キャリア網は分散AI推論基盤になれるのか？

～5G × IOWN APN × 分散GPUを使ったAI Grid実証でみえたこと～

株式会社 NTTドコモ
宮本 克真

自己紹介

宮本　克真 (Katsuma Miyamoto)

NTTドコモ コアネットワークデザイン部

経歴	• **NTT ネットワークサービスシステム研究所**
	  • Docker/Kubernetesを用いたNW仮想化の研究
	• **NTT ネットワークイノベーションセンタ**
	  • MEC向けゲートウェイの開発・維持管理
	  • FPGA/IPUを用いた高速パケット処理の研究
	• **NTTドコモ コアネットワークデザイン部（現職）**

JANOG登壇歴	• JANOG53（モバイルコアネットワークへのFPGA導入） • JANOG55（DPUで5GC UPFを実装してみた）
担当業務	• AI × GitOpsによる5GC設計構築自動化（**AI for NW**） • モバイルネットワーク × 分散AI基盤（**NW for AI**）
趣味	• クラフトビアバー巡り（最近はHazy IPA）

AI時代における通信事業者の役割とは？

【現状】 クラウドAI集中型・Agentic AI時代

- 通信事業者は帯域・接続を提供するだけのコモディティ化
- AIで生まれる付加価値はハイパースケーラに集中
- AIエージェント時代の遅延の積上げ（エージェント間通信はミリ秒レベルを要求）

【問い】AI時代における通信事業者の付加価値は何か？

- AI時代のキャリアNWの新しい役割を考える必要がある
- パケットを運ぶ事業者から、トークンを運ぶ事業者へ

通信事業者ならではのAI基盤を自分たちで作っていく必要がある

AI Gridとは

- NVIDIAが2026年3月のGTC2026で発表した**分散型AIインフラ構想**
- 地理的に分散・相互接続されたAI計算基盤を**単一の知能基盤として動作**させる仕組み
- ポイントは、**通信事業者・CDN事業者が主役**として位置づけられていること

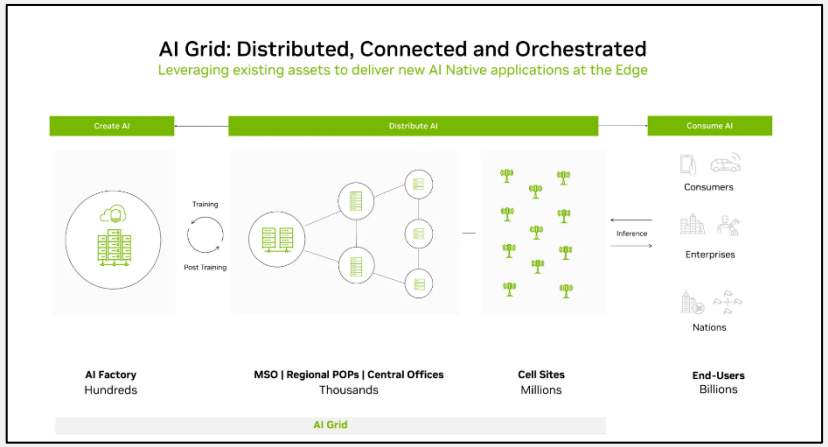
〈AI Gridを定義する4本の柱〉

① **分散（Distributed）**： Telco・CDN拠点のメッシュにGPUを配置

② **相互接続（Interconnected）**： 高速・低遅延NWでサイト間を接続

③ **統制（Orchestrated）**： SLA/遅延/コストを評価する知能制御プレーン

④ **統合（Unified）**： 中央DC・リージョナルエッジで同一動作



AI Grid（NVIDIA GTC San Jose 2026より）

AI Grid Explained: From Secure AI Factories to Distributed Intelligence S82010 | GTC San Jose 2026
Build With NVIDIA AI Grid Reference Design — AI Grid Documentation

NTTグループの技術でAI Grid実現に挑戦

- NTTグループの技術を組み合わせた**AI Gridに挑戦**
- 商用5G・dUPF・IOWN APNなど、**キャリア網ならではの構成要素**を検討
- 2026年6月の**Interop ShowNetでいち早く実証**してみた

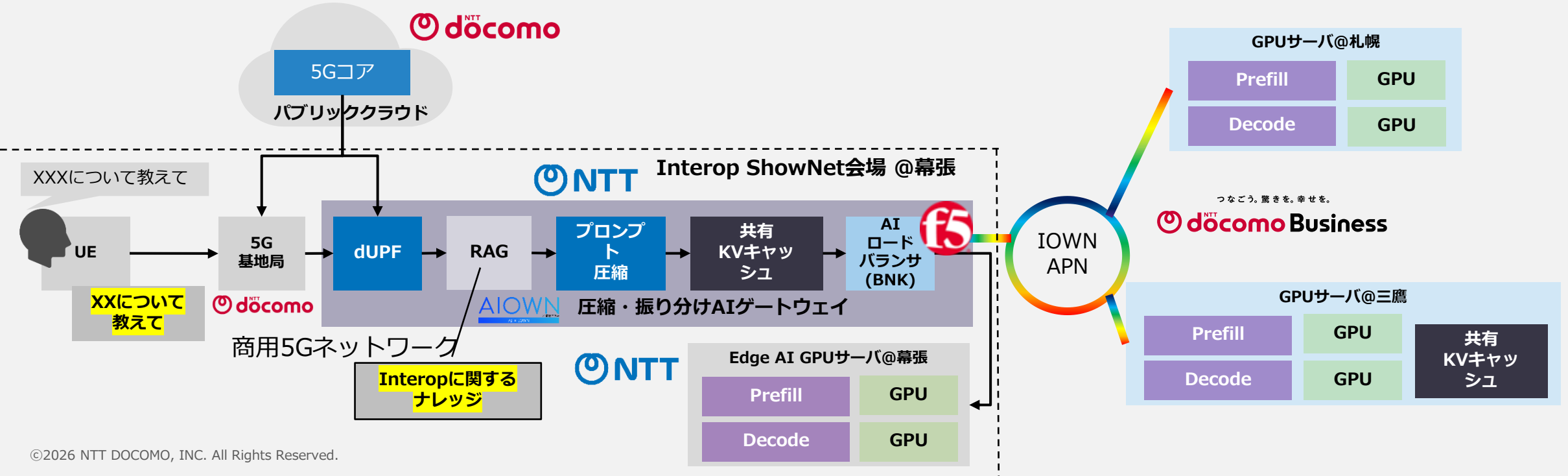
AI Gridの４本柱をNTTグループ技術での実現性を検討

Definitions	NTT Group Assets	期待される効果
①分散（Distributed）	全国NTT拠点に分散配置したGPU基盤	Edge + Regional + Farの全国展開
②相互接続（Interconnected）	IOWN APN (100Gbps RDMA)	拠点間の高速・低遅延接続
③統制（Orchestrated）	AI圧縮ゲートウェイ（プロンプト圧縮）	プロンプト圧縮による省メモリ化
	商用5G＋dUPF＋GPU動的振分	エッジ処理＋負荷に応じた動的振分
	F5 BNK との連携	SLA-aware Load Balancing
④統合（Unified）	NVIDIA Dynamo + KVキャッシュ共有 機能との連携	拠点間で推論コンテキストを統合

実証実験システム全体構成

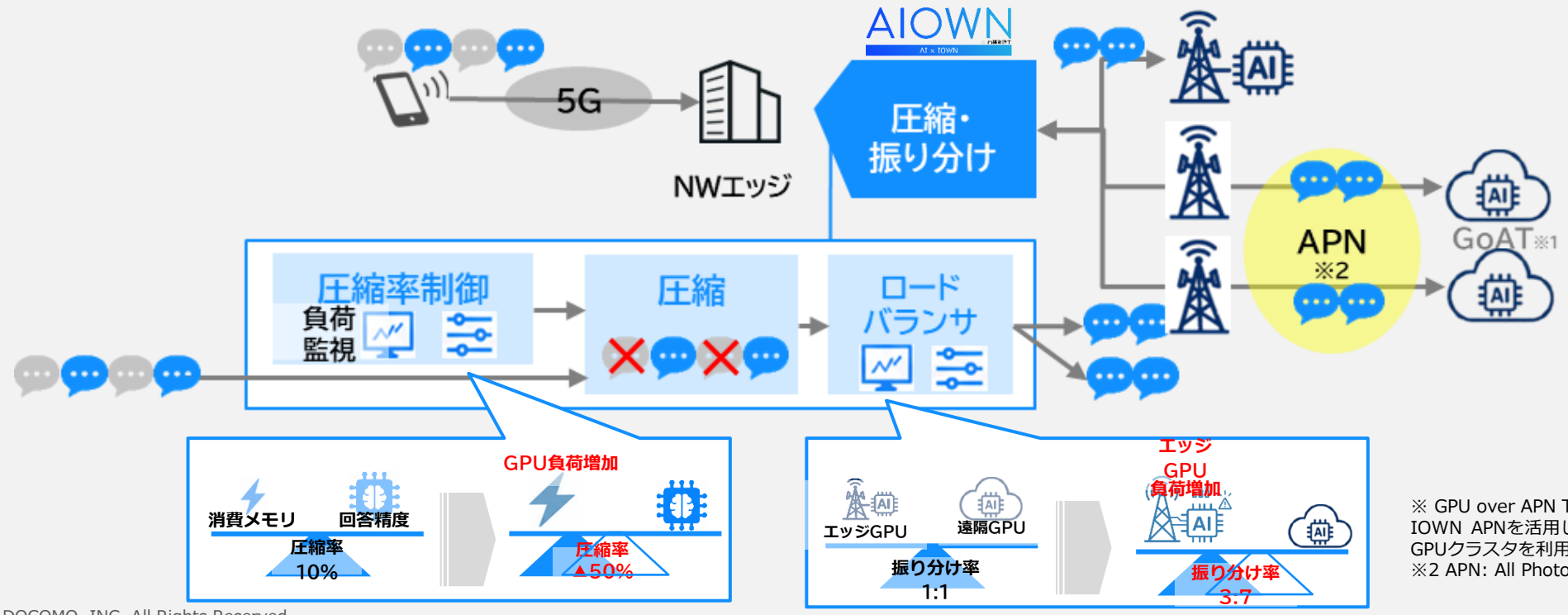
Interop ShowNetにおいて、商用5G・IOWN APN・分散GPUを用いたAI Grid実証実験を実施

- 圧縮振り分けゲートウェイをNWエッジ(幕張)に配置
- AI推論を、幕張のGPU、遠隔拠点のGPUに振り分ける
- 5G端末からAI推論リクエストに対する推論処理の効率化を検証
- ユースケースとしてInteropに関するAI音声チャットボット(AI Ask NOC)をデモ実証



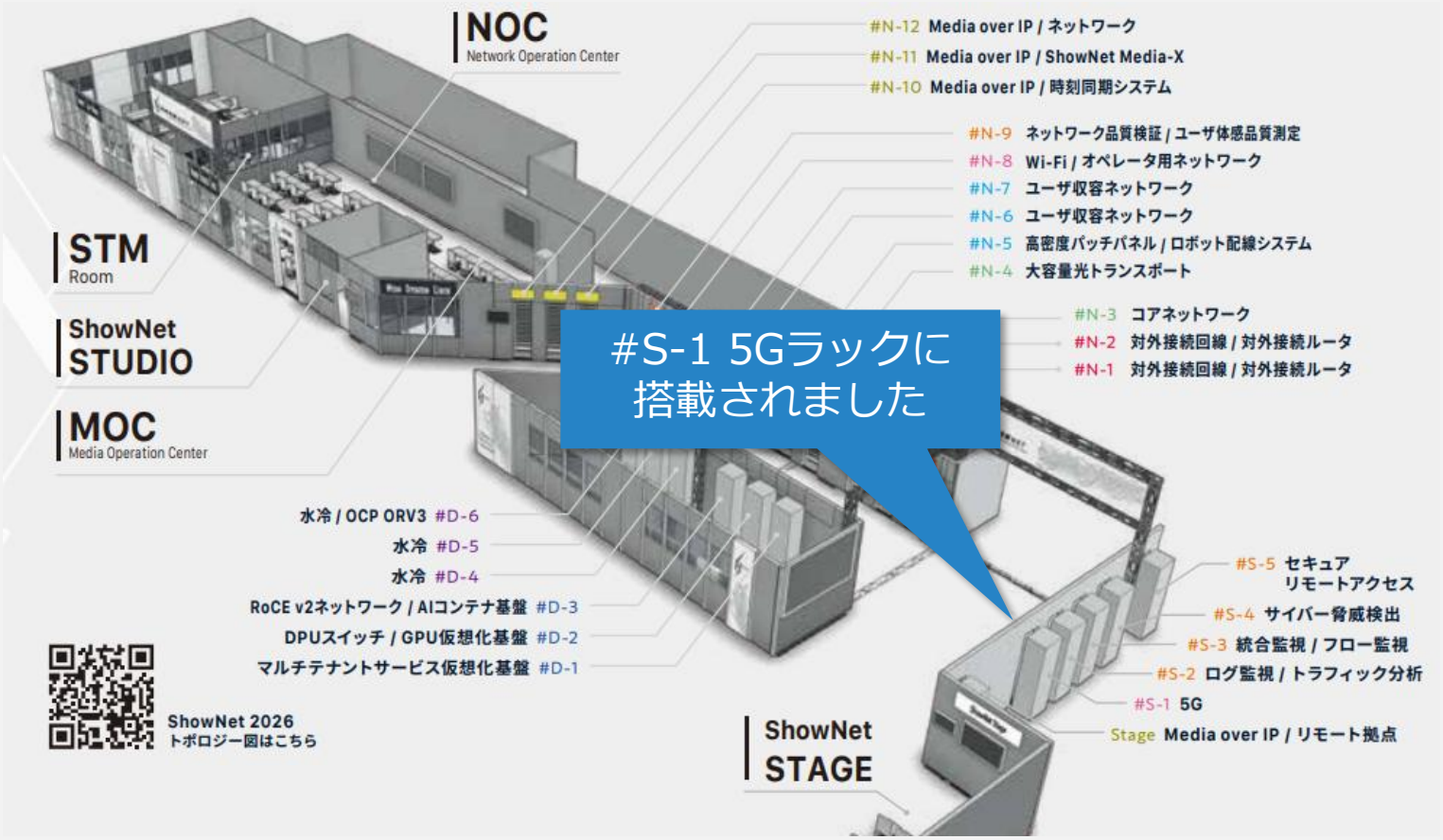
実証実験システム全体構成

- 圧縮はGPUの使用率に応じて動的に制御を行う
  - ex) GPU負荷が増えた(減った)場合には圧縮率を上げる(下げる)
- ロードバランサはGPU使用率に応じて振り分け割合を動的に制御
  - ex) エッジGPU負荷が増えた(減った)場合には遠隔GPUへの振り分けを増やす



※ GPU over APN Testbed
IOWN APNを活用し複数データセンターに分散した
GPUクラスタを利用できるAI向け分散GPUインフラ
※2 APN: All Photonics Network

ShowNetでの実機検証



#S-1 5Gラック

参考： GPU over APN Testbed（GoAT）

つなごう。驚きを。幸せを。
docomo Business

課題

AI需要の拡大により、GPUを拡充していきたいが、電力・冷却・床荷重の制約から単一拠点での増設には限界がある。

解決方法

IOWN APNで複数データセンターを繋ぎ、離れた拠点にあるGPUサーバをあたかも一体のGPU基盤として利用する。

提供価値

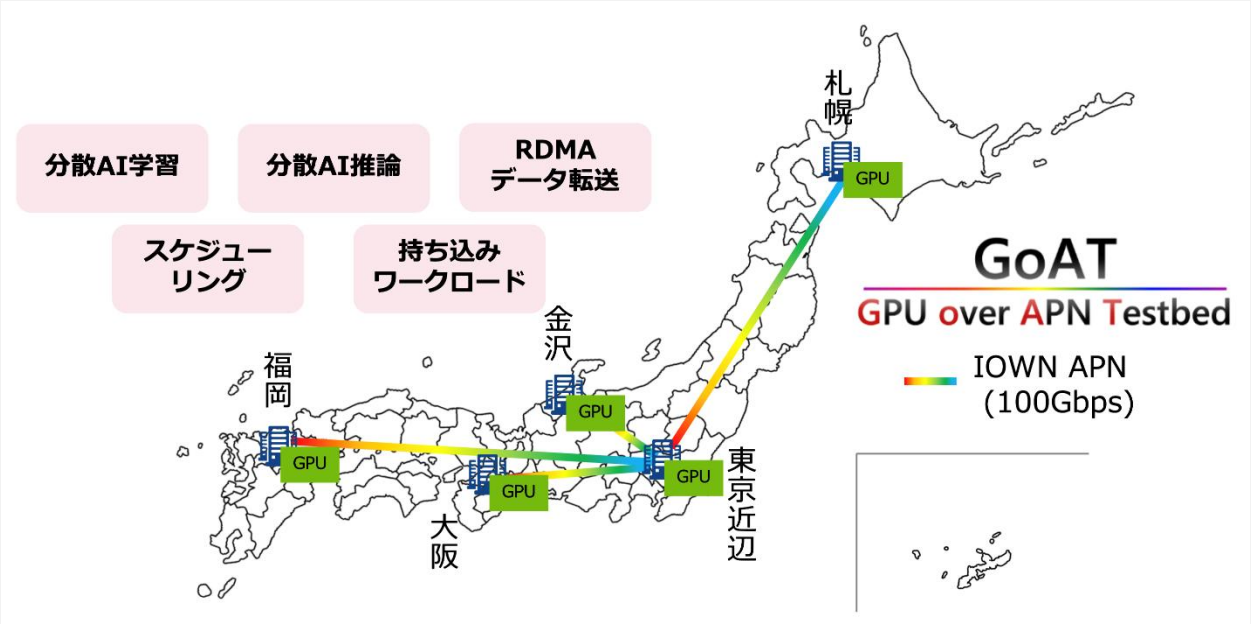
拠点に依存せずGPUを束ねて使えることで、分散配置を前提とした持続可能なAI基盤を実現。

想定利用シーン

AI開発者・事業者が、拠点を意識せず全国に分散したGPUを束ねて、学習や推論などをオンデマンドに実行。

全国広域分散GPU実証基盤

IOWN APNの低遅延・高帯域回線を活用し、地理的に離れた拠点でもGPUを一体的に利用できる環境。



参考： 圧縮・振分技術による省メモリ化・高速化

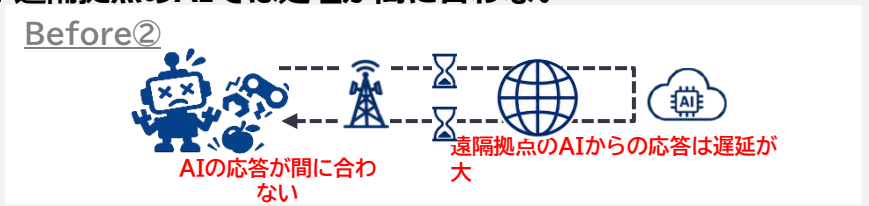
① 背景「AI負荷の増大。AIの応答速度不足」

※ GPU over APN Testbed
IOWN APNを活用し複数データセンターに分散したGPUクラスタを利用できるAI向け分散GPUインフラ

1. RAG／XRの普及により、AIが処理する情報が増え、リソース負荷が増大している。
2. フィジカルAIなど低遅延な応答を求めるアプリケーションにAIの応答が間に合わない

1. AI が「マルチモーダル」へ変容しリソース負荷が増大

2. 遠隔拠点のAIでは処理が間に合わない

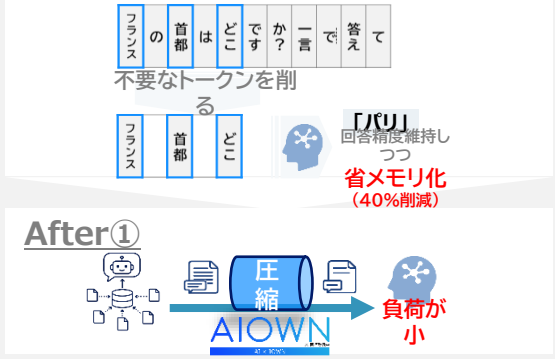


② 技術と効果「AIプロンプト圧縮で負荷削減＆近傍GPUへの振り分けで高速化」

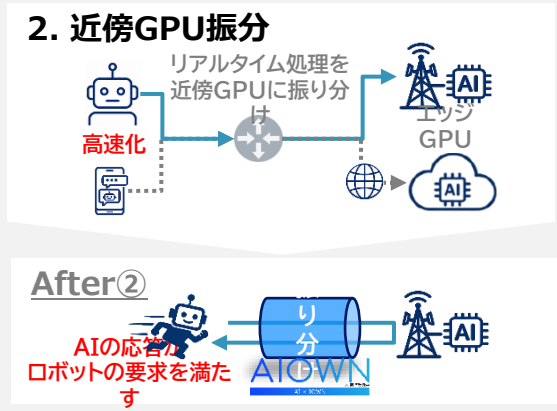
- 回答生成に不要なトークンを削る。リアルタイム処理を近傍GPUへ振り分ける。
- RAG アプリケーションを使って、「圧縮・振り分け」によるGPU省メモリ化のデモ展示をInteropにて行う。

機能説明と効果

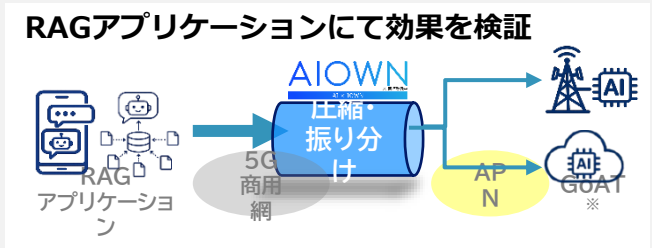
1. AIプロンプト圧縮



2. 近傍GPU振分

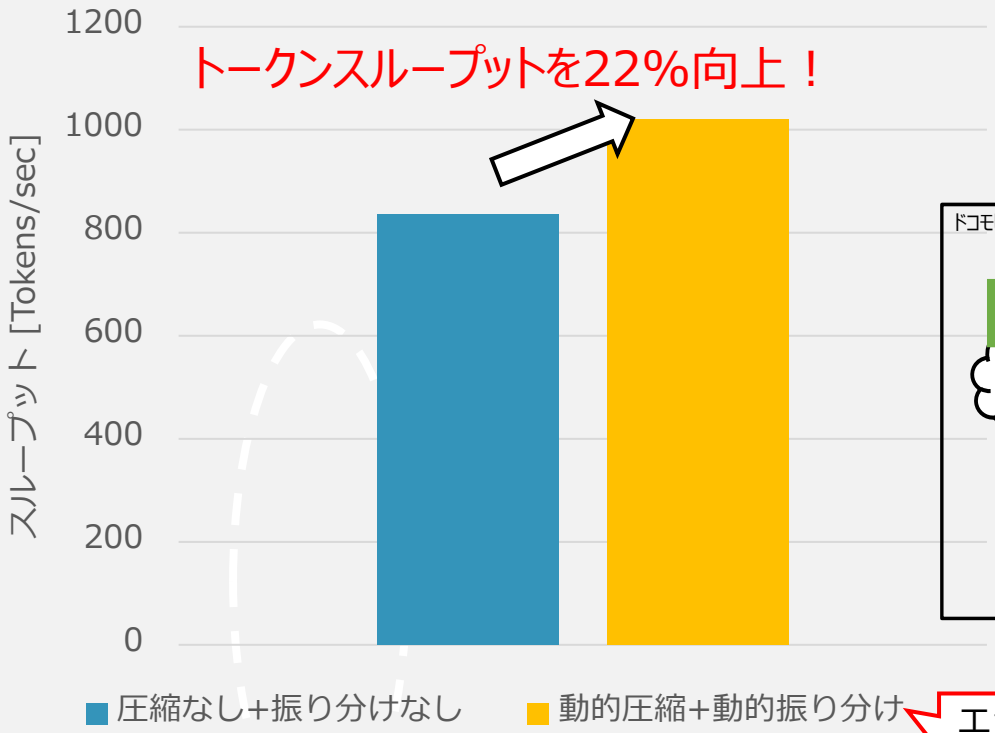


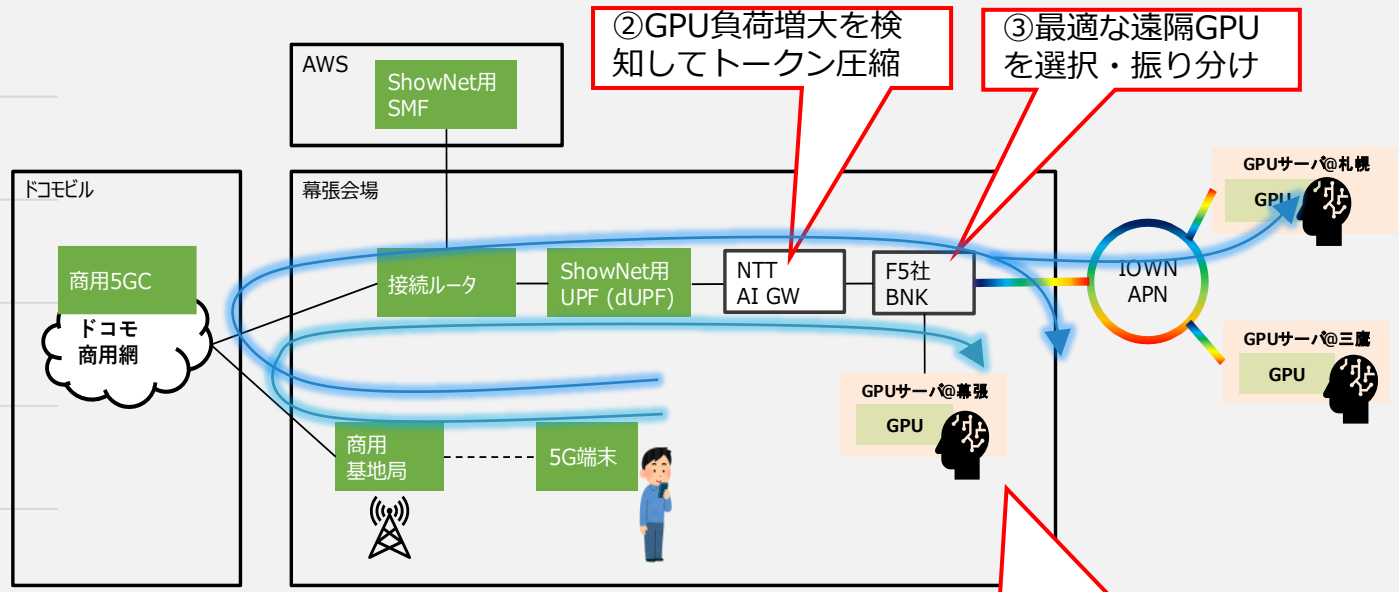
Interop展示イメージ



検証結果（スループット[Tokens/sec]）

- dUPFとIOWN APNを直結することで高速・低遅延・低ジッタで遠隔GPUと接続
- さらにNTT技術であるトークンの動的圧縮や、F5社のBNKと組合せることで、**AI Grid構成による遠隔GPUの利用においてAI処理能力（トークンスループット）の22%向上**を確認

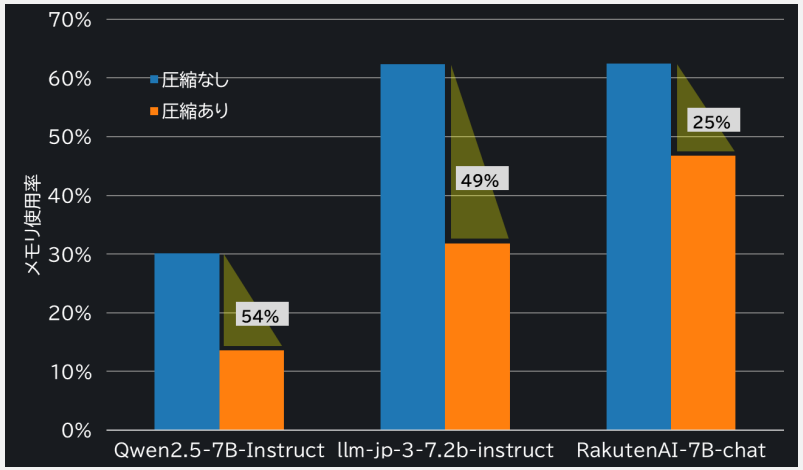




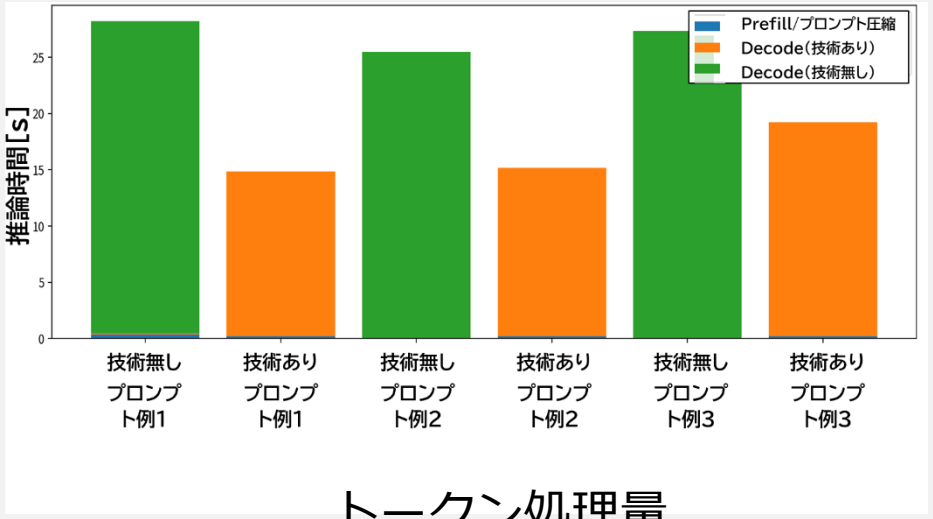
エッジGPU(幕張)の負荷状況に応じて
・不要なトークンを削除(圧縮)
・遠隔GPU(三鷹/札幌)を選択しトークンを振り分け

参考： 圧縮ゲートウェイの効果

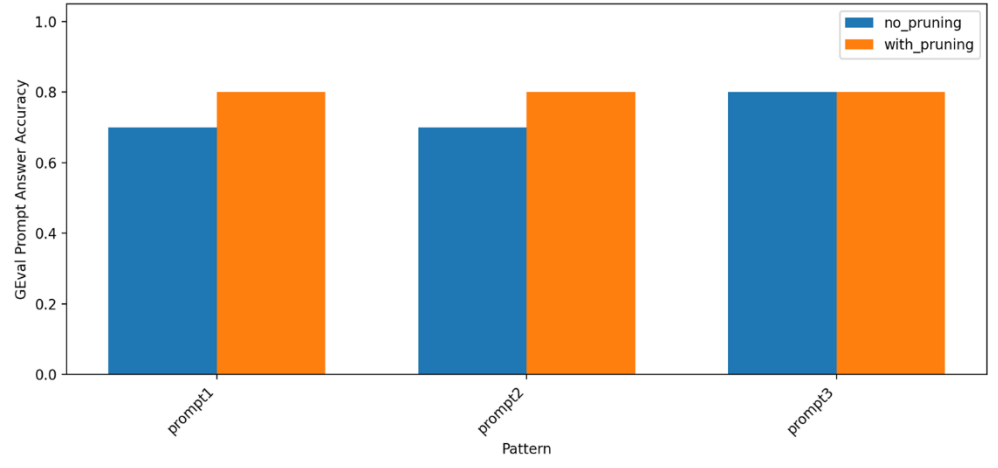
GPUメモリ使用率



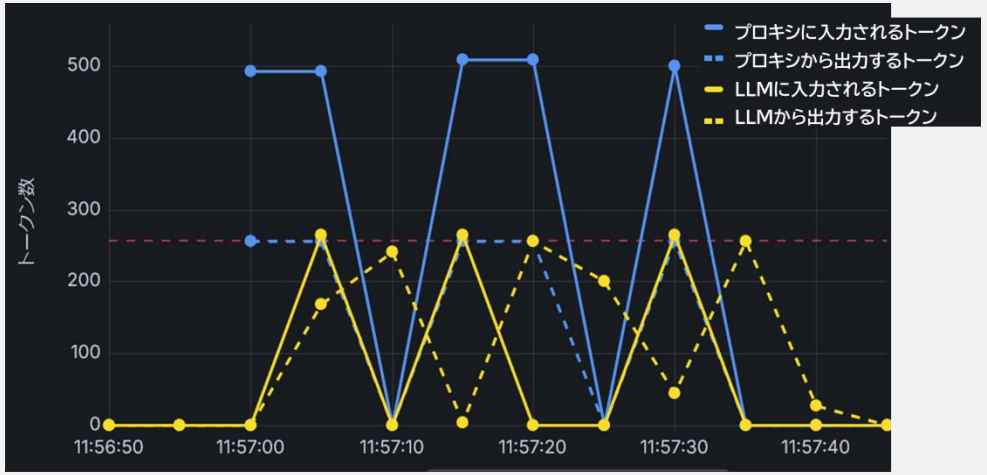
推論時間



推論精度



トークン処理量



まとめ

- AI時代では、通信事業者は**パケットではなくトークンを運ぶ時代**が到来？
- NVIDIAが発表した「AI Grid」について、NTT-Gの”**あり物の技術”を組み合わせて実現**した
- 性能面などある程度の実現性が見出された

＜気になっているところ＞

- 通信事業者はAI Gridに取り組むべきか？
- 分散AI基盤のユースケースとして、どういったものがあるのか？
- 分散化による設備コストの低減効果はどう考えるべきか（むしろ逆効果だったり？）
- 分散AI基盤の評価指標として何を重視すべきか？（使用率/遅延/電力/発電効率）

今後のJANOGで分散AI基盤に関する発表・議論が活発になると嬉しいです