Seeking Alpha

Transcripts

NVIDIA Corporation (NVDA) Q3 2026 Earnings Call Transcript

Nov. 19, 2025 10:13 PM ET | **NVIDIA Corporation (NVDA) Stock**, **NVDA:CA Stock**, **ZNVD:CA Stock**

SA Transcripts
157.12K Followers

Q3: 2025-11-19 Earnings Summary

Insights ▶

EPS of $1.30 beats by $0.04 | Revenue of $57.01B (62.49% Y/Y) beats by $1.91B

NVIDIA Corporation (NVDA) Q3 2026 Earnings Call November 19, 2025 5:00 PM EST

Company Participants

Toshiya Hari - Vice President of Investor Relations & Strategic Finance
Colette Kress - Executive VP & CFO
Jen-Hsun Huang - Co-Founder, CEO, President & Director

Conference Call Participants

Joseph Moore - Morgan Stanley, Research Division
Christopher Muse - Cantor Fitzgerald & Co., Research Division
Vivek Arya - BofA Securities, Research Division
Benjamin Reitzes - Melius Research LLC
James Schneider - Goldman Sachs Group, Inc., Research Division
Timothy Arcuri - UBS Investment Bank, Research Division
Stacy Rasgon - Sanford C. Bernstein & Co., LLC., Research Division
Aaron Rakers - Wells Fargo Securities, LLC, Research Division

Presentation

Operator

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

Toshiya Hari, you may begin your conference.

Toshiya Hari
Vice President of Investor Relations & Strategic Finance

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2026. 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 2026.

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 19, 2025, 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 measures to GAAP financial measures in our CFO Commentary which is posted on our website.

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

Colette Kress
Executive VP & CFO

Thank you, Toshiya. We delivered another outstanding quarter with revenue of $57 billion, up 62% and a record sequential revenue growth of $10 billion or 22%. Our customers continued to lean into 3 platform shifts, fueling exponential growth for accelerated computing, powerful AI models and agentic applications yet we are still in the early innings of these transitions that will impact our work across every industry.

We currently have visibility to $0.5 trillion in Blackwell and Rubin revenue from the start of this year through the end of calendar year 2026. By executing our annual product cadence and extending our performance leadership through full stack design, we believe NVIDIA will be the superior choice for the $3 trillion to $4 trillion in annual AI infrastructure build we estimate by the end of the decade.

Demand for AI infrastructure continues to exceed our expectations. The clouds are sold out and our GPU installed base, both new and previous generations, including Blackwell, Hopper and Ampere is fully utilized. Record Q3 data center revenue of $51 billion (sic) [$51.2 billion] increased 66% year-over-year, a significant feat at our scale. Compute grew 56% year-over-year, driven primarily by the GB300 ramp, while networking more than doubled, given the onset of NVLink scale up and robust double-digit growth across Spectrum-X Ethernet and Quantum-X InfiniBand.

The world hyperscalers, a trillion-dollar industry are transforming search, recommendations and content understanding from classical machine learning to generative AI. NVIDIA CUDA excels at both and is the ideal platform for this transition, driving infrastructure investment measured in hundreds of billions of dollars.

At Meta, AI recommendation systems are delivering higher quality and more relevant content, leading to more time spent on apps such as Facebook and Threads, any expectations for the top CSPs and hyperscalers in 2026, aggregate CapEx have continued to increase and now sit roughly at $600 billion, more than $200 billion higher relative to the start of the year.

We see the transition to accelerate computing and generative AI across current hyper workloads contributing toward roughly half of our long-term opportunity. Another growth pillar is the ongoing increase in compute spend driven by foundation model builders such as Anthropic, Mistral, OpenAI, Reflection, Safe Superintelligence, Thinking Machines Lab and xAI, all scaling, compute aggressively to scale intelligence.

The 3 scaling laws, pretraining post training and inference remain intact. In fact, we see a positive virtuous cycle emerging whereby the 3 scaling laws and access to compute are generating better intelligence and in turn, increasing adoption and profits.

OpenAI recently shared that their weekly user base has grown to 800 million. Enterprise customers has increased to 1 million and that their gross margins were healthy. While Anthropic recently reported that its annualized run rate revenue has reached $7 billion as of last month, up from $1 billion at the start of the year.

We are also witnessing a proliferation of agentic AI across various industries and tasks. Companies such as Cursor, Anthropic, OpenEvidence, Epic and Abridge are experiencing a surge in user growth as they supercharge the existing workforce, delivering unquestionable ROI for coders and health care professionals.

The world's most important enterprise software platforms like ServiceNow, CrowdStrike, and SAP are integrating NVIDIA's accelerated computing and AI stack. Our new partner, Palantir, is supercharging the incredibly popular oncology platform with NVIDIA CUDA-X libraries and AI models for the first time.

Previously, like most enterprise software platforms, Anthology runs only on CPUs. Lowe's is leveraging the platform to build supply chain agility, reducing costs and improving customer satisfaction. Enterprises broadly are leveraging AI to boost productivity, increase efficiency and reduce cost. RBC is leveraging agentic AI to drive significant analyst productivity slashing, report generation, time from hours to minutes. AI and digital twins are helping Unilever accelerate content creation by 2x and cut costs by 50%.

And Salesforce's engineering team has seen at least 30% productivity increase in new co-development after adopting Cursor. This past quarter, we announced AI factory and infrastructure projects amounting to an aggregate of 5 million GPUs. This demand spans every market, CSPs, sovereigns, modern builders, enterprises and supercomputing centers and includes multiple landmark build-outs. xAI's Colossus 2, the world's first gigawatt scale data center, Lilly's AI Factory for drug discovery, the pharmaceutical industry's most powerful data center.

And just today, AWS and HUMAIN expanded their partnership, including the deployment of up to 150,000 AI accelerators, including our GB300, xAI and HUMAIN also announced a partnership in which the 2 will jointly develop a network of world-class GPU data centers anchored by the flagship 500-megawatt facility.

Blackwell gained further momentum in Q3, as GB300 crossed over GB200 and contributed roughly 2/3 of the total Blackwell revenue. The transition to GB300 has been seamless with production shipments to the majority -- to the major cloud service providers, hyperscalers and [GP clouds] and is already driving their growth.

The Hopper platform in its 13th quarter since inception, recorded approximately $2 billion in revenue in Q3. H20 sales were approximately $50 million, sizable purchase orders never materialized in the quarter due to geopolitical issues and the increasingly competitive market in China. While we were disappointed in the current state that prevents us from shipping more competitive data center compute products to China, we are committed to continued engagement with the U.S. and China governments and will continue to advocate for America's ability to compete around the world.

To establish a sustainable leadership position in AI computing, America must win the support of every developer and be the platform of choice for every commercial business, including those in China. The Rubin platform is on track to ramp in the second half of 2026. Powered by 7 chips, the Vera Rubin platform will once again deliver an X-factor improvement in performance relative to Blackwell. We have received silicon back from our supply chain partners and are happy to report that NVIDIA teams across the world are executing to bring up beautifully.

Rubin is our third-generation rack-scale system substantially redefined the manufacturability while remaining compatible with Grace Blackwell. Our supply chain data center ecosystem and cloud partners have now mastered the build to installation process of NVIDIA's rack architecture. Our ecosystem will be ready for a fast Rubin ramp.

Our annual X factor performance leap increases performance per dollar while driving down computing costs for our customers. The long useful life of NVIDIA's CUDA GPUs is a significant TCO advantage over accelerators. CUDA's compatibility in our massive installed base, extend the life NVIDIA Systems well beyond their original estimated useful life. For more than 2 decades, we have optimized the CUDA ecosystem, improving existing workloads, accelerating new ones and increasing throughput with every software release.

Most accelerators without CUDA and NVIDIA's time-tested and versatile architecture became obsolete within a few years as model technologies evolve. Thanks to CUDA, the A100 GPUs we shipped 6 years ago are still running at full utilization today, powered by vastly improved software stack.

We have evolved over the past 25 years from a gaming GPU company to now an AI data center infrastructure company. Our ability to innovate across the CPU, the GPU, networking and software and ultimately drive down cost per token is unmatched across the industry. Our networking business purpose built for AI and now the largest in the world, generated revenue of $8.2 billion, up 162% year-over-year with NVLink, InfiniBand and Spectrum-X Ethernet, all contributing to growth.

We are winning in data center networking, as the majority of AI deployments now include our switches with Ethernet GPU attach rates roughly on par with InfiniBand. Meta, Microsoft, Oracle and xAI are building gigawatt AI factories with Spectrum-X Ethernet switches and each will run its operating system of choice, highlighting the flexibility and openness of our platform.

We recently introduced Spectrum-XGS, a scale across technology that enables gigascale AI factories NVIDIA is the only company with AI scale up, scale out and scale across platforms, reinforcing our unique position in the market as the AI infrastructure provider.

Customer interest in NVLink Fusion continues to grow. We announced a strategic collaboration with Fujitsu in October, where we will integrate Fujitsu's CPUs and NVIDIA GPUs via and NVLink Fusion, connecting our large ecosystems. We also announced a collaboration with Intel to develop multiple generations of custom data center and PC products, connecting NVIDIA and Intel's ecosystems using NVLink.

This week at Supercomputing '25, Arm announced that it will be integrating NVLink IP for customers to build CPU SoCs that connect with NVIDIA. Currently on its fifth generation, NVLink is the only proven scale up technology available on the market today. In the latest MLPerf training results, Blackwell Ultra delivered 5x faster time to train than Hopper. NVIDIA swept every benchmark.

Notably, NVIDIA is the only training platform to leverage FP4 while meeting the MLPerf's strict accuracy standards. In semi-analysis, InferenceMAX benchmark, Blackwell achieved the highest performance and lowest total cost of ownership across every model and use case. Particularly important is Blackwell's NVLinks performance on a mixture of experts, the architecture for the world's most popular reasoning models.

On DeepSeek-R1 Blackwell delivered 10x higher performance per watt and 10x lower cost per token versus H200, a huge generational leap fueled by our extreme co-design approach. NVIDIA Dynamo, an open source, low latency modular inference framework has now been adopted by every major cloud service provider, leveraging Dynamo's enablement and disaggregated inference, the resulting increase in performance of complex AI models, such as MoE models, AWS, Google Cloud, Microsoft Azure and OCI have boosted AI inference performance for enterprise cloud customers.

We are working on a strategic partnership with OpenAI, focused on helping them build and deploy at least 10 gigawatts of AI data centers. In addition, we have the opportunity to invest in the company. We serve OpenAI through their cloud partners, Microsoft Azure, OCI and CoreWeave. We will continue to do so for the foreseeable future. As they continue to scale, we are delighted to support the company to add self-build infrastructure, and we are working towards a definitive agreement and are excited to support OpenAI's growth.

Yesterday, we celebrated an announcement with Anthropic. For the first time, Anthropic is adopting NVIDIA, and we are establishing a deep technology partnership to support Anthropic's fast growth. We will collaborate to optimize Anthropic models for CUDA and deliver the best possible performance, efficiency and TCO. We will also optimize future NVIDIA architectures for Anthropic workloads. Anthropic's compute commitment is initially including up to 1 gigawatt of compute capacity with Grace Blackwell and Vera Rubin Systems.

Our strategic investments in Anthropic, Mistral, OpenAI, Reflection, Thinking Machines and other represent partnerships that grow the NVIDIA CUDA AI ecosystem and enable every model to run optimally on NVIDIAs everywhere. We will continue to invest strategically while preserving our disciplined approach to cash flow management. Physical AI is already a multibillion-dollar business addressing a multitrillion dollar opportunity on the next leg of growth for NVIDIA.

Leading U.S. manufacturers and robotics innovators are leveraging NVIDIA's 3 computer architecture to train on NVIDIA, test on Omniverse's computer and deploy real-world AI and just in robotic computers. PTC and Siemens introduced new services that bring Omniverse powered digital twin workflows to their extensive installed base of customers. Companies, including Belden, Caterpillar, Foxconn, Lucid Motors, Toyota, TSMC and Wistron are building Omniverse digital twin factories to accelerate AI-driven manufacturing and automation.

Agility Robotics, Amazon Robotics, Figure and Skild at AI are building our platform, tapping offerings such as NVIDIA, Cosmos, World Foundation Models for development, Omniverse for simulation and validation and Jetson to power next-generation intelligent robots.

We remain focused on building resiliency and redundancy in our global supply chain. Last month, in partnership with TSMC, we celebrated the first Blackwell wafer produced on U.S. soil. We will continue to work with Foxconn, Wistron, Amkor, SPIL and others to grow our presence in the U.S. over the next 4 years. Gaming revenue was $4.3 billion, up 30% year-on-year, driven by strong demand as Blackwell momentum continued.

End market sell-through remains robust and channel inventories are at normal levels heading into the holiday season. Steam recently broke its concurrent user record with 42 million gamers, while thousands of fans pack the GeForce Gamer Festival in South Korea to celebrate 25 years of GeForce.

NVIDIA pro visualization has evolved into computers for engineers and developers, whether for graphics or for AI. Professional Visualization revenue was $760 million, up 56% year-over-year, was another record. Growth was driven by DGX Spark, the world's smallest AI supercomputer, built on a small configuration of Grace Blackwell. Automotive revenue was $592 million, up 32% year-over-year, primarily driven by self-driving solutions. We are partnering with Uber to scale the world's largest Level 4 ready autonomous fleet built on the new NVIDIA Hyperion L4 robotaxi reference architecture.

Moving to the rest of the P&L. GAAP gross margins were 73.4% and non-GAAP gross margins was 73.6%, exceeding our outlook. Gross margins increased sequentially due to our data center mix, improved cycle time and cost structure. GAAP operating expenses were up 8% sequentially and up 11% on a non-GAAP basis. The growth was driven by infrastructure compute as well as higher compensation and benefits and engineering development costs.

Non-GAAP effective tax rate for the third quarter was just over 17% higher than our guidance of 16.5% due to the strong U.S. revenue. On our balance sheet, inventory grew 32% quarter-over-quarter, while supply commitments increased 63% sequentially. The we are preparing for significant growth ahead and feel good about our ability to execute against our opportunity set.

Okay. Let me turn to the outlook for the fourth quarter. Total revenue is expected to be $65 billion, plus or minus 2%. At the midpoint, our outlook implies 14% sequential growth driven by continued momentum in the Blackwell architecture. Consistent with last quarter, we are not assuming any data center compute revenue from China. GAAP and non-GAAP gross margins are expected to be 74.8% and 75%, respectively, plus or minus 50 basis points.

Looking ahead to fiscal year 2027, and input costs are on the rise, but we are working to hold gross margins in the mid-70s. GAAP and non-GAAP operating expenses are expected to be approximately $6.7 billion and $5 billion, respectively. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $500 million, excluding gains and losses from nonmarketable and publicly held equity securities. GAAP and non-GAAP tax rates are expected to be 17%, plus or minus 1%, excluding any discrete items.

At this time, let me turn the call over to Jensen for him to say a few words.

Jen-Hsun Huang
Co-Founder, CEO, President & Director

Thanks, Colette. There's been a lot of talk about an AI bubble. From our vantage point, we see something very different. As a reminder, NVIDIA is unlike any other accelerator. We excel at every phase of AI from pre-training and post training to inference. And with our 2-decade investment in CUDA-X acceleration libraries, we are also exceptional at science and engineering simulations, computer graphics, structured data processing to classical machine learning.

The world is going -- is undergoing 3 massive platform shifts at once. The first time since the dawn of Moore's Law, NVIDIA is uniquely addressing each of the 3 transformations. The first transition is from CPU general purpose computing to GPU accelerated computing and Moore's Law slows. The world has a massive investment in non-AI software from data processing to science and engineering simulations, representing hundreds of billions of dollars in compute -- cloud computing spend each year.

Many of these applications, which ran once exclusively on CPUs are now rapidly shifting to CUDA GPUs. Accelerated computing has reached a tipping point. Secondly, AI has also reached a tipping point and is transforming existing applications while enabling entirely new ones. For existing applications, generative AI is replacing classical machine learning in search ranking, recommender systems, ad targeting, click-through prediction to content moderation. The very foundations of hyperscale infrastructure.

Meta's GEM, a foundation model for ad recommendations trained on large-scale GPU clusters exemplifies this shift. In Q2, Meta reported over a 5% increase in ad conversions on Instagram and 3% gain on Facebook feed driven by generative AI-based GEM. Transitioning to generative AI represents substantial revenue gains for hyperscalers.

Now a new wave is rising, agentic AI systems capable of reasoning, planning and using tools from coding assistance like Cursor and Claude Code to radiology tools like Aidoc, legal assistants like Harvey and AI chauffeurs like Tesla FSD and Waymo. These systems mark the next frontier of computing, the fastest-growing companies in the world today, OpenAI, Anthropic, xAI, Google, Cursor, Lovable, Replit, Cognition AI, OpenEvidence, Abridge, Tesla are pioneering agentic AI.

So there are 3 massive platform shifts. The transition to accelerated computing is foundational and necessary, essential in a post-Moore's Law era. The transition to generative AI is transformational and necessary, supercharging existing applications and business models. And the transition to agentic and physical AI will be revolutionary, giving rise to new applications, companies, products and services.

As you consider infrastructure investments, consider these 3 fundamental dynamics, each will contribute to infrastructure growth in the coming years. NVIDIA is chosen because our singular architecture enables all 3 transitions. And thus so, for any form and modality of AI across all industries, across every phase of AI, across all of the diverse computing needs in the cloud and also from cloud to enterprise to robots, one architecture.

Toshiya, back to you.

Toshiya Hari
Vice President of Investor Relations & Strategic Finance

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

Question-and-Answer Session

Operator

[Operator Instructions] Your first question comes from Joseph Moore with Morgan Stanley.

Joseph Moore
Morgan Stanley, Research Division

I wonder if you could update us -- you talked about the $500 billion of revenue for Blackwell plus Rubin in '25 and '26 at GTC. At that time, you talked about $150 billion of that already having been shipped. So as the quarter is wrapped up, are those still kind of the general parameters that there's $350 billion in the next kind of 14 months or so? And I would assume over that time, you haven't seen all the demand that there is, there's any possibility of upside to those numbers as we move forward.

Colette Kress
Executive VP & CFO

I'll start first with a response here on that. Yes, that's correct. We are working into our $500 billion forecast. And we are on track for that as we have finished some of the quarters, and now we have several quarters now in front of us to take us through the end of calendar year '26. The number will grow. And we will achieve, I'm sure, additional needs for compute that will be shippable by fiscal year '26. So we shipped $50 billion this quarter, but we would be not finished if we didn't say that we'll probably be taking more orders.

For example, just even today, our announcements with KSA, and that agreement in itself is 400,000 to 600,000 more GPUs over 3 years. Anthropic is also net new. So there's definitely an opportunity for us to have more on top of the $500 billion that we announced.

Operator

The next question comes from CJ Muse with Cantor Fitzgerald.

Christopher Muse
Cantor Fitzgerald & Co., Research Division

There's clearly a great deal of consternation around the magnitude of AI infrastructure build-outs and the ability to fund such plans and the ROI yet at the same time, you're talking about being sold out, every stood up GP is taken. The AI world hasn't seen the enormous benefit yet from B300, never mind Rubin, and Gemini 3 just announced Grok 5 coming soon. And so the question is this, when you look at that as the backdrop do you see a realistic path for supply to catch up with demand over the next 12 to 18 months? Or do you think it can extend beyond that time frame?

Jen-Hsun Huang
Co-Founder, CEO, President & Director

Well, as you know, we've done a really good job planning our supply chain. NVIDIA supply chain basically includes every technology company in the world. And TSMC and their packaging and our memory vendors -- memory partners and all of our system ODMs have done a really good job planning with us. And we were planning for a big year.

We've seen for some time, the 3 transitions that I spoke about just a second ago, accelerated computing from general-purpose computing. And it's really important to recognize that AI is not just agentic AI but generative AI is transforming the way that hyperscalers did the work that they used to do on CPUs.

Generative AI made it possible for them to move search and recommender systems and ad recommendations and targeting, all of that has been moved to generative AI and still transitioning. And so whether you install NVIDIA GPUs for data processing or you did it for generative AI for your recommender system or you're building it for agentic chatbots and the type of AIs that most people see when they think about AI, all of those applications are accelerated by NVIDIA.

And so when you look at the totality of the spend, it's really important to think about each 1 of those layers. They're all growing. They're related, but not the same, but the wonderful thing is that they all run on NVIDIA GPUs.

Simultaneously, because the quality of the AI models are improving so incredibly. The adoption of it in the different use cases, whether it's in code assistance, which NVIDIA uses fairly exhaustively, and we're not the only one. I mean, the fastest-growing application in history, a combination of Cursor and Claude Code and code -- OpenAI's Codex and GitHub CoPilot.

These applications are the fastest-growing in history. And it's not just used for software engineers, it's used by -- because of wide coding is used by engineers and marketeers all over companies, supply chain planners, all over companies.

And so I think that that's just 1 example, and the list goes on, whether it's OpenEvidence and the work that they do in health care or the work that's being done in digital video editing, Runway and -- I mean a number of really, really exciting start-ups that are taking advantage of generative AI and agentic AI is growing quite rapidly. And not to mention we're all using it a lot more.

And so all of these exponentials, not to mention just today, I was reading a text from Denis. And he was saying that pre-training and post training are fully intact. And Gemini 3 takes advantage of the scaling laws and got to receive a huge jump in quality performance -- model performance. And so we're seeing all of these exponentials kind of running at the same time.

And just always go back to first principles and think about what's happening from each one of the dynamics that I mentioned before, general purpose computing to accelerated computing, generative AI replacing classical machine learning and, of course, agentic AI, which is a brand-new category.

Operator

The next question comes from Vivek Arya with Bank of America Securities.

Vivek Arya
BofA Securities, Research Division

I'm curious, what assumptions are you making on NVIDIA content per gigawatt in that $500 billion number? Because we have heard numbers as low as $25 billion per gigawatt of content to as high as $30 billion or $40 billion per gigawatt. So I'm curious what power and what dollar per gig assumptions you are making as part of that $500 billion number.

And then longer term, Jensen, the $3 trillion to $4 trillion in data center by 2030 was mentioned. How much of that do you think will require vendor financing? And how much of that can be supported by cash flows of your large customers or governments or enterprises?

Jen-Hsun Huang
Co-Founder, CEO, President & Director

In each generation, from Ampere to Hopper, from Hopper to Blackwell, Blackwell to Rubin, our part of the data center increases. And Hopper generation was probably something along the lines of 20-some-odd, 20 to 25. Blackwell generation, Grace Blackwell particularly is probably 30 to 30 to say, 30 plus or minus and then Rubin is probably higher than that.

And in each one of these generations, the speed up is X factors. And therefore, their TCO, the customer TCO, improves by X factors, and the most important thing is, in the end, you still only have 1 gigawatt of power. One gigawatt data centers, 1 gigawatt power. And therefore, performance per watt, the efficiency of your architecture is incredibly important. And the efficiency of your architecture can't be brute force. There is no brute forcing about it.

That 1 gigawatt translates directly. Your performance per watt translates directly absolutely directly to your revenues, which is the reason why choosing the right architecture matters so much now. The world doesn't have an excess of anything to squander. And so we have to be really, really -- we use this concept called codesign across our entire stack, across the frameworks and models, across the entire data center, even power and cooling, optimized across the entire supply chain or ecosystem.

And so each generation, our economic contribution will be greater. Our value delivered will be greater but the most important thing is our energy efficiency per watt is going to be extraordinary, every single generation. With respect to growing into -- continuing to grow, our customers' financing is up to them. We are -- we see the opportunity to grow for quite some time. And remember, today, most of the focus has been on the hyperscalers.

And one of the areas that is really misunderstood about the hyperscalers is that the investment on NVIDIA GPUs not only improves their scale, speed and cost for -- from general purpose computing. That's number 1, because Moore's Law saw scaling has really slowed. Moore's Law is about driving cost down. It's about deflationary cost, the incredible deflationary cost of computing over time. But that has slowed. Therefore, a new approach is necessary for them to keep driving the cost down. Going to NVIDIA GPU computing is really the best way to do so.

The second is revenue boosting in their current business models, recommender systems drive the world's hyperscalers. Every single -- whether it's watching short-form videos or recommending books or recommending the next item in your basket to recommending ads to recommending news to -- it's all about recommenders. The world has -- the Internet has trillions of pieces of content, how could they possibly figure out what to put in front of you and your little tiny screen, unless they have really sophisticated recommender systems to do so.

Well, that has gone generative AI, so the first 2 things that I've just said, hundreds of billions of dollars of CapEx that's going to have to be invested is fully cash flow funded. What is above it, therefore, is agentic AI. This is revenue -- this is net new, net new consumption, but it's also net new applications and some of the applications I mentioned before, but these are -- these new applications are also the fastest-growing applications in history, okay?

So I think that you're going to see that once people start to appreciate what is actually happening under the water, if you will, from the simplistic view of what's happening to CapEx investment, recognizing there's these 3 dynamics.

And then lastly, remember, we were just talking about the American CSPs. Each country will fund their own infrastructure. And you have multiple countries, you have multiple industries. Most of the world's industries haven't really engaged agentic AI yet, and they're about to. All the names of companies that you know we're working with, whether it's autonomous vehicle companies or digital twins for physical AI for factories and the number of factories and warehouses being built around the world, just a number of digital biology start-ups that are being funded so that we could accelerate drug discovery.

All of those different industries are now getting engaged, and they're going to do their own fundraising. And so don't just look at the hyperscalers as a way to build out for the future. You got to look at the world, you got to look at all the different industries and enterprise computing is going to fund their own industry.

Operator

The next question comes from Ben Reitzes with Melius.

Benjamin Reitzes
Melius Research LLC

Jensen, I wanted to ask you about cash. Speaking of $0.5 trillion, you may generate about $0.5 trillion in free cash flow over the next couple of years. What are your plans for that cash? How much goes to buyback versus investing in the ecosystem? And how do you look at investing in the ecosystem? I think there's just a lot of confusion out there about how these deals work and your criteria for doing those like the Anthropic, the OpenAI's, et cetera.

Jen-Hsun Huang
Co-Founder, CEO, President & Director

Yes, I appreciate the question. Of course, using cash to fund our growth, no company has grown at the scale that we're talking about and have the connection and the depth and the breadth of supply chain that NVIDIA has. The reason why our entire customer base can rely on us is because we've secured a really resilient supply chain, and we have the balance sheet to support them.

When we make purchases, our suppliers can take it to the bank. When we make forecast and we plan with them, they take us seriously because of our balance sheet. We're not making up the offtake. We know what our offtake is, and because they've been planning with us for so many years, our reputation and our credibility is incredible. And so it takes really strong balance sheet to do that, to support the level of growth and the rate of growth and the magnitude associated with that. So that's number one.

The second thing, of course, we're going to continue to do stock buybacks. We're going to continue to do that. But with respect to the investments, this is really, really important work that we do. All of the investments that we've done so far, all the period, is associated with expanding the reach of CUDA expanding the ecosystem. If you look at the work that -- the investments that we did with OpenAI.

Of course, that relationship we've had since 2016, I delivered the first AI supercomputer ever made to OpenAI. And so we've had a close and wonderful relationship with OpenAI since then. And everything that OpenAI does runs on NVIDIA today. So all the clouds that they deploy in, whether it's training and inference runs NVIDIA and we love working with them.

The partnership that we have with them is one, so that we could work even deeper from a technical perspective so that we could support their accelerated growth. This is a company that's growing incredibly fast. And don't just look at what is said in the press, look at all the ecosystem partners and all the developers that are connected to OpenAI, and they're all driving consumption of it. and the quality of the AI that's being produced, huge step-up since a year ago. And so the quality of response is extraordinary.

So we invest in OpenAI for a deep partnership in co-development to expand our ecosystem and support their growth. And of course, rather than giving up a share of our company, we get a share of their company. And we invested in them, in one of the most consequential once-in-a-generation companies that we have a share of. And so I fully expect that investment to translate to extraordinary returns.

Now in the case of Anthropic, this is the first time that Anthropic will be on NVIDIA's architecture. The first time Anthropic will be on NVIDIA's architecture is the second most successful AI in the world in terms of total number of users. But in enterprise, they're doing incredibly well. Claude Code is doing incredibly well. Claude Code is doing incredibly well all of the world's enterprise. And now we have the opportunity to have a deep partnership with them and bringing Claude onto the NVIDIA platform.

And so what do we have now? NVIDIA's architecture, taking a step back, NVIDIA's architecture, NVIDIA's platform is the singular platform in the world that runs every AI model. We run OpenAI, we run Anthropic, we run xAI because of our deep partnership with Elon and xAI, we were able to bring that opportunity to Saudi Arabia to the KSA so that HUMAIN could also be hosting opportunity for xAI. We run xAI, we run Gemini, we run Thinking Machines, let's see, what else do we run?

We've run them all. And so not to mention, we run the science models, the biology models, DNA models, gene models, chemical models and all the different fields around the world. It's not just cognitive AI that the world uses, AI is impacting every single industry.

And so we have the ability to the ecosystem investments that we make to partner with -- deeply partner on a technical basis with some of the best companies, most brilliant companies in the world, we are expanding the reach of our ecosystem, and we're getting a share and investment in what will be a very successful company, oftentimes once in a generation company. And so that basic -- that's our investment thesis.

Operator

The next question comes from Jim Schneider with Goldman Sachs.

James Schneider
Goldman Sachs Group, Inc., Research Division

In the past, you've talked about roughly 40% of your shipments tied to AI inference. I'm wondering, as you look forward into next year, where do you expect that percentage could go in, say, a year's time? And can you maybe address the Rubin CPX product you expect to introduce next year or contextualize that, how big of the overall TAM you expect that can take? And maybe talk about some of the target customer applications for that specific product.

Jen-Hsun Huang
Co-Founder, CEO, President & Director

CPX is designed for long context type of workload generation. And so long context, basically, before you start generating answers, you have to read a lot, basically long context. And it could be a bunch of PDFs. It could be watching a bunch of videos, studying 3D images, so on and so forth. You have to absorb the context. And so CPX is designed for a long context type of workloads. And it's perf-per-dollar excellent, it's perf-per-watt is excellent. And which -- maybe forget the first part of the question.

Colette Kress
Executive VP & CFO

Inferencing.

Jen-Hsun Huang
Co-Founder, CEO, President & Director

Oh inferencing, yes, there are 3 scaling laws that are scaling at the same time. The first scaling law called pretraining continues to be very effective. And the second is post-training. Post-training basically has found incredible algorithms for improving an AI's ability to break a problem down and solve a problem step by step. And post-training is scaling exponentially, basically, the more compute you apply to a model, the smarter it is, the more intelligent it is.

And then the third is inference. Inference because of chain of thought, because of reasoning capabilities, AIs are essentially reading, thinking before it answers. And the amount of computation necessary as a result of those 3 things has gone completely exponential. I think that it's hard to know exactly what the percentage of it will be at any given point in time and who.

But of course, our hope is that inference is a very large part of the market because if inference is large, then what it suggests is that people are using it in more applications and they're using it more frequently. And that's -- we should all hope for inference to be very large. And this is where Grace Blackwell is just an order of magnitude better more advanced than anything in the world.

The second best platform is H200, and it's very clear now that GB300, GB200and GB300 because of NVLink 72, the scale-up network that we have achieved. And you saw and Colette talked about in the semi analysis benchmark. It's the largest single inference benchmark ever done and GB200, NVLink 72 is 10x, 10 to 15x higher performance. And so that's a big step up. It's going to take a long time before somebody is able to take that on. And our leadership there is surely multiyear. And so I think I'm hoping that inference becomes a very big deal. Our leadership in inference is extraordinary.

Operator

The next question comes from Timothy Arcuri with UBS.

Timothy Arcuri
UBS Investment Bank, Research Division

Jensen, many of your customers are pursuing behind-the-meter power, but like what's the single biggest bottleneck that worries you that could constrain your growth? Is it power? Or maybe it's financing or maybe it's something else like memory or even foundry?

Jen-Hsun Huang

Co-Founder, CEO, President & Director

Well, these are all issues and they're all constraints. And the reason for that, when you're growing at the rate that we are and the scale that we are, how could anything be easy? What NVIDIA is doing obviously has never been done before. And we've created a whole new industry.

Now on the one hand, we are transitioning computing from general purpose and classical or traditional computing to accelerated computing and AI. That's on the one hand. On the other hand, we created a whole new industry called AI factories. The idea that in order for software to run, you need these factories to generate it, generate every single token instead of retrieving information that was pre-created. And so I think this whole transition requires extraordinary scale.

And all the way from the supply chain. Of course, the supply chain, we have much better visibility and control over because obviously, we're incredibly good at managing our supply chain. We have great partners that we've worked with for 33 years. And so the supply chain part of it, we're quite confident. Now looking down our supply chain, we've now established partnerships with so many players in land and power and shell.

And of course, financing. These things -- none of these things are easy, but they're all attractable and they're all solvable things. And the most important thing that we have to do is do a good job planning we plan up the supply chain, down the supply chain. We have established a whole lot of partners. And so we have a lot of routes to market. And very importantly, our architecture has to deliver the best value to the customers that we have.

And so at this point, I'm very confident that NVIDIA's architecture is the best performance per TCO, it is the best performance per watt. And therefore, for any amount of energy that is delivered, our architecture will drive the most revenues. And I think the increasing rate of our success, I think that we're more successful this year at this point than we were last year at this point.

The number of customers coming to us and the number of platforms coming to us after they've explored others, is increasing, not decreasing. And so I think the -- I think all of that is just -- all the things that I've been telling you over the years are really coming -- are coming through or becoming evident.

Operator

The next question comes from Stacy Rasgon with Bernstein Research.

Stacy Rasgon

Sanford C. Bernstein & Co., LLC., Research Division

Colette, I have some questions on margins. You said for next year, you're working to hold them in the mid-70s. So I guess, first of all, what are the biggest cost increases? Is it just memory or is it something else? What are you doing to work toward that? Is it -- how much is like cost optimizations versus prebuys versus pricing? And then also, how should we think about OpEx growth next year, given the revenues seem likely to grow materially from where we're running right now?

Colette Kress

Executive VP & CFO

Thanks, Stacy. Let me see if I can start with remembering where we were with the current fiscal year that we're in. Remember, earlier this year, we indicated that through cost improvements and mix that we would exit the year in our gross margins in the mid 70s. We've achieved that and getting ready to also execute that in Q4. So now it's time for us to communicate where are we working right now in terms of next year.

Next year, there are input prices that are well known in the industries that we need to work through. And our systems are by no means very easy to work with. There are tremendous amount of components in many different parts of it as we think about that. So we're taking all of that into account, but we do believe as we look at working again on cost improvement, cycle time and mix that we will work to try and hold at our gross margins in the mid-70s. So that's our overall plan for gross margin.

Your second question is around OpEx. And right now, our goal in terms of OpEx is to really make sure that we are innovating with our engineering teams with all of our business teams to create more and more systems for this market. As you know, right now, we have a new architecture coming out. And that means they are quite busy in order to meet that goal. And so we're going to continue to see our investments on innovating more and more both the software, both our systems and our hard work to do so. I'll leave it -- turn it to Jensen if he wants to add any couple of more comments.

Jen-Hsun Huang

Co-Founder, CEO, President & Director

Yes, that's spot on. I think the only thing that I would add is remember that we plan, we forecast, we plan and we negotiate with our supply chain well in advance. Our supply chain have known for quite a long time, our requirements. And they've known for quite a long time our demand, and we've been working with them and negotiating with them for quite a long time. And so I think the recent surge obviously quite significant.

But remember, our supply chain has been working with us for a very long time. So in many cases, we've secured a lot of supply for ourselves because, obviously, they're working with the largest company in the world in doing so. And we've also been working closely with them on the financial aspects of it and securing forecasts and plans and so on and so forth. So I think all of that has worked out well for us.

Operator

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

Aaron Rakers

Wells Fargo Securities, LLC, Research Division

Jensen, the question for you. As you think about the Anthropic deal that was announced and just the overall breadth of your customers, I'm curious if your thoughts around the role that AI ASICs or dedicated XPUs play in these architecture build-outs has changed at all? Have you seen, I think you've been fairly adamant in the past that some of these programs never really see deployments. But I'm curious if we're at a point where maybe that's even changed more in favor of just GPU architecture.

Jen-Hsun Huang

Co-Founder, CEO, President & Director

Yes. Thank you very much, and I really appreciate the question. So first of all, you're not competing against teams -- excuse me, against a company, you're competing against teams. And the -- there just aren't that many teams in the world who are built -- who are extraordinary at building these incredibly complicated things.

Back in the Hopper day and the Ampere days, we would build one GPU. That's the definition of an accelerated AI system. But today, we've got to build entire racks entire -- 3 different types of switches, scale up, scale out and scale across switch. And it takes a lot more than 1 chip to build a compute node anymore. Everything about that computing system because AI needs to have memory, AI didn't use to have memory at all.

Now it has to remember things, the amount of memory and context it has is gigantic. The memory architecture implication is incredible. The diversity of models from mixture of experts to dense models, to diffusion models that are aggressive not to mention biological models that are based on the laws of physics, the list of different types of models have exploded in the last several years.

And so the challenge is the complexity of the problem is much higher. The diversity of AI models is incredibly, incredibly large. And so this is where, if I will say, there are 5 things that makes us special, if you will. The first thing I would say that makes us special is that we accelerate every phase of that transition. That's the first space. CUDA allows us to have CUDA-X for transitioning from general purpose to accelerated computing.

We are incredibly good at generative AI. We're incredibly good at agentic AI. So every single phase of that -- every single layer of the transition, we are excellent at. You can invest in 1 architecture, use it across the board. You can use 1 architecture and not worry about the changes in the workload across those 3 phases. That's number one.

Number two, we're excellent at every phase of AI. Everybody's always known that we're incredibly good at pretraining. We're obviously very good at post-training, and we're incredibly good, as it turns out at inference because inference is really, really hard. How could thinking be easy? People think that inference is one shot and therefore, it's easy, anybody could approach the market that way. But it turns out to be the hardest of all because thinking as it turns out is quite hard. We're great at every phase of AI, the second thing.

The third thing is we're now the only architecture in the world that runs every AI model, every frontier AI model, we run open source AI models incredibly well. We run science models, biology models, robotics models. We run every single model. We're the only architecture in the world that can claim that. It doesn't matter whether you're auto regressive or diffusion based. We run everything and we run it for every major platform, as I just mentioned. So we run every model.

And then the fourth thing I would say is that we're in every cloud. The reason why developers love us is because we're literally everywhere. We're in every cloud, we're in every -- we can even make you a little tiny cloud called DGX Spark. And so we're in every computer, we're everywhere, from cloud to on-prem to robotic systems, edge devices, PCs, you name it. One architecture, things just work, it's incredible.

And then the last thing, and this is probably the most important thing, the fifth thing is, if you are a cloud service provider, if you're a new company like HUMAIN, if you're a new company like CoreWeave or Nscale or Nebius or OCI for that matter, the reason why NVIDIA is the best platform for you is because our offtake is so diverse. We can help you with offtake. It's not about just putting a random ASIC into a data center.

Where is the offtake coming from? Where is the diversity coming from? Where is the resilience coming from? The versatility of the architecture coming from, the diversity of capability coming from, NVIDIA has such incredibly good offtake because our ecosystem is so large. So these 5 things, every phase of acceleration and transition every phase of AI, every model, every cloud to on-prem. And of course, finally, it all leads to offtake.

Operator

Thank you. I will now turn the call to Toshiya Hari for closing remarks.

Toshiya Hari
Vice President of Investor Relations & Strategic Finance

In closing, please note, we will be at the UBS Global Technology and AI Conference on December 2 and our earnings call to discuss the results of our fourth quarter of fiscal 2026 is scheduled for February 25. Thank you for joining us today. Operator, please go ahead and close the call.

Operator

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

Read more current NVDA analysis and news

View all earnings call transcripts