S&P Global
Market Intelligence

NVIDIA Corporation NasdaqGS:NVDA

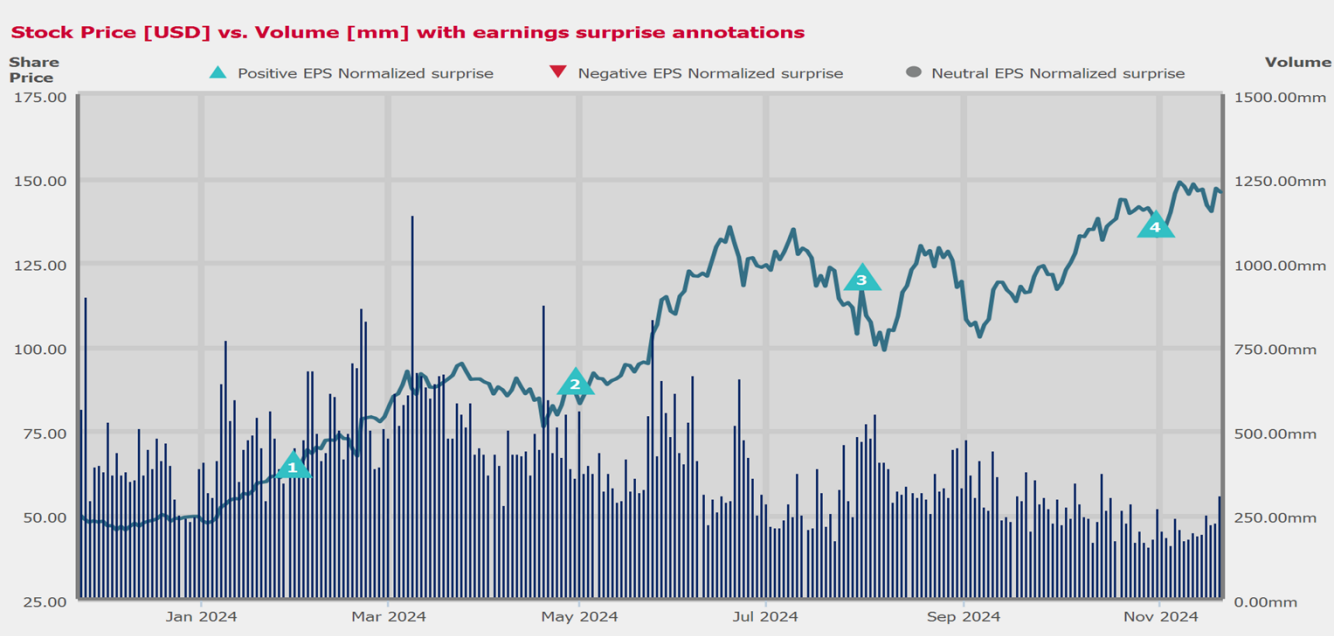
FQ3 2025 Earnings Call Transcripts

Wednesday, November 20, 2024 10:00 PM GMT

S&P Global Market Intelligence Estimates

	-FQ3 2025-			-FQ4 2025-	-FY 2025-	-FY 2026-
	CONSENSUS	ACTUAL	SURPRISE	CONSENSUS	CONSENSUS	CONSENSUS
EPS Normalized	0.75	0.81	8.00	0.81	2.86	4.21
Revenue (mm)	33134.31	35082.00	5.88	37025.37	126300.71	185037.12

Currency: USD
Consensus as of Nov-20-2024 9:52 PM GMT



- EPS NORMALIZED -

	CONSENSUS	ACTUAL	SURPRISE
FQ4 2024	0.46	0.52	1 13.04 %
FQ1 2025	0.56	0.61	2 8.93 %
FQ2 2025	0.64	0.68	3 6.25 %
FQ3 2025	0.75	0.81	4 8.00 %

Table of Contents

Call Participants

EXECUTIVES

Colette M. Kress
Executive VP & CFO

Jensen Huang

Stewart Stecker
Director of Investor Relations

ANALYSTS

Aaron Christopher Rakers
Wells Fargo Securities, LLC, Research Division

Atif Malik
Citigroup Inc., Research Division

Benjamin Alexander Reitzes
Melius Research LLC

Christopher James Muse
Cantor Fitzgerald & Co., Research Division

Joseph Lawrence Moore
Morgan Stanley, Research Division

Pierre C. Ferragu
New Street Research LLP

Stacy Aaron Rasgon
Sanford C. Bernstein & Co., LLC., Research Division

Timothy Michael Arcuri
UBS Investment Bank, Research Division

Toshiya Hari
Goldman Sachs Group, Inc., Research Division

Vivek Arya
BofA Securities, Research Division

Presentation

Operator

Good afternoon. My name is Jail, and I'll be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's third quarter earnings call. [Operator Instructions]

Thank you. Stewart Stecker, you may begin your conference.

Stewart Stecker
Director of Investor Relations

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2025. 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 of fiscal 2025. 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 20, 2024, 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 M. Kress
Executive VP & CFO

Thank you, Stewart. Q3 was another record quarter. We continue to deliver incredible growth. Revenue of $35.1 billion was up 17% sequentially and up 94% year-on-year and well above our outlook of $32.5 billion. All market platforms posted strong sequential and year-over-year growth, fueled by the adoption of NVIDIA accelerated computing and AI.

Starting with Data Center. Another record was achieved in Data Center. Revenue of $30.8 billion, up 17% sequential and up 112% year-on-year. NVIDIA Hopper demand is exceptional, and sequentially, NVIDIA H200 sales increased significantly to double-digit billions, the fastest prod ramp in our company's history. The H200 delivers up to 2x faster inference performance and up to 50% improved TCO. Cloud service providers were approximately half of our Data Center sales with revenue increasing more than 2x year-on-year.

CSPs deployed NVIDIA H200 infrastructure and high-speed networking with installations scaling to tens of thousands of GPUs to grow their business and serve rapidly rising demand for AI training and inference workloads. NVIDIA H200-powered cloud instances are now available from AWS, CoreWeave and Microsoft Azure with Google Cloud and OCI coming soon. Alongside significant growth from our large CSPs, NVIDIA GPU regional cloud revenue jumped 2x year-on-year as North America, India, and Asia Pacific regions ramped NVIDIA Cloud instances and sovereign cloud build-outs.

Consumer Internet revenue more than doubled year-on-year as companies scaled their NVIDIA Hopper infrastructure to support next-generation AI models, training, multimodal and agentic AI, deep learning recommender engines, and generative AI inference and content creation workloads. NVIDIA's Ampere and

Hopper infrastructures are fueling inference revenue growth for customers. NVIDIA is the largest inference platform in the world. Our large installed base and rich software ecosystem encourage developers to optimize for NVIDIA and deliver continued performance and TCO improvements.

Rapid advancements in NVIDIA software algorithms boosted Hopper inference throughput by an incredible 5x in 1 year and cut time to first token by 5x. Our upcoming release of NVIDIA NIM will boost Hopper inference performance by an additional 2.4x. Continuous performance optimizations are a hallmark of NVIDIA and drive increasingly economic returns for the entire NVIDIA installed base.

Blackwell is in full production after a successfully executed mask change. We shipped 13,000 GPU samples to customers in the third quarter, including one of the first Blackwell DGX engineering samples to OpenAI. Blackwell is a full stack, full infrastructure, AI data center scale system with customizable configurations needed to address a diverse and growing AI market from x86 to ARM, training to inferencing GPUs, InfiniBand to Ethernet switches, and NVLink and from liquid cooled to air cooled, every customer is racing to be the first to market.

Blackwell is now in the hands of all of our major partners, and they are working to bring up their data centers. We are integrating Blackwell systems into the diverse data center configurations of our customers. Blackwell demand is staggering, and we are racing to scale supply to meet the incredible demand customers are placing on us. Customers are gearing up to deploy Blackwell at scale. Oracle announced the world's first Zettascale AI cloud computing clusters that can scale to over 131,000 Blackwell GPUs to help enterprises train and deploy some of the most demanding next-generation AI models.

Yesterday, Microsoft announced they will be the first CSP to offer, in private preview, Blackwell-based cloud instances powered by NVIDIA GB200 and Quantum InfiniBand. Last week, Blackwell made its debut on the most recent round of MLPerf training results, sweeping the per GPU benchmarks and delivering a 2.2x leap in performance over Hopper. The results also demonstrate our relentless pursuit to drive down the cost of compute. The 64 Blackwell GPUs are required to run the GPT-3 benchmark compared to 256 H100s or a 4x reduction in cost.

NVIDIA Blackwell architecture with NVLink Switch enables up to 30x faster inference performance and a new level of inference scaling, throughput and response time that is excellent for running new reasoning inference applications like OpenAI's o1 model. With every new platform shift, a wave of start-ups is created. Hundreds of AI-native companies are already delivering AI services with great success.

Through Google, Meta, Microsoft, and OpenAI are the headliners. Anthropic, Perplexity, Mistral, Adobe Firefly, Runway, Midjourney, Lightricks, Harvey, Codeium, Cursor and the Bridge are seeing great success while thousands of AI-native start-ups are building new services. The next wave of AI are enterprise AI and industrial AI. Enterprise AI is in full throttle. NVIDIA AI Enterprise, which includes NVIDIA NeMo and NIM microservices is an operating platform of agentic AI. Industry leaders are using NVIDIA AI to build Copilots and agents.

Working with NVIDIA, Cadence, Cloudera, Cohesity, NetApp, Nutanix, Salesforce, SAP and ServiceNow are racing to accelerate development of these applications with the potential for billions of agents to be deployed in the coming years. Consulting leaders like Accenture and Deloitte are taking NVIDIA AI to the world's enterprises. Accenture launched a new business group with 30,000 professionals trained on NVIDIA AI technology to help facilitate this global build-out.

Additionally, Accenture with over 770,000 employees, is leveraging NVIDIA-powered agentic AI applications internally, including 1 case that cuts manual steps in marketing campaigns by 25% to 35%. Nearly 1,000 companies are using NVIDIA NIM, and the speed of its uptake is evident in NVIDIA AI enterprise monetization. We expect NVIDIA AI enterprise full year revenue to increase over 2x from last year and our pipeline continues to build. Overall, our software, service and support revenue is annualizing at $1.5 billion, and we expect to exit this year annualizing at over $2 billion.

Industrial AI and robotics are accelerating. This is triggered by breakthroughs in physical AI, foundation models that understand the physical world, like NVIDIA NeMo for enterprise AI agents. We built NVIDIA

Omniverse for developers to build, train, and operate industrial AI and robotics. Some of the largest industrial manufacturers in the world are adopting NVIDIA Omniverse to accelerate their businesses, automate their workflows, and to achieve new levels of operating efficiency. Foxconn, the world's largest electronics manufacturer, is using digital twins and industrial AI built on NVIDIA Omniverse to speed the bring-up of its Blackwell factories and drive new levels of efficiency. In its Mexico facility alone, Foxconn expects to reduce -- a reduction of over 30% in annual kilowatt hour usage.

From a geographic perspective, our data center revenue in China grew sequentially due to shipments of export-compliant Hopper products to industries. As a percentage of total data center revenue, it remained well below levels prior to the onset of export controls. We expect the market in China to remain very competitive going forward. We will continue to comply with export controls while serving our customers.

Our sovereign AI initiatives continue to gather momentum as countries embrace NVIDIA accelerated computing for a new industrial revolution powered by AI. India's leading CSPs include product communications and Yotta Data Services are building AI factories for tens of thousands of NVIDIA GPUs. By year-end, they will have boosted NVIDIA GPU deployments in the country by nearly 10x. Infosys, TSE (sic) [TCS], Wipro are adopting NVIDIA AI enterprise and upskilling nearly 0.5 million developers and consultants to help clients build and run AI agents on our platform.

In Japan, SoftBank is building the nation's most powerful AI supercomputer with NVIDIA DGX Blackwell and Quantum InfiniBand. SoftBank is also partnering with NVIDIA to transform the telecommunications network into a distributed AI network with NVIDIA AI Aerial and AI-RAN platform that can process both 5G RAN on AI on CUDA. We are launching the same in the U.S. with T-Mobile. Leaders across Japan, including Fujitsu, NEC and NTT are adopting NVIDIA AI Enterprise and major consulting companies, including EY Strategy and Consulting will help bring NVIDIA AI technology to Japan's industries.

Networking revenue increased 20% year-on-year. Areas of sequential revenue growth include InfiniBand and Ethernet switches, SmartNICs and BlueField DPUs. Though networking revenue was sequentially down, networking demand is strong and growing, and we anticipate sequential growth in Q4. CSPs and supercomputing centers are using and adopting the NVIDIA InfiniBand platform to power new H200 clusters. NVIDIA Spectrum-X Ethernet for AI revenue increased over 3x year-on-year. And our pipeline continues to build with multiple CSPs and consumer Internet companies planning large cluster deployments.

Traditional Ethernet was not designed for AI. NVIDIA Spectrum-X uniquely leverages technology previously exclusive to InfiniBand to enable customers to achieve massive scale of their GPU compute. Utilizing Spectrum-X, xAI's Colossus 100,000 Hopper supercomputer experienced 0 application latency degradation and maintained 95% data throughput versus 60% for traditional Ethernet.

Now moving to Gaming and AI PCs. Gaming revenue of $3.3 billion increased 14% sequentially and 15% year-on-year. Q3 was a great quarter for Gaming with notebook, console, and desktop revenue all growing sequentially and year-on-year. RTX end demand was fueled by strong back-to-school sales as consumers continue to choose GeForce RTX GPUs and devices to power gaming, creative and AI applications.

Channel inventory remains healthy and we are gearing up for the holiday season. We began shipping new GeForce RTX AI PC with up to 321 AI TOPS from ASUS and MSI with Microsoft's Copilot+ capabilities anticipated in Q4. These machines harness the power of RTX ray tracing and AI technologies to supercharge gaming, photo, and video editing, image generation and coding.

This past quarter, we celebrated the 25th anniversary of the GeForce 256, the world's first GPU. The transforming executing graphics to igniting the AI revolution. NVIDIA's GPUs have been the driving force behind some of the most consequential technologies of our time.

Moving to ProViz. Revenue of $486 million was up 7% sequentially and 17% year-on-year. NVIDIA RTX workstations continue to be the preferred choice to power professional graphics, design, and engineering-related workloads. Additionally, AI is emerging as a powerful demand driver, including autonomous vehicle simulation, generative AI model prototyping for productivity-related use cases and generative AI content creation in media and entertainment.

Moving to Automotive. Revenue was a record $449 million, up 30% sequentially and up 72% year-on-year. Strong growth was driven by self-driving brands of NVIDIA Orin and robust end market demand for NAVs. Volvo Cars is rolling out its fully electric SUV built on NVIDIA Orin and DriveOS.

Okay. Moving to the rest of the P&L. GAAP gross margin was 74.6% and non-GAAP gross margin was 75%, down sequentially, primarily driven by a mix shift of the H100 systems to more complex and higher-cost systems within Data Center. Sequentially, GAAP operating expenses and non-GAAP operating expenses were up 9% due to higher compute, infrastructure and engineering development costs for new product introductions.

In Q3, we returned $11.2 billion to shareholders in the form of share repurchases and cash dividends. Well, let me turn to the outlook for the fourth quarter. Total revenue is expected to be $37.5 billion, plus or minus 2%, which incorporates continued demand for Hopper architecture and the initial ramp of our Blackwell products. While demand greatly exceed supply, we are on track to exceed our previous Blackwell revenue estimate of several billion dollars as our visibility into supply continues to increase.

On Gaming, although sell-through was strong in Q3, we expect fourth quarter revenue to decline sequentially due to supply constraints. GAAP and non-GAAP gross margins are expected to be 73% and 73.5%, respectively, plus or minus 50 basis points. Blackwell is a customizable AI infrastructure with 7 different types of NVIDIA-built chips, multiple networking options and for air and liquid-cooled data centers. Our current focus is on ramping to strong demand, increasing system availability, and providing the optimal mix of configurations to our customer.

As Blackwell ramps, we expect gross margins to moderate to the low 70s. When fully ramped, we expect Blackwell margins to be in the mid-70s. GAAP and non-GAAP operating expenses are expected to be approximately $4.8 billion and $3.4 billion, respectively. We are a data center-scale AI infrastructure company. Our investments include building data centers for development of our hardware and software stacks and to support new introductions. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $400 million, excluding gains and losses from nonaffiliated investments. GAAP and non-GAAP tax rates are expected to be 16.5%, plus or minus 1%, excluding any discrete items. Further financial details are included in the CFO commentary and other information available on our IR website.

In closing, let me highlight upcoming events for the financial community. We will be attending the UBS Global Technology and AI Conference on December 3 in Scottsdale. Please join us at CES in Las Vegas, where Jensen will deliver a keynote on January 6. And we will host a Q&A session for financial analysts the next day on January 7.
Our earnings call to discuss results for the fourth quarter of fiscal 2025 is scheduled for February 26, 2025. We will now open the call for questions. Operator, can you poll for questions, please?

Question and Answer

Operator

[Operator Instructions] Your first question comes from the line of C.J. Muse of Cantor Fitzgerald.

Christopher James Muse
Cantor Fitzgerald & Co., Research Division

I guess, Jensen, a question for you on the debate around whether scaling for large language models have stalled. Obviously, we're very early here, but would love to hear your thoughts on this front. How are you helping your customers as they work through these issues? And then obviously, part of the context here is we're discussing clusters that have yet to benefit from Blackwell. So is this driving even greater demand for Blackwell?

Jensen Huang

Foundation model pretraining scaling is intact and it's continuing. As you know, this is an empirical law, not a fundamental physical law. But the evidence is that it continues to scale. What we're learning, however, is that it's not enough, that we've now discovered 2 other ways to scale. One is post-training scaling. Of course, the first generation of post-training was reinforcement learning human feedback, but now we have reinforcement learning AI feedback and all forms of synthetic data generated data that assists in post-training scaling.

And one of the biggest events and one of the most exciting developments is Strawberry, ChatGPT o1, OpenAI's o1, which does inference time scaling, what's called test time scaling. The longer it thinks, the better and higher-quality answer it produces. And it considers approaches like chain of thought and multi-path planning and all kinds of techniques necessary to reflect and so on and so forth. And it's -- intuitively, it's a little bit like us doing thinking in our head before we answer your question.

And so we now have 3 ways of scaling and we're seeing all 3 ways of scaling. And as a result of that, the demand for our infrastructure is really great. You see now that at the tail end of the last generation of foundation models were at about 100,000 Hoppers. The next generation starts at 100,000 Blackwells. And so that kind of gives you a sense of where the industry is moving with respect to pretraining scaling, post-training scaling, and then now very importantly, inference time scaling. And so the demand is really great for all of those reasons.

But remember, simultaneously, we're seeing inference really starting to scale up for our company. We are the largest inference platform in the world today because our installed base is so large. And everything that was trained on Amperes and Hoppers inference incredibly on Amperes and Hoppers. And as we move to Blackwells for training foundation models, it leads behind it a large installed base of extraordinary infrastructure for inference.

And so we're seeing inference demand go up. We're seeing inference time scaling go up. We see the number of AI-native companies continue to grow. And of course, we're starting to see enterprise adoption of agentic AI really is the latest rage. And so we're seeing a lot of demand coming from a lot of different places.

Operator

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

Toshiya Hari
Goldman Sachs Group, Inc., Research Division

Jensen, you executed the mass change earlier this year. There were some reports over the weekend about some heating issues. On the back of this, we've had investors ask about your ability to execute to the road map you presented at GTC this year with Ultra coming out next year and the transition to Rubin in '26.

Can you sort of speak to that? And some investors are questioning that, so if you can sort of speak to your ability to execute on time, that would be super helpful.

And then a quick part B. On supply constraints, is it a multitude of componentry that's causing this or is it specifically CoWoS, HBM? Is it supply constraints? Are the supply constraints getting better? Are they worsening? Any sort of color on that would be super helpful as well.

Jensen Huang

Yes, thanks. So let's see. Back to the first question. Blackwell production is in full steam. In fact, as Colette mentioned earlier, we will deliver this quarter more Blackwells than we had previously estimated. And so the supply chain team is doing an incredible job working with our supply partners to increase Blackwell, and we're going to continue to work hard to increase Blackwell through next year.

It is the case that demand exceeds our supply. And that's expected as we're in the beginnings of this generative AI revolution as we all know. And we're at the beginning of a new generation of foundation models that are able to do reasoning and able to do long thinking. And of course, one of the really exciting areas is physical AI, AI that now understands the structure of the physical world. And so Blackwell demand is very strong. Our execution is going well.

And there's obviously a lot of engineering that we're doing across the world. You see now systems that are being stood up by Dell and CoreWeave. I think you saw systems from Oracle stood up. You have systems from Microsoft, and they're about to preview their Grace Blackwell systems. You have systems that are at Google. And so all of these CSPs are racing to be first. The engineering that we do with them is, as you know, rather complicated.

And the reason for that is because although we build full stack and full infrastructure, we disaggregate all of this AI supercomputer and we integrate it into all of the custom data centers and architectures around the world. That integration process, it's something we've done several generations now. We're very good at it but still there's still a lot of engineering that happens at this point.

But as you see from all of the systems that are being stood up, Blackwell is in great shape. And as we mentioned earlier, the supply and what we're planning to ship this quarter is greater than our previous estimates. With respect to the supply chain, there are 7 different chips, 7 custom chips that we built in order for us to deliver the Blackwell systems. The Blackwell systems go in air-cooled or liquid-cooled, NVLink 8 or NVLink 72 or NVLink 8, NVLink 36, NVLink 72. We have x86 or Grace. And the integration of all of those systems into the world's data centers is nothing short of a miracle. And so the component supply chain necessary to ramp at this scale, you have to go back and take a look at how much Blackwell we shipped last quarter, which was 0. And in terms of how much Blackwell total systems will ship this quarter, which is measured in billions, the ramp is incredible.

And so almost every company in the world seems to be involved in our supply chain. And we've got great partners, everybody from, of course, TSMC and Amphenol, the connector company, incredible company. Vertiv and SK Hynix and Micron, [Spill], Amkor, KYEC, and there's Foxconn and the factories that they've built and Quanta and Wiwynn and, gosh, Dell and HP, and Super Micro, Lenovo and the number of companies is just really quite incredible, Quanta. And I'm sure I've missed partners that are involved in the ramping of Blackwell, which I really appreciate. And so anyways, I think we're in great shape with respect to the Blackwell ramp at this point.

And then lastly, your question about our execution of our road map. We're on an annual road map and we're expecting to continue to execute on our annual road map. And by doing so, we increased the performance, of course, of our platform. But it's also really important to realize that when we're able to increase performance and do so at X factors at a time, we're reducing the cost of training. We're reducing the cost of inferencing. We're reducing the cost of AI so that it could be much more accessible.

But the other factor that's very important to note is that when there's a data center of some fixed size and the data center always is of some fixed size. It could be, of course, tens of megawatts in the past, and now it's -- most data centers are now 100 megawatts to several hundred megawatts, and we're planning on gigawatt data centers, it doesn't really matter how large the data centers are. The power is limited.

And when you're in the power-limited data center, the best -- the highest performance per watt translates directly into the highest revenues for our partners.

And so on the one hand, our annual road map reduces cost. But on the other hand, because our perf per watt is so good compared to anything out there, we generate for our customers the greatest possible revenues. And so that annual rhythm is really important to us, and we have every intentions of continuing to do that. And everything is on track as far as I know.

Operator

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

Timothy Michael Arcuri
UBS Investment Bank, Research Division

I'm wondering if you can talk about the trajectory of how Blackwell is going to ramp this year. I know Jensen, you did just talk about Blackwell being better than I think you had said several billions of dollars in January. It sounds like you're going to do more than that. But I think in recent months also, you said that Blackwell crosses over Hopper in the April quarter. So I guess I had 2 questions.

First of all, is that still the right way to think about it, that Blackwell will cross over Hopper in April? And then Colette, you kind of talked about Blackwell bringing down gross margin to the low 70s as it ramps. So I guess if April is the crossover, is that the worst of the pressure on gross margin? So you're going to be kind of in the low 70s as soon as April. I'm just wondering if you can sort of shape that for us.

Jensen Huang

Colette, why don't you start?

Colette M. Kress
Executive VP & CFO

Sure, let me first start with your question, Tim, thank you, regarding our gross margins. And we discussed that our gross margins as we are ramping Blackwell in the very beginning and the many different configurations, the many different chips that we are bringing to market, we are going to focus on making sure we have the best experience for our customers as they stand that up.

We will start growing into our gross margins but we do believe those will be in the low 70s in that first part of the ramp. So you're correct, as you look at the quarters following after that, we will start increasing our gross margins, and we hope to get to the mid-70s quite quickly as part of that round.

Jensen Huang

Hopper demand will continue through next year, surely the first several quarters of the next year. And meanwhile, we will ship more Blackwells next quarter than this, and we'll ship more Blackwells the quarter after that than our first quarter. And so that kind of puts it in perspective. We are really at the beginnings of 2 fundamental shifts in computing that is really quite significant.

The first is moving from coding that runs on CPUs to machine learning that creates neural networks that runs on GPUs. And that fundamental shift from coding to machine learning is widespread at this point. There are no companies who are not going to do machine learning. And so machine learning is also what enables generative AI. And so on the one hand, the first thing that's happening is $1 trillion worth of computing systems and data centers around the world is now being modernized for machine learning.

On the other hand, secondarily, I guess, is that on top of these systems are going to be -- we're going to be creating a new type of capability called AI. And when we say generative AI, we're essentially saying that these data centers are really AI factories. They're generating something. Just like we generate electricity, we're now going to be generating AI. And if the number of customers is large, just as the number of consumers of electricity is large, these generators are going to be running 24/7. And today, many AI services are running 24/7, just like an AI factory.

And so we're going to see this new type of system come online, and I call it an AI factory because that's really as close to what it is. It's unlike a data center of the past. And so these 2 fundamental trends are really just beginning. And so we expect this to happen, this growth, this modernization and the creation of a new industry to go on for several years.

Operator

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

Vivek Arya
BofA Securities, Research Division

Colette, just to clarify, do you think it's a fair assumption to think NVIDIA could recover to kind of mid-70s gross margin in the back half of calendar '25? Just wanted to clarify that. And then Jensen, my main question, historically, when we have seen hardware deployment cycles, they have inevitably included some digestion along the way. When do you think we get to that phase? Or is it just too premature to discuss that because you're just at the start of Blackwell?

So how many quarters of shipments do you think is required to kind of satisfy this first wave? Can you continue to grow this into calendar '26? Just how should we be prepared to see what we have seen historically, right, a period of digestion along the way of a long-term kind of secular hardware deployment?

Colette M. Kress
Executive VP & CFO

Okay. Vivek, thank you for the question. Let me clarify your question regarding gross margins. Could we reach the mid-70s in the second half of next year? And yes, I think it is a reasonable assumption or goal for us to do, but we'll just have to see how that mix of ramp goes. But yes, it is definitely possible.

Jensen Huang

The way to think through that, Vivek, is I believe that there will be no digestion until we modernize $1 trillion with the data centers. Those -- if you just look at the world's data centers, the vast majority of it is built for a time when we wrote applications by hand and we ran them on CPUs. It's just not a sensible thing to do anymore. If you have -- if every company's CapEx, if they're ready to build data center tomorrow, they ought to build it for a future of machine learning and generative AI because they have plenty of old data centers.

And so what's going to happen over the course of the next X number of years, and let's assume that over the course of 4 years, the world's data centers could be modernized as we grow into IT, as you know, IT continues to grow about 20%, 30% a year, let's say. And so -- but let's say by 2030, the world's data centers for computing is, call it, a couple of trillion dollars. We have to grow into that. We have to modernize the data center from coding to machine learning. That's number one.

The second part of it is generative AI. And we're now producing a new type of capability the world's never known, a new market segment that the world's never had. If you look at OpenAI, it didn't replace anything. It's something that's completely brand new. It's in a lot of ways as when the iPhone came, it was completely brand new. It wasn't really replacing anything. And so we're going to see more and more companies like that. And they're going to create and generate, out of their services, essentially intelligence.

Some of it would be digital artist intelligence like Runway. Some of it would be basic intelligence like OpenAI. Some of it would be legal intelligence like Harvey. Digital marketing intelligence like [Writers], so on and so forth. And the number of these companies, these -- what are they called, AI-native companies, are just in hundreds. And almost every platform shift, there was -- there were Internet companies, as you recall. There were cloud-first companies. There were mobile-first companies. Now they're AI natives.

And so these companies are being created because people see that there's a platform shift, and there's a brand-new opportunity to do something completely new. And so my sense is that we're going to continue

to build out to modernize IT, modernize computing, number one; and then number two, create these AI factories that are going to be for a new industry for the production of artificial intelligence.

Operator

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

Stacy Aaron Rasgon
Sanford C. Bernstein & Co., LLC., Research Division

Colette, I had a clarification and a question for you. The clarification, just when you say low 70s gross margins, does -- is 73.5% count as low 70s or do you have something else in mind? And for my question, you're guiding total revenues, and so I mean, total data center revenues in the next quarter must be up "several billion dollars." But it sounds like Blackwell now should be up more than that.

But you also said Hopper was still strong, so like is Hopper down sequentially next quarter? And if it is, like why? Is it because of the supply constraints? Is -- China has been pretty strong. Is China is kind of rolling off a bit into Q4? So any color you can give us on sort of the Blackwell ramp and the Blackwell versus Hopper behavior into Q4 would be really helpful.

Colette M. Kress
Executive VP & CFO

So first, starting on your first question there, Stacy, regarding our gross margin and define low. Low, of course, is below the mids and let's say, we might be at 71%, maybe about 72%, 72.5%. We're going to be in that range. We could be higher than that as well. We're just going to have to see how it comes through. We do want to make sure that we are ramping and continuing that improvement, the improvement in terms of our yields, the improvement in terms of the product as we go through the rest of the year. So we'll get up to the mid-70s by that point.

The second statement was a question regarding our Hopper and what is our Hopper doing? We have seen substantial growth for H200 not only in terms of orders but the quickness in terms of those that are standing that out. It is an amazing product and it's the fastest growing and ramping that we've seen. We will continue to be selling Hopper in this quarter, in Q4 for sure. That is across the board in terms of all of our different configurations, and our configurations include what we may do in terms of China.

But keep that in mind that folks are also, at the same time, looking to build out their Blackwell. So we've got a little bit of both happening in Q4. But yes, is it possible for Hopper to grow between Q3 and Q4? It's possible but we'll just have to see.

Operator

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

Joseph Lawrence Moore
Morgan Stanley, Research Division

Great. I wonder if you could talk a little bit about what you're seeing in the inference market. You've talked about Strawberry and some of the ramifications of longer scaling inference projects. But you've also talked about the possibility that as some of these Hopper clusters age that you could use some of the Hopper latent chips for inference. So I guess do you expect inference to outgrow training in the next kind of 12 months time frame? And just generally, your thoughts there.

Jensen Huang

Our hopes and dreams is that someday, the world does a ton of inference. And that's when AI has really exceeded is when every single company is doing inference inside their companies for the marketing department and forecasting department and supply chain group and their legal department and engineering, of course, and coding of course. And so we hope that every company is doing inference 24/7.

And that there will be a whole bunch of AI native startups, thousands of AI native startups that are generating tokens and generating AI. And every aspect of your computer experience from using Outlook to PowerPointing or when you're sitting there with Excel, you're constantly generating tokens. And every time you read a PDF, open a PDF, it generated a whole bunch of tokens. One of my favorite applications is NotebookLM, this Google application that came out. I use the living daylights out of it just because it's fun.

And I put every PDF, every archived paper into it just to listen to it as well as scanning through it. And so I think that's the goal is to train these models so that people use it. And there's now a whole new era of AI, if you will, a whole new genre of AI called physical AI, just those large language models understand the human language and how the thinking process, if you will. Physical AI understands the physical world. And it understands the meaning of the structure and understands what's sensible and what's not and what could happen and what won't.

And not only does it understand but it can predict, roll out a short future. That capability is incredibly valuable for industrial AI and robotics. And so that's fired up so many AI native companies and robotics companies and physical AI companies that you're probably hearing about. And it's really the reason why we built Omniverse. Omniverse is so that we can enable these AIs to be created and learn in Omniverse and learn from synthetic data generation and reinforcement, learning physics feedback instead of just a human feedback, it's now physics feedback. To have these capabilities, Omniverse was created so that we can enable physical AI.

And so that -- the goal is to generate tokens. The goal is to inference, and we're starting to see that growth happening. So I'm super excited about that. Now let me just say one more thing. Inference is super hard. And the reason why inference is super hard is because you need the accuracy to be high on the one hand. You need the throughput to be high so that the cost could be as low as possible, but you also need the latency to be low.

And computers that are high-throughput as well as low latency is incredibly hard to build. And these applications have long context lengths because they want to understand. They want to be able to inference within understanding the context of what they're being asked to do. And so the context length is growing larger and larger. On the other hand, the models are getting larger. They're multimodality.

Just the number of dimensions that inference is innovating is incredible. And this innovation rate is what makes NVIDIA's architecture so great because we -- our ecosystem is fantastic. Everybody knows that if they innovate on top of CUDA and top of NVIDIA's architecture, they can innovate more quickly and they know that everything should work. And if something were to happen, it's probably likely their code and not ours.

And so that ability to innovate in every single direction at the same time, having a large installed base so that whatever you create could land on an NVIDIA computer and be deployed broadly all around the world in every single data center all the way out to the edge into robotic systems, that capability is really quite phenomenal.

Operator

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

Aaron Christopher Rakers
Wells Fargo Securities, LLC, Research Division

I wanted to ask you, as we kind of focus on the Blackwell cycle and think about the Data Center business. When I look at the results this last quarter, Colette, you mentioned that obviously, the networking business was down about 15% sequentially. But then your comments were that you were seeing very strong demand. You mentioned also that you had multiple cloud CSP design wins for these large-scale clusters. So I'm curious if you could unpack what's going on in the Networking business and where maybe you've seen some constraints and just your confidence in the pace of Spectrum-X progressing to that multiple billions of dollars that you previously had talked about.

Colette M. Kress

Executive VP & CFO

Let's first start with the networking. The growth year-over-year is tremendous. And our focus since the beginning of our acquisition of Mellanox has really been about building together the work that we do in terms of in the data center. The networking is such a critical part of that. Our ability to sell our networking with many of our systems that we are doing in data center is continuing to grow and do quite well.

So this quarter is just a slight dip down and we're going to be right back up in terms of growing. We're getting ready for Blackwell and more and more systems that will be using not only our existing networking but also the networking that is going to be incorporated in a lot of these large systems we are providing them to.

Operator

Your next question comes from the line of Atif Malik of Citi.

Atif Malik
Citigroup Inc., Research Division

I have 2 quick ones for Colette. Colette, on the last earnings call, you mentioned that sovereign demand is in low double-digit billions. Can you provide an update on that? And then can you explain the supply-constrained situation in gaming? Is that because you're shifting your supply towards data center?

Colette M. Kress
Executive VP & CFO

So first starting in terms of sovereign AI, such an important part of growth, something that is really surfaced with the onset of generative AI and building models in the individual countries around the world. And we see a lot of them, and we talked about a lot of them on the call today and the work that they are doing. So our sovereign AI and our pipeline going forward is still absolutely intact as those are working to build these foundational models in their own language, in their own culture, and working in terms of the enterprises within those countries.

And I think you'll continue to see this be growth opportunities that you may see with our regional clouds that are being stood up and/or those that are focusing in terms of AI factories for many parts of the sovereign AI. This is areas where this is growing not only in terms of in Europe, but you're also seeing this in terms of growth in terms of in the Asia Pac as well.

Let me flip to your second question that you asked regarding Gaming. So our Gaming right now from a supply, we're busy trying to make sure that we can ramp all of our different products. And in this case, our gaming supply, given what we saw selling through was moving quite fast. Now the challenge that we have is how fast could we get that supply getting ready into the market for this quarter. Not to worry, I think we'll be back on track with more supply as we turn the corner into the new calendar year. We're just going to be tight for this quarter.

Operator

Your next question comes from the line of Ben Reitzes of Melius Research.

Benjamin Alexander Reitzes
Melius Research LLC

I wanted to ask Colette and Jensen with regard to sequential growth. So very strong sequential growth this quarter, and you're guiding to about 7%. Do your comments on Blackwell imply that we reaccelerate from there as you get more supply? Just in the first half, it would seem that there would be some catch-ups. So I was wondering how prescriptive you could be there.

And then, Jensen, just overall with the change in administration that's going to take place here in the U.S. and the China situation, have you gotten any sense or any conversations about tariffs or anything with regard to your China business? Any sense of what may or may not go on? It's probably too early but wondering if you had any thoughts there.

Jensen Huang

We guide 1 quarter at a time.

Colette M. Kress
Executive VP & CFO

We are working right now on the quarter that we're in and building what we need to ship in terms of Blackwell. We have every supplier on the planet working seamlessly with us to do that. And once we get to next quarter, we'll help you understand in terms of that ramp that we'll see to the next quarter going after that.

Jensen Huang

Whatever the new administration decides, we will, of course, support the administration. And that's our -- the highest mandate. And then after that, do the best we can and just as we always do. And so we have to simultaneously and we will comply with any regulation that comes along fully and support our customers to the best of our abilities and to compete in the marketplace. We'll do all of these 3 things simultaneously.

Operator

Your final question comes from the line of Pierre Ferragu of New Street Research.

Pierre C. Ferragu
New Street Research LLP

Jensen, you mentioned in your comments, you have the pretraining, the actual language models and you have reinforcement learning that becomes more and more important in training and inference as well and then you have inference itself. And I was wondering if you have a sense like a high-level typical sense of out of an overall AI ecosystem, like maybe one of your clients or one of the large models that are out there. Today, how much of the compute goes into each of these buckets? How much for the pretraining? How much for the reinforcement? And how much into inference today? Do you have any sense for how it's splitting and where the growth is the most important as well?

Jensen Huang

Well, today, it's vastly in pretraining a foundation model because, as you know, post-training, the new technologies are just coming online. And whatever you could do in pretraining and post-training, you would try to do so that the inference cost could be as low as possible for everyone. However, there are only so many things that you could do a priority. And so you'll always have to do on-the-spot thinking and in context thinking and a reflection.

And so I think that the fact that all 3 are scaling is actually very sensible based on where we are. And in the area foundation model, now we have multimodality foundation models and the amount of petabytes video that these foundation models are going to be trained on, it's incredible. And so my expectation is that for the foreseeable future, we're going to be scaling pretraining, post-training as well as inference time scaling and which is the reason why I think we're going to need more and more compute. And we're going to have to drive as hard as we can to keep increasing the performance by X factors out of time so that we can continue to drive down the cost and continue to increase the revenues and get the AI revolution going.

Operator

I'll now turn the call back over to Jensen Huang for closing remarks.

Jensen Huang

Thank you. The tremendous growth in our business is being fueled by 2 fundamental trends that are driving global adoption of NVIDIA computing. First, the computing stack is undergoing a reinvention, a platform shift from coding to machine learning, from executing code on CPUs to processing neural

networks on GPUs. The $1 trillion installed base of traditional data center infrastructure is being rebuilt for Software 2.0, which applies machine learning to produce AI.

Second, the age of AI is in full steam. Generative AI is not just a new software capability but a new industry with AI factories manufacturing digital intelligence, a new industrial revolution that can create a multi-trillion-dollar AI industry. Demand for Hopper and anticipation for Blackwell, which is now in full production, are incredible for several reasons. There are more foundation model makers now than there were a year ago.

The computing scale of pretraining and post-training continues to grow exponentially. There are more AI native startups than ever and the number of successful inference services is rising. And with the introduction of ChatGPT o1, OpenAI o1, a new scaling law called test-time scaling has emerged. All of these consume a great deal of computing. AI is transforming every industry, company and country. Enterprises are adopting agentic AI to revolutionize workflows. Over time, AI coworkers will assist employees in performing their jobs faster and better.

Investments in industrial robotics are surging due to breakthroughs in physical AI, driving new training infrastructure demand as researchers train world foundation models on petabytes of video and Omniverse synthetically generated data. The age of robotics is coming. Countries across the world recognize the fundamental AI trends we are seeing and have awakened to the importance of developing their national AI infrastructure.

The age of AI is upon us and it's large and diverse. NVIDIA's expertise, scale, and ability to deliver full stack and full infrastructure lets us serve the entire multi-trillion-dollar AI and robotics opportunities ahead from every hyperscale cloud, enterprise private cloud to sovereign regional AI clouds, on-prem to industrial edge and robotics. Thanks for joining us today, and catch up next time.

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