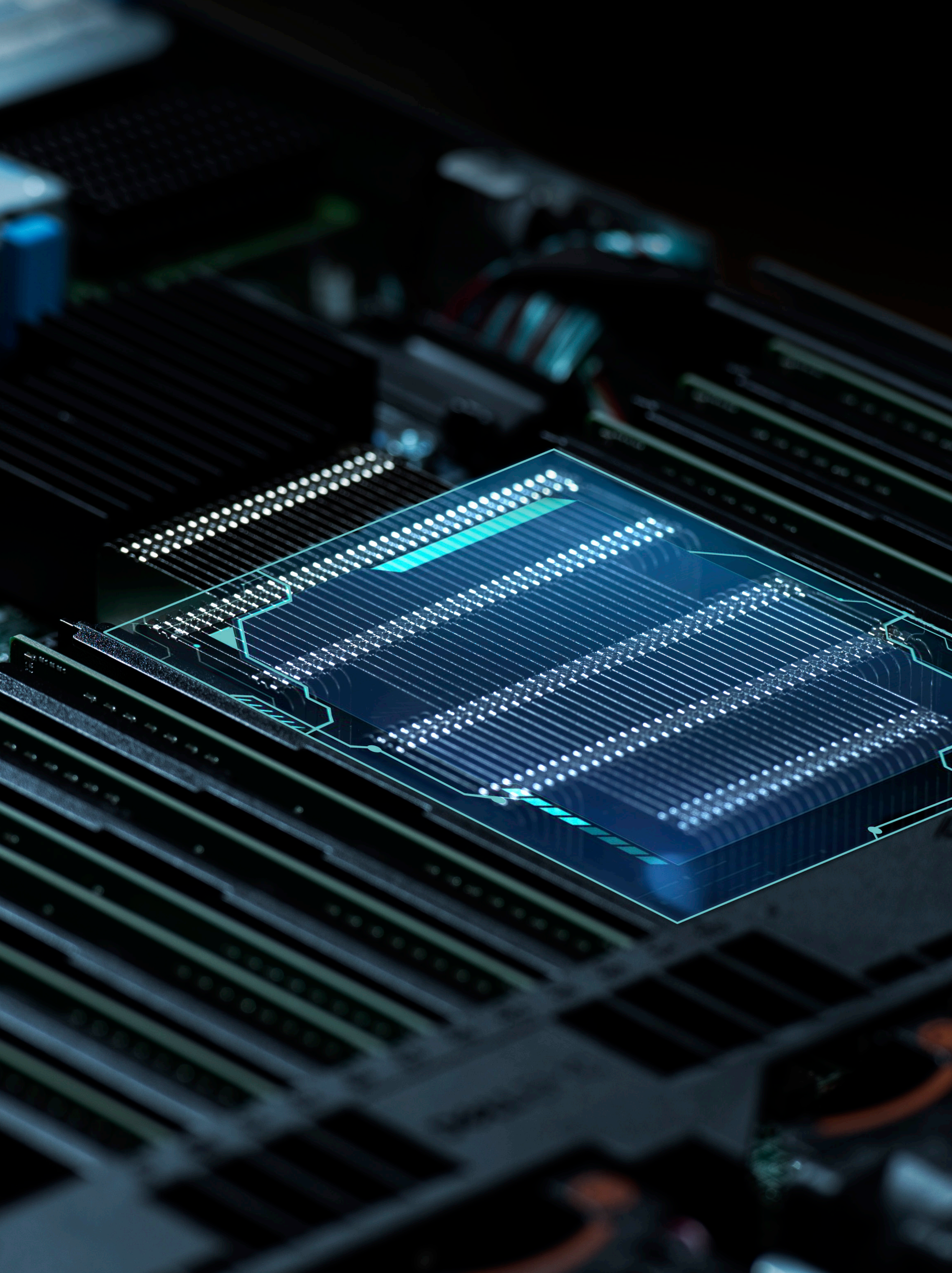
Cloud RAN Acceleration Technology

ERICSSON

verizon



Executive summary

After years of development, collaboration, and evolution, the path to the future of Cloud RAN* is becoming clear. It will be a combination of a common Cloud infrastructure supporting multiple workloads and best-in-breed software that is portable, flexible, and scalable.

Where acceleration takes place has been the subject of some discussion, essentially coming down to a choice between having a Selected Function Hardware Accelerator (sometimes known as 'look-aside' acceleration) and a Full Layer 1 Accelerator (in-line acceleration). Selected Function Hardware Acceleration is evolving into an integrated architecture, while Full Layer 1 Acceleration takes place on a separate card.

Considering the needs for energy efficiency, design flexibility, portability, and support for an ecosystem of Cloud RAN suppliers on common cloud infrastructure, Selected Function Hardware Acceleration currently offers the best option to build high-performing Cloud RAN networks.

Introduction

There is an increased interest in virtualization and cloud-native technologies in 5G Radio Access Networks (RAN) and beyond to meet the diverse and varied needs for more open, resilient, sustainable, and intelligent mobile networks.

Operators have several key needs that their Cloud RAN architecture must cater to. The chief among them are performance and efficiency and the disaggregation of hardware from software.

What is also desirable is support for an ecosystem of Cloud RAN suppliers who can offer their differentiated algorithms and product variants on a common cloud infrastructure, which simplifies deployment and operations wherever possible. This minimizes integration complexity and ensures achieving performance targets.

Another benefit comes from the ability to create more unified and common operations models across all network elements and vendors and to add an increased level of automation of network operations, thereby optimizing the total cost of ownership and performance for different deployments.

Overall, Cloud RAN offers the potential that operators increasingly require to run high-performing networks that are flexible, agile, and reliable.

Content

*Sometimes referred to as VRAN

1 About Cloud RAN

At its core, Cloud RAN disaggregates the RAN compute baseband software from the hardware, delivering the corresponding functionality through software running on commercial off-the-shelf (COTS) hardware. Cloud-native tools and processes are used to manage both software and hardware, with the software ideally running on any suitable COTS hardware, with or without integrated accelerators for improved performance, and with the goal of maximizing the portability of software on a range of different hardware.

This means that RAN software should be capable of being deployed in many different ways. It could be on cloud hardware on-site in what is called a Distributed RAN (D-RAN), or in a data center owned or leased by a Mobile Network Operator (MNO) to form a Centralized RAN (C-RAN) architecture.

Different parts of the RAN software stack have different requirements when it comes to processing and the time-critical nature of some elements, leading to a discussion as to which way forward is best. Within the compute platform of Cloud RAN, some of the most demanding computation acceleration is carried out by specialized hardware that accelerates compute-intensive functions.

In both the C-RAN and D-RAN architecture the DU and CU processing are making up the total Cloud RAN solution. To understand the processing requirements for achieving the best performance, we must understand the 5G RAN protocol stack. Of particular importance is the separation of the upper and lower parts of the RAN, where a higher-layer split is specified with a well-defined interface (F1) between two logical units - the centralized unit (CU) and the distributed unit (DU) (see figure 3).

Different parts of the RAN software stack have different requirements when it comes to processing and the time-critical nature of some elements, leading to a

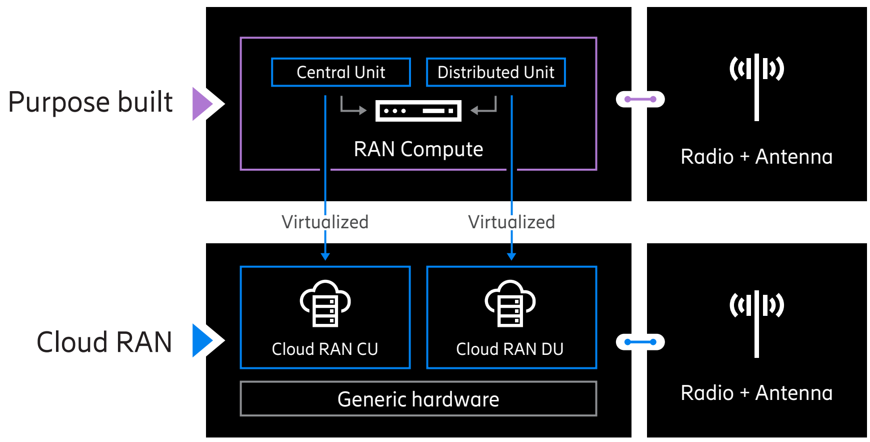


Figure 1

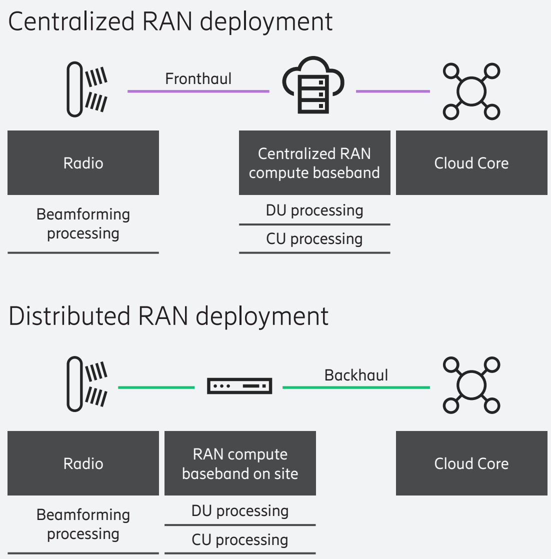
discussion as to which way forward is best. Within the compute platform of Cloud RAN, some of the most demanding computation acceleration is carried out by specialized hardware that accelerates compute-intensive functions. The lower you go in the protocol stack, the higher the demand on processing. Layers 1 and 2 combined comprise 90 percent of processing demand (see figure 3).

As a result, we must carefully consider and plan how to deploy the individual parts of the stack. This needs to be done from a hardware perspective to decide where the processing is located in the network and how it is performed.

The ultimate benefit of Cloud RAN is the flexibility and scalability it provides to MNOs. The wide array of deployment options allows operators to choose hardware and infrastructure that best suit their needs, budget, and business model. Choosing the right architecture and configuration of hardware and acceleration can help MNOs reap the full benefits of Cloud RAN. This creates conditions to use software from the industry's leading RAN solutions and match the performance of purpose-built hardware (RAN compute baseband) and software deployments.

Much has been learned in the process of developing such an infrastructure.

One of the main challenges is to address the high compute requirements in Layer 1 and Layer 2. The lower you go in the protocol stack, the higher the demand on processing. The wide carrier bandwidth of 5G mid and high bands together with massive MIMO technology exponentially increase the processing demand on Layer 1 and beamforming. In order to address this processing, acceleration technology is required. The options for this technology are discussed in the next section.

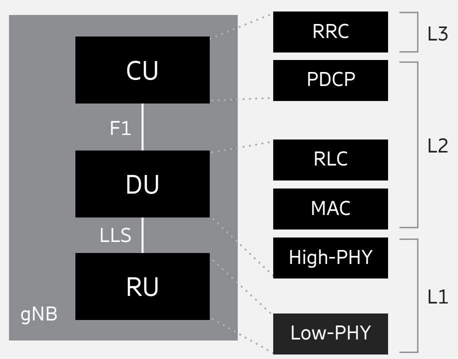


Centralized RAN vs. Distributed RAN

Commonly referred to as C-RAN, Centralized RAN is an alternative to today's most common deployment option, Distributed RAN (D-RAN). In C-RAN, a large part of the RAN processing is done at a common C-RAN hub for multiple antenna sites, compared with D-RAN, in which the majority of RAN processing is done locally at the antenna site.

In a C-RAN architecture the digital unit processing is done at the C-RAN Hub with a fronthaul interface to the radio as can be seen in figure 2. In D-RAN the fronthaul interface sits within the antenna site with a transport backhaul to the core network. The ability to run DU Processing at the C-RAN hub requires strict latency, jitter and bandwidth requirements.

Figure 2



Function	Subfunction	Main processing function
CU	Control plane (CU-C)	Radio Resource Control
	User plane (CU-U)	Packet Data Convergence Protocol
		Higher Layer Split (HLS / F1)
DU	Digital Unit	Radio Link Control
		Medium Access Protocol
		Physical layers
		Lower Layer Split (LLS)
RU	Radio Unit	Beamforming (if applicable)

Figure 3

2

Acceleration technology to meet Cloud RAN requirements

For extremely repetitive functions in Layer 1, the physical layer of the Radio Access Network, that are well-structured and require continuous processing, acceleration is needed with the goal of using the central processing unit (CPU) capacity for other complex RAN operations, thus enhancing performance.

In the case of Cloud RAN implementation, for example, one of the functions that place the greatest demand on processing power is L1 FEC (forward error correction, an error correction technique to detect and correct a limited number of errors in transmitted data from a user without the need for retransmission).

The L1 FEC could be implemented on a standard CPU with optimized software. However, in a high-capacity system, it would require a substantial amount of computing resources, and as a result2, it is highly desirable to offload that computing to a hardware accelerator. This can be seen as a similar approach to previous computing evolutions when floating point accelerators were introduced to offload a CPU.

As depicted in the block diagrams (Figures 5 and 6), there are two major acceleration technologies – the Selected Function Hardware Acceleration and the Full L1 Acceleration. Each has its unique approach to addressing the engineering challenges.

For operators, the implications of accelerator choice on the overall solution need to be considered. Energy efficiency, programmability to enable multi-vendor ecosystems on common infrastructure, minimizing integration complexity, and reducing the total cost of ownership (TCO) are key factors. These factors will now be further explored.

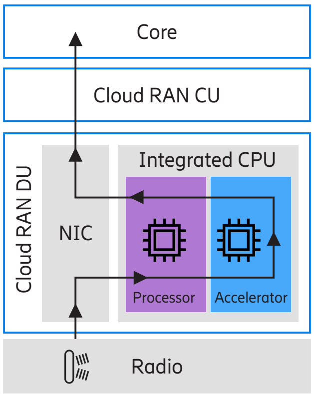


Figure 4

Selected Function Hardware Accelerator

Selected Function Hardware Accelerator (sometimes called "look-aside")
Using a Selected Function Hardware Accelerator leaves the CPU free to use its cycles to process other useful tasks while the accelerator is working on the selected functions. Once the CPU receives processed data back from the accelerator, it can switch back to the original processing context and continue the pipeline execution until the next function to be accelerated comes up. Selected Function accelerators require well-defined APIs to enable ecosystem adoption. To minimize data transfer, acceleration can be integrated within the CPU chip, as shown in the diagram.

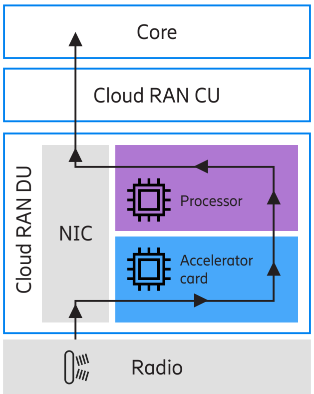


Figure 5

Full L1 Accelerator

Full L1 Accelerator card (also known as "inline")
In the Full L1 Acceleration case, some or all of the Layer 1 pipeline can be offloaded to the accelerator card, potentially allowing for a less data-heavy interface the between CPU and the accelerator card. This acceleration solution can in this case be a mix of programmable and "hard" blocks – again, there is a trade-off between flexibility and efficiency.

Energy efficiency

Energy efficiency is one of the key metrics for network performance, and it is generally measured in terms of the amount of system bandwidth processed relative to system energy consumption.

Both acceleration technologies are likely to improve as semiconductor technology evolves, but one major difference remains:

A Full L1 Accelerator card can save CPU core consumption, however, it requires a separate PCIe Card to be inserted in the server, which creates considerably more power consumption than a standard network interface card (NIC).

With Selected Function Hardware Accelerator, there is an opportunity for application design to use a larger pool of available CPU cores efficiently; with tighter integration of selected accelerators within the CPU, the potential for further efficiency improvements.

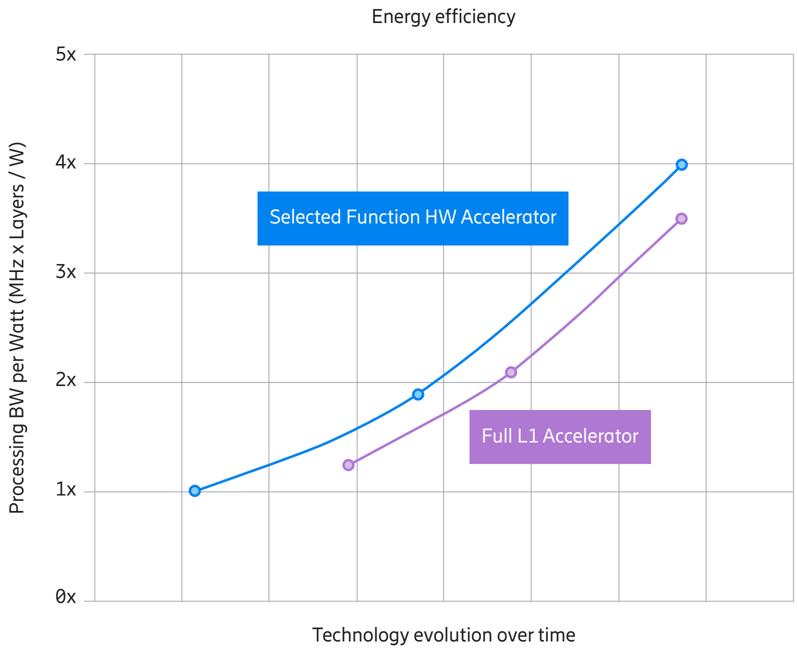


Figure 6 The 1x baseline is based on measurements on a DU workload supporting 600 MHz x 4 Layers on a single CPU with Selected Function HW Accelerator

To fairly evaluate the energy efficiency of a given technology, we have devised the following methodology

Server power: Considering the power consumption of all the hardware components, within the server, including the L1 card/chip

Processing bandwidth: Calculate the processing bandwidth in terms of max bandwidth multiplied by layers that can be processed

Energy-efficiency is calculated as the ratio between the processing bandwidth and the server power

Design flexibility and portability
Design flexibility is essential to the Cloud RAN concept, which disaggregates the software from the hardware. With the Selected Function Hardware Accelerator approach, only selected L1 RAN functions with well-defined algorithms and APIs are accelerated. This enables the Cloud RAN software provider to deliver best-in-class and differentiated algorithms, which is key to enabling innovation and high-performance solutions with coherent feature sets. They also co-exist on a common cloud infrastructure, which is critical for reducing operator integration complexity.

The common nature of the infrastructure and a familiar programming environment also make it easier to onboard multiple application suppliers onto the common hardware infrastructure.

A Full L1 Accelerator card requiring software specifically developed for that specific hardware component which will make disaggregation more challenging. Such cards often require software specifically developed for their hardware or chipset, which eliminates the possibility to create a common cloud compute infrastructure across the network and increases the risk of fragmentation.

Additionally, in cases where L1 software is provided by the accelerator card supplier, the industry would need to take on added integration complexity to maximize the benefit of their features and to be able to diagnose operational issues as they arise.

As such, using Full L1 Accelerator cards poses a challenge in terms of upholding some of the key principles in optimal Cloud RAN scenarios – namely, the strong desire for portability and flexibility.

Cloud nativeness
A key benefit of adopting Cloud RAN is the ability to employ common operational systems and practices across the network, simplifying deployment and the life cycle management (LCM) of resources.

Cloud-native applications require network services to be abstracted and easy to use without any platform dependencies, while delivering higher throughput and lower latencies. In Selective Function Hardware Acceleration, only the most compute-intensive, well-defined and latency-sensitive operations are accelerated, while the rest of the L1 functions are coded in the software application.

In order to maximize the potential of a platform-agnostic design that offers greater flexibility, this soft implementation also benefits from cloud-native principles like microservices architecture for easy life-cycle management, scalability and orchestration. The Selected Function Hardware Accelerator is abstracted using the standardized open-source Wireless Baseband Device Library (BBDev) framework (see figure 7) which also enables software portability across CPU variants.

In contrast, when using the Full L1 Accelerator card, the hardware on-boarding, configuration, management, abstraction and the accelerator software life cycle management using existing orchestration tools are some of the areas that still need development. The need for card-specific or application-specific software with Full L1 Acceleration runs counter to the Cloud RAN principles of portability and flexibility, effectively tying the MNO to a particular solution.

Considering these three aspects – the needs for energy efficiency, design flexibility and portability and the preference for cloud-native technologies - it is apparent at this stage of the technological evolution that the Selected Function Hardware Accelerator is a better option to deliver on Cloud.

BBDev

BBDev is a set of libraries and drivers specific for BB applications. It uses the Data Plane Development Kit DPDK networking libraries and drivers which, among other things, implement an abstraction layer between the OS & HW and the application.

DPDK and BBDev are industry de-facto standards (and adopted in O-RAN for FEC acceleration)

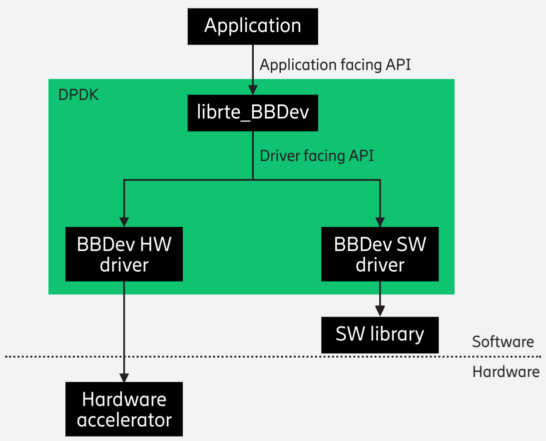


Figure 7

3 The future is in the cloud

There is enormous potential to roll out and leverage the benefits of Cloud RAN nationwide, and the nature of Cloud RAN necessitates a broader ecosystem approach if such rollouts are to be successful in the long term. Partnerships and standard collaborations spur development, ultimately maximizing the benefits as cloudification gets introduced from 5G towards the 6G era.

Strong partner ecosystems, such as those being pursued by Verizon and Ericsson, will drive and shape the best technology solutions in terms of standardization, integration, and security. This allows customers to select from the widest possible range of hardware and software options within the existing framework.

Essentially, the goal of operators is to maintain a high degree of flexibility, reduce time to market (TTM) for new services, and improve capacity efficiencies faster than can be done with customized hardware and software. This entails balancing solution flexibility and design efficiency in a way that not only supports a broad software ecosystem on the common compute platform, but also works across different CPU versions in the ecosystem. These needs are currently best met by the combination of best-in-breed software operating on a common infrastructure with a Selected Function Hardware Accelerator.

Ericsson and Verizon have concluded that, based on the current technology landscape, common Cloud infrastructure with Selected Function Hardware Acceleration currently offers the best path forward to deliver on energy efficiency, ecosystem support and flexibility at reduced complexity.

Abbreviations

a.k.a	Also Known As
API	Application Programming Interface
BBDev	Wireless Baseband Device
BW	Band Width
CI/CD	Continuous integration continuous delivery
COTS	Commercial of-the-shelf
CPU	Central Processing Unit
C-RAN	Centralized RAN
CU	Centralized Unit
CU-C	Centralized Unit - Control plane
CU-U	Centralized Unit - User plane
DevOps	Development and Operations
DPDK	Data Plane Development Kit
D-RAN	Distributed RAN
DU	Distributed Unit
FEC	Forward Error Correction
gNB	Next Generation Node B
High-PHY	Higher Physical layer
HLS	Higher Layer Split
HW	Hardware
L1	Layer 1
L2	Layer 2
LCM	Life Cycle Management
BBDev	wireless BaseBand Device library
LLS	Lower Layer Split
Low-PHY	Lower Physical layer
MAC	Medium Access Control
MHz	Mega Hertz
MNO	Mobile Network Operator
NIC	Network Interface Card
OS	Operating System
PDCP	Packet Data Convergence Protocol
RAN	Radio Access Network
RLC	Radio Link Control
RRC	Radio Resource Control
RU	Radio Unit
SW	Software
TTI	Transmission Time Interval
W	Watt

About Ericsson

Ericsson enables communications service providers and enterprises to capture the full value of connectivity. The company's portfolio spans the following business areas: Networks, Cloud Software and Services, Enterprise Wireless Solutions, Global Communications Platform, and Technologies and New Businesses. It is designed to help our customers go digital, increase efficiency and find new revenue streams. Ericsson's innovation investments have delivered the benefits of mobility and mobile broadband to billions of people globally. Ericsson stock is listed on Nasdaq Stockholm and on Nasdaq New York.

www.ericsson.com

About Verizon

Verizon Communications Inc. (NYSE, Nasdaq: VZ) was formed on June 30, 2000 and is one of the world's leading providers of technology and communications services. Headquartered in New York City and with a presence around the world, Verizon generated revenues of $133.6 billion in 2021. The company offers data, video and voice services and solutions on its award-winning networks and platforms, delivering on customers' demand for mobility, reliable network connectivity, security and control.

www.verizon.com

EN/LZT 3/28701-FGB1010990