JPMorgan Tech, Media and Telecom Conference

Company Participants

- Colette M. Kress, CFO & EVP

Other Participants

- Harlan Sur, Senior Analyst, Research Division, JP Morgan Chase & Co
- Unidentified Participant, Analyst, Unknown

Presentation

Harlan Sur {BIO 6539622 <GO>}

Okay. Why don't we see if everybody can take their seats so we can get started here? Okay. I think we need to go ahead and get started here.

So good morning. And again, welcome to JPMorgan's 45th Annual TMT Conference. My name is Harlan Sur. I'm the semiconductor and semiconductor capital equipment analyst. I'm very pleased to have Colette Kress, Chief Financial Officer of NVIDIA. And this is a special event because, I believe, this is the first time ever that NVIDIA has presented at our conference, as it's always been too close relative to their April earnings conference call. And so the good news is you had your earnings call a bit earlier. But the other reason why this is sort of a treat is that NVIDIA held its Annual GPU Technology Conference, or GTC, a couple of weeks back along with its Investor and Analyst Day.

So I've asked Colette to maybe start things off with an overview of what the team showcased back at GTC. And then we'll kick off the Q&A. So Colette, thank you very much for joining us today.

Colette M. Kress {BIO 18297352 <GO>}

No. Thank you for having us. And you're right, this is the first time that we've been here. And we're pleased to join the large group. And let me kind of start and kick off.

Last couple of weeks have been extremely, extremely busy times for us. We started with a small thing, which was announcing earnings for our First Quarter. And then we also held our annual GPU Technology Conference. This is our time that we pull in all of the different developers and ecosystem partners around the world and focusing on the latest and greatest in terms of GPU acceleration and computing. It's not similar to what we do in some of our other conferences where we talk about specifically gaming and the latest and greatest. But it's really, really focused on GPU

and GPU computing. A great event for us, probably our highest attendance that we have seen in the eight years of running. And also an extension, given that, in this last year, we had also used our GPU Technology Conference to go worldwide. And we held 7 other events worldwide in the fall, talking about the importance of GPU computing. This event brought industries from afar, everything from the top hyperscales in the world, top health care, top auto manufacturers and then some, many, many other industries represented at this conference.

So let's kind of talk through some of the key highlights in terms of what we announced at the overall conference. As much talked about and anticipated was the involvement of our overall Volta architecture. And so we came out with our latest in terms of understanding of our Volta architecture and announced our V100. V100 is our next GPU for the datacenter. So perfectly engineered for what we're seeing right now in the expansion of GPU computing, the expansion of accelerated computing with the key overall application in terms of for AI. We see a tremendous amount of interest, I think you see it in many of the events that you go to, of the importance of AI that will probably be the key transformational computing evolution that we're going to see over the next decade or 2 decades as we go forward. And the use of a GPU has been perfectly matched for some of the key problems that we're trying to solve in overall computing. With the overall slowdown of what we've seen in terms of Moore's Law and expected to go forth, the overall GPU really enables the ability for both parallel processing and overall expansion of your throughput by using a GPU and its parallel processing to do so.

So Volta 100 is one of our biggest undertakings, probably the most transistors and improvement in transistors from architecture to architecture that we have seen engineered for both deep learning training as well as for inferencing. We're very pleased with -- to have that available in the next couple of quarters coming forward.

Secondly, in terms of our announcements, our DGX AI supercomputer, our next-generation. We had launched that about a year ago. And that will be coming available on Volta as well. We are allowing the opportunity for those that buy DGX on our current architecture, Pascal, today that will also get the ability to receive Volta when it's available. So the seamless transition on the overall computer.

We also announced DGX Station. DGX Station is for those in research and R&D that is using overall GPUs and wants an actual server at their overall workstations. We have an ability, in terms of a form factor, of a mini AI supercomputer for you. That also allowed us to transition, to extend what we are doing with our AI supercomputers, which are complete end-to-end systems. End-to-end systems that are based absolutely on our overall GPU. But also the significant amount of software that is necessary to prime what you need for overall AI. We support all different frameworks that are currently running and are popular in the world. And think about AI worldwide, supporting what AI frameworks are here being used in the U.S. as well as those that are using in many of the other countries, including overall China.

So our AI supercomputer. And thinking about that evolving and the use of the cloud to also enable AI, is the introduction in terms of NVIDIA cloud services as we go

forward as well. We'll be able to extend exactly what people are working on in terms of with their AI supercomputers or what they're using with a Titan, what they're using a workstation. And extend that from an on premise to a cloud as we go forward.

That next took us to another piece of our overall GTC conference, where we talked about our latest developments in a key market industry where we're focused on for AI, which is automotive and autonomous driving. We announced a key design win in terms of with Toyota. Toyota, we've been working with for many, many years. And you should think about them as probably one of the key auto manufacturers, very, very focused on safety and overall quality in terms of what they use in terms of their car. So we're very excited in terms of that partnership and the future of autonomous driving in terms of working with Toyota.

So those are some of the key highlights that we announced at GTC. We followed that up with our Investor Day to better discuss the overall business models that we have for each of our 4 businesses. To remind everybody, we still have 4 key growth platforms that we're focused on: Gaming, Pro Visualization, high-end workstations, Datacenter and also Automotive that focuses both on our infotainment and autonomous driving cars.

Questions And Answers

Q - Harlan Sur {BIO 6539622 <GO>}

Great. No. That was a great overview of GTC and the Analyst Day. Looking back at the history of NVIDIA, it's interesting because the team has continuously reinvented itself. First, it was PC graphics with a team, started in the late '90s. It was the -- there was your parallel processing GPU platform that got you much success. 10 years later, kind of mid-2000 time frame, it was the move to GPU, GPU compute with your Pro Vis business and early entry into high-performance computing. Fast-forward to today, it's adoption of your parallel processing architecture into things like AI, deep learning, automotive. And so all of these dynamics, gaming, GPU, compute, AI and deep learning, are sort of all firing. Can you help us understand -- you've driven a 20% revenue CAGR over the past three years. Help us understand how you think about the growth profile for NVIDIA going forward. And maybe how does this profile, from a mix perspective, change over a period of time?

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

Yes. A lot of questions inside of that. But I think it's important to reflect in terms of where we are today and the speed of transformation, the speed of really thinking through how the evolution of computing will overall change and how we've adapted but also led, in many cases, some of the areas of the future. If you think about even 10 years ago, when we brought to the market overall CUDA. And CUDA is our overall development language that enables you to overall program the GPU, people were very confused and couldn't understand, why would you need that? Why would you enable it across every single one of your GPUs, whether that be a gaming, a PC or, in terms of the future, in terms of what we're using? When you think about that investment and decision more than 10 years ago, it seems perfect to look at in terms

of how GPU computing has now evolved and the importance of being able to write very, very specifically to the GPU and extracting the overall APIs to our overall software and applications that we're seeing. No one would have maybe predicted or would have thought to come forward about the slowdown of Moore's Law and the need of parallel processing in terms of uniquely finding the use of the GPU outside of the CPU's work, which is ever so important today, that, that expansion role would do. But our work right now is to continue to evolve into the industries that would be best in need of that additional overall compute capacity and focusing on that. Another way of saying that is, it's not to say -- to be a generalist across here. We are here to support all of the different frameworks. We are to support all types of AI. And where we've gotten very specific in terms of our overall software going forward is also in terms of key industries, which have, therefore, developed into key markets that we're going after as well. Markets that we think we're uniquely positioned to solve the problems that are very both sticky in nature, challenging problems. And then we think that we are probably the best ones to overall solve that. That would be what you would probably see in autonomous driving. Autonomous driving, probably 4, five years ago, we started to clearly understand that this was a computing challenge. This wasn't a problem that would be solved with smart cameras. This needed a significant amount of compute. And many of the partners that we had worked with over that time also saw that importance. We now know we're leading a big part of the industry's thought in terms of computing. And we've announced additional partnerships and design wins over the time. We're still in the very early stages of autonomous driving. There is 1 or 2 key customers on the roads today. But we know this will be a very big change as we think about 10 years from now. 10 years from now, will we have the same ability to drive our cars or will they be just self-driven? So when we think about that and we think about our growth, I think we've laid out a strategy across the continued growth of gaming, the significant expansion that we see in datacenter, the opportunity in datacenter and a brand-new, what we refer to as a start-up, in terms of autonomous driving that will lead us as we go forward. Being able to pin it down to an exact number in terms of growth rate, I think, is too challenging, or I think another way of saying it is we'd probably be wrong. And I think more focused on doing the best that we can to leverage our infrastructure, leverage our investments today to continue to expand those markets that we think are very large and a great opportunity for GPU.

Q - Harlan Sur {BIO 6539622 <GO>}

Well let's expand on that. And let's take that, the revenue growth. And turn it -- let's bring that back down to the margins. Gross margins for the team have expanded 450 basis points over the past three years. Your operating margins have expanded 1,600 basis points over that same period of time. Cash flow has doubled. But the team has never really put a stake in the ground around sort of margin or free cash flow targets. But from what you described at Analyst Day, team will continue to drive operating margin expansion as long as revenues continue to grow. On the gross margin front, though, I'm assuming that, as you continue to expect more of the incremental growth coming from your Datacenter businesses and your high-end enthusiast GPU businesses and, at some point, fully autonomous. But -- that as you expect to grow from these higher value-added segments of the market, that gross margin should continue to expand. But is there any way to think about sort of incremental margin? We always tend to think about, is there any way that we can sort of, back-of-the-envelope, think of some way of thinking about it? How should we

think about margin expansion as your revenue -- gross margin expansion as your revenues expand?

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

Yes. So we talked about a little bit, this, at Investor Day but let me try and expand a little bit more. So our gross margins. Gross margins is a very common metric in terms of what is used in terms of the semiconductor market. We love to look at what is the cost of the silicon. Let's talk about in terms of the ASP, what type of margin. But let's remember in terms of the transformation of where we've traveled as well as where we are going, meaning when you think about the value that we are delivering and the overall value that we have provided to our overall customers, yes, the hardware is perceived to be best-of-breed. The hardware is expected to be. But it's only one piece of it from the overall silicon portion. That significant amount of software and its presence of software has added a significant amount of value in terms of how customers use the overall TCO that they have in terms of using our products, getting work done tremendously faster and speeding up the ability for AI as we go forward. That means the investment that we have invested in that is essentially in our OpEx, essentially the development teams, the hardware and software teams. Software is now more than our overall size of our hardware development for the company. That's a pretty big state for an overall semiconductor company base in terms of there. So you're receiving that in terms of in the gross margin. Sure, we believe, if we continue down the aspect of improving the value that we can deliver with the GPU then computing and more and more time and resources on that software, we will again be able to, overall gross margins, expand across a lot of our different overall platforms. So then that comes into the case in terms of, how do we want to think about that from an overall OpEx and an OpEx investment? How much would we want to think about? We've outlined what we think we'll do in terms of OpEx growth over the next couple of quarters and horizon. We wanted people to remember that this is the opportunity for us to invest, for us to extrapolate the potential TAM and opportunity that we have in front of us. So growing in the high teens in terms of OpEx, we think, is a very, very good position for us, even as we focus on efficiencies or we think about our ability with a unified architecture that we have to use those investments overall widely. Our overall goal is that will benefit us from our overall operating margins, leveraging that work that we do once across many different overall platforms and, therefore, also being able to grow operating margins. It's not clear, though, that we'll do that perfectly. It's not something that we can sit here and align on. But we do see the great opportunity of the revenue growth in front of us and the TAMs that are there. It's out for us to go and get. And we will do our best just like we've done our best in the past.

Q - Harlan Sur {BIO 6539622 <GO>}

Yes, let's stick upon that, because that's actually an interesting point about the kind of near-term higher levels of R&D spending, right? The team has spent $11 billion in R&D over the past decade. You're investing, as you mentioned, more in software developments, relative to chip or hardware design. Datacenter is a great example where you're not only driving innovation of the platform itself. But you're actually helping to build the actual ecosystem, right, for your customers. And so what part of the R&D spend is actually spent on things like ecosystem enablement?

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

Yes. It's a very good question. When we think about our breakdown of just operating expenses alone, our R&D probably being the largest component of where we spend our investment. And that R&D takes on many different type of pieces, both in terms of enabling a GPU to just come to life. But also the work that we do with gaming developers, what we do in terms of enterprise partners and seeding the ability to work seamlessly with both new games that are coming out, new overall enterprise applications, from CAD, from Adobe, others, that, really, if you think about our customer set, they expect on day 1, when they use a GPU, for it to work seamlessly, both forward as well as backwards, as we think through the overall evolvement of software applications. And we will continue to focus on that. It's a differentiator that we have. It's our brand loyalty that we have built. And it is even more important if we think about the evolution of AI, being able to spread AI throughout the world as quickly as possible, very different than many things that have come to market over the last 20 years, the use of the cloud, the use of really thinking about a view that not any one framework should be solidified. But we will support all of them. And we'll endorse anything in terms of a provider that's out there and support them. So our overall goals for AI is to quickly bring it to market. And I think we have really proven with a lot of what we've brought to market to continue to advance that ecosystem.

Q - Harlan Sur {BIO 6539622 <GO>}

Let's talk about shareholder return. Over the past four years, the team has returned more than 85% of their free cash flow back to shareholders. Team is on track to return another $1.25 billion to shareholders this year. Should we assume roughly 85% free cash flow return over a broader time horizon? Or how should investors think about capital return metrics for NVIDIA?

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

Yes. So we take a very serious effort to focus on capital return and focus on also shareholder value over this period of time. Not only our capital return. But our overall performance on the P&L has provided really, really great results over this period of time. But when we think about our capital return, it is really making sure we've made the right investments that we do need to in the business, whether that means direct investment in terms of people, whether that means capital investments or really working on investments in terms of our partnerships and overall ecosystems. We're going to make sure that, that is front and center. So capital return will always be what we, therefore, can provide of our free cash flow back to shareholders. It is what we've given in terms of an absolute numbers. It eases rather than discussing the individual overall profit levels. But we will try to be the best-of-breed in terms of the highest percentage return. I think 85% is quite commendable against a lot of our peer set. And we'll look to leverage that cash the best. And if we think that best area is with our shareholders, then that's what we'll provide to it.

Q - Harlan Sur {BIO 6539622 <GO>}

Got it. Okay. Before I jump into some of the product categories, do we have any questions from the audience? If you do have a question, just -- if I could just ask you to please wait for the microphone. Right up here. We have one right up here.

Q - Unidentified Participant

Your CEO spoke about educating the developer workforce in one of the last calls, that it's been a challenge. Could you give us some guidance or some idea of how that's going? And are you able to get the information out adequately?

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

I think we've done a lot of work in terms of continuing to enable. If you think about CUDA and how much time we've worked in terms of teaching those that come out of higher education, very key in research houses, which have now moved to many of the key start-ups or large conglomerate enterprises that we see today, that work in terms of teaching CUDA throughout that 10-year period has been very important. Most recently now, those that have been advanced in terms of CUDA learning, which is over 500,000 developers worldwide, we have now moved that to focus on deep learning institute and deep learning training. We are looked at now as a place to also be taught to learn from NVIDIA in terms of what we've seen in deep neural networks, what we've seen in AI, how people are also using this for additional frameworks. So even ahead of our overall GTC Conference, more than 10,000 people being led in terms of there. We will likely, before the end of this calendar year, get to 100,000 people over this very, very short time, where we began this late last year, in terms of teaching them in the specifics of deep learning. An institute that allows you to not only hear firsthand what is deep learning and give the basics of it. But also, at the end of that development, that you truly understand how to build a neural net. You learn how to write your very first algorithms on either face recognition or video encoding or you can even do image detection over those. So I think we're on a really, really solid path. You saw a tremendous amount of those people at our GTC. And I think when we have our additional GTCs as we go forward in the year as well as the Institute here -- excuse me, back in Santa Clara, you can also be on premise with learning [ph] that as well.

Q - Harlan Sur {BIO 6539622 <GO>}

Any other questions? We have one right here. Just give us one second for the microphone.

Q - Unidentified Participant

On the last conference call, you guys talked about, in the Datacenter AI business, expanding into other verticals. So expanding to financial services and energy were 2 that you called out. Can you just talk a little bit about, for those clients, how far along they are in the process? Is it like sort of one team worked on this to try out. And they might expand it? Or has it gotten more widespread adoption in those industries? Or just how the clients are using that technology?

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

Sure, sure. So the question is really about our Datacenter business, our clients, our customers. What are you seeing? And how are they using these GPUs in terms of there? Let me start by first understanding what is within our Datacenter business. Our Datacenter business, in our Q1 results that we just reported, topped over $400

million and nearly tripled from a year ago. It grew sequentially about 38% as well. So it is -- it's been moving quite fast. A lot of discussion in terms of what are those customers. The customers within there is every hyperscale on the globe uses overall GPUs. They use them both for their internal use and research and applications that you use every single day as consumers. But they're also creating cloud instances. Cloud instances where many of the enterprises around the world will now have the immediate availability of a GPU instance. You begin their work on deep learning or overall virtualization, streaming, many other uses of overall GPU computing. So our business is made up of 5 different overall businesses: Number one, our traditional high-performance computing business that we have been working in for more than 10 years. This is where they are using accelerator in many of our largest supercomputers around the globe. Right now, our supercomputer universe, probably 30% of them use acceleration in order to improve the overall throughput of those large computers. And 70% of them use GPUs in those situations. Our overall high-performance computing business is also growing quite nicely and nearly doubled year-over-year from where we were a year ago. Our second business in there is essentially what we are selling for the overall AI area, whether that be deep learning training or whether that be overall inferencing. Our overall deep learning training has been the start of our AI business. We have a -- quite a good presence of training around the world where they are building very large deep neural nets in order to do the initial stages of training their overall data for overall applications. This started initially a couple of years ago with our overall hyperscales. And it's continuing to expand out, expanding out into industries if you think about the work in terms of automotive, the work in terms of manufacturing places, such as financial services, those in the room here as well as in e-tail, where there is a significant amount of data. Another way that you should think about deep learning training is follow the data. Where is the significant amount of data that they are looking to extrapolate additional and -- use of AI to solve many of the problems that they're looking at? The overall goal in terms of inferencing is our third business. Inferencing is still in its early stages of overall GPU use. That has traditionally been a CPU business. But the overall ability to work with GPU training and higher-end type of inference is very key and important and will also be continuing to expansion. As you think about many of our offerings in the Datacenter business in terms of Pascal, we're also enabling our inferencing. Our fourth business is our GRID business. That is a business where we are one-to-many, one GPU to overall many users. So we've virtualized approach to the GPU. This is being used where workstation use in the cloud or when they want the overall security protection of a PC in the cloud or you're looking in terms of streaming games or other types of high-end applications. They would use an overall GRID. And that business has been growing quite well as well, more than doubled last year. Then our last business is our DGX AI supercomputer. Those came out about a year ago. We're continuing to sell those into enterprises, automotives and well as now into some of the hyperscales, as you just saw Facebook purchase a string of about 128 of those all strung together for use in terms of in their datacenter. What these AI supercomputers are is they're a contained version of our overall GPU computing, a full end-to-end system, software-complete plug-it-in. And that enables them a quick start and the ability to put containers and containers together. We also saw the manufacturing arms in the Asia Pac area also look to string together more AI supercomputers. So our business is continuing to broaden, both from a breadth and depth. Our customers aren't just buying for their first projects. Some may be, some may be buying on the fifth, 10th, 20th project that they've been

overall working on. But there is a good amount of working on enabling the cloud. So more have that easy access. Something that wasn't necessarily there 20 years ago is now a great way to move quickly. And the adoption of AI will probably happen faster than many technologies did in the history.

Q - Unidentified Participant

We've seen very consistently over the last 30 years that the higher the opportunity in Silicon Valley, the higher the risks. And I'm hearing a lot of exciting opportunities. And I'm also recalling specific instances that you know. And we all know, about not managing through those risks, slipping customer expectations. We might start with my asking you, what percentage of your R&D is in the manufacturing area so that you don't have faults? And which is the bigger risk to you: customer expectations or external risks from others? How do you manage for the risks is really the whole thing.

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

Yes, I think the company spends a good amount of time considering all of the aspects of some of the risks that you outlined: redundancy and thinking about both where our development is and making sure that we can support no matter where we may be located. We do a significant amount of our development in the Santa Clara, the Silicon Valley area. But we do also a good percentage in the Asia Pac overseas area. Close in terms of where our silicon is manufactured. But we also do a significant amount of software development overseas. So we do have a good balance to minimize the risk of it all being completed in one area. In terms of the risk, in terms of -- for overall customers, our ability to have supported many of these customers over the 24 years of life that we've been here has established our overall presence of a very, very solid product, one that we will continue to support through multiple, multiple generations. So the customers continue to evolve. But some of our same customers that were with us 24 years ago are still our customers today and will likely be our customers as we go for the 10 to 15 years. What has been new is the expansion of all of the new customers that we are now supporting. But what is unique is our overall philosophy of how we're building it with a unified platform that allows us the consistency in terms of the product and the quality that we are putting out there. If you think about the quality level that is required for Datacenter or the quality level that is required for automotive, it is large. Being eight years plus in the automotive industry, we have gone through the cycles of how important your product must meet those overall standards and the standards to the utmost. And I think we're going into a generation of autonomous driving where the expectations are probably even higher in terms of how successful that is. So all good balancing acts are there. But this is not a speed. This is a thoughtful speed that I think we are in, both for datacenter, automotive and the future of overall gaming, to assure that our products all take in terms of those risks and that we keep the loyalty of our customers and the brand promise that we have.

Q - Harlan Sur {BIO 6539622 <GO>}

One question up here.

Q - Unidentified Participant

Would you be able to give us a dollar amount of your technology that would be in the typical autonomous car, either the software chips and -- or maybe separate from that, entertainment, too?

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

Yes, let me start first with maybe infotainment and maybe where our hopes and aspirations are in terms of autonomous. When you think about our infotainment business, which is the lion's share of our automotive business today, we have focused on the premium infotainment within many of the higher-end cars, where they truly appreciate the overall visual computing capabilities than the overall software that we've been able with those infotainment systems. On average, we are about $50 to $100 per module per car in those scenarios there. When we think about autonomous driving and we think about the importance of the software that we are designing with many of these car manufacturers and that overall programmability on algorithms that we'll necessarily need, we think that's going to be a very important value-add that is very hard to replicate and a lot of time on our engineers to do so. Our thought is that we will double, more than double, our overall averages piece [ph] if we go into autonomous driving is probably our initial thoughts in terms of there, maybe even more. But that's our initial thoughts.

Q - Unidentified Participant

(inaudible)

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

A dollar amount, if we're currently at $50 to $100, a dollar amount there. As you get up in terms of the additional levels, Level 3, Level 4, Level 5, I think it's still just unknown. Unknown in terms of where this will evolve to. I think it's still way in the early stages to see.

Q - Harlan Sur {BIO 6539622 <GO>}

All right. Well we are out of time. Colette, thank you very much for joining us. Hopefully, the team will be able to attend in the future years.

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

We'll be back. Just don't move your date.

Q - Harlan Sur {BIO 6539622 <GO>}

Thanks, Colette.

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

Thanks, Harlan. It's good to see you.