Nasdaq 42nd London Investor Conference

Company Participants

- Colette Kress, Executive VP & CFO
- Mark Lipacis, Analyst

Presentation

Mark Lipacis {BIO 2380059 <GO>}

Okay. I think that we're live. Welcome everybody, to the second day of the NASDAQ Conference virtually from London.

I would like -- I'm very delighted and honored to host NVIDIA's CFO, Colette Kress, for a fireside chat.

So I have a series of questions I'm going to ask Colette. And also there's an opportunity for you to ask questions on your interface and those will hit me by email, and we'll try to get to those also.

Questions And Answers

A - Mark Lipacis {BIO 2380059 <GO>}

So with that, we'll jump right into it.

And I would like to ask Colette, would you care to make -- start out with any introductory comments?

A - Colette Kress {BIO 18297352 <GO>}

Sure. Thanks so much for hosting us. We love this opportunity to get to a great wide audience on Europe as well as here in the U.S. It's been certainly a busy several months across the world, including NVIDIA. We have taken this time to announce our new architecture for the data center during this period of time. We have been able to announce the architecture, move it into full production and have it contribute to our Q1 results. Keep in mind, we just finished our Q1 of fiscal year '21, announced earnings several weeks ago. We are now into our Q2 and are continuing to ramp our overall Ampere architecture.

Not only is our data center business accelerating quite well with overall Ampere as well as just the use of accelerated computing, but we also have the opportunity with our other business lines, including gaming, to -- quite well during these very challenging times.

We had started off at the beginning of the year focusing on what the impact may have been with our overall COVID-19 that was impacting so much of the overall world. We were able to manage through both supply and demand challenges through our Q1 results and have provided guidance for Q2 incorporating that. What we're seeing right now is a strong desire to incorporate forms of accelerated computing in many of the different overall industries.

Our gaming is also a great area where we have seen advancements of number of gamers playing, those trying to game from -- at home during the overall lockdowns that we have seen.

But our overall innovations that we've done in different form factors to really help our overall gaming business as a whole.

So happy to answer a set of questions.

But as you can see, NVIDIA has been able to do quite well during this period of time that we're in now.

A - Mark Lipacis {BIO 2380059 <GO>}

Okay. Great. Thanks for that introduction, Colette.

I think what I would like to do is just to kind of level set here, just review what's transpired to bring NVIDIA where it is today.

I think if I go back, and I think about six years ago in 2013, the annual revenues for NVIDIA had been bouncing around $3.5 billion to $4 billion for the previous five years. PC chipsets accounted for about 1/2 of your business. Data center only accounted for about 3%. Gross margins were in the mid-50s. EBIT margins kind of struggled to get to 20%. That was six years ago.

So in contrast, this year, the Street's modeling revenues that are $10 billion higher than then, so about $14 billion. The gross margins are ten points higher to the mid-60s. Operating margins are 15 points higher to the mid-30s. What are the critical elements that have driven this transformation at NVIDIA?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So a great set of facts there in terms of the numbers and the overall growth that we've seen over the last decade. I have to step back and actually recall that many times, six, seven years ago, we had discussed the company and talking about our initial transformation that we were not a chip company, and we were not associated with overall PC shipments as being the largest overall correlation to our overall revenue.

After several years of discussing that, yes, we may have an interaction with general PCs, we were very specific in terms of the markets that we had chosen and the

markets that we saw in terms of growing. And we were more than just a PC company or a PC chip company.

You've seen us today transform, not only what we are doing in terms of with gaming and the overall gaming platform, but also what we have seen in terms of many other platforms used for overall computing. Whether that be servers or whether that thinks about the role of cloud computing in our everyday business today.

So as we've moved over this period of time, it is really getting a grasp on how the world would change for computing and the necessity of overall acceleration through that.

But it wasn't going to be a market that would be able to be built on just chips alone. We had focused and focused more than 10.5 years ago on also the overall software as an important piece.

The software begins with creating an ability to develop to ride on top of the overall GPU and stitching it together with many of the other components that we see in terms of in-use platforms. We created CUDA. CUDA is now available across every single one of our overall GPUs and is integral in terms of the work that we've done to expand to the markets that we've had.

So we've taken with the same overall unified architecture as a company, we have focused very specifically on areas that we knew acceleration would be very important and been able to expand.

We were quite fortunate to be in the right time at the right place for using acceleration to expand overall AI. AI for many years have been built as an area of focus for so many of the scientists and researchers that were looking on how AI would transform the world.

But the use of acceleration and the use of deep learning to build so much of what we see today in AI was a perfect match for the company.

We continue to focus not only on building superior chips, innovating with the overall design of our overall chips, but also very specific layers of software to support so many of the industries that we are.

We have also, over the last several years, realized that we also have an opportunity to accelerate beyond just the overall GPU and the GPU platform. As you've seen in the last year, now our focus was on the acquisition of Mellanox as well as several other pieces of the overall computing platform to look for acceleration possibilities and speed up of the work that can be overall complete in.

So I think through careful choice, careful choice of the overall markets that we have chosen but also a full look at a computing, we can now consider us more of a

platform play or more of a systems company than we actually look at as a chip company.

A - Mark Lipacis {BIO 2380059 <GO>}

Okay. That's a great explanation of how the transformation happened and what the drivers of it. I had the -- and that kind of leads me to the next question on the ecosystem.

I had the privilege of taking Jensen on a roadshow about five years ago. And in one of the meetings, when he was asked about competition, he said, AI is not a GPU, which I thought was a very zen thing to say at the time but it made me think about the importance of the ecosystem. And when we did the work, we found that in the PC era, the Wintel ecosystem captured about 80% of the value, generating the PC supply chain.

And then on the smartphone market, Apple is capturing about 80% of the value in the smartphone supply chain.

So my question is, do you think it's fair to say -- to compare NVIDIA's ecosystem to like a Wintel ecosystem in the PC era or the Apple's ecosystem in the smartphone era? Is that a fair comparison?

A - Colette Kress {BIO 18297352 <GO>}

There's some various similarities with some of the things that we saw back in that era. And for some that are just now entering, and maybe have entered the market since the overall Internet boom, can't have the full realization of some of those things that occurred at that time. It is a little bit different, though, now. It is different because there's also some underlying factors that are influencing what we see in terms of this next wave of overall computing.

When you think about overall NVIDIA, it is not that we have matched ourselves in the sense of the Microsoft and Intel time with a very specific platform or overall software. What this is, is the case of a very fast-moving industry.

The overall era of AI is continuing to transform every single month. And even in the last three years in terms of what we're seeing in terms of types of workloads and the types of solutions that are being built, it's very different.

So what you see is it's very hard to solidify on a static platform that takes you through all of those different changes, which, therefore, puts NVIDIA in quite a good position. NVIDIA's ability to be agnostic to all of the different types of computing solutions that are out there, the different CPUs, the different software stacks, the different middleware, all the different other components that may be inside of the data center, allows us and the overall user to have the most flexibility in an era where AI is changing.

So there was nothing that we did in terms of being specific and the only solution for AI. We are the solution, though, that is the broad-based for overall AI and allows from day one of thinking of AI to day 300 to continue to be used as the solution that allows them to expand their overall workloads. We're able to continue to work with all of the major frameworks that support AI. We're allowed to work with almost every single provider on the planet. Every single hyperscale, large and small across the world, is also focusing on the use of GPUs.

When you think about our A100 platform that we just announced, it's yet another evolution in understanding how people are taking accelerated computing into their overall data center. The A100 was built as a system architecture.

One, it comes with eight GPUs together with an underlying baseboard that allows it to be plugged into existing RAPIDS systems in an easy manner to improve the overall qualifications, allowing them to not have to swap out the entire pieces of all of the data center as we see the modernization of data centers that will be different pieces as a whole.

What we mean by that is you may have a stack of overall storage, a stack of GPUs and as well as a central area for your CPUs. In the form factor of A100, you're allowing an easy ability to put that incentive data center.

Additionally, A100 is also the ability to be elastic in terms of the type of use that you want to use accelerated computing. We know that AI -- we know that high-performance computing are important areas and workloads.

But many of our overall enterprises are focused, in terms of data analytics, are focused on the field of data science. And the usual platform that allows that scale out that scale up capability is very useful for both hyperscales, hyperscales for the cloud instances, as well as the enterprises.

What we mean by that is, at any point, you can choose to leverage the GPU for inferencing. You can virtualize a single GPU on that platform to seven unique instances for inferencing or you're able to stitch together all eight of the GPUs together to complete one large training data set as well.

You have the ability for the enterprises at the time of purchase to continue to split the use of the overall A100 across their business units, across their different workloads as they see and over a period of time that they only overall A100 to continue to revise for the different new market areas that they have.

So this is another era where our work and thinking through the future of accelerated computing continues to evolve. Probably very different since some of those early matches that we had seen back in the last 20 years in terms of the decade. It is a new world in terms of computing.

A - Mark Lipacis {BIO 2380059 <GO>}

What are the -- what would you consider the critical elements of the ecosystem you're providing to your customers? And what percentage of your R&D is spent on things other than chip design?

A - Colette Kress {BIO 18297352 <GO>}

So in our overall ecosystem, we can think about all of our different businesses and our connections with the ecosystem, each and every single one of our businesses has an ecosystem that we need to address that. That doesn't mean just thinking about the end customer and their use but the overall applications that allows them to use the overall compute underneath those applications and how that is put together and who are the key participants.

Starting already with overall gaming. We chose about five to ten years ago to focus on this ecosystem, to focus knowing that a gamer was not coming on board to look at their overall GPU card but actually was coming into the market because they wanted to play games. That ecosystem of how games are built and the types of games is where we had focused our overall energy and focus in terms of our connections and partnerships. Even when we brought overall real-time ray tracing to gaming, as we did two years ago, we were both the first entrant and we were very well pleased in terms of the ecosystem adoption and the thought about bringing real-time ray tracing to overall games. They needed a starter to do that. Meaning, the overall ecosystem would not have been able to manage bringing ray tracing to the market alone. They needed that partnership with the underlying compute, the underlying hardware to stitch that overall together.

Two years have passed. We now have every single game engine, every single game developer organization focused on how and to when to incorporate ray tracing in the overall futures of games. That puts us in, again, the top bar of both innovation, but really bringing what overall gamers like to see in the games. The more realistic, the more different types of games that can be built on ray tracing really expands the market of the type of entertainment and work that they can do.

That's one industry. If you think about that ecosystem even more importantly as we focus on data center and we focus on the types of applications that we use for accelerated computing. Our work began with high-performance computing. High-performance computing is really high computational, very difficult data sets that could, back ten years ago, take a significant amount of time to complete. Very important work, whether it be in oil and gas, whether it be in many of the medical fields as well in terms of the work that they had done.

But what we've seen now is understanding that we can use overall GPUs for inferring the overall answer -- overall AI. Our ability to allow AI to expand farther than just research and/or lab work, but for everyday use in terms of hyperscales as well as enterprises, had us focusing on the ecosystem and the top applications that could be overall accelerated. Now we don't build applications.

But what we do is we work together in both the middleware and creating the ease of with frameworks to attach ourselves to that ecosystem of applications so that the

user doesn't have to think about how that is put together.

Now our work is essential in this place. It has differentiated us from many of the overall players in the market that are really just coming to market with an underlying chip. Our focus is on these large workloads, carefully thinking them through.

Sure enough, the only way that they're going to be deployed and used is whether or not software has been stitched together.

Our company, although you may think of it as only a set of hardware engineers focused on designing the chip, working with our fab providers in terms of manufacturing, that's actually only a part of our overall engineering.

We have more software engineers than we actually have overall hardware engineers. That work together is both creating just base or system-level software, system-level software that allow someone to use any one of our GPUs for the work that they are doing but also the work that has to take place to stitch the overall application, rewrite sometimes that application to work with the overall GPU.

A - Mark Lipacis {BIO 2380059 <GO>}

On the data center side, I think there's a view out there that most of the coding on for data center acceleration platforms, and particularly in the AIs, done on higher level, what they call the deep learning frameworks like TensorFlow and Keras, which then could abstract the developer out of the NVIDIA's ecosystem effectively and it wouldn't make it easy to port any software developed on those deep learning frameworks to another platform should one present itself.

What are you saying to investors who present that concern to you? And what do you do to kind of insulate yourself from that happening?

A - Colette Kress {BIO 18297352 <GO>}

Yes. In the data center and in the field of overall AI, the overall open source frameworks that have been built are essential in terms of speeding up the ability to create overall applications specific for their overall industry. The hyperscales are very focused on these overall frameworks.

So what's interesting about them is there is a significant amount of work with those frameworks in terms of compiling the overall code, the data together that we do to allow those frameworks to work both efficiently and effectively on top of the GPUs.

We are essential underpinning below the overall frameworks. You're right, it's very common for developers to not understand the overall stitching together below, which we actually encourage. The thoughts that they can focus on their core competency in terms of working with the application and the data to solve overall problems and they are able to leverage a GPU to take them from workload to workload is great.

But it's not necessarily that easy to just lift and switch to some other piece of that, as you will see that through our overall frameworks as well as our frameworks with CUDA and CUDA-X, which is more middleware that's just together, those things are updated very, very often. And with the overall current discussion in terms of what is happening with frameworks.

It was interesting what you articulated as top frameworks today. Three years ago, that was a different set.

We continue to make sure that we are aligned with the most current overall frameworks as well as the most current form and use. Three years ago, we would have been talking about image detection and image categorization, and that would have been an important use of a certain overall framework. What we're looking at right now, which is a very hot overall workload, is the focus on conversational AI and the focus of recommendator engines. Very different overall frameworks, very different models that have been built to overall process natural language understanding. And the combination of these two together represent more than 50% of the type loads and workloads that are being done.

So anybody that's looking to take something that was built six months ago and put it to a different type of platform or chip is actually in the yesterday years because things are moving ever so fast to a new form of AI building larger, bigger data sets. The amount of work that is being changed on the overall frameworks as well as the underlying CUDA software is every single day changing.

So you're just really looking at people that are in the past and are not going to keep up with the speed of what we're seeing in terms of the adoption of AI.

A - Mark Lipacis {BIO 2380059 <GO>}

That's very helpful. I think another -- another kind of debate that comes up with investors is this, from a competitive standpoint and investors who think about how you could be disrupt -- how NVIDIA could be disruptive is this, the question about general purpose versus application-specific computing.

In one of the past analyst days, Jensen made the case that past computing eras were eras where we're dominated by general purpose computing platforms, and those are what became de facto standards as opposed to application-specific ones. And that NVIDIA's GPU are the accelerating computing platform, was best positioned to dominate because it was programmable and qualified under that kind of general-purpose computing platform idea.

The other side of the argument is -- which seems legitimate is that compute cycles are consolidating under seven -- like seven large cloud service providers, let's say. And those CSPs have massive economies of scale like no other end users in the computing industry in the past. And they can consolidate or centralize specific workloads and then benefit from an application-specific computing model by building custom processors or having merchant application-specific processors.

So the question is, how do you think about this? Why have the general purpose computing models dominated in the past? And does that apply today?

A - Colette Kress {BIO 18297352 <GO>}

Yes. A really interesting question. And it's been super interesting to look at history, why history has happened, and then also look at why NVIDIA has excelled through this overall period.

General purpose computing, I think it's an important perspective that says the ability for us to lead to adoption and lead to inherit the number of researchers and the number of developers that we have on our platform was because of the general purpose nature of it, right?

So we have reached to a point now that we have more than 1.8 million overall developers on our overall platform. What they have done is they have congregated toward the overall NVIDIA GPU platform because of its universal use.

Yes. There's a lot of different hyperscales. There's a lot of different types of platforms. But they know that they can program equally across overall software set that allows them that ability.

When you are that general purpose, it's an important that you demonstrate the overall developers and researchers to continue to expand into the workloads that you want to go into. That has led us not to be application-specific because the developers will choose the application to use the overall platform rather than the other way around.

So this developer base was extremely important to link together to many of the applications that they knew well and understanding the overall importance of accelerated computing in an era where Moore's Law was going to fall and become obsolete.

So the use of the developers, them overall thinking of the work that needs to be done to creative applications, allows us a leverage model about where to invest and where to spend our time. Even when you think about a world right now where you have a lot of hyperscales, that's great that the hyperscales -- and you think that they would have the ability to build out custom overall infrastructure.

But the reality is that's not always their core competency when they know that we are a full house of overall engineers that understand at the base chip design, overall all the aspects of building the thermals, all of the different uses and cases of how to improve the performance of the GPU. They turned to us to create that universal capabilities to use a GPU.

Some of them may have a very specific workload that they may use a custom chip that they have.

But serving all of the universe of developers out there, a universal platform but also a universal development platform, is essential for the breadth of the types of workloads that we're seeing that just wouldn't have been capable anybody that shows only application specific. You would have had to have invest tremendously in-house on a non-open application-specific era. And that's just not what we did. We kept things as about as open as possible, as agnostic as possible to support the growth of both AI high-performance computing and many of the workloads that you'll see in the future.

A - Mark Lipacis {BIO 2380059 <GO>}

Our field work indicates that NVIDIA accounts for over 85% of the acceleration instances in the market at the top cloud service providers. And at the same time, you have Amazon who is developing its own chip, Inferentia, that's in the market that -- our data suggests that's around 1% of instances available or 2%. Google has the TPU, which is about 1% of the instances available.

What is -- is this -- in your mind, is this kind of the biggest risk from the potential for other solutions to consume the acceleration cycles that are out there in the market? Is that your solutions from your own customers?

And what is your -- the data we have, we can only get what's in the public market, which are the instances available, what is your competitive intelligence telling you about the internal workloads? Do you think your -- as your biggest customers, do you think your share is equal to the external workloads on instances?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So it is an evolving era of the types of work -- the types of workloads that are being done by the overall hyperscales. We've been a very big part of the growth of overall AI with the overall hyperscales. They were the very first set of customers. Many of them did find an opportunity to try out their ability to create custom for some of the workloads that they have seen.

But you're correct, it's a very, very small share. And then one of the main reasons why that is, is most of the work that they've done on custom is a custom ASIC. They are focused in terms of -- and the design process to design something for a specific workload that they're seeing. Their speed to market on bringing an overall custom ASIC may be good, maybe a reasonable amount of time.

But keep in mind that the trends of the overall ASIC comes out, there's no ability to go back and revise the overall chip to revise for changes in terms of overall workloads.

So they need to be relatively confident in terms of the type of workloads that it will be able to be used for.

Even when you think about the overall A100, many years in the making of developing that architecture for it to probably be quite a big surprise for the overall

industry. Why? A100, just in three years, is a 20x improvement in performance versus the last generation of V100. Our core competency is working on top performance, not just top performance in speeds and feeds, but top performance on many of the different workloads. We've talked about it on the ability to, with its overall Tensor Cores and to really think about the floating point precision necessary for inferencing to dial up or down just depending on that type of workload. That would continue to gravitate people to using overall GPUs as a large percentage of the workloads that they're doing versus using an overall custom ASIC.

We believe we also have that ability for whether it be enterprises or overall cloud users to not only use the GPUs in those cloud instances, but also the importance of the edge.

Some of the instances that you talked about with AWS or Google are focused in a captive space versus our overall GPU to serve a very large market that we believe will exist on the edge is important. Not everything will be with the hyperscales, not everything will be in terms of a cloud incidents. Having a platform that is both scalable on the edge for the different types of workloads that will exist is an important piece, again, why people have continued to choose overall NVIDIA as well as NVIDIA GPUs in this case.

So from time to time, we'll see a custom ASIC. Oftentimes they're captive. They're probably focused on a very specific overall workload.

But the overall speed of adoption of AI and the growth focuses people to gravitate to a platform that will scale with the growth that we're seeing in that market.

A - Mark Lipacis {BIO 2380059 <GO>}

Yes. I want to kind of pick up on the point of Ampere because I think, if I think about what the solutions that have evolved or your platform over time, how it's involved, the initial products, maybe, let's just say, five or six years ago, you seem to be very well suited for training. And then you came out with some products that were -- seem to be focused on inferencing. And the Ampere solution seems to be -- it seems to be optimized for both, given its scalability.

From my standpoint, it seems like this solution fits completely into this idea of offering a general purpose of acceleration platform that does both training and inferencing.

So two questions on the Ampere then is, is that the right way to think about it, that this is -- that this is doing both, and it fits into this general purpose computing model? And at the end of the day, pick a time -- three, five years from now, ten years, is -- will all your chips be Ampere-like where they do both the inferencing and the -- offer best-in-class inferencing performance and best-in-class training performances?

And then your training-specific solutions go away and your inferencing-specific solutions go away, and you end up with a single acceleration platform on one

solution?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So really, a really important understanding of how we have entered into the overall inferencing market. Let's go back in terms of a couple of years ago.

We had discussed the size, opportunity in front of us in data center. And the size in front of us incorporated both training, high-performance computing, two areas where we had quite leadership position, but also our discussion of the importance of overall inferencing.

A lot of people discuss this that say, wait a minute, why? Why will people move to an overall GPU? People use a CPU for inferencing. You're going to have to need a different form factor. You're going to need to be smaller. You're going to need to focus in terms of the wattage that otherwise, you know what, the CPU will just be fine. We took on that piece as a very important understanding of how inferencing would change and how it has changed over the last couple of years.

Our thoughts in terms of working with many customers is the types of the inferencing that we're doing historically was really based on the type of compute they had. They said, well, the compute can do this. Therefore, I will create inferencing to support that. What we mean by that is rather overall simplistic types of inferencing for mass amounts of data but very binary types of responses as necessary.

What we saw is the output of training. The output of training for AI would create phenomenal overall data sets that new information coming in would need to go through that overall inferencing model, but a standard overall compute chip would likely be too slow to respond back for the overall needs of the overall inferencing workload that's there.

What we see right now is conversational AI -- for example, is a very phenomenal AI workload. It's an AI workload taking in one of the most challenging parts of overall understanding data, which is understanding natural language. Understanding natural language, what is said, deciphering in terms of the pieces of the meaning but also being able to respond in that language as well in a conversational manner. That takes training. That also, therefore, takes inferencing at the side.

But speed and performance are important. You have to have both the ability to speak multiple languages with your overall workload, but you're looking for a couple hundred milliseconds type of response.

So our entrance into inferencing was an extension of what we saw in terms of training. We have built now inferencing to solidly be in the double-digit percent of our overall data center business in just a couple of years. We have it growing more and doubling year-over-year in just this last quarter. With the ability to bring something like A100 to the market allows us to, again, not have the overall customer

choosing at purchase what the workload that they want to do and allow them that continuation of building their training models to move into overall inferencing.

As we go forward, it's going to be hard for us to determine specifically and model how much overall inferencing is of the workloads that we're selling, but we have the ability with overall A100 to meet the multiple needs of the end customer, whether that be a hyperscale or that be an enterprise and an enterprise on the edge as well.

Now will that mean that everything that we do moving forward would be only A100 types of platforms? Now we believe that the overall acceleration will meet so many of the different types of applications and servers going forward. And it someday, way in the future, almost everything will be accelerated. Over this period of time, we may still have different form factors outside of A100.

Keep in mind, we have A100, which is a platform. We also sell full systems. Full systems such as the DGXs. Why? And why do we do that? It is the ability for those that score competency is not to focus on the infrastructure to get an end-to-end configuration that allows them to plug-and-play, to focus on their application and focus on the work where they have their overall core competencies.

But similarly, we have -- may have the opposite, that we may say, we will provide you a specific overall GPU for specific workloads that you have in terms of overall volume.

So we'll see how the market plays out in terms of there, but I think we'll have a full host of different platform solutions and not everything will look just like an A100.

A - Mark Lipacis {BIO 2380059 <GO>}

Okay. That's clear. I want to shift gears a little bit to what's going on with -- on the macro front in China. How have the recent trade tensions between the U.S. and China impacted NVIDIA's business? And how big is China as a consumption market for you?

What's the risk that the U.S. government determines that NVIDIA GPUs are enabling something that's deemed to be politically unacceptable and then block shipments of NVIDIA products to China?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So let me first start off on China as a whole as a customer for NVIDIA. China is an important customer to us as a whole. They are more significant on our overall gaming business than they are on the rest of our overall businesses.

In our overall gaming business, the PC platform, the PC platform is a very integral part of the China household and many of the overall workloads across the overall nation. What we mean is, again, you kind of stem back to because we were universal in nature, people leveraged an overall PC to do much of the work from an overall

consumer user and is also an important piece of what they do in terms of overall business. They were some of the early starts of the high end type of gaming and using an overall PC platform. Now the PC platform has become a very important and probably the most built upon platform that people use for overall writing overall games.

But keep in mind, there is a statement there that says, it's a universal platform going into all of the areas of the world. It is just an underlying GPU with inside a PC and/or a laptop. It is not geared to any specific overall workload, which leads us to what we see in terms of in the data center.

In the data center, we have also worked with the overall hyperscales and many of the Tier 2 Internet provider -- consumer Internet providers in China. Not to the nature of what we have in the U.S. because the U.S. hyperscales have much larger overall CapEx budgets than what we see in China. We are an important part. We are an important part, again, as a universal platform also in terms of in the data center and the use of acceleration.

As the tensions between U.S. and China continue to go back and forth, it is more of an industry understanding than it is specific to overall NVIDIA. We're very similar to a memory supplier and/or any other type of part in terms of consuming into overall computing would be overall affected. We stand and watch to say there's nothing about our overall GPUs that are specific regarding that U.S. and China. It is more of just compute as a whole. We watch it carefully. On this part, we watch it with our overall industry leaders as well to focus on that.

But so far, our use is really just universal compute in the overall nation. And we have been able to continue to sell into overall China with the exceptions of specific overall entities that the U.S. government is concerned about, and we do make sure that we are following the U.S. laws regarding those and selling to those that are on the overall entity list.

A - Mark Lipacis {BIO 2380059 <GO>}

And I think this is -- that's very helpful is -- I think this is a related question on China. TSMC is a critical supplier to NVIDIA. If there was a political situation that developed where NVIDIA was blocked from using TSMC, what's NVIDIA's wafer foundry contingency plan?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So several years ago, probably more than five years ago, our path on creating multiple foundries or multiple foundry manufacturers was important for us. It was important to us from an enterprise risk standpoint. Enterprise risk that says, oftentimes in many different parts of the world, global conditions change, whether it be weather conditions, geopolitical things change and/or just basic oversupply and demand with many of them changes.

Our thoughts on creating another foundry was an important part of that. Now keep in mind, it wasn't an overnight decision nor was it something that we could put in place quickly.

So over those years, we now have a dual overall fab support model. We use TSMC. We also use Samsung. In many cases, we will take an architecture, and we will work in terms of the development of that architecture with both fabs. We can and could choose to split an architecture across our fabs and not just have singly on one overall fab provider.

To be clear, TSMC is a great partner. They have been with us since the beginning of time. We truly appreciate that partnership. That partnership is not just focused on their machines and their ability to spit out chips. It is really understanding their processes and how they work. And we're pleased to have now also built that with Samsung. And we have very similar but a different partnership and a different process as it relates to overall Samsung.

As we look forward, we like the ability to have overall two fabs. We will also continue to support something that may be pulling TSMC to a U.S.-based overall fab as well. We'd be supportive, and we're excited to see at the next term of where that is.

But our model today of -- at the high end that we are, the high end types of chips and the complexity of our chips, we're extremely pleased to have two fab providers.

A - Mark Lipacis {BIO 2380059 <GO>}

Excellent. We're coming at the end of the time, but I'm hoping that I can sneak in one more question here. Most investors I speak with really focus on the data center. And the data center market has historically been a lumpy one. And in 2018, it kind of shift -- you shift above trend line. And then the data center business declined dramatically in the first half of '19.

What is -- you talked about good visibility into the July quarter. What is -- can you talk about like the October quarter? Or can you just give us a framework?

Like when do you -- when does your out -- the next quarter of visibility, when does that start to kick in? And is there a risk that you're shifting it above trend line now and you have a -- you kind of have a reset just due to the natural lumpiness of that market?

A - Colette Kress {BIO 18297352 <GO>}

Yes. In the past -- and we've probably talked about this quite a bit over the last five years. People wanted to know when will we reach seasonality in data center, will it be reach seasonality? How can you assure that the growth, as we see, will continue quarter after quarter?

But there's nothing that we can guarantee from quarter-to-quarter sequential overall growth. We've done a pretty phenomenal job, almost consistently having growth.

But you're correct. There were periods of overall digestion, digestion in our industry. Even during that digestion, it was an important era where people wanted to know, well, when are we going to get back to growth? It was an era of time for us to focus on the engineers and continuing to expand our partnerships and our customer base to incorporate the work that we knew the enterprises we're working on and the growth that we saw in terms of the adds.

So where we sit right now at the end of Q1, our hyperscales, where originally, three or four years ago, were probably the majority, the lion's share of our overall data center business, are less than overall 50% of our overall business in data center. What's interesting on the other side is our enterprises, our research of high-performance computing, supercomputing is more than 50% of our business.

As we work in terms of building out breadth and depth of the type of workloads that are using acceleration and the customers, we will hopefully be able to see something smoother.

Right now, keep in mind it is still a stack of project by project. This is not a business model that we're building inventory and/or our customers are with inventory. They are specific on the projects that we are doing. What has been great over this period of time is we have seen more and more projects per customer. It's not just one project, and they are sometimes overlapping in the time frame in terms of their deployment.

Both our customers as well as us would like to improve the overall planning. Planning data centers is hard they know they have inefficiencies. We know that we have inefficiencies in terms of working with them, but working to clearly work through that road map on how they deploy the data centers is best-of-breed. We're ramping A100 right now. It is only a couple of months in, and it's got a long trajectory of years in terms of when it's being deployed. Even in the quarter before the launch of A100, we reached record levels of V100 and record levels of T4, even though it was in market for a couple of years.

So over this period of time, you will see us selling A100. You will also see us continue selling V100 and selling our overall specific inferencing product in T4. And as we look, we will continue to try and work through that visibility. Our visibility into Q2 is solid, very similar to the visibility that we've had in terms of Q1. We'll turn the corner as we get closer to Q3 and talk about that over visibility.

But I think it's important to step back and look that says, A100 is going to be with us for a while. And you know that we're not done with everything that we have coming on an overall architecture.

So there's great things to see. I think we're very pleased with the -- both the performance of the company as we started out today's call. That during COVID-19, the overall decision of the areas of focus and knowing that AI and accelerated computing is still a top priority to our customers during this very unique time in the industry and the world as a whole, I think, puts us in a great place as we move to both the second half.

But at this time, we're just going to have to wait until we get to second half before we can discuss that overall outlook that we have. Okay?

A - Mark Lipacis {BIO 2380059 <GO>}

That's very fair and very helpful. And we ran over. I think that will have to be the last word.

Colette, thank you very much for joining the fireside chat. That was extremely informative for me -- I'd imagine for everybody else.

Everybody else on the line, thanks for dialing in. Have a nice day.

A - Colette Kress {BIO 18297352 <GO>}

Thanks, Mark. Appreciate it. Take care.

A - Mark Lipacis {BIO 2380059 <GO>}

Bye-bye.

This transcript may not be 100 percent accurate and may contain misspellings and other inaccuracies. This transcript is provided "as is", without express or implied warranties of any kind. Bloomberg retains all rights to this transcript and provides it solely for your personal, non-commercial use. Bloomberg, its suppliers and third-party agents shall have no liability for errors in this transcript or for lost profits, losses, or direct, indirect, incidental, consequential, special or punitive damages in connection with the furnishing, performance or use of such transcript. Neither the information nor any opinion expressed in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP. © COPYRIGHT 2024, BLOOMBERG LP. All rights reserved. Any reproduction, redistribution or retransmission is expressly prohibited.