Citi's 2021 Global Technology Virtual Conference

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer

Other Participants

- Atif Malik, Citi

Presentation

Atif Malik {BIO 7312618 <GO>}

Good afternoon, everyone. Welcome to Day 1 of Citi's 2021 Virtual Global Technology Conference. It's my pleasure to welcome Colette Kress, CFO from Nvidia today. The format of our discussion is fireside chat. I'll kick it off with my questions and then we'll do the questions from the audience.

My introduction, I'm Atif Malik, I cover US Semiconductors and Capital Equipment Stocks here at Citi. Welcome, Colette.

Colette Kress {BIO 18297352 <GO>}

Thanks so much for having us, appreciate it.

Questions And Answers

Q - Atif Malik {BIO 7312618 <GO>}

(Question And Answer)

Colette, thanks for coming to the conference. Your keynote is always a big draw for our investors and I always find new and exciting topics to talk about. So, let's start with the Data Center business first. Data Center sales grew strongly 16% sequentially in the July quarter and you expect accelerating growth with Data Center being the lion's share of the October quarter sequential growth. Can you talk about where we are in A100 adoption curve and what is the approximate split of A100 versus V100 in the installed base?

A - Colette Kress {BIO 18297352 <GO>}

Right. So, first starting off, our A100 is our Ampere architecture for the Data Center. Our A100 has been in market for approximately a year or so, and it has been widely adopted by the major hyperscales, as well as the CSPs. Adoption in terms of use for

internal use but also for cloud instances. Our A100 is special. Our A100 comes with an HGX platform board and it enables our hyperscales to not have to choose how you want to use the A100. It is capable of doing both. Capable of doing both tuning[ph] as well as (Inaudible). So, it has been quite popular easy in terms of adoption. In terms of general availability, almost in all clouds worldwide.

But it also has encouraged enterprise adoption to be strong as well. Our NVIDIA-Certified Systems with our OEMs has helped accelerate the mainstream adoption that we're seeing. We have hundreds of thousands of enterprises already looking at A100 and or deploying A100. When you think about what percentage of our Data Center compute revenue is using A100, it is well over half. It is, yes, popular that we will be not only on our current architecture end market, but we will be using other of our architectures as they will be continued to be deployed both for hyperscale clouds, but also for enterprises, very common for people to add on to their current clusters, with some of our existing architectures of the past as well.

But A100 adoption, it's been driven by new AI use cases, new workloads, folks that are upgrading from V100 installed base, possibly some of them, but we still offer that V100 as well. Those new use cases are some of the most important areas of AI today. In terms of natural language processing and/or recommendator engines. Recommendator engines fuel everything from the hyperscale business models to consumer internet companies, as well as enterprises looking at ways to use that. So, the A100 is doing quite well.

Q - Atif Malik {BIO 7312618 <GO>}

Good to know. And Colette, how is the inference momentum with the addition of A30 going?

A - Colette Kress {BIO 18297352 <GO>}

Our inference platforms are doing quite well. Keep in mind, we have inference platforms that are specifically used for inferencing. We started this out a years ago with the introduction of our T4, our Turing 4 that was made specifically for inferencing and now we have come out with the A30. The A30 is an upgrade to our T4 and is also specifically used for inferencing.

Inferencing so far has outgrown the use of overall CPUs. Moving to GPUs and GPUs performance, both help them with latency, cost, and overall scaling considerations. In the early days, several years ago, when we introduced the idea of using GPUs, we were starting with an overall market size at zero. We have now seen such great adoption of our platforms for inferencing within Q2, for example, it makes up a very meaningful part of our revenue in Data Center compute. Additionally, the revenue is up 2x year-over-year.

So, we have great momentum going forward and we have lots of tools to assist those that want to use GPUs for inferencing, whether that be Triton. Triton adoption has occurred through many of our CSPs. This is an inference server that simplifies the deployment of AI models and scales into production. It powers open-source inference serving software and it also leads teams to deploy AI models from any

framework that is out there. Our TensorRT, for example, has been downloaded 2.5 million times across more than 27,000 companies. So, A30 provides 4x the inferencing performance of the T4. It also includes that MIG technology that you can also find in A100 and it supports a broad range of AI inferencing and mainstream enterprise compute workloads. So, the customer interest is high and it will continue to help power the AI adoption almost across all industries.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And are both hyperscale and enterprise verticals expected to grow in second half or is it a universal strength?

A - Colette Kress {BIO 18297352 <GO>}

With our guidance that we provided for Q3, we indicated that not only the Data Center acceleration that we saw in Q2 would also occur as we moved into your Q3, so we're expecting that acceleration. The demand is accelerating as we expect strength quarter-over-quarter from both our hyperscales and our vertical industries, including a ramp of enterprise that are using the specific NVIDIA-Certified Systems that our OEMs have developed. We expect strong growth overall in compute quarter-over-quarter and we have great momentum as we've discussed in terms of what the inferencing in those workloads, that had outgrown the CPUs.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And Colette, I was reading a blog somewhere saying that Artificial Intelligence economy will be nearly 4x larger than the mobile economy that drove the likes of Apple and iOS. You guys are the pioneers in accelerated computing and have employed deep learning in multiple new end markets. And one of the latest end markets at least from investor interest this year is Software. And Nvidia talked about a software subscription model for SuperPOD hardware and Omniverse enterprise at Computex in May earlier this year. How is the standalone software licensing adoption progressing and when can we start layering some of the sales from software in the future?

A - Colette Kress {BIO 18297352 <GO>}

Sure. We've made a great effort to date in terms of allowing as many as possible to easily adopt AI into their infrastructures and work. We started with the hyperscales. So hypersales are very well known to build up their infrastructures alone and often write some of their own software. The GPU is well positioned with its CUDA development platform, being able to work with all of the frameworks for many of those hyperscales to get started.

But we're talking about a new wave, a new wave in terms of adoption that will occur with our enterprise and our vertical industries. Now, what is different about them versus the overall hyperscales, it is very common for them to buy OEM systems or by what we are seeing right now as NVIDIA-Certified OEM Systems to begin their work on AI. What they are used to doing is, using enterprise software such as VMware as their underpinnings of system software to help them manage all of their Data Center systems.

We are coming to market now with standalone software to assist them in their AI deployment. NVIDIA AI Enterprise is one of the first. Dozens right now of our automotive, education, finance, healthcare, manufacturing and technology companies for example are looking and reviewing this software for their deployment. It's very common for enterprises to put their mission-critical software on a software program with overall support and work for maintenance as well. This is what we're doing here. So very like how they have sold overall VMware to them. This would be another part of that system software that helps them manage their entire fleet of servers, but also allow them to stitch together the work of AI directly to their applications that they would have as well, so that's one piece of edge.

Additionally, we can talk about Omniverse. Omniverse is another overall software opportunity that we have for us, an opportunity that more than 500 different companies are evaluating. What does Omniverse allow you to do? Omniverse really takes charge of what we're seeing as these virtual environments, connecting multiple and multiple virtual environments. We'll have our own domains about where we will live in that virtual environment. But the ability to move from environment to environment is what the key things that Omniverse would enable you to do. The software that will be licensed to each one of the overall users as a user's price in order to create those virtual environments.

What are these types of companies that may be evaluating it? For companies, such as BMW, Volvo, Lockheed Martin. Why are those important for them to look at it? These are companies that have a significant amount of design, collaboration, and large engineering forces that work together, both on the creative side of the overall design, that instead of having to be in person, they can just go to the virtual space to complete their overall design. We're quite excited about Omniverse, been working on it for many years. As you see it in terms of some of the important parts of the metaverse, that is also being talked about at the infrastructure layer. So here is an example of where we can both support the overall hardware infrastructure, but also the software that creates these virtual environments. Now each of these opportunities can be multi-billion dollars in nature, but it will probably take us one to two years for them to be more meaningful and will be able to provide metrics along the way that shows our tracking to our progress long-term.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And just staying on Omniverse. Because I've been getting a lot of questions on this area. You had about, 50,000 individual creators have downloaded Omniverse since its beta in December of last year and you have 2 million plus CUDA users. So, does your strong position in CUDA help you proliferate Omniverse? Because it is going to be a -- it's going to take a lot of collaboration across multiple areas to put out the -- no matter where is that, but does the CUDA position helped you in your efforts in Omniverse?

A - Colette Kress {BIO 18297352 <GO>}

Yes, really good question to think about some of the platforms, whether it'd be software and or hardware we bring to market, and how we think about it. We are rarely thinking at the full-end use case of how software and or the hardware may be

used. What comes to mind is much of our developers that are on CUDA, developers that are on many of the AI frameworks continue to find new use cases all the time. We see the benefits of Omniverse for the 20 million designers and engineers that are out there. But we are assured that also our CUDA, overall developers will even out onto that more. And onto it more to think about how to stitch together a full metaverse. How to provide it to more than just designers, engineers; provide it to folks that are just looking from a collaborative folks, could be folks like you and I in the future. I don't know, maybe we'll do this conference in our virtual environment, and I'll have my digital twin there for you to overalls (Inaudible) as well. So, there's great opportunities by just providing the opportunity and the overall programmability for folks to use Omniverse, CUDA, and many of our other products together to fund new use cases.

Q - Atif Malik {BIO 7312618 <GO>}

Great. Let's talk about the Grace CPU. You guys surprised us a bit this year in terms of your announcement on Grace CPU in 2023. How's the team progressing on the CPU development? And what has been the Data Center customer response on the combined multi-year GPU, DPU, and CPU roadmap so far?

A - Colette Kress {BIO 18297352 <GO>}

We had talked about our roadmap, that we wanted to focus on the modern data centers of the future. The modern data centers are racks of GPUs, racks of CPUs, and racks of DPUs. Now the Grace overall CPU that we have announced are Arm CPU, is a very specific CPU. It is a CPU that enables that connectivity between the CPU and GPU from the on store[ph] of data into the processors, that we can make sure that we are monitoring acceleration throughout the full process with the combination of the CPU and GPU together. So, it is a niche product that is focused on AI workloads, focused on high performance computing type of workloads, and maybe very key in terms of supercomputing as well. So, we're in the initial stages and we think in the next couple of years, this will come to market to be our first overall Data Center CPU.

Q - Atif Malik {BIO 7312618 <GO>}

Great. Let's switch to gaming. Gaming sales have more than doubled since April of last year. You guided sales up slightly, sequentially in the October quarter. How sustainable is the current boom in gaming, as environment recovers from pandemic?

A - Colette Kress {BIO 18297352 <GO>}

So our gaming demand is strong. Our gaming demand is strong, but our growth is continued to be gated by supply. Our channel inventory in the market is still low and we continue to work on providing more supply into the channel to improve those scenarios. The universe of gamers continues to expand. That universe expands because gamers are becoming more than just one-to-one playing with the game, it moved to becoming a very important social place for them to connect with their friends. But now it is a full entertainment sport. Entertainment that meets all different types of users, people that are actually playing the games, professional gamers, first-time gamers, people that are training others to be gamers, people that are broadcasting, and then, of course, there are those that are just watching others

game. The ones that are watching for example 500 million Esports of an audience. That's more than 700 million live streamers. PC gamers on Steam, very popular platform, is up more than 20% year on year.

So, during this last year and a half or longer, it's been an important time for people to find that entertainment, and that entertainment was gaming. And a very popular spot that all took place on. But as we all know, once a gamer, always a gamer. But it's important to look at some of the other things that have also driven a lot of the adoption towards gaming, a strong and powerful market of laptop gaming as well. Laptop gaming that infuses some of the best performance, but also in a thin and light offering. So, we're seeing folks not only build their own PCs with a desktop but also may have a second opportunity or their only opportunity which is a laptop. So, we have a record number of Ampere-based designs with the OEM and we are well-positioned heading into the holiday season with Ampere.

Now, right now, why is gaming popular as well? Is, we are using RTX across all of our notebooks, as well as our high-end in terms of desktop. What this means, as folks can realize some of the best games that are out there in real-time ray tracing that is just starting its second generation with RTX. We've got a great upgrade cycle ahead of us still. Ampere does make inroads into our installed base. But only 80% of our installed base still has not upgraded to RTX. So, we've got a great continued opportunity of Ampere in front of us. We're getting ready for the holiday season. And as we turn the corner to next year, we're still purchasing supply for long-term to fill this market.

Q - Atif Malik {BIO 7312618 <GO>}

Great. Colette, just on that RTX adoption. I was surprised to hear that number 80% seems like fairly early innings in terms of RTX adoption. And I know you guys don't talk about your next-generation roadmaps or products. But do you factor into where the adoption is -- of a current platform before you decide to put out your next-generation platform?

A - Colette Kress {BIO 18297352 <GO>}

So that is correct. RTX is an important movement right now, that are moving people to both upgrade and or entering into gaming. And so, 20% or less are upgraded, either 2 Ampere or are able to work on RTX types of cards. So, you're right, that we have a big opportunity in front of us. It's common that we may change into a new architecture and not have reached a large percentage of our installed base on the most current architecture. As you know, we're always working every day on new architecture, as well as architecture Next as well. So, we're going to make sure that whenever that time comes, we'll be ready to go, but it won't change anything, our plans are our plans. If the architecture is ready, we're happy to bring that to market.

Q - Atif Malik {BIO 7312618 <GO>}

Good. And then the supply chain, things have been tight across the board. Gaming availability has improved since May but remains tight. When can gamers get their hands on the RTX cards?

A - Colette Kress {BIO 18297352 <GO>}

Yes, channel inventory is still low. The overall prices in the market are a little bit higher than what I would say, manufactured suggested retail prices are. So, we know we have a way to go in providing more of that inventory. We're starting to see prices decrease and therefore make it more affordable and available to our gamers, but we're still working on that.

We work with all of our supply chain partners and ecosystems to improve that availability. It is a full piece of work across all of these different firms. Meaning, when we think about what we sell, it's very rare that we actually just sell chips. We are selling for platforms, so we have to think about all of the components, but we also need to think about the components that our partners and ecosystems must also procure to finish up those systems. So right now, we're working across the industry to make sure that we are all working together to get that inventory to the market and to gamers. We do our best to serve the overall needs of the gaming demand not only for gamers but again, for our overall graphics as it relates to our professional visualization as well. So this entire process includes a lot of different companies working on providing more supply.

Q - Atif Malik {BIO 7312618 <GO>}

Good. And then you've said the demand is going to exceed supply at -- through the end of this year on gaming. How is the team managing supply and where are the shortages most pronounced? Is it across wafers or substrates?

A - Colette Kress {BIO 18297352 <GO>}

That is correct. We do expect that we will not be able to serve all of the demand, but we're working on it through that. Now, the way that we are working on it is, if you've watched in terms of our procurement process, you'll see us more and more with longer-term commitments. Longer-term commitments across the stock in terms of for chips, for components, for all of the different necessities that we need, including contract manufacturers that help build out using the supply that we have. So, we're working on all of these different pieces right now in terms of procuring for a longer period of time.

Given the size of the company, we have quite a bit of optionality. We do have two high-end fabs that we are working with. We're probably one of the only large companies that has such a strong dual-fab process. We couldn't be more pleased in terms of our partnerships with both of the companies as that partnership and the long-term of working together has really helped us through this process.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And then the question on crypto mining. You guys have done a much better job this time around in terms of your strategy of offering new hash rate cards as well as CMPL as a separate category. Ethereum is moving to a proof-of-stake model, from proof-of-work and some investors are concerned this might result in mining demand

decreasing substantially. Given that bulk of Ethereum mining is being done by GPUs, do you think there is a risk of perhaps oversupply in secondhand GPUs?

A - Colette Kress {BIO 18297352 <GO>}

Yes. First -- let me first start with addressing what we see as the effectiveness of the low hash-rate cards. We established low hash-rate cards as a strategy to serve gamers with our GeForce cards. And then provide CMP cards to our miners as a whole. This we believe has been quite effective for the hash rate cards were new this time around and it really enabled us to market effectively GeForce for just the overall gamers.

The proof-of-stake has been rumored for quite some time and we have included in our forecast for Q3 that we would have a minimal amount of CMP revenue going forward. So, recent out[ph] activity could increase, but we're actually seeing several reasons why we believe it is not a significant risk to our overall gaming business. So one, gaming demand is so strong and the channel inventory is low. It doesn't matter what day of the week it is, but if there is some rumor out there that supply is showing up, the gamers are out there and standing in line to overall, get them.

Additionally, gamers want RTX, not just that they want a gaming card they want an RTX gaming card. And it would be difficult for miners at this point to buy a significant amount of those Ampere GPUs, as they're just not sold in large quantities. That's very unlikely that they have the ability to resell them into the market.

So, the CMP cards in quantity were provided to the professional miners and it is not desirable for gamers and probably would not see those in terms of resale either. So, given the availability and the mining performance considerations, it makes more sense for miners to purchase the CMPs and our gamers to continue to look for RTX as it becomes available.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And just moving on to auto (Inaudible). When is the inflection coming in Auto? And how would you differentiate your auto approach from other semiconductor competitors, like Mobileye and Qualcomm?

A - Colette Kress {BIO 18297352 <GO>}

So, there are so many different microprocessors inside of cars. But, I think it's important to note the difference in terms of our strategy versus the strategy of others. Our strategy is to build an end-to-end AV platform and AV Servers, soon to launch from Data Center to an in-car computer from data to training to simulation to on-road driving.

Many of our competitors are really focused on building fixed-function point solutions, which is very different from our overall strategy. So, our platform is open. We work with partners in many different ways or in any other different way that they are looking for. So, working whether or not they want the full end-to-end platform, that they want the software, or whether or not they want full options in terms of chips

or things for the Data Center. We are enabling a 10 trillion transportation industry, that will likely all become autonomous in the future. Our architecture and workflow from the cloud to the car is essential.

Now, we've talked about a pipeline, moving forward. A pipeline that takes us out to 2027 of the agreements that we've already signed with so many important car manufacturers out there, that pipeline is $8 billion by 2027. Now, we inspect[ph] inflection point of that pipeline in calendar year 2023, as folks are still working in terms of the development of these platforms to go into the fleets. These fleets can be passenger cars, robotaxis, and trucking.

Now, from time to time we'll continue to update this pipeline, as we are seeing many take interest of, one of our more important deals that we signed with Daimler. Daimler is using our end-to-end platform and is also sharing the software revenue that they receive in terms of keeping the autonomous software updated in terms of those passenger cars. So, as these new deals get added we'll continue to update our overall pipeline going forward as well.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And audience if you have any questions for Colette, feel free to click on the box and submit your question. Colette, on the Arm deal, you guys have made progress with some of your -- some of the fabulous semiconductor players like Marvell, MediaTek, Broadcom endorsing you guys, and you also acknowledged on the last earnings call that the regulatory discussions are taking a bit longer than initially expected an agreement with SoftBank allows until the end of 2022 to complete the acquisition. Investors ask, why NVIDIA still needs to acquire Arm?

A - Colette Kress {BIO 18297352 <GO>}

Well, we don't have to acquire Arm. We want to acquire Arm. Arm is a great company. Arm gives us an opportunity to address a much larger opportunity. It also gives us an opportunity to Arm on many different markets and workloads that they would like to get into. We want to see if we can expand Arm's IP licensing opportunities that allows us to offer Nvidia's technology in large end markets, including PCs, including mobile. But also allow us to help R&D roadmap at Arm and turbocharge that investment. They have -- would like to build out something similar to what they did with mobile, in the Data Center and within PCs. We know both of those markets well, and we would love to help them build out the ecosystem and the work that's going to be necessary.

We have the ability to reach more and more developers, developers at Nvidia, on CUDA, or otherwise to the more than 15 million developers that are available at Arm as well. So, we know that we can strengthen their work in terms of the Data Center and the PCs. Now, we believe regulators will also see and understand the value of Nvidia -- of what Nvidia can bring to Arm, to both the ecosystem and then how the customers can benefit from more competition and availability to those customers. So, we will be able to update as soon as we have more information, but we're still working through the regulatory process.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And a question on the financial model and the gross margins. How do you view your gross margins on a long-term basis given a push towards higher supply chain continuity, even higher supply chain cost, the foundry partner has talked about raising pricing, and just structurally higher inventory?

A - Colette Kress {BIO 18297352 <GO>}

Nvidia as a whole is not accustomed to some of the things that are occurring. Because our focus is in terms on innovation and continuing to build better and better platforms to meet the customers' needs. Better and better platforms generally mean some of the best and high-end types of solutions out there. So, costs have been continued to rise from the beginning of time for us. But today, we're able to offset most of that with the overall mix of what we're doing. We continue to take those platforms and build stronger and stronger systems, with our mix we're able to offset a good amount of those cost increases. Mix and software are probably the largest drivers of our margins built today, as well as in the long term in terms of what we expect our margins to do. So, we'll continue to keep an eye on it, but right now, these costs are nothing new to us.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And I've received a question from an investor on Auto. Auto pipeline is expected to grow to $8 billion in 2027? Please break down whether this is mainly software or hardware or something combined?

A - Colette Kress {BIO 18297352 <GO>}

Our pipeline of $8 billion for our Auto autonomous solutions out to 2027 is a combination -- is both the combination of hardware systems as -- that are inside the car, as well as the software opportunity -- system software and or drive software that we are adding for both passenger cars, robotaxis, and trucking. So, it includes both hardware and software.

Q - Atif Malik {BIO 7312618 <GO>}

Okay. And then the question on OpEx. You spend quite a bit of OpEx and rightfully because the size of the market you're after are very big. What is the split on OpEx on the software hiring or the programmers and application engineers versus the hardware engineers?

A - Colette Kress {BIO 18297352 <GO>}

Yes, a great question. So, our focus in terms of investments, the investment that you see within our operating expenses, a vast majority, a vast majority of that investment is new employees. New employees that we have welcomed through this pandemic and all -- many of them still continuing to work from home. Some of our top areas of focus in terms of hiring is both hardware and software. I'd say it's a race in terms of which one is hiring more. We are hiring software engineers and hardware engineers, for so much of our solutions.

Our solutions, if it was just hardware would not benefit, if we did not have all of the software solutions that we need. Keep in mind, remember, we have a unified architecture, an architecture that goes across all of our products from a hardware perspective, but that is also true in our software. So, our software engineers are working in terms of backward compatibility, but of course creating new capabilities for so many of our customers from hypersales, to enterprises, to gamers, to researchers, all of them and really focused on software that will enable them. So, both are very important hiring opportunities for us.

Q - Atif Malik {BIO 7312618 <GO>}

Great. And we're almost out of time, Colette. Thank you for being a key note at our conference. Thank you so much.

A - Colette Kress {BIO 18297352 <GO>}

Thank you again for having us. Appreciate it.