NVIDIA Corporation (NASDAQ:NVDA) Q3 2023 Results Conference Call November 16, 2022 5:00 PM ET

Company Participants

Simona Jankowski - IR

Jensen Huang - President and CEO

Colette Kress - EVP and CFO

Conference Call Participants

Vivek Arya - Bank of America Securities

C.J. Muse - Evercore

Timothy Arcuri - UBS

Stacy Rasgon - Bernstein

Mark Lipacis - Jefferies

Harlan Sur - JPMorgan

Aaron Rakers - Wells Fargo

Ambrish Srivastava - BMO

William Stein - Truist Securities

Matt Ramsay - Cowen

Joseph Moore - Morgan Stanley

Toshiya Hari - Goldman Sachs

Operator

Good afternoon. My name is Emma, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Third Quarter Earnings 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]

Simona Jankowski, you may begin your conference.

Simona Jankowski

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2023. With me 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 and fiscal 2023. 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 16, 2022, 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

Thanks, Simona. Q3 revenue was $5.93 billion, down 12% sequentially and down 17% year-on-year. We delivered record data center and automotive revenue, while our gaming and pro visualization platforms declined as we work through channel inventory corrections and challenging external conditions.

Starting with data center. Revenue of $3.83 billion was up 1% sequentially and 31% year-on-year. This reflects very solid performance in the face of macroeconomic challenges, new export controls and lingering supply chain disruptions. Year-on-year growth was driven primarily by leading U.S. cloud providers and a broadening set of

consumer internet companies for workloads such as large language models, recommendation systems and generative AI. As the number and scale of public cloud computing and internet service companies deploying NVIDIA AI grows our traditional hyperscale definition will need to be expanded to convey the different end market use cases. We will align our data center customer commentary going forward accordingly. Other vertical industries, such as automotive and energy, also contributed to growth with key workloads relating to autonomous driving, high-performance computing, simulations and analytics.

During the quarter, the U.S. government announced new restrictions impacting exports of our A100 and H100 based products to China, and any product destined for certain systems or entities in China. These restrictions impacted third quarter revenue, largely offset by sales of alternative products into China. That said, demand in China more broadly remains soft, and we expect that to continue in the current quarter.

We started shipping our flagship H100 data center GPU based on the new Hopper architecture in Q3. H100-based systems are available starting this month from leading server makers including Dell, Hewlett Packard Enterprise, Lenovo and Supermicro. Early next year, the first H100 based cloud instances will be available on Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure. H100 delivered the highest performance and workload versatility for both, AI training and inference in the latest MLPerf industry benchmarks. H100 also delivers incredible value compared to the previous generation for equivalent AI performance it offers 3x lower total cost of ownership while using 5x fewer server nodes and 3.5x less energy.

Earlier today, we announced a multiyear collaboration with Microsoft to build an advanced cloud-based AI supercomputer to help enterprises train, deploy and scale AI including large state-of-the-art models. Microsoft Azure will incorporate our complete AI stack, adding tens and thousands of A100 and H100 GPUs, Quantum-2 400 gigabit per second InfiniBand networking and the NVIDIA AI enterprise software suite to its platform.

Oracle and NVIDIA are also working together to offer AI training and inference at scale to thousands of enterprises. This includes bringing to Oracle cloud infrastructure the full NVIDIA accelerated computing stack and adding tens of thousands of NVIDIA GPUs, including the A100 and H100. Cloud-based high-performance [Technical Difficulty] is adopting NVIDIA AI enterprise and other software to address the industrial scientific communities' rising demand for AI in the cloud.

NVIDIA AI will bring new capability to rescale high-performance computing as a service offerings, which include simulation and engineering software used across industries. Networking posted strong growth driven by hyperscale customers and easing supply constraints. Our new Quantum-2 40 gigabit per second InfiniBand and Spectrum Ethernet networking platforms are building momentum.

We achieved an important milestone this quarter with VMware, whose leading server virtualization platform, vSphere, has been rearchitected over the last two years to run on DPUs and now supports our BlueField DPUs. Our joint enterprise AI platform is available first on Dell PowerEdge servers. The BlueField DPU design win pipeline is growing and the number of infrastructure software partners is expanding, including Arista, Check Point, Juniper, Palo Alto Networks and Red Hat.

The latest top 500 list of supercomputers released this week at Supercomputing '22 has the highest ever number of NVIDIA-powered systems, including 72% of the total and 90% of new systems on the list. Moreover, NVIDIA powers 23 of the top 30 of the Green500 list, demonstrating the energy efficiency of accelerated computing.

The number one most energy-efficient system is the Flatiron Institute's Henry, which is the first top 500 system featuring our H100 GPUs. At GTC, we announced the NVIDIA Omniverse Computing System, or OVX, reference designed featuring the new L40 GPU based on the Ada Lovelace architecture. These systems are designed to build and operate 3D virtual worlds using NVIDIA Omniverse Enterprise. NVIDIA OVX systems will be available from Inspur, Lenovo and Supermicro by early 2023. Lockheed Martin and Jaguar Land Rover will be among the first customers to receive OVX systems.

We are further expanding our AI software and services offerings with NVIDIA and BioNeMo large language model services, which are both entering early access this month. These enable developers to easily adopt large language models and deploy customized AI applications for content generation, tech summarization, chatbox, co-development, protein structure and biomolecular property predictions.

Moving to gaming. Revenue of $1.57 billion was down 23% sequentially and down 51% from a year ago, reflecting lower sell-in to partners to help align channel inventory levels with current demand expectations. We believe Channel inventories are on track to approach normal levels as we exit Q4.

Sell-through for our gaming products was relatively solid in the Americas and EMEA, but softer in Asia Pac as macroeconomic conditions and COVID lockdowns in China continued to weigh on consumer demand. Our new Ada Lovelace GPU architecture had

an exceptional launch. The first Ada GPU, the GeForce RTX 4090 became available in mid-October and a tremendous amount and positive feedback from the gaming community. We sold out quickly in many locations and are working hard to keep up with demand. The next member of the Ada family, RTX 4080 is available today.

The RTX 40 series GPUs features DLSS 3, the neural rendering technology that uses AI to generate entire frames for faster game play. Our third-generation RTX technology has raised the bar for computer graphics and helped supercharge gaming. For example, the 15-year old classic game Portal -- now reimagined with full ray tracing and DLSS 3 has made it on Steam's top 100 most wish-listed gains. The total number of RTX games and applications now exceeds 350.

There is tremendous energy in the gaming community that we believe will continue to fuel strong fundamentals over the long term. The number of simultaneous users on Steam just hit a record of 30 million, surpassing the prior peak of 28 million in January.

Activision's Call of Duty Modern Warfare 2 set a record for the franchise with more than $800 million in opening weekend sales, topping the combined box office openings of movie blockbusters, Top Gun: Maverick; and Dr. Strange in the Multiverse of Madness. And this month's League of Legends World Championship in San Francisco sold out in minutes with 18,000 Esports fans packed the arena where the Golden State Warriors play.

We continue to expand the GeForce NOW cloud gaming service. In Q3, we added over 85 games to the library, bringing the total to over 1,400. We also launched GeForce NOW on the new gaming devices, including Logitech, Cloud G, handheld, cloud gaming Chromebooks and Razr 5G Edge.

Moving to ProViz. Revenue of $200 million was down 60% sequentially and down 65% from a year ago, reflecting lower sell-in to partners to help align channel inventory levels with the current demand expectations. These dynamics are expected to continue in Q4.

Despite near-term challenges, we believe our long-term opportunity remains intact, fueled by AI, simulation, computationally intensive design and engineering workloads. At GTC, we announced NVIDIA Omniverse Cloud Services, our first software and infrastructure as a service offering, enabling artists, developers and enterprise teams to design, publish and operate metaverse applications from anywhere on any device. Omniverse Cloud Services runs on Omniverse cloud computer, a computing system comprised of NVIDIA OVX for graphics and physics simulation. NVIDIA HDX for AI

workloads and the NVIDIA graphics delivery network, a global scale distributed data center network for delivering low-latency metaverse graphics on the edge.

Leaders in some of the world's largest industries continue to adopt Omniverse. Home improvement retailer, Lowe's is using it to help design, build and operate digital twins for their stores. Charter Communications and advanced analytics company, HEAVY.AI are creating Omniverse power digital twins to optimize Charter's wireless network.

And Deutsche Bahn, operator of German National Railway is using Omniverse to create digital twins of its rail network and train AI models to monitor the network, increasing safety and reliability.

Moving to automotive. Revenue of $251 million increased 14% sequentially and 86% from a year ago. Growth was driven by an increase in AI automotive solutions as our customers' DRIVE Orin-based production ramps continue to scale. Automotive has great momentum and is on its way to be our next multibillion-dollar platform. Volvo Cars unveiled the all-new flagship Volvo EX90 SUV powered by the NVIDIA DRIVE platform. This is the first model to use Volvo's software-defined architecture with a centralized core computer containing both, DRIVE Orin and DRIVE Xavier, along with 30 sensors.

Other recently announced design wins and new model introductions include Hozon Auto, NIO, Polestar and XPeng.

At GTC, we also announced that NVIDIA DRIVE Thor Superchip, the successor to Orin in our automotive SoC roadmap. DRIVE Thor delivers up to 2,000 teraFLOPS of performance and leverages technologies introduced in our Grace Hopper and Ada architectures. It is capable of running both the automated drive and in-vehicle infotainment systems, simultaneously offering a leap of performance while reducing cost and energy consumption.

DRIVE Thor will be available for automakers 2025 models with Geely-owned automaker ZEEKR as the first announced customer.

Moving to the rest of the P&L. GAAP gross margins was 53.6% and non-GAAP gross margins was 56.1%. Gross margins reflect $702 million in inventory charges largely related to lower data center demand in China, partially offset by a warranty benefit of approximately $70 million. Year-on-year, GAAP operating expenses were up 31%, and non-GAAP operating expenses were up 30%, primarily due to higher compensation expenses related to headcount growth and salary increases and higher data center infrastructure expenses. Sequentially, both GAAP and non-GAAP operating expense

growth was in the single-digit percent, and we plan to keep it relatively flat at these levels over the coming quarters.

We returned $3.75 billion to shareholders in the form of share repurchases and cash dividends. At the end of Q3, we had approximately $8.3 billion remaining under our share repurchase authorization through December ‘23.

Let me turn to the outlook for the fourth quarter of fiscal 2023. We expect our data center revenue to reflect early production shipments of the H100, offset by continued softness in China. In gaming, we expect to resume sequential growth with our revenue still below end demand as we continue to work through the channel inventory correction. And in automotive, we expect the continued ramp of our Orin design wins. All-in, we expect modest sequential growth driven by automotive, gaming and data center.

Revenue is expected to be $6 billion, plus or minus 2%. GAAP and non-GAAP gross margins are expected to be $63.2 million and 66%, respectively, plus or minus 50 basis points. GAAP operating expenses are expected to be approximately $2.56 billion. Non-GAAP operating expenses are expected to be approximately $1.78 billion. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $40 million, excluding gains and losses on nonaffiliated investments. GAAP and non-GAAP tax rates are expected to be 9%, plus or minus 1%, excluding any discrete items. Capital expenditures are expected to be approximately $500 million to $550 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’ll be attending the Credit Suisse Conference in Phoenix on November 30th, the Arete Virtual Tech Conference on December 5th, and the JPMorgan Forum on January 5th in Las Vegas. Our earnings call to discuss the results of our fourth quarter and fiscal 2023 are scheduled for Wednesday, February 22.

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

Question-and-Answer Session

Operator

Thank you. [Operator Instructions] Your first question comes from the line of Vivek Arya with Bank of America Securities.

Vivek Arya

Colette, I just wanted to clarify first. I think last quarter, you gave us a sell-through rate for your gaming business at about $2.5 billion a quarter. I think you said China is somewhat weaker. So, I was hoping you could update us on what that sell-through rate is right now for gaming. And then, Jensen, the question for you. A lot of concerns about large hyperscalers cutting their spending and pointing to a slowdown. So if, let's say, U.S. cloud CapEx is flat or slightly down next year, do you think your business can still grow in the data center and why?

Colette Kress

Yes. Thanks for the question. Let me first start with the sell-through on our gaming business. We had indicated, if you put two quarters together, we would see approximately $5 billion in normalized sell-through for our business. Now, during the quarter, sell-through in Q3 was relatively solid. We've indicated that although China lockdowns continue to channel -- excuse me, challenge our overall China business, it was still relatively solid. Notebook sell-through was also quite solid and desktop, a bit softer, particularly in that China and Asia areas. We expect though stronger end demand, though, as we enter into Q4 driven by the upcoming holidays as well as the continuation of the Ada adoption

Jensen Huang

Vivek, our data center business is indexed to two fundamental dynamics. The first has to do with general purpose computing no longer scaling. And so, acceleration is necessary to achieve the necessary level of cost efficiency scale and energy efficiency scale, so that we can continue to increase workloads while saving money and saving power. Accelerated computing is recognized generally as the path forward as general purpose computing slows. The second dynamic is AI. And we're seeing surging demand in some very important sectors of AIs and important breakthroughs in AI. One is called deep recommender systems, which is quite essential now to the best content or item or product to recommend to somebody who's using a device that is like a cell phone or interacting with a computer just using voice. You need to really understand the nature, the context of the person making the request and make the appropriate recommendation to them.

The second has to do with large language models. This is -- this started several years ago with the invention of the Transformer, which led to BERT, which led to GPT-3, which led to a whole bunch of other models now associated with that. We now have the ability

to learn representations of languages of all kinds. It could be human language. It could be the language of biology. It could be language of chemistry. And recently, I just saw a breakthrough called genes LM, which is one of the first example of learning the language of human genomes.

The third has to do with generative AI. You know that the first 10 years, we've dedicated ourselves to perception AI. Now, the goal of perception, of course, is to understand context. But the ultimate goal of AI is to make a contribution to create something to generate product -- and this is now the beginning of the era of generative AI. You probably see it all over the place, whether they're generating images or generating videos or generating text of all kinds and the ability to augment our performance to enhance our performance to make productivity enhanced to reduce cost and improve whatever we do with whatever we have to work with, productivity is really more important than ever.

And so, you could see that our company is indexed to two things, both of which are more important than ever, which is power efficiency, cost efficiency and then, of course, productivity. And these things are more important than ever. And my expectation is that we're seeing all the strong demand and surging demand for AI and for these reasons.

Operator

Your next question comes from the line of C.J. Muse with Evercore.

C.J. Muse

You started to bundle on NVIDIA AI enterprise now with the H100. I'm curious if you can talk about how we should think about timing around software monetization? And how we should kind of see this flow through the model, particularly with the focus on the AI Enterprise and Omniverse side of things?

Jensen Huang

Yes. Thanks, CJ. We're making excellent progress in NVIDIA AI Enterprise. In fact, you saw probably that we made several announcements this quarter associated with clouds. You know that NVIDIA has a rich ecosystem. And over the years, our rich ecosystem and our software stack has been integrated into developers and start-ups of all kinds. But more so -- more than ever, we're at the tipping point of clouds. And that's fantastic because if we could get NVIDIA's architecture and our full stack into every single cloud, we could reach more customers more quickly. And this quarter, we announced several initiatives, one -- has several partnerships and collaborations, one

that we announced today, which has to do with Microsoft and our partnership there. It has everything to do with scaling up AI because we have so many start-ups clamoring for large installations of our GPU so that they could do large language model training and building their start-ups and scale out of AI to enterprise and all of the world's internet service providers.

Every company we're talking to would like to have the agility and the scale, flexibility of clouds. And so, over the last year or so, we've been working on moving all of our software stacks to the cloud – all of our platform and software stacks to the cloud. And so today, we announced that Microsoft and ourselves are going to standardize on the NVIDIA stack, for a very large part of the work that we're doing together so that we could take a full stack out to the world's enterprise. That's all software included.

We, a month ago, announced the same -- similar type of partnership with Oracle. You also saw that Rescale, a leader in high-performance computing cloud has integrated NVIDIA AI into their stack. Monite has been integrated into GCP. And we announced recently NeMo large language model and BioNeMo large language model to put NVIDIA software in the cloud. And we also announced Omniverse is now available in the cloud. The goal of all of this is to move the NVIDIA platform full stack software into the cloud, so that we can engage customers much, much more quickly and customers could engage our software. If they would like to use it in the cloud, it's per GPU instance hour; if they would like to utilize our software on-prem, they could do it through software license and so -- license and subscription. And so, in both cases, we now have software available practically everywhere you would like to engage it.

The partners that we work with are super excited about it because NVIDIA's rich ecosystem is global, and this could bring both, new consumption into their cloud for both them and ourselves, but also connect all of these new opportunities to the other APIs and other services that they offer. And so, our software stack is making really great progress.

Operator

Your next question comes from the line of [Chris Caso] (ph) with Credit Suisse.

Unidentified Analyst

Wonder if you could give some more color about the inventory charges you took in the quarter and then internal inventory in general. In the documentation, you talked about that being a portion of inventory on hand plus some purchase obligations. And you also spoke in your prepared remarks that some of this was due to China data centers. So if

you can clarify what was in those charges. And then in general, for your internal inventory, does that still need to be worked down? And what are the implications if that needs to be worked down over the next couple of quarters?

Colette Kress

Thanks for the question, Chris. So, as we highlighted in our prepared remarks, we booked an entry of $702 million for inventory reserves within the quarter. Most of that, primarily, all of it is related to our data center business, just due to the change in expected demand looking forward for China. So, when we look at the data center products, a good portion of this was also the A100, which we wrote down.

Now looking at our inventory that we have on hand and the inventory that has increased, a lot of that is just due to our upcoming architectures coming to market, our Ada architecture, our Hopper architecture and even more in terms of our networking business. We have been building for those architectures to come to market and as such to say. We are always looking at our inventory levels at the end of each quarter for our expected demand going forward. But I think we've done a solid job, at least in this quarter just based on that expectation going forward.

Operator

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

Timothy Arcuri

Colette, can you -- I have a two-part question. First, is there any effect of stockpiling in the data center guidance? I ask because you now have the A800 that is sort of a modified version of the A100 with the lower data transfer rate. So, one could imagine that customers might be stocking that while they can still get it. And I guess the second part of that is related to the inventory charge, can you just go into that a little bit more? Because last quarter, it made sense that you took a charge because revenue was less than you thought, but revenue came in pretty much in line. And it sounded like China was a net neutral. So, is the charge related to just working A100 inventory down faster? Is that what the charges related to?

Colette Kress

Sure. So, let me talk about the first statement that you indicated. Most of our data center business that we see is we're working with customers specifically on their needs to build out accelerated computing and AI. It's just not a business in terms of where

units are being held for that. They're usually four very, very specific products and projects that we see. So, I'm going to answer no, nothing that we can see.

Your second question regarding the inventory provisions. At the end of last quarter, we were beginning to see softness in China. We've always been looking at our needs long term. It's not a statement about the current quarter in inventory, as you can see. It usually takes two or three quarters for us to build product for the future demand. So, that's always a case of the inventory that we are ordering. So now looking at what we've seen in terms of continued lockdowns, continued economy challenges in China, it was time for us to take a hard look of what do we think we'll need for data center going forward, not led to our write-downs.

Operator

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

Stacy Rasgon

Collect, I had a question on the commentary you gave on the sequentials. It kind of sounded like data center maybe had some China softness issues. You said gaming resumed sequential growth. But then you said sequential growth for the company driven by auto, gaming and data center. How can all three of those grow sequentially if the overall guidance is kind of flattish? Are they all just like growing just a little bit, or is one of them actually down? Like how do we think about the segments into Q4 given that commentary?

Colette Kress

Yes. Thanks, Stacy. So, your question is regarding the sequentials from Q3 to our guidance that we provided for Q4. As we are seeing the numbers in terms of our guidance, you're correct, is only growing about $100 million. And we've indicated that three of those platforms will likely grow just a little bit. But our pro visualization business we think is going to be flattish and likely not growing as we're still working on correcting the channel inventory levels, to get to the right amount. It's very difficult to say which will have that increase. But again, we are planning for all three of those different market platforms to grow just a little bit.

Operator

Your next question comes from the line of Mark Lipacis with Jefferies.

Mark Lipacis

Jensen, I think for you, you've articulated a vision for the data center where a solution with an integrated solution set of a CPU, GPU and DPU is deployed for all workloads or most workloads, I think. Could you just give us a sense of -- or talk about where is this vision in the penetration cycle? And maybe talk about Grace -- Grace's importance for realizing that vision, what will Grace deliver versus an off-the-shelf x86, do you have a sense of where Grace will get embraced first or the fastest within that vision? Thank you.

Jensen Huang

Thanks Mark. Grace's data moving capability is off the charts. Grace also is memory coherent to our GPU, which allows our GPU to expand its effective GPU memory, fast GPU memory by a factor of 10. That's not possible without special capabilities that are designed between Hopper and Grace and the architecture of Grace. And so, it was designed -- Grace is designed for very large data processing at very high speeds. Those applications are related to -- for example, data processing is related for recommender systems, which operates on petabytes of live data at a time. It's all hot. It all needs to be fast, so that you can make a recommendation within milliseconds to hundreds of millions of people using your service. It is also quite effective at AI training, machine learning. And so, those kind of applications are really terrific.

We -- Grace, I think I've said before that we will have production samples in Q1, and we're still on track to do that.

Operator

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

Harlan Sur

Your data center networking business, I believe, is driving about $800 million per quarter in sales, very, very strong growth over the past few years, near term, as you guys pointed out, and the team is driving strong NIC and BlueField attached to your own compute solutions like DGX and more partner announcements like VMware. But we also know that networking has pretty large exposure to general purpose cloud and hyperscale compute spending trends. So, what's the visibility and growth outlook for the networking business over the next few quarters?

Jensen Huang

Yes, if I could take that. The -- first, thanks for your question. Our networking, as you know, is heavily indexed to high-performance computing. We're not -- we don't serve the vast majority of commodity networking. All of our networking solutions are very high end, and they're designed for data centers that move a lot of data. Now, if you have a hyperscale data center these days, and you are deploying a large number of AI applications, it is very likely that the network bandwidth that you provision has a substantial implication on the overall throughput of your data center.

So, the small incremental investment they make in high-performance networking translates to billions of dollars of savings frankly in provisioning the service or billions of dollars more throughput, which increases their economics.

And so, these days, with disaggregated and AI application -- AI provisioning and data centers, high-performance networking is really quite fantastic and it pays for itself right away. But that's where we are focused in high-performance networking and provisioning AI services in -- well, the AI applications that we focus on.

You might have noticed that NVIDIA and Microsoft are building one of the largest AI infrastructures in the world. And it is completely powered by NVIDIA's InfiniBand 400 gigabits per second network. And the reason for that is because that network pays for itself instantaneously. The investment that you're going to put into the infrastructure is so significant that if you were to be dragged by slow networks, obviously, the efficiency of the overall infrastructure is not as high. And so, in the places where we focus networking is really quite important.

It goes all the way back to when we first announced the acquisition of Mellanox. I think at the time, they were doing about a few hundred million dollars a quarter, about $400 million a quarter. And now, we're doing what they used to do in the old days, in a year, practically coming up in a quarter. And so, that kind of tells you about the growth of high-performance networking. It is an indexed to overall enterprise and data center spend but it is highly indexed to AI adoption.

Operator

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

Aaron Rakers

I want to expand on the networking question a little bit further. When we look at the Microsoft announcement today, we think about what Meta is doing on the AI footprint that they're deploying. Jensen, can you help us understand like where your InfiniBand

networking sits relative to like traditional data center switching? And maybe kind of build on that, how you're positioning spectrum for in the market, does that compete against a broader set of opportunities in the Ethernet world for AI fabric networking?

Jensen Huang

Yes. Thanks, Aaron. The math is like this. If you're going to spend $20 billion on an infrastructure and the efficiency of that overall data center is improved by 10%, the numbers are huge. And when we do these large language models and recommender systems, the processing is done across the entire data center. And so, we distribute the workload across multiple GPUs, multiple nodes and it runs for a very long time. And so, the importance of the network can be overemphasized. And so, the difference of 10% in overall improvement in efficiency, which is very easy to achieve, the difference between NVIDIA's InfiniBand, the entire software stack with what we call Magnum IO, which allows us to do computing in the network itself, a lot of software is running in the network itself, not just moving data around. We call it in-network computing because a ton of software is done at the edge of the -- within the network itself. We achieved significant differences in overall efficiency. And so, if you're spending billions of dollars on the infrastructure, or even hundreds of millions of dollars on the infrastructure, the difference is really quite profound.

Operator

Your next question comes from the line of Ambrish Srivastava with BMO.

Ambrish Srivastava

I actually had a couple of clarifications. Colette, on the data center side, is it a fair assumption that compute was down Q-over-Q in the reported quarter because the quarter before, Mellanox or the networking business was up as it was called out. And again, you said it grew quarter-over-quarter. So, is that a fair assumption? And then I had a clarification on the USG band. Initially, it was supposed to be a $400 million, really going to what the government was trying to firewall. Is the A800 -- I'm just trying to make sure I understand it. Isn't that against the spirit of what the government is trying to do, i.e., firewall, high-performance compute, or is A800 going to a different set of customers? Thank you.

Colette Kress

Thank you for the question. So, looking at our compute for the quarter is about flattish. Yes, we're seeing also growth, growth in terms of our networking, but you should look at our Q3, compute is about flattish with last quarter.

Jensen Huang

Ambrish, A800, the hardware, the hardware of A800 ensures that it always meets U.S. government's clear test for export control. And it cannot be customer reprogrammed or application reprogrammed to exceed it. It is hardware limited. It is in the hardware that determines A800's capabilities. And so, it meets the clear test in letter and in spirit. We raised the concern about the $400 million of A100s because we were uncertain about whether we could execute, the introduction of A800 to our customers and through our supply chain in time. The company did remarkable feeds to swarm this situation and make sure that our business was not affected and our customers were not affected. But A800 hardware surely ensures that it always meets U.S. government's clear tests for export control. .

Operator

Your next question comes from the line of William Stein with Truist Securities.

William Stein

I'm hoping you can discuss the pace of H100 growth as we progress over the next year. We've gotten a lot of questions as to whether the ramp in this product should look like a sort of traditional product cycle where there's quite a bit of pent-up demand for this significant improved performance product and that there's supply available as well. So, does this rollout sort of look relatively typical from that perspective, or should we expect a more perhaps delayed start of the growth trajectory where we see maybe substantially more growth in, let's say, second half of '23?

Jensen Huang

H100 ramp is different than the A100 ramp in several ways. The first is that the TCO, the cost benefits, the operational cost benefits because of the energy savings because every data center is now power limited, and because of this incredible transformer engine that's designed for the latest AI models. The performance over Ampere is so significant that I -- and because of the pent-up demand for Hopper because of these new models that are -- that I spoke about earlier, deep recommender systems and large language models and generative AI models. Customers are clamoring to ramp Hopper

as quickly as possible, and we are trying to do the same. We are all hands on deck to help the cloud service providers stand up the supercomputers.

Remember, NVIDIA is the only company in the world that produces and ships semi-custom supercomputers in high volume. It's a miracle to ship one supercomputer every three years. It's unheard of to ship supercomputers to every cloud service provider in a quarter. And so, we're working hand in glove with every one of them, and every one of them are racing to stand up Hoppers. We expect them to have Hopper cloud services stood up in Q1. And so, we are expecting to ship some volume -- we're expecting to ship production in Q4, and then we're expecting to ship large volumes in Q1. That's a faster transition than Ampere. And so, it's because of the dynamics that I described.

Operator

Your next question comes from the line of Matt Ramsay with Cowen.

Matt Ramsay

I guess, Colette, I heard in your script that you had you talked about maybe a new way of commenting on or reporting hyperscaler revenue in your data center business. And I wondered if you could maybe give us a little bit more detail about what you're thinking there and what sort of drove the decision? And I guess the derivative of that, Jensen, how -- that decision to talk about the data center business to hyperscalers differently. I mean what does that mean for the business that is just a reflection of where demand is and you're going to break things out differently, or is something changing about the mix of I guess, internal properties versus vertical industry demand within the hyperscale customer base. Thank you.

Colette Kress

Yes. Matt, thanks for the question. Let me clarify a little bit in terms of what we believe we should be looking at when we go forward and discussing our data center business. Our data center business is becoming larger and larger and our customers are complex. And when we talk about hyperscale, we tend to talk about 7, 8 different companies. But the reality is there's a lot of very large companies that we could add to that discussion based on what they're purchasing. Additionally, looking at the cloud, looking at our cloud purchases and what our customers are building for the cloud is an important area to focus on because this is really where our enterprise is, where our research is, where our higher education is also purchasing. So we're trying to look for a better way to describe the color of what we're seeing in the cloud and also give you a better understanding of some of these large installments that we're seeing in the hyperscales.

Jensen Huang

Yes. Let me double click on what Colette just said, which is absolutely right. There are two major dynamics that's happening. First, the adoption of NVIDIA AI in internet service companies around the world, the number and the scale by which they're doing it has grown a lot, internet service companies. And these are internet service companies that offer services, but they're not public cloud computing companies. The second factor has to do with cloud computing. We are now at the tipping point of cloud computing. Almost every enterprise in the world has both a cloud-first and a multi-cloud strategy. It is exactly the reason why all of the announcements that we made this year -- this quarter, this last quarter since GTC about all the new platforms that are now available in the cloud, a CSP, a hyperscaler is both -- are two things to us, therefore, a hyperscaler can be a sell to customer; they are also a sell with partner.

On the public cloud side of their business, because of the richness of NVIDIA's ecosystem because we have so many internet service customers and enterprise customers using NVIDIA's full stack, the public cloud side of their business really enjoys and values the partnership with us and the sell with relationship they have with us. And it's pretty clear now that for all of the hyperscalers, the public cloud side of their business would likely -- would very likely be the vast majority of their overall consumption. And so, because the world's CSPs, the world's public clouds is only at the early innings of their enterprise to -- lifting enterprise to the cloud world, it's very, very clear that the public cloud side of the business is going to be very large. And so, increasingly, our relationship with CSPs, our relationship with hyperscalers will include, of course, continuing to sell to them for internal consumption but very importantly, sell with for the public cloud side.

Operator

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

Joseph Moore

Great. Thank you. Wonder if you could talk to looking backward at the crypto impact. Obviously, that's gone from your numbers now, but do you see any potential for liquidation of GPUs that are in the mining network, any impact going forward? And do you foresee blockchain being an important part of your business at some point down the road?

Jensen Huang

We don't expect to see blockchain being an important part of our business down the road. There is always a resell market. If you look at any of the major resell sites, eBay, for example, there are secondhand graphics cards for sale all the time. And the reason for that is because a 3090 that somebody bought today, is upgraded to a 4090 or 3090 they bought a couple of years ago is upgraded to 4090 today. That 3090 could be sold to somebody and enjoyed if sold at the right price. And so, the volume of -- the availability of secondhand and used graphics cards has always been there. And the inventory is never zero. And when the inventory is larger than usual, like all supply demand, it would likely drift lower price and affect the lower ends of our market.

But my sense is that where we're going right now with Ada is targeting very clearly in the upper range, the top half of our market. And early signs are, and I'm sure you're also seeing it that the Ada launch was a homerun. That 4090 -- we shipped a large volume of 4090s because as you know, we were prepared for it. And yet within minutes, they were sold out around the world. And so, the reception of 4090 and the reception of 4080 today has been off the charts. And that says something about the strength and the health and the vibrancy of the gaming market. So, we're super enthusiastic about the Ada launch. We have many more Ada products to come.

Operator

Your last question today comes from the line of Toshiya Hari with Goldman Sachs.

Toshiya Hari

Great. Thank you so much for squeezing me in. I had two quick ones for Colette. On supply, I think there were some mixed messaging in your remarks. I think you talked about supply being a headwind at one point. And then when you were speaking to the networking business, I think you talked about supply easing. So, I was hoping you can kind of speak to supply if you're caught up to demand at this point. And then secondly, just on stock-based compensation, pretty mundane topic, I realize, but it is -- I think in the quarter, it was about $700 million. It's becoming a bigger piece of your OpEx. So, curious how we should be modeling that going forward. Thank you.

Colette Kress

Sure. When we look at our supply constraints that we have had in the past, each and every quarter, this is getting better. Networking was one of our issues probably a year ago, and it has taken us probably to this quarter, and next quarter to really see our supply improved so that we can support the pipeline that we have for our customers there. Now that's our supply. We've also made a discussion regarding our customers,

supply constraints issues. When setting up a data center, even getting a data center capacity has been very difficult. And therefore, that challenges them in their purchasing decisions as they're still looking for certain parts of that supply chain to come through. So, that hopefully clarifies what we were talking about regarding two areas of supply.

In our stock-based compensation, what we'll see, it's very difficult to predict what our stock-based compensation would be when it arrives. We have provided to our incoming employees but also once a year to our employees, and it's a single date in terms of when that is priced. So, it's difficult to determine. But stock-based compensation is an important part of our employees' compensation and will continue to be. So, we look at it from an overall compensation perspective. So up until now and when we do the focal, we'll see about the same size with a few additions for the reduced level of employee hiring that we have right now.

Operator

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

Jensen Huang

Thanks, everyone. We are quickly adapting to the macro environment, correcting inventory levels, offering alternative products to data center customers in China and keeping our OpEx flat for the next few quarters.

Our new platforms are off to a great start and formed the foundation for our resumed growth. NVIDIA RTX is reinventing 3D graphics with ray tracing and AI. The launch of Ada Lovelace is phenomenal. Gamers waited in long lines around the world, 4290 stocks sold out quickly. Hopper, with its revolutionary Transformer engine is just in time to meet the surging demand for recommender systems, large language models and generative AI. NVIDIA networking is synonymous with the highest data center throughput and enjoying record results. Orin is the world's first computing platform designed for AI-powered autonomous vehicles and robotics, and putting automotive on the road to be our next multibillion-dollar platform.

These computing platforms run NVIDIA AI and NVIDIA Omniverse, software libraries and engines that help the companies build and deploy AI to products and services. NVIDIA's pioneering work in accelerated computing is more vital than ever. Limited by business, general purpose commuting has slowed to a crawl just as AI demands more computing. Scaling through general purchase computing alone is no longer viable, both from a cost or power standpoint. Accelerated computing is the path forward. We look forward to updating you on our progress next quarter.

Operator

This concludes today’s conference call. Thank you for attending. You may disconnect.