CableLabs®

Technical Brief

February 2023

How 3GPP Supports Edge Compute Transformation

R&D Wireless

Arun Yerra, Principal Mobile Network Architect | a.yerra@cablelabs.com
Yunjung Yi, Principal Architect & Director of Wireless Standardization | y.yi@cablelabs.com
Rahil Gandotra, Senior Software Architect | r.gandotra@cablelabs.com

Executive Summary

This technical brief describes network architecture, procedures, and application frameworks defined by 3GPP to support edge computing applications. With end-user applications and use cases maturing rapidly to require computing resources closer to end-user devices, edge computing applications are quickly becoming the most prominent consumers of communication service provider (CSP) networks, and they are forecasted to grow tremendously in the near term (40%–70% CAGR by LF Edge 2021 Survey). In the long term, edge computing infrastructure demand will be driven by "edge native" use cases, such as extended reality (XR), autonomous vehicles, and V2X communications, that would function when edge computing capabilities are available. Edge clouds are expected to be deployed at different levels of distribution based on the latency requirements of applications like real-time, near real-time, and non-real-time latency requirements. Application servers can be hosted at on-premises data centers or micro data centers located at the cell towers or network HUB points, or at metro, regional, or core data centers.

Edge computing applications need edge networks, i.e., networks that can support stringent requirements of end-to-end network latency, jitter, bandwidth, application-specific QoS, reliability, and availability. Challenges in edge networking go beyond latency and jitter; the network should support local break out, dynamic insertion of data network attachment points, and re-routing of traffic. Features like application-specific QoS, end-to-end network slicing, auto scaling based on network function virtualization, and ease of deployment make 5G networks the essential drivers of edge computing adoption. Convergence of 3GPP and non-3GPP access networks enables residential customer edge compute use cases to have mobility by allowing seamless transition between MSO and MNO networks. Therefore, edge computing and access convergence work happening within 3GPP is essential for CSPs to transform their networks to support edge computing.

Introduction

In the short to medium term, edge compute infrastructure demand will be driven by cloud applications that are enhanced with edge computing capabilities to solve edge use cases. However, in the long term, the use of edge native (aware) applications for solving use cases can only function when edge computing capabilities are available. These edge native use cases depend on the maturation of key technologies like augmented and virtual reality (AR and VR) and autonomous systems, such as those used for closed-loop enterprise IT functions.

Use Cases

Edge computing applications are gaining traction across all of the marketing segments, but for now, mobile and residential customers, enterprise IT, and CSPs are the main drivers of the adoption of edge computing. It is forecasted that by 2028, 36.5% of the global infrastructure edge footprint will be used for use cases associated with mobile and residential consumers, and 11.9% will be used for enterprise IT. However, as infrastructure matures, edge computing applications will gain a substantial foothold in other markets like manufacturing, smart cities, commercial unmanned aerial vehicles (UAVs), healthcare, retail, and vehicle-to-everything (V2X). This section describes some of the use cases in CSP and residential and mobile customers market segments.

Residential and Mobile Customers

Some of the use cases in the residential and mobile customers market segment include smart homes, streaming services and content delivery, and augmented and virtual reality.

Smart homes—The smart homes category contains home automation use cases like connected security systems, smart speakers, and AI-enabled virtual assistants. These use cases generate a huge volume of data, and processing data closer to home improves reliability and responsiveness and also optimizes transport costs.

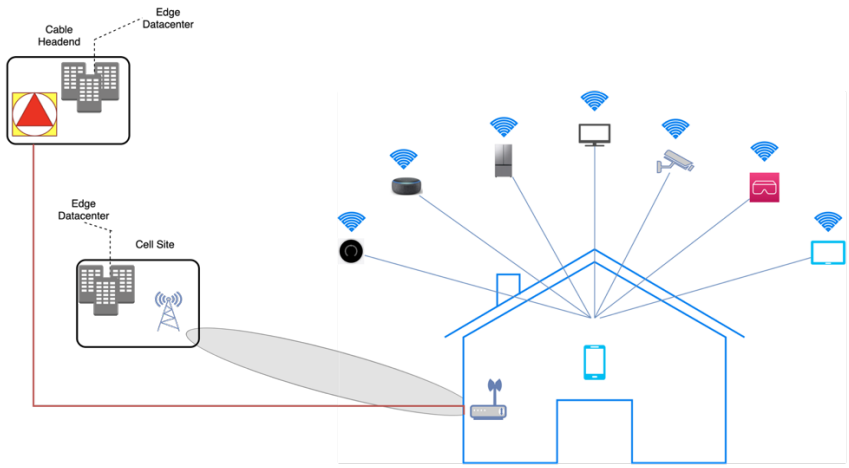

Figure 1. Edge Computing Use Cases for Residential Customers

Streaming services and content delivery—Edge computing applications can support low-latency requirements of video streaming and content delivery. They can also enhance the end-user experience for features such as search functions, content recommendations, personalized experiences, and interactive capabilities.

Augmented and virtual reality—AR and VR both require real-time processing of large data sets because any lag in analysis would delay subsequent actions. Delay in processing of AR and VR data sets may create a poor and sometimes unsafe user experience. Therefore, edge computing is essential for AR and VR, and it probably will be the first edge native application that gets wide adoption.

In-home network convergence—Network access convergence within a home network can enable end-user devices like smart phones, tablets, and smart appliances to use all available access networks and to share the bandwidth across all of the applications. Network telemetry data processing in real time can enable in-home smart devices and applications to seamlessly share the network's bandwidth.

Communication Service Providers

In addition to being the key providers of edge computing infrastructure in the near term, as well as in the long term, CSPs are also driving the early market momentum for the infrastructure edge as they virtualize their networks. It is forecast that in 2028, 10.9% of infrastructure edge deployments will support CSP use cases.

The advent of network virtualization, control and user plane separation (CUPS) architecture, and disaggregation of RAN networks is driving the edge compute infrastructure adoption by CSP networks. Disaggregation and virtualization of RAN components allow flexible deployment of components and use of edge-optimized open white box hardware closer to the radio, and some non-real-time functions are moved out to a central location. Servers that host radio components (like radio unit (RU) and distributed unit (DU)) can also host edge applications besides radio data path and control path processing. This hybrid model of deployment of 5G networks makes CSPs the early adopters of the edge computing paradigm.

Similarly, with residential customers, MSOs can transform their cable headend locations into edge data centers to host the edge applications. Besides offloading processing for use cases like AR and VR, home automation, and streaming services, MSOs can offload value-added services like intelligent parental controls, deep packet inspection (DPI), and real-time traffic steering.

Network Evolution for Edge Compute

Edge computing becomes possible when the underlying critical infrastructure is performant, highly available, and seamlessly integrated. Besides network latency and bandwidth, data sovereignty, security, control, and interconnection are also important. Applications will perform real-time computing at the edge and connect back to the core for less time-sensitive backend functions. Critical infrastructure for edge applications will keep moving from hyperscale data centers and regional data centers to micro edge data centers located in CSP networks, like at cell towers and cable headends. Edge compute infrastructure must be highly distributed and interconnected with strict latency and reliability requirements, and it requires multiple providers like CSPs, edge compute service providers (ECSPs), and application developers to work with each other. Therefore, edge computing needs an open, interoperable network connectivity between data centers and open, standards-based orchestration and operations management across multiple data centers.

Transitioning the CSP infrastructure to support edge compute use cases requires two distinct strategies. The first strategy involves upgrading existing CSP networks to be edge compute ready so that the current applications can service edge compute use cases. The second strategy involves developing brand-new edge compute-aware applications.

- Edge-unaware applications—Edge-unaware applications are cloud native applications that are needed to support an edge compute use case but are not capable of interfacing with edge compute infrastructure. These applications support client device mobility, and they need support from CSP infrastructure to deliver strict service requirements associated with edge compute use cases by connecting to the most optimal edge server.
- Edge-aware applications—Edge-aware applications utilize edge compute infrastructure features and application programming interfaces (APIs) and support stringent service requirements of edge compute use cases. Application clients can pick the most optimal server based on client device location and application service requirements. They can also support client device mobility and connect to new application server(s) with application context relocation.

3GPP SA2 and SA6 groups have defined standards-based networks and application architectures so that CSP networks can support edge compute use cases and the associated service requirements. SA2 defined a network architecture that can enable current cloud native applications to migrate to edge compute infrastructure. SA6 defined an edge compute application framework that enables applications to utilize 3GPP core network capabilities and support edge compute use cases (it provides a skeleton framework for edge native application development).

5GS Edge Compute Reference Architecture

CSP networks are evolving to support edge compute use cases and enable ECSPs to deploy edge solutions. 3GPP SA2 extended the 5GS network architecture to support edge compute applications. SA2 defined a new network function called Edge Application Server Discovery Function (EASDF) to support optimal server discovery and UE mobility events.

The following reference architecture for non-roaming and local breakout (LBO) roaming scenarios further depicts the relationship between the 5GS and a data network (DN), where edge application servers (EASs) are deployed. Edge application traffic can be routed to an EAS hosted in a data network-connected local site user plane function (UPF) by uplink (UL) classifier (CL).

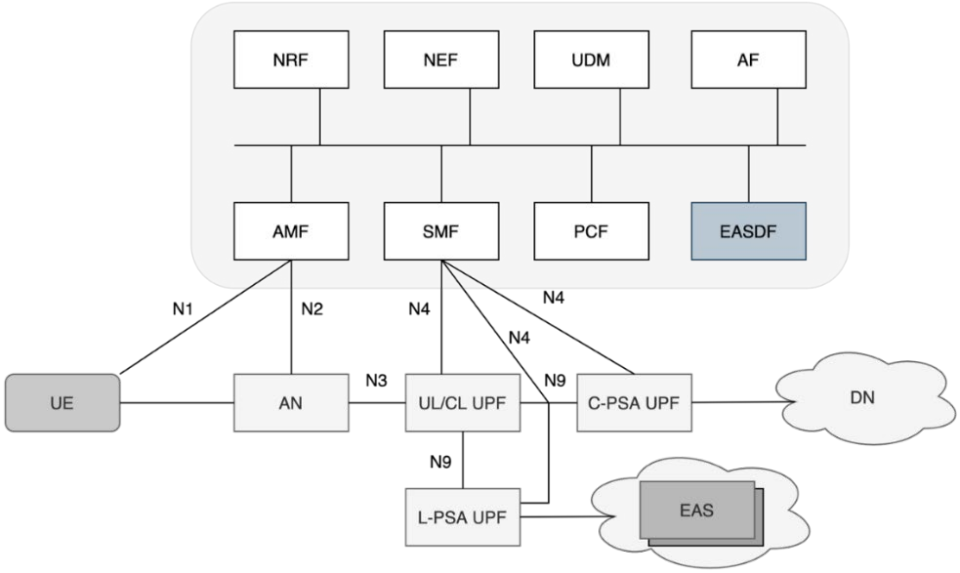

Figure 2. 5GS LBO Reference Architecture

SA2 also defined a framework for a client-side functionality that is needed to discover and connect to an optimal edge server during service provisioning or service updates. The application client (AC) is provisioned with application server(s) identified with fully qualified domain name (FQDN) address(es).

A 3GPP-defined edge domain name system (DNS) client (EDC) feature ensures that DNS requests from applications are sent to an edge-aware DNS (e.g., EASDF/DNS resolver), and EAS discovery and re-discovery procedures are utilized for identifying the optimal application server. The main extension of the DNS client within the EDC is the configuration of EASDF as DNS using the Protocol Configuration Option (PCO) field during protocol data unit (PDU) session establishment/modification or through Dynamic Host Configuration Protocol (DHCP).

The 5G core network is extended to include an EASDF that acts as a DNS resolver and notifies matched DNS messages to the session management function (SMF). The EASDF matches received DNS messages (queries and responses) against DNS message handling rules and DNS baseline patterns configured by the SMF and executes the corresponding actions, like forwarding the messages to a DNS and notifying the SMF, as well as inserting/replacing/removing the DNS ECS option.

SA6 Edge Compute Application Framework

3GPP SA6 has defined edge compute application architecture for edge-aware applications. The edge-aware client and server applications have the knowledge of edge infrastructure, and they could continually optimize applications based on the edge infrastructure state. Figure 3 shows the layers in 3GPP-defined architecture.

- Application layer—Edge-aware client and server applications developed by the application developers
- Edge enabler layer—Edge infrastructure components that enable edge-aware client and server modules to monitor and handle edge infrastructure events like network state changes, edge relocation, etc.
- Edge hosting environment—Edge compute optimized hardware infrastructure that hosts the edge applications, hosted in the edge data network (EDN) that is connected to the local UPF in the service provider network
- Edge management layer—Management and orchestration of edge applications and infrastructure to support highly dynamic edge application architecture
- 3GPP transport layer—Service provider network that connects edge clients with the backend servers

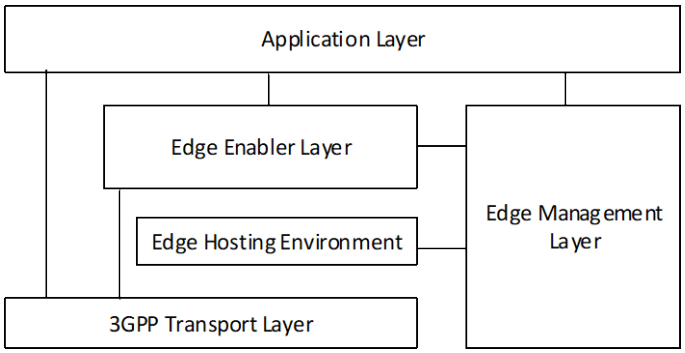

Figure 3. Edge Compute Application Architecture (Courtesy 3GPP TS 23.558)

The edge enabler layer consists of three main components: edge enabler client, edge enabler server, and edge configuration server.

Edge Enabler Client (EEC)	The EEC supports application clients to discover the application servers, detects UE mobility events, and retrieves configuration information to enable communication with application servers.
Edge Enabler Server (EES)	The EES enables ACs to connect to application servers by registering and dynamically instantiating the application servers as needed and supports application context relocation.
Edge Configuration Server (ECS)	The ECS provides identity and connectivity information of registered EAS servers to EECs. It supports registration of EES(s) and interfaces with the 3GPP core network for accessing the capabilities of network functions.

3GPP SA6 has defined edge reference architecture with reference points for enabling edge applications. SA6 defined the reference points and interfaces between the EEL and application layer components.

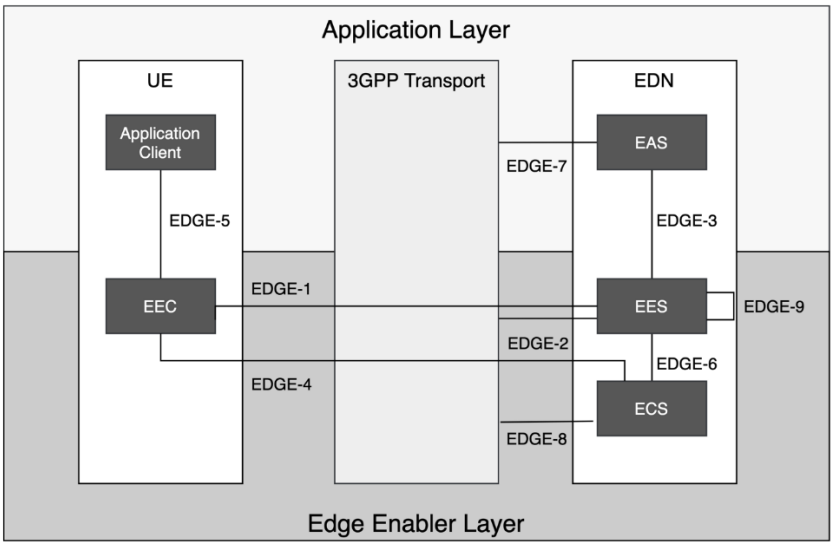

Figure 4. SA6 Edge Compute Application Reference Architecture

EDGE-1 and EDGE-4 reference points connect the EEC with the EES and ECS, respectively, for authentication and registration of each component with others and provisioning optimal EAS servers and EEC configuration provisioning. EDGE-6 refers to interactions between the ECS and the EES, which include registration and de-registration of the EES to the ECS and retrieval of configuration from the ECS. EDGE-2, EDGE-7, and EDGE-8 reference points identify the APIs with the 3GPP core network.

EDGE-3 refers to interactions between the EES and the EASs, which include registration and de-registration of the EAS with the EES and provides the EAS with network capability information. EDGE-5 refers to interactions between AC(s) and the EEC, which include providing EAS(s) with information and passing application profiles for EAS discovery.

Service Provisioning

This section discusses UE service provisioning procedures defined by both SA2 and SA6 working groups to support edge-unaware and edge-aware applications, respectively. Service provisioning procedures allow the edge compute application within a client device to connect to the most optimal backend server that satisfies the application service preferences and requirements.

SA2 Provisioning	SA6 Provisioning
The EAS discovery procedure is used to discover the optimal edge application server. The AC discovers the IP address(es) of the optimal edge application server(s) using the DNS.	UE service provisioning procedures are used to discover the optimal edge application server. The EEC sends EAS discovery filters to the EES to find the optimal servers after mutual authentication and registration.
The AC initiates EAS discovery procedures when an application session needs to be established and its DNS cache is clear. The AC uses the EDC client to send the DNS requests after establishing the PDU session or after PDU session modification.	The EEC initiates EAS discovery by communicating with the EES to obtain information about available EASs based on matching EAS discovery filters.
The DNS request will be sent to the EASDF, which acts as a DNS resolver (configured by the SMF during PDU establishment with the ePCO option).	The EAS discovery filters may include UE location and a list of preferred EAS(s) based on the UE's expected route. SA6 has defined the information model for EEC context, EAS discovery filters, and application service KPIs.
The EASDF matches received DNS messages (queries and responses) against DNS message handling rules and DNS baseline patterns configured by the SMF and executes the corresponding actions, like forwarding the messages to a DNS and notifying the SMF, as well as inserting/replacing/removing the DNS ECS option.	The EES will check that the EEC is authorized for EAS discovery, and if the UE's location is not available, the EES obtains the UE location using 3GPP core network capabilities. The EES identifies the EAS(s) based on the provided EAS discovery filters and the UE location.
The L-DNS or C-DNS will resolve the FQDN to the most optimal EAS server based on either the source IP address of the incoming DNS query or the ECS (EDNS client subnet) option, which identifies the UE IP address. How the DNSs are configured with DNS rules is outside the purview of SA2. ECSPs and MNOs configure these servers based on the location and EAS deployment information.	Upon successful application server discovery, the EES informs the EEC of the discovered EAS(s), application KPIs, supported regions, and ACR scenarios. The EES also may notify the application function (AF) (AF influence), which, in turn, can optimize the data path by using 3GPP network exposure function (NEF) provisioning APIs.
Upon receiving the DNS response, the EASDF forwards the message to the SMF, which, in turn, will identify the target DNAI based on the IP address of the EAS and may trigger data path optimization by dynamically inserting the L-PSA UPF and UL CL to steer the application traffic.	Upon receiving the EAS discovery response, the EEC uses the endpoint information for routing of the outgoing application data traffic to the EAS(s) and informs the application client of the discovered EAS.
Upon receiving the DNS response, the UE will cache the DNS record for subsequent use, and it continues to use the DNS record until the DNS record is deleted because of UE mobility or PDU session/IP session release.	The EEC will cache the EAS information for subsequent use, and the application client also may cache the discovered EAS.

Edge Relocation

When a UE moves to a new location, a different server can be more suitable for serving the applications than the current instance. Even in non-mobility cases, the serving application server can become sub-optimal based on the current network and server load or maintenance of the application server. To serve the application better, the client should transition from the current server instance (source EAS/S-EAS) to a new application server (target EAS/T-EAS). The feature that supports the transfer of the application server with minimal service interruption is called edge relocation or application context relocation.

SA2 Edge Relocation	SA6 Edge Relocation
Edge relocation requires selection of a new EAS(s), and it also may require reconfiguration of the data path between the UE and EAS(s) if the L-PSA UPF changes because of UE mobility.	SA6 identified the following roles to support ACR: Detection Entity, Decision-Making Entity, and Execution Entity. Based on the ACR scenario, different entities within a given edge deployment can play these roles.
Application server rediscovery happens with the same EAS discovery procedure used in service provisioning. The EAS discovery procedure is triggered whenever a new PDU session is established or when a session IP address is renewed. Hence, edge relocation can be supported by 3GPP-defined service continuity modes SSC Mode-2 and SSC Mode-3.	A detection entity detects the probable need for ACR by monitoring various aspects, such as the UE's location and server profiles. The decision-making entity consumes the detection entity events and decides if ACR is needed. The execution entity performs ACR after the decision-making entity decides in favor of ACR.
When the SSC mode is Mode-2, the PDU session may be released along with the session address (IP address). When the UE detects the PDU session has been released or a new IP prefix has been allocated within the PDU session, the EDC client within the UE initiates EAS discovery by sending a DNS query to discover the new EAS IP address.	Edge enablement layer entities (EEC, EES, and EAS) perform these tasks based on the edge relocation scenario. In some scenarios, only a new target EAS (T-EAS) needs to be discovered, whereas in some mobility scenarios, the EES will also change, and the T-EES has to be identified as well. Therefore, ACT (application context transfer) may involve S-EES and T-EES besides S-EAS and T-EAS.
If the SSC mode is Mode-3, the network allows the establishment of UE connectivity via a new PDU session anchor to the same data network before connectivity between the UE and the previous PDU session anchor is released. The EDC client within the UE will initiate EAS discovery with the new PDU session.	SA6 identified the following edge relocation scenarios, and in each scenario, the roles played by EEL components change. ▪ EEC with regular EAS discovery ▪ EEC executed ACR via S-EES ▪ S-EAS decided ACR ▪ S-EES executed ACR ▪ T-EES executed ACR
Like EAS discovery in service provisioning, optimal EAS selections happen within the C-DNS or L-DNS, and EAS selection criteria and framework is outside the scope of 3GPP SA2.	Upon successful ACT, the EES informs the EEC of the discovered EAS(s), application KPIs, supported regions, and ACR scenarios. The EES also may notify the AF (AF influence), which, in turn, can optimize the data path by using 3GPP NEF provisioning APIs.
The UE removes the old DNS cache related to old/removed IP address/prefixes during session release. Upon receiving the DNS response, the UE will cache the DNS record for subsequent use, and it continues to use the DNS record until the DNS record is deleted because of UE mobility or PDU session/IP session release.	The EEC will cache the EAS information for subsequent use, and the application client also may cache the discovered EAS.

Summary

Network and application architectures defined by 3GPP to support edge compute use cases within both SA2 and SA6 working groups can be complementary to each other. SA2 network architecture defines extensions to 3GPP 5G system architecture to enable cloud native and edge-unaware applications to support edge compute use cases. Application discovery with SA2 still requires the C-DNSs or L-DNSs to be configured with optimal EAS(s) for a given FQDN. SA6 framework can be used loosely to match AC KPIs with EAS capabilities to select the most optimal EAS. The SA6 edge compute application framework defines the framework for edge-aware applications to utilize and interface with the 3GPP 5G core so that applications continue to honor strict service requirements. UE data path optimization within the network dynamic L-PSA UPF and UL CL insertion requires support from 3GPP service continuity modes.

SA2 edge compute extensions also define procedures to enable SA6 edge enabler layer entities to communicate the configuration information. For example, the EEC entity within the UE can be configured with ECS connectivity information during PDU session establishment by the SMF (which fetches ECS address configuration information from the unified data management (UDM) together with session management and user subscription information). ECS address configuration information can be provided to 5GC through AF using NEF APIs. Similarly, an AF related to edge computing may guide the PCF on proper UE route selection policy (URSP) rules via application guidance for URSP rules determination mechanisms. The SA2 5GS system defined frameworks and network functions that provide well defined northbound APIs and reference points that allow a SA6 edge enabler layer to either query CSP network state or subscribe to CSP network events. SA6 architecture reference points EDGE 2/7/8 use these features defined by SA2 5GS to interface with CSP networks.

Essentially, both SA2 and SA6 define standards-based frameworks and reference architecture to transition CSP networks to support edge compute solutions. Additionally, these standards provide an open and collaborative space for CSPs, ECSPs, and ASPs to innovate and deploy next-generation edge applications.

Why CableLabs?

As edge computing is being adopted more rapidly across multiple market segments, infrastructure support is as important as technological breakthroughs in niche markets like AR/VR and autonomous vehicles. MNOs and MSOs play a key essential role in ramping up the edge compute infrastructure because they own the last-mile infrastructure and real estate to connect end-user devices to edge data centers. Edge computing will be a key driver in transforming MNO and MSO networks in the short term, as well as in the long term. Edge computing is a great example of collaboration between hyper scalers, CSPs, application developers, and niche startups, which is required to mature this market—only open, standards-based collaboration can take this forward, and the work happening between 3GPP and ETSI MEC is essential for edge compute infrastructure transformation.

CableLabs is participating in 3GPP edge initiatives and will continue to provide the latest updates to its members about the work happening in these organizations. Besides the standards development organizations (SDOs), CableLabs is also working with LF Edge initiatives and open-source projects like CAMARA, Adrenaline, and other initiatives like Network as a Platform. These initiatives will help the networks transform to be able to support edge computing solutions.

Acronyms

ACR	application context relocation
AF	application function
API	application programming interface
AR/XR	augmented reality/extended reality
CAGR	compound annual growth rate
C-DNS	central DNS server
CSP	communication service provider
CUPS	control and user plane separation
DN	data network
DNAI	data network access identifier
DPI	deep packet inspection
DU	distributed unit
ECSP	edge compute service provider
FQDN	fully qualified domain name
LBO	local breakout
L-DNS	local DNS server
LPWA	low powered wide area
NEF	network exposure function
PDU	protocol data unit
RAN	radio access network
RU	radio unit
SDO	standards development organization
SMF	session management function
UAV	unmanned aerial vehicle
UL	uplink
UL CL	uplink classifier
UPF	user plane function
URSP	UE route selection policy
V2X	vehicle to everything

References

3GPP TS 23.501, System Architecture for the 5G System (5GS)

3GPP TS 23.502, Procedures for the 5G System (5GS)

3GPP TS 23.548, 5G System Enhancements for Edge Computing; Stage 2

3GPP TR 23.758, Study on application architecture for enabling Edge Applications

3GPP TS 23.558, Architecture for enabling Edge Applications

State of the Edge 2021: A Market and Ecosystem Report for Edge Computing