つなごう。驚きを。幸せを。
docomo Business

第８５回ニッポンクラウドワーキンググループ会合（大阪開催）

2026年10月30日（金）

AIネイティブインフラ「AIOWN」

～AI時代を支える次世代インフラへの転換～

林　雅之

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
エバンジェリスト

AIOWN
AI×IOWN

「AIOWN」は、NTT株式会社で商用登録中です
「IOWN®」は、NTT株式会社の商標又は登録商標です

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

自己紹介



林 雅之

ブログを毎日更新

7,000日超／19年

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、技術戦略部門（兼務）
北海道次世代産業振興・まちづくりPT（兼務）

国際大学GLOCOM　客員研究員
埼玉工業大学　工学部情報システム学科　非常勤講師
岩手大学　地域協創教育センター　招へい講師

ITmediaオルタナティブ・ブログ『ビジネス2.0』の視点
NewsPicks　トピックスオーナー『デジタル政策と未来社会』
Voicy『ビジネス3.0』の視点　パーソナリティ
X：@masayukihayashi／@cloud_1topi

クラウドの基本 第2版

オープンデータ 超入門

スマートマシン 機械が考える時代

ほか

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

目次

01 **AIの計算需要は推論へ移り投資が拡大**
推論が主役になりAI工場とDC市場が広がる流れを示す

02 **AIインフラの制約は電力と熱に集中**
高密度化・液冷・直流給電と電力需要の逼迫を説明する

03 **IOWNは光で計算基盤の制約を解く**
APNと光電融合の仕組み、国際的な広がりを紹介する

04 **ワット・ビット連携は電源近くに分散**
政策の後押しと電源立地DCをAPNで結ぶ実証を示す

05 **AIOWNは全国の資源を一つの基盤にする**
AIOWNの価値と提供基盤、分散GPUの事例を紹介する

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

CHAPTER

01

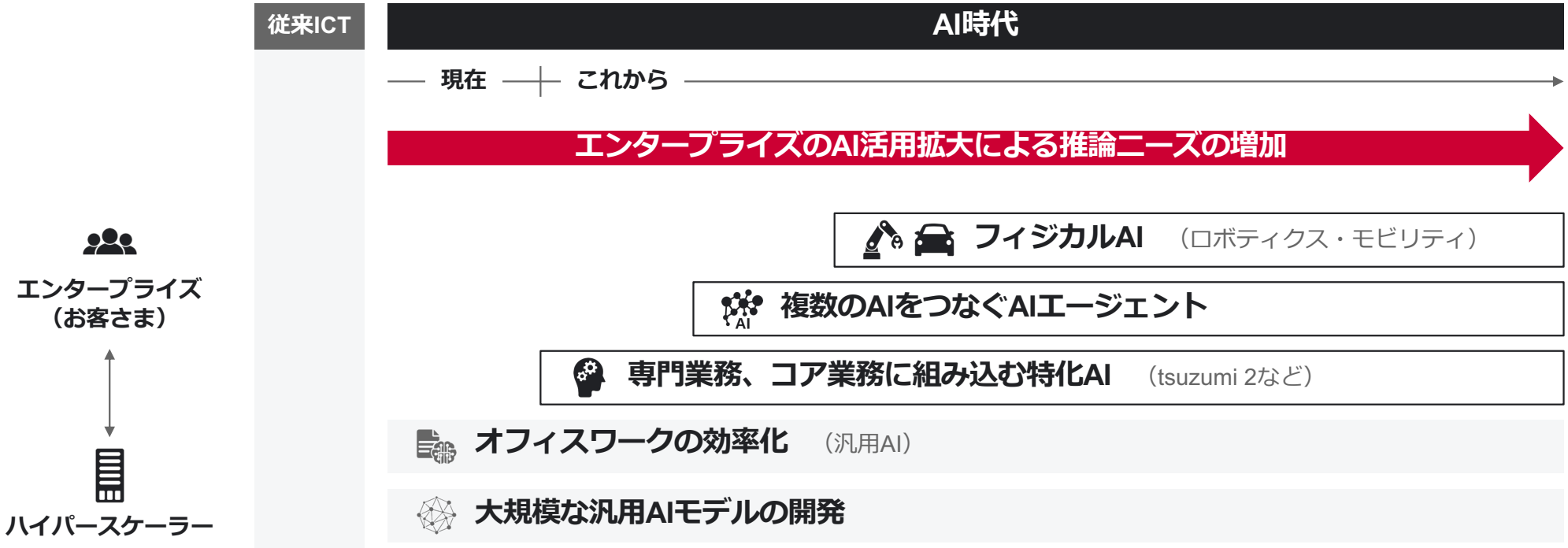
AIの計算需要は推論へ移り投資が拡大

推論が主役になりAI工場とDC市場が広がる流れを示す

AI活用の進展と領域拡大

オフィスワークの効率化から、コア業務・専門業務に組み込む特化AIとAIエージェントへ拡大

車やロボットと連携するフィジカルAIの登場で、GPUの用途は学習から推論中心へ



出所：NTT（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AIの進化

AI技術の進展により、業務代替・高度化の対象はホワイトカラー領域からエッセンシャルワークへと拡大。
これに伴い、従来はIT活用が限定的だった領域においても、新たなビジネス機会が広がる。



出所：OpenAI社、NVIDIA社等の発表を参考に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AI活用は汎用業務からコア業務・フィジカルAIへ

フィジカルAIが推論需要を押し上げ、2030年にはDC需要の4割超へ拡大する見通し

これまで		これから
活用領域｜汎用的な業務効率化	❯	**企業のコア業務・専門業務**
連携対象｜画面の中	❯	**車やロボットなどフィジカルAI** モノとの連携で推論需要が増加
ワークロード｜学習中心	❯	**推論中心** **2030年**にDC需要の**4割**超 （McKinsey推計）
投資の主体｜テック企業中心	❯	**事業会社へ拡大** 転換点は**2026年**（Gartner）

推論需要の大幅な拡大

フィジカルAIの市場と導入

60兆円
多用途ロボットの世界市場
（2040年見込み）

1,000万台規模
国内のAIロボット（2040年目標）

80%
フィジカルAIが支援する現場従業員

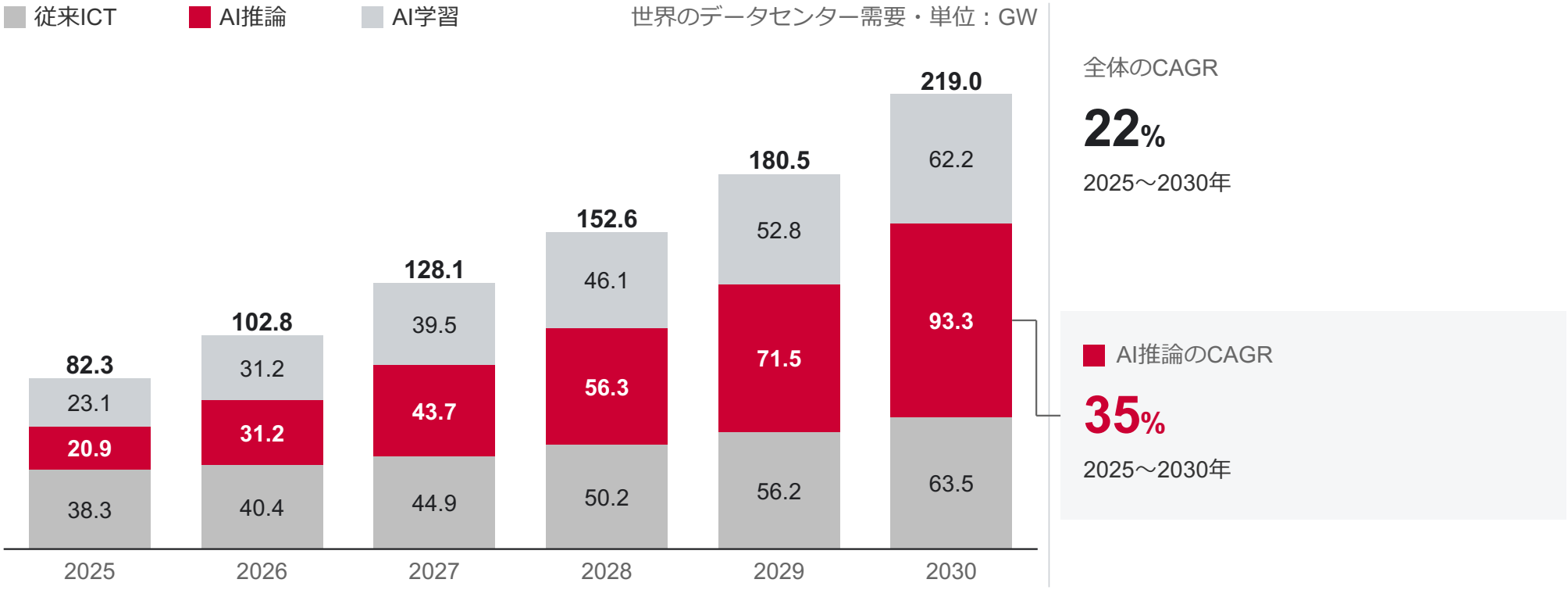
国際企業が対象（2030年、Gartner予測）

出所：NTT報道発表（2026年4月27日）、AIロボティクス戦略（2026年3月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AIワークロードは学習から推論へ

AI推論のワークロードは2025年から2030年で4倍以上、2030年には全体の4割超を占める



※AI Compute＝AI学習＋AI推論

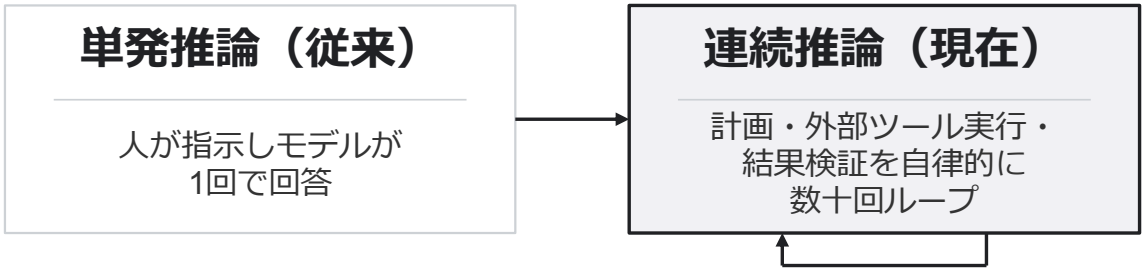
出所：McKinsey Data Center Demand Modelを基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

エージェンティックAIと連続推論シフトの現実

単発応答から数十回ループする自律型エージェントへ移行し、
世界のAI支出は2029年に1.3兆ドルに達する見通し

しくみ



業務アプリ搭載：タスク特化型AIエージェントを
載せる企業アプリは、
2025年の**5%**未満から2026年に**40%**へ（予測）

トークン処理量：2026年5月にGoogleは
月間**3,200兆**トークン超を処理し、1年で**約7倍**に増加

※トークン処理量はGoogle（2026年5月）による

市場へのインパクト

1.3兆ドル	2029年の世界AI IT支出（予測）
31.9%	AI支出の年平均成長率（2025～2029年）
10～100倍	連続推論化による1タスクあたりの推論計算需要

出所：IDC（2025年8月）、Gartner（2025年8月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

ワークロードの分岐：学習需要と推論需要の違い

推論の計算能力の伸びは2026年に122%増と、学習の56%増を大きく上回る見通しで、拠点は都市近郊やエッジへ広がる

	学習（Training）	推論（Inference）
規模・コスト構造	数千～数万GPU、初期設備投資（CapEx）が中心	**数ラック～数十ラック単位、運用費（OpEx）の大半が電力**
通信・立地	超広帯域の密結合（NVLink等）と電源立地のGW級キャンパス	**低遅延通信（P99.9テールレイテンシ）と都市近郊・エッジへの分散**
計算能力の伸び	前年比56%増（2026年、北米CSP5社予測）	**前年比122%増（学習の2倍超のペースで拡大）**
2030年の計算需要	62.2GW（2025年23.1GWから拡大）	**93.3GW（2025年20.9GWから4.5倍へ拡大）**
	発電所直結・GW級用地・液冷の大型案件	**都市型データセンター・コンテナ型エッジ拠点・推論専用ASICと光広域連携**

出所：TrendForce（2026年5月）、McKinsey（2026年2月）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AIファクトリーの台頭：電力をトークンに変える工場

従来の汎用ICTデータセンターとは設計思想が異なり、電力を投入して知能を量産するAI工場へ変わった

	従来の汎用データセンター（情報保管・業務処理）		AIファクトリー（知能・トークンの量産工場）
目的	Web・基幹システムの安定稼働とデータの保管	›	燃料・電力（Molecules）を投入し、 AIトークン（Tokens）を**24**時間量産。 目標は電力あたりのAI出力の最大化
サーバー構成	CPU中心、**1**台単位の独立サーバーと Ethernet接続	›	数万基のGPU/ASICをNVLink・ 光配線で結ぶ単一巨大コンピュータ
電力・冷却	**1**ラック**6**～**15kW**、空冷 （PUE **1.5**～**1.8**）、**480**V/**415**V交流給電	›	**1**ラック**100～250kW超** （将来**1MW**級）、直接液冷が必須 （PUE **1.1**～**1.2**）、**800**V直流給電へ
立地と規模	都心近郊の好アクセス立地、 受電容量**2**～**50MW**規模	›	発電所直結・GW級キャンパス （**100**～**1,000MW**）、IT機器が 建設費の**約60%**を占める（目安）

※ラック電力とPUEはVertiv等、IT機器の比率はPwC（2026年9月）による

出所：NVIDIA（2026年3月・8月・10月）、JERA・Dell・RHAELM（2026年10月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

推論シフトが変える基盤要件

推論の常時化で制約は遅延・電力・配置へ広がり、これらへの対応が設計条件

遅延	電力	配置
リアルタイム推論の常時化	ラック電力の上昇	拠点数の増加

常時化で起きる変化

利用者の近くでの応答	高密度GPUラックの導入	エッジまで広がる運用範囲
応答時間の制約を受ける処理	限られた電力での冷却	データを置く場所の選択

基盤に求められる要件

低遅延の拠点間ネットワーク	**チップ直結の液冷への移行**	**分散基盤の設計と処理配分**
利用者の近くへの分散配置	**受電容量の先行確保**	**データ主権の確保**

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

CHAPTER

02

AIインフラの制約は電力と熱に集中

高密度化・液冷・直流給電と電力需要の逼迫を説明する

AIラックの高密度化が冷却・配電設計を刷新

100kW超のラックでは空冷単独の限界が近づき、電力と熱と光と計算の協調制御が要件

	従来型データセンター	AIデータセンター
ラック密度	数kWから十数kW	**100kW超が閾値**
設計の中心	空調と気流の設計	液冷 / 冷却液分配装置 / 熱再利用 / 800V直流
評価指標	PUE が中心	水使用 / 漏えい / ラック重量 / 床荷重 / 作業安全 **まで設計対象**
更新周期	設備更新は 建屋の寿命に合わせる	**計算設備の更新周期と建屋の寿命が合わない**
	空冷を前提とした設計	**電力と熱と光と計算の協調制御**

出所：Schneider Electric（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

NVIDIAラック電力ロードマップ：800V直流で1MW級へ

2026年のVera Rubinまでは従来ラックの延長線上にとどまり、
2027年のKyber世代で800V直流・約600kWへ、2028年のFeynman世代で1MW級に達する

	GB300 2025年・出荷中	Vera Rubin 2026年8月出荷開始	Rubin Ultra 2027年予定	Feynman 2028年予定
ラック基盤	**Oberon**　交流配電→ラック内で54V直流へ変換	**Oberon／OCP MGX 1.2**　5,000Aの液冷バスバー	**Kyber**　建屋から800V直流配電へ全面転換	**Kyber次世代**　800V直流配電（1,500V直流もOCPで検討）
ラック電力	**最大142kW**　参照設計の上限（GB200は132kW）	**最大227kW**　GB300比1.6倍（MGX 1.2の仕様上限は250kW）	**約600kW**　GB300比約4.2倍	**1MW級**　想定（実値は非公表）
NVLink規模	**NVL72**　銅線（DAC）によるラック内の密結合	**NVL72**　NVLink 6：GPUあたり3.6TB/s	**NVL144**　将来はNVL576まで高密度化	**最大NVL1152**　光電融合（CPO）でラック間まで拡張
冷却・設置	**直接液冷＋一部空冷**	**全液冷**　ファンの無い計算トレイ・入口水温45℃	**液冷＋サイドカー**　電源・冷却で2ラック分の床を使う	**非公表**　OCPは二相式のチップ直接液冷を残る課題に挙げる

出所：NVIDIA（2026年8～10月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

冷却方式の世代交代：空冷の限界と液冷比率53%

空冷はリアドア併用でも30kWが限界となり、
AIチップの液冷採用比率は2026年に53%へ過半を超える見込み

AIチップの液冷採用比率：2024年14%→2025年約33%→2026年53%→2027年60%近く

第4世代 浸漬液冷（単相・二相） 最高効率の冷却。ただし運用負荷やPFAS規制（3Mが2025年末製造終了）は導入課題	**～200kW+**	**PUE 1.02～1.1**
第3世代 直接液冷DLC（現在標準） GB200（132kW）、Vera Rubin（最大227kW、MGX 1.2の上限250kW・入口水温45℃）の前提仕様	**～200kW+**	**PUE 1.1～1.2**
第2世代 リアドア空冷・間接液冷 ラック背面に熱交換器を置く過渡期の方式。インロー空調と併用し、30kWが上限	**～30kW/ラック**	**PUE 1.3～1.5**
第1世代 空冷（限界到来） 従来型DCの標準。外気冷房を併用しても15～20kWで排熱が限界に達する	**～15kW/ラック**	**PUE 1.5～1.8**

※電力使用効率（PUE）はDC全体の電力÷IT機器の電力

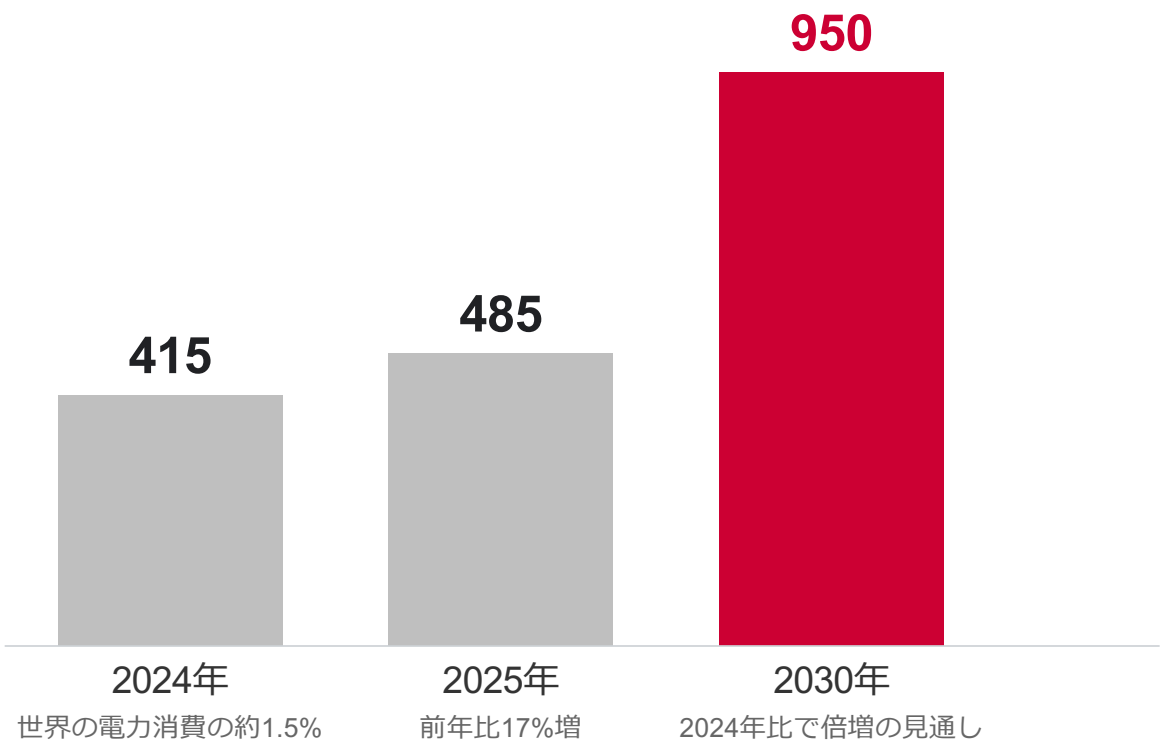
出所：TrendForce（2025年8月・2026年8月）、Vertiv等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

世界のデータセンターの電力消費

2030年に約950TWhへ倍増、世界の電力需要の約3%を占める見通し

世界のデータセンターの電力消費（TWh）



伸びの速さ

年率15%

2024～2030年の見通し。2017年以降は年率約12%で伸びてきた

AI特化型データセンター

3倍

2025～2030年の電力消費。2025年は前年比50%増

世界の電力需要に占める割合

約3%

2030年の見通し。950TWhは日本の総電力消費量をやや上回る規模にあたる

出所：IEA『Key Questions on Energy and AI』（2026年4月）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

CHAPTER

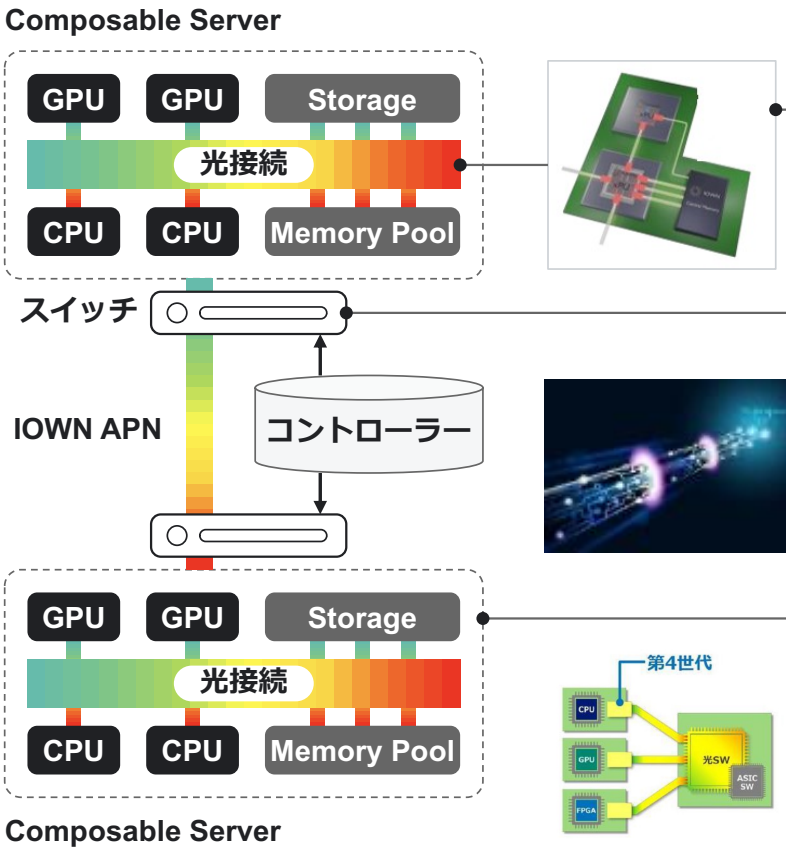
03

IOWNは
光で計算基盤の制約を解く

APNと光電融合の仕組み、国際的な広がりを紹介する

IOWNの主要技術

IOWNはデバイス内から拠点間まで光でつなぎ、別々のデータセンターを1つのように利用



1 光電融合デバイス（PEC）

(Photonics Electronics convergence devices)

- デバイス間・デバイス内の接続を**電気接続から光接続へ**

2 APN (All Photonics Network)

- **光の波長パス**で拠点間を接続し、
  高速、大容量、低遅延で通信
  （経路途中のスイッチでの電気信号⇔光信号変換を無くす）

3 分散型光コンピューティング

- 別々のデータセンターを**1つのデータセンターのように利用**
- 必要な分だけ効率的・柔軟かつ高速（光接続）で
  コンピュータリソースを利用

出所：NTT（2026年）を基に作成

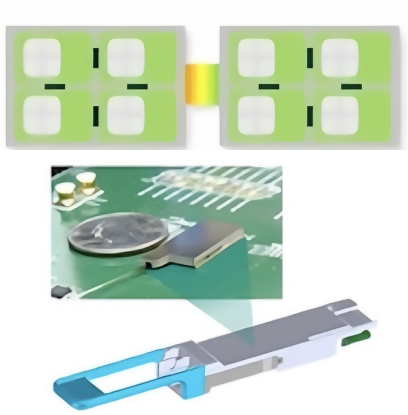
© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

光電融合技術（PEC）のロードマップ

Networking | **Computing**

IOWN 1.0

DC to DC



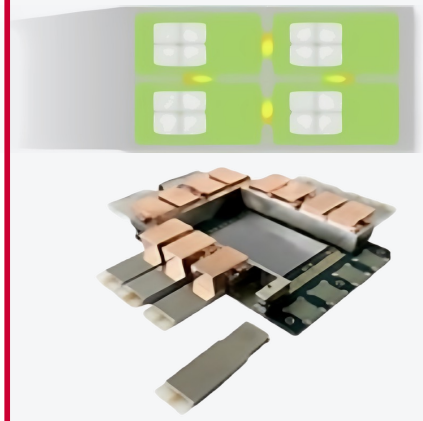
データセンター間接続
（PEC-1）

Scale Across

稼働中

IOWN 2.0

Board to Board



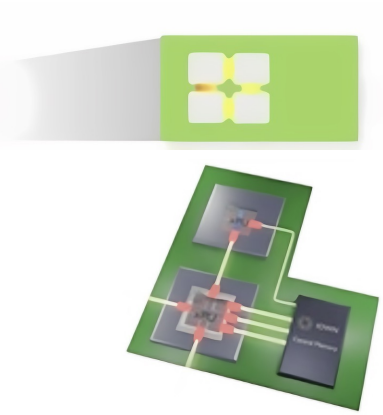
ボード接続
（PEC-2）

Scale Out

2026

IOWN 3.0

Chip to Chip



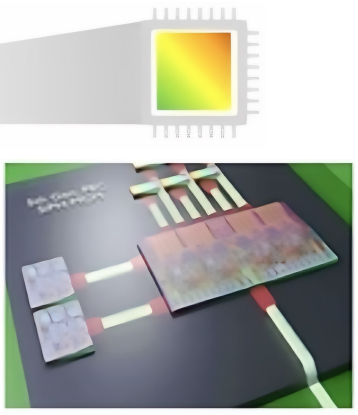
チップ間接続
（PEC-3）

Scale Up

2028

IOWN 4.0

Die to Die



チップ内光化
（PEC-4）

Scale In

2032

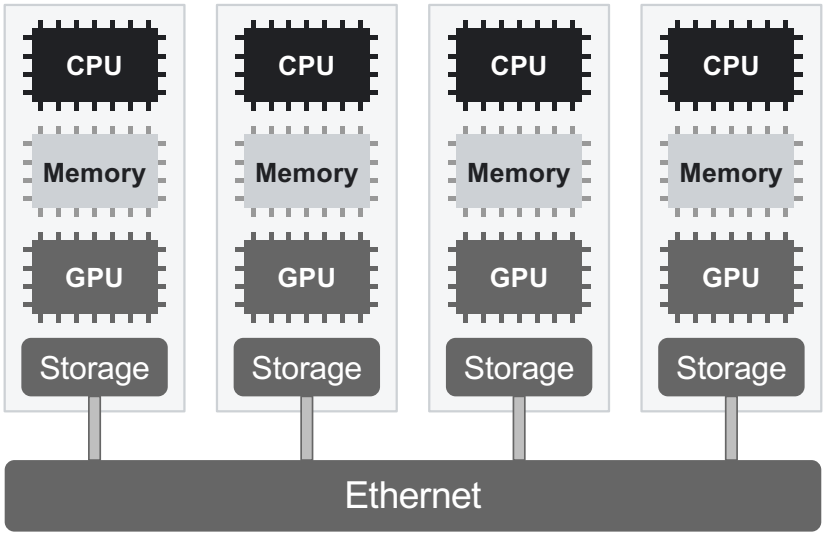
出所：NTT（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

IOWN光コンピューティング構成（将来イメージ）

光電融合デバイスの低電力化 × アーキテクチャの革新 × ソフトウェアの革新

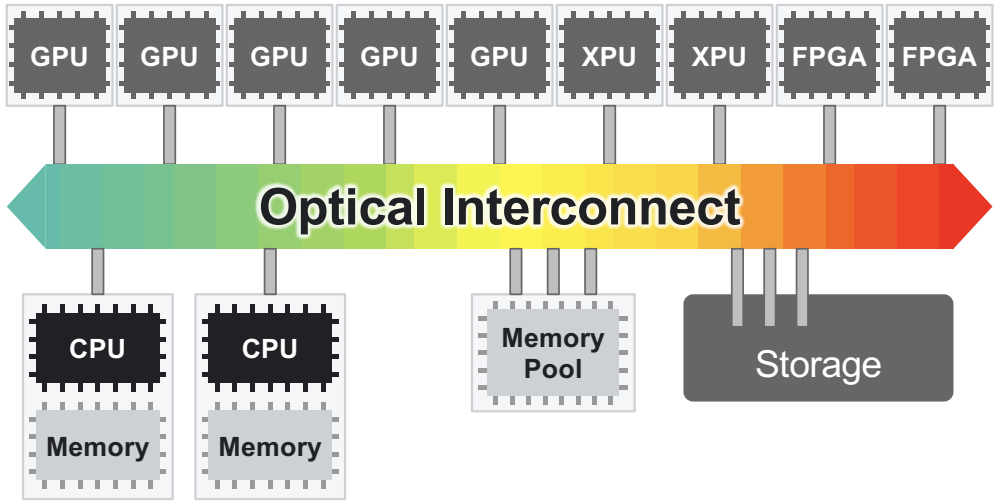
従来のアーキテクチャー



サーバー単位の構成

IOWN光コンピューティング

AIコンピューティング基盤/DCI（Data-Centric Infrastructure ）



リソースを必要なだけ光で接続
箱を超えたラックスケールコンピューティング

出典：NTTイノベーティブデバイス 2025年10月

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

IOWN Global Forum（IGF）の取り組み

2020年設立のIGFが仕様づくりの場、会員は3から175組織・団体へ拡大

2020年1月　米国で設立（NTT・インテル・ソニー）　　IOWN構想の実現に向けた国際団体

ユースケースの選出 → テクニカルロードマップの作成 → 技術仕様書の作成（アーキテクチャ・要件） → 普及・展開（標準化団体との連携）

ユースケースとアプリケーション

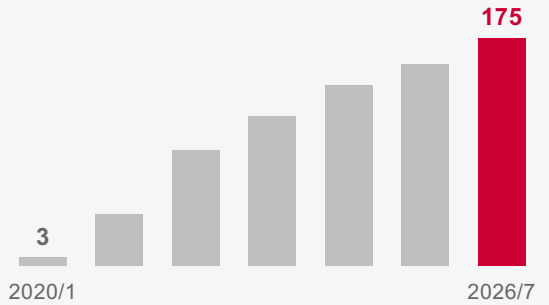
- スマートエネルギー
- スマート都市
- スマートモビリティ
- スマートファイナンス
- スマートエンターテインメント

テクニカルソリューション

- ネットワーク（光・ワイヤレス）
- 分散コンピューティング
- フォトニクス＆オプトエレクトロニクス
- デバイス・インターフェース端末

会員の加入状況（2026年7月時点）

175 組織・団体



※1 会員はアジア・米州・欧州に広がる。内訳はSponsor 35、General 111、Academic or Research 28、ほかAffiliate（2026年7月時点）

出所：IOWN Global Forum、NTTを基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

IOWN AI FundはIOWN関連のスタートアップに投資

日米韓台4か国・地域の運営パートナーと、グローバルなIOWNエコシステムの構築へ

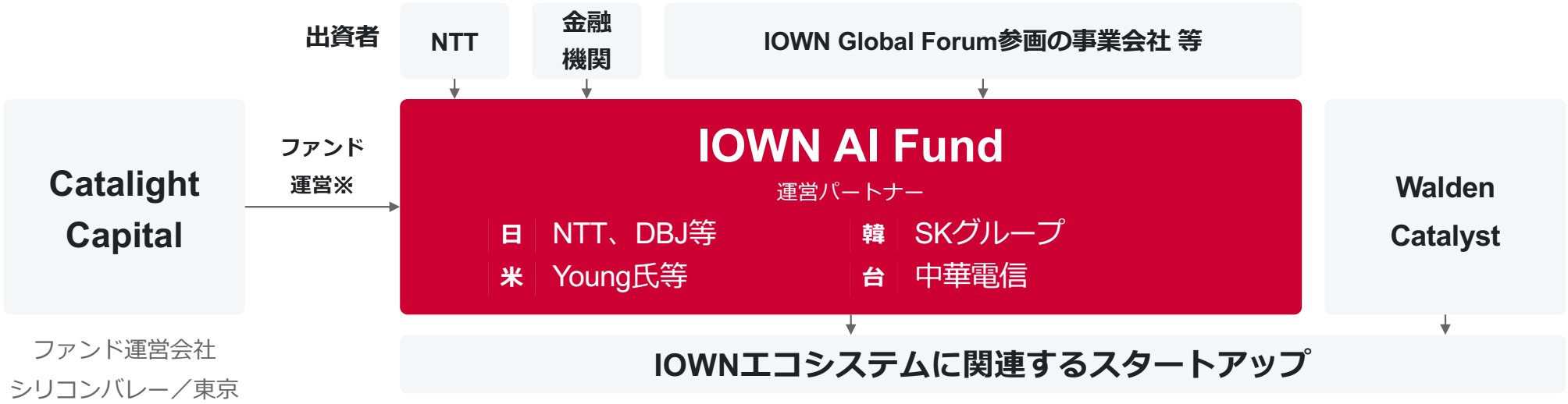
設立の背景

AIホットスポットのシフト
AIインフラの電力・容量制約

→ **光電融合・分散制御**の技術開発が加速し、シリコンバレーを中心に**スタートアップが多数誕生**

→ **AI時代の先端技術を取り込む**

ファンド体制



ファンド運営会社
シリコンバレー／東京

※ファンド運営会社の米国子会社として、Catalight Capital US, LLCを設立し、IOWN AI Fundへ投資助言する。

出所：NTT（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

IOWNの目指すエコシステム

IOWNは多様なパートナーとの連携により、オープンでグローバルなエコシステムを構築・拡大
スタートアップや先端技術を取り込み、共創で価値を最大化

エコシステムの拡大

アプリケーション・サービス		業界特化AI（医療／製造／金融…） ロボティクス／デジタルツイン
ソフトウェア		クラウド／分散システム AIオーケストレーション
AIモデル・推論		推論最適化／ 学習アルゴリズム／分散学習
光源・変調器		AIアクセラレーター／ チップレット／3D積層
AI向け半導体・パッケージング		光デバイス／ 光電融合モジュール
分散型AI基盤の制御		電力／冷却／運用の最適化 エッジ分散基盤・効率化技術
フォトニクス技術		光伝送／コヒーレント通信 光スイッチ／光ネットワーク

出所：NTT（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

CHAPTER

04

ワット・ビット連携は電源近くに分散

政策の後押しと電源立地DCをAPNで結ぶ実証を示す

日本のAI計算基盤政策と公的10兆円支援

日本政府はAI・半導体に2030年度まで公的10兆円以上を投じ、官民50兆円超の投資を誘発する

公的支援10兆円以上で、官民投資50兆円超を誘発

AI・半導体分野への2030年度までの集中支援枠組み

令和9年度のAI関連予算の概算要求は
2兆9,392億円

総務省はAPN研究開発・実装に
500億円、ワット・ビット連携に
264.5億円を要求

骨太方針**2026**では**17**戦略分野で
2040年度までに官民投資**370兆円超**

GPU供給支援（経産省）

経済安保推進法による特定重要物資のクラウド整備の半額助成（さくらに最大**約500億円**、KDDIに**約102億円**等を交付）

＋

DC整備ファンド（政投銀・DBJ）

AIデータセンター建設への協調融資・長期低利融資スキームの提供

＋

地域優遇・投資促進税制

投資額**35億円以上**・ROI **15%以上**で即時償却か税額控除**7%**。
自治体の立地補助金（北海道上限**15億円**等）

＋

通信・インフラ基金（総務省）

地方DC整備・国際海底ケーブル陸揚局の地方分散を支える**720億円**基金と補助金

出所：経済産業省（2025年12月）、内閣府・総務省（2026年8～9月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

DC立地は東京圏・大阪圏集中から地方分散へ

DC面積の88%が関東・関西に集中、2030年代は分散DCとワット・ビット連携へ

現状

民間企業の経済合理性で整備

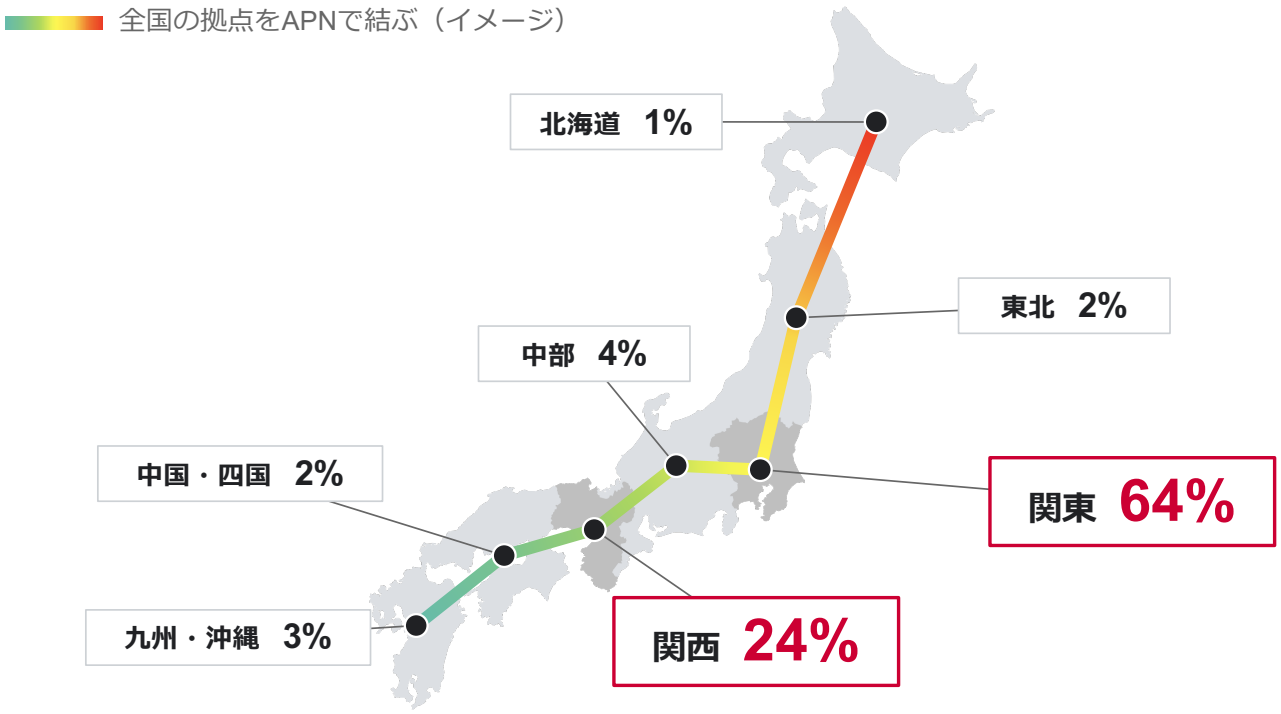
88%
が関東圏・関西圏に集中

地図の数字は地域別DC立地面積
（2023年）の比率

これから

首都直下・南海トラフへの備え

2030年代は、DC中核拠点と
DC拠点をオール光ネットワークで結ぶ



※比率は2023年の地域別DC立地面積に占める割合（単位は原資料に記載なし）　※濃い色は関東1都6県・関西2府4県。線は2030年代のイメージ

出所：ワット・ビット連携官民懇談会（第1回・2025年3月21日）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

ワット・ビット連携の概念図

電源立地の近くにデータセンターを分散配置し、利用者とは光（APN）で結ぶ

これまで

利用者		データセンター		電源立地（発電所）
都市部の企業と生活者	通信ネットワーク “ビット”＝通信	サーバー・NW機器	電力ケーブル “ワット”＝電力	風力・火力

ワット・ビット連携

利用者		同一エリア	
		データセンター	電源立地（発電所）
都市部の企業と生活者	**オールフォトニクスネットワーク（APN）** 電源の近くに拠点を置き、利用者とは光で結ぶ	サーバー・NW機器	風力・火力

※電力ケーブルに比べて、通信ケーブルの敷設・維持コストは100分の1とされる

出所：総務省・経済産業省（官民懇談会取りまとめ）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

分散型AI基盤とワット・ビット連携『取りまとめ1.0』

送電線より敷設・維持コストが100分の1とされる光通信網で電源立地DCと都市を結び、3段階で地方分散を推進する

足下のDC需要対策	DC集積拠点の造成	地方分散と高度運用
既存系統の最大活用とAPN実証	2030年頃～2030年代前半	2030年代前半～後半
オール光ネットワークの活用 電源と都市をAPNで結ぶ ユースケースの実証と拡充	**電力・通信の先行整備** GX戦略地域などGW級の集積地へ、 電力網と光通信網を先行整備	**APNワークロードシフト** 再エネ量や需給に応じ、APNで 拠点間のAI処理を移す（WLS）
系統余力公開と接続規律 ウェルカムゾーンマップと 接続ルールの見直しで既存設備を活用	**海底ケーブル・IX整備** 海底ケーブルの多ルート化と陸揚局の 地方分散で国外接続を確保	**蓄電池・DRの一体運用** 蓄電池・コジェネ・DRと小規模 DCを組み合わせ、系統負荷を軽減

※GX戦略地域の認定は経済産業省（2026年9月）による

出所：ワット・ビット連携官民懇談会『取りまとめ1.0』（2025年6月）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

GX2040ビジョンからGX戦略地域認定への政策工程

2025年2月のGX2040ビジョン閣議決定から1年半余りで、GX戦略地域（第1弾18地域）の認定まで政策が加速した



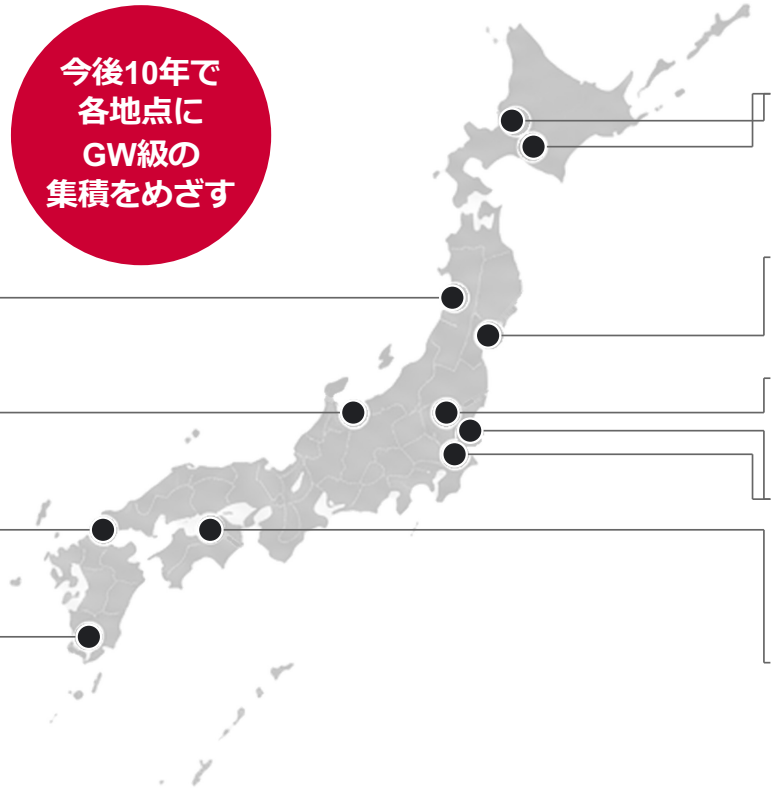
※認定は集積候補地の指定にとどまり、実際に送電系統と産業用地がいつ整うかは地点ごとに異なるため個別の見極めが不可欠

出所：経済産業省、時事通信（2026年9月）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

GX戦略地域（第1弾）：DC集積型9地域10地点

北海道（石狩・苫小牧）、秋田から鹿児島までの9地域10地点でGW級集積を目指す



北海道（2地点）
- ●石狩市（再エネ・寒冷地冷却）
- ●苫小牧市（工業地・港湾）

宮城県
- ●富谷市（東北の再エネ拠点）

栃木県
- ●矢板市（首都圏近郊・通信接続）

茨城県
- ●常総市・つくば市
（首都圏近郊・研究開発・通信接続）

香川県
- ●綾川町・坂出市・丸亀市
（産業団地の拡張）

秋田県
- ●秋田市（洋上風力・再エネ）

富山県
- ●南砺市（既存産業団地・水力）

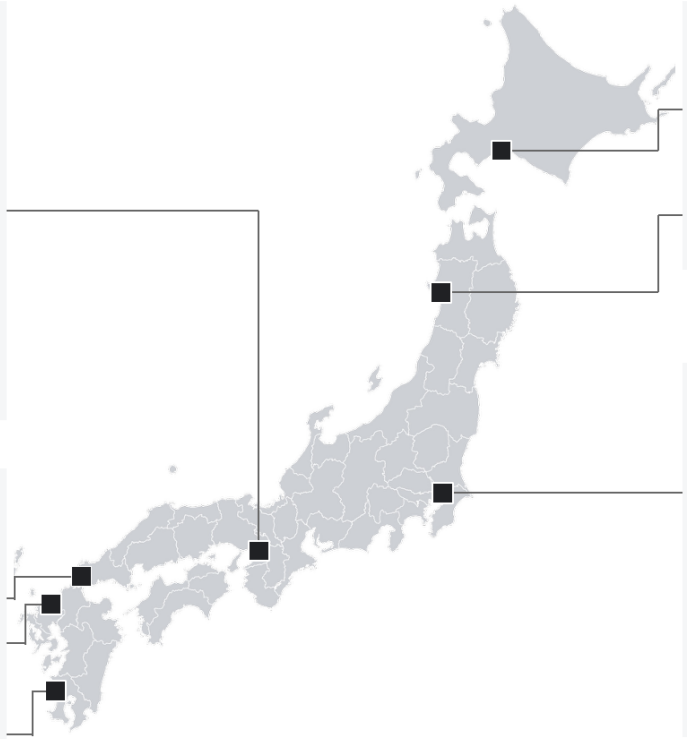
福岡県
- ●北九州市・直方市・鞍手町
（広域連携・新産業団地）

鹿児島県
- ●薩摩川内市（火力発電所跡地の
再開発）

出所：経済産業省「GX戦略地域（第1弾）」（2026年9月11日）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

今後の大型データセンターの計画



関西圏（大阪府）

NTTデータグループ 最大500MW
大阪府南部の湾岸｜建設費は最大1兆円
米Anthropic向けに最終調整

ソフトバンク 250MW
堺市｜150MWで2026年稼働目標

英コルト・ESR 130MW
箕面市｜2029年後半に第1期

九州

米APL
北九州市｜第1号案件、着工準備中 120MW
糸島市｜2029年から順次稼働 最大300MW

台湾の投資会社など 350MW
薩摩川内市｜2032年度に本格稼働

北海道・東北

ソフトバンク 300MW
苫小牧市｜2026年度に開業予定

米ビットグリット等・UAE 500MW
秋田市｜2030年代前半に稼働目標

関東圏（千葉県）

JERA・米Dell等 400MW
千葉市｜2028年ごろ運用開始

豪AirTrunk 300MW
印西市｜2021年に初期棟が開業

NTTデータグループ 200MW
白井市｜第1期は2030年以降

※計画規模は受電容量またはIT容量で、定義は計画ごとに異なる（2026年10月13日時点）　※北九州はAsia Pacific Land（2026年1月末時点）による

出所：ワット・ビット連携官民懇談会（2025年3月）、日本経済新聞（2026年10月13日）等を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

CHAPTER

05

AIOWNは
全国の資源を一つの基盤にする

AIOWNの価値と提供基盤、分散GPUの事例を紹介する

AIネイティブインフラ「AIOWN」の発表

AI活用の進展に合わせたリソース最適化・オペレーションを実現

2026 年 4 月 27 日

NTT 株式会社
株式会社 NTT データグループ
NTT ドコモビジネス株式会社

AI 活用の進展に合わせたリソース最適化・オペレーションを実現する
AI ネイティブインフラ「AIOWN」の展開

NTT 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）、株式会社 NTT データグループ（本社：東京都江東区、代表取締役社長：佐々木 裕、以下 NTT データグループ）、NTT ドコモビジネス株式会社（旧　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 ドコモビジネス)は、AI 利活用の急速な拡大を背景に、お客さまのニーズに合わせて最適な利用環境をお届けする AI ネイティブインフラ「AIOWN」を展開していきます。



出所：NTT報道発表「AIネイティブインフラ『AIOWN』の展開」（2026年4月27日）
https://www.ntt.com/about-us/press-releases/news/article/2026/0427_3.html

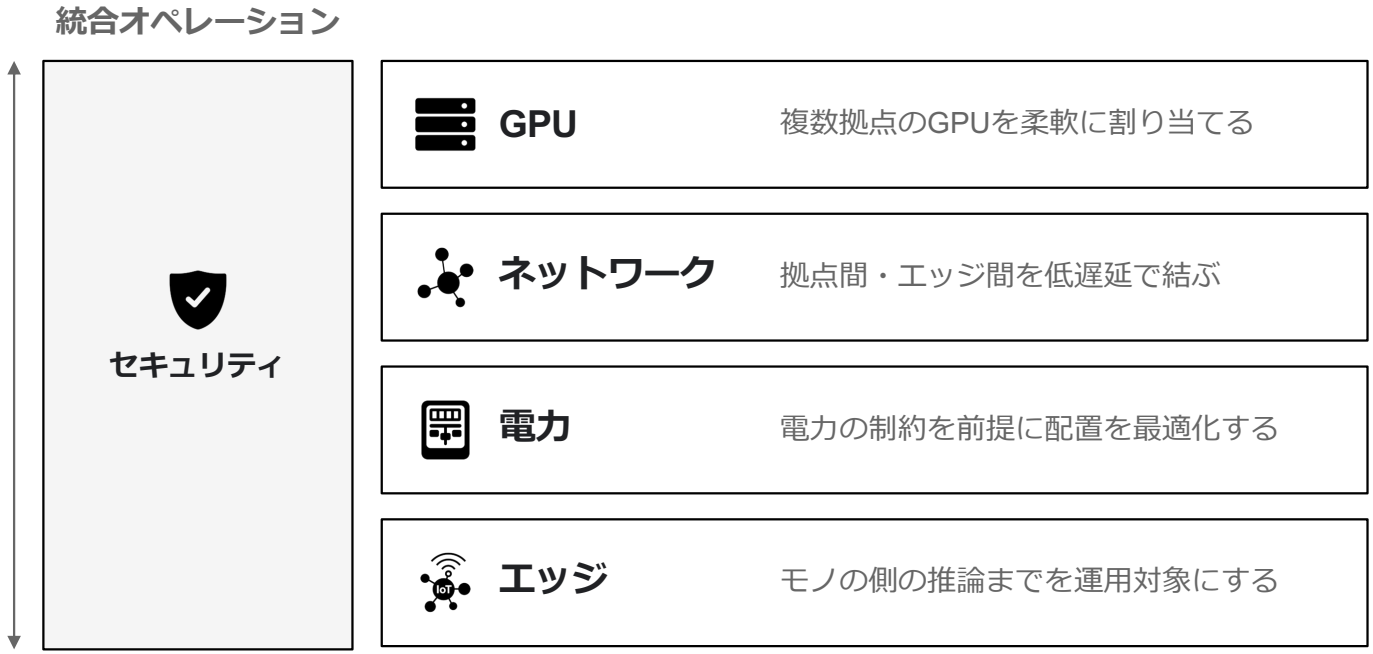
© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AIネイティブインフラ「AIOWN」

GPU・ネットワーク・電力といったリソースを最適化し、エッジを含めて統合運用。セキュリティを横断的に組み込み、全国に分散した資源を一つの基盤として扱う構想

AIOWN
AI x IOWN

全国に分散したデータセンターとエッジを、統合されたオペレーションで束ねるAIネイティブインフラ

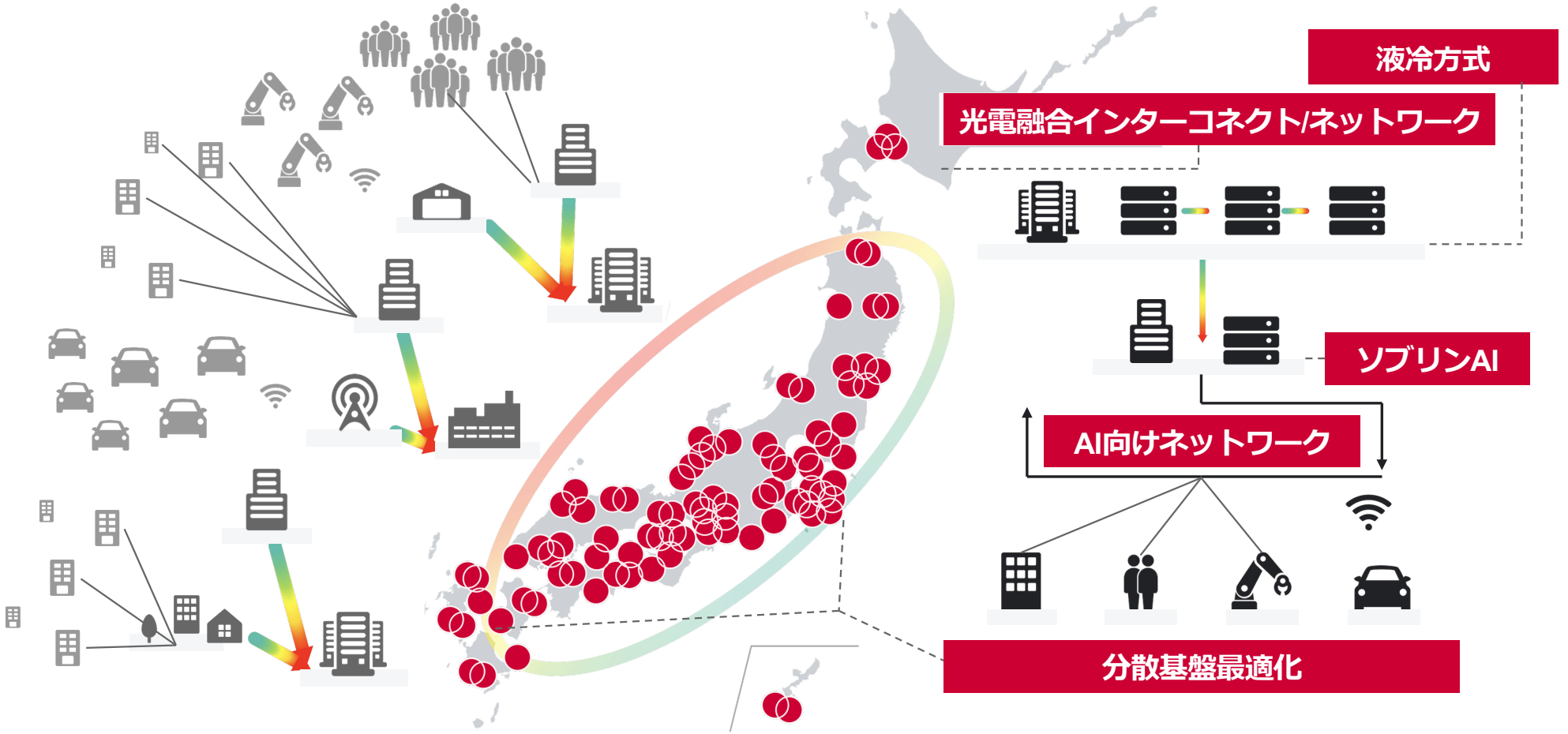


注：「AIOWN」は商標出願中

出所：NTT報道発表「AIネイティブインフラ『AIOWN』の展開」（2026年4月27日）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

NTTが取り組むAIネイティブインフラ



出所：NTT報道発表「AIネイティブインフラ『AIOWN』の展開」（2026年4月27日）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

ICTベースからAI時代に合わせたインフラ環境へ

機器構成はCPUからGPU中心へ、残る4項目は従来の要件にAI時代の要件が加わる





	従来ICT		AI時代
機器構成	CPU中心	➡	GPU中心
ネットワーク指標	帯域（Gbps）	＋	AIへの応答速度※・トークンスループット
データセンター要件	フロア面積・好アクセス	＋	電力密度・発熱対策
セキュリティ	データの保護	＋	ノウハウ/インテリジェンスの保護
コスト	サーバー・ストレージ・セキュリティ・データセンター・ネットワーク	＋	GPU（高コスト化）

コア業務へのAI適用、フィジカルAIなどさらなる普及により、求められるニーズは変化

※ P99/P99.9のテールレイテンシ（Tail Latency）：全リクエストのうち上位1%（P99）または0.1%（P99.9）の最遅リクエストの応答時間を指し、システム全体のうち「ほぼ最悪に近いケース」の遅延を評価する指標

出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

テールレイテンシとトークンスループットへの対応

演算リソース・光インターコネクト等のフルスタック統合で、低遅延・高いスループットと安定運用へ

テールレイテンシ（AIへの応答の遅さ）

上位1%（P99）／0.1%（P99.9）の最遅応答時間

「ほぼ最悪」の体感を評価

単位：**ms**

トークンスループット（処理量）

単位時間に生成・処理するトークン数

さばける処理量＝収容力

単位：**tokens/sec**

フルスタック統合

AIネイティブインフラ＝6つの要素

演算リソース

GPU性能＋HBMの
帯域・容量／ASIC

光インターコネクト

APN／光電融合／
NVLink／InfiniBand

推論ソフトウェア

推論エンジン・
フレームワーク

ストレージ／データ

高速読込・RAG・周辺処理

電力・冷却

高密度ラックと液冷

運用・最適化

持続的スループットの追求

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

お客さまニーズに応えるAIネイティブインフラ

AI活用の進展と領域拡大から生じる4つのニーズに、5つの技術要素で対応

ニーズとインフラの変化

AI活用の進展と領域拡大（多様なAI）

01	高密度ラックの発熱対応
02	DC／サーバー間の低遅延
03	機微なデータの扱い
04	エッジへの拡大

AIネイティブに求められる技術要素

Single-phase direct-to-chip liquid cooling	GPU高密度ラックを効率的かつコスト効率よく冷却する**液冷方式**
Photonics — electrics-converged interconnect/network	分散学習や推論を支える低遅延・超高速の**光電融合インターコネクト／ネットワーク**
Sovereign AI	ユーザーがデータ主権を持ち、安全性と法規制を確保する**ソブリンAI**
Distributed infrastructure design / Infrastructure Orchestration	GPU基盤やエッジ、オンプレミス環境などの**分散基盤最適化**
Network for AI	推論などのAIワークロードに最適化された**AI向けネットワーク**

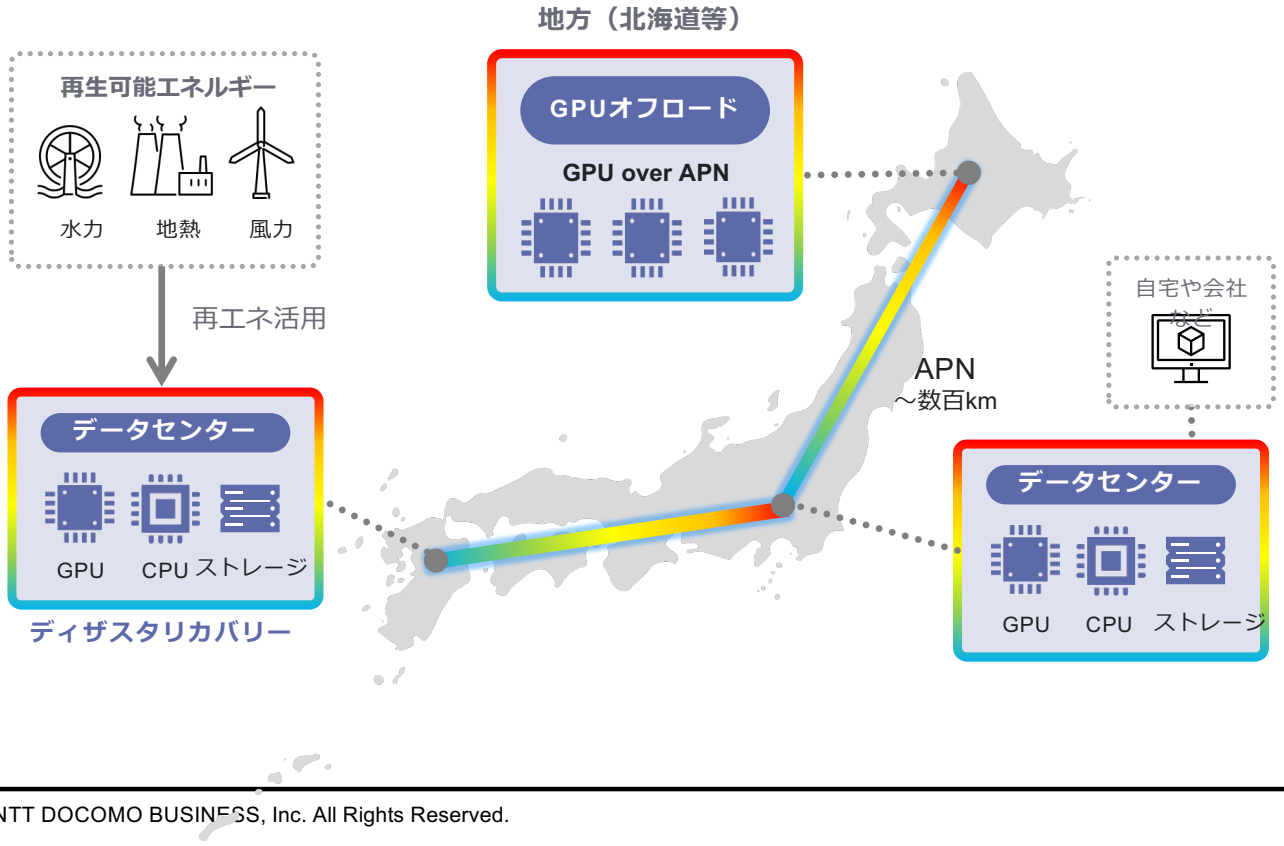
出所：NTT（2026年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AI時代を支える分散型コンピューティングインフラ

地方の再生可能エネルギーを活用し、計算資源とデータを最適に分散配置するAIインフラを実現

AIインフラ戦略を支える実装例



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の向上も確認

※NTTドコモビジネスによる実証数値

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

トヨタ自動車とNTTのモビリティAI基盤

モビリティAI基盤は計算基盤・AIモデル・通信基盤の3本柱で交通事故ゼロへ

2024年10月
交通事故ゼロ社会の実現に向け、
共同開発・運用を対外公表

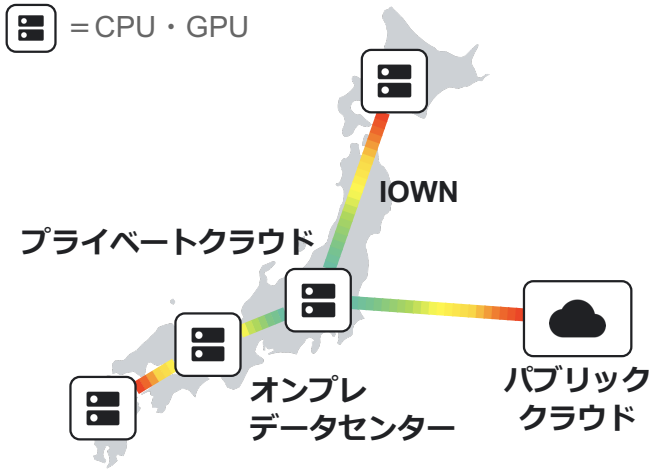
2028年頃～
パートナー各社とインフラ協調による
社会実装を推進

2030年以降
普及拡大をめざす

モビリティAI基盤

①分散型計算基盤

データセンターに分散型計算基盤を構築



②インフラ協調AIモデル

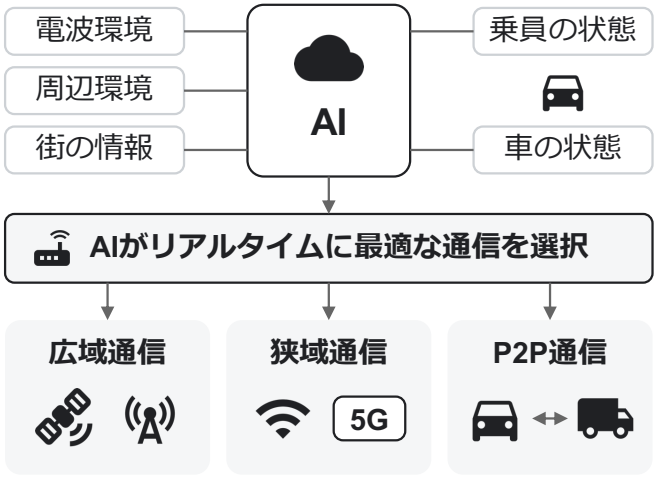
業界特有のデータを学習したモビリティAI

LMM（Large Mobility Data Model）

- 三位一体の事故ゼロ
- データドリブンで自動運転を高度化
- ヒトに寄り添うAIエージェント
- 省エネ・物流効率化

③インテリジェント通信基盤

インテリジェントで高信頼な通信の手段



出所：NTT（2025年）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

AIネイティブインフラ「AIOWN」の価値

あらゆるAIに効率的かつ効果的に対応するため、AI活用の4つの課題を4つの価値へ

	AI活用の課題	対応	AIOWN（AI×IOWN）の価値
1	増え続ける電力	光電融合／液冷データセンター	省電力化
2	トークン量の増大	トークン圧縮／リソース振り分け	トークン効率最大化
3	AIの接続先の多様化・増大	多様化する接続先にも対応したセキュリティ	セキュリティ
4	データの流れのブラックボックス化	データ主権の確保	ソブリンティ

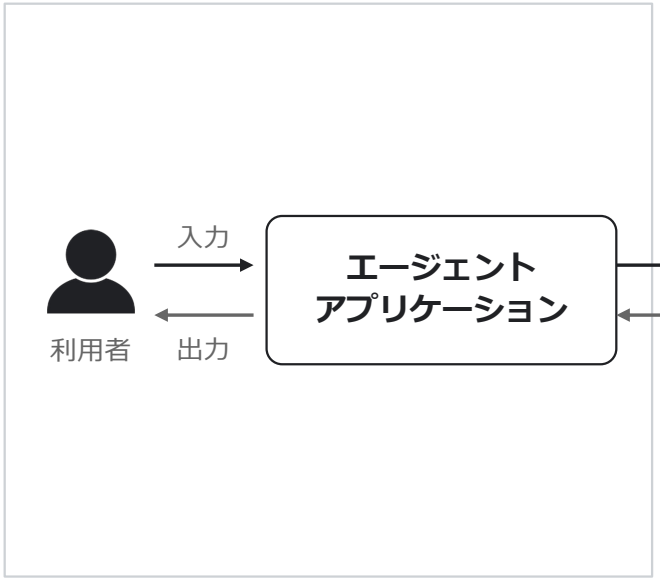
出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

トークン量は見えない内部処理で増大

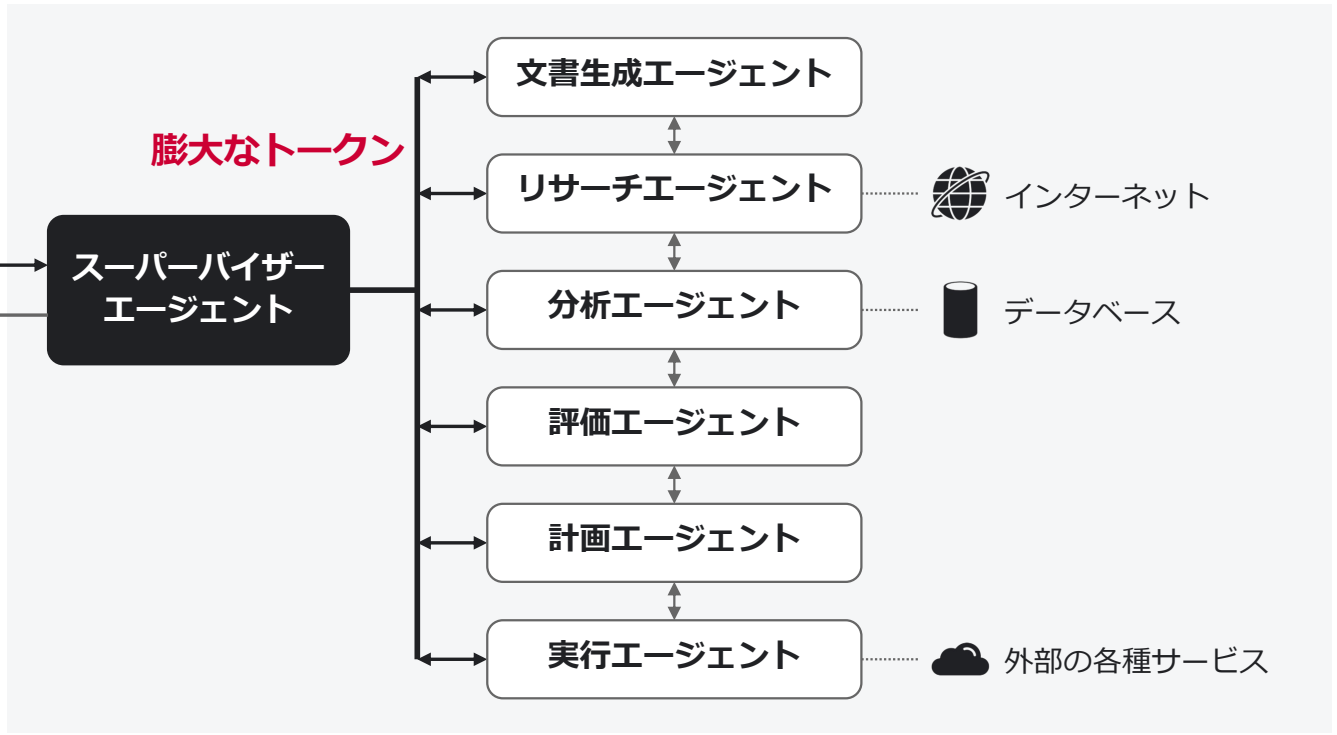
内部で6つのエージェントがやり取りを重ねるため、見えないままトークン量が増大

利用者から見える世界



AIエージェントの内部処理

利用者からは見えない



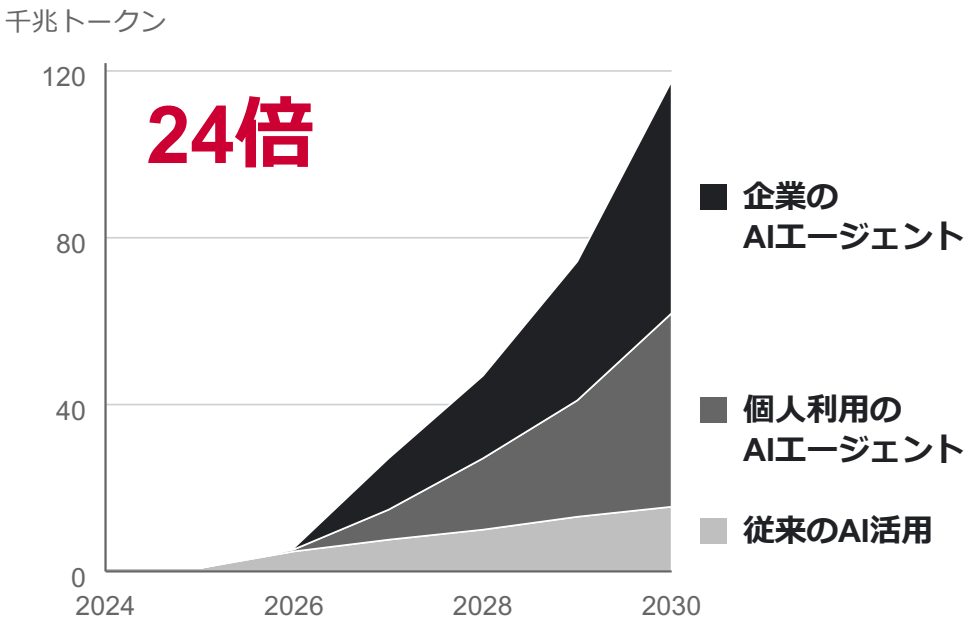
出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

トークン量の増大が企業のAI予算を圧迫

トークンの増大に伴い、企業のAI活用のコストは直近3カ月間で3～6倍に

トークン消費量は5年で24倍となる予測



企業の予想をはるかに超えるAIコスト

年間コーディング予算を年初4カ月で完全消費 — Uber

- AIコーディングツールに月額1,500ドルの上限を設定
- AI利用コストが、開発の投資利益率（ROI）を低下

OpenAI APIの年間予算を2カ月で消費 — 国内企業

- API利用でトークン量が大幅に増加
- 他AIモデルでは、トークン課金への変更で月額利用料が年2,000万円から年3～4億円になった例も

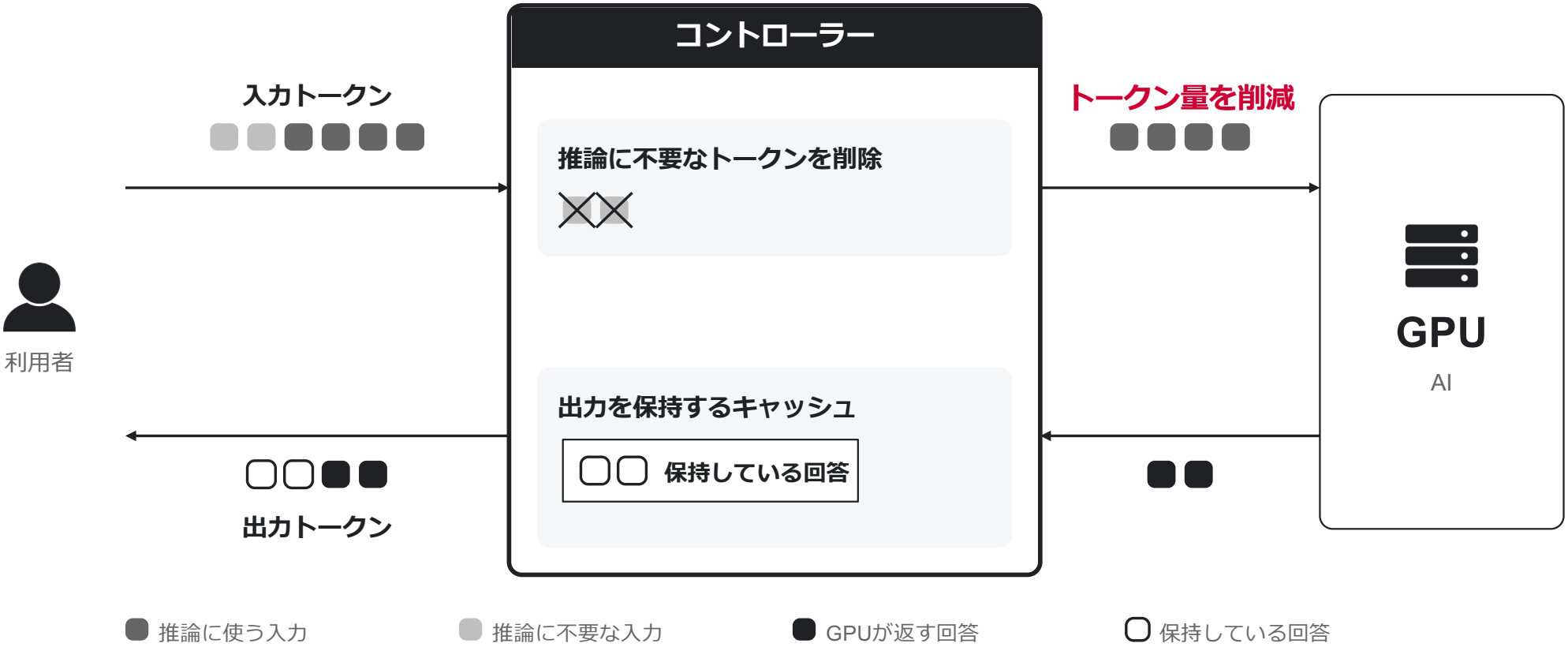
※左図はGoldman Sachs Research Estimates as of May 2026を基に作成　※レポート名："AI Agents Forecast to Boost Tech Cash Flow as Usage Soars"

出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

GPUに渡すトークン量をコントローラーで圧縮

入出力トークンを前段で加工・保持することで、GPUで処理するトークン量を削減



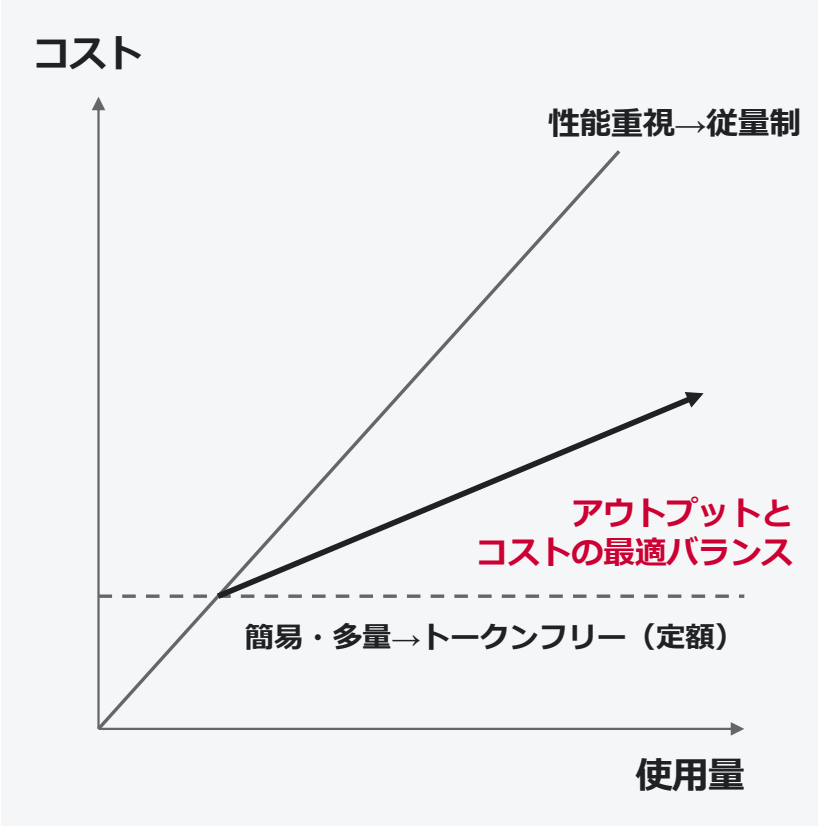
出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

モデルルーティングで従量制と定額を使い分け

業務やタスクごとにモデルを従量制と定額の2環境へ振り分け、トータルコストを低減





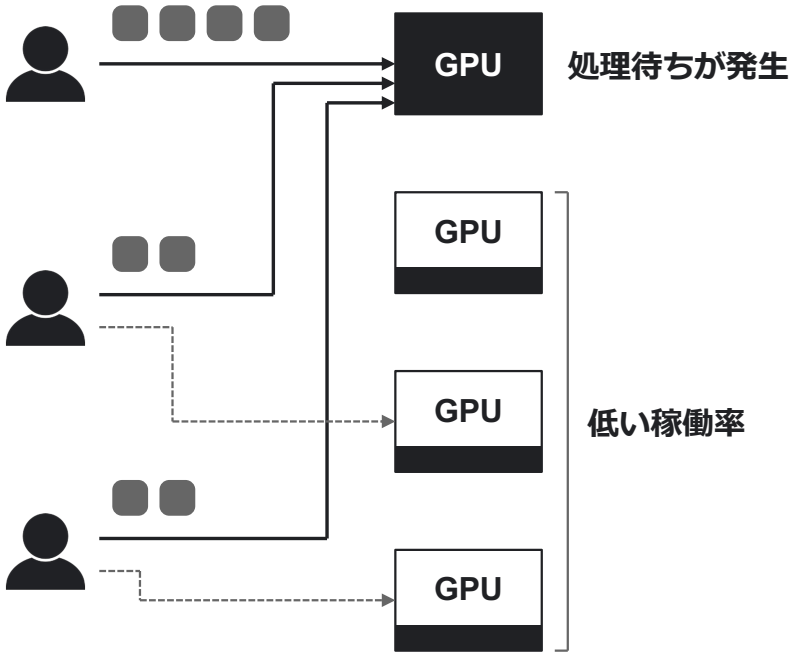
出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

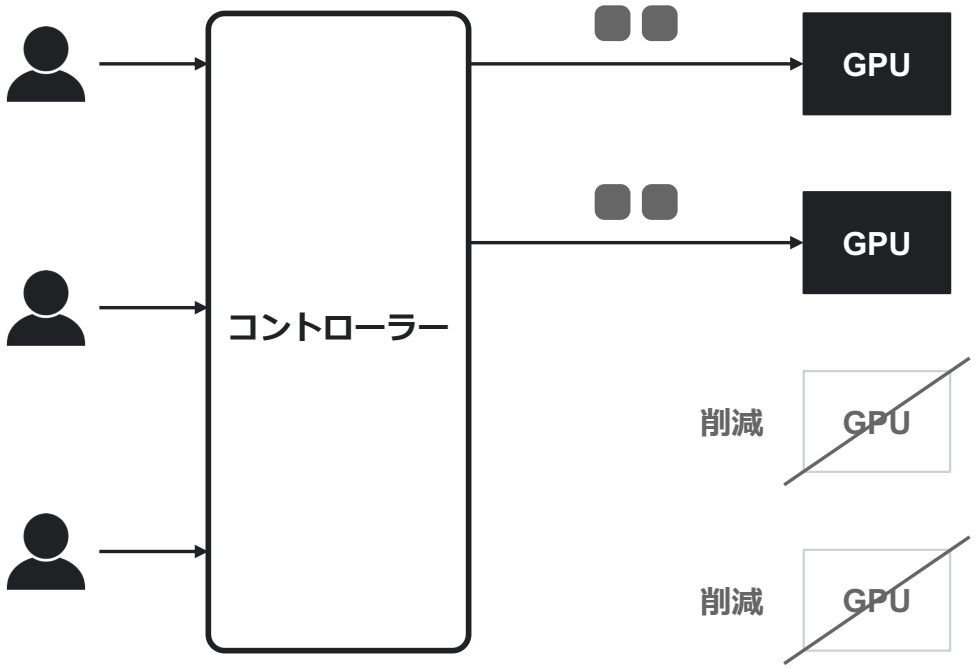
GPUの使い方は集中から振り分けへ

1台への集中をなくし、それぞれの稼働率を高めることで、少ないGPUで経済性を向上

現状



GPUの最適振り分け



出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

スループットとトークンマネジメントは両輪

エージェントで増えるトークンをさばくためには、収容力の増強とGPUに渡す量の削減

トークンスループット
フルスタック統合で収容力を増強

演算リソース	光インターコネクト	推論ソフトウェア
ストレージ／データ	電力・冷却	運用・最適化

低遅延・高スループット・安定運用

＋

トークンマネジメント
コントローラーでトークンの削減と振り分け

1. **前段で圧縮** 不要トークンの削除とキャッシュ
2. **モデルルーティング** 従量制と定額を使い分け
3. **GPUの振り分け** 集中をなくし、稼働率を向上

処理トークン・GPU総数・コストを削減

出所：NTT（2026）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

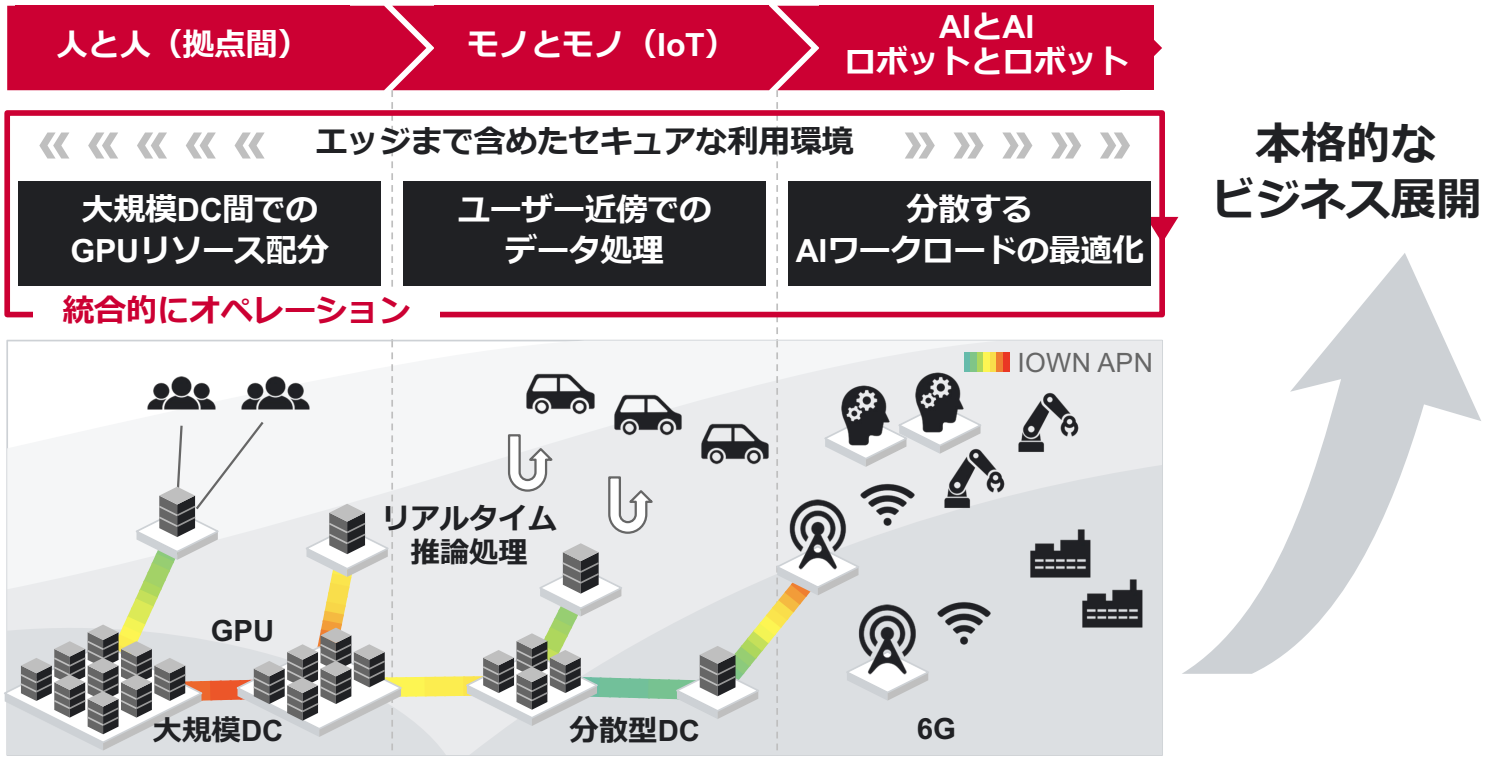
AIネイティブな次世代インフラへの転換

GPU・ネットワーク・電力といったリソースを最適化し、エッジまで含めたオペレーションを担う
AIネイティブインフラ「AIOWN」への転換を推進

AIOWN
AI x IOWN



「AIOWN」は商標出願中です



本格的な
ビジネス展開

出所：NTT 2025年度決算（2026.5.8）を基に作成

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

つなごう。驚きを。幸せを。

NTT docomo Business

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.