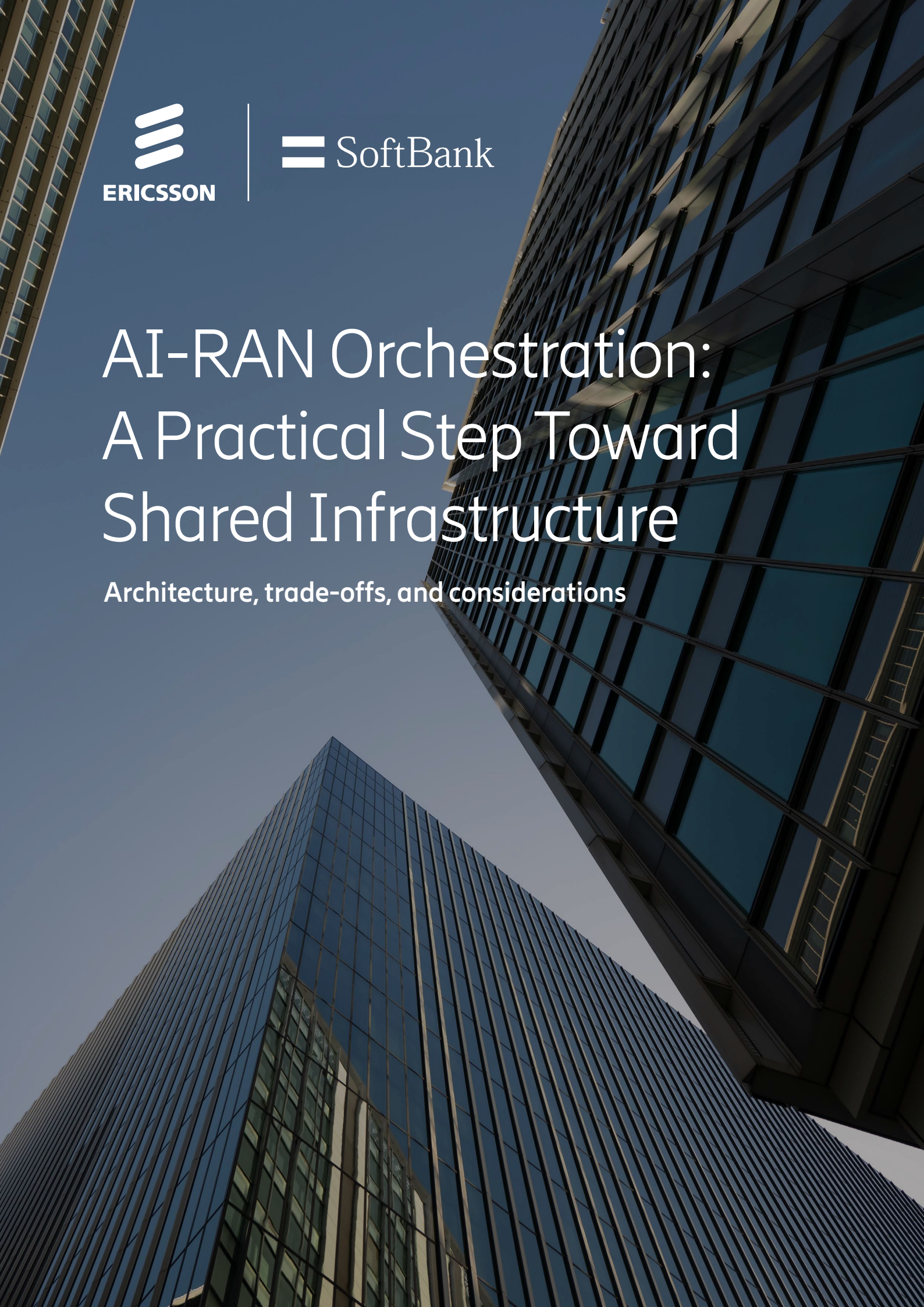


ERICSSON | SoftBank

AI-RAN Orchestration: A Practical Step Toward Shared Infrastructure

Architecture, trade-offs, and considerations

Introduction

AI-driven applications, ranging from immersive experiences to real-time decision systems place new demands on latency, uplink capacity, and orchestration across radio, transport, and compute. At the same time, the automation required to operate future networks, including functions such as intelligent network optimization, dynamic slicing, energy efficiency and lifecycle management, increasingly relies on AI techniques that benefit from tighter integration between network control and compute resources. These dual dynamics are reshaping the role of RAN as both an intelligent network and a critical substrate for AI execution.

This paper examines how RAN and AI can coexist within a shared architectural framework, and what capabilities are required on the network side to support emerging AI-driven use cases. The discussion focuses on architectural considerations for coexistence between RAN functions and AI workloads, including implications for automation, orchestration, and lifecycle management. Given that rApps provide a simple and flexible means to deploy and independently manage RAN automation functions, the paper assumes that the AI RAN workload considered here takes the form of an rApp [1] rather than a RAN baseband feature or other real-time RAN workload.

Within this context, the paper uses a concrete orchestration use case jointly developed by SoftBank and Ericsson. The use case demonstrates the potential of AI-RAN Orchestration through the interaction between SoftBank's AITRAS Orchestrator and Ericsson Intelligent Automation Platform (EIAP), which dynamically coordinate resource allocation in response to the computational demands of AI workloads, including rApps. Through this integration, the two platforms realize an AI-and-RAN use case in which AI workloads and RAN infrastructure operate in a coordinated manner. The use case also highlights how such coordination can further enable AI-on-RAN workloads as well as additional automation and optimization capabilities within the network itself (AI-for-RAN).

AI-and-RAN Orchestration

The concept of AI-and-RAN involves using a shared information technology infrastructure to host both AI workloads and RAN workloads, creating opportunities to improve infrastructure utilization and flexibility. Consumer/enterprise AI workloads typically span distinct execution phases. Training, retraining, and life-cycle management are generally centralized and optimized for high throughput and accelerator efficiency, while inference workloads range from batch processing to interactive services that are latency-sensitive and operate closer to the edge.

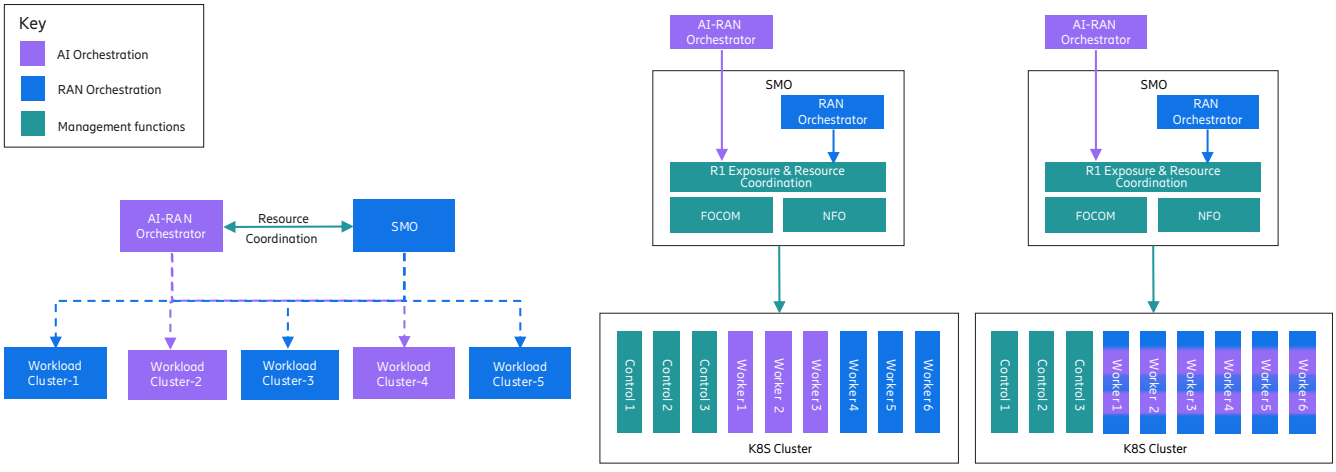
RAN workloads, by contrast, encompass functions with a broad range of timing and reliability requirements, from time-constrained processing in the real-time domain to more elastic functions operating outside the baseband processing path. While some RAN functions require tight latency bounds, others can tolerate greater variability and are naturally suited to execution in non-real-time environments.

Within this context, rApps represent a class of RAN workloads whose execution characteristics, particularly those associated with their machine learning models, are closer to those of AI workloads than to time-critical RAN processing. The compute requirements of these models are often more centralized and episodic, increasing during training or retraining phases and subsiding during steady-state operation. This makes rApps a concrete example for examining coexistence between AI workloads and RAN functions on shared infrastructure, which is explored in the following chapter.

Resource management scenarios and architectural implications

Given the coexistence of AI workloads and RAN workloads with different execution characteristics, resource management can be realized at different levels of sharing within the compute infrastructure. These levels reflect increasing degrees of coupling between workloads and corresponding trade-offs between isolation, flexibility, and operational complexity.

Figure 1. AI RAN Sharing Scenarios



Scenario 1: per-cluster allocation with east-west coordination

Scenarios 2 and 3: allocation of individual cluster resources (complete worker nodes or per-CPU core/GPU segments) with common coordination

Three representative scenarios can be considered. In Scenario 1, multiple clusters are deployed, with each cluster allocated to a specific workload domain, such as RAN or AI. In Scenario 2, a single cluster is shared across workload domains, while workloads remain isolated at a higher abstraction level. In Scenario 3, workloads share resources within the same server, enabling the finest level of sharing.

These scenarios form a progression from strong separation toward tighter integration. As sharing becomes more fine-grained, potential gains in utilization increase, but so do the demands on coordination, performance assurance, and fault isolation. The table below illustrates these trade-offs.

Dimension	Scenario 1: Multiple clusters	Scenario 2: Shared cluster	Scenario 3: Shared server
Level of resource sharing	None across clusters	Cluster-level	Server-level
Workload isolation	Very high	High	Limited
Operational complexity	Low	Medium	High
Performance predictability	High	Medium	Lower
Risk of cross-workload interference	Minimal	Limited	Higher
Fault isolation	Clear	Mostly localized	More complex
Resource Efficiency	Low	Medium	High
Alignment with current deployments	Strong	Partial	Limited
Business risk	Low	Moderate	High

Alignment with current deployment practices

Deploying multiple clusters dedicated to specific workload domains (Scenario 1) is closely aligned with how RAN and adjacent compute resources are deployed in current commercial networks. This approach provides clear isolation boundaries, predictable performance behavior, and straightforward fault containment, all of which are critical for maintaining network reliability. While this model does not maximize infrastructure sharing, it avoids introducing additional complexity in the absence of a clear business case that would justify tighter coupling between heterogeneous workloads.

From an operational perspective, this separation also allows each workload domain to evolve independently, using orchestration mechanisms and lifecycle processes tailored to its specific requirements, while still enabling infrastructure-level coordination where needed.

Considerations for shared clusters and shared servers

Sharing a single cluster or a single server across AI and RAN workloads (Scenarios 2 and 3) increases flexibility and the potential for improved utilization, but it also introduces tighter coupling between workload domains. In these scenarios, resource contention, fault propagation, and performance interactions become more difficult to reason about, particularly when workloads have different timing sensitivities and scaling behaviors.

Without a strong and explicit business driver, this added complexity may outweigh the benefits of finer-grained sharing, especially if it risks degrading RAN performance or increasing operational overhead. As a result, shared-cluster and shared-server models are better viewed as evolutionary options that may become more attractive as coordination mechanisms mature and as clearer value propositions emerge.

Orchestration implications across scenarios

In Scenario 1, AI software and RAN software can be orchestrated independently, since each workload domain is allocated its own cluster and associated orchestration functions. Infrastructure sharing, where required, can be realized through coordination mechanisms that allow orchestrators to request and release resources across clusters as demand fluctuates in each domain. This preserves clear ownership boundaries while enabling controlled sharing at the infrastructure level. This scenario is examined in more detail in the subsequent chapters.

In Scenarios 2 and 3, tighter coordination is required, as workloads are placed within a shared cluster or even on shared servers. In these cases, independent orchestration entities must coordinate their resource requests to avoid conflicts and unintended interactions. A key challenge is avoiding a "multi-master" situation, where multiple orchestrators independently assume control over the same resources.

One possible architectural direction is to separate domain-specific orchestration logic for AI and RAN workloads from the generic functionality responsible for coordinating resource allocation within Kubernetes clusters. This allows workload-specific requirements to be expressed independently, while relying on a common coordination point for cluster-level resource management. Such an approach is conceptually aligned with architectures such as the O-RAN Alliance Service Management and Orchestration framework, which distinguishes between workload orchestration functions and mechanisms used to coordinate changes toward the underlying infrastructure through standardized interfaces.

AI-and-RAN – demonstrating sharing edge resources for AI-RAN rApps and non-RAN workloads

In line with current deployment practices, an initial realization of AI-and-RAN can be explored through loosely coupled sharing of compute resources across workload domains. Rather than assuming fine-grained co-location of heterogeneous workloads, a more pragmatic approach is to enable controlled and temporary access to AI-optimized compute resources across domain boundaries, while preserving clear ownership and operational separation.

One representative example is the ability to allocate cloud-based compute resources with GPU capabilities, typically associated with an AI domain, to support RAN-related AI workloads during periods of lower utilization, such as off-peak hours. This allows AI model training or retraining workloads to be executed without permanently reserving resources within the RAN domain, and without introducing tight coupling between AI and RAN runtime environments.

Ericsson and SoftBank have collaborated to demonstrate such a use case, combining AI-for-RAN and AI-on-RAN environments to illustrate an AI-and-RAN scenario aligned with this deployment model. In the SoftBank laboratory environment, multiple single-node clusters are managed under an AI-RAN orchestrator, while RAN workloads, simulated for this purpose, are executed in an Ericsson laboratory under the control of an SMO, specifically Ericsson Intelligent Automation Platform. In this setup, an rApp acts as the RAN workload, performing optimization-related functions and serving as a concrete example of how AI-related RAN workloads can leverage shared compute resources without disrupting established RAN operational principles.

During normal operation, the rApp executes its AI model on a generic CPU-based Kubernetes cluster in the SMO environment in the Ericsson lab. The AI-RAN orchestrator in the SoftBank lab, meanwhile, is independently managing AI workloads for enterprise users on the SoftBank GPU-based infrastructure. The AI-RAN Orchestrator has formally registered an R1 extension API with EIAP to facilitate requests for additional resources.

At a certain point, the rApp running in the EIAP environment detects that its model is degrading in quality and requires retraining on new data. The rApp requests, via the AI-RAN orchestrator API registered via R1, access to a Kubernetes cluster with GPU support to run a training session with fresh network data.

The AI-RAN orchestrator evaluates its available cluster inventory and determines that it has an available cluster which is not currently in use by any AI application. It marks this cluster as reserved and passes the details to the rApp requiring training. The AI-RAN orchestrator can use various means to determine whether to make a cluster available, including time-based policies, predictions on future load, or other operational considerations.

Figure 2. AI-RAN Lab Demonstration Setup

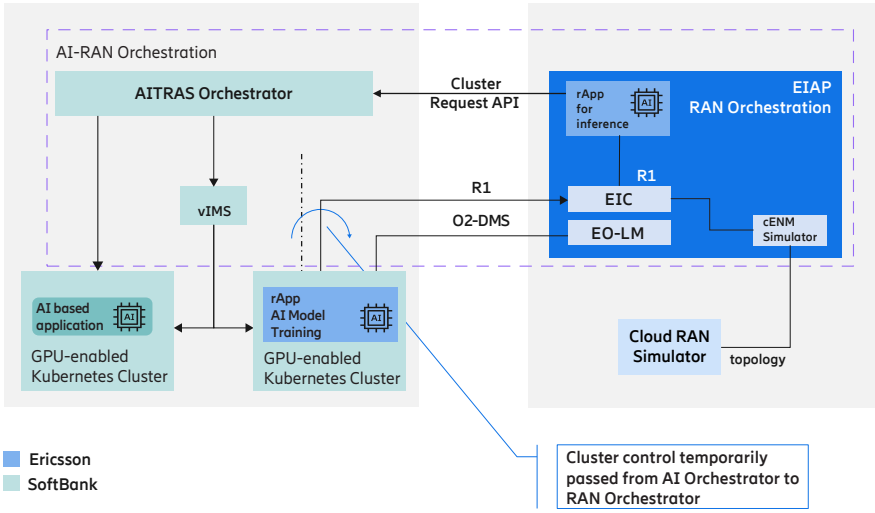
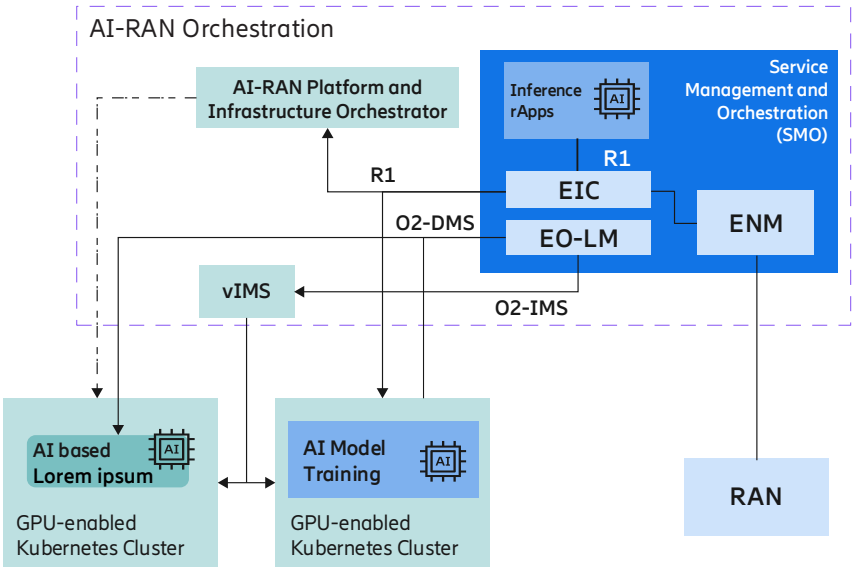


Figure 3. Future architecture, plan for development in commercial network



The rApp introduces this new cluster as a workload cluster into the SMO by registering it over R1 with the Federated O-Cloud Orchestration and Management (FOCOM) APIs. Subsequently, the rApp can deploy a model training workload over R1 via the Network Function Orchestration (NFO) APIs. Secure access to the required training dataset is provided to the workload through the Data Management and Exposure (DME) APIs over R1.

The model training workload can leverage the clusters' GPU to quickly iterate over a number of training epochs, improving model accuracy each time. At some point, either the model is adequately trained, or the AI-RAN orchestrator requests control to be returned for the cluster. Once the model reaches the desired performance level, or when the AI-RAN orchestrator requests the return of the cluster, the workload is terminated, and the updated model is made available to the rApp via the R1 DME APIs.

This use case is a practical first step towards demonstrating:

- Dynamic allocation of compute resources between RAN-related workloads and other edge AI applications, enabling more efficient hardware utilization and new revenue opportunities
- The ability for rApps AI models and non-RAN AI workloads to share edge infrastructure in a controlled manner
- How SoftBank's AI-RAN Platform and Infrastructure Orchestrator integrated with Ericsson's SMO can leverage R1 interfaces within the O-RAN architecture to coexist and collaborate with other workload orchestrators through loose coupling
- The extensibility of the O-RAN architecture, including the ability to expose NFO and FOCOM interfaces over R1 to support coordinated orchestration on shared infrastructure

Note that the architecture discussed and explored in this whitepaper is aligned to the reference architecture of AI-RAN orchestration defined in the AI-RAN Alliance [2], and that SoftBank's AITRAS Orchestrator corresponds to the AI-RAN Platform and Infrastructure Orchestrator defined in [3].

Future considerations: business and technology perspectives

As AI workloads and RAN workloads increasingly intersect at the infrastructure level, several considerations remain open, spanning both business and technology dimensions.

From a technology perspective, further investigation is needed to understand how heterogeneous workloads with different execution characteristics can coexist while preserving predictable performance and fault isolation.
This includes examining how resource abstraction, workload characterization, and interface evolution can support more flexible deployment options without introducing unnecessary coupling or operational risk. These considerations will shape how AI-and-RAN architectures evolve over time, informed by both operational experience and business rationale.

Key considerations emerging from this work include:

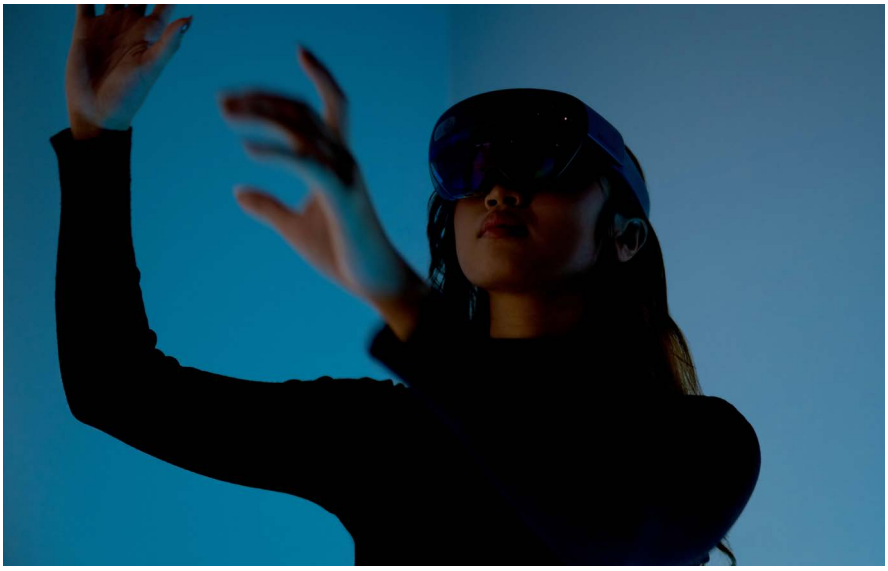
Business-oriented considerations

- Identifying concrete value drivers for infrastructure sharing beyond utilization gains, such as time-to-market or service differentiation
- Ensuring that added architectural complexity is aligned with a clear and measurable business and economic rationale
- Maintaining clear separation of responsibility and accountability across workload domains

Technology-oriented considerations

- Preserving performance predictability and fault isolation as a first-order design constraint
- Characterizing workloads in a way that supports informed placement and resource allocation decisions
- Allowing architectural evolution with progressively increasing coupling based on proof points and business needs

Together, these business and technology considerations underline the importance of aligning architectural choices with both operational realities and demonstrable value. As AI RAN architectures evolve, maintaining this balance will be key to enabling coexistence models that are sustainable, scalable, and appropriate for carrier-grade deployments.



Reference

1. An rApp is a RAN application deployed on top of the Ericsson Intelligent Controller (EIC), our Non-Real-Time RIC, leveraging data, models, and control interfaces to optimize and automate RAN behavior outside of time-critical processing.
2. AI-RAN Alliance, "AI-RAN Architecture v1.2," Feb. 2026. Available: https://ai-ran.org/documents/AI-RAN-Architecture-v1.2.pdf
3. AI-RAN Alliance WG2 (AI-and-RAN), "AI-RAN Platform and Infrastructure Orchestrator," Feb. 2026. Available: https://ai-ran.org/documents/AI-RAN-Platform-Infrastructure-Orchestrator.pdf

About Ericsson

Ericsson's high-performing networks provide connectivity for billions of people every day. For nearly 150 years, we've been pioneers in creating technology for communication. We offer mobile communication and connectivity solutions for service providers and enterprises. Together with our customers and partners, we make the digital world of tomorrow a reality.

www.ericsson.com

About SoftBank Corp.

Guided by the SoftBank Group's corporate philosophy, "Information Revolution — Happiness for everyone," SoftBank Corp. (TOKYO: 9434) operates telecommunications and IT businesses in Japan and globally. Building on its strong business foundation, SoftBank Corp. is aiming to activate the potential of AI across its businesses and drive implementation in line with its "Activate AI for Society" growth strategy. While further growing its telecom business, SoftBank is expanding its AI computing infrastructure and AI and Cloud service businesses with the aim of becoming a provider of Next-generation Social Infrastructure. To learn more, please visit.

www.softbank.jp/en/corp/