Goldman Sachs Communacopia & Technology Conference

Company Participants

- Manuvir Das, Vice President of Enterprise Computing

Other Participants

- Toshiya Hari, Analyst, Goldman Sachs
- Unidentified Participant

Presentation

Toshiya Hari {BIO 6770302 <GO>}

Hey, great. Good afternoon, everyone. Thank you so much for joining us. As expected, standing room only. My name is Toshiya Hari. I cover the semiconductor space at Goldman Sachs. Very pleased and very honored to have Manuvir Das, VP of Enterprise Computing. Manuvir, he leads a team working to democratize AI by bringing full-stack accelerated computing to every enterprise customer. He has more than 25 years of experience in the technology industry and prior to joining NVIDIA in 2019, he held a range of senior roles at both Dell and Microsoft. And at Microsoft. I believe you helped to create the Azure platform, Thank you so much for coming.

Manuvir Das

Yeah. Thank you, Toshiya. It's an honor to be here and thank you to everybody for taking the time. I'm just a small cog in the wheel, but happy to represent NVIDIA here.

Questions And Answers

Q - Toshiya Hari {BIO 6770302 <GO>}

That's awesome. So, Manuvir, before joining NVIDIA in 2019, again, you had a very successful career at both Microsoft and Dell. What initially attracted you to NVIDIA? How has the experience of the company played out so far relative to your original expectations? I think, I know the answer to that question, but I'll ask it anyway. And as the Head of Enterprise Computing, how do you spend your time? What are some of the key priorities for you?

A - Manuvir Das

Yeah, that's three great questions. So let me do them one-by-one. I think the reason I joined NVIDIA, Toshiya, because I grew by Microsoft, the ultimate software platform company. We understood that a platform is only as good as the applications that are developed on it. We had a big focus on developers. In fact, we had a whole thing called the developer division where we focused on developers.

And then I had my first conversation with Jensen who is the CEO of a chip company and all he would talk to me about was developers. And I really didn't get it. And then I watched his keynote at GTC Conference from the previous year. And he as the CEO spending three hours talking about use-case after use-case of accelerated computing that developers can embrace and build applications for, right? And it was an eye opening thing for me and I guess what I realized was at the time was, this company has this vision that there's a new form of computing coming, accelerated computing, but it's not a free thing.

You don't just take applications and move it to accelerated computing. You have to work on it one domain at a time and that means you need developers to embrace it, right? And that's the journey this company had been on really for 25 years already by then. And you could see it then, we were on the cusp of about a million developers who were using the NVIDIA platform, that number has now reached 4 million, right? So it's been remarkable.

So that's the reason I came to NVIDIA because I realized that these people are really seeing the world at the cusp of something, it's not about the chips. It's about the whole stack.

And then to your second question, I would say, what I've experienced differently since I got here was, NVIDIA has gone up a lot because what NVIDIA realized was that at the same time, so you have new technology, right, that is groundbreaking but people adopt in ways that they're familiar with adopting and there's a whole ecosystem in the enterprise for how you adopt technology whether it's hardware manufacturers like a Dell or an HPE or software platforms like VMware or an SAP or service integrators like a Deloitte or an Accenture.

And so we've spent a lot of time in the last few years at NVIDIA really getting that flywheel going. That's how we got to this point now. So I think that's what's been a little different for me.

And then. I think your third question is about what I do?

Q - Toshiya Hari {BIO 6770302 <GO>}

Key goals, priorities. How you prioritize.....

A - Manuvir Das

Yeah. So, I think at NVIDIA, we have a big team of people who do actually all the work. I'm just a talking head. But I think I do two things, right? One is. I work with the leadership team on our strategy, how we actually approach this opportunity via

enterprise, how we think about it. And then the second thing is this ecosystem I talked about, all the announcements you see with VMware with Snowflake, etcetera, so I spend a lot of my time working with this ecosystem so that everybody can win together, the customers, the partners in the ecosystem, and then of course, NVIDIA coming along.

Q - Toshiya Hari {BIO 6770302 <GO>}

That's great. Thanks for that. Jensen talks about the iPhone moment.

A - Manuvir Das

Yeah.

Q - Toshiya Hari {BIO 6770302 <GO>}

Of AI arriving. As sort of outsiders, we were exposed to ChatGPT and the likes.

A - Manuvir Das

Yeah.

Q - Toshiya Hari {BIO 6770302 <GO>}

Maybe late last year, maybe earlier this year depending on who you are, where fo you stand. I'm sure you saw this very earlier given what you do, hopefully. Can you describe what the AHA moment was for you and the broader NVIDIA team as it pertains to this big movement?

A - Manuvir Das

Yeah. I think I'll start by saying, firstly that I think for the world at large, really ChatGPT was the AHA moment, right, and NVIDIA works very closely with OpenAI, has for some time, we're very proud of those folks and really every customer conversation that we have about LLMs and generative AI, the first question we ask them is, have you looked at the tool kit that OpenAI provides. And if that's a great starting point for you, just go with that. Right? And they've done a great body of work, right? So I just want to put that out there.

I think for us, the AHA movement came, I would say probably about five years before that. And probably, yeah, September of 2019 is where we first put out the first version of a library called Megatron that NVIDIA built, that was really the framework for doing this kind of training for large language models. And the way to think about it is, you know, for years we had worked on all these AI use cases which are based on what is called supervised learning.

So basically you teach the model by giving it human-generated examples, right? Here is the photograph. I'm telling you there is a car in it. You know the photograph, I'm telling you that there is a dog in it, right? And so, humans have to go to the process of creating a lot of this data set that is fed into training and that creates a bottleneck and that creates this barrier-to-entry.

And then the AHA was, this advent of unsupervised learning, which is if you think about it, a lot of how people learn to, right? You don't always learn just by sitting in a college and a professor is teaching you. You learn by just absorbing and what's happening with LLMs is, it's unsupervised learning where you put a lot of data in front of it and the model just learns on its own. Right.

So that was really the AHA movement and so we built this framework on the one hand called Megatron to do the training. And then secondly, we realized that you needed different circuitry in the hardware, what we call transformers to really accelerate this form of real learning. And so in a roadmap for creating GPUs, we started pulling that circuitry in, the transformer circuitry. Right. And so that happened years before ChatGPT here.

Q - Toshiya Hari {BIO 6770302 <GO>}

Got it. In terms of customer engagements, being the leader of enterprise computing you must interact with thousands -- tens of thousands of big customers and potential customers who are either already deploying AI or are looking for ways to leverage the technology. Can you give us a feel as to how customer engagements have evolved since the beginning of the year? And what are your customers coming to you for today?

A - Manuvir Das

Yeah. I think I'd start by saying,. I think we have about 40,000 companies working with us on a technology. So, no. I do not actually meet with all of them. We have a bit challenge. But there has been a dramatic shift, Toshiya, and I would put it this way, right? I think in all the years we've been working with companies to date before this year, the customer conversation would always be about, well, what's the use-case that I as the company should care about. I as a customer should care about and we'd be pitching the use-case.

NVIDIA really in many ways made the AI market all these different use cases. We created them vertical by vertical and depending on which industry you're in which shows you why fraud detection, for example, would be a good use-case for you. So a lot of the conversation would be about that.

I would say this year when customers come to see us, now they already know what the use case is, right? It's the intelligent assistant, helping their employee, it's the customer interaction, what have you. And so the conversation is actually about, okay NVIDIA, what do I need to know and how can you help me and who can I work with to implement this, right?

And I think this is a good point to point out that NVIDIA by DNA, we're a platform company, right? We work with the ecosystem. We very rarely produce direct solutions for the customer ourselves. Right, We encourage our ecosystem to do that. So often the conversation is that educate you on the landscape on the technology stack. And then, let's show you who you can work with to implement. Right. And we're underneath everybody.

Q - Toshiya Hari {BIO 6770302 <GO>}

Got it. In terms of the long-term market potential for AI, it's extremely difficult from where we stand in predicting how big this could be. And perhaps it's challenging for you as well. You and Jensen, the broader team did throw out a couple of numbers at the Analyst Day. $300 billion in chips and systems, $150 billion in NVIDIA AI enterprise software, and another $150 billion in Omniverse Enterprise software. When you guys construct something like that, a long-term TAM, how do you go about it? Is it bottoms up? Is it tops down? Is it a bit of both? If you can kind of share that with us, that would be helpful.

A - Manuvir Das

Yeah. I think this is a great topic and maybe if it's okay, I'll take a couple of minutes to unpack it because I think it's very important, right? So firstly, we don't do TAM, right? So when we talk about all these numbers, we're really talking about the long-term market opportunity that we see. And then it'll will play out -- over the years to come as it will play-out. I think if you go back to the basics, the fundamental thing we see going on here, Toshiya, is that there is a new form of computing that is beginning its journey and that's what we call accelerated computing. Right,

And the whole point here is, if you think about the traditional computing systems based on CPU computing, what has changed over the decades is simply the location. You're doing it on-prem, you're doing in the cloud, you're doing it on your phone, but it's essentially the same style of computing. And as the world has evolved, more and more of the function of companies is being done in computing, which means you need more and more computing in the world, which means you need more datacenters, you need more energy, you need more horsepower and it's just not sustainable. It's just not on a sustainable trajectory, right?

And what we saw was, accelerated computing, which is this way where the same amount of footprint can do 10 times the work, 100 times the work, that was going to be the only way. So, the simplest way of thinking about NVIDIA is, we made this big bet. It's been decades in the making, that the way forward is going to be accelerated computing. Okay?

And so, how do we think about our market opportunity at a fundamental level, what we say is the footprint that is out there, there's about a $1 trillion of spend on datacenters, that footprint that already exists which is traditional computing, plus all the growth that is going to happen in the years ahead, that is all going to shift from traditional to accelerated computing and we've set ourselves at NVIDIA to be at the forefront of that shift.

So all our analysis of our market opportunity starts with that, that there's going to be this dramatic secular shift. We believe in it very strongly. Now we have the ultimate killer app for it, if you will, with generative AI and that is the beginning of our market opportunity.

And then the other stuff, the numbers that we did, a lot of which is bottom-up is to ask ourselves, okay, if I break that down, how much of that is systems and hardware, right? And you do a refresh cycle and that's where the $300 billion came from. And then the software opportunity.

The interesting thing for NVIDIA is for a long time, we've invested a lot in software. I would say probably 80% of our R&D over the last decade has been in software, not in hardware, but that software has just come along with the hardware. And the reason for that has been, because a lot of the early shift has been the develop ecosystem, the researchers, the R&D where you need the software, but it's sort of okay if I go through some pain adopting the software if you will, but we are now moving into the world of production.

We are moving into the world where enterprise companies are betting their business on the AI models that are running under the applications and so you need enterprise-grade software. And so the reason we put the other number for the software is because we see incrementally for NVIDIA in the years ahead, we have this big opportunity that we are actually the operating system of AI. We re the run-time of AI.

When you have a model and you take it with you and you run it under all your applications, that model needs to be running at 3 in the morning, right? It needs to be a supported enterprise-grade thing and so NVIDIA is the provider of that run-time, no matter where you are. And so that's where that incremental software opportunity comes from -- for us.

So on the one hand, we take the secular shift. We believe the secular shift is being driven by NVIDIA. So it's a big opportunity for us. And then on the other hand bottom-up, we look at these individual parts of the stack, if you will, to add it up.

Q - Toshiya Hari {BIO 6770302 <GO>}

Got it. So it sounds like you still feel pretty good about those numbers and more about timing.

A - Manuvir Das

I think we do. And I think we realize that it's a generational change. It's a multi-year transition that the industry is going to go through and we're here for the long haul.

Q - Toshiya Hari {BIO 6770302 <GO>}

Yeah. Got it, got it. The CSPs are very sophisticated. They are informed. In your field of enterprise. I'm sure there is a pretty broad range in terms of sophistication. You guys are doing quite a bit to democratize AI. I think you've got frameworks. You've got partnerships with the likes of Snowflake and VMware, Hugging Face, maybe talk about the significance of those partnerships and what you're doing to make it easier for your customers to deploy AI.

A - Manuvir Das

Yeah, this is the thingI work on and think about every day, Toshiya. I think there is a couple of different aspects of the democratizing, right? I think the first thing is we need to really understand the power of large language models or LLMs in democratization because I think we all understand from a used-case point-of-view, the value of this, right, that there is -- we see it every day. We all use ChatGPT, we see generative AI, generating images and all of that every day. But I think that's a very different way of thinking about it, right?

The thing about AI is, it's a new form of computing where basically what you do is, you have a corpus of data, you train a model with the corpus of data, it's learned and now you can ask your question. That sounds great. But the challenge for years has been that first the front-end part of the process has a very high bar to entry because you have to find all the data, the right data, you have to curate it, you have to go through this whole training process before you get a usable model, right?

The beautiful thing about large language models is, the thing called pre-trained foundation model, right, where some smaller set of people have done the work with hundreds of millions of dollars and many-many and servers large amounts of data and trained of these models and now they're ready to use, right, and you start from there, you fine tune with your own data and you use the model. So the whole front end of the process has been done for you. Right?

So the barrier to entry, we -- when we work with customers, we certainly have. You have the AI unicorns that are sitting there in large amount of training and then in every industry, you have a few companies that say, I'm at the forefront of this. I'm going to train my own giant models, right, like Bloomberg is a great example.

But I think for the bread and butter enterprise customer, a great place to start now is, let me pick up one of these pre-trained foundation models, right, whether it's somebody like OpenAI or it's a Llama 2 from Meta or NVIDIA's models or what have you. And so it's the ultimate democratization, because that front end of the process which is so difficult to do, you can now basically just jump ahead, right, and just start with the output of that. And do your fine tuning and your inference and embed it into your applications. So I think at the technology level, that's one kind of demonetization, right, that we work on and that we see.

And then the other one is at the ecosystem level. So for example, we made an announcement with VMware very recently. What was that about? What that was saying is, if you are one of these enterprise companies where you want to jump in at that sort of 75% point where you take a pre-trained model and now you just embed it in your applications, well, what do you need? You need to be able to take model, put in your briefcase and take it wherever your application is. That's what you need, right?

You need this -- VMware called it the Private AI, but basically, how am I empowered so that if I just pick a model, I have the right run-time and I use all the tools right now.

I already have a farm of servers deployed with VMware. My IT team already knows how to manage that, right, how to scale that, all of that, right?

I work with people who do my big transformations and deployments. I have a contract with Deloitte. I have a contract with Accenture, right? I have certain software platforms I use. And then the ultimate level of that democratization is, for example, the work we do with ServiceNow, that we've talked about, right? The ultimate democratization of AI is when we work with an enterprise company, that does not know they're doing AI. That does not know they're working with NVIDIA because they just upgraded to the newer version of ServiceNow that is infused with the AI work that NVIDIA and ServiceNow did together. And from their point of view, they just got a new functionality, right?

So to put it another way, we're a platform company. We believe in the network of networks and for us, the ultimate democratization of AI and where NVIDIA is going is the ecosystem of applications that are powered by NVIDIA's platform and this network of customers that all of these companies individually have, that we don't have a sales team for, but they have their sales teams for, right, and they're just taking NVIDIA technology to all of their customers and that's really the journey that we're on.

Q - Toshiya Hari {BIO 6770302 <GO>}

That's great. DGX Cloud, I had a couple of questions on that. I think you've recently introduced the concept of DGX Cloud. To the levels of the audience, can you describe what DGX Cloud is? How did it come together? How does it work? One of the questions I often get from investors is, do you actually end up competing with your customers? So if you can address that concern, that would be great.

A - Manuvir Das

So, I'll start by saying that we have no intention of competing with the CSPs. In fact, they are our best partners, right, among our best partners and we do a lot of business together and we work with a lot of customers to get them mutually with all of the CSPs, right? I think the simplest way to explain it would be, if you think about our server system that we built called DGX, right, that we've worked on for many years now on-premises, accelerating computing is a new kind of paradigm. Okay? It's not just about chips. It's not just about the systems. It's about the networking, the software, everything that goes into it.

And when we started this journey, the system manufacturers, the Dells, the HPs of the world, they sold a certain kind of system and their customers expected to procure certain kind of system from these manufacturers which was not really accelerated computing. And so, we built our own system to sort of show the way, to be the Scout team, right, and we established DGX with all these use cases and that helped the system manufacturers understand that this new computing model is actually interesting and there's a market here.

And as soon as they realized there was a market, we actually took the secret sauce of DGX. We actually took the internals of it and we productized it into an engineered

solution that we gave to the system manufacturer. And we said, you go sell your own systems now, build-your-own systems with your own IP, sell them. We're actually very happy when you sell one of your systems to the customer, because we are not going to scale this business with DGX. DGX is the Scout team, right, and we're always innovating and putting the next stuff into it.

Now, the way we came about with DGX Cloud was exactly the same thought process, but in the cloud, because we see more and more customers doing the work in the cloud, especially with AI. And so it's the same journey where what we said to the CSPs was, within each of your clouds, how about we mutually create footprints of NVIDIA DGX technology, where again we are putting in footprint for the next step in computing, right? And we need to do that because the networking, the storage, all of this is quite intertwined, because the computer is not a single computer now, it's the whole set of servers in the datacenter.

So we worked with them. We put this footprint in. Customers come in. They experienced the latest engaged advances and our expectation in fact is that the CSPs themselves watch that in operation and say, okay, thank you very much NVIDIA. Now I'm scaling that out, I've got it, right? And that's a really good outcome for everybody, for the customer, for the CSPs and [ph]you are out.

Jensen makes the statement that we want to be the best sales team for the CSPs, just like we have been for the System Manufacturers, right? So I think that's how we see it.

Q - Toshiya Hari {BIO 6770302 <GO>}

Okay. That makes sense. So it seems like the response so far has been very positive.

A - Manuvir Das

Yeah, I think it's a great relationship. We are working with all the CSPs and they are keen to expand the footprint.

Q - Toshiya Hari {BIO 6770302 <GO>}

Okay. Got it. I mean, you talked about DGX and how you work with your systems partners that definitely want to go there. So you recently introduced the L40S GPU. Maybe spend a couple of minutes describing the significance of the L40S and how the systems business or the OEM business has gone because I know that's an important route for enterprises as well.

A - Manuvir Das

I think you know, we've realized now that we've got to the point with accelerated computing where all the system manufacturers, they have their mainstream product line and then they've -- eventually is the point where they have a product line for accelerated computing with NVIDIA. And the L40S is really saying those two worlds don't have to be separate worlds anymore. Your mainstream server line, every server can have a GPU in it. And that's why we think datacenters are headed.

I want to say this, Toshiya, you know that, if you ask me what is the thing that makes NVIDIA special, okay, and how we build our GPUs, it's that we have one architecture. We have one programming model based on CUDA. And so what that means is that we see the market and we're able to produce a family of GPUs for different situations. But they are all programmed the same way. So, the developer ecosystem doesn't have to do different things.

So why do we do the L40S, okay? So you know about the Hopper generation, the H100. There's very high demand for that. Right. And obviously, we are working on that and there are certain ways in which the chip is packaged, that system is packaged. The beauty of the L40S is, it's really good for that back-end of the workflow I talked about, the fine-tuning, the influence, of course, you can do training with it, but it's really go to the back-end and it's not constructed the same way, right.

It doesn't have the same requirements for the chip-making process as a Hopper family. Right. So it creates another channel of supply, if you will, of this kind of computing infrastructure and everything in it is designed to fit into standard servers, the form factor, the power consumption, and all of that. So now we are in this journey. We just announced with Dell, with HPE, with Lenovo, a number several manufacturers. I think we have 100 odd systems coming online, where for customers we say, as your tech refreshing, your datacenters going-forward, you know, what do you rack and stack. Right. What's the standard server you rack and stack now, you put one of these servers with the GPU in it.

And the reason for that is because there are so many different use cases you can do, whether it's AI or data processing and there's a plethora of use cases and it just saves your money because one of the servers can do so much more than a single CPU server. Right. So it's all adding up to the secular movement. They step back and you say, you look inside a datacenter today and you see how many of the servers in a typical datacenter today have a GPU in them, it's single digit percentage, right? And what we expect to see in a few years down the road is the majority of those servers will have GPUs in them, getting the right systems of the L40S and processes like that is one part of it, working with the developer ecosystem, these 4 million developers to do domains and move more-and-more domains to accelerated computing, so there is the different use cases, I mean, this is the journey.

Q - Toshiya Hari {BIO 6770302 <GO>}

Right. I think you guys have been extremely dominant in training. I think the competitive setup for the landscape and inference is a little bit more nuanced, but something like the L40S, it sounds like, will be effective in addressing the inference market.

A - Manuvir Das

Yeah. I think it's of course the hardware. We're very proud of our chips and we innovate on our chips in a very rapid fashion and we are standing on a body of work, right? It's the hardware we've developed over time, all the circuitry. But remember, we're 80% of software company and our software and hardware team work in conjunction, right? The reason why the H100 does what it does so well for deep

learning and for LLM inference is because the software team at NVIDIA is at the forefront of figuring out what I actually want that chip to do, right, which allows me to build the right chip, right.

But with inference now, it's been clear to us for many years now that the training leads up to the inference. Right. So there is two movements going on. Okay? On the one hand, you have to understand training is not a one-off thing, okay, because what happens as you train a model, but a model is only as good as the data you trained it on. And the reality is, as you're running your business, you're getting new data every day. So training is not a one-and-done.

So just to be clear, right, that training will be done continuously. Any big company is going to have what we call an AI factory where you're basically doing training all the time. But clearly, this is all useful because you're using the model in your application's inference.

So we've spent the last few years really designing a hardware but especially the software stack for inference, right? And we think that's the -- a big opportunity going forward. And as I said, every place where you're doing inference, there is only one run-time. You can pick up and take with you in your briefcase along with the model today to run that model where you want to run it, in your own datacenter, in a colo, in a public cloud of your choice, your TO2 CSP, on your workstation, on your PC, there is one run-time, that the world has, that you can take with you to do this work anywhere, and that's NVIDIA software runtime for inference. Right. So it's both the hardware opportunity and a software opportunity.

Q - Toshiya Hari {BIO 6770302 <GO>}

Really fascinating. I'm going to pause here and see, I guess there is one person. Do we have mics in the room? Check that first.

Q - Unidentified Participant

Thank you. First of all, congratulations on all your success. I was struck by your comment about NVIDIA positioning itself as the operating system for AI applications. Given that and given your support of OpenAI, what responsibility you think companies like NVIDIA have to make certain at the AI is used in non-nefarious ways and what kind of services do you offer enterprise-grade customers who care about the ability to audit results, and to make certain that the results that have been put out are non-troublesome (multiple speakers)

A - Manuvir Das

It's a great question and I like the fact that you use the word responsibility because responsible AI, I think is the thing, right? So, we're a platform company, right? And our job is to provide foundational technology. So, we think of this in two ways. Right? The first way is that from the point of view of companies who are deploying AI, how do we empower them to run AI in the way that they want to, right? So how do we make it portable?

And this is why we produce the software stack that you can take with you, right? So you can run the AI in your building, on your premises. You have a compliance regime, right? Your data is sitting in certain places. So instead of making you come to the AI, we bring the AI to you. Right? So that's one part of it.

The second part of it in the hardware. We do work on confidential computing. We've got our -- latest generation of hardware has this built-in because if you think about it, the models are now the IP, right? The models are the software. So how do you protect the models when they are deployed and running? So, we've done the work in the hardware to create confidential computing in this world of AI. So that's one aspect of it.

The tooling, we provide software tools like we have a technology called Guardrails which allows you to control what the model actually does, what questions that should and shouldn't answer, how to control how much hallucination you get when you use the model, right? So I think that's one part of it.

The other aspect with responsible AI is, it all starts with the data, right, the IP that is in your model actually came from the data. So, where is the data coming from? How is it sourced? Are the right people who actually contributed the data, getting the economic credit for producing that data? And so this is why we work with companies like Getty Images, Shutterstock, who have licensed content, right?

And what we're doing as NVIDIA is, we're enabling them to create these models for generative AI, based on those data set that they have sourced responsibly. Right. And then we have, for example, we work with WPP as a company in the advertising realm, where they are then able to reliably take the output of those models that have been built by Getty and Shutterstock, etcetera, knowing that when they get assets out of that, that they're using in their campaigns, there is a complete knowledge as to where these came from and they were responsibly sourced, right.

So I think that's the other aspect of it, which is, where is the data coming from? Is it responsibly sourced? I think there's many aspects in general and we are just one company in the ecosystem, but certainly this is a big focus for NVIDIA.

Q - Unidentified Participant

Hi, Can you talk about the inference opportunity for NVIDIA as opposed to training? So would your share of inference be maybe three years down versus your training share?

A - Manuvir Das

That's a great question. So you will have noticed generally philosophically we're not a company that thinks about, should I have 25% share of something or should I have 40% share of something. I think our view on this is very simple that most computers, servers and workstations going forward, we'll be in a position where they'll be used for inference. And we believe that we have been working on the best hardware and software stack for inference. And so, that's my answer to you.

Q - Toshiya Hari {BIO 6770302 <GO>}

I wanted to squeeze in a supply questions. It's not the best place to end, but wanted to squeeze it in. It's no secret that there is a supply-demand mismatch today. I guess, a, how significant is the supply-demand mismatch to the extent you can quantify it? As the Head of Enterprise Computing, how do you plan your business when supply is so tight in GPUs and at what point would you expect supply-demand to meet, if you will?

A - Manuvir Das

Yeah, it's a great question and an interesting one to end on. I would say, firstly, as you said, yes, definitely we see more demand than the supply. We've talked about that in our earnings. We've been racing to increase supply. I think our partners in the supply chain have done an exceptional job working with us to increase supply.

We talked about the L40S just a minute ago. So we've also found ways to increase the supply in these other ways. I think it's a process we are in over multiple quarters in a year. I think the good thing is that our customers on the one hand have a journey to put new footprint into datacenters, that takes multiple quarters of its own. Meanwhile, we are in this part to increase supply over the next few quarters and so I think there's sort of a natural balance between those two things.

I cannot sit here and tell you exactly when what's going to happen, but I think we feel pretty good about the fact that we've taken the steps to increase our supply. And meanwhile, customers on their journey too and we see this sort of happening in conjunction, as we go. And as I said, it's a secular shift, right? It's not something we are focused on for just one quarter.

Q - Toshiya Hari {BIO 6770302 <GO>}

Sure, sure. I guess in the last 60 seconds that we have, obviously, you're playing in this big and growing market. You are at the core of everything. Is there anything that we have missed in this conversation or anything you want to highlight to the Group before we let you go?

A - Manuvir Das

I think the main thing I'd just point out is, if you zoom out enough, right, I know we're all focused on LLMs and generative AI. It's the killer application. But really what we're seeing is this, right, it's -- what's coming to fruition finally is, this move to a new computing platform that the world really needed which is accelerated computing and large language models and generative AI, this killer app that is convincing everybody to adopt this platform, but as they adopt this platform, it's going to have a much more far-reaching effects, because all your workloads like data processing that you normally have been doing on traditional computing, are going to be accelerated dramatically.

And in the same footprint, you're going to be able to do 10 times the amount of work, 100 times the amount of work that you could do in that existing footprint

today. And this is important not just for saving money, but for the energy footprint of the world. And so that's what we are on the cusp of and that's what we are truly excited about as NVIDIA, and that's what we see as the opportunity going forward. And it's just a confluence of these two things because you need the right application to really drive people to the new platform and then you open up all the opportunities that the new platform provides, right?

And so that's really what we are looking at going forward that we're super-excited about.

Q - Toshiya Hari {BIO 6770302 <GO>}

Amazing. Really enjoyed the conversation. Congratulations on everything and thank you so much. Really appreciate your time.

A - Manuvir Das

Thank you. Thank you for the time.