WHITEPAPER

SCALING AND OPTIMIZING OPEN RAN NETWORKS

Improving Cost-Effectiveness and Time-to-Market in Global Telco Deployments

TABLE OF CONTENTS

Executive Summary

The Open RAN landscape has matured significantly, reaching a pivotal stage where Tier 1 Mobile Network Operators (MNOs) are poised for large-scale deployment. This transition is underscored by major endorsements from industry giants like AT&T and Vodafone, reflecting the growing confidence in Open RAN technology. While initially led by greenfield operators, mainstream adoption has expanded to include legacy Tier 1 suppliers, showcasing the technology's broader applicability.

From a technological standpoint, Intel® played a crucial role in establishing the viability of 5G base stations on general-purpose processors. However, the ecosystem has diversified, with alternatives from AMD, ARM, and NVIDIA offering a range of solutions. The benefits of Open RAN and virtualization are driving industry objectives, promoting innovation, supplier diversity, and cost-effectiveness. The shift to cloud-native 5G architectures, while presenting challenges, brings scalability, agility, and significant operational benefits.

As the industry achieves performance parity with legacy solutions, attention turns to energy efficiency, dynamic workload management, and the ability to adapt to changing network traffic loads. Supermicro's optimized systems, ranging from Cloud RAN servers to cell site servers, enable operators to streamline Open RAN deployments with increased efficiency, scalability, and a lower total cost of ownership (TCO).

State of the Open RAN Ecosystem

Following a lengthy and tortuous gestation, Open RAN is considered ready for scale deployment by Tier 1 MNOs. This achievement was validated by AT&T's announcement and endorsement towards the end of 2023 and Vodafone's 2023 announcement of its intent to run a global RFP with around 30,000 cell sites intended to be Open RAN. Verizon has also reported having over 15,000 virtualized Cloud RAN sites, a close cousin of Open RAN. Notably, all these also involve legacy Tier 1 incumbent suppliers. While Open RAN reporting has tended to focus on the US and European markets, Japan has been a pioneer in Open RAN, with NTT DoCoMo and KDDI having early interoperable deployments.

Prior to this, greenfield operators, namely Rakuten and Dish Networks, have led the Open RAN charge. The pioneering work was done by new supplier entrants—Mavenir, Parallel Wireless, and Altiostar (later acquired by Rakuten)—although the primary standards steering body, the O-RAN Alliance, itself a merger of the xRAN Forum with the C-RAN Alliance, has always had heavy representations by industry stalwarts like Ericsson. The intent is for output from the O-RAN Alliance, as an operator-led body, to feed into the 3GPP standards organization.

From a technology perspective, Intel was the clear proponent and standards bearer. Intel advocated strongly that a 5G base station could be implemented in virtualized software running on standard general-purpose processors (GPPs). Here, a GPP is defined as an industry-standard, generally available processor that is also used for general computing applications. While meeting the required performance was clearly a stretch at the outset, thanks to some optimizations, initial hardware assists, and the inevitable performance gains from the march of Moore's law, we are at a point where GPPs can quickly implement 4G and lower-order 5G with acceleration and on the brink of supporting full Massive MIMO. Intel implemented a processor roadmap with the requisite enhancement that delivered the required performance. Intel also developed their FlexRAN™ software stack as a usable reference to implement Open RAN DU (Distributed Unit).

A key criticism of Open RAN has been that Intel was the only silicon supplier. Alternate COTS (commercial off-the-shelf) GPP processors for Open RAN are now available from AMD, ARM, and NVIDIA’s GPU-based solutions. One of the key optimizations was to add matrix vector processing. Intel has this with AVX-512, which AMD eventually also added. ARM has followed with its own version, the Scalable Vector Extension, or SVE2. GPUs that NVIDIA uses are inherently vector-processing capable.

Another contentious accelerator is one for FEC encoding, as this is the most compute-intensive processing task requiring hardware assist, especially for massive MIMO. This need led to an industry-wide debate about lookaside versus inline accelerators. Lookaside accelerators, as developed by Intel, accelerate just the FEC function, leaving the rest of the L1 processing to the CPU. In contrast, inline accelerators typically perform the complete Layer 1 DU functionality and are not perceived as COTS General Purpose Processors. In servers with the requisite PCIe expansion slots, inline processors can be mated with a GPP for Layer 2. The latest 5th Gen Intel® Xeon® processors with Intel vRAN Boost have integrated dedicated hardware FEC encoding acceleration within the CPU. Looking ahead, it is likely that either (1) FEC encoding can be done on the processor itself, as processors get more cores, or (2) a hardware abstraction driver will be developed, allowing software to call on different hardware accelerator implementations, even those from a different chipset vendor, in a standardized way.

In short, we are entering a phase where GPP performance has been validated as adequate for 5G DU implementation. Solutions are now available from multiple chipset suppliers for both GPPs (Intel, AMD, NVIDIA, and ARM) and Inline (Qualcomm, Marvell, NXP, and EdgeQ), enabling a broad ecosystem of hardware and software solution providers.

From an operator's perspective, the emphasis is now on cost, performance, and power efficiency, as these are all CapEx and OpEx drivers. There is a particular focus on energy consumption and the ability to dynamically match power consumption with network traffic loading.

Supermicro is uniquely positioned in that it has server offerings with all of the aforementioned chipset vendors in a variety of platform configurations, which can be optimally selected based on the operator's deployment scenario. All products have PCIe expansion slots and can optionally be provisioned with inline or other accelerators, such as AI modules for radio management and MEC applications. Supermicro is also an active participant in the O-RAN Alliance and an editor of WG7: The White-box Hardware Work Group.

Open RAN Recap

Before diving deeper into products and applications, it is worth a short recap on Open RAN and industry objectives.

Open RAN is defined by the O-RAN Alliance in Figure 1. The key is that the RAN had been segregated into discrete functional blocks (RIC, CU, DU, RU, AU) with industry standard interfaces between them. An often-understated change in Open RAN is separating the RIC (RAN Intelligent Controller) into its own entity with support for 3rd party apps that will allow for innovation in RAN resource management.

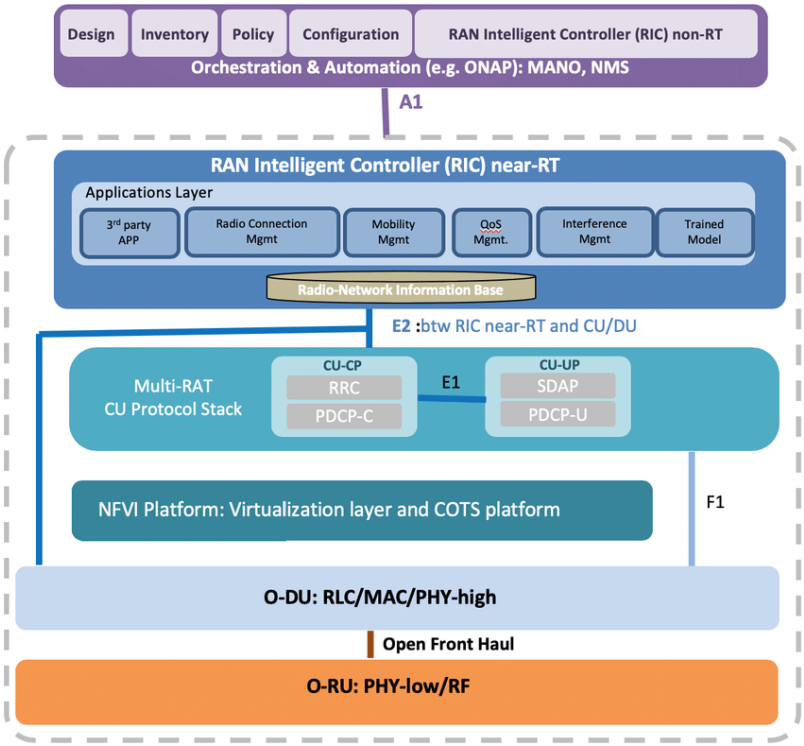

Figure 1 - O-RAN architecture. Image used courtesy of the O-RAN Alliance

One of the most critical interfaces is the front-haul interface and the split between DU and RU. The industry had defined various splits and primarily settled on 7.2x. However, Ericsson raised objections related to massive MIMO CAT-B radio uplink performance and is a proponent of putting the equalizer and channel estimation in the Radio Unit. Recently, an agreement was reached within the O-RAN Alliance for a Class A and Class B option that addresses the issue, and with that change, the industry now has full buy-in from the major system suppliers.

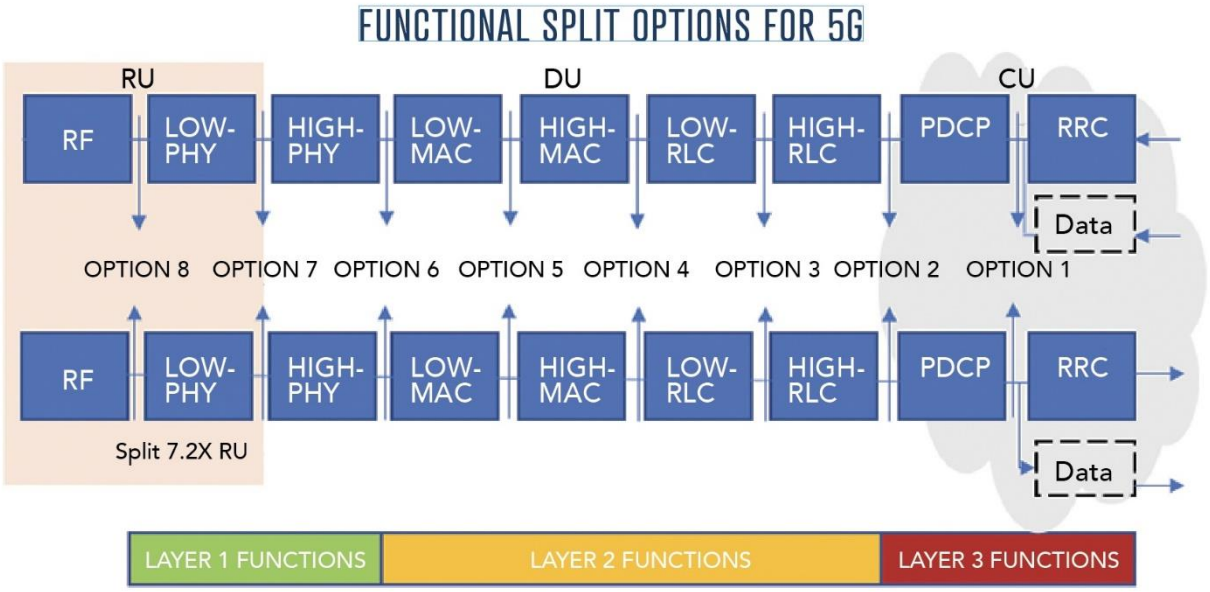

Figure 2- Open RAN Functional Splits

RAN "Openness"

When discussing Open RAN, various terms like Open RAN, OpenRAN, O-RAN are used. In general, Open RAN is the movement in wireless telecommunications to disaggregate hardware and software and to create open interfaces between them. The intention is to avoid vendor lock-in, diversify supply, and create more innovation.

Per Figure 3, there are two dimensions to Open RAN, leading to four quadrants,

The first openness is horizontal disaggregation, as defined by the O-RAN Alliance. Here, the NodeB (radio base station) is disaggregated into separate CU (Centralized Unit), DU (Distributed Unit), RU (Radio Unit), and RIC (RAN Intelligent Controller) Network Functions (NF) with standardized interfaces between them, allowing a mix-and-match of suppliers.

Note that the O-RAN Alliance does not mandate the implementation of the RAN architecture or require virtualization; it is primarily concerned with the disaggregation of the Base Station and standardizing the interfaces between parts.

The second is vertical openness through virtualization, where to align with the cloud-native 5G core and adopt the same operating model, it is desirable to virtualize the RAN Network Functions. OpenRAN comes from the Telecom Infra Project group, an initiative to define and build 2G, 3G, and 4G RAN solutions based on general-purpose, vendor-neutral hardware and software-defined technology. The NFs are virtualized into software running on COTS servers. This creates a software-defined RAN that benefits from cloud native DevOps agility and CI/CD (continuous integration and continuous deployment).

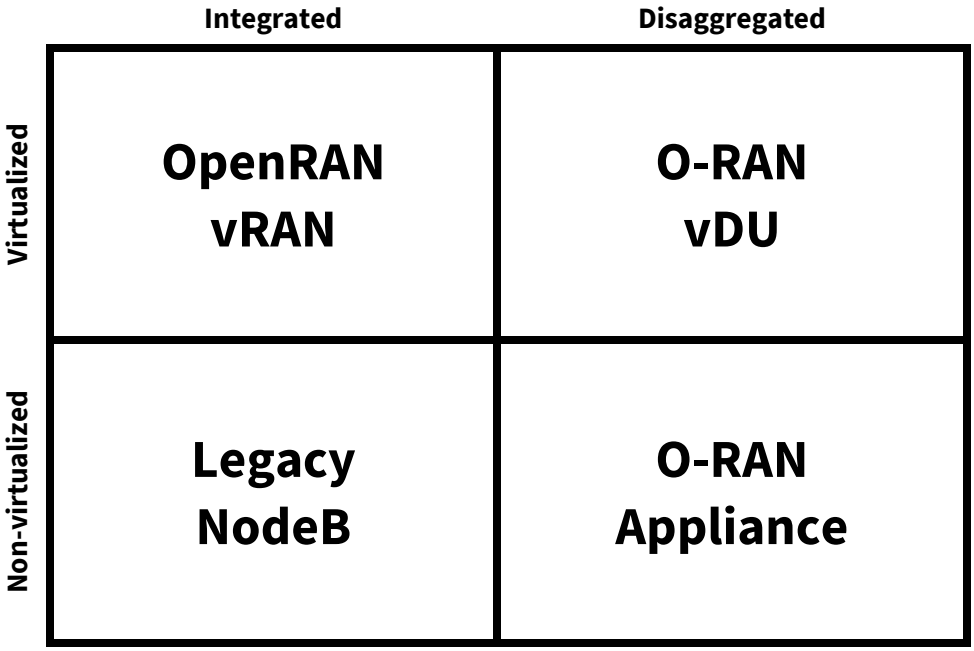

Figure 3 - Open RAN Implementations

These quadrants can be explained as follows:

1. Legacy NodeB: the traditional appliance base station is neither virtualized nor disaggregated. It uses proprietary firmware on custom silicon.
2. O-RAN Appliance: the legacy base station is disaggregated to conform to the O-RAN Alliance architecture and standard interface between components. Each component (CU, DU, RU) can be based on proprietary solutions, but units from different suppliers can be mixed and matched. This model was validated by NTT Docomo in their initial Open RAN deployments.
3. OpenRAN vRAN: virtualized but not disaggregated, often referred to as Open vRAN, with the emphasis on DU virtualization running on general purpose processors. These systems can still have proprietary interfaces and RAN architecture. Some early all-G implementations also fell into this category.

4. O-RAN vDU: disaggregated, virtualized, and O-RAN Alliance interface compliant. This combination is the holy grail. While there is some debate on how open CPUs with accelerators are, it can apply to both look-aside and inline acceleration, if the inline accelerator card is a commercially available, standardized card with supported software from multiple suppliers. The NVIDIA GPU card, which supports virtualization, would also fit here.

5G Virtualization and Cloud-native Benefits

A pertinent question is: "What are the benefits of Open RAN and virtualization?" as these are the industry's driving objectives.

An important goal of Open RAN is to stimulate greater innovation and supplier diversity. RAN is the most complex and costly part of the network, and operators look to reduce dependencies on a single supplier and avoid supplier lock-in. By disaggregating the base station system into separate components, different suppliers, both at the device and chipset level, could compete for a position in the network. Standardized interfaces would make it easier to substitute one supplier for another, or suppliers could focus on a subset of the total solution. Pulling the RIC out and creating an API for 3rd party applications brings real innovation to the RAN management and operation, including the ability to apply AI. In traditional appliance-based systems, the RIC functionality was not separately specified and was the secret sauce of the RAN vendor. The operator was dependent on the capabilities of their supplier.

Virtualization brings a related set of benefits. The 5G architecture was defined as a cloud-native service-based architecture (SBA), adopting the same principles as hyperscalers operating on cloud-like infrastructure. This means using containerized network functions running on standard COTS servers, with a high degree of DevOps style automation utilizing continuous integration and continuous deployment (CI/CD).

With virtualization, the need for customized hardware appliances was eliminated. Using standard COTS servers, the hardware could be sourced from volume server manufacturers, benefiting from industry scale and rate of evolution. The complexity of high-reliability hardware was replaced with resilience at the software level, and systems could scale seamlessly as network load increased.

Notwithstanding the many benefits, cloud-native 5G is not without its challenges, even more so on the RAN. These challenges include:

- Moving from a traditional, appliance-based network to a true cloud-native network requires a sizeable organizational mindset, cultural change, and a reskilling of the workforce. This shift has been one of the biggest hurdles to full standalone 5G deployment.
- In a legacy procurement model, a single vendor provided the solution, although, for business continuity, operators would typically have more than one supplier. With an open solution, especially if best-of-breed components are sourced from multiple suppliers, there is a need for a system integrator. Many operators are finding this a role that they best undertake and are part of the organizational transformation. However, in time, this can also be their best competitive differentiator.
- The technology is immature and going through a maturation and hardening process.
- Legacy procurement practices need to be reassessed. These typically call for long equipment lifecycles, whereas with cloud-native, servers are refreshed more frequently to take advantage of the superior performance and enhanced efficiency each new generation brings. Being GPP COTS, the servers can often be easily redeployed to different parts of the network, utilized for different workloads, or sold off.

5G RAN Virtualization

While the 5G core lent itself to cloud-native architecture, and an SBA (Service-Based Architecture) for the core had already been validated with 4G LTE EPC—albeit based on Virtual Machines (VMs)—to garner the same benefits, there was a desire to extend this to the RAN part of the network. However, the RAN has stringent real-time performance and latency requirements and utilizes complex Digital Signal Processing (DSP). These implementations traditionally required custom logic, embedded processors, and proprietary firmware.

Spearheaded by Intel with FlexRAN, it was proven that the DU RAN signal processing can be implemented on standard GPP, and this has now been further validated by AMD and ARM. NVIDIA has implemented their L1 Aerial RAN stack on the GPU, where any excess capacity can be used for AI and MEC applications.

Initial developments by both Mavenir and Parallel Wireless and Dish Networks and Rakuten deployments have validated that virtualized RAN can match appliance-based systems and operate cost effectively at scale.

The benefits of cloud-native for RAN have been extensively documented by Rakuten Symphony, running on Supermicro hardware. For more background information about automation in Open RAN, visit this TelecomTV webinar.

The benefits include:

- Fast-tracking development and testing new applications and updates, streamlining build, testing in staging, and advancement to production, contributing to agility.
- Software-based and policy-driven monitoring and actioning based on real-time insights, resulting in a more agile and reliable network.
- An automation framework that enables efficient lifecycle management for workloads, further increasing network agility.

The benefits span cost reduction, mitigation of errors when accepting new builds from vendors, and the ability to rapidly spin up new services in days or even hours versus months.

Now that performance parity with legacy solutions is in range and will improve with new-generation chipsets, and the spotlight has moved to energy efficiency. Arguably, at full network load, a custom silicon solution will be more power efficient, but networks only operate at full capacity for a small percentage of the time. With containerized Virtualized Network Functions, there is scope to vary the number of processor cores dynamically with network load. There is also consideration for deploying idle cores—say at night—for other purposes, thereby offsetting costs apportioned to the RAN workload.

The benefits of a virtualized Open RAN solution can be assessed by considering the Huawei case. Western governments determined Huawei equipment to be a security risk and required that it be removed, which necessitated a costly rip-and-replace of all installed equipment. Had the Huawei RAN been deployed using a virtualized, disaggregated Open RAN architecture, with industry standard interfaces, it would likely have been possible to keep the Radio Units (RUs), which make up the most expensive part of mobile network, as these primarily perform RF Signal processing. The DUs now being COTS GPP servers would simply require a different supplier's vDU container software stack to be ported and deployed, and this could potentially all be automated from a central location.

Open RAN Topologies

One of the complexities of the RAN is that the deployment topologies vary depending on the operator and whether it's urban or rural. A major determinant is the availability of fiber in the access network between the core and the cell tower. The topologies are shown in Figure 4 and fall into two distinct categories.

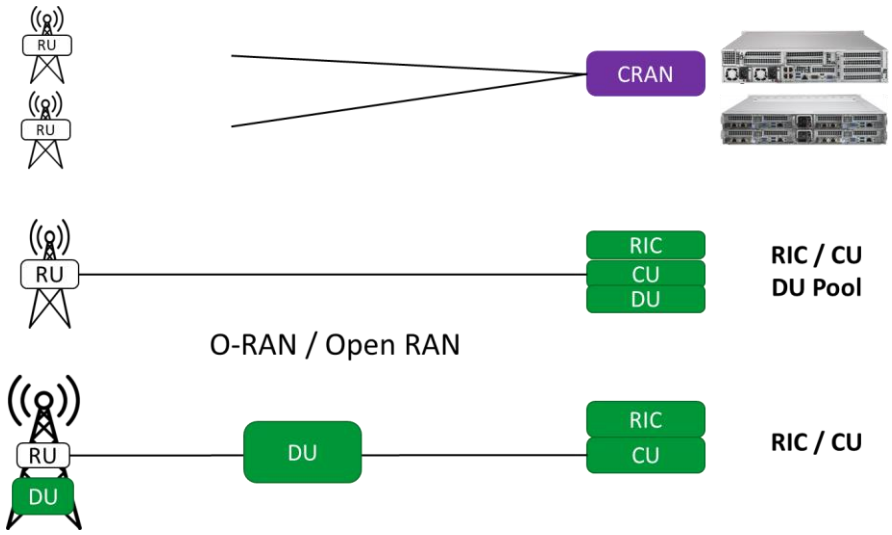

Figure 4 Open RAN Topologies

1. Distributed – DU at the cell site or access network.

   The primary challenges with the DU at the cell site are the extended temperature, vibration environmental requirements, and the form factor. Many cell-site cabinets, especially European ones, are very short in depth, requiring servers with sub-300mm depth. For cell-site deployments, eliminating the cell-site router or fronthaul gateway is a significant cost benefit. This setup requires the DU to have 6 to 12 RU optical connections and directly support 5G GNSS and IEEE 1588 Precision Time Protocol (PTP) Grand Master timing.

   If the DU is placed in the access network, a high-speed front-haul port connection and PTP support is required. Form factors and environmental requirements may be more relaxed and, depending on how many cell sites are served from the DU Node, multiple 1U or multi-node DU servers may be deployed.

2. Centralized – DU in a central office or edge data center

   C-RAN (Cloud RAN; also referred to as Centralized-RAN) deployments are typically resident in a refurbished telco central office or a modern data center. The form factor requirements are more relaxed, allowing for larger servers. With a C-RAN, many cell sites will feed to a common aggregation node. This situation allows for more complex and dynamic workload scheduling, e.g., when the network load is light, the workload can be concentrated on fewer servers. As this requires that containers be dynamically moved between servers without traffic interruption, it also lends itself to fault protection, where workloads on a failing server can simply be reassigned to another. Depending on the deployment, the traffic load will vary by time of day, such as a node that spans a residential neighborhood and business park so that the server capacity can be scaled to the aggregate peak workload rather than the sum of the workloads. For a lightly loaded network, any unused cores can be used to run other workloads.

Supermicro RAN Optimized Systems

Supermicro delivers the industry's broadest portfolio of workload-optimized solutions for Open RAN topologies and deployment scenarios. Additionally, Supermicro server options span all major CPU chipsets, allowing for further optimization depending on the technical or business requirements.

Supermicro offers 5G servers that scale from the cloud core and BSS/OSS to RAN, private 5G, and MEC applications. The servers are available in a variety of form factors to meet all operator's environmental requirements.

Cloud RAN Servers include:

- High core-count single processor and dual processor servers with PCIe Gen 5, supporting up to 400Gbps SmartNIC and DPU network cards. These systems have extensive SSD drive support and are ideal for CU, core, and OSS/BSS applications.
- Multi-node systems for Cloud RAN DU.
- NVIDIA MGX™ systems, which simultaneously support RAN, AI, and MEC applications.

Cell site servers include:

- Edge optimized multi-node systems for central office and access network feeding into a front haul network.
- Highly integrated extended-temperature systems situated at the cell site with direct connection to the RU.

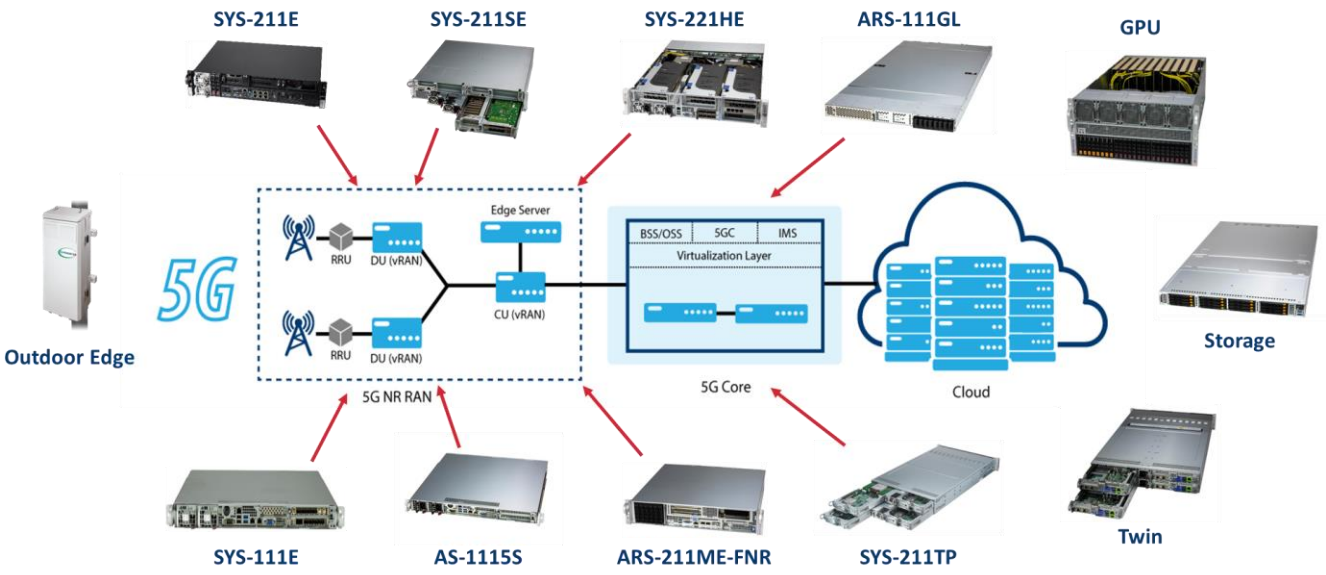

Figure 5 – Supermicro Systems for Open RAN Deployments

New All-in-One RAN-Optimized DU Server

Supermicro introduces a new server for the next generation of Open RAN deployments, designed for deployment at scale and reduction in both OpEx and CapEx. The SYS-211E-FRN(D)N13P is an all-in-one COTS DU server featuring a 5th/4th Gen Intel® Xeon® Scalable processor with Intel vRAN Boost.

The server has an onboard network interface and 12 SFP25G ports, eliminating the need for add-on cards and breakout cables, fully integrated timing support with eight hours of holding time, and a compact, long-life design. The Supermicro SYS-211E systems deliver a fully integrated server optimized for cost, size, and power usage, handling large volumes of traffic at the edge across multiple cell site configurations, including massive MIMO streams.

Visit Supermicro's website to learn more about the new X13 RAN-optimized Server.

Overview of Supermicro Systems for Open RAN

	X13 1U Edge/Telco Server	X13 2U Edge/Telco Server	X13 SuperEdge	X13 Hyper-E	H13 1U Edge/Telco Server	R13 Edge/Telco Server	G1 High-density GPU system
	SYS-111E-F(D)WTR	SYS-211E-FR(D)N2T	SYS-211SE-31A/D	SYS-221HE-FTNR(D)	AS -1115S-F(D)WTRT	ARS-211ME-FNR	ARS-111GL-NHR
Key Features	High-density 1U system for 5G networking	Ultra short-depth form factor for telco and edge deployments	Versatile, compact edge platform with three independent nodes	Powerful dual processor system in compact form factor	High-density 1U system for 5G Networking	Short-depth system for high single-thread performance	High-density 1U GPU system with integrated H100
Processor	5th/4th Gen Intel® Xeon® Scalable processor	5th/4th Gen Intel® Xeon® Scalable processor	5th/4th Gen Intel® Xeon® Scalable processor per node	Dual 5th/4th Gen Intel® Xeon® Scalable processor	AMD EPYC™ 8004 Series processor	Ampere® AmpereOne™ processor	NVIDIA GH200 Grace Hopper™ Superchip
RU	1U	2U	2U	2U	1U	2U	1U
System Depth	429mm / 16.9”	299mm / 11.8”	430mm / 16.9”	574m / 22.6”	429mm / 16.9”	432mm / 17”	940mm / 37”
Connectivity	2x 10GbE	2x 10GbE	1x 1GbE (1)	Up to 2x 100 GbE or 4x 10 GbE	2x 10GbE	2x 25GbE	1x 1GbE
Memory	Up to 2 TB DDR5-5600	Up to 2 TB DDR5-5600	Up to 2 TB DDR5-5600 (1)	Up to 8 TB of DDR5-5600	Up to 576GB DDR5-4800	Up to 4 TB DDR5-4800	Up to 480GB LPDDR5X and 96GB of HBM3
Expansion Slots	2x PCIe 5.0 x16 FHFL 1x PCIe 5.0 x16 LP	Up to 6x PCIe 5.0 slots (mixed sizes)	2x PCIe 5.0 FHHL 1x PCIe 5.0 HHHL (1)	Up to 8 PCIe 5.0 slots (mixed sizes)	2x PCIe 5.0 x16 FHFL 1x PCIe 5.0 x16 LP	6x PCIe 5.0 slots (mixed sizes)	2x PCIe 5.0 x16 FHFL
Power Supply	Redundant 800W AC or redundant 600W DC PSU	Redundant 800W AC or redundant 600W DC PSU	Redundant 2000W AC or redundant 2000 DC PSU	Redundant 2000W DC or redundant 1300W DC PSU	Redundant 800W AC or redundant 600W DC PSU	Redundant 1000W AC PSU	Redundant 2000W AC PSU

(1) Per Node

For more information about Supermicro’s systems for 5G and Open RAN deployments, visit www.supermicro.com/5g.

Glossary:

Cloud RAN

A virtualized RAN built on cloud-native properties, including microservices, CI/CD, and containerization.

Open RAN

Generic term for Open Radio Access Network architecture with disaggregated functionality, software-defined technology, and open, community-developed standards.

OpenRAN

An initiative based on TIP's (Telecom Infra Project) OpenRAN Project group.

O-RAN

O-RAN refers to the O-RAN Alliance Community and the specifications defined in their framework and interfaces.

vRAN

Refers to virtualized radio access networks, a way to run baseband functions as software.

About Supermicro

As a global leader in high performance, high efficiency server technology, and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

For more information, visit www.supermicro.com.