Q3 2022 Earnings Call

Company Participants

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

Other Participants

- Aaron Rakers, Analyst
- C.J. Muse, Analyst
- Mark Lipacis, Analyst
- Stacy Rasgon, Analyst
- Timothy Arcuri, Analyst
- Vivek Arya, Analyst

Presentation

Operator

Good afternoon. My name is Cindy, and I'll be your conference operator today. At this time, I would like to welcome everyone to the NVIDIA's Third Quarter Earnings Call. All lines have been placed on mute to prevent any background noise. After the speakers' presentation, there will be a question-and-answer session. (Operator Instructions) Thank you.

Simona Jankowski, you may begin your conference.

Simona Jankowski {BIO 7131672 <GO>}

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2022.

With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer.

I would like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the fourth quarter and fiscal year 2022. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent.

During this call, 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 the disclosure in today's earnings release, our most recent Form 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, November 17, 2021, based on information currently available to us. Except as required by law, we have seen no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website.

With that, let me turn the call over to Colette.

Colette Kress {BIO 18297352 <GO>}

Thanks, Simona. Q3 was an outstanding quarter with revenue of $7.1 billion and year-on-year growth of 50%. We set records for total revenue as well as for gaming, data center and professional visualization.

Starting with gaming, revenue of $3.2 billion was up 5% sequentially and up 42% from a year earlier. Demand was strong across the board. While we continue to increase desktop GPU supply, the new channel inventories remain low. Laptop GPUs also posted strong year-on-year growth, led by increased demand for high-end RTX laptops.

NVIDIA RTX technology is driving our biggest ever refresh cycle with gamers and continues to expand our base with creators. RTX introduced ground breaking real-time ray tracing and AI enabled super resolution capabilities, which are getting adopted by the new accelerating pace. More than 200 games and applications now support NVIDIA RTX, including 125 with NVIDIA DLSS. This quarter alone, 45 new games shipped will be a lesson.

In NVIDIA Reflex latency reducing technology is in top eSports titles, including Valorant, Fortnite, Apex Legends and Overwatch. In addition, the Reflex ecosystem continues to grow with Reflex technology now integrated in almost 50 gaming peripherals.

NVIDIA Studio for creators keeps expanding. Last month at the Adobe MAX Creativity Conference, Adobe announced two powerful AI features for Adobe Lightroom and Lightroom Classics, accelerated by NVIDIA RTX GPUs. In addition, several of our partners launched new studio systems, including Microsoft, HP and Asus. We estimate that a quarter of our installed base has adopted RTX GPUs. Looking ahead, we expect continued upgrades as well as growth from NVIDIA GeForce users given rapidly expanding RTX support and the growing popularity of gaming, eSports, content creation and streaming.

Our GPUs are capable of crypto mining, but we don't have visibility into how much this impacts our overall GPU demand. In Q3, nearly all of our Ampere architecture gaming,

desktop GPU shipments were Lite Hash Rate to help steer GeForce supply to gamers. Crypto mining processing revenue was $105 million, which is included in our OEM and other.

Our cloud gaming service, GeForce NOW has two major achievements this quarter. First, Electronic Arts brought more of its hits games to December. And, second, we announced a new GeForce NOW RTX 3080 membership tool priced at less than $100 for six months. GeForce NOW membership has more than doubled in this last year to over 14 million gamers that are streaming content on 30 data centers in more than 80 countries.

Moving to pro visualization. Q3 revenue of $577 million was up 11% sequentially and up 144% from the year ago quarter. The sequential raise was led by mobile workstations with the desktop workstations also growing as enterprises deployed systems to support hybrid work environments. Building on the strong initial ramp in Q2 and pure architecture service continued to grow leading verticals including media and entertainment, healthcare, public sector and automotive.

Last week, we announced general availability of Omniverse and Polaris 8 platform for simulating physically accurate 3D world and digital twins. Initial market reception to Omniverse has been incredible. Professionals at over 700 companies are evaluating the platform, including BMW, Ericsson, Lockheed Martin and Sony Pictures. More than 70,000 individual creators have downloaded Omniverse, since the open beta launch in December. There are approximately 40 million 3D designers in the global market.

Moving to automotive. Q3 revenue of $135 million, declined 11% sequentially and increased 8% from the year ago quarter. The sequential decline was primarily driven by AI Cockpit revenue, which was negatively impacted by automotive manufacturers' supply concerns. We announced self-driving truck start-up, Kodiak Robotics, automaker Lotus, autonomous bus manufacturers QCraft and EV startup WM Motor have adopted the NVIDIA DRIVE Orin platform for their next generation vehicles. They join a large and rapidly growing list of companies adopting and developing on NVIDIA DRIVE, including auto OEMs, Tier 1 suppliers and EVs, trucking companies, mobile taxies and software startups.

Moving to data center. Record revenue of $2.9 billion, grew 24% sequentially and 55% from the year ago quarter with record revenue across both hyperscale and vertical industries. Strong growth was led by hyperscale customers fueled by continued rapid adoption of Ampere architecture and support GPUs for both internal and external workloads. Hyperscale compute revenues doubled year-on-year, driven by the scale up of natural language processing and recommender models and cloud computing.

Vertical industry growth was also strong led by consumer Internet and broader cloud providers. For example, Oracle Cloud deployed NVIDIA GPUs for its launch of AI services, such as spec analysis, speech recognition, computer vision and anomaly detection.

We continue to achieve exceptional growth in Inference, which again outpaced our overall data center growth. We have transitioned our lineup of Inference purposed processers to

the Ampere architecture, such as the A30 GPU. We also released the latest version of Triton Inference Server software, enabling compute-intensive Inference workloads, such as large language models to scale across multiple GPUs and nodes with real-time performance.

Over 25,000 companies worldwide use NVIDIA AI Inference. A great new example is Microsoft Teams, which has nearly 250 million monthly active users. It uses NVIDIA AI to converts speech to text real time during video calls in 28 languages and across different API. We've reached 3 milestones to help drive more mainstream enterprise adoption of NVIDIA AI.

First, we announced the general availability of NVIDIA AI Enterprise, our comprehensive software suite, but AI tools and frameworks that enables the hundreds of thousands of companies running NVIDIA, running vSphere to virtualize AI workloads on NVIDIA-Certified Systems. Second, VMware announced a future update to vSphere with Tanzu that is fully optimized for NVIDIA AI. When its combined with NVIDIA AI Enterprise, enterprises can efficiently manage cloud-native AI development and deployment on mainstream data center servers and clouds with existing IT experience.

And further we expanded our LaunchPad program globally with Equinix as our first digital infrastructure partnering. NVIDIA LaunchPad is now available in nine locations worldwide, providing enterprises with immediate actions to NVIDIA's software and infrastructure to help them prototype and test data science and AI workloads. LaunchPad features NVIDIA-Certified Systems and NVIDIA DGX Systems running the entire NVIDIA AI software style.

In networking, revenue impacted as demand outstrips supply. We saw the momentum towards higher speed and new generation products, including ConnectX-5 and ConnectX-6. We announced the NVIDIA Quantum-2 400-gigabit per second end-to-end networking platform. Consisting of the Quantum-2 switch the ConnectX-7 network adaptor and the BlueField-3 DPU. The NVIDIA Quantum-2 switch is available from a wide range of leading infrastructure and system vendors around the world.

Earlier this week, the latest top 500 list of supercomputers showed continued momentum for our full stack computing approach, NVIDIA technologies accelerated over 70% of the systems on the list, including over 90% of all new systems and 23 of the top 25 most energy-efficient systems.

Turning to GTC. Last week we posted our GPU Technology Conference, which had over 270,000 registered attendees. Jensen's keynote has been viewed 25 million times over the past eight days. While our Spring GTC focused on new chips and system, this edition focused on softwares demonstrating our full continued stuff.

Let me cover some of the highlights. Our vision for Omniverse came to life at GTC. We significantly expanded its ecosystem and announced new capabilities. Omniverse Replicator is an engine for producing data should train robots, replicating augment real world data with massive diverse and physically accurate synthetic datasets to both accelerate development of high quality, high performance AI across computing domains.

NVIDIA Omniverse Avatar is a platform for generating interactive AI avatars. It connects several core NVIDIA SDKs including speech AI, computer vision, natural language understanding, recommendation engines and simulation, applications including automated customer service, virtual collaboration and content creation. Replicator and Avatar joined several other announced features and capabilities for Omniverse, including AI, AR, BR and simulation-based technologies.

We introduced 65 new and updated software development sets bringing our total to more than 150 serving industries from gaming and design to AI, cyber security, 5G and robotics. One of the SDKs is our first core licensed AI model, NVIDIA Riva, for building conversational AI applications. Companies using Riva during the open beta includes RingCentral for video conference like live capturing and AM [ph] for customer service chatbots. NVIDIA Riva Enterprise will be commercially available early next year for launch mode.

We introduced the NVIDIA NeMo Megatron framework optimized for training large language models on NVIDIA DGX SuperPOD infrastructure. This combination brings together production-ready, enterprise grade hardware and software to help vertical industries develop language and industry specific chatbots, personal assistance, content generation and summarization. Early adopters include Citi, JD.com and (inaudible).

We unveiled BlueField DOCA 1.2, the latest version of our DPU programming label with new cybersecurity capabilities. DOCA is to our DPUs as CUDA is to our GPUs. It enables developers to build applications and services on top of our BlueField DPUs. Our new capabilities make BlueField the ideal platform for the industry to build their own Euro Trust security platforms. The leading cybersecurity companies are working with us to provision their next generation firewall service on BlueField, including Check Point, Juniper, Fortinet, F5, Palo Alto Networks and VMware.

And we released Clara Holoscan, an Edge AI computing platform for medical instruments to improve decision-making tools in areas such as robot-assisted surgery, interventional radiology and radiation therapy planning. Other new or expanded SDK libraries unveiled at GTC include ReOpt for AI optimized logistics; Quantum for quantum computing; Morpheus for cybersecurity; Modulus for physical face machine learning; and cuNumeric, a data center scale mass library integrating to bring accelerated computing through the large and growing Python ecosystem. All in, NVIDIA's computing platform continues to expand as a broadening set of SDK enable more and more GPU accelerated application and industry uses.

CUDA has been downloaded 30 million times and our developer ecosystem is now nearing (inaudible). The applications they develop on top of our SDKs and the cloud edge computing platform are helping to transform multi-trillion dollar industries from healthcare to transportation to professional [ph] services, manufacturing, logistics and retail.

In automotive, we announced NVIDIA DRIVE Concierge and DRIVE Chauffeur, AI software platforms that enhance vehicles performance features and safety. DRIVE Concierge built

on Omniverse Avatar functions as an AI-based in-vehicle personal assistance, but enables automatic parking, summoning capabilities. It also enhances safety by monitoring the driver throughout the duration of the trip.

DRIVE Chauffeur offers autonomous capability, reminding the driver of constantly having to control the car. It will also perform address-to-address driving when combined with DRIVE Concierge AI [ph] platform.

For robotics, we announced Jetson AGX Orin, the world's smallest most powerful and energy-efficient AI supercomputers for robotics, autonomous machine and embedded computing at the Edge. Built on our Ampere architecture, Jetson AGX forum provides 6x processing of its predecessor and delivers 200 trillion operations per second, similar to a GPU enabled server that fits into the palm of your hand. Jetson AGX Orin will be available in the first quarter of calendar 2022.

Finally, we revealed plans to build Earth-2, the world's most powerful AI supercomputer dedicated to confronting climate change. The system would be the climate change counterpart to Cambridge-1, the UK's most powerful AI supercomputer that we built for corporate research. Earth-2 furnishes all the technologies we've invented up to this moment.

Let me discuss Arm. I'll provide you a brief update on our proposed acquisition of Arm. Arm with NVIDIA is a great opportunity for the industry and customers with NVIDIA's scale, capabilities and robust understanding of data center computing, acceleration and AI. We assessed Arm in expanding their reach into data center, IOT and PCs and advanced Arm's IP for decades to come. The combination of our companies can enhance competition in the industry as we work together on further building the world of AI.

Regulators at the US, FTC, have expressed concerns regarding the transaction and we are engaged in discussions with them regarding remedies to address those concerns. The transaction has been under review by China Antitrust Authority, pending the formal case initiation. Regulators in the UK and the EU have declined to approve the transaction in Phase 1 of their reviews on competition concerns. In the UK, they have also voiced national security concerns. We have begun the Phase 2 process in the EU and UK jurisdictions.

Despite these concerns and those raised by some Arm licensees, we continue to believe in the merits and the benefits of the acquisition to Arm, to its licensees and to the industry. We believe these concerns and those raised by some Arm licensees -- we continue to believe in the merits and benefits of the ongoing acquisition.

Moving to the rest of the P&L. GAAP gross margin for the third quarter was up 260 basis points from a year earlier, primarily due to higher end mix within desktop, notebook, GeForce GPUs. The year-on-year increase also benefited from a reduced impact of acquisition related costs. GAAP gross margin was up 40 basis points sequentially, driven by growth in our data center Ampere architecture products, which is particularly offset by mix in gaming. Non-gaming gross margin was up 150 basis points from a year earlier and up 30 basis points sequentially.

Q3 GAAP EPS is $0.97, 83% from a year earlier. Non-GAAP EPS was $1.17, up 60% from a year ago adjusting for our stock split.

Q3 cash flow from operations was $1.5 billion, up from $1.3 billion a year earlier and down from $2.7 billion in the prior quarter. The year-on-year increase primarily reflects higher operating income, particularly offset by prepayment for long-term supply agreement.

Let me turn to the outlook for the fourth quarter of fiscal 2022. We expect sequential growth to be driven by data center and gaming, more than offsetting a decline in CMP. Revenue is expected to be $7.4 billion plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 65.3% and 67%, respectively, plus or minus 50 basis points.

GAAP and non-GAAP operating expenses are expected to be approximately $2.02 billion and $1.43 billion, respectively. GAAP and non-GAAP other income and expenses are both expected to be an expense of approximately $60 million, excluding gains and losses on non-affiliated investments. GAAP and non-GAAP tax rates are both expected to be 11%, plus or minus 1% excluding discrete items. Capital expenditures are expected to be approximately $250 million to $275 million. Further financial details are included in the CFO commentary. Other information is also available on our IR website.

In closing, let me highlight upcoming events for the financial community. We will be attending the Credit Suisse 25th Annual Technology Conference in person on November 30th. We will also be at the Wells Fargo Fifth Annual TMT Summit virtually on December 1st, the UBS Global TMT Virtual Conference on December 6th, and the Deutsche Bank Virtual Auto Tech Conference on December 9th. Our earnings call to discuss our fourth quarter and fiscal year 2022 results is scheduled for Wednesday, February 16.

With that, we will now open the call for questions. Operator, will you please poll for these questions.

Questions And Answers

Operator

Yes. (Operator Instructions) For our first question, we have Aaron Rakers from Wells Fargo. Aaron, your line is open.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thanks for taking the question and congratulations on the results. I guess, I wanted to ask about Omniverse. Obviously, a lot of excitement around that. I guess the simple question is, Jensen, how do you define success in Omniverse as we look out over the next, let's call it, 12 months and how do we think about the subscription license opportunity for Omniverse. I know you've talked about $40 million total 3D designers, I think that actually doubled what you talked about back in August. So I'm just curious of how we at finance line should probably think about that opportunity materializing?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks. Omniverse success will be defined by, number one, developer engagement, connecting with developers around the world; two, applications being developed by enterprises; three, the connection -- designers and creators among themselves. Those are the nearest term I -- and I would say that in my type of definition, et cetera.

Near term also, it should be revenues and Omniverse has real immediate applications as I demonstrated at the keynote and I'll highlight a few of them right now. One of them, of course, is that, it serves as a way to connect 3D and digital design world. Think of Adobe as a world, think of the Autodesk as a world, think of Revit as a world. These are design world in the sense that people are doing things in it, they are creating things in it and they have to run day to day. We made it possible for these worlds to be connected for the very first time and for it to be shared like in cloud documents.

That's not been possible ever before and we can now share work with each other, you can see each other's work, you can collaborate and so in the world of remote working, Ominverse's collaboration capability is going to be really appreciated and that should happen right away. We would like to see that happen in very near term. And that drives of course more PC sales, more GPU sales, more workstation sales, more servers sales.

The second use case is digital twins. And we show in these following examples of how several companies using Omniverse to create a digital twin of a city so that they could optimize radio placements and radio energy used for beamforming. You saw BMW using it for their factories. You're going to see people using it for warehouse, logistics warehouse to plan and to optimize their warehouses and deploying the robots. And so digital twin applications are absolutely immediate.

And then remember robots has several clients. There is the physical robot that you saw and a physical robot would be a self-driving cars and physical robots would be the car itself turning it into a robot. So that it could be an intelligent assistant. But I demonstrated probably the -- in my explanation, the largest application of robots in the future and it's Avatars.

We built Omniverse Avatars to make it easy for people to integrate some amazing technology for computer vision, for speech recognition, natural language understanding, gesture recognition, facial animation and speech synthesis, recommender systems, all of that integrated into one system and running a real time. That Avatar system is essentially a robotic system and the way that you use that is, for example, with $25 million or so retail stores, restaurant, places like airports and train stations, office buildings and such, where you're going to have intelligent Avatars doing a lot of assistance. They might be doing check out, they might be doing check in, they might be doing customer support and all of that can be done with Avatars, as I've demonstrated.

So the virtual robotics application, digital buys of Avatars, it is going to be likely the largest robotics opportunity. So if you look at our licensing model, the way it basically works is that inside Omniverse is one of the main users and the main users could be one of the 20 million creators or 20 million designers and the 40 million creators and designers around

the world and they share Omniverse, each one of the main users would be a $1,000 user per year.

But don't forget that intelligent use or intelligent users that have been connected through Omniverse will likely be much larger as digital buyers than humans. So I mentioned 40 million, but are 100 million cars. In 100 million cars we'll all have -- we will have the capability to have something like in Omniverse Avatar and so those 100 million cars could be $1,000 per car per year. And in the case of the 25 million or so places where you would have a digital avatar as customer support or check out smart retail or smart warehouses or smart whatever it is, those avatars are also would each individually be a new account and so they would be $1,000 per Avatar per year. And so those are the immediate tangible opportunities for us and I demonstrate the applications in related keynotes.

And then of course behind all of that, call it a couple of hundred million digital agents, intelligent agents, some of them humans, some of them robots, some of them Avatars adds $1,000 per agent per year. Behind it are, NVIDIA GPUs and DPUs, NVIDIA GPU and the cloud, and NVIDIA GPUs and Omniverse servers and my guess would be that the hardware part of it is probably going to be about half and then the licensing part of it is probably about half of the time. So this is really going to be one of the largest graphics opportunities that we've ever seen. And the reason why it's taken so long for us to manifest is because it requires three fundamental technologies to come together, I guess four fundamentals technologies to come together.

First of all, it's video graphic, second is physics simulation, because we're talking about things in world that has to be believable. So it has to obey the laws of physics. And then third is artificial intelligence as I demonstrated and illustrated just now. And all of it runs on top of an Omniverse computer that has to do not just AI, not just physics, not just computer graphics, but all of it.

And so what long term people -- why people are so excited about it is, at the highest level what it basically means is that, that long-term when we engage in that, which is largely 2D today, long-term every query would be 3D and instead of just acquiring information we would query and interact with people in Avatars and claims in places and all of these things are in 3D. So hopefully one of these days that we will probably realize it as fast as we can every transaction that goes on to internet touches a GPU and today that's a very small percentage, but hopefully one of these days it will be a bit of a high percentage. So I hope that's helpful.

Operator

For our next question, we have Mark Lipacis from Jefferies. Mark, your line is open.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi. Thanks for taking my question. Jensen, it seems like every year there seems to be a new set of demand drivers for your accelerated platform, accelerated processing ecosystem, there's gaming, then neural network and AI and then blockchain and then ray tracing and five or six years ago you you guys showed a bunch of virtual reality demos, which were really exciting at your Analyst Day, excitement died down, now it seems to be

resurfacing particularly with Omniverse Avatar capability and Facebook shedding light on the opportunities. So the two questions from that are, how close is your Omniverse Avatar to morphing into like a mass market technology that everybody uses daily? You talk about like -- you said that everybody is going to be a gamers, everybody is going to be a Omniverse Avatar user. And maybe the bigger picture is, is it reasonable to think about new killer app coming out every year? Is there a parallel that we should think about with previous computing markets that we could think about for the computing area that we're entering right now? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. I really appreciate that. Chips are enablers, but chips don't create markets, software creates market. At this point, I explain over the years that accelerated computing is very different than general purpose computing and the reason for that is because you can't just write into compiler and compile Quantum business into a chip and it doesn't, you can't just compile Schrodinger's equation and have it distribute it across multiple GPUs, multiple nodes and have a new SaaS. You just -- you can't do that for computer graphics, you can't do that for artificial intelligence, you can't do that for robotics, you can't do that for the most of the interesting applications in the world and because we really run out of steam with GPUs and that people are saying that not because it's not true, it is abundantly clear that the amount of instruction (inaudible) that you can squeeze out of system is although not zero is incredibly hard, it's just incredibly hard.

And there is another approach and we have been advocating it's already computing for some time and now people really see the benefit of it, but it does quite a lot of work and yet the work basically says for every domain, for every application we have -- for every application in large domain that you have to have a whole stack. And so whenever you want to open a new market by accelerating those applications or that domain of applications, you have to come up with a new stack and the new stack is hard, because you have to un-form the application, you have to un-form the algorithms, the mathematics, you have to un-form computer science to distribute it across, to take from something that was single threaded and make it multi-threaded and make something that we've done sequentially, make it processing parallel. You break everything, you break storage, you break networking, you break everything.

And so it takes a fair amount of expertise and that's why we're saying that over the years, over the course of 30 years we have become a full-stack company, because we've been trying to solve this problem practically through decades. And so that's one. But the benefit once you have the ability, then you can open new markets and we played a really large role in democratizing artificial intelligence and making it possible for anybody to be able to do it.

Our greatest contribution is I hope when it's all said and done that we democratized scientific computing. So that researchers and scientists, computer scientists, data scientists, scientists of all kinds were able to get access to this incredibly powerful tool that we call computers to do advance research. And so every single year we're coming up with new stacks and we got a whole bunch of stacks we are working on and many of them are working on in plenty of sites, so you see it coming, you just have to connect it together.

One of the areas that we spoke about this time, of course, was Omniverse and you saw the pieces of it inbuilt in over time and it took half a decade to start building Omniverse, but it built on a quarter century of work. In the case of the Omniverse Avatar, you could literally point to MERLIN, the recommender; Megatron, the language -- large language model; Riva, the speech AI, all of our computer vision AI that have been demonstrating over the years, natural speech synthesis that we see every single year with I AM AI the opening credit, how we're using, developing an AI to be able to speak in the human way so that people feel more comfortable and more engaged with the AI.

Face, eye tracking, Maxine and all of these technologies are connected together. They were all built in pieces, but we integrated it, we have the intentions of integrating it and to create what it's called Omniverse Avatar. And now you asked the question how quickly will we deploy this, I believe Omniverse Avatar will be in drive tunes and restaurants, fast food restaurants, check out with restaurants, in retail stores all over to world within less than five years and we're going to need it in all kinds of different applications, because there is such a great shortage of labor and there is such a wonderful way that you can now engage in Avatar and it could -- it doesn't make mistakes, it doesn't get tired and it's always on and we made so that it's cloud native and so when you saw the keynote I hope you'd agree that the interaction is continuous and the conversational forum is so enjoyable.

And so anyway I think what you highlight is, one, accelerated computing is a full-stack challenge. Two, it takes software to open new markets. Chips can't open new markets. If you build another chip, you can steal somebody's share, but you can't open new market and it takes software to open new market. NVIDIA switch with software and that's one of the reasons why we could integrate such large market opportunities.

And then last with respect to Omniverse, I believe it's a near-term opportunity that we've be working on for some three, four, five years.

Operator

For our next question, we have C.J. Muse from Evercore ISI. C.J., your lines is open.

Q - C.J. Muse

Yeah. Good afternoon. Thank you for taking the question. And I guess not an Omniverse question, but I guess -- Jensen, I'd like your commitment that you will not use Omniverse to target the sell-side research industry.

As my real question, can you speak to your data center visibility into 2022 and beyond? And within this outlook, can you talk to traditional cloud versus industry verticals and then perhaps emerging opportunities like Omniverse and others? Would love to get a sense of kind of what you're seeing today. And then as part of that how you're planning to secure foundry and other supply to support that growth? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Thank you, C.J. First of all, we have secured guaranteed supply, very large amount of it, quite a spectacular amount of it from the world's leading foundry in substrate and packaging and testing certain companies, the integral part of our supply chain. And so we have done that and feel very good about our supply situation, particularly starting in the second half of this year and going forward. I think this whole last year was a wake up call for everybody to be much more mindful about not taking the supply chain for granted and we were fortunate to have such good partners, but nonetheless we've secured our future.

With respect to data center, about half of our data center business comes from the cloud and cloud service providers and the other half comes from enterprise, what we call enterprise companies and they're in all kinds of industries. And about 1% of it comes from supercomputing centers, because so 50% or so cloud, 50% or so enterprise and 1% supercomputing centers.

And we expect next year, the cloud, the cloud service providers to scale out their deep learning and their AI workloads really aggressively and we're seeing that right now. We built a really fantastic platform and -- number one. Number two, the work we've been doing with TensorRT, which has the run time that goes into the server that's called Triton is one of our best pieces of work. We're just so proud of it. And we said nearly 4 years ago, 3.5 years ago that Inference is going to be one of the great computer science storms and really prudent to do so. And the reason for that is, because sometimes it's too quick, sometimes with latency, sometimes with interactivity on the type of models with Inference. It's just all over the map, it's not just computer vision or this recognition, it's all over the map. And the reason for those that is, it's essentially different types of architectures, totally different ways to build different applications and so the application is fabricated.

And finally there is wonderful people working. We're now on our 8th generation on that. It's adopted all over the world. Some 25,000 companies are now using NVIDIA AI and recently at GTC we announced two very, very big things. One, we remind everybody that we -- just this month before we have try to support now just in every generation of NVIDIA GPUs, of which there are so many versions to be managing without trying how would you possibly deploy AI across the entire fleet of Nvidia servers, NVIDIA GPU servers that are all over world and so it's almost an essential tool just to operate and take advantage of all of NVIDIA's GPU that are in datacenter.

Two, we support CPUs and so there is no longer necessary for someone to have two Inference servers, we can just add one Inference servers, because the NVIDIA version is already essential now everybody could just use Triton and every single server in the data center could be part of the Inference capacity and then we did something else that was really big deal at GTC, which is the so-called Forced Inference Library, called FIL, that basically the most popular machine learning was in Inference models are based on Trees and Decision Trees and boosted gradient trees and people might know it as XGBoost and achieved [ph] all the place in fraud detection, in recommender systems and utilize the companies all over the world, because it's just self-explanatory. You can build upon it, you don't worry about regressions if we build these under the Trees, and we -- this GTC we announced that we support that as well.

And so all of the sudden, all of that workflows that runs on GPU is not only do they run on Triton, it becomes accelerated. In the last -- in the next year [ph] we will announce with the tremendous interest in large language models, Triton now also supports multi-GPU and multi-node Inference. So that we could take something like an open AIG53, a NVIDIA Megatron 530B or anybody's Triton model that's been developed all over the world in all these different languages, in all these different domains, in all these different fields of science and what -- in industry where we can now influence it in real time and I demonstrated it in one of the demos, there was a co-gen [ph] that the team built and it will be helpful to basically answer questions in real-time.

And so that is just a joint venture and these are the type of workloads that's going to make it possible for us to continue to scale out increases. Back to your original question, I think next year is going to be quite a good news for gaming centers. Customers are very mindful of securing their supply for their scale out and so we have lesser amount of visibility and more visibility coming than ever at data centers, but in addition to Triton is in adoption everywhere.

And then, finally, our brand new workloads, which is built on top of AI and graphics and simulation is Omniverse and we saw the examples that I gave, these are real companies doing real work and one of the areas that has severe shortages around the world is customer support, just genuine severe shortages all over the worlds and we think the answer is Omniverse Avatar. And it runs in data centers, you could easily adapt Omniverse Avatar to do drive throughs or retail check out or customer service, and I demonstrated that with Tokyo, a parking kiosk. You can use it for an tele-operated customer service and we've demonstrated that with Maxine and we demonstrated how you could use it even for video conferencing and then we demonstrated how we can use Omniverse Avatars for robotics, for example, to create a continues work what we call DRIVE Concierge where the car is turned into intelligent customer support, intelligent agent. I think Omniverse Avatar is going to be a really exciting driver for enterprises this next year and so next year is going to be a pretty perfect year for data center.

Operator

For our next question we have Stacy Rasgon from Bernstein Research. Stacy, your line is open.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi, guys. Thanks for taking my questions. I wanted to ask two of them on data center, both near term and then maybe a little longer term. On the near-term, Colette, you suggested guidance in the Q4 be driven by data center and gaming and you mentioned data center first. Does that mean that a bigger IP could just help us like parse the contribution of features into Q4? And then in the next year, given the commentary for the last question, again it sounds like you've got like a very strong outlook for data center both from hyperscale and enterprise. If I look at sort of the implied guidance you gave, are data center for you is probably likely to grow 50% year-over-year in this fiscal year. Would it be crazy to think given all the drivers that it could grow by a similar amount next year as well. Like, how should we be thinking about that given all of the drivers that you've been laying out.

A - Colette Kress {BIO 18297352 <GO>}

Okay. Thanks, Stacy, for the question. Let's first focus in terms of our guidance for Q4. Our statements that we made were just about driven by revenue growth from data center and gaming sequentially. We can probably expect our data center to grow faster than our gaming, probably both in terms of percentage wise and in absolute dollars. We also expect our CMP product to decline quarter-on-quarter to very negligible levels in Q4. So I hope that gives you a color on Q4.

Now in terms of next year, we'll certainly turn the corner into the new fiscal year. We certainly provide guidance one quarter out. We've given you some great discussions here about the opportunities in front of us, opportunities with the hyperscales, the opportunities with the verticals, Omniverse is a full stack opportunity in front of us. We are securing supply for next year, not just for the current year and Q4 to allow us to really grow into so much of this opportunity going forward. But, at this time, we're going to wait until next year to provide guidance.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Got it. That's helpful. I appreciate it. Thank you

Operator

For the next question we have Vivek Arya from BofA Securities. Vivek, your line is open.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question. Actually I had two quick ones. And so, Colette, you suggested the inventory purchase and supply agreements are up, I think, almost 68% year-on-year, does that provide some directional correlation with how you are preparing for growth over the next 12 to 24 months? So that's one question. And then the bigger question, Jensen, that I have for you is, where are we in the AI adoption cycle? What percentage of servers are accelerated in hyperscale and vertical industry today and where can those ratios get to?

A - Colette Kress {BIO 18297352 <GO>}

Thanks for the question. So let's first start in terms of supply or supply purchase agreement. You have noted that we are discussing that we have made payments towards some of those commitments. Not only are we procuring for what we need in the quarter, what we need next year and again we are planning for growth next year, so we have been planning that supply purchases, we are also doing long-term supply purchases. These are areas of capacity agreements and/or many of our different suppliers. We made a payment within this quarter of approximately $1.6 billion out of total long-term capacity agreement of about $3.4 billion. So we still have more payments to make and we were likely continue to be purchasing longer term to support our growth that we are planning for many years to come.

A - Jensen Huang {BIO 1782546 <GO>}

Every single server will be GPU accelerated somewhere. They of all the clouds and all the enterprise, less than 10%. They can give you a sense of where you are. In terms of the workloads, it is also consistent with that in the sense that that a lot of the workloads still only run on CPUs, which is the reason why in order for us to grow, we have to be a full stack company and we have to go find applications and we have find plenty of it, focus on the application that require acceleration or benefits tremendously from acceleration that if they were to get a million X speed up, which sounds insane but it's not. Mathematically I can prove it to you and historically I can even demonstrate it to you that in many areas we have seen million X speed up and has completely revolutionized those industries, computer graphics is of course one of them.

Omniverse would not be possible with that. And so the work that we're doing with digital biology, protein synthesis, which is likely going to be one of the large industries of the world that doesn't exist today at all. Protein engineering and the protein economy is likely going to be very, very large. You can't do that unless you are able to get million X speed up in the simulation of protein biology. And so those are -- and not to mention some of the most imperative comps that we have engaged, climate science needs million X, billion X speed up and we are at a point where we can actually tackle that.

And so in each one of these cases we have performed. We have to focus our resources to go and accelerate those applications and that translates to growth. Until then they run on GPUs and look at a lot of today speech synthesis and speech recognition system, it still uses fairly traditional or mixture of traditional and deep learning approaches for speech AI. NVIDIA Riva is the world's first, I believe, that is end to end neural network and we've worked with many companies in helping them advance their sort of -- they could move their clouds to our neural-based approaches. But that's one of the reasons why we do it, so that we could provide the reference, but we can also license it to enterprises around the world, so that they could advance it for their own use cases. And so one application after another we have to get it accelerated, one domain after another we have to get it accelerated.

One of the ones that we're excited about and something that we've been working on for so long is EDA, even our own industry, Electronic Design Automation, for the very first time we announced the EDA using GPU [ph] cloud computing, whether it's because of the artificial intelligence capability, because EDA is very large combinatorial optimization program and using artificial intelligence you could really improve the design quality and design time. So we're seeing for all the major game vendors from chip design to simulation to PCB design and optimization, design synthesis, moving towards artificial intelligence and the GPU acceleration in -- is very significant and then we see that with a mechanical chat and traditional chat application. Now also jumping on to GPU acceleration is getting very significant speed ups and so I'm super excited about the work that we're doing in each one of these domains. Because every time you do it, you open up brand new market and customers that never used NVIDIA GPUs now can, because ultimately people don't buy chips, it cannot solve problems. Without a full stack, without software, you can't really commence with enabling technology that the chip and ultimately solving the customers' problems.

Operator

Your final question comes from the line of Timothy Arcuri from UBS. Timothy, your line is open.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. Colette, I had a question about gross margin. Are there any margin headwinds maybe on the wafer pricing side that we should sort of think about normalizing out, because gross margin is pretty flat between fiscal Q4 to fiscal Q -- between fiscal Q2 and fiscal Q4, but I imagine that's kind of masking a strong underlying margin growth, especially as data center has been actually driving that growth. So I'm wondering if maybe there are some underlying factors that are sort of gating gross margin? Thanks.

A - Colette Kress {BIO 18297352 <GO>}

Yeah. So we have always been working on our gross margin and being able to absorb a lot of the cost changes along the way, architecture-to-architecture really. So that's always based into our gross margin. Our gross margins right now are largely stable. Our incremental revenue, for example, what we're expecting next quarter will likely align to our current gross margin levels that we finished in terms of Q3. Our largest driver always continues to be mix. We have a lot of different mix that has driven related to the high-end AI and RTX solutions, for example, and the software that is embedded in solutions have allowed us to increase our gross margin. As we look forward long-term software is sold separately can be another driver gross margin increases in future, but cost changes, cost increases are -- generally been a part of our gross margin figures.

Operator

Thank you. I will now turn the call over back to Jensen Huang for closing remarks.

A - Jensen Huang {BIO 1782546 <GO>}

Thank you. We had an outstanding quarter. Demand for NVIDIA AI is strong with hyperscalers and cloud services deploying at scale and enterprises broadening adoption. We now help more than 25,000 companies that are using NVIDIA AI. And with NVIDIA AI enterprise software suite, our collaboration with VMware and our collaboration with Equinix place NVIDIA LaunchPad across the world. Every enterprise has an easy arm length to NVIDIA AI. Gaming and Pro Vis are surging. RTX opportunity continues to expand with the growing market of gamers, creators, designers and now professionals building home workstations. We are working harder to increase supply for the overwhelming demand this holiday season.

Last week GTC showcase the expanding universe of NVIDIA accelerated computing. In combination with AI and data center scale computing, the model we pioneered is on the cusp of producing million X speed ups that will revolutionize many important fields; already AI and upcoming robotics, digital biology and what I hope climate signs. GTC highlighted our full stack expertise in action built on CUDA and our acceleration libraries in data processing, in simulation, graphics, artificial intelligence, market and domain specific software is needed to solve customer problems.

We also showed how software opens new growth opportunities for us. But the chips are the enablers, but it's the software that opens new growth opportunities. NVIDIA has 150 SDKs now addressed in many of the world's largest end markets.

One of the major themes of this GTC was Omniverse, our simulation platform for virtual worlds and digital twin. Our body of work and expertise in graphics, physics simulation, AI, robotics and full stack computing made Omniverse possible. At GTC, we showed how Omniverse is used to reinvent collaborative design, customer service avatars and video conferencing and digital twin to factories, processing plants and even entire cities. This is just the tip of the iceberg of what's to come. We look forward to updating you on our progress next quarter. Thank you.

Operator

Thank you. I will now turn over to Jensen for closing remarks.

A - Simona Jankowski {BIO 7131672 <GO>}

Well, I think we just heard the closing remarks. Thank you so much for joining us. We look forward to seeing everybody at the conferences that we have planned over the next few months and I'm sure we'll talk before the end of next earnings. Thanks again, everybody.

Operator

This concludes today's conference call. Thank you all for participating. You may now disconnect.