White Paper

AI-RAN：Telecom Infrastructure for the Age of AI

December 2024

Research Institute of Advanced Technology

SoftBank Corp.

SoftBank

Contents

Executive Summary

SoftBank's AI-RAN initiative aims to revolutionize the telecom industry by integrating Artificial Intelligence (AI) into Radio Access Network (RAN), transforming traditional networks from cost centers into intelligent, revenue-generating platforms. With mobile data traffic continuously growing, AI-RAN is expected to meet the dual challenges of rising infrastructure costs and intensifying market competition. This approach is expected to enable telecom operators to optimize network performance, reduce costs, and create new revenue streams through AI-enabled services.

AI-RAN may be implemented leveraging a software-defined, GPU-powered architecture called gRAN (GPU-based RAN). This advanced architecture supports high-performance network operations by utilizing the parallel processing power of GPUs. gRAN enables real-time data processing, intelligent resource management, and scalable multi-tenant operations. As the same platform supports both network and AI workloads, gRAN offers unparalleled flexibility, enabling seamless integration of RAN services and AI-native applications such as autonomous driving, real-time robotics, and edge computing.

SoftBank's AI-RAN product, AITRAS, exemplifies the convergence of AI and telecom infrastructure. AITRAS integrates RAN and AI workloads into a single, AI-native computing environment, offering carrier-grade RAN functionality with enhanced scalability and efficiency. The system supports multi-tenant operations, enabling network providers to run AI services alongside traditional network functions, creating new revenue opportunities. AITRAS is powered by NVIDIA GH200 Grace Hopper Superchip, which enable real-time AI inference and network management with optimal power efficiency.

Field and laboratory evaluations have confirmed AITRAS's ability to deliver carrier-grade stability, higher energy efficiency, and cost-efficient operations. In urban trials, the system successfully supported high-density traffic scenarios, while lab tests confirmed that its power consumption was comparable to that of current RAN systems, despite handling significantly higher workloads. This balance between performance and sustainability positions AI-RAN as a transformative force in telecom infrastructure.

To accelerate industry adoption, SoftBank played a leading role in establishing the AI-RAN Alliance in collaboration with major technology partners such as NVIDIA, Arm, Ericsson, Nokia, Samsung, and T-Mobile. This alliance is fostering innovation through collaborative research and development activities, advancing AI-RAN technologies while aligning with the global standards set by organizations like 3GPP and O-RAN Alliance.

SoftBank envisions a phased deployment roadmap for AITRAS, SoftBank's AI-RAN product, beginning with Over-the-Air pilot in a field area in 2024, followed by commercialization by 2026.

1. AI-RAN Vision: Shaping the Future of Telecom

The vision of SoftBank AI-RAN R&D is to revolutionize telecommunications by integrating AI into the core of RAN infrastructure, transforming traditional RAN into intelligent, adaptive, and revenue-generating platforms.

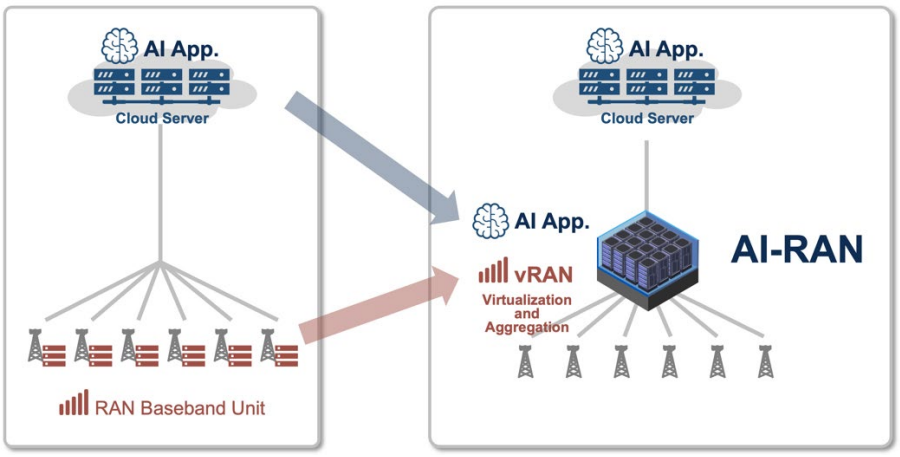

Figure 1. AI-RAN: AI and RAN integration

1.1 Telecom Challenges: Balancing Massive Capital Investments with ROI

The telecom industry is facing significant capital expenditure pressures due to rapidly evolving technologies and increasing data demands. The GSMA's The Mobile Economy 2024 report[1] reveals that in the global mobile market, total operator revenues are projected to grow from $1.11 trillion in 2023 to $1.25 trillion by 2030, representing a modest compound annual growth rate (CAGR) of 1.74%. However, total capital investments through 2030 are estimated at $1.5 trillion, exceeding total single-year revenues. This highlights a critical challenge faced by operators worldwide.

Foremost among these challenges is the substantial investment cost associated with 5G network deployment. New infrastructure requirements, such as the utilization of higher frequency bands and the mass deployment of MIMO antennas, necessitate significant funding. Additionally, the impact of increased traffic from generative AI applications like newly emerging Large Language Models (LLMs) on infrastructure must also be considered. The supply of equipment for this infrastructure is currently dependent on a few specific vendors, making it difficult to reduce costs and encourage commoditization. Additionally, the rapid proliferation of IoT devices and the growing popularity of high-definition video streaming necessitate continued network capacity expansion. Meanwhile, intense price competition in

[1] GSMA, The Mobile Economy 2024 Report: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/wp-content/uploads/2024/02/260224-The-Mobile-Economy-2024.pdf

telecom services makes it increasingly difficult to recover investments through traditional data service fee models. Furthermore, the emergence of Over-The-Top (OTT) providers who operate without their own communications infrastructure impacts telecom operators' profitability.

According to GSMA's The Mobile Economy 2024 report, despite discussions about a potential slowdown in growth, monthly global mobile data traffic per connection saw a significant increase from 10.2 GB in 2022 to 12.8 GB in 2023, representing the largest absolute growth since data tracking began in 2016. Looking forward, it is projected that total mobile data traffic will grow at an average annual rate of 23% between 2023 and 2030, and exceed 465 exabytes (EB) per month by the end of the decade. This network resource strain is forcing telecom operators to make substantial capital investments. Consequently, depending on their revenue models, operators face the risk of being unable to recover their increasing investment costs, presenting a critical management challenge.

Concurrently, price competition for telecom services has intensified, making it challenging to recoup investments through traditional data communication fee revenue models.

In this context, telecom operators are confronted with the challenge of improving investment efficiency. Specifically, they face two key issues: reducing infrastructure development costs and creating new revenue streams. This necessitates not only more efficient operation and greater cost reductions in network infrastructure, but also the development of value-added services and the establishment of new business models to secure additional revenue sources.

1.2 Opportunities: Transforming Network Infrastructure through AI-RAN

AI-RAN presents a unique opportunity to fundamentally transform network infrastructure, making it more adaptable, efficient, and capable of supporting new AI services. By leveraging AI, telecom operators can optimize network operations in real time, improve resource utilization, and introduce new revenue-generating opportunities.

One of the key opportunities offered by AI-RAN is its ability to shift from a static, hardware-dependent network architecture to a dynamic, AI and software-driven approach. AI allows for intelligent decision-making at the network edge, enabling real-time responses to traffic conditions, user demand, and service requirements. This level of adaptability ensures that networks will always operate at peak efficiency, provide better quality of service, and suppress energy consumption.

Furthermore, AI-RAN opens the door to new service offerings that were previously not feasible. For example, advanced network slicing, enabled by AI-driven resource management, allows operators to

create customized end-to-end virtual networks tailored to the specific needs of different customer segments, such as low-latency connections for gaming and LLM inferencing and high-reliability networks for enterprise mission critical applications. This ability to offer differentiated services not only enhances customer satisfaction but also creates new revenue streams for operators. New Edge AI inferencing services are also possible on the same AI-RAN infrastructure.

1.3 From Cost Center to Profit Center

AI-RAN is seen as a strong approach to enhancing the return on investment in network infrastructure for telecom operators. One of AI-RAN's key features, multi-tenancy, not only utilizes RAN resources for high-throughput broadband capacity, wireless quality improvement, and network optimization but also flexibly allocates resources for edge computing infrastructures that support AI training and inferencing. This multi-purpose capability enables operators to improve mobile network quality while creating new revenue opportunities.

By adopting AI-RAN, telecom operators can maximize the profitability of their network investments and establish sustainable growth models. This transformation converts traditional network infrastructure from a cost center into a profit center, enabling operators to achieve sustainable growth through new business models.

Figure 2. AI-RAN redefines telecom business

2. RAN Evolution: From dRAN, vRAN, Cloud RAN, Open RAN to AI-RAN

2.1 Key Developments in RAN Evolution

The RAN landscape has undergone a remarkable evolution, transitioning from traditional hardware-centric models to more advanced, AI and software-based architectures. This evolution can be characterized by the progression from Distributed RAN (dRAN) to Virtualized RAN (vRAN), Cloud RAN (C-

RAN), Open RAN, and ultimately AI-RAN.

2.1.1 Evolution from dRAN to vRAN, C-RAN, and Open RAN toward AI-RAN

dRAN represents the traditional RAN setup, where radio units, distributed across sites, are closely coupled with baseband units for signal processing. This setup often leads to increased costs including site assets, inefficient resource use, and delays in service evolution.

vRAN emerged as a response to these challenges by virtualizing baseband functions. With vRAN, network functions could be separated from dedicated hardware and deployed on commercial off-the-shelf (COTS) hardware, enhancing flexibility and scalability.

C-RAN further advanced this concept by centralizing baseband processing in a cloud environment. The centralized processing reduced hardware requirements at individual sites, allowing better pooling of resources and centralized management. It improved efficiency but required a robust backhaul to manage latency challenges.

Open RAN builds upon the virtualized and cloud-based approaches by introducing standardization and interoperability. It disaggregates RAN components, allowing operators to mix and match solutions from multiple vendors, breaking vendor lock-in, reducing costs, and encouraging innovation. This openness supports greater flexibility and adaptability in network deployments.

AI-RAN integrates AI capabilities into RAN operations over a common accelerated infrastructure and so represents the greatest advance. By providing AI and RAN, AI for RAN, and AI on RAN, operators can move beyond mere connectivity and make networks more intelligent, self-optimizing, and proactive.

2.1.2 Drivers of Transformation

The key drivers behind these transformations include cost optimization, performance optimization, improving flexibility with software, enhancing operational efficiency and capturing new monetization opportunities. Traditional RAN solutions require significant capital expenditure (CAPEX) for specialized hardware, while dRAN also faced scalability limitations and high operational costs. Moving to virtualized and cloud-based solutions addresses these challenges, allowing operators to minimize costs and fully utilize the scalability potential of the cloud infrastructure. Open RAN and AI-RAN take these benefits further by enabling flexibility through open interfaces, greater operational efficiency, and new monetization methods based on AI services.

2.2 AI-Native Networks: The Role of AI in RAN Transformation

AI is playing a transformative role in the evolution of RAN by providing advanced tools to optimize performance, automate resource allocation, and will ultimately transform how modern networks operate.

Improving Spectral Efficiency, Optimizing Performance and Resource Management: AI is fundamentally changing how RAN resources are managed by making operations more adaptive and efficient. In traditional RAN setups, resource allocation and management require manual configuration, making it hard to react to user demands. AI, however, enables a level of dynamic adaptability that was previously unachievable. For example, AI-native models can automatically allocate bandwidth based on real-time usage patterns, manage interference more effectively, and ensure optimal load balancing across the network. This improves spectral efficiency and optimizes performance and makes better use of the available infrastructure.

From Reactive to Predictive Models: One of the most significant contributions of AI to RAN is its ability to convert networks that merely react into those that can predict. Traditionally, network management responds to issues only after they occur. AI changes this paradigm as its predictive capabilities allow networks to anticipate problems and take preventive action. Machine learning algorithms can analyze vast amounts of network data to identify patterns and predict potential faults or congestion points before they impact service quality. This not only improves reliability but also helps minimize downtime and operational costs.

Unlocking Revenue Potential: AI-native networks are transforming RAN assets from traditional cost centers into revenue-generating centers. Generative AI introduces new user experiences by providing Edge AI inferencing and dynamic resource allocation, leading to better service quality and higher customer satisfaction. Generative AI also identifies opportunities for monetizing network capabilities, such as offering premium services, targeted advertising, and edge computing solutions. This shift not only maximizes the value of RAN investments but also positions networks as strategic assets driving profitability.

AI-RAN is thus at the forefront of a more proactive and efficient network management approach, transforming RAN into a key enabler of intelligent and autonomous network services.

2.3 AI-RAN Definitions

AI-RAN refers to the application of artificial intelligence technology to the RAN. It aims to improve mobile network efficiency and optimize power consumption, while enhancing the utilization of the existing infrastructure. The concept involves hosting both AI applications and virtual RAN (vRAN) software on the

same infrastructure, allowing telecom operators to generate revenue from both network access and AI services with a single capital investment.

The AI-RAN Alliance has established three items to address different aspects of AI integration in RAN:

AI-for-RAN: Focuses on using AI to enhance RAN performance. It explores how AI can improve operation efficiency, boost capacity, and achieve key performance targets in the radio access network.

AI-and-RAN: Investigates how to use the same infrastructure to run both RAN workloads and AI workloads simultaneously. The goal is to increase resource utilization and open up new revenue streams for telcos by hosting various AI applications on the same platforms that run network functions.

AI-on-RAN: Addresses solutions for running AI applications on the radio access network. It focuses on enhancing RAN to ensure it can handle the increasing demands of AI and generative AI applications without compromising key factors like latency and security.

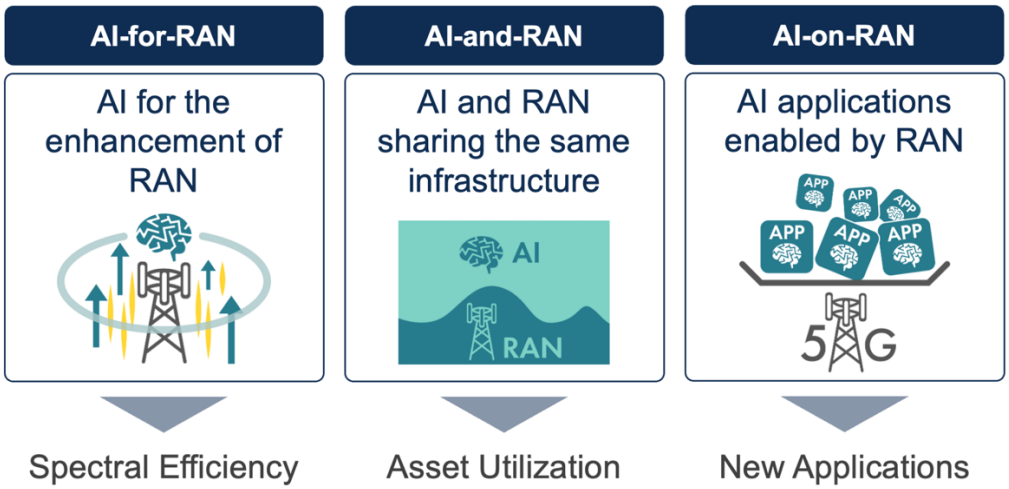

Figure 3. Three items to address different aspects of AI integration in RAN

These items collectively aim to integrate AI into the fabric of the radio access network, transforming networks into self-organizing, self-optimizing, and self-managing systems that can handle real-time changes, anticipate maintenance needs, and more efficiently manage resources.

3. History of SoftBank's AI-RAN R&D

3.1 Early Research and AI-RAN Development

SoftBank continues to explore new ways to create value by integrating traditional telecom infrastructure

with AI amidst the rapid innovation in AI technology. Recognizing that the full performance of 5G remains unrealized since its introduction, SoftBank began efforts to enhance 5G through AI and Machine Learning (ML).

SoftBank is leading the development of AI-RAN, a new architecture that integrates AI applications and software-based RAN into a single computer. AI-RAN enhances the capabilities and quality of RAN while also providing a shared computing platform for AI applications across various industries. SoftBank aims to deploy AI-RAN equipment in AI data centers distributed throughout Japan, directly connecting SoftBank base stations to these AI data centers to offer secure and low-latency AI services.

3.2 Applications of SoftBank AI-RAN Research

In the past, the primary strategy for achieving high-speed, high-capacity wireless communication was to increase the frequency bands used, as evidenced by the transition from 3G to LTE and 5G. However, the emergence of AI-RAN, which can enhance user experience without utilizing more frequency bands, holds great potential for effectively utilizing the finite public resource of the radio spectrum. With the transition from 5G to 6G networks approaching, the importance of AI-RAN is expected to strengthen even more.

The AI-RAN data center being developed by SoftBank will allow both "RAN operations" and "AI applications" to run simultaneously on the same server. This advancement enables telecom operators to secure two revenue streams—RAN and AI—with a single capital investment. Moreover, by integrating different services, operators can improve the operational efficiency of their infrastructure. Consequently, AI-RAN holds the potential to significantly improve the return on capital investment for telecom operators.

Case Study : Application of AI for channel interpolation in lower layers of wireless communication

In dense environments with multiple base stations and terminals, radio signals are often distorted by multipath fading. As a result, conventional signal processing technology may fail to accurately estimate wireless characteristics, leading to lower throughput.

To address this, we applied AI-native super-resolution technology, originally used in image analysis, to radio signal processing. Simulations were conducted to evaluate the potential uplink throughput improvements by reconstructing degraded signals using AI. After training the AI model with simulated radio signal data based on real-world environmental conditions and testing it with uplink signals, a 30% improvement in uplink throughput compared to conventional signal processing technology was observed (Figure 4).

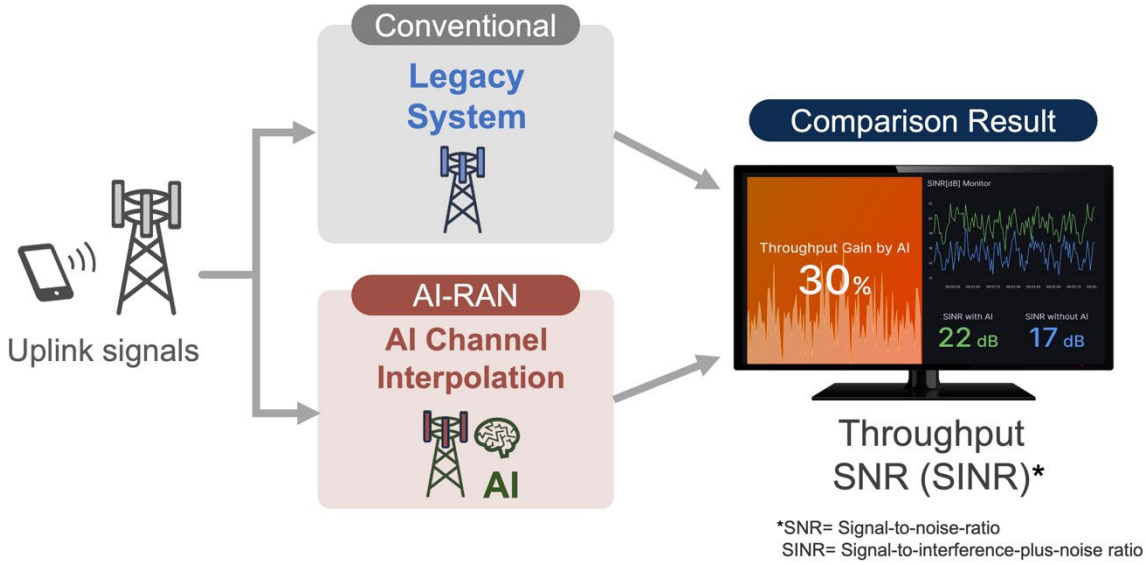

Figure 4. Comparison result of uplink signals: 30% throughput gain

3.3 Partnerships and Collaboration

Figure 5. AI-RAN Alliance launch ceremony at GSMA Mobile World Congress Barcelona 2024

SoftBank is accelerating the development of AI-RAN through its partnership with NVIDIA and other industry leaders, having begun the development of AI-RAN solutions on new hardware such as the NVIDIA Grace Hopper 200 Superchip (GH200), which is currently evolving into the NVIDIA Grace Blackwell platform.

To promote the widespread adoption and development of AI-RAN technology, SoftBank has partnered with industry leaders including NVIDIA, Arm, T-Mobile, Ericsson, Nokia, and Samsung to establish the AI-

RAN Alliance[2]. Since its launch at GSMA Mobile World Congress Barcelona 2024, the alliance has grown to 58 members (as of December 2024), encompassing a diverse mix of telecom operators, semiconductor companies, and academic institutions united by the mission of advancing RAN performance and capabilities through AI innovation.

SoftBank believes that AI-RAN has the potential to become the technology that will significantly impact not only the telecom industry but also society as a whole. With the imminent transition from 5G to 6G networks, the importance of AI-RAN is undeniable. SoftBank will continue to focus on continuing AI-RAN development to initiate a new paradigm shift for the 6G era.

SoftBank's AI-RAN R&D efforts have the potential to revolutionize telecom networks and create new business opportunities across various industries.

4. gRAN: GPU-based AI-RAN Architecture

gRAN, a term introduced by SoftBank, stands for GPU-based RAN that offers an architecture for deploying AI-RAN, considered the desirable evolutionary stage of RAN following vRAN, cRAN, and O-RAN. The introduction of gRAN marks a significant technological leap in the evolution of the radio access network. By leveraging the power of GPUs in addition to CPUs, gRAN enhances the efficiency, scalability, and flexibility of RAN infrastructures; it supports advanced AI-native functions and meets the ever-growing demands of AI applications and modern telecom networks.

4.1 Key Characteristics of gRAN

The transition from traditional RAN architectures to software-driven approaches with higher performance has paved the way for gRAN. As RAN becomes virtualized and open, and most importantly software-defined, it lays the foundation for porting RAN over GPU-based accelerated infrastructure, and bringing a new level of computational power and efficiency with gRAN.

4.1.1 Why GPUs for vRAN Evolution?

GPUs are well-suited for handling the highly parallel processing workloads common in modern RAN environments. Unlike CPUs, which are optimized for serial processing, GPUs excel at executing intensive matrix calculations and multiple tasks simultaneously, making them ideal for real-time RAN data and signal processing. This parallelism is particularly important in handling the complex algorithms required for 5G and future 6G technologies, such as massive MIMO, beamforming, and

[2] For more details about the AI-RAN Alliance's mission and initiatives, refer to their white paper at https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf

energy reduction. Utilizing GPUs in addition to CPUs allows telecom operators to handle these workloads more efficiently, reduce latency, and improve overall network performance and efficiency.

4.1.2 The Technological Leap with GPU-based vRAN (gRAN)

GPU-based RAN enables modern AI services like LLM inferencing by providing the computational power needed for real-time data processing and decision-making at the edge. This allows RAN to handle complex AI workloads efficiently, reducing latency and enhancing responsiveness. Additionally, GPUs facilitate dynamic resource optimization, such as advanced Self-Organizing Network (SON), by enabling rapid analysis and adaptation of network resources to changing demands, ensuring optimal performance and reliability in RAN environments. These attributes allow gRAN yield the more dynamic and responsive networks crucial for supporting emerging use cases such as Generative AI/LLM inferencing, augmented reality (AR), virtual reality (VR), and other data-intensive applications.

4.2 The Architecture of gRAN-based AI-RAN

The architecture of gRAN consists of several key components that work together to create a highly programmable, intelligent, and high performing network environment. The core elements of gRAN are the Radio Unit (RU), Distributed Unit (DU), Centralized Unit (CU), the integration of AI capabilities, and a multi-tenant and dynamic orchestrator.

Radio Unit (RU): The RU handles the radio frequency (RF) signals, converting them between analog and digital formats. It is responsible for communicating with user devices and serves as the mobile user's entry point to the RAN.

Distributed Unit (DU): The DU is responsible for lower-layer processing, including real-time tasks like scheduling, beamforming, and error correction. With AI integration, the DU can handle these complex tasks more efficiently, allowing for real-time optimization of network performance.

Centralized Unit (CU): The CU manages higher-layer processing, including packet scheduling, mobility management, and control plane functions. By centralizing these functions, the AI powered CU can provide better resource pooling and more efficient network management.

Integrated AI: The integration of AI capabilities in gRAN allows for intelligent resource management, predictive maintenance, and automation of network tasks. AI models can analyze network data in real time to optimize resource allocation, predict potential issues, and adjust network parameters dynamically.

AI Orchestrator for Multi-Tenancy and GPU Sharing: One of the key features of gRAN is its support for multi-tenancy, where multiple RAN functions and AI applications can run in parallel on the same GPU-based infrastructure. This is made possible by the immense parallel processing power of GPUs, which allows multiple network tasks to be executed concurrently without compromising performance. GPU sharing not only improves hardware utilization but also enables more efficient deployment of AI applications, leading to cost savings and enhanced operational efficiency.

gRAN bring significant advantages with its real-time parallel data processing, particularly for advanced RAN features such as massive MIMO, inter-cell coordination (e.g., 3GPP CoMP), multi-cell scheduling, precise positioning and many other advanced RAN resource optimization capabilities. Massive MIMO, which requires the handling of large numbers of antennas and processing data in parallel, benefits greatly from the GPU's parallel processing capabilities. Similarly, coordinated multipoint (e.g., CoMP) and precise positioning require real-time data exchange between multiple cells, and GPUs provide the computational power needed to manage these complex interactions effectively. The scalability of GPU systems ensures that the RAN can handle increased traffic demands, thus providing a more efficient and reliable network.

4.3 gRAN Case Study: NVIDIA AI Aerial

NVIDIA AI Aerial is a powerful accelerated computing platform, including hardware and software designed to support the development, simulation, and deployment of AI-RAN in 5G and future 6G networks. It includes several key components:

- **Aerial CUDA-Accelerated RAN** provides vRAN software to enable commercial-grade, software-defined, and cloud-native 5G RAN.
- **Aerial Omniverse Digital Twin** allows for physically accurate simulations of entire wireless systems, ranging from individual towers to large-scale city networks.
- **Aerial AI Radio Frameworks** offers integration with popular AI libraries and tools, enabling data generation, information capture, and large-scale training of AI and machine learning models for 5G and 6G research.

4.3.1 Converging vRAN and AI Edge Compute

As 5G and AI-native applications expand, the convergence of RAN and AI edge compute is becoming a critical technical evolution. NVIDIA's AI Aerial platform addresses these requirements by unifying the compute needs of both vRAN and AI edge workloads through a single system architecture.

vRAN Compute Requirements

Traditional vRAN compute for current and future generations of base station processing requires several key components:

- DSP (Digital Signal Processing) + CPU module to handle the baseband processing tasks.
- AI Processor module to integrate artificial intelligence for optimizing network performance.
- Software-defined module to ensure flexible and adaptable processing for dynamic network conditions.
- I/O (Input/Output) module to facilitate data communication and processing.

AI Edge Compute Requirements

AI edge compute systems also rely on a set of core components:

- CPU for general-purpose compute tasks.
- AI Processor to execute AI-native tasks at the network edge.
- Software-defined architecture for flexibility and dynamic control.
- I/O for managing data inputs and outputs.

Convergence for AI-RAN

In an AI-RAN converged architecture, these components are integrated into a unified system, combining the elements of both vRAN and AI edge compute:

- DSP + CPU to support baseband and general compute tasks.
- AI Processor for handling AI workloads.
- Software-defined architecture to dynamically optimize resource allocation.
- I/O to manage communication between components.

NVIDIA's solution elegantly maps these components onto their hardware platform, utilizing GPU, CPU, and NIC/DPU (Network Interface Card or Data Processing Unit) technology. These elements come together in the NVIDIA Grace Hopper and Grace Blackwell systems, designed to support the demands of both vRAN and AI edge compute in a highly efficient, software-defined manner.

The integration is powered by NVIDIA's AI Aerial platform, which provides the necessary software-defined framework to unify vRAN and AI edge compute, making it an ideal solution for network operators looking to harness the power of AI at the edge and optimize their RAN performance.

4.3.2 NVIDIA's AI Aerial Platform

NVIDIA AI Aerial leverages the parallel processing capabilities of GPUs, which are highly effective at managing the demanding workloads found in vRAN environments. These GPUs are ideal for

processing complex tasks such as massive MIMO, higher order modulation and coding and real-time inter-cell interference management, which require extensive computational resources. By utilizing NVIDIA AI Aerial, we can achieve enhanced scalability and agility in their RAN operations.

5. Introduction of AITRAS by SoftBank

AITRAS, SoftBank's AI-RAN product based on the gRAN architecture, offers a transformative opportunity to revolutionize telecommunication networks by integrating AI and RAN workloads on a single infrastructure. Its design enables telecom operators to run RAN and AI workloads concurrently, optimizing resource utilization while also bringing AI enhancements to the RAN.

5.1 Key Features of AITRAS

- Multi-tenancy for AI-and-RAN with AI-native orchestration, improving flexibility, cost efficiency, and resource utilization.
- 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 performance standards of traditional RAN systems.
- AI-driven enhancements in spectral efficiency and energy efficiency, significantly reducing operational costs and improving return on investment (ROI) while supporting higher performance levels.

5.2 Key Components of AITRAS

Physical System Components

AITRAS is built using advanced 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 on L1, L2, and L3 layers[3].
- Edge AI supporting AI applications.
- The orchestrator dynamically allocates the necessary computational resources for both AI and RAN applications so they function seamlessly.

[3] 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)"

Resource Management

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

RAN Functions

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

AI and Machine Learning Models in RAN

AITRAS is designed to enhance the full-stack performance of RAN, including spectral efficiency and 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.

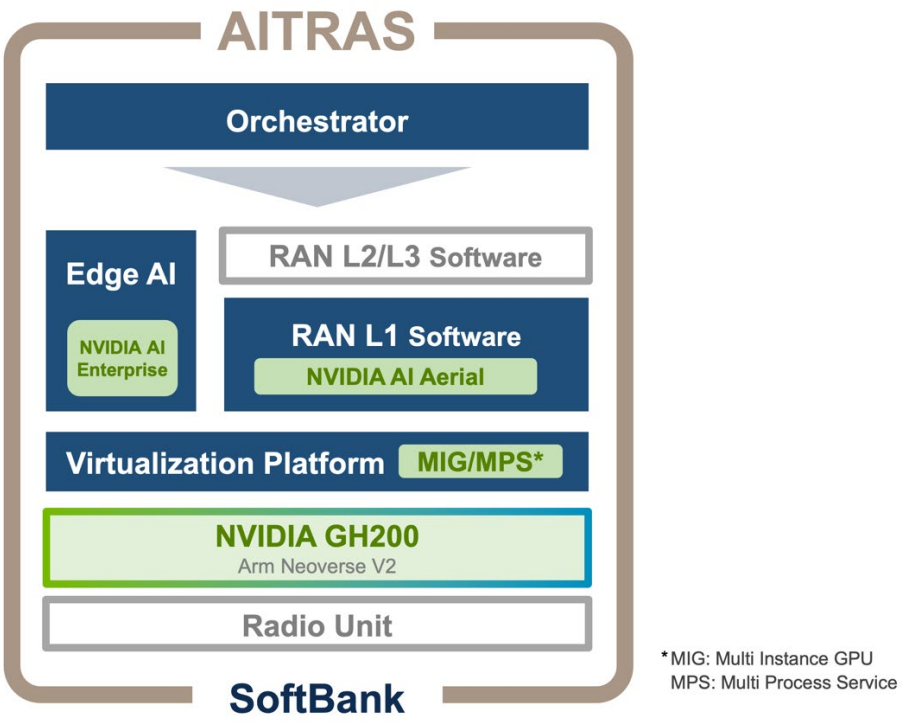

Figure 6. Components of AITRAS

5.3 AI-Native Orchestration

Orchestration Layer

The orchestrator is a central control mechanism for managing and optimizing RAN and AI workloads across the RAN and Edge AI services.

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

Multi-tenancy capabilities, powered by orchestration and GPU sharing technologies, allow AI and RAN workloads to efficiently share resources. This maximizes infrastructure utilization and enhances overall network efficiency and performance. Its key functionalities include:

- **Automatic Resource Allocation by AI**
  - The orchestrator dynamically assigns computing resources to meet varying RAN and AI workload demands.
  - It ensures efficient utilization of underlying computing resources by adapting to the needs of specific applications, such as RAN and AI (e.g., LLM Robots, LLM Autonomous Driving, and Advanced RAG (Retrieval-Augmented Generation) or 3rd party AI applications).

- **Dynamic Changes of Server Roles**

  The orchestrator allows servers to transition between roles based on workloads. For instance:
  - Switching between RAN servers and AI application servers.
  - Supporting different types of AI services including Edge AI services and NVIDIA serverless APIs.

- **Integration with NVIDIA AI Platform**
  - The orchestrator leverages NVIDIA serverless API and NVIDIA AI Enterprise software suite, enabling seamless operation of AI applications.
  - It supports spot AI demands from third parties, meaning it can provision services dynamically for external users requiring AI services.

- **Edge AI Services Management**
  - Acts as the backbone for managing and deploying Edge AI services.
  - Facilitates deployment of advanced services, such as LLM robots and autonomous driving solutions, while seamlessly maintaining RAN workloads.

- **AI-Driven Adaptability**
  - Continuously monitors system performance and adapts AI processes for optimal efficiency.
  - Ensures low latency, high throughput, and high availability.

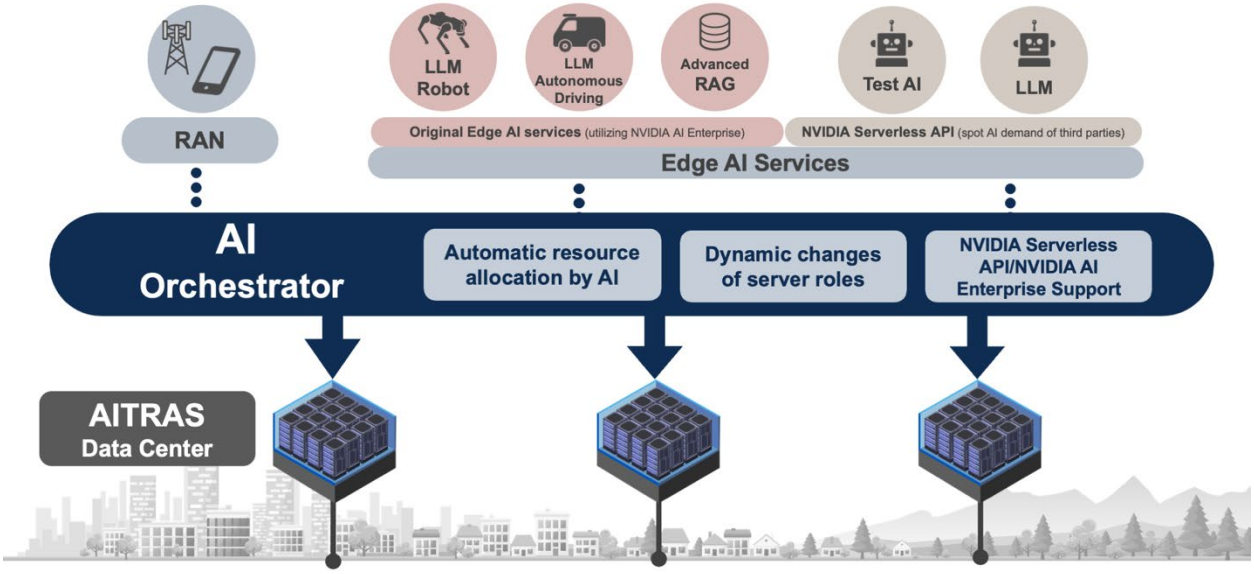

Figure 7. AITRAS orchestrator overview

5.4 Edge AI

SoftBank Edge AI solution delivers low-latency and data-secure AI applications over a 5G network with the additional benefits of guaranteed QoS. It integrates access to NVIDIA AI Enterprise software suite 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-native 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.

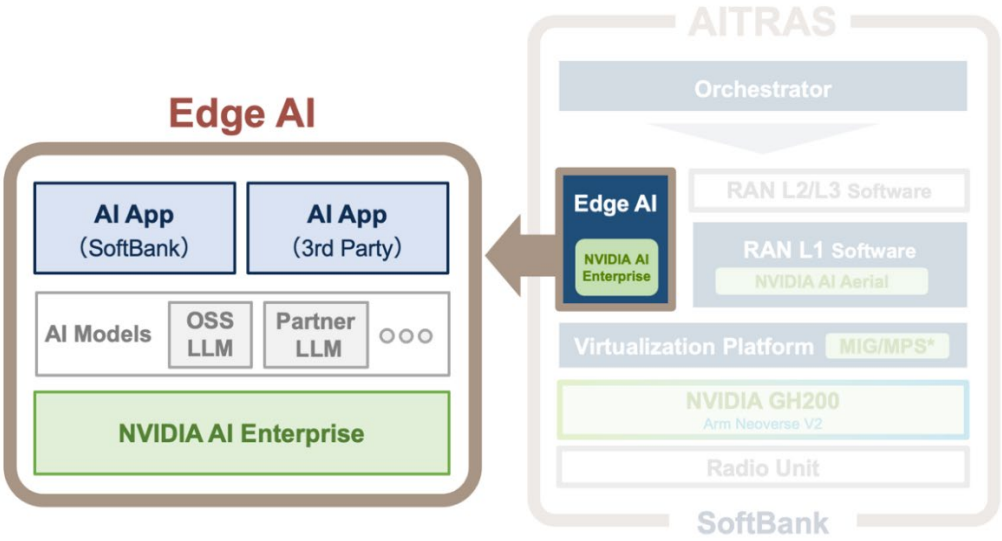

Figure 8. Edge AI in AITRAS

5.5 Key Benefits of AITRAS

Cost Reduction

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

Optimized Resource Utilization and Flexibility

Through AI-native 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 telecom operators quickly adapt to changing market demands.

New Revenue Opportunities

Combining Edge AI with RAN allows telecom operators to develop new 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 product maximizes RAN capacity while reducing power consumption. Additionally, by integrating AI into C-RAN (Cloud RAN), AITRAS enhances performance across multiple cells, ensuring superior network quality and efficiency.

SoftBank aims to roll out AITRAS across its own commercial network and globally for telecom operators from 2026. 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 SoftBank's AITRAS, allowing them to seamlessly trial AITRAS and experience its potential firsthand.

6. AITRAS Evaluation

6.1 Outdoor Testbed for AITRAS

SoftBank established an outdoor testbed to trial AITRAS in a real-world environment in Kanagawa Prefecture in Japan. The objective of this trial was to evaluate the performance of AITRAS under realistic network conditions and validate its potential to enhance network performance through AI-native automation and optimization.

Background and Objectives: The primary goal of SoftBank's AITRAS trial in Kanagawa was to assess stability as carrier grade in realistic environment. Carrier-grade stability refers to the ability of a cell to continue operating stably over a long period without shutting down. It is necessary to verify this stability by applying various loads to the cell. For this purpose, a high-interference environment, similar to an urban area, was constructed with five cells, each with 4 antennas, spread evenly over a 100 meter distance. The secondary goal was to evaluate AI-native network optimization in a real-world environment. Currently, evaluations are being conducted towards the primary goal.

Testbed Setup: The testbed involved integrating SoftBank's AITRAS with the existing network infrastructure. The NVIDIA GH200 Grace Hopper Superchip processes 20 5G cells on one server equipped as DU, supporting up to 4-layer MIMO with a bandwidth of 100MHz. Frequency band was 4.8GHz to 4.9GHz (n79).

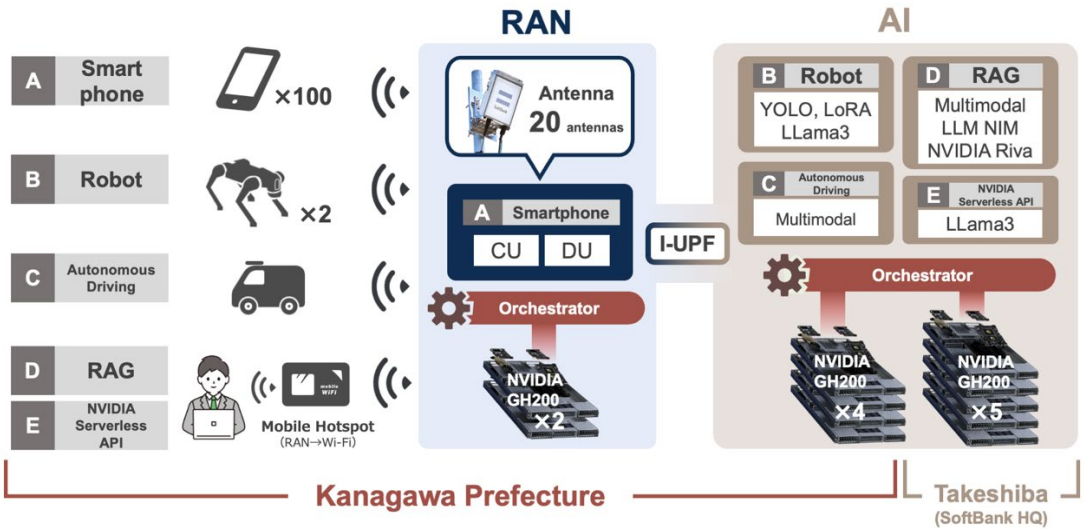

Figure 9. System architecture overview of AITRAS outdoor test

Trial Results - Stability Evaluation: The AITRAS trial in Kanagawa achieved carrier-grade stability. To demonstrate this, we conducted a test with 100 User Equipment (UEs) streaming videos simultaneously. Five UEs were placed at each of 20 locations; the results confirmed stable operation of the base stations under the simultaneous video streaming. This demonstration highlighted stability, and we are employing additional methods to further verify performance.

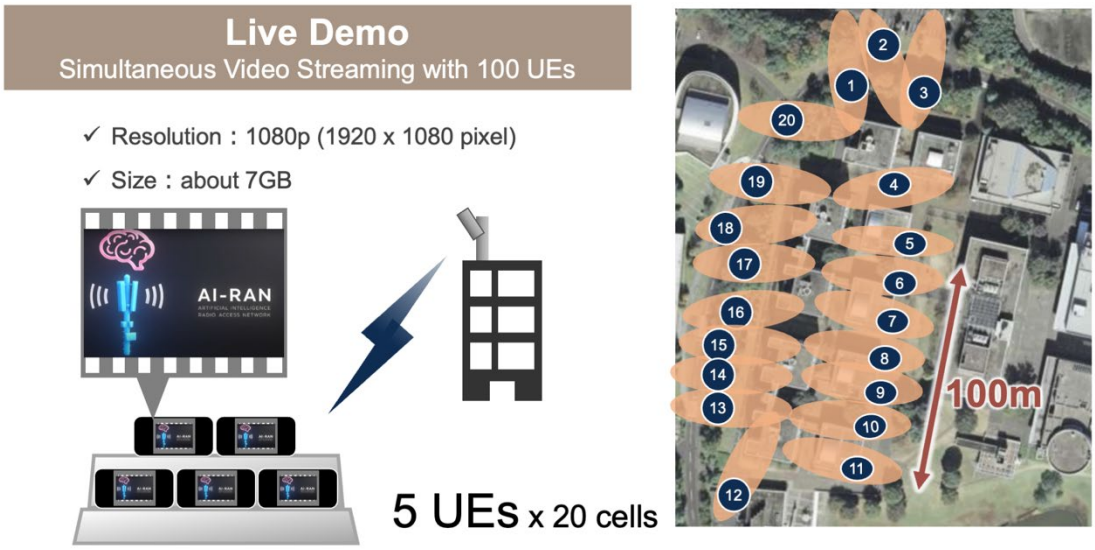

Figure 10. Live demo of simultaneous video streaming with 100 UEs

During the demonstration, we targeted key server performance metrics, particularly the fluctuations in

GPU usage. In the context of AI and RAN, GPU usage shifts according to RAN load changes (Figure 11). When RAN demands are low, the GPU can be repurposed for AI workloads. Gathering real-world data on traffic patterns and GPU usage is invaluable, as it provides essential insights on how the orchestrator can manage this switching effectively. This metric captures the entire process, where 100 devices stream videos simultaneously and then disconnect gradually over time.

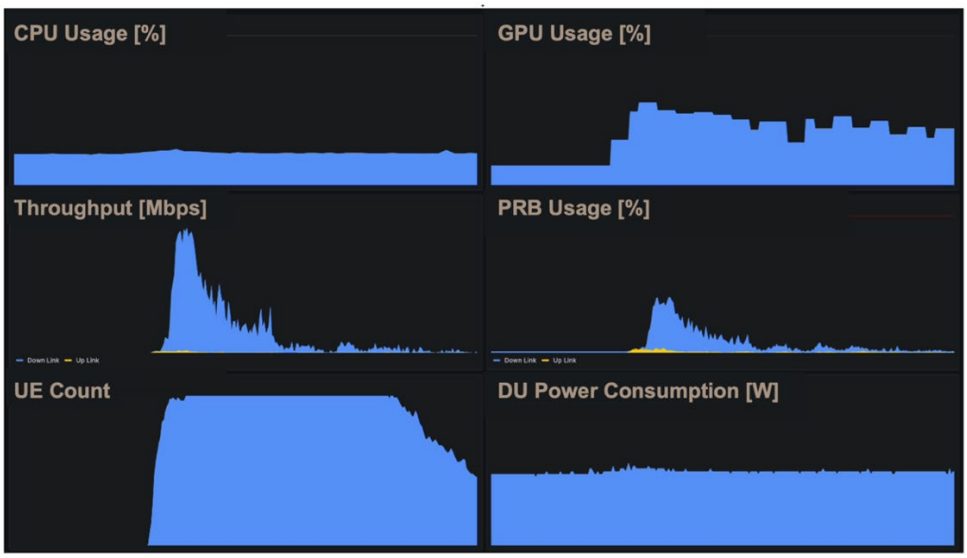

Figure 11. Resource monitoring of simultaneous video streaming with 100 UEs (traffic load and computational resources)

6.2 AITRAS Performance Evaluation

AITRAS is being evaluated out both outdoors and in a lab setting. One key metric of interest is the server's power consumption. While GPUs are often associated with high power usage, our lab tests revealed that processing 20 cells consumed around 500W, or roughly 25W per cell. This is comparable to current RANs and provides valuable insight when assessing total cost of ownership (TCO). A significant factor behind this low power consumption is the NVIDIA GH200, which combines the NVIDIA Hopper architecture and the Arm Neoverse V2-based NVIDIA Grace CPU in a single superchip. According to our tests, it uses about half the power of other companies' processors.

Additionally, lab evaluations are being conducted from the following perspectives as well:

- **UE Accumulation Evaluation**: By incrementally increasing the number of UEs, the load on the Control Plane (C-Plane) can be increased to verify the stability of the system. The rate at which UEs are added per unit of time also has a significant impact, making it an important evaluation parameter.
- **Radio Resource Occupation Evaluation**: By steadily occupying more of the available radio resources in the system, we apply load to verify the system's stability. Evaluation parameters include

scenarios such as downlink (DL) only, uplink (UL) only, both DL and UL, as well as the number of UEs and their mobility speeds.

6.3 SoftBank's L1 Enhancements in AITRAS

AITRAS's Layer 1 (L1) software, developed by SoftBank using NVIDIA AI Aerial, delivers high stability and high performance essential for a carrier-grade RAN through parallel signal processing and the optimization of task initiation timing. SoftBank will develop and implement further L1 enhancements by using AI to achieve maximized RAN capacity and reduced power consumption. In the future, SoftBank will implement not only L1 but also L2 / L3 enhancements by using AI.

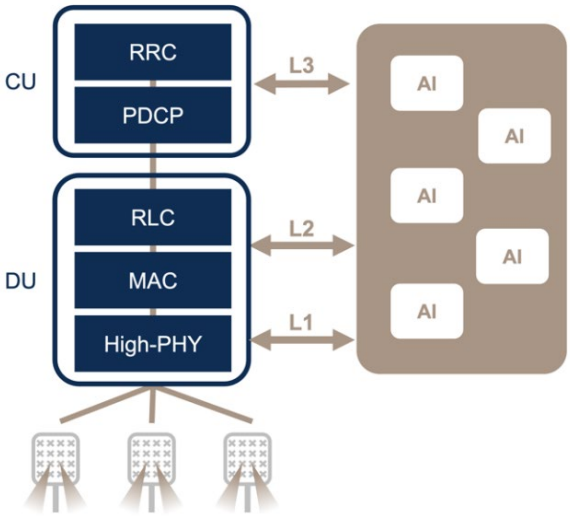

Figure 12. Full-layer optimization

Overall RAN Performance Improvements: The adoption of AITRAS will enable considerable improvements in SoftBank's network performance. By using GPU acceleration, SoftBank is able to achieve, simultaneously, higher processing speeds, reduced latency, and improved handling of multiple data streams.

Key Improvements in RAN Layer 1 Performance and Efficiency: AI-native L1 optimization has contributed significantly to improvements in key metrics such as cell capacity, throughput, and power consumption. By utilizing GPU capabilities to accelerate L1 processing, SoftBank's AITRAS is able to handle more users per cell without compromising performance. Additionally, throughput improvements have been achieved through more effective signal processing and real-time optimization of beamforming, while AI models have contributed to reducing power consumption by intelligently managing resources based on real-time network demand.

How gRAN-based RAN Outperforms Traditional RAN Architectures in Handling AI-for-RAN Workloads: Traditional RAN architectures rely on CPU-based processing, which is inefficient for handling the parallel processing requirements of AI workloads. In contrast, the gRAN architecture of AITRAS is capable of executing multiple AI algorithms simultaneously, making it well-suited for handling the complex workloads involved in AI-native RAN. The ability to run numerous operations concurrently allows AITRAS to achieve lower latency, improved energy efficiency, and enhanced scalability compared to traditional RAN architectures, thereby providing a superior foundation for meeting future 5G and beyond network requirements.

7. AI-and-RAN Virtualized Infrastructure in AITRAS

SoftBank is working to seamlessly integrate AI workloads into the telecom infrastructure, focusing on solving challenges like scalability, computational efficiency, and resource optimization aided by the rapid advances in AI technologies.

7.1 SoftBank AI-and-RAN Approach

The virtualized infrastructure of AITRAS consists of two major clusters: the Management Cluster and the Workload Clusters. The Management Cluster consolidates functions like multi-cluster management, GitOps, image registry, and central repositories, and is deployed at the core of the network. The Workload Clusters, on the other hand, are geographically distributed closer to the network edge for optimal performance and act as the execution environment for application workloads.

The AI-and-RAN concept is designed to address the growing computational demands imposed by AI workloads, particularly those involving inferencing. With the rise of LLMs and other AI applications, AI-and-RAN aims to provide a platform that can handle these heavy workloads alongside traditional RAN functions. This convergence is essential for optimizing resource usage and reducing both capital and operational expenditures in telecom networks.

7.2 Hardware and Resource Management

The AITRAS's AI-and-RAN virtualized infrastructure relies on both conventional servers and integrated CPU-GPU hardware, such as NVIDIA GH200, to efficiently manage RAN and AI workloads. The architecture prioritizes flexibility and scalability to meet the real-time and dynamic throughput requirements typical of RAN, driven by user demand and changing network conditions.

Three levels of resource management are implemented:

- Cluster Level: Distribution and allocation of workloads between different clusters.
- Node Level: Management of computing resources at individual nodes.
- Core Level: Fine-tuned management of resources at the hardware core level.

This multi-layered approach ensures that computational resources, particularly GPUs, are utilized to their fullest capacity, maximizing efficiency and reducing operational costs. Furthermore, AI-and-RAN employs priority-based resource management to support monetization efforts within the infrastructure by dynamically allocating resources according to workload importance.

The use of CPU-GPU integrated servers brings significant advantages in terms of computational power, especially for AI workloads, most of which require hardware acceleration. These servers integrate both CPUs and GPUs on a single board, providing the necessary power to handle complex AI inferencing tasks while maintaining cost-effectiveness. By relying on such hardware, AI-and-RAN is able to provide robust support for both AI and RAN functionalities, ensuring that computational resources are allocated efficiently based on real-time demands.

7.3 AITRAS AI-and-RAN Orchestrator

At the heart of the AITRAS's AI-and-RAN architecture is the orchestrator, responsible for the comprehensive management of AI and RAN workloads across the entire infrastructure. The orchestrator's functions, which include scheduling workloads, multi-cluster management, and handling manifests, ensure seamless interaction among the components of the infrastructure.

The orchestrator exposes a REST API that external systems can use to interact with the virtualized infrastructure, enabling automated and flexible workload deployment. It makes resource allocation efficient even under variable conditions by using advanced scheduling mechanisms based on node capability and workload requirements.

The scheduling process is divided into basic rules, which include filtering based on node capabilities, and extensive rules, which deal with more complex decisions, such as pod relocation and configuration adjustments. This extensible mechanism allows for continuous optimization and adaptation to handle to the evolution of network requirements.

Moreover, the orchestrator handles manifest files which describe the resources required by particular workloads. By managing manifests, the orchestrator ensures that workloads are deployed consistently across the AI-and-RAN virtualized infrastructure. The orchestrator also leverages GitOps practices, using

Git as the single source of truth for managing the state of the infrastructure. This approach provides a reliable and auditable way to manage infrastructure changes, improving the overall stability and security of the AI-and-RAN virtualized infrastructure.

The orchestrator includes monitoring functions and provides information such as:

- Locations of the managed servers.
- CPU and GPU utilization for RAN and for AI.
- GPU power consumption of each server.
- GPU utilization of each server.
- The number of managed servers and their roles as either RAN or AI.

Figure 13. Orchestrator monitoring result

7.4 Agentic AI - Serverless API Powered by NVIDIA AI Enterprise

Agentic AI refers to artificial intelligence systems that act as autonomous agents, capable of making decisions, performing tasks, and adapting to new situations with minimal human intervention. These systems are designed to operate independently within defined parameters, using data and algorithms to dynamically respond to changes in their environment and achieve specific goals. Examples of agentic AI include virtual assistants, task automation bots, and AI-native decision-making systems.

Agentic AI can be realized using serverless APIs by leveraging the flexibility and scalability of serverless computing. APIs allow developers to build and deploy individual functions or microservices that handle discrete tasks, such as data processing, decision-making, or communication, without creating or

managing dedicated infrastructures. Serverless APIs can serve as modular building blocks for agentic AI systems, enabling real-time responses to user input or environmental changes. By combining serverless architectures with machine learning models and APIs, developers can create lightweight, scalable, and cost-effective agentic AI solutions that are adaptable and responsive.

Serverless API powered by NVIDIA AI Enterprise provides a serverless API designed for deploying and managing AI workloads on GPUs at a global scale. Key benefits for developers include:

- **Serverless Architecture:** Focus on building AI applications without managing infrastructure.
- **GPU Acceleration:** Leverages NVIDIA's accelerated computing platforms for high-performance AI workloads.
- **Scalability:** Applications that scale easily as demand grows.
- Flexibility: Supports various deployment modes like containers and Helm charts.
- **Security and Reliability:** Provides built-in security and reliable operation.
- Versatile API Access: Supports HTTP polling, streaming, and gRPC for diverse communication needs.

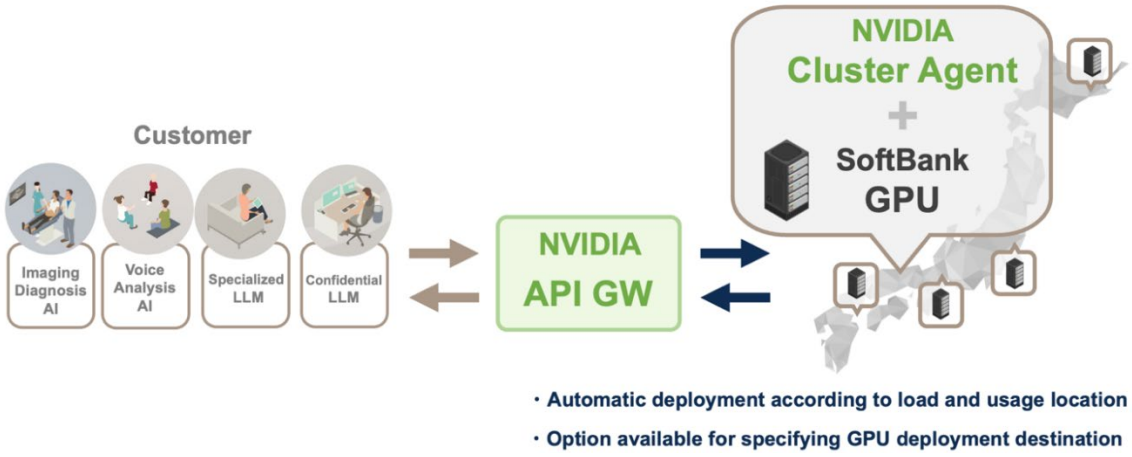

Figure 14. NVIDIA serverless API implementation solution for SoftBank infrastructure

NVIDIA serverless API is suitable for use cases like real-time AI inference, model fine-tuning, and running lightweight generative AI models. Developers benefit from easy deployment, cost optimization through scaling down during inactivity, and integration with cloud ecosystems. The orchestrator will enable the management and optimization of GPU resources within the virtualization platform. By allocating projected unused GPU resources to NVIDIA serverless API, GPU resources can be provided to users without impacting the other services running on AITRAS. Furthermore, by fully integrating Serverless API powered by NVIDIA AI Enterprise in AITRAS, base stations —previously viewed as cost centers—can be transformed into revenue-generating assets through the provision of AI applications and other services.

7.5 Meeting High Availability and Performance Standards

To meet the stringent availability requirements of telecom networks, the AI-and-RAN virtualized infrastructure in AITRAS is built to offer redundancy and high performance. The orchestrator is designed to ensure that even if it becomes unavailable, existing workloads on the clusters remain unaffected. Failover mechanisms, manual intervention options, and alternative onboarding paths help maintain consistent service availability.

The AI-and-RAN infrastructure incorporates several redundancy strategies to ensure uninterrupted service. For instance, the multi-cluster management capability allows workloads to be seamlessly transferred between clusters if failure occurs. Additionally, the orchestrator can cache critical information about cluster states, enabling it to resume operations smoothly after a disruption. This design ensures that any failure at the management level does not propagate to the workload level, preserving the reliability of ongoing services.

Moreover, the infrastructure supports vertical and horizontal scalability. Vertical expansion allows the addition of more clusters, nodes, or computing power, while horizontal expansion adds more computing resources to individual nodes, such as CPUs, GPUs, and memory. This dual scalability model ensures that the infrastructure can grow and adapt to changing network demands.

7.6 Sustainability and Energy Efficiency

AITRAS AI-and-RAN's approach, which leverages CPU-GPU integrated servers, introduces a unique requirement: balancing performance against energy efficiency. GPUs, while powerful, are energy-intensive, making it critical to evaluate and compare energy efficiency across different deployment models. SoftBank envisions AI-and-RAN as an environmentally sustainable solution, aiming to ensure that its energy consumption is either superior to or on par with traditional virtualized RAN systems. To address these requirements, AITRAS AI-and-RAN incorporates energy monitoring and optimization mechanisms that help minimize the power consumption of GPU-intensive workloads. By leveraging Kubernetes' built-in resource management capabilities, AITRAS can optimize the allocation of GPU resources, turning them off or scaling them down when they are not needed. This approach not only reduces energy consumption but also extends the lifespan of the hardware components, contributing to overall sustainability.

SoftBank is also exploring the use of renewable energy sources to power AI-and-RAN deployments, further reducing the carbon footprint of its infrastructure. By prioritizing energy efficiency and sustainability, AI-and-RAN is positioned as a responsible solution for the future of telecom networks, balancing the need for high computation power with environment protection.

AITRAS's AI-and-RAN virtualized infrastructure marks a significant breakthrough in the convergence of AI and telecommunications. This cutting-edge platform optimizes resource utilization while delivering unparalleled flexibility and efficiency. By leveraging virtualization, containerization, and advanced orchestration, AITRAS is the foundation for next-generation telecom networks that are highly adaptable, scalable, and primed to unlock the full potential of AI.

8. AITRAS AI Applications

8.1 The Shift to Computing-Centric Architecture

The evolution of telecom networks is increasingly influenced by the shift from a transmission-centric to a computing-centric architecture, driven by the demands of AI workloads and next-generation network capabilities. Traditionally, telecom networks were designed with a focus on data transmission and management, emphasizing the movement of large volumes of data through the network. However, with the integration of AI into RAN, there is a shift toward computing-centric architecture, where computation power and the ability to process data efficiently take precedence. This shift allows for the real-time execution of AI models, which are essential for enabling automation, predictive analytics, and advanced use cases like autonomous driving and intelligent customer service. The computing-centric approach ensures that network resources are used effectively to provide better service quality and adaptability.

8.2 Use Cases for the AITRAS AI-on-RAN

Integrating AI into the RAN infrastructure enhances network performance while addressing challenges posed by real-world AI applications. Unlike "AI-and-RAN," which describes the coexistence of AI and RAN workloads within AITRAS, or "AI-for-RAN," where AI is used to optimize and improve RAN operations, "AI-on-RAN" takes another step into the future. It involves embedding AI capabilities directly into the RAN infrastructure, enabling real-time decision-making and automation to support emerging AI-native applications. Here are some examples of such use cases, showcasing how these solutions can be applied in real-world scenarios to address specific challenges effectively.

8.2.1 Autonomous Driving

AITRAS’s AI-on-RAN technology plays a pivotal role in enabling autonomous driving by delivering the ultra-low latency and high reliability essential for autonomous vehicles related applications. With its ability to process data in real time, AITRAS ensures that autonomous vehicles receive critical time-sensitive information, such as traffic updates, road hazards, and real-time navigation support. This enhances both the safety and efficiency of autonomous driving.

SoftBank has developed a traffic understanding multimodal AI specifically designed for autonomous vehicle monitoring. By harnessing AITRAS's low-latency processing and high-infrastructure security, the multimodal AI operates in real time, offering a comprehensive understanding of vehicle status. This enables reliable remote support, ensuring seamless and safe autonomous driving.

A field trial of our traffic understanding multimodal AI was conducted in Kanagawa Prefecture, Japan. During the trial, the AI utilized AITRAS to instantly analyze potential driving risks by processing live footage transmitted from autonomous vehicles. It then translated these risks into actionable recommendations, providing real-time remote support. This advanced system allows autonomous vehicles to navigate safely, even in scenarios where they are unable to independently assess risks. Currently, remote operators issue instructions to autonomous vehicles based on information analyzed and verbalized by the traffic understanding multimodal AI. However, the ultimate goal is to achieve fully unmanned operations by allowing the traffic understanding multimodal AI to directly issue instructions to the autonomous vehicles.

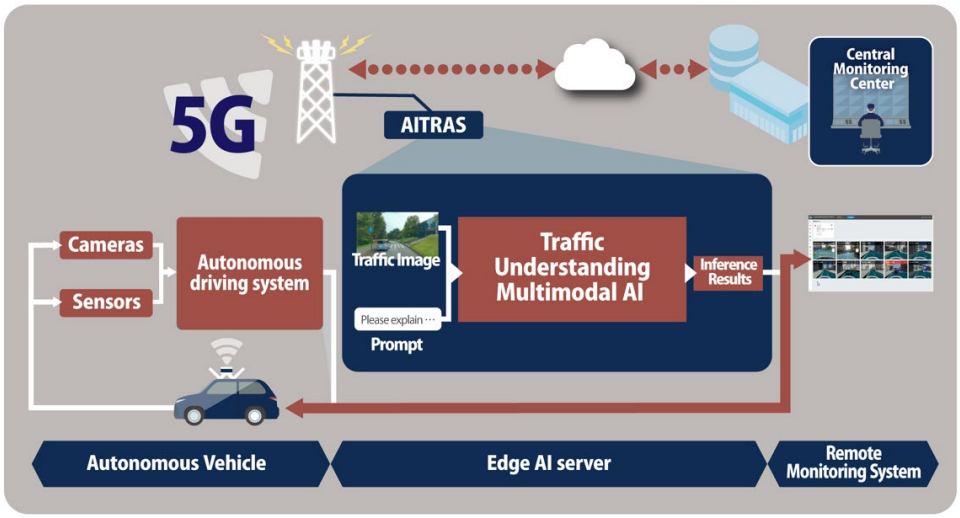

Figure 15. Ultra-low latency and highly secure autonomous driving remote support solution enabled by AITRAS

One scenario tested involved driving in a situation where a vehicle is stopped in front of a crosswalk. In this scenario, there is a risk of overlooking a person attempting to cross from behind the stopped vehicle, which could result in a collision with the pedestrian who is emerging from the vehicle's blind spot. According to Japanese traffic regulations, when approaching a signal-free crosswalk with a stopped vehicle in front, drivers are required to come to a complete stop before proceeding forward (Figure 16).

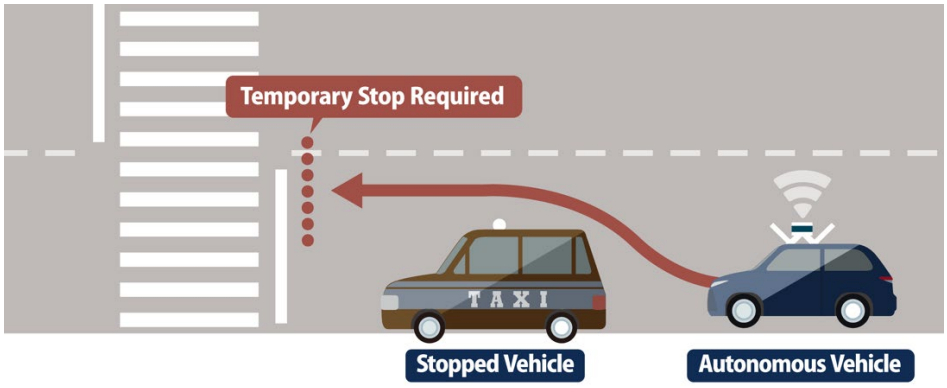

Figure 16. When there is a stopped vehicle in front of a crosswalk, a temporary stop is required

In this case, if the autonomous vehicle approaches the crosswalk at high speed or fails to perform a stop, there is a potential risk of accidents. When the autonomous vehicle drives normally, the traffic understanding multimodal AI generates instructions to continue to drive normally (Figure 17).

Figure 17. Inference results of multimodal AI when there are no significant risks on the road

When the autonomous vehicle approaches a stopped vehicle and a crosswalk, the risk level is high, the traffic understanding multimodal AI generates the instruction to stop in front of the crosswalk, as a pedestrian may suddenly appear (Figure 18). This confirmed the feasibility of using AITRAS to remotely support autonomous driving from an external location.

Figure 18. Inference results of multimodal AI when a vehicle is stopped in front of a crosswalk

8.2.2 Large Language Model (LLM) Robots:

LLM robots are defined as those that use LLMs in generating actions, enabling them to perform not only pre-programmed tasks, but also to autonomously understand and respond to their environment by collecting contextual information. Operating these LLMs on a larger computational infrastructure like AITRAS, as opposed to the limited computation resources of the robots themselves, allows for the utilization of more sophisticated LLMs, thereby enabling more flexible and adaptive behavior generation.

Typically, the process of generating outputs from LLMs after receiving prompts takes over one second. However, real-time robot control necessitates that both the inference time of the LLM and the transmission of data to and from the computing infrastructure on AITRAS be completed within approximately 0.1 seconds. To address this requirement, we have developed a high-speed control LLM and integrated it into the AITRAS infrastructure (Figure 19).

This development enables various robots, which leverage advanced decision-making through LLMs, to effectively operate across different sectors such as healthcare, education, and transportation, in any location where mobile network connectivity is available.

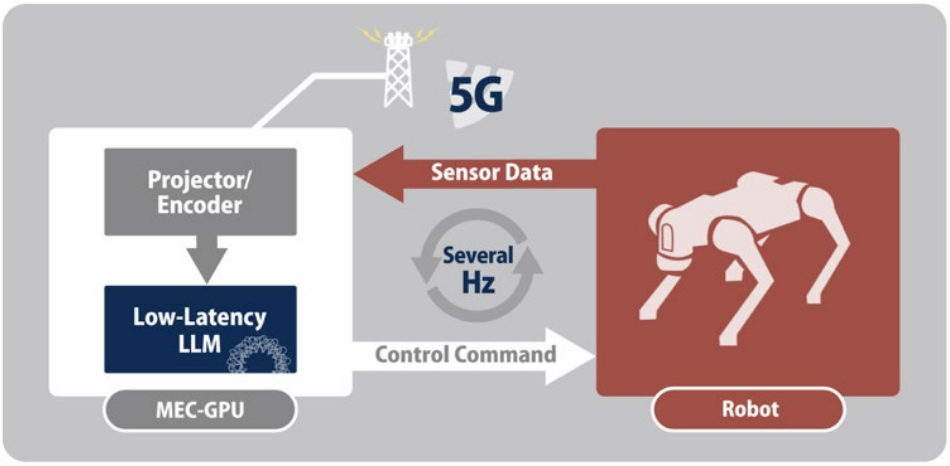

Figure 19. LLM robot operating with low latency

8.2.3 Retrieval-Augmented Generation (RAG) Chatbots for Customer Service and Network Support

AI-RAN also facilitates the deployment of Retrieval-Augmented Generation (RAG) chatbots, which combine LLMs with real-time data retrieval to offer accurate and context-aware responses. In customer service and network support roles, RAG chatbots leverage the AI-RAN infrastructure to provide quick and well-informed assistance, resulting in improved customer satisfaction. Since all data is processed exclusively in a closed environment, secure execution in a more secure environment is assured unlike cloud services over the Internet.

9. Strategic Business Models and Revenue Generation

9.1 Demand Forecasting, Customer Segmentation, and Business Models

As highlighted in the previous chapter, the telecom industry faces a critical challenge: deriving significant value from ever larger capital investments. However, integrating AI and RAN technologies can turn these investments into highly profitable business opportunities.

9.1.1 Cloud AI and GPUaaS Market Forecasts

The rapid expansion of the cloud AI market and its enabling GPUaaS market is noteworthy. GPUaaS provides scalable GPU resources for industries requiring advanced AI processing. By 2030, the global cloud AI market is projected to reach $397.8 billion[4], with the GPUaaS market contributing $25.5

[4] Fortune Business Insights, Cloud AI market size: https://x.gd/OUVxM

billion[5]. This growth underscores the widespread adoption of AI technologies and the surging demand for high-performance computing.

AI-RAN enhances the value of GPUaaS by leveraging low-latency edge computing and flexible AI capabilities. This is crucial for meeting the rising demands of industries such as manufacturing, healthcare, finance, and entertainment, all of which require real-time data processing.

9.1.2 GPUaaS Use Cases: Central Data Centers and Network Edge with AI-RAN

Cloud GPU services are divided between central data centers and network edge solutions like AI-RAN. Central data centers handle large-scale AI training, while AI-RAN, with its edge computing advantages, offers low latency, high capacity, enhanced security, and distributed processing.

Key applications for AI-RAN include:

- **Autonomous Driving and Robotics:** Ultra-low latency AI inferencing for high-definition image processing.
- **Manufacturing:** Real-time monitoring, predictive maintenance, and process optimization.
- **Entertainment:** AR/VR delivery with low-latency, high-resolution experiences.
- **Smart Cities:** Dynamic traffic signal management and urban optimization.
- **Healthcare:** Telemedicine and high-resolution image analysis.
- **Customer Service:** Real-time communication with intonation and volume.

AI-RAN complements existing cloud GPU services by addressing edge-specific demands while reducing the need to transfer large data volumes over networks.

9.1.3 Transforming Telecom Business Models with AI-RAN

AI-RAN enables telecom operators to reimagine their traditional business models, creating new revenue streams from unique value-added AI-enabled services:

- **Integrated Low-Latency AI Services:** Combining AI processing with telecom infrastructure to support emerging applications such as edge video analytics and AR/VR.
- **Data Sovereignty and Secure Processing:** Advanced AI processing within secure, domestic networks, ensuring data security compliance and protection.
- **AI Data Analytics and Prediction:** Providing enterprises with insights derived from anonymized data, supporting applications like tourism optimization, traffic management, and demand forecasting.

[5] Fortune Business Insights, GPUaaS market size: https://x.gd/fta9k

By transitioning from traditional telecom services to AI-native solutions, operators can unlock new growth opportunities while enhancing network utilization and efficiency.

9.2 AITRAS AI-and-RAN for New Revenue Generation

AI integration into RAN unlocks new opportunities for creating innovative and impactful services. For example, Edge AI inferencing enables real-time AI capabilities at the network edge that will support applications such as augmented reality (AR), virtual reality (VR), smart cities, real-time communication, and video analytics. These technologies allow for faster data processing and decision-making closer to the user, significantly improving performance and reducing latency.

Strategic investments in Edge AI platforms, such as AITRAS, along with infrastructure containerization, play a crucial role in this transformation. By leveraging these technologies, service providers can offer differentiated, revenue-generating services that cater to the growing demand for smarter, more responsive networks. This not only enhances user experience but also provides a competitive edge in an increasingly digital and connected world.

9.3 TCO Analysis

AITRAS's AI-and-RAN features impact total cost of ownership (TCO) by balancing capital expenditure (CAPEX) and operating expense (OPEX) considerations:

- **Reduced OPEX:** Automation of network management tasks lowers operational costs while improving service quality.
- **CAPEX Recovery through New Revenue Streams:** Long-term benefits of AI-native services justify initial infrastructure investments.

A comprehensive TCO analysis highlights the transformative potential of AITRAS for sustainable growth in the telecom sector. Detailed case studies will be explored in the following sections.

10. Case Study: AI-RAN TCO Analysis

10.1 AI-RAN Deployment Simulation in Urban Area, Tokyo

In this section, we present a case study focusing on Tokyo's Shibuya area where broadband wireless resources are essential for addressing mobile communication traffic demands.

Shibuya Station, a major terminal station located in one of Japan's busiest commercial districts, handles over 2.5 million passengers daily. This area represents one of Japan's most active and traffic-intensive urban zones, making it an ideal dense urban model for studying traffic management solutions. Based on the area coverage and anticipated traffic demands in the Shibuya area, our design encompasses 600 cells.

10.2 Regional Peak Traffic Variations

When examining traffic patterns in the Shibuya area during specific time periods, notable variations emerge. Traffic peaks intermittently throughout the 24-hour cycle. For instance, during evening rush hours, cells around the station experience near-peak traffic levels, while residential areas further from Shibuya Station show relatively low traffic. Similarly, in the early morning before work and late evenings after returning home, traffic levels in residential areas increase. Since many people change their living areas dynamically throughout the day, this alternation between peak traffic areas and low traffic areas occurs consistently, varying between weekdays and weekends.

Peak traffic Low traffic

Figure 20. Peak traffic and low traffic areas during daytime

Each cell is designed with the capacity to handle its specific peak-time traffic demands. This essential investment in infrastructure is crucial to meeting high-traffic demands such as high-definition video streaming, thereby ensuring high-quality services. However, traffic decreases significantly in off-peak times. As a result, when averaged across both peak and off-peak periods over a 24-hour cycle, the actual utilization rate of wireless resources (Downlink PRB[6]) averages approximately 30%[7].

[6] Physical Radio Block in 3GPP
[7] The utilization rate observed in this case study does not guarantee similar results in other areas

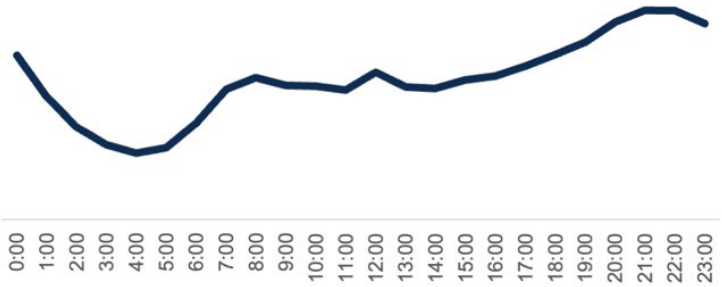

Figure 21. Hourly average downlink PRB utilization rate[8]

10.3 ROI Analysis of AI-RAN with NVIDIA GB200-NVL2

Considering the anticipated growth in future mobile traffic demands, continued investment in RAN capacity remains essential. By adopting the next-generation NVIDIA GB200-NVL2 platform, which features a powerful integration of CPU and GPU through the Grace and Blackwell architecture, AI-RAN can not only increase cell capacity but also enhance cost efficiency. This section examines the economic impacts of utilizing NVIDIA GB200-NVL2.

10.3.1 TCO

Considering the deployment of 600 cells in the Shibuya area, we compare the hardware, software, and operational costs of the existing ASIC-based custom BBU to AI-RAN using NVIDIA GB200-NVL2. Costs such as radio units, antennas, and installation costs at the cell site, as well as the 5G core network, network monitoring and billing, are assumed to be identical for both systems and are excluded from the analysis. The analysis period is five years, matching the depreciation period for hardware and software, with TCO calculated based on the operational costs over this period.

AI-RAN, leveraging the NVIDIA GB200-NVL2 platform, is expected to offer significant improvements in cell capacity and cost efficiency compared to traditional ASIC-based custom BBU. Consequently, the TCO for AI-RAN utilizing the NVIDIA GB200-NVL2 is shown to be lower than that of the ASIC-based custom BBU even in 100% "RAN-Only" mode, demonstrating a cost advantage. x86 architecture CPU does not demonstrate a clear cost advantage over ASIC-based custom BBU, highlighting the economic viability of AI-RAN.

[8] Actual data of average downlink PRB utilization rate of all cells in an urban area of Japan's capital region on weekdays (April 2023)

10.3.2 Revenue Potential

In a scenario where the NVIDIA GB200-NVL2 platform is utilized in 100% AI-only mode, each NVIDIA GB200-NVL2 server generates 25,000 tokens per second based on NVIDIA GB200-NVL2 AI performance benchmarks[9], potentially resulting in revenue of $20 per hour per server or $15,000 per month per server. We estimate the revenue potential in a scenario where one-third of the NVIDIA GB200-NVL2 resources are allocated for AI purposes, and the remaining two-thirds are utilized for RAN ("RAN-Heavy"). The mobile business revenue from communication services is assumed to be constant for both AI-RAN and ASIC-based custom BBU and is therefore excluded from the comparison. By allocating one-third of the NVIDIA GB200-NVL2 resources to AI, new revenue streams are generated. This not only allows AI-RAN to demonstrate cost advantages compared to ASIC-based custom BBUs but also offsets the TCO with the additional revenue from AI purposes, resulting in net profit.

10.3.3 ROI Analysis in the “RAN-Heavy”

When evaluating ROI[10] of utilizing the NVIDIA GB200-NVL2 in the "RAN-Heavy" mode, we project that AI-RAN can achieve a 33% ROI over five years. Additionally, the total revenue generated over these five years is estimated to be approximately twice the AI-RAN CAPEX investment.

10.3.4 Flexibility between "RAN-Heavy" and "AI-Heavy"

Previous calculations were based on the "RAN-Heavy" mode focused on RAN implementation. However, the performance of the NVIDIA GB200-NVL2 and the flexibility of AI-RAN allow for flexible allocation of resources to RAN and AI purposes. Therefore, by allocating more of the NVIDIA GB200-NVL2 resources to AI purposes, prioritizing revenue from AI becomes possible ("AI-Heavy"). We estimate the revenue potential of the "AI-Heavy" mode by utilizing two-thirds of the NVIDIA GB200-NVL2 resources for AI purposes and allocating the remaining one-third to RAN. As mentioned earlier, revenue from mobile business is excluded from this comparison.

The five-year TCO comparison reveals that the "AI-Heavy" mode using NVIDIA GB200-NVL2 has a higher TCO than the "RAN-Heavy" mode due to the increased number of servers required for 600 cells. However, the "AI-Heavy" mode demonstrates significantly higher revenue generation compared to the "RAN-Heavy" mode, with its revenue substantially exceeding the TCO.

10.3.5 ROI Analysis in the “AI-Heavy”

When evaluating ROI, we project that AI-RAN can achieve a maximum ROI of 219% over five years. Additionally, the total revenue generated over these five years is estimated to be approximately five times the AI-RAN CAPEX investment. This analysis demonstrates the potential for transforming

[9] The token AI workload used is Llama-3-70B FP4
[10] ROI % = (new AI revenues – TCO) / TCO

network infrastructure from a cost center into a profit center.

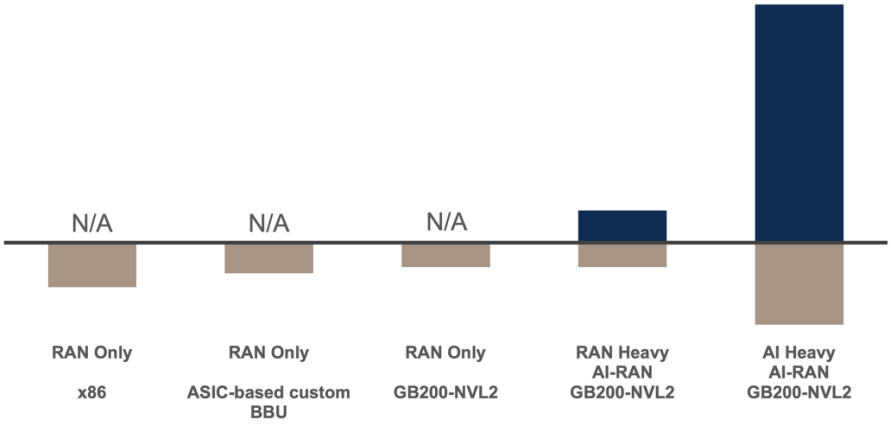

Figure 22. AI-RAN economics for covering an urban area in Tokyo with 600 cells

11. Conclusion

11.1 Charting the Future of Tomorrow's Networks

AITRAS, SoftBank's AI-RAN product presents a transformative solution to the challenge of managing large-scale capital investments in telecom while ensuring profitability. This white paper illustrates how AI-RAN can revolutionize network infrastructure, shifting it from a traditional cost center to a profit generator, that enables sustainable growth through innovative business models.

11.1.1 Economic Perspective

AI-RAN implementation offers substantial economic advantages for telecom operators. A SoftBank's simulation case study conducted for Tokyo's Shibuya area highlights how AI-RAN could optimize the operation of 600 cells while enhancing return on investment.

- **Cost Efficiency:** AI-RAN significantly reduces the TCO of hardware, software, and operations, as evidenced by a significant improvement in cell capacity and cost efficiency.
- **New Revenue Opportunities:** Beyond infrastructure efficiency, AI-RAN opens new revenue streams by providing computing resources for AI applications. Additionally, AI-RAN allows for flexible investment strategies, such as "RAN-Heavy" or "AI-Heavy," by flexibly allocating computing resources to both RAN and AI. This transforms the infrastructure from a cost center into a profit center.

The case study indicates that AI-RAN not only improves TCO compared to traditional ASIC-based custom BBU, but also, driven by new AI revenue, is projected to achieve a maximum ROI of 219% over five years.

11.1.2 Social Impact Perspective

Nationwide deployment of AITRAS accelerates digital transformation, fostering smart city initiatives and enhancing citizens' quality of life. Key impacts include:

- **Urban Efficiency:** Supporting applications such as autonomous vehicle communications, dynamic traffic management, and advanced telemedicine.
- **Bridging the Digital Divide:** Ensuring equitable access to high-quality network services across regions, thereby enabling remote work, distance learning, and new economic opportunities in areas that are now underserved.

AITRAS is evolving the telecom infrastructure into a foundational pillar of societal digitalization, offering growth opportunities for telecom operators while driving transformative social benefits.

11.2 Long-Term Vision and Sustainable Growth Strategies

The transition to AI-powered base stations such as SoftBank AITRAS is a pivotal challenge for the telecom industry. This transformation requires not only technological upgrades but also a rethinking of business models and organizational frameworks.

11.2.1 Strategies for Sustainable Growth

To secure a competitive edge and foster sustainable growth, telecom operators should focus on:

- **Sustainable Technology Investment:** Prioritize R&D in Edge AI and distributed computing to enhance network performance, reduce costs, and enable new services.
- **Partnership Development:** Collaborate with AI companies, cloud providers, and device manufacturers to expand technical capabilities and service offerings.
- **Talent Development and Organizational Reform:** Train specialists in AI and telecom and adapt organization structures to support agile decision-making and innovation.
- **Flexible Business Models:** Diversify revenue streams through Edge AI services and computation resource rental, creating new opportunities in AI-native networks.

11.2.2 Long-Term Vision: SoftBank AITRAS as Revolutionary Infrastructure

In the future, RAN will evolve into core platforms that integrate AI and communications, driving society-wide digital transformation. These advancements will enable:

- Smart cities with enhanced urban infrastructure.
- Widespread adoption of autonomous vehicles.
- Revolutionary services in healthcare and education.

Telecom operators must adopt a shared vision for this transformation, implementing strategies that balance technological innovation against ethical and social responsibilities. As the AI market grows, it is reshaping the telecom infrastructure. Telecom RANs are no longer just communication hardware; they are becoming critical AI platforms. Organizations that embrace this shift and focus on innovation will secure a dominant position in the next-generation telecom market. By seizing these opportunities, telecom operators can drive new value creation, expand markets, and support innovative social development.

References

[1] SoftBank, https://www.softbank.jp/en/corp/technology/research/story-event/057/

[2] SoftBank, https://www.softbank.jp/en/corp/technology/research/story-event/069/

[3] SoftBank, https://www.softbank.jp/en/corp/technology/research/news/052/

[4] SoftBank, https://www.softbank.jp/en/corp/technology/research/story-event/040/

[5] SoftBank, https://www.softbank.jp/en/corp/technology/research/story-event/041/

[6] SoftBank, https://www.softbank.jp/en/corp/technology/research/story-event/039/

[7] AI-RAN Alliance, https://ai-ran.org

[8] AI-RAN Alliance, https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf

[9] NVIDIA, https://developer.nvidia.com/blog/bringing-ai-ran-to-a-telco-near-you/

[10] NVIDIA, https://developer.nvidia.com/blog/ai-ran-goes-live-and-unlocks-a-new-ai-opportunity-for-telcos/

[11] Ericsson, https://www.ericsson.com/en/reports-and-papers/further-insights/ai-in-ran

[12] Telecoms.com, https://www.telecoms.com/ai/ai-ran-gathers-momentum-with-ericsson-softbank-tie-up

[13] RCR Wireless News, https://www.rcrwireless.com/20240917/fundamentals/what-is-the-ai-ran-alliance

[14] Fierce Network, https://www.fierce-network.com/wireless/nvidia-softbank-plot-big-rewards-ai-ran

[15] Fierce Network, https://www.fierce-network.com/ai/what-ai-ran

Acknowledgment

This white paper was authored by SoftBank's Research Institute of Advanced Technology team, with contributions from Ryuji Wakikawa, Alex Jinsung Choi, Hideto Funayoshi, Akinori Machida, Kiyoshi Nozaki, Shun Yamashina, Shinta Sugimoto, Koji Kusunoki, Jutatsu Sai, Koji Araki, Ayako Sato, Rie Maruno, Yuki Ota, Kenjiro Mori, Takuya Asakura, Koyo Enomoto, Hiroki Inaba, Sota Sugimura, Shu Anzai, Azumi Minami, and Sara Yellin.

We would like to express our deep gratitude to all the contributors who made this white paper possible. We also acknowledge the collaborative efforts of our technology partners, including NVIDIA, Arm, Fujitsu, Red Hat, Ericsson, and Nokia, whose cutting-edge innovations and partnerships were instrumental in shaping this research. We extend our appreciation to the many engineers, researchers, and developers whose relentless pursuit of technological excellence enabled the creation of AITRAS.

A sincere thanks to the AI-RAN Alliance for fostering an ecosystem of innovation, driving the adoption of AI-native RAN technologies, and supporting the development of industry standards.

Glossary

AI-RAN (Artificial Intelligence Radio Access Network): The integration of AI technologies into RAN to optimize network performance, automate resource management, and support new AI-native services.

AITRAS[11]: SoftBank's AI-RAN product based on a gRAN architecture and designed to merge AI workloads with telecom operations, enabling multi-tenant operations and enhanced service delivery.

gRAN (GPU-based RAN): An advanced RAN architecture utilizing GPUs for parallel processing to support AI-native functionalities, including real-time data analysis, low-latency processing, and large-scale AI applications.

Orchestrator: A central management system in AI-RAN environments that allocates, scales, and optimizes RAN and AI workloads dynamically, ensuring efficient resource usage and system performance.

AI-for-RAN: A framework focusing on using AI to enhance RAN performance, addressing areas like network efficiency, capacity, and optimization of radio resources.

[11] AITRAS is pronounced as /ˈaɪ-trʌs/, where "AI" rhymes with "eye" and "TRAS" rhymes with "plus." Note that AITRAS is not an acronym but a unique name for SoftBank's AI-RAN product.

AI-and-RAN: The integration of AI and RAN workloads on a shared infrastructure, enabling telecom operators to run both simultaneously for increased efficiency and new revenue opportunities.

AI-on-RAN: Embedding AI capabilities directly into the RAN infrastructure to enable real-time decision-making and automation, supporting advanced AI applications such as autonomous driving and robotics.

AI-RAN Alliance: A collaborative group of industry leaders, including SoftBank, NVIDIA, Arm, Ericsson, Nokia, Samsung, and others, aiming to accelerate the development and adoption of AI-native RAN technologies.

Serverless API: A cloud-based API that abstracts the underlying infrastructure, allowing developers to focus on building applications without managing servers. In AI-RAN, it enables the efficient deployment and scaling of AI tasks.

GPUaaS (GPU as a Service): A cloud-based service offering scalable GPU resources for AI and machine learning tasks, enabling high-performance data processing without the need for on-premise hardware.

Edge AI: AI computation performed at the network's edge, reducing data transfer latency and enabling real-time decision-making for applications like autonomous driving and intelligent robotics.

SoftBank

The Research Institute of Advanced Technology at SoftBank, established in April 2022, is committed to advancing cutting-edge technologies that drive societal progress. As an *activator* for innovation, it focuses on research in AI-RAN, 6G, HAPS (High Altitude Platform Station), autonomous driving, quantum technologies, and other forward-looking fields.

Our mission is to harness the power of technology to create solutions that address real-world challenges and shape a better future. Guided by SoftBank's corporate philosophy, *"Information Revolution - Happiness for everyone,"* we strive to lead the next wave of advancements that improve lives and inspire progress.

For more details about our latest technological developments and initiatives, please visit our website: https://www.softbank.jp/en/corp/technology/research/