Deloitte.

For Cloud Professionals
May 2021

For Cloud Professionals, part of the On Cloud Podcast

David Linthicum, Managing Director, Chief Cloud Strategy Officer, Deloitte Consulting LLP

Title: **How edge and cloud are powering next-generation technologies**

Description: There are many misconceptions about what edge computing is, how it's connected with cloud, and how organizations can leverage edge computing to create new products and services that delight their customers. In this episode, David Linthicum sits down with Deloitte's Myke Miller and Carnegie Mellon Professor Mahadev Satyanarayanan (Satya) to discuss all things edge—how it works with cloud and mobile devices, how it's evolving, architectural challenges, and how organizations across all industries will build an edge-native future that's brighter than we can imagine. Satya's advice to those companies embracing an edge-native future: understand your constraints and architect for today as well as for the future.

Duration: **00:25:01**

Operator:
This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to Deloitte.com/about. Welcome to On Cloud, the podcast for cloud professionals, where we break down the state of cloud computing today and how you can unleash the power of cloud for your enterprise. Now here is your host David Linthicum.

David Linthicum:
Welcome back to the On Cloud Podcast, your one place how to find out how to make cloud computing work for your enterprise. This is an objective discussion with industry thought leaders who provide their own unique perspective around the pragmatic use of cloud-based technology. Today on the show we are joined by Satya and Myke Miller, and he is an experimental computer scientist, an ACM and IEEE fellow and the Carnegie Group Professor of Computer Science at Carnegie Mellon University, and he is also a Deloitte Cloud Institute Fellow now. So, how are you doing, Satya?

Satya:
I'm doing great. Thank you very much for including me.

David Linthicum:
We wouldn't have it any other way. This is going to be a great discussion. Myke Miller is a Director at Deloitte Consulting LLP's Network Technology Leader with experience in leading complex transformations in multiple industries, and welcome, Myke.

Myke Miller:
Thanks, David. Looking forward to it.

David Linthicum:
So, I've got to ask you, Satya. I was an adjunct college professor for about eight years at a local university. What is a day in the life like of a university professor at Carnegie Mellon?

Satya:
Oh. *[Laughter]* It's a wonderful life. It's extremely busy. You have unlimited riches of exciting things to work on. The only thing that is a challenge is finding enough time to do all the things you want to do. We have brilliant students, both at the undergraduate and graduate level. My research team of PhD students and researchers are such a pleasure to work with. So, between teaching, working on research for the next generation of technologies, and publishing and participating in sharing my knowledge with people through events such as this one, it's a very exciting place to be.

David Linthicum:
So, how do you pick topics to work on? We're going to talk about edge-based computing, but, really, it's a target-rich environment as far as where you can put research and innovation and kind of where this technology is evolving. So, what do you look for to pick as really something you're going to focus on for—I guess a short period of time or a long period of time, but something that's going to provide enough research activity for the time being?

Satya:
Very good question, very good question. It's exactly the kind of question my PhD students ask me and even some post-doctoral fellows. What is the right way to pick these topics? I don't think there's an algorithm, and I don't think there's one size fits all, but let me share with you some things that I've found useful. First and foremost, we in the university shouldn't be doing what industry can do very well. So, problems that are very relevant and important to industry, they can solve themselves in many cases. So, we should be looking further out. The challenge is if we look too far out, then what we do may not actually be usable by the time we actually have results.

So, finding the sweet spot—I have found in my personal career somewhere between five- to ten-year lead time is the kind of topic that often has the highest impact because it takes us a few years to do the work. At the time we start, nobody may think it's very important other than us. By the time we are halfway through, interest in it is increasing. And by the time we are done, everyone says it's important. So, in a funny way, that actually applied to edge computing.

David Linthicum:
Yeah, and it's funny. Speaking of edge computing, this has kind of evolved in a very productive direction for people who are deploying this stuff. And one of the things I think early on in the edge-based systems is people looked at it as mutually-exclusive to cloud-based systems. And the reality is people who were in the cloud didn't buy that for a minute—I certainly didn't—because, ultimately, edge-based systems really were dependent on the cloud computing.

In fact, you wouldn't have something defined as an edge unless it was the edge of something, and, so, the architecture kind of evolved over time. And, so, things are blurring across the four technical computing architectures. You've got mobility, sensing devices, edge, cloud, continuous reporting sensing platforms, things like that. So, how are things evolving now, and how is there synergy between these different platforms moving forward? And, Satya, I'm going to go to you for that one.

Satya:
Sure. I think one of the big changes in the last three to four years is people have realized that you need both the cloud and the edge. I think you said it very well just a moment ago.

I think there's a growing recognition that it's not either/or. It's actually both. And picking what you do at the edge, what you do in the cloud, and somehow combining their strengths is really what is going to produce the optimal systems of the future. And I'm very happy to see that that line of thinking is now widely embraced in industry.

David Linthicum:
So, Myke, you see this every day. You're out there doing edge-based engagements, things like that. Is there still confusion among people who are looking at this technology as to where the dividing line is between cloud, mobility, edge-based systems, IoT, sensing devices, things like that?

Myke Miller:
So, David, the short answer is yes. I think there's still a perception that when you say cloud, you're talking about a datacenter that's very far away, very centrally-located, maybe covering acres. And the concept of being able to extend that to the edge is just—it's still something that people are learning to embrace and better understand how that can impact their businesses.

David Linthicum:
Yeah, absolutely. So, Satya, we have this notion of the intelligent edge which is starting to actually take off like a rocket now. And this has evolved over time because I think we've always had the concept that we're going to put intelligence as close to where the information is going to be gathered as possible. Some of the things that I worked on in my career would be a connected motorcycle where you have the ability to make decisions as to what the motorcycle is doing instantaneously onboard the motorcycle with a connection that's going into the backend systems. But you're putting the AI capabilities and ML capabilities—which you can do now, run on a Raspberry Pi or other devices, things like that—as close to the information that's being gathered. So, what's going on with the intelligent edge and what should the commercial enterprises out there be looking for or how this is evolving?

Satya:
So, you are correct that there's huge recognition by many parties that putting capabilities as close to the data as possible is a very powerful and valuable thing to do. So, the desire there is to put it as close to the point of data collection. In many cases this could be a smartphone; this could be a wearable device and so on. It is here that design meets reality.

The challenge is that mobility by its very nature makes demands on the attributes of the devices. It has to be light. It has to be small. It cannot get too hot and burn the user. It has to have decent battery life. Any mobile hardware designer has to wrestle with these challenges, and it makes the addition of these accelerators at the edge a challenge, because what it does is you are adding extreme amounts of processing power but they come with energy demand; they come with heat dissipation, et cetera.

So, the idea of perhaps pushing back on exactly how much compute you put at the mobile device but moving it a little deeper into the infrastructure, and so, we use the term cloudlet to mean a small cloud close by to the mobile device or IoT device. And, so, if I have a very power-hungry, heat-dissipating GPU, I can put it on a cloudlet. It's plugged into the power socket. It doesn't have to be particularly light, and it has air conditioning and so on. So, you are able to get the lightweight and all the other attributes needed for a mobile device along with the compute resources you need to do the kind of AI-centric tasks that were mentioned. So, it is this aspect, the ability to combine both the attributes of mobility and the compute demands of AI, that's really the challenging part in—having a one-hop wireless separation between those two makes all the difference.

David Linthicum:
Yeah, I think that's very profound. So, Satya, back to you. Isn't this really a tiering problem, a partitioning problem, ultimately, that we're dealing with as architects? And, so, in other words, we have the opportunity to put intelligence and to put information processing that exists very close to the data, because we have very powerful small devices that can exist in very hostile environments that you just mentioned. But trying to figure out where they exist becomes kind of the architectural issue to solve, and kind of as I'm building these things, and I'm building a few edge systems, you're ultimately looking at the tradeoffs of storing the information centrally and then storing it on the intelligent edge, or some edge device or IoT device that's closest to the information. But there's so many tradeoffs in doing that. You're kind of darned if you do, darned if you don't. So, how do you approach that problem?

Satya:
Yes, this is a familiar dilemma faced by any architect. The best way I can answer this question is really to use a metaphor. If you're an automobile engineer, long ago, the automotive space was segmented. You have trucks, you have passenger vehicles, you have light trucks, and then even among passenger vehicles you have a wide range. You have family-oriented minivans, you have sedans, and then you have the sports cars, right? So, the way to think about this is you have segmented the space into certain distinct groupings with optimization of specific attributes. And I give the example of automotive, but it applies to aircraft—it applies almost to any industry, that these spaces are usually so large that, rather than having an infinite number of design choices, you usually segment the world into a finite number of choices. And the specifics may vary over time, but those choices are usually on firm ground. The trucks of today are quite different from the trucks of 40 years ago, but they're still trucks. They still occupy the same niche.

So, in that vein, a couple of years ago, some colleagues of mine and I spent some time thinking, if you look at the world of distributed systems and mobile computing and IoT, and you have all these design choices facing us, is there a way to give some structure so that we can think about this more easily? We used actually an interesting metaphor, in fact, using metaphors again. You know, when you go to a high school chemistry or physics class, on the wall is usually the chart of elements, the periodic table. And it's a small data structure, but in a very compact way it encodes a tremendous amount of knowledge about the universe. So, this kind of distilling of properties into structure is valuable.

So, what we came up with is a way to organize the world of computing today in the 21st century into four tiers. The cloud represents one tier, where the economies of scale and the robust and secure mechanisms for safety are present. You have the mobile devices which are optimized to be small, lightweight, et cetera, and in the answer I gave a few minutes ago I spent some time on this. And then edge computing involves introducing a new tier between the mobile devices or IoT devices and the cloud, and their purpose is in fact to alleviate the extreme design pressure that hardware designers feel to support compute-intensive, bandwidth-intensive, latency-sensitive applications right close to the user. And this makes it possible to combine the attributes of mobility and compute-intensiveness that we spoke about a moment ago. And there's also a fourth tier. Most people will recognize this as things like RFID tags, though it's actually a much broader class today. These are battery-less computing devices which harvest energy, and you can leave them in place for 100 years, come back and query them, and they'll respond to them. There's no battery to run them.

So, this is something that we've found to be a very valuable way to organize our thinking. It's a couple of years old, and, so far it has been a very fruitful way to think about evolution. And all that has happened since essentially can be fitted into this way of thinking quite well.

David Linthicum:
So, Myke, I really like the thinking that Satya has, is that we're going to look at this through standardization, not necessarily rigid frameworks, but certainly concepts that we can agree upon in terms of how we're going to divide and architect these various systems. How do you think the businesses out there are going to start moving to more standardization of frameworks, architectural patterns, technology configurations around edge computing as you see these things evolve within the firm?

Myke Miller:

So, I think it's going to happen at a sector-by-sector level. I'll give you two examples. You know, within power and utilities we're seeing edge computing take off around things like smart meters and drones doing monitoring of powerlines. So, the business problem is driving the application of these edge technologies. The second example I can give is around smart factories where it's been—the promise of digital factories has been there for some time, but without having edge compute there to facilitate, the digitization of those factory floors it's just been incredibly hard to deal with the complexity of the model. So, David, to answer your question, it's going to be driven off of business problems.

David Linthicum:
Yeah, absolutely. I think the business problems kind of lead into the technology. And, so, Satya, I have really something that's bearing on my mind, and that's really kind of optimization of these architectures edge business cases, looking at the different ecosystems that are there, looking at the proximity, latency, bandwidth, performance, scalability, everything you have to get in there. In reality, as an architect by trade, and I've been an architect for 35 years, I'm always looking for something that close to 100-percent optimized. It's never going to be 100-percent optimized. But, ultimately, I see a lot of architectures out there that work, but aren't necessarily getting close to being optimized, and, therefore, they're kind of leaving a lot of money and efficiency on the table moving forward. What would your advice be to people who are designing these edge business systems out there in aligning their architectural choices so they do get more close to optimization?

Satya:
Yeah. So, the problem you describe is a very genuine—it's a very real one, and the reason is quite simple. At any given point in time, like right now, there is a certain level of technology. And if you want something built that works today, you have to live with those constraints. A year from now, five years from now, many of those constraints may be different. Perhaps the most cost-effective bandwidth that you can use in a certain setting might change significantly. 5G may be widespread, so that the bandwidth and latency assumptions you can make in certain parts of the architecture could change.

The difficult part is to figure out which portions of the architecture are going to have long life, and the design choices baked in are going to be valid, not yet for a year or two, but for a much longer period of time, and what facets of the architecture are able to accommodate improvements easily. And when I say easily, as you know, companies and individuals who use these architectures become dependent on specific aspects of them. Every single thing that you expose becomes a dependency, right? Every standard that is created reflects technology at a point in time. When CD ROMs were created, their size and their technology represented an optimal point. Today something that big is seen as pretty low-density. You could cram the whole thing into a much tinier form factor, yet CD ROMs and DVDs are the same size because of the need for compatibility.

So, I think the real genius of the best architects is identifying points that are stable and also identifying interfaces between those stable points that are somehow very mutable easily. So—and let me give you a simple example, just to sort of help you make things concrete. I used the term cloudlet to mean a small cloud close by, never specifying exactly what close means, because close is always relative to the demands of the application. For an augmented-reality application where the total end-to-end budget might be as low as 16 milliseconds, I have to have the edge very close by if I'm offloading to it. On the other hand, in terms of distance, physical distance, the speed of light at one millisecond, the speed of light in fiber gives you 200 kilometers. I mean, that's a wide range of cloudlet placement if you can have very good network connectivity.

So, if you design your system more in terms of times rather than physical distance in this case, as technology changes, the placement of the cloudlet may move. Maybe today it has to be in the same room as you are using wi-fi, but perhaps a few years hence it can be located in an aggregated location, because the technology has improved to that point. So, that change in the location of the cloudlet is actually invisible to the application. Either you use it, or the application knows that that change has happened. Those kinds of seamless moves and taking advantage of them to accommodate new technologies and improvements in technologies are the kinds of things that make architectures long-lived.

So, in general some of the lessons that we have learned over 40, 50 years of architecting computer systems include ideas such as on-demand caching, so that no matter where you are in the system provisioning it becomes a matter of data transfer in real time at the point of need, at the point of use, and it's an example of a mechanism that's very simple but very powerful, has many applications. So, building systems using these kinds of building blocks allows you to blur the boundaries in these systems, and I think that's going to be the way in which we accomplish a lot of what you have mentioned in the question.

David Linthicum:
So, Satya, moving forward, where are the innovations going to be coming from in world of edge computing, intelligent edge basically everything we discussed in this podcast? So, if we're going to look forward two or three years, what do you think we're going to be talking about in two or three years?

Satya:
I think in two or three years we will be talking about the exciting new edge-native applications that have been enabled by this new technology. The ability, for example, to have the kind of compute you could previously get only in the cloud, but now available on your mobile device with very low latency—what does that enable? What new applications can arise? So, things that you used to do offline, to be able to do them in real time. Instead of your drone flying out, capturing footage, coming back, uploading, processing, and discovering tomorrow that you need to send the drone out again, because, in fact, there's a key detail that you needed to get a closer look for.

Imagine the drone capturing the data, processing it in real time through edge computing, while it is still flying over the area of interest it decides it needs to drop down to lower altitude to capture more details because it has deemed the area interesting. So, this is the kind of transformation, things that used to be done offline, that used to take multiple rounds of individual steps are now done seamlessly, and brand new applications that leverage low latency, such as augmented reality, in many walks of life—elder care, manufacturing, and so on. I believe that this is where the next frontier is. Edge computing is only infrastructure. It's true value lies in the improvement in productivity, and also possibly the cost savings for much, much more efficient workflows in many different industries.

David Linthicum:
So, let's leave it there. I mean, this—a lot of this stuff we were talking about on the podcast is in the article, "The Edge of Cloud," written by myself, Myke, and Diana, and we're going to put it in the show notes so you can go ahead and read the article yourself. But, ultimately, this is basically the thoughts and the ambitions of Satya and his ability to define the space moving forward. So, I'm looking forward to this space evolving and, Satya, having you back on the

podcast and really kind of updating us as to what's going on. I thought your insights here were really, really, exciting to hear, and, ultimately, are going to lead us in the right path to make edge computing work for you.

So, if you enjoyed this podcast make sure to like and subscribe on iTunes or wherever you get your podcasts. Also don't forget to rate us. Also check out our past episodes including the On Cloud Podcast hosted by Mike Kavis and his show Architecting the Cloud. If you'd like to learn more about Deloitte's cloud capabilities, check out DeloitteCloudPodcast.com, all one word. And if you'd like to contact me directly you can reach me at DLinthicum@Deloitte.com, L-I-N-T-H-I-C-U-M. So, until next time, best of luck with your cloud projects. We'll talk again real soon. You guys stay real safe.

Operator:
Thank you for listening to On Cloud for Cloud Professionals with David Linthicum. Connect with David on Twitter and LinkedIn and visit the Deloitte On Cloud blog at www.deloitte.com/us/deloitte-on-cloud-blog. Be sure to rate and review the show on your favorite podcast app.

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast