Implications of 5G low-latency requirements on Hybrid Fiber-Coaxial Networks

A Technical Paper prepared for SCTE•ISBE by

Sanjay Dhawan
VP Strategy – Network Products and Solutions
Ericsson Inc.
6300 Legacy Dr., Plano, TX-75034
214-310-8779
sanjay.dhawan@ericsson.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

Introduction

With the finalization of initial 5G standards in 3GPP during Q1 2018, 5G deployments are starting to gain momentum globally, with a few initial commercial launches during 2018 For Cable MSO's, 5G adds new revenue opportunities, in terms of extending Mobile Broadband service to own users, and Fixed Wireless Access where there are challenges with DOCSIS/Fiber deployment, as well as IoT use cases.

One of the key attributes of 5G is the significant reduction in latency. Unlike traditional 4G LTE systems, latency requirements in 5G vary with use cases. As an example, for a traditional smartphone web browsing service, 15-20 ms round trip times may be acceptable. However, for a use case such as autonomous driving, round trip latency requirements need to be under 10 ms. Other use cases that require sub-10 ms round trip latency include industrial robotics control, drone control, web gaming, and connected, collaborative multi-site live concerts (band members playing a song across multiple locations). Typically, low latency demands tend to be localized with communication over a short distance. In order to fulfil the low latency requirements for such use cases, new architectures need to be implemented in the network. These include implementing control functionality and local switching at the edge, which in a DOCSIS network can even be a hub site. Micro servers that can support virtual applications will need to be deployed at hub sites or cell sites and in close proximity to users. In addition, while network slicing can support multiple use cases from each radio site and can fulfil use case specific routing, bandwidth and latency requirements will need to be deployed across the networks. Present DOCSIS networks are typically designed for a median latency requirement of ~10-15 ms, which can continue to work well for traditional Mobile Broadband use cases. Also, where network slicing with Edge Servers are deployed, the current cable infrastructure may be able address the 5G requirements.

5G also introduces a Virtual RAN architecture, where Layer 3 (higher layer) RAN functionality is centralized in the cloud. The one-way latency objective between the 5G radio site and the VRAN node is typically 5 ms. To fulfill such an objective, it becomes important to maximize fiber and optical switching in the access transport network. Layer 3 ethernet switching, which can add significant delays, can be deployed between the VRAN and the Core.

The 5G scheduler is hungry, which means that it will try to get the data it receives as soon as possible to the target user. For mmWave, the scheduler has a transmit time interval (window) of ~250 micro seconds which is extremely time sensitive. The faster the data can be transferred from Core to the radio, the faster it can be forwarded to the users.

Eliminating latency bottlenecks in the transport network will be key towards maximizing the overall throughput experience of 5G networks.

5G Requirements, Architecture and Use Cases

5G is gaining momentum with extensive interest from MSO's and MNO's. In fact, the race to be 1st to the market has already begun. 5G provides an evolution from current 4G smartphone services, while at the same time adding new revenue streams for service providers. Attributes such as Gigabit throughput experience enables Fixed Wireless in Urban areas, while ultra-low latency enables autonomous automotive control and remote robotic control for manufacturing automation.

5G will enable an enhanced user experience for industrial use cases. For the above manufacturing example, 5G would enable remote control of robots with round trip latency of 10 ms. For automotive and drones, autonomous control would be achieved via ultra-low latency complemented by distributed computing. Similarly, for energy and utilities, 5G would enable real time control and automation of grids.

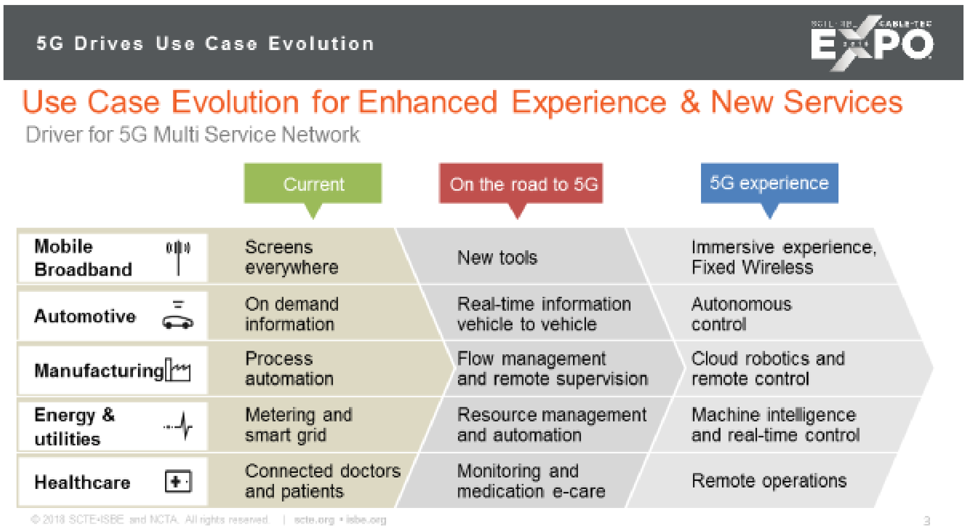


Figure 1 - Use Case Evolution with 5G

Given the initial momentum for 5G, Ericsson expects 48% of smartphones in North America to be 5G capable by 2023. Furthermore, as indicated in the Ericsson Mobility Report (2018), this is complemented by growth in IoT subscriptions from 100M today to 260M in the same time. From a user behavior standpoint, driven primarily by video, smartphone traffic is expected to grow 7 times from 2017 to 2023, to 49 GB/month.

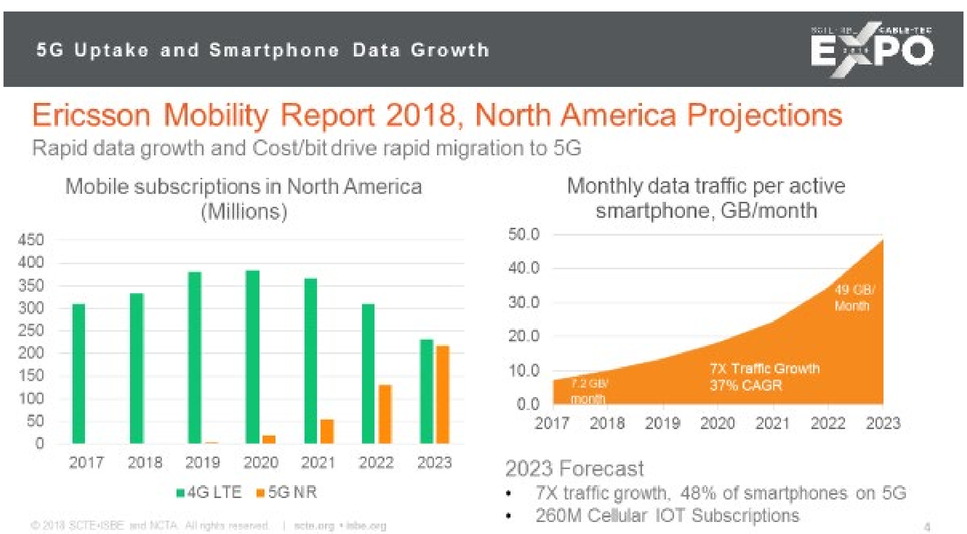


Figure 2 - 5G Subscription Uptake and Traffic Growth

5G enables performance boosts in multiple dimensions, which is key to supporting a diverse set of use cases. While the focus on initial use cases was peak Gigabit throughput, latency, reliability and positioning accuracy are emerging to be equally important along with new use cases. Several critical IoT use cases require sub 10 ms latency, 99.9999% reliability and even greater positioning accuracy than traditional GPS. On the other hand, the primary requirement for Fixed Wireless Access is a throughput experience that enables multiple 4K TVs in a home.

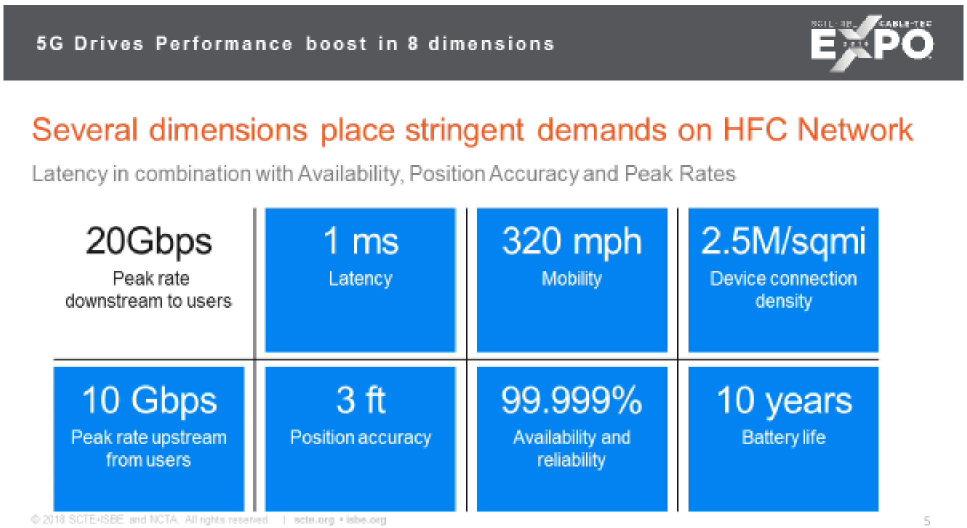


Figure 3 - 5G Drives Performance Boost in 8 Dimensions

Support for new IoT use cases complemented by exponential increase in traffic is expected to drive new ways of building networks. These networks are expected to be service slice aware, with ultra-dense small cell grids enabling Gigabit throughput along with ms latency, while macro networks provide ubiquitous coverage. Microcells leveraging mmWave (24, 28, 37 and 39 GHz) spectrum are ideal for such deployments, as mmWave provides 100 MHz – 1 GHz of spectrum per operator, and their low propagation characteristics enable an ultra-dense grid, where required.

Each class of use cases has different and distinct requirements on 5G from the perspective of coverage, bandwidth and latency. As an example, to support IoT use cases such as utility meters, ubiquitous coverage is required that includes even rural areas. Alternately, to support autonomous driving, a network that covers a wide area is required, while optimizing latency. To enable robotic manufacturing, localized indoor 5G optimized for ultra-low latency is required. Optimizing the combination of coverage with bandwidth, latency and reliability, is important for a well designed 5G network.

As can be seen from the figure below, a 5G network would require a combination of low, mid and high bands. The low bands would be ideal for providing wide coverage, while the high bands would be ideal for ultra-high capacity and ultra-low latency and require a dense grid. Similarly, mid bands are well suited for high capacity with moderate coverage. Tightly coupling 5G network elements serving low, mid and high bands would maximize the 5G experience for varying use cases.

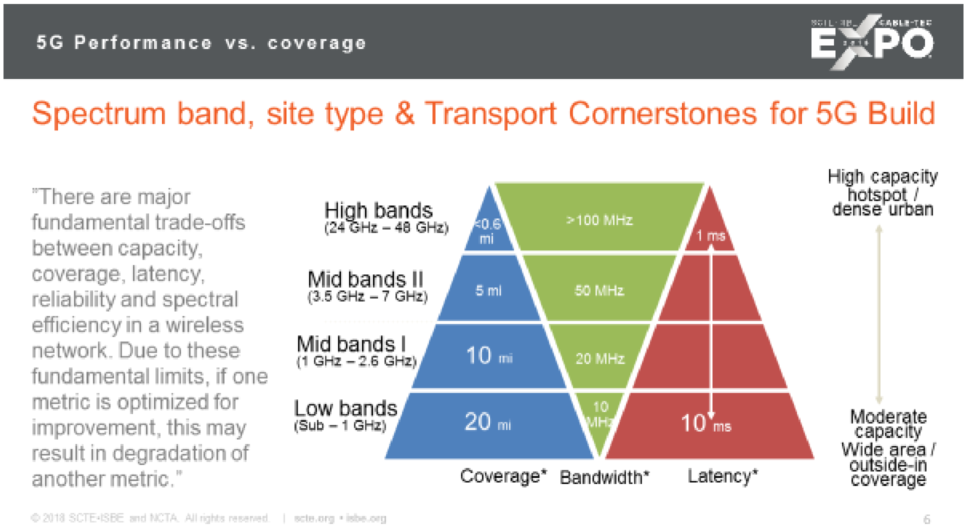


Figure 4 - Optimizing 5G Deployments for Coverage, Throughput and Latency

Deploying dense small cell grids are a strength for Cable MSO's, which can be leveraged for 5G, such that 5G small cells can be considered for deployment on an HFC/DOCSIS network with stand / pole sites, which include DOCSIS transport and power. As presented in the figure below, 5G macro coverage is achieved by deploying macro cells on existing cellular towers and complementing them with strand mount 5G pico base stations. The coordination between the macro and micro is achieved through a common Virtual RAN (Layer 3 functionality) and Core network. Such as architecture enables traffic steering, interference management and performance optimization across layers.

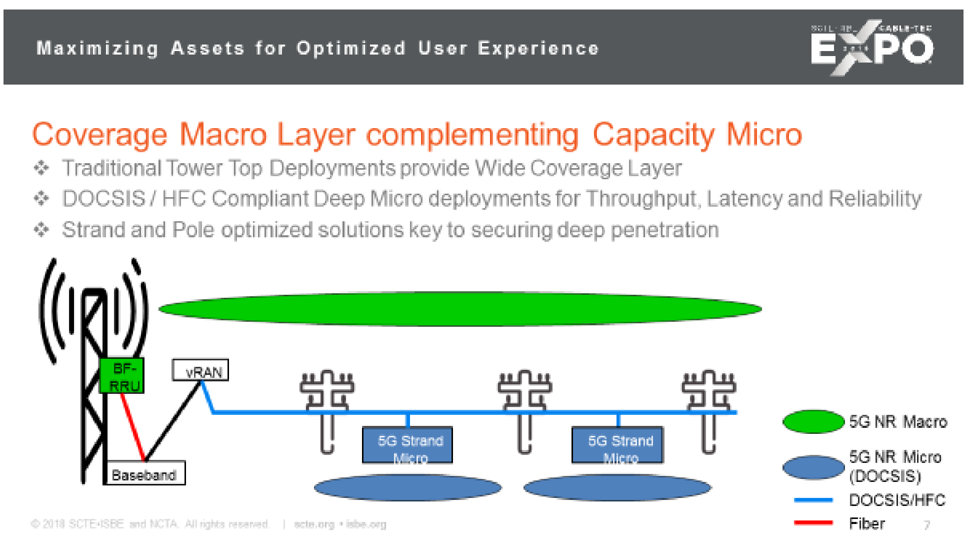


Figure 5 - Maximizing HFC/DOCSIS Assets for 5G Deployment

The radio KPIs place stringent requirements on the HFC transport network. As an example, to achieve a peak payload of 10 Gbps, a 360-degree coverage site would require a backhaul of 10-20 Gbps. Similarly, to achieve a round trip latency of <10 ms, an optical transport network is required with minimal or no L3 routing. The architecture considerations and associated transport requirements to achieve these radio performance KPIs are presented in the following sections.

As indicated above, for several 5G use cases, the most important attribute is sub 5-10 ms round trip latency, which in turn places <1ms latency requirements on the 5G Radio Access Network.

To enable low latency, several techniques are being implemented in 5G. To begin with, in mmWave, the scheduler has a time slot of 62.5-250 micro second. Within each slot the scheduler can serve 1 or more users, with Multi-User MIMO. Furthermore, techniques such as instant uplink access give ultra-low latency users instant access to the network for short data bursts, thereby keeping one-way latency to ~0.5-1.0 ms. Additional techniques such as mini-slot further reduce the transmit requirement to a subset of 1 timeslot. The scheduler is implemented as hungry, such that it will try to get the data out to users as quickly as possible, by maximizing the most important resource, i.e., spectrum, while managing users across excellent and poor radio conditions. As an example, if 100 MHz of spectrum is available, and there is only 1 user, it will be fully used for 1 or more slots, to get the buffered data out to the user as quickly as possible. If there are multiple users, the data push to users is optimized based on several factors, including Service QoS requirements, RF conditions, amount of data, etc. As radio conditions change rapidly, the scheduler needs to adapt. The adaptation of scheduling takes place on a timeslot basis.

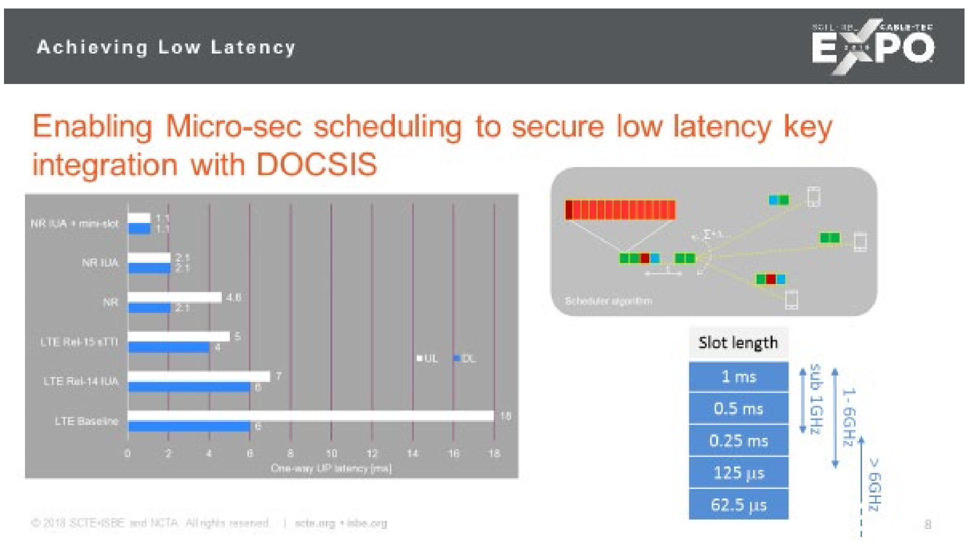


Figure 6 - Achieving Low Latency with 5G

An important architectural consideration in meeting diverse performance requirements is Network Slicing. Network slicing is akin to a VPN in IP networks, such that each VPN has its own bandwidth and QoS criteria. In 5G, where we can have hundreds of factories, each requiring dedicated bandwidth along with different QoS for different classes of devices (e.g., Robots, employees, etc.) in each factory, Network slicing is an optimal way to achieve such a requirement, without building dedicated networks for each factory. Essentially, each Network slice is a logical network serving a defined business purpose or customer, consisting of all required network resources, including Radio Access, Transport, Core and Cloud, configured together. It is created, changed and removed by management functions.

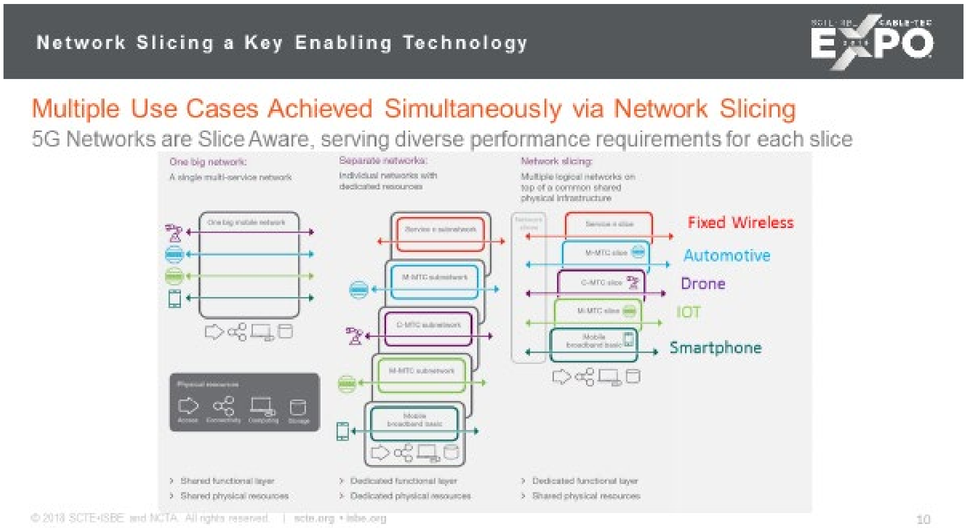


Figure 7 - Network Slicing a Key Enabling Technology

The ability to flexibly create slices places high demands on transport and network element capabilities. In principal, to be able to scale slices, it should be possible to create instances of required network elements for each slice dynamically. Furthermore, depending on the performance in fulfilling use case requirements, it should be possible to distribute functionality across the network. In the case of robot manufacturing, local switching would be required in the factory itself, implying the possibility to locate certain access and core nodes in the factory, without a major investment. This requires support of Cloud Native Virtualization of the network elements. Similarly, from a transport perspective, SDN becomes a key element. As slices start to scale, where 10's or even 100's of slices are defined in the network, it becomes important for the network to support network intelligence and automation in orchestration. A Functional View of a 5G Network Architecture that enables automatic slice management with distributed and virtualized capability is presented below.

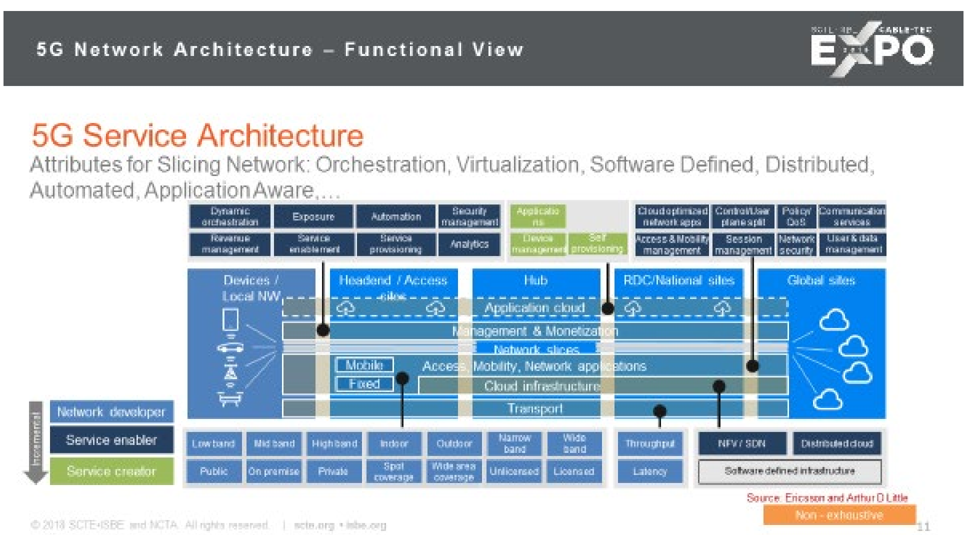


Figure 8 - 5G Network Architecture Functional View

A further zoom into an example of a Distributed Network Architecture with Virtual Network Functions and Application Servers in a Cable / HFC environment supporting three families of use cases (Critical IoT, enhanced Mobile Broadband and Massive IoT) is presented below.

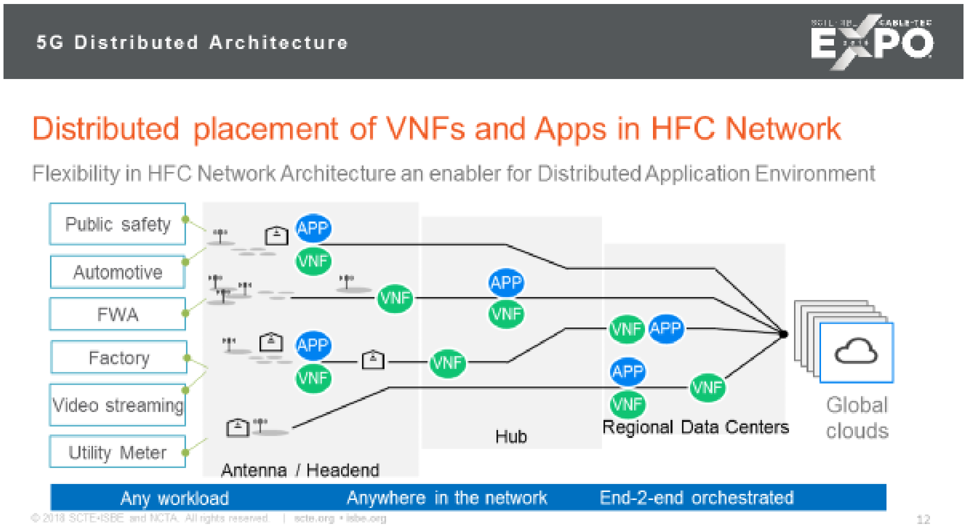


Figure 9 - 5G Distributed Architecture

Next, we consider the requirements of the distributed architecture on the HFC transport network. As shown in the figure below, for the case of critical IoT (also called Machine Type Communication, or MTC), where a radio processing function at a hub site serves a cluster of radio sites, the one-way latency between the radio site and the hub site would need to be 30 micro seconds and the bandwidth requirement would be 10-25 Gb/sec per radio. With eCPRI, ethernet support would be feasible for such an interface.

In the case of fixed or mobile broadband, each site would have the radio and the processing function collocated at the site, with the Virtual Controller at the Headend. In this scenario, the latency requirements would be less stringent and on the order of 75 micro seconds. As control signaling and payload is sent back to a server, the bandwidth requirement on the link is directly related to the payload and can be 10 Gbps or less. In such an architecture, DOCSIS / HFC network can serve as the access transport with a lesser degree of impact.

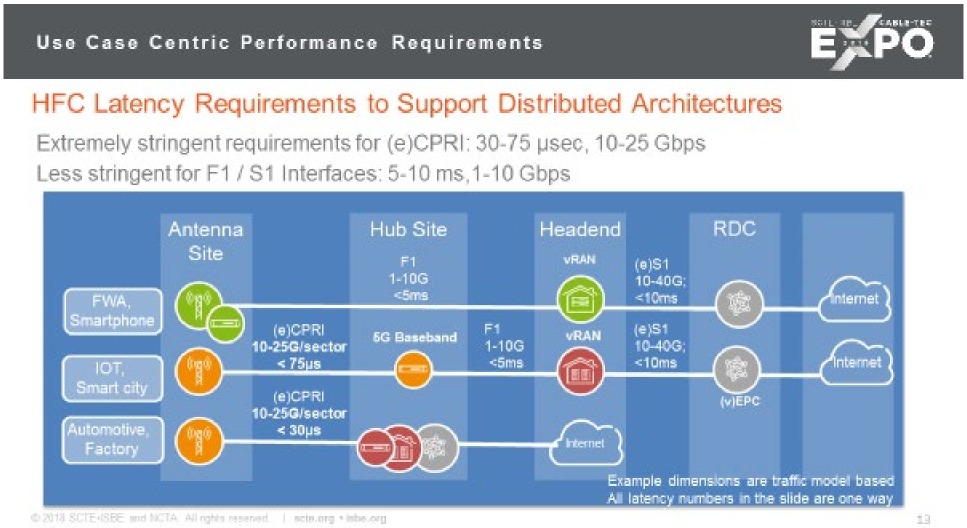


Figure 10 - HFC Throughput and Latency Requirements

In summary, an optimized 5G deployment requires multi-dimensional coordination with a DOCSIS/HFC network. To begin with extending site access, that presently supports hundreds of thousands of WiFi strand/pole mount radios, to 5G is a key factor. These sites also require extending the HFC transport and DOCSIS power to 5G pico base stations.

For several use cases, including Video Streaming to residential customers and autonomous automotive control, servers at Headend and Hub Sites capable of hosting VNFs and applications are expected to be deployed in scale.

Finally, an intelligent, fully orchestrated, Software Defined, Secure E2E network covering 5G RAN, Core and HFC are key to a fully automated and seamless service experience for the users.

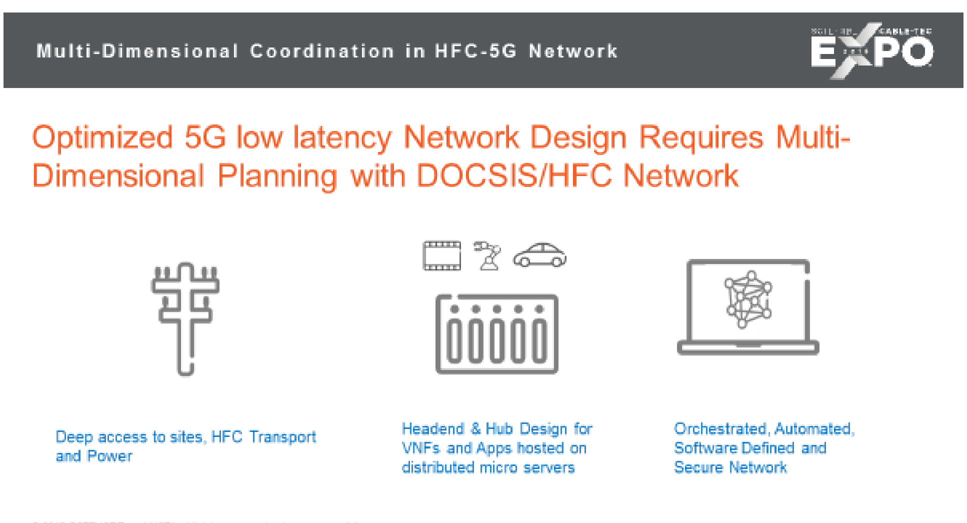
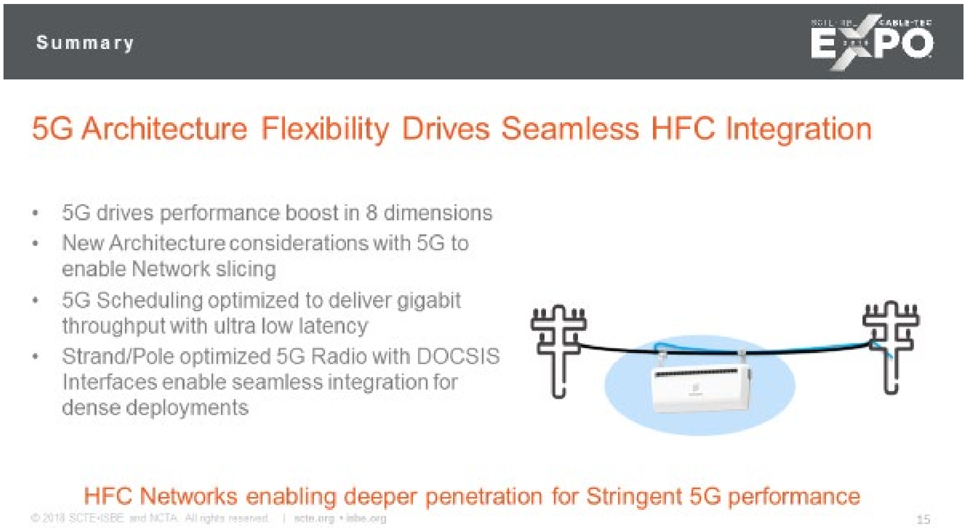


Figure 11 - Coordination Considerations between HFC and 5G

Conclusion



In summary, 5G is designed to deliver multi gigabit throughput experience, sub 5 ms latency, 99.9999% reliability and 3 ft positioning accuracy. This is achieved via highly optimized a network slice-aware distributed architecture complemented with microsecond level radio scheduling.

From a radio site perspective, for several of the use cases, the DOCSIS / HFC can be directly connected into an integrated Radio-Baseband unit and provide backhaul to a Headend or a Regional Distribution Center site. This will enable a deeper penetration of 5G small cells into urban and residential areas, thereby enhancing the coverage of IoT and broadband services.

Abbreviations

5G NR	5G New Radio
AAS	Adaptive Antenna System
BFF	Beam Forming Function
BPF	Baseband Processing Function
cMTC	Critical Machine Type Communication
CU	Central Unit
CU-C	Central Unit Control Plane
CU-U	Central Unit User Plane
DU	Distributed Unit
eCPRI	Packet based Common Public Radio Interface
FWA	Fixed Wireless Access
IUA	Instantaneous Uplink Access
KPI	Key Performance Indicator
MBB	Mobile Broadband
mMTC	Massive Machine Type Communication
mmWave	millimeter Wave
PPF	Packet Processing Function
RF	Radio Function

RPF	Radio Processing Function
RCF	Radio Control Function
TTI	Transmit Time Interval