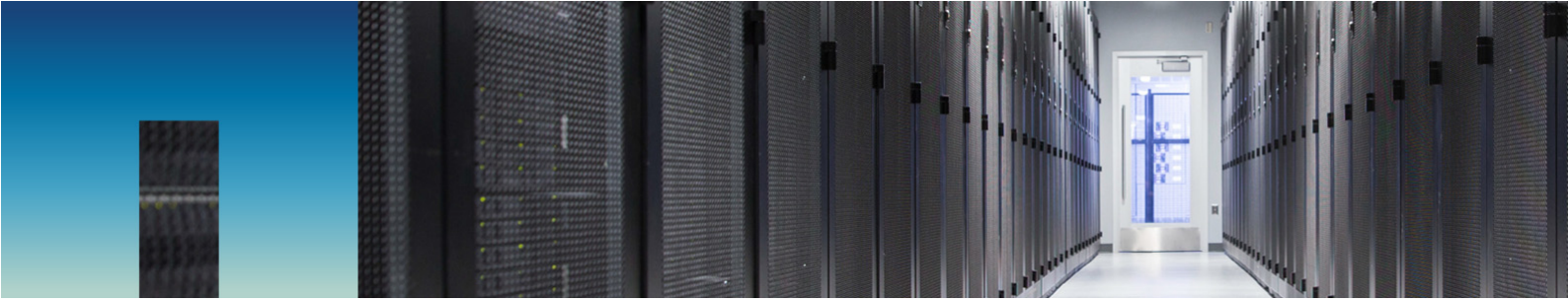


Technisches Whitepaper

Skalierbare KI-Infrastruktur

Entwürfe für reale Deep-Learning-Anwendungsfälle

Sundar Ranganathan, NetApp
Santosh Rao, NetApp

Juni 2018 | WP-7267

Unsere Partner

Executive Summary

Deep Learning (DL) macht schnelle Fortschritte bei einigen der größten heutigen Herausforderungen in der Wissenschaft möglich: In der Medizin werden bessere Möglichkeiten für die Krebsheilung entwickelt, in der Physik wird die Partikelerkennung und -klassifizierung vorangebracht, bei autonomen Fahrzeugen ist das Erreichen der Level-5-Autonomie möglich. Alle haben etwas gemeinsam – Daten. DL basiert auf Daten.

Grafikprozessoren (Graphics Processing Units, GPUs) bieten neue Einblicke, die zuvor nicht möglich waren. Um die strengen Anforderungen der GPUs in einer DL-Applikation zu erfüllen, müssen Storage-Systeme in der Lage sein, konstant Daten an die GPUs weiterzuleiten, und zwar mit niedriger Latenz und hohem Durchsatz, ganz gleich, ob es sich bei den Daten um Text, Bilder, Audio oder Videos handelt.

Wenn Unternehmen von kleineren DL-Implementierungen zur Produktion übergehen, ist es wichtig, eine Infrastruktur aufzubauen, die eine hohe Performance liefert und eine unabhängige und nahtlose Skalierung erlaubt. NVIDIA, als einer der führenden Anbieter von GPUs, bildet zusammen mit der NetApp Innovation der All-Flash-Storage-Systeme eine einzigartige Lösung, um DL-Applikationen zu beschleunigen.

INHALTSVERZEICHNIS

1 Einführung 1

2 Computing – NVIDIA DGX-1 1

2.1 GPU-beschleunigtes Computing 1

2.2 Punktlösung für Deep Learning 1

2.3 Pre-Optimized Systeme der Enterprise-Klasse 2

3 Storage – NetApp All Flash FAS 2

3.1 Hohe Performance 2

3.2 Skalierbarkeit 3

3.3 Robuste Plattformintegration 3

4 Referenzarchitektur 3

5 Rack Scale – klein beginnen und wachsen 5

6 Performance-Tests 6

7 Fazit 8

8 Anhang: Komponentenliste 9

LISTE DER TABELLEN

Tabelle 1) Kapazitäts- und Performance-Kennzahlen in Scale-out-Szenarien mit der A800. 5

Tabelle 2) Kapazitäts- und Performance-Kennzahlen in Scale-out-Szenarien mit den A700s. 5

Tabelle 3) Komponentenliste 9

ABBILDUNGSVERZEICHNIS

Abbildung 1) Referenzarchitektur in einer 1:5-Konfiguration 4

Abbildung 2) Netzwerkdiagramm mit der Verbindung auf Portebene bei einer 1:5-Konfiguration 4

Abbildung 3) Skalieren auf Rack-Ebene von einer 1:1-Konfiguration auf eine 1:5-Konfiguration mit der A800 6

Abbildung 4) Trainingsraten mit realen Daten und mit Bildverzerrungen 7

Abbildung5) Trainingsraten mit realen Daten und ohne Bildverzerrungen 7

Abbildung 6) GPU-Auslastung und A700 Lesedurchsatz mit ResNet-50 bei einer Rate von ~2500 Bildern/Sek. 8

1 Einführung

Wenn IT-Führungskräfte über den Mehrwert von Daten in ihren Unternehmen sprechen, sind die am häufigsten genannten Beispiele Deep Learning (DL) und, etwas allgemeiner, künstliche Intelligenz (KI). DL ist die Engine, die bereits Betrugserkennung im Finanzwesen, prädiktive Wartung in der Fertigung, Spracherkennung bei Kunden-Support-Systemen und verschiedene Autonomie-Level bei Fahrzeugen unterstützt.

In Zukunft werden Daten und DL-Applikationen genutzt werden, um die Produktivität zu verbessern, wichtige Muster zu ermitteln und richtungsweisende Services, Lösungen und Produkte in jeder denkbaren Branche zu entwerfen. Das Marktforschungsunternehmen IDC erwartet, dass die Ausgaben für KI für Software, Services und Hardware von 12 Milliarden US-Dollar im Jahr 2017 auf 57,6 Milliarden im Jahr 2021 steigen werden[1].

DL-Systeme nutzen deutlich verbesserte Algorithmen, da die Größe des neuralen Netzwerks sowie die Menge und Qualität der Daten, mit dem die Modelle geschult werden, gesteigert wurden. Je nach Applikation arbeiten DL-Modelle mit großen Mengen unterschiedlicher Daten, beispielsweise Text, Bilder, Audio, Video oder Zeitreihendaten. Deshalb benötigen DL-Applikationen eine hochperformante Infrastruktur, wie die hier vorgestellte Referenzarchitektur.

Die Implementierung von DL-Workloads auf einer Infrastruktur, die auf die einzigartigen und wachsenden Anforderungen von DL ausgelegt ist und gleichzeitig ein schnelleres Training ermöglicht, ist für den Erfolg der DL-Strategie eines Unternehmens unerlässlich.

Diese Infrastruktur sollte die folgenden grundlegenden Anforderungen erfüllen:

- hohe Computing-Leistung, um Modelle schneller zu trainieren
- hochperformanter Storage, um große Datensätze zu verarbeiten
- Computing und Storage reibungslos und unabhängig skalieren
- verschiedene Arten von Datenverkehr verarbeiten
- Kosten optimieren

Es ist unerlässlich, dass die Infrastruktur eine flexible Skalierung unterstützt. Wenn Computing- und Storage-Bedarf wachsen, verändern sich möglicherweise auch die Performance-Anforderungen und die Verhältnisse von Computing- und Storage-Systemen müssen sich dynamisch ohne Ausfallzeit anpassen lassen.

In diesem Dokument wird eine skalierbare Infrastruktur vorgestellt, die den NVIDIA DGX-1 Server basierend auf der neuen NVIDIA Tesla V100 GPU (Graphics Processing Unit)-Plattform[2] und das neue NetApp A800 All-Flash-Storage-System umfasst.

2 Computing – NVIDIA DGX-1

2.1 GPU-beschleunigtes Computing

Die von DL-Algorithmen durchgeführten Berechnungen erfordern eine immense Menge an Matrix-Multiplikationen, die gleichzeitig ausgeführt werden. Durch ihre enorm parallele Architektur sind moderne GPUs deutlich effizienter als nicht dedizierte Hauptprozessoren (CPUs) für Applikationen, bei denen die Datenverarbeitung gleichzeitig ausgeführt wird. Aufgrund der Fortschritte bei individuellen und geclusterten GPU-Architekturen sind sie nun die bevorzugte Plattform für Workloads wie High-Performance-Computing, DL und Analysen.

2.2 Punktlösung für Deep Learning

Die Zusammenstellung und Integration von Standard-Hardware- und Softwarekomponenten für DL von diversen Anbietern kann zu höherer Komplexität und längeren Implementierungszeiten führen, wodurch wertvolle Datenwissenschaftler viel Zeit mit der Systemintegration verbringen.

[1] Quelle: IDC, Worldwide Storage for Cognitive/AI Workloads Forecast, 2018-2022
[2] Unterstützt durch die NVIDIA Volta Architektur

Sobald die Lösung implementiert ist, stellen viele Unternehmen fest, dass sie sehr viel Zeit mit der Anpassung und Verfeinerung ihres Software-Stacks verbringen, während sich ihre Modelle weiterentwickeln. Als Reaktion darauf erstellte NVIDIA die DGX-1 Severplattform, ein vollständig integriertes und sofort einsatzbereites Hardware- und Softwaresystem, das speziell für DL-Workflows entwickelt wurde.

Jeder DGX-1 Server wird durch acht Tesla V100 GPUs unterstützt, die in einer hybriden Cube-Mesh-Topologie mit NVIDIA NVLink konfiguriert sind, das enorm hohe Bandbreite, eine Fabric mit geringer Latenz für die Kommunikation zwischen den GPUs bietet, was wichtig für das Training mehrerer GPUs ist, um den Engpass in Verbindung mit dem PCIe-basierten Interconnect zu verhindern. Der DGX-1 Server verfügt auch über Netzwerk-Interconnects mit niedriger Latenz und hoher Bandbreite für Multi-Node-Clustering über RDMA-fähige Fabrics.

2.3 Pre-Optimized Systeme der Enterprise-Klasse

Der DGX-1 Server nutzt GPU-optimierte Software-Container der NVIDIA GPU Cloud (NGC), einschließlich Container aller gängigsten DL-Frameworks. Die Deep-Learning-Container von NGC sind vorab auf jeder Ebene optimiert, einschließlich Treiber, Bibliotheken und Kommunikationsprimitive, und bieten maximale Performance für die NVIDA GPUs. Diese vorab integrierten Container isolieren Nutzer von den ständigen Umorganisationen, die typisch sind für die heutigen gängigen Open-Source-DL-Frameworks. IT-Teams verfügen so über ein stabiles, QA-getestetes Stack, auf dem sie DL-Applikationen der Enterprise-Klasse aufbauen können.

Diese vollständig integrierte Hardware- und Softwarelösung, die auf dem Know-how von NVIDIA beruht, beschleunigt die DL-Implementierungen für Applikationen, verkürzt die Trainingszeit von Wochen auf Tage oder Stunden und steigert die Produktivität der Datenforscher, da sie mehr Zeit mit dem Experimentieren anstatt mit Systemintegration und IT-Support verbringen können.

3 Storage – NetApp All Flash FAS

Da GPUs zunehmend schneller werden und die Größe und Komplexität der Datensätze ansteigt, ist das Verwenden von neuesten Storage-Systemen wichtig, um Engpässe zu verhindern und die Systemperformance zu maximieren. DL-Applikationen erfordern eine Storage-Lösung, die darauf ausgelegt ist, im großen Umfang parallel DL-Workloads zu verarbeiten. Sie erfordern ein hohes Maß an Parallelität bei der I/O-Verarbeitung, um das Verzögern von GPUs zu verhindern, während diese auf Daten warten.

In vielen DL-Applikationen umfasst der Datenverkehr die gesamte Daten-Pipeline des gesamten Datacenters (Edge, Core) und in der Cloud. Die Entwicklung einer Storage-Architektur erfordert einen ganzheitlichen Datenmanagementansatz von der Datenaufnahme und/oder Edge-Analyse über die Datenvorbereitung und das Training im Core-Datacenter bis zur Archivierung in der Cloud. Es ist wichtig, die Performance-Anforderungen, die Merkmale verschiedener Datensätze und die erforderlichen Datenservices zu kennen.

Bei DL-Workflows muss eine ideale Storage-Lösung in den folgenden grundlegenden Anforderungen überzeugen.

3.1 Hohe Performance

Die Engpässe in einer DL-Infrastruktur machen sich meistens in der Trainingsphase bemerkbar, wenn eine hohe I/O-Bandbreite mit einer enormen I/O-Parallelität erforderlich ist, um eine kontinuierlich hohe GPU-Auslastung sicherzustellen. Dies zeigt sich darin, wie gut die Storage-Architektur Performance mit hohem Durchsatz bei gleichzeitig niedriger Latenz liefern kann, was sich wiederum in der Unterstützung von ultraschnellen Netzwerkstrukturen zeigt.

Ein einzelnes NetApp A800 System unterstützt einen Durchsatz von 25 GB/s bei sequenziellem Lesen und 1 Million IOPS bei kleinen zufälligen Lesevorgängen mit einer Latenz von weniger als 500µs[3]. Die A800 besticht außerdem mit einer Netzwerkunterstützung von 100 GbE[4], die das Verschieben von Daten beschleunigen und das Trainingssystem insgesamt in Einklang bringen, da der DGX-1 100 GbE RDMA für Cluster Interconnect unterstützt. Das NetApp A700s System unterstützt diverse 40-GbE-Links für einen maximalen Durchsatz von 18 GB/s.

3.2 Skalierbarkeit

Große Datensätze sind wichtig, um die Modellgenauigkeit zu steigern. DL-Implementierungen, die klein beginnen (wenige Terabyte Storage), müssen möglicherweise bald horizontal auf mehrere Petabyte skaliert werden. Außerdem können die Performance-Anforderungen basierend auf dem verwendeten Trainingsmodell variieren und die Endapplikation kann eine unabhängige Skalierung von Computing- und/oder Storage-Ressourcen erfordern. Die Entwicklung einer robusten Systemarchitektur in einer Rack-Scale-Umgebung ermöglicht eine unabhängige Skalierung.

Die NetApp A800 und A700s Systeme lassen sich unabhängig, nahtlos und unterbrechungsfrei von einem 2-Node- (364,8 TB) auf einen 24-Node-Cluster (74,8 PB bei A800, 39,7 PB bei A700s) skalieren. Der Einsatz von ONTAP FlexGroup Volumes ermöglicht einfaches Datenmanagement in einem logischen Single Namespace Volume mit 20 PB und unterstützt so mehr als 400 Milliarden Dateien. Bei Clustern mit einer Kapazität von mehr als 20 PB lassen sich mehrere FlexGroups erstellen, um die erforderliche Kapazität abzudecken.

3.3 Robuste Plattformintegration

Da Unternehmen die Rate ihrer Datenerfassung beschleunigen müssen, ist der Automatisierungsbedarf rund um diese Daten offensichtlich. Ein Ansatz stellt das Verwenden von Containern dar. Es ermöglicht schnellere Implementierungen durch das Trennen von Applikationen vom OS und den Abhängigkeiten auf Gerätetreiberebene. Ein effizientes und einfaches Datenmanagement ist entscheidend für die Verkürzung der Trainingszeit.

Trident ist eine dynamische NetApp Storage-Orchestrierungslösung für Container Images, die vollständig in Docker und Kubernetes integriert ist. In Kombination mit NVIDIA GPU Cloud (NGC) und gängigen Orchestrierungslösungen wie Kubernetes oder Docker Swarm, können Unternehmen mit Trident nahtlos ihre KI/DL NGC Container Images auf NetApp Storage implementieren und so bei ihrem Einsatz von KI-Containern von einer Erfahrung der Enterprise-Klasse profitieren. Dies umfasst automatisierte Orchestrierung, Klonen für Test und Entwicklung, NGC Upgrade-Tests mit Klon-, Schutz- und Compliance-Kopien und viele weitere Datenmanagement-Anwendungsfälle für die NGC KI Container Images.

Datenverkehr bei DL kann aus Millionen von Dateien bestehen (Bilder, Video/Audio, Textdateien). Network File Systems (NFS) sind ideal, um hohe Performance über verschiedene Workloads hinweg zu erzielen: Sie verarbeiten sowohl zufälligen als auch sequenziellen I/O. In Kombination mit ONTAP FlexGroup Volumes kann NetApp All Flash FAS hohe Performance für kleine Datei-Workloads liefern, die über Storage-Systeme verteilt sind.

4 Referenzarchitektur

Das Erstellen einer Infrastruktur, die kontinuierlich hohen I/O-Durchsatz bei geringer Latenz liefern kann, um die I/O-Parallelität der GPUs zu unterstützen, und die unterbrechungsfreie Skalierung von Computing- und Storage-Systemen sind wichtig, um schnellere Trainingszeiten zu erzielen. Für diese Anforderungen ist ein Storage-System mit ultraschneller Netzwerkstruktur mit hoher Bandbreite und niedriger Latenz nötig. Nur so können die Systemperformance maximiert und verschiedene DGX-1 Server konstant mit Daten versorgt werden, die für jeden DL-Trainingsdurchlauf erforderlich sind.

[3] https://blog.netapp.com/the-future-is-here-ai-ready-cloud-connected-all-flash-storage-with-nvme/

[4] https://www.netapp.com/de/products/storage-systems/all-flash-array/aff-a-series.aspx#technical-specifications

Abbildung 1 zeigt eine NetApp Architektur in einer 1:5-Konfiguration, die aus fünf DGX-1 Servern besteht, die von einem A800 HA-Paar über zwei Switches bedient werden. Jeder DGX-1 Server ist mit jedem der beiden Switches über zwei 100-GbE-Links verbunden. Die A800 ist über vier 100-GbE-Links mit jedem Switch verbunden. Die Switches können über zwei bis vier 100-Gb-Inter-Switch-Links für Failover-Szenarien verfügen. Das High-Availability-Design ist Aktiv/Aktiv, damit der maximale Durchsatz über alle Netzwerkverbindungen hinweg aufrecht erhalten werden kann, wenn der Betrieb normal läuft.

Abbildung 1) Referenzarchitektur in einer 1:5-Konfiguration

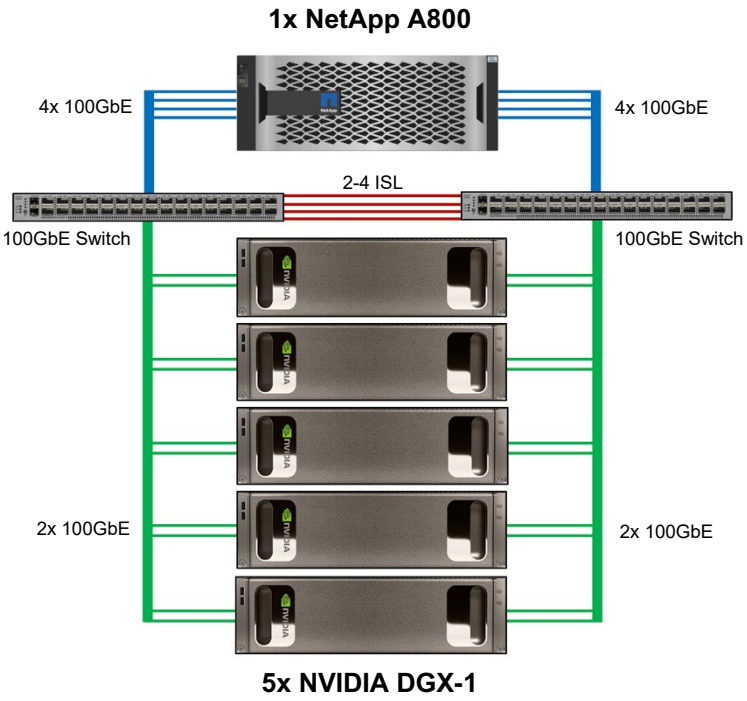
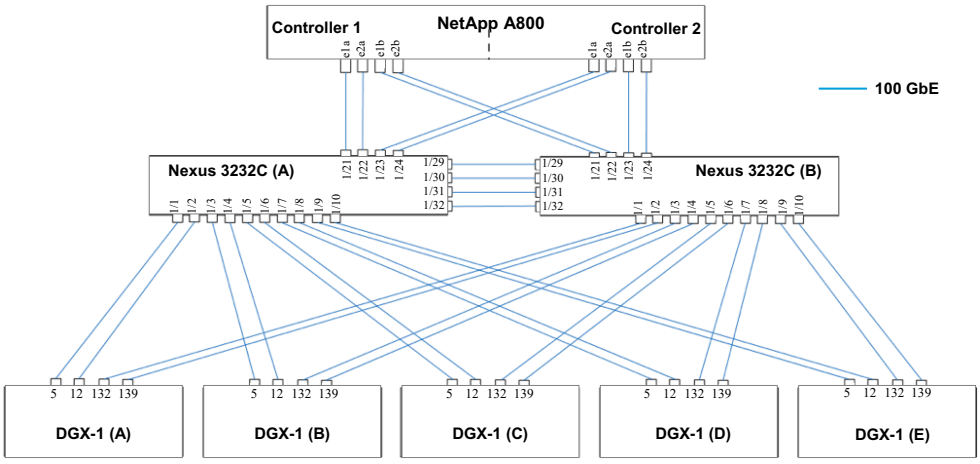


Abbildung 2 zeigt die Verbindung der Architektur auf Portebene. Diese Architektur besteht aus zwei Cisco Nexus 3232C 100-GbE-Switches sowie einem 1-GbE-Management-Switch (nicht abgebildet). Fünf DGX-1 Server sind mit jeweils vier 100-GbE-Links mit dem Netzwerk verbunden. Jeder DGX-1 ist mit zwei Links mit jedem Nexus Switch verbunden. Storage wird durch ein NetApp A800 HA-Paar bereitgestellt, bei dem jeder der beiden Storage-Controller mit jedem Nexus Switch mit zwei 100-GbE-Links verbunden ist.

Abbildung 2) Netzwerkdiagramm mit der Verbindung auf Portebene bei einer 1:5-Konfiguration



Mit den A700s wird aus der in Abbildung 1 gezeigten Architektur eine 1:4-Konfiguration (1 A700s : 4 DGX-1), mit vier 40-GbE-Links von A700s an jeden Switch auf der Storage-Seite und zwei 100-GbE-Links von jedem DGX-1 zu jedem der beiden Switches. Neben 100 GbE unterstützt das A800 System auch 40 GbE. Beide Architekturen lassen sich ohne Ausfallzeit vertikal und horizontal skalieren, wenn der Umfang der Datensätze steigt.

5 Rack Scale – klein beginnen und wachsen

Horizontale Skalierung bedeutet, dass bei wachsender Storage-Umgebung zusätzliche Storage-Kapazität und/oder Computing-Nodes nahtlos zum Ressourcen-Pool in der Shared-Storage-Infrastruktur hinzugefügt werden. Host- und Client-Verbindungen sowie Datastores können sich nahtlos im Ressourcen-Pool bewegen. So können bestehende Workloads einfach über verfügbare Ressourcen ausgeglichen werden und neue Workloads können einfach implementiert werden. Technologieaktualisierungen (Hinzufügen oder Ersetzen von Laufwerk-Shelfs und/oder Storage-Controllern) werden durchgeführt, wobei die Umgebung online bleiben und weiterhin Daten bereitstellen kann.

NetApp hat die Computing-Leistung von DGX-1 Servern mit der hochperformanten Architektur der A800 und A700s Systeme kombiniert, um eine überzeugende Lösung anzubieten, mit der Unternehmen DL-Workflows in wenigen Stunden implementieren und nach Bedarf nahtlos horizontal skalieren können.

Unternehmen, für die DL Neuland ist, beginnen möglicherweise mit einer 1:1-Konfiguration und skalieren diese horizontal, wenn die Daten zunehmen, auf eine 1:5-Konfiguration oder mehr im Scale-out-Modus. Tabelle 1 hebt die Kapazitäts- und Performance-Skalierung hervor, die mit einem Spektrum von Konfigurationen mit DGX-1 und A800 mithilfe von ONTAP 9.4 erzielt werden kann.

Tabelle 1) Kapazitäts- und Performance-Kennzahlen in Scale-out-Szenarien mit der A800.

Anzahl der **A800** Storage-Systeme	Anzahl der DGX-1 Server	Durchsatz	Typisch Bruttokapazität[5]	Bruttokapazität mit Erweiterung[5]
1 HA-Paar	5	25 GB/s	364,8 TB	6,2 PB

Die Informationen in Tabelle 1 basieren auf den Performance-Kennzahlen von A800 und ONTAP 9.4. Jede A800 bietet einen Durchsatz von 25 GB/s und ist in der Lage, Datenverkehr von 5 DGX-1 Systemen zu verarbeiten. Gleichzeitig besteht die Möglichkeit, eine Skalierung auf 6,2 PB Storage vorzunehmen.

Tabelle 2) Kapazitäts- und Performance-Kennzahlen in Scale-out-Szenarien mit den A700s.

Anzahl der **A700s** Storage-Systeme	Anzahl der DGX-1 Server	Durchsatz	Typisch Bruttokapazität[5]	Bruttokapazität mit Erweiterung[5]
1 HA-Paar	4	18 GB/s	367,2 TB	3,3 PB

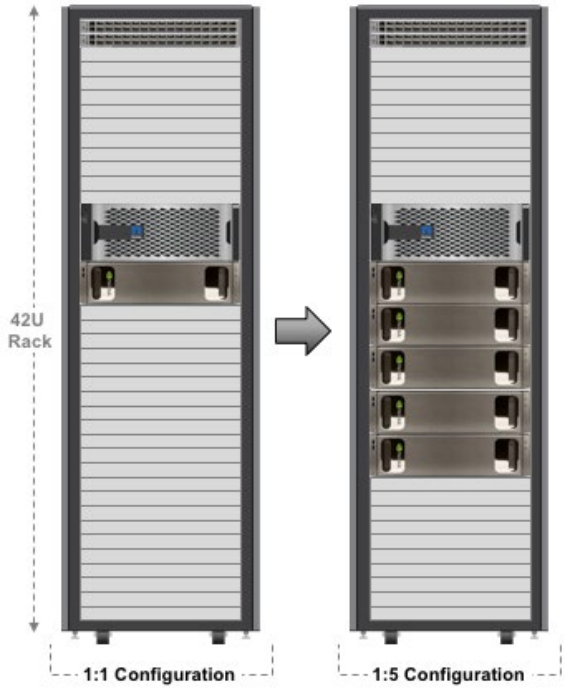
Die Informationen in Tabelle 2 basieren auf den Performance-Kennzahlen von A700s und ONTAP 9.4. Das A700s System unterstützt einen Durchsatz von 18 GB/s und eignet sich ideal als Einstieg für eine 1:4-Konfiguration.

Für Unternehmen, die mit einem niedrigeren Storage-Bedarf und geringeren Kosten einsteigen möchten, eignen sich die NetApp A300 oder A200 Storage-Systeme, die auch eine nahtlose Skalierbarkeit unterstützen.

Basierend auf den Skalierungsinformationen in Tabelle 1, zeigt Abbildung 3, wie sich eine 1:1-Konfiguration auf eine 1:5-Konfiguration in einem Datacenter skalieren lässt. Bei diesem Ansatz besteht die Möglichkeit, das Verhältnis zwischen Computing- und Storage-Ressourcen basierend auf der Größe des Data Lake, der verwendeten DL-Modelle und der erforderlichen Performance-Kennzahlen zu verändern.

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

Abbildung 3) Skalieren auf Rack-Ebene von einer 1:1-Konfiguration auf eine 1:5-Konfiguration mit der A800



Die Anzahl der DGX-1 Server und All Flash FAS Storage-Systeme pro Rack hängt von den Leistungs- und Kühlungsdaten des verwendeten Racks ab, während die letztendliche Platzierung des Systems von CFD-Analysen (Computational Fluid Dynamics), Luftstrommanagement und Datacenter-Design abhängt.

6 Performance-Tests

TensorFlow Benchmarks wurden in einer Installation mit einer 1:1-Konfiguration ausgeführt (ein DGX-1 Server und ein A700 Storage-System), wobei der ImageNet Datensatz (143 GB) auf einem FlexGroup Volume auf dem A700 System gespeichert war. NFSv3 war bei diesen Tests das Filesystem der Wahl.

Umgebungs-Setup:

- OS: Ubuntu 16.04 LTS
- Docker: 18.03.1-ce [9ee9f40]
- Dockerfile: nvcr.io/nvidia/tensorflow:18.04-py2
- Framework: Tensorflow 1.7.0
- Benchmarks: Tensorflow Benchmarks [26a8b0a]

Bei unseren anfänglichen Tests führten wir Benchmarks basierend auf synthetischen Daten aus, um die Performance der GPUs ohne potenzielle TensorFlow Pipeline oder Storage-basierte Engpässe zu untersuchen. Das Training wurde mit synthetischen Daten auf CUDA Cores und Tensor Cores für alle Modelle im TensorFlow Benchmark durchgeführt.

Im nächsten Schritt wurden reale Daten mit Verzerrung für alle Tests verwendet.

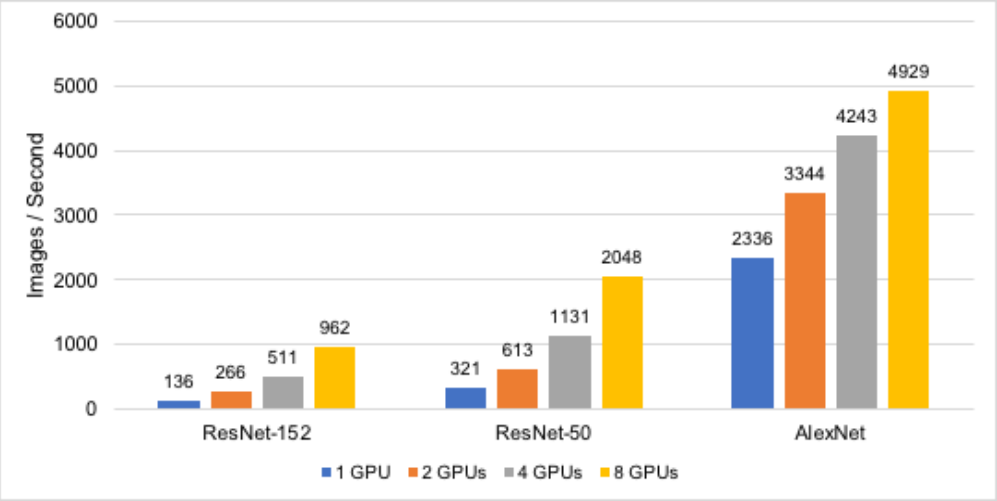
Die folgenden Punkte heben die zentralen Aspekte des Tests hervor:

- Die Trainings-Performance der Modelle wird als Anzahl der pro Sekunde verarbeiteten Bilder gemessen.
- Um die erzielbaren Trainingsraten zu demonstrieren, wurden drei gängige Modelle gewählt, die ein unterschiedliches Maß an Computing-Komplexität und prädiktiver Präzision repräsentieren – ResNet-152, ResNet-50 und AlexNet.

- Um Modell-Trainingsszenarien aus dem realen Leben zu simulieren, wurde Verzerrung (Bildvorverarbeitungsschritte) verwendet, um das System aus Sicht der Storage- und GPU-Verarbeitung zu belasten.
- Die Performance-Kennzahlen wurden mit unterschiedlich vielen aktivierten GPUs auf dem DGX-1 Server gemessen.
- Die GPU-Auslastungen lagen bei knapp 100 % während des gesamten Trainingszeitraums, was zeigt, dass das A700 System in der Lage ist, Daten schnell genug an die GPUs weiterzugeben, während hohe Trainingsraten beibehalten werden.
- AlexNet, das Modell mit den höchsten Storage-I/O-Anforderungen, wurde gewählt, um die Pipeline zu belasten und extreme Anwendungsfälle zu demonstrieren. Dieses Modell ist möglicherweise nicht das präziseste und ist für Einschränkungen bei Skalierungsszenarien bekannt.
- Die Batch-Größe betrug 64 für ResNet-152 und ResNet-50 und 512 für AlexNet.

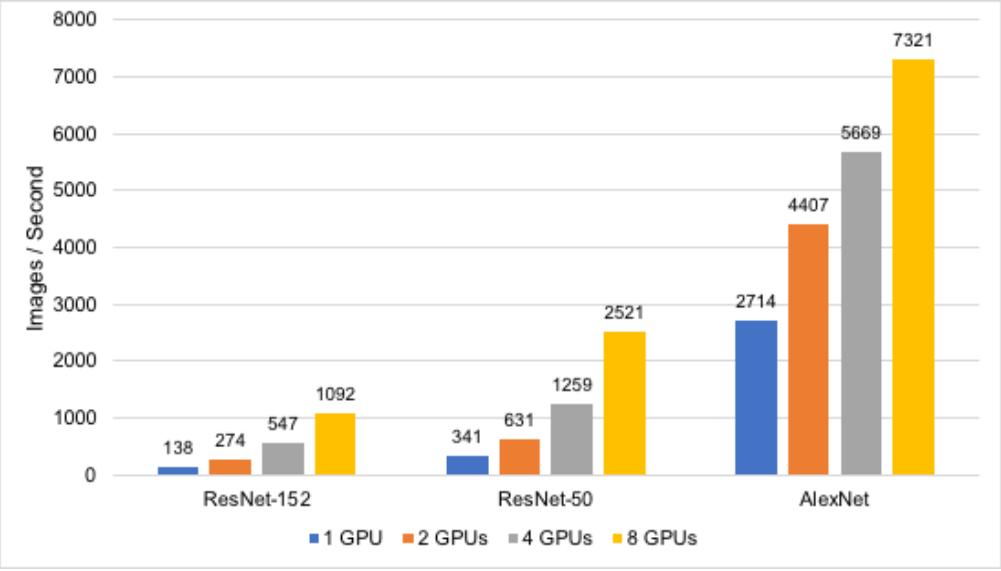
Die Abbildungen 4 und 5 fassen die Trainings-Performance zusammen, die mit jedem der drei DL-Modelle mit einer, zwei, vier und acht GPUs gemessen wurde.

Abbildung 4) Trainingsraten mit realen Daten und mit Bildverzerrungen



* Daten gerundet auf den nächsten Dezimalwert

Abbildung5) Trainingsraten mit realen Daten und ohne Bildverzerrungen



* Daten gerundet auf den nächsten Dezimalwert

Abbildung 6 zeigt die GPU-Auslastungen, die bei dem Training des ResNet-50 Modells mit acht GPUs erzielt wurde. Die grüne Kurve zeigt die Auslastungen aller acht GPUs und die orangefarbene Kurve zeigt den Lesedurchsatz von A700. Mit einem Lesedurchsatz von ~300 MB/s der A700 wird die hohe GPU-Auslastung beibehalten und eine Trainingsrate von ~2.500 Bildern/Sek. erzielt. Es dauerte ~500 Sekunden, um den 143 GB großen Datensatz in den DGX-1 Speicher zu laden. Vor und nach der 500-Sekunden-Marke sind die GPU-Auslastung und die Trainingsraten identisch. Das zeigt, dass es bei dieser Trainingsrate keine Storage-I/O- oder andere Engpässe in der Pipeline gab, die die GPUs bedienten.

Abbildung 6) GPU-Auslastung und A700 Lesedurchsatz mit ResNet-50 bei einer Rate von ~2500 Bildern/Sek.

7 Fazit

KI benötigt eine enorme Computing-Leistung und eine Infrastrukturarchitektur, die Schritt halten kann. Disziplinen wie DL erfordern extreme Performance, um GPUs zu bedienen, die datenhungrige Algorithmen unterstützen. Wenn KI mehr und mehr zur Kernfunktion wird, werden sich praktisch alle Unternehmen auf Einblicke verlassen, die aus den großen von ihnen generierten Datensätzen stammen. Um ihre Ziele zu erreichen, benötigen sie eine Infrastruktur, die schnell einsatzbereit ist, die erforderliche parallele Performance bietet, sich mühelos skalieren und einfach managen lässt.

Bei beschleunigten Trainingsgeschwindigkeiten der Modelle lassen sich über den DGX-1 Server Einblicke aus massiven Data Lakes sammeln. Durch die Kombination von neuster GPU-Hardware und GPU-optimierten Containern von NGC lassen sich DL-Applikationen schnell und effizient bereitstellen.

Als einer der Branchenführer im Bereich NFS verfügt NetApp über eine Reihe von Produkten wie die All Flash FAS Produktfamilie, ONTAP, FlexGroup, Trident und wichtige Storage-Effizienz-Funktionen, um die hohen Anforderungen an parallele Performance von DL-Applikationen zu erfüllen, und über praktische Kenntnisse bei dem Implementieren von KI-Lösungen.

NetApp hat gemeinsam mit NVIDIA eine Rack-Scale-Architektur entwickelt. Damit können Unternehmen mit einer kleinen Infrastruktur beginnen und diese nahtlos erweitern, sobald die Anzahl der Projekte und die Größe der Datensätze steigt. Diese Architektur soll Unternehmen vor zunehmender Infrastrukturkomplexität bewahren und ihnen helfen, sich auf das Entwickeln besserer DL-Applikationen zu konzentrieren. Mithilfe dieser KI-Lösungen erfüllen Unternehmen selbst noch so anspruchsvolle Performance-Anforderungen, was zu einer neuen Ära intelligenter Applikationen führt.

8 Anhang: Komponentenliste

Tabelle 3 führt die Komponenten für die in diesem Bericht beschriebenen verwendeten Architekturdesigns auf.

Tabelle 3) Komponentenliste

Komponente	Menge	Beschreibung
Basis-Server	1	Dual Intel Xeon CPU Motherboard mit x2 9,6 GT/s QPI, 8 Channel mit 2 DPC DDR4, Intel X99 Chipsatz, AST2400 BMC
	1	GPU-Baseboard, das 8 SXM2 Module (hybride Cube-Mesh) und 4 PCIE x16 Steckplätze für InfiniBand NICs unterstützt
Konnektivitätsports	1	10/100BASE-T-IPMI-Port
	1	Serieller RS232-Port
	2	3.0 USB-Ports
CPU	2	Intel Xeon E5-2698 v4; 20-Core; 2,2 GHz; 135 W
GPU	8	Tesla V100: 1 petaFLOPS, gemischte Präzision 32 GB Speicher pro GPU 40.960 NVIDIA CUDA Cores 5.120 NVIDIA Tensor Cores
Systemspeicher	16	32 GB DDR4 LRDIMM (insgesamt 512 GB)
SAS Raid-Controller	1	8-Port LSI SAS 3108 RAID Mezzanine
Storage (RAID 0) (Daten)	4	1,92 TB, 6 Gb/s, SATA 3.0 SSD
Storage (OS)	1	480 GB, 6 Gb/s, SATA 3.0 SSD
10 GbE NIC	1	Dual Port, 10 GBASE-T, Netzwerkadapter Mezzanine
Ethernet/InfiniBand EDR NIC	4	Single Port, x16 PCIe, Mellanox ConnectX-4 VPI MCX455A-ECAT
NetApp A800/A700/A300	1	All-Flash-Array, 1 HA-Paar
Cisco Nexus 3232C	2	100 Gb Ethernet Switches
Cisco Nexus 3048-TP	1	1 Gb Ethernet Management-Switch

Überprüfen Sie mithilfe des Interoperability-Matrix-Tools (IMT) auf der NetApp Support-Website, ob die in diesem Dokument angegebenen Produktversionen und Funktionen in Ihrer IT-Umgebung unterstützt werden. Das NetApp IMT ermittelt die Produktkomponenten und -versionen, die zu einer von NetApp unterstützten Konfiguration kombiniert werden können. Die jeweiligen Ergebnisse sind von der kundenspezifischen Installation bzw. den technischen Daten abhängig.

Copyright-Informationen

Markeninformationen