

GPU over APN

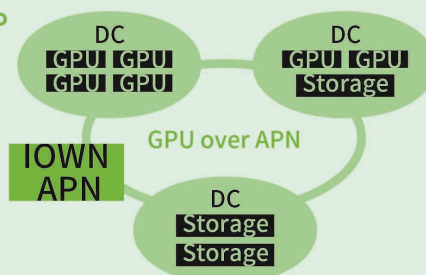
APNを活用して、AI需要に対応するGPUクラウドを柔軟かつセキュアに実現します

#業務効率化 #グリーントランスフォーメーション #地方創生

コンセプト/メリット Concept and Benefits

計算資源やデータの最適な分散配置を考慮した柔軟なGPUインフラ
Realizing a Flexible GPU Cloud with Optimized Distribution of Compute Resources and Data

- 1 処理量の変動に応じてオンデマンドにGPUリソースを確保
On-Demand GPU Provisioning: Dynamically allocate GPU resources to match workloads
- 2 利用者の拠点から移動できない機密度の高いデータの扱いが可能
Enhanced Data Security: Process sensitive data that must remain at the user's location
- 3 1つのデータセンターの床面積や電力供給能力に制限されない
Scalability Beyond Physical Constraints: Overcome limitations of floor space and power capacity in a single data center



実証実験 Experiments

- IOWN APNで接続した40km距離の2拠点にGPUクラスタを分散
Deployed single GPU cluster across two data centers 40km apart linked by IOWN APN
- APNとNVIDIA NeMo™を組み合わせる生成AI学習を実施
Performed Generative AI model training using IOWN APN and NVIDIA NeMo™
- 従来の単一拠点と比較して互角の性能を確認
Achieved performance equivalent to traditional single data center configuration

///技術課題

1つのDCにGPUクラウドを構築する場合、ワークロード増大時に柔軟に拡張することが困難です。また、企業データの所在を守る観点でも都心型DCを離れられず非効率です。

---要素技術

- ・ 低遅延・大容量の通信を実現するAPNと、セキュアな通信路を保証する耐量子セキュアトランスポートなどIOWN Pets技術
- ・ APNを用いたGPU分散基盤、それらを活用したアプリケーションの構築・運用技術
- ・ NVIDIA NeMo/GPU Direct 技術とAPNとを組みあわせるインテグレーション技術

---適用ビジネス

ビジネス領域：データセンタ事業者、GPU基盤利用者（AI／データ分析／メディア処理などでGPU基盤を利用するお客様）
ユースケース例：複数データセンタにまたがるGPU基盤を最適配置することで、伸縮性によるコスト最適化、DCキャパ・電力供給問題を解消、データレジデンシー問題を解消
提供時期：APN専用線：NTT Comより商用リリース済、GreenNexcenter®：NTT Comより2025年3月商用リリース予定
GPU over APN実証環境、共創パートナー募集中

///研究目標

低遅延・大容量のAPNとNVIDIA GPUの最新技術と組み合わせ、分散データセンタでも同一データセンタ相当の性能で生成AIの学習が行えることや、通信路もセキュアに保ちます。

---市中技術差異点

- ・ IOWN APNとNVIDIA NeMo FrameworkとLLM学習とを組み合わせた世界初の実証に成功
- ・ 耐量子セキュアトランスポートとも組み合わせるセキュリティ面で安心安全なGPUクラウドを提供