OCT 2021

A 5G Americas White Paper

5G EDGE AUTOMATION & INTELLIGENCE

Contents

Executive Summary

5G is the very first mobile wireless technology that seeks to connect anything anywhere, making communications both ubiquitous and as transparent as the air we breathe. A 5G ecosystem is emerging that is using pervasive, ever-present communications to expand the reach and scope of today's human-centric consumption patterns. One of 5G's greatest challenges ahead involves an increasingly complex, multi-dimensional, multi-variable world in terms of an ever-expanding multitude of service requirements and disparate traffic profiles. Additionally, a newly-introduced heterogenous mix of enabling technologies for both communications and compute, as well as their control, are adding to this emerging challenge.

An important part of the 5G ecosystem is edge computing. It facilitates radical new use cases that extend from the data center core to the network edges. Edge computing allows for compute and analytics to be moved closer to the data instead of exchanging ever growing amounts of data among cloud servers. As the ecosystem evolves, 5G and edge computing will further converge to enable edge network management, collect and capitalize on massive amounts of data while maintaining integrity and even ownership, and build pervasive intelligence for enabling various latency-sensitive, enterprise, and private services.

Two tasks are fundamental to address the complexity challenge and for the ecosystem to succeed: automation and optimization. Automation is required to cope with new scenarios in network and edge planning, operation, and management. Automation should ease the operation of network and compute infrastructure for rapidly growing vertical industries, transportation, and enterprise use cases that are bringing along with them new infrastructure owners. Optimization should also ease the extension of cloud computing and fast-growing Artificial Intelligence/Machine Learning (AI/ML) applications to the edge, as well as introduce self-optimization to best serve applications. Optimization strategies will ensure that every node in mobile networks can provide low latency, high reliability, and pervasive intelligence capabilities.

The first three chapters of this technical whitepaper details where and how intelligence can support the cross-over of communications in a 5G network and edge computing in a 5G edge network. Chapter 1 focuses on automation, while Chapter 2 looks at optimization of compute and communications. Chapter 3 provides insights on how to apply them.

The first two chapters will follow the same story line, starting with the specific background and introduction of the current state of the art and industry landscape. This is followed by a discussion on features and key technologies. Based on this discussion, the sections develop foreseeable requirements, illustrate emerging directions for network architectures, and conclude with system recommendations. Chapter 3 combines this understanding, with applications in potential use cases for autonomous industrial solutions, smart transport and energy, connected health, and digital twins – and more.

Core to the discussion include the technical capabilities of 5G and edge computing, where the intelligence of 5G network and edge computing can achieve interesting results. These include the automation of data collection, analysis and communication, and computing automation for ease of management. They may also include the optimization of network and computing to best support AI/ML applications at the edge, enabling pervasive intelligence at connected devices, distributed learning, situational networks, and collaborative edge intelligence.

This paper is the joint effort of experts from many different technical disciplines spanning research, development, operations, and applications. The result is a comprehensive discussion and guide that also shows how multi-dimensional complexity challenges can be tackled, bringing together expertise from multiple backgrounds united by a common goal: automating and optimizing 5G networks to capitalize on edge computing advances that serve future customer requirements and applications.

The combination of 5G and edge computing will create new capabilities and new business opportunities for the whole communication and computing industry. We hope this white paper will help the ecosystem achieve a consensus around the future of 5G and the edge technology roadmap, working closer to realize the vision, and bring a better automated, intelligent world together.

1. 5G Edge Automation

Cellular services have historically been delivered from a centralized location. Traffic from many cell sites is transported back to a central location, where it is hosted in a mobile packet core that provides connectivity to the wider Internet. However, the drawbacks of this approach have become apparent in the face of emerging applications that require low latency, high reliability, high bandwidth, and are characterized by localized communication among peers that share a common local domain. In many cases, rather than centralizing the entire service delivery, computation is needed for things such as automation of edge applications like video analytics, process control, self-driving cars, and more.

In this context, mobile network virtualization and cloudification of Radio Access Networks (RAN) have gained momentum in recent years. By decoupling hardware and software, network element protocol stacks can utilize the computing resources at the base stations or data centers near the radios and cell towers. This new paradigm creates new pools of computing resources at the edge of the network that can be utilized by other applications for different use cases, which has caused edge computing to become successful over the past decade.

The edge is defined as a continuum of edge zones, including such examples as:

- *Device edge – signal and data processing on the device*
- *Premise edge – processing that occurs on the premise (home, car, enterprise, etc.)*
- *Access edge – processing at cell sites or access Points of Presence (POP)*
- *Metro edge – upstream aggregation centers like Internet service providers, etc.*

This combination of different sizes of cloud data centers at global, national, local/regional, and potentially access locations are integrated into the network and operated by a central orchestration and management system. The exact specification of the infrastructure on the different sites may depend on the use cases and applications onboarded. In addition, there can be several infrastructure providers at the same site.

Automation at each of these edge zones has different forms and requirements. For example, an Internet of Things (IoT) device can perform autonomous and intelligent local computation based on its sensed environment. Automation can also impose a power reduction of IoT devices to conserve as much energy as possible. Another example involves self-driving cars, which can process data at the premise edge, as well as upload training data to the metro edge via the access edge for machine learning algorithms. Hence, various edge zones can be located at different locations, can execute different functions and/or decisions for different requirements, and may exchange information or data with each other.

1.1 Background

The term “Network Automation” or in short “Automation” has been used in the communication industry to describe a wide range of technologies that would help automate network system processes and service delivery with reduced or minimized human intervention. Human intervention is reduced by introducing predetermined decision criteria and related actions and embodying those predeterminations into the processes. Automation includes the use of various control systems for operating the network and services. Within this context, terms like “policy”, “control-loop”, and “autonomic decision-making” are used.

A policy governs the choices behavior (decisions) in a system. The functionality of a system, its invariant part, is called its “mechanism”. Policies are the variant part of a mechanism and are either static or non-static.

While static policies do not change over time or based on conditions, non-static policies change at runtime. There are three different non-static policy types: context--aware, adaptable, and adaptive. A context aware policy changes its decisions using context from its inputs (events), its own context, or outside information called external context. An adaptable policy changes its decisions based on any external stimuli (including context), such as with a particular configuration or parameters. An adaptive policy changes its decision-making behavior based on internal stimuli, for instance history of decisions or learning from previous decisions.

Adaptable and adaptive policies can also be based on Artificial Intelligence (AI) techniques, which are then typically categorized as intelligent policies. A policy has input and output interfaces, which are translated into interfaces towards some triggering mechanism for input and some actioning mechanism for output, both of which are outside the policy system. Consequently, one can connect any policy with any triggering system (and concrete protocol and communication) and any actioning system. Extending the interfaces for feedback is also possible. This is important especially in closed feedback control loops and can help to build self-stabilizing policy systems.

A closed-loop control system is an essential feature of the automatized work process. In its simplest form, within a closed-loop control system, a controller compares a measured output value of a system with a desired value. In case of a mismatch between these values, the controller decides on the actions needed to achieve the desired value by updating or changing its decision-making policy. Hence, closed-loop control is the enabler of adaptive decision-making. This process does not require any manual input or control and therefore, leads to automation.

Closed-loop control ensures that a deviation from a desired value is mitigated by updating a system’s policy, which is known as autonomic decision-making. From here, policies adapt to evolving or varying changes in the system. This adaptation can be realized by enhancing the capabilities of the policy itself or by automatically re-authoring a policy. Three aspects are important to achieve autonomic decision-making:

- *A policy should use a decision-making approach rather than a decision selection approach*
- *A policy should be able to use contextual information for its decision-making approach*
- *A policy should be able to change its decision-making process at runtime.*

Once these features are enabled in a dynamic system, automation can be achieved, and human intervention can be minimized, which is the anticipated goal of automation at the 5G network edge. As a result, automation in general will continue to advance rapidly in areas where there are real and tangible benefits (e.g., manufacturing, automotive, telecommunications, etc.). For instance, integrating AI advances into automation will enrich the human experience, such as allowing people to have enriched dialogue with personal devices and obtaining advice and guidance. AI continues to make steady progress in areas of speech recognition, decision making, and visual perception.

Automation is therefore needed to dynamically control (via control loops) and optimize (based on autonomic decision-making with adaptive policies) the heterogeneous networks of tomorrow. To support such automation, AI and Machine Learning (ML) are key ingredients. Figure 1 below demonstrates how autonomic decision-making can be impacted by external goals and context.

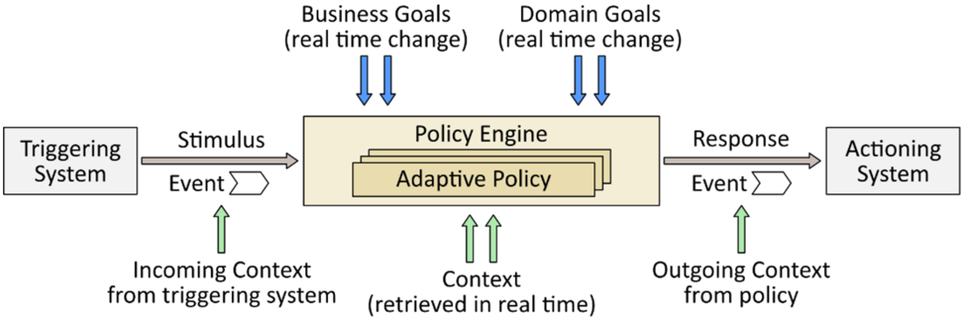


Figure 1 Autonomic decision-making driven by external goals and context.

There are some key advantages to automation in general. First, there will be an overall improvement in efficiency. Software agents are good at executing repetitive tasks to streamline production output, reduce or eliminate human errors and deliver high quality of assurance. AI data, collected and learned at the edge, focuses on the environment at that specific local edge and dynamic policies would adapt accordingly at that specific edge location. Second, shifting to automation and AI will mean technicians are working fewer hours in dangerous conditions, which will decrease workplace injuries. Third, intelligent automation will also provide interesting opportunities for workers to focus on more complex and innovative tasks.

Additionally, automation at the edge can be used to target workloads that are sensitive to stringent timing latencies. In this regard, context information that is exchanged between devices and agents deployed at the edge can help to optimize the workload-resulting in the reduction of completion time or improvement of algorithm performance.

Automation at the edge can be further improved with 5G's drastically increasing capabilities in data transfer rate, capacity, device density, and reduced latency and energy consumption. This powerful combination enables a wide range of use cases that will mark a significant transformation in our lives. From smart homes, smart cities, smart cars to Augmented Reality/Virtual Reality (AR/VR), 3D video, and e-health, the 5G network will become a ubiquitous and pervasive layer that touches every aspect of our daily lives.

In the meantime, innovative industrial applications enabled by 5G are being utilized to monitor, alert, diagnose and control activities across manufacturing, energy, utilities, transportation, smart grid, security, and public safety. The use of Software Defined Network (SDN), Network Function Virtualization (NFV), and edge computing in 5G increases flexibility by dynamically scaling resources, offering compute capability near the devices and exposing real time network measurements for the introduction of automation at the edge as well as AI-based decision-making.

Automation at the edge can facilitate new use cases that add value to the end customer. Near-real time analytics at the edge can provide timely insights to optimize end user performance by prioritizing radio resources automatically. If the Service Level Agreement (SLA) deteriorates or starts drifting, new resource management policies can be automatically created and applied to the system without human intervention. Another example where automation at the edge can be beneficial is to apply reinforcement learning for problems where the modelling of the environment is challenging or even currently impossible for instance, in network slice admission control or for problems where the optimal decision-making policy is not known due to the lack of a master/genie and must be intelligently approximated via trial and error.

Lastly, reinforcement learning algorithms need to evaluate the entire state of network resources, which might lead to scaling problems with increasing network size, and intelligently admit or reject these creation requests based on knowledge developed through exploration and exploitation of different decision options. Nevertheless, a superior level of "thinking" and "learning" will be needed to complement network automation on a path to realize the zero-touch network vision.

1.2 State of the Art and Industry Landscape

The RAN, Transport and Core ecosystem processes and technologies are deeply rooted in many standards organizations. These bodies establish standards that deal with automation and edge and include Zero-Touch Network and Service Management (ZSM), NFV Management and Orchestration (MANO), and Multi-Access Edge Computing (MEC). We provide a brief overview of the state of the art in edge automation and related standards and technologies and highlight 3rd Generation Partnership Project (3GPP) standards, such as Self-Organizing Networks (SON), Network Data Analytics Functions (NWDAF), as well as industry efforts like O-RAN and Tele-Management Forum (TMF), the European Telecommunications Standards Institute (ETSI) ZSM standard, and open-source activities.

1.2.1 Edge Automation related 3GPP Standards

As 5G networks are intended to support various new services such as IoT, cloud-based services, industrial control, autonomous driving, mission critical communications, etc. with ultra-low latency and high data capacity requirements, the 5G system architecture [1] supports edge computing to enable such services by applications that are hosted closer to the user equipment's (UE) access point of attachment in order to reduce the end-to-end latency and the load on the transport network. Additionally, edge computing deployment scenarios and use cases have been defined to guarantee end-to-end service requirements and discuss potential deployment solutions [2].

Furthermore, 5G system enhancement for edge computing capturing a reference architecture, connectivity models, procedures for supporting edge computing [3] as well as enhancements of edge computing management [4] have also been introduced. Finally, from the application point of view, a technical specification that provides an application layer architecture and related procedures for enabling edge applications over 3GPP networks have been defined [5]. Moreover, studies on architecture enhancements for 5G systems to support network data analytics services [6], enablers for network automation for 5G systems [7], enhancement of Management Data Analytics (MDA) [8], [9], enhancement for data collection for 5G New Radio (NR) and Evolved-Universal Terrestrial Radio Access New Radio Dual Connectivity (EN-DC) [10], [11], [12].

1.2.1.1 Self-Organizing Networks

The concept of SON plays a major role and is an integral part of legacy mobile radio access networks. SON is an automation technology to enable simpler and faster planning, configuration, management, optimization, and healing of the mobile network. SON is commonly divided into three architectural types: centralized SON (C-SON), distributed SON (D-SON), and hybrid SON (see Figure 2). C-SON functions are typically concentrated closer to higher-order network elements where network management systems are located to potentially allow

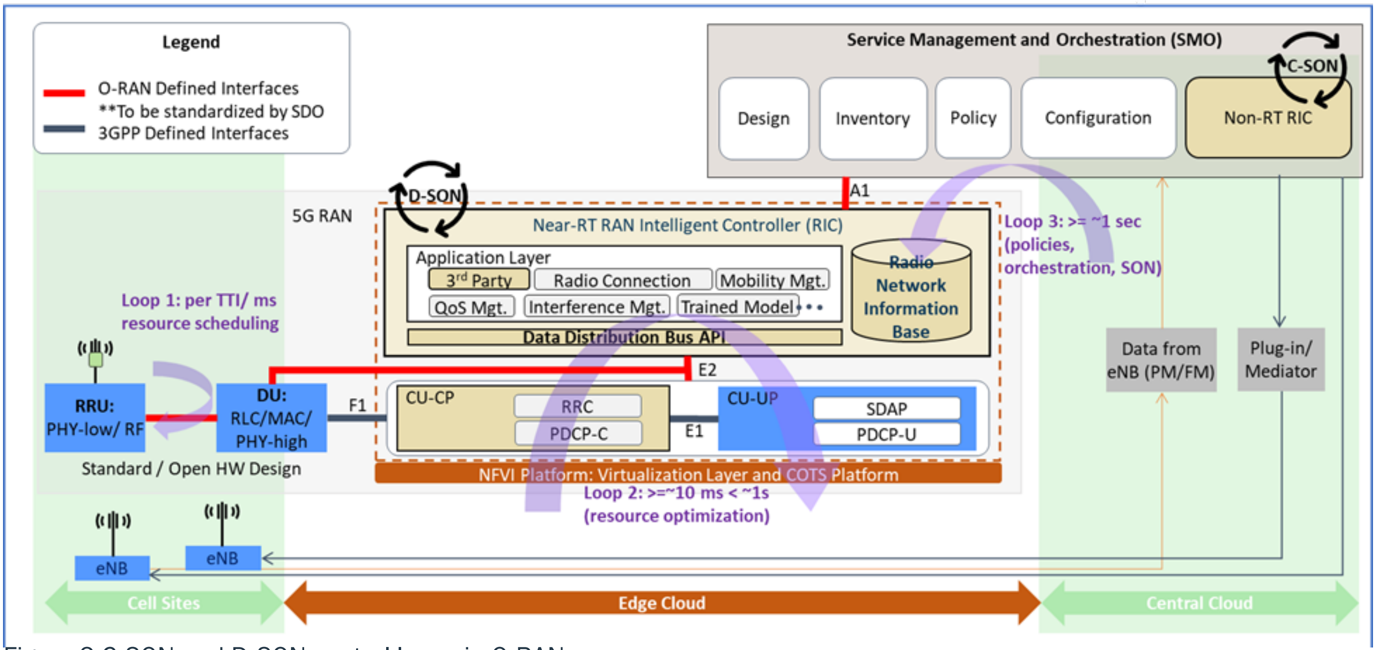

Legend
O-RAN Defined Interfaces
**To be standardized by SDO
3GPP Defined Interfaces
Service Management and Orchestration (SMO)
C-SON
Design
Inventory
Policy
Configuration
Non-RT RIC
A1
D-SON
5G RAN
Near-RT RAN Intelligent Controller (RIC)
Application Layer
3rd Party
Radio Connection
Mobility Mgt.
QoS Mgt.
Interference Mgt.
Trained Model
Data Distribution Bus API
Radio Network Information Base
Loop 3: >= ~1 sec (policies, orchestration, SON)
Loop 1: per TTI/ ms resource scheduling
E2
Data from eNB (PM/FM)
Plug-in/ Mediator
RRU: PHY-low/ RF
DU: RLC/MAC/ PHY-high
F1
CU-CP
RRC
PDCP-C
E1
CU-UP
SDAP
PDCP-U
Standard / Open HW Design
NFVI Platform: Virtualization Layer and COTS Platform
Loop 2: >=~10 ms < ~1s (resource optimization)
eNB
eNB
Cell Sites
Edge Cloud
Central Cloud

Figure 2 C-SON and D-SON control loops in O-RAN.

the autonomous control of a larger number of network elements and their coordination. In D-SON, functions are distributed among various network elements, including the edge, to potentially autonomously control various SON features, such as load balancing or antenna tilt configuration in a closed loop environment. A hybrid SON is a mix of C-SON and D-SON.

With the emergence of enhanced AI/ ML techniques, enhancements of SON with cognitive features are gaining momentum with the introduction of more agility in the network achieved by software/hardware decoupling, virtualization, increased compute, and decoupling of the protocol stack. While research in this context is ongoing, novel concepts and solutions are introduced and explored to handle the complexity of 5G networks with zero-touch network optimization and real-time problem solving.

1.2.1.2 NWDAF

On the 5G core network side, network automation and data analytics have been enabled with the introduction of the NWDAF in 3GPP Release-15. These operations have also been enhanced in subsequent releases [13]. NWDAF was introduced to provide analytics to 5G core network functions and to Operations Administrations and Management (OAM). Network policy decisions are made based on network analytics, which allows the Policy Control Function (PCF) or any other 5G core Network Function that has subscribed to NWDAF output to perform decisions, such as to update and/or adapt a policy by considering the analytics information provided by the NWDAF.

The PCF request may be triggered based on a request from other network functions, modification requests, or any changes in the network. The following analytics are relevant for policy decisions: “Load level information”, “Service Experience”, “Network Performance”, “Abnormal behavior”, “User Equipment (UE) Mobility”, “UE Communication”, “User Data Congestion”, “Data Dispersion”, and “WLAN performance”. As illustrated in Figure 3, such input data and analytics are collected by the NWDAF to make policy decisions. The output of the NWDAF serves the network functions and the OAM to decide how to use the data analytics provided by the NWDAF to improve the network performance, which reflects a closed-loop control system framework.

NWDAF is expected to have a distributed architecture providing analytics at the edge. Currently, studies on network automation enhancements are ongoing and focus on topics such as how to enable real-time or near-real time NWDAF, how to enable NWDAF-assisted user plane optimization and the interaction between NWDAF and AI model and training services. Within a network, NWDAF can be implemented in a centralized manner, distributed manner, or a hybrid of the two. When NWDAF is implemented in a distributed or in a hybrid manner, it is possible that distributed instances of NWDAF are placed at the edge to help with edge automation use cases. In this manner edge data would be stored, processed and analyzed locally. This would also help reduce latency and the overhead of carrying data across the network.

1.2.2 Open Radio Access Networks

Open Radio Access Networks are often abbreviated as “Open RAN”, “OpenRAN”, as well as “O-RAN”. For the purposes of this white paper, “Open RAN” (note the space between the words) refers to open and interoperable interfaces within and between various subcomponents of the RAN. Hence, it refers to the movement in wireless telecommunications to disaggregate hardware and software and to create open interfaces between them.

OpenRAN, on the other hand, refers to one of the two groups within the Telecom Infra Project (TIP), i.e., the OpenRAN project group, which is an initiative to define and build previous

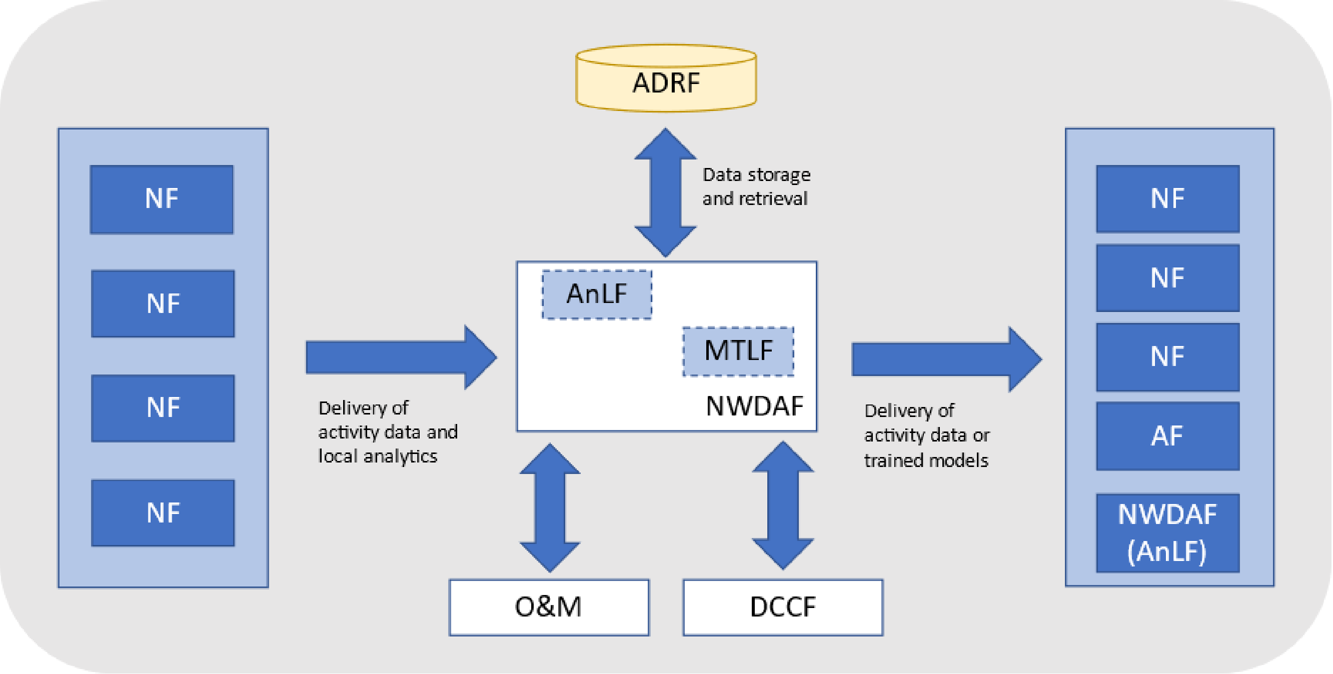

ADRF
Data storage and retrieval
NF
NF
NF
NF
Delivery of activity data and local analytics
AnLF
MTLF
NWDAF
Delivery of activity data or trained models
NF
NF
NF
AF
NWDAF (AnLF)
O&M
DCCF

Figure 3 General framework of 5G (core) network automation.

generations RAN solutions based on general-purpose, vendor-neutral hardware and software-defined technology, or the OpenRAN 5G NR project group which focuses on 5G NR.

O-RAN (or alternatively 'ORAN' in some cases), is an acronym for the O-RAN Alliance, which publishes new RAN specifications, releases open software for the RAN, and supports its members integration and testing of their implementations. Throughout this white paper, the abbreviations will be used based on these definitions.

Open RAN disaggregates the RAN. In O-RAN this is done by using open interfaces and incorporating the concept of RAN Intelligent Controllers (RICs) that can host smart applications (i.e., rApps and xApps) and perform radio resource management functions at a per UE level. These controllers extend new management and control interfaces to the RAN ecosystem (namely O1, O2, A1 and E2). Edge cloud servers will typically host the Near-Real Time RIC while centralized data centers will typically host the Non-Real Time RIC. Automation at the edge will involve both types of controllers, i.e., RICs, and orchestration engines. O-RAN is defining or clarifying the usage of the interfaces between the different parts of the RAN. These parts and their respective interfaces are identified and clarified in Figure 4.

- *Orchestrator and RIC component – A1 interface.*
- *RIC and Centralized Unit/ Distributed Unit (CU/DU) – E2 Interface.*
- *CU-CP (Control Plane) and CU-UP (User Plane) – E1 Interface.*
- *CU-DU – F1 interface.*
- *DU-RU (Radio Unit) – Open FrontHaul.*
- *Orchestrator and Cloud Platform (O-Cloud) – O2 Interface.*

1.2.3 Tele-Management Forum (TMF)

TMF has several projects focused on edge automation. For example, the Catalyst project [15] is constructing a standardized Edge Compute-as-a-Service (ECaaS) for realizing zero touch edge solutions. Another project, AI Operations (AIOps) [16] is targeting how AI can drive closed-loop service assurance in communications service provider's network services.

1.2.4 ETSI Zero Touch Network and Service Management (ZSM)

ETSI aims to deliver a framework within the ZSM industry specification group that enables automation of the end-to-end network management with minimal to zero human intervention. The scope of ZSM includes the RAN, transport, core, NFV, SDN, legacy, and everything in between, that makes up a communication service. The ZSM framework facilitates collaborative management interactions between all elements and all layers of the network enabled by closed-loop automation, AI, adaptive ML, and cognitive technology. The architecture is service-based, modular, flexible, scalable, and is defined in [17].

The ZSM framework defines standards interfaces that enable interactions between management domains, coordination between different closed loops, and interactions between AI components and closed loops, and hence, provides the glue allowing interactions of various components, enabling the autonomous management of the end-to-end network.

The complexity of the resources at the 5G edge can be abstracted by a management domain in the ZSM architecture. The interactions between management domains and the end-to-end service domain are defined by ZSM. Management domains in the ZSM architecture allow the separation of management concerns [18], and

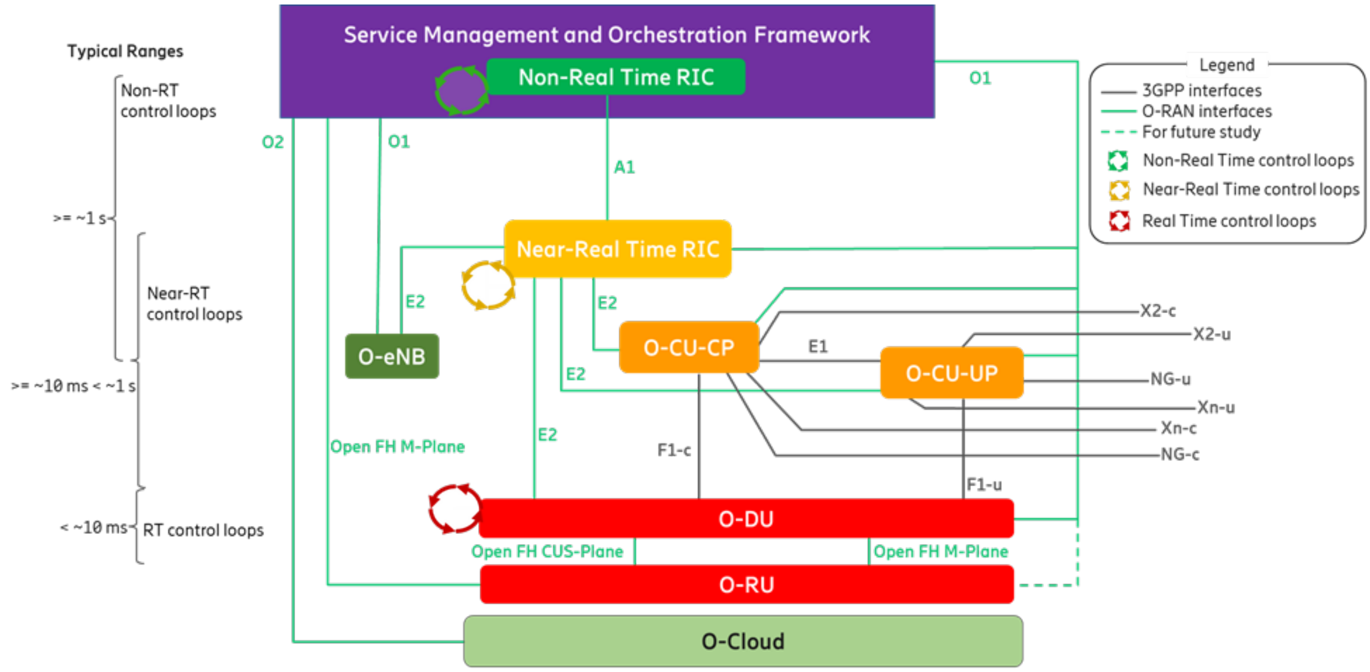


Figure 4 O-RAN architecture [14].

can consist of components in the 3GPP RAN, 3GPP Core, Transport domain, NFV components, O-RAN, Open Network Automation Platform (ONAP), etc.

1.2.5 Open Source

There are several open-source related projects relating edge clouds and automation. The Linux Foundation (LF) is driving several initiatives to bring together an ecosystem of open-source items. These items range from RAN orchestrators, policy guided RAN controllers to mobile packet core. LF aims to establish an open, interoperable framework for edge computing powered by AI/ML. Fostering cross-industry collaboration across IoT, Telecom, Enterprise and Cloud ecosystems is key to deliver value to end users.

At the 5G edge, an open-source software for carrier-scale edge computing applications that run in virtual machines and containers to support reliability and performance requirements has been introduced. Within the LF community, LF Edge, has several projects targeted for the edge cloud. LF Networking is leading a community-driven integration and proof of concept involving multiple open-source initiatives to show end-to-end use cases demonstrating various implementation architectures for end users. The 5G Super Blueprint covers RAN, edge, and core for enterprises and verticals. It spans a broad variety of use cases including 5G, AI, edge Infrastructure-as-a-Service/Platform-as-a-Service (IaaS/PaaS), IoT. The goal is to offer flexibility to scale edge cloud services quickly, to maximize the applications or subscribers supported on each server, and to help ensure the reliability of systems that must be functioning at all times.

1.3 Envisioned Features and Key Technologies

This section describes envisioned features and key technologies required for the implementation of edge automation and its enhancement as well as which can benefit from edge automation.

1.3.1 Distributed Data Collection, Normalization, and Real-Time Processing

Big data management is a challenging area of research. The problem becomes even more of a challenge when the data must be collected and processed to produce control signals that are sent back to the network in near real-time. A mix of data streaming technologies, in-memory and on-disk storage, and compute facilities will need to be located close to the edge of the network to minimize the latency and maximize the bandwidth of such processing. Standards for collecting operational and control data from the network (such as O-RAN Alliance's O1 and E2 interfaces [14]) are a must-have, but platforms must also provide the ability to implement customized normalization and machine-learning procedures to meet highly variable and rapidly changing business needs.

An important concern is how to meet the data collection and processing requirements in a manner that retains the ability of network operators and other stakeholders to mix-and-match solutions from different vendors, as well as the open-source community, so that they retain ownership and control of the data instead of being locked into proprietary database technologies. An open, high performance data streaming solution such as Kafka [19], Pulsar [20], and Rabbit-MQ [21] will have a role to play in any such architecture, along with data lake technologies that provide easy on-ramps and off-ramps for collected data and co-located compute facilities such as Spark [22] that will minimize the need to transport data from place to place when executing normalization activities or building insights.

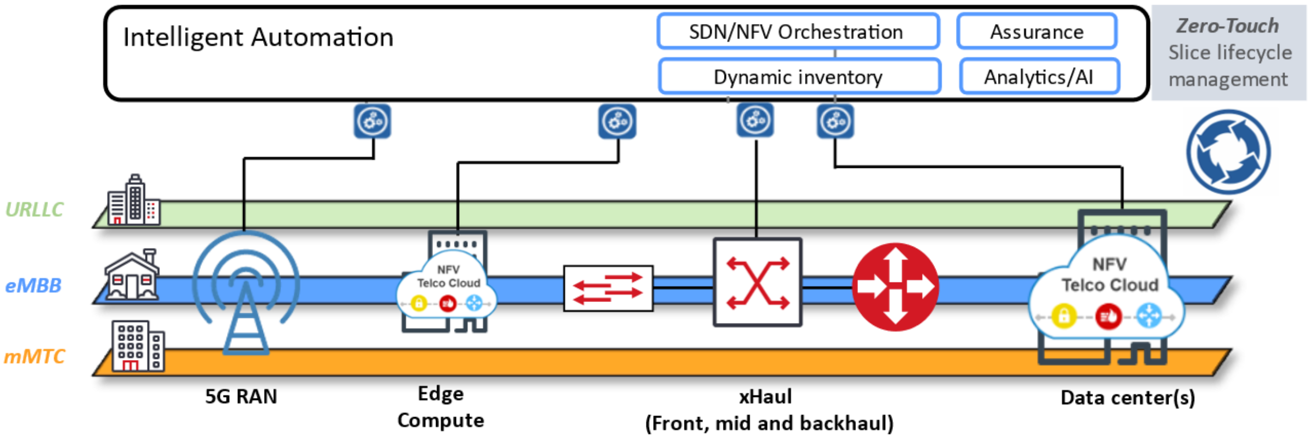


Figure 5 Intelligent Automation solution architecture.

1.3.2 Context Discovery and Situational Awareness

The ability to turn data from multiple sources into knowledge and exposing it as actionable insights is one of the key elements of understanding behavior while applying it to build situational awareness. This requires using computer vision, enhanced location services, network APIs exposing congestion data, and connection availability.

In addition to that, Application Program Interfaces (APIs) pairing application/device needs with matching available edge locations are necessary to offload and connect clients to best fitting edge cloud nodes. This should be based on geography, compute, storage, memory, and power requirements as well as other contextual insights. This bi-directional exchange of information should allow edge-enabled applications to be context-aware yet location independent.

Security policies can also be applied based on network and location insights. For example, an autonomous car would turn on Virtual Private Network (VPN) for a public Wi-Fi connection in a shopping mall before sending telemetry data on an identified fault versus when connected to a secure 5G slice on the road.

1.3.3 Network Slicing and Dynamics

A network slice is a logical end-to-end network defined over a common infrastructure comprised of physical and virtual resources. As the definition implies, a network slice supports end-to-end network connectivity for end-users, humans, and machines. Therefore, most of the actions performed on the traffic in a 5G network will take place at the edge, where content is created and consumed. Each slice is virtually isolated from another and is designed according to the specific needs of the application or end-user.

With network slicing being critical to the successful delivery of 5G services, mobile and wholesale operators alike should be able to plan, design, and activate thousands of customized network slices for their customers very quickly. They also should be able to modify and scale a slice up or down to address changing performance demands for optimized end-user experiences.

1.3.3.1 Intelligent Edge Automation for 5G Slicing

Intelligent, analytics-driven automation is more than just automating manual processes. It is the ability to take input from several sources, such as the network itself, analyze that input to generate actionable insights, and then execute upon them via intelligent actions. This type of automation is required in the complex end-to-end setting of 5G architecture. Intelligent automation is required at the edge of the network where most of the services will be provisioned, including network slices. Furthermore, by incorporating analytic and AI capability into edge automation process, critical insight can be extracted from network measurements and be used to generate optimal, dynamic slicing configurations allowing rapid service deployment and providing a framework with business agility and flexibility.

With intelligent edge automation, operators can implement zero-touch slice lifecycle management, which includes automating the design, creation, modification, and monitoring of end-to-end network slices as well as the provisioning of underlying resources to a slice, as and when required. The intelligent edge automation solution should also support the scaling and orchestration of network resources for 5G Core, xHaul (combination of backhaul, midhaul, and fronthaul), and RAN, along with the creation and operation of network slices. Figure 5 illustrates such an intelligently automated solution architecture.

Intelligent automation software is the key to the proper placement of Cloud-native Network Functions/Virtual Network Functions (CNFs/VNFs) within a mobile network and enabling Mobile Network Operators (MNOs) to maximize the utilization of network resources by re-allocating unused resources to other slices. With its advanced analytics, AI, ML, and automated orchestration features, an intelligent software solution enables the creation of a self-driving, self-healing, and self-optimizing 5G network with zero-touch capabilities.

1.3.3.2 Zero-Touch Network Slice Life Cycle Management

The massive number of network slices that will be required and the speed at which the services will have to be managed, make it impossible for mobile and wholesale operators to accomplish this process manually. Zero-touch automation capabilities are a necessity to efficiently manage the lifecycle of network slices. Figure 6 demonstrates what the network slice lifecycle management would look like in a 3GPP standardized network.

An intelligent edge automation software solution can automate the entire Lifecycle Management (LCM) of network slices and includes support for the Global System for Mobile Communications Alliance (GSMA) Slice Template for the initial design phase. Using this solution, operators can plan, design, and create new network slices, monitor and modify a slice to meet Quality of Service (QoS) or customer requirements, deactivate it when no longer required, and release associated resources back into a federated inventory system.

1.3.3.3 Rapid Service Deployment

As 5G adoption increases across new industry verticals, explosive growth in the number of services is expected. Users will demand faster deployment of new services that meet their specific end-to-end QoS requirements. An optimal 5G automation solution featuring network-wide correlated analytics, automated orchestration, and zero-touch capabilities can help operators reduce the time to plan, design, and deploy new services across the multi-layer, multi-vendor, multi-domain network from weeks or months to a few minutes. The rapid deployment capability reduces the time to market for service offerings, increases customer satisfaction, and shortens the time to revenue for the operator.

1.3.3.4 An Open and Standards-Based Solution

A 5G automation solution should support a wide range of industry standards initiatives, open-source projects, and Open APIs from the TMF, Mobile Ecosystem Forum (MEF), Open Networking Foundation (ONF) and others. Aligned with this approach, a zero-touch slice lifecycle management solution should support the Communication Service Management Function (CSMF), the Network Slice Management Function (NSMF), and the Network Slice Subnet Management Function (NSSMF) standard, as detailed in the 3GPP specifications, also referenced in the ETSI NFV MANO framework.

In order to adhere to emerging 5G standards, increase flexibility, and support a multi-vendor network, an intelligent analytics-driven edge automation solution should be designed and developed as a cloud-native application built upon a containerized, microservices-based architecture.

1.4 Requirement Analysis

Automation itself and the techniques to automate processes are not different in the edge compared to other places in the network. Edge automation will require event processing, analytics, closed loop control, and policies the same way as Transport or Core. However, the circumstances at the edge will require a specific way of applying these techniques, with some degree of uniqueness. Devices (UEs and others), sessions (network and application) and users now consume resources such as compute, storage, sensing, services, and applications close to their current location in the network (or at least definitely closer than before). This requires a very different approach to automate resource management, while also maintaining “classic” (non-edge) usage scenarios:

Edge resource management: This is usually the first target for automation and should cover all relevant zones (device, premise, access, metro) and must guarantee that availability and quality access to all resources for any given number of consumers (devices, sessions, and users) active in those zones. In static scenarios, where neither the consumers nor the resources or the network conditions change, “classic” automation techniques should be sufficient. Any dynamicity – for example devices moving in and out of edge zones, sessions migrate between devices, compute and storage being changed, applications being updated or moved – can result in rather complex scenarios.

Edge infrastructure: compute and storage, or clouds, are heterogeneous hardware and software environments. Automation will need to be able to deploy, monitor, and repair applications and sessions using them on a wide range of clouds simultaneously. Edge automation should be independent of the underlying cloud providing multi-cloud mechanisms.

Location-aware vs. location-dependent: One goal of edge computing is to provide resources at specific locations close to the consumers. However, said consumers

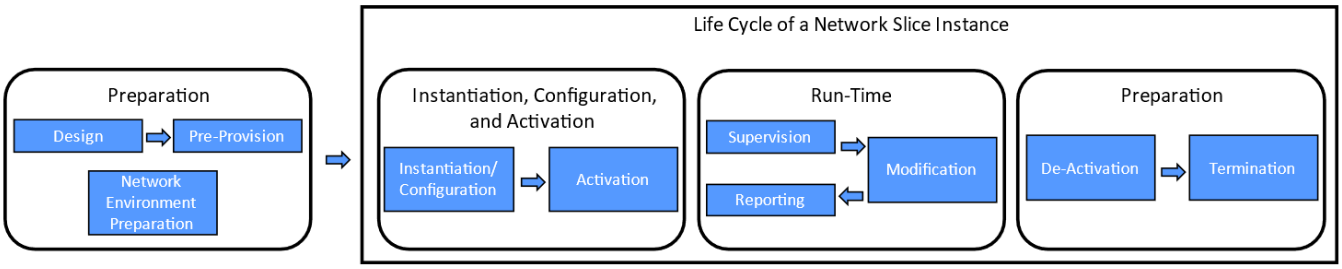

Life Cycle of a Network Slice Instance
Preparation
Design
Pre-Provision
Network Environment Preparation
Instantiation, Configuration, and Activation
Instantiation/ Configuration
Activation
Run-Time
Supervision
Reporting
Modification
Preparation
De-Activation
Termination

Figure 6 3GPP network slice lifecycle management [23].

can and will move and are probably not interested in accessing these specific locations but resources with defined QoS. Edge automation should be location-aware where location defines the automation target yet not be location-dependent.

Mobility management and application-dependency: a core property of a mobile network involves facilitating consumer and, to a certain degree, resource mobility. A common scenario is a mobile UE connected (via “long” tunnels or slices) to relatively static and aggregated endpoints (e.g., packet gateways in the Core at a few central locations). At the edge, the connections are “short”, endpoints are no longer aggregated centrally, and the nature of end points changes from some sort of gateway to applications. Edge automation must cope with application dependency while supporting consumer and resource mobility, for example, a hand-over of an application between different edge zones to follow a connected UE.

Fine-grained time-sensitivity: one reason to use edge computing is to support time-sensitive services with QoS impossible to achieve otherwise. There are many time-sensitive properties, low latency being one of them. Consumers will define their time requirements and the edge will have to cope with very fine-grained requests. Edge automation should create, maintain, and remove resources based on many specific time requirements rather than a few generic service classes or QoS parameters.

1.5 Architecture Direction

The telecommunications industry is amid several transformational shifts, including the adoption of 5G technology as well as momentum in the marketplace from closed, single vendor RAN to ORAN—standardized, software-defined interfaces that are open and interoperable. While these are first steps towards introducing automation and control into the 5G network, further (architectural) enhancements must be introduced to achieve automation at the 5G edge. To this end, an indicative end-to-end architecture is shown in Figure 7 that encompasses an edge automation and control framework, service management and orchestration layer, intelligent edge applications layer and a distributed edge infrastructure layer.

An Edge Automation and Control framework is envisioned to include various components and to support standardized networks and entities, e.g., 5G NR, private 5G, Open RAN RIC etc. Such a framework should integrate a wider architecture aiming for end-to-end observability, control, and optimization of Open RAN and the 5G edge ultimately extending towards transport and core. Note that Open RAN is one of the many potential implementation baselines for an edge automation framework.

An edge automation and control framework may include:

- *An Open RAN RIC (intelligent RAN control),*
- *An Intelligent Access Controller (multi-mode access support, access control),*
- *An Intelligent Core Controller (control at the core and transport network e.g., SD-WAN),*
- *An Intelligent Edge Controller (main coordination and optimization function),*
- *Edge Intelligence (for recursive monitoring, intelligent decision-making, automation, and optimization)*
- *A Time-sensitive communications data plane that can be dynamically programmed as driven by application needs*

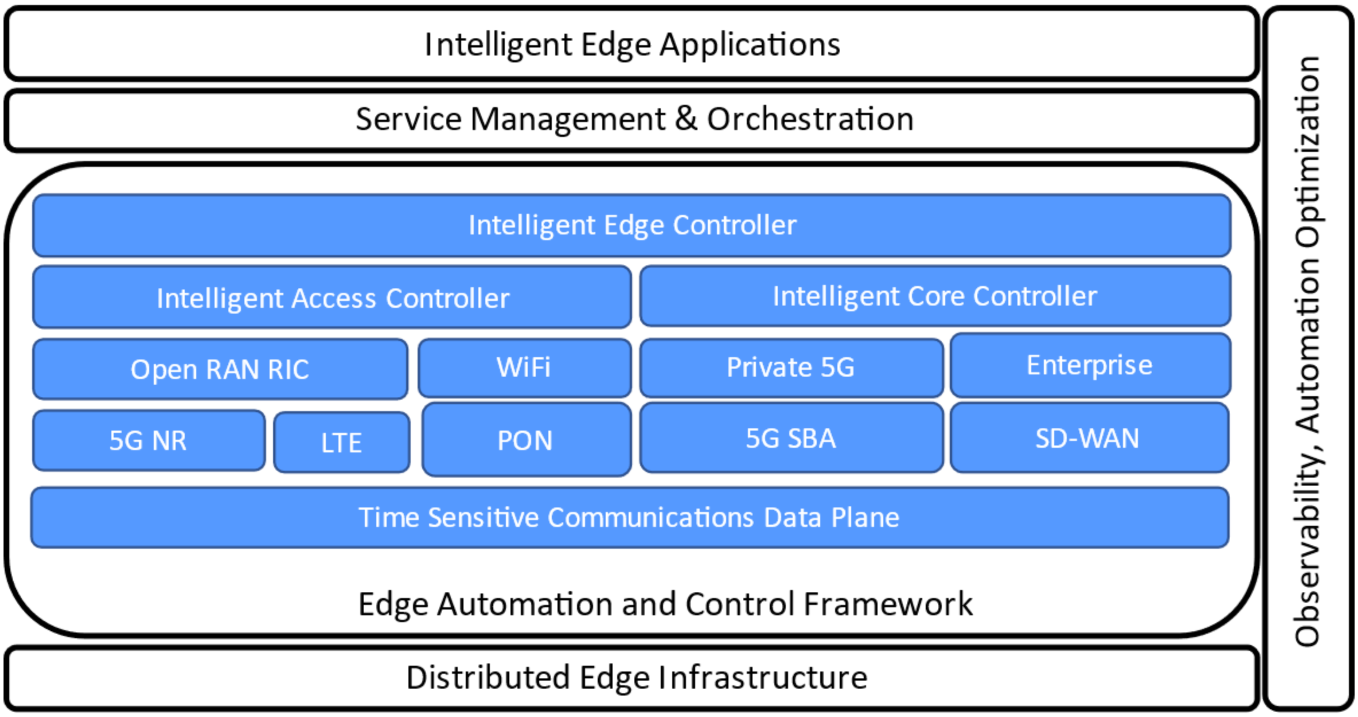

Intelligent Edge Applications
Service Management & Orchestration
Intelligent Edge Controller
Intelligent Access Controller
Intelligent Core Controller
Open RAN RIC
WiFi
Private 5G
Enterprise
5G NR
LTE
PON
5G SBA
SD-WAN
Time Sensitive Communications Data Plane
Edge Automation and Control Framework
Distributed Edge Infrastructure
Observability, Automation Optimization

Figure 7 Edge automation and control framework.

Intelligent Edge Controllers with Time-Sensitive Applications

The introduction of edge controllers for the end-to-end control and automation to support various emerging application including time-sensitive applications, are expected to provide the required degree of control, management, and orchestration for the 5G edge automation. However, such a framework will have its own challenges. Two main challenges in this context are security and a new marketplace.

Security: Through enhanced visibility of key interfaces over open interfaces, AI/ML powered tools provide automated security analytics. As Open RAN evolves to incorporate new capabilities such as real-time and non-real-time RIC that use pre-trained AI models, new threats emerge related to algorithmic conflicts, adversarial attacks, and data exfiltration. Open RAN is evolving towards providing real-time security at the edge. Traditional implementations might take additional time and manual diagnosis to curb such threats possibly leaving the network exposed.

O-RAN, with its virtualization, disaggregation, automation, and intelligence, is expected to be a complementary part of 5G's broader progression to greater security:

- *Open interfaces ensure interoperability of protocols and security features*
- *Disaggregation establishes diversity of supply chain*
- *Cloud-native applications provide isolation*

An Open RAN architecture specified by the O-RAN Alliance, builds upon 3GPP RAN specifications with additional interfaces and functions. However, these additional interfaces and functions introduce additional security risks due to the expanded attack surface. Internal and external attacks can exploit vulnerabilities in the network architecture and cloud infrastructure, while 5G use cases have decreased risk tolerance.

Marketplace: The introduction of intelligent edge applications opens a new and exciting marketplace of applications. Much like the plethora of apps that are available in our cell phones today, the Open RAN marketplace will provide a rich ecosystem for innovators to develop various RAN optimization applications. Diverse types of intelligent edge applications (rApps and xApps) can be developed to optimize both at a macro cell level or surgically at a per UE level. Non-RT RIC applications (rApps) operate at a time granularity of 10s of seconds while near-real-time RIC applications (xApps) operate at a much faster timescale in the order of 10s of milliseconds.

Implementing Open RAN technologies could spur innovation and potentially provide additional benefits of increased flexibility, agility, and resilience in the RAN. Decoupling the hardware and software of the RAN not only creates possible opportunities for new businesses, both small and large, to enter the market, but it also could decrease the probability of the vendor lock-in that can occur. The modular nature also encourages the development of "best-of-breed" solutions due to increased vendor competition. Finally, a disaggregated and open ecosystem could provide resilience and agility benefits.

1.6 System Recommendations for ML-driven Automation

The capabilities introduced by 5G make it possible to take ML applications to the next level, providing a better user experience and open new opportunities, for instance in healthcare, security, and finance applications. However, making those advances means that the underlying system must support unprecedented amounts of data being constantly collected and processed in a timely manner as well as large amounts of computation must be executed promptly on the edge devices or on edge zones (i.e., premise, access, or metro).

To achieve these goals, instead of adhering to the traditional worst-case planning, even the most basic functionalities should be amenable to automation and adapt to evolving situations. To achieve reliable systems and responsible resource consumption and scalability of those systems, we need ML solutions to both automate the design and provide fundamental support for growing ML applications.

1.6.1 ML for Systems

Machine Learning for systems explores how we can leverage machine learning tools and advances to improve systems. Several approaches can help to automate and manage future networks, including advanced 5G networks.

1.6.1.1 Smart Edge Configuration

Every edge device has its own capabilities and resource constraints. For example, a cellphone is battery powered (unless connected to a power source), whereas a smart refrigerator is connected to a constant power supply but may have weaker hardware configuration and capabilities due to profitability considerations. Many such devices may participate in joint data collection and processing tasks and even communicate among themselves. Each such device should have automation capabilities to coordinate an expectations negotiation for different tasks and assess its limitations and ability to execute a given application or task.

For example, a server or an application coordinator may produce sample application benchmarks for each device to execute. Then, according to collected benchmark results, an ML model can be used to determine the suitable hardware configuration and functionality for each participating device. This may lead to an automatic configuration of devices, better use of their resources, and better system utilization.

1.6.1.2 Smart Edge Monitoring

There are several challenges involving edge device monitoring, statistics collection, and different fault and anomaly detection needs.

Statistics collection: When considering each device separately, storage and communication are of major concern. For instance, a device can have automation capabilities to adjust the resolution of its samples (e.g., images), or to recognize what information is more critical and time-sensitive and therefore should be immediately dispatched (e.g., healthcare application).

Anomaly Detection: Automated anomaly detection includes a wide range of applications such as finance, surveillance, health care, intrusion detection, fault detection in safety-critical systems, and medical diagnosis. For example, anomalies in network traffic could mean that a hacked device is sending out sensitive data to an unauthorized destination; anomalies in a credit card transaction could indicate credit card or identity theft; and anomaly readings from various sensors could signify a faulty behavior in hardware or a software component.

Edge devices are usually resource-constrained in terms of compute, communication, and memory, while anomaly detection applications usually require detection of anomalies as fast as possible. Therefore, it is of increasing interest to develop, support and deploy resource-efficient anomaly detection ML models on such edge devices. 5G can offer improved cross-device connectivity which can be leveraged to improve anomaly detection performance, for instance by employing efficient distributed data collection.

Predict Infrastructure Failure: Previously, failed infrastructure meant that edge devices might have lost connectivity and application accesses for the failure or takeover duration, introducing inconvenience and worse. However, considering the extent of future uses and applications, such an infrastructure failure can result in disastrous scenarios, leading to a huge financial damage (e.g., an automatic order to sell or buy stocks) and even a loss of life (e.g., medical and life-supporting applications). Keeping available backups ready to take over may prevent such scenarios. A complimentary and scalable solution is the ability to predict infrastructure failure. A device or neighboring devices' ability to predict upcoming malfunctioning relies on lessons learned from the past. Automating such failure predictions is a big step toward system reliability on a larger scale with life-supporting and critical applications.

Predict Network Overload: Network overload is a known problem. For example, an audience of tens and even hundreds of thousands of people may lose connectivity at a sports event. Predicting such overloads at specific times and locations, either by the network infrastructure or the edge devices, can be used to issue notices alarming users from the possibility of such an event. It may be necessary for some users and application to take this possibility into account and take measures accordingly.

1.6.2 Systems for ML

Running ML applications over edge devices is highly challenging in terms of resource consumption. With the increase of ML applications' demands and usages, available amounts of data, and growth in the number of users, these challenges will only increase. The ability to meet these challenges largely relies on how we operate and perform communication and computation for ML applications.

Systems for ML explores how the system design itself can improve ML performance and resource usage to performance tradeoffs by exploiting domain knowledge. Automation of systems to offer the appropriate tools and configurations to ML applications is an essential step towards better ML practices.

1.6.2.1 Support for ML Applications

ML applications often require a lot of communication, computation, and data collection. For example, in federated learning, participating devices perform potentially computationally-extensive operations on their local data, exchange parameter updates, and must be concerned with the privacy of their data.

Exposing suitable system APIs to such ML applications can off-load many tasks and enable better resource usage and scalability. For example, a system may expose interfaces that will allow for efficient in-network aggregation, privacy (e.g., differential privacy and secure aggregation), and even encoding, decoding, and down-sampling capabilities when an application can indicate that lossy information suffices.

1.6.2.2 Support for Distributed Data Collection

Collecting data from billions of devices and for different applications and use-cases is a major challenge. Having a human expert to fine-tune this process for each use case is impractical. A desirable design goal is to have a system API that supports data collection in a way that allows users and applications to specify which data is relevant and collect only relevant data. For example, an application can indicate through an API metrics by which the system can perform in-network data filtering and aggregation during the collection process.

Another step in automated monitoring and statistics collection is the coordination of such activities over a set of devices. For example, coordinated monitoring and statistics collection of several devices may result in better information and resource usage. Likewise, eliminating redundancy in data collection process executed by many devices can improve both bandwidth and computational resources.

2. 5G Edge Optimization, Intelligence, and Analytics

The world of computing and communication has been going through a fundamental paradigm shift in recent years. Moore's Law has increased both computing and communication capacity of network nodes and devices tremendously. The end user/edge devices are becoming more and more compute-intensive to run sophisticated optimization approaches and/or AI/ML workloads to drive actionable insights at the edge. At the same time many new time-sensitive, mission critical services require distributed, collaborative processing at near edge users as opposed to centralized/cloud-based processing.

The combination of 5G connectivity and AI/ML computing capability at the edge enables more intelligent applications for the network edge nodes and devices as well. Devices with limited processing power can leverage 5G network edge node resources or other more capable devices nearby to gain intelligence. Network edge nodes can also leverage devices to collect real-time data from devices' sensors and create a joint perception of the environment. They can also collaborate in sharing sensor data, AI/ML processing, and coordinate actions thus enable collaborative intelligence.

This shift creates new business opportunities for 5G edge computing and communications technologies as the ability to run AI/ML algorithms on the edge infrastructure that are connected to and can serve edge users. Such edge infrastructures are lucrative for circumventing bandwidth, latency, and cost concerns of cloud computing with the global AI edge chipset revenue forecast to grow to $51.9B by 2025 [24].

2.1 Background

The 5G NR air interface is vastly improved to address the requirements of various emerging use cases. To guarantee QoS requirements for such use cases, other network architectural concepts have been introduced that support management and orchestration, optimization, and AI/ML-based analytics. A new split RAN architecture and edge computing are two significant architectural changes in 5G networks to reduce the overall latency and guarantee QoS.

With the migration to cloud-based 5G networks, there is a need for a collaborative and intelligent approach to optimize the fragmented ecosystem of edge computing. While mobile network operators can open up their network as a distributed cloud to non-telco workloads, enterprises need to optimize their applications for the new distributed architecture, which provides an unprecedented opportunity for distribution and processing of the massive amounts of data and its analytics at the edge. 5G networks allow us to fully exploit edge computing by moving the data collection, compute, and analytics closer to the end points, where data is generated and consumed, rather than sending the data to and from servers in cloud data centers thus essentially leading to significant reduction in end-to-end latency in data analytics and delivery , 5G and beyond edge networks inherently have the intelligence needed for smartly moving, storing, and processing data on the fly.

With both massive data and the processing power with AI/ML capabilities at 5G edge nodes, including mobile edge computing nodes and mobile nodes, distributed learning [25] and collaborative intelligence [26] will become possible and be able to support real-time intelligence and collaboration. Edge nodes can work together to share sensor data from each other to obtain joint perception with collaborative AI/ML in a dynamic environment and to act together with group decisions for the improvement of efficiency, productivity, and safety for various 5G edge applications like intelligent transportation system, smart factory, smart energy, and smart homes.

2.2 State of the Art and Industry Landscape

5G calls for a new level of flexibility in architecting, scaling, and deploying telecom networks. Cloud technology offers new innovative alternatives for such RAN deployments complementing existing proven purpose-built solutions. Cloud RAN refers to realizing RAN functions over a generic computing platform instead of a purpose-built hardware platform and managing the RAN application virtualization using cloud-native principles. Cloudification of the RAN begins with running selected 5G RAN network functions in containers through Commercial Off-The-Shelf (COTS) hardware platforms. It starts with the control plane and user plane in the CU and continues with latency-sensitive radio processing functions in the DU. By pushing distributed units to the edge, mobile networks provide low latency services and a pool of processing and other computing resources to support mobile user processor off-loading use cases [27].

2.2.1 The Industrial Cloud Ecosystem

The combination of different sizes of cloud data centers, namely edge zones, at global, national, local/regional, and potentially access locations are integrated into the network and operated by a central orchestration and management system. The exact specification of the infrastructure on different sites may depend on the use cases and applications onboarded. In addition, there can be several infrastructure providers on the same site and these distributed computing resources, including MEC nodes and edge devices like the Roadside Unit (RSU) which can provide AI/ML services [28] or enable collaborative intelligence.

Many companies participate in the ecosystem, from hardware vendors, platform companies to applications developers, System Integrator (SI) companies, and Cloud Service Providers (CSPs). Two other key players in the ecosystem are the Hyperscale Cloud Providers (HCPs) and Operational Technology (OT) vendors.

Hyperscale Cloud Providers, such as AWS, Microsoft Azure, Google, and AliCloud all have a core business to provide cloud infrastructure and platforms. They own application ecosystems with thousands of contributing developers and can serve multiple enterprises in several sectors globally. HCPs are keen to be ecosystem drivers for edge computing. As a part of this approach, HCPs start offering on-premises compute, storage, database, and other services run locally on dedicated platforms provided to the customers at the edge. This approach is ideal for workloads and applications that require low latency and access to on-premises systems, enabling edge workloads to extend their reach to the cloud as needed.

OT vendors have IoT platforms and applications, supported by edge computing components. Some examples of these companies include Siemens, General Electric, BMW, and ABB. They have strong enterprise relationships, especially in the manufacturing sector. Companies looking to do an intelligent manufacturing deployment are likely to partner with an OT vendor to a certain extent. OT vendors establish relationships with HCPs for global deployments of their solutions, access to the application development ecosystem, and an environment to create and deploy their applications.

SI companies have a wide range of capabilities to address enterprise pain points related to solution implementation and integrating offerings from

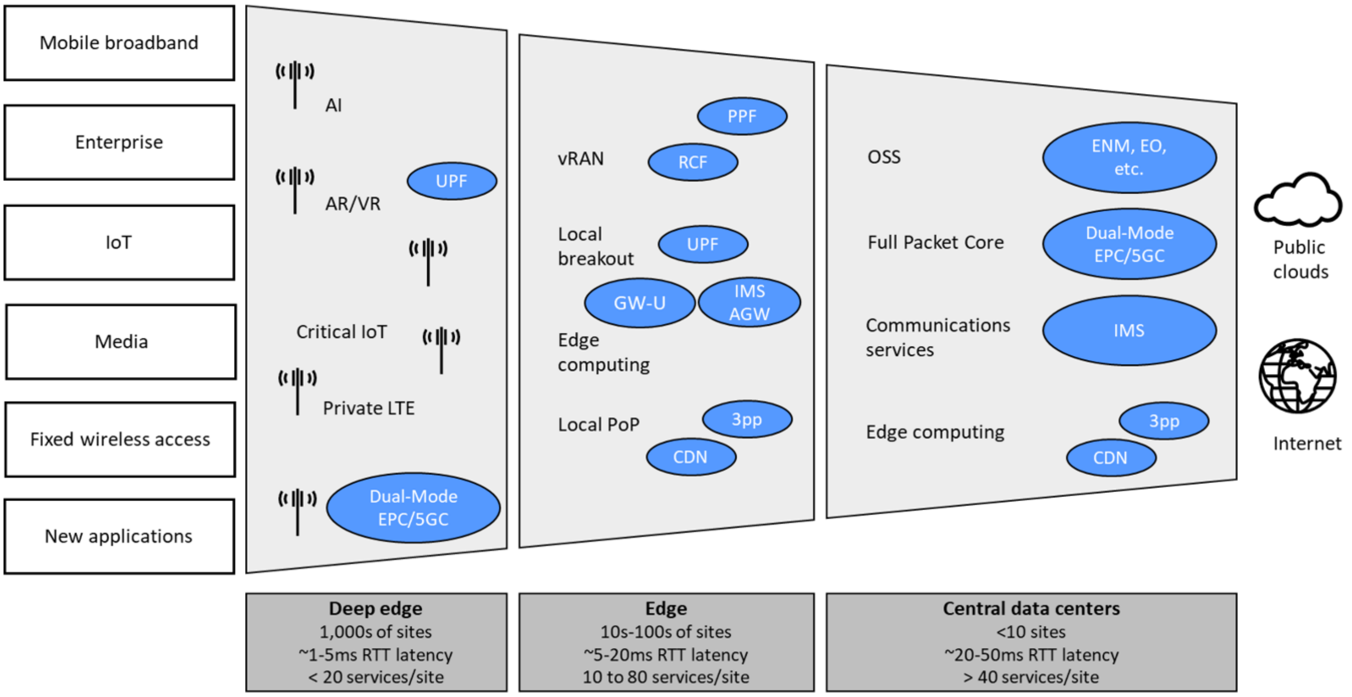


Figure 8 Distributed edge to meet the need of 5G applications.

different ecosystem companies. SI companies can be both global and local and are likely to be present in most solution implementations in one way or another. Apart from specialized SI companies, other companies can also take an SI role in solution implementation, for example, OT vendors or HCPs.

Figure 8 provides a better look at the overall cloud and edge ecosystem, as it pertains to 5G applications.

2.2.2 Centralized, Decentralized, and Hybrid Network at Edge

Edge intelligence requires non-traditional hierarchy, a hybrid and distributed network for edge devices. The introduction of 5G NR sidelink [29] and Integrated Access and Backhaul (IAB) [30] in 3GPP standards has enabled both direct device-to-device and multi-hop communication at the edge of networks. They enable a hybrid network topology of both centralized and distributed (mesh) networks at the edge, allowing the edge nodes to communicate without going through the hierarchy of the core network. Such heterogeneous connections and networking technology can thus enable low latency (at milli-second level), high reliability (up to 99.999%) and high throughput (from 1Gbps up to 10Gbps) at the edge, which can further enable edge applications to provide localized, real-time, safety related AI/ML services.

2.2.3 Data Collection and Analysis at Edge

One key benefit of the edge intelligence is the accessibility of real-time, local data that is captured by nearby devices or sensors connected to the edge or mobile edge computing node. 3GPP has defined data collection features like [31]. In the O-RAN architecture, data can be collected at the edge nodes including base station, mobile edge computing node, and user devices, to support near real-time network analysis for RAN optimization. Similarly, other sensor data like video, audio, temperature, and others could be collected by various edge nodes and can be used for various edge AI/ML applications to help provide near-real time analysis and decision making, to enforce safety, security or efficiency in verticals such as intelligent transportation systems, smart factory cases, and others.

2.2.4 RAN Intelligence Controllers

In a typical RAN, there are millions of decisions taken every second about which user to serve over the radio interface and how. Each of these decisions contributes to the service quality and the prioritization among users and services in case of conflicts. Traditionally, these micro-decisions are governed by a combination of supplier design choices and network configuration parameter settings done by the service provider. In the relatively simple 2G systems, the effect of a configuration change was mainly possible to understand. In today's more sophisticated multi-service 5G networks, it is virtually impossible, in a cost-efficient manner, to predict the effect a given set of configuration changes will have on the end-user services.

However, the intent of the RAN remains the same: to offer connectivity to the service providers' customers in a profitable way. The idea of intent-based management for RAN is to evolve the RAN configuration from setting technical parameters and instead allowing service providers to specify the connectivity service itself, prioritizing across users and services based on business intent and devices capabilities.

The non-real-time RIC is a concept developed by the O-RAN Alliance to realize intent-based management, built on principles of automation and AI and ML. The non-real-time RIC brings genuinely novel capabilities to the system and addresses use cases that were previously out of reach, with the ability to set policies per user and data enrichment information for RAN optimization. Intent-based management based on non-real-time RIC can be applied to Cloud RAN to enable a high degree of network programmability and can equally well be applied to purpose-built RAN to enable a wide variety of automation and optimization use-cases that are not possible today. This approach can be extended to the resources at the edge and combined with computing resource allocation to the end-users.

The non-real-time RIC is part of the Service Management and Orchestration system (SMO) and consists of a platform plus a set of microservices (named rAPPs by O-RAN Alliance) representing the network intelligence. The system's design is based on the following principles:

- *Access to information: There is a wealth of contextual information – not available in the RAN – with the potential to improve radio–resource management, RAN performance, and user experience. This includes application-level information, cross-domain information, UE positions, mobility trajectories, UE computation capabilities and external information.*
- *Dynamic optimization: Traditionally, management and orchestration have been performed on the timescale of hours. With automation and improved interfaces, the non-real-time RIC can optimize the RAN on a time scale down to half a second.*
- *User-level service assurance: Optimizing the RAN on a user level (in addition to the per-node level) enables the non-real-time RIC to address a broad set of use-cases that were previously out of reach.*
- *AI/ML over-engineered programs: The intelligence in RAN control is gradually moving to AI/ML-based software, and the non-real-time RIC is designed for AI/ML from day one.*
- *Innovation for openness: It is possible to build an open eco-system of intelligent controller software where applications (rAPPs) feed each other with data and insights.*

2.3 Envisioned Features and Key Technologies

Multiple international organizations have defined the expected requirements, features, and key technologies in the context of 5G edge. ESTI has published a specification [32] on an MEC framework and a reference architecture, as well as many other specifications as summarized in the MEC in their 5G Network white paper [33]. 3GPP has defined even more key enabling technologies of 5G to support MEC [34]. In addition, concepts such as Explainable AI, Named Data Network, innovative transport layer protocol, joint optimization of communication and computing, and distributed machine learning have been studied by various academic and industry organizations. These new features and key technologies will play an important role in future edge optimization, intelligence, and data analytics. The following sections will give an overview of some which may be growing in importance.

2.3.1 Explainable AI

AI has achieved growing momentum in its application in many fields to deal with increased complexity, scalability, and automation, which also permeates digital networks today. A rapid surge in the complexity and sophistication of AI-powered systems has evolved to such an extent that humans do not understand the complex mechanisms by which AI systems work or how they make certain decisions — something that is particularly a challenge when AI-based systems compute outputs that are unexpected or seemingly unpredictable. This especially holds for opaque decision-making systems, such as those using Deep Neural Networks (DNNs), which are considered complex black box models.

The inability for humans to see inside black boxes can result in AI adoption (and even its further development) being hindered, which is why growing levels of autonomy, complexity, and ambiguity in AI methods continues to increase the need for interpretability, transparency, understandability, and explainability of AI products/outputs (such as predictions, decisions, actions, and recommendations). These elements are crucial to ensuring that humans can understand and — consequently — trust AI-based systems. Explainable AI (XAI) refers to methods and techniques that produce accurate, explainable models of why and how an AI algorithm arrives at a specific decision so that AI solution results can be understood by humans.

Without explanations behind an AI model's internal functionalities and its decisions, there is a risk that the model would not be considered trustworthy or legitimate. XAI provides the needed understandability and transparency to enable greater trust toward AI- based solutions. Thus, XAI is acknowledged as a crucial feature for the practical deployment of AI models in systems and, more importantly, for satisfying the fundamental rights of AI users related to AI-based decision-making (according to European Commission ethical guidelines for trustworthy AI). Standardization bodies such as the ETSI and the Institute of Electrical and Electronics Engineers (IEEE) also emphasize the importance of XAI where AI models are deployed, indicating XAI's growing importance in the future. AI deployers and developers must comply with these ethical guidelines and regulations to ensure their AI solutions are explainable and trustable.

However, there are significant challenges in developing explainability methods. One of them is the trade-off between attaining the simplicity of algorithm transparency and impacting the high-performing nature of complex but opaque models (when one increases the transparency aspect, privacy and the security of sensitive data come into question).

Yet another challenge is to identify the correct information for the user, where different levels of knowledge will come into play. Beyond selecting the level of knowledge retained by the user, generating a concise (simple but meaningful) explanation also becomes a challenge. Researchers attempt to integrate knowledge-based systems so that the explanation becomes relevant to its application's context [35].

XAI helps deliver trust by supporting with the following properties:

- *Trustworthiness, to attain the trust of humans on the AI model by explaining the characteristics and rationale of the AI output*
- *Transferability, where the explanation of an AI model allows a better understanding of it so that it can be transferred to another problem or domain/ application properly*
- *Informativeness, relating to informing a user regarding how an AI model works to avoid misconception (this is also related to human agency and autonomy, which ensures humans understand AI outcomes and can take intervening actions on that basis)*
- *Confidence, which is achieved through having a model that is robust, stable, and explainable to support human confidence in deploying an AI model*
- *Privacy awareness, ensuring that the AI and XAI methods do not expose private data (which can be done through data anonymization)*
- *Actionability, with XAI providing indications regarding how a user could change an action to yield a different outcome in addition to providing the rationale for an outcome*
- *Tailored (user-focused) explanations, allowing humans — as AI system users of different knowledge backgrounds — to understand the behavior and predictions made by AI-based systems through tailored descriptions based on their roles, goals, and preferences*

It is vital to incorporate interpretability and explainability at different levels of complex AI techniques. The XAI framework is tightly linked with providing explanations for both different AI techniques (ML and Machine Reasoning (MR) techniques) and the environment through properly defined interfaces.

The main components of such a framework center on explanations, explainability for data, explainability for ML, and explainability for MR (see purple parts of Figure 9). The distinctive approach that we are taking is to apply explainability to ML and MR and the interplay between ML and MR by feeding the output of an ML model (both its predictions and explanations) into our MR techniques and applying it XAI to generate explanations. This proactive placement provides the right AI trustworthiness early on rather than relying on reactive fixes. Furthermore, this framework allows the integration of new XAI algorithms into the respective explainability components. In the future, newly developed XAI techniques for ML/MR can be easily deployed within the explainability for ML/MR components.

2.3.2 Multi-Access for the 5G Edge

Multi-access traffic management at the edge is vital for addressing ever increasing performance requirements for current and future applications. However, it is not possible to achieve,

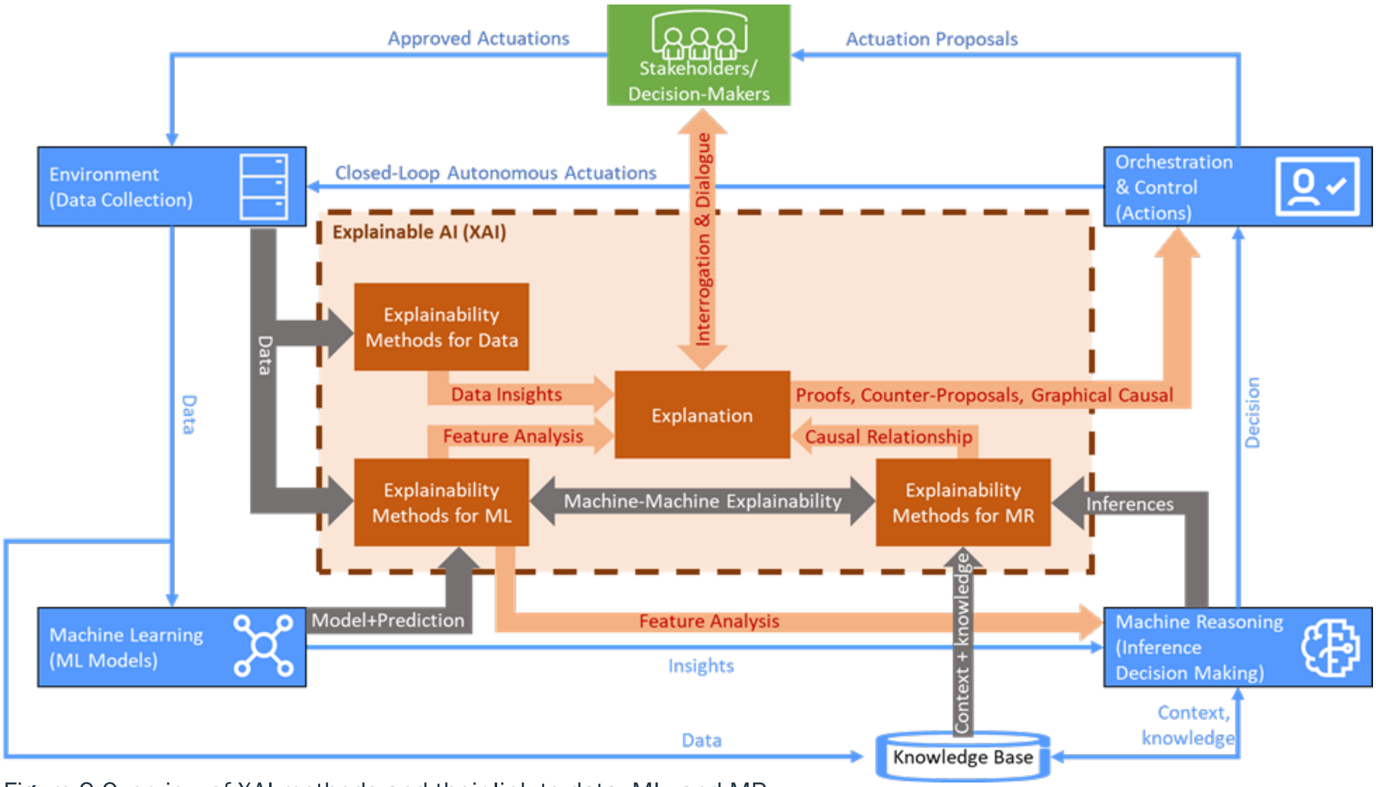

Approved Actuations
Stakeholders/ Decision-Makers
Actuation Proposals
Environment (Data Collection)
Closed-Loop Autonomous Actuations
Orchestration & Control (Actions)
Explainable AI (XAI)
Interrogation & Dialogue
Data
Explainability Methods for Data
Data Insights
Feature Analysis
Explanation
Proofs, Counter-Proposals, Graphical Causal
Causal Relationship
Data
Decision
Explainability Methods for ML
Machine-Machine Explainability
Explainability Methods for MR
Inferences
Model+Prediction
Feature Analysis
Context + knowledge
Machine Learning (ML Models)
Insights
Machine Reasoning (Inference Decision Making)
Context, knowledge
Data
Knowledge Base

Figure 9 Overview of XAI methods and their link to data, ML, and MR

for example, the maximum data rate and the highest reliability at the same time. Hence, the multi-access challenge is to manage data traffic across all available access networks and meet diverse application requirements in coverage, rate, latency, and reliability. To address this challenge, the following key questions need to be answered:

- *How to acquire individual application requirements?*
- *How to support multi-access traffic management?*
- *What measurements are needed for making smart decisions?*

Applications may have different QoS requirements, and Traffic Management (TM) service [36] recently introduced in the ETSI MEC reference architecture [37] allows applications to get informed of various capabilities and multi-access network connection information, and to provide requirements such as delay, throughput, and loss for influencing traffic management operations at the edge.

Multi-access traffic management requires a set of new protocols between client and network. Recently, multiple access management service [38] has been proposed. In parallel, 3GPP has developed the access traffic steering, switching, and splitting [39] feature. Both provide mechanisms for flexible selection of network paths, and leverage network intelligence and policies to dynamically adapt traffic distribution across selected paths under changing network/link conditions. Figure 10 shows the multi-access protocol stack which consists of the following two sublayers:

- *Convergence sublayer: This layer performs multi-access specific tasks such as access (path) selection, multi-link (path) aggregation, splitting/ re-ordering, lossless switching, keep-alive, and probing. Generic Routing Encapsulation (GRE) [40] may be used to encode additional control information, e.g., sequence number, at this sub-layer.*
- *Adaptation sublayer: This layer performs functions to handle tunneling, network layer security, and Network Address Translation (NAT). Existing protocols, including User Datagram Protocol (UDP) and Internet Protocol Security (IPSec), can be re-used.*

To take full advantage of multi-access connectivity, we should distribute traffic load intelligently across available access links in a manner that improves user experience with efficient radio resource usage. To achieve this goal, measurements that reflect the connectivity conditions of different access networks should be incorporated while determining multi-access traffic distribution. For example, the end-to-end packet delay measurements can be used to identify which access provides better latency performance.

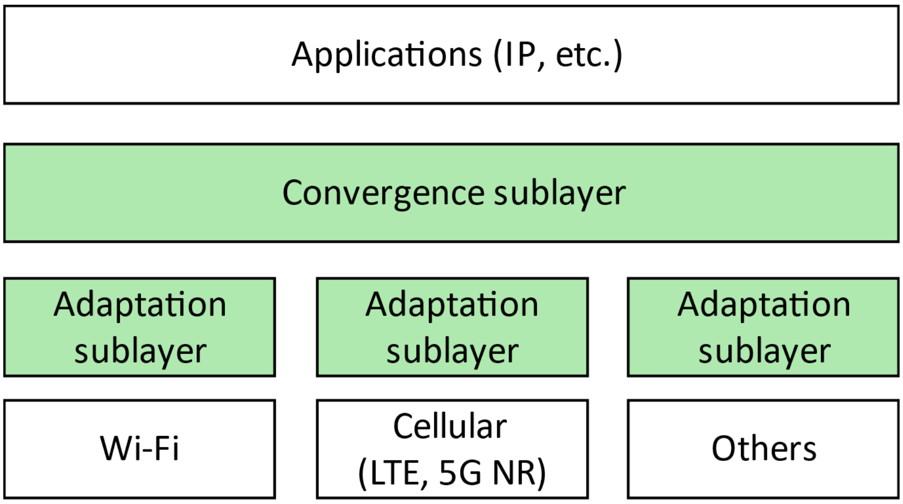


Figure 10 Multi-Access Protocol Stack.

When serving QoS flows requiring high reliability, packet drop ratio measurements give a good indication of whether redundant transmission over multiple access networks is required. In addition to end-to-end packet statistics, RAN measurements can indicate network quality degradation caused by deteriorating radio link quality or access network congestion, in a timely fashion. Moreover, ML/AI based algorithms may be developed to automatically configure and manage data traffic across all the available networks and improve end-user experiences.

Using these building blocks, we can unleash the full potential of multi-access/multi-connectivity at the edge to address the performance requirements of applications now and in the future.

2.3.3 Situational Network at the Edge

The emerging architecture being designed to accommodate 5G applications is known as Multi-Access Edge Computing (MEC) and several challenges must be solved before it can become a mainstream reality.

First, devices that might need edge computing must be recognized at the beginning of a session when user plane resources are being established. A local uplink classifier or branching point or packet session anchor must be allocated for these devices. Recognition of such devices may not be straightforward and could present an interesting machine learning challenge in and of itself. Second, protocols and techniques must be created for devices to discover any edge services available to them, including discovery of peer devices that may also be attached to the network as clients. Third, mobility management must be developed for the unique edge environment that preserves the IP address associated with a given session when necessary but that also recognizes that IP addresses may need to be reassigned under some circumstances. Finally, the interactions between edge computing and slicing must be dealt with. The concept of multiple Network Slice Instances (NSIs), each implementing the same slice but for different geographic locations, will be important here.

2.3.4 Situation-aware Transport Layer Protocol

Transport layer protocols such as Transmission Control Protocol (TCP) rely on the end host to handle congestion control, flow control and end-to-end reliability with the underlying assumption of ‘in-order byte’ delivery. However, dynamic, self-organized situational 5G networks at the edge, which may be running diverse applications (from end users to edge) in a collaborative, multi-hop fashion with involvement of multiple edge infrastructures, may not necessarily stick to the in-order byte delivery paradigm as the network topology itself may be dynamically changing [41], [42].

On the other hand, TCP relies on its reactive congestion control which kicks in only after the packet has traversed full path from the source node to the destination and back to the source/host edge infrastructure. Such a reactive approach leads to delays. Furthermore, the traditional end-host networking-based approach may be prone to single point of failure (or attacks by external jammers or misbehavior by existing end-users/subscribers of the network services) which may overload the networks, thus resulting in highly unreliable and latent networks at the edge. Thus, the added dynamics due to mobility, changing deployment and reconfiguring topology need to be taken into consideration while designing Transport Layer Protocols.

To this end, there is an inherent need for design of a situation-aware transport protocol. Additionally, some awareness of the network can be brought into consideration while designing such Transport Layer Protocols. For instance, one such protocol may involve in-band, telemetry-based network awareness building at the transport layer by, for example, exposing the congestion-related meta data. Additionally, this mechanism can be used in conjunction with the application awareness for building a stateful way of serving dynamic and situational-aware network.

2.3.5 Joint Optimization of Communication and Computing

With the proliferation of smart autonomous devices equipped with computing, communication, and storage, we are observing data explosion at the edge, generated from many new applications such as self-driving cars, AR/VR, panoramic telephony, holographic interaction, digital twins, etc. 5G edge computing is essentially bringing computing, intelligence, and communication together by providing a platform to process data close to source for these services, thereby reducing end-to-end latency and network traffic overhead. Both communication and edge computing resources, however, are limited, variable, and shared by several nodes in the vicinity running multiple applications with different performance requirements.

The existing computing and communication layers have mostly been designed either in an isolated manner or via loose coupling with coarse-grained information sharing between the layers. The expectation with 5G is that it will increase both users and traffic by an order of magnitude [43]. For this reason, it is of paramount importance to jointly optimize the computing and communication process to help avoid underutilization and oversubscription of resources, while also significantly improving an application’s performance.

While the concept of computing-communication co-optimization has been under consideration by academia and industry for a while, there are several challenges and directions that require closer attention to unleash the full potential of the joint optimization. First, a sound theoretical framework is indispensable to study the optimal joint resource allocation and resulting performance benefit for a given topology, Radio Access Technology (RAT), and computing resource (trusted peers, edge, cloud).

The framework should consider realistic modeling of communication and computing processes.

Second, it is important to have an in-depth characterization study of the emerging edge workload and design a generic computing resource representation or abstraction to assess the computing storage requirements. While AI/ML is likely to be the dominant workload to enable edge intelligence, the computing and communication resource requirements for distributed training (e.g., federated learning) and inference are different. The network also needs to collect real-time status or telemetry on the usage of current computing and communication resources in an efficient manner.

Third, the applications are increasingly becoming distributed and being implemented using cloud-native constructs. Additionally, applications can be decomposed dynamically (such as distributed inference [44]) where the real-time availability of computing and communication resources will help determine the right decomposition and placement/off-loading of decomposed units.

Fourth, the traditional design of scheduling computing and communication resources should be re-thought separately. Depending on the underlying hardware (e.g., Central Processing Unit (CPU), Graphical Processing Unit (GPU), Field Programmable Gate Arrays (FPGA), custom accelerators) used for computing and location (e.g., on-device or edge) of computing, the execution time may vary and put variable budget for the communication process to finish and meet end-to-end application requirement. A joint computing-communication resource scheduler can leverage the real-time information [45], use AI/ML-based techniques to predict the arrival of input data for computing, and finally schedule both resources to keep the computing instances (e.g., container) up and running for processing data upon arrival.

Last but not the least, academia and industry can work together to develop a realistic simulator/emulator leveraging open-source edge computing frameworks (e.g., O-RAN, OpenNESS), AI/ML frameworks (e.g., Intel Distribution of OpenVINO), AI/ML libraries, and real-world sensor datasets, which will be handy to validate the scalability of the aspects mentioned above and to generate key performance indicators.

The joint optimization problem poses additional challenges and complexity in the case of edge-enabled applications such as control systems and robotic applications, where time-sensitive functions are offloaded from robots to the edge for compute acceleration, energy efficiency and for leveraging advanced AI capabilities. Guaranteed latency and determinism are critical for reliable control loop operation, especially while off-loading robotic functions (such as perception, planning, cognition) on the edge at scale. As a result, robotic control also needs to be jointly optimized along with communications and computing for reliable and efficient robotic operation. The state of the wireless network (latency, packet errors etc.) can be used to adapt robotic control to changing network conditions and available compute resources. Similarly, the state of robots and their environment can be used to dynamically adapt, provision and schedule compute resources (CPU cores, memory, etc.) and communications resources (packet scheduling, reliability) to the changing needs of robotic tasks.

2.3.6 Distributed Learning at Edge

Distributed Learning [46] is one of the key enablers of edge intelligence that focuses on both training and inference with private and sensitive data at the edge, while avoiding the communication and latency cost of moving data for centralized processing in the datacenter. At the same time, it enables efficient use of compute capabilities available at the network edge or across a group of on-premises devices pooled together. Either of such available compute capabilities can thus host services by harnessing data close to the generation points and by leveraging disaggregated resources for compute and processing.

Distributed learning approaches have evolved to address several issues of centralized cloud-based learning and are especially relevant to the problems at the 5G edge. In particular, distributed learning uses are inherently collaborative at typical endpoint or edge nodes which have only a partial view of the data required for learning. Collaboration in distributed learning may take different forms, such as in Federated Learning or Fully Distributed (Decentralized) Learning, [46]. In Federated Learning, the collaboration is managed with the help of a central coordinator that combines the learnings from nodes processing over their own data. Whereas in Fully Distributed Learning, there is no central coordinator, and nodes must collaborate in a peer-to-peer manner.

An additional advantage of distributed learning that is relevant to the 5G edge is the ability to adapt the learning models to the local situational context of the edge. For example, a 5G edge infrastructure for a factory implementation would have different requirements compared to a 5G edge network deployed to service urban residential clients. In this case, a distributed learning framework deployed to optimize the 5G edge for functions such as traffic cell prediction, QoS management may need to be adapted differently for the two distinct deployments. Distributed ML allows for such model personalization to these distinct deployments, while still benefiting from learning common features through collaboration across all the nodes. With distributed learning local context or situational awareness can easily be included within the locally trained ML models to improve model accuracy for the local 5G edge context, when compared to a centrally trained model.

Distributed learning has many challenges, particularly when applied over the 5G wireless edge [46]. We list some of the challenges below:

- *Statistical properties of data distribution: Different nodes at the 5G edge may have different data collection and storage abilities, as well as may only have a limited view of the overall data distribution. The diverse non-Independent and Identically Distributed (non-IID) data across clients can lead to slow convergence of ML model training slow down AI model training.*
- *Heterogeneous Communication and Computational Costs: Devices on the wireless edge may have diverse computational and communications capabilities causing 'straggler effects' where poorly resourced clients on inefficient links relay their data infrequently or have high error rates. This leads to problems with model convergence, accuracy and fairness.*
- *Scalability: Collaboration to learn very large models across a large number of users can lead to poor scaling.*
- *Privacy and Security: While distributed learning solutions avoid data sharing across devices, sharing of model parameters during learning can still leak data privacy. Studies have shown that it is possible to reconstruct user's data through model inversion attacks.*
- *Security risks: Adversarial or malicious clients can corrupt model training by inserting false updates.*
- *Need for Self-Learning. Another challenge involves learning with limited data labeling, as access to annotated data with limited human support is one of the important challenges for the edge environment.*
- *Continuous learning: Constant new updates to the model may lead to a catastrophic forgetting of the model's earlier learning.*

There are several promising techniques that are being explored to mitigate the above issues and may be found in [46] [47] [48] [49]. We further note that ML computations are one of the key compute workloads that must be supported over the 5G Edge. Hence, several compute-communication co-optimization approaches discussed in earlier section are also relevant.

Currently there are various efforts to apply distributed learning solutions to applications in wireless networking [50] [49]. While there are recent efforts to introduce distributed learning techniques in both O-RAN and 3GPP, such efforts are nascent. Further work will be required to evolve 5G edge architecture framework to support distributed learning as a critical workload, in the design of the next generation of wireless standards.

2.4 Requirement Analysis

The ecosystem for edge computing is fragmented and is quickly evolving. Technical solutions, interfaces, standards, and business models are not set. Several players must be involved to create end-to-end solutions and CSPs must carefully consider in which industries they can expand their offerings beyond connectivity.

The edge application ecosystem is driven by third-party applications outside of the telecom domain since solutions for new use cases require specific domain knowledge from industry players outside the telecom space. Edge infrastructure will therefore be accessible to third party application providers and developers and will host a multitude of applications, each with specific characteristics and needs.

The edge application environment enables mobile network operators to host non-telco workloads and open up the network as a distributed cloud resource. Enterprises can develop applications, deploy, and manage them flexibly via orchestration logic towards a 'landing zone' that accesses the distributed cloud infrastructure and leverages services exposed through APIs for consumption. Below is a brief overview of the functional components needed to create an edge computing solution [51].

Connectivity: Once the development environment is installed, connectivity will be configured by the application developer. The application running on the network edge may have connectivity requirements on bandwidth, throughput, mobility, and/or latency within its components (for example, deployed on different hardware in a redundant setup) or with the external world, such as an internet connection, and the user equipment or the application session. Traffic routing for applications deployed further out in the network topology will need new mobile network solutions, such as distributed anchor, session breakout, and multiple sessions, and in some cases coordination between application server selection and usage of these mobile network solutions.

Application runtime execution environment: The very basic functionality that an edge computing service may provide is the runtime execution environment for VNF and non-telco workloads. An execution environment should be able to host applications and harmonize the requirements of the development communities. Many applications may use edge computing with different characteristics and functional requirements and require different platform components. Therefore, the operator provides a generic or multiple execution environments on the network edge that application developers can later customize.

Dynamic orchestration and management: Centralized orchestration is required to maintain consistency between possible traffic breakout points (where user plane gateway functionality is deployed) and the applications (which consume the traffic) in the network edge. The central orchestration and management functionalities need to be aware of the network topology and the available resources in the distributed cloud infrastructure. This orchestration layer will provide a harmonized single orchestration and management functionality over the different orchestration functions present. One of its purposes is to manage the platforms for non-telco workloads and VNFs according to service level agreements.

Service exposure: Exposure is a crucial function of defining and developing new capabilities (APIs) and securely exposing them to non-telco workloads. The exposure server exposes the core capabilities available internally within the operator or to a partner with a commercial agreement. The exposed core capabilities add value to internal or external users, for example, connectivity, optimization, identity, security, data, and analytics.

Optimization: 5G and edge computing techniques provide several opportunities for smart network optimization, which can be theoretical, heuristic, or AI/ML-based. AI/ML techniques can 1) detect changes in demand, deterioration or drifting of SLAs, and inefficiencies or problems in the network, 2) diagnose such issues and identify the root cause, and 3) predict the response of the network to workload redistribution, deployment of new resources (e.g., network slices), configuration changes, and changes in management policies. The system can then select and implement the best response to changing conditions. Operators can use AI/ML techniques to gain useful and timely insight into their networks and optimize the management, operation, and/or orchestration of the network.

To realize these functionalities at the edge, there is paramount need for building aggregate yet distributed knowledge of the applications, situations, or workloads, across all the devices and entities involved in such movement, storage and/or processing of the data at the edge. Inherently, such knowledge building would require intelligence and judicious collaboration between all the network entities/ devices involved thus requiring the need for collaborative intelligence at the edge. Towards this end, a few key innovations in the areas of distributed learning, edge data analysis for driving scalable intelligence across multiple applications/workloads with bounded latency and guaranteed high reliability are needed.

2.5 Architecture Direction

To realize the 5G edge optimization, edge intelligence, and data analytics, architectural innovations are required to support the envisioned features and key technologies at 5G edge. First, at edge network level, traditional 5G connectivity provides centralized, hierarchical architecture plus limited device-to-device connectivity support. Due to the high dynamic nature of the edge, it is needed to have highly flexible situational networks at the edge to connect the edge devices on the move to edge nodes, and connect the edge devices to each other, with or without centralized core networks support. Such situational network at the edge will allow sharing of data, sensor resource, and computing processing capability with low latency, high reliability, and high flexibility at the edge for various use cases. Secondly, to support the ever-increasing AI/ML computing needs and exponentially rise in AI/ ML workloads, distributed learning at the edge is needed to leverage the distributed data as well as AI/ ML processing capability of scattered edge network nodes and edge devices. This calls for a new distributed AI/ ML architecture to make optimal use of the available communication and computing resources while meeting the latency, security, and privacy requirements the 5G edge. The following sections discuss these architecture directions in detail.

2.5.1 Situational Network Architecture

Edge computing provides an ideal platform to enable many critical and time-sensitive applications that require huge sensor capabilities and computing resources to process sensor data in near real-time. Furthermore, intelligent data movement in a bandwidth-efficient manner and data utilization to make intelligent and timely decisions by running AI/ML algorithms at the edge must be accommodated at the edge. Situational awareness becomes critical here to take maximum advantage of potentially big data generated by sensor system utilizing compute resources available at edge in a time-critical way while keeping the bandwidth requirements manageable. An edge situation network enables intelligent collaboration among sensors, infrastructure nodes, and local compute nodes to process data closer to its source or point of service delivery.

Situational awareness can be acquired and maintained by mainly two types of systems: (i) a dynamic context-based discovery system, (ii) an Intelligence, Surveillance, and Reconnaissance (ISR) system. In dynamic context discovery, each node in situational edge network continuously acquires and maintains updated situation perception and network context in the proximity and network by frequently sharing information such as environment perception, Node's own status and information (device type/role, location, orientation, etc.), perceived communication environment, compute capability, and sensing capability and configuration. Context discovery enables the edge situation network to form an intelligent collaborative group for efficient and intelligent optimization of sensing, caching, communication, and compute requirements in the network. Context information maintained by dynamic context discovery is then utilized to create a collaborative ISR system by pushing compute at the edge to realize data to decision concept in a time-sensitive way. ISR optimizes the utilization of sensor assets, compute resources and network resources to collect and fuse actionable information to provide reliable high-quality situational information for assessing options, threats, and consequences of decisions. A situational aware model also needs to provide meaningful representations of actionable context and situational information so that network and users can readily consume the information in optimizing various operations and time-critical decision making.

2.5.2 Collaborative Edge Intelligence

Edge intelligence, where intelligent compute devices are needed for moving, storing, and processing data closer to its source or point of service delivery is paramount for prediction, preparation, and response in an accelerated manner to deliver near real-time services. Edge intelligence is paramount for real-time services and can accelerate AI/ML computations and workloads and offload centralized systems (e.g., cloud-based) that require higher bandwidths and lower latencies. Many applications leveraging AI/ML at the edge, however, are real-time and collaborative in nature where the form of collaboration can be fluid and application-and-context dependent (e.g., local relevance). For example, multi-camera video analytics at smart intersections requires real-time communication with the smart AI cameras (that have the most relevant field of view) for sharing information such as location, frame, output of local processing and analytics.

To enable such near real-time application-and-context or situation-aware collaboration among the intelligent edge devices and edge infrastructures, the underlying 5G networks needs to offer ultra-low latency and ultra-highly reliable communication. When these intelligent edge users are connected over a 5G network they can, together with the edge infrastructure, create a locally-available shared and distributed computing substrate that can be leveraged for time-sensitive collective computing or analytics tasks that may be comprised of a chain of AI/ML inferences. Such collaboration among intelligent nodes, if done intelligently and judiciously, can enable faster and more accurate decision-making processes for several high-stake applications such as traffic management, emergency response, drones, AR/VR, and autonomous systems comprising of connected autonomous robots, connected autonomous vehicles, among other systems.

To this end, the need for building collective knowledge and sharing such knowledge by forming networks of collaborative intelligent nodes is paramount. Such collective intelligence with connected network of intelligent agents can be termed as Collaborative Edge Intelligence (CEI). The emergence of CEI can thus address diverse application requirements, for instance, real-time event detection, action classification, and collaborative decision making with a comprehensive application-and-context or situation awareness of the edge environment in which such intelligent agents operate.

The CEI provides a paradigm for structured collaboration among the intelligent edge agents so that the joint edge intelligence can be realized to attain an overall objective of delivering real-time response service for, say, data and analytics delivery to edge users. A fundamental enabler for CEI is intelligent networked computing to enable near real-time collaboration among heterogeneous computation capable edge servers and edge users. Hence, networked computing framework, algorithms, application-aware communication-compute protocol built on top of ultra-low latency, guaranteed highly reliable (and always available) communication substrate, are fundamental enablers to realize CEI.

For realizing such CEI, the diverse workloads across AI, media, and network, all together converge onto a common infrastructure which must deliver optimization, efficiency, and lower cost of ownership. With modern state-of-the-art robust packages and tools, telecommunications, semiconductors, and other ecosystem partners, would require the development of such converged edge applications with AI and 5G networking capabilities [52].

2.6 System Recommendations for ML-driven Optimization

The joint complexity of different edge devices, network components, communication protocols, mechanisms over the different layers and applications create a reality of very complex interactions and mutual influence. Manual optimization or optimization based on classical algorithms and approaches requires ever-growing domain expertise and human resources.

An ML-driven optimization is thus an appealing tool for such modern systems. This is the case for 5G, where low latency and high bandwidth networks allow for many edge devices to perform numerous communications over a constantly growing variety of applications. ML-driven optimization has the potential to adapt to evolving situations, conditions and heterogeneous environments as well as seeing through complex interaction between the various components of a system and optimize resource usage in a way that is rarely accessible, even with domain expertise.

2.6.1 ML for Systems

There are many ways in which ML can be used to optimize 5G. Data delivery through the network may be required to be lossless (e.g., state-dependent encryption) or can sustain losses (e.g., audio or video streaming). For lossless communications, it is desired to have intelligent control over the way data is processed and paced through the network and, finally, delivered to its destination. For lossy communications, it may be smartly controlling the loss (e.g., lower the resolution of a video stream upon congestion).

2.6.1.1 ML for Congestion Control

Congestion control refers to a mechanism that determines the pace at which a sender injects new data into the network. The traditional congestion control mechanism, employed over TCP, relies on an intuitive mechanism and possibly with theoretical guarantees, such as "Additive Increase/Multiplicative Decrease" (AIMD), along with "slow start" and "congestion avoidance." More recent mechanisms introduce rate-based methods (e.g., CUBIC), incorporate feedback from network switches (e.g., DCTCP), and even work cross layers (e.g., QUICK).

The quality of the congestion control is well-known to have a critical impact on both network throughput (e.g., due to packet losses and retransmissions) and latency. Thus, replacing such classic solutions with ML-driven algorithms holds the potential to lower latency and increase throughput by having the mechanisms themselves adapt to evolving network conditions and considering the traffic itself.

2.6.1.2 ML for Data Streaming

For video, audio, and gaming applications, it is often the case that latency is of greater importance than throughput. Smart ML-driven data streaming aims to adjust the signal's quality (e.g., video resolution) to achieve better client experience and resource usage. The advantage of ML-based methods over classical solutions is in resolving when and where to introduce the loss dynamically. For example, in audio and video conferences, ML can help in introducing loss in less important data items (e.g., silent moments or video fragments with homogeneous background). The network can use an ML API exposed by edge devices and network stations that can guide it to a better use of data loss.

2.6.1.3 ML for Caching

Caching is a well-known mechanism to improve data locality, resource usage, client experience and dramatically reducing latency and applications response time. Traditional cache management mechanisms rely on intuitive methods and often with theoretical guarantees (e.g., begin online competitive). Such methods include the Least-Recently-Used (LRU) policies, their approximations, and related more recent variants (e.g., Time-aware LRU (TLRU), Least-Frequently-Used (LFU)). None of them is optimal for specific application and usage patterns. Moreover, all these policies are reactive because they rely only on previous data usage patterns.

ML-based cache management may enable two significant advances that can lead to improved resource usage. First, the policy can be made adaptive, evolving, and learning as data keeps arriving. Second, it may learn not only the data patterns but also take into consideration other events. For example, it may be absolutely time-dependent and react to real-time events, such as parsing news sites and social media to predict what data will be consumed shortly and prepare for it.

2.6.1.4 ML for Scheduling and Load balancing

Scheduling and load balancing are at the heart of the management of any distributed system. The main goal of these algorithms is to distribute and time the work in a way that optimizes some target metrics such as job completion times, tail latency, or maximum oversubscription of a processing or a network element.

It is often the case that these problems are computationally intractable, so modern solutions heavily rely on heuristics. Moreover, the metrics to optimize are usually not sufficiently simple and the systems include different participants with different optimization goals (e.g., throughput vs. delay-sensitive applications). Thus, even designing simple heuristics becomes highly challenging and requires domain expertise across many layers. ML-based scheduling and load balancing solutions have the potential to capture complex structures and dependencies among metrics and participants and offer better performance and with limited domain expertise in each specific application.

2.6.2 Systems for ML

Different layers of the infrastructure (i.e., network and compute) offer different abstraction levels. Usually, it is the case that support of APIs exposed by the infrastructure can offer better resource efficiency. For example, while it is possible to implement lossless packet delivery in the application layer (e.g., over UDP) the efficiency and the usage of network resources is better managed where such delivery is supported by the network itself. Therefore, having the 5G infrastructure to support and expose APIs to ML application is a key step towards resource efficiency.

2.6.2.1 Systems Support for Caching ML Application Data

Data prefetch is a well-known technique to accelerate response times and increase resource efficiency both in hardware and software. For some applications the access patterns are evident and data pre-fetch works well, and for some others it may be more challenging. It is therefore of interest to expose API to the ML applications that will guide the prefetch mechanisms towards which data to prepare. Note that we have earlier discussed ML-based caching, however, we may be also interested in classic data structure with significantly faster processing speed that only exposes parameters to ML applications.

2.6.2.2 Systems Support for Federated Learning

In a federated learning procedure, many edge devices participate in a construction of an ML model. Usually, the process involves a coordinator and a parameter server (which may be a single entity, centralized or distributed) that coordinates a training procedure. At each training round, the coordinator picks a subset of available devices that in turn derive parameter updates based on their local data and send their updates to the parameter server. The parameter server processes all updates and computes updated model parameters.

A main idea behind federated learning is to protect the privacy of participants' data. Federated learning imposes several challenges in privacy as well as bandwidth and compute, especially for edge devices. 5G has the potential to take federated learning a step forward by providing support for better privacy (e.g., differential privacy, secure aggregation), support for in-network processing of updates (e.g., filtering, averaging, shared-randomness) and incremental computation alleviating the burden on the recourse-limited edge devices.

2.6.2.3 Systems Support f or ML SLAs

Different ML models trained towards the same main optimization goal may differ in properties such as expected accuracy and inference time. To adhere to SLAs and provide better user experience, the infrastructure can direct queries to those models that would best benefit the users' needs and SLAs. For example, it may decide to direct a query towards a faster model due to network congestion to reduce the response time.

2.6.2.4 Systems Support for Collaborative Intelligence

One of the exiting directions for nowadays and future ML applications is Collaborative Intelligence where the edge devices communicate and interact directly for ML purposes. One example for such interaction would be smart cars that communicate and exchange information about traffic conditions and fast evolving situations to prevent life threatening events.

System support for collaborative intelligence may offer short communication paths, privacy, security, and authenticity of data and even help in processing as discussed earlier.

3. Application of 5G Edge Automation and Edge Intelligence

With the enablement of automation, optimization, and intelligent decision-making for network and compute resource allocation, network function selection, as well as workload optimization at the 5G edge, various use cases can be realized with guaranteed QoS. This chapter provides a brief list of use cases together with their main challenges and describes how they can benefit from the application of automation and intelligence at the 5G edge.

3.1 Autonomous Industrial Solutions

Industry 4.0 and new autonomous industrial use cases for intelligent factories bring strict requirements for computing and networking domains, challenging broadly utilized cellular technologies. Connectivity requirements will play a vital role in such autonomous systems. Future networks will need to support more than 37 billion connected [53] industrial IoT devices, from sensors through operational analytics to high-definition video analytics. From the autonomous industry connectivity perspective, six requirements will play a critical role to enable envisioned use cases:

- *Latency – where worst-case latency requirements will drive manufacturing automation, electrical grind and intelligent homes use cases*
- *High-speed bandwidth – to stream hundreds of video streams and run distributed real-time analytics is the critical for the manufacturing to process data*
- *Multi-tenancy –to provide secure access to the factory for multiple vendors at the same time*
- *Autonomy / Self-configurability – for instance, where a network can self-monitor and configure autonomously without human intervention, such as when a freshly discovered robot can join a network on-demand and get the required bandwidth and latency performance. Another case could involve the autonomous re-configuring of a failed node or broken segment of a network link.*
- *Security – where each part of the system, from the control plane to the user data, must be secure and preserve privacy at each stage of data processing*

With the delivery of technologies to address the above requirements, autonomous industrial systems will bring a new generation of tools and technologies to increase productivity, efficiency, and safety. Tools like computer vision, machine learning, combined with sophisticated sensing capabilities, will improve the productivity, the inventory tracking, and the safety monitoring.

Challenges and potential approaches

While some data transfer challenges for autonomous systems can be addressed by deploying new 5G equipment, the latency and bandwidth for industrial applications will need to be managed through a new form of autonomous orchestration fitted for industrial sectors. In addition to software frameworks, new deployments will need to move away from the standard manufacturing pyramid, to provide a unified Internet Technology (IT) and OT solutions through a Cyber-Physical Systems overlay. Other challenges and approaches include:

Information Centric Networking - A machine-to-machine protocol like Open Platform Communications Unified Architecture (OPC UA) [54] will play a key role in data distribution in industrial systems. However, the overhead of TCP protocol when mapping names to IP addresses and its weak support for multicast communication are pushing academia and industrial partners to look into other approaches. One of the proposed solutions is Information-Centric Networking (ICN) [55], where a unified control and data plane can simplify the onboarding, addressing, and communication patterns of industrial devices. ICN can also address increased dynamicity challenges where drones, collaboration robots ("cobots"), and mobile robots will need to move across the factory floor.

Time Sensitive Networking - IEEE 802.1 Time-Sensitive Networking (TSN) standards and integration with 5G systems defined in 3GPP Rel. 16 will enable TSN time synchronization and scheduled traffic to be delivered between TSN devices connected across a 5G system. Therefore, to provide effective mechanisms to schedule traffic between TSN capable networks and 5G will be an important capability for services such as cobots that will autonomously work side-by-side with human workers or other cobots. Intelligent Edge co-optimized with wireless TSN can also enable compute and battery constrained collaborative mobile robots, to effectively leverage the Edge for advanced AI (perception, learning, cognition and adaptation), while meeting tight end-to-end latency and reliability requirements.

Co-system Determinism - Finally, co-systems determinism for network and compute where subsystems will need to meet the expected delivery deadline every time will create new challenges as data will need to be transported often over long distances, potentially traversing multiple operators. New determinism requirements will need to be also addressed to deliver optimal solutions for factory multitenant setup.

The confluence of the mentioned industrial edge, OPC UA, TSN, industrial cloud ('on prem'), and industry 4.0 use cases lead to significant architectural changes in Industrial Control Systems (ICS) that will directly influence 5G edge automation. On the other hand, the next iteration of industrial systems that will utilize 5G edge intelligence will be able to move away from a typical industrial multi-level factory overlay [56] into the autonomous and distributed factory automation system.

3.2 Intelligent Transport Systems

Transportation systems are struggling to keep pace with the demands of our globally connected economy. The increasing trend toward urbanization is creating unprecedented challenges for city leaders around transportation infrastructure. In addition, today's cities account for 71 to 76 percent of CO2 emissions [57], 67 to 76 percent of global energy use, road safety challenges with 1.3 million deaths annually [58], and an economic impact of $305 billion due to congestion. As a result, rapidly growing cities are under pressure to address pedestrian safety, congestion, environmental issues, and resulting economic impact.

Smart Roadways and Intelligent Transportation Systems (ITS) are one of the most practical and near-future applications of edge automation and edge intelligence. These technologies help enable smart cities to overcome some of the most pressing operational challenges, such as increasing urbanization, energy efficiency, and congestion, that impact the daily lives of each citizen. edge automation and edge intelligence serve as the lifeblood for realizing an intelligent roadways vision through the deployment of multimodal sensors and real-time processing. Together, these combine to create efficient traffic management services, data collection, and real-time analytics for road users and pedestrian safety.

Edge computing and automation software frameworks, together with smart edge infrastructure may play a key role in orchestrating and managing the ITS applications in a world of distributed edge computing [59]. Thus, it is pivotal to create common platforms and architectures to help cities merge their IoT and networking workloads to achieve greater synergy and optimize their hardware solutions for a world in which expanding cities creates strain on transportation networks. To enhance the safety, reliability, efficiency, customer experience, and quality of a city's transportation infrastructure, ITS technologies bring forth unique requirements in the form of confluence of computation and communication at the edge, enabling critical services for the infrastructure, vehicles, and other users of the transportation system. In this way, roadside edge computing infrastructure forms the basis for realizing the ITS vision of the future—Intelligent, connected roadway infrastructure that is resilient and can adapt to the needs of a growing and changing city.

To this end, 5G connectivity (Vehicle-to-Infrastructure (V2I), Infrastructure-to-Pedestrian (I2P), Private wireless, etc.) combined with compute demanding capabilities such as multimodal sensing, accurate positioning/localization, are to be deployed at the roadside edge infrastructure to meet the bandwidth and latency requirements for transportation. Such capabilities would help to improve the efficiency, safety, and experience of all the road users, as well as generating an overall net positive impact on environmental greenhouse gas emissions. The V2I and I2P communications enable vehicles and other road users to communicate with static or movable road infrastructure nodes, sharing data that can improve operational coordination and efficiency. In ITS solutions, RSUs, vehicles, and other road users generate large amounts of time-sensitive data to be used for a variety of applications and use cases. Furthermore, as the network spectrum resources available for the ITS is limited, the data between the nodes need to be shared in a timely and reliable manner via bandwidth-efficient communication.

Some of the important use case examples for ITS include, but are not limited to:

- *Sharing of perception, maneuver, and AI-workload models efficiently among entities using the road*
- *Enhancing Vulnerable Road User (VRU) safety*
- *Offering value-added services*
- *Roadside virtual environments via digital twins with on-demand service orchestration at the edge infrastructure*

Addressing the unique challenges posed by such use cases is key to offering reliability, safety, and efficiency in transportation. Thus, there is immense ongoing industrial efforts for enabling ITS with active collaboration with policy makers, automakers, manufacturers, cellular infrastructure operators (supporting C-V2X, 5G NR V2X, Dedicated Short-Range Communications (DSRC), NextGen Wi-Fi, and beyond 5G networks) around the world. For seamless and unified integration of all such technologies in common socio-economic manner, it is crucial to actively advance the standards and technical bodies with the singular goal of using the benefits of technology across edge automation and intelligence to improve the lives of citizens across smart cities and transportation systems.

3.3 Smart Energy and Smart Homes

Smart energy involves electricity, water, and gas delivered to customers through smart meters, which provide critical data to maximize the value of home automation systems and related IoT devices. A Smart Home is the integration of the utility smart meters and in-home devices enabled by an internal wireless radio link embedded in the smart meter. There are consumer level new services and products such as smart appliances, communicating thermostats, Heating, Ventilation, and Air Conditioning (HVAC) vent zone controllers, remote smart phone monitoring applications, sprinklers, and electric vehicle charging stations, etc.

A Smart Home helps conserve energy (electricity, gas and water), limit peak demand and increase overall delivery as well as endpoint efficiency. Home automation uses computers or smart phones to control basic home functions and features automatically, as well as allowing vital home functions to be controlled remotely from anywhere in the world through the Internet. A Smart Home should provide comfort, security, and the most cost-effective use of electricity, gas and water. Some home automation can include scheduling and automatic operation of water sprinkling, setting rules, swimming pool conditioning, heating and air conditioning, window coverings, security systems, lighting, food preparation, clothes washing and drying appliances, electrical vehicle charging – and many more tasks.

To achieve these, the automation system must have access to several sets of data, which may include: total real time energy consumption, total accumulated energy usage, individual smart appliance and major load real time energy consumption, pricing information, customer preferences in terms of comfort parameters and cost containment, usage patterns, and many more. Sources of home automation data can be from several sources, including: the utility company and its billing systems, smart meters for real time and accumulated energy consumption and energy quality, as well as smart appliances such as thermostats, electric vehicle charging stations and smart sensors.

Additionally, there are also challenges for home automation. As more devices are included in a Smart Home, more data is transmitted over the network, which may result in latency concerns if the data is transmitted to the cloud or the remote server is far away. Another concern may involve security and privacy risk when home data is transmitted to a public cloud.

Edge computing should help home automation in multiple ways. The edge has powerful computing resources, which off-loads the computing task from homes. Edge networks can store more data than a home does, and the rich data is useful for AI/ML in the Edge, which can in turn help home automation. Moreover, as the edge is closer to the home user than a public cloud, it reduces communication latency. Edge may also help relieve the security concern with closed loops between edge compute servers and UE. In addition, with edge automation, the Smart Home can manage resources better, for example, to monitor battery power, and utilize the data in AI/ML training for predictive maintenance.

To better improve the home automation efficiency, AI at the edge can be implemented partially in edge devices and partially in the hubs and gateways through which they connect. With the joint AI capability, decision is more local to the user, without sending inquiry/query to and waiting for decision from an edge node or a device which is far away. The following are some examples where the performance can be influenced or improved by joint edge automation with device/hub/gateway.

- *Self-healing*
- *Monitor signal strength and prevent outage*
- *Detect jamming attack on the meters*
- *Automatic door opening*

Beyond edge automation, edge can also help optimize the operations of components of Smart Energy and Smart Homes. As Smart Energy is distributed to homes from central locations such as factories or other energy generation facilities, it can be monitored for usage at the residence, thus allowing the whole energy distribution system to be optimized via load forecasting, distribution automation, and other energy grid optimization techniques.

3.4 Connected Health

Accelerated by the global COVID-19 pandemic, telehealth solutions range from simple video conferencing sessions with medical professionals to sophisticated in-home monitoring, allowing healthcare workers to track progress and adjust treatments for remote patients. Secure, low-latency connectivity, wearable devices, smart in-home sensors, and computer vision make virtual health management possible while reducing costs and increasing accessibility to healthcare. Additionally, experiments with remote surgery show promising results while edge-based computer vision enable flagging of health conditions in real-time. Edge automation enables these types of solutions to make better and faster decisions while keeping computer vision models local where possible therefore maintaining compliance with privacy regulations.

3.5 Enabling Location Information

Location information is essential for key services and applications with use cases such as precise e911, fraud prevention and mitigation, hyper-local customer applications such as weather, instant couponing or experience zones. It is also used for real-time network optimization.

While Global Positioning System (GPS) or device-provided location is commonly used, uplink-based location tracking has the advantage of better UE battery management and user control of the location info.

For the uplink-based method, Next Generation (NG)-RAN, an O-RAN node includes Next Generation NodeB (gNB) and real- and non-real-time RICs that enhance traditional network functions with embedded intelligence. The gNB controls multiple Transmission Reception Points (TRPs) an antenna array with two or more antenna elements located at a specific geographical location for an area. Using Sounding Reference Signal (SRS) measurements at different TRPs, the uplink location can be calculated in real-time at a high level of accuracy – as low as sub-20 meters.

In addition to the mainstream use cases, this method enables velocity measurement that could provide crash detection capabilities as well as angle of arrival as an alternate method to barometric pressure for z-axis positioning which are critical for applications in automotive or industrial automation environments.

3.6 Cloud and Edge Gaming

The demand for high quality, high throughput, and low latency is continuously increasing as a requirement for gaming platforms. For example, modern games expect gaming infrastructure to support vast amount of data processing to render frames at the highest quality at the highest frame rate possible, support for hardware-accelerated ray tracing, and AI capabilities plus a vast amount of storage space (e.g., 150GB of storage space per game install).

The introduction of cloud gaming platforms enabled gamers to play on any device connected to the Internet without the need to upgrade their client gaming platforms. Cloud gaming offers two models of operation, frame streaming and command streaming model. The frame streaming model eliminates the need for high-performance GPU and storage in the client platform by rendering the game in the cloud instance and stream the encoded frames to the client device. The command streaming model eliminates the storage needs at the client by launching the game in the cloud instance and stream the GPU APIs and data to render a frame using the client GPU. Though the frame streaming model offers better throughput, the end-to-end user input to display latency in cloud gaming is much higher than the client gaming platforms and potentially degrades the overall gaming experience. Additionally, the transcoding and streaming content from the cloud to the client under varying network conditions brings visual quality variations.

As low latency requirements, link variability, and available bandwidth are still challenging, new approaches related to the 5G edge and MEC space are being considered to dynamically distribute rendering between cloud, edge, and client based on latency budget, bandwidth, or acceleration requirements of the gaming workload. The 5G plane enables a new way of accessing streamed game content inside and outside the home giving new experiences to the users. Edge automation can therefore lower the latency, as well as manage additional latency requirements added by additional network or compute cycles. It provides a path to bring a true end-to-end gaming experience that can leverage edge to perform a hybrid of frame streaming and command streaming models to deliver highly responsive gaming experience without sacrificing quality and frame rate.

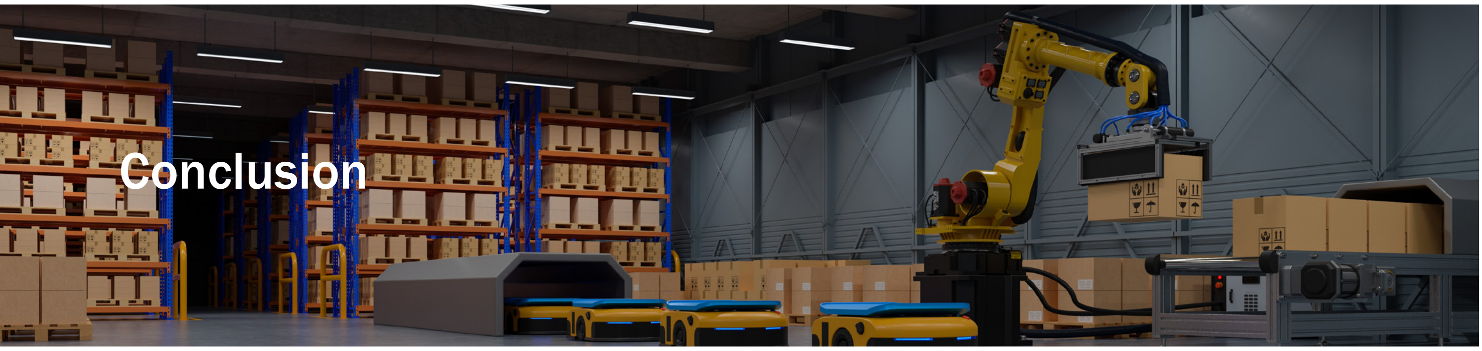
3.7 Scalable Digital Twin

A Digital Twin (DT) is a real-time virtual representation of a physical entity such as an object, a system, or a process. Using connected sensors, this cyber-physical technology permits connectivity and synchronization between the physical components and their digital counterparts. Further, through analytics and simulations using the digital model, the Digital Twin technology can produce future predictions with rich insights about the physical entity.

The unique characteristics of DT technology has several potential applications in the fields of infrastructure, smart cities, manufacturing, natural resources, healthcare, etc. For example, in ITS, DTs can accurately simulate transport network of a city and can optimize traffic efficiency, planning and development of transport infrastructure. In manufacturing, the DT technology has demonstrated disruptive impact on handling complex processes like product lifecycle management, asset maintenance, production line efficiency optimization, etc.

So far, the DT technology has seen limited adoption in the industry due to its stringent requirements on communication and computing infrastructure. In large scale applications such as in smart city, ITS, or manufacturing, the DT technology needs to collect data from a large number of connected sensors, and it is not yet viable to transport all these data to the data center for processing. On the other hand, DT also requires powerful computation resources to analyze the sensor data, simulate complex digital models, and generate future predictions in real-time.

The combination of 5G and edge technologies can provide feasible and cost-effective solutions to realize DT applications. Edge computing allows processing of sensor data close to the source, thus avoiding the need to transport large amounts of data over long distances. Alongside, the low-latency communications offered by 5G can help to achieve real-time service requirements of the DT applications. On the computational front, microservices based architecture can provide a scalable and flexible solution for the DT system to keep up with the analysis and simulation tasks in real time.



Conclusion

5G and edge computing are two intertwined technologies that will converge and work together to significantly improve the performance of applications and enable massive amounts of data to be processed in near real-time at different locations (edge zones). Edge automation and optimization with AI/ML can help to automate and optimize network system processes and service delivery at each available edge zone or throughout multiple edge zones. The ultimate goals of this symbiosis between 5G and edge involve increased performance guarantees, enhanced workload balancing, improved processing capabilities and times via 5G edge automation and optimization, reduced human intervention up to zero-touch management and orchestration. Edge intelligence, based on the low latency, high reliability 5G connection at edge and the AI/ML processing power provided by edge computing enables pervasive intelligence on all connected edge devices, as well as distributed data analysis and distributed learning on connected edge devices.

Equipped with AI/ML-driven capabilities, the 5G edge can be further augmented. Differentiation is crucial between AI/ML-based solutions for the network to control, manage, and orchestrate resources and functions and how systems should be designed to improve the performance and resource utilization of AI/ML-based solutions. Integrating AI/ML advances to 5G edge automation will enrich human experience, enable autonomic decision-making with adaptive policies, and reduce or eliminate human errors. The implementation of AI/ML-driven optimization at the 5G edge enables the adaptation to evolving situations, conditions, and diverse environments as well as seeing through the complex interactions between various components of a system and optimizing the resource usage.

5G edge automation and optimization can enhance the 5G edge with various new features and enable multiple key technologies. In this white paper, some key features and technologies related to data collection and processing, context discovery and situational awareness, how to handle (network) dynamics, explainable AI, multi-access, distributed learning, and achieving the joint optimization of communication and computing have been discussed. An analysis of requirements followed discussion around potential directions for network architecture to demonstrate the gaps and needs for enabling the discussed key features and technologies. Finally, a list of selected use cases demonstrates key benefits and challenges facing industries today regarding 5G edge automation and optimization.

Acronyms

3GPP: 3rd Generation Partnership Project

5G: 5th Generation

AI: Artificial Intelligence

IAIOps: AI Operations

AIMD: Additive Increase/ Multiplicative Decrease

API: Application Program Interface

AR: Augmented Reality

BBU: BaseBand Unit

C-SON: Centralized SON

CEI: Collaborative Edge Intelligence

CNF: Cloud-native Network Function

CPU: Central Processing Unit

COTS: Commercial Off-The-Shelf

CSP: Cloud Service Provider

CSMF: Communication Service Management Function

CU: Centralized Unit

DNN: Deep Neural Network

D-SON: Distributed SON

DSRC: Dedicated Short-Range Communications

DT: Digital Twin

DU: Distributed Unit

ECaaS: Edge Compute-as-a-Service

EN-DC: Evolved-Universal Terrestrial Radio Access New Radio Dual Connectivity

ETSI: European Telecommunications Standards Institute

FPGA: Field Programmable Gate Arrays

GMA: General Multi-Access

gNB: Next Generation NodeB

GPU: Graphical Processing Unit

GPS: Global Positioning System

GSMA: Global System for Mobile Communications Alliance

HCP: Hyperscale Cloud Provider

HVAC: Heating, Ventilation, and Air Conditioning

I2P: Infrastructure-to-Pedestrian

IaaS: Infrastructure-as-a-Service

IAB: Integrated Access and Backhaul

ICN: Information Centric Networking

ICS: Industrial Control Systems

IEEE: Institute of Electrical & Electronic Engineers

IID: Independent and Identically Distributed

IoT: Internet of Things

IP: Internet Protocol

IPSec: Internet Protocol Security

ISR: Intelligence, Surveillance, and Reconnaissance

IT: Internet Technology

ITS: Intelligent Transportation Systems

LCM: Life Cycle Management

LF: Linux Foundation

LFU: Least-Frequently-Used

LLS: Lower-Layer Split

LRU: Least-Recently-Used

MANO: Management and Orchestration

MDA: Management Data Analytics

MEC: Multi-Access Edge Computing

MEF: Mobile Ecosystem Forum

ML: Machine Learning

MNO: Mobile Network Operator

MR: Machine Reasoning

NAT: Network Address Translation

NDN: Named Data Networking

NFV: Network Function Virtualization

NG: Next Generation

NR: New Radio

NSI: Network Slice Instance

NSMF: Network Slice Management Function

NSSMF: Network Slice Subnet Management Function

NWDAF: Network Data Analytics Function

Acronyms

OAM: Operations Administrations and Management

O-Cloud: Orchestrator and Cloud Platform

ONAP: Open Networking Automation Platform

ONF: Open Networking Foundation

O-RAN: Open Radio Access Network

OT: Operational Technology

PaaS: Platform-as-a-Service

PCF: Policy Control Function

PoP: Point of Presence

QoS: Quality of Service

RAN: Radio Access Network

RAT: Radio Access Technology

RIC: RAN Intelligent Controller

RSU: RoadSide Units

RU: Radio Unit

SDN: Software Defined Network

SFG: Security Focus Group

SI: System Integrator

SLA: Service Level Agreement

SON: Self-Optimizing/Organizing Network

SRS: Sounding Reference Signal

TCP: Transmission Control Protocol

TIP: Telecom Infrastructure Project

TLRU:
Time-aware Least-Recently-Used

TM: Traffic Management

TMF: Tele-Management Forum

TRP: Transmission Reception Points

TSG: Technical Specification Group

TSN: Time Sensitive Networks

UDP: User Datagram Protocol

UE: User Equipment

UP: User Plane

UPF: User Plane Function

UWB: Ultra-Wide Band

V2I: Vehicle-to -Infrastructure

VNF: Virtual Network Function

VPN: Virtual Private Network

VR: Virtual Reality

VRU: Vulnerable Road User

XAI: Explainable AI

ZSM: Zero-Touch Network and Service Management

References

[1] 3GPP, System Architecture for the 5G System (5GS), V.17.1.1, 3GPP TS 23.501, June 2021.

[2] 3GPP, Study on management aspects of edge computing, V16.0.1, 3GPP TR 28.803, Sept. 2019.

[3] 3GPP, 5G System Enhancements for Edge Computing, V1.0.0,, 3GPP TS 23.54, June 2021.

[4] 3GPP, Study on enhancements of edge computing management, V1.0.0, 3GPP TS 28.814, June 2021.

[5] 3GPP, Architecture for enabling Edge Applications, V.17.0.0, 3GPP TS 23.558, June 2021.

[6] 3GPP, Architecture enhancements for 5G System (5GS) to support network data analytics services, V17.1.0, 3GPP TS 23.288, June 2021.

[7] 3GPP, Study on enablers for network automation for the 5G System (5GS)", V.17.0.0, 3GPP TR 23.700-91, Dec. 2020.

[8] 3GPP, Study on enhancement of Management Data Analytics (MDA), V17.0.0, 3GPP TR 28.809, March 2021.

[9] 3GPP, Management and orchestration; Management Data Analytics (MDA), V0.0.0, 3GPP TS 28.104, April 2021.

[10] 3GPP, Study on enhancement for Data Collection for NR and EN-DC, V.0.1.0, 3GPP TR 37.817, Jan. 2021.

[11] 3GPP, Management and orchestration; Edge Computing Management,, 3GPP TS 28.538, June 2021.

[12] 3GPP, Study on Security Aspects of Enhancement of Support for Edge Computing in 5GC, V0.6.0, 3GPP TR 33.839, May 2021.

[13] 3GPP, 5G System; Network Data Analytics Services; Stage 3, V.17.3.0, 3GPP TS 29.520, Jun. 2021.

[14] O-RAN, "O-RAN Architecture Description", v05.00, O-RAN, WG1, 2021.

[15] T. Forum. [Online]. Available: https://www.tmforum.org/catalysts/the-edge-in-automation/.

[16] T. Forum. [Online]. Available: https://www.tmforum.org/collaboration/catalyst-program/artificial-intelligence-operations-aiops/.

[17] ETSI, "Zero-touch network and Service Management (ZSM); Reference Architecture, ETSI GS ZSM 002 V1.1.1," 2019.

[18] ETSI, "Zero-touch network and Service Management (ZSM); Closed-Loop Automation; Part 1: Enablers, ESTI ZSM009-1 V1.1.1," 2021.

[19] Apache, "Apache Kafka," [Online]. Available: https://kafka.apache.org/.

[20] Apache, "Apache Pulsar," [Online]. Available: https://pulsar.apache.org/.

[21] RabbitMQ, "Rabbit-MQ," [Online]. Available: https://www.rabbitmq.com/.

[22] Apache, "Apache Spark," [Online]. Available: https://spark.apache.org/.

[23] 3GPP, ""Study on management and orchestration of network slicing for next generation network", V15.1.0, 3GPP TR 28.801," 2018.

[24] Omdia, "Artificial Intelligence for Edge Devices," Available: https://omdia.tech.informa.com/OM011942/Artificial-Intelligence-for-Edge-Devices, 2020.

[25] P. S. Dutta, N. R. Jennings and L. Moreau, "Cooperative Information Sharing to Improve Distributed Learning in Multi-Agent Systems," Journal of Artificial Intelligence Research, vol. 24, p. 407–463, 2005.

[26] V. I. Bajić, L. Weisi and T. Yonghong, "Collaborative intelligence: Challenges and opportunities," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[27] Ericsson, "Edge Computing and 5G," [Online]. Available: https://www.ericsson.com/49d80b/assets/local/digital-services/trending/edge-computing/edge-computing-5g-report.pdf.

[28] J. Zhang and K. B. Letaief, "Mobile Edge Intelligence and Computing for the Internet of Vehicles," Proceedings of the IEEE, vol. 108, no. 2, pp. 246-261, 2020.

[29] 3GPP, ""Overall Description of Radio Access Network (RAN) Aspects for Vehicle-to-Everything (V2X) based on LTE

and NR", 3GPP TR 37.985, V.16.0.0," 2020.

[30] 3GPP, """Integrated Access and Backhaul Radio Transmission and Reception", 3GPP TS 38.174, V.16.3.0," 2021.

[31] 3GPP, """Study on RAN-centric data collection and utilization for LTE and NR", 3GPP TR 37.816, V.16.0.0," 2019.

[32] ETSI, "Mobile Edge Computing (MEC); Framework and Reference Architecture, ETSI GS MEC 003, V2.1.1," Jan. 2019.

[33] ETSI, "MEC in 5G Networks, ESTI White Ppaer No.28," Jun. 2018.

[34] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15)," June 2018.

[35] Ericsson, "Explainable AI – How Humans Can Trust AI," [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/explainable-ai--how-humans-can-trust-ai.

[36] ETSI, "MEC Framework and Reference Architecture," [Online]. Available: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/02.01.01_60/gs_MEC003v020101p.pdf.

[37] ETSI, "Traffic Management APIs," [Online]. Available: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/015/02.01.01_60/gs_mec015v020101p.pdf.

[38] R. 8743, "Multi-Access Management Service," [Online]. Available: https://www.rfc-editor.org/rfc/rfc8743.txt.

[39] 3GPP, "5G System; Access Traffic Steering, Switching and Splitting (ATSSS); Stage 3," [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3607.

[40] IETF, "RFC 2890 "Key and Sequence Number Extensions to GRE"," [Online]. Available: https://www.rfc-editor.org/rfc/rfc2890.txt.

[41] K. Sundaresan, S. Park and R. Sivakumar, "Transport Layer Protocols in Ad Hoc Networks," in Ad Hoc Networks, Springer, Boston, MA, 2005.

[42] I. K. a. K. Wehrle, "Transport Protocol Issues of In-Network Computing Systems," IETF, 2020.

[43] L. Peterson, T. Anderson, S. Katti, N. McKeown, G. Parulkar and J. Rexford, "Democratizing the Network Edge," in SIGCOMM CCR, 2019.

[44] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars and L. Tang, "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," in 22nd International Conference on Architectural Support for Programming Languages and Operating Systems, 2017.

[45] M. Eisen, K. Arjun, A. S. Baxi and D. Cavalcanti, "Network Performance Adaptation in Wireless Control with Reinforcement Learning," in 54th Asilomar Conference on Signals, Systems, and Computers, 2020.

[46] P. Kairouz, H. B. McMahan and e. al., "Advances and open problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, 2021.

[47] R. Balakrishnan, M. Akdeniz, S. Dhakal, A. Anand, A. Zeira and N. Himayat, "Resource Management and Model Personalization for Federated Learning over Wireless Edge Networks," Journal of Sensor and Actuator Networks, vol. 10, no. 17, 2021.

[48] A. Anand, S. Dhakal, M. Akdeniz, B. Edwards and N. Himayat, "Differentially Private Coded Federated Linear Regression," in IIEE SP Data Sciences and Learning workshop, 2021.

[49] M. Isaksson and K. Norrman, "Secure Federated Learning in 5G Mobile Networks," in IEEE Global Communications Conference, 2020.

[50] S. Niknam, H. S. Dhillon and J. H. Reed, "Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges," IEEE Communications Magazine, vol. 58, no. 6, pp. 46-51, 2020.

[51] Ericsson, "Edge computing and deployment strategies for communication service providers," [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/edge-computing-and-deployment-strategies-for-communication-service-providers.

[52] Intel, "Intel's Converged Edge Insights," [Online]. Available: https://www.intel.com/content/www/us/en/edge-computing/edge-software-hub-converged-edge-insights.html.

[53] J. Research, "INDUSTRIAL IOT: FUTURE MARKET OUTLOOK, TECHNOLOGY ANALYSIS & KEY PLAYERS 2020-2025,"

Oct. 2020.

[54] OPC, “OPC UA,” [Online]. Available: https://opcfoundation.org/about/opc-technologies/opc-ua/.

[55] IETF, “ICN IETF,” [Online]. Available: https://datatracker.ietf.org/rg/icnrg/about/ .

[56] ISA, “ISA 95,” [Online]. Available: https://www.isa.org/isa95/.

[57] C. Cities, “https://www.c40.org/why_cities”.

[58] WHO, “Road traffic injuries, https://www.who.int/health-topics/road-safety#tab=tab_1,” WHO, 2021.

[59] Intel, “The Future of Smart Road Infrastructure,” [Online]. Available: https://www.intel.com/content/www/us/en/transportation/smart-road-infrastructure.html.

[60] J. Clayton, “Crafting a Powerful Executive Summary,” Harvard Business School, 8 Sept. 2003. [Online]. Available: https://hbswk.hbs.edu/archive/crafting-a-powerful-executive-summary. [Accessed 13 Feb. 2020].

[61] O. Mayr, The Origins of Feedback Control, Clinton, MA USA: The Colonial Press, Inc., 1970.

[62] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang and B. Zhang, “Named Data Networking,” in ACM SIGCOMM Computer Communication Review (CCR), July 2014.

[63] C. Yi, J. Abraham, A. Afanasyev, L. Wang, B. Zhang and L. Zhang, “On the Role of Routing in Named Data Networking,” in ACM Conference on Information-Centric Networking, 2014.

[64] R. Pirmagomedov, S. Srikanteswara, D. Moltchanov, G. Arrobo, Y. Zhang, N. Himayat and Y. Koucheryavy, “Augmented Computing at the Edge Using Named Data Networking,” in IEEE Globecom Workshops (GC Wkshps), Dec. 2020.

[65] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang and B. Zhang, “Named Data Networking,” in ACM SIGCOMM Computer Communication Review (CCR), 2014.

[66] M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb and M. Zorzi, “Machine Learning at the Edge: A Data-Driven Architecture with Applications to 5G Cellular Networks,” IEEE Transactions on Mobile Computing, vol. doi: 10.1109/TMC.2020.2999852.

[67] S. Prakash, S. Dhakal, M. Akdeniz, Y. Yona, S. Talwar and N. Himayat, “Coded computing for low latency Federated Learning over Wireless Edge Networks,” IEEE Journal of Selected Areas in Communication, Special issue on ML for Communications and Networking, vol. 39, no. 1, 2021.

Acknowledgments

5G Americas' Mission Statement: 5G Americas facilitates and advocates for the advancement and transformation of LTE, 5G and beyond throughout the Americas.

5G Americas' Board of Governors members include Antel, Airspan Networks, AT&T, Cable & Wireless, Ciena, Cisco, Crown Castle, Ericsson, Intel, Liberty Latin America, Mavenir, Nokia, Qualcomm Incorporated, Samsung, Shaw Communications Inc., T-Mobile USA, Inc., Telefónica, VMware and WOM.

5G Americas would like to recognize the significant project leadership and important contributions of group leaders Meryem Simsek of VMWare and Clark Chen 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.