SCTE CABLE-TEC EXPO FAST FORWARD 2021
VIRTUAL EXPERIENCE
OCTOBER 11-14

2021 Fall Technical Forum
SCTE • NCTA • CABLELABS

Edge Computing Architecture

A Technical Paper prepared for SCTE by

Umamaheswar (Achari) Kakinada
Director, Wireless R&D
Charter Communications, Inc
6360 S Fiddlers Green, Greenwood Village, CO 80111
847-544-6560
Achari.Kakinada@charter.com

Deh-Min Richard Wu
Director, Wireless R&D
Charter Communications, Inc
6360 S Fiddlers Green, Greenwood Village, CO 80111
256-763-1202
deh-minrichard.wu@charter.com

Curt Wong
Senior Director, Wireless R&D
Charter Communications, Inc
6360 S Fiddlers Green, Greenwood Village, CO 80111
425-395-4379
Curt.Wong@charter.com

Yildirim Sahin
Director, Wireless R&D
Charter Communications, Inc
6360 S Fiddlers Green, Greenwood Village, CO 80111
720-536-9394
Yildirim.Sahin@charter.com

Table of Contents

Title **Page Number**

1. Introduction ... 3
2. Motivation For Edge Computing ... 3
   - 2.1. Latency ... 4
   - 2.2. Privacy & Security ... 4
   - 2.3. Data Volume And Backhaul Bandwidth ... 4
   - 2.4. Need For Autonomy ... 5
3. Edge Computing Architecture & Standards ... 5
   - 3.1. 3GPP ... 5
   - 3.2. ETSI ... 7
   - 3.3. IETF ... 8
4. Dimensions Of Edge Computing ... 8
5. Edge Computing Continuum ... 9
6. Analytics And Intelligence At The Edge ... 11
7. Edge Orchestration And Deployment ... 11
8. Conclusion ... 12

Abbreviations ... 13

Bibliography & References ... 15

List of Figures

Title **Page Number**

Figure 1 - Key motivators for Edge Computing ... 3
Figure 2 - 3GPP Edge Computing Architecture - Roaming with UL CL/BP ... 5
Figure 3 - 3GPP Edge Computing Connectivity Model ... 6
Figure 4 - ETSI Multi Access Edge Computing framework ... 7
Figure 5 - IETF Beyond Edge Computing ... 8
Figure 6 - Verticals that may benefit from Edge Computing ... 9
Figure 7 - Edge Computing Continuum ... 10
Figure 8 - Edge Deployment Continuum ... 10
Figure 9 I & II - ETSI MEC Deployment ... 12

1. Introduction

Trillions of gigabytes of data is being generated/captured by devices and network systems, which need to be analyzed and processed. Forbes estimates [1] that in the year of 2025, 150 zettabytes of data will need to be analyzed and processed. Over half of this data is expected to come from the edge of the network (79.4 Zetta bytes by 2025 [2]). The world's most valuable resource is no longer oil, but data [3, 4]. While oil is a limited natural resource which needs to be preserved and conserved. On the contrary data is experiencing explosive growth and showing no signs of slowing down anytime soon. Data need to be managed, processed and analyzed to derive value from it. Conventionally, this data is sent to the centralized systems usually in the cloud for processing, analyzing and deriving insights. It is an enormous task at hand. This method of processing the data incurs significant latency and huge amount transport cost. Which often renders the derived stale insights not useful for most latency sensitive applications.

Edge computing (EC) addresses and mitigates some of these issues. In this paper we look at different aspects of EC, comprising of addressing the latency in data processing, analyzing and deriving insights; architecture, standards and deployment considerations.

2. Motivation For Edge Computing

EC supports placement of compute, storage and other processing resources needed for performing analytics and deriving insights in a given scenario, along a wide spectrum of deployment scenarios ranging from centralized data center to individual devices or somewhere in between either extreme, to address the latency, storage and any special processing needs of a given application or service. The key motivators of EC are latency reduction, enhanced privacy and security, backhaul bandwidth optimization and enabling autonomous decision making at the edge of the networks [5]. This is depicted in Figure 1.

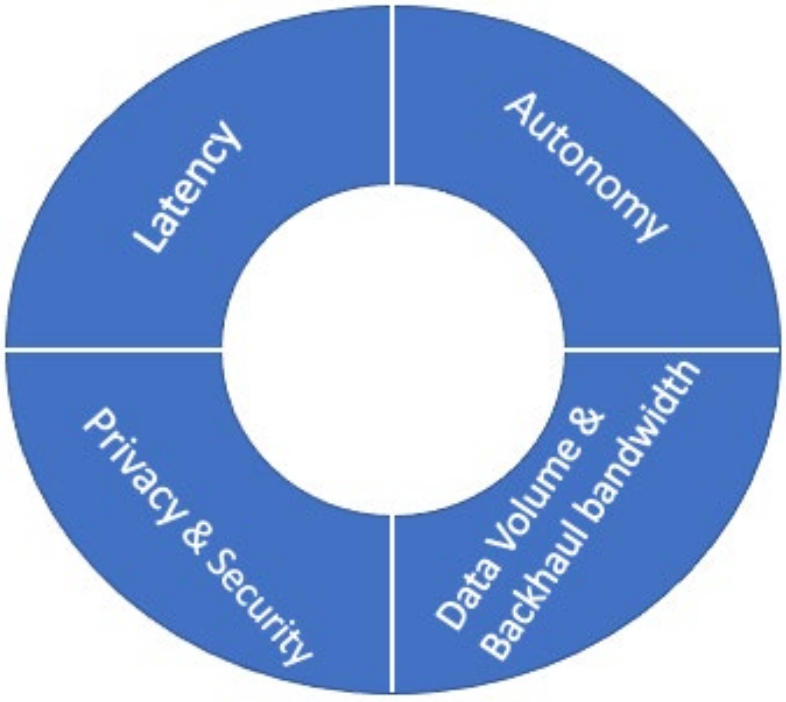

Figure 1 - Key motivators for Edge Computing

2.1. Latency

Zettabytes of data is being generated [2] at the network edge. Often there is a time-critical need to process and analyze these data, derive insights and take actions based on the learnings from these insights. It may not be viable to transfer this data to a central core and wait for the decisions to be made as there may be opportunity costs and safety issues because of this latency. Reducing latency is imperative[9] for many applications in Industry 4.0, healthcare, smart cities, aviation, autonomous driving, enterprises, entertainment, and augmented reality/virtual reality (AR/VR). Data need to be processed with a deterministic latency in a timely manner and EC addresses this critical need. We will look into various EC architectures and deployment scenarios currently being explored in various standards bodies and in the industry at large to meet different latency requirements. The EC enables agile service response and facilitates logic execution closer to the end users and devices both temporally and spatially.

2.2. Privacy & Security

In many centralized data processing scenarios such as cloud-based services, the user and/or device-generated data is transported to the central data center for processing, deriving analytics and taking actions. There are many regulatory requirements to ensure the privacy of the data as to origin, identity of the users/devices, etc. For instance, the energy/water usage reports from individual meters, if associated with an individually identifiable user may impact the privacy and security of this person. Another instance can be getting a count of number of vehicles in a road transport network without sending individually identifiable information about any vehicle or person, aggregating their number per segment/area, deriving analytics and insights to formulate a strategy for optimal traffic management. Further, these regulations vary widely across different sates/counties/municipalities, each requiring compliance with its own set of rules and regulations. This phenomena in not unique to smart cities scenario; it extends to other verticals such as healthcare and enterprises.

EC can act as an intermediary between the user data and centralized servers in the cloud or on premises. The EC can provide the desired anonymity for the sensitive data of the individual devices and users; and aggregate such data to minimize the processing at the central servers and reduce the costs of raw data transportation. It also can facilitate implementation of local rules and policies, and ensure compliance with regulations of the individual administrative domain (cities, healthcare systems, enterprises etc.). Additionally, EC provides autonomy and control often needed by these entities which are responsible for the data generated by users and devices in their respective domains.

2.3. Data Volume And Backhaul Bandwidth

Large, continuous data streams from huge number of devices/end-points can be burdensome on backhaul networks [5]. The overhead associated with the transfer of data from the devices to the central application servers, increases manyfold. Each data point being transported incurs overhead at each of the protocol layers, often these are small amounts of data compared to the transport overhead associated with it. In many instances, applications can benefit from aggregation of these data in edge nodes before being sent to the central servers. For example, a water meter can continuously collect the consumption data, while it probably needs to send them to the central server few times a day to meet the desired application processing needs. In some industrial 4.0 applications, the control may need to be gathered and logged, but may not be need to be reported to central servers for each individual data point. However, any exceptions, such as the temperature of a control unit becoming unusually high and needing immediate attention, may be sent immediately as a high-priority message by the edge node. Also, during normal course of operation, many systems collect large number of data points which are of interest for a short duration, after short life this data may lose it relevance, EC can address processing this type of data effectively. The

EC can act in this manner and conserve bandwidth usage on the backhaul, reduce the transportation costs and decongest the transport network while effectively meeting the application needs.

2.4. Need For Autonomy

Often the regulatory requirements for data autonomy, where the data is generated vary widely. This is the case for smart cities in which each city and local governments are responsible to define the regulation for maintaining the data and ensuring that it is secured. This calls for autonomy to do certain processing locally where the data is generated. Also, many applications in areas such as industry 4.0, need local context aware data processing, which are not suitable for central cloud computing. The EC enables meeting the local processing needs, also facilitates data processing per local policies and compliance with local regulations. The same needs exist in many other verticals, such as hospitality, health care, industry 4.0, enterprises and smart cities.

With the above-mentioned characteristics, EC is an enabler and plays a pivotal role in transitioning from centralized to distributed computing. It also makes it a key component of transition from centralized Industry 3.0 to distributed, data-driven and latency-sensitive Industry 4.0 paradigm. The same can be applicable to many other verticals. It minimizes time-to-insight, time-to-action by an order of magnitude, also minimizes cost-of-insight, in the process enabling timely and effective decision making and opening many new avenues of opportunities.

3. Edge Computing Architecture & Standards

3.1. 3GPP

Figure 2 and Figure 3 [14] provide a reference architecture for supporting EC per 3GPP specifications. The Figure 2 reference architecture depicts 5GS architecture for roaming scenario supporting EC with uplink classifier/branch point (UL CL/BP) support.

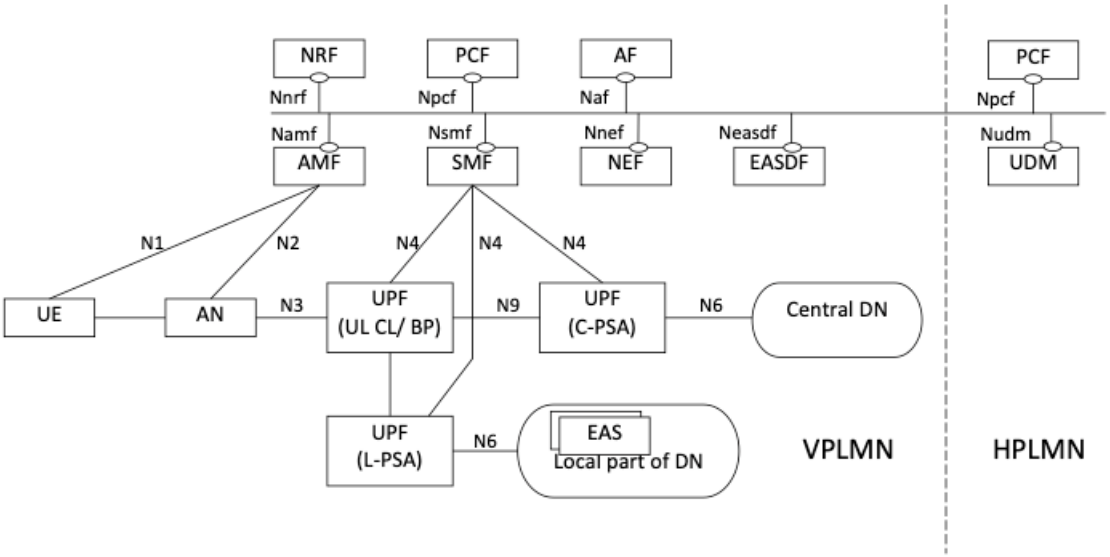

Figure 2 - 3GPP Edge Computing Architecture - Roaming with UL CL/BP

The 5G core per 3GPP standards specifications [14] supports the following connectivity models to enable EC:

- *Distributed anchor point*: For a protocol data unit (PDU) session, the PDU session anchor user plane function (PSA UPF) is in a local site, i.e., close to the UE location. The PSA UPF may be changed due to UE mobility.
- *Session breakout:* A PDU session has a PSA UPF in a central site (C-PSA UPF) and one or more PSA UPF in the local site (L-PSA UPF). The C-PSA UPF provides the IP Anchor Point when UL classifier is used. The EC application traffic is selectively diverted to the L-PSA UPF using UL classifier or multi-homing branching point mechanisms. The L-PSA UPF may be changed due to UE mobility.
- *Multiple PDU sessions:* EC applications use PDU Session(s) with a PSA UPF(s) in local site(s). The rest of applications use PDU session(s) with PSA UPF(s) in the central site(s). Any PSA UPF may be changed due to e.g., UE mobility and using session and service continuity (SSC) mode 3 with multiple PDU Sessions.

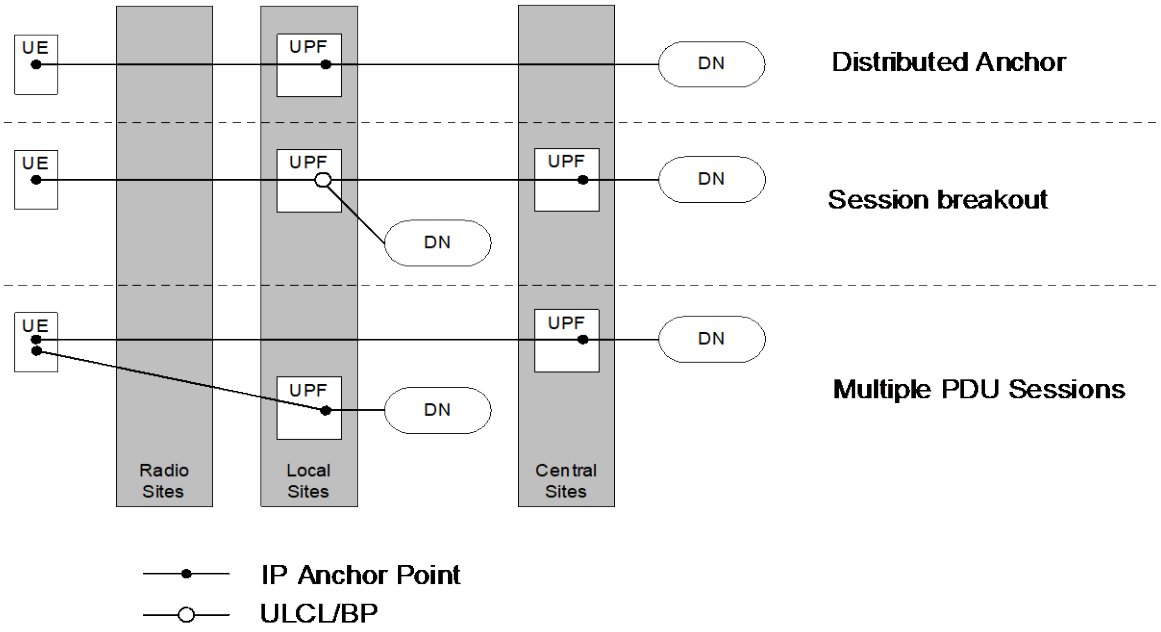

Figure 3 - 3GPP Edge Computing Connectivity Model

The following are some of the salient features of 5GS [14] support for EC:

- Flexible placement of UPFs in the network, which provide IP anchor point or different branch points in the network
- Simultaneous connections to multiple data networks
- Support for multi-homed PDU sessions, using either UL CL or BP
- Enhanced SSC: a connection through a new PDU session anchor point is established before the previous connection is terminated
- The application function (AF) can make a request to influence traffic routing for a given UE

Standardization efforts are currently ongoing for the 3GPP Rel.17 of the specifications [14]. The scope for Rel.17 includes:

- Edge application server (EAS) discovery and re-discovery
- Edge relocation
- Network exposure to EAS
- Support of 3GPP application layer architecture for enabling EC
- Services of EAS discovery function (EASDF) for EAS discovery, DNS etc.

3.2. ETSI

Figure 4 provides the framework of ETSI multi access mobile edge computing (MEC) for the deployment in a network functions virtualization infrastructure and virtualized network functions (NFVI/VNF) environment [15]. MEC offers to application developers and content providers cloud-computing capabilities and an information technology service environment at the edge of the network. MEC is access agnostic, providing flexibility in the operator network; managing different types of sites where the location of the edge will depend on the use case and needed activities to be performed. MEC system incorporates two levels: the MEC host level and system level. The former consists of the MEC host, the platform and the virtualization infrastructure manager, while the latter is composed by the MEC orchestrator, the operations support system.

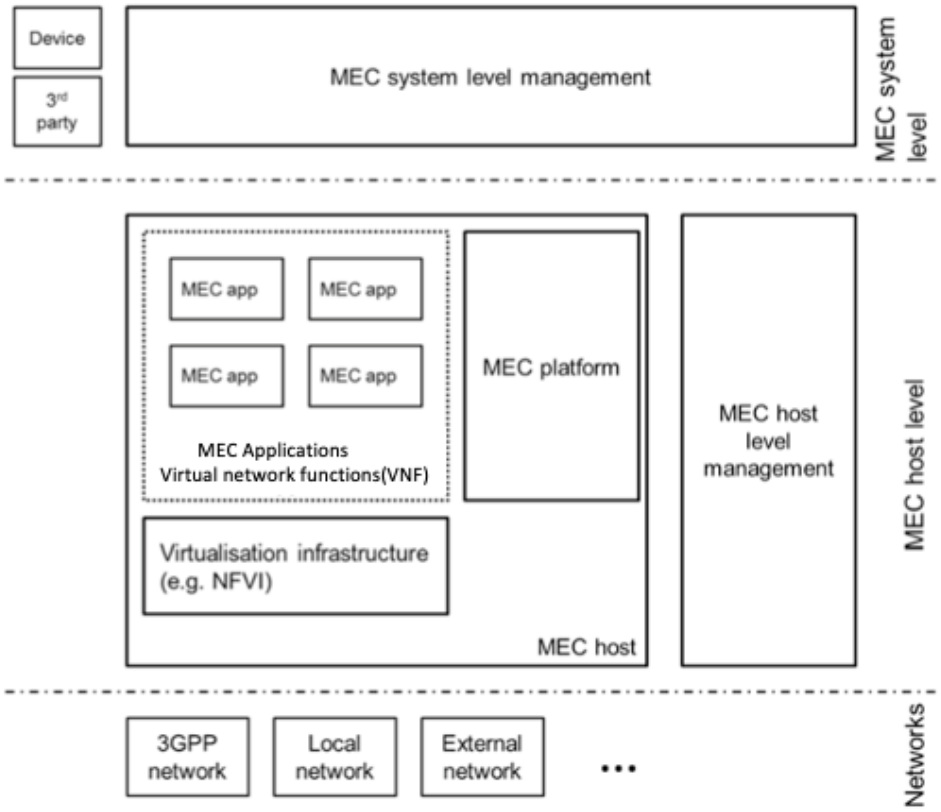

Figure 4 - ETSI Multi Access Edge Computing framework

3.3. IETF

The motto of IETF for beyond edge computing (BEC) methodology is – "Distribute as much as you can, centralized only if you must" [13]. IETF BEC aims to further research and standardize the protocol between multiple BEC gateways, common API across various BEC platforms, user mobility: edge to edge, edge device configuration/management, light-weight virtualization technologies (container/uni-kernel) and local edge security. BEC platform is depicted in Figure 5. The approach here is push the applications and use cases which are latency sensitive towards the edge of the network i.e., edgification of the system. On the other hand, system which scale well, benefit from clustered or centralized processing pushed deeper into the core of the network, i.e., cloudification approach for the system is adapted.

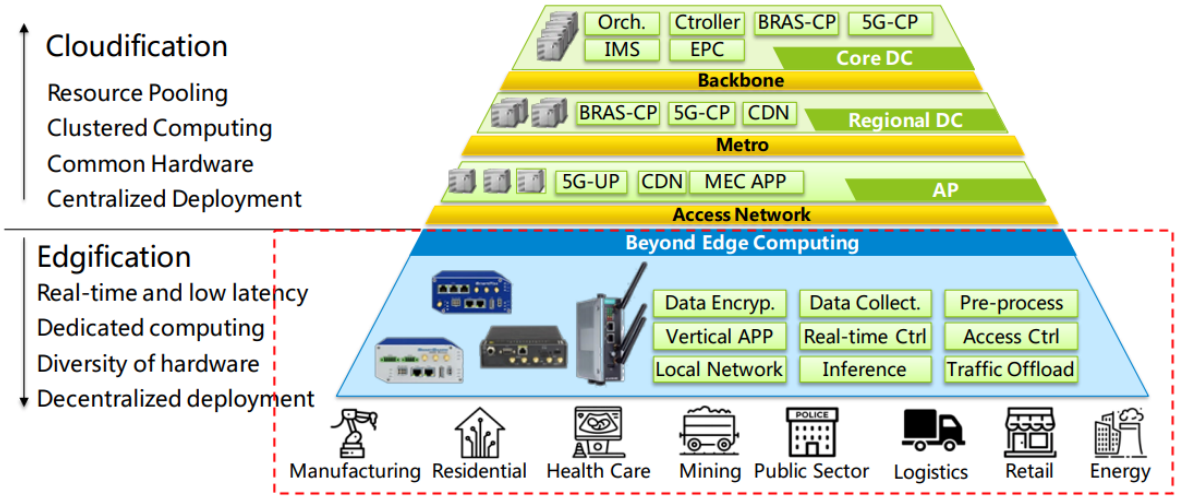

Figure 5 - IETF Beyond Edge Computing

4. Dimensions Of Edge Computing

Figure 6 below (adapted from [10]) depicts different dimensions of EC, the degree of importance these dimensions assume in catering to the needs of different verticals differ. The dimensions of EC are:

- *Time scale/Responsiveness* – real-time, near-real-time, and non-real-time.
- *Communication span* – device-2-device (D2D), device to near edge, device to center of network. How many other entities a given entity interacts with to perform its function, determines the amount of data exchanged and mutual dependence between the systems. Some of the devices may act independently and report their state (e.g., temperature sensors, humidity sensors etc.); while a cooling system may take all this info from multiple sensors and coordinate with coolant, fans etc. to achieve desired temperature control. Similarly, traffic sensors and traffic control systems act in tandem but with different degrees communication span.
- *Degree of control/autonomy* – fully autonomous entities which act independently based on local info such robotics systems, connected vehicles etc.; Systems with some autonomy and local decision making based on predefined policies and rules; and centrally controlled system which collect the data and report to central system and act as per instructions from central system.
- *Location* – immediate vicinity/local/confined to the entity, regional and global. The sphere of influence of a given entity in performing its function.

To this end while connected vehicle, critical healthcare, Industry 4.0 equipment may need real-time responsiveness together with a large degree of autonomy to respond to changes, other applications such as traffic control systems may function well with near real-time response and regionally coordinated control, some enterprise applications may be processed with centralized services as to the response time and processing needs. A given EC deployment for specific requirement need to be tuned based on the processing needs, need for autonomy, sensitivity to response times and specificity for location.

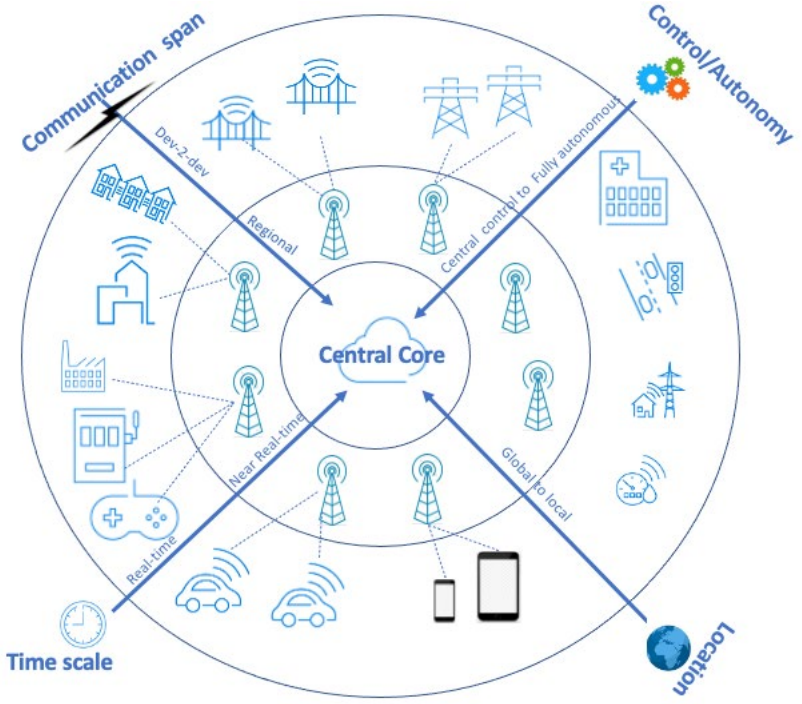

Figure 6 - Verticals that may benefit from Edge Computing

5. Edge Computing Continuum

EC is the key technology, that can support innovative services for a wide ecosystem, stakeholders for EC range from operators, infrastructure providers, application and content providers in the continuum as depicted below. It can span a variety of network locations, form factors, network, and application functions. Centralized computing is performed deeper into the network possibly in the cloud if it suits a given application, and is often used for latency tolerant bulk processing for large number of users. In contrast for latency sensitive tasks requiring some autonomy and local decision-making EC works better. In EC storage and analytics of data is performed closer to end users and devices, in close proximity to where such data is generated. This may differ for different applications, sometimes for the same application there may be multiple points of EC for different types of data in the continuum of the EC depicted above. The 3GPP and ETSI architectures depicted in Figures 2-3-4 enable realization of any combination of this as needed for an application, service or enterprise use case such as industrial control, healthcare, hospitality industry, video analytics, smart city functions, AR/VR/XR applications. Figure 7 describes the EC continuum based on latency and processing needs, assumes ultra reliable low latency communication (URLLC) for access network.

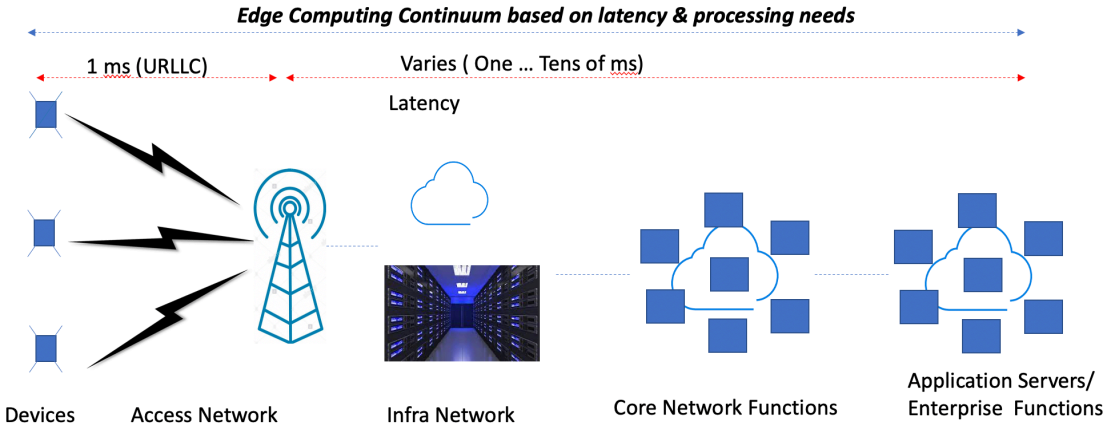

Figure 7 - Edge Computing Continuum

The spatial and temporal proximity needs between devices and application services and systems offering these services is determined by the characteristics of the services under consideration. These include real-time responsiveness, mobility, interactivity, criticality of the function (Industry 4.0, healthcare etc.). These characteristics together with the costs and affordability of a given solution shall largely influence EC deployment in the continuum depicted above. The Figure 8 adapted from [5] depicts continuum of EC from physical end point to central data center, and potential ways to instantiate and scale these. For instance, the edge instances from physical endpoint through compute edge may be singular nodes, the edge deployment from local data center through central cloud could be a cluster of nodes or cloud, based on resource needs and cost considerations.

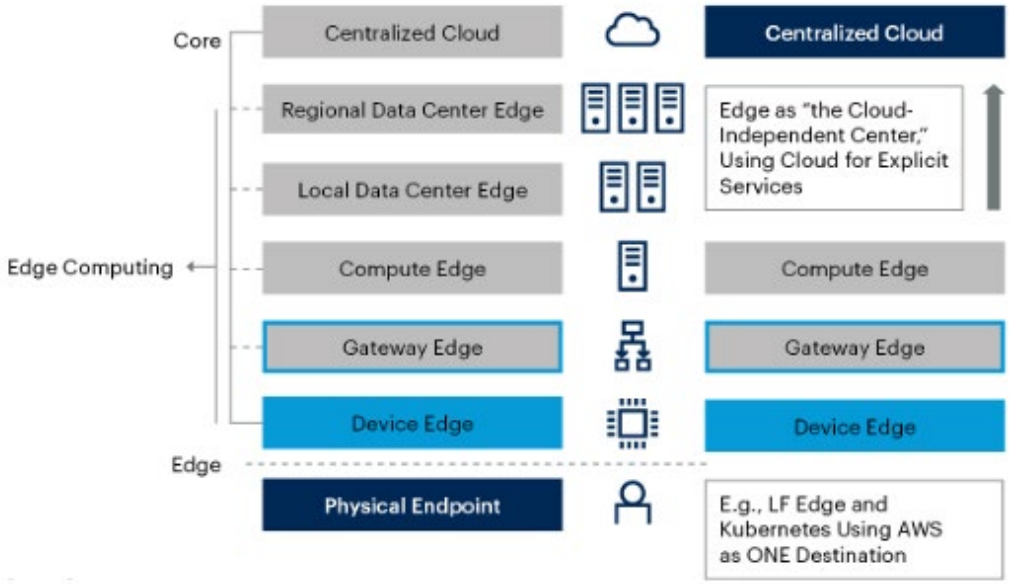

Figure 8 - Edge Deployment Continuum

The original vision for EC is to provide compute and storage resources closer to the user in open standards and in a ubiquitous manner [11,12]. EC is a crucial computing paradigm for multiple verticals IoT, AR/VR/XR, Industry 4.0, and smart cities. The basic characteristics of EC, compute, storage and latency vary widely among these verticals. The specifications from standards bodies like 3GPP and solution offerings from service providers and equipment vendors also have evolved to fulfill these needs. EC specifications from 3GPP currently can support positioning of multiple instances of UPF function at different points network to address latency and data processing needs. Similarly, solutions from vendors provide edge cloud services (e.g., AWS outpost, Azure edge etc.) and standalone servers conducive for EC from many providers.

6. Analytics And Intelligence At The Edge

There is a huge amount of data being generated by the large number of devices connected to the network both through wired and wireless networks. This data is often transported to the central core for processing, analysis and to discern insights and act on them. Often due to the latency of the transport networks and delay in processing, the full potential of this data remains untapped. Therefore, the processing of this data at the edge of the network using various analytics and deep learning (DL) techniques enables deriving insights performing timely actions to realize the value.

Edge computing enables incorporation of DL analytics technology such as computer vision (CV) – in a parking lot to detect expiry of parking duration for a car, to detect availability of parking spot, and natural language processing (NLP) – to provide context sensitive localized information about certain operations. EC can also be beneficial to many applications such as AR/VR/XR, gaming, Industry 4.0 applications, smart cities, health care and hospitality are only to name a few. Incorporation of DL into edge is a huge enabler, can make it possible many previously not feasible applications and use cases, also enhance the quality of experience for existing applications e.g., online gaming.

To realize this potential, there needs to be a match between the capability of the EC nodes, and services, DL processing needs and adequacy of the offered accuracy, latency, energy, memory footprint [9]. The available processing power, energy (battery powered devices/small nodes) and memory in the EC are often the bottleneck. Some of them can be addressed by training the models in a central location in the cloud doing much of the heavy lifting and deploying the trained models at the edge to process the locally generated data and derive insights. The key is tailoring the EC resource and DL models to match one another and meet the needs of the applications. 3GPP has intimated an effort for analytics and machine learning based insights using a federated learning (FL) model, which can potentially fulfill some of these needs.

7. Edge Orchestration And Deployment

EC imposes a few unique challenges to orchestration. Namely:

- *Resource constraints:* resources such as power (e.g., battery powered edge devices/systems), CPU, storage may be severely constrained at the edge
- *Scale:* the number of edge instances may be large, ensuring consistency and coherence across several instances may be difficult
- *Autonomy:* due to resource constraints and loss of communication may necessitate autonomous operation at the edge for extended periods of time

The degree of importance of above-mentioned factors needs to be taken into account in orchestration and deployment of an EC instance. The leading orchestration tools (OpenStack, Kubernetes, ONAP) are

currently focusing on large scale cloud deployments, some these are being adapted for edge orchestration and are still evolving.

ETSI MEC deployment model [16] allows EC to be accessible to a wide range of mobile devices with reduced latency. Figure 9 below depicts, how to make the edge services accessible to large number of devices even when the devices' current access network does not offer the service. While the device during its mobility reaches the target network, which does not offer the application services previously availed by the device, as depicted in Figure 9-II (b) the target MEC host can reach to the application services in prior serving network or alternatively as depicted in Figure9-II (c) the UPF in the target network can reach the application server through the UPF in the prior serving network.

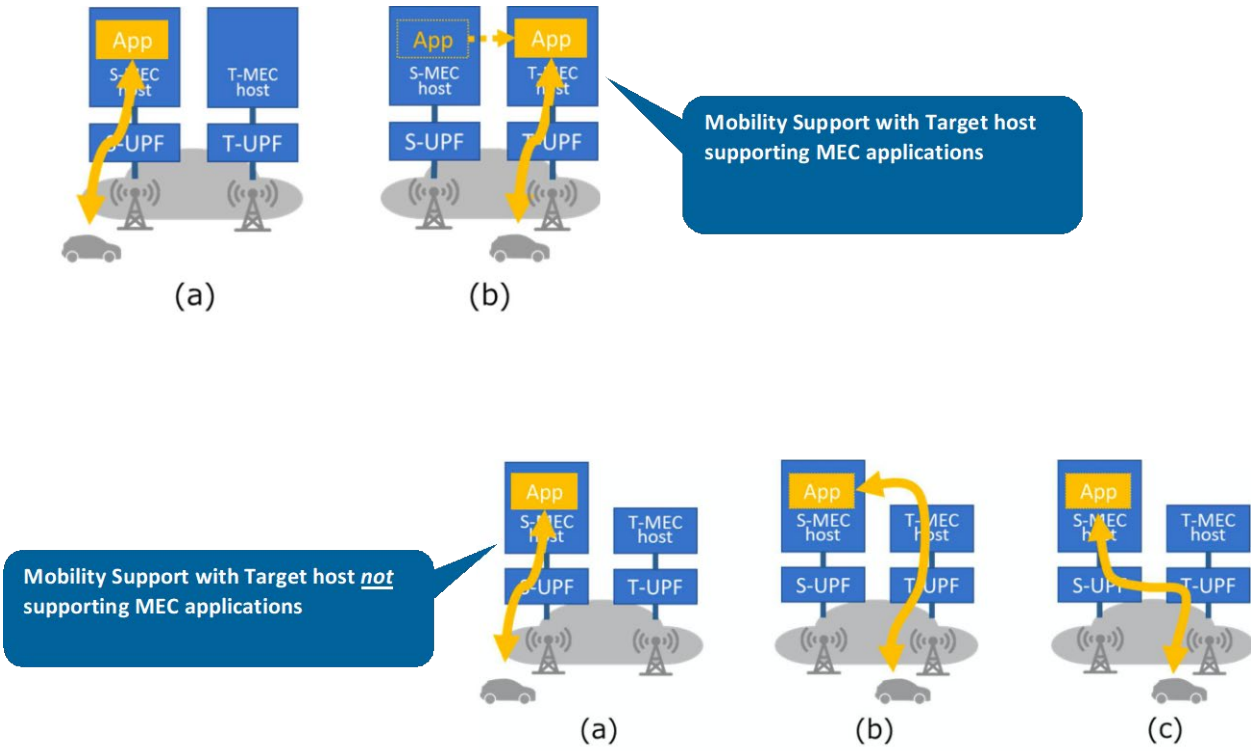

Figure 9 I & II - ETSI MEC Deployment

8. Conclusion

In the last few years significant strides have been made in enhancing the EC architecture, yet there is a need for more improvement. Currently some of EC efforts appear to be in specific silos marked for specific applications, software stacks, sources of data being used, specific cloud and network service providers. This fragmentation across different silos, service providers and multitude of software stacks constrain the stakeholders from realizing the full potential of EC. There is a need for holistic integration of these diverse domains through standardization, industry alliances and market forces. Some of it is being addressed in various standards bodies such as 3GPP, ETSI, IETF and various industry alliances and projects (MANO, ONAP, Kubernetes, etc.). It is imperative on all the stake holders to harmonize and accelerate these efforts across different standards bodies and industry alliances to realize the full potential of EC. EC is already delivering on its promise by significantly optimizing on time-to-insight, time-to-action and cost-of-insight, in the process enabling timely and effective decision making and opening new avenues of opportunities.

Abbreviations

3GPP	3rd Generation Partnership Project
5G-CP	5G control plane
AF	application function
AMF	access management function
AN	access network
AR/VR/XR	augmented reality/virtual reality/extended reality
BEC	beyond edge computing
BRAS-CP	broadband remote access server- control plane
CDN	content delivery network
C-PSA	central PSA
CV	computer vision
D2D	device to device
DL	deep learning
DN	data network
EAS	edge application server
EASDF	edge application server discovery function
EC	edge computing
EPC	enhanced packet core
ETSI	European Telecommunications Standards Institute
FL	federated learning
HPLMN	home PLMN
IETF	Internet Engineering Task Force
L-PSA	local PSA
MANO	management orchestration
MEC	multi access mobile edge computing
NEF	network exposure function
NFV	network functions virtualization
NFVI	network functions virtualization infrastructure
NFVO	NFV orchestrator
NLP	natural language processing
NRF	network repository function
ONAP	open network automation platform
OSS	operations support system
PCF	policy control function
PDU	protocol data unit
PLMN	public land mobile network
PSA	PDU session anchor
SSC	session and service continuity
SMF	session management function
UDM	unified data management
UE	user equipment
UPF	user plane function
UL CL/BP	uplink classifier/branch point

URLLC	ultra reliable low latency communications
VIM	virtualization infrastructure manager
VNF	virtualized network function
VPLMN	visiting PLMN

Bibliography & References

[1] https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/?sh=6a0cc89120d7

[2] https://www.iotacommunications.com/blog/iot-big-data/

[3] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[4] https://www.weforum.org/agenda/2018/01/data-is-not-the-new-oil/

[5] Gartner, 2021 Strategic Roadmap for Edge Computing, https://www.gartner.com/doc/reprints?id=1-24JFAZOO&ct=201104&st=sb

[6] https://www.usenix.org/sites/default/files/conference/protected-files/hotedge18_slides_bhardwaj.pdf

[7] https://www.usenix.org/system/files/conference/hotedge18/hotedge18-papers-bhardwaj.pdf

[8] Edge Exchange, Bhardwaj et al. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8790194

[9] Convergence of Edge Computing and Deep Learning: A Comprehensive Survey, Xiaofei Wang, Yiwen Han, Victor C. M. Leung , Dusit Niyato, Xueqiang Yan, and Xu Chen, IEEE

[10] Fog Computing: Principles, Architectures, and Applications, Amir Vahid Dastjerdi, Harshit Gupta, Rodrigo N. Calheiros, Soumya K. Ghosh, and Rajkumar Buyya

[11] All One Needs to Know about Fog Computing and Related Edge Computing Paradigms: A complete Survey, A. Yousefpour, C. Fung, T. Nguyen, K. Kadiyala, F. Jalali, A. Niakanlahiji, J. Kong, and J. P. Jue, *J. Syst. Archit.*, vol. 98, pp. 289–330, Sep. 2019

[12] OpenEdgeConsortium: About - The Who, What, and How, http://openedgecomputing.org/about.html, Technical Report, OpenEdge Computing

[13] IETF: Problem Statement of Edge Computing Beyond Access Network for Industrial IoT, draft-geng-iiot-edge-computing-problem-statement-00

[14] 5G System Enhancements for Edge Computing, Stage 2 (3GPP TS 23.548)

[15] Multi Access Edge Computing (MEC): Framework and Reference Architecture, ETSI GS MEC 003 V2.2.1

[16] Multi Access Edge Computing (MEC): MEC 5G Integration, ETSI GR MEC 031 V2.1.1

