Intel

Reference Implementations- Free, Customer-Deployable Apps for Customer Use Cases

CORPORATE PARTICIPANTS

Xiaojun (Shawn) Li
Intel – Sales Director, Next Wave OEM & eODM

Hassnaa Moustafa
Intel Corporation – Principal Engineer

PRESENTATION

Xiaojun Li

Welcome everyone to the Intel Network Builders Insights Series. I'm Shawn Li, Sales Director, Next Wave OEM and eODM. Network Communications is a sales organization at Intel Corporation, and I am your host for today's webinar.

Thank you for taking the time to join us today for our webinar titled Reference Implementations – Free, Customer-Deployable Apps for Customer Use Cases.

Before we get started, I want to point out some of the features of the BrightTALK tool that may improve your experience. There is a Questions tab below your viewer. I encourage our live audience to please ask questions at any time. Our presenter will hold answering them until the end of the presentation. Below your viewer screen, you will also find the Attachments tab with additional documentation and reference materials, which pertain to 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 will use the information to improve our future webinars.

Intel Network Builders Insights Series takes place live every month. So, please check the channel to see what's coming and access our growing library of the recorded content. In addition to the resources you see here, we also offer comprehensive NFV and 5G training programs through the Intel Network Builders University. You will find the link to the program in the Attachments tab as well as a link to the Intel Network Builders Newsletter.

Today, we are pleased to welcome Hassnaa Moustafa. Hassnaa Moustafa is a Principle Engineer at Intel, currently working at Edge Computing and AI Solutions across IoT segments and network edge infrastructure. Previously at Intel, Hassnaa led Car-to-Cloud Solutions for connected autonomous vehicles, connectivity technologies across IoT segments.

Before joining Intel, Hassnaa was a senior R&D engineer at Orange in France, where she contributed to cost-efficient wireless network solutions for Europe and the Middle East. And led engineering efforts in the video and multimedia services optimization over wireless networks. Hassnaa earned her Ph.D. in mobile networks from Telecom Paris Tech and her master's degree from the University of ParisXI.

Welcome, Hassnaa, and thank you for taking the time to join us today.

Hassnaa Moustafa

Thank you, Shawn.

Xiaojun Li

I will hand it over to you.

Hassnaa Moustafa

Thank you, Shawn.

Xiaojun Li

I will hand it over to you to start off. Thank you.

Hassnaa Moustafa

Thank you, Shawn. And thank you everyone, and welcome to this session. And I'm pleased to share with you what we have today on reference implementations. Thank you so much.

So, yes, as Shawn explained, we want to share with you today in this webinar the efforts that we do at Intel for reference implementations for edge solutions, which helps in providing to our partners, to ISVs, to different edge solution builders sort of free and customer-deployable apps that you can use and deploy for different use cases that you have.

So, first, let me start by the motivation. Why do we do this? Why do we do it? We do it because we have different challenges and also opportunity with the edge. With the edge itself, we have a big opportunity, but this opportunity is accompanied by a challenge. What is this challenge from a software perspective, from edge software solutions?

The challenge we have is... the opportunity we have is edge computing is bringing the cloud three sources to the edge, which is great. But at the same time, it requires that every edge solution be a cloud-native solution working at the edge theme, as it were, in the cloud. The challenge here is the edge has a diverse capability compared to a cloud deployment. If I connect to the cloud for any service, I have a cloud cluster, I have an environment, which is a homogenous environment, a high-performance environment, which is known. The edge is another story. We can observe edge deployment composed of diverse compute capabilities, from a platform which has core CPU capabilities, up to a platform having server capabilities. So, we have diverse compute capabilities at the edge. That's one of the differences which builds a challenge for us.

Other challenge is edge locations. If I connect to the cloud, again, we have the cloud. We have the data centers. We know what it is. For the edge computing or edge cloud, location can vary. Location can be an on-premise edge in a factory. It can be an edge device on a roadside unit for smart cities. It can be an edge compute platform in a regional data center of a telco. So, the location itself changes, which creates different requirements. Different requirements, security perspective. Different requirements, compute perspective, latency perspective.

So, what we conclude is there is no one edge solution that can fit all. From an edge solution builder or a software solution builder for the edge, the challenge is you can create your solution once and be sure that it will work on any edge device, on any edge location. That's why we're here to bridge the gap between the cloud and the edge, and the reference implementations we provide is assured to do this.

What do we do here with these reference implementations? We offer a comprehensive set of software toolkits and applications using these toolkits in a way that abstracts underlying hardware complexity. From a developer perspective, from an edge solution builder perspective, you wouldn't feel this diversity of the hardware, all is abstracted for you and you will focus on building your application.

Our software frameworks and tools we offer with our reference implementation allows you to deploy your solution on any edge platform without constraints of the edge location itself in a modular approach as I will explain in the coming part of the presentation.

So, that, to start, is why we do it. We do it to help you, to help you to bridge the gap between the cloud and the edge to alleviate the challenges that we see with the edge deployment, and to make your focus on your solution, build it, don't worry about the underlying complexity of the hardware. And work better with Intel architecture, of course, with optimized performance.

So, here, if we look at the edge from a holistic view, the big picture, it will iterate on what I mentioned, you find that location is different. Location can vary from an on-premise edge for factories, for healthcare segment, retail. Up until the network infrastructure edge, which can be either access edge at the access network level, or even near edge beyond the access network level, or regional data center of a telco. So, different edge locations and different latency requirements, as you see showing the latency from one millisecond up until 40 milliseconds. With the network edge infrastructure, latency itself will change, so this gives an idea on if you have an application that is very, very latency-sensitive then an on-premise edge will be a better choice for your solution.

Key challenges that I explained across this heterogeneous edge location and edge capabilities is how will you scale your solution? How you build the solution once and then you scale it to work on an on-premise edge or access edge or even regional data center edge? And how to have the needed optimization to meet the latency need as well for your application.

So, what we will show today is how to leverage our software building blocks from Intel frameworks. They work with optimization on IA, and you can build your solution to overcome this challenge. And we give you a reference implementation as a reference solution as well, one you can use as a jump start to reduce your development time.

If we look here, as I explained, we follow a cloud-native approach, of course. Cloud-native microservices approach. Our whole idea is to enable edge cloud. What is followed in the cloud world from a solution perspective, how they are built, microservices, cloud-native, that's what we are following. We are following a modular approach to build – we call it One Edge Reference Implementations, meaning you have reference implementations that can work on any type of edge that I explained.

If we have a bottom-up overview here, it all starts by the edge platform. Edge platform Intel recommended hardware, we recommend specification for the hardware to fit any location. For an on-premise edge versus network infrastructure edge, there are different power requirements, for example, compute requirements. So, we provide recommended hardware that fits any edge location.

And on top of this, we have Intel Smart Edge Open, which is a software framework offering you cloud-native capabilities, orchestration capabilities using a Kubernetes-based approach. It's based on Kubernetes, and it allows to have optimization on IA. It expands Kubernetes with several optimizations on IA. And Intel Smart Edge Open is certified by the CNCF, Cloud Native Computing Foundation. It's certified as a Kubernetes product.

So, Intel Smart Edge Open acts as the bridge between the platform and all what's on top. It abstracts underlying hardware complexity of the platform. It allows ease of use for any accelerator on the platform, whether accelerator for network workload or accelerator for your application, or accelerator for even ethernet kind of communication. All this is abstracted to the application.

And if we see here, we are putting Developer Experience Kit. We have different flavors for Smart Edge Open. One flavor is for developers is to use and build their applications. That's the Developer Experience Kit. We have other flavors for private wireless deployment and mainly targeting RAN – 5G RAN and 5G core vendors who want to bring their RAN and 5G core software on a cloud-native platform.

For this talk today, we're focusing on Developer Experience Kit, because this is targeting developers, software solution builders for the edge, IoT solution developers as well.

So, on top of Intel Smart Edge Open, when I say on top meaning Smart Edge Open on-board, on-board all the blue box components that you see in this picture. On-board what? On-board the different toolkits and capabilities for like Edge Insights for Industrial is a software framework by Intel enabling data ingestion, enabling analytics, and enabling even insights sharing. This is onboarded as cloud-native microservices with Smart Edge Open. We have different libraries and toolkits from Intel like SVT-HEVC, which is scalable video transcoding for HEVC video optimization on Intel Xeon platforms. We have it onboarded as a microservice with Smart Edge Open.

We have Intel WebRTC toolkit, which is WebRTC framework not like the open-source framework, but it's by Intel, optimized to work with better performance on IA. And we release it in a regular cadence. So, this is as well onboarded with Smart Edge Open. We have network AI applications even offered and onboarded to help you if you want to build an application for network AI, for network analytics, you will find it. And we have several components for AI inference, optimization, for AI workload mainly based on OpenVINO toolkit. We have OpenCV Deep Learning Streamer. We have OpenVINO Model Server allowing you to send requests to OVMS, OpenVINO Model Server for inference. Meaning if you have a camera – a smart camera, for example, which cannot do the processing for video analytics, so the camera or a small IoT device can send requests to an edge device to OVMS, OpenVINO Model Server for inference requests. And then OVMS will provide the inference data, the insights, and send back to the IoT device. So, it allows a very scalable deployment.

We have different components, different microservices with Smart Edge Open, enabling you, as a solution builder, to leverage and build your own solution. We call them even AI-inference-as-a-service, real-time communication, through WebRTC-as-a-service, data-ingestion-as-a-service, data-insights-as-a-service through AII framework. We have even RAN Intelligent Controller-as-a-service to help you build network AI applications.

These are the building blocks, the modular building blocks that you can make use of.

In addition, we build edge reference implementations that makes use of these building blocks. So, as an example of an application solution, which is available for you and making use of these building blocks and working on Smart Edge Open edge node. So, this is something that we handoff, we provide, we share, and enable you to leverage and use and expand.

These reference implementations, they are available in an easy way through Intel Developer Catalog. If you go to Intel Developer Catalog, it's like a one-stop shop for all our software, and you can download reference implementations. You can see the user guide on how to make it work and to install it on your platform.

So, that's how we are thinking of it to make sure that we are contributing to this new world of building solutions for the edge and bridging the gap between the cloud and the edge. And easing for you the leverage of Intel hardware and Intel software capabilities.

What is the value, again, I can iterate of these reference implementations? Because I think it's important that you know the value because we can have different audiences. Different audiences, different target audience for the reference implementation, and each target audience will find a value from it.

So, first, it's reduced development cost. If you want to build an edge solution in the media domain, smart cities domain, retail, industrial, you can use one of our reference implementations as your starting point, and start to expand, adapt, change and build your own application. So, it reduces your development cost. It, of course, accelerates time-to-market. You can go fast. They can serve as the jump start for you. So, it will accelerate time-to-market.

It helps also to scale across the edge portfolio, meaning what? Meaning if you have – I told you the story of abstraction of the underlying hardware complexity, making you focus on building your application. So, at the end of the day, you can scale. You can scale across different edge platforms with different capabilities.

And other important point is the developer outreach. We are – this is for us to do an outreach to the developer community. If we keep offering great software capabilities without doing an outreach, maybe there will be a missing point here. So, we do the effort, we are proactive, we are offering this through the Intel Developer Catalog. It's sort of developer outreach as well.

And by the way, there are several hackathons, even where Intel contributes to and we're using reference implementations in these hackathons. And developers they join, and they work, they expand. And we have very interesting feedback from these events.

Here is what are the reference implementations available today where you can go and download and use. But you know this is a continuous effort and we expand. So, what I'm saying today is what we have today, but by the end of the year, you will see more and more and so on.

We have Defect Detection for industrial environment, PCB, which is Printed Circuit Board. We have different partnerships with our industrial partners, and we noticed that in factories, specifically when it comes to chipset design and so forth, the printed circuit board, of course, there is a lot of yield aimed at the end. It can cost, it can be a heavy and expensive operation if we don't do defect detection in a very advanced way. Instead of having someone keep looking and examining each printed circuit board, that's an automated way using AI, using video analytics with the different optimizations I explained to do printed circuit board defect detection. And we have feedback about the saving in terms of time in this process thanks to this automated approach.

And here, these are hyperlinks, you will get the material, so you can go click the link and you will go on where to download.

In the health segment as well, with our IoT health segment at Intel and with our partners, we see a big need for telepathology. Telepathology means it's the lab analytics, but making it automated. We all experience different lab analytics in our life, and we see how it's a heavy process, it takes time. It takes time, it's not a real tight process. So, what happens here, this is a reference implementation which allows also to use image analytics and do automated analytics for the samples collected by the images. And the value we see also, it can be shared across multiple medical staff even to do – check after the insights coming from AI. And it allows contribution from medical staff who are collocated in different premises.

So, it's very efficient and we have this deployed as a pilot with different partners we have in the health segment. And it's getting a big success. And we're asked also to do – we're learning. We're learning requirements, for example, such reference implementation looks for security. High-level security because we are dealing with patients' records, information, samples. So, we're learning also the different requirements we need to incorporate.

Telehealth is another reference implementation that we are having as well, allowing remote health. Remote health like video session happening between the medical staff and the patients. And we also – during this pandemic time, remote health is important. So, that's a reference implementation enabling secure communication, secure video sessions between the medical staff and the remote patient. It can be applied also within the hospitals, that's the requirement we got where even patients who are kept in the hospital and in their rooms, or intensive care or sensitive situation, you cannot have a doctor checking every minute on the patient. So, having cameras, and monitoring, and doing analytics, and seeing the status of patients in the hospital, and sending alerts to medical staff in the hospital in real-time, that's very valuable. So, this reference implementation can serve as well.

Immersive media is another one where we are having – we are offering AR/VR experience in a distributed way where you have cameras. I'll take an example like a stadium. A stadium with interactive experience, sport event – important sport event and the stadium providing an interactive experience to the spectator in the stadium, or even remote spectators. So, cameras will be everywhere and camera streams will provide video streams to the edge platform. And then video stitching happens at the edge platform to provide 360-video. And this 360-video with like region of interest highlighting also will be provided to the end user with an AR/VR headset. It allows interactive experience where the end user device itself can ask for a specific field of view, can ask the edge platform, can send a request to the edge platform for a specific field of view and get the content based on that. So, it's not only 360-video experience, it's also interactive video experience.

This one is observing very big traction for different scenarios, for entertainment, sports event, and also industrial environment, if you wish, for remote debugging, for example, also it can help a lot.

Smart city, or smart intersection is mainly targeting for smart cities. And this enables edge platforms on intersections, on road intersections to monitor through video analytics coming from camera feeds, monitor the status of intersections. Density of vehicles. Density of pedestrians. And is getting very big traction in crowded cities where there is a need to learn about high-risk intersections, and also a need to learn about traffic peak hours. How the cities can better regulate traffic lights, for example, based on the density in peak hours. So, this reference implementation is helping as well.

This last one here is mainly network AI. It's intelligent connection management. It enables what? It enables automated handover to happen. It enables the telco to trigger automated handover from cell to cell based on the connectivity and based on the quality of service needs.

Usually, how the handover happens today in the cellular network when you connect with your mobile. Your mobile experience hand over from cellular tower to another, there are fixed configurations that the telco is doing. But AI now is permitting the telcos to do it in an automated way, in an advanced way not based on fixed rules, but based on real-time situation of the network. So, here it's heavy deep learning involved in this. We're using a lot of optimization from the different building blocks I showed in earlier slides.

So, these are like – as you saw, is covering a wide variety of segments even up until the network AI, leveraging the modular approach I explained, leveraging Smart Edge Open, and the combination of different microservices I showed in my earlier slide.

One more here in addition to the reference implementations, I want to spend a minute here to show you the capabilities we have with Intel oneAPI software tools for AI And analytics to tell you how you can leverage these powerful tools in building your solutions.

So, we have one Intel toolkit – oneAPI toolkits, it offers several types of sub-frameworks. It's a big framework which has different sub-frameworks. We have, for example, oneAPI AI analytics toolkit, and this one is targeting data scientists and AI researchers who are building their AI algorithms and they need performance optimization for the AI analytics. So, that's the target audience. We have oneAPI base toolkit, which is composed of different libraries, oneDNN, oneCCL, and oneDAL. And the target audience here is deep learning framework developers.

If you are building a framework for deep learning, and you want optimization in this framework, you can leverage this toolkit.

We have OpenVINO as part of the oneAPI software tools. OpenVINO is a part. And OpenVINO is mainly for AI inference optimization at the edge when you deploy your solution. You have an AI solution deployed at the edge and doing inference. OpenVINO will allow AI inference optimization. And it supports several neural networks that can be optimized, several neural network models.

Here, a quick glance on the Intel Developer Catalog that I explained. It's the one-stop shop where you can go, download the reference implementations, download any software capabilities that can help you. It covers a wide variety of segments for IoT as you see here. You can have even OpenVINO by its own, you can download it by its own. You have Intel Smart Edge, you can download it by its own. You have several media experience toolkits, Open Visual Cloud. You can pick and choose what you want and build your solution.

Besides the reference implementation we provide, you can also find at Intel Developer Catalog plenty, plenty of standalone frameworks you can download and use.

That's a look inside. To see how it looks like if you didn't have the chance to visit Intel Developer Catalog before, you have search by keywords, and you can get different options for whether implementations or toolkits, and you can pick the package you need, then customize for your need. You can pick the OS of your choice. You can pick the platform of your choice. Whether it's a core CPU platform or server. There is a quick license, which is clickthrough license. It's nothing. It's a non-monetized license, but it's an evaluation license. And then you will get specification for some recommended hardware, also hardware specification for some components and you can download.

So, we have so far very good feedback. It's a smooth experience and it's a very important venue where we can share, we can do the outreach to developers and software solution builders for the edge.

Yes, I think, we can leave also some time for interactive discussion. But here is – you have here access also several means to Intel Developer Catalog. If you get the materials, we will be glad that you experience it, and let us know your feedback.

So, I think, Shawn, we can provide time here for interactive discussions, for questions.

Xiaojun Li

Sure. Thank you. Thank you. And we have some questions here. "How does Intel Smart Edge Open help the deployment of reference implementations and why are the reference implementations important to be released with Smart Edge Open?"

Hassnaa Moustafa

Yes, that's a great point. Yes, I'll speak about the first part first. How Smart Edge Open is used to build reference implementations.

Firstly, if you remember, I spoke earlier, I said as the cloud world is moving to the edge, meaning what? Meaning we need to create a similar cloud environment at the edge. So, we need a cloud-native framework for microservices, onboarding, and orchestration. We cannot get out of this, that's the way to do it, and that's the trend. We need a cloud-native framework for orchestration and management of workload. And the workload is mainly microservices.

So, if we don't offer and show how – Smart Edge Open is a cloud-native framework. Smart Edge Open is a cloud-native framework for the edge, optimized on IA. So, if we don't like – if you don't have the cloud-native framework, you will build a great solution that's great. You will build it. But after that, how will you deploy it? How will you deploy it on different edge nodes and how will you orchestrate the workload and manage it? You need the cloud-native framework.

So, we offer you Smart Edge Open as a cloud-native framework for this. Even with a bonus a lot of optimization on IA, and abstraction of the underlying hardware complexity.

If you go to Vanilla Kubernetes, you will have a framework for orchestration, but you wouldn't have all the optimization we're offering on IA, and you wouldn't have all the hardware complexity abstraction we offer.

Why we do reference implementations ourselves? To help you. To help you see how a solution for the edge can be built and can land on Smart Edge Open edge nodes and can leverage the different software capabilities, optimization tools you are having. If we don't show it to you and offer it to you as an example, you need to deal with it yourself and discover. The task will be not as easy as how we do it now.

Xiaojun Li

Great, thank you. Thank you. The next question. "Can the reference implementations be used as a base for the edge solution in solution builders to expand and commercialize?"

Hassnaa Moustafa

Yes, that's an interesting question as well. Yes, that's the whole idea. We offer it, it's like – we don't monetize it. It's free like open sources. It can serve multiple purposes. If the purpose is test and evaluation of Intel capabilities, that's the purpose. If the purpose is just to experience an edge solution and see how it works, that's the purpose. If the purpose is to take it, expand it, and commercialize it, of course, we don't have any obligation or restriction here. We don't have – if our partner or if the software solution builder will go for commercialization, the point is we don't have a support channel for commercial deployment. That's the point. Then our partner or the edge solution builder is interested by taking the reference implementation, of course, they take and expand and commercialize, but we don't support commercial deployment. That's the point. There is a need also for the partner to have a support model for the commercial deployment.

If something happens in the field, for example, Intel cannot commit to follow-up here as we provide this as a reference. But of course, nothing restricts going commercially, yes.

Xiaojun Li

Great, thanks. Thanks, Hassnaa. And one more question. "How do we see the value of the reference implementations with Intel private 5G solutions?"

Hassnaa Moustafa

Oh, that's a good point. Yes, if you remember earlier I spoke about Intel Smart Edge Open Developer Experience Kit, that's what we offer here. That was the topic of the talk to demonstrate the reference implementation. We didn't speak about the 5G piece. But of course, these reference implementations can run on Intel Smart Edge Open edge nodes with private 5G. And we observe this a lot for private 5G, in general, as we know it's the low-hanging fruit for 5G. And that's what we observe today with our partners. They are looking to deploy private 5G for IoT, mainly for IoT, for industrial segment, for smart cities, for different segments of IoT, the healthcare. They are looking for private 5G.

So, when we say "Private 5G", it's happening on an on-premise edge for IoT, meaning in a factory, you will have an edge node hosting the 5G functions. And on this edge node, you will have a co-existence of edge applications, meaning that's what we say convergence of the network functions and the services. We will have edge nodes hosting both, hosting the private 5G components, the RAN software

stack, and the 5G core software stack, and hosting at the same time the edge applications. That can be any application for analytics or example of what I mentioned.

So, co-existence is there. Co-existence of network functions and services is there. So, for this reference implementation, they can simply land with a private 5G workload on the same edge platform. And we tested this with several of our partners.

The only caveat is your platform resources because your hardware needs to support 5G workload with high determinism and very, very low latency. And at the same time, support edge services that can have AI processing that can need also CPU – a high number of CPU cores. But we offer guidance.

That's why when I mentioned recommended hardware. We sit with our partners and we speak and we understand their workload needs, and we say for such a deployment, you need such number of CPU cores. You need, for example, accelerators for the RAN, vRAN, virtual RAN workload of 5G. You need specific accelerators.

We do this guidance, of course, and even we publish them with Intel Smart Edge Open Private Wireless Experience Kit release. So, Intel Smart Edge Open helps also in a way where it can help assigning a specific number of CPU cores for the workload.

For example, when you have 5G workload, you can use the capabilities in Intel Smart Edge Open to assign a dedicated number of CPU cores for the 5G workload. And if you want, you can assign a dedicated number of CPU cores for the AI processing workload, for example. So, you can ensure that all will work with the performance you need. And in a convergent way, you put the network function, the services on the same edge platform.

Xiaojun Li

Great. Great. Thank you, Hassnaa, for sharing such great information.

Thank you all for joining us today and please do not forget to give our team a rating from the live – for the live recording, so we can continue to improve our quality of our webinars.

Please be sure to join us next time, Wednesday, August 3rd at 8 a.m. Pacific Time for Intel Ethernet 800 Series: Delivering High Timing Accuracy for 5G vRAN. You can find the link for the registration in the Attachments tab.

Thank you again. This concludes our webcast.

Hassnaa Moustafa

Thank you. Thank you, Shawn. Thank you, everyone. Thank you.

Xiaojun Li

Thank you. Bye.