DELL Technologies

Station de travail Dell Precision pour la science des données : points de référence et pratiques d'excellence pour l'IA, v2.0

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

NVIDIA RTX

Nous publions une version 2.0 de ce document pour présenter les conclusions résultant de la mise à niveau d'Ubuntu Linux par Canonical (de la version 18.04 vers la version 20.04), de la mise à niveau du logiciel NVIDIA Data Science par NVIDIA (de la version 2.4.0 vers la version 2.8.0) et de l'ajout du processeur graphique NVIDIA RTX A6000. La description, la configuration et les conclusions pour Ubuntu 20.04, le logiciel NV Data Science 2.8.0 et le processeur graphique RTX A6000 se trouvent à l'Annexe 1.

Alors que les calculs intensifs rendus possibles par l'intelligence artificielle (IA) deviennent un élément essentiel des modèles économiques de la plupart des organisations, il est de plus en plus vital de trouver les outils, les technologies et les techniques appropriés pour les réaliser efficacement. Pour démontrer comment la sélection des composants appropriés peut optimiser l'efficacité des tâches d'IA, Dell a comparé diverses configurations de stations de travail Dell Precision pour la science des données dans le cadre de workflows de Deep Learning et d'apprentissage automatique.

L'analyse de Dell démontre clairement les avantages considérables de l'accélération par processeur graphique et inclut toutes les données d'origine. Par conséquent, les tiers peuvent tester et valider les conclusions. Un point de référence d'entraînement d'un modèle d'apprentissage automatique révèle que l'exécution sur un processeur prend 6,4 fois plus de temps que sur une configuration avec processeur graphique. Un autre point de référence de Deep Learning affiche des résultats jusqu'à 4,74 fois plus rapides du fait de l'accélération par plusieurs processeurs graphiques.

La seconde moitié de ce document utilise ces conclusions pour fournir un guide simple des pratiques d'excellence permettant de sélectionner les composants optimaux pour n'importe quel workflow.

Introduction : L'environnement de l'IA

L'entreprise a rapidement changé au cours de la dernière décennie. L'Internet of Things et son équivalent industriel produisent ainsi de grandes quantités de données. Les estimations d'IDC suggèrent que ces dernières atteindront 175 zettaoctets dans le monde d'ici 2025, soit dix fois plus qu'en 2017.

Par conséquent, les besoins en puissance de traitement augmentent et, dans le même temps, les avancées en matière d'algorithmes, de logiciels Open Source et d'accélérateurs matériels spécialisés font exploser l'adoption de l'intelligence artificielle (IA).

Le domaine de l'IA est vaste et les sous-ensembles de l'IA que sont l'apprentissage automatique et le Deep Learning sont considérés comme les approches basées sur les données les plus avancées qui soient pour résoudre bon nombre des problèmes complexes d'aujourd'hui.

Ces techniques libèrent la valeur cachée dans d'énormes quantités de données provenant de sources aussi diverses que les dépenses des clients, les vidéos YouTube et les machines des usines. Mais les choses ne sont pas si simples et il faut les technologies, les compétences et l'équipement appropriés pour extraire des renseignements exploitables de cette matière brute.

L'apprentissage automatique et le Deep Learning présentent chacun leurs propres cas d'utilisation et défis. L'apprentissage automatique est moins sophistiqué et généralement utilisé sur des données structurées, telles que les données tabulaires, traitées à l'aide d'algorithmes bien connus tels que la régression linéaire, la régression logistique, le classifieur bayésien naïf et XGBoost.

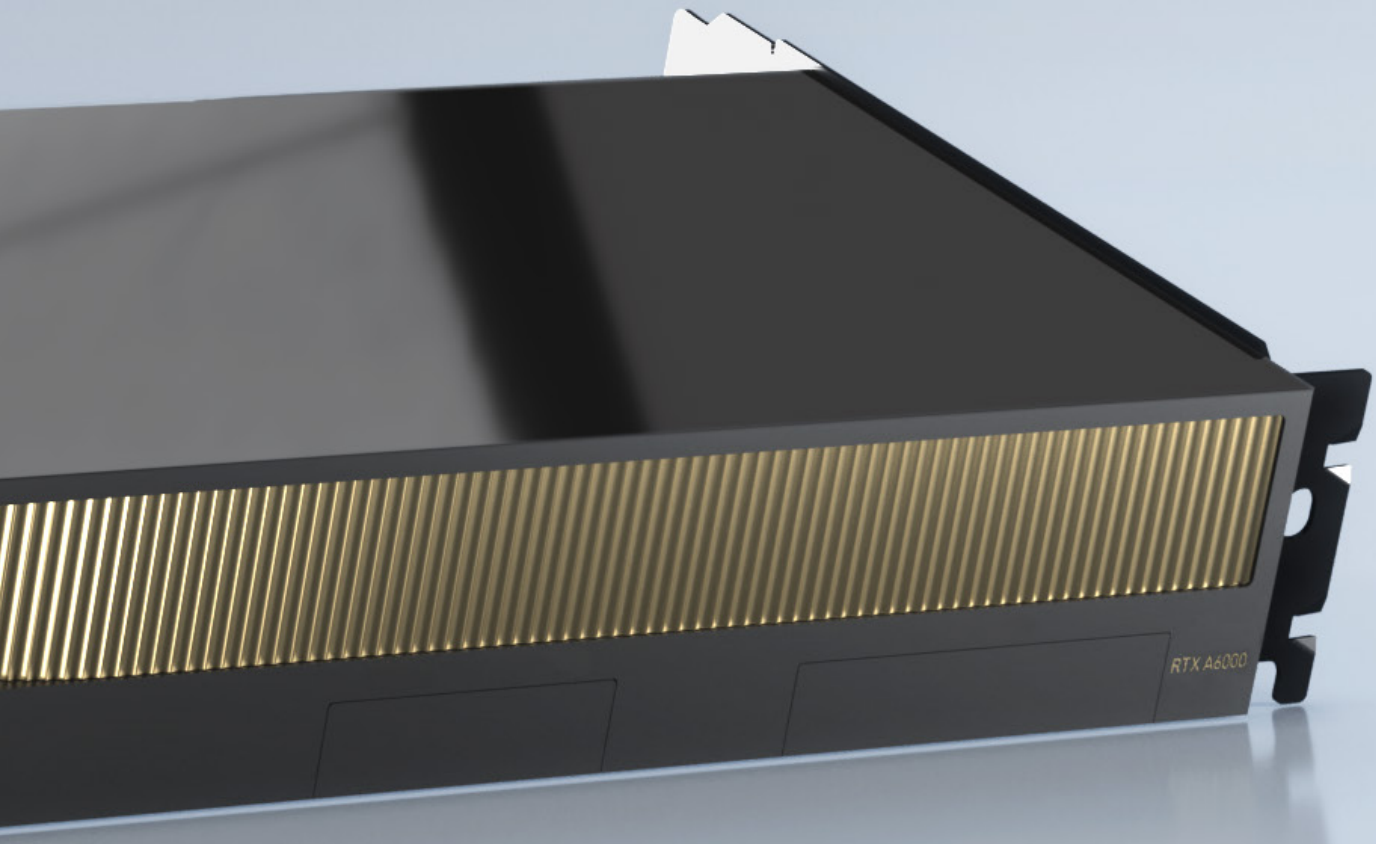
Ceux-ci peuvent s'exécuter sur des plates-formes de calcul uniquement équipées de processeurs ou accélérées par des processeurs graphiques, en fonction du cas d'utilisation.

Le Deep Learning, quant à lui, est l'une des techniques les plus efficaces de l'arsenal IA. Il fonctionne particulièrement bien sur les données non structurées, telles que les images, la vidéo ou la voix. Il est profondément sophistiqué : il repose sur des approches de réseaux de neurones artificiels, inspirées par la structure et l'activité des neurones au sein du cerveau humain.

Complexité et rôle des scientifiques des données

La montée en puissance de ces techniques d'IA et les différents défis qui les accompagnent font que les entreprises s'empressent d'embaucher des scientifiques des données. Ces professionnels, chargés d'acquérir et d'organiser les données pour découvrir leur valeur cachée, sont face à une tâche complexe.

Celle-ci couvre plusieurs disciplines et un large éventail de domaines, notamment l'analytique et l'apprentissage automatique. De plus, elle nécessite un matériel puissant, validé, mais également flexible, fourni avec des outils et des environnements logiciels prêts à l'emploi. Toutefois, cela est plus facile à dire qu'à faire, et génère deux principaux défis liés à la complexité.



Le premier défi concerne l'intégration.

L'installation et l'intégration manuelles de composants matériels et logiciels sont longues et fastidieuses. D'une part, les accélérateurs matériels (par exemple, les processeurs graphiques NVIDIA RTX) introduisent de nouvelles améliorations spécialisées à chaque génération. D'autre part, les outils logiciels sont constamment optimisés pour adopter les derniers algorithmes, cadres et bibliothèques. Il est donc difficile de rassembler ces nouveaux éléments, de garantir la compatibilité entre toutes les variables et de transformer facilement les scientifiques des données en administrateurs système, d'autant que cela n'est pas faire bon usage de leurs compétences.

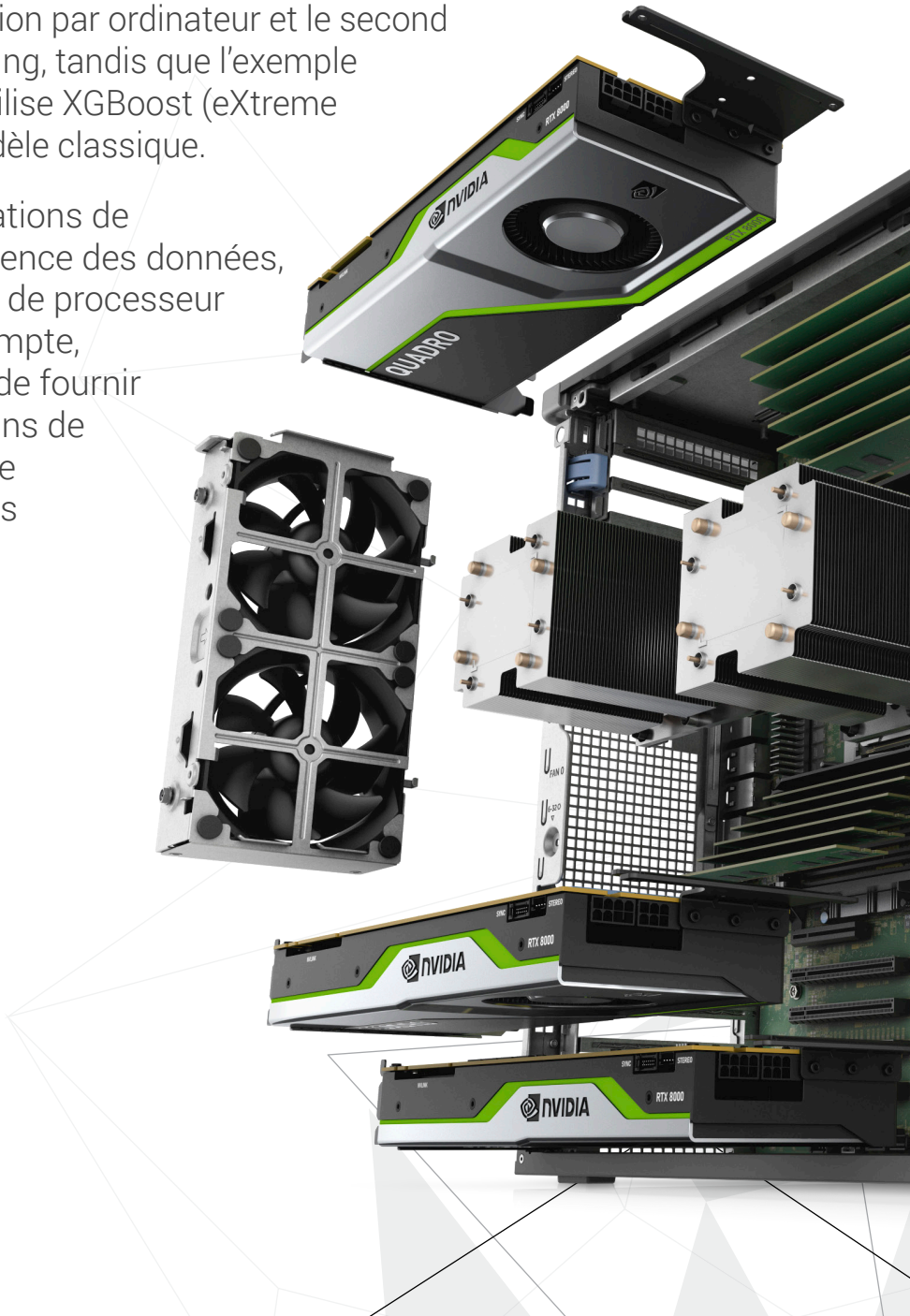
Le second défi concerne la configuration.

Le choix du processeur graphique à utiliser dépend de la charge applicative concernée. Par exemple, les exigences de la vision par ordinateur peuvent différer de celles des tâches de traitement du langage naturel (NLP), qui sont elles-mêmes différentes des exigences de l'apprentissage automatique classique. Les différents types de jeux de données impliqueront également différentes exigences de plates-formes de calcul. Il est donc difficile de choisir la plate-forme et le processeur graphique qui conviennent le mieux à différents cas d'utilisation.

Nos processus d'analyse comparative

Ce document examine trois cas d'utilisation : le Deep Learning (à travers deux instances) et l'apprentissage automatique. Le premier exemple de Deep Learning porte sur la classification d'images via la vision par ordinateur et le second va plus loin avec BERT Fine Tuning, tandis que l'exemple d'apprentissage automatique utilise XGBoost (eXtreme Gradient Boosting) dans un modèle classique.

Dans tous les cas, plusieurs stations de travail Dell Precision pour la science des données, avec différentes configurations de processeur graphique, ont été prises en compte, comparées et présentées, afin de fournir des conseils sur les accélérations de performances attendues pour le développement de modèles. Les détails de ces opérations sont fournis dans les sections pertinentes ci-dessous.



De nouvelles technologies pour relever les défis modernes

Les stations de travail Dell Precision pour la science des données (STSD) font partie d'une ligne de produits spécialement conçue pour relever le défi de l'intégration. Elles regroupent un ensemble d'équipements matériels de dernière génération optimisés par des processeurs graphiques NVIDIA, ainsi que la pile logicielle NVIDIA Data Science. Cette dernière fournit les derniers cadres et bibliothèques accélérés par des processeurs graphiques qui ont été validés à l'usine Dell.

Cela signifie que les scientifiques des données peuvent désormais se lancer dans une transition vers l'IA sans avoir à craindre de passer des heures ou des jours à déterminer quels équipements matériels et logiciels fonctionnent ensemble. Toutefois, le défi de la configuration reste à la charge des départements IT, des scientifiques des données et de leurs sociétés. Dell a établi ce point de référence pour vous aider à résoudre ce problème.

Facteurs à prendre en compte lors de la configuration pour la science des données

Afin de dimensionner et de configurer correctement une station de travail Dell pour la science des données en fonction des besoins d'un scientifique des données, il faut suivre trois grandes étapes : (1) déterminer la taille du jeu de données et le meilleur modèle IA ; (2) faire correspondre le processeur graphique et sa mémoire au jeu de données ; et (3) décider de la configuration appropriée pour le processeur et sa mémoire.

ÉTAPE 1

Consiste à dimensionner le jeu de données et à choisir l'approche de développement du modèle IA. La solution universelle n'existe pas. Les approches et les modèles varient largement dans le secteur.

Le dilemme est généralement le suivant :

- La mémoire du système est trop petite, donc les résultats sont déformés ou compromis.
- La mémoire du système est trop grande et coûteuse, ou ne tient pas physiquement dans un seul système.

Il est possible de diviser le jeu de données en petits morceaux, appelés mini-lots, pour s'adapter à la mémoire disponible du système.

ÉTAPE 2

Il s'agit de faire correspondre le processeur graphique le plus adapté et la meilleure taille de mémoire de processeur graphique à ce jeu de données. Lorsque l'utilisation du processeur graphique approche 100 % (tel qu'on peut le voir via l'utilitaire de ligne de commande « nvidia-smi »), cela indique qu'un processeur graphique avec davantage de mémoire ou qu'une configuration à deux/trois processeurs graphiques est requis(e).

ÉTAPE 3

Implique de décider de la configuration optimale pour le processeur et sa mémoire afin que le côté processeur puisse alimenter le côté processeur graphique du système. Enfin, il faut sélectionner la configuration et le dimensionnement du stockage.

Prenons l'exemple d'un point de référence BERT (Bidirectional Encoder Representations). Nous avons remarqué au cours de nos points de référence BERT que les configurations RTX 6000 surpassent les configurations RTX 8000 lors de l'utilisation de longueurs de séquences et de tailles de lots inférieures.

Cependant, une fois que la longueur de séquence et la taille de lot augmentent à un certain seuil, la carte RTX 6000 n'a pas suffisamment de mémoire pour exécuter la tâche.

Le jeu de données

En fonction des exigences de taille du jeu de données, ce dernier peut tenir entièrement dans la mémoire utilisée lors de l'entraînement du modèle. Dans un système uniquement équipé d'un ou de plusieurs processeurs, le jeu de données est placé dans la mémoire DDR du ou des processeurs et le modèle est entraîné à l'aide de ces derniers.

Dans un système comme la station de travail pour la science des données, contenant des processeurs graphiques, le modèle est entraîné à l'aide des ressources de processeur graphique, et le jeu de données réside dans la mémoire du ou des processeurs graphiques. De nombreux modèles et jeux de données peuvent tenir dans les mémoires des cartes de processeur graphique modernes.

Dans certains cas, lorsque la taille du jeu de données dépasse la mémoire d'un processeur graphique, des astuces telles que l'utilisation de bibliothèques distribuées entre plusieurs processeurs graphiques (par exemple, DASK) peuvent permettre de charger le jeu de données sur les différents processeurs graphiques.

Sélection et dimensionnement du processeur graphique qui accélère le mieux la modélisation IA

Remarque : reportez-vous au « Tableau 2 : Processeurs graphiques utilisés pour l'analyse comparative » pour connaître la mémoire tampon de chacun des processeurs graphiques NVIDIA disponibles au moment de la publication.

Le RTX A6000 dispose de 48 Go de mémoire GDDR6 à 768 Go/s, de 10 752 cœurs CUDA de classe Ampere, de 84 cœurs RT de classe Ampere et de 336 cœurs Tensor de classe Ampere. La combinaison de deux processeurs graphiques RTX A6000 à l'aide de NVLink double le nombre de cœurs et la quantité mémoire, en combinant les deux cartes de mémoire de processeur graphique dans 96 Go de mémoire unifiée disponible pour le jeu de données.

Le RTX 8000 dispose de 48 Go de mémoire GDDR6 à 672 Go/s, de 4 608 cœurs CUDA, de 72 cœurs RT et de 576 cœurs Tensor. La combinaison de deux processeurs graphiques RTX 8000 à l'aide de NVLink double le nombre de cœurs et la quantité mémoire, en combinant les deux cartes de mémoire de processeur graphique dans 96 Go de mémoire unifiée disponible pour le jeu de données.

Le processeur graphique RTX 6000 possède le même nombre de cœurs que le RTX 8000, 24 Go de mémoire graphique GDDR6 à 672 Go/s, et peut également utiliser NVLink pour se connecter à un second RTX 6000 afin d'offrir une mémoire unifiée combinée de 48 Go.

Le RTX 5000 dispose de 3 072 cœurs CUDA, de 48 cœurs RT, de 384 cœurs Tensor et de 16 Go de mémoire GDDR6 à 448 Go/s.

Pourquoi différentes configurations et charges applicatives ont-elles été utilisées pour ce point de référence ?

Dell a choisi et a exécuté divers points de référence pour rapporter les différences de performances entre plusieurs plates-formes et configurations de stations de travail Precision pour la science des données. Les plates-formes pour la science des données vont des stations de travail mobiles Dell, à savoir les modèles mobiles 15" 7550 et 17" 7750, aux stations de travail fixes Dell, c'est-à-dire les tours Dell 5820 et 7920, jusqu'à la station de travail Dell 7920 au format rack.

Au sein de ces plates-formes, nous avons sélectionné différentes configurations de mémoire de processeur classiques et de processeur graphique, dans le but de refléter un éventail de ressources de calcul adaptées aux charges applicatives d'entraînement des modèles IA. Les processeurs graphiques NVIDIA comparés étaient les processeurs graphiques mobiles Quadro RTX 5000 dans les stations de travail mobiles pour la science des données, et les processeurs graphiques Quadro RTX 6000 et RTX 8000 dans les stations de travail tours et rack pour la science des données.

Toutes ces plates-formes et configurations ont été équipées du même système d'exploitation, à savoir Ubuntu Linux, et de la pile logicielle NVIDIA Data Science, qui est l'offre logicielle groupée validée et expédiée avec chaque station de travail Dell pour la science des données.

La pile logicielle NVIDIA Data Science inclut les logiciels TensorFlow, Python, XGBoost et Jupyter Notebook optimisés en usine fournis à chaque utilisateur ; ceux-ci ayant été comparés avec le matériel sous-jacent. Les comparatifs entre les différentes plates-formes, configurations et informations de processeur graphique sont présentés dans les tableaux fournis sur cette page :

Modèle	Precision 7550 mobile	Precision 7750 mobile	Precision 5820 Fixe	Precision 5820 Fixe	Precision 7920 Fixe	Precision 7920 rack
Processeur	Intel® Xeon W-10885M (8 cœurs, 16 Mo de mémoire cache, 2,4 GHz à 5,3 GHz, 45 W, vPro)	Intel® Xeon W-10885M (8 cœurs, 16 Mo de mémoire cache, 2,4 GHz à 5,3 GHz, 45 W, vPro)	Intel® Xeon W-2175 (14 cœurs, 19,25 Mo, 2,5 GHz à 4,3 GHz, 140 W, vPro)	Intel® Xeon W-2245 (8 cœurs, 11 Mo, 3,9 GHz à 4,7 GHz, 155 W, vPro)	Double processeur Intel® Xeon Gold 6134 (8 cœurs, 24,75 Mo, 3,2 GHz à 3,7 GHz, 130 W, vPro)	Double processeur Intel® Xeon Gold 6244 (8 cœurs, 24,75 Mo, 3,6 GHz à 4,4 GHz, 150 W, vPro)
Carte graphique	NVIDIA® Quadro RTX 5000 avec 16 Go de mémoire GDDR6	NVIDIA® Quadro RTX 5000 avec 16 Go de mémoire GDDR6	Simple/ double carte NVIDIA® Quadro RTX 6000 avec 24 Go de mémoire GDDR6	Simple carte NVIDIA® Quadro RTX 8000 avec 48 Go de mémoire GDDR6	Double/triple carte NVIDIA® Quadro RTX 6000 avec 24 Go de mémoire GDDR6	Double/triple carte NVIDIA® Quadro RTX 6000 avec 24 Go de mémoire GDDR6
Mémoire	64 Go (4 x 16 Go) de mémoire DDR4 non ECC à 2 933 MHz	64 Go (4 x 16 Go) de mémoire DDR4 non ECC à 2 933 MHz	128 Go (8 x 16 Go) de mémoire DDR4 RDIMM ECC à 2 666 MHz	256 Go (4 x 64 Go) de mémoire DDR4 RDIMM ECC à 2 933 MHz	128 Go (8 x 16 Go) de mémoire DDR4 RDIMM ECC à 2 666 MHz	128 Go (8 x 16 Go) de mémoire DDR4 RDIMM ECC à 2 666 MHz
Stockage	2 disques SSD PCIe NVMe M.2 de classe 40 1 To	2 disques SSD PCIe NVMe M.2 de classe 50 2 To	1 disque SSD PCIe NVMe M.2 de classe 40 1 To	1 disque SSD PCIe NVMe M.2 de classe 50 1 To	1 disque SSD PCIe NVMe M.2 de classe 40 1 To	1 disque SSD SATA 2,5" de classe 20 1 To + 1 disque SSD SATA 2,5" de classe 20 512 Go
Système d'exploitation	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

Tableau 1 : Configurations STSD testées

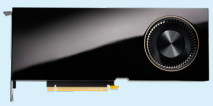
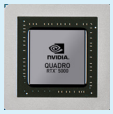
	Processeurs graphiques de bureau			Processeurs graphiques mobiles
	RTX A6000	RTX 8000	RTX 6000	RTX 5000
				
Cœurs CUDA	10 752 Ampere	4 608	4 608	3 072
Cœurs RT	84 Ampere	72	72	48
Cœurs Tensor	336 Ampere*	576	576	384
Mémoire	48 Go GDDR6 à 768 Go/s	48 Go GDDR6 jusqu'à 672 Go/s	24 Go GDDR6 jusqu'à 672 Go/s	16 Go GDDR6 jusqu'à 448 Go/s

Tableau 2 : Processeurs graphiques utilisés pour l'analyse comparative

Pour les conclusions de point de référence du processeur graphique RTX A6000, reportez-vous à l'Annexe

Deep Learning : Classification d'images

La première analyse comparative a été effectuée à l'aide du tf_cnn_benchmark de TensorFlow. Il est basé sur le réseau neuronal convolutif (CNN) qui est considéré comme la base des tâches de vision par ordinateur, telles que les classifications d'images, la détection d'objets, les segmentations d'images, les réseaux adverses génératifs et d'autres. Les réseaux résiduels (ResNets), qui sont l'une des nombreuses topologies CNN, se composent de blocs résiduels.

Chaque bloc possède deux branches : l'une qui alimente l'entrée directement vers la sortie et l'autre qui effectue deux à trois convolutions. Les deux branches sont ajoutées et servies au bloc suivant. Cette approche génère un CNN hautes performances puisque nous empilons des couches ensemble, tout en évitant le problème de disparition des gradients. Par conséquent, ResNet50 (composé de 50 couches) est considéré comme la norme pour l'analyse comparative de la classification d'images.

Nous avons utilisé tf_cnn_benchmark avec ResNet50 comme topologie pour CNN. Le référentiel tf_cnn_benchmark contient des scripts qui exécutent l'entraînement et l'inférence de modèles de classification d'images standard sur des images synthétiques, ainsi que d'autres jeux de données publics tels qu'ImageNet.

Dans ce cas, nous avons utilisé le jeu de données par défaut, qui est un jeu de données synthétique. Ce point de référence est également capable de s'exécuter sur un ou plusieurs processeurs graphiques au sein d'un nœud de station de travail ou sur plusieurs nœuds de station de travail. Pour plus d'informations sur tf_cnn_benchmark, reportez-vous à GitHub.

Une autre pratique courante des professionnels du Deep Learning est l'utilisation de la fonctionnalité AMP (Automatic Mixed Precision). Celle-ci implique l'utilisation d'opérations mathématiques à simple précision (FP32) et demi-précision (FP16) pour accélérer l'entraînement du réseau neuronal sans perte de précision.

Les derniers processeurs graphiques NVIDIA Quadro, en commençant par l'architecture Volta et en continuant avec les processeurs graphiques les plus récents, incluent des cœurs Tensor spécialisés pour utiliser l'AMP. Dans notre cas, nous avons utilisé FP16 lorsque nous avons exécuté ce point de référence.

La dernière variable que nous avons utilisée est la taille de lot (BS). La taille de lot et le taux d'apprentissage sont sans doute les deux hyperparamètres les plus importants lors de l'exécution d'expériences différentes une fois le modèle de réseau neuronal déterminé. En outre, ces deux paramètres sont souvent étroitement liés. La flexibilité permettant d'exécuter différentes tailles de lots donne aux scientifiques des données la possibilité d'explorer davantage de fonctionnalités et, dans certains cas, réduit le temps d'entraînement.

Le tf_cnn_benchmark utilisé ici est destiné à l'entraînement de modèles et a été exécuté pour un nombre défini d'itérations, tandis que la vitesse moyenne a été mesurée en images/s. Cet exemple de script concerne un double processeur graphique, avec une taille de lot de 32 et une précision FP16 :

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

Dans cette optique, nous avons testé différentes tailles de lots sur différents processeurs graphiques, du Quadro Mobile RTX 5000 16 Go aux simple, double et triple Quadro RTX 6000 24 Go, installés dans des tours Precision Performance 5820 et 7920. Les deux dernières sont des tours Precision hautes performances respectivement équipées de plates-formes de processeurs Intel Xeon Cascade Lake à un et deux sockets.

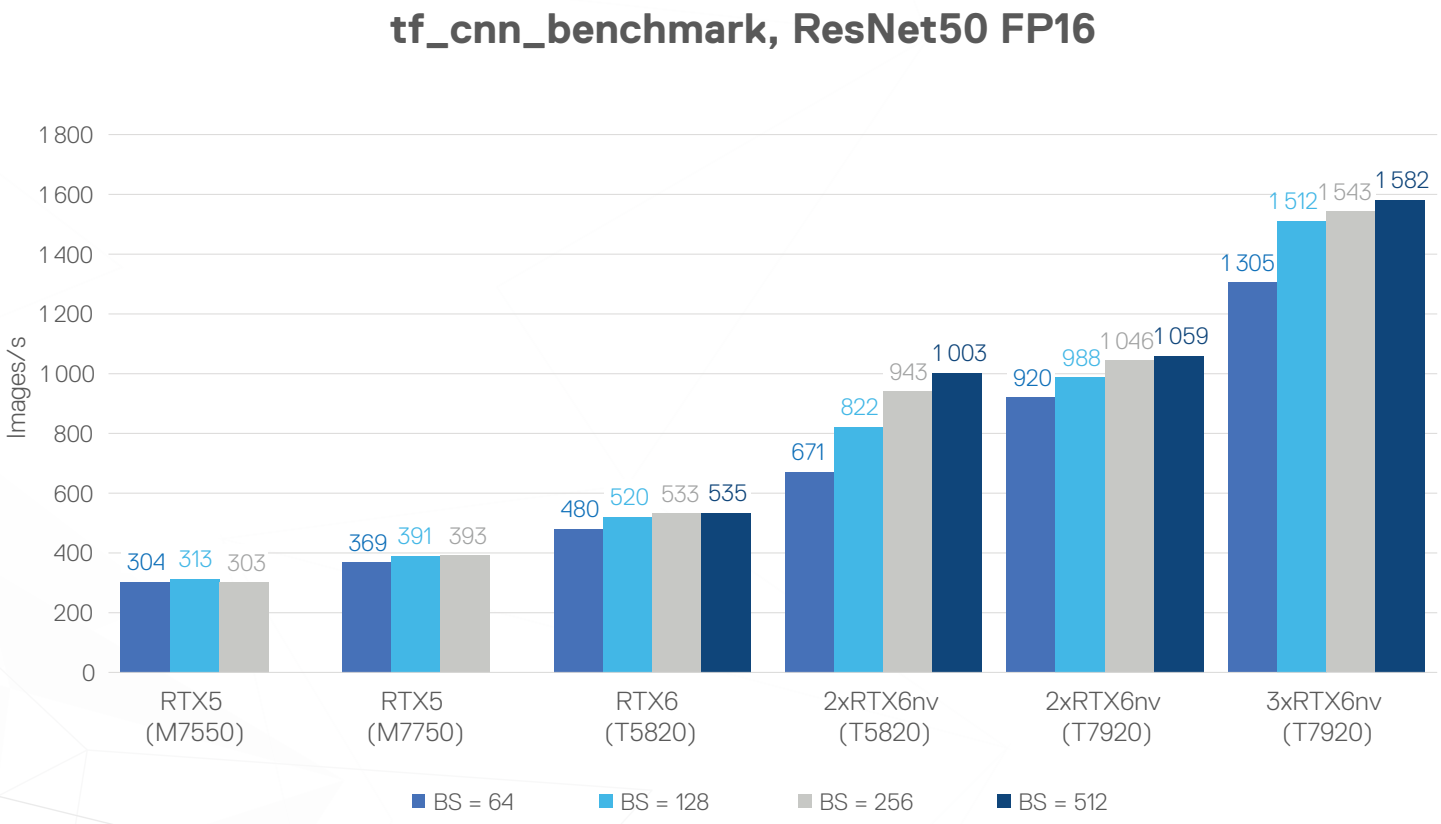


Figure 1

Deep Learning : Classification d'images (suite)

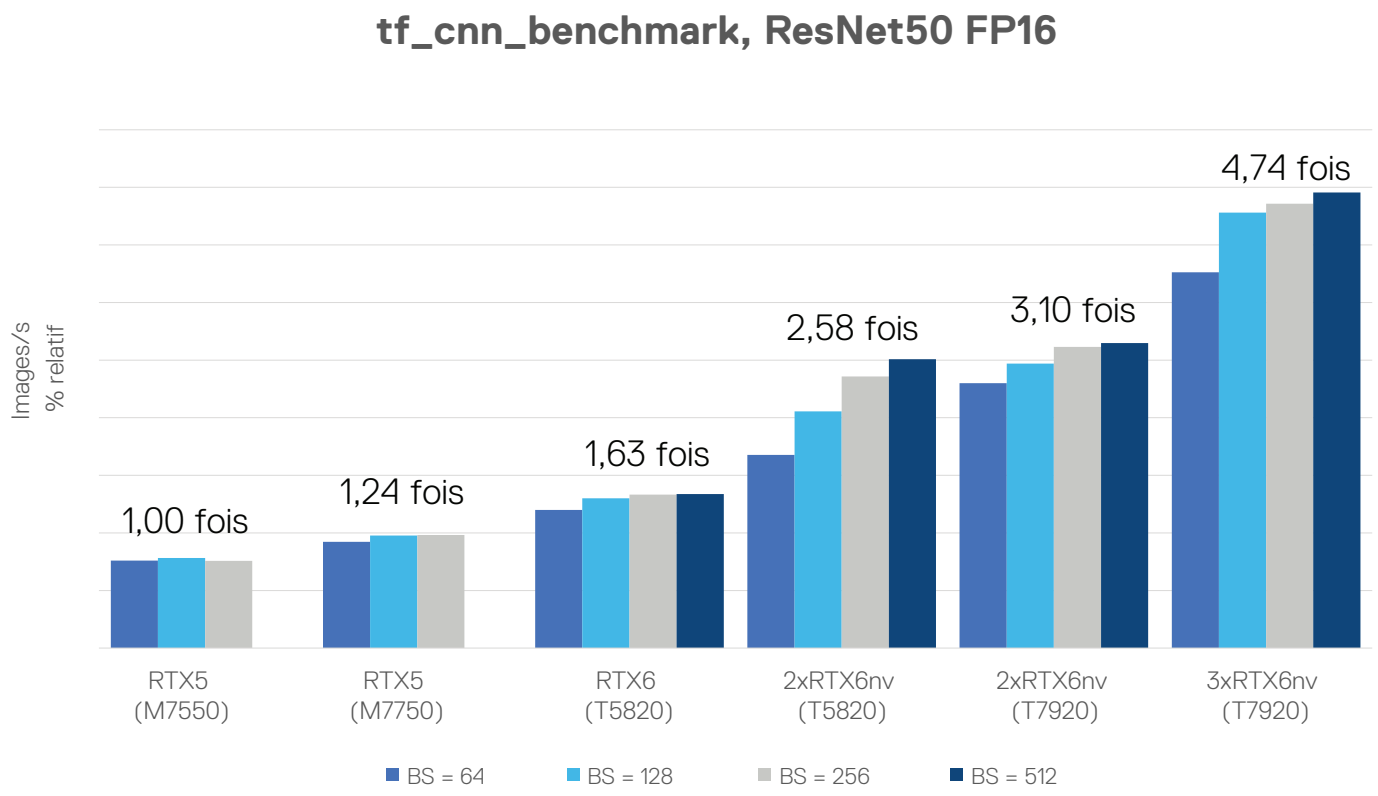


Figure 2

Le graphique montre que les stations de travail mobiles Precision 7550 et 7750 sont plus que capables d'effectuer ces tâches avec un nombre respectable d'images par seconde. Il est important de noter que le passage de la station de travail mobile 15" 7550 à la station de travail mobile 17" 7750 implique une amélioration des performances d'environ 24 %, alors que les deux plates-formes utilisent le même Quadro RTX 5000.

En effet, le boîtier plus grand de la 7750 offre une meilleure solution thermique et de refroidissement, ce qui permet au RTX 5000 de fonctionner à des vitesses d'horloge plus élevées (c'est-à-dire, une utilisation plus élevée).

En passant à la tour 5820 avec un seul RTX 6000, nous avons constaté une amélioration de 63 % du débit, en raison d'un plus grand nombre de cœurs CUDA, de cœurs Tensor et d'une quantité supérieure de mémoire vidéo.

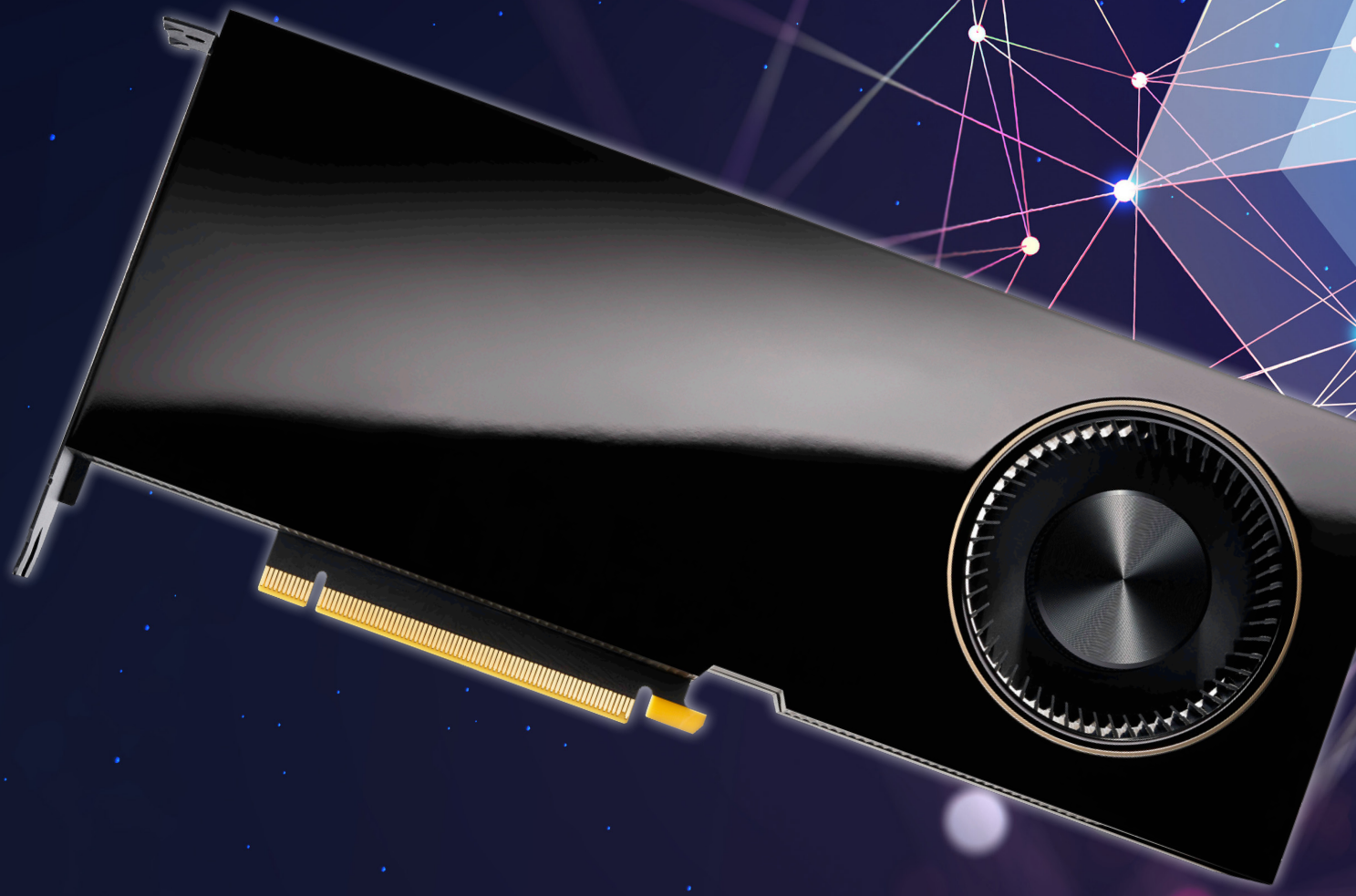
Le double RTX 6000 avec NVLink double pratiquement le débit du RTX 6000 seul, en particulier sur la tour 7920 (3,1 fois contre 1,63 fois).

En plus de disposer d'un second processeur graphique, la tour 7920 offre également un second microprocesseur multicœur, fournissant davantage de cœurs et de canaux de mémoire, nécessaires pour éviter de bloquer ces deux processeurs graphiques hautes performances.

L'introduction d'un troisième processeur graphique offre un coup de pouce supplémentaire présentant une évolution presque linéaire. À l'aide de la commande nvidia-smi, nous avons pu confirmer que lors de l'exécution de ces analyses comparatives et d'autres dans les domaines du traitement du langage naturel (NLP), tous les processeurs graphiques (double et triple) sont toujours entièrement utilisés. En d'autres termes, les processeurs graphiques ont systématiquement été alimentés avec les données par le processeur, la mémoire du processeur et les systèmes de stockage. Aucune performance ni aucun débit n'est resté inutilisé. Aucun goulot d'étranglement de l'architecture système n'est apparu.

Il est également apparu clairement que non seulement les processeurs graphiques Quadro RTX des tours surpassent leurs homologues mobiles, mais qu'ils ont en plus permis de traiter des tailles de lots plus importantes (p. ex., BS = 512). En effet, le RTX 6000 offre une mémoire vidéo plus importante et un plus grand nombre de cœurs CUDA et de cœurs Tensor, comme indiqué précédemment.

Le double processeur graphique avec NVLink a également fourni des performances de débit plus élevées, en raison de la communication directe entre les processeurs graphiques, ce qui évite le bus PCIe comparativement lent.



Deep Learning : Traitement du langage naturel

La deuxième analyse comparative a été effectuée à l'aide du point de référence BERT Fine Tuning NLP de TensorFlow. Le traitement du langage naturel (NLP) utilise l'IA pour permettre à un ordinateur de comprendre, d'analyser, de manipuler et générer le langage humain.

BERT (Bidirectional Encoder Representations from Transformers) est un puissant outil qui a été développé par Google fin 2018 et qui permet aux ordinateurs de traiter, d'analyser et de « comprendre » le langage humain. Il s'est imposé dans différentes applications NLP, telles que les systèmes de questions-réponses, la reconnaissance d'entités nommées, l'inférence du langage naturel et la classification de textes. Auparavant, tous les modèles de langue (par exemple, Skip-gram et CBOW) étaient unidirectionnels. Ils ne pouvaient parcourir la fenêtre contextuelle du mot que de gauche à droite ou de droite à gauche. BERT utilise la modélisation bidirectionnelle du langage pour comprendre le contexte d'un mot, c'est-à-dire que le modèle apprend le contexte d'un mot en fonction de tout son environnement.

Pré-entraînement BERT

La procédure de pré-entraînement suit largement la documentation existante sur le pré-entraînement du modèle de langue. Pour le corpus de pré-entraînement, BERT utilise le BooksCorpus (800 millions de mots) et Wikipédia en anglais (2 500 millions de mots). Comme vous pouvez l'imaginer, tout cela aurait pris plusieurs semaines avec les stations de travail. Nous nous sommes donc concentrés sur les points de référence BERT Fine Tuning.

BERT Fine-Tuning

Le modèle pré-entraîné est utilisé comme base (apprentissage par transfert) avec les mêmes poids, tout en ajoutant quelques couches spécifiques à la tâche NLP concernée. Dans notre cas, il s'agissait de questions/réponses. Cette approche est couramment utilisée pour créer de nouvelles tâches NLP spécifiques et réduire considérablement la complexité du réglage.

Nous avons utilisé des scripts de point de référence finetune_train_benchmar.sh du NVIDIA NGC Repository BERT pour TensorFlow. À l'aide du script, nous avons pu tester sur le jeu de données SQuAD v1.1 à l'aide d'une précision fp16 ou fp32.

Des longueurs de séquences de 128 et 384, et des tailles de lots de 1, 2, 4, 8, 16, 32 et 64 ont été exécutées au cours de ce script. Afin d'exécuter toutes les tailles de lots, nous avons modifié le fichier finetune_train_benchmark.sh sur la ligne 81 et ajouté les tailles de lots 8, 16, 32 et 64 à la ligne suivante.

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

La configuration de l'entraînement BERT Fine Tuning offre les options Base ou Large. BERT LARGE (L=24, H=1 024, A=16, Nbre total de paramètres=340 millions) et BERT BASE (L=12, H=768, A=12, Nbre total de paramètres=110 millions). Nous avons choisi BERT Large pour nos points de référence et avons utilisé la commande suivante :
BERT# scripts/finetune_train_benchmark large true <num_gpus> squad.

Configurations des stations de travail Precision

System	Modèle Precision	Processeur	Fréquence	cœurs	Mémoire	Processeur graphique
Config. 1	T5820	W-2245	3,9 à 4,7 GHz	8 cœurs	256 Go	1 à 2 RTX 8000
Config. 2	T5820	W-2245	3,9 à 4,7 GHz	8 cœurs	256 Go	1 à 2 RTX 6000
Config. 3	T7920	5217	3 à 3,7 GHz	8 cœurs	196 Go	3 RTX 8000
Config. 4	T7920	5217	3 à 3,7 GHz	8 cœurs	196 Go	3 RTX 6000
Config. 5	T5820	W-2175	2,5 à 3,4 GHz	14 cœurs	64 Go	2 GV100

Tableau 3

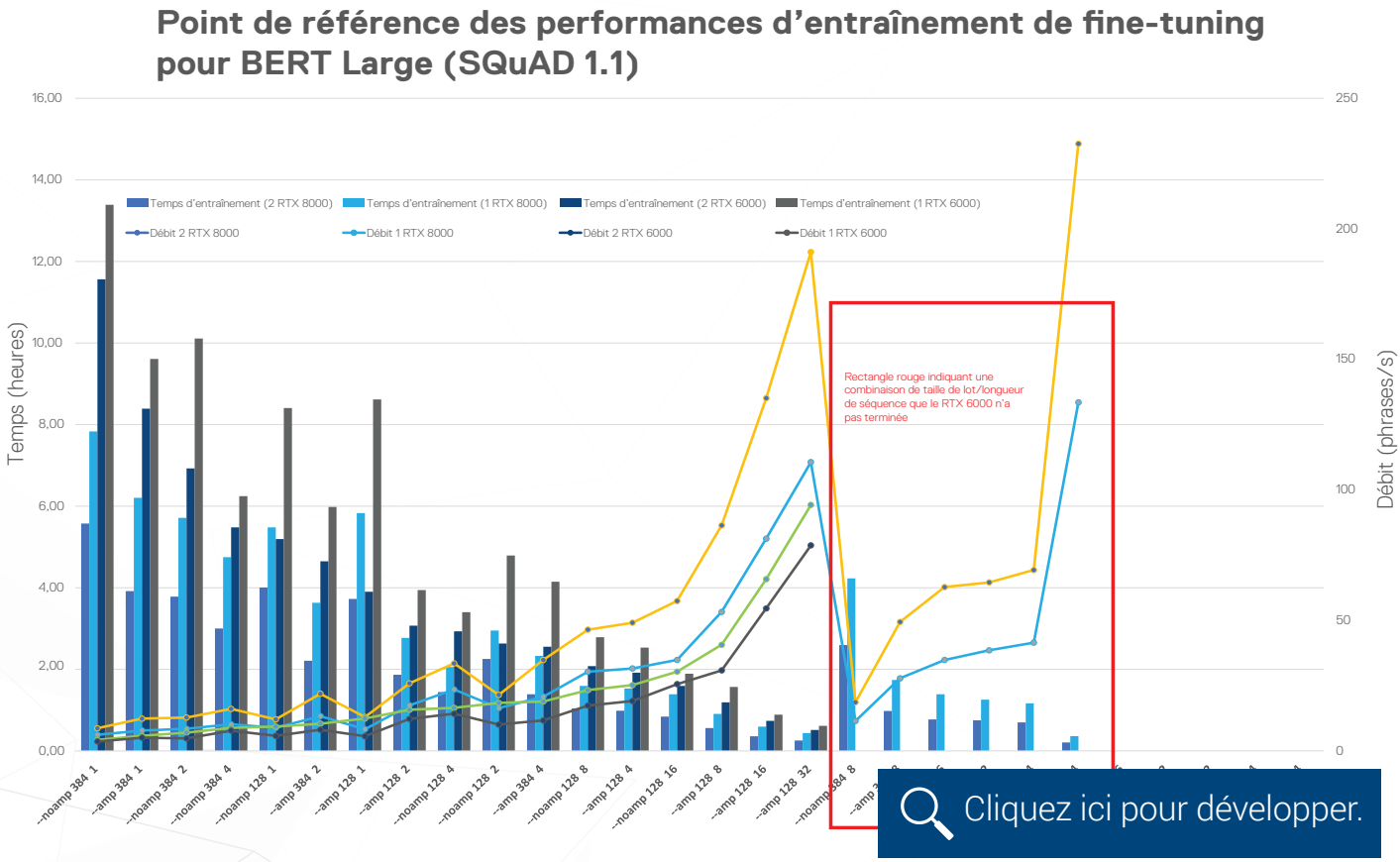


Figure 3

Deep Learning : Traitement du langage naturel (suite)

Différentes exécutions sont affichées sur l'axe horizontal où :

--amp seq_len BS signifie FP16, longueur de séquence et BS = taille de lot, et

--nonamp seq_len BS signifie FP32, longueur de séquence et BS = taille de lot

Sur la figure 3, classée par débit, les 2 RTX 6000 (VERT) ont surpassé le RTX 6000 seul (GRIS) en termes de temps d'entraînement et de débit (phrases/s). Cela s'est amplifié à mesure que la longueur de séquence et la taille de lot augmentaient.

Exemple 1 :

FP16, longueur de séquence 128, taille de lot 1.

1 RTX 6000 Phrases par seconde 8,82

2 RTX 6000 Phrases par seconde 13,04

Résultat : 2 RTX 6000 ont fait augmenté les performances de 47 %.

Exemple 2 :

FP16, longueur de séquence 384, taille de lot 2.

1 RTX 6000 Phrases par seconde 13,77

2 RTX 6000 Phrases par seconde 22,08

Résultat : 2 RTX 6000 ont fait augmenté les performances de 60 %.

Notez les exemples 1 et 2 ci-dessous. Nous avons également remarqué que le RTX 6000 exécute les combinaisons FP32 128 32, mais qu'il n'a pas complètement exécuté les FP16/32 64. Il a également été observé que le RTX 6000 n'a pas complètement exécuté les FP16 384 8 ou de taille de lot supérieure. Ces combinaisons sont affichées dans le rectangle rouge du graphique ci-dessus. Les RTX 8000 ont pu exécuter toutes les combinaisons de longueur de séquence et de taille de lot jusqu'à FP32 384 8 (exemple 3).

Exemple 3 :

FP16, longueur de séquence 384, taille de lot 8.

2 RTX 6000 Phrases par seconde : incomplet

2 RTX 8000 Phrases par seconde 49,38

Résultat : 2 RTX 6000 : incomplet

Deep Learning : Traitement du langage naturel (suite)

Performances d'entraînement	Longueur de séquence	Taille de lot	Test 1 RTX 8000 (ECC)		Test 2 RTX 6000 (ECC)	
			T5820/W-2245/256 Go		T5820/W-2245/256 Go	
			2 GPU	1 GPU	2 GPU	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				

Synthèse

Le choix d'un processeur graphique (GPU) pour les tâches NLP doit être déterminé en fonct
lots. Les RTX 8000 exécutent des combinaisons de virgules flottantes, de longueurs de séc
longueurs de séquences et des tailles de lots pendant le fine-tuning, l'utilisation de deux RT

Dans le tableau 4, plusieurs zones incomplètes sont représentées par des cellules ombrées en bleu. La majorité d'entre elles sont liées aux systèmes RTX 6000. Les triples RTX 8000/RTX 6000 sur les stations 7920 ont été plus lents que les doubles RTX 8000/RTX 6000 sur certaines des exécutions.

Lorsque les cartes graphiques étaient identiques, les systèmes dotés d'une meilleure fréquence de base du processeur et d'une mémoire supérieure ont surpassé ceux disposant d'une fréquence de base de processeur et d'une mémoire inférieures.

Test 3 RTX 6000		Test 4 RTX 8000 (ECC)			Test 5 RTX 6000 (ECC)		Test 6 GV100 (ECC)	
T5820/W-2245/256 Go		T7920/5217/196 Go					T5820/W-2175/64 Go	
2 GPU	1 GPU	3 GPU	2 GPU	1 GPU	2 GPU	1GPU	2 GPU	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				

Tableau 4

tion de l'importance de la virgule flottante, de la longueur de séquence et des tailles de quences et de tailles de lots supérieures. Même si un RTX 8000 exécute bon nombre des X 8000 apparaît comme la solution performante.

Apprentissage automatique : Classification

La troisième analyse comparative a été effectuée à l'aide de la bibliothèque XGBoost. XGBoost (eXtreme Gradient Boosting) est la dernière évolution des algorithmes basés sur des arbres de décision. Il s'appuie sur un ensemble de méthodes d'arborescence qui appliquent le principe visant à stimuler les apprenants faibles. Pour les données tabulaires et structurées, il est considéré comme l'une des meilleures techniques de sa catégorie.

Pour cet exercice, nous avons utilisé un jeu de données numériques synthétiques de 6 millions de lignes x 501 colonnes (dont l'une correspondait à la sortie) ; toutes en précision complète (FP32). Le véritable notebook Python Jupyter a été utilisé à partir du notebook de démo disponible sur Rapids.ai. RAPIDS est la suite Open Source de bibliothèques d'apprentissage automatique accélérées par processeur graphique de NVIDIA, qui inclut XGBoost.

La station de travail Dell pour la science des données est préconfigurée avec les bibliothèques de science des données les plus courantes, y compris RAPIDS et XGBoost. Aucune installation logicielle supplémentaire n'est donc nécessaire. De plus, la bibliothèque XGBoost est désormais accélérée par processeur graphique.

Le test a été effectué en comparant le fonctionnement avec processeur uniquement au fonctionnement avec différents processeurs graphiques, comme avant, mais cette fois-ci, nous avons inclus les processeurs graphiques RTX 5000 mobiles, ainsi que les modèles RTX 6000, RTX 8000 et RTX GV100 destinés aux tours.

Le test prend uniquement en compte le temps nécessaire pour entraîner le jeu de données susmentionné à l'aide de XGBoost avec les mêmes paramètres.

Nous avons augmenté le paramètre num_round à 100, défini le rapport entraînement/validation sur 90 %/10 % et laissé le reste des paramètres sur les valeurs par défaut du notebook de démonstration.

Le graphique ci-dessous montre que le temps d'entraînement est considérablement réduit lors de l'utilisation d'un processeur graphique par rapport à un simple processeur. En déplaçant simplement l'entraînement d'un processeur Xeon W-10885M vers le processeur graphique RTX 5000, l'exécution a pris 50 % de temps en moins.

Avec un processeur graphique HBM2 32 Go plus élevé (RTX GV100 T5820), la vitesse était plus importante encore (6,4 fois supérieure).

Ainsi, des entraînements qui prenaient auparavant une semaine peuvent peut-être être exécutés en une journée. Les modèles RTX 6000, RTX 8000 et RTX GV100 bénéficient considérablement non seulement de leur grande quantité de mémoire vidéo, ce qui permet à l'ensemble du jeu de données d'y tenir, mais également de leur plus grand nombre de cœurs CUDA et Tensor, ce qui a un impact direct sur le temps de traitement et de convergence.

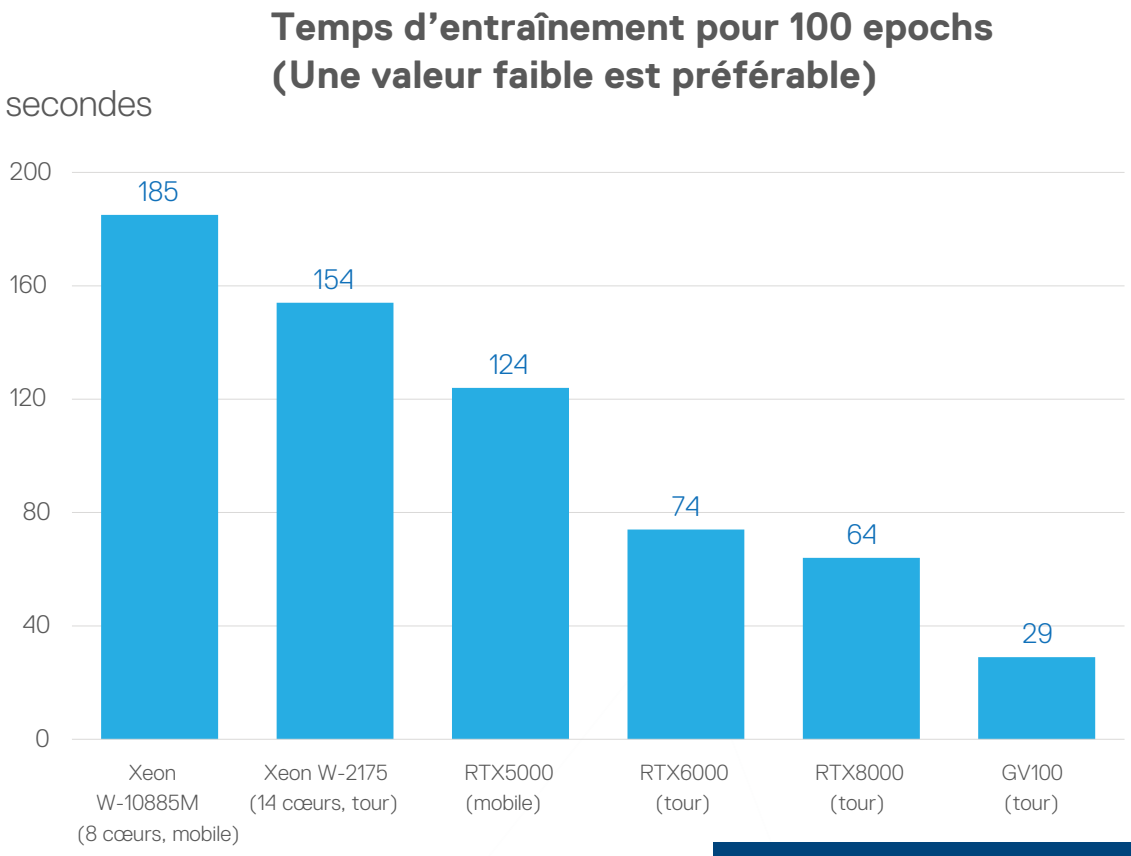


Figure 4

Cliquez ici pour développer.

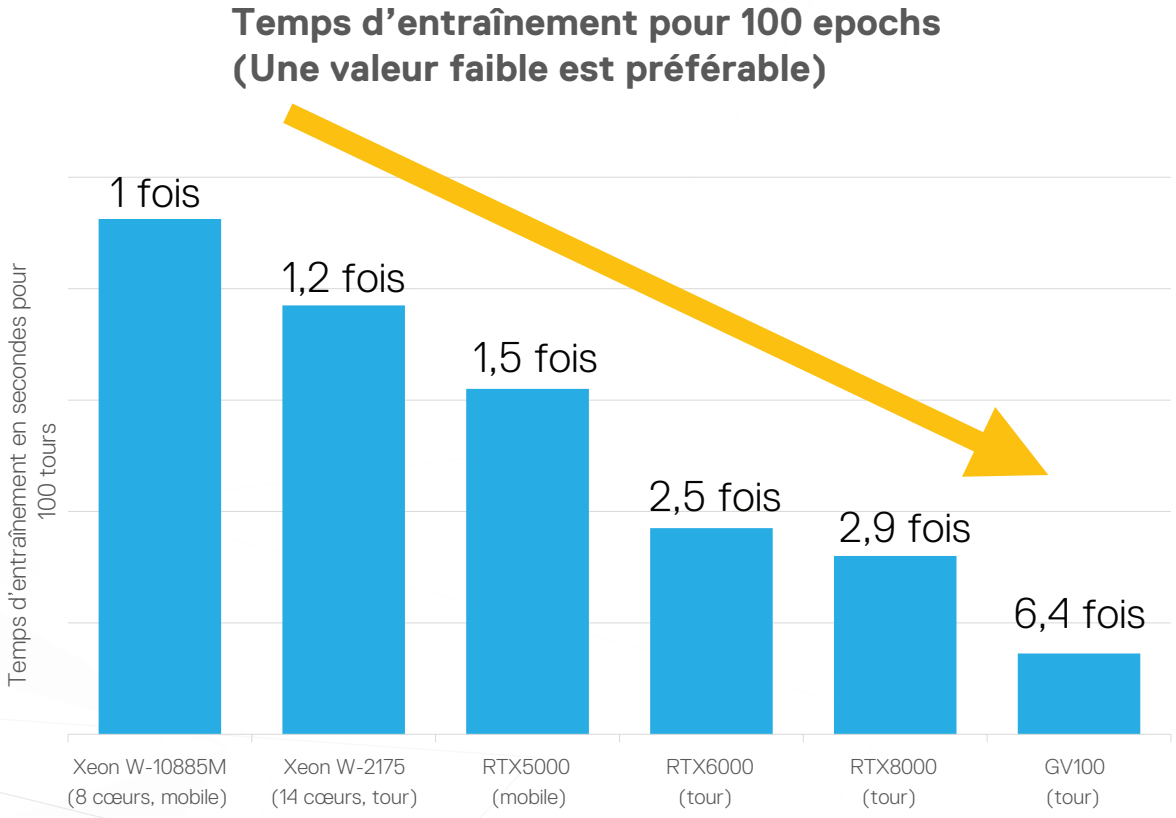


Figure 5

Cliquez ici pour développer.

Comment dimensionner correctement une STSD

Il faut suivre trois grandes étapes pour dimensionner et configurer correctement des stations de travail accélérées par processeur graphique afin de les utiliser dans le développement de modèles IA :

1. Déterminez le type et la taille de la mémoire du modèle IA à utiliser et évaluez la taille du jeu de données sur lequel ce modèle sera utilisé.
2. Dimensionnez le processeur graphique et sa mémoire.
3. Dimensionnez le processeur, sa mémoire et le stockage de masse.

ÉTAPE 1

Détermination de la taille du jeu de données et de l'approche de développement du modèle IA

La solution universelle n'existe pas dans le secteur de l'IA et de la science des données, car les jeux de données et les modèles varient énormément. Les approches classiques de modélisation de l'IA fonctionnent avec des jeux de données de taille relativement réduite. Toutefois, bon nombre des modèles d'apprentissage automatique et de Deep Learning sont sélectionnés pour résoudre des problèmes qui concernent des jeux de données très volumineux et non structurés.

Ils peuvent être gigantesques si les données analysées sont des images ou des vidéos. Que l'approche de modélisation à utiliser soit l'IA classique, l'apprentissage automatique ou le Deep Learning, la quantité de mémoire idéalement nécessaire pour héberger l'intégralité du jeu de données et le modèle dépasse largement la quantité de RAM pouvant être physiquement placée dans la plate-forme. Par conséquent, dans de nombreuses situations, le jeu de données doit être divisé en lots de taille inférieure. La plupart des scientifiques des données choisissent d'optimiser la taille des lots en fonction de la mémoire physique disponible.

Dans une station de travail IA accélérée par processeur graphique, ce sont le processeur graphique et sa mémoire qui hébergent et exécutent l'entraînement du modèle, pas le processeur et sa mémoire. Par conséquent, après avoir déterminé l'approche de modélisation et évalué la taille du jeu de données utilisé, l'étape suivante consiste à dimensionner le processeur graphique et sa mémoire.

ÉTAPE 2

Dimensionnement du processeur graphique et de sa mémoire

La mémoire DDR du processeur graphique, utilisée dans les mémoires tampons vidéo du processeur graphique, est plus performante et plus rapide que celle utilisée dans le processeur. En effet, la tâche de la mémoire DDR du processeur graphique est de suivre l'appétit vorace des cœurs de calcul du traitement parallèle dans le processeur graphique.

- **Stations de travail mobiles Dell Precision 7550 et 7750 pour la science des données**
  Elles utilisent le processeur graphique RTX 5000 de NVIDIA qui contient 3 072 cœurs CUDA, 48 cœurs RT, 384 cœurs Tensor et 16 Go de mémoire GDDR6 à 448 Go/s. Cette mémoire tampon de processeur graphique est adaptée aux jeux de données ou aux tailles de lots allant jusqu'à 16 Go.
- **Stations de travail fixes Dell Precision 5820 et 7920 pour la science des données**
  Elles utilisent les processeurs graphiques RTX A6000, RTX 6000 et RTX 8000 de NVIDIA. Le processeur graphique RTX 6000 contient 4 608 cœurs CUDA, 72 cœurs RT, 576 cœurs Tensor et une mémoire GDDR6 à 672 Go/s.
  Pour les modèles RTX 6000 et RTX 8000, seule la taille de mémoire tampon diffère. Le RTX 6000 dispose de 24 Go, tandis que le RTX 8000 en a 48. Ainsi, le RTX 6000 convient aux jeux de données ou aux tailles de lots allant jusqu'à 24 Go et le RTX 8000 convient aux jeux de données ou aux tailles de lots allant jusqu'à 48 Go.

La STSD 5820 est une plate-forme au format tour qui peut être configurée avec un ou deux processeurs graphiques RTX A6000, RTX 8000 ou RTX 6000. Si les deux cartes de processeur graphique sont connectées via NVLink, la taille de la mémoire tampon vidéo du processeur graphique est doublée. Cela permet à la STSD 5820 de prendre en charge des jeux de données ou des tailles de lots allant jusqu'à 96 Go.

La STSD 7920 est une plate-forme au format tour ou rack qui peut être configurée avec un, deux ou trois processeurs graphiques RTX A6000, RTX 8000 ou RTX 6000, fournissant jusqu'à 144 Go de mémoire provenant de trois processeurs graphiques RTX A6000 ou RTX 8000.

Sur la STSD 7920 au format tour, deux de ces trois cartes de processeur graphique peuvent être connectées via NVLink, ce qui double la mémoire tampon vidéo de processeur graphique.

Ainsi, la STSD 7920 peut prendre en charge des jeux de données ou des tailles de lots allant jusqu'à 96 Go, plus jusqu'à 48 Go de mémoire supplémentaire provenant du troisième processeur graphique.

ÉTAPE 3

Sélection de la configuration optimale pour le processeur et sa mémoire afin qu'ils puissent répondre aux besoins de calcul et de données du côté processeur graphique du système

L'objectif est d'atteindre un taux d'utilisation du processeur graphique proche de 100 % lors des exécutions d'entraînement du modèle. Enfin, il faut configurer et dimensionner le stockage.

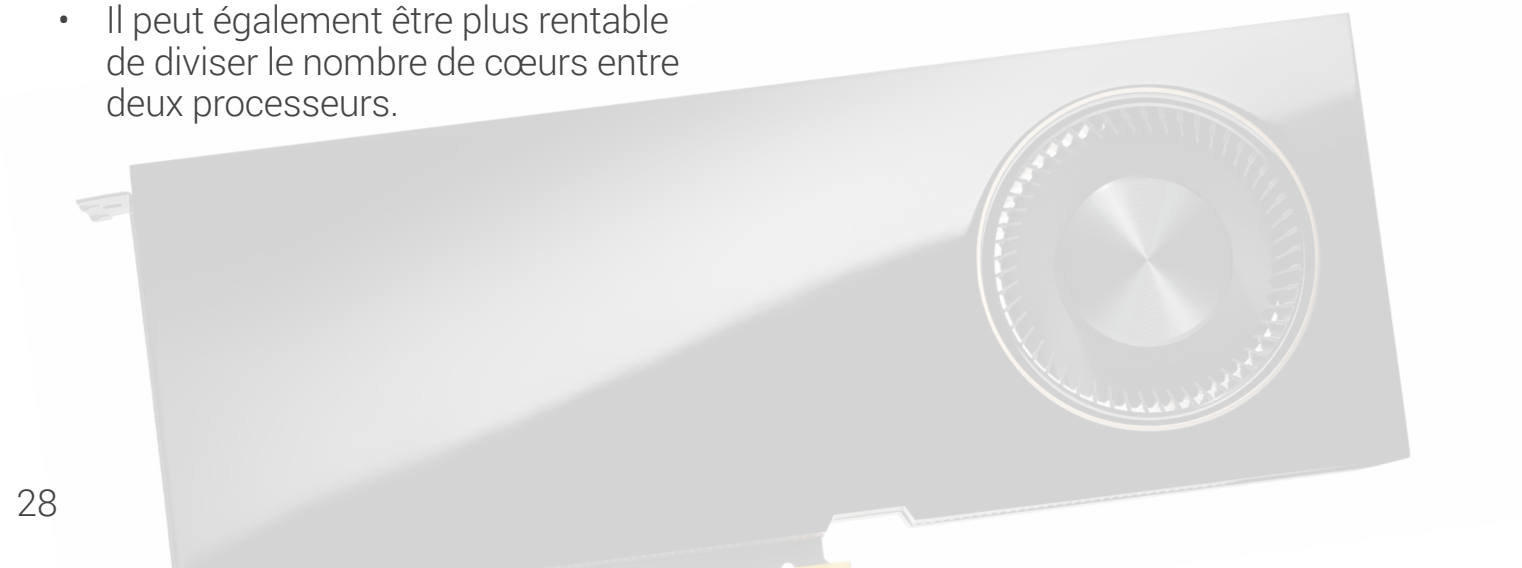
D'après notre expérience en matière de configuration et de mesure des performances sur de nombreuses charges applicatives accélérées par processeur graphique, une pratique d'excellence en matière de dimensionnement de la mémoire du processeur dans le but d'utiliser pleinement le côté processeur graphique du système consiste à configurer deux fois plus de mémoire DDR de processeur que de mémoire de processeur graphique.

Par exemple, dans une STSD 5820 au format tour équipée de deux processeurs graphiques RTX A6000, configurez la mémoire du processeur pour qu'elle soit de 192 Go. Pour une STSD 7550 mobile avec un processeur graphique RTX 5000, configurez la mémoire du processeur pour qu'elle soit d'au moins 32 Go.

D'après ce que nous avons appris des scientifiques des données et de notre propre expérience dans l'entraînement de différents modèles d'apprentissage automatique et en particulier de Deep Learning, lors de la configuration d'une station de travail d'IA accélérée par processeur graphique, il faut plusieurs processeurs multicœurs pour utiliser pleinement deux ou trois processeurs graphiques. C'est pourquoi nous vous recommandons la pratique d'excellence suivante :

- Prévoyez un cœur de traitement pour 8 à 16 Go de mémoire DDR de processeur.
- Lorsqu'un processeur ne dispose pas de suffisamment de cœurs, divisez le nombre de cœurs entre les deux processeurs.
- Il peut également être plus rentable de diviser le nombre de cœurs entre deux processeurs.

Par exemple, si vous disposez de deux processeurs graphiques RTX A6000 dans une tour STSD 7920, configurez cette dernière avec 192 Go de mémoire DDR de processeur et prévoyez d'utiliser deux processeurs Xeon contenant chacun environ 6 à 12 cœurs.



Comment configurer votre mémoire DDR de processeur

Il est important de tirer pleinement parti de tous les canaux de mémoire du contrôleur de mémoire de votre processeur. Cela optimise la bande passante de la mémoire du processeur disponible du côté processeur du système afin d'utiliser pleinement le côté processeur graphique du système. Si vous ne le faites pas, une grande quantité de performances système sera inexploitée :

- Pour les STSD mobiles 7550 et 7750, cela implique de remplir les sockets DIMM DDR par multiples de quatre
- Pour la STSD 5820 au format tour, cela implique de remplir les sockets DIMM DDR par multiples de quatre
- Pour les STSD 7920 au format tour et 7920 au format rack, cela implique de remplir les sockets DIMM DDR de chaque processeur par multiples de six

Enfin, dimensionnez le stockage de masse du système. Les pratiques d'excellence que nous avons recueillies auprès des scientifiques des données impliquent de séparer le stockage de masse des données « chaudes » du système du stockage de masse utilisé en tant que disque de démarrage. Pour les charges applicatives d'IA, le disque de démarrage et le ou les disques de données doivent être de type SSD. Les disques SSD de classe 50 sont recommandés, mais la classe 40 peut être un bon compromis si la rentabilité est un facteur à prendre en compte ou si cela libère plus de budget pour des processeurs comportant davantage de cœurs.

La taille et la complexité des données augmentant considérablement dans les charges applicatives d'IA, la taille des disques de données est en train de s'accroître. À la date de publication de ce document, les disques SSD de 1 à 2 To répondent à la plupart des exigences de données en matière de développement de modèles IA pour les stations de travail.

Certains scientifiques des données choisissent d'inclure du stockage pour les données « froides » afin de soutenir les disques SSD de données chaudes. Il peut s'agir de disques SSD également, mais des disques SATA plus rentables peuvent aussi être utilisés. Nous vous recommandons cependant d'utiliser ce type de disque SATA uniquement pour le stockage de données froides.

Conclusion

Les renseignements obtenus lors de l'exécution des points de référence de ce livre blanc indiquent que la sélection du processeur graphique et de la plate-forme appropriés dépend de la charge applicative.

Les stations de travail mobiles offrent des fonctionnalités qui permettent d'effectuer des tâches de vision par ordinateur et de NLP de partout. Pour les charges applicatives plus exigeantes, telles que la classification d'images médicales haute résolution ou le NLP avec un grand modèle utilisant une longueur de séquence et une taille de lot élevées, des processeurs graphiques hautes performances tels que les RTX 6000/8000/A6000 des stations de travail Dell au format tour seront parfaits.

Dans les modèles NLP de pointe tels que BERT LARGE avec plus de 340 millions de paramètres, le test d'une combinaison de longueur de séquence et de taille de lot plus élevées n'a été possible qu'avec le RTX A6000 ou le RTX 8000 ; ces modèles disposant de 48 Go de mémoire GDDR6 et d'un grand nombre de cœurs CUDA.

Dans le cas des tâches d'apprentissage automatique, telles que l'utilisation de jeux de données tabulaires et structurés, les processeurs graphiques haut de gamme tels que RTX 6000/8000/A6000 peuvent stimuler le chargement du jeu de données dans la mémoire du processeur graphique et booster l'entraînement lors de l'utilisation de bibliothèques accélérées par processeur graphique telles que XGBoost.

Dans la dernière section de ce document, nous avons présenté des instructions générales et des pratiques d'excellence sur la façon de dimensionner correctement les stations de travail Dell Precision pour la science des données.

Plus d'informations et sujets connexes

Vous trouverez ici les informations et les liens mentionnés dans ce document, ainsi que d'autres renseignement et liens simples vers des stations de travail Dell Precision pour la science des données, préconfigurées, et prêtes à l'achat et à la personnalisation. En outre, consultez cette page de destination pour obtenir les informations les plus récentes, qui ne sont pas disponibles au moment de cette publication.

Liens Dell :

Livre blanc sur les performances de la STSD Dell et de la solution NAS Isilon H400

Guide de référence rapide sur l'IA

Guide d'installation d'une STSD

Présentation du secteur IA/STSD

Présentation des éléments d'une STSD

Liaisons NVIDIA :

Pile NVIDIA Data Science

Éducation et formation sur l'IA accélérée par processeur graphique NVIDIA

Liens Canonical :

https://certification.ubuntu.com

https://ubuntu.com/dell

https://ubuntu.com/contact-us

Annexe 1

Résultats des performances du processeur graphique RTX A6000

Cette annexe fournit les résultats de l'exécution d'analyses comparatives des performances sur le processeur graphique RTX A6000 de NVIDIA. Nous avons suivi la même méthodologie d'analyse comparative que celle décrite dans la partie principale de ce livre blanc et appliquée aux processeurs graphiques RTX 5000, RTX 6000, RTX 8000 et GV100. Dans les analyses comparatives des performances exécutées cette fois-ci, le système d'exploitation Linux a été mis à jour d'Ubuntu 18.04 vers Ubuntu 20.04, et la pile logicielle NV Data Science a été mise à jour de la version 2.4.0 vers la version 2.8.0. Par conséquent, les résultats reflètent non seulement les améliorations de performances du matériel de processeur graphique RTX A6000, mais également celles des optimisations d'Ubuntu 20.04 et des bibliothèques et pilotes de périphériques de la pile logicielle NVIDIA Data Science 2.8.0.

Plates-formes et configurations matérielles des stations de travail Dell pour la science des données

Nous avons réutilisé et analysé les STSD Precision 5820 et 7920 au format tour, et installé des processeurs graphiques RTX A6000 simples, doubles et triples, dont des paires de RTX A6000 avec et sans NVLinking. Nous avons également réexécuté des analyses comparatives sur les processeurs graphiques RTX 6000 et RTX 8000 simples, doubles et triples, car des mises à jour ont été apportées au système d'exploitation Linux Ubuntu et à la pile logicielle NVIDIA Data Science depuis que nous avons exécuté les analyses comparatives initialement décrites dans la partie principale de ce document.

Les points de référence

Nous avons exécuté les versions actuelles des points de référence utilisés précédemment pour recueillir les performances des STSD Dell sur les trois charges applicatives différentes :

- Apprentissage automatique : XGBoost sur des données épidémiologiques (structurées)
- Deep Learning pour la classification d'images : imagerie tf_cnn sur ResNet50 (données non structurées)
- Deep Learning pour le traitement du langage naturel (NLP) : BERT Large sur le jeu de données SQuAD v1.1 (données non structurées)

Conclusions pour l'A6000 : apprentissage automatique à l'aide de XGBoost sur données épidémiologiques (structurées)

XGBoost (eXtreme Gradient Boosting) est la dernière évolution des algorithmes basés sur des arbres de décision. Il s'appuie sur un ensemble de méthodes d'arborescence qui appliquent le principe visant à stimuler les apprenants faibles. Il s'agit d'une technique de pointe pour les données tabulaires et structurées.

Nous avons utilisé un jeu de données simulé pour la population du Royaume-Uni, synthétisé à partir de données de recensement officielles du Royaume-Uni, pour prédire la probabilité d'infection d'une personne par un virus simulé. Le jeu de données a été utilisé dans l'entraînement dli RAPIDS. Nous avons utilisé un notebook de démo de Rapids.ai pour exécuter le test. RAPIDS et XGBoost sont inclus dans la pile logicielle Data Science de NVIDIA. Nous avons utilisé la partie RAPIDS 0.18 par défaut de la pile Data Science 2.8.0.

Le jeu de données que nous avons utilisé comportait 6 millions de lignes x 501 colonnes (la 501e correspondant à la sortie indiquant si la personne est infectée ou non), le tout en précision FP32. L'entraînement du modèle a été effectué sur des RTX A6000 et RTX 6000 afin de comparer la durée nécessaire pour terminer 500 itérations.

La figure 6 présente nos résultats. Le RTX 6000 a mis 23,7 secondes pour terminer 100 époques, tandis que le RTX A6000 n'a mis que 14,5 secondes, ce qui est 38 % plus rapide. Le GV100 a été plus rapide, en ne mettant que 9,8 secondes.

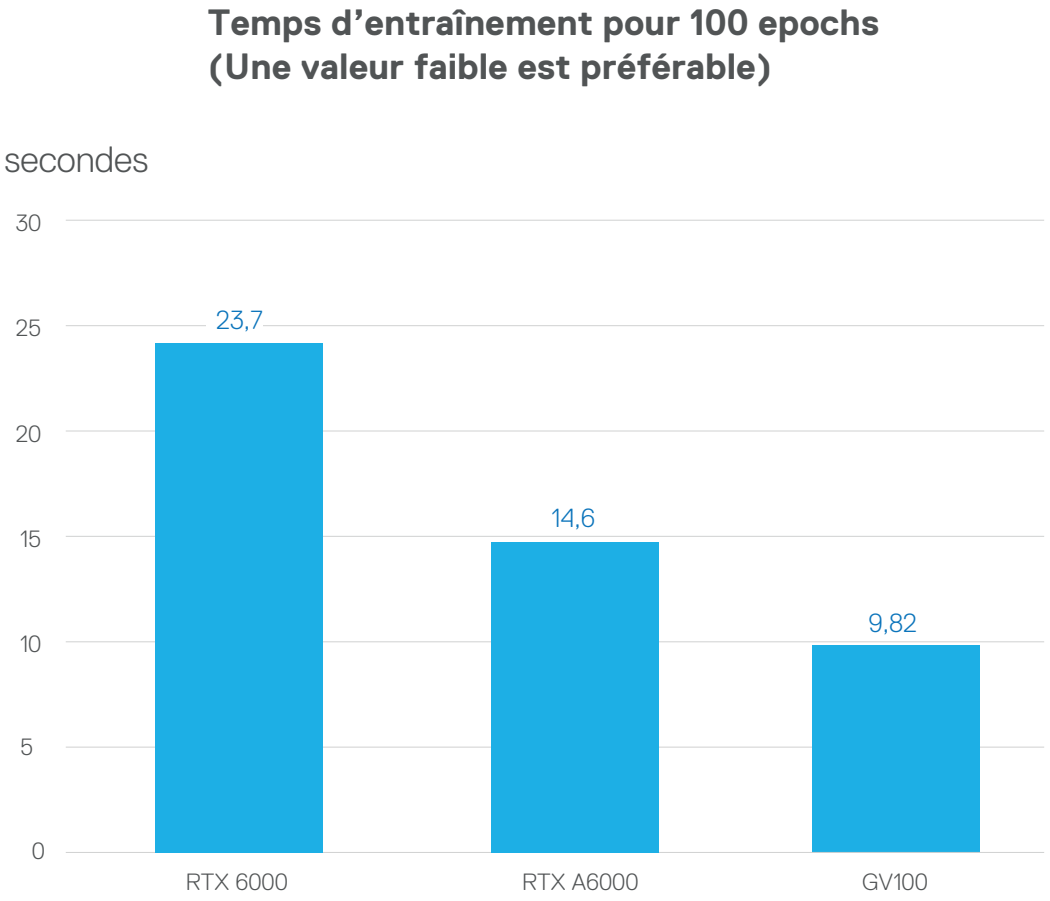


Figure 6

Annexe 1 (suite)

Conclusions pour l'A6000 : Deep Learning pour la classification d'images ; imagerie tf_cnn sur ResNet50 (données non structurées)

Le test a été effectué à l'aide du tf_cnn_benchmarks officiel de TensorFlow. Le référentiel contient des scripts qui exécutent un entraînement de modèles de classification d'images standard sur des images synthétiques, ainsi que sur des jeux de données ImageNet. L'entraînement est exécuté pour un nombre défini d'itérations, tandis que la vitesse moyenne est mesurée en images/seconde. Pour tester pleinement les fonctionnalités des stations de travail, nous avons exécuté le point de référence avec diverses tailles de lots (BS) et différents nombres de processeurs graphiques. Nous avons utilisé tf_cnn pour entraîner le jeu de données d'images synthétiques ResNet50. Nous avons utilisé la demi-précision (FP16).

La figure 7 illustre les résultats de la classification d'images tf_cnn sur Resnet50. La figure montre que les STSD Dell configurées avec le processeur graphique RTX A6000 Ampere ont considérablement surpassé les processeurs graphiques RTX 6000 et RTX 8000, et même le GV100. Elle indique également qu'une mémoire volumineuse était nécessaire pour exécuter les grandes tailles de lots (1 024) que le RTX 6000 n'avait tout simplement pas pu traiter. Les deux processeurs graphiques NVLink n'ont pas considérablement amélioré les performances pour la classification d'images à l'aide du point de référence tf_cnn. Il en a été déduit que le bus PCIe3 est tout simplement suffisant pour gérer le trafic entre les deux cartes. Dans la plupart des cas, les performances des deux RTX A6000 étaient deux fois supérieures à celles du RTX A6000 seul. L'ajout d'un troisième RTX A6000 a offert une évolution linéaire qui a permis d'entraîner encore plus d'échantillons et en moins de temps. Cela se traduit par une réduction du temps nécessaire pour converger vers le modèle de Deep Learning final.

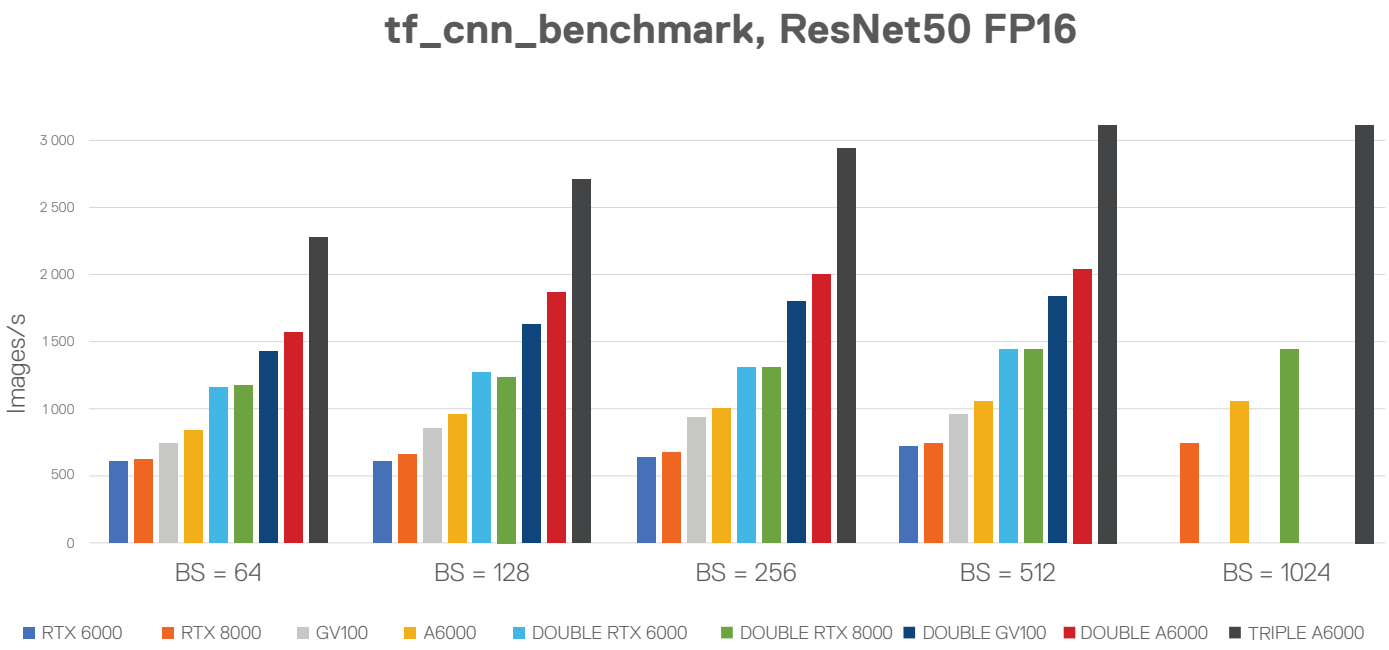


Figure 7

Cliquez ici pour développer.

Conclusions pour l'A6000 : traitement du langage naturel (NLP) ; BERT Large sur jeu de données SQuAD v1.1 (données non structurées)

Le traitement du langage naturel (NLP) est utilisé pour comprendre et générer le langage humain. Pour les performances NLP, nous avons utilisé BERT (Bidirectional Encoder Representations from Transformers) Large.

BERT est devenu une mesure standard du NLP. BERT utilise la modélisation bidirectionnelle du langage pour comprendre le contexte d'un mot (c'est-à-dire que le modèle apprend le contexte d'un mot en fonction de tout son environnement). Pour nos mesures, nous avons choisi l'entraînement BERT Large sur le jeu de données SQuAD v1.1.

À l'aide des STSD Dell, nous avons couvert les paramètres BERT suivants :

- Précision : FP16 et FP32
- Longueur de séquence : 128, 384
- Taille de lot : 1, 2, 4, 8, 16, 32 et 64

La figure 8 présente les résultats, exprimés en « entraînements par seconde ». Une évolution linéaire significative se produit de un à deux, puis à trois RTX A6000. Le RTX A6000 a pu gérer une plus grande combinaison de longueurs de séquences et de tailles de lots. Dès que l'on passe à des séquences de données plus longues, des tailles de lots plus importantes et une précision supérieure (par exemple, FP32), l'évolution est de plus en plus marquée entre les processeurs graphiques simples, doubles, puis triples. Dans ces cas précis, le RTX 6000 (simple, double ou triple) n'a pas pu exécuter les combinaisons en raison de sa mémoire limitée à 24 Go ; le RTX A6000 ayant quant à lui 48 Go de mémoire.

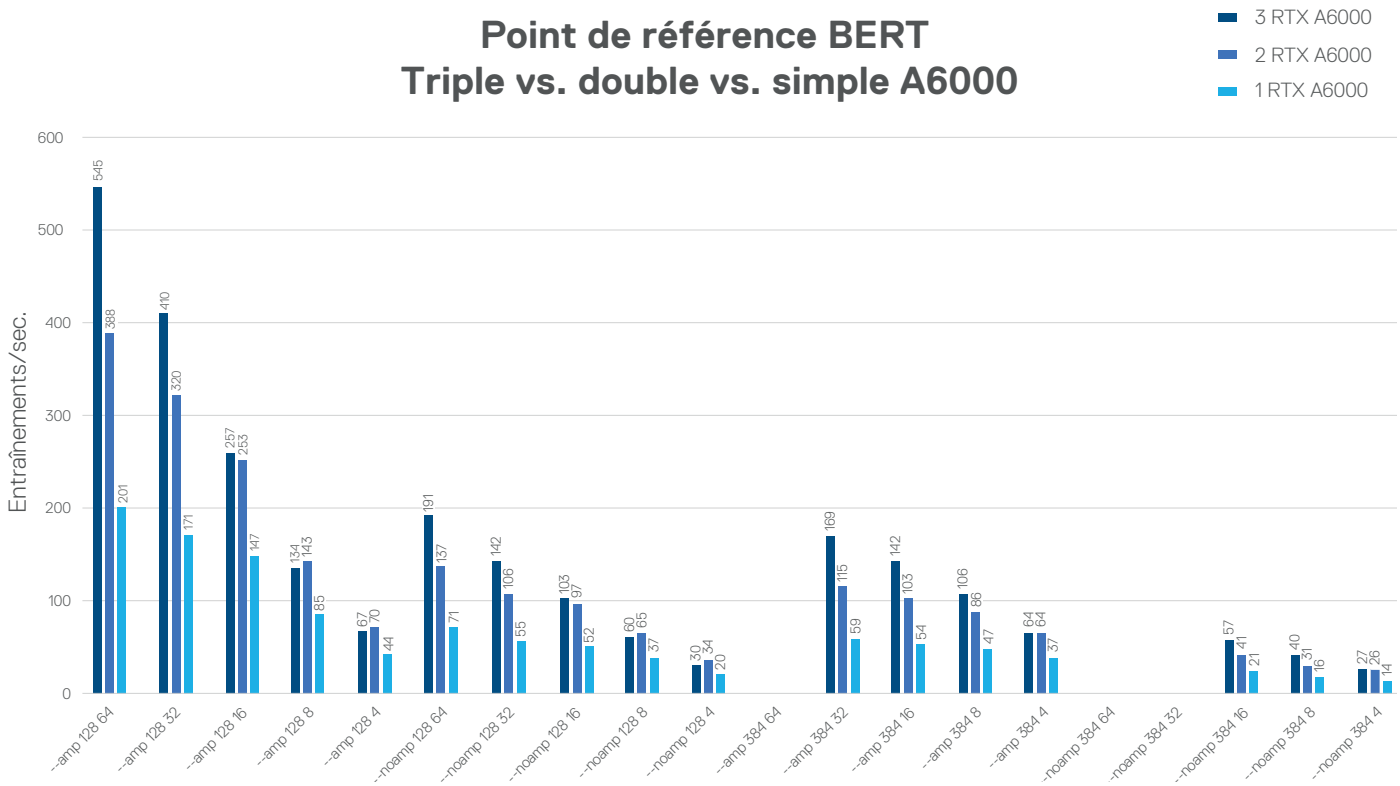


Figure 8

Cliquez ici pour développer.

Annexe 1 (suite)

Conclusions pour l'A6000 : avantage de NVLink entre deux RTX A6000

La figure 9 présente les résultats de l'analyse comparative BERT Large d'une paire de processeurs graphiques RTX A6000 avec et sans NVLink. Un connecteur NVLink utilisé entre deux processeurs graphiques RTX permet de voir les mémoires de chaque carte comme une seule mémoire unifiée, ce qui double la quantité de mémoire et la quantité de cœurs Tensor fonctionnant sur cette mémoire totale pour l'entraînement du modèle. Tout aussi important, les données et la communication entre les deux cartes de processeur graphique peuvent passer par NVLink plutôt que par le bus PCIe beaucoup plus lent. Cela peut considérablement augmenter les performances sur les grands modèles de Deep Learning ou d'apprentissage automatique entraînés à l'aide de jeux de données très volumineux. C'est le cas avec les entraînements BERT sur NLP comme ceux que nous avons utilisés. Le graphique montre qu'une paire de processeurs graphiques RTX A6000 avec NVLink offre des performances beaucoup plus élevées que sans NVLink. Les performances sont entre 2,5 et 9 fois plus rapides que si les deux processeurs graphiques ne sont connectés que par le bus PCIe 3.2.

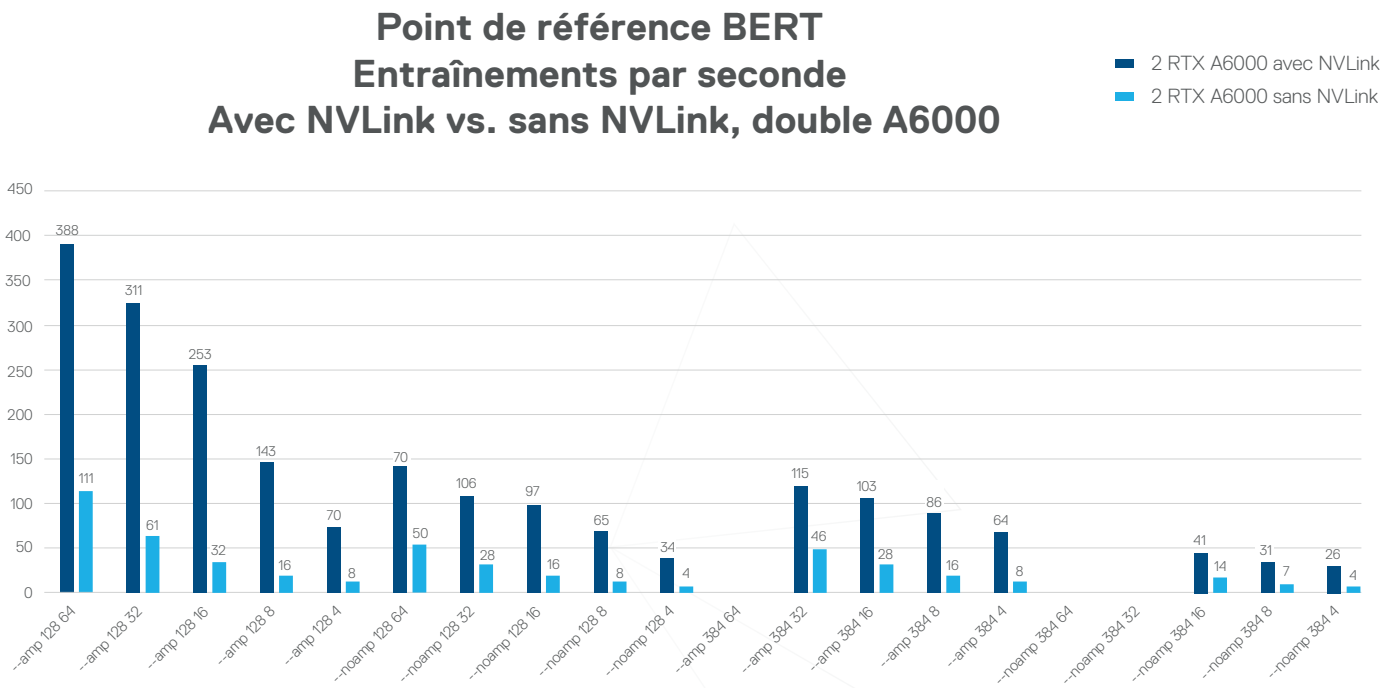


Figure 9

Cliquez ici pour développer.


Precision 7920
Precision 5820