The Deloitte On Cloud Podcast

Gary Arora, Chief Architect of Cloud and AI Solutions at Deloitte

Title: **Mind the gaps: Red Hat's Kevin Tunks on bridging legacy systems gaps to accelerate AI adoption**

Description: In this episode, Gary Arora and Red Hat's Kevin Tunks discuss recent shifts in enterprise technology, from AI's rapid growth to virtualization disruption and the return of on-premise workloads. Kevin highlights common architectural gaps and the role of open-source AI tools. He also discusses key steps to get legacy systems cloud and AI ready. Finally, Kevin gives practical advice for incrementally integrating AI and ensuring consistency across environments.

Duration: **00:26:56**

Gary Arora: Welcome back to On Cloud. My name is Gary Arora. I'm a chief architect for cloud and AI solutions at Deloitte. Something big is shifting in enterprise technology. AI is accelerating faster than most architectures can keep up with. Leaders everywhere are asking that one big question. Are we actually ready for what's coming? Today, we are talking about what is changing, the hidden architecture gaps, the challenges of creating consistency across environments, and what it actually takes to bring AI into legacy systems without breaking everything. To help us unpack it all, we have Kevin Tunks, chief architect and national technology advisor at Red Hat, someone who guides some of the largest enterprises through these exact challenges every day. Kevin, thank you so much for joining the show.

Kevin Tunks: Gary, thank you so much for having me. I'm excited to be here.

Gary Arora: Likewise. Let's start right at the top. We are at an unusual moment right now. AI is here. Most architectures are still legacy, unable to support it. Virtualization markets are being disrupted. Cloud strategy is being rewritten in real time. We are seeing a lot of workloads actually move back to on-prem from cloud. From your seat advising these large enterprises, what's the fundamental shift happening right now that technology leaders just cannot afford to misread? Why does this moment matter more than, let's say, the last decade of cloud transformation that we have had?

Kevin Tunks:
It's a great setup, and it's a big question. The big things that we're hearing about this inflection moment are there's this move to AI. AI has kind of moved from this point of interest in science fiction and some pretty good predictive models to generative models suddenly kind of shifted what that pace of capability is, what the realities are of making that part of your business.

Business models are being reconsidered and how we look at technology and hosting and how we deliver services are all over the place is really, really changing very, very quickly. That said, an awful lot of customers have not really gotten to the point where they see their specific workload, the specific use case that's going to change things for them. There's a lot of experimentation that's happening, a lot of innovation that's happening around new AI offerings, new capabilities, how much you have to build yourself versus how much you can buy in the marketplace. All of that fluctuation is creating this, "Ooh, am I ready to buy in or is it a time to kind of hold and prepare?" That's one space that we have a lot of conversations around.

The other space that we have a lot of conversations around is the big driver of enterprise change right now is really the virtualization disruption in the market. The legacy provider that was almost ubiquitous virtualization technology in the space changed hands and changed prices. That's really forced people to rethink, how do I want to move my data center? What's my true enterprise capability going to look like? How do I use this moment to either just do a virtualization to virtualization, a V to V kind of migration? Is that my only problem? Or am I solving a V to V problem? Also preparing myself for the

future with AI and other technologies coming at work. Those two drivers, how much do I want to just solve my immediate problem, but how much do I prepare for the future are kind of the two big drivers that we're excited about and seeing come to bear.

Gary Arora: I want to take it one level deeper. When you're in the trenches with the CIOs and the architects, you get a very different view of what their systems can actually handle. A lot of the leaders believe they are cloud ready until they actually try to run AI or modern workloads at scale. Where do you see the biggest gap between what enterprises think their architectures can support versus what it can actually handle?

Kevin Tunks: The big ones that we see are this idea of AI as a bolt-on. AI is just a sort of plug-in kind of concept, as opposed to really having it be infused all the way into your processes. For customers who are just purely experimenting, AI is a bit of a bolt-on. It's a sandbox environment. It's something that's just experimental in terms of what's happening. It's those customers that are then saying, "Ooh, I've got something that I want to take out of the sandbox, out of the science kind of activities, and I want to make this really applied." Now, suddenly governance and scalability and infrastructure architecture, and where am I going to apply this? Am I going to apply this in my data center? Am I going to apply this in an edge environment, or is it going to be hybrid with a cloud piece on that? Those are the questions that people haven't really thought through yet as things come out of the sandboxes.

The customers who are at that point, some very large financial firms, some very big neo clouds, and others that are starting to really think about that governance and security and compliance considerations, that's where those conversations are shifting, not to mention the difference between training initial models and then doing what are called VLLM models. When you're doing the compression around your models, when you're doing the full inferencing and deploying that into production, those are the pieces that senior leadership is still getting their arms wrapped around. What does that mean? What's the cost evaluation of that? How much change needs to take place in order to take advantage of that good idea that came out of the sandbox?

Gary Arora: I love how you framed it. AI as a bolt-on is certainly the mindset that I feel is the elephant in the room. But how do we get out of this? How do we shift from AI as a bolt-on to infusing AI into legacy systems? Because a lot of the leaders, they do want AI everywhere, but their systems were never designed for it. How do you help these enterprise customers to realistically integrate AI into 10 or 20 or 30 year old systems without triggering a multiyear rewrite.

Kevin Tunks: There are some timing issues that come with this. If it's not an "all at once" kind of thing, it does take a change over time. Most of that is really finding individual business processes that have a scope limit to them. Finding if those applications have they moved on a cloud native journey? Are they starting to modernize in whole or in part? Are they API driven? Is there a way to infuse information into those applications, those legacy business or business applications that can change the real tone of what that business outcome is that it does, whether that's accelerating the pace of being able to accomplish something, like I want to process my invoices faster, or I want to detect fraud more effectively, or I want to do X, Y, Z, I want to move ships or goods or to where they need to go faster.

Any of those kinds of use cases tend to be wrapped into business applications that are there. If we can find AI solutions that are specific, measurable, and identifiable that you can plug in and say, hey, I've built this extra piece of knowledge, whether that's an agentic concept or otherwise, and now I can, I have a way to infuse that incrementally into my application. I can do a small thing that is scope bounded, kind of what we did with microservices and microservice architectures, I can start using those Agentic AI concepts and start incrementally adding them into my core applications. That gives us a way to sort of say, I took an action, I saw what the impact was, I can measure it, I can look at it, I can see how it goes, and then I can take another action and another action and another action. With that incremental process, really make a huge transformation, frankly, very quickly, but having good feedback loops along the way.

Gary Arora: You brought up specific agentic-solution-solving measurable problems. There are a lot of open source tooling in this space. Of course, the AI conversation isn't complete without talking about open source. Red Hat does sit at the intersection of open source and enterprise-grade reliability. What do you think is something that's misunderstood about open source AI in enterprise? How do you advise leaders to use open source in places where it speeds up innovation, but also without introducing risk?

Kevin Tunks: There's always some amount of risk in any move forward. You have to be comfortable with the risks that you're taking. When you're in early stage, when you're sandboxing, when you're looking at what are those use cases? How can I define that problem well? How can I apply a specific solution that is then measurable, and I can really look at it as part of my business outcome? Look, a lot of people are going to find ways to minimize their cost when doing that type of activity. Totally get that. If you don't know the outcome, you want it to cost the least amount possible for your MVP for a minimum viable product kind of concept on it. Totally understand that.

However, in that transition out of the sandbox, if you'll take my analogy, into a production environment where you have governance, and compliance, and security requirements, and the ability to operate at large scale in an environment and have competence, and trust and consistency with how you do it, that's when there's this shift between pure upstream, superfast innovation, and an enterprise-supported capability, like what Red Hat brings to the table. For some companies that are saying, "Hey, this is not my first, I've got a series of capabilities that I want to put in AI power. I know that these are going to go into production. I have high confidence that these are going to move all their way into production." Well, then in that case, you really want to retire the complexity of that interaction, that integration as early as you can.

You start with the technologies that you're going to end up finishing running on. For companies that are saying, "Hey, I really don't know if this thing is going to go to production,' you're going to go with more upstream open source for that experimental phase. But as you move to the other side of it, that's really where Red Hat comes in because it is the enterprise standard that people are incredibly comfortable with. We bring that trust and security to it that's when you want to wrap that framework so it kind of doesn't get away from you too early.

Gary Arora: Speaking about enterprise standard, let's talk about another trend in enterprise technology. For better part of the last decade, I helped enterprises move their workloads into cloud, the data into cloud stores. Lately, we have been noticing a little bit of the reverse, where the workloads are

coming back to on-prem, especially now with AI, as more enterprises are exploring, running large language models, or even small language models on-prem, either because of data gravity or closer governance or even cost predictability. The big question we are trying to answer is: What does an enterprise AI platform look like today? From where you sit, are there any patterns you're seeing from customers who want on-prem AI, but still expect cloud-like elasticity or security and automation?

Kevin Tunks: For sure there are. It really does come all the way down to the hardware that you're using. To your point, there's no lack of workloads moving to the hyperscalers. They're still growing at an incredible pace, hands down. However, with some of these capabilities, we're hitting a maturity point where some capabilities are coming back on-prem. Like you said, there's a number of drivers. Data gravity is certainly one of them. Data sovereignty kind of flows into that same piece when you get some of the national boundaries associated with things, the legal constructs of where data actually exists, where things from a legal framework, how liability looks like in prosecuting these different concerns, how much independence certain companies want in certain industries to have those frameworks in place.

We're probably starting to see a more mature, slightly more sophisticated set of differences. People are really putting applications and workloads, not in sort of a one hit wonder motion where it's all going to cloud or all going to be on-prem or all at the edge, but really this nuanced capability of true hybrid or multi-cloud where it's, I want to put the right applications and investments in the right location to service up what they're doing. Those are a lot of the big drivers that I talked to folks about. The piece about how to make that work is when you're doing it on-prem in particular, you're probably going to be as close to the bare metal as possible.

Performance, particularly when doing any kind of training activity or any kind of inferencing activity, you want that to be a highly performant for all the economic reasons of making these things viable. What we're seeing a lot of are solutions that are componentized solutions, meaning they're not HCI, they're not like really, really, really tightly bundled kind of appliance pieces, but they're more flexible than that. You want sort of prebaked architectures to retire out a lot of complexity in that space. Then from there, you want application platforming capabilities, to really drive outcomes, give you that flexibility so that you can constantly swap out the components that you're looking at. If one particular group within your company has a problem space, they have a set of tooling, they want to control over their set of tooling, you can still use a consistent app platform capability that gives you the ability for each one of those different divisions, each one of those different projects to really choose their tools wisely, and have that flexibility, many of which can not only run on-prem, but then also run in the cloud. Those are usually the conversations that we're having at the C-suite level.

Gary Arora: In terms of the next-Generative AI native architecture, we are seeing more of hybrid pattern, where an AI platform might be spread across multiple vendors. It makes sense to use the best of tech choices available and not be locked in. But it does open up the issue of having a consistent experience, not to mention security and governance and the maintenance of all across the multi-public clouds, across edge, across private cloud on-prem, across your virtualized platforms. First of all, in your experience, what does consistency actually mean at an architecture level? Are there any shortcuts or some secrets here that we can implement so we are not starting from scratch every time?

Kevin Tunks: You're right. The number of hybrid experiences, the desire to be able to change where you do things have that flexibility or I want to run at my own edge or have combinations. All those permutations, like, really, really complex and they all end up becoming really economically driven decisions more than they should have to be technology-driven decisions. The conversation around that is what gives you the ability to handle the complexity at each one of those different environments, whether you're on-prem and you're on a mixed set of hardware from different vendors.

If I have something that's consistent across all of those, then I'm really making an economic-based decision. That's exciting. Because that means my teams are more portable. I can deploy them wherever they need to go. That's good for both my internal corporate teams, but also my consulting teams that are coming in to advise. They have a common familiar set of environments that they're working through. I have this ability to really define at different phases of maturity of my application of all training is going to be on-prem because of the data gravity. I want my—from a networking perspective—I don't want anything doing ingress and egress across the cloud lines. That's totally fair.

But for inferencing and execution of my model, maybe I want to have the most distributed set of those capabilities from a security standpoint, from that consistency standpoint that you brought up, having the ability to have the consistency of that platform that you're landing on. In our case, that would be Red Hat OpenShift. But having that consistency as you use it in all those different areas becomes a very, very valuable piece. It's hard to have that kind of platform that is patched and secure and understandable that is also consistent across all those different environments.

Gary Arora: That's something we are noticing that modernization today is becoming an ecosystem sport. Something that no single vendor or product can do alone for the right reasons. What role do partners play in delivering consistent modernization experiences across cloud virtualizations and AI workloads.

Kevin Tunks:
I use the analogy a lot of any modernization, anything that you're doing, it's kind of like going in for surgery when you're, like a knee replacement or an upgrade or a shoulder, hip or shoulder or something like that. You want to try to do the least amount of damage possible so that the customer's really getting the most out of what their specific investment is trying to do. You want to minimize that recovery gap to get across to the other side. That means do the least amount of damage possible. From a vendor standpoint, this multi-vendor kind of concept, you've got networking vendors and storage vendors and data protection vendors and observability vendors, and this very complex group of ecosystem that comes together for these solutions. Your data has to live somewhere.

Very few organizations are completely homogeneous when it comes to them. They're going to have different storage providers, they're going to have different hardware providers, they're going to have all these different things. With all that being said, if you have something that glues all of those together, that becomes a really important piece. It's kind of like having a point guard. If you go with my second analogy and the same answer. It's like having a point guard. They pass the ball. They're able to get the best out of all the different players on the field.

That's where we want, that's where we play a big role in that. But bringing the best of storage, bringing the best of compute, bringing the best of the chip sets and the accelerators, wherever they're from. Having that investment so you can take advantage of all those different innovations because they're all innovating at just a phenomenal pace but not being too far behind with any one of them. That's what the architectural gaps start to look like and where you're looking at, "Okay, I get that there's complexity, I get that I want to do this, but this also gives you that adaptability over time so that you can take advantage of and go well, I want more of this in my portfolio, I want more of that in my portfolio and change and adapt over time."

Gary Arora: Is there a balance one needs to strike here with multiple vendors coming into the ecosystem? Is there something like too many vendors? Because we do see that many enterprises want more vendor choice, but their architectures are usually pushing them towards a single vendor because a single vendor is now offering multiple capabilities, and they are launching new capabilities. If you take some of the large data platforms, for example, and then you have separate vendors, which when you combine, you get one holistic solution with multiple capabilities. Is there a decision tree that supports one or the other? What are you seeing from your vantage point?

Kevin Tunks: Again, I'd go back to this concept of early innovation activity, sandboxing activities, is the easiest point with which you can control variables and you can have sort of single-vendor solutions to prove out a concept. But once you get into any scale out of that concept, almost inevitably, you're going to want to be able to evolve it over time. Early on, minimizing variability is a great place to go as you start to go into more of a production environment. Just know that plan for the fact that you're going to have more people there than if you practice on your own, but you ultimately go. You play games as a team because it's a team sport. It does emphasize one of the big points of intersection and all the stuff.

Systems integrators and global systems integrators play an incredibly important role in helping navigate through where people are having success. Groups like systems integrators, and the teams that they create amongst in that environment really do become this powerful spectrum of balancing at what point in time do I take on the added complexity of having multiple vendors? It's, like I said, it's usually when you hit that tipping point of, we're really going to put this into production. Now, it's got to work inside of the entire ecosystem of my data center and my legacy hardware and my chip sets and the new things that I'm buying along the way. It's got to orchestrate amongst all of that.

Gary Arora:
Different muscle and rigor needed for proving our concepts than for scaling out for production workloads?

Kevin Tunks: That's right.

Gary Arora: Let's wrap up on an optimistic future. If you look a couple of years out, and in technology terms, a couple of years can be decades, but let's say you look two or three years out. Beyond today's hybrid cloud shakeup or virtualization shakeup with the early waves of AI integration that we are in, what's the next major architectural breakthrough you think enterprises should prepare for? Is there anything that excites you most about where this is headed?

Kevin Tunks: There's two things that I'll kind of put out there. It's maybe just outside of your two to three year, but not much outside of it. It's worth putting on the long-range radar screen. Quantum compute is going to have a phenomenal impact on certain industries in particular. The industries like manufacturing, health care, any kind of like large transportation kind of planning activities. Those are going to be and there are a whole bunch of public sector activities that's really going to have a big impact on. I mentioned that one more because it creates a real clarity around why we want to make sure we're getting our arms around AI. We're getting our arms around application platforming in that space now.

The amount of data that a quantum system puts out is really not something that humans and human-level processes without AI assistants are going to be able to take advantage of. That's one that is right on that like 2029. It's really like three years at this point in time. That is a piece that we've got to start thinking about: How do we rigorously handle workplace preparation, business process design activities, IT infrastructures to be able to support the virtualized legacy applications, the containerized modern applications for new experiences, the AI infusion into all of that, so that we're ready to take advantage of those next capabilities?

Those are the ones that probably excite me the most. Certainly, that there's all the things that are happening with robotics and more independent robotics that will have a really interesting space on that. I'm still very bullish that someday we will have more autonomous driving vehicles, which I would strangely put into the robotics field. Basically, you're just riding inside of a robot. But some of that technology is going down really, really, really interesting places.

Gary Arora: Really interesting space. We're already seeing autonomous driving in multiple cities. A lot of cool new infusions with AI and quantum. Well, that's all folks. This has been an incredibly grounding conversation. Thank you so much, Kevin, for helping cut through the noise and show what's actually changing in this new wave of platform decisions. If you found this episode helpful, leave us a review and check out our other episodes on On Cloud for more conversations on how AI and emerging tech are reshaping the enterprise. I'm Gary Arora. Thank you for joining us, and we'll see you on the next episode of On Cloud.

Operator:

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