WHITE PAPER

Laying the 5G Foundation with an Evolved Mobile Transport Network

Introduction

Mobile Network Operators (MNOs) are embarking on a phased deployment of 5G. Commercial launches of 5G for mobility began among early adopters in late 2018 to early 2019, and these were preceded by pre-standard Fixed Wireless Access (FWA) offerings. These initial deployments are characterized by MNOs rushing to become first in their respective markets to introduce 5G NR (New Radio) technology and apply it to delivering their current portfolio of mobile broadband services. Mobile industry conferences and publications focus on the many new and exciting applications and use cases 5G will enable. However, it will take time to realize the full potential of 5G technology and, as with any major new build, the foundation must come first. This foundation must be capable of supporting the following:

- Five-fold increase in traffic volume by 2024
- Applications requiring an order of magnitude lower latency than today's network delivers
- Over 8.9 billion mobile subscriptions and four billion IoT connections globally by 2024

Early 5G deployments

While the new foundation must support future use cases, current 5G deployments are primarily focused on delivering higher capacity for enhanced Mobile Broadband (eMBB) services by leveraging the existing Radio Access Network (RAN) infrastructure from 4G Long-Term Evolution (LTE) networks and newer releases such as LTE-Advanced and LTE-Advanced Pro. These deployments are also utilizing the existing 4G Evolved Packet Core (EPC). In other words, they are based on what the 3GPP defines as Non-Standalone (NSA) deployments defined in Release 15 Early Drops and Main Drops of the 5G NR standards. With NSA, 5G NRs are deployed alongside LTE radios and leverage the existing LTE radio control infrastructure. LTE procedures are used for location updates, subscriber authentication, and user attachment to the network. Importantly for Voice-over-LTE (VoLTE) services, both control and media bearers route through LTE radios only. Most of these deployments will leverage a Distributed RAN (D-RAN) architecture, where baseband processing functions are collocated with Remote Radio Heads (RRH). The use of existing infrastructure for these initial 5G rollouts includes the mobile transport (wireline) network. While capacity upgrades and incremental enhancements to the transport network would likely have taken place to support LTE-Advanced and LTE-Advanced Pro features, and further capacity upgrades are required to accommodate 5G, the same fundamental transport architecture is being leveraged. This initial modernization deployment phase is referred to as "Early 5G".

Moving from Early 5G to Full 5G

Full 5G refers to a completely updated end-to-end network that supports all the increased performance promises of 5G. As this move from Early 5G to Full 5G occurs, fundamental changes to both the RAN architecture and transport network architecture are required. Full 5G deployments will commence once MNOs start deploying a 5G Core (5GC) network. These deployments will typically involve the 5GC coexisting with the EPC, which will entail additional NSA use cases, but Standalone (SA) mode deployments will also appear. Among early adopter MNOs, Full 5G deployments are expected to begin in 2020, though for many MNOs, this will not occur for many years.

The deployment of Full 5G will enable new applications, use cases, and features defined in the corresponding 3GPP specifications. These deployments will be based on the latest drop of Release 15, as well as Release 16. 5GC deployment

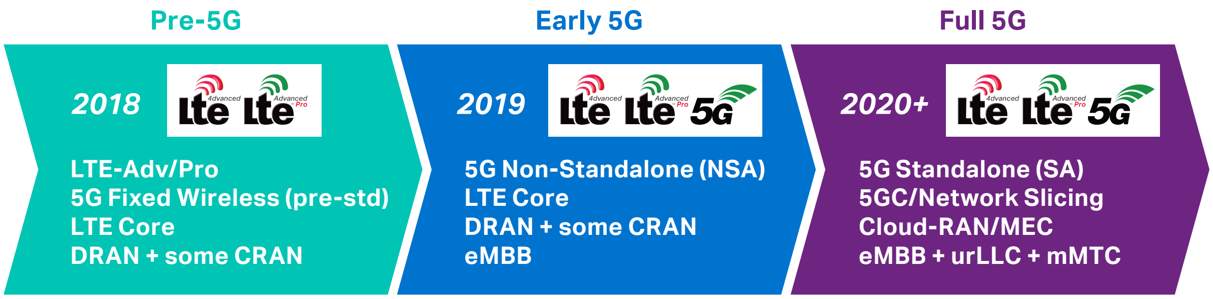


Figure 1. Phased deployments of 5G

will enable new capabilities, such as 3GPP Network Slicing, as well as additional applications like ultra-reliable Low-Latency Communications (urLLC), massive Machine-Type Communications (mMTC) and, for the first time, native Ethernet services in the wireless domain via 5G NRs.

With the introduction, maturation, and increased adoption of technologies such as Software-Defined Networking (SDN) and Network Functions Virtualization (NFV), 5G design principles were defined to take full advantage of these software-driven innovations. This translates to the virtualization and disaggregation of many RAN and mobile core functions. Therefore, there will be an increased adoption of a Cloud-RAN approach, which has significant architectural implications for the transport network as well. Of course, the deployment of Multi-access Edge Computing (MEC), required for low-latency applications, is expected to take place during the Full 5G phase, and must be considered when looking at the underlying transport architecture.

The challenge is that the new foundation for Full 5G must be cost-effective while handling these new application and bandwidth demands. How can an MNO deploy a Full 5G network using a single infrastructure capable of supporting both massive bandwidth latency-tolerant services and ultra-reliable and bounded low-latency services in the same infrastructure? Despite the higher bandwidths enabled by 5G, Average Revenue Per User (ARPU) will prove challenging to grow, and the cost of new spectrum is high. Therefore, the foundation must be cost-optimized in multiple areas, including the following:

- Open architecture with the virtualization of various functions, instead of utilizing expensive proprietary hardware
- Mobile transport network capacity increases, while reducing price per unit bandwidth (Moore's Law-like economics)
- Open, best-of-breed, and high-volume networking silicon allowing flexible customization and cost optimization
- Network automation and intelligence optimization and expansion

A new 5G service model and a focus on expanding the addressable market to new verticals is also required to ensure better ROI for 5G infrastructure investments.

What is 5G?

The Impacts of RAN and mobile transport network disaggregation

MNOs want to centralize the RAN control functions to improve overall performance, gain efficiencies, and reduce costs. Centralizing RAN elements enables a many-to-one relationship between Broadband Unit (BBU) control functions and the RRHs. Capabilities such as carrier aggregation (aggregation of spectrum), Coordinated Multi-Point (CoMP), and X2 hand-over are simplified when RAN functions are centralized. This also results in improved performance for mobile users, both humans and machines (things). RAN centralization is not unique to 5G, as centralized deployments exist in 4G LTE networks where BBUs are physically centralized in BBU hotels, or Centralized/Cloud RAN (C-RAN) hub locations. However, these centralized BBU deployments are typically based on proprietary hardware from RAN vendors. With 5G's design principles focused on a software-driven approach, the goal is to virtualize these centralized and disaggregated functions and run them on Commercial Off-The-Shelf (COTS) server platforms.

In a 3GPP radio system, baseband functions include both a real-time processing part and a non-real-time processing part. Real-time baseband handles radio functions such as dynamic resource allocation (scheduler), gNB measurement, configuration and provisioning, radio admission control, and others. Non-real-time baseband handles radio functions like inter-cell Radio Resource Management (RRM), Resource Block (RB) Control, and connection mobility and continuity functions.

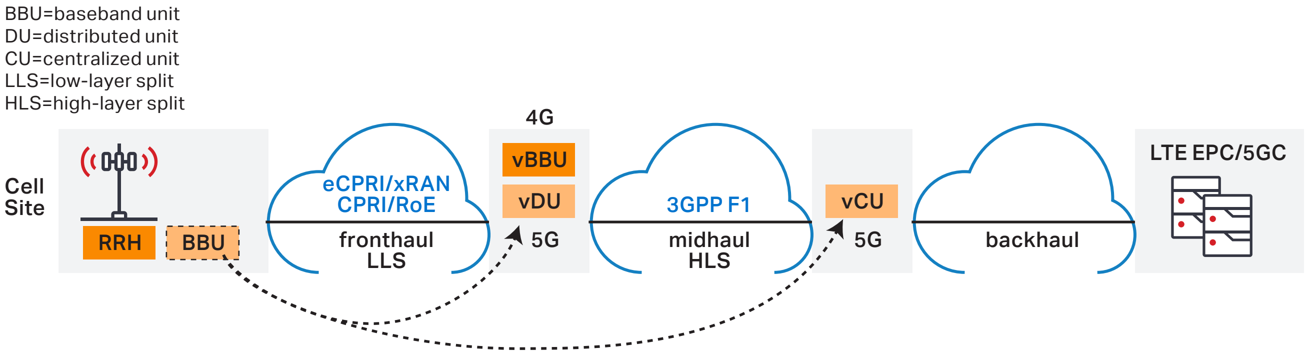


Figure 2. Split RAN architecture

Baseband functions are mapped into Distributed Units (DUs) for real-time baseband processing and Centralized Unit (CU) for non-real-time processing. Figure 2 illustrates this split, or disaggregated architecture.

While there are many possible RAN functional split options within the 3GPP RAN sub-layers, one can generalize them into a Lower Layer Split (LLS) and a Higher Layer Split (HLS). This results in new partitions of the mobile transport network as well, namely fronthaul and midhaul, which have their own unique transport characteristics and requirements. The fronthaul network between RRHs and BBUs/DUs carry extremely latency-sensitive Common Public Radio Interface (CPRI) radio control traffic for 4G LTE radios and eCPRI or Open RAN (O-RAN) fronthaul radio control messages for 5G NR and ng-LTE radios. CPRI fronthaul traffic is a serialized constant bitstream technology traditionally carried over dedicated fiber or WDM.

The decline of CPRI

CPRI requires more than an order of magnitude higher bandwidth than it delivers in effective user throughput. Since this bandwidth inefficiency does not scale with higher-capacity radios and Multiple-Input Multiple-Output (MIMO) antenna technology, new technologies have emerged to replace CPRI, such as IEEE 1914.3 Radio-over-Ethernet (RoE), O-RAN Fronthaul, and eCPRI. The use of RoE Structure-Aware mapping, combined with additional CPRI Layer 1 offloading, can reduce the required fronthaul bandwidth by a factor of five or more. With 5G, eCPRI and O-RAN fronthaul were defined as a more bandwidth-efficient packetized fronthaul interface from inception, supporting both Ethernet-based or UDP-based transport mechanisms.

“Closed” eCPRI fronthaul networking vs. “open” O-RAN fronthaul networking

The eCPRI specification was initially published by RAN vendors, including Ericsson, Huawei, NEC, and Nokia, to define the message structure for eCPRI and the transport layer of carrying eCPRI data streams. However, the specification did not completely standardize the radio control messages within the eCPRI application layer—known as the eCPRI protocol layer in the eCPRI specifications—required to ensure a fully open and interoperable implementation between the RRH/RU (Radio Unit) and DU. Therefore, MNOs became concerned that eCPRI may still result in lock-in by RAN vendors, forcing them to deploy the same vendor for the entire RAN, instead of enabling a best-of-breed disaggregated approach. This is one of the key drivers for the formation of the O-RAN Alliance, which was a merger of xRAN Forum and C-RAN Alliance and is led by MNOs to define open interfaces and standardized radio control messages between RAN elements. The goal is to enable an ecosystem where they can mix-and-match RAN vendors in a cloud-native 5G architecture. The O-RAN Alliance has a number of Working Groups, including WG4, which published version 1.0 of their open fronthaul specification in March of 2019, which significantly leverages the previously published xRAN fronthaul specification.

The midhaul network carries traffic between the DU and CU and has tighter latency requirements (1ms to 5ms) when compared to backhaul traffic (less than 20ms), but not nearly as stringent as fronthaul (150μs to 200μs). This is accomplished over the 3GPP standardized F1 interface, which utilizes standards-based Ethernet and IP encapsulation for the transport layer. Figure 3 summarizes the different transport network requirements between fronthaul, midhaul, and backhaul. While midhaul has a somewhat smaller latency budget than backhaul,

RU	DU	CU	5GC
	Fronthaul	Midhaul (PDCP/RLC)	Backhaul (CU with SDAP)
Bandwidth	4G: 2.5 (CPRI 3) to 10G (CPRI 7/7a) 5g: N x 10GE/25GE/50GE	10GE/25GE/50GE	10GE/25GE/50GE/100GE
Latency (Round-Trip)	4G: 150us~200us (Bounded) 5G eMBB: 150us or Less (Bounded)	1ms~5ms (Bounded)	Less than 20ms
Radio Protocol(s) Processing	O-RAN, eCPRI, CPRI CPRI with 1914.3 RoE	Not Required (Transport is IP/Ethernet)	Not Required (Transport is IP/Ethernet)
Statistical Multiplexing	CPRI/RoE Structure Agnostic: No xRAN, eCPRI, RoE FDM: Yes (Marginal)	Yes	Yes
Network Slicing	Not Required	Yes Criteria: Based on S-NSSAI (Single-Network Slice Selection Assistance Information), QoS Flow Indictcator, QoS Flow Level Parameters, DRB ID (Data Radio Bearer ID), etc IEs	Yes Criteria: Based on NSI (Network Slice Instance), IMEI, IMSI, IMEI/IMSI Ranges, PLMN-ID, etc IEs
Reach	Less than 10km	Less than 20km	Less than 100km
Packet Timing/Sync	4G & 5G: 1ns PTP Timestamp Accuracy	4G LTE-A Pro: 15ns~20ns PTP Timestamp Accuracy 5G: 1ns PTP Timestamp Accuracy	4G LTE-A Pro: 15ns~20ns PTP Timestamp Accuracy 5G: 1ns PTP Timestamp Accuracy
Topology	Hub & Spoke, Ring	Hub & Spoke, Mesh, Ring	Hub & Spoke, Mesh, Ring
Transport Technologies	L1: P2P Fiber, Packet Optical (Flex-E/G.mtn) L2: Ethernet/TSN L3: IP/Ethernet (RU Remote Mgmt. Only	L1: Optical, Packet Optical (Flex-E/G.mtn) L2: Ethernet/TSN L3: IP/MPLS, EVPN, Segment Routing	L1: Optical, Packet Optical (Flex-E/G.mtn) L2: Ethernet/TSN L3: IP/MPLS, EVPN, Segment Routing
OAM	CPRI L1 & L2 OAM, 1914.3 RoE OAM (round trip delay, etc.)	802.1ag CFM, Y.1731, TRAMP, RFC 2544/Y.1564, 802.3ag EFM; VCCV BFD; G.mtn OAM (in progress)	802.1ag CFM, Y.1731, TRAMP, RFC 2544/Y.1564, 802.3ag EFM; VCCV BFD; G.mtn OAM (in progress)

Figure 3. Mobile transport network characteristics by network segment

the required transport network architecture and technologies are very similar between the two, as shown in the table below. In fact, in many cases, midhaul and backhaul traffic are expected to combine many portions of the network, given the different deployment scenarios that MNOs will employ, even within a given metro (such as combinations of D-RAN and C-RAN).

In the more latency-sensitive fronthaul network, care must be taken to deliver the required performance, especially in situations where traffic from 4G and 5G RRH is mixed. As eCPRI was defined to utilize a native packet transport, it is somewhat more tolerant to jitter than CPRI, which is natively a time domain-oriented constant bitstream. Therefore, when CPRI is packetized via technologies such as IEEE 1914.3 RoE Structure-Agnostic mapping and combined with eCPRI or F1 over the same fiber, special provisions are required to guarantee its low-latency and low-jitter delivery to the BBU. New technologies, such as FlexE with ITU-T G.mtn (Metro Transport Networking) enhancements as well as Time-Sensitive Networking (TSN), have emerged as tools to provide these latency and jitter guarantees (more specifically, IEEE 802.1CM defines how TSN, specifically frame pre-emption, should be applied to fronthaul).

Target architecture for 5G deployments and virtualization of the RAN

As MNOs centralize the DU and CU, to achieve the aforementioned performance and efficiency benefits, they are also seeking to virtualize these functions. They want to move away from closed, proprietary hardware and leverage COTS-like hardware. Therefore, Central Offices (COs) and C-RAN hubs are evolving to become more data center-like to host the virtualized RAN elements (hence the term Cloud RAN). The cloudification of 5G mobile networks further paves the way for software programmability, adaptive networking, and the deployment of new mobile applications in these cloud environments.

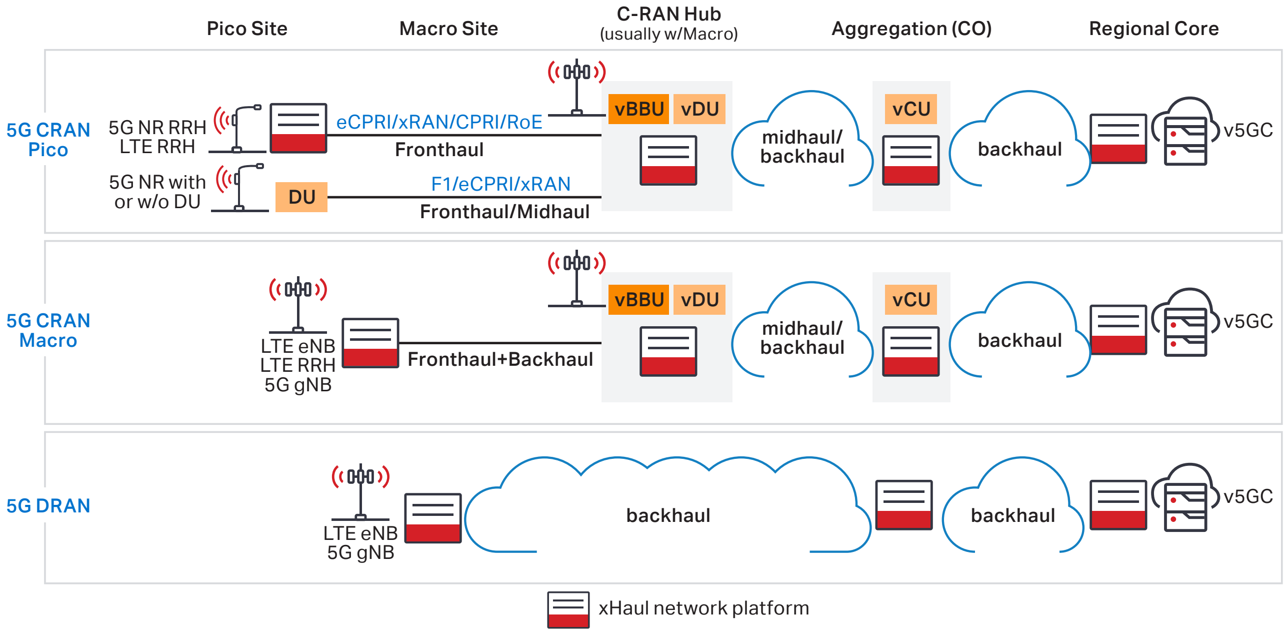


Figure 4. Deployment models for Full 5G

While the adoption of C-RANs will have significant impacts to the transport network, it is important to recognize that deployment of this architecture will not happen throughout mobile networks. Even major Tier 1 MNOs who deploy Cloud/Centralized-RANs will do so in more densely populated metro areas where they will benefit the most from this approach. In suburban or rural locations with lower population density, the number of users and associated traffic demands may not justify the investment in this type of architecture. Therefore D-RANs will likely continue to be deployed in these less-populated geographic environments. In large metros and their surrounding areas, there will likely be combinations of D-RAN and C-RAN environments being served by common portions of backhaul infrastructure and a common mobile core. Figure 4 illustrates a few high-level examples of deployments models for Full 5G, including C-RAN and D-RAN models.

Figure 4 shows different deployment models for Full 5G across the various layers of a network, from tower locations on the left (small and macro cell sites) to the 5G core on the right. At various locations in the network, there will be “xHaul platforms”, supporting combinations of fronthaul, midhaul, and backhaul network capabilities. There may be combinations of the three example deployment models in a single mobile network, as previously indicated, based on varying geographic factors and user density. The C-RAN hub, aggregation location, and regional core are all potential targets for deployment of compute for the purposes of hosting virtualized RAN functions such as the vBBU/vDU/vCU or other MEC applications. Details of this compute environment and data center-like infrastructure are not shown in this diagram, given the mobile transport network focus. The following delves deeper into each of these deployment models to see what is required in each location:

5G C-RAN Small Model: Small cells with 5G NR will be used for densification to augment existing macro cell coverage areas. Some Tier 1 MNOs in the U.S. have announced plans to deploy 5G with millimeter wave (mmWave) in the high-band spectrum (26GHz to 40GHz), which has a smaller coverage area, albeit with high capacity, and therefore requires a larger number of small cells. For cells that operate with 100MHz of spectrum bandwidth or more, these 5G NR deployments may be deployed with an integrated, hardened virtualized DU (vDU) at RRHs and deliver the F1 midhaul interface to the centralized and virtualized CU infrastructure. For small cells that operate with less than 100MHz of spectrum bandwidth, this category of 5G NR RRHs will be deployed with eCPRI, xRAN, and IEEE 1914.3 RoE fronthaul interface from the RRH to the DU, where the latter can also be a virtualized infrastructure. 5G NR will be combined with 4G LTE radios in some cases, even on the tower. Unless dedicated fiber or wavelengths are available for each RRH, traffic will need to be combined and multiplexed onto the same fiber. This means a pole-mount packet-platform will be required to packetize the CPRI via RoE and combine it with eCPRI/xRAN onto the same fronthaul fiber toward the C-RAN. Due to tight latency-bounded CPRI/RoE traffic, care must be taken to guarantee low-latency delivery, which requires features above and beyond standard Ethernet/IP switching; these are explored in more detail later

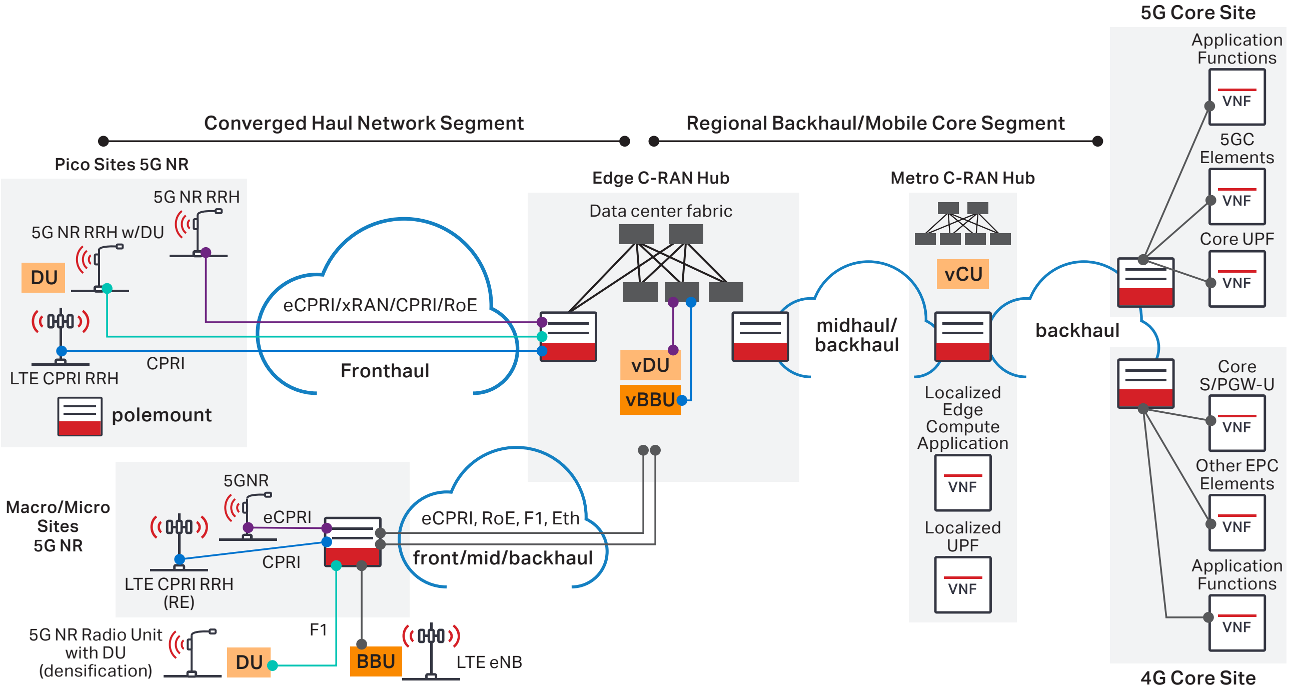


Figure 5. Combined view of Full 5G target architecture

in this document. Due to latency requirements, the distance between RRHs and C-RAN hubs are typically kept from 10km to 15km. At the C-RAN hub, fronthaul traffic is received by a fronthaul-capable xHaul platform from small cells and delivered to the vBBU/vDUrunning on COTS servers. If CPRI traffic has not yet been packetized at the tower (that is, if coming from an LTE RRH over direct fiber), it must now be mapped into RoE by this xHaul platform for delivery to the vBBU, as there would be no native CPRI support on the COTS servers.

5G C-RAN Macro Model: 5G NR will also be deployed in macro cells using mid-band and low-band spectrum for broad coverage, often overlapping with 4G LTE macro coverage areas, but driving higher capacities. These macro cells would typically include a BBU from the LTE eNB deployment and/or connections from other macros that were being subtended. In other words, traffic from 4G LTE macro cell sites were traditionally "backhaul" based on a distributed RAN architecture. However, as MNOs deploy additional radio heads for 5G NR, and/or implement densification via small cells connected to these macro sites, the macro may become increasingly space- and power-constrained. Therefore, there will be 5G and 4G LTE RRH deployed without a localized base station, meaning that fronthaul will also be delivered from these macro sites. Hence, there will be combinations of fronthaul and backhaul, or perhaps even fronthaul + midhaul + backhaul (xHaul), from a single macro site and over a single fiber pair. Similar to the small cell case above, at the C-RAN hub site, the xHaul platform must be able to combine and/or packetize fronthaul traffic from these macro towers for the local vBBU/vDU while also directing the midhaul/backhaul traffic towards the mobile core.

5G D-RAN: 5G NR will also be deployed in a D-RAN architecture, which was the predominant model for 4G LTE networks. Early 5G deployments using NSA mode are also likely to follow more of a D-RAN approach. MNOs having C-RAN in portions of their network will deploy D-RAN in other portions of their network. Other MNOs may only deploy D-RAN for years before adopting C-RAN, if they adopt it at all. The approach taken may vary in different parts of the world based on regional factors. However, one key distinction between an 4G LTE D-RAN and a 5G D-RAN is that, once a 5G core is deployed, MNOs may benefit from 3GPP network slicing. 3GPP Network Slicing will have impacts in how traffic is backhauled with the transport network or, more specifically, in how traffic is isolated and provided with unique guarantees or treatments relative to the other slices in the network. The same is true of the prior two 5G deployment models. 3GPP network slicing will apply throughout midhaul and backhaul portions of the network and, in some cases, fronthaul, although the need for this is not as apparent given that fronthaul typically consists of dedicated point-to-point connections from RRH to BBU/DU.

By combining the different deployment models described above into one common network view, this can result in an architecture such as the one shown in Figure 5.

Gain more 5G insights →

Implications for wholesale network operators

In some markets, such as in North America, a significant amount of mobile transport occurs via wholesale providers. Given the wireless coverage required to be a leading MNO in a geographically vast North American continent, these MNOs must rely on wireline wholesale providers for many parts of the country, since their own wireline network footprint covers only a portion of their wireless footprint. This wholesale model exists in some other markets globally, such as the United Kingdom, but is most prevalent in North America. Because the 4G LTE deployments were predominantly based on a D-RAN architecture, the wholesale offerings are largely focused on backhaul network services.

With the 5G architecture changes discussed in the previous section, there will be impacts to the wholesale provider offerings and business models. In a split RAN architecture, MNOs will want to tightly control the fronthaul segment of the network because of performance and latency requirements—and because they may want to provide their own timing distribution from the C-RAN hubs. This implies they will likely seek dark fiber from wholesale operators to connect the RRH to their C-RAN hubs. There are some wholesale operators looking into "lit" wholesale offerings for fronthaul; these are feasible provided the wholesale operator implements some of the new technologies and innovations discussed in the upcoming section of this paper.

Midhaul and backhaul portions of the network can generally be addressed by wholesale providers with current technology and product offerings with early 5G deployments. However, as MNOs move to Full 5G, the requirements to support 3GPP Network Slicing—and its corresponding implication for the transport system—must be considered by the wholesalers. As discussed in the upcoming section, there are many potential use cases and implementation options for 3GPP Network Slicing. Wholesale providers will have an opportunity to differentiate in their midhaul and backhaul wholesale offerings through value-added capabilities to support 3GPP Network Slicing.

xHaul network infrastructure requirements to support 5G architecture

Delivering Full 5G requires architecture changes and new technology innovations, including the following:

- Packetized and deterministic xHaul
- Evolution to virtualization and C-RAN
- Transport networks resources instantiated as part of 3GPP Network Slice Instance (NSI)
- Higher-speed interfaces and packet-optical integration, which will include increasing use of 25GbE, 50GbE, and 100GbE interfaces in fronthaul and midhaul networks with 400GbE in midhaul/backhaul, along with corresponding mapping into the photonic layer via coherent optics
- Synchronization—As with more advanced LTE features (such as CoMP), 5G will require time/phase synchronization in addition to frequency synchronization, and even more stringent timing precision for features such as MIMO transmission diversity. A deeper exploration of 5G synchronization is beyond the scope of this paper.

This section explores the first three of these technology innovations and associated architecture impacts in further detail.

Packetized and deterministic xHaul: As previously noted, bandwidth growth from 4G LTE and 5G radios requires fronthaul traffic be packetized to support the scale required. For LTE, CPRI traffic is packetized via IEEE 1914.3 RoE technology, which supports structure-agnostic and structure-aware mapping modes. Structure-aware mapping reduces required fronthaul bandwidth to some degree (from 100 percent of CPRI bandwidth to ~ 85 percent) by discarding non-utilized and idle portions of CPRI traffic. However, more significant gains are achieved by performing CPRI Layer 1 offload, which corresponds to the Low-PHY portion of the RAN functional split—achieving a much greater bandwidth reduction (~ 20 percent of CPRI bandwidth) for fronthaul traffic. This is accomplished by implementing Intra-PHY functional split, which is the adopted split option defined in eCPRI and xRAN/ORAN fronthaul specifications, as illustrated in Figure 6.

It should be noted that bandwidth percentages are approximate, as these will vary by implementation and traffic pattern. Layer 1 offload entails processing the PHY lower layer of 4G LTE and 5G NR systems. This includes performing (Fast Fourier Transform (FFT), inverse FFT, Cyclic Prefix insertion and removal, beamforming, Physical Random-Access Channel (PRACH) filtering, and precoding, thereby converting any time domain

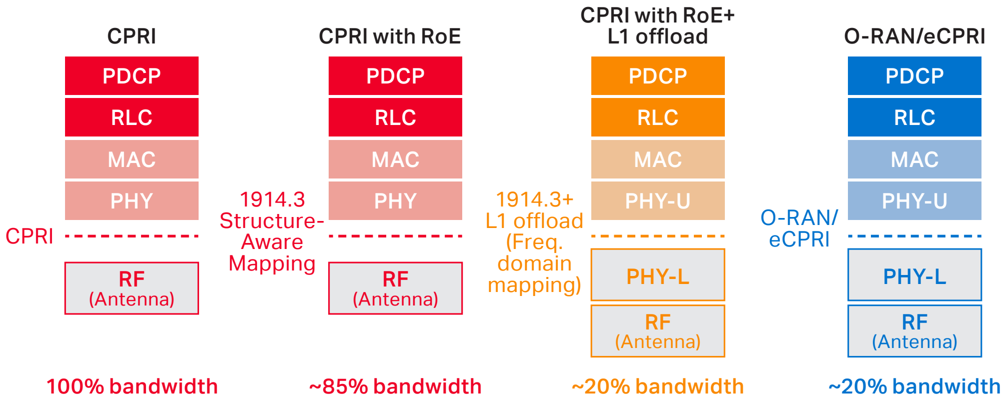


Figure 6. eCPRI and xRAN/ORAN fronthaul split options and specifications

fronthaul bitstream into frequency domain mapped data payloads. This provides significant bandwidth savings on the fronthaul interface. The eCPRI v2.0 specification that was published in May 2019 includes an Inter-Working Function (IWF) to convert from CPRI to eCPRI, including support for frequency domain mapping, as described above. However, it acknowledges that full interworking is still dependent on vendor-specific implementations (unlike an open standard like O-RAN).

Since 4G LTE and 5G radios will be deployed together in the same cell site locations, this can result in a mix of packetized CPRI and eCPRI/O-RAN traffic. Due to the latency sensitivity of CPRI traffic between RRHs and BBUs, regardless of whether it has been packetized, care must be taken to ensure that the bounded latency requirements are met. As noted in Figure 3, 4G LTE round-trip times from RRHs to BBUs are as low as 150μs. With distances from RRHs to a C-RAN hub—where BBUs reside—being typically around 10km, optical transmission delays from RRHs to BBUs alone are around 100μs round-trip, leaving only about 50μs for additional delays. If the fronthaul transport rate between RRHs and the C-RAN is 10GbE, the delay of CPRI/RoE traffic waiting for frame transmission of another traffic type, such as eCPRI, can be significant, even if CPRI/RoE has been placed in a strict priority queue. The delay having to wait for an eCPRI frame once it has started transmission can be greater than 20μs for jumbo frames. If fronthaul transport rates are 100GbE, then the delay incurred by CPRI/RoE becomes negligible. Delay with rates in between these (such as 25GbE) may still be considered significant.

There are two technologies in the fronthaul that can be applied to mitigate latency impacts on CPRI/RoE when combined with other traffic. One such mechanism is **Flex Ethernet** (FlexE), standardized in the OIF, which supports channelization as one of its use cases. This means FlexE can partition a 50GbE interface into 15 Gb/s + 35 Gb/s channels, for example, wherein each channel is scheduled in a TDM-like fashion. By mapping CPRI/RoE into one of these channels with dedicated TDM-like scheduling, its latency and jitter will not be impacted by traffic in the other channel and bounded low-latency delivery can be ensured.

The other technology is **TSN**, specifically its ability to provide Time-Aware Scheduling (standardized in IEEE 802.1Qbv) with Frame Pre-emption (standardized in IEEE 802.1Qbu). IEEE 802.1Qbv-compliant Ethernet switches have a time gate control logic associated with all eight Ethernet queues, whereby the gate opening and closing time for frame transmission can be programmed in nanoseconds granularity. Frame pre-emption works by fragmenting lower-priority frames, such as non-fronthaul traffic, to immediately service CPRI/RoE frames without incurring further delay. Combining IEEE 802.1Qbv and IEEE 802.1Qbu ensures high-priority frames assigned to a particular queue always have bounded latency and jitter performance, regardless of packet sizes of low-priority frames.

The challenge with time-aware scheduling and frame pre-emption is that it is supported by very few, specialized Ethernet silicon devices. Current available devices have limited scale and functionality to meet broader mobile transport networking requirements, such as those requiring converged xHaul deployment scenarios. As described in the 5G C-RAN macro cell site deployment model, fronthaul traffic may be combined with backhaul and even midhaul traffic. This means a macro cell site platform must support backhaul transport technologies as well as fronthaul. These backhaul technologies typically include L2 and L3 VPNs running over MPLS. Today, MPLS backhaul networks are based on traditional protocols, such as Label Distribution Protocol (LDP) or Resource Reservation Protocol-Traffic Engineering (RSVP-TE), but increasingly operators are evolving to Segment Routing MPLS, as an SDN-driven packet underlay. In these converged scenarios, FlexE is better suited to ensure low-latency delivery of packetized CPRI, as it is more widely supported together with scalable backhaul technologies, such as Segment Routing MPLS.

Evolution to virtualization and C-RAN: Many technology enablers are required to enable virtualization of the RAN elements, and thus deployment of a C-RAN. When functions such as vBBU and vDU are deployed in a C-RAN hub, they run on servers with Ethernet interfaces typically supporting 10GbE and 25GbE. Native CPRI interfaces are not supported, which is the other key reason packetization of CPRI is required, in addition to the aforementioned bandwidth savings. If CPRI has not already been packetized at the cell site tower location, it must be packetized once it reaches the C-RAN hub and before delivery to the vBBU.

For a successful vBBU/vDU implementation, two technical challenges on x86 COTS servers needs to be resolved. First and foremost, x86 CPUs are general purpose processors, and are not designed to handle real-time radio scheduling tasks. To overcome this first challenge, one or more "FPGA-based Hardware Accelerator" interfaces is required on the COTS servers for handling the radio scheduling and fronthaul processing tasks. In that case, the x86 CPU resources are freed up for other virtualization functions. The second challenge is the PCIe bus bandwidth and protocol bottleneck. The PCIe bus and protocol was never designed to handle extremely high amount of constant bit rate time domain streams, like CPRI IQ data and Control and Management (C&M) Fast channels.

Since fronthaul protocols are not designed to perform any retransmission when fronthaul frames are dropped due to bus congestion issues, this PCIe bus and protocol bandwidth limitation needs to be resolved as well to realize the possibility of vBBU/vDU. This is where the Fronthaul Gateway with the L1 Offload capability will help. With L1 offload processing enabled on the Fronthaul Gateway, the constant bit stream of time domain CPRI IQ and C&M (Fast) control channels are adapted to a variable bit rate frequency domain data stream, thus reducing the effective fronthaul data stream rate hitting the PCIe bus. Over time, newer generations of x86 CPUs will have 100GbE SerDes connectivity, as well as supporting Cache Coherent interfaces like Intel's CXL (Compute Express Link). These enhancements are all meant to eliminate the PCIe bandwidth bottleneck and allow high-performance network connection between the x86 CPU and various hardware accelerators, either FPGA- or ASIC-based.

MNOs are also virtualizing the mobile core functions. While most mobile core functions are deployed higher up in the network, some functions may be distributed closer to the edge. For example, there are synergies in collocating some virtual core functions with the vCU location. In some cases, the vCU may even be collocated with the vDU. One of the core functions that MNOs are looking to virtualize and distribute is the User Plane Function (UPF), which handles the routing and forwarding of user data packets in the mobile core (in LTE, this was handled by the Serving Gateway (S-GW) and Packet Data Network Gateway (P-GW). Distributing this function helps with scalability and performance of a virtualized mobile core, as this is a packet processing-intensive function that is more challenging to implement in compute if completely centralized. There are advantages to leverage hardware acceleration to pre-process traffic before delivering it to the virtualized UPF. Therefore, there is synergy in running the vUPF in locations where the mobile transport network and xHaul platforms are deployed, and in locations where a compute infrastructure is not already in place (such as a traditional central office), it can be advantageous to integrate this compute directly into the xHaul transport platforms themselves.

Orchestrating virtual functions will be an important part of a virtualized RAN and core solution for 5G. The orchestrator must be able to allocate the necessary compute, memory, and storage to these virtual functions. This allocation of compute resources must be adaptive and dynamic in nature given that user demands and scaling will change. The orchestrator should further coordinate with the mobile transport network as necessary to provision the networking connectivity between the xHaul transport equipment and the virtualized functions.

3GPP Network Slicing: Once 5GC is deployed, MNOs will be able to leverage this new capability defined in the 3GPP standards. 5G is intended to support a wide range of applications and business needs, each with their respective performance, scale, and reliability requirements. 3GPP Network Slicing was defined in the standards to add the flexibility and scale to efficiently support this more diverse set of requirements concurrently over the same infrastructure. Furthermore, since ARPU is difficult to grow, despite the higher bandwidth enabled by 5G, 3GPP Network Slicing is viewed as a technique to increase the addressable market size for 5G by addressing new application spaces with premium Service Level Agreements (SLA) requirements. There is no de facto list of target use cases for 3GPP Network Slicing, although there are many different viewpoints and candidate applications being discussed in the industry.

Sample 3GPP Network Slicing use cases

Some potential use cases and applications for 3GPP Network Slicing that often arise in discussion include:

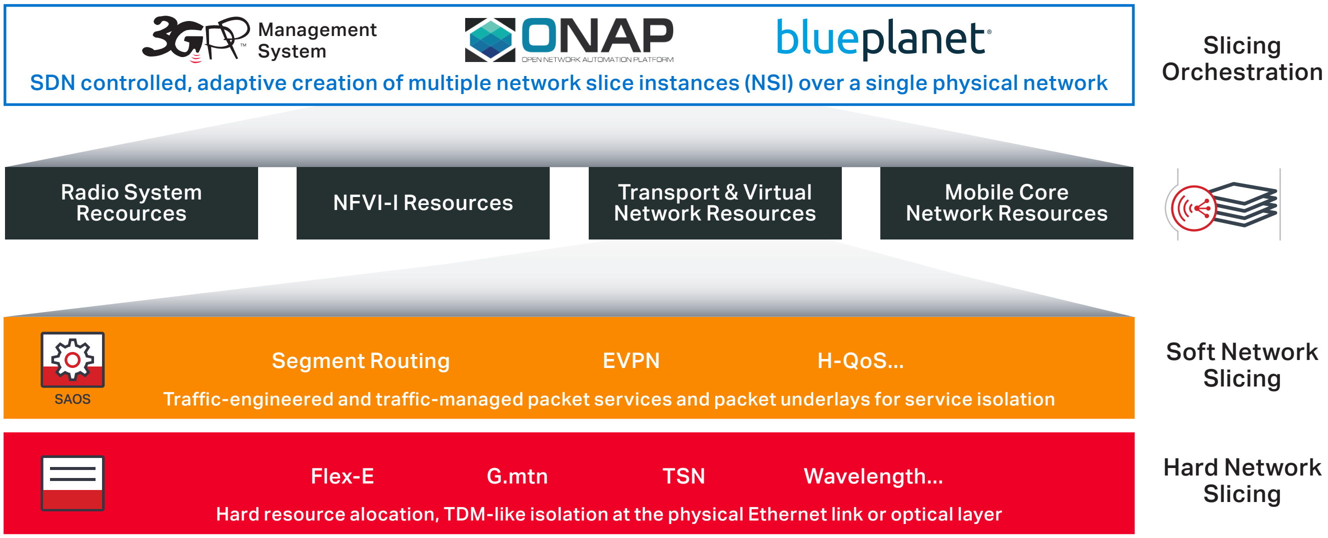
- Emergency services for emergency responders
- Entertainment
- Manufacturing (such as industrial robotics)
- Enterprise private networks (including Fixed Wireless Access). The special case of utilities (water, electricity, gas) control and monitoring networks often comes up to replace existing costly dedicated networks delivering highly reliable communication to many remote locations
- Automotive (Cellular Vehicle-to-Everything, also referred to as CV2X)
- Medical
- Shipping and logistics
- Autonomous vehicles and smart transportation

It is not yet clear how many slice types will practically be deployed—opinions range from just a handful to thousands.

When implementing 3GPP Network Slicing in the mobile network, it is important to take a comprehensive approach, which includes the orchestration and provisioning of network slices, as well as how they are implemented in the network layer via both soft and hard slicing mechanisms. Figure 7 shows an overview diagram of this comprehensive approach and possible slicing technologies.

3GPP Network Slicing will span from the wireless radio system, where spectrum is sliced, to the wireline transport network, to the mobile core. These slices must be coordinated end to end across these resources. In the network layer, the use of hard or soft slicing will depend on the requirements and application of the slice user. Soft slicing can be used to provide traffic-engineered and traffic-managed isolation of resources. Technologies such as Segment Routing MPLS (SR-MPLS) can be used to provide multiple SDN-controlled traffic engineered Label Switched Paths (LSPs), representing different slices, with policies at ingress to map traffic into the appropriate path or slice. The Segment Routing paths or tunnels can be established based on various constraints/parameters and policies such as bandwidth, latency, resiliency requirements, transport, peering costs, and more. Tunnels can be mapped in the xHaul transport platform to specific QoS treatments. This should include dedicated queuing and scheduling resources, with reserved buffer allocation, to provide resource partitioning when slices are sharing ports. Figure 8 depicts the interaction between a 3GPP System (3GPP Management System, RAN Network Functions, and Core Network Functions) and the transport network to form an end-to-end 3GPP Network Slice Instance.

It is important that the packet network and technologies, such as Segment Routing, be implemented in an adaptive and dynamic way. As slices will be sold on the basis of achieving a premium SLA, it is important to leverage telemetry and analytics to monitor the network given that conditions may vary over time, requiring adaptive changes to ensure the slice SLAs are met on an ongoing basis. Furthermore, some MNOs



3GPP Management System
ONAP
OPEN NETWORK AUTOMATION PLATFORM
blueplanet
SDN controlled, adaptive creation of multiple network slice instances (NSI) over a single physical network
Slicing Orchestration
Radio System Recources
NFVI-I Resources
Transport & Virtual Network Resources
Mobile Core Network Resources
SAOS
Segment Routing
EVPN
H-QoS...
Traffic-engineered and traffic-managed packet services and packet underlays for service isolation
Soft Network Slicing
Flex-E
G.mtn
TSN
Wavelength...
Hard resource alocation, TDM-like isolation at the physical Ethernet link or optical layer
Hard Network Slicing

Figure 7. Holistic view of 3GPP Network Slicing

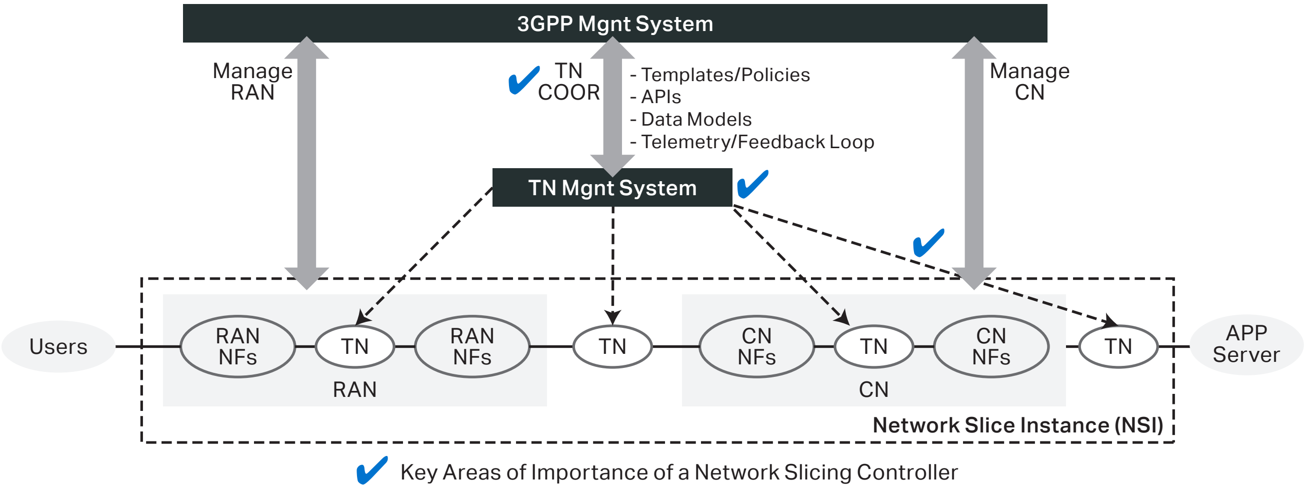


Figure 8. Transport Network (TN) resources are an important part of the 3GPP NSI

want Network Slices themselves to be dynamically provisioned and/or removed based on predefined policies like subscription durations or on-demand provisioning. SDN control is an important component for achieving this dynamic behavior, which is a key reason Segment Routing is often cited as the packet underlay technology of choice for 5G.

Hard slicing delivers strict isolation of resources without relying on statistical multiplexing mechanisms or virtualized partitioning. From an OSI model perspective, hard slicing is thought of as being implemented at Layer 1. Hard slicing can be applied, for example, where the end-user would otherwise be served by a "private network" build. Since the xHaul network will have packet and optical transport technology, one option is to dedicate wavelengths to slices. With most modern optical networks supporting flexible grid DWDM technology and programmable modulation of coherent modems, the wavelengths can be "right-sized" depending on the requirements of that slice. Where the packet/IP network infrastructure connects into the optical layer, technologies such as FlexE can be utilized to provide hard (TDM-like) channelization, sub-rating, or bonding of the Ethernet PHY, and this can be applied to match the wavelength bandwidth. There are new technologies such as G.mtn, defined in ITU study group 15, which extends the FlexE channel construct end to end across the network through a new Slicing Channel Layer (SCL). With SCL, the FlexE channels can be cross-connected at intermediate nodes at the lowest possible latency since this cross connection is not based on a full Ethernet or MPLS frame for its switching intelligence.

Foundation technologies for 5G mobile networks

5G promises significant end-to-end network performance gains compared to 4G LTE mobile networks. These gains require significant updates and modernization of existing wireless and wireline network technologies, which means 5G is about far more than a simple update. 5G will be a multi-year journey, with MNOs leveraging a vast array of new and emerging technologies to deliver far better performance than anyone experiences today. What subscribers, both humans and machines, will ultimately experience will depend on a variety of factors, most notable the economics of 5G services. As history has shown time and again, when improved network performance is made available, it is always voraciously consumed, often by new use cases and applications yet to be defined. 5G services will be deployed globally on a broadening scale over the coming years, and users will soon ask themselves how they ever lived without it.

Visit the Ciena Community
Get answers to your questions