NVIDIA DGX-2
DAS LEISTUNGSFÄHIGSTE DEEP-LEARNING-SYSTEM DER WELT FÜR DIE KOMPLEXESTEN KI-HERAUSFORDERUNGEN

Skalierung für modernes KI und Deep Learning – eine echte Herausforderung

Deep Neural Networks werden innerhalb sehr kurzer Zeit immer umfangreicher und komplexer, damit sie die wichtigsten Herausforderungen in Wirtschaft und Forschung meistern können. Mit der für moderne KI-Workloads erforderlichen Rechenkapazität können traditionelle Rechenzentrumsarchitekturen nicht mehr mithalten. Moderne Methoden, in denen zunehmend parallele Modelle genutzt werden, stoßen an die Grenzen, die ihnen durch die Bandbreite der Verbindungen zwischen GPUs auferlegt sind. Denn Entwickler richten immer größere Cluster für beschleunigtes Computing ein und definieren damit die Grenzen der Kapazität von Rechenzentren neu. Hier ist ein neuer Ansatz erforderlich – einer, der nahezu uneingeschränkte Skalierung für KI-Computing ermöglicht. Dadurch lassen sich die bisherigen Einschränkungen überwinden und schneller die Erkenntnisse gewinnen, die unsere Welt verändern können.

Leistung für das Training bisher untrainierbarer Modelle

KI wird immer komplexer und erfordert daher mehr Rechenleistung als je zuvor. NVIDIA® DGX-2™ ist das weltweit erste System mit 2 PetaFLOPS, das die Leistung von 16 der fortschrittlichsten Grafikprozessoren der Welt in sich vereint. Es beschleunigt die neuesten Arten von Deep-Learning-Modellen, die bisher nicht trainierbar waren. Dank der bahnbrechenden Grafikprozessorskalierung lassen sich nun viermal größere Modelle auf einem einzelnen Knoten trainieren. Im Vergleich zu älteren x86-Architekturen würde die Fähigkeit der DGX-2, ResNet-50 zu trainieren, das Äquivalent von 300 Servern mit zwei Intel Xeon Gold CPUs erfordern, die über 2,7 Millionen Dollar kosten würden.

NVIDIA NVSwitch – revolutionäres KI-Netzwerk-Fabric

Für bahnbrechende Forschung sind hohe Flexibilität zur Nutzung paralleler Modelle sowie Bandbreite zwischen den Grafikprozessoren auf einem völlig neuen Niveau erforderlich. NVSwitch ist die Antwort von NVIDIA auf diese Herausforderung. NVSwitch bietet schon heute das Netzwerk-Fabric der Zukunft – vergleichbar mit der Entwicklung von Dial-Up-Verbindungen bis zum Ultra-High-Speed-Breitband. Mit NVIDIA DGX-2 sind die Komplexität und Größe von Modellen nicht mehr durch die Grenzen herkömmlicher Architekturen eingeschränkt. Mit einem Netzwerk-Fabric in DGX-2, das 2,4 TB/s Bisektionsbandbreite für eine 24-fache Steigerung gegenüber den Vorgängergenerationen bietet, können Sie mit dieser Lösung ohne Kompromisse auf Training mit parallelen Modellen setzen. Diese neue Datenschnellspur bietet unendlich viele Möglichkeiten für Modelltypen, die von auf 16 Grafikprozessoren gleichzeitig verteiltem Training profitieren.

SYSTEMSPEZIFIKATIONEN

Grafikprozessoren	**16× NVIDIA® Tesla V100**
Grafikprozessorspeicher	**Insgesamt 512 GB**
Leistung	**2 PetaFLOPS**
NVIDIA CUDA®-Recheneinheiten	**81.920**
NVIDIA Tensor-Recheneinheiten	**10.240**
NVSwitches	**12**
Max. Leistungsaufnahme	**10 kW**
CPU	**Dual Intel Xeon Platinum 8168, 2,7 GHz, 24 Kerne**
Arbeitsspeicher	**1,5 TB**
Netzwerk	**8× 100 Gb/s Infiniband/100 GigE Dual 10/25 Gb/s Ethernet**
Datenspeicher	**Betriebssystem: 2× 960 GB NVME-SSDs. Interner Speicher: 30 TB (8× 3,84 TB) NVME SSDs**
Software	**Ubuntu Linux Einzelheiten, siehe Zusatzsoftware**
Systemgewicht	**154,2 kg**
Systemabmessungen	**Höhe: 440,0 mm Breite: 482,3 mm Länge: 795,4 mm – Ohne Rahmen an Vorderseite 834,0 mm – Mit Rahmen an Vorderseite**
Betriebstemperatur	**5 °C bis 35 °C**

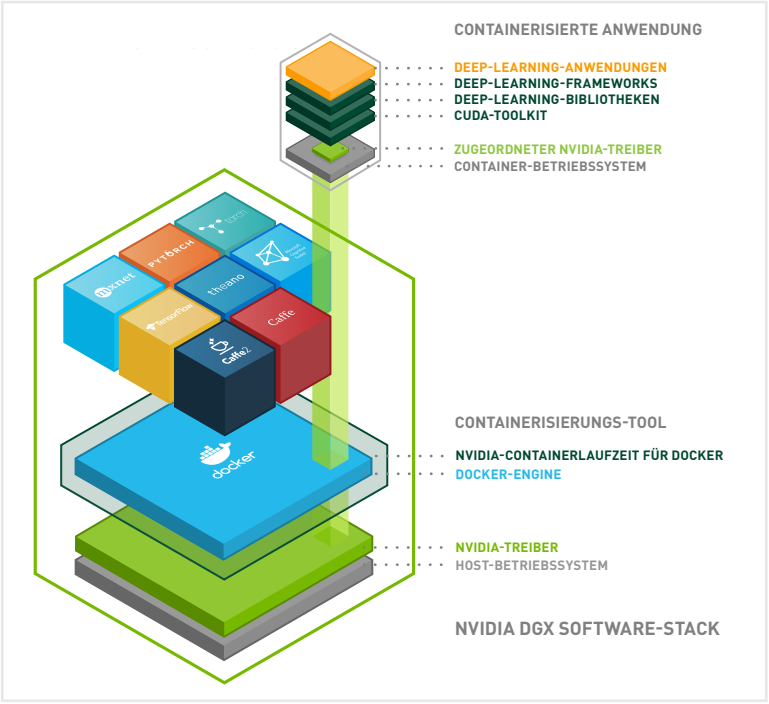
KI-Skalierung auf völlig neuem Niveau

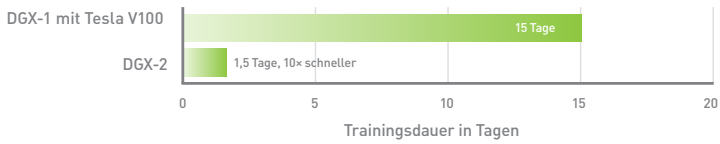
Moderne Unternehmen müssen KI-Leistung gemäß ihren geschäftlichen Vorgaben schnell bereitstellen, also KI ohne gleichzeitige Kosten- oder Komplexitätssteigerungen hochskalieren. Daher haben wir DGX-2 mit der DGX-Software entwickelt, die schnellere Bereitstellung und vereinfachte Operationen im benötigten Umfang ermöglicht. DGX-2 ist eine sofort einsatzbereite Lösung und die schnellste Option zum Hochskalieren von KI. Sie bietet Unterstützung für Virtualisierung und ermöglicht die Einrichtung einer privaten KI-Cloud für Unternehmensanforderungen. Unternehmen profitieren nun von uneingeschränkter KI-Leistung in einer Lösung, die sich mühelos mit nur einem Bruchteil der Netzwerkinfrastruktur skalieren lässt, der sonst zum Kombinieren der Ressourcen für beschleunigtes Computing erforderlich ist. Dank des beschleunigten Bereitstellungsmodells und der speziell auf einfache Skalierbarkeit ausgelegten Architektur kann sich Ihr Team verstärkt darauf konzentrieren, wichtige Erkenntnisse zu gewinnen, und muss weniger Zeit für den Aufbau der Infrastruktur aufwenden.

KI-Infrastruktur für Unternehmensanforderungen

Wenn Ihre KI-Plattform geschäftskritisch für Ihr Unternehmen ist, muss sie die RAS-Kriterien erfüllen: hohe Zuverlässigkeit, Verfügbarkeit und Wartungsfreundlichkeit. DGX-2 ist eine Unternehmenslösung zur kompromisslosen Ausführung von KI-Operationen rund um die Uhr. Sie ist speziell auf die RAS-Kriterien ausgelegt und zeichnet sich daher durch niedrige ungeplante Ausfallzeiten, optimierte Wartungsoptionen und kontinuierlichen Betrieb aus.

Mit ihr verwenden Sie weniger Zeit auf Feineinstellung und Optimierung und mehr Zeit auf Entdeckungen. Der Support der Enterprise-Klasse von NVIDIA erspart Ihnen die zeitaufwendige Aufgabe der Fehlerbehebung bei Hardware und Open-Source-Software. Mit jedem DGX-System, einer integrierten Lösung mit Software, Tools und NVIDIA-Expertise, profitieren Sie bei der Einführung und beim Trainieren von beeindruckender Schnelligkeit, die nachhaltig bestehen bleibt.





Workload: FairSeq, 55 Epochen bis zur Lösung PyTorch-Trainingsleistung.

Weitere Informationen finden Sie unter **www.nvidia.de/dgx-2**

NVIDIA DGX-2
DAS LEISTUNGSFÄHIGSTE DEEP-LEARNING-SYSTEM DER WELT FÜR DIE KOMPLEXESTEN KI-HERAUSFORDERUNGEN

Skalierung für modernes KI und Deep Learning – eine echte Herausforderung

Deep Neural Networks werden innerhalb sehr kurzer Zeit immer umfangreicher und komplexer, damit sie die wichtigsten Herausforderungen in Wirtschaft und Forschung meistern können. Mit der für moderne KI-Workloads erforderlichen Rechenkapazität können traditionelle Rechenzentrumsarchitekturen nicht mehr mithalten. Moderne Methoden, in denen zunehmend parallele Modelle genutzt werden, stoßen an die Grenzen, die ihnen durch die Bandbreite der Verbindungen zwischen GPUs auferlegt sind. Denn Entwickler richten immer größere Cluster für beschleunigtes Computing ein und definieren damit die Grenzen der Kapazität von Rechenzentren neu. Hier ist ein neuer Ansatz erforderlich – einer, der nahezu uneingeschränkte Skalierung für KI-Computing ermöglicht. Dadurch lassen sich die bisherigen Einschränkungen überwinden und schneller die Erkenntnisse gewinnen, die unsere Welt verändern können.

Leistung für das Training bisher untrainierbarer Modelle

KI wird immer komplexer und erfordert daher mehr Rechenleistung als je zuvor. NVIDIA® DGX-2™ ist das weltweit erste System mit 2 PetaFLOPS, das die Leistung von 16 der fortschrittlichsten Grafikprozessoren der Welt in sich vereint. Es beschleunigt die neuesten Arten von Deep-Learning-Modellen, die bisher nicht trainierbar waren. Dank der bahnbrechenden Grafikprozessorskalierung lassen sich nun viermal größere Modelle auf einem einzelnen Knoten trainieren. Im Vergleich zu älteren x86-Architekturen würde die Fähigkeit der DGX-2, ResNet-50 zu trainieren, das Äquivalent von 300 Servern mit zwei Intel Xeon Gold CPUs erfordern, die über 2,7 Millionen Dollar kosten würden.

NVIDIA NVSwitch – revolutionäres KI-Netzwerk-Fabric

Für bahnbrechende Forschung sind hohe Flexibilität zur Nutzung paralleler Modelle sowie Bandbreite zwischen den Grafikprozessoren auf einem völlig neuen Niveau erforderlich. NVSwitch ist die Antwort von NVIDIA auf diese Herausforderung. NVSwitch bietet schon heute das Netzwerk-Fabric der Zukunft – vergleichbar mit der Entwicklung von Dial-Up-Verbindungen bis zum Ultra-High-Speed-Breitband. Mit NVIDIA DGX-2 sind die Komplexität und Größe von Modellen nicht mehr durch die Grenzen herkömmlicher Architekturen eingeschränkt. Mit einem Netzwerk-Fabric in DGX-2, das 2,4 TB/s Bisektionsbandbreite für eine 24-fache Steigerung gegenüber den Vorgängergenerationen bietet, können Sie mit dieser Lösung ohne Kompromisse auf Training mit parallelen Modellen setzen. Diese neue Datenschnellspur bietet unendlich viele Möglichkeiten für Modelltypen, die von auf 16 Grafikprozessoren gleichzeitig verteiltem Training profitieren.

SYSTEMSPEZIFIKATIONEN

Grafikprozessoren	**16× NVIDIA® Tesla V100**
Grafikprozessorspeicher	**Insgesamt 512 GB**
Leistung	**2 PetaFLOPS**
NVIDIA CUDA®-Recheneinheiten	**81.920**
NVIDIA Tensor-Recheneinheiten	**10.240**
NVSwitches	**12**
Max. Leistungsaufnahme	**10 kW**
CPU	**Dual Intel Xeon Platinum 8168, 2,7 GHz, 24 Kerne**
Arbeitsspeicher	**1,5 TB**
Netzwerk	**8× 100 Gb/s Infiniband/100 GigE Dual 10/25 Gb/s Ethernet**
Datenspeicher	**Betriebssystem: 2× 960 GB NVME-SSDs. Interner Speicher: 30 TB (8× 3,84 TB) NVME SSDs**
Software	**Ubuntu Linux Einzelheiten, siehe Zusatzsoftware**
Systemgewicht	**154,2 kg**
Systemabmessungen	**Höhe: 440,0 mm Breite: 482,3 mm Länge: 795,4 mm – Ohne Rahmen an Vorderseite 834,0 mm – Mit Rahmen an Vorderseite**
Betriebstemperatur	**5 °C bis 35 °C**

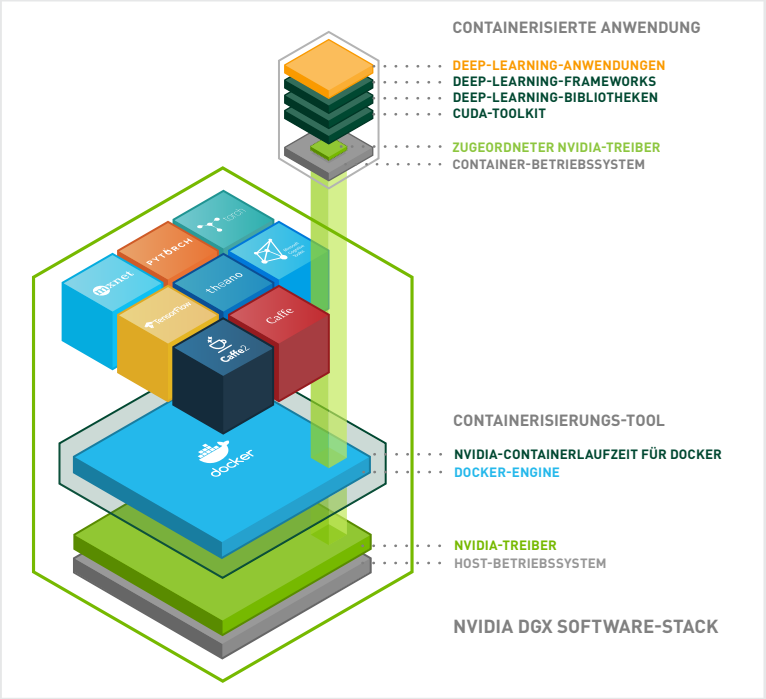
KI-Skalierung auf völlig neuem Niveau

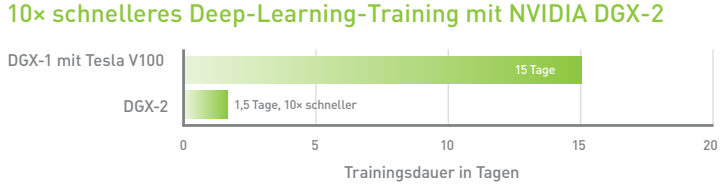
Moderne Unternehmen müssen KI-Leistung gemäß ihren geschäftlichen Vorgaben schnell bereitstellen, also KI ohne gleichzeitige Kosten- oder Komplexitätssteigerungen hochskalieren. Daher haben wir DGX-2 mit der DGX-Software entwickelt, die schnellere Bereitstellung und vereinfachte Operationen im benötigten Umfang ermöglicht. DGX-2 ist eine sofort einsatzbereite Lösung und die schnellste Option zum Hochskalieren von KI. Sie bietet Unterstützung für Virtualisierung und ermöglicht die Einrichtung einer privaten KI-Cloud für Unternehmensanforderungen. Unternehmen profitieren nun von uneingeschränkter KI-Leistung in einer Lösung, die sich mühelos mit nur einem Bruchteil der Netzwerkinfrastruktur skalieren lässt, der sonst zum Kombinieren der Ressourcen für beschleunigtes Computing erforderlich ist. Dank des beschleunigten Bereitstellungsmodells und der speziell auf einfache Skalierbarkeit ausgelegten Architektur kann sich Ihr Team verstärkt darauf konzentrieren, wichtige Erkenntnisse zu gewinnen, und muss weniger Zeit für den Aufbau der Infrastruktur aufwenden.

KI-Infrastruktur für Unternehmensanforderungen

Wenn Ihre KI-Plattform geschäftskritisch für Ihr Unternehmen ist, muss sie die RAS-Kriterien erfüllen: hohe Zuverlässigkeit, Verfügbarkeit und Wartungsfreundlichkeit. DGX-2 ist eine Unternehmenslösung zur kompromisslosen Ausführung von KI-Operationen rund um die Uhr. Sie ist speziell auf die RAS-Kriterien ausgelegt und zeichnet sich daher durch niedrige ungeplante Ausfallzeiten, optimierte Wartungsoptionen und kontinuierlichen Betrieb aus.

Mit ihr verwenden Sie weniger Zeit auf Feineinstellung und Optimierung und mehr Zeit auf Entdeckungen. Der Support der Enterprise-Klasse von NVIDIA erspart Ihnen die zeitaufwendige Aufgabe der Fehlerbehebung bei Hardware und Open-Source-Software. Mit jedem DGX-System, einer integrierten Lösung mit Software, Tools und NVIDIA-Expertise, profitieren Sie bei der Einführung und beim Trainieren von beeindruckender Schnelligkeit, die nachhaltig bestehen bleibt.





Workload: FairSeq, 55 Epochen bis zur Lösung PyTorch-Trainingsleistung.

Weitere Informationen finden Sie unter **www.nvidia.de/dgx-2**