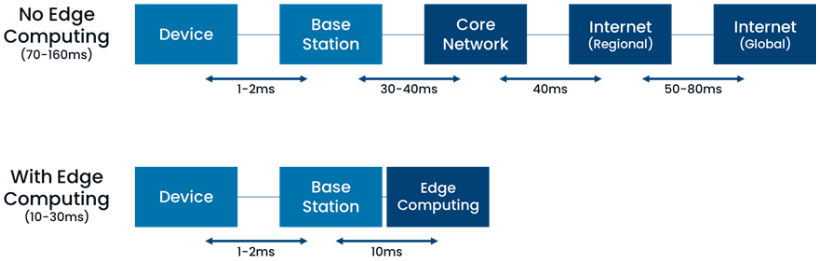
Edge Computing

Owned by David Cherney DISH •••
Last updated: Jul 10, 2023 • Add Workflow

Contents

The latency of applications can be reduced by reducing the distance data travels. More importantly, number of routers encountered is reduced In the same way: placing hardware close to the user.



In particular, hosting applications accessed by UE close to the UE instead of hosting those apps on data networks like the internet, reduces latency by an order of magnitude. This kind of mechanism has become essential to latency sensitive applications like multiplayer gaming.

The Edges

To minimize latency and networking load, the closer to the UE that aspects of applications run the better. The following definitions apply to 5G communications.

Definition: The infrastructure edge consists of compute resources for apps placed at the same location as base stations.

Definition: The device edge consists of compute resources for apps within the UE.

(Note that what is meant by “edge” is the edge of the access network, as shown below. Indeed 3GPP defines edge computing as “computing capabilities at the edge of an access network”. I argue that the phrase “edge of the core” is not appropriate to refer to this concept, and that the phrase is better used to describe aspects of the core dealing with roaming and dealing with security.)

Multi-Access Edge Computing

Computation on the edge used to be called “mobile edge computing”. However, since edge computing is helpful for both wireless and fixed access networks, and since 5G is meant to advance fixed mobile convergence, the term has been replaced with the multi-access edge computing, MEC.

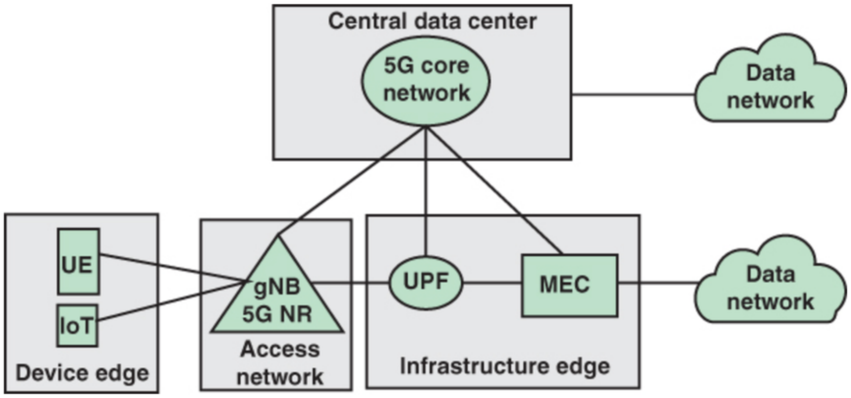


FIGURE 10.3 MEC Integrated into a 5G Network

MEC is a form of enhancement of telecommunications via cloud computing. Telecommunication is communication over distance, and cloud computing is the use of other people's computers. Cloud service providers (like AWS) build data centers (DCs) at many locations and provide fast fiber connections between those data centers. Some DC are large and expensive ($3B for a national data center, NDC) and some are small. There are many times more small data centers than large data centers.

MEC uses this geographically distributed infrastructure to enhance telecommunications service to UE. Carefully articulating how;

- A UE might be using an app that could benefit from infrastructure edge computation
- user plane data for that app will flow through a service data flow (SDF)
- That SDF will be in a PDU session
- That PDU session will have an anchoring UPF
- a network operator (like DISH) will put UPF instances at many data centers (DCs), both central DCs and edge DCs
- the control plane of the core will choose which UPF instance to use as a PDU session anchor
- the control plane can be made aware of the SDFs preference for infrastructure edge computing
- the context aware control plane will choose a UPF instance running in a data center geographically close to the access point (gNB) of the UE.
- this close placement reduces the physical distance and number of routers (hops) between GNB and UPF, and thus reduces latency
- aspects of the app can run in the same data center as the UPF, minimizing the hops between UPF and the app for those aspects of the app.

The service for an app can be broken into three parts:

1. aspects of the app running on the infrastructure edge (edge data centers)
2. aspects running on the device edge (on the device)
3. aspects running at more distant locations (central data centers).

How close is close?

UPF Close to User

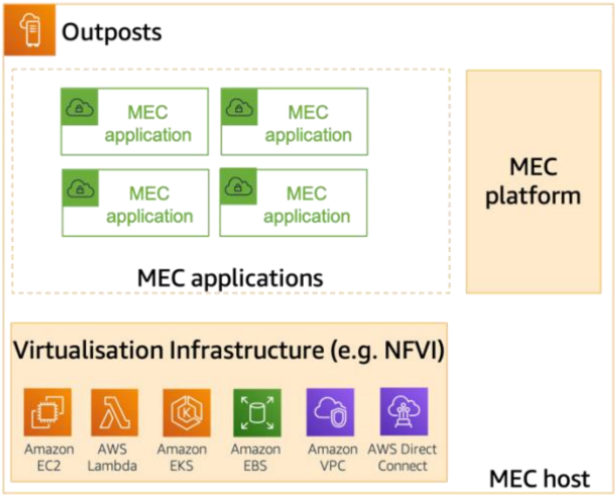
Smaller and smaller data centers have been invented as time has gone on. AWS offers

- local data centers,
  - close to large populations that are far from large data centers
- passthrough edge data centers
  - aggregation centers for local data centers
- breakout edge data centers
  - host CU's from gNBs

- Wavelength Data Regions; AWS places physical compute resources inside a network operator's infrastructure, like a AWS computer in a DISH base station, or more specifically an AWS shelf on a DISH rack, like pictured below. Such a shelf from AWS is called an AWS Outpost.

UPF Close to App

This rack might serve a single gNB. In particular, a single CU for that gNB, which is a virtual network function, might be on this rack. That CU will serve multiple DUs and thus multiple cells. The anchoring UPF might also be on that rack running in EKS. The application might also be on that rack. By having all three, CU, UPF and app running on the same rack, single digit millisecond scale latency is enabled.



Running on the same rack might mean two boxes connected by a cable, but it also can mean running on the same physical computer. That computer will be running virtualization infrastructure software (to create virtual machines or containerized environments like Kubernetes). The UPF and MEC apps can run on the same virtual environment. Close can go so far as to mean that a MEC app is running in the same Kubernetes cluster as the UPF. This effectively puts the app in the core. This is why you see AF (for application function) in diagrams of the 5G core.

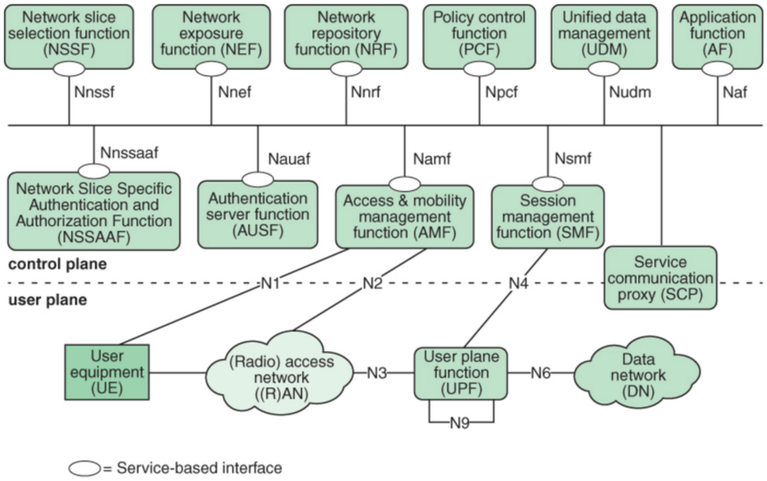

Network slice selection function (NSSF)
Network exposure function (NEF)
Network repository function (NRF)
Policy control function (PCF)
Unified data management (UDM)
Application function (AF)
Nnssf
Nnef
Nnrf
Npcf
Nudm
Naf
Nnssaaf
Nauaf
Namf
Nsmf
Network Slice Specific Authentication and Authorization Function (NSSAAF)
Authentication server function (AUSF)
Access & mobility management function (AMF)
Session management function (SMF)
Service communication proxy (SCP)
control plane
user plane
N1
N2
N4
User equipment (UE)
(Radio) access network ((R)AN)
N3
User plane function (UPF)
N6
Data network (DN)
N9
= Service-based interface

FIGURE 9.4 Non-Roaming 5G System Architecture

Enhancements

ITU does not require any particular architecture for MEC. However, 3GPP does define tools that enhance MEC.

AF Influence on Traffic Routing

Application functions can inform the 5G core (specifically the PCF and SMF) that an application requires edge resources. This allows the core to choose a UPF instance located at MEC host that is both close to the UE and is equipped with the right MEC services.

Edge Application Server Discovery Function (EASDF)

For the AFs to proactively inform the core in this way, they need access to information about the capabilities of the MEC hosts. The AFs do not access this information directly from the MEC hosts, but rather use the 5G core as a proxy.

- MEC hosts register a description of their identity, location, and abilities with the 5G core function called the edge application server discovery function, EASDF (defined in release 17).
- This allows the EASDF to act as a DNS server (giving names to the addresses of the MEC hosts.)
- An app running on a UE can then send UE location information to the EASDF and obtain a list of qualified MEC hosts. This is called edge application server discovery.
- The UE, advised by the informed app, can then request a PDU session to contain the SDF for the application, and this request can include information about preference of location of UPF
- The SMF in charge of establishing that PDU session (by being in charge of choosing a UPF to serve as anchor) is informed of the preferences, and makes its choice.

UE Route Selection Policy

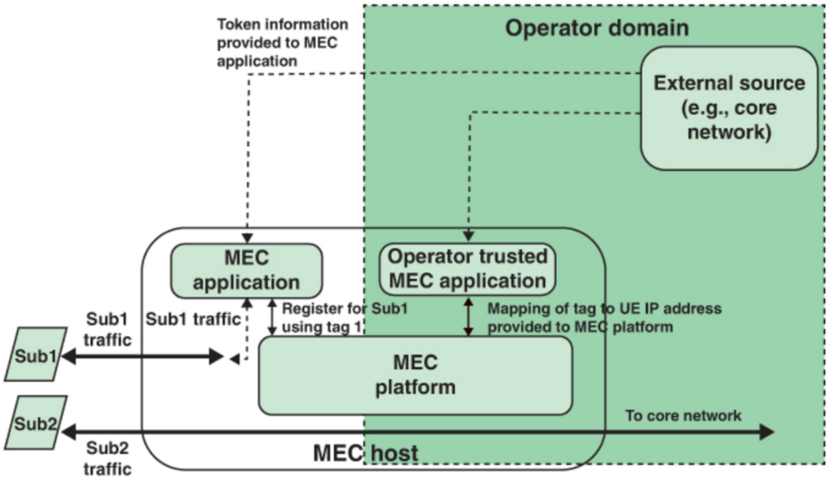
Another mechanism to allow apps to control the selection of UPF is called UE route selection policy (URSP). These policies are stored in the PCF and are delivered to the UE by the AMF on request. They then reside in the UE and are used by the UE for choosing which existing PDU sessions shall host a new SDF for an app. The policies are informed of the edge computation needs of the apps. If no existing PDU session is appropriate, a new PDU session is created.

Operator Trusted MEC Applications

It is one thing for a data flow for an app to be directed to a MEC host based on the app; what about directing a flow to a MEC host based on the user? For example, a targeted advertising app selects a group of users, or more specifically the data flows of those users, and sends them advertisements. To do this, the app must obtain information from the network about users and their data flows. This is done through MEC hosts.

In particular, such MEC applications need a mapping of the addresses of UEs' PDU sessions (and thereby the SDFs within) to users. This information resides in the core’s UDR and UDSF. If a MEC application running on a MEC host has the privilege of being a operated trusted MEC application, then it can get this information from the core through the NEF.

In the diagram below, the MEC host (a computer) contains a router and MEC platform (the collection of MEC functionalities on that computer). The MEC platform informs routing as follows. An operator trusted MEC application has obtained a map of PDU session addresses to subscribers from the core through the NEF. That trusted MEC app applies tags to PDU sessions according to a rule; apply the tag if the subscriber meets a description, otherwise do not apply the tag. These tags are sent back to the core via the NEF. A non-trusted MEC application obtains this tagging information from the core through the NEF. Subscriber 2 does not meet the rule, so sub2’s PDU session does not carry the tag, and is routed to a UPF instance at a large data center for anchoring. Sub1 does meet the rule, so sub2’s PDU session does carry the tag, and so traffic for Sub 1 is routed to a MEC application at this MEC host. This MEC app sends Sub1 a targeted ad through that PDU session.

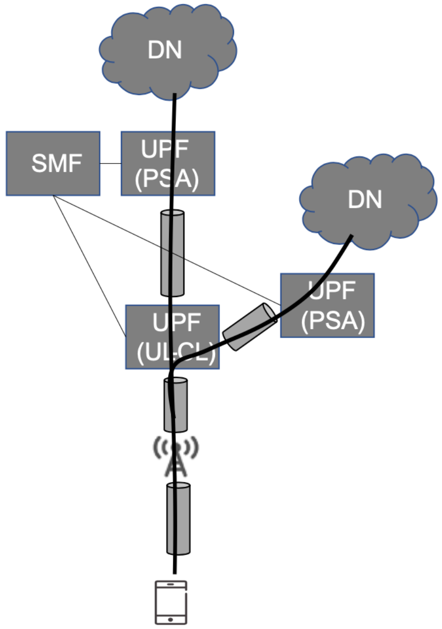

Token information provided to MEC application
Operator domain
External source (e.g., core network)
MEC application
Operator trusted MEC application
Sub1 traffic
Sub1 traffic
Register for Sub1 using tag 1
Mapping of tag to UE IP address provided to MEC platform
Sub1
MEC platform
Sub2
To core network
Sub2 traffic
MEC host

UL-CL/BP

The infrastructure edge data centers that are farthest out on the edge (and closest to UE) are both the most desirable and the smallest. Due to their small size, the compute power available at them can be quite small. This creates the problem of a edge data center with its resources exhausted.

To address this problem, a new functional role for a UPF (besides PSA and I-UPF, both already introduced in this introduction) has been created to allow a single PDU session to have anchoring UPFs (PSAs) at two data centers.

In the uplink direction, a decision must be made about which packets are sent to which PSA, and the new role is called uplink-classifier (UL-CL). In the downlink direction data flows from the two PSAs must be merged and the role is called branching-point (BP). A UPF that plays this role is called a UL-CL/BP.



Presumably, the reason for this design is motivated by the structure of PDU sessions; within a single PDU session there may be two kinds of service data flows (SDFs)

- SFDs that require edge compute,
- SDFs that do not require edge compute.

The UL-CL/BP mechanism allows the SDFs that require edge compute to go to edge DCs, while sending the other SDFs to larger DCs. (I have not seen this explicitly stated anywhere.)

MEC and Mobility

What if a UE using edge compute for a SDF moves far from the edge data center serving it? Take for example someone watching a movie on a phone for 2 hours while in a car, moving about 100 miles in the process. As another example, a drone streaming video to a computer vision model deployed on the infrastructure edge that travels many miles. There needs to be a mechanism to support edge service continuity under mobility. To discuss this possibility, it will help to introduce new terminology.

Definition: A MEC application is an application that is intended to be run on the infrastructure edge.

Definition: A MEC host is an entity that can host a virtual infrastructure to provide compute, storage, and network to MEC applications.

Definition: A MEC platform is a collection of functionality that is required to run MEC applications.

Definition: A MEC service is a service provided by either a MEC platform or a MEC application.

Definition: A MEC system is a collection of MEC hosts, each MEC host being at a different physical site.

Definition: MEC management is management of a MEC system.

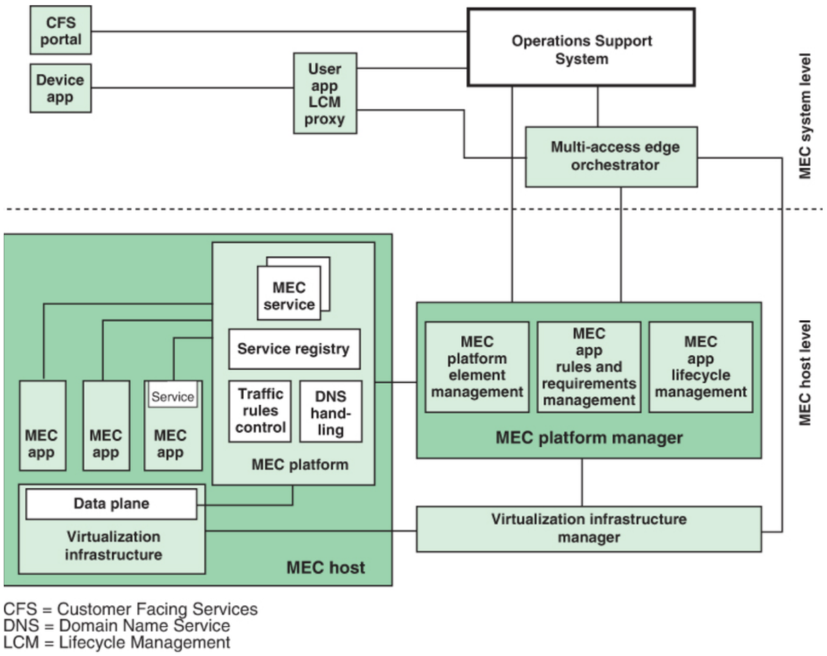

CFS portal
Device app
User app LCM proxy
Operations Support System
Multi-access edge orchestrator
MEC system level
MEC service
Service registry
Traffic rules control
DNS hand-ling
MEC platform
MEC app
MEC app
Service
MEC app
MEC platform element management
MEC app rules and requirements management
MEC app lifecycle management
MEC platform manager
MEC host level
Data plane
Virtualization infrastructure
MEC host
Virtualization infrastructure manager
CFS = Customer Facing Services
DNS = Domain Name Service
LCM = Lifecycle Management

FIGURE 10.5 Multi-Access Edge System Reference Architecture

If a UE is connected to a MEC application running MEC host 1, and the UE moves out of range of MEC host 1 and into the range of MEC host 2, then MEC management uses MEC services to change UPF instances serving as UL-CL/BP and those serving as (edge) PSA so that MEC host 2 becomes the host of the MEC application.

Multiple UPF Roles

Note that the three roles a UPF may play (PSA, I-UPF and UL-CL/BP) are not mutually exclusive; a UPF instance at a particular data center can simultaneously play the role of a PSA for one PDU session and the role of I-UPF for another PSA. It could also play all three roles, and the number of PDU sessions the UPF is involved in can by any number in {0,1,2,…}.

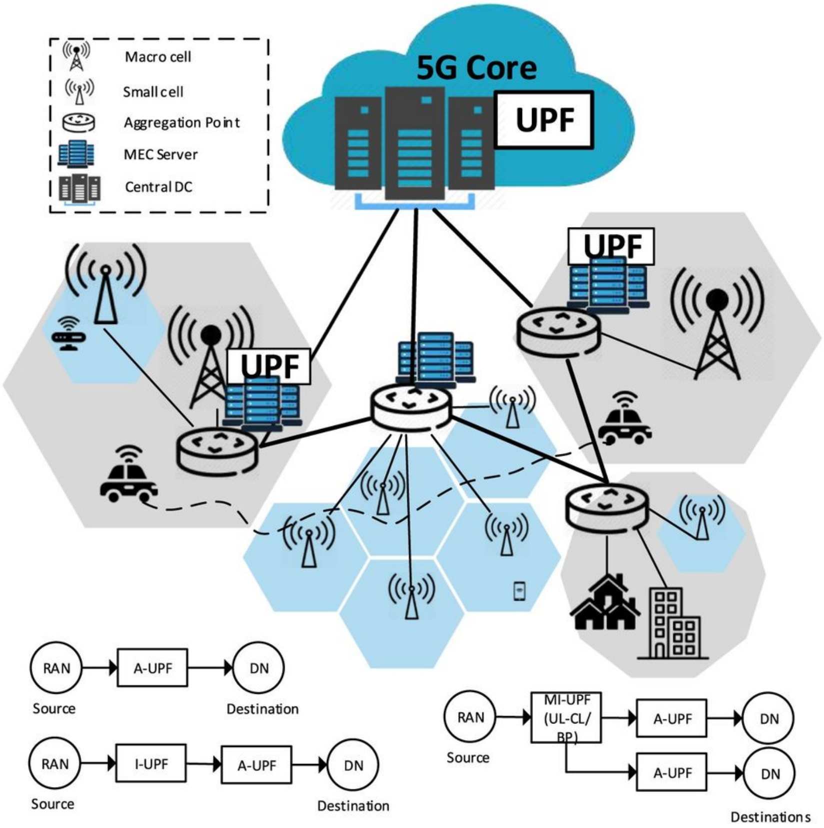
(I suspect that in the diagram above, the best configuration is for the same edge UPF instance to play the role of both UL-CL/BP and PSA, while a UPF at larger data center plays the role of the other PSA.)

UPC Problem

The various roles that a UPF can play in facilitating

1. PDU sessions as a PSA
2. mobility as an I-UPF
3. edge computing as a UL-CL/BP

introduce an extremely difficult problem; where in the available data centers should UPF instances be placed, and what cloud resources should be provided to them. This is called the UPF placement and chaining problem (UPCP). In the diagram below PSAs are labeled A-UPF.


Macro cell
Small cell
Aggregation Point
MEC Server
Central DC
5G Core
UPF
UPF
UPF
RAN
A-UPF
DN
Source
Destination
RAN
I-UPF
A-UPF
DN
Source
Destination
RAN
MI-UPF (UL-CL/ BP)
A-UPF
DN
A-UPF
DN
Source
Destinations