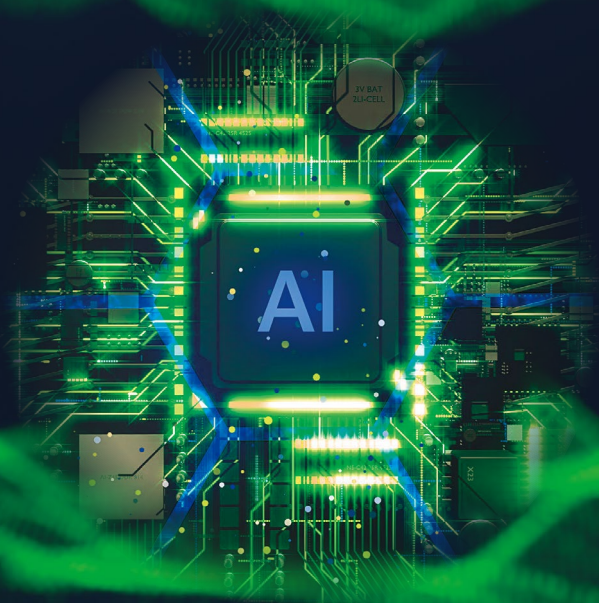


AI-RAN

AI-RAN
ALLIANCE
1st
ANNIVERSARY



SPECIAL REPORT

SoftBank Implements “LLM Foundation Model” for Telecom Operators on “AITRAS”

Accelerating the AI-Native RAN Era

[COLUMN] What Is “AITRAS,” a Converged AI-RAN Solution?

Refinement and inquiry.

Is that all there is to research?

Technology has always been a catalyst for social change. But without the implementation of technology into society, technology cannot activate social change.

We're the agents activating that change. That's why we do research.

To break down barriers while taking the shortest possible path.

That's why we do research.

Holding steadfast to our wildest dreams.

That's why we do research.

So we can break with the preconceived notions of the past, and pioneer new ones for the future.

We're researchers.

We'll reinvigorate society through technology.

With our research, we'll move the world forward.

We are Activators



SoftBank Implements “LLM Foundation Model” for Telecom Operators on “AITRAS”

Accelerating the AI-Native RAN Era

Interviewer: Impress SmartGrid Newsletter Editorial Team

On February 26, 2024, SoftBank formed the “AI-RAN Alliance”^[1] with global mobile network operators, telecommunications equipment vendors, and universities in preparation for the next 6G network era. Soon after, SoftBank unveiled “AITRAS,”^[2] a tangible product based on the AI-RAN concept (refer to the column at the end of the document).

In March 2025, it was announced that an “LLM^[3] foundational model” for telecommunications operators would be developed and implemented into AITRAS. The “AI model” utilizing this fundamental model can be applied flexibly in the field of wireless communications. We interviewed Ryuji Wakikawa, Head of Research Institute of Advanced Technology, SoftBank Corp., who has been leading this initiative, to learn more about the status of the development of the “LLM foundational model” based on his experience.

AI-RAN Alliance: Pioneering AI-native RAN Innovation

The AI-RAN Alliance was established on February 26, 2024, with the goal of realizing a new architecture that converges AI applications and the Radio Access Network (RAN) on a unified computing platform. This initiative aims to enhance performance, optimize resource utilization, and generate new revenue streams.

The AI-RAN Alliance collaborates with a diverse array of industry actors, including global

telecommunications operators like SoftBank and vendors. Together with key partners, the alliance is driving network transformation towards 5G (Fifth Generation) and beyond to 6G (Sixth Generation) through AI-RAN initiatives that facilitate advanced applications.

The AI-RAN Alliance’s work is structured around three primary technological domains, as illustrated in Figure 1.

- **AI-and-RAN** : This initiative integrates RAN and AI infrastructure to maximize infrastructure utilization and enable improved capital investment efficiency.

▼ [1]

AI-RAN Alliance website: <https://ai-ran.org/>

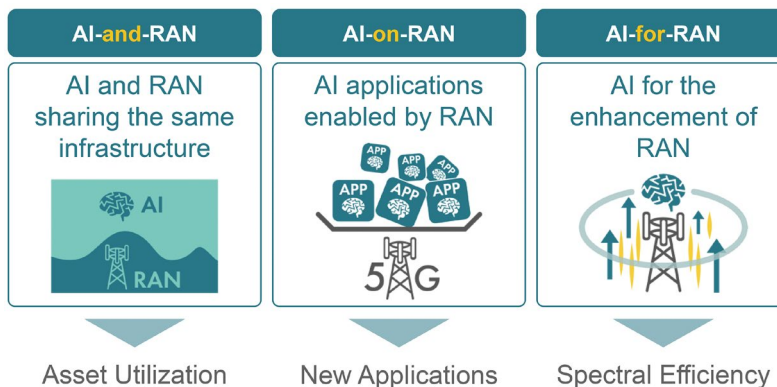
▼ [2]

AITRAS is a converged solution that enables the operation of RAN (Radio Access Network) and AI (Artificial Intelligence) on the same NVIDIA platform. https://www.softbank.jp/en/corp/news/press/sbkk/2024/20241113_06/

▼ [3]

Large Language Models (LLMs) are language models built using massive datasets and deep learning techniques. They have been a key driver behind the widespread adoption of generative AI.

Figure 1 Three Technological Domains of the AI-RAN Alliance



Source: SoftBank Research Institute of Advanced Technology



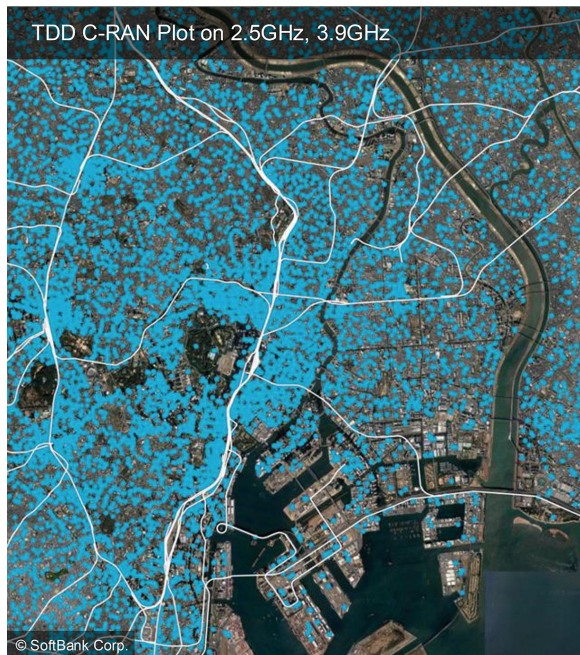
Ryuji Wakikawa

Ryuji Wakikawa, Head of Research Institute of Advanced Technology, SoftBank Corp., stated that “the development of the foundational model will benefit both telecommunications equipment vendors and network operators.”

Photo: Hiroyuki Matsumoto

- **AI-on-RAN** : This area focuses on deploying AI services at the network edge.
- **AI-for-RAN** : This field concentrates on using AI to enhance the core capabilities of RANs, such as spectral efficiency, energy optimization, and performance improvements.

Figure 2 Antennas Installed in High Density (Central Tokyo, blue dots indicate antennas)



Source: SoftBank Research Institute of Advanced Technology

Challenges of Wireless Networks in the 5G/6G Era and the Emergence of Generative AI and LLM

(1) New Opportunities Through Generative AI and LLM

Currently, wireless communication systems are transitioning to the 5G/6G era, which is characterized by an increasing demand for higher speed, greater capacity, and enhanced performance. Furthermore, as wireless communication systems consist of a large number of base stations and operate across various frequency bands, the complexity of the network continues to increase (see Figure 2).

Additionally, the capacity for mobile data traffic is expanding on a daily basis.

In order to address these challenges, telecommunications operators are making substantial investments and incurring considerable operational costs. However, these trends are expected to intensify further in the future. As a result, telecommunications operators are faced with the challenge of developing new approaches for network deployment and operational management, forcing them to transform themselves.

Meanwhile, the advent of generative AI (e.g., ChatGPT) that utilizes large-scale language models (LLMs) presents a new opportunity to reimagine the methodologies for the deployment and operational management of wireless networks.

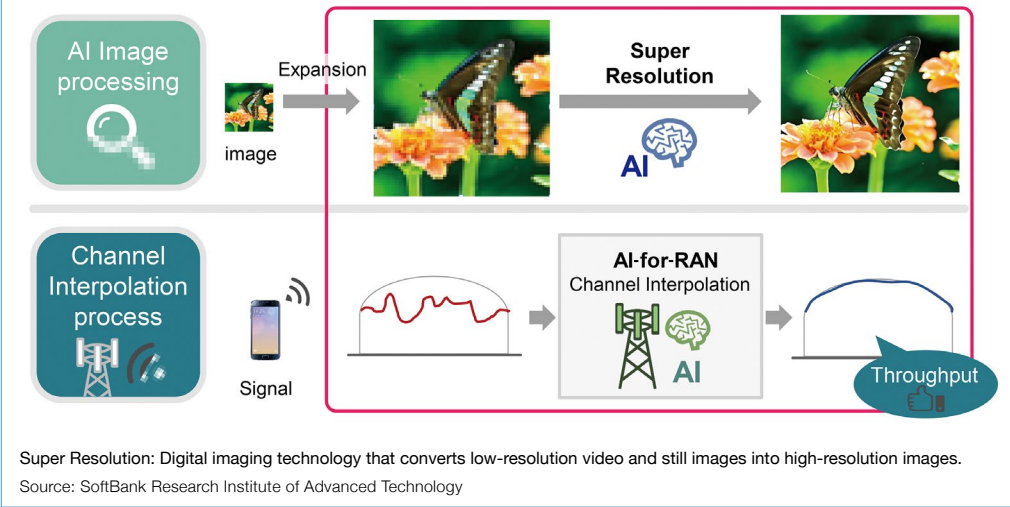
(2) Examples of AI-Powered Signal Processing

Let’s take a look at an example of signal processing in wireless communications using AI.

In wireless communications, incoming signal distortion can occur during transmission due to radio interference or insufficient received power. By leveraging AI’s predictive capabilities to interpolate for these distortions, it becomes possible to prevent throughput degradation (Use of AI-for-RAN: Lower part of Figure 3).

This AI-driven signal interpolation process is similar to super-resolution technology, where low-resolution video, as shown in the upper sec-

Figure 3 Super-Resolution Technology and Signal Interpolation in AI Image Processing Are Similar



▼ [4]
SRS (Sounding Reference Signal) is one of the uplink reference signals in 5G. It is used by the base station to estimate the uplink Channel State Information (CSI), which represents the propagation properties of the wireless channel through which the signal has traveled.

tion of Figure 3, is converted to high-resolution video by interpolating missing pixels through AI-based image processing.

AI-for-RAN: Enhancing RAN Performance Through AI

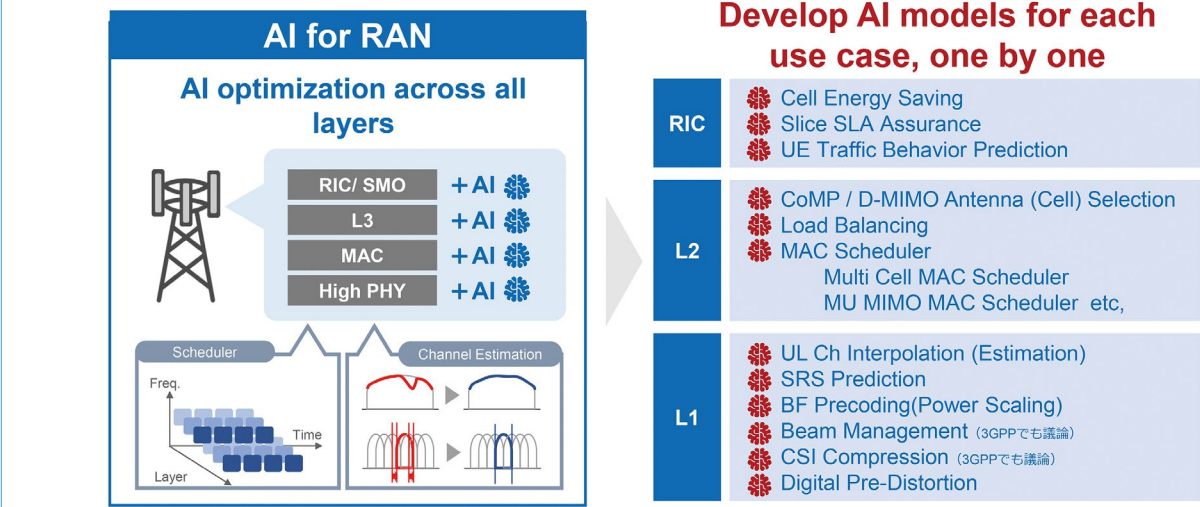
As previously mentioned, AI-for-RAN develops technologies that enhance RAN and improve

performance using AI.

The current approach to the development of AI models for wireless communications is depicted in Figure 4. The present approach focuses on optimizing layer-specific processing with AI and developing AI models tailored to specific use cases. Examples of such AI models include the following:

- **L1 (Physical Layer):** Channel interpolation processing, SRS^[4] prediction, beamforming/

Figure 4 The Current Approach to the Development of AI Models for Wireless Communications



SLA: Service Level Agreement CoMP: Coordinated Multiple Point transmission/reception
 D-MIMO: Distributed-MIMO Multi Input Multi Output MU MIMO: Multi-User Multiple Input, Multiple Output
 SRS: Sounding Reference Signal CSI: Channel State Information 3GPP: 3rd Generation Partnership Project
 Digital Pre-Distortion (DPD): One of the components used to increase the efficiency of power amplifiers in wireless communications.
 Source: SoftBank Research Institute of Advanced Technology

- ▼ [5]
CoMP stands for “Coordinated Multiple Point transmission/reception,” which is a technology that enables coordination between multiple cells for transmission and reception.
- ▼ [6]
D-MIMO stands for Distributed-MIMO (Multi Input Multi Output), which is a communication technology using multiple antennas.
- ▼ [7]
RIC (RAN Intelligent Controller) is a component responsible for controlling and optimizing RAN functions.
- ▼ [8]
SLA (Service Level Agreement) is a contract that defines service quality standards and other criteria agreed upon between service providers and users across various services, including network services.

- precoding, and beam management.
- **L2 (Data Link Layer):** CoMP^[5]/D-MIMO^[6] antenna selection, load balancing, and scheduler optimization.
- **RIC^[7] (RAN Intelligent Controller):** Cell power optimization (energy savings), SLA^[8] assurance, and user equipment behavior prediction.

Developing Versatile, Scalable Foundation Models

(1) Creating Foundation models for Enhanced Performance

SoftBank is not only continuing with its con-

ventional practice of building AI models for each use case (see Figure 4), but has also begun adopting a new approach that collects and trains on a variety of wireless communication-related data, including base station parameter values and terrain data, in order to develop telecommunications-specialized generative AI foundational models (see Figure 5).

Compared to earlier deep learning methods, the method excels in areas like scalability, handling of long-range contexts, and overall learning capability and versatility. Among these, the most important point is its simplicity of expansion to large-scale models.

Instead of building distinct AI models for

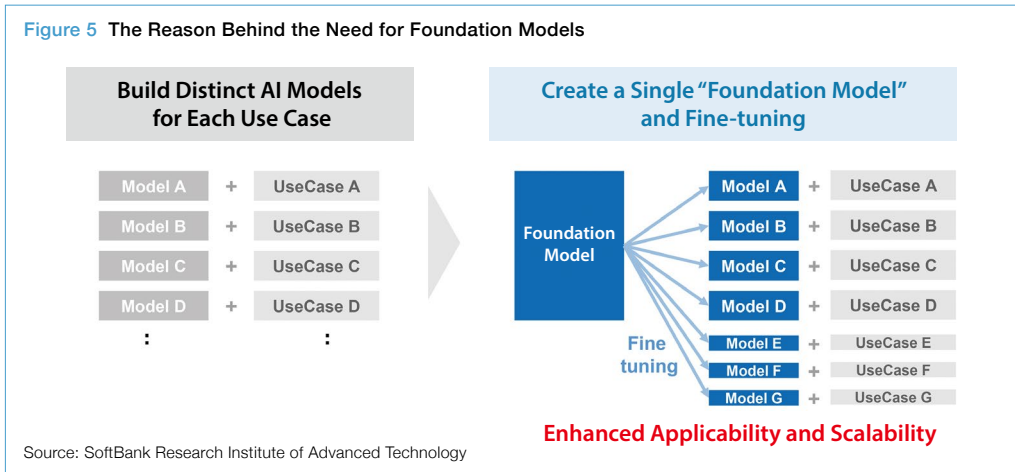
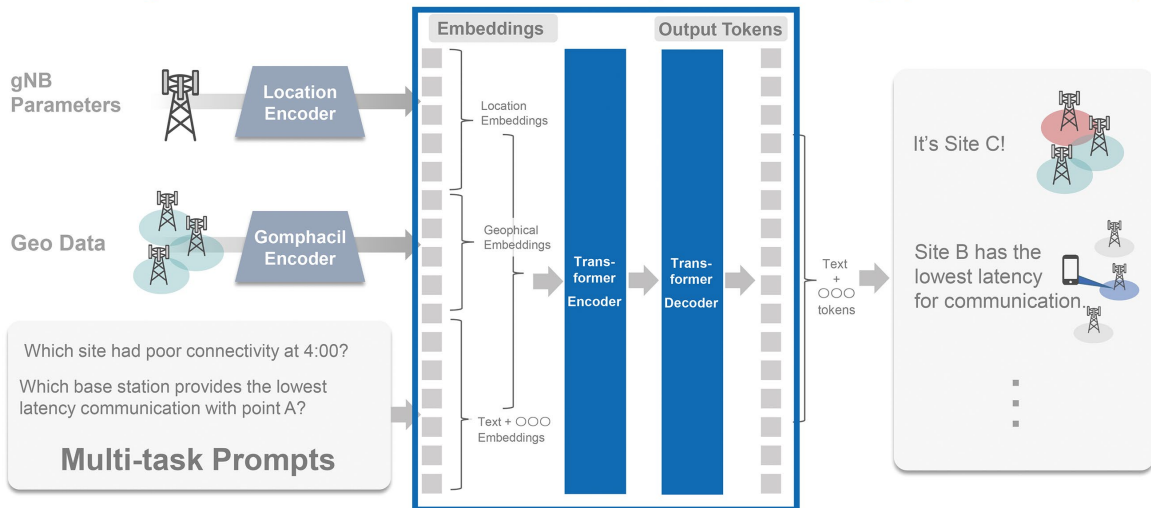


Figure 6 Training Methods for the LLM (Transformers)

Learning from various datasets through embedding (vectorization)



Source: SoftBank Research Institute of Advanced Technology

each use case as shown on the left side of Figure 5, SoftBank creates a single “foundation model,” as illustrated on the right side of Figure 5. In this foundation model, large volumes of data are trained, and subsequently fine-tuning^[9] is carried out to generate AI models that address various tasks (Figure 5, Figure 6). This broadens the scope of applications while also ensuring scalability.

(2) Comparing the Conventional Approach and SoftBank’s Approach

Figure 7 compares the conventional approach with SoftBank’s approach. In the conventional approach (Figure 7, left), the AI is developed by integrating “open data” (such as data from 3GPP and O-RAN)^[10] into the foundation of an existing LLM.

In contrast, SoftBank’s newly developed foundational model (Figure 7, right) incorporates not only open data but also terabytes of SoftBank’s communications-related data and retrains these data, while ensuring rigorous consideration for personal data protection. As a result, this model is capable of responding to queries related to SoftBank’s network.

Unlike general AI models, SoftBank’s telecommunications-specialized generative AI foundational model, which has been further trained using real-world telecommunications

data and equipment configuration parameters—excluding personal information—gained from its business experience, is expected to be applicable across various use cases in the telecommunications industry.

Use Cases of AI Models Built on Telecom-Specific Foundation Models

Figure 8 (page 8) illustrates how the AI model developed from the previously described foundational model for telecommunications operators can be applied.

A large volume of data related to wireless communications—such as base station data, wireless specifications (parameters, codes, etc.), and digital twin data (Figure 8, left)—is annotated^[11] and afterwards fully incorporated into the foundational model for training.

The fine-tuned AI model, built on this foundation, can be applied in various areas that many telecommunications operators focus on, including design and operational support, application-specific communication optimization, and performance enhancement across multiple base stations or individual stations.

Now, let’s explore specific use cases.

▼ [9]

Fine-tuning is the process of making additional training adjustments to a previously trained model using a new dataset.

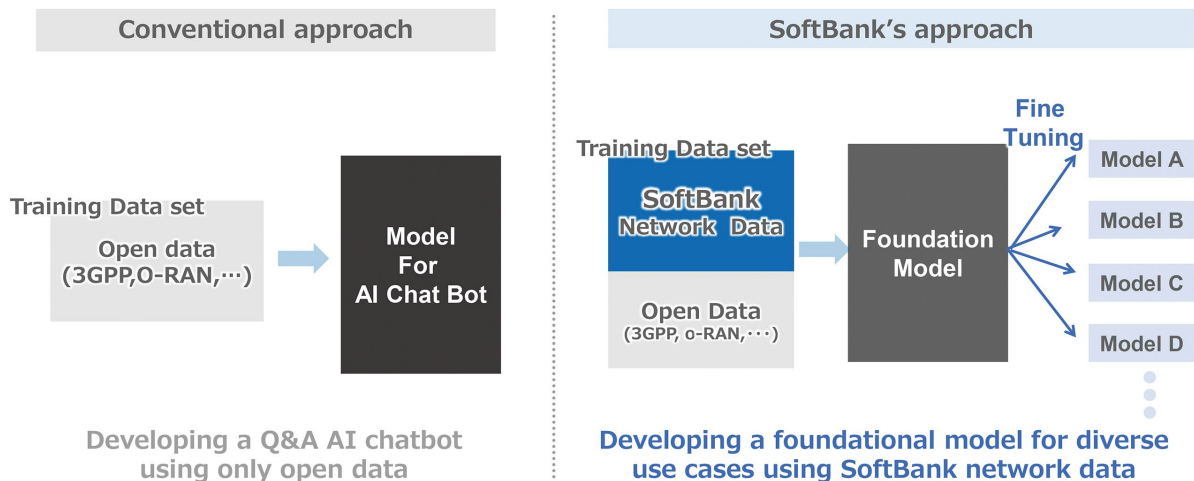
▼ [10]

The current approach also incorporates standardization data and documents from organizations such as ITU (International Telecommunication Union), IETF (Internet Engineering Task Force), and 3GPP (Third Generation Partnership Project), enabling the model to learn specification-related information.

▼ [11]

Annotation originally means adding notes or explanations. In the context of AI, annotating refers to the act of assigning tags (labels) to data, such as text, audio, images, or videos, to structure and organize it so that the AI can properly process the information.

Figure 7 Comparing the Conventional Approach and SoftBank’s Approach



Source: SoftBank Research Institute of Advanced Technology

▼ [12]

An AI agent is a program that leverages AI to process information and make decisions in place of humans.

[Use Case 1] AI Models for Operators

The red-boxed area in Figure 8 represents the domain of AI models designed for operators. In other words, this area relates to LLMs (AI models) that comprehend and generate human language. The following example illustrates a typical interaction within this model, where the AI agent^[12] generates responses to the operational administrator’s inputs.

Operational Administrator: “We are planning to deploy a new base station. Are there any issues with building a new base station in this area?”

AI Agent: “Yes, there is an issue. This area frequently experiences congestion (network overload).”

Operational Administrator: “Provide the optimal configuration parameters to reduce congestion in this area.”

AI Agent: “I get it. XX are the optimal configuration parameters.”

Mr. Wakikawa also presented the following case in addition to this example:

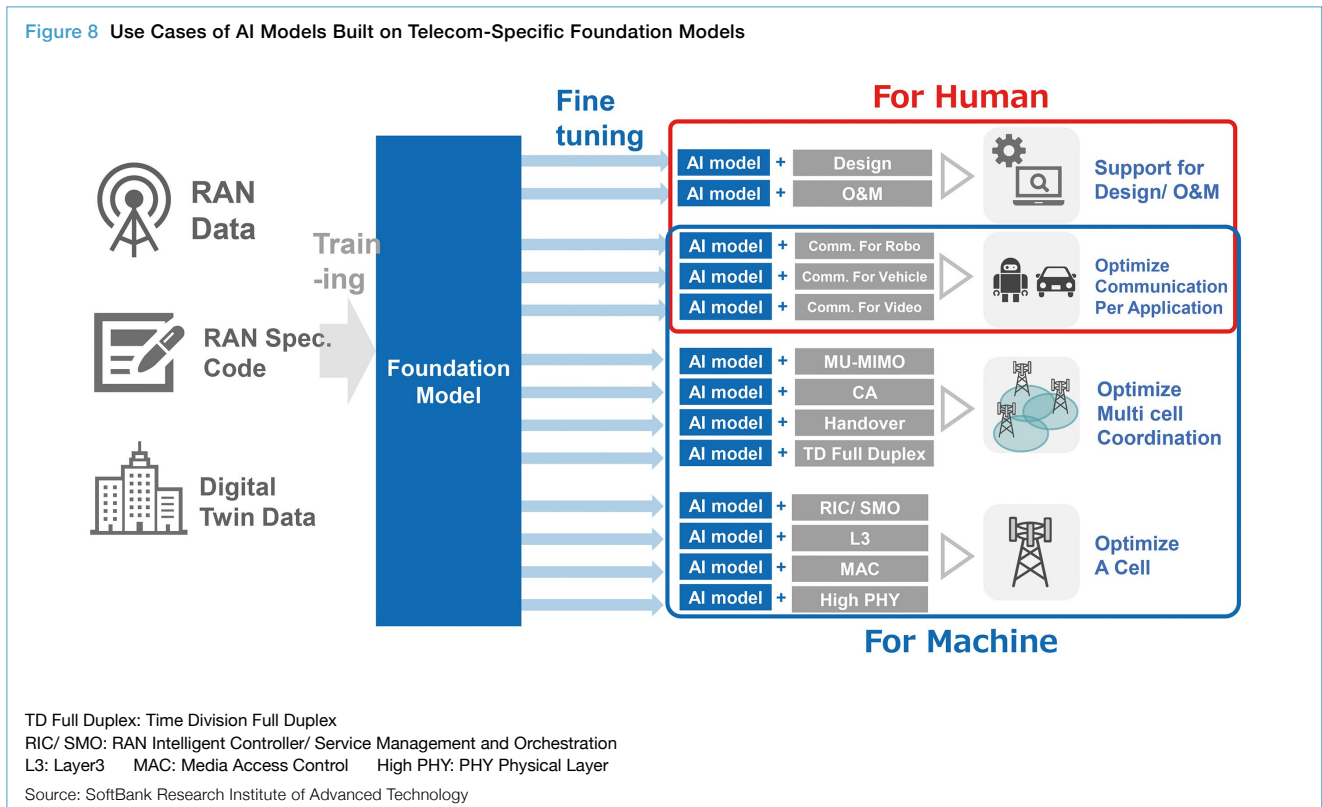
“If I want to operate 30 robots in this factory with low latency (200 milliseconds) while ensuring an effective communication speed of 20 Mbps per unit, I would ask where we should place the base station to create the optimal wireless network. Then, the AI agent provides an appropriate response, offering recommendations for optimal base station location.”

[Use Case 2] AI Models for Machines

The blue-boxed area in Figure 8 represents the domain of AI models designed for machines. In this field, processes such as base station adjustments are fully automated, with communication and optimization occurring directly between base stations and devices without human involvement.

Mr. Wakikawa explained this area as follows:

“We utilize foundation models to identify relationships between various numerical data points. This allows us to understand which KPIs and configuration parameters affect network operations. This knowledge helps the AI agent to derive and apply optimal configuration adjustments automatically.



Therefore, in this domain, AI does not even need to produce inference results in natural language.”

Why is AI Vital for Telecom Operators?

Mr. Wakikawa also emphasized the financial value of AI adoption for telecommunication operators in relation to Figure 8.

“From a different standpoint, the red-boxed area in Figure 8 is closely tied to OPEX (Operating Expenditure). By leveraging AI in this field—for customer support, training, troubleshooting, and setup assistance—operators can manage their infrastructure more efficiently.”

He also noted that reducing the number of operational sites and equipment during the design phase can help to lower operational costs such as land expenses, labor costs, and electricity usage.

Mr. Wakikawa further explained: “In contrast, the blue-boxed area in Figure 8 is closely related to CAPEX (Capital Expenditure). By improving spectral efficiency, we can reduce the demand for additional base stations and frequency allocations that would otherwise be required to accommodate increasing traffic. As illustrated in Figure 2, AI technologies such as machine learning and deep learning enable more effective allocation and management of radio frequencies. By maximizing spectrum utilization and eliminating waste, these developments ultimately contribute to cost reduction.”

Telecommunications systems operate as a

complex system where multiple elements interact. For example, changing the configuration of a single base station can affect surrounding base stations. If one station increases its output, it may interfere with adjacent stations, potentially reducing overall throughput. Traditionally, optimizing these parameters has relied on operator expertise, with engineers making incremental adjustments through trial and error to identify the optimal configuration.

The foundation model developed by SoftBank has been trained on comprehensive telecoms data. This enables the model to find relationships among various data points within the system. As a result, it can generate optimal solutions for wireless network operations. “SoftBank has deployed around 230,000 high-density wireless base stations throughout Japan, a staggering number even by global standards. Furthermore, our network quality is outstanding. We feel we have produced an extremely crucial foundation model by training it on this massive number of base stations, the enormous volume of traffic, and high-quality network data,” Mr. Wakikawa stated, underlining its effectiveness.

Three Key Elements of Foundation Models

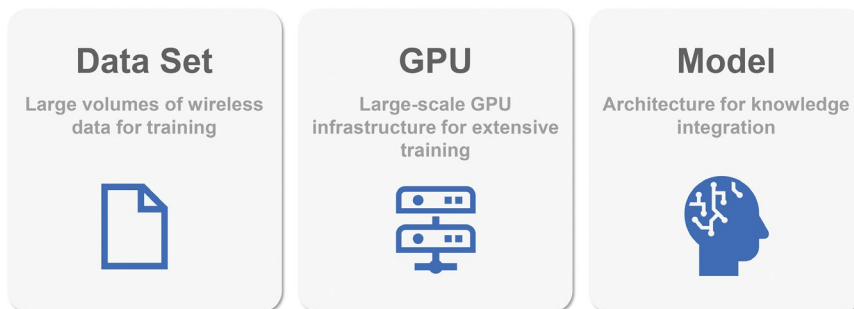
To develop a foundation model like this, three key elements are required (Figure 9).

The first is the dataset^[13]. A vast amount of wireless communication data is necessary for training. In general, foundation models for

▼ [13]

A dataset is a structured collection of data intended for processing by computers. In the context of AI development, a dataset specifically refers to an aggregate of data used for training and evaluating AI models.

Figure 9 Three Key Elements of Foundation Models



Source: SoftBank Research Institute of Advanced Technology

generative AI require massive datasets for training, and the quality of the dataset determines the accuracy of the AI model directly. Therefore, having high-quality, reliable data is crucial.

The second crucial element is the GPU (Graphics Processing Unit), which is essential for large-scale training. A single GPU or even a few are insufficient for this level of learning. SoftBank has a substantial number of GPUs, which makes it possible to do efficient large-scale training.

The third key element is the model. The model refers to the architecture designed to assimilate knowledge, and its structural design is critically important. Existing generic pre-trained models have already been trained on diverse datasets, making it difficult to significantly alter their existing weight priorities even when incorporating SoftBank’s vast telecoms data afterward. Consequently, it is essential to design an architecture from the outset that effectively learns and integrates SoftBank-specific data.

By integrating these three elements, SoftBank can finally develop the foundation model for telecommunications operators that it envisions (Figure 9).
“SoftBank plans to integrate this foundation into AITRAS. Traditionally, enhancing RAN performance required replacing base station hardware

with newer models or updating software. However, in the future, we aim to improve performance not just through these traditional methods, but also by expanding our foundation model to optimize network performance,” said Mr. Wakikawa.

Practical Use Cases of AI in Telecom

Figure 10 shows the specific use case SoftBank is currently developing.

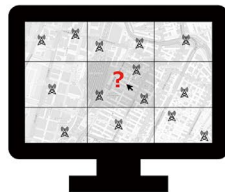
The upper section of Figure 10 illustrates a scenario in which a network operations manager is assessing the placement of a new 5G base station (gNB) using a geographic map. First, the operator asks, “What would be the optimal placement (configuration) of a new base station at this location?” In response, the AI agent processes the inquiry and provides an optimized configuration, stating, “The optimal placement for the new 5G base station is {A:1, B:2, ...}”

In the lower section of Figure 10, another scenario is illustrated where the operational manager asks, “If traffic increases, how should the existing 5G base station ‘X’ be reconfigured?” In response, the AI agent says, “The configuration should be adjusted to parameters like {A:1, B:2, ...}”

Foundation models can be used for a variety

Figure 10 The Specific Use Case SoftBank Is Currently Developing

Creating configuration of existing gNB

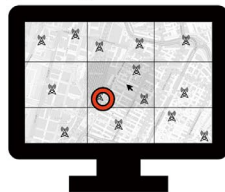


What would be the optimal config for the new gNB given the data?



The optimal config for the new gNB would be {A:1, B:2, ...}

Updating configuration of existing gNB



Can I optimize gNB X’s config to handle increased capacity demand?



Here is the list of configs to change: {A:1, B:2, ...}

gNB: gNodeB (5G base station)

Source: SoftBank Research Institute of Advanced Technology

of applications in addition to the ones that have been previously mentioned. For instance, they can resolve challenges that were previously unresolved, such as identifying trends in cell capacity utilization rates and establishing standard values for wireless communication parameters. These functionalities are highly valuable for telecommunications operators.

Challenges and Prospects for the 6G Era

In addition to the points discussed so far, there is another critical challenge: a significant portion of professionals in the telecommunications sector are on the brink of retirement.

Mr. Wakikawa added, *“There is a scarcity of next-generation expertise to succeed the industry veterans who have built wireless networks. This is a global issue, since the number of maintenance and network operations managers continues to decline. Experienced professionals in the field have firsthand knowledge and expertise in network design and operations, having previously managed these processes manually. However, these tasks are now largely automated. The next generation of professionals, having entered the industry in an era when networks are already automated (basically black-boxed), tend to have a limited understanding of the fundamental concepts underlying*

system operations. In reality, this operational expertise is becoming a critical issue for the future. While network reliability continues to improve, the challenge lies in closing the gap caused by the decline in operational management skills. Effectively supplementing this gap with AI will be essential for the future of telecommunications network operations.”

Telecommunications operators possess an invaluable and vast amount of network and operational management data. In the era of AI-driven network transformation, the ability to generate flexible datasets from this enormous volume of data is an essential challenge that cannot be overlooked.

Under these circumstances, SoftBank has developed a foundation model trained on vast amounts of telecommunications-related data. This foundation model aims to contribute to RAN operations design and infrastructure reconstruction through adaptable AI models that can address a wide range of use cases. This is supposed to hasten the AI-native change of the whole telecom industry.

Mr. Wakikawa concluded, *“The development of foundation models will bring benefits to both telecommunications equipment vendors and operators.”*

We look forward to the emergence of an AI-native RAN world in the coming 6G era.

Profile



Photo: Hiroyuki Matsumoto

Ryuji Wakikawa, Vice President, Head of Research Institute of Advanced Technology, SoftBank Corp.

He is also a Project Professor at the Global Research Institute at Keio University, Japan, where he received his Ph.D. in 2004. In 2007, he was awarded the Ericsson Young Scientist Award.

He focuses on the development of advanced technologies, architecture design, and shaping future visions. The research areas are 5G Advanced, 6G, High Altitude Platform Stations (HAPS), Non-Terrestrial Networks (NTN), LLM/Transformer, AI-RAN, quantum technologies, and autonomous driving technologies.

References

- AI-RAN Alliance, <https://ai-ran.org/>
- <https://www.softbank.jp/en/corp/technology/research/news/064/>
- <https://www.softbank.jp/en/corp/technology/research/>

What Is “AITRAS,” a Converged AI-RAN Solution?

▼ [1]

Carrier-grade refers to a high standard that meets the stringent requirements for network deployment and operation by telecommunications operators.

▼ [2]

GPU (Graphics Processing Unit) is a processor designed for processing 3D graphics and video. It is capable of processing large amounts of data at high speed. GPUs have also been utilized for large data analysis and AI computations in recent years.

▼ [3]

Single User MIMO (SU-MIMO) is a MIMO communication method in which the transmitting and receiving sides always maintain a 1:1 relationship.

▼ [4]

L1 refers to the physical layer (Layer 1) in the OSI reference model within the vRAN (virtualized RAN) software architecture. Similarly, L2 corresponds to the data link layer (Layer 2), and L3 corresponds to the network layer (Layer 3). OSI (Open Systems Interconnection) is a seven-layer reference model that organizes the functions required for communication across diverse computer systems into hierarchical layers.

AITRAS: Integrating RAN and AI on a Single Computer Platform

Following the announcement of the AI-RAN concept, SoftBank unveiled “AITRAS” as its “next move.” AITRAS is a converged solution allowing RAN (Radio Access Network) and AI (Artificial Intelligence) to operate on the same computer platform. It delivers carrier-grade^[1] RAN with high capacity, high performance, and high quality. At the same time, it is being developed to efficiently process and manage AI applications like generative AI. With this product, SoftBank aims to provide not only communication services but also various AI-enhanced communication services and standalone AI processing services.

The system configuration of AITRAS is shown in Figure 1. It is built on the newly designed NVIDIA GH200 Grace Hopper Superchip for large-scale AI processing. On this platform, MPS (Multi Process Service) for parallel processing and MIG (Multi Instance GPU^[2]) for virtualization are constructed. Above these lay-

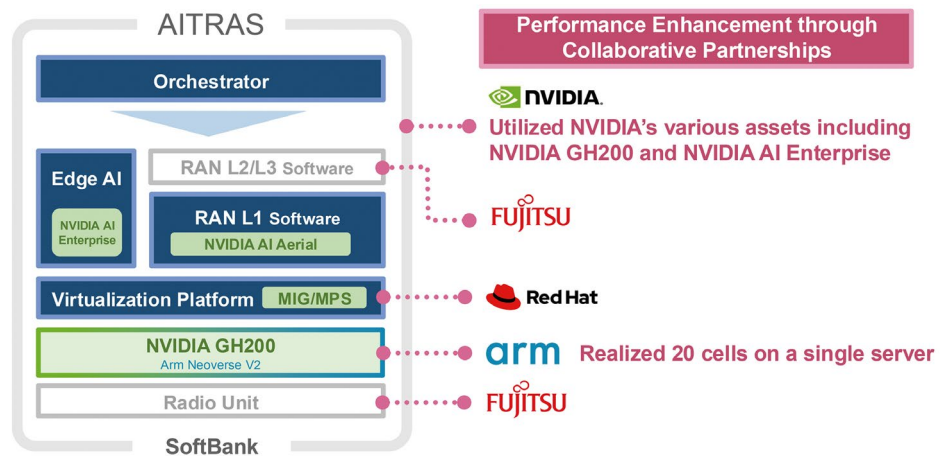
ers, the L1 to L3 radio signal processing software and the Edge AI service platform are deployed. Both wireless communication and AI services are managed and controlled by an orchestrator, which automates their operation and administration.

SoftBank primarily develops the navy blue-colored components depicted in Figure 1, while the remaining parts are jointly developed in collaboration with partner companies.

RAN Test Environment Deployed and Operational at University Campus

The AITRAS demonstration environment has been built at Keio University’s Shonan Fujisawa Campus (SFC) in Kanagawa Prefecture and is operated in concert with resources at SoftBank’s Takeshiba headquarters in Tokyo (Figures 2 and 3). This setup includes an experimental local 5G network operating in the 4.8–4.9 GHz frequency band with a maximum bandwidth of 100 MHz. It supports a maximum of 4 layers of Single User MIMO^[3] and features a 4T4R (four transmit

Figure 1 AITRAS System Elements



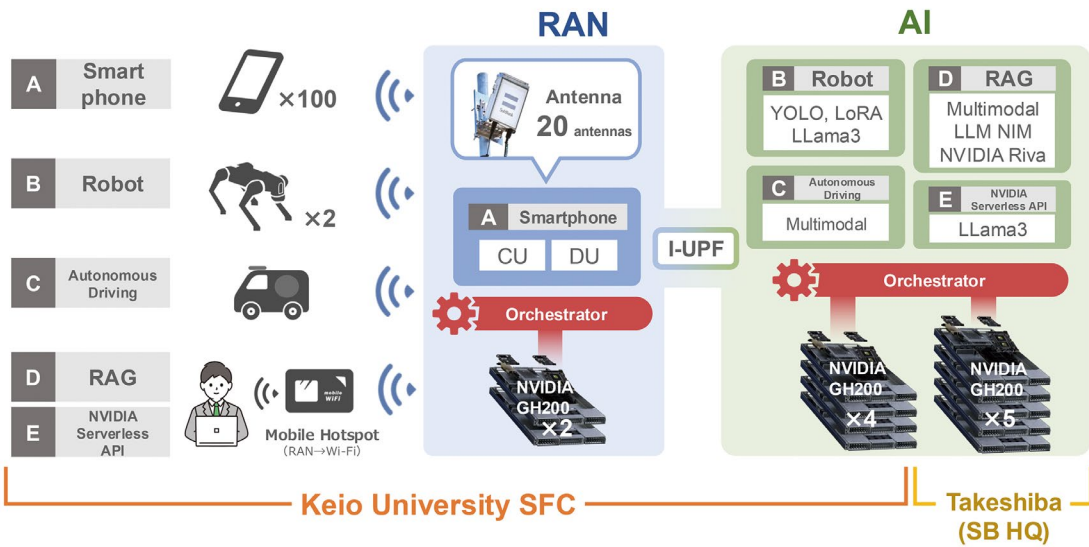
Source: SoftBank Research Institute of Advanced Technology

antennas, four receive antennas) configuration with 20 antennas processed by a single NVIDIA GH200 Grace Hopper Superchip server. However, since the L2 and L3 protocol layers^[4] of the RAN require parallel processing, two servers

featuring the NVIDIA GH200 Grace Hopper Superchip are deployed for RAN operations.

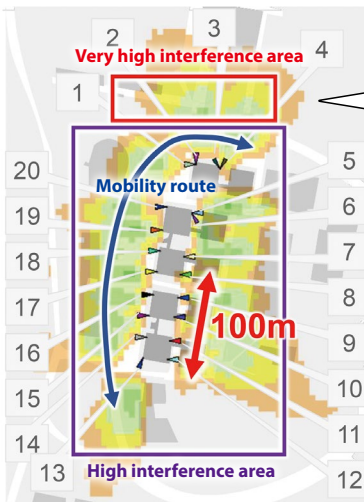
As shown in Figure 3, antennas at SFC are intentionally arranged in an interference-prone configuration to replicate urban conditions. In

Figure 2 AITRAS System Architecture Overview



CU: Control Unit, controls the operation of the CPU.
 DU: Data Unit, performs arithmetic and logic operations on data.
 YOLO: You Only Look Once, an image recognition algorithm for detecting objects in an image.
 LoRa: Deep Learning Model, a method to tune AI for image generation
 Llama3: An open source large-scale language model developed by Meta, Inc.
 NIM: NVIDIA Inference Microservice, a service for deploying inference environments for AI models in production environments
 NVIDIA Riva: A service for building multilingual language/translation AI
 Source: SoftBank Research Institute of Advanced Technology

Figure 3 Outdoor Testbed for AITRAS at Keio University Shonan Fujisawa Campus (SFC)



* Coverage of the roadway for mobility evaluation.
 * The very high interference area (upper part of the figure) is intentionally reproduced for evaluation of the emphasis function between base stations.



Base station clusters arranged in high density to simulate an urban area



Base station antennas



AITRAS Server with NVIDIA GH200

[RAN Overview (Main Specifications, As of Nov. 2024)]

Supported Frequency Bands	4.8GHz~4.9GHz (TDD)
Supported Bandwidth	Max. 100MHz bandwidth
Single User MIMO	Max. 4 layers
Antenna	4T4R
# of Cells on a Single Server	20 cells/NVIDIA GH200
EIRP	Max. 46.2dBm

TDD: Time Division Duplex
 Single User MIMO: A Multiple Input Multiple Output (MIMO) communication method in which the transmitter and receiver always have a 1:1 relationship.
 4T4R: 4T (4 Tranceiver) / 4R (4 Receiver)
 EIRP: Equivalent Isotropic Radiated Power (or Effective Isotropically Radiated Power)
 dBm: Decibel milliwatt. Unit of power for radio waves.

Created by the editorial team based on materials provided by SoftBank Corp.

▼ [5]

Large Language Models, or LLMs, are a type of language model that is scaled up in three key aspects: "computational volume" by computers, the "amount of training data" for learning, and the "number of parameters" that define the model's complexity. A "language model" is one that quantifies the human language (natural language) used in everyday life so that computers can understand it. Large-scale language models achieve significantly higher performance compared to traditional language models due to their large scale.

▼ [6]

Containerization is a virtualization technology that arranges server environments to efficiently manage application processing and other functions.

this setup, 100 smartphones are used to perform simultaneous access, allowing for analysis of traffic conditions and power consumption.

Mr. Wakikawa said of the current development situation, "We are making adjustments to the extreme in order to achieve the carrier-grade standards we are aiming for in terms of stability, communication capacity, and power consumption."

The "RAN part" introduced so far and the "Edge AI part," which will be discussed later, are directly connected via I-UPF (Intermediary User Plane Function, a data transfer layer; see Figure 2) to provide functions and performance tailored to the respective needs of communication and AI.

Edge AI Handles AI Processing

The servers equipped with NVIDIA GH200 Grace Hopper Superchip for AI service processing are utilized as a testing environment, with four units deployed at SFC and an additional five units at SoftBank's headquarters in Takeshiba.

The Edge AI responsible for AI processing in this setup is compatible with NVIDIA AI Enterprise (Figure 4). NVIDIA AI Enterprise is

a software platform designed for AI processing, supporting tasks such as the development of large language models (LLMs)^[5].

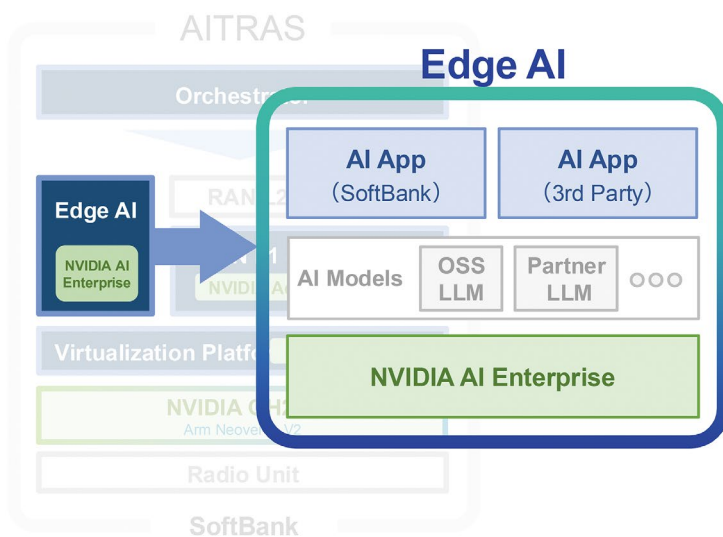
Moreover, the system features the NVIDIA Serverless API, which enables dynamic allocation of GPU resources as required. By leveraging this API to containerize^[6] GPU resources and provide them to AI services, the system can efficiently accommodate the temporary AI needs from enterprises.

AITRAS Reflects SoftBank's Core Commitment

(1) Low-Power Arm Neoverse V2 Processor

Although the GH200's nominal power consumption is 1,500W, SoftBank's measurements indicate that its actual power consumption is approximately 25W per cell. Consequently, even when controlling 20 antennas with a single GH200, the total power consumption is maintained around 500W (25W × 20 antennas). This energy efficiency is largely attributed to the Arm Neoverse V2 processor. According to SoftBank's power consumption study, the Arm Neoverse V2 processor consumes approximately half the

Figure 4 AITRAS: Edge AI Overview



Original AI Service

- Autonomous driving monitoring app
- Cloud robot
- RAG Menu @edge, etc.

NVIDIA Serverless API

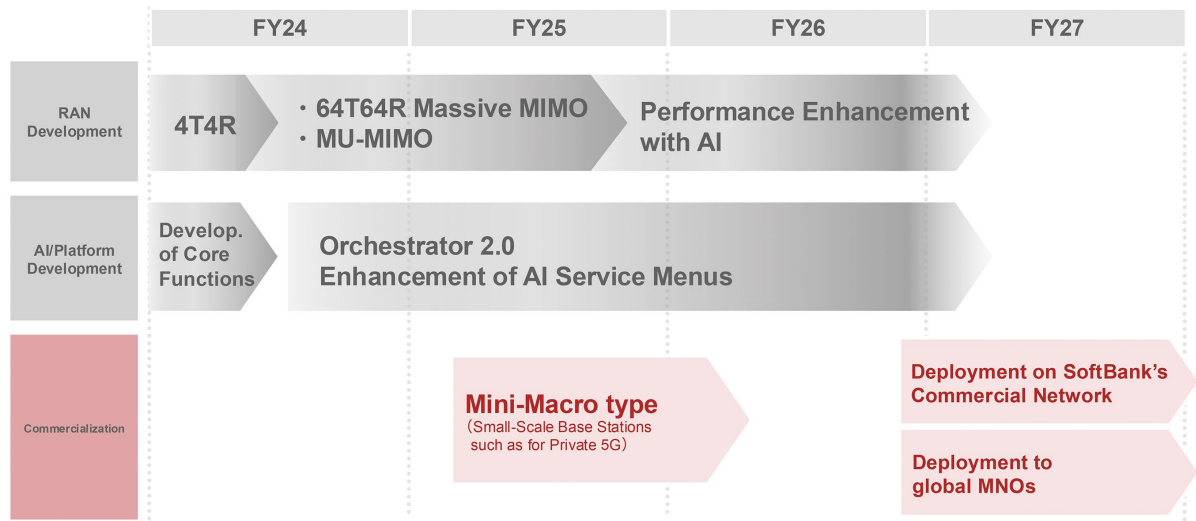
Spot allocation of available resources based on customer demand

NVIDIA AI Enterprise

Industry-standard AI framework

Source: SoftBank Research Institute of Advanced Technology

Figure 5 AITRAS Development Roadmap



Source: SoftBank Research Institute of Advanced Technology

power of the comparable CPU.

(2) Inter-Base Station Control for Interference Mitigation

AITRAS performs all signal processing entirely in software, with all signals centrally processed by the GH200 at the central base station. This approach is designed to minimize excessive interference between antennas.

(3) AI Orchestrator Targeting 100% Utilization Rate

AI-RAN incorporates the concept of "AI-and-RAN," which involves reallocating idle data center processing capacity—such as during off-peak hours at night—to AI workloads. This approach aims to ensure that telecommunication infrastructure operates at full capacity at all times. The key technology SoftBank is developing to enable this is the AI Orchestrator.

AITRAS: Commercial Launch on SoftBank Network in FY2026

The greater demand for AI services is expected to be primarily in high-density urban areas. In terms of frequency bands, the focus is on the capacity band^[7] for 5G TDD. SoftBank plans to deploy AITRAS in this frequency band by fiscal year 2026, with a phased development to other frequency bands to follow.

The commercialization of AITRAS is initially being prepared for deployment in mini-macro type for private 5G networks. A key feature of AITRAS is its ability to offer a converged solution that combines the radio access network (RAN) and GPU servers.

Following this initial rollout, SoftBank plans to expand deployment to its commercial network and international mobile network operators between 2026 and 2027 (Figure 5).

▼ [7]

A capacity band is a frequency range utilized in security of communication capacity. The 1,000MHz (1GHz) frequency band is a frequency band that is used for high-capacity multimedia mobile communications.

References

- https://www.softbank.jp/en/corp/news/press/sbkk/2024/20241113_06/
- <https://www.softbank.jp/en/corp/technology/research/news/052/>
- <https://www.softbank.jp/en/corp/technology/research/story-event/069/>



SoftBank

R&D

