O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Research Report on Scalable and User-Centric RAN Architecture: Service Requirements and Design Considerations

Report ID: RR-2025-03

Contributors:

KDDI
LGU+
NVIDIA
University of York

Release date: 2026.01

Authors

Amr Amrallah, KDDI (Editor-in-Chief)

Akio Ikami, KDDI

Riichiro Nagareda, KDDI

Hyosun Yang, LGU+

Lopamudra Kundu, NVIDIA

Alister Burr, University of York

Reviewers

Salvatore D’Oro, Northeastern University

Hiroshi Miyata, Sumitomo Electric Industries

Vikas Dixit, Reliance Jio

Jan Plachy, DTAG

Nurit Sprecher, Nokia

Michael Garyantes, QCM

Disclaimer

The content of this document reflects the view of the authors listed above. It does not reflect the views of the O-RAN ALLIANCE as a community. The materials and information included in this document have been prepared or assembled by the above-mentioned authors, and are intended for informational purposes only. The above-mentioned authors shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of this document subject to any liability which is mandatory due to applicable law. The information in this document is provided ‘as is,' and no guarantee or warranty is given that the information is fit for any particular purpose.

Copyright

Executive summary

As the telecommunications industry progresses into the Sixth Generation (6G) era, Radio Access Networks (RAN) are expected to undergo transformative changes in design and management driven by unprecedented service diversity, higher performance targets, and the need for more flexible, intelligent, and energy-efficiency architectures. Unlike Fifth Generation (5G) and 5G-Advanced (5GA) mobile networks, which primarily addressed enhanced mobile broadband, massive IoT, and ultra-reliable low-latency communication within largely static deployments, 6G is anticipated to support pervasive connectivity across heterogeneous devices, immersive services such as extended reality, large-scale digital twins, and ubiquitous AI-driven automation. These new services will generate highly dynamic, location-dependent traffic patterns and require networks that can adapt resources on demand with minimal latency.

In this context, scalability is essential to efficiently accommodate a rapidly growing number of devices and services without disproportionate increases in cost or complexity, while user-centricity ensures that the network can tailor performance to individual service requirements, improving quality of experience (QoE) across diverse scenarios. Addressing these needs demands rethinking RAN architecture to allow elastic resource allocation, distributed intelligence, and seamless integration of computation and communication.

This report investigates advancements in scalable and user-centric RAN architectures aimed at meeting those evolving requirements. By incorporating cutting-edge technologies such as Artificial Intelligence / Machine Learning (AI/ML), energy-efficient systems, and adaptive frameworks, this report provides a focused overview of innovations in user-centric RAN, highlighting opportunities to enhance network scalability, flexibility, and energy efficiency. These advancements aim to address current challenges, such as managing highly variable traffic patterns, supporting heterogeneous service demands, and reducing operational costs, that cannot be fully met with existing RAN architectures.

Furthermore, the report highlights key advancements in RAN architecture, focusing on leveraging cloud-native frameworks, virtualization, and distributed resources to enhance the flexibility and efficiency of RAN deployments. Energy efficiency can be achieved through the use of AI/ML algorithms that dynamically consolidate O-RAN Radio Units (O-RUs) onto fewer O-RAN Distributed Units (O-DUs) during low-traffic periods, temporarily deactivating idle O-DUs to minimize power consumption. In current deployments, this approach is constrained by the lack of mature, standardized mechanisms for real-time orchestration across multi-vendor O-RUs and O-DUs, as well as limitations in fronthaul reconfiguration latency and service continuity assurance. Implementing such consolidation requires seamless remapping of O-RU traffic to alternative O-DUs without disrupting ongoing sessions, capabilities that are still under development in O-RAN management and orchestration frameworks. From a scalability perspective, these energy savings reduce infrastructure and operational overhead, allowing the network to accommodate future traffic growth without a proportional increase in active hardware or costs. From a user-centric perspective, conserving energy frees computational and radio resources that can be reallocated to users or

services with higher performance demands, enabling more personalized and context-aware quality-of-experience (QoE) delivery. From the O-RU perspective, consolidation can alter fronthaul routing paths, change synchronization sources, and trigger reconfiguration events, potentially impacting timing accuracy and radio frequency (RF) calibration. Ensuring that these transitions are handled smoothly, without degrading user experience or violating latency targets, is therefore a key design consideration in scalable, user-centric RAN architectures.

Adaptive and resilient systems in this context go beyond conventional automation and intelligent controllers by integrating real-time network monitoring with AI/ML driven predictive fault detection and rapid recovery mechanisms that operate across distributed and multi-vendor O-RAN environments. From a scalability perspective, these capabilities enable the network to maintain service continuity and performance as the number of connected devices and services grows, without requiring proportional increases in manual intervention or centralized resources. From a user-centric perspective, they enable proactive detection and resolution of issues that could degrade individual user experiences, dynamically reallocating resources or adjusting service parameters to preserve QoE in real time.

Additionally, the report addresses operational and design challenges, focusing on complexities like system integration, latency, and scalability, which are tackled through standardized interfaces and advanced orchestration tools.

This scalable, user-centric RAN framework lays a strong foundation for future telecommunications. By integrating AI/ML, energy-efficient technologies, and adaptive designs, the framework addresses the complex demands of 6G networks, such as ultra-low latency for mission-critical applications, extreme reliability for autonomous and industrial use cases, massive connectivity for large-scale IoT ecosystems, high spectral efficiency for dense urban environments, and dynamic adaptability to diverse service contexts and user requirements. From a user-centric perspective, it enables per-user and context-aware optimization of computing, radio, and fronthaul resources, ensuring tailored performance and consistent QoE across varied scenarios. In doing so, this framework not only meets emerging technical requirements but also unlocks new opportunities for innovation and sustainability in next-generation RAN deployments.

Table of Contents

List of abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
5GA	5G-Advanced
6G	Sixth Generation
AF	Application Function
AGV	Automated Guided Vehicles
AI	Artificial Intelligence
AIaaS	AI-as-a-Service
AMCOP	Aarna Multi Cluster Orchestration Platform
API	Application Programming Interface
App	Third-party application
AR	Augmented Reality
ASIC	Application-Specific Integrated Circuit
BBU	Baseband Unit
CAPEX	Capital Expenditure
CCIN	Communication and Computing Integrated Networking
CD	Continuous Delivery/Deployment
CF-mMIMO	Cell-Free massive MIMO
CI	Continuous Integration
CNF	Cloud-Native Network Function
CoMP	Coordinated Multipoint
COTS	Commercial-Off-The-Shelf
CPU	Central Processing Unit
C-RAN	Centralized RAN
D-MIMO	Distributed MIMO
DDC	Digital Down Conversion
DMS	Deployment Management Services
DPDK	Data Plane Development Kit
DPU	Data Processing Unit
D-RAN	Distributed RAN
DD-MIMO	Decentralized Distributed MIMO
DSP	Digital Signal Processing
DUC	Digital Up Conversion
eCPRI	enhanced Common Public Radio Interface
FFT	Fast Fourier Transform
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
IMS	Infrastructure Management Services
IoT	Internet of Things
ISAC	Integrated Sensing and Communication
LCM	Lifecycle Management
LLM	Large Language Model

MIG	Multi-Instance GPU
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MNO	Mobile Network Operator
multi-TRP	multi-Transmission and Reception Point
NDT	Network Digital Twin
NF	Network Function
NIC	Network Interface Card
Near-RT RIC	Near-Real-Time RIC
Non-RT RIC	Non-Real-Time RIC
O-CU	O-RAN Central Unit
O-DU	O-RAN Distributed Unit
O-RU	O-RAN Radio Unit
OPAE	Open Programmable Acceleration Engine
OPEX	Operational Expenditure
OSC	O-RAN Software Community
PHY	Physical Layer
PRB	Physical Resource Block
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
rApp	Non-RT RIC Application
RF	Radio Frequency
RIC	RAN Intelligent Controller
SDN	Software-Defined Networking
SMO	Service Management and Orchestration
SoC	System-on-Chip
SR-IOV	Single Root I/O Virtualization
UAV	Unmanned Aerial Vehicle
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communication
vCU	virtual Centralized Unit
vDU	virtual Distributed Unit
VR	Virtual Reality
vRAN	virtual RAN
xApp	Near-RT RIC Application
XR	Extended Reality

List of figures

1 Introduction

In the era of the Sixth Generation (6G) mobile networks, Radio Access Networks (RAN) are expected to evolve into cloud-native architectures [1], building on today's partially virtualized deployments where O-RAN Distributed Units (O-DUs) and O-RAN Central Units (O-CUs) may run on Commercial-Off-The-Shelf (COTS) hardware but O-RAN Radio Units (O-RUs) remain largely hardware-specific. While full radio frequency (RF) cloudification of O-RUs is unlikely, their control and baseband-adjacent functions are expected to be increasingly virtualized for more agile configuration and integration. In addition, while similar vertical-specific services were anticipated in Fifth Generation (5G) mobile networks, particularly through network slicing for industrial applications, the anticipated use cases for 6G extend further to encompass not only human communications but also a broader range of AI-driven, real-time interactions between heterogeneous devices, such as robots and Unmanned Aerial Vehicles (UAVs) [2]. This is expected to lead to greater diversity in network service requirements and traffic patterns, driven by the heterogeneous needs of both human users and intelligent machines.

In this report, "user-centric" refers to the ability of the RAN to adapt performance, resource allocation, and energy usage to the specific needs and contexts of individual users or services. This principle underpins scalability, by allowing efficient capacity expansion, and supports energy efficiency by aligning resource activation with actual demand. To address this diversity, the technical report ITU-R M.2516 outlines the concept of user-centric and on-demand deployment within a fully distributed and decentralized architecture [3], where distributed designs are preferred over centralized ones for minimizing latency and enabling localized, per-user optimization. The objective of this research report is to identify and analyze pivotal questions concerning user-centric and scalable RAN architecture, as well as the emerging service requirements that contribute to its realization. By clarifying how these principles interact and why they matter for 6G, the report provides a framework for understanding the limitations of today's approaches and the opportunities for next-generation designs.

Therefore, this research report aims to address the dynamic nature of demand from both human users and non-human endpoints (e.g., devices, robots, and UAVs), both spatially and temporally, which remains insufficiently managed by current RAN architectures, thus motivating approaches that improve the efficiency and sustainability of RAN. Additionally, this report examines deployment options, including centralized and distributed RAN variants and edge- versus central-cloud placements, and explains how these choices enable scalable, user-centric operation.

2 What is scalable and user-centric RAN architecture?

2.1 Motivation

The advent of the 6G mobile network heralds a transformative shift in the RAN towards cloud-native architectures, a change driven by the need to accommodate an increasingly diverse array of use cases. Unlike previous generations, 6G mobile networks are poised to support not only human communications but also complex interactions between a multitude of devices, including robots and UAVs [2]. This evolution necessitates a RAN infrastructure that is both agile and efficient, capable of dynamically responding to the increased spatial and temporal variability in resource consumption.

The anticipated cloud-native deployment of the O-RAN logical nodes, i.e., O-DUs and O-CUs, each defined as an O-RAN Network Function (NF) across distributed cloud platforms underscores the need for a robust and flexible RAN architecture [4]. In this context, an O-DU or O-CU can be implemented as a set of NFs, while the O-RU remains a specialized hardware element handling the RF front-end but potentially integrating more cloud-managed control and signal processing functions in the future. Deployment models may include public or private cloud hosting of these NFs, as well as neutral host infrastructures that enable shared use of physical and virtualized resources by multiple operators.

As the 3rd Generation Partnership Project (3GPP) advances the air interface with increased Multiple Input Multiple Output (MIMO) multiplexes and multi-Transmission and Reception Point (multi-TRP) configurations, the computational demands on radio signal processing are expected to rise significantly due to both the complexity of these features and the dynamic, agile nature of next-generation communications [5]. This escalation necessitates more flexible and scalable processing environments, where virtualization decouples NFs from dedicated hardware and cloudification enables these virtualized NFs to be dynamically deployed and orchestrated across distributed computing resources. In this model, user devices generate diverse and time-varying traffic patterns that trigger dynamic resource adjustments, while O-RUs handle the RF front-end and initial signal processing before forwarding data to virtualized O-DUs and O-CUs hosted in cloud or edge environments. This approach allows the network to scale resources efficiently and respond in real time to changing user and service demands, as advocated by initiatives like Hexa-X and the Next G Alliance [6], [7].

Moreover, as documented in several technical reports and academic studies, the current efforts to enhance RAN resource utilization within fully distributed architectures highlight the critical need for a user-centric approach [3], [8]. In this context, “user-centric” means designing the RAN to allocate and adapt resources based on the specific needs, locations, and service contexts of individual users or devices, rather than relying solely on cell-centric or static allocation models. This enables more precise QoE management, supports highly diverse service requirements, and reduces inefficiencies caused by over-provisioning. The need for such adaptability is driven by the increasing heterogeneity of traffic patterns in 6G, from high-throughput extended reality (XR) applications to low-latency industrial control, where static architectures cannot meet performance targets for all users simultaneously. The O-RAN ALLIANCE focus on improving resource utilization on general-purpose infrastructure by developing relevant features in O-Cloud, including the integration of Communication and Computing Integrated Networking (CCIN) [9]. By enabling flexible allocation of available compute and network capacity, these features make it possible to dynamically host not only RAN functions but also vertical applications close to the end user. Virtualization is the key enabler for this adaptability, as it decouples these functions from dedicated hardware, allowing them to be instantiated, scaled, or relocated on demand in response to changing service requirements.

By realizing a scalable and user-centric RAN on general-purpose computing platforms, there is an opportunity to allocate resources flexible for both traditional RAN functions and emerging workloads. One example is supporting the deployment of third-party Artificial Intelligence / Machine Learning (AI/ML) models on the RAN infrastructure, enabling operators to process data closer to the user and optimize resource utilization. This integration could provide new monetization opportunities through vertical applications and enhanced adaptability. However, such capabilities are not yet widely available due to gaps in standardization for secure onboarding, lifecycle management, and resource isolation of third-party applications, as well as latency and determinism challenges for AI/ML inference in shared environments. While some proprietary solutions exist, broad interoperability and compliance with O-RAN ALLIANCE specifications remain limited. In the end-to-end traffic path, the user device connects via the O-RU, which handles RF transmission/reception and initial signal processing before forwarding data to virtualized O-DUs and O-CUs, potentially co-hosted with AI/ML applications in cloud or edge environments, where resource allocation and service adaptation can occur.

2.2 Concept and structure of scalable and user-centric RAN architecture

The scalable and user-centric RAN architecture represents a paradigm shift in network design, focusing on creating a flexible, efficient, and responsive system. In this context, “user-centric” refers to making the network aware of individual user and their applications, enabling performance and resource allocation to be tailored to specific service requirements in real time. Traditional network models often emphasize maximizing throughput and minimizing costs, generally catering to broad efficiency metrics. However, this new approach emphasizes adaptability and responsiveness, ensuring efficient network performance tailored to individual user needs in real-time [5]. While user-centricity focuses on tailoring services and resource allocation to specific user or application needs, scalability addresses the network’s ability to accommodate growth in devices, services, and traffic volumes without proportional increases in cost or complexity. Both aspects are complementary and should be considered together to deliver sustainable, high-quality performance at scale.

Figure 2-1 illustrates the proposed novel architecture. In response to high user demand, as depicted on the left side, the network scales-out to meet user requirements. This is achieved by deploying multiple instances of the virtual DU (vDU) as a Cloud-Native Network Function (CNF) within the Edge cloud, which manages newly formulated user clusters. The vDUs are supported by hardware accelerators such, as Graphics Processing Units (GPUs) and System-on-Chips (SoCs), which integrate specialized processing components (e.g., digital signa processors (DSPs), AI engines, or Field-Programmable Gate Arrays (FPGA) fabric) to efficiently handle the high computational load of real-time signal processing tasks such as channel coding, MIMO operations, and scheduling. These accelerators enable virtualized implementations to achieve performance levels comparable to dedicated hardware, while also improving scalability and energy efficiency in cloud-based RAN deployments. Additionally, the central cloud adapts by deploying multiple instances of the virtual CU (vCU) as a CNF to accommodate the increased numbers of vDUs. It also hosts some third-party Applications (Apps) that leverage AI/ML techniques to enhance network optimization and improve user experience. Conversely, during periods of low user demand, as shown on the right side, the network scales in by deploying fewer instances of vDU and vCU as CNFs in the Edge and Central clouds, respectively. Moreover, due to the availability of ample computational resources resulting from low user demand, some Apps can operate on the Edge cloud. This scaling-in contributes to power saving, as hardware can be optimized for energy efficiency, repurposed, or fully powered down when demand is minimal.

Not only in operator-owned private clouds but also in shared or neutral-host infrastructures, where underutilized workloads can be migrated to other nodes, enabling some servers to be consolidated or deactivated for improved energy efficiency. Today, such capabilities remain limited due to gaps in technology (e.g., seamless workload migration across distributed nodes, advanced orchestration frameworks, and AI/ML-driven predictive scaling) and in standardization (e.g., missing Application Programming Interfaces (APIs) for energy-aware orchestration and cross-vendor interoperability). In 6G networks, we expect these gaps to be addressed through mature energy-aware orchestration frameworks, standardized interfaces in O-RAN/O-Cloud, and advanced SoC-based platforms that allow fine-grained control of power states and workload placement.

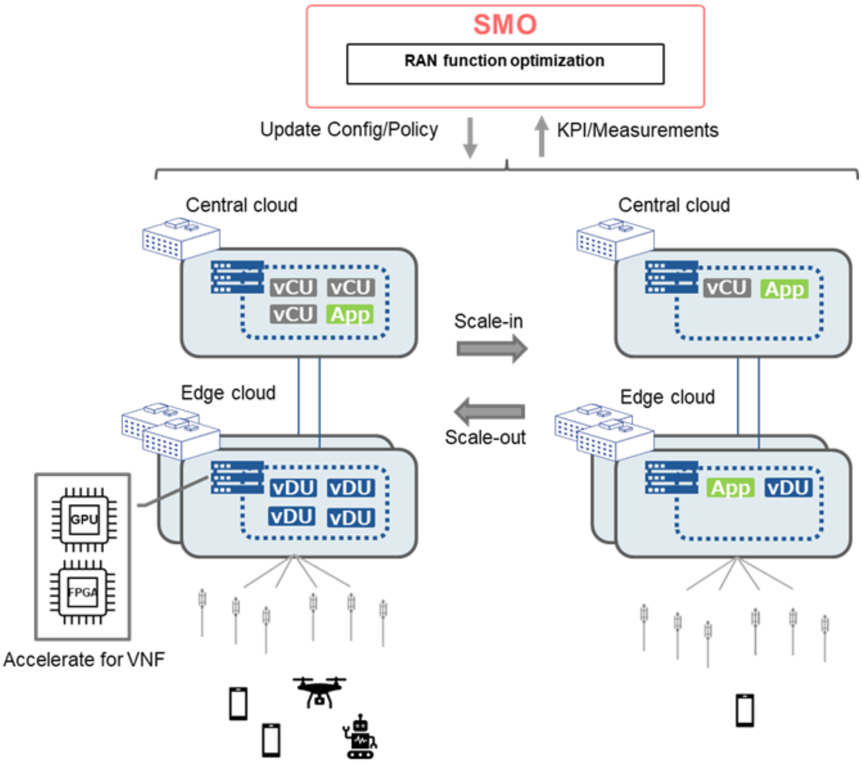

Figure 2-1 The concept of scalable and user-centric RAN architecture.

At the heart of this architecture is the principle of adaptability. The network is crafted to seamlessly scale and adjust according to real-time traffic and user demands. This is achieved by leveraging a modular design, where components are decoupled and can be managed independently. Modularity allows for the rapid deployment of additional resources or the reallocation of existing resources without overall network disruption. The result is a system that remains agile, capable of responding swiftly to changing user patterns and emerging technologies. The architecture emphasizes modulization and flexibility. With network functions being disaggregated, each part of the system can be optimized individually. However, local optimizations must be coordinated through system-level orchestration to avoid fragmented or sub-optimal outcomes. Overall performance is ensured by orchestration frameworks that continuously monitor traffic conditions and user requirements, triggering optimization actions, such as scaling, migration, or resource reallocation, in real time or based on predictive analytics. This balance between local flexibility and global coordination enables both immediate responsiveness and long-term efficiency.

Dynamic resource allocation forms a crucial element in this architecture. By utilizing advanced algorithms and real-time analytics, the network can predict and respond to user demands, ensuring that radio resources (spectrum, physical resource blocks (PRBs), beams, and transmission power), computing resources (central processing unit (CPU)/GPU/SoC capacity at DU/CU nodes), and transport resources (fronthaul/backhaul bandwidth and latency budgets) are distributed where they are needed the most. These resources are managed at different levels: radio scheduling in the DU, compute orchestration in the edge or central cloud, and transport coordination via software-defined networking (SDN). Real-time analytics rely on diverse inputs such as traffic load measurements, user mobility patterns, channel state information, and QoE indicators, which are processed by AI/ML models to forecast demand fluctuations, trigger scaling, and reallocate workloads across nodes. This not only enhances the user experience but also improves overall network efficiency by minimizing wastage and avoiding bottlenecks. While 5G and 5G-Advance (5GA) mobile networks already provide mechanisms such as dynamic scheduling, network slicing, and load balancing, they remain limited in terms of per-user granularity, cross-domain coordination, and standardized APIs for AI-driven orchestration. In 6G, we expect these gaps to be addressed through unified frameworks that integrate radio, compute, and transport domains with predictive analytics, enabling truly user-centric and proactive resource allocation.

The user-centric approach is driven by the need to provide personalized and high-quality service for every user. Unlike 5G network slicing, which primarily provides pre-defined virtual networks tailored to broad service categories, a user-centric RAN dynamically adapts to the unique requirements of individual users, devices, and applications in real time. The architecture is designed to understand and anticipate the unique requirements of various applications and devices. Whether users are engaging in bandwidth-intensive activities like video streaming or low-latency applications like gaming, the network adapts to meet these specific demands. This design also supports a wide variety of use cases, from consumer-level demands to more complex industrial applications, such as Internet of Things (IoT) and smart city solutions. By prioritizing the individual needs of each use case, the architecture ensures reliable and consistent service quality across diverse scenarios.

Furthermore, the architecture for the scalable and user-centric RAN is designed with future adaptability in mind. Its emphasis on flexibility (e.g., being software-defined) modularization, and disaggregation ensures that new infrastructure technologies, such as emerging hardware accelerators, advanced transport solutions, or orchestration platforms, can be integrated through standardized interfaces without redesigning the entire system. As new applications and user behaviors emerge, orchestration frameworks and open APIs allow components to be updated or replaced independently, ensuring that the architecture evolves in step with technological progress. Moreover, this architecture provides a foundation for integrating advanced technologies as they become viable. This approach not only extends the longevity and relevance of the system but also provides a clear path for adopting innovations in both hardware and software, while maintaining consistent performance and user-centric quality of service (QoS). Building on this conceptual foundation, the next step is to consider emerging service paradigms that illustrate how these concepts can be realized in practice.

2.3 Emerging service paradigms for realizing scalable and user-centric RAN

2.3.1 Scalability for energy efficiency

Energy efficiency of cellular systems has been a main topic for all in the industry. O-RAN has helped Mobile Network Operators (MNOs) to reduce costs by decoupling hardware and software, so that the baseband application software can now be deployed on Commercial Off-The-Shelf (COTS) hardware. The traditional Baseband Unit (BBU) stack can be divided into O-DUs and O-CUs and allows the centralization in data centers and localized hubs. These new deployment scenarios can leverage existing data center energy efficiency programs. The distribution of NFs across physical server pools allows for additional potential energy efficiencies. Server pool sizes can be adjusted dynamically with no end-user impact [7], [10].

Many MNOs are deploying 5G mobile network primarily for its higher capacity, so that they can offer higher speeds to consumers and gain a market advantage. But 5G was also developed to support low latency, and the industry has so far been slow to take advantage of it. One challenge with providing low-latency applications has been the need to deploy edge computing economically. Until now, MNOs have had to consider deploying an extra server if they wanted edge computing capabilities, which typically requires a lot of extra Capital Expenditure (CAPEX), power, and space. One area of potential opportunity for them is the investigation of latency-sensitive applications to understand their specific latency and reliability requirements, and then to design RAN solutions that can efficiently meet them. In this context, near-real-time RAN Intelligent Controller (near-RT RIC) Applications (xApps) and AI capabilities serve as optimization mechanisms that dynamically managing spectrum, scheduling, and edge resources to ensure latency-sensitive services can be delivered efficiently. Some manufacturers have developed integrated network solutions that combine energy-efficient hardware with advanced traffic management enabled by intelligent NFs. For instance, these systems leverage real-time analytics to dynamically adjust network traffic, resulting in significant power savings and reduced overall ownership costs, while delivering measurable improvements in key performance metrics [11].

2.3.2 RAN and AI

Scalable and user-centric RAN architecture leverages the cloud-native, compute-oriented infrastructure that underpins next-generation networks, enabling the transformation of traditional, communication-centric RAN into a compute-communication converged system. This convergence is primarily driven by the availability of flexible computing resources in distributed cloud environments, while scalability and user-centricity ensure that these resources are allocated efficiently according to user demands and heterogeneous workloads. This paradigm shift evolves conventional, single-purpose RAN infrastructure into an overarching RAN-AI architecture that integrates RAN and AI workloads on the same underlying computing platform. Bringing AI workloads into the RAN compute domain ushers AI-as-a-Service (AIaaS) capabilities for telcos, unlocking new monetization opportunities and reducing Operational Expenditure (OPEX) through improved network efficiency and asset utilization.

Integration of RAN and AI workloads exemplifies in three distinct forms: *AI-for-RAN*, *RAN-for-AI* and *CCIN*.

- **AI-for-RAN:** Integration of AI at the protocol/architectural levels of RAN to enhance network performance, such as improving spectral and operational efficiency, enabling predictive maintenance and optimizing radio resource management is the foundational principle of AI-for-RAN [12]. In addition, network automation has become essential with the rise of virtualized RAN. Within this context, AI-driven models, trained on historical data and continuously learning in real-time, are driving this zero-touch evolution, by automating tasks across the network’s Lifecycle Management (LCM) [13]. O-RAN ALLIANCE has been at the forefront of this zero-touch journey, harnessing AI capabilities in open RAN through RAN Intelligent Controller (RIC), which includes the non-real-time RIC (non-RT RIC) and the near-real-time RIC (near-RT RIC). The former hosts various AI-driven applications (rApps) at cloud/central site, within the Service Management and Orchestration (SMO) framework, with deployment options at cloud or central sites, operating on a ~second-level timescale, while the latter supports applications (xApps) deployed closer to the edge cloud, operating on a ~10ms timescale. Beyond the RIC, AI-for-RAN also extends to the air interface itself, enabling adaptive modulation, beamforming, channel prediction, and interference management. The adoption of AI-for-RAN by O-RAN ALLIANCE facilitates optimized RAN operations, including dynamic spectrum management, load balancing, energy savings, handover optimization and interference mitigation.

- **RAN-for-AI:** Utilizing RAN infrastructure, to support vertical (non-RAN) AI applications, is a growing trend across various industry verticals such as smart manufacturing, smart cities and IoT. This use case, often referred to RAN-for-AI, fosters a mutually beneficial relationship between two traditionally separate domains, i.e., RAN and AI, in the telecom space [13]. By virtue of the open, modular, and cloud-native architecture of O-RAN with open nodes and interfaces, O-RAN ALLIANCE has created a cloud-native network infrastructure blueprint with open interfaces that can natively support plug-and-play integration of vertical AI applications in the edge cloud. As one example, the potential of RAN-for-AI has been demonstrated by Softbank in November 2024, as part of their AITRAS solution (with O-RAN components), to show how AI integration with virtual RAN (vRAN) software can enable low latency and improve performance of a robotic dog instructed to follow a human [14]. However, realizing RAN-for-AI at scale in 6G will require several foundational enablers: standardized open interfaces between the RAN and external applications, secure and isolated handling of third-party workloads within shared infrastructure, and coordinated low-latency orchestration across radio, compute, and transport domains.

- **CCIN:** Facilitating the management and resource sharing of computing resources between RAN NFs and non-RAN AI applications is the key in incentivizing RAN-AI deployments in commercial telecommunications networks in a sustainable, cost-effective, and efficient way. By dynamically orchestrating RAN and AI workloads on a single infrastructure, resources can be optimally utilized, thereby maximizing asset utilization and infrastructure value. O-RAN ALLIANCE and its vibrant software community (OSC) have been championing the cloudification of RAN with disaggregation of traditional RAN components and virtualization of network functions to be deployable in the O-Cloud, which is a perfect infrastructure for enabling CCIN. Aarna Networks' Aarna Multi Cluster Orchestration Platform (AMCOP), developed by Aarna Networks (a vendor of cloud-native orchestration for 5G and edge computing), is one practical example of CCIN, based on O-RAN SMO. AMCOP not only automates the orchestration of RAN workloads on GPU clouds but also monitors resource usage to dynamically assign idle GPU cycles to AI applications [13].

2.3.3 Adaptive and resilient network

Resilience is a critical factor in networks, ensuring network reliability and performance even amid changing demands and potential disruptions. Achieving this requires a flexible and robust infrastructure, as depicted in Figure 2-2.

In traditional RAN architectures, service quality can be degraded by factors such as mobility, high traffic loads, and network failure events. On the other hand, the scalable and user-centric RAN in cloud architecture offers a potential solution to enhance resilience through its inherent adaptability and flexibility. This section describes how the architecture enhances network resilience through improved mobility management, load balancing, flexible network configuration, and failure recovery mechanisms.

- **Mobility management:** In conventional networks, performance may suffer from degradation as users move, leading to frequent handovers and poor signal quality at cell edges. The scalable and user-centric RAN addresses these challenges by dynamically assigning O-RUs to O-DUs, ensuring seamless connectivity. Here, "dynamic assignment" denotes rapid O-RU/O-DU reassignment at the O-RU granularity for load balancing and fault recovery; per-UE simultaneous anchoring at multiple O-DUs is not assumed under the current O-RAN connectivity model. This process requires fast fronthaul reassignment, coordinated scheduling across units, additional reassociation signaling, and precise time synchronization to maintain timing and phase alignment. Optional packet duplication and flow control across parallel links can help keep sessions intact during transitions. Additionally, RU-to-RU mobility is addressed through O-DU/O-CU-coordinated multipoint cooperation (with details provided later in this report), which mitigates cell-edge degradation and reduces handover interruptions, thereby improving overall network performance. In practice, the O-DU and O-CU coordinate the cooperating O-RUs by sharing channel state information and user equipment (UE) context, aligning beams and physical resource blocks across nodes, and transmitting the same user data simultaneously from multiple O-RUs. Control-plane coordination over existing interfaces and tight time/frequency synchronization keep symbols aligned so the UE can combine the parallel signals, improving edge throughput and reliability without resorting to hard handovers.

- **Load balancing:** During periods of high traffic, network congestion can lead to service quality degradation. In a user-centric RAN, traffic steering and resource allocation are decided on a per-user and per-application basis, guided by policy targets and conditioned on instantaneous RF conditions and transport capacity/latency. In practice, the dominant constraints are the user's radio conditions and the capacity of the O-RU to O-DU transport link. Accordingly, coordination focuses first on radio-side actions within the UE's feasible RF neighborhood (e.g., resource and beam steering and, where available, selective traffic split via dual connectivity or inter-frequency options). Compute- or transport-side steering (such as O-RU/O-DU reassignment or workload relocation) is applied only when the path has sufficient headroom and service-continuity constraints can be met. Shifting a user to a different O-RU is considered only in deployments with overlapping coverage (e.g., dense urban or indoor); otherwise, per-user scheduling and prioritization mechanisms are used. While the air interface (spectrum and bandwidth), which is per-user, coordination-aware approach, remains the primary limiting resource, this flexible resource allocation mechanism enables the network to adapt efficiently to changing traffic demands, maintaining network stability even under heavy load and ensuring optimal utilization of both radio and compute resources in evolving service scenarios. Although 5G and 5GA include features such as dynamic scheduling, traffic steering, and carrier aggregation, their deployment is often constrained by vendor heterogeneity, limited cross-domain control access radio, compute, and transport layers, and insufficient fine-timescale, per-user, context-aware coordination—challenges that 6G aims to overcome.

- **Flexible network component configuration:** A cloud-based RAN offers the flexibility to dynamically configure network components for optimized performance.

This dynamic configuration applies primarily to cloud-native O-DUs and O-CUs, as well as application functions; the O-RU however, remains an RF element whose operational parameters—such as association, beamforming power control, and functional-split configuration—are managed through orchestration interfaces rather than instantiated as virtual functions. Within the scalable and user-centric RAN, NFs and Application Functions (AFs)—representing application-level functions hosted in edge or central clouds that interact with the RAN through exposure or management interfaces to influence resource behavior (e.g., analytics or vertical applications)—can be allocated in real-time to meet diverse service requirements. Here, the term "cloud-based" refers to the deployment environment (edge or central clouds), whereas "user-centric" denotes the policy objective of per-user or per-application adaptation. This dynamic placement not only improves resource utilization but also adapts efficiently to changing user and service demands [15].

- **Network failure recovery:** Ensuring service continuity in the event of network failures is crucial. The scalable and user-centric RAN enhances resilience by enabling flexible network configurations that dynamically replace and reconfigure failed RAN elements. This adaptability allows uninterrupted service continuity by redirecting traffic and reallocating resources to operational nodes in real time. Moreover, automated failover mechanisms and cloud-based redundancy strategies further strengthen network stability, allowing seamless recovery from disruptions while maintaining optimal performance. While public cloud environments typically include built-in redundancy mechanisms, comparable high-availability strategies must be explicitly designed and managed within private cloud deployments.

The scalable and user-centric RAN represents a transformative approach to enhancing network resilience. By leveraging cloud-based capabilities, this architecture effectively addresses challenges related to mobility, load balancing, evolving service demands, and network failures. However, widespread adoption remains constrained by current technology readiness—specifically, the need for fast radio reassignment mechanisms, precise time synchronization, scalable automation, mature multi-vendor interoperability, standardized interfaces, and consistent deployment of edge-cloud and orchestration capabilities. As this approach continues to evolve, future 6G networks are expected to exhibit even greater resilience, ensuring seamless connectivity and enhanced service quality.

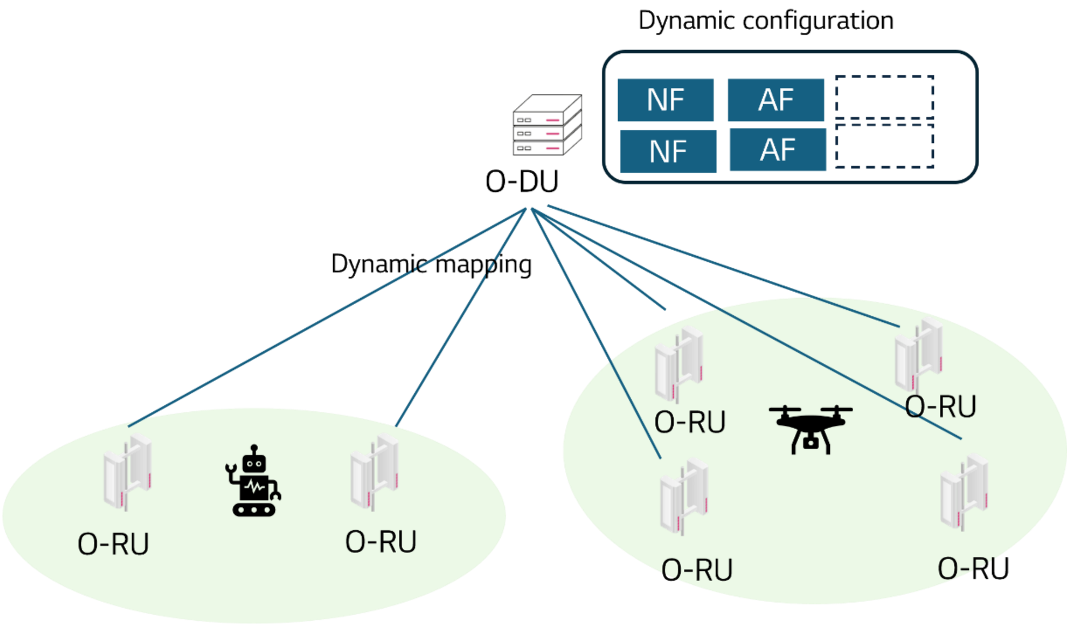

Figure 2-2 Flexible network configuration.

3 Requirements and principles

3.1 Introduction

The advancements in 5G mobile network technology introduced substantial computational challenges for RANs. In 5G mobile networks, the O-DU should manage complex functions like real-time signal processing, data encapsulation, decoding, and scheduling to meet the high bandwidth and low latency requirements. These tasks demand significant computational power, prompting the integration of hardware accelerators, such as CPUs, GPUs, and SoCs, to enhance both of efficiency and processing speed. Looking ahead, 6G mobile networks are expected to further intensify these demands due to increased service heterogeneity, higher data rates, and the integration of intelligent applications involving real-time communication between diverse devices, including XR, robots, and UAVs. In this context, scalability becomes even more critical. Relying solely on software processing would necessitate the addition of numerous servers, leading to higher costs and operational complexity. By integrating hardware accelerators, operators can achieve higher throughput within the same server footprint and power envelope, improving both of scalability and cost-effectiveness. This section examines the general aspects of hardware acceleration in user-centric RANs, covering types of accelerators, their architectures, realistic RAN topologies, and effective placement patterns for O-DU and O-CU units. It also addresses constraints, requirements, and associated pros and cons, providing a comprehensive understanding of how hardware acceleration meets the performance and scalability needs of emerging 6G mobile networks.

3.2 Acceleration for scalable and user-centric RAN

In the rapidly evolving landscape of mobile networks, efficient processing hardware has become a cornerstone for optimizing RAN performance. As network demands escalate with the expansion of 5G mobile networks and the advent of user-centric RAN technologies, the need for robust processing of complex RAN functions intensifies. Various types of processing units, including general-purpose CPUs and specialized accelerators like GPUs, SoCs and FPGAs, where algorithm agility is required, are employed to handle these demands. These processors work in tandem to enhance throughput, reduce latency, and improve energy efficiency by managing compute-intensive tasks within the RAN architecture [16]. The primary types of processing hardware used in RAN are:

- **Central Processing Unit (CPU)** is often referred to as the brain of the computer. It is the primary component that executes instructions and processes data in a computing system. Designed for versatility, CPUs are capable of handling a broad spectrum of tasks, making them indispensable in general-purpose computing. In the context of RAN, CPUs predominantly manage control plane functions. The control plane is responsible for network signaling, management, and configuration tasks that maintain the network’s operation and facilitate communication between devices. CPUs excel in this role due to their ability to handle complex decision-making processes, execute sequential operations, and manage intricate protocols. For example, tasks such as authentication and security procedures, mobility management, and resource allocation are typically handled by CPUs. They process commands from the core network, manage connections with the UEs, and oversee the setup and teardown of communication sessions. Additionally, CPUs handle fewer demanding user plane operations, which involve the actual transmission of user data through the network. While CPUs can process user plane tasks, they are not optimized for the high-throughput, low-latency requirements of intensive signal processing found in the physical layer of modern networks like 5G.

- **Graphics Processing Unit (GPU)** was originally developed to accelerate the rendering of graphics and images by handling multiple calculations simultaneously. GPUs are engineered for parallel processing, containing thousands of smaller, specialized cores designed to perform many operations concurrently. This makes them highly effective for tasks that can be broken down into smaller, repetitive computations. In modern computing, GPUs have evolved beyond graphics rendering and are now widely used for General-Purpose GPU computing. This involves leveraging their parallel architecture for a variety of computational tasks that benefit from parallelism. Within RAN architectures, GPUs are utilized to accelerate compute-intensive user plane operations, particularly in the physical layer:
  - **Signal processing:** GPUs process complex mathematical computations required for modulation and demodulation, filtering, and spectrum analysis.
  - **Massive MIMO:** They handle the parallel processing of signals from multiple antennas, increasing network capacity and spectral efficiency.
  - **Beamforming:** GPUs calculate the optimal transmission paths, enhancing signal strength and reducing interference.
  - **Channel coding and decoding:** They efficiently perform error detection and correction algorithms, ensuring data integrity over the air interface.

By offloading these tasks to GPUs, RANs can achieve higher throughput, reduced latency, and improved user experiences. GPUs also enable networks to support bandwidth-intensive applications like high-definition video streaming, virtual reality (VR), and large-scale IoT deployments.

- **System-on-Chip (SoC)** integrates multiple processing engines (e.g., CPU cores, graphics/AI engines, and baseband-oriented digital signal units) on a single chip, delivering high performance per watt with low inter-engine latency. In RAN deployments, SoCs commonly accelerate physical-layer workloads (such as channel coding/decoding and beamforming), assist scheduling, and run AI inference close to the data path. Many modern SoCs also include limited programmable logic blocks, providing some adaptability while retaining energy efficiency suited for large-scale 6G rollouts. In RAN architectures, SoCs are used to:
    - **Accelerate physical-layer workloads:** They handle channel coding/decoding, beamforming and MIMO processing, channel estimation, filtering, and FFT efficiently.
    - **Run time-critical control and AI inference:** They assist scheduling and manage rapid retransmissions by making quick success or failure decisions, temporarily storing and combining re-sent bits, and promptly triggering the next transmission near the radio path with minimal added latency.
    - **Improve energy efficiency** They apply fine-grained power management (such as power gating and dynamic frequency control) to lower energy per bit while sustaining throughput.
    - **Support secure, virtualized deployment:** hey provide hardware isolation so workloads can be safely placed in edge or central clouds alongside O-DU/O-CU functions.

- **Field-Programmable Gate Array (FPGA)** is a type of programmable logic device that allows hardware customization after manufacturing. FPGAs consist of an array of configurable logic blocks connected by programmable interconnects, enabling designers to implement custom digital circuits tailored to specific applications. In essence, FPGAs bridge the gap between the flexibility of software and the performance of hardware. They offer the ability to create highly specialized hardware accelerators that can be updated or reprogrammed to meet evolving standards and requirements. In RAN architectures, FPGAs are instrumental in accelerating tasks that demand deterministic performance and ultra-low latency:
    - **Custom signal processing pipelines:** FPGAs can implement specific algorithms for Fast Fourier Transforms (FFT), Digital Signal Processing (DSP), filtering, and other mathematical operations critical in wireless communication.
    - **Protocol acceleration:** They can handle protocol-specific functions at hardware speed, such as framing, synchronization, and time-critical control functions.

  - **Encryption and decryption:** FPGAs can perform cryptographic operations efficiently, enhancing security without compromising performance.
  - **Radio interface functions:** Tasks like Digital Up/Down Conversion (DUC/DDC) and pre-distortion can be offloaded to FPGAs to optimize radio performance.

By deploying specialized accelerators (with SoCs as the default choice and FPGAs where programmability is essential), network operators can achieve hardware-level acceleration for specialized tasks, significantly improving throughput and reducing latency compared to software-based implementations on CPUs or GPUs. However, in large-scale deployments, FPGAs often entail higher device cost, higher energy per bit, and specialized development toolchains, which limits their cost-effectiveness and scalability; they are therefore best suited to prototyping, low-volume rollouts, rapidly evolving functions, or cases requiring strict determinism. FPGAs are particularly valuable in scenarios where processing needs are not only intensive but also require precise timing, such as in Ultra-Reliable Low-Latency Communications (URLLC) defined in 5G standards, provided that the associated cost and power trade-offs are acceptable. From an energy perspective, offloading these pipelines to dedicated logic lowers energy per bit/operation relative to general-purpose CPUs; for mature, fixed workloads, SoCs typically deliver higher performance per watt, while FPGAs retain flexibility for evolving algorithms.

The integration of acceleration hardware into the RAN is accomplished through different acceleration architectures, primarily inline and lookaside acceleration [17]. The choice of architecture affects how data flows through the network and how processing tasks are managed.

- **Inline acceleration** embeds hardware accelerators directly into the data processing path. Data packets flow through the accelerator, which performs necessary computations without deviating from the main path. This seamless integration minimizes delays which is ideal for tasks that require real-time processing and ultra-low latency, primarily in the physical layer. Inline acceleration ensures deterministic performance and high throughput but can complicate system design and pose scalability challenges. This yields deterministic latency and high throughput for time-critical physical-layer functions; trade-offs include greater integration complexity, tighter coupling to the processing pipeline, and reduced flexibility for upgrades at scale, where changes can be disruptive.

- **Lookaside acceleration** places hardware accelerators parallel to the main data path. Data requiring specialized processing is diverted to the accelerator and then returned to the main flow upon completion. This configuration allows for selective offloading without interrupting primary data transmission. Lookaside acceleration is well-suited for tasks where latency is less critical, often in the data link layer and higher layers. It offers greater flexibility and resource optimization but may introduce additional latency and requires complex control logic. It also enables sharing of accelerators across services and easier technology insertion as demands evolve; the trade-offs are extra dispatch/return latency, more complex control and buffering, and bandwidth overhead from detouring traffic, so it is less suitable for immediate, ultra-low-latency functions.

- **Integration challenges and strategic approaches for acceleration technologies:**
  - Challenges associated with acceleration technologies:
    - The integration of CPUs, GPUs, and FPGAs for RAN acceleration introduces increased complexity in system design that necessitates specialized development skills, which can complicate maintenance and updates.
    - CPUs may encounter performance limitations when managing intensive computational tasks in the physical layer, leading to potential bottlenecks and scalability issues.
    - Although GPUs offer significant performance improvements, they come with higher initial costs due to specialized hardware requirements. Additionally, they entail high operational cost including non-trivial power/cooling overheads, and their implementation requires expertise in parallel programming frameworks, which can increase development effort and time. Furthermore, GPUs may consume more power, potentially affecting energy efficiency in certain applications.
    - FPGAs involve a complex development process that necessitates proficiency in hardware description languages and skilled personnel for effective implementation and typically entail higher device cost, which can limit cost-effectiveness at scale.
    - The integration of GPUs and FPGAs may also necessitate significant architectural changes, leading to potential integration difficulties. Furthermore, the limited ecosystem and resources available for FPGA development can pose additional challenges. Overall, for both GPUs and FPGAs, total cost of ownership (acquisition, power/cooling, and operational complexity) should be weighed against performance gains, particularly for large deployments.

  - Challenges and strategies for integrating acceleration hardware in vRANs within O-Cloud environments:
    - The lack of standardized interfaces for accelerator integration hinders interoperability and complicates the deployment of diverse technologies. This has the potential to lead to vendor lock-in, which restricts flexibility in technology choices.
    - Current cloud orchestration tools, such as Kubernetes, often struggle to manage specialized accelerators efficiently; this results in suboptimal resource allocation and underutilization of hardware. Therefore, enhanced orchestration tools must integrate accelerator-

aware scheduling into platforms like Kubernetes to ensure efficient resource management and dynamic scaling.

- The process of virtualization introduces additional latency, which undermines the benefits of hardware acceleration, particularly for latency-sensitive applications. Additionally, the overhead associated with virtualization can significantly impede the expected performance gains from acceleration technologies.
- It is essential to adopt standardized interfaces and application programming interfaces (APIs). Furthermore, utilizing common frameworks, such as the Open Programmable Acceleration Engine (OPAE), can enhance interoperability and simplify integration across various platforms.
- Implementing latency minimization strategies, such as Single Root I/O Virtualization (SR-IOV) and the Data Plane Development Kit (DPDK), can help reduce virtualization overhead while maintaining low-latency processing paths [18].

Security and isolation in multi-tenant environments create further complexities in the deployment of acceleration hardware. Hence, robust security measures, including strict access controls and effective isolation mechanisms, are imperative for protecting against unauthorized access and safeguarding data integrity in multi-tenant environments.

3.3 Effective placement for virtualized O-DU/O-CUs

In the design and deployment of mobile networks, the topology and placement of O-DU and O-CU units play critical roles in ensuring efficient network performance and scalability. Various topologies and placement patterns are employed to meet the diverse requirements of modern networks [4]. Among the most common are ring topology, tree topology, and configurations like Centralized RAN (C-RAN) and Distributed RAN (D-RAN). Understanding these topologies and their implications is vital for optimizing network design, enhancing coverage, and ensuring seamless connectivity. In the context of mobile networks, the strategic and dynamic placement of O-DU and O-CU units is critical for optimizing network performance, ensuring low latency, and maximizing resource efficiency. For clarity, this subsection focuses on O-DU/O-CU topology and placement in the transport/cloud domain. Constraints tied to the physical distance between radio heads and an O-DU, and the resulting fronthaul timing budgets, are acknowledged but not analyzed here. Standards should enable a range of interoperable placement options. However, the actual placement of O-DU/O-CU remains implementation- and deployment-specific. Accordingly, we focus on key considerations, such as latency/transport constraints, availability of acceleration, and operational/cost factors, rather than prescribing a single placement. Different placement patterns and strategies are adopted to achieve these goals, each with its own set of constraints, requirements, and trade-offs. This section explores effective placement for edge cloud and low latency services, effective placement for resource optimization, and the associated gap analysis and requirements for optimizing O-Cloud environments.

- **Ring topology** is a network configuration where each node is connected to exactly two other nodes, forming a circular data path. This topology provides a simple yet robust design that offers several advantages in deployment. The inherent redundancy of ring topology ensures continuous operation, as data can be rerouted in the opposite direction if a single link fails. This enhances network reliability, crucial for maintaining service quality in mobile networks. Furthermore, adding new nodes is straightforward, as new O-DUs or O-CUs can be integrated into the existing ring structure without significantly disrupting the network. However, potential latency issues arise due to the path length, especially in larger rings, and multiple node or link failures can still cause significant disruptions.

- **Tree topology** is a hierarchical network structure that resembles a tree, with a central root node branching out to multiple levels of subordinate nodes. This topology is highly effective for organizing and managing large networks. Its structured hierarchy facilitates efficient data routing and management, making it easy to scale by adding new nodes at appropriate levels without disrupting the overall network. Tree topology also simplifies network management and troubleshooting due to the clear parent-child relationships between nodes. However, the central or intermediate nodes pose single points of failure, potentially isolating entire subtrees if a node fails, necessitating robust redundancy and fault-tolerant mechanisms.

- **C-RAN** centralizes the baseband processing units in a central location, while remote radio heads are distributed across the network. This approach improves resource utilization and centralized management, reducing the overall number of required baseband units and enabling advanced features like Coordinated Multipoint (CoMP) and centralized interference management. However, C-RAN can introduce latency issues associated with centralized signal processing and require high-capacity fronthaul links, increasing deployment costs.

- **D-RAN** distributes both radio and baseband processing closer to end users. This lowers latency and supports local processing, improving responsiveness and user experience. The trade-offs are higher site-level complexity and operational cost, since more locations host processing resources and require management and maintenance.

- **Examples illustrating the requirements and principles:**
  - Optimizing O-DU and O-CU Placement at network edges:
    - Placing O-DUs and O-CUs closer to end-users is essential for delivering edge cloud services and maintaining low latency. This strategic proximity reduces the physical distance that data must travel, thereby reducing RAN-side processing and access latency for critical applications that require real-time processing, such as augmented reality (AR), VR, UAVs, and industrial automation.

    - Placing O-DUs and O-CUs at the network edge supports distributed processing, enabling the offloading of computational tasks from central locations to edge nodes. This improves network efficiency by balancing the processing load and reducing congestion in the core network. However, this deployment necessitates substantial infrastructure investment, as each edge node must be equipped with the required hardware and connectivity resources. It should be noted that the end-to-end latency is minimized when latency-critical application servers are co-located in the same edge cloud. Otherwise, traffic must still traverse the wider network to reach the application.
    - These requirements can increase operational complexity and costs, as the maintenance and management of a larger number of distributed units pose notable challenges. Furthermore, the implementation of robust monitoring and maintenance protocols is necessary to ensure consistent performance and reliability across all edge nodes.

  - Optimizing O-DU and O-CU placement in O-Cloud environments:
    - The lack of standardized interfaces and protocols for integrating diverse acceleration technologies complicates multi-vendor deployments for the placement of O-DUs and O-CUs at O-Cloud environments.
    - Current orchestration tools are often limited in their ability to manage specialized hardware accelerators efficiently, hindering dynamic resource allocation in O-Cloud environments.

3.4 Scalable architecture for cell-free and distributed MIMO

The proposed scalable architecture provides a valuable opportunity to implement a new paradigm for wireless networks which extends the concepts of CoMP and multi-TRP introduced in earlier generations of 3GPP standards [19], [20], towards the full Network MIMO concept [21]. In recent years, a variety of terms have been used for this paradigm, including cell-free massive MIMO (CF-mMIMO) [22], distributed massive MIMO, and more simply distributed MIMO (D-MIMO) [23]. In its most basic form CF-mMIMO assumes that all UEs are served by all Access Points (APs) in the network – where AP denotes the radio transmission point. In 3GPP terms, this aligns with the TRP (which may be a component of a gNB), and in O-RAN ALLIANCE it most closely maps to the O-RU. A gNB is a broader logical node, often split into O-CU/O-DU/O-RU, and is not equivalent to a single AP. All baseband processing is then performed by a CPU located at a single central point, with signals being transferred between all APs and the CPU over AP-to-CPU transport links (for example, the O-RAN ALLIANCE Open Fronthaul interface when the processing entity is an O-DU). At the gNB coordination and scheduling level, this removes conventional per-cell boundaries (hence the term “cell-free”). Thus, the cell-free concept enables multiple APs to jointly serve a UE; however, it does not abolish frequency-reuse planning (“cells”) in the spectrum-management sense. The resulting system may be regarded as a large-scale distributed multi-user MIMO system in which all the antennas on all APs throughout the network cooperate to serve all UEs–hence the terms “Network MIMO” and “Distributed MIMO”. Note that, the O-RAN ALLIANCE Open Fronthaul interface does not define a “cell”; the “cell-free” notion applies to RAN operation/association rather than to the fronthaul interface. In this section we will henceforth use the term D-MIMO as a general term for such a RAN architecture.

Since it encompasses all other possible RAN architectures, this paradigm allows optimal performance over all such architectures involving the same APs. Moreover, it assumes that APs and UEs cooperate within a single cooperative domain without fixed cell boundaries (“cell-free”), thereby avoiding intercellular interference through joint processing and coordination, which constitutes the main limitation on the capacity of conventional cellular systems. It can therefore be expected to outperform conventional cellular RAN architectures. Nevertheless, spectrum remains the fundamental scarce resource: spatial reuse is still required, and in practice each UE is served by a small, user-centric cluster of nearby APs, allowing geographically separated UEs to reuse the same frequencies. However, the system still fails to scale efficiently as the number of served UEs and serving APs becomes large, as the number of fronthaul connections and the processing per UE in principle grow unbounded. The reason can be summarized in three points: first, spectrum-reuse and pilot/CSI overheads impose practical limits; second, the number and capacity requirements of AP-to-processing transport links scale rapidly with network size; third, tight synchronization constraints and per-UE joint precoding/decoding complexity both increase sharply as the network expands.

Hence, the proposed scalable and user-centric network architecture described in this document provides an opportunity to realize a scalable and user-centric version of D-MIMO, which has the potential to significantly improve capacity. While a variety of proposals for scalable D-MIMO (or CF-mMIMO) systems have been presented in recent literature [24], [25], here we describe a straightforward scheme for its implementation which we refer to as Decentralized Distributed MIMO (DD-MIMO) [26]. This scheme has the advantage that aligns naturally with the O-RAN architecture, in which baseband processing takes place within a O-DU that may be flexibly located, rather than relying on a centralized or distant cloud-based processing entities.

The scheme is also user-centric, in that it allows an O-DU to be selected for signal processing for each user based on user performance criteria. The most important of these are likely to be latency and physical layer (PHY) performance, in terms of achievable rate for given signal and noise power. There may in practice be a trade-off between latency and PHY layer performance, since the more APs that can cooperate to serve a given UE the better the PHY layer performance is likely to be – however increasing this number also increases the number of UEs whose signals are processed in the same O-DU, requiring it to be located further from some of the UEs and increasing latency, as well as increasing processing load in that O-DU.

There is likely to be an optimal trade-off between PHY layer performance and latency, occurs when a small cluster of UEs is served by a given O-DU. The uplink signals from this cluster of served UEs are received at a significant level at a limited subset of the

O-RUs in the network. By uplink-downlink reciprocity, The O-RUs that receive significant uplink energy typically also transmit to the same set of UEs on the downlink. We refer to this subset as the participating set of and to the area enclosing them as the coordination region. Figure 3-1, the coordination region for the O-DU under discussion corresponds to the red circle; the blue and green circles depict coordination regions of neighboring O-DUs, and the overlaps indicate O-RU sharing across O-DUs. Since the coordination region is more extensive than the service region in which the cluster of served UEs lie the participating set for each O-DU is highly likely to overlap with those for neighboring O-DUs, which implies that some O-RUs should be connected to more than one O-DU, leading to the architecture illustrated in Figure 3-1.

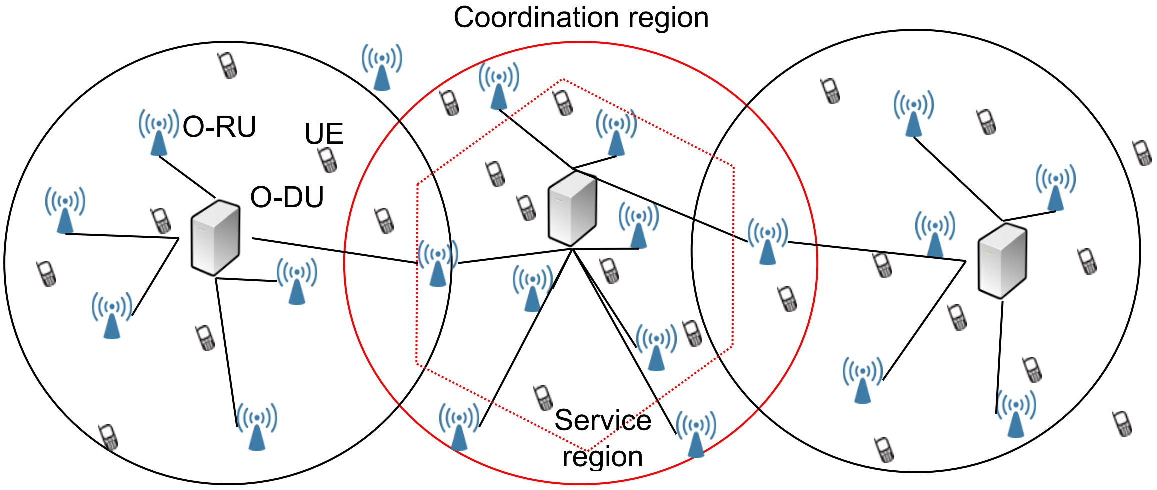

Figure 3-1 Decentralized Distributed MIMO.

However, it should be noted that there is typically no sharp distinction between O-RUs that receive significant uplink power from the served UEs and those for which this received power is entirely negligible. Consequently, the size of the participating set for each O-DU is a design choice which determines the trade-off between the PHY layer performance and the complexity of and load upon the fronthaul network. The larger the participating set, the larger the total signal power available to the RAN (on the uplink), or to the UE (on the downlink), but also the more O-RUs connected to a single O-DU and the more O-RUs connected to more than one O-DU.

The benefit of this architecture arises from the increased number of degrees of freedom in the MU-MIMO system it creates: hence it requires a sufficient number of cooperating O-RUs and hence also of served UEs in a cluster. Thus the benefit is not available in a fully distributed RAN, where the O-DU is connected to and collocated with a single O-RU. However, the increase in path loss results in diminishing returns in this respect beyond a relatively small number of O-RUs, while the network complexity increases approximately as the square of the radius of the coordination region: this clearly creates an optimum trade-off with a relatively small number of served UEs for a given O-DU, as mentioned above.

Note also, the importance of channel estimation in such a system and the effect this has on the D-MIMO architecture. D-MIMO is effectively a multi-user MIMO (MU-MIMO) system, in which all the O-RUs in the participating set provide a composite MIMO antenna array to serve the UE cluster, and thus requires accurate knowledge of the MIMO channel between each UE and this composite antenna array. In both the LTE and 5GNR standards this channel knowledge is obtained at the network side using the demodulation reference sequences (DM-RSs) transmitted by each user. 5GNR (in current releases) provides 12 distinct and orthogonal DM-RS sequences, in principle allowing 12 users to be distinguished in any given region of coverage. This of course requires DM-RSs to be re-used across the network: the re-use of the DM-RS being used by a member of the served UE cluster of interest may lead to *pilot contamination*, in which the channel from the nearest user of the same DM-RS contaminates the channel estimate of the wanted user, which in turn may give rise to interference between the two UEs. Now a UE outside the served cluster but lying within the coordination region may contribute significant signal power and potentially also cause pilot contamination. Hence, we should avoid re-using a DM-RS not only within the UE cluster served by the same O-DU but also in its coordination region – across which region choice of DM-RS needs to be coordinated: hence the name. These considerations also limit the number of UEs served by a given O-DU, and the number of O-RUs in the participating set.

Figure 3-1 suggests that O-DUs should be located in specific locations, but this does not in fact preclude a more flexible implementation of the sort described in section 2.2, although placing a O-DU close to the service region containing UEs it serves is likely to minimize latency. Again, there is a trade-off between minimizing latency and providing flexibility for the network to adapt to varying demands.

3.5 Scalability for energy efficiency

Dynamic resource allocation uses AI/ML to predict O-RU traffic, anticipates peaks and troughs in traffic volume, and assigns O-RUs to vDUs/vCUs dynamically via the fronthaul network switch (Open Fronthaul over Ethernet using the enhanced Common Public Radio Interface (eCPRI) protocol). As a result, O-RUs can be consolidated onto a smaller number of vDUs/vCUs, and any idle vDUs/vCUs can be temporarily deactivated to reduce power consumption during low-traffic periods as shown in Figure 3-2. Eight O-RUs are assigned to four vDUs/vCUs on the left side of Figure 3-2. If traffic load goes down a certain threshold, all vDUs/vCUs that have no more active O-RUs can be deactivated, as the right side of Figure 3-2. These inactive vDUs/vCUs are then reactivated when capacity demand increases. Dormant vDU/vCU capacity that can be software activated on demand also improves overall system reliability and resilience, by keeping a small pool of quickly activatable capacity and enabling rapid re-homing and recovery when a fault occurs. Consolidation does not eliminate work, the UE load is shifted, so savings arise only when unused servers can enter deep-idle or power-off states; idling containers alone provides limited benefit. Dynamic resource allocation can assign vDUs to vCUs dynamically too. AI technology anticipates changes in traffic patterns and uses this data to predictively scale capacity within latency and service-continuity constraints. In this discussion, the transport path and its energy use are treated as roughly constant; any routing adjustments are handled by orchestration and detailed reconfiguration costs are out of scope of this report. Finally, radio transmission power remains the largest energy component at the site; the consolidation described here targets the compute and cooling share and is intended to complement radio-layer energy-saving features.

Modern GPUs often incorporate Multi-Instance GPU (MIG) technology, which enhances energy efficiency by selectively deactivating and powering off idle or unused cores during periods of low demand [27]. Additionally, the converged GPU card can partition its resources into several instances, each fully isolated with its own high-bandwidth memory, cache, and compute cores. This capability can be very helpful for MNOs. By activating just a few of the cores, the MNO can minimize OPEX while retaining the option to increase processing capacity automatically as they grow, and as traffic demands increase.

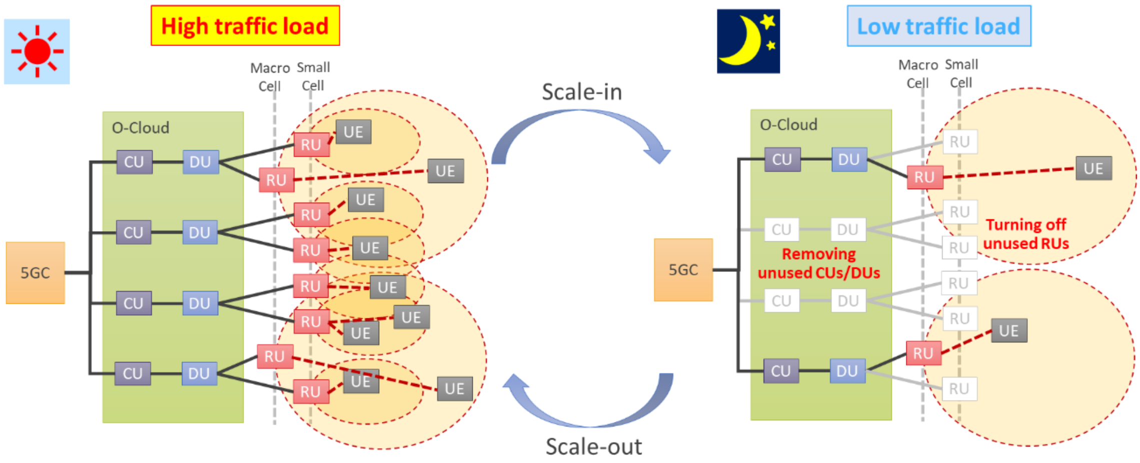

Figure 3-2 Deactivating vDU for reducing power consumption.

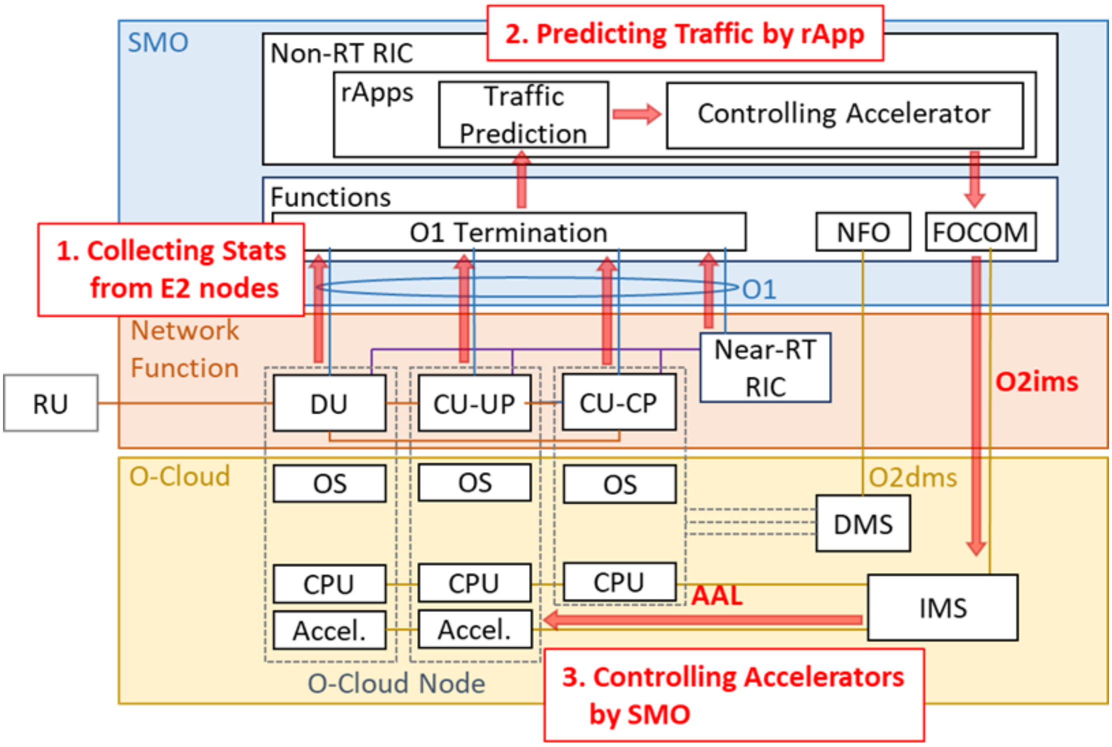

Figure 3-3 Procedures for controlling accelerators on O-RAN architecture.

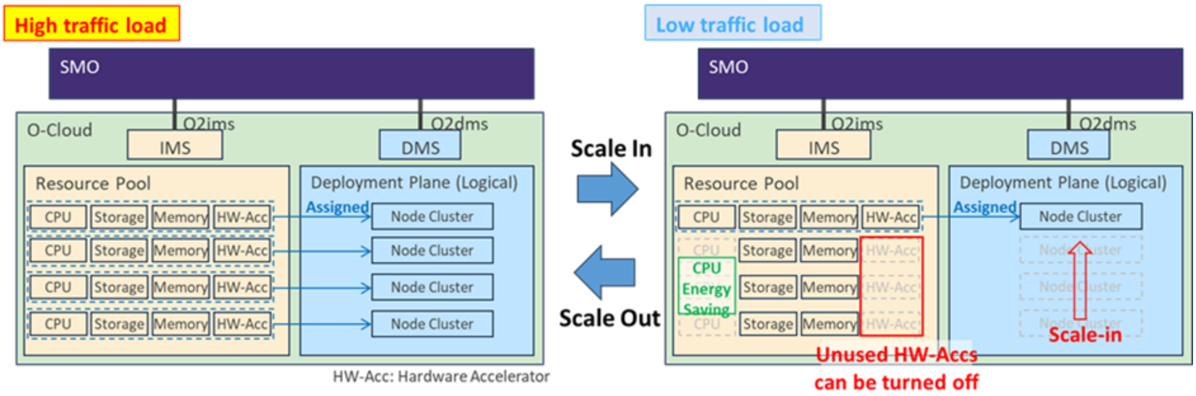

Figure 3-4 Resource mapping to O-Cloud nodes and networks.

Dynamic resource allocation can be realized by O-RAN interfaces. A high-level architecture of O-RAN is shown in Figure 3-3. O-Cloud is a cloud computing platform for O-RAN and contains resources, such as CPU, accelerators, memory, and storage. O-RAN NFs, such as O-CU and O-DU, are deployed onto O-Cloud. Resources of O-Cloud can be mapped to each NF shown as Figure 3-4. Deployment Management Services (DMS) is a function of O-Cloud and can deploy (activate) and remove (deactivate) each NF. On the other hand, Infrastructure Management Services (IMS) is also a function of O-Cloud and can manage and control each resource. Both DMS and IMS can be controlled by SMO via O2dms and O2ims interfaces respectively. rApps of Non-RT RIC in SMO are applications for optimizing the vRAN and O-Cloud by utilizing the statistics from O1 and O2 (ims/dms) interfaces.

For overall optimization of O-Cloud energy efficiency, workload assignment and power status for each resource in O-Cloud should be controlled by SMO. The following four functions are required for SMO. Figure 3-3 describes the relationship between functions and interfaces standardized by O-RAN.

- SMO can periodically collect statistical information about traffic load and power consumption of O-Cloud by using O1 and O2 (ims/dms) interfaces.
- rApps of Non-RT RIC in SMO can leverage statistical data to estimate near future RAN workload and power consumption, providing predictions from approximately few second up to several minutes ahead.
- SMO can utilize IMS via O2ims and AAL for controlling each resource's workload allocation and power status.
- SMO can utilize DMS via O2dms for deploying and removing NFs.

If NF templates with and without utilizing specific computing resources such as accelerators are prepared, NFs without the computing resource can be utilized with low traffic demand and the energy consumption of virtualized RAN will be reduced. Some extension of O2ims, O2dms, and AAL interfaces would be required for controlling workload allocation and power status of each resource from SMO.

3.6 RAN and AI

The foundation of RAN and AI convergence can be built upon a set of technology enablers and principles that are fundamental to building scalable, user-centric, high-performance, and intelligent RAN which seamlessly integrate communication functions with advanced computational capabilities. In this section, “RAN-AI” is used as an umbrella spanning AI-for-RAN, where AI models optimize RAN functions; RAN-for-AI, where the RAN platform hosts vertical AI applications at the edge; and CCIN, where RAN and non-RAN AI workloads share and are co-managed on the same platform. To that end, the key requirements, and principles of realizing RAN-AI are described below [13].

- **Cloud native architecture**: RAN-AI needs a paradigm shift from the conventional RAN (a hardware centric evolution approach from one generation to another) to a software-centric, cloud native design which offers scalability, flexibility, and agility. In a software-defined architectural framework, the NFs running on RAN-AI infrastructure are decoupled from the underlying hardware platform, enhancing flexibility, portability, and programmability. RAN NFs implemented as software on general purpose processors (like CPU or GPU) instead of special-purpose, custom-build hardware (like Application-Specific Integrated Circuit (ASIC) or FPGA) allow dynamic instantiation as Cloud Native NFs (CNFs), facilitating dynamic resource allocation and orchestration. This software-based approach enables quick updates and reconfigurations of NFs running on RAN-AI without hardware changes or upgrades as in traditional RAN. At the same time, emerging edge AI applications and other user-centric services can be readily integrated into such agile network infrastructure, serving efficiently both RAN and AI workloads on the same platform.

One of the key tenets of RAN-AI is to support elastic scaling of both computing and communication resources, which, in turn requires adaptability to demand by adding or removing CNFs as needed. The cloud-native architecture of RAN-AI is fundamental to achieving such elasticity and efficiency. In particular, the following characteristics of cloud-native architecture are paramount to building RAN-AI:

- Microservices Architecture: CNFs are decomposed into microservices, each handling specific functions and adapting to varying workloads.
- Containerization: technologies like Docker and Kubernetes are useful for packaging and deploying CNFs as microservices, enabling efficient resource utilization in multi-tenant environments.
- Multi-tenancy Support: multiple RAN and AI services can run on the same shared infrastructure, maximizing resource utilization and efficiency.

Another important tool of cloud computing that would be beneficial for RAN-AI is continuous integration and continuous delivery/deployment (CI/CD). Embracing CI/CD practices for both RAN and AI services would allow RAN-AI for rapid deployment of new features and upgrades and maintaining service continuity with minimal disruption during network updates, upgrades and deployments of new AI services and NFs.

- **Joint orchestration of RAN and AI workloads:** RAN-AI's multi-tenancy requires seamless coordination between communication and computing resources, which can be achieved through a unified orchestration platform. This platform enables the real-time and on-demand allocation of resources based on factors such as network traffic, user demand, and SLAs. The joint orchestrator should manage computational resources (e.g., GPU, CPU, memory, NIC) alongside communication resources (e.g., time, frequency, bandwidth) to meet the specific needs of various workloads. AI models should be integrated into the orchestration platform to make intelligent, data-driven decisions.

  Agility is essential for adapting to fluctuations in network conditions and computing demands. The orchestrator must scale resources up or down as needed to efficiently handle varying traffic loads during peak and off-peak times. It should leverage the capabilities of the underlying hardware to enable dynamic scaling. For example, techniques like MIG can partition a single physical GPU into multiple instances with memory and fault isolations. This allows the orchestrator to support simultaneous processing of RAN workloads and edge AI applications using shared GPU resources, ensuring efficient utilization and flexibility.

- **Accelerated computing:** RAN-AI requires more versatile, software-defined accelerators like GPUs to handle intensive workloads, including both RAN computations and AI model training and inference. The infrastructure must be optimized for both training AI models as well as deploying the trained models for inference at the edge. While large-scale AI model training may occur in the cloud, real-time, online training and inference need to happen at the edge nodes with integrated computing capabilities. To handle the vast amount of data generated by

RAN and AI services in real-time, accelerated computing infrastructure with massively parallel processing capability is essential. This capability enhances AI algorithm performance, baseband signal processing and computing speed for intensive workloads such as massive MIMO beamforming, multi-cell scheduling, integrated sensing and communication (ISAC).

Additionally, the infrastructure should feature high-speed, low-latency interconnects to ensure rapid and seamless communication across network nodes. Network Interface Cards (NICs), especially smart NICs and Data Processing Units (DPUs) enable distributed AI-driven network operations, enhance real-time AI processing, optimize network traffic, and ensure seamless integration between cloud and edge computing environments – thereby enabling high-speed, low-latency data transfer between RAN-AI computing elements, storage and network nodes.

- **Network digital twin:** In order to enable native-AI support on RAN-AI infrastructure, Network Digital Twins (NDTs) will be crucial for providing a risk-free, experimental sandbox for simulating and testing AI algorithms and network strategies before deploying in real RAN-AI networks. NDTs would allow for pre-validation of AI-generated optimization strategies for RAN-AI, testing of new AI services and network configurations and evaluation of network performance under various loading conditions (with RAN and AI workloads co-existing and time-varying). NDT can simulate RAN-AI networks at various scales – all the way from micro deployments to large, city-scale models with hundreds of base stations. This scalability, combined with configurability and customization options in NDT allows testing of various AI algorithms for RAN in diverse use case scenarios as well as adaptation of AI models to specific network requirements and conditions, ensuring efficient integration of AI capabilities into the RAN domain in RAN-AI. The fidelity of AI-models enhancing RAN performance relies on the availability of large-scale datasets collected from base stations, user devices and network nodes for training and inference. Such diverse and rich dataset may not always be available. NDTs can fill in this gap, serving as the source of high-quality synthetic data generation. Alongside data generation, NDT can provide a safe environment for development and refinement of AI models (e.g., training and validation of Large Language Models (LLMs) and other AI applications), the key enablers of RAN-AI.

3.7 Adaptive and resilient network

To effectively manage the scalable and user-centric RAN, a transition to intelligent network management approach is required. Unlike traditional network management approaches, dynamic network configuration can be supported in response to UE mobility, real-time user demands and traffic load. This necessitates an advanced management system that autonomously detects and optimizes network configurations.

This section describes the requirements for efficiently managing a dynamically changing network, with a focus on real-time network status monitoring, network configuration optimization, and proactive failure prediction mechanisms.

- **Real-time network status monitoring:** In a scalable and user-centric RAN architecture, the network components, including O-RU, O-DU, NF, and AF configurations, can dynamically configure based on user location, traffic loads, and service requirements. Therefore, a network management system should be able to continuously monitor the status of these components and determine optimal network configurations in real time.

  To support this, each network component (O-RU, O-DU, NF, etc.) should report network state information periodically using standardized interfaces while instantly notifying the management system of any changes. The network status information can include traffic volume, resource utilization, and fault occurrences.

  To minimize the overhead of reporting this information, AI-based data filtering and compression techniques can be used. This ensures that only essential data is transmitted, reducing unnecessary network load while improving operational efficiency.

- **Network configuration optimization:** Within the scalable and user-centric RAN architecture, network topology and component configurations can be dynamically adapted to reflect real-time user mobility, traffic load, and service demands. To achieve this, the network management system can leverage AI/ML to analyze real-time data and optimize network configuration.

  AI/ML can be used to analyze real-time network conditions, such as traffic load, mobility patterns, and available resources, to determine the optimal network configuration.

- **Fault prediction and recovery:** Ensuring uninterrupted network operations requires proactive fault detection and rapid failure response mechanisms. AI-based fault detection can be further enhanced through Digital Twin technology, which provides a virtual replica of the physical network for real-time simulation and predictive analytics.

  By continuously analyzing network operation, the digital twin can detect anomalies and predict failure events before they occur. Using this information, the network can make preemptive network adjustments by simulating various fault scenarios. Also, when failures are predicted, the network can autonomously reconfigure by rerouting traffic through alternative paths or leveraging adjacent nodes for rapid recovery.

  Therefore, and as depicted in Figure 3-5, by integrating AI-based predictive analytics with Digital Twin simulations, networks can reduce downtime, minimize service disruptions, and improve operational resilience [29].

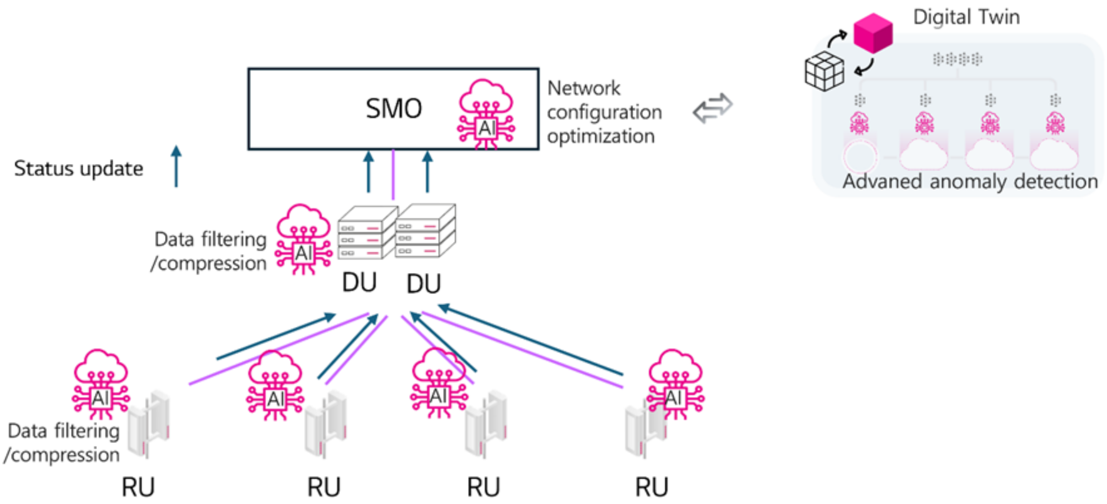

Figure 3-5 Network management in scalable and user-centric RAN architecture.

4 Conclusion

This report has delineated the key elements and design approaches central to achieving a scalable and user-centric RAN architecture for next-generation mobile networks. By embracing cloud-native paradigms and virtualization, the report illustrates how the disaggregation of traditional baseband processing can be transformed into a flexible, adaptive framework. The strategic placement of O-DUs and O-CUs enables dynamic resource allocation that can respond to variable traffic conditions and diverse service demands. Such an architecture not only addresses the complexities of fluctuating network loads but also facilitates more efficient, energy-conscious operations. In particular, the report has surveyed deployment options, such as C-RAN and D-RAN variants and edge- versus central-cloud placements, to enable the practical deployment of a scalable, user-centric RAN.

Furthermore, the integration of advanced hardware accelerators, such as GPUs and FPGAs, combined with AI/ML-driven orchestration techniques, enhances network performance while promoting energy efficiency. The exploration of various AI integration models, including AI-for-RAN, RAN-for-AI, and CCIN, demonstrates the potential for embedding intelligence into both network control and data processing layers. This incorporation of AI provides a pathway to real-time network adaptation and intelligent resource management, which is essential for optimizing performance and maintaining resilience in ever-evolving operational environments.

Overall, the report establishes a comprehensive conceptual foundation that bridges traditional RAN implementations with emerging digital paradigms. By highlighting the importance of modularity, dynamic management, and integrated intelligence, the proposed framework offers guidance for future technical developments. As the mobile networking landscape continues to evolve, the insights presented here serve as a roadmap for developing more agile, efficient, and user-centric RAN solutions that can meet the demands of tomorrow’s connected world.

References

[1] Q. Li, Z. Ding, X. Tong, G. Wu, S. Stojanovski, T. Luetzenkirchen, A. Kolekar, S. Bangolae and S. Palat, “6G cloud-native system: Vision, challenges, architecture framework and enabling technologies,” *IEEE Access,* vol. 10, pp. 96602-96625, 2002.

[2] One6G, “6G and Robotics: A Methodology to Identify Potential Service Requirements for 6G-empowered Robotic Use Cases,” One6G, Zurich, 2024.

[3] ITU-R, “Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond,” R-REP-M.2516, 2022.

[4] M. Klinkowski, “Optimized Planning of DU/CU Placement and Flow Routing in 5G Packet Xhaul Networks,” *IEEE Transactions on Network and Service Management,* vol. 21, no. 1, pp. 232-248, 2024.

[5] A. Ikami, A. Amrallah and H. Yang, “Research report on Use Case and Gap Analysis for Radio Quality Assurance,” O-RAN Alliance, nGRG, 2024.

[6] Hexa-X, “Architecture for B5G/6G networks,” Hexa-X, 2023. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2023/07/Hexa-X-D1.4-Final.pdf.

[7] Next G Alliance, “6G Technologies for Wide-Area Cloud Evolution,” Next G Alliance, [Online]. Available: https://nextgalliance.org/white_papers/6g-technologies-for-wide-area-cloud-evolution/.

[8] Beyond 5G Promotion Consortium, “Beyond 5G White Paper ～Message to the 2030s～,” Ver. 2, 2023. [Online]. Available: https://b5g.jp/en/output/.

[9] O-RAN.WG1.CCIN-R004-v01.00, “Communication and Computing Integrated Networks,” O-RAN Alliance, Alfter, 2024.

[10] Intel, “Reducing Energy Use and Carbon Footprint of Open RAN Networks,” 2023. [Online]. Available: https://builders.intel.com/solutionslibrary/reducing-energy-use-and-carbon-footprint-of-open-ran-networks.

[11] “AI-enabled computing for O-RAN increases utilization, reduces power consumption and cost,” Fujitsu, [Online]. Available: https://networkblog.global.fujitsu.com/2023/02/22/ai-enabled-computing-for-o-ran-increases-utilization-reduces-power-consumption-and-cost/.

[12] AI-RAN Alliance, “AI-RAN Alliance Vision and Mission White paper,” 2025. [Online]. Available: https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf.

[13] L. Kundu, X. Lin, R. Gadiyar, J.-F. Lacasse and S. Chowdhury, “AI-RAN: Transforming RAN with AI-driven Computing Infrastructure,” *arXiv preprint arXiv:2501.09007,* 2025.

[14] Fujitsu, “Leveraging AI-RAN to transform the future of Radio Access Networks,” 2025. [Online]. Available: https://networkblog.global.fujitsu.com/2025/02/03/leveraging-ai-ran-to-transform-the-future-of-radio-access-networks/.

[15] Y. Xiao and M. Krunz, “Dynamic Network Slicing for Scalable Fog Computing Systems With Energy Harvesting,” *IEEE Journal on Selected Areas in Communications,* vol. 36, no. 12, pp. 2640-2654, 2018.

[16] L. Kundu, X. Lin, E. Agostini, V. Ditya and T. Martin, “Hardware acceleration for open radio access networks: A contemporary overview,” *IEEE Communications Magazine,* vol. 62, no. 9, pp. 160-167, 2023.

[17] Ericsson, “Cloud RAN Acceleration Technology,” 2022. [Online]. Available: https://www.ericsson.com/4ae403/assets/local/ran/doc/cloud-ran-acceleration-technology-positioning-paper.pdf.

[18] Cisco, “Understand NFVIS Virtual Networks: OVS, DPDK and SR-IOV,” 13 Feb 2024. [Online]. Available: https://www.cisco.com/c/en/us/support/docs/routers/enterprise-nfv-infrastructure-software/221679-understand-nfvis-virtual-networks-ovs.html. [Accessed 15 Apr 2025].

[19] 3GPP, “Coordinated multi-point operation for LTE physical layer aspects,” TR36.819 V11.20, 2013.

[20] 3GPP, “Enhancements on MIMO for NR,” 3GPP TSG RAN Meeting #87e RP-200474, 2020.

[21] S. Venkatesan, A. Lozano and R. Valenzuela, “Network MIMO: Overcoming Intercell Interference in Indoor Wireless Systems,” in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, 2007.

[22] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, “Cell-Free Massive MIMO Versus Small Cells,” *IEEE Transactions on Wireless Communications,* vol. 16, no. 3, pp. 1834-1850, 2017.

[23] O. Haliloglu et al., “Distributed MIMO Systems for 6G,” in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2023.

[24] E. Björnson and L. Sanguinetti, “Scalable Cell-Free Massive MIMO Systems,” *IEEE Transactions on Communications,* vol. 68, no. 7, pp. 4247-4261, 2020.

[25] G. Interdonato, P. Frenger and E. G. Larsson, “Scalability Aspects of Cell-Free Massive MIMO,” in *2019 IEEE International Conference on Communications (ICC)*, 2019.

[26] J. Zhao, “Decentralised Distributed Massive MIMO,” PhD thesis, University of York, 2023. [Online]. Available: https://etheses.whiterose.ac.uk/id/eprint/34161/. [Accessed 22 May 2024].

[27] NVIDIA, “NVIDIA Multi-Instance GPU User Guide,” 31 Mar 2025. [Online]. Available: https://docs.nvidia.com/datacenter/tesla/mig-user-guide/. [Accessed 15 Apr 2025].

[28] P. Almasan et al., “Network Digital Twin: Context, Enabling Technologies, and Opportunities,” *IEEE Communications Magazine,* vol. 60, no. 11, pp. 22-27, 2022.

[29] ITU-R, “Future technology trends of terrestrial international mobile telecommunications systems towards 2023 and beyond,” International Telecommunications Union, Radiocommunication Sector (ITU-R), Geneva, 2022.

[30] O-RAN.WG6.CADS-v08.00 TR, “O-RAN Working Group 6, Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN,” O-RAN Alliance, 2024.