Transcripts

NVIDIA Corporation (NVDA) Q2 2025 Earnings Call Transcript

Aug. 28, 2024 9:04 PM ET | **NVIDIA Corporation (NVDA) Stock**, **NVDA:CA Stock** | 29 Comments |

SA Transcripts
149.34K Followers

Summary

Play Call **Press Release 10-Q**

EPS of $0.68 beats by $0.04 | Revenue of $30.04B (122.40% Y/Y) beats by $1.29B
2 Likes

Q2: 2024-08-28 Earnings

NVIDIA Corporation (NASDAQ:NVDA) Q2 2025 Earnings Conference Call August 28, 2024 5:00 PM ET

Company Participants

Stewart Stecker - Investor Relations
Colette Kress - Executive Vice President and Chief Financial Officer
Jensen Huang - President and Chief Executive Officer

Conference Call Participants

Vivek Arya - Bank of America Securities
Toshiya Hari - Goldman Sachs
Joe Moore - Morgan Stanley
Matt Ramsay - TD Cowen
Timothy Arcuri - UBS

Stacy Rasgon - Bernstein Research
Ben Reitzes - Melius
C.J. Muse - Cantor Fitzgerald
Aaron Rakers - Wells Fargo

Operator

Good afternoon. My name is Abby and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Second 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] Thank you.

And Mr. Stewart Stecker, you may begin your conference.

Stewart Stecker

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the second 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 would like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the third quarter of fiscal 2025. The content of today's call is NVIDIA's property. It cannot be reproduced or transcribed without prior written consent. During this call, we may make forward-looking statements based on current expectations. These are subject to a number of risks, 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, August 28th, 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.

Let me highlight an upcoming event for the financial community. We will be attending the Goldman Sachs Communacopia and Technology Conference on September 11 in San Francisco, where Jensen will participate in a keynote fireside chat. Our earnings call to discuss the results of our third quarter of fiscal 2025 is scheduled for Wednesday, November 20th, 2024.

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

Colette Kress

Thanks, Stewart. Q2 was another record quarter. Revenue of $30 billion was up 15% sequentially and up 122% year-on-year and well above our outlook of $28 billion. Starting with data center, data center revenue of $26.3 billion was a record, up 16% sequentially and up 154% year-on year, driven by strong demand for NVIDIA Hopper, GPU computing, and our networking platforms.

Compute revenue grew more than 2.5 times, networking revenue grew more than 2 times from the last year. Cloud service providers represented roughly 45% for our data center revenue and more than 50% stemmed from the consumer, Internet, and enterprise companies.

Customers continue to accelerate their Hopper architecture purchases, while gearing up to adopt Blackwell. Key workloads driving our data center growth include generative AI, model training, and inferencing. Video, image, and text data pre and post-processing with CUDA and AI workloads, synthetic data generation, AI-powered recommender systems, SQL, and vector database processing as well.

Next-generation models will require 10 to 20 times more compute to train with significantly more data. The trend is expected to continue. Over the trailing four quarters, we estimate that inference drove more than 40% of our data center revenue. CSPs, consumer Internet companies, and enterprises benefit from the incredible throughput and efficiency of NVIDIA's inference platform. Demand for NVIDIA is coming from frontier model makers, consumer Internet services, and tens of thousands of companies and startups building generative AI applications for consumers, advertising, education, enterprise and healthcare, and robotics. Developers desire NVIDIA's rich ecosystem and availability in every cloud. CSPs appreciate the broad adoption of NVIDIA and are growing their NVIDIA capacity given the high demand.

NVIDIA H200 platform began ramping in Q2, shipping to large CSPs, consumer Internet, and enterprise company. The NVIDIA H200 builds upon the strength of our Hopper architecture and

offering, over 40% more memory bandwidth compared to the H100.

Our data center revenue in China grew sequentially in Q2 and is a significant contributor to our data center revenue. As a percentage of total data center revenue, it remains below levels seen prior to the imposition of export controls. We continue to expect the China market to be very competitive going-forward. The latest round of MLPerf inference benchmarks highlighted NVIDIA's inference leadership with both NVIDIA, Hopper, and Blackwell platforms combining to win gold medals on all tasks.

At Computex, NVIDIA with the top computer manufacturers unveiled an array of Blackwell architecture-powered systems and NVIDIA networking for building AI factories and data centers. With the NVIDIA MGX modular reference architecture, our OEMs and ODM partners are building more than 100 Blackwell-based systems designed quickly and cost-effectively.

The NVIDIA Blackwell platform brings together multiple GPU, CPU, DPU, NVLink, NVLink switch, and the networking chips systems, and NVIDIA CUDA software to power the next-generation of AI across the cases, industries, and countries. The NVIDIA GB 200 NVL72 system with the fifth generation NVLink enables all 72 GPUs to act as a single GPU and deliver up to 30 times faster inference for LLMs, workloads, and unlocking the ability to run trillion parameter models in real time.

Hopper demand is strong and Blackwell is widely sampling. We executed a change to the Blackwell GPU mass to improve production yields. Blackwell production ramp is scheduled to begin in the fourth quarter and continue into fiscal year '26.

In Q4, we expect to ship several billion dollars in Blackwell revenue. Hopper shipments are expected to increase in the second half of fiscal 2025. Hopper supply and availability have improved. Demand for Blackwell platforms is well above supply, and we expect this to continue into next year. Networking revenue increased 16% sequentially.
Our Ethernet for AI revenue, which includes our Spectrum-X end-to-end Ethernet platform, doubled sequentially with hundreds of customers adopting our Ethernet offerings. Spectrum-X has broad market support from OEM and ODM partners and is being adopted by CSPs, GPU cloud providers, and enterprise, including X-AI to connect the largest GPU compute cluster in the world.

Spectrum-X supercharges Ethernet for AI processing and delivers 1.6 times the performance of traditional Ethernet. We plan to launch new Spectrum-X products every year to support demand

for scaling compute clusters from tens of thousands of DPUs today to millions of GPUs in the near future. Spectrum-X is well on-track to begin a multi-billion dollar product line within a year. Our sovereign AI opportunities continue to expand as countries recognize AI expertise and infrastructure at national imperatives for their society and industries.

Japan's National Institute of Advanced Industrial Science and Technology is building its AI bridging cloud infrastructure 3.0 supercomputer with NVIDIA. We believe sovereign AI revenue will reach low-double-digit billions this year. The enterprise AI wave has started. Enterprises also drove sequential revenue growth in the quarter. We are working with most of the Fortune 100 companies on AI initiatives across industries and geographies. A range of applications are fueling our growth, including AI-powered chatbots, generative AI copilots, and agents to build new monetizable business applications and enhance employee productivity.

Amdocs is using NVIDIA generative AI for their smart agent, transforming the customer experience and reducing customer service costs by 30%. ServiceNow is using NVIDIA for its Now Assist offering, the fastest-growing new product in the company's history. SAP is using NVIDIA to build dual Co-pilots. Cohesity is using NVIDIA to build their generative AI agent and lower generative AI development costs. Snowflake serves over 3 billion queries a day for over 10,000 enterprise customers is working with NVIDIA to build Copilots. And lastly, Wistron is using NVIDIA AI Omniverse to reduce end-to-end cycle times for their factories by 50%.

Automotive was a key growth driver for the quarter as every automaker developing autonomous vehicle technology is using NVIDIA in their data centers. Automotive will drive multi-billion dollars in revenue across on-prem and cloud consumption and will grow as next-generation AV models require significantly more compute. Healthcare is also on its way to being a multi-billion dollar business as AI revolutionizes medical imaging, surgical robots, patient care, electronic health record processing, and drug discovery.
During the quarter, we announced a new NVIDIA AI foundry service to supercharge generative AI for the world's enterprises with Meta's Llama 3.1, collection of models. This marks a watershed moment for enterprise AI. Companies for the first time can leverage the capabilities of an open source frontier-level model to develop customized AI applications to encode their institutional knowledge into an AI flywheel to automate and accelerate their business.

Accenture is the first to adopt the new service to build custom Llama 3.1 models for both its own use and to assist clients seeking to deploy generative AI applications. Nvidia NIM accelerate and simplify model deployment. Companies across healthcare, energy, financial services, retail,

transportation, and telecommunications are adopting NIMs, including Aramco, Lowe's, and Uber. AT&T realized 70% cost savings and 8 times latency reduction after moving into NIMs for generative AI, call transcription, and classification.

Over 150 partners are embedding NIMs across every layer of the AI ecosystem. We announced NIM agent Blueprints, a catalog of customizable reference applications that include a full suite of software for building and deploying enterprise generative AI applications. With NIM agent blueprints, enterprises can refine their AI applications overtime, creating a data-driven AI flywheel. The first NIM agent blueprints include workloads for customer service, computer-aided drug discovery, and enterprise retrieval augmented generation.

Our system integrators, technology solution providers, and system builders are bringing NVIDIA NIM agent blueprints to enterprises. NVIDIA NIM and NIM agent blueprints are available through the NVIDIA AI enterprise software platform, which has great momentum. We expect our software, SaaS and support revenue to approach a $2 billion annual run rate exiting this year, with NVIDIA AI Enterprise notably contributing to growth.

Moving to gaming and AI PCs. Gaming revenue of $2.88 billion increased 9% sequentially and 16% year-on-year. We saw sequential growth in console, notebook, and desktop revenue and demand is strong and growing and channel inventory remains healthy. Every PC with RTX is an AIPC. RTX PCs can deliver up to 1,300 AI tops and there are now over 200 RTX AI laptops designs from leading PC manufacturers.

With 600 AI-powered applications and games and an installed base of 100 million devices, RTX is set to revolutionize consumer experiences with generative AI. NVIDIA ACE, a suite of generative AI technologies is available for RTX, AI PCs. Mecha BREAK is the first game to use NVIDIA ACE, including our small large -- small language model, Minitron-4B optimized on device inference. The NVIDIA gaming ecosystem continues to grow, recently added RTX and DLSS titles including Indiana Jones and the Great Circle, Dune Awakening, and Dragon Age: The Veilguard. The GeForce NOW library continues to expand with total catalog size of over 2,000 titles, the most content of any cloud gaming service.

Moving to Pro visualization. Revenue of $454 million was up 6% sequentially and 20% year-on year. Demand is being driven by AI and graphic use cases, including model fine-tuning and Omniverse-related workloads. Automotive and manufacturing were among the key industry verticals driving growth this quarter. Companies are racing to digitalize workflows to drive

efficiency across their operations.

The world's largest electronics manufacturer, Foxconn is using NVIDIA Omniverse to power digital twins of the physical plants that produce NVIDIA Blackwell systems. And several large global enterprises, including Mercedes-Benz signed multi-year contracts for NVIDIA Omniverse Cloud to build industrial digital twins for factories. We announced new NVIDIA USD NIMs and connectors to open Omniverse to new industries and enable developers to incorporate generative AI Copilots and agents into USD workflows, accelerating their ability to build highly accurate virtual worlds. WPP is implementing USD NIM microservices in its generative AI-enabled content creation pipeline for customers such as the Coca-Cola company.

Moving to automotive and robotics, revenue was $346 million, up 5% sequentially and up 37% year-on-year. Year-on-year growth was driven by the new customer ramps in self-driving platforms and increased demand for AI cockpit solutions. At the consumer -- at the Computer Vision and Pattern Recognition conference, NVIDIA won the Autonomous Brand Challenge in the end-to-end driving at-scale category, outperforming more than 400 entries worldwide. Austin Dynamics, BYD Electronics, Figure, Intrinsic, Siemens, Skilled ADI, and Teradyne Robotics are using the NVIDIA Isaac Robotics platform for autonomous robot arms, humanoids, and mobile robots.

Now moving to the rest of the P&L. GAAP gross margins were 75.1% and non-GAAP gross margins were 75.7%, down sequentially due to a higher mix of new products within data center and inventory provisions for low-yielding Blackwell material. Sequentially, GAAP and non-GAAP operating expenses were up 12%, primarily reflecting higher compensation-related costs. Cash flow from operations was $14.5 billion.
In Q2, we utilized cash of $7.4 billion towards shareholder returns in the form of share repurchases and cash dividends, reflecting the increase in dividend per share. Our Board of Directors recently approved a $50 billion share repurchase authorization to add to our remaining $7.5 billion of authorization at the end of Q2.

Let me turn the outlook for the third quarter. Total revenue is expected to be $32.5 billion, plus or minus 2%. Our third-quarter revenue outlook incorporates continued growth of our Hopper architecture and sampling of our Blackwell products. We expect Blackwell production ramp in Q4. GAAP and non-GAAP gross margins are expected to be 74.4% and 75%, respectively, plus or minus 50 basis points.

As our data center mix continues to shift to new products, we expect this trend to continue into the fourth quarter of fiscal 2025. For the full-year, we expect gross margins to be in the mid-70% range. GAAP and non-GAAP operating expenses are expected to be approximately $4.3 billion and $3.0 billion, respectively.

Full-year operating expenses are expected to grow in the mid-to-upper 40% range as we work on developing our next generation of products. GAAP and non-GAAP other income and expenses are expected to be about $350 million, including gains and losses from non-affiliated investments and publicly-held equity securities.

GAAP and non-GAAP tax rates are expected to be 17%, plus or minus 1%, excluding any discrete items. Further financial detail are included in the CFO commentary and other information available on our IR website. We are now going to open the call for questions.

Operator, would you please help us poll for questions.

Question-and-Answer Session

Operator

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

Vivek Arya
Thanks for taking my question. Jensen, you mentioned in the prepared comments that there is a change in the Blackwell GPU mask. I'm curious, are there any other incremental changes in back end packaging or anything else? And I think related, you suggested that you could ship several billion dollars of Blackwell in Q4 despite the change in the design. Is it because all these issues will be solved by then? Just help us size what is the overall impact of any changes in Blackwell timing? What that means to your kind of revenue profile and how are customers reacting to it?

Jensen Huang

Yes. Thanks, Vivek. The change to the mask is complete. There were no functional changes necessary. And so we're sampling functional samples of Blackwell -- Grace Blackwell in a variety of system configurations as we speak. There are something like 100 different types of Blackwell based systems that are built that were shown at Computex. And we're enabling our ecosystem to

start sampling those. The functionality of Blackwell is as it is, and we expect to start production in Q4.

Operator

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

Toshiya Hari

Hi, thank you so much for taking the question. Jensen, I had a relatively longer-term question. As you may know, there's a pretty heated debate in the market on your customers and customer's customers return on investment and what that means for the sustainability of CapEx going forward. Internally at NVIDIA, like what are you guys watching? What's on your dashboard as you try to gauge customer return and how that impacts CapEx?

And then a quick follow-up maybe for Colette. I think your sovereign AI number for the full-year went up maybe a couple of billion. What's driving the improved outlook? And how should we think about fiscal '26? Thank you.

Jensen Huang

Thanks, Toshiya. First of all, when I said ship production in Q4, I mean shipping out. I don't mean starting to ship, but I mean -- I don't mean starting production, but shipping out. On the longer term question, let's take a step-back and you've heard me say that we're going through two simultaneous platform transitions at the same time. The first one is transitioning from accelerated computing to -- from general-purpose computing to accelerated computing. And the reason for that is because CPU scaling has been known to be slowing for some time. And it is slow to a crawl. And yet the amount of computing demand continues to grow quite significantly. You could maybe even estimate it to be doubling every single year.

And so if we don't have a new approach, computing inflation would be driving up the cost for every company, and it would be driving up the energy consumption of data centers around the world. In fact, you're seeing that. And so the answer is accelerated computing. We know that accelerated computing, of course, speeds up applications. It also enables you to do computing at a much larger scale, for example, scientific simulations or database processing. But what that translates directly to is lower cost and lower energy consumed.

And in fact, this week, there's a blog that came out that talked about a whole bunch of new

libraries that we offer. And that's really the core of the first platform transition going from general purpose computing to accelerated computing. And it's not unusual to see someone save 90% of their computing cost. And the reason for that is, of course, you just sped up an application 50x, you would expect the computing cost to decline quite significantly.

The second was enabled by accelerated computing because we drove down the cost of training large language models or training deep learning so incredibly, that it is now possible to have gigantic scale models, multi-trillion parameter models, and train it on -- pre-train it on just about the world's knowledge corpus and let the model go figure out how to understand a human represent -- human language representation and how to codify knowledge into its neural networks and how to learn reasoning, and so -- which caused the generative AI revolution.

Now generative AI, taking a step back about why it is that we went so deeply into it is because it's not just a feature, it's not just a capability, it's a fundamental new way of doing software. Instead of human-engineered algorithms, we now have data. We tell the AI, we tell the model, we tell the computer what's the -- what are the expected answers, What are our previous observations. And then for it to figure out what the algorithm is, what's the function. It learns a universal -- AI is a bit of a universal function approximator and it learns the function.
And so you could learn the function of almost anything, you know. And anything that you have that's predictable, anything that has structure, anything that you have previous examples of. And so now here we are with generative AI. It's a fundamental new form of computer science. It's affecting how every layer of computing is done from CPU to GPU, from human-engineered algorithms to machine-learned algorithms. And the type of applications you could now develop and produce is fundamentally remarkable.

And there are several things that are happening in generative AI. So the first thing that's happening is the frontier models are growing in quite substantial scale. And they're still seeing -- we're still all seeing the benefits of scaling. And whenever you double the size of a model, you also have to more than double the size of the dataset to go train it. And so the amount of flops necessary in order to create that model goes up quadratically. And so it's not unusual -- it's not unexpected to see that the next-generation models could take 20 -- 10, 20, 40 times more compute than last generation.

So we have to continue to drive the generational performance up quite significantly, so we can drive down the energy consumed and drive down the cost necessary to do it. So the first one is, there are larger frontier models trained on more modalities and surprisingly, there are more

frontier model makers than last year. And so you have more on more on more. That's one of the dynamics going on in generative AI. The second is although it's below the tip of the iceberg. What we see are ChatGPT, image generators, we see coding. We use a generative AI for coding quite extensively here at NVIDIA now.

We, of course, have a lot of digital designers and things like that. But those are kind of the tip of the iceberg. What's below the iceberg are the largest systems -- largest computing systems in the world today, which are -- and you've heard me talk about this in the past, which are recommender systems moving from CPUs, it's now moving from CPUs to generative AI. So recommended systems, ad generation, custom ad generation targeting ads at very large scale and quite hyper targeting search and user-generated content. These are all very large-scale applications have now evolved to generative AI.

Of course, the number of generative AI startups is generating tens of billions of dollars of cloud renting opportunities for our cloud partners and sovereign AI. Countries that are now realizing that their data is their natural and national resource and they have to use -- they have to use AI, build their own AI infrastructure so that they could have their own digital intelligence.
Enterprise AI, as Colette mentioned earlier, is starting and you might have seen our announcement that the world's leading IT companies are joining us to take the NVIDIA AI enterprise platform to the world's enterprises. The companies that we're talking to. So many of them are just so incredibly excited to drive more productivity out of their company.

And then I -- and then General Robotics. The big transformation last year as we are able to now learn physical AI from watching video and human demonstration and synthetic data generation from reinforcement learning from systems like Omniverse. We are now able to work with just about every robotics companies now to start thinking about start building on general robotics. And so you can see that there are just so many different directions that generative AI is going. And so we're actually seeing the momentum of generative AI accelerating.

Colette Kress

And Toshiya, to answer your question regarding sovereign AI and our goals in terms of growth, in terms of revenue, it certainly is a unique and growing opportunity, something that surfaced with generative AI and the desires of countries around the world to have their own generative AI that would be able to incorporate their own language, incorporate their own culture, incorporate their own data in that country. So more and more excitement around these models and what they can

be specific for those countries. So yes, we are seeing some growth opportunity in front of us.

Operator

And your next question comes from the line of Joe Moore with Morgan Stanley. Your line is open.

Joe Moore

Great. Thank you. Jensen, in the press release, you talked about Blackwell anticipation being incredible, but it seems like Hopper demand is also really strong. I mean, you're guiding for a very strong quarter without Blackwell in October. So how long do you see sort of coexisting strong demand for both? And can you talk about the transition to Blackwell? Do you see people intermixing clusters? Do you think most of the Blackwell activity is new clusters? Just some sense of what that transition looks like?

Jensen Huang

Yes. Thanks, Joe. The demand for Hopper is really strong and it's true. The demand for Blackwell is incredible. There's a couple of reasons for that. The first reason is, if you just look at the world's cloud service providers and the amount of GPU capacity they have available, it's basically none. And the reason for that is because they're either being deployed internally for accelerating their own workloads, data processing, for example.

Data processing, we hardly ever talk about it because it's mundane. It's not very cool because it doesn't generate a picture or generate words, but almost every single company in the world processes data in the background. And NVIDIA's GPUs are the only accelerators on the planet that process and accelerate data. SQL data, Pandas data science, toolkits like Pandas and the new one, Polars, these are the most popular data processing platforms in the world.

And aside from CPUs, which as I've mentioned before, really running out of steam, Nvidia's accelerated computing is really the only way to get boosting performance out of that. And so that's number one is the primary -- the number one use-case long before generative AI came along is the migration of applications one after another to accelerated computing.

The second is, of course, the rentals. Their renting capacity to model makers or renting it to startup companies and a generative AI company spends the vast majority of their invested capital into infrastructure so that they could use an AI to help them create products. And so these companies need it now. They just simply can't afford -- you just raise money, you -- they want you

to put it to use now. You have processing that you have to do. You can't do it next year, you got to do it today. And so there's a fair -- that's one reason.

The second reason for Hopper demand right now is because of the race to the next plateau. So the first person to the next plateau, it gets to be -- gets to introduce a revolutionary level of AI. So the second person who gets there is incrementally better or about the same. And so the ability to systematically and consistently race to the next plateau and be the first one there, is how you establish leadership. NVIDIA is constantly doing that and we show that to the world and the GPUs we make and AI factories that we make, the networking systems that we make, the SOCs we create.

I mean, we want to set the pace. We want to be consistently the world's best. And that's the reason why, we drive ourselves so hard. And of course, we also want to see our dreams come true, and all of the capabilities that we imagine in the future, and the benefits that we can bring to society. We want to see all that come true. And so these model makers are the same. They're -- of course, they want to be the world's best, they want to be the world's first. And although Blackwell will start shipping out in billions of dollars at the end of this year, the standing up of the capacity is still probably weeks and a month or so away. And so between now and then is a lot of generative AI market dynamic.

And so everybody is just really in a hurry. It's either operational reasons that they need it, they need accelerated computing. They don't want to build any more general-purpose computing infrastructure and even Hopper, you know, of course, H200 state-of-the-art. Hopper, if you have a choice between building CPU infrastructure right now for business or Hopper infrastructure for business right now, that decision is relatively clear. And so I think people are just clamoring to transition the trillion dollars of established installed infrastructure to a modern infrastructure in Hopper's state-of-the-art.

Operator

And your next question comes from the line of Matt Ramsay with TD Cowen. Your line is open.

Matt Ramsay

Thank you very much. Good afternoon, everybody. I wanted to kind of circle back to an earlier question about the debate that investors are having about, I don't know, the ROI on all of this CapEx, and hopefully this question and the distinction will make some sense. But what I'm having discussions about is with like the percentage of folks that you see that are spending all of this

money and looking to sort of push the frontier towards AGI convergence and as you just said, a new plateau and capability.

And they're going to spend regardless to get to that level of capability because it opens up so many doors for the industry and for their company versus customers that are really, really focused today on CapEx versus ROI. I don't know if that distinction makes sense. I'm just trying to get a sense of how you're seeing the priorities of people that are putting the dollars in the ground on this new technology and what their priorities are and their time frames are for that investment? Thanks.

Jensen Huang
Thanks, Matt. The people who are investing in NVIDIA infrastructure are getting returns on it right away. It's the best ROI infrastructure -- computing infrastructure investment you can make today. And so one way to think through it, probably the most -- the easiest way to think through is just go back to first principles. You have trillion dollars' worth of general-purpose computing infrastructure and the question is, do you want to build more of that or not? And for every $1 billion worth of general CPU-based infrastructure that you stand up, you probably rent it for less than $1 billion. And so because it's commoditized. There's already a trillion dollars on the ground. What's the point of getting more?

And so the people who are clamoring to get this infrastructure, one, when they build-out Hopper based infrastructure and soon Blackwell-based infrastructure, they start saving money. That's tremendous return on investment. And the reason why they start saving money is because data processing saves money, data processing is pricing, just a giant part of it already. And so recommended system save money, so on and so forth, okay? And so you start saving money.

The second thing is everything you stand-up are going to get rented because so many companies are being founded to create generative AI. And so your capacity gets rented right away and the return on investment of that is really good. And then the third reason is your own business. Do you want to either create the next frontier yourself or your own Internet services benefit from a next-generation ad system or a next-generation recommended system or next-generation search system?

So for your own services, for your own stores, for your own user-generated content, social media platforms for your own services, generative AI is also a fast ROI. And so there's a lot of ways you could think through it. But at the core, it's because it is the best computing infrastructure you could

put in the ground today. The world of general-purpose computing is shifting to accelerated computing. The world of human-engineered software is moving to generative AI software. If you were to build infrastructure to modernize your cloud and your data centers, build it with accelerated computing NVIDIA. That's the best way to do it.

Operator

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

open. **Timothy Arcuri**

Thanks a lot. I had a question on the shape of the revenue growth both near and longer-term. I know, Colette, you did increase OpEx for the year. And if I look at the increase in your purchase commitments and your supply obligations, that's also quite bullish. On the other hand, there's some school of thought that not that many customers really seem ready for liquid cooling and I do recognize that some of these racks can be air-cooled. But Jensen, is that something to consider sort of on the shape of how Blackwell is going to ramp?

And then I guess when you look beyond next year, which is obviously going to be a great year and you look into '26, do you worry about any other gating factors like, say, the power, supply chain or at some point, models start to get smaller? I'm just wondering if you could speak to that? Thanks.

Jensen Huang

I'm going to work backwards. I really appreciate the question, Tim. So remember, the world is moving from general purpose computing to accelerated computing. And the world builds about $1 trillion dollars' worth of data centers -- $1 trillion dollars' worth of data centers in a few years will be all accelerated computing. In the past, no GPUs are in data centers, just CPUs. In the future, every single data center will have GPUs. And the reason for that is very clear is because we need to accelerate workloads so that we can continue to be sustainable, continue to drive down the cost of computing so that when we do more computing our -- we don't experience computing inflation.

Second, we need GPUs for a new computing model called generative AI that we can all acknowledge is going to be quite transformative to the future of computing. And so I think working backwards, the way to think about that is the next trillion dollars of the world's infrastructure will

clearly be different than the last trillion, and it will be vastly accelerated.

With respect to the shape of our ramp, we offer multiple configurations of Blackwell. Blackwell comes in either a Blackwell classic, if you will, that uses the HGX form factor that we pioneered with Volta, and I think it was Volta. And so we've been shipping the HGX form factor for some time. It is air-cooled. The Grace Blackwell is liquid-cooled. However, the number of data centers that want to go liquid-cooled is quite significant.
And the reason for that is because we can in a liquid-cooled data center, in any data center -- power limited data center, whatever size data center you choose, you could install and deploy anywhere from 3 times to 5 times, the AI throughput compared to the past. And so liquid cooling is cheaper, liquid cooling TCO is better, and liquid cooling allows you to have the benefit of this capability we call NVLink , which allows us to expand it to 72 Grace Blackwell packages, which has essentially 144 GPUs.

And so imagine 144 GPUs connected in NVLink and that is -- we're increasingly showing you the benefits of that. And the next click is obviously a very low-latency, very-high throughput, large language model inference. And the large domain is going to be a game-changer for that. And so I think people are very comfortable deploying both. And so almost every CSP we're working with are deploying some of both. And so I -- I'm pretty confident that we'll ramp it up just fine.

Your second question out of the third is that looking-forward, yes, next year is going to be a great year. We expect to grow our data center business quite significantly next year. Blackwell is going to be going to be a complete game-changer for the industry. And Blackwell is going to carry into the following year. And as I mentioned earlier, working backwards from first principles.

Remember that computing is going through two platform transitions at the same time and that's just really, really important to keep your head on -- your mind focused on, which is general purpose computing is shifting to accelerated computing and human engineered software is going to transition to generative AI or artificial intelligence learned software.

Operator

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

Stacy Rasgon

Hi guys. Thanks for taking my questions. I have two short questions for Colette. The first, several billion dollars of Blackwell revenue in Q4. I guess is that additive? You said you expected Hopper demand to strengthen in the second half. Does that mean Hopper strengthens Q3 to Q4 as well on top of Blackwell adding several billion dollars?
And the second question on gross margins, if I have mid-70s for the year, explaining where I want to draw that. If I have 75% for the year, I'd be something like 71% to 72% for Q4 somewhere in that range. Is that the kind of exit rate for gross margins that you're expecting? And how should we think about the drivers of gross margin evolution into next year as Blackwell ramps? And I mean, hopefully, I guess the yields and the inventory reserves and everything come up.

Colette Kress

Yes. So Stacy, let's first take your question that you had about Hopper and Blackwell. So we believe our Hopper will continue to grow into the second half. We have many new products for Hopper, our existing products for Hopper that we believe will start continuing to ramp, in the next quarters, including our Q3 and those new products moving to Q4. So let's say Hopper there for versus H1 is a growth opportunity for that. Additionally, we have the Blackwell on top of that, and the Blackwell starting of -- ramping in Q4. So hope that helps you on those two pieces.

Your second piece is in terms of on our gross margin. We provided gross margin for our Q3. We provided our gross margin on a non-GAAP at about 75%. We'll work with all the different transitions that we're going through, but we do believe we can do that 75% in Q3. We provided that we're still on track for the full-year also in the mid-70s or approximately the 75%. So we're going to see some slight difference possibly in Q4.

Again, with our transitions and the different cost structures that we have on our new product introductions. However, I'm not in the same number that you are there. We don't have exactly guidance, but I do believe you're lower than where we are.

Operator

And your next question comes from the line of Ben Reitzes with Melius. Your line is open.

Ben Reitzes

Yes, hey, thanks a lot for the question, Jensen and Colette. I wanted to ask about the

geographies. There was the 10-Q that came out and the United States was down sequentially while several Asian geographies were up a lot sequentially. Just wondering what the dynamics are there? Obviously, China did very well. You mentioned it in your remarks, what are the puts and takes?
And then I just wanted to clarify from Stacy's question, if that means the sequential overall revenue growth rates for the company accelerate in the fourth quarter given all those favorable revenue dynamics? Thanks.

Colette Kress

Let me talk about a bit in terms of our disclosure in terms of the 10-Q, a required disclosure, and a choice of geographies. Very challenging sometimes to create that right disclosure as we have to come up with one key piece. Piece is in terms of we have in terms of who we sell to and/or specifically who we invoice to. And so what you're seeing in terms of there is who we invoice. That's not necessarily where the product will eventually be, and where it may even travel to the end-customer. These are just moving to our OEMs, our ODMs, and our system integrators for the most part across our product portfolio.

So what you're seeing there is sometimes just a swift shift in terms of who they are using to complete their full configuration before those things are going into the data center, going into notebooks and those pieces of it. And that shift happens from time-to-time. But yes, our China number there are inverse into China, keep in mind that is incorporating both gaming, also data center, also automotive in those numbers that we have.

Going back to your statement in regarding gross margin and also what we're seeing in terms of what we're looking at for Hopper and Blackwell in terms of revenue. Hopper will continue to grow in the second half. We'll continue to grow from what we are currently seeing. During -- determining that exact mix in each Q3 and Q4, we don't have here. We are not here to guide yet in terms of Q4. But we do see right now the demand expectations. We do see the visibility that will be a growth opportunity in Q4. On top of that, we will have our Blackwell architecture.

Operator

And your next question comes from the line of C.J. Muse with Cantor Fitzgerald. Your line is open.

C.J. Muse

Yes, good afternoon. Thank you for taking the question. You've embarked on a remarkable annual product cadence with challenges only likely becoming more and more given rising complexity and a radical limit in advanced package world. So curious, if you take a step back, how does this backdrop alter your thinking around potentially greater vertical integration, supply chain partnerships and then thinking through consequential impact to your margin profile? Thank you.

Jensen Huang

Yes, thanks. Let's see. I think the first answer to your -- the answer to your first question is that the reason why our velocity is so high is simultaneously because, the complexity of the model is growing and we want to continue to drive its cost down. It's growing, so we want to continue to increase its scale. And we believe that by continuing to scale the AI models that will reach a level of extraordinary usefulness and that it would open up, I realize the next industrial revolution. We believe it.

And so we're going to drive ourselves really hard to continue to go up that scale. We have the ability fairly uniquely to integrate -- to design an AI factory because we have all the parts. It's not possible to come up with a new AI factory every year, unless you have all the parts. And so we have -- next year, we're going to ship a lot more CPUs than we've ever had in the history of our company, more GPUs, of course, but also NVLink switches, CX DPUs, ConnectX EPU for East and West, Bluefield DPUs for North and South and data and storage processing to InfiniBand for supercomputing centers to Ethernet, which is a brand-new product for us, which is well on its way to becoming a multi-billion dollar business to bring AI to Ethernet.

And so the fact that we could build -- we have access to all of this. We have one architectural stack, as you know. It allows us to introduce new capabilities to the market as we complete it. Otherwise, what happens is, you ship these parts, you go find customers to sell it to and then you've got to build -- somebody has got to build up an AI factory. And the AI factory has got a mountain of software.

And so it's not about who integrates it. We love the fact that our supply chain is disintegrated in the sense that we could service Quanta, Foxconn, HP, Dell, Lenovo, Supermicro. We used to be able to serve as ZT. They were recently purchased and so on and so forth. And so the number of ecosystem partners that we have, Gigabyte, ASUS, the number of ecosystem partners that we have that allows us allows us to -- allows them to take our architecture, which all works, but integrated in a bespoke way into all of the world's cloud service providers, enterprise data centers.

The scale and reach necessary from our ODMs and our integrator supply-chain is vast and gigantic because the world is huge. And so that part we don't want to do and we're not good at doing. And I -- but we know how to design the AI infrastructure, provided the way that customers would like it and lets the ecosystem integrate it. Well, yes. So anyways, that's the reason why.

Operator
And your final question comes from the line of Aaron Rakers with Wells Fargo. Your line is open.

Aaron Rakers

Yes, thanks for taking the questions. I wanted to go back into the Blackwell product cycle. One of the questions that we tend to get asked is how you see the Rack Scale system mix dynamic as you think about leveraging NVLink? You think about GB, NVL72, and how that go-to-market dynamic looks as far as the Blackwell product cycle? I guess put this distinctly, how do you see that mix of Rack Scale systems as we start to think about the Blackwell cycle playing out?

Jensen Huang

Yes, Aaron, thanks. The Blackwell Rack system, it's designed and architected as a rack, but it's sold, in a disaggregated system components. We don't sell the whole rack. And the reason for that is because everybody's rack is a little different, surprisingly. You know, some of them are OCP standards, some of them are not, some of them are enterprise, and the power limits for everybody could be a little different. Choice of CDUs, the choice of power bus bars, the configuration and integration into people's data centers, all different.

And so the way we designed it, we architected the whole rack. The software is going to work perfectly across the whole rack. And then we provide the system components, like for example, the CPU and GPU compute board is then integrated into an MGX, it's a modular system architecture. MGX is completely ingenious. And we have MGX ODMs and integrators and OEMs all over the plant.

And so just about any configuration you would like, where you would like that 3,000 pound rack to be delivered. It's got to be close to -- it has to be integrated and assembled close to the data center because it's fairly heavy. And so everything from the supply chain from the moment that we ship the GPU, CPUs, the switches, the NICs. From that point forward, the integration is done quite close to the location of the CSPs and the locations of the data centers. And so you can

imagine how many data centers in the world there are and how many logistics hubs, we've scaled out to with our ODM partners.

And so I think that because we show it as one rack and because it's always rendered that way and shown that way, we might have left the impression that we're doing the integration. Our customers hate that we do integration. The supply-chain hates us doing integration. They want to do the integration. That's their value-added. There's a final design in, if you will. It's not quite as simple as shimmy into a data center, but the design fit in is really complicated.
And so the install -- the design fit-in, the installation, the bring-up, the repair and replace, that entire cycle is done all over the world. And we have a sprawling network of ODM and OEM partners that does this incredibly well. So integration is not the reason why we're doing racks. It's the anti-reason of doing it. The way we don't want to be an integrator. We want to be a technology provider.

Operator

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

Jensen Huang

Thank you. Let me make a couple more -- make a couple of comments that I made earlier again. That data center worldwide are in full steam to modernize the entire computing stack with accelerated computing and generative AI. Hopper demand remains strong and the anticipation for Blackwell is incredible. Let me highlight the top-five things, the top-five things of our company. Accelerated computing has reached a tipping point, CPU scaling slows, developers must accelerate everything possible. Accelerated computing starts with CUDA-X libraries. New libraries open new markets for NVIDIA.

We released many new libraries, including accelerated Polars, Pandas, and Spark, the leading data science and data processing libraries, QVS for vector product vector databases. This is incredibly hot right now. Aerial and Sionna for 5G wireless base station, a whole suite of -- a whole world of data centers that we can go into now. Parabricks for gene sequencing and Alpha fold two for protein structure prediction is now CUDA accelerated.

We are at the beginning of our journey to modernize a $1 trillion dollars' worth of data centers from general-purpose computing to accelerated computing. That's number-one. Number two, Blackwell is a step-function leap over Hopper. Blackwell is an AI infrastructure platform, not just a

GPU. It also happens to be in the name of our GPU, but it's an AI infrastructure platform.

As we reveal more of Blackwell and sample systems to our partners and customers, the extent of Blackwell's lead becomes clear. The Blackwell Vision took nearly five years and seven one-of-a kind chips to realize. The Grace CPU, the Blackwell dual GPU, and a coVS package, ConnectX DPU for East-West traffic, BlueField DPU for North-South and storage traffic, NVLink switch for all-to-all GPU communications, and Quantum and Spectrum-X for both InfiniBand and Ethernet can support the massive burst traffic of AI. Blackwell AI factories are building-sized computers.

The third is NVLink. This is a very big deal with its all-to-all GPU switch is game-changing. The Blackwell system lets us connect 144 GPUs in 72-GB 200 packages into one NVLink domain, with an aggregate NVLink bandwidth of 259 terabytes per second in one rack. Just to put that in perspective, that's about 10 times higher than Hopper, 259 terabytes per second, it kind of makes sense because you need to boost the training of multi-trillion parameter models on trillions of tokens. And so that natural amount of data needs to be moved around from GPU to GPU.

For inference, NVLink is vital for low-latency, high-throughput, large language model, token generation. We now have three networking platforms, NVLink for GPU scale-up, Quantum InfiniBand for supercomputing and dedicated AI factories, and Spectrum-X for AI on Ethernet. NVIDIA's networking footprint is much bigger than before.

Generative AI momentum is accelerating. Generative AI frontier model makers are racing to scale to the next AI plateau to increase model safety and IQ. We're also scaling to understand more modalities from text, images, and video to 3D, physics, chemistry, and biology. Chatbots,

coding AIs, and image generators are growing fast, but it's just the tip of the iceberg.

Internet services are deploying generative AI for large-scale recommenders, ad targeting, and search systems. AI start-ups are consuming tens of billions of dollars yearly of CSP's cloud capacity and countries are recognizing the importance of AI and investing in sovereign AI infrastructure. And NVIDIA AI and NVIDIA Omniverse is opening up the next era of AI general robotics.

And now the enterprise AI wave has started and we're poised to help companies transform their businesses. The NVIDIA AI Enterprise platform consists of NeMo, NIMS, NIM agent blueprints, and AI Foundry, that our ecosystem partners the world-leading IT companies used to help customer -- companies customize AI models and build bespoke AI applications. Enterprises can then deploy on NVIDIA AI Enterprise runtime and at $4,500 per GPU per year, NVIDIA AI Enterprise is an exceptional value for deploying AI anywhere. And for NVIDIA's software TAM can be significant as the CUDA-compatible GPU installed-base grows from millions to tens of millions. And as Colette mentioned, NVIDIA software will exit the year at a $2 billion run rate. Thank you all for joining us today.

Operator

Ladies and gentlemen, this concludes today's call and we thank you for your participation. You may now disconnect.