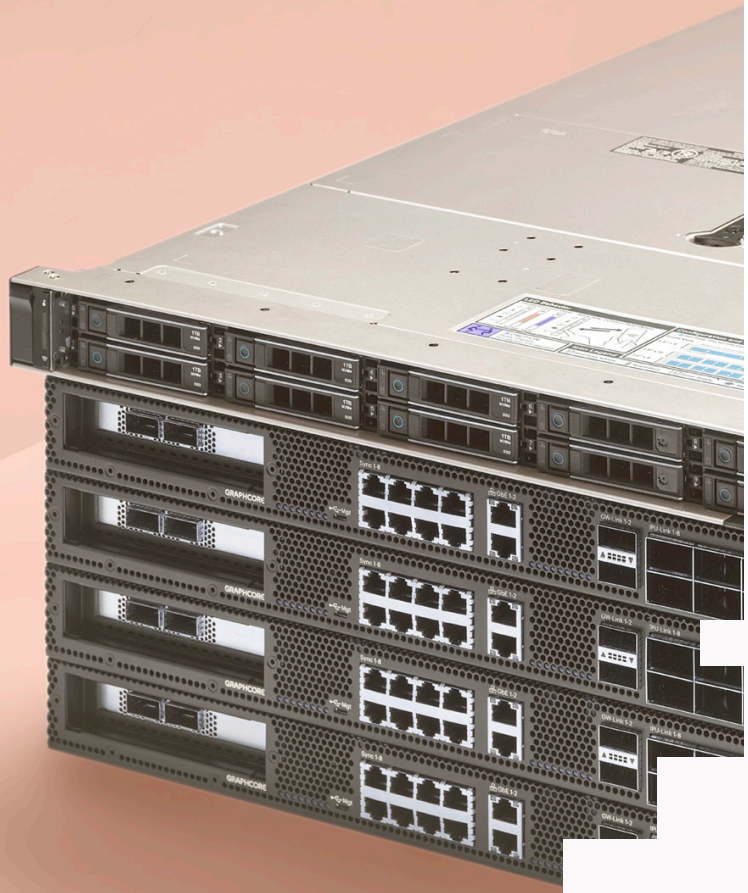
GRAPHCORE

BOW POD16



探索|构建|发展

Bow Pod系统为大规模机器智能部署提供高性能和高效率，旨在加速当今的大型复杂模型，同时也为创新者提供一个探索和发明未来解决方案的平台。

Bow Pod16系统非常适合探索。它提供了快速跟踪原型制作阶段、快速投入生产所需的所有功能、性能和灵活性。Bow Pod16是使用IPU构建更富创新的人工智能解决方案的不二之选，适用于任何机器学习领域，包括语言和视觉，探索GNN和LSTM，或开拓全新的领域。

最新一代IPU

Bow Pod16系统配备4台Bow-2000机器，每台机器都包含4个全新的开创性Bow IPU处理器。这一创新的IPU是世界上首款使用Wafer-on-Wafer（WoW）技术制造的处理器，将经过验证的IPU技术优势提升到一个新的水平。

性能和效率

Bow Pod16提供高达5.6 petaFLOPS的人工智能计算以及行业领先的效率，这一切都归功于在部署解决方案时使用了创新的硅技术、专注于效率和横向扩展的计算和存储架构，以及软件和应用优先的方法。

平滑部署和快速上市

整个系统，包括硬件和软件，已经被构建在一起。Bow Pod64支持所有标准的框架和协议，可以直接集成到现有数据中心环境以及私有云和公有云中。除了较短的系统聚合器配置和部署时间外，还有一系列已经经过测试和验证的、专为人工智能设计的、市场领先的服务器平台和高性能存储设备可供选择。创新者可以专注于使用熟悉的工具和框架大规模部署他们的人工智能工作负载，同时解锁前沿性能和效率。

系统规格

处理器	16个Bow IPU
1U刀片单元	4台Bow-2000机器
独立核心	23552个
线程	141312个
性能	5.6 petaFLOPS FP16.16 1.4 petaFLOPS FP32
存储	14.4 GB处理器内存储（In-Processor-Memory™）最高可达1 TB流存储（Streaming Memory™）

软件	Poplar® SDK
Host-Link	100 GE RoCEv2
系统重量	66千克+主机服务器和交换机
系统尺寸	4U+主机服务器和交换机
主机服务器	来自Graphcore®合作伙伴的经批准的主机服务器
储存	来自Graphcore合作伙伴的经批准的解决方案
散热	风冷

BOW POD16

定制计算的分解

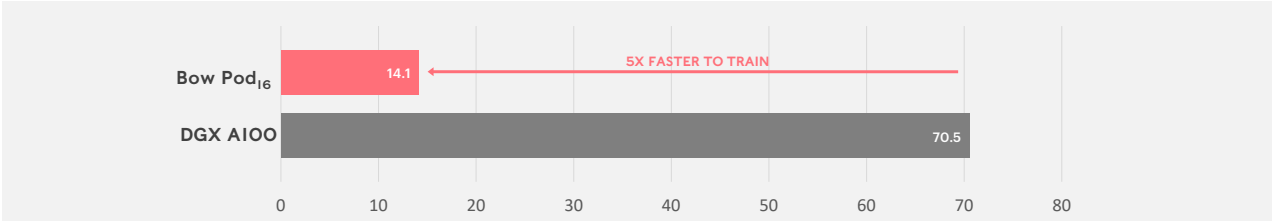
机器智能工作负载具有非常多样化的计算需求。对于生产部署，优化人工智能与主机计算的比率有助于最大限度地提高性能，同时改善总体拥有成本。Bow Pod系统允许将服务器和交换机的数量灵活映射到所需Bow-2000机器的数量，因此部署更适合生产人工智能工作负载。Bow Pod16支持多种服务器配置。

面向扩展构建的通信架构

高效的数据访问和传输可以解锁更高的人工智能性能。IPU-Fabric是一种创新的通信架构，用于系统范围的数据传输，在各个Bow IPU内、跨Bow-2000、Bow Pod之间和整个数据中心扩展高速互连。IPU-Fabric提供高性能、低时延的通信，以最大限度地提高人工智能应用程序效率，并与标准数据中心通信技术配合使用。

面向人工智能开发人员的平台

支持TensorFlow、PyTorch、PaddlePaddle和许多其他流行的机器学习框架，并作为开源提供，同时配有全面的PopLibs™库，用于社区驱动的协作和创新。对于希望完全控制以发挥最大性能的开发人员，Graphcore Poplar SDK支持使用C++直接进行IPU编程



EfficientNet-B4 Training Time To Train Performance
IPU-POD Platforms | Preliminary Results (Pre-SDK2.5) | G16-EfficientNet-B4 Training
DGX A100 (A100-SXM4-80GB) | TensorFlow | Mixed Precision | https://developer.nvidia.com/deep-learning-performance-training-inference

为大规模部署设计

带有Poplar SDK工具和框架图像的预构建Docker容器让创新者能够快速启动和运行。还支持用于容器编排、平台可视化和配置的各种通用框架，包括Slurm、Kubernetes和OpenStack。

软件优先

完全集成和IPU优化的Poplar软件利用IPU架构的独特特性来构建具有无与伦比的性能和灵活性的人工智能应用程序。Poplar支持在不增加开发复杂性的情况下轻松地将模型从一个IPU扩展到数千个，从而使创新者能够专注于应用程序的准确性和性能。

获取人工智能专业知识

Graphcore的人工智能专家和我们的精英合作伙伴网络在全球范围内为安装、生产和应用程序开发提供丰富的经验和支持。

准备好体验新一代的机器智能了吗？

与我们的合作伙伴联系，评估您的人工智能基础架构要求和解决方案的适用性。
如有问题，请直接通过info_china@graphcore.ai联系Graphcore