2023 Evercore ISI Semiconductor & Semiconductor Equipment Conference

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer

Other Participants

- Analyst
- Matthew Prisco, Evercore

Presentation

Analyst

Good morning. Great. Good morning. Welcome all of you to Evercore's Semiconductor & Semiconductor Equipment Conference 2023. It's our pleasure to have all of you here today. And importantly, it's a pleasure to have some of the most influential and visionary leaders in the technology sector with us over the next two days to discuss what I would argue is probably the most complex and dynamic set of trends that any industry that I'm aware of is facing, which makes it a fertile ground for creating alpha as investors. The industry's importance only continues to grow. Many of the generation-defining advancements that have become everyday topics, generative AI to quantum computing, robotics, smart manufacturing, autonomous driving, clean energy, are all powered by this industry.

The market cap, and I think the investors have recognized that the market cap of the semiconductor industry in aggregate has grown to over $5 trillion. And we have our first trillion-dollar market cap company that had its roots in the semiconductor industry, notice I was very careful about that, it's no longer just a semiconductor company, and that's just been a remarkable transformation to watch more on that in a second.

In addition to understanding the secular trends that are driving the industry, navigating the geopolitical sensitivities is a critical part of the investment thesis. To that end, hopefully, our program over the next couple of days will shed some light on those events. We do have and we're delighted to have Todd Fisher, who is the head of the CHIPS Act, for a lunch program today. Our Founding Chairman, Roger Altman, will moderate that session. And obviously, the tech sovereignty dynamics are going to be a hugely influential factor in investing in this sector.

A quick word on Evercore; like many of you, we stayed very invested into this down cycle and that has proven to be quite lucrative for us, like it has for many of you. On

the advisory side, which I lead, we've hired 11 new partners this year, which has been one of our most significant recruiting years, including hiring Tammy Kiely, who is a Former Head of Technology and Head of Semiconductors at Goldman, to join our partner; Tom Stokes in leading our semiconductor effort at Evercore. And importantly, we continue to be highly invested in our equities platform. Evercore ISI is a critical part of Evercore strategy going forward. I think, fusing our banking relationships, our capital markets capabilities with Evercore ISI's world-class research, we think, is a winning strategy and we're highly committed to that.

Finally, I was talking to Ed Hyman, who leads Evercore ISI -- founded Evercore ISI, I think, during the course of August, and he said something that stayed with me. He said there's only three Js that matter when it comes to investing for the rest of the year. The first was Jackson Hole. Of course, he was referring to the Fed policy. The second was Jinping, which was Xi and China. And the third was Jensen. And I think he was spot-on on all fronts, particularly on the Jensen front. So, with -- and as I mentioned earlier, I think the transformation we've seen over the last decade -- couple of decades with NVIDIA, and of course, it's an overnight success a couple of decades in the making has been just incredible to watch.

So, with that in mind, I'm delighted to welcome Colette Kress, CFO of NVIDIA to the stage. Thank you and enjoy the conference.

Matthew Prisco {BIO 20801520 <GO>}

So, good morning, all, and welcome to the Evercore ISI Semi & Semi Equipment Conference.

I am Matthew Prisco, semiconductor analyst here. And this morning, we have the pleasure of welcoming Colette Kress, CFO of NVIDIA accompanying with me for a valuable discussion. I believe Colette is celebrating her 10th anniversary as NVIDIA CFO this month, a period over which NVIDIA has seen some modest gains as it's transitioned from a PC unit play in the eyes of many to the accelerated compute powerhouse that it is today. So, welcome, Colette. Thank you for joining us and congratulations on a hugely successful journey.

Colette Kress {BIO 18297352 <GO>}

Thank you.

Questions And Answers

Q - Analyst

(Question And Answer)

Q - Matthew Prisco {BIO 20801520 <GO>}

So, as for format of this chat, I have a number of questions to run through, but we'll save some time at the end for audience Q&A. But to start, there are two main

buckets, where we're fielding the most questions today. These are around data center sustainability and supply. So, kicking it off on the sustainability side, particularly the data center revenues in 3Q position has tripled from 1Q levels. Would love to hear your thoughts on what's driving confidence in the sequential growth from here. And maybe how close are you working with customers on the infrastructure build-out strategies? And how are you thinking about potential pent-up demand for data center GPUs today?

A - Colette Kress {BIO 18297352 <GO>}

Okay. Thanks for having us here. I do have to make an opening statement here --

Q - Matthew Prisco {BIO 20801520 <GO>}

Sure.

A - Colette Kress {BIO 18297352 <GO>}

-- that says, 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. Okay.

So, let's first talk about some of the things regarding demand. That's our sustainability of demand in terms of what we're seeing. We do have very strong visibility as well as strong interest in terms of our products. A lot of this stemmed after the opening of OpenAI's ChatGPT over the holidays and it's been an increasingly interesting time related to that. People really started to understand the simplicity of how using AI in so many of our enterprises, whether they think about it from a monetization standpoint of new products that they could do or whether or not they can just see AI in terms of efficiency in everything that they are doing.

That demand that we are seeing requires us to spend quite a bit of time in planning with many of our customers. Our customers that we see today are customers that we have also been with sometimes for a decade or so, working on their work in terms of inside of data centers. I think there were some important opening statements that says what we are supplying is not a chip. We look at ourselves as a data center computing company and helping them create the data centers of the future. The data centers of the future that we've talked about are not only associated with AI and some of this very important killer app that we see, but really more along the accelerated computing that we think will take off consistently as we see in the future.

That growth that we see is really looking at how they can improve both the efficiency of their data center, working on the sustainability of their data center, the use of energy, but also really allowing them to do work that they've just not been able to accomplish before without accelerated computing. So, yes, we are working with many companies on helping them in terms of their data center builds, helping them in terms of the planning as we look in terms of solutions.

Large language models right now are very front-and-center as they are using our products to help them in building those large language models; and then in the

future you'll see the inferencing that is related to the models that they have built. But keep in mind that's only a couple of the different use cases, because we also see the use of recommendator engines, the use of many other forms of extracting a substantial amount of data and using that data to accelerate their computing as well. So, those are some of the things that we're seeing across our customers.

Q - Matthew Prisco {BIO 20801520 <GO>}

That's perfect. And now foundational model build-outs are clearly a focus area today. How will you think about the number of these models that will need to be built and how long could that drive training demand? And then maybe once these initial models are built out and training focus shifts towards fine-tuned iterations for specific applications or just model maintenance adjusting for drift and whatnot, will there be enough training demand to fully utilize the current build-out that we're seeing?

A - Colette Kress {BIO 18297352 <GO>}

So, the discussion on foundational models has come front-and-center. There's one and two and three more different foundational models that have been created. There's more that will be created for both different countries, different regions of the world, as well as very specific areas. But those are only one types of the models that we'll see. We'll see additional models that will often be related specifically to a company. Now the company may look at two different types as well. They will have those that they have internally that are using the data, the information that helps fuel their inside company, but also what they can do to help customers in terms of call centers and building models, how to answer those types of calls.

So, not only are these foundational models are important part of training, but you'll continue to see models being built otherwise. Now, our architecture that we have built in terms of a lot of our systems that we bring to market today are associated with the capabilities to do both training and then moving into that inferencing stage after that training process. That has extremely helped the efficiencies of data centers, because it is a dual use. You'll see them build the models, work in terms of the inferencing and may even come back that you have to make adjustments to the model over time and continue to build into that. That leaves them with a great infrastructure of both the best training that you can do, but also a very, very sizable inferencing market in the future as well. So, we do think you'll see both of those.

The inferencing will be related to massive amounts of data that may be there with the consumer Internet companies and how they use it. You see companies such as Microsoft considering that use of our inferencing platform to help them. You also see GCP, for example, looking at our offerings to do that as well. So, both inferencing and training will be important as we go forward.

Q - Matthew Prisco {BIO 20801520 <GO>}

Okay. So, 2023 is seemingly the year of the training build, build that comes with extreme compute requirements and lofty ASPs. But as these models begin to increasingly move into production and mix, skews more towards that inference side, how should we think about the impact to NVIDIA? Is there any risk from pressure in

the fall of ASPs or just any type of digestion period from customers as they kind of take that training build and leverage it for infrastructure?

A - Colette Kress {BIO 18297352 <GO>}

When you think of the work that's being done now on training, a lot of them have spent their time thinking about the size of their models, thinking about the parameters that would be necessary in the type of training. And so that has determined the amount that they are interested in terms of purchasing. Our position in helping companies even in all forms of AI or accelerated computing is really about the TCO savings that they can realize. There is nothing better than the infrastructure that we have to both save money, save energy in their work. That is what's determined our pricing. Our pricing looks and say of the value that we will help them save not on a single chip, but on the data center as a whole. Those savings we both pass to them and part of the work that we've done on the pricing and then we'll have some that we use for reinvestment back into our company.

When we set price, we set price at the very beginning and we relatively keep that price throughout the ownership and the selling of those two different architectures. So, what do we see going forward? Although we're in this position of AI and looking at generative AI, it's a very important piece right now. This is probably a very important inflection as people are now understanding the need to focus on the $1 trillion installed base of x86 CPU type of infrastructure that's there. This is the time to think about how do you make that even more efficient; how do you think about moving that installed base to accelerated computing, not necessarily just for generative AI, but all different forms of computing that can happen. We'll likely see that transition as we move through generative AI to moving to accelerated computing in the long term as well. So, there's the benefit of thinking about purchasing now as you move towards that accelerated computing as the ownership of the data center as a total.

Q - Matthew Prisco {BIO 20801520 <GO>}

Makes sense. Now moving to the proliferation of these large language models; how are you seeing the split between customers creating their own models from scratch versus those that are leveraging the foundational models and fine-tuning to meet their own needs? Then maybe how do each of those different scenarios impact NVIDIA differently?

A - Colette Kress {BIO 18297352 <GO>}

Yeah. The foundational models, using our infrastructure as many large companies build foundational models and not only one foundational model, but upgrades to the foundational models over time will be there. The other thing will be enterprises and companies building their own different models. We continue to support them as well. They may be supported in many different factors, but one of the things that has been very helpful to them are models -- pre-trained models that we either have available and/or things such as NeMo, or BioNeMo that helps them in terms of optimizing their models as well.

We have software, we have services that continue to support them as many of these companies build out their models. We receive requests all the time. My model is in great shape. It's working at about 80%. Can you help us in terms of optimize that? So, NVIDIA has not only that key infrastructure that they need, but we also have the software to support them.

Q - Matthew Prisco {BIO 20801520 <GO>}

That's perfect. And then when thinking about the inference market in particular, majority today processed by CPU, but as these generative-AI-based models move into production, how should we think about the GPU opportunity or maybe put differently, how to think about the percentage of new inference-based build-outs, which are best supported by GPUs and kind of what's driving that view?

A - Colette Kress {BIO 18297352 <GO>}

So, when you look at the infrastructure that you can do both, it's a little bit difficult for us to differentiate and determine the exact amount of time that is being spent on training and inferencing. However, we do believe inferencing is a very large market, likely larger long term in terms of what we're seeing. And we'll continue not only selling our systems that do both, but we also have specific systems that can be leveraged specifically for inferencing. And that's an important piece that folks are looking at and thinking through the cost of every step of that inferencing platform. Cost of course is going to be important and also the sustainability of energy, what is the lowest form of energy that they can use to move through that inferencing position. So, we do have platforms that help in terms of both of those and I think we'll continue to be a driver of the inferencing.

The inferencing has become extremely complex over time. Inferencing of 20-30 years ago and using CPUs to work, it was rather a binary type of decision in the inferencing. Today, you have such complex models and that inferencing and the latency that's required for a lot of time the end use is a very important piece and GPUs are just perfectly set up to help that.

Q - Matthew Prisco {BIO 20801520 <GO>}

All right. Very clear. So, I'd like to move over to supply now, other major area that we're fielding questions on. And maybe you can help us understand the disconnect between the approach many are taking in correlating cost capacity growth to NVIDIA's GPU shipment growth capabilities. Obviously, the correlation is not holding at all currently. So, what is being missed here? And maybe just help us better understand how we should be thinking through the supply backdrop and potential expansion from here?

A - Colette Kress {BIO 18297352 <GO>}

So, we highlighted at our earnings our plans for supply. This continued ramp of supply for the quarters moving forward even as we go into our fiscal year '25. We've been working across the board in all of our different suppliers to help improve our supply position and support many of our customers and the demand that they put in front of us. That means looking at a lot of different things, adding additional

suppliers, adding additional capacity, qualifying, but also looking in terms of the time spent and improving the cycle time as a whole across that supply chain. Those are the things that we've worked on. Colos is an important piece in terms of what we are putting together. It's a form of packaging that is really for a lot of the high-end types of chips and the types of things that we do. But keep in mind, there are multiple suppliers that we can add to our colos to improve our overall size of supply, and we've done that. And you'll see that being a part of our ramp as we bring on more and more suppliers to do that.

Q - Matthew Prisco {BIO 20801520 <GO>}

Okay. So, now you had mentioned this development and qualification of new suppliers for key steps in the manufacturing process. So, first, can you add maybe some more color on NVIDIA's involvement and investments into developing that supply chain? And second, as you build out your supplier base, how do we think about that qualification process and how you mitigate maybe quality concerns from utilization of less mature sources?

A - Colette Kress {BIO 18297352 <GO>}

Okay. So first, starting with our partnership with suppliers, many of our partners have been with us for all 30 years at the company, as well as even in multiple decades, and that partnerships have been very important. We try and continue to treat our suppliers as best as we can. In some cases, we may be going faster than our suppliers. So, helping work together in terms of how we can improve supply scenarios is our work.

When you look at how we think of supply, you may look at only the inventory that we have. But the reality is looking at our purchase commitments as well as sometimes our prepaids, which are both helping our suppliers quickly bring up capacity or quickly bring up supply are going to be some of those things. The more that they can get an understanding of our long term with our purchase commitments and our capacity agreements has been very helpful and we'll continue to do that.

As we move forward, going forward and taking folks that need to be qualified or additional help in terms of bringing that, that's a process that we work together with them. We are right there with them both on a quality standpoint. Most of our suppliers are not new suppliers. They've just been serving other parts of the business or they've been serving other different customers. So, we feel very good about the types of customers that we will see in the supply chain and the quality really hasn't been an issue. They're all right there with us.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. As we think about the potential supply limitations, is this causing a heavier mix of Hopper overall than you would have expected? And over here, maybe you could offer some color on where we stand in that Hopper adoption curve today, how that compares to prior cycles and how you think about that moving forward?

A - Colette Kress {BIO 18297352 <GO>}

So, Hopper; Hopper is our current generation architecture that we have in the data center. And it has been in market probably close to a year. It's something that we had launched last fall. But keep in mind, even when we launched it last fall, this was something that we continued to improve our relationships with our customers, knowing it's coming, helping them both qualify, helping them really understand the engineering behind it, so that it is equally adopted quickly as we see in a lot of our other products. So, Hopper is an important architecture. You'll continue to see us for some time using Hopper. But what's interesting and something we've always seen as well is that our prior architecture is often sold at the very same time. So, we are selling Hopper and Ampere. Our Ampere architecture, it's also the second best architecture out there in the market.

And so why? Why are both of them being sold? Well, many people have already qualified on Ampere. Some of them are still qualifying on Hopper. It's a great opportunity for them to add additional Ampere to some of the projects that they're doing. They may move into the second stage of a project and changing an architecture may not be that ideal for them continuing with Ampere. So, even in the second quarter, we sold and from a volume standpoint just about equal between our Ampere and our Hopper. But I know we'll see Hopper continue to ramp even more as we see in the next couple of quarters.

Q - Matthew Prisco {BIO 20801520 <GO>}

All right. Perfect. So, prior to the data center revenue explosion over the past two quarters, you guys had highlighted a vision for acceleration attach of roughly mid-single digits, growing to about 100% over the next 10 years. Given the recent uptick, how are you thinking about where we now stand today and what does that journey to 100% look like? Maybe any color you could offer in terms of workload transitions or timing of potential milestones, anything would be helpful.

A - Colette Kress {BIO 18297352 <GO>}

Yes. So, the statements we're making here is looking at the single digits of what percent currently we are within that data center. Data centers across the world, you have to look at it in its full view in terms of understanding modern data centers right now are quite disaggregated. So, our goal is to be a percentage of the data centers as a whole, not necessarily counting servers or counting individual types of chips. I think it's the right way to look at things from a data center perspective. You're right. Our vision focuses on accelerated computing will be within all data centers and will be an important piece of that.

This movement of generative AI, the easy understanding of it has really influenced folks looking at accelerated computing, both for this key AI application, but also in terms of the long term that says conserving energy and finding a way to just densify the work is really important from a TCO value to them. So, we're on that journey. How fast does that get to that 100%, that's going to be really, really hard to determine. But again, this is a powerful inflection point for us to continue even past generative AI to move to accelerated computing.

Q - Matthew Prisco {BIO 20801520 <GO>}

And on that path to 100% of data centers being accelerated, are there other limiters when thinking about the full data center system that could pressure the slope of the adoption curve, whether it's hardware, software, ecosystem? And if so, what steps are being taken to kind of push on those fronts?

A - Colette Kress {BIO 18297352 <GO>}

It's an important piece to understand. When we think of us as a data center computing company and what we provide, sure, the GPU systems have been important, but keep in mind, we also have the ability now to adopt a CPU to help in the overall acceleration process. Our acquisition of Mellanox and the use of networking was to really understand the importance of networking to many of the high volume systems that are there. The amount of traffic that comes inside of a data center is such unique and focusing on network and an acceleration is important. That's why we step back and look at the moment data has probably entered into the data center, what can we do to accelerate that work? Our networking, our NVLink, our CPU, our GPUs, but importantly, our software, our development platform, our end-to-end stack to help them all the way to the application is very key in terms of that adoption of accelerated computing.

Without that help, working on such important time to get applications re-routed for using accelerated computing is work. And I think that full platform is influencing the ease of adoption and the ease of moving to accelerated computing faster than anything that we've seen.

Q - Matthew Prisco {BIO 20801520 <GO>}

Okay. And as we envision this world of full acceleration, how are you thinking about the split between GPUs, ASICs, other potential accelerators or maybe what percent of that total compute is best served through GPUs? And what are some areas that potentially will remain better suited for alternatives?

A - Colette Kress {BIO 18297352 <GO>}

It's tough to determine the success of a custom ASIC or a different type of accelerator, but I think it's important to understand that we are about data center acceleration. And sure, we have some form of accelerated chips, but there's a lot of different types of accelerators out there. Not all of them are accomplishing the same work that we are doing. Some of it is actually very difficult to determine the size of that success. In a world, where things are moving quite quickly, changing quite quickly as almost all types of applications going forward will have some form of AI in them. AI is still in the very early of that journey. Customizing a specific ASIC, hard coding a specific ASIC makes it difficult, because things are moving very fast for that to be beneficial and it probably takes you several years to accomplish a custom ASIC.

But there will be others that may be for a specific workload of size, a very static type of workload that you could see customizing some form of accelerator for it. That's something that has always existed. There will be possibly other small different types of options. But our focus is more of how can we get the adoption of the platform at

all different parts of our platform and that gives the customer as much of a choice as possible as they think through what they will build in the future.

Q - Matthew Prisco {BIO 20801520 <GO>}

Okay. And now we're clearly in a multiyear investment cycle here with a lot of spend going into this accelerated compute platform. But as we think of the attach rate growing, even to 20%, the amount of GPU spend becomes quite significant. So how are you thinking about the growth in end-use cases to justify these levels of investment? And do you think there'll be strong enough pull for customers to continue to pay up for this additional AI functionality?

A - Colette Kress {BIO 18297352 <GO>}

What we're seeing today, something as simple as generative AI has a lot of different pieces of really explaining to enterprises the simplicity of how using AI and many different forms of AI can improve their business. So, long term, you will see more and more applications both move to accelerated and/or incorporate AI. And that journey has begun and has been a part of us probably for more than a half a decade as well. As you see consumer Internet companies, you see very, very large cloud companies also working in terms of recommendator engines, working on moving a significant amount of their inferencing work that they do for sites such as search cam ads, things that they need to do for marketing of their products, being very, very key to accelerated computing and using of AI.

We're just at the beginning. We work right now in terms of software that help support many of the individual industries. You've seen right now many of the enterprises turning to us to both help them in the software and in the services of what they can provide in helping them rebuild their models, rebuild how they are supporting customers as well. So, this is not something that says, hey, this is a window of today. This is really that whole journey going forward that you'll see of an increase.

Q - Matthew Prisco {BIO 20801520 <GO>}

Right. So, I'd like to move over to the software side for a little bit. The announcements and the software offerings and partnerships, including VMware, Hugging Face, ServiceNow all very positive in our view and support the solidifying of NVIDIA's moat and reach. That said, the quantification of software revenues hasn't really changed all that much in the past 18 months. So what is driving that apparent disconnect between all the goodness we're hearing on the software side and the actual flow through to the P&L? And then how are you thinking about the software growth trajectory from here?

A - Colette Kress {BIO 18297352 <GO>}

Yes. Let's talk about we have much versions of software that help many different enterprises and help any customer that is using our platform. Keep in mind, there's a lot of software that is both just embedded in the systems that we provide to our customers. Just from the onset of their ability to have a full development platform, a full optimization platform embedded in the systems is extremely helpful. But you're

correct, we also sell it separately. This is an important piece for enterprises. As enterprises turn towards AI and accelerated computing, having somebody that is running that software, keeping it current, keeping with security there is an important risk factor that enterprises want to see. They want to purchase that. You've seen already in terms of our announcement, our second announcement with VMware, continuing to work with their platform as well.

How enterprises have VMware being connected to VMware, so that this can be visible inside of enterprises' data centers are very key. Now our overall software that we are selling separately, selling separately will likely reach near $1 billion this year. So, it is scaling, it is scaling quickly. This is key for NVIDIA AI Enterprise. That's a key part of it. We also have our Omniverse platform. And then long term, you're also going to see autonomous driving be a key factor of our software revenue as well.

Q - Matthew Prisco {BIO 20801520 <GO>}

Great. So, maybe on NVIDIA AI Enterprise for a moment, the company made the decision to directly package the solution in with the H100 systems. And recognizing that it is baked into that H100 price, maybe you could talk about the strategic rationale and the inclusion, and how has customer reception been to that dynamic?

A - Colette Kress {BIO 18297352 <GO>}

If a customer buys H100, depending on the different forms of the H100 they have, some of them also have that opportunity to buy NVIDIA AIE. There's many different places to buy NVIDIA AIE. If you, for example, are a user of the cloud, and you've got cloud instances with one of our many cloud providers, many of the market stores within those clouds also sell NVIDIA AIE that you can get so that you can remain in the cloud to get all the things that you need as an enterprise for that software and services with us as well. Additionally, as you remember, we've also included DGX Cloud, which gives you the same type of offering, but flipped in the way that we have the infrastructure, we support them in that software and services, but there's an underlying cloud instance in terms of the infrastructure with the H100. That's another way that you can purchase NVIDIA AIE.

If you purchase our DGX infrastructure, it comes inclusive of all of the software. That's a very important reference architecture for our enterprises or many other different types of customers. We want what NVIDIA has full stack, everything that we have and we can sell NVIDIA AIE within there. And then separately, if we are working with OEMs and ODMs as they build out server configurations using Hopper or some of our other products, you can also get NVIDIA AIE in terms of that way. So, there's many different ways. It's not always just with our Hopper. Many of our solutions provide the access to that software.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. And then before ChatGPT and the build out of these generative AI infrastructure took over all of mindshare, Omniverse was a key focus area, and theoretically, a killer app for next wave of AI as the collaborative platform for all AI tools. So, with that said, I would love an update on just where we stand today in terms of adoption and customer interest and maybe as we think through the

portfolio as a whole out of the next five years, how meaningful of a contributor is Omniverse to that pie?

A - Colette Kress {BIO 18297352 <GO>}

Omniverse is a great look at what we'll see in the future in terms of a 3D Internet. We have worked decades now looking at a 2D. But that transformation of a high-level 3D view is going to be extremely important. The work continues. The amount of adoptions from the creatives, those in terms of designers, really working in that 3D world is key. So, when you think about Omniverse, you have to also think about the importance of the infrastructure that's need, the workstations that need and/or the cloud infrastructure as well that they may use in either ways. That comes as a very important part. And then as we sell licenses for Omniverse as well, Omniverse at an enterprise license helps a group of designers work together in terms of creating that 3D environment. More and more upgrades as we see, more people coming on board, including that in both a USD form factor, but just a full content of what we can do in Omniverse has been a great increase. So, it's still there and an important piece to us.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. So, I want to pause for a moment and see if there are any questions from the audience at this point. All right, so we'll keep going. If anybody has any questions, just let us know. But in DGX Cloud, just moving there next. It seems like an evolution in the NVIDIA go-to-market strategy, and the partnership with Hugging Face seems like a great step in the right direction there. But given previous commentary that the ideal mix of DGX Cloud is 10% versus CSP Cloud 9D, why is the ratio being limited there, particularly given the incremental economic benefit of it?

A - Colette Kress {BIO 18297352 <GO>}

So, there is DGX infrastructure, which is our full systems that we would sell, but then there is also DGX Cloud, which as it takes all of the quality, all the quantity of the infrastructure and software and providing it through the cloud. Our DGX infrastructure, yes, maybe a reference architecture that may approximately be about 10% of what we sell in terms of there. But DGX Cloud is just a different form to receive the same thing. That's not in the same limits. That's an ability for us working with our cloud providers to hone in on different infrastructure and software and services that we can do on top of that.

So, it has much of an interest as we work with enterprises building models, they want to build models with our assistance with our software and services. They love to be on the infrastructure that we're using every day with the cloud providers to assist them and we work hand-in-hand with them. So, it's off to a great start, much interest for it, but we're still setting up that infrastructure with the cloud providers right now.

Q - Matthew Prisco {BIO 20801520 <GO>}

Great. That makes sense. And with the Grace Hopper Superchip shipping this quarter, can you give us an update on early customer demand reads and how you think about that ramp over the next one to two years? And then maybe how large of

a contributor could this be to the data center business over time, given the robust growth we've seen on the GPU side of things?

A - Colette Kress {BIO 18297352 <GO>}

Yes. Our Grace Hopper 200 is coming to market in the second half. We're very excited to include Grace connected with our Hopper architecture. Those things being architected together to really solve a lot of the work on AI and accelerated computing as it relates to processing data as it comes into the data center. Really thinking about the time that that data has to be processed, getting ready for the acceleration process is an important part of that connection with the CPU.

It also allows a very large memory position as well that assists both with the size of the data, processing the data and putting that to market. Types of customers, you'll see interest stemming both from our CSPs, as well as companies looking to create a full cluster that's available for them doing training models and/or completing inferencing at really, really good performance level, as well as a very good cost to it. So, we're excited to bring it to market. We'll probably be giving you some more updates as we finish in terms of our Q3.

Q - Matthew Prisco {BIO 20801520 <GO>}

That sounds great. And then on the networking side, we've seen some quite robust growth there, particularly last quarter. How are you thinking about that growth profile from here, maybe relative to the compute data center business? And could you help us in rank ordering the primary drivers of that business today, whether InfiniBand, DPU, Ethernet, whatever may be in there?

A - Colette Kress {BIO 18297352 <GO>}

Yes, our networking business and our acquisition of Mellanox has been really, really a great turnout for us. As we understood, thinking about data center computing how important the networking was to many of our customers. And our culture and similarity with Mellanox couldn't have been stronger to really help there. This is an area that not only are we working in terms of the acceleration of networking, but working hand-in-hand with the products that we're building in GPUs to leverage the capabilities of networking to improve that.

When you break it down in terms of what is successful, there's no debate out there that InfiniBand is one of the most important types of networking infrastructure that you would need for AI workloads. And many of our top customers choose InfiniBand for that purpose. It really helps with all different types of traffic that comes into the data center. And it is the preferred process that they use for AI. However, we have also focused in terms of Ethernet. And you'll see us now coming out with Spectrum-X. Spectrum-X along with the SmartNIC capability in our DPU will be an important process as well for enterprises or CSPs that are interested in the multi-tenancy, but also that staging position that they need with the DPU. So, it's hard to say what is our priority of what is important. But we know that often we are moving together with what we are doing in data center systems and networking together. So, it's been important to us to have both of these options to help the end customers fully build out their data centers.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. So, moving below the line, how should we think about gross margin expansion from here as data center increasingly becomes a more dominant portion of the mix and the interest segment mix within data center continues to improve with more software and customers buying up the stack? Is there kind of a fair baseline we should be considering for annual expansion or other moving parts to be thinking about?

A - Colette Kress {BIO 18297352 <GO>}

Gross margin, we've stayed pretty consistent that our largest driver of gross margin is mix. You're correct, there's mix in terms of our different market platforms that we sell into, but there is also mix within our data center business. When you think of our pricing as we've discussed, our pricing is focused on the TCO that we add in terms of value and many of our systems. The higher the performance, you also have a higher amount of software and services that are incorporated with that. So, that has improved our gross margin, but there are still parts of our data center business that have a very different mix associated with not as high a performance of some of our larger systems.

So, we'll continue with software being additive on top of that. That will be long term probably a key driver for us in the future. We already are seeing in terms of our outlook for Q3 that we provided that again our gross margin will likely increase there. So, it's something that we're focused on in terms of both providing the right solutions to the customers, but also helping us rethink through software investments and hardware investments that we see that are not necessarily in that gross margin. And that's what you're seeing that is enabling our gross margin today.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. Question?

Q - Analyst

I have a question. So, I'm curious about the L40 product. How did that come about? And it sounds like it's being used for fine-tuning. So, curious if that's like a new market (inaudible) and how should we think about the SaaS opportunity?

A - Colette Kress {BIO 18297352 <GO>}

Yes, so that's another great product that we have here. The L40S is probably one of the key ones to think about there. That allows you some great features and capabilities. One, not targeting any specific, but enterprises can really benefit from this. Why? You can now work with hundreds of different OEM providers in different form factors that this can be inserted in an existing server platform with four different L40Ss incorporated in there.

Now you have the ability to install that in a standard data center configuration, lining up with all the different servers that you have. Capabilities also therefore adding the

software capabilities to help them, as you discussed, in the initial models, often in terms of the child models or down from the parent models that they're creating to do that training and then also to the inferencing. This is a great option for enterprises, but we're also seeing this for large companies, consumer Internet companies and/or in terms of the CSPs, looking at this for also an inferencing type of platform that you could put together. It's a great option.

Q - Matthew Prisco {BIO 20801520 <GO>}

I guess just this robust revenue trajectory we're talking about, how do we think about OpEx growth within that in order to support it, especially as we enter this period of hiring? And what are the strategic priority of R&D dollars today?

A - Colette Kress {BIO 18297352 <GO>}

Right now, our trajectory on OpEx and our investments are important to us as we think not about what is in market today, but what we do believe we know we need to build and to bring into market. So, absolutely investing. Our revenue right now is growing a little faster than our OpEx is, but that doesn't mean that we're not focused on investment. Investment is in a lot of different areas, certainly on the engineering front and keeping track with the hiring that we are able to do, but also the compute infrastructure, where we are helping other customers by pre-looking at these configurations before their overall coming to market.

Most of our research and our engineers are using our compute internally to assist them in that. So, we'll continue to make investments in that area. We'll make investments working with both suppliers, partners and others out there as well. So, our number one focus right now is as much investment as that we can, prioritizing that for both the near term new products that are coming, but also definitely for the long term.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. And then, before you know it, you're going to have quite the pile of cash in the upcoming years. So, how does the company think about best deploying that capital today? I know the $25 billion in share repurchase authorization signals continued intent to buy back opportunistically. But overall, how should we think about NVIDIA's prioritization of spend?

A - Colette Kress {BIO 18297352 <GO>}

When we think about our capital, number one thing is going to be investment back into the business, as we've discussed. That can be certainly from an OpEx standpoint. That can be in terms of partnerships. It could be small and medium M&A through the lens of looking for opportunities to bolt-on to the work that we're doing. We've been successful on small or medium ones. And if we find something, that would be something that we would do.

Outside of that, focused in terms of employee equity dilution and making sure that we offset that, that is what our stock repurchases are mainly about. Our goal there is to make sure that we can limit that dilution. Our authorizations were coming towards

an end, and so, we refreshed our authorization for $25 billion that can take us further into the future of offsetting that dilution. So that would be another use in terms of our capital.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. Well, with only a minute left, I'd like to cede the floor to you for any closing comments or areas you'd like to highlight that we may have missed.

A - Colette Kress {BIO 18297352 <GO>}

So, a lot of focus right now certainly on generative AI. But remember, we're still in the early stages of AI as we see it today. I know it seems as such amazing work that the AI that has been demonstrated can do, but we do see this future not only of increased AI capabilities, but definitely accelerated computing for the long term. When you think through the $1 trillion installed base that is out there of x86 and the thoughts of how quickly people move to accelerated computing, it's front-and-center and a great opportunity for them now as they are moving to key workloads of AI, using generative AI, and continuing that path probably further and moving to fully accelerated computing throughout their data centers. That's what I think we see.

Q - Matthew Prisco {BIO 20801520 <GO>}

Perfect. Well, with that, we are unfortunately out of time. But Colette, it was a pleasure chatting with you today and thank you very much for joining us.

A - Colette Kress {BIO 18297352 <GO>}

Thank you. Thanks so much.