白皮书

NVIDIA DGX-1 系统架构

超快速的深度学习平台

目录

1.0 简介 2
2.0 NVIDIA DGX-1 系统架构 3
2.1 DGX-1 系统技术 5
3.0 NVLink 和 InfiniBand 支持的多 GPU 及多系统扩展 6
3.1 DGX-1 NVLink 网络拓扑可实现高效的应用程序扩展 7
3.2 借助 NVLink 扩展深度学习训练 10
3.3 借助 InfiniBand 实现 DGX-1 多系统扩展 13
4.0 DGX-1 软件 16
4.1 NVIDIA CUDA Toolkit 17
4.2 NVIDIA Docker 18
4.3 NVIDIA 深度学习 SDK 19
4.4 NCCL 20
5.0 DGX-1 适用的深度学习框架 22
5.1 NVIDIA Caffe 22
5.2 Microsoft Cognitive Toolkit 23
5.3 MXNet 23
5.4 TensorFlow 24
5.5 Theano 25
5.6 Torch 25
5.7 DIGITS 26
6.0 结果：为实现最高深度学习性能而打造的 DGX-1 27

摘要

NVIDIA® DGX-1™（图 1）是专为深度学习打造的集成系统。DGX-1 配有 8 个 NVIDIA® Tesla® P100 GPU 加速器，在混合立体网络中通过 NVIDIA® NVLink™（NVIDIA 高性能 GPU 互联技术）互相连接。DGX-1 可搭配双插槽 Intel Xeon CPU 及四块 100 Gb InfiniBand 网络适配器使用，为深度学习训练提供惊人的强大性能。此外，DGX-1 系统软件及强大软件库均经过调试，可在 Tesla P100 GPU 网络中扩展深度学习，同时在生产和研究领域提供灵活、可扩展的深度学习应用平台。

图 1　NVIDIA DGX-1

1 简介

深度学习正迅速改变计算机科学领域的发展，并对 Google [Metz 2015]、Facebook [Statt 2016] 及 Amazon [Finley 2016] 等大型企业的经济因素产生了巨大影响。即使此新兴学科与技术已成为主流，其迅猛发展之势仍然丝毫不减。为服务日益复杂的应用程序及提供高质量结果，神经网络正呈指数地不断深化和复杂化。与此同时，主流交互式应用程序中部署的神经网络也受到推动，以实现更快速的推理和结果预测。

因此，深度学习性能需求也在日益增长。Facebook CTO Mike Schroepfer 近日指出：

- Facebook 部署的神经网络每秒可处理 6 百多万预测
- Facebook 当前 25% 的工程师正使用 AI 和机器学习 API 及基础架构
- Facebook 内部部署有计算能力超过 40 PFLOP/s 的 GPU，以充分支持组织内的深度学习 [Schroepfer 2016:6:54]

在 2016 年 4 月的一次采访中，百度首席科学家吴恩达表示，百度在机器学习方面持 HPC （High Performance Computing，高性能计算）相关观点：

> *网络训练速度越快，数据集及模型就可进行更多迭代，而迭代次数越多，机器学习就越先进。[Ng 2016a]。*

吴恩达指出，训练百度的某个语音模型即需 10 exaflops 的计算能力 [Ng 2016b:6:50]。

神经网络的不断深化与复杂化会极大地提升准确性，例如微软的深度残差网络 (Deep Residual Networks) [He et al.2015])，但训练此类高度准确的神经网络需要大量计算时间，而网络的复杂性亦增加了预测延迟。

NVIDIA DGX-1（图 1）是一款新的人工智能 (AI) 超级计算机，旨在满足深度学习训练对性能永无止境的巨大需求。NVIDIA 对 DGX-1 的期望是，打造世界上超快速的深度神经网络训练平台，为深度学习研究者及数据科学家实现网络快速部署和即插即用的特点。DGX-1 架构汲取了 NVIDIA 在高性能计算领域的丰富经验，以及与各大重要云服务提供商和众多《财富》1000 强公司通力合作所获得的在 NVIDIA GPU 上深度学习框架优化的相关知识。

2 NVIDIA DGX-1 系统架构

NVIDIA® DGX-1™ 是一款深度学习系统，专为实现高吞吐量和高互联带宽而构建，可最大程度提升神经网络训练性能。系统核心组件是 8 个 Tesla P100 GPU ，通过混合立方 NVLink 网络拓扑互联，如第 3 节中所述。除了这 8 个 GPU 之外，DGX-1 还内置两块 CPU 以供启动、存储管理及深度学习框架协调之用。DGX-1 内嵌于 3 机架单位 (3U) 机箱中，集电源、冷却、网络、多系统互联及 SSD 文件系统缓存于一体，各组件相互协调，为您优化吞吐量和深度学习训练时间。

NVLink 是高带宽、低能耗的互联技术，借助该技术，可实现 NVIDIA® Pascal™ 架构 GPU 与同代 GPU 间或节点内其他设备间的互联，每个 GPU 的双向总带宽可达 160 GB/s，约为当前 PCIe Gen3 x16 互联的 5 倍。NVLink 互联及 DGX-1 架构的混合立体 GPU 网络拓扑，可在一组 8 个 Tesla P100 GPU 间实现最高带宽的数据交换。

Tesla P100 分页转移引擎可实现 GPU 与大容量主机内存间高带宽、低用度的数据共享 [NVIDIA Corporation 2016]。同时，DGX-1 通过 InfiniBand (IB) 网络提供系统至系统的高带宽（详见第 3.3 节）扩展至多节点高性能集群。

图 2 为 DGX-1 系统组件图。

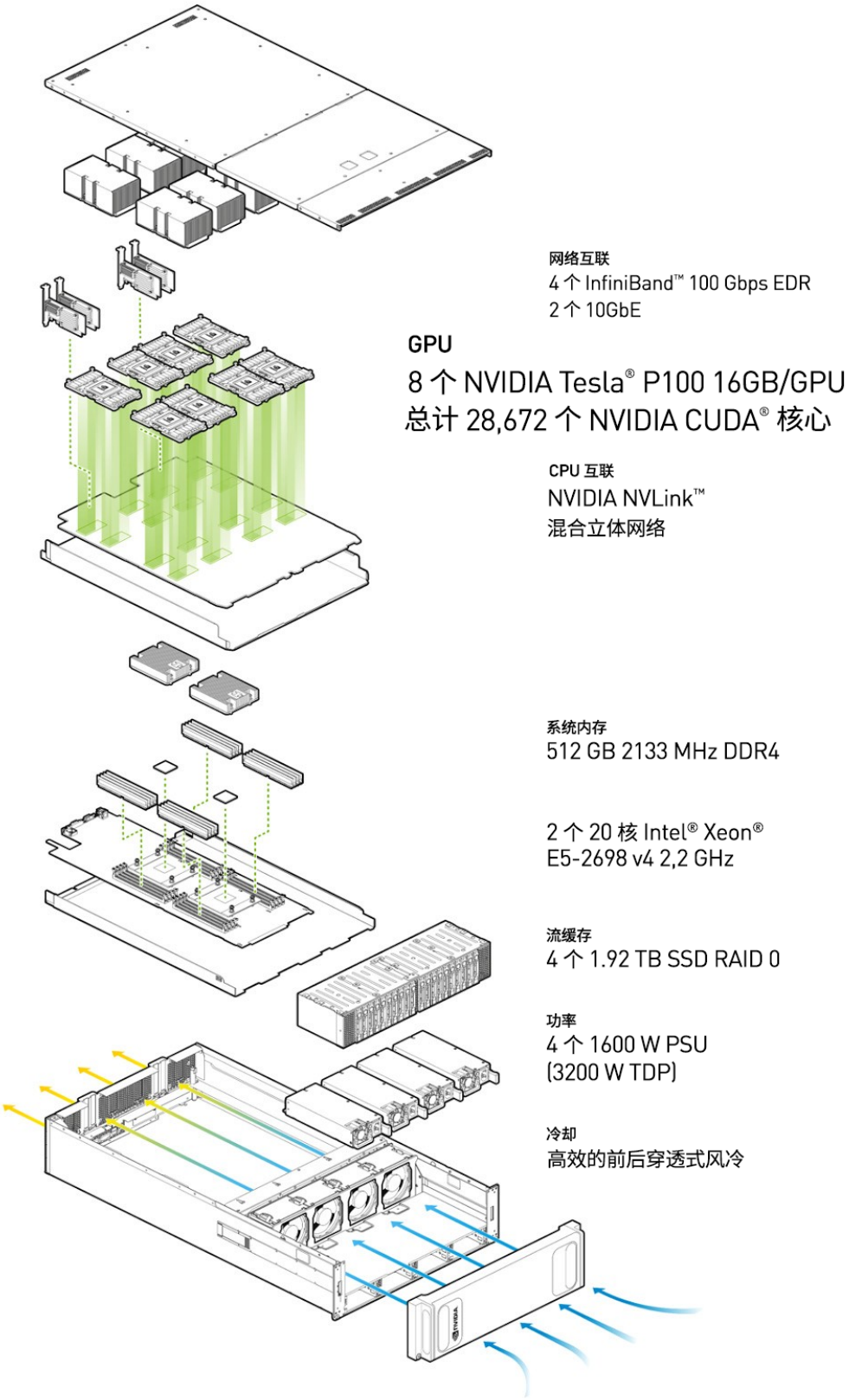
网络互联
4 个 InfiniBand™ 100 Gbps EDR
2 个 10GbE
GPU
8 个 NVIDIA Tesla® P100 16GB/GPU
总计 28,672 个 NVIDIA CUDA® 核心
CPU 互联
NVIDIA NVLink™
混合立体网络
系统内存
512 GB 2133 MHz DDR4
2 个 20 核 Intel® Xeon®
E5-2698 v4 2,2 GHz
流缓存
4 个 1.92 TB SSD RAID 0
功率
4 个 1600 W PSU
(3200 W TDP)
冷却
高效的前后穿透式风冷

图 2　DGX-1 系统组件

2.1 DGX-1 系统技术

Tesla P100（图 3）是 NVIDIA 最新款加速器，专为高性能计算及深度学习应用程序而设计 [NVIDIA Corporation 2016]。P100 配有的 GP100 GPU 包含 56 个流多元处理器 (SM)，每个处理器均包含如下配置：

- 64 个单精度 (FP32) NVIDIA® CUDA® 核心
- 32 个双精度 (FP64) CUDA 核心
- 256KB 寄存器堆 (RF)
- 64KB 共享内存

Tesla P100 峰值计算吞吐量为：

- 双精度性能为 5.3 TFLOP/s
- 单精度性能为 10.6 TFLOP/s
- 半精度性能为 21.2 TFLOP/s [NVIDIA Corporation 2016]

图 3　Tesla P100 加速器

为了支持如此高的计算吞吐量，Tesla P100 是首个搭载 HBM2 技术（第二代高带宽内存技术）的 GPU 架构。P100 内置 16 GB HBM2 堆栈式内存，带宽为 732 GB/s，远超传统 GPU 使用的 GDDR5 RAM 带宽。HBM2 堆栈式内存与 GPU 位于同一物理封装内，相对于传统的 GDDR5 而言能够节省大量空间。节省下的空间使我们能够打造出类似 DGX-1 的高密度 GPU。

每个 Tesla P100 拥有 4 个 NVLink 连接，双向总带宽达 160 GB/s。NVLink 与 DGX-1 互联拓扑及其作用详见第 3 节。

GPU 与 CPU 间的 PCIe 连接可访问 CPU 大容量 DRAM 内存，从而为工作集和数据集实现通向或来自 GPU 的流式传输。CPU 内存容量经配置为 GPU 的 4 倍，可简化深度学习工作负载的缓冲管理并实现平衡。通常情况下，两倍 GPU 内存占用足以管理后台数据移动和后台缓冲；四倍 GPU 内存占用则可在管理内存工作集及数据流移动方面获得更大的灵活性。除 512 GB 的系统 DRAM 外，每个 Tesla P100 GPU 还包含 16 GB 的 HBM2 复合封装内存，总计有 128 GB HBM2 内存，GPU 内存净带宽为 8 × 732 GB/s，即 5.86 TB/s。

DGX-1 配有以太网和 InfiniBand (IB) 两种网络接口。两个万兆以太网接口可方便远程用户存取系统和远程存储。为连接多个 DGX-1 系统，每个系统均配备四个高带宽、低延迟的 EDR IB（扩展数据速率 InfiniBand）端口，总通信传输速率达到 400 Gb/s。每个 DGX-1 系统配有 8 个 P100 GPU 和 4 个 EDR IB 端口，即每两个 GPU 对应一个 IB 端口。此外，EDR IB 支持 NVIDIA GPUDirect 远程直接内存存取 (RDMA) 协议，数据可直接从某一系统的 GPU 内存传输至另一系统的 GPU 内存，无需涉及 CPU 或系统内存。

对于深度学习系统 DGX-1 的性能而言，高效、高带宽的训练数据流至关重要，而可靠、低失败率的存储也非常重要。每个系统均配有一个 480 GB boot OS SSD，四个 1.92 TB SAS SSD（共计 7.6 TB）作为 RAID-0 磁盘组，可确保出色性能。对于大于 7 TB 的工作集，数据将在一个训练回合内通过 SSD 于后台得到暂存及缓冲。

除了高计算吞吐量及内存带宽，训练深度神经网络还需要磁盘存储器拥有高数据流速率。DGX-1 专为支持数百千兆字节至数兆兆字节等大型训练数据集而设计。四个 SAS SSD 共可提供 2 GB/s 流式数据读取带宽，可确保训练期间不会出现流式传输磁盘瓶颈。

DGX-1 的热设计功耗 (TDP) [1] 为 3.2KW，实际功耗根据工作负载呈动态变化。为方便部署，在多数数据中心环境中，进风温度为 5°C 至 35°C，DGX-1 采用空气冷却。

3 NVLink 和 InfiniBand 支持的多 GPU 及多系统扩展

随着越来越多的开发者公布、利用其应用程序中的可用并行性，越来越多的服务器采用一个 CPU 对应两个或两个以上 GPU 的设计。密集型 GPU 系统为单节点性能扩展提供绝佳方式，多 GPU 应用程序效率受 GPU 间 PCIe（Peripheral Component Interconnect Express，外设组件互连传输通道）总线连接的制约。同样地，随着数据中心应用程序日益革新，多个互联系统也需要高效扩展。为应对上述需求，DGX-1 采用全新 NVLink 高速 GPU 互联技术，可在一个系统内实现多 GPU 扩展，还有多个 EDR InfiniBand 接口，可在多个互联的 DGX-1 系统间提供超高带宽。

考虑到通信运营成本较高，开发者必须通过计算重叠数据传输或通过 PCIe 互联仔细编排 GPU 访问以最大化提升性能。随着 GPU 运行速度越来越快，以及 GPU/CPU 比率的上升，更高性能的 GPU 互联技术也得以出现。

1. TDP 对运行应用程序时真实的最大系统功率做了规定，数据中心也以 TDP 作为功率及冷却要求的正确衡量标准。

这项挑战促使了 NVLink 高速互联技术的问世，该技术可实现 NVIDIA Pascal GPU 与同代 GPU 或支持 NVLink 的 CPU 以及节点内其他设备之间的连接。NVLink 支持 GPU ISA，这意味着运行于以 NVLink 连接 GPU 的项目，可基于其他 GPU 内存数据及本地内存数据而直接执行。GPU 还可在远程 GPU 内存地址上执行原子存储操作，以实现更加紧密的数据共享与优化的应用程序扩展。

NVLink 使用 NVIDIA 全新高速信号互联技术 (NVHS)。NVHS 通过差分对传输数据，速率高达 20 Gb/s。其中 8 个差分连接组成“子链路”（子链路负责一个方向的数据传输），两个子链路（一个子链路对应一个方向）组成一个 “链路”（一个链路可连接两个处理器，如 GPU 和 GPU 或 GPU 和 CPU）。单个链路支持端点间高达 40 GB/s 的双向带宽。多个链路可整合至一起，以实现处理器间更高的带宽。Tesla P100 采用的 NVLink 可支持多达 4 个链路，实现理论上的最大双向总带宽，即 160 GB/s。

3.1 DGX-1 NVLink 网络拓扑可实现高效的应用程序扩展

通常情况下，高性能应用程序通过以下两种方式扩展计算能力，即*强扩展和弱扩展*[2]。应用在固定问题总大小的并行处理器数量增加时，扩展性越强，解决问题的时间就会越短。理想情况下的强扩展性，获得的加速与所使用处理器的数量相当。

另一方面，应用在*每处理器*固定大小问题的并行处理器数量增加时，扩展性越弱，解决问题的时间就会越短。换而言之，问题大小随处理器数量的增长而增长。因此随着问题大小（及处理器数量）增长，执行时间趋于相对稳定。理想情况下的弱扩展性，问题大小及处理器数量按比例呈线性增长，但解决方案时间未增长。

随着单个处理器及处理器集群变宽（拥有更多并行处理组件），部分问题的弱扩展效益逐渐减退，直至最终这些问题可能结束并行。在此情况下，这些问题不得不进入强扩展。但实际情况是，大多数应用程序呈现一定程度的强扩展时，通常并非完全的线性增长。

其中一个主要原因就是通信成本。在不断增加的处理器上强扩展某个问题，则每个处理器工作量将不断减少，同时相对增加了这些处理器的通信成本。在强扩展体系中，为这些互联所调整的快速互联技术和通信原语至关重要。

2. 请注意，这里 “弱” 并非指 “强” 的下级形式，两者均为重要的实际衡量标准。

为了实现极限的超高计算密度，DGX-1 内置 8 个 NVIDIA Tesla P100 加速器。此类以多个高度并行 GPU 为基础的应用程序扩展，目前的限制在于 PCIe 互联技术。NVLink 可提供能够实现出色深度学习和其他应用程序（强和弱）扩展所需的通信性能。每个 Tesla P100 GPU 拥有四个 NVLink 连接点，均以点到点的方式连接至其他 GPU，峰值带宽可达 20 GB/s。多个 NVLink 连接可整合使用，成倍增加特定两个 GPU 间的可用互联带宽。这样 NVLink 便可提供灵活的互联，从而构建多个 GPU 间的各类网络拓扑。Pascal 还支持 16 通道的 PCIe 3.0。这些总线在 DGX-1 中负责 CPU 与 GPU 间的连接。PCIe 还应用于高速网络接口卡。

DGX-1 的 NVLink 网络拓扑设计旨在优化一系列影响因素，包括各种点到点及群集通信原语可获得的带宽、拓扑结构灵活性及其在 GPU 子集中的性能表现。在设计期间，NVIDIA 工程师就各应用程序的强扩展与弱扩展进行了建模预测，如深度学习、分类、快速傅立叶变换 (FFT)、分子动力学、图形分析、计算流体动力学、地震成像及光线追踪等。此白皮书重点分析深度学习训练扩展。

混合立体网络拓扑（图 4）可以看作一个立方体，8 个角为 8 个 GPU，12 条边代表 NVLink 连接，6 个面中有 2 个面的对角线互相连接。还可以看作两个交织的环，每个环代表一个 NVLink 连接。

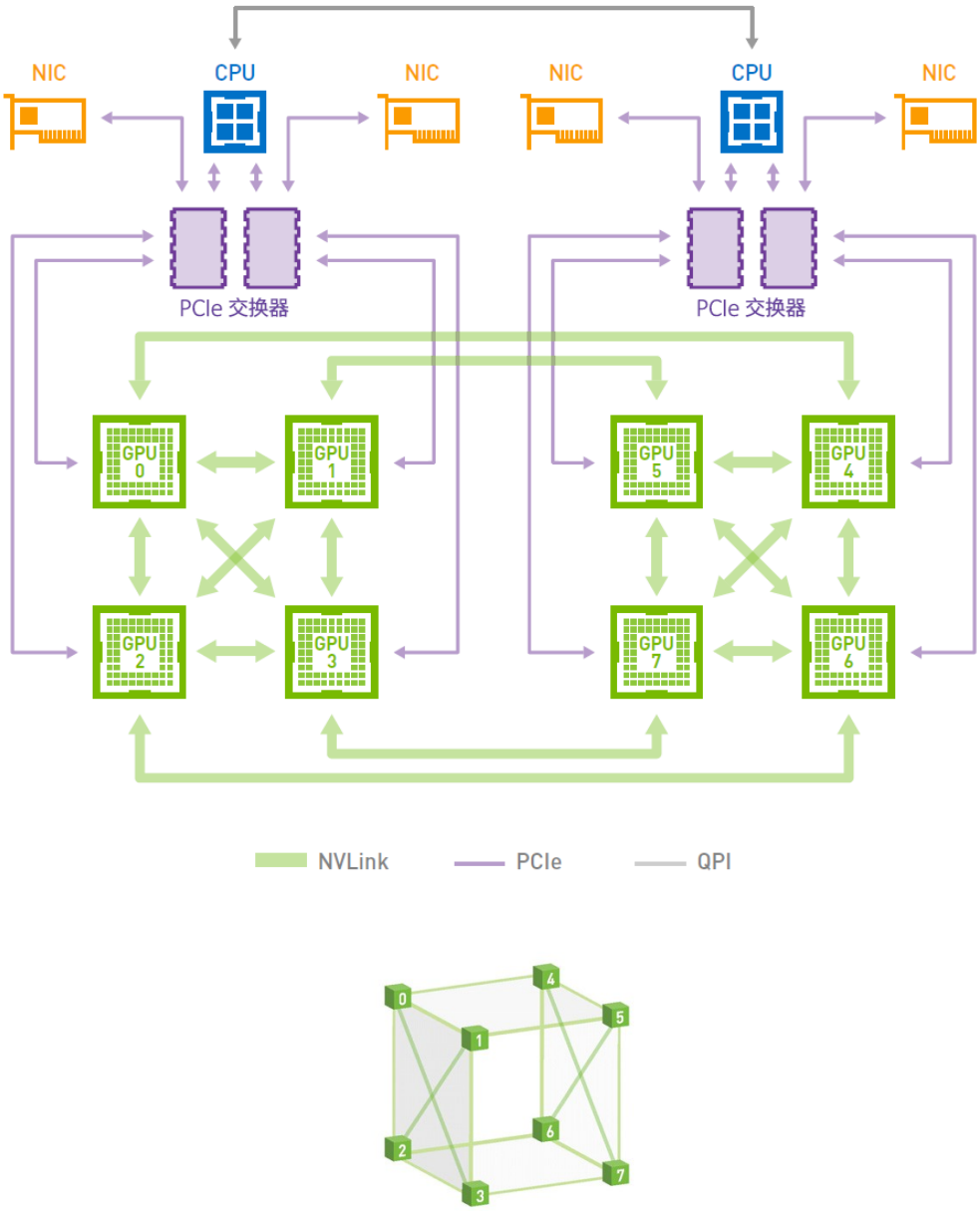

图 4　　DGX-1 采用 8-GPU 的混合立体互联网络拓扑。
立体网状连接面的各个角与 PCIe 树形网络相连，通过后者实现与 CPU 和 NIC 的连接。

立体网络拓扑可提供超高带宽，适用任何 8-GPU NVLink 拓扑，可实现多集群通信原语，其中包括对深度学习而言至关重要的 broadcast、gather 及 all-gather。采用 NVLink 连接以跨越两个 GPU 集群（四个 GPU 为一个集群）可减轻 PCIe 总线及 CPU 间 SMP 连接的压力，以及避免在两集群传输时出现系统内存中的分段传输现象。此外，此 8-GPU 立体网络拓扑还可轻松再分为两个完整的半互联结构，或是以 NVLink 相连的一对 GPU 。灵活性在多应用程序共享同一服务器时至关重要。

3.2 借助 NVLink 扩展深度学习训练

深度神经网络会学习多个抽象层级，既有简单也有复杂概念。深度模型的优势在于性能强大且可不断学习。深度神经网络的训练方式为，向其输入大量数据经过层层计算，以生成与已知正确答案进行比较的输出数据。计算输出误差之后，误差将通过反向传播在网络中回传。在每一步反向中，模型参数会按照同一目的进行调整，即尝试使用名为随机梯度下降 (SGD) 的数值优化方法减少误差。这一程序会对数据进行全面处理，同时改进模型。

在多 GPU 和（或）多节点间并行训练深度神经网络，需要分配以下任一项：

- 输入数据（“数据并行”）：在数据并行方法中，单个并行任务必须相对在反向传播中计算的模型，定期重新同步其梯度，以使模型参数在并行任务中保持同步。这相当于 all-reduce 操作。
- 受训练的模型（“并行模型”）：在模型并行方法中，或是一次选择一个并行任务相对输入数据传播其梯度，或是对数据（连续子集）梯度使用 all-gather 操作，以使所有 GPU 的传出带宽并行运作。
- 上述两者的结合 [Wu et al.2015] [Krizhevsky 2014]。

在弱扩展深度神经网络训练中，随着 GPU 数量增长，全局有效 SGD 最小批量大小也会增加。并不出人意料的是，弱扩展方式拥有更高的并行效率，即使 GPU 间互联相对缓慢也是如此。然而，在 SGD 优化收敛受负影响之前，最小批量大小仅可扩展至一定水平。各网络对增加的弱扩展性的相对误差因各网络而异。

下图为 DGX-1 训练性能扩展图示，图 5 和 图 6 中的柱形图表示 ResNet-50 深度神经网络架构使用 Microsoft Cognitive Toolkit (CNTK) 的每秒图像训练性能，斜线代表 2、4 或 8 个 P100 GPU 对比单 GPU 的并行加速。图 5 表示每 GPU 的最小批量大小为 64 张图像时的弱扩展性能，而图 6 表示总最小批量大小为 64 张图像（在所有活动 GPU 之间分配）时的强扩展性能。

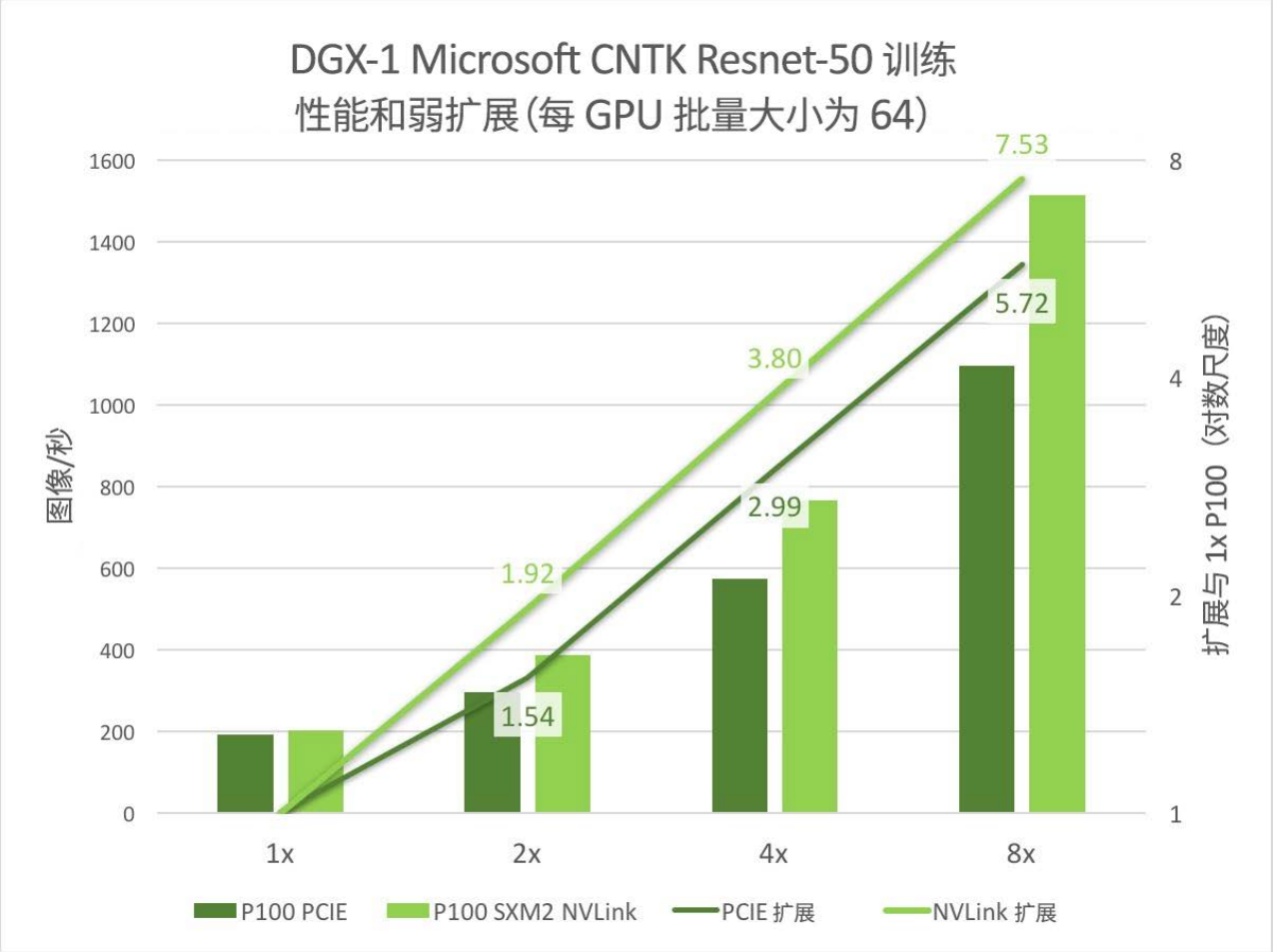

图 5　当每 GPU 的批量大小为 64 时，DGX-1 使用 Microsoft Cognitive Toolkit (CNTK) 训练 ResNet-50 神经网络架构时的弱扩展结果和性能。该柱形图表示 DGX-1 分别配备 1、2、4 和 8 个 Tesla P100 GPU 并采用 NVLink 作为 GPU 间通信方式时的性能表现（浅绿色），与配备 8 个 Tesla P100 GPU 的现成系统采用 PCIe 通信方式时的性能表现（深绿色）的对比情况。这些折线表示相较单一 GPU 的速度提升情况。如果使用 8 个 GPU，NVLink 的训练性能约为 PCIe 的 1.4 倍（每秒 1513 张图像对比每秒 1096 张图像）。测试采用 NVIDIA DGX 容器版本 16.12，在 cuDNN 6.0.5、NCCL 1.6.1、gradbits=32 的条件下处理真实数据。

如图 5 所示，对比配备 1、2、4 和 8 个 GPU 采用 PCIe（树状拓扑）通信方式的深度学习训练与 DGX-1 的 8-GPU 混合立体网络 NVLink 互联方式，NVLink 在量产型 DGX-1 硬件方面表现出明显的优势。在配备 4 个和 8 个 GPU 的情况下，NVLink 的表现非常出众，这里 DGX-1 能够以 PCIe 无法做到的方式聚集多个 NVLink 连接，总速度提升相当于 PCIe 的 1.4 倍左右。DGX-1 架构的 NVLink 互联技术能实现比 PCIe 更优的弱扩展，而且与其他 NVLink 网络配置（比如环型拓扑）相比，NVLink 混合立体网络拓扑技术能针对深度学习提供最好的整体弱扩展。

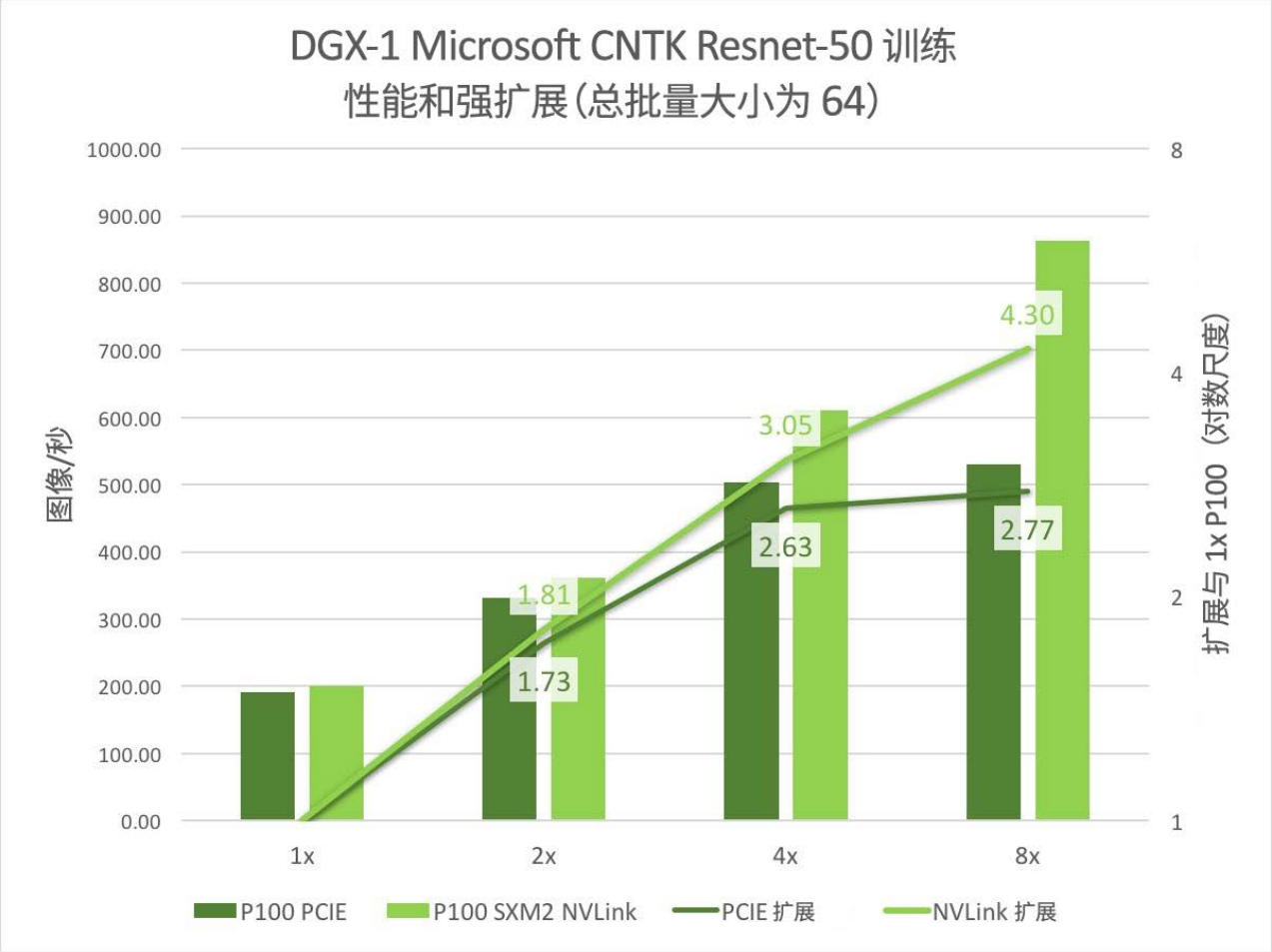

图 6　DGX-1 分别配备 1、2、4 和 8 个 Tesla P100 GPU 且采用 NVLink 互联技术时的 ResNet-50 训练性能和强扩展速度提升情况，相较于现成系统配备相同数量的 P100 GPU 且采用 PCIe 互联技术时的扩展情况。训练采用专为 DGX-1 优化的 Microsoft Cognitive Toolkit (CNTK)，总批量大小为 64 张图像。柱形图表示训练吞吐量，以每秒处理的图像数为单位，折线表示与单一 P100 GPU 对比的强扩展速度提升。NVLink 的扩展效率更高，相较采用 8 个以 PCIe 方式相连的 P100 GPU 的系统，DGX-1 的训练性能可提高至 1.6 倍以上（863 张图像/秒对比 531 张图像/秒）。测试采用 NVIDIA DGX 容器版本 16.12，在 cuDNN 6.0.5、NCCL 1.6.1、gradbits=32 的条件下处理真实数据。

在强扩展中，全局有效的最小批量大小固定不变，与 GPU 的数量多寡无关。强扩展中的并行效率对通信带宽更加敏感。在这种情形下，将 CNTK 上的 ResNet-50 训练以 64 总批量大小扩展为 8 个仅通过 PCIe 连接的 GPU 会有很大损失，8-GPU-PCIe 系统的并行效率不到 35%。相比之下，DGX-1 架构的 8-GPU NVLink 混合立体拓扑结构在同一测试中可实现约 54% 的效率，如图 6 所示。

DGX-1 架构的 NVLink 混合立体互联拓扑结构可提供较 PCIe 互联更优的强扩展。DGX-1 配备 8 个 Tesla P100 GPU，与配备 8 个 P100 GPU 但透过 PCIe 连接（采用树形拓扑）的系统相比，其 ResNet-50 训练性能可以提高至 1.6 倍以上。

3.3 借助 InfiniBand 实现 DGX-1 系统的多系统扩展

最新计算工作负载（尤其是深度学习）的多系统扩展，需要在 GPU 之间进行系统内部及跨系统的强通信，以匹配各系统的强大 GPU 性能。除能够实现 GPU 间高速内部通信的 NVLink 技术外，DGX-1 还采用了 Mellanox ConnectX-4 EDR InfiniBand 端口，为系统间提供大量带宽，同时减少瓶颈。DGX-1 中配置最新的 InfiniBand 标准，即 EDR IB，可实现：

- 每个端口有 8 条数据通道，每通道可以 25 Gb/s 或总计 200 Gb/s 的速率运作（4 条输入通道 (100 Gb/s) 和 4 条输出通道 (100 Gb/s) 同步运作）；
- 采用低延时通信及内置原语和群集，可以加速多个系统间的大型运算；
- 高性能网络拓扑支持，可以同时在多个系统之间进行数据传输，将数据冲突情况降至最小；
- 跨 InfiniBand 结构的 NVIDIA GPUDirect RDMA[3] 支持多系统中 GPU 间的直接传输。

DGX-1 配有四个 EDR IB 端口，速率可达 800 Gb/s（同时可达 400 Gb/s 的系统输入输出速率），以用于构建 DGX-1 系统的高速集群。四个 EDR IB 端口平衡节点内部及节点之间的带宽，且在特定用例中，节点间通信可以充分利用完这些带宽。与以太网这类典型的联网技术相比，InfiniBand 在大型多系统集群之间可提供相当于 20 倍的带宽和 4 倍的低延迟（请参阅表 1）。

表 1　采用 InfiniBand 的多系统 DGX 集群可将网络性能提高 20 倍。

	普通多系统集群	**DGX-1 多系统集群**
系统数量	124	124
技术	双 10 Gb 以太网	四 EDR IB

3. 通过 NVIDIA GPUDirect 远程直接数据存取 (RDMA) 协议，可以跨 InfiniBand 网络在两个不同系统中的 GPU 间直接传输数据，而无需涉及 CPU 或系统内存。这可以显著减少延迟，提高多系统的性能表现。

表 1　采用 InfiniBand 的多系统 DGX 集群可将网络性能提高 20 倍。（续）

	普通多系统集群	DGX-1 多系统集群
单一系统峰值网络带宽	40 Gb/s	800 Gb/s
全集群峰值对分带宽	310 GB/s	6,012 GB/s[a]
系统范围的延迟	5 us	1.28 us[b]

a. 6,012 GB/s 的 InfiniBand 全集群峰值带宽是根据 124 个系统计算得出，每系统有 4 个 EDR IB 端口，速率为 200 Gb/s（同时有 100 Gb/s 输入和 100 Gb/s 输出速度） * 64/66 位编码除以 8 位/字节以转换成字节数，乘以 0.5，计算出对分带宽。

b. 1.28 us 的 Infiniband 延时基于 1.01 us 的系统到系统延时，加上 0.09 us 时的 3 次交换器转换。

最新的 DGX-1 多系统集群采用基于胖树拓扑结构的网络，可在系统之间提供路线良好、可预测且无冲突的通信（请查阅图 7）。胖树结构是一种树状网络拓扑结构，叶上的系统通过多个转换器层级连接至一个中央顶层交换器。胖树结构中的每个层级都有同样数量的链路，提供同等的带宽。胖树拓扑结构可确保最高的通信对分带宽[4]和最低延时，适用于计算和深度学习应用中常见的 all-to-all 或 all-gather 类型群集。

此外，内部立体网络以能够提供最佳性能的方式连接至外部的 InfiniBand 网络。DGX-1 系统中的 EDR IB 端口设于 NVLink 混合立体网络的转折处，可以通过 InfiniBand 实现 GPU 间的直接 RDMA 通信。每个 100 Gb IB 链路都有与 x16 PCIe gen 3 接口相似的带宽，所以系统间的相互连接可得到很好的平衡。从应用程序的观点来看，GPUDirect RDMA 及这种独特的网络设计可让任何 GPU 内核都能直接存取网络中任何其他 GPU 的内存，同时将用度、延时和冲突降至最低。

4. 对分带宽是一个网络集群的两部分总的可用带宽。其确定方法是沿中心将系统网络拆分，然后加上所有拆分后链路的带宽。

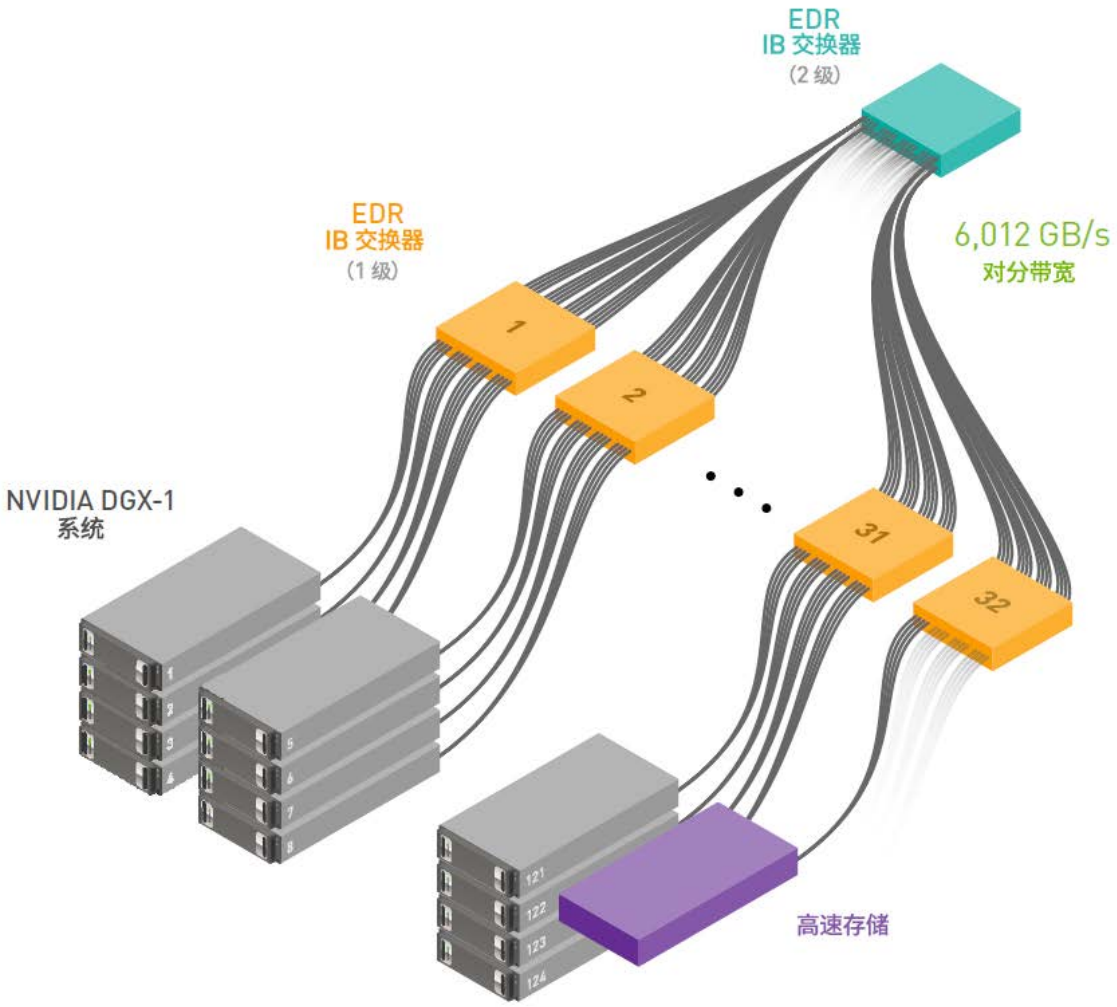

图 7　　专为深度学习设计的 124 个 DGX-1 系统的多系统集群示例

虽然图 7 中的参考系统架构基于 124 个 DGX-1 系统，但采用的胖树拓扑结构可扩展为具有更多交换器的更大配置，且仍保留 InfiniBand 的高性能特征。对于图 7 中的设计，如果每个 1 级交换器支持四个 DGX-1 系统，则可以配置至多达 32 个 1 级交换器和 16 个 2 级交换器，以支持最多 128 个系统。要支持更大的系统，可以新增第三个交换器层级。

4 DGX-1 软件

目前已有可以大规模运行深度学习的 DGX-1 软件。其主要目标是让从业者能够在 DGX-1 上部署深度学习框架和应用程序，同时将设置工作减至最少。该平台软件的设计理念为最大限度地减少服务器上安装的操作系统和驱动程序，并通过由 NVIDIA 维护的 DGX Container Registry[5] 在 NVIDIA Docker（请参阅第 4.2 节）容器内配置全部应用程序和 SDK 软件。DGX-1 的可用容器包括多个经优化的深度学习框架、NVIDIA DIGITS 深度学习训练应用程序、第三方加速解决方案及 NVIDIA CUDA Toolkit。图 8 显示的是 DGX-1 深度学习软件堆栈。

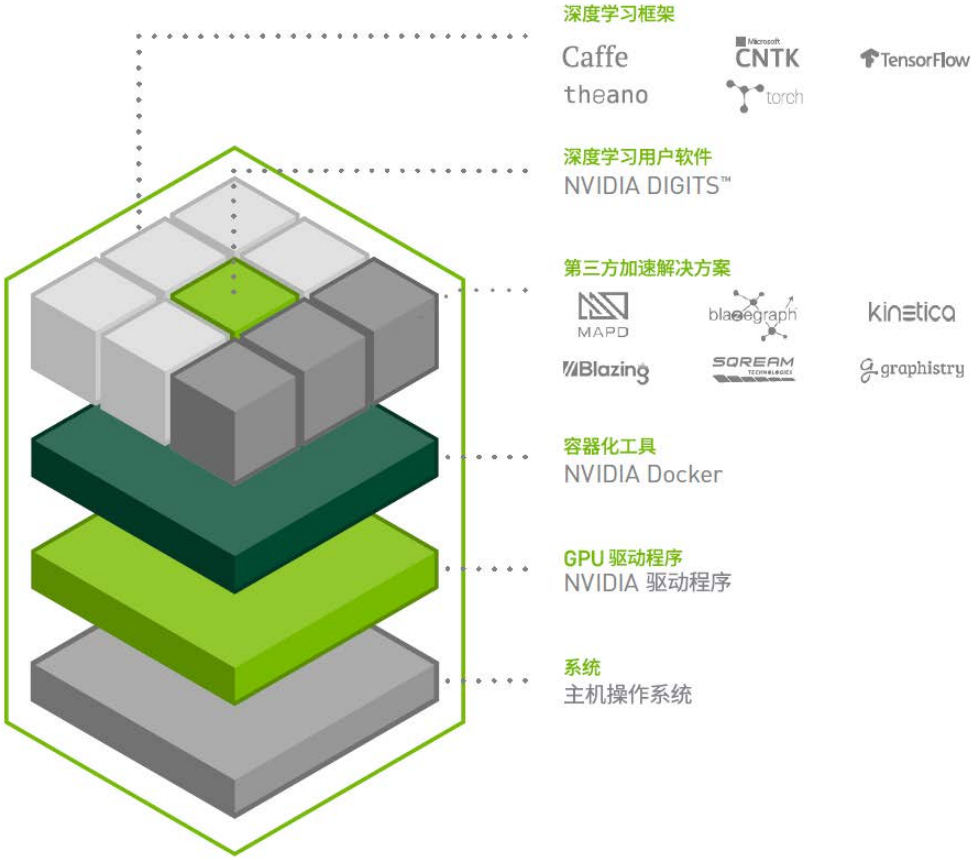

图 8 DGX-1 深度学习软件堆栈

5. NVIDIA 提供的 Docker 注册表服务。请参阅 http://docs.nvidia.com/dgx/dgx-registry-guide/

该软件架构具有很多优势：

- 每个深度学习框架都位于单独的容器内，所以每个框架都能使用不同版本的库，比如 libc、cuDNN 等，并且不会相互影响。
- 为提高性能或修复问题，深度学习框架经过多次改进，现在 DGX Container Registry 中已有新版本容器。
- 系统易于维护，且由于应用程序并非直接安装于操作系统上，所以操作系统镜像非常干净。
- 可无缝提供安全更新、驱动程序更新及操作系统补丁。

这些深度学习框架和 CUDA Toolkit 都包含经定制调整的库，可以在 DGX-1 上提供较高的多 GPU 性能。

本节的其余篇幅将详细介绍 DGX-1 软件堆栈的主要组件（在 GPU 计算软件驱动程序层之上）。第 5 节将详细介绍 DGX-1 针对深度学习框架进行的优化。

4.1 NVIDIA CUDA Toolkit

CUDA 是 NVIDIA 建立的并行计算平台和编程模型，允许应用程序开发人员利用 GPU 强大的并行处理能力。CUDA 是深度学习以及很多其他计算及存储密集型应用的 GPU 加速基础，包括天文学、分子动力学模拟和计算金融学等。目前利用 CUDA 并行计算平台的 GPU 加速应用程序已超过 400 款 [NVIDIA 2017b]。DGX-1 不仅是速度超快的深度学习平台，而且是适合种类广泛的 GPU 加速应用程序的超先进 CUDA 平台。

NVIDIA CUDA Toolkit 为 C 和 C++ 开发人员构建 GPU 加速应用程序提供全面的开发环境。CUDA Toolkit 包含 NVCC（面向 NVIDIA GPU 的 CUDA C++ 编译器）、一套 GPU 加速算法库、调试和分析工具、范例以及综合编程指南和文档。CUDA Toolkit 可以直接安装于 DGX-1，也可以作为 NVIDIA Docker 容器镜像提供，以用作任何容器化 CUDA 应用程序的基础层（如图 9 所示）。此外，整个 CUDA Toolkit 会内嵌于各深度学习框架的容器镜像中。

4.2 NVIDIA Docker

近些年来，软件容器的应用迅速增加，以简化数据中心应用程序的大规模部署工作。容器可以压缩封装应用程序的依赖项，提供可复制且可靠的应用程序和服务器的执行工作及服务，同时避免完整虚拟机的使用。

Docker 容器提供了一个机制，用于将 Linux 应用程序与其所有的库、配置文件及环境变量捆绑在一起，这样不论在什么 Linux 系统上运行或是在相同主机实例间，其执行环境始终相同（请参阅图 9）。Docker 容器为仅用户模式，所有从容器调用的内核都由主机系统内核处理。DGX-1 将 Docker 容器作为部署深度学习框架的机制。

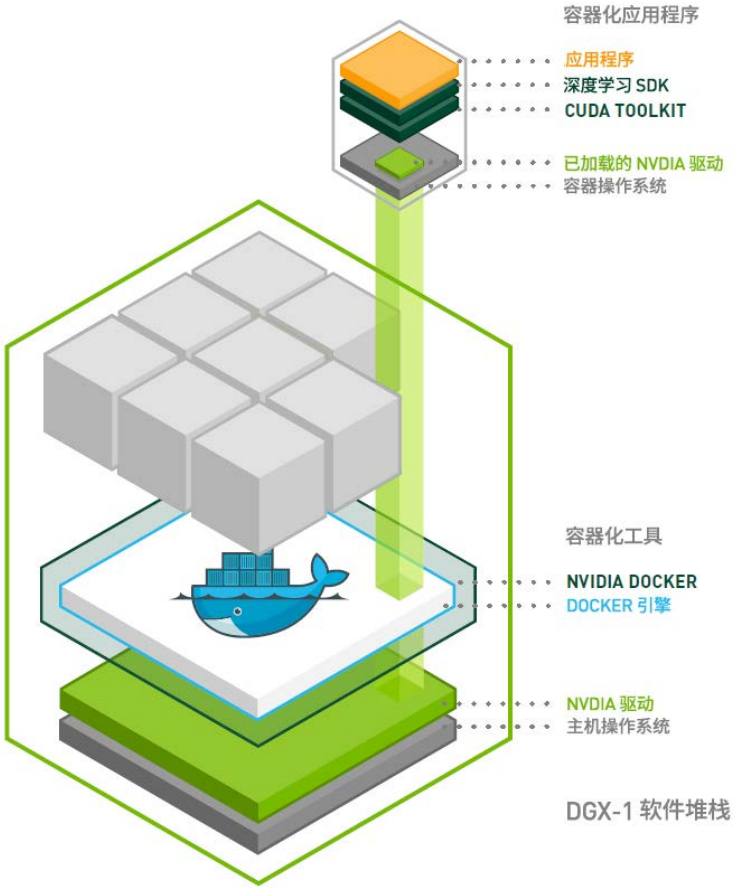

图 9　Docker 容器压缩封装应用程序依赖项，以实现可复制且可靠的应用程序执行。NVIDIA Docker 会在启动时将 NVIDIA 驱动程序和 GPU 的用户模式组件加载到 Docker 容器中。

Docker 容器是与平台和硬件无关的，通过将（容器中的）用户模式代码与（在主机上存取的）内核模式代码进行分离实现这一点。在使用 NVIDIA GPU 等专用硬件时，这种分离就会产生问题，因为 GPU 驱动程序包含的用户模式和内核模式组相互匹配。早期围绕这一问题做的工作是，在容器中完整安装 NVIDIA 驱动程序，然后在启动时映射到 NVIDIA GPU 对应的字符设备中（例如 /dev/nvidia0）。由于主机驱动程序的版本必须要与容器中所安装驱动程序的版本完全匹配，因此这种解决方案不太可靠。这个要求会显著降低这些早期容器的可移植性，削弱 Docker 的这一重要特性。

为实现利用 GPU 的 Docker 镜像的可移植性，NVIDIA 开发了开源项目 NVIDIA Docker [NVIDIA Corporation 2015]，提供了一个命令行工具，用于在启动时就将 NVIDIA 驱动程序的用户模式组件和 GPU 全部加载到 Docker 容器中，如图 9 所示。为了使此解决方案发挥作用，开发人员不能在构建 Docker 时将 NVIDIA 驱动程序安装到 Docker 镜像中，这一点很重要。

nvidia-docker 本质上是一个围绕 Docker 的包装器，可以透明方式为容器配置必要的组件，以执行 GPU 上的代码。

4.3 NVIDIA 深度学习 SDK

NVIDIA 可提供构建于 CUDA 并行计算平台之上的全套 GPU 加速库。下面两种库可提供用于深度神经网络的 GPU 加速原语：

- CUDA 基础线性代数子例程库 (cuBLAS)：cuBLAS 是完整的标准 BLAS 库的 GPU 加速版，可大幅提高 GPU 上的运行速度。cuBLAS 广义矩阵乘法 (GEMM) 例程是深度神经网络应用的关键计算，例如应用于全连接层计算。
- CUDA 深度神经网络库 (cuDNN)：cuDNN 可大幅优化标准例程（例如向前和向后的卷积层、池化层、归一化层和激活层）的实施。

在使用面向 DGX-1 的 NVIDIA Docker 容器部署时，深度学习框架会自动配置为使用专为 DGX-1 中 Tesla P100 GPU 架构优化的并行例程。

4.4 NCCL

NVIDIA 集合通信库（NCCL，发音与 “Nickel” 相同）是一个多 GPU 集合通信原语库，具备拓扑感知，可轻松集成于应用程序中。NCCL 最初是作为开源研究项目开发的[6]，其设计轻巧，仅依赖于常见的 C++ 库和 CUDA 库。NCCL 可部署于单进程或多进程应用程序，以透明方式处理必要的进程间通信。NCCL API 的设计目的是让任何用过 MPI 群集（例如 broadcast、reduce、gather、scatter、all-gather、all-reduce 或 all-to-all）的用户都感到熟悉。

面向 DGX-1 的 NVIDIA Docker 容器包含某版本的 NCCL，可针对 DGX-1 架构的 8-GPU 混合立体 NVLink 网络优化这些群集。使用这些容器部署时，诸如 NVIDIA Caffe、Torch、CNTK (Microsoft Cognitive Toolkit) 和 TensorFlow 等深度学习框架如果要在多个 GPU 上运行，则会自动使用该版本的 NCCL。

高效实现群集的方法有很多。但关键是，我们的实现需要将处理器间互联的拓扑结构考虑在内。为说明这一点，请思考将 GPU0 中的数据传播至图 10 中 PCIe 树形拓扑结构中的所有其他 GPU。

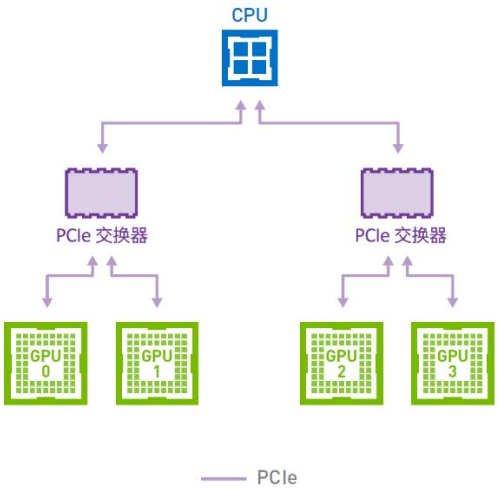

图 10　通用 PCIe 拓扑结构，适用于单一 CPU 附加 4 个 GPU 的情况。紫色箭头表示 PCIe x16 连接。

一种方法是包含两个步骤的树形算法：第一步是将数据从 GPU0 传送至第二个 GPU，第二步是这两者将数据传送至其余处理器。不过，还有另一种方法可选。在第一步中，将数据从 GPU0 传送至 GPU1，然后在第二步中从 GPU0 传送至 GPU2，从 GPU1 传送至 GPU3，或是先从 GPU0 复制数据至 GPU2，然后在第二步中从 GPU0 复制至 GPU1，从 GPU2 复制至 GPU3。观察这个拓扑结构，显然可以看出我们倾向于第二种方法，因为同时从 GPU0 传送数据至 GPU2 以及从 GPU1 到 GPU3 会导致上层 PCIe 链路出现冲突，将该步骤的有效带宽减半。总之，群集想要有出色的性能表现，需要特别注意互联拓扑结构。

6. http://github.com/NVIDIA/nccl

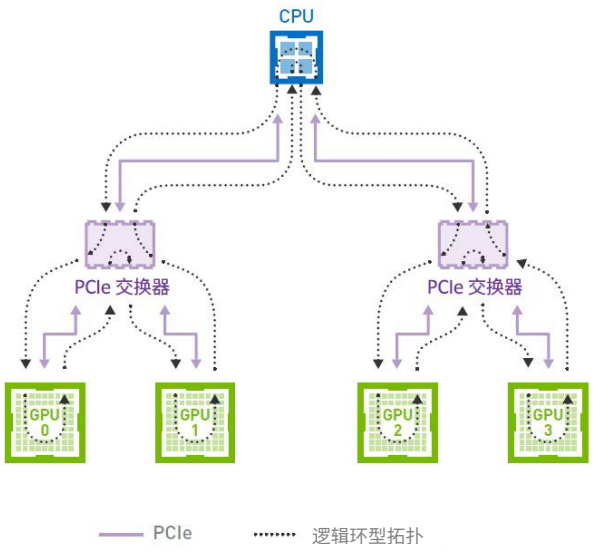

图 11　PCIe 树状结构中 GPU 的环形顺序

为优化传播带宽，更好的方法是将 PCIe 树状拓扑结构作为一个圆环，如图 11 所示。然后，围绕这个圆环从 GPU0 到 GPU3 转播小块输入数据，以此执行数据传播。有趣的是，环形算法可为几乎全部标准群集运算提供近乎最优的带宽，甚至在应用于树状 PCIe 拓扑结构时亦如此。但要注意，选择正确的环形顺序仍很重要。

为提供最大带宽，NCCL 实施了环型群集。实际上，NCCL 以隐含方式将 GPU 编入最佳环形顺序。这可以让应用程序有出色的性能表现，同时让开发人员免于担心特定的硬件配置。

配备 8-GPU 的混合立体网络可以被视作单个 NVLink 连接的两个相互交织的双向环形，如图 12 所示。以这种方式看待拓扑结构，可确保除 all-to-all 外群集的性能大致相等。

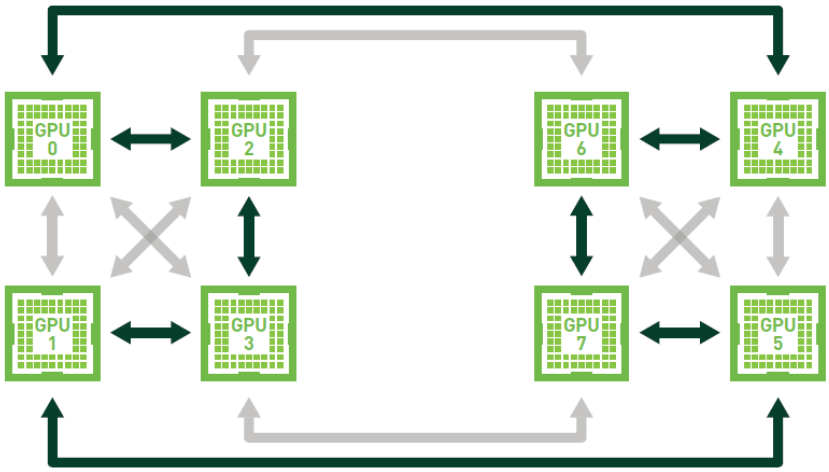

图 12　DGX-1 NVLink 混合立体拓扑可被视作两个交织的单 NVLink 连接双向环；此处以灰色和深绿色显示。

通过这两个独特的双向环执行深度学习的 GPU 间转移。两个双向环连接全部的 8 个 GPU，并且共同使用两个方向中每个 Pascal GPU 的所有四条链路。相比在前几代硬件上使用 PCIe 所获得的 10 GB/s

速度，通过此方法，能以超过 60 GB/s 的速度执行简化和传播操作。此性能对于实现高扩展性的深度学习训练至关重要。

5 DGX-1 适用的深度学习框架

NVIDIA 深度学习 SDK 可加速广泛使用的深度学习框架，例如 Caffe、CNTK、MXNet、TensorFlow、Theano 和 Torch。以下章节将介绍可在 DGX-1 上运行的上述各框架。

DGX-1 软件堆栈提供专为系统优化的框架容器化版本。这些框架（包括所有必要的依赖项）皆经过预建、测试并且可随时运行。对于在建立定制深度学习解决方案时需要更多灵活性的用户，每个框架容器镜像还包含支持定制修改和增强功能的框架源代码，以及完整的软件开发堆栈，详情请参阅第 4 节。

5.1 NVIDIA Caffe

Caffe[7] 深度学习框架在设计时已将灵活性、速度和模块化考虑在内。该框架最初是由伯克利视觉和学习中心 (BVLC) 及社区贡献者共同开发。

NVIDIA Caffe [NVIDIA Corporation 2017a] 是由 NVIDIA 维护的 BVLC Caffe 分支版本，专门针对 NVIDIA GPU（尤其是多 GPU 配置）进行调整。NVIDIA Caffe 包括多精度支持以及其他 NVIDIA 增强功能，并且提供专门针对 NVIDIA DGX-1 调整的性能。

以下总结列出了 NVIDIA Caffe 的优化和变更内容。

- 采用最新版的 cuDNN
- 集成支持 NVLink 的最新版 NCCL，可提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍
- 采用并行解析器以提升 I/O 性能
- 微调性能
- 支持 16 位浮点 (FP16) 计算和存储，可降低一半的内存和存储要求
- 自动选择最佳卷积算法

7. http://caffe.berkeleyvision.org/

5.2 Microsoft Cognitive Toolkit

Microsoft Cognitive Toolkit[8] (CNTK) 是统一的深度学习工具包，让用户可轻松实现并合并前馈深度神经网络 (DNN)、卷积神经网络 (CNN) 和递归神经网络 (RNN) 等热门模型类型。CNTK 跨多个 GPU 和服务器进行自动微分和并行化，以执行随机梯度下降 (SGD) 学习 [Microsoft Corporation 2017]。CNTK 可称之为 Python 或 C++ 应用程序库，或是作为使用 BrainScript 模型描述语言的独立执行工具。

NVIDIA 已和 Microsoft 紧密合作，加快 Cognitive Toolkit 在 DGX-1 和 Azure N 系列虚拟机等 GPU 系统中的运行速度。双方的合作为初创公司和大型企业等提供了极大的易用性和可扩展性，这是因为单个框架可用于 DGX-1 内部部署的首批训练模型，之后可再将这些模型部署到 Microsoft Azure 云 [9]。

以下总结列出了 DGX-1 CNTK 的优化和变更内容。

- 采用最新版的 cuDNN
- 集成支持 NVLink 的最新版 NCCL，可提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍
- 改进图像阅读器管道，允许 AlexNet [Krizhevsky et al.2012] 以超过 12,000 张图片/秒的速度进行训练
- 为多 GPU 训练的每块 GPU 可减少高达 2 GB 的 GPU 内存用度
- 扩大卷积支持

5.3 MXNet

MXNet[10] 是专为提高效率和灵活性而设计的深度学习框架，让您可以混合符号编程和命令编程，以最大限度提升效率和生产力。MXNet 的核心是动态依赖调度程序，可自动即时并行处理符号式操作和命令式操作。在调度程序之上增加图形优化层可加快符号执行速度并提高内存效率。MXNet 轻量便携，并且可扩展至多个 GPU 和多台机器上。

8. https://www.microsoft.com/en-us/research/product/cognitive-toolkit/
9. 如需在 Azure 中使用 CNTK 的相关信息，请参阅 https://github.com/Microsoft/CNTK/wiki/CNTK-on-Azure
10. http://mxnet.io

以下总结列出了 DGX-1 MXNet 的优化和变更内容。

- 采用最新版的 cuDNN
- 改进输入管道，便于处理图像
- 优化嵌入层 CUDA 内核
- 优化张量传播和简化 CUDA 内核

5.4 TensorFlow

TensorFlow[11] 是使用数据流图表进行数值计算的开源软件库。图像中的节点代表数学运算，而图像边缘则代表节点间流动的多维数据阵列（张量）。您可通过此灵活的架构，将计算部署到桌面、服务器或移动设备中的一个或多个 CPU 或 GPU，而无需重写代码。

TensorFlow 最初由 Google Brain 团队（隶属于 Google 机器智能研究组织）的研究人员和工程师开发，旨在进行机器学习和深度神经网络研究。该系统具有极佳的通用性，也适用于其他各种领域。

以下总结列出了 DGX-1 TensorFlow 的优化和变更内容。

- 采用最新版的 cuDNN
- 以 libjpeg-turbo 替换 libjpeg
- 集成支持 NVLink 的最新版 NCCL，可提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍
- 支持 ImageNet 预处理脚本

11. https://www.tensorflow.org

5.5 Theano

Theano[12] 是一个 Python 库，能让您有效定义、优化和评估涉及多维阵列的数学表达式。自 2007 年起，Theano 便一直为大规模计算密集型科学调查提供支持。

以下总结列出了 DGX-1 Theano 的优化和变更内容：

- 采用最新版的 cuDNN
- 运行时代码生成：更快速地评估表达式
- 广泛的单元测试和自我验证：检测与诊断多种类型的错误

5.6 Torch

Torch[13] 是一种科学计算框架，可为深度学习算法提供广泛支持。由于简单且快速的脚本语言 Lua 以及底层 C/CUDA 实现，Torch 易于使用且效率很高。Torch 提供热门的神经网络和优化库，不但容易使用，而且还提供超大灵活性，以构建复杂的神经网络拓扑。

以下总结列出了 DGX-1 Torch 的优化和变更内容。

- 采用最新版的 cuDNN
- 集成最新版、采用 NVLink 支持的 NCCL，以提升多 GPU 扩展性。使用数据并行 SGD 时，支持 NVLink 的 NCCL 可将 ResNet-50 训练性能提升 2 倍
- 缓冲待 NCCL 通信的参数，以降低延迟用度
- 针对递归网络（RNN、GRU、LSTM）的 cuDNN 绑定（包括持久性版本），可大幅提升小批量训练的性能
- 扩大卷积支持
- 支持向 cuDNN 例程的 16 和 32 位浮点（FP16 和 FP32）数据输入
- 支持对 FP16 张量的操作（使用 FP32 算法）

12. http://deeplearning.net/software/theano/
13. http://torch.ch

5.7 DIGITS

NVIDIA 深度学习 GPU 训练系统 (DIGITS)[14] 将深度学习的强大功能交由工程师和数据科学家掌控。

DIGITS 可用于快速训练高度准确的深度神经网络 (DNN)，以执行图像分类、分割以及物体检测等任务。DIGITS 可简化常见的深度学习任务，例如管理数据、在多 GPU 系统上设计和训练神经网络、使用高级可视化技术实时监控性能，以及从结果浏览器中选择性能更好的模型用于部署。DIGITS 采用完全交互式的设计，因此数据科学家可以专注于设计和训练网络，而不必进行编程和调试。

以下总结列出了 DGX-1 DIGITS 的优化和变更内容。

- DIGITS 运行于针对 DGX-1 进行优化的 Caffe 和 Torch 框架之上

14. https://developer.nvidia.com/digits

6 结果：为实现最高深度学习性能而打造的 DGX-1

DGX-1 在训练热门深度神经网络方面的性能表现有力证明了深度学习集成系统的价值。图 13 中的图表显示相较于现有系统（相同 GPU 且采用 Microsoft Cognitive Toolkit、TensorFlow 和 Torch 训练 ResNet-50 和 ResNet-152 深度神经网络），DGX-1 的训练加速情况。该图表显示出两个明显的优势：

1. 与用于深度学习训练的前代 NVIDIA Tesla M40 GPU 相比，DGX-1 中 P100 GPU 的吞吐量明显增加。
2. 与 8 个 Tesla P100 GPU 通过 PCIe 互联的类似系统相比，DGX-1 的性能明显提升。

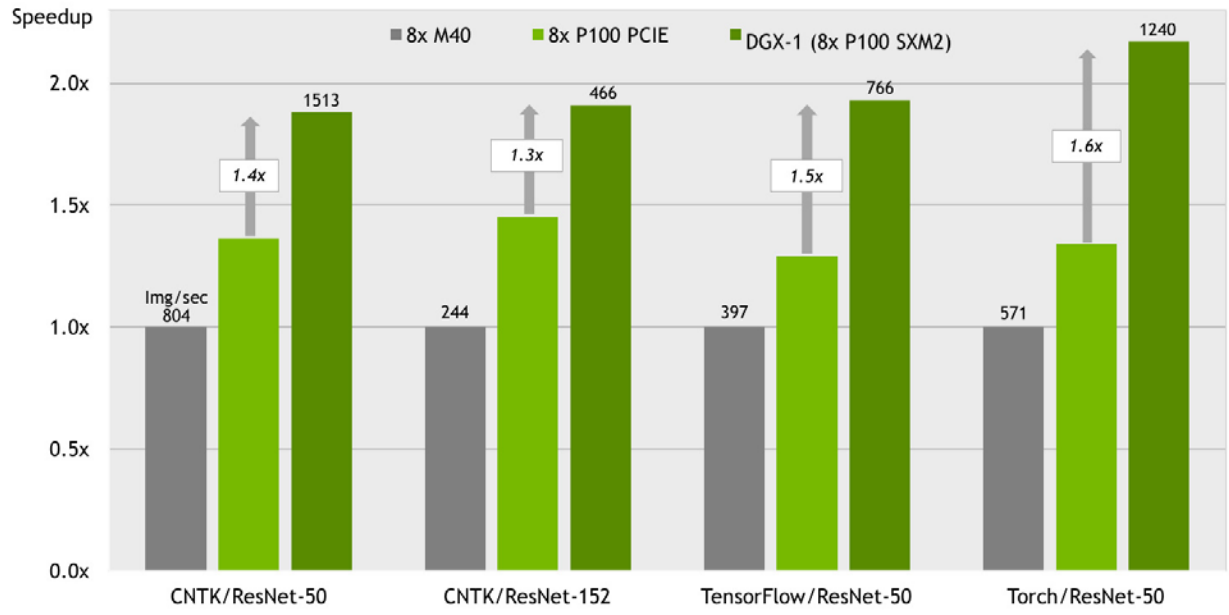

图 13　在热门 CNTK (2.0 Beta5)、TensorFlow (0.12-dev) 和 Torch (11-08-16) 深度学习框架中，针对 ResNet-50 和 Resnet-152 深度神经网络架构，使用 8 块 Tesla P100 的 DGX-1 与 8-GPU Tesla M40 系统、使用 PCI-e 互联的 8-GPU P100 系统之深度学习训练加速比较。训练使用 32 位浮点算法，ResNet-50 和 ResNet-152 的总批量大小分别为 512 和 128。其他软件：NCCL 1.6.1、CUDA 8.0.54、cuDNN 6.0.5、Ubuntu 14.04。NVIDIA Linux 显示驱动程序 375.30。8x M40 和 8x P100 PCIe 服务器是搭载双 Intel Xeon E5-2698v4 CPU 和 256GB DDR4-2133 RAM（DGX-1 拥有 512GB DDR4-2133）的 SMC 4028GR 服务器。测试使用 NVIDIA DGX 容器版本 16.12。

8 个 Tesla P100 GPU 之间的 NVLink 混合立体网络互联只是 DGX-1 提供高性能的部分原因，其背后的助力因素远不止如此。DGX-1 是一个集成系统，并且拥有专门针对深度学习的完整软件平台（如第 4 和 5 节所述），而这才是 DGX-1 性能优势的主要来源。DGX-1 在深度学习方面的优势还包括深度学习框架的优化，例如对 NVIDIA Caffe、cuBLAS、cuDNN 和其他 GPU 加速库的优化，以及通过 NVLink 调整的 NCCL 集合通信。此集成软件平台与 Tesla P100 和 NVLink 的组合，可确保 DGX-1 的性能优于类似的现有系统。

如要深入了解 NVIDIA DGX-1，请访问
http://www.nvidia.cn/object/deep-learning-system-cn.html

参考资料

Finley, K.2016. Amazon's Giving Away the AI Behind Its Product Recommendations http://www.wired.com/2016/05/amazons-giving-away-ai-behind-product-recommendations/

He, K., Zhang, X., Ren, S., and Sun, J.2015.Deep Residual Learning for Image Recognition. *arXiv [cs.CV]*. http://arxiv.org/abs/1512.03385.

Krizhevsky, A.2014. One weird trick for parallelizing convolutional neural networks. *arXiv [cs.NE]*. http://arxiv.org/abs/1404.5997.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. 2012.ImageNet Classification with Deep Convolutional Neural Networks.In: F.Pereira, C.J.C.Burges, L.Bottou and K.Q.Weinberger, eds., *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc.,1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Metz, C.2015. TensorFlow, Google's Open Source AI, Signals Big Changes in Hardware Too. http://www.wired.com/2015/11/googles-open-source-ai-tensorflow-signals-fast-changing-hardware-world/

Microsoft Corporation. 2017. The Microsoft Cognitive Toolkit (CNTK). Github. https://github.com/Microsoft/CNTK#what-is-the-microsoft-cognitive-toolkit.

Ng, A.2016a. Baidu's Chief Scientist on Intersection of Supercomputing, Machine Learning. http://www.nextplatform.com/2016/04/01/baidus-chief-scientist-intersection-supercomputing-machine-learning/.

Ng, A.2016b. AI: The New Electricity.https://www.youtube.com/watch?v=4eJhcxfYR4I.

NVIDIA Corporation. 2015.nvidia-docker software. *Github*. https://github.com/NVIDIA/nvidia-docker.

NVIDIA Corporation. 2016.NVIDIA® Tesla® P100 – The Most Advanced Data Center Accelerator Ever Built.Featuring Pascal P100, the World's Fastest GPU. http://www.nvidia.com/object/pascal-architecture-whitepaper.html

NVIDIA Corporation. 2017a. *NVIDIA Caffe branch*. https://github.com/NVIDIA/caffe

NVIDIA Corporation. 2017b. *GPU-Accelerated Applications*. http://www.nvidia.com/content/gpu-applications/PDF/gpu-applications-catalog.pdf

Schroepfer, M.2016.F8 2016 Day 1 Keynote.https://developers.facebook.com/videos/f8-2016/keynote/.

Statt, N.2016. Exploring Facebook's massive, picture-painting AI brain. http://www.theverge.com/2016/7/13/12172904/facebook-ai-big-sur-machine-learning-prineville-data-center

Wu, R.,Yan, S., Shan, Y., Dang, Q., and Sun, G.2015.Deep Image: Scaling up Image Recognition.*arXiv [cs.CV]*. http://arxiv.org/abs/1501.02876.

通知

本白皮书所述的所有信息，包括评论、意见、NVIDIA 设计规范、参考板、文件、图纸、诊断信息、列表和其他文档（统称与单称均为“资料”）均“如实”提供。NVIDIA 并未作出与资料相关的明示、暗示、法定或其他形式的保证，并明确否认与非侵权、适销性和特定用途适用性相关的所有暗示保证。

NVIDIA 保留随时对这一规范进行纠正、更改、增强、改进以及其他改动和/或终止任何产品或服务的权利，恕不另行通知。客户在下订单之前应获取最新的相关规范并验证这些信息是否为当前信息以及是否完整。

除非 NVIDIA 授权代表与客户另行签署销售协议，否则 NVIDIA 产品的销售受订单确认时所提供的 NVIDIA 标准销售条款与条件的制约。就购买这一规范中提到的 NVIDIA 产品而言，NVIDIA 在此明确拒绝应用客户的任何一般条款与条件。

NVIDIA 产品并非针对医学、军事、航空、航天或生命保障设备而设计，并未授权用于也不保证适合用于上述设备，亦不得用于 NVIDIA 产品之失效或故障合理预计会造成人身伤亡或财产或环境破坏的应用场合。客户如果在此类设备或应用场合中融入和/或使用 NVIDIA 产品，NVIDIA 不承担任何相关责任，风险由客户自行承担。

在未经进一步测试或改动的情况下，NVIDIA 并不表示，也不担保基于这些规范的产品适合任何具体用途。每款产品所有参数的测试不一定由 NVIDIA 进行。确保产品适合客户所计划的应用场合并针对该应用场合进行必要的测试以避免应用场合出现问题或产品失灵，是客户单方面的责任。客户产品设计中的缺点可能会影响 NVIDIA 产品的质量和可靠性，并且可能会导致超出本规范以外的额外或不同的条件和/或要求。NVIDIA 不承担因下列情况造成失灵、损坏、成本或问题相关的任何责任：(i) 以违反本规范的方式使用 NVIDIA 产品或 (ii) 客户产品设计。

本规范不对 NVIDIA 专利权、版权或其他 NVIDIA 知识产权作任何明示或暗示的许可。NVIDIA 所发布的有关第三方产品或服务的信息并不构成 NVIDIA 对于使用该产品或服务的许可，亦不构成担保或支持。使用此类信息可能需要获得第三方的专利权或其他知识产权的许可，或者需要获得 NVIDIA 的专利权或其他知识产权的许可。只有在获得 NVIDIA 书面批准的情况下才可以复制本规范中的信息，而且必须毫无改动地复制并附带所有相应的条件、限制条款和通知。

所有 NVIDIA 设计规范、参考板、文件、图纸、诊断信息、列表和其他文档（统称与单称均为“资料”）均“如实”提供。NVIDIA 并未作出与资料相关的明示、暗示、法定或其他形式的保证，并明确否认与非侵权、适销性和特定用途适用性
相关的所有暗示保证。尽管客户可能会因任何原因造成损失，但是 NVIDIA 针对本文所述产品向客户承担的全部责任应仅限于该产品的 NVIDIA 销售条款与条件

商标

NVIDIA、NVIDIA 徽标、CUDA、Pascal、Tesla、NVLink 和 DGX-1 是 NVIDIA Corporation 在美国和其他国家/地区的商标和/或注册商标。其他公司和产品名称可能是其各关联公司的商标。

版权所有

© 2017 NVIDIA Corporation. 保留所有权利。