CORPORATE PARTICIPANTS

Lilian Veras
Moderator

Thomas Vits
MobiledgeX – Technical Partnerships

Francesc Guim
Intel CTO Office NEX – Principal Architect and Edge Chief

PRESENTATION

Lilian Veras

Welcome everyone to the Intel Network Builders webinar program. Thank you for taking the time to join us today for a presentation titled: “Building On-Demand Edge Clouds with Mixed Infrastructure”.

Before we get started, I want to point out some of the features of the BrightTALK tool that may improve your experience. There's a Questions tab below your viewer. I encourage our live audience to please ask questions at any time. Our presenters will hold answering them until the end of the presentation. Below your viewing screen you will also find an Attachments tab with additional documentation and reference materials, including a number of websites and documents mentioned in this presentation. Finally, at the end of the presentation, please take the time to provide feedback using the Rating tab. We value your thoughts and we will use the information to improve our future webinars.

Intel Network Builders Webinar Series takes place live twice a month, so check the channel to see what's upcoming and access our growing library of recorded content. In addition to the resources you see here from our partners, we also offer a comprehensive NFV and SDN training program through Intel Network Builders University. You can find the link to this program in the Attachments tab, as well as a link to the Intel Network Builders newsletter.

Intel Network Builders partners have been working to accelerate network innovation by optimizing their solutions on Intel technologies. These industry leaders are recognized in our Winners’ Circle program and MobiledgeX is a Gold partner. Learn more about our INB Winners’ Circle program by clicking on the link in the Attachments tab.

Today we're pleased to welcome Thomas Vits from MobiledgeX and Francesc Guim from Intel. Thomas Vits is a product management consultant from MobiledgeX. He has over 15 years of mobile industry product development and product management experience, with the last five years focused on helping telcos and application developers realize the value of network edge computing. Currently at MobiledgeX, his focus is working with telcos, their ISV ecosystem partners, and relevant standard bodies to build standards-based federated edge infrastructures, and to combine edge computing and network capabilities to enable differentiated user experiences.

Cesc has a PhD in computer architecture, and is currently Intel principal architect inside NEX organization. He's driving creation and definition of current and next edge end-to-end system architectures.

Welcome Thomas and Cesc, and thank you, again, for joining us today. I will hand over to Cesc to start off. Thank you.

Francesc Guim

Cool, thanks a lot for the intros. So let's get started with the webinar.

Okay, so we know that patches is no no longer a single location. When I look at kind of-- when we started working on edge architectures maybe six, seven years ago, most of the POCs, or trials, or deployments that we were doing, they were looking at edge a single box or single cluster that is placed in a specific location, like a central office or a static [00:03:35] cabinet, and where all the compute is performed at that location. And the reality is that obviously, over time, and as the technologies have been maturing and evolving, both

from software and hardware perspective, that has allowed the ecosystem in general to think a little bit more broadly, and now looking at edge at something that goes from device to the cloud. And potentially, if you look at many of the deployment models that happen today, you will find some of them that are still in a particular location, but even within that location, you will have multiple tiers.

And the other thing is that each of these tiers that you have where you can potentially deploy edge workloads, they have different types of requirements and different types of characteristics, and one of the important works that we have to do as architects and as solution providers from an edge perspective is to understand which are the different deployment models that you may have from a particular edge provider or edge customer, and obviously, as you go more close to the device, here on the left, you have lower latency, but you have more restrictions in terms of how much compute you can put in there, what are thermal requirements, how you manage those systems. And now as you're moving more towards the public cloud, you have more commodity hardware that probably is going to be providing you the lowest TCO in principle, and that's going to be capable to aggregate more compute. And in some cases, it's not black and white, you may have some intermediate point where you have different types of tiers that will have some intermediate type of requirements.

Now, one of the important things is to understand where, what, and why and basically, the work that we've been doing with MobiledgeX folks, and Thomas here on the call, is to try to understand what are the different use cases that you may have from edge perspective, and where you are placing them? And you are placing them because either there's latency requirements, you have to be closer to the user, you have security requirements or data sovereignty requirements, so that has to be within a particular edge deployment domain, or because you need certain reliability, and that means that you have to have this thing, or the use case deployed into a particular location before reaching the public cloud.

And the second aspect that is-- and that's from an orchestration perspective, it's important because the ISVs developing solutions, they know where they want to-- their requirements in terms of latency, security, and stuff like that, but in many cases, they don't need to be aware on what hardware under each of these locations is, and more importantly, what is the right location for deploying that use case that has certain requirements, and that's the first part of where the orchestration comes into play. So I have a use case, I want to deploy it, I know the latency requirements, I need an orchestration capability, or an orchestration architecture that needs taking care on where my content is going to land.

And the second one, which is important as well, is now you go and you have edge as a... it's a live entity, and what I mean by that is that the loads, the network infrastructure changes over time. You will have more content delivery happening at nights, so that means that there is more pressure into the backhaul of the network, you will have probably more video analytics during the day, and you will have less backhaul network problems, and maybe you can upload some of these video analytics more on the cloud. And that means that also your orchestration architecture has to be aware on the dynamicity of the edge as an infrastructure that has multiple tiers where you can deploy, and that's managed by the orchestration.

The second very important part is that you don't only need to understand where things run, and why, based on the deployment requirements that come from the end user or edge customer. You also need to understand the different types of use cases that you are deploying in this edge infrastructure, and what are the requirements that those use cases have. And part of the work that we do is to-- here you can see stuff that goes across manufacturing network functions, smart cities, speech recognition, and other type of verticals, and you need to understand how these use cases behave, the type of requirements that they will have in terms of compute, acceleration in terms of storage, in terms of network, and so on, so forth. So that when you are deploying-- let's say, again, that an end user wants to deploy, let's say, a content delivery or augmented reality use case, and that use case is going to tell you, hey, for this container, I need 10-millisecond roundtrip latency, and then in many cases, it's not going to be capable to tell you I need this type of memory, this type of compute, this type of storage, and so on, so forth.

From orchestration and system architecture perspective, you need to understand that very likely this content delivery workload will require good I/O connectivity to the storage, will require good network capabilities to manage the quality of service, but also we are required to have good memory capacity to have a good caching engine. So when someone is applying a use case that is mapped into a content delivery type of use case, for example, or in the case of AR/VR that requires a GPU for example for the new acceleration, now

the orchestration layer has to understand how to contemplate the use case, tap into the edge location that has the right hardware to basically execute and realize that use case. And within this context is where the multi-verticality aspect is very relevant, because you need to understand the different use cases and verticals, and how you are orchestrating them.

Now, when you combine the two aspects that I just mentioned, so the device to the cloud multi-tiered architecture and now the use cases, what we're trying to do from Intel's perspective, and we've been engaging with MobiledgeX on trying to understand how to realize that with their software stack, is how you can have an architecture that basically allows you to converge different types of accesses, because in some cases, you may have 5G or 4G connectivity to access to that content delivery, but maybe some other users are accessing that content delivery through regular type B ethernet type of connectivity, or through MPLS, or things like that. So you need to have this convergence on the transport perspective, and the same thing from the different verticals. So you need to have an end-to-end system architecture that is capable to comprehend different tiers of deployments that have different requirements and different restrictions, needs to understand the different users' use cases that you may have, and the requirements that it has in terms of compute, but deployment of the use cases as well, and you need to understand how the bytes are going to be moving from the device that is sending information to be processed by that service, or gathering data from that service, like from the content delivery perspective.

So the other important aspect is not only in the architecture as well, and that's an area that's very important and where MobiledgeX is working with different MNOs worldwide, it becomes a key aspect on a scale, is now that you have edge with multi-tier, multi-vertical, but the other thing is that is multi-providers, because at the end of the day, if you're accessing the content delivery, for example, and you're in Spain, your content provider, maybe it's the same, but the edge provider is going to be different, and if that content provider wants to deploy the same container in another country, how those APIs that we were talking about for deploying the services, and understanding the deployment requirements and how they are orchestrated, how those APIs look like. They have to be different, or they should be similar or consistent.

So one of the important aspects on this area is actually how we can standardize a little bit more how the services - [00:13:39] that if I'm a developer from edge perspective, but I don't need to implement 10 different APIs to enable my application for edge. So, I have one consistent API that I can use to develop my use case, and if I go into any of these edge providers, I don't need to either quote again my application or have come complex implementations that have switches and stuff like that, depending on what the edge provider is.

So with this context, I'm going to hand it over to you, Thomas, to take it from here and tell us the journey that we've been doing together.

Thomas Vits

Great, thanks a lot, Cesc, for this introduction. I highly appreciate it.

So as you were already describing the opportunity, let me recap that specifically looking at telco operators. So there is this ever-growing demand of applications and services which do require local edge cloud infrastructures, and on the other hand, the network operators do have the relevant assets, the points of presence and the networks themselves, and to some extent the customer access to help realizing that, and what we do as MobiledgeX is to provide exactly that orchestration platform that allows to distribute those different types of workloads to the respective infrastructures, and allows in a very easy way for the developers to utilize and orchestrate these workload deployments. And I'm going to talk a little bit about how we designed our solution and also specifically talk about the work we're doing in that regards together with Intel.

If we look at the use cases, and Cesc has been already providing a great overview in terms of verticals which can benefit from edge computing, we look here at four use case categories more from a characteristics perspective, and we essentially see four different types of what we call edge-native service application categories, starting with the network functions themselves. So elements of the 4G, 5G, or in the future 6G networks that are deployed on the telco edge cloud themselves.

The second category is what we call flow applications. So this refers to applications like video analytics, or machine learning AI type of applications, where basically you have a high-bandwidth data stream originating from, for instance, a camera or other types of sensors

that need to be processed ideally close to the device in order to reduce the need for network transport, and also to have a lower latency response for certain use cases.

A third type of use case we see is what we call immersive and pervasive types of applications and services. So here, you can think of any type of the so called metaverse services, highly personalized and interactive local content, which needs to be delivered to a wide range of different devices through different access networks.

And a fourth type of edge-native services is autonomous vehicles, so connected cars or drones and other types of vehicles which require a real-time exchange, either amongst themselves or with some control units aiding or helping these vehicles, for instance, for safety or other means.

All these use cases can, of course, be combined to some extent, but they all have their specific characteristics and that's why we look at them in this way.

So, what do we do as MobiledgeX? So, we basically provide four core values. The first and foremost is, as mentioned in the beginning, the easy deployment and management for the developer of those edge-native applications or edge-enhanced services. So, basically, the interface we design towards the developer is leveraging as much as possible known paradigms. So, from a development perspective, it becomes super easy to basically, additionally, use the edge for those types of your applications or application microservices, which do require the edge.

The second aspect we cater for is what we call the Hybrid Cloud Orchestration. So, because-- I mean, we work with telco operators across the globe, and all of them do have different preferences or also due to other considerations, different types of stacks they build the edge infrastructure on. So, what we do is we extract on all these different types of stacks and harmonize it to expose it then, again, to the developer in a unified form. And by that, allow applications to distribute their workloads easily across all those different types of stacks.

The third aspect is, as I mentioned before, the network functions as one of the key use case areas to cater for the specific needs for deployment of network functions. For instance, one of the most relevant examples or straightforward examples is the deployment of a UPF directly via the platform itself and, therefore, being able then to break out the traffic at that very location.

The fourth aspect is-- and that is becoming very relevant these days-- is the network exposure. So, that means besides the pure orchestration of edge workloads towards the different infrastructures, we integrate also with the exposed network capabilities, for instance, prioritization for QoS or other latency management features on the network paths, and more specifically on the access network paths. So, that the combination of the edge orchestration, the placement and, therefore, reduction of the data paths, then together with these control points towards the network truly enables low latency and low jitter type of applications or services.

And this is also part of what we are doing together with Intel. So, as Cesc was already explaining, the end-to-end control and orchestration capabilities over the multi-cloud edge infrastructure is key to also unlock the end-to-end latency advantages. And therefore, we are working with Intel to understand A. on the use case side, the specific requirements and characteristics towards then the specific infrastructure, as well as also establishing the means to interact and integrate with the specific hardware, to utilize it in the best possible way. So, that the utilization of the underlying hardware is optimized, then also the monitoring capabilities are utilized, so that the end-to-end orchestration can take all this information into account and create the optimum in terms of overall system, depending on the specific requirements of the edge workloads, which can be different and have multiple objectives.

And so, in a nutshell, we are working with Intel to integrate with the latest reference architectures and software building blocks, as well as the chipset accelerators, as mentioned, to optimize the utilization of those.

This shows the overall architecture, which we built to enable this orchestration and management of application workloads. So, the key aspect here is that the control plane, which we are building, has visibility and control points from the device, so that can be a smartphone, but that can be also a car or a drone or whatever it is, by means of our SDKs and APIs, which are specifically designed for the client-side. Over the network, by integrating with the exposed network capabilities so, for instance, in a 5G set up, talking to the

NEF, the network exposure function, to the actual infrastructure stack and eventually to the hardware itself. So, what is summarized here as Edge Node Awareness. So, knowing, not only the location and general characteristics of the hardware, but also the current status in terms of load, in terms of other aspects, which may be relevant for the dynamic orchestration of edge workloads.

On the application side, northbound, we provide an interface for the application developer to declare his requirements. And that is, on the one hand, the compute requirements he sees for his workload, but also other requirements like latency, obviously, but that can be eventually also extended with aspects related to sustainability, for instance. So, orchestrating towards energy efficiency or low emissions, or other commercial objectives which can be built into this placement optimization.

And then through the control plane and with the understanding of the application requirements, actual deployments are then executed by the platform based on certain triggers and policies. So, for instance, they could be triggered by actual clients in a specific region, so that when you define for a specific city, for instance, a certain threshold of clients, then automatically the platform will spin up a backend service in the edge cloud to serve these specific clients in that specific region.

That happens fully dynamically, especially also on a per session or connection basis between a client and a backend, which also enables the mobility case, so that when you have a car going through the country, that there is a mechanism to hand over the session between different backends to always ensure the shortest possible path between the car and the backend from a network topology perspective.

So, by having this control plane and these mechanisms as described, we can basically enable the three main aspects of the telco edge cloud, which is allowing for a differentiated quality of experience. We can also ensure service locality and that can be, for different reasons, not only latency, but also other, for instance, regulatory aspects. And in the end, for lower cost, because the data transport is being minimized and depending on the cancellation [00:28:29], moving or offloading computation either from the device or existing onsite or on-prem type of edge deployments to the telco edge, actually leads to a better cost structure.

Another important aspect we have been working on for quite some time now is the aspect of federation across different operators and across different operator edge platforms. So, what that means is that by the efforts of the GSMA Telco Edge Cloud Group and the related Operator Platform Group, OPG, which defines the principles of federation, we have been working with the GSMA groups and other stakeholders to actually implement this type of federation, which means, essentially, that an application which I-- or not I, but an application developer onboarded through a specific operator, so in this example here, or actually in a trial which we recently executed with Telefónica becomes then also available in the federated operator infrastructure.

So, in that case, it was a federation between Telefónica and Bridge Alliance, and the specific POC was with Singtel. So, that means that then a subscriber of Telefónica going to Singapore and roaming in the Singtel network is then served by an edge infrastructure of Singtel, so the specific application backend is deployed in the Singtel edge platform, which is not a MobiledgeX platform, so we are using the east-westbound interfaces as defined by GSMA commonly in the ecosystem to allow this type of federation, which we believe is essential in order to build a virtually global footprint for the deployment of telco edge workloads.

So, in summary, the benefits which we unlock for the telco operators. First of all, full flexibility. So, we allow telcos to deploy and build edge cloud resources with any stack or provider of their preference, including public cloud providers or any combination. So, not limiting their-- any decision or any selection by the telcos.

But we still provide control to the operator by having this aggregated layer on top of these different stacks and, therefore, allowing the operator to keep full control of the applications and workloads, which are being executed on their infrastructure.

And the third aspect of profitability is obvious through means of the federation and easy exposure of edge infrastructure to the application developers who are in need of telco edge infrastructure. The telcos can monetize their investment into the edge infrastructure. In combination, by the way, with the exposure of the network capabilities. And I think this is a very important aspect that we see more and more the combination of exposing network capabilities related to QoS and related to latency and the edge orchestration together are the necessary ingredients to really have a powerful and sustainable value proposition towards the application developers.

An example use case we have been working on with Intel is with a software partner called Kibernetika, so this is about, basically, creating heatmaps for people walking, for instance, through a retail store, so that analysis can be done where the buyers are actually mostly present and, therefore, you can do some optimizations or take some measures if you experience some issues there.

And what we did there is, basically, using the Kibernetika workload, which is using the Intel OpenVINO toolkit to utilize the specific Intel accelerators being available in a specific deployment over the MobiledgeX platform, orchestrated then on-demand to the locations where currently some requirements for this analysis exist.

This is an example here of the actual heatmaps being produced. So, you can basically see, as mentioned, where the people are in a retail are, and based on that can do some optimization or generate some metrics like just tracking people and counting them.

This is another example of the actual heatmaps being created. And this is just one example out of many of those flow type of edge-native services we mentioned, and-- which are then dynamically orchestrated to a specific edge infrastructure.

Another example of use cases, this time here for network functions is an activity we did together with Deutsche Telekom and Mavenir where, basically, UPF, obviously connected to a 5G core, was dynamically deployed to an Intel infrastructure over the MobiledgeX platform. And the dynamic deployment of a UPF is a very interesting aspect, specifically for optimizing roaming, but also for in-network optimization of the traffic breakout. For instance, if you have specific short-term demand of doing that at specific locations, and this ability to do that dynamically, basically, gives you then the opportunity to optimize for cost.

So, in summary, the ability to define these federated edge clouds based on any combination of different edge infrastructure stacks including public cloud infrastructure is an ongoing opportunity for telcos as we see it, because as mentioned in the beginning, there is a demand from different types of application workloads. And especially then in combination with exposing specific network capabilities, this is a unique opportunity for telcos.

We, as MobiledgeX, provide a flexible and universal orchestration platform and control plane for exactly this kind of topology and optimize it for the performance on Intel architecture and Intel hardware. And this combination allows telcos to offer their edge cloud services with an optimized TCO and service levels while retaining full control and having full flexibility. So, this, I think, is very important that with this approach, telcos are fully free to decide upon their stacks and how they want to lay out their edge infrastructure, but keep the full control especially also if desired about the application developer, application customer relationship.

So, in a nutshell, on-demand edge clouds have the potential to deliver new revenue streams while improving the utilization of capital investments through the already existing infrastructures which are there to serve the internal network functions.

Thank you very much. So, that was an overview of what we, at MobiledgeX, have been doing with Intel and I'd like to thank, again, Cesc and the whole team for the great collaboration. And with that, back to you, Lilian.

Lilian Veras

Fabulous. Thank you, Thomas and Cesc, for such a great presentation. We do have a few questions that have come in while you were presenting, so let’s get started on those.

The first question we have here, “What are the use cases for telco edge that are relevant today and how do you see these going forward?”

Thomas Vits

So, maybe I’ll take it first and then, Cesc, you can extend on that.

So, what we see today is, basically, a focus on those use cases where you have a defined connectivity, and these are then, for instance, use cases like video analytics, so the flow type of applications I was mentioning. What we see, to some extent, is also content delivery type of applications. While then if you look into the future, the aspect of federation and cross-operator comes into play. And this will then unlock the full range of different use cases, including then, in the end, B2C type of services like gaming et cetera.

So, in summary, it's a full range and Cesc has been showing that nicely with all the different verticals, but it will play out differently over time. For now, as a starting point, we see, as mentioned, the video analytics and content delivery type of services as a starting point.

Francesc Guim

I fully agree with what Thomas just said. And maybe something to comment on the use cases and use cases implementation is that I think that as the technology becomes more mature, we will see how the use cases are evolving over time.

And for example, this week we are heading to the Intertraffic Conference in Amsterdam where we are seeing a lot of video analytics use cases as Thomas is pointing out. And we are seeing some-- how of these use cases that have been implemented today with video analytics like neural network based video analytics implementations, how they are incorporating new elements such as Lidar or other types of sensors that are being utilized for improving the use cases itself, either in accuracy or even in removing false positives and stuff like that. That's certainly dictating as well how the architecture has to change.

And I think that as these technologies become more mature and more deployed across the different geolocations or countries of where edge has a footprint, that's going to change as well how the use cases that are available, and the technologies that are underneath.

But to Thomas's point, I think that today we see a lot of footprint in terms of video analytics and content delivery and gaming starts to ramp as well.

Lilian Veras

Awesome. Thanks to you both. A second question we have here, "The hyperscalers have their own edge offerings. How does MobiledgeX differentiate from that?"

Thomas Vits

So, indeed, I think there are two specific aspects, as mentioned also during the presentation. So, one key aspect is the federation which we enable. And we've done now several trials, you can read that up on the GSMA website, across different operators. And as we were talking about the use cases going forward, this is really key for a wide range of use cases, for instance, automotive, so in the area of connected cars for safety applications and others, you need to be able to run services across the networks, because you cannot assume that the users all sit on the same network. So, that is the one aspect.

And the second aspect is the deep integration with the network itself, and the exposed capabilities of those networks. So, the ability to, together with the actual edge orchestration, have control points towards the network to ensure that on the access network and transport network towards the edge, the benefits of edge by having these reduced data paths are preserved. And this is something, again, we are working very closely with the GSMA groups and the whole telco ecosystem to do that type of integration, but bringing this altogether is what, I believe, is necessary for truly unlocking the value of telco edge.

Francesc Guim

I just want to add, I'm 100% concurring with what Thomas said. I think that from an ecosystem perspective and to truly provide the value of edge into the solution providers, ISPs as they develop use cases, we have to democratize what edge is and making as much as possible accessible-- edge accessible to these developers regardless to the location and what technologies are underneath.

To Thomas's point, typically, what we are seeing is that there are lots of - [00:46:05], I don't want to say small but medium players, that they are developing edge solutions, and they're investing their time on optimizing the application from their domain perspective. They have to focus on that and not the other aspects. They're not providing value-add on the other types of architecture. So, I think that by providing these common layers that democratize all the different edge providers and allow this movement from one place to the other, we are also allowing the ecosystem to focus on the areas that each of the players can provide value, and want to provide value.

Lilian Veras

Awesome. Thank you. The last question we have here from the audience, “Do you envision an environment where edge cloud located GPUs make it possible to deploy cheaper camera endpoints for video analytic apps? For example, for public safety or retail op use cases?”

Thomas Vits

Yes, generally, accelerators and specifically GPUs allow for offloading-- either offloading functions which are today executed on devices, whether that’s a camera or whether that’s some other type of embedded device to the edge, and that allows and obviously for producing these devices with a lower bill of materials and improved form factors. So, we have been working in other areas as well, for instance, for AR headsets where you offload certain functions to the edge using accelerators, but cameras is another great example. I definitely believe that this will happen.

What is here also important to say is that we always see, basically, a three-tiered architecture, and what I mean by that is there is still something-- there will be something happening on device, there will be something happening at the edge, and there will be still something happening at the central cloud. So, I'm not saying that everything moves from the device to the edge, but certain functions which are relevant and need the edge will go there, while others, as mentioned, remain on device and in a public cloud.

Francesc Guim

I just want to add that I think one of the key aspects here is to really understand, again, to my previous comments at the beginning of the presentation that you need to understand-- and that now goes back into the use case implementation, you always will have a trade-off of your application, how much is general purpose or business logic of the application, and how much you can accelerate with those accelerators or if you have a video application, and video analytics application, very likely you will have 80 to 90% of the application that can be accelerated depending on the use case like with the GPS, you may have some other applications that may utilize, maybe 40% of the GPU and 60% remaining, it’s going to go into the general purpose CPU.

So, I think that from that perspective, we have to really understand and that’s an ongoing effort to understand how the different use cases behave in this ratio of acceleration versus general compute and make this matching on where the accelerators make sense.

To Thomas’s point, you have different tiers and depending on where you are mapping the use cases and how they look like, you will have some places where it makes more sense to have GPU acceleration, and some others that will say, “No, it’s more preferable to help from TCO perspective, like CPU only”. And the point is that you really need to understand the use cases and how they are implemented and how that translates into the system architecture and utilization.

Lilian Veras

That’s great. I'd like to thank you both again for such an insightful presentation. That was our last question, and I would like to ask our audience to please not forget to give our team a rating for the live recording, so that we may continuously improve the quality of our webinars.

Thank you so much again, Thomas and Cesc, and I’ll see you next time.

Francesc Guim

Cool, thank you.

Lilian Veras

Thank you.

Thomas Vits

Thank you guys. Thanks, Thomas.

Lilian Veras

Bye.

Francesc Guim

Bye-bye.