Path towards scaling Open RAN Architecture

Joint whitepaper by

Vodafone and NTT DOCOMO

C1 Public

Executive Summary

This white paper outlines the diverse deployment scenarios supported by Open RAN. The benefits and challenges associated with these scenarios can vary significantly depending on the environments and conditions in which Mobile Network Operators (MNOs) operate. Therefore, MNOs must actively explore these considerations to identify the most suitable network solutions tailored to their specific needs.

The deployment scenarios discussed include not only the distributed system architectures traditionally prevalent in RAN networks but also centralised architectures that leverage virtualisation and cloud technologies. Moreover, MNOs should carefully consider the extent of centralisation, as the flexibility to distribute or centralise RAN functions presents opportunities for effectively leveraging Open RAN.

In this paper, we introduce typical Open RAN deployment scenarios and detail the associated advantages, challenges, and considerations for each. It is our hope that this white paper serves as a valuable resource in your evaluation process and supports MNOs in accelerating the adoption of Open RAN within the context of their surrounding network landscapes.

Once MNOs have grasped the fundamental elements presented in this document, we encourage them to engage with the Open RAN community particularly as global Open RAN implementation cases continue to emerge. Building an ecosystem and identifying optimal solutions will be crucial for Open RAN's success. This paper aims to serve as a catalyst for promoting a collaborative spirit among operators, fostering innovation in 5G and beyond.

C1 Public

Table of Contents

C1 Public

1. Introduction

1.1. Industry Drivers for Open RAN and Challenges for Operators

The vision for Open RAN is to create a disaggregated open platform to drive innovation, competition and diversity in supply markets by leveraging the expansive IT cloud computing ecosystem to develop best-in-class RAN products. While traditional RAN, LTE and 5G, have adopted a distributed system architecture, cloud-based RAN platforms enable new opportunities. One example supported by O-RAN ALLIANCE standards would be a fully centralised RAN, but it requires low-latency transport network. High dark fibre costs are challenging in many markets where an intermediate step aggregating Centralised Unit (CU) functions is also possible.

The architecture of traditional LTE and 5G base station networks has been driven by the needs of customers and services. Distributed RAN simplifies low latency communication between network functions and maximise performance without network bottlenecks. On the other hand, cloud computing platforms create opportunities to improve capital and operational costs through increased supplier competition, acting as a catalyst for innovations such as energy efficiency and software operations.

Open RAN offers flexibility to distribute or centralise RAN functions close to the end user at cell site, edge, regional or central data centre locations create opportunities for Open RAN. For example, more centralisation creates synergies with central and regional cloud platforms for the core network and Multi-access Edge Computing (MEC) and Artificial Intelligence (AI) as part of an end-to-end cloud strategy.

By identifying industry drivers and avoiding fragmented Open RAN solutions, Open RAN can provide scale benefits to realize lower costs for the industry while ensuring supply market diversity and promoting innovation.

1.2. Purpose of This Paper

The scope of this technical report is to survey and explore architectural options and methods with the purpose to inform future Open RAN deployments of potential for cost reduction and innovation.

Open RAN deployment scenarios are diverse, and the benefits or challenges associated with these solutions can vary based on the environment and conditions of the networks operated by MNOs. MNOs need to actively explore these considerations to find the most suitable network solution, and this document is intended to serve as a useful resource in that process.

C1 Public

2. Definitions of Open RAN Architecture

Before going into the detail, this chapter introduces typical Open RAN deployment architectures. Although different configurations may arise in the future as technology evolves, the configuration illustrated in Figure 1 are generally considered feasible according to current technology trends.

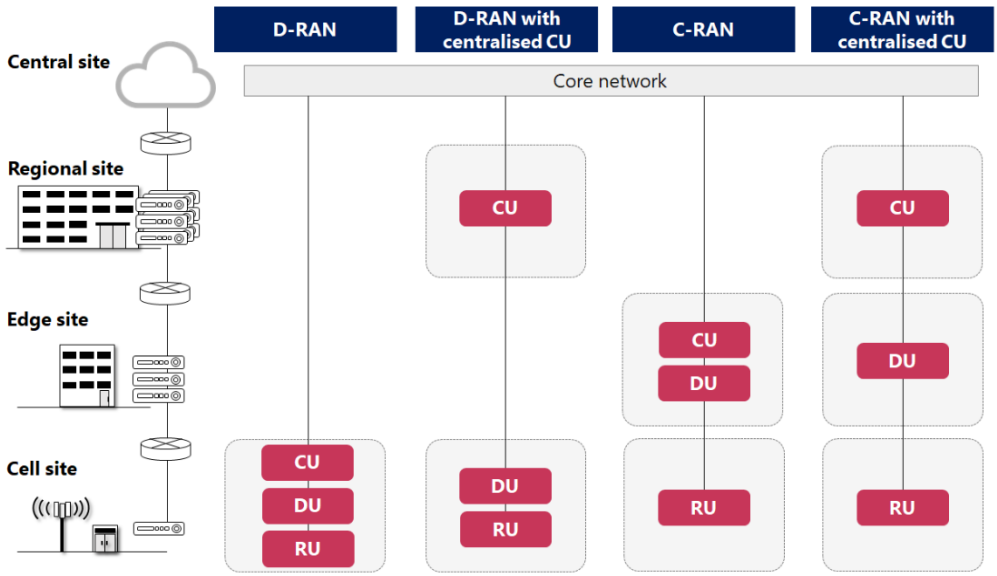

Figure 1 Open RAN deployment architectures

In D-RAN, the CU and DU are deployed in the same cell site as the RU. Typical cell sites are towers or common buildings where the CU and DU are located in shelters or indoor spaces. Compared to more controlled environments such as data centres, servers needs to be optimised for this environment - such as supporting NEBS3 (Network Equipment-Building System) compliance, and may have constraints in terms of capacity and security.

With the introduction of virtualised RAN, the operation of disaggregated CU and DU in different locations has become easier, promoting the adoption of CU and/or DU aggregation. Figure 1 illustrates three centralised architectures based on the location of the CU and DU. The edge site represents potential DU aggregation sites, typically small-scale data centres situated within 30 kilometres from the cell site to guarantee a fronthaul transmission delay of 150 microseconds. The regional site represents potential CU aggregation sites located within a few hundred kilometres from the cell site. These regional sites are typically larger data centres than those at edge sites.

In D-RAN with centralised CU, the DU and RU are deployed at the cell site, and the CU is centralised and located at the regional site.

C1 Public

In C-RAN, the RU is deployed at the cell site, and the CU and DU are centralised and located at the edge site.

In C-RAN with centralised CU, the RU is deployed at the cell site, the DU is centralised and located at the edge site, and the CU is further centralised and located at the regional site.

This paper highlights key considerations from various perspectives, with a primary focus on D-RAN, D-RAN with centralised CU and C-RAN architectures which are commonly used configurations.

C1 Public

3. Conditions and Considerations for Each Deployment Scenarios

This section provides the main technical considerations within the RAN Macro Network domain to be considered when selecting an architecture option. Topics related to the impact of the transport domain are identified in the context of overall networking architecture and are expected to be a subject of future white papers.

3.1. Dimensioning

The introduction of virtualised RAN revolutionises dimensioning strategies, as it enables best in class compute servers and the pooling of computational resources across multiple of cells. This allows for a dynamic adjustment of resource allocation based on each cell's traffic trends, thereby enhancing efficiency and effectiveness in dimensioning. The architectural placement of DU and CU functions at cell sites, edge, or data centres creates differences in the ability to pool and scale resources. For example, centralising functions and/or more powerful compute can significantly reduce or eliminate cell site compute load, leading to a site cost reduction or potential deferral of future capacity investments. Therefore, it is crucial to recognise that virtualised RAN plays a pivotal role in optimizing resource management within the network.
A key concept of RAN dimensioning of hardware resources is based on the principle of pooling gain or statistical multiplexing gains (SMG). In cellular networks, SMG exploits the spatial and temporal distribution of downlink and uplink user traffic allowing efficiency in compute resources [Zhang, 2019]. Traditionally, resources corresponding to the total of the maximum traffic values for each cell in terms of CU and DU are required. However, since the busy hours for each cell are different, centralising network functions may allow for the realization of several tens of percent of SMG. Determination of SMG is specific to the type of architecture, network technology and traffic design where compute resources are typically dimensioned based upon busy-hour traffic statistics with market forecasts.

$$\mathrm{SMG} = \frac{\text{Processing resources in D}-\text{RAN}}{\text{Processing resources in C}-\text{RAN}}$$

Furthermore, Control and User Plane Separation (CUPS) can be applied to packet-based CU functions with benefits of independent scaling of control plane (CU-CP) and user plane (CU-UP) functions.

To assess DU and CU hardware costs and savings for different Open RAN architectures, the following design inputs are required and listed:

- Architecture: Which technologies (GSM, UMTS, LTE or 5G NR) will adopt centralising of functions such as DU or CU.

C1 Public

- Software Efficiency: Dimensioning compute resources for RAN workloads are bound by the capacity and capabilities of both hardware and software. Efficient software baseband designs, which utilize hardware accelerators, can significantly increase capacity.
- Number of cells: Clarification of the number of cells per target hardware DU and CU server units, along with the quantification of baseband compute resources based on Radio Unit and backhaul connectivity requirements.
- Traffic parameters: Targets for downlink and uplink user and cell throughputs, control plane signalling traffic load, and associated forecasts. Independent dimensioning of CU-CP and CU-UP should be applied.
- Redundancy: Resilience support of local and geographic redundancy of CU functions.
- Dimensioning of current and future network features and capabilities: RAN sharing, Uplink Coordinated Multi-Point (ULCOMP), User Plane Integrity Protection (UPIP), network slicing, advanced security, and more.

Finally, dimensioning compute resources for Open RAN requires new skills, processes and responsibilities to design and dimension compared to traditional RAN.

3.2. Operations and Lifecycle Management

Reducing OPEX is a common concern among MNOs. It is worth considering benefits and efficiency in the context of operation and maintenance throughout the lifecycle of RAN for each deployment scenario. In a D-RAN scenario, the large number of servers distributed across numerous cell sites may increase the number of site maintenance visits compared to a centralised CU. However, with centralised CU or DU, server clusters in fewer data centers than cell sites are expected to reduce the number of sites to be visited, thereby reducing the workforce required for unexpected maintenance tasks.

Optimal use of the resource is also essential. In a D-RAN case, it would be sufficient to deploy servers that can serve the users within the cell site's coverage. However, deploying a full set of CU, DU and RU even in areas with low traffic demand could result in a cell site operating with surplus resources. Additionally, the changes in traffic over time could lead to resource shortcomings or excesses. On the other hand, centralised CU and DU with virtualised RAN allows resource pooling, which facilitate flexible automatic and manual scaling as well as healing. The resource pools that serve a wider area across multiple cell sites make it possible to flexibly scale resources and take recovery actions based on traffic demands and fault conditions. For virtualised DU that processes with a hardware accelerator, additional steps would be required when scaling in, scaling out and healing compared to virtualised CU, because accelerator

C1 Public

settings, such as RU address and fronthaul switch information, need to be configured in addition to deploying the DU application. Additionally, existing active sessions cannot be transferred to another CU or DU pod without disruption. Scaling out can be triggered based on an anticipated increase in session numbers, allowing new sessions to be allocated to the newly added pods, while scaling in can occur when a decrease in session numbers is forecasted. The ability to auto-scale depends on the product.

During the construction phase, centralisation allows for the procurement, design and construction of hardware for serving a wide area to be done at once. This consequently cuts down on the number of configuration tasks and tests, reducing the lead time to service provision.

In the case of D-RAN with centralised CU, the benefits of centralisation could be limited, because visits to cell sites for DU maintenance are still necessary. In addition, transitioning to a centralised CU architecture can impact the procedures related to site planning, design, and operation of physical infrastructure implications, potentially introducing complexity.

3.3. Mobility Performance

In the case of D-RAN deployments a single stack for both DU and CU within a server instance can offer low latency services executed locally minimising control plane delay particularly for intra-site cell handovers.

Centralising CU-CP and CU-UP functions allows the aggregation of DUs under a single gNB ID instance, creating what we call a super gNB or cluster. Whilst intra-gNB handovers will be moderately extended, conversely all inter-gNB handovers evolve to become an intra-gNB handovers within the cluster to allow faster handovers and reduces signalling connection loads towards the core network for more mobile users. D-RAN with centralised CU and C-RAN also avoid tromboning effects, X2 packet forwarding from source to target gNB via security gateway, on transport at handovers and dual connectivity mobility use cases. For D-RAN, direct Xn/X2 IPSEC which avoids routing via security gateway can mitigate this effect.

C1 Public

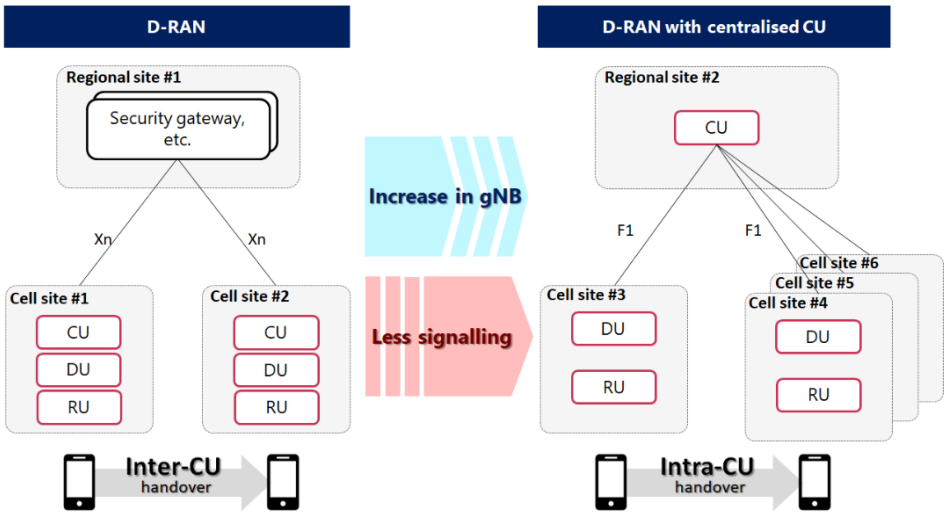

Figure 2. D-RAN with centralised CU inter and intra mobility use case

The placement of DU and CU functions at different locations beyond the cell site can impact the control plane signalling. Mobility procedures between the DU and CU can contribute to prolonged latency dependent on the number of call flow transactions. If the distance between the CU and DU is 1000 km over the F1 interface, the one-way latency would be approximately 5 milliseconds, would minimise the aggregate increase in mobility procedures as shown in Table 1. For example, there is no noticeable impact on end user experience considering the average reaction time to audio or visual stimuli is between 200-250 milliseconds [Jain,2015]. The table below summarises the performance impact of different architectures on mobility use cases based upon technical specifications [ORAN, 2024] and [3GPP,2019]. Centralising CU will impact control plane signalling latency but with no impact expected to customer user traffic, network KPIs or user experience in general.

Table 1. Latency and performance impact of D-RAN with centralised CU RAN

	Procedure	Additional F1 delay* by centralizing CUs	End user impact
C-plane	Service Request	35 ms	Minor impact
	Intra-gNB handover	30 ms	
	Inter-gNB handover	20 ms	
	Inactive to Connected	20 ms	
U-plane	Uplink/Downlink Throughput	No user experience impact	

* The number of transactions may vary depending on the radio environment & configuration, which could also lead to differences in delay values.

C1 Public

Centralised CU with hybrid of 5G centralised CU and LTE D-RAN architectures leads to mobility exceptions with LTE and 5G NR interworking. Figure 3 illustrates an example of 5G Non-standalone (NSA) E-UTRA-NR Dual Connectivity (EN-DC) scenario where the 5G NR CU-CP and CU-UP functions are centralized, while 5G NR DU and LTE eNB are located at the cell site. In this scenario, the Master Cell Group (MCG) radio resources of LTE are processed by PDCP (Packet Data Convergence Protocol) in 5G CU. Therefore, when centralising 5G CU, the potential increase in X2 latency between 5G CU and eNB needs to be considered. From the perspective of interaction with the core network, whilst S1-C is managed by LTE eNB, S1-U is managed by 5G CU for EN-DC cases. Thus, the impact of potentially different latencies on the control plane and user plane should also be taken into account.

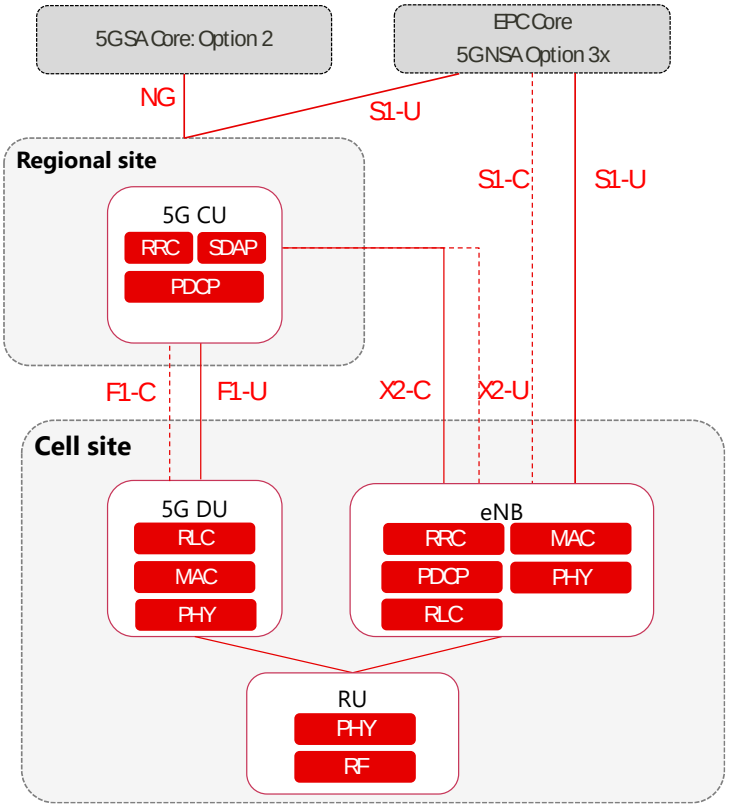

Figure 3. Example of 5G NSA with D-RAN with centralised CU architecture

For centralised CU and C-RAN, centralising network functions may improve performance through greater co-operative radio techniques and simplification of mobility use cases. When CU is centralised, packets are buffered at the PDCP layer which may lead to improved performance during mobility. However, it is essential to consider the fluctuations caused by delays between the CU and DU in the transport network. In scenarios where multiple technologies coexist and different architectures are adopted for each technology, there may be performance effects from the transport

C1 Public

topology. Ultimately, MNOs must determine the optimal scenario based on the configuration of their network.

3.4. High Availability

When considering the architecture, it is important to assess the impact of failure and consider how to achieve high availability. In a D-RAN, the impact of a single CU or DU failure is limited to the service area of each cell site. On the other hand, centralised CU or DU may become a single point of failure that affects larger number of cells than in a D-RAN case.

To minimize the impact of centralisation, measures need to be taken to enhance resilience, such as implementing CU redundancy locally within a data centre or across geographically separated data centres. These measures can help to ensure the continuity of services. In addition, it might be necessary to intentionally restrict the number of users per CU or per DU in the design phase to mitigate the impact. Furthermore, in areas where services are offered on multiple frequency bands, one option to enhance overall availability of the area is to assign cells operating on different frequency bands to different CUs and DUs hosted on different servers.

When it comes to minimizing downtime, the fact that COTS (Commercial Off-The-Shelf) servers in C-RAN and D-RAN with centralised CU are pooled as O-Cloud resources provides a benefit that D-RAN doesn't have. With pooled resources, a virtualised CU is not bound to a specific server and can be created anywhere, as long as transport latency conditions are met. This enables the re-creation of virtualised CU and service recovery in the event of failure. It should be noted that, when re-instantiating a virtualised DU on another server for service recovery, reconfiguration of the accelerator may be necessary, and there may be cases where the accelerator settings cannot be migrated to another server using deployment features of container orchestration tools. While additional standby servers are necessary at data centres for failover purposes, it is more resource-efficient than having redundant servers at every cell site or higher availability single servers. Furthermore, data centres generally offer more controlled facilities and access to a stable power supply compared to a cell site, making them more resilient to disasters such as earthquakes and better suited for deploying spare resources.

3.5. Security

An architectural topic is to understand the impact of security controls upon compute workloads. In a D-RAN or C-RAN architecture, the DU and CU are co-located at the cell or edge site. This architecture enables cell or edge processing of all gNB functions in one location. This will mitigate the security requirements of the F1 interface as both are processed usually within the same hardware.

C1 Public

In the centralised CU architecture, CUs are deployed in regional or central data centres, which are protected by robust buildings, thereby reducing risks compared to distributed sites. On the other hand, the physical separation between the CUs and DUs necessitates addressing potential such as data tampering due to man-in-the-middle attacks.

3GPP TS 38.323 Release 16 introduced full rate integrity protection requirement for user plane, which protects the integrity of the user plane data between the UE and the gNB in 5G SA (Standalone) and NR-Dual Connectivity deployments.

The integrity protection feature for 5G is processed within the PDCP stack which is part of CU and requires significant computational resources, that both RAN and UE vendors must accommodate, while the use of hardware cryptographic security accelerators for the integrity protection can free up CPU capacity.

A centralised CU architecture may have a favourable compatibility with security accelerators. This is primarily because, in the event that a security algorithm needs to be changed, it is generally easier to swap devices and implement processing within a data centre compared to a cell site.

Furthermore, if there is a need to implement more powerful encryption algorithms, which may require double the computational power, the addition of such equipment can be performed more flexibly in a data centre compared to a cell site. This flexibility enhances the ability to adapt to evolving security requirements efficiently.

Attention to the control plane (F1-C) is essential, in addition to user plane considerations. The link between the DU and the CU could require protective measures, such as IPsec or DTLS.

Overall, understanding the relationship between security controls and compute workloads is crucial in both D-RAN and centralised CU architectures. The introduction of integrity protection described in this section highlights the need for robust computational resources to ensure the security of communications. The integration of cryptographic accelerators, along with the flexibility to quickly adapt to update security requirements, enhances the resilience of next-generation networks against security risks.

3.6. Hardware Evolution

3.6.1. Silicon architecture

The Open RAN deployment scenario emphasizes the importance of hardware accelerators to enhance server processing performance and capacity. This allows vDUs to process received data quickly and minimize latency. Moreover, accelerators can

C1 Public

optimize communication between DUs and RUs, improving network efficiency. In short, accelerators can play a role in efficiently deploying RAN by enhancing server capacity.

The use of specialised compute or hardware accelerators can offload processing from applications running on a General-Purpose Processor (GPP) to importantly support 5G Massive MIMO or other advanced RAN technologies and provide efficiencies for RAN workloads. Hardware accelerators are available in an array of technologies and combinations, for example:

1. Field programmable gate arrays (FPGA),
2. Fixed function application specific integrated circuits (ASIC)
3. Specialised digital signal processors (DSP) and system on a chip (SoC)
4. Graphic processing units (GPU)

Thus, different types of accelerators are available in the market today, but in this paper, we focus on the two main types: Inline and Look-aside. Inline accelerators are isolated with acceleration functions that process all High-PHY layer functionalities. On the other hand, Look-aside accelerators are integrated into CPUs or PCIes, processing some of the High-PHY functions while leaving the rest to the CPU.

Each type has key considerations in various aspects. Regarding cell capacity and scalability, Inline accelerators can support a large number of cells and offload all High-PHY processing to reduce CPU workload. Multiple accelerators can be installed in a server. In terms of energy efficiency, Look-aside accelerators have lower total power consumption compared to Inline accelerators, as they don't require a separate accelerator card. In terms of integration, both types require collaboration between vDU software vendors, CPU vendors, and accelerator vendors. However, the complexity of this collaboration depends on the allocation of L1 functions between vDU and accelerator.

Considering the pros and cons of each type, MNOs need to have the ability to select best-in-class silicon products based on the deployment scenario. Although it may be a challenging task, it presents an opportunity to maximize network performance, flexibility, and energy efficiency.

3.6.2. Hardware requirements for cell sites and outdoor installation in a D-RAN environment

In a D-RAN scenario, the CU, DU and RU are placed at the cell sites. With the introduction of the virtualized Open RAN, there is a potential increase in components to be installed compared to the dedicated devices used in traditional RAN. It is essential to carefully consider the optimal placement of these components. Figure 4 illustrates examples of cell sites.

C1 Public

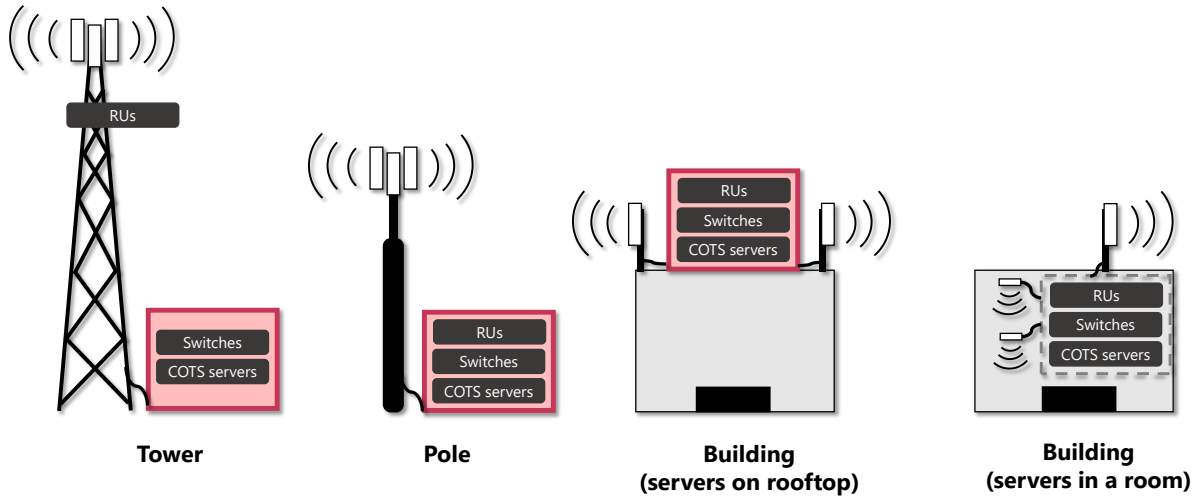

Figure 4 Examples of cell sites

Cell sites, for example, face constraints regarding installation space and power capacity. Battery backups must also be considered to ensure uninterrupted operation. Furthermore, the deployment of COTS servers, conventionally housed in indoor data centres, to areas susceptible to outdoor influences introduces new requirements; typically, the European GR-63-CORE and GR-1089-CORE or the US NEBS3 compliant COTS hardware. The environment at cell sites can be harsh, posing various challenges such as high and low temperature, condensation due to temperature changes, humidity, water, dust, vibrations, and damage from salt.

In response to these challenges, MNOs face the task of selecting not only appropriate RAN components but also ensuring suitable protective housing and temperature management solutions. Additionally, measures such as minimizing noise from cooling fans can be necessary to avoid any negative impact on surroundings. By addressing these considerations, MNOs can achieve a more effective and community-friendly implementation of virtualised Open RAN at cell sites.

C1 Public

4. Conclusion

This paper has examined the benefits of open platforms in supporting various architectural options, ranging from distributed to centralised RAN. Open RAN offers several advantages, including the separation of hardware and software, which enhances vendor diversity, fosters innovation, improves energy efficiency, and facilitates software operationalisation. This separation makes it easier to transition to next-generation platforms, resulting in better computing performance. However, the extent of these benefits can vary depending on the environments and conditions of the networks operated by MNOs.

From an architectural perspective, D-RAN is advantageous for simplifying operations and transport, while C-RAN and centralised CU excels in scalability, centralised security management, and efficient resource utilisation. The choice depends on the specific needs for latency, security, and scalability in the network deployment. MNOs must continuously and actively consider these factors, along with economic and operational considerations, to identify the most suitable network solutions.

Additionally, it is advisable for MNOs to engage on Open RAN in discussions. Building an ecosystem and identifying optimal solutions will be crucial for success. This collaborative approach not only enhances competitiveness but also drives innovation in evolving telecommunications infrastructure with Open RAN.

Ultimately, Open RAN provides the flexibility and agility to adapt functions and network architectures, enabling mobile operators to deliver the best possible service to their customers.

C1 Public

References

[Zhang, 2019] Z. Zhang et al., "Statistical Multiplexing Gain Analysis of Processing Resources in Centralized Radio Access Networks," in IEEE Access, vol. 7, pp. 23343-23353, 2019

[Jain, 2015] Jain A., Bansal R., Kumar A., Singh K.D. A Comparative Study of Visual and Auditory Reaction Times on the Basis of Gender and Physical Activity Levels of Medical First Year Students. Int. J. Appl. Basic Med. Res. 2015;5:124–127

[ORAN, 2024] O-RAN Alliance, 2024. TS O-RAN Open F1/W1/E1/X2/Xn Interfaces Working Group, NR C-plane profile. O-RAN.WG5.C.1-R004-v13.00

[3GPP, 2019] TS 38.401 version 15.7.0 Release 15 - 5G; NG-RAN; Architecture description. 3rd Generation Partnership Project (3GPP)

C1 Public