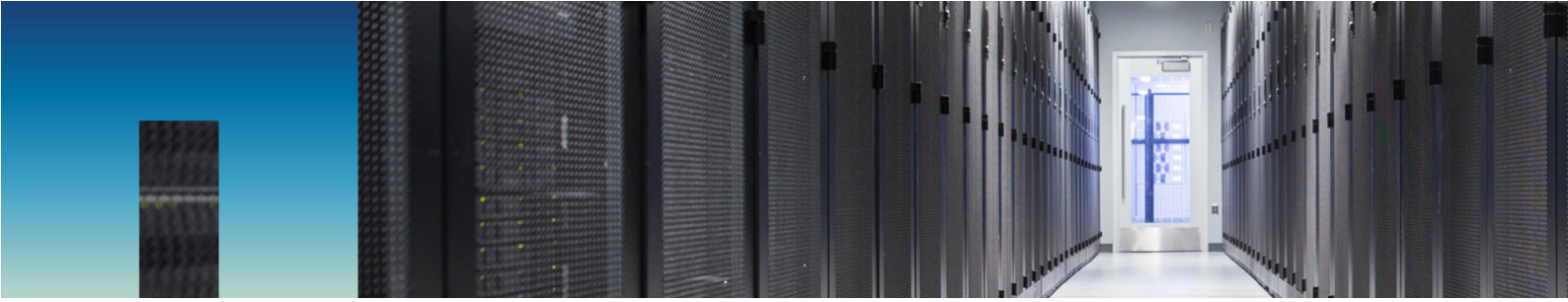


技术白皮书

可扩展的 AI 基础架构

专为实际深度学习使用情形而设计

NetApp 公司 Sundar Ranganathan
NetApp 公司 Santosh Rao

2018 年 6 月 | WP-7267

合作方

NVIDIA

内容提要

如今，在深度学习 (Deep learning, DL) 的推动之下，一些面临着巨大挑战的科学领域取得了突飞猛进的发展：在医学领域发现更好的癌症治疗方法，在物理学领域实现粒子探测与分类，在自动汽车领域实现 5 级无人驾驶。这些成就有一个共同的元素，那就是数据。深度学习从根本上说是数据驱动的。

图形处理单元 (GPU) 可帮助获得以前不可能获得的新洞见。为了满足深度学习应用程序中对 GPU 的严苛需求，存储系统必须能以低延迟和高吞吐量不断地为 GPU 馈送数据，无论数据是文本、图像、音频还是视频。

随着企业从小规模深度学习部署向生产方向发展，设计一个能够提供高性能并且支持独立无缝扩展的基础架构变得尤为重要。NVIDIA 在 GPU 领域的领先地位与 NetApp 在全闪存存储系统领域的创新实力完美结合，共同打造了旨在加快深度学习应用程序响应速度的独特解决方案。

NetApp

目录

1 引言 1

2 计算 — NVIDIA DGX-1 1

2.1 GPU 加速计算 1

2.2 专为深度学习构建 1

2.3 经过预先优化的企业级功能 2

3 存储 — NetApp AFF 2

3.1 高性能 2

3.2 可扩展性 2

3.3 强大的平台集成 2

4 参考架构 3

5 机架级扩展 — 从小规模入手，逐步扩展 4

6 性能测试 5

7 结束语 7

8 附录：组件列表 8

表格目录

表 1) 使用 A800 横向扩展时的容量和性能指标 4

表 2) 使用 A700s 横向扩展时的容量和性能指标 4

表 3) 组件列表 8

插图目录

图 1) 采用 1:5 配置的参考架构 3

图 2) 显示 1:5 配置的端口级别连接的网络示意图 4

图 3) 使用 A800 在机架级别从 1:1 扩展到 1:5 配置 5

图 4) 使用真实数据并启用图像失真时的训练速度 6

图 5) 使用真实数据并禁用图像失真时的训练速度 6

图 6) 以每秒约 2500 个图像的速度训练 ResNet-50 时的 GPU 利用率和 A700 读取吞吐量 7

1 引言

当 IT 主管谈起自己企业中数据的价值时，最常提及的示例就是深度学习 (DL) 和更广泛的人工智能 (AI)。深度学习已成为助力金融行业发现欺诈、制造业进行预测性维护、客户支持机器人的语音识别以及汽车领域各级别无人驾驶的强大引擎。

今后，数据和深度学习应用程序将被用于提高生产率，识别基本模式，并在每个可以想象的行业中设计颠覆性服务、解决方案和产品。据市场研究机构 IDC 预测，用于 AI 的软件、服务和硬件支出将从 2017 年的 120 亿美元增长到 2021 年的 576 亿美元[1]。

深度学习系统所利用的算法可通过扩大神经网络的规模以及训练模型所使用数据的数量和质量而得到大幅改进。根据特定的应用程序，深度学习模型会处理大量不同类型的数据，例如文本、图像、音频、视频或时间序列数据。这正是深度学习应用程序需要高性能基础架构（例如本文介绍的参考架构）的原因所在。

如果企业想要成功实施深度学习战略，那么用于部署深度学习工作负载的基础架构必须经过专门设计，能够在处理不断增长的独特深度学习需求的同时，以最短的时间完成训练。

此基础架构应满足以下高级别基础架构要求：

- 强大的计算能力，可以在更短时间内训练模型
- 能够处理大型数据集的高性能存储
- 无缝独立地扩展计算和存储
- 处理不同类型的数据流量
- 优化成本

拥有一个支持灵活地进行纵向扩展和横向扩展的基础架构至关重要。随着计算和存储的扩展，性能需求可能会不断变化，这就要求能够在不停机的情况下动态调整计算系统与存储系统的比例。

本白皮书将介绍一种可扩展的基础架构，其中包括基于全新 NVIDIA® Tesla® V100 图形处理单元 (GPU) 平台[2] 构建的 NVIDIA® DGX-1™ 服务器和全新的 NetApp® A800™ 全闪存存储系统。

2 计算 — NVIDIA DGX-1

2.1 GPU 加速计算

深度学习算法中执行的计算涉及并行运行的巨量矩阵乘法。相较于通用中央处理单元 (CPU)，现代化 GPU 的高度并行架构可大幅提高并行处理数据的应用程序的效率。单个 CPU 架构和集群 GPU 架构的进步之处使其成为高性能计算、深度学习和分析这类工作负载的首选平台。

2.2 专为深度学习构建

组装和集成来自多家供应商的现成深度学习硬件和软件组件，会增加复杂性和部署时间，从而导致宝贵的数据科学资源在系统集成工作上花费大量精力。

许多企业在部署解决方案后发现，随着模型的发展演变，他们在调整和优化软件堆栈上花费了过多的周期。意识到这一点之后，NVIDIA 创建了 DGX-1 服务器平台，这是一个专门为深度学习工作流构建且软硬件全面集成的统包系统。

每台 DGX-1 服务器均由 8 个 Tesla V100 GPU 提供支持，配置在混合式立方体网格拓扑结构中，该拓扑结构使用 NVIDIA NVLink™ 为多 GPU 训练所需的 GPU 间通信提供超高带宽、低延迟网络结构，可消除与基于 PCIe 的互连相关的瓶颈。DGX-1 服务器还配备了低延迟、高带宽网络互连，用于在支持 RDMA 的网络结构上实现多节点集群。

[1] 来源：IDC，2018 年至 2022 年全球用于认知/AI 工作负载的存储的预测
[2] 由 NVIDIA Volta 架构提供支持

2.3 经过预先优化的企业级功能

DGX-1 服务器利用来自 NVIDIA GPU Cloud (NGC) 的 GPU 优化的软件容器，包括用于所有最流行的深度学习框架的容器。NGC 深度学习容器在每个层都经过预先优化，包括驱动程序、库和通信原语，而且可为 NVIDA GPU 提供最高性能。目前流行的开源深度学习框架通常会不断地产生干扰，这些预集成容器可以让用户免受其扰，为团队提供经过 QA 测试的稳定堆栈，用于构建企业级深度学习应用程序。

这款全面集成的软硬件解决方案由 NVIDIA 丰富的专业知识作为后盾，可加快深度学习应用程序部署，将训练时间从数周缩短为几天或几小时，并且可提高数据科学家的工作效率，使他们能够花更多的时间进行实验，而不是浪费在系统集成和 IT 支持上。

3 存储 — NetApp AFF

随着 GPU 速度越来越快，数据集规模和复杂性日益增加，必须使用一流的存储系统来消除瓶颈，最大限度地提高系统性能。深度学习应用程序需要专为处理大量并行深度学习工作负载而设计的存储解决方案，这些工作负载需要高度的并行 I/O 处理能力，才能避免 GPU 因等待数据而发生停滞。

许多深度学习应用程序中的数据流量跨越从边缘到核心再到云的整个数据管道。设计存储架构需要采用全面的数据管理方法，范围包括从数据载入和（或）边缘分析到在核心数据中心内的数据预处理和训练，再到在云中归档。因此，必须了解性能要求、多种数据集的特征以及所需的数据服务。

对于深度学习工作流而言，理想的存储解决方案必须能够出色地满足以下高标准要求。

3.1 高性能

深度学习基础架构中的瓶颈最常出现在训练阶段，此时需要具有高 I/O 带宽和大量并行 I/O 处理能力，才能维持高 GPU 利用率。这就要求存储架构能够在提供高吞吐量性能的同时保持低延迟状态，反过来这意味着需要支持高速网络结构。

单个 NetApp A800 系统支持吞吐量为 25 GB/秒的顺序读取和 100 万次 IOPS 的小型随机读取，同时保持低于 500 微秒 (µs) 的延迟[3]。此外，A800 的与众不同之处在于其 100GbE[4] 网络支持，不仅可以加快数据移动速度，而且能够促进整个训练系统中的平衡，因为 DGX-1 支持 100GbE RDMA 的集群互连。NetApp A700s 系统支持利用多条 40GbE 链路实现最高达 18 GB/秒的吞吐量。

3.2 可扩展性

大型数据集对于提高模型准确性具有重要作用。小规模（几 TB 存储）的深度学习部署可能很快就需要横向扩展到数 PB。而且，性能需求会因所使用的训练模型以及终端应用程序而变化，这就需要独立扩展计算和（或）存储。在机架级环境中设计强大的系统架构即可实现独立扩展。

NetApp A800 和 A700s 系统可以独立、无缝且无中断地从 2 个节点 (364.8 TB) 扩展为包含 24 个节点的集群（使用 A800 时可达 74.8 PB，使用 A700s 时可达 39.7 PB）。使用 ONTAP® FlexGroup™ 卷，可在 20 PB 的单一命名空间逻辑卷中轻松地管理数据，支持超过 4000 亿个文件。如果集群容量大于 20 PB，可以创建多个 FlexGroup 以涵盖所需的容量。

3.3 强大的平台集成

随着企业加快数据收集速度，引入数据自动化操作的需求变得显而易见。使用容器是实现此目的一种方式；这种方式将应用程序与操作系统和设备-驱动程序层分离开，可以加快部署速度。高效和简单的数据管理对于缩短训练时间至关重要。

Trident™ 是 NetApp 提供的用于容器映像的动态存储业务流程协调程序，它与 Docker™ 和 Kubernetes™ 实现了全面集成。Trident 与 NVIDIA GPU Cloud (NGC) 和 Kubernetes 或 Docker Swarm 等常见业务流程协调程序相结合，支持客户将 AI/DL NGC 容器映像无缝部署到 NetApp 存储上，从而获得企业级 AI 容器部署体验。其中包括自动流程编排、用于测试和开发的克隆、使用克隆进行 NGC 升级测试、用于保护和满足合规性要求的副本以及针对 NGC AI 容器映像的更多数据管理用例。

[3] https://blog.netapp.com/the-future-is-here-ai-ready-cloud-connected-all-flash-storage-with-nvme/
[4] https://www.netapp.com/cn/products/storage-systems/all-flash-array/aff-a-series.aspx#technical-specifications

深度学习中的数据流量可包含数百万个文件（图像、视频/音频、文本文件）。网络文件系统 (NFS) 非常适合在各种各样的工作负载之间提供高性能；它们可以很好地处理随机和顺序 I/O。与 ONTAP FlexGroup 卷结合使用时，NetApp AFF 可以为分布在各个存储系统上的小型文件工作负载提供高性能。

4 参考架构

构建一个能够实现持续高 I/O 吞吐量并且保持低延迟的基础架构，以利用 GPU 的 I/O 并行处理机制并且无中断扩展计算和存储系统，才是缩短训练时间的关键所在。这些要求意味着，存储系统必须支持高带宽、低延迟的高速网络结构，才能最大限度地提高系统性能并持续不断地为多台 DGX-1 服务器馈送运行每个深度学习训练所需的数据。

图 1 显示了采用 1:5 配置的 NetApp 架构，它由 5 台 DGX-1 服务器和 1 个 A800 高可用性 (HA) 对组成，后者通过 2 个交换机为服务器馈送数据。每台 DGX-1 服务器通过 2 条 100GbE 链路分别连接到每个交换机。A800 通过 4 条 100GbE 链路连接到每个交换机。两个交换机均具有专门为故障转移情形而设计的 2 到 4 条 100 Gb 交换机间链路。HA 采用双活设计，因此可以在所有网络连接之间保持最大吞吐量，而且不会出现故障。

图 1) 采用 1:5 配置的参考架构

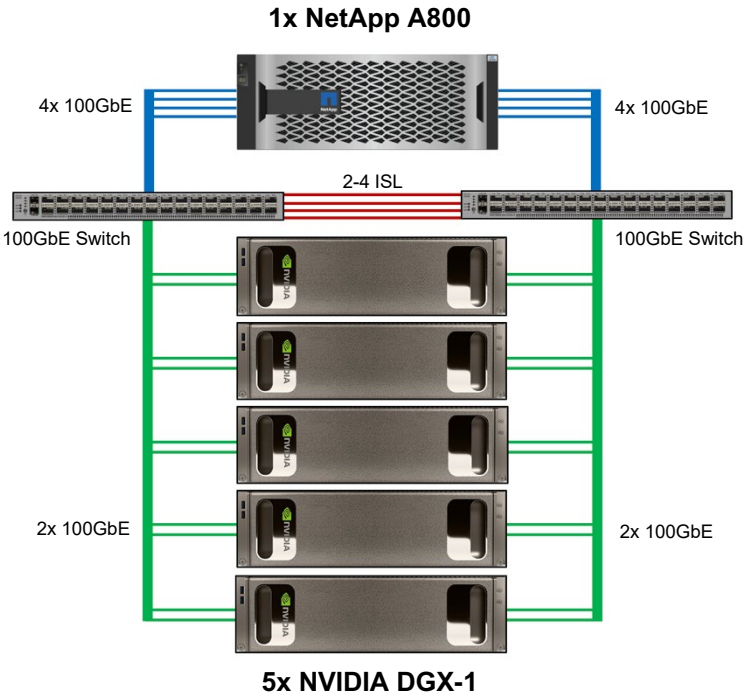
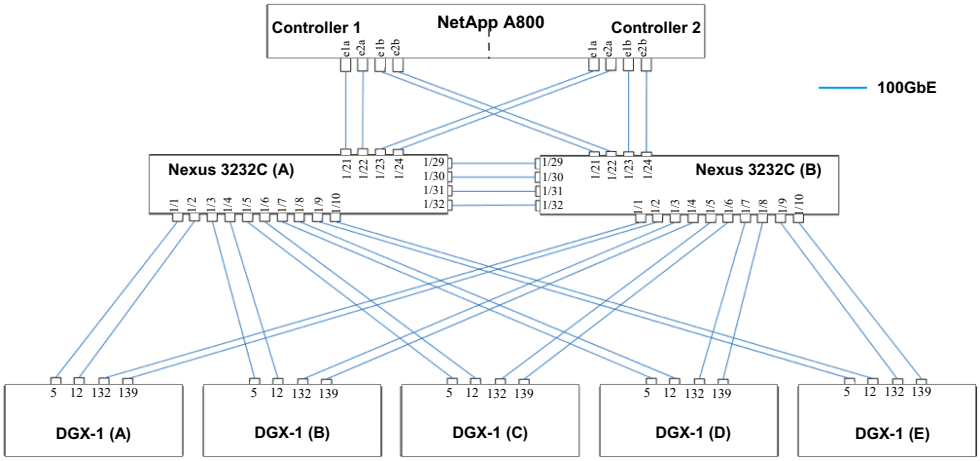


图 2 显示了该架构的端口级别连接。此架构由 2 个 Cisco Nexus 3232C 100GbE 交换机和 1 个 1GbE 管理交换机（未显示）组成。5 台 DGX-1 服务器各自使用 4 条 100GbE 链路连接到网络，而且每台 DGX-1 均通过两条链路连接到每台 Nexus 交换机。存储由 NetApp A800 HA 对提供，2 个存储控制器各自使用 2 条 100GbE 链路与每个 Nexus 交换机相连。

图 2) 显示 1:5 配置的端口级别连接的网络示意图



如果使用 A700s，图 1 中所示架构更改为 1:4 配置（1 个 A700s， 4 台 DGX-1），A700s 通过 4 条 40GbE 链路与存储端的每个交换机相连，每个 DGX-1 通过 2 条 100GbE 链路与每个交换机（共两个）相连。除 100GbE 之外，A800 系统还支持 40GbE。随着数据集规模不断增长，这两种架构均可进行纵向和横向扩展，而且无需停机。

5 机架级扩展 — 从小规模入手，逐步扩展

横向扩展意味着，随着存储环境不断增长，可以在驻留于共享存储基础架构上的资源池中无缝添加更多存储容量和（或）计算节点。主机和客户端连接以及数据存储库均可在资源池中无缝地任意移动。因此，可以轻松地在可用资源上平衡现有工作负载，并且可以轻松部署新工作负载。技术更新（添加或更换驱动器架和/或存储控制器）可以在环境保持联机并持续提供数据时完成。

NetApp 将 DGX-1 服务器的计算能力与 A800 和 A700s 系统的高性能架构相结合，打造了一个极具吸引力的解决方案，支持企业在几小时内完成深度学习工作流部署并根据需要无缝地进行横向扩展。

准备部署深度学习的企业，可以从 1:1 配置入手，然后随着数据增长以横向扩展模式扩展到 1:5 甚至更高比例的配置。表 1 重点说明了使用一系列采用 DGX-1 和 A800（使用 ONTAP 9.4）的配置可实现的容量和性能扩展。

表 1) 使用 A800 横向扩展时的容量和性能指标

A800 存储系统数量	DGX-1 服务器数量	吞吐量	典型原始容量[5]	原始容量（含扩展）[5]
1 个 HA 对	5	25 GB/秒	364.8 TB	6.2 PB

表 1 中的信息基于 A800 和 ONTAP 9.4 性能指标。每个 A800 提供 25 GB/秒的吞吐量，可以在处理来自 5 个 DGX-1 系统的流量的同时，提供扩展至 6.2 PB 存储的选项。

表 2) 使用 A700s 横向扩展时的容量和性能指标

A700s 存储系统数量	DGX-1 服务器数量	吞吐量	典型原始容量[5]	原始容量（含扩展）[5]
1 个 HA 对	4	18 GB/秒	367.2 TB	3.3 PB

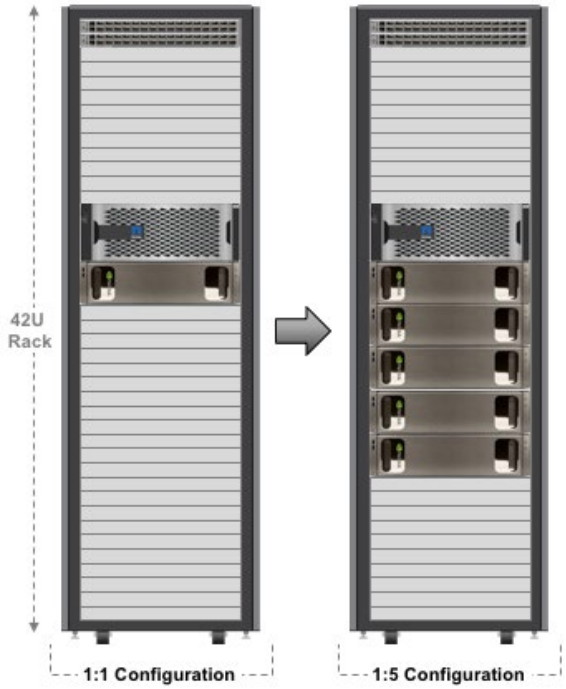
表 2 中的信息基于 A700s 和 ONTAP 9.4 性能指标。A700s 系统可支持 18 GB/秒的吞吐量，适合起步阶段的 1:4 配置。

[5] https://www.netapp.com/cn/media/ds-3582.pdf

需要从更低存储占用空间和成本起步的企业，可使用 NetApp A300 或 A200 存储系统，它们同样可以无缝扩展。

根据表 1 中的扩展信息，图 3 说明了如何在数据中心将 1:1 配置扩展到 1:5 配置部署。采用这种方法，可以根据数据湖大小、所使用深度学习模型以及所需性能指标灵活调整计算与存储的比例。

图 3) 使用 A800 在机架级别从 1:1 扩展到 1:5 配置



每个机架中 DGX-1 服务器和 AFF 存储系统的数量取决于所使用机架的电源和散热规格。系统的最终位置取决于计算流体动力学分析、气流控制和数据中心设计。

6 性能测试

TensorFlow 基准测试是基于 1:1 配置设置（1 台 DGX-1 服务器和 1 个 A700 存储系统）进行的，其中 ImageNet 数据集 (143 GB) 存储在 A700 系统上的 FlexGroup 卷中。为这些测试选择了 NFSv3 作为文件系统。

环境设置：

- 操作系统：Ubuntu 16.04 LTS
- Docker：18.03.1-ce [9ee9f40]
- Docker 文件：nvcr.io/nvidia/tensorflow:18.04-py2
- 框架：Tensorflow 1.7.0
- 基准测试：Tensorflow 基准测试 [26a8b0a]

作为初始测试的一部分，我们基于合成数据运行了基准测试，研究在没有潜在 TensorFlow 管道或存储相关瓶颈情况下的 GPU 性能。在 TensorFlow 基准测试中，我们在 CUDA 核心和 Tensor 核心上针对所有模型使用合成数据运行了训练。

接下来，我们使用失真的真实数据进行了所有测试。

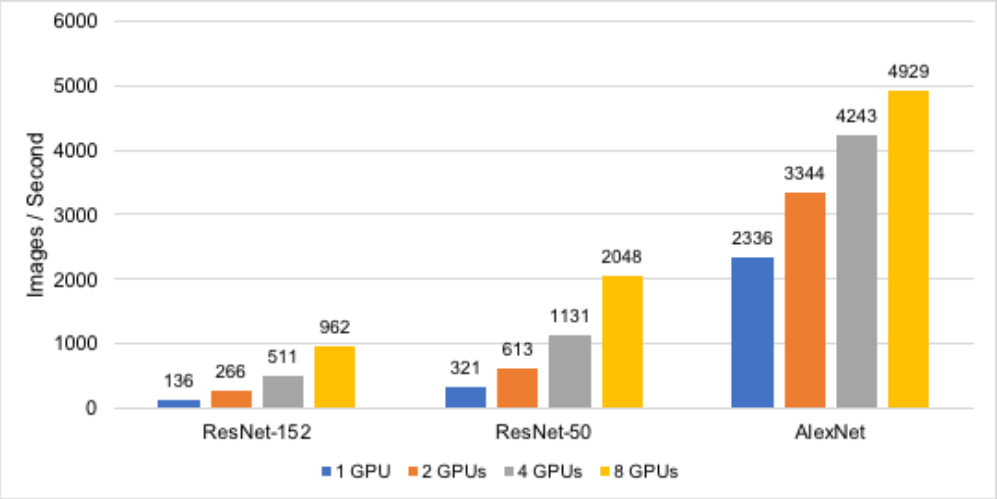
以下几点着重强调了这些测试的突出特点：

- 模型训练性能按每秒处理的图像数量来衡量。
- 为了证明可实现的训练速度，选用了代表不同程度的计算复杂性和预测准确性的三个常见模型：ResNet-152、ResNet-50 和 AlexNet。

- 为了体现真实的模型训练场景，启用了失真（图像预处理步骤）来从存储和 GPU 处理角度对系统施加压力。
- 性能指标使用在 DGX-1 服务器上启用的不同 GPU 数量来衡量。
- 在整个训练阶段，GPU 利用率保持在接近 100% 的水平，这表明 A700 系统在保持高训练速度的同时，能够足够快速地为 GPU 馈送数据。
- AlexNet 作为存储 I/O 最密集的模型，被选择用来对管道施加压力，展示极端使用情形。此模型可能不是最准确的，而且已知在扩展情形下存在一定的限制。
- 使用的批大小 — ResNet-152 和 ResNet-50 为 64，AlexNet 为 512。

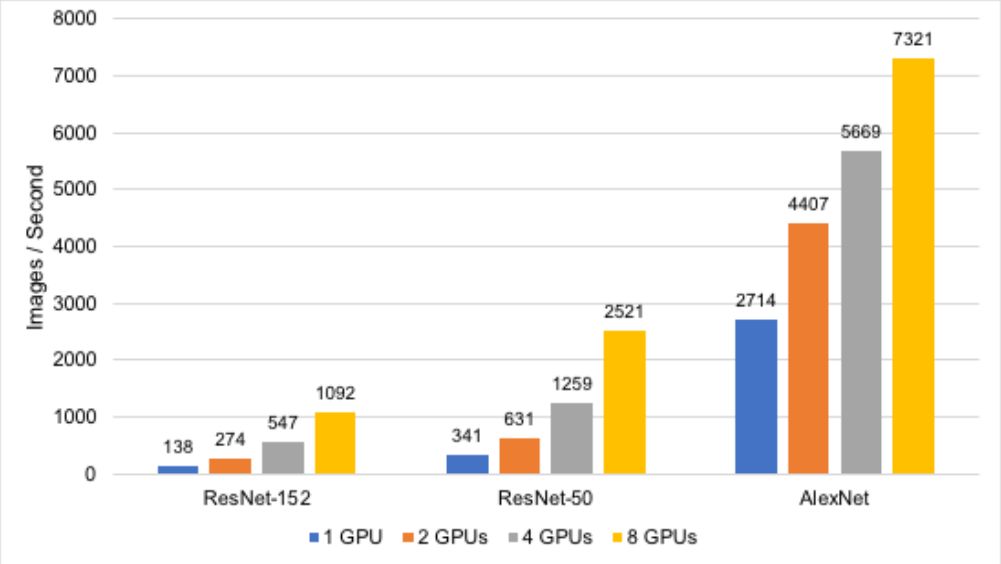
图 4 和图 5 总结了使用 1 个、2 个、4 个和 8 个 GPU 对每个深度学习模型进行测试所得到的训练性能。

图 4) 使用真实数据并启用图像失真时的训练速度



*数据四舍五入到最近的十进制值

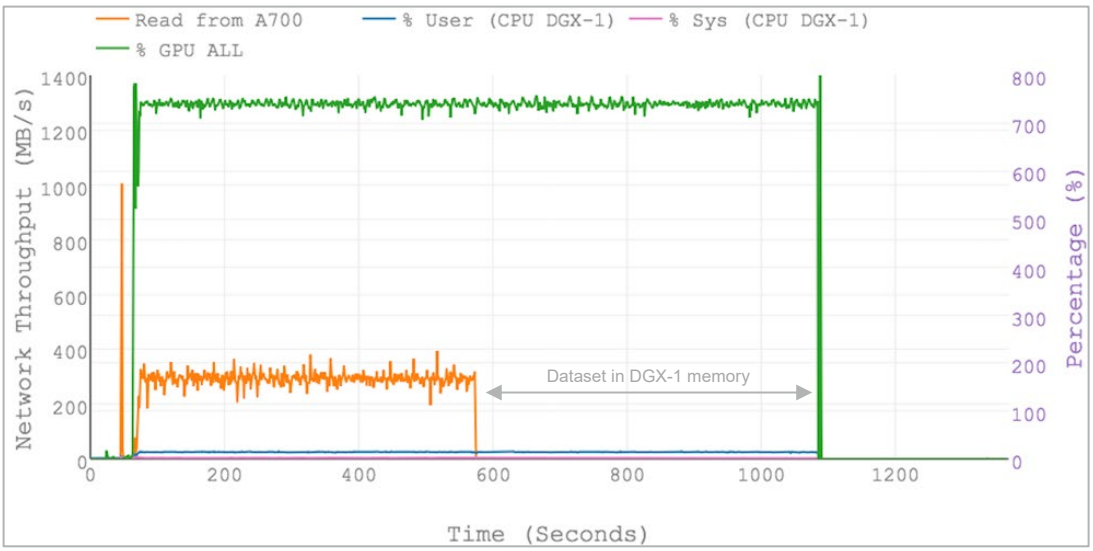
图 5) 使用真实数据并禁用图像失真时的训练速度



*数据四舍五入到最近的十进制值

图 6 说明了使用 8 个 GPU 来训练 ResNet-50 模型时实现的 GPU 利用率。绿色曲线显示了全部 8 个 GPU 的利用率之和，橙色曲线表示 A700 的读取吞吐量。为了保持高 GPU 利用率和每秒约 2500 个图像的训练速度，A700 的读取吞吐量达到约 300 MB/秒。将大小为 143 GB 的数据集加载到 DGX-1 内存中用时约 500 秒。在 500 秒前后的时间点，GPU 利用率和训练速度完全相同。这一点表明，在此训练速度下，为 GPU 馈送数据的管道中不存在存储 I/O 瓶颈或其他瓶颈。

图 6) 以每秒约 2500 个图像的速度训练 ResNet-50 时的 GPU 利用率和 A700 读取吞吐量



7 结束语

AI 需要巨大的计算能力和并驾齐驱的基础架构。深度学习等领域需要极致性能来满足 GPU 的需求，从而为耗用大量数据的算法提供动力。随着 AI 成为一种核心业务能力，几乎所有企业都将依赖于它们生成的大型数据集所产生的洞察力。为了实现目标，企业需要能够快速建立起来且能提供所需并行性能，并能毫不费力地进行扩展和易于管理的基础架构。

DGX-1 服务器可加快模型训练速度，因此可以从大型数据湖中获得洞见。将来自 NGC 的一流 GPU 硬件和经过 GPU 优化的容器相结合，可以快速高效地部署深度学习应用程序。

作为 NFS 领域的行业领导者，NetApp 不仅拥有包括 AFF 产品系列、ONTAP、FlexGroup、Trident 和领先存储效率功能在内的一系列产品，而且具有部署 AI 解决方案的实际专业知识，可满足深度学习应用程序的大规模并行性能要求。

NetApp 与 NVIDIA 合作推出了机架级架构，可帮助企业从小规模入手，然后随着项目数量和数据集大小的增长无缝地对基础架构进行扩展。该架构旨在减轻企业处理日益复杂基础架构的负担，帮助他们集中精力开发更好的深度学习应用程序。采用此类 AI 解决方案可增强企业满足最严苛性能要求的能力，开启全新的智能应用时代。

8 附录：组件列表

表 3 列出了本报告所述架构设计使用的组件。

表 3) 组件列表

组件	数量	说明
基础服务器	1	带 x2 9.6 GT/s QPI 的双 Intel Xeon CPU 主板，8 通道双 DPC DDR4，Intel X99 芯片组，AST2400 BMC
	1	GPU 基板，支持 8 个 SXM2 模块（混合式立方体网格）和 4 个用于 InfiniBand NIC 的 PCIE x16 插槽
连接端口	1	10/100BASE-T IPMI 端口
	1	RS232 串行端口
	2	USB 3.0 端口
CPU	2	Intel Xeon E5-2698 v4，20 个核心，2.2GHz，135W
GPU	8	Tesla V100： 每秒 1000 万亿次浮点运算，混合精度 每个 GPU 32 GB 内存 40,960 个 NVIDIA CUDA® 核心 5120 个 NVIDIA Tensor 核心
系统内存	16	32 GB DDR4 LRDIMM（共 512 GB）
SAS RAID 控制器	1	8 端口 LSI SAS 3108 RAID 夹层卡
存储 (RAID 0)（数据）	4	1.92 TB，6 Gb/秒，SATA 3.0 SSD
存储（操作系统）	1	480 GB，6 Gb/秒，SATA 3.0 SSD
10GbE NIC	1	双端口，10GBASE-T，网络适配器夹层卡
以太网/InfiniBand EDR NIC	4	单端口，x16 PCIe，Mellanox ConnectX-4 VPI MCX455A-ECAT
NetApp A800/A700/A300	1	全闪存阵列，1 个 HA 对
Cisco Nexus 3232C	2	100GbE 以太网交换机
Cisco Nexus 3048-TP	1	1GbE 管理交换机

要验证您的特定环境是否支持本文档所述的确切产品和功能版本，请参见 NetApp 支持站点上的互操作性表工具 (Interoperability Matrix Tool, IMT)。NetApp IMT 中定义的产品组件和版本可用于构建 NetApp 所支持的配置。具体的配置结果取决于每个客户如何依照所发布规格进行安装。

版权信息

商标信息