NetAppとLenovoによるエッジAI推論

NetApp artificial intelligence solutions

NetApp
February 12, 2026

目次

NetAppとLenovoによるエッジAI推論

TR-4886: エッジAI推論 - NetAppとLenovo ThinkSystem - ソリューション設計

Sathish Thyagarajan、 NetApp Miroslav Hodak、Lenovo

このドキュメントでは、新たなアプリケーション シナリオに対応するエッジ環境のNetAppストレージ コントローラと Lenovo ThinkSystem サーバーに GPU ベースの人工知能 (AI) 推論を展開するためのコンピューティングおよびストレージ アーキテクチャについて説明します。

まとめ

先進運転支援システム (ADAS)、インダストリー 4.0、スマート シティ、モノのインターネット (IoT) などのいくつかの新しいアプリケーション シナリオでは、ほぼゼロの遅延で継続的なデータ ストリームを処理する必要があります。このドキュメントでは、これらの要件を満たすエッジ環境のNetAppストレージ コントローラと Lenovo ThinkSystem サーバーに GPU ベースの人工知能 (AI) 推論を展開するためのコンピューティングおよびストレージ アーキテクチャについて説明します。このドキュメントでは、 NVIDIA T4 GPU を搭載したエッジ サーバー上のさまざまな推論タスクを評価する、業界標準の MLPerf 推論ベンチマークのパフォーマンス データも提供します。オフライン、シングル ストリーム、マルチストリームの推論シナリオのパフォーマンスを調査し、コスト効率の高い共有ネットワーク ストレージ システムを備えたアーキテクチャが非常に高性能であり、複数のエッジ サーバーのデータとモデルの管理の中心点となることを示します。

はじめに

企業はネットワーク エッジで膨大な量のデータを生成するようになっています。スマート センサーと IoT データから最大限の価値を引き出すために、組織はエッジ コンピューティングを可能にするリアルタイム イベント ストリーミング ソリューションを求めています。そのため、計算負荷の高いジョブは、データセンター外のエッジで実行されることが増えています。 AI 推論はこのトレンドを推進する要因の 1 つです。エッジ サーバーは、特にアクセラレータを使用する場合、これらのワークロードに十分な計算能力を提供しますが、特にマルチサーバー環境では、ストレージの制限が問題になることがよくあります。このドキュメントでは、エッジ環境に共有ストレージ システムを導入する方法と、パフォーマンスを低下させることなく AI 推論ワークロードにどのようなメリットをもたらすかについて説明します。

このドキュメントでは、エッジでの AI 推論のリファレンス アーキテクチャについて説明します。複数の Lenovo ThinkSystem エッジ サーバーとNetAppストレージ システムを組み合わせて、導入と管理が容易なソリューションを作成します。これは、複数のカメラと産業用センサーを備えた工場現場、小売取引における販売時点管理 (POS) システム、自律走行車の視覚異常を識別する完全自動運転 (FSD) システムなど、さまざまな状況での実際の展開のためのベースライン ガイドとなることを目的としています。

このドキュメントでは、Lenovo ThinkSystem SE350 Edge Server とエントリーレベルのNetApp AFFおよびEF シリーズ ストレージ システムで構成されるコンピューティングおよびストレージ構成のテストと検証について説明します。リファレンス アーキテクチャは、AI 導入のための効率的でコスト効率の高いソリューションを提供するとともに、 NetApp ONTAPおよびNetApp SANtricityデータ管理ソフトウェアによる包括的なデータ サービス、統合データ保護、シームレスなスケーラビリティ、クラウド接続データ ストレージも提供します。

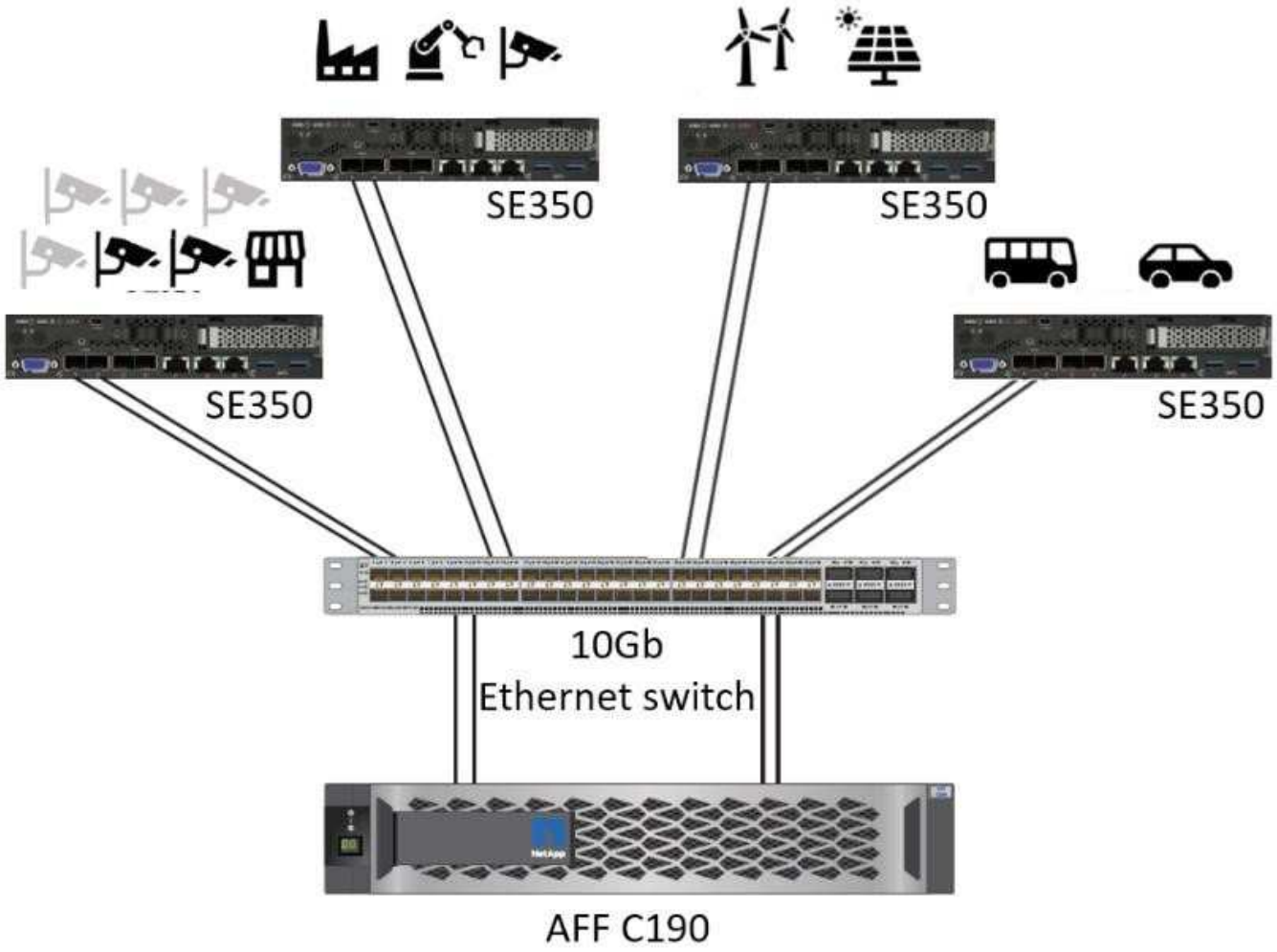
対象

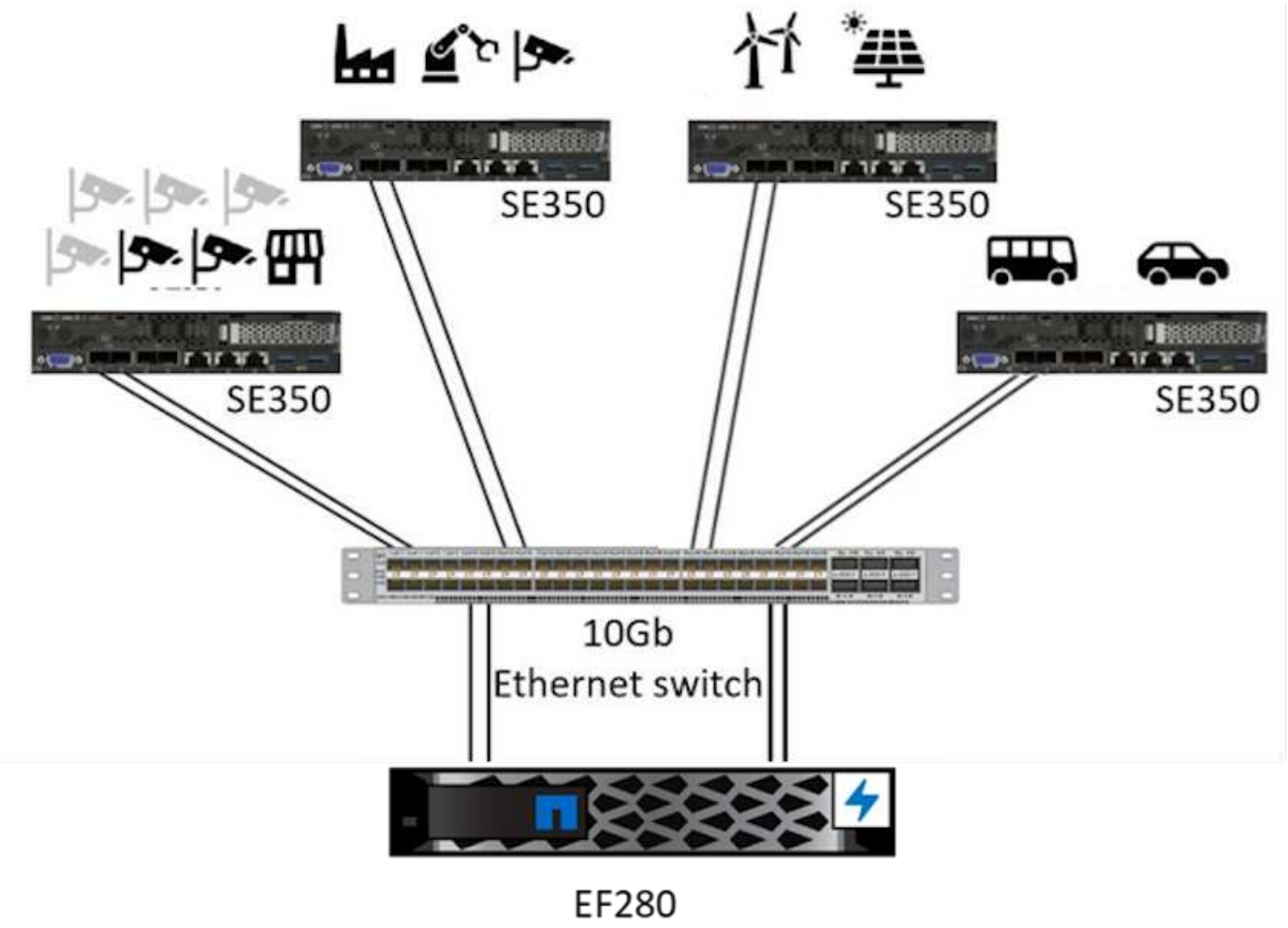
このドキュメントは次の読者を対象としています。

- エッジでの AI の製品化を希望するビジネス リーダーおよびエンタープライズ アーキテクト。
- データ サイエンティスト、データ エンジニア、AI/機械学習 (ML) 研究者、AI システムの開発者。
- AI/ML モデルとアプリケーションの開発のためのソリューションを設計するエンタープライズ アーキテクト。
- ディープラーニング (DL) および ML モデルを効率的に展開する方法を探しているデータ サイエンティストおよび AI エンジニア。
- エッジ推論モデルの展開と管理を担当するエッジ デバイス マネージャーとエッジ サーバー管理者。

ソリューションアーキテクチャ

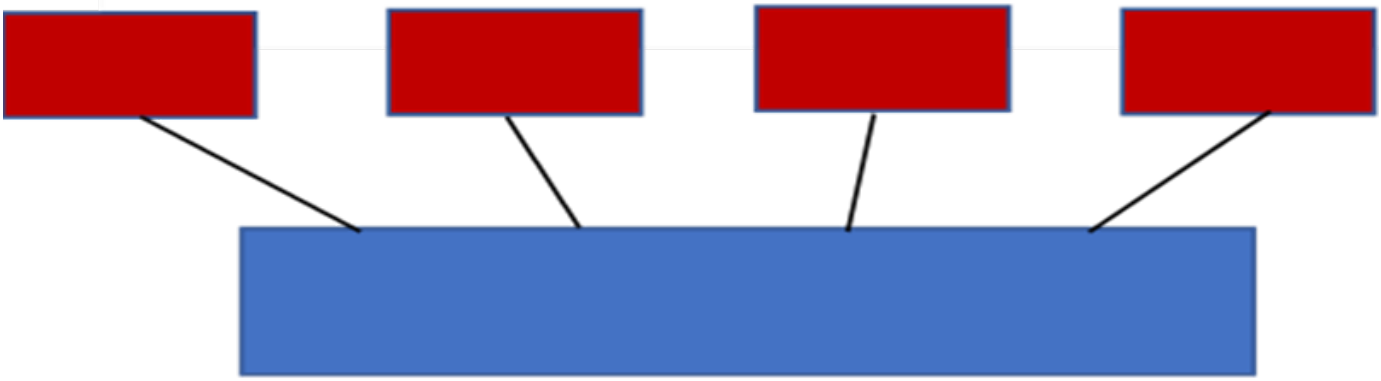
この Lenovo ThinkSystem サーバーとNetApp ONTAPまたはNetApp SANtricityストレージ ソリューションは、従来の CPU と並んで GPU の処理能力を使用して、大規模なデータセットでの AI 推論を処理するように設計されています。この検証では、次の 2 つの図に示すように、単一のNetApp AFFストレージ システムと相互接続された単一または複数の Lenovo SR350 エッジ サーバーを使用するアーキテクチャによる、高いパフォーマンスと最適なデータ管理が実証されています。





次の図の論理アーキテクチャの概要は、このアーキテクチャにおけるコンピューティング要素とストレージ要素の役割を示しています。具体的には、次の内容が表示されます。

- カメラやセンサーなどから受信したデータに対して推論を実行するエッジ コンピューティング デバイス。
- 複数の目的を果たす共有ストレージ要素:
  - 推論モデルと推論の実行に必要なその他のデータを一元的に保存する場所を提供します。コンピューティング サーバーはストレージに直接アクセスし、推論モデルをローカルにコピーする必要なく、ネットワーク経由で推論モデルを使用します。
  - 更新されたモデルはここにプッシュされます。
  - エッジ サーバーが受信した入力データをアーカイブし、後で分析できるようにします。たとえば、エッジ デバイスがカメラに接続されている場合、ストレージ要素はカメラでキャプチャされたビデオを保存します。



赤	青
レノボのコンピューティングシステム	NetApp AFFストレージシステム
カメラやセンサーなどからの入力に基づいて推論を実行するエッジ デバイス。	後で分析するために、エッジ デバイスからの推論モデルとデータを保持する共有ストレージ。

このNetAppと Lenovo のソリューションは、主に次の利点を提供します。

- エッジでの GPU アクセラレーション コンピューティング。
- 共有ストレージからバックアップおよび管理される複数のエッジ サーバーの展開。
- データ損失なしで低い復旧ポイント目標 (RPO) と復旧時間目標 (RTO) を満たす強力なデータ保護。
- NetAppスナップショット コピーとクローンを使用してデータ管理を最適化し、開発ワークフローを効率化します。

このアーキテクチャの使い方

このドキュメントでは、提案されたアーキテクチャの設計とパフォーマンスを検証します。ただし、コンテナー、ワークロード、モデル管理、クラウドまたはオンプレミスのデータセンターとのデータ同期など、特定のソフトウェア レベルの部分は、展開シナリオに固有のものであるため、テストしていません。ここでは複数の選択肢が存在します。

コンテナ管理レベルでは、Kubernetes コンテナ管理は適切な選択肢であり、完全なアップストリーム バージョン (Canonical) またはエンタープライズ展開に適した修正バージョン (Red Hat) のいずれかで適切にサポートされています。その"NetApp AI コントロール プレーン"NetApp Tridentと新たに追加された "NetApp DataOps ツールキット"データ サイエンティストやデータ エンジニアがNetAppストレージと統合するための組み込みのトレーサビリティ、データ管理機能、インターフェイス、およびツールを提供します。Kubernetes 用の ML ツールキットである Kubeflow は、TensorFlow Serving やNVIDIA Triton Inference Server などの複数のプラットフォームでのモデルのバージョン管理と KFServing のサポートに加えて、追加の AI 機能を提供します。もう 1 つのオプションは、GPU 対応 AI 推論コンテナーのカタログへのアクセスとともにワークロード管理を提供するNVIDIA EGX プラットフォームです。ただし、これらのオプションを本番環境に導入するには多大な労力と専門知識が必要になる可能性があり、サードパーティの独立系ソフトウェア ベンダー (ISV) またはコンサルタントの支援が必要になる場合もあります。

ソリューション領域

AI 推論とエッジ コンピューティングの主な利点は、デバイスが遅延なく高品質のデータを計算、処理、分析できることです。エッジ コンピューティングの使用例は非常に多く、このドキュメントですべてを説明することはできませんが、ここでは代表的なものをいくつか紹介します。

自動車：自動運転車

エッジ コンピューティングの典型的な例は、自律走行車 (AV) の先進運転支援システム (ADAS) です。無人運転車の AI は、安全な運転を実現するために、カメラやセンサーからの大量のデータを迅速に処理する必要があります。物体と人間の間の解釈に時間がかかりすぎると生死に関わる可能性があるため、そのデータをできるだけ車両の近くで処理できることが重要です。この場合、1 つ以上のエッジ コンピューティング サーバーがカメラ、RADAR、LiDAR、その他のセンサーからの入力を処理し、共有ストレージが推論モデルを保持し、センサーからの入力データを保存します。

ヘルスケア：患者モニタリング

AI とエッジ コンピューティングの最大の影響の 1 つは、在宅ケアと集中治療室 (ICU) の両方で慢性疾患の患者の継続的なモニタリングを強化できることです。インスリンレベル、呼吸、神経活動、心拍リズム、胃腸機能を監視するエッジデバイスから得られるデータは、人の命を救うために行動できる時間が限られているため、即座に行動を起こさなければならない即時のデータの分析が必要です。

小売：レジなし決済

エッジ コンピューティングは AI と ML を強化して、小売業者がチェックアウト時間を短縮し、客足を増やすのに役立ちます。レジなしシステムは、次のようなさまざまなコンポーネントをサポートします。

- 認証とアクセス。実店舗の買い物客を検証済みのアカウントに接続し、小売スペースへのアクセスを許可します。
- 在庫監視。センサー、RFID タグ、コンピューター ビジョン システムを使用して、買い物客によるアイテムの選択または選択解除を確認します。

  ここでは、各エッジ サーバーが各チェックアウト カウンターを処理し、共有ストレージ システムが中央の同期ポイントとして機能します。

金融サービス：キオスクにおける人間の安全と詐欺防止

銀行組織は AI とエッジ コンピューティングを使用して革新を起こし、パーソナライズされた銀行エクスペリエンスを生み出しています。リアルタイムのデータ分析と AI 推論を使用するインタラクティブ キオスクにより、ATM は顧客の現金引き出しをサポートするだけでなく、カメラで撮影した画像を通じてキオスクを積極的に監視し、人間の安全に対するリスクや不正行為を特定できるようになりました。このシナリオでは、エッジ コンピューティング サーバーと共有ストレージ システムがインタラクティブ キオスクやカメラに接続され、銀行が AI 推論モデルを使用してデータを収集および処理できるようになります。

製造業：インダストリー**4.0**

スマートファクトリーや 3D プリンティングなどの新たなトレンドとともに、第 4 次産業革命 (インダストリー 4.0) が始まりました。データ主導の未来に備えるために、大規模なマシンツーマシン (M2M) 通信と IoT を統合し、人間の介入を必要とせずに自動化を強化します。製造業はすでに高度に自動化されており、AI 機能の追加は長期的なトレンドの自然な流れです。 AI により、コンピューター ビジョンやその他の AI 機能を利用して自動化できる操作を自動化できます。品質管理や、人間の視覚や意思決定に依存するタスクを自動化して、工場現場の組立ラインで材料の分析を高速化し、製造工場が必要な ISO 規格と品質管理基準を満たすのに役立ちます。ここで、各コンピューティング エッジ サーバーは、製造プロセスを監視するセンサーの配列に接続され、更新された推論モデルが必要に応じて共有ストレージにプッシュされます。

通信：錆検出、鉄塔検査、ネットワーク最適化

通信業界では、コンピューター ビジョンと AI 技術を使用して画像を処理することで、錆を自動的に検出し、腐食が発生しているためさらに検査が必要な携帯電話基地局を特定します。近年、ドローン画像と AI モデル

を使用して塔の特定の領域を特定し、錆、表面のひび割れ、腐食を分析するケースが増加しています。通信インフラや携帯電話基地局を効率的に点検し、定期的に劣化状況を評価して、必要に応じて速やかに修理することを可能にする AI 技術の需要は高まり続けています。

さらに、通信分野で新たに出現しているユースケースとして、AI および ML アルゴリズムを使用してデータ トラフィック パターンを予測し、5G 対応デバイスを検出し、MIMO (複数入力複数出力) エネルギー管理を自動化および拡張するというものがあります。 MIMO ハードウェアは、ネットワーク容量を増やすために無線塔で使用されますが、これには追加のエネルギー コストがかかります。セル サイトに展開された「MIMO スリープ モード」の ML モデルは、無線の効率的な使用を予測し、モバイル ネットワーク オペレーター (MNO) のエネルギー消費コストの削減に役立ちます。 AI 推論およびエッジ コンピューティング ソリューションは、MNO がデータセンターとの間で送受信されるデータの量を削減し、TCO を下げ、ネットワーク運用を最適化し、エンドユーザーの全体的なパフォーマンスを向上させるのに役立ちます。

まとめ

AI を活用した自動化とエッジ コンピューティングは、ビジネス組織がデジタル変革を実現し、運用の効率と安全性を最大限に高めるのに役立つ主要なアプローチです。エッジ コンピューティングでは、データをデータセンターとの間で移動する必要がないため、データの処理速度が大幅に向上します。したがって、データセンターやクラウドとの間でデータを送受信する際に発生するコストが削減されます。企業がエッジに展開された AI 推論モデルを使用してほぼリアルタイムで意思決定を行う必要がある場合、レイテンシの低減と速度の向上が有益となります。

NetAppストレージ システムは、ローカル SSD ストレージと同等以上のパフォーマンスを提供し、データ サイエンティスト、データ エンジニア、AI/ML 開発者、ビジネスまたは IT の意思決定者に次のメリットをもたらします。

- AI システム、分析、その他の重要なビジネス システム間でデータを簡単に共有できます。このデータ共有により、インフラストラクチャのオーバーヘッドが削減され、パフォーマンスが向上し、企業全体のデータ管理が合理化されます。
- コストを最小限に抑え、リソースの使用率を向上させるために、コンピューティングとストレージを独立して拡張できます。
- 統合されたスナップショット コピーとクローンを使用して合理化された開発および展開ワークフローにより、瞬時にスペース効率の高いユーザー ワークスペース、統合バージョン管理、および自動展開が実現します。
- 災害復旧と事業継続性のためのエンタープライズ グレードのデータ保護。このドキュメントで紹介されているNetAppと Lenovo のソリューションは、エッジでのエンタープライズ グレードの AI 推論の展開に最適な、柔軟なスケールアウト アーキテクチャです。

謝辞

- JJ Lenovo、HPC および AI ソリューション担当シニア マネージャー、ファルカンガー氏
- Dave Arnette、 NetAppテクニカル マーケティング エンジニア
- Joey Parnell、 NetApp EシリーズAIソリューション担当テクニカルリード
- Cody Harryman、 NetAppの QA エンジニア

詳細情報の入手方法

このドキュメントに記載されている情報の詳細については、次のドキュメントや Web サイトを参照してください。

- NetApp AFF Aシリーズアレイ製品ページ

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

- NetApp ONTAPデータ管理ソフトウェア - ONTAP 9 情報ライブラリ

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

- TR-4727: NetApp EFシリーズの紹介

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

- NetApp EシリーズSANtricityソフトウェア データシート

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

- NetAppコンテナ向け永続ストレージ - 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/"
- NetApp BlueXPコピーと同期

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

- TensorFlowベンチマーク

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

- Lenovo ThinkSystem SE350 エッジサーバー

  "https://lenovopress.com/lp1168"

- Lenovo ThinkSystem DM5100F 統合フラッシュストレージアレイ

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