

UBS GLobal Technology Conference

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer

Other Participants

- Tim Arcuri, Analyst, UBS

Presentation

Tim Arcuri {BIO 3824613 <GO>}

Great. We're going to get started. Hi, I am Tim Arcuri, I'm the Semiconductor Analyst here at UBS. Very pleased to have NVIDIA for our lunch keynote. We have Colette Kress. Everyone knows Colette, she's the CFO. And before I start to ask questions, I was going to let you have some opening remarks.

Colette Kress {BIO 18297352 <GO>}

Okay. Well, he had a great list of things that he wants to discuss, but I thought we should give a good overview. We announced our earnings right before Thanksgiving. So coming to this conference will be our first conference kind of highlighting. But as we all saw our results for the quarter, we're quite strong in terms of both the company-level as well as each and every single one of our market platforms, but most of the standout was our Data Center business.

Our Data Center business, we're seeing a -- continued growth in two key areas, both in terms of the growth of accelerated computing, but also the growth of generative AI in this place. That comes in many different forms. One, our customers sat, our customer sat expansion continues. Our consumer Internet companies and our enterprises are now on-board in terms of the work, in terms of generative AI, and the continuation of accelerated computing and was a strong contributor to our quarter.

Our CSPs, each and every one of our key CSPs have also stood up our HGX H100 architecture and are ready to serve customers with that computing mobile. Their growth was also a great contribution to the quarter as well. We continue to work in terms of new types of workloads, each and every day across this customer sat. These are important areas that we've been working on in terms of both large language models but also generative AI applications, recommended or engines are still an important part of our business in an accelerated computing as well and that continues to grow.

Additionally, we've added some key important areas that we are seeing around the world. We've talked about our AI factories that we see being built, but also the expansion in terms of sovereign AI around the world as many that involved in terms of what they need regionally for both their accelerated computing and work in terms of generative AI. So those results within the quarter continues as we provided our outlook for the next quarter in terms of growth.

We are planning in terms of supply increases each quarter, both of this current year that we're in, as well as the year after that in terms of increasing supply to support our customers as we go forward. So earlier this morning as we look up and down the West Coast of the US, up and Seattle area is where one of our customers is, but was hosting in Las Vegas, the announcement of our GH200 Superchips that is now available and with shipping within Q3. But one of our key partners, AWS, who we've been a partnership from the very onset was the company that set-up in the cloud the first time in GPUs in the cloud and have more than 2 million GPUs in the Cloud from NVIDIA.

But now will be the first customer of our GH200 Superchips and have enabled that not only for the use of other customers in terms of building a great supercomputer, fastest supercomputer planned in the future, but also will be a DGX cloud customer for us. This allows us to work with so many of our enterprises in terms of the software and the services that they need to support their AI work, but we're excited to have AWS again as part of our family here.

I'll turn it over to you.

Tim Arcuri {BIO 3824613 <GO>}

Great. Thank you, Colette. I guess I'll start-off with the question that or a version of the question that I asked on the call that you called a Tim question last night. Everyone I think in this room is obviously super impressed by how fast you're growing. And at the same time, I think everyone's concerned in terms of how sustainable that is? And whether you can keep on growing in 2025, I'm not asking you to guide but it sure sounds to me like you're pretty optimistic that you can continue to grow even through 2025. You know, you're putting your money where -- mouth is with your with your purchase commitments, they were up \$6.5 billion Q-on-Q, they were only up \$3.2 billion last quarter. They are seeing an acceleration in your -- in the commitments that you're making in terms of supply, it's hard to believe that you'd be doing that if you thought that 2025 was some kind of a peak -- sorry, 2024 was some kind of a peak. So can you just talk about all that?

Colette Kress {BIO 18297352 <GO>}

Yeah. So as we finished here at the end of 2023, one of the things that you've heard of over the last several months now in terms of talking about, one of it is really speeding up and broadening our ability to bring new products to the market. We think this is an essential piece of our growth going forward. Why? The overall expansion of this market since the onset of generative AI has continued to both

increase the TAM associated with that, but also the many different variants that we're seeing in AI.

We're bringing new products to market as a faster rate than probably what we've seen in prior architecture generations. One of the things that you saw was our H200 come to market. This is an important product that's bringing an increased memory size that really helps us a lot of these large models and customers are very excited to talk about the availability of H200 and to bring not only to our CSPs but many of our other end customers that are out there as well.

We've also talked in terms of the next architecture as well that will be available. This is an important area as people start that planning, planning out a data center. If you think about the time that it takes to just build the data center after planning can be, if you're World-class close to six months. So when you think about the time that's necessary to plan for the type of infrastructure that you want to build, the type of projects they want you could be in there about a year.

(inaudible) right now is helping them understand our product roadmap, helping NIM start that planning because of the advancement of accelerated computing and this AI that we're seeing. We think this is a big driver of growth as we go forward. So yes, there is an opportunity for us to continue that growth, but with just fueling them and helping them understand our product roadmap, we think that is a key driver of it as well.

Tim Arcuri {BIO 3824613 <GO>}

And there has -- in the past during years where you've had a big product transition, there has been some digestion that has sometimes happened, where customers say, well, I'm not going by the current-generation I'm going to wait for the nextgen. It seems that because year on allocation today that that's, you know, less of a risk than it, you know, would be otherwise, but can you talk about that? Is that something that we should be worried about that a customer would realistically wait to buy an H100 waiting for a Blackwell or --

Colette Kress {BIO 18297352 <GO>}

Yeah. During the year, we certainly had a surge of an immediate demand for folks wanting as quickly as possible to start their work and nerve fueling of AI solutions for many of the enterprises. So, yes, our supply and our ability to add supply has been our top priority to serve this market. And no, we're not done with that yet. We do believe as we move into the new calendar year that we will see substantial improvement in terms of our demand versus supply piece of that. But you're correct. Right now, we are still working on building more supply particularly one of the key areas would be our (inaudible) but each and every quarter we have advanced start. As you've seen in terms of how we've given our outlook and we'll continue to do that.

So when we know that we already have a significant amount of demand and already starting to help them plan for the next generation, we just believe this expanded market has contributed a lot of that interest and will still be with us as we look in terms of growth going forward.

Tim Arcuri {BIO 3824613 <GO>}

So, just relative to where we are in the adoption curve for AI, I think a lot of us are, you know, obviously we see all the capacity that's been installed and a lot of us are waiting for proof points from your end customers ultimately that they're getting a return on investment for all the money that they're (Technical Difficulty). So can you talk about that? Can you talk about maybe some of the KPIs that you're watching to, you know than to leave it there Is this next wave of elasticity coming as they actually monetize what they've already bought?

Colette Kress {BIO 18297352 <GO>}

Yeah, it's a really great way to think about when you are investing in the solutions with NVIDIA. You've got a plan in terms of the work that you're going to do. Some of them are in the early stages, some of them are already planning the monetization and/or the efficiency and productivity. It's important to note that both of those things exist. How do I monetize with specific applications or how am I looking for increased productivity that AI can provide me going forward that eliminates additional investment that they would have to make.

But already what we're seeing with the inferencing side of AI in improvement already this quarter from both launching H200 also tends for RT capabilities for LLM has improved a 4x improvement in terms of the cost-savings that they would have and the productivity that they've had on the inferencing. So what kind of applications do we see entering in terms of the market?

Well, the top ones that you see right off the bat are just these co-pilot's thoughts. You've seen the largest software company on the planet, think about what they can initially do with co-pilot with all of their installed base. That's a huge return on investment and one that you will likely see return on investment for each dollar that they are investing. They may get a return of investment of five, that's a huge easy math for them to do and why they're also interested in copilots.

You've seen many others that had been focusing on co-pilots and these assistance that will probably generate throughout the world very quickly and that they can monetize quickly and modernize software productivity applications now with that co-pilot. (inaudible) already received a tremendous amount of return on investment from the work they've done. Time and time, there are new capabilities of (inaudible) inferencing capability of our GPUs. But it has been world-known across much of our search customers that we have, recommender engines were just yet another one on that inferencing.

Marketing as it exists today has changed. The digital marketing and the use of recommenders is what is going to be necessary for all going forward. But then we can also look at the creative side of us, look at Adobe, Firefly, and the work that they have done now transforming what they've done to all the AI-generated. These are in their online a complete return on investment almost at the moment that they get it up and running and ready to go. We know more enterprises just see the opportunity to do a me too. Software, databases, those with lots of data you will continue to see that uses generative AI as we go forward.

Tim Arcuri {BIO 3824613 <GO>}

And I think just to that point, we're all looking at these highly visible, you know, end applications for, you know, for, you know, sign as to whether this is getting monetized, but ultimately you're taking your software stack and you're embedding it in some of these software platforms that are then going out into the world to then help to proliferate AI that went to.

Colette Kress {BIO 18297352 <GO>}

Absolutely. So if you think about our software platform, it is an important piece to many of these enterprises. Those enterprises are working out their level of the application. But when they think about the work that we've done on our full stack not infrastructure but hardware of course, is a key component. But our software that enables them A large language models, examples that they have, system software that's helping them to manage the overall work when it's up and running and many more STKs and things available, so that's an important part in working with our enterprises.

When you think about DGX Cloud or you just think about the offering of NVIDIA AI-E, which is just such a key component of that, that helps all enterprises around the world, no matter what their workforce is available to them. They can leverage out to begin that work very quickly.

Tim Arcuri {BIO 3824613 <GO>}

There was another thing that I thought was interesting last call. I think I counted six times when you talked about sovereign AI clouds maybe it was five times, but it was a lot. And it seems that we're all looking at the data center CapEx TAM and we're saying, well, it's a \$250 billion TAM and they're getting so big and they're going to be such a big percentage of that TAM, how can they keep growing. But when you start to think about it from a country perspective that this is becoming strategic from a country perspective and I think there was a reason why you mentioned that five times on the call that the TAM might ultimately be much larger than how we're looking at it today. So can you talk about that?

Colette Kress {BIO 18297352 <GO>}

Yeah. So let's just kind of break that down. When we talk about what did we mean by sovereign AI. Today, one of the most popular generative AI applications ChatGPT

really centered regarding the country we're in right now, regarding the US. US culture, US data, US speak in terms of that matter. Each country will look to develop AI, in their own language, their own culture, what they need in terms of foundation models. Those foundation models will be fueled also with proprietary data, likely from different enterprises taking more of that foundational model for what they can use within their enterprises as well. That is continuing to grow.

What you see being stood up is what we refer to as those regional specialized CSPs around the world. They are setting up different from the large CSPs that we see that have all different types of flavor of compute, but really focused on accelerated computing or just AI supercomputing together that they can serve either a whole nation, a whole large enterprise in many of these regions. This is growth that we see continuing. You've heard us talk about our work in terms of with India. Working with India, which has some of the best software engineering, best services there are interested in. How do I create that foundation model here within India? How do I create services for the many companies that they support?

Tim Arcuri {BIO 3824613 <GO>}

Great. Maybe you can talk about China, you know, obviously, there's been a lot of news lately with the ban and the potential for you to add back some revenue to come up with some SKUs that work, you know, within the new rules. Sounds like that's maybe a little bit delayed. But can you talk just about the potential for you to do what you did last time, which was come out with A800 and H800 and offer your, you know, Chinese customers a alternative that was -- that sat within the band?

Colette Kress {BIO 18297352 <GO>}

Yes. So we have new US export controls in place, and certainly, we will follow them on every piece of it. We've discussed a little bit in terms of the more complexity that there are with these export controls in terms of different levels. Different levels that says at a level of above, these would not be available within China. There's a second level that says it will be available or an opportunity with notification and discussion with the US government. And then there are some things at a lower level of performance that can be done without communications with the US government.

We will bring products to market. We will bring products to market with China. We will work with the US government. Keep in mind, not only do we want to make sure we have that relationship with the US government, your customers do as well as we set out on that path. So nothing unusual in terms of that process and we're working as quickly as possible. But you're correct. There will be new products available for the market to serve.

Tim Arcuri {BIO 3824613 <GO>}

And it seems a little bit -- you know, correct me if I'm wrong, but it seems a little bit different than last time where, okay, here are the lines in the sand. And then you came out with SKUs that just sat beyond those lines in the sand. They're sort of offering in the documentation. If you read it, it's sort of more of a we're willing to

work with you to allow export of even a SKU that is technically banned as long as you can include technology that can track its use case, that can guarantee its use case.

So it seems a bit different or it seems a bit, I don't want to say open-ended, but it seems a bit more like they're willing to negotiate as long as you can embed some technology that would ensure its use case.

Colette Kress {BIO 18297352 <GO>}

Yes. I see that part in the rules where they are seeking out a desire to understand from companies. Do they think that could be an option for them? Not necessarily put in place yet more of under review of some new type of way to discuss in terms of the customer use. I'd say that is still on the side and just a possibility but not something that we are focused on right now.

We are focused right now in terms of products that we know the US government fully understands and would be ready for China as soon as possible. So that's going to be our top of mind right now.

Tim Arcuri {BIO 3824613 <GO>}

Got it. Let's talk about capital allocation for a second. You generated \$7 billion in free cash flow last quarter. Only a little bit more than half of it was returned. I think the repo you've said is really just meant to offset dilution. You know, so you're going to have a lot of cash. I have you having \$40 billion in cash this time next year. What are your thoughts around M&A? There doesn't seem to be that much to buy because you've got a very unified architecture and platform, so it would have to be a pretty unique thing for it to fit in your unified architecture and platform. So how do you think about the, you know, usage of all this cash that you're generating?

Colette Kress {BIO 18297352 <GO>}

We focus on capital allocation both at the company level with our board and also with the most interest in terms of our investors, in terms of how we use the capital. First and foremost, our investments back into the business for any of our future product solutions is very, very key. Whether that be capital investments, whether that be hiring, all those different things, of course, are always included in terms of our use of capital.

We have done that opportunity to make sure we're returning to our shareholders by not establishing any type of dilution that's going to be front and center, because we do provide equity to our employees. It's an important part of their compensation, but that is another piece of it.

The next place that we turn to are different types of investments, investments that can also be forms of M&A as well. We have always been very thoughtful in terms of the M&A that we've done. Probably the most important key M&A that we did was Mellanox, probably one of our most favored M&A that we've had, and really being

such a great win for us, a great working together, something that just took off and did just phenomenal. Working together not only on accelerated computing with a GPU but the entire data center now as we understand that importance of networking.

We hope to find another great gem of that. We'll continue to look, but that doesn't mean there's also not small M&A that we'll look at from time to time as well to do.

Tim Arcuri {BIO 3824613 <GO>}

Great. Can we talk about software? We're, you know, 22 minutes in and we haven't talked about software yet. But I thought what was interesting I know, we get so caught up in how big the, you know, numbers are getting, but what I thought was interesting was that you did say that your software and your services have now crossed over \$1 billion annualized. So can you talk about that? Is there a way to think about how much growth we could see in that segment? All the customers we talked to for DGX Cloud, they all love the service. You've talked about AI foundry. There's a lot of different growth angles for this to take.

Colette Kress {BIO 18297352 <GO>}

Yes. So we announced that our software and services business as well as our SaaS businesses, is approaching \$1 billion within this quarter that we're in here in terms of the fourth quarter. We're excited in terms of the progress, we know how important our software has been to so many of our different customers that we work with. You're right. Our DGX Cloud is working in terms of expansion even now with AWS, joining the pack is a great opportunity for customers to work directly with us on so much of that software and planning solutions with them together.

We'll continue as we move forward as our infrastructure part of our sales, meaning our hardware sales, will be the underpinnings of what future drives that software, because that software is essential for the enterprises to do the work that they're going to do. So I think those two things will both scale and we'll continue to provide updates as we move going forward. But three key areas of that software NVIDIA AI Enterprise looking at it mostly as just the operating system for AI.

The operating system that every enterprise wants to have that says we have someone that's managing that software, keeping it both current forward compatibility, backward compatibility, no matter what part of our hardware platforms that they use and move to, but also giving them access to so many different SDK, system software that really helps them. There's already large language models ready for them to use to start with, written out, start with these models, great overall exceptional software for them to overall use.

Omniverse is the next one that's very key to us. Omniverse is a new area really focusing on that 3D Internet, that overall 3D look in terms of both building out factories or manufacturing. Last piece of it is our automotive software, our AV

software. It will be here soon as we have many design wins that we have focused on to use our AV software. And that, again will be an essential part.

Tim Arcuri {BIO 3824613 <GO>}

Great. I thought a really cool announcement last quarter was AI Foundry. And can you talk about this? I think you announced it on Azure on the earnings call. And it seems to me to be a pretty big deal where you can reduce the friction for an enterprise to bring their data to you, and the enterprise still owns their data, and you're -- you know, you're sort of operating in a way that TSMC would, they make chips for you.

I mean, you know, they bring their data to you, and you help them, you know, create models with their data. So can you talk about that and sort of how the early customer feedback has been on?

Colette Kress {BIO 18297352 <GO>}

Yeah, it's still in the early days, but it's an important piece. You've characterized it well. Even with TSMC, what we did with them was we didn't need to be that full manufacturing owner. We provided them the design. They're bringing data together. Look at that from an Azure standpoint, look at that from a large CSP or a regional CSP could do the same. Create that Foundry system that they have created a supercomputer bringing in their data. That data is owned within there.

And we can build them all of their AI solutions using our software stack and piece of it. It's an important way that we think people will absorb AI going forward to have that capability rather than all being both either on prem or being in the cloud where they do not have that support. So the support is an important part for them to establish their work.

Tim Arcuri {BIO 3824613 <GO>}

And you see that continuing to be hosted in public cloud. You don't see a scenario where you would ultimately build your own data centers to host that yourself.

Colette Kress {BIO 18297352 <GO>}

Yeah, we don't have plans for us to be in that business. We love our partnerships that we have around the world with all of the major CSPs. Each and every one of them have been tremendous partners with us in terms of their creation of helping the enterprises. That's good for them. Similar to -- we're not here to do TSMC's work either. So that's an important part of why this whole process works. They do that. We've added in terms of the software and support to help them get there and it's a great partnership.

Tim Arcuri {BIO 3824613 <GO>}

And can you talk a little bit about, you know, obviously, you have a large foundry partner. There are other alternatives that you could use on the Foundry side. Is there a scenario where if, you know, Intel was -- did build a successful Foundry business, that that's an alternative that you would consider?

Colette Kress {BIO 18297352 <GO>}

I think there are a lot of great Foundry partners. TSMC has been a great one. As you know, we also use Samsung today. Would we love a third one? Sure. We would love a third one. And that takes a work of what are they interested in terms of the services? Keep in mind, there is other ones that may come to the US. TSMC in the US may be an option for us as well. Not necessarily different, but again, in terms of a different region, nothing that stops us from potentially adding another Foundry.

Tim Arcuri {BIO 3824613 <GO>}

Got it. And maybe one -- there was a question asked on the call that you made a comment that guidance could have been better were it not for the China bands, which then -- I got a bunch of emails saying, well, that would suggest that they're not as -- they're not constrained. But I think what you were referring to was more L40S, which is -- which was banned, but is not itself constrained because it doesn't use co-host. Is that what --

Colette Kress {BIO 18297352 <GO>}

That is correct. So on the call, we asked (inaudible) that says, what if not? Could you have when you think about your outlook? As much as we would love to say that we've continued and improved supply over this period of time, each and every quarter, we still have some work to do. So we had indicated that although the China export restrictions have removed many products for us to say because of our strong demand, it wouldn't necessarily impact us in the near term.

Could it have increased a little bit? Yes. We have some products, L40S would be a great example that says had that not been restricted. Sure, we could have gotten more of an outlook for the quarter to do that. Keep in mind we're still working, we just need a little bit of time to get ready for our China customers and bring them some products that they could find tremendously beneficial. And I know we'll get that done.

Tim Arcuri {BIO 3824613 <GO>}

Got it. Maybe as a last question, is there a way that you can look at the networking attach rate? So for each dollar of compute you're shipping, is there a way to quantify what your networking attaches to each dollar of your shipping?

Colette Kress {BIO 18297352 <GO>}

Depending on the compute and who we are working with, it is very, very common that our networking is a very important part of that process as well. They have working not only in terms of designing the type of compute that they want to build, whether that be the memory around that, whether that be the CPU around it, the networking is essential. Makes sense when you think about the focus on AI is not about its time, just with the GPU, it's the moment that data enters the data center.

We have certainly visibility with these customers in terms of how much of the networking that they need. And often they are purchasing our full stack of networking on the same piece of it. Whether that be the switch, whether it be the optical cables, the adapters, it is a full swing. InfiniBand has been a very important part of what we're seeing in terms of AI. It is the gold standard for an AI build because it's dealing with all different types of traffic going in and out of that data center. And many have certified that for all of their AI solutions.

So, yes, we can see that and that has influenced a lot of the InfiniBand. And I think it's important to mention that going into this next calendar year, Spectrum-X for Ethernet again taking the same learnings that we understand of what is important in terms of speeds for that networking and how we can use Ethernet for also that AI and be world-class that is expected in Q1.

Tim Arcuri {BIO 3824613 <GO>}

Great. We've run out of time, but thank you, Colette.

Colette Kress {BIO 18297352 <GO>}

Absolutely. Thanks, Tim. Great to see you.

Tim Arcuri {BIO 3824613 <GO>}

Thank you.

This transcript may not be 100 percent accurate and may contain misspellings and other inaccuracies. This transcript is provided "as is", without express or implied warranties of any kind. Bloomberg retains all rights to this transcript and provides it solely for your personal, non-commercial use. Bloomberg, its suppliers and third-party agents shall have no liability for errors in this transcript or for lost profits, losses, or direct, indirect, incidental, consequential, special or punitive damages in connection with the furnishing, performance or use of such transcript. Neither the information nor any opinion expressed in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP. © COPYRIGHT 2024, BLOOMBERG LP. All rights reserved. Any reproduction, redistribution or retransmission is expressly prohibited.