企業のAI活用を阻む「インフラの壁」を乗り越える方法とは？

GPU、データセンター、ネットワークを駆使した統合的なアプローチ

ビジネスにおける AI の活用は急速に拡大している一方で、AI を用いた大規模なシステムの導入・活用には、高性能な GPU だけでなく、データセンター、ネットワーク、ストレージといった複雑なインフラ整備が不可欠であり、多くの企業が課題に直面している。本稿では、AI インフラの整備に求められる技術要素を踏まえ、AI 活用を成功に導くための最適なインフラ構築・運用のアプローチを解説する。

AI 活用の本格化に伴い生じた「インフラ構築の課題」

生成 AI や機械学習の活用が急速に拡大する現在、AI のためのインフラ整備は企業競争力に直結する重要な要素となっている。また、AI インフラの構築・運用に求められる技術的要件に変化が生じている。AI モデルの高度化に伴い、高性能な GPU や、こうした GPU の高発熱に耐えられるデータセンター、広帯域ネットワークへの需要が急増。さらに GPU においては、膨大な消費電力を必要とするという課題があり、2018 年の 100W から 2024 年には 700W 超、将来的には 15kW 規模まで増加すると見込まれている。

もちろん、対策するべきはこうした GPU や電力の問題だけではない。AI インフラ構築においてはネットワーク、ストレージなどを含む総合的なアプローチが不可欠である。さらに、製造業や半導体業界など設計情報をはじめとする機密性の高い情報を扱う企業では、セキュリティの観点からパブリッククラウドの利用をためらうケースも少なくない。そのため、秘匿性の高いデータはオンプレミスやプライベートクラウド、システム構築のスピードや柔軟性が求められるデータはパブリッククラウドといったように、ハイブリッドな環境を構築し、用途に応じて使い分けることが重要となっている。

インフラからソフトまで広範な専門スキルが必須

AI インフラの整備は、多くの専門的なスキルが求められる。先述のように、GPU やデータセンター、ネットワーク、ストレージといったハード面の整備はもちろん、その上のミドルウェアやソフトウェアまで多岐にわたる要素があるため、それらに関して自社のビジネス課題、利用用途に応じた最適な構成を検討・設計する必要がある。それは決して簡単ではなく、時間とコストがかかるため、検討の初期段階でつまずくケースがある。構築はもちろん、その後の運用工程も複雑化しており、これらをすべて自社内のリソースのみで行うことは困難な状況だ。

例えば GPU という要素 1 つを取ってみても、その性能を最大限に引き出すためのアーキテクチャー設計や、複数の GPU を連携させるためのクラスター構成、それに伴うネットワーク構成、ストレージ設計、さらにはソフトウェア部分まで、求められる専門領域が多岐にわたる。

さらに、AI インフラは先述した電力という問題があるため、他の IT システムのインフラとは異なり、特にファシリティ面で注意するべき点が多い。プライベート環境での構築を望む企業は、自前のデータセンターや自社ビルへの設置を検討するが、AI インフラは発熱の冷却対応、大量の電力供給が必要不可欠であり、これを自社だけでカバーするのは困難だろう。

多彩な技術で包括的な支援を提供する「AI インフラソリューション」

こうした複雑化する AI インフラ構築の課題を解決するため、NTT ドコモビジネスが提供しているのが「AI インフラソリューション」である。同ソリューションは、GPU サーバー、ネットワーク、ストレージ、超省エネ型データセンターなど、AI 活用に必要なインフラ基盤をワンストップで提供し、企業の AI インフラの設計・構築・運用を包括的にサポートするものだ。代表的なサービスや支援内容として以下の 4 つが挙げられる。

1 つめは、高性能 GPU を専有型で利用できる「Smart Data Platform(SDPF) GPU プラットフォーム」だ。これは、国内データセンターで専有型の GPU 基盤を IaaS として提

供するものだ。月額定額制で最新の GPU を利用でき、たとえば、大規模言語モデルの事前学習やファインチューニングなどの GPU 稼働率の高い用途では、プライベートクラウドよりも安価なコストで利用が可能となっている。また、高性能なネットワーク構成とマネージドサービスでの提供により、運用の安定性を確保している。

2つめは、液冷方式のサーバー機器に対応したデータセンター「Green Nexcenter®」だ。液体冷媒を使った効率的な冷却方式により、従来型データセンターでは冷却の限界と言われていた 1 ラックあたり 20kW 以上の電力消費に対応し、現在サービス提供している大阪第 7 データセンターでは最大 80kW まで対応。冷却用の消費電力は従来比で 30%以上削減可能だ。こうした超省エネ型データセンターにより、高性能な GPU サーバーを高密度で配置できる。

3 つめは、NTT が推進する IOWN® 構想（最先端の光関連技術および情報処理技術を活用したネットワーク・情報処理基盤の構想）にもとづき APN（オールフォトニクス・ネットワーク）技術を取り入れた「docomo business APN Plus powered by IOWN®」だ。AI 活用の拡大に伴う利用形態の変化や多様性に応じ、大容量、低遅延かつ揺らぎのない通信品質の特長に加え、柔軟かつ迅速に帯域や経路変更ができるネットワークを実現。

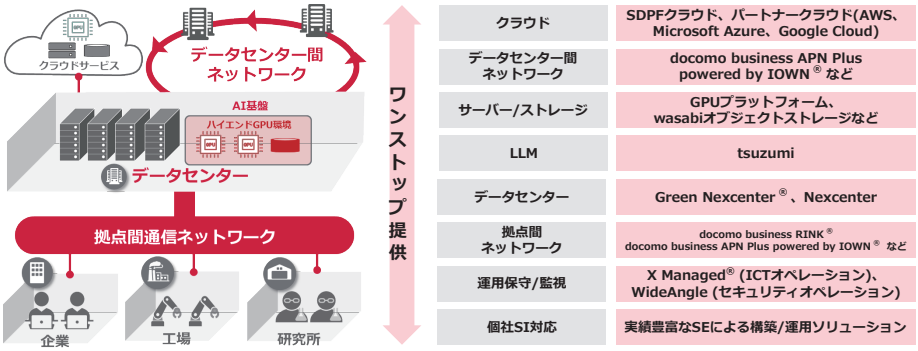
そして 4 つめは、「専門家による技術支援」だ。NTT ドコモビジネスは、NVIDIA の GPU サーバー「DGX シリーズ」の設置資格を持つコロケーションパートナーであるほか、AI インフラ基盤の設計 / 構築に知見をもつグループ会社の NTTPC コミュニケーションズとの連携も可能だ。同社は、NVIDIA のパートナー認定制度 NPN (NVIDIA パートナー ネットワーク) において「DGX AI Compute Systems」「Compute」コンピテンシーで最上位レベルのエリートパートナー として認定を受けている。高い技術力に基づき、AI 用途に最適化したアーキテクチャー設計・構築・運用を提供し、最適な AI 環境の実現をサポートする。

導入から運用まで AI 活用の定着を促進

企業は AI インフラソリューションを活用することで、複数社にまたがるさまざまな技術要素を選定し、SIer などにインテグレーションを依頼して構築するといった従来のシステム構築の煩雑さを解消できる。

さらに同ソリューションは、導入段階だけでなく利用用途に応じた最適なインフラ設計や運用チューニングを支援することで、企業の AI 活用の定着とスケーラビリティ確保を実現する。構成要素が複雑で専門スキルが多岐にわたる AI インフラにおいて、この「ワンストップでの提供」と「フルマネージドでのカバー」が、NTT ドコモビジネスの独自価値となっている。

AI インフラの活用例は多岐にわたり、現在製造業や研究開発機関などさまざまな事例が登場している。しかし、利用目的やデータの特性に応じて最適なインフラ構成は異なり、構築にあたっては多くの検討事項がある。その中で NTT ドコモビジネスは、AI インフラの全領域をワンストップで提供し、顧客それぞれに応じて最適な AI 活用を強力に支援する。



AI インフラソリューションは LLM からクラウド、データセンター、ネットワーク、マネージドまでワンストップで提供可能

NTTドコモビジネス株式会社

〒100-8019 東京都千代田区大手町2-3-1　大手町プレイスウエストタワー
プラットフォームサービス本部 クラウド&ネットワークサービス部　E-mail：digitalsol-cn@ntt.com