

GPU over APN

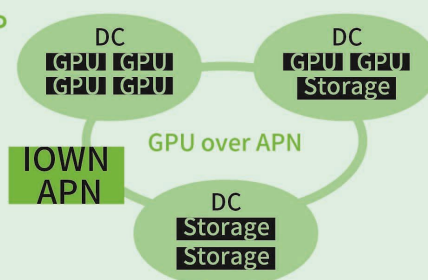
Realizing an elastic and secure GPU cloud using APN for
Increasing AI demands

#Green Transformation #Productivity Improvement #Regional Revitalization

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

///Technical Issue

Deploying a GPU cloud in a single data center limits scalability during workload spikes and complicates data residency compliance.

///Research Goal

Low-latency APN and NVIDIA GPUs enable distributed data centers to perform AI training with high performance and secure communications.

---Technology

GPU distributed infrastructure utilizing APN, and techniques for building and operating applications leveraging these technologies
Integration expertise combining NVIDIA NeMo/GPU Direct technologies with APN.

---Applicable Business

Business Area: Data center operators, GPU infrastructure users
Use Cases: Benefits of Multi-Data Center GPU Infrastructure will include, Cost optimization through scalability, Resolves data center capacity and power supply issues, Addresses data residency concerns.
Availability: GPU over APN proof-of-concept environment: Seeking co-creation partners

---Novelty

We successfully conducted the world's first proof of concept combining:
IOOWN APN, NVIDIA NeMo Framework, LLM training
Additionally, by integrating quantum-resistant secure transport, we can provide a GPU cloud that is both secure and reliable.