Architecting the Cloud, part of the On Cloud Podcast

Mike Kavis, Managing Director, Deloitte Consulting LLP

Title:	**Cloud isn't just another data center; it's a revolution**
Description:	Even as cloud adoption rates soar, some companies are still hesitant to make the leap. Many cite security as a reason, and others just think it's simply another datacenter, not understanding the true transformative power of cloud. In this episode, Mike Kavis and industry thought leader, Naidu Annamaneni, talk about why cloud isn't just another datacenter, but instead is a revolution in how computing is designed, secured, and implemented. They also cover how cloud breaks the constraints of traditional computing and gives companies the security, cost-effectiveness, and freedom they need to design apps that truly meet their needs.
Duration:	**00:28:58**

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 Architecting the Cloud, part of the On Cloud Podcast, where we get real about Cloud Technology what works, what doesn't and why. Now here is your host Mike Kavis.

Mike Kavis:
Hey, everyone, and welcome back to the Architecting the Cloud Podcast, where we get real about cloud technology. We discuss all the hot topics around cloud computing, but most importantly with the people in the field who do the work every day. I'm your host, Mike Kavis, chief cloud architect over at Deloitte. And today I'm joined by an old friend, Naidu Annamaneni, who – I'm not going to tell you how old we were, but when we were working together he was writing printer drivers, which no one does anymore, right? *[Laughter]* So, that dates us a little bit. But he is now a thought leader and expert in cloud and probably in every layer from chips up to the application stack. So, I'm going to let him tell us what he's been up to the last few years, a little bit about his background, and we're going to have an exciting conversation today. So, welcome to the show, Naidu.

Naidu Annamaneni:
Thank you, Mike. It is a pleasure talking to you and all the memories you bring back – old memories from eSilicon So, yeah, a little bit about myself, you know, we worked together in – back in Catalina five years I was doing the systems development; you were on the apps side. No, basically my group was

doing the real time systems where I was the main architect doing the application server and the drivers and you know 25 different databases (Inaudible) right? Since then I moved to the Bay Area to start my own startup and we were too advanced. And then I moved on to eSilicon as a founding architect, took on the increasing responsibilities, and the last ten years I was their global CIO.

And I've done major transformation of eSilicon from everything on-prem, on-prem applications and, yeah, high performance computing to SaaS-based applications and cloud-native applications. So, we became the first company in the world to do end-to-end chip design in the cloud at a lower cost than what it was from our previous engagement. And currently – we sold the company, and currently I am advising various startups as well as some of the companies doing the multicloud cost-effective silicon design solution in the cloud.

Mike Kavis:
Yeah, cool. So, we're going to hit two topics today that I picked out from a couple really good blog posts you wrote. The first one – I'll pull it up here – basically says – and I've argued this for a number of years – is your IT is more secure in the cloud than in your own datacenters. And I'd probably put an asterisk by that: if you do it right. And you laid out a few key points out there. So, let's discuss that. I think the first point you hit on was talking about your current layers of security can move with you to the cloud, so explain what you mean by that.

Naidu Annamaneni:
Yeah, so what it is, is that historically, people are running in the datacenter that they have the firewalls to protect it. And then they have their own, the network layer security as well as the applications, right, and the data security, the various tools. And then your single sign-on with the multifactor authentication and all those. You know, there are cloud-native security tools, so what my argument is that if you architect in such a way that, you know, for example, the simple thing is Palo Alto firewalls, right?

When you're moving to the cloud, that means you don’t have to give up all that, and Palo Alto has virtual firewalls within your, you know, cloud framework so that you still can have that perimeter concept. You know, over time you can get rid of and you know implement a at scale or wherever you are, you are looking at the application-level security. But that's what I mean, so that your stack, entire stack that it is here, you move it so that it gives you the sense of confidence: oh, you are secure here. So, that's – the same thing in our case, the same, right? We have quoted multifactor authentication tools from on-prem. In our case we were using (Inaudible) move to the cloud.

Mike Kavis:
Yeah. One of the challenges, though, is sometimes people take the exact implementation from the datacenter and try to implement that in the cloud, right? So, if you look at the cloud, they have the concept of VPCs, right? And I've seen some companies say, "Well, we have this network segmentation strategy and we want this same exact thing in the cloud." We're like, "You can solve the same requirements in the cloud but leverage the VPC." And I see too often they try to bring all their tools or technologies or methodologies, and cloud is a little more than just a datacenter, right? So, how do you help people make that transition? You know, you still have the same controls and policies to satisfy. Now you have an appliance you can use. But the implementation, there might be a better way to leverage some of the cloud capabilities along with that appliance to solve that problem.

Naidu Annamaneni:
Yeah, yeah. So, that's an excellent point, Mike. Actually, with the VPC concept, right, your flexibilities than almost compared to the, if you're using any of the top network vendors, right? The plumbing is increasingly complex. So, it is insane for anyone to move the network layer, right? *[Laughter]* That is another thing, that concepts may be there, your policies may be there, but you don't import everything as is. So, VPC (Inaudible) your segmentation-wise or even the implementation-wise is, it's much more simpler and basically everything is software driven vs. you know going into the hardware, right? So, that's the flexibility. But to embrace the flexibility of the cloud and flexibility of VPCs to rearchitect, not reinvent but rearchitect.

Mike Kavis:
Right. Yeah, and your next point, and I was glad you brought this up. It says, "The cloud gives you a world-class cybersecurity team," and I totally agree with that because the cloud providers’ cloud security is a core competence, right? If I'm a bank, cloud security is not a core competence, right? So, talk about the world-class capabilities and the teams that these cloud providers bring to the plate when you move to the cloud.

Naidu Annamaneni:
Yeah, so typically you know, Mike, I spent almost 20 years in working with various roles, right, where I had both software stack, you the applications and the hardware guys and the network engineers and security engineers, right? But my ability to hire is very limited because only a couple of guys. Translate that into the cloud guys, right? So, public cloud is one thing, but they have other offerings where in addition to their own usage - so, take for example Google, right? So, they have a billion users, over a billion users on the Gmail. So, to ensure that they have direct teaming in addition to their own, every application, how they code, right? So, how they monitor, 24/7 monitoring, and penetration testing, and ethical hacking.

They themselves try to do that and (Inaudible). None of this is possible for me as a – right? I may do the penetration testing once a year because we have to comply with certain customer requirements and vendor requirements, but these guys do that for a living on a 24/7 basis. And that's the argument that it is. No enterprise can afford to have, where any cloud provider… If you take all three amigos, right? Each one of them has in excess of 2,000 people on their security teams. And even if you are a big company, a Fortune 10 company, you cannot have 2,000 security people, and with all the skills. And there is a skill gap, right? So, you cannot hire – even if you have the budget to hire them, you cannot possibly hire all those people. And their incentive is not to come and work for you but they want to go somewhere where their skills are tested in a more dynamic way.

Mike Kavis:
Yep. And then the last point you made is the cloud gives you centralized control of your data. So, talk about that.

Naidu Annamaneni:
So, what happens is that, so in any enterprise, right, if you are going in and – the enterprises, they may have evolved over the last tens of years, of fifty years or so, the older it is, the bad your data management is because everywhere, in the users' laptop or you know in subgroups and all right? So, the controls are, even if you have controls, it is very hard to manage and make sure that centralized, right, I can confidently say that I have a control on our data. So, when you are moving to the cloud, you cannot have all those bad practices, right? And so, you have only one place where you can have the group

policies and who sees what and all. But the IT administration can have that, and say that, “Okay, rather than simple things, right, a chip design, and every engineer is doing their design work on their own laptops or desktop, you are centralizing everything into one environment where they are coming in and doing their design – data never leaves the cloud. That's what we have done. We have done that after, when I took over several years ago at eSilicon, we had this same problem with different datacenters but we consolidated and made sure that our design happens only in the datacenter – data never leaves. That means the fact that we were doing designs for other silicon companies, our IP is our security, right? And so, the same practices that we are lucky that we enforced so that we didn't have the (Inaudible) problem, but other customers can embrace the same concept and providing the VDI solution to their engineers or end users and the data is managed within the cloud environment, within the governance of the IT, or end users for that matter.

Mike Kavis:
Yeah, I think back when we used to work together. Every application team just kind of had their own environment, their own policies, their own rules. And when you go to the cloud and, like – we'll use the example of AWS; everyone's using S3 – you can enforce rules and policies on S3 that any user who uses S3 is compliant with the same policy. You couldn't do that back in that day when everyone owned their own app. It was up to the hygiene of whoever owned that app, which in our case didn't end well a lot, right? So, yeah, that's a good point. So, we're going to pivot to the next one. Your name of the article was "High-Performance Computing in the Cloud can be Cost-Effective." And we were talking about this before we got on, and the first point here talks about the mindset. So, without stealing your thunder, kind of paint the picture of the problem we're trying to solve here, and we're talking about the semiconductor industry here. And then talk about that mindset and how that needs to change.

Naidu Annamaneni:
So, yeah, that mindset is not necessarily applicable only for semiconductors, but any, CIO everything, right? So, what happens when you are running in the datacenter? Some people have stacks of servers and some people are maybe able to virtualize it and use it, right? But still it's a static environment, especially in the semiconductor chip design. So, everyone has a concept of grid, so there are tools, scheduling tools that enables – so you have 10,000 servers with however many cores so you statically configure them and said that, okay – announce them to the grid engine, saying that, okay, these 1,000 to 10,000 servers are configured in that way and then engineers (Inaudible).

That is static, right? But your cost is – you bought them and now it is there. So, quite often what happens is the semiconductor workloads are not static, their requirements are similar to the retail industry, right, where during the Christmas season or Thanksgiving season you have lots of consumers come into your store, or especially on the online, right, or even in the physical world, you have a mad rush. So, your datacenter (Inaudible) peak demand. In semiconductor world there is something called (Inaudible) typically during the (Inaudible), that is the fabrication. You are getting your design ready to send it to the fabrication. And so, two to three months of timeframe where your workloads will be 4X more compute intensive than in a typical day. So, you have to have the datacenter, the peak demand, expecting that, okay, I'm going to equip my datacenter (Inaudible).

Then the remaining time it is idle, right? So, if you move to the cloud, what happens is that rather than making it as static, your mindset needs to shift, saying that now cloud is elastic. It is, you know, infrastructure as a code, because normally, you are statically configuring all the VMs and all, right? It is not. So, you have the flexible APIs as needed. You develop your architecture such that autoscaling based on engineers submitting the jobs, you'll start the VMs, job run set, and then as soon as the job is finished you shut it down.

Then on top of that, you know, so there are other ways of optimizing it. Quite often what happens is that when the engineers submit the jobs for the discrete environment quite often it's a guesstimate. Okay, I need, Mike, to run my job 16 cores, 512 GB. In theory that may be okay, but in practice that is highly suboptimal. That means the actual cost may be 8 cores and it only requires 256 GB. So, all this data can be data-mined and then say that, okay, apply the AI/ML and say that to only make it as what is actually needed for the job, and also it eliminates the friction that engineers are constantly guessing it, right? When they move from one design to the other design, parameters change. Then you're not longer guessing again, right? So, the system takes care of it. So, that is the mindset. Embrace the elasticity, both, up and down, and also optimize using the other tools.

Mike Kavis:
Yeah, the other opportunity the cloud presents – and back in my startup days, when we first built our application there wasn't a lot of users, right? So, we picked, back then you didn't have 100 different compute instances to choose from. You had about a dozen. But we picked what we thought made sense. And as the customer base grew, we started realizing, hey, for this layer we actually need a more memory-intensive server, and in this layer we need a CPU. And we were able to, in test environments, throw huge loads at our different layers of our architecture and figure out which server type made the most sense. And it was just changing an API call to get it.

That wasn't even feasible in the old world. You had to guess before you built it. You bought servers and that's what you had, and then you were constrained to that choice. And in the case like in my startup, we had no idea. You know, until you've got a huge customer base, you have no idea which server type. So, I think that's a huge advantage in the cloud, the ability to just change server types with an API call. It just opens the world up to things, yeah.

Naidu Annamaneni:
Yeah, infrastructure as code. Excellent point, Mike. You know, the same is true even in the chip design. What happens is that as the designs grow and during the final phase full chip design, one of the things you know, your core may be there, right? So, I had in my own datacenter from 196 to 2 terabyte machines. I can, the one – the 34 cores with 196 gig, all the way 256, 512, 1 terabyte, 2 terabyte wishes. Here I don't have to do anything. It's the API call And, also the (Inaudible) it doesn't matter. So, you, it doesn’t matter; the engineer is figuring it out and requesting things appropriately. So, that's – yeah.

Mike Kavis:
Yeah, that's pretty cool. So, I want to stay on the topic of cost estimation but – or cost optimization, but let's move up the stack a little. And I was listening to the Werner keynote a few weeks ago, right before Christmas I think it was. And the thing I liked most about his keynote – a lot of keynotes are all about what did we just release. His was all about rethinking how you architect in the cloud, which is a topic I love, right? And he was saying fault tolerant is not enough; you need to design for dependability.

And I think what he means there is a lot of people build to autoscale, right? So, like if this server's reaching 80 percent, we're going to fire up three more. That's great. You need to do that, but that's not enough. So, when I look up what dependability is, when he's saying design for dependability, it's it has to be secure. It has to be safe. And when he's talking about safe, it's safety of the end user. So, if the system breaks, it doesn't create bad consequences for the end user, right? It has to be available, reliable, and resilient.

But I would add one more in the cloud, and that's cost-optimized, right, because now you have so many choices. Architects have to really think about the economics of those choices. Like when we were making these choices back in the day, this is the infrastructure we had. We didn't have choices. This is your – we have a constraint and we have to work within the constraint. Now you're kind of unconstrained and you have to make the right architectural choices that are feasible and cost-effective. So, talk about that. You know, I don't think a lot of architects are classically trained to think like that.

Naidu Annamaneni:
No, it's an excellent point. *[Laughter]* You know, my guys were so – I was fortunate to have a team that were incredibly smart. One of the things what we did in the, is that the cost optimization side, in addition to applying the straight AI/ML, that these cloud guys come up with so many variables, right? So, in the case of (Inaudible) instances, you have price on demand. Then you have committed instances, the one year commit , three year commit, longer, right? So, these are all the variables, but you need in addition to the configurations the number of cores, memory, because based on the application, you need to know. In our case, we took that friction away yeah, AI/ML engine, based on all our datamining.

That means you need to have the foresight of datamining saying that, okay, this data is golden and we are going to use this. So, in addition to that, all these variables on the pricing side or cost side needs to be factored into the design. Let's say that you committed let's say 2,000 cores for three year commit which comes at a lower cost, and then another thousand cores for one year commit, than the on-demand instances, right? If the jobs are coming in, we orchestrated our algorithms such that the first thing is that we consume all the 2,000 3 year commit because that's my lowest cost. Then we dip into the thousand cores right? If both of them are exhausted – because this is a grid. (Inaudible) all the queues are coming in. Then our guys go on and rather than waiting for those to be available. And if the project timelines are there, so we'll go ahead and create the on-demand instances and then schedule the jobs. So, these variables need to be factored in, in addition to the configurations, making sure that the overall solution is cost-effective.

Mike Kavis:
Yeah. The other thing is because we build systems that autoscale, we have to protect against logic that may make us autoscale in a faulty manner. So, let's just say there's a bug that creates for some reason – or a DDoS attack or something that creates a flood of transaction requests that shouldn't happen. If you don't build some kind of regulator in there, your bill could go – I mean, you could just be running all this compute, and your monitors are going to say everything's healthy, right? So, it's just – it's a different mindset. You've really got to consider things. We were so constrained back in the day. You know, we complained about it but it protected us in some extent. So, trust these things!

Naidu Annamaneni:
Yeah, that is an excellent point, Mike. Again, that is called governance, right? So, governance is something that, cloud gives you the enormous freedom. It doesn't mean that you want to go and bankrupt your company, right? Because the minute you are not there, (Inaudible) So, what we had is we had several controls. You know, the controls have been – the centralizing controls are managed by my team, the IT teams, right? So, where do you have that global level, what was more core count? We didn't put it to be unlimited, saying that, okay, this is our peak is less than our steady state demand is about 10,000 cores.

And we configurated such that our max limit on the cloud side is about 30,000 cores so that you have the 20,000 variability. And then at a project level also we had limits. That is being controlled by the design managers so that no project is monopolizing it. So, you have these safeguards built in so that your bill will never go up, and then if there is a true demand we'll work with the cloud providers to increase that upper limit from 30,000 to 50,000. But you still have the project level set safeguards. In addition to that, we are in multi locations and geographies, right? We had one in Iowa, Singapore, and Europe, right? So, you have those limits also come in. So, you need to layer these levels so you are not bankrupting the company with any DDoS attack or any other attacks or general bug in the code.

Mike Kavis:
So, last question, and this one just popped into my mind out of the blue. We didn't really talk about it beforehand but I want to talk a little bit about IoT and edge processing, because when I think about it now, you have all these devices out in the – I'll call it out in the wild, whether they're in the ag-tech field or in a retail store shelf or on a manufacturer line, and there's very little compute and storage on it. And this reminds me of the work we were doing back in the day where we had an old little OS2 box with very little capacity. And on your side of the farm, you guys were trying to squeeze blood out of a turnip, and on our side we had these huge boxes of Windows, and unfortunately a lot of people didn't design real effectively there.

And I think we've seen that with edge, too, right? You really have to design differently and you really have to figure out what am I going to do on the edge versus what am I going to bring back and do here? So, any – I don't really have a question there, other than as you look out on the edge, what are some of the parallels you've seen from the work we did in retail with distributed computing? And what are some of the different approaches an architect's going to have to take when designing on the edge versus in the cloud?

Naidu Annamaneni:
No, you bring back the memories – oh, my God. *[Laughter]*

Mike Kavis:
Hopefully some of them are good, too, but –

Naidu Annamaneni:
Well, so, if you remember, that communication, in addition to the device drivers and, we had a large team. But I personally wrote that software communicating between our datacenters to the store. That piece of code was written by me. You know, so we had to be tolerant. Remember the way we used to call 22,000 satellites? We were operating remotely because see, everything was remote, nothing, right? We were able to, communication, reboot, and all that logic in addition to the core problem, right? Same thing on the edge side. So, one of the things that you have to be – this ties back to the cost also, right?

You know, so the cloud – if you are going to communicate a lot of datasets to yours, again your egress charges, every cloud provider is going to be charging a ton of money, right? Your egress may be – that bill may exceed your compute. So, what it is, is that you – and most of the people are building the analytics onto the edge, where edge devices are getting smarter and doing the processing. You process this and whatever is needed, the summary results, send back to the main, right? So, that is the paradigm shift, which is exactly what we were doing 20-plus years ago. We used to process all the things, whatever it is. Then we used to send you the summaries and –

Mike Kavis:
Through dial-up, right? *[Laughter]*

Naidu Annamaneni:
Yep, dial-up modems. And now you have the broadband and also with 5G, that is, you know, your bandwidth is unlimited, but you still need to make sure that your cloud provider egress charges. Just because you have the bandwidth you don't send it. And not in the sending it, but the communication, what that the cloud providers are sending a lot of data to you, that's going to cost you.

Mike Kavis:
Yeah, and the big difference now is everything – compute's cheaper, storage is cheaper, bandwidth is bigger. You can do these things in real time or near real time, where we were turning decisions around in 24 hours, so it's pretty exciting. And then the other thing is one of the teams I worked on was figuring out the loyalty marketing algorithms. Now it's just an API. The providers provide, like, what do you buy next? They have an API for that. So, this whole team and all these servers we used to have that used to figure all this stuff out, I don't even have to do that anymore. At the end of the day it's all the same. It's just it's evolved to bigger, faster, cheaper. And at the end of the day it's the same architectural constraints, right?

Naidu Annamaneni:
Yeah, yeah. No, it is cheaper, and also, like I said, this is the elasticity and also the – it's enabling – the cloud is enabling so many – your cost of entry, right, is so low. And if you want to spin out a startup, like you know right now, that you – I'm working on a digital health startup. It wouldn't be possible 20 years ago. Now we can do that because without any cost we can have the cloud platform, right? So, that's the ability of the cloud, flexibility and low cost, low barrier to entry.

Mike Kavis:
Yeah. When we were pitching at the Amazon startup in 2010, the name of our pitch was the $86.00 Demo. So, we took a couple guys you know from where we worked and we built a point of sale on the cloud – not so much point of sale, but a redemption engine in the cloud connected to the point of sale, and our bill was 86 bucks.

Naidu Annamaneni:
Yeah. *[Laughter]* Yeah, that's the –

Mike Kavis:
And we had no money, so – but if it wasn't for the cloud, we wouldn't have even tried that startup. Step one is buy a datacenter. We're already done, right? We had no money. So, it's pretty – the power of the cloud is pretty incredible. So, that's all the time we got today. Great to catch up – haven't talked to you in a few years. Where can we find stuff like your blog posts? Are you on Twitter, LinkedIn or stuff? Where can the listeners kind of see all your content?

Naidu Annamaneni:
Yeah, so LinkedIn – I've been organizing most of the various articles that were published around my cloud work and the blogs. And I am also working on this governance blog. So, LinkedIn is my preferred way of communicating and sharing my knowledge and thought leadership with the rest of the world.

Mike Kavis:
Cool. Well, great talking to you again.

Naidu Annamaneni:
Thank you, Mike.

Mike Kavis:
So, that's it for today's episode of Architecting the Cloud. To learn more about Deloitte or read today's show notes, head over to www.DeloitteCloudPodcast.com where you can find more podcasts by me and my colleague David Linthicum just by searching for Deloitte On Cloud Podcast on iTunes or wherever you get your podcasts. Again, I'm Mike Kavis, your host. If you want to contact me directly I'm at MKavis@Deloitte.com and always @MadGreek65 on Twitter. Thanks for listening. We'll see you next time on Architecting the Cloud.

Operator:
Thank you for listening to Architecting the Cloud, part of the On Cloud Podcast with Mike Kavis. Connect with Mike on Twitter, 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

About Deloitte

As used in this podcast, “Deloitte” means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.
Please see www.deloitte.com/about to learn more about our global network of member firms.