AITRAS

AI-RAN統合ソリューション

1. はじめに

AIは、産業や企業活動、そして私たちの日常生活における利便性や生産性に新たな変革をもたらしつつあります。AI技術は日々飛躍的に進化しており、生成AIなどの人に対してのサービスに留まらず、今後はロボット、ドローン、車両などのデバイスをリアルタイムで制御するためにも活用が広がっていくでしょう。

これらのユースケースに共通して求められるのは、広範囲にわたる接続性や、高い信頼性と安全性を備えた超高速な通信環境です。この新しいAIに基づいたトラフィックに対応するためにネットワークも進化していかなければなりません。通信事業者は、既存の中央および分散インフラストラクチャに加えてAIの学習と推論のための基盤も構築する必要があります。AIとRAN（Radio Access Network/無線アクセスネットワーク）を統合する革新的なアプローチであるAI-RANは、モバイルインフラの中でAI推論に最適な環境の提供も可能にするため、今後のモバイルコミュニケーションを大きく変革していきます。

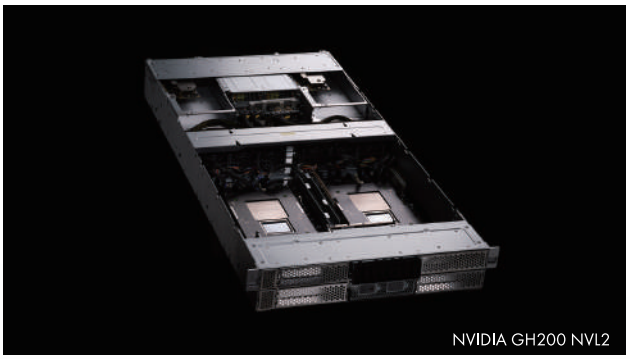
ソフトバンクのAI-RANソリューション「AITRAS」は、無線およびAIの両方のワークロードを同時に共有可能なGPUベースのインフラストラクチャです。 これにより、通信事業者はモバイルインフラをコストセンターから収益源へと効果的に転換することが可能となります。「AITRAS」は、ネットワークインフラ利用効率の向上、コスト削減だけでなく、テレコミュニケーションに新しいビジネスモデルも導入することで、新たな収益源の模索の機会を提供することができるため、モバイル事業の変革に向けた重要なファクタとなります。

2.「AITRAS」概要

AI-RANコンセプトに準拠したソフトバンクの「AITRAS」は、AIとRANを単一のインフラで統合することにより、テレコミュニケーションネットワークを革新的に変える機会を提供します。このGPUベースの統合型インフラは、通信事業者がRANとAIのワークロードを同時に実行できるよう設計されており、リソースの効率を最大化しながら運用の無駄を削減します。

「AITRAS」は統合インフラとして主に以下の特長があります。
- AIとRANのマルチテナントおよびAIベースのオーケストレーションにより、柔軟性を向上させ、コストとリソース利用を効率化し、無駄を低減
- 各種AIアプリケーションの開発、展開および収益化が可能
- 高機能、高性能、高品質なキャリアグレードのRANを実現
- AIにより、エネルギー効率を向上させ、ROIの改善と全体的な運用コストの削減を実現

ソフトバンクは、「AITRAS」を2026年以降にグローバル展開することを目指しています。また、2025年以降には、他の通信事業者がAI-RANコンセプトの実用性とその価値を評価できるように「AITRAS」のリファレンスキットも提供する予定です。これにより、通信事業者は、容易に「AITRAS」の性能を試すことが可能になります。



3.「AITRAS」主な構成と機能

物理システム構成

NVIDIA GH200 Grace Hopper Superchip、Radio Unit、ネットワークスイッチ群を主なパーツとして構成されます。

論理システム構成

GPUベースのNVIDIA GH200上に、仮想化基盤、L1/L2/L3※で構成されるRANと、AI等が動作するEdge AI、そして、AIおよびRANのアプリケーションが正常に動作するために必要なコンピューターリソースを提供するオーケストレーターで構成されます。

※L1/L2/L3:　RANソフトウェア構造におけるOSI参照モデル「物理層（第1層）」「データリンク層（第2層）」と「ネットワーク層（第3層）」

リソース管理機能

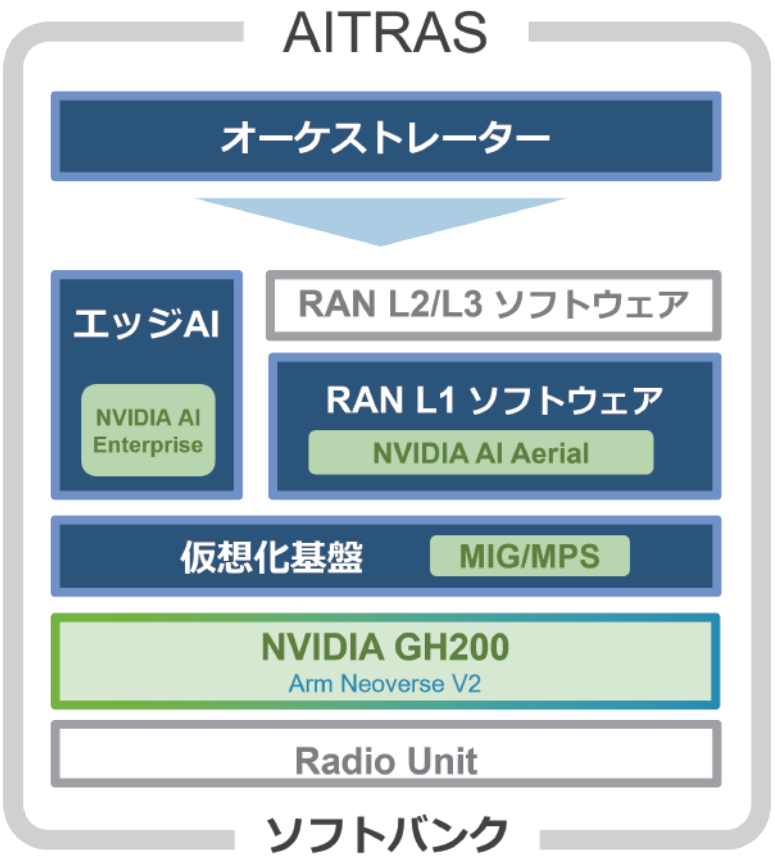
複数の拠点間やサーバー間であっても、最適化されたデータフローメカニズムを介して、AIとRANのワークロードとコンピューターリソースが同時管理され、AIによって効率的なリソース制御が行われます。

RAN機能

高機能、高性能およびスケーラビリティを実現するために、ネットワーク機能はNVIDIA GH200上でソフトウェアのみで完全に仮想化されており、RANソフトウェアも柔軟に展開および管理可能です。また、RANの機能を高度化していくことも容易になります。

RANにおけるAIと機械学習モデル

「AITRAS」はさまざまなAI機械学習モデルを活用することで、エネルギー効率を含むRANの全スタック性能を向上させるよう設計されています。これらのモデルは、チャネル推定、変調、エラー訂正などの無線信号処理に適用され、RANの効率と性能を大幅に改善します。



4.「AITRAS」基盤の設計ポリシー

AIを活用したオーケストレーター

AIを活用したオーケストレーター
オーケストレーターによってAIおよびRANのワークロードが管理され、効率的なリソース利用が保証されます。オーケストレーターは、需要に基づいて動的にリソースを割り当てるAIモデルを使用することで、性能を最適化し、遅延を最小限に抑えます。

マルチアクセスエッジコンピューティング（MEC）の統合

AIアプリケーションがMEC機能を含んだEdge AIでサポートされることで、シームレスなユーザー体験を提供します。「AITRAS」は、データをユーザーの近くで処理するため遅延を低減させ、自動運転サポートやリアルタイムビデオ分析などのアプリケーションの性能を向上させます。

クラウドネイティブ設計

仮想ネットワーク機能およびクラウドネイティブネットワーク機能の両方に最適化されたKubernetesベースのコンテナ化ネットワーク機能を有しています。「AITRAS」は分散型エッジデータセンターを想定したソリューションですが、このクラウドネイティブアプローチにより、中央集中型データセンターにも柔軟に展開が可能です。

スケーラビリティと柔軟性

「AITRAS」は、分散型、集中型などのさまざまなRAN構成の展開シナリオに対応できるように設計されています。このアーキテクチャにより、通信事業者は需要に基づいてネットワークをスケールさせることができ、ネットワークの計画および展開に高い柔軟性を提供します。

5. AIに基づいたRANの管理と自動化

ネットワーク機能のライフサイクル管理

AIに基づいたオーケストレーションシステムにより、AIベースのネットワーク機能の展開、スケーリング、障害復旧、およびアップグレードが自動化されます。これにより、ネットワークは需要に応じて自動的に適応し、運用の複雑さを軽減します。

自動プロビジョニング

RANのゼロタッチプロビジョニング（ZTP）により手動によるオペレーション工数を削減します。ZTPにより、RANの迅速な展開が可能となり、新サービスの商用化までの時間やネットワークメンテナンスに要する時間を短縮します。

予測保守

AIを活用した能動的な監視と保守により、ネットワークの信頼性を確保します。予測保守では、機械学習アルゴリズムを使用して潜在的な問題を事前に特定し、ダウンタイムを最小限に抑えるとともに、ネットワークの可用性を向上させます。

リソース割り当てとトラフィック管理

最適なネットワーク性能を実現するため、リアルタイムで動的なリソース割り当てを行い、MIG（マルチインスタンスGPU）機能を使用したマルチテナント機能を備えています。これにより、AIとRANのワークロードが効率的にリソースを共有し、インフラの利用効率が最大化され、ネットワーク全体の性能が向上します。

6. セキュリティとコンプライアンス

マルチテナント環境

「AITRAS」は、ワークロードごとに隔離された安全なリソース環境を提供し、キャリアグレードの信頼性を確保します。アプリケーションごとに独立したインスタンスGPU単位でリソースを割り当てることで、マルチテナントをサポートし、AIとRANの間、さらには異なるAIの間でもワークロードが独立して処理されます。

データプライバシーとコンプライアンス

データのプライバシー保護と規制要件の順守を実現するための組み込みメカニズムに対応しています。「AITRAS」には、データ保護のための暗号化技術やアクセス制御機能が含まれており、GDPR（EU一般データ保護規則）などの規制にも準拠しています。

7. Edge AI

Edge AIは、低遅延で安全性の高いAIアプリケーションを提供します。また、NVIDIA AI Enterpriseを活用し、企業やユーザーがAIアプリケーションを容易に開発、展開できる環境を実現します。ソフトバンクが開発したAIアプリケーションの具体例として以下が挙げられます。

マルチモーダルAIによる自動運転遠隔サポート

車載カメラの映像などを5GネットワークをEdge AI上で動作するマルチモーダルAIに伝送することで、自動運転をサポートします。このAIはリアルタイムで交通状況を分析しリスク評価を行います。その結果、遠隔監視者や車両に対しリスク回避のための推奨アクションをチャットインターフェースを通じて指示します。

Edge RAGによる業務の効率化

オフィスや工場、建設現場などの企業データを5GをEdge AI上で動作するRAG（検索拡張生成）に入力することで、企業は自社の内部情報に基づく高精度な検索結果を得られます。さらに、自社業務に特化したタスクを生成AIに任せることが可能となり、業務や工程の進捗状況の可視化などが自動的に実現されます。なお、企業データは、インターネット上のクラウドではなくEdge AIに保存されるため、機密性とデータ主権が確保されます。

リアルタイムロボット制御

ロボット搭載カメラの映像などを5GをEdge AI上で動作するロボット制御AIに伝送し、ロボットが人の指示や動きに応じて動作します。Edge AIによる制御AIの反応時間は、インターネット上のクラウドと比較し、低遅延でかつ安定的であるため、リアルタイムでのロボット制御に適しています。

8.「AITRAS」の導入メリット

コスト削減

AIとRANのワークロードをGPUベースの「AITRAS」に統合することで、AIとRANそれぞれの専用のハードウェアは不要となり、投資コストと運用コストを大幅に削減できます。

インフラリソースの効率化／柔軟性

オーケストレーションにより、AIとRANのワークロード間でコンピューターリソースを動的に割り当てることが可能となり、インフラの利用効率を向上させます、これにより、迅速かつ柔軟なサービス提供を実現します。

新たな収益の創出

Edge AIをRANと組み合わせて活用することで、通信事業者は、AIをベースとした新たなビジネスを創出し、変化する市場需要に対応可能な競争優位性を確立することが可能になります。

キャリアグレードRANの性能向上

NVIDIA GH200と、ソフトバンクがNVIDIA AI Aerialをベースに開発したL1ソフトウエアを活用することで、キャリアグレードの安定性と性能を備えたRANを提供します。また、RAN容量の最大化や消費電力の削減も実現されます。さらには、C-RAN（集中型RAN）にAIの効果を組み合わせることで、複数セルにおける高性能化も実現できます。

「AITRAS」は、

通信事業者に新たなビジネス創出の機会をもたらす

革新的なソリューションです。

また、AIと通信の融合を加速させることで

「AI共存社会」を支える次世代社会インフラの実現において、

欠かせない重要な役割を担っていきます。

SoftBank