OCPJ Meet Up 2025夏

IOWN技術を活用した AI向けGPUインフラの最新事例

2025年7月8日

NTTドコモビジネス株式会社
イノベーションセンター
IOWN推進室

社名変更について

つなごう。驚きを。幸せを。
NTT docomo Business

最先端テクノロジー・イノベーションを通じ、企業と地域が持続的に成長できる
自律分散型社会を支える「産業・地域DXのプラットフォーマー」へ

NTTコミュニケーションズ株式会社

NTTドコモビジネス株式会社

アジェンダ

つなごう。驚きを。幸せを。
NTT docomo Business

01 IOWNとNTTドコモビジネスの取り組み

02 AI向けGPUインフラの要件と課題

03 最新事例：GPU over APN 実証実験

01 IOWNとNTTドコモビジネスの取り組み

「IOWN®」は、NTT株式会社の商標又は登録商標です

IOWNに取り組む背景

つなごう。驚きを。幸せを。
docomo Business

現状

生成AI等のICTの進展による
けた違いのデータ量増加など、
電力消費の急増

世界のデータセンタの年間消費電力
（2018年→2050年）
約2,600倍※

既存の情報通信システムでは
伝送能力と処理能力の双方に限界に
＋半導体競争の過熱

2032年に向けて

持続可能な社会を実現するため
技術革新による新たな価値創造

環境負荷が大幅に下がる
ネットワークやコンピューティング
インフラの高度化の実現にむけて、

ネットワークから端末まであらゆる場所
に**光電融合**デバイスなどの
光（フォトニクス）関連技術を活用した
次世代のインフラへ

IOWN®

※出典：国立研究開発法人科学技術振興機構低炭素社会戦略センター

IOWN（Innovative Optical and Wireless Network）とは

つなごう。驚きを。幸せを。
docomo Business

スマートな社会の実現に向けた**光（フォトニクス）関連技術を活用した次世代情報通信基盤**

実世界とデジタル世界の掛け合わせによる未来予測等を実現

コグニティブ・
ファウンデーション（CF）

マルチオーケストレーター

あらゆるものをつなぎ、
その制御を実現

IOWN光コンピューティング

光電融合技術を活用した次世代コンピューティング
Data Centric Infrastructure（DCI）

オールフォトニクス・ネットワーク(APN)

ネットワークから端末まで、すべてにフォトニクス（光）ベースの技術を導入

光電融合技術
（光プロセッサ）

2024年3月1日
APN専用線プラン
powered by IOWN
提供開始

APNの特徴

つなごう。驚きを。幸せを。
docomo Business

通信ネットワークのすべての区間で光波長を占有することで**「大容量」「低遅延」「低消費電力」**を実現

IOWN APN

All Photonics Network

	低遅延 エンドエンド遅延※1	**大容量高品質** 伝送容量※2	**低消費電力** 電力効率※3
最終目標（2032年）	**1/200**	**125倍**	**100倍**

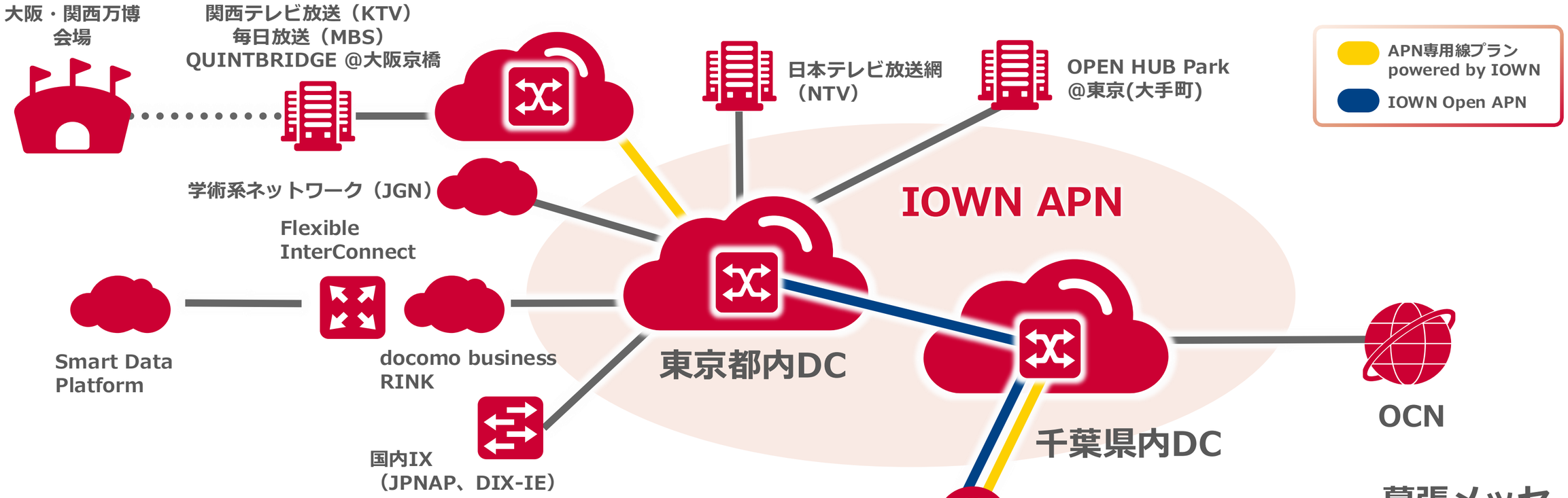
※1 同一県内で圧縮処理が不要となる映像トラヒックでのエンドエンドの遅延の目標値 ※2 光ファイバー1本あたりの通信容量の目標値 ※3 フォトニクス技術適用部分の電力効率の目標値

出典：NTT株式会社を基に編集

ShowNet 2025

Interop Tokyo 2025 ShowNetでの取り組み

つなごう。驚きを。幸せを。
docomo Business



Interop Tokyo 2025出展APN構成の特長

- 光波長多重化による大容量伝送
- All-Photonics Network技術による低遅延・低ジッタ接続
- マルチベンダな装置による波長を収容可能な構成
- ShowNetへ対外接続を提供
- 遠隔拠点との連携デモなどの実アプリケーション

おかげさまでグランプリを受賞しました

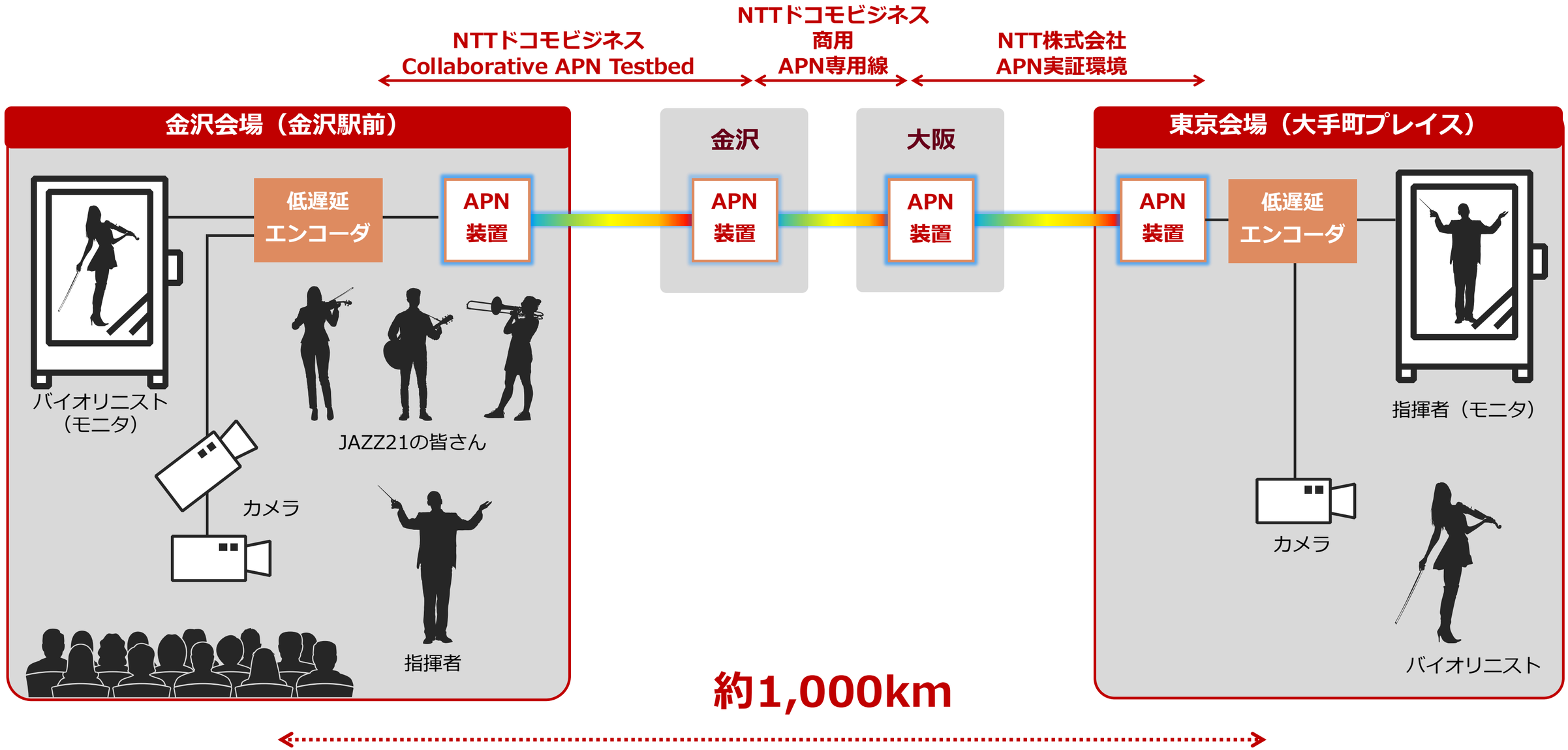
Interop2025 Best of Show Award

APN(All Photonics Network)部門

APN専用線プラン powered by IOWN

リアルタイム遠隔音楽ライブの実証実験（構成イメージ）

つなごう。驚きを。幸せを。
docomo Business



製薬・創薬研究拠点 × IOWN：湘南アイパーク

つなごう。驚きを。幸せを。
docomo Business

三菱商事 NTT NTT Communications MORGENROT iPark MC DIGITAL REALTY A Mitsubishi Corporation and Digital Realty JV

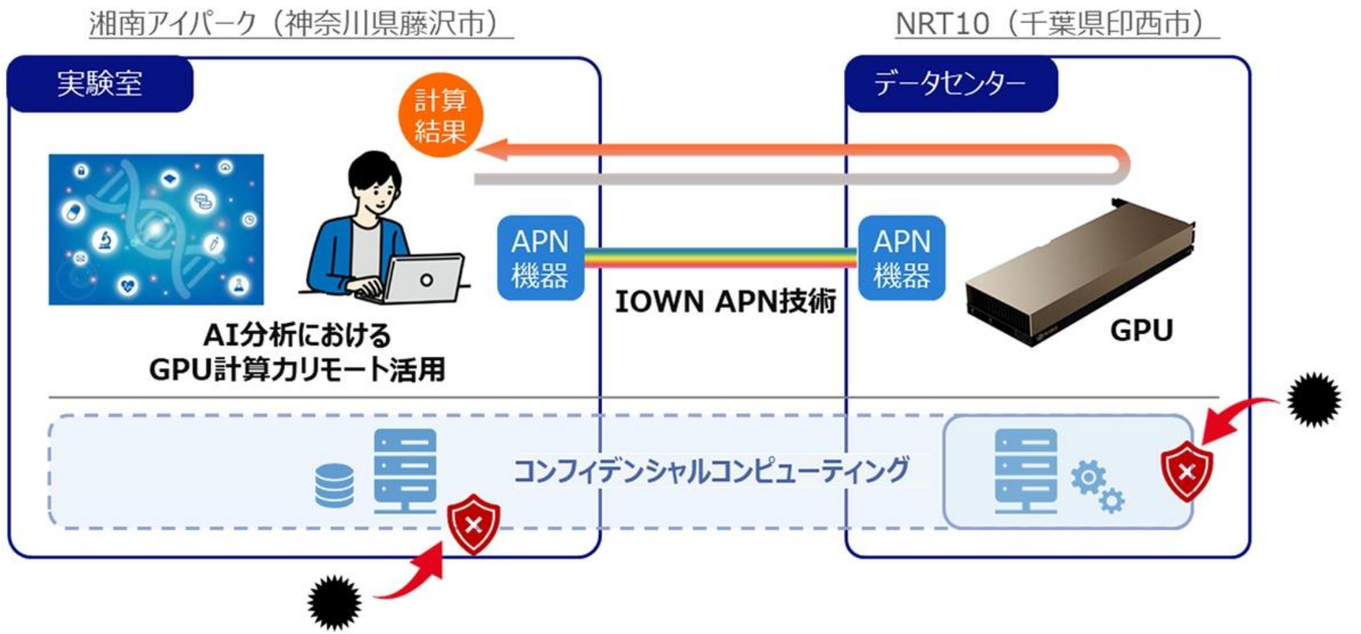
2025年2月17日

三菱商事株式会社
日本電信電話株式会社
NTTコミュニケーションズ株式会社
モルゲンロット株式会社
アイパークインスティチュート株式会社
MCデジタル・リアルティ株式会社

GPU計算力リモート提供の共同実証実験を開始

～IOWN APNを活用し、湘南アイパークの製薬・創薬研究データを安全に分析～

三菱商事株式会社(本社：東京都千代田区、代表取締役社長：中西 勝也、以下「三菱商事」)、日本電信電話株式会社(本社：東京都千代田区、代表取締役社長：島田 明、以下「NTT」)、NTTコミュニケーションズ株式会社(本社：東京都千代田区、代表取締役社長：小島 克重、以下「NTT Com」)、モルゲンロット株式会社(本社：東京都千代田区、代表取締役社長：森本 竜英、以下「モルゲンロット」)、アイパークインスティチュート株式会社(本社：神奈川県藤沢市、代表取締役社長：藤本利夫、以下「アイパークインスティチュート」)の5社は共同で、湘南ヘルスイノベーションパーク(以下「湘南アイパーク」)及びMCデジタル・リアルティ株式会社(本社：東京都港区、代表取締役社長：畠山孝成、以下「MCDR」)が運用するデータセンターにおけるGPU※1計算力リモート提供に関する共同実証実験を開始しました。



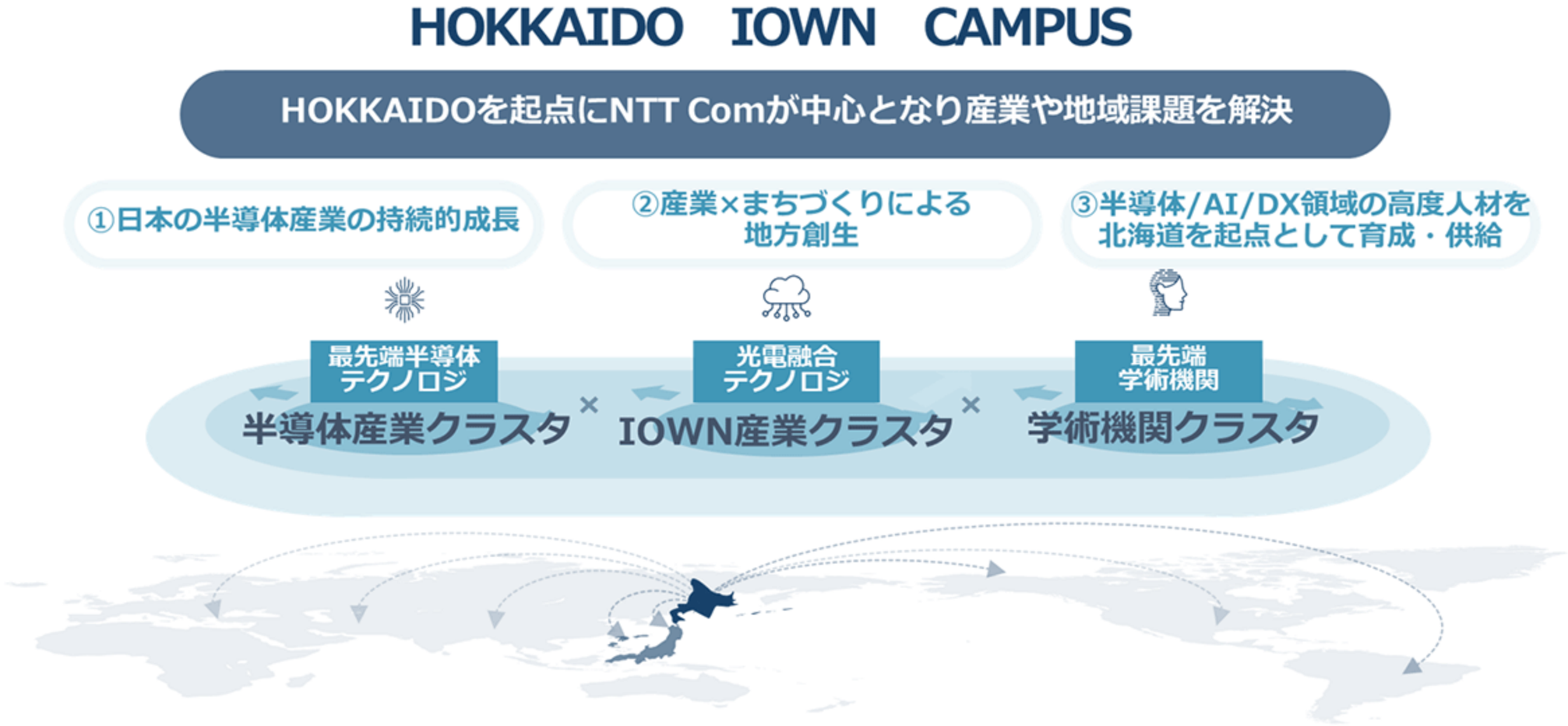
- 創薬AI分析などの産業向けAIプロセスに適合するワークロードの実証
- 遅延・フレームロスなどのネットワーク性能劣化がコンピューティング処理への影響検証
- 製薬・創薬業界が求めるセキュリティ要件を満たせるかのコンフィデンシャルコンピューティングの実行可能性検証

ニュース 2025年2月17日:GPU計算力リモート提供の共同実証実験を開始｜NTTドコモビジネス 企業情報

「HOKKAIDO IOWN CAMPUS」の取組

つなごう。驚きを。幸せを。
docomo Business

北海道を起点とし、さまざまな産業や地域課題を解決する事業コンセプト「HOKKAIDO IOWN CAMPUS」を発表します。産業振興×まちづくりにIOWNを活用し、半導体産業をはじめとするさまざまな産業/地域課題を解決し、北海道の地方創生に貢献



ニュース 2024年8月1日:さまざまな産業や地域課題を解決する事業コンセプト「HOKKAIDO IOWN CAMPUS」の発表について｜NTTドコモビジネス 企業情報

IOWN産業クラスタ：千歳市の自動運転バス実証実験

つなごう。驚きを。幸せを。
NTT docomo Business

2024.11.15 報道発表

実証実験の目的

全国的に少子高齢化、人口減少が進む中で、公共交通分野においては、運転手の高齢化、人手不足が深刻化しています。本市においても、運転手不足などから、路線バスにおいて減便されており、今後、運転手の年齢構成から、運転手不足はさらに加速するものと推測しています。

そこで、本市においては、バスの運転を自動化することにより、運転に係る人数を削減し、持続可能な地域公共交通を確保するため、自動運転バスの実証実験を実施しています。

実施体制

実施主体：千歳市 City of Chitose

参画企業：

A-Drive / AISAN / 北海道中央バス / 千歳相互観光バス株式会社 / docomo business / NTT Communications / TOKAI RIKA / SOMPOリスクマネジメント / NDS 日本データーサービス株式会社

路線バスの運転手不足に対応する
IOWNや5Gワイドなどを活用した
路線バス自動運転実証を
千歳市で実施

公共交通機関のさらなる
利便性向上・運転手不足
対応に向けて、

IOWNなどの
最先端デジタルテクノロジー
を活用

その先のさらに住みよい
まちづくりに貢献

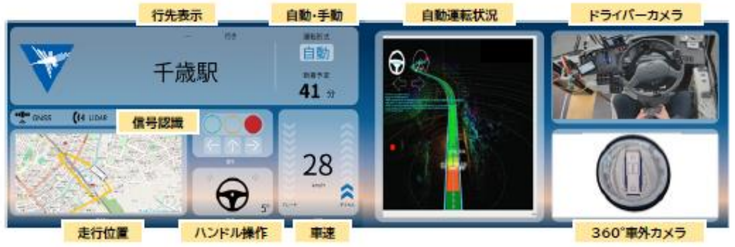
遠隔監視システム

- 自動運転バスの車室内外で撮影されたカメラ映像をモバイル回線(5G，LTE等)を介し遠隔地に伝送します。その際に課題となる回線速度による映像伝送の遅延やデータ圧縮による画像劣化に対し、カメラ映像を画像処理技術により統合化することで、高画質で遅延のない伝送が可能です。

実証期間の10:00~16:00、千歳駅東口待合室、
千歳市役所 市民ロビーに設置

デジタルサイネージ

- 乗客に対し、自動運転制御状況を分かりやすく知らせるため、センシング結果やドライバーの手元を撮影したカメラ映像などを表示するディスプレイを車室内に搭載し、自動運転バスと連携する車両情報(アクセル/ブレーキ，舵角等)を表示します。



ニュース 2024年11月15日:路線バスの運転手不足に対応するIOWNや5Gワイドなどを活用した路線バス自動運転実証を千歳市で実施｜NTTドコモビジネス 企業情報

お問い合わせ案件の多い業種・業界

つなごう。驚きを。幸せを。
NTT docomo Business

2030年をターゲットに
中長期ビジョン／デジタル戦略／インフラ戦略／共創
新たな収益獲得、ビジネスモデルなどを視野に

スマートシティ

・スマートサービスを支える基盤
・デジタルツイン
・エンターテイメント

金融

・取引の公平性の担保（確定遅延）
・遠隔拠点間の大量データ取引
・事業継続（BCP）への対応

製造

・ FA機器の遠隔リアルタイム制御
・工場拠点間ネットワーク

放送・メディア

・リモートプロダクション
・高精細映像配信
・没入型メディア

IT事業者

・次世代ネットワーク
・データセンター間接続
・次世代コンピューティング

電子・電機

・デバイス×光電融合技術
・光半導体のエコシステム

今後の活用イメージ（将来像）

02 AI向けGPUインフラの要件と課題

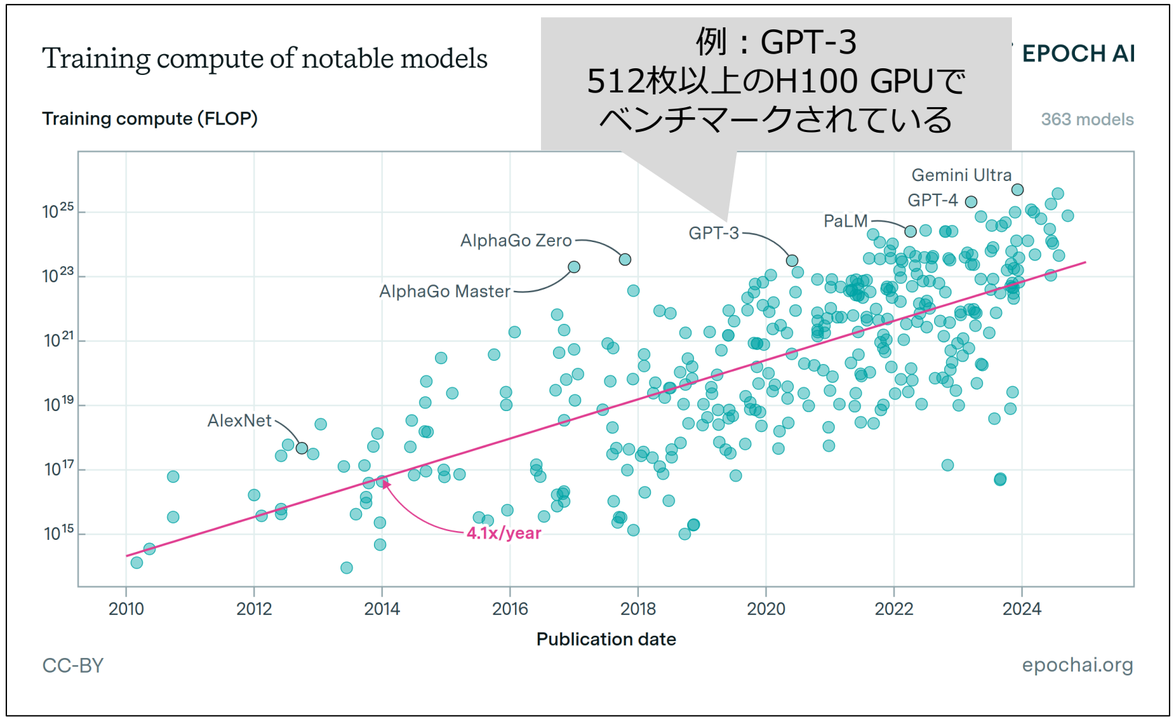
AI向けGPUインフラ需要増のトレンド

つなごう。驚きを。幸せを。
NTT docomo Business

生成AI、データ利活用、画像処理などの分野で、**多くの計算資源が必要**となっている

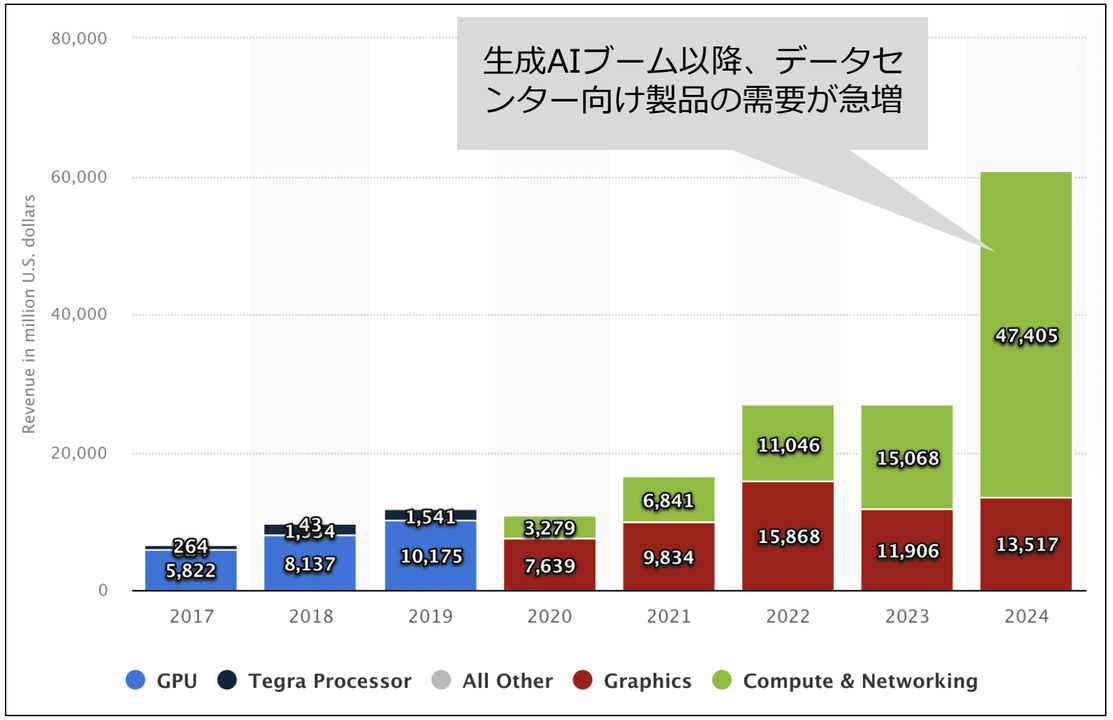
また、1台のGPUサーバーでは搭載できるGPUの数に限りがあるため、
複数台のGPUサーバーを並べて同時に使うシーンがますます増えている

AIモデル学習にかかる計算量の推移



https://epochai.org/data/notable-ai-modelsより引用・注釈追加

GPU業界最大手NVIDIA社の売上高の推移



https://www.statista.com/statistics/988034/nvidia-revenue-by-segment/より引用・注釈追加

AI向けGPUインフラの特徴

つなごう。驚きを。幸せを。
docomo Business

学習・推論等のAIワークロードのためのICTインフラの特徴

1. **高い計算能力と並列処理性能**
   GPUや専用アクセラレータを多用し、膨大な演算を高速に処理する

2. **大量の電力消費と発熱**
   特に学習時は消費電力・排熱が大きく、インフラ側の電源・冷却設計が重要

3. **高速インターコネクトと大容量ストレージ**
   巨大なデータセットやモデルの保存・転送のため、ストレージやネットワーク帯域も高性能が求められる

4. **柔軟なスケーラビリティと運用管理**
   ワークロードの変動や新技術への対応のため、インフラの拡張性・自動運用機能が重視される

AI向けGPUインフラの要件と課題 1/2

つなごう。驚きを。幸せを。
docomo Business

インフラ制約の観点では

ラックあたり電力密度	最新AIラックは250-900kW/ラックに達する（2026–27年予測※1） →従来DCの電力供給能力を超過
冷却能力	通常7-20kW/ラックの空冷ユニットでは120-140kWの熱除去が不可 →液冷対応不可時は負荷分散が必要
床荷重制限	液冷ラックの荷重が、従来DCの標準床荷重を逼迫または超過する可能性

→「やむをえず」分散へ

※1. https://www.tweaktown.com/news/101799/ai-servers-of-the-future-rack-density-1000kw-with-nvidias-next-gen-rubin-ultra-gpus/index.html

AI向けGPUインフラの要件と課題 2/2

つなごう。驚きを。幸せを。
docomo Business

事業継続・災害復旧の観点では

AIワークロードの継続性確保	拠点障害時に学習や推論を即時再開し、サービス停止や学習中断リスクを最小化したい
巨大なAIデータセットやモデルのバックアップ	複数拠点にデータやモデルを分散配置し、災害時のデータ消失リスクを低減したい
データ主権要件や法規制への柔軟な対応	企業ごとや地域ごとにデータやモデルを管理し、企業/国の規制やガバナンスに適合させたい
リソース有効活用と分散処理による精度向上	型落ちや零細なリソースも含めて活用することでAIサービスの効率性を向上させたい

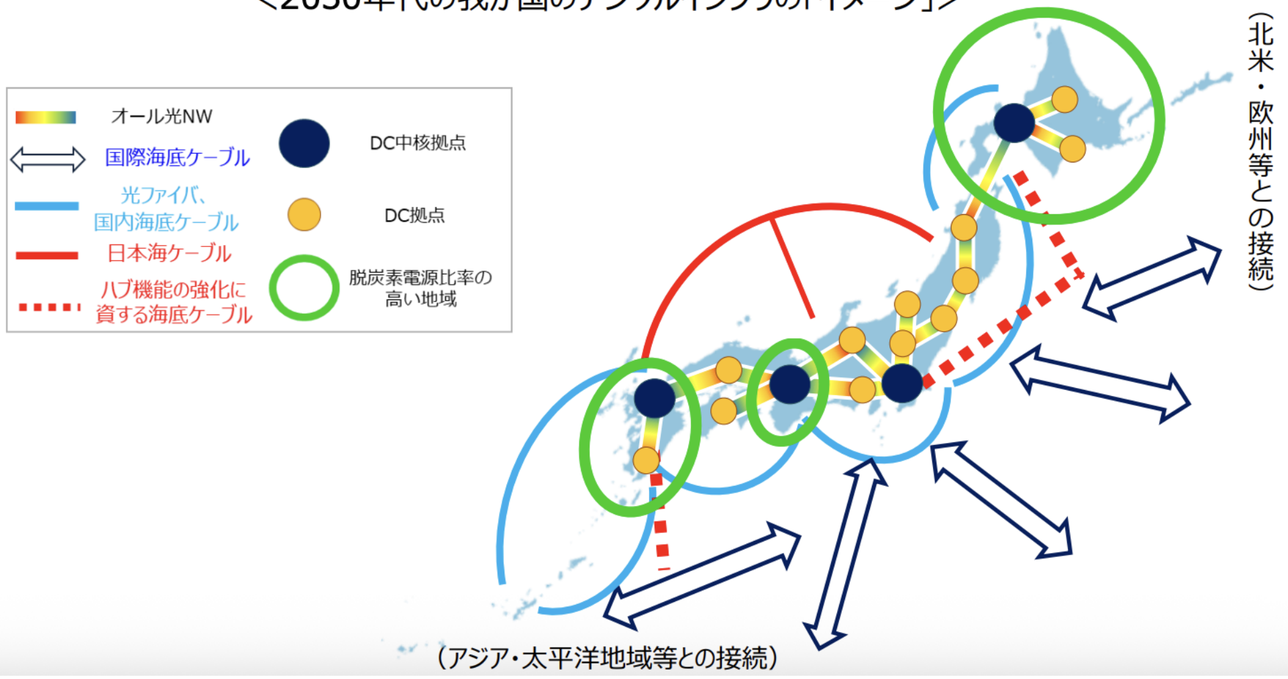
→「あえて」分散へ

[参考] AIインフラにかかわる総務省見解

つなごう。驚きを。幸せを。
docomo Business

- デジタルインフラは、「社会インフラのインフラ」として、**我が国にとって必要不可欠な礎**。政府として、デジタルインフラの整備に向けて、**本提言を受けての施策を早急に検討**し、**具体化**することが重要。
- AI・半導体・オール光ネットワーク・量子コンピューター等の技術を軸に、AI革命に続く**大きなパラダイムシフトが数年以内に起こり得ることを念頭に置いた柔軟性の確保**が重要。
- 今後は、**AIの利活用や人材育成・研究開発等にも目を向けていく**ことが重要。
- **関連する他の政策の枠組みや検討の動向を注視**しつつ、**取組の進捗状況等についてフォローアップ**を実施。関係省庁や事業者等とも連携しながら**更なる戦略の検討や必要に応じた適時の見直し**を行っていくことが必要。

<2030年代の我が国のデジタルインフラの「イメージ」>

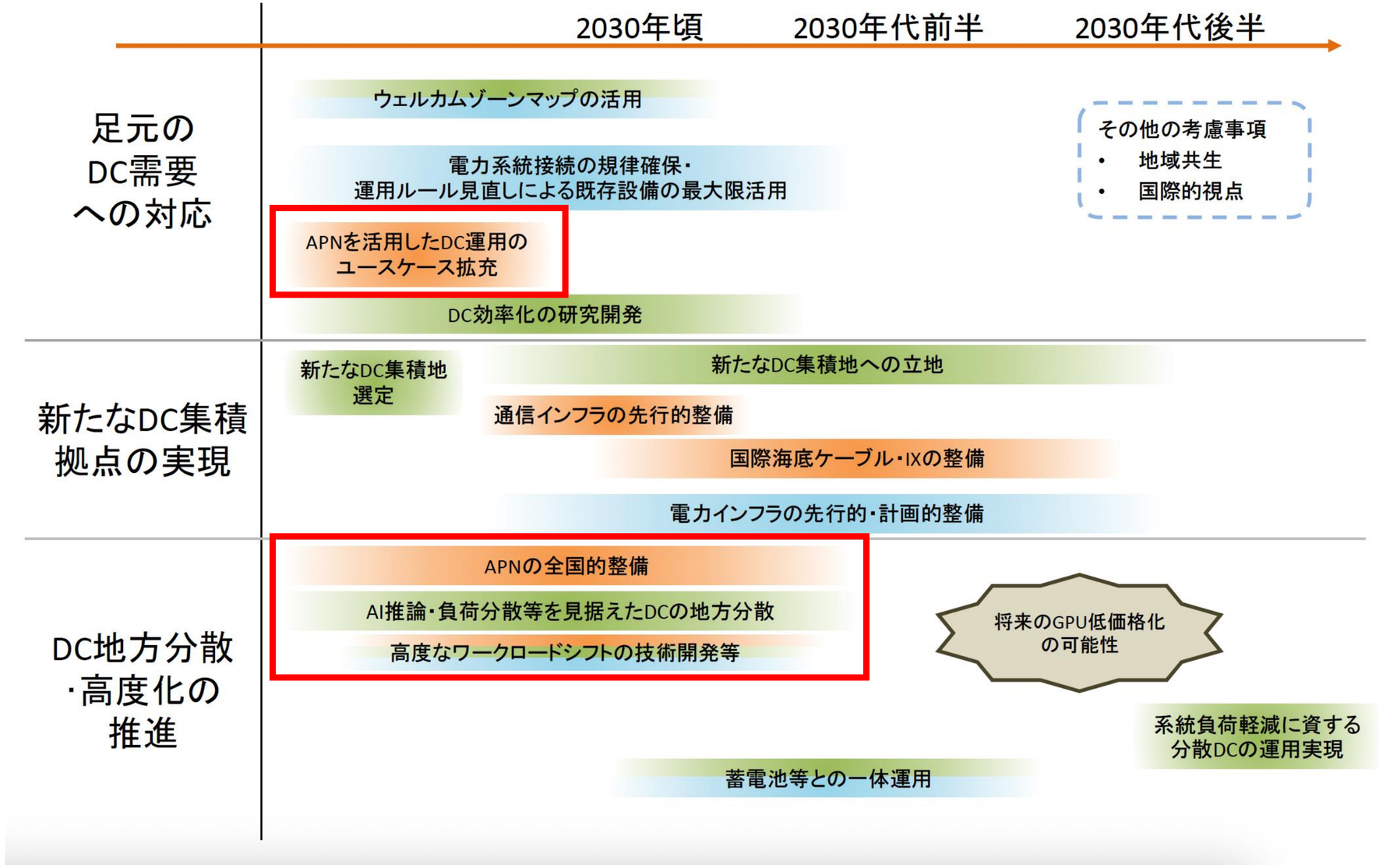


デジタルインフラ（DC等）整備に関する有識者会合 中間取りまとめ3.0（令和6年10月4日）より引用

[参考] AIインフラにかかわる総務省見解

ワット・ビット連携の実現に向けた進め方のイメージ

電力　DC　通信



ワット・ビット連携官民懇談会 取りまとめ1.0（令和7年6月12日）より引用

03 最新事例：GPU over APN 実証実験

GPU over APN コンセプト

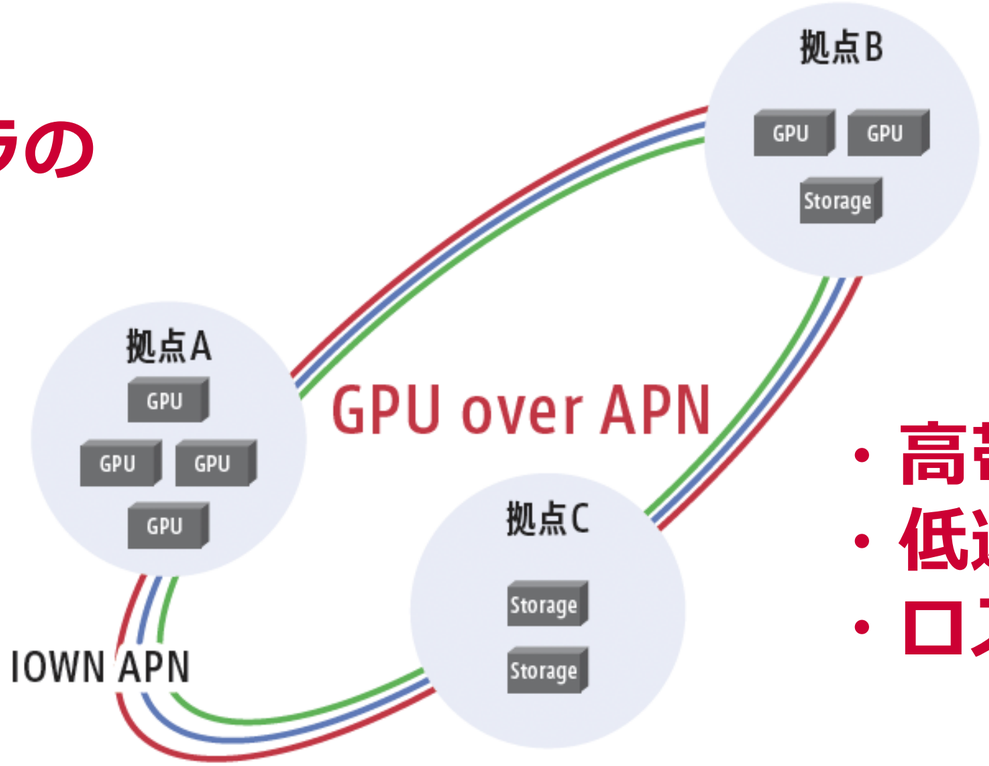
つなごう。驚きを。幸せを。
docomo Business

コンセプト/
訴求ポイント

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クラウドを実現

- 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
- 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可能
- 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

AI向けGPUインフラの
分散ニーズに対して
選択肢を提供したい

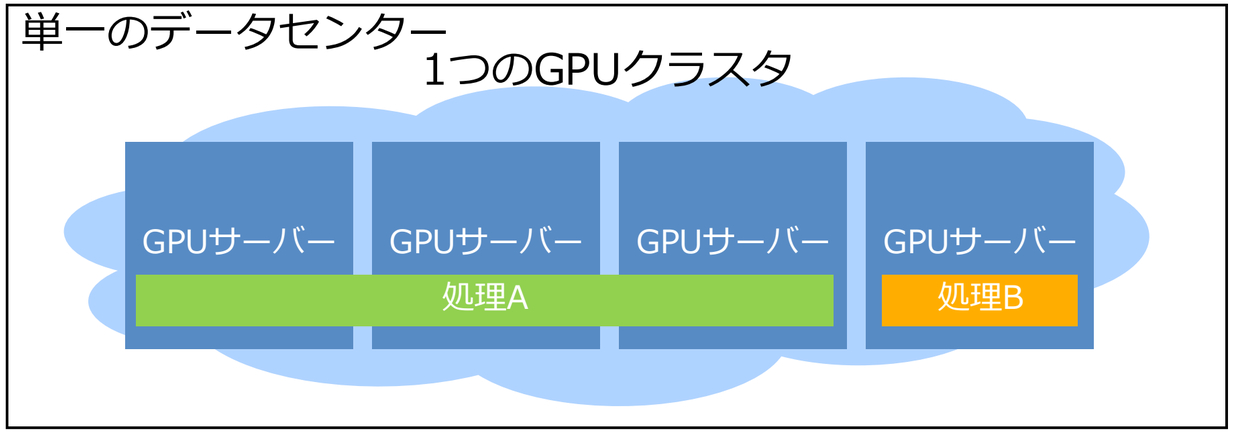


- **高帯域**
- **低遅延**
- **ロスレス**

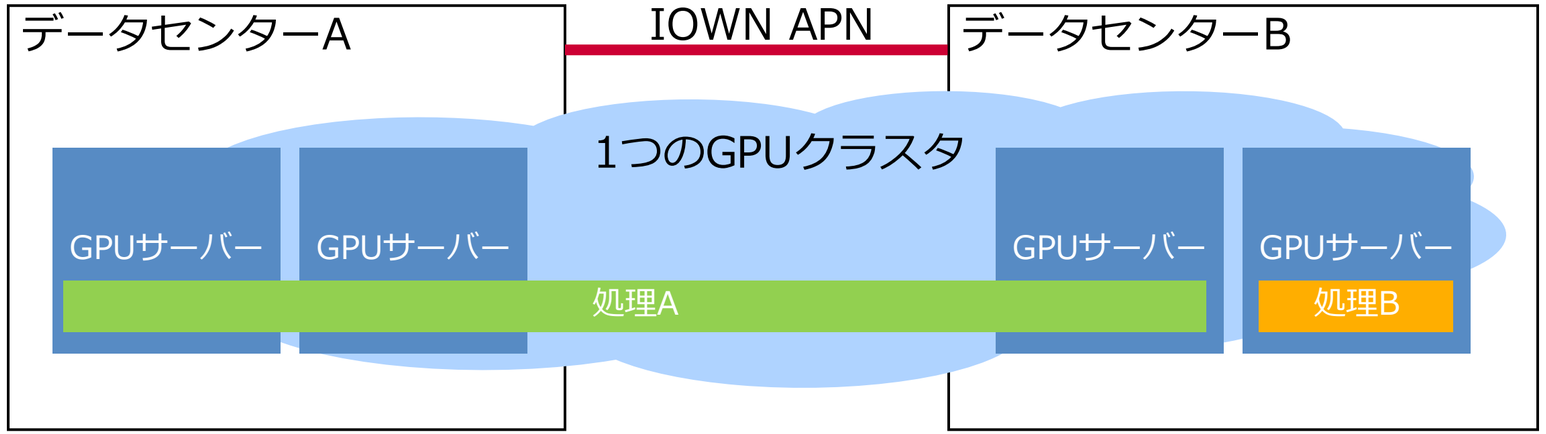
GPUクラスタを分散するとは

本実証におけるGPUクラスタの分散配置とは、
「1つの」GPUクラスタを複数のデータセンターにまたがって配置することを指す

従来



本実証の構想



GPU over APN ユースケース

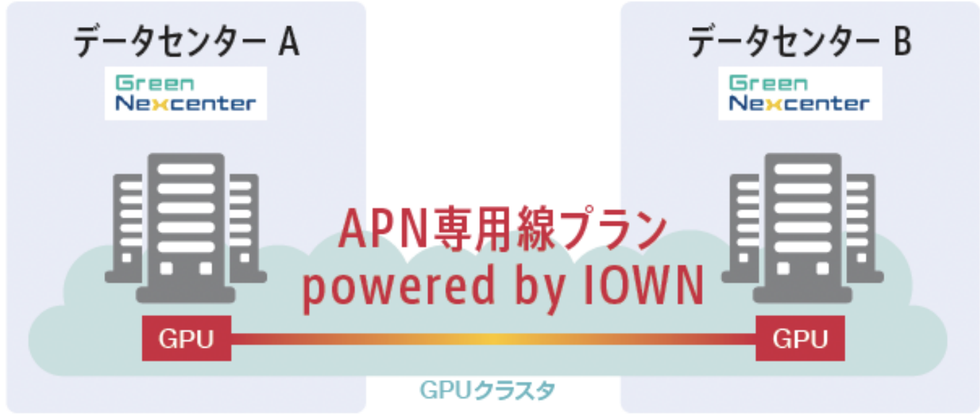
つなごう。驚きを。幸せを。
docomo Business

以下大きく2つのユースケースを想定し、実証を行った。

ユースケース1

複数のデータセンターで
1つのGPUクラスタを構成し、
クラスタ内高速通信を前提と
して計算を実施

クラスタ拡張
拠点分散



GPUサーバ同士を
引き離してみる

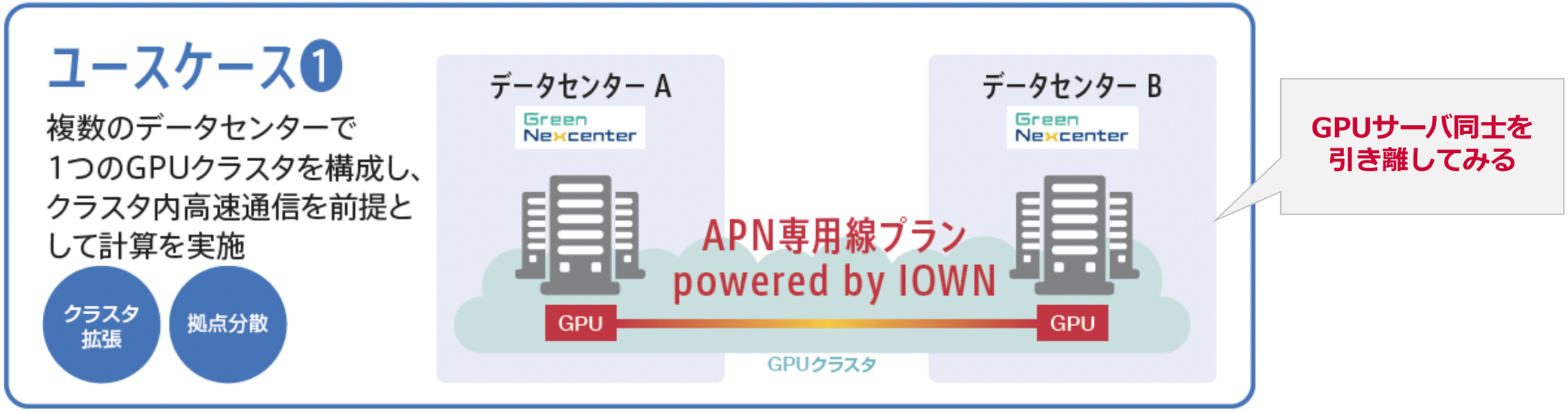
ユースケース2

他拠点に置いたストレージ
から高速にデータの読み書き
を実施

LLM学習
大規模データ

GPUサーバと
ストレージを
引き離してみる

ユースケース①



AI学習 のワークロード

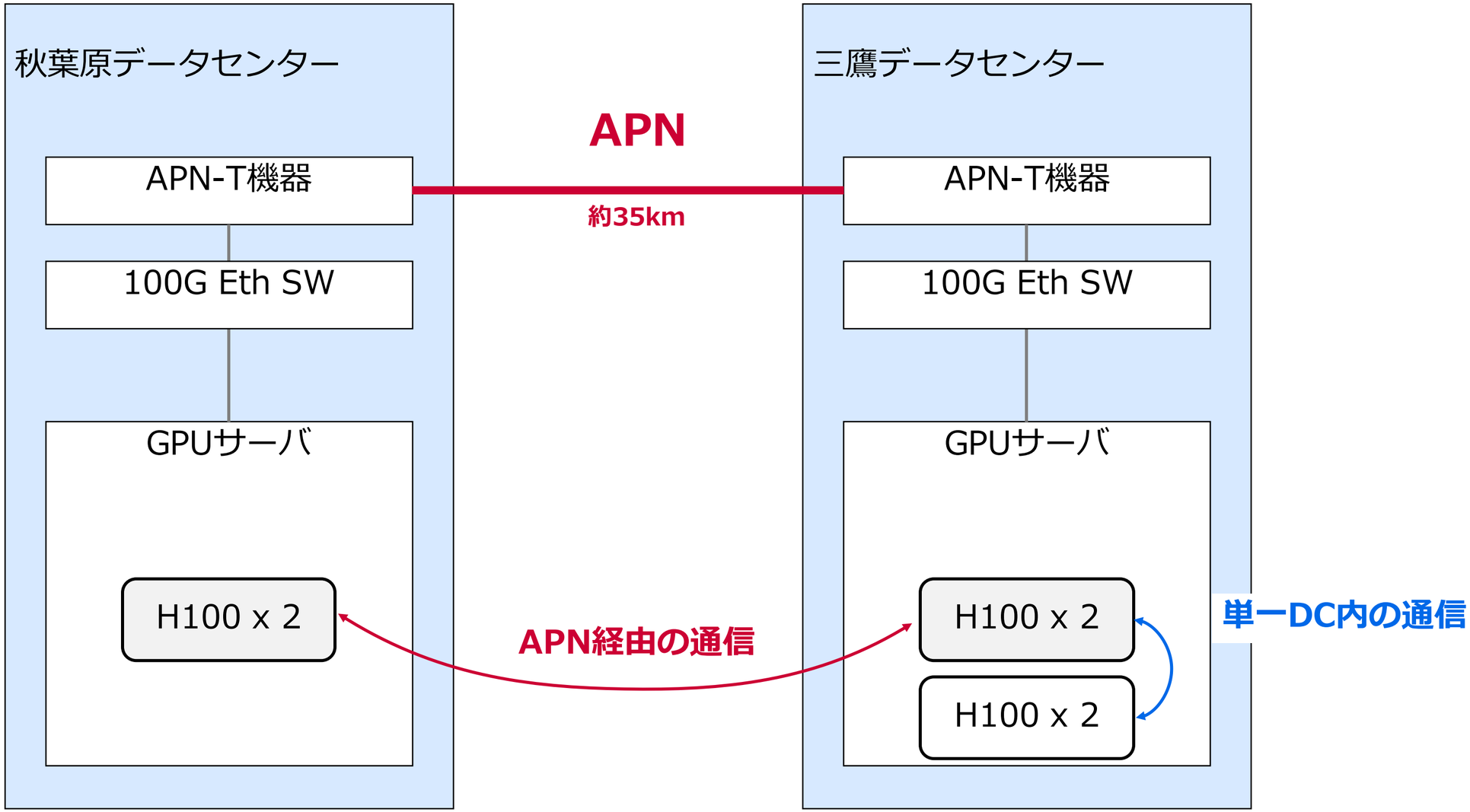
【0: ベース実験】秋葉原・三鷹の **2拠点** におけるAIモデル学習のスループット性能を測定
【1: 多拠点対応】秋葉原・三鷹・川崎の **3拠点** におけるAIモデル学習のスループット性能を測定
【2: 遠距離対応】3000kmの **超遠距離** を模擬した2拠点間でAIモデル学習のスループット性能を測定

AI推論 のワークロード

【3: 分散推論対応】2.の環境にて **AI推論を並列実行** し推論の精度とスループット性能を測定

0: 2拠点で分散学習 実証環境

つなごう。驚きを。幸せを。
docomo Business



2ノード4枚 vs. **2ノード4枚**

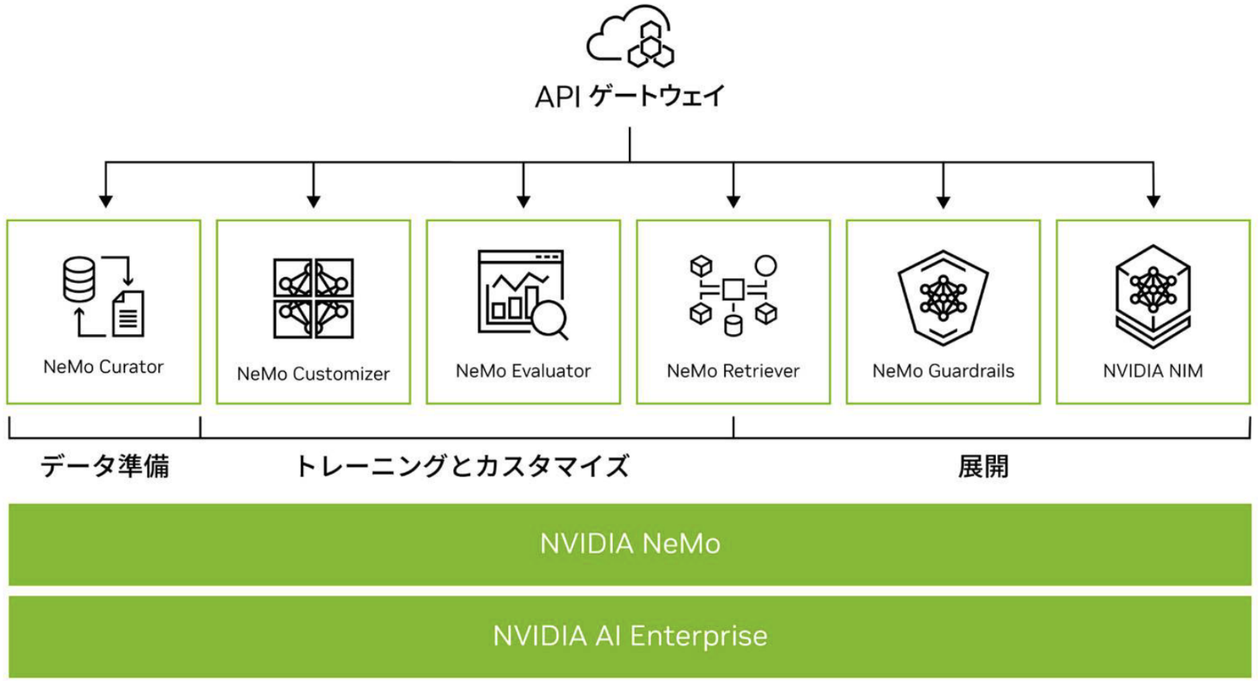
0: 2拠点で分散学習 実証環境

つなごう。驚きを。幸せを。
docomo Business

複数GPUサーバーを用いた分散学習に対応したNVIDIA NeMo™ を活用して、実証環境を構築・利用した

NVIDIA NeMo™ :

- あらゆる環境でカスタム生成AIを開発できるエンドツーエンドのプラットフォーム
- 正確なデータキュレーション、最先端のカスタマイズ、検索拡張生成（Retrieval Augmented Generation：RAG)、高速化されたパフォーマンスを備えたエンタープライズ対応モデルを提供

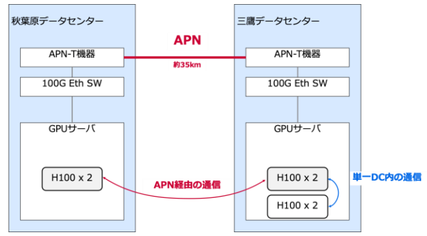


https://resources.nvidia.com/ja-jp-nemo/nemo-jp より引用

0: 2拠点で分散学習 実証結果

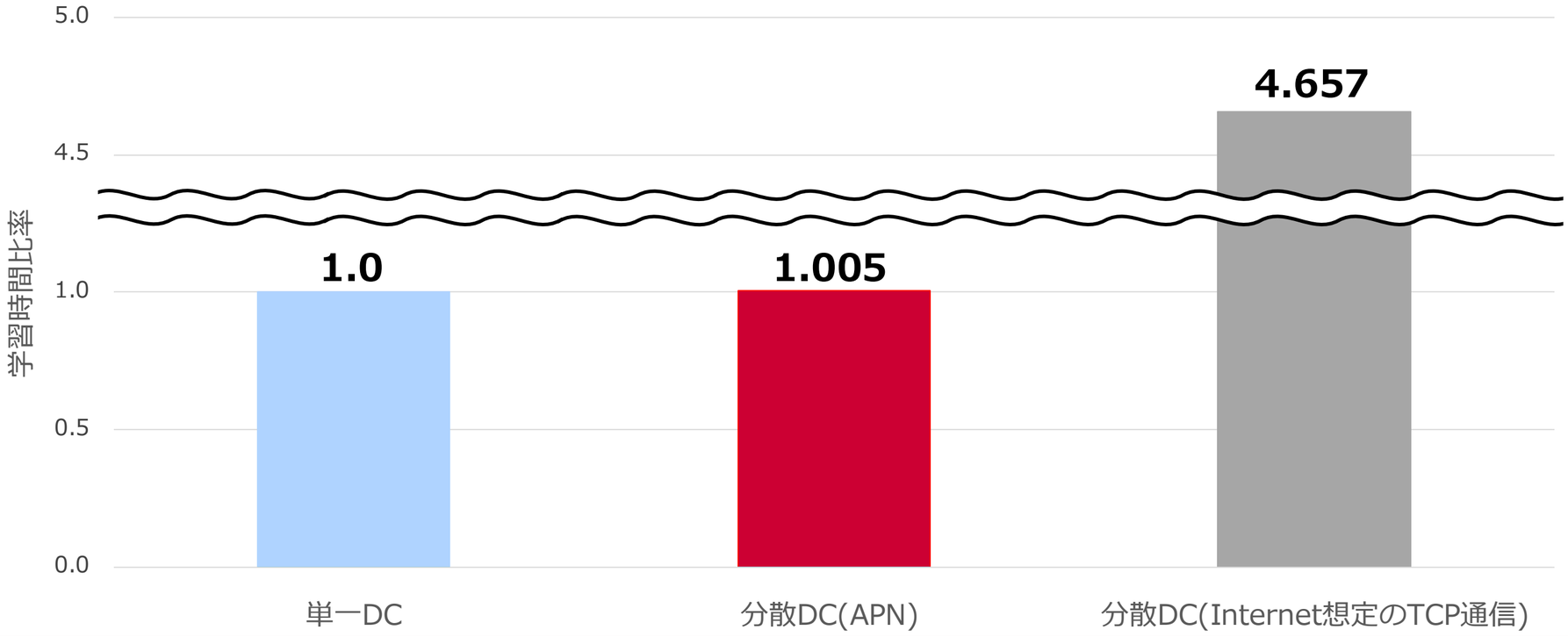
つなごう。驚きを。幸せを。
NTT docomo Business

- **LLM(tsuzumi 7B)の事前学習** を、NVIDIA H100 Tensor コア GPU x2基 x2ノードで実施
- 処理完了の所要時間（1ステップあたりの時間 x 所定ステップ数）を計算
- 単一DC、分散DC(APN経由※0)、分散DC(Internet想定のTCP通信※1)、の三者の性能を比較



単一DCでの学習の所要時間を1とした場合、分散DC(APN) は **約1.005倍** とほぼ互角

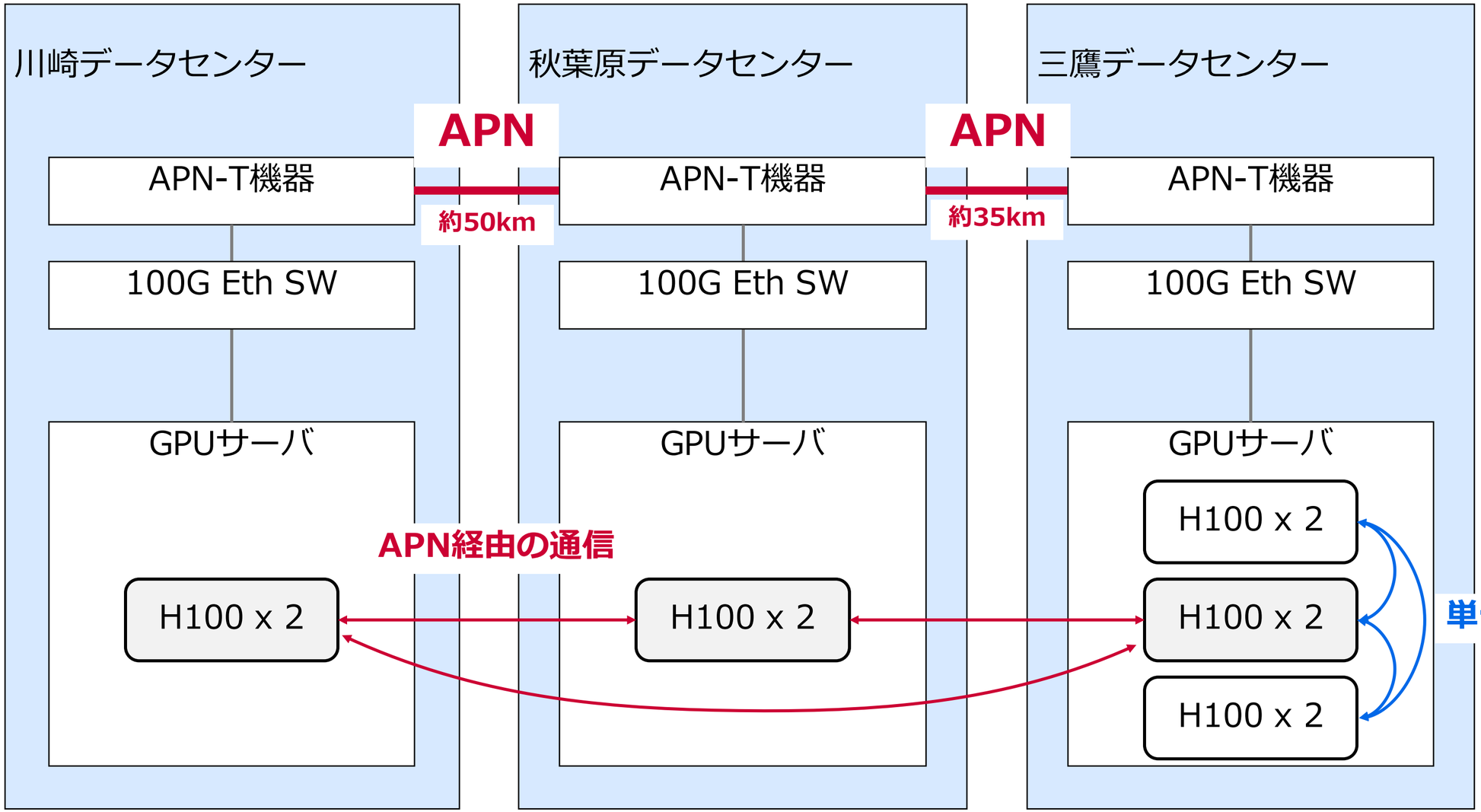
分散DC(Internet想定のTCP通信)では **約4.7倍** かかる



※0
Bandwidth=100Gbps
Ping RTT=0.526ms

※1
Bandwidth=300Mbps
(tcコマンドで制限)
Ping RTT=15.0ms
(tcコマンドで両端7.25ms挿入)

1: 多拠点で分散学習 実証環境

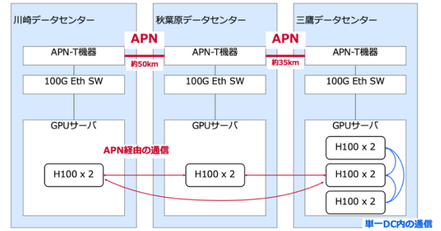


3ノード6枚 vs. **3ノード6枚**

1: 多拠点で分散学習 実証結果

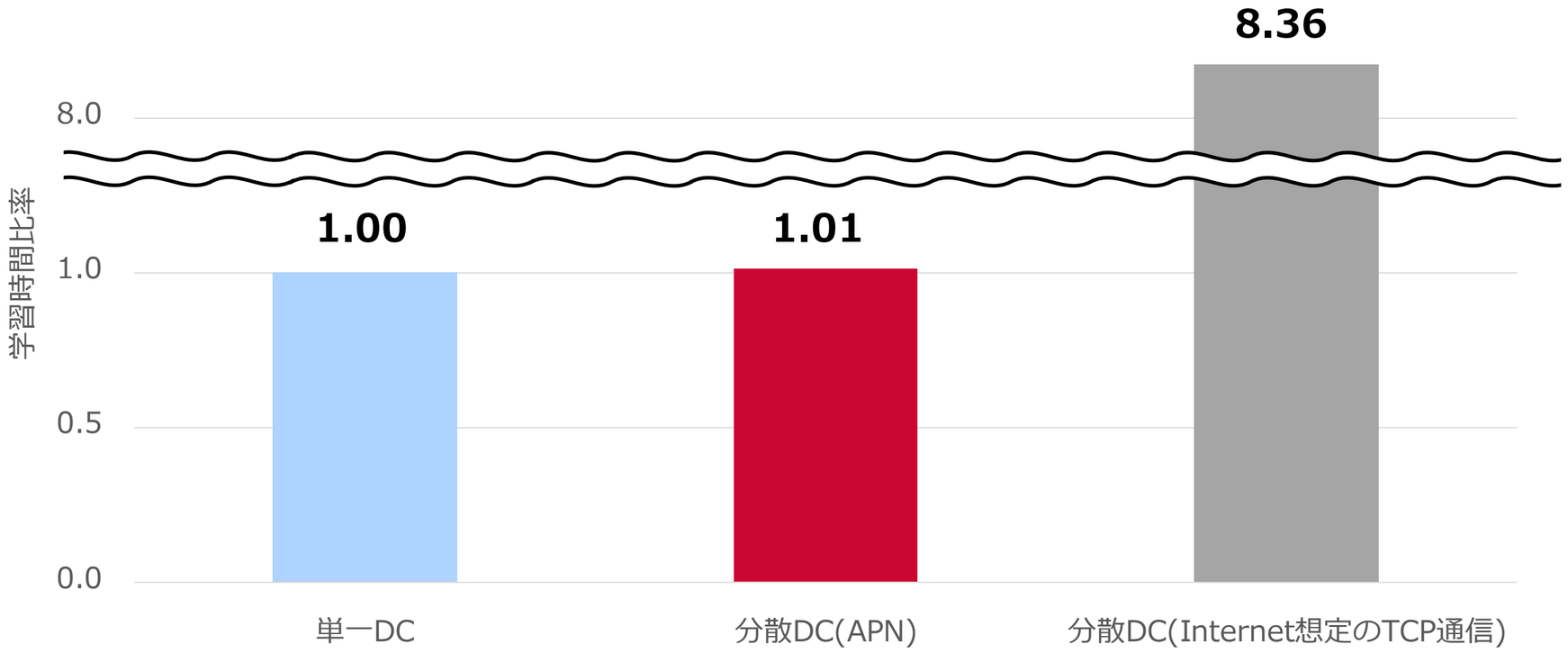
つなごう。驚きを。幸せを。
NTT docomo Business

- **LLM(tsuzumi 7B)の事前学習** を、NVIDIA H100 Tensor コア GPU x2基 x3ノードで実施
- 処理完了の所要時間（1ステップあたりの時間 x 所定ステップ数）を計算
- 単一DC、分散DC(APN経由※0)、分散DC(Internet想定のTCP通信※1)、の三者の性能を比較



単一DCでの学習の所要時間を1とした場合、分散DC(APN) は **約1.01倍** とほぼ互角

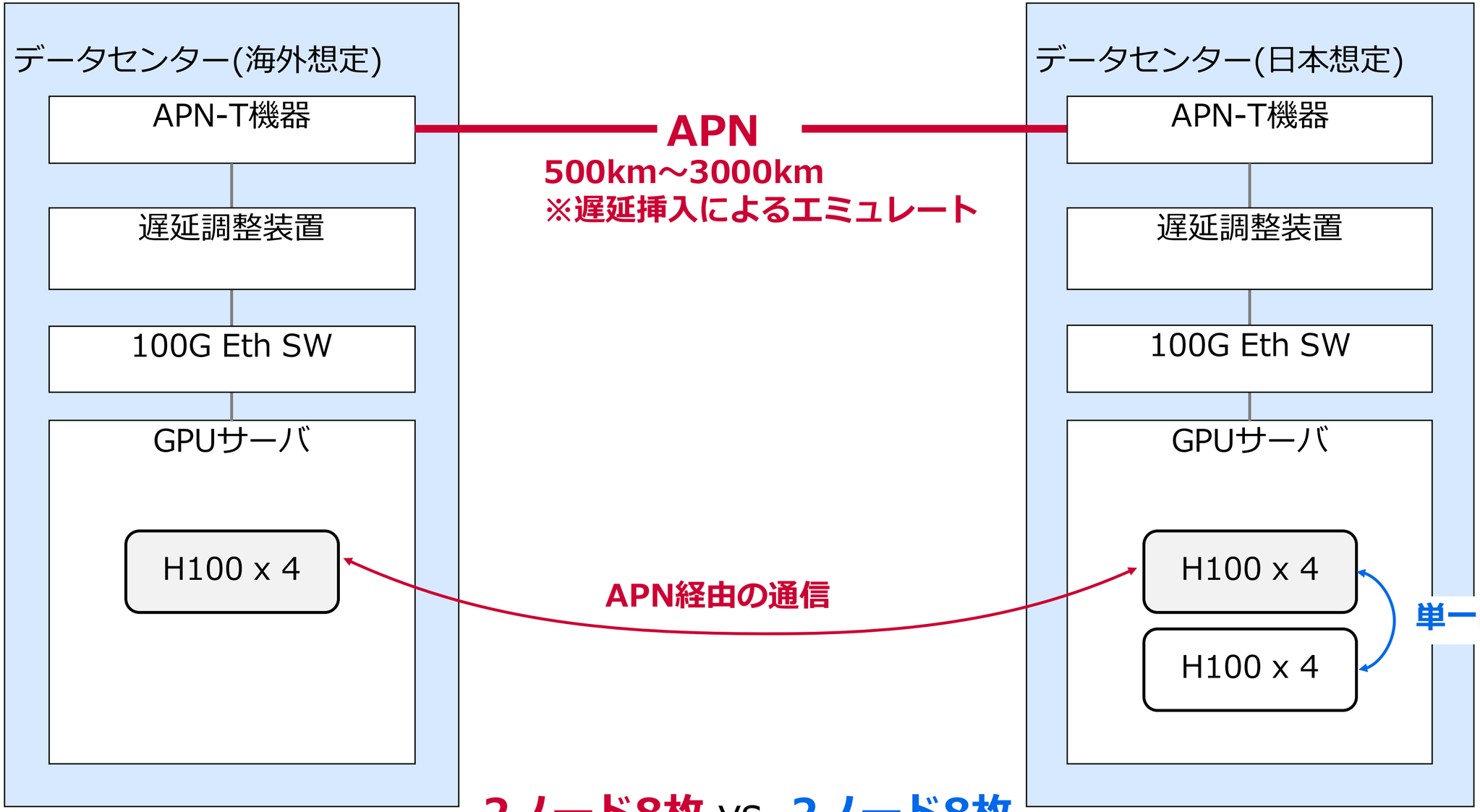
分散DC(Internet想定のTCP通信)では **約8.36倍** かかる



※1
Bandwidth=300Mbps
(tcコマンドで制限)
Ping RTT=15ms
(tcコマンドで各拠点間に
遅延挿入)

2: 遠距離で分散学習 実証環境

つなごう。驚きを。幸せを。
NTT docomo Business

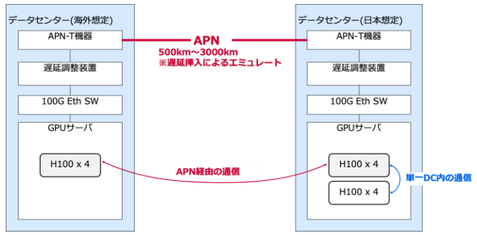


2ノード8枚 vs. **2ノード8枚**

2: 遠距離で分散学習 実証結果

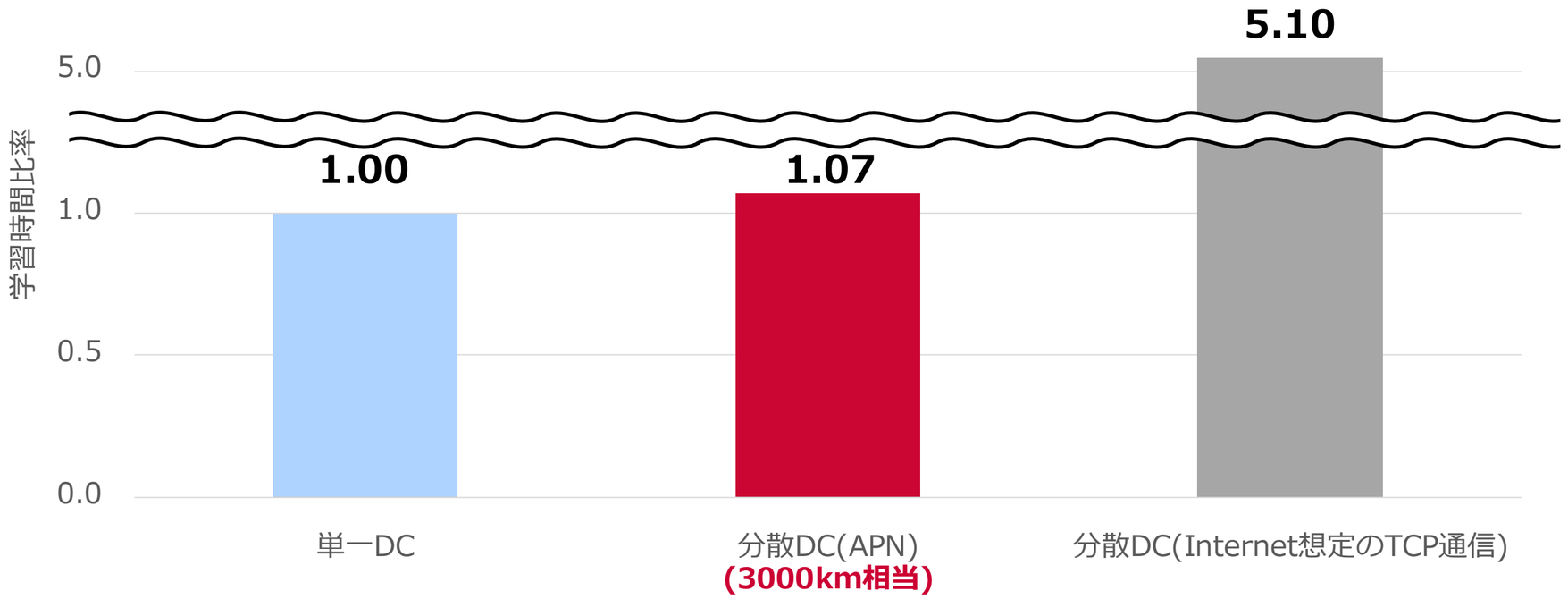
つなごう。驚きを。幸せを。
NTT docomo Business

- **LLM(tsuzumi 7B) の事前学習** を、 NVIDIA H100 Tensor コア GPU x4基 x2ノードで実施
- 処理完了の所要時間（1ステップあたりの時間 x 所定ステップ数）を計算
- 単一DC、分散DC(APN経由※0)、分散DC(Internet想定のTCP通信※1)、の三者の性能を比較



単一DCでの学習の所要時間を1とした場合、分散DC(APN) は **約1.07倍** とほぼ互角

分散DC(Internet想定のTCP通信)では **約5.10倍** かかる



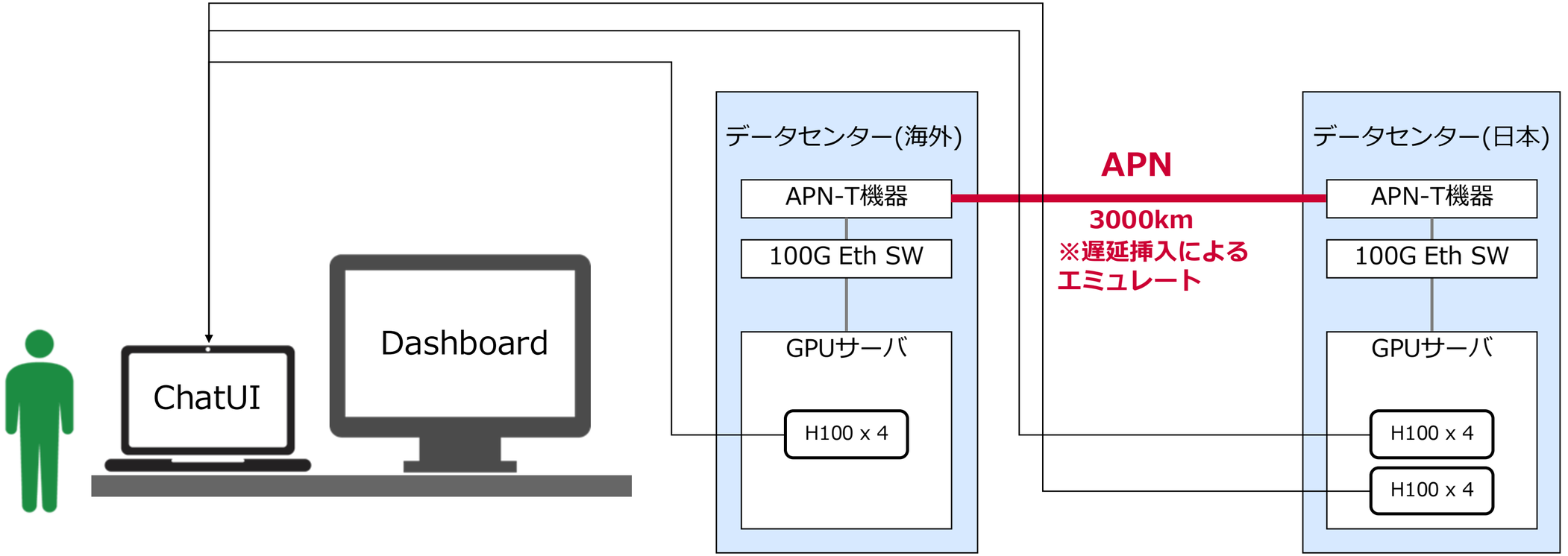
※0
Bandwidth=100Gbps
Ping RTT=30.6ms
(遅延調整装置で両端15ms挿入,
1km=5μs換算)

※1
Bandwidth=300Mbps
(tcコマンドで制限)
Ping RTT=50.6ms
(tcコマンドで両端25ms挿入)

3: 遠距離で分散推論 実証環境

つなごう。驚きを。幸せを。
docomo Business

ChatUIと可視化ダッシュボードをデモ用に実装



前節「2: 遠距離で分散学習」の実証環境(3000km模擬)を流用

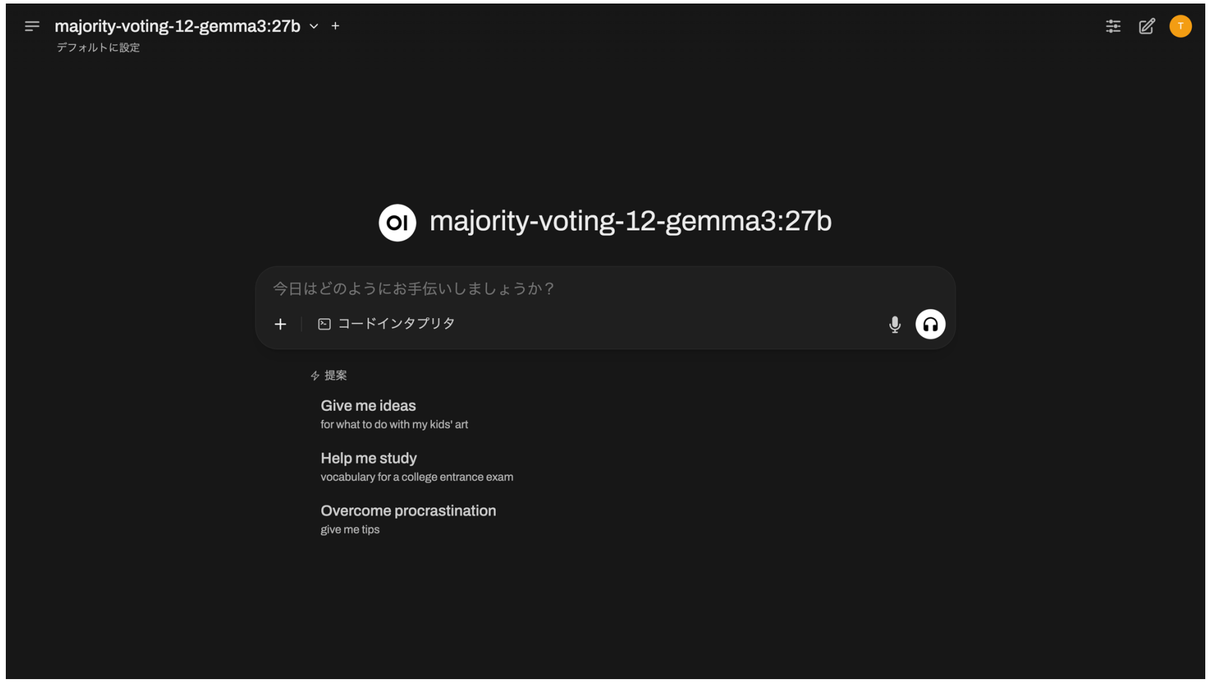
推論精度比較： **3ノード12枚** vs. **1ノード1枚**
スループット比較： **2ノード8枚** vs. **2ノード8枚**

3: 遠距離で分散推論 実証環境

つなごう。驚きを。幸せを。
NTT docomo Business

ChatUI

OpenAI APIを並列で動かす
python scriptを実装(FastAPI利用)



推論させたい問題を投入し回答を得る

Dashboard

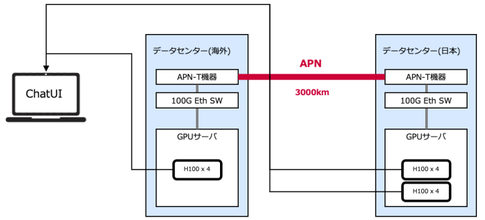
Prometheus + Grafanaで構築
メトリクスを収集・可視化

2拠点のGPU利用率の変化を観測

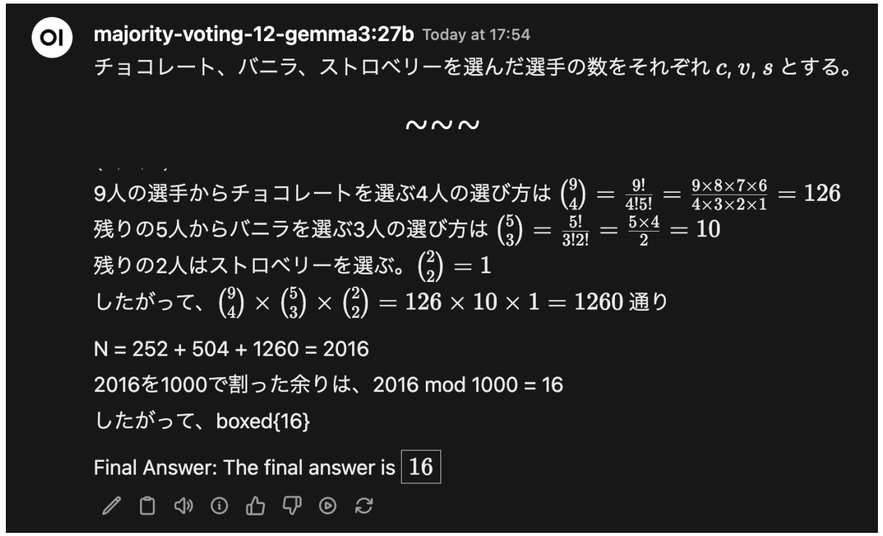
3: 遠距離で分散推論 実証結果1/2

つなごう。驚きを。幸せを。
NTT docomo Business

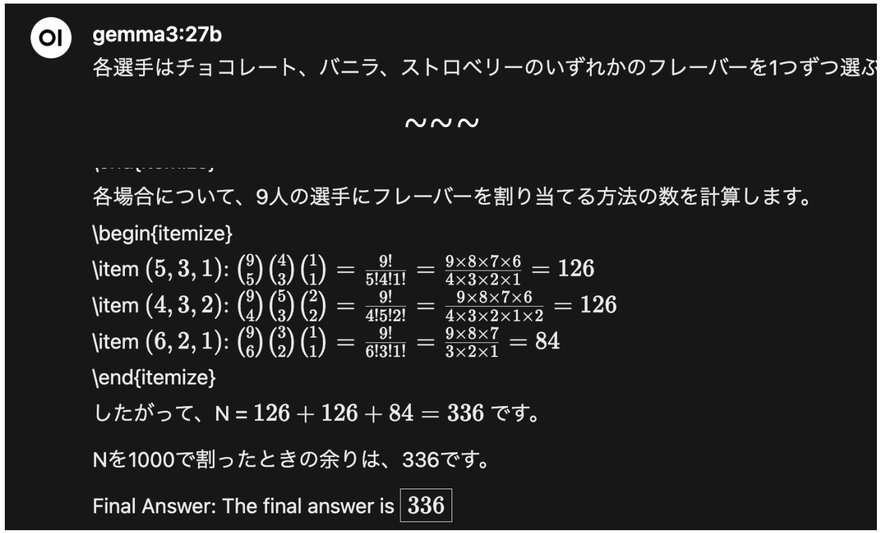
- 遠距離(約3000km)分散環境を利用し、NVIDIA H100 Tensor コアGPU x4基 x3ノードで実施
- アメリカ数学選抜試験(AIME)の問題※1を入力として推論性能を比較
- LLMモデルを**複数並列(12モデル)**で実行し、多数決をとり最頻出の回答を採用※3,4
  - 多様なアプローチを網羅し、その上で最も「一貫性のある」答えを選ぶ
- 単一推論、遠距離(約3000km)分散推論、の2者を比較



回答性能を比較した結果、遠距離(約3000km)分散推論は**正答**
単一推論は確率的に誤答



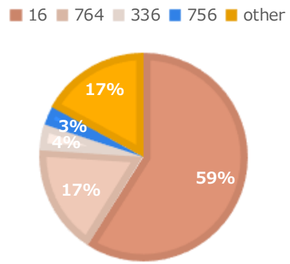
遠距離(約3000km)分散推論
正答



単一推論
確率的に誤答※2

※1
2025 AIME I
Problem 3
正答は「16」

※2
例：100回試行の出力内訳

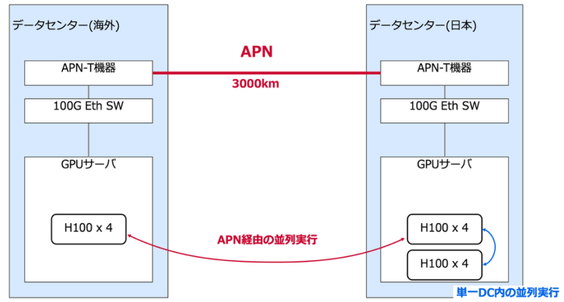


※3 Snell, Charlie, et al. "Scaling llm test-time compute optimally can be more effective than scaling model parameters." arXiv preprint arXiv:2408.03314 (2024).
※4 Zeng, Zhiyuan, et al. "Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities?." *arXiv preprint arXiv:2502.12215* (2025).

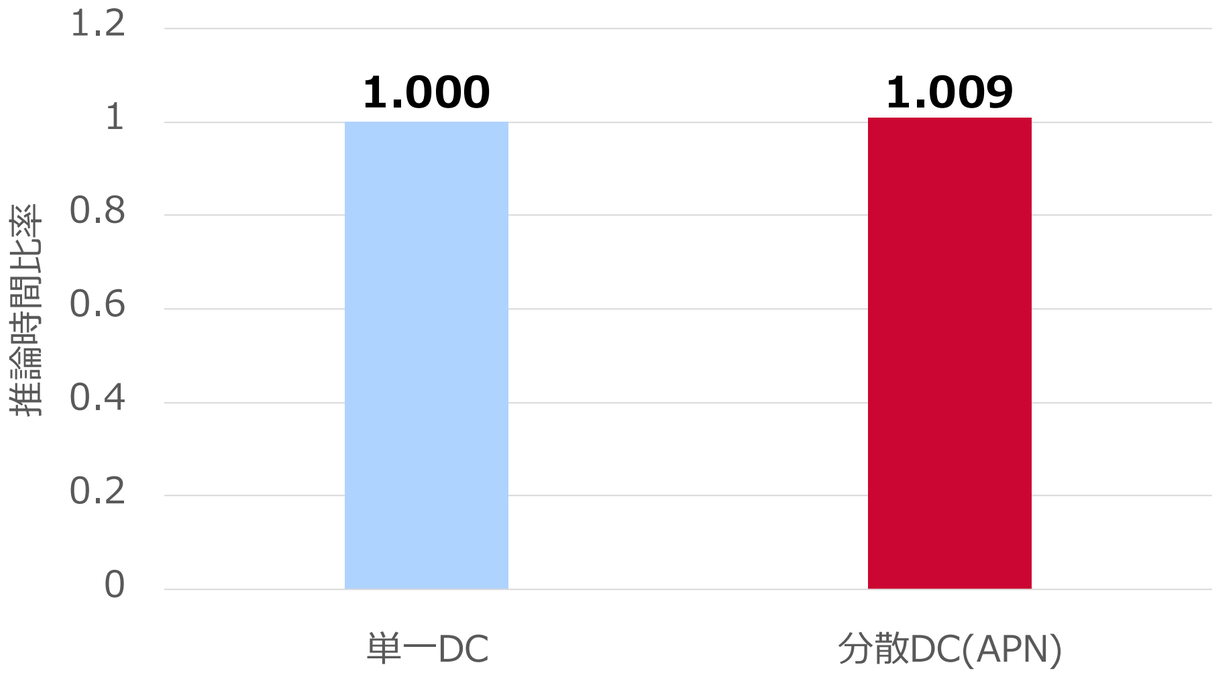
3: 遠距離で分散推論 実証結果2/2

つなごう。驚きを。幸せを。
NTT docomo Business

- 遠距離(約3000km)分散環境を利用し、NVIDIA H100 Tensor コアGPU x4基 x2ノードで実施
- アメリカ数学選抜試験(AIME)の問題を入力として推論時間(e2e_latency)を比較
- LLMモデルを**複数並列(8モデル)**で実行し、多数決をとり最頻出の回答を採用※1,2
  - 多様なアプローチを網羅し、その上で最も「一貫性のある」答えを選ぶ
- 単一DC、分散DC(APN)経由、の2者を比較

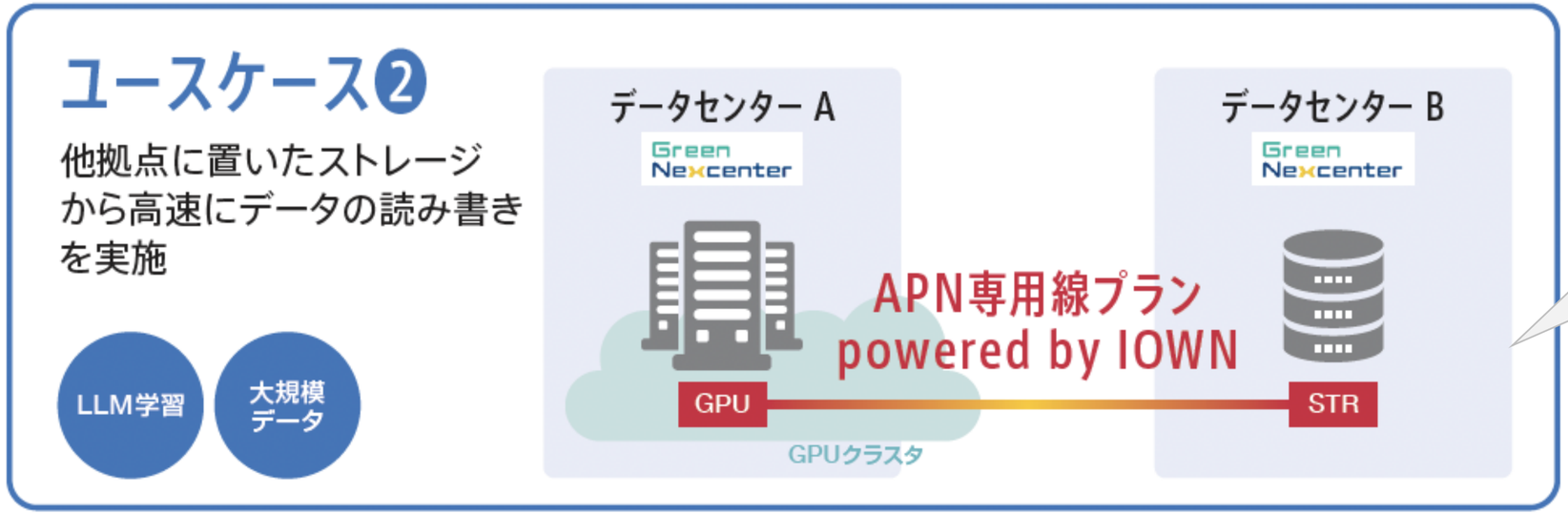


単一DCでの推論時間を1とした場合、分散DC(APN) は **約1.009倍** とほぼ互角



※1 Snell, Charlie, et al. "Scaling llm test-time compute optimally can be more effective than scaling model parameters." arXiv preprint arXiv:2408.03314 (2024).
※2 Zeng, Zhiyuan, et al. "Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities?." *arXiv preprint arXiv:2502.12215* (2025).

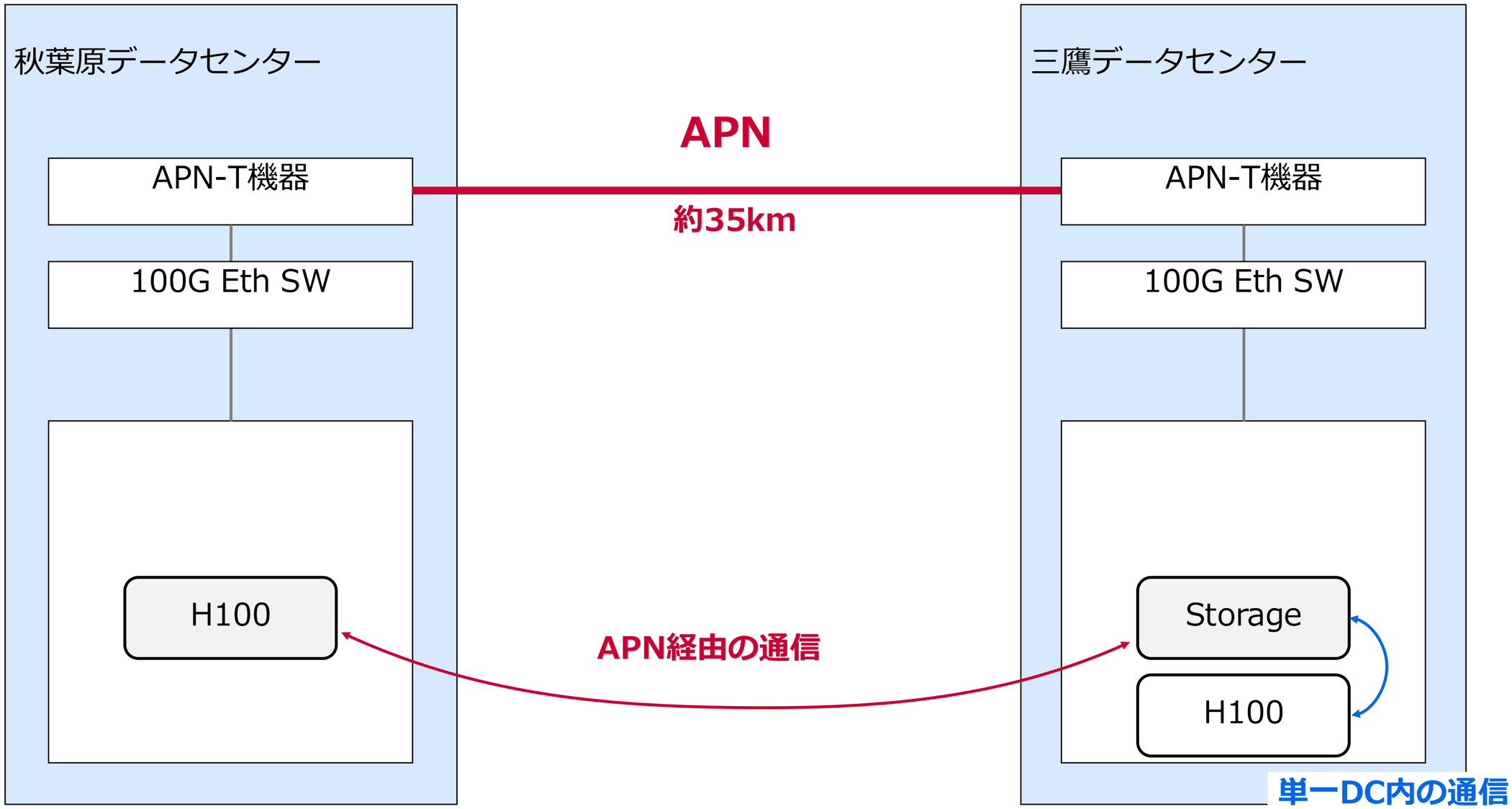
ユースケース②



遠隔ストレージ

【4: ベース実験】秋葉原・三鷹の **2拠点間** におけるストレージアクセス性能を測定

4: 遠隔ストレージ性能 実証環境



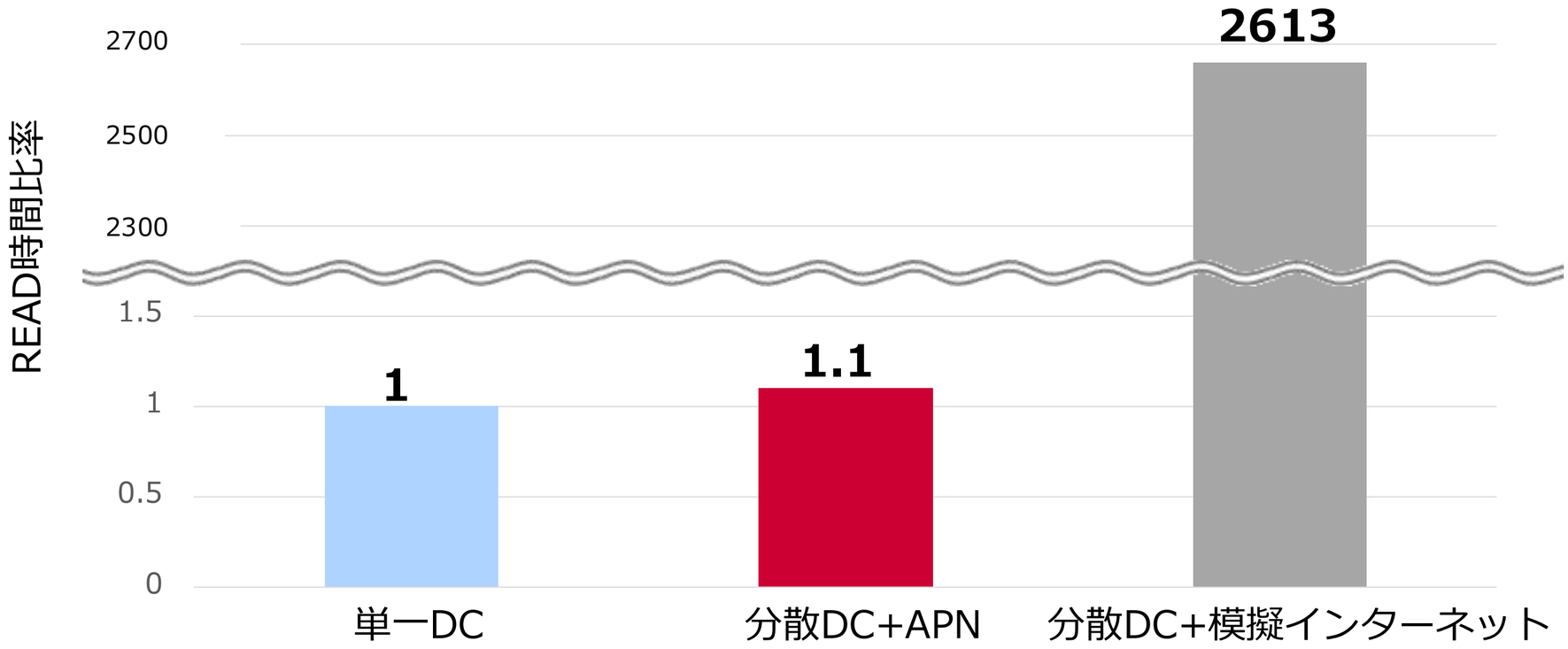
4: 遠隔ストレージ性能 実証結果1/2

つなごう。驚きを。幸せを。
NTT docomo Business

- GPUサーバからストレージサーバへのNFSアクセスの性能を
  GPU Direct Storage ベンチマークツールgdsio により測定
- 単一DC、分散DC+APN経由、分散DC+模擬インターネット回線※1、の三者の性能を比較

単一DCでのデータ読み出し所要時間を1とした場合、分散DC+APNは **約1.1倍** とほぼ互角※2

分散DC+模擬インターネットでは **約2613倍** かかる



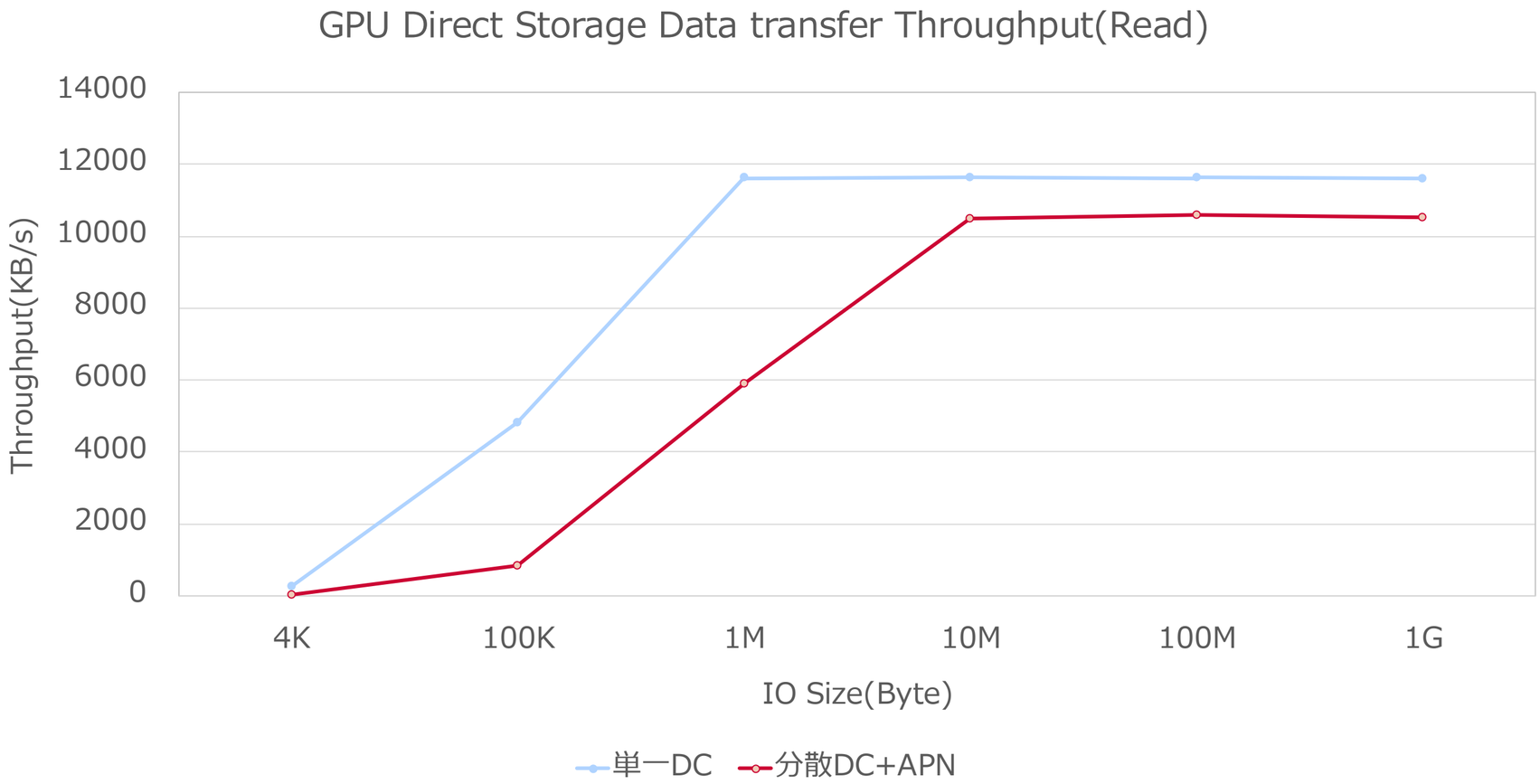
※1
模擬インターネット回線：
Bandwidth=300Mbps
(SWにてシェーピングして制限)
Ping RTT=0.526ms

※2
測定条件
- ✓ スレッド数: 8
- ✓ データサイズ：4KB ~ 1GB
- ✓ スループット：IOサイズを変化させた場合の最良値を採用

4: 遠隔ストレージ性能 実証結果2/2

つなごう。驚きを。幸せを。
docomo Business

- GPUサーバからストレージサーバへのNFS over RDMAを使ったアクセス性能を調べるため
  ベンチマークツールgdsioで様々な条件下でのREADスループットを測定
- **IOサイズ※を大きくしていくと分散DCが単一DCに接近するスループットを発揮**
  ※1回のI/Oリクエストで転送するデータ量、ブロックサイズ。

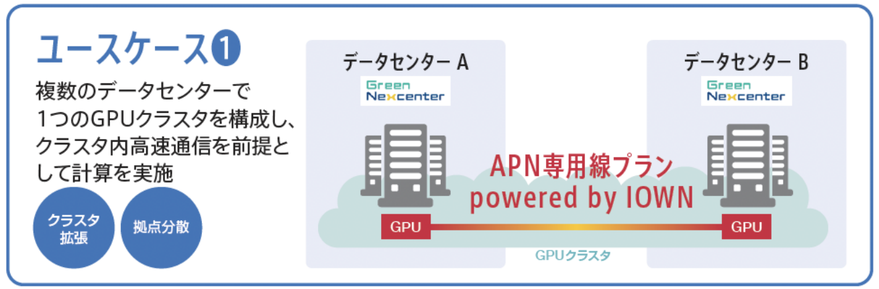


※測定条件
- ✓ スレッド数: 8
- ✓ データサイズ：4KB ~ 1GB
- ✓ 総データ量：2GB

実証結果まとめ

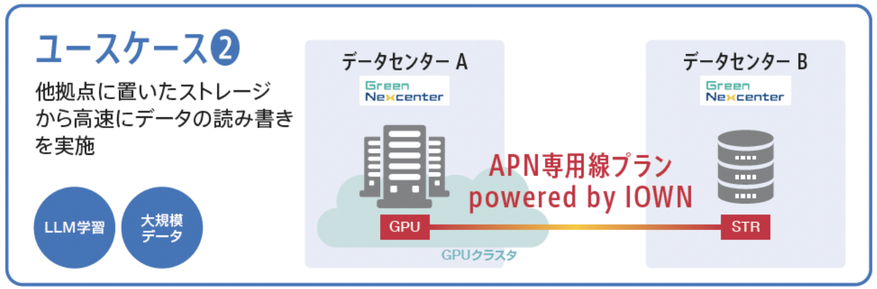
つなごう。驚きを。幸せを。
docomo Business

- 2つのユースケースのどちらでも、**分散DC+APNの実用性や、APNの優位性を確認**
- IOWN APNの高速大容量・低遅延の特徴を活かして、GPUサーバ間のデータ転送が迅速かつ効率的に行われるため
- AI向けGPUインフラの最適分散配置により、より柔軟かつ効率的なリソース利用の可能性が広がる



AI学習性能：小規模LLM(tsuzumi 7B)の事前学習
近距離3拠点で所要時間 **1.01倍**
超遠距離2拠点で所要時間 **1.07倍**

AI推論性能：LLM(gemma3 27B)を用いた並列推論
超遠距離2拠点で所要時間 **1.009倍**
リソース増による推論精度向上も確認



ストレージアクセス性能：GPU Direct Storageベンチマーク
近距離2拠点間でREAD所要時間 **1.1倍**

NTTドコモビジネスのAPN専用線/データセンタ

つなごう。驚きを。幸せを。
docomo Business

- ✓ NTTドコモビジネスは全国において約70か所のデータセンターを運用
- ✓ NTTドコモビジネスのAPN専用線は、全国で1,300回線（起点・終点別）以上の実績あり

関東圏の主な
データセンター

Green
Nexcenter
大阪第7データセンター

STNetとNTT docomo Business
協業のデータセンター

Nexcenter
高松第2データセンター

横浜第１データセンター
Green
Nexcenter

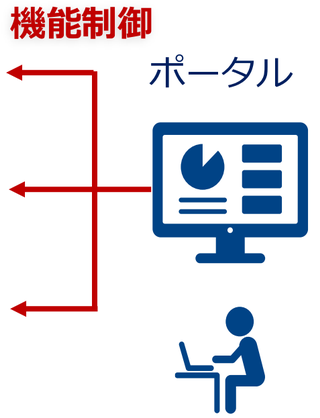
京阪奈データセンター(仮称)
Green
Nexcenter

IOWN APN

NTTドコモビジネスのGPUaaS

つなごう。驚きを。幸せを。
NTT docomo Business

- 専有型のIaaS提供に加え、オンデマンド型のGPUaaSを提供予定
- SDPFクラウドにおける**ベアメタルサーバーの提供**から始め、**仮想サーバー/コンテナへと対応領域**を拡大
- AI-Centric ICTプラットフォームの一部として展開し、ポータルからSD化されたNWとセットで統一的な制御を実現



NW含めて
統一的な制御が可能

ネットワーク(NaaS=NW as a Service×NW as a Sensor)
(docomo business RINK/新IoTサービス/IOWN APN等)

今後の展望

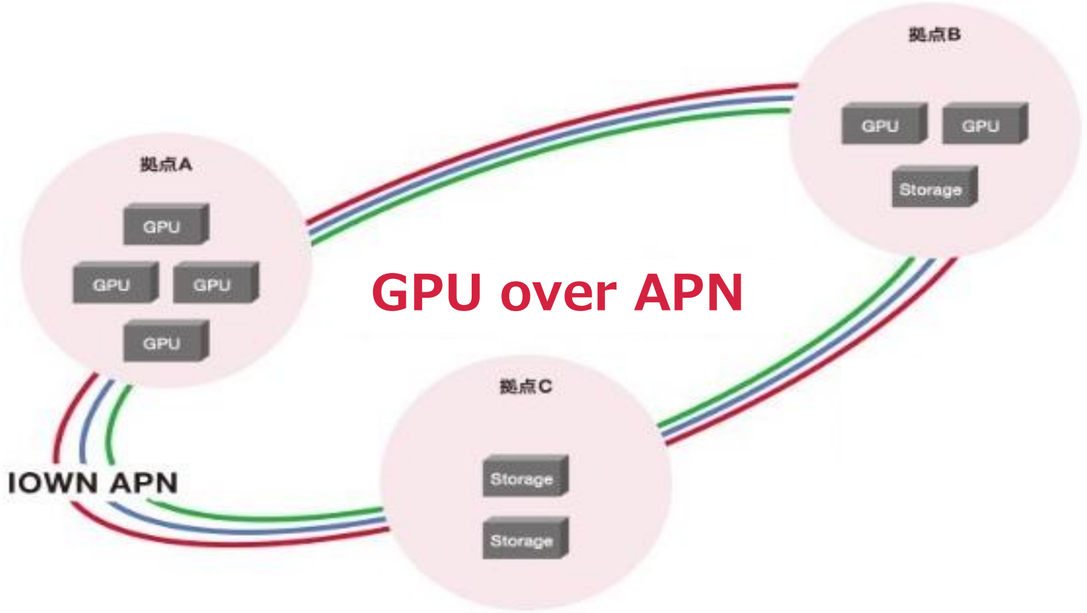
つなごう。驚きを。幸せを。
NTT docomo Business

- 本実証に興味、課題感を持っているパートナーを募集し、共創戦略のもとで共にPoCを実施

- 「GPU over APN」の実証環境を順次拡充し、更に可能性を広げる
  - より大規模なクラスタへ適用した際の課題の特定や解決
  - GPU利用者や提供者のニーズのより具体的な汲み取り
- 実証で得られたノウハウを商用サービスへ反映し、複合的ソリューションとして事業化をめざす

NTT docomo Business GPU over APN 実証環境

AI/ML	データ分析	ベンチマーク	持込みアプリ
ジョブスケジューラ（slurm, k8s, など）			
GPUサーバ (H100, A100, P100)	高速フラッシュストレージ		持込みHW
データセンタ	APN回線		



つなごう。驚きを。幸せを。

NTT docomo Business