最新世代のDell PowerEdge XR7620サーバーによる機械学習 (ML) 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, ISG

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 サーバーに期待される、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 XR7620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。XR7620は、製造、小売、防衛、通信など、エッジでのAI/ML推論機能を必要とする主要なワークロードをターゲットとしています。

最大2基の300WのGPUアクセラレーターカードを搭載し、最も要求の厳しいエッジワークロードに対応するXR7620は、エンタープライズエッジでのML/AIシナリオ向けに、前世代のDell PowerEdge XR 12（300W GPUアクセラレーターを1基のみ搭載可）と比較して、画像分類ワークロードを45%高速化します。低レイテンシーと高い処理能力の組み合わせにより、より迅速で効率的なデータ分析が可能となり、企業はより多くのビジネスチャンスに向けたリアルタイムの意思決定を行うことができる。

エッジ コンピューティング

エッジ コンピューティングとは、一言で言えばデータの発生源にコンピューティング パワーを近づけることです。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、エッジ コンピューティングの重要性は、ますます高まっています。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。そのため、堅牢で、目的にあわせ専用設計され、コンパクトかつ高速なエッジサーバーは、そうした環境でのサーバー導入に理想的です。Dell PowerEdge XR7620サーバは、これらの条件をすべて満たしています。最も要求の厳しいワークロードに対応する高性能、大容量のサーバーでありつつ、-5℃から55℃の高温度な環境と埃っぽい環境で動作が認定されており、450mm（耳からラック背面まで）クラスの奥行きの短いフォームファクターで提供されます。

MLPerf Inference ワークロードの概要

MLPerf は、さまざまなワークロードタイプと処理シナリオをベンチマークする多面的なベンチマークスイートです。5 つのワークロードと 3 つの処理シナリオがあります。ワークロードは以下の通りです。

- 画像分類
- 物体検出

- 医用画像のセグメンテーション
- 音声テキスト変換
- 言語処理

シナリオは、シングルストリーム (SS)、マルチストリーム (MS)、およびオフラインです。

タスクは一目瞭然ですが、使用したデータセット、使用したMLモデルおよびその説明とともに以下の表に示します。シングルストリームテストでは、90パーセンタイルの結果が報告されました。マルチストリームテストでは、99パーセンタイルの結果が報告されました。

表 1. MLPerf Inferenceのベンチマーク シナリオ

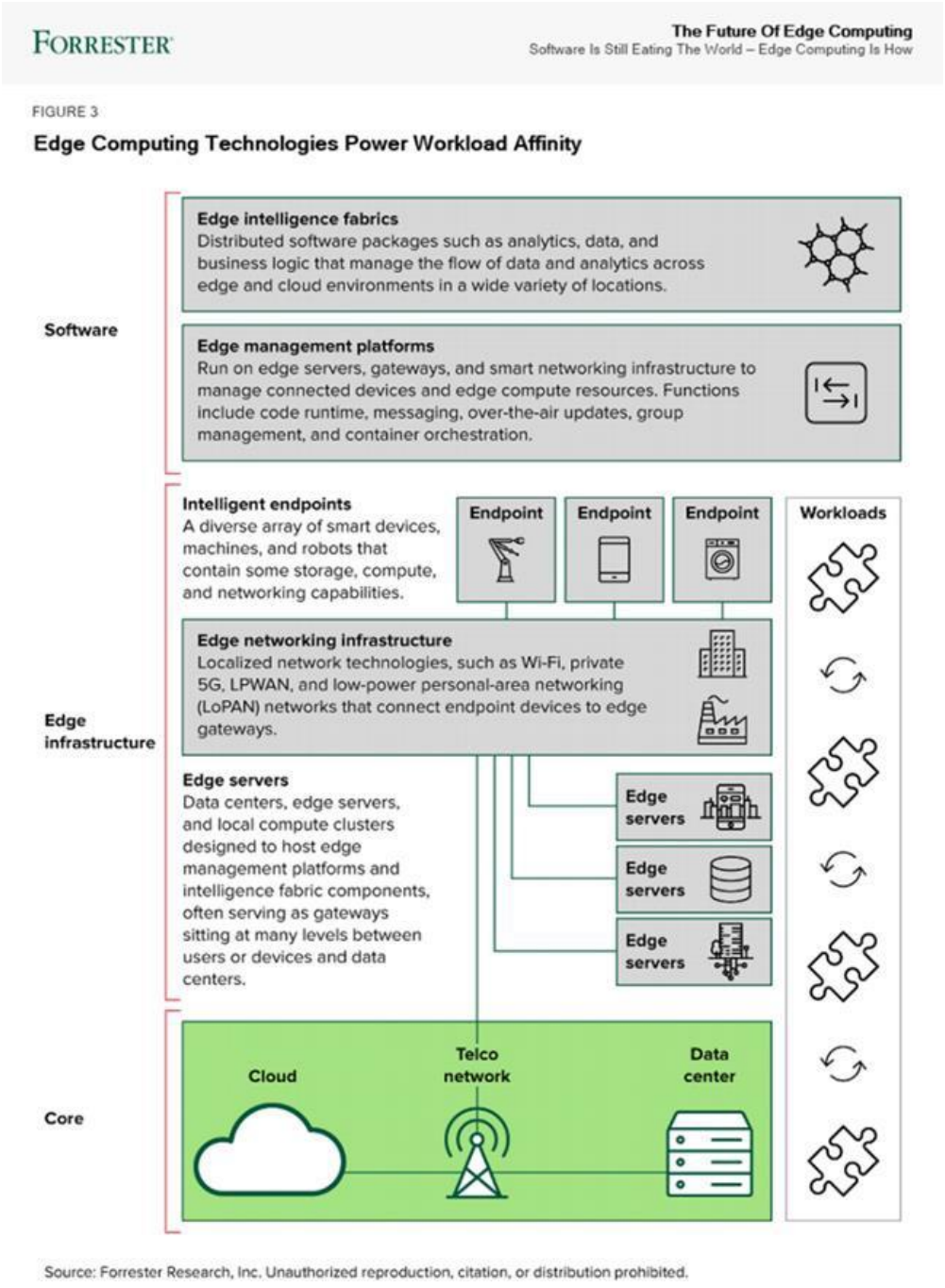
シナリオ	パフォーマンス メトリクス	ユースケースの例
シングル ストリーム	90パーセンタイルのレイテンシ	Google 音声検索: クエリが入力されるまで待機し、検索結果を返します。
オフライン	スループット測定	バッチ処理、別名オフライン処理。 Google フォトは、写真を識別し、タグ付けし、特定の人物と場所/イベントを含むアルバムをオフラインで生成します。
マルチストリーム	99パーセンタイルのレイテンシ	例1：マルチカメラによる監視と迅速な意思決定。MultiStreamは、不審な行動を特定するために複数のリアルタイムストリームを処理するCCTVバックエンドシステムのようなものです。 例2：自動運転カメラは複数のカメラからの入力を統合し、リアルタイムで運転判断を行います。

表 2. 推論ベンチマーク用のMLPerfスイート

Area	Task	Model	Dataset	QSL Size	Quality
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)
Vision	Object detection	Retinanet	OpenImages (800x800)	64	99% of FP32 (0.3755 mAP)
Vision	Medical image segmentation	3D UNET	KiTS 2019	42	99% of FP32 and 99.9% of FP32 (0.86330 mean DICE score)
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)
Language	Language processing	BERT	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 (f1_score=90.874%)

エッジ コンピューティングの未来に関する業界レポート

Forrester社のレポート（「5つのテクノロジー要素が、4つのビジネスシナリオでワークロードの親和性を実現する」）によると、現在のほとんどのシステムは、1つの場所でソフトウェアを実行するように設計されています。そのため、工場に多くのセンサーが設置されたり、イベントに多くの人が集まったり、カメラに多くのビデオフィードが送られるなど状況が変化すると、パフォーマンスに限界が生じます。革新的なAI/ML、アナリティクス、IoT、コンテナ ソリューションは、新しいアプリケーション、デプロイメント オプション、ソフトウェア設計戦略を可能にします。将来、システムは、その時々のニーズに応じて、さまざまな場所でソフトウェアを実行する場所を選択するようになるでしょう。



ML/AI推論パフォーマンス

表 3. Dell PowerEdge XR7620の主な仕様

MLPerf system suite type	Edge
CPU	CentOS 8.2.2004
オペレーティングシステム	第4世代インテル Xeon スケーラブル・プロセッサー MCC SKU
メモリ	512GB
GPU	NVIDIA A2
GPU搭載数	1
ネットワーキング	1x ConnectX-5 IB EDR 100Gb/Sec
ソフトウェアスタック	TensorRT 8.4.2 CUDA 11.6 cuDNN 8.4.1 Driver 510.73.08 DALI 0.31.0

図 1.　Dell PowerEdge XR7620: 2U 2S

表 4. テストに利用したNVIDIA GPU

ブランド	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						
NVIDIA	A2	16GB GDDR6	60W	SW, HHHL or FHHL	n/a	AI Inferencing, Edge, VDI
NVIDIA	A30	24GB HBM2	165W	DW, FHFL	Y	AI Inferencing, AI Training
NVIDIA	A100	80GB HBM2e	300W	DW, FHFL	Y, Y	AI Training, HPC, AI Inferencing

エッジ サーバーは、画像処理を GPU にオフロードします。サーバーの価格レベル・性能レベルがさまざまな要件に合わせて異なる点は、GPUも同様です。XR7620は、最大2枚のDW（ダブルワイド） 300W GPUまたは4枚のSW（シングルワイド） 150W GPUをサポートし、Dell PowerEdgeサーバー ポートフォリオが提供する、絶えず進化するスケーラビリティと柔軟性の一部となっています。これに対して前世代のPowerEdge XR11では、最大2枚のSW GPUがサポートされていました。

エッジサーバー vs データセンター用サーバーの比較[1]

オフラインシナリオでNVIDIA A100 GPUを使用したテストでは、Dell XR7620は、前世代のDell PowerEdgeラック型サーバーと比較して、1%未満の差しかないパフォーマンスを提供しました。AI推論シナリオにおいて、奥行き430mmのXR7620エッジサーバーが、ラック型サーバーと同等のパフォーマンスを提供することができるのです。図2を参照ください。

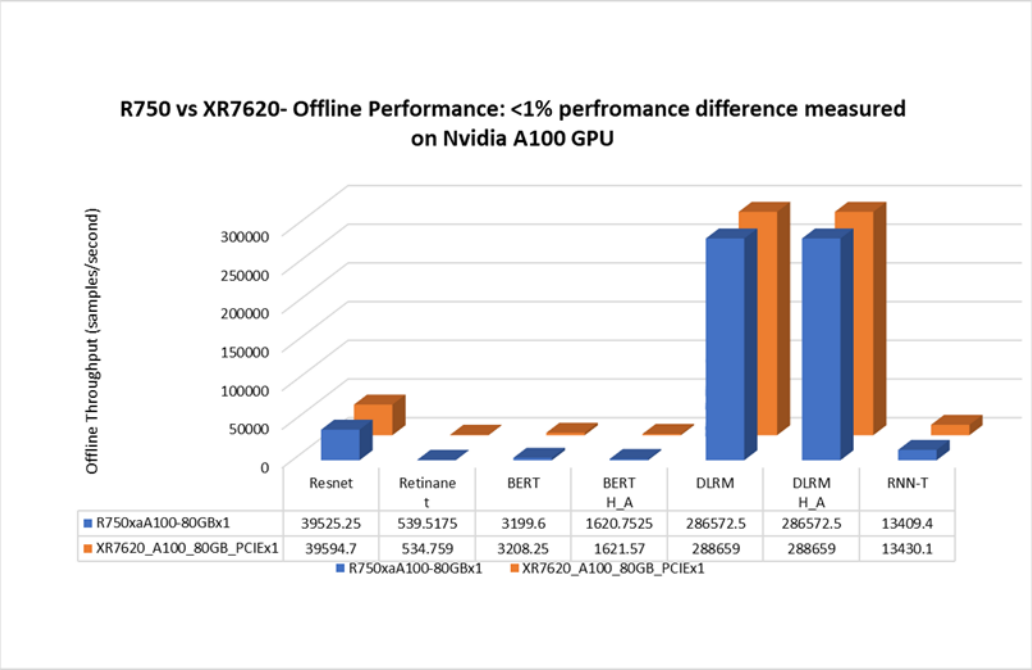


図 2. ラック型サーバー vs エッジサーバーのMLPerf のオフライン パフォーマンス比較

NVIDIA A2 GPUを搭載したXR7620のパフォーマンス

またXR7620 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搭載時のテストも行われました。結果は図3をご確認ください。

[1] 2023年1月、Dell Cloud and Emerging Technologyラボでのテストに基づく。

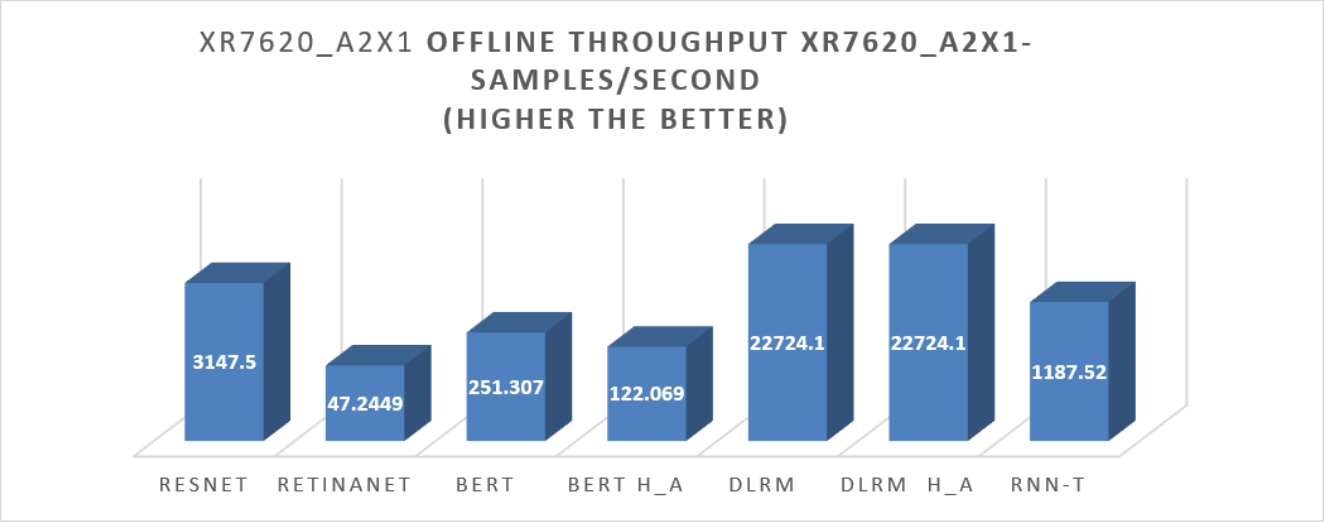


図 3. XR7620 オフラインパフォーマンスのテスト結果

NVIDIA A2 GPU搭載のXR7620 は、シングル ストリーム シナリオの MLPerf ワークロードでも全範囲でテストされました。結果は図 4をご覧ください。

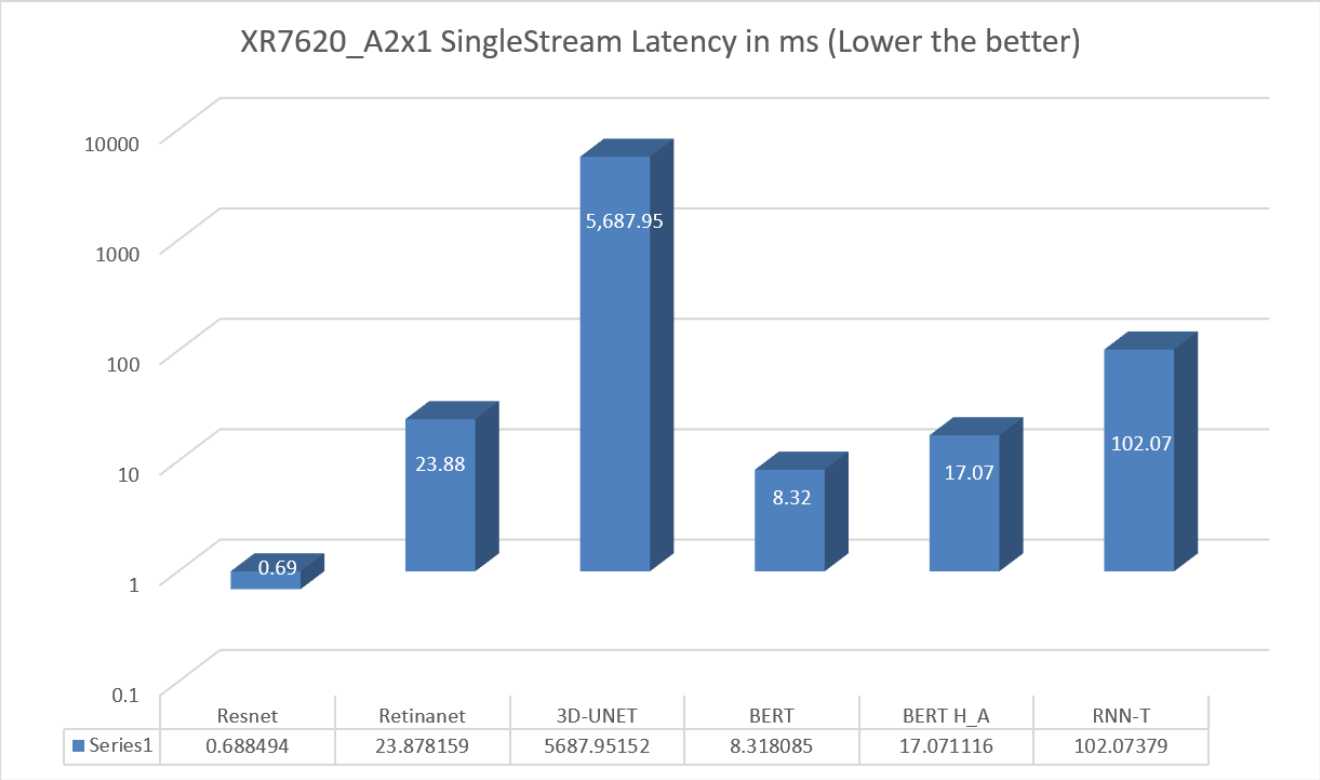


図 4. XR7620シングルストリームパフォーマンスの結果

XR7620 は、オフライン シナリオの MLPerf ワークロードの全範囲に対して、NVIDIA A30 GPU でもテストされました。図 5をご覧ください。

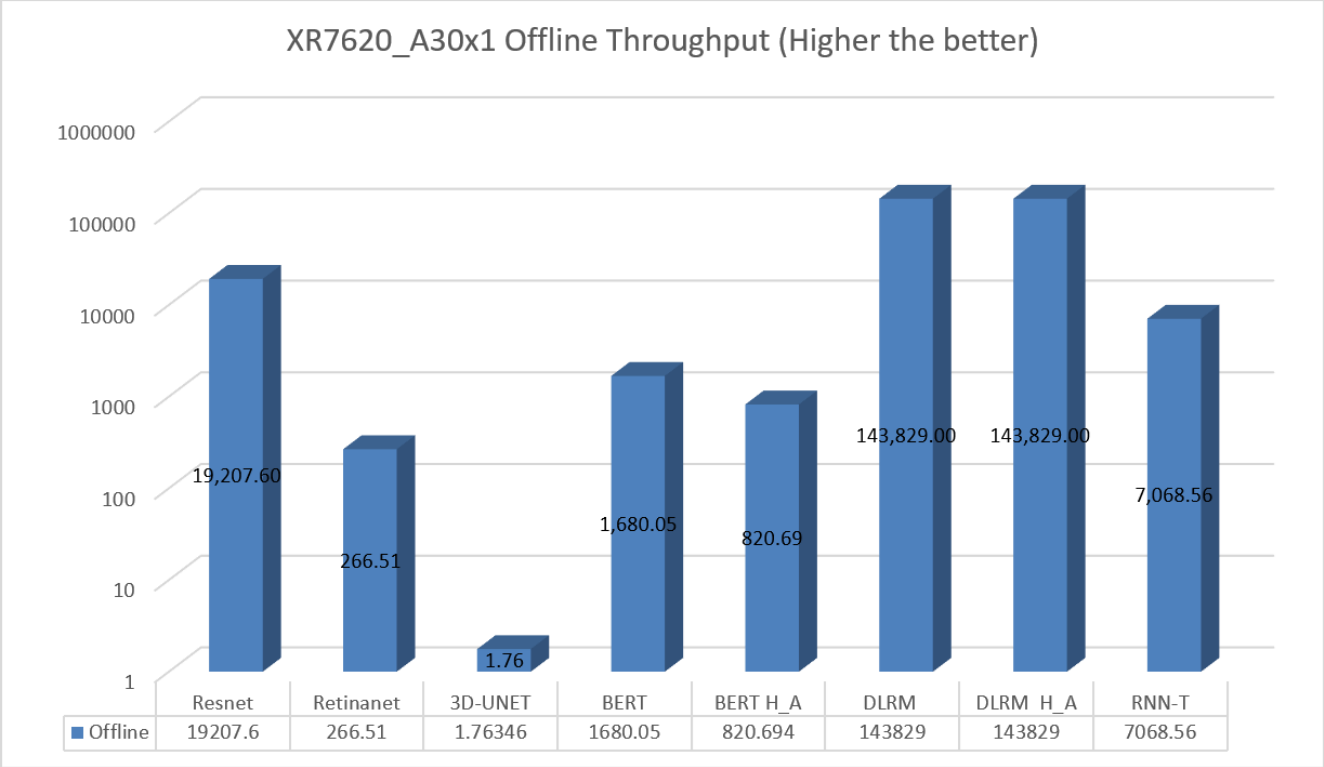


図 5. A30 GPU搭載時のXR7620のオフラインパフォーマンス検証結果

また、XR7620 は NVIDIA A30 GPU を使用し、シングルシナリオにおける MLPerf ワークロードの全範囲についてテストも行いました。図 6 を参照ください。

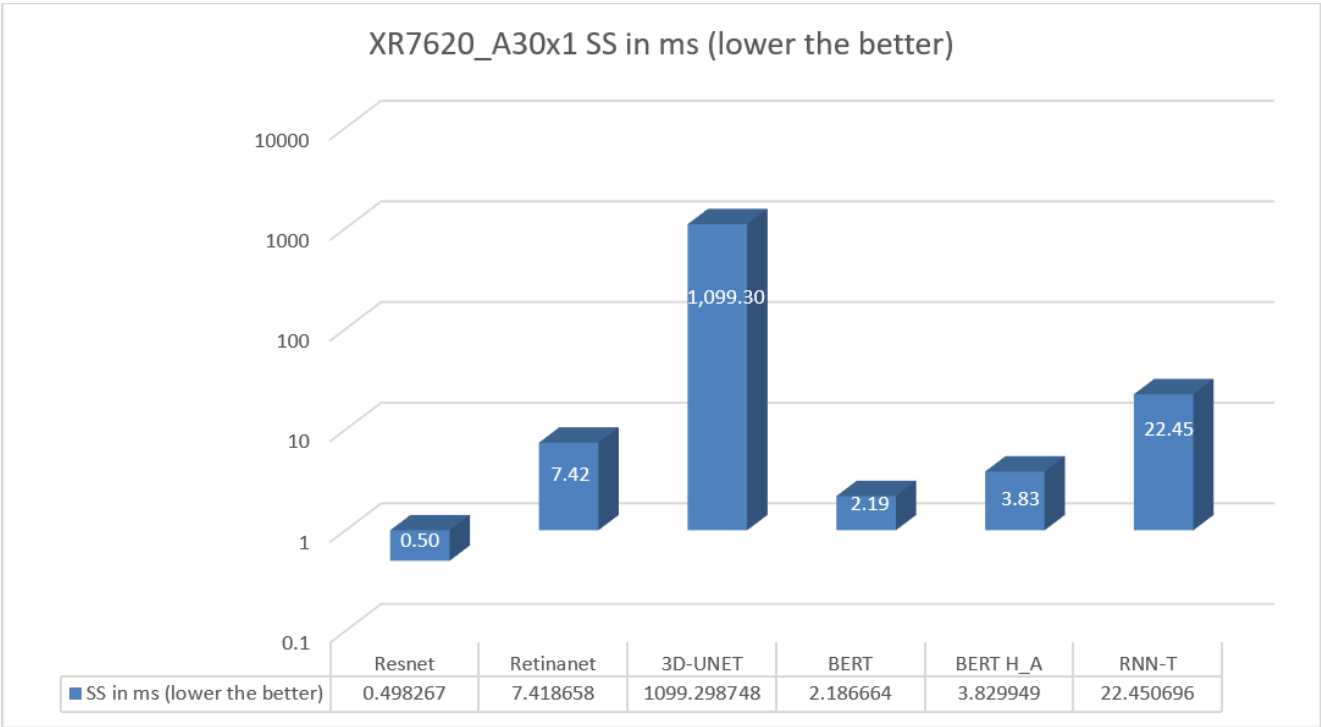


図 6. A30 GPU活用時のXR7625のシングルストリーム パフォーマンス

一部のシナリオでは、最新世代のPowerEdgeサーバーは、PCIe Gen 5など最新テクノロジーの統合による旧世代からの改善も見られました。

音声テキスト変換

Dell XR7620は、前世代のDellサーバーと比較して、スループットが16%向上しました。図 7 を参照ください。

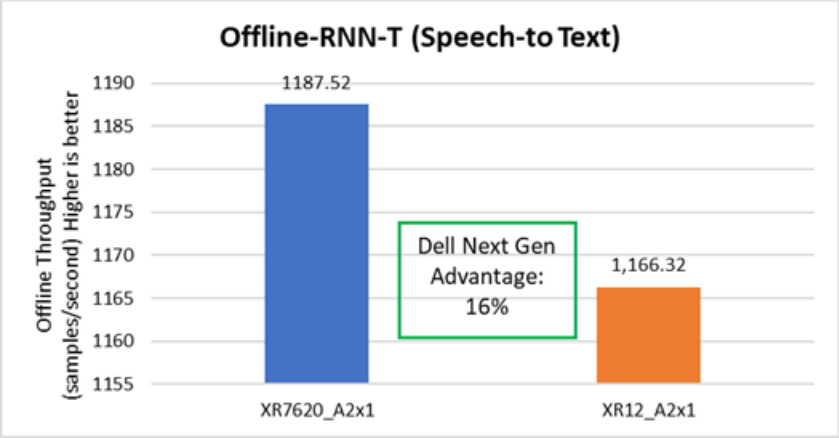


図 7. XR7620でのオフライン音声テキスト変換のパフォーマンスの改善

画像分類

Dell XR7620は、前世代のDellサーバーと比較して、レイテンシが45%改善されました。図 8 を参照ください。

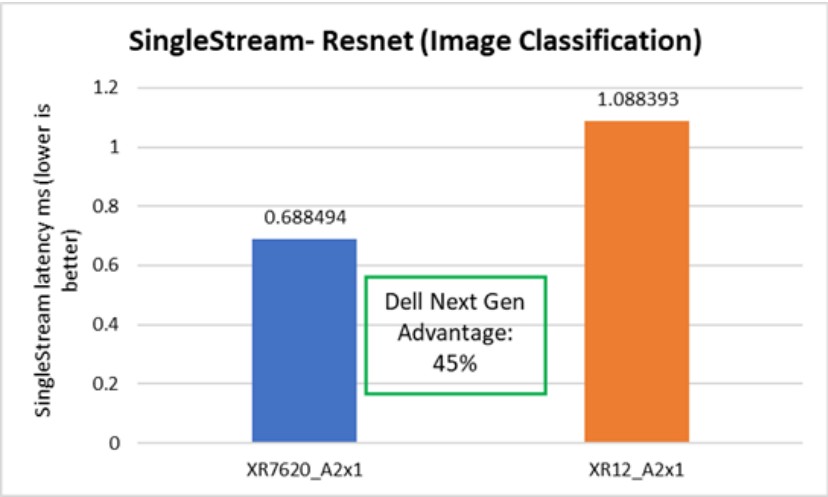


図 8. SS Image Classification performance improvement on XR7620

結論

Dell 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て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。今回のMLPerf の結果は、AI 推論をおこなうサーバーのエッジでの推論に関する、現実的なシナリオを提供するものとなります。本書の結果に基づき、Dellサーバは引き続き完全なソリューションを提供します。

参考情報

- MLPerf Inference Benchmark

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。
- 未検証のMLPerf v2.1 推論。結果はMLCommons協会によって検証されていません。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。無断使用は固くお断りします。より詳細な情報は www.mlcommons.org でご確認いただけます。