5G Fronthaul Over DOCSIS

Transporting O-RAN's Split 7-2x over DOCSIS

A Technical Paper

Mark Grayson
Distinguished Engineer
Cisco Systems
10 New Square Park, Feltham, Middlesex, TW14 8HA, United Kingdom
+44 20 882 43403
mgrayson@cisco.com

John Chapman
Fellow, CTO Broadband Technologies
Cisco Systems
408 526-7651
jchapman@cisco.com

Table of Contents

List of Figures

List of Tables

1. Introduction

The transition to open RAN (Radio Access Network) based on interoperable splits is gaining significant momentum across the mobile industry. Conventional composed base stations are being decomposed into separate radio units (RU), distributed units (DU) and centralized units (CU), as illustrated in Figure 1. However, where best to split the open RAN is a complex compromise between RU simplification, DU functionality, CU functionality, support of advanced co-ordinated multipoint RF capabilities, consequential limitations on transport delay budgets as well as interface bandwidth expansion.

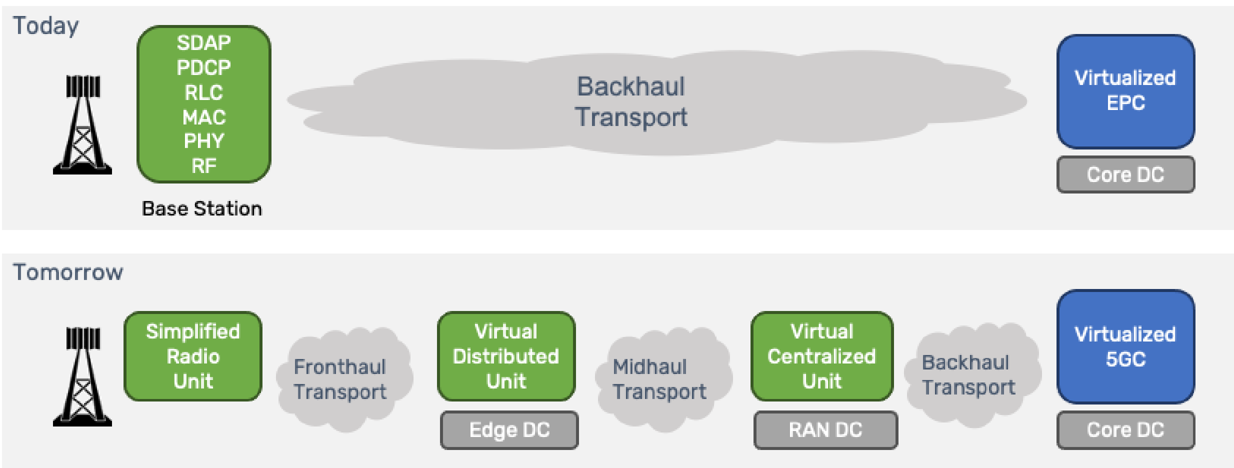

Figure 1 - The Transition to a Decomposed Radio Access Network

During the study into 5G's Radio Architecture, several alternative splits were analysed [1]. These were broadly categorized into "higher layer splits" (HLS) and "lower layer splits" (LLS). The demarcation between these two categories is the location of the scheduler, with the term LLS being applicable to deployments with lower latency transport where it is possible to realize enhanced performance through centralized scheduling, and HLS being applicable to deployment with higher transport latencies that operate with a distributed scheduler.

To help in comparing alternative options, different splits have been assigned numbers with higher numbers representing splits "lower down" in the protocol stack, meaning less functionality being deployed "below" the split in the RU. Lower layer splits occur below the medium access control (MAC) layer in the protocol stack which contains the scheduler functionality. Several lower later splits are possible, including Split 6 - between the MAC and physical layers, Split 7 - within the physical layer, and Split 8 - between the physical layer and the RF functionality. Interestingly, Split 6 is analogous to the Remote PHY interface as defined at CableLabs [2].

Before 5G's analysis into splits, the de facto approach to split the RAN was to use an interface based on the Common Public Radio Interface (CPRI). Back in 2003, the CPRI industry co-operation (www.cpri.info) had defined an interface between a Radio Equipment Control (REC) element implementing all the RAN baseband functions and a Radio Equipment (RE) element implementing the RF functions, to enable the RE to be located at the top of a cell tower and the

REC to be located at the base of the cell tower. This interface was subsequently repurposed to support relocation of the REC to a centralized location that could serve multiple cell towers via a fronthaul transport network. Using the split numerology introduced in 3GPP 38.801, with the RE implementing the RF and REC implementing the physical layer and above, the CPRI-based split is identified as a split 8 approach.

Split 8 has been characterized as requiring up to a 30-fold bandwidth expansion compared to HLS approaches [2]. Furthermore, in LTE the transport latency budget of all LLS approaches is constrained by the operation of hybrid automatic repeat request (HARQ) functionality in the MAC layer. This results in the oft-quoted delay requirement of 250 micro-seconds for one way transport delay budget between the radio and the element implementing the MAC layer's uplink HARQ functionality. Finally, with 5G's increasing focus on multi-antenna systems, the CPRI split 8 approach is hampered by the need to scale linearly with the number of RF elements.

With such extreme requirements, it is evident why the focus of DOCSIS based transport has been on HLS approaches to RAN decomposition, where there is nominal bandwidth expansion compared to conventional RAN backhaul systems and delay budgets can be measured in milli-seconds instead of micro-seconds. This paper takes an alternative view and compares the requirements for supporting two different lower layer splits, namely the network functional application platform interface (nFAPI) Split 6 as defined by the Small Cell Forum (www.smallcellforum.org) and the Split 7-2x as defined by the O-RAN Alliance (www.o-ran.org). Whereas the pre-conception is that lower layer splits are incompatible with fronthaul being transported using DOCSIS, this paper examines the LLS requirements associated with these splits and demonstrates how, given correct configuration, fronthaul deployments can be compatible with DOCSIS based transport.

2. Small Cell Splits

The Small Cell Forum (SCF) took the initial lead in defining a multivendor lower layer split, taking its FAPI platform application programming interface (API) that had been used as an informative split of functionality between small cell silicon providers and the small cell RAN protocol stack providers, and enabling this to be "networked" over an IP transport.

This "networked" FAPI, or nFAPI, enables the Physical Network Function (PNF) implementing the small cell RF and physical layer to be remotely located from the Virtual Network Function (VNF) implementing the small cell MAC layer and upper layer RAN protocols. First published by the SCF in 2016, the specification of the MAC/PHY split has since been labelled as "Split 6" by 3GPP TR38.801 that studied 5G's New Radio access technology and architectures. The nFAPI architecture is shown in Figure 2.

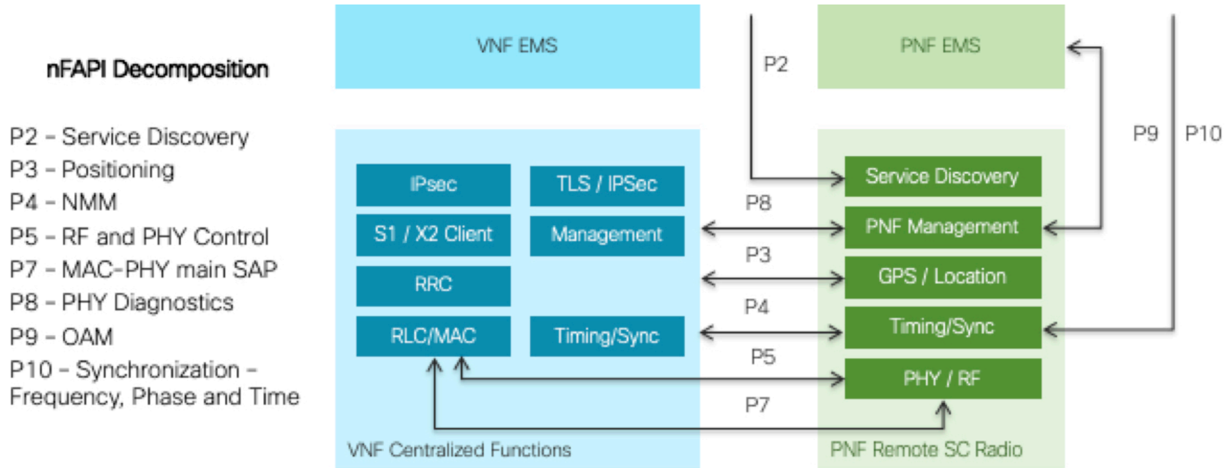

Figure 2 - Small Cell Forum's nFAPI MAC/PHY Split

The initial SCF nFAPI program delivered important capabilities that enabled small cells to be virtualized. Importantly, when comparing the transport bandwidth requirements for the fronthaul interface, nFAPI/Split 6 does not significantly expand the bandwidth required compared to more conventional small cell backhaul deployments. Moreover, just like the backhaul traffic, the nFAPI transport bandwidth varies according to served traffic, enabling statistical multiplexing to be used over the fronthaul IP network. This can be contrasted with the alternative CPRI/Split 8 that requires bandwidth expansion up to 30-fold and a constant bit rate connection, even if there is no traffic being served in a cell.

3. HARQ Latency Constraints

LTE contains a retransmit mechanism called the hybrid automatic repeat request (HARQ). DOCSIS does not have an equivalent mechanism. There is an uplink synchronous HARQ which has tight timing constraints and a downlink asynchronous HARQ that does not. The timing constraints discussed here are unique to LTE and do not apply to 5G. This is important as in the Cable industry, some operators may only consider deploying 5G on their DOCSIS network.

Whereas nFAPI/Split 6 offers significant benefits over CPRI/Split 8 in terms of bandwidth expansion, both splits are below the hybrid automatic repeat request (HARQ) functionality in the MAC layer that is responsible for constraining the transport delay budget for LTE fronthaul solutions. LTE-based Split 6, Split 7 and Split 8 all have a common delay constraint equivalent to 3 milliseconds between when uplink data is received by the RF to the time when the corresponding downlink ACK/NAK needs to be ready to be transmitted by the RF, as illustrated in Figure 3.

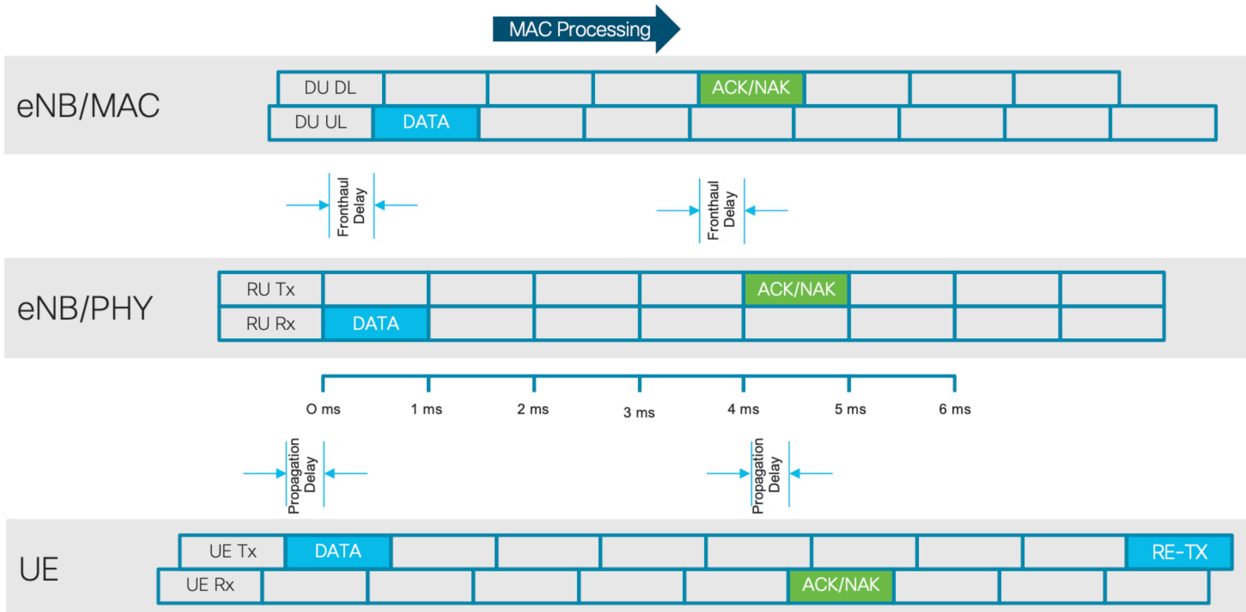

Figure 3 - HARQ Latency Constraints for LTE

These 3 milliseconds need to be allocated to HARQ processing by the MAC layer and fronthaul transport delay, with a common assumption being that 2.5 milliseconds are allocated to processing, leaving 0.5 milliseconds allocated to round trip transport, or 250 micro-seconds for one way transport delay budget between the element implementing the RF and the element implementing the MAC layer's uplink HARQ functionality.

The Small Cell Forum acknowledges such limitations when using its nFAPI split. Because the 250 micro-seconds one way transport budget severely constrains nFAPI deployments, SCF defines the use of HARQ interleaving that leverages standardized signalling to defer HARQ buffer emptying, enabling higher latency fronthaul links to be accommodated. Although HARQ interleaving buys additional transport delay budget, the operation has a severe impact on single user equipment (UE) throughput; as soon as the delay budget exceeds the constraint described above, the per UE maximum throughput is immediately decreased by 50%, with further decreases as delays in the transport network increase, as illustrated in Figure 4.

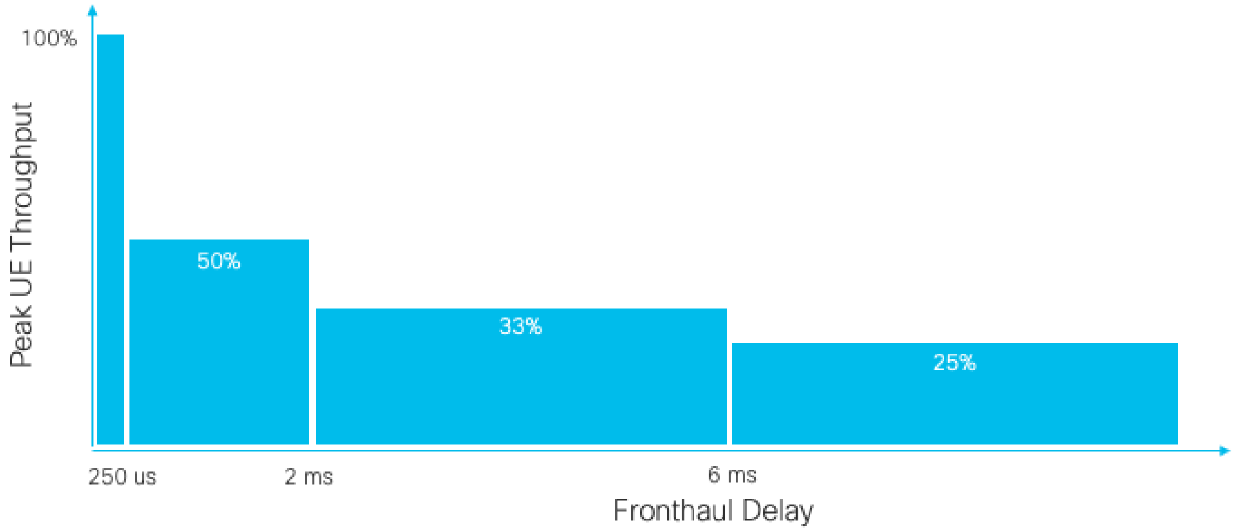

Figure 4 - Impact of HARQ interleaving on peak UE throughput

Some proponents have advocated that in certain LTE deployment scenarios, the operation of HARQ can be disabled to avoid the associated 250 micro-second delay constraint. However, analysis indicates that HARQ operation plays an important role in improving performance as signal to node ratio (SNR) falls below 8 dB [4], meaning that a significant number of cell edge users will likely be impacted if HARQ is disabled.

The restriction associated with the operation of LTE uplink HARQ is due to the definition of HARQ operation as being synchronous, whereby the identity of a HARQ process is derived from its transmission timing. This can be contrasted to the operation of the LTE downlink HARQ that defines the signalling of a HARQ process identity along with the data which means that there are no equivalent downlink timing constraints.

Importantly, 5G new radio (NR) does not implement the same synchronous uplink HARQ procedures, instead defining the use of HARQ process identities in both the downlink and the uplink. Consequently, 5G fronthaul systems do not suffer the same HARQ-based transport delay constraints as LTE. Instead, the limiting factor constraining the transport budget in 5G fronthaul systems is the operation of the windowing during the random-access procedure.

When a UE wants to establish a connection, it first recovers information on the system information broadcast type 2 (SIB2) message broadcast by the cell. SIB2 includes information about the random-access channel (RACH) configuration. The UE uses the RACH configuration to determine the time, frequency, preamble identity and repetition information to use when sending the physical random-access channel (PRACH) preamble, sometimes referred to as “message 1” (MSG1). If received correctly by the network, the base station will transmit a random-access response (RAR) message to the UE. The UE will monitor the physical downlink control channel (PDCCH) for reception of the RAR message sent by the network, sometimes referred to as “message 2” (MSG2). Importantly, the monitoring period is controlled by a parameter termed *raResponseWindow* which has a maximum value of 10 milliseconds [5], as illustrated in Figure 5.

This maximum value of 10 milliseconds needs to be partitioned between the PRACH preamble processing by the MAC layer, round-trip transport delays and over the air delays. Allocating 2.5 milliseconds to PRACH preamble processing leaves 7.5 milliseconds to be allocated between the over the air transmissions and round-trip transport delay. This effectively means that fronthaul round-trip transport delays of up to 5 milliseconds can be accommodated without impacting the currently defined RACH processing.

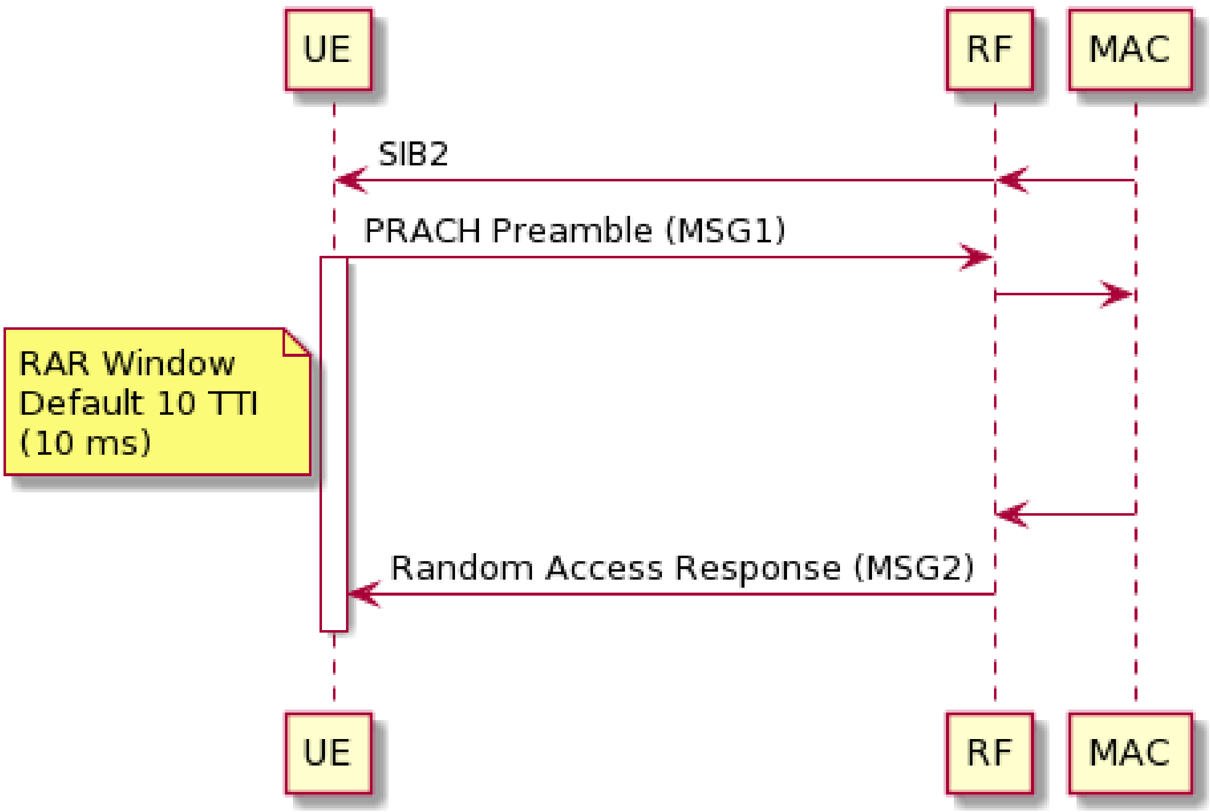

Figure 5 - Random Access Timing

Importantly, whereas the 500 microsecond round-trip transport delay requirement necessitated by LTE's synchronous uplink HARQ cannot be met by the DOCSIS system, the ~5 millisecond round-trip transport delay for 5G might be achieved with an optimized DOCSIS configuration, and thus splits 6 through 8, will work for DOCSIS, at least from the HARQ viewpoint. The Low Latency Xhaul (LLX) feature is targeted at getting the majority mobile traffic to be within 5 milliseconds round trip.

4. Split PHY Alternatives

Unlike in nFAPI/split 6 where there is a clear demarcation between MAC and PHY layers, the newer split 7 intra PHY approach can have multiple realizations. 3GPP 38.801 describes three alternative realizations:

- Split 7-1: whereby the FFT, Cyclic Prefix handling and uplink PRACH processing are distributed into the RU
- Split 7-2: In addition to those functions defined in Split 7-1, the RU additionally includes layer mapping/de-mapping and optionally the precoding functionality
- Split 7-3: In addition to those functions defined in Split 7-2, in the downlink, the RU additionally includes the modulator

These split PHY alternatives, together with the nFAPI/split 6 and the CPRI/split 8 are illustrated in Figure 6.

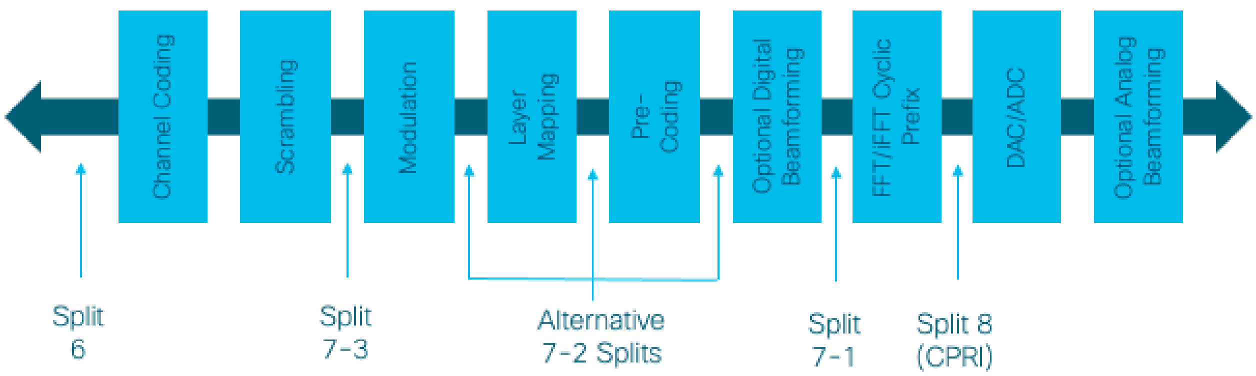

Figure 6 - Alternative Lower Layer Splits

As reported by the Small Cell Forum [2], these alternative split PHY approaches offer benefits in terms of advanced RF combining capabilities. Table 1 is taken from the Small Cell Forum's study into virtualization which highlights that "the lower down in the protocol stack the decomposition occurs, the greater the ability to benefit from the enhanced co-ordination techniques".

However, the alternative approaches to split PHY realizations risk fragmenting the industry in its effort to define a multi-vendor interoperable split PHY approach. This issue was taken up in 2016 by a group of operators and vendors in the xRAN Forum. The xRAN Forum worked on comparing alternative vendor views on split PHY realization and managed to coalesce these into a single approach. Then in 2018, the xRAN Forum announced its merger with the C-RAN Alliance to form a world-wide, carrier-led effort to drive new levels of openness in the radio access network of next-generation wireless systems named the O-RAN Alliance.

Table 1 - Comparing Advanced RF Combining Capabilities of Lower Layer Splits

Advanced RF Combining Capability	PDCP/ RLC	Split MAC	MAC/ PHY	Split PHY
Carrier Aggregation		x	x	x
Cross Carrier Scheduling		x	x	x
Higher order MIMO				x
Downlink Joint Processing- Joint Transmission			x	x
Uplink Joint Reception independent PHY decoding			x	x
Uplink Joint Reception joint equalization PHY decoding				x
Joint Processing – Dynamic point Selection		x	x	x
Co-ordinated Scheduling/Beamforming up and downlink	x	x	x	x

5. O-RAN Alliance's Lower Layer Split

Taking its lead from earlier xRAN work, the O-RAN Alliance published its "7-2x" Split PHY specification in February 2019 [6]. All Split 7 alternatives offer significant benefits over the legacy CPRI/Split 8, which includes avoiding split 8 requirements to scale fronthaul bandwidth on a per antenna basis, as well as introducing transport bandwidth requirements that vary with served traffic in the cell, compared to Split 8 which has a near constant network data rate even when there is no cell traffic. Moreover, when compared to Split 6, the O-RAN lower layer Split 7-2x supports all advanced RF combining techniques, including the higher order multiple-input, multiple-output (MIMO) capability that is viewed as a key enabling technology for 5G deployments.

However, instead of supporting individual transport channels over the nFAPI interface, Split 7-2x defines the transport of frequency domain IQ defined spatial streams or MIMO layers across the lower layer fronthaul interface. The use of frequency domain IQ symbols can lead to a significant increase in fronthaul bandwidth when compared to the original transport channels.

Figure 7 illustrates the bandwidth expansion due to split 7-2 occurring "below" the modulation function, where the original 4 bits "1110" to be transmitted are expanded to over 18 bits after 16-QAM modulation is applied, even when using the block floating point compression scheme defined by O-RAN Alliance.

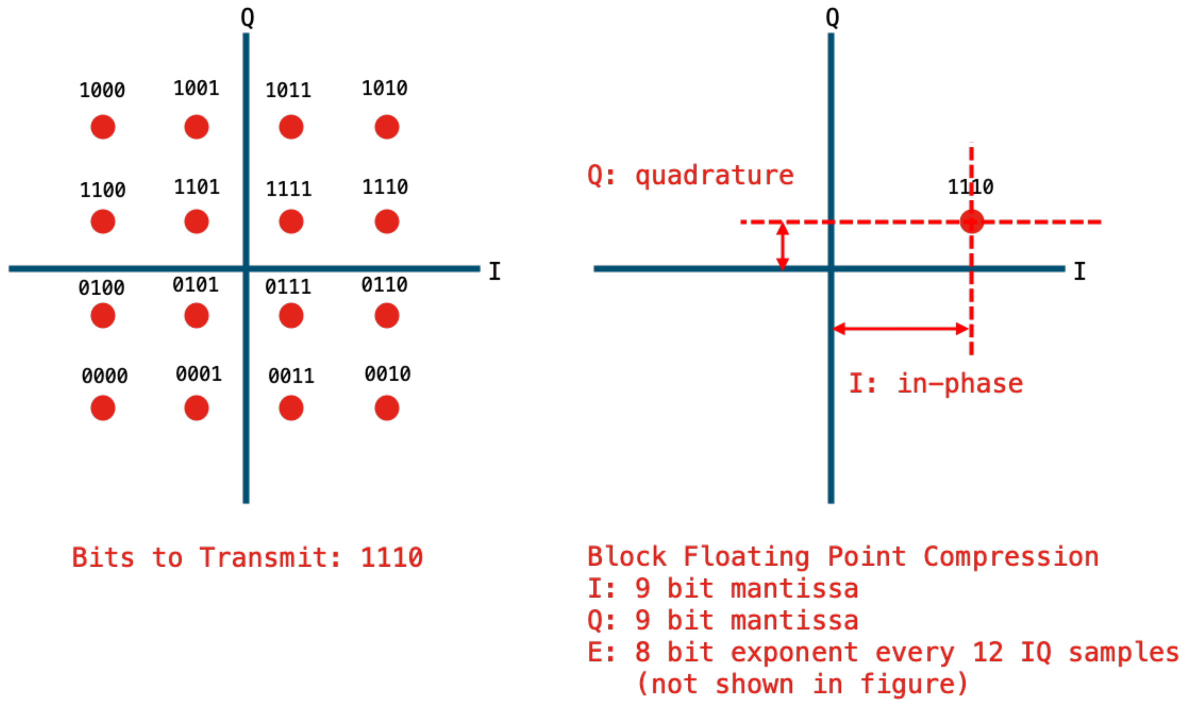

Figure 7 - Bandwidth Expansion with Block Floating Point Compressed Split 7-2x

The bandwidth expansion is a function of the modulation scheme, with higher expansion required for lower order modulation, as shown in Table 2.

Table 2 - Bandwidth Expansion for Split 7-2x with Block Floating Point Compression compared to Split 7-3

Modulation Scheme	Bits to modulate	Block floating point bits	Bandwidth expansion ratio
16 QAM	4	18.67	4.67
256 QAM	8	18.67	2.33

Such a bandwidth expansion was one of the reasons that proponents of the so called Split 7-3 advocated a split that occurred “above” the modulation/demodulation function. To address such issues, and the possible fragmentation of different Split 7 solutions, the O-RAN Alliance lower layer split includes the definition of a technique termed *modulation compression*.

The operation of modulation compression of a 16-QAM modulated waveform is illustrated in Figure 8. The conventional Split 7-2 modulated constellation diagram is shifted to enable the modulation points to lie on a grid that then allows the I and Q components to be represented in binary instead of floating-point numbers. Additional scaling information is required to be signalled across the fronthaul interface to be able to recover the original modulated constellation points in the RU, but this only needs to be sent once per data section.

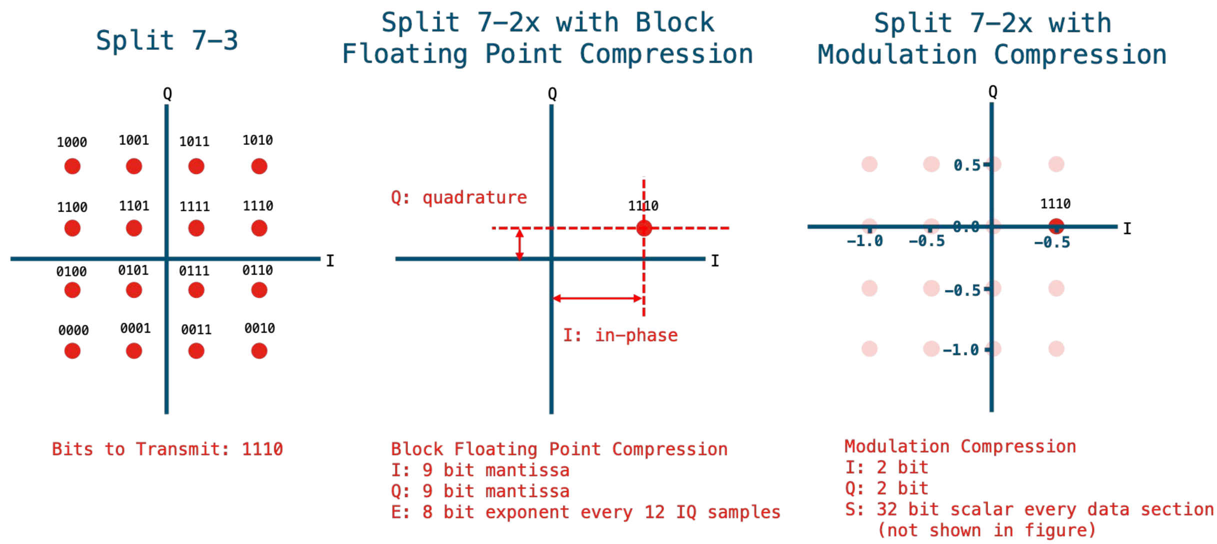

Figure 8 - User Plane Bandwidth Reduction Using Modulation Compression with Split 7-2

Because modulation compression requires the in-phase and quadrature points to be perfectly aligned with the constellation grid, it can only be used in the downlink. However, when used, it decreases the bandwidth expansion ratio of Split 7-2x, where the expansion compared to Split 7-3 is now only due to the additional scaling and constellation shift information. This information is encoded as 4 octets and sent every data section, meaning the bandwidth expansion ratio will vary according to how many Physical Resource Blocks (PRBs) are included in each data section. This value can range from a single PRB up to 255 PRBs, with Table 3 showing the corresponding Split 7-2x bandwidth expansion ratio over Split 7-3 is effectively unity when operating using large data sections.

Table 3 - Bandwidth Expansion for Split 7-2x with Modulation Compression compared to Split 7-3

Modulation Scheme	Bits to modulate	7-2x Block FP Compression BW Expansion Ratio	PRBs per data section	Modulation Compression (ModComp) bits	7-2x ModComp BW Expansion Ratio
16 QAM	4	4.67	1	6.67	1.67
			10	4.27	1.07
			255	4.01	1.00
256 QAM	8	2.33	1	10.67	1.33
			10	8.27	1.03
			255	8.01	1.00

Note, even though modulation compression is only applicable to the downlink (DL), the shift of new frequency allocations to Time Division Duplex (TDD) enables a balancing of effective fronthaul throughput between uplink (UL) and downlink. For example, in LTE, 4 of the 7 possible TDD configurations have more slots allocated to downlink traffic, compared to 2 possible configuration that have more slots allocated in the uplink. Using a typical 12-to-6 DL/UL configuration, with 256-QAM and 10 PRBs per data section, the overall balance of bitrates for modulation compression in the downlink and block floating point compression in the uplink will be (1.03 x 12) to (2.33 x 6), or 12.40:13.98, i.e., resulting in a relatively balanced link as it relates to overall bandwidth.

A more comprehensive analysis by the O-RAN Alliance has examined control and user-plane scaling requirements for Split 7-2x with modulation compression and compared the figures with those for Split 7-3. When taking into account other overheads, this analysis indicated that the difference in downlink bandwidth between Split 7-3 and Split 7-2x with Modulation Compression was estimated to be around 7%. Using such analysis, it is evident why the O-RAN Alliance chose not to define a Split 7-3, instead advocating a converged approach based on Split 7-2x that can be configured to address a variety of lower layer split deployment scenarios.

6. Comparing Split 7-2x and nFAPI

Material from the SCF clearly demonstrates that, in contrast to Split 7, their nFAPI/Split 6 approach is challenged in supporting massive MIMO functionality that is viewed as a key enabling technology for 5G deployments. However, massive MIMO is more applicable to outdoor macro-cellular coverage, where it can be used to handle high mobility and suppress cell-edge interference use cases. Hence, there may be a subset of 5G deployments where massive MIMO support is not required, such as 5G fronthaul over DOCSIS, so let's compare the other attributes.

With both O-RAN's Split 7-2x and SCF's nFAPI lower layer split occurring below the HARQ processing in the MAC layer, both are constrained by exactly the same delay requirements as it relates to LTE HARQ processing and fronthaul transport budgets. Both O-RAN's Split 7-2x and SCF's nFAPI lower layer split permit the fronthaul traffic load to match the served cell traffic, enabling statistical multiplexing of traffic to be used within the fronthaul network. Both O-RAN's Split 7-2x and SCF's nFAPI split support transport using a packet transport network between the RU and the DU.

The managed object for the SCF's PNF includes the ability for a single PNF to support multiple PNF Services. A PNF service can correspond to a cell, meaning that a PNF can be shared between multiple operators, whereby the PNF operator is responsible for provisioning the individual cells. This provides a foundation for implementing Neutral Host. More recently, the O-RAN Alliance's Fronthaul Working Group has approved a work item to enhance the O-RAN lower layer split to support a "shared O-RAN Radio Unit" that can be parented to DUs from different operators, thus facilitating multi-operator deployment.

Both SCF and O-RAN Split 7-2x solutions have been influenced by the Distributed Antenna System (DAS) architectures that are the primary solution for bringing the RAN to indoor locations. The SCF leveraged the approach to DAS management when defining its approach to shared PNF operation. In contrast, O-RAN's Split 7-2x has standardized enhanced "shared cell" functionality where multiple RUs are used in creating a single cell. This effectively uses the eCPRI based fronthaul to replicate functionality normally associated with digital DAS deployments.

Comparing fronthaul bandwidth requirements, it's evident that the 30-fold bandwidth expansion of CPRI was one of the main reasons for SCF to embark on its nFAPI specification program. However, the above analysis highlights how O-RAN has delivered important capabilities in its Split 7-2x to limit the necessary bandwidth expansion and avoid fragmentation of the lower layer split market between alternative split PHY approaches.

The final aspect when comparing these alternatives is how much the bandwidth is expanded when going from Split 6 to Split 7-2x. Figure 6 illustrates that the bandwidth expansion between Split 6 and Split 7-3 is due to the operation of channel coding. With O-RAN having already estimated that Split 7-3 offers a 7% bandwidth savings compared to Split 7-2x with Modulation Compression, we can use typical channel coding rates to estimate the bandwidth expansion between Split 6 and Split 7-2x.

Table 4 uses typical LTE coding rates for 64-QAM modulation to calculate the bandwidth expansion due to channel coding rate, where the coding rate is the ratio of the useful data transmitted in a subframe to the total amount of data transmitted. This is combined with the additional 7% expansion due to Modulation Compression to estimate the differences in required bandwidth. This table shows that the difference in bandwidth between nFAPI/Split 6 and Split 7-2x is a function of channel coding rate and can be as high as 93% for 64QAM with $^{1}/_{2}$ rate code, and as low as 16% for 64 QAM with an $^{11}/_{12}$ rate code.

Table 4 - Example LTE 64QAM Channel Coding Bandwidth Expansion

Name	Effective Code Rate	Channel Coding BW Expansion	Channel Coding Expansion plus 7%
64-QAM $^{1}/_{2}$	0.554	1.81	1.93
64-QAM $^{3}/_{5}$	0.650	1.54	1.64
64-QAM $^{3}/_{4}$	0.754	1.33	1.42
64-QAM $^{5}/_{6}$	0.852	1.17	1.26
64-QAM $^{11}/_{12}$	0.926	1.08	1.16

Whereas the above analysis indicates that the cost of implementing the Channel Coding above the RU in Split 7-2x is a nominal increase in bandwidth, the benefit to such an approach is the significant simplification of the RU by removing the need to perform channel decoding. Critically, the channel decoder requires highly complex arithmetic and can become the bottleneck in physical layer processing. Often, this results in the use of dedicated hardware accelerators that can add significant complexity and cost to the Split 6 Radio Unit. In contrast, O-RAN's split 7-2x allows the decoding functionality to be centralized, where it is expected that it can benefit from increased utilization and associated efficiencies, while simplifying the design of the O-RAN Radio Unit. A summary of these comparisons is illustrated in Table 5.

Table 5 – Summarizing Differences Between nFAPI and Split 7-2x

Characteristic	nFAPI	Split 7-2x	Comment
Advanced RF Techniques	Supports 6 out of 8 RF techniques	Supports 8 out of 8 RF techniques	Split 7-2x supports higher order MIMO
Round-trip Transport Latency for LTE	Hard limit of 0.5 milliseconds	Hard limit of 0.5 milliseconds	Identical delay constraints as both splits are below HARQ
Round-trip Transport Latency for NR	Soft limit of ~5 milliseconds	Soft limit of ~5 milliseconds	Identical delay constraint
Bandwidth Expansion compared with HLS	Limited bandwidth expansion	~16-93% bandwidth expansion for 64 QAM	Split 7-2x has lower bandwidth expansion for higher modulation rates
MIMO Layer Bandwidth Expansion	None	Bandwidth scales with MIMO layers (Cat-B)	Key delta in bandwidth is due to expansion due to MIMO layers
Statistical Multiplexing in Transport	Yes	Yes	Both splits enable statistical multiplexing
RU Complexity	Similar to composed base station	Removes requirement for channel decoder in RU	Split 7-2x enables RU simplification

7. MIMO Layer Optimization

Both LTE and 5G define the use of MIMO that use multiple transmitting and receiving antennas and exploit multi-path to enable multiple MIMO layers to be supported. In cable, an analogy would be that each port on a fully segmented fiber node is a MIMO layer. The comparison above indicates that there may be limited bandwidth expansion possible with Split 7-2x compared with Split 6 when a single MIMO Layer is being sent over the fronthaul interface. However, where the RAN is configured to support multiple MIMO layers, then this will be the primary parameter that governs the effective bandwidth expansion of Split 6 versus Split 7-2x deployments.

The separate MIMO layers can be used in various configuration, including spatial multiplexing where different layers are used to transmit separate information in order to increase the capacity of the channel, and transmit diversity where different layers are used to transmit the same information in order to enhance the quality of the received signal. Earlier analysis of LTE field trials has compared the transmission modes used in a congested multi-cell environment, contrasting spatial multiplexing, transmit diversity and massive MIMO techniques [7]. These results indicate that the most common multi-antenna technique operated in the network is transmit diversity. Significantly, the O-RAN Split 7-2x supports 3 different transmission schemes:

- Spatial Multiplexing with Cyclic Delay Diversity
- Spatial Multiplexing without Cyclic Delay Diversity
- Transmit Diversity

Specifically, with transmit diversity, the two or four transmit diversity MIMO layers can be packed into a single physical resource block sent over the Split 7-2x interface, meaning that the fronthaul bandwidth is not expanded compared with Split 6, even when transmitting multiple MIMO layers. This is illustrated below for the downlink direction, showing a single set of resource elements (shown in yellow) being used to drive multiple antenna streams (green for antenna 0, blue for antenna 1) and where the individual antenna reference symbols necessary for the operation of MIMO are time multiplexed across the fronthaul interface and then unpacked into the separate streams used to transmit over the respective antennas.

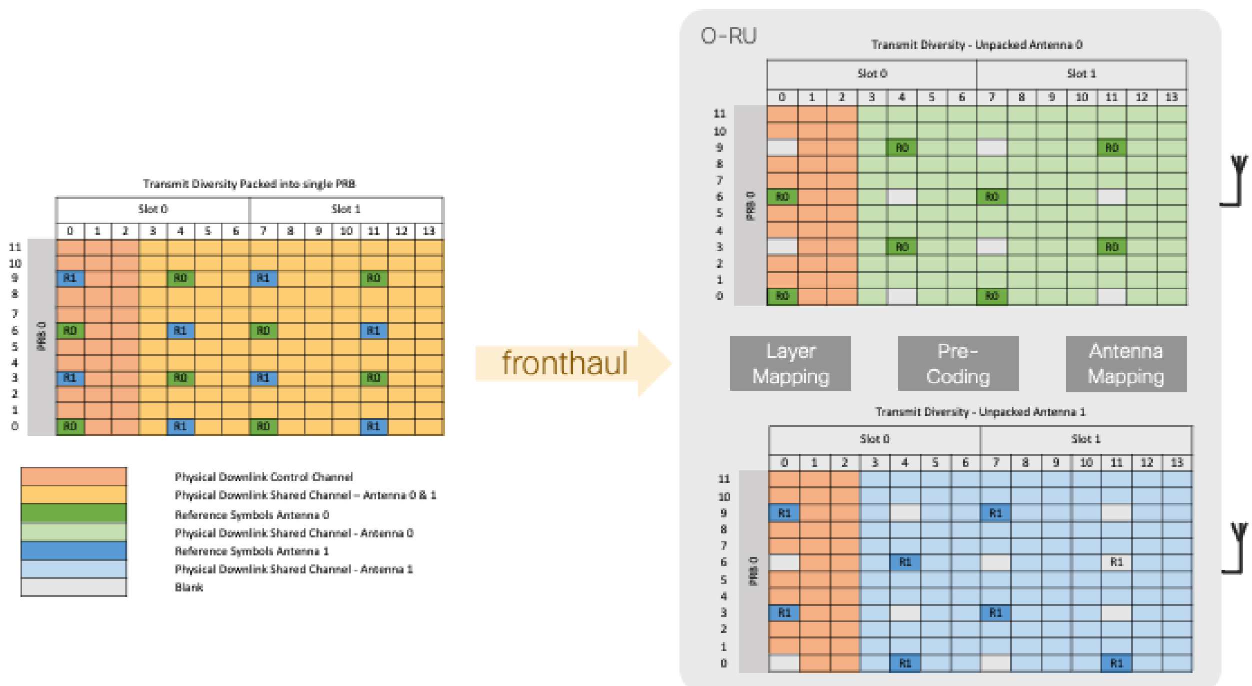

Figure 9 - Packing Multiple Transmit Diversity Layers into a Single Physical Resource Block

Hence, in scenarios where transmit diversity is the most common multi-antenna technique, the options available in O-RAN's split 7-2x avoids any additional fronthaul bandwidth expansion compared to Split 6.

8. Minimizing Transport Delays with Co-operative Scheduling

The above analysis has shown the variety of techniques embedded within the O-RAN fronthaul specification that are aimed at reducing the bandwidth requirements when transporting the multi-vendor interoperable Split 7-2x interface. The final aspect covered in this section deals with the

O-RAN Alliance specification for optimizing fronthaul deployments transported using resource allocation-based transport networks such as DOCSIS and passive optical networking (PON).

Heavily influenced by earlier CableLabs low latency mobile xhaul (LLX) technology [8][9] , O-RAN has specified a "co-operative transport interface" (CTI) targeted at minimizing transport delay requirements through the coordination of resource scheduling between the RAN elements and the transport network. Leveraging the same concepts in CableLabs' bandwidth report (BWR) concept, the mobile scheduler in O-RAN's O-DU can signal in real-time information such as fronthaul bandwidth and associated latency or QoS requirements to the transport scheduling entity in the transport node (TN). Here the reported information can be used to expedite resource allocation for the uplink RAN traffic, including adapting the transport control traffic sent to the distributed transport unit (TU) used to support the packetized xhaul traffic.

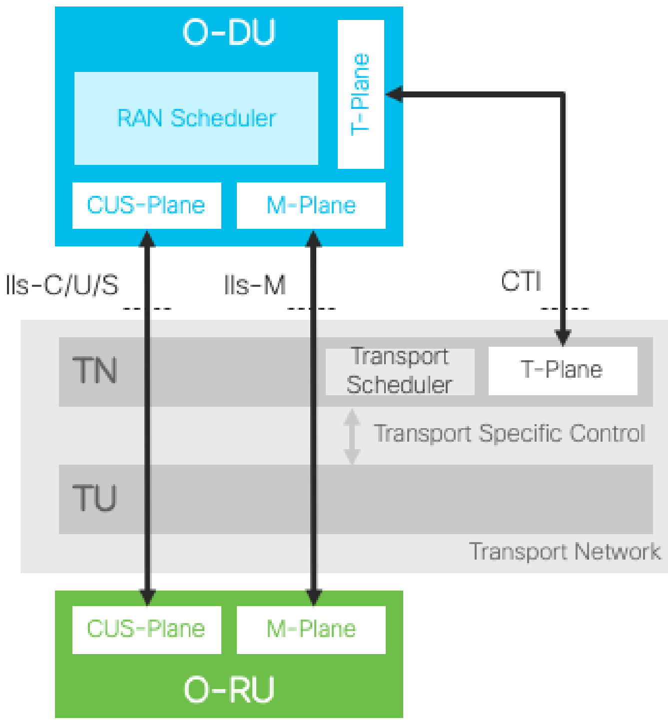

Figure 10 - O-RAN's Co-operative Transport Interface

Earlier trials of the BWR system have demonstrated the benefits of scheduler co-ordination [10]. Trials have shown that even when channel utilization is high and many users are trying to access the channel, BWR ensures a 1-2 millisecond latency with a higher DOCSIS traffic priority applied to the BWR flow. Significantly, at the 95th percentile, BWR has been demonstrated to reduce DOCSIS upstream latency by almost an order of magnitude, from 22 milliseconds to 2.5 milliseconds.

With 5G new radio (NR) avoiding the strict sub-millisecond latency constraints associated with LTE's synchronous uplink HARQ procedures, the specification of CTI by the O-RAN alliance, coupled with the above delay measurement results, indicate that DOCSIS will be able to support the transport delay requirements associated with O-RAN's 5G fronthaul interface.

9. Scaling For Fronthaul Bandwidth

Network densification is a key driver for enabling 5G. This will see conventional cell tower sites be upgraded with 5G capability, but also a raft of new sites being deployed. These new sites will likely be targeted at delivering "hotspot" capacity to meet consumers' ever-increasing demand for more mobile broadband, or targeted coverage for supporting new vertical value chains which 5G aims to support. The 5G radio market can then be viewed as split into conventional "on-tower" based and newer "off-tower" based deployments.

The on-tower radios will be attempting to deliver increased capacity across the tower's coverage area through the use of MIMO. These MIMO systems could range from 8x8 MIMO up to 64x64 massive MIMO systems. Due to their larger radius, there will also be more UEs per radio which means the average traffic rate will be higher. The 5G on-tower market will almost always use a fiber-based fronthaul transport, although some instances do have radio backhaul.

The off-tower market is the emerging small cell market. The small cell will have a smaller radius with a less dense MIMO (2x2 or 4x4) and with fewer UEs, which means the average traffic rate will be lower. These small cells may be pole/building mounted, strand mounted, or indoor mounted. This is the market for HFC connectivity using the fiber or coax side of the hybrid fiber-coax (HFC) plant. The coax side of the HFC plant would use the DOCSIS protocol. This market segment is somewhat analogous to Wi-Fi with a potentially larger radius.

In order to understand the impact of fronthaul on the off-tower market, we first define the expected bandwidth model of the RAN system that will be backhauled over DOCSIS. Then we compare the bandwidth load to the service level of a cable modem.

With so many variations in how to configure an LTE or 5G system, it is further complicated by different approaches of how to configure the fronthaul. We will focus our attention on the split 7-2x Modulation Compression profiles approach that leverage the bandwidth saving capabilities described above.

Moreover, we will re-use a TDD test profile that looks to be applicable to the majority of new spectrum being allocated to 5G, using the so called "DDDSU" frame structure configuration. In this TDD frame configuration, D represents a slot configured for downlink operation, U represents a slot configured for uplink operation, and S represents a special slot that includes a number of symbols for downlink, a number of symbols for uplink and a number of symbols for a guard period between the downlink and uplink symbols. In an example the 14 symbol special slot can be configured as 10:2:2 (D:G:U) where G is guard time. So, this is equivalent to a time multiplexing ratio of approximately 4:1 for DL:UL.

An example LTE profile would be for 2x20MHz radio that uses 2x2 MIMO, for example corresponding to a CBRS deployment that uses 4 x 10 MHz licenses. Using Modulation Compression in the downlink and block floating point with 9-bit mantissa in the uplink, our calculations are that when considering control plane and transport overheads, a peak of 330

Mbit/s of throughput will be required on the downlink and a peak of 320 Mbit/s of throughput will be required on the uplink. Note, because of the unequal fronthaul compression techniques applied to the UL and DL, the 4:1 time multiplex results in a roughly symmetrical fronthaul bandwidth being transported over Ethernet.

An example 5G NR profile using the same 40 MHz of spectrum would be for a 40 MHz radio that uses 4x4 MIMO. Using Modulation Compression in the downlink and block floating point with 9-bit mantissa in the uplink, our calculations are that when considering control plane and transport overheads, a peak of 950 Mbit/s of throughput will be required on the downlink and a peak of 690 Mbit/s of throughput will be required on the uplink.

Whereas these rates represent the peak uplink and downlink speeds, it is recognized that real-world deployments operate with a non-uniform spatial distribution of traffic such that not all radios will be simultaneously operating at their peak capacity. The SCF analyzed such a phenomenon as part of their nFAPI virtualization deliverables [12]. While the exact peak-to-mean spatial distribution across a RAN will be a function of the deployment, e.g., including the use case being address (e.g., residential/enterprise/urban), the SCF report that a value of a peak-to-mean ratio of 3.5-to-4.0 across a 200 radio node "off-tower" network can be used to help dimension the virtualized RAN.

The bandwidth of the DOCSIS plant is well documented in [13] and results are show in Figure 11 with 96 channels (6 MHz per channel) of MPEG video, and in Figure 12 with no MPEG video. Figure 11 represents a typical case today on the HFC plant while Figure 12 represents a point in 3 to 5 years when MPEG video has been retired from the HFC plant and video services are all video over IP over DOCSIS.

The bandwidth numbers in the table represent the peak bandwidth of the DOCSIS spectrum. Each cable modem (CM) will be provisioned with some value less than this. The Distributed Access Architecture (DAA) nodes are described by the number of unique DOCSIS ports in the downstream (DS) and upstream (US) direction. A 2x4 DAA node has two unique DOCSIS DS (equivalent to a downlink) and four unique DOCSIS US (equivalent to an uplink) ports.

2 x 4 Node Capacity	User	DOCSIS 3.1 with video					DOCSIS 4.0 ESD with video					
Scenario	Calc	1	2	3	4	5	6	7	8	9	10	11
DS End MHz	1002	1002	1218	1002	1218	1218	1794	1794	1794	1794	1794	1794
DS Start MHz	54	54	54	108	108	258	108	258	372	492	606	834
US Rtn Path MHz	42	42	42	85	85	204	85	204	300	396	492	684
VOD/SDV MPEG-TS	0	32	32	32	32	32	32	32	32	32	32	32
Linear Video MPEG-TS	0	64	64	64	64	64	64	64	64	64	64	64
DOCSIS DS port Gbps	8.7	2.9	5.1	2.6	4.8	3.3	10.5	9.0	7.9	6.7	5.6	3.3
DOCSIS US port Gbps	0.10	0.10	0.10	0.47	0.47	1.48	0.47	1.48	2.11	2.93	3.75	5.39
Ethernet DS Gbps	17.3	10.7	15.0	10.1	14.4	11.4	25.8	22.8	20.6	18.2	15.9	11.4
Ethernet US Gbps	0.4	0.4	0.4	1.9	1.9	5.9	1.9	5.9	8.4	11.7	15	22

2	DS ports per Node	4096	OFDM Modulation	24	DS MHz skipped below 108 MHz	32	VOD/SDV MPEG-TS
4	US ports per Node	2048	OFDMA Modulation	YES	Video in FDX Transition Band	64	Linear Video MPEG-TS
24	SC-QAM 6 MHz ch	256	SC-QAM Modulation	1794	D4.0 Stop Frequency (MHz)	ESD	D4.0 FDX or ESD
4	ATDMA 6.4 MHz ch	64	ATDMA Modulation	16.4	US Start Freq (MHz)	120	FDX Trans Band (MHz)

Figure 11 - DOCSIS Bandwith for a 2x4 DAA Node with Video

2 x 4 Node Capacity	User	DOCSIS 3.1 with no video					DOCSIS 4.0 ESD with no video					
Scenario	**Calc**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
DS End MHz	1002	1002	1218	1002	1218	1218	1794	1794	1794	1794	1794	1794
DS Start MHz	54	474	54	108	108	258	108	258	372	492	606	834
US Rtn Path MHz	42	42	42	85	85	204	85	204	300	396	492	684
VOD/SDV MPEG-TS	0	0	0	0	0	0	0	0	0	0	0	0
Linear Video MPEG-TS	0	0	0	0	0	0	0	0	0	0	0	0
DOCSIS DS port Gbps	8.7	4.7	10.8	8.4	10.5	9.0	16.2	14.8	13.6	12.4	11.3	9.0
DOCSIS US port Gbps	0.10	0.10	0.10	0.47	0.47	1.48	0.47	1.48	2.11	2.93	3.75	5.39
Ethernet DS Gbps	17.3	9.4	21.6	16.7	21.0	18.0	32.5	29.5	27.2	24.8	22.6	18.0
Ethernet US Gbps	0.4	0.4	0.4	1.9	1.9	5.9	1.9	5.9	8.4	11.7	15	22

2	DS ports per Node	4096	OFDM Modulation	24	DS MHz skipped below 108 MHz	0	VOD/SDV MPEG-TS		
4	US ports per Node	2048	OFDMA Modulation	YES	Video in FDX Transition Band	0	Linear Video MPEG-TS		
24	SC-QAM 6 MHz ch	256	SC-QAM Modulation	1794	D4.0 Stop Frequency (MHz)	ESD	D4.0 FDX or ESD		
4	ATDMA 6.4 MHz ch	64	ATDMA Modulation	16.4	US Start Freq (MHz)	120	FDX Trans Band (MHz)		

Figure 12 - DOCSIS Bandwidth for a 2x4 DAA Node

The DOCSIS plant, as are all networks, is built with over-subscription in mind. In Figure 11, scenario 1, the DOCSIS plant supports 3 Gbps DS and 100 Mbps US. This is a typical deployment in 2021. As a typical example, this bandwidth may support 200 CMs with highest provisioned rate of 1 Gbps DS and 20 Mbps US for about 10% to 20% of the CMs, while the majority of the CMs are 200 Mbps x 10 Mbps and 100 Mbps x 5 Mbps. This defines an over-subscription case that recognizes that not all CMs will transmit or receive at the same time.

As discussed before, the small cell will not use all four of its MIMO channels simultaneously all the time. Nor will it use its full spectral capacity all the time. In fact, it depends upon the reach of the small cell and how many subscribers it connects to. Since the CM is part of a large service group (SG) of say 200 homes, the reach of the small cell and any other small cells on that node, will have the same geographical footprint. An analysis of the number of small cells per fiber node and small cells per macro-cell can be found in [14].

So, at a first pass, if a DOCSIS SG shares the same region as a set of small cells, and hence the same customers and same traffic load, then their traffic patterns will be similar. In this scenario, think of Wi-Fi connected laptops to a DOCSIS network, mixed with cell phones that are either Wi-Fi to DOCSIS connected or mobile connected.

At a second pass, the small cell fronthaul downstream bandwidth (950 Mbps) matches the max CM bandwidth (1 Gbps) and thus fits. The small cell fronthaul upstream (650 Mbps) does not match the upstream bandwidth of scenario 1 (100 Mbps). Instead, the DOCSIS upstream spectrum will have to increase from 42 MHz return (100 Mbps) to 204 MHz return (1.48 Gbps). This assumes a common spectrum starting point of 16.4 MHz.

The tables show that the DOCSIS DS bandwidth can be increased dramatically to about 10 Gbps with DOCSIS 3.1 by removing the legacy MPEG video. DOCSIS 4.0 with extended spectrum DOCSIS (ESD) can take the DS limit further as well as increase the upstream bandwidth.

Another issue that arises in support small cells, especially if they are in the home, is the support of IEEE 1588 timing. The DOCSIS Time Protocol (DTP) provides this service and is defined in [15] with further support in [16][17][18][19][20].

10. Conclusion

Fronthaul is perceived as being synonymous with fiber-based transport systems; extreme bandwidth requirements and sub-millisecond latency requirements may cause many to reach the conclusion that DOCSIS is only suitable for transporting higher-layer splits. However, in this paper we have described the enhanced capabilities in O-RAN's Split 7-2x lower-layer split architecture that are targeted at minimizing the impact on transport networking requirements.

Standardized techniques have been described that enable the bandwidth expansion of a single fronthaul stream to be reduced to low percentages when compared to alternative higher-layer split alternatives. Moreover, spatial stream optimization techniques have been described that enable a single transport stream to drive certain MIMO antenna configurations. The hard delay requirements for LTE fronthaul are deprecated in favour of soft delay requirements for 5G New Radio, with delay constraints now measured in milliseconds instead of microseconds.

Taking the lead from earlier CableLabs BWR concepts, O-RAN has fully specified a co-operative transport interface to link the RAN and transport schedulers, an approach that has already demonstrated low millisecond transport latencies and, at the 95th percentile, a reduction in DOCSIS upstream latency of almost an order of magnitude, from 22 milliseconds to 2.5 milliseconds.

Finally, we use example LTE and 5G profiles to set the parameters used in fronthaul bandwidth calculation and compare them with existing and evolving DOCSIS bandwidths. It was clear that the DOCSIS downstream had ample bandwidth, but the HFC plant needs to be upgraded to 204 MHz to allow for small cell fronthaul.

Abbreviations

ACK	acknowledgement
ADC	analog to digital conversion
BWR	bandwidth report
CB	coordinated beamforming
CM	cable modem
CMTS	cable modem termination system
CPRI	Common Public Radio Interface
CS	coordinated scheduling
CTI	co-operative transport interface
DAC	digital to analog conversion
DAS	distributed antenna system
DL	downlink
DOCSIS	data over cable service interface specifications
DPS	dynamic point selection
DS	downstream
DTP	DOCSIS Time Protocol
FAPI	functional application platform interface
FFT	fast Fourier transform
HARQ	hybrid automatic repeat request
HLS	higher layer split
iFFT	inverse fast Fourier transform
IQ	in-phase and quadrature
JR	joint reception
JT	joint transmission
LLS	lower layer split
LTE	long term evolution
MAC	medium access control
MIMO	multiple-input multiple-output
MPEG	Moving Picture Experts Group
NAK	negative acknowledgement
nFAPI	networked functional application platform interface
NMM	network monitor mode
NR	new radio
OAM	operations and maintenance
PDCCH	physical downlink control channel
PDCP	packet data convergence protocol
PNF	physical network function
PON	passive optical networking
PRACH	physical random-access channel
PRB	physical resource block
QAM	quadrature amplitude modulation
RACH	random-access channel
RAN	radio access network
RAR	random-access response
RE	radio equipment

REC	radio equipment control
RF	radio frequency
RLC	radio link control
RRC	radio resource control
RU	radio unit
SCF	Small Cell Forum
SG	service group
TDD	time division duplex
UE	user equipment
UL	uplink
US	upstream
VNF	virtual network function

Bibliography & References

[1] *Study on new radio access technology: Radio access architecture and interfaces*, 3GPP 38.801, https://www.3gpp.org/DynaReport/38801.htm

[2] *Virtualization for small cells: Overview*, Small Cable Television Laboratories, Inc., "Remote PHY Specifications for DAA", CM-SP-R-PHY, CableLabs. [link]

[3] Cell Forum, SCF106, https://scf.io/en/documents/106__Virtualization_for_small_cells_Overview.php

[4] *LTE physical layer: Performance analysis and evaluation*, Applied Computing and Informatics Volume 15, Issue 1, January 2019, Pages 34-44, https://www.sciencedirect.com/science/article/pii/S2210832717301990

[5] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification*, 3GPP 36.321, https://www.3gpp.org/DynaReport/36321.htm

[6] *O-RAN Fronthaul Control, User and Synchronization Plane Specification*, O-RAN Alliance, https://www.o-ran.org/specifications

[7] *Performance Evaluation of a Live Multi-Site LTE Network*, IEEE Access, Volume 6, 2018, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8453797

[8] *Low Latency Mobile Xhaul over DOCSIS® Technology*, Data Over Cable Service Interface Specification Mobile Applications, CM-SP-LLX-I01-190628 [link]

[9] John T. Chapman, Jennifer Andreoli-Fang, *Low Latency Techniques for Mobile Backhaul over DOCSIS*, SCTE Cable-Tec Expo Fall Technical Forum, Denver, October, 2017. [link]

[10] Jennifer Andreoli-Fang, John T Chapman, Tong Liu, Damian Poltz, *Blueprint for Mobile Xhaul over DOCSIS*, SCTE Cable-Tec Expo Fall Technical Forum, Sep, 2019. [link]

[11] *O-RAN Fronthaul Interoperability Test Specification (IOT)*, O-RAN Alliance, https://www.o-ran.org/specifications

[12] *Network Aspects of Virtualized Small*, Small Cell Forum, SCF-161, https://scf.io

[13] John T. Chapman, "The Path to 100 Gbps DAA Nodes", *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, Oct, 2021. [link]

[14] John T. Chapman, "Small Cell Traffic Engineering," *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, Oct, 2020. [link]

[15] Cable Television Laboratories, Inc., "Synchronization Techniques for DOCSIS Technology Specification," CM-SP-SYNC, CableLabs. [link]

[16] Roy Sun, John T. Chapman, et. al., "DOCSIS Time Protocol Proof of Concept", *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, Oct, 2021. [link]

[17] Roy Sun, John T. Chapman, Rahil Gandotra, "Designing a Cloud-based DOCSIS Time Protocol Calibration Database", *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, Oct, 2021. [link]

[18] Elias Chavarria Reyes, John T. Chapman, "How the DOCSIS Time Protocol makes the SYNC Specification Tick," *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, Oct, 2020. [link]

[19] Jennifer Andreoli-Fang, John T. Chapman, "Mobile Backhaul Synchronization Architecture," *SCTE Cable-Tec Expo Fall Technical Forum*, Denver, October, 2017. [link]

[20] John T. Chapman, Rakesh Chopra, Laurent Montini., "The DOCSIS Timing Protocol (DTP), Generating precision timing services from a DOCSIS system," *INTX/SCTE Spring Technical Forum*, 2011. [link]