Transcript:

Rajiv Ramaswami: I am optimistic about technology and its role in the economy because regardless of where I think the economy is, I think companies are going to invest in technology. They're going to invest in software. If you look at the rates, I mean whatever GDP growth, I think tech spending will be growing faster than that. And then software spending within tech is probably going to grow faster than that just because of the nature of how companies are going digital. It's a business imperative. Companies struggle with getting good software talent. There isn't enough to go around. AI talent is even harder to get. The set of tools that you need to put together AI applications and get them going to market, that's not easy either. And then finally, on top of that, there is a shortage of hardware – GPUs – today in the market. Even if you want to do it, you've got everything to do it, you can't get a hold of the hardware.

Jason Lopez: You might be wondering, didn't the speaker start by saying he's optimistic about technology? This is the Tech Barometer podcast. I'm Jason Lopez. What follows are comments from Rajiv Ramaswami, CEO of Nutanix. These come from a conversation captured by an audio recorder in the room. So we thought it would be interesting to whittle this down into a short podcast to hear him essentially riffing on the topic of AI in the enterprise, which has become essential for companies trying to empower workers or streamline or automate.

Rajiv Ramaswami: Whether there is automation of business processes, whether it's customer support, whether it's document analysis and search or processing credit card applications or market applications or what may be, or making software developers more productive through co-piloting or enabling more automated capabilities like better fraud detection, et cetera. So these are all big things that can move the needle in a very significant way.

[Related: 9 Predictions for IT in Age of AI]

Jason Lopez: Rajiv sees momentum building, at least on the development side, in the next year or two. There's an expectation for the emergence of the first wave of generative AI applications, building on the ecosystem. But the challenge will be to overcome a few things like GPU storage, cost benefit analysis and getting higher quality results.

Rajiv Ramaswami: If you're, for example, using some sort of a copilot to generate snippets of code automated, for example, how good is the code? Can you trust it? Can you go send it out? Whatever use case may a customer service use case or where you're looking for document search and retrieval. So getting the accuracy to a reasonable level where you feel comfortable. That's another thing that you have to do and you have to keep training until you get to that level before you can actually put it into production and do inferencing.

Jason Lopez: AI implementation isn't inexpensive. The cost of compute clusters for training can be very high, and Rajiv sees companies having to really clarify the ROI and make a business case for AI adoption.

Rajiv Ramaswami: And so I think people are also going to go through that life cycle of, okay, there's a lot of interest in AI, and then they're going to go implement and they're going to say it's expensive to implement and therefore they're going to be, again, looking at business cases like everything else. I think we'll go through that phase as we move forward here with AI, I think there's a lot of potential and to realize the potential, you also have to pay attention to what it's going to take to get it done and how much it's going to cost and make sure you're getting a

benefit. And then you've got to go get it done. You have to go implement it, which means putting together the team of people, figuring out how to get your data in order, how to choose the right large language model, how to train it properly, how to reduce the size of the model to what you need, and then of course, train it and get into production, get the fidelity of the data results and then put it into production.

[Related: Generative AI Propels IT Modernization]

Jason Lopez: And this presumes you have the talent on board to make it happen.

Rajiv Ramaswami: It takes data scientists, it takes AI engineers and machine learning operational engineers, and then it takes good infrastructure people along with the developers who build the app. So it takes all of these skill sets to go bring this to life.

Jason Lopez: The other challenges of AI implementation revolve mainly around data management. And as we've already heard, it requires high fidelity data in the right place for effective algorithm functioning.

Rajiv Ramaswami: Getting the data together in the right place itself is a massive task for many customers. And the second is you have to run the AI algorithms where your data is because data has gravity and you need to protect that data. You don't want to give up your IP when you run a general purpose AI LLM, for example, on your data. So you have to protect that as well. So that's the other set of considerations that emerge when people are running these AI applications. And then we hear about large language models that are in the trillions of parameters and they can do great things, but they're also very expensive. And so the question you have to understand is, what is it that I need from my application? Do I need that big a model? Can it do a smaller model? And I think there's a lot of optimization that has to happen there also.

[Related: Developing in the Age of AI and Multicloud]

Jason Lopez: Nutanix has gone through a process of AI implementation and uses a cloud platform known as GPT in a box. It provides an integrated solution with storage and machine learning toolkits.

Rajiv Ramaswami: So we can help our customers run their AI applications using a platform that's close to being a turnkey platform, and it can be used wherever their data is sent. And so this is a platform that we call GPT-in-a-Box. It's our usual cloud platform. It's what we all use it for, running every other application. In this case, we also include files and object storage with our unified storage because all these applications need a lot of storage for the data. And then we include some commonly needed machine learning and operational toolkits as part of this model, because again, customers don't have the where with all to go integrate everything that's needed to run an AI application. So we try to integrate everything that's under the covers needed to run an AI application so the customers can focus on the application itself, and they can rely on our platform to go run it wherever they'd like to run, wherever the data is present.

[Related: Generative AI Moves Beyond Hype for IT Operations]

Jason Lopez: Twenty-twenty-three was the year of large language models.

Rajiv Ramaswami: Twenty-four, I think, is going to be the year where it becomes more real. People start building applications, especially not consumer applications like chat and writing poems to your friends using ChatGPT, but real enterprise business applications that could be running. So this is probably going to be the year where people start using these for good business use cases. I think that's the case there.

Jason Lopez: Rajiv Ramaswami is the CEO of Nutanix. This is the Tech Barometer podcast. I'm Jason Lopez. Check out another Tech Barometer podcast featuring Rajiv's comments on the Nutanix roadmap in the story why hybrid multi-cloud matters. You can find it and other tech stories at theforecastbynutanix.com. Tech Barometer is a production of The Forecast. Thanks for listening.