Key Enabling Features for the Evolution of vRAN and the Path to AI-RAN

March 2026

SK TELECOM & NTT DOCOMO WHITE PAPER

Table of Contents

List of abbreviations

CAPEX: Capital Expenditures

CU: Central Unit

DU: Distributed Unit

L1: Layer 1

L2: Layer 2

LCM: Life Cycle Management

MEC: Multi-access Edge Computing

MIG: Multi-Instance GPU

MNO: Mobile Network Operator

MPS: Multi-Process Service

OPEX: Operational Expenditures

RAN: Radio Access Network

RU: Radio Unit

SMO: Service Management and Orchestration

TCO: Total Cost of Ownership

vRAN: virtualized RAN

vUPF: virtual User Plane Function

WL: Workload

1 Executive summary

This white paper serves as a continuation of the first vRAN Joint White Paper published in February 2024 by SK Telecom and NTT DOCOMO [1]. Over the past two years, the vRAN ecosystem has made steady progress across various areas—including Layer 1 accelerators and management/orchestration technologies—and has also expanded in terms of commercialization scale. However, the deployment of core virtualization-specific capabilities—such as resource pooling and scaling—remains limited as many development efforts continue to prioritize basic system stability.

To maximize the benefits of vRAN for MNOs, this paper highlights three key enabling features:

- **Hardware/Software Strict Separation**: Decoupling RAN software from specific hardware or virtualization platforms to eliminate vendor lock-in and enable rapid software-driven innovation. This allows MNOs to upgrade RAN technologies (e.g., from 5G to 6G) simply by replacing software on existing hardware platforms.
- **Resource Pooling**: Enabling flexible utilization of computing resources of vRAN servers within and across multiple servers. This minimizes idle capacity and optimizes infrastructure utilization, allowing MNOs to maintain service quality while reducing the required number of servers and overall power consumption. As such, resource pooling is a key differentiator that enables improvement of vRAN TCO.
- **Enabling AI Computing with vRAN**: Evolving RAN from a communication-only system into an integrated AI platform. By leveraging vRAN's high-performance hardware and xPUs (e.g., CPU, GPU, NPU, etc.), MNOs can not only enhance RAN workload performance but also support AI workloads while ensuring RAN service quality through an intelligent and SLA-aware orchestration.

These features are essential for validating the economic and operational advantages of vRAN and for laying the foundation for AI-RAN. It is critical that vendors prioritize items to develop these functionalities and that the broader ecosystem collaborates to deliver these capabilities. By actively adopting these features, MNOs can establish vRAN as the core, future-proof infrastructure for next-generation networks.

2 Introduction

This chapter provides the necessary context for understanding the evolution of vRAN. It begins by defining the core values and global adoption trends of vRAN, highlighting the shift from proprietary hardware to flexible, software-defined architectures. Building upon the foundational considerations established in the first joint white paper published in 2024 [1], this section examines how the vRAN ecosystem has matured through advancements in acceleration, management, and AI integration. Finally, it outlines the specific scope and purpose of this document, which focuses on identifying and realizing the essential differentiating features required to unlock the full potential of vRAN in commercial environments.

2.1 Overview of vRAN

As the mobile communications industry rapidly evolves, network operators are required to deliver faster data rates, higher network capacity, and increasingly sophisticated services to meet rising user expectations and support new applications. The advent of 5G, and the continued evolution of LTE, have significantly increased both the volume and complexity of base-station processing. Traditionally, these demands were addressed through highly specialized hardware and tightly coupled software, optimized for maximum performance in dedicated base-station equipment. However, in parallel, the broader IT landscape has witnessed remarkable innovation in cloud computing, virtualization, and general-purpose hardware performance. This stark contrast in design philosophies has opened an opportunity for transformation in the RAN domain.

The concept of a vRAN represents a shift away from proprietary, monolithic base-station architectures towards flexible, disaggregated systems enabled by virtualization and cloud-computing technologies. In vRAN, base-station functions—classically intertwined with dedicated hardware—are decoupled from the underlying physical infrastructure and instantiated instead as software on generic, high-performance hardware platforms. This decoupling allows MNOs to leverage advances in commodity IT hardware, benefit from fast innovation cycles, and respond dynamically to fluctuating traffic demands by scaling resources elastically.

The transition to vRAN is motivated by the potential for multiple, significant benefits:

- Cost Optimization: By adopting general-purpose hardware, mobile operators can achieve economies of scale, reduce vendor lock-in, and optimize infrastructure investment. Frequent hardware refreshes can also bring performance and energy efficiency gains.
- Operational Agility: Decoupling hardware from software enables rapid innovation and flexible network upgrades, with streamlined deployment of new vendors or features via software alone. Virtualization and automation contribute to faster rollouts and less on-site installation work.

- Intelligent Management: By leveraging software-defined principles, vRAN supports centralized and intelligent resource orchestration, empowering MNOs to automatically adjust capacities, recover from failures, and manage upgrades end-to-end—from the edge to the network core.
- Network Slicing & Service Innovation: The programmability of vRAN forms a basis for supporting differentiated services, including dynamic network slices optimized for diverse application requirements—spanning enhanced mobile broadband, IoT, and ultra-reliable low-latency communications.

The adoption of vRAN solutions is progressing globally, as many MNOs pursue network modernization. Across various regions—including Asia, Europe, and North America—leading operators are conducting large-scale vRAN deployments, promoting open ecosystems, and collaborating with technology partners to validate new approaches. Commercial vRAN rollouts and pilot projects are increasingly common, aiming to achieve cost reduction, diversify vendor sources, and enhance operational flexibility.

These global trends highlight the growing recognition of vRAN's benefits, such as improved efficiency and agile network management. While technical and operational challenges remain, vRAN is expected to play an essential role in supporting the evolution of next-generation mobile networks worldwide.

Nevertheless, introducing vRAN entails significant technical challenges specific to the unique, real-time processing requirements of the RAN. These include ensuring ultra-low latency and high-throughput radio-layer processing on general-purpose hardware, supporting diverse, distributed site conditions, and unifying operations for a heterogeneous mix of physical and virtual resources. There is a critical need to balance the flexibility and efficiency promised by virtualization with the uncompromising performance expected in mobile wireless networks.

In recognition of these imperatives, leading MNOs have initiated collaborative efforts to standardize and commercialize vRAN solutions. Initiatives such as OREX® bring together technology partners to accelerate the development, integration, and deployment of open and interoperable vRAN systems. Leveraging experience gained through virtualizing core networks, these efforts aim to address technical and operational challenges, promote ecosystem growth, and realize the full promise of vRAN: enabling future-proof, cloud-native, and highly efficient radio access networks for 5G and beyond.

2.2 Key Considerations for vRAN

Our Joint White Paper [1] describes six critical considerations that were proposed:

- The evolution of the L1 accelerator is essential for enhancing performance of cell capacity and power consumption

- Key features for vRAN such as resource pooling, scaling and auto-healing should be prioritized
- Energy efficiency is an essential KPI for MNOs to move towards vRAN
- The enhancement of management technologies, tools, and procedures plays a crucial role in improving the integration of vRAN
- TCO, network controllability, and standardization should be considered for future network based on cloud and AI native
- Security of vRAN equipment is crucial

Looking back over the two years since the publication of the joint white paper, vRAN technologies and the surrounding ecosystem have continued to evolve in close alignment with the above considerations. Regarding L1 accelerator, newer generations of accelerators featuring increased core counts and enhanced acceleration performance have been introduced. Furthermore, L1 processing based on general-purpose accelerators, such as GPUs, has recently been adopted, expanding the range of approaches beyond RAN-specific acceleration solutions.

For an energy efficiency, ongoing advancements in both hardware and software have steadily improved the power efficiency of vRAN deployments. In relation to vRAN management technologies, the standardization of SMO and O-Cloud—driven by O-RAN ALLIANCE—and related technology development by vendors have made vRAN management and integration easier. For realizing intelligent RAN utilizing AI capability—which is also known as AI-RAN, AI-Centric RAN and AI-native RAN—(hereinafter called “AI-RAN”), vRAN has established itself as a core enabling technology, which has recently gained considerable industry attention. On a security perspective, security requirements have been actively advanced in line with the principles of zero-trust architecture, such as the activities of O-RAN ALLIANCE work group 11.

However, the development and delivery of core vRAN features—such as resource pooling, scaling, and auto-healing—remain slow, with limited vendor progress compared to other areas. Moreover, leveraging virtualization to enable base stations to provide AI computing capabilities has also emerged as a new approach, though it remains at an early stage.

2.3 Scope and purpose of this white paper

The purpose of this white paper is to present key enabling features for the evolution of vRAN and the path to AI-RAN that can maximize its benefits for MNOs. While the performance and TCO of vRAN continue to improve through advances in hardware and software technologies, realizing its full potential requires dedicated support for virtualization-specific capabilities.

Currently, vendor development and support for such capabilities—those that distinguish vRAN from legacy, non-virtualized RAN—remain limited. This is largely

because most development efforts prioritize improving the performance and stability of the vRAN system over introducing new, differentiated capabilities.

From an MNO perspective, this white paper identifies the essential differentiating features required for vRAN evolution and its progression toward AI-RAN, and explains their criticality. This paper provides the concept, challenges, requirements, detailed components, and expected benefits of each feature, thereby providing strategic guidance for the continued advancement of vRAN.

3 Key Enabling Features

This chapter identifies and explores the essential differentiating features required for the evolution of vRAN to maximize its value for MNOs. While an evolution of hardware and software continue to improve baseline performance, the full potential of vRAN is realized through specific capabilities that distinguish it from legacy, non-virtualized RAN. For each feature, the following sections provide a comprehensive analysis of the core concepts, technical challenges, requirements, and benefits, offering strategic guidance for the continued advancement of vRAN in commercial networks.

3.1 Hardware/Software strict separation

The equipment of legacy base station has relied on “purpose-build” hardware, which is specifically designed for RAN to meet strict requirements. This will result in tightly integrated hardware and software from a single vendor and limiting MNOs’ flexibility.

vRAN allows MNOs to flexibly tailor functions and performance to their needs by combining general-purpose servers with software from any vendor, thereby optimizing network performance and reducing costs.

This chapter outlines the requirements, challenges, and benefits associated with achieving hardware and software separation in base station equipment.

3.1.1 Overview

Historically, base station equipment has been required to perform real-time, low-latency processing to meet the specific requirements of RAN. To meet the requirement, legacy base station has been implemented by purpose build hardware, which is specific to RAN, and the hardware and software are tightly coupled, integrated by single vendor. It has limited MNOs’ flexibility in selecting functions and performance according to their needs—resulting in vendor lock-in.

Meanwhile, the IT industry has witnessed remarkable technological evolution, including the use of general-purpose hardware by improvements on hardware

performance and the spread of virtualization technologies, which enables the separation of hardware and software. These innovations have made it possible to realize vRAN. Through the vRAN, MNOs can construct networks that flexibly adapt functions based on their requirements. Specifically, by utilizing general-purpose servers for hardware and selecting software from any vendor, operators can optimize network performance and reduce costs.

3.1.2 Challenges and Requirements

To realize vRAN hardware and software separation, MNOs procure hardware and software separately. While each vendor guarantees the performance and quality of their respective components (servers, virtualization platform, vRAN software, hardware accelerators, etc.), the MNO or system integrator has to perform system integration—combining these components and guarantee the performance and quality of the base station as a whole.

NTT DOCOMO has taken the initiative to act as a system integrator, leading the construction and deployment of vRAN through separate procurement of hardware and software. By assembling components with servers, accelerators, routers, and other necessary hardware and software from entirely different vendors, NTT DOCOMO has successfully implemented and introduced multi-vendor solutions.

One of the challenges is to necessitate enhanced mutual understanding of specifications, visualization and alignment of the process and time plan of the integration, and effective communication across numerous stakeholders and players in both hardware and software domain.

The other challenge, which is a requirement for vendors, is that complete separation of hardware and software has not yet been fully achieved. For example, certain software solutions require specific accelerators, or can only be deployed on a particular virtualization platform. Eliminating these constraints would enable MNOs to procure “best of breed” equipment with greater flexibility, while vendors could expand into new market segments, further accelerating the adoption of vRAN.

3.1.3 Benefits

By implementing base station systems that separate hardware and software through vRAN, competitive principles can be applied for the base station market, which was previously subject to vendor lock-in. It is expected to improve both CAPEX and OPEX. The adoption of hardware–software separation structure enables more efficient implementation of resource pooling functionality, which further reduces the cost.

Moreover, the separation makes software update easier. One of the use cases is that MNOs could pre-build hardware and subsequently upgrade RAN technologies from 5G to 6G simply by replacing the software.

Even in rapidly evolving fields such as artificial intelligence (AI), leveraging general-purpose CPUs and GPUs in vRAN can foster the creation of new business opportunities and the development of service platforms.

3.2 Resource Pooling for vRAN

Resource pooling is a key enabler for maximizing TCO efficiency in vRAN. Although vRAN on general-purpose hardware may seem less competitive in capacity and power consumption than dedicated hardware, pooling can match or surpass conventional base stations by flexibly aggregating computing resources across cells and functions. This chapter outlines its background, concept, deployment options, challenges, and requirements, along with case studies illustrating potential TCO savings and its value for vRAN evolution.

3.2.1 Overview

In conventional RAN deployments, MNOs tend to over-dimension the computing capacity of CU and DU. This provisioning strategy is intended to accommodate peak-hour traffic and maintain network stability during special events such as large public gatherings.

However, this approach often results in excessive CAPEX investments and prolonged periods of resource under-utilization. A substantial portion of base station computing assets remain idle outside of peak periods, contributing to both capital and operational inefficiencies.

vRAN resource pooling can address this by virtualizing and sharing computing resources across base stations, minimizing idle capacity and optimizing infrastructure utilization.

From a definitional perspective, this feature can be described as follows:

- Resource: A collective term for computing and RAN resources.
  - Computing resource: The physical computational assets of a server, such as CPU, accelerators, memory, and related hardware components.
  - RAN resource: The logical assets and functions in the RAN, such as Layer 1 (PHY), Layer 2 (MAC, RLC), Layer 3 (PDCP, RRC), OAM, etc.
- Pooling: A technology that organizes a set of homogeneous resources into a shared and centrally managed pool, enabling dynamic allocation and flexible usage based on demand.

- Resource pooling for vRAN: A feature that enables RAN resources to share the computing resources of a vRAN infrastructure, allowing efficient scaling and utilization across multiple cells or functions.

The unique characteristics of vRAN make it inherently suitable for implementing resource pooling. Key enabling aspects include:

- Decoupling hardware and software via virtualization: vRAN removes the tight binding between computing and RAN resources, allowing computing capacity to be flexibly assigned where needed.
- Virtualization platforms with elastic scaling: Modern virtualization platforms, such as Kubernetes, provide the ability to scale up/down (within a server) and scale in/out (across servers), enabling dynamic allocation of RAN resources in line with demand.
- Standardization-based management through O-RAN: O-RAN ALLIANCE work group 6 is standardizing the O-Cloud layer, while SMO frameworks establish a standards-based foundation for managing vRAN at scale.

Resource pooling yields clear benefits for MNOs in multiple operational scenarios, including 1) overload control, 2) power saving, and 3) reduction in the number of servers.

Overload control scenario refers to situations where RAN resource utilization on a specific server suddenly increases, leading to a shortage of computing resources. In such cases, RAN workloads can be offloaded to idle servers during events to maintain stable service quality shown in [Figure 1.(a)]. This ensures that MNOs can deliver consistent service performance to subscribers even during sudden traffic surges.

Power saving scenario applies in low-traffic conditions, where RAN resources from servers with low computing resource utilization are migrated to other active servers, allowing unused servers to be powered down [Figure 1.(b)]. This reduces overall base station power consumption and lowers OPEX.

Reduction in the number of servers scenario leverages flexible RAN resource pooling to increase the utilization of computing resources, thereby reducing the total number of required servers [Figure 1.(c)]. For example, consider a service region where the average PRB utilization is just 15%, and the probability of the majority of cells simultaneously reaching peak PRB utilization is low (which becomes more reasonable as the target service region size increases). In vRAN deployments without resource pooling, cell capacity is typically dimensioned based on peak traffic conditions (e.g., 100% PRB utilization), and a fixed number of servers must be deployed regardless of actual PRB usage. In contrast, with resource pooling in vRAN, computing resources can be dynamically shared among RAN functions overload, enabling the region to be served with fewer servers—resulting in measurable CAPEX savings.

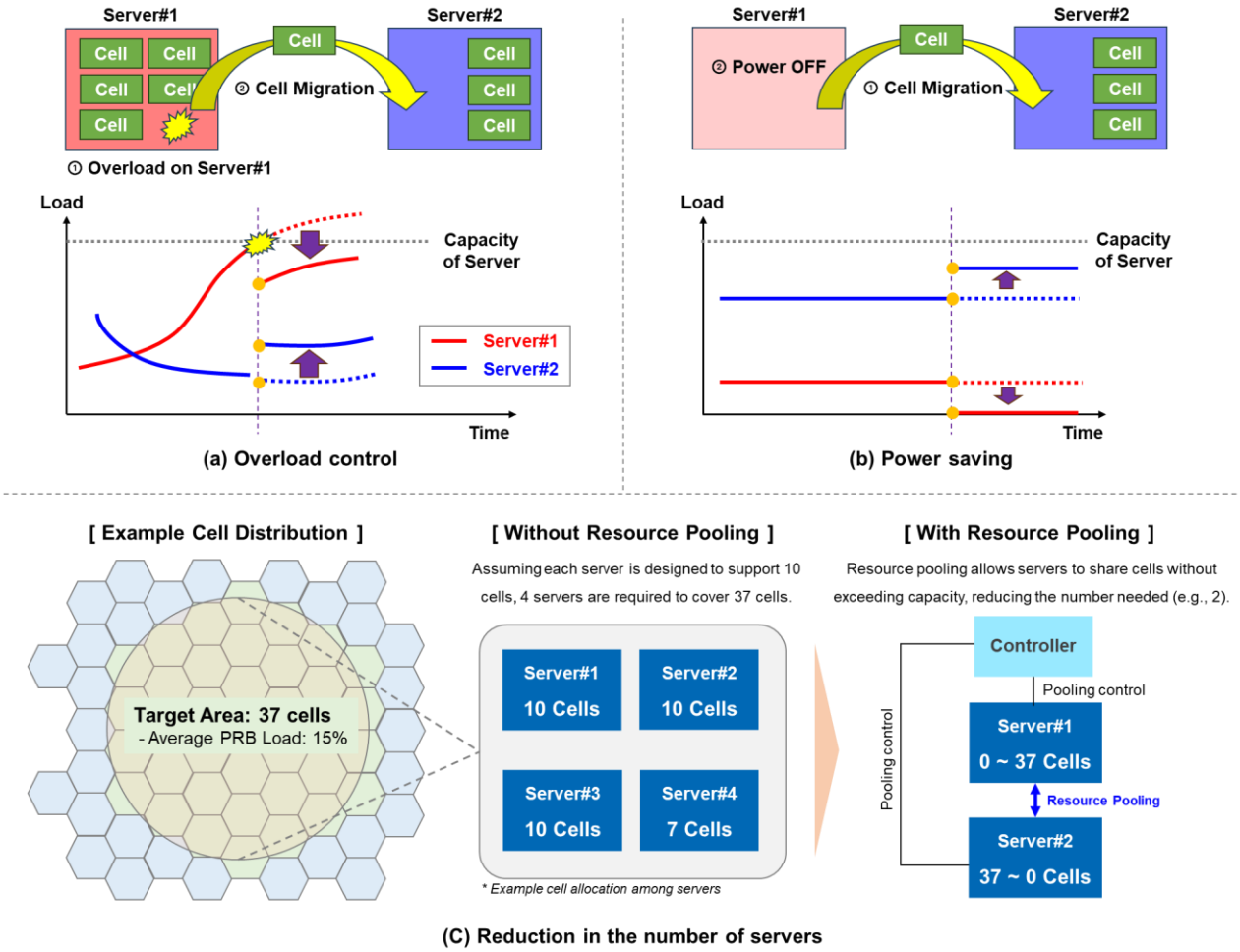


Figure 1. Resource Pooling Scenario Examples

3.2.2 Deployment Options for Resource Pooling in vRAN

Resource pooling in vRAN can be implemented through a range of models, operational modes, and configurations of RAN resource type and granularity. The choice among these options dependent on network architecture, operational scenarios, and deployment constraints. The chosen approach impacts pooling efficiency, implementation complexity, and orchestration requirements. This section describes the pooling models, their applicability to different network architectures, associated fronthaul considerations, operational modes, and key design factors.

Resource pooling models can be broadly classified into the following categories:

- Vertical Pooling (Intra-server): Computing resources within a single server are shared among multiple RAN resources. The allocation of these resources is adjusted according to varying demand.
- Horizontal Pooling (Inter-server): Computing resources across multiple servers are aggregated into a shared resource pool, which is accessible to multiple RAN resources regardless of server boundaries.
- Hybrid Pooling: Combination of vertical and horizontal pooling to maximize utilization within and across servers.

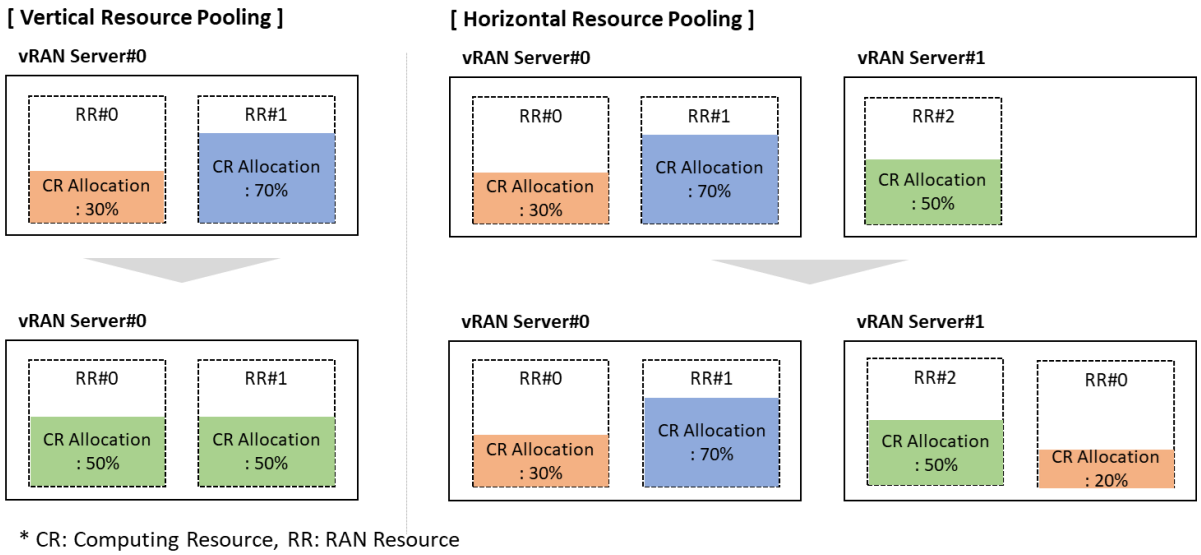


Figure 2. Example scenario of vertical and horizontal resource pooling

The applicability of these models varies with network architecture. Major examples of deployment scenario are described in [Figure 3].

In Distributed RAN (D-RAN), where baseband units and radio units are co-located at each cell site, the limited number of servers per site typically favors vertical pooling. In Centralized RAN (C-RAN), multiple baseband units are located in a central office, making hybrid pooling a viable choice thanks to co-location of multiple servers. In a mixed D-RAN/C-RAN configuration—where DUs are located at cell sites while CUs are centralized—vertical pooling can be applied to DUs and hybrid pooling to CUs.

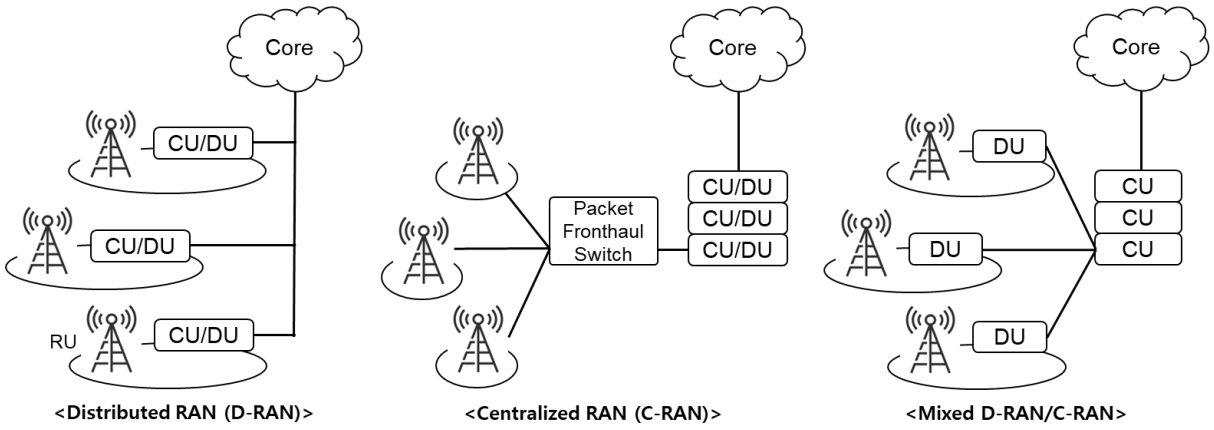


Figure 3. RAN deployment architecture

Horizontal pooling in C-RAN necessitates specific fronthaul capabilities. Changes in the DU server assigned to process an RU's signal require a fronthaul network capable of flexible RU-DU connectivity. This is typically achieved with packet-based fronthaul solutions such as eCPRI switches, provisioned for sufficient capacity. The fronthaul switch topology—such as star, tree, or other suitable configurations—should be engineered according to server pool composition and capacity requirements.

Effective pooling requires selecting the appropriate RAN resource type and pooling granularity shown in [Figure 4]. Lower-layer pooling (e.g., PHY) can offer high efficiency gains but is limited by strict latency requirements. Larger granularity (e.g., DU) simplifies management but reduces flexibility, while smaller granularity enables fine control with added complexity. The chosen combination should align with vRAN architecture, performance goals, and implementation feasibility.

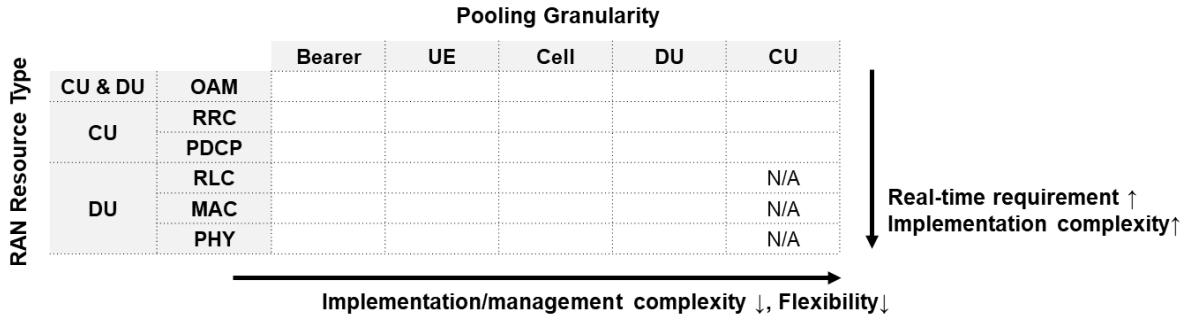


Figure 4. Characteristics of RAN resource type and pooling granularity combinations (exemplary)

3.2.3 Challenges and Requirements

Although resource pooling is one of the key features capable of maximizing the benefits of vRAN, there are multiple challenges for MNOs to adopt the feature into their network. This section addresses the main challenges in deploying resource pooling in commercial networks and outlines MNO-oriented requirements for its implementation.

Hardware and Software Dependency

While vRAN generally reduces the tight coupling between hardware and software compared to traditional purpose-built RAN systems, certain dependencies still exist. For example, software may rely on specific hardware elements such as Layer 1 accelerators, limiting the full flexibility of resource pooling, as resources from servers equipped with different Layer 1 accelerators cannot be utilized. Achieving complete decoupling of hardware and software is essential to support unrestricted pooling operations. Industry standardization efforts, such as the Acceleration Abstraction Layer under O-RAN ALLIANCE, may play a key role in enabling this separation. This requirement aligns closely with the hardware/software strict separation capability discussed in Section 3.1.

Potential Service Impact During Pooling Operations

Reallocation of computing resources across RAN resources can introduce service impacts, depending on the implementation method. For example, during the reallocation of specific cell resources to another server, there may be a temporary

inability for that cell to provide service. For MNOs, even short service interruptions can be critical. Therefore, it is vital for resource pooling solutions to be designed and operated in a way that eliminates—or at least minimizes—service disruption during transitions. Vendors are expected to deliver solutions with maximum feasible flexibility while ensuring service continuity, and additionally, MNOs are also required to have mechanisms and operations in place to maintain service continuity.

Network Architecture Limitations

From the MNO's perspective, an existing D-RAN-centric network architecture can be a bottleneck to the full adoption of resource pooling. The greatest benefits are achieved in C-RAN, where horizontal pooling can be deployed across co-located servers. This advantage should be considered when evaluating evolution paths toward C-RAN. Vendors should also ensure that their solutions provide sufficient fronthaul distance from baseband units to RUs.

Implementation Priority

The most critical requirement is the actual availability of resource pooling functionality from vRAN vendors. At present, driven by feature prioritization, implementation complexity, and market considerations, many vendors adopt a cautious stance toward delivering resource pooling capabilities. However, given its potential to significantly enhance the value of vRAN deployments, the provision of this feature should be regarded as a high-priority development objective.

In summary, overcoming the hardware-software dependency, minimizing service impact, adapting network architecture, and ensuring vendor commitment are essential steps to enable successful deployment of resource pooling in vRAN ecosystems. The next chapter presents case studies demonstrating benefits achieved through deployment of resource pooling.

3.2.4 Quantitative Analysis of Benefits

This section quantitatively analyzes the benefits of adopting resource pooling in vRAN, focusing on real-world commercial network scenarios. Multiple case studies are used to demonstrate the effectiveness and necessity of the feature.

As discussed in Section 3.2.1, resource pooling can reduce the number of servers required to operate a vRAN and lower overall power consumption—both of which are closely tied to CAPEX and OPEX. Based on actual commercial network data, this analysis aims to quantify:

1. The reduction in server count required to serve a target region.
2. The savings in power consumption.

These are compared between cases with and without resource pooling to demonstrate why the feature is essential for vRAN deployment.

Key analysis assumptions:

- Resource Pooling Model: Hybrid; RAN Architecture: C-RAN; Pooled RAN Resource Types: Layer 1 and Layer 2.
- Computing resources for Layer 1 scale linearly with PRB utilization, while Layer 2 resources scale with the number of active UEs. Other factors affecting computing demand are excluded (conservative gain estimation).
- Max capacity is defined as PRB utilization at 100% with 600 active UEs per cell. At this capacity, vRAN compute demand per resource type is assumed to be: Layer 1: 50%, Layer 2: 30%, Others (Layer 3, OAM, etc.): 20%.
- Server power consumption is assumed to remain constant regardless of the number of cells or traffic load, with powered-off servers consuming 0 W.

Dataset: The analysis targets one central LTE DU hub (1,539 cells) and one 5G DU hub (709 cells) in dense urban areas. Based on SK Telecom's commercial network data, one week of 5-minute interval measurements includes downlink PRB usage and active UE counts.

Reference Case (No Pooling): As described in Section 3.2.1, MNOs dimension their networks conservatively to handle peak-hour traffic and maintain stability during special events (e.g., large public gatherings). Server count is thus sized for maximum capacity conditions.

Pooling Scenarios: Two idealized scenarios are analyzed—(1) Layer 1 pooling only, and (2) combined Layer 1 and Layer 2 pooling, assuming full flexibility with no performance degradation or overhead. For power savings, workloads are consolidated to maximize server utilization, and any idle servers are powered off.

[Table 1] summarizes results for each case:

- LTE DU Hub: Layer 1 pooling reduces server count to 64% of baseline (power to 56%), while Layer 1+2 pooling lowers it to 36% (power to 27%).
- 5G DU Hub: Gains are greater—server counts fall to 55% and 26% of baseline, with power usage reduced to 52% and 23%, driven by higher PRB and UE loads in LTE cells than in 5G cells in this region.

Resource Pooling	LTE DU Hub		5G DU Hub	
	Required Server Count	Total Power Consumption	Required Server Count	Total Power Consumption
Disabled (Reference)	100% (Ref.)	100% (Ref.)	100% (Ref.)	100% (Ref.)
Enabled (L1 Only)	64%	56%	55%	52%
Enabled (L1 and L2)	36%	27%	26%	23%

Table 1. Effect of Resource Pooling on Server Count and Power Savings

Results show that resource pooling can significantly reduce network investment and operational costs. Targeting suburban or rural regions, where PRB usage and active UE counts are lower, may yield even greater relative gains in server count and power savings. Although the findings are based on an idealized pooling model and gains may vary with the traffic characteristics of the target region or changes in the underlying analysis assumptions, results highlight resource pooling as a key enabler for vRAN TCO reduction and a major differentiator from conventional RAN, making it essential for vRAN evolution and large-scale deployment.

3.3 Enabling AI computing with vRAN

In the 6G era, mobile network systems are expected to undergo fundamental changes across multiple dimensions. From the perspective of users and the broader service ecosystem, the widespread adoption of AI-driven applications—such as Physical AI—will introduce traffic characteristics that differ markedly from those of traditional RAN. In particular, the expansion of large-scale AI model usage will lead to a greater need for distributed computing structures capable of processing AI data efficiently.

From the MNO's perspective, the growing adoption of vRAN creates a need to enhance RAN performance by incorporating high-performance hardware, xPU (e.g., CPU, GPU, NPU, etc.). At the same time, global ecosystem players such as NVIDIA are proposing architectural models that integrate hardware accelerators and AI-optimized frameworks enabling vRAN system to process AI workloads natively. These changes not only promise tangible improvements in network performance through the integration of AI technologies, but also suggest that the RAN—traditionally dedicated solely to mobile access functions—may evolve into an edge computing platform capable of supporting AI services.

Along with this backdrop, it is increasingly important to examine what MNOs will prepare in order to support AI workloads within the RAN system, as well as the technical considerations required to enable this capability. This section elaborates the concept of "Enabling AI Computing in vRAN" and provides a structured analysis of the architectural and technological requirements for its realization.

3.3.1 Overview

Enabling AI computing in vRAN refers to the functionality that allows MNOs to execute AI computations from external AI service providers within MNO-managed RAN infrastructure.

This expands the role of RAN—previously limited to communication functions—into that of an AI computing platform. By utilizing computing resources located at base stations closest to the end users, AI inference and processing can be delivered with

significantly lower latency, improving service responsiveness and reducing traffic load.

This concept may be similar to MEC, but it is differentiated in two key aspects:

First, it goes beyond simply placing the service provider's servers in close physical proximity, by seeking ways to share essential AI computing resources with existing base station infrastructure. Second, it aims to provide differentiated functions and changes within base stations, such as localized vUPF.

The detailed scope of this feature includes:

Computing Support for AI Workload

Providing AI workload processing capabilities directly within the base station infrastructure.

Maximizing Hardware Resource Utilization in vRAN

Fully leveraging hardware resources (e.g., CPU, GPU, etc.) deployed in vRAN servers through efficient resource-sharing mechanisms to support heterogeneous workloads.

Supporting orchestration inside and between servers for efficient handling of heterogeneous workloads.

vRAN Architecture Design changes

Expanding orchestration capabilities and adding an entity (e.g., vUPF) for low latency.

SLA Compliance and RAN Service Quality Assurance for Each Workload

Ensuring both RAN and AI workloads meet their respective SLAs and enforcing these guarantees through orchestrator-driven decisions.

3.3.2 Challenges and Requirements

To support RAN workloads for communication services and AI workloads for AI services, changes to the existing vRAN system will be considered. This section introduces related considerations and requirements

Impact analysis on vRAN architecture

Integrating AI computing capabilities into vRAN requires a comprehensive assessment of the impact on DU/CU components and the entire RAN architecture.

New capability additions should not degrade existing radio access functions or harm RAN performance, and minimizing structural changes can be recommended. Additionally, compatibility with DU/CU resource management, timing control, and

operational automation functions are maintained to prevent conflicts with AI workloads.

Establishing architectural consistency ensures that RAN can maintain its original performance guarantees while seamlessly integrating AI computing capabilities.

Defining orchestration criteria for heterogeneous workload delivery and prioritizing RAN workloads

In order to provide RAN and AI workloads in the RAN system, resource orchestration for each workload's operation is required, along with the proper orchestration criteria.

The orchestration criteria may include system status, SLAs for each workload, traffic predictions, and etc. However, in cases where SLAs for heterogeneous workloads conflict, it is reasonable to prioritize quality of services for RAN in order to maintain essential network quality.

Support for heterogeneous systems

To support AI workloads in parallel, changes in the hardware configurations of existing DU/CU servers are needed, and it is impossible for all servers to have the same configuration. The types and manufacturers of xPU installed in each server vary, and depending on various environmental factors, the configurations of servers supporting AI workloads also vary.

Therefore, orchestration for multiple heterogeneous systems will be performed based on standardized interfaces.

Hardware resource-sharing technologies to ensure seamless RAN/AI service

Since CPU, GPU, and various accelerator resources can be shared between RAN and AI workloads within vRAN servers, advanced hardware resource-sharing technologies are essential. The two types of workloads have significantly different and time-varying resource demands. To meet SLA requirements, fixed allocation methods are insufficient.

Therefore, technologies are needed to flexibly allocate GPU and CPU resources according to workload characteristics, along with real-time adjustment functions to reflect changes in demand.

3.3.3 vRAN architecture to enabling AI Computing

This section describes the requirements identified in 3.3.2 into a concrete, operator-oriented architecture that allows RAN and AI workloads to coexist on the same vRAN infrastructure without compromising carrier-grade performance.

At a high level, the architecture decomposes orchestration responsibilities and introduces an in-server Resource Management layer to arbitrate heterogeneous compute (CPU/GPU/other xPUs) between RAN and AI workloads.

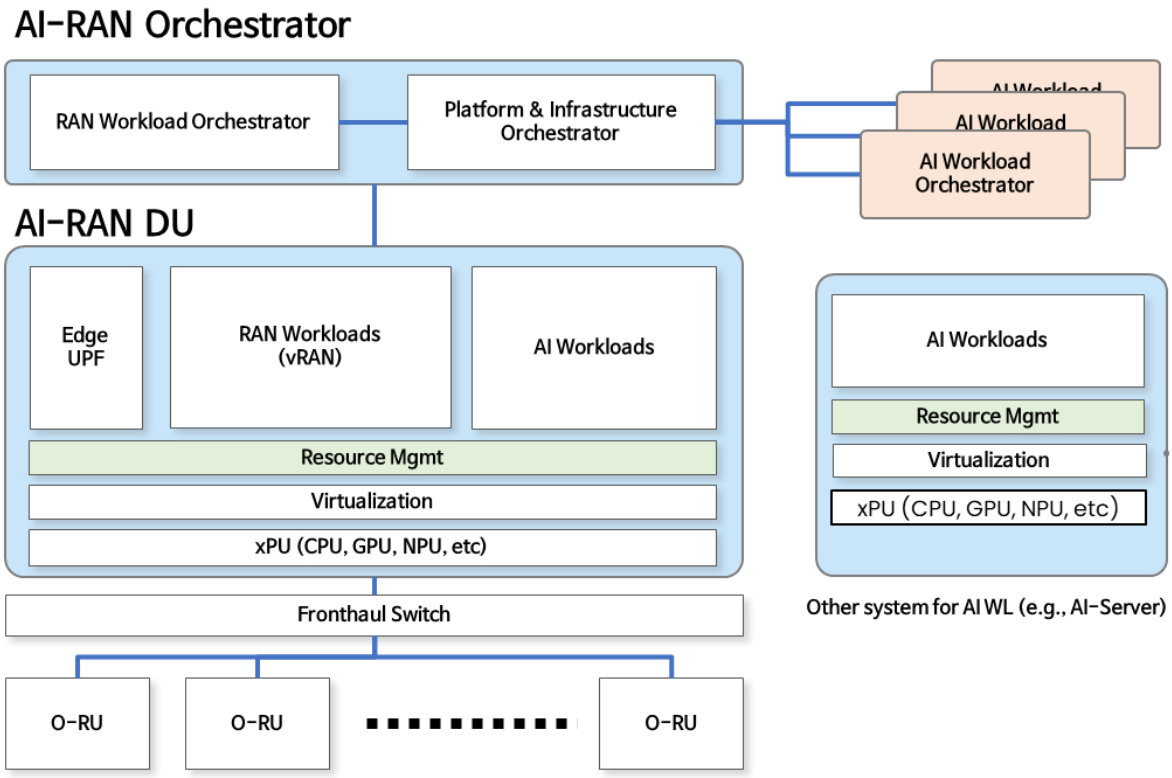


Figure 5. Architecture of vRAN to enabling AI Computing (exemplary)

Here are the main components and roles which are described below.

Platform and Infrastructure Orchestrator

The Platform and Infrastructure Orchestrator serves as the central entity for high-level resource management, mainly focusing on the roles of resource orchestration and deployment decision. It acts as an umbrella orchestrator that converges workload-specific orchestrators, interconnecting with both the RAN WL Orchestrator and the AI WL Orchestrator.

When a deployment decision based on orchestration criteria is performed here, the LCM and policy enforcement for each workload are delegated to workload-specific orchestrators.

Workload-Specific Orchestrator

To address the distinct requirements of different network functions, orchestration is segmented into specialized domains:

- RAN WL Orchestrator**:** Enforces carrier-grade reliability (i.e., SLA) for RAN workloads and manage their LCM to guard minimum guaranteed SLA for RAN services.
- AI WL Orchestrator**:** Manages LCM for AI workloads, and handles advanced features for AI services (e.g., dynamic offload decisions to AI Data Centers)

based on AI workload's compute and latency requirements. AI data traffic handled by AI WL orchestrator would forward to AI workloads in AI-RAN DU directly through localized vUPF within the vRAN sever, in order to facilitate low-latency AI services.

Resource Management

Central to the coexistence of telecommunications and intelligence is the AI-RAN Resource Management layer. It controls the hardware resource allocation between RAN workloads and AI workloads. By dynamically arbitrating resources, it ensures that critical RAN functions maintain performance stability while maximizing the utility of available compute for AI applications. For example, in an initial configuration, a 3G MIG instance may be allocated to the RAN workload, while a 2G MIG instance is assigned to the AI workload. When an increase in RAN traffic is anticipated, the orchestration system can proactively expand the xPU allocation for the RAN workload to a 4G MIG instance by correspondingly reducing the xPU resources assigned to the AI workload, thereby prioritizing RAN service continuity.

Several variations for heterogeneous systems

To cope with multi-generation and multi-vendor xPU mixes, the architecture needs to support diverse hardware requirements with unified way:

- Localized vUPF: To facilitate low-latency AI services, a localized vUPF is deployed. It is responsible for processing and delivering AI data traffic directly within the vRAN server, minimizing transport latency.
- Optimized xPUs: The vRAN architecture integrates optimized processing units—comprising CPUs, NPUs, and GPUs—to support demanding AI computing tasks effectively.
- Unified Configuration Management: Orchestrators perform comprehensive management of server configurations including software, middleware and/or device drivers. This capability is essential for addressing the complexities arising from multi-generation and multi-vendor environments and varying server deployment timelines, ensuring consistent operation across a diverse hardware infrastructure.

3.3.4 Orchestration of RAN and AI workloads

With the architectural roles defined in 3.3.3, this section specifies orchestration criteria, decision loops, and in-server allocation mechanisms that allow AI jobs to opportunistically use vRAN compute while preserving strict RAN SLAs.

Orchestration Criteria

The main criteria of orchestration may include:

- SLA for both workloads to ensure service quality
- Hardware resource usage (e.g., base station-deployed individual server resources) for workload deployment decision
- RAN traffic prediction and AI traffic prediction to enable fast orchestration decisions
- Operator policy

Especially, since guaranteeing the QoS of RAN traffic in mobile networks is very important, analyzing the expected RAN traffic volume by time and by cell and utilizing these prediction values is very meaningful.

Resource allocation within a vRAN server

Resource allocation technology should be considered for various xPUs. In this subclause, resource allocation for GPU is described. Dynamic GPU allocation allows multiple workloads to share one physical GPU by partitioning computing resources in time and space, with adjustable allocation during execution. It can be implemented by logical GPU separation, priority control at kernel or processing level, or limiting memory/compute to reduce interference.

Design requires understanding partition/reclaim units, whether reallocation disrupts workloads, and setting adjustment period and scale based on the requirement of latency sensitivity and performance tolerance.

3.3.5 Benefits

In this section, based on RAN traffic prediction, we will examine the feasibility of xPU (e.g., GPU) resource allocation between RAN WL and AI WL, and analyze the benefits of this feature.

Traffic Pattern of Active Base Station

[Figure 6] below shows the analysis of RAN traffic by time of day. As expected, traffic stays relatively high during the daytime and drops significantly at night. This demonstrates the possibility of utilizing the remaining resources of RAN servers for AI WL through RAN traffic prediction.

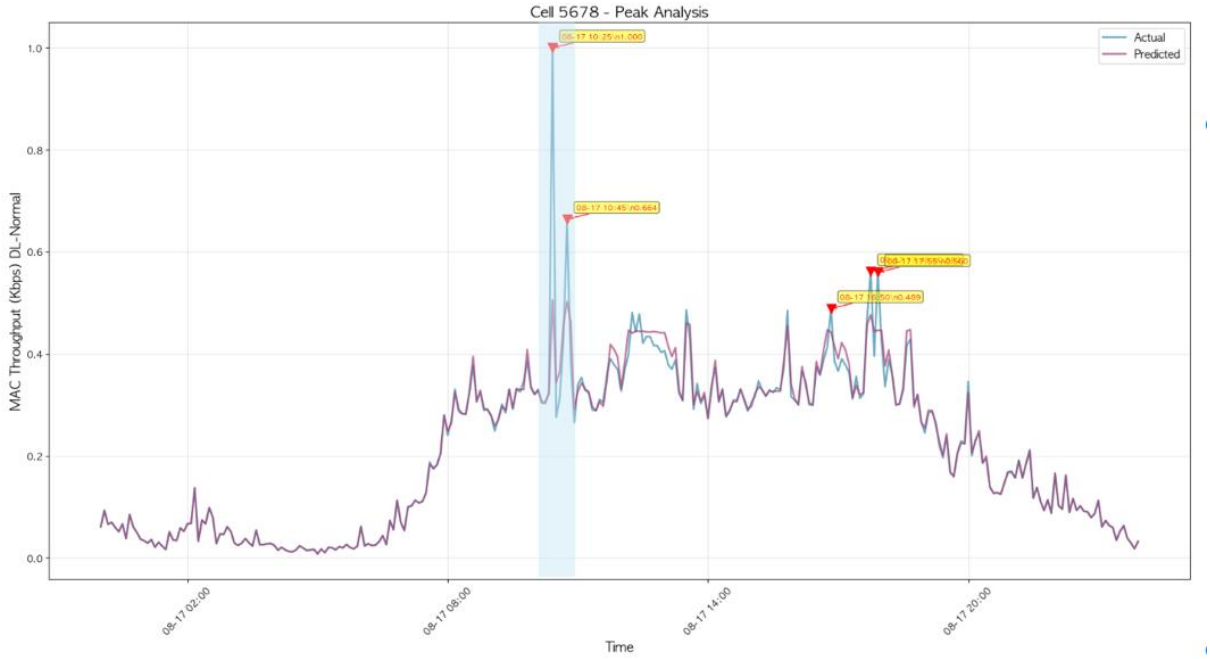


Figure 6. Traffic Pattern of Active Base Station

Reactive AI WL Orchestration

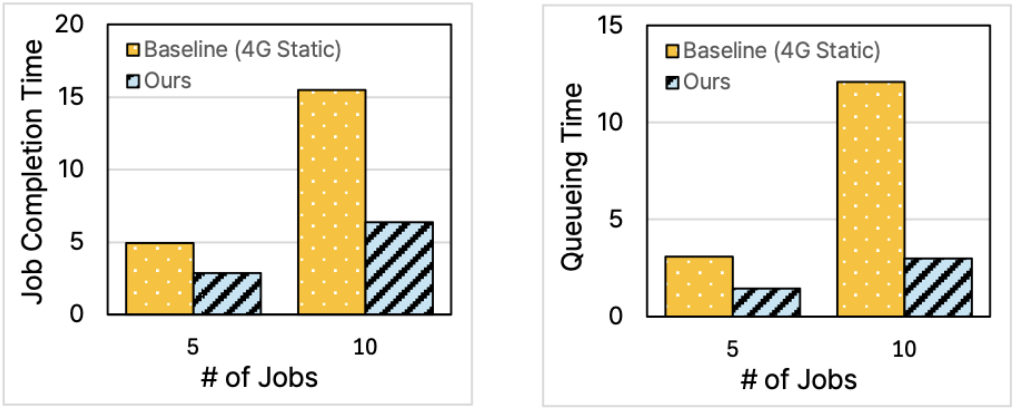
※ The tests were conducted using NVIDIA GPUs.

[Figure 7] shows the result of reconfiguring the initially allocated GPU MIG instances for RAN workload and AI workload according to orchestration criteria such as RAN traffic prediction in a GH200 server environment.

Figure 7. Test result for GPU reconfiguration

When reconfiguring MIG instances, the existing MIG instances are removed and new MIG instances are created, and during this process, GPU usage is typically interrupted. This interruption can potentially affect the service quality of the RAN workload, so it requires careful consideration.

As shown in [Figure 8] below, the goal of orchestration is to minimize the job completion time for both AI and RAN workloads deployed in the cluster, thereby improving resource utilization efficiency. For the AI workload, Large Language Model inference is selected as the representative use case. Upon the arrival of a new job, the orchestrator re-evaluates all active workloads—N total jobs, including the new arrival—and determines a placement plan that minimizes the overall makespan (the maximum job duration across all jobs). This method ensures optimal job scheduling while dynamically adapting to workload changes, delivering balanced performance between RAN traffic-handling and AI inference tasks.



- Job Completion Time: The total time from when a job is submitted until it is completed
- Queuing Time: The time spent waiting during the process of job completion.

Figure 8. Job Completion Time and Queuing Time

The test result conceptually validates the feasibility of supporting heterogeneous workloads on a single infrastructure using base station resources, while also indicating the potential for new service models and future revenue opportunities for mobile operators. Technically, the results demonstrate that RAN and AI workloads can coexist under limited infrastructure resources through appropriate orchestration. Going forward, further refinement of orchestration criteria—such as workload-specific SLA definitions—and the development of more accurate traffic prediction models will be required, along with extensions to heterogeneous accelerator environments beyond GPUs to enable flexible resource placement and reallocation.

4 Conclusion

vRAN has reached a pivotal turning point where it needs to demonstrate its economic efficiency and differentiated value in commercial networks. The enabling features analyzed in this paper represent the essential requirements for vRAN to surpass the performance and flexibility of traditional, purpose-built RAN architectures and to evolve toward AI-RAN.

The primary conclusions and recommendations for the ecosystem are as follows:

- **Realizing HW/SW Strict Separation as the Foundation:** Achieving a truly open environment requires the complete elimination of dependencies between RAN software and specific hardware or virtualization platforms. HW/SW strict separation is the fundamental pillar that ensures the long-term flexibility and openness promised by virtualization. Operators should be able to maintain their hardware investments while evolving the network—such as upgrading from 5G to 6G—through software updates alone.
- **Implementing Resource Pooling for TCO Efficiency:** Resource pooling is a key enabler for optimizing network operations, allowing MNOs to efficiently handle fluctuating traffic demands with fewer resources while maintaining service quality. This capability directly contributes to lowering both CAPEX and OPEX by improving infrastructure utilization. Vendors are strongly encouraged to prioritize the delivery of these flexible pooling capabilities in their development roadmaps.
- **Evolution toward AI-RAN Platforms:** Integrating AI computing into vRAN creates new business opportunities by transforming the network into an integrated AI platform. By sharing high-performance hardware between RAN and AI workloads through intelligent orchestration, operators can maximize infrastructure utilization while meeting strict SLA requirements.

The success of vRAN depends on the ecosystem's ability to move beyond basic stability and deliver these high-value, differentiating capabilities. Equipment vendors should align their product roadmaps with these essential features, while technology partners and standardization bodies are expected to accelerate the creation of open interfaces and interoperable systems. By collectively prioritizing the delivery and adoption of these features, the industry can ensure that vRAN becomes the core, future-proof infrastructure for highly efficient and intelligent 5G-Advanced and 6G networks.

References

[1] SK Telecom & NTT DOCOMO, "Key Considerations for vRAN: Insights from SK Telecom and NTT DOCOMO," White Paper, February 2024.