揭秘 FACEBOOK 未来的机器学习平台

人工智能3 月 19 日

粗看上去，世界上的超大规模用户和云构建商制造的东西通常看上去和感觉上去都像超级计算机，但如果你仔细观察，就常会看到一些相当大的差异。差异之一是，他们的机器并不是为了实现最高性能而不惜一切代价去设计，而是在性能和成本之间实现了最佳平衡。

简而言之，这就是为什么社交网络巨头 Facebook（世界上最大的人工智能用户之一）大量订购英伟达的 HGX-1 和 HGX-2 系统用于机器学习训练，然后就到此为止了。（HGX-1 和 HGX-2 系统是 GPU 加速器制造商英伟达的 DGX 系列的超大规模用户版本。）

这并不是巧合，为什么微软、谷歌、亚马逊网络服务、阿里巴巴、腾讯、百度，以及中国第四大巨头（中国移动或京东）同样设计自己的服务器，或是使用 Facebook 在 2011 年创建的开放计算项目（OCP）中的设计，或是在 OCP 启动六个月后由阿里巴巴、百度和腾讯发起了天蝎计划项目。在某些情况下，他们甚至设计自己的 ASIC 或在 FPGA 上运行专门用于机器学习的算法。

公平地说，Facebook 确实在 2017 年 6 月安装了英伟达 DGX-1 CPU-GPU 混合系统的半定制实现，该系统有 124 个节点，峰值双精度性能为 4.9 petaflops，在 HPC 常用的 Linpack 并行 Fortran 基准测试中的评价为 3.31petaflops。但这是个例外，不是常规。

但是，Facebook 喜欢设计自己的硬件，然后将其开源，试图围绕这些设计构建一个生态系统，以降低工程和制造成本，并降低供应链风险，因为越来越多的公司进入了开放计算领域。这与微软几年前加入 OCP 并将一系列完全不同的开源基础设施设计（从服务器到存储到交换）抛入 OCP 生态系统的原因相同。这增加了创新，但也导致了供应链分叉。

在本周于圣何塞举行的 OCP 全球峰会上，Facebook 展示了针对机器学习训练和基础设施的未来系统设计，让世界有机会看到针对现代数据中心的这两个日益重要的工作负载的成本优化设备的至少一个潜在的未来。这些设计非常有趣，表明 Facebook 热衷于创建能够容纳尽可能多的供应商的不同类型计算的系统，再次降低成本和供应链风险。

不是基本训练

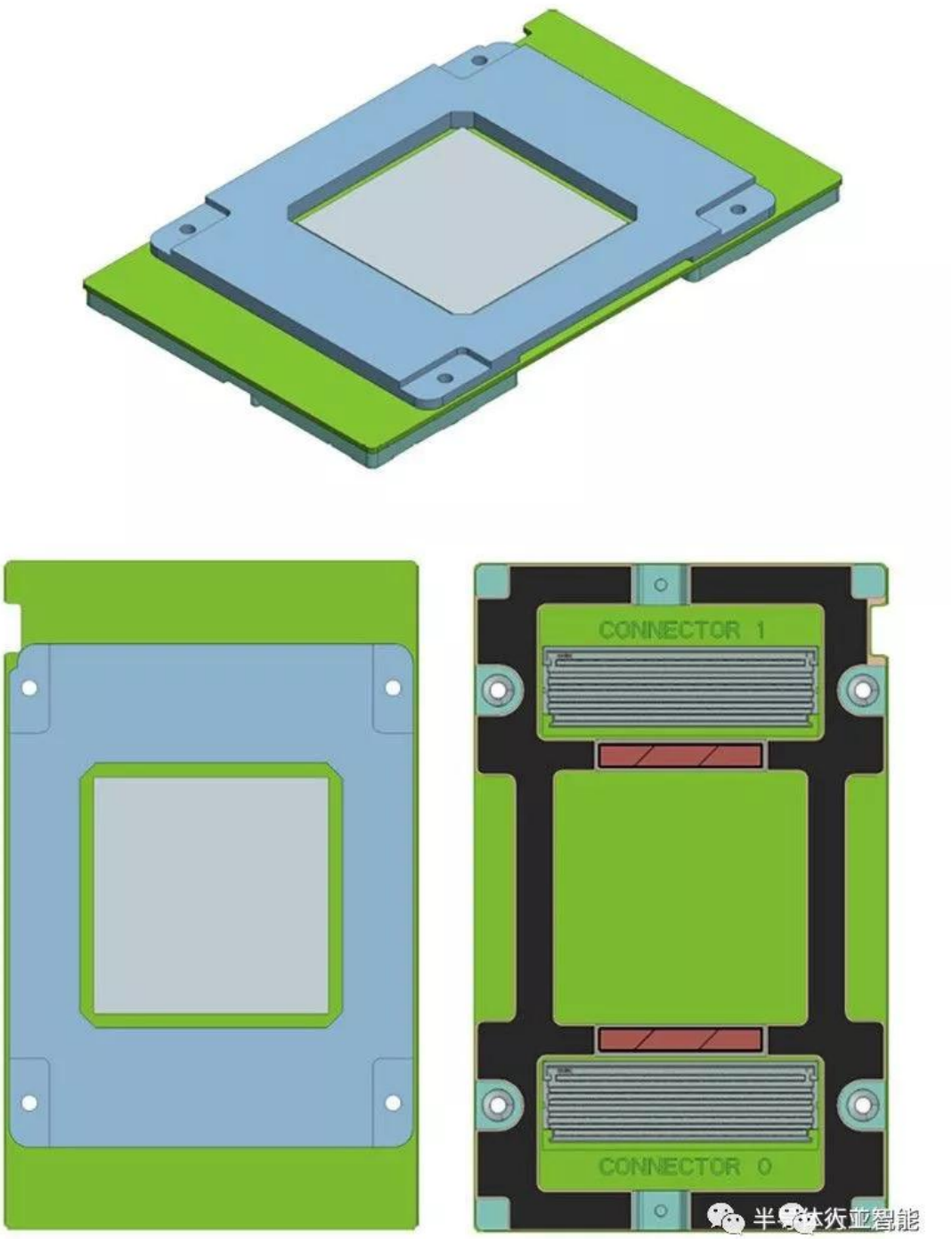
第一台新机器代号为“Zion”，它的目标是 Facebook 上的机器学习训练工作负载。Zion 系统由两个不同的子系统组成，就像英伟达的 DGX-1 和微软的 HGX-1，也包括 DGX-2 和 HGX-2，以及 ODM 和 OEM 厂商为客户制造的各种等价产品。 Zion 系统是两年前 Facebook 在 OCP 峰会上与微软的 HGX-1 一起发布的“Big Basin” ceepie-geepie 系统的继承者，这两个系统的设计都为 OCP 做出了贡献。Big Basin 机器的主机支持多达 8 个英伟达的“Pascal”GP100 或“Volta”GV100 GPU 加速器，以及两个英特尔 Xeon CPU。巧妙之处在于 CPU 计算和 GPU 计算是分开的，分别位于不同的主板和不同的机箱中，因此它们可以单独升级。具体取决于品牌和型号。

Big Basin 是对其前身“Big Sur”的彻底改进，后者是一款密度较低的设计，基于单个主板，配备两个 Xeon CPU 和多达 8 个 PCI-Express Nvidia Tesla 加速器（M40 或 K80 是最受欢迎的）。Big Sur 于 2015 年 12 月曝光。Facebook 在谈到设计时表示，开发工作已经基本完成，还没有投入生产，这意味着 Zion 机器还没有投入生产，但很快就会问世。（我们在 2018 年 1 月讨论了 Facebook 不断演变的 AI 工作负载，以及运行这些工作负载的机器。)Zion 机器的变化显示了 Facebook 在混合 CPU-GPU 机器上的想法的变迁，这些想法是我们许多人都想不到的。

Zion 机器的两个子系统被称为“Emerald Pools”和“Angels Landing”，分别指的是 GPU 和 CPU 子系统。尽管 facebook 多年来一直表示，其服务器设计的目的是允许选择处理器或加速器，但在这个例子中，facebook 和微软合作提出了一种独特的封装和主板插接方法，称为 OCP 加速器模块（简称 OAM），该方法允许使用具有不同插座和热量的加速器，可以选择 250 瓦至 350 瓦不等的风冷，未来则可以选择高达 700 瓦的水冷，但就硬件形式而言，所有这些都一致部署在这些加速系统中。

超大规模用户谷歌、阿里巴巴和腾讯将与 Facebook 和微软一起推广 OAM 封装，芯片制造商 AMD、英特尔、Xilinx、Habana、高通和 Graphcore 也是如此。系统制造商 IBM、联想、浪潮、广达电脑、企鹅计算、华为技术、WiWynn、Molex 和 BittWare 也都支持 OAM。毫无疑问，其它公司也将效仿它们的芯片和系统——惠普和戴尔显然是缺席的 OEM，而富士康和 Inventec 则是缺席的主要 ODM。

通过 OAM，加速器被插入一个便携式插座，它的管脚在一侧，然后是一组标准的并行管脚，它在概念上类似于英伟达的 SXM2 插座，用于 Pascal 和 Volta GPU 上的 NVLink，从模块上取下并插入主板上匹配的端口中。下图说明了它的原理：



任何插入 Emerald Pools 机箱的特定加速器都会有散热器，散热器具有不同数量的鳍片和不同的材料，可用于冷却其下方的设备，但高度一致，因此无论哪种加速器插入插槽，散热器都能以一致的方式保持整个机箱中的气流不变。虽然 Facebook 没有

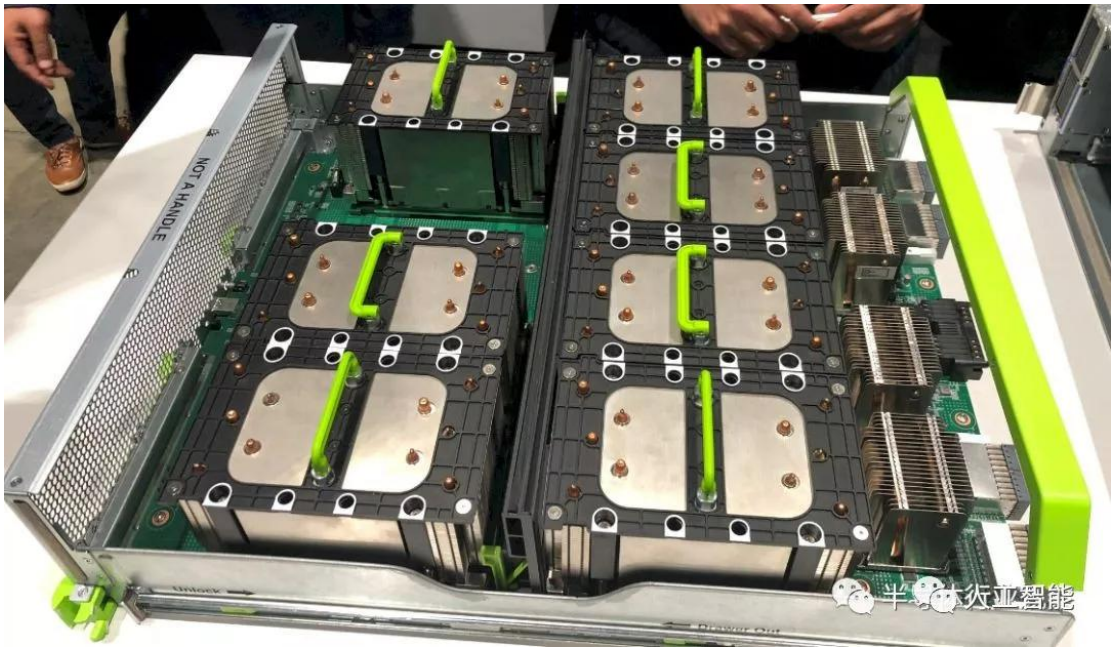
这么说，但没有理由不能将多个不兼容的加速器插入 Emerald Pools 机箱，并使用该机箱中实现的 PCI-Express 交换结构相互连接并与主机 CPU 连接。下图是 OAM 的外观：

它看起来很像小型汽车电池，不是吗？

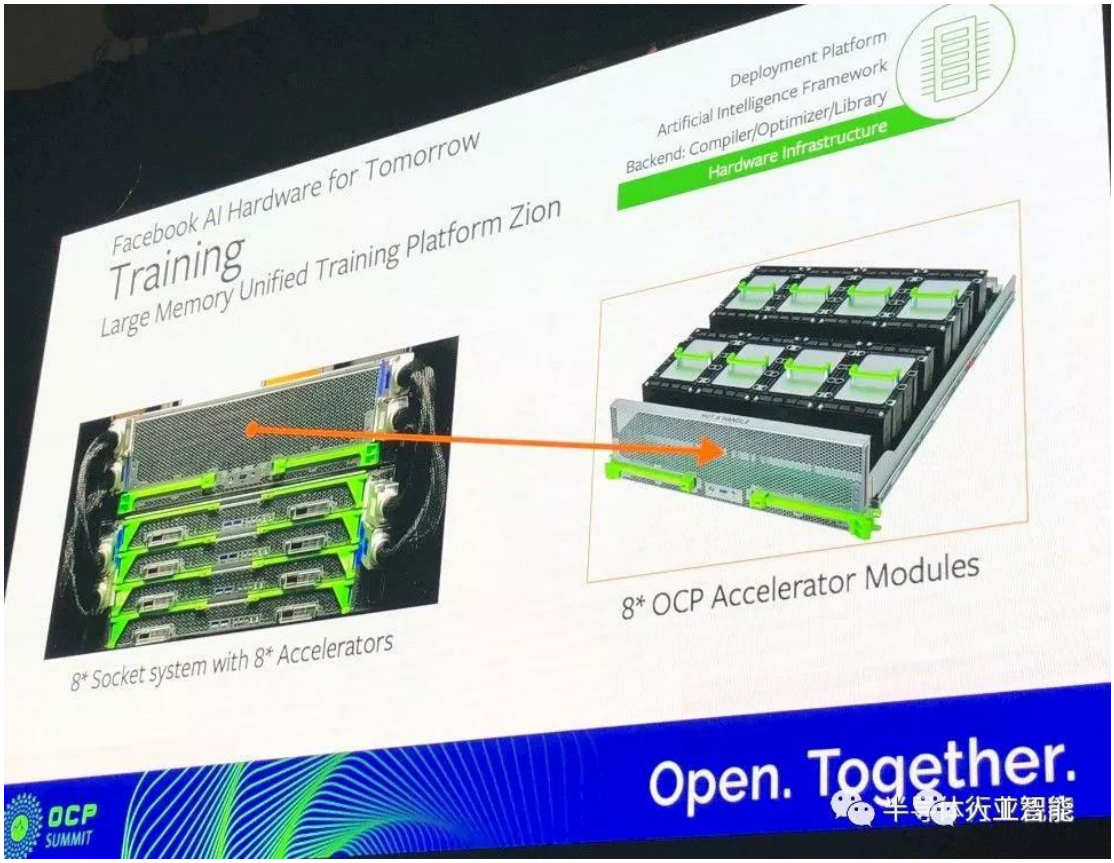
每个 OAM 的尺寸为 102 毫米×165 毫米，足够容纳我们认为未来将会越来越大的多芯片模块。对于耗电量高达 350 瓦的设备，OAM 可支持 12 伏特的输入；对于需要驱动高达 700 瓦的设备，OAM 可支持 48 伏特的输入；风冷的散热能力预计将在 450

瓦左右。当前的 OAM 规范允许在加速器和主机之间提供一个或两个 PCI-Express 3.0 x16 插槽，而且很显然，更快的 PCI-Express 4.0 和 5.0 插槽已在规划图中。这样就剩下 6 到 7 个 PCI-Express 链路用于交叉耦合加速器。顺便说一句，这些链路可以分成两部分，以提供更多的互连链路，并可以增加或减少任意给定链路的通道数量。

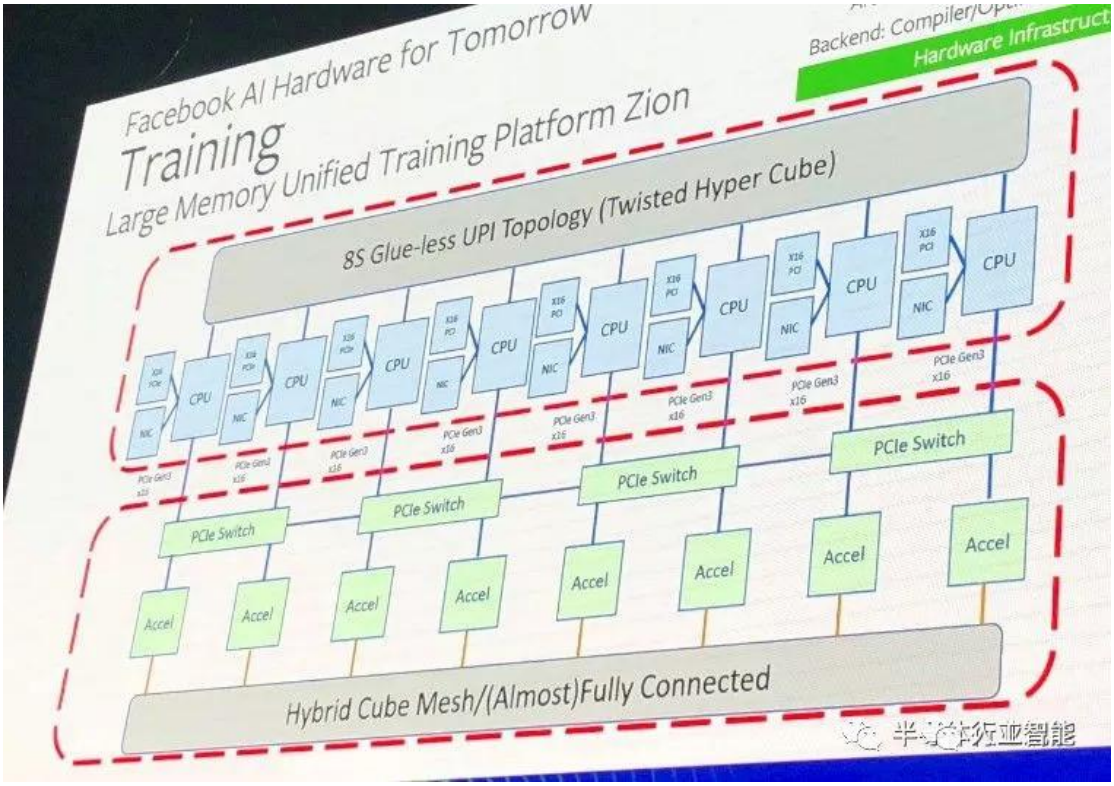
下图是 Emerald Pools 机箱，里面插了 8 个加速器中的 7 个。



Emerald Pools 底座后面有四个 PCI-Express 交换机，位于图片的右侧，每个交换机都插入对应的 Angels Landing CPU 机箱（即 Zion 系统的另一半）上的配套 PCI-Express 交换机。该系统的 CPU 部分没有在 Facebook 展位上展出，但 Facebook 技术项目经理、设计其 AI 系统的工程师之一 Sam Naghshineh 在一次演讲中展示了这台机器：



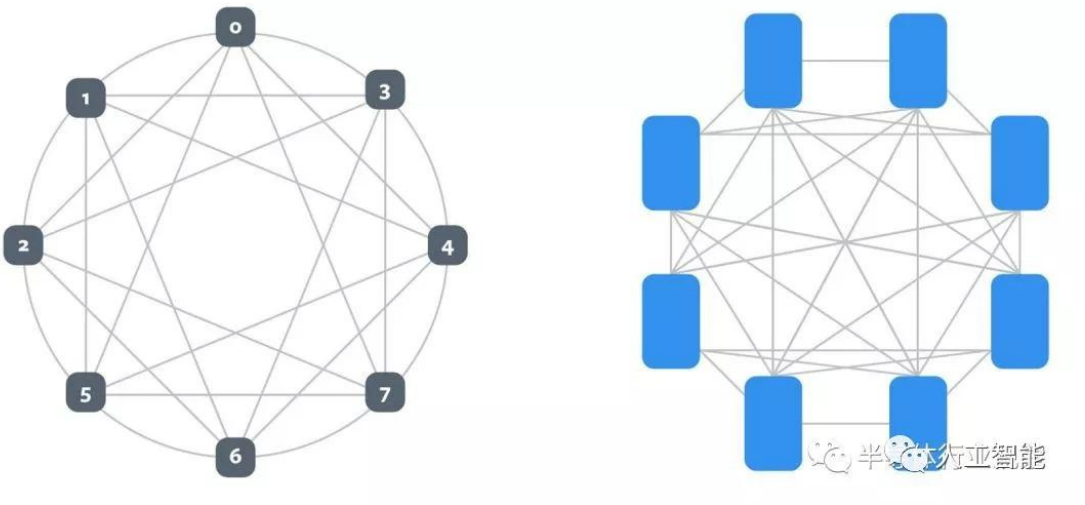
你可以看到，4 个 PCI-Express 3.0 管线从加速器底座和 CPU 底座上出来，将它们连接在一起。关于 Angels Landing 有趣的一点不是它总共有 4 个服务器底座，每个都有一对 Xeon SP 处理器，这是超大规模数据中心的常规设计。巧妙之处在于，由于在系统的 CPU 端进行机器学习训练期间，对数据密集处理的需求不断增加，于是它使用处理器上的 UltraPath Interconnect（UPI）链接将这 4 个双插槽机器捆绑在一起，以创建一个 8 插槽共享内存节点。按照 Naghshineh 的说法，从技术上讲，这称为扭曲超立方体拓扑：


Facebook AI Hardware for Tomorrow
Training
Large Memory Unified Training Platform Zion
8S Glue-less UPI Topology (Twisted Hyper Cube)
Backend: Compiler/Op
Hardware Infrastruct
CPU
NIC
PCIe Gen3 x16
PCIe Switch
Accel
Hybrid Cube Mesh/(Almost)Fully Connected

这个大 CPU 节点设计为拥有 2 TB 的 DRAM 主内存，而无需使用大内存条或 Optan3D XPoint 主内存，而且重要的是，该节点可在系统的 CPU 端提供足够的内存带宽，从而无需使用 HBM 内存。（这并不是说英特尔或 AMD CPU 还拥有 HBM 内存，但某

些场合它们确实拥有 HBM 内存，尤其是对于 HPC 和 AI 工作负载而言。）这 8 个插槽的 DRAM 内存带宽和容量一样重要。

如你所见，Angels Landing CPU 机箱中的每个 CPU 都有自己的网络接口卡以及 PCI-Express 3.0 x16 插槽，用于将 CPU 连接到 PCI-Express 交换机结构，该交换机结构将加速器计算复合体连接在一起，并连接到 CPU。这些加速器链接在上图中几乎完全连接的混合立方体网格中，但还可以支持其他拓扑，如下所示：



左图中，每个加速器有 6 个端口，8 个加速器连接在一个混合立方体网格中。右图中，仍然有 8 个设备，但是每个设备都有一个额外的端口（总共 7 个），这些设备可以按照 all-to-all 的互连方式进行链接。显然还有其他选择，重点是不同的神经网络在不同的互连拓扑结构中效果更好，这将允许 Facebook 和其他公司改变互连的拓扑结构，以满足神经网络的需求。

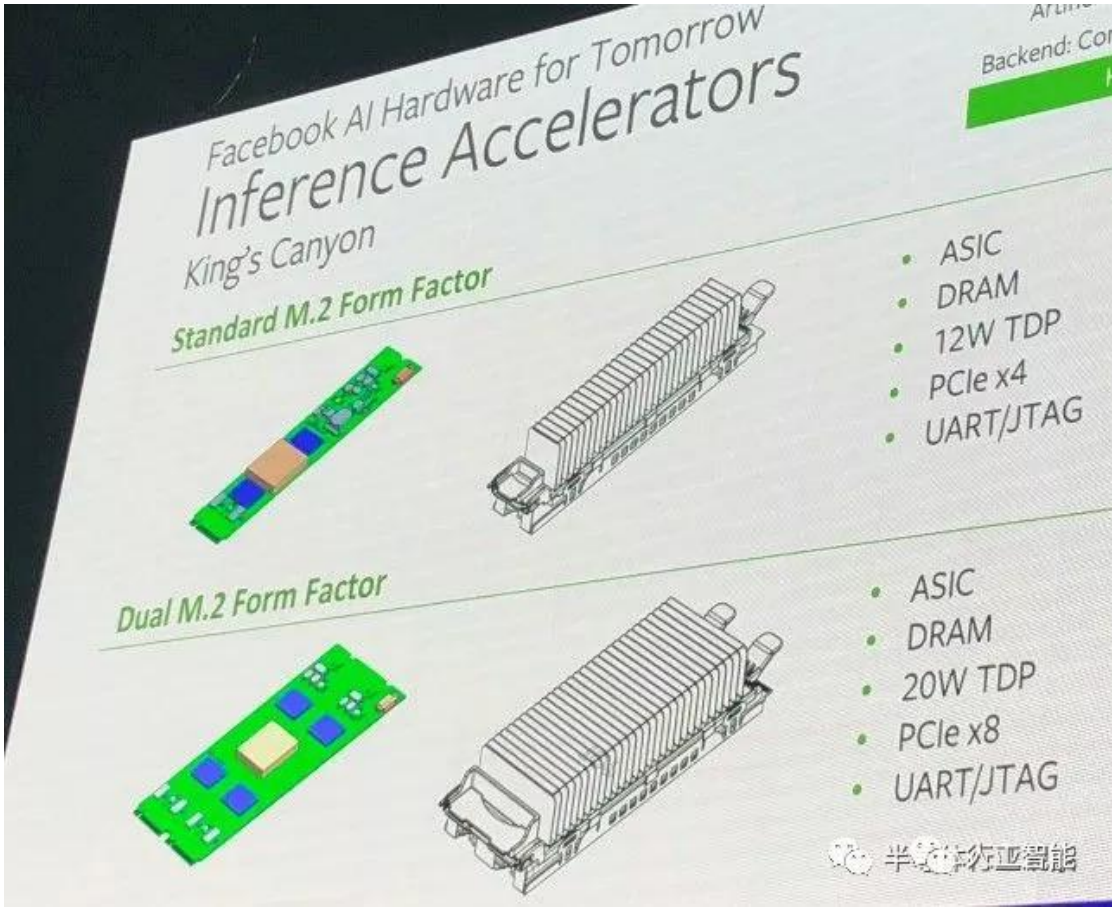
推理的未来

Facebook 毫不掩饰地表示，它希望拥有比目前市场上更高效的推理机，这是 Facebook 去年在一篇论文中讨论的一个话题。在本周的 OCP 全球峰会上，Facebook 公司高层概述了机器学习推理硬件的未来。

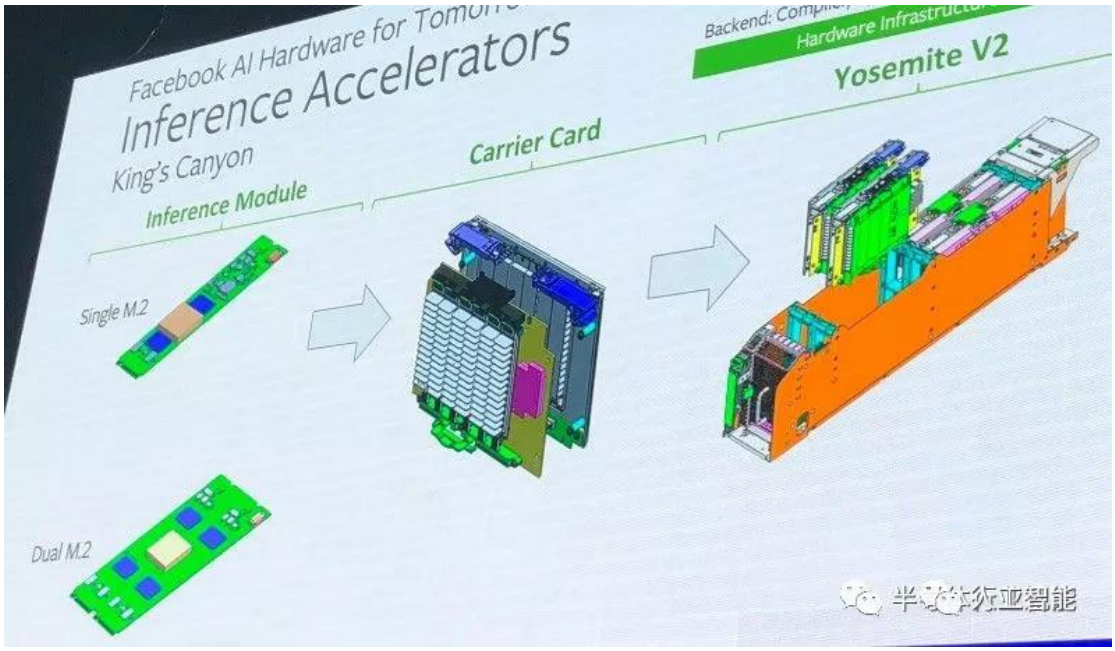
Facebook 技术和战略主管 Vijay Rao 提醒大家，早在 1980 年，英特尔就为 8086 系列处理器设计了 8087 数学协处理器，这些处理器如今是客户端的核心芯片和服务器上的 Xeon 芯片的前身。这些机器可以在 2.4 瓦的热度范围内实现 50 kiloflops（32 位单精度），达到相当惊人的每瓦 20.8 kiloflops。Facebook 的目标是使用像 INT8 这样的低精度数学运算，来达到接近每瓦 5 teraflops，如果你看看英伟达的 GV100，它可以达到每瓦特 0.4 teraflops。

Rao 在他的主题演讲中解释说：“我们一直在与许多合作伙伴密切合作，设计用于推理的 ASIC。与传统 CPU 相比，在加速器中运行推理的吞吐量增加是值得的。在我们的情况下，应该是每瓦特 10 倍左右。”

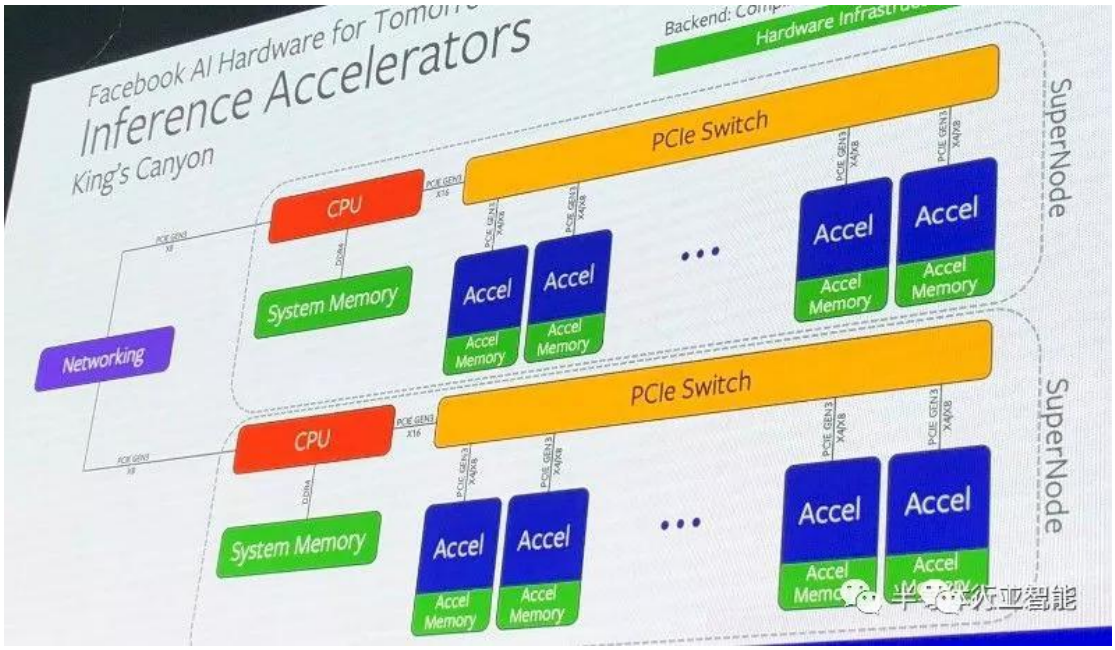
Rao 大致谈到了将 M.2 推理引擎组合到微服务器卡上，然后将它们插入到 2015 年创建的“Yosemite”服务器机箱中，Facebook 设计该机箱是为了完成基本的基础设施工作。但当天晚些时候，Naghshineh 实际展示了它的实现方法。以下是 M.2 推理引擎的“Kings Canyon”系列：



Facebook 正试图鼓励推理芯片制造商支持两种不同的形式。一个是单个的宽 M.2 单元，最大支持 12 瓦，并带有一个 PCI-Express x4 接口，另一个具有两倍大的内存、20 瓦的热度范围，一对 PCI-Express x4 端口，可以单独使用或捆绑使用。这些 M.2 推理卡中的多个被插入“Glacier Point”载卡中，该载卡插入真正的 PCI-Express x16 插槽，最多可以有 4 个载卡被插入 Yosemite 机箱，如下所示：



群集推理引擎的框图如下所示：



这样做的唯一原因与使用低核心计数、高频率、单插槽的微型服务器来运行电子设计自动化（EDA）工作负载相同，英特尔就是这样做的，尽管它想要向世界销售双插槽服务器。推理工作负载类似于 Web 服务和 EDA 验证：你可以将整个较小规模的工作分派到大量松散耦合（几乎没有耦合，完全不是真正耦合）计算单元中的一个，然后

一次执行大量的这些任务，并同时完成大量工作。对一位数据的推断决不依赖于对无数其他工作的推断。机器学习训练则不同，它更像传统的 HPC 仿真和建模，在不同的程度和频率下，对一个计算元素进行的任何处理都依赖于其他计算元素的结果。

因此，我们所看到的用于机器学习训练和推理的截然不同的硬件设计都来自 Facebook。我们可以肯定的是，Facebook 希望能够采用它认为适合框架的任何类型的 CPU 和加速器进行训练，以及任何价格低廉的芯片推理引擎，在任意给定的时间内，它的性能都比 CPU 好 10 倍。今天在 Facebook 运行在 X86 服务器上的推理业务是英特尔的失败。或许也未必，没准 Facebook 会决定在今年晚些时候推出 M.2 Nervana NNP 推理引擎。我们将会看到推理是如何流过 Kings Canyon 的。

本文来源：半导体行业观察 (ID:icbank)，作者：nextplatform