Barclays Global Technology, Media and Telecommunications Conference

Company Participants

- Colette M. Kress, Executive VP & CFO

Other Participants

- Blayne Peter Curtis, Director & Senior Research Analyst, Barclays Bank PLC, Research Division

Presentation

Blayne Peter Curtis {BIO 15302785 <GO>}

Thanks for joining. I'm Blayne Curtis, U.S. semiconductor analyst. Apologize up ahead for the voice. It's kept active, (inaudible) and raspy. But I told Collette, if I run out of voice, she's just going to adlib the rest. But welcome to the first of 2 lunch keynotes, very happy to have from NVIDIA, Colette Kress, CFO.

Questions And Answers

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

I thought maybe a good way to start out here is, we're thinking of the broader audience that may not be as familiar with NVIDIA. This is the second day of our conference. And I'm sure themes like AI, like autonomous driving, robots, gaming are themes that (inaudible) from other companies. When you look at that, maybe a good way to start is from strategically, you have a core (play) technologies. But you're sort of increasingly even more on markets with AI. So maybe just talk about the strategy of NVIDIA and how these new markets are serving in with AI.

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

Okay. We've definitely changed as a company over the last 25 years. We are really focused as a company and a strategy as an accelerated computing company. The ability to take accelerated computing to 4 key market drivers, including gaming, pro visualization, data center including AI as well as autonomous vehicles. Those probably are very front and center. We have been known in each of these markets to be growing more than double digits across each and every single one of them. We tend to have the highest market share in each of these markets as well as also probably the best technology through and through. This has been through long work of really taking the GPU in the first overall killer app in terms of gaming that was very, very essential in terms of bringing high-end gaming to the market. But we

now have seen so many other extremely important use cases for this as we go forward.

The overall area of accelerated computing has expanded tremendously over the last couple of years because of 2 very, very important market changes. The first one of that is the ending of overall Moore's Law. Moore's Law has shown where the overall computing performance has been able to grow 100x over 10 years and is now coming down to probably a growth rate of only 2x over that same period of time. The use of therefore accelerated computing using overall GPUs has the ability to expand this and grow overall computing 1000x over that same 10x period of time.

Additionally, the increased use of AI and the use of an overall GPU to perform some of these very, very deep computational work has also expanded, which has really expanded our market. So we still believe we have tremendous overall market growth in front of us and extending overall portfolio for these markets.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

You know I can maybe generally note segments, the data center. This is a business that over the last three years has gone -- it's over a $3 billion run rate. You think -- last Analyst Day you talked about handling $50 billion, up from $30 billion, huge numbers. You look at $3 billion. Obviously, a lot more to go. You put the financial analyst hat on and you see the law of large numbers growth here. And whether it's 200% year-over-year and now it's 40% or 50%. And people are trying to figure out where we are on that curve. So maybe just your perspective. I mean, you have 2 parts of it people may not be familiar with. You have the training and learning and then you have the inference side. Where we are on the first wave of kind of the learning and figuring out what you want to do and then where is NVIDIA on the second part of deploying the inference and (inaudible).

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

So that's a really interesting question when you think about where we start right now in terms of the data center business. If you think about our actual performance of what we've seen over this last year, what we've even seen over the last couple of years, we're still, in terms of the number of servers that are actually deployed in the world, less than maybe low single digits in terms of that overall installed base. Equally in terms of when you look at the overall hyperscale. So the overall CSPs, in terms of their overall use of CapEx, we're still extremely small overall percentage. But we've reached a run rate that exceeds probably more than $3 billion a year in terms of our work of overall data center.

I believe it's continued to expand the overall workloads that we address, we expand the type of compute that we are doing. Just most recently, for example, we introduced the overall T4, which you have heard being so right for the area of overall machine learning, the overall use in terms of overall inferencing, which continues to overall expand our workloads. We have traditionally looked at 3 different types of markets that we have focused on in terms of data center. One, focused in terms of the work in terms of supercomputing, high-performance computing, the work in terms of WISP, CSPs and hyperscales. Those 2 have now expanded with the addition

of the overall T4 to address the overall enterprise market, as we enter into expanding to new markets, such as machine learning and data analytics as well.

So we now have the ability outside of just supercomputing, CSPs to hit the overall enterprise, to expand the workload past deep learning, machine learning, supercomputing and mostly in terms of these new workloads. The T4 is essential in terms of both it bringing what we will consider to be a scale out overall infrastructure, it is our second generation of using our overall Tensor Cores. So now we have the ability for overall mass volume for the hyperscales and for enterprises using the overall T4.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

Even as it's -- the T4 is your inference pass through that could be plugged into (inaudible) server. And to date, you've seen the deployment of inference just being done on a CPU and you can do it, it's not going to be deficient. But you're still early days. And I've begged -- everybody has been waiting for this, when is this inflection point going to happen? Or (inaudible) make a decision to put that extra processor in the servers and as you look out into next year, is that something that you think will start?

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

Absolutely. The T4 is very important and is actually doing quite well already. As you know, Google has -- is one of the first to adopt it and put it in terms of its Google Cloud platform. And the other overall CSPs are working quite fast to actually move to also incorporate T4. Establishing additional spaces within data centers for the mass use of the overall T4 is clearly in terms of the next direction that we'll go. You have both a form factor and an absolutely solid overall software stack in order for them to deploy that overall T4 throughout their overall data centers, throughout their overall workload. We do believe that the hyperscales long term will likely get to almost every single one of their overall servers in the data center will be accelerated.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

Even as the other part of inference is, there's the data center and then there is edge applications and you'll play in some of those. You've been a leader in autos. You're always very supportive of auto bus tours. So thank you. On Tuesday, we saw the head of that business. And I felt like the message was a little different in that everyone has been focused on level 4, 5. It's taking longer to kind of deploy those. But there is this wave of, whatever you want to call it, C-plus and Tesla-type autopilot that actually may be coming quicker. So just kind of your perspective on that shift and what that means to NVIDIA?

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

Yes, I think it's important to understand whether it be looking at a robotaxi, which is focused on the level 4 and level 5 or looking at a level 2. What is very, very clear is the amount of compute that is necessary to complete either one of these sizes of levels in terms of autonomous vehicles. We have learned that over the years we have been very supportive of all of the different OEMs and the Tier 1s that are out there.

What is most important is we have now an architecture, we have delivery of a full solution and an end-to-end autonomous platform that they can use in terms of both testing and development today, referred to as Xavier, as well as in terms of with its overall drive software that's incorporated in there.

That is essential in terms of the work that they need to do to complete that testing and complete the overall development work. Now whether or not the level 4, level 5 happens first, whether or not the high-end level 2, there are 2 streams that are absolutely happening. The level 4, level 5 is a much more complex type of problem. It needs the overall redundancy for it to be able to function no matter where the car is because it would have no driver, it would have no overall steering wheel.

The work in terms of level 2 and high-end level 2, we were the first to bring level -- high-end level 2 to the market. Now we have a platform here and our very first win with Volvo in terms of mainstream passenger cars is also in front of us. So we know that both of these are competing in terms of their time frames. But testing and having the actual silicon today to actually do that work is absolutely instrumental to being sure that we can be in the market in the next couple of years.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

Maybe it's interesting and you see (inaudible) you also have a full, particularly in auto, the entire full stack. You have your own fleet of autonomous cars, you have your own teams labeling. Can you give a perspective as to what exactly your range of offerings are versus maybe if you wanted, you and OEM, you're working with a Tier 1 and you wanted to go it alone. What do you bring to the table versus what others have?

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

Yes. It's correct. We have a full stack. We have the ability to provide many different configurations and even on the overall compute standpoint. We can create a compute platform that is with a SoC with a GPU, you could have 2 SoCs, 2 GPUs. It depends in terms of how they want to overall configure inside the types of cars that they're doing. But even additionally there, we can from end-to-end complete that full stack and take on that work holistically in terms of the overall software. In many cases, we are likely partnered with them, working jointly in terms of how they want to define autonomous vehicles for their level 2s or for their robotaxis we'll be working with them.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

I think it was interesting in the meeting that (Rob Saunders) headed autos with them, actually now you have a lot of other edge inference markets. You spend a lot of time just talking about whether it's trucking or actually last-mile delivery. Can you maybe just give us -- I think there is maybe 3 silos here, there is auto, there is this kind of robotics, there is even health care to talk about. Can you kind of outline these opportunities and where you are seeing new things pop up?

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

Yes. We believe the world of autonomous machines, as we move forward, is going to be a material amount of overall compute necessary and very essential in terms of how these machines function in terms of the future. So we have developed as 2 different platforms to address those markets. We've talked about Xavier, which is here for overall autonomous vehicles. But we also have our Jetson platform that we are centering around robotics, centered around drones, centered in terms of any machines, whether those be medical devices scanned and the work that they need to do with that. In that case, we can look at the overall AI required for them to be autonomous as being part of an open source platform that we have versus the Xavier platform still functioning in terms of our algorithms with overall CUDA. So both of those markets, we know, have a tremendous amount of devices. We will continue to see and broaden the arena for this overall market. But, yes, we are addressing both of them.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

And putting your developer hat on in terms of when this may actually impact the models, some of the use cases, whether it's within a warehouse and it's automated forklift, I get that, that makes sense. Some of these use cases that, I mean, you're saying, like, in China, you could get these things last mile, seems really too futuristic. So (inaudible) is another one where it makes total sense. It's the better solution. So when you think about all of these opportunities layering in, is that a next year story or you think it's a little bit longer term?

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

I think the work that is happening is realtime. Even though it may sound futuristic to think about that last mile or think about trucks driving across the U.S. by themselves in the middle of the night, there's a tremendous amount of work happening, with many of the key overall trucking industry to think about the delivery. The absence of trucking, those that can actually be employed to deliver so much of the goods that we are doing in the cyberspace is a very, very large problem that they need to address.

The overall safety concerns are also front and center in terms of what they want to do. So what you're going to see is a lot of work in terms of the development to manage the overall computing that needs to happen. That means they need to practice with a significant amount of data. That means we're addressing them with work in our data center in terms of helping them with infrastructure that they are able to manage their data, they are able to mine their overall data to come up with a solution that will therefore be on the roads in the next couple of years. So we do see it today. Maybe the end consumer doesn't see it today. But from a development side, we're busily working with them.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

(Honestly), on the competitive landscape. And it may be more going back to the data center. And I've been doing (semis) a long time and I haven't seen this much investment in one particular area and we (do the best story) we had last month. And I think there's nearly 50 companies that have seen significant funding and are all attacking this $50 billion opportunity. I think it's also interesting, you've seen some

vertical in some of your customers, Google is the first. Amazon last week had an announcement of a solution. And maybe can you think about -- Google is a great example. They have their own. But also one of your largest customers. How that has worked out? And maybe if not, maybe there's multiple solutions can exist at the same time?

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

Yes. The ability for an underlying architecture to get overall mass scale is its ability to be spread throughout an overall data center and the consistency across. We have the ability now with our overall data center products to address almost every single type of workload that would be available in terms of the data center and with a consistent overall software platform. When you think about the amount of CUDA downloads, the amount of developers that we have, both writing to the platform as well as also writing additionally to the overall applications they may work with, we are expanding for the rapid movement that you are also seeing in this industry.

When we think about our work over the last five years and what we began in terms of AI at that time and the workloads that we were working on at that point and the expansion that we've seen, starting additionally with just image detection to the advancement of video encoding, the advancement in terms of natural language processing, the ability to even come out now with recommendation engines, the overall workloads continue to change. We also get the ability to add more and more workloads collectively together into one overall AI solution. That is going to be essential. A simple one workload may not last in its form factor very long to where it has to have that continuous change. Our overall both platform and software together allows us to feed so much of that market.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

If we then change gears to the gaming side, you launched ray tracing, one of the biggest advancements in your gaming history. When you look at -- you're now working through some inventory in gaming. So I think maybe if you can just give us a perspective of -- I mean, people may not even know what ray tracing is. So maybe quickly what even that is. Then if you look at that generation, how are you feeling about the sales of that ecosystem of games? I think there's at least one game that supports it. But I think it takes time to get software that can actually use the functionality you've added in. Then on flip side, you're working through and you were (upfront) about the amount of the older generation you're working through. What's the status of that?

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

So let's first start with Turing. Extremely excited about the architecture that we launched in terms of Turing. Turing brought ray tracing to the overall gaming as well as in terms of the overall pro visualization business. What do we mean by ray tracing? The overall ability for graphics to look about as real as they can is the overall use of light, the overall use of shadows. The better the overall lighting and shadow, the more three-dimensional, the more realistic those appearing.

We were talking about bringing ray tracing pretty much to the same thing that we had done probably more than 10 years ago when we brought programmable shading into the overall GPU and brought it into the overall graphics world. Ray tracing essentially is something that we have assimilated over these last 10 years and now we are doing it in realtime. The realtime of the reflection of light, the use in the bouncing off of the rays of light within the overall pictures are done in realtime. This is probably something that people didn't feel that would be brought to graphics for more than 10 years in terms of the future.

So bringing it right now to the market and incorporating it with so many of the games and so many of the enterprise applications is an exciting factor for the next generation of gaming that we have. We bring it together in terms of Turing. It continues to outperform into the existing games as well. Even overall Pascal architecture it's 25%, 30% better in terms of what we have. It's even better in terms of -- astronomically 10x better than our overall Maxwell version as well. We have tremendous amount of gaming ability to upgrade to the overall Turing architecture to take place for existing games. But also to buy into the future of all the new games that will come out with ray tracing.

In the future, we'll be probably in a space that there will be the haves and have-nots, the games that have the overall ray tracing, the ones that look so photorealistic, almost like film, to ones that will not. So you get 2 things with that, the increased performance today but also the generation of ray tracing that will come forward.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

And in terms of that catalyst, I'm just curious. You launch products at the high end first and then it trickles down. The highest volume SKUs are not probably at the high end. You're going to eventually launch more mid-range products and that's why you're clearing out the inventory. I'm just trying to understand, when you look out over the next year, is the catalyst really to get the titles and get the demand for ray tracing? Or is it just a function of getting that mid-range out there which will be the bigger volume?

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

It's a little bit of both. We do have to work through our mid-range Pascal overall inventory. Prices following the end of cryptocurrency were just too high for too long of a period of time. They've normalized now. And we have got a great set of overall promotions set up for the overall holiday season and even beyond. So we feel that this is a finite amount of time for us to work through that overall inventory, correct, for us to continue our path of bringing great products to market and restarting what we think we need to do in terms of the mid-range. It is a combination of extending out the new architecture, the additional performance. But also breathing into the overall ray tracing that we can bring forth, too.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

We have actually a few more minutes. The other area where ray tracing you've done a great deal, you have pro viz and people also may not know what exactly that is. But (inaudible) numbers, how do you drill that down into what actually the TAM is for the

pro viz? And you start talking about rendering video and such like that applications that you can now enable in realtime. There's probably light-feature applications we're don't even know about. But you can -- wrap that into what the opportunity is and how that will benefit you financially?

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

So when you think about the market of pro visualization, this is what we refer to as our overall enterprise graphics. There's a tremendous amount of work, whether it be regarding TVs, commercial, film industry, what we do in terms of photoshopping, creating catalogs, a lot of that work is actually rendered graphics or special effects in terms of that. In order to develop the overall special effects that we see in so many of these end use cases is what we refer to as the rendering process. So they are rasterizing all together the overall graphics that have been built.

With the overall ability, with overall ray tracing, we have now been able to take RTX to a part of the market that we haven't addressed before, where essentially the work in terms of the special effects takes place in what we'll call an overnight or the next day where they send it to rendering farms to come back with the full frames fully completed and then do the editing from there. What we've been able to do is insert ourselves into that process and address what we're seeing right now is about 1.5 million overall servers or a several billion dollar overall business but it's generally just been overall CPU farms. Transforming those to overall GPU massively improves their work, massively improves their time in order to create so much of this overall special effects and improve in terms of the overall quality of what they can do. So that's how we are -- overall size in this business.

Q - Blayne Peter Curtis {BIO 15302785 <GO>}

Thank you. With that, we're out of time. And my voice lasted. I appreciate you joining us.

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

Thank you. So much.