5th Annual Wells Fargo TMT Virtual Summit

Company Participants

- Manuvir Das, Vice President of Enterprise Computing
- Simona Jankowski, Investor Relations

Other Participants

- Aaron Rakers, Analyst, Wells Fargo

Presentation

Aaron Rakers {BIO 6649630 <GO>}

Good afternoon, everybody. I'm Aaron Rakers, the IT Hardware and Semiconductor Analyst here at Wells Fargo, extremely pleased to host the Conference Call -- Knovio Conference Call, a virtual meeting with NVIDIA, Manuvir Das, who is the Head of NVIDIA's Enterprise Computing business, and we'll get into discussions here.

Before we get there, I think Simona Jankowski from the IR team wanted to read a quick disclaimer on forward-looking statements.

Simona Jankowski {BIO 7131672 <GO>}

Yes. Thank you, Aaron, for hosting us. Just as a quick reminder, this presentation contains forward-looking statements, and investors are advised to read our reports filed with the SEC for information related to risks and uncertainties facing our business.

Aaron Rakers {BIO 6649630 <GO>}

Perfect. Good deal. So, Manuvir, again thank you for joining us. There is a lot to delve into here with NVIDIA that the Company has been on fire in the data center business, so it's a great opportunity to have you speak to some of the key trends.

So -- but maybe to levels at Enterprise Computing, it's broad-based, it's increasingly more and more diverse. I think you and I talked about the BlueField product in the past, but maybe the level set the discussion, just a real quick overview of your responsibilities within the Company, and then we'll go in the questions from there.

Manuvir Das

Yeah. Thank you so much for having me, Aaron. And on behalf of NVIDIA, it's a pleasure to be here. In NVIDIA, we operate as one NVIDIA, one business. There are no people other than the CEO who think of themselves as owners of a particular business, right. So from myself and my peers who report to Jensen, we do three different things for him. One is, we're each responsible for some of the products.

In my case, I'm responsible for DGX, which is our flagship server for AI, as well as EGX, which is the more mainstream hardware that is built by our OEM partners with GPUs in them. And then I'm responsible on the software side for our GPU accelerated data science software and -- as well as our new push now for enterprise-grade software for AI that we're making available to enterprise customers and all sort of the middleware if you will, of AI. So that's one way of thinking about my role.

The second way is we serve, of course, different customer segments, and the customer segment that I focus on is really enterprise customers, and in particular, when they're doing their work outside of the public clouds -- a lot of them work in the public cloud. And then finally, perhaps most importantly for this call, we all think about different areas of NVIDIA's strategy, and the part that I work with Jensen on is our push to democratize AI across enterprise companies which means really the move from the thousands of enterprise companies that are using AI today to the hundreds of thousands that are actually out there, right. So that's really my mission.

Questions And Answers

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. That's a great overview and exactly the topics that we want to hit on here. So AI adoption, it's becoming clear that we move well beyond the hype cycle, right. It's an important driver of businesses. It's underpinning of productivity growth in many instances across different verticals. I think NVIDIA recently mentioned, I think 25 -- more than 25,000 companies using NVIDIA AI for inferencing globally. How do you think about what's next -- the next steps of enterprise adoption? And then we'll go into some of the product things that you're involved in that really looks to democratize that adoption rate.

A - Manuvir Das

Yeah. That's a great question, Aaron. And I just clarify that the statement we made about 25,000 numbers for using NVIDIA AI technology in general for both training and inference, right. And some of them use it in the cloud, some of them use it in their own buildings and so on, right.

I think, Aaron, when we think about the inflection point we are at right now, we are at an inflection point, okay. And I can describe that to you in three different ways. One way of thinking about it is the success we've had with AI to date has been companies in particular verticals where there was a clear use case very related to the business they're in. For example, I'm on an online shopping site and I need to recommend what's the next thing for somebody to buy. I'm a healthcare company and I'm doing

research on drug discovery. So very specific to the use case, we've had a lot of success there, and we'll see that adoption continue.

I think the inflection now is on horizontal use cases. So, for example, conversational AI is a way to create chatbots for customer service, and it doesn't matter what industry or companies in, as long as you're interacting with customers, you can benefit from having an AI-powered chatbot to communicate with your customers. So I think that's the first inflection which is from moving from these very vertical use cases to more horizontal use cases. So that's one.

I think the second inflection is if you think of the AI journey for company, training comes first, influence comes next because training is the process by which you hire data scientists and researchers and you say, okay, take on AI and show me that we can produce models that will help our business. And then when that work is done, then the company gets through the hardest step of I'm willing to put this into production to use in my business, and that's when inference is done.

So, in fact, if you look across the enterprise base, you'll find a lot of companies today who have done the training stage but are very early in doing the inference and actually putting this into production. And so I think in the next few years, we think you will see a dramatic change there, as this flood of companies gets to the stage of now actually using the AI in production by deploying it into inference. So I think that's a second way of thinking about the inflection.

And then I think the third way of thinking about the inflection which is perhaps a little further out. Today, when we talk about an enterprise company using AI, they need to understand what AI is. They need to learn how to do AI, how to do training how to do inference. But going forward, I think you'll see more companies adopt AI in a transparent manner. And what I mean by that is, I have some application I use today from some vendor, right, maybe it's a document editing platform like Office 365, or it's an ERP system. And the vendor just says to me, hey, there's a new version of this thing, guess what, it's now infused with AI and so it's better because of that. A great example is, you know, if you're editing in Word today, you start typing a sentence and it will suggest a completion for the sentence, and that is using AI. So in this case, the company that's adopting it does not have to learn AI technology at all. They're just deploying the next version of an existing product that has been infused with AI. So it's that vendor that has learned about AI and worked with NVIDIA to adopt AI, rather than the end customer, right. And that's really true marketization because that makes it so easy for everybody to adopt. So I think three different ways you can think about the journey forward for enterprise AI.

Q - Aaron Rakers {BIO 6649630 <GO>}

In that latter point, just kind to dovetail a question of that. Does that infusion of AI at the application layer, does that necessarily say, hey, that application layer is going to increasingly need a parallel compute -- accelerated compute underneath of it? Or is that, hey, this application infused with AI can run on general-purpose processors as effectively as a historical application? I'm just kind of curious of that -- you know, where that sits?

A - Manuvir Das

Yeah. And the key phrase that you use there was as effectively because, yes, inference can certainly run on general-purpose servers with CPUs, but it is increasingly the case that the more effective way, both in terms of cost and performance and capability to run the inference is on GPUs, not just the training but the inference. And this is why we believe that more and more servers -- regular servers and data centers will have GPUs in them because more and more applications running on the servers will be infused with AI, and they will benefit significantly from having a GPU in the server.

It's no different than saying, can you do some graphics work with just built-in graphics on the CPU versus than having a discrete GPU? Perhaps in some cases, you can but it's way more effective if you have a GPU, right? And it's no different here for the application infused with AI, we already see now that they benefit greatly from having a GPU in the server. And this trend is only going to get more dramatic, Aaron because the AI models that are being used within the application are becoming larger and larger and more complicated, and that's again where the GPU really makes a big difference.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. So one of the questions I often come back to is that NVIDIA, and I think this last earnings call, Jensen had mentioned 10% of servers deployed globally today, incorporated GPU. I'm going to kind of tie two questions together. So the comment was also made that over time, majority of servers could deploy a GPU. But when I talk to enterprises, ourselves being one of them, --

A - Manuvir Das

Right.

Q - Aaron Rakers {BIO 6649630 <GO>}

-- it doesn't seem like the traditional on-premise enterprise is going to have a big GPU attach rate. So what are we missing there? AI workloads running cloud, do they run on-premise data center? I'm just -- how does -- how do you guys see that?

A - Manuvir Das

Yeah. It's a great question and a great one to unpack. So number one, it is absolutely the case today that if you just look at how many servers are shipped, new servers that are shipped and going into enterprise data centers, servers with GPUs in them are in the single-digit percentages, right, that is the case today, and we expect that, that will change dramatically.

The reason why it is the way it is today is because there has been a general mental model that you need a special kind of server to do AI. A server that you pack it to the gills with GPUs. You put four GPUs, eight GPUs in the server, it's a bespoke thing. It's not a server that costs $10,000. It's a server that costs $100,000 or $200,000, and it's

like a Ferrari. And when you buy one of those, you better be running AI on it 24/7, or you wasted your money. That's been the general model and that's why the penetration has been what it is.

The shift that we are working on now with our customers is taking a general-purpose server, a server you would put inside of VMware farm, a server you would rack and stack into your data center as a multipurpose server that you can run SAP on it, you can put a database on it and what have you. You take that $10,000 server, you put one GPU in it that cost a few thousand dollars, and now you've got a server that on one hand can run your traditional workloads, and when you're only using for that, there is no regrets because it's a modestly priced server, but on the other hand, it can also run your accelerated workloads like AI.

So it's a multipurpose server, and when that begins to happen, we expect it will drive a significant shift in this percentage because it's the mainstream volume servers that will begin to ship with GPUs in them. And as Jensen said, we fully expect that in the fullness of time, this will become the standard. And so if you look enough years out, the majority of servers will have GPUs in them. All of them will have GPUs in them, but it is a journey, right. And so -- but I think this is a fundamental point that the difference between where we are at now and where we're headed is that today, the AI servers with GPUs in them are bespoke. They're putting a corner of the data center. They serve one purpose only, whereas the servers we are working with OEMs are now, along with the software we've built, these are multipurpose the new workhouse of your data center. The new server that you're going to rack and stack into every new footprint that you build so that your footprint can be future-proof, and the reason is because more and more of the workloads that you run in your data center are going to require a GPU because they've been infused with AI and they need that acceleration.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. No, that's an interesting point, right, because I think myself and a lot of investors I talk to look at it and to your point, there is eight GPUs on average in terms of training, there's four maybe for referencing. But what we're really talking about is just somewhat elasticity of bringing price points down to drive coupled with the application ecosystem evolving with more AI infusion.

A - Manuvir Das

That's absolutely right, Aaron. It's elasticity in the price point and its elasticity in the usage point. This is why we did the work with VMware, right, that you already have a farm of servers that are administered via VMware that are being shared out for different users and different workloads in your data center, and AI should just be taught of as another such workload. Your data scientists should just be thought of as another set of users who can come and access the same pool, and your IT team should be able to just expand the pool naturally to these use cases, right, and that's why we did the work with VMware.

Q - Aaron Rakers {BIO 6649630 <GO>}

So you've driven a bunch of other questions, I can add on there. But I'm going to kind of try and keep with my script here a little bit. So how does that involve, or how does that entail what NVIDIA is doing as far as a full-stack strategy company? You're responsible for DGX, EGX as you mentioned. Do you evolve down that path in those environments where elasticity starts to drive more attach rate of GPUs? Do you want to be that full -- are you that full-stack provider in that type of environment ecosystem?

A - Manuvir Das

Yes, absolutely. And I just want to be -- I'll be careful answering your question because I also don't want to miscommunicate here.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah.

A - Manuvir Das

Nearly -- the people who have done the best work in creating hyperscale multi-tenant compute environments are the public clouds, and they've done a fantastic job of that and they really changed computing quite significantly. We have very deep partnerships with all of them. We work very closely with all of them. In that case, we are providing them hardware and software that they incorporate into their clouds, right, to provide a multi-tenant hyperscale environment.

That said the modern datacenter regardless of whether it's an enterprise data center or a colo or a cloud is going to look more and more like that, where what you're trying to do is you have a shared environment of compute storage and networking, that is used for multiple applications. That's just how the industry has evolved, and NVIDIA is very committed to being a provider of that full-stack, okay. And that means the hardware, the networking, the middleware software, the operating system of this infrastructure if you will, as well as the AI software that you can run on top of that. We are absolutely working on all those fronts. We are a full-stack company.

Now, we are a platform company, and so what that means is, we are happy to interact with people at any level. You can acquire just the hardware and build everything else on top. You're not going to acquire our middleware software and do your own AI software on top. You can acquire our AI software and just go from there, right. And we are happy to work with you at any level, but we certainly have a viewpoint on what the full-stack should look like. We are working on the full stack and we make it available in different ways. For example, we announced a program this year where we are working with Equinix, and we are pre-deploying that full-stack already for customers in Equinix data centers, so they can just choose to consume it there. All they can learn from that and build their own.

So -- and the reason we do this, Aaron, it's important because it is a full-stack problem. Accelerated computing is not easy. You have to optimize the different parts of the stack together in order to get the best benefit. I'll give you a very small example of that. You think about AI and you produce these models, and models can

be large, and so you have to split the model up when you actually do the inference. Now and you have to think about what are the hardware capabilities of each of my servers, so I can be smart about in my software how to split up the mark. So you cannot do this an isolation one without the other, and that's why we've done all this work together.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah, that's a great overview. And you touched on that a little bit with the AI launched that product with Equinix. But the other thing that you've done is the -- I think the tagline was democratization of AI, when you announced this product, which is the AI Enterprise software suite.

A - Manuvir Das

Yes.

Q - Aaron Rakers {BIO 6649630 <GO>}

And can you help us appreciate what that is? How that's -- I believe it's a subscription license, how we think about that is democratizing AI and ultimately, representing a pretty attractive reportable soft [ph] for a revenue stream for the company?

A - Manuvir Das

Yeah. We're very excited about NVIDIA AI Enterprise. Of course, there is a revenue stream that we foresee that we are very excited about. But I think more than that, we're excited about what it means for enterprise company, right. So I'd put it to you this way that one of the barriers to AI adoption today for enterprise companies, I think it starts from there.

The first barrier is that, AI software is complex and difficult to create. There is a huge ecosystem today of start-ups who provide a piece here, a piece there, et cetera, but it's largely a DIY effort to coble software together to do AI. The reason we created NVIDIA AI Enterprise, as we said, we as NVIDIA have actually built most of this software. We've spent years building this software. We've matured the software. So now we can package it all together and say to an enterprise company, just procure this piece of business software, it has everything you need. It's enterprise-grade. It has a support SLA. It's like Windows. It's like VMware vSphere. It's a platform. It's the operating system of AI that you can go deploy on all servers in your data center. So that was the first point of it to make it easier to adopt AI.

The second point of it was if you think of the dilemma or gap that you have with AI for enterprises today, it's there is two personas. There's the data scientist. The data scientist is doing the research, creating new models, Jupyter Notebooks, all of that. Every night, they tweak the data a little bit, they get a better model, they like this model now, right, more than the one they had yesterday. On the other hand, you've got IT. IT are the people who control the data centers. More importantly, IT are the people who are accountable for the software that is running in production that the business depends on. Where if something happens to that software tomorrow,

whether it's compromised or not functioning, the Board wants to know what happened, right, my company's reputation is suffering.

So these two distinct entities. And in the early adoption of AI, the data scientists have essentially worked around IT. They've either done their work in the cloud or they built their own environments and going forward. So you have to break this gap somehow to truly democratize AI. And so we built NVIDIA AI Enterprise in this way, where we said, if you're enterprise IT and you are in the business of creating a farm of compute for your workloads and your users, then NVIDIA AI Enterprise is a thing that you add to your farm so that your compute farm is ready to do AI. And then you can go to your data scientists and researchers and say, come here and do your AI work here on this compute farm because we'll be able to provide that to you. There'll be a center of excellence where everybody can come and get their resources, and we will have one place where working together, so we can take the results of your work and put it into production.

And so that's the work we did with VMware, where basically we integrated NVIDIA AI Enterprise with VMware vSphere. So any IT admin who is familiar with VMware today, and they say, okay, I'll create templates of different kinds to spin up VMs in my VM farm, they can very easily create templates and they can say to a data scientist, here is a Jupyter Notebook template. Come and use this one and now you're good to go. So that's the second reason.

I'll give you a great example of this, Aaron. You know, we've been -- NVIDIA Enterprise has been available generally for a couple of months now in general availability. Obviously, we've had our initial traction and successes. There is one where I will highlight to you without the name where to meet it was the economical example of the shift that we were expecting that we saw.

There was an IT team. They read about NVIDIA Enterprise. They contacted us. They got into a launchpad. They had the experience of setting up an AI environment. They were happy with that. They set it up. They went to their data science team and they said, look, we set up an environment for you. Can you come try it out and tell us if it's good enough for you? So they brought the data scientists, they used it, they spun off Jupyter Notebooks, et cetera. They said this is great. This is exactly what I want. And the IT team came back to us and do a procurement for the software. So now, they are in the loop, and they are not outside of the AI Adoption. They are now actually part of it. And they are serving the AI to their data scientist. So that's kind of our expectation here.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Remind me just real briefly how it's licensed?

A - Manuvir Das

Yeah. So we chose to license based on the way that VMware licenses vSphere because this product as we've done it today is based on top of VMware vSphere. So it's the same licensing model, roughly speaking, it's about $2,000 per year, per superior socket, okay, is the license of the model, which -- for those who are familiar

with VMware will say, that actually looks very much like the pricing for VMware vSphere, and it's also -- it's model the same way in terms of support period and SLA and all of that. So from the point of view of procurement, it feels just like a natural extension. That's the simple way of thinking about it. It's about $2,000 per CPU socket per year on a subscription basis.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. That's perfect. A couple of other topics in a little bit of time we have left, I want to touch on is that, I've written a lot about it. I still feel like investors aren't fully either understanding or appreciative of another layer of acceleration, this idea of disaggregation, you and I talked about BlueField in the past.

A - Manuvir Das

Yeah.

Q - Aaron Rakers {BIO 6649630 <GO>}

So remind us again, how you see BlueField data processing units evolve? When we might see an inflection? And I think tied to that Project Monterey with VMware, the importance of that as we think about things over the next couple of quarters here or so?

A - Manuvir Das

Yeah. And I put it to you this way, right? The modern data center, as we discussed is multi-tenant, right. We have multiple people and workloads sharing the same equipment. And so the world, the infrastructure is becoming zero trust, which means that you can't just put firewalls on the outside and let everybody party inside. Every application, every user, every workload must be thought of as an adversary with respect to everything else that's running in the data center. And it turns out that in order to properly support a zero-trust environment, you need the capability of a DPU. That's really the only feasible way -- a tractable way of implementing a zero-trust environment because the amount of information going across the network that you have to process and analyze, and manage efficiently to create the secure environment requires those capabilities of having actual processing on the networking interface.

So we believe that regardless of whether you're a public cloud or a colo, an enterprise data center, the trend is very clear, every server will have to have a DPU in it in order to be part of a secure zero-trust environment, right, that's why we are so keen on the DPU. That's why you'll notice, Aaron that when we talk about our DPU BlueField, we do not talk about BlueField 2, which is our current BlueField. We talk about BlueField 2 and BlueField 3 and BlueField 4 because there is a journey here. This is an inflection that the whole industry is going to go through. It's going to take multiple generations of the hardware and software for that journey to occur. And furthermore, there is going to be a lot of different players in this system, cyber security providers, virtualization platform providers, et cetera.

And this is why we took the same approach with the DPU that we took with the GPU which is we created an SDK. It's called DOCA. And the idea is that all of the software developers from these different areas have one way to program to the DPU to take advantage of it, right, because otherwise, you'd have a piece of silicon that does one thing and you wouldn't really be able to support of the use cases.

So I think what I was going to convey there is why do we believe the DPU is so important because we think it will be in every server and it has to be in every server because the industry shifting to multi-tenant environments. It's going to be too expensive going forward to have siloed environments like this is my footprint that does one thing and that's my footprint that does another thing, right. That's going to be too expensive. That's going to be too siloed. That's going to be too inflexible, and the cloud has shown that. This is the beautiful thing about the cloud that if you have one pool of infrastructure that can be used in multiple ways, that's the winning proposition and so that's what we believe in.

Q - Aaron Rakers {BIO 6649630 <GO>}

I think to your point, the evolution of this disaggregation accelerated architecture isolation evolution, if you will, the roadmap, as you mentioned, BlueField 2, BlueField 3, BlueField 4, eventually involves actually some convergence of GPU and DPU together. Is that -- should I think about that theoretically as that being maybe the inflection of really driving this, you know, elasticity where these things converge? Or do you think it may be happens quicker than that?

A - Manuvir Das

Yeah, I think it happens quicker than that. So that's not reflecting. I'll come back to that. I think you talked about VMware Project Monterey, right. Again, that in itself is the way to think about the inflection. What is happening with VMware Project Monterey is if your server has a DPU inside it, then a lot of this work of enforcing zero-trust and security, et cetera, is now put onto the DPU, so you cannot actually have these servers run in this multi-tenant fashion whereas the CPU is freed up to actually run the workload.

What's happened today is that more and more of this functionality has gone into the software and it has to run on the same CPU that the workloads need to run on. So we move into the DPU, free up the CPU and more importantly, the DPU is built in such a way that it does that work much more efficiently than the CPU does, right, otherwise you're just shifting the problem.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah.

A - Manuvir Das

So we're very keen on Monterey, Aaron, because I think that's what's going to drive the inflection. The next stage of it, as you said, what we are uniquely doing in NVIDIA with the addition of GPU and AI capability into the DPU, just takes things to another

level because now you can use AI to understand what is happening in the network -- directly on the network, right, and so --

Q - Aaron Rakers {BIO 6649630 <GO>}

Okay. So is that necessarily that convergence does away with the GPU as a discrete component?

A - Manuvir Das

No, no, not at all.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Okay, that's perfect. That's helpful. The final -- in two minutes, and I know it's two minutes probably way too short to talk about this. But Omniverse Enterprise, I think any call that you've probably been on NVIDIA recently is Metaverse and how you play into that. For your responsibilities, what does that mean in a brief two-minute comment?

A - Manuvir Das

I think for me, as I mentioned, we're interested in providing enterprise-grade software to enterprise companies. And we have an offering with Omniverse. It's called Omniverse Enterprise. Basically, [ph] in general availability, we've seen great adoption for this, super excited because really Omniverse is a way for people to collaborate when they're doing 3D design work. And the number of people who are looking to collaborate in this way is massive. There's 40 million people out there who need this kind of collaboration.

And so we've created an Omniverse Enterprise. It has a simple licensing model again based on subscription. You can think of it in terms of thousands of dollars per such person per year. We have more detailed information, of course, in our documents, but you can build some simple math in your head of the opportunity here. We are also working on something called Omniverse Avatar for enterprise companies which is it's not just about the creators and users, but it's about the work that company does where you have a digital form of it within Omniverse, and that is its own journey and opportunity, both with financially and what it means for companies, right.

So what is encouraging to us, Aaron, is this, we worked for many years to produce Omniverse. And what we found in the year that we really now pushed it forward and brought together our newest enterprise, the traction has been amazing. The appetite for this thing has been quite amazing. So it's another one of these things, where I can sit and say to you in theory conceptually, the hypothesis is everybody is going to need it, it's going to change every enterprise company. But I can also say to you that in terms of the traction we've seen too far, it certainly seems to match that hypothesis. We see great appetite and interest in Omniverse Enterprise. Yeah.

Q - Aaron Rakers {BIO 6649630 <GO>}

I can keep going on, but I think we are unfortunately out of time, Manuvir. I appreciate you helping us understand the story a little bit better and thank you so much.

A - Manuvir Das

It was my pleasure. Again on behalf of NVIDIA, thank you so much for giving us the opportunity to have this discussion.

Q - Aaron Rakers {BIO 6649630 <GO>}

Perfect. Have a great day.

A - Manuvir Das

Yeah, you too.