DELL Technologies

Dell Precision Data Science Workstation: Benchmarks und Best Practice für KI, R2.0

Autoren: Steven Starrett, Raed Hijer, Brandon Kranz, Kyle Harper

NVIDIA RTX

Wir haben die Version 2.0 dieses Dokuments veröffentlicht, die Ergebnisse des Canonical-Upgrades von Ubuntu Linux von 18.04 auf 20.04, des NVIDIA-Upgrades der NVIDIA Data Science Software von 2.4.0 auf 2.8.0 sowie der hinzugefügten NVIDIA RTX A6000-GPU enthält. Die Beschreibung, das Setup und die Ergebnisse für Ubuntu 20.04, die NV Data Science Software 2.8.0 und RTX A6000 finden Sie in Anhang 1.

In den meisten Unternehmen ist die Verarbeitung von großen Datenmengen – die dank der künstlichen Intelligenz (KI) möglich ist – zu einem wichtigen Bestandteil der Geschäftsmodelle geworden. Daher wird es immer wichtiger, die richtigen Tools, Technologien und Techniken zu finden, um diese Aufgabe effektiv zu erledigen. Um zu veranschaulichen, wie die Auswahl der richtigen Komponenten die Effektivität von KI-Aufgaben maximieren kann, hat Dell verschiedene Konfigurationen der Dell Precision Data Science Workstation (DSW) mit Workflows für Deep Learning (DL) und maschinelles Lernen (ML) verglichen.

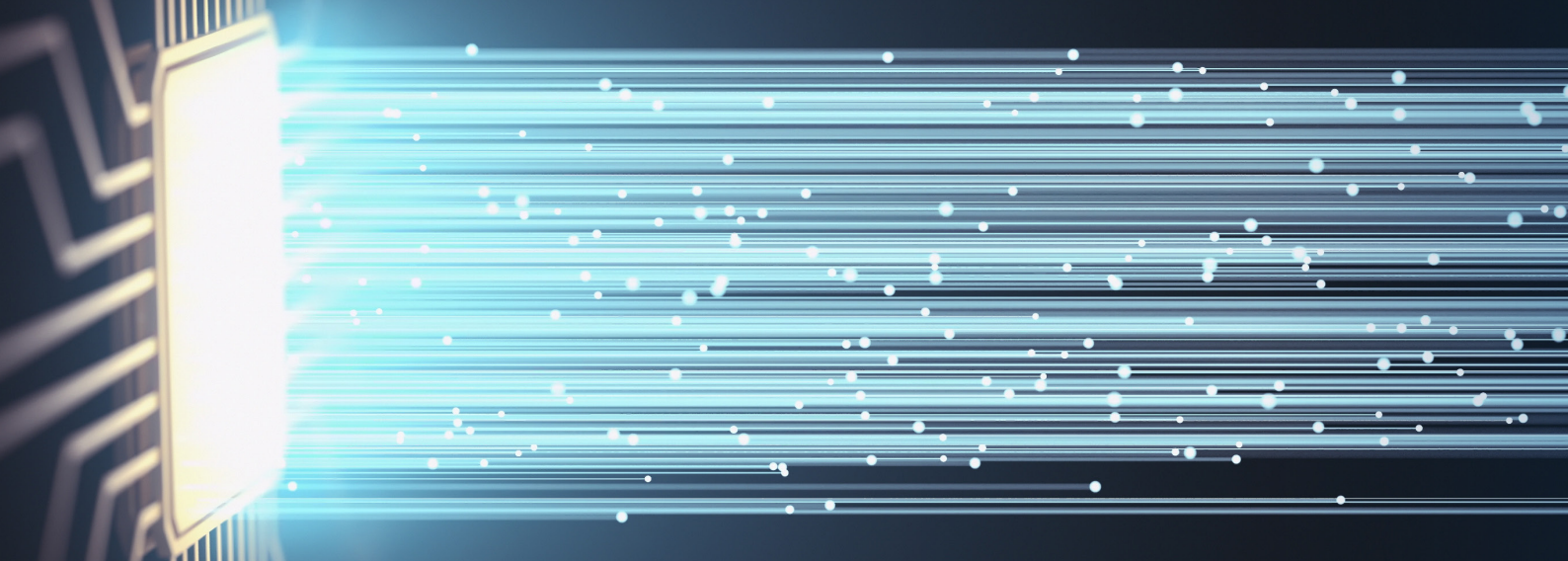
Die Dell Analyse zeigt eindeutig die substanziellen Vorteile der GPU-Beschleunigung auf. Sie bindet zudem sämtliche Originaldaten ein, sodass die Ergebnisse auch von Dritten getestet und validiert werden können. Eine Benchmark aus dem ML-Modelltraining zeigt, dass die Ausführung auf einer CPU 6,4-mal länger dauert als mit einer GPU-Konfiguration. Eine andere DL-Benchmark hingegen ist aufgrund der Multi-GPU-Beschleunigung bis zu 4,74-mal schneller.

In der zweiten Hälfte dieses Whitepaper wird anhand dieser Ergebnisse ein einfacher Best-Practice-Leitfaden zur Auswahl der optimalen Komponenten für jeden Workflow bereitgestellt.

Einführung: Die KI-Landschaft

Die Geschäftswelt hat sich in den letzten zehn Jahren rasant verändert, da das Internet der Dinge und sein industrielles Pendant enorme Datenmengen produzieren. IDC-Schätzungen zufolge werden die Daten bis 2025 weltweit 175 Zettabyte erreichen. Das ist eine 10-fache Steigerung gegenüber 2017.

Daraus sind höhere Anforderungen bei der Verarbeitungsgeschwindigkeit entstanden, die mit den gleichzeitigen Weiterentwicklungen bei Algorithmen, Open-Source-Software und speziellen Hardwarebeschleunigern zu einer explosionsähnlichen Nutzung der künstlichen Intelligenz (KI) geführt haben.



Im großen Bereich der KI gelten die KI-Untergruppen des maschinellen Lernens (ML) und des Deep Learning (DL) als die modernsten datengestützten Ansätze, um zahlreiche der heutzutage bestehenden komplexen Probleme zu lösen.

Diese Techniken erschließen den Wert, der in den riesigen Datenmengen steckt, die aus völlig unterschiedlichen Quellen (wie z. B. Kundenausgaben, YouTube-Videos und Maschinen im Werk) einfließen. Das ist jedoch nicht einfach und erfordert Technologie, Kompetenz und das richtige Equipment, damit sich aus diesem Rohmaterial verwertbare Erkenntnisse gewinnen lassen.

Maschinelles Lernen und Deep Learning haben jeweils ihre ganz eigenen Anwendungsfälle und Herausforderungen. Maschinelles Lernen ist weniger ausgefeilt und wird in der Regel bei strukturierten Daten verwendet, wie z. B. Tabellendaten. Die Verarbeitung erfolgt mithilfe von bekannten Algorithmen wie lineare oder logistische Regression, Naive Bayes und XGBoost.

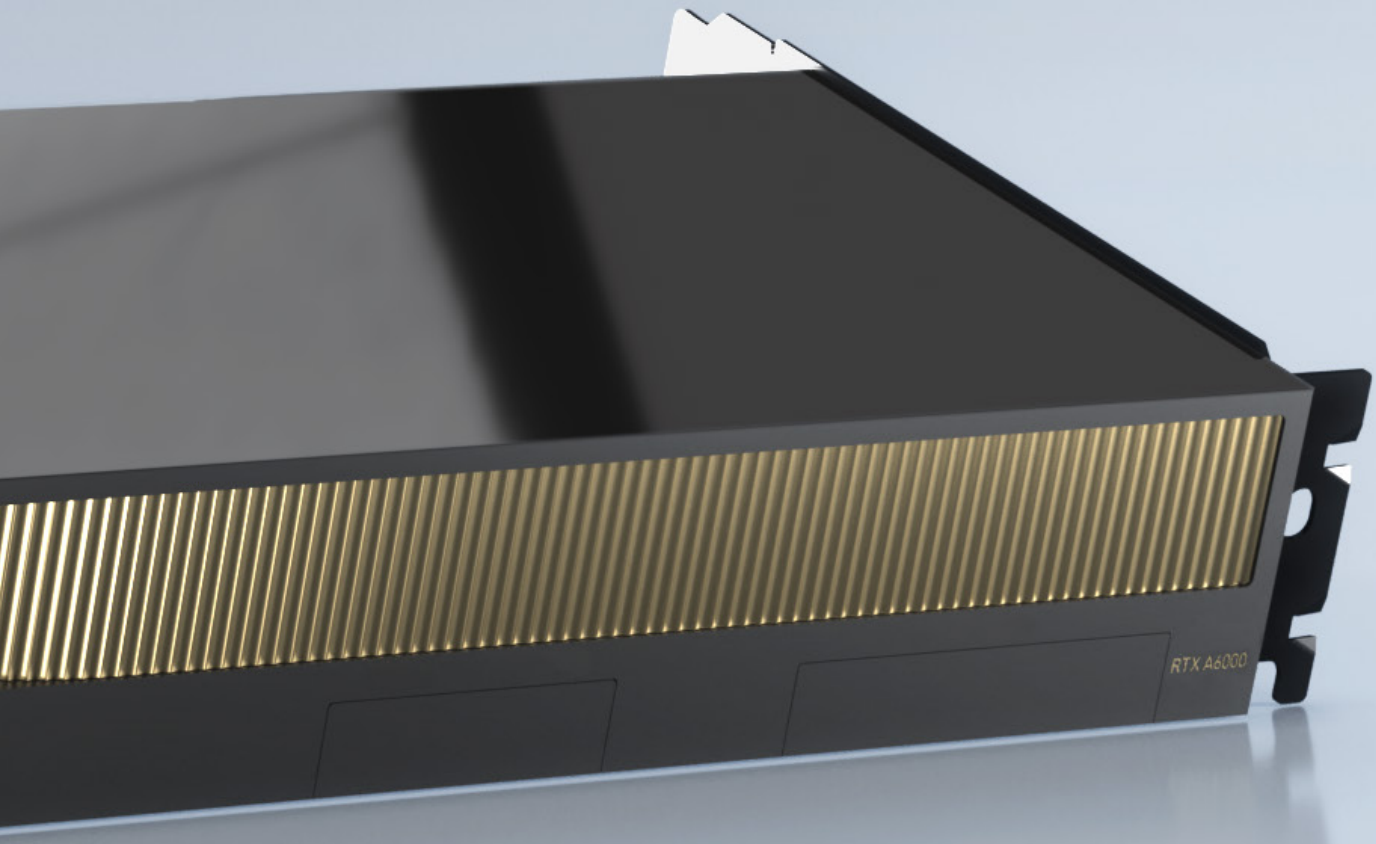
Diese können je nach Anwendungsfall nur auf reinen CPU- oder auf GPU-beschleunigten Compute-Plattformen ausgeführt werden.

Deep Learning hingegen ist eine der effektivsten Techniken im KI-Instrumentarium und eignet sich besonders gut für unstrukturierte Daten (wie z. B. Bilder, Videos oder Sprache). Es ist ausgesprochen ausgereift und basiert auf Ansätzen für künstliche neuronale Netze, die von der Struktur und der Aktivität von Neuronen im menschlichen Gehirn inspiriert sind.

Komplexität und die Rolle von DatenwissenschaftlerInnen

Aufgrund der Zunahme dieser KI-Techniken und ihrer unterschiedlichen Herausforderungen haben viele Unternehmen DatenwissenschaftlerInnen eingestellt. Sie haben die sehr komplexe Aufgabe, die Daten mit dem Ziel, deren verborgenen Wert zu erschließen, einzuspeisen und zu kuratieren.

Dies ist ein multidisziplinäres Feld, das eine Vielzahl von Bereichen abdeckt, darunter Analysen und maschinelles Lernen. Folglich wird dazu eine leistungsstarke, validierte und dennoch flexible Hardware in Kombination mit einsatzbereiten Softwareumgebungen und Tools benötigt. Das ist allerdings einfacher gesagt als getan und bringt zwei wichtige Komplexitätsherausforderungen mit sich.



Herausforderung 1: die Integration

Die manuelle Installation und Integration von Hardware- und Softwarekomponenten ist umständlich, mühsam und zeitaufwendig. Einerseits führen Hardware-beschleuniger (wie z. B. NVIDIA RTX-GPUs) mit jeder Generation neue spezialisierte Verbesserungen ein. Andererseits werden Softwaretools ständig für die neuesten Algorithmen, Frameworks und Bibliotheken optimiert. All diese neuen Elemente zusammenzuführen und die Kompatibilität zwischen sämtlichen Variablen zu gewährleisten, ist schwierig – und kann schnell dazu führen, dass DatenwissenschaftlerInnen als SystemadministratorInnen arbeiten und ihre Fähigkeiten nicht richtig nutzen.

Herausforderung 2: die Konfiguration

Die Entscheidung, welche GPU eingesetzt werden soll, hängt wiederum von der verwendeten Workload ab. So können sich z. B. die Computer-Vision-Anforderungen von denen der NLP-Aufgaben (Natural Language Processing, Verarbeitung natürlicher Sprache) unterscheiden, die selbst schon andere Anforderungen haben als das klassische maschinelle Lernen. Verschiedene Dataset-Typen stellen eben auch unterschiedliche Anforderungen an die Compute-Plattform. All das macht es schwierig, die am besten geeignete Plattform und die GPU für die verschiedenen Anwendungsfälle auszuwählen.

Unsere Benchmarkingprozesse

In diesem Whitepaper werden drei Anwendungsbeispiele untersucht: Deep Learning – auf zwei Instanzen – und maschinelles Lernen. Beim ersten Deep-Learning-Beispiel wird die Computer-Vision-Bildklassifizierung analysiert, beim zweiten kommt noch BERT Finetuning hinzu. Beim Beispiel für maschinelles Lernen wird XGBoost (eXtreme Gradient Boosting) in einem klassischen Modell verwendet.

In allen Fällen werden mehrere Dell Precision Data Science Workstations mit verschiedenen GPU-Konfigurationen eingesetzt, mit Benchmarks versehen und präsentiert, um Hinweise zu den erwarteten Performancegeschwindigkeiten für die Modellentwicklung zu geben. Die Details dazu werden in den entsprechenden Abschnitten weiter unten behandelt.

Neue Technologien für moderne Herausforderungen

Die Dell Precision Data Science Workstation (DSW) ist eine speziell entwickelte Produktlinie, die die Herausforderung der Integration bewältigt. Sie kuratiert die neueste Hardware mit NVIDIA-GPUs sowie dem NVIDIA Data Science Software-Stack. Dieser bietet die neuesten GPU-beschleunigten Frameworks und Bibliotheken, die im Dell Werk validiert wurden.

Damit können DatenwissenschaftlerInnen nun den KI-Pfad einschlagen und müssen sich keine Gedanken mehr darüber machen, wie viele unzählige Stunden oder Tage sie damit verbringen müssen, herauszufinden, welche Hardware und Software zusammenarbeiten. Die Herausforderung der Konfiguration liegt jedoch weiterhin bei den IT-Abteilungen, den DatenwissenschaftlerInnen und deren Unternehmen. Dell hat diese Benchmark zur Lösung des Problems zusammengestellt.

Konfiguration einer Dell Precision Data Science W

Für die korrekte Dimensionierung und Konfiguration einer Dell Data Science Workstation auf Basis der Anforderungen von DatenwissenschaftlerInnen sind drei allgemeine Schritte erforderlich. Hierbei handelt es sich um (1) die Bestimmung der Datasetgröße und des besten KI-Modells, (2) die Auswahl der passenden GPU und des GPU-Speichers für das Dataset und (3) die Auswahl der richtigen CPU und der CPU-Speicherkonfiguration.

SCHRITT 1

Hier geht es um die Dimensionierung des Datasets und die Auswahl des Entwicklungsansatzes für das KI-Modell. In dieser Branche gibt es keine Universallösung, da die Ansätze und Modelle sehr unterschiedlich sind.

Meist besteht das Dilemma in einem der folgenden beiden Punkte:

- Der Arbeitsspeicher des Systems ist zu klein, sodass die Ergebnisse verzerrt oder beeinträchtigt sind.
- Der Arbeitsspeicher des Systems ist zu groß und teuer oder kann nicht physisch in einem System untergebracht werden.

Das Dataset kann in kleinere Segmente unterteilt werden – sogenannte Mini-Batches –, um in den verfügbaren Speicher des Systems zu passen.

SCHRITT 2

Hier geht es darum, die beste GPU und die GPU-Speichergröße für das Dataset zu bestimmen. Wenn sich die GPU-Auslastung 100 % nähert (das wird im Befehlszeilendienstprogramm „nvidia-smi“ angezeigt), ist dies ein Hinweis darauf, dass eine größere Speicher-GPU oder eine Konfiguration mit zwei oder drei GPUs erforderlich ist.

SCHRITT 3

Hier geht es um die Auswahl der optimalen CPU und der CPU-Speicherkonfiguration auf der CPU-Seite, um die GPU-Seite des Systems versorgen zu können. Abschließend werden die Storage-Konfiguration und die -Dimensionierung ausgewählt.

Nehmen wir eine BERT (Bidirectional Encoder Representations) Benchmark als Beispiel. Im Rahmen der BERT Benchmarks haben wir festgestellt, dass die RTX 6000-Konfigurationen bei geringeren Sequenzlängen und Batchgrößen die RTX 8000-Konfigurationen leistungsmäßig übertreffen.

Sobald die Sequenzlänge und die Batchgröße jedoch einen bestimmten Schwellenwert übersteigen, hat die RTX 6000-Karte zu wenig Speicher, um die Aufgabe auszuführen.

Workstation – zu berücksichtigende Faktoren

Das Dataset

Je nachdem, wie groß das Dataset sein soll, passt es möglicherweise vollständig in den Speicher, der zum Training des Modells verwendet wird. In einem reinen CPU-System wird das Dataset im DDR-Speicher der CPU platziert und das Modell wird unter Verwendung der CPU(s) trainiert.

In einem System mit GPUs – wie der DSW – wird das Modell unter Nutzung der GPU-Ressourcen trainiert und das Dataset verbleibt im GPU-Speicher. Der Speicher moderner GPU-Karten ist für viele Modelle und Datasets geeignet.

In einigen Fällen, in denen die Datasetgröße den Speicher einer einzelnen GPU übersteigt, können Tricks wie der Einsatz von verteilten Multi-GPU-Bibliotheken (z. B. DASK) zum Laden des Datasets in die GPUs verwendet werden.

Auswahl und Dimensionierung der GPU zur optimalen Beschleunigung der KI-Modellierung

Hinweis: In der „Tabelle 2: Zum Benchmarking verwendete GPUs“ ist der Speicherpuffer für jede NVIDIA-GPU angegeben, die zum Zeitpunkt der Veröffentlichung verfügbar war.

Die RTX A6000 verfügt über 48 GB GDDR6-Speicher mit bis zu 768 Gbit/s, 10.752 CUDA-Cores der Ampere-Klasse, 84 RT-Cores der Ampere-Klasse und 336 Tensor-Cores der Ampere-Klasse. Die Kombination von zwei RTX A6000-GPUs über NVLink verdoppelt die Anzahl der Cores und den Speicher, sodass die beiden Karten gemeinsam 96 GB einheitlichen GPU-Speicher für das Dataset zur Verfügung stellen.

Die RTX 8000 verfügt über 48 GB GDDR6-Speicher mit einem Durchsatz von bis zu 672 Gbit/s, 4.608 CUDA-Cores, 72 RT-Cores und 576 Tensor-Cores. Die Kombination von zwei RTX 8000-GPUs über NVLink verdoppelt die Anzahl der Cores und den Speicher, sodass die beiden Karten gemeinsam 96 GB einheitlichen GPU-Speicher für das Dataset zur Verfügung stellen.

Die RTX 6000-GPU verfügt über dieselbe Anzahl von Cores wie die RTX 8000 sowie über 24 GB GDDR6-Grafikspeicher mit einem Durchsatz von bis zu 672 Gbit/s. Die RTX 6000 kann über NVLink mit einer zweiten RTX 6000 verbunden werden und so 48 GB an kombiniertem einheitlichen Speicher bereitstellen.

Die RTX 5000 verfügt über 3.072 CUDA-Kerne, 48 RT-Kerne, 384 Tensor-Cores und 16 GB GDDR6-Speicher mit einem Durchsatz von bis zu 448 Gbit/s.

Warum wurden für diese Benchmark verschiedene Setups und Workloads verwendet?

Dell hat Benchmarks ausgewählt und durchgeführt, um die Performanceunterschiede zwischen mehreren Dell Precision DSW-Plattformen und -Konfigurationen darzustellen. Die Plattformen reichen von Dell Mobile DSWs (7550 Mobile mit 15" und 7750 Mobile mit 17") über Dell Tower DSWs (Dell 5820 und 7920 Tower) bis zur Dell 7920 Rack DSW.

Für diese Plattformen haben wir verschiedene gängige CPU-Speicherkonfigurationen und GPU-Konfigurationen ausgewählt, die das Spektrum der Compute-Ressourcen abbilden, die für Workloads im KI-Modelltraining geeignet sind. An NVIDIA-GPUs wurden im Benchmarking die mobile Quadro RTX 5000-GPU in Mobile DSWs sowie die Quadro RTX 6000-GPU und die RTX 8000-GPU in den Tower und Rack DSWs eingesetzt.

All diese Plattformen und Konfigurationen wurden mit identischen Ubuntu Linux-Betriebssystemen und dem NVIDIA Data Science Software-Stack geladen. Dieses Software-Bundle wird als Teil jeder DSW von Dell validiert und ausgeliefert.

Der NVIDIA Data Science Software-Stack umfasst die werkseitig optimierte TensorFlow-, Python-, XGBoost- und Jupyter Notebook-Software, die alle NutzerInnen erhalten, und wurde zusammen mit der zugrunde liegenden Hardware bewertet. Ein Vergleich der verschiedenen Plattformen, Konfigurationen und GPU-Details finden Sie in den Tabellen auf dieser Seite:

Modell	Dell Precision 7550 Mobile	Dell Precision 7750 Mobile	Dell Precision 5820 Tower	Dell Precision 5820 Tower	Dell Precision 7920 Tower	Dell Precision 7920 Rack
Prozessor	Intel® Xeon W-10885M (8 Cores, 16 MB Cache, 2,40 GHz bis 5,30 GHz, 45 W, vPro)	Intel® Xeon W-10885M (8 Cores, 16 MB Cache, 2,40 GHz bis 5,30 GHz, 45 W, vPro)	Intel® Xeon Prozessor W-2175 (14 Cores, 19,25 MB, 2,5–4,3 GHz, 140 W, vPro)	Intel® Xeon Prozessor W-2245 (8 Cores, 11 MB, 3,9–4,7 GHz, 155 W, vPro)	Dual Intel® Xeon Prozessor Gold 6134 (8 Cores, 24,75 MB, 3,2–3,7 GHz, 130 W, vPro)	Dual Intel® Xeon Gold 6244 (8 Cores, 24,75 MB, 3,6–4,4 GHz, 150 W, vPro)
Grafikkarte	NVIDIA® Quadro RTX 5000 mit 16 GB GDDR6	NVIDIA® Quadro RTX 5000 mit 16 GB GDDR6	Single/Dual NVIDIA® Quadro RTX 6000 mit 24 GB GDDR6	Single NVIDIA® Quadro RTX 8000 mit 48 GB GDDR6	Dual/Triple NVIDIA® Quadro RTX 6000 mit 24 GB GDDR6	Dual/Triple NVIDIA® Quadro RTX 6000 mit 24 GB GDDR6
Speicher	64 GB, (4 x 16 GB) DDR4 2.933 Mhz Non-ECC	64 GB, (4 x 16 GB) DDR4 2.933 Mhz Non-ECC	128 GB (8 x 16 GB) DDR4 2.666 MHz RDIMM ECC	256 GB (4 x 64 GB) DDR4 2.933 MHz RDIMM ECC	128 GB (8 x 16 GB) DDR4 2.666 MHz RDIMM ECC	128 GB (8 x 16 GB) DDR4 2.666 MHz RDIMM ECC
Storage	2 x M.2-NVMe-PCIe-SSD, Klasse 40, 1 TB	2 x M.2-NVMe-PCIe-SSD, Klasse 50, 2 TB	1 x M.2-NVMe-PCIe-SSD, Klasse 40, 1 TB	1 x M.2-NVMe-PCIe-SSD, Klasse 50, 1 TB	1 x M.2-NVMe-PCIe-SSD, Klasse 40, 1 TB	1 x 2,5"-SATA-SSD, Klasse 20, 1 TB + 1 x 2,5"-SATA-SSD, Klasse 20, 512 GB
Betriebssystem	Ubuntu Linux 18.04.5 LTS	Ubuntu Linux 18.04.5 LTS	Ubuntu Linux 18.04.5 LTS	Ubuntu Linux 18.04.5 LTS	Ubuntu Linux 18.04.5 LTS	Ubuntu Linux 18.04.5 LTS
TensorFlow	1.14	1.14	1.14	1.14	1.14	1.14
Python	3.7.6	3.7.6	3.7.6	3.7.6	3.7.6	3.7.6
XGBoost	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0
Jupyter Notebook	6.0.3	6.0.3	6.0.3	6.0.3	6.0.3	6.0.3

Tabelle 1: Getestete DSW-Konfigurationen

	Desktop-PC-GPUs			Mobile GPUs
	RTX A6000	RTX 8000	RTX 6000	RTX 5000
				
CUDA-Cores	10.752 Ampere	4.608	4.608	3.072
RT-Cores	84 Ampere	72	72	48
Tensor-Cores	336 Ampere*	576	576	384
Arbeitsspeicher	48 GB GDDR6 768 Gbit/s	48 GB GDDR6 bis zu 672 Gbit/s	24 GB GDDR6 bis zu 672 Gbit/s	16 GB GDDR6 bis zu 448 Gbit/s

Tabelle 2: Zum Benchmarking verwendete GPUs

Im Anhang finden Sie alle Benchmarkergebnisse für die RTX A6000-GPU.

Deep Learning: Bildklassifizierung

Das erste Benchmarking wurde mit dem Skript tf_cnn_benchmark von TensorFlow durchgeführt. Das basiert auf dem Convolutional Neural Network (CNN), das als Grundlage für Computer-Vision-Aufgaben wie z. B. Bildklassifizierungen, Objekterkennung, Bildsegmentierungen und Generative Adversarial Network (GAN) gilt. Residual Networks (ResNets), eine von vielen CNN-Topologien, bestehen aus Restblöcken.

Jeder Block hat zwei Zweige, einer führt den Input direkt zum Output, der andere führt zwei bis drei Konvolutionen (Faltungen) durch. Die beiden Zweige werden zusammengeführt und in den nächsten Block eingespeist. Dieser Ansatz führt zu einem CNN mit hoher Performance, da Layer gestapelt werden und gleichzeitig das Problem der verschwindenden Gradienten vermieden wird. Daher gilt ResNet50 (bestehend aus 50 Layern) als Standard für das Benchmarking der Bildklassifizierung.

Wir haben tf_cnn_benchmark mit ResNet50 als Topologie für CNN verwendet. Das Repository von tf_cnn_benchmark enthält Skripte, die Trainings und Inferenzen von standardmäßigen Bildklassifizierungsmodellen anhand von synthetischen Bildern und anderen öffentlichen Datasets wie ImageNet ausführen.

In diesem Fall haben wir das Standarddataset verwendet, das ein synthetisches Dataset ist. Diese Benchmark kann auch auf einzelnen oder mehreren GPUs auf einem oder mehreren Workstation-Nodes ausgeführt werden. Weitere Informationen zu tf_cnn_benchmark finden Sie unter GitHub.

Eine weitere gängige Praxis von Deep-Learning-ExpertInnen ist die Verwendung von Automatic Mixed Precision (AMP). Das umfasst die Verwendung der Gleitkommaformate für einfache Genauigkeit (FP32) und halbe Genauigkeit (FP16), um das Training des neuronalen Netzes ohne Genauigkeitsverlust zu beschleunigen.

Die neuesten NVIDIA Quadro-GPUs, beginnend mit der Volta-Architektur bis zu den neueren GPUs, enthalten spezielle Tensor-Cores für die AMP-Nutzung. In unserem Fall haben wir FP16 für diese Benchmark verwendet.

Die letzte Variable, die wir eingesetzt haben, ist die Batchgröße (BS). Die Batchgröße und die Lernrate sind wohl die beiden wichtigsten Hyperparameter bei der Durchführung verschiedener Experimente, sobald das neuronale Netzmodell ermittelt wurde. Zudem sind diese beiden Parameter oft eng verknüpft. Die Flexibilität, verschiedene Batchgrößen ausführen zu können, ermöglicht es DatenwissenschaftlerInnen, mehr Funktionen zu erkunden und – in einigen Fällen – die Trainingszeit zu verkürzen.

Das hier verwendete Skript tf_cnn_benchmark wird im Modelltraining eingesetzt und für eine festgelegte Anzahl von Iterationen bei durchgeführt, während die durchschnittliche Geschwindigkeit in Bildern/Sek. gemessen wird. Dieses Beispielskript ist für zwei GPUs mit einer Batchgröße von 32 und FP16:

```
python tf_cnn_benchmarks.py --num_gpus=2 --batch_size=32 --model=resnet50 –use_fp16=true
```

Vor diesem Hintergrund haben wir mit verschiedenen Batchgrößen auf verschiedenen GPUs experimentiert, von der Quadro Mobile RTX 5000 mit 16 GB bis zu einer, zwei oder drei Quadro RTX 6000-GPUs mit 24 GB, die im Dell Precision Performance 5820 Tower und 7920 Tower verbaut sind. Die letzten beiden sind Dell Precision Performance Tower mit Intel Xeon Cascade Lake Prozessorplattformen und einem bzw. zwei Sockeln.

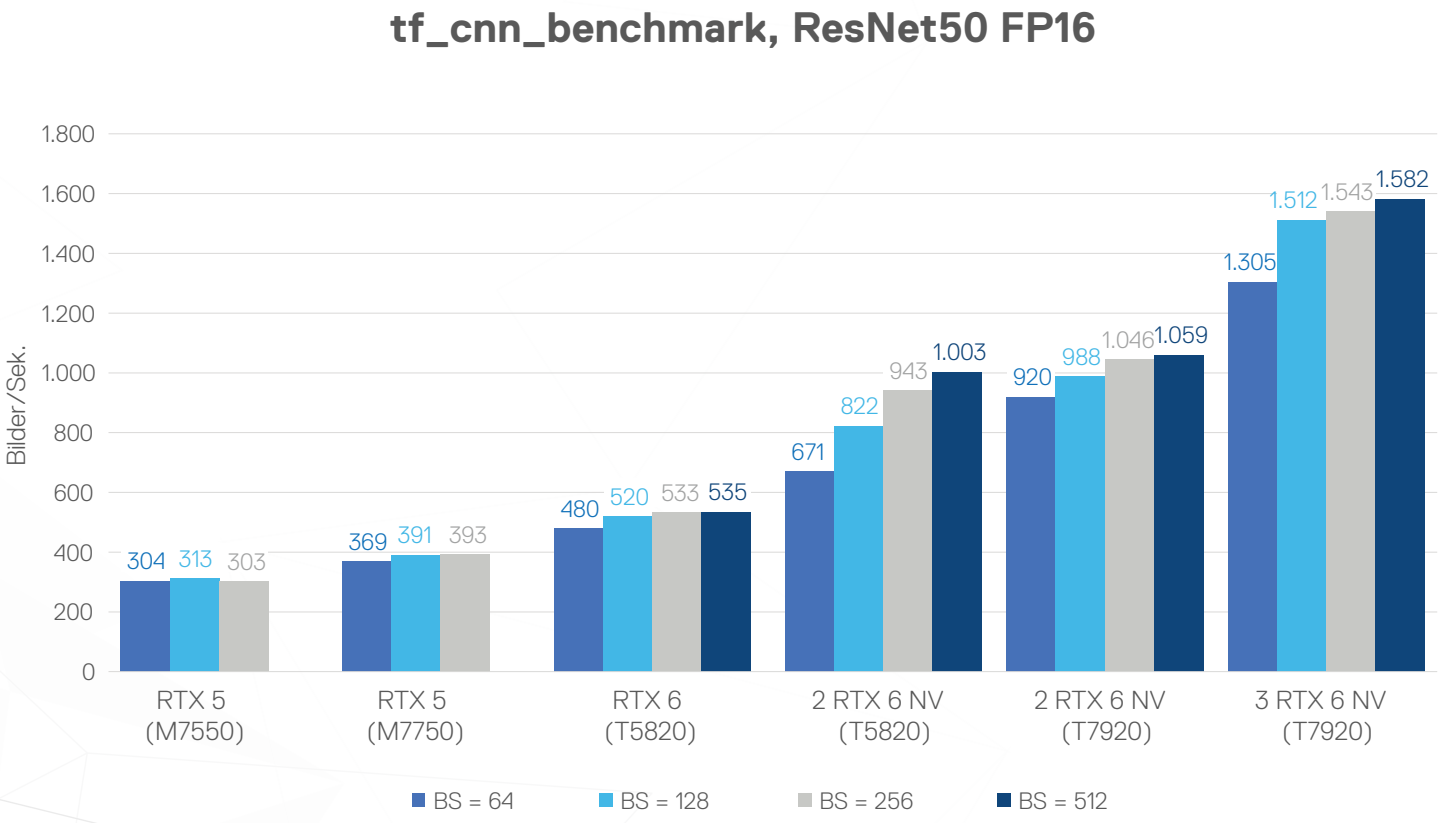


Abbildung 1

Deep Learning: Bildklassifizierung (Fortsetzung)

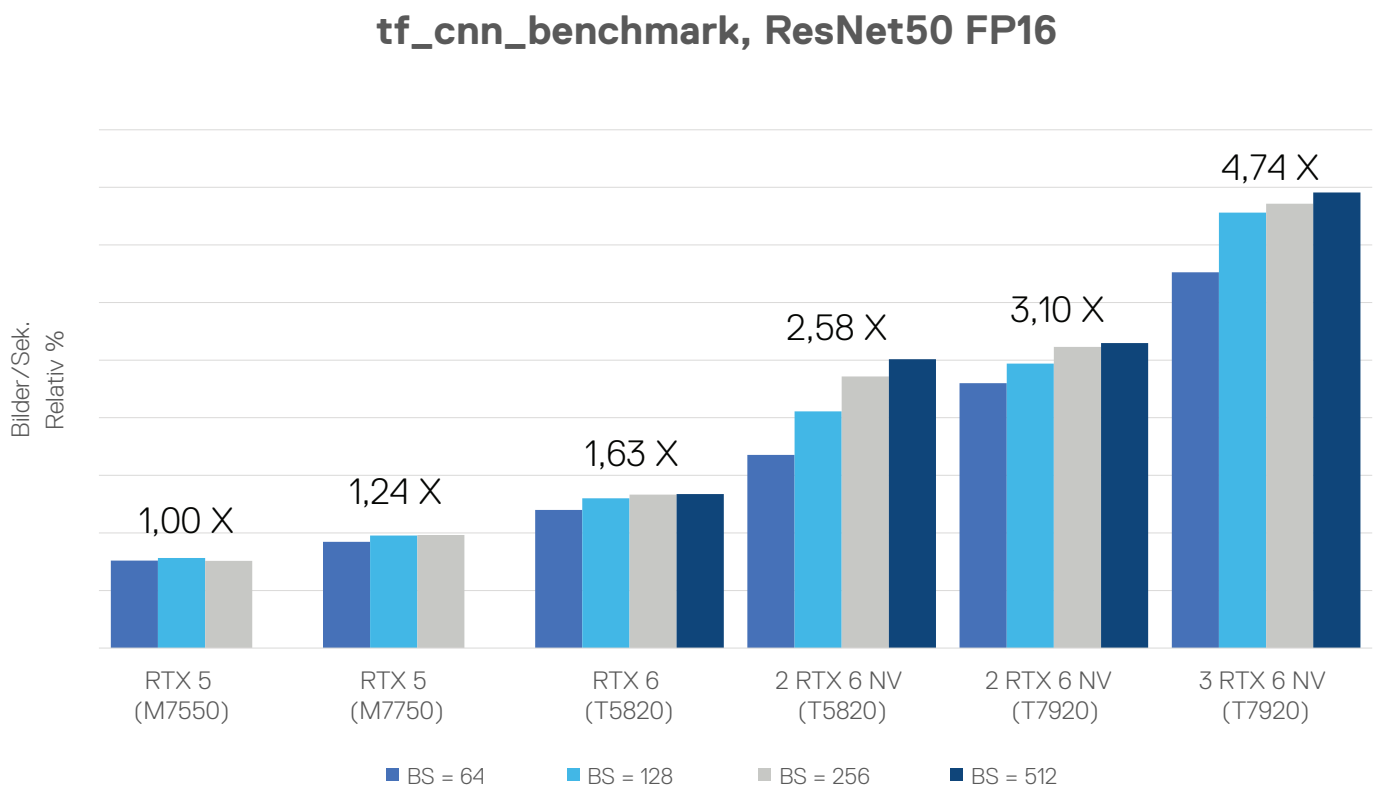


Abbildung 2

Das Diagramm zeigt, dass die Dell Precision 7550 und die 7750 Mobile Workstation diese Aufgaben mit einem sehr guten Wert an Bildern/Sekunden ausführen können. Erwähnenswert ist, dass beim Wechsel von der 7550 Mobile Workstation mit 15" zur 7750 Mobile Workstation mit 17" eine Performanceverbesserung von rund 24 % erzielt wird, obwohl beide Plattformen die gleiche Quadro RTX 5000 verwenden.

Das liegt am größeren Gehäuse der 7750 mit besserer Kühlung, dank der die RTX 5000 mit höheren Taktraten (d. h. einer höheren Auslastung) laufen kann.

Beim 5820 Tower mit einer RTX 6000 stellten wir eine Durchsatzverbesserung von 63 % fest, was auf mehr CUDA-Cores, Tensor-Cores und Videospeicher zurückzuführen ist.

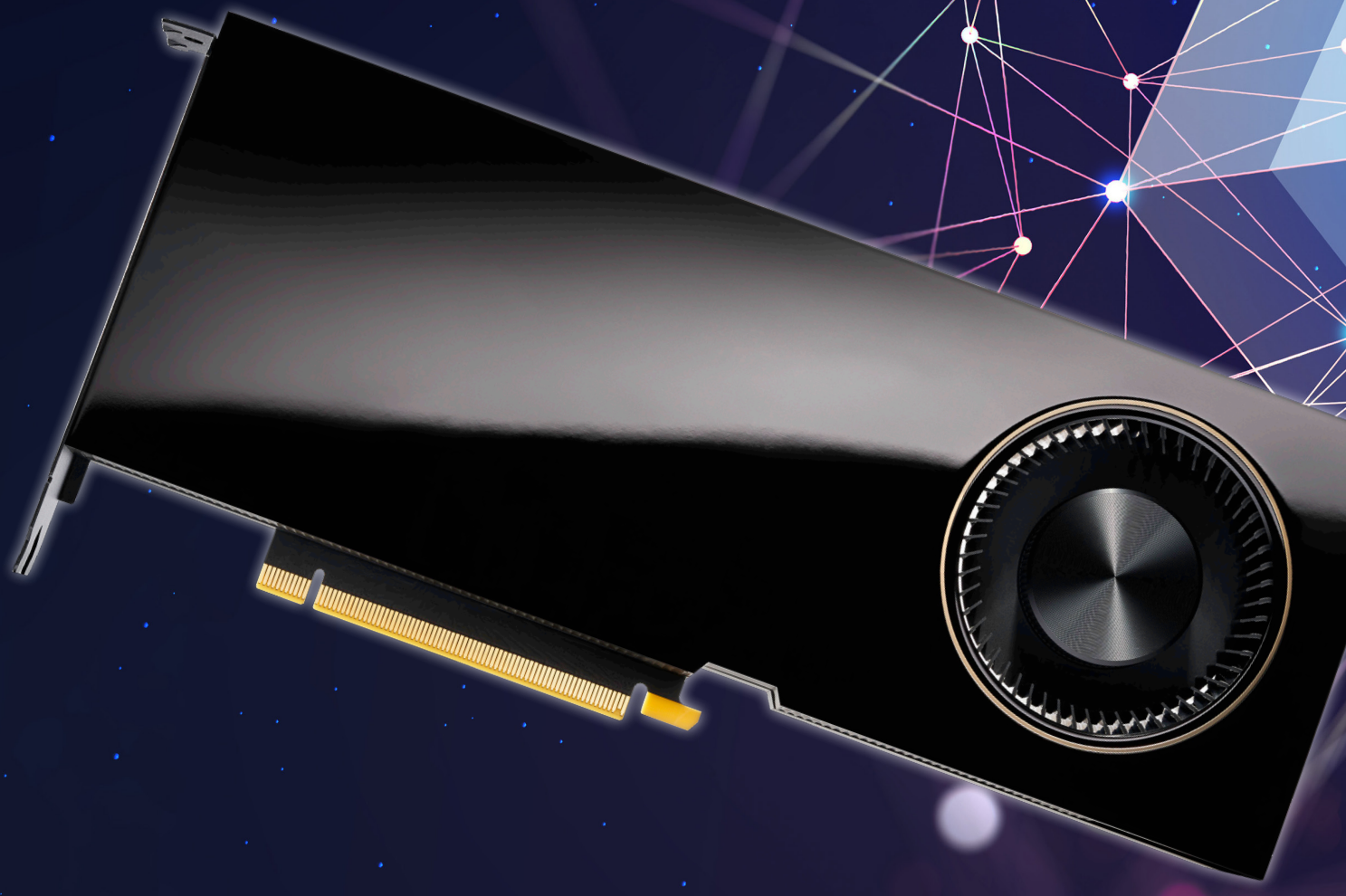
Mit zwei RTX 6000-GPUs und NVLink lässt sich der Durchsatz einer RTX 6000 nahezu verdoppeln, insbesondere bei der 7920 Tower-Plattform (3,1-fach im Vergleich zu 1,63-fach).

Neben der zweiten GPU verfügt der 7920 Tower auch über einen zweiten CPU-Multi-Core-Mikroprozessor. Diese zusätzlichen Cores und Speicherchannel sind erforderlich, damit die beiden Hochleistungs-GPUs vernünftig arbeiten können.

Die dritte GPU bietet eine zusätzliche Steigerung auf eine nahezu lineare Skalierung. Mit dem Befehl „nvidia-smi“ konnten wir bestätigen, dass beim Ausführen dieser Benchmarks sowie weiterer Benchmarks im NLP-Bereich alle GPUs (also zwei und drei) vollständig ausgelastet bleiben. Mit anderen Worten: Die GPUs wurden stets mit Daten der CPU, des CPU-Speichers und der Storage-Systeme „gefüttert“. Es gibt weder ungenutzte Performance noch ungenutzten Durchsatz und keine Engpässe bei der Architektur.

Es war auch klar, dass die Quadro RTX-GPUs der Tower nicht nur ihre Mobile-Gegenstücke übertreffen, sondern dass mit ihnen auch größere Batches möglich sind, wie z. B. BS = 512. Das liegt nicht nur am größeren Videospeicher, sondern auch an den zusätzlichen CUDA-Cores und Tensor-Cores in der RTX 6000 (wie bereits erwähnt).

Zwei GPUs mit NVLink lieferten aufgrund der direkten GPU-GPU-Kommunikation auch einen höheren Durchsatz (ohne Nutzung des vergleichsweise langsamen PCIe-Bus).



Deep Learning: Natural Language Processing

Das zweite Benchmarking wurde im BERT Finetuning mit der NLP-Benchmark von TensorFlow durchgeführt. Beim Natural Language Processing (NLP, Verarbeitung natürlicher Sprache) wird KI eingesetzt, um einem Computer die Fähigkeit zu „verleihen", menschliche Sprache verstehen, analysieren, ändern und erzeugen zu können.

BERT (Bidirectional Encoder Representations from Transformers) ist ein leistungsstarkes Tool, das Ende 2018 von Google entwickelt wurde, mit dem Computer menschliche Sprache verarbeiten, analysieren und „verstehen" können. Das Tool ist zum Standard für verschiedene NLP-Anwendungen geworden, wie z. B. die Beantwortung von Fragen, die Erkennung benannter Entitäten, Inferenz natürlicher Sprache und Textklassifizierung. Zuvor waren alle Sprachmodelle (also Skip-Gram und Continuous Bag-of-Words) unidirektional. Sie konnten das Kontextfenster eines Worts nur von links nach rechts oder von rechts nach links durchlaufen. BERT nutzt eine bidirektionale Sprachmodellierung, um den Kontext eines Worts zu verstehen, das heißt, das Modell „erlernt" den Kontext eines Worts basierend auf der gesamten Umgebung.

BERT Pre-Training

Das Pre-Training-Verfahren folgt weitgehend der vorhandenen Literatur zum Pre-Training von Sprachmodellen. Für den Pre-Training-Wortkorpus nutzt BERT sowohl BooksCorpus (800 Mio. Wörter) als auch die englische Wikipedia-Ausgabe (2,5 Mrd. Wörter). Wie Sie sich vorstellen können, würde dies mit Workstations mehrere Wochen dauern. Daher haben wir uns auf die Benchmarks des BERT Finetuning konzentriert.

BERT Finetuning

Das vortrainierte Modell wird als Basis (Transfer Learning) mit den gleichen Gewichtungen verwendet. Dann werden einige Layer für die spezielle anstehende NLP-Aufgabe hinzugefügt. In unserem Fall bestand die Aufgabe aus Fragen und Antworten. Dies ist ein gängiger Ansatz, um neue NLP-spezifische Aufgaben zu erstellen und die Finetuning-Komplexität deutlich zu reduzieren.

Wir verwendeten das Benchmarkskript finetune_train_benchmar.sh aus dem NVIDIA NGC-Repository BERT for TensorFlow. Mit dem Skript konnten wir das SQuAD v1.1-Dataset mit FP16 oder FP32 testen.

Die Sequenzlängen 128 und 384 sowie die Batchgrößen 1, 2, 4, 8, 16, 32 und 64 wurden während des Skripts ausgeführt. Um alle Batchgrößen auszuführen, haben wir die Datei finetune_train_benchmark.sh in der Zeile 81 bearbeitet und die Batchgrößen 8, 16, 32 und 64 zur folgenden Zeile hinzugefügt:

for batch_size in 1, 2, 4, 8, 16, 32, 64: do

Das Trainingssetup für das BERT Finetuning bietet die Optionen „Base“ oder „Large“. BERT LARGE (L = 24, H = 1.024, A = 16, Summe Parameter = 340 Mio.) und BERT BASE (L = 12, H = 768, A = 12, Summe Parameter = 110 Mio.). Wir haben BERT Large für unsere Benchmarks ausgewählt und den folgenden Befehl verwendet: BERT# scripts/finetune_train_benchmark large true <num_gpus> squad.

Konfigurationen von Dell Precision Workstations

System	Dell Precision Modell	Prozessor	Frequenz	Cores	Arbeitsspeicher	GPU
Konfig. 1	T5820	W-2245	3,9–4,7 GHz	8 Cores	256 GB Speicher	1–2 x RTX 8000
Konfig. 2	T5820	W-2245	3,9–4,7 GHz	8 Cores	256 GB Arbeitsspeicher	1–2 x RTX 6000
Konfig. 3	T7920	5.217	3,0–3,7 GHz	8 Cores	196 GB Arbeitsspeicher	3 x RTX 8000
Konfig. 4	T7920	5.217	3,0–3,7 GHz	8 Cores	196 GB Arbeitsspeicher	3 x RTX 6000
Konfig. 5	T5820	W-2175	2,5–3,4 GHz	14 Cores	64 GB Arbeitsspeicher	2 x GV100

Tabelle 3

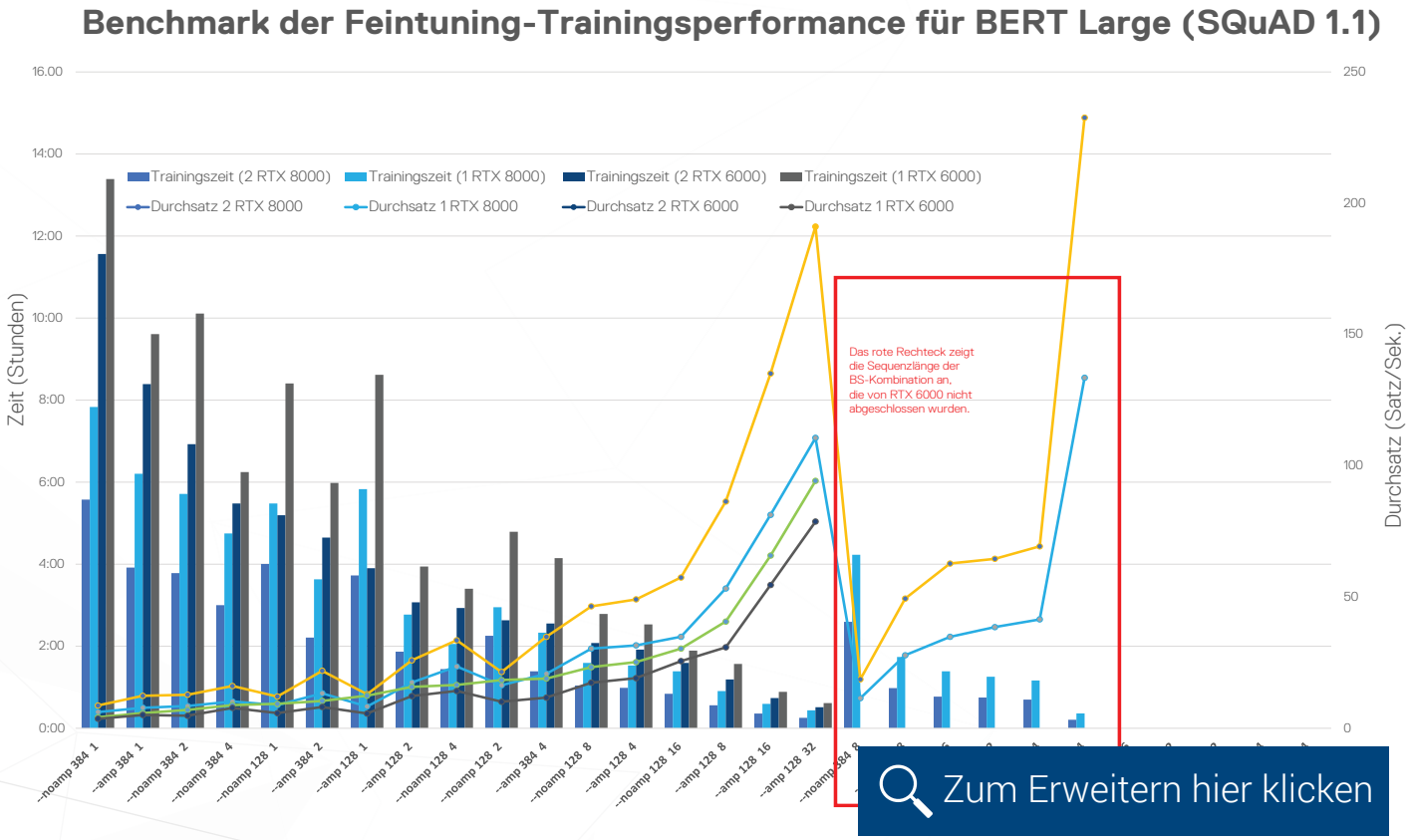


Abbildung 3

Deep Learning: Natural Language Processing (Fortsetzung)

Verschiedene Durchläufe werden auf der horizontalen Achse mit Folgendem angezeigt:

--amp seq_len BS steht für FP16, Sequenzlänge und BS = Batchgröße sowie

--nonamp seq_len BS steht für FP32, Sequenzlänge und BS = Batchgröße

In Abbildung 3, sortiert nach Durchsatz, übertrafen die zwei RTX 6000-GPUs (GRÜN) die einzelne RTX 6000 (GRAU) bei Trainingszeit und Durchsatz (Sätze/Sek.). Dies verstärkte sich bei zunehmender Sequenzlänge und Batchgröße.

Beispiel 1:

FP16, Sequenzlänge 128, Batchgröße 1

1 x RTX 6000 Sätze pro Sekunde: 8,82

2 x RTX 6000 Sätze pro Sekunde: 13,04

Ergebnis: 2 RTX 6000 mit einer Performancesteigerung von 47 %

Beispiel 2:

FP16, Sequenzlänge 384, Batchgröße 2

1 x RTX 6000 Sätze pro Sekunde: 13,77

2 x RTX 6000 Sätze pro Sekunde: 22,08

Ergebnis: 2 RTX 6000 mit einer Performancesteigerung von 60 %

Beachten Sie die Beispiele 1 und 2. Wir haben ebenfalls festgestellt, dass die RTX 6000 FP32 128 32 ausführte, aber FP16/32 64 nicht abgeschlossen wurde. Zudem waren die RTX 6000-Ausführungen FP16 384 8 oder höher ebenfalls unvollständig. Diese werden im obigen Diagramm im roten Rechteck abgebildet. Die RTX 8000 konnte alle Sequenzlängen und Batchgrößen bis zu FP32 384 8 beenden (Beispiel 3).

Beispiel 3:

FP16, Sequenzlänge 384, Batchgröße 8

2 x RTX 6000 Sätze pro Sekunde: unvollständig

2 x RTX 8000 Sätze pro Sekunde: 49,38

Ergebnis: 2 x RTX 6000 – unvollständig

Deep Learning: Natural Language Processing (Fortsetzung)

Trainings-performance	Sequenzlänge	Batchgröße	„Test 1 RTX 8000 (ECC)“		„Test 2 RTX 6000 (ECC)“		
			T5820/W-2245/256 GB		T5820/W-2245/256 GB		T
			2 GPUs	1 GPU	2 GPUs	1 GPU	
FP16	128	1	13,07	8,66	13,04	8,82	
FP32	128	1	12,41	10,27	12,52	10,9	
FP16	128	2	26,26	17,05	26,08	17,11	
FP32	128	2	21,98	17,08	22,13	16,96	
FP16	128	4	50,15	31,59	49,82	31,77	
FP32	128	4	34,86	24,34	35,31	24,59	
FP16	128	8	88,23	54,53	88,01	54,43	
FP32	128	8	49,44	31,27	50,7	31,76	
FP16	128	16	138,19	83,45	138,88	84,25	
FP32	128	16	61,23	35,95	64,18	36,74	
FP16	128	32	196,17	112,5	198,96	114,98	
FP32	128	32	71,09	39,01			
FP16	128	64	246,55	136,15			
FP32	128	64	78,21	41,73			
FP16	384	1	12,53	7,98	12,52	7,89	
FP32	384	1	9,12	6,47	9,2113	6,5	
FP16	384	2	22,14	13,66	22,08	13,77	
FP32	384	2	13,61	8,81	13,89	8,92	
FP16	384	4	35,59	21,14	35,64	21,33	
FP32	384	4	17,55	10,41	18,08	10,57	
FP16	384	8	51,25	29,66			
FP32	384	8	20,79	11,56			
FP16	384	16	65,8	36,12			
FP32	384	16					
FP16	384	32					
FP32	384	32					
FP16	384	64					
FP32	384	64					

Zusammenfassung

Die Auswahl der GPU für NLP-Aufgaben sollte basierend auf der Wichtigkeit von Gleitkomm
Gleitkommazahlen sowie höhere Sequenzlängen und Batchgrößen beenden. Obwohl eine F
scheint der Einsatz von zwei RTX 8000-GPUs die beste Performancelösung zu sein.

In Tabelle 4 gibt es mehrere unvollständige Bereiche, die blau markiert sind. Die meisten davon stammen von RTX 6000-Systemen. Die Performance von drei RTX 8000/RTX 6000 auf den 7920-Systemen war bei einigen Durchläufen langsamer als zwei RTX 8000/RTX 6000.

Bei identischen Grafikkarten übertrafen die Systeme mit einer höheren Prozessorbasisfrequenz und mehr Speicher die Systeme mit einer niedrigeren Prozessorbasisfrequenz und weniger Arbeitsspeicher.

„Test 3 RTX 6000“		„Test 4 RTX 8000 (ECC)“			„Test 5 RTX 6000 (ECC)“		„Test 6 GV100 (ECC)“	
5820/W-2245/256 GB		T7920/5217/196 GB					T5820/W-2175/64 G	
2 GPUs	1 GPU	3 GPUs	2 GPUs	1 GPU	2 GPUs	1 GPU	2 GPUs	1 GPU
13,11	9,64	9,01	12,97	8,29	12,39	5,61	13,25	8,14
12,5	11,35	5,88	12,07	8,82	9,31	5,75	13,44	10,15
26,18	18,89	18,03	25,89	17,45	15,74	12,26	26,63	16,44
22,19	18,5	11,26	21,43	16,38	18,36	10,09	22,46	16,44
50,12	34,89	35,39	49,08	31,58	25,19	19,1	49,94	30,15
35,39	26,03	20,74	33,51	23,51	16,48	14,21	35,21	23,19
88,32	59,75	66,76	86,37	53,26	40,64	30,82	88,46	50,43
50,66	33,2	35,51	46,42	30,33	23,27	17,35	48,52	30,26
139,25	91,14	116,87	135,11	81,31	65,77	54,57	136,72	80,44
63,09	37,82	54,7	57,44	34,84	30,33	25,56	58,26	34,5
198,81	124,1	188,34	191,11	110,61	94,22	78,72	194,62	111,38
		73,84	64,54	38,53			67,3	39,91
		265,85	232,6	133,5				
		91,84	69,29	41,47				
12,54	8,66	8,96	12,34	7,79	5,76	5,03	12,49	7,53
9,21	6,91	5,31	8,67	6,17	4,18	3,61	9,77	6,35
22,21	15,2	16,84	21,87	13,31	10,4	8,08	22,41	12,86
13,9	9,36	9,39	12,78	8,46	6,98	4,78	13,56	8,55
35,77	23,32	29,95	34,83	20,74	18,92	11,65	36,22	20,64
18,02	11,04	14,83	16,1	10,18	8,82	7,74	17,13	10,31
		48,83	49,38	27,8			51,17	30,09
		21,07	18,63	11,43			20,06	11,66
		71,79	62,74	34,81				

Tabelle 4

nazahlen, Sequenzlänge und Batchgrößen erfolgen. Die RTX 8000 kann mehr
RTX 8000 viele der Sequenzlängen und Batchgrößen im Finetuning abschließt,

Maschinelles Lernen: Klassifizierung

Das dritte Benchmarking wurde mit der XGBoost-Bibliothek durchgeführt. XGBoost (eXtreme Gradient Boosting) ist die neueste Entwicklung von auf Entscheidungsstrukturen basierenden Algorithmen. Sie baut auf einem Ensemble aus Baummethoden auf, die das Prinzip der Förderung von schwachen Lernenden anwenden. Bei tabellarischen und strukturierten Daten zählt dies zu den besten Techniken.

Für diese Übung haben wir ein synthetisches numerisches Dataset mit 6 Mio. Zeilen x 501 Spalten verwendet (eine davon ist die Ausgabe), alle mit voller Genauigkeit (FP32). Das tatsächliche Python Jupyter Notebook stammt aus dem Demo-Notebook in Rapids.ai. RAPIDS ist die Open-Source-Suite von NVIDIA mit GPU-beschleunigten Bibliotheken für maschinelles Lernen, einschließlich XGBoost.

Die Dell Data Science Workstation ist mit den gängigsten Data-Science-Bibliotheken, einschließlich RAPIDS und XGBoost, vorkonfiguriert. Daher ist keine zusätzliche Softwareinstallation erforderlich. Zudem ist XGBoost nun GPU-beschleunigt.

Im Test wurden wie zuvor reine CPUs mit verschiedenen GPUs verglichen, und zwar diesmal mit der RTX 5000-GPU für Mobile sowie RTX 6000-, RTX 8000- und RTX GV100-GPUs für die Tower DSW.

Der Test berücksichtigt nur die Trainingszeit mit dem oben angegebenen Dataset unter Verwendung von XGBoost mit denselben Parametern.

Wir haben num_round auf 100 erhöht und das Training-zu-Validierung-Verhältnis auf 90-10 % festgelegt. Bei den restlichen Parametern wurden die Standardeinstellungen des Demo-Notebooks übernommen.

Das folgende Diagramm veranschaulicht eine drastische Reduzierung der Trainingszeit bei den GPUs im Vergleich zu reinen CPUs. Alleine bei dem Wechsel des Trainings von der Xeon W-10885M-CPU zur RTX 5000-GPU benötigte die CPU 50 % mehr Zeit für den Durchlauf als die GPU.

Bei 32 GB HBM2 (RTX GV100 T5820) gab es eine erheblich erhöhte Geschwindigkeit (6,4-mal).

Damit lassen sich wochenlange Trainings auf vielleicht einen Tag reduzieren. Die bessere Performance von RTX 6000, RTX 8000 und RTX GV100 liegt nicht nur am großen Videospeicher, in den das gesamte Dataset passt, sondern auch an der höheren Anzahl von CUDA- und Tensor-Cores, die sich direkt auf die Verarbeitungs- und Konvergenzzeit auswirkt.

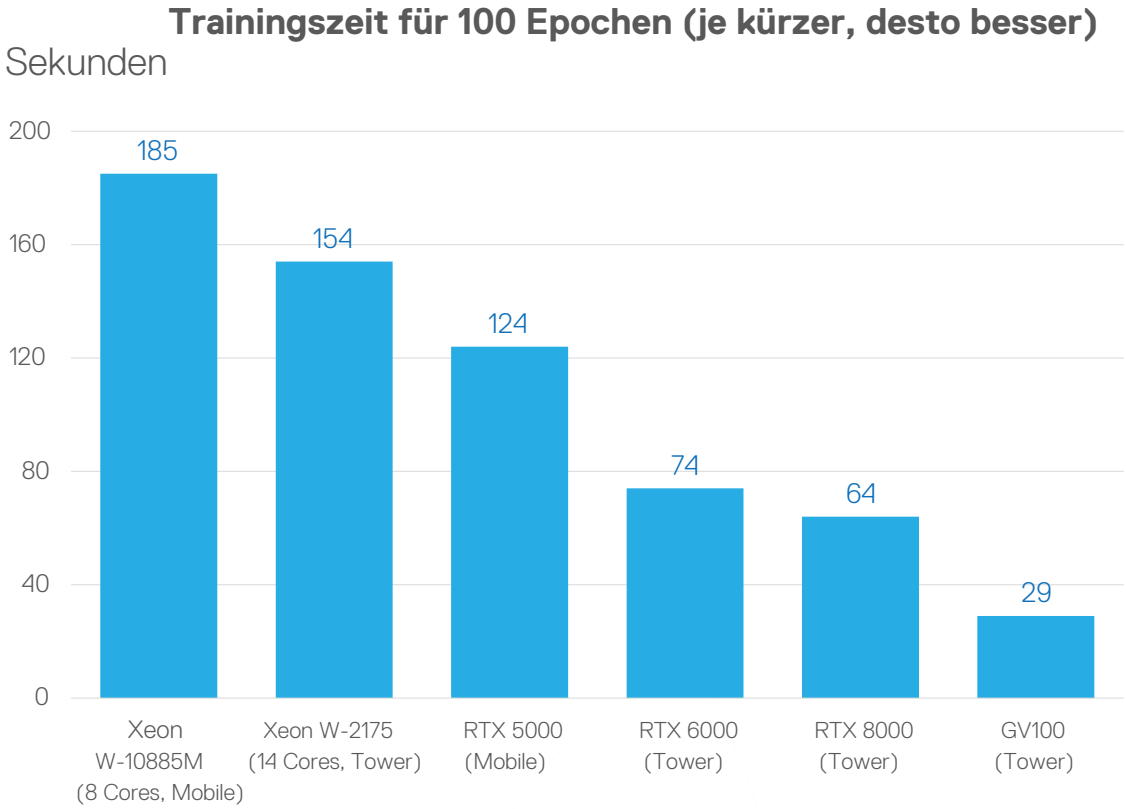


Abbildung 4

Zum Erweitern hier klicken

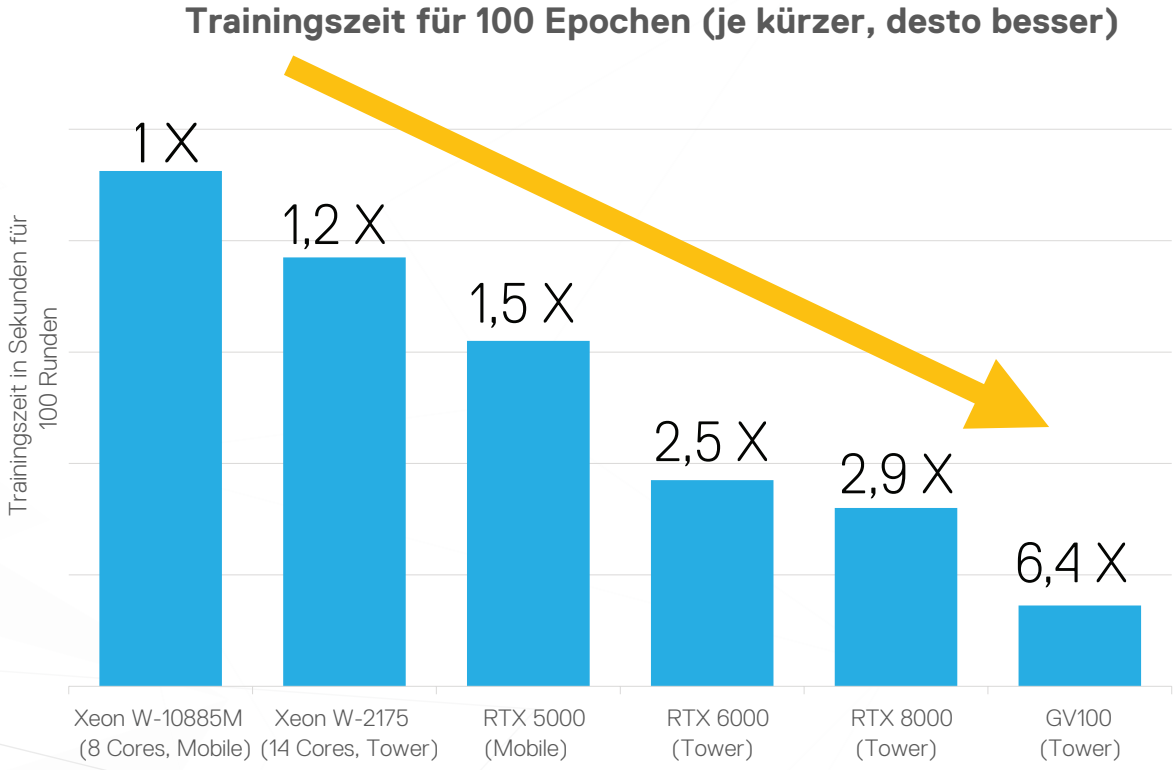


Abbildung 5

Zum Erweitern hier klicken

Richtige Dimensionierung einer DSW

Es gibt drei allgemeine Schritte zur korrekten Dimensionierung und Konfiguration von GPU-beschleunigten Workstations für die KI-Modellentwicklung:

1. Bestimmen Sie den Typ sowie die Speichergröße des zu verwendenden KI-Modells und bewerten Sie die Größe des Datasets, auf dem das Modell ausgeführt werden soll.
2. Legen Sie die Größe für die GPU und den GPU-Speicher fest.
3. Legen Sie die Größe für die CPU, den CPU-Speicher und den Massen-Storage fest.

SCHRITT 1

Legen Sie Datasetgröße und den Ansatz für die KI-Modellentwicklung fest.

Im KI- und Data-Science-Bereich gibt es keine Universallösung, da Datasets und Modelle sehr unterschiedlich sind. Bei klassischen KI-Modellierungsansätzen werden relativ kleine Datasets verwendet. Viele der Probleme, die ML- und DL-Modelle lösen sollen, bestehen jedoch bei sehr umfangreichen und unstrukturierten Datasets.

Diese können ausgesprochen groß sein, wenn es sich bei den zu analysierenden Daten um Bilddaten oder Videos handelt. Unabhängig davon, ob es sich bei dem gewählten Modellierungsansatz um klassische KI, maschinelles Lernen oder Deep Learning handelt: Der Speicher, der idealerweise für die Unterbringung des gesamten Datasets und des Modells benötigt wird, übersteigt bei Weiten den RAM, der physisch in der Plattform verbaut werden kann. Daher ist es in vielen Situationen erforderlich, das Dataset in kleinere Batchgrößen zu unterteilen. Die meisten DatenwissenschaftlerInnen entscheiden sich dafür, die Batchgröße gemäß dem verfügbaren physischen Arbeitsspeicher zu maximieren.

In einer GPU-beschleunigten KI-Workstation erfolgen das Housing und die Durchführung des Modelltrainings im GPU- und GPU-Speichersystem (und nicht im CPU- und CPU-Speichersystem). Daher besteht der nächste Schritt – nachdem Sie den Modellierungsansatz bestimmt und die Größe des verwendeten Datasets bewertet haben – in der Dimensionierung der GPU und des GPU-Speichers.

SCHRITT 2

Wählen Sie die am besten geeignete GPU und die GPU-Speichergröße aus.

GPU-DDR-Speicher, der in GPU-Videopuffern verwendet wird, ist leistungsstärker und schneller als der CPU-DDR-Speicher. Der Grund dafür ist, dass der GPU-DDR-Speicher stets die hohen Auslastungsanforderungen der parallel verarbeitenden Compute-Cores in der GPU erfüllen muss.

- **Dell Precision 7550 und 7750 Mobile Data Science Workstations**
  Sie nutzen die NVIDIA RTX 5000-GPU mit 3.072 CUDA-Cores, 48 RT-Cores, 384 Tensor-Cores und 16 GB GDDR6 mit einer Übertragungsrate von bis zu 448 Gbit/s. Dieser GPU-Speicherpuffer eignet sich für Datasets oder Batchgrößen von bis zu 16 GB.

- **Dell Precision 5820 und 7920 Tower Data Science Workstations**
  Sie verwenden NVIDIA RTX A6000-, RTX 6000- und RTX 8000-GPUs. Die RTX 6000-GPU umfasst 4.608 CUDA-Cores, 72 RT-Cores, 576 Tensor-Cores und GDDR6-Speicher mit einer Übertragungsrate von bis zu 672 Gbit/s.
  Die RTX 6000 und die RTX 8000 unterscheiden sich nur in ihrer jeweiligen GPU-Speicherpuffergröße. Die RTX 6000 hat 24 GB und die RTX 8000 48 GB. Dadurch ist jede RTX 6000 für Datasets oder Batchgrößen von bis zu 24 GB und jede RTX 8000 entsprechend bis zu 48 GB geeignet.

Die 5820 DSW ist eine Plattform mit Tower-Formfaktor und kann mit einer oder zwei RTX A6000-, RTX 8000- oder RTX 6000-GPUs konfiguriert werden. Werden die beiden GPU-Karten über NVLink verbunden, verdoppelt sich die Größe des GPU-Videopuffers. Damit kann die 5820 DSW Datasets oder Batchgrößen von bis zu 96 GB unterstützen.

Die 7920 DSW ist als Plattform mit Tower- oder Rack-Formfaktor erhältlich und kann mit einer, zwei oder drei RTX A6000-, RTX 8000- oder RTX 6000-GPUs konfiguriert werden. Sie kann mit drei RTX A6000- oder RTX 8000-GPUs bis zu 144 GB an GPU-Speicher bereitstellen.

Auf der 7920 Tower DSW können zwei dieser drei GPU-Karten über NVLink verbunden werden, wodurch sich die Größe des GPU-Videopuffers verdoppelt.

Auf diese Weise unterstützt die 7920 DSW Datasets oder Batchgrößen von bis zu 96 GB – plus bis zu 48 GB zusätzlichen GPU-Speicher von der dritten GPU.

Beobachtungen und Best Practices

SCHRITT 3

Wählen Sie die optimale CPU-Speicher- und CPU-Prozessorkonfiguration für die CPU-Seite aus, um stets die Compute- und Datenanforderungen der GPU-Seite des Systems zu erfüllen.

Das Ziel ist, während der Durchläufe des Modelltrainings eine GPU-Auslastung von nahezu 100 % zu erreichen. Abschließend geht es um die Storage-Konfiguration und -Dimensionierung.

Unsere Erfahrung bei der Performancekonfiguration und -messung von zahlreichen GPU-beschleunigten Workloads hat folgende Best Practice ergeben: Wenn die CPU-Speicherdimensionierung zu einer konstant vollen Auslastung der GPU-Seite des Systems führen soll, müssen Sie doppelt so viel CPU-DDR-Speicher wie GPU-Speicher haben.

Beispiel: In einer 5820 Tower DSW mit zwei RTX A6000-GPUs konfigurieren Sie einen CPU-Speicher von 192 GB. Bei einer 7550 Mobile DSW mit einer RTX 5000-GPU konfigurieren Sie einen CPU-Speicher von mindestens 32 GB.

Die Aussagen der DatenwissenschaftlerInnen decken sich mit unserer eigenen Erfahrung beim Training von verschiedenen ML- und, insbesondere, DL-Modellen: Bei der Konfiguration von GPU-beschleunigten KI-Workstations ist mehr als ein Multi-Core-Prozessor erforderlich, um Systeme mit zwei oder drei konfigurierten GPUs stets voll auszulasten. Aus diesem Grund empfehlen wir folgende Best Practice:

- Planen Sie einen Prozessor-Core pro 8–16 GB an CPU-DDR-Speicher ein.
- Falls ein Prozessor zu wenig Cores hat, teilen Sie die Anzahl der Cores zwischen den beiden CPU-Prozessoren auf.
- Es kann auch wirtschaftlich sinnvoll sein, die Core-Anzahl zwischen zwei Prozessoren aufzuteilen.

Beispiel: Bei zwei RTX A6000-GPUs in einer 7920 Tower DSW konfigurieren Sie die 7920 Tower DSW mit 192 GB CPU-DDR-Speicher und planen die Verwendung von zwei Xeon Prozessoren mit jeweils etwa 6–12 Cores ein.

So konfigurieren Sie CPU-DDR-Speicher

Es ist wichtig, alle Speicherchannel im Speichercontroller des Prozessors vollständig zu nutzen. So wird die auf der CPU-Seite des Systems verfügbare CPU-Speicherbandbreite maximiert, damit die GPU-Seite des Systems stets voll ausgelastet bleibt. Wenn dies nicht geschieht, bleibt ein beträchtlicher Teil der Systemperformance ungenutzt:

- Bei der 7550 und 7750 Mobile DSW müssen dafür die DDR-DIMM-Sockel mit einem Vielfachen von vier bestückt werden.
- Bei der 5820 Tower DSW müssen dafür die DDR-DIMM-Sockel mit einem Vielfachen von vier bestückt werden.
- Bei der 7920 Tower und der 7920 Rack DSW müssen dafür die DDR-DIMM-Sockel jedes Prozessors mit einem Vielfachen von sechs bestückt werden.

Abschließend geht es um die Größe des Massen-Storage im System. Laut den Best Practices, die uns die DatenwissenschaftlerInnen mitgeteilt haben, sollte der im System für „warme Daten“ verwendete Massen-Storage von dem Massen-Storage, der als Startlaufwerk verwendet wird, getrennt werden. Bei KI-Workloads sollten Solid-State-Laufwerke (SSDs) für sowohl das Startlaufwerk als auch die Datenlaufwerke eingesetzt werden. SSDs der Klasse 50 werden empfohlen, aber auch SSDs der Klasse 40 können ein guter Kompromiss sein, wenn es um Kosteneffizienz geht oder wenn dadurch mehr Budget für CPUs mit höherer Core-Anzahl freigesetzt wird.

Da die Größe und Komplexität von Daten in KI-Workloads dramatisch zunimmt, steigt auch die Größe der Datenlaufwerke an. Zum Zeitpunkt der Veröffentlichung dieses Dokuments erfüllen SSDs mit 1–2 TB die meisten Datenanforderungen für die KI-Modellentwicklung mit Workstations.

Einige DatenwissenschaftlerInnen möchten auch Storage für „kalte Daten“ als Unterstützung der SSDs für „warme Daten“ hinzufügen. Natürlich können das ebenfalls SSD-Festplatten sein, aber dafür lassen sich auch kosteneffiziente SATA-Laufwerke einbauen. Wir empfehlen aber, Storage-Laufwerke mit SATA-Performance nur für diesen Storage zu verwenden.

Fazit

Die Erkenntnisse aus der Durchführung der Benchmarks in diesem Whitepaper zeigen, dass die Auswahl der richtigen GPU und Plattform von der Workload abhängig ist.

Mobile Workstations bieten Funktionen, um Aufgaben der Bereiche Computer Vision und NLP ortsunabhängig auszuführen. Für anspruchsvollere Workloads – wie z. B. hochauflösende medizinische Bildklassifizierung oder große NLP-Modelle mit umfangreicher Sequenzlänge und Batchgröße – sind leistungsstärkere GPUs wie die RTX 6000/8000/A6000 in Dell Tower Workstations die richtige Lösung.

In den neuesten NLP-Modellen (wie z. B. BERT LARGE mit mehr als 340 Mio. Parametern) war das Experimentieren mit einer Kombination aus höherer Sequenzlänge und Batchgröße nur mit der RTX A6000 oder der RTX 8000 aufgrund des GDDR6-Speichers von 48 GB und der hohen Anzahl an CUDA-Cores möglich.

Bei ML-Aufgaben, wie z. B. die Verarbeitung von tabellarischen und strukturierten Datasets, können High-End-GPUs wie die RTX 6000/8000/A6000 das Laden des Datasets in den GPU-Speicher sowie das Training beschleunigen (bei der Verwendung von GPU-beschleunigten Bibliotheken wie XGBoost).

Im letzten Abschnitt dieses Whitepapers wurden allgemeine Anleitungen und Best Practices zur richtigen Dimensionierung der Dell Precision Data Science Workstation vorgestellt.

Weitere Informationen und verwandte Themen

Hier finden Sie Informationen und Links, die in diesem Whitepaper referenziert werden, sowie weiterführende Informationen und einfach zu verwendende Links zu vorkonfigurierten Dell Precision Data Science Workstations, die Sie erwerben und anpassen können. Auf dieser Landingpage finden Sie zudem die neuesten Informationen, die zum Zeitpunkt dieser Veröffentlichung noch nicht verfügbar waren.

Dell – Links:

Dell DSW und Isilon H400-NAS-Lösungsperformance – Whitepaper

KI-Kurzübersicht

DSW – Installationshandbuch

KI/DSW – Branchenübersicht

DSW – Komponentenübersicht

NVIDIA – Links:

NVIDIA Data Science Stack

NVIDIA GPU-beschleunigtes KI-Training

Canonical – Links:

https://certification.ubuntu.com

https://ubuntu.com/dell

https://ubuntu.com/contact-us

Anhang 1

Performanceergebnisse für die RTX A6000-GPU

In diesem Anhang werden die Ergebnisse der Ausführung von Benchmarks zur Performanceanalyse der NVIDIA RTX A6000-GPU ergänzt. Wir haben uns an dieselbe Benchmarkingmethodik gehalten, die wie im Text dieses Whitepapers beschrieben für die GPUs RTX 5000, RTX 6000, RTX 8000 und GV100 verwendet wurde. Bei den hier durchgeführten Benchmarks zur Performanceanalyse erfolgte ein Update des Linux-Betriebssystems von Ubuntu 18.04 auf Ubuntu 20.04 sowie ein Update des NV Data Science Software-Stacks von 2.4.0 auf 2.8.0. Aus diesem Grund spiegeln die Ergebnisse nicht nur die Performanceverbesserungen der RTX A6000-GPU-Hardware wider, sondern auch die Optimierungen in Ubuntu 20.04 sowie die Bibliotheken und Gerätetreiber im NVIDIA Data Science Software-Stack 2.8.0.

Die Dell DSW-Hardwareplattformen und -Konfigurationen

Wir haben wieder die Dell Precision 5820 Tower DSWs und die Dell Precision 7920 Tower DSWs für die Benchmarkausführung verwendet, und zwar mit einer, zwei oder drei installierten RTX A6000-GPUs mit und ohne per NVLink verbundenen Paaren aus RTX A6000-GPUs. Zudem haben wir die Benchmarks für ein, zwei und drei RTX 6000- und RTX 8000-GPUs erneut durchgeführt, da es seit den ursprünglichen Benchmarks, die im Hauptteil dieses Dokuments beschrieben wurden, Updates für das Ubuntu Linux-Betriebssystem und den NVIDIA Data Science Software-Stack gab.

Die Benchmarks

Wir haben die aktuellen Versionen der zuvor durchgeführten Benchmarks verwendet, um die Dell DSW-Performance für drei verschiedene Workloads zu erfassen:

- Maschinelles Lernen mit XGBoost für epidemiologische Daten (strukturierte Daten)
- Deep Learning zur Bildklassifizierung – tf_cnn für ResNet50-Imaging (unstrukturierte Daten)
- Deep Learning für Natural Language Processing (NLP) – BERT Large für das Dataset SQuAD v1.1 (unstrukturierte Daten)

A6000-Ergebnisse: maschinelles Lernen mit XGBoost für epidemiologische Daten (strukturierte Daten)

XGBoost (eXtreme Gradient Boosting) ist die neueste Entwicklung von auf Entscheidungsstrukturen basierenden Algorithmen. Sie baut auf einem Ensemble aus Baummethoden auf, die das Prinzip der Förderung von schwachen Lernenden anwenden. Bei tabellarischen und strukturierten Daten gilt dies als die beste Technik.

Wir haben ein simuliertes Dataset der britischen Bevölkerung verwendet, das aus offiziellen Daten der britischen Volkszählung erstellt wurde, um die Wahrscheinlichkeit vorherzusagen, mit der sich eine Person mit einem simulierten Virus infiziert. Das Dataset wurde im RAPIDS dli-Training verwendet. Für die Testausführung haben wir ein Demo-Notebook von Rapids.ai verwendet. RAPIDS und XGBoost sind im NVIDIA Data Science Software-Stack enthalten. Wir haben das RAPIDS 0.18-Standardsegment des Data Science Stack 2.8.0 genutzt.

Das von uns verwendete Dataset umfasste 6 Mio. Zeilen x 501 Spalten (die Ausgabe in Spalte 501 gibt an, ob die Person infiziert ist oder nicht), alles mit dem Gleitkommaformat FP32. Das Modelltraining erfolgte auf RTX A6000-GPUs und der RTX 6000, um die Dauer bis zur Beendigung von 500 Iterationen zu vergleichen.

In Abbildung 6 werden unsere Ergebnisse dargestellt. Die RTX 6000 benötigte 23,7 Sekunden für den Abschluss von 100 Epochen, während die RTX A6000 nur 14,5 Sekunden benötigte und somit 38 % schneller war. Die GV100 war am schnellsten, sie brauchte nur 9,8 Sekunden.

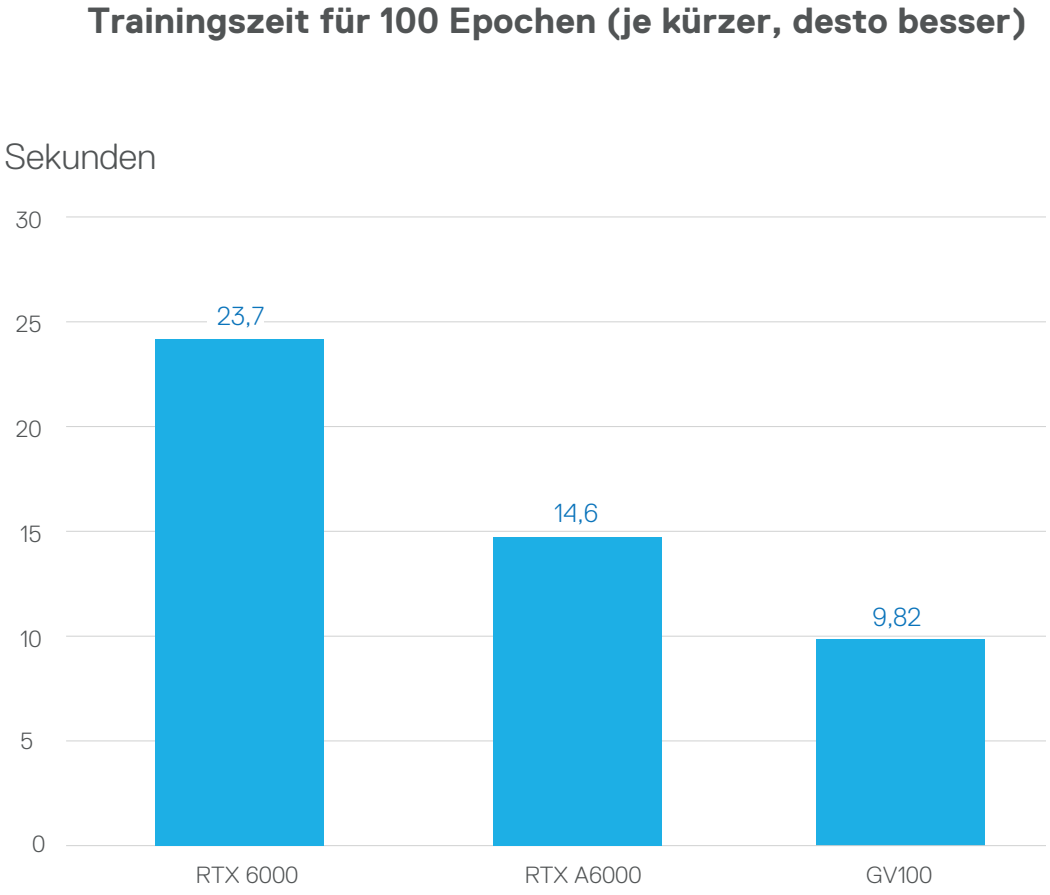


Abbildung 6

Anhang 1 (Fortsetzung)

A6000-Ergebnisse: Deep Learning zur Bildklassifizierung – tf_cnn für ResNet50-Imaging (unstrukturierte Daten)

Der Test wurde mit dem offiziellen Skript tf_cnn_benchmarks von TensorFlow durchgeführt. Das Repository enthält Skripte für das Training von standardmäßigen Bildklassifizierungsmodellen anhand von synthetischen Bildern und von ImageNet-Datasets. Die Trainingsdurchführung erfolgt für eine festgelegte Anzahl an Iterationen, dabei wird die durchschnittliche Geschwindigkeit in Bildern/Sekunde gemessen. Um die Funktionen der Workstation vollständig zu testen, führten wir die Benchmark mit verschiedenen Batchgrößen (BS) und GPU-Anzahlen durch. Für das ResNet50-Training mit dem synthetischen Bilddatasets haben wir tf_cnn verwendet. Wir haben das Gleitkommaformat der halben Genauigkeit (FP16) eingesetzt.

In Abbildung 7 werden die Ergebnisse der ResNet50-Bildklassifizierung mit tf_cnn dargestellt. Die Abbildung zeigt, dass die mit der Ampere RTX A6000 konfigurierten Dell DSWs sowohl die RTX 6000- als auch die RTX 8000-GPU und sogar die GV100-GPU deutlich übertrafen. Zudem verdeutlicht sie, dass sehr viel Speicher benötigt wurde, um Durchläufe mit umfangreichen Batchgrößen (Batchgröße = 1.024) zu beenden. Diese konnten von der RTX 6000 gar nicht verarbeitet werden. Zwei per NVLink gekoppelte GPUs lieferten bei der Bildklassifizierung mit tf_cnn Benchmark keine nennenswert bessere Performance. Abgeleitet wurde, dass der PCIe3-Bus einfach ausreicht, um den Datenverkehr zwischen den beiden Karten zu verarbeiten. Die Performance von zwei RTX A6000 war in den meisten Fällen doppelt so hoch wie die einer einzelnen RTX A6000-GPU. Eine hinzugefügte dritte RTX A6000 sorgte für lineare Skalierung, sodass im Training noch mehr Samples in weniger Zeit möglich sind. Folglich wird auch die Zeitspanne, die zum Erreichen des endgültigen Deep-Learning-Modells erforderlich ist, verkürzt.

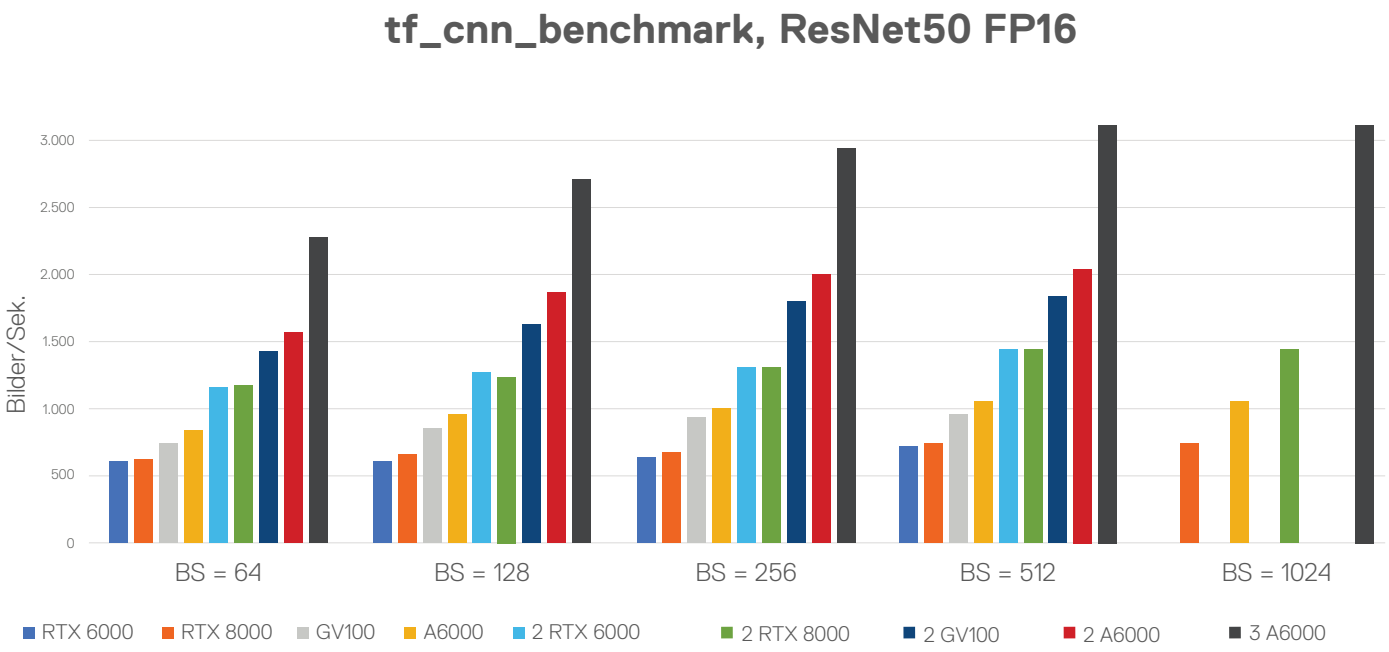


Abbildung 7

Zum Erweitern hier klicken

A6000-Ergebnisse: Natural Language Processing (NLP) – BERT Large für das Dataset SQuAD v1.1 (unstrukturierte Daten)

Natural Language Processing (NLP) wird verwendet, um menschliche Sprache zu verstehen und zu erzeugen. Für die NLP-Performance haben wir BERT (Bidirectional Encoder Representations from Transformers) Large verwendet.

BERT ist zu einer Standardmessung für NLP geworden. BERT nutzt eine bidirektionale Sprachmodellierung, um den Kontext eines Worts zu verstehen, das heißt, das Modell „erlernt" den Kontext eines Worts basierend auf der gesamten Umgebung. Für unsere Messungen haben wir BERT Large für das Training mit dem SQuAD v1.1-Dataset ausgewählt.

Mit den Dell DSWs haben wir folgende BERT Parameter durchlaufen:

- Gleitkommaformat: FP16 und FP32
- Sequenzlänge: 128, 384
- Batchgröße: 1, 2, 4, 8, 16, 32 und 64

In Abbildung 8 werden die Ergebnisse als „Training pro Sekunde" dargestellt. Eine erhebliche lineare Skalierung erfolgt von einer einzelnen zu zwei und drei RTX A6000-GPUs. Die RTX A6000 konnte eine höhere Kombination aus Sequenzlänge und Batchgrößen verarbeiten. Bei längeren Datensequenzen und umfangreicheren Batchgrößen sowie höherer Genauigkeit (d. h. Gleitkommaformat FP32) wird die Skalierung von einer GPU zu zwei und drei GPUs noch deutlicher. In diesen Fällen konnte die RTX A6000 (in einzelner oder mehreren Konfigurationen) den Durchlauf wegen der Beschränkung auf 24 GB Speicher – im Vergleich zu 48 GB bei 2 RTX A6000 – nicht ausführen.

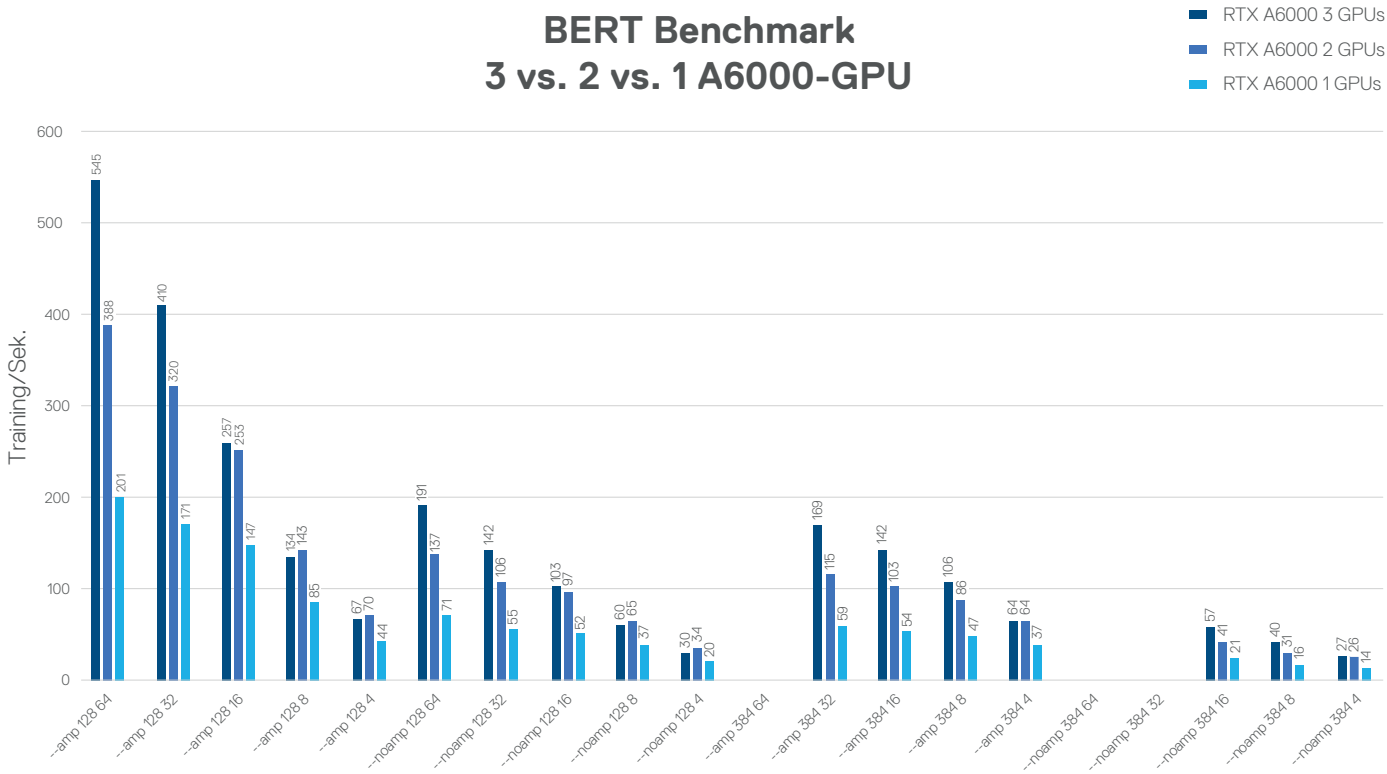


Abbildung 8

Zum Erweitern hier klicken

Anhang 1 (Fortsetzung)

A6000-Ergebnisse: Vorteil der NVLink-Verbindung von zwei RTX A6000-GPUs zu einem Paar

In Abbildung 9 werden die Ergebnisse des BERT Large-Benchmarkings dargestellt, in dem ein Paar aus RTX A6000-GPUs mit und ohne NVLink verglichen wurde. Durch eine NVLink-Verbindung zwischen einem RTX-GPU-Paar kann der Speicher jeder GPU-Karte als ein einheitlicher Speicher betrachtet werden. Dadurch verdoppelt sich sowohl der Speicher als auch die Anzahl der Tensor-Cores, die in diesem Gesamtspeicher für das Modelltraining genutzt werden. Ebenso wichtig ist, dass Daten und Kommunikation zwischen den beiden GPU-Karten über den NVLink erfolgen können (und nicht über den viel langsameren PCIe-Bus). Das kann zu einer deutlich höheren Performance bei DL- oder ML-Modellen führen, die mit sehr großen Datasets trainiert werden. Das ist der Fall bei BERT für solche NLP-Trainingsdurchläufe wie die, die wir durchgeführt haben. Das Diagramm veranschaulicht, dass ein über NVLink verbundenes Paar aus RTX A6000-GPUs eine viel höhere Performance bietet als zwei RTX A6000-GPUs ohne NVLink-Verbindung. Die Performance ist im Vergleich zu einer Verbindung der beiden GPUs über einen PCIe3.2-Bus um 2,5–9-mal höher.

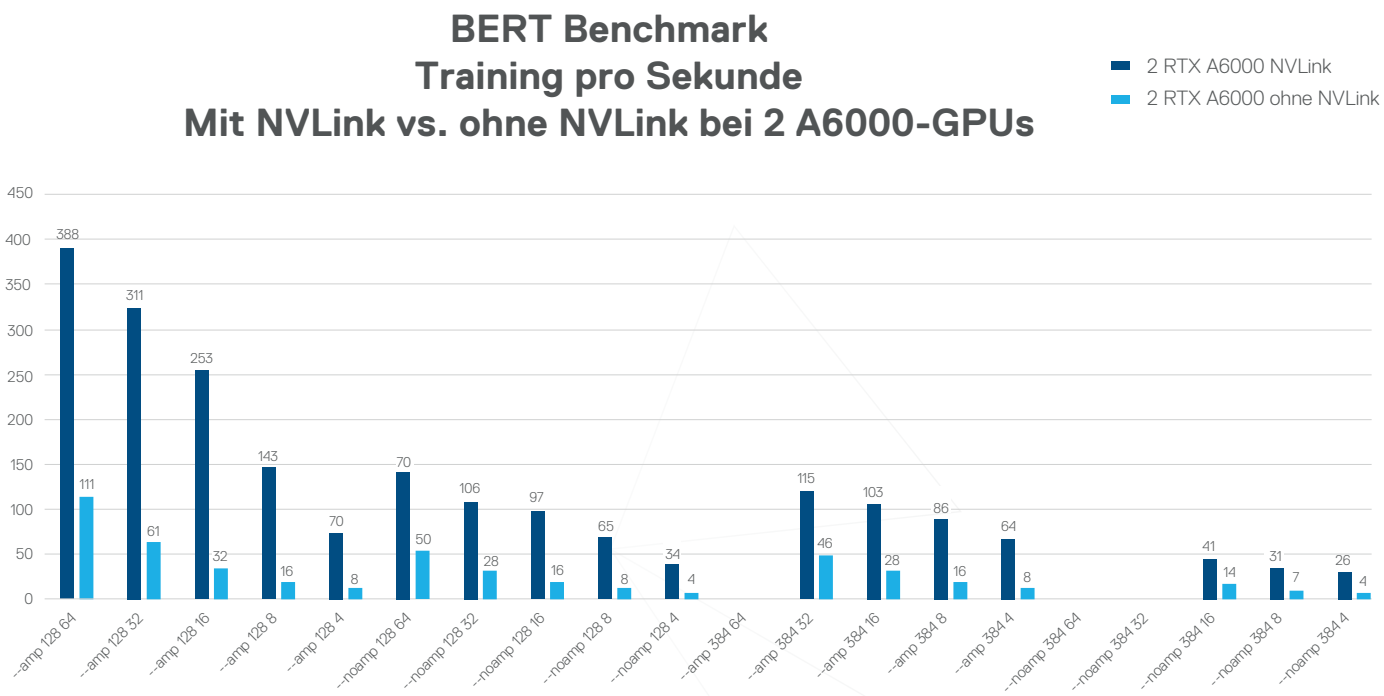


Abbildung 9

Zum Erweitern hier klicken


Precision 7920
Precision 5820

NVIDIA und NVIDIA Quadro sind Marken und/oder eingetragene Marken der NVIDIA Corporation in den USA und/oder weiteren Ländern.
© 2021 Dell Inc. oder deren Tochtergesellschaften.
Alle Rechte vorbehalten. Dell Technologies, Dell EMC, Dell und andere Marken sind Marken von Dell Inc. oder deren Tochtergesellschaften.
Andere Marken können Marken ihrer jeweiligen Inhaber sein.
Die Produkte können von den Abbildungen abweichen.