Speaker 1
0s

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. All lines have been placed on mute to prevent any background noise. After the speaker's remarks, there will be a question-and-answer session. If you would like to ask a question during this time, simply press star, followed by the number one on your telephone keypad. If you would like to withdraw your question, press star one again. Thank you. Toshiya Hari, you may begin your conference.

Speaker 3
33s

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.

Speaker 3
1m 53s

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website. With that, let me turn the call over to Colette.

Speaker 2
2m 8s

Thank you, Toshiya. We delivered another outstanding quarter with revenue of $57 billion, up 62% year over year, and a record sequential revenue growth of $10 billion, or 22%. Our customers continue to lean into three 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 $500 billion 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-$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 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.

Speaker 2
3m 58s

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. Analyst 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 accelerated computing and generative AI across current hyperscale 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 three scaling laws, pre-training, post-training, and inference remain intact. In fact, we see a positive virtuous cycle emerging whereby the three 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 have increased to 1 million, and that their gross margins were healthy.

Speaker 2
6m 2s

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, Open Evidence, Epic, and Abridge are experiencing a surge in user growth as they supercharge the existing workforce, delivering unquestionable ROI for coders and healthcare 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 ontology platform with NVIDIA CUDA X libraries and AI models for the first time. Previously, like most enterprise software platforms, Ontology 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 costs. 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%. Salesforce's engineering team has seen at least a 30% productivity increase in new code development after adopting Cursor.

Speaker 2
8m 5s

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, model builders, enterprises, and supercomputing centers, and includes multiple landmark buildouts. 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. Just today, AWS and Humane expanded their partnership, including the deployment of up to 150,000 AI accelerators, including our GB300. xAI and Humane also announced a partnership in which the two 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 two-thirds of the total Blackwell revenue. The transition to GB300 has been seamless, with production shipments to the major cloud service providers, hyperscalers, and GPU 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.

Speaker 2
10m 19s

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 seven 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 the 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 and our massive installed base extend the life of NVIDIA systems well beyond their original estimated useful life. For more than two decades, we have optimized the CUDA ecosystem, improving existing workloads, accelerating new ones, and increasing throughput with every software release.

Speaker 2
12m 9s

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 six 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.

Speaker 2
14m 10s

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 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 bridge FP4 while meeting MLPerf's strict accuracy standards. In semi-analysis inference max benchmark, Blackwell achieved the highest performance and lowest total cost of ownership across every model and use case. Particularly important is Blackwell's NVLink's 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 code design approach.

Speaker 2
16m 8s

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 toward 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.

Speaker 2
18m 1s

Our strategic investments in Anthropic, Mistral, OpenAI, Reflection, Thinking Machines, and others represent partnerships that grow the NVIDIA CUDA AI ecosystem and enable every model to run optimally on NVIDIA's everywhere. We will continue to invest strategically while preserving our disciplined approach to cash flow management. Physical AI is already a multi-billion dollar business addressing a multi-trillion dollar opportunity and the next leg of growth for NVIDIA. Leading U.S. Manufacturers and robotics innovators are leveraging NVIDIA's three-computer architecture to train on NVIDIA, test on Omniverse computer, and deploy real-world AI on Jetson 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 Skilled 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.

Speaker 2
19m 55s

We will continue to work with Foxconn, Wistron, Amcor, Spill, and others to grow our presence in the U.S. over the next four 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 packed 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 were 73.6%, exceeding our outlook. Gross margins increased sequentially due to our data center mix, improved cycle time, and cost structure.

Speaker 2
21m 53s

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 in 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. 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, 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.

Speaker 2
23m 41s

GAAP and non-GAAP other income and expenses are expected to be an income of approximately $500 million, excluding gains and losses from non-marketable 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. Thanks, Colette. There has 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. With our two-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 undergoing three massive platform shifts at once, the first time since the dawn of Moore's Law. NVIDIA is uniquely addressing each of the three transformations. The first transition is from CPU general-purpose computing to GPU accelerated computing as 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.

Speaker 2
25m 37s

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 assistants like Cursor and Claude Code to radiology tools like iDoc, 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, Replet, Cognition AI, Open Evidence, Abridge, Tesla—are pioneering agentic AI. There are three massive platform shifts. The transition to accelerated computing is foundational and necessary, essential in a post-Moore's Law era.

Speaker 2
27m 57s

The transition to generative AI is transformational and necessary, supercharging existing applications and business models. 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 three fundamental dynamics. Each will contribute to infrastructure growth in the coming years. NVIDIA is chosen because our singular architecture enables all three 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 a cloud, and also from cloud to enterprise to robots. One architecture. Toshiya, back to you. We will now open the call for questions. Operator, would you please pull for questions? Thank you. At this time, I would like to remind everyone in order to ask a question, press star, then the number one on your telephone keypad. We'll pause for just a moment to compile the Q&A roster. As a reminder, please limit yourself to one question. Thank you. Your first question comes from Joseph Moore with Morgan Stanley. Your line is open. Great. Thank you. I wonder if you could update us. You talked about the $500 billion of revenue for Blackwell plus Rubin in 2025 and 2026 at GTC.

Speaker 2
29m 46s

At that time, you had talked about $150 billion of that already having been shipped. As the quarter's wrapped up, are those still kind of the general parameters that there's $350 billion in the next kind of 14 months or so? I would assume over that time, you haven't seen all the demand, but there is any possibility of upside to those numbers as we move forward. Yeah. Thanks, Joe. 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. We have several quarters now in front of us to take us through the end of calendar year 2026. The number will grow, and we will achieve, I'm sure, additional needs for compute that will be shippable by fiscal year 2026. We shipped $50 billion this quarter, but we would be not finished if we did not say that we will probably be taking more orders. For example, just even today, our announcements with KSA and that agreement in itself is 400,000-600,000 more GPUs over three years. Anthropic is also not new. There is definitely an opportunity for us to have more on top of the $500 billion that we announced.

Speaker 2
31m 15s

The next question comes from C.J. Muse with Cantor Fitzgerald. Your line is open. Yeah. Good afternoon. Thank you for taking the question. There's clearly a great deal of consternation around the magnitude of AI infrastructure buildouts 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 GPU is taken. The AI world hasn't seen the enormous benefit yet from B300, never mind Rubin. Gemini 3 just announced Grok 5 coming soon. 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 timeframe? As you know, we've done a really good job planning our supply chain. NVIDIA's supply chain basically includes every technology company in the world. TSMC and their packaging and our memory vendors and memory partners and all of our system ODMs have done a really good job planning with us. We were planning for a big year. We've seen for some time the three transitions that I spoke about just a second ago: accelerated computing from general-purpose computing.

Speaker 2
32m 37s

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 add recommendations and targeting. All of that has been moved to generative AI and is still transitioning. Whether you installed 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. When you look at the totality of the spend, it's really important to think about each one of those layers. They're all growing. They're related, but not the same. 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 OpenAI's Codex and GitHub Copilot, these applications are the fastest-growing in history.

Speaker 2
34m 16s

It's not just used for software engineers. It's used because of vibe coding. It's used by engineers and marketeers all over companies, supply chain planners all over companies. I think that that's just one example, and the list goes on, whether it's open evidence and the work that they do in healthcare or the work that's being done in digital video editing, runway. I mean, the number of really, really exciting startups that are taking advantage of generative AI and agentic AI is growing quite rapidly. Not to mention, we're all using it a lot more. All of these exponentials, not to mention, just today, I was reading a text from Demis, and he was saying that pre-training and post-training are fully intact. Gemini 3 takes advantage of the scaling laws and got a received a huge jump in quality performance and model performance. We are seeing all of these exponentials kind of running at the same time. I 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.

Speaker 2
35m 40s

The next question comes from Vivek Aria with Bank of America Securities. Your line is open. Thanks for taking my question. 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 or $40 billion per gigawatt. I am curious what power and what dollar per gigawatt assumptions you are making as part of that $500 billion number. Longer-term, Jensen, the $3 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? Thank you. In each generation, from Ampere to Hopper, from Hopper to Blackwell, Blackwell to Rubin, our part of the data center increases. Hopper generation was probably something along the lines of 20-somewhat, 20-25. Blackwell generation, Grace Blackwell particularly, is probably 30-30, say 30 plus or minus. Rubin is probably higher than that. In each one of these generations, the speedup is X factors. Therefore, their TCO, the customer TCO, improves by X factors.

Speaker 2
37m 19s

The most important thing is, in the end, you still only have one gigawatt of power, one gigawatt data centers, one gigawatt of power. Therefore, performance per watt, the efficiency of your architecture, is incredibly important. The efficiency of your architecture can't be brute forced. There is no brute forcing about it. That one 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. We have to be really, really—we use this concept called co-design across our entire stack, across the frameworks and models, across the entire data center, even power and cooling optimized across the entire supply chain in our ecosystem. Each generation, our economic contribution will be greater. Our value delivered will be greater. 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 see the opportunity to grow for quite some time. Remember, today, most of the focus has been on the hyperscalers. 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 from general-purpose computing—that is number one—because Moore's Law scaling has really slowed.

Speaker 2
39m 16s

Moore's Law is about driving cost down. It's about deflationary cost, the incredible deflationary cost of computing over time. 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 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? That has gone generative AI. The first two things that I've just said, hundreds of billions of dollars of CapEx is going to have to be invested, is fully cash flow funded. What is above it, therefore, is agentic AI. This is net new, net new consumption, but it's also net new applications. And some of the applications I mentioned before, but these new applications are also the fastest-growing applications in history.

Speaker 2
40m 39s

Okay? 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 three dynamics. Lastly, remember, we were just talking about the American CSPs. Each country will fund their own infrastructure. 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 the number of digital biology startups 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. Do not 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. Enterprise computing is going to fund their own industry. The next question comes from Ben Ritzes with Melius.

Speaker 2
42m 4s

Your line is open. Hey, thanks a lot. Jensen, I wanted to ask you about cash. Speaking of half a trillion, you may generate about half a 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? 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 OpenAIs, etc. Thanks a lot. Yeah, 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 forecasts 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.

Speaker 2
43m 35s

Because they've been planning with us for so many years, our reputation and our credibility is incredible. 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. 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. With respect to the investments, this is really, really important work that we do. All of the investments that we've done so far, well, all the period, is associated with expanding the reach of CUDA, expanding the ecosystem. If you look at the work, the investments that we did with OpenAI, it's, of course, that relationship we've had since 2016. I delivered the first AI supercomputer ever made to OpenAI. We have had a close and wonderful relationship with OpenAI since then. Everything that OpenAI does runs on NVIDIA today. 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.

Speaker 2
44m 59s

This is a company that's growing incredibly fast. Do not just look at what is said in the press. Look at all the ecosystem partners and all the developers that are connected to OpenAI. They are all driving consumption of it. The quality of the AI that is being produced is a huge step up since a year ago. The quality of response is extraordinary. We invest in OpenAI for a deep partnership in co-development to expand our ecosystem and to support their growth. Of course, rather than giving up a share of our company, we get a share of their company. We invested in them in one of the most consequential once-in-a-generation companies, once-in-a-generation company that we have a share of. 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. In enterprise, they're doing incredibly well. Claude Code is doing incredibly well. Claude is doing incredibly well all over the world's enterprise. Now we have the opportunity to have a deep partnership with them and bringing Claude onto the NVIDIA platform.

Speaker 2
46m 23s

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 Humane 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 run them all. 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. We have the ability, through 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. That's our investment thesis. The next question comes from Jim Schneider with Goldman Sachs.

Speaker 2
48m 1s

Your line is open. Good afternoon. Thanks for taking my question. 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? Can you maybe address the Rubin CPX product you expect to introduce next year and 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? Thank you. CPX is designed for long-context type of workload generation. Long-context, basically, before you start generating answers, you have to read a lot, basically long-context. 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. CPX is designed for long-context type of workloads. It's perf per dollars. Its perf per dollar is excellent. Its perf per watt is excellent. Which made me forget the first part of the question. Inferencing. Oh, inference. Yeah. There are three scaling laws that are scaling at the same time. The first scaling law called pretraining continues to be very effective.

Speaker 2
49m 36s

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. The amount of computation necessary as a result of those three things has gone completely exponential. I think that it's hard to know exactly what the percentage will be at any given point in time and who. Of course, our hope is that inference is a very large part of the market. If inference is large, then what it suggests is that people are using it in more applications, and they're using it more frequently. We should all hope for inference to be very large. 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. It's very clear now that GB300, GB200, and 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.

Speaker 2
51m 20s

GB200, NVLink 72, is 10 times, 10-15 times higher performance. That is a big step up. It is going to take a long time before somebody is able to take that on. Our leadership there is surely multi-year. Yeah. I think I am hoping that inference becomes a very big deal. Our leadership in inference is extraordinary. The next question comes from Timothy Arcury with UBS. Your line is open. Thanks a lot. Jensen, many of your customers are pursuing behind-the-meter power. What is the single biggest bottleneck that worries you that could constrain your growth? Is it power, or maybe it is financing, or maybe it is something else like memory or even foundry? Thanks a lot. These are all issues, and they are all constraints. 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. We have created a whole new industry. 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.

Speaker 2
52m 45s

The idea that in order for software to run, you need these factories to generate it, every single token instead of retrieving information that was pre-created. I think this whole transition requires extraordinary scale. All the way from the supply chain, of course, the supply chain, we have much better visibility and control over it because, obviously, we're incredibly good at managing our supply chain. We have great partners that we've worked with for 33 years. 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. None of these things are easy, but they're all tractable, and they're all solvable things. 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. We have a lot of routes to market. Very importantly, our architecture has to deliver the best value to the customers that we have. 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.

Speaker 2
54m 28s

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. I think all of that is just all the things that I've been telling you over the years are really coming true and are becoming evident. The next question comes from Stacey Raskin with Bernstein Research. Your line is open. Questions. Colette, I have some questions on margins. You said for next year, you're working to hold them in the mid-70s. 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? How much is cost optimizations versus pre-buys versus pricing? 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? Thanks, Stacey. 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 and our gross margins in the mid-70s.

Speaker 2
56m 3s

We've achieved that and getting ready to also execute that in Q4. 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. Our systems are by no means very easy to work with. There are a tremendous amount of components, many different parts of it, as we think about that. We are taking all of that into account. We do believe, as we look at working again on cost improvements, cycle time, and mix, that we will work to try and hold at our gross margins in the mid-70%. That is our overall plan for gross margin. Your second question is around OpEx. 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. That means they are quite busy in order to meet that goal. We're going to continue to see our investments on innovating more and more, both our software, both our systems, and our hardware to do so.

Speaker 2
57m 26s

I'll leave it turned to Jensen if he wants to add any couple more comments. Yeah. 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 has known for quite a long time our requirements. And they've known for quite a long time our demand. We've been working with them and negotiating with them for quite a long time. I think the recent surge, obviously, quite significant. Remember, our supply chain has been working with us for a very long time. 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. 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. I think all of that has worked out well for us. Your final question comes from the line of Aaron Rakers with Wells Fargo. Your line is open. Yeah. Thanks for taking the question. 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 buildouts has changed at all.

Speaker 2
58m 57s

Have you seen? I think you've been fairly adamant in the past that some of these programs never really see deployments. I'm curious if we're at a point where maybe that's even changed more in favor of just GPU architecture. Thank you. Thank you very much. I really appreciate the question. First of all, you're not competing against teams. Excuse me. Again, as a company, you're competing against teams. There just aren't that many teams in the world 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. Today, we've got to build entire racks, entire three different types of switches: a scale-up, a scale-out, and a scale-across switch. It takes a lot more than one chip to build a compute node anymore. Everything about that computing system, because AI needs to have memory, AI didn't used 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 a mixture of experts to dense models to diffusion models to autoregressive, not to mention biological models that obey the laws of physics.

Speaker 2
1h 0m

The list of different types of models has exploded in the last several years. The challenge is the complexity of the problem is much higher. The diversity of AI models is incredibly large. This is where, if I will say, the five things that make 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 phase. That 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. Every single phase of that, every single layer of that transition, we are excellent at. You can invest in one architecture, use it across the board. You can use one architecture and not worry about the changes in the workload across those three 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.

Speaker 2
1h 1m

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 autoregressive or diffusion-based. We run everything. We run it for every major platform, as I just mentioned. We run every model. 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 could even make you a little tiny cloud called DGX Spark. 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. 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 Humane, if you're a new company like CoreWeave or NSCALE or Nevius, or OCI for that matter, the reason why NVIDIA is the best platform for you is because our offtake is so diverse.

Speaker 2
1h 3m

We can help you with offtake. It's not about just putting a random ASIC into a data center. Where's the offtake coming from? Where's the diversity coming from? Where's 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 five 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. Thank you. I will now turn the call to Toshiya Hari for closing remarks. In closing, please note we will be at the UBS Global Technology and AI Conference on December 2nd. And our earnings call to discuss the results of our fourth quarter of fiscal 2026 is scheduled for February 25th. Thank you for joining us today. Operator, please go ahead and close the call. Thank you. This concludes today's conference call. You may now disconnect.