Independent market research and competitive analysis of next-generation business and technology solutions for service providers and vendors

HEAVY READING

WHITE PAPER

New Transport Network Architectures for 5G RAN

A Heavy Reading white paper produced for Fujitsu

FUJITSU

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

5G RAN NEW REQUIREMENTS; NEW ARCHITECTURES

Advanced operators are preparing for rapid deployment of 5G networks. Their aim is to evolve today's mobile broadband services, through lower unit costs and improved end-user performance, and to address new business segments using 5G system capabilities. To achieve this, operators are designing network architectures that will scale to device and traffic densities far beyond what is commonplace in LTE networks today and meet the latency and reliability requirements of demanding new service types. This is critical preparation for the 5G super-cycle that will enable operators to extend their reach into diverse markets over the next 10 years.

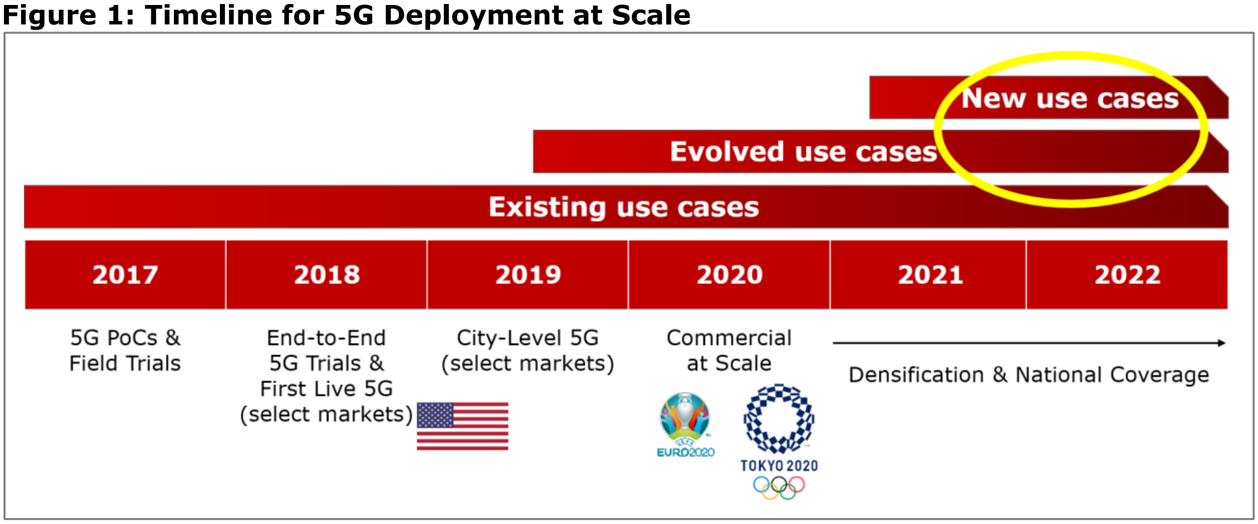
An essential part of the new design is the transport network that provides connectivity between radio sites, edge data centers and cloud applications. This white paper investigates this new transport architecture and the role that packet-based connectivity fabrics play in the performance of the 5G RAN itself and on revenue-generating end-user services. Because 5G will be introduced into large, well-optimized and commercially productive LTE networks, integration into existing environments is important; therefore, this paper also discusses the importance of integrated 4G/5G transport.

The paper argues that 5G RAN transport and connectivity is more than a necessary cost. For network operators, it is an investment in critical assets that will generate unique advantages in the way customers experience, and interact with, services. It is an architecture for the long-term development of the 5G service offer.

Market Outlook & Deployment Strategies

The first specifications for the 5G system, covering radio and core, are now available. This gives the industry the confidence to accelerate commercial launch, and several high-profile operators have stated that they expect to launch commercially in 2018 and 2019; we then expect many operators to follow in 2020 and 2021. Within five years of launch, it is plausible, in advanced markets, that 50 percent of an operator's subscribers will be 5G customers. **Figure 1** shows a timeline for commercial launch, coverage expansion and operation at scale. The first commercial launch to be announced is Verizon's 5G Home service, a fixed wireless broadband service that will go live in October 2018.

Figure 1: Timeline for 5G Deployment at Scale



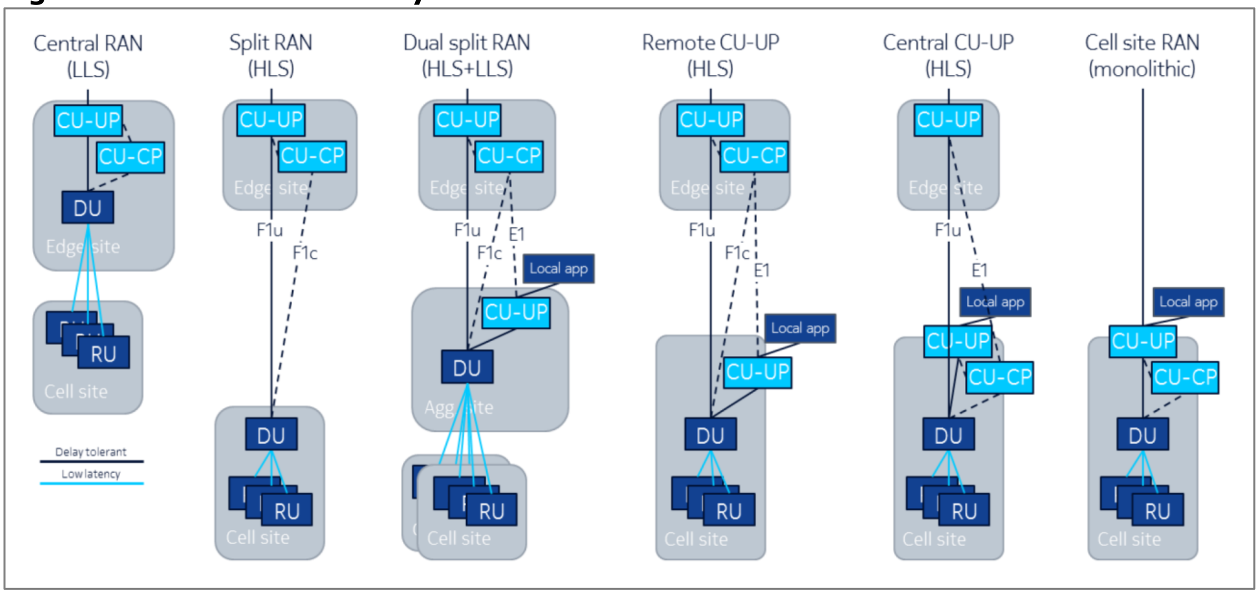
Source: Heavy Reading

5G RAN Architecture

The 3GPP 5G RAN architecture – specified in Release 15 and known as NG-RAN – introduces new terminology, interfaces and functional modules. The NG-RAN consists of a set of radio base stations (known as gNBs) connected to the 5G core network (5GC) and to each other. The gNB incorporates three main functional modules: the Centralized Unit (CU), the Distributed Unit (DU), and the Radio Unit (RU), which can be deployed in multiple combinations.

The primary new interface is the F1 interface between DU and CU This are expected to be interoperable across vendors. Standardization of a further lower-layer interface between DU and RU is under consideration, but progress is likely to occur outside the 3GPP in the first instance. The CU can be further disaggregated into the CU user plane (CU-UP) and CU control plane (CU-CP), both of which connect to the DU over F1-U and F1-C interfaces respectively. This new 5G RAN architecture is described in 3GPP TS 38.401.

Figure 2: Location Flexibility for 5G RAN Functional Units



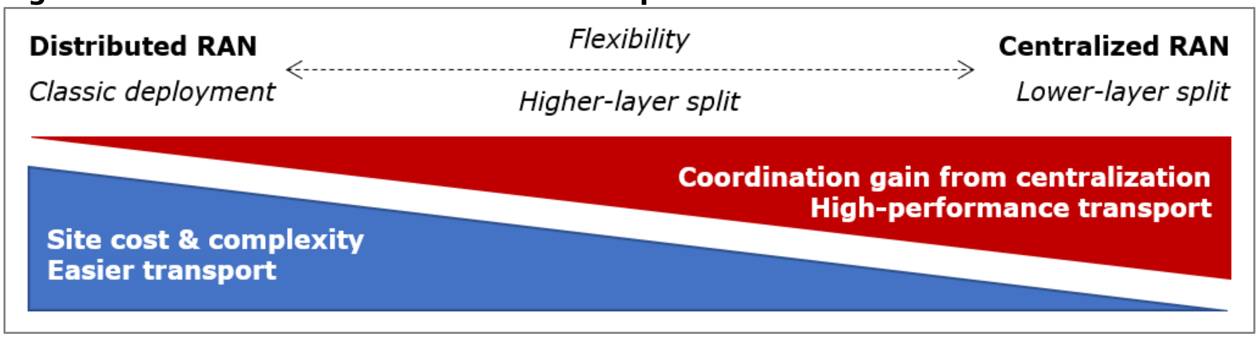
Source: NGMN, 2018

As with all 3GPP standards, NG-RAN is a logical architecture that can be implemented and deployed in different ways, according to an operator's requirements and preferences. As shown to the right of **Figure 2**, the base station can be deployed as a monolithic unit deployed at the cell site, as in classic cellular networks, or split between the CU, DU and RU. The CU-DU interface is a higher-layer split (HLS), which is more tolerant to delay. The DU-RU interface, which is not yet standardized, is a lower-layer split (LLS), which is more latency-sensitive and demanding on bandwidth, but may offer improved radio performance across a coverage area due to coordination gain. CUs, DUs and RUs can be deployed at locations such as cell sites (including towers, rooftops and associated cabinets and shelters), transport aggregation sites and "edge sites" (e.g., central offices or local exchange sites).

Which 5G RAN Deployment Architecture?

The choice between lower-layer and higher-layer splits is a critical decision in NG-RAN architecture. The challenge is that the trade-offs are not always clear-cut, and it can be hard to determine which model will be economically optimal for each carrier. Moreover, it may make sense to use different models for different regions – for example, rural coverage versus urban capacity. There will also be variation for different use cases – for example, an ultra-low-latency factory automation network would require DU/CU close to, or integrated with, the RU. **Figure 3** shows the high-level trade-off between coordination gain from centralization and the latency and bandwidth requirements in the transport (fronthaul) network.

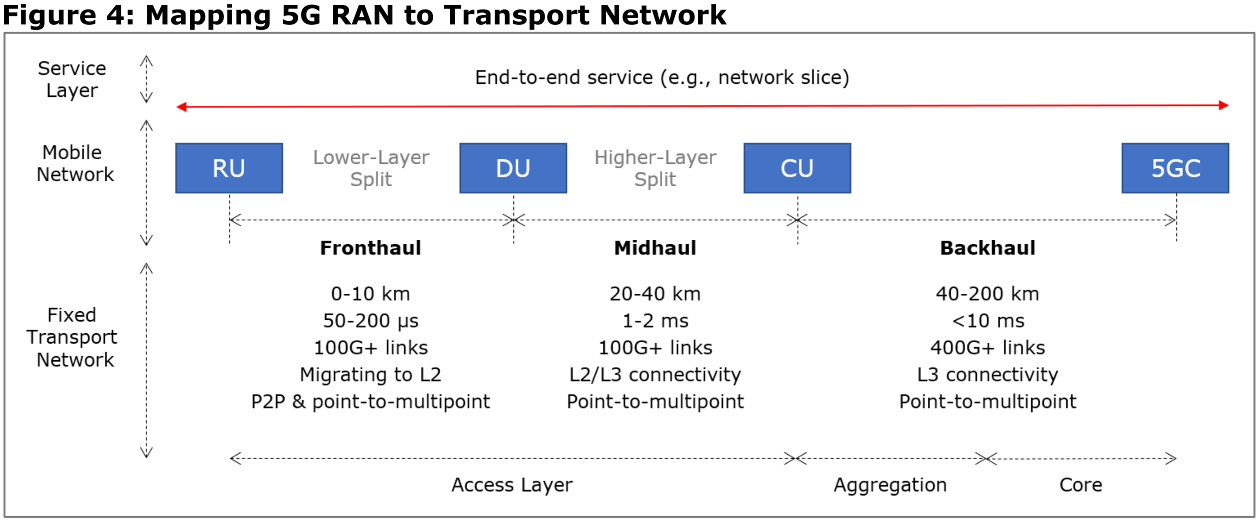
Figure 3: Coordination Gain & Functional Splits



Source: Heavy Reading

It is not straightforward to determine the optimal RAN topology and associated operating model. Usage patterns, device capabilities, operating costs, RF strategy and the existing network footprint and capabilities are all influential on the decision. The transport network is especially important because it provides connectivity between the NG-RAN functional modules and therefore determines, in the first instance, which deployment topologies are possible, and then, which are cost-effective. **Figure 4** shows the primary functional elements of the 5G RAN (RU, DU and CU) and how they are logically mapped to the transport network.

Figure 4: Mapping 5G RAN to Transport Network



Source: Heavy Reading

A lower-layer split between the DU and RU is demanding in terms of bandwidth, latency and packet loss, and to deploy a DU at a distance from the cell site typically requires the operator to invest in fronthaul fiber infrastructure. A higher-layer RAN split, with DUs deployed at the cell site, is more tolerant in terms of transport performance, but perhaps offers less coordination gain and, potentially, higher opex in the longer run.

And RAN technology decisions alone do not determine the architecture: If the operator wants to support ultra-low-latency and ultra-reliable services, then the network will need to be designed accordingly – for example, for certain applications the radio compute (a.k.a. the DU/CU) may need to be deployed at the cell site, or on-premises, to meet latency and availability targets of the end-user service. The 5G RAN deployment architecture is therefore codependent on RAN design, the transport network and end-user services.

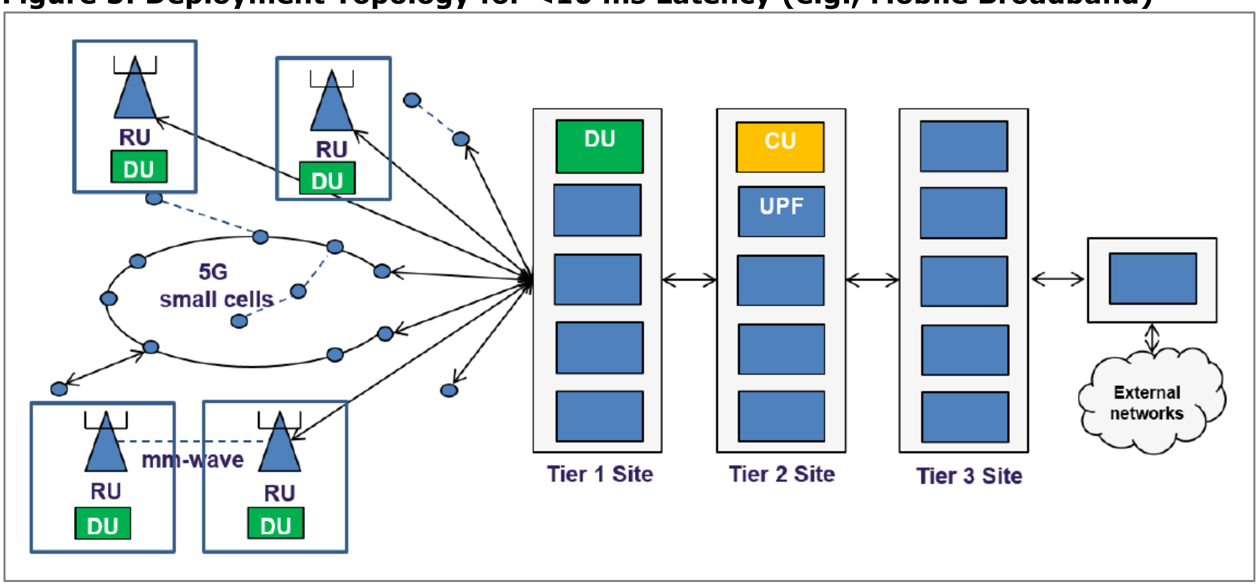
Mapping Functional Modules to the Physical Network

Decisions on where to place 5G RAN and core network functions depend on service requirements and the transport and compute capabilities deployed in the network. **Figures 6, 7** and **8** show example deployment topologies. Macro cells and small cells connect to external networks through user plane functions (UPFs) in the 5G core via a series of transport aggregation sites. These aggregation sites decrease in number toward the core, with a corresponding increase in transport capacity – for example, for a network with 1,000 Tier 1 sites, there might be 100 Tier 2 sites and 10 Tier 3 sites.

Where specifically RAN and core functions are deployed will vary between, and within, operators. End-to-end network latency probably the critical determinant. The Next Generation Mobile Network (NGMN) Alliance guidelines for one-way latency from the cell site to a Tier 1, Tier 2 and Tier 3 aggregation site are 0.6 ms, 1.2 ms and 4.2 ms respectively.

Therefore, for a mobile broadband service with tolerance for a 10 ms roundtrip delay, the deployment model shown in **Figure 5** may be optimal. In this case, the RAN CU and the core UPF are colocated at the Tier 2 site (e.g., a central office), while the DU may be distributed to a Tier 1 aggregation site or to the cell site itself (or be located close by, in a cabinet or shelter).

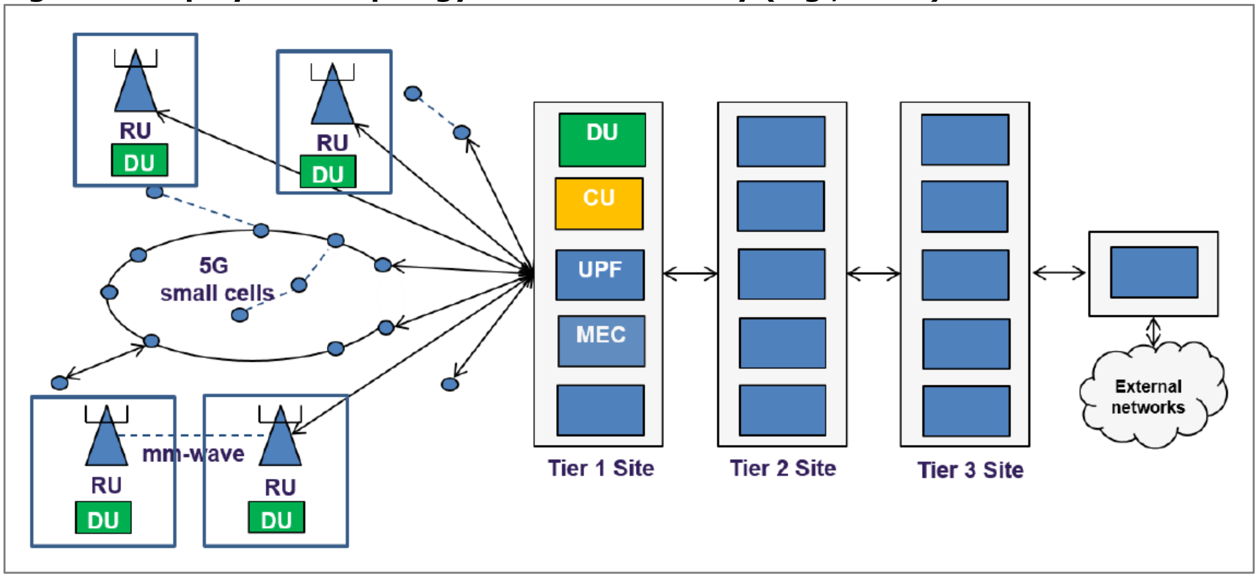
Figure 5: Deployment Topology for <10 ms Latency (e.g., Mobile Broadband)



Source: NGMN

In **Figure 6** the UPF is deployed closer to the edge at the Tier 1 site, to give around 5 ms roundtrip latency. This edge facility may also host a multi-access edge computing (MEC) platform to host and run latency-sensitive applications.

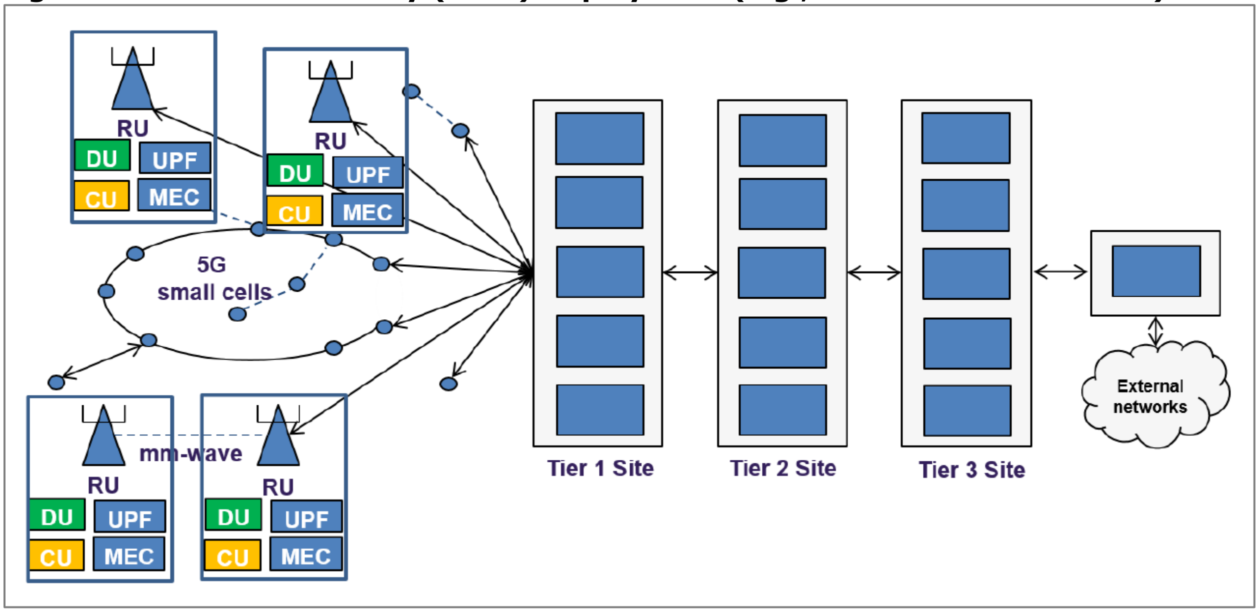
Figure 6: Deployment Topology for <5 ms Latency (e.g., AGVs)



Source: NGMN

In **Figure 7** the CU and DU are colocated with the RU, or even integrated into a single unit, at the cell site. Because very low latency is required, the core network (UPF) and MEC server may also be deployed at the cell site to enable applications to run locally. An example service may a next-gen factory automation using 5G for robotic motion control in an industrial production line.

Figure 7: Ultra-Low-Latency (1 ms) Deployment (e.g., Industrial Automation)



Source: NGMN

AN OPEN X-HAUL TRANSPORT NETWORK

The new 5G RAN and end-user services requirements drive the need for a new transport network. While many 4G assets can be reused, it will not be sufficient to overlay 5G on a 4G backhaul architecture, and even where an operator has deployed an advanced C-RAN architecture, new investment is required to meet 5G requirements.

5G RAN introduces new physical topologies, more functional split options, and ultra-low-latency services, which demand shorter distances between user and compute. Therefore, there is a need for new, flexible, multipoint packet networks in the RAN that can scale and flex to the many deployment models expected for 5G.

Open X-Haul

This new RAN transport should support fronthaul, midhaul and backhaul – a.k.a. "X-Haul" – applications. Fiber and SDN are the building blocks of the architecture, with operators investing in fiber access, metro SDN and edge data centers to create a "transport services fabric" for the 5G RAN. This is a strategic investment in next-gen transport that will improve operating economics and underpin the 5G service offer.

The major requirements for Open X-Haul 5G RAN transport can summarized as:

- **Standards-based** – Leading to interoperability between segments and/or components of the network. For example, using NETCONF/YANG between equipment from one vendor and another vendor's SDN controller, or one vendor's RU to interconnect with another vendor's DU.
- **Packet-based system** – To enable meshed connectivity between RAN components, to replace CPRI fronthaul over wavelengths, and to create a more programmable, adaptable network.
- **Sync** – 5G is primarily TDD-based with tight sync requirements, even for classic cell site deployments; as multipoint radio and cloud RAN are adopted, these sync requirements become even more stringent.
- **Latency** – Around 8 ms roundtrip time is needed for mobile broadband services in the macro network; for ultra-low-latency services, this drops to in the region of 1 ms. For fronthaul schemes, between 50-100 μs is required (per IEEE 1914.3).
- **Multiple fronthaul protocols –** LTE fronthaul is associated with CPRI. In the 5G era, combinations of new interfaces and protocols including eCPRI, Time Sensitive Networking (TSN), xRAN fronthaul, Telecom Infra Project (TIP) fronthaul, and the proposed "FX fronthaul" interface, may need to be supported.
- **Bandwidth & 100-Gbit/s access** – Major macro cell sites today are often connected by gigabit or 10-Gbit/s Ethernet backhaul. 5G will drive a need for 100-Gbit/s interfaces deeper into the access network. This requires high-volume, cost-effective optics.
- **Meshed L3 connectivity/routing** – Offers the flexibility to deploy RAN functions in the optimal location from cell sites, aggregation points and edge data centers, and to move functions as needs evolve. A meshed architecture is also advantageous to reliability and resiliency.

- **Programmable & automated** – Constraining operational costs while radically increasing the customer offer is fundamental to the 5G business case. To deliver increased RAN transport capability therefore requires extreme automation.
- **Integrated with cloud infrastructure** – The network that connects physical sites must also support virtual network overlays to connect users and applications. SDN-driven micro-segmentation is important for offering network slices, and this requires integration between the underlay transport and the (virtual) cloud network.

Fronthaul Options (RU-DU Interface)

The most demanding interface in 5G RAN transport is between the RU and DU. The RU (radio) provides functions such as analog to digital conversion, filtering, power amplification and TX/RX; with the use of massive multiple input/multiple output (MIMO), the RU is integrated with the antenna. The baseband unit (the DU) provides digital processing, including signal modulation, encoding, scheduling, etc. The two systems must be tightly integrated for optimal system performance and, classically, are delivered by a single vendor over a proprietary or quasi-proprietary interface – in 4G networks this is typically via CPRI, but increasingly eCPRI or Radio over Ethernet.

4G fronthaul typically uses CPRI, which was designed for fiber runs from the cabinet at the base of a cell tower to a tower-mounted remote radio head (typically a few tens of meters), but has since been extended over long distances (up to tens of kilometers) to enable centralized RAN architectures. The challenge with CPRI over long distances is that a) it is very demanding on transport and requires wavelengths or dark fiber pairs; b) it is, in effect, proprietary to the RAN vendor; and c) it doesn't scale well to massive MIMO and some other new RAN technologies.

More recently, new radio innovations have started to open up the DU-RU interface, enabling operation over packet transport networks. Specifically, the industry has developed methods for sending digital radio frames over Ethernet and has created functional splits with more Layer 1 processing in the radio to relax the requirements on packet transport. As of 2018 this RU-DU interface is not specified by the 3GPP; however, it is under study for inclusion in a later release. Meanwhile, there are several industry proposals to address fronthaul and midhaul interfaces and transport methods. These are shown in **Figure 8**.

Figure 8: Fronthaul Interfaces & Transport Initiatives

	Description	Comments
CPRI	Digital RF over optical interface between radio & baseband Designed for towers; extended to C-RAN Only 8-16 percent efficient without compression	Quasi-proprietary and demanding on transport; losing ground as operators seek greater flexibility
eCPRI	Enhancement to run CPRI over switched Ethernet 10X reduction in bandwidth compared to CPRI Does not sufficiently address management plane	Insufficiently standardized; too many variables and options
xRAN Fronthaul Interface	Uses eCPRI format over IP Interoperable user plane & management plane Expected to be adopted by the ORAN Alliance	Emerging as the specification most likely to be adopted for lower-layer splits; gaining support

	Description	Comments
Time Sensitive Networking	IEEE 1914.3 is mapping radio interface to TSN IEEE 802.1CM is a "TSN for fronthaul" profile Supports multiple higher- and lower-layer splits Supports tight time synchronization	Increasing interest in TSN; emerging as a preferred Ethernet transport for 5G RAN; standardization expected end 2018
TIP Open Fronthaul	vRAN fronthaul project for lower-layer split Will function over non-ideal fronthaul Focused on LTE rather than 5G	Useful contribution; looks likely to be folded into xRAN/ORAN
3GPP Higher-Layer Split	HLS interface (F1) between DU-CU E1 interface between CU-UP and CU-CP Approved in TS 38.47x	F1 and E1 are open interfaces specified by 3GPP; main standard for midhaul
TTA	Proposed Fx open interface for lower-layer split between gNB-CU and gNB-DU South Korean-led initiative; first proposed by SK Telecom	Little international support currently relative to other lower-layer splits

Source: Heavy Reading

Two example combinations that appear to be have the most momentum at this stage are:

- **For the higher-layer split (midhaul)** – the 3GPP functional split 7x (corresponds to the F1 interface) over IP/Ethernet transport, which is relatively easier to deploy; will probably dominate most early split-RAN 5G architectures.
- **For the lower-layer split (fronthaul)** – xRAN interface over TSN-based Ethernet transport. This appears an attractive model, but it requires further development.

Time Sensitive Networking

TSN is part of the 802.1Q family of standards and is designed to provide deterministic messaging on standard Ethernet networks. It is a Layer 2 technology that is centrally managed and uses coordinated scheduling to ensure performance for real-time applications. Real-time, deterministic communication is important to many industries – for example, aerospace, automotive, transportation and utilities, and of course, manufacturing. In many of these industries, TSN is emerging as the baseline for real-time networking.

Features that make TSN an attractive candidate for 5G RAN transport include zero congestion loss, bounded ultra-low-latency, hitless 1+1 redundancy and the ability to carry sync accurate to <1 µs. "802.1CM TSN for Fronthaul" is a collaborative effort by the CPRI and IEEE organizations. Two profiles applicable to both CPRI and eCPRI are under development. Profile A sends user data (IQ data) as a high-priority traffic class above control and management data. Profile B uses "frame preemption," a TSN feature, to prioritize different traffic types (sync, user data, management data, etc.).

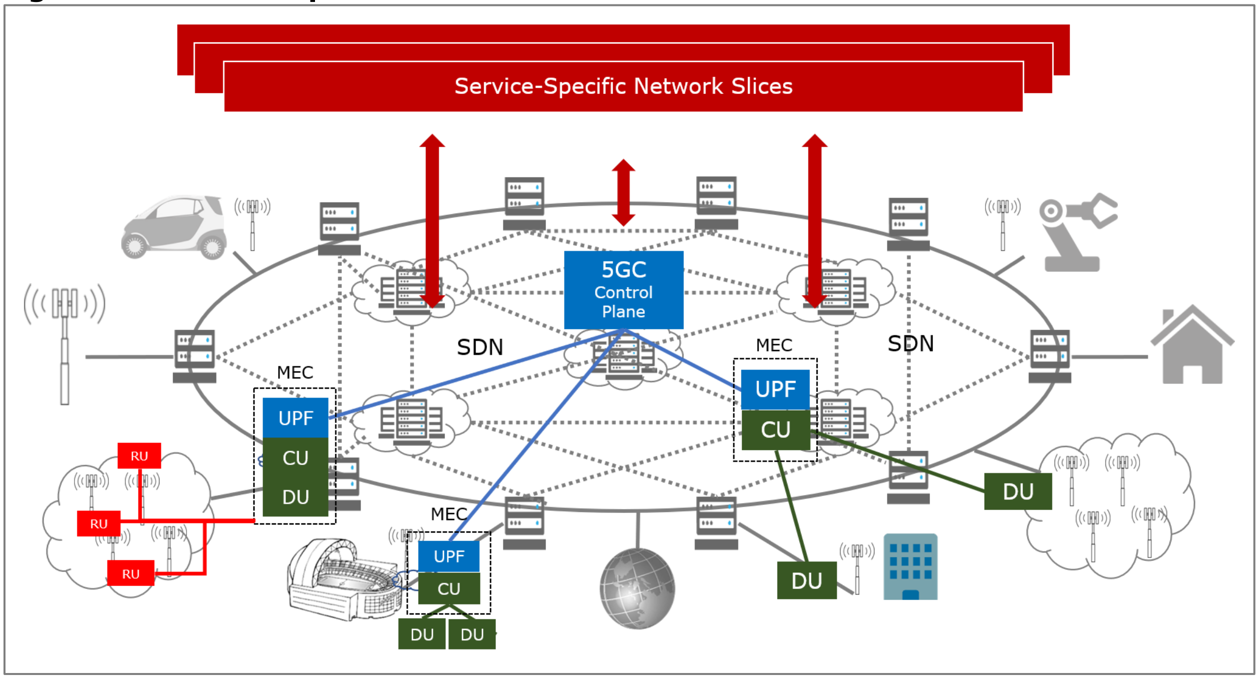
An advantage of TSN is that time-critical and background flows share the same packet network. And because it is Ethernet and is being adopted by other industries, there are a wide variety of applications and users, contributing scale to the technology relative to a mobile-only standard. A common X-Haul transport network for 5G RAN, serving fronthaul, midhaul and backhaul, can therefore be developed using TSN.

METRO-AREA 5G TRANSPORT

RAN requirements drive the design of 5G transport in the access. However, the 5G transport network must also connect compute and storage clouds deployed at edge locations, such as central offices or aggregation points. In many cases, the operator is also thinking strategically about developing a metro edge using integrated packet-optical platforms to create a universal aggregation network that can serve many access types and customer segments. In this sense, the RAN transport decision is strategically important beyond 5G itself.

When applied to 5G, **Figure 9** shows a scenario where the 5G core control plane is deployed (somewhat centrally), with UPF distributed to edge locations. In some cases, these same locations will also host CUs, DUs and services in various combinations. In this chart, RUs are shown in red connected over fronthaul to an integrated UPF/CU/DU located, for example, in a central office. The transport network should provide meshed connectivity to allow the operator the flexibility to deploy multiple models, and critically, to adopt new deployment models over time.

Figure 9: A 5G Transport "Service Mesh"



Source: Heavy Reading

To stand up infrastructure is one thing; more important is to deliver the end-user service. In 5G the intent is to use network slices, configured to customer-specific and/or service-specific requirements, to provide an end-to-end service across domains: RAN, transport, core and cloud. A simple example might be a fixed wireless access slice on 5G routed discretely to, say, a mobile broadband slice. A more granular example might be a connected car, which plausibly needs slices for infotainment, telemetry, assisted driving, etc., each with different prioritization and reliability needs, and different business models. In each case, the operator must map slices in the 5G domain (using QoS and slice identifiers in the 5G RAN and core) into the transport network, and in some cases, the cloud platform, which hosts the application. Multidomain orchestration is required to create and manage these services.

Operations & Automation

The expectation is that this new metro-edge network will be SDN-based, programmable, and open to multi-domain orchestration. From an operations perspective, this means using modern tool chains to configure, monitor and scale the network. It is well-established that traditional network management, driven by vendor element management systems and operator OSS stacks, will not be able to meet the demands of a dynamically configurable, cloud-native network constructed with multi-vendor equipment. The focus on low-cost operations and automation is especially critical at the network edge, because there will be a significantly greater number of elements and functions to manage, and because equipment may be deployed in locations that are not staffed and may be difficult to access.

There are many networking initiatives focused on automation of operations across physical and software functions. SDN-based dynamic path computation, policy-based routing, cross-domain orchestration, etc., can all make contributions. Model-driven, intent-based configuration is of primary importance to lowering opex.

The NETCONF protocol, and its associated modeling language YANG, have an important role in programmability and automation, specifically at the management plane. The combination of NETCONF/YANG is generally associated with configuration of switches and routers and is gaining broad support as operators migrate from CLI-based device management to API-driven devices. APIs provide an abstraction of the underlying technology, so that a (network) programmer can disregard unnecessary details but have access to: what data is available (attributes); how the data is organized (schema); and how the data can be accessed (methods).

From a 5G radio perspective, the xRAN Forum has adopted the use of NETCONF/YANG for programmatically configuring and managing the RUs in its lower-layer split RAN architecture. Use of such a standardized framework and common modeling language simplifies integration between DU and RU, even (and especially!) in multi-vendor scenarios. Using YANG, an operator is able to model the configuration and operational state of its 5G RUs, potentially enabling them to programmatically manage capacity pooling and load-balancing across a RAN coverage area, or to allocate resources to a network slice.

This lays the foundation for cross-domain orchestration of the RAN with other domains that have already adopted NETCONF/YANG. With a full-mesh packet network architecture, this will allow operators to support end-to-end services, with the appropriate resources allocated in the radio, core and transport. An operator could use this tool chain to integrate access, transport and cloud services into a virtual network slice – and will be able to do so dynamically, for example to set up a mission-critical network slice. This capability will massively expand the addressable market for operators' 5G services.