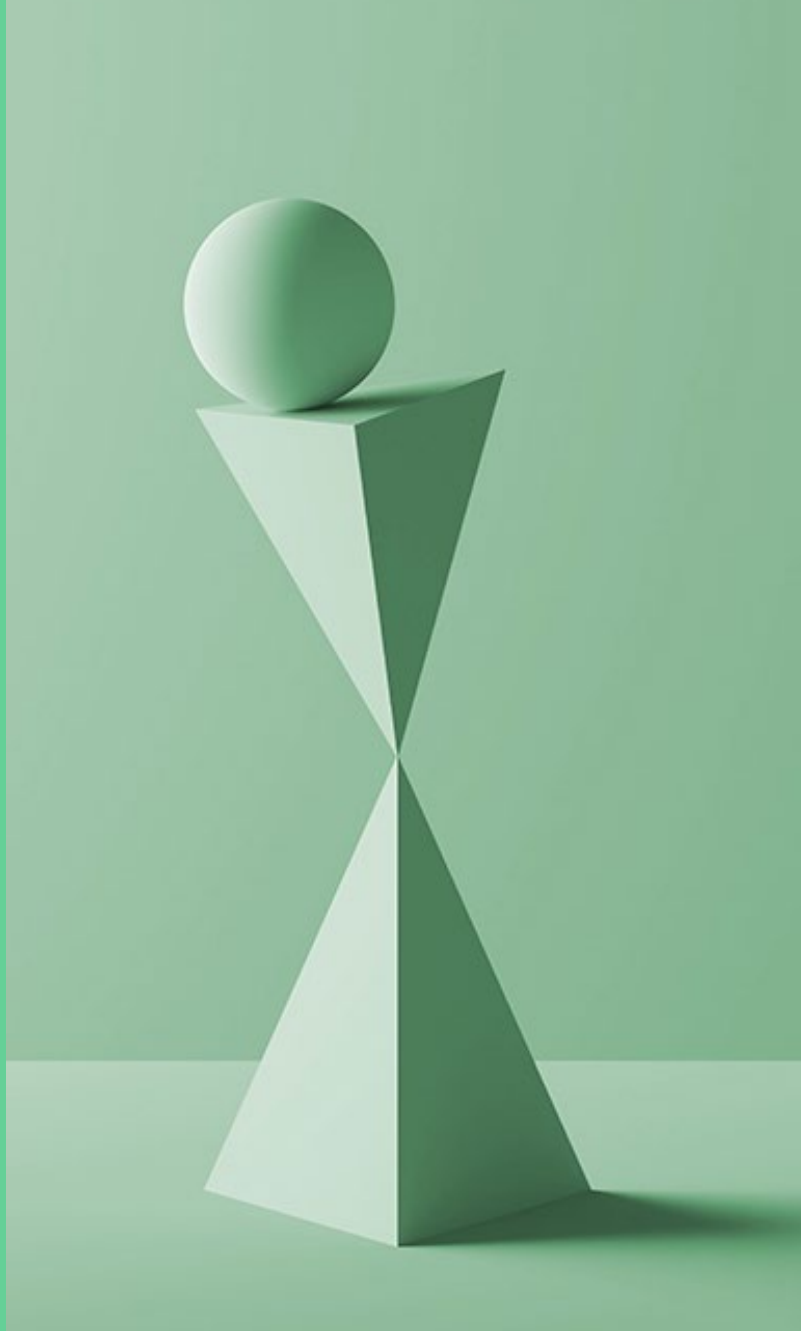
LÖSUNGSÜBERBLICK

# ONTAP AI

Vereinfachen, beschleunigen und integrieren Sie mit NetApp und NVIDIA Ihre Datenpipeline für ML und DL



NetApp | NVIDIA

**Herausforderungen bei der KI-Infrastruktur**
Mithilfe von künstlicher Intelligenz (KI), Machine Learning (ML) und Deep Learning (DL) können Unternehmen Betrugsversuche erkennen, Kundenbeziehungen verbessern, die Lieferkette optimieren und in einem immer härter umkämpften Markt innovative Produkte und Dienstleistungen anbieten. Womöglich gehört Ihr Unternehmen zu den vielen, die neue DL-Ansätze nutzen möchten, um den digitalen Wandel voranzutreiben und sich einen Wettbewerbsvorteil zu sichern. Um im größtmöglichen Umfang von Deep Learning zu profitieren, müssen Sie zunächst mehrere wichtige Hürden überwinden.

Eine Integration in Eigenregie durchzuführen, ist ein komplexes Unterfangen. Die Zusammenstellung und Integration von Compute-, Storage-, Netzwerk- und Software-Standardkomponenten für ML und DL kann die Komplexität erhöhen und die Implementierungszeiten verlängern. Dadurch verschwenden Data Scientists wertvolle Zeit auf die Systemintegration.

Es ist schwer, eine vorhersagbare und skalierbare Performance zu erreichen. Gemäß den Best Practices für DL sollten Unternehmen klein anfangen und ihre Ressourcen erst mit der Zeit skalieren. Traditionell wurden Computing und Direct-Attached Storage eingesetzt, um Daten in KI-Workflows einzuspeisen. Die Skalierung mit herkömmlichem Storage kann allerdings zu Unterbrechungen und Ausfallzeiten für die laufenden Vorgänge führen.

Unterbrechungen wirken sich negativ auf die Produktivität der Data Scientists aus. Bei der ML- und DL-Infrastruktur bestehen zahlreiche Abhängigkeiten zwischen Hard- und Software; zum Aufrechterhalten einer funktionierenden DL-Infrastruktur braucht man fundierte, umfassende KI-Kenntnisse. Ausfallzeiten oder eine langsame KI-Performance können eine Kettenreaktion auslösen, die die Entwicklerproduktivität in Mitleidenschaft zieht und die Betriebskosten unkontrolliert in die Höhe schnellen lässt.

**Die Lösung**
Jetzt können Sie das Versprechen von KI, ML und DL vollständig verwirklichen – vereinfachen, beschleunigen und integrieren Sie Ihre Datenpipeline mit der bewährten Architektur von NetApp ONTAP AI, gestützt auf die Power von NVIDIA DGX-Systemen und NetApp All-Flash-Storage mit Cloud-Integration. Mittels einer Data-Fabric-Architektur optimieren Sie den Weg der Daten zuverlässig zwischen ihrem Entstehungsort (Edge, Peripherie), dem Rechenzentrum (Core) und der Cloud und beschleunigen Analysen, Training und Inferenz.

**Die wichtigsten Vorteile**

**Geringere Risiken dank flexibler validierter Lösungen**
- Designkomplexität und Unsicherheiten ausschließen
- Konfiguration und Implementierung durch verfügbare vorkonfigurierte Lösungen optimieren

**Passende Performance und Skalierbarkeit**
- Klein beginnen und unterbrechungsfrei wachsen
- Mit einer hochperformanten Lösung schneller Ergebnisse erhalten

**Aufbau einer integrierten Datenpipeline**
- Daten mit einer integrierten Pipeline zwischen Entstehungsort, Rechenzentrum und Cloud intelligent managen
- Auf KI-Fachwissen und einfache Support-Optionen gestützte Lösung implementieren

**Vereinheitlichte KI-Workloads**
- Infrastruktursilos beseitigen
- Flexibel auf Geschäftsanforderungen reagieren

NetApp ONTAP AI verfügt als einer der ersten konvergenten Infrastruktur-Stacks über NVIDIA DGX A100, das weltweit erste KI-System mit 5 PetaFLOPs, sowie über hochperformante NVIDIA Mellanox Ethernet-Switches. Sie profitieren dadurch von einheitlichen KI-Workloads, einer einfacheren Implementierung und einem schnellen Return on Investment.

„Deep Learning revolutioniert nahezu jeden Markt, in dem wir aktiv sind. Wir setzen Deep Learning in verschiedenartigen Märkten ein und schaffen damit völlig neue Möglichkeiten. Die auf NVIDIA DGX-Systemen und NetApp All-Flash-Storage basierende Lösung NetApp ONTAP AI vereinfacht und beschleunigt die Datenpipeline für Deep-Learning-Prozesse."

**Tim Ensor, Director of Artificial Intelligence Cambridge Consultants**

Abbildung 1: ONTAP AI Architekturen mit DGX A100; Konfigurationen mit zwei, vier und acht Nodes

**Geringere Risiken dank flexibler validierter Lösungen**
Die hohe Innovationsgeschwindigkeit im KI-Bereich macht die Gestaltung einer effektiven KI-Infrastruktur zu einer Herausforderung. Mit ONTAP AI können Sie Unsicherheiten ausräumen und durch Einsatz einer praxiserprobten Referenzarchitektur schneller starten. Alternativ steht Ihnen eine vorkonfigurierte integrierte Lösung zur Verfügung, die sich unkompliziert beschaffen und bereitstellen lässt – damit werden Design und Management unkomplizierter.

Die integrierte Lösung ONTAP AI ist in vier vorkonfigurierten Optionen mit Kapazitätserweiterung und optional erweiterter Software verfügbar. Bei dieser integrierten Lösung wird die Komplexität noch stärker verringert – durch Installation vor Ort und umfassenden Support mit einer zentralen Rufnummer für alle Schritte von der Problemmeldung bis zur Problemlösung.

**Passende Performance und Skalierbarkeit**
DL-Trainingsroutinen erfordern eine enorme Rechenleistung. Ein schnelleres Image-Training kann die Gesamtrechenkosten senken und gleichzeitig KI-Innovation und Produktivität steigern.

Das DGX-A100-System basiert auf der neuen NVIDIA Ampere-Architektur und liefert im Vergleich zur vorherigen Generation eine bis zu 6-mal höhere Trainings-Performance. Sie erhalten damit die Computing-Infrastruktur eines Datacenters für Analyse, Training und Inferenz – jetzt konsolidiert in einem Einzelsystem. Im Vergleich zu CPU-Systemen benötigt das DGX-A100-System 1/25 der Stellfläche und 1/20 der Energie – bei nur 1/10 der Kosten.

Die Investition in hochmodernes Computing erfordert ebenso hochmodernen Storage, der Tausende von Training-Images pro Sekunde verarbeiten kann. Dazu ist eine hochperformante Datenservices-Lösung nötig, die mit den anspruchsvollsten DL-Trainings-Workloads Schritt hält.

Bei NetApp All-Flash-Storage können Sie einen kontinuierlichen Durchsatz von 2 GB/s (5 GB/s Spitze) erwarten. Die Latenz liegt deutlich unter 1 Millisekunde, während die GPUs mit einer Auslastung von über 95 % arbeiten. Ein einzelnes NetApp AFF A800 System unterstützt einen Durchsatz von 25 GB/s bei sequenziellen Lesevorgängen und 1 Million IOPS bei kleinen zufälligen Lesevorgängen, mit einer Latenz von weniger als 500 Mikrosekunden bei NAS-Workloads.

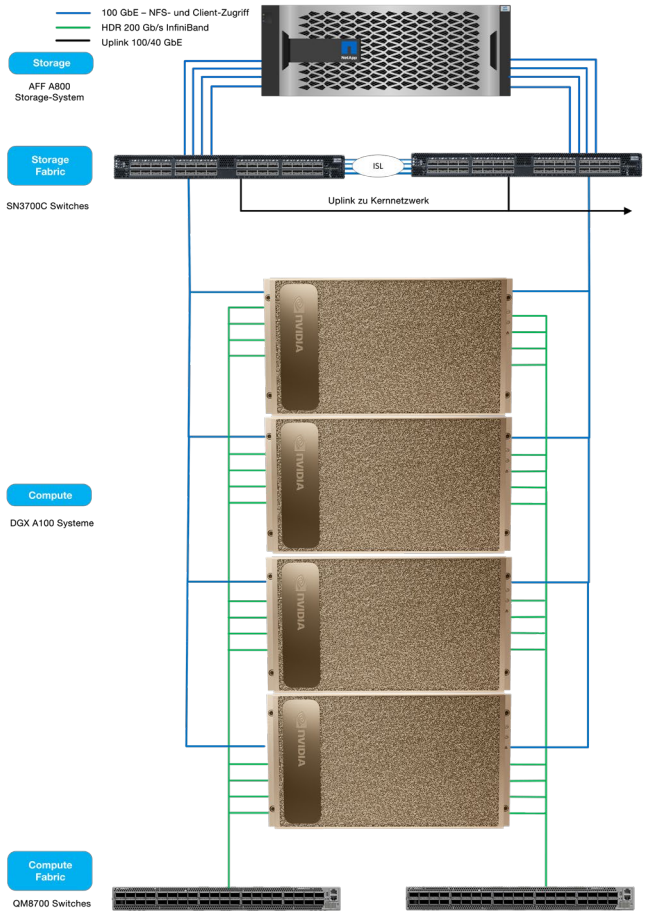



Abbildung 2: ONTAP AI Konfiguration mit vier Nodes und Mellanox Spectrum 100-GbE-Switches

Mit der Rack-Scale-Architektur von NetApp haben Sie in Ihrem Unternehmen die Möglichkeit, von einigen hundert Terabyte auf ein All-Flash-System im zweistelligen Petabyte-Bereich zu skalieren. Dank NetApp ONTAP FlexGroup kann außerdem ein Single Namespace mit bis zu 20 PB mehr als 400 Milliarden Dateien verarbeiten.

## Aufbau einer integrierten Datenpipeline zwischen Edge, Core und Cloud

ONTAP AI nutzt Ihre Data Fabric, um das Management von Daten in der gesamten Datenpipeline auf einer einzigen Plattform zusammenzuführen. Dieselben Tools können Sie dazu einsetzen, um Ihre aktiven und ruhenden Daten zu kontrollieren und zu schützen. Nebenbei erfüllen Sie zuverlässig Ihre Compliance-Anforderungen. Sollte in Ihrer DL-Umgebung ein Problem entstehen, können Sie darauf vertrauen, dass unser bewährtes Support-Modell Sie bei der Fehlerbehebung unterstützt und berät.

## Vereinheitlichte KI-Workloads

Sie haben in Ihrem Unternehmen jetzt die Möglichkeit, Infrastruktursilos zu beseitigen, die ungenutzt sind oder Engpässe bei KI-Workloads verursachen. Mit ONTAP AI erhalten Sie ausgehend von DGX-A100-Systemen eine universelle KI-Infrastrukturlösung, mit der Analysen, Training und Inferenz auf einer Plattform konsolidiert werden. Die Plattform kann flexibel auf Ihre Geschäftsanforderungen reagieren, und die Gesamtbetriebskosten sind niedriger als bei älteren Architekturen.

## NetApp und NVIDIA: Gemeinsam Innovation fördern

Herzstück von ONTAP AI ist das DGX-A100-System als universeller Baustein für Datacenter-KI, der Training, Inferenz, Data Science und andere High-Performance-Workloads unterstützt. Jedes DGX-A100-System verfügt über acht NVIDIA A100 Tensor Core GPUs und zwei AMD EPYC Prozessoren der 2. Generation. Die neuesten ultraschnellen NVIDIA-Mellanox-ConnectX-6-Interconnects mit Kompatibilität zu 100/200-Gb-Ethernet und InfiniBand sind ebenfalls in jedes System integriert.

Die Partitionierung des DGX-A100-Systems in bis zu 56 Instanzen pro System erlaubt dank neuester NVIDIA GPU-Multi-Instanztechnologie (MIG) mehrere kleinere Workloads schneller zu verarbeiten. Durch diese Beschleunigung ist die GPU-Performance in ONTAP AI effizient zuweisbar. Ihre Data Science-Teams können so überall in Ihrem Unternehmen schnellere Iterationen durchführen, die Reproduzierbarkeit automatisieren und KI-Projekte bei höherer Qualität bis zu drei Monate früher abschließen.

NetApp All Flash FAS Systeme setzen den schnellsten und flexibelsten All-Flash-Storage der Branche mit den weltweit ersten End-to-End-NVMe-Technologien ein, um den Datenfluss zu ML- und DL-Prozessen aufrechtzuerhalten. Das AFF A800 System kann Daten bis zu viermal schneller in DGX-Systeme einspeisen als Lösungen von Mitbewerbern.[1]

Die in die ONTAP AI Lösung integrierten Mellanox Spectrum Ethernet-Switches bieten die in KI-Umgebungen geforderte niedrige Latenz, hohe Dichte, hohe Performance und Energieeffizienz.

1. Lesedurchsatz von bis zu 300 GB/s pro All-Flash-Cluster im Vergleich zu 75 GB/s bei einem führenden Mitbewerber.

Eine Data Fabric auf der Basis von NetApp ermöglicht Datenmanagement und Cloud-Integration auf höchstem Niveau und trägt damit zur Beschleunigung von DL bei gleichzeitig hohem Schutz Ihrer kritischen Daten bei. ONTAP bietet eine beispiellose Datenreduktionsrate von insgesamt 22:1 und bis zu 54 % niedrigere Gesamtbetriebskosten als Direct-Attached Storage.

Das DGX-A100-System basiert auf dem NVIDIA DGX Software-Stack mit optimierter Software für KI- und Data-Science-Workloads. Durch die deutlich höhere Performance machen sich Ihre Investitionen in die KI-Infrastruktur schneller bezahlt.

Die NetApp AI Control Plane trägt durch Integration von Kubernetes und Kubeflow in die NetApp Data Fabric zu einem einfacheren KI-Datenmanagement bei. Diese integrierte Lösung gibt Ihnen optimale Datenverfügbarkeit und -portabilität zwischen Edge, Core und Cloud. Ergänzt wird die AI Control Plane durch das NetApp DataOps Toolkit – eine Python-Bibliothek, die Ihren Data Scientists und Data Engineers zahlreiche Datenmanagement-Aufgaben erleichtert. Das Provisionieren eines neuen Daten-Volumes ist ebenso möglich wie beispielsweise das blitzschnelle Klonen von Daten-Volumes oder das Erstellen von NetApp Snapshot Kopien zur Nachverfolgung und Bestimmung von Baselines.

Erfolg ist nur mit den richtigen Tools möglich. Deshalb ist ONTAP AI mit führender MLOps-Software (Machine Learning Operations) wie Domino Data Lab und Iguazio validiert. Ihre Teams können auf die vertrauten Tools zurückgreifen, sodass Ihre KI-Umgebung optimal genutzt wird und schneller verwertbare Erkenntnisse liefert.

**Lösungskomponenten**

- NVIDIA DGX-A100-Systeme
- NetApp Storage-Systeme der AFF A-Serie mit ONTAP 9
- NVIDIA Mellanox Spectrum SN3700C, NVIDIA Mellanox Quantum QM8700 und/oder NVIDIA Mellanox Spectrum SN3700-V
- NVIDIA DGX-Software-Stack
- NetApp AI Control Plane
- NetApp DataOps Toolkit

**Referenzarchitekturen**

NetApp hat die folgenden Referenzarchitekturen auf Basis von ONTAP AI veröffentlicht, die in verschiedenen Anwendungsfällen in bestimmten Branchen eingesetzt werden können:

- ONTAP AI Referenzarchitektur für das Gesundheitswesen: Diagnostische Bildgebung
- ONTAP AI Referenzarchitektur für Workloads für das autonome Fahren: Lösungsdesign
- ONTAP AI Referenzarchitektur für Financial-Services-Workloads: Lösungsdesign

---

**Über NetApp**

In einer Welt voller Generalisten beweist sich NetApp als Spezialist. Wir haben ein Ziel fest im Blick: Ihr Unternehmen darin zu unterstützen, Ihre Daten optimal zu nutzen. NetApp bringt die Datenservices, denen Sie vertrauen, in die Cloud und die Einfachheit und Flexibilität der Cloud in Ihr Datacenter. Selbst bei höchsten Ansprüchen lassen sich die branchenführenden NetApp Lösungen in unterschiedlichsten Kundenumgebungen und den weltweit führenden Public Clouds einsetzen.

Als Cloud- und Daten-orientierter Softwareanbieter stellt nur NetApp alle Technologien bereit, mit denen Sie Ihre eigene maßgeschneiderte Data Fabric aufbauen, Ihre Clouds vereinfachen, Ihre Public Clouds anbinden und so die richtigen Daten, Services und Applikationen sicher bereitstellen können – immer und überall. www.netapp.de

**Über NVIDIA**

Als NVIDIA 1999 die GPU, d. h. den separaten Grafikprozessor, erfand, löste das Unternehmen den Boom des Marktes für PC-Spiele aus, setzte neue Maßstäbe für moderne Computergrafik und revolutionierte die parallele Datenverarbeitung. GPUs mit Deep Learning legten in jüngerer Zeit den Grundstein für die moderne KI und damit für das nächste Zeitalter der Datenverarbeitung. Die GPU agiert als das Gehirn von Computern, Robotern und autonomen Fahrzeugen, die dazu in der Lage sind, die Welt wahrzunehmen und zu verstehen. Weitere Informationen hierzu finden Sie unter www.nvidia.de.

NetApp