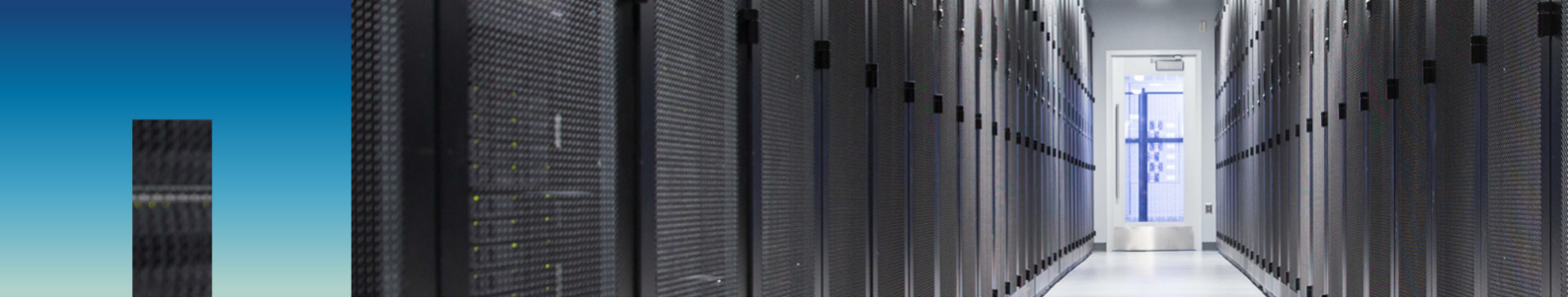


Architecture vérifiée NetApp

NetApp ONTAP AI, optimisée par NVIDIA

Infrastructure d'IA évolutive : cas d'utilisation réels d'apprentissage profond

David Arnette, Sundar Ranganathan, Amit Borulkar, Sung-Han Lin et Santosh Rao, NetApp Août 2018 | NVA-1121

En partenariat avec

NVIDIA

SOMMAIRE

LISTE DES TABLEAUX

LISTE DES FIGURES

1 Synthèse

Ce document contient des informations concernant l'architecture décrite dans le livre blanc L'infrastructure d'IA évolutive (WP-7267). Cette architecture repose sur un système FAS 100 % Flash NetApp® AFF A800, des serveurs NVIDIA® DGX-1™ et des switchs Ethernet 100 Gb Cisco® Nexus® 3232C. Le système a été soumis à des bancs d'essai standard pour en vérifier le fonctionnement et les performances. Les résultats des tests de validation indiquent que cette architecture se révèle très performante en matière d'entraînement et d'inférence. Ils montrent également qu'elle offre des capacités de stockage appropriées pour prendre en charge plusieurs serveurs DGX-1. Le système testé permet en outre de de dimensionner indépendamment les ressources de calcul et de stockage, passant d'une configuration demi-rack à une configuration multi-rack avec des performances prévisibles pour répondre aux exigences des charges de travail d'apprentissage machine.

2 Récapitulatif du programme

2.1 Programme d'architecture vérifiée NetApp

Le programme d'architecture vérifiée NetApp propose une architecture validée pour les solutions NetApp qui est :

- Testée en profondeur
- Normative par nature
- Conçue pour minimiser les risques de déploiement
- Optimisée pour accélérer la mise en service

Ce document est destiné aux ingénieurs de solutions partenaires et NetApp, et aux décideurs stratégiques clients. Il présente les critères de conception de l'architecture utilisés afin de déterminer l'équipement, le câblage et les modèles de configuration requis dans un environnement spécifique.

2.2 Solution NetApp ONTAP AI

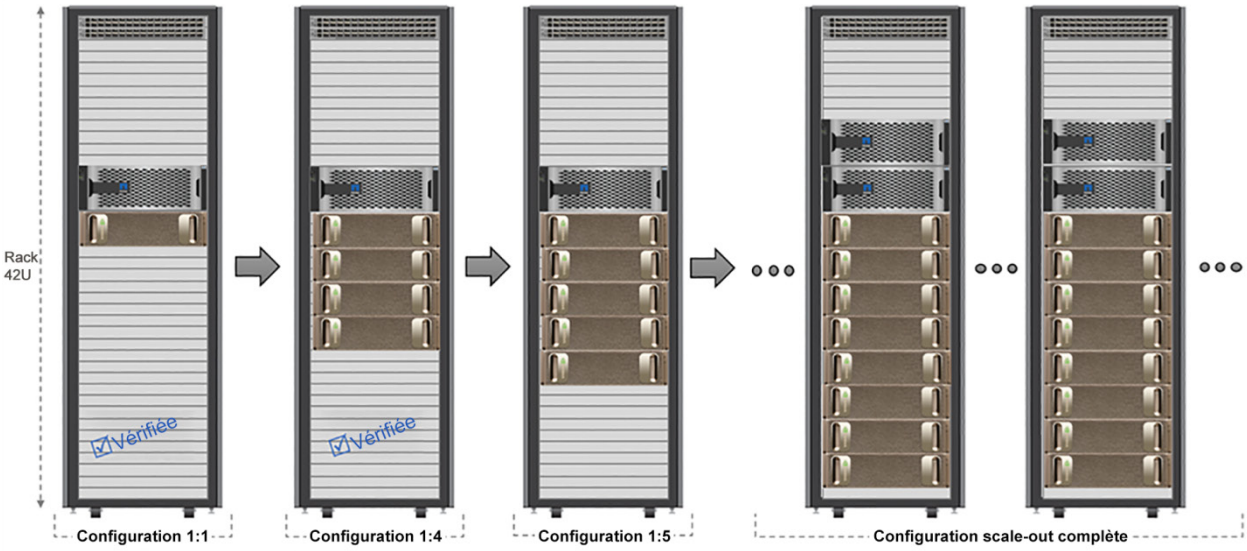
L'infrastructure convergée NetApp ONTAP® AI, optimisée par les serveurs NVIDIA DGX-1 et le système de stockage NetApp connecté au cloud, est une architecture développée et vérifiée par NetApp et NVIDIA. Il s'agit d'une architecture normative qui :

- Simplifie la conception
- Offre une évolutivité indépendante des ressources de calcul et de stockage
- Permet de commencer avec un déploiement de petite envergure, puis d'évoluer de manière fluide
- Propose plusieurs options de stockage pour répondre à des exigences variées de coûts et de performance

NetApp ONTAP AI intègre des serveurs NVIDIA DGX-1, des processeurs graphiques (GPU) NVIDIA TESLA® V100, et un système NetApp AFF A800 avec une connectivité réseau optimale. Cette architecture élimine la complexité et les approximations, et simplifie les déploiements d'intelligence artificielle (IA). Elle permet en outre de commencer avec une petite infrastructure, puis d'évoluer de manière non disruptive tout en gérant intelligemment les données de la périphérie, jusqu'au cœur et au cloud, et inversement.

La Figure 1 montre l'évolutivité de la solution NetApp ONTAP AI. Le système AFF A800 a été testé avec quatre serveurs DGX-1. Les résultats des tests révèlent une marge de performance supplémentaire permettant de prendre en charge cinq serveurs DGX-1 ou plus, sans impact sur la vitesse de stockage ou la latence. De plus, en ajoutant davantage de paires de switchs réseau et de switchs de stockage au cluster ONTAP, la solution peut être étendue à plusieurs racks pour fournir un débit extrêmement élevé, accélérant ainsi l'entraînement et l'inférence. Cette approche flexible permet d'ajuster la répartition entre le calcul et le stockage de manière indépendante, en fonction de la taille du data lake, des modèles d'apprentissage profond utilisés et des metrics de performance requis.

Figure 1) Architecture rack de la solution NetApp ONTAP AI



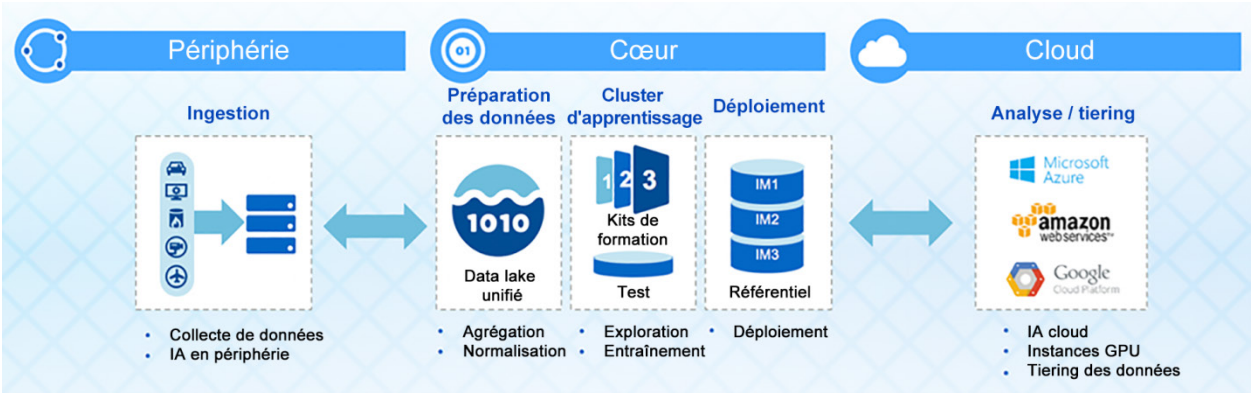
Le nombre de serveurs DGX-1 et de systèmes AFF par rack dépend des spécifications d'alimentation et de refroidissement du rack utilisé. La configuration finale des systèmes tient compte de l'analyse dynamique des ressources de calcul, de la gestion du flux d'air et de l'architecture du data center.

3 Pipeline de traitement de données d'apprentissage profond

Avec l'apprentissage profond, ou deep learning, il est possible de détecter les fraudes, d'améliorer les relations clients, d'optimiser la chaîne d'approvisionnement, et de fournir des produits et des services innovants sur un marché de plus en plus concurrentiel. Les performances et la précision des modèles d'apprentissage profond sont nettement accrues par le développement de la taille et de la complexité des réseaux neuronaux, de même que par l'augmentation de la quantité et de la qualité des données utilisées pour l'entraînement des modèles.

Compte tenu des datasets massifs, il est essentiel de concevoir une infrastructure permettant les déploiements dans plusieurs environnements. Lors du déploiement complet d'applications d'apprentissage profond, les données progressent dans un pipeline composé de trois sections : la périphérie (ingestion de données), le cœur (clusters d'apprentissage et data lake) et le cloud (archivage, tiering et développement/test). Ce schéma est très courant, notamment pour les applications telles que l'Internet des objets (IoT).

Figure 2) Pipeline de traitement de données de la périphérie au cœur, jusqu'au cloud.



La Figure 2 présente un aperçu des composants des trois étapes :

- **Ingestion des données.** L'ingestion a généralement lieu en périphérie, par exemple lors de la capture de données, affluant depuis des véhicules autonomes ou des terminaux de points de vente. Selon l'utilisation, une infrastructure IT peut se révéler nécessaire au niveau du point d'ingestion ou à proximité. Une chaîne de magasins peut avoir besoin d'une empreinte réduite dans chaque point de vente afin de consolider les données de plusieurs terminaux.
- **Préparation des données.** Un prétraitement des données est nécessaire pour les normaliser et les nettoyer avant de procéder à l'entraînement. Ce prétraitement a lieu dans un data lake qui se trouve soit dans le cloud, sous la forme d'un tier Amazon S3, soit dans des systèmes de stockage locaux sous la forme d'un magasin de fichiers ou d'objets.
- **Formation.** Pour la phase d'entraînement critique de l'apprentissage profond, les données sont généralement copiées à intervalles réguliers du data lake vers le cluster d'apprentissage. Les serveurs utilisés lors de cette phase ont recours à des GPU pour paralléliser les calculs, dont l'appétit en données est gigantesque. Il est crucial de répondre aux besoins en bande passante d'E/S brute pour maintenir un taux d'utilisation élevé des GPU.
- **Inférence.** Les modèles entraînés sont testés et déployés en production. Ils peuvent également être renvoyés vers le data lake pour affiner la pondération ou être déployés à la périphérie sur des terminaux intelligents s'il s'agit d'applications IoT.
- **Archivage, tiering.** Les données inactives d'itérations passées peuvent être conservées indéfiniment. Nombre d'équipes spécialisées en IA préfèrent archiver les données inactives dans un système de stockage objet dans un cloud public ou privé.

Selon l'application, les modèles d'apprentissage profond peuvent utiliser un grand nombre de données, structurées ou non structurées, de plusieurs types. Du fait de ces différences, le système de stockage sous-jacent doit être en mesure de répondre à un large éventail d'exigences quant à la taille des données stockées et au nombre de fichiers dans le dataset.

Ces exigences incluent :

- La possibilité de stocker et de récupérer des millions de fichiers simultanément
- Le stockage et la récupération de divers objets de données tels que des images, des données audio, vidéo et des données de différentes séries chronologiques
- Un parallélisme poussé à faible latence qui ne ralentit pas les vitesses de traitement des GPU
- Des services de données et une gestion des données transparentes couvrant la périphérie, le cœur et le cloud

Associés aux fonctionnalités SDS et d'intégration au cloud de NetApp ONTAP, les systèmes AFF couvrent l'ensemble des pipelines de traitement de données de la périphérie au cœur jusqu'au cloud pour l'apprentissage profond. Ce document est axé sur les solutions pour les composants d'entraînement et d'inférence du pipeline de traitement de données.

4 La solution

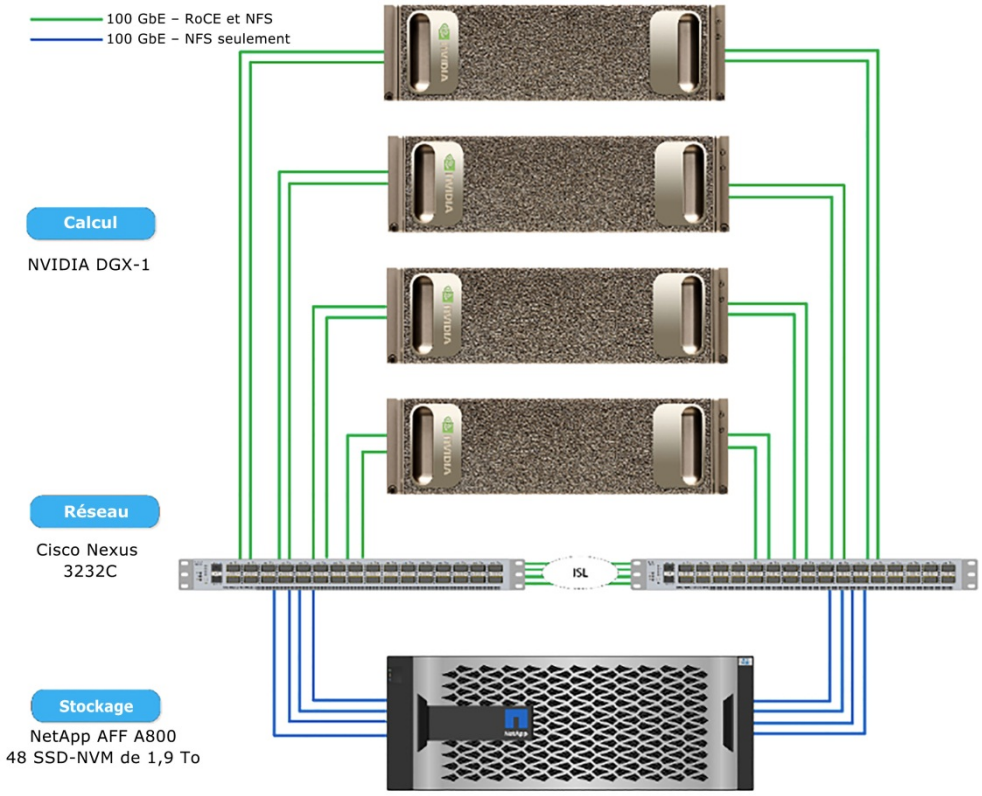
Les systèmes d'apprentissage profond exploitent des algorithmes à forte intensité de calcul, parfaitement adaptés à l'architecture des GPU NVIDIA. Les calculs réalisés par les algorithmes d'apprentissage profond font appel à un nombre immense de produits matriciels exécutés en parallèle. L'architecture hautement parallélisée des GPU modernes les rend plus efficaces que les processeurs classiques, qui sont moins spécialisés pour les applications telles que l'apprentissage profond, dont le traitement de données est effectué en parallèle. Les innovations apportées aux architectures de calcul des GPU NVIDIA en cluster et individuels basées sur le serveur DGX-1 en ont fait les plateformes privilégiées pour les charges de travail d'informatique haute performance (HPC), d'apprentissage profond et d'analytique. Pour fournir des performances optimisées dans ces environnements, l'infrastructure de prise en charge doit être capable de maintenir les GPU NVIDIA alimentés en données. L'accès aux datasets doit donc être assuré à des latences ultra-faibles avec une bande passante élevée.

4.1 Technologie de la solution

Cette solution a été implémentée avec un système NetApp AFF A800, quatre serveurs NVIDIA DGX-1 et deux switchs Ethernet Cisco Nexus 3232C de 100 Gb. Chaque serveur DGX-1 est relié aux switchs Nexus par des liaisons de 100 GbE utilisées pour les communications entre les GPU via le protocole RoCE (RDMA over Converged Ethernet). Les communications IP classiques pour l'accès au stockage NFS s'effectuent également sur ces liaisons. Chaque contrôleur de stockage est relié aux switchs réseau par quatre liaisons de 100 GbE.

Les infrastructures HPC classiques utilisent le protocole RDMA over InfiniBand (IB) pour la connectivité entre nœuds, car il combine une bande passante élevée et une faible latence. Comme la technologie Ethernet permet désormais d'atteindre des niveaux de performance qui n'étaient auparavant possibles qu'avec IB, l'utilisation du protocole RoCE facilite l'adoption de ces fonctionnalités. En effet, les technologies Ethernet sont bien connues et très répandues dans les data centers d'entreprise. La Figure 3 illustre l'architecture de base de la solution.

Figure 3) Architecture vérifiée de la solution NetApp ONTAP AI



4.2 Serveurs NVIDIA DGX-1

Le serveur DGX-1 est un système matériel et logiciel clé en main entièrement intégré, conçu spécialement pour les flux de travail d'apprentissage profond. Chaque serveur DGX-1 est alimenté par huit GPU Tesla V100 configurés en topologie hybride cube-maillage. La technologie NVIDIA NVLink™ qui fournit une bande passante ultra-rapide et un fabric à faible latence pour les communications entre les GPU. Cette topologie est essentielle pour l'entraînement de plusieurs GPU et élimine les goulots d'étranglement caractéristiques des connecteurs PCIe qui ne peuvent pas assurer la linéarité des performances à mesure que le nombre de GPU augmente. Le serveur DGX-1 est également équipé d'interconnecteurs réseau à large bande passante et faible latence pour la mise en cluster de plusieurs nœuds via des fabrics compatibles RDMA.

Le DGX-1 est alimenté par NVIDIA GPU Cloud (NGC), le registre de conteneurs basé sur le cloud de NVIDIA pour les logiciels accélérés par GPU. NGC fournit des conteneurs pour les frameworks d'apprentissage profond les plus populaires, tels que Caffe2, TensorFlow, PyTorch, MXNet et TensorRT, optimisés pour les GPU NVIDIA. Les conteneurs intègrent la structure ou l'application, les

pilotes, les bibliothèques et les primitives de communication nécessaires. Ils sont optimisés sur l'ensemble de la pile par NVIDIA pour accélérer au maximum les performances grâce aux GPU. Les conteneurs NGC incorporent la boîte à outils NVIDIA CUDA, qui fournit la bibliothèque de sous-programmes d'algèbre linéaire de base NVIDIA CUDA (cuBLAS), la bibliothèque de réseau de neurones profonds NVIDIA CUDA et bien plus encore. Les conteneurs NGC comprennent également la bibliothèque NCCL (NVIDIA Collective Communications Library) pour les primitives de communication collective multi-GPU et multi-nœuds, permettant la prise en compte de la topologie pour l'entraînement d'apprentissage profond. NCCL permet la communication entre les GPU dans un seul comme dans plusieurs serveurs DGX-1.

4.3 Systèmes NetApp AFF

NetApp AFF est un système de stockage de pointe qui répond aux besoins des entreprises grâce à ses excellentes performances, sa flexibilité supérieure, son intégration dans le cloud et sa gestion de données optimale. Conçues spécifiquement pour les systèmes Flash, les baies AFF contribuent à accélérer, gérer et protéger les données stratégiques.

Le système NetApp AFF A800 est la première solution NVMe de bout en bout du secteur. Pour les charges de travail NAS, un seul AFF A800 prend en charge un débit de 25 Gbit/s pour les lectures séquentielles et 1 million d'IOPS pour les lectures aléatoires de petite taille à des latences inférieures à 500 microsecondes. Les caractéristiques des systèmes AFF A800 sont :

- Débit massif allant jusqu'à 300 Gbit/s et 11,4 millions d'IOPS dans un cluster à 24 nœuds
- Ethernet 100 Gb avec connectivité FC 32 Gb
- Disques SSD 30 To avec écriture multi-flux (MSW)
- Haute densité de 2 Po dans un tiroir 2U
- Évolutivité de 364 To (2 nœuds) à 74 Po (24 nœuds)
- NetApp ONTAP 9.4, avec une suite complète de fonctionnalités de protection et de réplication des données pour une gestion des données inégalée

Le système de stockage AFF A700s arrive en deuxième position en termes de performances : il prend en charge un débit de 18 Gbit/s pour les charges de travail NAS et le transport 40 GbE. Les systèmes AFF A300 et AFF A220 offrent des performances suffisantes à moindres coûts.

4.4 NetApp ONTAP 9

ONTAP 9 est la dernière génération de logiciel de gestion du stockage de NetApp qui permet de moderniser l'infrastructure et de passer à un data center prêt pour le cloud. Avec des capacités de gestion des données à la pointe du secteur, ONTAP permet de gérer et de protéger les données avec un seul ensemble d'outils, quel que soit leur emplacement. Les données peuvent aussi être déplacées librement partout où elles sont nécessaires, que ce soit en périphérie, au cœur ou dans le cloud. ONTAP 9 comprend de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données stratégiques et pérennisent l'infrastructure sur toutes les architectures de cloud hybride.

Simplification de la gestion des données

La gestion des données est essentielle pour les opérations IT, car elle permet d'utiliser les ressources appropriées pour les applications et les jeux de données. ONTAP inclut les fonctionnalités suivantes pour rationaliser et simplifier les opérations et réduire le coût total de possession :

- **Compaction des données à la volée et déduplication étendue** La compaction des données réduit le gaspillage d'espace à l'intérieur des blocs de stockage, et la déduplication augmente considérablement la capacité effective.
- **Qualité de service (QoS) minimale, maximale et adaptative.** Les contrôles QoS granulaires permettent de maintenir les niveaux de performance des applications critiques dans des environnements hautement partagés.
- **FabricPool ONTAP.** Cette fonctionnalité offre une hiérarchisation automatique des données inactives vers des options de stockage en cloud public et privé, notamment Amazon Web Services (AWS), Azure et la solution NetApp StorageGRID®.

Accélération et protection des données

ONTAP offre des niveaux supérieurs de performances et de protection des données et étend ces fonctionnalités grâce à :

- **Des performances élevées et une faible latence.** ONTAP offre le débit le plus élevé possible à la latence la plus faible possible.
- **NetApp ONTAP FlexGroup.** Un volume FlexGroup est un conteneur de données haute performance pouvant évoluer de manière linéaire jusqu'à 20 Po et 400 milliards de fichiers, fournissant un namespace unique qui simplifie la gestion des données.
- **Protection des données.** ONTAP fournit des fonctionnalités de protection des données intégrées avec une gestion commune sur toutes les plateformes.
- **NetApp Volume Encryption (NVE).** ONTAP offre un chiffrement natif au niveau du volume avec un support de gestion des clés interne et externe.

Infrastructure pérenne

ONTAP 9 aide à répondre aux besoins métier en constante évolution :

- **Évolutivité transparente et opérations non disruptives.** ONTAP prend en charge l'ajout non disruptif de capacité aux contrôleurs et l'évolution scale-out des clusters. Vous pouvez effectuer la mise à niveau vers les technologies les plus récentes, telles que NVMe et FC 32 Gb, sans migration des données ni panne coûteuse.
- **Connexion cloud.** ONTAP est le logiciel de gestion de stockage le plus connecté au cloud, avec des options de stockage SDS (ONTAP Select) et des instances natives de cloud (NetApp Cloud Volumes Service) dans tous les clouds publics.
- **Intégration avec les applications émergentes** ONTAP fournit des services de données d'entreprise pour les plateformes et applications nouvelle génération, telles qu'OpenStack, Hadoop et MongoDB, en utilisant la même infrastructure prenant en charge les applications d'entreprise existantes.

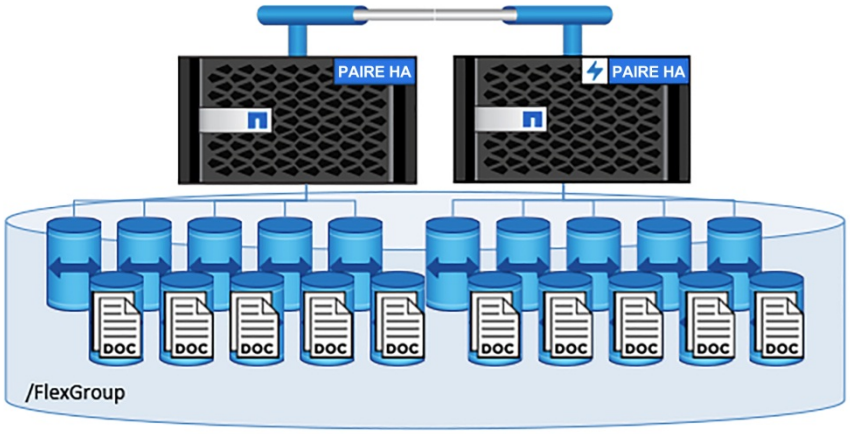
4.5 Volumes NetApp FlexGroup

Le dataset d'entraînement regroupe généralement un grand nombre de fichiers (potentiellement des milliards). Les fichiers peuvent inclure du texte, de l'audio, de la vidéo et d'autres formes de données non structurées qui doivent être stockées et traitées pour être lues en parallèle. Le système de stockage doit stocker un grand nombre de petits fichiers (potentiellement des milliards) et doit lire ces fichiers en parallèle pour les E/S séquentielles et aléatoires.

Un volume FlexGroup (Figure 4) est un namespace unique composé de plusieurs volumes de membres constitutifs qui est géré et agit comme un volume NetApp FlexVol® pour les administrateurs de stockage. Les fichiers du volume FlexGroup sont alloués aux volumes de membres individuels et ne sont pas répartis entre les volumes ou les nœuds. Ils présentent de nombreux atouts :

- Les volumes FlexGroup permettent une capacité massive (plusieurs pétaoctets) et une faible latence prévisible pour les charges de travail comportant un grand nombre de métadonnées.
- Ils prennent en charge des centaines de milliards de fichiers dans le même namespace.
- Ils prennent en charge les opérations parallélisées dans les charges de travail NAS sur les processeurs, les nœuds, les agrégats et les volumes FlexVol constitutifs.

Figure 4) Volumes NetApp FlexGroup.



4.6 NVIDIA GPU Cloud et Trident

Le NVIDIA GPU Cloud (NGC) fournit un catalogue d'images Docker entièrement intégrées et conçues pour la performance, qui tirent pleinement parti des GPU NVIDIA. Ces images incluent toutes les dépendances nécessaires, telles que les bibliothèques NVIDIA CUDA Toolkit et NVIDIA DL. Ces images sont testées, réglées et certifiées par NVIDIA pour être utilisées sur les serveurs NVIDIA DGX-1. En outre, pour permettre la portabilité des images exploitant les GPU, NVIDIA a développé NVIDIA Container Runtime pour Docker, qui vous permet de monter les composants en mode utilisateur des pilotes et GPU NVIDIA dans le conteneur Docker au lancement.

Trident, de NetApp, est un fournisseur de stockage dynamique open source pour Docker et Kubernetes. Associé aux NGC et à des orchestrateurs courants comme Kubernetes ou Docker Swarm, Trident vous permet de déployer de manière transparente vos images de conteneur d'apprentissage profond NGC sur le stockage NetApp, pour que vos déploiements de conteneurs d'IA bénéficient de performances élevées. Ces déploiements incluent une orchestration automatisée, le clonage pour les tests et le développement, des tests mis à niveau qui utilisent le clonage, des copies de protection et de conformité et de nombreux autres cas d'utilisation des données pour les images de conteneur NGC d'IA et d'apprentissage profond.

4.7 Switchs réseau Cisco Nexus 3232C

Le Nexus 3232C de Cisco (Figure 5) est un switch 100 Gbit/s à faible latence, dense, haute performance et écoénergétique spécifiquement conçu pour le data center. Ce modèle compact à 1 rack (1RU) offre une commutation de couche 2 et 3 sur tous les ports avec une latence de 450 nanosecondes. Ce switch fait partie de la plateforme Cisco Nexus 3200 et exécute le logiciel Cisco NX-OS, leader du marché, vous offrant des fonctionnalités complètes largement déployées. Le Cisco Nexus 3232C est un switch Quad Small Form-Factor Pluggable (QSFP) avec 32 ports QSFP28. Chaque port QSFP28 peut fonctionner à 10, 25, 40, 50 et 100 Gbit/s, avec un maximum de 128 ports à 25 Gbit/s.

Figure 5) Switchs Cisco Nexus avec prise en charge NX-OS des normes Converged Enhanced Ethernet et RoCE v1 et v2.

La solution testée n'utilise que la moitié des ports disponibles sur chaque switch réseau. Chaque switch prend en charge jusqu'à huit serveurs DGX-1 avec des ports d'accès de stockage supplémentaires pour améliorer la puissance des GPU. Pour des implémentations de plus grande envergure, le Cisco Nexus 7000 prend en charge jusqu'à 192 ports de 100 GbE par switch. Alternativement, une topologie Leaf-Spine pourrait être implémentée avec plusieurs paires de switchs Nexus 3000 reliés à un switch Spine central.

4.8 RDMA over Converged Ethernet (RoCE)

L'accès direct à la mémoire (DMA) permet aux sous-systèmes matériels tels que les contrôleurs de lecteur de disque, les cartes son, les cartes graphiques et les cartes réseau d'accéder à la mémoire système pour lire/écrire des données sans utiliser les cycles de traitement du processeur. RDMA étend cette fonctionnalité en permettant aux cartes réseau d'effectuer un transfert de données de serveur à serveur entre les mémoires des applications. Ceci s'effectue à l'aide d'une fonctionnalité sans copie, qui n'implique pas le système d'exploitation ni les pilotes des terminaux. Cette approche réduit considérablement l'espace système utilisé et la latence du processeur en contournant le noyau pour les opérations de lecture/écriture et d'envoi/réception.

RoCE est l'implémentation la plus largement déployée de RDMA over Ethernet et s'appuie sur les nouvelles normes CEE (Converged Enhanced Ethernet). Elle est devenue une fonction standard dans de nombreuses cartes réseau haut de gamme, ports CNA (Converged Network Adapter) et switchs réseau. L'Ethernet classique utilise un mécanisme de distribution optimal pour le trafic réseau, mais il ne convient pas à la faible latence et à la bande passante élevée requises pour les communications entre les nœuds GPU. Le CEE créé un support de réseau physique sans perte et permet d'attribuer la bande passante à n'importe quel flux de trafic spécifique sur le réseau.

Pour une livraison garantie sans perte de paquets Ethernet, les réseaux CEE utilisent le contrôle de flux prioritaire (PFC) et la sélection de transmission améliorée (ETS). Le PFC permet d'envoyer des trames de pause pour chaque classe de service (CoS) spécifique. Ceci limite le trafic réseau spécifique tout en permettant à d'autres flux de circuler librement. L'ETS permet une allocation de bande passante spécifique pour chaque CoS pour affiner la granularité du contrôle de l'utilisation du réseau.

La possibilité de hiérarchiser le RoCE sur tout autre trafic permet d'utiliser les liaisons 100 GbE à la fois pour le trafic RoCE et le trafic IP classique, tel que le trafic d'accès au stockage NFS démontré dans cette solution.

5 Exigences technologiques

Cette section indique le matériel et les logiciels utilisés pour la validation de cette solution. Tous les tests documentés dans la section 7, Vérification de la solution, ont été effectués avec le matériel et les logiciels indiqués ci-après.

Remarque : la configuration vérifiée dans cette architecture de référence repose sur la disponibilité des équipements de laboratoire et non sur les exigences ou les limitations du matériel testé.

5.1 Configuration matérielle

Les composants matériels utilisés pour valider cette solution sont répertoriés dans le Tableau 1. Les composants matériels utilisés dans une implémentation spécifique de cette solution peuvent varier en fonction des besoins.

Tableau 1) Configuration matérielle requise.

Matériel	Quantité
Serveurs NVIDIA DGX-1 GPU	4
Système NetApp AFF A800	1 paire haute disponibilité (HA), incluant 48 disques SSD NVMe de 1,92 To
Switchs réseau Cisco Nexus 3232C	2

5.2 Configuration logicielle

Les composants logiciels requis pour implémenter la solution sont répertoriés dans le Tableau 2. Les composants logiciels utilisés dans une implémentation spécifique de cette solution peuvent varier en fonction des besoins.

Tableau 2) Configuration logicielle requise.

Logiciel	Version
NetApp ONTAP	9.4
Firmware du switch Cisco NX-OS	7.0(3)I6(1)
Système d'exploitation DGX-1 pour NVIDIA	Ubuntu 16.04 LTS
Plateforme de mise en conteneurs Docker	18.03.1-ce [9ee9f40]
Version du conteneur	netapp_1.7.0.2 basé sur nvcr.io/nvidia/tensorflow:18.04-py2
Framework d'apprentissage machine (ML)	TensorFlow 1.7.0
Horovod	0.11.3
OpenMPI	3.1.0
Logiciel de banc d'essai	Bancs d'essai TensorFlow [1b1ca8a]

6 Architecture de la solution

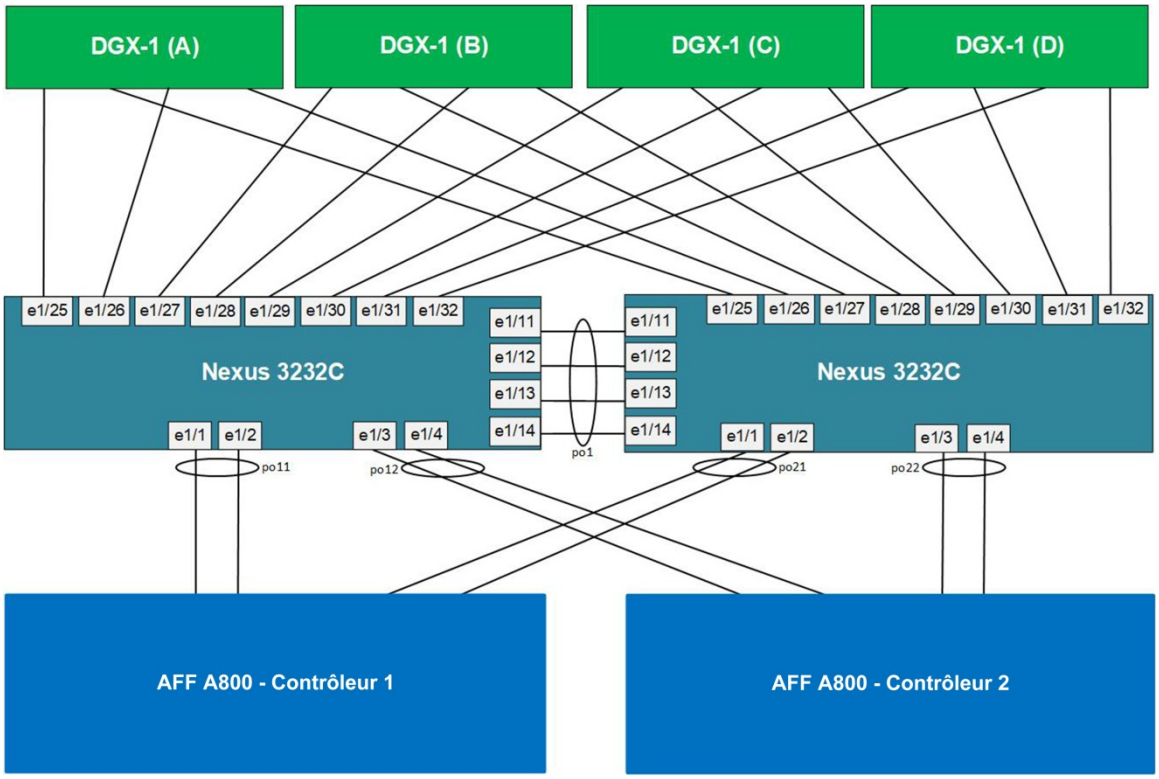
L'architecture vérifiée répond aux exigences de l'exécution des charges de travail d'apprentissage profond. Cette vérification permet aux data scientists de déployer des infrastructures et des applications d'apprentissage profond sur une infrastructure prévalidée. Elle contribue à éliminer les risques et permet aux entreprises de se concentrer sur leurs données pour les transformer en informations exploitables. Cette architecture peut également offrir des performances de stockage exceptionnelles pour les autres charges de travail HPC sans nécessiter ni modification ni réglage de l'infrastructure.

6.1 Topologie du réseau et configuration du switch

Pour cette solution, le standard RoCE est utilisé à la place d'IB afin d'assurer la connectivité à bande passante élevée et à faible latence requise pour la communication entre les serveurs DGX-1. Les switchs Cisco Nexus prennent en charge RoCE en implémentant PFC, ce qui permet aux utilisateurs de donner la priorité au trafic RoCE sur le trafic IP classique sur une liaison partagée. Ils offrent également la possibilité d'utiliser simultanément les liaisons 100 GbE pour RoCE et IP.

Cette architecture utilise une paire de switchs Ethernet Cisco Nexus 3232C de 100 Gb pour le réseau d'accès inter-cluster et de stockage principal. Ces switchs sont reliés par quatre ports réseau de 100 Gb configurés en tant que canaux de port standard. Ce canal de port ISL (Inter Switch Link) permet au trafic de circuler entre les switchs en cas de pannes de liaison de l'hôte ou du système de stockage. Chaque hôte est relié aux switchs Nexus par une paire de liaisons actif-passif et, pour assurer la redondance de la couche de liaison, chaque contrôleur de stockage est relié à chaque switch Nexus par un canal LACP à deux ports. La configuration switch réseau-port est illustrée par la Figure 6.

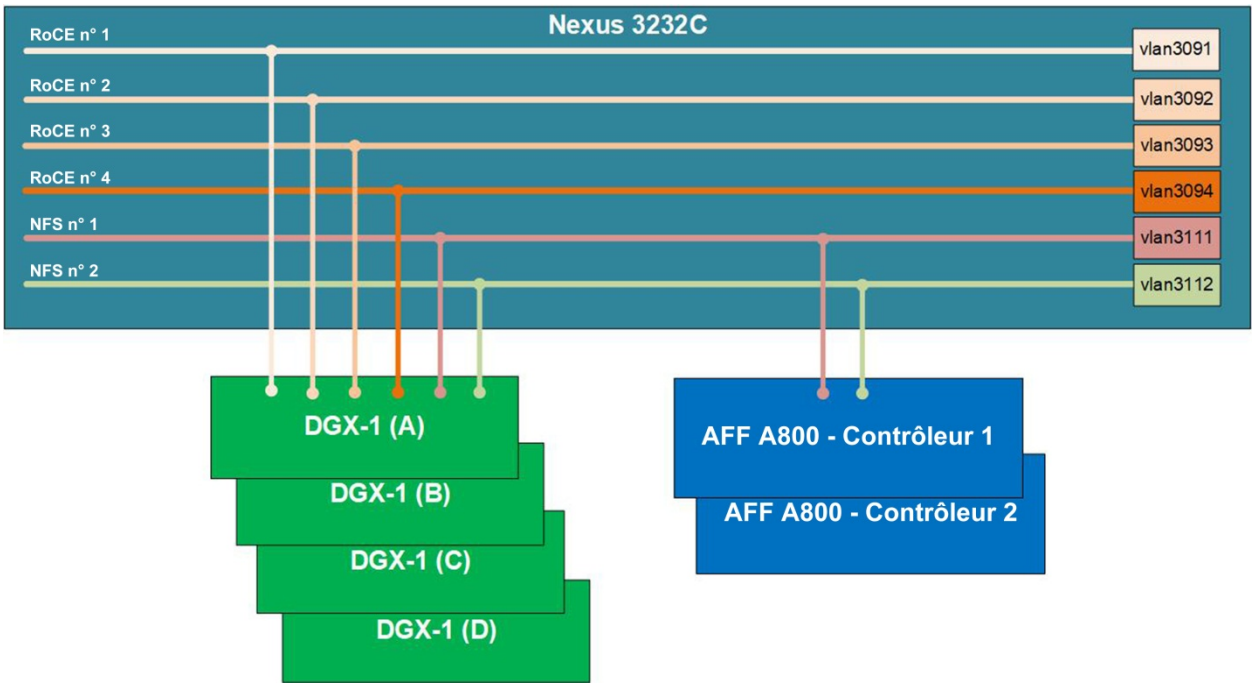
Figure 6) Configuration switch réseau-port.



Plusieurs réseaux locaux virtuels (VLAN) ont été configurés pour prendre en charge le trafic de stockage RoCE et NFS. Quatre VLAN sont dédiés au trafic RoCE et deux sont dédiés au trafic de stockage NFS. Quatre réseaux locaux virtuels et plages d'adresses IP discrets sont utilisés pour assurer un routage symétrique pour chaque connexion RoCE. La pile logicielle NVIDIA gère ces connexions pour l'agrégation de la bande passante et la tolérance aux pannes. Cette solution utilise NFSv3 pour l'accès au stockage, celui-ci ne prenant pas en charge les chemins d'accès multiples, deux VLAN sont utilisés pour activer plusieurs montages NFS dédiés. Cette approche n'offre aucune tolérance de panne supplémentaire, mais permet d'utiliser plusieurs liaisons pour augmenter la bande passante disponible. Le PFC est configuré sur chaque switch pour attribuer les quatre VLAN RoCE à la classe prioritaire et les VLAN NFS à la classe de type « best effort » par défaut. Tous les VLAN sont configurés pour des trames jumbo avec une limite d'unité de transmission maximale (MTU) de 9 000.

Les ports/switchs pour les serveurs DGX-1 sont configurés en tant que ports de jonction et tous les VLAN RoCE et NFS sont autorisés. Les ports/canaux configurés pour les contrôleurs du système de stockage sont également des ports de jonction, mais seuls les VLAN NFS sont autorisés. La connectivité VLAN pour le serveur DGX-1 et les ports du système de stockage est représentée à la Figure 7.

Figure 7) Connectivité VLAN pour les ports DGX-1 et de stockage.



Pour fournir un service prioritaire pour le trafic RoCE, la carte réseau hôte attribue une valeur de CoS de 4 au trafic sur chaque VLAN RoCE. Le switch est configuré avec une stratégie QoS qui assure un trafic constant avec cette valeur CoS. Le trafic NFS se voit attribuer la valeur CoS par défaut de 0, ce qui fait partie de la stratégie QoS par défaut sur le switch et fournit un service optimal.

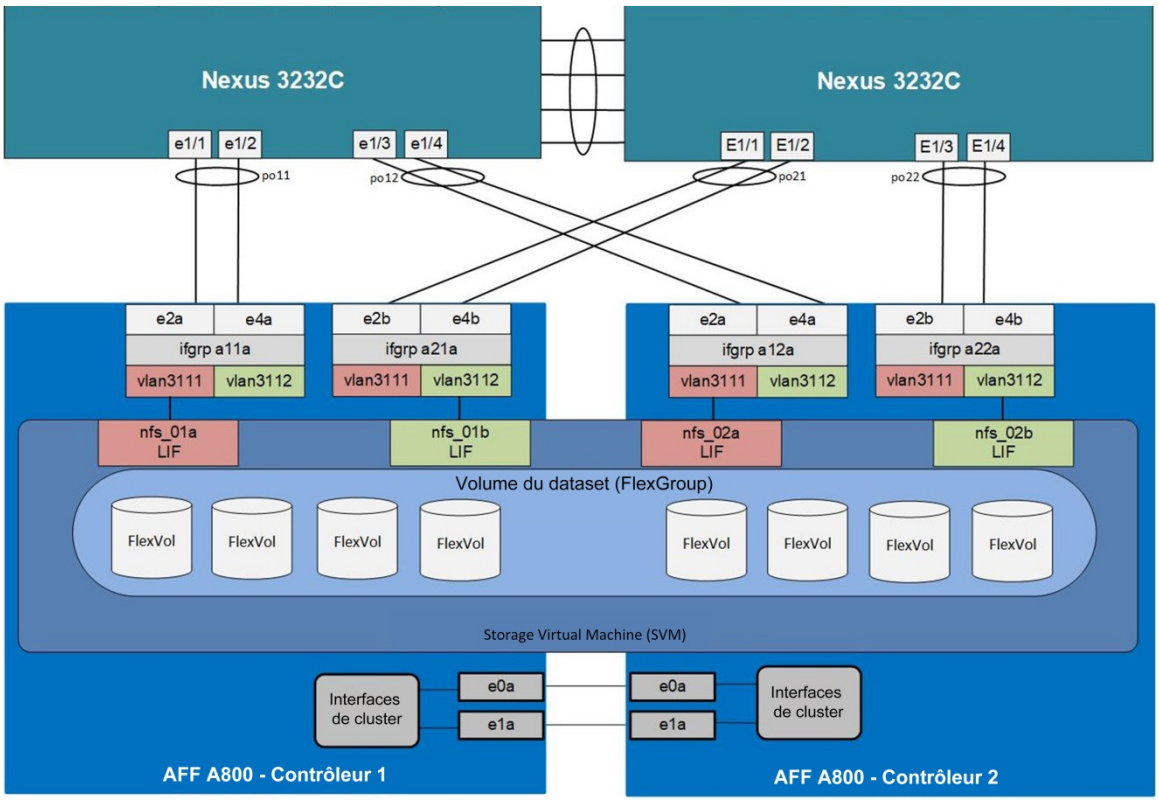
Le PFC est ensuite activé sur chaque port DGX-1, ce qui permet au port du switch d'envoyer des trames de pause pour des classes de service spécifiques afin d'éliminer l'encombrement du switch. En utilisant l'ETS pour allouer 95 % de la bande passante au trafic RoCE en cas de congestion, cette configuration permet une allocation dynamique des ressources entre le trafic RoCE et NFS tout en donnant la priorité aux communications entre nœuds. Vous pouvez également modifier l'allocation de bande passante de manière dynamique pour optimiser les charges de travail nécessitant des performances de stockage plus élevées et une communication interne réduite.

6.2 Configuration du système de stockage

Pour prendre en charge les exigences de réseau de stockage de toute charge de travail potentielle sur cette architecture, chaque contrôleur de stockage est doté de quatre ports 100 GbE en plus des ports intégrés requis pour l'interconnexion des clusters de stockage. La configuration du système de stockage est illustrée à la Figure 8. Chaque contrôleur est configuré avec un groupe d'interface LACP à deux ports (ifgrp sur la Figure 8) pour chaque switch. Ces groupes d'interface offrent une connectivité résiliente pouvant atteindre 200 Gbit/s à chaque switch pour l'accès aux données. Deux VLAN sont provisionnés pour l'accès au stockage NFS et les deux VLAN de stockage sont partagés entre les switchs et chacun de ces groupes d'interface. Cette configuration permet un accès simultané de chaque hôte aux données via plusieurs interfaces, ce qui améliore la bande passante disponible pour chaque hôte.

Tout accès aux données à partir du système de stockage s'effectue via un accès NFS à partir d'une machine virtuelle de stockage (SVM) dédiée à cette charge de travail. La SVM est configurée avec un total de quatre interfaces logiques (LIF) avec deux LIF sur chaque VLAN de stockage. Chaque groupe d'interface héberge un seul fichier LIF, résultant en un LIF par VLAN sur chaque contrôleur avec un groupe d'interface dédié pour chaque VLAN. Cependant, les deux réseaux locaux virtuels sont partagés entre les deux groupes d'interface sur chaque contrôleur. Cette configuration permet à chaque LIF de basculer vers un autre groupe d'interface sur le même contrôleur, de sorte que les deux contrôleurs restent actifs en cas de défaillance du réseau.

Figure 8) Configuration du système de stockage.

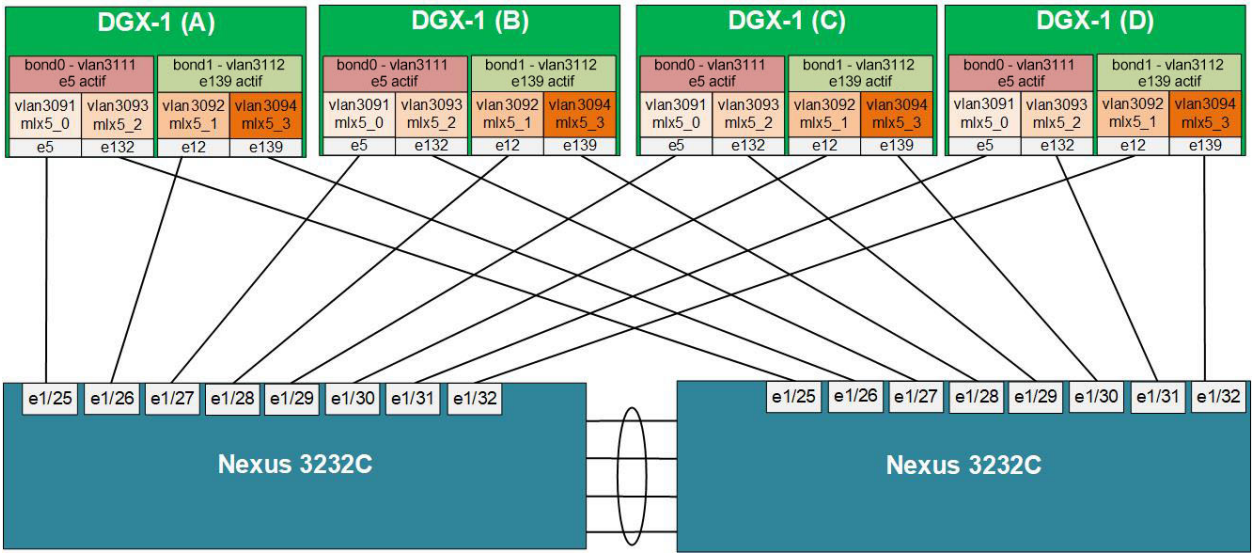


Pour le provisionnement de stockage logique, la solution utilise un volume FlexGroup afin de constituer un pool de stockage réparti sur les nœuds du cluster de stockage. Chaque contrôleur héberge un total de 46 partitions de disque, les deux contrôleurs partageant chaque disque. Lorsque le FlexGroup est déployé sur la SVM de données, un certain nombre de volumes FlexVol sont provisionnés sur chaque agrégat, puis combinés dans le FlexGroup. Cette approche permet au système de stockage de proposer un pool de stockage capable de s'adapter à la capacité maximale de la baie et d'offrir des performances exceptionnelles en exploitant simultanément tous les disques SSD de la baie. Les clients NFS accèdent au groupe FlexGroup en tant que point de montage unique via l'un des fichiers LIF mis à disposition pour la SVM. Pour augmenter la capacité et la bande passante d'accès client, il suffit d'ajouter d'autres nœuds au cluster de stockage.

6.3 Configuration de l'hôte

Pour la connectivité réseau, chaque DGX-1 est doté de quatre cartes d'interface réseau à port unique Mellanox ConnectX4. Ces cartes fonctionnent à des vitesses Ethernet atteignant 100 Gb et prennent en charge RoCE, offrant une alternative moins coûteuse à IB pour les applications d'interconnexion de cluster. Chaque port 100 Gb est configuré en tant que port de jonction sur le switch approprié, avec quatre VLAN et deux VLAN NFS autorisés sur chacun. La configuration VLAN et des ports réseau des hôtes DGX-1 est illustrée à la Figure 9.

Figure 9) Configuration des VLAN et des ports réseau des hôtes DGX-1.



Pour la connectivité RoCE, chaque port physique héberge une interface VLAN et une adresse IP sur l'un des quatre VLAN RoCE. Les pilotes Mellanox sont configurés pour appliquer une valeur de CoS réseau de 4 à chacun des VLAN RoCE. Le PFC est configuré sur les switchs pour assurer un service sans perte prioritaire à la classe RoCE. RoCE ne prend pas en charge l'agrégation de plusieurs liaisons en une seule connexion logique, mais le logiciel de communication NVIDIA NCCL peut utiliser plusieurs liaisons pour l'agrégation de la bande passante et la tolérance aux pannes.

Pour l'accès au stockage NFS, deux liaisons actif-passif sont créées à l'aide d'un lien vers chaque switch. Chaque liaison héberge une interface VLAN et une adresse IP sur l'un des deux VLAN NFS, et le port actif de chaque liaison est relié à un autre switch. Cette configuration offre jusqu'à 100 Go de bande passante dans chaque VLAN NFS et assure une redondance en cas de liaison hôte ou de défaillance d'un switch. Pour que les connexions RoCE bénéficient de performances optimales, tout le trafic NFS est affecté par défaut à la classe de qualité de service Best Effort. Toutes les interfaces physiques et les interfaces de liaison sont configurées avec un MTU de 9 000.

Pour augmenter les performances d'accès aux données, plusieurs montages NFSv3 sont effectués depuis le serveur DGX-1 vers le système de stockage. Chaque serveur DGX-1 est configuré avec deux VLAN NFS, avec une interface IP sur chaque VLAN. Le volume FlexGroup de l'AFF A800 est monté sur chaque VLAN de chaque DGX-1, avec des connexions totalement indépendantes du serveur au système de stockage. Même si un seul montage NFS est capable de fournir les performances requises pour cette charge de travail, plusieurs points de montage sont définis pour permettre l'utilisation d'une bande passante d'accès au stockage supplémentaire pour d'autres charges de travail plus gourmandes en stockage.

7 Vérification de la solution

Cette section décrit les tests que nous avons effectués afin de valider le fonctionnement et les performances de cette solution. Nous avons utilisé pour ces tests les équipements et logiciels spécifiques répertoriés dans la section 5, Exigences technologiques.

7.1 Plan du test de validation

Cette solution a été vérifiée en utilisant des tests standard avec un certain nombre de configurations de calcul afin de démontrer l'évolutivité de l'architecture. Le dataset ImageNet a été hébergé sur l'AFF A800 en utilisant un seul volume FlexGroup auquel NFSv3 pouvait accéder par quatre serveurs DGX-1 au maximum, comme recommandé par NVIDIA pour l'accès au stockage externe. TensorFlow a été utilisé comme structure d'apprentissage machine pour tous les modèles testés, et des mesures de performance de calcul et de stockage ont été relevées pour chaque scénario de test. Les principaux résultats sont présentés à la section 7.2, Résultats du test de validation.

Les modèles de réseaux de neurones convolutifs (CNN) suivants, avec différents niveaux de complexité de calcul et de stockage, ont été utilisés pour tester les taux d'entraînement :

- **ResNet-152** est généralement considéré comme le modèle d'entraînement le plus précis.
- **ResNet-50** offre une meilleure précision qu'AlexNet avec un traitement plus rapide.
- **VGG16** offre la meilleure communication entre les GPU.
- **Inception-v3** est un autre modèle de TensorFlow courant.

Chacun de ces modèles a été testé avec différentes configurations matérielles et logicielles pour étudier les effets de chaque option sur les performances :

- Nous avons testé chaque modèle avec des données synthétiques et le dataset de référence ImageNet. Des tests avec des GPU supplémentaires internes au DGX-1 et sur plusieurs serveurs DGX-1 ont également été réalisés pour évaluer l'évolutivité du cluster de calcul et l'évaluation des performances d'accès au stockage.
- Nous avons utilisé des données ImageNet sans appliquer la distorsion pour réduire la charge de traitement du processeur avant de copier des données dans la mémoire GPU.
- Nous avons testé chaque modèle en utilisant des cœurs Tensor et des cœurs CUDA pour vérifier les améliorations de performances apportées par les cœurs Tensor.
- L'augmentation des performances GPU a eu pour effet d'accroître les exigences d'accès au stockage. Les résultats des tests montrent que l'AFF A800 est tout à fait en mesure de répondre à ces exigences.
- Nous avons testé chaque modèle d'apprentissage profond avec plusieurs tailles de batchs. L'augmentation de la taille de batchs a plusieurs effets sur le système : hausse des taux d'entraînement, baisse des exigences de communication entre les GPU et accroissement des besoins en bande passante de stockage. Nous avons testé les tailles de batchs suivantes avec chaque modèle :
  - 64, 128 et 256 pour ResNet-50
  - 64 et 128 pour tous les autres modèles
- Chaque modèle a été testé avec un, deux et quatre serveurs DGX-1 pour étudier leur évolutivité sur plusieurs GPU utilisant RoCE comme interconnexion (via Horovod).
- L'inférence a été testée à l'aide de tous les modèles avec les tailles de batchs les plus importantes (256 pour ResNet-50 et 128 pour tous les autres modèles), 32 GPU (cœurs Tensor et cœurs CUDA) et le dataset ImageNet.
- Toutes les mesures de performance ont été recueillies après au moins deux séries de tests. Nous avons observé des résultats de performance légèrement meilleurs lors d'entraînements sur plusieurs séries de tests. Chaque test a été exécuté cinq fois et la moyenne des mesures de performance observées a été notée.

7.2 Résultats du test de validation

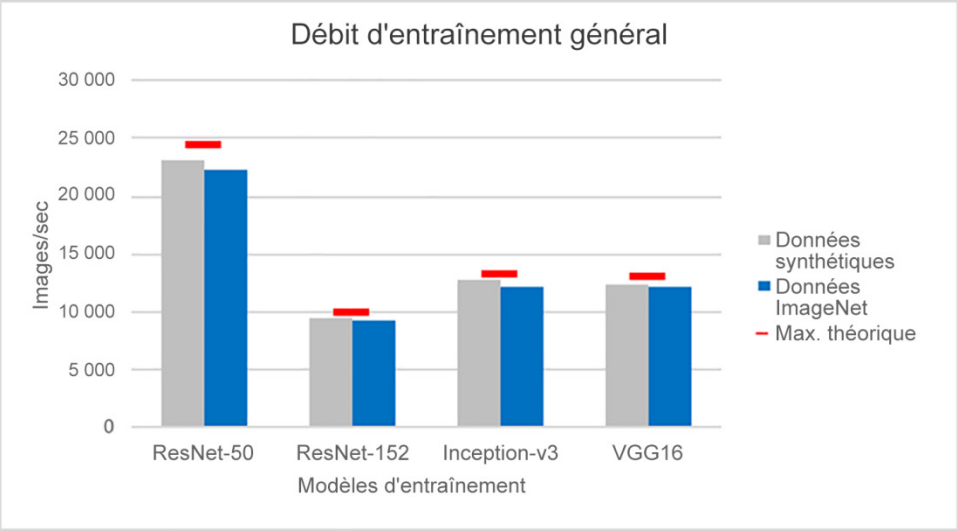
Comme décrit précédemment, nous avons effectué plusieurs tests pour évaluer le fonctionnement général et les performances de cette solution. Cette section présente les principaux résultats relatifs aux données de performances de calcul et de stockage Les résultats complets des tests figurent en annexe. Les informations suivantes concernant les données présentées dans les prochaines sous-sections de ce rapport sont à prendre en compte lors de l'analyse des résultats :

- La performance de l'entraînement des modèles est mesurée en images par seconde.
- Les performances de stockage sont mesurées à l'aide du débit (Mbit/s) et de la latence (µs). Les données du processeur du système de stockage ont également été mesurées afin d'évaluer la capacité de performance restante du système de stockage.
- Chaque système a été testé avec plusieurs tailles de batchs. Des tailles de batchs plus importantes augmentent le débit d'entraînement général. Seule la plus grande taille de batchs testée pour chaque modèle est indiquée. Les données pour chaque taille de batchs testée sont disponibles en annexe :
  - Les tests ResNet-50 ont été réalisés avec une taille de batchs de 256.
  - Les tests ResNet-152, Inception-v3 et VGG16 ont été réalisés avec une taille de batchs de 128.

Débit d'entraînement général

La Figure 10 montre le nombre maximal d'images d'entraînement par seconde obtenues avec chacun des modèles testés en utilisant des cœurs Tensor pour des performances maximales. La Figure 10 compare le débit d'entraînement obtenu avec 32 GPU en utilisant les données ImageNet et les données synthétiques pour la comparaison de base. Elle montre également le maximum théorique réalisable, dans lequel tous les GPU entraînent des données synthétiques de manière indépendante sans mettre à jour les paramètres entre eux. Comme indiqué à la Figure 10, le débit atteint pour les données ImageNet est très proche du débit des données synthétiques.

Figure 10) Débit d'entraînement pour tous les modèles

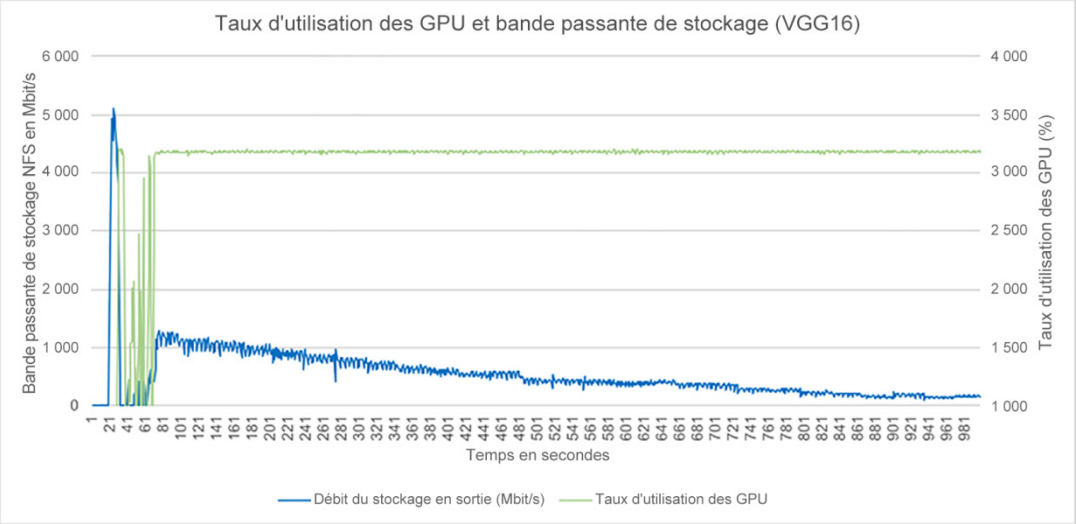


Performances de la charge de travail du GPU

La série de données suivante démontre la capacité du système de stockage à répondre aux exigences du serveur DGX-1 sous une charge complète. La Figure 11 montre le taux d'utilisation des GPU des serveurs DGX-1 et la bande passante de stockage générée lors de l'exécution de chaque modèle avec 32 GPU. Comme l'indique le graphique, la bande passante de stockage commence très haut lorsque les données initiales sont lues depuis le stockage dans le cache TensorFlow. Elle diminue ensuite progressivement à mesure que le dataset intègre la mémoire DGX-1.

Une fois que toutes les données sont dans la mémoire locale, l'accès au stockage devient presque nul. Les GPU DGX-1 commencent à traiter les données presque immédiatement, et taux d'utilisation des processeurs graphiques reste cohérente tout au long du test. Ce graphique montre les résultats du modèle VGG16 avec une taille de batchs de 128, lequel a permis d'atteindre le taux d'utilisation des GPU le plus élevé au cours de nos tests. Les graphiques des autres modèles sont disponibles en annexe. Notez que l'échelle des taux d'utilisation des GPU est la somme des taux d'utilisation de tous les GPU. Dans ce cas, avec 32 GPU testés, le taux d'utilisation maximal possible est de 3 200 %.

Figure 11) Utilisation des GPU et bande passante de stockage (VGG16).



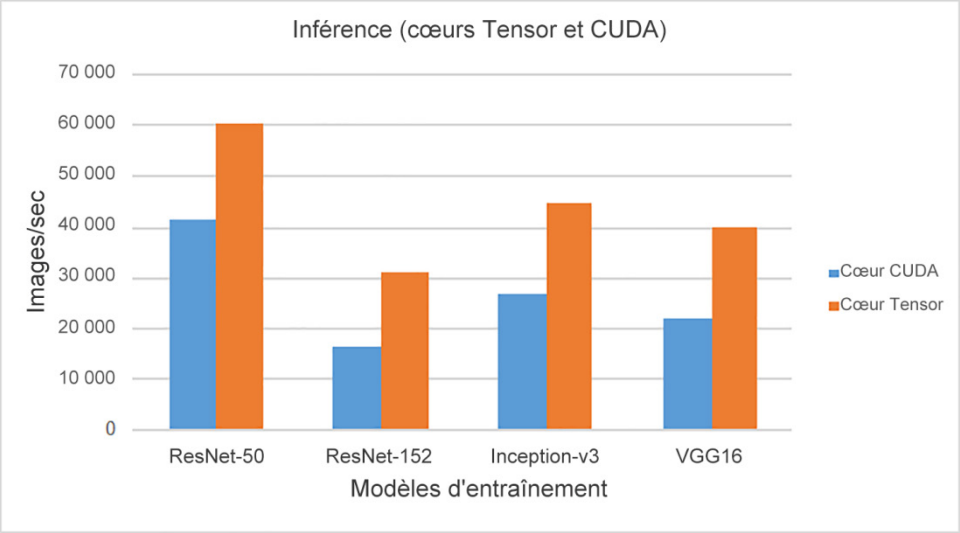
Comme le montre la Figure 11, le taux d'utilisation des GPU reste supérieur à 95 % pour les 32 GPU. Elle reste également constante, quelle que soit la quantité de données provenant du système de stockage. Le système de stockage assure un débit de 5 Gbit/s de données, puis passe d'environ 2 Gbit/s à presque rien pendant le reste de la période d'entraînement. Ce résultat démontre qu'aucun goulot d'étranglement au niveau de l'accès au stockage n'affecte les performances des GPU avec cette charge de travail. Avec un plus grand dataset excédant la capacité de mémoire locale, les performances d'accès au stockage afficheraient un débit constant jusqu'à la fin de l'entraînement. La Figure 11 montre également une comparaison du taux d'utilisation des GPU en fonction de la bande passante de stockage. Elle n'indique pas le temps requis pour l'intégralité de la phase d'entraînement, car la bande passante de stockage diminue progressivement jusqu'à devenir presque nulle.

Inférence avec les GPU

L'inférence consiste à déployer un modèle d'apprentissage profond afin d'évaluer un nouvel ensemble d'objets et de faire des prévisions avec des niveaux de précision d'analyse prédictive similaires à ceux obtenus pendant les phases d'entraînement. Dans une application avec un dataset d'images, l'inférence vise à classer les images en entrée et à répondre aux demandeurs aussi rapidement que possible. Outre un débit élevé, une latence faible est également primordiale.

NetApp ONTAP AI a été utilisé pour tester les performances d'inférence et mesurer les metrics de débit dans cette phase. La Figure 12 affiche le nombre d'images pouvant être traitées par seconde lors de l'inférence. Ce test compare le débit obtenu avec 32 GPU utilisant des données ImageNet sur chacun des modèles testés à l'aide de cœurs Tensor et de cœurs CUDA. Grâce à la puissance de NetApp ONTAP AI, les cœurs Tensor peuvent être utilisés pour classifier instantanément un nombre important d'images.

Figure 12) Inférence pour tous les modèles (cœurs Tensor et CUDA).



Performances de l'AFF A800 avec des charges de travail d'entraînement d'IA

La bande passante de stockage, la latence et la marge CPU ont été recueillies pour étudier les performances du système de stockage avec chacun des modèles testés. Les Figure 13, Figure 14 et Figure 15 montrent les mesures du système de stockage pour chaque modèle testé avec des données réelles. Ces tests ont été réalisés avec des tailles de batchs plus élevées pour augmenter la charge de travail du stockage et couvrir le pire des cas possibles.

Notez que dans chaque metric, la charge de travail totale générée par chaque modèle avec 32 GPU se situe bien dans la fourchette de performance du système AFF A800. Comme cadre de référence pour la charge de travail d'entraînement, une charge de travail artificielle a été générée en utilisant des E/S flexibles avec un profil d'E/S de lecture séquentielle de 64 000. Pour la charge de travail générée avec des E/S flexibles, le débit culminait à plus de 15 Gbit/s, la latence de lecture restait nettement inférieure à 1 ms et le taux d'utilisation du processeur était légèrement inférieur à 50 %. Pour atteindre le débit maximal possible avec le nombre limité de serveurs DGX-1 disponibles, des montages NFS supplémentaires et plusieurs tâches d'E/S flexibles ont été utilisées sur chaque serveur.

Remarque : Les tests montrent qu'une paire HA NetApp AFF A800 prend en charge un débit pouvant atteindre 25 Gbit/s avec une latence inférieure à 1 ms pour les charges de travail NAS.

Figure 13) Bande passante de stockage pour tous les modèles.

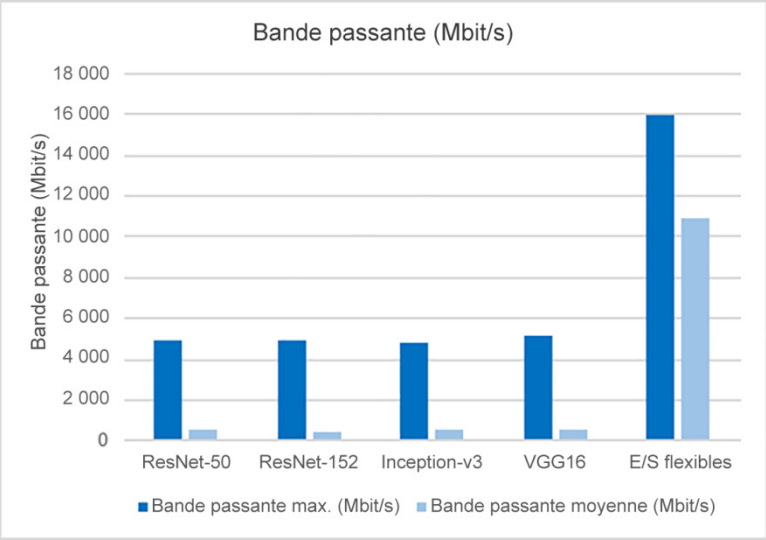


Figure 14) Latence de stockage pour tous les modèles.

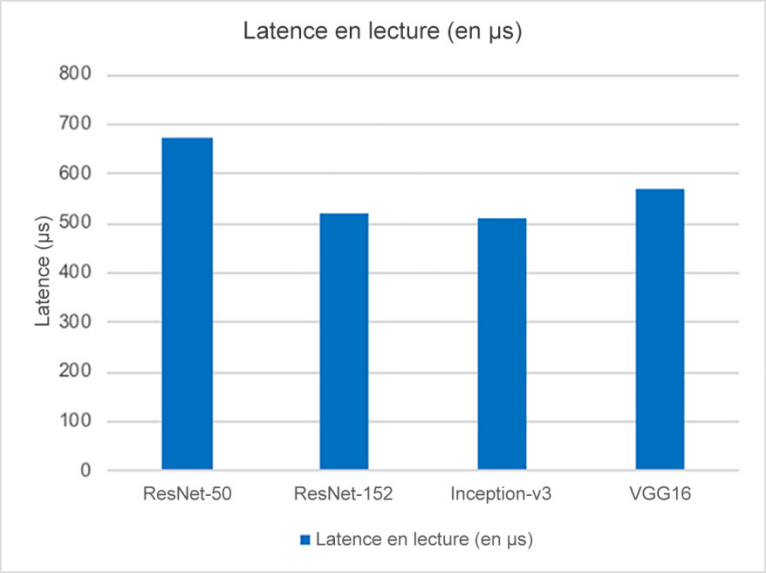
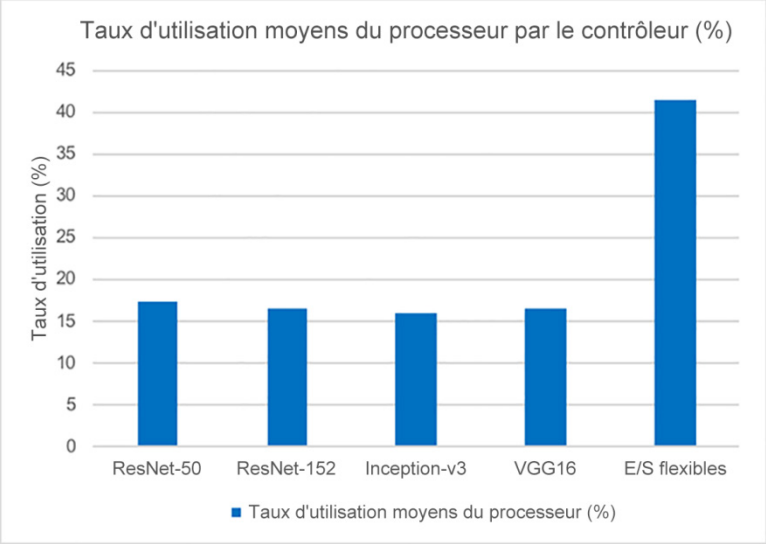


Figure 15) Taux d'utilisation du processeur de stockage pour tous les modèles.



7.3 Conseils de dimensionnement de la solution

Cette architecture doit servir de référence aux clients et partenaires qui souhaitent mettre en œuvre une infrastructure HPC avec des serveurs NVIDIA DGX-1 et un système NetApp AFF.

Comme le montre cette validation, le système AFF A800 prend facilement en charge le workload d'entraînement d'apprentissage profond généré par quatre serveurs DGX-1, avec une marge d'environ 70 % sur la paire HA. Par conséquent, le système AFF A800 est également capable de prendre en charge des serveurs DGX-1 supplémentaires. Pour des déploiements de plus grande envergure avec des exigences de performances de stockage encore plus élevées, des systèmes AFF A800 supplémentaires peuvent être ajoutés au cluster NetApp ONTAP. ONTAP 9 prend en charge jusqu'à 12 paires HA (24 nœuds) dans un seul cluster. Avec la technologie FlexGroup validée dans cette solution, il peut atteindre plus de 20 Po dans un seul volume. Le dataset que nous avons utilisé dans cette validation était relativement petit. Cependant, ONTAP 9 peut augmenter sa capacité de manière considérable avec une évolution linéaire des performances, car chaque paire HA offre des performances comparables au niveau vérifié dans ce document.

Pour les clusters DGX-1 plus petits, un système AFF A220 ou AFF A300 se révèle suffisamment performant pour un prix inférieur. Comme ONTAP 9 prend en charge les clusters à modèles mixtes, il est possible de commencer avec une empreinte réduite et d'ajouter au cluster des systèmes de stockage plus nombreux ou plus grands selon l'évolution des besoins en capacité et en performance.

Du point de vue du réseau, cette architecture vérifiée n'utilise que 16 des 32 ports disponibles sur chaque switch Nexus 3232C. Chaque switch prend en charge jusqu'à huit serveurs DGX-1 avec des ports d'accès au stockage supplémentaires pour augmenter considérablement la puissance de calcul sans ressource réseau supplémentaire. Pour des mises en œuvre plus importantes, le Cisco Nexus 7000 prend en charge jusqu'à 192 ports à 100 GbE par switch. Vous pouvez aussi implémenter une topologie Leaf-Spine avec plusieurs paires de switchs Nexus 3000 reliés à un switch central.

Sur la base des tests de validation effectués avec cette charge de travail d'IA, chaque DGX-1 nécessite un débit d'environ 2 Gbit/s. Comme il a été démontré que le système AFF A800 peut offrir un débit de 25 Gbit/s avec une charge de travail similaire générée par d'autres moyens, cette architecture pourrait prendre en charge neuf serveurs DGX-1 ou plus par paire HA AFF A800.

8 Conclusion

Le serveur DGX-1 est une plateforme d'apprentissage profond extrêmement puissante qui bénéficie d'une infrastructure de stockage et réseau tout aussi puissante pour des performances maximales. En combinant les systèmes NetApp AFF avec les switchs Cisco Nexus, cette architecture vérifiée peut être implémentée dans presque toutes les tailles, d'un seul serveur DGX-1 avec un système AFF A220 jusqu'à 96 serveurs DGX-1 avec un cluster AFF A800 à 12 nœuds. Associés aux fonctionnalités SDS et d'intégration au cloud de NetApp ONTAP, les systèmes AFF couvrent l'ensemble des pipelines de traitement de données de la périphérie au cœur jusqu'au cloud pour l'apprentissage profond.

Remerciements

Nous remercions nos collègues de NVIDIA, Darrin Johnson, Tony Paikeday, Robert Sohigian et James Mauro pour leurs contributions à cette architecture vérifiée NetApp. Nous n'aurions pas pu réaliser cette étude sans le soutien et les conseils des principaux membres de l'équipe NetApp, Robert Franz et Kesari Mishra.

Nous tenons à remercier sincèrement tous ces contributeurs dont les connaissances et l'expertise de pointe ont étayé les recherches menées dans le cadre de cette validation.

Sources d'informations complémentaires

Pour en savoir plus sur les informations données dans ce document, consultez les ressources suivantes :

- Serveurs NVIDIA DGX-1
  - Serveurs NVIDIA DGX-1
    https://www.nvidia.fr/data-center/dgx-1/
  - GPU NVIDIA Tesla V100 à cœurs Tensor
    https://www.nvidia.fr/data-center/tesla-v100/
  - NVIDIA GPU Cloud
    https://www.nvidia.fr/gpu-cloud/
- Systèmes NetApp AFF
  - Fiche technique AFF
    https://www.netapp.com/fr/media/ds-3582.pdf
  - NetApp Flash Advantage for AFF
    https://www.netapp.com/us/media/ds-3733.pdf

  - Documentation ONTAP 9.x
    http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286
  - Rapport technique NetApp FlexGroup
    https://www.netapp.com/fr/media/tr-4557.pdf
- Matrice d'interopérabilité NetApp :
  - NetApp Interoperability Matrix Tool
    http://support.netapp.com/matrix
- Connectivité réseau Cisco Nexus

  Les liens suivants fournissent des informations supplémentaires sur les switchs de la gamme Cisco Nexus 3232C :
  - switchs de la gamme Cisco Nexus 3232C
    https://www.cisco.com/c/en/us/products/switches/nexus-3232c-switch/index.html
  - Guide de configuration Cisco Nexus 3232C
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-installation-and-configuration-guides-list.html
  - Référence de ligne de commande Cisco Nexus 3232C
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-command-reference-list.html
- Structure d'apprentissage machine (ML) :
  - TensorFlow : Un framework d'apprentissage machine Open source pour tous
    https://www.tensorflow.org/
  - Horovod Le framework d'apprentissage profond Open source distribué d'Uber pour TensorFlow
    https://eng.uber.com/horovod/
  - Optimisation des GPU dans l'écosystème d'exécution du conteneur
    https://devblogs.nvidia.com/gpu-containers-runtime/
- Dataset et bancs d'essai :
  - ImageNet
    http://www.image-net.org/
  - Bancs d'essai TensorFlow
    https://www.tensorflow.org/performance/benchmarks

Annexe

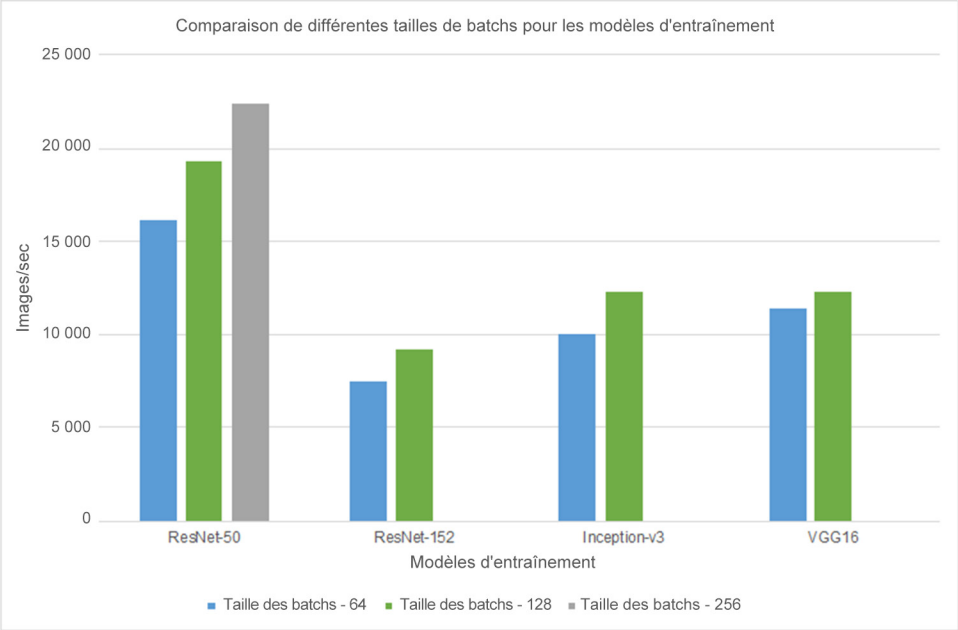
Cette section contient des résultats supplémentaires pour les tests réalisés avec cette architecture.

Taux d'entraînement pour différentes tailles de batchs pour chaque modèle

La Figure 16 montre une comparaison des différentes tailles de batchs pour les différents modèles de formation utilisant les composants suivants :

- Nombre de GPU : 32 (4 serveurs DGX-1)
- Cœurs : Tensor Tensor cores
- Batch sizes: Tailles de batchs : 64 128 et 256 pour ResNet-50, 64 et 128 pour les autres modèles

Figure 16) Comparaison des tailles de batchs pour les modèles d'entraînement.



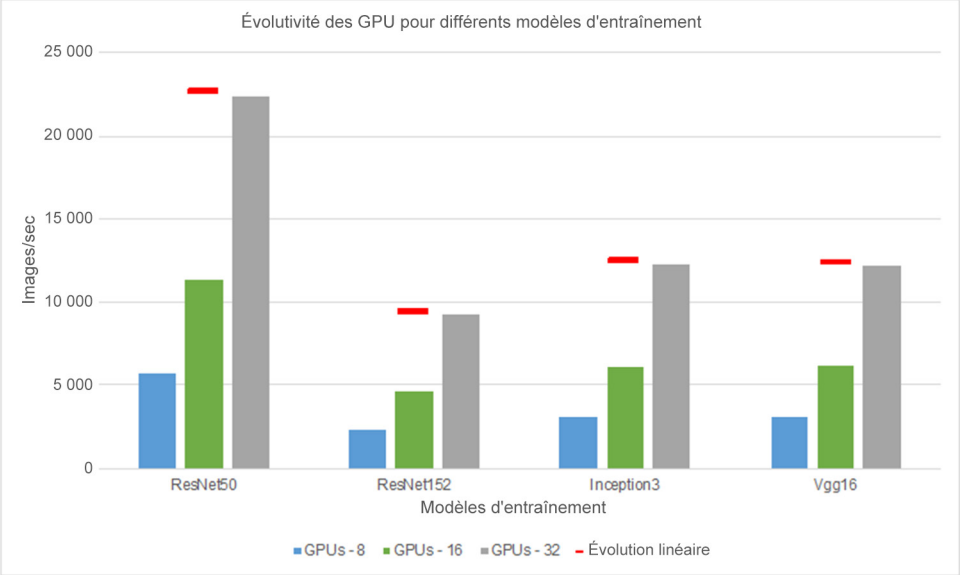
Conclusion : le débit d'entraînement augmente lorsque la taille de batchs passe à 256 ou 128.

Comparaison de l'évolutivité des GPU pour chaque modèle

La Figure 17 montre l'évolutivité des GPU pour les modèles d'entraînement utilisant les composants suivants :

- Nombre de GPU : 8 (1 serveur DGX-1), 16 (2 serveurs DGX-1) et 32 (4 serveurs DGX-1)
- Cœurs : Tensor Tensor cores
- Batch sizes: Tailles des batchs : 256 pour ResNet-50 et 128 pour les autres modèles

Figure 17) Évolutivité des GPU pour différents modèles d'entraînement.



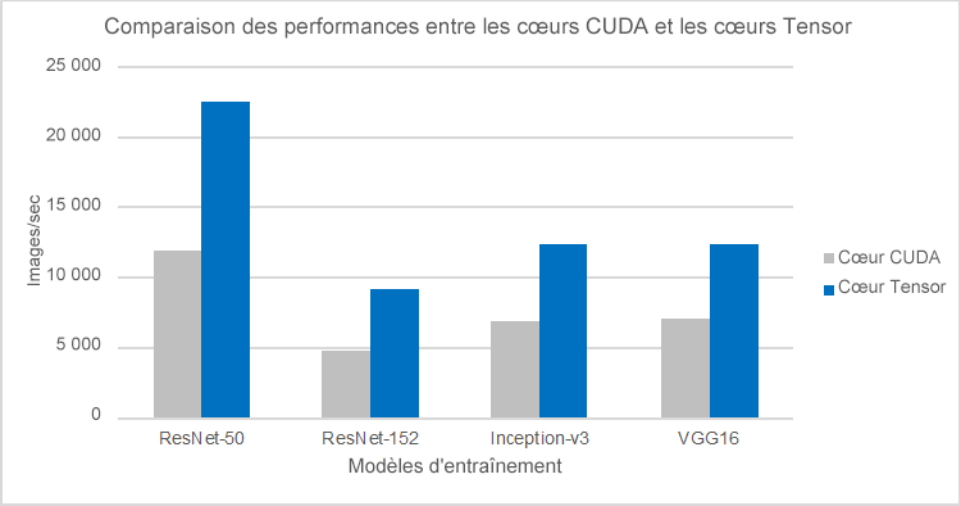
- Conclusion : l'évolutivité linéaire des GPU est observée dans tous les modèles d'entraînement.

Comparaison des cœurs Tensor et CUDA

La Figure 18 montre une comparaison des performances entre les cœurs CUDA et les cœurs Tensor utilisant les composants suivants :

- Nombre de GPU : 32 (4 serveurs DGX-1)
- Cœurs : Tensor et CUDA
- Tailles des batchs : 256 pour ResNet-50 et 128 pour les autres modèles

Figure 18) Comparaison des performances des cœurs CUDA et des cœurs Tensor.



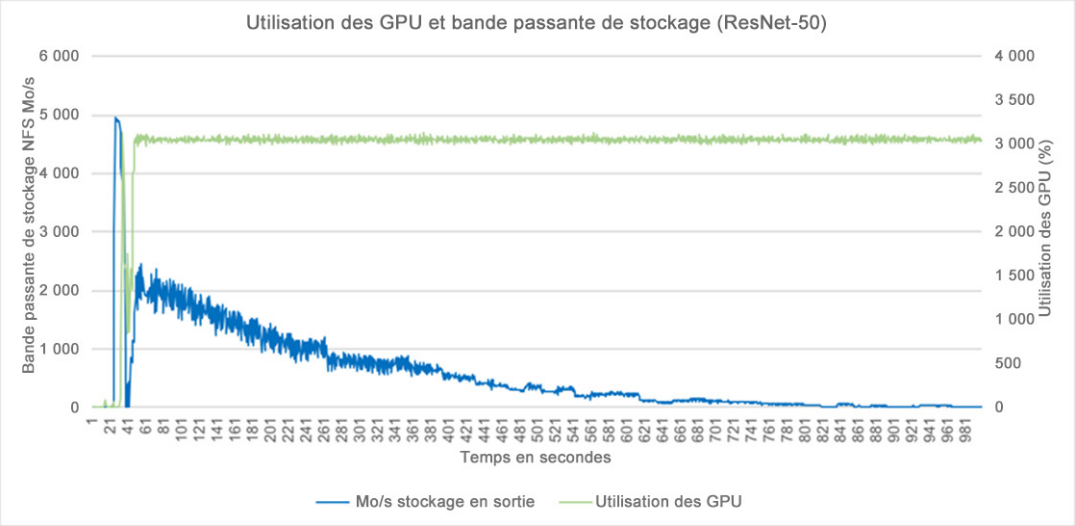
Conclusion: les cœurs Tensor offrent des performances supérieures à celles des cœurs CUDA.

Charge de travail des GPU pour tous les modèles

Les Figure 19, Figure 20 et Figure 21 montrent le taux d'utilisation et la bande passante du processeur graphique pour les modèles ResNet-50, ResNet-152 et Inception-v3, qui ont utilisé les composants suivants :

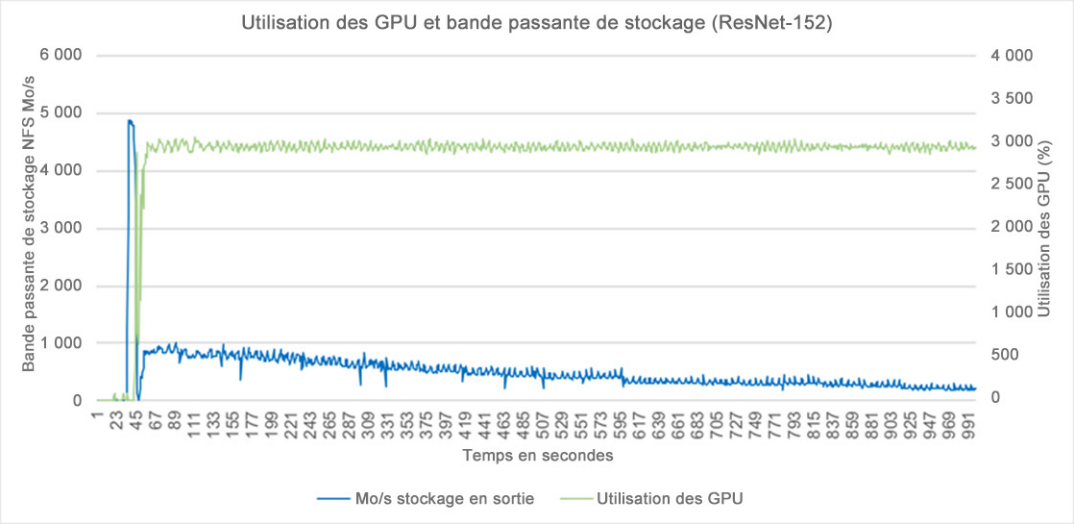
- Nombre de GPU : 32 (4 serveurs DGX-1)
- Cœurs : Tensor Tensor cores
- Tailles des batchs : 256 pour ResNet-50 et 128 pour les autres modèles

Figure 19) Utilisation des GPU et bande passante de stockage pour ResNet-50.



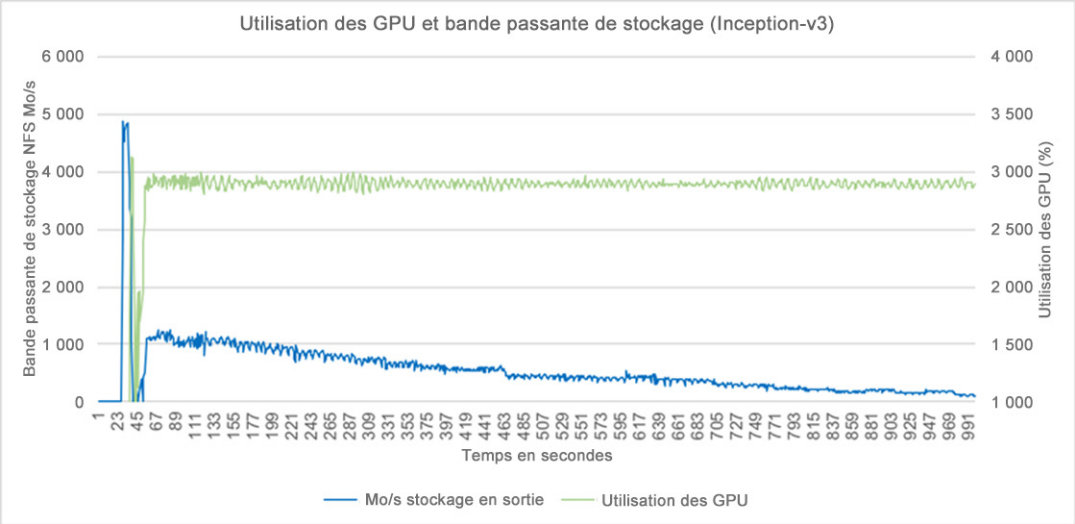
La Figure 20 montre le taux d'utilisation des GPU et la bande passante pour ResNet-152.

Figure 20) Taux d'utilisation des GPU et bande passante de stockage pour ResNet-152.



La Figure 21 montre le taux d’utilisation des GPU et la bande passante pour Inception-v3.

Figure 21) Taux d’utilisation des GPU et bande passante de stockage pour Inception-v3.



Reportez-vous à la matrice d'interopérabilité (IMT) sur le site de support NetApp pour vous assurer que les versions de produits et de fonctionnalités mentionnées dans le présent document sont prises en charge par votre environnement. La matrice d'interopérabilité de NetApp définit les composants et les versions de produits qu'il est possible d'utiliser pour créer des configurations prises en charge par NetApp. Les résultats dépendent des installations de chaque client et de leur conformité aux spécifications publiées.

Informations sur le copyright

Informations sur les marques commerciales