

Comment la mémoire persistante peut-elle profiter aux applications d'intelligence artificielle et de Machine Learning ?

Les centres de données devraient profiter des avantages de la mémoire rémanente pour éliminer les goulets d'étranglement d'entrée et de sortie et accélérer les performances des technologies de l'intelligence artificielle et du Machine Learning.



par **Arthur Sainio**, directeur du Marketing produit, **SMART Modular Technologies**

Dans les centres de données des entreprises actuels, la faible capacité de mémoire et les performances d'entrée/sortie (E/S) des périphériques de stockage de masse constituent les deux principales causes de ralentissement. Ces deux points noirs ont été considérés tour à tour selon différentes perspectives informatiques : la mémoire est un espace de stockage temporaire pour le code et les données, permettant de faire tourner une application, tandis que les disques durs et autres périphériques de stockage permanent permettent de conserver des données sur le long terme. Quand une application doit accéder à des données stockées sur un disque (ce qui se produit fréquemment avec les ensembles volumineux de données qui ne peuvent être stockés en mémoire vive), la lenteur du temps d'accès freine significativement les performances de l'application. L'apparition de la mémoire rémanente a marqué un tournant dans la hiérarchisation de la mémoire et du stockage traditionnels dans les centres de données grâce à l'émergence d'une nouvelle architecture hyperconvergente unifiée, qui accélère de façon significative les performances des serveurs de stockage.

Croissance des applications IA et ML

La prolifération des données a entraîné un essor considérable des applications d'intelligence artificielle (IA) et de Machine Learning (ML, soit « apprentissage machine »), alors que les systèmes traditionnels ne sont pas conçus pour permettre d'accéder à des ensembles de données aussi volumineux. La principale difficulté que rencontrent les applications IA et ML pour se généraliser, est le temps global de recherche et de fourniture d'informations basé sur des solutions ETL (*Extract, Transform, Load*), très consommatrices en données, et des opérations de *checkpointing* (sauvegarde régulière de l'état du système). Les applications IA et ML créent des E/S et des performances de calcul particulièrement exigeantes pour les solutions ETL accélérées par GPU (unité de traitement graphique). La variation des E/S et les performances de calcul sont déterminées par la bande passante et la latence. Les analyses de données ultra performantes, nécessaires pour les applications IA et ML, exigent des systèmes avec la meilleure bande passante et la plus faible latence possibles.

Selon une étude publiée par l'International Data Corporation (IDC), à savoir le [Worldwide Artificial Intelligence Spending Guide](#), les dépenses concernant les systèmes IA et ML atteindront 97,9 milliards de dollars en 2023, soit 2 fois et demie les 37,5 milliards qui devraient y être consacré en 2019. De leur côté, les besoins nécessaires pour le traitement des données enregistrant une telle expansion devraient connaître une croissance exponentielle. Les solutions de mémoire conventionnelles actuelles manquent d'une composante essentielle pour répondre à cette croissance : la non-volatilité, y compris comme architecture parallèle, va s'imposer comme la réponse aux futurs besoins des centres de données. Cependant, avec ces nouvelles architectures, les coupures d'alimentation pourraient engendrer des pertes se chiffrant à plusieurs millions de dollars pour les centres de données. C'est pourquoi le besoin en mémoire non-volatile est imminent.

Placer la mémoire non-volatile au plus près des CPU

Les opérations de *checkpointing* consistent à sauvegarder l'état du réseau en cours d'apprentissage, pour s'assurer que le résultat des données apprises n'est pas perdu en cas de coupure de l'alimentation. Ces opérations de *checkpointing* représentent un défi pour les applications IA et ML car elles utilisent la capacité de traitement et consomment beaucoup d'énergie, sans offrir aucun avantage direct à l'application en elle-même. Les traitements effectués dans d'autres nœuds du réseau peuvent également être arrêtés lors de l'écriture de données dans un emplacement central. Ce type d'opération demande également beaucoup d'écritures, ce qui aggrave le problème dans certaines situations, car le stockage conventionnel, tels que les disques durs, se révèle inefficace quand des données y sont en cours d'écriture.

Comme les opérations de *checkpointing* sur une mémoire centrale peuvent réduire significativement la vitesse utilisée par les applications IA et ML, les ingénieurs cherchent désormais à placer la mémoire non-volatile au plus près du CPU, afin de minimiser l'impact de ce processus essentiel. On atteint ainsi un meilleur équilibre entre les data et le calcul, ce qui permet au système de fournir la puissance générale requise.

Les mémoires NVDIMM sur les applications IA et ML

La mémoire rémanente, sous forme de module de mémoire DDIMM non-volatile (NVDIMM, soit *Non-Volatile Dual-Inline Memory Module*), est utilisée pour accroître les performances des applications sensibles à la latence d'écriture, offrant un modèle de stockage persistant efficace avec les performances d'une DRAM. Les centres de données ont une chance unique de tirer parti des modules NVDIMM pour atteindre une faible latence et les exigences de performances accrues des applications IA et ML, sans perturbation technologique majeure.

Penguin Relion 4118GTS – AI / ML Performance at Scale	
Processor/ Chipset	(2) Intel® Xeon® Scalable Processor family, TDP up to 205W *Support for Cascade Lake
Data and Storage Layer	(24) DIMM slots, 6 DPC, DDR4 - DRAM - NVDIMM - SCM (10) 2.5" hot-swappable HDD/SSD bays - (4) U.2 (Secure) NVMe devices only - (6) 2.5" (Secure) SATA/SAS devices
PCIe Accelerated GPU and Networking	(8) NVIDIA V100 SXM2 w/ NVLINK (4) 100G Low Latency High Speed Network (2) 25GbE Ethernet
Workloads and Verticals	     

Figure 1. Exemple d'un serveur à mémoire persistante optimisé pour les applications IA et ML.

Quand des mémoires NVDIMM sont connectées à un serveur, elles sont reliées par le BIOS à une sous-section de mémoire rémanente en sein de la mémoire principale. L'application est alors libre d'utiliser cette mémoire rémanente pour accélérer l'opération de *checkpointing*. L'approche traditionnelle constitue une autre solution dans laquelle les données de checkpoint sont transférées via la pile d'E/S, sur le module NVM, puis sauvegardées dans un SSD. Cette méthode a une incidence sur la latence de la pile d'E/S et de la Flash NAND.

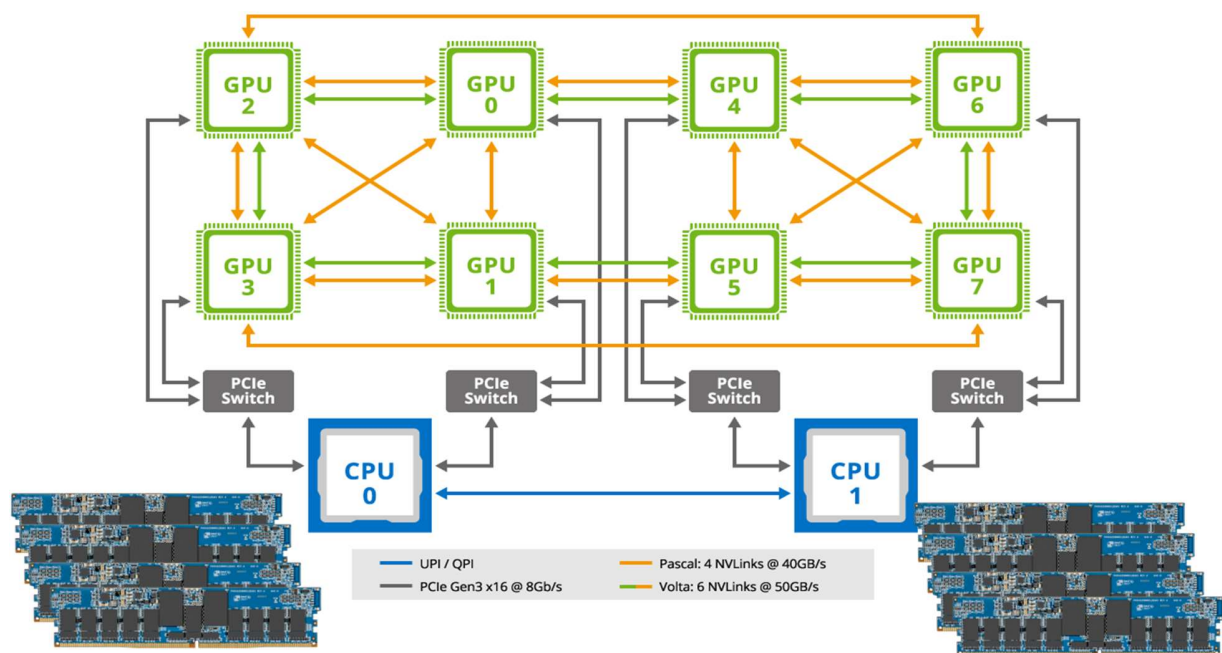


Figure 2. Quatre modules NVDIMM 32 Go sont utilisés pour chaque processeur, fournissant la totalité de la mémoire rémanente et rapide, adressable par octet.

Les modules NVDIMM constituent une solution idéale pour les serveurs IA et ML ultra performants. Les solutions ETL très consommatrices en données et les opérations de *checkpointing* peuvent utiliser la zone de mémoire rémanente de la mémoire principale, ce qui leur permet d'opérer avec une latence (<100 ns) et une bande passante (25,6 Go/s) équivalentes à une DRAM.

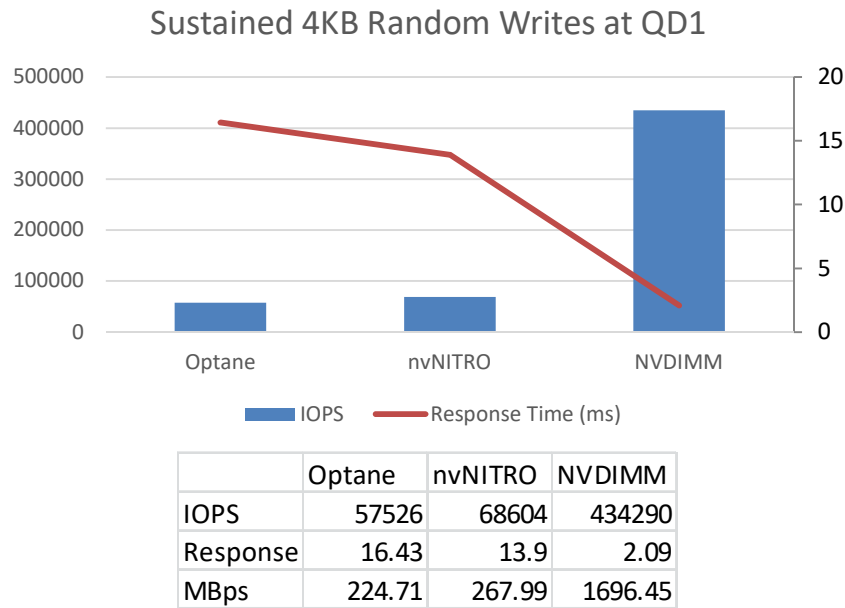


Figure 3. Graphique montrant les résultats des tests de comparaison ezFIO entre un module Intel® Optane™ NVMe 2.5" SSD, un module MRAM NVMe U.2 SSD et un module NVDIMM.

Tandis que les modules NVDIMM sont utilisés pour accélérer les opérations de *checkpointing* pour les applications IA, ils peuvent également se révéler utiles pour le ML, pour accroître les performances et protéger les données collectées par les algorithmes. Les serveurs de stockage configurés par unité GPU fonctionnent avec des algorithmes, tout comme la simulation et le ML. Les modules NVDIMM sont utilisés pour protéger les serveurs avec GPU contre la perte des données de simulation. Les tailles typiques des ensembles de données des algorithmes varient de quelques kilo-octets (ko) à plusieurs Téra-octets (To). C'est pourquoi toute perte de données entraînerait la nécessité de reprendre tout le travail à zéro. Lorsque quatre serveurs sont configurés avec des modules NVDIMM, les tailles des ensembles de données atteignant 1 To peuvent utiliser la mémoire rémanente, par opposition au mode de stockage traditionnel, et ainsi améliorer de façon spectaculaire les performances, sans aucun risque de perte de données.

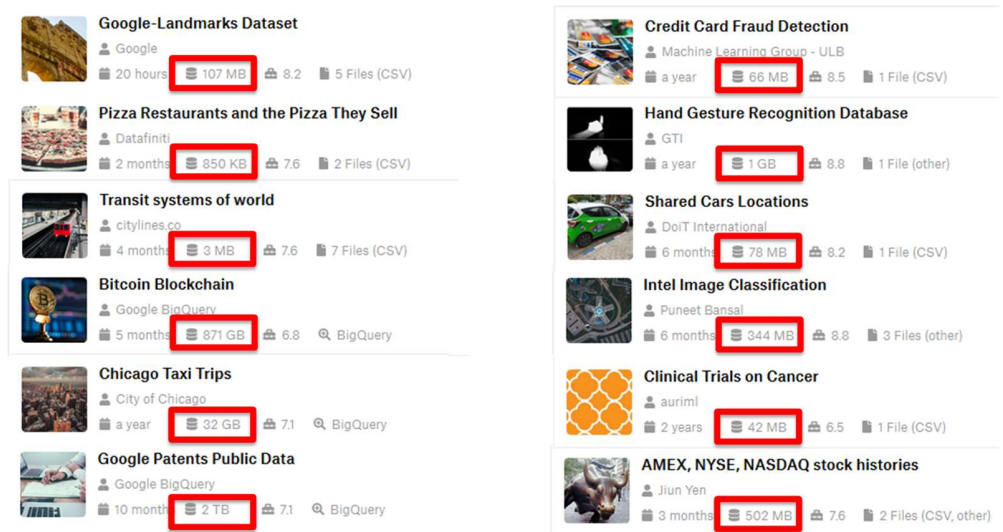


Figure 4. Exemples d'ensembles de données d'apprentissage machine (ML) allant de 850 ko à 2 To.

La méthode la plus fréquemment utilisée pour traiter les ensembles de données IA, ML et de simulation (qui possèdent tous des caractéristiques communes) consiste à envoyer les ensembles de données sur le réseau via InfiniBand ou Ethernet, vers le serveur IA/ML, où elles sont mises en cache dans le SSD, pour éliminer tout risque de perte de données. Une partie des données est ensuite placée dans la DRAM par l'unité GPU, où les calculs peuvent alors être effectués. Par exemple, ce procédé peut permettre d'effectuer des calculs sur un ensemble de données pour déterminer si ces données représentent l'image d'un chien ou d'un chat. Une fois le calcul effectué, la réponse est renvoyée vers le réseau. Cependant, en cas de panne système au cours du processus, tous les calculs sont perdus. En optant pour des modules NVDIMM, ce processus peut être amélioré de façon spectaculaire. Nul besoin de mettre les ensembles de données en mémoire cache dans les SSD. Les ensembles de données peuvent en effet être directement placés dans la DRAM, où l'unité GPU peut immédiatement commencer ses calculs. La réponse visant à déterminer si un ensemble de données spécifique représente l'image d'un chien ou d'un chat peut intervenir dans un délai bien plus rapide. En même temps, il n'y a aucun risque de perdre ni les données ni les calculs, car les modules NVDIMM sont rémanents.

Cependant, les modules NVDIMM ne sont pas seulement adaptés aux applications IA et ML : ils peuvent également être utilisés pour les applications financières couramment appelées FinTech. Les applications FinTech exigent un degré de performances très élevé (latence réduite et vitesses de transaction élevées) car « le temps, c'est de l'argent ». Les transactions traitées doivent être enregistrées de façon chronologique et synchronisée avant que la transaction suivante puisse être lancée. Cette fonction d'horodatage synchronisée, tout en étant critique en cas d'audit, crée par ailleurs un goulet d'étranglement d'entrée/sortie significatif pour de nombreux systèmes, ralentissant la vitesse des transactions. En utilisant les modules NVDIMM, le processus actuellement employé, consistant à enregistrer les données sur un disque SATA ou des modules NVM SSD, peut être éliminé. Au lieu d'envoyer les données du journal via des E/S vers le Flash SSD, elles peuvent être stockées directement

dans une DRAM ultra rapide rendue rémanente grâce à l'utilisation de modules NVDIMM. Les modules NVDIMM permettent au système de commencer la transaction suivante avec la certitude que l'horodatage de la transaction précédente est enregistré dans un emplacement sécurisé, sans aucun risque de perte des données.

Alors que les modules NVDIMM ont fait leur apparition il y a plus de 10 ans, les avantages de l'utilisation de ce type de mémoire rémanente pour les applications IA et ML font toujours l'objet de recherches dans différents secteurs comme ceux de la banque et de la distribution, les industries de transformation et de production, la santé et les services professionnels. L'écosystème qui soutient les modules NVDIMM, y compris les systèmes d'exploitation, la compatibilité matérielle et la normalisation JEDEC, est le fruit du travail commun de nombreuses entreprises pour favoriser l'adoption de la mémoire rémanente. Les modules NVDIMM, associés à l'essor de l'IA et du ML, tracent une voie idéale pour augmenter les performances système.