NVIDIA DGX-2
全球领先的深度学习系统助您应对最复杂的 AI 挑战

解决现代 AI 和深度学习需求的扩展难题

为了应对商业应用和科学研究中最紧迫的挑战，深度神经网络在规模和复杂性上发展迅速。传统的数据中心架构已无法满足支持现代 AI 工作负载所需的计算能力。新技术如增加模型并行使用率与 GPU 之间的带宽限制相冲突，因为开发人员构建了越来越大的加速计算集群，从而限制了数据中心的规模扩展。人们需要一种新方法来提供几近无限的 AI 计算规模，以便突破障碍，加速获取可以改变世界的见解。

变不可能为可能的训练性能

日益复杂的 AI 渴求前所未有的计算水平。NVIDIA® DGX-2™ 是世界上第一个 2 petaFLOPS 系统，配备 16 块极为先进的 GPU，得以为先前无法训练的最新深度学习模型类型提供加速。凭借开创性的 GPU 规模，您可以在单个节点训练 4 倍规模的模型。与传统的 x86 架构相比，DGX-2 训练 ResNet-50 的性能相当于 300 台配备双路英特尔至强 Gold CPU 服务器的性能，而后者成本超过 270 万美元。

NVIDIA NVSwitch - 革命性的 AI 网络结构

前沿研究要求自由地利用模型并行性，并且需要前所未有的 GPU 间带宽。NVIDIA 开发了 NVSwitch 以解决这一需求。正如从拨号上网到超高速宽带的革新，NVSwitch 把属于未来的网络结构带到了今天。有了 NVIDIA DGX-2，模型的复杂性和规模不再受传统架构限制的约束。在 DGX-2 中采用网络结构进行模型并行训练，可提供 2.4TB/ 秒的对分带宽，比前几代增加 24 倍。这种新的互连“超高速公路”为模型类型赋予了无限可能，现在用户可同时在 16 块 GPU 间进行分布式训练，强大的计算能力得以最大程度地释放出来。

系统规格	
GPUs	**16块 NVIDIA® Tesla V100**
GPU 显存	**共 512GB**
性能	**2 petaFLOPS**
NVIDIA CUDA® 核心数量	**81920**
NVIDIA Tensor 核心数量	**10240**
NVSwitches	**12**
最大功率	**10 kW**
CPU	**双路英特尔至强 Platinum 8168, 2.7 GHz, 24 核**
系统内存	**1.5TB**
网络	**8X 100Gb/秒 Infiniband/100GigE 双 10/25Gb/秒 Ethernet**
存储空间	**操作系统：两块 960GB NVME SSD 内部存储：30TB (8块 3.84TB) NVME SSD**
软件	**Ubuntu Linux OS 如需详细信息，请参阅软件堆栈**
系统重量	**154.2 千克**
系统尺寸	**高：440 毫米 宽：482.3 毫米 长：795.4 毫米 - 无框架 834 毫米 - 有框架**
运行温度范围	**5°C 至 35°C**

更大规模的 AI

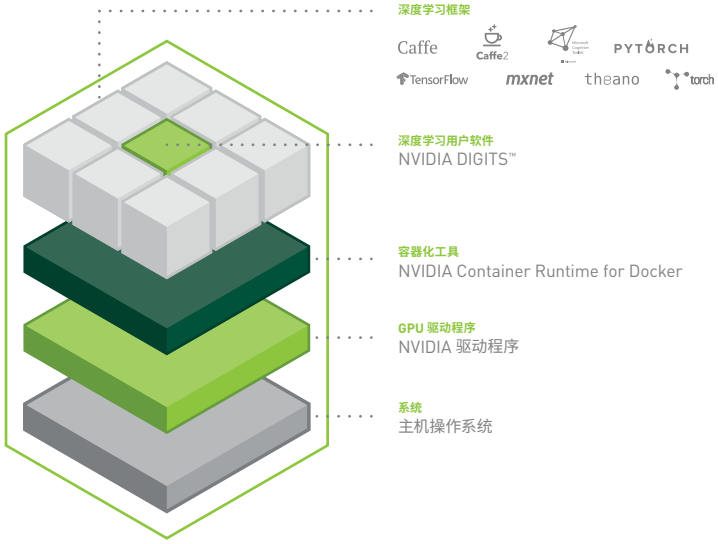
AI 现代企业需要根据业务需求迅速部署 AI，同时还需要在不增加成本或复杂性的前提下扩展 AI 规模。我们构建了 DGX-2，并为其配备了 DGX 软件，从而实现大规模的加速部署和操作简化。DGX-2 提供的解决方案是实现扩展 AI 最快的路径，加上虚拟化支持，使您可以建立自己的企业级 AI 云。现在，企业可以在一个解决方案中充分利用不受限制的 AI 算力，这种解决方案轻松地扩展了将加速计算资源结合在一起所需的网络基础设施部分。利用加速部署模型和专门为易扩展性构建的架构，您的团队可以将更多的时间用于获取见解，而节省花费在构建基础设施的时间。

企业级 AI 基础设施

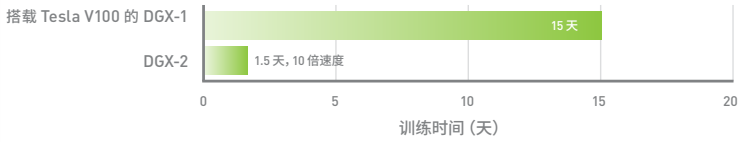
如果 AI 平台对您的业务至关重要，那么您需要考虑使用专为可靠性、可用性和可服务性 (RAS) 设计的平台。企业级的 DGX-2 专为严苛的全天候 AI 业务和 RAS 而构建，可减少非计划停机时间、简化维护，及保持运行的持续性。

节省调试和优化时间，增加专注于探索的时间。NVIDIA 企业级支持让您无需耗费时间对硬件和开源软件进行问题排查。借助每一个 DGX 系统，用户可利用包括软件、工具和 NVIDIA 专业知识的集成解决方案，更快地入门、训练和运行。

NVIDIA DGX-1 软件堆栈



NVIDIA DGX-2 可将深度学习训练速度提高 10 倍



测试网络: FairSeq，55 epochs，PyTorch。

如需了解更多信息，请访问 www.nvidia.cn/DGX-2