Citi's 2023 Global Technology Conference

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer

Other Participants

- Atif Malik, Analyst, Citi

Presentation

Atif Malik {BIO 15866921 <GO>}

(Starts Abruptly) of Citi Global Technology Conference. My name is Atif Malik. I cover US semiconductors, semiconductor equipment and communication equipment stock here at Citi. It's my pleasure to welcome NVIDIA and Colette Kress as a keynote to our conference. Colette has been a very loyal keynote presenter at this conference for multiple years. She is going to walk us through a few slides at the start and then we'll dive into the fireside chat

Colette.

Colette Kress {BIO 18297352 <GO>}

All right. Thank you. Thanks all for joining us. Let me start with we've been on the road, I've been discussing with many different investors so I thought I could take an opportunity to start with some key statements that we've been discussing with many. First, remember that we will be likely making forward-looking statements. We strongly suggest you read our SEC filings that we did for further information about all risks and uncertainties. So I just wanted to open up with that.

This should look familiar to you. This is our revenue that we just completed in our second quarter of fiscal year '24, in the last four quarters in total that what we've seen. We've seen tremendous demand. Most recently, as generative AI has traveled through the world and an understanding of what the capabilities in a very simple format is capable for many.

We finished the quarter at $13.5 billion. A very big portion of that as you can see was our data center business. We have provided an outlook and guidance for the Q3 that we're in today that is even stronger at $16 billion, again of what we expect to be influenced by our data center business and for our data center systems that we create. Additionally, our revenue is being applied worldwide. However, we also have a very strong concentration in the US, which is now near 50%.

We're at the tipping point in the data center. We will likely when you annualize our Q2 results in terms of our revenue within data center will surpass the summation of what you see in terms of CPU servers in the data center if you annualize. That's one way to look at it so we probably look at it in terms of a little bit different manner.

We are a data center computing company. We are focusing on systems and solutions and a core end-to-end stack of software and services to build the modern data centers to build what you need for accelerated computing. We think this is a better view in terms of how big the market is in front of us and what our aspirations are for accelerated computing throughout.

You can see that we are a portion of the spend. We are a portion in terms of influencing the efficiency of how you build data centers as well as the performance that is capable of using accelerated computing and GPUs for AI and many additional workloads.

Now, when you look at us, we talk about us as a full-stack computing company. If you look here in terms of on the left, this focus is on the number of classical computing solutions in terms of building a data center with probably more than 900 to 1,000 different CPUs and (inaudible) servers. Given some of our -- performance of our top systems such as H100, we can reduce that down to just two GPU servers in that data center as a whole. That solves a lot of different solutions, not only from the performance, the amount of data you can process through the data center, but it also helps in terms of their total TCO value. You'll hear us talk quite a bit in terms of the more you buy, the more you save, but this is a case that not only do you save in terms of the cost of the infrastructure of the server alone, but what you have in terms of the data center, all of the cabling, all that you have in terms of the brick-and-mortar, the different build-outs. We save costs but also have one of the most energy-efficient solutions using accelerated computing. In this current day and edge, the overall making sure the use of energy is the most efficient case that they can do that, and using accelerating computing and AI is that.

It's also been a very important acquisition that we did with the acquisition for Mellanox and the improvement. In just the time frame that they had been a part of NVIDIA, we've increased the revenue along with their work with us in developing products by 4X. Additionally, when you think about the InfiniBand architecture and what that has provided for AI, we've reached a 7X versus that period.

The overall systems that are working in networking or the switches, the cabling, the speeds to influence the whole workload of accelerating computing from the minute the data enters into a data center is why our purchasing -- Mellanox and our work together is so important for accelerated computing.

We look at this slide because we give a lot of questions regarding what do we see in terms of supply, how are you thinking about suppliers you go forward with the demand that's in front of us. Yes, we do want to make sure we can move to support

the demand that's been presented and we've been working quite effectively with all of our suppliers, all of our different partners in terms of improving supply.

As this slide shows, we've been fairly flat in terms of the inventory that we have on hand at the end of each quarter, but this is really what we are looking at when we think about our supply. As you've seen a significant increase in the middle section here, which refers to our purchase commitments that we have been working around the world, which both our existing suppliers increasing capacity, adding new suppliers in terms of capacity.

Qualifying additional suppliers is an important part of our process every single day to make sure we can serve this demand. We also have long-term purchase commitments and prepaids with some of our providers to make sure that we can help them as they stand-up capacity for them. So when you really look at supply, these are the numbers that we look out to help understand that.

I want to end here before we move into additional questions that we have held, our TAM. What we look at in terms of our opportunity, we discussed this previously regarding our TAM that we see forth $1 trillion. $1 trillion that we see forth, when you look at both our chips and systems at about $300 million. We look at our software that we may apply to the enterprises of $150 million, as well as what we can do for Omniverse gaming and automotive. These are all very important opportunistic. We believe that the fueling of generative AI derives that inflection that starts us to reach our TAM opportunities that we see in front of us.

So all of this was in our plans as we thought about bringing more and more AI to the market and the future still will continue with more use of accelerated computing to fuel the data science.

Questions And Answers

Q - Atif Malik {BIO 15866921 <GO>}

Great. Thank you for those slides, Colette, and it's going to be helpful to talk to those slides as we go over some of these questions. And Colette, NVIDIA is seeing unprecedented data center demand, and as you pointed out at our dinner last night, AI is the killer application for accelerated computing just like gaming was for graphics. While the strong AI demand might be a surprise to some investors, I feel like you guys have talked about this over the last six-seven years with your full stack data center compute accelerating GPU model, adding networking through Mellanox on the way, and homegrown NVLink DPUs, and finally, CPUs with Grace. So a great job on the vision and execution, and thank you for taking us along on this fascinating journey.

My question is -- are going to based on the acronym you used at the dinner last night, which was SAS, S stands for supply, A stands for allocation, and S stands for sustainability. I thought that was a very clever acronym to summarize, and most of the questions we're getting from investors from here onwards. Obviously, the

demand outlook is very strong. When we talked to the startups in the Silicon Valley, we get a number around $25 billion to $30 billion based on just $250 million in the GPU opportunities under 2,000 GPUs. And obviously, the hyperscalers are not included in it and that's on top of it and then you have even governments asking for GPUs.

So on the supply side, every morning we get up and we hear some news on, so what's packaging, advanced packaging? A foundry is talking about (inaudible) And as you walk through some of these slides, can you help us understand how comfortable you feel on the supply situation as we try to meet this enormous demand?

A - Colette Kress {BIO 18297352 <GO>}

Yeah. So we talked a bit on -- you can see the bar that has increased quite a bit in terms of the purchase amendments, but kind of stepping back and understanding where did we start our work and how can you increase supply for the size of ramp that you have both seen between Q1 and Q2, and the extension that you see moving into Q3.

There has been discussions, correct, regarding CoWoS which is just such an essential important process force for advanced packaging, putting together our chips and memory together. There has been a great foundry partner of ours that has been working with us for many years as the development of CoWoS. But we have also worked now to both -- influence both additional capacity with them but also providing other suppliers that can help us in this process of setting up demand.

That supply that we have done with additional CoWoS suppliers, you're going to see that step-up, as we've indicated that we expect supply to increase each quarter, even as we move into fiscal year '25, and do expect there to be certain large step-ups as we increase this overall CoWoS capacity. And we couldn't be more pleased that often, our long-term partners are right here, helping us that we can help position you in that right position.

Some of the other areas of focus in the networking, we've done a fabulous job of often working and selling our networking along with our GPUs systems. We can get into situations given the large demand for AI large systems that InfiniBand and some of the cabling can be a little difficult in terms of all the SKUs that we have to work through. Again, this is an area that we worked very quickly in establishing more partnerships in terms of the optic cabling that we need and we feel like we're also in a really great place in terms of that.

So as we look forward, our focus right now not only is what do we see in terms of the long-term virtue. That's a very big part of our work today, but we do want to make sure we're serving our customers as best as we can.

Q - Atif Malik {BIO 15866921 <GO>}

Great. So the next topic is allocation. And obviously, you guys are working across your ecosystem partners with all the major hyperscalers, and how do you think about allocation of your finite supply? I thought last night, your description of looking at software companies and security companies, and what they're trying to do in AI and it was helpful. Can you share with us how you're thinking about allocation?

A - Colette Kress {BIO 18297352 <GO>}

So our allocation given the astounding demand worldwide, every region, every type of industry, every type of customer, as we have described in our earnings for Q2, our CSPs, our cloud service providers are more than the majority of the revenue that we received in data center. After that step, we move to the consumer Internet companies, some very large consumer Internet companies as well as a large tail of them.

Our third category therefore moves into the enterprises. Let's not forget though, your CSPs and your sales to your CSPs are also selling to the enterprises as well as they stand-up compute for research, standing up for large universities and also setting up for enterprises. But how we started this process from making sure that we were able to reach all different types of companies versus focusing on we need the PO, okay? Very helpful for us to start our planning if we are included in terms of their planning. We worked with many large companies for many years. They do help us understand that planning process and work, and that is one piece of our process that says help us in terms of that PO.

Another piece of it is also understanding are you ready to receive it, okay? Setting up data centers is not a very quick process. It takes time. It takes planning. They are also looking at the compute and the networking in some of the later stages of that. So we are looking in terms of exactly when do you expect to need this to provide it with inside of your data centers.

Thirdly, we have been working with many of these companies to understand the strategic plans that they have for us. We're often helping them what size of models are you looking for, what size of compute do you therefore need. That is helpful for us as well in terms of their planning. So as you can see, we have a broad set of customers, a continued advanced group of our customers really focusing on AI, and we're going to do the best that we can to serve them all as quickly as possible.

Q - Atif Malik {BIO 15866921 <GO>}

Okay. And in terms of demand, Colette, what are the signposts that investors should be watching in terms of the sustainability of generative AI demand, whether it's sales of Copilot and Microsoft or OpenAI? And what are the things that you guys are watching internally?

A - Colette Kress {BIO 18297352 <GO>}

When we think about the future, we knew there would be at some point, some great inflection there happen that the understanding of AI would be upon all of us, and the simplicity of ChatGPT was really that. There's not a person that couldn't

understand possible use cases of what you could do with something such as a model that could therefore solve problems of doing things manually or with humans to get some of that information across. That though is only the starting point where people have thought through, not only AI solutions, not only just generative AI, but also everything that we see in recommender [ph] engines.

What we see in terms of simulations, prototyping, all of that done in terms of using accelerated computing and AI, rather than the traditional methods. Many workloads have already started that transition, very large enterprises looking for their key applications, and what you should look to follow is where is the data. Large amounts of data are very, very prime for things that can be accelerated in the future.

Going forward, we think, accelerated computing with this inflection point is now more on people's minds for various reasons. One, the absence of Moore's Law or Moore's Law dying has really caused a view of what do we do with our CPU servers. Is that an upgradable solution, or is this the time to move to accelerated computing for the future to make sure that the efficiency of using that capital is well done? And this is where we see the market moving in this inflection point.

It's also been very key, not only from just a return on investment on that infrastructure, the performance improvement through productivity improvement that you have, but also thinking through that sustainability that is going to be necessary for each and every company to think through. While sustainability that says, you need energy efficiency, you have to be able to do the performance that you've done in the past using less energy, doing it faster than that, and there is nothing better than the overall systems that GPUs provide to do that. So we think this is just the beginning along with our TAM to see more and more just falling over to that accelerated computing in the future.

Q - Atif Malik {BIO 15866921 <GO>}

Great. And you recently highlighted new specialized GPU cloud providers such as CoreWeave as emerging customers, and in fact, you also have a minority stake in CoreWeave. Could you give us some more context around the extent to which they are contributing to revenues, and how that relationship works in general?

A - Colette Kress {BIO 18297352 <GO>}

Yeah. It's been interesting to see that along with our work in terms of accelerated computing, new types of CSPs have informed. New CSPs that in sometimes aren't doing the exact same thing that the large CSPs are doing. A large -- I've got any type of infrastructure that you want. I have CPUs here. I have different types of configs.

CoreWeave, as an example, specialized really in just accelerated GPU computing, and that has been their goal. Now CoreWeave has also quite some skills in terms of just their speed of adoption, their speed in terms of setting things up. We have a very small minority passive investment in them, and they have been working not only with us but very many other large customers in terms of standing up compute for them. They are small. They did have some allocation, but again, it's very small.

Q - Atif Malik {BIO 15866921 <GO>}

Right. And Colette, part of the sustainability equation is competition and your CUDA advantage is clearly strong. And can you talk about how do you feel about your competitive strengths in terms of your data center products?

A - Colette Kress {BIO 18297352 <GO>}

So when you think about the possibility of how competition views in terms of our work, this is not anything new that we haven't seen over the last decade of what folks want to look at is in terms of alternative that could possibly allow them that monetization. It won't surprise us that some different types of ASICs -- custom ASICs will be designed, but again, it's often not very comparable to what we have provided.

What I mean by that is when you think about what we have developed at data center scale and a full end-to-end platform with the software, with the development platform, it becomes not easy to match to it. We'll see custom ASICs likely hit the market. It's a challenging process that they learned. It's difficult to get up to the same performance of 30 years of our business that we've done. And our four teams are focused solely on accelerated in GPU computing. But if they find a standard not moving workload that may be able to be a hard quoted for that, that could be a possibility.

So we stand to really be that platform solution for accelerated computing. There will be other different players along the line, but we do really focus in terms of how do we support the platform for accelerated computing going forward.

Q - Atif Malik {BIO 15866921 <GO>}

Great. Moving on to the TAM. The slide you showed, you talked about a $1 trillion TAM across different end markets in autos and chips. Jensen has defined your AI TAM as the x86 installed base that could potentially get converted to a different model of data centers, the data centers that are more dense with GPUs in them. So can you just kind of elaborate what will happen to those general-purpose data centers as this new AI server model takes off?

A - Colette Kress {BIO 18297352 <GO>}

It's going to be top-of-mind for many companies to think through that installed base. What we're referring to here is, we believe there is probably $1 trillion of x86 CPU servers that are installed. Some of them may be halfway through their lives, some may be in terms of their depreciable lives coming to the end. Is that something that people want to renew, go out, buy a new CPU server? Or is this the time that you can take a different approach to accelerated computing, start your work of including GPUs to really begin that transition to more performance and to more TCO value going forward?

That's the question. Capital is important. Those are things that you want to make sure you're getting the best performance, but also the best return on that investment. So

we'll likely see that how people think about the size of AI. They're beginning to have a meaningful amount of projects so that could be a start of transitioning many of the data centers to accelerated computing.

Q - Atif Malik {BIO 15866921 <GO>}

Got it. But I get this question quite often is, how big is your infants opportunity relative to training and what is your share in the inference market versus training?

A - Colette Kress {BIO 18297352 <GO>}

Yes. Our training market, there has been no confusion that we have a very material percentage of the training market. Most training because of the strong amount of data that is necessary, our solutions have been very helpful for that. Now, what we did several architectures prior to our Hopper architecture is we helped a lot of the customers work through this position, where our systems that we have created now are capable, at the same time doing training and moving to an inferencing stage in terms of that. [ph] What that helped was both the decision-making challenge that said, what system, what do I need to use because you now have a system that can do, but you also have saved a tremendous amount of TCO. I don't need to transition to a new cluster -- create a new cluster for that inferencing. So at the same time, we believe we've helped and put them together. We can get feedback in terms of how they're leveraging this for inferencing.

Now, inferencing is not new. Inferencing is decades old. However, inferencing is changing. The inferencing of 30 years ago, very common to be a binary solution that I need a yes, no answer, maybe something in terms of the prototyping and qualifying. That can still stay as it is. That's not where we will probably operate in. But when you see the complexity of the types of algorithms now, and the response time that's necessary often in milliseconds, the complexity of doing that is perfect for GPU computing. And we'll likely see more and more transition to that.

You can talk through in terms of companies like Microsoft that is looking at that Copilot attachment to their Office 365. That's it's a very important inferencing solution as they think through the use of GPUs through and through in terms of the work that they do. There is other large recommender engines, search commands, all of these things that can benefit and are using that inferencing has been very important solutions for them.

So as everything starts advancing more and more, we do believe it will be a material part of inferencing because we think it's a very large market as well in front of us.

Q - Atif Malik {BIO 15866921 <GO>}

And then how should we think about your system sales versus component sales versus software sales as an emerging area for you?

A - Colette Kress {BIO 18297352 <GO>}

When we think through our sales today, we do laugh a little bit and say, well, we're not exactly a chip company. If you see in some of these systems, several of us may not be able to carry them, and they are enormous. And if you think through our H100 systems, 35,000 different components are used to make one of those systems.

So our systems approach in that full stock has been extremely helpful to our customers in terms of the ease of adoption as all of those different things have been qualified, but more importantly, we have focused from the time information enters the data center to the end to focus on accelerating every different piece of that.

So right now in our data center revenue that we have systems are more than the majority. They are going to be the lion's share of what we have. We'll probably see that continue for quarters. But additionally, we have come to market and we have new products coming to market to also assist many of our enterprises. That likely have a difference set-up at the current time in terms of their data centers.

This is an area where we can provide them, for example, a L40S. L40S is available into the market this quarter and all the way through for the second half. Why is that an important product for them? One, you can take an OEM, ODM server. We probably have 100 of them coming out that they will be able to put in four L40Ss inside of that configuration. That's a great -- just right there server for a small model in terms of training, but also during the inferencing.

With the networking solutions, linking that together, you have also a great opportunity for the enterprises. One, it can be consistent with what they have inside of the data center and start their work on-premise on many of their AI -- generative AI solutions that they're working on. We'll also come to market, for example, with Grace Hopper 200. It's also another great system. Now what's different, we're introducing Grace.

Grace is an important piece of our vision of accelerating at everything that we can. It is one of the best-performing types of CPUs out there. It's always been focused on arm and focused on low-energy, and that's an important stage when you think about your total TCO of everything that you do. That's going to allow the ability to be well-engineered for all of your AI types of work that you will do in accelerated computing. Those two have been engineered together. Bringing not market also with a great memory solution is also a great system solution now.

So there will be more than just the systems that the large systems we see, more and more new products will come to really tailor for some of these very specific types of workloads that we're going to see.

Q - Atif Malik {BIO 15866921 <GO>}

Another topic that comes across with investors is China. What is the long-term impact to NVIDIA's business from continued US government restrictions on China and certain countries in Middle East?

A - Colette Kress {BIO 18297352 <GO>}

So the US government has been working for decades in terms of working on protecting the US. They are concerned only regarding national security, and that work has led to many types of products having some forms of restrictions. We are following those restrictions very closely in terms of what we see. And they have a version of our products that is just not the same in terms of what we have here for many of our US customers and other regions.

We continue our discussions with them. There has been rumors out there regarding possible additional changes. We don't know what that will be -- if there will be in that perspective. We do believe the export restrictions that have been put in place are effective. They are effective in the work that they've done, both in terms of those restrictions. There is also a list of different types of companies that you can't sell to in China as well. Additionally, as read, our US government also asked that for any product built in the US and shipped to the Middle East do so with a license and would need a license for the US government. That's not a meaningful amount of revenue to us. However, we have been informed by the US government on that piece.

We've discussed our data center revenue that is associated with our China business is approximately 20% to 25% of any one of our quarter, and we continue to work to see what additional changes there will be. But probably most important, it's a very important market to many companies, the industry as a whole. And so, as we carefully think that through, changes will not necessarily affect anything in the short-term for us as our demand is quite strong. But thinking about it for the very long-term, depending on any changes, it will impact the industry and it would impact us.

Q - Atif Malik {BIO 15866921 <GO>}

Great. And let's spend a few minutes on gaming. It's still 30% of your sales. And how is the gaming seasonality changing because of notebooks, and how would it impact NVIDIA going forward?

A - Colette Kress {BIO 18297352 <GO>}

Yeah. Our gaming business has been both a combination of high-performance desktop solutions. Self building gaming solutions is a very big key for gamers. They love all of those different components. However what has also been great to see is laptops and the growth that laptops have. We have helped for nearly most of this decade with the OEMs that allows a configuration called Max-Q that helps them include the highest performing GPU and the thin and light that many of us know have with most of our laptops. That important configuration has really fueled a very large market now for gamers that use notebooks for that. It allows them to game anywhere they are, whether they are in classes, at college, back in their dorm rooms or anytime that they have their laptop with them.

What we see in some regions is we're actually selling more laptops than we actually are in terms of desktop solutions. That changes a little bit of the seasonality. Working with our OEMs and what they do to bring laptops to market and the distribution, you

see strong quarters in your second and third quarter as they get ready for back-to-school and they get back to building those systems for the holiday as well.

So this might change to be those being strong quarters. Our H2 usually is a stronger half, just in general, with the holidays that are there so we'll probably see that continue as well.

Q - Atif Malik {BIO 15866921 <GO>}

So one question on auto. You have a very strong position in the cloud, which is training the fleet and -- but your chips and position in the cars is lower than -- in the cloud. Can you just talk about how the adoption of large language models would kind of play into your hands as your ramp your auto pipeline, which I believe is $14 billion or so?

A - Colette Kress {BIO 18297352 <GO>}

Yes. We've talked about our pipeline as we look forward out about six years or so. We'll probably have about $14 billion from what we procured in design wins with many of these AV companies as well as what we're seeing in terms of the EV companies. This is an important area where they are using our computing platform as a very important part of the builds of electrical vehicles that you can really restart the buildable -- manufacturing those cars.

That compute platform has been very helpful, not only for all of the electronics, all of the software in the car, but also helping them with AV, both of those capabilities with a standard platform. Our own architecture is great for that. Additionally, we have long-standing agreements with Daimler, with JLR in terms of helping them, not only with the infrastructure inside of the car, but a software platform for high-end AV across their entire fleet. We will see that come to market in the years forward, and that also now gives us an opportunity for a software stream as we move forward.

So right now, you see a good amount of our revenue just from our cars today, our design-win, working with the EVs, and in the future, you will see more as we focus on the software.

Q - Atif Malik {BIO 15866921 <GO>}

Great. Let me pause here and see if there are any questions in the audience. It's okay. We will keep going. And another question on capital allocation. You guys surprised investors with a large $25 billion share repurchase announcement on the last earnings call, and you plan to buy your shares through this year. And what is the long-term capital allocation M&A strategy?

A - Colette Kress {BIO 18297352 <GO>}

Yeah. Capital allocation, all companies want to look to say how do we leverage and use that capital that we see. Always it's going to be our number one goal is investment back into our business. Investment in that business can be anything from our suppliers to our R&D teams to adding more employees is very, very important,

but also in terms of M&A. Small and medium M&A will always be an area that we research. We take a strong look at these types of companies that we think could help bolt-on and bolt-on and assist us in the work that we're doing on accelerated computing. But sometimes those are hard to find, but it is still part of our number-one strategy on capital.

Second thing in terms of using capital, we'll look through the dilution factor for providing employees equity in the Company. We want to make sure we balance that, not driving the dilution and therefore repurchasing our stock will be there.

What we did in terms of executing an authorization of $25 billion? It was getting to the point that our authorization level was getting low and we restarted our authorization that can take us forward. There is no timeline of where that ends, but we do have that authorization for the dilution part of our business. Now keep in mind though, we will continue our investments. We've got great plans. We're here today, but as you know, we have a tremendous amount of engineers working on continuing to expand our work that we do.

Q - Atif Malik {BIO 15866921 <GO>}

Great. And we're almost out of time. Colette, thank you for coming to the Citi conference.

A - Colette Kress {BIO 18297352 <GO>}

Thank you.