Empowering the Edge Everywhere

Nitin Dahad (00:06):

This is the Smarter World Podcast, focusing on the technology and issues behind today's connected world. I'm host Nitin Dahad, technology journalist and editor at EE Times and embedded.com.

(00:19):

So the piece of edge computing is advancing so fast that today's breakthrough will be eclipsed a short while later with another breakthrough. While we see traces of early IoT concepts in devices we use every day, innovations in edge computing are rapidly transforming how we interact in the world around us, whether it's a wearable or in industrial robot. We'll discuss some of these trends with Ron Martino, senior vice president and general manager of the Edge processing business at NXP. Welcome, Ron

Ron Martino (00:49):

Thank you, Nitin. It's a pleasure to talk to you today and happy 2021 to you.

Nitin Dahad (00:54):

Great. Yes, thank you. And you too. So edge computing, I've written lots on this and this has many definitions and means different things to different people at its core. However, it's the ability for efficient processing closer to the user, which can bring quicker insights to data. Can you maybe give me your definition of NXP way of looking at edge computing, and then maybe we can go into a deeper conversation on that.

Ron Martino (01:19):

Sure, I'd be happy to. Edge computing simply is distributed local computation and sensory capability. It's capability that effectively interprets, analyzes, and acts on the sensory data to perform a set of meaningful functions in these functions vary depending on the given application or setting that the devices are in. They can be single edge devices, or they can be the aggregation of multiple devices to create an aggregated intelligence, so to speak. It's also local compute capability, which is complimentary and interactive with cloud computing. So it doesn't try to be a replacement or an alternative to cloud. It becomes complimentary capability. And then to take it a little bit further, it's also evolving over time. Early rollout of IoT types of devices really started with traditional devices collecting data and then aggregating and sending that up to a cloud for processing. And this is classically seen in terms of voice assistance where a lot of data is sent to the cloud and the higher capacity of compute is used to really enhance the experience.

(02:32):

Edge computing is evolving to become smarter and then more intelligent. And in what we mean by those expressions is smart edge computing is balancing the use of local computing with central or cloud computing, simplifying the data at the source of the data and then only using the cloud when it's necessary and using it efficiently. And that specifically is driven in many aspects by cost of ownership. And then as this evolves even further, there's more intelligence where we want the end devices to have more capability to do the interpretation, the analysis, and then make decisions locally. And that involves a whole set of content that we can discuss in terms of what NXP is doing in edge processing solutions. That's really helping this transformation.

Nitin Dahad (03:25):

We'll agree that technology is impacting society in a meaningful way. Energy and efficient smart building systems can monitor occupancy and controlled temperatures. We have more safer automated driving features in our cars. How do you anticipate the industry will build on some of these examples? And maybe you want to give some examples around enabling better productivity, safety and greenness.

Ron Martino (03:48):

Exactly. And as you said, those are the aspects of the impact of edge computing on society and also involving high security of data. In the case of productivity, a great example is enhanced workforce. Leveraging edge processing or wearable devices that are enabled with vision and machine learning where a worker can diagnose a problem and more rapidly repair it, whether it's in your home or whether it's in your factory. When it comes to safer world, these intelligent edge devices really can enhance safety by recognizing various danger signals. They can recognize alarms, they can recognize a person who has fallen or glass breaking, and it can apply more intelligence in terms of determining what the source of that issue is through various sensing means and various computation means whether it uses radar sensing devices that NXP is developing, whether it is using vision capability or just the interpretations of audio input into the device, you can act upon a set of information that's collected by the computer or that computing system and then take action.

(05:04):

It can prevent theft of a car from your property, or it can save or get more rapid help to a person who is injured within your home. And then if we go to greener and energy conscious concepts, there are many ways that one can focus on optimization of accomplishing these tasks. One is this concept of vampire power where you plug in devices and they're doing nothing but they're consuming wattage, they're consuming power. How do we eliminate unnecessary use of power and only turn on the capability that's necessary at the time that it's needed? And the devices that NXP are creating are focused on this very efficient use of processing and sensing in doing it in a way to collect and process things efficiently, which includes the concept of bringing it up to cloud computing, but doing it in an efficient way of sending only the necessary workloads that you want to take place on the cloud.

(06:11):

And then conceptually, we take it to the next level, which is this concept of an aware edge where devices really respond more in a human-like behavior. They start to understand their environment, they aggregate inputs and interact with other devices to collect information and understand the context of the situation and then take decisions accordingly. And just a simple practical example of this is traffic patterns with local capability that can interpret crowds and different congestion points and really optimize the situation locally by observing and sensing the number of cars and the conditions of the environment to make just driving more efficient so that you don't waste time.

Nitin Dahad (07:00):

Yeah, that's quite interesting. I did just write an article on something with the San Francisco Metropolitan Transport Authority doing some trials in that area, and as you said, monitoring the traffic patterns and then figuring out when you need to increase the traffic signal sort of frequency or decrease. So it's quite interesting how you can make them aware.

Ron Martino (07:21):

Yeah, yeah. It's an entire industrial engineering application that is evolving into edge computing.

Nitin Dahad (07:28):

Moving on from a technology perspective, what do you need to have and what are the pieces that make up that intelligent edge or as you are going towards aware edge, the aware properties requires more the concepts of mesh networks and self-healing networks. But let's start basically from the intelligence and what you need.

Ron Martino (07:45):

So let's start with the foundation. You need to have compute platforms and they need to scale, they need to be energy efficient. And what's changed, I've been in this industry for 32 years and it used to be single core performance and frequency as the main determining factor of is it a good compute device or processor? And now it's really about multiple independent heterogeneous compute subsystems. Okay. It's a lot of words there, but it's basically having a GPU, having a CPU, having a neural net processing unit, having a video processing unit, having a DSP, and then how do you optimize these different hardware accelerators and compute devices and optimize it for again, a given end application. And that's where NXP really excels and being able to scale from low-end controllers that are in the form of CPU in a single digit DMIPS and milliamps all the way up to hundreds of thousands of DMIPS and tens of watts having this scalable range of compute that has all these other elements of capability around graphics, around processing DSPs, and then fast real-time micro controllers.

(09:05):

NXP is introduced gigahertz, MCU, and we are going to continue to extend that. That's just on the real-time core. And then integration of other very optimized hardware accelerators or capabilities that are targeted at voice applications, at human machine interaction that involves both vision and voice together. And then do this in a way where it's truly involving ultra-low leakage with operational modes that can adjust to really optimize the energy usage, the area under the curve, even with these large on chip memories, which are necessary as you look at some of the workloads.

(09:47):

And then this goes on to optimization of machine learning capability, security integration with the highest levels of coverage for many different attack surfaces, efficient connectivity capability in terms of energy usage as well as open standards. And it can also take advantage of some great technology NXP offerings such as high accuracy distance measurement, whether it uses our UWB technology for locating in a very accurate way, the physical location of a given person or tracking device. That becomes very interesting use cases when you're looking at asset tracking or you're looking at security concepts such as access to a given secure room. And then the final thing is wrapping this all on a seamless user experience because if it's not easy to use and it's not natural to use, then it won't be used. So getting a seamless comfortable experience in place is absolutely critical.

Nitin Dahad (10:49):

So you did it outline quite a lot in this area. In terms of the other pieces that you need, I know NXP, and many other companies build reference designs that sort of comprise all of these. If you wanted to build something with that heterogeneous compute architecture and you wanted some micro controllers, then you wanted some security, I guess you've got things that people can use.

Ron Martino (11:10):

Yeah, that's correct. What we do is from just a basic offering of a processor or a microcontroller all the way up to a reference platform that's pre optimized for local voice or second for vision detection and inference capability or the combination of those, we put together reference platforms that a customer can purchase, such as our RT family of devices. We have face recognition offering that can be purchased, and that's a fully enabled and design of the system that a customer can take as a starting point in modifying customized to their needs where they want to specialize or where they want to brand it.

Nitin Dahad (11:57):

I remember seeing them, I'm not sure if it was Embedded World in Nuremberg in 2020 or 2019, but I definitely remember seeing quite a lot on the NXP booth. Great. Most would agree that intelligent devices and systems in our homes and work are gaining traction. What are some of the technology differences between the plain old IoT, as you would say, and the industrial markets?

Ron Martino (12:19):

Sure, it's a great question, but let's start with just basic differences. In the industrial front, the connectivity standards, the environmental requirements, longevity needs, the time in which that they want to purchase and use the part which can be 15 plus years, and the safety requirements are much more extensive and demanding in the industrial space when you compare it to the IoT world, an example of this which could is sometimes referenced around an industrial 4.0 topic is the evolution of these interconnected machines and the use of higher bandwidth, higher throughput connectivity. One of the areas NXP is investing in this area is on time sensitive networking or TSN and the integration of both the Mac and Switch into a whole set of devices that can support daisy chain setups of multiple machines or can support endpoint functions leveraging this more deterministic TSN backbone, which also supports much higher data rates and throughput that many of the legacy standards are converging on in terms of achieving the next level of performance.

(13:35):

When you compare this to IoT markets, there's a much broader need for extreme energy efficiency, a higher use of voice HMI, high use of wireless connectivity and shorter lifecycle for the applications like smart homes and wearables. And just to expand a little bit on the wearable front, there is an interesting world where you want great user experience, but you want the longest battery life. So the optimization of these edge devices to perform their functions but then shut down and preserve battery life is very important. And that really rich user experience has to be done in the most efficient way because that's the time that it's burning the battery.

Nitin Dahad (14:16):

I can believe that I bought some electronics over Christmas for monitoring health parameters, and it's interesting to see how long some of these have to last.

Ron Martino (14:24):

Yeah, no, it's true. And it's a great area when you look at wearables and what they can do to make our life safer, whether it's just personal health monitoring or whether it's applied to conditions that were experienced today where you can monitor for signs of disease and leverage that for containment. So it's very relevant in the world that we're living in today.

Nitin Dahad (14:45):

Yes. And talking personal devices and home devices, one of the things I think is quite frustrating is in the smart home area, you can't take a smart home product from one company and make it work with other devices. Well, at least I don't think it's that easy for the sort of average person. I understand that NXP is trying to change that with an open standard for smart home wireless interoperability. Can you tell us more about this initiative and why it's important for NXP?

Ron Martino (15:12):

Sure. I'd be happy to. As you said, when you look at take home devices, just as a main example, the standards and the interoperability capabilities is very fragmented across many well-known and strong companies in the market. What we have today is a project that it's named CHIP, which is connected home over IP standards project, and it has NXP as well as a number of these leaders in the other leaders in the industry working together to try to consolidate, not to proprietary standard, but to an open standard that is common across all of the industry and allows people to build on this open standard.

(15:57):

And as you said, as you try to connect smart home devices, it's not easy. It's using these different standards, but the focus of this project is to build on the many years of work that NXP and others have done around ZigBee and thread in the ZigBee Alliance, and then build upon that with an upper layer capability leveraging the technologies that the Amazons, the Apples and the Googles have rolled up to build this open standard, which we reference as CHIP and establish this common linkage between devices.

(16:32):

So when you plug something in it will be very simple to connect. NXPs plan is to have actual products in the market later this year with the first versions of the CHIP standard and many of these standards that will continue to evolve and enhance over time and will continue to drive this with the other leaders that we're working with in the industry.

Nitin Dahad (16:55):

Okay. Now, I'm going to come to Edge Computing and then we'll talk about AI and machine learning. So there's two areas I want to discuss and one, the first one is around complexity. Even being an electronic engineer myself, when somebody says machine learning, that sort of makes me think of my first-year electronics degree course where it was all physics and lots of equations and things, which I couldn't understand, shouldn't really been admitting that. But this brings up a visual of complex models, graphs, and code. Is it really that complex? So we'll touch complexity now and then we'll talk about cost in a little while.

Ron Martino (17:30):

Sure. You're absolutely right. For many, when you talk about AI and ML, it's a very complex abstract concept. As we look at how it's going to be applied, there are projections that 90% of all edge devices will use some form of machine learning or artificial intelligence by 2025. And we truly believe that's the case. And we are rolling out products that are optimized around this, and it's a combination of what we do to optimize the hardware that's used, the processors and the micro controllers to run this capability. But for the end user, it's more about how complex is it to deploy practical ML that's meaningful to the end use case. And what we've been focused on here as well as others in the industry is how do we create ML models such as cloud providers from Amazon, Google, Microsoft. They're offering a range of these models that can be used, they can be trained with your own data and augmented and adopted for your use case, or there are many companies out there that want to collect data and create their own models.

(18:36):

So what NXP is focused on is how do we enable a cloud agnostic capability that allows flexibility in a simple user interface or development environment? And that's what we've recently announced with our investments with Au-Zone. And what we're going to be rolling out in 2021 is an enhanced development environment where you can choose the type of content that you bring in your own data models that you have, or models that you have chosen to acquire through your favorite source or cloud provider and bring that and then optimize it and deploy it onto the end device because it's that optimization using something that's raw data could be very heavy in terms of what it requires to run it, but if you optimize it and tune it for a given edge processor and you deploy it specifically on one of those subsystems that I was describing earlier, which there are many of now, if you do that in the most effective way in this development environment, it's more easily to understand and apply. It becomes a much easier task. And that's the focus that NXP has.

Nitin Dahad (19:45):

That's right. And I have written about at EE Times. We've written about a few companies, including startups in this area who are trying to make that easier so that you don't need that knowledge. So I think that's quite important. Let's come on to cost. How does machine learning increase the cost of the end solution?

Ron Martino (20:02):

Sure. I think it touches on what we just mentioned. If you have a very complex heavy machine learning model or capability, that's going to require much higher computational capability, and the higher the computational capability, the more expensive it's going to be. You can choose to do that on an edge processor, or you can choose to deploy that to a cloud and have it operate. When we try to tune these use cases or these models for a specific use case, you can then become very efficient and then you can leverage the traditional technology scaling and Moore's law to really add hardware acceleration specific for ML. That doesn't take up much silicon area, so it becomes a small cost adder, but a very optimum ability to perform that given work that you want to, whether it's detecting people and identifying who they are locally as one example.

(21:02):

You can do that very efficiently on a microcontroller now that's optimized with very efficient silicon implementation. And then you can make it scalable as well with some of our processors where you can scale to an external higher performance neural net processor, or you work in a complimentary fashion with the cloud. Again, all have a cost to them, and it depends on the complexity of the task, but it can be very efficient to very complex capability that you can roll out.

Nitin Dahad (21:35):

Okay. One of the things, whether it's edge or whether it's cloud, I think one of the things that there's increasing concern about machine learning models is around biases. And I guess if you're constrained, your biases can increase. What do you think the role of industry or and specifically NXP has in helping ensure ethics in devices that use AI or artificial intelligence?

Ron Martino (21:57):

Sure. No, it's a very interesting topic and it's growing in terms of the industry discussion. The simple definition of ethics is moral principles that govern a person's behavior are conducting of an activity. That's one definition. In this context, it's governing the principles for machines to do right and not wrong actions. And there's many examples of this. You don't want a machine to injure or cause a fatality when it's performing its task and it's focused on its task is the most important thing for it to get done and needs to know when to stop and that it's going to cause an injury.

(22:35):

It needs clear transparency of operation, whether it's simple concepts around, I want to know it's listening to me or watching me, but also how is it determining its conclusion to then take an action becomes very important security standards to make sure that the systems are secure and don't have backdoor access or other sensitivities or vulnerabilities in terms of their attack surface, such that somebody can access an AI system and then influence it to do certain things or make certain decisions that may be favorable to the person who is attacking the system.

(23:16):

And then you get into the whole concept of model development and how do you implement AI systems that don't have a preset bias that from a principle base is wrong. So when we look at NXP and what we're doing, we've rolled it out or launched an AI ethics initiative, which underscores our commitment to this ethical development. And within that, we talk about being good. We talk about preserving human-centric AI, which really is around avoiding subordination to or co coercion by an AI system, as well as this transparency, the highest standards of scientific excellence as well as trust in AI systems. And we're working with those in the industry that are working on standards, working on the general principles as well as the compliance environments or capabilities that one can certify a system towards to continue to drive this ongoing initiative of what is the right and wrong in terms of machine capability and how it's used.

Nitin Dahad (24:22):

It's a huge, huge topic, I guess. Yeah, we could talk about lots of different things, but we're move on. The subject of our conversation is edge technology. To conclude, what challenges do you say remain with edge technology and implementing edge technology?

Ron Martino (24:37):

Sure. This is an ongoing activity, it's the age of edge computing, and there are many areas for continued optimization, energy efficiency, and really driving and leveraging energy harvesting concepts and near threshold operation of devices is a continued investment by many in the industry. Security, this need for protecting data and continuing to advance this, is an ongoing activity. Investments in silicon specific signatures and different types of cryptography and ways of performing computing in a protected way such as homomorphic encryption, performing calculations in an encrypted environment and never decrypting it. And then extending that around connectivity, throughput and latency requirements as well as the power consumption and optimization of that will continue to optimize connectivity and bring that into these edge devices in more and more efficient ways.

(25:42):

And then finally, this whole concept of aware and intelligence. We're in a third generation of developing and implementing neural network processors or subsystems that go into our processors, and that is driving improvements in efficiency and scaling. But there is continued research in this area in terms of driving higher levels of efficiency with accelerators and different technologies around spike neural nets as well as we can go very far into quantum AI. But in the near term, clearly we'll see a continued evolution around more traditional accelerators and the integration of those into these scalable processors that NXP is bringing to the market.

Nitin Dahad (26:28):

Edge computing and edge processing is very topic of now, and it has been for a few years. I'm sure we could talk lots and I'm sure we're going to do another episode on this fairly soon. But until then, Ron, thank you very much.

Ron Martino (26:40):

All right. Thank you, Nitin, been a pleasure talking to you.

Nitin Dahad (26:43):

This has been the Smarter World Podcast with me, Nitin Dahad. Thanks for listening and see you next time.