Inférence d'IA Edge avec NetApp et Lenovo

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Sommaire

Inférence d’IA Edge avec NetApp et Lenovo

TR-4886 : Inférence IA en périphérie – NetApp avec Lenovo ThinkSystem – Conception de solutions

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Ce document décrit une architecture de calcul et de stockage pour déployer l’inférence d’intelligence artificielle (IA) basée sur le GPU sur les contrôleurs de stockage NetApp et les serveurs Lenovo ThinkSystem dans un environnement de périphérie qui répond aux scénarios d’application émergents.

Résumé

Plusieurs scénarios d’application émergents, tels que les systèmes avancés d’assistance à la conduite (ADAS), l’industrie 4.0, les villes intelligentes et l’Internet des objets (IoT), nécessitent le traitement de flux de données continus avec une latence proche de zéro. Ce document décrit une architecture de calcul et de stockage pour déployer l’inférence d’intelligence artificielle (IA) basée sur le GPU sur les contrôleurs de stockage NetApp et les serveurs Lenovo ThinkSystem dans un environnement de périphérie qui répond à ces exigences. Ce document fournit également des données de performances pour le benchmark d’inférence MLPerf standard de l’industrie, évaluant diverses tâches d’inférence sur des serveurs Edge équipés de GPU NVIDIA T4. Nous étudions les performances des scénarios d’inférence hors ligne, à flux unique et à flux multiples et montrons que l’architecture avec un système de stockage en réseau partagé rentable est très performante et fournit un point central pour la gestion des données et des modèles pour plusieurs serveurs périphériques.

Introduction

Les entreprises génèrent de plus en plus de volumes massifs de données à la périphérie du réseau. Pour tirer le meilleur parti des capteurs intelligents et des données IoT, les organisations recherchent une solution de streaming d’événements en temps réel qui permet l’informatique de pointe. Les tâches exigeantes en termes de calcul sont donc de plus en plus effectuées en périphérie, en dehors des centres de données. L’inférence de l’IA est l’un des moteurs de cette tendance. Les serveurs Edge fournissent une puissance de calcul suffisante pour ces charges de travail, en particulier lors de l’utilisation d’accélérateurs, mais le stockage limité est souvent un problème, en particulier dans les environnements multiserveurs. Dans ce document, nous montrons comment vous pouvez déployer un système de stockage partagé dans l’environnement Edge et comment il profite aux charges de travail d’inférence IA sans imposer de pénalité de performances.

Ce document décrit une architecture de référence pour l’inférence de l’IA à la périphérie. Il combine plusieurs serveurs Edge Lenovo ThinkSystem avec un système de stockage NetApp pour créer une solution facile à déployer et à gérer. Il est destiné à être un guide de base pour les déploiements pratiques dans diverses situations, telles que l’usine avec plusieurs caméras et capteurs industriels, les systèmes de point de vente (POS) dans les transactions de détail ou les systèmes de conduite entièrement autonome (FSD) qui identifient les anomalies visuelles dans les véhicules autonomes.

Ce document couvre les tests et la validation d’une configuration de calcul et de stockage composée d’un serveur Lenovo ThinkSystem SE350 Edge et d’un système de stockage NetApp AFF et EF-Series d’entrée de gamme. Les architectures de référence offrent une solution efficace et rentable pour les déploiements d’IA tout en fournissant des services de données complets, une protection des données intégrée, une évolutivité transparente et un stockage de données connecté au cloud avec les logiciels de gestion de données NetApp ONTAP et NetApp SANtricity .

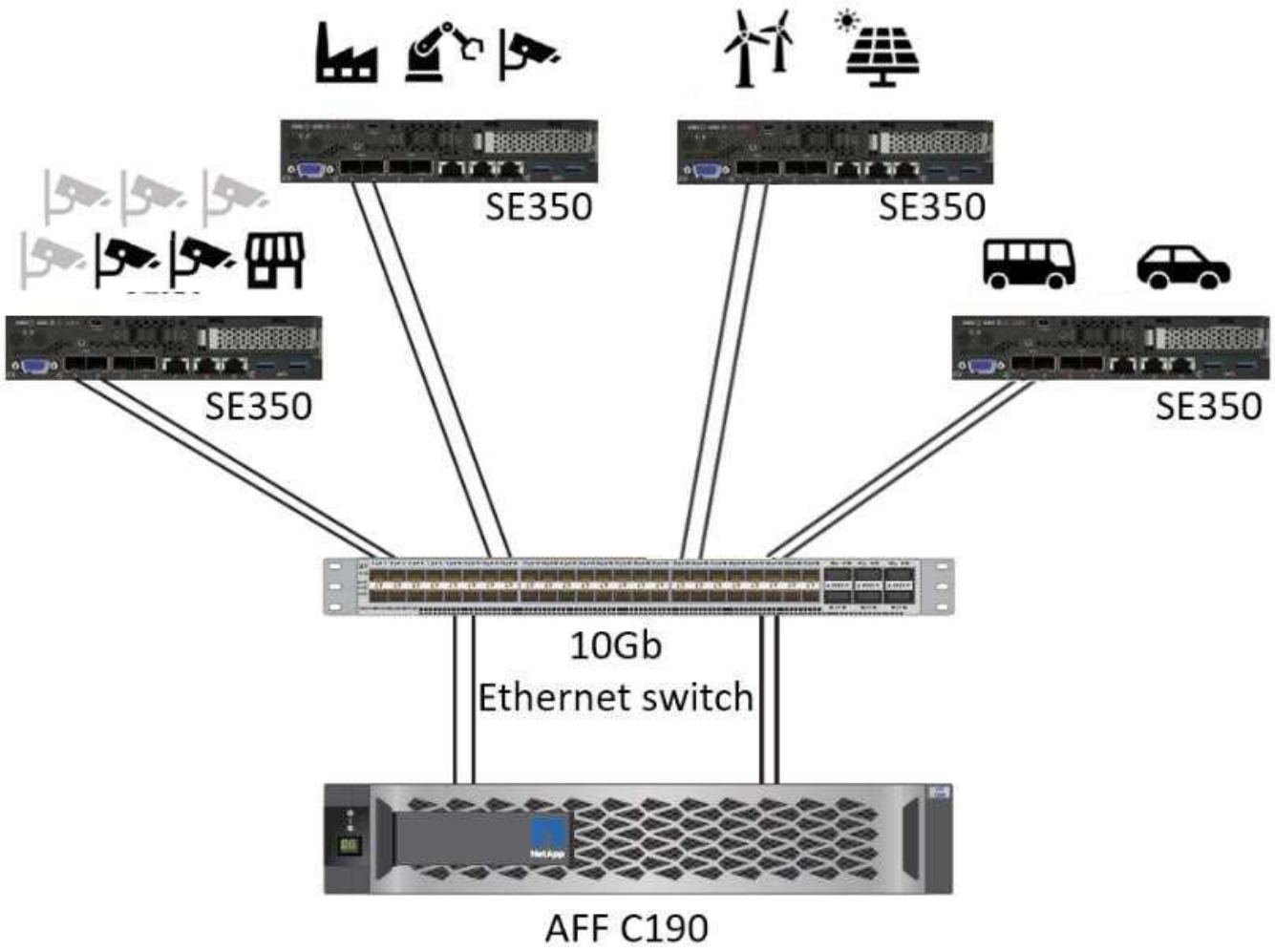
Public cible

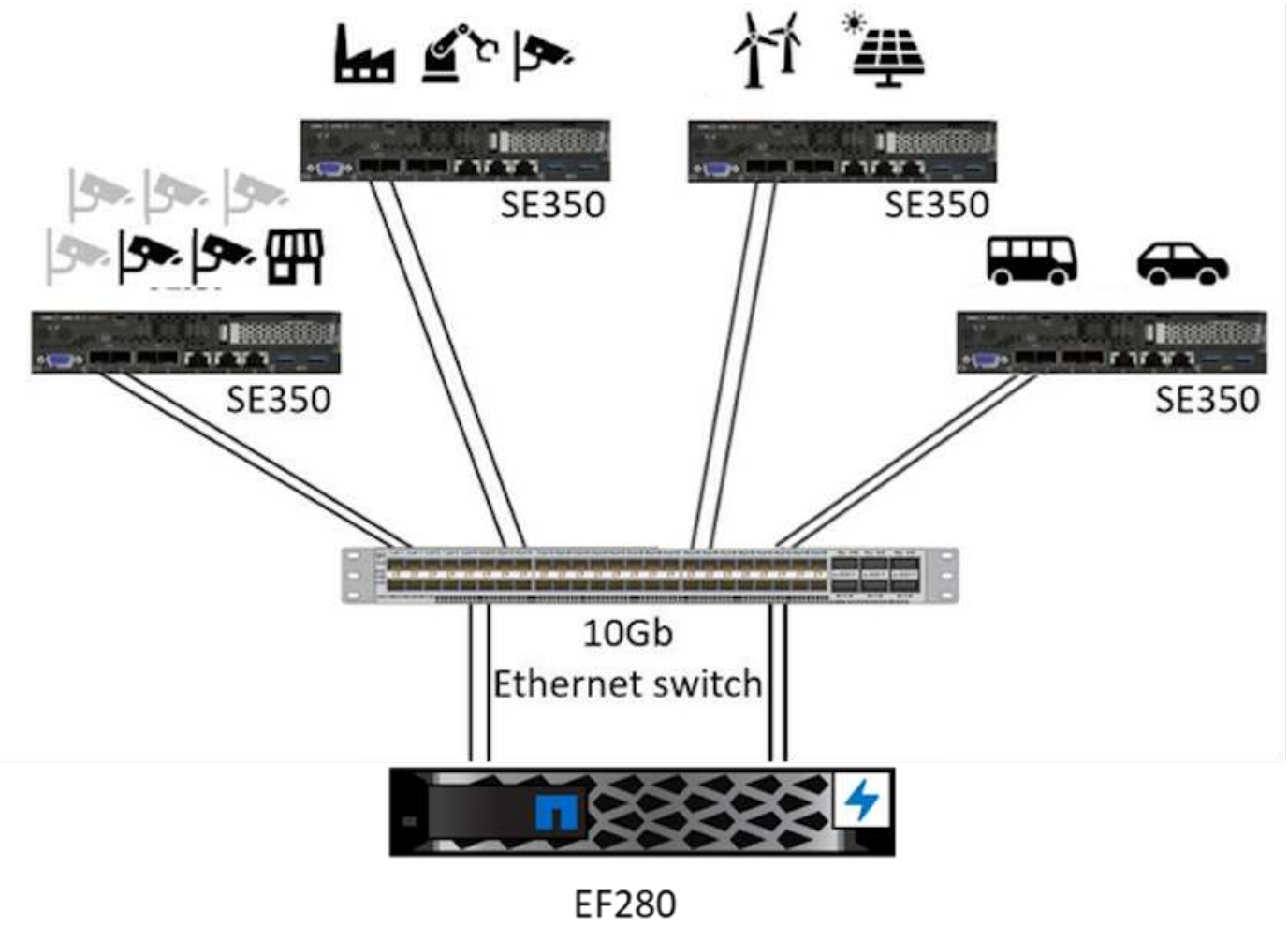
Ce document est destiné aux publics suivants :

- Les chefs d'entreprise et les architectes d'entreprise qui souhaitent produire l'IA à la périphérie.
- Scientifiques des données, ingénieurs des données, chercheurs en IA/apprentissage automatique (ML) et développeurs de systèmes d'IA.
- Architectes d'entreprise qui conçoivent des solutions pour le développement de modèles et d'applications d'IA/ML.
- Les scientifiques des données et les ingénieurs en IA recherchent des moyens efficaces de déployer des modèles d'apprentissage profond (DL) et de ML.
- Gestionnaires de périphériques Edge et administrateurs de serveurs Edge responsables du déploiement et de la gestion des modèles d'inférence Edge.

Architecture de la solution

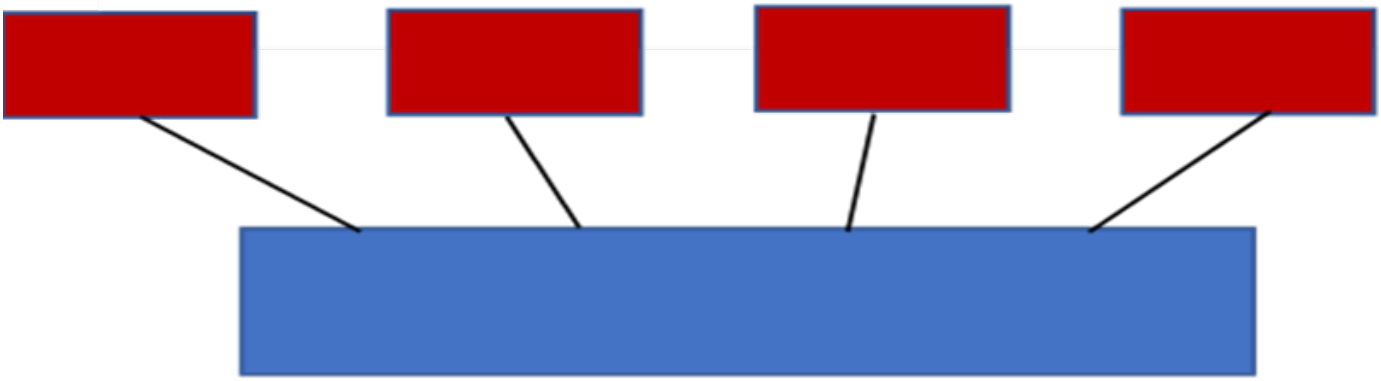
Cette solution de serveur Lenovo ThinkSystem et de stockage NetApp ONTAP ou NetApp SANtricity est conçue pour gérer l'inférence IA sur de grands ensembles de données en utilisant la puissance de traitement des GPU aux côtés des CPU traditionnels. Cette validation démontre des performances élevées et une gestion optimale des données avec une architecture qui utilise un ou plusieurs serveurs Edge Lenovo SR350 interconnectés avec un seul système de stockage NetApp AFF , comme illustré dans les deux figures suivantes.





L’aperçu de l’architecture logique dans la figure suivante montre les rôles des éléments de calcul et de stockage dans cette architecture. Plus précisément, il montre ce qui suit :

- Appareils de calcul Edge effectuant des inférences sur les données qu’ils reçoivent des caméras, des capteurs, etc.
- Un élément de stockage partagé qui sert à plusieurs fins :
  - Fournit un emplacement central pour les modèles d’inférence et d’autres données nécessaires pour effectuer l’inférence. Les serveurs de calcul accèdent directement au stockage et utilisent des modèles d’inférence sur le réseau sans avoir besoin de les copier localement.
  - Les modèles mis à jour sont poussés ici.
  - Les archives saisissent les données que les serveurs Edge reçoivent pour une analyse ultérieure. Par exemple, si les périphériques périphériques sont connectés à des caméras, l’élément de stockage conserve les vidéos capturées par les caméras.



rouge	bleu
Système informatique Lenovo	Système de stockage NetApp AFF
Appareils de périphérie effectuant des inférences sur les entrées provenant de caméras, de capteurs, etc.	Stockage partagé contenant des modèles d'inférence et des données provenant d'appareils périphériques pour une analyse ultérieure.

Cette solution NetApp et Lenovo offre les principaux avantages suivants :

- Calcul accéléré par GPU à la périphérie.
- Déploiement de plusieurs serveurs Edge sauvegardés et gérés à partir d'un stockage partagé.
- Protection robuste des données pour atteindre les objectifs de points de récupération (RPO) et de temps de récupération (RTO) faibles sans perte de données.
- Gestion optimisée des données avec des copies et des clones NetApp Snapshot pour rationaliser les flux de travail de développement.

Comment utiliser cette architecture

Ce document valide la conception et les performances de l'architecture proposée. Cependant, nous n'avons pas testé certains éléments au niveau logiciel, tels que la gestion des conteneurs, des charges de travail ou des modèles et la synchronisation des données avec le cloud ou le centre de données sur site, car ils sont spécifiques à un scénario de déploiement. Ici, plusieurs choix existent.

Au niveau de la gestion des conteneurs, la gestion des conteneurs Kubernetes est un bon choix et est bien prise en charge soit dans une version entièrement en amont (Canonical), soit dans une version modifiée adaptée aux déploiements d'entreprise (Red Hat). Le"Plan de contrôle de l'IA NetApp" qui utilise NetApp Trident et le nouveau "Boîte à outils NetApp DataOps" fournit une traçabilité intégrée, des fonctions de gestion des données, des interfaces et des outils permettant aux scientifiques et aux ingénieurs de données de s'intégrer au stockage NetApp . Kubeflow, la boîte à outils ML pour Kubernetes, fournit des fonctionnalités d'IA supplémentaires ainsi qu'une prise en charge du contrôle de version des modèles et de KFServing sur plusieurs plates-formes telles que TensorFlow Serving ou NVIDIA Triton Inference Server. Une autre option est la plate-forme NVIDIA EGX, qui fournit une gestion de la charge de travail ainsi qu'un accès à un catalogue de conteneurs d'inférence d'IA compatibles GPU. Cependant, ces options peuvent nécessiter des efforts et une expertise considérables pour les mettre en production et peuvent nécessiter l'assistance d'un fournisseur de logiciels indépendant (ISV) ou d'un consultant tiers.

Domaines de solutions

Le principal avantage de l'inférence de l'IA et de l'informatique de pointe est la capacité des appareils à calculer, traiter et analyser des données avec un niveau de qualité élevé sans latence. Il existe beaucoup trop

d’exemples de cas d’utilisation de l’informatique de pointe pour les décrire dans ce document, mais en voici quelques-uns parmi les plus importants :

Automobiles : Véhicules autonomes

L’illustration classique de l’informatique de pointe se trouve dans les systèmes avancés d’assistance à la conduite (ADAS) dans les véhicules autonomes (VA). L’IA des voitures sans conducteur doit traiter rapidement de nombreuses données provenant de caméras et de capteurs pour être un conducteur sûr et performant. Prendre trop de temps pour interpréter la différence entre un objet et un humain peut signifier la vie ou la mort. Il est donc crucial de pouvoir traiter ces données au plus près du véhicule. Dans ce cas, un ou plusieurs serveurs de calcul Edge gèrent les entrées des caméras, du RADAR, du LiDAR et d’autres capteurs, tandis que le stockage partagé contient les modèles d’inférence et stocke les données d’entrée des capteurs.

Santé : Suivi des patients

L’un des plus grands impacts de l’IA et de l’informatique de pointe est sa capacité à améliorer la surveillance continue des patients atteints de maladies chroniques, tant dans les soins à domicile que dans les unités de soins intensifs (USI). Les données provenant des appareils de pointe qui surveillent les niveaux d’insuline, la respiration, l’activité neurologique, le rythme cardiaque et les fonctions gastro-intestinales nécessitent une analyse instantanée des données sur lesquelles il faut agir immédiatement car le temps est limité pour agir afin de sauver la vie de quelqu’un.

Commerce de détail : paiement sans caissier

L’informatique de pointe peut alimenter l’IA et le ML pour aider les détaillants à réduire le temps de paiement et à augmenter le trafic piétonnier. Les systèmes sans caissier prennent en charge divers composants, tels que les suivants :

- Authentification et accès. Connecter l’acheteur physique à un compte validé et lui permettre d’accéder à l’espace de vente.
- Suivi des stocks. Utilisation de capteurs, d’étiquettes RFID et de systèmes de vision par ordinateur pour aider à confirmer la sélection ou la désélection d’articles par les acheteurs.

  Ici, chacun des serveurs périphériques gère chaque comptoir de caisse et le système de stockage partagé sert de point de synchronisation central.

Services financiers : sécurité humaine aux guichets et prévention de la fraude

Les organisations bancaires utilisent l’IA et l’informatique de pointe pour innover et créer des expériences bancaires personnalisées. Les bornes interactives utilisant l’analyse de données en temps réel et l’inférence de l’IA permettent désormais aux distributeurs automatiques de billets non seulement d’aider les clients à retirer de l’argent, mais également de surveiller de manière proactive les bornes grâce aux images capturées par les caméras pour identifier les risques pour la sécurité humaine ou les comportements frauduleux. Dans ce scénario, les serveurs de calcul Edge et les systèmes de stockage partagés sont connectés à des bornes interactives et à des caméras pour aider les banques à collecter et à traiter des données avec des modèles d’inférence d’IA.

Fabrication : Industrie 4.0

La quatrième révolution industrielle (Industrie 4.0) a commencé, avec l’émergence de nouvelles tendances telles que l’usine intelligente et l’impression 3D. Pour préparer un avenir axé sur les données, la communication machine à machine (M2M) à grande échelle et l’IoT sont intégrés pour une automatisation accrue sans intervention humaine. La fabrication est déjà hautement automatisée et l’ajout de fonctionnalités d’IA est une continuation naturelle de la tendance à long terme. L’IA permet d’automatiser des opérations qui

peuvent être automatisées à l'aide de la vision par ordinateur et d'autres capacités d'IA. Vous pouvez automatiser le contrôle qualité ou les tâches qui reposent sur la vision humaine ou la prise de décision pour effectuer des analyses plus rapides des matériaux sur les chaînes de montage dans les usines afin d'aider les usines de fabrication à respecter les normes ISO requises en matière de sécurité et de gestion de la qualité. Ici, chaque serveur de calcul Edge est connecté à un ensemble de capteurs surveillant le processus de fabrication et les modèles d'inférence mis à jour sont poussés vers le stockage partagé, selon les besoins.

Télécommunications : détection de rouille, inspection des tours et optimisation du réseau

L'industrie des télécommunications utilise des techniques de vision par ordinateur et d'IA pour traiter des images qui détectent automatiquement la rouille et identifient les tours cellulaires qui contiennent de la corrosion et nécessitent donc une inspection plus approfondie. L'utilisation d'images de drones et de modèles d'IA pour identifier des régions distinctes d'une tour afin d'analyser la rouille, les fissures de surface et la corrosion a augmenté ces dernières années. La demande continue de croître pour les technologies d'IA qui permettent d'inspecter efficacement les infrastructures de télécommunication et les tours de téléphonie mobile, d'évaluer régulièrement leur dégradation et de les réparer rapidement si nécessaire.

En outre, un autre cas d'utilisation émergent dans les télécommunications est l'utilisation d'algorithmes d'IA et de ML pour prédire les modèles de trafic de données, détecter les appareils compatibles 5G et automatiser et augmenter la gestion de l'énergie à entrées et sorties multiples (MIMO). Le matériel MIMO est utilisé dans les tours radio pour augmenter la capacité du réseau ; cependant, cela entraîne des coûts énergétiques supplémentaires. Les modèles ML pour le « mode veille MIMO » déployés sur les sites cellulaires peuvent prédire l'utilisation efficace des radios et aider à réduire les coûts de consommation d'énergie pour les opérateurs de réseaux mobiles (MNO). Les solutions d'inférence d'IA et d'informatique de pointe aident les opérateurs de réseaux mobiles à réduire la quantité de données transmises dans les deux sens vers les centres de données, à réduire leur coût total de possession, à optimiser les opérations réseau et à améliorer les performances globales pour les utilisateurs finaux.

Conclusion

L'automatisation basée sur l'IA et l'informatique de pointe constituent une approche de pointe pour aider les entreprises à réaliser leur transformation numérique et à maximiser l'efficacité et la sécurité opérationnelles. Avec l'informatique de pointe, les données sont traitées beaucoup plus rapidement car elles n'ont pas besoin de voyager vers et depuis un centre de données. Par conséquent, le coût associé à l'envoi et au retour de données vers les centres de données ou le cloud est réduit. Une latence plus faible et une vitesse accrue peuvent être bénéfiques lorsque les entreprises doivent prendre des décisions en temps quasi réel à l'aide de modèles d'inférence d'IA déployés en périphérie.

Les systèmes de stockage NetApp offrent des performances identiques ou supérieures à celles du stockage SSD local et offrent les avantages suivants aux scientifiques des données, aux ingénieurs de données, aux développeurs d'IA/ML et aux décideurs commerciaux ou informatiques :

- Partage de données sans effort entre les systèmes d'IA, les analyses et d'autres systèmes commerciaux critiques. Ce partage de données réduit les frais généraux de l'infrastructure, améliore les performances et rationalise la gestion des données dans l'ensemble de l'entreprise.
- Calcul et stockage évolutifs indépendamment pour minimiser les coûts et améliorer l'utilisation des ressources.
- Flux de travail de développement et de déploiement rationalisés à l'aide de copies et de clones Snapshot intégrés pour des espaces de travail utilisateur instantanés et peu encombrants, un contrôle de version intégré et un déploiement automatisé.

- Protection des données de niveau entreprise pour la reprise après sinistre et la continuité des activités. La solution NetApp et Lenovo présentée dans ce document est une architecture flexible et évolutive, idéale pour les déploiements d'inférence d'IA de niveau entreprise à la périphérie.

Remerciements

- JJ Falkanger, directeur principal, solutions HPC et IA, Lenovo
- Dave Arnette, ingénieur marketing technique, NetApp
- Joey Parnell, responsable technique des solutions d'IA de la série E, NetApp
- Cody Harryman, ingénieur assurance qualité, NetApp

Où trouver des informations supplémentaires

Pour en savoir plus sur les informations décrites dans ce document, reportez-vous aux documents et/ou sites Web suivants :

- Page produit des baies NetApp AFF série A

  "https://www.netapp.com/data-storage/aff-a-series/"

- Logiciel de gestion de données NetApp ONTAP — Bibliothèque d'informations ONTAP 9

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- TR-4727 : Présentation de la série NetApp EF

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- Fiche technique du logiciel SANtricity NetApp E-Series

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- Stockage persistant NetApp pour conteneurs — NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- Copie et synchronisation NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Test de référence TensorFlow

  "https://github.com/tensorflow/benchmarks"

- Serveur Edge Lenovo ThinkSystem SE350

  "https://lenovopress.com/lp1168"

- Baie de stockage flash unifiée Lenovo ThinkSystem DM5100F

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

Informations sur le copyright

Informations sur les marques commerciales