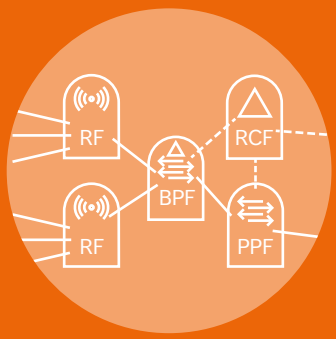
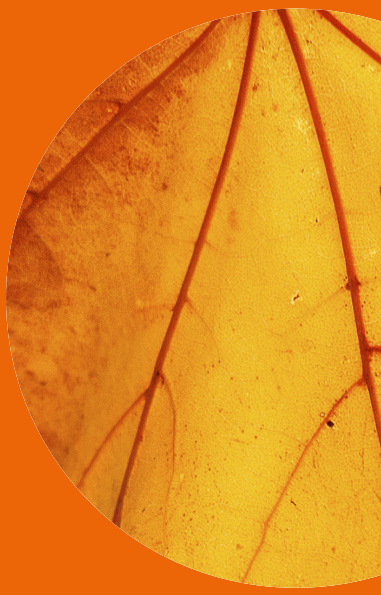
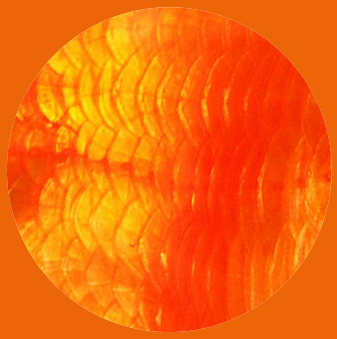
CHARTING THE FUTURE OF INNOVATION VOLUME 93 | #6·2016

ERICSSON TECHNOLOGY Review





4G/5G RAN ARCHITECTURE: HOW A SPLIT CAN MAKE THE DIFFERENCE



ERICSSON

4G/5G RAN architecture

HOW A SPLIT CAN MAKE THE DIFFERENCE

As 5G evolves, many innovative services like extreme mobile broadband and long-range massive MTC will come into play. In line with the evolution of 4G and the introduction of 5G, RAN architecture is undergoing a transformation: to increase deployment flexibility and network dynamicity, enabling networks to meet increasing performance requirements, while at the same time keeping a lid on the total cost of ownership. The proposed future-proof software-configurable split architecture will be able to support new services, deployed on general-purpose and specialized hardware, with functions ideally placed to maximize scalability, spectrum, and energy efficiency, all while supporting the concept of network slicing.

ERIK WESTERBERG

THE REQUIREMENTS created by extreme MBB, the IoT, and massive MTC call for an alternative to today's deployment architectures. The changes to architecture include the capability to place selected functions closer to the network edge, for example, and the ability to increase RAN resilience. Cost is naturally a factor, as spectrum availability and site infrastructure continue to dominate operator expenditure for wide-area systems. The evolution of RAN architecture therefore needs to include measures for enhanced spectrum efficiency that are harmonized with other improvements in the areas of hardware performance and energy efficiency.

In light of these cost and performance requirements, a number of capabilities are shaping the evolution path of RAN architecture:

Seamless Radio Resource Management

The best combination of any radio beam within reach of a user should be used for connectivity across all access network technologies, antenna points, and sites. This capability will be achieved by applying carrier aggregation, dual connectivity, CoMP, and a number of MIMO and beamforming schemes.

Functional split

Some 5G requirements — such as ultra-low latency and ultra-high throughput — require highly flexible RAN architecture and topology. This will be enabled by splitting RAN functions, including the separation of the user plane (UP) and the control plane (CP) in higher layers.

Dynamic and software-defined RAN

The capability to configure, scale, and reconfigure logical nodes through software commands enables the RAN to dynamically adjust to changing traffic conditions, hardware faults, as well as new service requirements. This capability will be achieved by separating out logical nodes suitable for virtualization (on a GPP) and designing functions that require specialized hardware to be dynamically (re)configurable on an SPP.

> THE CAPABILITY TO CONFIGURE, SCALE, AND RECONFIGURE LOGICAL NODES THROUGH SOFTWARE COMMANDS ENABLES THE RAN TO DYNAMICALLY ADJUST TO CHANGING TRAFFIC

Deployment flexibility

Deployment flexibility enables an operator to deploy and configure the RAN with maximum spectrum efficiency and service performance regardless of the site topology, transport network characteristics, and spectrum scenario.

This is achieved through a correct split of the RAN architecture into logical nodes, combined with the future-proof freedom to deploy each node type in the sites that are most appropriate given the physical topology and service requirements.

The process to reach the target architecture with the correct split involves a number of steps:

- determining the logical functions that comprise 5G RAN on a level below 3GPP
- identifying the latency-critical functions that need to be placed within a few TTIs to antenna elements
- identifying which functions have more relaxed latency requirements

Terms and abbreviations

BPF — baseband processing function | **CO** — central office | **CoMP** — coordinated multipoint | **CP** — control plane | **CPRI** — Common Public Radio Interface | **C-RAN** — cloud RAN | **DL** — downlink | **EPC** — Evolved Packet Core | **E2E** — end-to-end | **GPP** — general purpose processor | **HARQ** — hybrid automatic repeat request | **IoT** — Internet of Things | **MAC** — media access controller | **MBB** — mobile broadband | **MIMO** — multiple-input, multiple-output | **MTC** — machine-type communications | **NFV** — Network Functions Virtualization | **NR** — next generation RAT | **OSS** — operations support systems | **PDCP** — packet data convergence protocol | **PDU** — protocol data unit | **PGW** — packet data network gateway | **PHY** — physical interface transceiver | **PPF** — packet processing function | **RAT** — radio-access technology | **RCF** — radio control function | **RDC** — regional data center | **RF** — radio function | **RLC** — Radio Link Control | **RRM** — Radio Resource Management | **SON** — self-organizing networks | **SPP** — special purpose processor | **S-RRM** — server RRM | **TTI** — time transmit interval | **UE** — user equipment | **UL** — uplink | **UP** — user plane | **U-RRM** — user RRM | **VNF** — Virtualized Network Function

- determining where to place anchor points for soft combining, carrier aggregation, and dual connectivity — among the user-plane functions
- identifying which nodes can be implemented as VNFs

This process is illustrated in *Figures 1*, *2*, and *3*. *Figure 1* shows the 4G/5G logical architecture at a level below 3GPP; *Figure 2* shows today's 4G split into an RU and a DU; and *Figure 3* shows the target split architecture. Throughout the process, and as a result of the functional decomposition, new inter-node interfaces emerge, whose characteristics need to be taken into consideration to ensure that the underlying transport network can support the various deployment scenarios.

The logical 4G/5G RAN architecture

The external interfaces of the RAN domain (except to the OSS) are standardized under 3GPP, as is the functional behavior of the RAN domain as a whole. Below the high-level specification, 3GPP leaves room for innovation to enhance the network with RAN-internal value-add features — a flexibility that has over a number of years resulted in continuous improvement in many areas, including spectrum efficiency (in the form of scheduling algorithms, power control algorithms, and various RRM features), energy efficiency, and enhancements to service characteristics such as lower latencies. To determine the optimal architectural split, however, the RAN architecture needs to be examined with a finer level of granularity than that offered by 3GPP.

RAN anchor points

Figure 1 illustrates the logical RAN architecture, which for the purposes of simplification shows the UL and DL instances of each function combined, and the solid lines indicate user plane functions. In the downlink, PDUs enter the RAN domain over the S1-U interface (right) and are delivered to devices (on the left) over the radio interface. The multipath-handling function is the anchor point for

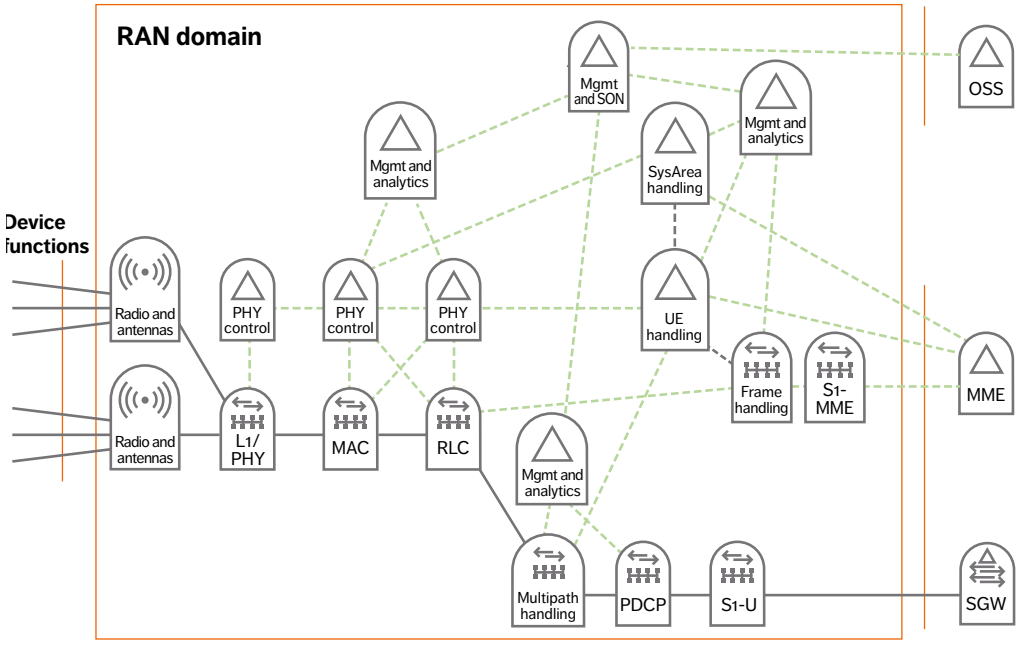


Figure 1: The logical 4G/5G RAN architecture — one level below 3GPP

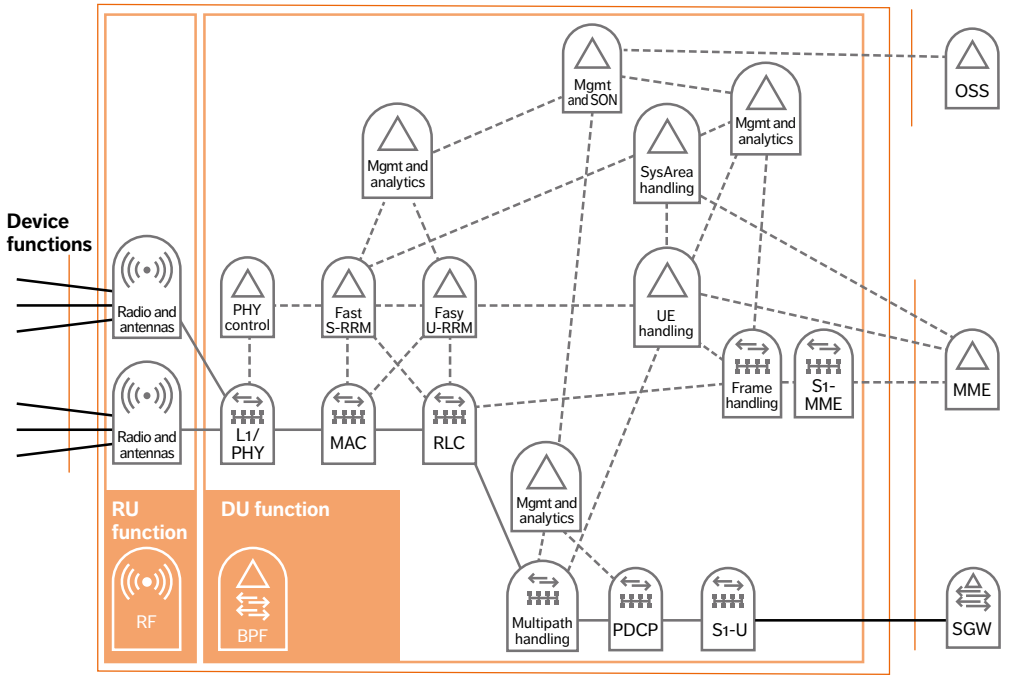


Figure 2: Logical RAN architecture — present split into an RU function and a DU function

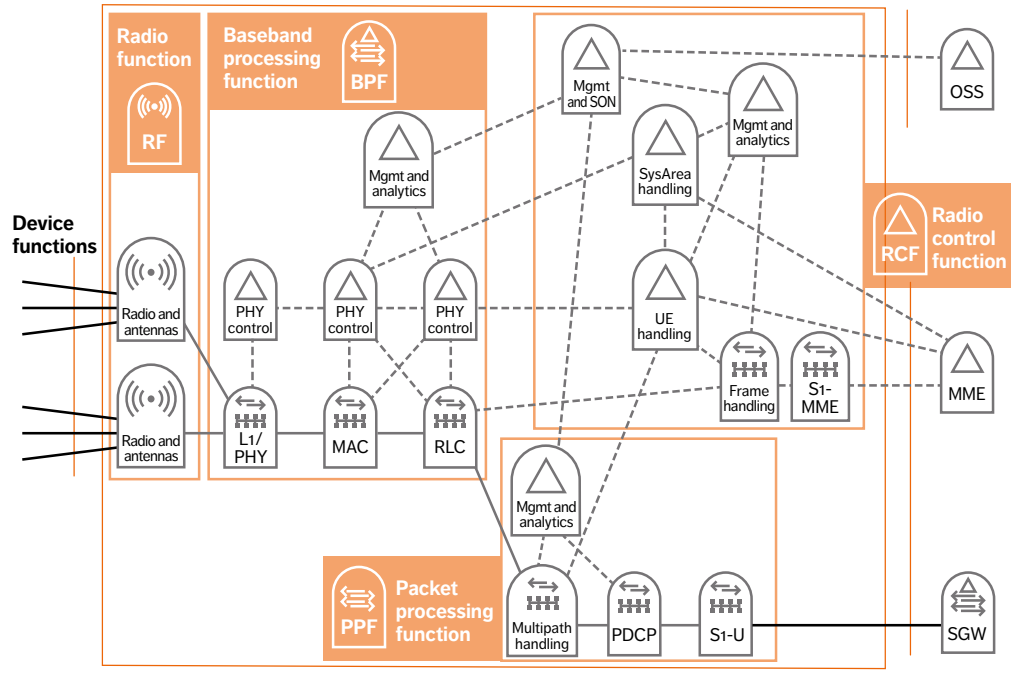


Figure 3: Logical RAN architecture — target split into the functions RF, BPF, PPF, and RCFt

dual connectivity, which schedules individual PDCP PDUs in the same user-data stream to different RLC instances, possibly on different RATs. In this way, a single UE can simultaneously receive and send data over different radio channels — for example, one NR and one LTE channel — that are connected to different sites. The MAC function is the anchor point for carrier aggregation, which schedules MAC PDUs to each user over a multitude of 4G or 5G carriers. The MAC function handles CoMP and multi-beam transmissions. In the uplink, the L1/PHY function performs soft combining, the MAC function aggregates UL data in carrier aggregation, and multipath handling aggregates data received from dual connectivity UL data streams.

The HARQ loop

3GPP specifies retransmission periods and response times at the radio level between the UE and the network as the HARQ loop, which includes: the air-interface transmission time, completion times for RF-L1/PHY-MAC functions, and RF-L1/PHY-MAC functions in the UE.

The loop results in a standardized round-trip latency budget of 3ms for LTE, and down to 200µs for NR. The RAN functions participating in the loop work synchronously with the air-interface TTI, while the PDCP and multipath handling function feed and receive packets traveling to and from the RLC layer asynchronously — which has implications for the split architecture.

Control plane functions

Runtime control functions can generally be divided into three categories, depending on whether they: act on a per-user basis (U-RRM), control spectrum on a system level (S-RRM), or manage infrastructure and other common resources.

The U-RRM functions include measurement reporting, selection of modulation and coding schemes, per-UE bearer handling, and handover execution. The functions in fast U-RRM act within the HARQ loop and serve the scheduler with processed real-time per-UE information. In contrast, the U-RRM function (UE-handling) works on a time scale of 10ms and above, including bearer handling, per-UE policy handling, handoff control, and more.

Functions that control spectrum on the system level include radio scheduling, distribution of the power budget across active UEs, and system-initiated load-sharing handovers. The fast S-RRM operating within the HARQ loop is responsible for radio scheduling and functions in tight coordination with the MAC and RLC. System-area handlers — such as load sharing, system information control, and dual-connectivity control — control spectrum on a 10ms time scale, or slower.

Functions that control infrastructure and common resources — other than spectrum — include handling of transport, connectivity, hardware, and energy. By allowing the control functions for spectrum, transport, infrastructure, and connectivity to interact, a holistic control system for RAN resources can be built.

Interface characteristics

The RAN HARQ loop time budget of three TTIs is divided into time for processing, and time for signals and data to traverse the various inter-function interfaces. The less time spent on interface signaling, the more time is available for processing, which translates into lower cost for hardware and for energy consumption. To minimize signaling latency, and thus maximize hardware efficiency, the MAC, RLC, and fast-U/S-RRM functions should run on the same hardware instance. As traffic moves to the right in *Figure 4*, the requirement on interface latency gradually relaxes.

The CPRI scales with effective carrier bandwidth and the number of antenna elements. An inter-site CPRI interface with joint combining at the central office (CO) can, in many deployments, result in a gain in uplink spectrum. For LTE, the CPRI can be inter-site (a few Gbps), but in NR — with wider carriers and more antenna elements — an inter-site CPRI would be challenging from both a latency and bandwidth perspective.

The interface between the RLC and the multipath handler scales with user data and has a latency tolerance in the order of several milliseconds. This interface is limited by the performance of the dual connectivity feature, which degrades gracefully as

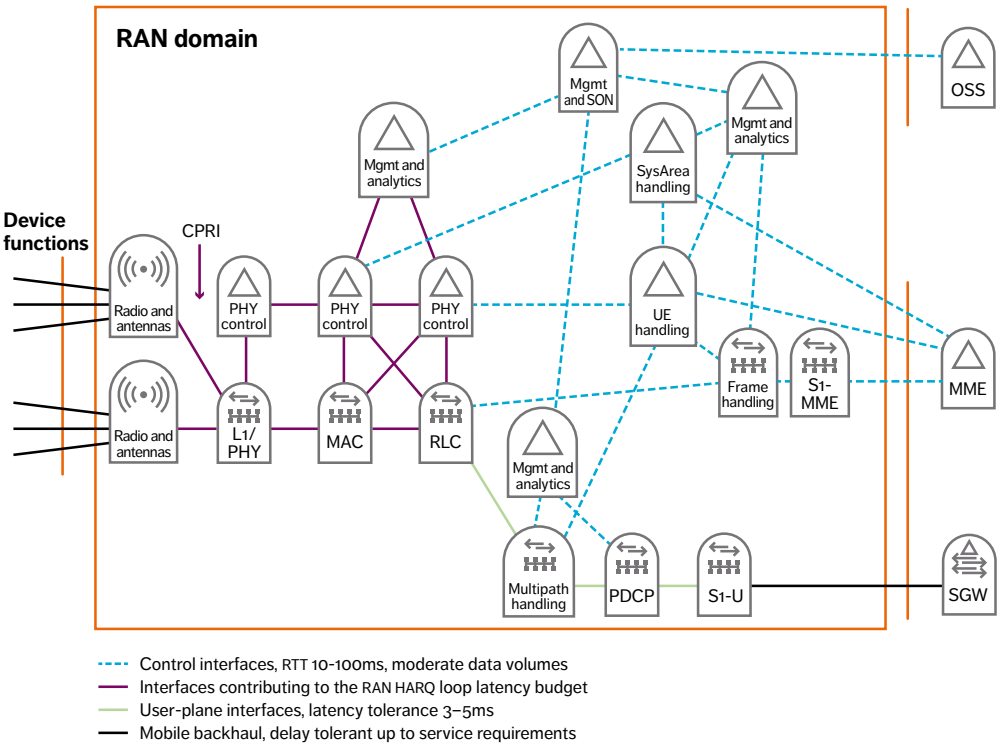


Figure 4: The logical interfaces in the RAN architecture and their characteristics requirements

the interface latency increases. This interface can, therefore, either be node internal or a networked interface between nodes and even between sites that are not more than 3-5ms apart.

The remaining interfaces in Figure 4 carry either slow control data (indicated by the blue lines) or user data between the RAN and the EPC (S1-U interface). At over 10ms, the latency requirements on these interfaces are quite relaxed.

Hardware requirements

Control functions that are asynchronous to the radio interface tend to be suitable for virtualization and VNF deployments, as they are transaction based and do not involve heavy packet processing.

On the other hand, functions like multipath handling, PDCP, and S1-U termination involve packet processing (encapsulation, header reading/creation, encryption/decryption, and routing) and can be challenging to virtualize. However, if the underlying hardware contains ciphering offload and packet-processing accelerators, virtualization is possible without performance degradation, and so these functions can be virtualized and deployed in an NFV environment.

Most RAN processing cycles occur in the HARQ synchronous functions. Tasks like uplink radio decoding and scheduling are, for example, highly processing intense. And so, the more processing that can be carried out in the uplink decoding, the better the uplink sensitivity, and the more processing that can be allocated to the scheduler, the better the use of spectrum in both the uplink and the downlink.

Reducing processing in the HARQ synchronous

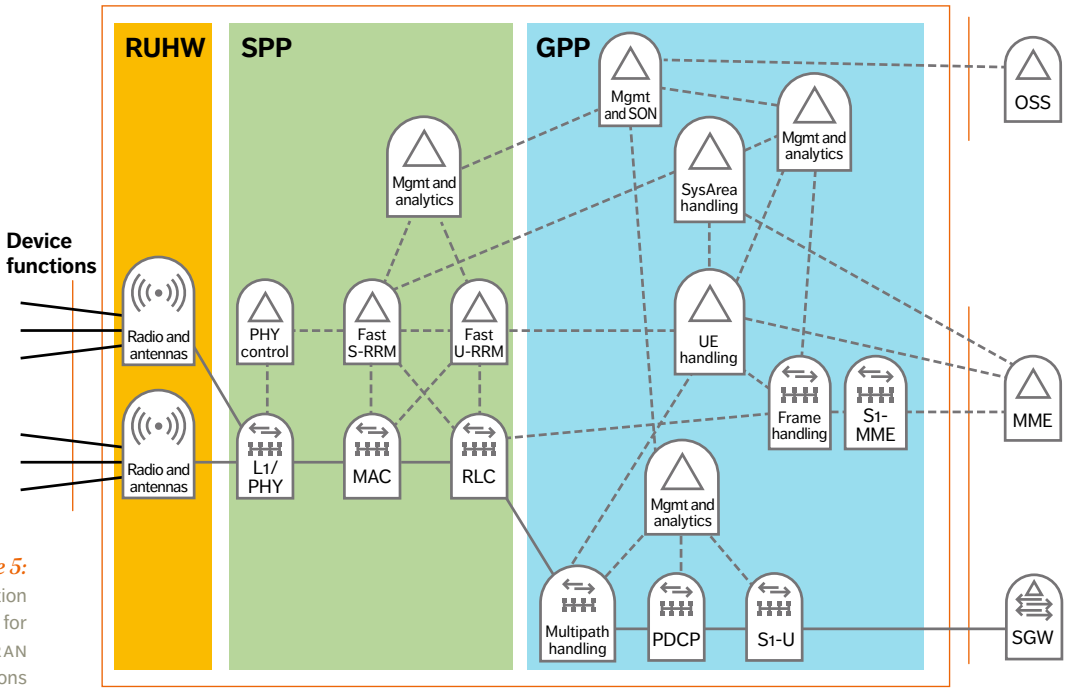


Figure 5: Preferred execution environments for the various RAN functions

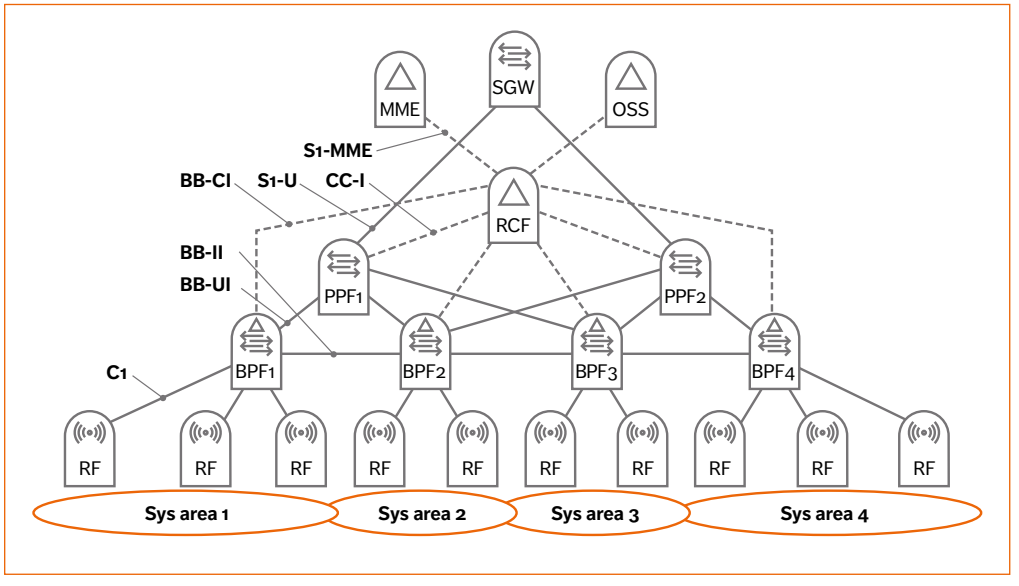


Figure 6: Hierarchy, instantiation, and inter-node interfaces in the RAN split architecture

functions is not a good idea if spectrum efficiency is to be maintained or improved. Special-purpose multi-core hardware is best suited to this type of processing, as its price-performance ratio is presently five times that of single-core hardware. And so, HARQ synchronous functions are likely to continue to run on special purpose processor (SPP) hardware for at least one or two more generations of RAN hardware.

To avoid flow control issues between the MAC and the RLC — and given the level of interaction between the scheduler and RLC — the MAC, RLC, and the fast RRM should run on the same hardware instance. The resulting hardware environment is shown in *Figure 5*.

The functions on the right in the illustration (blue area) run on general purpose processors (GPPs) that include hardware accelerators and ciphering offload for multipath handling and PDCP. The functions in the middle (green area) run on SPPs with multi-core hardware suitable for supporting functions in the HARQ loop, and the radio hardware is on the left (yellow area).

Resulting split architecture

Based on function and interface characteristics, preferred execution environment, and spectrum efficiency, the target functional composition, which is shown in *Figure 6*, includes the following logical RAN nodes:

Packet processing function — PPF

The PPF, which is suitable for virtualization, contains user-plane functions that are asynchronous to the HARQ loop, and includes the PDCP layer — such as encryption — and the multipath handling function for the dual connectivity anchor point and data scheduling.

Baseband processing function — BPF

Given the stringent requirements for spectrum efficiency, the BPF benefits from being placed on an SPP. The BPF includes user-plane functions that are synchronous to the HARQ loop, including the RLC, MAC, and L1/PHY, and it is also the anchor point for carrier aggregation (MAC) and soft combining (L1/PHY). The BPF contains the per-TTI RRM (fast radio scheduler), and is also responsible for the CoMP, for the selection of the MIMO scheme, and for beam and antenna elements.

Radio function — RF

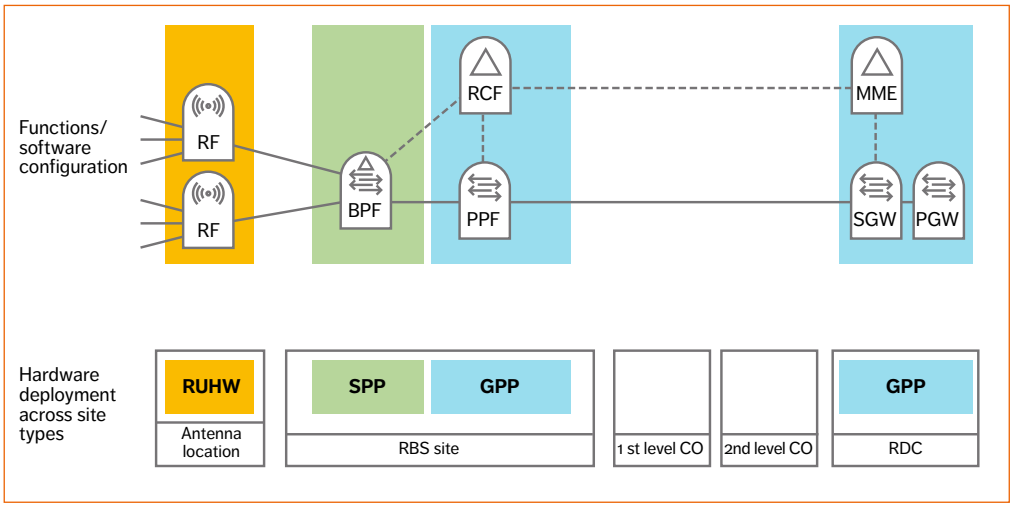
The RF requires special radio hardware and includes functions such as modulation, D/A conversion, filtering, and signal amplification.

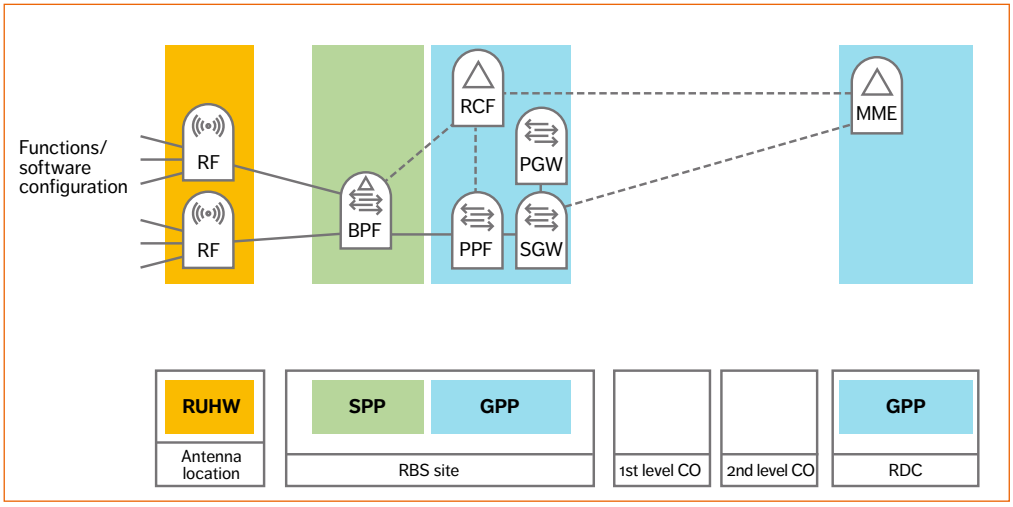
Radio control function — RCF

The RCF handles load sharing among system areas and different radio technologies, as well as the use of policies to control the schedulers in the BPFs and PPFs. At the user and bearer level, the RCF negotiates QoS and other policies with other domains, and is responsible for the associated SLA enforcement in the RAN. The RCF controls the overall RAN performance relative to the service requirement, creates and manages analytics data, and is responsible for the RAN SON functions. Like the PPF, the RCF is suitable for virtualization.

The logical interfaces are:

- C1 — an evolution of the CPRI interface, this interface scales with effective carrier bandwidth (carrier bandwidth × number of antenna streams) with a latency requirement of around one TTI (1ms for LTE and down to 67µs for NR).
- BB-UI — the user-plane interface between the PPF and the BPF, it carries PDCP PDUs, which scale according to the amount of user data sent by the BPF instance to the system area.
- BB-II — the interface between two BPF instances, it carries user data for scenarios that use inter-BPF carrier aggregation (carrier aggregation over two carriers controlled by two different BPF instances). This interface also carries control data for the inter-BPF CoMP, scales in line with the amount of user data transmitted, and the latency requirement is the same as for C1.
- BB-CI — the control-plane interface for the BPFs, which carries control and analytics data from the BPF to the RCF. This interface primarily scales with the volume of analytics data, and at 10ms or more, the latency requirements are quite relaxed.
- CC-I — the control-plane interface for the PPFs, which





Figures 7a and 7b: Deployment of hardware and nodes across site types in a classic main-remote deployment — valid for both LTE and NR

carries control and analytics data from the PPF to the RCF. Like BB-C1, this interface scales primarily according to the volume of analytics data transmitted, and has relaxed latency requirements of 10ms or more.

In a deployment, each function (RF, BPF, PPF, and RFC) is instantiated. An instance of the radio functions will be associated with a number of antenna elements at an antenna site, and a set of ***n*** RF instances are connected to one instance of the BPF.

Each antenna element (RF) is associated with one BPF. Hence, a BPF instance handles the cells corresponding to the RF antenna elements it is associated with. The set of cells under the control of a BPF instance is referred to as a system area. By definition, one BPF instance handles the radio transmission and reception of traffic in its system area. Within its system area, a BPF instance also controls beamforming, power, spectrum, scheduling, load sharing, and fast U/S-RRM.

Mobility within a system area is hidden under the BPF and not visible to the PPF or the EPC. A multitude of BPF instances are connected to one instance of the PPF. One PPF instance is therefore associated with a large number of BPFs, as well as the traffic sent to and by users in a large set of system areas. As the S1-U terminates in the PPF, a user can move around within this set of system areas or cells without causing an S1 or X2 handover. Each instance of the RCF can handle a small or a large set of PPFs — and all the associated BPFs. In this way, the RCF can keep a holistic view of an area that is just a single cell, up to an area consisting of thousands of cells. With this architecture, RRM coordination and spectrum efficiency within a system area can be maximized, using the full suite of RRM features available, which lays the foundation for the application of future innovations and RRM technologies.

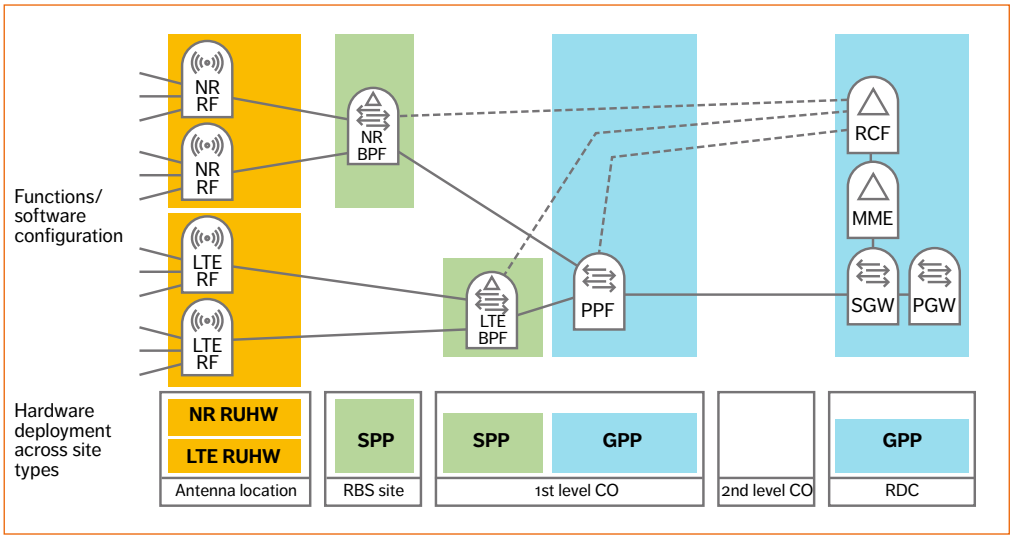
Deployment alternatives and examples

Having defined the split RAN architecture, it is possible to describe some of the deployment alternatives it supports and the associated dynamics of the software-defined radio network. *Figures 7a* and *7b* show a distributed main-remote macro site. The hardware deployed at each site is shown in the bottom half of each figure, while the top part shows the configuration of network functions across that hardware. The hardware deployment is semi-static, but the network functions are (re)configured using commands or machine instructions, and are therefore software defined.

The antenna location contains radio unit hardware that hosts the RFs. In the distributed main-remote deployment, a fiber link connects the antenna location to the RBS main site over a CPRI (C1 interface). The RBS site is configured with an SPP for the BPF, and a GPP for the PPF and RCF, and is connected over S1 mobile backhaul transport to EPC gateways residing in a more centrally located regional data center (RDC). In addition to providing an execution environment for the PPF and RCF, the GPP hardware offers a virtualization environment for VNFs and applications.

This deployment enables computing at the mobile edge for selected applications, users, or data streams to be moved to the RBS site by means of a system reconfiguration — or even automatically based on policies and traffic triggers. In Figure 7b, a virtual packet gateway — deployed as a VNF — is located at the RBS site running in the virtualization environment of the GPP. While Figures 7a and 7b describe the same hardware installations, the configuration of 7b supports computing at the mobile edge in the RBS site — due to the ability to break out traffic in the local packet gateway function. In 7a, all the traffic is backhauled to the gateway in the RDC. As the hardware deployments in 7a and 7b are the same, the radio-network architecture difference is created through dynamic software commands. In this way, the network architecture can adapt according to policies and varying traffic conditions, either through software configuration or automatically (automation and SON) — which is the basis for software-defined RAN.

A second example, illustrated in *Figure 8*, shows a typical LTE C-RAN deployment extended with main-remote NR. As *Figure 8a* illustrates, the radio unit hardware is located at the antenna location, while the SPP and GPP are deployed in the central office sites. The SPP hardware runs one or several instances of the BPF, and the GPP provides the NFV



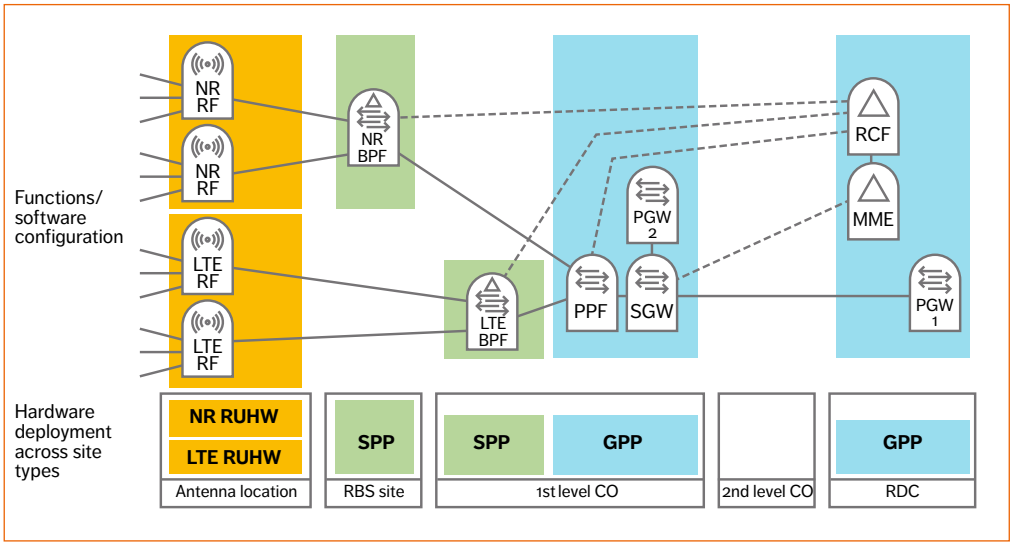


Figure 8a and 8b: C-RAN deployment with (instance *a*) a centralized packet gateway and (instance *b*) a second instance of the PGW distributed to the CO site for local breakout of selected traffic

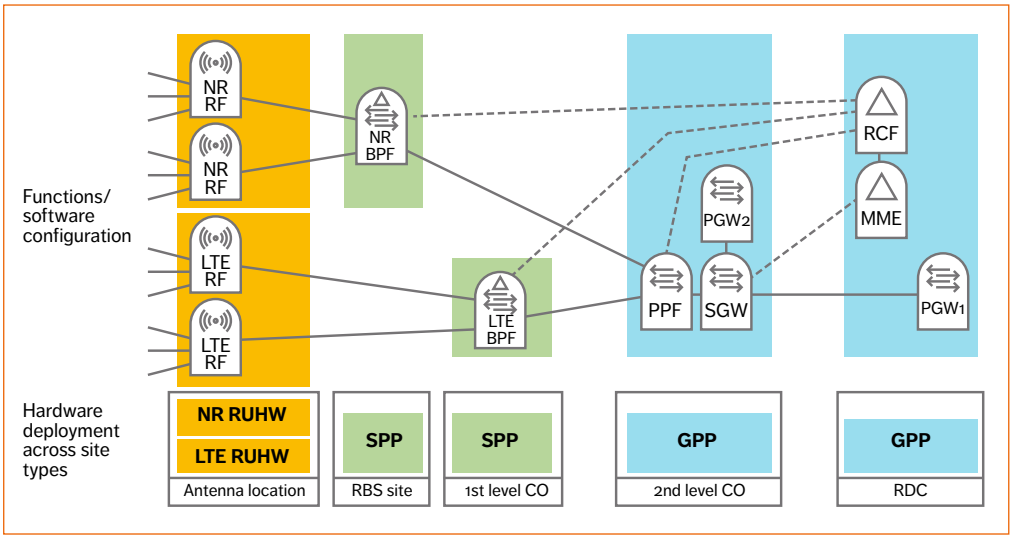


Figure 9: Flexible C-RAN architecture second-level CO for better dual connectivity performance

environment for the PPF and the RCF. Compared with the distributed macro deployment shown in Figure 7a, the LTE BPF in the C-RAN deployment holds a more centralized position. Each instance of the BPF covers more antenna locations, which results in high spectrum efficiency, as the BPF can use the complete set of fast RRM features — including joint combining, multipoint transmission, and coordinated scheduling — across all the antenna points in the system area covered by the first-level central office site.

The C-RAN deployment requires high-bandwidth fiber for the C1 interface between the BPF and the RF (CPRI fronthaul). Like the distributed macro architecture of 7b, the C-RAN deployment offers a GPP environment for VNFs and applications in the central office. The GPP may be part of a distributed data center, in which case, the PPF and RCF are deployed as VNFs, while the SPP is deployed as standalone hardware. Alternatively, the SPP can be part of the distributed data center, providing specialized hardware for the BPF — similar to the way other types of data-center specialized hardware can be used by applications with special needs, such as packet processing or firewalling. In this way, next generation central offices can be turned into a combination of mobile-edge sites and baseband hotels.

The flexible C-RAN architecture illustrated in *Figure 9* supports more centralized deployments compared with the C-RAN configuration shown in Figure 8. The GPP in the flexible C-RAN architecture is provisioned at the next level in the hierarchy, moving the PPF and hence the dual-connectivity anchor point to a more central position, which enables smooth dual connectivity mobility. Spectrum efficiency is the same or slightly improved, the number of distributed data centers drops, and the mobile edge is slightly more centralized. The transport requirements of both C-RAN configurations are similar.

Figure 10 shows an on-premises architecture that is fully self-contained using pico base stations and an on-premises data center hub that stores content and carries out processing locally. In this case, all four RAN nodes (RF, BPF, PPF, and RCF) are integrated onto the same chip.

The ideal split architecture contains pools of

hardware (an SPP and GPP) strategically deployed in selected RBS and CO sites. Instances of the three function types BPF, PPF, and RCF — and any VNFs in the mobile network — are dynamically created, modified, scaled, and terminated based on need and operator policies. Building mobile networks in this software-defined way results in an architecture that:

- is reliable and resilient — as it enables functions to relocate following a hardware, site, or link failure
- is flexible and energy efficient — as it automatically adapts to peaks and troughs in traffic patterns and service usage
- reduces time to market for new services and network features
- can maximize spectrum efficiency (minimize cost for spectrum and sites) given the geography and site topology of the network, through its flexibility to dynamically distribute the hub points for joint combining, multipoint transmission, COMP, carrier aggregation, and dual connectivity

Conclusions

The design of the 4G/5G split RAN architecture focuses on increased spectrum efficiency, full deployment flexibility, and elasticity; processing is carried out where resources are available and needed. The split RAN architecture consists of the two user-plane network functions: a packet processing function (PPF) and a baseband processing function (BPF), together with the antenna-near radio function (RF), and the control-plane radio control function (RCF).

The PPFs and RCFs can each be deployed either in classical pre-integrated nodes or in fully virtualized environments as VNFs or any combination thereof. Both functions are suitable for virtualization with existing technology, with benefits to the PPF brought by packet accelerators and ciphering support in the underlying hardware.

In principle, the BPF can also be virtualized. However, special-purpose hardware is still about five to seven times more efficient than general-purpose hardware for the type of processing the BPF performs, and so it is expected that the BPF will be deployed on an SPP for one or two more generations of hardware.

The RCF takes holistic responsibility for Radio Resource Management, RAN analytics and SON, it maintains policy and bearer information, and interworks with non-RAN domains such as the EPC

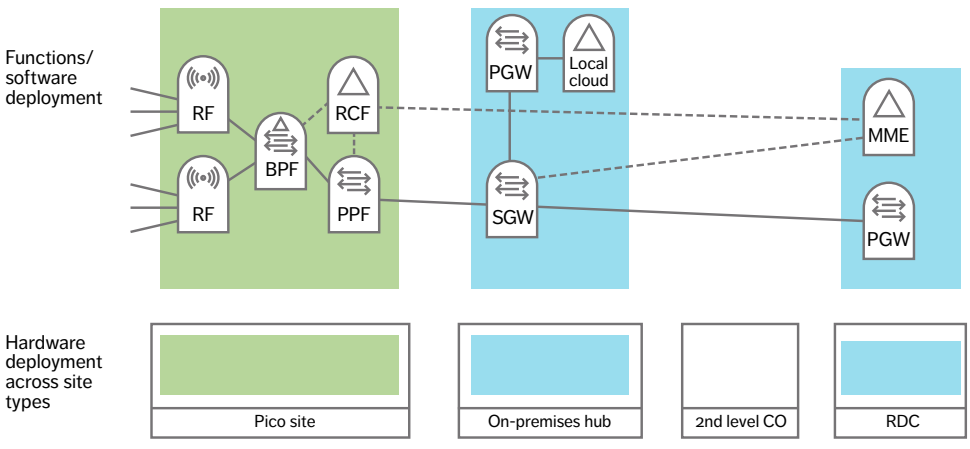


Figure 10: Factory deployment examples using pico 4G/5G base stations (green) and with on-premises breakout possibility

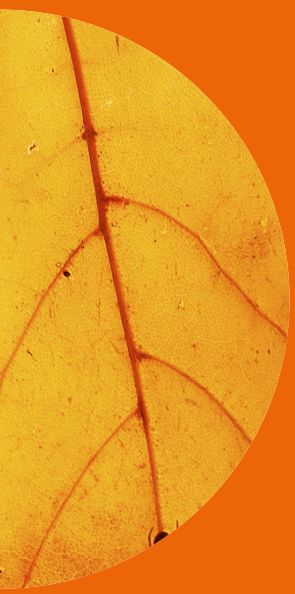
and resource orchestration layers. The RCF can be centralized or distributed in a closed on-premises factory network, for example. Deploying the user-plane BPF (processing synchronous to the TTI) and PPF (asynchronous packet processing) can be achieved in a variety of ways, as long as the BPF is within one to two TTIs from the antenna points. The PPF, on the other hand, can be more centralized, with a distance of up to 5-7ms from the radio functions. And so user-plane functions can be deployed to match service requirements, and maximize spectrum efficiency according to the spectrum, transport, and site availability, as well as the particular local geography.

The split architecture results in the necessary scaling dimensions to support 5G use cases and traffic structures in a cost-efficient way. Its flexibility and decoupling of hardware from software enables a software-defined elastic resilient RAN. It also guarantees that RAN architecture is future-proof. As an evolution of 4G RAN, the split can be gradually introduced in line with business needs. ✱

AUTHOR

Erik Westerberg

◆ joined Ericsson from MIT, Massachusetts, the US, in 1996 and is a senior expert in system and network architecture. During his first 10 years at Ericsson, he worked with development of the mobile broadband systems before broadening his field to include the full network architecture. He presently holds the position of chief network architect in Ericsson's Business Unit Network Products. He holds a Ph.D. in quantum physics from Stockholm University, Sweden.



ISSN 0014-0171
284 23-3283 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000