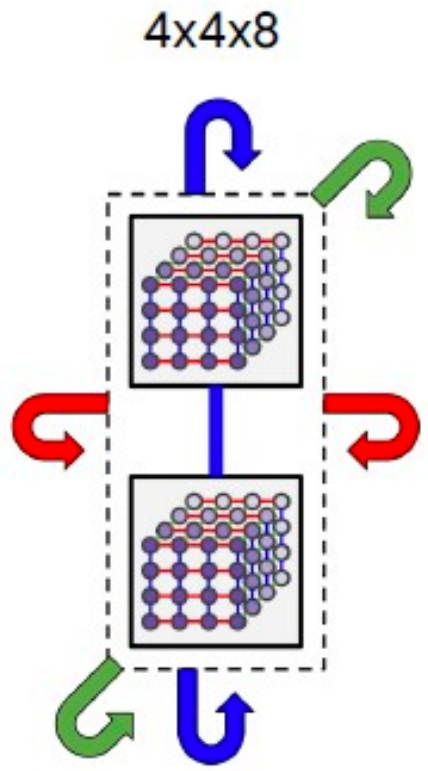
- TPUv4 首次亮相的新 3D torus 互联方式，具有更多带宽和更高基数，它可以紧密耦合 4,096 个 TPUv4 引擎，在 TPU v4 POD 总计提供 1.126 exaflops 的 BF16 峰值算力。



TPUv4 Overview

- Slice now with 3D torus
- Some Specific slices:
  - 2x2x1(v4-8)
  - 4x4x4(v4-128)
  - 4x4x8(v4-256)
  - 4x8x8(v4-512)
  - 8x8x8(v4-1024)
  - 8x8x16(v4-2048)
  - 8x16x16(v4-4096)



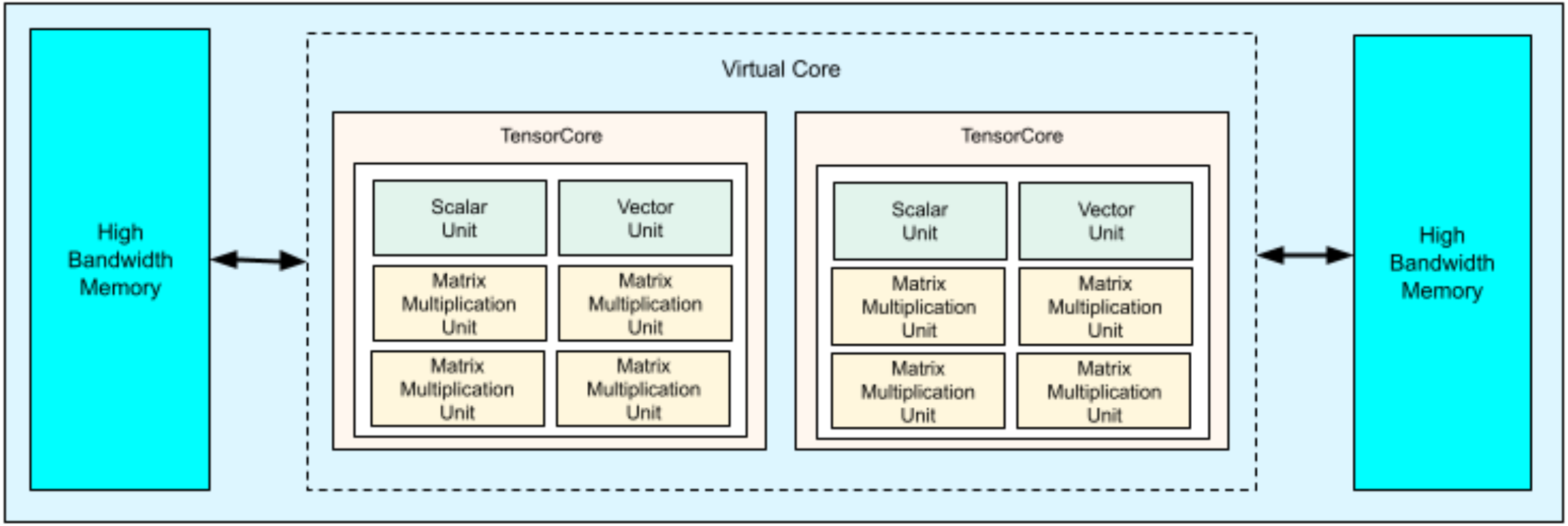


TPU v4

关键规范	v4 Pod 值
每个芯片的峰值计算次数	275 万亿次浮点数（bf16 或 int8）
HBM2 容量和带宽	32 GiB、1200 GBps
测量的最小值/平均值/最大值	90/170/192 瓦
TPU Pod 大小	4096 条状标签
互连拓扑	3D 环形图
每个 Pod 的峰值计算次数	1.1 Exaflops（bf16 或 int8）
每个 Pod 的全宽带宽	1.1 PB/秒
每个 Pod 的对分带宽	24 TB/秒

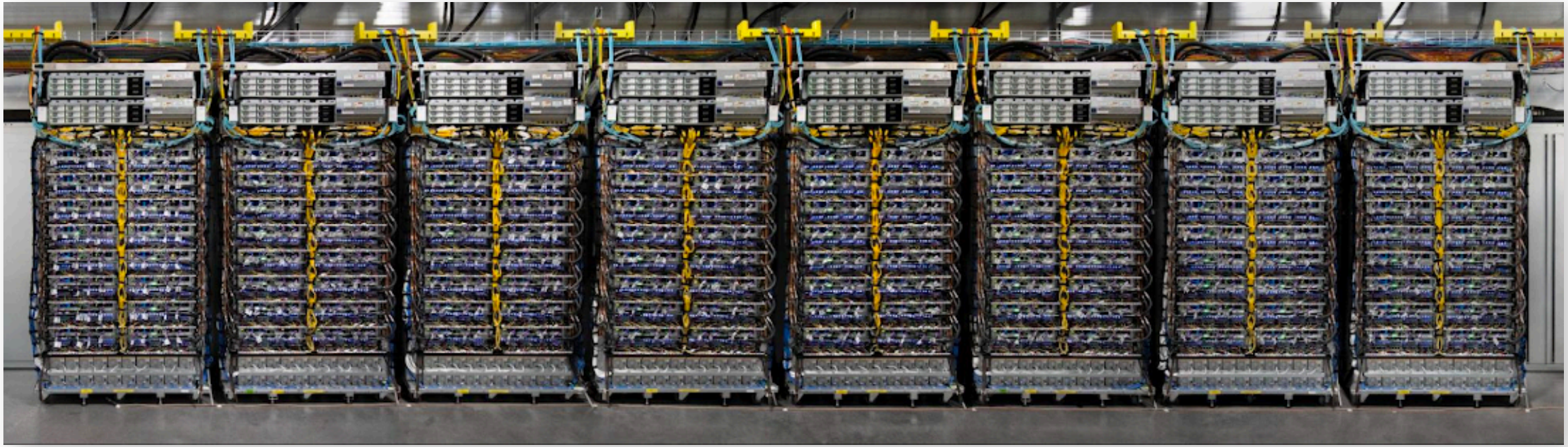
TPU v4

- 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。下表显示了 v4 TPU Pod 的架构图。



TPU v4 POD

- Google用TPU集群构建出Pod超级计算机，单台TPU v4 Pod包含4096块v4芯片，每台Pod的芯片间互连带宽是其他互连技术的10倍，因此，TPU v4 Pod的算力可达1 ExaFLOP，即每秒执行10的18次方浮点运算。



思考

- GPU 与 TPU 最大的区别在哪里？

思考

- 软件栈（AI框架、AI编译器）
- 互联方式（机间互联、芯片互联）
- 芯片架构（架构演进、数据驱动）

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