2024 年 10 月 7 日

NTT コミュニケーションズ株式会社

NTT Com、IOWN APN を活用した分散データセンターでの生成 AI 学習実証実験に世界で初めて成功

ドコモグループの法人事業ブランド「ドコモビジネス」を展開する NTT コミュニケーションズ株式会社(以下 NTT Com)は、超高速かつ超低消費電力を実現する IOWN 構想※1 の主要技術であるオールフォトニクス・ネットワーク（以下 APN）で接続した複数のデータセンターに NVIDIA GPU 搭載サーバーを分散配置した環境で、NVIDIA AI Enterprise プラットフォームの一部である NVIDIA NeMo™※2 を用いた生成 AI モデル学習の実証実験(以下 本実証)に世界で初めて成功しました。

1.背景

生成 AI、データ利活用、画像処理などの分野で GPU クラスタの重要性が高まる中、サービス提供事業者や利用者にとって従来は単一のデータセンター内で GPU クラスタを構築・利用することが一般的でした。しかし、単一のデータセンターでは、生成 AI のモデルサイズ増大に伴う処理量の変動に応じてオンデマンドに GPU リソースを入手できないことや、1 拠点のデータセンターのキャパシティや電力供給に制限があること、利用者の拠点から移動できない機密度の高いデータの取り扱いが難しいことが課題でした。

本実証により、IOWN APN を用いた分散データセンターにおける、GPU クラスタでの処理の有効性を確認することで、GPU クラスタ利用者や提供事業者の課題解決に貢献します。

2.本実証の概要

NVIDIA GPU 搭載サーバーを約 40km 離れた三鷹と秋葉原のデータセンターに分散配置し、データセンター間を 100Gbps 回線の IOWN APN で接続しました。NVIDIA NeMo™ 使用して、両拠点の GPU サーバーを連携させ、生成 AI モデルの分散学習を実施しました。なお、本実証はデル・テクノロジーズ株式会社による GPU サーバーやストレージなどの機器提供および協力のもとで実施しました。

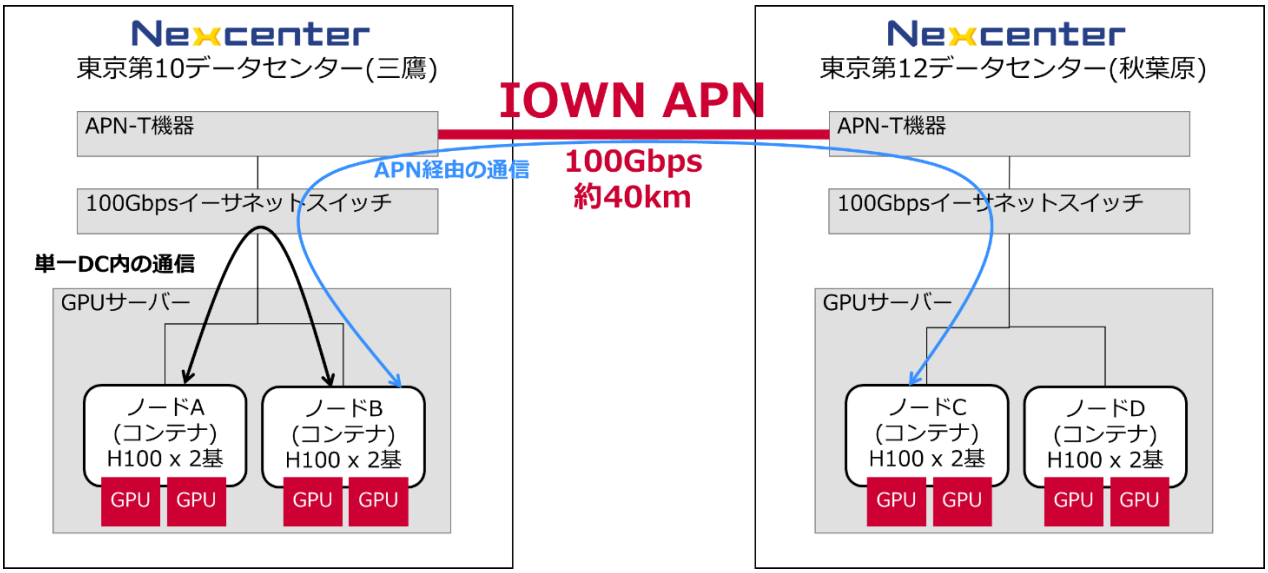
本実証で用いた技術の主な特長は以下の通りです。

(1) IOWN APN

IOWN APN の高速大容量・低遅延接続により、GPU サーバー間のデータ転送が迅速かつ効率的に行われ、小規模な AI モデルの事前学習や追加学習などの比較的軽量な処理に対して、単一のデータセンターと遜色ない性能を発揮できます。これによって、複数のデータセンター環境で柔軟に GPU クラスタを構築し、効率的なリソース利用を実現することが可能です。

(２) NVIDIA NeMo™

分散学習に対応した大規模言語モデルの学習、カスタマイズ、展開のためのエンド ツー エンド プラットフォームである NVIDIA NeMo™ を活用しました。今回の実証で扱った Llama 2 7B※3 以外のモデルなど、将来的にさまざまな生成 AI の処理に対応可能です。



<実証のイメージ>

3.本実証の成果

本実証は世界で初めて、高速大容量・低遅延な接続を可能とする IOWN APN と NVIDIA NeMo™を組み合わせた環境で、生成 AI のモデル学習（Llama 2 7B の事前学習※4）を動作させることに成功しました。

単一のデータセンターで学習させる場合の所要時間と比較して、インターネット経由の分散データセンターでは 29 倍の時間がかかるが、IOWN APN 経由の分散データセンターでは 1.006 倍と、単一のデータセンターとほぼ同等の性能を発揮できることを確認しました。

4.今後の展開

本実証の成果をもとに、IOWN APN で接続された分散データセンターにおける GPU クラスタの可能性をさらに広げ、国内 70 拠点以上のデータセンター間などを接続可能な「APN 専用線プラン powered by IOWN」や、液冷方式サーバーに対応した超省エネ型データセンターサービス「Green Nexcenter®」、などを組み合せた GPU クラウドソリューションとしてお客さまへ提供をめざします。

5．docomo business Forum'24 出展情報

2024 年 10 月 10 日(木)～11 日(金)に開催する「docomo business Forum'24」にて、本実証を展示予定です。公式 Web サイトの展示情報よりご確認ください。

公式 Web サイト：https://www.ntt.com/business/go-event.html?ir=nr

■展示名：その瞬間を感じる IOWN の世界

■展示番号：IV-01

＊会場：ザ・プリンス パークタワー東京 B2 フロア

＊日時：2024 年 10 月 10 日(木)～11 日(金) 9:30～17:30

＊参加方法：公式 Web サイトより事前に来場お申し込みをお願いします

＊参加費用：無料

※1：IOWN (Innovative Optical and Wireless Network)構想とは、NTT が提唱する次世代情報通信基盤です。
https://group.ntt/jp/group/iown/ 「IOWN®」は、日本電信電話株式会社の商標又は登録商標です。

※2：NVIDIA NeMo™とは、生成 AI モデルを構築・カスタマイズ・デプロイするための開発プラットフォームです。
https://docs.nvidia.com/nemo-framework/index.html

※3：Llama 2 7B とは、Meta 社が公開している大規模言語モデル（LLM）の 1 つで、パラメータ数が 70 億のものです。

※4：事前学習（Pre-training）とは、大規模なデータセットを使用してモデルに基本的な知識を習得させるプロセスのことです。