AITRAS

SoftBank AI-RAN Solution Overview

1. Executive Summary

AI is driving transformative changes across industries, enterprises, and user experiences, significantly enhancing convenience and productivity. Its applications are expanding beyond services like ChatGPT to real-time control of devices such as robots, drones, and vehicles. These diverse use cases share a critical demand for extensive, reliable, secure, and ultra-fast connectivity. To meet the needs of this new AI-driven traffic, telecommunication networks must evolve.

Telecom operators now face a strategic imperative to build infrastructures that not only support traditional operations but also enable AI training and inferencing. By integrating AI capabilities into their existing central and distributed infrastructures, operators can transform their networks into robust platforms for AI innovation.

AI-RAN (Artificial Intelligence Radio Access Network) offers a groundbreaking approach, providing an ideal environment for AI inferencing within mobile infrastructure. "AITRAS," SoftBank's AI-RAN solution, represents a GPU-based infrastructure that simultaneously handles both RAN and AI workloads. This integration effectively shifts mobile infrastructure from being a cost center to a revenue-generating asset.

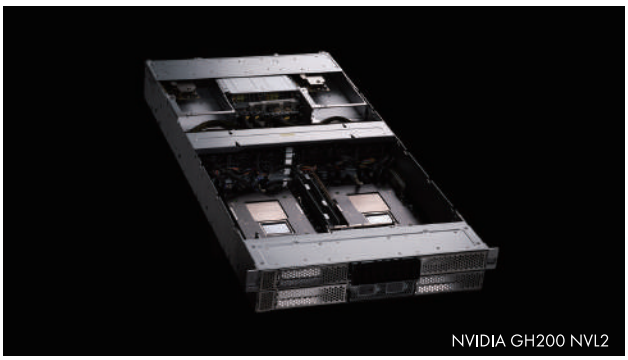
AITRAS enhances network efficiency, reduces operational costs, and introduces innovative business models for telecommunications. By leveraging this solution, telecom operators can not only optimize their infrastructure usage but also explore new revenue streams, positioning AI-RAN as a cornerstone of future network transformation.

2. Solution Overview

SoftBank's "AITRAS," based on the AI-RAN concept, offers a transformative opportunity to revolutionize telecommunication networks by integrating AI and RAN on a single infrastructure. This GPU-based unified infrastructure is designed to enable telecommunication operators to run RAN and AI workloads concurrently, optimizing resource utilization while reducing operational waste.
Key features of AITRAS as an integrated infrastructure include:

- **Multi-tenancy for AI and RAN with AI-based orchestration,** improving flexibility, cost efficiency, and resource utilization while minimizing waste.
- **Support for the development, deployment, and monetization of various AI applications,** allowing operators to expand their service offerings.
- **Carrier-grade RAN performance,** delivering high functionality, performance, and quality that meet the stringent requirements of traditional RAN systems.
- **AI-driven enhancements in energy efficiency and ROI,** significantly reducing operational costs while maintaining high performance levels.

SoftBank aims to roll out AITRAS globally from 2026 onward. To support early adoption, a reference kit will be available starting in 2025. This kit will provide telecom operators with the essential hardware and software components to evaluate the practicality and value of the AI-RAN concept, allowing them to seamlessly trial AITRAS and experience its potential firsthand.



3. Key Components of AITRAS

Physical System Components

AITRAS is built using key hardware components, including the NVIDIA GH200 Grace Hopper™ Superchip, Radio Units, and network switches.

Logical System Architecture

AITRAS operates on a GPU-based NVIDIA GH200 platform and consists of:

- **A virtualization platform**
- **RAN functions** structured with L1, L2, and L3 layers*.

  *L1/L2/L3 refer to the OSI reference model layers in RAN software architecture: "Physical Layer (Layer 1)," "Data Link Layer (Layer 2)," and "Network Layer (Layer 3)."
- **Edge AI,** supporting AI applications.
- **An orchestrator,** providing the necessary computational resources for both AI and RAN applications to function seamlessly.

Resource Management

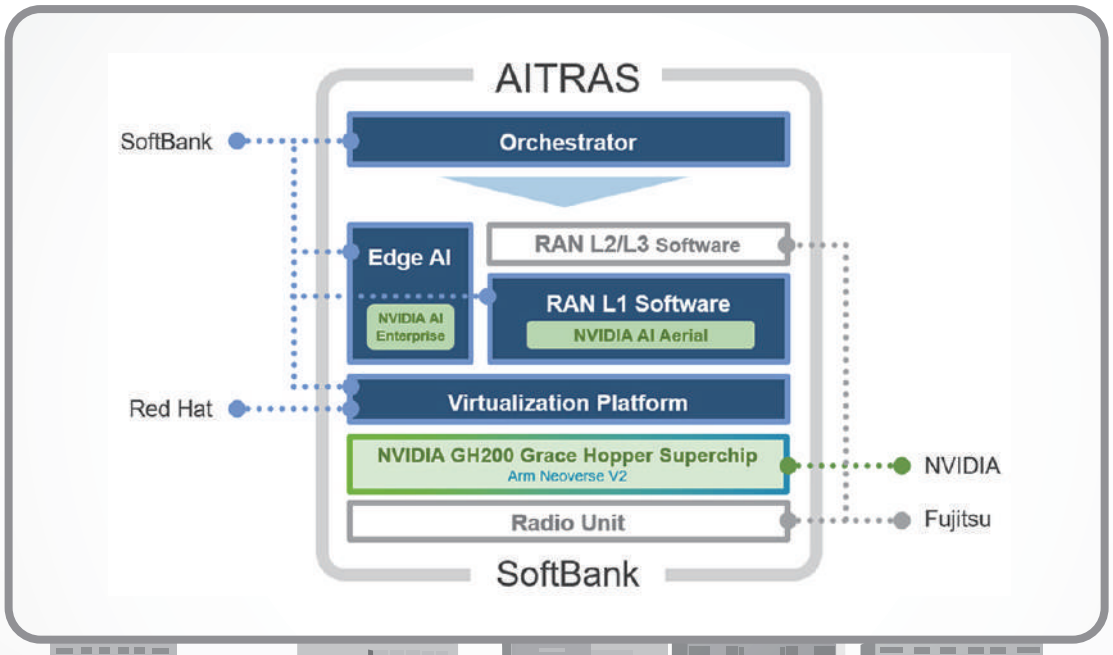
Even across multiple sites or servers, AI and RAN workloads, along with computing resources, are simultaneously managed through an optimized data flow mechanism. This enables AI-driven efficient resource control.

RAN Functions

To achieve high functionality, performance, and scalability, network functions are fully virtualized in software on NVIDIA GH200. This allows RAN software to be flexibly deployed and managed, making it easier to enhance and advance RAN capabilities.

AI and Machine Learning Models in RAN

AITRAS is designed to enhance the full-stack performance of RAN, including energy efficiency, through the application of various AI/ML models. These models enhance wireless signal processing tasks such as channel estimation, modulation, and error correction, significantly boosting network efficiency and performance.



4. Platform Architecture of AITRAS

AI-Enabled Orchestration Layer

AITRAS incorporates an orchestrator to manage AI and RAN workloads, ensuring efficient resource utilization. The orchestrator leverages AI-driven algorithms to dynamically allocate resources based on real-time demand, optimizing system performance while minimizing latency.

Multi-Access Edge Computing (MEC) Integration

By supporting AI applications with Edge AI, including MEC functionality, a seamless user experience is delivered. AITRAS processes data closer to the user, reducing latency and enhancing the performance of applications such as autonomous driving support and real-time video analytics.

Cloud-Native Design

AITRAS features Kubernetes-based containerized network functions optimized for both Virtual Network Functions and Cloud-Native Network Functions. This cloud-native design enables flexible deployment in various environments, ranging from distributed edge data centers to centralized cloud infrastructures.

Scalability and Flexibility

The modular architecture of AITRAS supports a variety of deployment scenarios, including distributed, centralized, and aggregated RAN configurations. This design allows telecom operators to scale their networks efficiently based on demand, offering enhanced flexibility in network planning and deployment.

5. AI-Driven RAN Management and Automation

Lifecycle Management of Network Functions (NFs)

The deployment, scaling, healing, and upgrading of AI-based network functions are fully automated through an AI-driven orchestration system. This enables the network to adapt dynamically to changes in demand, reducing operational complexity and minimizing the need for manual intervention.

Automated Provisioning

Zero-Touch Provisioning (ZTP) for RAN elements streamlines operations by minimizing manual efforts. ZTP facilitates the rapid deployment of network components, accelerating the rollout of new services and significantly reducing network maintenance time.

Predictive Maintenance

AITRAS employs AI-powered proactive monitoring and maintenance to ensure network reliability. Predictive maintenance leverages machine learning algorithms to identify potential issues before they escalate, reducing downtime and enhancing network availability.

Resource Allocation and Traffic Management

AITRAS enables real-time, dynamic resource allocation to optimize network performance. Multi-tenancy capabilities, powered by Multi-Instance GPU (MIG) technology, allow AI and RAN workloads to efficiently share resources. This maximizes infrastructure utilization and enhances overall network efficiency and performance.

6. Security and Compliance

Multi-Tenancy Environment

AITRAS ensures carrier-grade reliability by providing secure, isolated resource environments for each workload. It supports multi-tenancy by allocating independent GPU instances for each application, ensuring that AI and RAN workloads, as well as different AI applications, are processed independently without interference.

Data Privacy and Compliance

To protect data privacy and meet regulatory requirements, AITRAS includes built-in mechanisms such as encryption and access controls. These features safeguard sensitive data and ensure compliance with regulations like GDPR, providing a secure and compliant platform for operators.

7. Edge AI

Edge AI delivers low-latency and data-secure AI applications. Additionally, it leverages NVIDIA AI Enterprise to enable businesses and users to easily develop and deploy AI applications. Examples of implemented AI applications include the following:

Multi-Modal AI for Remote Autonomous Vehicle Support

Edge AI supports autonomous driving by transmitting data such as video from onboard cameras via 5G to a multi-modal AI running on Edge AI infrastructure. The AI performs real-time traffic analysis and risk assessment, providing recommendations to remote supervisors or directly to the vehicle through a chat interface.

Operational Efficiency with Edge RAG (Retrieval-Augmented Generation)

Businesses such as offices, factories, and construction sites can input company-specific data into Edge AI-based RAG systems via 5G. This allows for highly accurate search results tailored to company-specific information. Tasks unique to the business can be assigned to generative AI, automating processes like progress tracking and data visualization. Data sovereignty is ensured as all sensitive company data is stored locally on Edge AI rather than in a public cloud.

Real-Time Robotic Control

Video feeds from cameras mounted on robots are transmitted via 5G to an AI-based control system running on Edge AI. The robots respond to human commands and motions with low-latency, real-time actions. Compared to cloud-based systems, Edge AI significantly reduces response times, making it ideal for real-time robotic control in environments requiring instant reactions.

8. Benefits of Implementing AITRAS

Cost Reduction

By consolidating AI and RAN workloads onto the GPU-based AITRAS infrastructure, operators can eliminate the need for separate hardware, significantly reducing both capital and operational expenditures.

Optimized Resource Utilization and Flexibility

Through AI-driven orchestration, AITRAS dynamically allocates computing resources between AI and RAN workloads. This not only enhances infrastructure utilization but also enables fast and flexible service delivery, helping operators quickly adapt to changing market demands.

New Revenue Opportunities

Combining Edge AI with RAN allows telecom operators to develop new AI-based business models. This capability provides a competitive advantage, enabling operators to respond effectively to evolving market needs and create additional revenue streams.

Enhanced Carrier-Grade RAN Performance

AITRAS leverages NVIDIA GH200 Grace Hopper Superchip and SoftBank's custom L1 software, developed using NVIDIA AI Aerial, to deliver highly stable and high-performance carrier-grade RAN. This solution maximizes RAN capacity while reducing power consumption. Additionally, by integrating AI into C-RAN (Centralized RAN), AITRAS enhances performance across multiple cells, ensuring superior network quality and efficiency.

"AITRAS"
is a revolutionary solution that offers
telecommunications operators opportunities
to create new business ventures.
It also plays a crucial role in accelerating
the convergence of AI and telecommunications,
serving as an essential next-generation
social infrastructure to support
an "AI-coexistent society."

SoftBank