**Steve McDowell:** I think we're all getting AI fatigue. Every product briefing I go to, if it doesn't have the word AI in it, I'm in the wrong room. That's all anybody wants to, whether it's relevant or not. And I think a lot of these technology transitions, whether it's the internet, whether it's the smartphone, whether it's the PC, going way back, there's a brief burst of time where we're very excited about the technology pieces, but the value of technology comes in how I use it. So have we hit a wall on AI? No, but I think it's time to change the conversation. Let's stop talking about GPUs. We'll stop talking about who the providers are, and let's start talking about what are we going to do with it and how's it going to change my business? Because that's the interesting conversation. The way I think about AI and we talk about AI, now I'm going to talk about generative AI, but I'll just call it AI. What's driving the current moment? I mean, we've had AI for a decade, kind of the modern form of AI for a decade. We use it. We use it for predictive analytics, image recognition, retail, whatever. But what's really impacting what's about to impact enterprise is generative AI.

We're seeing a couple of things, and I think 2023 was really the year we figured out how to make these models. And right now we're going through a kind of rapid phase of how do we make it safe for enterprises? And I look forward over the next 18, 24, even 36 months, and it's really how do we deploy that in the enterprise from an IT perspective? That means a couple of things. One, I need to pick a partner who's going to be my generative AI provider. I have a core set of functionality. There's only a handful of companies in the world that could train these models. It's OpenAI, it's meta with Llama, it's Anthropic, just a handful of these companies. And the way it deploys in the enterprise is I take this large language model and I fine tune it with my own data.

So if I'm an IT guy right now, I'm going down the path of how do I, I'm just trying to deploy. I'm trying to enable my enterprise to use it. And then when it comes to using it, there's really two pieces. There's the piece that's very business focused. How am I going to use AI to enable the next or the next iteration of digital transformation? It's going to change all our lives, but it's also, how do I use it to make my own IT operations more efficient? I've been to 15 conferences this year, and the theme for 2024 is AIOps, right? Even Nutanix is announcing capabilities around AIOps.

[Related: Pivot Past the Enterprise AI and Cloud Native Hype]

Do I trust it? How do I trust it, how do I deploy it? Where do I use it? A lot of decisions are happening for the poor enterprise IT architect. So what's happening with ai, we're trying to figure it out. We're at the phase where we've invented the technology right now, we're in the enablement phase, then we're going to start to use it, and then it's going to drive the transformation. And I look at every kind of big technology, every transformational technology kind of follows this path. Although the timelines are getting much shorter, right? It's understand it, it's enable it, it's play with it, and then it's going to drive change.

It doesn't have to be expensive to get into AI as a user and consumer of the technology. And I think where it's disrupting the tech industry and driving and forcing a lot of this conversation about the technology is it's very expensive. It's very complex. It doesn't look like anything I've touched before as an IT guy, or it looks like scientific computing. The cloud guys are solving this for me. They're managing the infrastructure. They're buying these expensive GPUs. They're amortizing the cost over multiple users. Things that I don't have a budget or capability to do as an enterprise, where that's causing disruption in the industry is, well, if AI is driving the industry right now and cloud is taking all those dollars, and my company's name is Dell or HPE or Lenovo, where am I getting my revenue? I mean, the server market was already down. And if those dollars now are being prioritized to cloud, that creates a real dilemma.

[Related: Cloud Vendor Shakeup Puts Focus on IT Resilience]

If I'm looking at this as an IT practitioner saying, where's the value of AI to my enterprise? I don't care about hardware. That's why I like GPT-in-a-Box that the Nutanix is delivering because that's a software set of capabilities. I can deploy that if that makes sense, right? I can deploy that at the edge, if that makes sense, or I can roll that out in the cloud, if that makes sense. So I think a lot of tech companies are trying to prove their relevance around AI. And I'm not saying they're irrelevant. It's just causing a lot of disruption. It's going to change the way that we think about infrastructure.

There are two pieces of AI. There's training and there's inference, right? As a business user, the value is on the inference. An inference is when I take an AI model and I throw some of my own data against it, and it gives me back results, the speed, the time to value, the time to decision is the AI. I say, the closer to your data that the AI is, the faster I'm going to get time to value. And where we look at where, and this is not even a generative AI thing so much when we talk about AI, the most prevalent use of AI is image processing, whether it's for manufacturing, whether it's for retail, whether it's for automotive. If you have a car made in the last 10 years, you have so many sensors in your car, it's not efficient to take the cameras in your car or the cameras at the seven 11, send those up to the cloud to be processed and send them back down.

[Related: Living Workflows of AI at the Edge]

If I'm running a retail establishment, I might not have internet. It may go down, there may be a storm. I can't shut my business down when I lose the internet. So by moving those inference functions to the edge, I get all the value of that AI where it makes the most sense. And there's a couple of things that came together to make this the moment in time where that happens. One is we've been talking about the value 5G is going to bring to the world. I get all of this high bandwidth wireless everywhere. We rolled that out without really a killer application. And then we started all this generative AI stuff, and we got really good at AI. And now part of what that did was all the focus on high-end ai, kind of the natural curve of technology is all of the AI inference capabilities required for mid-range, low level ai, dirt

sheet, dirt sheet, right? The camera you recorded me on right now can probably track my face. It's doing inference and facial recognition on a $30 processor. It's AI at the edge, it's practical, and it brings true business value. Now, where it gets complicated when we talk about edge, edge is anything outside of the data center where there's no IT guy, let's call it that.

[Related: Building a GenAI App to Improve Customer Support]

It's a couple of edge segments. There's robo, remote office branch office. To me, those are extensions of your data center. They're well controlled environments where it becomes really interesting or kind of mass deployments, whether it's a convenience store, whether it's factories, whether it's smart cities, and I have sensors on the lights for traffic control. Managing all of those becomes a challenge. So I'm deploying AI at the edge for these kinds of use cases, but I also have to manage security patches, updates. I got to push new models down. Sometimes I got to bring data back. So what's the framework look like for that? Well, guess what it, it's not that different from how I'm doing hybrid multicloud. The boxes are just a lot smaller. So we did this paper and it's like we're driving AI to the edge for all the reasons I just said. But then it was really more about, okay, we're pushing AI to the edge. How do we manage it and how do we make it efficient to manage? And that requires, again, it looks a lot like multicloud, but it may be a more unconstrained environment.