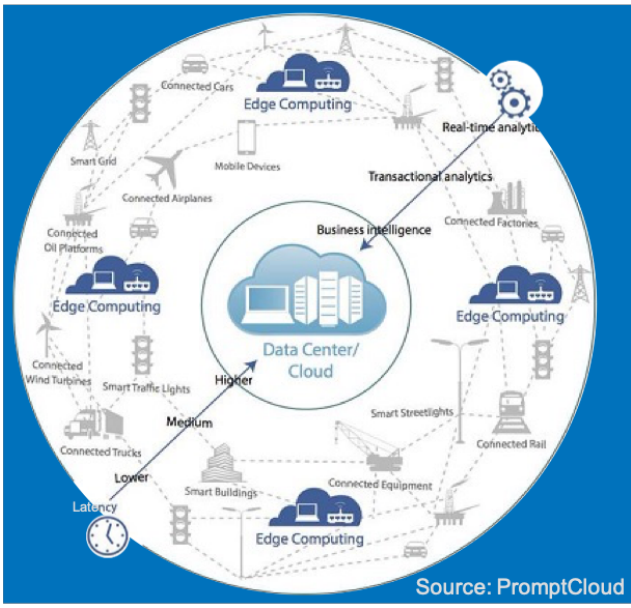
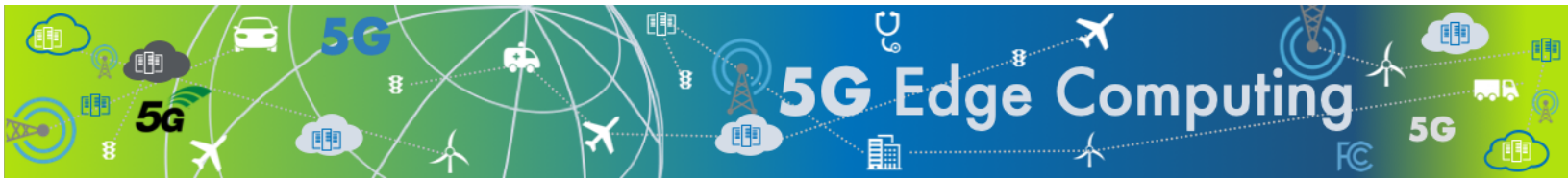

# 5G EDGE COMPUTING WHITEPAPER




## FCC Technological Advisory Council

## 5G IoT Working Group

Walter Johnston, Kevin Sparks, Brian Daly, Russ Gyurek, Kumar Balachandran, John Barnhill, Lynn Merrill, Brian Markwalter, Adam Drobot, Jeff Foerster, Dale Hatfield




# TABLE OF CONTENTS


## 5G Architecture View of Edge Computing

Edge computing refers to locating applications – and the general-purpose compute, storage, and associated switching and control functions needed to run them - relatively close to end users and/or IoT endpoints. This greatly benefits applications performance and associated QoE, and it can also improve efficiency and thus the economics depending on the nature of the specific application.

Distributed edge computing is analogous to, and can be regarded as an extension of, the evolution of content distribution over the last few decades. To improve the performance and efficiency of delivering content – especially the video content dominating broadband traffic – global CDN operators' nodes are widely distributed down to network peering points, and in many cases down to CDN appliances located well inside BIAS networks. Broadband service providers also typically deploy distributed CDNs for their own in-network content distribution. While distributed CDNs mostly revolve around storage caches, enabling applications with edge compute extends this to both compute and storage, and the more general cloud service stack necessary to on-board and run 3rd party applications.

Enabling applications to be localized in edge compute close to end users first and foremost improves network transit latency. Latency is a significant driver in improved performance as is high reliability, e.g. through setting up radio bearers that allow low block error rate tolerance. Edge compute combined with the optimized latency performance of 5G NR air interface and 5G Core processing can reduce round-trip-time by up to two orders of magnitude in situations where there is tight control over all parts of the communication chain. This will enable new classes of cloud applications, in such areas as industrial robotic/drone automation, V2X, and AR/VR infotainment, and associated innovative business models.

Edge compute is also important for localization of data and efficient data processing. Industry and Government regulations may require localization of data for security and privacy reasons. Certain application scenarios may pose restrictions on the use of excessive transport bandwidth or may require transport to external sites to be scheduled by time-of-day, requiring local storage or caching of information. Additionally, there may need to be local processing of information to reduce the volume of traffic over transport resources.

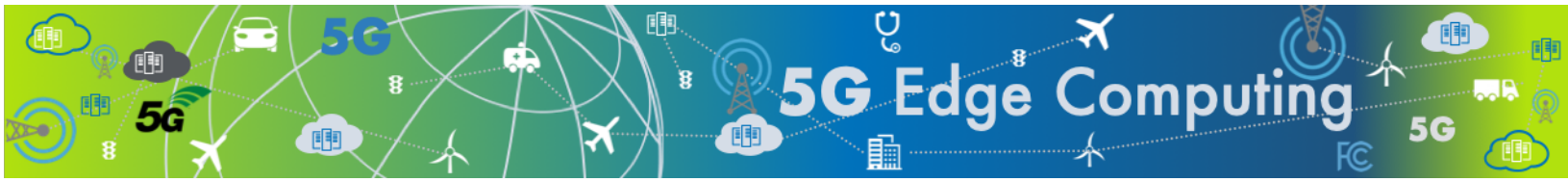
Edge clouds are expected to be deployed at different levels of distribution, which may be phased in over time. Core data centers exist in networks today, typically at regional levels (a


few per country), and will continue to host centralized network functions. Metro level edge clouds, both network operator owned as well as operator-neutral entities, will host low latency broad-based consumer and enterprise applications. 'Far-edge clouds will be located within a few 10's of km of end users in network Central Offices or cell towers and will host ultra-low latency and/or high reliability applications and may initially be targeted opportunistically at high value industrial automation and IoT users. On-premise clouds sit within enterprise locations and serve similarly stringent applications; they exist as fully-owned private clouds today but are expected to increasingly be addressed by webscale cloud operators and network operators extending their cloud capabilities to the very edge.

In addition to hosting new 5G era services, the other major network operator driver for edge compute and edge clouds is deploying virtualized network infrastructure, replacing many dedicated hardware-based elements with virtual network functions (VNFs) running on general purpose edge compute. Even portions of access networks are being virtualized, and many of these functions need to be deployed close to end users. The combination of these infrastructure and applications drivers is a major reason that so much of 5G era network transformation resolves around edge cloud distribution.

According to a new Gartner report[1], "Around 10% of enterprise-generated data is created and processed outside a traditional centralized data center or cloud. By 2022, Gartner predicts this figure will reach 75%". Gartner defines edge computing as solutions that facilitate data processing at or near the source of data generation. For example, in the context of the Internet of Things (IoT), the sources of data generation are usually things with sensors or embedded devices. Edge computing serves as the decentralized extension of the campus networks, cellular networks, data center networks or the cloud.

[1] *https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/*




## Technologies & Standards: Definition and Work Related to Edge Computing

Edge compute for 5G era networks builds on innovations from many different parts of the information and communications technology (ICT) sector. Modern compute, storage and switching technologies are, of course, the hardware foundation of any type of cloud implementation. Cloud virtualization, orchestration and management software is similarly essential to be able to offer and on-board any cloud services. And Software Defined Networking (SDN) is also key to virtualize the interconnectivity of functions within and between clouds.

More specifically for network edge clouds, Network Functions Virtualization (NFV) enables cloud levels of dynamics and flexibility for network implementation, which in turn is a key enabler for providing dynamic network slicing vital for 5G services. Many of these network functions, as well as the applications running in the edge cloud, require hardware acceleration (in the form of network processors, GPUs, ARM processor arrays, and/or even dedicated ASICs, depending on functionality) to handle the high computational, signal processing, throughput and low latency demands.

A key architectural innovation for the packet core (vEPC and 5GC) is 'control & user plane separation' (CUPS), which allows multiple levels of user plane gateways corresponding to multiple levels of edge cloud distribution and applications placement. Further, the ETSI Multi-Access Edge Compute (MEC) ISG has defined enablement functions to support application placement in distributed edge clouds. This includes an application hosting environment and APIs to provide network intelligence to applications (e.g. current loading levels on different access types, mobility event triggers for applications that need to transfer state to another application instance in a new serving edge cloud).

## Use Cases Related to Mobile Edge Computing

Edge computing and processing aren't new concepts, so why are we talking about the edge? Existing and upcoming next-generation technologies such as the Internet of Things (IoT), software-defined networking (SDN), blockchain, and 5G are fueling innovations in the development of software applications across several industries. These emerging technologies require massive amounts of near real-time computation to deliver content to users and relay real-time data to centralized computing centers. To adapt and digitally transform, enterprises must develop effective strategies for navigating the opportunities and challenges of edge intelligence.

Edge computing brings multiple benefits to telecommunications companies[2]:

- reducing backhaul traffic by keeping right content at the edge,
- maintaining Quality of Experience (QoE) to subscribers with edge processing,
- reducing TCO by decomposing and dis-aggregating access functions,
- reducing cost by optimizing the current infrastructure hosted in central offices with low cost edge solutions,
- improving the reliability of the network by distributing content between edge and centralized datacenters,
- creating an opportunity for 3rd party cloud providers to host their edge clouds on the telco real estate.

The computational resources can be distributed geographically in a variety of location types (e.g., central offices, public buildings, customer premises, etc.,) depending on the use case requirements.

As more computing power is deployed in technologies at the network edge, it's clear that computing resources will become more widely distributed across the networking landscape. Centralized cloud compute environments will continue to operate and will be augmented with edge computing resources, which will be reliant on network capacity that supports edge technologies' traffic and services.

Enterprises are deriving benefits from edge computing in the form of enhanced security, lower latency that enables faster analysis and decision-making, and more efficient utilization of

[2] *https://about.att.com/content/dam/innovationdocs/Edge_Compute_White_Paper%20FINAL2.pdf*

network capacity enabled by sending less data to the centralized data center for processing. However, for these advantages to be achieved, network capacity needs to be easy to manage, flexible and agile so it can be dimensioned to support the computing needs of the enterprise efficiently.

As more computing power is sent to the network edge, it will need a foundation in order to be utilized. A software-defined infrastructure may be the launch pad to a fully virtualized network and functions. A virtualized network is dynamic, flexible and supports the rapid instantiation of functions to support customer demands.

Many industry experts are pushing back on the notion that cloud and edge computing are in competition with each other. Instead, forward-looking organizations and, even many public cloud service providers, are beginning to consider how to selectively employ both. While cloud adoption remains a critical focus for many organizations, a new era of connected devices is simultaneously transferring data collection and computing power to the edge of networks.

Both cloud and edge computing have their advantages and challenges. The next hurdle for IT teams is determining how to get the best of both. While cloud adoption remains a critical focus for many organizations, a new era of connected devices is simultaneously transferring data collection and computing power to the edge of networks.

Small-scale data centers offer another approach. By deploying these data centers in strategic geographic locations, companies can move data processing closer to the end-user or device. Doing so provides similar benefits as edge computing, while still maintaining the centralized management benefits that enterprises love about the cloud.

The strategy is certainly gaining momentum. Sales of so-called micro-modular data centers (MMDCs) may reach nearly $30 million this year, up from $18 million in 2017, according to 451 Research[3]. The report notes that while the overall spend may seem small, MMDCs are playing a significant role in thousands of expensive projects aimed at localizing computer processing power.

Edge computing isn't an all-or-nothing proposition. Centralized cloud services aren't going anywhere, but there is a need for complementary edge computing capabilities to enable next-generation devices. It's possible to process most important data at the edge, and then shift

[3] *https://www.networkworld.com/article/3238476/data-center/micro-modular-data-centers-set-to-multiply.html*

remaining data to centralized facilities. A hybrid solution can allow an industry such as financial services to thrive: edge technologies deliver real-time, fast experiences to customers and provide the flexibility to meet industry requirements with centralized data storage.

For enterprises, the data deluge will continue. Going forward, edge technologies will often be part of the solution stack for organizations overwhelmed by their computing needs – but likely not the only answer.

Today's applications – and those just on the horizon – are high-performance and power hungry. They generate significant amounts of data and require real-time computing power. Consider how much computation will be required to put self-driving cars on the road. Certain systems, like braking, will be controlled by the car's internal systems and require immediate responses. With traditional networks, a device sends information to a data center that may be hundreds of miles away. Data takes time to travel across large physical distances. As a result, delays can occur. With edge computing, critical functions can be processed at the network's edge in real-time. Data from secondary systems, such as updating the car's maps or managing the onboard infotainment system, can be processed in the cloud.

Edge technologies make it feel like every device is a supercomputer. Digital processes become lightning fast. Critical data is processed the edge of the network, right on the device. Secondary systems and less urgent data are sent to the cloud and processed there. With SDN, organizations have more flexibility to define rules on where and how data is processed to optimize application performance and the user experience.

When paired with 5G, which promises faster speeds and lower latency, edge computing offers a future with near real-time, back-and-forth connections.

Moving data processing closer to the network edge has security implications. With software-defined networking, it's possible to develop a multi-layered approach to security that takes the communication layer, hardware layer and cloud security into consideration simultaneously.

There are multiple edge open source and standard initiatives (e.g., ONAP, Open Stack, ONF, CNCF, ETSI MEC, OPNFV, Open Compute Project, LNF Akraino, 3GPP, etc.,) that are converging to create an ecosystem that will support edge computing and services.


Use cases where edge computing can bring new value[4]:

- Autonomous Vehicles
  - Self-driving cars need to be able to learn things without having to connect back to the cloud to process data
  - According to some third-party estimates, self-driving cars will generate as much as 3.6 terabytes of data per hour from the clusters of cameras and other sensors. Some functions like braking, turning and acceleration will likely always be managed by the computer systems in the cars themselves. But what if we could offload some of the secondary systems to the cloud? These include things like updating and accessing detailed navigation maps.
- Industrial Automation
  - Help create machines that sense, detect, learn things without having to be programmed
  - Edge computing could spark the next generation of robotic manufacturing. The future 5G service could play a vital role in what's called "Industry 4.0 – Digital Manufacturing". The anticipated low-latency wireless connections could eliminate traditional wired connections to robotic assemblers. Updates come quicker. Products can get to market faster.
- Augmented reality (AR) and virtual reality (VR)
  - Creating entirely virtual worlds or overlaying digital images and graphics on top of the real world in a convincing way also requires a lot of processing power. Even when phones can deliver that horsepower, the tradeoff is extremely short battery life.
  - Edge computing addresses those obstacles by moving the computation into the cloud in a way that feels seamless. It's like having a wireless supercomputer follow you everywhere.
- Retail
  - Creating more immersive in-store environments with technologies like AR
- Connected homes and offices
  - Complete tasks like turning on lights on command or changing the temperature. With edge computing, it will be possible for them to happen in near real-time
- Predictive Maintenance
  - Help detect machines that are in danger of breaking, and find the right fix before they do
- Video monitoring
  - Handle data at the edge rather than sending to the cloud
- Software-defined networking
  - Require local processing to find the best route to send data at each point of the journey
- Fog computing
  - Uses edge devices to connect to a distributed computing model

---

[4] *https://www.zdnet.com/google-amp/article/10-scenarios-where-edge-computing-can-bring-new-value/*


## How Edge Computing Will Evolve the Communications Network

Because of the twin drivers of network function virtualization and new latency sensitive end user applications both requiring cloud infrastructure distributed to the edge of the network, edge computing is having a central transformational impact on the way the networks are implemented. Edge clouds will host virtualized access functions (e.g. Cloud RAN baseband processing and control) and the core user plane and service chaining functions needed to terminate traffic destined for applications present at each cloud level.

It is likely that the low latency - high reliability applications at the metro cloud level, and the ultra-low latency/high reliability applications at the far-edge and on-premise levels, will represent only a fraction of overall applications. The larger share of relatively latency-insensitive generic applications is expected to continue to be hosted in large centralized clouds with their economies of scale. However, the more specialized 5G era applications promise to be high value use cases that will drive innovative business models and new transformative value creation. An implication of this is that partnership business models between 'application and content providers' (ACPs) and network providers and/or edge cloud providers will become very important to realizing this new 5G era services potential.

## Glossary - Definitions

| Term | Description |
|---|---|
| 2G | Second Generation Mobile Network |
| 3G | Third Generation Mobile Network GSM |
| 3GPP | Third Generation Partnership Project |
| 4G | Fourth Generation Mobile Network |
| 5G | Fifth Generation Mobile Network |
| 5G NR | 5G New Radio |
| 5GC | 5G Core Network |
| 5GNB | Fifth Generation NodeB |
| 5GPPP | Fifth Generation Private Public Partnership |
| 5GS | Fifth Generation System IMT |
| ACP | Application and Content Provider |
| API | Application Program Interface |
| AR | Augmented Reality |
| ARM | Advanced RISC Machine |
| ASIC | Application-Specific Integrated Circuit |
| BIAS | Broadband Internet Access Service |
| CDN | Content Delivery Network |
| CNCF | Cloud Native Computing Foundation |
| CUPS | Control Plane – User Plane Separation |
| ETSI | European Telecommunications Standards Institute |
| fog | extended concept of cloud computing at the network edge |
| GPU | Graphics Processing Unit |
| ICT | Information and Communications Technology |
| IoT | Internet of Things |
| ISG | Industry Specification Group (ETSI) |
| LNF Akraino | Linus Foundation Software Stack Supporting High-Availability Cloud Services Optimized for Edge |
| MEC | Multi-Access Edge Compute |

| | |
|---|---|
| MMDC | Micro-Modular Data Center |
| NFV | Network Function Virtualization |
| ONAP | Open Networking Automation Platform |
| ONF | Open Networking Foundation |
| OPNFV | Open Platform for Network Function Virtualization |
| OTT | Over the Top |
| PSTN | Public Switched Telephone Network |
| QoE | Quality of Experience |
| QoS | Quality of Service |
| RAN | Radio Access Network |
| SDN | Software-Defined Network |
| TCO | Total Cost of Ownership |
| TDM | Time Division Multiplexing |
| V2X | Vehicle to Vehicle or Infrastructure |
| vEPC | virtual Evolved Packet Core |
| VLAN | Virtual Local Area Network |
| VNF | Virtual Network Function |
| VPN | Virtual Private Network |
| VR | Virtual Reality |