Full transcript

Laurel Ruma: From MIT Technology Review, I'm Laurel Ruma and this is Business Lab. The show that helps business leaders make sense of new technologies coming out of the lab and into the marketplace. Our topic today is edge computing and 5G. With immense amounts of data being created and analyzed on devices and at the edge, network speed must also be a priority to process that data in real time, so users benefit in the moment from insights.

Two words for you, 5G everywhere.

My guest is Nick McKeown, who is the senior vice president and general manager of the network and edge group at Intel Corporation. He's also a professor of electrical engineering and computer science at Stanford University. Nick has founded five networking companies and received more than 25 industry awards, including the 2021 IEEE Alexander Graham Bell Medal.

This episode of Business Lab is produced in association with Intel Corporation.

Welcome, Nick.

Nick McKeown: Thank you. It's great to be here.

Laurel: So, you recently joined Intel in a new role that brought together the network platforms group, the internet of things group, and the connectivity group into one single business unit. How did you merge these groups and then prioritize workflow, culture, as well as innovation?

Nick: That's a great question. As you said from my background, I came to this role both as an entrepreneur from having started a number of networking companies, as well as being a professor at Stanford, but also helping to challenge the networking community over a long period to think more in terms of software, the software that drives the infrastructure. In fact, I always credit MIT Technology Review for the invention of the term software-defined networking, which was used as a term to describe a project that we were doing at Stanford about 15 years ago. And it captures the way in which the networking industry has moved in that time.

So, as I came into this role, I was looking at three businesses that we already have in place. And the network platforms group is really our 5G and our private and public network technology and products. Our internet things group is really an enterprise internet of things—so, things like factory automation, support for the transportation industry—and our connectivity group is really cloud networking. And this is all of the networking that takes place in big cloud data centers. So, in some ways, three very different businesses that work in very different ways. But on the other hand, all having this common thread of networking, networking technology, and things that connect to it and take advantage of that network.

Three very solid businesses that are doing an extremely good job already with senior leaders who have a very deep understanding of the technology and the way in which those businesses are evolving.

So, on the face of it, a relatively simple task to come in with such an established set of leaders, strongly collaborating together already, particularly between our network and our IoT, because what we traditionally think of as mobile network operators, many of the things that they have developed and evolved in the last few years with 5G are now becoming very relevant to the premises of edge customers. People doing factory automation is a good example, retail applications, where there's more analysis that's being done out at the edge. And in some cases where they want the communication technology that we've developed for the mobile operators, 5G, private 5G with new parts of the spectrum that are available. And so there's a lot of commonality between those.

Similarly, between our 5G networking business and our cloud networking business, there's a lot of commonality because the telco industry as a whole is really in a state of flux right now. 5G was the first really software-driven, software-defined technology where the walled gardens of old are crumbling, and as they do so, the telco industry is going through a change. The cloud service providers are now moving in and trying to figure out how they can help, maybe how they can take some of that business for themselves. And so there's a lot of turmoil and new strategic initiatives between them. In terms of the technology that we provide, we love the fact that there is a huge amount of innovation going on. We supply the technology to the mobile operators, to those building the public internet, as well as the cloud service providers. So, as they figure out new business relationships between them, we try to provide them with the agility and the programmability that allows them to morph that business as they figure out the new ways to build it.

And we have strong customer collaboration. Many of the customers that we work with are common across these businesses between networking and an edge. You may have seen that we recently announced a very close co-development of our new infrastructure processing units with Google. Well, those IPUs will be very useful for carrying communication workloads at the edge as well. So, we're partnering with communication service providers. We work very closely with Rakuten. We've announced that recently, and we work very closely with companies like Audi, who are deploying new AI inference on the factory floor in tight collaboration with compute that resides either on that floor or nearby in a co-lo facility. So, communications, processing at the edge, AI inference, all coming together under this common framework.

Laurel: And AI inference is that ability to use computer vision to scan, say, cars coming off the factory floor, parts or pieces, and to see what could be incorrect with them right then and there to fix that problem.

Nick: That's right. So, it turns out to be a very big and interesting application of machine learning where one example could be, if a robotic welder is welding the frame of a car and is doing many, many welds, clearly you need those to be done quickly to be efficient, and you also need them to be done with high quality. And so in the past, it's required a lot of manual intervention and manual checking to make sure that those welds were of sufficient quality. Now, what we can do is not only have a camera that is watching that welder in order to look at the quality of the weld, but in real time, be able to react and fix a weld, or very quickly reject a weld and bring in a human to be able to check and then to fix it if need be.

So, using inference as a way of understanding what a good weld looks like through training, and then through inference very, very quickly identifying that problem. So that would be a typical example. Or it could be a little bit more mundane, a camera that is in a shop that's understanding the movement within a shop in order to be able to understand where to place merchandise, infantry management, things like that.

Laurel: So plenty of opportunity here. What is intelligent edge computing and what are some of those technological advances driving it?

Nick: Roughly speaking, broadly defined, edge computing is taking the technology resources that we've been developing over many years for the computing industry and using them to analyze and process data at the edge, perhaps store data closely, so that it's more private, and so we have more sovereignty over the data. But we’re placing that data and the compute close to each other, where they're generated and consumed at the edge. So that's it, roughly. While it would be tempting to take data that we produce from cameras, etc., at the edge and move it all the way up into the cloud, that's often not the right thing to do. It may take too long. So, it may be a latency constraint where we don't have the tightness of control that we need, or it may be just too expensive. And thirdly, we may worry about if we move it out of the location that it was generated, what's going to become of it from a privacy or a security point of view?

Intel itself is a fairly good example. In our factories, we typically have two networks. We have an IT network, which is a traditional network that you would see in an enterprise. And then we have the operational network, which is used to both control all the machines and monitor everything that's going on. And the data, the operational data, that's traveling over that network is Intel's super-secret sauce. This is our differentiating knowledge about how to do manufacturing. We would never to send that to the cloud. And so we would want to keep that, process it, analyze it.

And that kind of factory automation is typical of anyone who has a modern manufacturing facility. In order to be able to do that close by, you can have higher data rates between whatever it is that's gathering the data— cameras, sensors, etc., and very tight control in terms of low latency back to the actuators. So, if you're moving a robot, determining how a robot arm moves, you may only have one or two milliseconds in which to make a decision. And so you need that proximity, because you couldn't do that if you had to go off to a site that was further away. So if you combine that low latency, that high data rate, and that privacy, then you end up with a solution which is almost self-contained at the edge. Clearly, it communicates to the outside world, but from a computation and data perspective, almost all of that is taking place at the edge. So much so that we are expecting that within a few years, say by 2025, more than three quarters of all the data that's being created will be created out towards the edge instead of in centralized data centers. And that's just because of this huge emergence of this kind of application at the edge.

Laurel: Yeah. That's a nice stat from the Gartner team, really showing the change of data being processed in data centers to actually the edge, wherever that may be: on device, on oil platforms, on factory floors, as you mentioned.

Nick: That's right. Yeah.

Laurel: So you mentioned earlier that partnership with Google, with infrastructure processing units or IPUs. Why are they important in today's cloud data center? What's that differentiator right there that people will start hearing more about an IPU?

Nick: Yeah, that's a great question. The IPU, or the infrastructure processing unit, is really a new class of device that Intel has recently introduced. They often get confused with what people used to call SmartNICs, or smart network interfaces. I'll explain in a moment why they're radically different from that. And in fact, this term SmartNIC is a bit of a misnomer. The IPU is helpful initially for someone who is operating a large data center or a cloud. So, consider a company like Google, they have to have software and hardware that actually implements the cloud. And then they have servers that run their tenants, their customers' software workloads, on top of that cloud. Now, when we look at a data center, what we see are rows and rows of servers. And so we think, oh yeah, of course they're going to run the infrastructure code that operates the cloud, as well as their tenant workload on the same servers. It kind of makes sense, right? That would be the most efficient way to do it.

The problem is that if you do that, you spend a huge amount of time, effort, and resources trying to make sure that the tenant's workload, over which they have no control (they're just renting out the compute), that they don't know what that workload is doing. They're trying to make sure that it doesn't disrupt either the infrastructure itself or other tenants, because they've got to maintain the isolation between the tenants, but also within the infrastructure itself. I mean, it would be terrible if the tenant workload was to actually bring down the infrastructure and bring down the entire cloud, and then of course, nobody gets anything done at all. So, they put a lot of work and effort, and a lot of resources into trying to do that.

What the IPU does is it allows them to run that infrastructure code that operates the data center in a separate and secure and isolated set of CPU cores. And that way they can use all the servers that are the other side of the PCIe bus or whatever the bus is that connects that IPU to the server. They can use all of those, too, for their tenants. And it makes a much simpler model for them and a much more secure and isolated model.

So, this is the primary interest for the cloud service providers. Other cloud service providers are heading in the same direction. My guess is that we will, in fact I'm very convinced of this, look back in five or six years' time, we'll see that there has been a change in the way that the cloud data centers are moved, such that the IPU is the coordinator of traffic that comes in from the outside, determines which CPU or accelerator or memory that it goes to. And then its part of the communication that goes on between those to coordinate it. And so, it's almost like a coordinating device as well to make sure the communication takes place in a secure way, but also extremely fast and with low latency so that it doesn't negatively impact the performance of the cloud.

Laurel: So certainly that's important for the cloud service providers and something maybe most folks won't necessarily see on the front lines, but to know that your data is actually in a more secure environment, that certainly, and isolated is one of those things that would help businesses perhaps choose cloud providers as they go forward, because they want to make sure that as cybersecurity attacks become more frequent, that somehow their data is safe, correct?

Nick: That's right. First of all, having that confidence the cloud service provider's infrastructure is going to stay solid and isn't going to go down, that obviously gives peace of mind to the tenant, because you don't want to be part of a cloud that is constantly going down for security or perhaps it gets attacked from either a tenant workload or from the outside. So that kind of isolation gives you a lot more peace of mind and comfort that that it's not going to happen. Second thing is, if you are running a workload in a cloud, then you obviously want as high a performance as you can get in terms of the networking capacity between the different compute elements that you've rented or leased from the cloud. And the IPU helps construct the microservices that most modern applications are constructed from.

So those microservices are small self-contained pieces of code that are offering a service, a well-described service that could be spread over tens, hundreds or thousands of servers. The IPU help stitch those together with low latency, secure high-bandwidth pipes between those different workloads that make up the overall tenant software application that they've developed. And so the IPU is really helpful to the tenant as well.

Laurel: So getting back to 5G, what do you see the role of 5G in edge computing? What are we going to see more of?

Nick: Well, there's a number of ways in which 5G is going to play out. When we think of 5G for most of us, it's just that 5G logo we see on the top right corner of our phone as it begins to appear. And so, for the end user, for a client with a phone and maybe with a laptop in the near term, we will see that primarily as higher data rates. And so that's the obvious way in which we will see that. Early indications from Korea and from China indicate that when consumers have 5G, they typically increase the amount of data they're downloading per month by about threefold, about 3x. And that's largely because they're getting quicker access to more video material. That video material will be higher quality because now we have higher quality screens on our phones. So, the consumption of data is certainly going up as a consequence of that high data rate.

And so therefore the infrastructure itself provided by the operators, they need to match that by rolling out their 5G networks. And those 5G networks need to be very high capacity. And whether they are a mobile operator of the more traditional form, the national telco operators around the world, and of course they're some of the earliest and are the ones with the greatest need to roll out that infrastructure, but as they do so, there are opportunities that start to emerge because 5G, the data rate, the latency, the control that you have over the 5G network means that we can start using it for applications we would not have previously thought suitable for a cellular technology. So, in other words, things that we would not have done with 2G, 3G, 4G in the past.

For example, that robot arm that I was talking about earlier, if you want to actually control that robot arm in real time, you either need to have a cable, a wire, an ethernet cable that connects to it in order to be able to give you the guarantee that you've got connectivity, the data rate that you need, and the low latency control, or you need to replace it with a wireless link. Now imagine that that robot is moving around. You really don't want a wire trailing around on the floor for other robots to trip over. You'd really like it to be a wireless link. And the problem is that the Wi-Fi hasn't really got there just yet in terms of the quality that you would want. What 5G offers, in particular private 5G, is a much more reliable, much lower latency, much more controlled-by-software experience.

And so now what you can do is have a very high-quality link that is comparable to the wire that you've just replaced it with. And that will actually open up a huge number of new possibilities and new applications. So, if a robot is moving around on the floor at a few miles per hour, you may only have a one or two milliseconds in which to change its direction. You need a high likelihood that you can both observe it, analyze that movement, and then control it from the outside. In order to be able to do that, you need links of the quality the private 5G will provide. So this is where we think that one of the early applications of private 5G at the edge will take place.

Laurel: And you're doing quite a bit of work in your research with 5G and connected edge-to-cloud opportunities here, including something called Project Pronto. What is that? And with Project Pronto in mind, what kind of long-term ideas do you have about programmable forwarding and advancements in 5G itself?

Nick: Yeah. Networking generally, whether it's the public internet, private networks, cloud networking, and mobile networking like 5G, they always used to be very, very distinct. They operated in different ways. They had different standards, different companies produced the equipment. They were essentially walled gardens, or at least they operated in different silos. That really has changed in the last four or five years, as there's a a common understanding that it's all coming together around the idea that the network itself, whether it's the network in my home, the network in a factory, the network in a cloud, is all becoming more software-defined under software control. And as that happens, it gets you to ask a number of questions. First of all, if the network is software controlled, can I modify it and change it to do things that I want to do that I haven't been able to do in the past? In the past, all the functions of networks were really locked down by and determined by standards and equipment manufacturers who had very little incentive to change.

Once it's all based on software, you can start to try out new ideas. And some of the new ideas that people have been looking at are to do with having greater observability to be able to see what the network is doing at a very fine timescale. Observe what it's doing and then when you need to take corrective action to fix it, and the fixing could be things as mundane as a broken link, a broken piece of equipment, but they could actually be a functional incorrectness in the software that is controlling it. If you can actually monitor and see that in real time and provide a closed-loop control at a number of different levels, at the low level for things that have just broken, to a high level for things that are just functional or structural things that are incorrect. Then you can start to have a network that is more autonomous, that is more automatic, that is able to understand what it's doing, and then compare that against your original intent, your original aspirations for that network.

I know this sounds very lofty, and 10 years ago, frankly, it would've been considered absurd and ridiculous that you could even contemplate such a thing. Well, networking technology has moved along a lot in the last few years. Functions in the network like firewalls and load balances and VPNs, things like this, that used to be in fixed function have moved up into software. With mobile infrastructure, 5G is really the first example of a network infrastructure that has moved from fixed function hardware up into software, and now all of the digital signal processing that used to take place on specialized devices takes place in software. The switches, the network interfaces, these new IPUs, they're all moved from being fixed function to being programmable. So, their behavior is defined in software. So now we're in this situation where the entire network is defined in software, programmable from end-to-end, as well as the control plane that controls it from top to bottom.

So now it really is a platform. Once you've got a software platform that you can change its behavior, you can start introducing these previously absurd sounding ideas, these fanciful ideas of automatic, real-time, closed-loop control of an entire network, whether that is inside a cloud or whether it is over the entire country. What we were doing in Project Pronto was to develop a prototype, to show the government, to show the world that it was possible to do this with technology that is available today. And that we would do so with software that was predominantly open source. So, we partnered with the Open Networking Foundation (ONF), and it was funded by DARPA, by the Department of Defense, as a showcase that this was now possible. So ONF developed Aether, an open source, private 4G/5G connected-edge platform that is a cloud-managed all-in-software programmable platform that will allow us to do this.

ONF is doing it. A number of companies are deploying it as experimental in their labs. Universities at Stanford, Cornell, and Princeton are part of developing new research ideas that they can demonstrate. Once it's all in software, it becomes much easier for graduate students and programmers to try out their new ideas on top of this platform. And one of these key ideas is to verify in real time the network is operating according to a specification, formally checking against that specification in real time, as packets fly around in the network. And this has never been done before, and so it's a research project. It will take a while to prove out. But this, I think, is the direction that networks will go. We will no longer think of them as fixed-function entities determined by standards bodies. We will think of them as software platforms, where we program them to do what we need them to do.

Laurel: Yeah. And that certainly is crucial, clearly as companies and technologies evolve and are demanding these next evolutions of ideas, and what is today's research project obviously could be tomorrow's product that people are investing in. So, back to thinking about 5G and how that is evolving at the edge, and if most of the data is now coming from the edge, when we think about securing the edge, what current or emerging technologies will help address those concerns?

Nick: Yeah, first of all, to put this into perspective, we're asking, what are the things that our customers and users are most worried about protecting at the edge? One of the big, newly emerging concerns is the data models that our customers or our customers' customers in many cases have generated, which is based on an understanding of their context. So, it could be a data model where they have trained a model for understanding the particular layout of a factory floor and the movements that take place within it. And they may have developed a secret sauce, which is knowledge of how to automate a particular process for which they've trained a model. And that model becomes extremely valuable. It could have been very expensive to create. It could be tens of millions of dollars to develop that model in the first place, but it's just code, right? At the end, it's a model that's represented just by the model itself.

And so it becomes a very valuable asset of whomever has created that. So this is a new era for AI, for machine learning, and for the world of technology. And they would like to be able to train that model somewhere. It could be at the edge, it could be in the cloud, or anywhere in between. And they want to be able to securely and safely move that valuable model out to the edge where they can then run the analysis in real time with data streaming, for example, off a camera or off a set of sensors. So, in order to be able to do that, a lot of care must be taken in order to move that model, because it's very hard if someone was to be able to get access to it, to tamper with it. It's rather hard to tell whether these models have been tampered with, or whether someone has acquired one or stolen one, then selling that on to others.

So, we've been developing a lot of security products. Intel's SGX and TDX products have been developed specifically with this in mind of protecting models to make sure or that when they're in transit, they can be secured. This sort of data privacy and protection of this asset is going to be very important in the future as these models become more tightly intertwined with the way that we do business.

Laurel: And that security-by-design kind of philosophy really takes precedence, doesn't it?

Nick: Yeah, exactly. That's right. The security of, whether it's the inference models or other private data that companies have shown a concern or a lack of willingness of moving to the cloud, figuring out how you secure that, whether it stays at the edge or whether it's moved to and from the cloud is going to be so important over the next few years.

Laurel: And speaking of the next few years, how do you see edge computing evolving? What are some of those more tangible aspects that we'll start seeing? For example, you go into a supermarket now and you can pay as you go with a handheld device. That's a very common experience here in the United States.

But then with autonomous cars, as you mentioned, factory floors, will we start seeing this effect of real-time processing more and more at the consumer, and then maybe more immediate business level?

Nick: The simple answer is, in terms of the combination of new IoT apps, combined with both public and private 5G, whatever we think that is going to happen, it will shock us that people will come up with applications that we won't think of. And that's because it's the wild pioneering west. And it's wonderful, it's exciting, it's terrifying, it's growing, it's expanding. And it's a very, very healthy area of massive amounts of innovation, entrepreneurship, and competition. And it's just super exciting to watch. Every day, every week, I see a number of different use cases that our customers or their customers have put in place that we would never have thought of.

So, you may have seen things like these smart delivery bots. Frankly, if you told me a few years ago that we would be seeing delivery within towns and cities, where there would be delivery that was taking place through autonomous vehicles that would walk down the sidewalk, climb the stairs and deliver right to someone's door, I would have said, OK, maybe 15 or 20 years. But those are being tested and rolled out right now. We showed an example of that at our Intel Innovation event last week for the Roxo smart bot that we developed in collaboration with FedEx.

This is a good example of something which is in some ways ahead of what people would have predicted, but it’s just the tip of the iceberg for the things that people are doing. In that particular case, it's able to exploit the reliable high data rate of 5G, and then IoT inference applications that are running on the sensors and the actuators for that device in order to understand where it is to make sure that it's safe as it moves around.

But that is just one visible example that we’ll all see. When you go into a warehouse, into a factory, the sort of places that not many of us go; typically, you'll find that the control and automation that takes place because of those sensors, we're seeing the combination of the actuators plus a network based on 5G, is going to create a Cambrian explosion of new ideas that, if we were to try and predict, we would get it wrong, frankly. Our job, as Intel, is always to produce the hardware technology and the programmable technology that allows our customers to do things that we wouldn't have thought of. And that really is the right way to think of us and for us to think of our role. We're creating the software, the hardware, the platforms, that enable them to develop those exciting new applications on top.

Laurel: Yeah. And that is what is so opportunistic about all of these factors coming together at this time, including, as you mentioned earlier, the reemergence of software as a force in networking, right? So how is software coming back into networking? Mostly because I think people think of networking as mostly hardware and perhaps that's not what we need to think of anymore.

Nick: Yeah. When the internet was first defined back in the late 60s and early 70s, there was this saying that instead of having traditional, slow-moving standards bodies, the internet would be defined by a loose consensus in running code. What that tells you is there was an attempt to move away from the rigid, slow-moving standard bodies, to a time where you could actually define functionality in terms of code. So, it was a great idea, but it didn't happen, right? Instead, the internet became bogged down in way too many standards, way too many committers, way too slow-moving. And through the needs of having high performance and the unexpected growth of the internet, a lot of it moved to fixed-function hardware. And that was in part to get the performance that people needed, as well as the low cost and the low power that was needed, particularly for the public internet in the big exchange points.

So, we went through this era in the 90s and 2000s, when that should've been open and simple, fast-moving and agile, to instead being bogged down and ossified, and very slowly moving. And then there were a number of things that started to happen, because it meant that the internet was no longer innovating. And when I say internet, I mean, networking broadly defined, whether it's in our homes, in the cellular networks, in Wi-Fi, in enterprises, in the public internet, as well as inside cloud data centers. And really, it started with two things that happened at the same time. First of all, was the realization that a lot of functions were wrapped up in fixed-function hardware, like mentioned earlier, firewalls, load balances, gateways, even spam-detection devices, etc., could actually be placed in software, where you could scale them out by replication of the software when needed during times of surge, and then be able to change and modify them as you needed. And this was known as network function virtualization, or NFV.

This started in about 2010, and it coincided with this software-defined networking movement, which was really about turning the closed proprietary equipment into software that was running on software platforms. And that's how the big cloud service providers have built their networks ever since. What they do is, instead of using fixed-function devices, they buy silicon, they program it, they run software on top (that they write), and then they control it in a manner that allows them to control the reliability, the security, and the new features they need over time. More recently, the same thing has happened with vRAN or virtual ran, where the 5G infrastructure, the radio access networks, have moved up into software. Intel produces software called FlexRAN that runs on our Xeon processors, that moves 5G into software running on those Xeon processors.

And more recently at the edge, the functions that were being baked into the hardware at the edge, have moved into AI inference models running on our OpenVINO platform, which is an inference software that allows developers to develop models and then use those models very, very efficiently at the edge. There are many more examples of this, and I could probably go on all day. It's one of my favorite topics, but essentially all these things that were being thought of as baked into hardware, or in specialized accelerators, or in custom hardware, are lifted up and out into software. What this means is, it becomes all about what the customer, or the end customer, wants that system to do. It's no longer determined by us. I always like to say, no chip designer ever operated a big network. Why do I say that? Well, if you bake the function into hardware, then the functionality of the entire network was determined by a chip designer. But they've never operated such a network. So how the heck can we expect them to get it right? Of course, they're not going to get it right.

And so, everyone was super frustrated that this functionality was baked in. But those who had to operate networks for a living, couldn't do it in an efficient manner where they could fix it, improve it for themselves. By moving it up into software, the chip designer is now creating a programmable infrastructure, then it moves the definition of its behavior up to those who own and operate networks, or inference devices for a living. So, it becomes a software problem. And that means it can move at a much faster rate and is much more likely to solve problems that that chip designer never even knew existed in the first place. But what's more, those software developers will then create beautiful, new ideas on top of that platform they conceived of, in need of the problems they were trying to solve. And that means it will not only innovate faster, but it'll innovate better, as a consequence.

Laurel: And that's what we all want. Nick, thank you so much for joining us today on the Business Lab.

Nick: It's a pleasure to be here. Nice talking to you.

Laurel: That was Nick McKeown, senior vice president at Intel, who I spoke with from Cambridge, Massachusetts, the home of MIT and MIT Technology Review, overlooking the Charles River.

That's it for this episode of Business Lab. I'm your host, Laurel Ruma. I'm the director of Insights, the custom publishing division of MIT Technology Review. We were founded in 1899 at the Massachusetts Institute of Technology. And you can find us in print, on the web, and at events each year around the world. For more information about us and the show, please check out our website at technologyreview.com.

This show is available wherever you get your podcasts. If you enjoy this episode, we hope you'll take a moment to rate and review us. Business Lab is a product of MIT Technology Review. This episode was produced by Collective Next. Thanks for listening.

This podcast episode was produced by Insights, the custom content arm of MIT Technology Review. It was not written by MIT Technology Review's editorial staff.