最新世代サーバー、Dell PowerEdge XR5610の機械学習パフォーマンス

著者: Frank Han, HPC and AI Innovation Lab
Rakshith Vasudev, HPC and AI Innovation Lab
Manya Rastogi, Technical Marketing Engineering

概要

デル・テクノロジーズはこのほど、高度なパフォーマンスを提供しエネルギー効率に優れた設計を実現した最新世代のDell PowerEdge サーバーを発表しました。

今回の「Direct from Development Tech Note」では、この最新世代PowerEdge サーバーに期待できる、新しい能力を解説します。業界標準のMLPerf Inference v2.1ベンチマーク・スイートを使用した、PowerEdge XR5610の機械学習（ML）のパフォーマンステストと、その結果をご説明します。XR5610は、ネットワーク・通信、エンタープライズ・エッジ、軍事、防衛といった、エッジでのAI/ML推論機能を必要とする、すべての主要なワークロードをターゲットとしています。

1Uでシングルソケット搭載のXR5610は、MCC SKUスタックの第4世代インテル® Xeon® Scalableプロセッサーを搭載し、エッジに最適化された奥行きの浅い堅牢な1Uサーバーです。最大8本のDDR5メモリ スロットと、2つのPCIe Gen5 x16カードスロットなど、最新世代のテクノロジーを搭載したこの製品は、前世代のPowerEdge XR12と比較して、画像分類ワークロードを46%高速化（レイテンシーを46%削減）できます。

PowerEdge XR5610：エッジ環境に向けた専用設計

エッジ コンピューティングは本質的に、計算能力をデータの発生源に近づけます。モノのインターネット（IoT）のエンドポイントやその他のデバイスが、より多くのタイムセンシティブなデータを生成するにつれて、エッジ コンピューティングがますます重要になっています。機械学習（ML）や人工知能（AI）アプリケーションは、特にエッジ コンピューティングの導入に適しています。エッジ コンピューティングの環境条件は通常、集中型データセンターとはかなり異なります。エッジ コンピューティングの設置サイトの環境には、空調設備が一切ないかあってもせいぜい最小限の空調設備と、通信クローゼット程度で構成されていることがあります。

Dell PowerEdge XR5610 は、堅牢で奥行きの短い (400 mm クラス) エッジ用 1U サーバーで、スペースや環境面で制約のある場所に導入できるよう設計されています。このサーバーは、-5C～55C（23F～131F）の高温環境での動作に適しており、通信テレコム領域のvRANワークロード、軍事および防衛の展開、ビデオ監視、IoTデバイスの集約、PoS分析を含むリテールAIにおいて、優れた性能を発揮するよう設計されています。

図1.　Dell PowerEdge XR5610 – 1U

エッジでのインテリジェンスという新興テクノロジー

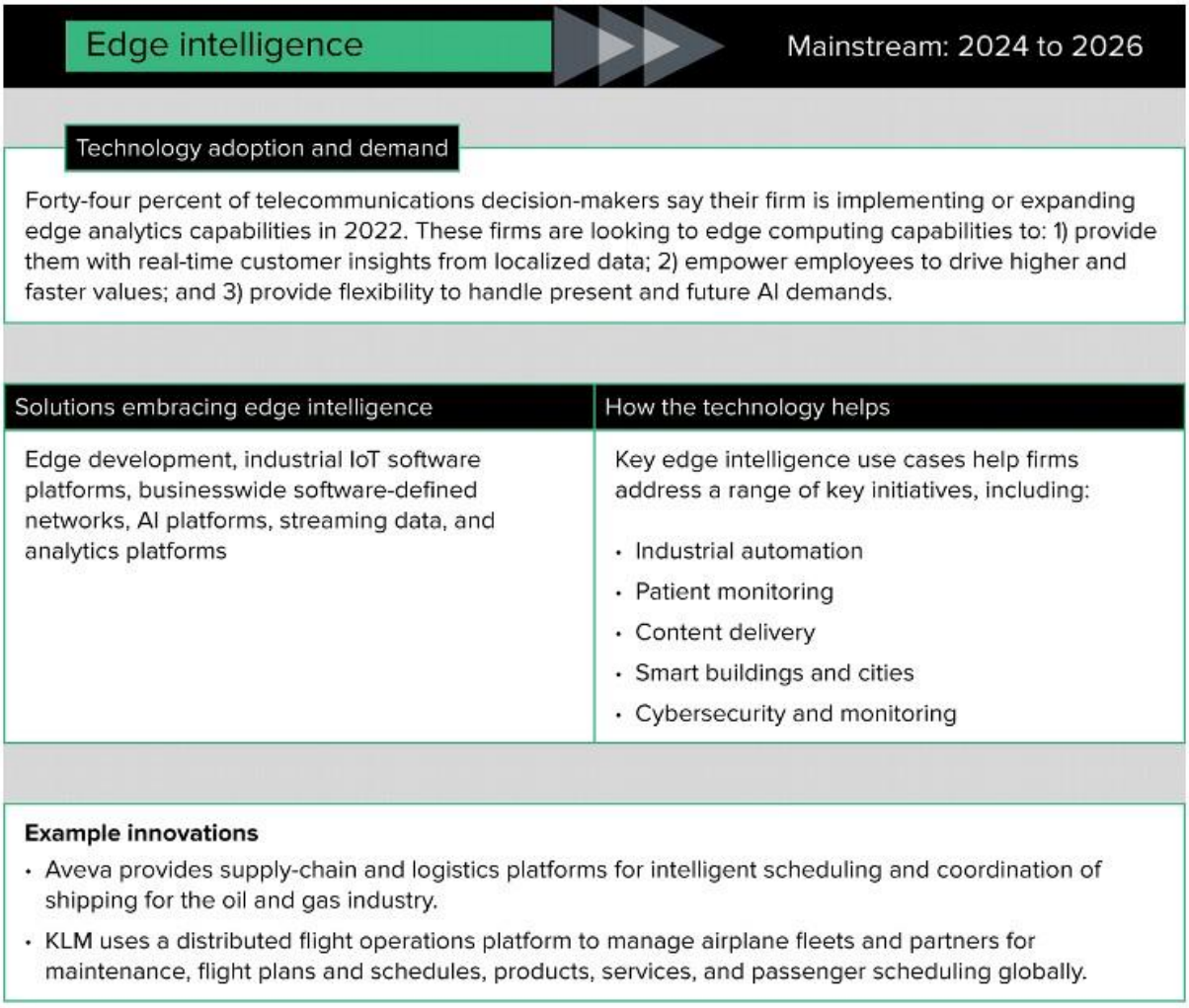
Forresterの最新レポート によれば、「エッジ インテリジェンスは2022年の新興テクノロジー トップ10にランクインしており、アプリケーション、デバイス、通信エコシステムのリアルタイム ネットワークでデータを取得し、推論を組み込み、インサイトに結びつけるのに役立つ」とされています。

FORRESTER®

Prepare For Emerging Edge Intelligence Momentum
Use Edge Intelligence To Enable Real-Time Insight And Contextual Awareness

FIGURE 1

Edge Intelligence Will Be Mainstream In The Next Two To Four Years



Base: 4,049 telecommunications decision-makers
Source: Forrester's Networks And Telecom Survey, 2022

Source: Forrester's Networks And Telecom Survey, 2022

図2. Forresterのレポートより抜粋（許諾を得て掲載）

MLPerfにおける推論ワークロード：概要

MLPerf Inference は、4つの異なるワークロードの種類と 3 つの処理シナリオを測定する多面的なベンチマーク フレームワークです。4つのワークロードは、画像分類、オブジェクト検出、医療画像、音声テキスト変換、自然言語処理 (BERT) です。次の表で概説されている処理シナリオは、シングル ストリーム、マルチストリーム、およびオフラインです。

表 1. MLPerf Inferenceのベンチマーク シナリオ

シナリオ	パフォーマンス メトリクス	ユースケース
シングル ストリーム	90パーセンタイルのレイテンシ	検索結果。クエリが実行されるまで待機し、検索結果を返します。 例: Google音声検索
マルチストリーム	99パーセンタイルのレイテンシ	マルチカメラの監視と迅速な意思決定。複数のリアルタイムストリームを処理し、不審な行動を特定するCCTVバックエンドシステムのような役割を果たします。 例: すべての複数カメラ入力を統合し、リアルタイムで運転判断を行う自動運転車両
オフライン	スループット測定	バッチ処理 (オフライン処理とも呼ばれます)。 例: 写真を識別し、人物をタグ付けし、特定の人物や場所、またはイベントをオフラインで記録したアルバムを生成する、Google フォト サービス

推論用の MLPerf スイートには、次のベンチマークが含まれています。

表2. 推論ベンチマーク用のMLPerfスイート

エリア	タスク	モデル	データセット	QSL サイズ	クオリティ
ビジョン	画像分類	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)
ビジョン	オブジェクト検出	Retinanet	OpenImages (800x800)	64	99% of FP32 (0.3755 mAP)
ビジョン	医用画像セグメンテーション	3D UNET	KiTS 2019	42	99% of FP32 and 99.9% of FP32 (0.86330 mean DICE score)
スピーチ	自動音声認識	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 – WER, where WER=7.452253714852645%)
言語	言語処理	BERT	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 (f1_score=90.874%)

MLPerfにおける推論ワークロード

次の表は、MLPerf 推論テスト スイートに使用したPowerEdge XR5610 の主な仕様概要です。

表3. MLPerf Inference テストスイートに利用したDell PowerEdge XR5610の主な仕様

コンポーネント	スペック
CPU	第4世代インテル Xeon スケーラブル・プロセッサー MCC SKU
オペレーティングシステム	CentOS 8.2.2004
メモリ	256 GB
GPU	NVIDIA A2
GPU搭載数	1
ネットワーキング	1x ConnectX-5 IB EDR 100 Gbps
ソフトウェアスタック	• TensorRT 8.4.2 • CUDA 11.6 • cuDNN 8.4.1 • GPU driver 510.73.08 • DALI 0.31.0
ストレージ	NVMe SSD 1.8 TB

表 4 に、ベンチマークテストで使用した NVIDIA GPU の仕様を示します。

表4. テストに利用したNVIDIA GPU

GPUモデル	GPUメモリ	最大消費電力	フォームファクター	2-wayブリッジ	推奨ワークロード
PCIe adapter form factor					
A2	16 GB GDDR6	60 W	シングルワイド、フルハイト ハーフレングスまたはフルハイト ハーフレングス	対象外	AI推論、エッジ、VDI

このエッジサーバーでは、画像処理はGPUにオフロードされます。サーバーの価格と性能のレベルがさまざまな要件に合わせて異なるように、GPUもまた、然りです。XR5610では、前世代のXR11と同様、最大2枚のシングルワイドGPUがサポートされます。

今回のXR5610 は、オフラインシナリオにおける MLPerf ワークロードの全範囲について、NVIDIA A2 GPUを使用してテストされました。次の図は、テスト結果を示しています。

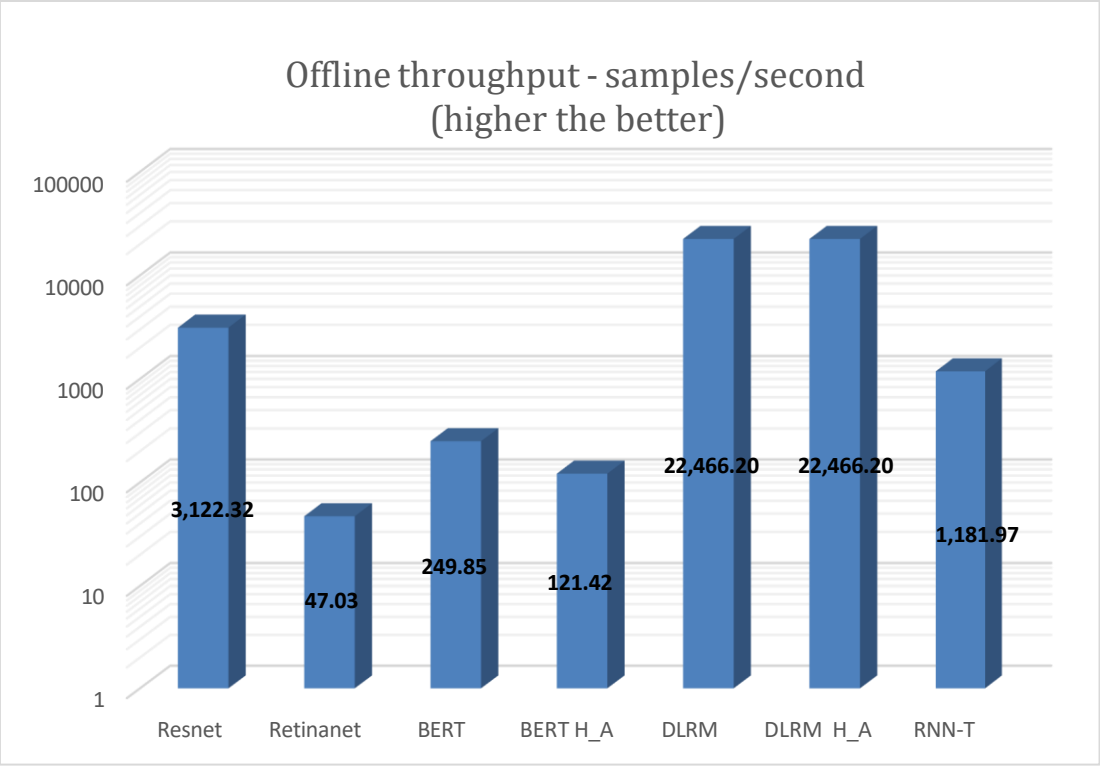


図3. MLPerf オフラインシナリオにおける NVIDIA A2 GPU テスト結果

またXR5610 は、NVIDIA A2 GPU を使用して、シングルストリームシナリオの MLPerf ワークロードの全範囲についてもテストされました。次の図は、そのテスト結果を示しています。

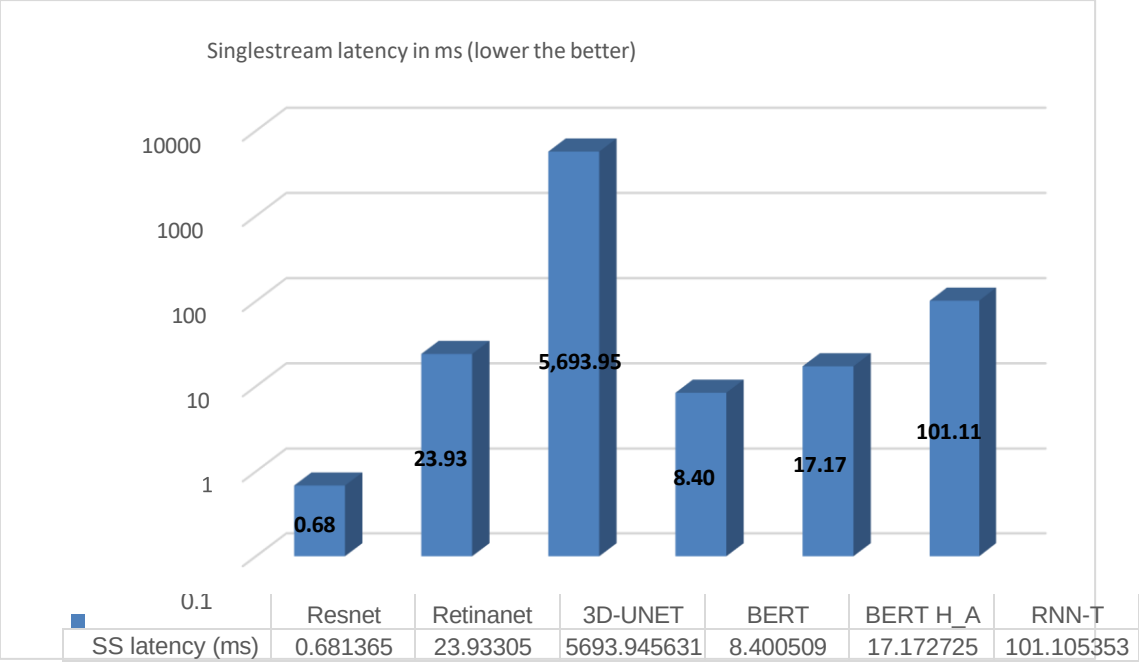


図4. MLPerf シングル ストリーム シナリオの NVIDIA A2 GPU テスト結果

一部のタスク/ワークロードでは、PCIe Gen5などの新テクノロジーに起因すると思われる前世代機からの改善もXR5610に見られました。

.

イメージ分類

次の図に示すように、PowerEdge XR5610 は前世代のサーバーと比較して、イメージ分類の待機時間が 46% 向上しました。

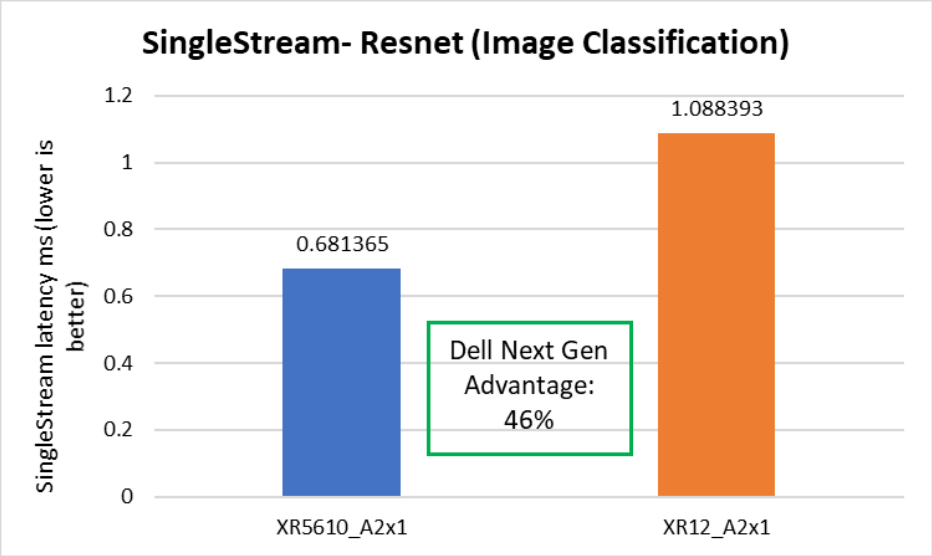


図5. 画像分類の待ち時間： XR5610および旧世代のPowerEdgeサーバー

自動音声認識

Dell XR5610 は、次の図に示すように、前世代の PowerEdge サーバーと比較して、音声テキスト変換スループットが 15% 向上しました。

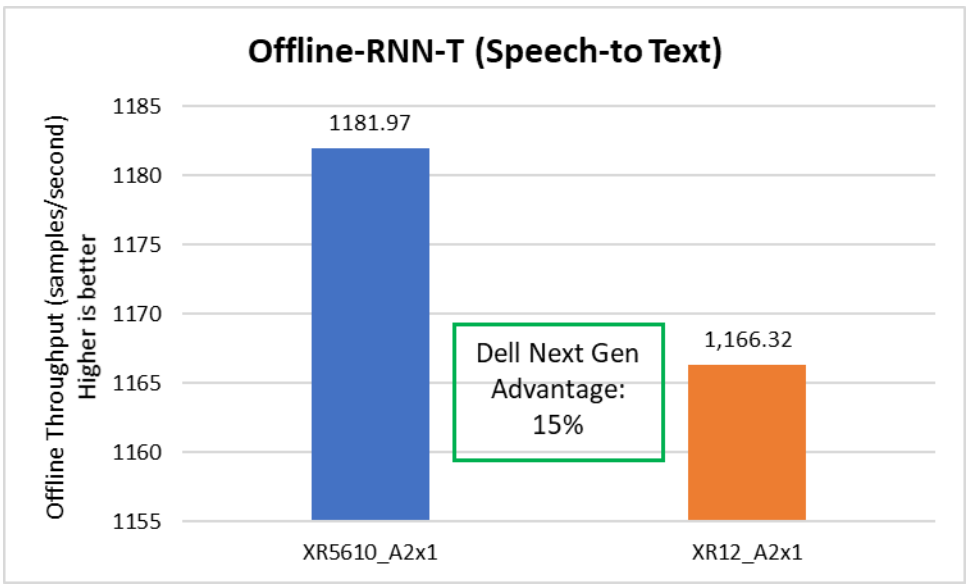


図6. 音声テキスト変換の待機時間: XR5610と前世代のPowerEdge サーバー

結論

PowerEdge XR ポートフォリオは、異なるユースケースに基づき、さまざまなエッジおよびテレコム領域の導入オプションに向け、合理的なアプローチを提供し続けます。PowerEdge XRは、業界標準の堅牢性認証（NEBS）を取得し、-5℃～+55℃の温度範囲で拡張性と柔軟性を備えたコンパクトなソリューションを提供します。

参考情報

MLPerf Inferenceベンチマーク

注:

- 2023 年 1 月に Dell Cloud and Emerging Technology ラボで実施したテストに基づく。結果は 24 年度第 2 四半期に MLPerf に提出予定。
- MLPerf2 MLCommons Association による検証結果ではありません。MLPerf の名称およびロゴは、米国およびその他の国における MLCommons Association の商標です。無断転載を禁じます。無断使用厳禁。詳細は www.mlcommons.org をご参照ください。