Q2 2021 Earnings Call

Company Participants

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

Other Participants

- Aaron Rakers, Analyst
- CJ Muse, Analyst
- Joseph Moore, Analyst
- Stacy Rasgon, Analyst
- Timothy Arcuri, Analyst
- Toshiya Hari, Analyst
- Vivek Arya, Analyst

Presentation

Operator

Good afternoon. My name is David and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Financial Results Conference Call. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, 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 Second Quarter of Fiscal 2021. With me on the call today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer.

I'd 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 third quarter of fiscal 2021. 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, August 19 2020, based on information currently available to us. Except as required by law, we assume 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. Q2 was another extraordinary quarter. The world continue to battle the COVID-19 pandemic and most of our employees continue to work from home. But through the teams' agility and dedication, we successfully combined Mellanox into NVIDIA, while also delivering a very strong quarter. Revenue was $3.87 billion, up 50% year-on-year up 26% sequentially and well ahead of our outlook.

Starting with gaming. Revenue was $1.65 billion was up 26% year-on-year and up 24% sequentially, significantly ahead of our expectations. The upside is broad based across geographic regions, products and channels. Gaming growth amidst the pandemic, highlights the emergence of a leading form of entertainment worldwide. For example, the number of daily gamers on Steam on leading PC game online distributor is up 25% from pre-pandemic levels. And MPD reported that US consumer spending on video games grew 30% in the second calendar quarter to a record $11 billion.

NVIDIA's PCs and laptops are ideal for the millions of people, who are now working, learning and gaming at home. At the outset of the pandemic, many retail outlets proposed and demand shifting to online channels. As the quarter progressed and the stores reopened, retail demand picked up. Our cafes

Largely reopened and online sales continue to thrive.

Gaming laptop demand is very strong, as students and professionals turned to GeForce based systems to improve how they work, learn and game from home. We've ramped over 100 new models with our OEM partners, focused on both premium and mainstream price points. In the premium laptop segment, we delivered unparalleled performance with the GeForce RTX 2080 and the 2070 Super GPUs in thin and light form factors. We also brought ray tracing to gaming laptops for the first time at price point as low as $999, with the GeForce RTX 2060.

In the mainstream segment, we brought the GeForce GTX to laptop price points as low as $699. Momentum continues for our Turing architecture, which enables stunning new digital effects in games and its driving powerful upgrade cycle among gamers. It's RTX

technology has ray tracing and AI to programmable sharing and is quickly redefined the new standard for computer graphics. DLSS use the AI capabilities of Turing to boost frame rates by almost 2x, while generating crisp image quality.

RTX support in blockbuster games continues to grow, including mega-hit, Death Stranding, the high anticipated Cyberpunk, 2077 and the upcoming release of Watch Dogs. These games join Minecraft and other major titles that support NVIDIA, RTX ray tracing and DLSH. We're in the midst of a 21 day countdown campaign promoting a GeForce special event on September 1. With each day highlighting a year in the history of GeForce. We don't want to spoil the surprise, but we encourage you to tune in.

We are very pleased with the traction our GeForce NOW cloud gaming service, now in its second quarter are commercially availability. GFN offers the richest content to any game streaming service, through partnerships with leading digital game stores, including Valve Steam, Epic Games and Ubisoft Uplay. GeForce NOW enables users with underpowered PC, Mac or android devices to access powerful GPUs to play their library of PC games in the cloud, expanding the universe of gamers that we can reach with GeForce.

Just yesterday, we announced that GFN is now supported on Chromebooks, further expanding our reach into 10s of millions of users. In addition to NVIDIA's insurance, GFN is available or coming soon to a number of telecom partners around the world, including SoftBank, and KDDI in Japan, Rostelecom and Beeline in Russia, LGu Plus in South Korea and Taiwan Mobile.

Moving to progress. In Q2 was $203 million in revenue down 30% year-on-year and down 34% sequentially with declines in both mobile and desktop workstations. Sales were hurt by lower enterprise demand and with the closure of many offices around the world. Industries negatively impact during the quarter, include automotive, architectural engineering and constructions, manufacturing, media and entertainment and oil and gas. Initiatives by enterprises to enable remote workers drove demand for virtual and cloud-based graphic solutions. Accordingly, our Q2 V-GPU bookings accelerated increasing 60% year-on-year. Despite near term challenges, we are winning new business and areas such as healthcare, including Siemens, Philips and General Electric and the public sector.

We continue to expand our market opportunity with over 50 leading design and creative applications that are NVIDIA RTX enabled, include the latest release from foundry, Chaos Group and Maxon. These applications provide faster ray tracing and accelerated performance, improving creators design workflows. The pandemic will have a lasting impact on how we work. Our revenue mix going forward would likely reflect this evolution in enterprise workforce trends. With a greater focus on technologies, such as NVIDIA laptops and virtual workstations that enable remote work and virtual collaboration.

Moving to automotive. Automotive revenue was $111 million, down 47% year-on-year and down 28% sequentially. This was slightly better than our outlook of a 40% sequential decline, as the impact of the pandemic was less pronounced than expected, with auto production volumes starting to recover after bottoming in April. Some of the declines also

due to the roll off of legacy infotainment revenue, which will remain a headwind in future quarters.

In June, we announced a landmark partnership with Mercedes-Benz, which starting in 2024, will launch software defined intelligent vehicles across the entire fleet and using end-to-end NVIDIA technology. Mercedes utilizing NVIDIA's full technology stack including the DRIVE AGX computer, DRIVE AGX autonomous starting software and NVIDIA's AI infrastructure, standing from the car to the cloud.

Centralizing and unifying computing in the car will make it easier to integrate and upgrade advanced software features as they are developed. With over-the-year update, vehicles can receive the latest autonomous driving and intelligent cockpit features, increasing value and extending the joy of ownership with each software upgrade. This is a transformative announcement for the automotive industry, making the turning point of traditional vehicles becoming high performance, update-able data centers on wheels, this also transformative announcement from NVIDIA to evolving business model, as the software content of our platforms grows. Positioning us to build a recurring revenue stream.

Moving to data center. Data center is a diverse consist of cloud service providers, public cloud providers, supercomputing centers, enterprises, telecom and industrial edge. Future revenue was a record $1.75 billion up of 167% year-on-year and up 54% sequentially. In Q2, we incorporated a full quarter of contribution from the Mellanox acquisition, which closed on April 27 first day of our quarter. Non-ops contributed approximately 14% of company revenue and just over 30% of data center revenue, both compute and networking within data center shut has occurred, with accelerating year-on-year growth.

The biggest news in data center this quarter was the launch of our Ampere architecture. We are very proud of the team's execution in launching and ramping this technology more marvel, especially amidst the pandemic. The A-100 is the largest chip ever made with 54 billion transistors, it went on full software stack for accelerating the most compute-intensive work loads.

Our software releases include CUDA 11, the new version of over 50 CUDA-X libraries and a new application frameworks from Major AI workloads, such as Jarvis, the conversational AI and Merlin for deep recommending learning systems. The A100 deliveries NVIDIA's greatest generational leap ever, using AI performance by 20x over its predecessor, it is also our first Universal Accelerator, unifying AI training and inference and powering workloads such as data analytics, scientific computing, genomics, edge video analytics, 5G [ph] services and graphics.

The first Ampere -GPU A100 has been widely adopted by all major server vendors and cloud service providers. Google Cloud platform was the first cloud customer to bring it to market, making it the fastest GPU to come to the cloud in our history. And just this morning Microsoft Azure announced the availability of processing scalable AI clusters, which are based on the A100 and interconnected with 200 gigabytes per second

Mellanox InfiniBand networking. A100 is also getting incorporated into offerings from AWS, Alibaba cloud, Baidu Cloud and Tencent Cloud.

And we announced that the A100 is going to market with more than 50 servers from leading vendors around the world, including Cisco, Dell, Hewlett Packard Enterprise and Lenovo. Adoption of the A100 into leading server makers offering is faster than any prior launch, with 30 systems expected this summer and over 40 more by the end of the year. The A100 is already winning industry recognition in the latest A100 training benchmark in Oculus 0.7, NVIDIA set 16 record, its sweeping all categories for commercially available solutions in both per chip and out scale performance, based on the A100 and MLPerf offers the industries first and only objective AI benchmark.

Since the benchmark was introduced 2 years ago, NVIDIA has consistently delivered leading results and record performance for both training and inference. NVIDIA also top the charge in the latest Top 500 list of the fastest supercomputers, the ranking released in June showed that 8 of the world's top 10 supercomputers using NVIDIA GPUs, NVIDIA networking or both. They include the most powerful systems in the US and Europe. NVIDIA now combined with Mellanox powers 2/3s of the Top 500 systems on the list compared with just less than a half for the two companies in total two years ago.

In energy efficiency, systems using NVIDIA GPUs are pulling away from the path. On average, they're nearly 2.8x more powerful sufficient than systems without NVIDIA GPUs measured by gigaflops per watt. The incredible performance and efficiency of the A100 GPU is fast amplified by NVIDIA's own new Selene supercomputer, which debuted as number 7 on the Top 500 list and is the only top 100 system to cross the 20 gigaflops per watt barrier. Our engineers were able to assemble Selene in less than 4 weeks using NVIDIA's open modular DGX SuperPOD reference architecture. Instead of the typical build trying of months or even years. This is the system that we will used to win the MLPerf benchmarks. And it is a reference design, it's available for our customers to quickly build a world-class supercomputer.

We also brought GPU acceleration to data analytics, one of the largest and fastest growing enterprise workload, we enabled end-to-end acceleration of the entire data analytics workload pipeline for the first round with NVIDIA's GPUs and software stack in the latest version of Apache Spark released in June. Spark is the world's leading data analytics platform used by more than 500,000 data scientists and 16,000 enterprises worldwide.

And we have two major milestones to share. We have now shipped a cumulative total of 1 billion CUDA GPUs and the total number of developers in the NVIDIA ecosystem just reached 2 million. It took over a decade to reach the firm with first million and less than years to reach the second million. Now we -- have fantastic results across the board in its quarter as part of NVIDIA. Now to revenue, growth accelerated the strength across Ethernet and InfiniBand products.

Our Ethernet shipments reached a new record, major hyperscale builds drove the upside in the quarter, as growth in cloud computing and AI is fueling increased demand for high-

performance networking. Mellanox networking was a critical part of several of our major new product introductions this quarter, these include the GTX AI system, the DGX SuperPOD clusters for our Selene supercomputer and the EGX edge AI platform. We also launched the Mellanox Connect X6, Ethernet NIC, the 11th generation product of the Connect 6 family and it's designed to meet the needs of modern cloud and hyperscale data centers were 25, 50 and a 100 gigabytes per second is becoming the standard.

We expanded our switch networking capabilities with the addition of Cumulus Networks, a privately held network software company that we purchased in June. Cumulus augment our Mellanox acquisition and building out open modern data center. The combination of NVIDIA accelerated computing Mellanox networking and Cumulus software enables data centers that are accelerated, disaggregated and software defined to meet the exponential growth in AI, cloud and high performance computing.

Moving to the rest of the P&L, Q2 GAAP gross margin was 58.8% and Non-GAAP gross margin was 66%. GAAP gross margin declined year-on-year and sequentially due to cost associated with Mellanox acquisition. Non-GAAP gross margins increased by almost 6 points year-on-year, reflecting a shift in product mix with higher data center sales and lower automotive sales. Q2 GAAP operating expenses were $1.62 billion and non-GAAP operating expenses were $1.0 4 billion, up 57% and 38% from a year ago, respectively. Q2 GAAP EPS as $0.99, up 10% from a year earlier. Non-GAAP EPS was $2.18 up 76% from a year ago. Q2 cash flow from operations was $1.57 billion.

With that let me turn to the outlook for the third quarter of fiscal 2021. We expect revenue to be $4.4 billion plus or minus 2%. With that, we expect gaming to be up just over 25% sequentially, with data center to be up in the low to mid single-digits sequentially. We expect both Pro Viz and Auto to be at similar levels out in Q2. GAAP and non GAAP gross margins are expected to be 62.5% and 65.5% respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $1.54 billion and $1.09 billion, respectively. Full year GAAP and non-GAAP OpEx is tracking in line with our outlook of $5.7 billion and $4.1 billion, respectively.

GAAP and non-GAAP OI&E are both expected to be expense of approximately $55 million. GAAP and non-GAAP tax rates are both expected to be 8%, plus or minus 1%, excluding discrete items. Capital expenditures are expected to be approximately $225 million to $250 million.

Further financial details are included in the CFO commentary and other information available on our IR website. In closing, let me highlight upcoming events for the financial community. We will be at the BMO Virtual Technology Summit on August 25. Citi's 2020 Global Technology Conference on September 9, Deutsche Bank's Technology Conference on September 14 and the Evercore's Virtual Memo forum, the future of mobility on September 23. We will also host a financial analyst q&a with Jensen on October 5, as part of our next virtual GTC. Our earnings call to discuss our third quarter results is scheduled for Wednesday, November 18.

We will now open the call for questions. Operator, would you please call for questions? Thank you.

Questions And Answers

Operator

(Operator Instructions) Your first question comes from the line of Vivek Arya with Bank of America. Your line is open.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question. And congratulations on the strong growth and execution. Jensen, I'm wondering how much of the strength that you're seeing in gaming and data center is maybe more temporary because of COVID or some customer pull-ins in the data center or so forth. And how much of it is more structural and more secular? And that can continue, even once we get into hopefully, sooner rather than later into a more normalized period for the industry?

A - Jensen Huang {BIO 1782546 <GO>}

Yes. Vivek, thank you. So first of all, we didn't see colons. And we're in the beginning of our brand new product cycle with Ampere. And so the vast majority of the data center growth came from that. The gaming industry, with all that's happening around the world and it's really unfortunate. But it's made gaming the largest entertainment medium in the world. More than ever people are spending time digitally, spending on their time in video games. The thing that the people haven't realized about video games. Is that it's not just the game itself anymore.

The variety of different ways that you can play weather you can hang out with your friends in fortnight, go to a concert in fortnight, building virtual worlds in Minecraft. You're spending time with your friends, you're using it to create, to realize your imaginations. People are using it for broadcast, for sharing ideas and techniques with other people and so.

And then of course, it's just an incredibly fun way to spend time even by consumption of a video game. And so there's just so much that you could do with video games now. And I think that this way of enjoying entertainment digitally has been accelerated as a result of the pandemic. But I don't think it's going to return. Video game adoption has been going up over time and pretty steadily. PC is now the single largest entertainment on platform, it is the largest gaming platform and GeForce is now the largest gaming platform in the world.

And as I mentioned, it's not just about gaming, there's just so many different ways that you can enjoy games. With data center, the things like the cultural change that's happening in data center are a couple of different dynamics that are happening at the same time. The first dynamic, of course, is the movement to the cloud. The way that a cloud data center is built in a way that enterprise data center or cluster is built, is

fundamentally different. And, it's really, really beneficial to have the ability to accelerate applications that cloud service providers would like to offer which is basically everything. And we know that one of the most important applications of today is, artificial intelligence. It's a type of software that really wants acceleration and NVIDIA's GPU acceleration is the perfect medium and perfect platform for it.

And then the last reason about the data center is the architectural change, from hosting applications to hosting services that's driving this new type of architecture called disaggregation versus hyper converge. And the original name of hyper scalars is a large data center of a whole bunch of hyper converged computers. But the computers of today are really, really just aggregated. A single application service could be running on multiple servers at the same time, which generates a ton of East West traffic. And a lot of it is artificial intelligence, neural network models.

And so because of this type of architecture, two components or two types of technologies are really important feature of cloud. One of them, as I mentioned was -- is the acceleration and our GPU is ideal for it. And then the other one is high speed networking. And the reason for that is because the servers now disaggregated, the application is fractionalized and broken up into a bunch of small pieces that are running across the data center. And whenever an application you could send a parts of the answer to another server for the microservice to run, that transmission is called East West traffic. And the most important thing you could you could possibly do for yourself, is to buy really high speed, low latency networking and that's what Mellanox was fantastic.

And so we find ourselves really in this perfect condition, where the future is going to be more virtual, more digital and that's one of the -- that's the reason why GeForce is so successful. And then we find ourselves in a world where the future is going to be more autonomus and more AI driven and that's the benefit of our GPUs. And then lastly, cloud, micro service transactions, really benefit high speed networking and that's where Mellanox comes in. And so I think that this is the dynamics that I'm describing are permanent. And it's just been accelerated to the present, because of everything that's happening to us, but this is the future and it's not there's no going back and we just found everything accelerated.

Operator

Your next question comes from the line of Timothy Arcuri with UBS. Your line is open.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. Jensen, I guess, I had a question on both architecture and also manufacturing. And I think on the manufacturing side, even where to go now for some time and when you did ask in the past about, moving to more of a tiled or tripled approach, usefully made light of that, that the CPU guys are clearly taking that approach. So I guess the question is, why do you think you won't have to make a similar move? And then on the side of architecture, the theme of hardship this week was really how compute demand is far outstripping computing power. And then we see this talk about you in

ARM. So I guess, can you talk about whether GPU is the future and maybe the prior opportunity to integrate CPU and GPU? Thanks.

A - Jensen Huang {BIO 1782546 <GO>}

Yes, we push architecture really hard. And the way we push architecture is, we start from the system. We believe that the future computer company as a data center scale company. The computing unit is no longer a microprocessor or even a server or even a cluster. The computing unit is an entire data center now. And as I was explaining to Vivek just now that a micro service that you're enjoying hundreds of billions of transactions a day, but those are broken up into a whole bunch of micro services that are running across the entire data center.

And so the data center is running -- the entire data center is running an application. I mean, that's pretty remarkable thing. And that's happened in the last several years. We were ahead of this trend and we recognize that as a computing company, we have to be a data center scale company and we architect from that strength starting. If you look at our architecture, we were the first in the world to create the concept of an NVlink where the eight processors being fully synchronized across a computing node and we created the DGX.

We recognize the importance of high speed networking and low latency networking and that's why we bought Mellanox. And the amount of software that we invented along the way to make it possible for low latency communications, whether it's a GPU direct or recently the invention of GPU direct storage, all of that technology was inspired by the idea that you have to think about the data center, all in one holistic way. And then in the last -- in this current generation with Ampere.

We invented the world's first multi-instance GPU, we invented the world's first multi-instance GPU which means that our Ampere GPU could simultaneously be one GPU or with NVlink, eight GPUs combined working together. So you could, think of them as being titled, so those eight GPUs are working anonymously together or each one of the GPUs that fraction life itself, if you don't need that much GPU working on your workload, fractionalize into a multi GPU instance we call them MIG, a little tiny incidence.

And each one of those time incidence are more powerful and more performance than our entire Volta GPU lap time. And so, what do you like to fractionalize a GPU which happens oftentimes, create a larger GPU using NVlink or create an even larger GPU the size of DGX pipe connected together with high-speed, low latency networking with Mellanox. We could architect it anyway you like.

You made a comment about -- you asked the question about ARM, we've been a long-term partner of ARM. And we use ARM in a whole bunch of applications. And whether it's an autonomous driving or robotic application, the Nintendo Switch, console business that we're in. And then recently, we brought CUDA to ARM and to bring accelerated computing to ARM. And so, so we work with the ARM team very closely, they're really great guys. And one of the special about the ARM architecture that you know very well, either it's incredibly energy efficient and because of energy efficient, it has the headwind

to scale into very high performance levels over time. And so anyways we love working with ARM guys.

Operator

Your next question comes from the line of Aaron Rakers with Wells Fargo. Your line is open.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Thanks for taking the question and congratulations on the quarter. Just building on some prior questions. The first one, I was just curious if you could help us appreciate kind of the installed base of the gaming GPU business relative to where we're at the Turing upgrade cycle? What do we see still on prior generations, be it Pascal or before?

And then secondly, I was wondering -- collect. Could you just give me any kind of updated commentary or views on visibility in the data center business? how that -- has that changed over the last three months? What does that look like as you look through the back half of the calendar year? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yes. Thanks a lot, Aaron. We are still in the ramping of the RTX generation. Turing, the current generation that we're in, is the world's first ray tracing GPU and infused is the RTX technology fuses three fundamental technologies. The Programmable shader that we introduced a long time ago that revolutionized computer graphics and we added two new technology.

One technology is a ray tracing acceleration core that makes the tracing of rays and looking for intersections within the ray and the same objects in sync super, super fast. And that's -- it's a complicated problem, it's a super complicated problem, we wanted to be running concurrently to shape. So the rate reversal and the stating of the pixels could be done independently and concurrently.

The second thing, as we invented this technology to bring AI, Artificial Intelligence using this new type of algorithm called deep learning to computer graphics. And one example of its capability is the algorithm we introduced called DLSS Deep Learning Super Sampling, which allows us to essentially synthesize and by learning from previous examples. Essentially learning from previous examples of images and remembering it or remembering what beautiful images look like. So that when we take the low resolution image and you run it through the deep neural network, it synthesizes a high resolution image that's really, really beautiful.

And people have commented that it's even more beautiful than native rendered images at the native resolution. And the benefit is not only the beautiful, it's also super-fast. We essentially nearly double the performance of RTX as a result of doing that. So you can have the benefit of ray tracing as well as very high resolution and very high speed. And so that's called RTX.

And Turing is probably not even close, not even 1/3 of the total installed base, of all of our GeForce GPUs, which is, as you know, the single largest installed base of gaming platforms in the world. And so we support this large installed base and we're in the process and bringing them to the future with RTX. And now with the new console generation coming, every single game developer on the planet is going to be doing ray tracing and they're going to be creating much richer content. And because of multi-platform, cross platform play, and because of the size of the gaming platform, PC gaming platform, it's really important that these game developers bring the latest generation content to PC's, which is great for us.

Q - Aaron Rakers {BIO 6649630 <GO>}

And then on the data center visibility?

A - Colette Kress {BIO 18297352 <GO>}

Yes. Let me see, if I could answer this one for you. Yes, we have been talking about our visibility of data center and as you've seen in our Q2 results, you can see that our overall adoption of the NVIDIA computing portfolio has accelerated quite nicely. But keep in mind, we're still really early in the product cycle, A100 is ramping, it's ramping very strong into our existing installed bases, but also into new markets.

Right now A100 probably represents less than a quarter of our data center revenues. So we still have a lot to grow. We have good visibility looking into Q3 with our hyperscales. We have a little bit more of a mixed outlook in terms of our vertical industries, given a lot of the uncertainty in the market and in terms of the overall economy.

On-premises are challenged, because of the overall COVID. But remember industries are quickly and continuing to adopt and move to the overall cloud. But overall, we do expect a very strong Q3.

Operator

Your next question comes from the line of CJ Muse with Evercore ISI. Your line is open.

Q - CJ Muse

Yes. Hi, thank you for taking the question. I guess two questions. If I look at your outstanding inventory purchase obligations grew 17% sequentially. Is that, as you prepare for the September one launch and can you kind of comment on gaming visibility into the back half of the year? And then the second question. Jensen, I know you're very focused on platforms and driving recurring revenues. Would not be here, if there is any particular platforms over the last three months we've made real headway or gets you excited. Whether a Jarvis, Marlins, Sparks or whatever? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Thanks a lot CJ. We're expecting a really strong second half for gaming. I think, this may very well be one of the best gaming seasons ever. And the reason for that is because,

because PC gaming has become such a large format. The combination of major games like Fortnite and Minecraft, because of the way people game now, they're gaming and their E-spoiling, even F1 is in E-sport now. They are hanging out with friends, they're using it to create other content. They're using game captures to create art, they're sharing it with the community. It's a broadcast medium.

The number of different ways you could game as just really, really exploded. And it works on PCs, because all of the things that I described require cameras or keyboards or streaming systems or -- but it requires an open system that is multi-tasking. And so the PC have just become a large platform for gaming.

And the second thing is, is that RTX. It's a home run. We really raise the bar with computer graphics in the games are so beautiful. It's really going to the next level. It's not been just amazing since we introduced programmable shaders about 15 years ago. And so for last 15 years, we've been making programmer shares better and better and better and it hasn't getting better, but its never been a giant leap like this, and RTX brought both artificial intelligence, as well as ray tracing to PC game. And then the third factor is the console launch, those people are really -- the game developers are really gearing up for a big leap and because of the gaming, because how vibrant the gaming market is right now and how many people around the world is depending on gaming at home. I think it's going to be the most amazing season ever.

We're already seeing amazing numbers from our console partner Nintendo. The switch has about to sell more than Super Nintendo, more than all the Famicom. I mean, which was one of the best growing console of all time. I mean, they are on their way to make Switch the most successful gaming platform of all time. And so I'm super excited for them. And so I think it's going to be a quite a huge second half for gaming.

Operator

Next question comes from the line Toshiya Hari with Goldman Sachs. Your line is open.

A - Jensen Huang {BIO 1782546 <GO>}

Colette, I feel like, I missed the CJ's second question. Can we jump on and answer it?

A - Colette Kress {BIO 18297352 <GO>}

I think the question was regarding our inventory purchases on that piece. Is that the part of your -- referring to -- Keep in mind CJ that when you think about the complexity of the products that we are building, we have extremely long lead times, both in terms of what we produce for the data center. Our full systems that we need to do, as well as what you are seeing now between the sequential growth between Q2 and Q3 for overall gaming. So all of that is in preparation for the second half. Nothing unusual about it, other than just, we've got to hit those our revenue numbers that are in our Q3 guidance.

Q - CJ Muse

Okay.

Operator

Your next question comes from the line of Toshiya Hari with Goldman Sachs. Your line is open.

Q - Toshiya Hari {BIO 6770302 <GO>}

Hi, good afternoon and thank you so much for taking the question. I had one for Jensen and another one for Colette. Jensen, just following up on the data center business, as you probably know quite a few of your peers have been talking about potential digestion of capacity on the part of their hyperscale customers over the next, call it, 6 to 12 months. Curious, is that something that you think about, worry about in your data center business or do you have enough idiosyncratic growth drivers, like the A100 ramp. I guess the breadth that you've built within the data center business across computing networking or are those enough to -- for you to buck the trend within data center over the next 6 to 12 months?

And then the second one for Colette. Just on gross margins, you're guiding October quarter gross margins down 50 basis points sequentially, based on the color that you provided for the individual segments. It looks like the mix remains pretty positive. So just curious what's driving the marginal decline in gross margins in the October quarter? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yes. Thank you. So and thanks for the question. The -- our data center trend is really tied to a few factors. One is the proliferation of using deep learning and artificial intelligence in all the services that are in -- by the cloud service providers. And I think it's fair to say that, over the last several years, the number of breakthroughs in artificial intelligence has been really terrific. And we're seeing anywhere from 10 times, 10x more computational requirement each year to more than that. And so in the last three years, we've seen somewhere between a 1,000 to 3,000 times increase in the size of model of the computational requirement necessary to create the AI models and to deploy these AI models.

And so the number one trend that we are -- probably indexed to, is the breakthroughs of AI and the usefulness of AI and how people are using it? And one of the -- and I remember the CJ question there is -- I'll answer this along with that. One of the things that we look for and you should look for is help how -- what kind of breakthroughs are based on deep learning and based on AI that these services all demand. And there are three big ones, just gigantic ones.

Of course, one of them is natural language understanding. The ability to take very complicated text and use deep learning to create essentially a -- the mentioned reduction called deep embedding, the mentioned reduction on that body of text, so that you could use that vector as a way to teach a recommender system, which is the second major breakthrough, the recommender system how to predict and make a recommendation to somebody.

Recommendation on adds and videos and there are trillions of videos on the web. You need ways to recommend them. Both the news and just the amount of information that is going to -- that is in true dynamic form, require these recommenders to be instantaneous. And so the first one is natural language understanding, the second one is the recommender system, gigantic breakthroughs in the last several years.

And third is conversation with AI and we're going to have conversational agents that are just super clever and they can predict what you're about to ask. They're going to predict the right answer for you, make recommendations to you. Based on the three pillars that I've just described. And I haven't even started talking about robotics. The breakthroughs that are happening there with all the factories need to automate. And the breakthroughs that we're seeing in self-driving cars, so the models there are, are really improving fast.

And so the answer to you, Toshiya and CJ are kind of similar. That on the first one, we're index to AI, the second we're index to breakthroughs of AI. So that you can continue to consume more-and-more capability and more technology. And then the third thing that we're indexed to, is the movement of workloads to the cloud. It is now possible to do rendering in to cloud. Remote graphics workstations in the cloud. And NVIDIA virtual workstation is in every single cloud.

You could do big data analytics in the cloud and these applications are just giving you a few applications, where you can do scientific computing in the cloud. These applications all have fundamentally different computing architectures. NVIDIA is the only accelerated architecture that allows you to do micro services for conversational AI and other types of AI applications to still allow applications like high performance computing, training, big data analytics to virtualized applications like workstation. Our platform is universal and these three facts that I just described are supremely complex, virtualized, micro services based and scale up based. And so these bare metal scale up and these are complicated and it's one of the reasons why we bought Mellanox, because they're at the core and at the intersection of all of that. The storage, the networking, the security, the virtualization, they're at the intersection of all of that.

And I just described three dynamic that are very, very powerful and are at the early stages yet. And so those are the things that we're really index to and then lastly, when somebody adopts, they'll introduce a new platform like Ampere and/or in the beginning of a multi-year product cycle. Ampere is such a gigantic gigantic breakthrough. It's the first universal GPU we've ever created. It is both able to scale up as well as scale out, scale up as a modern GPU scale out as fractionalization, multi instance GPUs. And it reduce -- it saves money, tremendous amount of money for people who use it.

It speeds up their application and reduces their TCO. Their TCO value just goes through the roof. And so, we're in the beginning of this multi-year cycle and the enthusiasm has been fantastic. This is the fastest ramp we've ever had. And so we're going to keep on racing through the second half.

A - Colette Kress {BIO 18297352 <GO>}

Okay. And Toshiya you asked a question regarding our guidance going forward, regarding gross margin. And within our Q3 guidance, we have just a smaller decline in our gross margin from Q2. Most of that is really associated with mix, but also a little bit in terms of the ramping on our new Ampere architecture products that we have. To keep in mind, our data center will likely be a lower percentage of total revenue, given the strong overall gaming growth that we expect between Q2 and Q3. Within that gaming growth keep in mind, consoles are also included, which will continue to be below our company totals, average gross margin and that is expected to be up strongly quarter-over-quarter for our overall console shipment. We're going to be ramping those new architectures over time, we have the ability to expand our gross margin as Ampere GPUs mature too.

Operator

Your next question comes from the line of Stacy Rasgon with Bernstein Research. Your line is open.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi guys. Thanks for taking my question. I wanted to dig in to the data center a little bit, this is a question for Colette. So in the quarter ex Mellanox data center was up, core data center maybe 6%, 7%. The guide looks to be roughly similar to that in the Q3. Can you talk to us a little bit about what's driving the trajectory. Are you more demand or more supply limited at this point? What is your supply situation look like and what are the lead times, especially on the A100 products for data center, a look like at this point. But if you had more capacity available, do you think you have like a stronger trajectory than you have right now?

A - Colette Kress {BIO 18297352 <GO>}

Yes. Stacy, thanks for the question. Let me first start on our Q3 outlook and what we're seeing and when we think about our demand and our supply, we're very comfortable with the supply that we have. Keep in mind, our products are quite complex. And a lot of our time is spent in terms of procuring every aspect of that supply over multiple quarters previously. So that's how we work, but we are very confident with the overall supply that we have across the board in data center. Keep in mind, that's not just A100, we are continuing to sell our V100 our T4 and we're also bring new versions of the A100, are coming to overall market. So I hope that helps you understand our statements on where we have in terms of Q3 guidance. I'll see if Jensen wants to add a little bit more to that.

A - Jensen Huang {BIO 1782546 <GO>}

Well, when we're ramping, we sure love to have more and sooner and but this is our plan and we're executing to the plan. It is a very complicated product as Colette mentioned, it is the most --

Q - Stacy Rasgon {BIO 16423886 <GO>}

Got it. And just a quick follow-up. Within the data center guidance, how do you think about like the core data center sequential growth versus Mellanox?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So in terms of moving from Q2 to Q3, we believe that most of the actual growth that we will receive in that single -- both single digits to mid single digit growth will likely stem from NVIDIA compute, will be the largest driver of that.

Operator

Your next question comes from the line of Joseph Moore with Morgan Stanley. Your line is open.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. I wonder, if I could ask a longer-term question about the -- how you guys see the importance of process technology. There has been a lot of discussion around that in the CPU domain. But I know you guys haven't really felt the need to be first on 7-nanometer and you've done very well. Just how important do you think it is to be early in a new process node and how does that factor into the cycle of innovation in NVIDIA?

A - Jensen Huang {BIO 1782546 <GO>}

Yes. Good. Thanks, Joe. The process technology is a lot more complex than a number. I think people has simplified it down to almost a ridiculous level. And it's a process technology, we have a really awesome process engineering team. World-class, everybody will recognize it, absolutely world-class. And we work with the foundries, we work with TSMC really closely, to make sure that we engineer transistors that are ideal for us. That we engineered mineralization systems it's ideal for us. It's a complicated thing and we do it at high point.

Then the second part of it is, is their architecture, where the process technology and the rest of the design process. The architecture of the chip in the final analysis what NVIDIA paid for is architecture, not procurement of transistors. We're paid for architecture and there is a vast difference between our architecture and the second best architecture and the rest of the architectures.

The difference is incredible, we are easily twice the energy efficiency all the time. Irrespective of the number of -- in the transistor side. And so it must be more complicated than that. And so we are -- we put a lot of energy into that. And then last thing I would say is that, going forward, it's really about data center scale computing. Going forward, you optimize at the data center scale and the reason why I know this for a fact, is because if you were a software engineer you would be sitting at home right now and you will write a piece of software that runs on the entire data center in the cloud. We have no idea what's underneath it, nor do you care.

And so what you really want, is to make sure that, that data centers is high throughput as possible, there are a lot of code in there. And so what NVIDIA has decided to do over the years, is to take our game to a new level. Of course we start building the world's best processors and we use the world's best foundries and we partner with them very closely

to engineer the best process for us. We partner with the best packaging companies to create the world's best packaging. We're the world's first user of cobots. And whether it's -- I think we're -- I'm pretty sure we're still the highest volume by far of 2.5D and 3D packaging.

And so we start from a great chip. We start from great chip. But we don't end there. That's just the beginning for us. We take this being all the way through systems to systems software, algorithms, networking, all the way up to the entire data center. And the difference is absolutely shocking. We built our data center Selene and it took us 4 weeks. We put up this -- put up Selene in 4 weeks time it is the 7th fastest supercomputer in the world. One of the fastest AI supercomputers in the world. It's the most energy efficient supercomputer in the world. And it broke every single record in MLPerf. And I can assure you something about the scale that we work and the complexity of the work that we do. This is the future, it's for -- this future is about data centers.

Operator

And there are no questions at this time. Jensen Huang, I turn the call back over to you.

A - Jensen Huang {BIO 1782546 <GO>}

Thank you. We accelerated computing model we pioneered has clearly past the tipping point. Adopting of NVIDIA computing is accelerating. On this foundation and leveraging one architecture, we have transformed our company in three dimensions. First, NVIDIA is a full stack computing platform company. Often the world's most dynamic industry which chips, system, software and libraries like NVIDIA AI to tackle the most pressing challenges. And second NVIDIA's data center scale company with capabilities to architect, build and operate the most advanced data centers. But data center is the new computing unit, with this capability we can create modern data center architectures that are computer maker partners and then scale out to the world's industries.

Third NVIDIA is a software defined company today, with rich software content by GeForce NOW. NVIDIA virtual workstation in the cloud. NVIDIA AI and NVIDIA DRIVE that will add recurring software revenue to our business models. In the coming years, AI will revolutionize software. Robotics will automate machine, and the virtual and physical worlds will become increasingly integrated through VR and AR. Industry advancements will accelerate and NVIDIA accelerated computing will play an important role. Our next GTC will be coming on October 5, again from my kitchen. Joining me -- I have some exciting developments to share with you. Thanks everyone.

Operator

This concludes today's conference call. You may now disconnect.
