MEC (Multi-Access Edge Computing) Commonly called “Mobile Edge Computing”

MEC is an idea for network architectures that brings what are normally described as cloud computing services to the “edge” of a mobile network – generally at the base station, or at a point of aggregation of several base stations.

The vision of MEC is to allow companies to operate application-specific servers close to mobile devices (both geographically, and from a latency perspective). The intended benefits of MEC are to reduce network congestion and latency, and therefore give users a better experience.

The Technology:

From a basic technical perspective, the idea of MEC is for some network requests coming from a user’s device to be directed to the local MEC resources, which are located on-site at the base station, or are otherwise nearby, as opposed to sending these requests through the full operator core network to an internet gateway site, which might be at the opposite end of the country. This incurs a latency, and also requires significant transit network capacity to funnel data to central locations.

The primary benefit of MEC is generally heralded to be low latency – by having traffic leave a 5G network locally, at the base station, and being served by equipment located at (or near to) the base station, the network latency of routing traffic back through to the mobile operator’s core network is eliminated. Similarly, the latency of then routing traffic from the operator’s core, through the internet, to the network of the application provider (or their data centre provider) can be eliminated.

When considering the widely advertised 1ms latencies of 5G networks, and more broadly the introduction of ultra-low-latency communications, MEC begins to make more sense. Firstly, we should assume that all latencies are measured based on round-trip times (i.e. the time to send a

message back and forth between start and end-point). This is because, absent a return message, we have effectively created an infinite latency channel, since we cannot tell if the message was delivered, or act upon it. This means that the latency for sending a message, without a return path, is relatively meaningless.

Allowing for the speed of light in an optical fibre to be 200 km/ms, and noting that we require bidirectional communications, this means light travels 100 km in a fibre in one millisecond. This assumes that the light simply travels down a single strand of fibre, without switching, boosting, routing, or any other network-based process. Clearly this is not going to be achieved in the real world, as this would require us to lay dedicated fibres between every source and destination, and lose the economies of scale of aggregating together radios and carrying out packet switching in networks.

A typical mobile network's backhaul transit network will have 4 or 5 hops, involving packet-switching of the fibre traffic, and therefore adds latency of perhaps as much as 10 microseconds each time[1]. After allowing for this latency in each direction, and depending on the switches used in the network and level of traffic, it is likely that the transit latency would mean only 50 to 80 km of fibre was realistic. This also does not account for the air-interface latency, the baseband processing of the traffic, and the latency of getting to the data centre or other server hosting facility used by an application provider.

5G networks aim to improve air interface latency using several techniques, but these are not magical, and are still subject to the laws of physics. In a 5G network, prioritisation of traffic can firstly be used to ensure Ultra-Reliable Low-Latency Communications (URLLC) take precedence over Enhanced Mobile Broadband (eMBB) and other lower priority, less urgent, traffic. Optimised hybrid automatic repeat request (HARQ) protocols on TDD bands will allow for a re-transmission to take place within 1ms of the original failed transmission.

The grant-free uplink processes in 5G[2] make it possible for uplink transmissions to take place. In a conventional 4G system, the process for a handset carrying out an uplink transmission is:

1. Handset has data it wishes to transmit, prepares it into a packet to send
2. Handset waits for its scheduled chance to transmit a scheduling request
3. Handset sends scheduling request to base station
4. Base station sends back a scheduling grant via Physical Downlink Control Channel (PDCCH)
5. User Equipment (UE) responds to grant, transmitting data on the Physical Uplink Shared Channel (PUSCH)

[1] https://www.cisco.com/c/en/us/products/collateral/switches/nexus-5020-switch/white_paper_c11-465436.html
[2] https://www.mdpi.com/1424-8220/19/16/3575/pdf

This process clearly involves a lot of waiting for an available uplink time slot, and then waiting for a response granting permission to carry out the actual transmission itself. This is the equivalent of requiring each phone to wait for their turn to put their hand up to say they have something to say, then wait to be called upon, then actually say what they wanted to say in the first place. The goal of low-latency 5G is to avoid phones having to go through this iterative process, and instead simply say what they wanted to say in the first place!

In 5G, where ultra-low-latency services are involved, it will be possible for the handset or UE to be pre-configured with a grant-free uplink configuration in advance, which allows the UE to simply transmit when it has data to send, rather than waiting on resource assignment. This should reduce uplink latency to that of typical downlink latency. Ericsson[3] has modelled example latencies, and based on a 120 kHz TDD configuration with 14 symbols per slot, the latency for grant-free uplink traffic (and downlink traffic) should be around 536 microseconds.

Noting that a downlink and grant-free uplink have the same latency, the minimum realistic round-trip latency in a 5G network is therefore in excess of 1 millisecond, before the data has even left the base station.

In order to approach a 1ms end-to-end latency, it is clear that we need to re-think where we measure latency from and to. MEC offers a way to do this – by having servers located at the base station, or alternatively at a nearby aggregation site, where 5G Data Units (DUs) are sited, latency from handset to service can be reduced, compared with traffic running over the internet[4].

It is important to note, however, that in an FDD network (where uplink and downlink frequencies are not shared through timeslots), lower latencies can be achieved – grant-free latencies of as low as 89 microseconds can be achieved in 120 kHz FDD configurations (i.e. in mini slots). It is worth noting however that the vast majority of European 5G networks to date have deployed in TDD spectrum (i.e. 3.4 to 3.6 GHz). There is less available FDD spectrum for use, meaning that FDD-based services have lower available capacity and throughput. This means there will be an inherent trade-off between capacity and low latency in 5G networks.

By making computer resources available close to the user in the network, the need for traffic to be transferred through the network core is avoided. To implement this, user plane traffic needs to be split – some of it should be directed towards local (MEC) servers, while regular internet traffic will be routed in the conventional way. Note that where lower latency internet connectivity is desired, internet traffic may egress the network without traversing back to the core network – this is often referred to as CUPS (Control User Plane Separation), but really is simply what can be achieved when the 5G UPF (User Plane Function) is located near the base station in the network.

[3] https://cscn2017.ieee-cscn.org/files/2017/08/Janne_Peisa_Ericsson_CSCN2017.pdf
[4] https://medium.com/5g-nr/cloud-ran-and-ecpri-fronthaul-in-5g-networks-a1f63d13df67

This only makes sense in some circumstances, like where there is good upstream/backhaul internet connectivity available, but the site is a long distance from the operator's core network.

This is not new – the 5G RuralFirst DCMS T&T project implemented and demonstrated this in their Orkney network, experimenting with local traffic break-out and other innovative ways to improve the availability of internet services to trialists, rather than being constrained by a traditional MNO architecture of routing all traffic to a single central core location.

MEC takes advantage of the capabilities of CUPS and UPF components in order to route the traffic which can be served by MEC servers directly to them, while still forwarding regular traffic to the internet. By carrying out the user plane function (UPF) near the edge of the network, this routing is efficient and lower latency, compared with sending it back to the core.

MEC can be used to deliver static web content (i.e. a local copy of a content delivery network), as well as to enable local processing of data and interactive services (such as locally-resident Microsoft Office 365 functions for a more seamless experience, or other application-specific functionality).1

Traffic steering and policy-based routing can be used to route certain traffic to the local MEC servers via the local UPF, while sending unrelated traffic to the regular core network for egress as usual. The SMF (Session Management Function) and PCRF (Policy and Charging Rules Function) of the 5G core carries out traffic steering between base stations and UPFs, by passing policies to the UPF. These can allow for the "steering" of traffic destined for certain IP address ranges to another UPF, and then to a local MEC network.

A MEC edge facility can be located at a base station, or indeed any other part of an operator network, including at base station aggregation sites.

The Questions to Address

Anti-trust and anti-competitive measures:

Latency is inherently network-originated, and an MNO can therefore prioritise/de-prioritise traffic at will (as a commercial differentiator, notwithstanding net neutrality rules). An MNO could de-prioritise less profitable internet/cloud service traffic to drive sales of low-latency MEC resources/services to internet/cloud service customers and providers. How do we regulate this behaviour, who regulates it, and who decides where the line lies between reasonable network traffic management practices, and unreasonable anti-competitive behaviour?

Specifically, there could be potential for antitrust issues, were someone to argue that MNOs were abusing a dominant position as a mobile network operator to gain a lucrative foothold in the separate cloud computing market (specifically taking advantage the captive nature of the

edge processing market)? Does a regulator have the capability and access to suitable information to even find this line?

To make this clearer – if you want to access the "edge", you need to deal with the MNO, or their chosen designate. There is no realistic prospect of competitive edge service provision, since MNO RANs are private networks which do not peer with other networks at the edge. This means if you want to provide edge services, your only option is the MNO or their chosen commercial partners (who themselves deal with the MNO directly) who can give you access to edge computing resources or connectivity. Even if there were a pretence of a competitive market, with multiple cloud providers offering edge access, all would be inherently paying the price the MNO asked for, in order to be able to have a presence at the edge.

In addition, if you want to access ¼ of the UK market, you need to deal with 1 MNO. To be able to serve everyone, you need to deal with all 4 MNOs. This means that there is no competitive provision, or indeed market competition at all in the edge market – it is entirely vertically integrated by MNO, and you need to partner with each MNO (at their price, or not at all) to reach their subscribers.

Perhaps, were MEC to proceed like this, portions of MNO businesses would need to become regulated utility services in light of this, to prevent legal challenge by the CMA.

Impact on open innovation and competition:

MNOs have gained a dominant market position in telecoms, and the barrier to new entrants is high, as a result of spectrum pricing and availability. There is a risk of "MEC-capture", whereby an innovative company could only gain access to provide MEC-based services through a commercial agreement with a rival, who will be of infinitely greater scale. This is particularly relevant when considering the wider "enterprise communications services" business models that MNOs are increasingly adopting. This could hold back the market, in the same way a new ISP being unable to gain equitable wholesale access to public-funded "wires and ducts" infrastructure holds back the wired telecoms market. This issue is highlighted by the legal principle of "essential facilities".

Limited Enablement of the oft-heralded "MEC" use-cases:

There is a famous and often-quoted point by John Carmack, an early pioneer of VR technology with Oculus, that "I can send an IP packet to Europe faster than I can send a pixel to the screen. How f'd up is that?"[5].

The same holds true with 5G – we are constrained by end devices, operating system latency, and screen latency and technology. Humans generally do not perceive latency below 20ms either, even for VR. For the AR/MR/VR use-cases, which MEC is often heralded to enable, it is important

[5] https://twitter.com/id_aa_carmack/status/193480622533120001

to also consider how these technologies work. For immersive VR to work, you need at least 60 fps of video (more ideally 90 fps or above), with minimal latency (below 20ms).

When the user moves their head, the VR content displayed needs to change to reflect this movement within that 20ms "latency budget", to avoid the experience being jarring. This normally means that the full 3D video would need to be stored either on the device, or within a 10ms latency of the device itself. This is where MEC could help. Modern VR tends to stream at least 4K video for each eye, although there is significant correlation between both videos, and this should be compressible to a single 4K video.

Clearly MEC servers, with suitably low latency, could enable the streaming of VR video to devices. The question is whether users really want to consume VR content when "out and about", with an immersive headset isolating them from the world around them. This means that AR/MR are more likely to be used outside of the confines of high-bandwidth home networks, but these have lower throughput requirements since they are not streaming a full 4K feed per eye.

The RF throughput is the limiting factor, not the backhaul:

There are 2 "pipes" of relevance to each base station – one pipe is the radio channel from the base station to phones and other devices. The other pipe is the backhaul channel, connecting the base station to the rest of the mobile network. The size (capacity) of the pipe determines the amount of information (bandwidth) that can flow through it. Fibre backhaul has a high capacity (large pipe), generally much higher than the radio channel between phones and the base station (narrower pipe) itself.

Because of the laws of physics, there is a limited amount of spectrum available to use for each operator's base stations, so the choice of pipe size is limited for connections from phones to base station. The pipe must be shared by all users connected to the base station. This limits total throughput per user. Using MEC reduces the amount of data needing to flow through the backhaul "pipe". MEC doesn't change the amount of data needing to flow across the narrower radio "pipe" from phone to base station. This means that MEC has not increased the capacity of the individual cell site at all.

It is relatively cheap to add more backhaul capacity if a fibre is already installed – technology like DWDM and faster interfaces can increase the backhaul to more than 400 Gbit/s – more than enough for the vastly foreseeable future. However, it is not cheap to increase the size of the pipe between phone and base station – that remains the limiting factor, regardless of MEC. MEC can reduce the amount of time it takes data to be processed, by bringing processing to nearer the base station, but importantly, the same amount of finite pipe capacity "over the air" is still needed for MEC.

We must also be careful not to "over-use" low latency as an opportunity to sell better connections – to achieve low latency connectivity (the main advantage of MEC), each end device

needs to have finite radio resources (i.e. pipe capacity) set aside for it, whether it is used or not (to enable grant-free uplinks). Note that downlink is already low-latency in 4G, and therefore is not significantly different in 5G. This ultimately reduces the ultimate available capacity (throughput) of the finite radio pipe, if finite radio "pipe" capacity is being set aside for users, just in case they need it.

MEC can make sense for content caching (like a CDN, for example) or local computing in some circumstances, where there is very limited backhaul capacity available, no real prospect for improving it, and a clear use-case that can be facilitated through cached or locally computed content. For example, a train or plane could have a small MEC cache hosting selected Video-on-Demand service content (i.e. iPlayer, Netflix, etc), like we see today with Wi-Fi-based movie streaming on trains/planes. Similarly, Microsoft could host a local cached copy of their Office 365 productivity suite to enable better performance on the limited bandwidth train or plane connection. There are still challenges with either of these approaches – the MEC cache still needs to be updated somehow, and in the case of Office 365, if users were storing content on the cache, there would be significant complexities around synchronisation of documents when a MEC cache was offline for some time. Therefore, it is likely that such approaches are better suited to "static" content CDNs (like the files required to let a user access cloud-based features of the Microsoft Office suite), but not for user content being uploaded (i.e. their documents stored on OneDrive).

The average user will not observe any performance increase from MEC in normal web browsing or other similar normal content consumption use-cases. Connection latency is rarely, if ever, a constraining factor there – per Robert Miller's "*Response time in man-computer conversational transactions*", a latency of below 100ms in a mainframe system is generally perceived to "instantaneous" by most users in most scenarios.

Who is pushing MEC? Cost versus benefits:

It is unclear at this point exactly who is demanding MEC, and what use-cases are enabled by it, which cannot happen without it. One primary driver of MEC is MNOs looking to create additional revenues by extending their downstream reach from beyond the network itself, to the content which the network is conveying (i.e. reaching into the application layer). In other words, monetising the data transfer itself across their network, rather than simply being a "pipe" that transfers data. This enables MNOs to turn over-the-top and internet-based service providers into direct paying customers, and "own" the whole value chain to the consumer. This is in line with the kinds of gatekeeping-based business models being outlawed or heavily regulated by the EU's forthcoming Digital Services Act, and subject to antitrust and competition investigations on both sides of the Atlantic.

MNOs appear to want to lease access to "cloud" style services (i.e. partnership with AWS as seen with VZW[6]) in order to become gatekeepers to edge-cloud services. The benefit of MEC however is unclear other than a few very specific and isolated use-cases, because an MNO edge compute node will invariably (due to market economics) have a price premium attached due to being closer to the user, compared with a traditional cloud service.

Cloud is already commoditised and a highly competitive market (due to open peering between network providers, and no "lock-in" customer capture requiring you to use any particular cloud provider). This means that MEC can only really be viable for niche use-cases requiring low latency. A careful balance must be struck, to prevent the potential risks of significant market power in the telecoms sector acting to distort the (functioning and highly competitive) cloud computing market, by virtue of the concentration of mobile network operation in a small number of providers, each with exclusivity over their own networks.

Understanding the differences between MEC and Cloud:

The fundamental way the internet works today is that network traffic is carried and exchanged at no charge between large network operators in internet exchanges (such as LINX in London). These deals, where money does not change hands according to traffic passed, are commonly referred to as "peering agreements". Peering agreements reduce the cost of connectivity to an internet service provider, as they have a direct inter-connection with those they need to reach (i.e. users or businesses), and therefore do not need to lease capacity from commercial transit network providers.

These networks then provide connectivity to large enterprise/operator clients, who buy "access ports" into a very large pipe, effectively. Innovation and growth on the internet are facilitated as a direct consequence of this process – all traffic is treated equally, and operators do not discriminate between traffic when transferring it. This avoids the creation of an information underclass, and helps to democratise the opportunities afforded by the internet to be available to all.

The cloud computing model (AWS, Azure, etc.) has driven attempts to normalise the metering and charging for network capacity (bandwidth) as a significant profit source. Cloud computing models, which MEC looks to replicate, charge as much as $0.09 per GB of data transferred to the internet. This is a price which bears very little relation to the cost of the bandwidth, but can become a significant barrier to adoption and innovation if it were replicated at the "edge" of networks, as it would allow operators to take over the full value chain end-to-end, from user and handset, through to the server infrastructure and content delivery. By way of comparison, a leased server at £40 per month often comes with unlimited bandwidth on a Gigabit-speed connection. This shows the significant profit margins in cloud provider bandwidth prices. It is likely that MEC operators will seek to increase bandwidth prices even further, despite being

[6] https://www.verizon.com/about/news/verizon-aws-mobile-edge-computing

nearer to the user, and thus incurring lower costs – bearing in mind that MEC reduces the bandwidth requirements of the network provider.

One key commercial difference between regular cloud provision and MNO MEC is that an MNO is already charging their users for a data allowance – a user buys 5, 10, 20, 100 GB of data (or indeed "unlimited", subject to fair use policies), and is then able to use this how they please. They can access any website or service, and have paid their provider a fair price for the bandwidth used. Cloud providers and other internet-based companies are able to deliver content to these users are low or zero cost (where they peer with the MNO's network, or one of their peers), since the traffic is able to pass between networks at internet exchanges. Importantly however, there is no direct payment required by an MNO in order to reach their customers. This means MNOs do not act as "gatekeepers" to users.

Under a MEC scenario, there are a number of questions for those wishing to provide services using MEC – would they need to pay the MNO for the backhaul traffic used to connect their MEC node to the internet to update its content and control it? Will they be required to pay for metered data transferred from their MEC node to users (at the edge), or will this be "free" of charge and included in their MEC node access? Will MNOs offer differential "sender-pays" pricing, whereby certain MEC partner services are "zero rated" to consumers, and therefore more commercially attractive to users, since they do not consume their monthly data allowance (which detriments commercial rivals to these services which can afford to rent MEC capacity with each MNO)?

This is particularly relevant, in light of recent EU jurisprudence on the legality or otherwise of selective traffic zero-rating, and the impacts of this on net neutrality[7].

There is already precedent to suggest that some operators may be tempted to "double-dip" and charge both providers and users for the same data – femtocell users (and Wi-Fi calling users) provide their own backhaul connectivity to reach the operator, yet are billed at the same rate for services used, even when not consuming any mobile spectrum capacity on the base station. Such practices may themselves separately be worthwhile topics of future competition research and analysis, since all 4 UK operators operate the same commercial model for these bring your own backhaul (BYOB) services.

Security Challenges into the future of MEC:

There will clearly be new security challenges on the edge – it is notoriously hard to secure remote devices. Attestation-based trusted platforms would seem to be nearly essential in achieving this. This is an area of the IT world that is rapidly evolving, and not one that MNOs are generally that familiar with, although this is likely to change over time. There is a risk, in the meantime, that MEC could become the so-called soft underbelly of the network.

[7] https://www.lw.com/thoughtLeadership/Zero-Rating-and-EU-Net-Neutrality-Rules

If traffic is being dynamically routed at the edge in order to enable MEC, there is potential for unintended traffic to be mis-directed to the MEC node, if a rogue user/attacker is able to gain access to the traffic steering component. This is a general threat for any SDN or similar "software-based" networking system.

There is also a requirement for access rules be suitably defined and enforced to enable security monitoring of which IP ranges/services are being presented locally from the MEC system, and the remediation of any unexpected IP ranges being offered.

Also, in light of the significant low-latency advantages being heralded for MEC, there are open questions as to how MNOs and providers will undertake security monitoring (which may require malicious packets/traffic to be blocked from reaching MEC nodes), without adding the latency that MEC seeks to reduce. Techniques such as cut-through switching, commonly used to reduce latency in network equipment, are inherently incompatible with stateful packet inspection firewalls, which look at the full packet (rather than just the source and destination headers) before allowing it to pass. In particular, application layer firewalls and web application firewalls (which are increasingly essential to protect web and HTTP-based services from attack) will add considerable extra latency, and still be required in a MEC scenario to protect the MEC assets.

From a security perspective, there are also challenges to be addressed around use of TLS and PKI certificates, and how to ensure the security of these if they are deployed "at the edge", especially where such certificates could be used to intercept any traffic going to a service (rather than just traffic to a given edge node) – it seems prudent for security standards to evolve in this space, to ensure that providers do not deploy wildcard certificates to edge devices, and rather constrain these to dedicated hostnames for specific MEC server instances. How this would be used to implement (for example) web services would be a user experience and usability problem – clearly the "*.google.com" wildcard certificate should not be stored at the edge on MEC servers, yet users do not want to have to visit "google-com.mec-lac00293.mnc004.mcc235.3gppnetwork.net" to access Google!

Connected and Autonomous Vehicles, and the need for Inter-Provider Edge Interoperability:

There are some potential applications of MEC in CAV (Connected and Autonomous Vehicle) scenarios, where vehicles may benefit from being able to communicate in a low-latency manner to warn other vehicles about hazards, collisions, and braking etc.

However, for these kinds of situations, peer-to-peer style architectures of communication, i.e. V2V, V2X, V2R (road), V2I (infrastructure) will deliver lower latencies, and work on the vast majority of miles of UK roads where MNOs will not have deployed high-capacity, MEC-enabled 5G coverage (as evidenced by the 4G roll-outs). Given the UK's notorious 4G coverage of roads

outside of major conurbations, it is difficult to see a future where ubiquitous 5G coverage is able to enable these kinds of use-cases.

Where MEC infrastructure and architectures are used to deliver services like this, it is unclear how vehicles connected to different MNOs would interoperate at the edge, and until this is addressed, it seems that there is a very limited use for MEC to enable CAV applications that work between all vehicles, unless MNOs were to engage in much more widespread RAN and active infrastructure sharing (i.e. neutral hosting) nationwide. Otherwise, an "EE"-powered car would not be able to warn a car with an "O2" SIM of rapid braking ahead.

In addition, any convergence of independent RANs, or other increased cross-dependency of operators would impact on the UK's general CNI resilience, and would be a priority to NCSC. In particular, inter-connection at the edge of networks would be a particular security concern, due to the lower levels of monitoring and visibility of traffic passing further from the network core. It may not be feasible at all to enable this kind of inter-connection at the network edge, simply from a security and resilience perspective.

Shared Rural Networks (SRNs) and the Digital Divide:

SRN presents a residual risk when discussing MEC and 5G, since SRN appears to only apply to 4G connectivity, but many people believe that it will cover "all future generations, including 5G". SRN is therefore highly unlikely to deliver MEC to begin with.

Looking to the future, the question then becomes "how could MEC possibly operate under an SRN-style scenario of this type?" Since the SRN is designed around each operator still deploying their own conventional infrastructure, even if it were to be 5G capable (which it is not), this means that each operator would need to deploy MEC equipment at each site. If SRN is needed to reach an area, it is hardly feasible to justify investment in MEC.

If operators were to do this, and MEC were to be successful, the next "digital divide" will be over who lives in an area with MEC service, and who the new "have-nots" will be, noting the even higher cost of equipment for MEC (on top of 5G SA) means that there will be more "have-nots".

Further Information:

Dave Happy & Dr. Greig Paul
Md@Telint.biz & greig.paul@strath.ac.uk

Glossary of Terms used

- Km/ms – kilometers per millisecond
- URLLC – Ultra-reliable low latency communications
- eMBB – Enhanced mobile broadband
- HARQ – Hybrid automatic repeat request
- PDCCH – Physical downlink control channel
- PUSH – Physical uplink shared channel
- CUPS – Control and user plane separation
- SMF – Session management Function
- CMA – Competition and markets authority
- AWS – Amazon Web Services
- VZW – Verizon (a US Operator)
- SRN – Shared Rural Network

5G RuralDorset

www.5gruraldorset.org
@5gruraldorset
LinkedIn
Rural 5G Group

5GEM
5G Enabled Manufacturing
https://www.5g-em.org