Morgan Stanley Technology, Media & Telecom Conference

Company Participants

- Colette M. Kress, Executive Vice President and Chief Financial Officer

Other Participants

- Joseph Moore, Analyst, Morgan Stanley
- Unidentified Participant

Presentation

Joseph Moore {BIO 17644779 <GO>}

Alright. I think we're ready to get going. I'm Joe Moore from Morgan Stanley. Very happy to have with us today, Colette Kress, the CFO of NVIDIA. I think Colette wanted to make some opening hedge comments and then we can go straight into it.

Colette M. Kress {BIO 18297352 <GO>}

Absolutely. Okay, I have a statement to read. As a 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.

Questions And Answers

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. Well, I feel like we can go for a couple of hours, given how much stuff has been going on for you guys. But maybe we could start, you and I've been talking in this forum about various forms of large language models for four years. But I feel like 2023 is an important year where this is now becoming mainstream, it's becoming every hyperscaler kind of talking about their capabilities as table stakes going forward.

So I wonder if you could just talk to that? And you made the comments on the earnings call that in the last 60 days has all this enthusiasm has built, that you do see that transforming into demand for product on the road. Could you just talk about what those conversations are like right now?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. So really good way to start, talking about where we ended with our earnings release and some of the great statements that we've made. But let me kind of step-back a bit and talk about when we entered into this calendar year, what did we expect, moving and probably important to understand is, we do believe AI is such an important part of computing, an important part of both accelerated computing and really what we're about. But we've entered a stage now where we actually believe AI is an inflection point. And what is stemming that inflection point, not only our focus in terms of large language models, recommend data engines or natural language processing, that we've done, but we have now incurred a point in time with generative AI, particularly with ChatGPT, that folks understand in just some of the most simple cases, how this can benefit them. Benefit them from a use case as a consumer or an enterprise on thinking about how they can develop AI within their universe as well, both for monetization capabilities or just sheer efficiency and improvement.

So we're at an important stage, we're at an important stage now for AI, and our work that we have done over the years is both bringing a high end platforms. But more than just thinking about something as the H100 or the A100 from a computing platform is the data center as a whole, and what we can do in terms of improving efficiencies and use of acceleration. A key use case there would be AI and we're seeing that everyday. So the interest that we've seen over the holidays is, we're talking about large enterprises. CEOs of large enterprises, where they've originally focused on, we need to concentrate on AI, now both the leadership team. Now, the CEOs, all understand what is possible, whether it could be something as easy as a chatbot, whether it be simulation of their factories, whether it could be even more use cases that will continue to develop overtime.

Start-ups are very interested in this as well, because this is where those large language models have been built and we'll likely see a lot more large language models start to influence cases. Things not only here in the U.S., but things worldwide as well. Whether those are foundational models or whether are specific to industries. But this therefore fuels CSPs interest. CSPs have service agreements with many of these companies, need to set-up -- compute for them to start their work, if they want to start on large language models and generative AI. So all of these things continue to fuel the demand that we see.

Q - Joseph Moore {BIO 17644779 <GO>}

I think the demand creation from it, it's pretty clear on the one-hand, where -- when I watched CNBC during the day, CEO after CEO comes on and says these are important developments for us. We need to figure out how to use this better. How does that manifest for you? Is that people than -- the people who are already developing these types of models? Are looking at much higher levels of complexity, looking at scaling got out, is it driving new people to develop models from the ground-up? Is it generating interest for your own models that you guys were talking about? Just in general, how do you take that excitement and enthusiasm, how quickly does that turn into CapEx and NVIDIA opportunity...

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. It's a really good question in terms of how does that process start. What is interesting is all those places that you referred to is exactly what we see, all different types of it, because we focus on the full stock and because we have focused on a model, using our ecosystem to get to enterprises and get to customers, we are seeing them at every stage of where they are with AI. Haven't started. They want to focus on AI, they need some help in terms of building out how they want to design that architecture. Two, they are already working with an existing model, need more optimization. And what we can do in terms of at the optimization level with CUDA, and/or working with their framework. Or even we're in the situation where we maybe exactly helping them with their model. Starting them with one of our own models or optimizing their existing model with efficiencies, that can help in both ways, focused on the hardware, focused on the software. So we're really all in -- on any type of AI, help that we can do. Now at the end-of-the day, what that does is, there is many different opportunities for them to procure. Okay, they can procure directly with CSPs, they compare with procure with one of our partners as well. So the opportunity for them to absorb AI because of our full platform in many different ways.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. And maybe you can help us understand how this overlaps with a kind of declining budget environment from the hyperscalers? And you've alluded to this, you saw some of that in your last quarter but everyone's kind of thinking about scaling back to budget. Do you see them protecting and growing graphics within those budgets over-time? Given the importance of these workloads and just how do you think, how do we transition from the budget cut being the focus to the kind of focus on the growth that you guys are generating?

A - Colette M. Kress {BIO 18297352 <GO>}

When you think about these economic times, it's both the time for folks to focus on their budgets or focus on what they're spending on. However, they're still working on efficiencies of how they're using their money or how they're using their capital. The focus on accelerated computing, no matter how you look at it is always going to be an improvement of efficiency and the use of their money. The amount of money that they save in terms of moving to accelerated, not only is it more efficient just from a computing but you're spending less costs. Less costs versus a existing CPU type of infrastructure or any other version of legacy type of computing.

So we tend to be front and center of the priorities, even in the time that they are working across the enterprise to be more efficient, you are going to focus on efficiency. Our work therefore is helping them design those different pieces. Although there were some areas within our Q4, where they needed to push some things out, nothing was lost, things just were taking a little bit more time for whatever different reasons they may have in terms of setting up at CSPs or other types of data centers. But the important part is, the world has understood that Moore's Law is not with us anymore, we will not drive that efficiency, you have the overall need for sustainability as well. So focusing on accelerated computing is going to be a priority. Accelerated computing, that is also demonstrating AI is also another added bonus for those that are focusing.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. And I guess another part of this is going to be the capabilities of your Hopper products that you're ramping now, which obviously were designed with these types of models in mind, you talked about 6X, the transformer performance at the time you launched it. How important is that hopper transition and how quickly do you see a transitioning from Ampere?

A - Colette M. Kress {BIO 18297352 <GO>}

Yes. Our Hopper architecture has been in the design phase for many years. We've been working not only internally, but keep in mind, we work with some of the key generative AI partners. Open AI, which created ChatGPT, has been with us and working with us for many years. We will continue working with many of these different partners, that helps us in the design of what we brought to market. There is no unusualness about that H100, includes the transformer engine -- a transforming engine, which is really modeled towards large language models.

Yes, we've been talking about it from years and years, but now you're seeing a very important use case. We'll continue on that path to influence, not only the hardware specifics, but the software, not only with the transformer engine from an infrastructure standpoint is included. We've also made huge inroads in terms of what we can do in terms of software. Hardware alone is probably about a 5x to 6x improvement from the last generation, from our software or even an inferencing standpoint you're dealing with, maybe close to a 30x improvement. So Hopper is an important piece, it just allows more efficiency. So we believe as we ramp H100 those focusing on some of these key workloads will really look to use Hopper. However, we still also ship our A100 which is also the second-best, that is out there in the market and still may be a great opportunities for those existing clusters that they may have and adding on to it, because it's already qualified, we will probably see both H100 and A100 this year.

Q - Joseph Moore {BIO 17644779 <GO>}

And one of the surprising elements at least for me around the quarter was that H100 was bigger than A100. Where does that leave you for the rest of the year? You've talked about an accelerating growth in data center, but I was sort thinking you would get a big value uptick because you have H100 ramping through the year. It seems like that mix might be a little bit more static from here. So how do you think of that in terms of how that drive visibility through the rest of the year?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. So Q4 represented probably the first significant quarter with H100. We started doing production at the tail-end of Q3, Q4 was there. So we're still in that ramping perspective. But as we've discussed, both of them will be important H100 and A100 as we see this moving forward. We expect growth as we head into Q1. The quarter that we're in right now from a sequential standpoint, we expect strong growth and we also expect a little bit of year-over-year growth. So our focus is on accelerating that growth rate for the rest of the year and that can be an important piece for H100, but can also be for A100 and our other products in there as well.

Q - Joseph Moore {BIO 17644779 <GO>}

Okay. Great. And then I guess, accelerating growth, just to define that you're sort of talking about year-on-year growth each quarter growing a little bit faster, the concern that challenging in July. So is it -- can you just talk to, is that a steep ramp through the year? Or just how are you seeing that today? It seems like maybe a little bit more visibility to the back-half of the year than the first-half?

A - Colette M. Kress {BIO 18297352 <GO>}

Yes, the growth that we're expecting for data center, yes, from a year-over-year perspective, we expect that acceleration after we leave Q1. What type of growth that is, I think, we probably look more at the size of the TAM that's in front of us, the opportunity with so many different offerings availability to do that. So it's really hard to say, how fast that growth rate is going to be, but we know that this inflection point is a key focus area for us.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. And then on the inference side of it, you mentioned the inference performance of Hopper being 30 times. But it's really interesting now that this is Microsoft and Google kind of bickering in each other over who has the lowest cost per query like this esoteric question of inference cost is suddenly like one of the most important costs. Can you size how big your inference business is today? And can you talk to the role of your inference as you think specifically two large language models down the road?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. As everyone focus on efficiencies with inside of their data center, they're all looking at what are things costing. But at the same time that you focus on cost, you have to focus on what are you seeing in terms of utilization across everything in terms of the data center. So let's kind of step-back that we have been putting in market, a platform that allows you to do both, training and inferencing on the same overall platform. Why did we do that? We did that, such that, at the time of both procurement and the time that you trying to optimize from the training side to the inferencing, you heard have synergies across that and you could probably benefit.

So at the same time that you could be spending now working on a large language model, training it, next week you could mark, in terms of the inferencing side of it. We have both improvement from the training side and the inferencing side in terms of improving productivity through that. But what's also important is being able to capture the power that is improperly being utilized and accelerated computing does that as well. So at the same time that we may improve training and with a better H100 to do that, we also improved the overall inferencing as well. And we improved the whole system -- and so we see people moving to training and inferencing on the same platform. The more complexity that you are seeing in inferencing really stems from those large language models and those using accelerated computing for inferencing as well. So with software, with the utilization of power, with its overall accelerated computing, you'll probably see both training and inferencing costs come down over time.

Q - Joseph Moore {BIO 17644779 <GO>}

But it's hard for you to classify, which is -- which I guess, if it's the same parts?

A - Colette M. Kress {BIO 18297352 <GO>}

Correct. We can't determine at any point in time how much has been done for training and inferencing on a individual box. We do believe training is important part today. We've made great inroads on a meaningful amount of inferencing as well and this is a great -- bigger opportunity as we see going forward too.

Q - Joseph Moore {BIO 17644779 <GO>}

Okay. Great. Thank you. Maybe if we could shift a little bit to a new initiative that you talked about on the call, the DGX Cloud. Can you kind of explain to us what that is and what the scope of NVIDIA's ambitions are in terms of having your own kind of cloud service capabilities offered within bigger cloud service providers?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. DGX Cloud, our new offering that we're going to bring to market really focused on AI as a service, essentially from your browser. When we work with enterprises, we're really trying to help all enterprises set-up computing, no matter where they want to do that. Whether they want to do it with the CSPs cloud, whether they want a hybrid environment, whether they want, on-premise. What you get with DJX cloud is now the opportunity for an instance in the cloud that NVIDIA helps with a full stack of software, a full stack of services, access to our models, access to both our engineering to help you do that, where we will pay the CSP for the instance. So we provide them the opportunity, but they can choose which CSP they want to host that out, and we will pay that piece.

There is also still an opportunity that enterprises work with the CSPs, but also purchase our full stack with the CSPs as well. So we can be selling with the CSPs, we can work directly with the enterprises and host have help them see their CSP instance equally. This allows them to have the most flexibility to be up and at speed as quickly as possible, as they are working and continue to work that they are working with us in terms of designing their models to designing their infrastructure.

Q - Joseph Moore {BIO 17644779 <GO>}

And in terms of NVIDIA providing -- NVIDIA created models around large language models, things like that. What's the scope of that? Would you -- does that put you in competition with your cloud customers in some instances?

A - Colette M. Kress {BIO 18297352 <GO>}

So our models -- that we are creating our models, whether it'd be a large language model can be anything from just a footprint for them to start their work on or we can actually help in the full design of how they do that? We do that today in terms of helping optimize their models for efficiencies and we'll continue to do that. The CSPs but also may host models, but we still have that opportunity working with many of

those customers in that model as well. So there's no difference, our work has over the years with the CSPs has continued to grow significantly their optimizing infrastructure in terms of standing it up. We're optimizing infrastructure for acceleration with the models that we're doing.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. Before I leave data center, the question of Inspur has come up a lot with investors, added to the entity list of Chinese server company last week. Can you talk to how much impact that could have to NVIDIA?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. So Inspur is a partner for us, when we indicated partner, they are helping us stand up computing for end of the end customers. And as we work forward we will probably be working with other partners, for them to stand-up compute within the Asia-Pac region or even other parts of the world. But again, our most important focus is focusing on the law and making sure that we follow exports controls very closely. So in this case, we will look in terms of other partners to help us.

Q - Joseph Moore {BIO 17644779 <GO>}

It's been a busy year for your lawyers, I guess, on these.

A - Colette M. Kress {BIO 18297352 <GO>}

(inaudible)

Q - Joseph Moore {BIO 17644779 <GO>}

Okay. Great. And then maybe if you could talk a little bit to software as a business model for you guys. The revenue generating capabilities of that can you just talk to how that scales out overtime?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. Our software business is an important part with our enterprise work. Enterprises are not staffed with the engineers that many of our CSPs are or even some of the largest companies on the world being able to do. So we have established a full stack of software, it doesn't start with just -- it doesn't start and end with just CUDA. It is all of the CUDA, DNN, it's all the different libraries, it's all the system software, everything that really gets you close to that application. And we've established both a full stack of software that we can license to enterprises, so they can start instantaneously with what they have already for x86 types of computing. They can now have that same capability, but all the needs that they have for AI, that is work that they need to influence, how they will incorporate into their applications that they use. Our industry applications, our stock actually helps put that all together for them. They still in that case, have access to the software, but they also have access to the service that we can do as we continue to improve software overtime.

Now, our software business right now is in the hundreds of millions and we look at this as still a growth opportunities we go forward. We've got three different areas that we focus on for software, both our NVIDIA AIE, enterprise stock that gives them a full stack to AI solutions, that they can use in the enterprise. We also use this for Omniverse that we can also sell separately. Omniverse is our great path to the 3D Internet and focusing in terms of simulating factories, manufacturing and works of such. And then thirdly, we will see going forward, automotive and our derived platform. We've heard from Mercedes in this last week, as they talked about their use of our drive software and sharing that software with us. So these are three key areas, but there is a lot more software opportunities and also our work in terms of software as a service or our overall AI as a surface will also be in addition to.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. So I wanted to ask about gaming and then I'll see if there is questions for the audience. Any sense of gaming sell-through now, you're at about a $7 billion run-rate sell through of your business. I think at one point you had talked about kind of a $10 billion run-rate for last -- second-half last year. But then, China has been weaker than that, but also you have a new product cycle. So just in general, how are you thinking about it seems pretty clear that you are under shipping end-demand now, but any sense of how much?

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. So we finished Q4, we feel really good going into Q1 that most of our inventory corrections that we needed to do was behind us or essentially our normalized channel, we'll probably hit us within this Q1. So our channel will be normalized, this gives us an opportunity to continue selling in our newest architecture Ada. Over the last quarter, we were able to launch our Ada and able to see the excitement of the gamers in terms of the new architecture, our 4090, 4080, 4070 Ti, all three doing really well during this period.

We watch our overall sell-through, and watch our sell-through to look at demand and we have articulated, probably about $2.5 billion a quarter, give or take the seasonality of any specific quarter. Our H2 is usually larger from a demand than our H1. So we feel we're on-track to really get to a position that we can actually sell-in closer and start getting closer to that and sell-through of what we sell-in and then grow from there, from a gaming perspective, that's all on track. Sure. We did have a little bit of volatility with China and the COVID that they had there, but again solid demand in gaming still absolutely exists worldwide.

Q - Joseph Moore {BIO 17644779 <GO>}

And as you think about this business, I mean, I look back over 10 years and you can see a disruptive influence in cryptocurrency in six of those 10 years and causing this kind of shortage (inaudible) cycle. if you don't have that going forward, is this just a much more steady growth type business that doesn't have that volatility? Or is there always going to continue to be somewhat more of a choppy trend, I mean, growth, obviously, but a choppier growth trend?

A - Colette M. Kress {BIO 18297352 <GO>}

Probably what is consistent over many years is the focus on gamers wanting the best infrastructure that they can have for their gaming experience, as well as the innovation that has come with that, whether it'd be programmable shading more than 10 years ago to where we are today in terms of ray tracing. These important pieces have fueled gaming, but gaming has transformed over that 10 years. Transformed to an entertainment industry, not just a gamer with his PC at home, it's bigger than that, it's a social platform as well. So I think we will continue to see this, the growth we believe for the most part, the key parts of crypto are behind us. And again, this is it for us to win the hearts of so many of the different gamers out there.

We have the ability through multiple platforms to also reach gamers, not only in terms of just the desktop, but you've seen the transformation of laptops, notebooks that can really take place and drive the same type of gaming that you would have on your own desktop device, that is really with the Max-Q technology, that is putting the top-end GPUs inside of a laptop and the influence that ray tracing has enabled as well. Ray tracing really helped game developers really add more realistic capabilities, it's infused with AI to help those game developers, build the best types of real games going forward. More than 350 games now on DLSS going forward. And we're going to continue to add more-and-more innovation. So yes, we think this is an opportunity for us going forward.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. So let me pause there and see if there are questions from the audience. I can keep calling up now. Sir, mic is somewhere.

Q - Unidentified Participant

Thank you. There is a report last week that Amazon Web Services doesn't have enough compute capacity for all the demand to launch some new AI products. So curious on the data center chip side, are you worried about running into shortages? Are you able to serve all the demand from hyperscale customers? Maybe just address that? Thank you.

A - Colette M. Kress {BIO 18297352 <GO>}

Yeah. It's a -- each and every CSP has had to go through capacity challenges and it's all different parts of the data center. It's not any one piece of it. As they are focused on building out the most efficient capacity, they can run into problems along that way. So we do know that there has been challenges and some of them are may not be ready to accept even GPU, some of those things that we saw at the end of Q4. But sure enough, we know that they will spend that time getting an understanding of when they will be able to accept it and start building out that. We're still in that ramping process of H100. Demand is strong. So there could be times, any week, any quarter that we have to be careful in terms of whether or not we're going to be able to make that strong demand. So we'll stay focused on it. But yes, demand right now is certainly at a important part for us right now and we've got a good process on supply, but let's see how it works.

Q - Joseph Moore {BIO 17644779 <GO>}

But when you hear that type of an anecdote, is your first thinking that's probably a Hopper challenge because it may seems like A100, there is more inventory out there?

A - Colette M. Kress {BIO 18297352 <GO>}

Correct. As we ramped H100, that is probably going to be the key area of our focus. But yes, we do have existing architectures that we can also sell during that.

Q - Joseph Moore {BIO 17644779 <GO>}

Question -- are there questions from the audience?

If not, you mentioned the Mercedes thing, and I know you've been talking about this for years, but if anyone hasn't, we've gone back and look at the Mercedes analyst day slides. I mean, NVIDIA is unlike every other slide like it's really impressive. The degree to which you've influencer program, they've essentially endorsed the economics that you guys have been talking about, which is economic split with $1 billion plus mid decade and multi-billion kind of end of the decade. Can you just talk about that relationship? And is that as you guys think about the automotive opportunity, how much of it is things like that? Where you're doing full system development, really good economics as opposed to selling cards or chips, which I know you also have a lot of those wins as well. What do you think the biggest economic upside for NVIDIA will be over-time as that opportunity develops?

A - Colette M. Kress {BIO 18297352 <GO>}

Our Mercedes work that we're doing is phenomenal. I think they both understood being a lighthouse leading company in automotive, the importance of AV for both safety and just the transformation of the automotive industry for software to be an important part of the car going forward. And so are working together for many years and been a part of their fleet, that will come to market in their calendar 2025 is an important piece. We continue to think about these types of business models and work with many other industries as well. At the same time that we're working with Mercedes for inside of the car, as you know, we're also looking at key lighthouse on helping them on their manufacturing floor.

Now you're talking about a very different business model as we focus on Omniverse. And what they can do to generate efficiencies on that factory floor. So the opportunity is always thinking about the industry, the industry's top players and helping them from end to end, that can be a great business model for us. But again we'll take all different sorts of that. There will be a case that we may sell somebody just the infrastructure. There may be a case that we may sell them the full software and they may change in terms of different components or different parts of the data center infrastructure. All of these are important in terms of building out our ecosystem, building out that platform approach, but nothing that we would turn away from, because it all just helps, get AI and accelerated computing out there.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. We have a lot of opportunities in front of you. Thank you so much for your time today.

A - Colette M. Kress {BIO 18297352 <GO>}

Great. Thank you.