Research @ Citi Podcast, Episode 28: Quantum & Chips — NVIDIA Conference Takeaways

Recorded: March 26, 2025

Published: April 2, 2025

Host: Rob Rowe, U.S. Regional Director of Research

Guest: Atif Malik, U.S. Specialty Semiconductors Analyst

Transcript:

Atif Malik (00:00)

Physical AI, the robots and the cars start to adopt AI more meaningful way. So those are the applications that we like, we are more bullish on the use of AI in science and medicine and then in cars and other robots.

Rob Rowe (00:15)

Hi, everyone, welcome to our Research @ Citi podcast. I'm Rob Rowe, U.S. Regional Director of Research, and we have a really exciting topic today. I have Atif Malik, who is one of our senior technology analysts covering hardware. Both of us were at the NVIDIA GTC conference, or the GPU Tech Conference, this prior week and there were a lot of takeaways and it was a very interesting conference in terms of development on AI, development in quantum computing, and other subjects. But before we go on, Atif, thanks for being on the podcast with us today.

Atif Malik (00:52)

My pleasure, Rob.

Rob Rowe (00:54)

How about we start by, maybe you could give us an idea of your chief takeaways from the conference?

Atif Malik (01:01)

Yes, it was great to see you and, like you were talking before this session, that the conference itself has changed from a very technical conference to a forum where you see multiple companies, from Amazon, Google, and Meta to, to many enterprises now trying to find out what's next in AI. And as the NVIDIA banner said in front of the building, “What's next in AI starts here,” with NVIDIA. Fundamentally, I think, NVIDIA is trying to tell us that we are in what they're calling Phase Three of AI. Phase One was a perception AI. So think of Google Assistant, Alexa by Amazon offering, five, six years ago for speech recognition. Then we moved to a period called generative AI two years ago with a focus on ChatGPT, digital markets, content creation. And now we are at the cusp of Phase Three, which is agentic AI. This is the use of agents, coding assistance for software, customer services, patient care, and the next chapter is going to be the physical AI, self-driving cars and robotics.

So I think there's a lot of focus on agentic AI and physical AI at this conference with generative AI more in the rearview mirror. And just stepping back, the big message on the adoption of AI was that the industry is moving from training to inference or reasoning. We have spent the last two, three years building the infrastructure — big clusters, GPUs and compute — to train large language models like ChatGPT, like DeepSeek, and now the focus is to monetize those models into use of inference and reasoning, and that's in front of us.

Rob Rowe (02:59)

Atif, how do they define that? I mean, it's kind of funny to me because I guess I understand what reasoning is in my own head. But how do we define that in AI?

Atif Malik (03:06)

Yeah, it's interesting. The reasoning is defined as once you've trained a model to look for patterns, now that model is still quite black and white. When you ask for it to infer, it can give you many, many answers. What is the weather going to be like tomorrow? It could be sunny, it could give you many choices or options, and that's inference. When you get an answer, sometimes it's called getting a token. That's inference. The reasoning part is where you are thinking, the model is thinking which answer suits you best. If you're looking for— I think Jensen Huang, the founder, gave a very good example of wedding planning. If you are planning a wedding and you're trying to figure out who are you going to sit at a table? Are you going to sit with your mother-in-law and, you know, with your friends or a separate table, the best answer or the reasoning is where the world is moving to. So reasoning is more like thinking with respect to the multiple options you have and just giving you the best outcome.

Rob Rowe (04:14)

Mm. And how far out do they think that is? I did attend one session where they were talking about the challenges of reasoning and that there were other hurdles such as visual hurdles that also have to be attained by AI.

Atif Malik (04:29)

Yeah, it's happening now. The big breakthrough this year on reasoning was the DeepSeek announcement that shook the AI world a couple of months ago, and DeepSeek is a reasoning model. They used a pre trained model called Llama, which is an open model from Meta that did all the training and then they built the reasoning on it and they were able to exceed the inference performance of other models in a big way. So reasoning models are happening now and other companies like Meta and Google, they're all working on reasoning models.

Rob Rowe (05:05)

Interesting. And what were some of the other themes that came out? I know prior to our podcast, you were talking about power.

Atif Malik (05:13)

Yeah, I think power was, it was an interesting topic. You know, you had companies like Vertiv present, and these guys are making the data centers, and then they're kind of planning the data centers in terms of power, in terms of, cooling and all those things, and... It's interesting, I think the message was that we are moving from a data-limited state to a computer power limited situation where when we spoke to companies who are in the business of venting GPUs, and we asked them, why are you building your data centers in New Jersey? They will say, because we have the power availability around that. Obviously, New York is close enough for big demand and financial services point of view, but it's becoming very obvious that power is become the limiter in terms of how many tokens or the revenues or the inference you can do. So Vertiv talked about 108 gigawatts of new data center power required by 2028, that's a very big number. In fact, that number is above Citi's own forecast. So power was a topic that came up a lot in terms of the data center demand and obviously, NVIDIA is constantly trying to reduce its power performance metric on its new hardware products, but it's definitely a limiter right now in terms of meeting the demand.

Rob Rowe (06:36)

And Atif, right now, when we think about one thing about AI that I think everybody is sort of, I would say markets are probably expecting or demanding, is that we start to see results in certain industries. We start to see the benefit of AI showing up in various industries. And are there particular industries right now where we've identified the gains or the efficiencies that AI is producing and what other industries do you think we'll see in the future?

Atif Malik (07:14)

Yeah, it's a good question, Rob. I think that the application of the industries that came up the most in terms of where AI could get more breakthroughs, the first area was in science and medicine. There was a lot of discussion on understanding protein folding. This common mention that the MRI for mammograms, the time to do those could be cut by four because you can recognize the image and make sense of those images a lot faster using AI. So mammograms or pre-screening for cancer, that time could be cut by four. Another area that was mentioned a lot was self-driving cars. You are seeing a lot more obviously level four type cars, Waymo and others on at least in San Francisco where I live. But level five, where the precision and the reliability of the car starts to get close to perfection is not too far away. So, it goes back to what I mentioned earlier in terms of physical AI, the robots and the cars start to adopt AI more meaningful way. So those are the applications that feel like, we are more bullish on the use of AI in science and medicine and then in cars and other robots.

Rob Rowe (08:30)

And aside from the discussions on reasoning, are there particular hurdles that the conference was worried about, say, or still thinking about?

Atif Malik (08:43)

Yeah. I think there is still a big concern on the use of AI to create a deep fake attacks and how do we make AI smart enough to decipher if a video is a fake video versus a real video and countermeasures. And I think Meta spent a lot of time in kind of indicating that they have models that can distinguish a fake video from a real video, but the privacy and the security and that was definitely an area of AI that came up in the discussions.

Rob Rowe (09:17)

And actually, I know both of us attended the quantum computing session, I guess Jensen Huang dedicated Thursday as Quantum Computing Day. And that was quite an impressive session. It ran about two and a half hours, and I think it was standing room only at the San Jose Civic Center, and he interviewed about 15 different firms — some big, some small, some private, some public — on quantum computing. What were your takeaways on that session and why do you think Jensen Huang did it?

Atif Malik (09:51)

Yeah, Rob, I thought that was a very interesting session and also quite controversial given what Jensen had said earlier where at CES in January, he made a statement that quantum computing was 15 years away. I feel like he kind of backtracked that statement by inviting everyone in the quantum computing world. And NVIDIA will not have these companies on the stage if NVIDIA did not believe quantum computing was happening and not too far away and a promise that he's going to have a Quantum Computing Day every year at GTC. So the main takeaways for me was that it doesn't seem like there's a single type of technology that's going to be a winner on quantum computing, but it did appear that superconducting and annealing

methods maybe have an earlier lead. Now, there's a lot of focus on noise or error correction in terms of getting the right qubits out of quantum computing. That remains an area of focus, but there's certain end markets and certain applications like quantum chemicals and cryptography for financial services, which are the early adopters of quantum computing. But the use of classical computing with quantum computing will remain key, at least in the near to midterm. Yeah. It was very animated at times where Jensen was getting defensive on what he had said. But it was very interesting, and to me, quantum computing definitely is happening a lot faster than what people expected a few years back.

Rob Rowe (11:25)

I thought one interesting observation that Jensen Huang made, and he set the narrative, was that in many respects, you might want to think of quantum computing not as a computer, but as an instrument. Because there's a lot of, I guess, a lot of people are thinking of quantum computing is having a computing construct which is more associated with a computer. In other words, storage, data, that whole thing. Whereas in many respects, this is a more problem-solving, solution-oriented instrument that is also can be very much a part of classical computing. In other words, it's not classical computing versus quantum computing. It's more that we're going to see a lot more combinations where both are working together, where AI and classical computing would work together. Do you agree?

Atif Malik (12:18)

Yeah, that's a good point, Rob, and I thought he was helping the quantum computing companies by telling them to set a lower bar by not saying this is a quantum computer because the computer comes a lot of other things like storage, like networking, like software. So I think his message was that you look at it as an instrument or something where classical computing can assist in those areas where— quantum computing is still, there's a lot of challenges. And so I thought he was trying to help companies by saying that you can change the narrative by saying this is more of an apparatus or instrument than a new sort of computer. But it was more in the messaging and that was a good point by Jensen too, to change the messaging from quantum computer to an instrument. Because the end market is still quite selective for quantum computing in terms of where it lends itself to most utility.

Rob Rowe (13:17)

Yeah, and I think a number of the participants on the panels talked about some of the applications that they're already working on, but I do know that he said that next year when they have their Quantum Computing Day, they will expect everyone to come back with the executable projects that they've been working on.

Atif Malik (13:38)

Right. And I think he made a very good point. He said, look, do we care if our Uber car comes in a few seconds earlier or the DoorDash delivery happens a minute earlier? We don't need quantum computing to do all that, but we do need quantum computing to do other tasks like cryptography…

Rob Rowe (13:59)

…and folding proteins.

Atif Malik (14:00)

Folding proteins, yes.

Rob Rowe (14:02)

Well, thank you, Atif. This has been tremendously insightful and I look forward actually going to that conference again next year. It's a great way of learning about all the AI applications. When you go through those exhibit halls, you know, the exhibition halls, there's plenty of small entrepreneurial companies that have all sorts of great inventions that they're coming up with or ways or applications for AI. I thought it was rather fascinating.

Atif Malik (14:31)

Great, Rob, and I look forward to seeing you in the green jacket that I saw you buying at the NVIDIA merchandise store, the green NVIDIA logo jacket from the merchandise store next year. So you can blend in. *[laughs]*

Rob Rowe (14:43)

I actually bought— I bought some t-shirts for my son, which were rather fascinating. I think what's funny about the t-shirts is they had little program— they had little mathematical formulas where it says, let's not get down to this, and I have no idea what that formula means. But anyway, it's fun to wear it. Thanks again, Atif.

Atif Malik (15:01) Thanks, Rob.

Rob Rowe (15:02)

Our next podcast will be on European Views on tariffs with Arnaud Marès and Elise Badoy. This podcast was recorded on March 26th, 2025.

[Disclaimer] (15:14)