5G Backhaul/Fronthaul Opportunities and Challenges

A Technical Paper prepared for SCTE•ISBE by

Joe Mocerino
Global Solution Architect
Fujitsu Network Communications

Table of Contents

List of Figures

Introduction

Planning and deploying a mobile network to support a myriad of 5G applications will be no easy feat, considering the complexities of these new architectures and the interdependencies between the RAN and transport network.

RAN transport rates for 5G will be over 15 times greater than those available in 4G LTE and its variants. However, using the same operation of the 4G RAN for 5G would yield transport rates for optics and platforms price prohibitive. To mitigate this situation, the 3GPP standardized a new RAN model splitting the processing functionality of the 5G BBU into several blocks, thereby reducing the transport rate requirements.

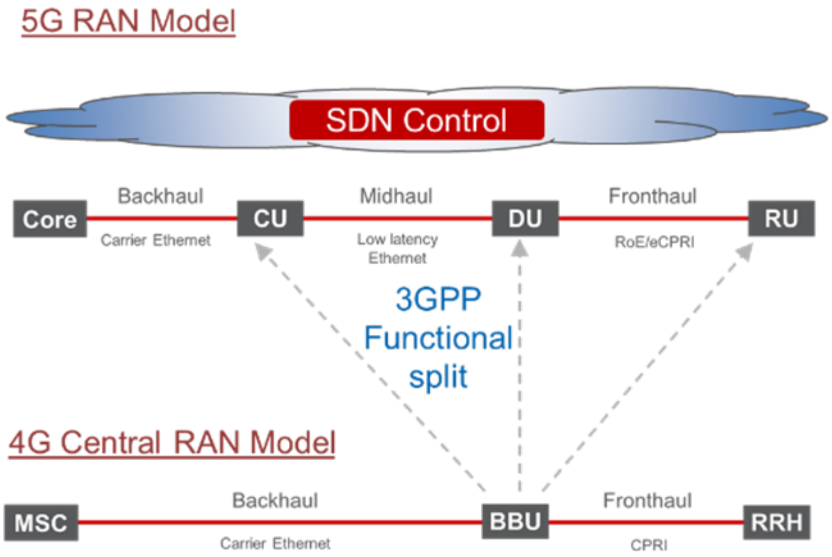

Figure 1 - 4G RAN evolution to 5G

The key building blocks of the Next Generation RAN (NG-RAN) architecture are the centralized unit (CU), distributed unit (DU) and remote radio unit (RU). Fronthaul transport between the RU and DU will use the more efficient eCPRI protocol which provides higher performance at a lower cost per bit than CPRI used for 4G services.

The IEEE has standardized the latency budgets for the new 5G RAN. These budgets are similar to the original 4G fronthaul and backhaul segments with the exception of the new ultra-reliable-low-latency connectivity (uRLLC) use cases. In the fronthaul, these result in a 50 microsecond latency budget. Backhaul remains at 10 milliseconds and fronthaul at 100 microseconds for all but uRLLC applications. The new area of transport is the “midhaul” or next generation fronthaul II (NGFH II) section, which will vary from one to three milliseconds in latency budget as per the IEEE 1914.3.

With 5G services, a new form of RAN topology is emerging. The predominate topology in the RAN today is the distributed RAN. The distributed RAN consists of all the 4G elements- remote radio head (RRH) and baseband unit (BBU) at the cell site. This topology has the lowest latency. Next is the centralized RAN where the BBU is centralized at a location within 20 kilometers of the cell site. The centralized RAN configuration enables the BBUs at the central location to pool resources to address the demands of the cell sites. This eliminates the risk of over or under engineering the individual cell site with a specific capacity of BBU. Cell site aggregation also enables two or more cell sites to address demands of an individual mobile user.

The "virtualized" RAN is the new model for 5G. The processing elements of the RAN, i.e. the DU and CU- will ultimately be virtualized because vertical network slicing will initiate in the DU. This is the most flexible topology as it can be dynamically repurposed. The Next Generation Mobile Networks (NGMN) consortium of service providers has developed several RAN topologies. These models vary from a distributed RAN with site cost and complexity balanced with less demanding transport to a centralized RAN, which provides a coordination gain and yields a high-performance transport layer.

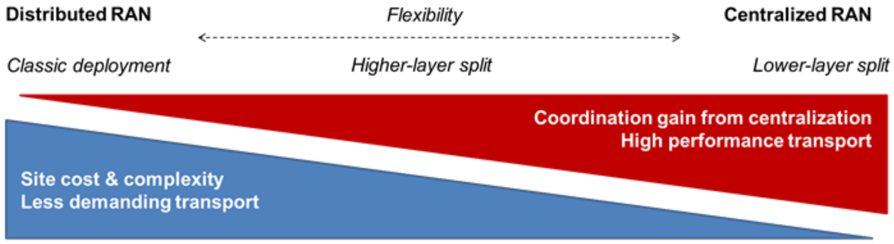

Figure 2 - Comparison of Distributed to Centralized RAN Functional Splits

This flexibility is accomplished by splitting the functionality of the 5G RAN elements to deliver the performance requirements needed for the upcoming 5G use cases. The virtual RAN and transport topology will work very closely together. Service providers will be able to develop a single infrastructure to address the upcoming use cases and multi-tenant operation without having to dedicate assets to one topology type or having multiple network element overlays.

Radio Type Drives Optimum Transport Option

The radio types used for 5G are millimeter wave (mmWave) and sub-6 gigahertz. Each has its pros and cons. The mmWave radio is using the higher frequencies offering high capacity service but coverage is limited to about 100 meters because its high frequencies are challenged with walls and obstacles. Therefore, separate indoor and outdoor RAN networks will be used. In the outdoor environment a densification strategy is needed resulting in deployments on streetlight and utility poles, sides and tops of buildings much like small cell installations. These non-traditional sites will require much more fiber facilities and have very limited power and footprint available.

Sub-6 gigahertz radios have less capacity than mmWave but have better coverage of about one kilometer. They do not have the penetration issues of mmWave and are installed on traditional cell towers.

Optimum Transport for mmWave

Mobile Network Operators are deploying both radio types for 5G applications. Today these radios are used for fixed wireless access (FWA) to offer high speed Internet to residential and small to medium businesses. Transport options include dedicated dark fiber (DDF) as the first choice, but lacks integrated remote visibility. When DDF is in short supply, a multiplexing capability is needed to extend the capacity of the fiber. Two such multiplexers are the traditional wave division mux (WDM) and the new time sensitive networking (TSN) otherwise referenced as an Ethernet mux. Both technologies transport 4G CPRI (10 Gbps), 5G eCPRI (10 and 25 Gbps) and gigabit Ethernet up to 25G.

The WDM has several variants ranging from an all-passive technology, to semi-passive to a full-active technology. The all-passive technology is the lowest cost but lacks any remote visibility for performance monitoring. In semi-passive, the cell site end is full passive while the hub end uses active transponder technology. Smart optics are used in the semi-passive to establish self-tuning automation and a communications channel for remote visibility. Intelligence at the hub end provides integrated optical DDM, OTDR, performance monitoring and latency measurement capabilities working with the smart optics communications channel. Full-active WDM is the most expensive means of WDM transport but offers the most capability including full remote visibility and topology options for self-healing operation.

For mmWave deployments, given the challenges at the cell site in power and footprint, the semi-passive system is optimum in terms of providing a cost effective solution with a level of remote visibility. The cell site end is fully passive and does not require any power. The footprint for the outside plant WDM and enclosure is just a little bigger than a tall coffee container saving on footprint where it is at a premium. Linear or point to multi-point topology is available on the semi passive system for ease of deployment.

TSN or Full Active WDM for Sub 6 GHz Radio Fronthaul

When deploying sub 6 GHz radios at traditional cell towers based on the coverage capability, there will be multiple sectors resulting in many channels for transport to a central location. Service providers will need to transport in the fronthaul the new 5G services along with the legacy 4G channels, which are highly inefficient. This presents a major challenge in the total number of channels to be transported.

Using the traditional WDM approach would require many expensive 10 Gbps and 25 Gbps optics to drive services over the fronthaul. This results in high cost and large footprint for the transport layer. However, up to a mix of 40 channels of: 4G CPRI, 5G eCPRI and Ethernet channels can be transported over a single fiber strand, justifying a WDM approach when channel counts are high and fiber assets are near depleted. Alternatively, if the capacity of traffic for transport is such that minimal fiber is required, the time sensitive networking (TSN) is a more cost effective option. The TSN approach will utilize a packet to multiplex channels in the fronthaul and can also translate the inefficient CPRI to eCPRI protocols reducing the total bandwidth capacity. The ORAN Alliance has specified a functional split, 7.2x, to translate highly inefficient CPRI traffic to eCPRI using the low order physical layer 1 processing (Low PHY). This function would reduce CPRI bandwidth capacity up to 5:1. Another translation approach for CPRI to eCPRI is the CPRI Cooperation's eCPRI v2.0. The eCPRI v2.0 does not do a full translation, instead it reformats the IQ data in the CPRI frame to that of eCPRI. This is useful in evolving distributed 4G cell sites to 5G while maintaining the legacy 4G service. When evolving the 4G cell site to 5G the backhaul capacity will increase from 1 GE to 10 GE.

Figure 3 - Add 5G to 4G Distributed RAN

The compact TSN mux offers high density transport for 4G, 5G and Ethernet services, translation from CPRI to eCPRI, at costs lower and in a smaller footprint than full active WDM using one or more 100 GE network connections.

Virtual Networks via Network Slicing

The virtual RAN will offer the service provider the greatest flexibility via a single transport infrastructure with multiple micro-services. This is accomplished using a mix of software-defined networking (SDN), network functions virtualization (NFV), and end-to-end network slicing for the RAN, edge transport and core networks. This RAN virtualization sets the groundwork for a single physical network infrastructure representing multiple virtual network configurations each representing a network slice, hence the term "network slicing." Each network slice is a complete virtual network within the infrastructure.

The edge transport network establishes a common infrastructure using programmable and disaggregated network elements. Edge transport routers are used from the DU, where the network slice point begins, to the core offering dynamic multipoint connectivity. To assist in maintaining a predictive low-latency operation, MPLS segment routing (MPLS-SR) is the most common infrastructure technology used to facilitate network slicing.

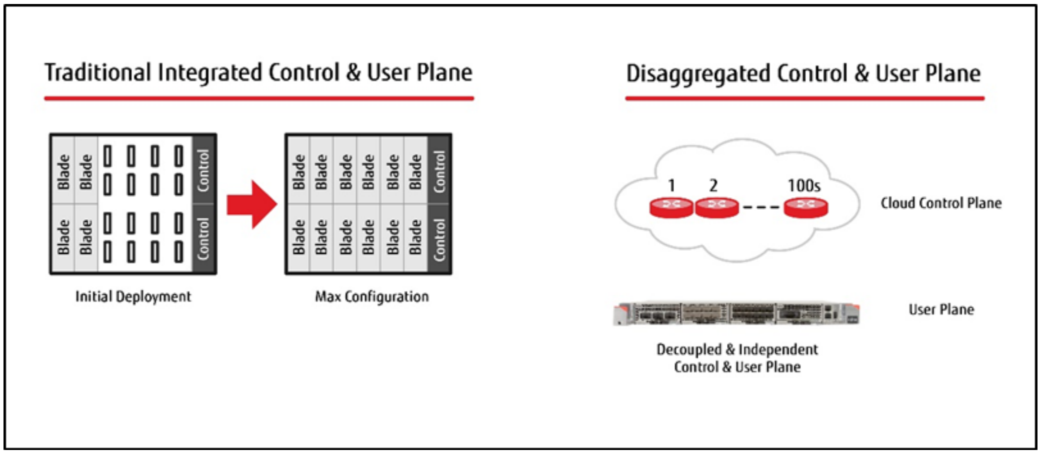

Figure 4 - Traditional vs. Cloud Control Plane Router

Traditional router architectures are vertically integrated, self-contained network elements. They consist of a chassis with line cards deployed in predefined slots along with switch fabric and control cards in other slots. Connectivity between line cards and switch cards is enabled via electrical backplane traces commonly referred to as serializer/deserializer (SerDes). The number of traces between slots and the speed with which the traces are clocked determines the maximum inter-slot communication capacity. This architecture requires the alignment of three hardware components: the line card, the switch fabric cards, and the backplane. Service providers are challenged in three areas when specifying a router platform for their 5G network:

- Determining the right capacity and performance for the site demands
- Minimizing the physical and environmental allocations, and
- Scaling platform capacity and performance for the long term.

Control in the Cloud

Router vendors typically offer a mixture of low-, medium-, and high-capacity performance units. Sizing the integrated router capacity is challenging because the control plane, backplane speed, and chassis capacity limit the performance and scaling of the user plane blades. Under-allocating the router performance can risk loss of opportunity, whereas over-allocated router performance results in capex inefficiency.

During initial installation only 20% to 30% of the router capacity is utilized but the chassis footprint, power, and thermal reserve all have to be fully allocated, resulting in cost-inefficiencies. Anytime the capacity of the slots is increased, all three elements must move in lockstep.

Because service providers loathe the idea of forklifting the chassis/backplane, vendors try to future-proof their node designs to support capacity expansions, including cooling, power, and backplane traces. However, because any chassis design utilizes the most cost-effective, commercially technology available at the time, there are limits to how far vendors can future-proof the network element. Once these measures are exhausted, additional capacity enhancements require replacement with a newer chassis.

To resolve the limitations of a traditional router, the next-generation router will employ a programmable disaggregated control and user plane architecture. The control plane is completely independent of the user plane, and in advanced models it is hosted and executed in the cloud. Incorporating cloud native technology and routing protocol isolation into the disaggregated router via a cloud control plane resulting in a single 1RU blade element capable of dynamically producing hundreds of router instances for RAN services and customer isolation. The virtual routing segments, quality of service (QoS), and resiliency requirements are provisioned in the cloud using automation for the virtualized service.

Figure 5: Network Slicing Using Cloud Control Plane Router

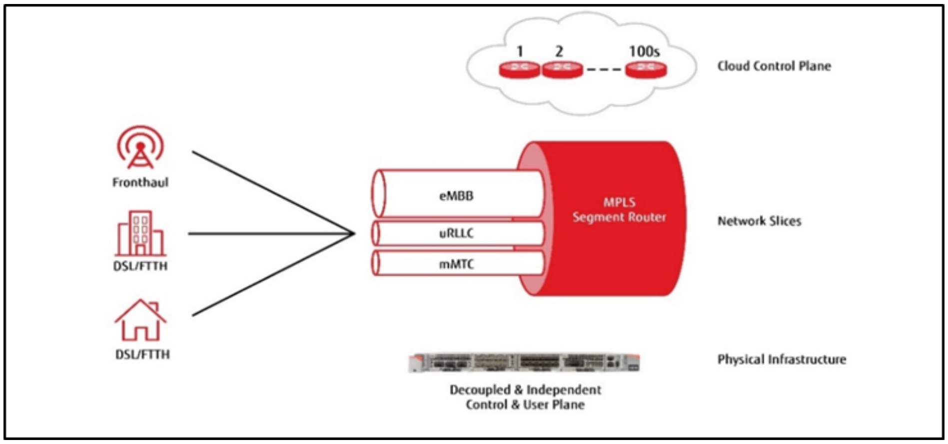

Figure 5 - Network Slicing Using Cloud Control Plane Router

Once the cloud control plane calculates mapping for each service, the control information is then pushed down to the router user plane infrastructure. If a physical site has a catastrophic failure, its virtual routing profile can be moved in the cloud control plane to another physical site, simplifying resiliency operations. Applying this architecture to the router optimizes physical/environmental cost-efficiencies, simplifies network engineering, reduces infrastructure capacity risks, and offers superior performance scaling.

Scalable Slices

The network orchestrator coordinates this ecosystem between the core, edge transport, and RAN elements. As network slices are established via DU asset allocations and multiple CU-UP terminations, the transport network establishes router instances to support these individual services and customers providing the transport QoS guarantees.

A traditional router architecture with integrated control and user plane is initially cost-inefficient, has risks of over or under performance based on chassis size, and has limited scaling functionality over the long term. On the other hand, the disaggregated cloud control and user plane router approach establishes a single transport infrastructure with the ability to dynamically virtualize multiple networks cost-effectively in a highly scalable fashion. As today's networks continue to evolve, this dynamic flexibility will be key to allowing tomorrow's architecture to meet diverse needs for capacity, latency, and performance, fulfilling the 5G promise.