Cloud RAN: A Guide to Acceleration Options

NOKIA

Introduction

Future Radio Access Networks (RAN) are expected to evolve gradually towards Cloud RAN based solutions, which will be mainly deployed alongside existing purpose-built Classic RAN networks. The difference between Cloud RAN and Classic RAN is in the so-called baseband computing, consisting of the Distributed Unit (DU) and the Centralized Unit (CU), whereby the Radio Units (RU) are the same for Cloud RAN and Classic RAN. Furthermore, the term Cloud RAN is used here to reflect the target state where DU and CU software is fully cloud-native rather than only virtualized ("vRAN") and merely enabling the use of "Commercial-off-the-Shelf" (COTS) server hardware. In fact, such a vRAN should rather be called Server RAN than Cloud RAN, because it does not deliver many of the benefits of cloud computing as they are known from e.g. IT or Core Network workloads.

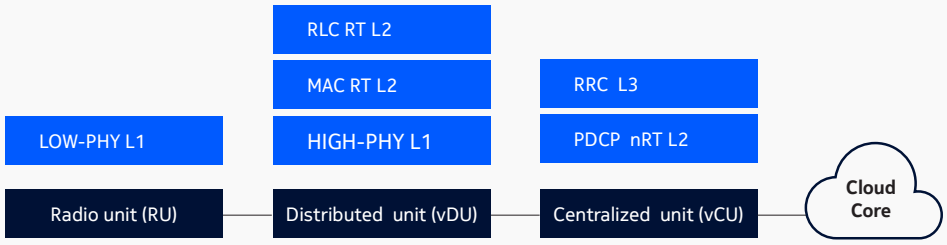
Hardware acceleration technology is a critical element in Cloud RAN performance. The two primary options for implementing Cloud RAN acceleration are the so-called "Look-Aside" and "In-Line" architectures. Early deployments of Cloud RAN were done based on the available Look-Aside acceleration solutions. In-Line acceleration solutions are rapidly gaining momentum and have been launched by several companies.

This paper objectively analyses the two main acceleration option categories and concludes that the overall benefits of In-Line acceleration outweigh those of Look-Aside for any commercial Cloud RAN deployment aiming for highest performance, lowest cost, lowest power consumption and greatest flexibility.

Cloud RAN hardware acceleration

Cloud RAN introduces the vertical disaggregation of the RAN baseband software from purpose-built baseband processing hardware. Open RAN represents the other dimension of disaggregation, splitting a base station horizontally to three parts; the RU, DU and CU, where the DU and CU are the baseband functions. The vertical disaggregation enables the use of COTS server hardware which can be purchased separately from suppliers other than the DU/CU software suppliers. Cloud-native DU/CU software and Container-as-a-Service (CaaS) software together with COTS hardware are expected to bring typical cloud computing benefits. The targeted benefits of fully cloud-native Cloud RAN include orchestration and automation of RAN functions and their management to yield operational savings, capacity elasticity to maximize the efficiency of computing resources usage, and the ability to introduce new network features and functionalities faster.

3GPP has functionally split the baseband unit into the real-time DU and the non-real time CU. In Cloud RAN, both are obviously virtualized and called the virtualized Distributed Unit (vDU) and the virtualized Centralized Unit (vCU).



In RAN, the vDU L1 processing is highly complex with sophisticated algorithms, including forward error correction (FEC), channel estimation, modulation, and layer mapping. These functions require specialized computing beyond the capabilities of general-purpose processors (GPP) used in COTS hardware, in order to cope with the extreme computing capacity needs of L1 and very tight latency requirements in end-to-end Mobile Networks.

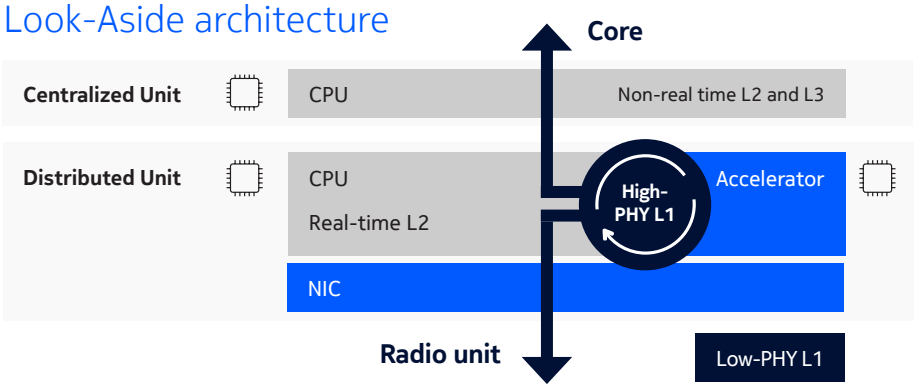
On the other hand, the real-time (RT) L2, non-real time (nRT) L2, and L3 workloads can be effectively computed on GPPs (such as x86 or ARM) generally used in COTS server hardware. The L2 and L3 software in vDU and vCU is simpler to disaggregate from the hardware than the L1 software that leverages specialized silicon to meet the necessary performance, cost, and power requirements.

Specialized hardware accelerators are, therefore, required for L1 processing in Cloud RAN vDUs. These can be implemented as either Look-Aside or In-Line architectures, operating in conjunction with COTS server hardware otherwise using GPPs.

Look-Aside and In-Line architecture options

In the Look-Aside architecture option (referred to by some as the Selected Function Hardware Accelerator), the general-purpose computing CPU acts as the master for processing L1, with selected key functions (e.g. FEC) sent back-and-forth to the hardware accelerator. The hardware accelerator can be either a separate Peripheral Component Interconnect Express (PCIe) card in the server, or be located on the same die (integrated "on-die") alongside the CPU in which case it is no longer a general-purpose processor (GPP). Such a "GPP" would carry cost and power consumption overhead in any other application use case and should not even be called a GPP, it is rather a custom SoC for Cloud RAN. In both of these Look-Aside acceleration cases, the CPU still processes many of the L1 real-time computations, for which it is inefficient, as well as the L2 and L3 processing for which the GPP is better suited.

Look-Aside architecture



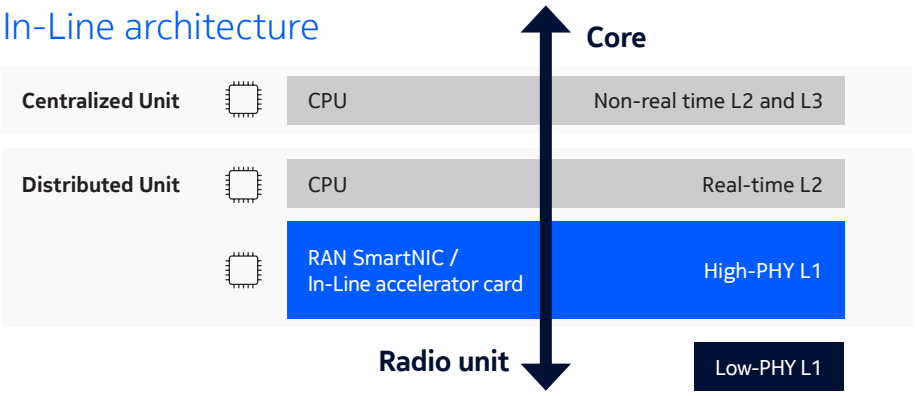
With the In-Line architecture option (referred to by some as the Full L1 Accelerator), all or part of the L1 processing is off-loaded from the CPU with a RAN SmartNIC PCIe card. SmartNICs (Network Interface Card) are today commonly used as accelerators in public and private cloud data centers[1]. In-Line acceleration SmartNICs use dedicated and optimized silicon technology for L1 processing and fully relieve the general-purpose CPUs (GPPs) from the ultra-high L1 processing demands.

This frees up valuable CPU resources enabling higher performance for L2 and L3 application processing. With an In-Line SmartNIC, less complex and less costly non-accelerated CPUs (GPPs) can be used for L2 and L3 processing where they are better suited.

[1] SmartNICs are commonly used in Cloud data centers to offload certain functions including networking, storage, and security functions from the server processor, freeing the general-purpose CPU to focus on application performance.

The In-Line architecture implements the best-of-breed technology with a specialized high-performance SmartNIC combined with the typical benefits of COTS server hardware. The efficiency, capacity and connectivity of this optimized solution is higher, and the power consumption lower than COTS server Look-Aside solutions can deliver.

In-Line architecture



The processing power and capabilities of both the In-Line and Look-Aside architecture options will increase over time. However, it is important to understand that L1 processing demands will increase significantly due to higher radio capacity needs, as well as due to lower latency requirements, in evolution from 5G to 5G-Advanced and towards 6G.

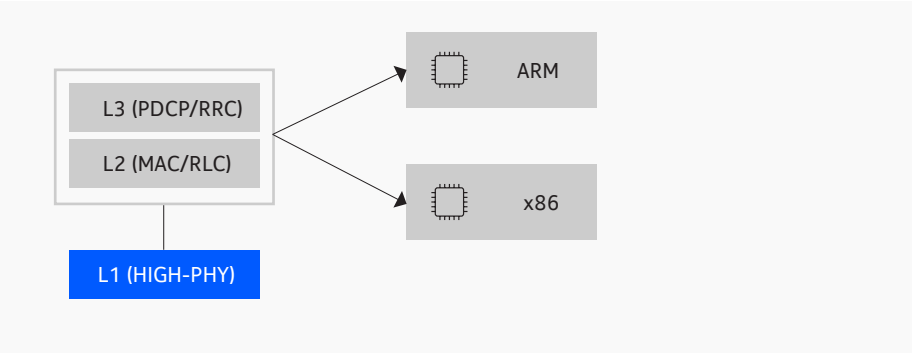
System scalability

The L1 and L2/L3 capacity requirements scale along different vectors. L1 scales based on the used resources on the air interface (e.g. bandwidth), whilst L2 and L3 scale largely based on the number of users, connections, and traffic.

With In-Line acceleration, L1 capacity can be increased by adding, in a targeted and cost-efficient manner, SmartNICs independently from the CPUs (GPPs) that are processing L2/L3. And vice-versa, CPU (GPP) capacity can be increased independently from In-Line SmartNIC capacity. With Look-Aside acceleration, all capacity enhancements require increasing the number of CPUs even if not all the CPU functions are uniformly needed for the processing of the different software layers (L1, L2, L3).

System Architecture

In-Line acceleration simplifies the entire architecture with a clean L1-L2 interface. The O-RAN Alliance aspires to open the interface between L1 and L2. In-Line SmartNICs use the standard PCIe interface and integrate with all compute and cloud platforms, with the optionality of x86 or ARM processors for disaggregated L2 and L3 processing. This provides great overall solution design flexibility and a greater choice of cloud server hardware providers.



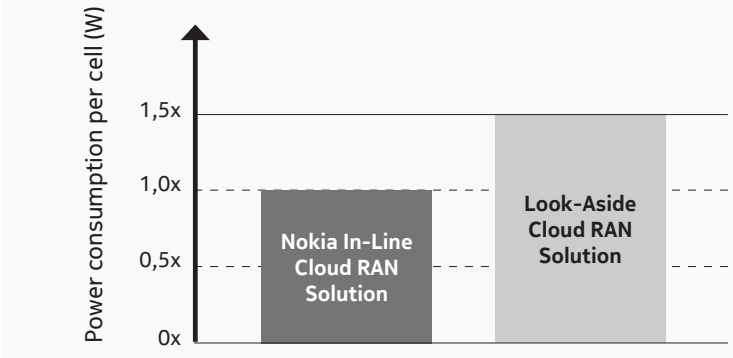
With the In-Line accelerator handling all the L1 compute, as well as the fronthaul interface between DU and RU, changes in the fronthaul architecture and specifications (e.g. in the specifics of the RU-DU fronthaul split) can be at least partly managed by the SmartNIC, thereby isolating, and reducing the changes needed in the higher layers running on COTS server hardware. Also, the fast two-to-three-year renewal and refresh cycles of typical COTS server hardware are avoided for the SmartNIC that follows rather the rhythm and lifecycle of mobile network developments with subscribers in mind.

Energy Efficiency

It is widely recognized that for highly demanding workloads such as L1 processing, purpose-built silicon technology provides higher performance and is also more energy efficient.

Typically, In-Line solutions use energy efficient ARM-based silicon that is also commonly used in Classic RAN networks and increasingly across all networks, including webscalers' cloud data centers.

A Cloud RAN (vDU+vCU) configuration based on technology available in 2023 and included in roadmaps for the future has been benchmarked for a high-capacity Cloud RAN configuration. The result is a significantly lower power consumption per cell with an In-Line solution compared to a Look-Aside solution. This translates into a lower cost per cell for the In-Line solution, in addition to the aforementioned capacity and efficiency benefits.



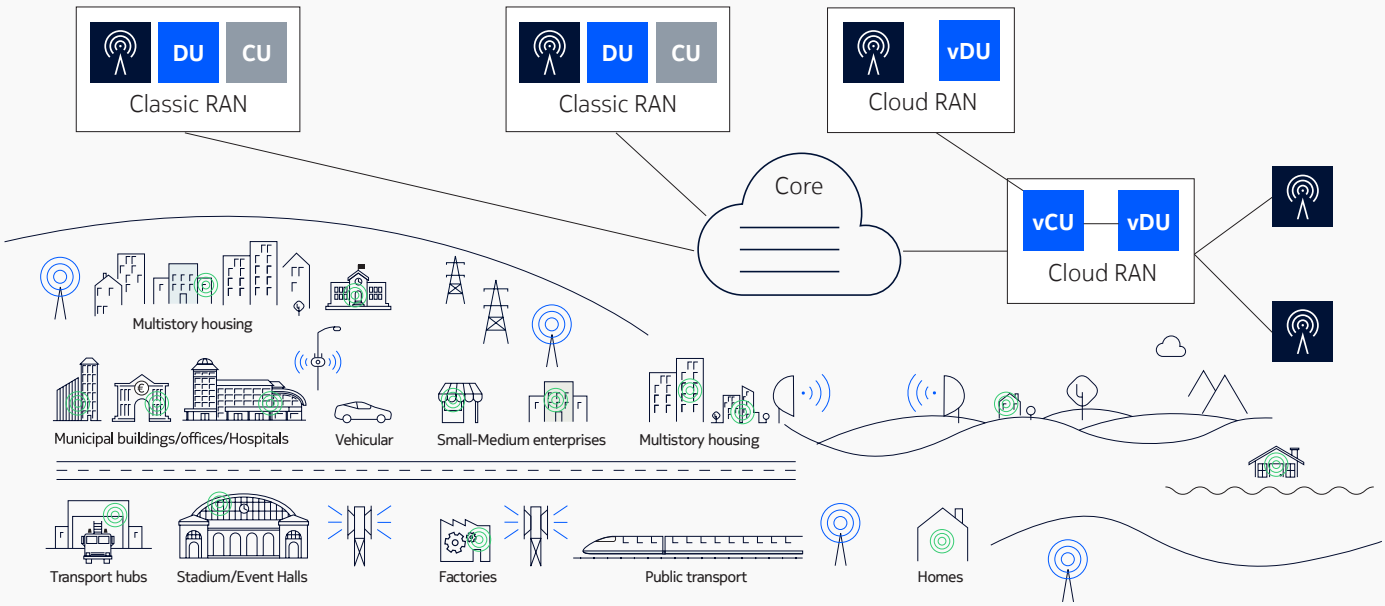
In-Line Cloud RAN with Nokia SmartNIC
COTS server for L2/L3 with Nokia SmartNIC L1 In-Line accelerator.

Look-Aside Cloud RAN
COTS server for L1/L2/L3 with NIC for the Fronthaul.

Cloud RAN and Classic RAN in mobile networks

RAN networks are expected to evolve gradually towards Cloud-based RAN solutions. According to Dell'Oro, Cloud RAN solutions have the potential to make up approximately 20 percent of the total RAN market by 2027[2]. Volume deployments are expected to start gradually accelerating in 2024-2025. While the limited number of greenfield operators can build their networks from a clean slate based on Cloud RAN technology, all other operators will choose to adopt a hybrid network approach so Cloud RAN will have to co-exist with Classic RAN for many years. This hybrid mobile network will evolve over time with Cloud RAN being introduced for specific use cases and/or geographies.

For example, most CSPs will focus their Cloud RAN deployments on high-capacity configurations with high performance and energy efficiency requirements that are best fulfilled with the In-Line architecture option. Further, the high-capacity configurations can best benefit from geographical centralization of vCUs.



[2] Dell'Oro Mobile RAN Five Year Forecast Report 2022-2027 (January 2023)

In-Line accelerated Cloud RAN architecture provides the unique opportunity to use common L1 System-on-a-Chip (SoC) technology for both Cloud RAN and Classic RAN. This brings feature and performance parity with the same software release management cadence, providing consistent end-user experience across the whole mobile network and enabling an optimal network design encompassing the co-existence of Cloud RAN and Classic RAN.

Look-Aside Cloud RAN can, of course, also be deployed alongside Classic RAN. However, it requires different L1 software. This leads to parallel development and challenges in reaching feature parity with different product releases, ultimately leading to cumbersome roll-out constraints with inconsistent end-user experience and a higher total cost of ownership.

Cloud-nativeness

The In-Line SmartNIC with cloud native vDU and vCU software supports the commonly accepted cloud-nativeness principles including, among others, data being decoupled from the applications, loosely coupled microservices, elastic and horizontal scaling, automated lifecycle management, and continuous software integration and delivery. For example, it supports the scalability and flexible deployment of containerized vDU application functions to various nodes managed by the container orchestration platform, such as Kubernetes.

Cloud efficiency is higher with the In-Line accelerated Cloud RAN architecture as it relaxes the latency requirements on the Containers-as-a-Service (CaaS) layer running on the CPU (GPPs). This leads to savings on the real-time features of CaaS.

The In-Line accelerator will be in most, if not all cases, a SmartNIC. As described earlier, SmartNICs are commonly used today as accelerators in server and cloud infrastructure environments, including those of webscalers. Whilst SmartNICs have their own specific drivers and management operations, there are common procedures and methods for integrating and managing SmartNICs in scalable server and cloud environments.

For example, Kubernetes operators and plug-ins[3] manage the RAN SmartNIC with the cloud-native tools used for software deployment and for managing the cloud environment in which the RAN SmartNIC resides.

[3] Device Plugin: https://kubernetes.io/docs/concepts/extend-kubernetes/compute-storage-net/device-plugins/

Conclusion

To conclude, In-Line acceleration with a Cloud RAN SmartNIC is seen to be most suitable for all RAN deployments, including the highest capacity mobile networks, whilst Look-Aside acceleration may be more suitable for lower capacity, higher latency mobile networks, if energy efficiency is of little concern.

The advantages of In-Line acceleration technology extend to enabling feature and performance parity with Classic RAN, higher capacity and connectivity, better energy efficiency, lower comparable TCO, ease of integration into any server and cloud environment, and flexibility in choosing higher layer processing computing architectures and associated server hardware providers.

There is a healthy competitive environment for commercial In-Line SmartNICs with several suppliers driving continuous product improvements, cost benefits and innovations. By comparison, there is no competition within the Look-Aside technology development, which typically brings risks of excessive supplier market power and roadmap dependencies.

The ideal Cloud RAN solution supports a choice of technology suppliers and operating environments. With an In-Line RAN SmartNIC, and truly cloud native vDU and vCU software, a Cloud RAN solution can operate with any mainstream CaaS layer, server maker and webscaler cloud environment. The various deployment preferences of Cloud RAN customers can thus be fulfilled very well by a solution based on the In-Line acceleration architecture.

Since 2017, Nokia has been at forefront of developing Cloud RAN technology solutions with our customers in close collaboration with our semiconductor, server maker, hyperscaler and managed cloud service partners.

Visit Nokia AirScale Cloud RAN page to find out more.

Glossary

3GPP	The 3rd Generation Partnership Project
CaaS	Containers-as-a-Service
COTS	Commercial-off-the-shelf
CPU	Central Processing Unit
CSP	Communications Service Provider
CU	Centralized Unit
DU	Distributed Unit
eCPRI	Enhanced Common Public Radio Interface
FEC	Forward Error Correction
GPP	General Purpose Processing
HIGH-PHY	High Physical Layer
L1	Layer 1
L2	Layer 2
L3	Layer 3
LOW-PHY	Low Physical Layer
MAC	Medium Access Protocol
NIC	Network Interface Card
nRT	non-Real Time
PCIe	Peripheral Component Interconnect Express
PDCP	Packet Data Convergence Protocol
RAN	Radio Access Network
RLC	Radio Link Control
RRC	Radio Resource Control
RT	Real-Time
RU	Radio Unit
SoC	System-on-a-Chip
vDU	Virtualized Distributed Unit
vCU	Virtualized Centralized Unit

Nokia OYJ
Karakaari 7
02610 Espoo
Finland

Tel. +358 (0) 10 44 88 000

CID: 213050

nokia.com

At Nokia, we create technology that helps the world act together.

As a B2B technology innovation leader, we are pioneering the future where networks meet cloud to realize the full potential of digital in every industry.

Through networks that sense, think and act, we work with our customers and partners to create the digital services and applications of the future.