25-Feb-2026

NVIDIA Corp. (NVDA)

Q4 2026 Earnings Call

CORPORATE PARTICIPANTS

Toshiya Hari
Vice President, Investor Relations & Strategic Finance, NVIDIA Corp.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

OTHER PARTICIPANTS

Vivek Arya
Analyst, BofA Securities, Inc.

Ben Reitzes
Analyst, Melius Research LLC

Joe Moore
Analyst, Morgan Stanley & Co. LLC

Antoine Chkaiban
Analyst, New Street Research LLP

Harlan Sur
Analyst, JPMorgan Securities LLC

Mark Lipacis
Analyst, Evercore Group LLC

C.J. Muse
Analyst, Cantor Fitzgerald & Co.

Aaron Rakers
Analyst, Wells Fargo Securities LLC

Stacy A. Rasgon
Analyst, Bernstein Research

Timothy Arcuri
Analyst, UBS Securities LLC

Atif Malik
Analyst, Citigroup Global Markets, Inc.

James Edward Schneider
Analyst, Goldman Sachs & Co. LLC

MANAGEMENT DISCUSSION SECTION

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 Fourth Quarter Earnings Call. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer session. [Operator Instructions] Thank you.

Toshiya Hari, you may begin your conference.

Toshiya Hari

Vice President, Investor Relations & Strategic Finance, NVIDIA Corp.

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the fourth 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.

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 first quarter of fiscal 2027. 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, February 25th, 2026, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website.

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

Colette M. Kress

Chief Financial Officer & Executive Vice President, NVIDIA Corp.

Thanks, Toshiya. We delivered another outstanding quarter with record revenue, operating income and free cash flow. Total revenue of $68 billion was up 73% year-over-year, accelerating from Q3. Growth on a sequential basis was also a record, as we added $11 billion in data center revenue across a diverse and expanding set of customers, including cloud providers, hyperscalers, AI model makers, enterprises and sovereign nations. Demand for our Blackwell architecture extreme co-designed at data center scale continues to strengthen as inference deployments grow, in addition to training. The transition to accelerated computing and the infusion of AI across existing hyperscale workloads continue to fuel our growth. Agentic and physical AI applications built on increasingly smarter and multimodal models are beginning to drive our financial performance.

On a full year basis, data center generated revenue of $194 billion, up 68% year-over-year. We have now scaled our data center business by nearly 13x since the emergence of ChatGPT in fiscal 2023. We look ahead, we expect sequential revenue growth throughout calendar 2026, exceeding what was included in the $500 billion

Blackwell and Rubin revenue opportunity we shared last year. We believe we have inventory and supply commitments in place to address future demand, including shipments, extending into calendar 2027.

Every data center is power constrained. Customers make critical architectural decisions based on performance per watt, given these constraints and the need to maximize AI factory revenue. SemiAnalysis declared NVIDIA Inference King as recent results from InferenceX reinforced our inference leadership, with GB300 NVL72 achieving up to 50x performance per watt and 35x lower cost per token, compared with Hopper. And continuous optimization of CUDA software helped deliver up to five times better performance on GB200 NVL72 just within four months. NVIDIA produces the lowest cost per token and data centers running on NVIDIA generate the highest revenues.

Our pace of innovation, particularly at our scale, is unmatched, fueled by an annual R&D budget approaching $20 billion and our ability to extreme co-designed across compute and networking across chips, systems, algorithms and softwares. We intend to deliver X factor leaps in performance per watt every generation and extend our leadership position over the long term.

Q4 data center revenue of $62 billion increased 75% year-over-year and 22% sequentially, driven primarily by sustained strength in Blackwell and the Blackwell Ultra ramp. With NVIDIA infrastructure in high demand, even Hopper and much of the six year old Ampere-based products are sold out in the cloud.

Nearly a year has passed since the release of our Grace Blackwell NVL72 systems. Today, nearly nine gigawatts of infrastructure on Blackwell are deployed and consumed by the major cloud service providers, hyperscalers, AI model makers and enterprises.

Networking, a cornerstone of our data center scale infrastructure offering, was a standout this quarter, generating $11 billion in revenue, up more than 3.5x year-over-year. Demand for our scale up and scale out technologies reached record levels, both growing double digit sequentially, driven by strong adoption of NVLink, Spectrum-X Ethernet and InfiniBand. On a year-over-year basis, growth was driven primarily by NVLink 72 scale up switches as Grace Blackwell Systems accounted for roughly two-thirds of data center revenue in the quarter.

NVLink scale up fabric has revolutionized computing and demonstrates the power of extreme co-design across all of the chips of the supercomputer and the full stack. In Q4, we announced that we will enable AWS with NVLink to integrate with their custom silicon.

Momentum is strong with our Spectrum-X Ethernet scale up and scale across networking, as customers work to unify distributed data centers into integrated giga scale AI factories. For the full year, our networking business exceeded $31 billion in revenue, up more than 10-x compared to fiscal 2021, the year we acquired Mellanox.

Our demand profile is broad, diverse and expanding beyond just chatbots. First, there's a fundamental platform shift from classical machine learning to generative AI. Strong evidence of ROI as hyperscalers upgrade massive traditional workloads to generative AI, including search, ad generation and content recommender systems, is encouraging our largest customers to accelerate their capital spending.

For example, at Meta, advancements in their GEM model drove a 3.5 increase in ad clicks on Facebook and more than 1% gain in conversations on Instagram, translating into meaningful revenue growth. With the same NVIDIA infrastructure, Meta Superintelligence Labs can train and deploy their frontier agentic AI systems. Frontier agentic systems have reached an inflection point. Claude Code, Claude Cowork and OpenAI Codex have

achieved useful intelligence. Adoption is skyrocketing, and tokens are profitable, driving extreme urgency to scale up compute. Compute directly translate to intelligence and revenue growth.

Analysts expectations for 2026 CapEx across the top five cloud providers and hyperscalers who collectively account for little over 50% of our data center revenue, are up nearly $120 billion since the start of the year and approaching $700 billion. We continue to expect the transition of classic data center workloads to GPU accelerated computing, and the use of AI to enhance today's hyperscale workloads and contribute toward roughly half of our long-term opportunity.

Every country will build and operate some parts of its AI infrastructure, just like with electricity and Internet today. In fiscal year 2026, our sovereign AI business, more than tripled year-over-year and over $30 billion, driven primarily by customers based in Canada, France, the Netherlands, Singapore and the UK. Over the long run, we expect our sovereign opportunity to grow at least in line with the AI infrastructure market, as countries spend on AI proportional to their GDP.

While small amounts of H200 products for China-based customers were approved by the US government, we have yet to generate any revenue and we do not know whether any imports will be allowed into China. Our competitors in China, bolstered by recent IPOs, are making progress and have the potential to disrupt the structure of the global AI industry over the long-term. To sustain its leadership position in AI compute, America must engage every developer and be the platform for choice for every commercial business, including those in China.

We will continue to engage with the US and China governments and advocate for America's ability to compete around the world. We unveiled the Rubin platform last month at CES, comprised of six new chips; the Vera CPU, Rubin GPU, NVLink, 6 Switch, ConnectX-9, SuperNIC, BlueField-4 DPU, and Spectrum-6 Ethernet Switch. The platform will train MoE models with one-fourth the number of GPUs and reduce inference token costs by up to 10x compared to Blackwell.

We shipped our first Vera Rubin samples to customers earlier this week and we remain on track to commence production shipments in the second half of the year. Based on its modular cable-free tray design, Rubin will deliver improved resiliency and serviceability relative to Blackwell. We expect every cloud model builder to deploy Vera Rubin.

Moving to gaming. Gaming revenue of $3.7 billion increased 47% year-on-year, driven by strong Blackwell demand and improved supply. GeForce RTX is the leading platform for PC gamers, creators, and developers.

In Q4, we added several new technologies and advancements, including DLSS 4.5, which uses AI to bring game visuals to a new level. G-SYNC Pulsar, bringing incredible clear graphics even in motion, and 35% faster LLM inference across leading AI PC frameworks.

Looking ahead, while end demand for our products remains strong and channel inventory levels are healthy, we expect supply constraints to be the headwind to gaming in Q1 and beyond.

For Professional Visualization, it crossed the $1 billion mark for the first time, with revenue of $1.3 billion, up 159% year-over-year and 74% sequentially. During the quarter, we launched the RTX PRO 5000 Blackwell Workstation with 72 gigabytes of fast memory for AI developers running LLMs and agentic workflows.

Automotive revenue of $604 million was up 6% year-over-year and was driven by robust demand for self-driving solutions. At CES, we introduced Alpamayo, the world's first open portfolio of reasoning, vision, language, action models, simulation blueprints and data sets enabling vehicles that can think.

The first passenger car featuring Alpamayo built on NVIDIA Drive will be on the road soon in the new Mercedes-Benz CLA.

Physical AI is here, having already contributed north of $6 billion in NVIDIA revenue in fiscal year 2026. Robotaxi rides are growing exponentially with commercial fleets from Waymo, Tesla, Uber, WeRide and Zoox, and many others are expected to scale from thousands of vehicles in 2025 to millions over the next decade, creating a market poised to generate hundreds of billions of dollars of revenue.

This expansion will demand orders of magnitude more compute with every major OEM and service provider developing on NVIDIA's platform. We continue to advance robotics development with the new NVIDIA Cosmos and Isaac GROOT open models, frameworks and NVIDIA's powered robots and autonomous machines for leading companies, including Boston Dynamics, Caterpillar, Franka Robotics, LG Electronics and NEURA Robotics. To accelerate industrial physical AI adoption, we also announced new expanding partnerships with Dassault Systèmes, Siemens and Synopsys to bring NVIDIA AI infrastructure, Omniverse digital twins, world models and CUDA-X libraries to millions of researchers, designers and engineers building the world's industries.

Let's move to the rest of the P&L. GAAP gross margin was 75%, and non-GAAP gross margin was 75.2%, increasing sequentially as Blackwell continued to ramp. GAAP operating expenses were up 16% sequentially and up 21% on a non-GAAP basis, related to new product introductions and compute and infrastructure costs. Non-GAAP effective tax rate for the fourth quarter was 15.4%, below our outlook for the quarter, primarily due to the impact of a one-time tax benefit.

Inventory grew 8% quarter-over-quarter, while purchase commitments also increased significantly, as we have strategically secured inventory and capacity to meet demand beyond the next several quarters. This is further out in time than usual and reflects the longer demand visibility we have. While we expect tightness in the supply for our advanced architectures to persist, we remain confident in our ability to capitalize on the growth opportunity ahead with our scale, expansive supply chain, and the longstanding partnerships continuing to serve us well.

We generated free cash flow of $35 billion in Q4 and $97 billion in fiscal year 2026. For the year, we returned $41 billion, or 43% of free cash flow to our shareholders in the form of share repurchases and dividends. We continue to invest in our technology and our ecosystem to cultivate market development, drive long-term growth, and ultimately yield total shareholder returns superior to the market or our peer group. Importantly, we will continue to run a strategic and disciplined process as it relates to our investments, and we remain committed to returning capital to our shareholders.

Let me turn to the outlook for the first quarter. Starting this quarter, we will be including stock-based compensation expense in our non-GAAP results. Stock-based compensation is a foundational component of our compensation program to attract and retain world class talent. Let me first start with revenue. Total revenue is expected to be $78 billion, plus or minus 2%. We expect most of our growth to be driven by Data Center. Consistent with last quarter, we are not assuming any Data Center compute revenue from China in our outlook. GAAP and non-GAAP gross margins are expected to be 74.9% and 75%, respectively, plus or minus 50 basis points. For the full year, we continue to see gross margins in the mid-70s. We will keep you updated on our progress as we prepare for the Vera Rubin transition.

GAAP and non-GAAP operating expenses are expected to be approximately $7.7 billion and $7.5 billion, respectively, including stock-based compensation expense of $1.9 billion.

For the full year, we expect non-GAAP operating expenses to grow in the low 40s on a year-over-year basis as we continue to invest in our expanding opportunity set. For the full year fiscal year 2027, we expect GAAP and non-GAAP tax rates to be in between 7% and 19%, excluding any discrete items and material changes to our tax environment.

With that, let me turn the call over to Jensen. I think he has a few words for us.

Jen Hsun Huang

Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

This quarter we significantly deepened and expanded our partnerships with leading frontier model makers. We recently celebrated OpenAI's launch of GPT-5.3-Codex, trained with and inferencing on Grace Blackwell NVLink 72 systems.

GPT-5.3-Codex can take on long running tasks that involve research, tool use, and complex execution. 5.3-Codex is deployed broadly inside NVIDIA, our engineers love it.

We continue to work with OpenAI toward a partnership agreement and believe we are close. We are thrilled with our ongoing partnership with OpenAI, a once in a generation company we've had the pleasure of partnering with since their first days.

Meta Superintelligence Labs is scaling up at lightning speed. Last week, we announced that Meta is deploying millions of Blackwells and Rubin GPUs, NVIDIA CPUs, and Spectrum-X Ethernet for training and inference. This quarter, we announced a partnership with Anthropic and a $10 billion investment in their company. Anthropic will train an inference on Grace Blackwell and Vera Rubin systems.

Anthropic's Claude Cowork agent platform is revolutionary and has opened a floodgates for enterprise AI adoption. Between Claude Cowork and OpenClaw, compute demand is skyrocketing and ChatGPT moment of agentic AI has arrived.

With partnerships spanning Anthropic, Meta, OpenAI, and xAI, NVIDIA deployed across every cloud, and with our ability to build full-stack AI infrastructure from the ground up or support them in the cloud, we're uniquely positioned to partner with frontier model builders at every stage; training, inference, and AI factory scale out.

Finally, we recently entered into a non-exclusive licensing agreement with Groq for its low-latency inference technology, and welcomed a team of brilliant engineers to NVIDIA. As we did with Mellanox, we will extend NVIDIA's architecture with Groq's innovations to enable new levels of AI infrastructure performance and value. We look forward to sharing more at GTC next month.

Okay, back to you.

Toshiya Hari

Vice President, Investor Relations & Strategic Finance, NVIDIA Corp.

We will now transition to Q&A. Operator, please poll for questions.

QUESTION AND ANSWER SECTION

Operator: [Operator Instructions] Your first question comes from Vivek Arya with Bank of America Securities. Your line is open.

Vivek Arya
Analyst, BofA Securities, Inc.

Q

Thanks for taking my question. I think you mentioned that you now have growth visibility into calendar 2027 also, and I think your purchase commitments kind of reflect that confidence. But Jensen, I'm curious, when you look at your top cloud customers, cloud CapEx close to $700 billion this year, many investors are concerned that it would be harder for this level to grow into next year. And for several of them, their cash flow generation capability is also getting compressed. So I know you're very confident about your roadmap, right, and your purchase commitments and whatnot, but how confident are you about your customers' ability to continue to grow their CapEx? And if their CapEx doesn't grow, can NVIDIA still find a way to grow in that envelope? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

I am confident in their cash flow growing. And the reason for that is very simple. We have now seen the inflection of agentic AI and the usefulness of agents across the world in enterprises everywhere. You're seeing incredible compute demand because of it. In this new world of AI, compute is revenues. Without compute, there's no way to generate tokens. Without tokens, there's no way to grow revenues. So in this new world of AI, compute equals revenues. And I am certain that at this point, with the productive use of Codex and Claude Code and the excitement around Claude Cowork and, just the incredible enthusiasm about OpenClaw and the enterprise versions of them. All of the enterprise ISVs who are now working on agentic systems on top of their tools platforms, I am certain at this point that we are at the inflection point. We've reached the inflection point, and we're generating profitable tokens that are productive for customers and profitable for the cloud service providers. And so the simple logic of it, the simple way to think about it is computing has changed. What used to be software running on computers, modest amount of computers, call it $300 billion or $400 billion worth of CapEx each year has now gone into AI and AI in order to have, in order to generate tokens, you need compute capacity. And that translates directly to growth and that translates directly to revenues.

Operator: Your next question comes from Joe Moore with Morgan Stanley. Your line is open.

Joe Moore
Analyst, Morgan Stanley & Co. LLC

Q

Great. Thank you. And congratulations on the numbers. You talked about some of the strategic investments that you've made into Anthropic and potentially OpenAI [ph] Claude as well (26:23) but also partners Intel, Nokia. Synopsys. You're clearly at the center of everything. Can you talk about the role of those investments and kind of how do you view the balance sheet as a tool to kind of grow the NVIDIA's position, the ecosystem and, and participate in that growth.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

As you know, fundamentally, at the core of everything NVIDIA is our ecosystem. That's what everybody loves about our business. The richness of our ecosystem, just about every startup in the world is working on NVIDIA's

platform, we are in every cloud, we're in every on-prem data center, we're all over the world's edge and robotic systems, thousands of AI natives are built on top of NVIDIA. We want to take the great opportunity that we have as we're in the beginning of this new computing era, this new computing platform shift to put everybody on NVIDIA. Everything is already built on CUDA and so we're starting from a really terrific starting point. But as we build out the entire AI ecosystem, whether it's in AI for language or physical AI or AI physics or biology or robotics or manufacturing, we want all of these ecosystems to be built on top of NVIDIA. And this is such a wonderful opportunity for us to invest into the ecosystem across the entire stack.

Our ecosystem is also richer today than it used to be. We used to be largely a computing platform on GPUs, but now we're a computing AI infrastructure company, and we have computing platforms on, well, every aspect of that.

And everything from computing to AI models to networking to our DPU, all of that has computing stacks on top of it. And as I mentioned before, whether it's an enterprise or in manufacturing, industrial or science or robotics, each one of these ecosystems have different stacks, and we want to make sure that we continue to invest into our ecosystem. So, our investments are focused very squarely, strategically on expanding and deepening our ecosystem reach.

Operator: Your next question comes from Harlan Sur with JPMorgan. Your line is open.

Harlan Sur
Analyst, JPMorgan Securities LLC

Q

Good afternoon. Thanks for taking my question. Networking continues to rise as a percentage of your overall data center profile, right. Through fiscal 2026, your networking revenues accelerated on a year-over-year basis every single quarter, right. With 3.6x growth, as you guys mentioned year-over-year growth in Q4, obviously, on the strength of your scale up and scale out networking product portfolio, I would seem to remember that first half of last year, your annualized run rate on your Spectrum-X Ethernet Switching platform was around $10 billion annualized. It looks like that may have stepped up to around $11 billion, $12 billion in the second half of last year.

Jensen, looking at your order book, especially with Spectrum-XGS, upcoming 102T Spectrum-6 Switching platforms launching soon. Where is the Spectrum run rate trending now and as you foresee exiting sort of this calendar year?

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah. As you know, we see ourselves as an AI infrastructure company and the AI computing infrastructure includes CPUs, GPUs, and we invented NVLink to scale up the one computing node into a giant computing rack.

We invented the idea of a rack scale computer. We don't ship nodes of computers, we ship racks of computers. And those – that NVLink Switch scale up system is then scaled out using Spectrum-X and InfiniBand, we support both. And then further, we also scale across data centers using Spectrum-X scale across.

And so the way we think about networking is really an extension, it's – we offer everything openly, so that people could decide to mix and match in different scale, and however they would like to integrate it into their bespoke data center. But in the final analysis, it's all one big part of our platform.

And the invention of NVLink really turbocharged our networking business. Every rack comes with nine nodes of switches, and each one of them has two chips in it and in the future they'll have more. And so the amount of switching that we do per rack is really quite incredible.

We're also now the largest networking company in the world. And if you look at Ethernet, we came into the Ethernet market about a couple of years ago into Ethernet switching. And I think that we're probably the largest Ethernet networking company in the world today, and surely will be soon. And so Spectrum-X Ethernet has been a home run for us. But we're open to however people want to do networking. Some people just really love the low latency and the scale up capability of InfiniBand. And we will continue to support that, of course. And some people love to integrate their networking across their data center based on Ethernet. And we created an Ethernet capability that extends Ethernet with artificial intelligence, way of processing in the data center and we're incredibly good at that and our Spectrum-X performance really shows it.

The difference of when you built a $10 billion or $20 billion AI factory, the difference of 10%, and it could be easily 20% on the effectiveness and the utilization of your network for your data center that translates to real money. And so NVIDIA's networking business is really, really growing fast. And it's, I think it's just because we built the AI infrastructure so effectively and the AI infrastructure business is growing incredibly fast.

Operator: Your next question comes from C.J. Muse with Cantor Fitzgerald. Your line is open.

C.J. Muse
Analyst, Cantor Fitzgerald & Co.

Q

Yeah. Good afternoon. Thank you for taking the question. I guess with CPX for large context Windows and Groq likely adding a decode specific solution. Curious how we should think about your future roadmap. Should we be thinking about customized silicon either by workload or customer, as an increasing focus by NVIDIA, particularly helped by your move to a dielet architecture? Thanks so much.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

We don't use – we want to – everybody should want to extend, push out dielet as long as they can. And the reason for that is because every time you cross a dielet you have a dielet, you have to cross an interface. Every time you cross an interface, you add latency, you add power unnecessarily. We're not allergic to dielet. We use dielets already. But we try to use dielets only when we absolutely have no choice but to do so. And so we – if you look at the Grace Blackwell architecture and the Rubin architecture, we used two giant reticle limited dies [ph] and we abut them (33:53) and that reduces the amount of architecture crossing.

The [ph] dielet tax (33:59) shows up in the architecture effectiveness of the competitors. If you look at NVIDIA, people call it our software advantage. But where software starts and architecture starts and ends is kind of hard to tell. It's, our software is effective because our architecture is so good. And so the CUDA architecture is unquestionably more effective, more efficient, delivers more performance per flop, per watt than any computing architecture out there and it's because of the way we architect.

With respect to, how we think about Groq and the low latency decoder, I've got some great ideas that I'd like to share with you at GTC, but the simple idea is that our infrastructure is incredibly versatile because of CUDA. And we're going to continue to do that. All of our GPUs are architecturally compatible, which means that when I'm working on optimizing models today for Blackwell, all of that work and all of that dedication to optimizing software stacks and new models, also benefit Hopper and also benefit Ampere. It's the reason why A100 continues to feel

fresh and continues to stay performant years after we've deployed it into the world. Architecture compatibility allows us to do that. It allows us to invest enormously in software engineering and optimization, knowing that our entire installed base in the cloud, on-prem, everywhere from generations of architectures of GPUs will all benefit. And so we'll continue to do that.

And allows us to extend the useful life, allows us to have innovation, flexibility, and velocity, which translates the performance and very importantly, performance per dollar and performance per watt for our customers. And so what we'll do with Groq is, you'll come to see GTC, but what we'll do is we'll extend our architecture with Groq as an accelerator in very much the ways that we extended NVIDIA's architecture with Mellanox.

Operator: The next question comes from Stacy Rasgon with Bernstein Research. Your line is open.

Stacy A. Rasgon
Analyst, Bernstein Research

Q

Hi, guys. Thanks for taking my questions. Colette, I wanted to dig a little bit into the call for sequential growth through the year. So, I mean, you grew this quarter more than $10 billion sequentially in data center, and the guide seems to imply the bulk of the increase $10 billion sequential in data centers. So, how do you see that as we go through the year, especially as Rubin ramps into the back half? Blackwell has been a pretty massive acceleration with a sequential growth. Should we expect something similar as we get to Rubin?

And then I was also just hoping you could comment on, your expectations for gaming? I understand the memory issues and everything else. Do you think gaming can still grow year-over-year in fiscal 2027, or will that be under more pressure given memory? So, those two questions, please. Thank you.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

A

Thanks, Stacy. Let me start with the revenue going forward. Again, we're trying to look at revenue quarter-by-quarter. As you think about the full year, we are absolutely going to be still selling and providing Blackwell probably at the same time that we're also seeing Vera Rubin come to market.

This is a very great architecture that helps them just today quickly standing up and have already planned on many different orders across the different customers to provide that.

It's too early yet to determine how much in terms of that Vera Rubin, that beginning ramp will start in the second half and we'll get through it. But no confusion in terms of the strong demand and the interest. We do expect pretty much every single customer to be purchasing Vera Rubin. The question is, are, how soon are we in market and how soon are they able to stand that up in terms of in their data centers? That was your first part.

The second part was focusing on our gaming. As much as we would love to have additional more supply, we do believe for a couple quarters, it is going to be very tight. If things improve by the end of the year, there is an opportunity to think about what that is from a year-over-year growth. But it's still too early for us to know at this time. And we'll get back to you as soon as we can.

Operator: Your next question comes from Atif Malik with Citi. Your line is open.

Atif Malik
Analyst, Citigroup Global Markets, Inc.

Q

Thank you for taking my question. Jensen, I'm curious if you can touch on the importance of CUDA as now more of the investment dollars in AI are coming from inference workloads?

Jen Hsun Huang

Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Without CUDA, we wouldn't know what to do with inference. The entire stack from TensorRT LLM that we introduced a few years ago, which is still the most performant inference stack in the world. Optimizing it for NVLink, requires us to discover and invent new parallelization algorithms that sits on top of CUDA, to distribute the workload and the inferencing to take advantage of the aggregate bandwidth across NVLink 72. NVLink 72 has enabled us to deliver generationally 50 times more performance per watt. It's just an incredible leap and it's sensible. NVLink 72 is a great invention. It was hard to do, the creation of the switching technology, disaggregating the switches, building the system racks, all of that. We did it all in plain sight and everybody knew how hard it was for us to do. And, but the results are incredible. So performance per watt is 50 times, performance per dollar, 35 times and so the leap in inference is incredible.

It's very important – it's really important to realize that inference equals revenues now for our customers because agents are generating so many tokens and the results are so effective. When the agents are coding, it's off generating thousands, tens of thousands, hundreds of thousands because they're running for minutes to hours. And so these systems, these agentic systems are spawning off different agents working as a team.

The number of tokens that are being generated has really, really gone exponential. And so, we need to inference at a much higher speed. And when you're inferencing at a much higher speed, and each one of those tokens are dollarized, it directly translates into revenues. And so inference equals, inference performance equals revenues for our customers.

For the Data Centers, inference tokens per watt translates directly to the revenues of the CSPs. And the reason for that is because everybody is power limited. And so, I mean, no matter how many data centers you have, each data center, 100 megawatts or 1 gigawatt has power limits. So the architecture that has the best performance per watt translates because each token, the performance tokens per watt, each token is dollarized, tokens per watt translates to dollars per watt, which translates in a gigawatt directly to revenues.

And so you could see that every CSP understands this now, every hyperscaler understands this, that CapEx translates to compute, compute with the right architecture translates to maximizing revenues and compute equals revenues. Without investing capacity today, without investing in compute, there cannot be revenue growth and that I think everybody understands. Compute equals revenues. Choosing the right architecture is incredibly important, is more than strategic now, it directly affects their earnings. And choosing the right architecture, the one with the best performance per watt is literally everything.

Operator: Your next question comes from Ben Reitzes with Melius Research. Your line is open.

Ben Reitzes

Analyst, Melius Research LLC

Q

Yeah. Hey, thanks. First, let me say kudos on including the stock-comp in non-GAAP. I think that's a great move, but that isn't my question. My question is around gross margins and the sustainability of the mid-70s long-term. Should we read into the visibility on supply being available into calendar 2027 that it's sustainable until then?

And then Jensen, what about after that? Are there innovations in memory consumption you can unveil that makes us feel better about the ability to keep margins at that level for a long time? Thanks.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

The single most important lever of our gross margins is actually delivering generational leaps to our customers. That is the single most important thing. If we could deliver generationally performance per watt that exceeds dramatically what Moore's Law can do. If we can deliver performance per dollar dramatically more than the cost of our systems, than the price of our systems, then we can continue to sustain our gross margins. That's the simple, most important concept.

Every – the reason why we're moving so fast is because number one, the demand for tokens in the world as a result of the inflection points that we've gone through has now – has gone completely exponential. I think we're all seeing that, to the point where even our six-year old GPUs in the cloud are completely consumed, and the pricing is going up.

And so we know that the amount of computation necessary – the amount of compute necessary for the modern way of doing software is growing exponentially. And so our strategy is to deliver an entire AI infrastructure every single year. This year we introduced six new chips, Rubin next-generation will do many new chips as well. And every single generation, we are committed to deliver many X factors of performance per watt and performance per dollar. And that pace and our ability to do extreme co-design, allows us to deliver that value and that benefit to the customers. And that is the single most vital thing as it relates to our value delivered.

Operator: Your next question comes from Antoine Chkaiban with New Street Research. Your line is open.

Antoine Chkaiban
Analyst, New Street Research LLP

Q

Hi. Thanks a lot for taking my question. I'd like to ask about space data centers, which some of your customers are considering. How feasible do you think that is and what kind of horizon? And what do the economics look like today? And how do you think that could evolve over time? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Well, the economics are poor today, but it's going to improve over time. As you know, the way that space works is radically different than how it works down here. There's an abundance of energy. But solar panels are large, but there's plenty of space in space.

The heat dissipation -- it's cold in space. However, there's no airflow. And so the only way to dissipate heat is through conduction. And the radiators that you need to create are fairly large. Liquid cooling is obviously out of the question because it's kind of -- it's heavy and freezes. And so the methods that we use here on Earth are a little different than the way we would do it in space. But there are many different computing problems that really wants to be done in space. And so, NVIDIA is already the world's first GPU in space, Hopper's in space.

And one of the best use cases of GPUs in space is imaging, to be able to image at extremely high resolutions using, of course, optics and artificial intelligence and to be able to do that computation of reprojection of different angles and be able to upres and do noise reduction and just be able to see, be able to image at very large, very high resolutions, extremely large scales, and very, very fast. It's hard to do that by sending petabytes and

petabytes of imaging data back here on Earth and doing that work. It's easier just to do it out in space. And then, ignore all of the data collected and processed until you see something interesting. And so artificial intelligence in space will have very good, very interesting applications.

Operator: Your next question comes from Mark Lipacis with Evercore ISI. Your line is open.

Mark Lipacis
Analyst, Evercore Group LLC

Q

Hi. Thanks for taking my question. I want to pick up with the comment you made on the script about revenue diversification. I believe, Colette, you said that hyperscalers were over 50% of revenues, but growth was led by the rest of your Data Center customers. And as a clarification, I just want to make sure I understood that, does that imply your non-hyperscale customers grew faster. And if so, what are the – can you help us understand what are the non-hyperscalers doing different. Are they doing different things than the hyperscalers or the same things on a different scale and does this – do you expect this trend to continue? Would you expect your customer base to evolve to a point where non-hyperscalers are becoming a bigger part of your, the larger part of your business? Thank you.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

A

Yes. Let's see if we can help on this question. So when you think about our top five, as we articulated as being our CSPs, our hyperscalers, and they have right now, [ph] as I said (49:15) about 50% of our total revenue. There's a big organization, therefore, of diversity of all different other types of companies that we are working with. That it goes through our AI model makers, that goes through our enterprises, that goes to supercomputing, it goes to our sovereigns. There's a lot of other different facts on there. But you are correct, it's a very fast growing area as well.

We have a strong position in terms of all of our different cloud providers on our platform, and now we also have a extreme diversity of different customers that we are seeing all the way across the world. And this will really benefit, seeing that diversity and being able to serve all of those parts. Let me see if Jensen wants to add a bit more

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah, this is one of the advantages that we have with our ecosystem. I'll build on top of CUDA. We have -- we're the only accelerated computing platform that is in every cloud, that's available through every single computer maker, available at the edge and, we're now cultivating telecommunications. Obviously, the future radios will all be AI driven radios and the future wireless network would also be a computing platform. That is a foregone conclusion. But somebody has to go and invent the technologies to make that possible. And we created a platform called Aerial to go do that.

We are in just about every single robot, every single self-driving car. Our ability – CUDA's ability to have the benefit of the performance of specialized processors, one the one hand, with the Tensor cores inside our GPUs, on the other hand, the flexibility of CUDA allows us to solve language problems, computer vision problems, robotics problems, to biology problems, physics problems, and just about all kinds of AI and all kinds of computation algorithms. And so the diversity of our customer base is one of the greatest strengths that we have.

The second thing, of course, is without our own ecosystem, even if our processor was programmable, if we didn't cultivate our ecosystem and talking about some of the things that we're doing today, investing in our future ecosystem and continuing to enhance our ecosystem, without our ecosystem, it's hard for us to grow beyond what design wins we capture for somebody else's ecosystem. And so we could grow and expand our ecosystem very naturally because of our – the platform that we created.

And then lastly, one of the things that's really important is the partnerships that we have with OpenAI and Anthropic, with xAI, with Meta now makes – and of course, just about every single open source in the world. There's 1.5 million AI models on Hugging Face, all of it runs on NVIDIA CUDA. And so an open source in totality probably represents the largest – the second largest model in the world, OpenAI is the largest, second largest, probably all the collection of all the open sources.

And so NVIDIA's ability to run all of that makes our platform super fungible, super easy to use, and really safe to invest into. And so that creates the diversity of customers and the diversity of the platforms and available in every single country and – because we support the whole world's ecosystem.

Operator: Your next question comes from Aaron Rakers with Wells Fargo. Your line is open.

Aaron Rakers
Analyst, Wells Fargo Securities LLC

Q

Yeah. Thanks for taking the question. I guess, sticking with the idea as a platform and extreme co-design, some of the news over this last quarter has obviously been NVIDIA's ability or push to bring Vera CPUs to market on a standalone solution basis.

So, I guess, Jensen, I'm curious of what's the importance that Vera plays in this architecture evolution as we move forward? Is this being driven more by the proliferation or the heterogeneity of inference workloads? I'm just curious of how you see that evolving for NVIDIA, particularly on a standalone CPU basis? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah, thanks. And I'll tell you some more about it at GTC. But at the highest level, we made fundamentally different architecture decisions about our CPUs compared to the rest of the world's CPUs.

It's the only data center CPU that supports LPDDR5. It is designed to be focused on very high data processing capabilities. And the reason for that is because most of the computing problems that we're interested in are data-driven, artificial intelligence being one.

And the single threaded performance and this ratio with bandwidth is just off the charts. And we made those architectural decisions because in the entire phase – the different phases of AI from data processing before you even do training, you have to do data processing. So, you have data processing, pre-training, and in post-training now the AI's are learning how to use tools and the usage of tools, many of those tools run in CPU-only environments or they run in CPU or GPU accelerated environments. And Vera was designed to be an excellent CPU for post-training and so some of the use cases in the entire pipeline of artificial intelligence includes using a lot of CPUs.

We love CPUs as well as GPUs. And when you accelerate the algorithms to the limit, as we have, Amdahl's law would suggest that you need really, really fast single threaded CPUs. And that's the reason why we built, Grace to be a extraordinarily great at single threaded performance and Vera is off the charts better than that.

Operator: Your next question comes from Tim Arcuri with UBS. Your line is open.

Timothy Arcuri
Analyst, UBS Securities LLC

Q

Thanks a lot. Colette, I was wondering if you can talk about the deployment of capital. I know that you really jacked up the purchase commits. But it sounds like maybe you're over the hump on this, and you're going to probably generate about $100 billion in cash this year.

So, and, pretty much no matter how good the results have been, the stock hasn't really gone up much. So I would think that you probably feel like this is a pretty good price to be buying back a bunch of it here. So I was wondering if you can talk about that like, question being, why not put a big stake in the ground and just, have a huge share repo here? Thanks.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

A

So, thanks for the question. We look at our capital return, very, very carefully. And we do believe that one of the most important things that we can do is really supporting the extreme ecosystem that's in front of us. That stems from everywhere, from our suppliers and the work that we need to do to assure that we can have the supply that's needed and help them from a capacity, all the way that we are in terms of the early developers of the AI solutions, that will be on our platform.

So we will continue to make this a very important part of our process and strategic investments. But of course, we are still repurchasing our stock. We are still with our dividend as well. And we will continue to find the right unique opportunities within the year for doing those different purchases.

Operator: Your final question comes from Jim Schneider with Goldman Sachs. Your line is open.

James Edward Schneider
Analyst, Goldman Sachs & Co. LLC

Q

Thank you for taking my question. Jensen, you've previously outlined the potential to get to $3 trillion to $4 trillion of data center CapEx by 2030, which implies a potential acceleration in growth rates, which you sort of guided to this, at least this next quarter.

The question is, what are some of the key application areas that you believe are most likely to drive that inflection? Is that physical AI, agentic or something else? And do you still feel good about that $3 trillion to $4 trillion envelope? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah. Let's back that up and just reason through it from a few different ways. So the first way is on first principles. The way that software is done in the future using AI is token driven. And I think everybody talks about tokenomics and, talks about data centers generating tokens and inference is about generating tokens. And, we generate

tokens. We're just talking about tokens. How NVIDIA's NVLink 72 enabled us to generate tokens at 50 times better performance per unit energy than the previous generation. And so token generation is at the center of almost everything that relates to software in the future and relates to computing.

If you look at the way we use computing in the past, however, the amount of computation demand for software in the past is a tiny fraction of what is necessary in the future. And AI is here. AI is not going to go back. AI is only going to – only get better from here. And so if you think about it and you said, okay, well, the world was investing about $300 billion to $400 billion a year in classical computing. And now AI is here and the amount of computation necessary is a thousand times higher than the way we used to do computing, the computing demand is just a lot higher.

And so if we continue to believe there's value in it, and we'll talk about that in a second, then the world will invest to produce that token. And so the amount of token generation capability that the world needs is a lot more than $700 billion.

And I'm fairly confident that we're going to continue to generate tokens. We're going to continue to invest in compute capacity from this point out. And fundamentally, because every single company depends on software, every software will depend on AI, and so every company will produce tokens. And that's the reason why I call them AI factories.

And whether you're a company in the cloud data centers, you have AI factories to generate tokens for your revenues. If you're a enterprise software company, you're going to generate tokens for the agentic systems that are on top of your tools. If you are a robotics factory and self-driving car's first indication of that, you have huge supercomputers, which are basically AI factories to generate tokens that goes into your cars, that becomes its AI. And then you also have to put computers inside the cars to continuously generate tokens. And so, we're fairly sure now that this is the future of computing.

Now, why is it so certain that this is the future of computing? And the reason for that is because the way we used to do software was pre-recorded. Everything was captured a priori. We pre-compile the software, we pre-write the content, we pre-record the videos. But now everything is generative in real-time. And when it's generated in real-time, it can take into context of the person, the situation, the query, and the intentions could all be taken into consideration to generate the outcome of this new software we call AI, agentic AI.

And so the amount of computation necessary is far, far greater than prerecorded, just as a computer has a lot more computation capability than a DVD recorder, a DVD player that was pre-recorded, artificial intelligence needs a lot more computing capability than the way we used to do software in the past.

Now, the question about computation, about sustainability at the first level is just at the computer science level, this is the way computing is going to be done. Now, from an industrial level, because all of our companies, in the final analysis are powered by software and the cloud companies are powered by software. And if the new software requires tokens to be generated and the tokens are monetized, then it stands to reason that their data center build-out directly drives their revenues. And so compute drives revenues. And I think they all understand that. I think people are increasingly starting to understand that as well.

And then lastly, the benefits that AI produces for the world ultimately has to generate revenues. And we're seeing right in front, right being developed as we see – as we stand here, agentic AI has turned an inflection point, and it literally happened in the last couple of two, three months. Of course, inside the industry, we've been seeing it for a while, probably six months or so, but the world is now awaken to the agentic AI inflection. The agents are super

smart. They're solving real problems. Coding is obviously supported by agentic systems now. And all of our coders here at NVIDIA are using agentic systems, either Claude Code or OpenAI Codex enormously to – and oftentimes both and Cursor, oftentimes all three, depends on the use case. But they have agents and co-design partners, engineering partners to help them solve problems.

And you could see their revenues skyrocketing. These companies, in the case of Anthropic, I think their revenues 10xed in a year. And they are severely capacity constrained because demand is just incredible. And the token demand is incredible. The token generation rate is growing exponentially. And the same thing with, of course, OpenAI, their demand is incredible. And so the more compute that they can stand online, bring online, the faster their revenues will grow. And that goes back to the comment that I was saying that inference is revenues, that compute equals revenues now in this new world.

And in a lot of ways, that's the reason why we say it's a new industrial revolution. There are new factories, new infrastructure being built and this new way of doing computing is not going to go back. And so to the extent that we believe that producing tokens is going to be the future of computing, which I believe, and I think largely the industry believes, then we're going to be building out this capacity from this point forward and continue to expand from here.

Now the thing that is – the wave that we're seeing now is the agentic AI inflection and the next inflection beyond that is physical AI, where we take AI and these agentic systems into the physical applications such as manufacturing, such as robotics. And so that's a giant opportunity ahead. Okay.

Operator: This concludes the question-and-answer session. I'll turn the call to Toshiya Hari.

Toshiya Hari

Vice President, Investor Relations & Strategic Finance, NVIDIA Corp.

In closing, please note, Jensen will be participating in a fireside chat at the Morgan Stanley TMT Conference in San Francisco on March 4. He'll also be giving a keynote at GTC in San Jose on March 16. Our earnings call to discuss the results of our first quarter of fiscal 2027 is scheduled for May 20. 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.