Q3 2021 Earnings Call

Company Participants

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

Other Participants

- Aaron Rakers, Analyst
- Ambrish Srivastava, Analyst
- C.J. Muse, Analyst
- Harlan Sur, Analyst
- John Pitzer, Analyst
- Stacy Rasgon, Analyst
- Timothy Arcuri, Analyst
- Vivek Arya, Analyst
- William Stein, Analyst

Presentation

Operator

Good afternoon. My name is Jason and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Third Quarter 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 third 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 fourth 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 Forms 10-K and 10-Q and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, November 18, 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>}

Thank you, Simona. Q3 was another exceptional quarter with record revenue of $4.73 billion, up 57% year-on-year, up 22% sequentially and well above our outlook. Our new NVIDIA Ampere GPU architecture is ramping with excellent demand across our major market platforms. Q3 was also a landmark quarter both for us and the industry as a whole. As we announced plans to acquire Arm from SoftBank for $40 billion, we are incredibly excited about the combined companies opportunities and we are working through the regulatory approval process. For today, we will focus our remarks on our quarterly performance.

Starting with Gaming. Revenue was a record $2.27 billion, up 37% year-on-year, up 37% sequentially and ahead of our high expectations. Driving strong growth was our new NVIDIA Ampere architecture based GeForce RTX 30 Series of gaming GPUs. The GeForce RTX 3070, 3080 and 3090 GPUs offer up to two times the performance and two times the power efficiency over the previous turning based generation. Our second-generation NVIDIA RTX combines ray tracing and AI to deliver the greatest ever generational leap in performance. First announced on September 1 and ranging in price from $499 to $1,499, these GPUs have generated amazing reviews and overwhelming demand. PCWorld called them staggeringly powerful, while Newegg cited more traffic than Black Friday. Many of our retail and e-tail partners sold out instantly. The RTX 30 series drove our biggest ever launch. But we had anticipated strong demand. It exceeded even our bullish expectations. Given industry-wide capacity constraints and long cycle times, it may take a few more months for product availability to catch up with demand.

In addition to the NVIDIA Ampere GPU architecture, we announced powerful new tools for gamers as well as for tens of millions of live streamers, broadcasters, Esports professionals, artists and creators. NVIDIA Reflex is a new technology that improves reaction time in games, reducing system latency by up to 58%. NVIDIA Reflex is being integrated into popular Esports games, such as Apex Legends, Call of Duty Warzone, Fortnite and Valorant. NVIDIA Broadcast is a universal plug-in for video conferencing and live streaming applications that enhances the quality of microphones, speakers and webcams with NVIDIA AI effects such as audio noise removal, virtual background effects

and webcam audio frame. With it remote workers and live streamers can turn any room into a broadcast studio.

Blockbuster games continue to adopt NVIDIA's RTX ray tracing and AI technologies at the games announced at Fortnite, which has more than 350 million players worldwide is adding NVIDIA RTX real-time ray tracing, NVIDIA DLSS AI super resolution and NVIDIA Reflex, making the game more beautiful and even more responsive. Other major new titles featuring RTX this holiday season include Watch Dog Legion, Call of Duty, Black Ops Cold War and much anticipated Cyberpunk 2077. Gaming laptop demand was also strong with double-digit year-on-year growth for the 11th quarter in a row. NVIDIA GeForce Laptop Support, the most demanding applications for creators and designers while doubling as a powerful gaming race by night. We also had record gaming console revenue on strong demand for the Nintendo Switch. And we continue to grow our cloud gaming service, GeForce NOW, which has doubled in the past seven months to reach over 5 million registered users.

GeForce NOW is unique as an open platform that connects to popular game stores including Steam, Epic Games and Ubisoft Connect, allowing gamers access to the titles they already own. 750 games are currently available on GFN, the most of any cloud gaming platform, including 75 free to play games with more games added every Thursday. GFN supports many popular clients including PCs, Macs and Chromebook. Stay tuned for more devices to come in the near future. In addition, GFN's reach continues to expand through our telco partners in a growing list of countries including Japan, Korea, Taiwan, Russia and Saudi Arabia. We are also providing technology that enables the cloud in services to an expanding number of partners. Following our earlier announcement with Tencent, Amazon and Facebook are beginning to offer cloud gaming services powered by NVIDIA.

Moving to ProVis, Q3 revenue was $236 million, down 27% year-on-year and up 16% sequentially, ahead of our expectations. Sequential growth was driven by strength in notebooks, which posted record revenue boosted by work from home initiatives and the shift to send and mobile workstations. This is particularly offset by decline in desktop workstations, which continue to be impacted by the pandemic and drove the year-on-year decline. From an industry demand perspective, stronger verticals including healthcare, public sector, higher education and research and financial services. We continue to win new business in a number of areas. In healthcare we added Medtronic for visual surgical applications and (inaudible) for medical imaging.

In technology and media and entertainment, we gained wins for design, rendering and broadcast applications. During the quarter, we announced that Omniverse, the world's first 3D collaboration and simulation platform has entered open beta. Omniverse enables the tens of millions of designers, architects and creators to collaborate real time on-premises or remotely, using the virtual and physical world on Omniverse brings together NVIDIA breakthroughs in Graphics, Simulation and AI. It will help enterprises address evolving requirements as work forces become increasingly distributed. Initial market response from this transformative platform has been phenomenal. Over 400 individual creators and developers in diverse industries have been evaluating Omniverse, and early adopters, including Ericsson, BMW, Foster + Partners, and Lucasfilm.

The pandemic is accelerating development of AR, VR and mixed reality technologies, which will have a profound impact on how we work and play. For example, our work with NASCAR to enable a variety of AR and VR services at the edge is revolutionizing the racing experience for millions of fans across the globe. With our industry-leading real time ray tracing graphics, AI and simulation hardware and software stacks, NVIDIA is in a unique position to enable the future of blending the physical and virtual world.

Moving to Automotive, Q3 revenue was $125 million, down 23% year-on-year and up 13% sequentially. Sequential growth was driven by a recovery in global automotive production volumes, as well as continued growth in AI cockpit revenue. The year-on-year decline was due to the expected ramp down of legacy infotainment revenue. In September, Mercedes-Benz debuted its redesign of S-Class sedan, featuring an all new NVIDIA powered Amdocs AI cockpit system with an augmented reality heads up display. AI voice assistance and rich interactive graphics to enable every passenger in the vehicle to enjoy personalized intelligent features.

Also in September, Li Auto, a leading electric car brand in China announced that it will develop its next-generation of vehicles using the software defined NVIDIA DRIVE AGX Orin platform. Orin delivers nearly seven times the performance and three times the energy efficiency of our previous generation SoC making it uniquely capable to power next generation autonomous electric vehicles. We have excellent traction with the start ups. Finally, last week, NVIDIA and Hyundai Motor Group announced the automakers entire lineup of Hyundai, Kia and Genesis models will come standard with NVIDIA DRIVE in-vehicle infotainment systems starting in 2022. This feature rich software defined computing platform will allow vehicles to be perpetually upgraded with the latest AI cockpit features.

Now moving to Data Center, revenue was a record $1.9 billion, up 162% year-over-year and up 8% sequentially. Driving growth was the strong ramp of our A100-based platforms, continued growth with Mellanox and record T4 shipments for inference. Let me give you a little bit of color on each. Our new NVIDIA Ampere architecture gained further adoption by cloud and hyperscale customers and started ramping into vertical industries. Over the past weeks Amazon Web Services, Oracle cloud infrastructure and Alibaba Cloud announced general availability of the A100 following Google Cloud platform and Microsoft Azure. A100 adoption by vertical industries drove strong growth.

As we began shipments to server OEM partners who is broad enterprise channels reach a large number of end customers. We also ramped the GE EGX A100 server and began shipping NVIDIA DGX SuperPOD, the first turnkey AI infrastructure. These range from 20 to 140 DGX A100 systems interconnected with Mellanox's HDR InfiniBand networking and enabled customers to install incredibly powerful AI supercomputers in just a few weeks time. In fact, we have announced plans to build an 80 node DDX SuperPOD with 400 kind of plots of AI performance. Okay, which one, which will be in the UK's fastest AI supercomputer. That will be used by NVIDIA researchers for collaborative research within the UK's, AI and healthcare community across academia, industry and startups. It joins other systems in NVIDIA's complex of AI supercomputers powered by our R&D and autonomous vehicles, conversational AI, robotics, graphics, HPC and other domains. This includes Selene, now the world fifth fastest supercomputer and fastest commercial

supercomputer and a new NVIDIA DGX SuperPOD, which ranked first on the Green 500 list of the world most energy efficient supercomputers.

A great example of the tremendous opportunities for AI in healthcare is our new partnership with GSK for applying computational to the drug and vaccine discovery process. GSK, London based AI hub will utilize biomedical data, AI methods and advanced computing platforms to unlock genetic and clinical data with increased precision and scale. In addition to this investment in NVIDIA's DGX A100 system, GSK will have access to NVIDIA's Cambridge-1, the NVIDIA Clara discovery software and NVIDIA scientists.

In Q3, the A100 swept the industry standard ml per benchmark for AI inference performance. Following our suite in the prior quarters amount for benchmark for AI training. Notably, our performance led in AI inference actually extended compared with last year's benchmark. For example, in the ResNet-50 test for image recognition, or A100 GPU beat CPU, only systems by 30 times this year versus six times last year. Additionally, A100 outperformed CPUs by up to 237 times in the newly added recommender test, which represent some of the most complex and widely used AI models on the Internet. Our winning performance in AI inference is translating to continued strong revenue growth. Alongside the continued ramp of the A100 T4 sales set a record, as the NVIDIA AI inference adoption is in full throttle.

We estimate that NVIDIA's installed GPU capacity for inference across the seven largest public cloud now exceeds that of the aggregate CPU capacity in the cloud, testament to the tremendous performance and TCO advantage of our GPUs. Hundreds of companies now operate AI-enabled services on NVIDIA's inference platform, including the A100 or T4 GPU and our Triton inference serving software. For example, Tencent uses NVIDIA AI inference to recommend video, music, news and apps, supporting billions of queries per day. Microsoft uses NVIDIA AI inference for grammar correction in Microsoft Office, supporting half a trillion queries a year and American Express uses it for real-time fraud detection.

We also gain tremendous traction in supercomputing. We announced that NVIDIA technology including Ampere architecture GPUs and HDR InfiniBand networking will power five systems awarded by EuroHPC, a European initiative to build excess scale supercomputing. This includes CINECA, a university consortium in Italy and one of the world's most important supercomputing center, which will use NVIDIA's accelerated computing platform to build the world's fastest AI supercomputer. CINECA supercomputer's name Leonardo advances the age of excess scale AI delivering 10 extra plots of AI performance to enable AI and high performance computing converged applications use cases. It is built with nearing 14,000 NVIDIA Ampere architecture based GPUs and Mellanox, HDR 200-gigabit per second, InfiniBand Networking. And just the released top 500 list of supercomputers show that NVIDIA GPUs or networking powered nearly 70% and eight of the 10 top supercomputers on the list.

Mellanox had another record quarter with double-digit sequential growth, well ahead of our expectations, contributing 13% of overall company revenue. The upside reflected sales to a China OEM that will not recur in Q4. As a result, we expect a meaningful

sequential revenue decline for Mellanox in Q4, but still growing 30% from last year. Mellanox reached record revenue in both InfiniBand and Ethernet, driven by cloud, enterprise and supercomputing customers. Strong demand for high-performance interconnects where Mellanox as leader is being fueled by AI increasingly complex applications which demand faster, smarter, more scalable networks. As the data center becomes the new unit of computing in the age of AI, Mellanox networking is foundational to modern scale out architectures.

At GTC in October, we unveiled the BlueField-2 DPU, our Data Processing Unit, a new kind of processor which offloads critical networking, storage and security tasks from the CPU. A single BlueField-2 DPU can deliver the same data center services that consume up 125 CPU cores. This frees up valuable CPU cores to run a wide range of other enterprise applications. In addition, it enabled zero trust security features to prevent data breaches and cyber attacks and accelerates overall performance. VMware announced that it will offload, accelerate and isolate its industry-leading ESXi Hypervisor with NVIDIA's BlueField-2 DPU, boosting vSphere and data center performance and efficiencies. We also unveiled our three-year DPU roadmap, unified Mellanox's leading network capabilities with NVIDIA's GPUs and the new NVIDIA DOCA or Data Center on a Chip Architecture. Software development kit for building DPU accelerated applications. We believe that over time DPUs will ship a millions of servers unlocking a $10 billion total addressable market. BlueField-2 is sampling now with major hyperscale customers and will be integrated into the enterprise server offerings of major OEMs.

This was our busy period for product launches. Earlier this week at Supercomputing 20, we announced a new double capacity A100 80-gigabyte GPUs and GTX Systems for organizations to build, train, and deploy massive AI models. We also announced the new DGX station A100, a powerful work group server with four A100 GPUs and a massive 320-gigabyte GPU memory for data scientists and AI researchers working in offices, research facilities, labs or at home. All these additions to the NVIDIA Ampere architecture family of products will be available early next year. At SC20, we also announced the next generation NVIDIA Mellanox 400-gigabit per second InfiniBand architecture giving AI developers and scientific researchers the fastest available networking performance. This double data throughput and as new in network computing engines to provide additional acceleration. Solutions based on this new architecture are expected to sample in the second quarter of calendar 2021.

Moving to the rest of the P&L. Q3 GAAP gross margin was 62.6% and non-GAAP gross margin was 65.5%. GAAP gross margin declined year-on-year primarily due to charges related to the Mellanox acquisition, partially offset by product mix. The sequential increase was driven by the absence of non-recurring inventory step-up expense related to the Mellanox acquisition. Non-GAAP gross margins increased by 140 basis points year-on-year, reflecting a shift in product mix with higher data to our sales, including the contribution from Mellanox. Non-GAAP gross margin was down 50 basis points sequentially, in line with our expectations, driven by product mix. Q3 GAAP operating expenses were $1.56 billion and non-GAAP operating expenses were $1.1 billion, up 6% and 42% from a year ago, respectively. Q3 GAAP EPS was $2.12, up 46% from a year earlier and non-GAAP EPS was $2.91, up 63% from year ago. Q3 Cash flow from operations was $1.28 billion.

With that let me turn to the outlook for the fourth quarter of fiscal 2021. As a reminder, Q4 includes a 14th week, which we expect to be incrementally addition to revenue and operating expenses. We expect Gaming to be up sequentially in what is typically a seasonally down quarter as we continue to ramp up our new RTX 30 series products. We expect Data Center to be down slightly versus Q3. With that, we expect computing products to grow in the mid-single digits sequentially, more than offset by sequential decline in Mellanox. We expect continued sequential growth in Auto and ProVis though not yet returning to year-on-year growth and we expect a seasonal decline in OEM.

Revenue is expected to be $4.8 billion plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 62.8% and 65.5%, respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $1.64 billion and $1.18 billion, respectively. GAAP and non-GAAP other income and expenses are both expected to be an 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 $300 million to $325 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 will be virtually attending the Credit Suisse Technology Conference on November 30, Wells Fargo TMT Summit, December 1, and the UBS TMT Conference on December 7. Our earnings call to discuss the fourth quarter and full year results is scheduled for Wednesday, February 24.

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

Questions And Answers

Operator

Certainly. (Operator Instructions) Your first question comes from the line of John Pitzer from Credit Suisse. Your line is open.

Q - John Pitzer {BIO 1541792 <GO>}

Can you hear me?

Operator

Yes.

Q - John Pitzer {BIO 1541792 <GO>}

Yeah, hey guys. Congratulations on solid results. Thank you for taking my questions. Just, Colette, going back to your commentary around Mellanox, it seems like you're guiding the January quarter to about $500 million, which means the core data center business is still growing nicely, call it 6%, 7% sequentially. I'm just kind of curious when you look at the core data center business, I know there's not a direct correlation to server business

but we're clearly going through a cloud digestion in server and core vertical markets, enterprise for servers are weak. When you look at your core data center business, do you feel as though that's having an impact and this is sort of the digestion that you saw kind of in late fiscal '20 -- sorry fiscal '19 into '20, but you're doing it, still growing significantly year-over-year ro how would you characterize the macro backdrop?

A - Colette Kress {BIO 18297352 <GO>}

Sure. Let me -- clarify for those also on the call. Yes. We expect our data center revenue in total to be down slightly quarter-over-quarter. The computing products, NVIDIA computing products is expected to grow in the mid-single digit quarter-over-quarter, as we continue the NVIDIA AI adoption and particularly as A100 continues to ramp. Our networking, our Mellanox networking is expected to decline meaningful quarter-over-quarter as sales to that China OEM will not recur in Q4, though, we still expect the results to be growth of 30% or more year-over-year. The timing of some of this business therefore shifted from Q4 to Q3 but overall H2 is quite strong. So, in referring to overall digestion, the hyperscale business remains extremely strong. We expect hyperscale to grow quarter-over-quarter in computing products as A100 continues to ramp. The A100 continues to gain adoption not only across those hyperscale customers, but again, we're also receiving great momentum in inferencing with the A100 and the T4.

I'll turn it over here to Jensen to see if he has more that he would like to add.

A - Jensen Huang {BIO 1782546 <GO>}

Colette, captured it very well. The only thing that I would add is our inference initiative is really gaining great momentum. Inference is one of the hardest computer science problems, compiling these gigantic neural network computational graphs into a target device is really, really has proven to be really, really hard. The models are diverse when you convert vision to language to speech and there are so many different types of models been created, the model sizes are doubling every couple of months. The length of your expectations are increasing all the time or latencies decreasing all the time.

And so the pressure on inference is really great, the technology pressures is really great. And our leadership there is really pulling ahead. We're in our seventh generation TensorRT. We over the course of last couple of years developed an inference server, it's called Triton, has been adopted all over the place. We have several hundred customers now using NVIDIA AI to deploy their AI services. This is some of the early innings. And I think this is going to be our largest near-term growth opportunity. We are really firing on all cylinders there between A100s ramping in the cloud, A100s beginning to ramp in the enterprise and all of our inference initiatives are really doing great.

Q - John Pitzer {BIO 1541792 <GO>}

Jensen, maybe to follow-on there, just on the vertical markets. Clearly, work from home and COVID this year, kind of presented a headwind to new technology deployments on-prem. I'm kind of curious if we expect sort of an enterprise recovery in general next year, how do you think that will translate into your vertical market strategy? And is there

anything else above and beyond that you can do to help accelerate penetration of AI to that end market?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah, John, that's good point. I mean, it's very clear that the inability to go to work is slowing down the adoption of new technology in some of the verticals. Of course, we're seeing rapid adoption in certain verticals like for example, using AI in healthcare to rapidly discover new vaccines and early detection of outbreaks and robotic applications. So, warehouses, digital retail, last mile delivery, they are seeing just really, really great enthusiasm adopting new AI and robotics technology. But in some of the old -- some of the more traditional industries, the new capabilities and new technologies are slowly to deploy. One of the areas that I'm really super excited about is the work that we're doing in remote work and making it possible for people to collaborate remotely. We have a platform called Omniverse, it's, an early data. The feedback from marketplace has been really great. And so, I've got a lot more to report to you guys in the upcoming months around Omniverse.

And so but anyways, I think when the industry recover, we serve -- our fundamental purpose as a company is to solve the greatest challenges that impact industry where ordinary computers can't. And these challenges are -- serve some of the most important applications in the verticals that we address. And they're not commodity applications, they are really impactful, needle moving applications. So, I have every confidence when the industries recover, things will get designed, cars will be designed and planes will be designed and ships will be designed and buildings would be designed in and we're going to see a lot of design and can see a lot of simulation, we are going to see a lot of robotics applications.

Operator

(Operator Instructions) Your next question comes from the line of C.J. Muse from Evercore. Your line is open.

Q - C.J. Muse

Yeah, good afternoon. Thank you for taking the question. You talked about in your prepared remarks limited availability of capacity components. You suggested perhaps a few months to catch up. Curious if you can speak to the visibility that you have for both Gaming and Data Center into your April quarter?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Colette, you want me to take that real quick and then you can help me out.

A - Colette Kress {BIO 18297352 <GO>}

Yes, absolutely.

A - Jensen Huang {BIO 1782546 <GO>}

So, C.J., first of all, we have a lot of visibility into the channel as you know, especially for gaming. And we know how many weeks of inventory is in what parts of the channel. We've been draining down the channel inventory opportunity for some time and meanwhile, we've also expect a very, very successful launch with Ampere. And even, even with our bullish demand expectations and all of the anchors that we built, which is one of the fastest ramps ever. The demand is overwhelming. And I guess in a lot of ways, it's kind of expected. The circumstances are -- it's been a decade since we've invented a new type of computer graphics. Until years ago we invented a programmable shader and it set the industry on the course to create a type of images that we see today.

But it's very clear that the future is going to look something much, much more beautiful and we invented and NVIDIA RTX to do that and has two capabilities; one, based on ray tracing and the other one is based on artificial intelligence in its generation. The combination of those two capabilities is creating images that people are pretty excited about. And at this point it's defined the next generation content. And so when we -- it took us 10 years to invent it. We launched it two years ago. And took our second-generation to really achieve the level of quality and performance that the industry -- that the customers really, really expect. And now -- now the demand is just overwhelming. And so we're going to continue to ramp fast and this is going to be one of our most successful ramps ever. And it gives our installed base of some 200 million plus GeForce gamers the best reason to upgrade in over a decade. And so this is going to be a very large generation (inaudible) is my guess.

And then with respect to data center, they are ramping into A100. A100 is our first generation of GPUs that does several things at the same time. It's universal, we positioned it as a universal because it's able to do all of the applications that we in the past had to have multiple GPUs to do it. It does training well, it does inference incredibly well. It does high performance computing. It does data analytics and so it's abled -- the Ampure architecture is able to do all of this at the same time. And so the utilization for data centers is -- and the utility is really, really fantastic and the reception has been great. And so we're going to ramp into all of the world's cloud. I think starting this quarter, we're now in every, every major cloud provider in the world, including Alibaba, Oracle and of course the giants, the Amazons, the Azure, and the Google One. And we're going to continue to ramp into that. And then of course, we're starting to ramp into enterprise which in my estimation, long-term, will still be the largest growth opportunity for us, turning every industry into an AI, turning every company into AI and augmented with AI and (inaudible) the iPhone moment to all of the world's largest industries. And so we're ramping into that and we are seeing great deal of our enthusiasm.

Operator

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

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi guys, thanks for taking my question. You said that the extra week was contributing incrementally to revenue and OpEx. Can you give us some feeling for how much it's contributing to revenue and OpEx in Q4? And does that impact, at least on the revenue

side differ say between gaming and data center? And then how should we think about it impacting seasonality in the Q1, as that extra week close off?

A - Colette Kress {BIO 18297352 <GO>}

Sure, let me try this one Jensen. Yes, we did corporate about 14th week into our guidance for both revenue and OpEx. We will likely have incrementally positive impact on revenue, although it is tough to quantify. Okay. Our outlook also reflects incremental OpEx for Q4 in primarily two different areas in terms of compensation and depreciation. And given that our employees are such a material powered of our OpEx, it can be close to one-fourteenth of the quarter. Now, when we look a little bit farther, we should think about the incremental positive in both gaming and data center from that extra week as we are hopefully will be extra supply. But not likely as much as one-fourteenth of the quarter of revenue as enterprise demand is essentially project-based and gaming demand though is tied to the number of gamers that might be shopping for the overall holiday. So again, still very hard for us to determine at this time. Normally between Q4 and Q1, there is seasonality in gaming, big seasonality downward, but we'll just have to see as we are still supply constrained within this Q4 to see what that looks like. From an OpEx standpoint, we will probably expect our OpEx to be relatively flattish as we move from Q4 to Q1.

Operator

Your next question comes from the line of Vivek Arya from Bank of America. Your line is open.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question and congratulations on the strong growth. Jensen, my question is on competition from internally designed products by some of your larger cloud customers, Amazon and Google One and others. We hear about competition from time to time and I wanted to get your perspective. Is this a manageable risk? Is the right way to think that they are perhaps using more of your product in the public cloud but they are moving to internal products for internal workloads? Just how should we think about this risk going forward? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Thanks, Vivek. Most of the cloud vendors, in fact, I believe all of the cloud vendors use the same infrastructures largely for their internal cloud and external clouds or have the ability to or largely due. And there is a -- the competition we want to be really good and the reason for that is this. It's just that acceleration makes it very clear that acceleration is the path forward for training and an influence. The vast majority of the world's training models are doubling in size of a couple of months. And it's one of the reasons why our demand is so great. The second is, is inference. The vast majority of the world's inference is done on CPUs and nothing is better than the whole world recognizing that the best way forward is to do inference on accelerators and when that happens, our accelerators are the most versatile, it has the highest performance, we move the fastest. Our rate of innovation is the, is the fastest, because we're also the most dedicated towards, we are most committed towards, we have the largest team in the world to it. Our stack is the most advanced giving us the greatest (inaudible) and performance.

And so we see a sponsor of announcement here now. But they're also our largest customers. And as you know that we're ramping quite nicely and ramping quite nicely in Amazon and Microsoft and Oracle and others. And so I think the big takeaway is that the great opportunity for us to be, if you look at the vast amount of workload, AI workload in the world, the vast majority of it today is still on CPUs and it's very clear now that this is going to be an accelerator workload and we are the best accelerator in the world. And this is going to be a really big growth opportunity for us in the near term. In fact, we believe, it's our largest growth opportunity in near term and we are in the early innings of it.

Operator

Your next question comes from the line of Harlan Sur from JPMorgan. Your line is open.

Q - Harlan Sur {BIO 6539622 <GO>}

Good afternoon and thanks for taking my question and great job on the quarterly execution. The Mellanox networking connectivity business was up 80% year-over-year. I think it was up about 13%, 14% sequentially. I know there was upside in October from one China customer, but it did grow 70% year-over-year last quarter, and you're still expecting 30% year-over-year growth next quarter. If I remember correctly, I think InfiniBand is about 40% of that business; Ethernet cloud is about 60%. Jensen, what are the big drivers especially since we're in the midst of cloud spending digestion cycle? And I just thought that the team announced their next-gen 400-gig InfiniBand solution, which should drive another strong adoption cycle like with your supercomputer customers, when does this upgrade cycle start to fire?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Let's see. Our data center business consists of of supercomputing centers, which is small, high performance computing, which is a much larger part of supercomputing, much larger than supercomputing, and then hyperscale and enterprise which -- which about 50/50. Of the data center business, the accelerated computing part is not very much associated with digestion and others, it's much more associated with workloads and our new product cycles, the TCO that we bring in our inference, the type of models that the cloud service providers are deploying whether they're -- whether they are deploying new AI models based on deep learning and how much of that we -- how much of those workloads that we've -- we've completed according to our accelerators and the readiness for deployment. And so those are the factors associated with accelerated computing. It's really about the apps, it's really about the workloads, really driven by AI.

On the other hand, the networking part of our business is more connected to CPU business because they are much more broad based. The networking part of our business is driven by this idea of new hyperscale data center architecture called disaggregation, software disaggregation not necessarily hardware disaggregation, software disaggregation, where this type of software (inaudible) orchestrate microservices that are deployed across the data center. So, one service, one application is a monolithic running on one computer anymore, it's distributed across multiple computers and multiple nodes, so that the hyperscale data centers can more easily scale up and scale out according to

the workloads and according to the demand on the data center. And so this aggregation has caused the networking between the compute nodes to be of all vital importance and because (inaudible) is the lowest latency, highest performance, highest bandwidth network that you can get. The TCO benefit at the data center scale is really fantastic. And so when they are building our data centers, (inaudible) is going to be much more, much more connected to that.

In the enterprise, depending on new CPU cycles, it could affect them. If a CPU cycle were to delay a little bit, it would affect them by quarter, they were deploying by quarter, it would affect them by calling in the quarter. And so those are kind of the dynamics of it. I think the net-net of it is that it's a forgone conclusion at this point that, that AI is is going to be the future of the way software is written. AI is the most powerful technology force of our time. And acceleration is the best path forward. And so that's what drives our computing business. On the networking business has everything to do with the way architecture of data centers, cloud data centers which is not architected with microservices now. And that's what foundationally drives our networking business demand. And so we're really well positioned in these two fundamental dynamics because as we know, AI is the future and cloud computing is the future. Both of those dynamics are very favorable to us.

Operator

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

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. I wanted to ask a question that was asked before in a bit of different way. If I look at the core business excluding Mellanox, the core data center business, it was up about 6% sequentially the past few quarters and your guidance sort of implies up about that much again in January, which is certainly good and it's in cloud digestion. But of course, you have Ampere still ramping as well which should be a pretty good tailwind. So, there seems to be some offsetting factors. So, I guess I wonder if you feel like your core data center revenue is still being constrained right now by some market digestion and kind of how you sort of balance or handicap these two factors? Thanks.

A - Jensen Huang {BIO 1782546 <GO>}

Our growth is in the near term is more affected by the cycle time of manufacturing and flexibility of supply. We are in a good shape to and all of our supply is informs our guidance. But we would appreciate shorter cycle times, we would appreciate more agile on supply chains. But the world is constrained at the moment. And so, so we just have to make the best of it. That even in that condition -- even in that condition, we've -- all of that is working for our guidance and we expect to grow.

Operator

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

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah, thanks for taking the question. And also congratulations on the quarter. I wanted to go back to kind of the Mellanox question. I know prior to the acquisition Mellanox was growing maybe in the mid-to-high 20% range. These last two quarters, it's grown over 75%. I guess the simple question is how do you think about the growth rate for Mellanox going forward? And that topic we've started to hear you talk more about BlueField and data processing units. I think in your commentary you alluded to server OEM design wins incorporating these DPUs. What are you looking at or when should we think about the DPU business really starting to inflect and become a material driver for the business? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Long term, long-term every computer in the world will be built like a data center. And every node of a data center is going to be a data center in itself. And the reason for that is because we want the apex surface to be basically zero. And today, most of the data centers are all protected as a periphery. But in the future if you would like cloud computing to be the architecture for everything and every data center is multi-tenant, every center is secure, then you're going to have to secure every single node. And each one of those nodes are going to be -- have software defined networking, software defined storage and it's going to have per application security. And so the processing that is -- that it will need to offload the CPU is really quite significant. In fact, we believe that somewhere between 20% to 40% of today's data centers -- cloud data centers is the capacity, the throughput the computational load is consumed running basically the infrastructure overhead. And that's what the DPUs intended was designed to do. We're going to, we're going to offload out number one and number two, we are going to make every single application secure and companies, zero trust computing. I will become a reality.

And so the important really quite tremendous. And I believe therefore that every single server in the world will have a DPU inside somewhere, just because we care so much about security and just because we care so much about throughput in TCL. It is really -- it's really the most cost-effective way of getting the data center. And so I expected our DP business to be quite large. And so that's the reason why we're putting so much energy into it. It's a programmable data center on a chip, data center infrastructure on chip. It is the reason why we're working with VMware on taking the system with the data center. The software defined system data center and putting it on BlueField. And so this is a very important initiative for us. I'm pretty excited about it as you can imagine.

Operator

Your next question comes from the line of Ambrish Srivastava from BMO Capital Markets. Your line is open.

Q - Ambrish Srivastava {BIO 4109276 <GO>}

Thank you very much. Colette and I apologize if I missed it, but for Mellanox, do you expect it to get back to that growth trajectory on a sequential basis in the April quarter?

And I'm assuming that the shortfall in the current quarter is from a pulling from from all way?

A - Colette Kress {BIO 18297352 <GO>}

So, our impact to our Q4 guidance for Mellanox is impacted by a sale to China OEM for Mellanox that will not recur end-Q4. As we look forward into Q1 of April, we're going to take this a quarter at a time and provide thoughts on guidance for that once we turn the corner to the new fiscal year.

Q - Ambrish Srivastava {BIO 4109276 <GO>}

At the highest level cloud, I think the, it's safe to say that high-speed networking is going to be one of the most important things in cloud data centers as we go forward. And the vast majority of the world's data center is still built for the traditional hyper-converged architecture, which is all moving over to microservices based disaggregate software defined disaggregated architectures and that journey is still in its early days. And so I fully expect future cloud data centers -- all future data centres are going to be connected with high-speed networking inside, they call it East West traffic and all of the traffic will be secured. And so imagine building firewalls into every single server and imagine every single transaction, every single transmission inside the data center to be high-speed and fully encrypted. And so pretty amazing amount of computation is going to have to be installed into future data centers. But that's an accepted requirement now. And I think I think our networking business of Mellanox is in the early innings of growth.

Operator

Your final question today comes from the line of William Stein from Truist Securities. Your line is open.

Q - William Stein {BIO 15106707 <GO>}

Great, thanks for taking my question. You've given us some pieces of this puzzle. But I'm hoping maybe you can address directly the sort of SKU by SKU roll out of Ampere. We know that we didn't have a ton of SKUs, last quarter, there were more in this quarter that you announced and now you're doing sort of this refresh, it sounds like with double the memory on the A100. Is the T4 going to be refreshed? And if so, when does that happen? And are there other either systems or chips that are still waiting for the Ampere refresh that could potentially contribute to an extended cycle as we look at the next year.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah, in terms of the total number of SKUs that we've ramped of Ampere were probably somewhere along a third to a half of this -- at the SKUs at this point, maybe a little bit less. Yeah, it's less. The way that you can think though it. You could reverse engineer it is like this. You know what our gaming lineup looks like for desktops. And so, traditionally we try to have a new architecture in every single segment. And we have not, we've not gone below -- gone below 4.99 yet. And so, so there is a very big part of the marketplace that we're still in the process of addressing. And then the second thing is laptops. None of those, none of the Ampere architecture has launch for laptops. And then there's

workstations, do the same thing with desktops and workstations and laptops workstations. and none of those have gone on and then there's data center. And our data center business for cloud, you've seen some of the early versions of it, A100 but then there is cloud computing for graphics, there's cloud gaming, there's enterprise, edge enterprise applications, enterprise data analytics applications and so there is a fair, fair number of exciting new products we still have in progress.

Operator

That concludes our Q&A for today. I now turn the call back to Ms. Jankowski for closing remarks.

A - Simona Jankowski {BIO 7131672 <GO>}

Actually, that would be for Jensen.

Operator

My apologies.

A - Jensen Huang {BIO 1782546 <GO>}

Okay. Thank you, Simona. This was a terrific quarter. NVIDIA is firing on all cylinders, NVIDIA RTX has reinvented graphics and has made real-time ray tracing the standard of next generation content. Creating the best ever reason to upgrade for hundreds of millions of NVIDIA gamers. AI where software write software, no humans can is the most powerful technology force of our time and is impacting every industry. NVIDIA AI against swept and now curve training and now inference as well, extending our leadership in this important new way of doing computing.

NVIDIA AI's new Triton influence server, a platform that -- is that, I will speak a lot more about in the future. And a lot more frequently because of it's important and our full stack optimize platform are gaining rapid adoption to operate many of the world's most popular AI enhanced services, opening a major growth opportunity. Data centers are the new unit of computing. Someday we believe there will be millions of autonomous data centers distributed all over the globe. NVIDIA's BlueField DPU programmable data center on a chip in our rich software stack will help place AI data centers in factories, warehouses, 5G base stations and even on wheels. And with our pending acquisition of Arm, the company that builds the most -- the world's most popular CPU we will create the computing company for the age of AI with computing extending from the cloud to trillions of devices.

Thank you for joining us today and wish all of you happy holidays and please do stay safe and I look forward to seeing you guys next call.

Operator

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