GTC Financial Analyst Q&A

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer
- Jensen Huang, Founder, President and Chief Executive Officer
- Simona Jankowski, Head of Investor Relations

Other Participants

- Aaron Rakers, Analyst, Wells Fargo
- C.J. Muse, Analyst, Evercore ISI
- Harlan Sur, Analyst, J.P. Morgan
- Joseph Moore, Analyst, Morgan Stanley
- Mark Lipacis, Analyst, Jefferies
- Matthew Ramsay, Analyst, Cowen and Company
- Rajvindra Gill, Analyst, Needham & Company
- Stacy Rasgon, Analyst, Bernstein Research
- Timothy Arcuri, Analyst, UBS
- Toshiya Hari, Analyst, Goldman Sachs
- Vivek Arya, Analyst, Bank of America Merrill Lynch
- William Stein, Analyst, Truist Securities

Presentation

Simona Jankowski {BIO 7131672 <GO>}

Hi everyone and welcome to GTC. This is Simona Jankowski, Head of Investor Relations at NVIDIA. I hope you all had a chance to view GTC's news fact keynote this morning. We also published several press releases and vlogs detailing today's announcements. Over the next hour or so we'll have an opportunity to unpack and discuss today's news with our CEO, Jensen Huang and our CFO, Colette Kress in an open Q&A session with financial analysts.

Before we begin, let me quickly cover our Safe Harbor statement. During today's discussion, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to our most recent Forms 10-K and 10-Q and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

Questions And Answers

A - Simona Jankowski {BIO 7131672 <GO>}

With that, I would like to welcome you to the Q&A session with Jensen and Colette. We'll be taking questions over Zoom. I think you're all familiar with the interface, but as a quick reminder, please use the raise hand feature on Zoom if you'd like to ask a question. Then unmute yourself when called upon. Let me pause for a moment here to review the queue before we approach our first question.

And our first question will come from Toshiya Hari with Goldman Sachs. Please go ahead.

Q - Toshiya Hari {BIO 6770302 <GO>}

Hi. Can you hear me okay?

A - Simona Jankowski {BIO 7131672 <GO>}

Yes, we can.

Q - Toshiya Hari {BIO 6770302 <GO>}

Okay. Great. Jensen, first of all thank you so much for the keynote. I'll probably have to wanted another 10 times to fully digest everything that was announced there. But just on the Omniverse it seems like your customer engagement has broadened very significantly since you provided us with an update last time. Can you remind us how you're thinking about the contributions to your P&L? Am I realize this is relatively nascent and this is a very long-term strategy for you guys? But how should we think about -- as analysts how should we think about the contributions to revenue growth and your profitability going forward?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks for that, Toshiya. There are three components to the Omniverse platform. The first component is the simulation platform, it's called OVX, the Omniverse computer. And OVX had a first generation prototype, but the volume scale out is based on second generation OVX-2 which is powered by Ada, L40s, and CX-7 and BlueField-3. And without belaboring the reason why in terms of performance is a giant leap over Ampere as you saw in the keynote. It is all based on your rendering and the performance is really incredible. And so we're in full production with that now, we're ramping and trying to get the OVX computers to customers as fast as we can. We have quite a large number of customers signed up to receive OVX computers.

The second component has to do with the Omniverse applications that are built on top of it. And there are applications that span the entire range of design, build and operate inside a company. So when you're designing your car or designing your factory or designing your product all the way to manufacturing it to operating it, Omniverse will be involved. I believe that Omniverse will be one of the first

enterprise applications next to the web browser, if you will, next to web applications that spans practically every organization. In fact, we showed you few examples where design was involved, marketing was involved, product configuration was involved, manufacturing was involved in simulation and operations as well.

And so it's just about every single organization has the opportunity to engage Omniverse and work on a single source of truth. So the second is applications, they're typically by user. The third application is a new type of database and it's called Omniverse Nucleus, and we have just put that up in the cloud and we're going to host that as a managed service. Think of Nucleus as a database, but is a new type of database because it's interactive, it's shared. And so you pull you bring in your -- the data -- the three dimensional data or metadata or whatever or behavior data, physics data or relationship data, supplier data which component goes with and which supplier, et cetera. And you bring in all of that data associated with a product or a building where it could be a (inaudible), it could be JT files, which characterizes the 3D geometry. It could be a ERP system to associate which vendor goes with which component. And you bring in all of that data into Nucleus. Nucleus is shared by all the people that use it, it's an active distributed database, it is in 3D, it's the world's first large scale USD database. And that business model will probably be like a cloud database business model. And the more people that are connected to it, there'll be a storage component associated where there'll be a used component associated with it.

And so Nuclear is -- think of it as a cloud database and we just we announced today that, Omniverse Cloud would be hosted in the cloud and that Omniverse cloud is basically the database. So we're in the process of fine-tuning the pricing associated with each one of the -- in the case of the system that part of is very well understood. In the case of the pricing of the applications and services we're in the process of fine tuning that in the case of the nucleus database in the cloud, were in the process of fine tuning that as well. But from my sense is that it will be very -- if you will, conventional compared to the likes of things that I've just described.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. Our next question will be from Tim Arcuri with UBS.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. Can you are hear me?

A - Simona Jankowski {BIO 7131672 <GO>}

Yes.

A - Jensen Huang {BIO 1782546 <GO>}

Yes, Tim.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Perfect. Awesome. So Jensen, I think a big theme here are really these new infrastructure as a service offerings with the BioNeMo and the Omniverse Cloud. And I guess I had two questions. First, maybe Colette, update us on -- I think you said back in March, I think you said $100 million a year revenue run rate that you gave for recurring software and services revenue. So first question is, can you give us an update on that?

And then second of all, Jensen, I'm sort of curious about the business model for this stuff. Are you going to just offer instances on AWS and other clouds or ultimately it seems like maybe you can offer your own cloud for this? And it seems like maybe this is sort of an inflection for you to look more like CSP or a hyperscaler in itself. Thanks a lot.

A - Colette Kress {BIO 18297352 <GO>}

So, Tim, thanks for the question. Let me first start with your first part of the question, as we had talked earlier about, we already have a run rate of software and services that we provide. That number is in the couple of $100 million and we're going to continue to grow from that. We have great new offerings even today, but also offerings that we've been working on for some time as well.

I'll move to Jensen and he can talk a little bit more.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Remember accelerated computing is a full stack computing approach. The method of using brute force transistors and the advances of Moore's Law has largely ran its course. Going forward, the opportunities for continuing to variety, price performance curve of Moore's Law has ended. And so if you wanted to be able to do larger scale computing and to do it in a cost effective way, after 15 years -- almost 20 years of pursuing is already computing. I think the very, very broadly, almost it's conventional wisdom that accelerated computing is really the path forward and it's an opportunity for us to not just stay with Moore's Law but go into a much more turbocharged accelerated computing law. And artificial intelligence, of course has benefited from that, molecular dynamics has benefited from that, weather and climate simulations is going to benefit from that. There is so many different things -- different fields of ray tracing NVIDIA's own business, core business of computer graphics has benefited tremendously. And so we -- the first part, it's a full stack challenge.

The -- our architecture is available in every cloud and our partnership with cloud vendors CSPs is really in two parts. There is the internal consumption part, which is about using NVIDIA's accelerated computing stacks to accelerate their workloads. It could be recommender systems, it could be speech AI, it could be very large scale queries or and now of course the emergence of large language models, which is on question with the most important AI model of the decade. And if not the most important AI applications of the decade. And so that's for internal consumption. For external consumption in the public cloud, NVIDIA is a partner. And if you will, maybe even an extended sales and marketing force of all the CSPs because we -- through our ecosystem, through our evangelism of the platform and through the software

developers and all the startups to the 12,000, 13,000 startups that are work -- that are built on NVIDIA. And all of these companies that are using NVIDIA's accelerated computing go into the cloud. We are essentially the business attractors of a very, very significant part of their public cloud service. And we're going to continue to do that.

There is several areas where we believe that we might be able to simplify further and democratize to reach of NVIDIA's accelerated platforms, because it's fairly complicated to put these systems together into -- and couple it up yourself in the public cloud. And those two areas associated with large language models, which we announced today, that we will have NVIDIA managed services. And these managed services would basically be, think of it as AI training, but is a domain specific AI training is designed to be super good at large language models. And so the effort and the cost of training with NVIDIA services running in these clouds and we'll run initially in -- we're going to run our services in cloud as we can. And when they use our service, they could substantially reduce the cost of training a large language model or training what is called a prompt, how to tune that large language model for your specific application.

And then the second thing is Omniverse which requires just a giant amount of engineering we've been working on for a year to bring Omniverse into the cloud and we have currently running on AWS. But our goal is to have it run in all the cloud, so that all of our partnerships with the CSPs have the benefit and opportunity to attract customers that are using Omniverse. And so, in a couple of different areas, large language models per se for -- looking for English, well for language and also for chemistry and biology. And then also Omniverse those platforms are so complicated. We thought we would stand up NVIDIA managed services and make it a lot easier and cost effective for people to use it. And we'll probably do a lot more of that going forward, especially in the areas that are very, very hard to do.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. Our next question will come from Will Stein with Truist.

Q - William Stein {BIO 15106707 <GO>}

Great. Thank you for taking my question and I'll add my thanks for all the incredible announcements you made today. Some of them, maybe a little bit confusing to us, so many in there -- quite technical. But Jensen you made one that was curious where in -- I think related to autonomous driving transitioning from (inaudible) Thor. Can you detail what conditions led to that decision? And maybe also remind us of your progress with your big customer announcement from about a year ago, I think what you talked about Mercedes-Benz adopting the technology and maybe remind us if there have been similar relationships that have been built to the same level with other OEMs that we might have missed. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot Will. I'll roll backwards. Mercedes-Benz ships the first car in 2025, late '25. And we also announced that JLR, all of the brands of JLR in their entire fleet of all the brands are going to also be powered by NVIDIA's full stack and that is

shortly after 2025. And so we're pushing forward in both of those. We also have Orin designed into about 40 different cars and companies. And not to mention medical instruments and IoT Edge AI servers and in robotics of all kinds. And so Orin and the chip that is in the self-driving -- in our self-driving car stack is really just a phenomenal homerun and it started ramping a couple of quarters ago and it's going to be ramping quite fast going forward from here.

And so I think there is something like $11 billion with the pipeline in the next several years, that is associated with Orin and the systems associated within the software associated with it. But the reason why we decided to change Atlin [ph], which was a next generation Orin to a brand new architecture is because the three processors that are inside a robotics processor. One of them is a GPU, one of them is a CPU, and one of them is our Tensor Core. These three processors made such enormous leaps in the last two years.

We made the hard decision to swarm for and put these three new technologies into it, rather than waiting another two years. Because robotic system has a cycle time about two years. Every two years, we announced a major leap. And so we just, we didn't want to just miss it. With Atlin, that the previous version, which is based on the next generation of Orin, what was just missed it. And so we decided to bite the bullet and really just hunker down and work hard and it was just too much for us. We couldn't bear waiting another two years, I mean that's kind of the simple fact of it.

All of these projects, their projects at the heart and you love it and you have to love it building it. It's incredibly hard work and there is a lot of imagination that goes into it, a lot of passion and hard work that goes into it. And I just couldn't imagine waiting another two years to get Hopper in there and Grace in there and Ada in there. And so -- so we decided to just do it now. And so that's the reason why it's is probably -- it was probably if I could say 60% passion and just couldn't bear not bringing that technology to the world and 40% just hard work, it was a lot of hard work, but the passion overcame the hard work.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you, Jensen. Next question will come from Aaron Rakers with Wells Fargo.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thank you. Can you hear me?

A - Jensen Huang {BIO 1782546 <GO>}

Yes, Aaron, nice to hear.

Q - Aaron Rakers {BIO 6649630 <GO>}

Thank you. Thanks for doing this call. I want to go down now that you've talked about today Hopper being in full production which accounts like, I guess, maybe first question on that is just to confirm that the pace of the Hopper ran [ph] it is now largely as anticipated, given some of the questions around the China export

situation. But then the second question to that is that, when we think about the Hopper product cycle and you think about the breadth of the platform strategy that NVIDIA is built out, how can we think about the pricing strategy of Hopper relative to Ampere? And the opportunity they just continue to take a bigger piece of that data center footprint as Hopper materializes likely over the next several quarters?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. So first of all, Hopper is about five times the throughput [ph] and because of a new type of engine called Transformer Engine. And Transformers is the most important model of today. Really largely replaced (inaudible) and LCMs and are surely will also replace a lot of the computer vision-based algorithms. And so it's a 5x speed -- it's a 5x increase in throughput, it's a 3x increase reduction -- excuse me, 3x reduction in total cost of ownership. The reason why is because, inside the data center, aside from our Hopper, which is not a chip as you know. If you look at the Hopper system, it's an entire system that we built. And then we disassembled it and took out the most complex part is called the Hopper HGX Board which has eight Hoppers in it, tens of thousands of components. And it's just insanely complex and we ship the whole thing as Hopper (inaudible). And that system goes into a data center, and the data center includes cables and networking and switches and power supplies and storage and so on and so forth in power delivery and the infrastructure and so on so forth. And so we have 5x to throughput, we have three times lower TCO which implies that Hopper's price is higher than Ampere. And -- but I think the value proposition is so fantastic. That is it net reduces the cost compared to last generation substantially as I just mentioned.

Let's see the ramp, we are in full production when we ship some quantity this quarter when we ship most of the quantity next quarter. The signals from the CSPs and the OEMs and the enterprises are really solid. The reason, one of the primary drivers. And Ampere didn't have this benefit at its ramp, but Hopper has this benefit, which is a brand new revolutionary new model called Transformers. And you probably have been looking at seeing on the web, the impact of the unreasonable performance, if you will, of large language models and being able to learn new skills with just a few shots of learning a few examples of learning. And secondarily, its ability to generate images. The number of applications that are coming out all of the industries is really quite incredible. And so we have the benefit this time of having almost an industrial wide recognition that large language models has democratized AI, made it easy for almost anybody to use AI. And so Hopper is really, is going to -- not only advance large language models but to democratize the application of it, because the inference cost is so much lower.

And then not to mention one of the most vibrant industries at the moment and probably the most incredibly well funded segments of the worlds startups is digital biology. And all of you that are watching other industries you probably recognize a digital biology just went through its revolution between the cost of gene sequencing and the breakthrough in predicting structure of proteins and structure of chemistries and to be able to understand the language of biology and chemistry. But is has turbocharged drug discovery segments and therapeutic segments. And so you're pricing a lot of start-ups go in there and then remains really, really vibrant. And so anyways Hopper is going to be revolutionary for all of these applications.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from C.J. Muse with Evercore. Please go ahead.

Q - C.J. Muse

Good morning. Thank you for taking the question and congrats on the formal launch of Ada Lovelace. On that front would love to stick with gaming, Jensen. Now that you on your third generation -- ray tracing. Can you talk about the competitive landscape? And then I guess, maybe bigger picture considering kind of the inventory correction that we're going through. Can you speak to kind of your anticipated ramp timing for 4090, 4080 and then perhaps how we should be thinking about 4070 mainstream coming online? Thanks so much.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. We are -- we took action in the last quarter and this quarter to prepare our channel for a great launch. And as you know, we've substantially reduced selling of Ampere to allow the channel to normalize. And the market for -- that the end markets are soft but they're not so soft that it wouldn't allow us the opportunity to sell through the excess inventory we had in the channel. We took specific action in marketing programs to particularly reduce the segment that Ada is going into, initially. We typically --and Ada will be no different, typically ramp from the top down. That's where the enthusiast would like to see brand new products and the customers that refresh more frequently every couple of years or so, I would like to see their new products. It's also the segment where we need to ramp for Omniverse and Omniverse workstations and Omniverse servers and so on and so forth. And so it was a sensible place for us to ramp first. And so we'll ramp Ada nicely, starting a little bit this quarter but largely next quarter and very, very robustly going into, leaving the year and going into next year. So that's our current execution plan, but we're in a really good place at the moment.

I didn't mentioned, you asked about competition. Let's see. I think it's fairly well known that we're quite far ahead in ray tracing. RTX is about two things, about three things I guess. The first is programmable shading that we invented some 20 years ago, 20 somewhat years ago. We augmented programmable shading with RTX which has two new processors. One is a hardware ray tracing processor and we're in our third generation of that. And the second is our Tensor Cores, our AI processors. And I guess we're in our fifth generation of that.

The AI work that's necessary for neural rendering, neural graphics, infusing artificial intelligence and ray tracing and programmable shading. We are just miles ahead of. And you saw some of the keynote today, to keep in mind that everything was rendered and everything was rendered in real time. There were not offline renderings of like movies, these are computer graphics simulations. And to be able to see something like Razor-X [ph] with no offline rendering and everything is a no-baking that most video games do today and to be able to do that in real-time is really quite just unbelievable. And so I think the Ada is a giant quantum leap forward and it was -- it's a bigger leap, if you look at all the performance, the power, the efficiency, every aspect of it. It's a bigger leap from Ampere, then Ampere was from

Turing. And so that kind of tells you something, the third generation get a lot of things right and pulled a lot of things together.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from the Vivek Arya with Bank of America.

Q - Vivek Arya {BIO 6781604 <GO>}

Great. Thank you so much, Jensen and Colette, for the keynote and for the opportunity to ask a question.

Just wanted to clarify Jensen, just based on the comments you made about gaming is $2.5 billion still kind of the right run rate of end demand. I understand you were shipping below that. So I just wanted to clarify if that is still a reasonable number to use? Or what the actual end demand as of your gaming product? But my question is, on the Hopper -- on the Grace CPU, what are NVIDIA's ambitions with that? Are you planning to kind of just focus on the HPC segment? Or do you see the chance to take it across the entire 30 billion right time for server CPUs that is out there? And then if the ambition is kind of more narrower than that, is it because of a technology issue? Is it a cost issue, right? Is it a supply issue? Like what prevents NVIDIA from going after the entire 30 billion or so TAM [ph] for server CPUs?

A - Colette Kress {BIO 18297352 <GO>}

So let me see if I can start, Jensen, and kind of clarify for the (inaudible) what's the 2.5 billion of demand for gaming. In our earnings release, we discussed looking at a normalized demand for our gaming. Given that we were planning on under shipping for the quarter. And so looking out Q2, Q3 combined, we believe that the underlying sell-through demand is approximately valued at about $5 billion. And then yes, you can split up between two quarters. Seasonality sometimes outplays into other quarters, so approximately 2.5 billion plus and minus. So I just wanted to make sure we had that background for those on the call.

A - Jensen Huang {BIO 1782546 <GO>}

Grace Hopper, I highlighted a particular problem, that is giant in scale and probably the most valuable software in the world today. A long-time ago Windows was the most important software -- it was the most valuable software. And then you could argue that a decade or so later Page Rank was one of the most valuable pieces of software on the planet. And then now it's a recommender system. It drives the vast majority of the worlds e-commerce and just buy everything that is put in front of our small screen from the trillions of things that it could have selected, it was recommended and ranked for us by a recommender system. This -- that the amount of data you could just imagine is enormous and there are many recommender systems, there is just not just one recommender system, but almost everything that pops of an ad or puts up a product or puts a price up or ranks a movie or book in front of you, or a blog or a short form video or long form video, just about everything that is put in front of you came out of the recommender system.

Every single company has it, we recently just did a recommender system for an investment banker who had a product service that improved its results significantly and they were so pleased by it for the products, for the services that they offer. And so in the future almost every website vibrant recommender systems and we're making it, so that it's on the one hand much, much easier to scale using SDK, we call Merlin. On the other hand, making it a lot easier for people to engage it. And so I chose the recommender system because it represents today's AI factory if you will, almost every one of the large data centers that are running 24x7 are constantly recommending something and constantly collecting new data to go refine our recommender. And so I chose probably the one application that is simultaneously very large scale. I'm also very different in computing profile than just about everything that we've done so far in AI.

And so I used that as a way to feature the uniqueness of Grace Hopper. But there is a whole bunch of other examples like that, Spark which is the leading data analytics engine in the world, used by probably 80% of the worlds enterprise. And the most popular data analytics platform in the cloud is going to -- Grace Hopper is going to be ideal for that. You're going to use it for very large data scientific computing. And I just I selected -- in fact, I just gave you a few examples of probably the most valuable enterprise applications. It's going to be really great for cloud computing, because the energy density is so high. So you could put a lot more CPUs in a rack because the energy efficiency is so great. And so there are lot of different use cases, I just chose a few just to highlight what makes Grace so special. Even in that one example, Grace Hopper is probably seven times the performance of Hopper, which is the most powerful computer in the world today. And so the fact that, we can create an architecture that is so much such a big leap forward for a particular application, I thought that was worthy to highlight.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Stacy Rasgon with Bernstein. Please go ahead.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi guys, thanks for taking my question. Can you hear me?

A - Simona Jankowski {BIO 7131672 <GO>}

Yeah.

A - Jensen Huang {BIO 1782546 <GO>}

Yes, Stacy. That's good.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Great. Thank you. So I wanted to ask about how you view supply and availability of Ada at launch, given what happened during the Ampere lunch? And more importantly, how are you gauging demand for Ada in an environment like this, where obviously the channels flushing out people were worried about GPUs formerly used in crypto mining getting dumped on the market. I mean, just given all the noise, how

do you gauge demand for that product? So I guess supply and demand is what I mean asking.

A - Jensen Huang {BIO 1782546 <GO>}

Well, the best work we delayed, I think it's fairly broadly known that we delayed the launch of Ada, to give the channel an opportunity to clear the 3080 to 3090s, the 3080 Ti the 3090 Ti, but basically the high-end segments. And gave us an opportunity to work with our partners to put marketing programs in place to move the pricing into a segment that even as Ada comes would still be really good value for anybody who bought it. And so we are -- we prepared ourselves into two quarters of pretty harsh medicine in order to prepare for Ada. Ada is going to come into a segment that is going to be well above anything, that is going to be affected by crypto or any of the reactions that you were referring to. Stacy, I think it's -- if you look at where 4080 is going to come out at, it targets a very large segment of our GPUs because that's where the enthusiast are and that's where the core gamers are and they tend not to be affected by market conditions here or there. And so I expect Ada's launch to be very successful. I believe that we've prepared our cells in our channel to welcome Ada with open arms and we stay clear of and just follow all of the dynamics that you mentioned. So we should have a great launch.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Matt Ramsay with Cowen. Please go ahead.

Q - Matthew Ramsay {BIO 17978411 <GO>}

Thank you very much. Good morning team. Can you guys hear me okay?

A - Jensen Huang {BIO 1782546 <GO>}

Yes, Matt.

Q - Matthew Ramsay {BIO 17978411 <GO>}

Awesome. Thanks for letting me ask the question and thanks for all the information today. Jensen, I think a lot of us are still trying to get our heads around some of the restrictions that came in with China. And the intentions of the governments and also the ramifications, secondary and tertiary of what's been announced. So my question is around some of the things that you guys announced and today on the software side, whether it's GeForce NOW in gaming or drives in the auto business and obviously the Omniverse stuff across enterprise?

How do you see the environment for those software opportunity is in China, those -- some restriction on your hardware being shipped into China. Accelerate those opportunities for you in software for Chinese customers where you can host the compute and the data yourselves. Do you fear that there might be restrictions on your software business in China? I honestly don't have a clue as to how to do sort of a risk award analysis of that for your software businesses. So, any thoughts there would be really appreciated. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Sure. The restrictions are very specific. Restrictions are very specific. And it's a -- it requires license for a specific level of compute, combined with a specific level of inter-chip connection bandwidth. And within the restrictions we will offer our customers and they will have plenty of choices of alternative products that are within the envelope that requires, that are not restricted. And if a restriction -- but if a customer requires that very specific product, we will seek a license. And so I think the U.S. government would like to know, who in the world are using products of that nature. But for most of our customers alternative products are going to be just fine. And so we're working hard and working fast to offer our customers alternative products. And my expectation is that, outside of proper execution, which the team is working really hard on, we should be able to offer and our customers would accept alternative products, that are excellent.

Regarding software, it has no incremental relationship with the restriction. And the reason for that is this, you've heard me say before, that accelerated computing is at full stack. You can't just go from a compiler -- C++ compiler or C compiler, in compiler software that runs really nicely on a GPU, that's just not the way it works. And the reason for that is, because the CPU was designed to be compliable, it was designed to be forgiving of bad code. And it was designed in a single threaded way so that software is easy to write. GPUs are designed in a multi-threaded way and instead of one or two or four threads of execution, inside our GPU there could be 15,000, 25,000 threads of execution. In the 25,000 things running around, 20,000 things running around it's hard for any human to keep track of. And so we created a programming language, a programming model and a whole bunch of libraries on top and a whole bunch of algorithms and runtime engine that sits on top of that. And we call those platforms. And those platforms whenever I announced new software, I will tell you about when you deduct the new software on specific purpose, out of the 300 SDKs the question is why did I select the ones that I selected. And the reason for that mostly is one of several vectors.

One, there is a new application field that is very important and then I think you should know about, our developer should know about, out scientists should know about, our ecosystem should know about. Like for example, for the very first time we were able to take the most valuable, probably the most recent modern valuable database called the graph database. And we formulated, sample it into deep learning and use deep learning to learn predictive patterns out of the graph, we called that cuGraph. And I spoke about it last time, I spoke around this time and want to speak about next time because it's just the radical importance of it in so many different industries.

I might have spoken about operational research and the work that we do with cuOpt, the work that we're doing with large language models and the work that we're doing with quantum computing. All of these things expand our market. And so that's the second reason why I tell you is, because I highlight these SDKs to help you understand how accelerated computing is going into new markets. Markets that are well beyond deep learning maybe or enabled by deep learning in the case of graphs. And so those libraries I spoke about to highlight its importance, new use

case, new market or new breakthroughs. And it's -- so in a lot of ways the best way to look at NVIDIA's business growth is to look at our SDKs. And the reason for that is because accelerated computing is a full stack thing. And so if there is new software there is new consumers, new demand for our hardware. And so it's always software before hardware, but not the other way around.

Operator

Our next question will come from Mark Lipacis with Jefferies.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi. Thanks for taking the question and thanks for the great presentation. Jensen, you used the expression of full stack computing a lot and I think there is a lot that goes behind that description of what you guys are delivering. Can you talk about, how you've transformed NVIDIA, like what does that mean internally for the company that you've kind of become a full stack computing company?

And can you compare what you are doing today in this accelerating computing era, to what was the model before? Or is it fair to say that, what you are to accelerating computing like Microsoft and Intel and PCs, is that a fair analogy or is that underselling what you guys are delivering? And if you could kind of put it in perspective, what you guys are really delivering out there relative to previous on computing areas? I think that would be interesting. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Mark, that's a great question. The PC industry was actually the only unique industry and it created the horizontal business model. And notice cloud computing is not horizontal cloud computing, is in fact vertical. You have the cloud service providers who are platforms as they have SaaS, they have PaaS, they have database as a service to sort of DBaaS and then they have IaaS, you know. And so there is a lot of as a service. And you could think of accelerated computing in basically the same way. That in order to be successful in accelerated computing, it's not so much that we're doing what other people used to do and we aggregated into one vertically integrated company. It's just that in this vertically in this field of computing, if you're not vertically integrated, you're not going to be successful. Nobody is going to write your operating system. Nobody is going to write your runtime engine. Nobody is going to develop your -- even your distributed operating system, whether it's in the cloud or supercomputing or enterprise, nobody is going to write it for you. And so you really have to go do it yourself. You have no choice but to be a storage company and a networking company because the storage and networking and cybersecurity in the world of multi-tenant public clouds. And where it interacts with our accelerated computing which is data center scale.

Your network, your storage, your cybersecurity is really part of your computing fabric. We have no choice. And so it's not about building the chip. None of my customers buy the chip. They need -- ultimately they might place a PO on a whole bunch of chips, but what they're really buying is the NVIDIA computing stack. And which explains the reason why we're so broadly used in so many different clouds

and computer makers and so on and so forth, that the stack just -- wherever you happen to be using it, it does what we promise, that was going to do. It delivers speed of well beyond reason and well beyond the cost of adding it by many orders and many factors. So anyways it's a full stack computing problem.

We talk about basically four stacks, Mark. There is a -- the first stack NVIDIA put together was our graphics stack and today it's called NVIDIA RTX. And computer graphics is now a full stack problem. If you don't design and develop the artificial intelligence or the physics engine or the ray tracing engine and all the software. And then you just -- nobody is going to do it for you. It's now part of OpenGL, it doesn't exist. And so if you don't build that full stack, people can't use your platform. The second is NVIDIA HPC, that's our scientific computing stack, this is where quantum chemistry, molecular dynamics, so on and so forth, fluid dynamics and so on so forth.

And then the third is NVIDIA AI, which has all of our end-to-end run times and engines. It's NVIDIA AI is essentially the operating system of modern artificial intelligence. And it goes -- it starts from data ingestion, data processing with RAPIDS, into deep learning into now with a cuGraph or graph analytics and graph learning systems, all the way to inference train [ph]. And so that end-to-end platform is part of NVIDIA AI. And if you're doing machine learning or artificial intelligence of any type of model, anywhere you could use NVIDIA AI. And then the last one is NVIDIA Omniverse. And so NVIDIA Omniverse is the next wave of AI, where the artificial intelligence has to interact with the physical world. And you need a way of providing ground truth for the artificial intelligence and Omniverse is designed for that. And so we gave these four platform names but they all include operating systems, networking, storage, data center scale stuff, all the libraries, all the run times and that they are represented as four platforms.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Harlan Sur with J.P. Morgan. Please go ahead.

Q - Harlan Sur {BIO 6539622 <GO>}

Yeah. Can you guys hear me?

A - Jensen Huang {BIO 1782546 <GO>}

Yes, Harlan.

Q - Harlan Sur {BIO 6539622 <GO>}

Yeah. Thanks. Good morning. (inaudible) and thank you for hosting this call. Jensen, on these new managed services offerings that you unveiled and will be unveiling going for NeMo, BioNeMo, LLM, you guys are further accelerating customers' time to market with large language model training, $2 million in deployment. What's the monetization model here, is that subscription based or consumption based? And then I also noticed that the L40 announcement going into your OVX platform, when should we expect the launch of the rest of your L Series ProViz line of GPUs solutions?

A - Jensen Huang {BIO 1782546 <GO>}

I'll take the second one first. We announced RTX 6000 and it will complement Ampere in the workstation lineup. Our typical rhythm is desktop first, a couple of quarters later maybe less than that this time the notebook and all the thin and lights and all of those things come shortly after. So you could expect that, that basic rhythm. In the case of L40, your question was, what about the other versions? Well, I wouldn't want to ruin the surprise, but L40 was really quite the perfect one to launch this time because customers are really clamoring for the Omniverse computers. And we've got -- we're working with all the OEMs in the world and getting (inaudible) shipped out to the enterprises.

Our business model for the NVIDIA managed cloud services is going to be by consumption. So for example, in the case of NeMo, large language model, we've made it so that -- it's much, much easier for you to learn the prompts associated with adapting that language model for your own application. And so everything has been stood up and it will likely be a model that's very similar to a cloud service providers per GPU hour, except it would be value-added per GPU hour. And net of the value-added price, you should still see much, much lower cost and at the very minimum much more effective and a much easier way to train your model. And let's see, I think that's spot right, Omniverse is going to be a consumption model as well and so is the training, that large language model training.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Raji Gill with Needham. Please go ahead.

Q - Rajvindra Gill {BIO 16383656 <GO>}

Yes. Thank you for taking my questions and thank you for this keynote. Well, just a quick question on the near term if I can. I appreciate the fact that you are under shipping a gaming demand by quite a large amount in order to clear out excess inventory, the channel to prepare for your next generation gaming architecture. And so that is and then you also indicated that you have about $5 billion kind of run rate, if you combine the two quarters. That would imply that you have pretty good visibility still into end market demand. And I'm just wondering about the overall end market demand, if roughly two-thirds of the gaming market is tied to China and Europe, if I have those statistics correct. Both economies are weakening fairly significantly. I don't know if China is pushing a stimulus plan or not. But -- so the question really is about visibility into to that end market demand, that level of confidence. I appreciate that you're taking the large inventory correction, I think that makes a lot of sense, but just any insight there would be super helpful. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Did you want to take that? I'm happy to either way.

A - Colette Kress {BIO 18297352 <GO>}

No, I'll start off, that says, (inaudible) quite accurate, that we've got a good amount of business around the world in each of the regions. You can almost divide each of the

regions, one-third, one-third, one-third of what we're seeing. Each of them are dealing with different macro conditions right now. But underlying the ability to game, the form of gaming is under attainment is still driving solid sell-through in terms of our products. So this underselling that we are doing is a bright time before we produce Ada going forward. But we watch this carefully, we are still seeing the solid demand. Now, how do we see that, do things change going forward? Possibly, but right now it looks solid.

I move to Jensen, to see if you have anything more to add.

A - Jensen Huang {BIO 1782546 <GO>}

I thought that was perfect.

A - Simona Jankowski {BIO 7131672 <GO>}

Okay. I think we have time for one last question and that will come from Joe Moore with Morgan Stanley. Please go ahead.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you and thanks for the presentation. I wonder if you could talk about pricing in the Gaming business. I noticed the 4080 comes out at a higher price, materially higher than the 3080. And I know the suggested retail price in 3080 was never really achieved, it was always about that. But can you just talk generally about how we should think about, how cost inflation is affecting you guys through the full stack, and to the extent that I know you provide more value each generation and the prices have generally been drifting up. But is this cycle is going to be fairly normal or is the general inflation going to drive the prices higher than normal?

A - Jensen Huang {BIO 1782546 <GO>}

Well, first of all, the value proposition of Ada is after charts. We have never had a generational leap this great. And the new architecture is -- we fundamentally advanced all three processors and the breakthroughs in neural rendering is just off the charts. It's really unlike anything that's ever been possible before. And so the value is incredible. The adoption of our platform across the worlds game developers is incredible. Ada is going to come out, it's going to be incredible, so the value is incredible. The pricing is higher than last generation, but I would say comparable. And the gross margins to us is comparable to the last generation and so that kind of frames it. But I think the most important thing is, is at each price point the value that gamers going to get from there, there our new Ada graphics card is just going to be off the charts.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you, Jensen. I think that's all the time we have for questions today. Are there any closing comments you'd like to make before we get off the air.

A - Jensen Huang {BIO 1782546 <GO>}

Well, let me do this. We covered a lot of ground and when somebody comes to see NVIDIA keynote, unlike a chip keynote of speeds and feeds and teraFLOPS and gigaflops, we have plenty of those as well. But you get to hear about a whole bunch of new applications and new industries and particularly new software stack. And so let me just frame very simply the four things we talked about, it was really in four categories. The first is just a whole bunch of new chips, were in a new product cycle, were kicked off multiple product cycles at the same time. Ada is for gaming, it's for workstation and very new it's for the very first time it's also for Omniverse.

Our Hopper, our Transformer engine is a giant leap forward, it's in full production. Orin unlike Xavier, like Xavier before that, which was really only for AVs, Orin has been just a home run with respect to all of the customer adoption. But it's also for robotics, it's also for industrial edge and it's also for our medical instruments. All of these are going to be robotic systems. And then next up, I really appreciate the question with Grace CPU and Thor, they're next up, ready to go, and so for the next round. And so the first is just a whole bunch of new chips and new hardware that's associated with it.

The second is to recognize that accelerated computing is a full stack and the data center scale computing approach. Moore's Law has slowed, and to be even more blunt, that's really stocked in every measurable way. And it is now broadly accepted that, accelerated computing is the path forward and NVIDIA recognizing that, this is a full accelerated, full stack challenge has put together four platforms. I mentioned RTX, I mentioned scientific computing, I mentioned AI and Omniverse. And during this GTC keynote we spoke about three of them. RTX showcased newer rendering and DLSS 3.

Our NVIDIA AI spoke about the SDKs, it's much, much more than deep learning and even deep learning has taken a giant leap forward with large language models and recommender systems. And we highlight of RAPIDS for data analytics, Spark for data analytics, Jacks [ph] the work that we're doing with DeepMind and Google Brain. The next major framework cuGraph or graph analytics and train, which is used by 35,000 companies around the world, is about to rollout for large language models, which allows you to do train on distributed computing. It's a technological marvel.

And just as AI needed ground truth to learn from, you ultimately these -- even if they're unsupervised they have to learn from some form of ground truth. The next wave of AI where AI meets the physical world needs, its ground truth and its ground truth is impossible to collect from the physical world. And so we have to generate ground truth from the virtual world, we call that Omniverse. Omniverse will be as vital to the future of artificial intelligence, as the first generation TensorFlow and PyTorch, just were to the first generation of artificial intelligence. And this is where the digital world meets the physical world and the next wave of AI happens.

We announced 150 connectors which opens up Omniverse to all of these industries because they would like to find a way to automate their industry. And one of the largest ecosystems is called Siemens JT and we're delighted to have a connector for that. But we have 150 others. We extended NVIDIA AI and NVIDIA Omniverse into

the cloud and we had just spoken about that, so that we could make it easier for some -- with some of these really complicated workloads, that are really in the domains of some of the largest companies, we made it easier, so that we can democratize these -- some of the areas of artificial intelligence, so that every researcher, every scientist could take advantage of it. And by doing, we did that by extending it into the cloud and we'll have with Omniverse will have the Nucleus, if you will Omniverse super cloud. And it's going to be a database in the cloud, a brand new type of database.

And just as relational databases were revolution along several decades ago and graph databases were resolution about a decade ago. And they're all these different types and now we hope that Omniverse is going to, in the cloud will enable just about everybody, every designer, every industry, to be able to connect to it, work among each other. And so that was the third announcement, the third category. And then finally, this is the area and I've been speaking about this for some time. This is really the era of enterprises taking advantage and applying AI to revolutionize their products and services, but also to revolutionize themselves to bring automation into their companies.

You -- all of you tracking IT for a long time, know that a long time ago there used to be one IT department in most companies, but now there are four IT departments and these IT departments all have different reasons, that they do what they do. There is sales marketing IT department as we know, that emerge kind of about a decade ago. But the two new ones has to do with artificial intelligence and deploying artificial intelligence workloads. One has to do with MLOps and the creation of AI, we say that you need two computers to deploy AI, one to develop the AI and simulate the AI, and then you need one to deploy, and the word that's used is inference the AI, apply the AI into your products and services. These two organizations both have IT departments and the way that you apply this new form of software is rather arcane and it is not extremely easy. For all of us I have been doing for a long time, this is the new way we do software. But it's time for the world's enterprise to be able to take advantage of this capability. And there are so many enterprises.

And so this time at this GTC I'm so delighted to announce the Deloitte, the world's largest professional services firm with their 350,000 professionals are going to build practice on NVIDIA AI and NVIDIA Omniverse, so that we can together take these platforms NVIDIA AI, NVIDIA Omniverse to the world's enterprises and help them create new price and services and help them revolutionize themselves. So four major categories. New chips, whole bunch of new software, new services and new partners.

And so thank you all for joining GTC, this was surely a news packed GTC. So I appreciate your attention and I look forward to seeing you soon.

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.