Operators can choose one of three implementation strategies for deploying AI-native RAN

October 2025

Adaora Okeleke and Caroline Gabriel

AI-native radio access network (RAN) architecture offers a route for operators to take advantage of the immense potential promised by AI. The technology can improve the economics associated with the RAN while expanding its operational capabilities to provide a much-needed boost to a sector that has struggled to achieve revenue growth. Deployment of the full AI-native RAN is not an easy road, and is likely to prove extremely challenging for most operators. Not only are the upfront costs a substantial obstacle for many operators, but there is a risk of major disruption brought about by introducing AI functions into the same compute environment as RAN functions.

Analysys Mason's report *AI-native RAN: implementation strategies* identifies three AI-native RAN implementation strategies that provide operators with the flexibility to start their AI-native RAN journey at a pace that suits their business goals and smoothes the migration path from existing RAN architectures to AI-native RAN. This article provides a summary of the implementation strategies and their implications for operators and RAN vendors.

AI-native RAN promises several benefits but has challenges that could slow down adoption

Operators are exploring AI-native RAN architecture in the hope that it may transform current RAN economics and increase operational and resource efficiency. However, realising these benefits and aspirations will require that operators understand what changes need to occur in existing RAN environments to allow migration to the new AI-native architecture, and the impact of these changes on existing RAN deployments.

A key change needed for the migration is the inclusion of graphical processing units (GPUs) – which are the most common processors for running advanced AI such as large language models (LLMs) – in existing RAN equipment, potentially as far out as some cell sites. Analysys Mason's recent survey of 67 operators around the globe about their future RAN architecture plans indicates several concerns regarding this change.[1] These concerns include the high upfront costs of acquiring or accessing the GPUs, the significant power consumption of these AI processors, and the ongoing costs and performance trade-offs that may come with managing both AI and RAN workloads within the same environment.

Operators' concerns could slow down the pace of transition towards the AI-native RAN architecture and prevent them from enjoying its potential benefits. Nonetheless, operators must move quickly to define a roadmap to AI-native RAN because future 6G networks are likely to be based on this architecture. Operators, therefore, need to identify strategies that they can adopt to transition towards AI-native RAN architecture while minimising risks.

[1] For more information, see Analysys Mason's report *Operators' requirements for their next-generation RANs: survey results and analysis*.

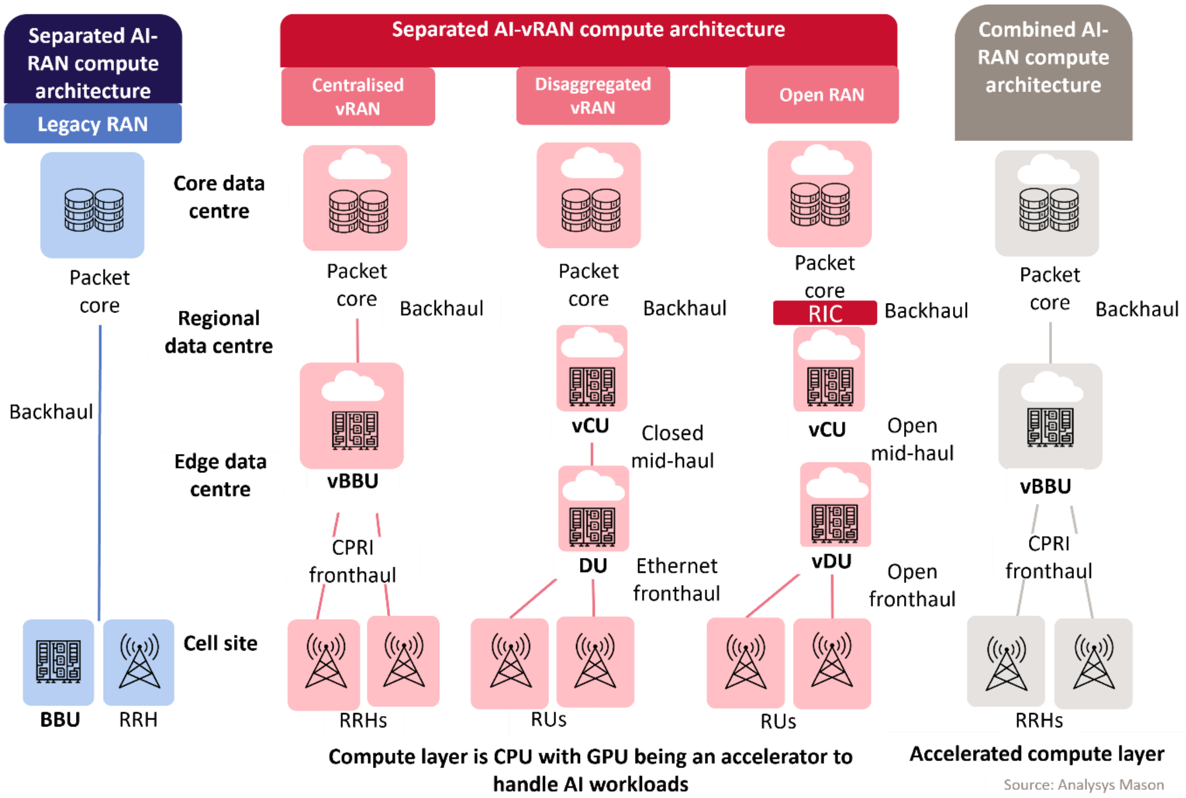
Analysys Mason's research has identified three implementation models for the realisation of AI-native RAN

The ultimate goal when planning AI-native RAN is a cloud-native, software-defined virtualised RAN (vRAN) running on a compute architecture that supports both RAN and AI acceleration. Therefore, the starting point for any AI-native RAN architecture needs to be a vRAN. However, most operators have not commercially deployed vRAN, and these operators need to consider how far they can achieve some AI-RAN benefits while they still have a conventional RAN in which the baseband unit is a specialised appliance. Even some operators that have deployed vRAN will be unwilling to replace this equipment with AI-native systems so early in the asset lifecycle.

We have identified three AI-native RAN deployment models that operators are most likely to adopt, depending on their current RAN (Figure 1). Each of these models has a different compute architecture that would support AI processing. These models are:

- the separated AI-RAN compute architecture
- the separated AI-vRAN compute architecture
- the combined AI-RAN compute architecture

Figure 1: Analysys Mason's three AI-native RAN architectures[2]



Source: Analysys Mason

[2] CPU = central processing unit; BBU = baseband unit; RRH = remote radio head; vBBU = virtual baseband unit; vCU = virtualised centralised unit; DU = distributed unit; RU = radio unit; vDU = virtualised distributed unit

The separated AI-RAN applies AI-RAN control to an existing traditional RAN with both AI and RAN applications running on separate compute platforms. This first step will ease operators' transition to the AI-native RAN because it does not disrupt existing RAN operations and requires lower upfront cost than the other two strategies. However, lower upfront cost is a short-term advantage because running two architectures will involve greater operating costs, complexity and integration challenges.

Operators with vRAN deployments can choose from one of two options: the separated AI-vRAN and the combined AI-RAN compute architectures. The decision will depend on the maturity of the operator's vRAN and their willingness to increase capital expenditure.

The separated AI-vRAN allows coexistence between an existing vRAN and an AI-centric RAN. The RAN network functions will run in the existing compute environment, while the AI applications will run in a separate compute environment, with dedicated AI accelerators. Key benefits of this strategy are a well-planned transition to the AI-native RAN without wasting recent vRAN investments, and the flexibility to map RAN and AI processing resources to existing network topology. However, this strategy comes with a high-cost risk, given the overheads of running and harmonising two separate platforms.

The combined AI-RAN will be completely software-driven and cloud-native. The RAN and AI applications will run on the same compute platform, with acceleration that combines the sophisticated processing capabilities of the GPU with the CPU and other relevant chipsets required to run both AI and RAN workloads. This architecture gives operators an opportunity to optimise their own RAN performance, prepare for 6G and play their role in meeting the increasing consumer and enterprise demand for AI capabilities. However, the cost of adopting the combined AI-RAN approach is higher than that of the other strategies in the short term, and where AI processing is relatively centralised, there will be a need for high-quality fibre connectivity to connect cell sites to the AI-native central basebands.

The three AI-native RAN implementation options offer operators flexibility but place huge demands on the RAN ecosystem

Having multiple AI-RAN implementation options allows operators to develop a transition roadmap to a fully AI-native RAN environment that aligns with their networks and goals.

However, the flexibility that operators derive from having these implementation options places a huge demand on the RAN ecosystem. RAN vendors, system integrators and the rest of the RAN ecosystem must be prepared to support any one of the three strategies, given operators' varying levels of RAN maturity. These players should be as flexible as possible in offering operator customers and prospects multiple blueprints for AI-native RAN that align with operators' network and business priorities. They should also be prepared to support operators' transition from one implementation strategy to another over time.

This article draws on the third report in a series of Analysys Mason publications on the topic of AI-native RAN. Analysys Mason brings extensive experience in the areas of AI and wireless technologies and infrastructure, gained through research and customer projects. This expertise positions us to support stakeholders in understanding how to capture the opportunities that AI-native RAN presents. For further information, get in touch with Adaora Okeleke or Caroline Gabriel.