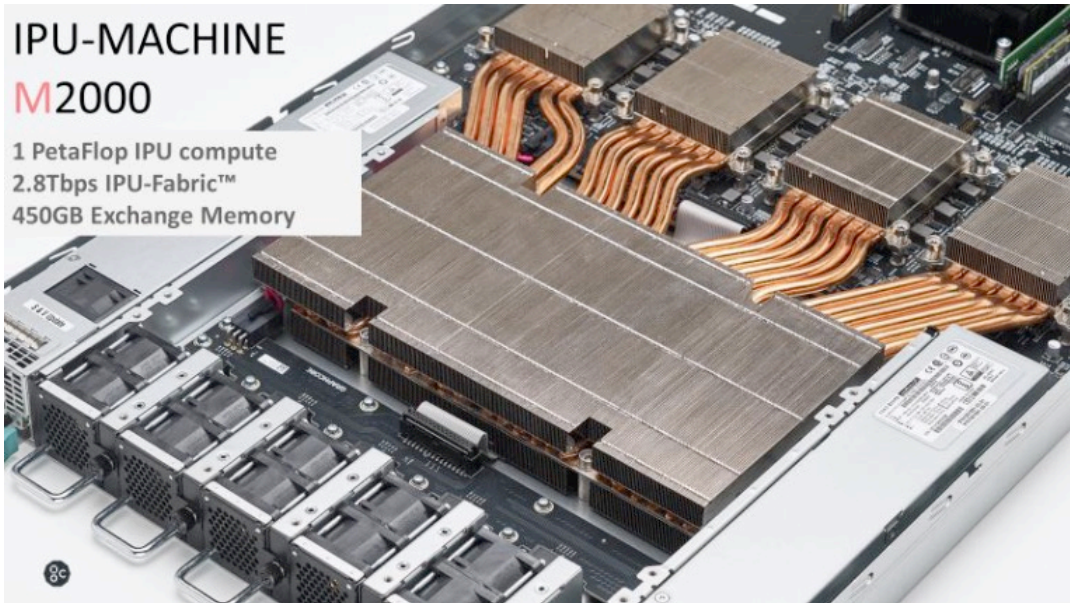
GRAPHCORE 第二代 IPU

简介

总部位于英国的初创公司 Graphcore 于 2018 年推出了用于人工智能加速的智能处理器（IPU）。如今 Graphcore 推出了 IPU-Machine。这个第二代平台具有更强大的处理能力、更多的内存和内置的可扩展性，可以应对极大的并行处理工作负载。这家资金雄厚的创业公司拥有一流的工程师、顾问和投资者，估值接近 20 亿美元。它的第一代硬件目前已在微软 Azure 云以及 Dell-EMC 服务器中使用，两者皆为其投资者。Graphcore 现在将自己的未来倾注在第二代平台上：一个目前在业界独一无二的、具有广泛扩展性的即插即用构建块。

图 1：Graphcore IPU-Machine



Graphcore IPU-Machine 包括四个 IPU、集成的 100GbE 横向扩展结构、PCIe 和附加的 DDR 内存。每个 1U 设备都可以提供高达 1 PFLOP 的 AI 性能和 450GB 的存储。

Source: Graphcore
来源：Graphcore

Copyright ©2020 Moor Insights & Strategy

在该公司 Poplar 软件栈的支持下，该新平台的亮点在于 4 通道 IPU-Machine 设备中的 7 纳米 Colossus MK2，并且可在 1024 个机架上扩展到最多 64,000 个 IPU。完全配置的 AI 超级计算机可以提供大约 16 EFLOPS 的 AI（16 位 FP）性能，与第一代产品一样，该公司的重点是实现对大规模高性能并行计算的简化。

本研究论文将探索这个新平台，并将之与其不断增加的潜在竞争对手进行比较，评估其优势和劣势。

THE COLOSSUS MK2 IPU (GC200)

由台积电制造的新 MK2 部件是大规模并行的 594 亿晶体管处理器，可跨 1472 个核提供约 250 万亿次每秒运算（TOPS），并通过 2.8Tb/秒 的低时延结构互连 900MB 的处理器内存储。MK1 一代的绝大多数架构设计都可以延续到 MK2 平台，其处理图块包含核和片上 SRAM，它们在同一结构上互连，可以向片外扩展，以与其他 IPU 域通信。

在将原始 IPU 设计缩小到 7 纳米时，Graphcore 的设计者选择性能和内存的最大化，而非降低成本，这与人们普遍的认知相一致，即：在数据中心训练应用中，AI 应用程序主要在性能上受限，对成本并不敏感；而边缘 AI 推理处理对成本和功耗更为敏感。因此，与之前的产品相比，MK2 实现了核数量增加 20%，片上 SRAM 增加 333%，可扩展性提升 16 倍。新的系统软件可以增强可扩展性、部署和管理。

在系统级别上，高达 448GB 的“流存储”DRAM 补充了片上 IPU 内存。MK2 IPU 还通过实施由 Graphcore 开发的一组称为 AI-Float 的新颖的浮点实施技术来提高性能。AI-Float 技术用于调整 AI 计算的能量和性能，使用标准 IEEE FP16 格式，（Graphcore？）以多种方式对 AI-Float 进行了优化，其随机舍入特性可以使 FP16 能够匹配主权重上的 FP32 性能，使 FP16.16 可以匹配 FP16.32，进行正向和反向传播（每个芯片提供 250 TFlop/秒的运算）。该芯片还支持 62.5 TFlop/秒单精度 FP32。

许多、甚至于大多数 AI 模型产生具有高度稀疏性的模型参数。不乘以零元素可以将性能提高两倍或更多。MK2 具有针对稀疏模式的新稀疏性优化，包括块、标量、静态和动态稀疏性。挑战在于，需要提早知道什么时候不做乘法。Graphcore 能够将稀疏性优化有效地编译到计算图顶点小代码中。

Copyright ©2020 Moor Insights & Strategy

THE IPU MACHINE (M2000)

对于加速器相而言，将 AI 硅片作为系统而不是芯片来交付已变得越来越普遍，因为与定制化的依赖于 OEM 或 ODM 的设计和测试相比，它可以将上市时间缩短六个月至一年。Graphcore 在一个双 IPU PCIe 主板上实现了 MK1 硅片的产品化，以简化采用并缩短上市时间。借助 MK2 版本，Graphcore 又迈出了重要的一步，交付了包含四个 IPU 设备的装置，称为 IPU-Machine （M2000）。1U 披萨盒可通过搭载了 ROCE（聚合以太网上的 RDMA）的 100Gb 以太网访问，以实现低时延访问。使用以太网避免了 PCIe 连接器的瓶颈和成本，并实现了灵活的 CPU 来加速比率。

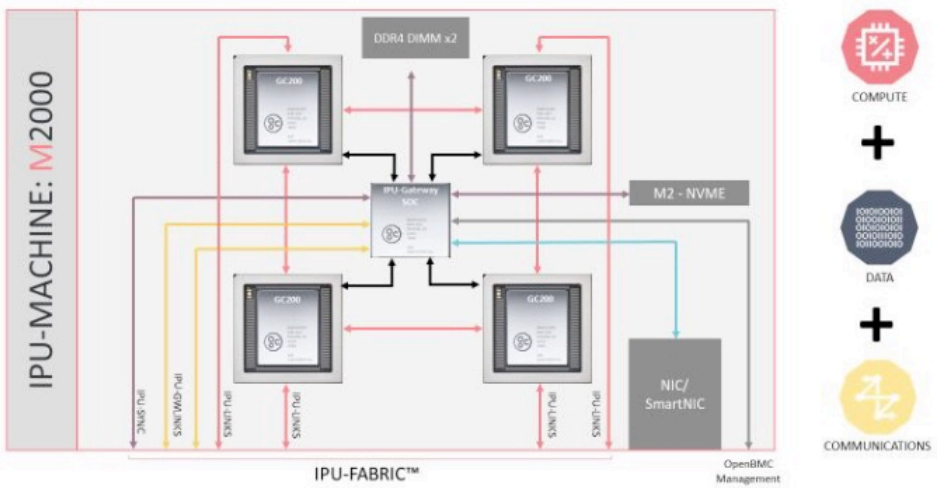
盒中包括集成横向扩展网络，使用户能够轻松地从小型开发系统扩展到大规模机架部署，所有这些均通过标准网络进行联网，且成本低于使用 InfiniBand。IPU 结构通过以太网隧道连接图块 tile 和其它 IPU，无论部署的大小如何，都保持相同的编程模型和批量同步处理（BSP）同步。IPU-Machine 标价$ 32,450，听起来可能很昂贵，但是该平台的性能与和竞争对手进行的比较充分说明这款平台代表了很高的价值。

除了即插即用扩展功能外，IPU-Machine 还通过可供四个 IPU 使用的 DDR 内存来补充处理器内存储。客户对 MK1 的反馈表明，下一代将需要多得多的内存来运行正在开发的极大型的模型。新的 IPU-Machine 提供 450 GB 的内存来处理这些更大的模型。据 OpenAI.org 称，由于模型大小每 3.5 个月增加一倍，因此该内存架构可能会改变游戏规则，提供的带宽是高带宽内存（HBM2）所能提供的带宽的 100 倍，容量是其 10 倍，与此同时成本显著降低。

IPU-Machine 的存储模型也与 CPU 或许多 AI 加速器（例如 GPU）中的存储模型有很大不同。Poplar Graph Compiler 不需要在存储层次结构中将数据和代码从主机内存存储区交换到加速器的存储，而是在图块 tile 上的存储和 IPU-Machine 上的 DDR 内存中创建确定性的代码-存储关系。计算图顶点中的 IPU 图块 tile 知道数据驻留的位置并直接访问它。没有缓存，没有交换，没有预取，也没有因此增加延迟。实际上，IPU-Machine 可以在独立模式下用于推理处理，而无需连接到主机服务器。而且，由于在 MK1 中首次引入的 BSP 模型可以同时编译计算和通信，因此与并行处理常用的传统消息处理或共享存储结构相比，网络通信费用可以保持在最低水平。IPU-Machine 包括一个可以访问 DRAM 的网关芯片，两个 100Gbps IPU-Fabric Link，一个用于标准 SmartNIC 的 PCIe 插槽，两个 1GbE OpenBMC 管理接口，以及可以访问一个 M.2 插槽。

Copyright ©2020 Moor Insights & Strategy

图 2：IPU-Machine M2000 架构图



Graphcore 将通过 1U 的 IPU-Machine 设备交付 IPU 芯片，以实现轻松扩展和部署。

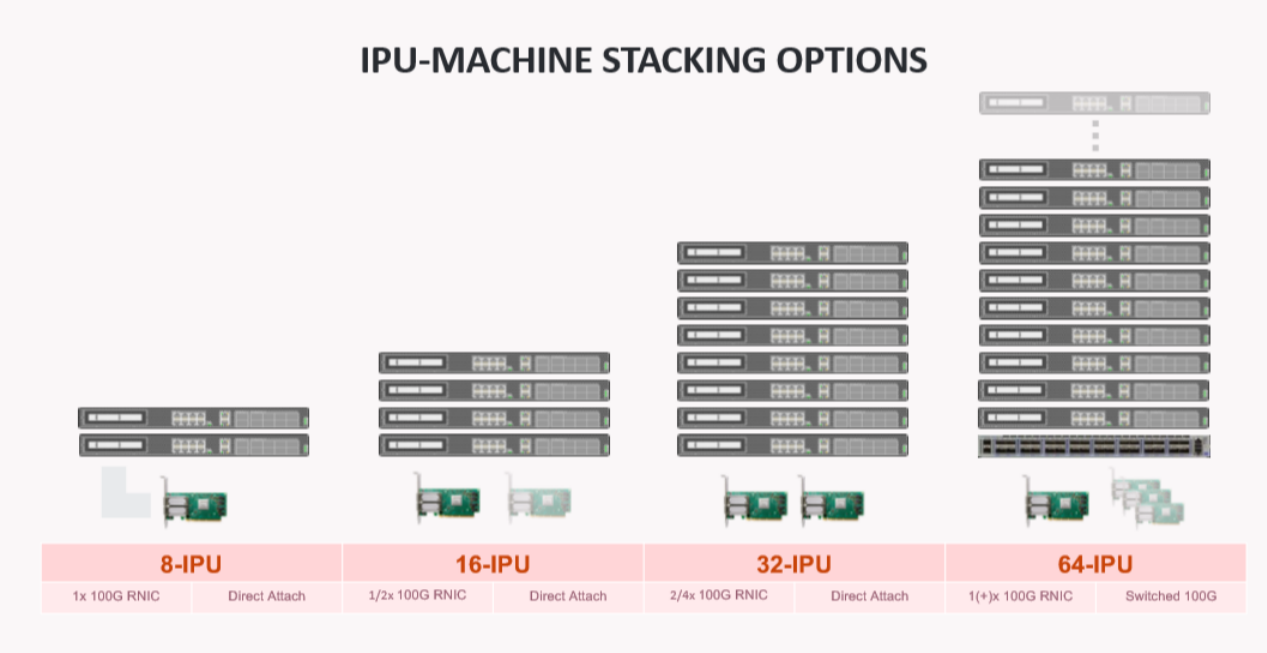
Source: Graphcore
来源：Graphcore

第二代 IPU 结构

因为模型尺寸急剧扩大，其中一些甚至包含数十亿个参数，内置结构已成为 AI 加速器的必需品。这些大型模型必须分布在数百或数千个处理器中，才能在合理的时间内解决问题。为了满足这一需求，一些公司在其加速器上提供了专有的结构，例如 NVIDIA 的 NVLink，而其他公司，例如在 2019 年被英特尔收购的 Habana Labs，则依赖于标准以太网。Graphcore 的混合模型使用专有的 IPU-Link 结构在 IPU 中的图块 tile 与机架中相邻 IPU 之间进行通信，同时在标准 100GbE 上隧道传送 IPU-Link 协议，从而实现支持更大配置的机架到机架的横向扩展。 IPU-Machine 设计在这里发挥巨大作用，无论是否带有主机 CPU（用于推理处理），都可以在大型基础设施中实现即插即用的扩展。

图 3：IPU-Machine 堆叠选项

Copyright ©2020 Moor Insights & Strategy



The IPU-Machine enables plug and play scaling, and Graphcore is providing reference architectures for users to ensure tested configurations.

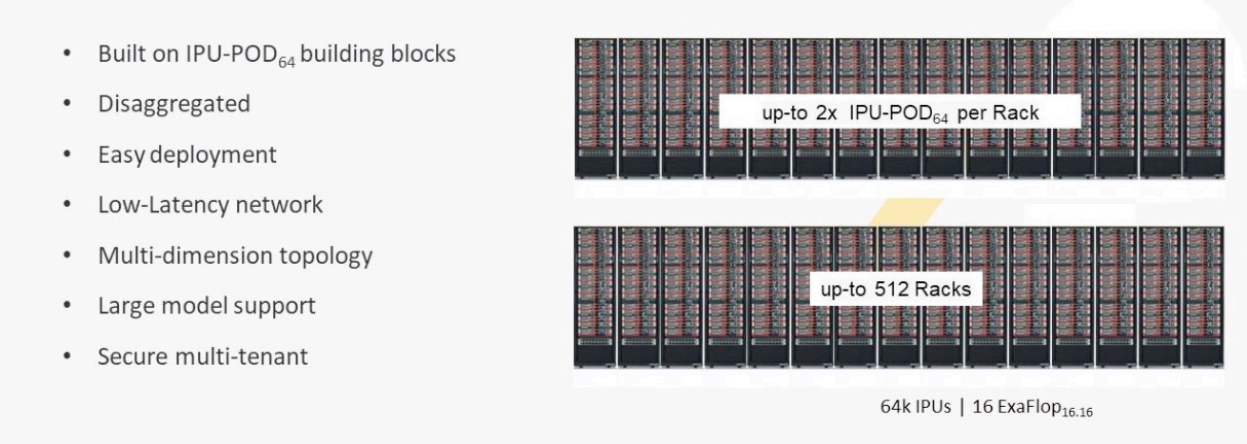
IPU 机器可实现即插即用的扩展，而 Graphcore 正在为用户提供参考架构，以确保配置经过了测试。
Source: Graphcore
来源：Graphcore

Graphcore 结构实现了灵活的分解模型，使用户可以在运行过程中配置加速器阵列，而不受大多数其他系统中 CPU 与加速器的固定比例的约束。例如，自然语言模型需要很少的 CPU 交互和利用率，其中 CPU 与加速器的比例为 1-100 就足够。 另一方面，卷积神经网络可能需要 1-4 或 1-8 的比率，因为这些功能（例如求平均值）发生在标量 CPU 内核上。通过利用 100Gb 以太网，弹性配置易于部署和使用，可扩展至 64,000 个 IPU。 我们认为，这种分解的扩展模型可能是第二代 Graphcore IPU 平台最重要的功能。

机器学习在三个维度上彰显了并行性：数据（或批）并行性，用于转换、池化等的层并行性，以及跨多个实例的张量并行性。理想情况下，这是在多维结构间实现的，并且 IPU-Fabric 支持的 3D-Ring 拓扑通过一对一的直接通信以极低的延迟有效地实现了这一点。该结构还公开了诸如全缩减，全聚集和广播之类的集合，以促进应用程序的开发。

图 4：用于超级计算规模的 IPU-POD

Copyright ©2020 Moor Insights & Strategy



Graphcore envisions entire data centers filled with interconnected IPU-Machines in interconnected IPU-PODs. Multi-tenancy support could make this attractive to cloud-scale service providers such as Microsoft Azure.

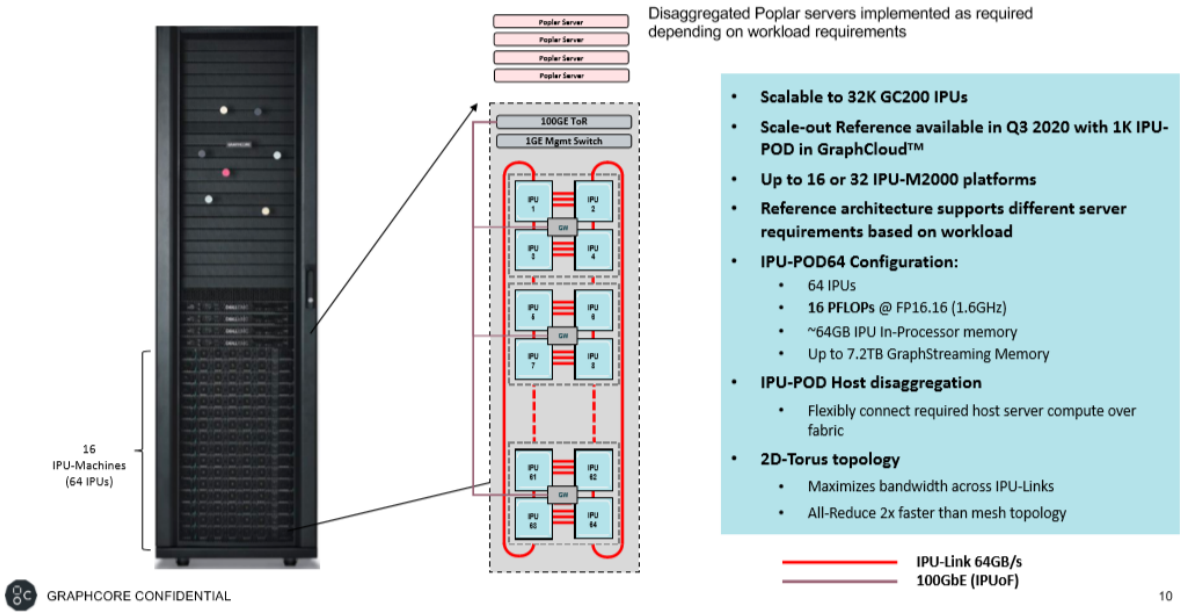
Graphcore 设想了布满了在互连 IPU-POD 中互相连接的 IPU-Machine 的整个数据中心。多租户支持可能使它对诸如微软 Azure 这样的云规模服务提供商具有吸引力。

Source: Graphcore
来源：Graphcore

分解的可扩展模型支持广泛的部署选项，其中一些已通过参考实施实现了标准化。尽管 Graphcore 尚未披露正在构建全规模 16 Exaflop 实施方案的客户，但我们怀疑有些客户正在探索这种级别的部署。

图 5：IPU-Pod 64 参考架构

Copyright ©2020 Moor Insights & Strategy



使用现有的以太网设备，结构链接 fabric links 支持机架内和整个数据中心范围的高效通信。

Source: Graphcore
来源：Graphcore

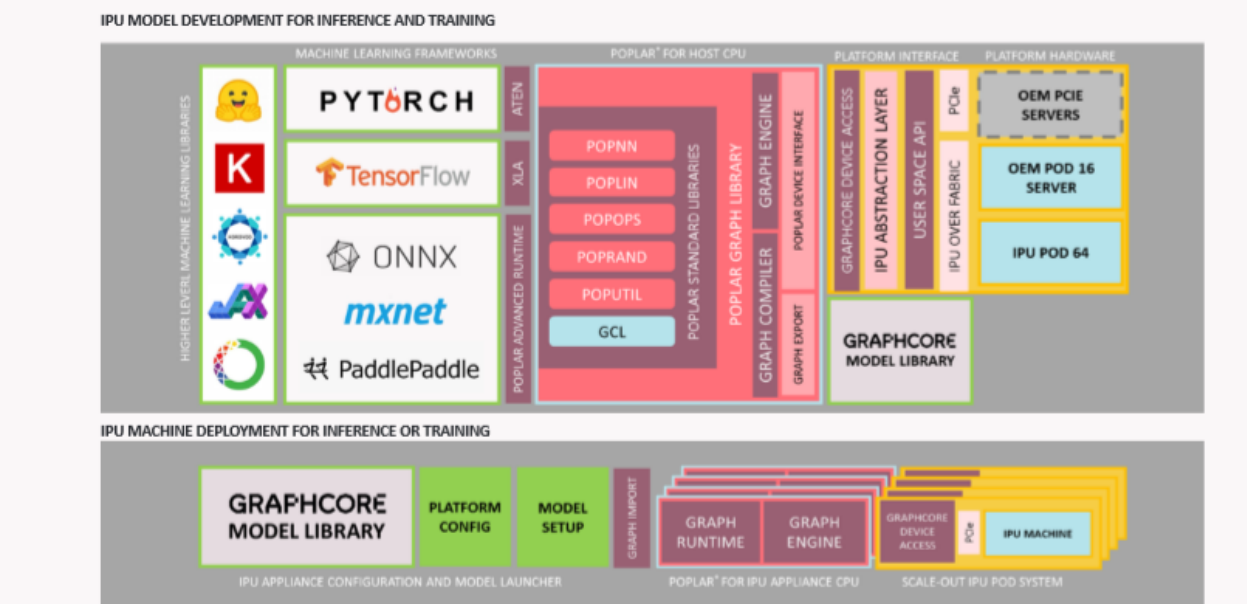
IPU-Fabric™和 BSP 编程模型可实现扩展性。正如我们在 Poplar 软件堆栈研究中所探索的那样，编译成“虚拟 IPU”实例、然后部署在任意配置上的能力对于大规模并行应用程序的设计者来说应该很有吸引力，这些应用程序包括但不限于深度神经网络模型。IPU-Machine 机架中的 2D 圆环互连可实现高效的全减少和其他集体操作。

Graphcore 软件栈的更新

Graphcore 已对 Poplar 软件堆栈作出多项改进。很自然，当尝试扩展到 MK2 提供的水平时，高级管理软件就变得至关重要。Graphcore 已扩展其软件堆栈，以包括 Graphcore 通信库（GCL）和行业标准工具的插件，这些行业标准工具用于调度、虚拟化、安全性和监控，诸如 SLURM，Kubernetes，Grafana 和 Prometheus。控制平面软件为虚拟 IPU 提供接口，将物理容器划分为多个虚拟容器。云服务供应商很高兴看到对多租户的支持，通过禁用虚拟容器之间的链接来隔离用户，与此同时，叠加网络提供了实例隔离。

图 6：Poplar SDK

Copyright ©2020 Moor Insights & Strategy



Poplar has been enhanced to support the IPU-Machine concept and a communications library (GCL) to support scale-out communications. Changes to Poplar are highlighted in blue.

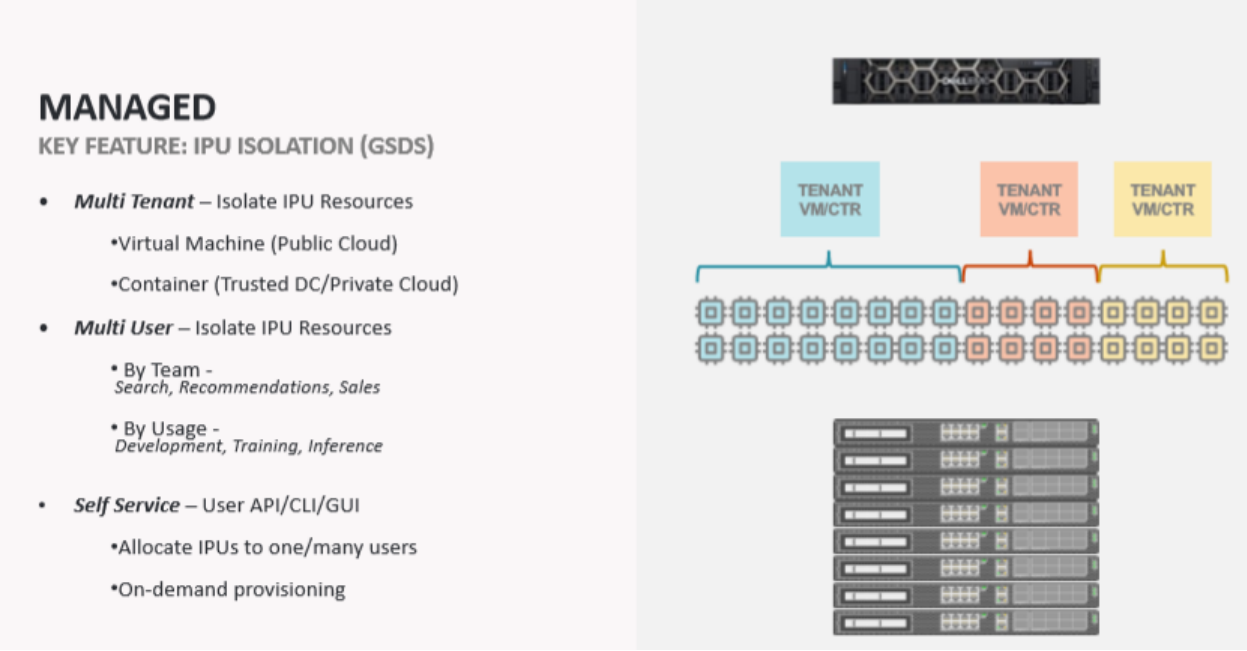
增强的 Poplar 支持 IPU-Machine 概念和通信库（GCL），从而对横向扩展通信加以支持。对 Poplar 的更改以蓝色标出。

Source: Graphcore
来源：Graphcore

Graphcore 还添加了自动自我发现和管理功能，以支持 IPU-Machine 每一单元的部署和执行。用于多租户和多用户使用模型的新自助服务接口和 API 应该有助于 Graphcore 渗透例如微软 Azure 这样的云服务提供商。微软 Azure 也是 Graphcore 的早期采用者和投资者。

图 7：IPU-POD 对多用户的支持

Copyright ©2020 Moor Insights & Strategy



API, CLI, and GUI interfaces are now available to support multi-tenancy in the IPU-Pods, critical for providing a secure and private shared infrastructure.

现在可以使用 API，CLI 和 GUI 接口来支持 IPU-PODs 中的多租户，这对于提供安全和私有的共享基础设施至关重要。

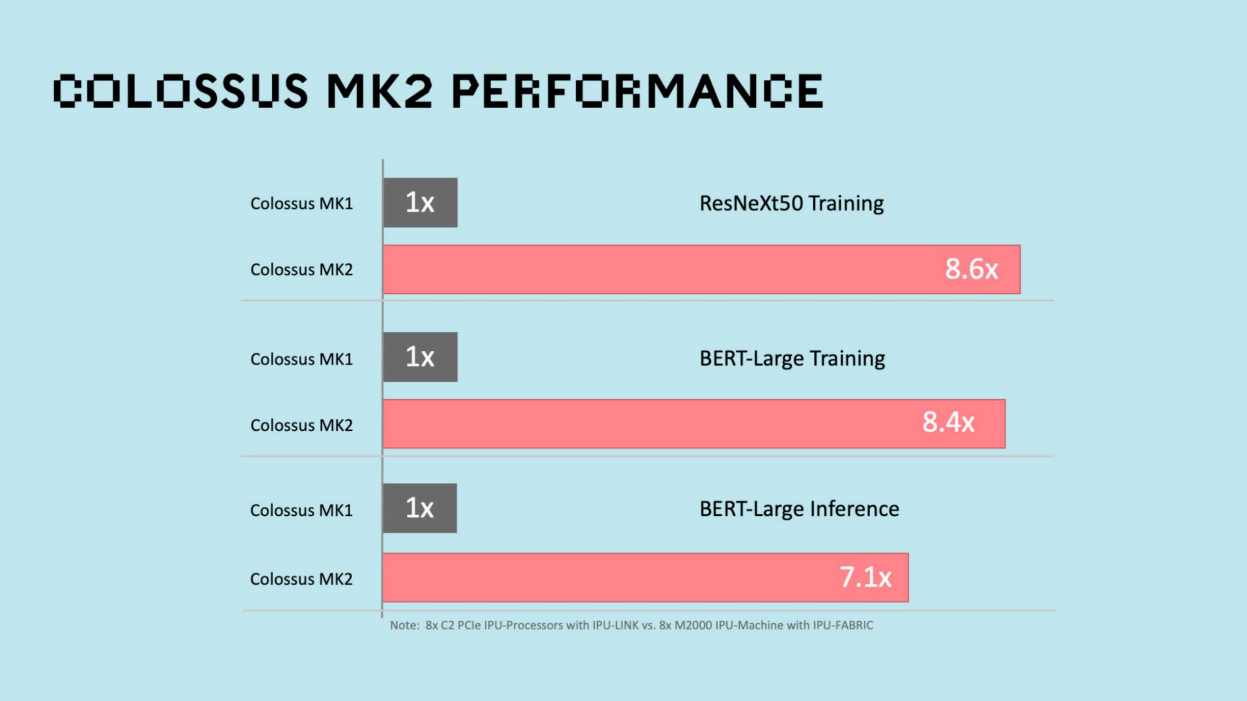
Source: Graphcore
来源：Graphcore

性能

Graphcore 表示，新的 4 芯片 IPU-Machine 在训练神经网络中的性能是 2 芯片前身 PCIe 卡的 7 至 9 倍以上，推理处理的性能则是 8 倍以上。因此，芯片与芯片之间的比较可能会表明 MK2 的性能提高了 3 到 4 倍。这些性能的提升已足以令人印象深刻，但在我们看来，用于较大模型的内存的大量增加，以及 IPU-Machine 和 IPU-Pod 的即插即用可扩展性、结构和管理更为令人惊叹。

图 8：Colossus MK2 的性能

Copyright ©2020 Moor Insights & Strategy



来源：Graphcore

结论

在 Poplar 和 IPU-Machine 提供的即插即用硬件和软件平台中提供新的 Colossus MK2 IPU 看起来是令 Graphcore 技术被广泛采用的绝佳策略。该公司资金雄厚，除拥有创新的处理器设计团队外，还拥有软件和系统工程团队。我们认为，它们这支稳定的专业队伍提供了从实际经验中得出的洞见，这些洞见将最大程度地减少团队遇到困境的情况，并增加了 Graphcore 技术的吸引力。Graphcore 工程团队仔细考虑了用户将要面对的开发、部署和优化任务。我们相信 Graphcore 是少数几个可以挑战 AI 加速现状并实质性推进并行处理和 AI 发展的创业公司之一。Graphcore 能否扩展到推理处理并最终扩展到边缘 AI，还有待进一步观察，这是因为在这些领域，对于任何将性能和规模视为基本原则的初创公司而言，成本效益和软件生态系统都可能构成它们的潜在障碍。

Copyright ©2020 Moor Insights & Strategy

有关此论文的重要信息

CONTRIBUTOR 投稿人

Karl Freund, Moor Insights & Strategy 高级分析师

PUBLISHER 出版人

Patrick Moorhead, Moor Insights & Strategy 公司创始人，总裁兼首席分析师

INQUIRIES 垂询

如果您想讨论此报告，请与我们联系，Moor Insights＆Strategy 会迅速做出回应。

CITATIONS 引用

本论文可以由经认可的媒体和分析人员引用，但必须在上下文中引用，并标示作者的姓名、职务以及“ Moor Insights＆Strategy”。非新闻和非分析人员必须获得 Moor Insights＆Strategy 的事先书面许可，方可进行任何引用。

LICENSING 许可

本文档（包括所有支持材料）归 Moor Insights＆Strategy 所有。未经 Moor Insights＆Strategy 事先书面许可，不得以任何形式复制，分发或共享本出版物。

DISCLOSURES 披露

本文是由 Graphcore 委托编写的。Moor Insights＆Strategy 为文中提到的许多高科技公司提供研究、分析、建议和咨询服务。我公司中没有任何雇员持有在本文中引用的任何公司的股权。

DISCLAIMER 免责

本文档中提供的信息仅供参考，可能包含技术上的不准确性、遗漏和印刷错误。 Moor Insights＆Strategy 对此类信息的准确性、完整性或充分性不做任何保证，并且对于此类信息的错误、遗漏或不足之处不承担任何责任。本文档包含 Moor Insights＆Strategy 的观点，不应解释为事实陈述。本文所表达的观点如有更改，恕不另行通知。

将预测和前瞻性陈述作为指示性指标，而不是对未来事件的精确预测。 尽管我们的预测和前瞻性陈述代表了我们在当下对未来前景的判断，但它们仍存在风险和不确定性，可能导致实际结果出现重大差异。 提醒您不要过分依赖这些预测和前瞻性陈述，这些预测和前瞻性陈述仅反映了本文发布之日时我们的观点。请记住，我们没有义务根据新信息或未来事件修改这些预测和前瞻性陈述，也没有义务公开发布任何修改结果。

©2020 Moor Insights＆Strategy 公司和产品名称仅供参考，可能是其各自所有者的商标。