5G AT THE EDGE

Executive Summary

October 2019

5G americas

5G at the Edge

Executive Summary

5G and mobile edge computing will go hand-in-hand. Mobile edge computing refers to the buildout of edge computing systems in mobile telecommunications systems, particularly in 5G scenarios. The reason for edge computing technology partnering with 5G is that 5G will vitalize the 4th industrial revolution – Industry 4.0 – and reinvent entire industries through the Internet of Things in the coming years. With new 5G use cases and applications crossing all areas of horizontal and vertical industries, there will be an even greater need for resolving the vast and varied requirements in data traffic, cell densification and spectrum bands. 5G and Internet of Things (IoT) applications need large bandwidth, strict latency, faster data rates and high reliability, and these requirements can only be secured with efficient architectures. Edge computing is critical for 5G networks to enable new applications and thus begins what promises to be an excellent marriage of technologies.

In a 5G mobile network, the edge is the network located as close as possible to the source of data. What the 'edge' is depends upon the use case. In mobile wireless services, the edge might be a cell phone or perhaps a cell tower; in the Internet of Things, a vehicle connected to everything (V2X) could be the edge; or in manufacturing Machine-to-Machine (M2M) could be the edge; or finally a laptop might be the edge in an enterprise setting. IoT applications are expected to benefit significantly from edge.

Essentially, edge connectivity is a highly distributed computing environment used by applications to store and process content as close as possible to the end user. Often times, these applications require real-time radio and network information to offer a personalized experience to their user. Edge devices can be any device that produces data, such as sensors, industrial machines and other device that collects data.

For years, networks have been evolving to provide reliable communications and computing capabilities at the edge. Key benefits include the enablement of lower latency, enhanced security and backhaul cost savings.

5G and the Cloud

5G will now need to support heterogeneous mobility networks with advanced distributed cloud services, by combining wireless networks with an edge cloud that is agile, virtualized and software-defined. Edge architectures must be redesigned to satisfy the stringent SLAs (Service Level Agreements) of emerging applications. Compute resources must move closer to the Edge to satisfy latency and bandwidth constraints, which requires the entire architecture to be highly distributed. Small cell-based networks will need to be deployed in order to enable high speed data for shorter ranges to a cloud or data center.

Networks are evolving to primarily utilize Software Defined Networking (SDN), Network Function Virtualization (NFV) and cloud-native architectures to enable disaggregation and virtualization of primary functions. This leads to separation of control plane and user plane and introduces capabilities such as network slicing and mobile edge computing. 5G also shifts to a Services-Based Architecture (SBA), which moves from a response-request method of communication to a producer-consumer type model. 5G

architecture is much more distributed than previous wireless generations, needing many more cells which enable much more distributed processing.

Cloud applications today are typically built using a client-server architecture. “Front end” developers implement the client software which executes in a web browser or natively on the device. “Back end developers” implement the server software which runs in a cloud data center. “Infrastructure developers” are responsible for back end computing and network connectivity infrastructure.

Cloud developers have come to expect easy to use software/tools, pre-packaged and easy deployment, seamless management, and data privacy and security benefits to enable services to be easily developed and deployed at a high level of abstraction. For example, to abstract service to service communication details, a “service mesh” of software-based “sidecar” proxies in their own infrastructure layer and is typically employed to route requests between services.

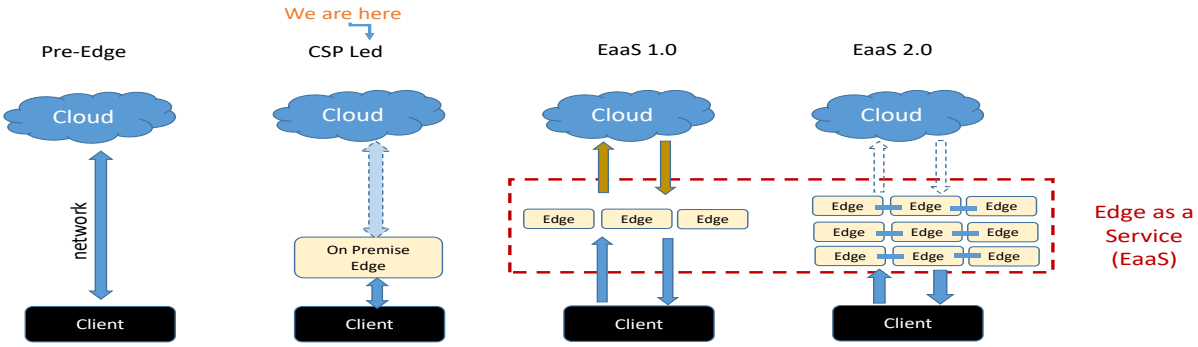

Figure 1. Expected Edge Evolution.

The “pre-edge” environment, as shown in Figure 1, has a 2-tier architecture where Communications Service Providers (CoSP) provide the access network infrastructure to connect clients and on-premises networks to the CSP infrastructure. CoSPs, burdened by heavy investments required to modernize their access networks and increasingly dissatisfied at being relegated to being merely a transport service provider, represent another important EaaS (Edge-as-a-Service) target customer opportunity in addition to the Cloud Service Providers (CSPs).

5G architecture is essentially designed to take advantage of cloud-native concepts – the ability to leverage self-contained functions within or across data centers (the cloud), communicate in a micro-services environment, and work together to deliver services and applications. Disaggregation and virtualization are two key elements of 5G cloud-native architecture.

Disaggregation and Virtualization

5G networks take advantage of cloud-native concepts, such as containerization and micro-services through techniques like software-defined networking (SDN) and network function virtualization (NFV), even as they grow and become more distributed.

The use of SDN and NFV basically allows the disaggregation and virtualization of many of the telecommunications and mobility functions like S/P-GW (Serving/Public-Gateway), MME (Mobility Management Entity), Radio Access Network (RAN) CU/DU (Central Unit/Distributed Unit), TDF (Traffic

Detection Function), Internet Protocol (IP) Routing, and Ethernet Encapsulation/Switching. These functions are hosted as software services and dynamically instantiated in different parts of the network segments; thus, the overall 5G network is designed to be software configurable.

Control Plane and User Plane Separation (CUPS) is the concept of disaggregation that allows these two planes to exist on separate devices or at separate locations within the network. As an example, in the core, CUPS separates the user plane functionality from control plane functionality in the Serving Gateway (S-GW), Public Data Network Gateway (P-GW) and Traffic Detection Functions (TDF). Separating the control plane from the user plane allows the two planes to scale independently, without having to augment the resources of one plane when additional resources are only required in the other plane. And, the separation allows planes to operate at a distance from each other—they're no longer required to be co-located.

From a functional disaggregation perspective, the Service, Control, Data and Management Planes separation are already being realized on transport systems using SDN. From the direction provided by the latest standard specifications for Long Term Evolution-Advanced (LTE-A) and 5G, functional disaggregation also takes place on the mobile network element layer. For example, 5G RAN is disaggregated into CU (Central Unit) and DU (Distributed Unit) functions; and within the CU, they are disaggregated into CU-CP (Control Plane) and CU-DP (Data Plane). When all data plane functions of different network elements are disaggregated, the data plane is distributed using a consolidated set of protocols. The data plane functions could either be realized via a Virtual Network Function (VNF) construct Multi-Access Edge Computing (MEC) platform or as a programmable Application-Specific Integrated Circuit (ASIC) construct (Programmable Transport Underlay). The transport control plane and data plane protocols are expected to consolidate and simplify as network systems adopt a cloud-native construct.

In the Radio Access Network (RAN), cloudification allows the disaggregation of the Remote Radio Unit (RRU) from the Baseband Unit (BBU.) By separating these functions, it becomes possible to create a pool of BBU resources that supports several distributed RRUs. This is referred to as a C-RAN, therefore, Cloud-RAN, where elements of the RAN can be centralized and implemented in the cloud as well. Doing this allows a more efficient use of resources in the RAN. It also creates some challenges, such as the need for fronthaul connectivity between the RRUs and the BBUs. This challenge is being addressed by architectures that define the splits at different locations in the RAN, with the different architectures having trade-offs between bandwidth requirements and the ability to centralize resources.

Machine Learning and Artificial Intelligence

Edge computing is particularly important for machine learning and other forms of artificial intelligence. 5G incorporates edge computing into wireless networks with emerging open source initiatives and standards to manage data across the network, from radio access, transport, to the core - enabling powerful new capabilities like network slicing. Edge computing uses innovative artificial intelligence and machine learning technologies to improve the management of data workloads across networks.

- Edge networks will be designed as autonomous and intelligent systems that sense the context around their environment and application, applying network resources in real-time.
- Key areas for artificial intelligence and machine learning include intelligent platforms, radio access network optimizations, end-to-end application delivery, network analytics and management, and subscriber and service management.

- Data from 5G wireless networks will be used for edge computing-based training of sophisticated artificial intelligence algorithms, enhancing new forms of emerging autonomous services and applications.

New 5G Spectrum = New Advanced Antenna Systems

In order to reduce the need for transport bandwidth for all the antenna signals between the baseband, radio and the antenna itself, 3GPP standardized an alternative architecture for the gNodeB (Next Generation NodeB) allowing digital processing for antenna beam-forming closer to the antenna elements, while also defining the rest of the processing in separate control and user plane functions.

It can be noted that the increase in the number of sites needed for 5G New Radio (NR), also increases the need to closely coordinate the signals from multiple antennas to reduce interference and maximize capacity for the Cells/User Equipment (UE). This drives the need for more processing close to the antennas.

All of this requires an architecture where lower layer baseband functions working closely with the radio can be separated architecturally from higher layer baseband functions. 3GPP standards facilitate this by a new interface, called 'F1' in 3GPP Release 15 (Rel-15), that allows the higher layer part of the protocol stack to be processed separately from the lower layer baseband (Distributed Unit -DU) functions.

This paves way for running the higher layer Centralized Unit (CU) for control and user plane functions on a generic server in a cloud environment. This would also allow the control and user plane to individually scale for optimal deployment to meet Edge capacity and coverage requirements for new use cases that place higher demands on high capacity and low latency. Open Source bodies such as O-RAN Alliance are working to realize similar virtualization objectives, starting with disaggregated (DU and Radio) and integrated (DU+Radio) deployments for small cells and indoor/outdoor pico/micro/ femto deployments.

Looking forward, there is an opportunity to deploy servers for part of the baseband processing further out in the network edge, including on-premise and enterprise locations. With this architecture option, Edge computing can then enable new use cases for operators and enterprises for mutual benefit.

Open Source

Different parts of a 5G network – even spectrum - can be shared or use open-source hardware or software, due to the separation of the control and user functions, allowing for new applications and services. Open source initiatives are critical to the development of standards in how these apps and services will be created.

A new reference architecture for edge computing-enabled 5G systems is being shaped that will have broad implications for how wireless networks operate in the future. This reference architecture will combine the elements of network function disaggregation with new models of programmability, new interconnections between radio access, transport and core networks, and implement artificial intelligence for self-organization.

These changes may lead to entirely new Internet architectures that can include options for information-centric networks (or their hybrids), overlay which can interoperate on top of existing models for Wi-Fi, Bluetooth, cellular, Transmission Control Protocol (TCP)/Internet Protocol (IP) or Ethernet, or even more exotic options that stack different capabilities.

Early 5G deployments will typically involve the 5GC (5G Core) co-existing with the EPC (Evolved Packet Core), which entails additional Non-Standalone (NSA) use cases. Standalone (SA) mode deployments are also emerging. Figure 2 shows an estimated timeline.

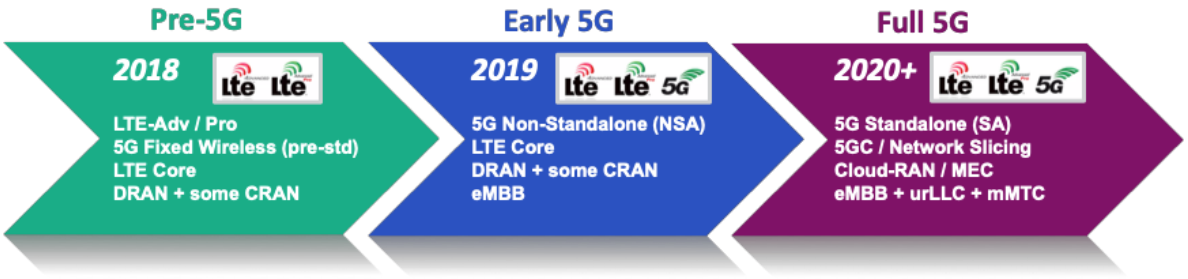

Figure 5.1. Phased Deployments of 5G.

The 5GC deployment will enable new Network Slicing capabilities. It will also enable additional applications such as URLLC, massive machine-type communications (mMTC) and for the first time, delivering native Ethernet services over the 5G NR. With the introduction and maturation of technologies such as SDN and NFV during the last decade, 5G design principles were defined to take full advantage of these software-driven innovations.[1] This translates to the virtualization and disaggregation of many of the RAN and mobile core functions. Therefore, there will be an increased use of a Cloud-RAN approach which has significant architectural implications for the network architecture.

To this end, 3GPP Rel-15 has introduced Split RAN Architecture which disaggregates the RAN baseband functions into functional blocks which could be optimally distributed to maximize spectral efficiency while minimizing operational cost. The baseband functions include both a real-time processing part and a non-real time processing part. The real-time baseband handles radio functions like Dynamic Resource Allocation (Scheduler), gNB Measurement, Configuration and Provisioning, and Radio Admission Control.

The non-real-time baseband handles radio functions like Inter-Cell Radio Resource Management (RRM), Resource Block (RB) Control and Connection Mobility & Continuity functions. These baseband functions are mapped into the Distributed Unit (DU) for real-time baseband processing; and Centralized Unit (CU) for non-real-time processing, respectively. 3GPP Split RAN architecture is depicted in Figure 3.

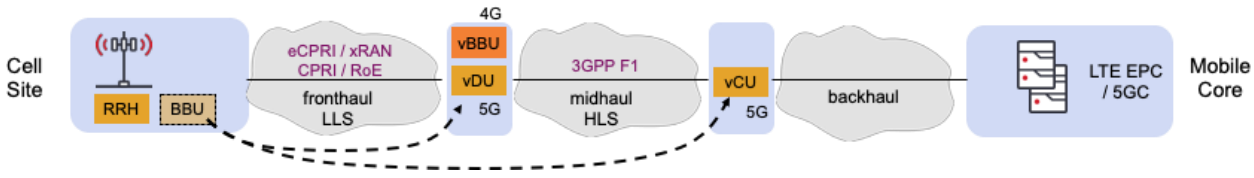

Figure 3. Split RAN Architecture.

Future Directions

The current internet architecture is based upon a host-centric communication model. Having an address to connect to and establishing a session between the host and the client is a pre-requisite for receiving data. Internet usage has evolved however, and there is a paradigm shift in Internet model along with a

[1] *NGMN 5G Whitepaper*, February 2015.

need for security and mobility support. This has led researchers into considering a name-based architecture. Routing, forwarding, caching and data-transfer operations are performed on topology-independent content names rather than on IP addresses. Naming data chunks allows the ICN (Information-Centric Network) to directly interpret and treat content per its semantics without the need for deep packet inspection (DPI) or delegation to the application layer.

Historically the evolution of ICN has been primarily for data. However, ICN can also be used to orchestrate compute. Instead of sending an interest packet for a piece of information, the user (device) can request the execution of a function by its name. The network then routes the information to the closest resource that can compute the desired functions and returns the processed data back.

This is a powerful paradigm where the entire edge can be a compute server. If the user is in a new environment, without any knowledge of the closest edge server, the device can still request the network to orchestrate the compute. Named Function Networking (NFN), Named Function as a Service (NFaaS) and Remote Method Invocation over ICN (RICE) are examples of implementing dynamic and distributed compute within the network. In Named Function Networking (NFN), the network's role is to resolve names to computations by reducing λ-expressions. NFaaS builds on very lightweight Virtual Machines (VM) and allows for dynamic execution of custom code. RICE presents a unified approach to remote function invocation in ICN that exploits the attractive ICN properties of name-based routing, receiver-driven flow and congestion control, flow balance, and object-oriented security.

While edge computing today is performed in real time, the orchestration is largely performed out of band through centralized architectures. However, the compute and edge service requirement could change dynamically at the edge due to mobility, wireless links being up or down, changing contexts and etcetera. ICN's fundamental architecture is distributed and decentralized, which is well aligned with the needs of dynamic orchestration. Further protocols like NFN and RICE enable orchestration for deep learning or federated learning at the edge by tying the orchestration with the execution.

Conclusion

With the growth mobile data traffic and diverse mobile use cases, there is a universal acceptance to adopt increased virtualization in mobile networks with more software-driven equipment that is flexible, intelligent and energy efficient to facilitate efficient radio access and converged network design.

Society is on the verge of groundbreaking changes in 5G wireless capabilities to serve new services and applications of the immersed connected society. Edge computing can facilitate advancements in mobility, computation, storage and acceleration features which all depend on 5G. Likewise, the future direction of networking requires that the next generation of Edge reference architecture prioritizes its relationship fully with 5G.