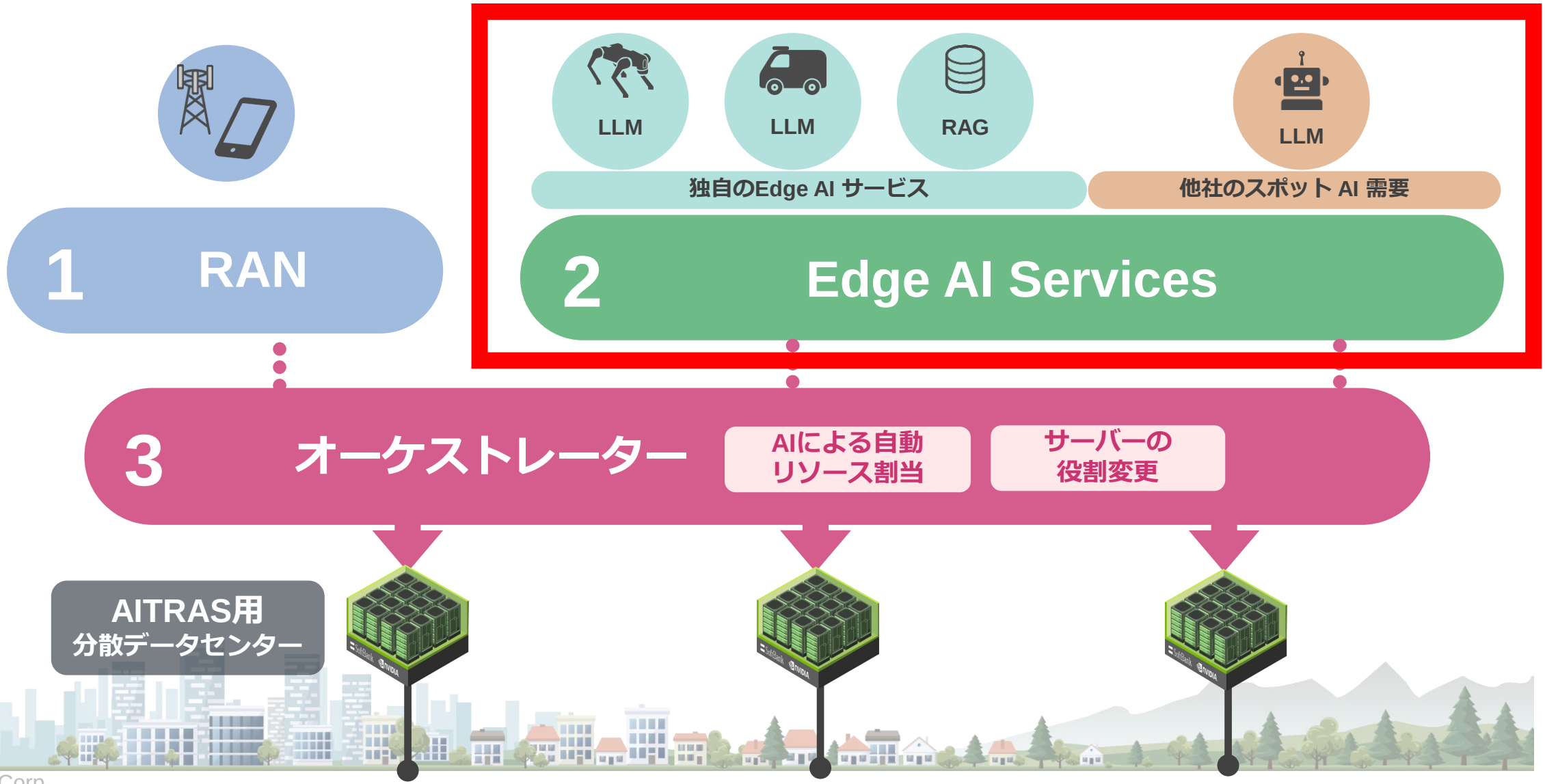
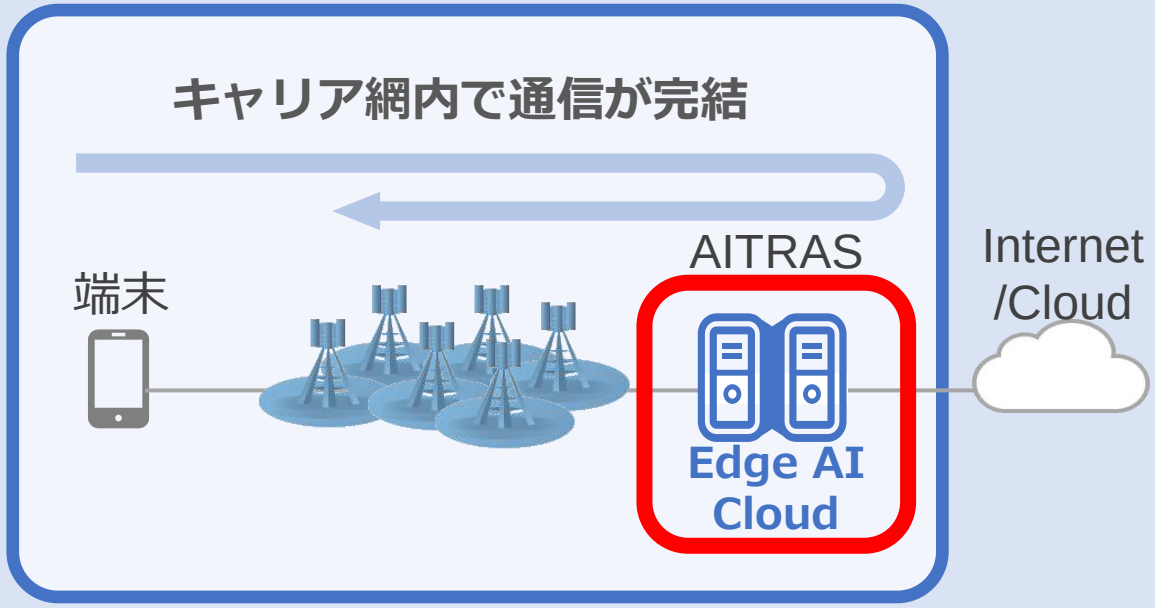
AITRASの要素技術



AITRASのEdge AIの特徴
～革新的なAIサービス実現をサポート～



低遅延 & 高セキュリティな
Edge AI Cloudを提供

NVIDIA AI Enterpriseと連携した
スピーディな開発の実現

キャリアだからこそできる
低遅延＆高セキュリティな Edge AI Cloud

AITRASのEdge AIの特徴
～革新的なAIサービス実現をサポート～

低遅延 & 高セキュリティな
Edge AI Cloudを提供

NVIDIA AI Enterpriseと連携した
スピーディな開発の実現

NVIDIA AI Enterprise と連携した スピーディな開発の実現

ビジネスロジック、企業データ

NVIDIA AI Enterprise

NIMS* Microservice | CUDA-x Microservice

AI Application Framework

End-to-End AI Development Tools

Infrastructure Management and Optimization

Edge AI Cloud

AITRAS上でスピーディな AIサービス開発が可能に

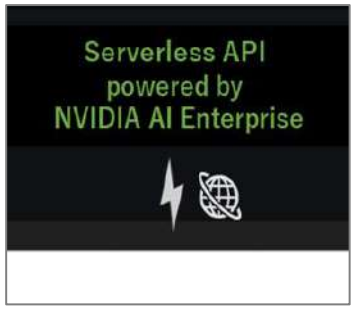
NVIDIA AI Enterpriseと連携 AI開発を強力に支援する機能群

低遅延＆高セキュリティ Edge AI Cloud

*NVIDIA Inferencing Microservices

AITRASのエッジAIユースケース

後ほどデモにて詳細をご説明



AITRAS

オーケストレーター

エッジAI

NVIDIA AI Enterprise

RAN L2/L3 ソフトウェア

RAN L1 ソフトウェア

NVIDIA AI Aerial

仮想化基盤 MIG/MPS

NVIDIA GH200

Arm Neoverse V2

Radio Unit

SoftBankオリジナル AI サービス

- ◆ 自動運転遠隔サポート
- ◆ Cloud LLMロボ
- ◆ RAG Menu@Edge

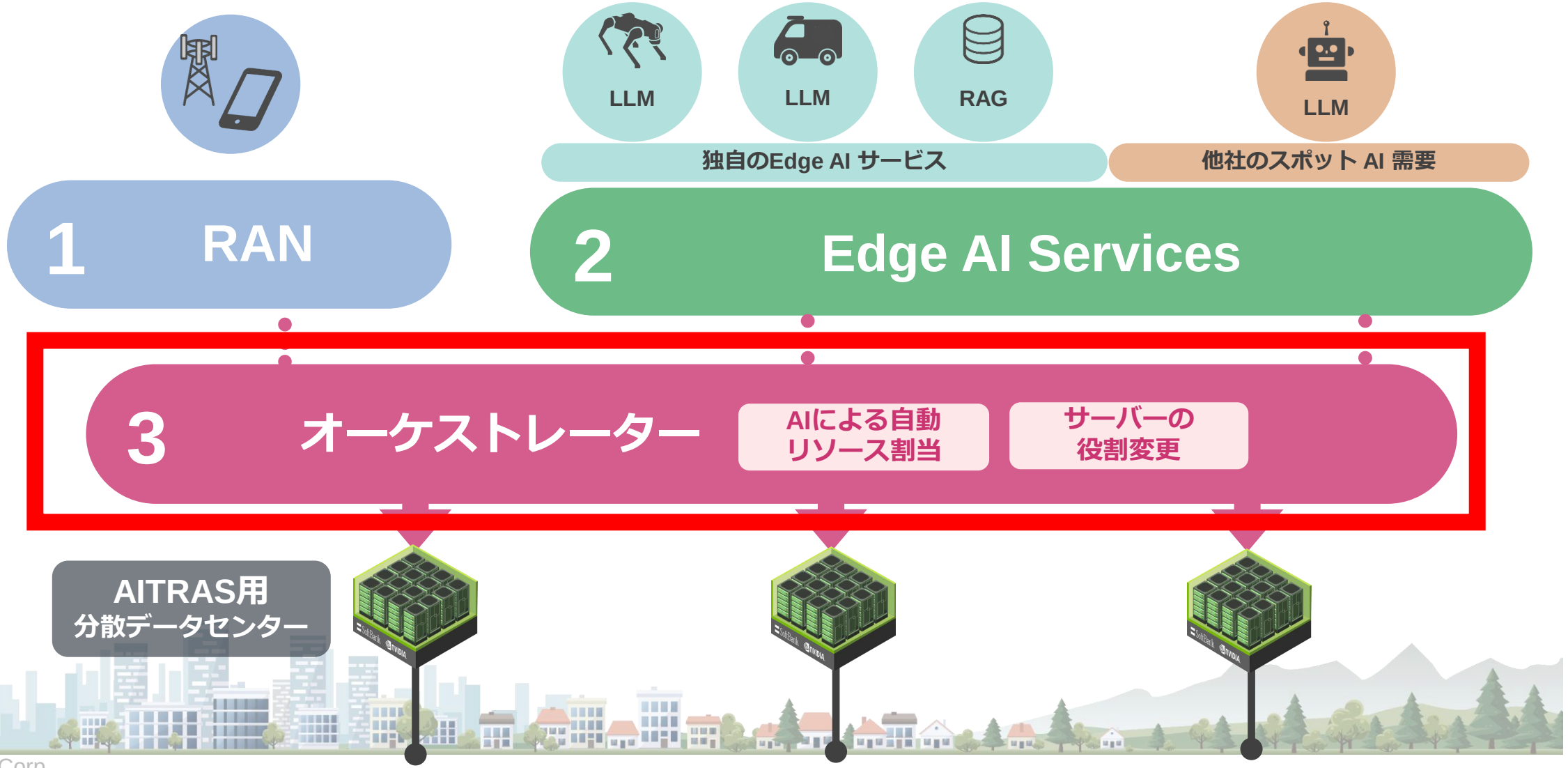
3rd Party向け Edge AI Cloud

- ◆ Serverless API Powered by NVIDIA AI Enterprise @Edge

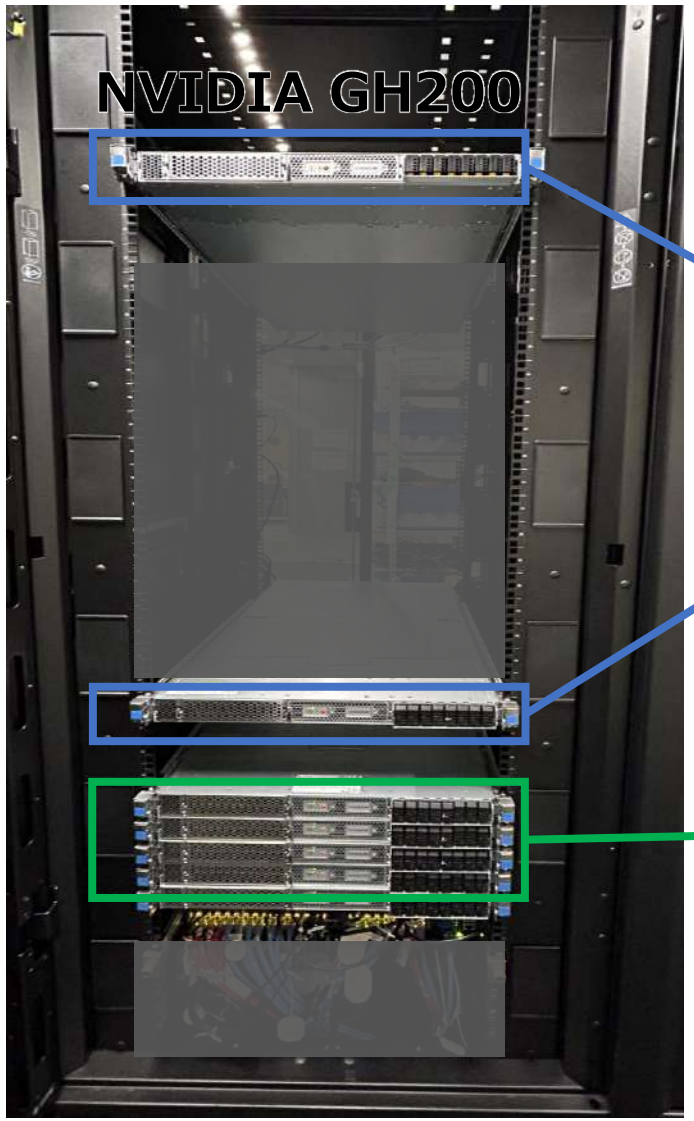
AI Models　OSS LLM　Partner LLM　○○○

NVIDIA AI Enterprise

AITRASの要素技術

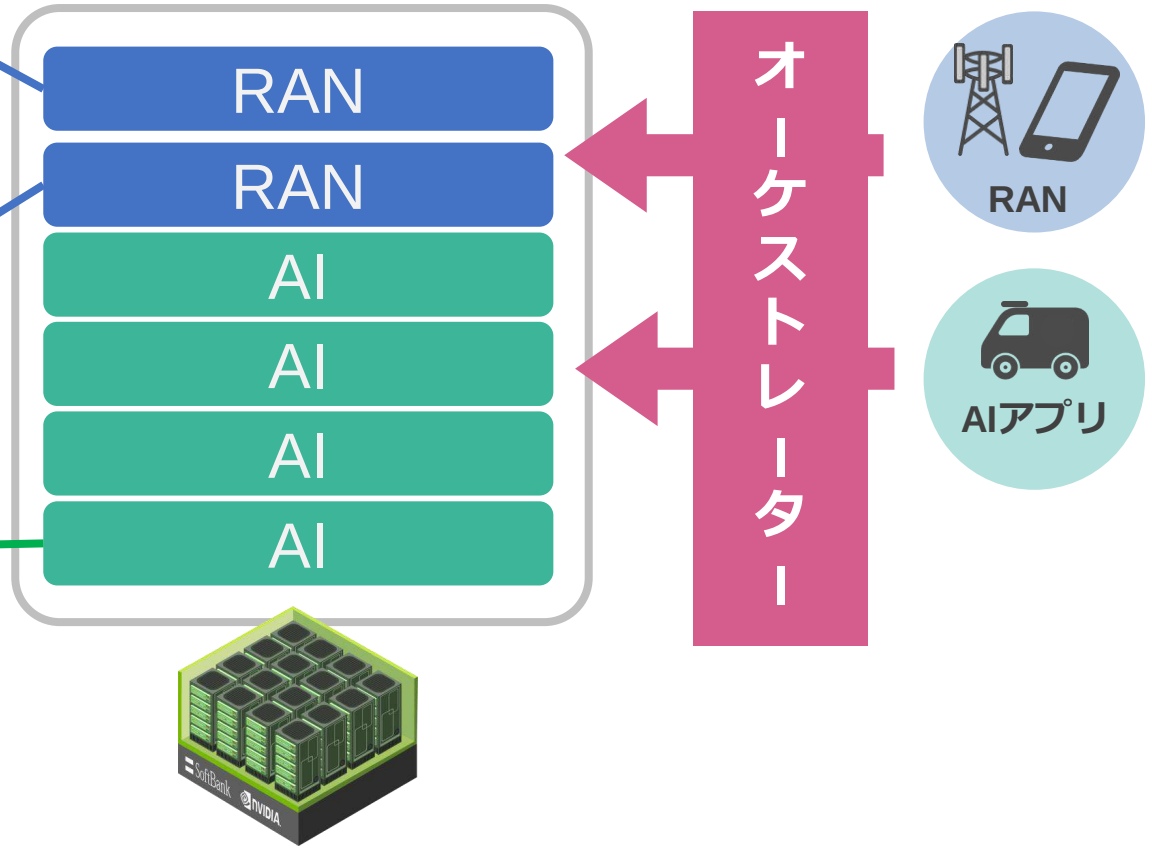


AITRAS オーケストレーター基礎的な機能



✓サーバーごとに“RAN”か“AI”の役割を切替

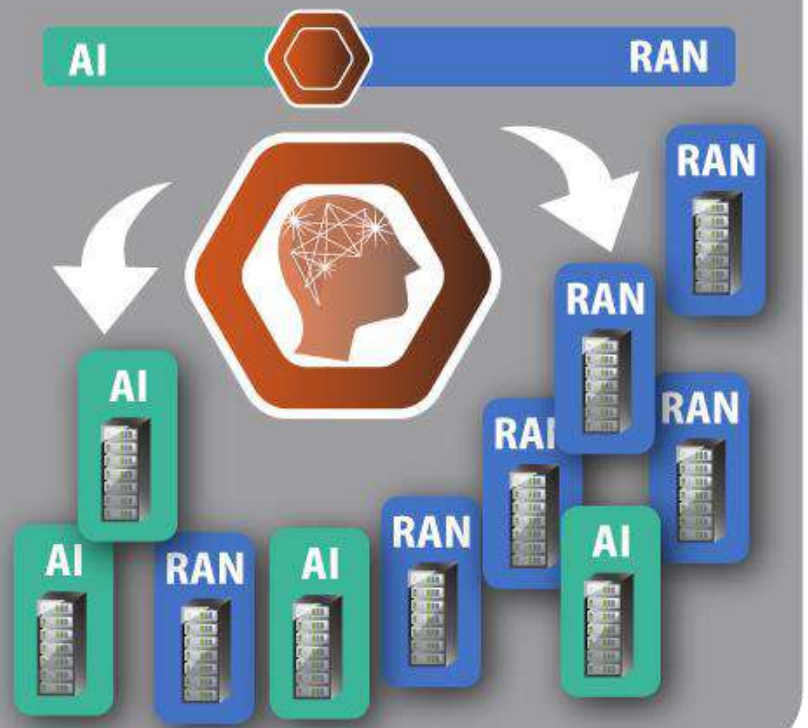
✓役割に応じたアプリをデプロイ



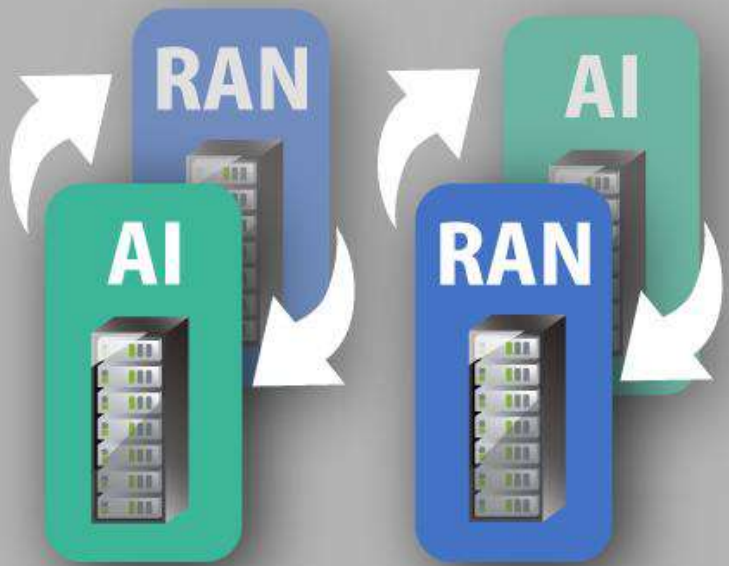
AITRAS オーケストレーターの大事な2つの機能

SoftBank

AIとRANの需要を最適なリソースに割り当て



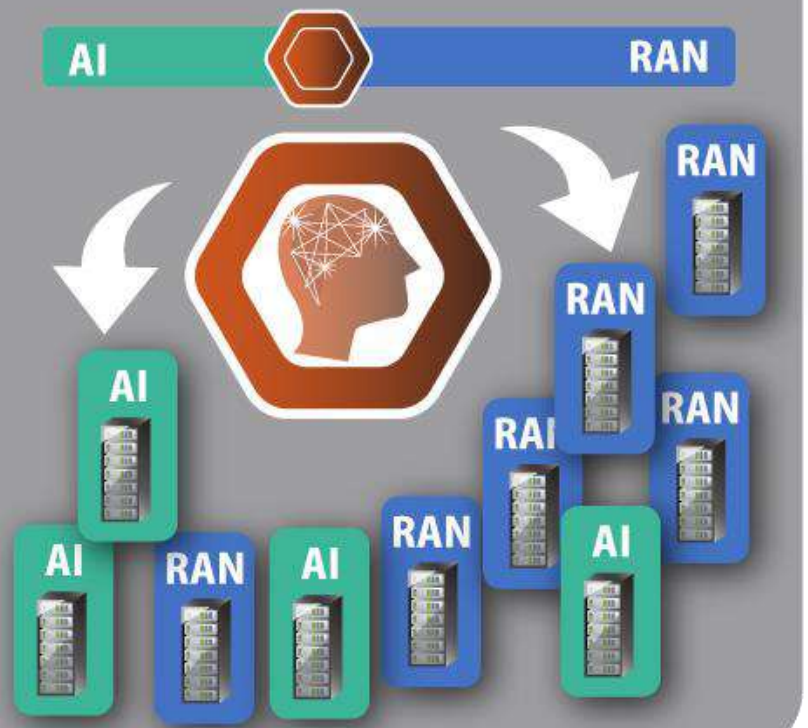
サーバーの役割の動的な変更



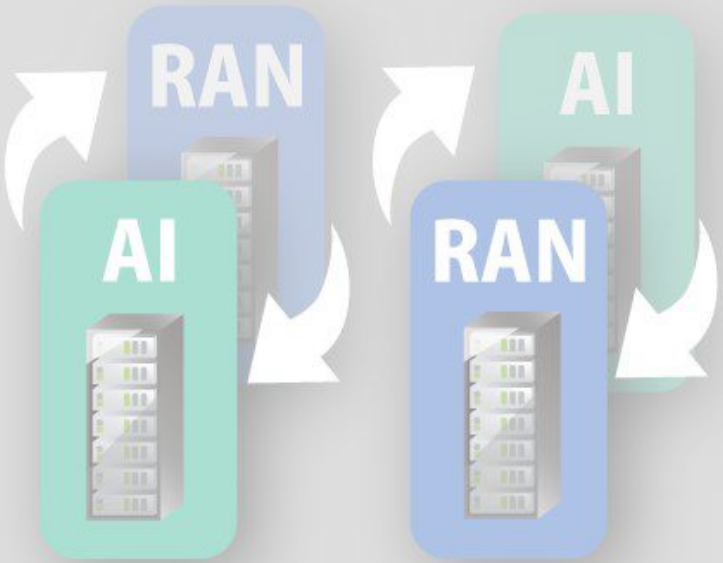
AITRAS オーケストレーターの大事な2つの機能

SoftBank

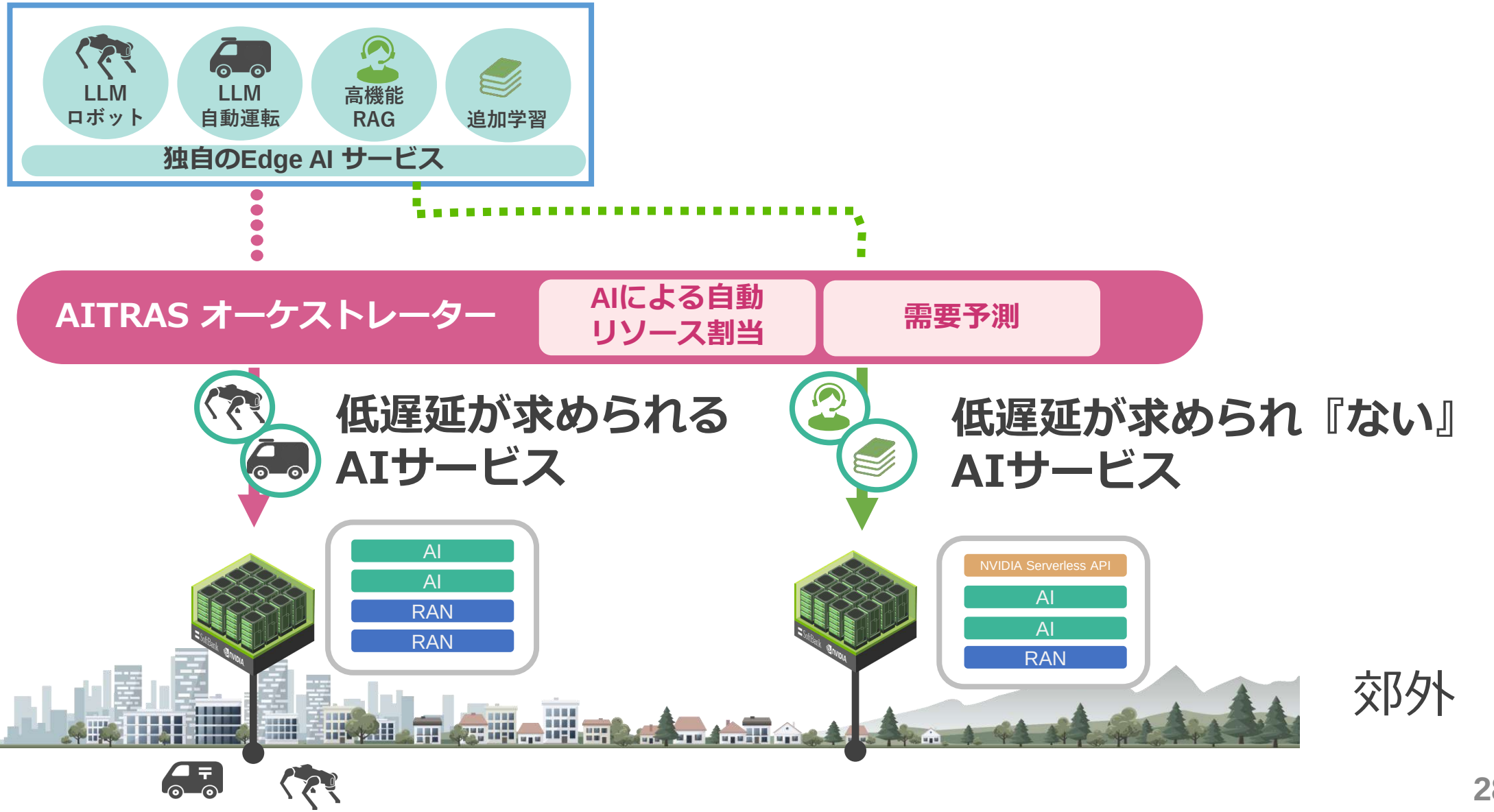
AIとRANの需要を最適なリソースに割り当て



サーバーの役割の動的な変更

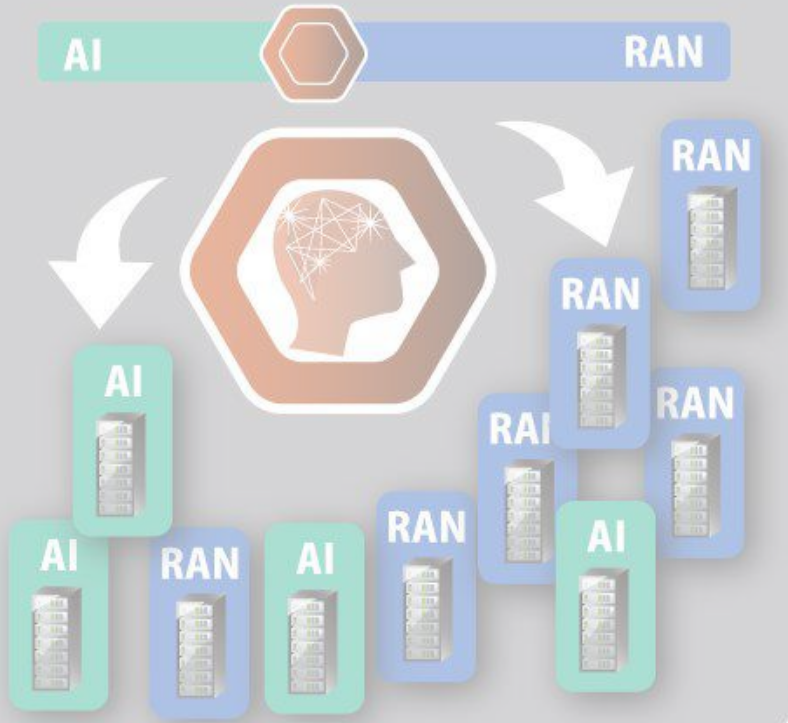


用途に応じてサービスを振り分け

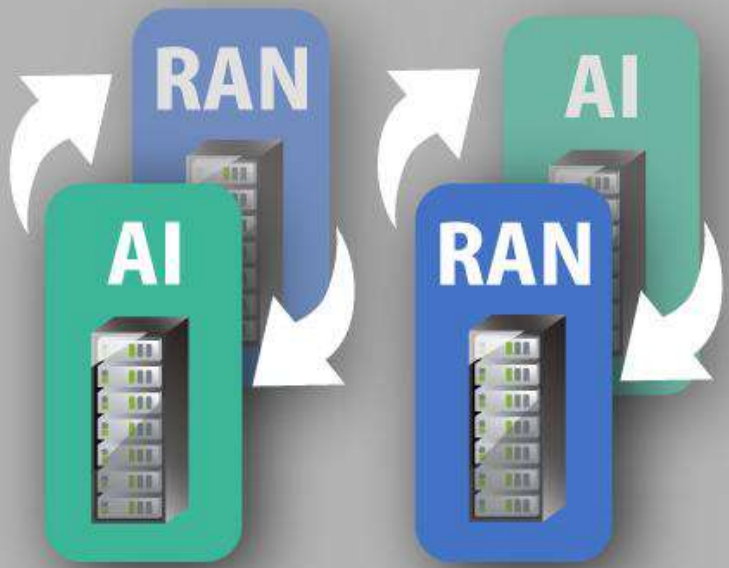


AITRAS オーケストレーターの大事な2つの機能

AIとRANの需要を最適なリソースに割り当て



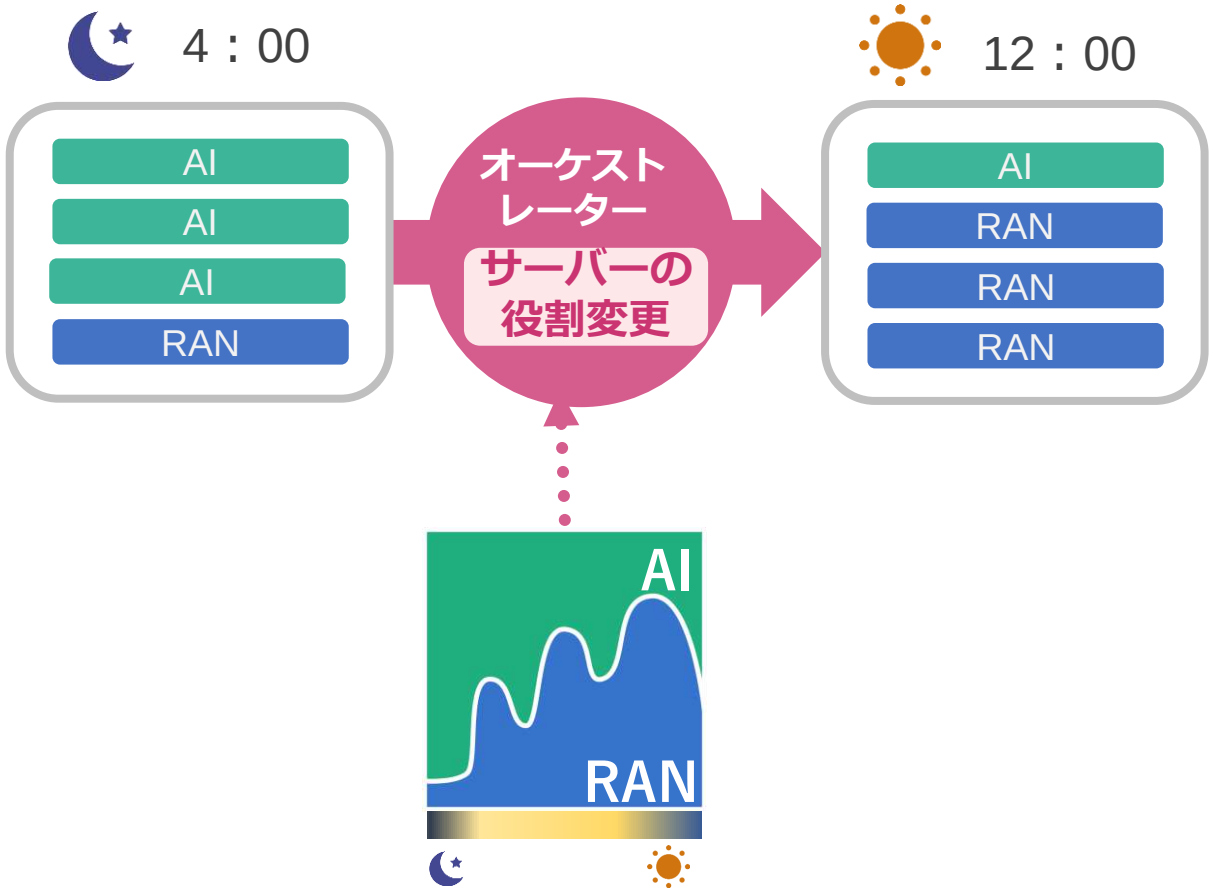
サーバーの役割の動的な変更



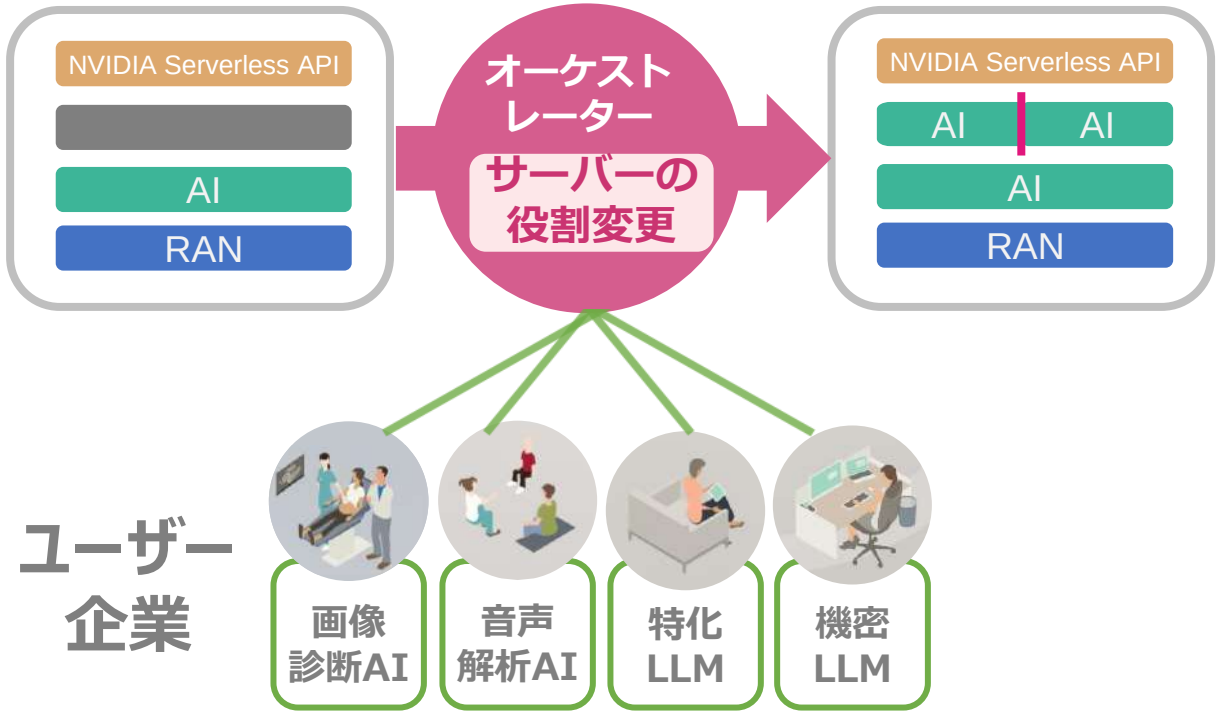
サーバの役割を最適に変える

SoftBank

RANの需要に合わせサーバーの役割変更
（AI⇔RAN）



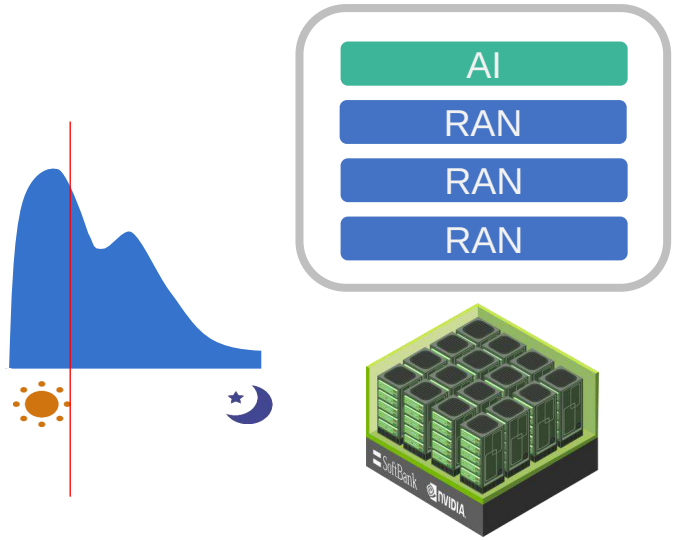
需要に応じてGPUを分割
(MIG)



デモ

デモシナリオ

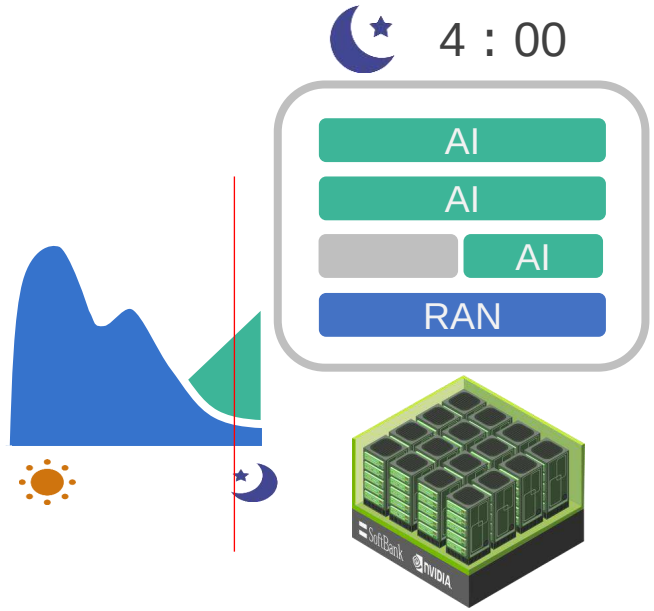
昼はRANが忙しい



サーバーの
役割変更

AIによる自動
リソース割当

深夜帯になりRANの需給の減少
サーバーを適切な役割に変更しAIアプリ実行

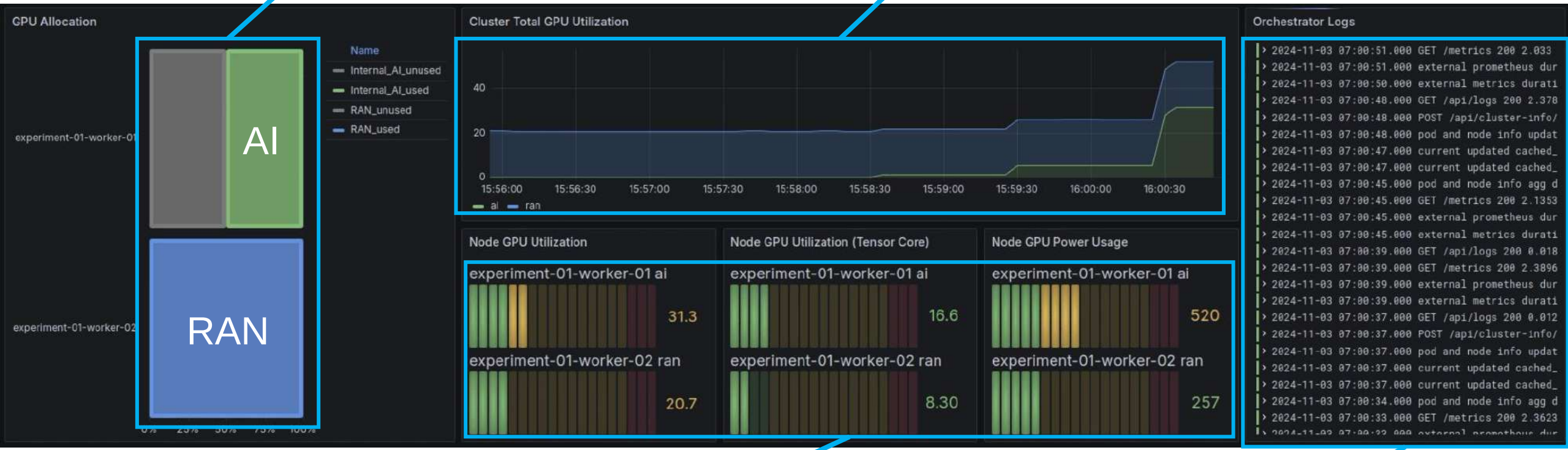


適切な大きさの
GPUを
AIに割り当て
&AIアプリ実行

デモ画面のご紹介

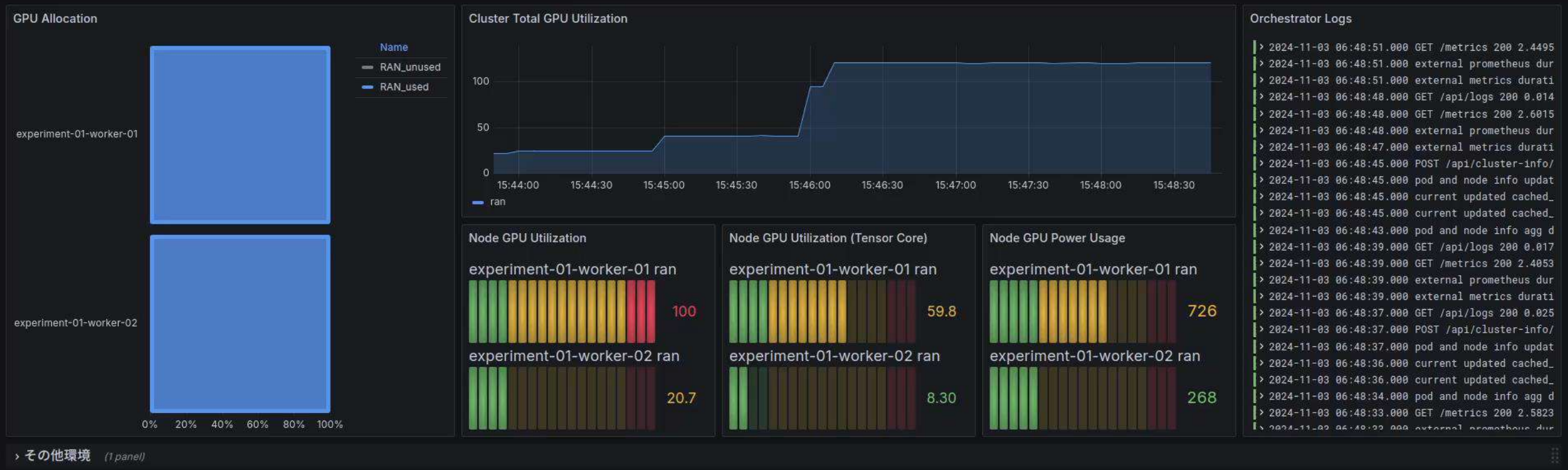
ノードの種別・GPU割当状況

GPU使用状況のグラフ



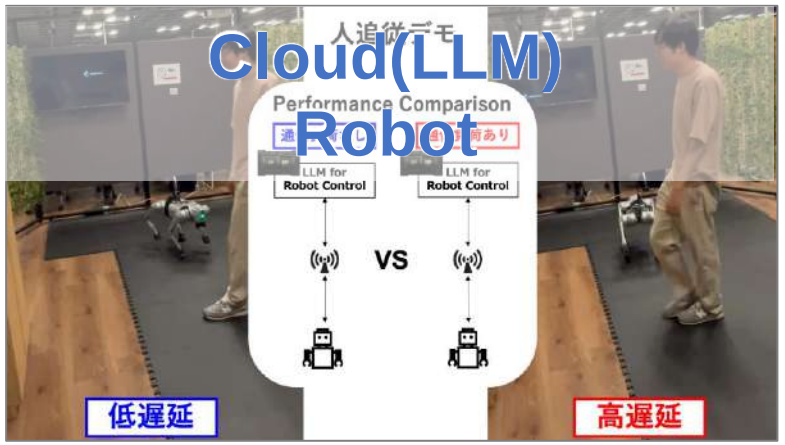
CUDAコア、Tensorコア、
ノードの消費電力の即時値

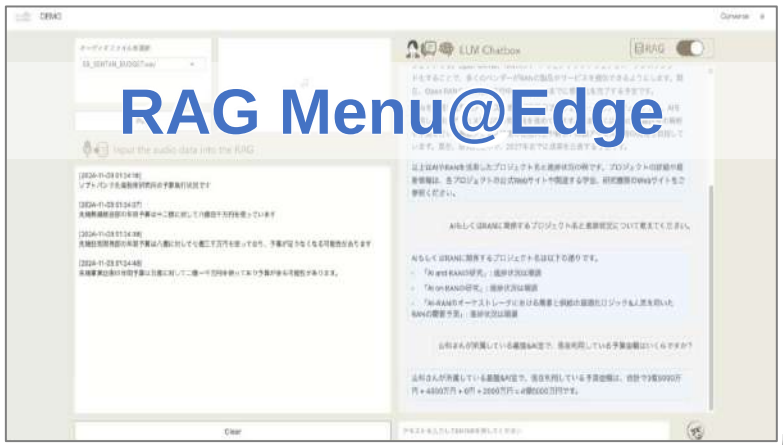
オーケストレーターの
操作ログ



AI-RANオーケストレーターが管理するクラスターに2つのノードが存在し、
RANのアプリケーションがそれぞれデプロイされています。

NVIDIA AI Enterprise@AITRAS を用いたユースケース





本日デモで
ご説明予定

Edge AI Services

- **AI App**（SoftBank）
- **AI App**（3rd Party）
- AI Models: OSS LLM / Partner LLM / ○○○
- NVIDIA AI Enterprise

Serverless API powered by NVIDIA AI Enterprise @edge