White Paper

vRAN

The Journey to a Cloud-native, Fully Software-defined vRAN Architecture

How vRAN / OpenRAN can increase vendor independence and deliver the future of 5G mobile communications using an open interface architecture running on industry-standard hardware

Authors

Adnan Boustany, Director – vRAN Infrastructure Customer Programs, Intel Corporation

Edel Curley, Market Development Manager for EMEA Communication Service Providers, Intel Corporation

Andy Dunkin, HW Development Manager OpenRAN, Vodafone

Devang Solanky, OpenRAN Product Manager, Vodafone

Table of Contents

Executive Summary

With the telco industry's transformation to virtualized RAN now underway, OpenRAN has matured from a nascent technology to become a robust, ultramodern network architecture. Today, a growing ecosystem of solutions providers is powering real commercial deployments by leading communications service providers (CoSPs), creating a momentum that is undeniable.

Vodafone and Intel have been OpenRAN pioneers, and the multi-year ongoing collaboration between these industry leaders has made significant contributions towards the rapid development of this technology. Vodafone has already started commercial deployments in a key market such as UK with Intel-based general purpose processor (GPP) platforms. As they look toward scaling these deployments to other markets, this paper reflects on the progress Vodafone and Intel have made thus far. It also highlights key lessons learned, illustrating how OpenRAN platforms are becoming more software-defined and evolving further towards a cloud-native architecture. It also provides insight into how future investments in technologies (silicon and software) will make OpenRAN solutions progressively more efficient and easier to deploy.

Introduction

Operators around the world are moving to implement OpenRAN platforms, as evidenced by numerous lab trials, field pilots and first deployments, including Europe's first commercial OpenRAN roll-out by Vodafone. With these deployments, the industry's experience and understanding grows. These learnings drive vendors and operators to look deeper into the technical platforms and ensure that compute, RF and all other platform components are fully optimized. In doing so, this focus is increasingly placed at a component and software subsystem level to ensure the two are properly implemented for optimum performance. The necessary synergy required at this level has been identified as critical.

In traditional Radio Access Networks (RAN) deployments, these complex systems require specialized hardware and software in order to deliver the performance and reliability mobile communications demand. Much of the system is custom designed, with hardware and software tightly coupled and offered by one or a few vendors. Competition is limited and innovation confined by the commitments of vendors.

With the development and continuing advancements of high-performance general purpose processors, such as the family of Intel® Xeon® Scalable processors, network processing operations can now be accomplished with common-off-the-shelf (COTS) servers. The introduction of virtualization into communications systems enhances operators' flexibility to deploy many network functions across a single platform. This, in turn, creates more opportunities for efficiency and commonality with an operator's IT systems. OpenRAN has been developed to run on GPP-based servers, and as the industry moves to the latest generation silicon, Vodafone has seen increasing evidence that OpenRAN's performance and reliability can match that of the traditional RAN solutions.

Vodafone has been a primary proponent of OpenRAN for several years and holds seats on both the ORAN Alliance and Telecom Infra Project boards. Working with its strategic vendors, Vodafone is building Europe's first commercial deployment of an OpenRAN network. The commitment includes 2,500 OpenRAN-based sites, with its first deployment already completed in Bath, Somerset, U.K.

Traditional RAN vs. vRAN / OpenRAN

GPP-based systems and virtualization have enabled this new approach to cellular network delivery, referred to as vRAN and embraced in OpenRAN architecture and specifications, which is taking RAN towards a more software-defined architecture model (Figure 1). vRAN has already enabled greater flexibility, lower cost, and new innovations in the compute side of the baseband unit (BBU).

With OpenRAN, standardization of the *interfaces* across the entire RAN deployment driven by the O-RAN Alliance enables a fully open architecture. As seen across other industries, an open, standards-based architecture invites competition and accelerates innovation from developers and vendors. More suppliers expand the ecosystem, creating greater price flexibility. Overall, a fully open approach enables greater business and service agility, reduced cost, and faster deployment of new services. This allows communications operators—both large and small—to select the right combination of interoperable systems and services from a wider range of vendors to meet their customers' needs.

OpenRAN also enables evolution to a fully software-defined, cloud-native model of communications, from the centralized unit (CU) and distributed units (DU) to the RAN sites. A cloud-native approach allows providers to deploy new services or upgrade existing services easily and quickly as new innovative solutions become available.

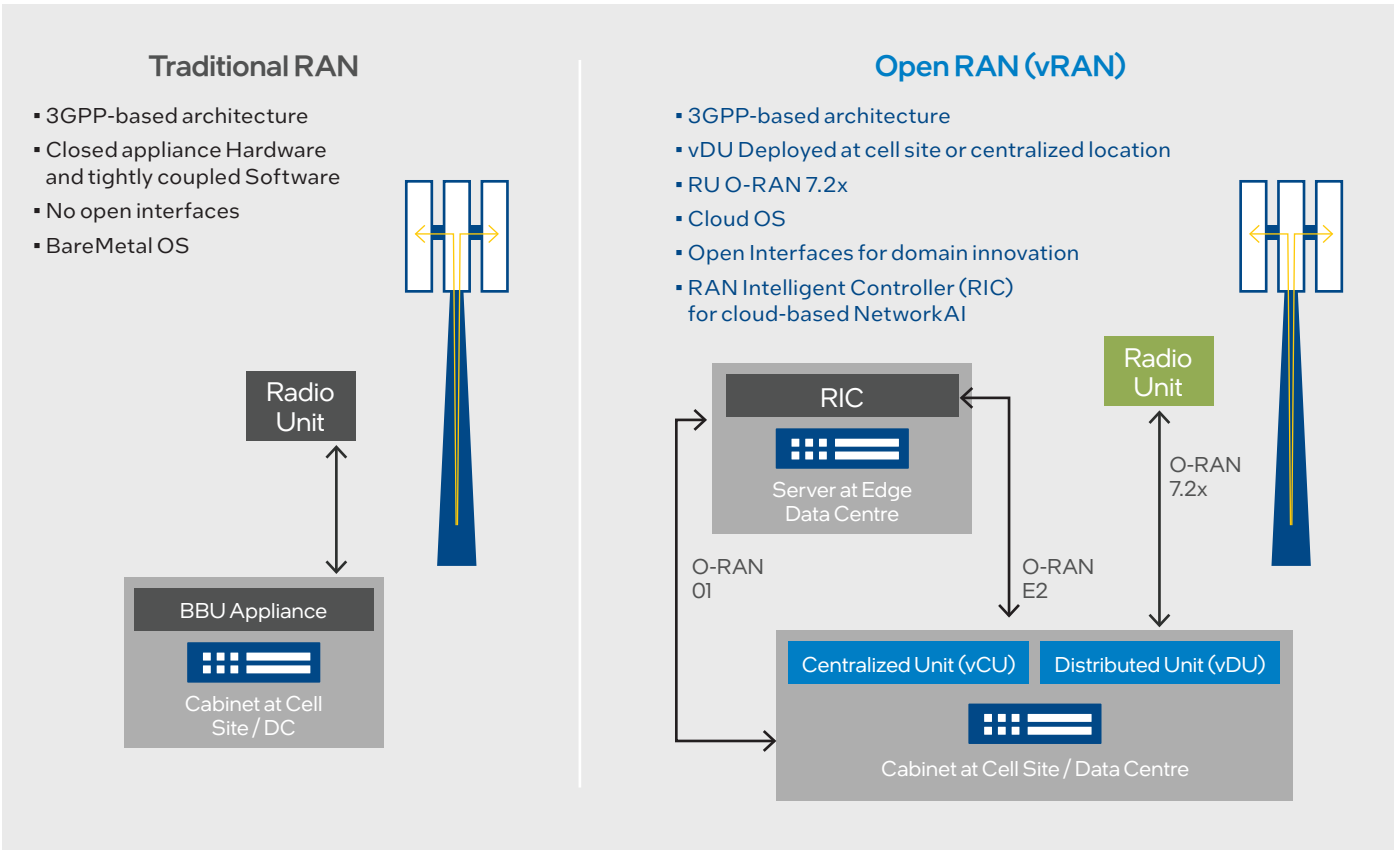


Figure 1. Traditional RAN deployments versus vRAN and OpenRAN architectures

Current Status of OpenRAN and Key Learnings from Early Deployments

In January 2022, Vodafone switched on its first OpenRAN site in Bath, Somerset, U.K. and announced its strategic vendor partners, whose servers are powered by 3rd Gen Intel® Xeon® Scalable processors, using O-RAN Timing and Sync-capable NIC cards for fronthaul, and Intel Dedicated vRAN Accelerator ACC100 adapter for FEC offload. The design is O-RAN compliant, based on 7.2x interface radio and uses of COTS hardware in a containerised software. In late 2022, Vodafone announced its first Golden Cluster, based on its extensive end-to-end network integration and detailed testing. There have been significant learnings already with more expected as Vodafone moves into full operations.

Vodafone has also published a whitepaper[1] on the importance of System Integration, which is very important to be addressed by the industry together.

OpenRAN brings in lots of additional potentials with advancements of silicon ahead. However, using the disaggregated approach, the industry also **needs to ensure that both hardware and software leverage each other's highest potential to bring in the maximum efficiencies possible in OpenRAN**. Together in this white paper, Vodafone and Intel will highlight the evolutions and potentials that the industry needs to leverage for OpenRAN.

How Intel is Driving vRAN/OpenRAN Platform Innovation to Deliver Gen-on-Gen Improvements Required to Advance OpenRAN

Traditional RAN products are largely based on custom, purpose-built technologies to implement Layer 1 (L1) processing, including specialized processors, DSPs, and accelerators. L2 and L3 processing is typically done on CPU cores with accompanying accelerators for certain tasks. Over the last ten years, Intel has been investing in the FlexRAN™ reference architecture, which includes both hardware and software investments to help easily deploy RAN on industry-standard technologies. On the hardware side, multiple generations of the Intel® Xeon® processor family have integrated technologies formerly done by external devices and included new and evolving instruction sets that enable efficient implementation of L1 processing, in addition to L2 and L3 workloads. Intel is also driving platform Innovation in other key areas.

Openness: Building a RAN software stack on a GPP platform such as Intel Xeon processors makes it easy for a wide variety of ecosystem vendors to develop their own differentiated solutions tailored to operators' needs. The x86 Xeon platform has a broad and mature developer ecosystem. At the same time, the Intel Xeon processor platform integrates a host of instruction sets that can be used for implementing optimized RAN workloads, such as AES-NI for encryption or AVX-512 for L1 baseband processing. Additionally, Intel's FlexRAN reference software and its comprehensive suite of features provides operators and developers with an example implementation, which builds on top of existing open source components, as seen in Figure 2. This enables RAN software vendors to rapidly develop RAN stack, either on their own or by leveraging FlexRAN stack.

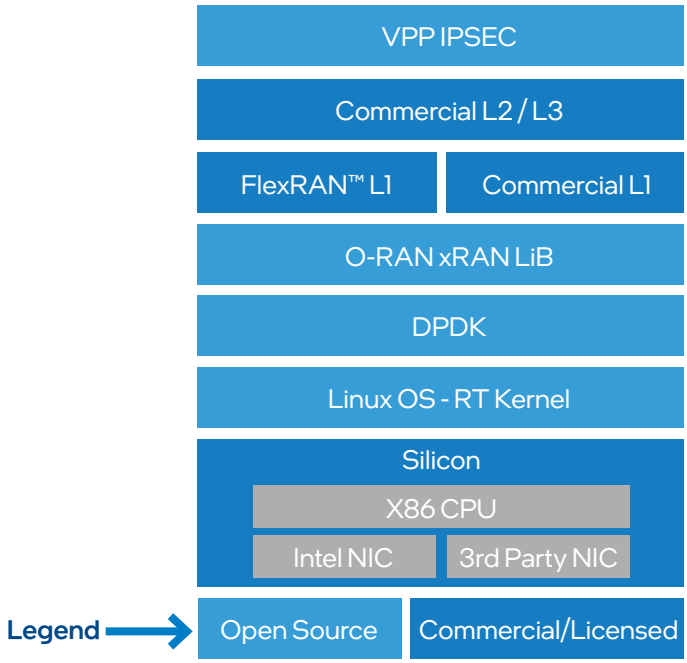


Figure 2. FlexRAN™ Software Architecture enabling an ecosystem for vRAN solutions

Integration: Integrating technologies into the processor is a design philosophy Intel has followed for years. This integration delivers an elegant solution that can improve performance and reduce dependency on external cards and other hardware. Reduction in hardware enables more physically compact and less costly solutions.

Software Enabling: Intel hardware and software investments have enabled efficient execution of RAN workloads on Intel Xeon Scalable processors. Each new generation of these processors has demonstrated frequency-scaling headroom, new performance levels, and technology which includes integration of vRAN acceleration into the processor. Simultaneous 2G, 4G and 5G RAN software pipelines can run inline within the GPP, eliminating the need for specialized external components.

Intel® Xeon® Scalable Processors for vRAN

Intel Xeon processor-based solutions have supported various vRAN/OpenRAN trials on the 1st Gen Intel Xeon Scalable processors, followed by wide-scale commercial deployments on 2nd Gen Intel Xeon Scalable processors. Today, 3rd Gen Intel Xeon Scalable processors support deployments with higher network capacities. It is the processor of choice in Vodafone's effort to bring 2,500 OpenRAN sites in the UK. Please see Figure 3.

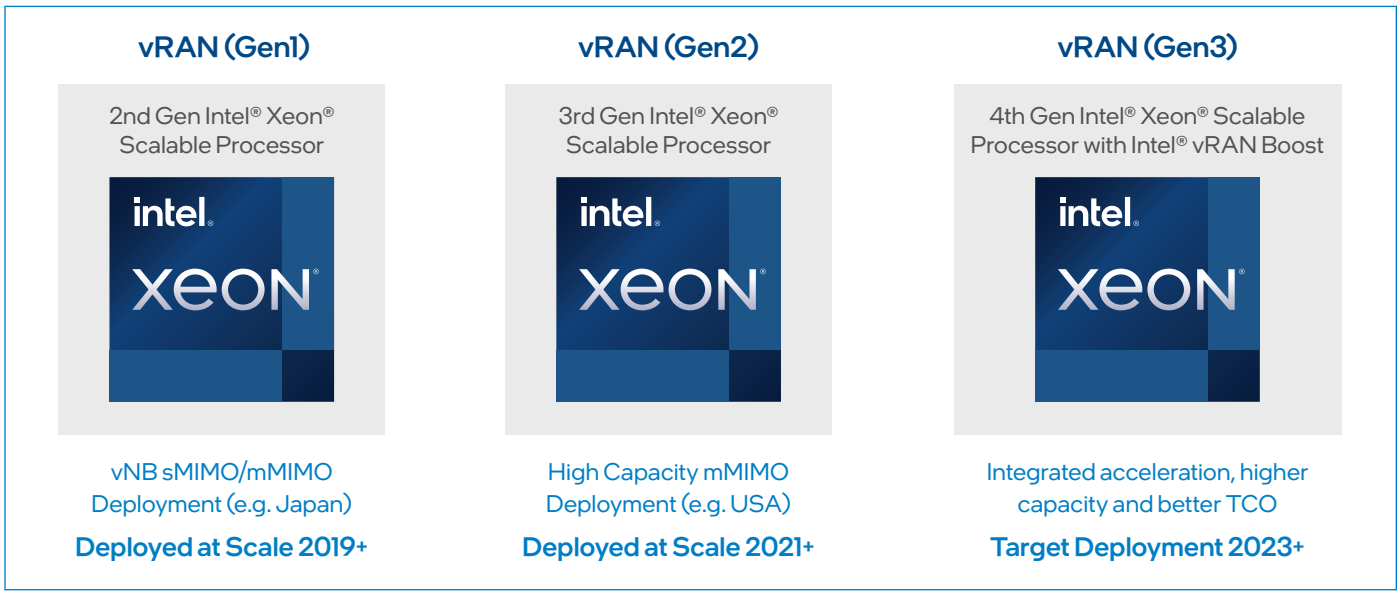


Figure 3. The multi-generation view of Intel® Xeon® Scalable Processors

Intel® Xeon® Scalable Processor Instruction Set Innovations for vRAN/OpenRAN

Intel® AVX Accelerates Floating Point Operations

RAN workloads require high-performance of floating-point calculations and bit manipulation, executed as parallel, Single Instruction, Multiple Data (SIMD) processes. Intel has enhanced its floating-point technology through years of evolution, from Intel® Streaming SIMD Extensions (Intel® SSE) to Intel® Advanced Vector Extensions (Intel® AVX), Intel® AVX2, Intel® AVX-512[2] and more. These instructions are not only applicable to vRAN/OpenRAN, but also can be utilized across many other network segments. See Figure 4.

Intel AVX-512 technology is optimized for high-performance, parallel execution of floating-point processing and bit manipulation. It offers significant performance gains over previous generations of Intel AVX. Enhancements include doubling register width and the number of available registers, and generally offering a more flexible instruction set compared to its predecessors Intel AVX has been integrated in all generations of Intel Xeon Scalable processors. It is now further optimized in the latest third generation processors with compelling performance benefits for vRAN/OpenRAN.

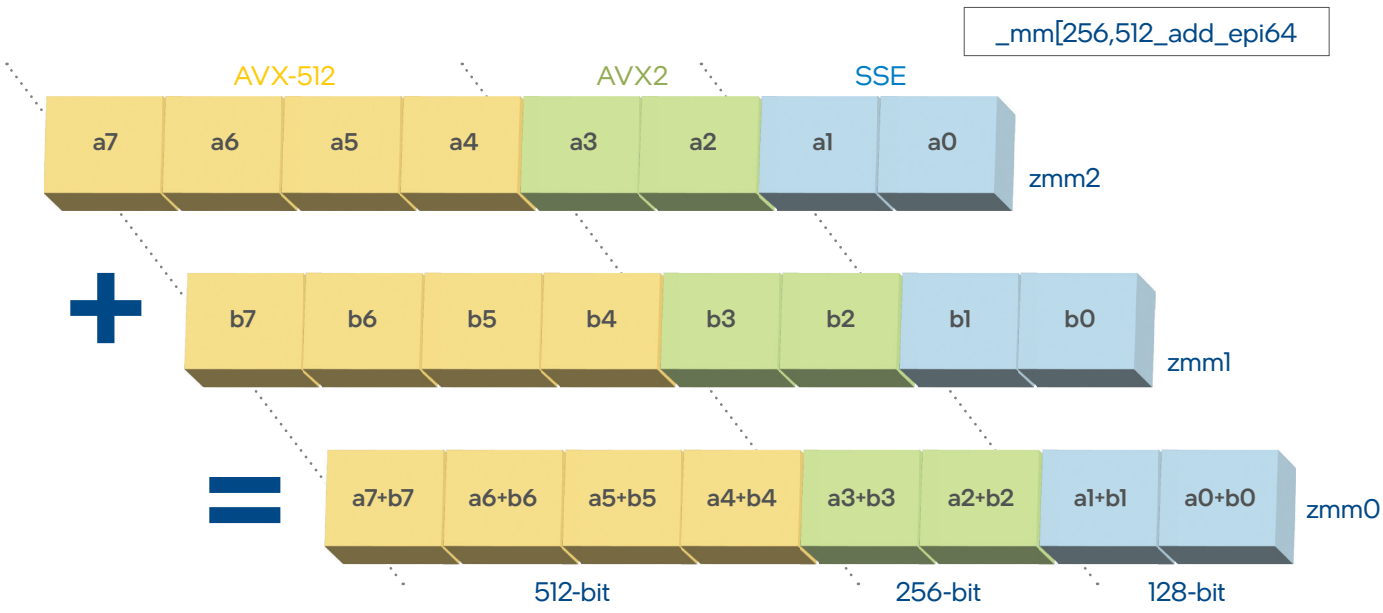


Figure 4. Example vector addition with Intel® SSE, Intel® AVX2, and Intel® AVX-512 instruction sets

Intel AVX Enhancements Accelerate Half-precision Floating Point Operations

4th Gen Intel Xeon Scalable processors integrate an extension to Intel AVX-512 optimized for vRAN. This set of instructions supports a wide range of general-purpose numeric operations for 16-bit, half-precision IEEE-754 floating-point (Figure 5). It complements the existing 32-bit and 64-bit floating-point instructions already available in Intel® Xeon® processor-based products.

The new enhanced instructions for vRAN are ideal for numeric operations where reduced precision can be used, such as signal and media processing. For example, wireless signal processing operations, including beamforming, precoding, and minimum mean squared error (MMSE), can take advantage of AVX-512 for vRAN, accelerating results. The addition of these instruction extensions enables GPPs to compete effectively with DSP performance.

Figure 5. Intel® AVX-512 for vRAN supports faster calculation of smaller floating-point numbers

Integrating vRAN Acceleration

L1 forward error correction (FEC) is a compute-intensive 4G and 5G workload. FEC resolves data transmission errors over noisy channels. FEC techniques detect and correct a limited number of errors in 4G or 5G data without the need for retransmission.

Initially, FEC was implemented in PCIe add-in cards to accelerate processing in hardware. The 4th Gen Intel Xeon processor family will include a SKU with Intel vRAN Boost, a feature that eliminates the need for an external accelerator card by fully integrating vRAN acceleration directly into the Intel Xeon processor. Intel vRAN Boost helps reduce operators' component requirements, which translates to reduced system complexity and can provide energy efficiency advantages. Additional benefits of utilizing a GPP platform with fully integrated vRAN acceleration are described in a recent Intel blog entitled, "Meeting Future Data Demand with vRAN Processing Capacity, Energy Efficiency."[3]

Software as an Enabler of Efficient vRAN Implementation

Intel investments give operators the solutions they need to realize the goals of vRAN and OpenRAN. Performance and feature enhancements over the generations of the Intel Xeon processor family prove the commitment by Intel to continue to enhance the performance-per-watt and price/performance of this GPP processor family to meet the communications industry's needs. By running on GPPs, OpenRAN can benefit from these performance improvements on the latest silicon technologies, including:

IPC: Intel delivers performance gains, gen-on-gen, which can be measured via a number of parameters. One such parameter is the Instruction Per Cycle (IPC) gain. **This means that software can be compiled for the new silicon generation and get the IPC benefits, without re-writing the code.**

Intel® AVX: In addition to IPC, there are other enhancements within the CPU, such as the AVX-512 extension for vRAN, that software vendors can utilize for achieving even higher performance than solely relying on IPC gains.

Hyper-Threading: Another example of how the software is architected to use well known technology features within the CPU is Hyper-Threading. Intel® Hyper-Threading technology enables more parallel processing without adding more cores to run the workload.

Figure 6 illustrates how Hyper-Threading works with two threads (in dark blue and light blue in diagram) that share a physical core. The Y axis is the time it takes for both threads to complete processing, and the X axis is available hardware resources in the core occupied by the threads. In this example, it is assumed that both threads have the same priority and don't depend on each other.

- **With Hyper-Threading off**, both threads are pinned to the same physical core and run in a serial fashion. As shown in Figure 6, there are many hardware resources unused that lead to stalls (in grey). These are wasted cycles.
- **With Hyper-Threading on**, each thread is pinned to a logical core under the same physical core. Therefore, there are few stall cycles and the hardware blocks are filled with both threads.

By utilizing all the technology improvements in the latest silicon generation, communications service providers can realize the highest performance mobile networks.

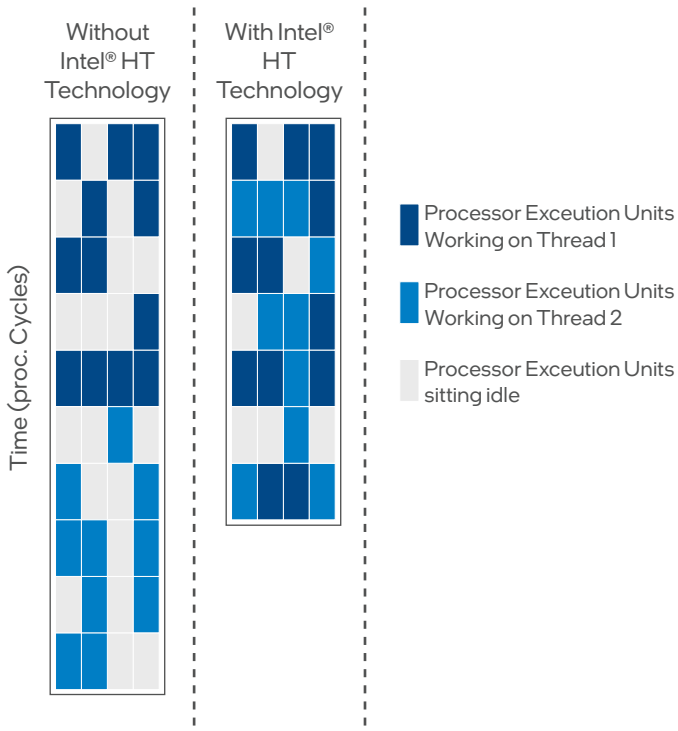


Figure 6. Example of Thread Processing Pipeline

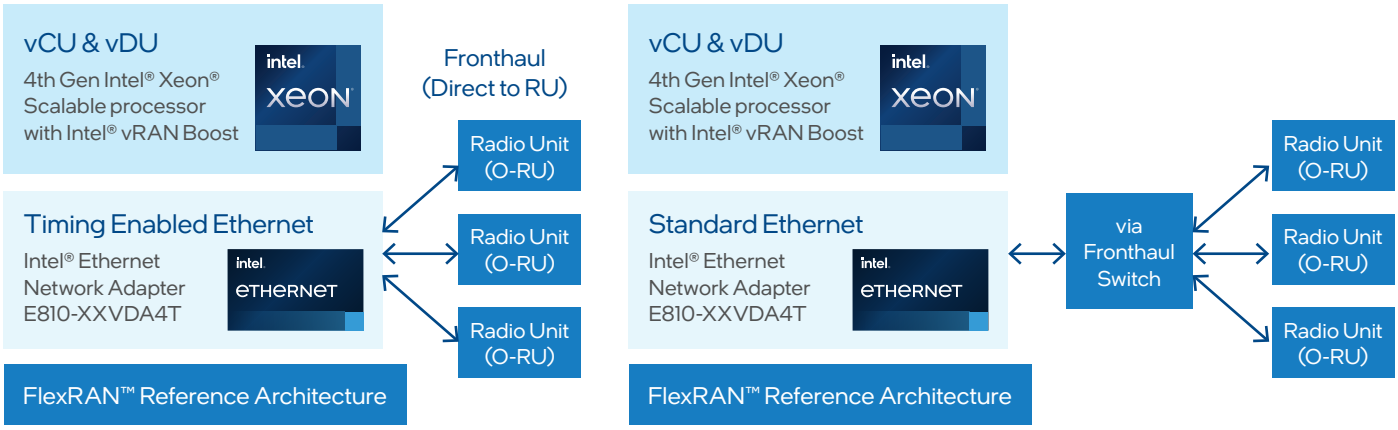


Figure 7. Fronthaul Connectivity Options

RAN Synchronization

O-RAN 7.2x defines tight timing requirements. The IEEE 1588 Precision Time Protocol (1588 PTP), in conjunction with Synchronous Ethernet (SyncE), enables timing synchronization across the network. This functionality can be included in standard PCIe Ethernet adapters with the appropriate functionality, which provides greater flexibility at a lower price point than deploying special-purpose timing appliances.

In 2022, Intel launched the Intel® Ethernet Network Adapter E810-XXVDA4T and the Intel Ethernet Network Adapter E810-CQDA2T, which supports 1588 PTP and SyncE. These adapters integrate a high precision oscillator for greater phase accuracy and extended holdover, and it includes dual SMA connectors to connect to external time sources. They also support an optional Global Navigation Satellite Systems (GNSS) module to allow frequency, phase, and time-of-day synchronization with multiple GNSS constellations. Intel and Vodafone have helped to evolve these solutions, which are critical to deliver competitive one-box solutions for saving space, power and money.

The type of connections from the DU to the RU, dictates the type of ethernet technology needed, as seen in Figure 7.

Innovations in Power Management

Delivering higher performance and lower power consumption

BBU power consumption is an area that is ripe for innovation. BBU sites are typically provisioned for peak throughput. However, for major portions of the day and night, they do not run at full capacity. Power optimization must be a priority for the industry; the ecosystem is exploring various solutions to solve this critical issue. Power savings come in the form of two categories. The first is related to understanding what is happening in the network (**Network-level Power Optimization**), and the second is related to the underlying silicon platform and what it can do (**Platform-level Optimizations**).

Network-level Power Optimization: The data center industry uses wide instrumentation and power management software to enjoy power savings in times of low usage. A fully virtualized OpenRAN architecture can leverage these power saving capabilities. To fully comprehend what is happening in the network, detailed information must be available on an on-demand basis, including information about network load and performance, as well as the infrastructure at each node. For example, utilization of radio resources enables the Service Management and Orchestration (SMO) to make decisions on powering down radios not needed for a projected period.

Platform-level Optimizations: Operators can move from a purely peak/off-peak mindset to microsecond control by using software to enable greater power management on an application-specific basis. This can be done at a system or platform level. Intel devices feature comprehensive telemetry capabilities that include processor-level characteristics, such as core utilization and power consumed, and network information, such as packet receive and drop rates. Intel devices also provide a range of power management knobs that can be activated to dynamically modulate power based on changing processing performance requirements.

Frequency modulation features called P-states enable a privileged server agent executing on the cloud to change the platform-wide power by changing core frequencies or other IP blocks as the traffic load varies.

Application power management: At an application level, very rapid core state transitions supported by C6 states in Intel CPUs may be leveraged to put cores into sleep states for very short time periods (in the order of microseconds). This is done based on anticipated processing workload in the next slot interval.

Other C-states provide capabilities to control core states supporting varying degrees of transjtion times with associated power savings variations. Applying cloud native principles, core scaling may be used to dynamically vary the number of cores running a given workload based on traffic load. This approach can also be extended across multiple servers, so when traffic load is low, all workloads can be consolidated on a few servers, and some servers can be shut down when anticipated. Laboratory testing of C-state management has shown the potential to generate power savings of approximately 30%.[4]

Industry coming together: Vodafone has been working with a number of partners on a standardized method for measuring the energy consumption of OpenRAN. At ORAN Plugfest Spring 2022, Vodafone, collaborating with Wind River, Intel, Keysight Technologies, and Radisys, have demonstrated the use of P-states to show a reduction in power consumption of an OpenRAN infrastructure by 9% and 12%, during high and low mobile traffic peak scenarios respectively.[5]

The Open RAN architecture developed by the O-RAN Alliance—shown in Figure 8—features standardized hardware and software interfaces aimed at enabling a broad multi-vendor ecosystem. This ecosystem provides operators various options for network management and orchestration capabilities such as power management, traffic management and steering, network slicing, and radio resource management that may also leverage AI/ML-based algorithms.

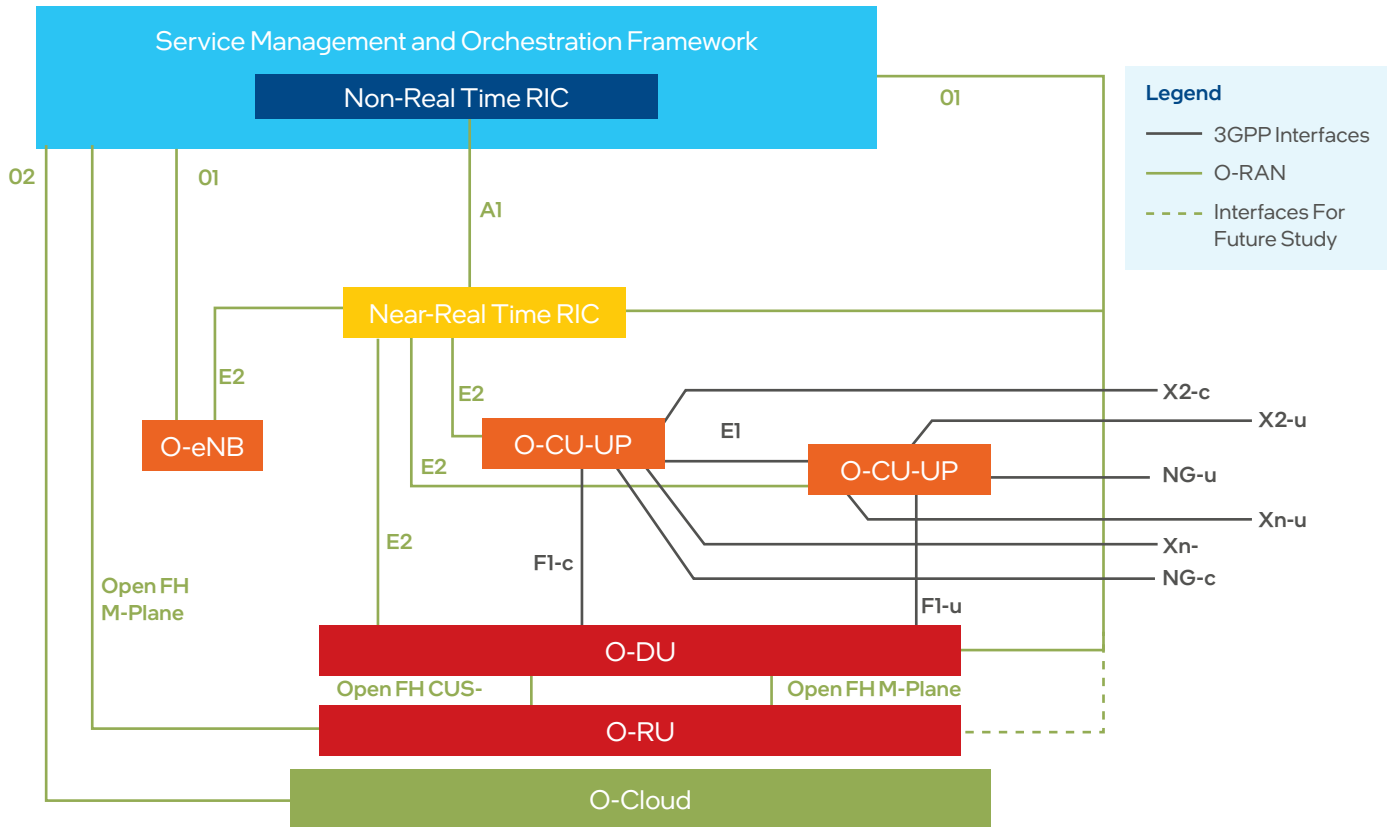


Figure 8. O-RAN Architecture

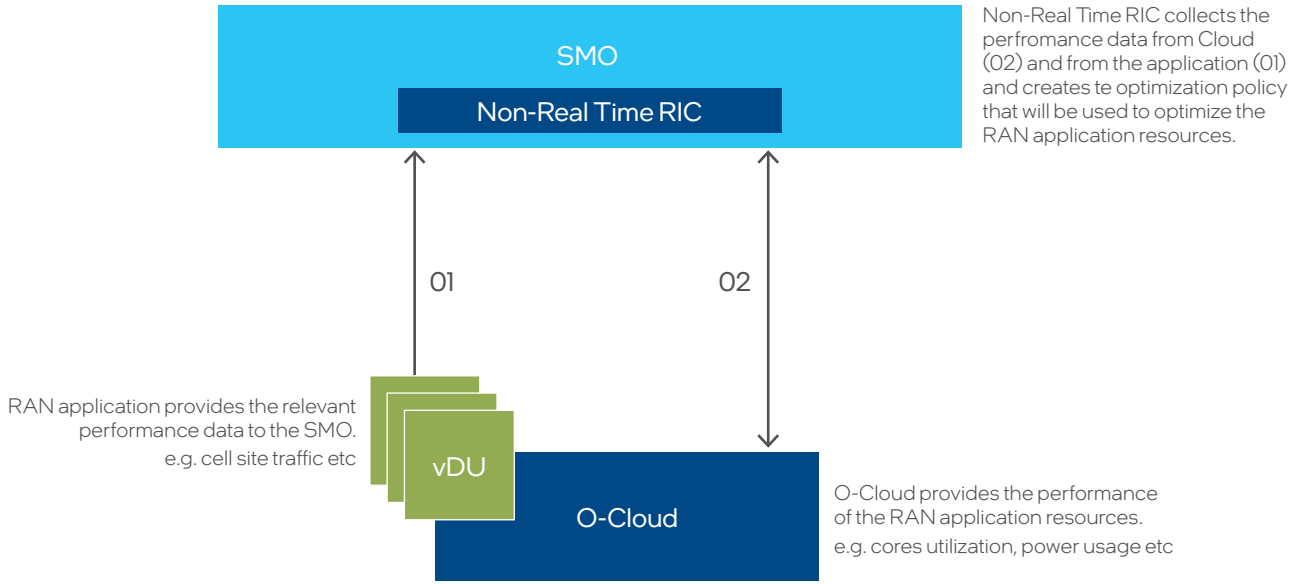


Figure 9. OpenRAN software stack continues to develop through open source contributions by vendors and developers

Figure 9 illustrates an example of how power management can be achieved in the O-RAN architecture. The O-Cloud provides the relevant performance data about the resources used by the application to the SMO, e.g. power consumption data, core utilization. This data is analyzed by the non-real time RIC to derive the optimization policy that the O-Cloud will use for optimizing the O-Cloud resources used by the RAN application. An rApp running in the non-real time RIC may be used to further fine-tune the optimization policy.

Telco Servers for OpenRAN/vRAN

Telco servers have distinct physical and environmental requirements given their physical location in the network. Macro D-RAN servers must meet Telecordia NEBS level 3 specifications. All ports, including power supply, must be accessible from the front panel with front-to-rear airflow. To achieve optimal power efficiency, careful consideration must be given to the thermal design. The airflow impedance shall be minimized to enable heat dissipation with optimum fan duty cycle.

There are a number of competing priorities with respect to telco servers:

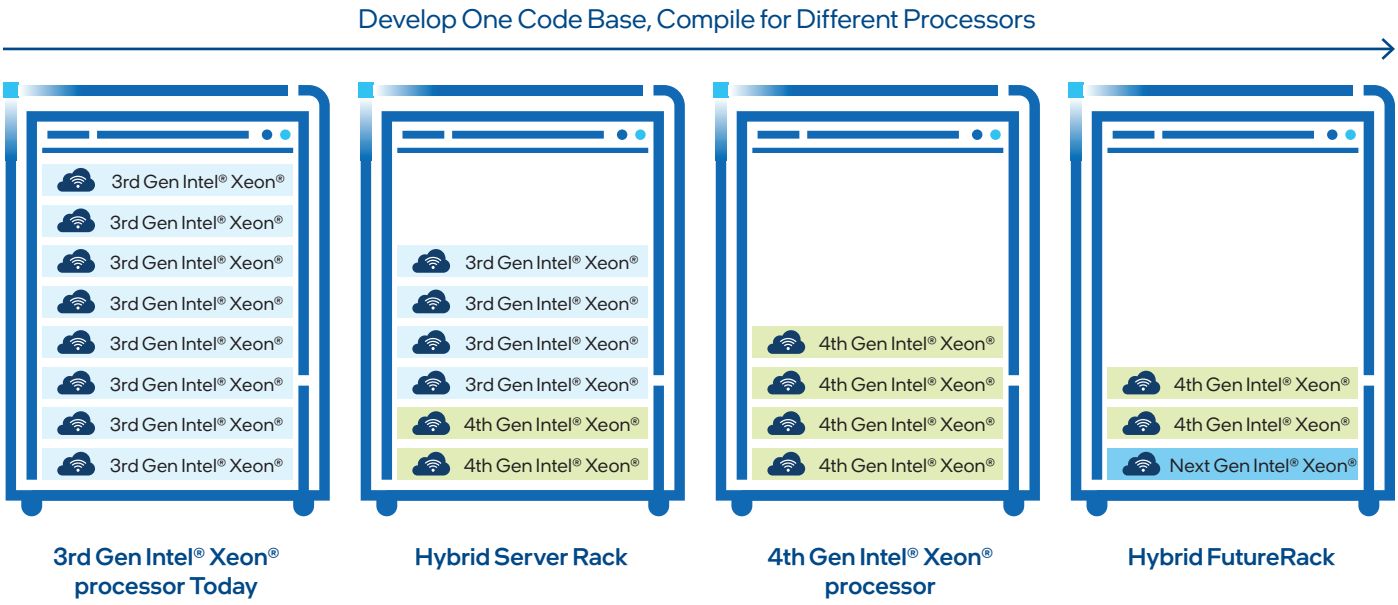
- Compute density (or cell capacity) per rack unit
- Airflow impedance from front to rear
- Front panel port density (particularly for narrowband D-RAN)
- Fixed ethernet on server vs. modular server supporting Ethernet NICs.

Server OEMs are developing a variety of telco form factors with priority emphasis on different parameters:

- Dense compute implementations in the form of 1U2N, 2U2N, and 2U4N offers greatest compute density in 19" Racks.
- 1U1N (one compute node in full width 1U of rack) provides front-panel real estate for PCIe add-in cards, or a large number of discrete Ethernet ports for efficient front-haul connectivity.
- 2U2N form factor consists of two half-width 2U sleds; this form factor can support higher powered CPU SKUs due to it being able to accommodate taller heat-sink while supporting PCIe add-in cards
- Shallow depth server options are desirable in some brown-field D-RAN scenarios where existing rack infrastructure has real estate available for 300mm equipment.

Multi-generation Hardware Support

To help RAN software developers re-use software investments across multiple generations of Intel processors, Intel has invested in a library to help developers code once and then recompile across multiple generations of Intel Xeon processors. A good example of this is the development of Intel C libraries referred to as "dvec," which aims to assist with forward compatibility on Intel CPU generations. This means that RAN software vendors can maintain a single code base and deploy it on multiple Intel Xeon processor generations, as illustrated in Figure 10.



Intel C++ Class libraries allow developers to simplify code development, increase forward and backward protability, and decrease maintenance costs.

Figure 10. The Intel® C++ compiler enables development for multi-generation hardware platforms

Summary and the Importance of Collaboration to OpenRAN's Future

OpenRAN, a new way of building radio networks, has been evolving since its early inception in the O-RAN Alliance and the Telecom Infra Project. Solutions today are **meeting the requirements of leading Communications Service Providers**. Those requirements are also evolving, and OpenRAN solutions must evolve with them to ensure they remain competitive and OpenRAN becomes a mainstream technology.

One of the main tenets of OpenRAN is **disaggregation** – breaking the complete RAN solution into different parts so that different ecosystem players can focus on one part (or some or all). This builds the foundation for a more diverse and specialized ecosystem that is now able to innovate rapidly at a component level. Standardized interfaces, a key focus area for O-RAN, enables these components to plug and play, alleviating time-consuming integration cycles.

As communication service providers such as Vodafone work with selected partners to build a complete solution based on their respective components, **it is critical that any efficiencies and improvements made at the component level (e.g. new instruction sets) can be leveraged at the solution level**. Therefore, open cooperation and collaboration among partners is a key tenet of OpenRAN, and imperative to ensure that the end result is more than the sum of its parts.

Vodafone and Intel have been closely collaborating, and in this paper, we have shared some of the key technologies and innovation in Intel-based OpenRAN platforms that have led to the solution that is being deployed today, as well as laying out the path to the next generation platforms that will meet Vodafone's requirements in the future.

vRAN/OpenRAN today and into the future

- vRAN is deployed for Macro Networks
- vRAN investments and innovations will continue to enhance the TCO
- vRAN solutions need Industry co-operation with software taking advantage of the Silicon capabilities.
- O-RAN and TIP to continue bringing the ecosystem together on technologies and use cases.

As a result of this, Intel-based OpenRAN platforms are form factor optimized, meet the environmental requirements and have specific optimizations for RAN workloads from the functionality point of view, as well as efficiency and performance. As next steps, Intel is continuing to build on the existing platforms and continuing to make significant steps forward on improving energy efficiency and overall TCO. Intel is open and keen to work with ecosystem partners to ensure that Vodafone leverages these innovations in their solutions.

The industry is delivering strong technology evolution roadmaps and innovation. In parallel, it is critical to remain fully focused on operators' key metrics, cost, performance, and efficiency, as these are critical to industry success. Now is the time to concentrate on collaboration and common goals leveraging operators' experience and feedback from first commercial deployments. By aligning these industry efforts through the O-RAN Alliance and Telecom Infra Project communities, the industry can deliver the necessary cutting-edge compute platforms necessary to support OpenRAN as it expands into more widespread global deployments at scale.

More Information

https://www.intel.com/content/www/us/en/wireless-network/5g-network/radio-access-network.html

[1] https://www.vodafone.com/sites/default/files/2022-05/open-ran-system-integration-white-paper.pdf.

[2] https://builders.intel.com/docs/networkbuilders/intel-avx-512-fp16-instruction-set-for-intel-xeon-processor-based-products-technology-guide-1651874188.pdf

[3] https://community.intel.com/t5/Blogs/Tech-Innovation/Edge-5G/Meeting-Future-Data-Demand-with-vRAN-Processing-Capacity-Energy/post/1415977

[4] https://builders.intel.com/docs/networkbuilders/benefits-of-virtualizing-the-layer-1-in-a-ran-stack-1664444022.pdf

[5] https://www.vodafone.com/news/technology/vodafone-wind-river-intel-keysight-technologies-radisys-test-green-open-ran

1222/DA/MESH/PDF 353767-001US