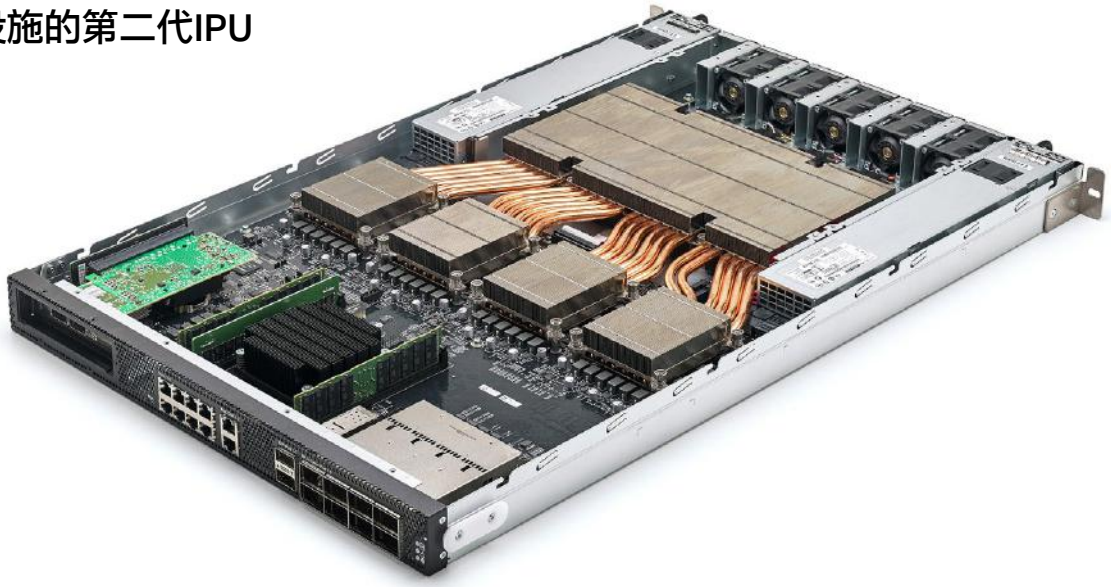
GRAPHCORE 拟未

IPU-MACHINE: M2000

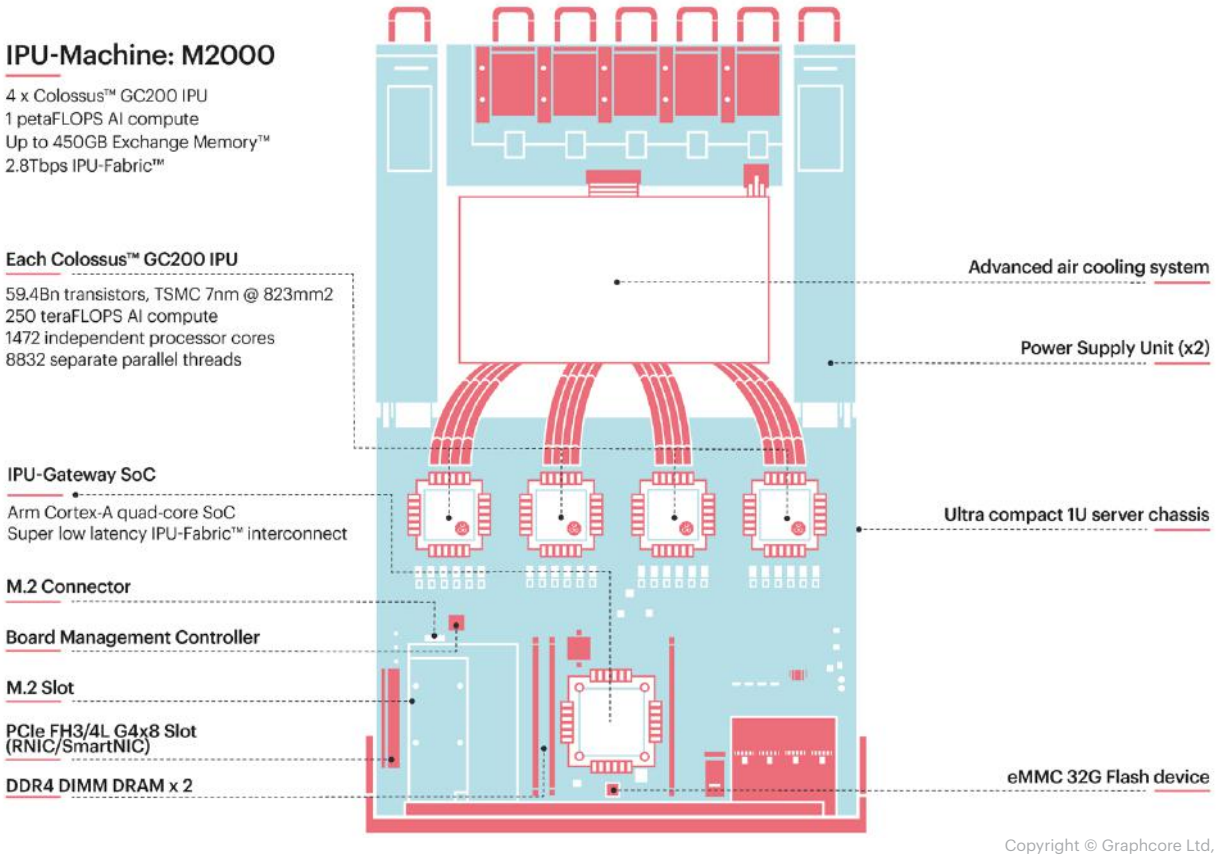
适用于大规模AI基础设施的第二代IPU系统



作为 AI 基础设施的核心和新的构建模块，IPU-M2000 由 4 个 Graphcore 第二代7纳米 IPU—— Colossus Mk2 GC200 驱动。它以纤薄的 1U 尺寸提供了 1 PetaFLOP的AI计算和高达450GB的Exchange-Memory™和2.8Tbps IPU-Fabric，可实现超低时延通信，以满足最苛刻的机器智能工作负载。

IPU-M2000具有灵活的模块化设计，提供从一个到数千个的扩展性。借助于内置的2.8Tbps高带宽以及接近零时延的内置于机箱的IPU-Fabric™互连架构，IPU-M2000可以作为一个独立的系统工作，或者8个堆叠在一起，也可以在IPU-POD$_{64}$系统中将16个紧密互连的IPU-M2000机架扩展到超级计算规模。

IPU-M2000专为高性能训练和推理工作负载而从零设计，可使将您的AI基础设施整合为一体，以最大程度地利用数据中心。您可以从开发和实验开始，然后扩大到全规模生产。IPU-M2000即日起接受预订。



GRAPHCORE 拟未

Poplar软件

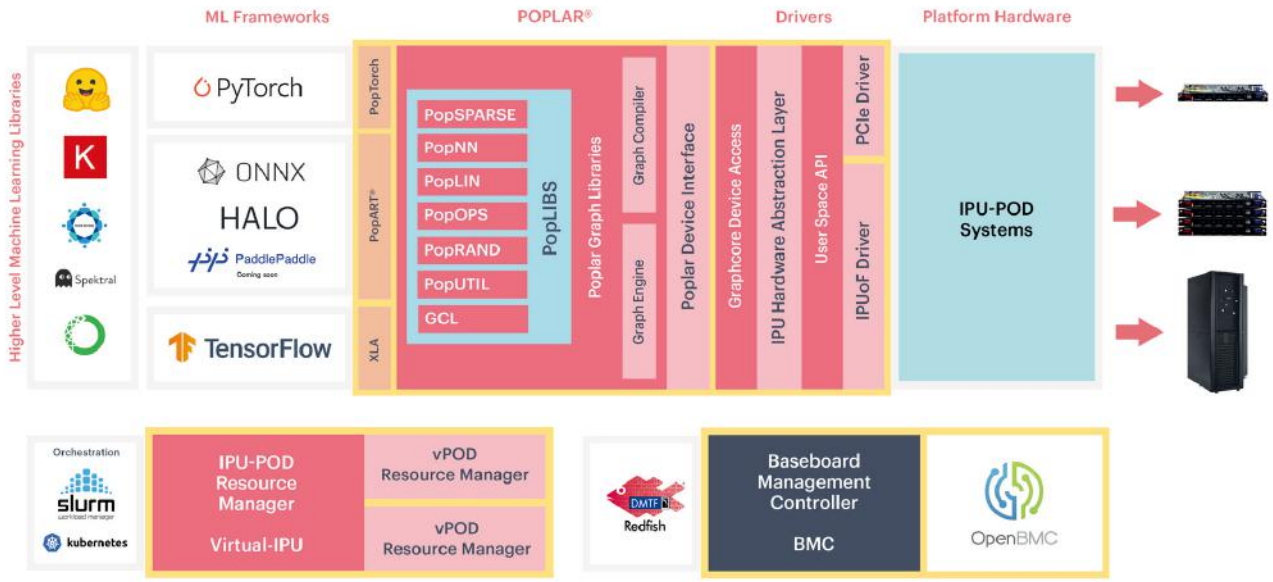
借助Poplar，管理大规模IPU就像对单个设备进行编程一样简单，使您可以只专注于数据和结果。

我们先进的编译器通过处理大型模型的所有调度和工作分区（包括内存控制），简化了IPU编程，而Graph Engine则构建了运行时，在你所有的可用IP-U、IPU-Machine或IPU-POD之间高效地执行工作负载。

除了跨大型IPU配置运行大型模型之外，使用Graphcore的Virtual IPU软件，动态共享您的AI计算也成为可能。在数十、数百甚至数千个IPU一起工作，进行模型训练的同时，您可以将其余的IPU-M2000机器分配给推理和生产部署。

Poplar支持包括TensorFlow、PyTorch、ONNX和PaddlePaddle在内的标准机器学习框架，以及行业标准的融合基础设施管理工具，因此易于部署，包括OpenBMC，Redfish，Docker容器，以及通过Slurm和Kubernetes进行的编排。而且，我们一直在增加对更多平台的支持。

您可以直接从Graphcore AI工程师那里获得专家支持，并帮您快速启动并运行。



IPU-M2000™的主要性能

计算

- 4个Colossus™ Mk2 GC200 IPU
- 1 PetaFlop AI计算
- 5888个独立的处理器核

存储

- 高达450GB的Exchange Memory™
- 180TB/秒的Exchange Memory™带宽

通信

- 2.8Tbps超低延迟IPU-Fabric™
- 直接连接或者通过以太网交换机连接
- 集合和全缩减操作支持

IPU Gateway SoC

- Arm Cortex quad-core A-series SoC

外形

- 行业标准1U

软件

- Poplar SDK
- PopVision可视化和分析工具

融合基础设施支持

- Virtual-IPU全面虚拟化和工作任务管理支持
- 支持SLURM工作任务管理
- 支持Kubernetes编排
- 内置的OpenBMC管理
- Grafana系统监控工具接口

想即刻开始？

通过联系info_china@graphcore.ai与我们的专家联系，以评估您的AI基础设施要求和解决方案的适用性。

GRAPHCORE.CN