5G AT THE EDGE

5G Americas Whitepaper

October 2019

TABLE OF CONTENTS

1. INTRODUCTION

Over the next few years, 5G is expected to reinvent entire industries with new use cases, business models, and organizations that will emerge in response to shifting technology and business landscapes. The growth of 5G wireless technologies are necessitating approaches that include Edge computing architectures.

Today, emerging 5G markets including AR/VR (Augmented Reality and Virtual Reality), V2X (Vehicle-to-Everything), transportation, manufacturing, health and education are being tooled with applications that operate in a time-sensitive fashion, requiring a range of data bandwidth, varying degrees of cell densification and spectrum operating range. Unlike previous generations, 5G platforms are relying on strong distributed cloud foundations of network and compute transformation that will lead operators to new market growth.

For years, networks have been evolving to provide reliable communications and computing capabilities at the Edge. Edge Computing brings compute, storage and networking closer to applications, devices and users. Some of its key benefits include the enablement of lower latency, enhanced security and backhaul cost savings. Edge computing architectures can also be complex as it is not a one size fits all. Operators will typically deploy Edge computing elements to address specific services, applications and use cases.

This new mobile era will be powered by emergent applications that will span across a ubiquity of mobile devices and edge clouds. The combination of 5G and the Internet of Things (IoT) are adding newer applications that need connectivity not just between people, but also between things. Many of these applications have large bandwidth needs and strict latency requirements like video traffic, gaming, AR/VR and connected cars.

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.

Of course, there are new architecture opportunities and challenges that are being addressed that also require open interfaces to provide edge cloud technologies such as acceleration, analytics, AI and ML (Artificial Intelligence and Machine Learning) engines to complement 5G. Fortunately, a convergence of several overlapping technology trends is giving rise to new solutions that are necessary for numerous IoT (Internet of Things) applications.

One of the latest developments in 5G is O-RAN (Open Radio Access Network), which directly addresses service agility and cost considerations of networks. As mobile traffic increases, both mobile networks and the equipment running them must become more flexible, software-driven, virtualized, intelligent and energy efficient. Work on O-RAN will lead to reference RAN and converged network design that is more open and will include smart features that define real-time control and analytics.

The new reference design will also consider embedded machine learning systems and artificial intelligence back-end modules to empower network intelligence and increase service agility. Technologies from open source and open white-box network elements will provide key software and hardware components for these reference designs.

Another development is the growth of cloud computing, which has been growing as a key driver for the Internet Technology (IT) industry over the past decade. Today, cloud computing has come of age, and is

scalable and integral to any agile IT architecture in which enterprises use public, on-premises private clouds. Cloud computing also addresses various hybrid models with high-bandwidth applications supported by end devices as well. The convergence of Edge with the Cloud is being driven by requirements that support applications which are easy to operate, simple, agile, elastic and adaptable.

New groundbreaking possibilities are also emerging with the combination of 5G and artificial intelligence. AI can play a significant role in improving the network operations especially at the edge of the network and create new service opportunities. It is increasingly being viewed as a platform for operators to provide open edge services and developers to create applications supporting consumers, enterprises and multiple verticals.

The latest trend in this direction is creating intelligence at the edge by utilizing distributed architectures for AI systems where the end devices are configured to make time sensitive decisions, while the cloud is used for training and fine-tuning of AI. Additionally, sensitivity to privacy of data is also resulting in end devices playing a larger role in AI training, building on their inference capabilities.

Therefore, a 5G network leads to a highly distributed and decentralized system where AI features can be embedded in several layers of the network architecture to enable local data generation and processing, as well as central data consolidation. Intelligence built into the 5G system can allow better management of services where intelligent functions can be customized for specific services allowing them to operate resiliently, securely and efficiently. 5G communications tie it all together, provisioning the communication platform between Cloud and Edge.

This paper provides additional detail on the evolution of 5G architecture, its adaptability to existing Cloud architecture, and the various methodologies that are currently being adopted for Cloud-native applications. It covers a detailed panorama of emerging use cases and their requirements for 5G networks that can facilitate advanced mobility, compute, storage and acceleration features for applications with ranging latency considerations.

Most importantly, this paper supplies an in-depth view of the various industry initiatives in defining the EDGE architectures while keeping in view of the emerging networks that are planned to serve 5G applications. Overall, it defines the next generation Edge reference architecture and explores future directions in networking.

2. NEXT GENERATION 5G TECHNOLOGY FOR EDGE COMPUTING

As the next generation of wireless technology, 5G differs starkly from previous wireless generations. Where previous wireless generations were designed to connect people to people and to connect people to the Internet, 5G extends even more broadly. It connects things to people, to the Internet, and to other things. It also addresses an expanded set of industry verticals necessitating the networks to be more flexible in meeting a wider range of requirements on latency, cell density, spectrum of radio frequencies and data rates.

Consequently, 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 focuses on supporting three broad categories of applications to enable unprecedented use-cases: eMBB (enhanced Mobile Broadband), mMTC (massive Machine-Type Communication) and URLLC (Ultra-Reliable Low-Latency Communication.) 5G architecture is much more distributed than previous wireless generations, needing many more cells which enable much more distributed processing.

The following subsections describe the key characteristics and features of 5G architecture in its interaction with the cloud and in supporting broader sets of vertical domains.

2.1 5G CLOUD-NATIVE ARCHITECTURE: DISAGGREGATION AND VIRTUALIZATION

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.

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), 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 TDF functions. 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.

This has also led to initiatives such as xRAN**.** In 5G, fronthaul will move away from CPRI-centric (Common Public Radio Interface) interfaces to one of several types of packetized fronthaul. Multiple SDOs (Standards Definition Organizations) are releasing specifications for packetized fronthaul for usage with 5G and LTE-A Pro networks. These include: eCPRI 1.0 from CPRI.info; Institute of Electrical and Electronics Engineers (IEEE) 1914.3 RoE (Radio over Ethernet); and xRAN.org's xRAN Fronthaul 1.0.

All these fronthaul specifications are focused on realizing the fronthaul in the physical layer of the RAN. In particular, xRAN Fronthaul 1.0 drives the possibility of open RE (Radio Equipment) and REC (Radio Equipment Controller), where multi-vendor radio controllers and active antennas could interoperate; as well as providing an open ecosystem of radio control applications like Hybrid Load Balancing, eICIC (Enhanced Intercell Interference Coordination) and SON (Self-Organizing Networks), and etcetera. Note the Open RAN Alliance (O-RAN) has recently formed to continue the C-RAN/xRAN work. O-RAN is updating and extending the xRAN specifications based on an architecture of RAN Intelligent Controller (RIC) <-> O-CU <-> O-DU <-> O-RU.

Another major difference between 5G and previous versions of wireless networks lies in the radio spectrum being used. 5G requires much larger amounts of bandwidth, and therefore the quantity of spectrum is also increased and its locations differ as well. 5G uses some frequency ranges below 6 GHz, while a move into the mmWave (millimeter wave) range provides access to much larger amount of contiguous spectrum.

One disadvantage of the higher frequency ranges is the shorter distances the signals can propagate, and their inability to penetrate fixed objects and susceptibility to rain fade. These limitations therefore impact the quantity and location of base stations needed for proper coverage. This can be solved with a larger number of smaller base stations (therefore, Small cells, Micro cells, Femto/Pico cells). A larger quantity of sites needing interconnection produces a more distributed architecture, which is described later in this section.

Network slicing is a term applied to the technique of isolating the performance of a portion of the network compared to another. The purpose is to be able to deliver the stringent characteristics that 5G offers (eMBB, MTC, URLLC) for specific subsets of users, operators or applications. Use cases that would take advantage of a network slice vary, but industrial, law enforcement or emergency responders are just some of the examples that would make good use of a slice. Current implementations differentiate between soft slices and hard slices. Soft slices use control plane software-oriented techniques to slice the network, techniques like VPNs and Segment Routing. Hard slicing uses techniques much more closely coupled to the hardware itself, such as Optical Transport Network (OTN) switching or leveraging FlexE (Flexible Ethernet).

2.2 DISTRIBUTED ARCHITECTURE

5G will have a much more distributed architecture than previous wireless versions. A distributed architecture is needed because: 1) the variety of frequency ranges used to deliver high bandwidth and the density requirements of 5G can require many more base stations, or small cells; and 2) the latency requirements of 5G requires portions of application processing much closer to the user. There are other reasons that contribute to the distributed architecture of 5G:

Densification of Radios—the need for increased coverage and higher data transfer rates on user devices will mandate an increased number of radios. The introduction of millimetre wave radios and

the continued miniaturization of radio technology will contribute to significantly more radios being deployed.

Increased C-RAN (Cloud/Centralized RAN) deployment—with a Split CU (Centralized Unit) and DU (Distributed Unit) architecture being defined in 3GPP 5G specifications, there will be an increase in the centralization of baseband processing and radio control functions in many locations. The advantages of RAN centralization include more efficient support of Carrier Aggregation (CA), Network MIMO (Multiple-Input/Multiple-Output), Downlink CoMP (Co-ordinated Multi-Point), Uplink L1 CoMP Joint Processing, and other features. With densification, many of those dense sites will not have the real estate nor cost efficiency for traditional base station systems.

Introduction of Edge Compute—the distribution of compute power closer to the edge will provide a better quality of experience and fulfil new use cases. Distributed compute will also host any virtualized RAN functions, such as virtual CU and DU and other mobility control functions (for example, Automatic Protection Switching (APS) Mode Mismatch Failure (AMF) as part of a CUPS architecture.

Introduction of New Applications—in addition to standard voice and broadband data, 5G introduces new applications, primarily based on D2D (device-to-device) communications and IoT, that will drive new traffic patterns for data and control traffic. These applications also dictate a widely varying set of performance characteristics in the air interface and transport, and in many cases leverage edge compute mentioned above. With applications now driving traffic patterns, the network and networking protocols should maintain minimal state information to allow stability, scalability, simplicity and agility in operations. Source routing protocols will be strategically important in a 5G network.

Expected Bandwidth Increases—With the opening up of additional spectrum in the mmWave band and Band 42 (3.5GHz) for 5G NR, coupled with antenna techniques such as Massive MIMO, the amount of transport bandwidth is going to increase at the Fronthaul LLS (Low Layer Split), Fronthaul HLS (High Layer Split) and Backhaul segments. New transport physical interface distributions will also emerge as follows: 10GE ~ 25GE for Fronthaul LLS, 1GE ~ 10GE for Fronthaul LLS (3GPP F1 interface) and 50GE ~ 100GE for C-RAN Hub Site Backhaul.

Deployment of Deterministic - Behavior Packet Transport Systems—Handling low latency applications and guaranteeing their upper bound latency limits with fixed jitter and low packet error loss rate will be important. However, it will be complicated from an engineering standpoint, and expensive to just reply on L2/L3 VPNs (Virtual Private Network) without any forms of hard pipe enforcements. Different forms of Deterministic Networking Packet Transport technologies will become important for 5G transport systems and they will be deployed in different segments of the networks. For instance, IEEE 802.1 TSN (Time Sensitive Network) is a good packet access technology, while Flex-Ethernet (Flex-E) and OTN are possible options to deploy in core packet networks, especially when the core networks involve data center to data center connectivity.

2.3 EDGE CONSIDERATIONS FOR RADIO ACCESS

5G radio access networks, based on the New Radio (NR) standard, will provide new levels of capacity, peak data rates and low latency, all simultaneously to tens of thousands of users which is far beyond the capabilities of previous cellular generations.

The high capacity is principally due to the allocation of new spectrum, particularly mmWave and the sub 6 GHz allocation, allowing the use of new radio technology that can significantly increase spectral efficiency and capacity. The higher the frequency of spectrum, the smaller the antenna requirement (proportional to

the wavelength). Due to their propagation characteristics, higher frequency carriers have limited distance in which to reach the radios in an access network, indicating that more sites will be needed to take advantage of the spectrum.

Multiple-Input, Multiple-Output (MIMO) multi-antenna transmission technology uses antenna physical size versus frequency to deliver on capacity needs. MIMO achieves this by significantly increasing the number of transmission ports to boost network capacity and data throughput at the cost of needing higher processing resources for all the antenna signals. However, without further enhancements, this places a tremendous burden on the transport capacity in the RAN network due to the need to transport all individual antenna signals to the baseband processing function.

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 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.

3. 5G AND EDGE COMPUTE USE CASES

In recent years cloud computing has been the dominant source of growth of data center deployments. A small number of hyper scale cloud-based digital platforms, the *Super 7 Cloud Service Providers*" (CSPs), continue to grow in power and influence. Cloud computing economics have been compelling due to significant economies of scale and scope enabling attractive pay as you go pricing models with highly elastic capacity due to large scale sharing of reusable compute, network, and storage resources. By 2020, more than 70 percent of applications published will be targeted for cloud/web deployment.[1]

3.1 CURRENT LANDSCAPE

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

[1] *Cisco Global Cloud Index 2019, Forecast & Methodology, 2016-2021,* white paper by Cisco. 19 November 2018.

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.

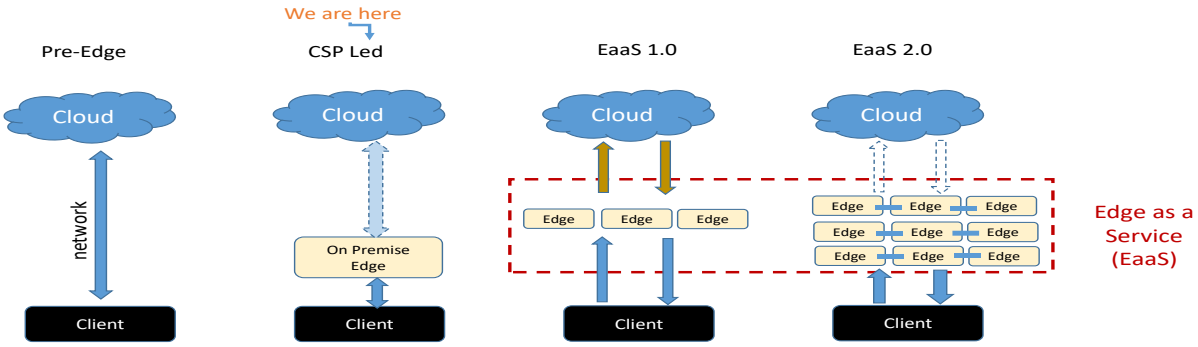

Figure 3.1. Expected Edge Evolution.

As shown in Figure 3.1, the "pre-edge" environment has a 2-tier architecture. In this architecture, 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 class of EaaS (Edge-as-a-Service) target customer opportunity in addition to the CSPs.

Edge computing is particularly important for machine learning and other forms of artificial intelligence, such as image recognition, speech analysis, and large-scale use of sensors. Specific use cases may include video security surveillance, automated driving, connected industrial robots, traffic flow and congestion prediction for smart city, and so on. In the case of industrial IoT or self-driving cars, a processing delay between the device and the cloud can mean disaster.

Many AI applications that are enabled by edge computing lie in the categories of access network analytics, video analytics, Machine-to-Machine (M2M) analytics, augmented reality and location services, among others. As for the wireless radio communications to the client by 5G and emerging new innovations in core architecture, much of the excitement is envisioning where and how the technology can improve existing services or enable entirely new ones. It's impossible to enumerate all the possible improvements or new services. The following list gives some of the best example use-cases and how their extreme characteristics are supported by 5G.

- **Autonomous vehicle control**. Autonomous cars require very low latency and high reliability, for vehicle-to-vehicle, vehicle-to-people and vehicle-to-sensors systems. Any self-driving applications that use video would likely require high data rates as well.

- **Emergency communication**. In an emergency, high reliability is critical to re-establishing communications infrastructure that may have been damaged and allowing first responders to support rescue efforts. Energy efficiency is also critical in areas where the power grid has been

affected. Extreme density (as few base nodes will be available immediately) and extreme distance from the base nodes are requirements.

- **Factory cell automation/Smart Factory**. Devices communicating within a factory require very low latency and high reliability, as the commands from a controller must reach actuators with tight time constraints to keep the assembly line moving. Combining low latency and high reliability communications with intelligence of machine learning or artificial intelligence creates a “light outs” factory where humans aren’t even required.

- **Large outdoor event/stadium**. Large outdoor events require temporary access to large amount of bandwidth, all within a relatively small geographical area, thus producing a high density of connectivity.

- **Massive amount of geographically spread devices/IoT**. This scenario constitutes connectivity to a massive number of devices spread over a geographical area. This is also referred to as “Internet of Things”, or IoT, and is accomplished by use of sensors and actuators. Because the number of devices will be high, each device needs to be inexpensive and have a long battery life.

- **Remote surgery and examination**. The Tactile Internet has been described as a network that allows an operator to see and feel the physical world from a remote distance. Remote surgery is an extreme example of this, where very low latency communications with high reliability is critical to enabling a potentially “life-at-stake” service.

- **Smart city**. Connecting the sensors and actuators across a city allow intelligent decisions to be made and allows the city to become “smarter.” This can be anything from utility meters (gas, electrical, water) to trash cans that broadcast their need to be emptied. This requires a high density of low power devices. Cloud services in a “smart office” will require high data rates at low latency, whereas small devices (like wearables) can tolerable moderate latency with at lower data rates.

- **Smart grid network**. Digital technology enhancements to the electrical grid allow for large amounts of telemetry data to be available. Electrical grids require precise synchronization to keep the grid stable during unpredictable loads or demands. Smart grids will have the ability to detect changes, intelligently understand what actions need to be taken, and then perform these actions within milliseconds or less. Bandwidth and low latency communications are required here.

- **Traffic information**. Related to the Smart City use-case, the ability to collect and process data about traffic and traffic jams requires a huge number of sensors measuring and collecting data, therefore, speed or rate data. This information could also be combined with emergency services coordination, whereas traffic information could be fed to local responders to guide them to or from disasters (for example, routing injured people to hospitals or treatment centers).

- **Virtual Reality, Augmented Reality and Gaming**. Much information has been written about the possibility of overlaying interesting/important information on top of a live video stream, or the ability to be completely immersed in a virtual world. These use-cases will potentially require massive amounts of bandwidth, but within certain latencies, also make the reaction appear “real-time” to humans.

3.2 USE CASE DESCRIPTIONS

In this section, further detail is provided on some of the compelling use cases that leverage 5G capabilities and new developments in EDGE computing.

3.2.1 AUGMENTED REALITY

Augmented reality (AR) use cases mostly involve image recognition (face, objects, and etcetera) and their analytics. The client devices are typically in the form of wearables or mobile phones with a camera. A typical usage scenario is when a user points camera device to an object and sees useful annotations about the object. There are already several smartphone applications that offer AR services such as ViewAR, Vivino, Augment, and etcetera. Each of the applications or wearables are designed for providing AR for specific purposes. As an example, Vivino provides useful information about a wine when a picture of the wine bottle's label is captured and provided. Similarly, ViewAR helps with 3D visualization for packaging and also planning the space in a room.

An AR scenario typically consists of three stages: 1) sending a captured image from a device, 2) image recognition and object identification, and 3) performing analysis and producing a result (in the form of annotations of useful information found from the internet, object identification, and etcetera). Steps 2 and 3 are generally the most time-consuming parts. An AR service that is dedicated for specific functionalities (such as Vivino for wine), will specify certain targets to an image recognition service. The image recognition service will identify the image and look for these targets. The AR service will then use the result to perform further operations such as analytics or finding more information from the internet. The results are then displayed to the user.

It is common that the AR service uses another service for image recognition. For example, all the above-mentioned apps use Vuforia as the main engine for its Application Program Interface (API) so that the image recognition can find the specified targets. The image recognition can be performed both on the device and on the cloud. Performing it on device with an on-device database is not power efficient, while performing the same over the cloud takes longer time.

Another example is the Catchoom cloud image recognition service. The service allows AR applications to send requests to their service in the cloud through a Representational State Transfer conforming Web Services (RESTful) API and responds with the image recognition. For instance, PhooDi is an application that uses the Catchoom cloud image recognition service to provide users with nutritional information about food products. The AR application can finally perform the third step from the client device after the image is recognized – by looking it up the information on the internet and presenting it to the user.

Image recognition services such as Vuforia and Catchoom have several AR applications using their service. This can only be expected to rise the in the future as both the number of new AR applications and their users increase. Since their services are currently hosted in the cloud, the latency and turnaround time of image recognition is high. For example, it takes a couple of seconds to find information about a wine through Vivino. This would be unacceptable in scenarios where wearables are involved, time-critical annotations are required, and for which real time movement needs to be accounted. Therefore, moving the image recognition services from the cloud to the edge can improve the total turnaround time giving users seamless experience.

3.2.2 VIDEO ANALYTICS AT THE EDGE

Video analytics has a significant role to play in a variety of industries and use cases. For example, face recognition from traffic and security cameras is already playing an essential role in law and order. Several other types of analytics can be performed on video content such as object tracking, motion detection, event detection, flame and smoke detection, AI learning of patterns in live stream or archive of videos, and etcetera. Presently, video analytics is done on the cloud or on dedicated private servers depending upon the need and the functions to be established. Performing video analytics at the edge poses as both a requirement as well as an opportunity for several fields. For example:

- Surveillance and public safety: Processing live video streams almost instantaneously at the edge can lead to better surveillance and enforcing law and order. Two examples of this use case are face detection and incident identification and triggering, which would allow law enforcement officers to take immediate actions involving an incident.
- Supporting connecting cars and self-driving: Live streaming of the scene as seen by a self-driving car needs be analyzed in very short time to decide the actions to be taken by the car. A self-driving vehicle could already contain resources to process the scene instantaneously. Edge video analytics can serve for processing (or preprocessing) farther scenes or post-processing video scenes for continual training and feedback.
- Enabling smart cities and IoT: Video analytics at the edge is an important element to enable smart cities. For example, traffic video analysis can be used to route traffic in the most efficient way. Fire or smoke detection in an area can be identified instantaneously and ensure no traffic is continued towards the danger zone by sending feedback to both the city infrastructure as well as to connected cars in an area.
- Enhanced infotainment services: Video analytics at the edge can be used to enhance the real-life experience of event audiences such as sports, concerts and other shows. Videos from different camera angles at an event can be analyzed and applied with AR/VR functions and presented to a live audience through large screens, smart phones, VR devices, and etcetera.

3.2.3 CONTENT DISTRIBUTION NETWORKING AND CONTENT CACHING AT THE EDGE

According to the *Global Mobile Data Traffic Forecast Update 2016-2021*, global mobile data traffic grew 63 percent in 2016 reaching 7.2 exabytes per month at the end of 2016. Videos accounted for about 60 percent of mobile data traffic. By 2021, the global mobile data traffic is expected to reach 49 exabytes with 78 percent of the annual traffic expected to be video.

As much of this traffic is video content, the possibility of redundant content being delivered to users in the same region is high. According to caching software provider Qwilt, over 80 percent of the video traffic only consists of 10 percent of the titles. Therefore, duplicates of the videos are being repeated, increasing the backhaul traffic and OPEX costs.

Consumers of the traffic are mainly users of handheld devices such as smartphones, tablets, laptops and etcetera. Therefore, the edge cloud becomes an appropriate infrastructure to cache content, which can significantly reduce the backhaul traffic. The potential to save on Operational Expenditures (OPEX) costs for the TSPs (Telecommunications Service Provider) is increased when considering that content caching may not be limited to only videos - its scope widens to other data types such as music and documents.

There are three primary ways to perform content caching:

1. Content caching based on traffic learning: This caches content in a region based on its popularity and growing requests/traffic. Along with it, content that is similar to specific popular content can also be cached proactively.
2. Targeted content caching: This caches content at the edge for a target audience. For example, an audience in a stadium or a gathering.
3. User guided caching: The user indicates the content that is to be cached (for a service fee or a part of the data plan of the user). For example, videos that the user adds as "watch later" in YouTube or puts in a favorite list in Netflix could be candidates for caching. Since there may be several users in the same region having similar content interests, caching this content paves the way for saving on backhaul traffic costs.

Content caching as a use case has different workloads that may run together to perform the caching. Some of them are content caching algorithms, data aggregators, machine learning codes, content traffic analytics, web servers, and more.

3.2.3.1 EDGE ACCELERATED WEB

The edge accelerated web is a use case that allows edge cloud services to be used for every smartphone user. Under most conditions, page load times are dominated by the front-end operations rather than the server in normal networks. The browser performs operations such as content evaluation and rendering. This not only consumes time but also power, which is essential for power-critical end devices such as mobile phones. By performing these operations at the edge, users can experience a quality browsing experience and as well save the battery power on their devices. Operations may include ad- block, rendering, content evaluation, video transcoding and more.

3.2.3.2 HEALTH CARE

As the financial and human costs stemming from non-communicable diseases such as cancer and diabetes continue to rise globally, regular screening methods are important for the world's population. With fewer health specialists and more rural dwellers compared to urban populations in many countries, mobile and edge networks play an important role in mitigating the rise of non-communicable disease (NCD).

Artificial Intelligence (AI) based technology has shown promising success in terms of quality in screening of NCDs/CDs. Many countries have begun adopting AI to augment the abilities of health care specialists in managing large scale deployment of their practices and provide access to these services through AI assisted tele-screening in rural areas (for example, China).

With AI assisted tele-screening, early intervention for a large number of NCDs (and CDs) can be greatly improved versus inflicting great human suffering and straining the resources of any nation. Early screening of NCDs/CDs can help provide better treatment plans and a better quality of life post treatment. These services could be made available at remote locations in rural areas without patients having to travel long distances for screening. These AI assisted disease screening techniques can be based on imaging such as X-ray, Ultrasound, Fundus, Optic Coherence Tomography (OCT), Positron Emission Tomography (PET), Magnetic Resonance Imagery (MRI), Infrared, Computed Tomography (CT) and non-imaging modalities for physiological body parameters such as height, weight, Body Mass Index (BMI), heart rate, Electrocardiogram (ECG), blood pressure, blood sugar, and etcetera.

Another dimension that is steadily gaining prominence is the need to migrate to health care from sick care. This migration requires continuous real-time analytics of at-risk patients who can be monitored for early warning signals. This can be achieved through wearable medical devices that can continuously monitor certain physiological parameters in order to alert healthcare service providers with timely inputs.

Another usage is continuous real-time analytics of senior citizens who can be monitored for early warning signals. These analytics may also include audio-visual data apart from physiological parameters monitored through wearable devices, which may be captured from sick patients who are recovering through in-home or in-hospital care.

3.2.4 SPEECH ANALYTICS AND DERIVED WORKLOADS

The speech analytics landscape is generally comprised of four components: 1) speech recognition, 2) machine translation, 3) text-to-speech and 4) natural language understanding. Major Tier 1 providers include Baidu, Microsoft, Google, Amazon, Apple and IBM – all of whom offer APIs that cover these areas. In addition, many big companies like Google, IBM, Microsoft, and Baidu also offer cloud-based speech analytics solutions.

At the edge, there are also device-level speech analytics solutions like Apple Siri or Amazon Alexa. With Language User Interface (LUI) gaining ground as a more natural way of interfacing with the user, more speech analytics applications such as chatbots will continue to be developed in the foreseeable future.

3.2.5 DATA PROCESSING AT THE EDGE FOR IOT

With IoT devices soon expected to produce trillions of gigabytes of data daily, the Internet of Things is expected to be both the biggest producer and consumer of data. Billions of IoT devices will include components in a variety of uses, including smart city, smart retail, smart vehicles, smart homes, and more.

Edge devices are, in theory, IoT devices – and video analytics and AR/VR will play an important part of the IoT. For example, a face detection workload may be run for a device in a smart city setting, or for checkout in a smart retail shop, or as a part of AR for a private user. IoT workloads will also generally include all the AI workloads in terms of processing a data point.

One specific IoT-related workload is the IoT Gateway. With all the IoT data needing to be processed differently at different latencies for varying purposes, compute capability to process all this data at different locations to fulfill the varying latency requirements is necessary. Thus, the edge cloud is an ideal location for performing the following:

1. Data pre-processing (bidirectional), such as filtering, changing formats, and etcetera
2. Data processing for latency critical use-cases and scenarios with connected components
3. Partial data processing and storage

Data organization and processing will be an important operation at the edge cloud. Fundamentally, data organization entities range widely in complexity, from simple key-value stores that are designed for very fast access to data to complex analytics operations.

3.2.6 VIDEO SURVEILLANCE AND SECURITY APPLICATIONS

Figure 3.2 shows an end-to-end use case providing video surveillance to cities, enterprises or neighborhoods over the network. MEC (Multi-Access Edge Computing) is used for analyzing video streams from nearby surveillance IP cameras to conduct targeted searches in order to detect, recognize, count and track pedestrians, faces, vehicles, license plates, abnormal events / behaviors and other types of content in the video. Analysis and processing happen closer to the point of capture, thereby conserving video transmission bandwidth and reducing the amount of data routed through the core network.

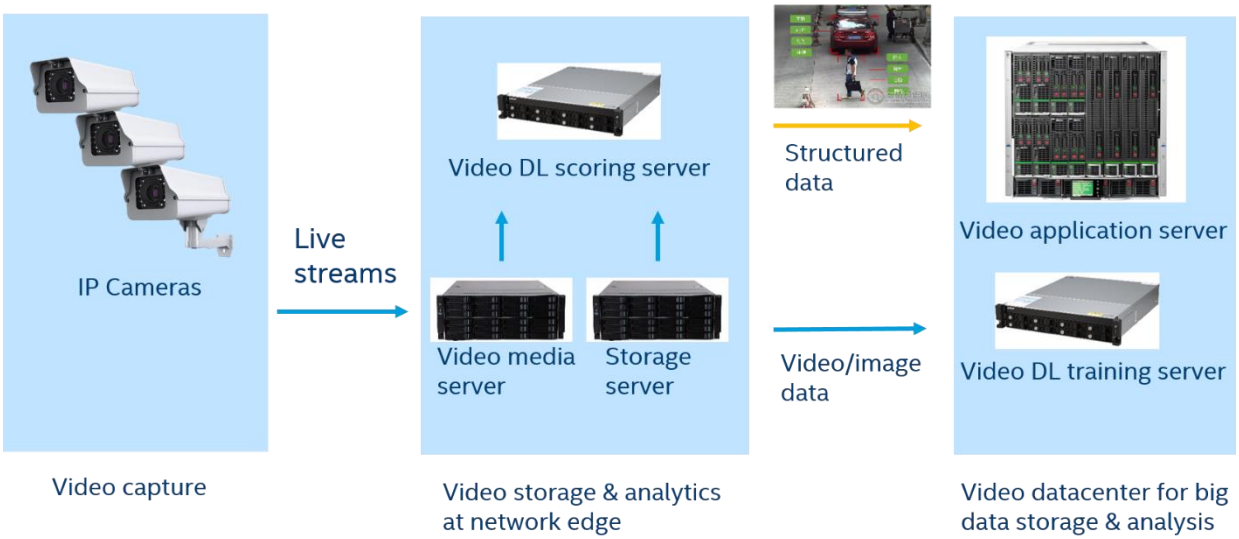

Figure 3.2. Smart City Video Analytics System.

This new application will allow consumers to make payments for retail or entertainment purchases using MEC, 5G solutions and advanced facial recognition technology.

3.2.7 CONNECTING EVENT ATTENDEES TO VIDEO AND VIRTUAL REALITY APPLICATIONS

Virtual reality (VR) video streaming is another compelling MEC use case. People who attend large events, like conventions or sporting events, often struggle with basic access to data services via LTE or 4G networks. If basic data connectivity is a frustration, those mobile customers will be unable to conduct data-intensive tasks, such as video streaming or VR applications. Network operators have proven that MEC can easily support high-quality, data-intensive VR applications involving high resolution and 360-degree video.[2]

3.2.8 REMOTE MONITORING, NETWORK TROUBLESHOOTING AND VIRTUAL MACHINES

Artificial Intelligence is also applied to network operations at the edge. For example, CommSPs are using network analytics to monitor the behavior of virtual machines. When any issues or degradations are detected, network administrators can make quick decisions on how to handle the issue. Software-defined networking allows the distribution of network intelligence to the edge. By being able to detect an issue or

[2] *How 5G is Set to Change Broadcasting and Sports Landscape*, Telecoms Tech News. 27 September 2017.

anomaly and address it quickly, rather than a delayed response in 10 or 20 minutes, the quality of service for the user will be much improved.

4. ROLE OF ARTIFICIAL INTELLIGENCE / MACHINE LEARNING IN NEXT GENERATION EDGE SYSTEMS

The growing role of AI/ML (Artificial Intelligence/ Machine Learning) is an accelerating trend in designing next generation edge networks. Sophisticated AI/ML techniques are being applied to address the complex, stringent and diverse requirements posed by emerging autonomous, immersive, and multi-modal sensory applications.

4.1 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING FOR ENHANCING AND AUTOMATING EDGE SYSTEMS

The application of AI/ML tools is expected to permeate all aspects of edge system design, ranging from AI/ML-enabled Client devices, to the Radio Access Network, to the Cloud infrastructure. While the role of AI/ML in enhancing wireless/edge system design is nascent, a breadth of recent work applies ML techniques to all layers of the protocol stack including the Physical Layer (PHY), Medium Access Control Layer (MAC), Network Layer (NET), and Application Layer (APP) towards: optimizing End-to-End (E2E) delivery of diverse applications, the deployment and management of dense, hierarchical, multi-radio wireless and edge infrastructure, and the dynamic adaptation and orchestration of software-defined cloud and service infrastructure to meet the stringent requirements of advanced applications.[3,4,5]

There is also a growing appreciation that an edge network should be designed as an autonomous and intelligent system. That system should be able to sense its environmental and application context, and have the ability to interpret and act on the contextual information in real-time.[6] The concept of an autonomous network mandates pervasive, distributed intelligence and AI/ML capabilities across the entire network, in order to drive real-time context prediction and responsive network adaptation. Additionally, AI/ML techniques are also expected to be critical in off-line network analytics, capacity planning, network/service orchestration, and subscriber management, as well as for training a bulk of AI/ML services deployed at the edge/clients.

Figure 4.1 is illustrative of the breadth of AI/ML applications across the E2E mobile edge network, which promise increased automation in designing, deploying and maintaining next generation networks and service delivery.

[3] *Machine Learning for Communications,* IEEE Communications Society, Research Library.
[4] ITU-T Focus group on Machine Learning for Future Networks including 5G (FG-ML5G), ML5G-I-118-R8.
[5] *Evolution to an Artificial Intelligence Enabled Network*, ATIS. September 2018.
[6] *Self-X Design of Wireless Networks: Exploiting Artificial Intelligence and Guided Learning*, Erma Perenda, Samurdhi Karunaratne, Ramy Atawia, Haris Gacanin. CoRR abs/1805.06247. 2018.

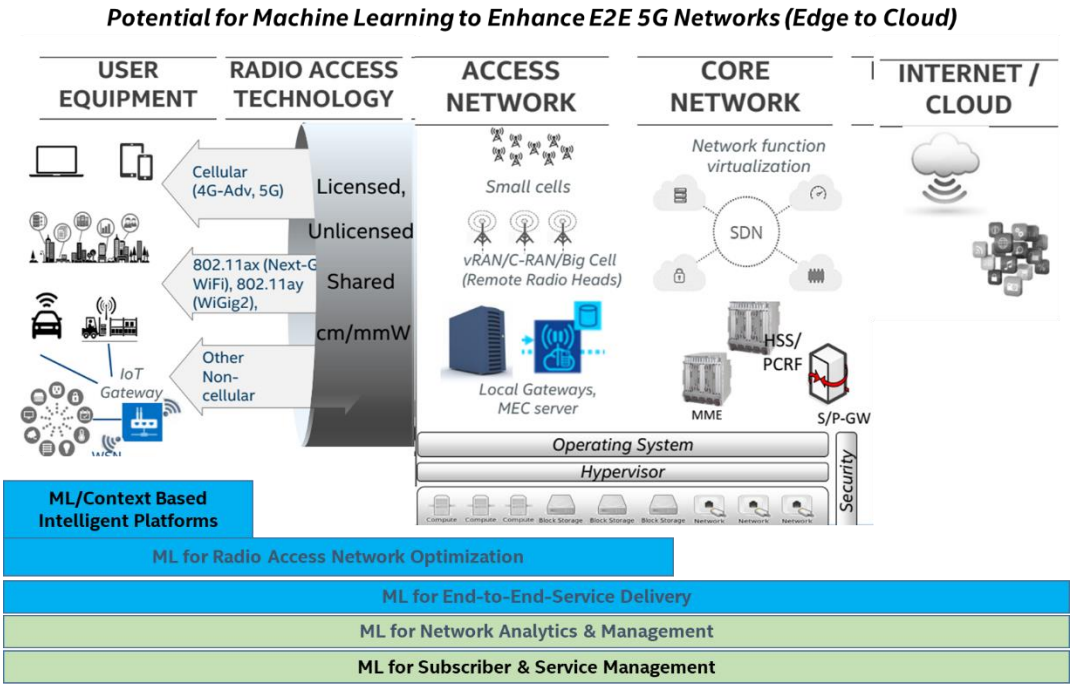

Figure 4.1. ML and AI Based Techniques Permeate all Aspects of E2E Wireless System Design, Service Management and Delivery.

The following are some key application areas of AI/ML in edge networking.

- **ML/Context Based Intelligent Platforms.** ML for modem design, MI (Machine Intelligence) techniques for context prediction and context aware optimization of platform performance
- **Radio Access Network Optimizations:** Includes novel approaches to air interface design at the physical layer, MAC/Network layer radio resource management, interference management, network load balancing, mobility management, network routing and etcetera
- **E2E Application Delivery:** Example areas include service delivery optimizations to meet the stringent, application-specific Quality of Experience (QoE) requirements of immersive media, autonomous systems and tactile Internet applications
- **Network Analytics and Management:** Techniques for traffic/capacity analysis, detection of anomalous network conditions, self-healing networking, and enhancement of E2E network management and operation, and etcetera
- **Subscriber & Service Management:** Encompasses predictive analytics for service differentiation, analytics for subscriber behavior, customer support and etcetera

The application of AI/ML frameworks for edge networking promises flexibility, scalability, software/hardware reuse and automated system design. However, several challenges remain when applying AI frameworks for edge systems comprising wireless networks. Foremost is the ability of AI/ML solutions to develop predictive solutions and adapt in real-time to fast changing dynamics of the wireless edge, typically with limited amount of data.

4.2 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING AS A WORKLOAD FOR NEXT GENERATION EDGE NETWORKS

The use of AI/ML techniques can enhance edge network performance, making them more capable of managing emergent data-intensive, real-time autonomous and immersive applications. There is a significant opportunity to leverage the ubiquitously available compute and storage resources of wireless edge networks for localized, real-time, and on-demand distributed learning.

AI/ML-based processing of sensory information collected in the network can be used for deriving contextual information that is important to local processing of AI/ML workloads, such as those required for image processing, object detection and classification, localization, and etcetera.

However, so far, a significant focus of edge computing for ML has been on local inference or offload of inference tasks from the client to the edge to the cloud, while the tasks of training learning models have largely been executed within the cloud. Training data used for AI/ML, including video images, health related measurements, traffic/crowd statistics, and more, is currently typically located at wireless edge devices. Transferring these local data sets to the central cloud for creating ML models incurs significant communications cost, processing delays and privacy concerns.

As distributed AI/ML over wireless networks gains in popularity, there will be broad motivation to use the wireless edge as an integrated compute-communication engine for distributed learning. Given the ubiquity of wireless infrastructure, one can envisage an increasing trend to use the wireless edge for ML training, going beyond the task of ML inference and support of compute offload.

The trend towards edge based training emerges in the Federated Learning model,[7,8] where a central server orchestrates local ML training across a large number of clients and then aggregates the local models to develop a more sophisticated global model without requiring the clients to share their private data.

It is expected that the need for local training will continue to accelerate as learning models shift towards adaptive, online and real-time learning, to address the need of emerging real-time services. Figure 4.2 illustrates the potential for distributing AI/ML workloads across wireless edge networks, which includes client-based AI training and inference.

[7] *Federated Learning: Strategies for Improving Communication Efficiency*, Jakub Kone, H. Brendan McMahan, Felix X. Yu, Ananda Theertha Suresh & Dave Bacon, Google. 2017.

[8] *Communication-Efficient Learning of Deep Networks from Decentralized Data*, H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas. 2017.

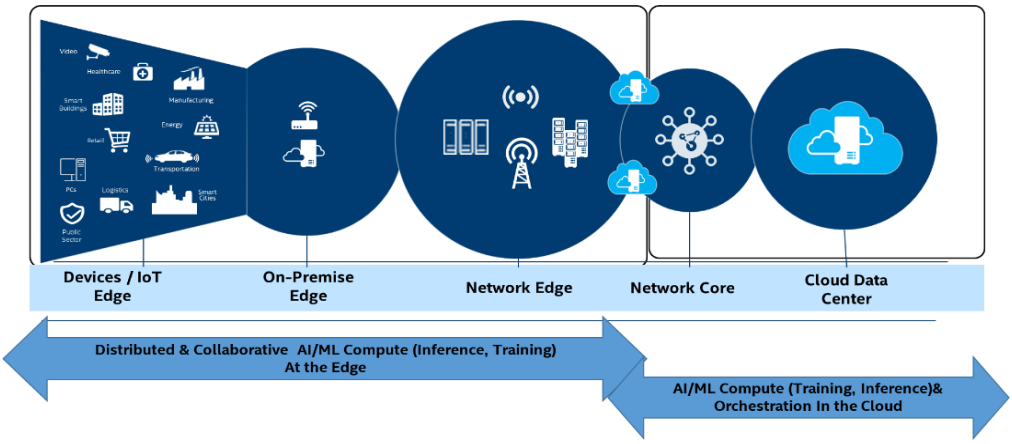

Figure 4.2. Edge Computing Extends to AI/ML Across the Entire Edge Network Assisted by Training and Orchestration Via the Cloud.

Limitations of wireless edge networks and clients must be addressed to fully enable pervasive AI/ML (training and inference) at the edge. Unreliable and dynamically changing wireless connectivity, coupled with mobility of edge devices, creates challenges in distributing ML workloads to the cloud or to other edge devices. Additionally, wireless nodes have heterogeneous sensing, thus compute, storage capabilities and data generated across different devices is highly skewed, and only reflective of partial observations. This makes it difficult for learning models to converge to the true model based on just local data sets, when performing ML training at the edge.

Therefore, there is a need for devices to develop the necessary techniques to collaborate or leverage network assistance to enable reliable learning while comprehending the limitations of wireless communication.

4.3 IMPLICATIONS FOR NEXT GENERATION EDGE ARCHITECTURE AND DESIGN

AI and ML approaches will play an important role in enhancing and automating next generation wireless edge networks. They will also enable a ubiquitously available edge learning engine to facilitate the real-time learning required for emerging autonomous/immersive services.

Enabling reliable and real-time learning over wireless edge networks, will require a cross-disciplinary approach, capable of understanding the fundamental theory of AI/ML techniques, adapting ML approaches for wireless applications and comprehending the uncertain, dynamic nature of learning over wireless channels. Also of importance are understanding the sensing and storage limitations and their impact on available data sets, as well as addressing compute and privacy concerns in moving the compute resources and data sets across the network. We expect that synergistic and integrated design of wireless networking with edge AI/ML will be key to addressing these challenges.

5. 5G ARCHITECTURE, CURRENT STATE ANALYSIS

3GPP is currently focused on specification of fifth generation of wireless systems, 5G, which builds on ITU IMT-2020 requirements. Phase 1 of 5G specifications which targeted early 5G deployments with limited capabilities have been published as part of 3GPP Release 15. Release 15 compliant 5G systems have started to deploy as of 2019 with broad scale rollout expected in 2020 timeframe.

5.1 3GPP SPLIT RAN/O-RAN ARCHITECTURE

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 5.1 shows an estimated timeline.

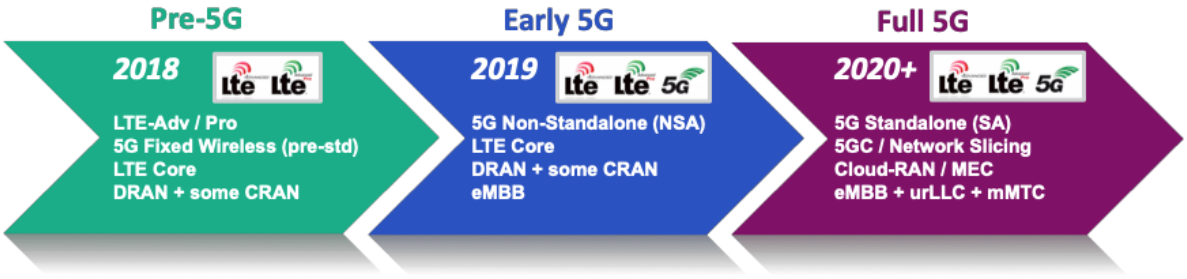

Figure 5.1. Phased Deployments of 5G.

The deployment of a 5GC will enable new applications, use cases, and features defined in the corresponding 3GPP specifications. These deployments will be based on the late drop of Release 15 as well as Release 16.[9]

The 5GC deployment will enable new capabilities such as 3GPP Network Slicing. 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.[10] This translates to the virtualization and disaggregation of many of the RAN and mobile core functions. Therefore, we will see increasing use of a Cloud-RAN approach which has significant architectural implications for the network architecture, such as:

- Open architecture with virtualization of functions instead of utilizing expensive proprietary hardware
- Mobile transport network capacity increases while reducing price per unit bandwidth (Moore's Law-like economics)
    - Accomplished by leveraging best-of-breed, high volume networking silicon
- Increased network automation and intelligence

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 5.2.

[9] 3GPP specifications.
[10] *NGMN 5G Whitepaper*, February 2015.

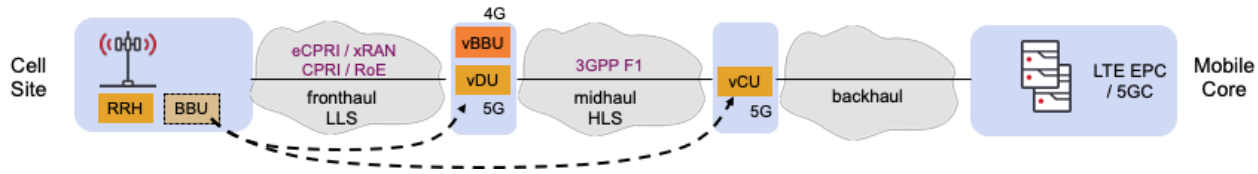

Figure 5.2. Split RAN Architecture.

The O-RAN Alliance is an operator-led industry group that aims to develop an Open RAN architecture building on the 3GPP Split RAN architecture, with two key objectives:

- Openness: specifying open interfaces and additionally disaggregated architecture that will allow operators to introduce additional vendors and best-of-breed systems into their RAN networks to address their specific needs and reduce their expenses
- Intelligence: defining capability in the architecture that allows for the introduction of artificial intelligence/machine learning to improve the automation of their networks to support increased scale, complexity and flexibility for highly dynamic 5G services.

The O-RAN high level architecture is shown in Figure 5.3.

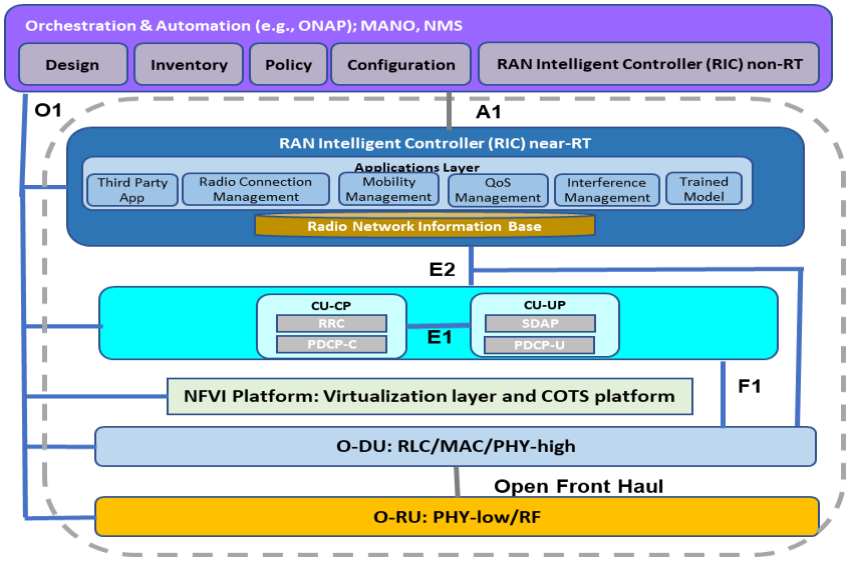

Figure 5.3. O-RAN Architecture.

O-RAN architecture includes some key features, which include:

- Separation of the O-RAN Radio Unit (O-RU) and the O-RAN Distributed Unit (O-DU). This allows for the introduction of separate RU and DU components with a standardized Open Fronthaul interface connecting them for multi-vendor interoperability

- Introduction of cloudification to support the replacement of proprietary hardware with virtualized network functions for the non-RU components of the architecture. This reduces costs and allows agile deployment of processing and storage capacity based on traffic demands
    - Includes virtualization of the O-CU protocol stack supporting 4G, 5G and other protocol processing, allowing distribution of processing capacity across multiple platforms and

responsiveness to control commands issued by the Near-Real Time (RT) RAN Intelligence Controller (RIC)

- Introduction of Near-RT and Non-RT RICs. This further decouples Control and User-Planes to allow greater scalability of control functions. Additionally, introduction of advanced artificial intelligence/machine learning technology for faster response and improved optimization of resources using closed loop automation
    - Non-Real Time RIC functions include service and policy management, analytics and model training for AI/ML models to be deployed on the Near-RT RIC, using the A1 interface
    - The Near-Real Time RIC provides next generation radio resource management functions, leveraging embedded intelligence but also providing a platform for on-boarding of third-party control applications

- Introduction of standard interfaces to support multi-vendor interoperability including: the Open Fronthaul interface between O-RU and O-DU; the E2 interface between the Near-RT RIC and the O-DU and O-CU for streaming of measurements and configuration commands; the A1 interface between Non- and Near-RT RIC; and the O1 interface for standardized management of all O-RAN components

5.2 5G TRANSPORT

There are various possibilities of RAN functional split options that may result from the 3GPP Split RAN architecture. However, one can generalize these into a Lower-Layer Split (LLS) and a Higher-Layer Split (HLS). This results in new partitions of the mobile transport network, fronthaul and midhaul, which have their own unique transport characteristics and requirements.

The fronthaul network between the RRH (Remote Radio Head) and BBU/DU (Baseband Unit / Distributed Unit) carries extremely latency-sensitive Common Public Radio Interface (CPRI) radio control traffic for LTE radios and eCPRI or O-RAN Fronthaul radio control messages for 5G NR and ng-LTE radios. The CPRI fronthaul traffic is a serialized constant bitstream technology that is traditionally carried over dedicated fiber or Wavelength Division Multiplexing (WDM) technology.

The eCPRI specification was published to define the message structure for eCPRI and the transport layer of carrying the eCPRI data streams. However, the specification did not completely standardize the radio control messages within the eCPRI application layer (known as the eCPRI protocol layer in the eCPRI specifications) that is required to ensure a fully open and interoperable implementation between the Remote Radio Head (RRH)/ Radio Unit (RU) and Distributed Unit (DU).

The O-RAN Alliance has a number of working groups, including WG4 that has published version 1.0 of their open fronthaul specification in March of 2019, which significantly leverages the previously published xRAN fronthaul specification.

The midhaul network carries traffic between the DU and CU and has tighter latency requirements such as ~1-5 milliseconds (ms) than backhaul traffic (<20 ms) but not nearly as stringent as fronthaul (150-200 μs). This is accomplished over the 3GPP standardized F1 interface which utilizes standards-based Ethernet & IP encapsulation for the transport layer.

Figure 5.4 summarizes the different transport network requirements between fronthaul, midhaul and backhaul. While the midhaul network has a somewhat smaller latency budget than backhaul, the required

transport network architecture and technologies are very similar as shown in Table 5.1. In fact, in many cases, midhaul and backhaul traffic are expected to combine in many portions of the network given the different deployment scenarios that operators will employ even within a given metro (for example, combinations of Distributed-RAN (D-RAN) and Centralized-RAN/Cloud-RAN (C-RAN/Cloud-RAN).

Table 5.1. Mobile Transport Network Characteristics by Network Segment.

	RU / Fronthaul	DU / Midhaul (PDCP/RLC)	CU / Backhaul (CU with SDAP)	5GC
Bandwidth	4G: 2.5G (CPRI 3) to 10G (CPRI 7/7a) 5G: N x 10GE / 25GE / 50GE	10GE / 25GE / 50GE	10GE / 25GE / 50GE / 100GE	
Latency (Round-Trip)	4G: 150us ~ 200us (Bounded) 5G eMBB: 150us or Less (Bounded)	1ms ~ 5ms (Bounded)	Less than 20ms	
Radio Protocol(s) Processing	O-RAN, eCPRI, CPRI CPRI with 1914.3 RoE	Not Required (Transport is IP/Ethernet)	Not Required (Transport is IP/Ethernet)	
Statistical Multiplexing	CPRI / RoE Structure Agnostic: No xRAN, eCPRI, RoE FDM: Yes (Marginal)	Yes	Yes	
Network Slicing	Not Required	Yes Criteria: Based on S-NSSAI (Single – Network Slice Selection Assistance Information), QoS Flow Indicator, QoS Flow Level Parameters, DRB ID (Data Radio Bearer ID), etc IEs	Yes Criteria: Based on NSI (Network Slice Instance), IMEI, IMSI, IMEI/IMSI Ranges, PLMN-ID, etc IEs	
Reach	Less than 10km	Less than 20km	Less than 100km	
Packet Timing / Sync	4G & 5G: 1ns PTP Timestamp Accuracy	4G LTE-A Pro: 15ns ~ 20ns PTP Timestamp Accuracy 5G: 1ns PTP Timestamp Accuracy	4G LTE-A Pro: 15ns ~ 20ns PTP Timestamp Accuracy 5G: 1ns PTP Timestamp Accuracy	
Topology	Hub & Spoke, Ring	Hub & Spoke, Mesh, Ring	Hub & Spoke, Mesh, Ring	
Transport Technologies	L1: P2P Fiber, Packet Optical (Flex-E/G.mtn) L2: Ethernet / TSN L3: IP/Ethernet (RU Remote Mgmt. Only)	L1: Optical, Packet Optical (Flex-E/G.mtn) L2: Ethernet / TSN L3: IP/MPLS, EVPN, Segment Routing	L1: Optical, Packet Optical (Flex-E/G.mtn) L2: Ethernet / TSN L3: IP/MPLS, EVPN, Segment Routing	
OAM	CPRI L1 & L2 OAM, 1914.3 RoE OAM (round trip delay, etc.)	802.1ag CFM, Y.1731, TWAMP, RFC 2544/Y.1564, 802.3ag EFM; VCCV BFD; G.mtn OAM (in progress)	802.1ag CFM, Y.1731, TWAMP, RFC 2544/Y.1564, 802.3ag EFM; VCCV BFD; G.mtn OAM (in progress)	

5.3 NETWORK TRANSPORT TO SUPPORT THE 5G TARGET ARCHITECTURE

Delivering on 5G Transport will require architecture changes and new technology innovations. These mobile network improvements will include:

- Packetized and Deterministic xHaul
- Transport networks resources instantiated as part of 3GPP Network Slice Instance (NSI)
- Higher speed interfaces and packet-optical integration
  - This will include increasing use of 25 Gigabit Ethernet (GE), 50 GE, 100 GE interfaces in fronthaul/midhaul networks and 400 GE in midhaul/backhaul along with corresponding mapping into the photonic layer via coherent modem technology
- Time/Phase and Frequency Synchronization
  - As with more advanced LTE features such as Coordinated Multi-Point (CoMP), 5G will require time/phase synchronization in addition to frequency synchronization. It will also require even more stringent timing precision for features such as MIMO transmission diversity. However, deeper discussion of synchronization is beyond the scope of this paper

5.3.1 PACKETIZED AND DETERMINISTIC XHAUL

The bandwidth growth from LTE and 5G radios requires that fronthaul traffic be packetized to support the scale required.

For LTE, CPRI traffic is packetized via IEEE 1914.3 RoE (Radio over Ethernet) technology which supports Structure Agnostic and Structure Aware mapping modes. Structure Aware mapping can reduce the required fronthaul bandwidth to some degree (from 100 percent of CPRI bandwidth to ~85 percent) by discarding non-utilized / idle portions of CPRI traffic.

However, more significant gains are achieved by performing CPRI Layer 1 offload which corresponds to the Low-Phy portion of the RAN functional split. This achieves a much greater bandwidth reduction (to ~20 percent of CPRI bandwidth) for fronthaul traffic. This is accomplished by implementing Intra-PHY functional split which is the adopted split option defined in eCPRI and xRAN/O-RAN fronthaul specifications (see Figure 5.4).

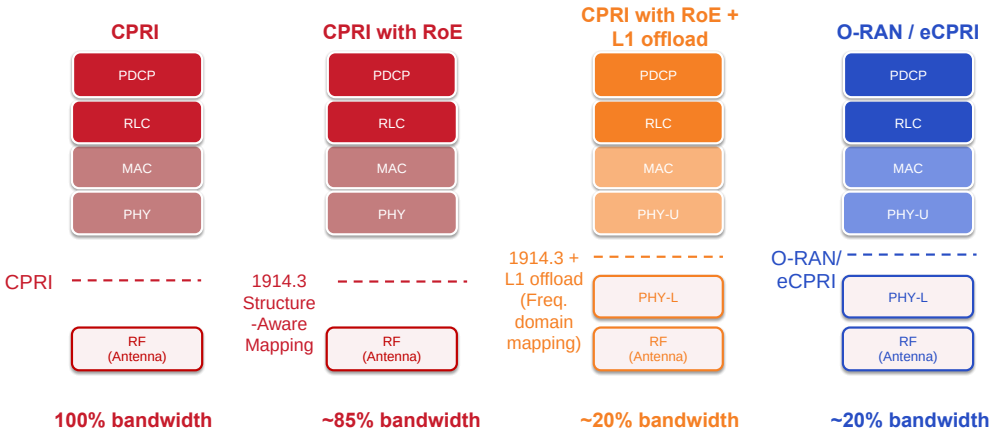

Figure 5.4. Fronthaul Bandwidth Reduction.

There are two technologies in the fronthaul that can be applied to mitigate the latency impacts on CPRI/RoE when combined with other traffic.

One of these mechanisms is FlexE (Flexible Ethernet) which supports channelization as one of its use cases. This means that FlexE can, for example, partition a 50 GE interface into 35 Gbps + 15 Gbps channels, wherein each of these channels is scheduled in a Time Division Multiplexing-like (TDM) fashion. By mapping CPRI/RoE into one of these channels with dedicated TDM-like scheduling, its latency and jitter will not be impacted by traffic in the other channel and bounded low-latency delivery can be ensured.

The other technology is Time Sensitive Networking (TSN), and specifically its ability to provide Time Aware Scheduling (standardized in 802.1Qbv) with Frame Pre-emption (standardized in 802.1Qbu). 802.1Qbv compliant Ethernet switches have a time gate control logic associated with all 8 of the Ethernet queues and whereby the gate opening time and closing time for frame transmission can be programmed in nano-seconds granularity.

Frame pre-emption works by fragmenting lower priority frames (such as non-Fronthaul traffic) in order to immediately service CPRI/RoE frames without incurring further delay. Combining 802.1Qbv and 802.1Qbu

ensures that high priority frames assigned to a queue always have bounded latency and jitter performance regardless of the packet sizes of the low priority frames.

Fronthaul traffic may be combined with backhaul traffic and even midhaul traffic. This means that a macro cell-site platform must support backhaul transport technologies as well as fronthaul. These backhaul technologies typically include L2 and L3 VPNs running over MPLS. Today, these MPLS backhaul networks are based on protocols such as Label Distribution Protocol (LDP) or Resource Reservation Protocol – Traffic Engineering (RSVP-TE), but increasingly architectures are evolving to Segment Routing MPLS as an SDN-driven packet underlay.

5.4 5G NETWORK SLICING

Once the 5GC is deployed, Mobile Network Operators (MNO) will be able to leverage this new capability defined in the 3GPP standards. 5G is intended to support a wide range of applications and business needs, each with their respective performance, scale, reliability requirements. 3GPP Network Slicing was defined in the standards to add the flexibility and scale to efficiently support this more diverse set of requirements, concurrently over the same infrastructure. While there is no de facto list of target use cases for 3GPP Network Slicing, and there are many different viewpoints and candidate applications being discussed in the industry.

When implementing 3GPP Network Slicing in the mobile network, it will be important to take a comprehensive approach. This approach must include the orchestration and provisioning of slices as well as how they are implemented into the network layer via both soft and hard slicing mechanisms. An overview diagram of this comprehensive approach and possible slicing technologies is shown in Figure 5.5.

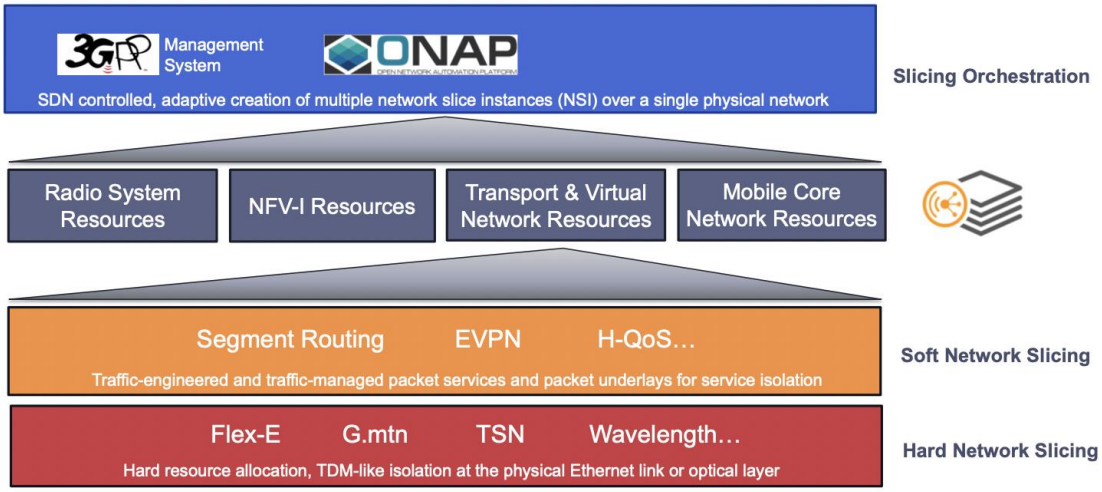

Figure 5.5. Holistic View of Network Slicing.

3GPP Network slicing will span from the radio system (where spectrum will be sliced), to the transport network, to the mobile core. These slices will need to be coordinated end-to-end across these resources. In the network layer, the use of hard or soft slicing will depend on the requirements and application of the slice user. Soft slicing can be used to provide traffic-engineered and traffic-managed isolation of resources.

Technologies such as Segment Routing Multi-Protocol Label Switching (SR-MPLS) can be utilized to provide multiple SDN-controlled traffic engineered paths or Label Switched Paths (LSP) representing different slices, with policies at ingress to map traffic into the appropriate path or slice. The Segment Routing paths or tunnels can be established based on various constraints/parameters and policies such as bandwidth, latency, resiliency requirements, transport / peering costs, and more. The tunnels can be mapped in the xHaul transport platform to specific QoS treatments. This should include dedicated queuing and scheduling resources with reserved buffer allocation to provide resource partitioning when slices are sharing ports.

It is important that the packet network and technologies such as Segment Routing be implemented in an adaptive and dynamic way. As slices may be sold based on achieving a premium SLA, it is important to leverage telemetry and analytics to monitor the network given that conditions may vary over time requiring adaptive changes to ensure the slice SLAs are met on an ongoing basis. Further, some MNOs want Network Slices themselves to be dynamically provisioned and/or removed based on predefined policies like subscription durations or on-demand provisioning.

SDN control is an important component for achieving this dynamic behavior which is a key reason why Segment Routing is often cited as the packet underlay technology of choice for 5G.

Hard slicing delivers strict isolation of resources without relying on statistical multiplexing mechanisms or virtualized partitioning. From an Open Systems Interconnection (OSI) model perspective, hard slicing is thought of as being implemented at Layer 1. Hard slicing can be applied, for example, where the end user would otherwise be served by a “private network” build.

Since the xHaul network will have packet and optical transport technology, one option is to dedicate wavelengths to slices. With most modern optical networks supporting flexible grid Dense Wavelength Division Multiplexing (DWDM) technology and programmable modulation of coherent modems, the wavelengths can be “right-sized” depending on the requirements of that slice. Where the packet/IP network infrastructure connects into the optical layer, technologies such as FlexE can be utilized to provide hard (TDM-like) channelization, sub-rating, or bonding of the Ethernet PHY, and this can be applied to match the wavelength bandwidth.

There are also new technologies such as G.mtn, defined in International Telecommunication Union (ITU) study group 15, which extends the FlexE channel construct end-to-end across the network through a new Slicing Channel Layer (SCL). The standardization process for G.mtn began in late 2018 and is expected to be completed during 2019, with some component vendors in the industry already announcing support of this technology. With SCL, the Flex-E channels can now be cross connected at intermediate nodes at the lowest possible latency since this cross connection is not based on a full Ethernet frame or a full MPLS frame for its switching intelligence.

6. EDGE ARCHITECTURES - CURRENT STATE ANALYSIS

A broad set of transformations are taking place at the edge as part of the re-architecting required for 5G. There are business transformations associated with the monetization of services, including over-the-top (OTT) services and the desire to achieve faster time-to-market (TTM.) There are also technical transformations, which include a variety of areas such as quality of service (QoE), ultra-low latency techniques, the integrations of SDN, NFV and an open edge cloud, data collection and analysis leveraging edge analytics, virtualization of 5G components and automation.

The architecture is transforming to include 5G Novel Radio Multiservice Adaptive NORMA[11]-like network architecture cloud concepts, ECOMP (Enhanced Control, Orchestration, Management and Policy) and a flexible architecture composed of RAN, core, Content Distribution Network (CDN), application delivery, automation and IoT. The final transformation is happening in the industrial sector, where IT (Information Technology) intersects with OT (Operational Technology.) This encompasses Information, Communication Technology & Electronic (ICT&E), SCADA (Supervisory Control and Data Acquisition) systems, ICS (Industrial Control Systems) and IoT, all needing low latency and high security.

To efficiently and effectively deploy 5G network supporting ultra-low latency and high bandwidth mobile network, a variety of applications and workloads at the edge and close to the mobile end user devices (UE or IoT) must be deployed. That would include various virtualized RAN and core network elements, content and various applications as previously described.

Near-real time network optimization and customer experience / UE performance enhancement applications at the edge might also be deployed. Edge cloud must support the deployment of third-party applications, for example, value-added optional services, marketing, advertising, and etcetera. Mechanisms to collect process, summarize, anonymize, etcetera, the real time radio network information (for example, geo-location data) and to and make this available to third-party applications will be deployed either at the edge or a central location or outside the service provider environment. Edge data collection could also be used for training machine learning models and fully trained models can be deployed at the edge to support network optimization.

Although Cloud, Fog and Edge may appear similar, they exist in different parts of the architecture and therefore perform slightly different functions.

- **Cloud**: The collection of connect, compute and storage pervasive throughout a network. Generally speaking, large portions of the cloud are implemented in large, centralized data centers

- **Fog**: A Fog environment places intelligence at the local area network, closer to the creation of the data than the Cloud. Data is then transmitted from endpoints to a gateway, where it is then transmitted for processing (therefore, to the Cloud) and then returned

- **Edge**: The Edge is the closest point to the creation of the data, and it is characterized by intelligence and data processing embedded within an edge device (such as an IoT device or embedded automation controller)

[11] *NORMA: Novel Radio Multiservice adaptive network Architecture*, a project is committed to design a network architecture that enables new business opportunities.

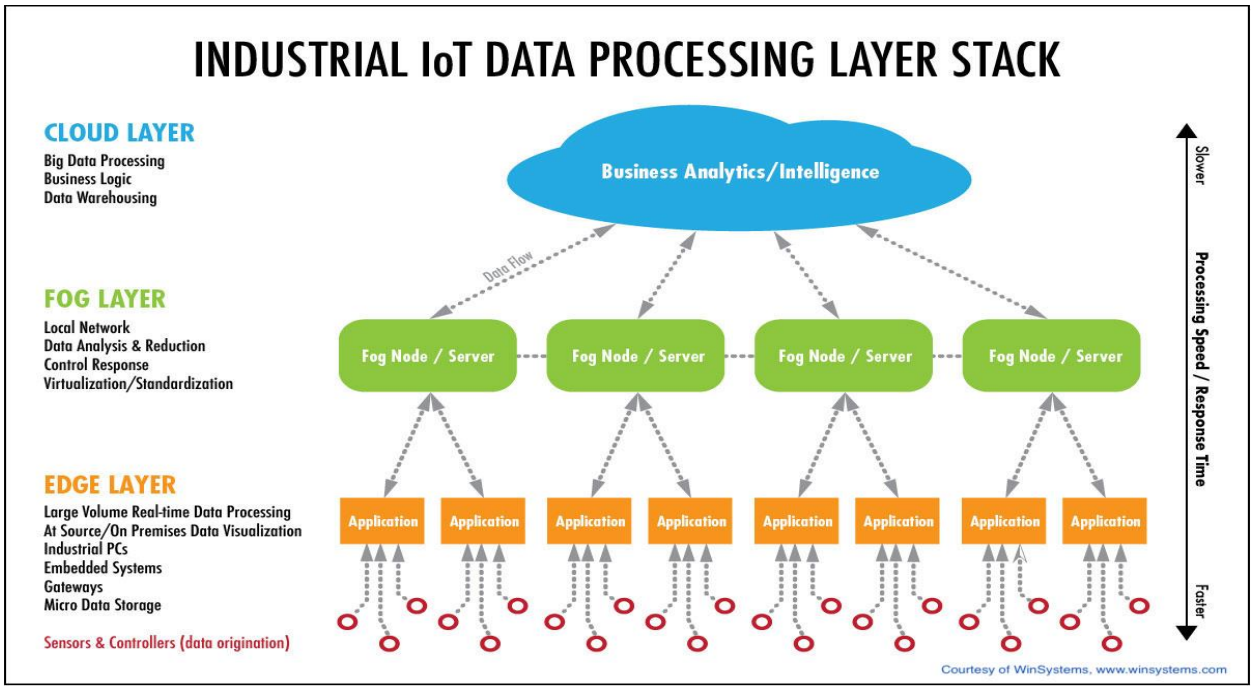

Figure 6.1. Cloud, Fog, Edge illustrated.

6.1 CURRENT INDUSTRY INITIATIVES

The Edge is characterized by a variety of requirements from fixed and mobile access networks, driven by different themes and use-cases, deployment scenarios, and business models. Two key technology paradigms overlap these edge requirements—5G and Cloud Computing. Access, Cloud and IoT market segments are converging with compute and intelligence capabilities residing at several locations: on a mobile user device (for example, a vehicle or handset); located in a home (for example, a home automation appliance); or an enterprise (for example, a local service network); or positioned in the network at a cell tower, or a Central Office.

As it results from the amalgamation of different market segments, this technology suite is currently being referred to by different names, as defined by the contributing market segment, for example, the Telco industry has landed on the term Multi-Access Edge (MEC), whereas IoT professionals call it Open Fog.

Several industry initiatives, currently underway, are exploring and/or developing different facets of this technology suite as driven by the specific needs of the originating market segment. ETSI NFV has a MEC specific workgroup but other SDOs that work on the mobile system architecture such as 3GPP and the O-RAN alliance do not specify architectures that are edge-specific. There are also several open source projects that draw from cloud concepts to apply to the Edge. Notable amongst them are Linux Foundation Edge and OpenStack Edge initiatives. All of this has necessitated adapting and/or creating edge-focused requirements and open source projects in order to create a guard-railed field to innovate. The Standards Development Organizations activities are described in section 9.2.

One of the key elements of 5G technology evolution is the role being played by open source initiatives in combination with Standards Development Organizations (SDO). The foremost drivers for the use of open source are:

- Promoting a multi-vendor ecosystem by lowering the interoperability barriers

- Increasing the innovation velocity by leveraging open source software
- Realizing cloud scale economics through virtualization

6.1.1 OPEN SOURCE INITIATIVES

"Open source" is an umbrella term that is used to capture multiple initiatives with different objectives. At a high level, open source initiatives fall into three categories.

- **Open Interfaces**: Driven by the need to promote multi-vendor interoperability and introduce competition towards best-of-breed procurement. SDOs define and publish architectures, interfaces, protocols and information models. Examples include new interfaces such as A1, O1, E2, Open eCPRI specified under the O-RAN Alliance, and ETSI NFV architecture for Edge

- **Open Source Software and Open APIs**: Open sourcing infrastructure (common/platform) software is motivated by the need to cut down software development time and facilitate faster innovation. Open APIs allow context exposure from network functions to third party applications. Examples include various projects under Linux Foundation Edge

- **Open Reference Design**: By publishing COTS hardware reference specifications and blueprints for deployment scenarios and use cases, open reference designs (under initiatives such as Akraino) accelerate the virtualization of edge network functions towards a cloud native, service-agile operational environment

OpenStack: OpenStack has established itself as the de-facto tool for open-source cloud infrastructure control, or control of large pools of compute, storage and networking resources. It was originally designed to manage centralized hyperscale datacenters, but has since evolved to control heterogenous infrastructures including the network edge. A collaboration with ETSI MEC and Office of Emergency Communications (OEC) is currently underway for edge-optimized implementations of OpenStack better suited for distributed computing.

OpenStack Foundation- Edge Computing Group (OSF- Edge): OSF Edge's objective is to define infrastructure systems needed to support applications distributed over a broad geographic area, with potentially thousands of sites, located as close as possible to discrete data sources, physical elements or end users. The working group will identify use cases, develop requirements, and produce viable architecture options and tests for evaluating new and existing solutions across different industries and global constituencies.

Linux Foundation Edge (LF Edge): LF Edge is an umbrella organization that aims to establish an open, interoperable framework for edge computing independent of hardware, silicon, cloud, or operating system. LF Edge is initially compromised of five projects that will support emerging edge applications in the area of non-traditional video and connected things that require lower latency, faster processing and mobility: Akraino Edge Stack, EdgeX Foundry, Open Glossary of Edge Computing, Home Edge Project and Project EVE (Edge Virtualization Engine).

Akraino: This open-source initiative aims to create open-source software for an edge stack, designed for carrier-scale edge computing applications running on virtual machines and containers to support reliability and performance requirements.

EdgeX Foundry: EdgeX Foundry is a vendor-neutral open source project hosted by the Linux Foundation to build a common open framework for IoT edge computing. It is composed of an interoperability framework

hosted within a full hardware- and Operating System (OS)-agnostic reference software platform. The reference platform is designed to enable an ecosystem of plug-and-play components for the deployment of IoT solutions. The EdgeX Foundry is an open platform for developers to build custom IoT solutions. They may do so by either feeding in data from their own devices and sensors, or consuming and processing the data produced.

Edge Virtualization Engine (Project EVE): Project EVE develops the open source Edge Virtualization Engine (EVE) for deployments on bare metal device hardware. It also provides system and orchestration services and a container runtime for platform consistency. EVE allows cloud-native development practices in IoT and edge applications.

Open Glossary of Edge Computing: This project provides a concise collection of terms related to the field of edge computing. It aims to improve communication and accelerates innovation through a shared vocabulary.

Home Edge: Aims to enable an open source residential edge computing framework, platform and ecosystem**.**

Open Network Automation Platform (ONAP): ONAP is an open-source initiative that provides a platform for real-time, policy-driven orchestration and automation of physical and virtual network functions. 5G and Edge Automation working committees are working to evolve ONAP for programmable control of edge compute.

Open Platform for NFV (OPNFV): OPNFV is a collaborative open-source project under the Linux Foundation with the goal of defining a common open source NFV platform. OPNFV works closely with ETSI to promote consistent open NFV standards.

Cloud Native Computing Foundation (CNCF): CNCF uses an open source software stack to deploy applications as microservices, packaging each part into its own container, and dynamically orchestrating those containers to optimize resource utilization. It contains several open source projects such as Kubernetes, Prometheus, and etcetera.

Acumos: Acumos, under the Linux Foundation, provides an open source framework to build, share, and deploy AI apps. Acumos standardizes the infrastructure stack and components required to run an out-of-the-box general AI environment.

Open Networking Foundation (ONF) Central Office Re-architected as a Datacenter (CORD): The CORD platform leverages SDN, NFV and Cloud technologies to build agile datacenters for the network edge. One key goal for this initiative is an open reference implementation to create a complete operational edge datacenter with built-in service capabilities and commodity hardware using the latest in cloud-native design principles. CORD originally targeted Residential Access, Mobile Access and Enterprise user segments, however, it is now evolving toward Multi-Access Edge.

ONF's Converged Multi-Access and Core (COMAC): COMAC is an open-source project to bring convergence to Operators' mobile and broadband access and core networks. By leveraging and unifying both access and core projects, COMAC will enable greater infrastructure efficiencies as well as common subscriber authentication and service delivery capabilities so users can roam seamlessly between mobile and fixed environments while experiencing a unified experience. COMAC is composed of the following component projects: SDN-Enabled Broadband Access (SEBA), Virtual Optical Line Termination Hardware Abstraction (VOLTHA), Radio-Central Office Re-architected as a Datacenter (R-CORD), Trellis and Open Mobile Evolved Core (OMEC).

Open19 Foundation: The Open19 Project is an open platform for any 19" rack environment for servers, storage and networking. Its goal is to build a community for a new generation of open data centers and edge solutions. It is an industry specification that defines a cross-industry common server form factor, creating a flexible and economic data center and edge solution for operators of all sizes.

6.1.2 EDGE COLLABORATIVE CONSORTIA

Interest in the Edge has given rise to multiple consortia with a focus on specific issues to deployment.

Telecom Infra Project (TIP): TIP aims at the democratization of the telecommunications infrastructure. Many groups exist under TIP, including an Edge Computing project group which focuses on lab and field implementations for services/applications at the network (wireless and fixed) edge. The project group will leverage architectures and standards (for example, ETSI MEC, CORD, and etcetera) and will focus on implementations, each driven by use-case needs.

Automotive Edge Computing Consortium (AECC): AECC drives the evolution of edge network architectures and computing infrastructures to support high volume data services for an efficient connected-vehicle future. Its strategic goal is to support large automotive data sets between vehicles and the cloud by using edge computing and efficient system design.

Industrial Internet Consortium (IIC) and OpenFog: IIC and OpenFog combined at the beginning of 2019 with a focus on industrial IoT, fog and edge computing. OpenFog originally targeted IoT services; it has now evolved toward a horizontal, system-level architecture that distributes computing, storage, control and networking functions closer to the users along a cloud-to-thing continuum.

Open Edge Computing (OEC) and Living Edge Lab (LEL): LEL is a live testbed for edge computing with leading edge application and user acceptance testing. OEC is a collective effort of multiple companies exploring technologies surrounding edge computing; it is also a common communication and exchange platform to promote LEL results. It is a partnership with Carnegie Mellon University.

Radio Edge Cloud (REC): REC is an appliance tuned to support the O-RAN Radio Access Network Intelligent Controller (RIC). The REC will make it possible for third parties to develop applications and have access to the RAN. REC is part of the Telco Appliance blueprint within the Akraino project. Today, most of the RANs are managed by the vendors that make telco gear. REC opens the RIC to applications (called xApps) and those xApps can be used to manage the RAN.

Kinetic Edge Alliance (KEA): The KEA is an industry group working on an edge reference architecture. This architecture includes an SDN-enabled compute fabric federated into a single, logical programmable entity that strand etches across geographically distributed locations.

7. NEXT GENERATION EDGE REFERENCE ARCHITECTURE

5G is anticipated to be revolutionary in many aspects and that will require new considerations for plan, design and operation of 5G systems. For example, optimal functional placement emerges as a key requirement to efficiently and effectively deploying 5G networks supporting ultra-low latency and high bandwidth applications. The placement of virtualized RAN and core network elements, and the application workloads (AR/VR, industrial automation, connected cars, and etcetera) will be dictated by the speed of light as shown in Figure 7.1.

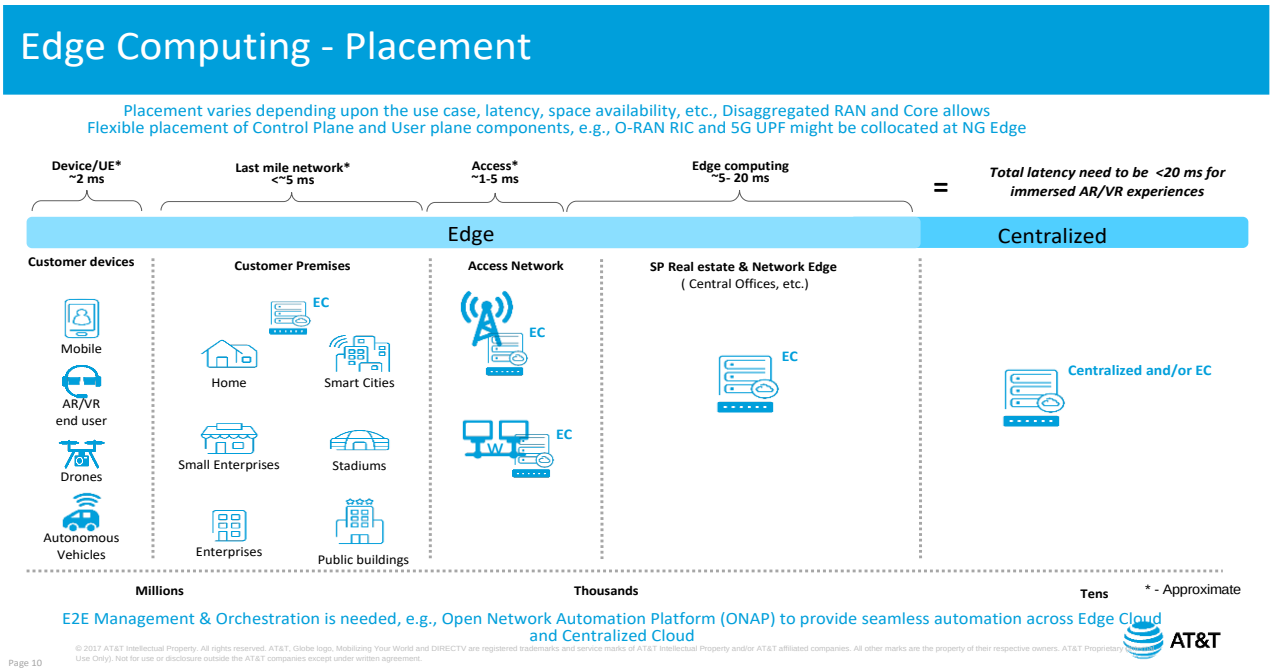

Figure 7.1. Edge Computing Placement.

Following are some of the applications that might be deployed on the Edge along with their associated time scales:

- Non-real time apps/3rd party analytic applications (>~500 ms) that support a broad scope including slice monitoring, performance analysis, fault analysis, root cause analysis, and centralized SON applications, ML methodologies for various applications, Policy and optimization applications (for example, Video optimization, Drive Test Minimization, and etcetera)

- Near-real time (~50-100 ms) UE / Area optimization applications/ 3rd party apps: These are in service path optimization applications and run in open Centralized Unit-Control Plane (High) (CU-CP-H) platforms (also known as RAN Intelligent Controller or SD-RAN controller). These applications include load balancing, link set-up, policies for L1-3 functions, admission control and leverage standard interfaces defined by O-RAN Alliance between network information bases (or context databases) and third-party applications. Data collection is through B1 and implemented using x technology.

- Third party applications that are value added services, for example, advertising, marketing and etcetera: These applications don't fall into network optimization (as is the case with the previous two examples) or operations automation (second example), but rather value-added services. MEC or Edge Data Collection, Analytics and Events (DCAE) can provide needed data (for example, geo location, anonymized customer data, and etcetera) via standard sets of APIs.

  Applications really sit outside ONAP or network infrastructure, for example: special services in stadiums, exhibitions, malls, deployed in combination with Small Cells; enterprise campuses, deployed in combination with Small Cell and Macro Base Transceiver Station (BTS), such as Remote Radio Head (RRH), and Distributed Antenna System (DAS); city-wide applications such as IoT applications deployed as part of Smart City initiatives, or services for city residents and visitors deployed at metro aggregation sites and baseband hotels.

Network-wide applications such as essential network functions and ubiquitous services require a consistent experience / performance deployment in combination with radio cloud or specific deployment patterns for uses like Vehicle-to-Everything communication (V2X) along roads. Therefore, security, trust, metering / charging model, and more must be considered. These applications have different latency requirements and could offer multiple deployment options at the edge, central or vendor location.

- Third party applications that directly interact with the UEs, like AR/VR, factory automation, drone control, and etcetera: In this case, messages, requests or measurements go directly from the UE via User Plan Function (UPF) or GWs to the applications and the applications respond back. ONAP can deploy these applications and manage them just like any other network element, or they could be un-managed applications like Access Point Names (APN) in today's world.

Functional requirements resulting from this disparate set of applications is driving a fundamental change in design and deployment of mobile systems. Key elements of this pivot are:

- The need to federate heterogenous service environments, for example, telco access services and Cloud Data Center services, stitching disparate control systems in order to distribute 'Resources' (including compute, storage, networking) and 'Intelligence' in a spatio-temporal manner as depicted in Figure 7.1.
- The need to support heterogeneous compute environments, including Graphical Processing Units (GPU), highly programmable network accelerators, and etcetera, in addition to traditional compute, storage, and more

- The need for new functional elements that enable collection, monitoring and accumulation of sensor data, as well as processing and dissemination of control information from a wide array of end user devices ranging from traditional mobile devices of today to cyber-physical systems of the future. This includes mechanisms to collect real time radio network information, process it in real-time (for example, Geo Location data), summarize, anonymize, and etcetera and make it available to third party applications deployed at the edge, a central location or outside the service provider's environment

- The need for distribution of compute that allows sensor 'data' to be processed locally into 'information', and then into 'knowledge'; and finally, into 'wisdom' when processed with global context over a sustained period of time; the broader the data set, the deeper the wisdom discerned

In order to effectively align with this pivot, edge systems must evolve toward a recursive model that helps to address complexity while enabling extreme flexibility. Edge systems must become increasingly adaptive in order to enable a system of ultra-lean modular compositions that maximizes extensibility. This must be done with minimal redundancy, and with the goal to eliminate hardware and software dependencies.

SDN, with its aim to separate control plane and user plane functions, is an important first step to enabling data plane programmability. SDN has proved to be a game-changer, as it's helped to establish concrete proof points that the control plane and the data plane can be separated, and that it's possible to use external programmatic controls.

NFV is also a step toward the application of SDN in the telecommunications environment, however, straight virtualization of monolith gateways isn't the answer. It is ineffective when closed proprietary controls remain preserved, merely transitioning from a physical to a virtual implementation of the same closed function. Key

to enabling effective programmatic control is the ability for an external system to control the various applications contributing to a particular user flow.

Introduced in Figure 7.2 is an architecture pattern that replaces the current monolithic systems with simple control applications. With this approach, creating a user flow becomes a matter of stitching together these control applications with east-west and north-south interfaces. These control applications could be implemented with a set of Control Plane components associated with respective User Plane (UP) capabilities and resources applicable to the domain they serve, for example, the Mobile Access Components associated with Mobile Access User Plane for the RAN, and similar associations for control applications representing other domains.

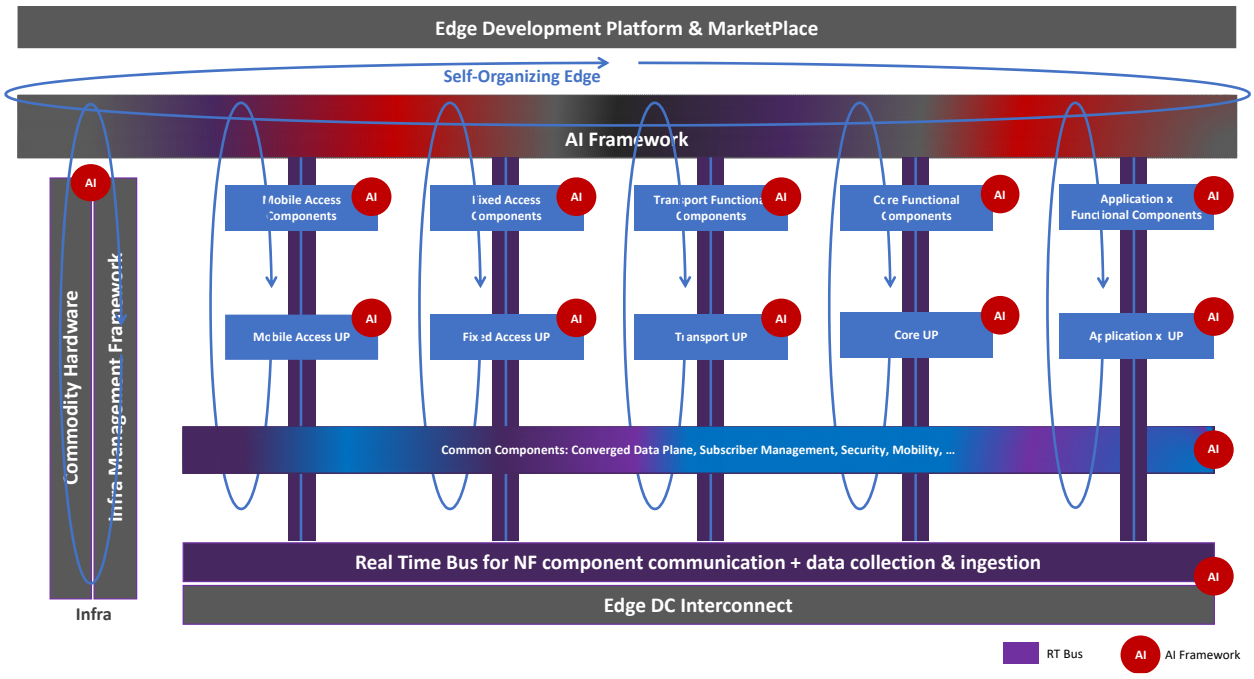

Figure 7.2. Unified Edge Architecture for Wireline and Wireless Access

Principle elements of this architecture pattern are disaggregation, programmability, and disaggregation of latency constrained network functions as well as distribution and interconnection of disaggregated functional components, and distribution of intelligence for self-optimizing systems.

7.1 DISAGGREGATION

The virtualization of network functions and the disaggregation of traditional network equipment presents an opportunity to build networks in fundamentally different ways than they are built today.

Disaggregation is the concept of breaking apart a tightly integrated system into its individual components. The purpose is to: 1) allow the system to use any of many available components for a specific function, and 2) allow the disaggregated components to be recombined in a more efficient manner.

By providing the flexibility to choose which components are used, the component with the “best” attributes such as cost, scalability, latency, and etcetera can be used. In an open, standardized system, these components can come from any provider and thus avoid vendor lock-in. The flexible interchange of

components also allows the ability to introduce new functions into the system, with the possibility of creating entirely new services, without having to reconstruct the entire system.

Disaggregation has shown it can be very disruptive to the industry to which it is applied. One famous example is how the personal computer disrupted the mainframe computer market. These are the same principles that ultimately lead to Sun Microsystem's loss of the workstation market.

The main issue with a disaggregated composition is the complexity introduced when the components are no longer part of an integrated system. Specifically, the main challenges are, how the components communicate with each other, and how packets' flows are mapped through the various components. In order to achieve this, a virtualized and modularized network OS is required that allows the use of APIs between disaggregated software components.

Disaggregation has the following three advantages for integrated devices:

1. **Cost**: Use of open-source hardware and software, as well as slowing competition amount the components, will lead to cost reductions

2. **Feature Flexibility**: Feature and functionality in networking equipment is locked into a cycle of software releases. Upgrading the software on an integrated device often requires taking the device completely offline. In a disaggregated architecture, software can be added or changed on the system as fast as the software modules are developed

3. **Scalability**: In a disaggregated architecture, only the components that need to be are upgraded. If for instance the forwarding/user plane needs to be upgraded for bandwidth or density issues, then this can be accomplished without replacing or upgrading the control plane

7.2 PROGRAMMABILITY

In addition to the advantages of disaggregation, concepts such as Software Defined Networking (SDN) can fundamentally change network architectures, by allowing networks to be built around a centralized control instead of traditional distributed control. SDN separates the control plane from the forwarding plane, allowing the network to be composed of a (logically) centralized control system that manages multiple, dispersed elements. SDN enables the creation of a network platform, over which network applications and an ecosystem of functions and applications can be further built. This further supports the ability to achieve new functions by re-architecting the flow between network components.

There is also value in the ability to program the forwarding plane. Programming the forwarding plane with protocols such as P4 or NPL (National Physical Laboratory) allow the implementations of new functions, or features for specific needs, all at DevOps speed and contained within the forwarding plane. This allows the service providers to implement new functions and enable them to open the network devices to their customers, allowing the customers to run custom designed functions.

For example, a programmable switch could be used to incorporate security functions (therefore, firewall-like filtering functions) directly into the packet switches themselves. This would theoretically eliminate the need for separate firewall appliances or virtual firewall functions, and thereby integrate security into the network fabric itself. Another example is vNF (Virtual Network Function) offloading, with functions such as BNG (Broadband Network Gateway) included in the packet switches to free up the resources on general compute servers.

7.3 DISAGGREGATION OF LATENCY CONSTRAINED NETWORK FUNCTIONS

A key tenet of this architecture is replacement of monolithic applications with disaggregated applications made up of multiple reusable components.

When a monolithic application is broken down into its component parts, a potential pitfall exists for applications with real-time communication requirements between the components. This is especially evident with components that have control loops at different timescales that require the close coordination that was possible in the monolithic construct. For example, these types of control loops exist in the 4G and 5G RAN and Edge. The objective is to ensure that the disaggregated application achieves the same performance as it had as a monolith.

One obvious solution is to ensure that components aren't placed any further apart than the latency would allow, but this is not enough to ensure the prioritization of messages at different time scales. In order to support this requirement, one possible option is a "real-time bus" used for inter-component and extra-component communication. The real-time bus is a carefully choreographed balance between hardware, software, and network engineering. The main participants in the real-time bus would be the compute hosting the components and the top-of-rack switch.

Hardware requirements include packet acceleration techniques within the container host(s) such as Field-Programmable Gate Arrays (FPGA) and offload hardware such as smart Network Interface Cards (NIC). Software could include Vector Packet Processing (VPP), Single Root Input/Output (I/O) Virtualization (SR-IOV), Data Plan Development Kit (DPDK), Netmap, PF_RING, and a host of other accelerators as well as queue managers.

The real-time bus would work together to ensure that the time-domains of inter-component messages are respected such that none of the messages at any of the timescales get delayed beyond what is acceptable for that time-domain.

7.4 DISTRIBUTION AND INTERCONNECTION OF DISAGGREGATED FUNCTIONAL COMPONENTS

A distributed edge cloud presents some challenges for management and service deployment. One approach that can be used to simplify the distributed edge-cloud is to make it appear as a single borderless cloud. There are different ways that this can be achieved using either a centralized mechanism or, preferably, a distributed mechanism.

An edge-cloud is composed of three key resources: compute, store, and connect (networking). These three resources are distributed throughout the edge-cloud and serve as a platform for the delivery of services. These services can be composed of interconnected components running on different compute platforms using storage and collaborating over connect.

The aggregate of all these resources, regardless of their distributed nature, can be viewed as a single fabric used to deliver services. Each resource has two fundamental behaviors that can be leveraged to assemble this distributed fabric. First, a resource can be queried (sensed) to gather information on its current state and, second, each resource can be requested to take a specific action (act). Using these fundamental properties, each resource can be equipped with a software process that has a control loop that continuously senses the state of the resource, discerns key information, infers and decides on some action and then requests that the resource take an action.

These software processes also possess the ability to communicate with each other, which enables them to collaborate and construct fabric-based services from the individual resources. Each of these processes is independently able to receive an intent request and then work with peer processes representing other

resources to assemble the service represented by the intent. This fully distributed model allows new resources to be added or existing resources to be removed with very little operational overhead. The new resource "looks" for other peer processes and automatically integrates itself into the overall fabric.

Using this approach, there is no concern for the individual components, where they're located or how they're interconnected. The software processes are able, through collaboration, to establish the most optimized placement and interconnection of workloads to meet the requested intent if the required resources are available.

7.5 DISTRIBUTION OF INTELLIGENCE FOR SELF-OPTIMIZING SYSTEMS

Optimal distribution of Intelligence is a key challenge for 5G Edge systems expected to serve a multitude of disparate autonomic systems. In-turn, they may be comprised of device swarms that contribute localized autonomic control to their respective systems.

Currently prevalent automation systems rely mostly on static policies driven mainly by back-end data analysis, which is typical of human-in-the-loop control systems. This works just fine with simpler systems of today, but when it comes to nested control of disparate autonomic systems, and stringent latency constraints anticipated with 5G era control applications, advanced automation techniques such as 'self-learning policies' and 'self-driving control' are deemed essential.

One of the first uses of self-optimizing or self-governing systems came about in radio systems, with the SON (Self Optimizing Networks) capabilities specified by NGMN and 3GPP for optimization of resources across heterogeneous access technologies. These systems, however, are limited in functional scope.

Similarly, Artificial Intelligence and Machine Learning have been around for a while but use of such techniques in telecommunications systems has just recently shown some promise. When combined with flexibility afforded by SDN and NFV, application of AI and ML for autonomic system control provides a perfect breeding ground for self-optimizing systems that are dynamically composed and continuously adapt themselves with a fabric of recursive control defined in terms of Sense (detect what's happening), Discern (interpret senses), Infer (understand implications), Decide (choose a course of action), and Act (take action).

8. DEPLOYMENT CONSIDERATIONS

Edge Cloud (EC) sites constitute a shared resource for delivery of a variety of applications such as AR/VR, autonomous cars, smart highways, connected homes, and factories and interactive services. Edge Cloud sites are implemented as a suite of virtual network functions (vNF), each with its own resource and location requirements of EC sites. Similarly, different RAN function splits also present their requirements at EC sites. EC site selection and desirable network topology depends critically on the composite requirements of important attributes, spanning all supported network functions including the following:

- Latency – UE - EC
- Subscriber distribution
- Required geographic coverage
- Traffic characteristics – time of day variation, mobility of UE
- Throughput, compute and store requirements

- Connectivity with network based common functions
- Resiliency and load balancing across EC sites
- Availability, restoration and network reliability

The following characteristics are critical to how well EC sites and the underlying transport network can meet these requirements.

Ubiquity

An efficient EC design is enabled by ubiquitous availability of candidate sites to match traffic/subscriber distribution connected by a resilient topology. Depending on latency requirements and geographic distribution of UE locations, appropriate availability of EC sites sufficiently close to the UEs are required.

Infrastructure

A potential EC site may be qualified in terms of infrastructure resources such as survivable data-center quality rack space, power, cooling and networking bandwidth (BW). Depending on availability of these resources, a site may be qualified for a subset of vNF that may be served at the site. Such qualification expands utility of even smaller sites.

For example, a vNF requiring low latency and low compute and storage resources, in support of a RAN function split, may be served from most available sites. This makes it easier to satisfy the low latency requirements for most UEs. A packet or data intensive function may require a large amount of store and compute, with corresponding rack space, power and cooling available in fewer sites; however it may tolerate high latency. Common centralized functions may require high availability characteristic of highly connected, resilient backbone sites. Load balancing may also be used to distribute compute workloads among several nearby smaller EC sites.

Connectivity and Restoration

Topology and routing protocols of each 1-3 network layer (optical, data, IP/network) as shown in Figure 8.1, as well as effective restoration mechanisms at each 1-3 network layer and their efficient interworking, can play a critical role in the delivery of requirements of latency under normal conditions as well as conditions with network impairment. Generally, the more “meshy” or connected the network the better.

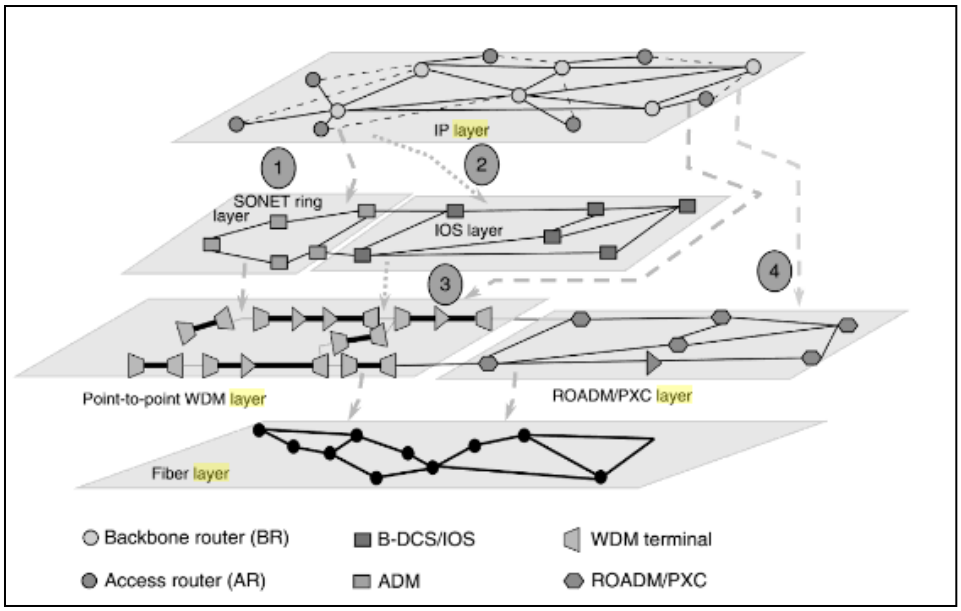

Figure 8.1. Illustration of L1-3 network layer connectivity.[12]

At the two extremes of the topology are a "skinny" tree network and a fully connected network (Figure 8.2). Most robust networks have a mesh topology with a high average node degree and efficient routing and restoration protocols at each network layers. It is also a common practice to have a robust, better connected core backbone network to support critical common functions and provide sufficient path diversity to restore failures at the edge of the network. A minimum degree of 2 or higher per network node may also be used to ensure multiple survivable paths exist to EC nodes under conditions of network impairment.

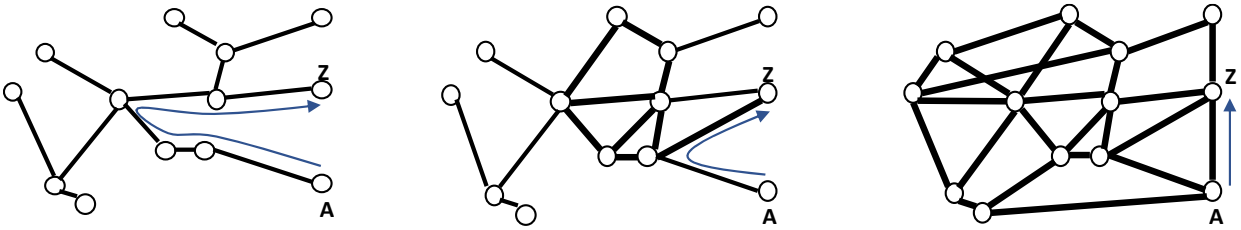

Figure 8.2. Network Topology Illustrations.

8.1 EDGE COMPUTING NETWORK DESIGN

Selection of Edge Computing (EC) sites and distribution of workloads to EC sites is a complex optimization problem. SLA requirements for latency, other resources for different vNFs, topological and other resource constraints and multiple criteria including cost and availability are challenges. At heart it is a large constrained combinatorial problem. Several clustering and mathematical programming approaches are

[12] *Commercial Optical Networks, Overlay Networks and Services*, Robert Doverspike & Peter Magill, from Optical Fiber Telecommunications VB: Systems and Networks. Ed. Ivan Kaminow, et.el. Elsevier. 2010. p 531.

employed for design and many efficient heuristics have been developed. Design methodology details are specific to a provider's network and resources.

Figure 8.3 illustrates an EC placement for a large mobile network with M cell sites homed to C nodes in a network with N nodes and L links with a minimal 2-connected mesh topology. A nested EC placement satisfying vNFs with three latency levels, 1 ms, 3ms and 5 ms and minimum coverage of 95 percent of cell sites is shown. The resulting placement has n1 EC sites for 1 ms SLA, n2 3 ms SLA and n3 5 ms SLA.

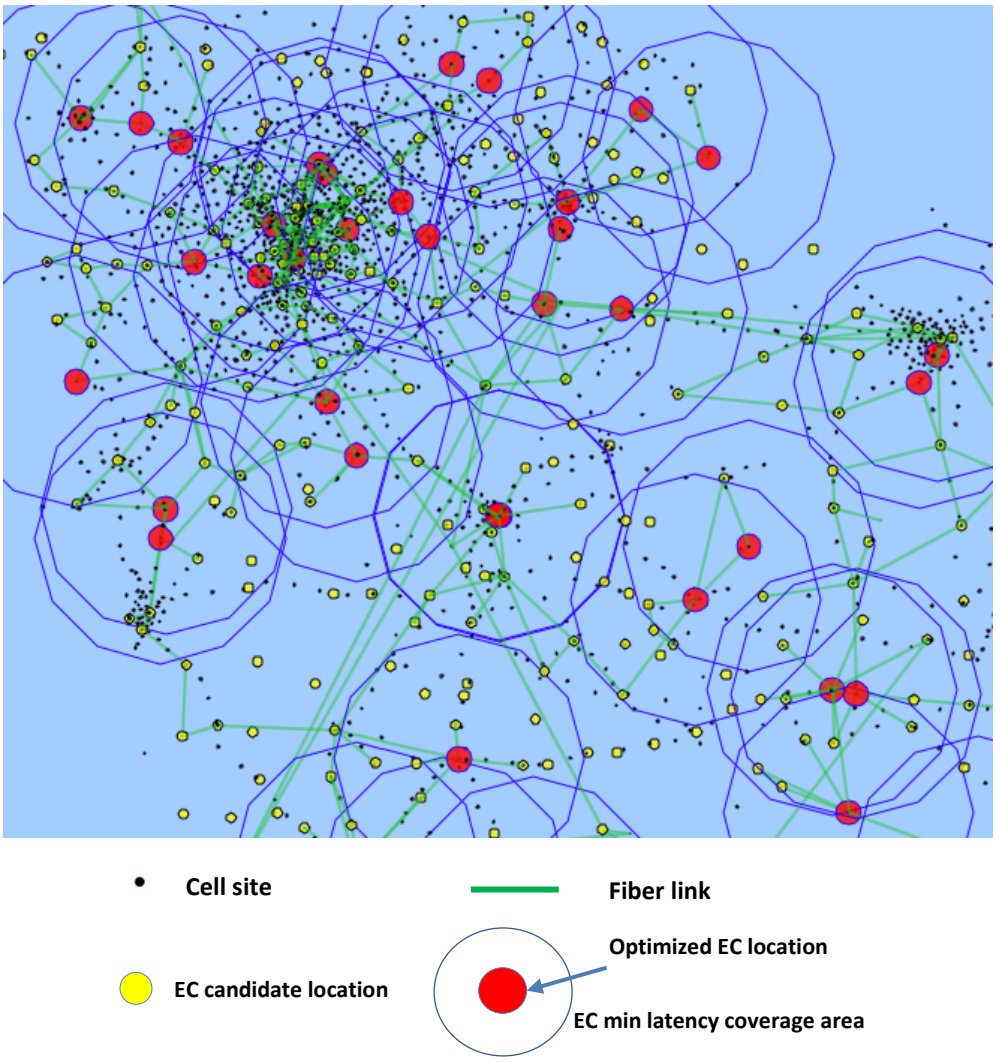

Figure 8.3. Illustration of an EC Site Selection for a Large Mobility Network with 1 ms Latency, 95% Cell Site Primary and Failover EC Coverage Requirements.

9. ROLE OF OPEN SOURCE AND STANDARDS

Edge is characterized by a variety of requirements from fixed and mobile access networks, driven by different themes and use-cases, deployment scenarios, and business models. Two key technology paradigms overlap these edge requirements - 5G and Cloud Computing.

ETSI NFV has a MEC specific workgroup but other SDOs that work on the Mobile system architecture such as 3GPP and O-RAN Alliance do not specify architectures that are Edge-specific. There are also several open source projects that draw from cloud concepts to apply to edge. Notable among them are Linux Foundation Edge and OpenStack Edge initiatives.

In order to create an open field of innovation with guard rails, organizations have had to adapt and/or create edge-focused requirements and open source projects. One of the key elements of 5G technology evolution is the role being played by Open Source initiatives in combination with Standards Development Organizations (SDO).

The foremost drivers for the use of Open Source include promoting multivendor ecosystem by lowering the interoperability barriers, increasing the innovation velocity by leveraging open source software, and realizing cloud scale economics through virtualization.

Open Source itself is an umbrella term that is used to capture multiple initiatives with different objectives. At a high level, open source initiatives fall into 3 categories:

- Open Interfaces: Driven by the need to promote multivendor interoperability and introducing competition towards best of breed procurement, SDOs define and publish architectures, interfaces, protocols and information models. Examples include architecture and new interfaces such as A1, O1, E2 and Open eCPRI specified under O-RAN alliance and ETSI NFV architecture for Edge

- Open Source Software and Open APIs: Open sourcing infrastructure (common/platform) software is motivated by the need to cut down software development time and facilitate faster innovation. Open APIs allow context exposure from network functions to third party applications. Examples include various Open Source Software projects under Linux Foundation Edge (LF Edge)

- Open Reference Design: By publishing COTS hardware reference specifications and blueprints for deployment scenarios and use cases, open reference designs (under initiatives such as Akraino) accelerate the virtualization of edge network functions towards a cloud native, service-agile operational environment

The field of Edge technology development spans multiple open source bodies, Standards Development Organizations and Industry consortia.:

9.1 OPEN SOURCE INITIATIVES

There are various open source initiatives that all focus on various elements. Some of the following were explained in section 6.1.1. However, the following list provides further explanation on the various open source initiatives taking place throughout the industry.

OpenStack Foundation- Edge Computing Group (OSF- Edge):

OpenStack provides fundamental building blocks for the virtualization infrastructure that can be deployed anywhere, including the edge of the network. As a highly distributed infrastructure software, OpenStack is running in thousands of data centers around the world. The modular nature of OpenStack is leveraged to help run the minimal services required at the edge, yet provide robust support for bare metal, container technologies and virtual machines. The user group consisting of telecom and retail industries are working to advance the edge computing use cases with OpenStack.

OSF Edge Working Group's objective is to define infrastructure systems needed to support applications distributed over a broad geographic area, with potentially thousands of sites, located as close as possible to discrete data sources, physical elements or end users. The assumption is that network connectivity is over a WAN.

The OSF Edge Working Group identifies in its mission statement that it will identify use cases, develop requirements, and produce viable architecture options and tests for evaluating new and existing solutions, across different industries and global constituencies. This is aimed to enable development activities for Open Infrastructure and other Open Source community projects to support edge use cases.

The group also connects and works with different open source projects such as Glance, Keystone, Ironic, Airship and more, with adjacent open source communities such as ONAP, OPNFV ETSI MEC, Akraino and etcetera.

Linux Foundation Edge (LF Edge):

LF Edge is an umbrella organization that aims to establish an open, interoperable framework for edge computing independent of hardware, silicon, cloud, or operating system.

The projects included under LF Edge are:

- **Akraino Edge Stack:** Akraino Edge Stack aims to create an open source software stack that supports high-availability cloud services optimized for edge computing systems and applications. The Akraino Edge Stack is designed to improve the state of edge cloud infrastructure for enterprise edge, OTT edge, and carrier edge networks.

  It will offer users new levels of flexibility to scale edge cloud services quickly, to maximize the applications and functions supported at the edge, and to help ensure the reliability of systems that must be up at all times. Akraino Release 1 delivers a deployable and fully functional edge stack for edge use cases ranging from Industrial IoT, Telco 5G Core and vRAN (Virtual Radio Access Network), uCPE (Universal Customer Premises Equipment), SDWAN (Software-Defined Wide Area Network), edge media processing, and carrier edge media processing and creates framework for defining and standardizing APIs across stacks, via upstream/downstream collaboration.

  Akraino is currently comprised of 11+ blueprint families that includes 19+ specific blueprints under development to support a variety of edge use cases. The community tests and validates the blueprints on real hardware labs supported by users and community members.

- **EdgeX foundry:** EdgeX Foundry is a vendor-neutral open source project hosted by The Linux Foundation to build a common open framework for IoT edge computing. At the heart of the project is an interoperability framework hosted within a full hardware- and OS-agnostic reference software platform. The reference platform helps enable an ecosystem of plug-and-play components that unifies the marketplace and accelerates the deployment of IoT solutions.

  The EdgeX Foundry is an open platform for developers to build custom IoT solutions, either by feeding data into it, from their own devices and sensors, or consuming and processing data coming out.

- **Edge Virtualization Engine (Project EVE):** Project EVE develops the open source Edge Virtualization Engine (EVE) for deployments on bare metal device hardware. It also provides system and orchestration services, and a container runtime for platform consistency. EVE allows cloud-native development practices in IOT and edge applications. Automating remote operations mean developers can handle far larger fleets of devices.

- **Open Glossary of Edge Computing:** This project provides a concise collection of terms related to the field of edge computing and aims to improve communication and accelerates innovation through a shared vocabulary.

- **Home Edge:** Aims to enable residential edge computing open source framework, platform and ecosystem

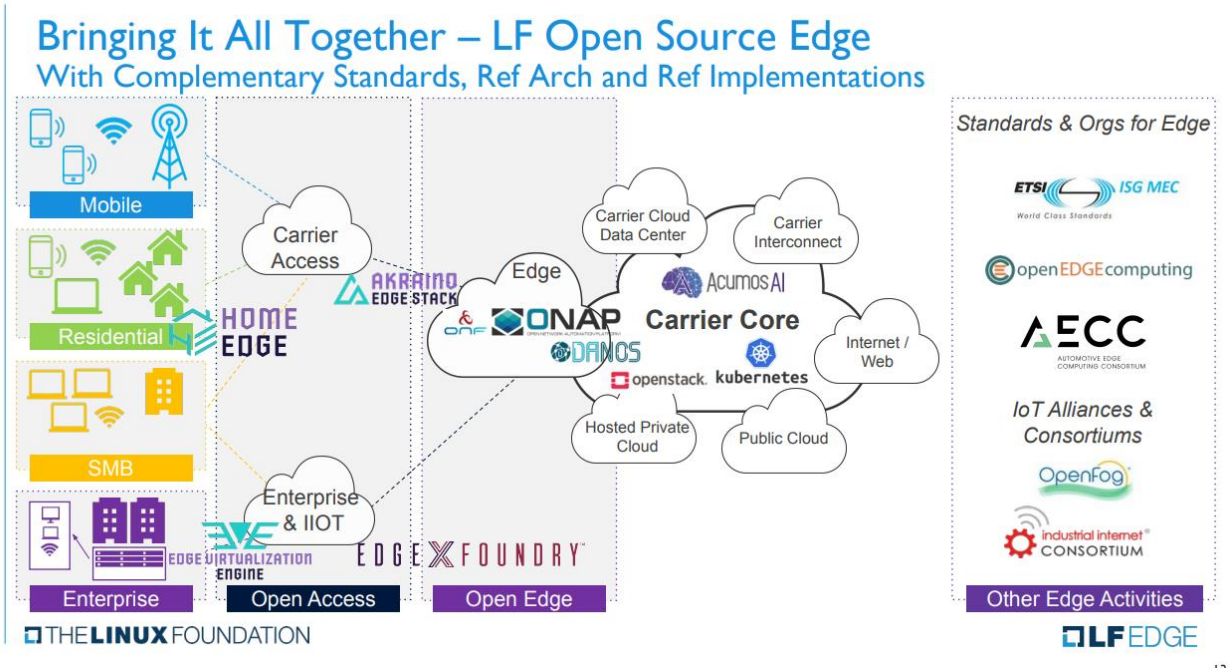

Figure 9.1. Complementary Standards, Reference Architecture and Implementations.

Other Open Source Projects which are not specifically Edge-focused, but which provide key Edge building blocks are:

Open Network Automation Platform (ONAP):

ONAP under the Linux Foundation provides ONAP provides an open platform for real-time, policy-driven orchestration and automation of physical and virtual network functions to rapidly automate new services and support complete lifecycle management.

Open Platform for NFV (OPNFV):

OPNFV is a collaborative open-source project under the Linux Foundation that helps define a common open source NFV platform. OPNFV works closely with ETSI to promote consistent open NFV standard. It works closely with ETSI NFV to allow a common NFVI, Continuous Integration (CI) with upstream projects, stand-alone testing toolsets, and a compliance and verification program. OPNFV continuously release

versions that have been integrated, tested, and are deployment-ready. OpenDaylight, a sub-project, provides an Open SDN-based network control with a loose collection of independent open source projects such as Openstack, Openflow, OpenVSwitch and etcetera.

Cloud Native Computing Foundation (CNCF):

CNCF uses an open source software stack to deploy applications as microservices, packaging each part into its own container, and dynamically orchestrating those containers to optimize resource utilization. It contains several fast-growing open source projects such as Kubernetes, Prometheus and etcetera.

Acumos:

Acumos under Linux Foundation provides an open source framework that makes it easy to build, share, and deploy AI apps. Acumos standardizes the infrastructure stack and components required to run an out-of-the-box general AI environment.

9.2 STANDARDS DEVELOPMENT ORGANIZATIONS

Standards Development Organizations (SDOs) play a key role in standardizing the deployments for Global scale. SDOs fulfill a key role by taking an end-to-end view of the network ecosystem to help define and develop scalable, backward compatible, future-proof technologies with a long-term outlook. This renders developing specifications for standardization a slower and complex process than, for instance developing applications with readily available open source platform code. Initiatives such as O-RAN alliance are trying to address this by focusing specifically on one part of the network - the RAN – for faster innovation.

Edge is an active technology domain of significant industry interest; this gets reflected in the multitude of activities in various open source forums and SDOs. The result is a crowded space today in terms of edge focused initiatives, open source projects, standards development and industry consortiums. It is important to secure that the work done in these projects, standards and industry bodies are coordinated and feed into each other to avoid long term redundant work activities and inefficient overlaps.

Open Source Projects and Standardization work need to work in combination and develop a strong technology base for innovative products and applications at a scale that favor global realization of the technology benefits. It is important to note that Open Source and Standards need to work hand-in-hand in order to bring innovation to global scale deployments, avoid costly fragmentation of the industry so the benefits of 5G, Cloud and Edge are accessible to all sections of the global society.

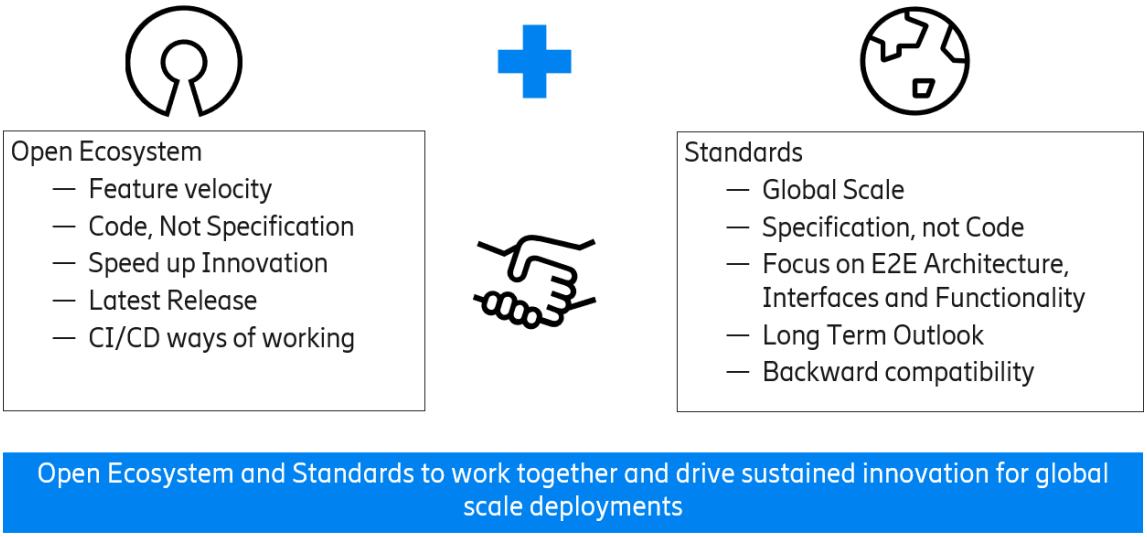

Figure 9.2. Open Ecosystem and Standards

ETSI NFV Multi-access Edge Computing (MEC) ISG: This initiative started out as Mobile Edge Compute; it has evolved with support of SDN/NFV toward Multi-Access Edge Compute. It offers cloud computing capabilities at the edge of the network. MEC provides a new ecosystem and value chain at the edge of the wireless network and is an evolution towards the convergence of IT and telecommunications networking. A key technology element is the MEC Server (referred to as MEC Host framework) that provides RAN API services for development of network optimization applications: Radio Network Information Service (RNIS); Location Information Service; and Bandwidth Manager Service. ETSI MEC working group defines reference architecture and APIs for multi-access edge computing (MEC). It provides reference architecture and API specification with a focus towards Telco cloud applications. It specifies the elements that are required to enable applications to be hosted in a multi-vendor multi-access edge computing environment.

IEEE Edge Automation Platform (EAP): A high-level perspective and projection of how the industry could evolve, with highlights of common needs, the challenges to achieving those needs, and the potential solutions to those challenges.

3GPP: The 3rd Generation Partnership Project develops architecture and specifications that cover cellular telecommunications technologies, including radio access, core network and service capabilities. They provide a complete system description for mobile telecommunications. They also provide hooks for non-radio access to the core network, and for interworking with non-3GPP networks. While 3GPP does not define Edge-specific architectures or how the 3GPP functions are implemented (physical, virtual and etcetera) the work done in SA2- Architecture group contains key constructs and architectural aspects relevant for Edge such as disaggregated mobile core, control plane user plane separation (CUPS), URLLC and access convergence.

O-RAN ALLIANCE: O-RAN Alliance, founded in 2018 by Operator members across the globe, has stated its mission as “Leading the industry towards open, interoperable interfaces and RAN virtualization”. As of September 2019, the O-RAN community consists of 21 operator members and 82 contributing vendor members. The Open RAN Alliance is driven by telecommunications service providers to move radio access networks towards open interfaces and machine learning/artificial intelligence. Disaggregated RAN systems will provide an extensible, software-based service delivery platform capable of rapidly responding to

changing user, application and business needs. The O-RAN architecture is based on well-defined, standardized interfaces to enable an open, interoperable supply chain ecosystem in full support of, and complimentary to, standards promoted by 3GPP and other industry standards organizations.

10. 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, with most users mainly interested in accessing large amounts of information, irrespective of its physical location.

10.1 NEW INTERNET ARCHITECTURES

This paradigm shift in the usage model of the Internet, along with the need for security and mobility support, 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.

Transdisciplinary Research Institute for Advancing Data Science (TRIAD) was one of the first efforts that proposed extending the Internet with content routing capabilities, but Data Oriented Network Architecture (DONA) from UC Berkeley is one of the first complete ICN architectures, that radically changed naming by replacing the hierarchical Uniform Resource Locators (URL) with flat names. Van Jacobsen's seminal talk on ICN in 2006 spurred a lot of activity including the development of Content Centric Networking.

The National Science Foundation's (NSF) subsequent investment in the Future Internet Architectures (FIA) program really help solidify these efforts and led to the development of multiple ICN flavors. Named Data Networking (NDN) has continued its momentum in developing a community and tools that can be freely downloaded.

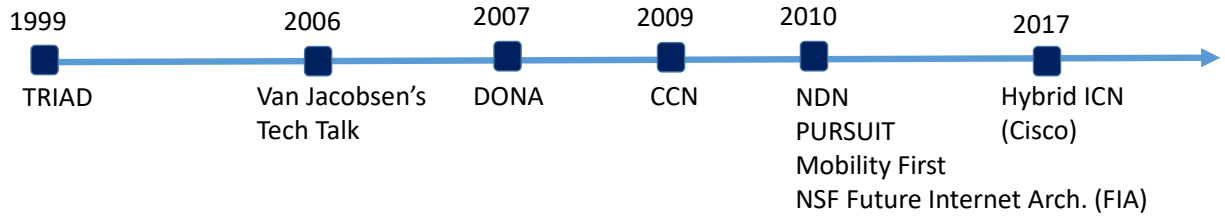

Figure 10.1. Timeline of Internet Architectures.

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.

10.2 IMPLEMENTATION OPTIONS FOR ICN

ICN is a network layer protocol meaning a layer 3 protocol in the OSI (Open Systems Interconnection) model. It essentially replaces the thin waist of IP with named data chunks. When ICN is implemented directly over layer 2 without IP, it is called a native implementation. NDN currently provides open source software that can be implemented natively in layer 3. However, since IP networks are typically present in all networks, replacing IP with ICN is unlikely and not recommended. ICN can co-exist with IP and some of the options are overlay, hybrid ICN, and dual stack.

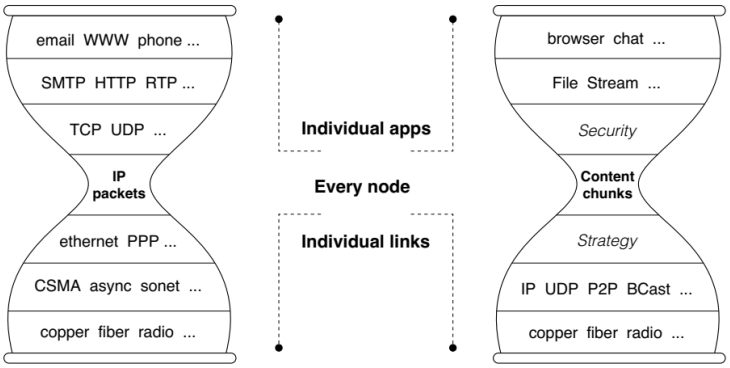

Figure 10.2. IP vs ICN Stack.

10.3 OVERLAY

Like IP, NDN is a universal overlay: NDN can run over anything that can forward datagrams (Ethernet, Wi-Fi, Bluetooth, cellular, IP, Transmission Control Protocol (TCP), and etcetera). NDN can simply run over the deployed IP infrastructure, rather than trying to replace or change it

In Figure 10.3, the solid nodes have NDN over IP while the empty nodes are regular IP nodes. In such an overlay, node A is connected to node F via node B. In NDN terms, node A has two faces, one to node C and one to node F. So, when node A has an NDN packet for node F, the IP layer below forwards the packet to node F, at which point it is received by the NDN layer and decoded. If node A has a packet for node C, it forwards the packet to face C. The IP layer then creates the one hop route to node C at which point the NDN layer receives the packet. Overlay creates a simple way to implement NDN without changing the whole network.

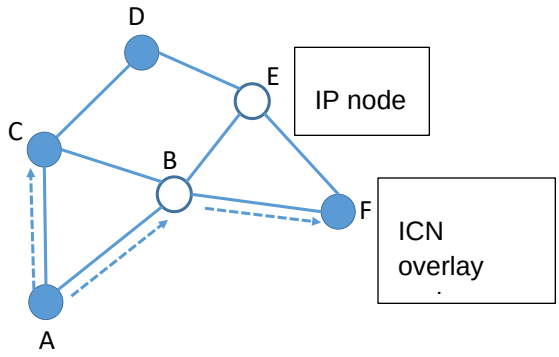

Figure 10.3. NDN Overlay.

10.4 HYBRID ICN

Cisco has designed Hybrid ICN (hICN) as shown in Figure 10.4 as a solution for deployment of ICN inside IP, rather than over IP. hICN maps names into IP addresses. An IP packet is modified such that the destination field carries the name of the content in a hICN interest packet. Regular IP routers are modified by adding a hICN forwarding module (for ICN processing). Regular IP routers forward the packet to the destination address. If a hICN router encounters the interest packet, it retrieves the data from its cache (if it has been cached) and serves the request while inserting the name of the data in the source address.

The benefit of hICN is that it can be integrated into existing IP networks and does not need a separate ICN to IP translator. However, it greatly limits the namespace design by confining it to IP addresses. Furthermore, other fields provided by the ICN packet formats cannot be used which also limits how many ICN features can be incorporated.

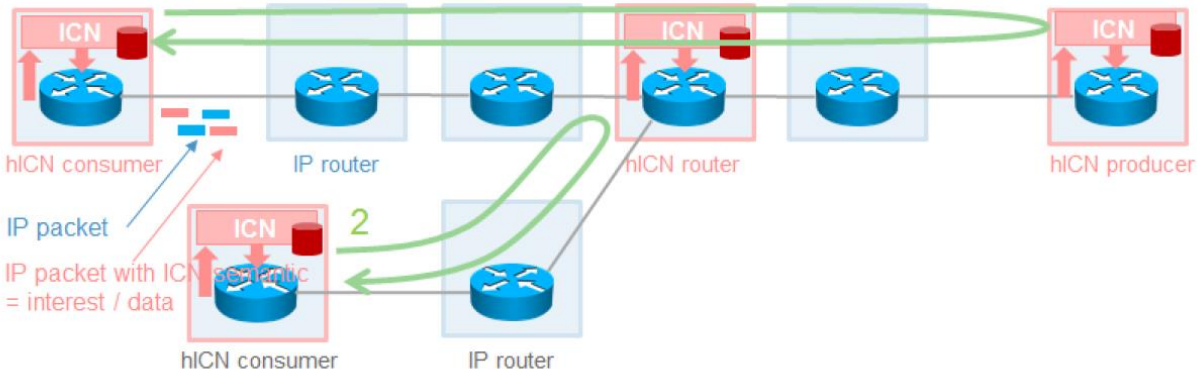

Figure 10.4. Hybrid ICN Architecture.

10.5 DUAL STACK

If ICN is used in an island with a well-defined gateway, then native implementations of ICN (directly over layer 2) or dual stack is an option. If ICN-based protocols become more popular, this approach can be used more widely in the network. However, proper schemes for interworking between ICN and IP layers need to be defined.

Operators are seeking new architectural approaches to optimize delivery of digital services processed at the edge of the network domain. Edge Computing is a natural development in the evolution towards the convergence of IT and telecommunication networking, which is being driven by software based platforms and the usage of IT virtualization technology for telecommunications infrastructure, functions and applications.

10.6 RECURSIVE INTER NETWORK ARCHITECTURE

RINA,[13] the Recursive Inter Network Architecture, is an innovative network architecture that relies on the premise that networking is distributed Inter-Process Communication (IPC). RINA decomposes networks into layers of generic protocols that can be configured via policies to optimally serve their operational requirements.

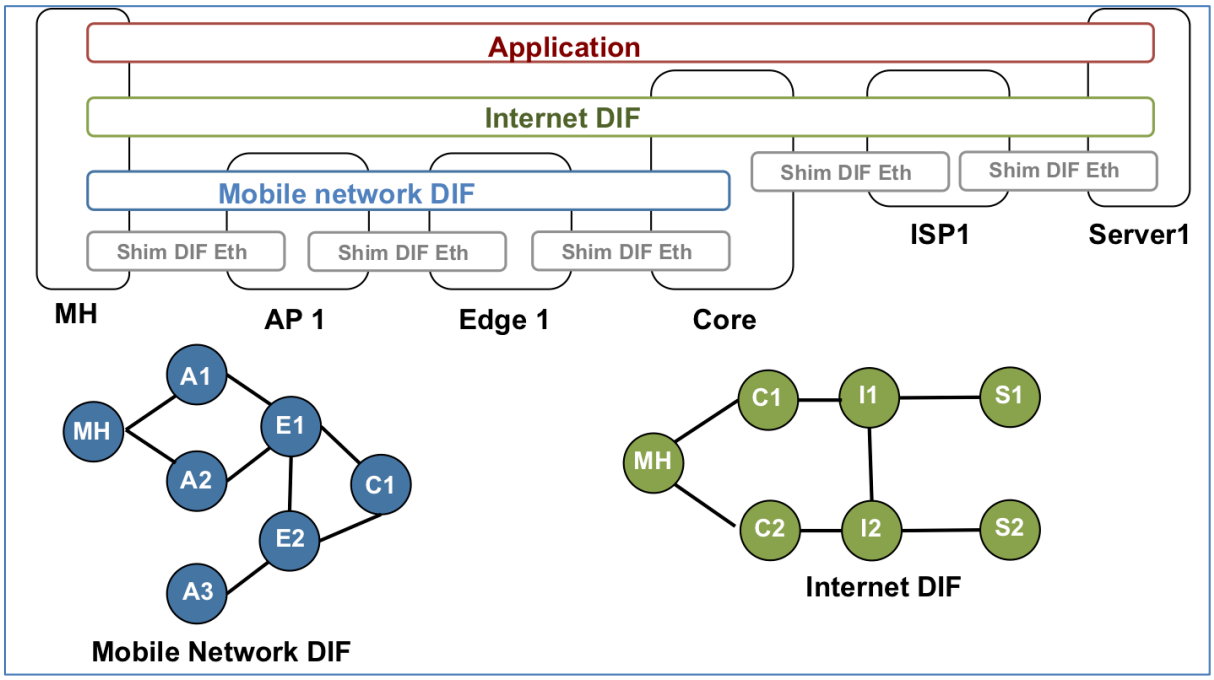

Figure 10.6. RINA Edge Computing.

As Figure 10.6. shows, in RINA there is a single type of layer—called a Distributed IPC Facility (DIF)—that is recursive, therefore, it repeats as many times as the network design requires.

RINA provides properties to simplify and optimise Edge Computing. RINA:

- Enables multi-homing without the need of special protocols
- Performs distributed mobility management
- Supports application naming and discovery through multiple networks[14]

[13] ETSI GR NGP 009 V1.1.1 (2019-02), *Next Generation Protocols (NGP); An example of a non-IP network protocol architecture based on RINA design principles*.
[14] *Layer discovery in RINA networks,in Computer Aided Modeling and Design of Communication Links and Networks (CAMAD),* 2012 IEEE 17th International Workshop. 2012. pp. 368–372.

- Delivers network slicing as an inherent capability of the architecture—a slice is seen as a 'QoS cube' within a DIF or sometimes a dedicated DIF with specific QoS requirements[15]
- Assures security policies including authentication, access control and confidentiality for each DIF[16]

Mobility Management at the Edge must be simple, fast and synchronized with the network database. In RINA, mobility of the application or user is not restricted to a particular layer, since all layers in RINA have the same infrastructure and protocol frameworks, and all of them are capable of dealing with systems that move using the same tools. This property allows network designers to scale both up and down to adapt to the highly variable density of users at the Edge; and to very low latency high speed mobile 'handover' that some network services must support.

RINA provides fully dynamic and efficient application discovery support—both within a layer and across layers. This generic capability of the architecture leverages all types of applications and eliminates the need for dedicated protocols to provide registration and discovery capabilities for every subset of applications.

RINA can support multi-domain, multi-operator end-to-end service orchestration. It enables the dynamic orchestration of DIFs across multiple providers by interacting with each provider's Network Management System to grow/reduce/modify the connectivity graph of such DIFs in response to flow allocation requests from applications. Recent experiments using RINA[17] demonstrate this orchestration for network operators, application developers and end-users.

The principles that underlie RINA lead to gains for network configuration management and mobility management. RINA networks can deliver services that demand fast failure recovery and guaranteed end-to-end QoS. Network renumbering is simplified and can be used to maximize the aggregation of addresses in routing tables. These enhancements are key to realize the next generation of the Edge Computing.

11. CONCLUSION

Computing and communications have evolved over time and yet we are once again at the precipice of a fundamental shift in its redefinition, as Edge empowers existing applications and enables a new business models, and revenue streams. Over the last few years, there has been enormous development in terms of the variety of use cases using the latest technology innovations in 5G and other media capabilities requiring robust mobile communications and computing capabilities at the Edge.

The role of Edge computing is to process data as close as possible to the source. In a 5G mobile network, the edge is the network that is located as close as possible to the end-user. Edge connectivity in its basic conception is a highly distributed computing environment that is used for applications to store and process content in close proximity to mobile users. Applications more often require real-time radio and network information and can offer a personalized and contextualized experience to the mobile subscriber. Edge computing can be complex and operators will deploy Edge computing elements to address specific services, applications and use cases.

Many 5G and IoT applications have large bandwidth needs, strict latency and high reliability requirements, like video traffic, gaming, AR/VR and connected cars. As new use cases and applications are powered by the 5G technologies spanning across a multitude of devices and edge clouds with stringent latency

[15] *Challenges of network slicing*, P.T.Neil Davies,,,IEEE SDN Newsletter. January 2017.
[16] *From protecting protocols to protecting layers: designing, implementing and experimenting with security policies in RINA*, E. Grasa, O. Rysavy, O. Lichtner, H. Asgari, J. Day, and L. Chitkushev IEEE ICC 2016, Communications and Information Systems Security Symposium. 2016.
[17] *Execution of experiments, analysis of results and benchmarking against KPIs,* Arcfire D4.4. June 2018.

requirements, the need for efficient architectures is paramount to addressing the other challenges connected with acceleration, analytics, AI and ML engines.

With the growth in mobile traffic increases, 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. One of the latest developments in this sphere, O-RAN addresses the service agility and cost considerations of networks. In this connection, several key issues including the emerging innovations in combinations of 5G and Artificial Intelligence technologies have been discussed in this white paper.

Sample use cases and requirements for 5G networks are also provided. It has been demonstrated how 5G can facilitate advanced mobility, compute, storage and acceleration features for applications with ranging latency considerations. The paper also covered application of 5G, AI and Edge Computing domains and details the benefits for emerging new use cases and services. It presented an overview of the vision and path to next generation Edge networks, examined the current state of 5G and Edge architectures, and reviewed the emerging role of AI and ML.

In defining the next generation Edge reference architecture and exploring future directions in networking, 5G has truly been demonstrated to be at the Edge of computing.

APPENDIX

ACRONYMS

Abbreviation	Meaning
3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GC	5G Core
AECC	Automotive Edge Computing Consortium
AI	Artificial Intelligence
AMF	APS Mode Mismatch Failure
API	Application Program Interface
APN	Access Point Name
APP	Application Layer
APS	Automatic Protection Switching
AR	Augmented Reality
ASIC	Application Specific Integrated Circuit
ATIS	Alliance for Telecommunications Industry Solutions
BBU	BaseBand Unit
BMI	Body Mass Index
BNG	Broadband Network Gateway
BTS	Base Transceiver Station
BW	Bandwidth
CA	Carrier Aggregation
CD	Communicable Disease
CDN	Content Distribution Network
CI	Continuous Integration
CN	Core Network
CNCF	Cloud Native Computing Function
COMAC	Converged Multi-Access and Core
CoMP	Coordinated Multi-Point
CORD	Central Office Re-architected as a Data center
CoSP	Communication Service Provider
CPRI	Common Public Radio Interface
CSP	Cloud Service Provider

CT	Computed Tomography
CU	Centralized Unit
CU-CP-H	Centralized Unit- Control Plane (high)
CUPS	Control Plane and User Plane Separation
DAS	Distributed Antenna System
DCAE	Data Collection Analytics & Events
DIF	Distributed Inter-Process Communication (IPC) Facility
DNS	Domain Name System
DPDK	Data Plan Development Kit
DPI	Deep Packet Inspection
D-RAN	Distributed Radio Access Network
DU	Distributed Unit
DWDM	Dense Wavelength Division Multiplexing
E2E	End-to-End
EaaS	Edge-as-a-Service
EAP	Edge Automation Platform
EC	Edge Cloud
ECG	Electrocardiogram
ECOMP	Enhanced Control Orchestration Management & Policy
eCON	evolution to Content Optimized Network
eMBB	Enhanced Mobile Broadband
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards institute
ETSI MEC	ETSI Multi-access Edge Computing
EVE	Edge Virtualization Engine
FIA	Future Internet Architecture
FIB	Forwarding Information Base
FlexE	Flexible Ethernet
FPGA	Field Programmable Gate Arrays
GE	Gigabit Ethernet
GHz	Gigahertz
gNodeB	Next Generation NodeB
GPU	Graphical Processing Unit
GS	Group Specification
HD	High Definition

hICN	Hybrid Information Centric Networking (ICN)
HLS	Higher Layer Split
HTTP	Hyper Text Transfer Protocol
ICN	Information Centric Networking
ICNRG	ICN Research Group
ICS	Industrial Control Systems
ICT&E	Information Communication Technology & Electronic
IEEE	Institute of Electrical & Electronic Engineers
IETF	Internet Engineering Task Force
IIC	Industrial Internet Consortium
IMT	International Mobile Telecommunication
I/O	Input/Output
IoT	Internet of Things
IP	Internet Protocol
IPC	Inter-Process Communication
IPTV	IP Television
IRTF	Internet Research Task Force
ISG	Industry Specification Group
IT	Internet Technology
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
KEA	Kinetic Edge Alliance
KPI	Key Performance Indicator
L1	Layer 1
L2	Layer 2
LEL	Living Edge Lab
LF Edge	Linux Foundation Edge
LLS	Lower-Layer Split
LSP	Label Switched Path
LTE	Long Term Evolution
LUI	Language User Interface
M2M	Machine-to-Machine
MAC	Medium Access Control
MCORD	Mobile edge platform for CORD
MEC	Mobile Edge Computing

MI	Machine Intelligence
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MME	Mobility Management Entity
mMTC	massive Machine-Type Communications
mmWave	Millimeter Wave
MNO	Mobile Network Operator
MPLS	Multi-Protocol Label Switching
MRI	Magnetic Resonance Imaging
ms	Millisecond
NCD	Non-Communicable Disease
NDN	Named Data Networking
NFaaS	Named Function-as-a-Service
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
ngLTE	Next Generation Long Term Evolution
NGMN	Next Generation Mobile Networks
NIC	Network Interface Card
NPL	National Physical Laboratory
NR	New Radio
NSA	Non-Standalone
NSF	National Science Foundation
NSI	Network Slice Instance
OCT	Optic Coherence Tomography
O-DU	ORAN Distributed Unit
OEC	Open Edge Computing and Office of Emergency Communications
OMEC	Open Mobile Evolved Core
ONAP	Open Networking Automation Platform
ONF	Open Networking Foundation
ONOS	Open Network Operating System
OPEX	Operational Expenditure
OPNFV	Open Platform for NFV
O-RAN	Open Radio Access Network
O-RU	ORAN Radio Unit
OS	Operating System

OSI	Operating System Interconnection
OST-Edge	OpenStack Foundation – Edge Computing Group
OT	Operational Technology
OTN	Optical Transport Network
OTT	Over-the-Top
PET	Positron Emission Tomography
PHY	Physical Layer
PIT	Pending Interest Table
PoC	Proof of Concept
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
REC	Radio Edge Cloud
Rel-X	3GPP Release (number, for example Rel-15)
RESTful	Representational State Transfer conforming Web server
RFC	Request for Comments
RIC	RAN Intelligent Controller
RICE	Remote Method of Invocation over ICN
RINA	Recursive Inter-Network Architecture
RNIS	Radio Network Information Services
RoE	Radio over Ethernet
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Remote Radio Unit
RSVP-TE	Resource Reservation Protocol – Traffic Engineering
RT-RIC	Real Time – RAN Intelligent Controller
SA	System Architecture or Standalone
SBA	Service-Based Architecture
SCL	Slicing Channel Layer
SDN	Software Defined Network
SDO	Standards Development Organization
SDWAN	Software Defined- Wide Area Network
S-GW	Serving Gateway

SLA	Service Level Agreement
SON	Self-Optimizing/Organizing Network
S/P-GW	Serving/Public Gateway
SR	Segment Routing
DONA	Data-Oriented Network Architecture
SR-IoV	Single Root- Input/Output (I/O) Virtualization
TCP	Transmission Control Protocol
TDF	Traffic Detection Function
TDM	Time-Division Multiplexing
TIP	Telecom Infrastructure Project
TLS	Transport Layer Security
TLV	Type Length Value
TR	Technical Report
TRIAD	Transdisciplinary Research Institute for Advancing Data Science
TSG	Technical Specification Group
TSP	Telecom Service Provider
uCPE	Universal Customer Premises Equipment
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URI	Uniform Resource Identifier
URLLC	Ultra-Reliable Low Latency Communication
V2X	Vehicle-to -Everything
VNF	Virtual Network Function
VoLTE	Voice over LTE
VM	Virtual Machine
VPP	Vector Packet Processing
VPN	Virtual Private Network
VR	Virtual Reality
WDM	Wavelength-Division Multiplexing
WID	Work Item Description

ACKNOWLEDGEMENTS

The mission of 5G Americas is to advocate for and facilitate the advancement of 5G and the transformation of LTE networks throughout the Americas region. 5G Americas is invested in developing a connected wireless community for the many economic and social benefits this will bring to all those living in the region.

5G Americas' Board of Governors members include AT&T, Cable & Wireless, Ciena, Cisco, CommScope, Ericsson, Intel, Kathrein, Mavenir, Nokia, Qualcomm Incorporated, Samsung, Shaw Communications Inc., Sprint, T-Mobile USA, Inc., Telefónica and WOM.

5G Americas would like to recognize the significant project leadership and important contributions of leaders Paul Smith of AT&T and Rao Yallapragada of Intel, along with many representatives from member companies on 5G Americas' Board of Governors who participated in the development of this white paper. The contents of this document reflect the research, analysis, and conclusions of 5G Americas and may not necessarily represent the comprehensive opinions and individual viewpoints of each particular 5G Americas member company. 5G Americas provides this document and the information contained herein for informational purposes only, for use at your sole risk. 5G Americas assumes no responsibility for errors or omissions in this document. This document is subject to revision or removal at any time without notice. No representations or warranties (whether expressed or implied) are made by 5G Americas and 5G Americas is not liable for and hereby disclaims any direct, indirect, punitive, special, incidental, consequential, or exemplary damages arising out of or in connection with the use of this document and any information contained in this document.