Copyright© 2024 SK Telecom. All Rights Reserved.

SK telecom

SK Telecom 6G White Paper

View on Future AI Telco Infrastructure



ICT Infra

Copyright© 2024 SK Telecom. All Rights Reserved.

SK Telecom 6G White Paper

View on Future AI Telco Infrastructure

ICT Infra

SK telecom

Contents

[Executive Summary] 03

1. Introduction 04

2. Considerations on 6G Architecture Design 06

2.1 Observations on the 5G Vision 07
2.2 Overarching Aspects of the 6G Framework, Ubiquitous Intelligence 08
2.3 AI: Catalyst for the Next Generation of ICT Evolution 09

3. Perspective on AI Telco Infrastructure 11

3.1 AI Company Vision 11
3.2 Evolution of Wireless and Wired Networks: AI Telco Infrastructure 12

4. Telco Edge AI 14

4.1 Why Telco Edge AI? 14
4.2 What is Telco Edge AI? 16

5. Evolution Directions of Telco Edge AI 20

5.1 Ecosystem Creation 20
5.1.1 Operator/Vendor Cooperation 20
5.1.2 Standardization and Academic Research Activities 21
5.2 Business Model Discovery 23
5.3 Key Technologies 25
5.3.1 Radio Access Network 25
5.3.2 Core Network 27
5.3.3 Transport Network 28
5.3.4 Device 31
5.3.5 AI orchestration 32

6. Closing Remarks 36

References 37

Abbreviations 38

Change History 39

Executive Summary

[Introductory Words]

- Telco's business model, represented as a cycle, increased traffic demand → revenue generation → infrastructure investment, has stagnated in growth, necessitating new monetization efforts through architectural innovation of AI-based telco infrastructure to re-rate the value of telco infrastructure.

[Consideration on 6G Architecture Design]

- While 5G was designed to support a broader range of usage scenarios, such as enhanced Mobile BroadBand (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC), it primarily focused on enhancing performance of communication connectivity itself.
- The 6G Framework Recommendation emphasizes the importance of ubiquitous intelligence as an overarching aspect for 6G, and it is vital to consider a balance between connectivity and AI when designing the 6G architecture.

[Perspectives on AI Telco Infrastructure]

- SK Telecom has unveiled its AI pyramid strategy as its vision to transform into an AI company, outlining three key layers: ▲ AI infrastructure (1st floor) ▲ AIX (2nd floor) and ▲ AI services (3rd floor).
- Within the AI infrastructure, the AI Telco Infrastructure will deliver ▲ AI-Assisted Network that enhances network efficiency, and ▲ AI-Enabling Infra $^{\text{Telco Edge AI}}$ that creates new business models and values.

[Telco Edge AI]

- The future telco infrastructure will evolve toward Telco Edge AI, which provides both connectivity and AI computing, offering differentiated value by combining the strengths of both large-scale AI datacenter (AI DC) and On-Device AI.
- Telco Edge AI infrastructure will be established by leveraging the existing spaces and facilities owned by the telco and introducing new equipment. Its key components include ▲ Co-location ▲ AI-Server and ▲ AI-RAN.

[Evolution Direction of Telco Edge AI]

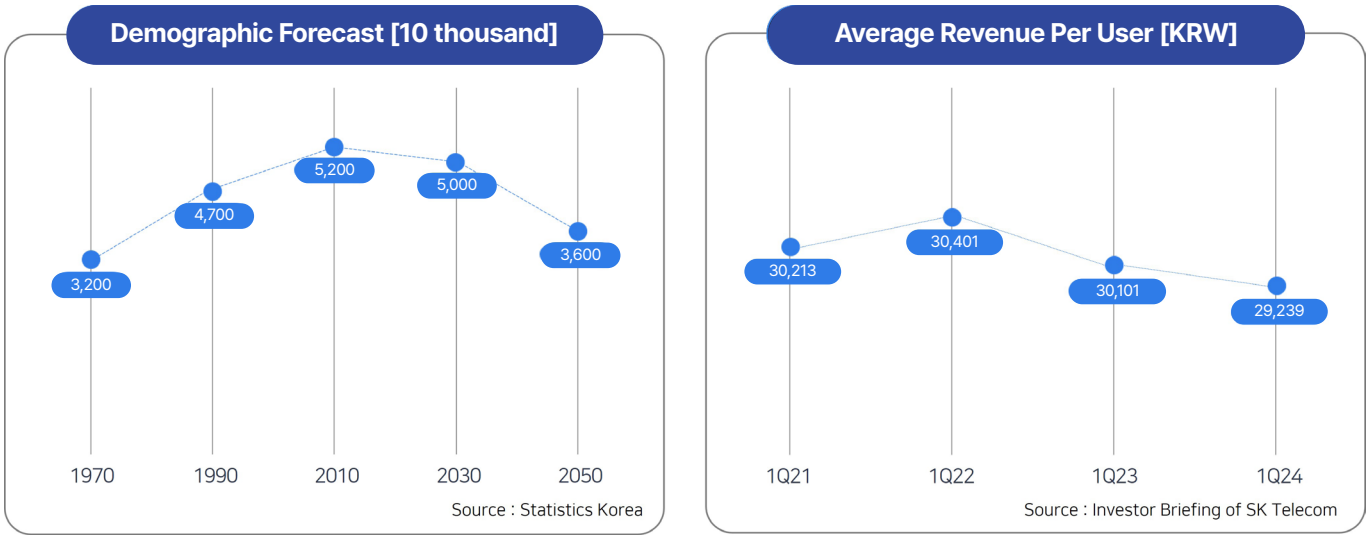
- For the successful evolution of Telco Edge AI, it is of utmost importance to develop and expand a global ecosystem that benefits vendors, operators, and customers, and to discover promising business models.
- SK Telecom plans to outline detailed implementation strategies for key technology areas, including radio access networks, core networks, transport networks, devices, and AI orchestration, and to secure necessary technologies to bring these plans to fruition.

[Concluding Remarks]

- To successfully evolve into Telco Edge AI, we aim to enhance the sustainability of the telco business by building meaningful success stories through cooperation with various partners.

1. Introduction

The mobile telecommunications industry truly began to flourish after the commercialization of 3G, followed by the emergence of smartphones and the launch of LTE in the late 2000s. The arrival of the mobile internet era served as a catalyst for the industry's growth, but at the same time, it raised concerns that mobile telecommunication networks might become mere "dumb pipes". For instance, the launch of Skype in 2003, the advent of messaging services like WhatsApp launched in 2009, and the explosive increase in mobile data traffic in the mid-2000s due to video streaming services like Netflix and YouTube has posed major challenges to the mobile network ecosystem. Particularly in South Korea, securing new customers is becoming increasingly difficult due to population decline, while average revenue per user (ARPU) is stagnating. Therefore, if we only consider the traditional generational change model for mobile network evolution, there is a concern that investment capacity will be insufficient and the mobile network ecosystem will shrink accordingly.



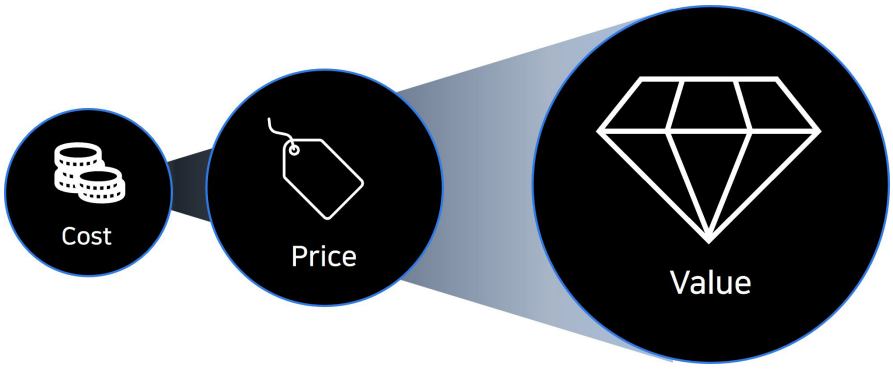
[Figure 1] Demographic Forecast and ARPU for SK Telecom

To overcome these challenges, there are ongoing attempts to expand the existing business-to-customer (B2C)-focused mobile telecommunications service market into various vertical domains. Under the catchphrase "Beyond Connectivity", efforts are being made to enter adjacent information and communication technology (ICT) areas utilizing telecommunications, such as cloud services, smart city solutions, Internet of Things services, digital advertising and marketing, and to collaborate with various partners in various fields to pioneer new businesses including autonomous driving and Urban Air Mobility (UAM).

As the world evolves at an unprecedented pace, the demand for more advanced and efficient communication is increasing. While 5G has provided the foundation for mobile communications to accommodate a wider range of usage scenarios, such as enhanced Mobile BroadBand (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC), compared to LTE only emphasizing only broadband mobile communications, it is true that 5G is still primarily focused on evolving the technical

performance of the mobile network itself, such as data rate or capacity, and does not deviate much from the goals pursued by previous generations, such as 3G and LTE [1]. Meanwhile, the entire ICT industry is expected to demand more advanced communication infrastructure capable of providing higher levels of performance and accommodating a wider range of usage scenarios [2]. In addition, research and investment are on the rise to accelerate in applying AI to network infrastructure and utilizing their resources to provide AI services, to meet various needs, such as improving network efficiency and sustainability, as well as generating new revenue by utilizing the existing space and computing resources of telco network infrastructure.

As such, the evolution direction for telcos is the pursuit of infrastructure evolution and AI business monetization. To achieve this, it is necessary to create customer value and new business models by incorporating innovative AI technologies into telco infrastructure. In other words, it is crucial to provide value that goes beyond the cost of deploying and operating infrastructure and the price paid by customers. To make this possible, it is very important to significantly increase the value perceived by customers from the perspective of the next-generation infrastructure evolution.



[Figure 2] Objective of Telco Infrastructure (Cost < Price < Value)

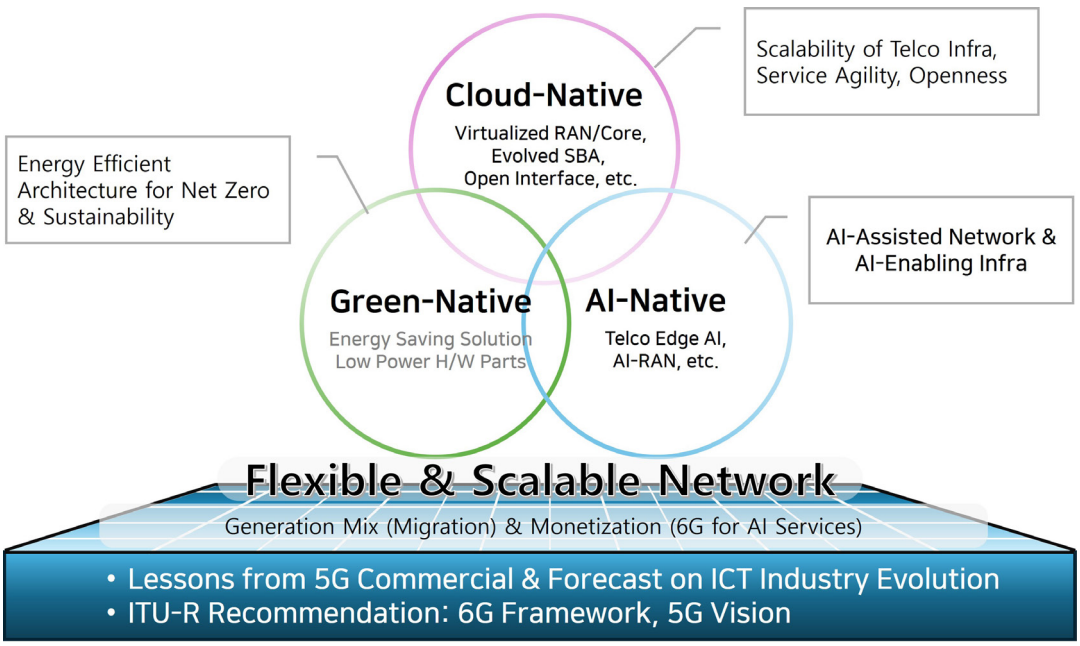
In particular, AI services triggered by generative AI, recently led by ChatGPT, are expected to provide high value for cost as a full-fledged productivity improvement tool that go beyond the level of personal assistant applications in daily life, and are being introduced in various fields, such as offices, laboratories, and industrial sites. Accordingly, while existing hyperscalers are accelerating competition to provide differentiated value from existing data centers by building AI DC equipped with high-performance chipsets specialized for learning and inference required for AI services, the mobile communication industry is utilizing its large-scale network infrastructure only for communication services (connectivity).

In this white paper, we aim at presenting the evolution direction toward Telco Edge AI, which can create new revenue streams and re-rate the value of network infrastructure by innovating the network architecture currently used for limited purposes. This evolution builds on the AI-Native nature of 6G networks, a critical aspect of future mobile networks [1], by providing differentiated value at the edge, at the customer touchpoint, compared to AI DCs.

2. Considerations on 6G Architecture Design

The International Telecommunication Union Radiocommunication Sector (ITU-R) approved the ITU-R M.2160 document in Nov. 2023, which is a recommendation that sets out the framework and overall objectives for future 6G development [3]. This recommendation is the first milestone in the official 6G standardization activities, as the technical performance requirements of the 6G network are set according to the usage scenarios and capabilities that should be accommodated in the 6G era, and it serves as a reference for standards development organizations, such as 3GPP to develop practical 6G standards.

SK Telecom is currently investigating crucial elements of 6G architecture design as we evolve our existing networks into 6G. As shown in the below figure [3], considering the overall relationship with the 5G vision and 6G framework recommendations, lessons learned from the currently operating 5G, and prospects for future network development, we are conducting the research on the evolution towards Flexible & Scalable Network as the foundation for Cloud-Native, Green-Native, and AI-Native, all of which are important elements that should be reflected in the architecture design of the 6G [1].



[Figure 3] Key Considerations in Designing an Overall 6G Architecture

First, Cloud-Native is a design principle, that should be considered when designing a next-generation network infrastructure, to flexibly respond to rapidly changing requirements and to achieve high availability, scalability, and operational cost reduction by providing communication services based on cloud technology away from the existing physical communication network environment.

Second, Green-Native is a design principle, that should be reflected in the future network infrastructure, to ensure the sustainability of the mobile communication business, as well as participate in reducing carbon emissions to respond to climate change which is a global hot topic. The technical factors to be considered in the radio access network and core network have been described in the 6G white paper published in 2023 [1].

Third, AI-Native is a design principle that should be considered in designing the next-generation network infrastructure to secure the performance, efficiency, and stability of the network by integrating AI across all areas of the network, such as radio access networks, core networks, and transport networks [1], [4]. Rather than simply adopting AI partially, this involves leveraging AI to make the network more intelligent, thereby improving customer experience, maximizing operational stability and efficiency based on AI across all areas of network, and creating new revenue opportunities by utilizing existing telco infrastructure assets to provide AI services.

Meanwhile, we are focusing on flexible and scalable network architecture as the fundamental principle for 6G architecture design, which is the basis of the above three elements. Generation Mix, one of the keywords for this, is an important topic of research for SK Telecom from a reflection on Generation Change, which aimed at a clean slate architecture that focused only on improving the technical performance of next-generation mobile communications and lacked consideration of the previous generation. Generation Mix is a key consideration factor for enhancing the sustainability of the mobile communication business through cost-effective investment by enabling smooth migration through coexistence with existing network systems as well as demand-based deployment to build infrastructure based on demand for 6G services.

This white paper presents our views on Telco Edge AI, which we are focusing on as an essential element of the evolution to AI-Native among the critical elements of 6G network architecture.

2.1 Observations on the 5G Vision

The 5G Vision Recommendation published by the ITU-R in 2015 laid out the framework and objectives for the development of mobile communication systems to be put into operation after 2020, with three use scenarios (eMBB, URLLC, and mMTC) and key performance indicators as important factors [5]. However, such key performance indicators are still limited to defining the performance objectives of traditional mobile communication itself.

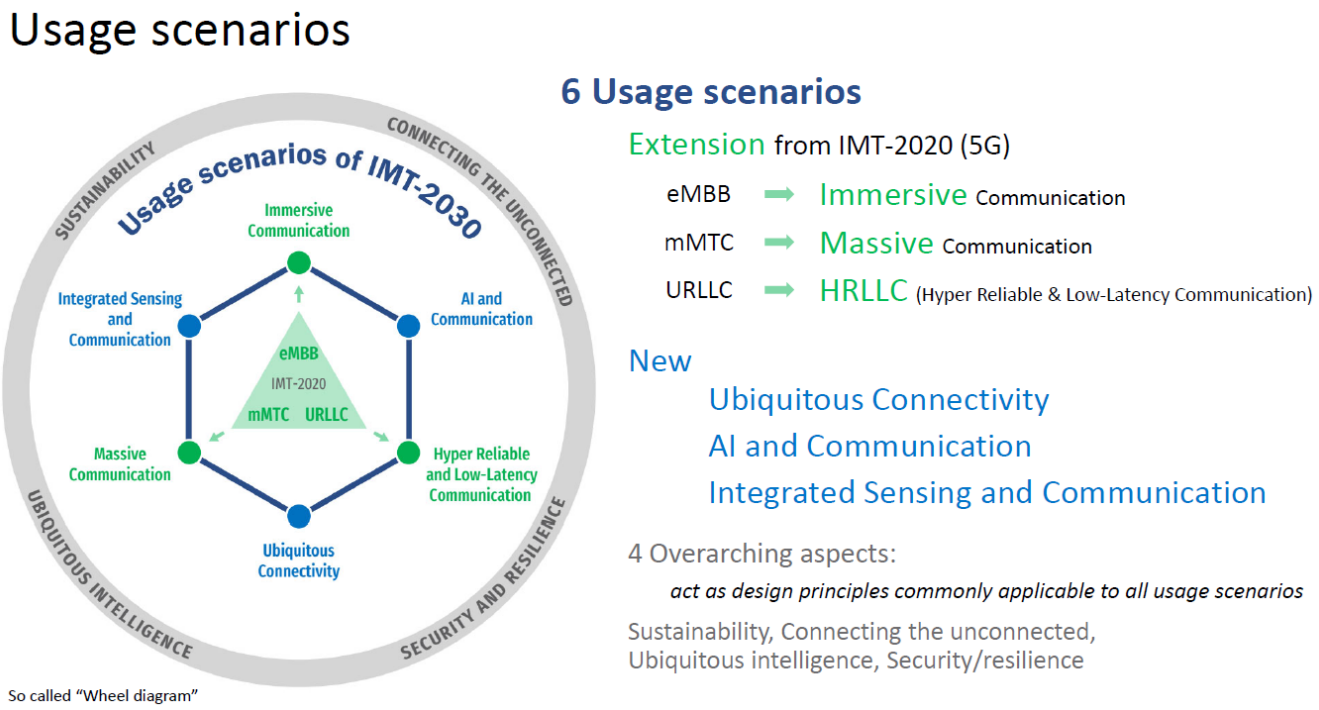
On the other hand, the vision of the future after 2020, which was predicted when the 5G vision was announced in 2015, is changing at a faster pace than predicted. The historic Go match between AlphaGo and Lee Sedol in 2016 marked a turning point, demonstrating the potential of AI based on deep learning and reinforcement learning. In contrast, the emergence of ChatGPT, a generative AI based on large language

models in 2023, is leading a revolutionary change to the extent that AI is now reshaping the foundations of the ICT ecosystem paradigm [6]. AI is already upon us at a pace far beyond everyone's expectations when the 5G vision was established.

Ahead of the 5G rollout, the mobile communication industry focused only on improving the performance of the mobile communication itself, which may be an inevitable consequence due to the fact that such rapid change was not expected. However, we should prepare with the attitude that it is not too late to seriously think about the evolution direction of the new mobile communication network infrastructure in line with the era of AI, which is already in front of all of us.

Therefore, in order to overcome the limitations of the current 5G in evolving into an AI-Native architecture, we will briefly review the direction proposed in the 6G Framework Recommendation, and through this, we will examine aspects related to AI as important characteristics that the 6G network should have in the future.

2.2 Overarching Aspects of the 6G Framework, Ubiquitous Intelligence



[Figure 4] Usage Scenarios and Overarching Aspects of IMT-2030

* Source: ITU website: IMT towards 2030 and beyond (itu.int)

First, we take a look at a usage scenario that can intuitively check the comparison with the 5G vision in the 6G framework recommendation approved by the ITU-R in Nov. 2023 [3].

The usage scenarios for 6G presented by the ITU-R include three scenarios that have been expanded from 5G, and three scenarios have been newly added, among which AI and communication scenarios can be noted. More notably, four overarching aspects are proposed as design principles that will be common to all six usage scenarios, including how ubiquitous intelligence should differentiate 6G networks from 5G.

Here's a quick overview of the ubiquitous intelligence described in the recommendation:

With the steady progress and fast spread of technologies in AI and particularly machine learning (ML), it is expected that intelligence would be present in every part of the communication system to support the building of smart cities and communities. Future connected devices may become fully context-aware for more intuitive and efficient interactions among humans, machines, and the environment. Possible autonomous management of networks by AI/ML may also be capable of, to some extent, performing self-monitoring, self-organization, self-optimization and self-healing without human intervention. It is expected that the air-interface would be enhanced by AI models.

IMT-2030 could serve as an AI-enabling infrastructure capable of providing services for intelligent applications. AI-enabled air interface along with distributed computing and intelligence could potentially allow for end-to-end AI applicability and the convergence of communication and computing. These systems would have functions of inferences, model training, model deployment, as well as computing distributed across networks and devices.

In summary, it is explained that it has two main objectives: intellectualization/efficiency of network infrastructure using AI, and network infrastructure that can provide AI application services. In other words, the evolution of AI-Assisted Network and AI-Enabling Infra, which will be discussed hereafter in this white paper, should be one of the key elements of the 6G network architecture.

2.3 AI: Catalyst for the Next Generation of ICT Evolution

AI is expected to play an important catalytic role in the next generation of ICT. AI-Native networks are a form in which AI is deeply integrated into the management and operation of telecommunications networks,

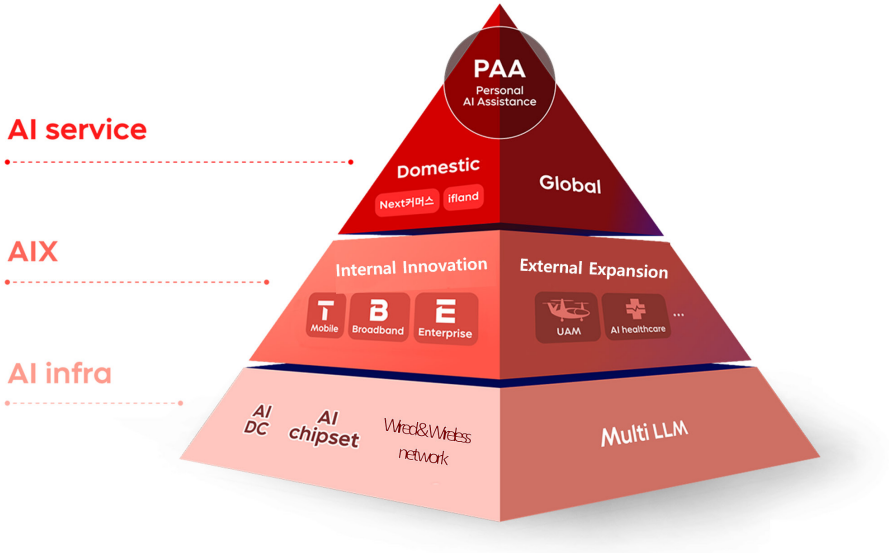
enabling autonomous and intelligent network management. Specifically, AI-Native networks can include functions, such as autonomous network management, intelligent traffic management, prediction and analytics, network security enhancement, and energy efficiency management. AI monitors network status in real-time and optimizes network performance through anomaly detection and automatic adjustments. In addition, AI algorithms efficiently distribute data traffic to minimize congestion and ensure quality of service (QoS). AI can predict network traffic patterns, proactively identify and respond to potential issues, and strengthen cyberattack detection and response systems to identify and block threats in real time. AI also optimizes the energy consumption of the network, making it more sustainable. Due to the diverse applications of AI, leading ICT companies worldwide have recognized the importance of AI-Native networks and are actively pursuing investments and development in this area.

According to the market research firm Value Market Research, the demand for AI in the mobile communications market is estimated to reach $20 billion by 2030 [7]. It is expected to grow exponentially by a factor of 10 from the $2 billion surveyed in 2022. The global mobile communication network equipment market is estimated to be $48 billion by 2024, of which the network equipment market for wireless access networks accounts for a very large share of $36 billion [8]. Therefore, if the current network equipment is replaced by AI-Native networks, it will have a huge impact that will change the game of the next-generation ICT market.

In conclusion, AI is a key element of the next-generation communications infrastructure, playing a vital role in a variety of areas, including autonomous network management, performance optimization, operational cost reduction, and user experience improvement. The evolution to AI-Native networks will maximize the efficiency and reliability of communication networks and enable innovative ICT services and applications in the future.

3. Perspective on AI Telco Infrastructure

As AI emerges as a key technology determining the future of businesses, SK Telecom has established an AI company vision to leap forward as a global AI enterprise in response to this change. This chapter presents SK Telecom's perspective on the AI Pyramid Strategy, which is at the core of the AI company vision, and the evolution towards AI infrastructure, which forms the foundation of this strategy.



[Figure 5] SK Telecom's AI Pyramid Strategy

3.1 AI Company Vision

In Sep. 2023, SK Telecom unveiled its blueprint to become a global AI leader through the 'AI Pyramid Strategy', which aims to advance AI technology. The AI Pyramid Strategy plans to enhance AI technology, strengthen customer relationships through AI services, and execute through collaboration centered on AI alliances. The AI Pyramid Strategy, consisting of three main areas - AI Infra, AIX (AI Transformation), and AI Services - is closely linked to SK Telecom's core businesses, aiming to innovate across all areas of industry and life.

1) AI Infra
AI Infra is the area where SK Telecom's technological capabilities are concentrated, including Telco LLM (Large Language Model), AI wireless and wired networks, and AI DCs. SK Telecom is pursuing the development of telecom-specific AI models and cooperation through the Global Telco AI Alliance (GTAA). Additionally, it aims to enhance network efficiency and stability by applying AI to wireless and wired infrastructure through investment and cooperation in AI DCs.

2) AIX (AI Transformation)
AIX is responsible for innovating productivity and customer experience by applying AI to SK Telecom's main business areas. It is streamlining business operations in mobile, broadband, enterprise solutions,

and other areas using AI technology, while also discovering new growth engines. AIX is being realized through innovations, such as AI-based customer centers, the introduction of personalized broadcasting service 'AI Btv', and the provision of customized AI solutions for enterprises. Furthermore, SK Telecom is actively pursuing AI-driven innovation in Urban Air Mobility (UAM) and AI healthcare sectors.

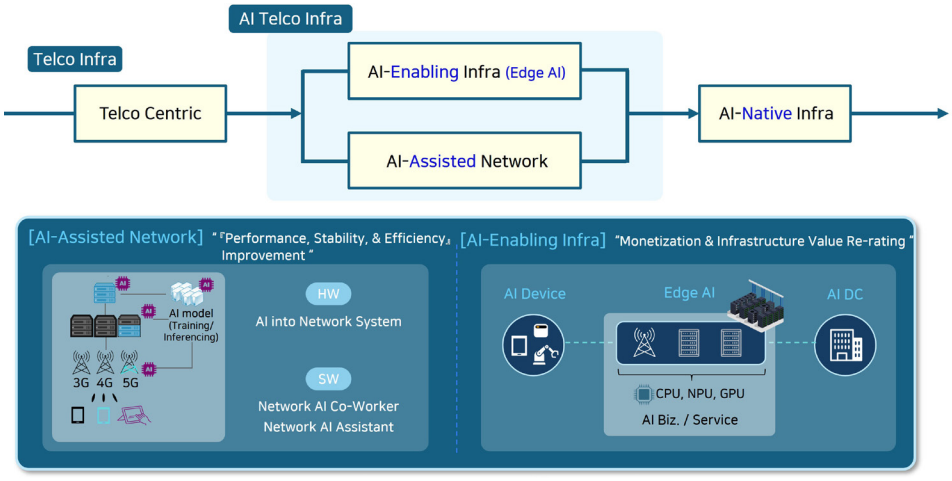
3) AI Service

The AI service is a strategy to secure competitiveness in the global market based on SK Telecom's AI technology. SK Telecom is innovating user experience through its personal AI assistant (PAA) 'A. (A dot)' and aims to expand this service to the global market to broaden its presence in the AI agent field and secure leadership in the global AI market.

To successfully implement the AI Pyramid Strategy, SK Telecom has established 'T.H.E. AI' as its AI governance principle. 'T.H.E.' stands for 'by Telco, for Humanity, with Ethics', embodying principles that ensure AI technology interconnects customers and stakeholders (Telco), aims for diversity, inclusion, and the advancement of human welfare (Humanity), and guarantees transparent and ethical AI decision-making processes (Ethics). These AI governance principles provide guidelines for ensuring reliability and stability across SK Telecom's AI business, through which SK Telecom aims to develop trustworthy, transparent, and safe AI.

3.2 Evolution of Wireless and Wired Networks: AI Telco Infrastructure

SK Telecom is pursuing evolution towards AI wireless and wired infrastructure, moving beyond traditional infrastructure dedicated solely to communication services, with the ultimate goal of transitioning to AI-Native Infrastructure. This infrastructure evolution requires a different approach from the past, seeking transformation of both telco infrastructure and core AI infrastructure. AI Telco Infrastructure can be broadly categorized into AI-Assisted Network and AI-Enabling Infrastructure based on its utilization aspects.



[Figure 6] Two Aspects of AI Telco Infra

1) AI-Assisted Network

AI-Assisted Network is a concept that maximizes the quality improvement, operational stability, and efficiency of mobile communication networks. This involves introducing various AI-based solutions, such as simulation and Vision AI into network systems, applying them to investment and asset management, and detecting anomalies in operational field networks. It can also contribute to enhancing customer-perceived quality through optimization of nationwide field issues.

To create a robust and resilient network which maintains the quality of customer experience, it is necessary to be able to immediately take action and recover from any faults or overloads through utilizing AI to monitor all detailed indicators that affect customer experiences. Additionally, In-Network Computing technology enables efficient deep data collection and analysis, allowing for effective processing of substantial data traffic. This can be particularly beneficial for AI-based anomaly detection and root cause analysis can be efficiently utilized.

2) AI-Enabling Infra (Telco Edge AI)

AI-Enabling Infra plays a role in connecting devices and AI DCs by utilizing the infrastructure assets of mobile network operators. Recently, with the proliferation of On-Device AI technology, inference-oriented AI services at the device level are being activated, while simultaneously developing large-scale data-based learning models in ultra-large-scale AI DCs. Between these two, the concept of providing business-to-business (B2B)/B2C services by utilizing network infrastructure and computing resources owned by mobile network operators to create new monetization and recreate infrastructure value is termed Telco Edge AI.

The goals of Telco Edge AI can be divided into infrastructure evolution and AI business monetization. In terms of infrastructure evolution, this includes concretizing implementation plans, discovering key and differentiating technologies, and creating and leading a global ecosystem. For AI business monetization, it pursues increasing infrastructure value through concretizing the Edge AI infrastructure vision, maximizing the utilization of telco assets, and discovering and concretizing new business models.

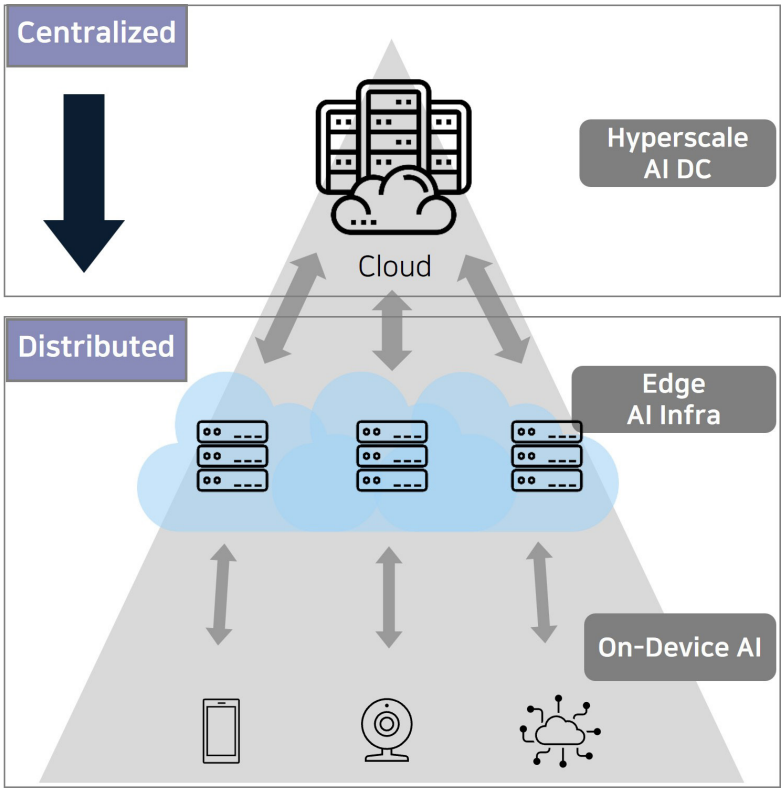
SK Telecom continues its efforts to create customer value and new businesses through Edge AI innovation of Telco infrastructure, aiming to proactively suggest timely evolution directions for communication networks.

4. Telco Edge AI

4.1 Why Telco Edge AI?

With advances in AI technologies and the explosive demand for AI services, the need for high-performance AI computing is also skyrocketing. In addition, the demand for low-latency training and inference, as well as large-scale computing power, is expected to grow as AI services are rolled out more extensively in the future. Therefore, AI telco infrastructure should also evolve to meet these demands and requirements. Traditional AI telco infrastructure has typically been centralized, relying on cloud-based, large-scale AI DCs (Hyperscale AI DCs). While these are well-suited for large-scale and high-performance AI models by providing sufficient computing power using high-performance GPUs, they come with drawbacks, such as high maintenance costs – due to space and power requirements, high latency, which is inherent in the centralized infrastructure, and increased vulnerability to information leakage in case of security breach situation.

Recently, On-Device AI technologies and services utilizing them have been getting significant attention in the mobile industry. The AI inference performance of mobile application processors (AP) has been continuously improving, and with the advancements in AI model lightweighting, small-scale models that offer performance comparable to large-scale models have emerged in specific applications. As a result, mobile devices are now able to deliver AI services tailored to their limited environments. On-Device AI is expected to continue evolving as a key tool for enhancing the customer experience.



[Figure 7] The Concept of AI Infrastructure

SK Telecom envisions that the future telco infrastructure will deliver complementary and differentiated AI services by combining the strengths of large-scale AI DCs and On-Device AI. While existing telco infrastructure has evolved to provide the enhanced connectivity, the future telco infrastructure will shift towards Telco Edge AI[1] infrastructure, offering both connectivity and AI computing power. By distributing centralized, large-scale AI workloads to the edge, this infrastructure will enable AI services with low latency and enhanced security.

Category	On-Device AI	Telco Edge AI Infrastructure	Hyperscale AI DC
Location	Terminal	Telecommunications Infrastructure	Centralized
Configuration	Low-End GPU/NPU/CPU	xPU[2] (GPU, NPU, CPU, etc.)	High-End GPU
Scale	Built-In Terminal	Small Server Racks Built-In Communication Equipment	Large Server Racks
Intended Use Cases	Small AI Services (Some Inferences)	AI Services at Scale + Communication Services	Large-Scale AI Services
Low Latency, Security, Agility	High (Processing by Device)	Middle	Low (Via Multiple Nodes)

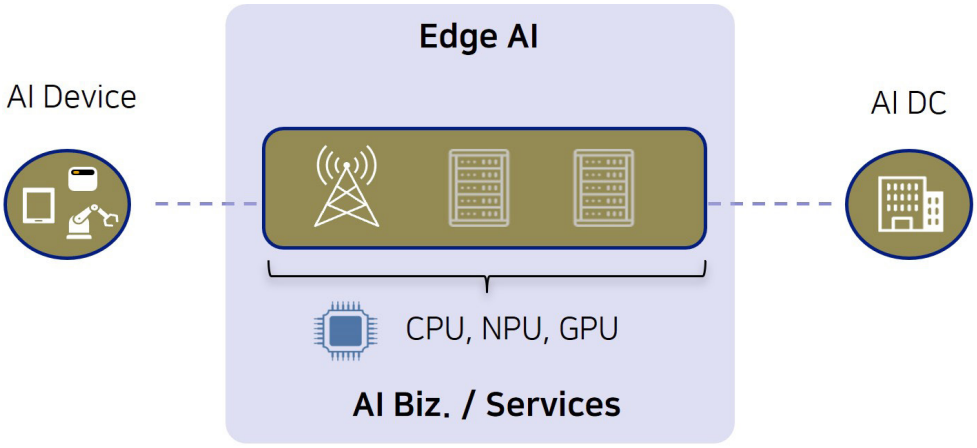
[Table 1] The Differences between On-Device AI, Telco Edge AI, and Hyperscale AI DC

[1] Edge AI is distinct from the existing centralized AI DC, and refers to the form in which AI computation is performed at the edge of the network, which is close to the data, and it is a concept that includes On-Device AI. However, Edge AI in this white paper is distinguished from large-scale AI DC / On-Device AI, and refers to a form that provides AI services while located in the communication infrastructure.

[2] xPU: A dedicated processing unit (e.g. CPU/GPU/NPU) designed based on a specialized instruction set/architecture to optimize for different processing needs.

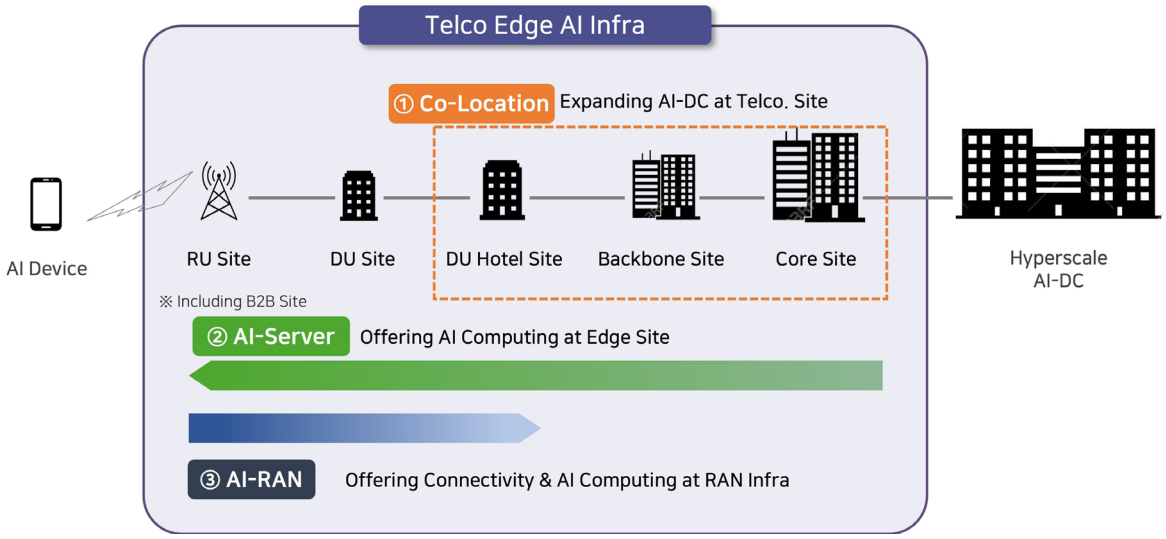
4.2 What is Telco Edge AI?

Telco Edge AI infrastructure is an infrastructure that adds new value to the traditional telco infrastructure, which focuses only on connectivity, by offering not only communication services but also the AI computing power necessary for AI services. This is achieved by leveraging existing infrastructure, facilities, and equipment of operators.



[Figure 8] Telco Edge AI between On-Device AI and AI DC

Mobile network operators are deploying and operating infrastructures and facilities to provide telecommunication services nationwide, making it difficult to replace them with Telco Edge AI infrastructure all at once. Therefore, SK Telecom classifies its Telco Edge AI infrastructure into the following three areas based on business and service characteristics.



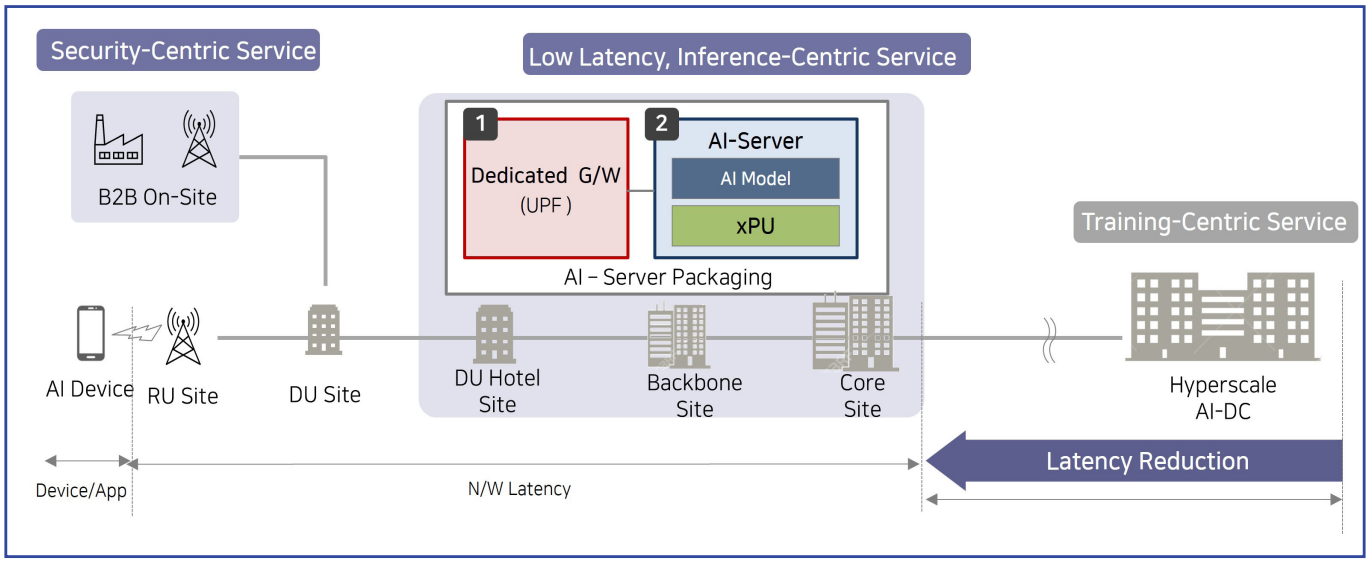
[Figure 9] Three Areas of Telco Edge AI

1) Co-location

The first area is “co-location”, which involves transforming the unused space in telco infrastructure owned by mobile network operators into infrastructure that provides AI computing and AI services cost-effectively. xPU-based servers, which provide AI computing power, consume more energy and generate substantial heat, accelerating the aging of chips compared to traditional telecommunications equipment. Consequently, even if there are available spaces in the telco infrastructure, introducing xPU-based servers without proper preparations is challenging. Therefore, as an alternative to hyperscale AI DCs, which require excessive time and costs to build, we can expand high-performance, energy-efficient cooling systems and power supply systems in existing telco facilities to proactively prepare, allowing for the flexible and rapid deployment for AI services.

2) AI-Server

The second area is “AI-Server”, which is to provide AI computing power at the edge and improve communication performance by placing AI servers further forward based on the cooling and power facilities prepared on the co-location. AI-Server does not directly provide connectivity. Instead, it can provide AI workloads for improving the performance of communication services, such as network automation, optimization, and energy saving to facilities concentrating Data Unit (DU) or higher network side. AI workloads for services could be developed by third parties or companies. The AI-Server can provide AI services at the edge, providing customers with relatively low latency and high security compared to large-scale AI DCs.



[Figure 10] AI-Server

The main purpose of AI-Server is not to provide AI computing power at the level of large-scale AI DCs, but to provide low-latency, inference-centric AI services. Therefore, it is necessary 1) to select an xPU optimized for the AI workload of the service to be provided, and 2) to consider building a user plane function (UPF) to provide low-latency services between AI server and device.

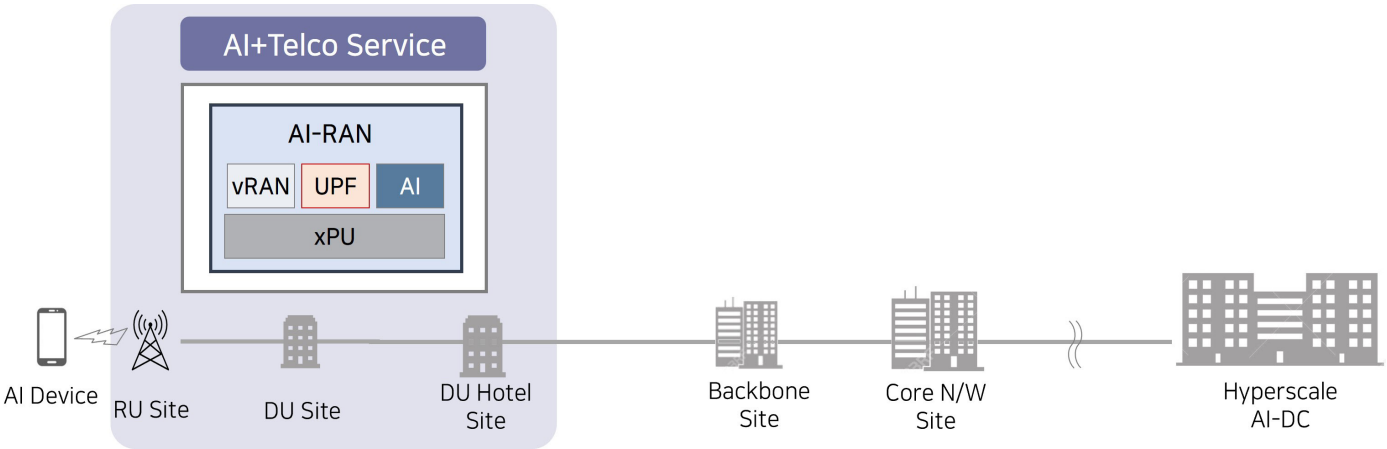
3) AI-RAN

The third area is "AI-RAN", which refers to the area where RAN functions that provide communication services and AI computing computations for AI services are performed at the same xPU-based server platform. AI-RAN is one of the key technologies for realizing AI-Native, which is being discussed in 6G. The difference from AI-Server is that it provides AI and communication services simultaneously on a single equipment.

In AI-RAN, AI computing power can be used in two main areas. It is a field that 1) improves RAN performance using AI models to simplify physical layer signal processing and wireless resource allocation functions that require complex computation, which are the main workloads of traditional RAN, and 2) provides AI services at the same time by utilizing the computing resources left after providing RAN services for connectivity.

AI-RAN can be introduced as a replacement for existing base station equipment. Base station equipment has been deployed in a wide range of operator facilities with a tremendous quantity, and from a technical point of view, it must fundamentally provide stable communication services. Therefore, there are many considerations that need to be addressed from a mid to long-term perspective.

However, general-purpose commercial on-the-shelf (COTS)-based RAN equipment are still inefficient in terms of cell capacity, power consumption and stability compared to traditional RAN equipment. Besides,



[Figure 11] AI-RAN

technology in the field of AI is changing very rapidly, and new AI chips and platforms continue to appear on the market. However, the technical support and maintenance period for AI chips and platform products is relatively short, which is different from a traditional telecommunication equipment that requires long-term and stable operation at the operator's facility. If the xPU-based AI-RAN equipment does not guarantee sufficient technical maintenance, there are concerns not only for AI services but also for providing stable communication services, so it is necessary to consider the life cycle management perspective of the equipment.

The technology that should be considered in the Telco Edge AI infrastructure is to provide both AI computing power and stably providing communication services. In order to achieve this mission, it is important to adopt proper xPU that takes into account the requirements of AI applications and communication services to Telco Edge AI infrastructure.

High-performance xPUs provide sufficient AI computing power, but they can cause excessive power consumption which generates substantial heat that accelerate the aging process of the chips. In comparison, AI chips for On-Device AI aim for low power consumption, but the AI computing power they can provide is limited. In recent years, chip technologies to provide AI computing power have recently been competitively advancing, including CPUs, GPUs, and NPUs. Therefore, if we can build an optimized Telco Edge AI infrastructure with proper xPUs, we can maximize the efficiency and performance of the telco infrastructure, which will lay the foundation for providing innovative ICT services and applications.

AI-RAN, like AI-Server, can also provide low latency and high security because AI computing resources are located at the edge. In addition, Telco Edge AI, such as AI-Server and AI-RAN is expected to provide customers with additional value that only mobile operators can provide compared to hyperscalers, such as efficient traffic control in congested situations, by combining AI services with network-specific functions, such as network slicing and QoS provisioning.

5. Evolution Directions of Telco Edge AI

To successfully advance toward Telco Edge AI, it is crucial to concretize implementation plans and secure differentiated technologies to support them. Additionally, it is important to continuously explore specialized business models and foster the development of a global ecosystem where vendors, operators, and customers can all benefit.

In addition, AI-Assisted Network, another aspect of AI infrastructure utilization discussed in Section 3.2, should not be overlooked in order to increase the performance and efficiency of the network infrastructure that mobile operators are currently operating or will deploy in the future. AI-Assisted Network serves as an essential intermediate step in the future architectural evolution toward Telco Edge AI.

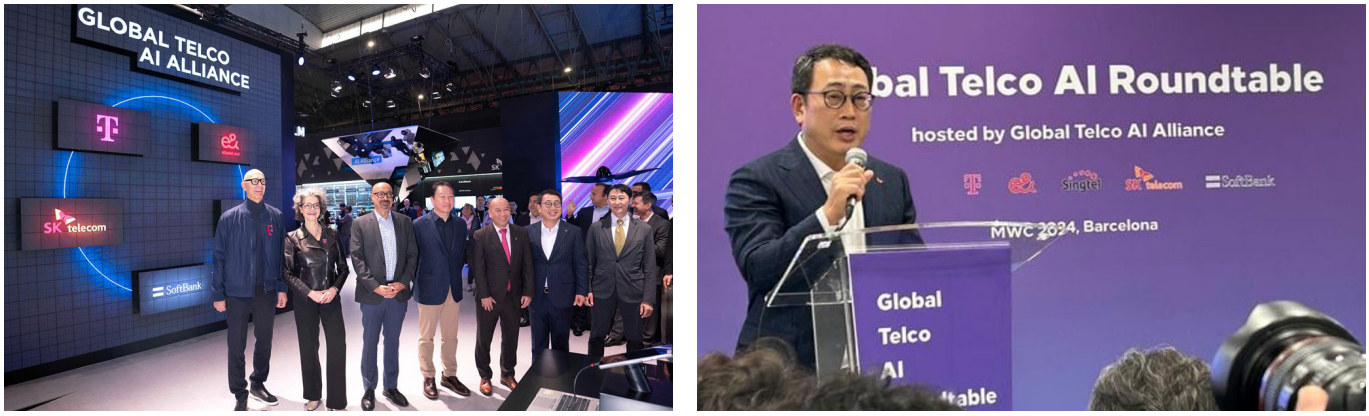
Therefore, this chapter will first look at the activities for the creation/expansion of the Telco Edge AI ecosystem, present examples of Telco Edge AI-specialized business models, and finally describe the key technologies required for the evolution toward Telco Edge AI, including the two utilization aspects of AI infrastructure.

5.1 Ecosystem Creation

5.1.1 Operator/Vendor Cooperation

To transform the traditional telecommunication infrastructure into one that offers differentiated value through Telco Edge AI, building a supporting ecosystem is crucial.

To achieve this, cooperation among mobile operators who are looking for new business models and ways to generate revenue is essential. To this end, SK Telecom plans to share its vision for the evolution of future



[Figure 12] MWC24 GTAA Inaugural Meeting

mobile communication network infrastructure with major global mobile operators and form partnerships to build a global ecosystem. In Feb. 2024, SK Telecom announced its strategy to transform into a 'Global AI Company' by establishing the Global Telco AI Alliance (GTAA) at MWC24 and jointly developing a telco LLM (Telco Large Language Model) with Deutsche Telekom, E&Group, Singtel Group, and SoftBank to accelerate AI technology innovation and application through cooperation among global telecom operators. Based on this, SK Telecom will expand the scope of cooperation to Telco Edge AI and establish a global operator alliance to share the technology and know-how possessed by operators, and promote cooperation and solidarity between operators to secure Telco Edge AI-based services and solutions.

In addition, SK Telecom will strive to build a global ecosystem through partnerships with major equipment and chipset vendors. In the past, virtualization technology began to be introduced into core networks and radio access networks, and some IT companies entered existing ecosystem dominated by telecom-specialized vendors, causing changes in the telco ecosystem. The convergence of telecommunications and AI is emerging as a key element of 6G, which is expected to arrive around 2030, and major AI server and chipset vendors are attempting to enter the telco market with similar strategies. In particular, competition is intensifying, for example, several big tech companies have formed an alliance to counter companies that have already taken the lead in AI computing infrastructure, such as GPUs.

In particular, for the success of Telco Edge AI, it is essential to have optimized Telco-Grade infrastructure, which is differentiated from the systems and solutions used in the existing IT industry. Incidentally, the ultra-high-end xPUs and servers used for AI training and inference at hyper-scale in AI DC are not only inefficient in terms of total cost of ownership (TCO) in Telco Edge AI, but also can be over-specified. Therefore, joint efforts are essential to expand the ecosystem of Telco Edge AI through close communication and cooperation among players in various fields, such as devices, chipsets, servers, networks/storage, platforms, and contents.

SK Telecom is collaborating with major equipment and chipset vendors to create an ecosystem that encompasses not only telecommunications and IT technologies but also AI technologies, to design/build/manage Telco Edge AI infrastructure optimized for the company's network environment, and to secure SLA (reliability, life cycle management, technical support) and TCO considering the characteristics of telecommunications equipment. Besides, as another effort to build/expand the ecosystem of Telco Edge AI through cooperation with government, SK Telecom is participating in the national projects, 6G core technology development project (2021~2025, funding amount: 191.7 billion KRW) and next generation network technology development project (2024~2028, funding amount: 440.7 billion KRW), that have been proposed by Ministry of Science and ICT in South Korea.

5.1.2 Standardization and Academic Research Activities

Standardization of technologies is also essential to realize economies of scale by expanding the ecosystem and

avoiding fragmentation, such as ensuring 1) seamless working of Telco Edge AI infrastructure with hardware and software from a variety of vendors, and 2) interoperability between the various devices and systems consisting of Telco Edge AI. SK Telecom participated as a panelist in the "3GPP Stage 1 Workshop on IMT 2030 Use Cases" held in early May 2024 and suggested the importance and need for standardization of Telco Edge AI infrastructure, and proposed the standardization of Telco Edge AI infrastructure, one of the core technologies for the evolution of next-generation infrastructure, at the 3GPP SA1#106 meeting [9] / O-RAN Alliance standardization meeting [10] held in late May/June 2024. In the future, SK Telecom will continue to propose related standardization items to SDO, such as 3GPP, O-RAN Alliance, and ETSI, and proceed with standardization work to promote the evolution of Telco Edge AI infrastructure that can provide innovative services and new values.



[Figure 13] Activities in 3GPP Stage 1 Workshop IMT-2030 Use Cases and O-RAN Alliance Meeting

In addition, as part of securing intellectual property rights, we proposed to define equipment nodes, components within nodes for Telco Edge AI infrastructure, and also suggested links/interfaces for interworking with existing equipment and exchange messages required for supporting AI inference/learning, which are for a basic structure for the evolution of Telco Edge AI infrastructure. In addition, when an edge node supports communication and AI services at the same time, we proposed optimized resource utilization of infrastructure by defining the resource management/optimization structures and interfaces between multiple edge nodes.

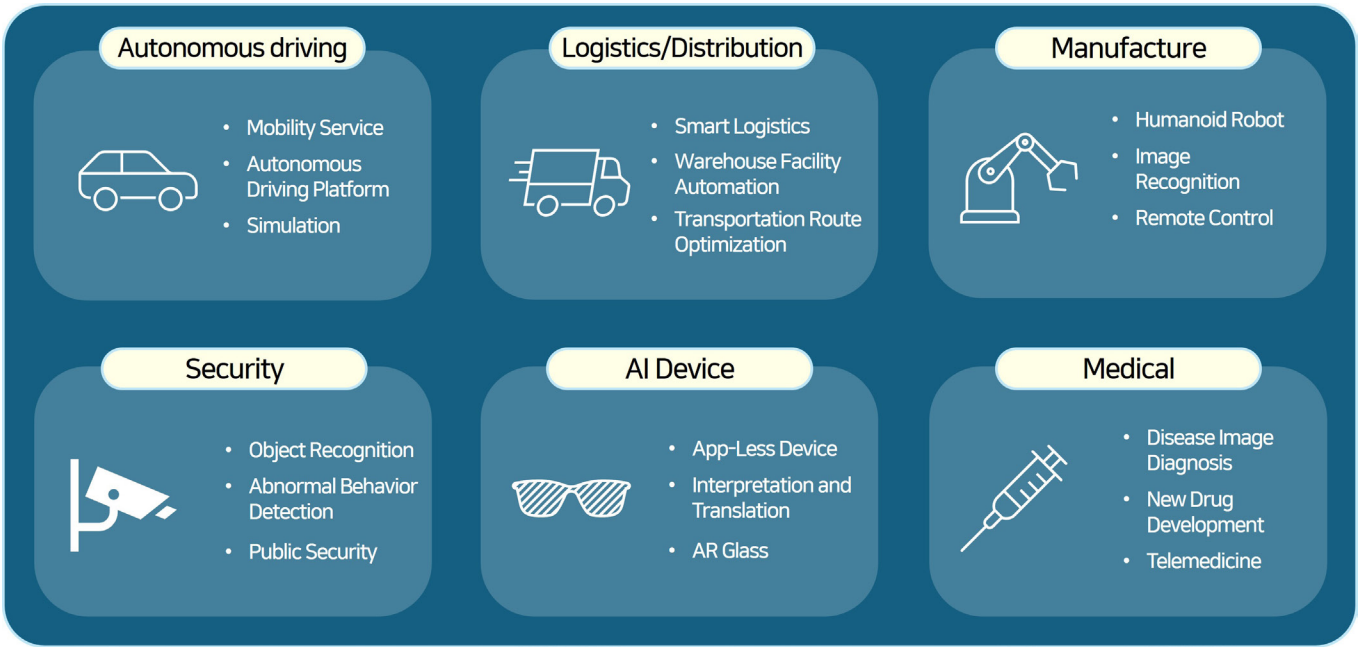
Based on the results of such research activities, we presented the overview and direction of AI-Native radio access network research and optimal network design based on 6G Simulator at the Korea Institute of Communications and Information Sciences (KICS) conference in Jun. 2024, and published in IEEE Communications Magazine in Aug. 2024, and will present a paper at the International Conference on ICT Convergence (ICTC) in Oct. 2024 [11], [12], [13], [14].

In addition, we are making efforts to expand the Telco Edge AI ecosystem from R&D results through various industry-academia cooperation activities. For example, the development of AI-RAN S/W including AI use

cases and the quantitative evaluation of the utility of the xPU server-based Telco Edge AI infrastructure will be conducted when AI workload is applied. As another example of industry-academia cooperation, designing/implementing FPGA for transceiver modem of ISAC technology, one of usage scenarios in IMT-2030 Framework Recommendation [3], will be performed with connecting to an xPU-based server for AI training/inference in order to test/verify provision of sensing and communication services.

5.2 Business Model Discovery

As mentioned earlier, a business model based on Telco Edge AI infrastructure is an essential element to re-rate the value of telco infrastructure and evolve it into a monetizable infrastructure. In this chapter, we present examples of business models that are expected to benefit from Telco Edge AI infrastructure.



[Figure 14] Examples of Telco Edge AI Business Models

1) Autonomous driving

So far, vehicle manufacturers are focusing on developing On-Device based autonomous driving platforms that can be mounted on vehicles. V2X communication, including low latency and high security, is essential for autonomous driving, and a business model that provides large-scale space and server provision using AI servers and computing resources dedicated to real-time inference using AI-RAN can be considered.

2) Logistics/Distribution

In the case of logistics/distribution, it can be considered that it is easy to introduce AI technology due to the relatively simple and standardized nature of work. By utilizing Telco Edge AI infrastructure in areas, such as inventory management/automation of warehouse facilities and optimization of logistics transportation routes, it is possible to meet business needs for automation and real-time business processing.

3) Manufacture

In the case of the manufacturing industry, the trend of digital transformation (DX) and automation/unmanned is accelerating, and it is possible to seek efficiency through big data analysis/learning for logistics transportation, facility robots, and processes in isolated spaces in factories, and new business models can be discovered by utilizing Telco Edge AI infrastructure in these areas.

4) Security

In the case of the security area, the use of AI in areas, such as video analysis, object detection, and intelligent monitoring can enable real-time response with increased efficiency and low latency, and Telco Edge AI infrastructure can be used to support this.

5) AI Device

Recently, with the rise of On-Device AI technology that performs AI calculations on the terminal device itself, customers' expectations for AI devices are rising. However, due to limitations in terms of use cases and the current AI performance of terminals, the spread of AI devices is limited. Therefore, business models, such as improving AI performance or reducing latency of AI devices using Telco Edge AI infrastructure, or providing AI services using Telco Edge AI infrastructure for low-cost devices with minimal functions for AI support may be considered.

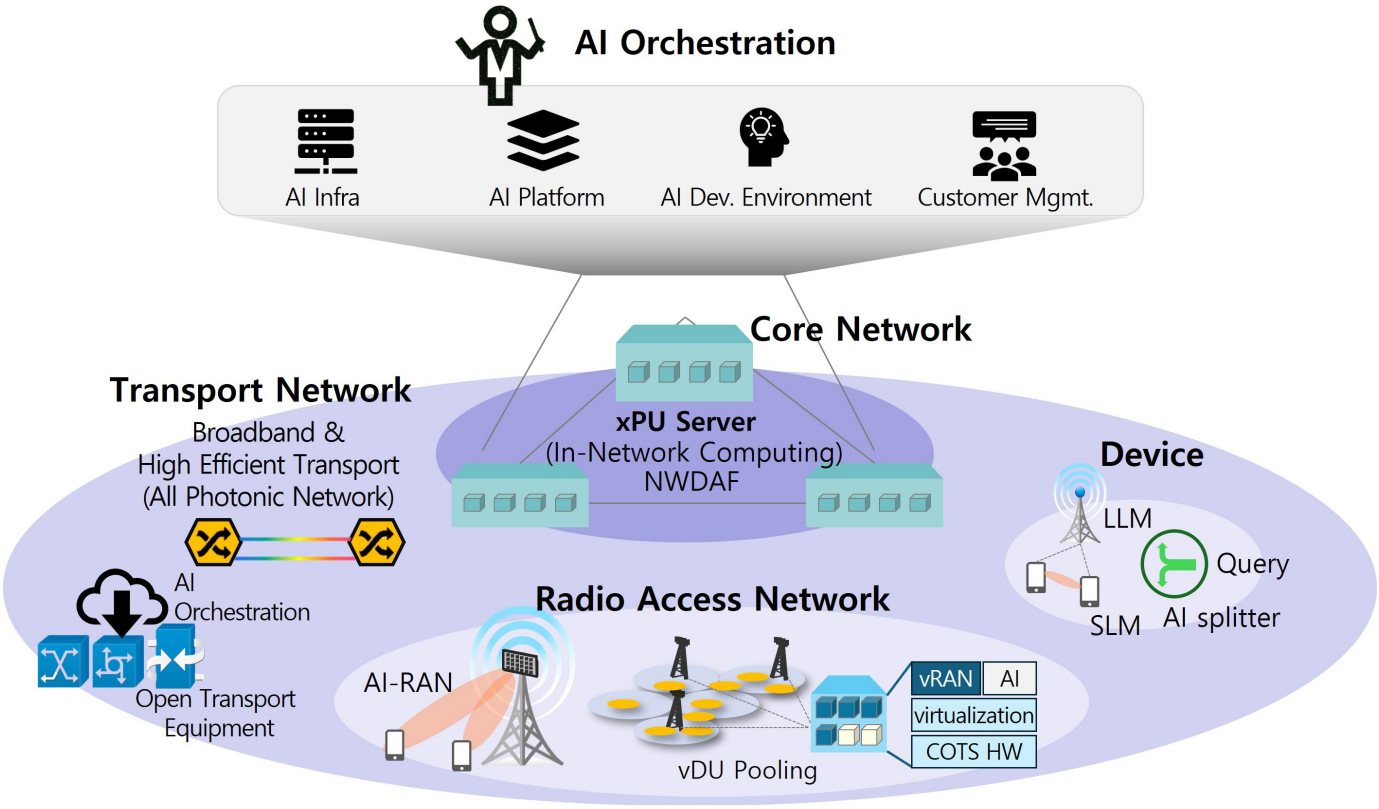
6) Medical

In the medical field, the use of AI in various areas, such as AI-based analysis of vast medical data and diagnosis, treatment, and prevention of diseases is also expected. In particular, since high security may be required from the privacy perspective of personal medical data, it is necessary to consider discovering a medical-related business model using Telco Edge AI infrastructure.

In addition to the above business models, it is important to discover new promising business models that are expected to utilize Telco Edge AI infrastructure, and based on this, it will be possible to promote the evolution of communication infrastructure to AI Telco infrastructure. In order to accelerate the discovery of promising business models, which is the most important factor for the successful evolution of Telco Edge AI, SK Telecom plans to conduct Proof of Concepts (PoCs) on various use cases to verify the possibility of securing meaningful results.

5.3 Key Technologies

This chapter introduces key technologies that SK Telecom is focusing on currently researching and developing for the evolution toward Telco Edge AI. The technologies introduced here aim at meeting the requirements for Telco Edge AI, such as efficient allocation, deployment, and management of AI computing resources, and technologies for ultra-low latency communication between Edge-Client and Edge-Edge. Some of these technologies can also be utilized for AI-Assisted Networks in which case the expected effects of their use in AI-Assisted Network are described. The key technologies related to Telco Edge AI, viewed from a comprehensive perspective across the entire network, are as follows.

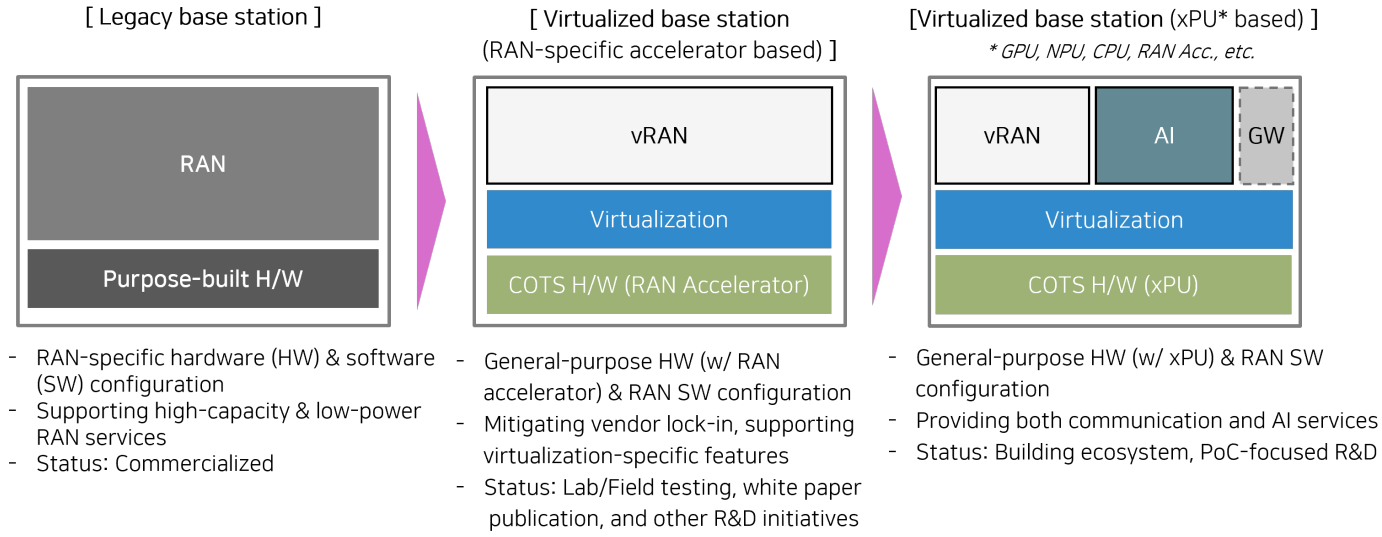


[Figure 15] Key Technologies for Telco Edge AI in Each Area

5.3.1 Radio Access Network

Traditional base stations are designed and configured with purpose-built hardware optimized for operating the vendor's RAN software, making it challenging to realize AI-RAN, which needs to provide both RAN and AI services at the same time. In contrast, virtualized base stations are equipped with xPU to handle RAN and AI workloads on general-purpose hardware, and through the virtualization layer between hardware and

software, xPU resources can be shared for RAN software acceleration from base station vendors and AI software acceleration from third-party vendors. It is a core technology for the realization of AI-RAN, which aims to provide RAN and AI services simultaneously on a single equipment.



[Figure 16] Evolution Direction of Base Station Architecture for Realizing AI-RAN

Meanwhile, the current virtualized base station has some limitations in terms of cell capacity and power consumption compared to the existing purpose-built hardware-based base station. SK Telecom has developed a variety of xPU-based virtualized base stations, including CPU, RAN accelerator, and GPU, and is conducting R&D to identify and improve performance, capacity, power consumption, and TCO through testing in lab and field environments.

In Feb. 2022, SK Telecom announced the results of its research on 5G virtualized base stations with major vendors at MWC 2022, and in Jan. 2023, it successfully demonstrated 5G virtualized base stations in a field environment. Since then, as RAN-specific accelerator technology has diversified and evolved, in Nov. 2023, through cooperation with major vendors, SK Telecom has verified improvements in terms of cell capacity and power consumption of new RAN-specific accelerator-based virtualized base stations compared to existing virtualized base stations in a laboratory environment.

Based on these R&D achievements in virtualized base stations, SK Telecom jointly published a white paper on virtualized base station technologies with NTT DOCOMO in Feb. 2024, suggesting key considerations and evolution directions for virtualized base stations [15]. SK Telecom plans to continue mid-to-long term R&D on virtualized base station to secure the foundation for architectural evolution toward AI-RAN by promoting the development and verification of various virtualized base stations based on xPU, such as GPUs as well as RAN-specific accelerators.

In addition, SK Telecom is actively conducting R&D to improve performance and quality from communication services perspective by applying AI technology to various functions of base stations. As a representative example, we are conducting research on the application of AI technology to the air interface between base station and user equipment (UE) through cooperation with global operators and vendors. While conventional mobile communication systems use reference signals (RS) for communications between transceivers, the spectral efficiency was improved by minimizing the RS load through AI-based modulation and demodulation technology. Through a PoC, we verified the possibility of the throughput gain in a laboratory environment and presented detailed R&D outcomes at MWC 2024. SK Telecom is also conducting R&D activities for various other use cases, primarily considering the usefulness from the perspective of an operator actually operating the commercial network.

5.3.2 Core Network

Core network serves as the gateway for all voice and data traffic generated by UE to connect to the Internet. It handles terminal access, authentication, registration, IP address allocation, data session management, handover, QoS, and billing processing. This is achieved through various interconnected NF (Network Function) equipment designed for each specific role.

Core network plays an important role in ensuring high-capacity & low-latency processing, stability, and customer quality. However, transitioning to a Cloud-Native architecture can present challenges, such as the potential for increased delays due to the separation of software components. Core network is increasing in capacity as server performance improves, and it requires high stability, so it is vital to predict and detect signs of failure and deterioration of network quality in advance for automatic or pre-emptive action.

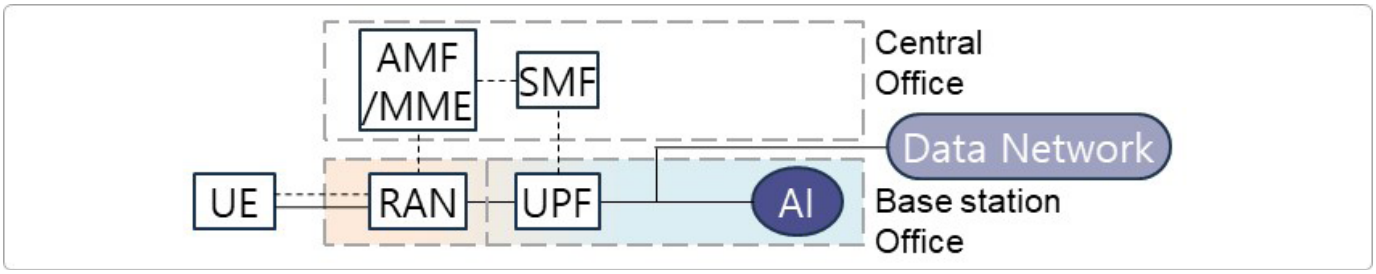
It is crucial to improve the system performance with the aid of servers equipped with xPU and software technology. SK Telecom is currently advancing towards 6G Core technology for handling large-capacity and low-latency using xPUs through collaboration with global partners. As a result of our research so far, we have published a white paper titled ‘Inline Service Mesh’ on a technology that can reduce NF communication latency by up to 70% in Jan. 2024 [16], and we expect it to be reflected in 3GPP [17].

By utilizing In-Network Computing technology, we are investigating methods to effectively access signal and packet data quality even under heavy data traffic conditions. We are also focusing on developing AI-based technologies for anomaly detection and root cause analysis, and continue to publish academic papers [18] and conduct research activities.

NWDAF (Network Data Analytics Function) is a standard equipment that analyzes data produced by NF with the aim of optimizing equipment functionality, enhancing customer quality, improving performance, and automating operations. SK Telecom has successfully developed network congestion prediction to

provide the best customer quality, and continues to research various applications, including detecting abnormal UE activity.

For Telco Edge AI, we are investigating technologies that operate on xPUs and distribute UPF deployment to base stations closer to customers. Unlike voice and data traffic, UPF deployed at the edge connects AI service traffic with AI computing resources, allowing mobile network operators to deliver inference services closer to customers.



[Figure 17] Concepts of Traffic Handling for Edge-Deployed & Distributed UPFs

5.3.3 Transport Network

With the introduction of distributed AI DCs suitable for generative AI learning and services, the architecture of transport networks is changing. AI learning is primarily conducted through GPU servers in DCs, requiring high-capacity and lossless transmission technologies.

Due to the high price of GPUs and the increased power consumption, AI services are suitable for AI inference on customers' devices or on-premise. Accordingly, transport networks that support distributed AI DCs require the role of an inter-connectivity between Edge AI DCs and ultra-high-capacity, lossless, low-latency, and low-power connectivity services within Edge AI DCs.

5.3.3.1 Broadband and high-efficiency transmission technology

Due to the nature of Edge AI, which aims for distributed DCs, it is essential to secure high-capacity and high-efficiency connectivity between Edge AI DCs in order to train large amounts of data. In order to enhance power efficiency, it is necessary to develop towards minimizing optical-electrical conversion within or between Edge AI DCs. One of the major organizations conducting research and development on this topic is IOWN Global Forum, which is working on achieving ultra-wideband and ultra-low latency through end-to-end optical transmission, referred to as APN (All Photonic Network).

APN separates existing integrated optical transport system (e.g., ROADM, Reconfigurable Optical Add-Drop Multiplexer) into functional blocks according to their purpose (transponder, multiplexer, and wavelength selective switch) and classifies them into APN-T, APN-G, and APN-I, respectively, while striving to build an open equipment ecosystem.

Open transport equipment can be used by separating hardware and software and installing Open Network operating system software on general-purpose switches/routers and optical transport equipment. By applying common units according to the switching capacity of each region by utilizing such open transport equipment, it is possible to flexibly respond to hardware replacement or expansion, and software can be installed on an on-demand basis according to the functionality to efficiently use only the features optimized for AI Telco infrastructure.

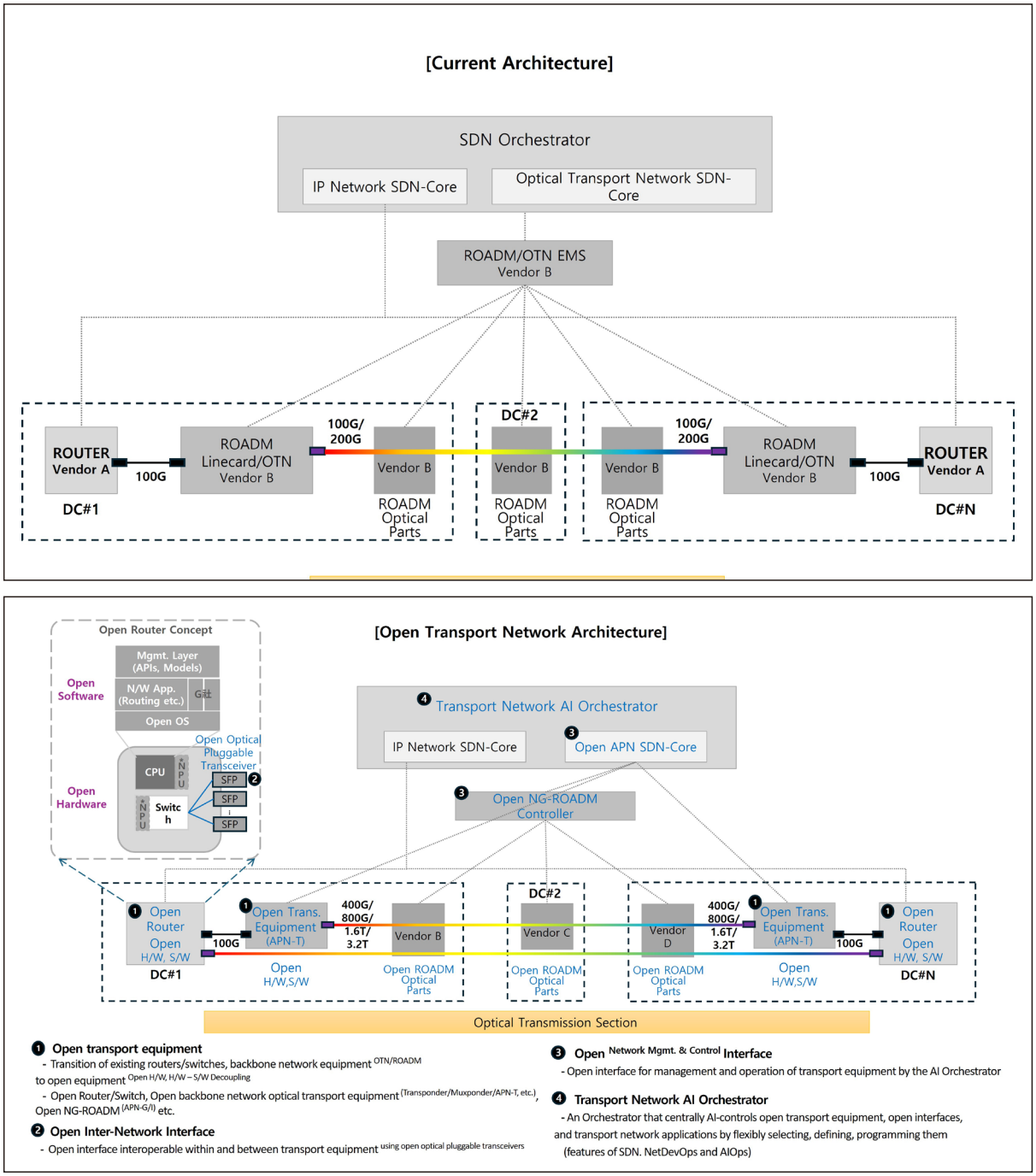
CPO (Co-Package Optics) technology is being actively researched and developed as a potential candidate for technological evolution within Edge AI DC, and open optical pluggable transceiver technology capable of transmitting over 400Gbps and over 500km between Edge AI DCs is also important.

One of the candidate technologies is optical switching technology to achieve an all-photonic network within and between Edge AI DCs. While it is challenging to completely replace packet switching, which has the advantage of high-speed switching, it is possible to realize a network that satisfies high scalability, low latency, lossless, and low power through a network architecture that combines packet switching and optical switching.

5.3.3.2 Transport Network AI Orchestration

For the networking of the Telco Edge AI infrastructure, it is necessary to have a transport network AI Orchestrator technology that controls and manages all the above-mentioned open transport equipment, optical pluggable transceivers, and optical switching as one network. In addition to accommodating existing vendor-specific equipment, we are conducting research and development to directly control and manage open transport equipment and optical pluggable transceivers, and increase the level of automation using telemetry data collection and AI control.

Through the above activities, we are ultimately reducing the TCO of the transport network, securing the completeness of transport network equipment, increasing operational stability, and promoting AI advancement.



[Figure 18] Open Transport Network Configuration for Transition to AI Telco Infrastructure

Through open transport equipment R&D, we are securing the transport network technology required for AI telco infrastructure while implementing high-capacity, ultra-low-latency technologies essential for AI services.

In AI Telco infrastructure, the transport network consists of an open IP switch for the connection within Edge AI, and an open router/open optical transport network equipment for the connection between Edge AI.

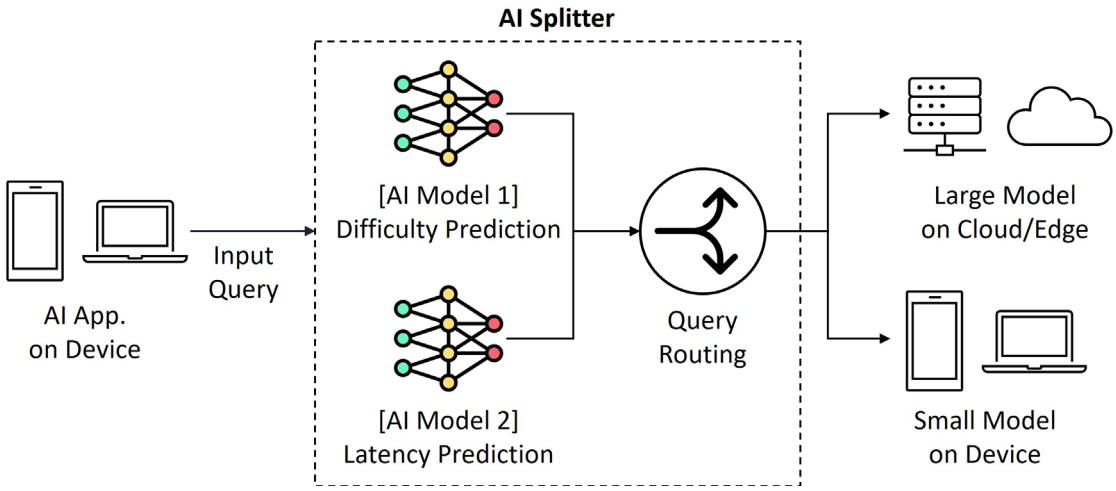
The biggest differences from networks composed of existing transport equipment are high-capacity through integration of optical coherent technology, ultra-low latency via minimization of optical-electrical conversion, openness of transport equipment, and evolution of network management technology including AI and intent-based automation. We have been developing and demonstrating technologies related to open transport equipment, open long-distance optical pluggable transceivers, open optical path switching technology, and transport network AI orchestrators in response to these changes.

5.3.4 Device

Even in the 6G era, where AI will be at the forefront, devices are expected to play a very important role in forming contact point with customers. Accordingly, SK Telecom is actively conducting R&D in terms of devices for the effective convergence of communication and AI in the Telco Edge AI environment.

5.3.4.1 Device-based Hybrid AI Optimization

Within the trend of AI technology advancement, we are conducting research to enhance the quality and efficiency of AI services by strengthening the role of devices. Recently, with the continuous improvement of AI computing power of mobile APs, small language models that can run on devices are emerging. Google's Gemini Nano, Microsoft's Pi3 Mini, and Meta's Mobile LLM are prime examples. They have similar levels of performance in specific scenarios, such as Q&A compared to LLMs running in the cloud, and they are cost-effective in terms of infrastructure construction/operation. Therefore, in the hybrid AI environment (i.e.,coexistence of device, Telco Edge AI, and cloud) that is gradually becoming a reality, the challenge is to effectively determine which node is most advantageous for handling user requests. SK Telecom is developing an AI splitter technology, that analyses and predicts various characteristics of a user's request using Telco Edge AI and On-Device AI and then assigns them to the optimal node, and plans to verify its effectiveness through future PoCs.



[Figure 19] Conceptual Diagram of AI Splitter

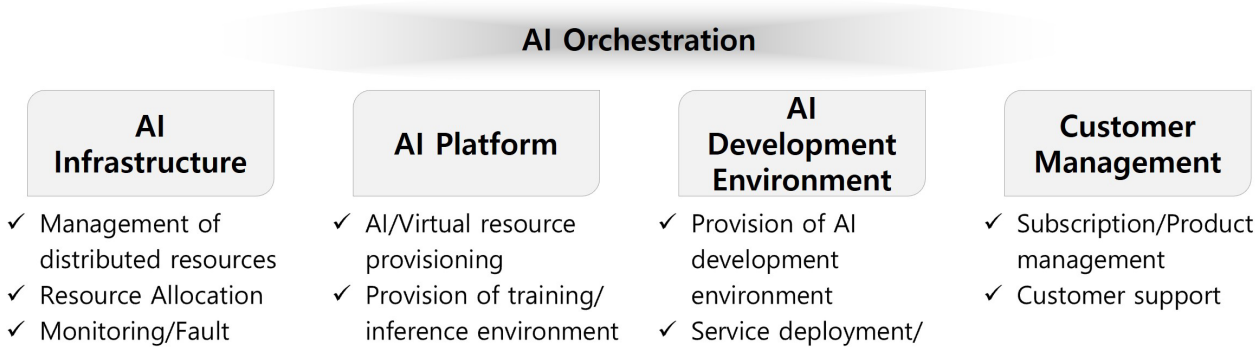
5.3.4.2 New devices with Telco Edge AI

We are conducting R&D activities for devices to discover and support new business models and use cases based on 6G and Telco Edge AI. Recently, new devices with various form factors are emerging to effectively utilize AI technology. Rabbit's R1 and AI Pendant, which provide personal assistant services based on voice and gestures, and Meta's smart glasses and Quest 3, which can provide new experiences to customers by combining AI with AR/VR, are representative examples. Such new types of devices are currently adopting server-based AI services due to processor and memory requirements. However, such AI services that are only used for servers have limitations in terms of customer experience and information protection. Therefore, in order to secure low latency and security, we plan to identify AI-specialized devices that can comprehensively utilize Telco Edge AI and On-Device AI along with 6G technology to be introduced in the future. In addition, we also plan to continue to provide new user experiences in the AI era by developing related AI services.

5.3.5 AI orchestration

We have been conducting R&D to manage AI computing resources in end-to-end manner to be used in Telco Edge AI. AI orchestration is necessary to provide customers with an efficient and fast AI development environment and to offer AI services. We are developing technologies in areas, such as providing AI infrastructure and platforms to customers, helping the AI service development environment, and managing to provide customers with the information they need.

AI orchestration provides the AI computing resources needed to deliver AI services. In Telco Edge AI environment, AI computing resources are distributed among small DU sites or core sites. Hence, it is necessary to effectively manage distributed AI computing resources and provide them to customers to improve the user experience. AI orchestration integrates and manages distributed AI computing resources, and helps allocate and deploy resources to meet customer needs. It also provides an environment for customers to develop and commercially deploy AI services. AI orchestration plays an important role in the areas of AI infrastructure management, platform management, development environment management, and customer management.



[Figure 20] Key Roles of AI Orchestration

AI computing resources used for training and inference in AI services must support ultra-low-latency broadband networks and storage space. AI computing resources need to be allocated and connected in real time, and as services evolve in diversity, complex AI computing resources need to be delivered flexibly and cost-effectively. It is necessary to manage the status of distributed AI computing resources and manage the optimal use of resources. In addition, in order to improve the quality of AI services for customers, it is necessary to integrate and manage various AI computing resources, and to distribute and schedule AI service workloads.

SK Telecom is developing SDI (Software Defined Infrastructure) technology and applying it to commercial networks in the face of virtualization. SDI is a solution for installing and managing virtualized computing resources. In addition, it has AI computing resource management, AI platform, AI development environment, and customer management (billing optimization) solutions. Research is underway to provide AI orchestration through collaboration with overseas solutions and integration between solutions. In what follows, we will take a look at the technology areas that SK Telecom is focusing on.

5.3.5.1 AI Infrastructure Management

As a service provider that provides AI computing resources, it is necessary to build and continuously manage AI computing resources. As an AI computing resource, server/GPU models, performance, networks, storage space, and virtual environments should be built remotely and automatedly in distributed regions. In addition, it should be able to keep them up-to-date and available to customers through regular upgrades.

In order to provide uninterrupted service to customers, temperature, hardware and software failure monitoring, and alarms should be provided. In order to automatically recover from failures and automatically upgrade resources, technology is needed to transfer AI computing resources being provided to customers without interruption or wait for use suspension.

Since AI services require high processing power, service quality degradation or interruption may occur if AI computing resources, such as GPUs, networks, and storage space are not efficiently managed. For AI service quality, AI computing resources including GPUs, should be optimally allocated, and managed to prevent bottlenecks. In other words, it is crucial to optimize resource management by automatically relocating AI workloads to idle GPUs to reliably deliver high-performance AI services.

5.3.5.2 AI Platform Management

AI orchestration allows efficient utilization of physical AI computing resources by converting them into virtual resources. According to the customer's request, it should support the creation and deployment of VM environments and container environments. VM virtualization can provide an independent workspace

and provides a virtualization environment in the form of a container. Container management platforms, such as Kubernetes support container deployment, expansion, and operation automation.

AI platform management is necessary to maximize the flexibility and efficiency of AI computing resources. Container virtualization technology is essential to effectively meet the diverse requirements of AI services. Through AI platform management technology, it effectively operates various AI workloads, provides automated management, and offers stable and reliable services.

Since AI computing resources play a key role in training and inference work of AI services, it is important to flexibly connect AI computing resources to virtualization platforms and provide them to customers' service environments. AI computing resource scheduling, which optimally distributes and manages AI computing resources to maximize AI service performance by shortening AI workload processing time while maximizing the use of AI computing resources, is important.

In particular, clustering technology is required to bundle multiple GPUs for large-scale AI service training/inference. NVIDIA offers GDR (GPU Direct RDMA) as a memory technology for processing training data with large bandwidth, NVLink and Infiniband as network technologies, and GPU Direct Storage (GDS) as a technology for storage devices. Other network technologies include UALink (Ultra Accelerator Link) and RoCE (RDMA over Converged Ethernet) which are used in Ethernet. However, although it can be used within the same building, it is difficult to use AI computing resources more than a few tens of kilometers away due to physical limitations, such as increased delays and reduced speeds, so we are investigating various ways to overcome this.

Furthermore, a technology is needed to virtualize and dynamically segment the GPU cluster itself. In AI services, if Infiniband or RoCEv2 resources are not allocated to match the GPU allocation, the network becomes a bottleneck, lowering GPU utilization. AI platform dynamically builds and isolates a virtual network dedicated to GPU clusters according to amount of the customer's resource request, taking into account the message exchange pattern between GPU servers and the physical network topology with a rail-optimized structure.

Clustering of AI computing resources at the distributed edge should be overcome by software-based, parallel, distributed processing. Depending on the size of the training data and the model, data parallel processing and model parallel processing are used. The well-known open-source distributed learning framework Horovod, which supports Keras, Tensorflow, PyTorch, and MxNet, supports both data and model parallel processing faster and easier than single framework-based distributed processing. GPU clustering requires AI computing resources, data and model optimization to improve learning performance, and in an Edge environment where AI computing resources are distributed, AI orchestration technology to integrate software for GPU clustering is important.

5.3.5.3 Manage your AI development environment

AI orchestration provides a development and validation environment for the development of AI services. It provides data processing and management tools to efficiently process the data required for AI model training. Data processing and model deployment tools improve development efficiency by automating data processing and model deployment that form the basis of AI model.

Integrated development environment enables AI developers to quickly develop and validate AI models. AI development tools and framework support are needed to reduce the complexity and increase the efficiency of the AI development process. It supports MLOps environments to minimize repetitive tasks and continuously improve model performance by automating the training and deployment functions of AI models. The AI development environment shortens the AI development cycle and continuously improves the quality of AI services.

5.3.5.4 Customer Care

In order to use AI computing resources, it is necessary to support product management and purchase support. AI orchestration provides a portal functionality that allows customers to access and use AI services, monitor AI computing resources, and manage AI services. The customer portal should be equipped with a variety of AI computing resources, development environments, and AI service products for customers to choose from and use.

Through the customer portal, customers should be able to see which AI computing resources are available for purchase, and where they can manage and monitor the AI computing resources they have purchased. It is also important for customers to be able to anticipate and evaluate the use of AI computing resources on their own, and to suggest the optimal use of AI computing resources to customers. It also provides customer support services, such as technical support, troubleshooting, maintenance, and community.

Customers need to be able to effectively leverage AI computing resources and receive technical support for achieving their business goals. Continuous technical support and maintenance should be provided for offering stable and reliable AI services. Customers should be able to easily manage and monitor AI computing resources, flexibly adjust contract terms, and use services cost-effectively.

6. Closing Remarks

SK Telecom has a vision of transforming its Telco network infrastructure into a Telco Edge AI infrastructure, and through this, it expects to create new value and business models that customers can experience, ultimately enhancing the sustainability of its telco business.

SK Telecom has been honored to be listed as an IEEE milestone in 2024, recognizing its efforts in innovation following its history of numerous challenges since achieving the world's first commercialization of CDMA.



[Figure 21] 'CDMA Commercialization' listed as an IEEE Milestone for the first time in Korea

Looking back at SK Telecom's history of continuous challenges and innovations over the long term since the world's first commercialization of CDMA in 1996, the innovation and transformation of the telco network will not be a task that can be completed in a short period of time. However, we believe that this moment, where we aim to accumulate even small but meaningful quick-win cases through collaboration with numerous partners in the ICT sector and establish systematic milestones to achieve our mid- to long-term goals, will serve as a new starting point that marks another significant milestone in the mobile communications industry.

References

[1] SK Telecom, "5G Lessons Learned, 6G Key Requirements, 6G Network Evolution, and 6G Spectrum," SK Telecom 6G White Paper, Aug. 2023. [Online]. Available: https://bit.ly/4a27B9I

[2] Deloitte Insights Report, "The Evolution of the Telecommunications Industry Creates New High-Level Value," Apr. 2024.

[3] ITU-R M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond", Nov. 2023

[4] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. -J. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," IEEE Communications Magazine, vol. 57, no. 8, pp. 84-90, Aug. 2019

[5] ITU-R M.2083-0, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond", Sep. 2015

[6] McKinsey & Company, "The economic potential of generative AI: The next productivity frontier", Jun. 2023

[7] Value Market Research, "Global AI In Telecommunication Market Research Report –Industry Analysis, Size, Share, Growth, Trends and Forecast 2023 to 2030," Nov. 2023

[8] "Competitiveness Diagnosis and Policy Direction of the Communication Equipment Industry by Value Chain," KIET Monthly Industrial Economics, vol. 285, pp. 68-80, June 2022

[9] SK Telecom, S1-241246, "SK Telecom's View on Future Telco. Infrastructure", 3GPP SA1#106, May 2024

[10] SK Telecom, "The path to AI Telco Infra Evolution", 6th O-RAN 6G Workshop: The Road to 6G Standards: Views from Industry and Academia, Jun. 2024

[11] K. Lee, et al., "Overview and Direction of AI Native Radio Access Network Research", Korean Telecommunications Society Summer Conference, Jun. 2024

[12] J. Oh et al., "Optimal Network Design Based on 6G Simulator", Korean Telecommunications Society Summer Conference, Jun. 2024

[13] M. Na, et al, "Operator's Perspective on 6G: 6G Services, Vision, and Spectrum", IEEE Communications Magazine, vol. 62, no. 8, pp. 178-184, Aug. 2024

[14] H. Lee, et al, "The Evolution Path of Telco Infrastructure: From Open RAN to 6G AI-RAN (Telco Edge AI)", in the 15th International Conference on ICT Convergence (ICTC) Workshop, Oct. 2024

[15] SK Telecom and NTT DOCOMO, "Key Considerations for vRAN: Insights from SK Telecom and NTT DOCOMO", Feb. 2024

[16] SK Telecom and Intel, "Toward 6G Core Architecture Using an Inline Service Mesh", Jan, 2024

[17] SK Telecom, S1-242019, "6G Area of Interest", 3GPP SA1#107, Aug. 2024

[18] H. Kim et al. "Frequency-Aware and Telemetry Item-Wise In-band Network Telemetry for Per-flow Measurement", IEEE INFOCOM, May. 2024

Abbreviations

Abbr.	Full Name	Abbr.	Full Name
3GPP	3rd Generation Partnership Project	**ITU-R**	International Telecommunication Union Radiocommunication Sector
AI	Artificial Intelligence	**LCM**	Life Cycle Management
AI DC	Artificial Intelligence Data Center	**LLM**	Large Language Model
AI-RAN	Artificial Intelligent-Radio Access Network	**LTE**	Long Term Evolution
AIX	AI Transformation	**ML**	Machine Learning
AP	Application Processor	**MLOps**	Machine Learning Operations
APN	All Photonic Network	**mMTC**	massive Machine Type Communications
APN-T	All Photonic Network Transceiver	**NPU**	Neural Processing Unit
APN-G	All Photonic Network Gateway	**NWDAF**	Network Data Analytics Function
APN-I	All Photonic Network Interchange	**PAA**	Personal AI Assistant
AR	Augmented Reality	**PoC**	Proof of Concept
ARPU	Average Revenue Per User	**QoS**	Quality of Service
B2B	Business-to-Business	**RAN**	Radio Access Network
B2C	Business-to-Customer	**RDMA**	Remote Direct Memory Access
CDMA	Code Division Multiple Access	**RoCE**	RDMA over Converged Ethernet
CPU	Central Processing Unit	**RU**	Radio Unit
DCI	Data Center Interconnect	**SDI**	Software Defined Infrastructure
DU	Digital Unit	**SLA**	Service Level Agreement
DX	Digital Transformation	**TCO**	Total Cost of Ownership
eMBB	enhanced Mobile BroadBand	**UALink**	Ultra Accessible Link
ETSI	European Telecommunications Standards Institute	**UAM**	Urban Air Mobility
FPGA	Field Programmable Gate Array	**UPF**	User Plane Function
GDR	GPU Direct RDMA	**URLLC**	Ultra Reliable Low Latency Communications
GDS	GPU Direct Storage	**V2X**	Vehicle to Everything
GPU	Graphic Processing Unit	**VM**	Virtual Machine
GTAA	Global Telco AI Alliance	**VR**	Virtual Reality
ICT	Information & Communication Technology		
IEEE	Institute of Electrical and Electronics Engineers		
IMT	International Mobile Telecommunication		
IOWN	Innovative Optical and Wireless Network		
IP	Internet Protocol		
ISAC	Integrated Sensing And Communication		

Change History

Date	Descriptions	Version
October 15th, 2024	Initial Release	1.0


SK telecom