NVDA Earnings Call – FY2025 Q4

Generated by discountingcashflows.com

Date: February 26, 2025

Christa

Good afternoon. My name is Christa, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA Corporation'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. If you would like to ask a question during this time, simply press star followed by the number one on your telephone keypad. And if you would like to withdraw your question, Thank you. Stewart Stecker. You may begin your conference. Thank you.

Stewart Stecker

Good afternoon, everyone, and welcome to NVIDIA Corporation's conference call for the fourth quarter of fiscal 2025. With me today from NVIDIA Corporation 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 Corporation's Investor website. Webcast will be available for replay until the conference call discuss our financial results, the first quarter of fiscal 2026. The content of today's call is NVIDIA Corporation's property. It can't be reproduced or transcribed without prior written consent. During this call, we may make forward-looking statements based on current expectations, these are subject to a number of significant risks and uncertainties and our actual results may differ materially. 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 26, 2025, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements. During this call, we will discuss non-GAAP financial measures. Confined and reconciliation of these non-GAAP financial measures GAAP financial measures in our CFO commentary, which is posted on our website. With that, let me turn the call over to Colette.

Colette Kress

Thanks, Stewart. Q4 was another record quarter. Revenue of $39.3 billion was up 12% sequentially and up 78% year on year. And above our outlook, of $37.5 billion. For fiscal 2025, revenue was $130.5 billion. Up 114% in the prior year. Let's start with data center. Data center revenue for fiscal 2025 was $115.2 billion. More than doubling from the prior year. In the fourth quarter, it is in a revenue of $35.6 billion was a record, up 16% sequentially and 93% year on year. As the Blackwell ramp commenced, and Hopper 200 continued to contribute growth, In Q4, Blackwell sales exceeded our expectations. We delivered $11 billion of Blackwell revenue to meet strong demand. This is the fastest product ramp in our company's history. Unprecedented in its speed and scale. Blackwell production is in full gear across multiple configurations, and we are increasing supply quickly. Expanding customer adoption. Our Q4 data center compute revenue jumped 18% sequentially and over 2x year on year. Customers are racing to scale infrastructure to train the next generation of cutting-edge models and unlock the next level of AI capabilities. With Blackwell, it will be common for these clusters to start with 100,000 GPUs or more. Shipments have already started for multiple infrastructures of this size. Post-training and model customization are fueling demand for NVIDIA Corporation infrastructure and software as developers and enterprises leverage techniques such as fine-tuning, reinforcement learning, and distillation to tailor models for domain-specific use cases. Hugging Face alone hosts over 90,000 derivatives traded from the Llama Foundation model. The scale of post-training and model customization is massive and can collectively demand orders of magnitude more compute than pretraining. Our inference demand is accelerating. Driven by test time scaling and new reasoning models. Like OpenAI's O3, DeepSeq R1, and Grok 3. Long-thinking reasoning AI can require 100x more compute per task compared to one-shot inferences. Blackwell was architected for reasoning AI inference. Blackwell supercharges reasoning AI models with up to 25x higher token throughput and 20x lower cost versus Hopper 100. It is revolutionary. Transformer engine is built for LLM. And mixer of experts inference. And its NVLink domain delivers 14x the throughput of PCIe Gen 5. Ensuring the response time, throughput, and cost efficiency needed to tackle the growing complexity of inferences scale. Companies across industries are tapping into NVIDIA Corporation's full-stack inference platform to boost performance and slash cost. Now tripled inference throughput and cut cost by 66% using NVIDIA Corporation TensorRT for its screenshot feature. Perplexity sees 435 million monthly queries and reduced its inference costs 3x with NVIDIA Corporation Triton inference server and TensorRT LLM. Microsoft Bing achieved a 5x speedup at major TCO savings for visual search across billions of images with NVIDIA Corporation, TensorRT, and acceleration libraries. Blackwell has great demand for inference. Many of the early GV200 deployments are earmarked for inference. A first for a new architecture. Blackwell addresses the entire AI market from pretraining, post-training, to inference across clouds, to on-premise, to enterprise. Its programmable architecture accelerates every AI model and over 4,400 applications ensuring large infrastructure investments against obsolescence in rapidly evolving markets. Our performance and pace of innovation are unmatched. We're driven to a 200% reduction in inference cost in just the last two years. We delivered the lowest TCO and the highest ROI. And full-stack optimizations for NVIDIA Corporation and our large ecosystem including 5.9 million developers continuously improve our customers' economics. In Q4, large CSPs represented about half of our data center revenue. And these sales increased nearly 2x year on year. Large CSPs were some of the first to stand up Blackwell, with Azure, GCP, AWS, and OCI, bringing GV200 systems to cloud regions around the world to meet surging customer demand for AI. Regional cloud hosting NVIDIA Corporation GPUs increased as a percentage of data center revenue. Reflecting continued AI factory build-outs globally and rapidly rising demand for AI reasoning models and agents. Coreweave launched a 100,000 GV200 cluster-based instance with NVLink switch and Quantum-2 InfiniBand. Consumer Internet revenue grew 3x year on year. Driven by an expanding set of generative AI and deep learning use cases. These include recommender systems, vision language understanding, synthetic data generation search, and agentic AI. For example, XAI is adopting the GV200 to train and inference its next generation of Grok AI models. Meta's cutting-edge Andromeda, advertising engine runs on NVIDIA Corporation's Grace Hopper Superchip. Serving vast quantities of ads across Instagram, Facebook applications. Andromeda harnesses Grace Hopper's fast interconnect and large memory to boost inference, throughput by 3x. Enhanced ad personalization, and deliver meaningful jumps in monetization and ROI. Enterprise revenue increased nearly 2x year on accelerating demand model fine-tuning. Agentic AI workflows. And GPU-accelerated data processing. We introduced NVIDIA Corporation Llama Numitron model family nodes to help developers create and deploy AI agents across a range of applications, including customer support, fraud detection, and product supply chain and inventory management. Leading AI agent platform providers, including SAP and ServiceNow, are among the first to use new models. Health care leaders, IQVIA, and Lumenon. And Mayo Clinic as well as ARC and Institute are using NVIDIA Corporation AI to speed drug discovery enhance genomic research, and pioneer advanced health care services with generative and agentic AI. As AI expands beyond the digital world, NVIDIA Corporation infrastructure and software platforms are increasingly being adopted to power robotics and physical AI development. One of the early and largest robotics applications and autonomous vehicles were virtually every AV company is developing on NVIDIA Corporation, in the data center. NVIDIA Corporation's automotive vertical revenue is expected to grow to approximately $5 billion this fiscal year. At CES, Hyundai Motor Group announced it is adopting NVIDIA Corporation Technologies to accelerate AV and robotics development and smart factory initiatives. Vision transformers, self-supervised learning, multimodal sensor fusion, and high-fidelity simulation are driving breakthroughs in AV development and will require 10x more compute. At TDX, we announced the NVIDIA

Corporation Cosmos World foundation model platform. Just as language foundation models have revolutionized language AI, Cosmos is a physical AI to revolutionize robotics. The robotics and automotive companies, including ride-sharing giant Uber, are among the first to adopt the platform. From a geographic perspective, sequential growth in our data center revenue was strongest in the US, driven by the initial ramp of Blackwell. Countries across the globe are building their AI ecosystems and demand for compute infrastructure is surging. France's €200 billion AI investment and the EU's €200 billion Invest AI initiative offer a glimpse into the build-out that will redefine global AI infrastructure in the coming years. Now as a percentage of total data center revenue, data center sales in China remained well below levels seen onset of export controls. China shipments absent any change in regulations, we believe that will remain roughly at the current percentage. The market in China for data center solutions remained very competitive. We will continue to comply with export controls while serving our customers. Networking revenue declined 3% sequentially. Our networking attached to GPU compute systems is robust at over 75%. We are transitioning from small NVLink 8 with InfiniBand to large NVLink 72. The Spectrum X. Spectrum X and NVLink switch revenue increased and represents a major new growth sector. We expect networking a return to growth in Q1. AI requires a new class of networking. NVIDIA Corporation offers NVLink switch systems for scale of compute. For scale out, we offer Quantum InfiniBand for HPC supercomputers, and SpectrumX for Ethernet environments. Spectrum X enhances the Ethernet for AI computing and has been a huge success. Microsoft Azure OCI, Fortease, and others are building large AI factories with SpectrumX. The first Stargate data centers will use Spectrum X. Yesterday, Cisco announced integrating Spectrum X into their networking portfolio to help enterprises build AI infrastructure. With its large enterprise footprint and global reach, Cisco will bring NVIDIA Corporation Ethernet to every industry. Now moving to gaming and AR PCs. Gaming revenue of $2.5 billion decreased 22% sequentially and 11% year on year. Full year revenue of $11.4 billion increased 9% year on year. And demand remains strong throughout the holiday. However, Q4 shipments were impacted by supply constraints. We expect strong sequential growth in Q1 as supply increases. The new GeForce RTX 50 series desktop and laptop GPUs are here. Built for gamers, creators, and developers, they fuse AI and graphics, redefining visual computing. Powered by the Blackwell architecture, fifth-generation tensor cores, and fourth-generation RT cores and featuring up to 3,400 AI TOPS. These GPUs deliver a 2x performance leap and new AI-driven rendering, including neural shaders, digital human technologies, geometry, and lighting. The new DLSS 4 boosts frame rates up to 8x with AI-driven frame generation turning one rendered frame into three. It also features the industry's first real-time application of transformer models packing 2x more parameters and 4x to compute unprecedented visual fidelity. We also announced a wave of GeForce Blackwell laptop GPUs with new NVIDIA Corporation Max-Q technology that extends battery life, by up to an incredible 40%. These laptops will be available starting in March from the world's top manufacturers. Moving to our professional visualization business. Revenue of $511 million was up 5% sequentially and 10% year on year. Full year revenue of $1.9 billion increased 21% year on year. Key industry verticals driving demand include automotive and health care. NVIDIA Corporation Technologies and generative AI are reshaping design engineering, and simulation workloads. Increasingly, these technologies are being leveraged in leading software platforms. From ANSYS, Cadence, and Siemens fueling demand for NVIDIA Corporation RTX workstations. Now moving to automotive. Revenue was a record $570 million, up 27% sequentially and up 103% year on year. Full year revenue of $1.7 billion increased 55% year on year. Strong growth was driven by the continued ramp in autonomous vehicles, including cars and robotaxis. At CES, we announced Toyota the world's largest automaker will build its next generation vehicles on NVIDIA Corporation Oren, running the safety-certified NVIDIA Corporation Drive OS. We announced Aurora and Continental. Will deploy driverless trucks at scale powered by NVIDIA Corporation Drive 4. Finally, our end-to-end autonomous vehicle platform, NVIDIA Corporation DRIVE Hyperion, has passed industry safety assessments by Ryland, two of the industry's foremost authorities, automotive-grade safety and cybersecurity, NVIDIA Corporation is the first AV platform to receive a comprehensive set of third-party assessments. Moving to the rest of the P&L. GAAP gross margins, was 73%. And non-GAAP gross margins were 73.5%. Down sequentially as expected with our first deliveries of the Blackwell architecture. As discussed last quarter, Blackwell is a customizable AI infrastructure with several different types of NVIDIA Corporation build chips. Multiple networking options, and for air and liquid-cooled data center. We exceeded our expectations in Q4, in ramping Blackwell, increasing system availability, providing several configurations to our customers. As Blackwell ramps, we expect gross margins to be in the low seventies. We initially, we are focused on expediting the manufacture as they race to build out Blackwell infrastructure. When fully ramped, we have many opportunities to improve the cost and gross margin. Will improve and return to the mid-seventies. Late this fiscal year. Sequentially, GAAP operating expenses were up 9% and non-GAAP operating expenses were 11%, reflecting higher engineering development costs and higher compute and infrastructure costs for new product introductions. In Q4, we returned $8.1 billion to shareholders, the form of share repurchases cash dividends. Let me turn to the outlook in the first quarter. Total revenue is expected to be $43 billion. Plus or minus 2%. Continuing with its strong demand, we expect a significant ramp of Blackwell in Q1. We expect sequential growth. In both data center and gaming. Within data center, we expect sequential growth from both. Compute and networking. GAAP and non-GAAP gross margins are expected to be 70.6%. And 71% respectively. Plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $5.2 billion and $3.6 billion. We expect full year fiscal year 2026 operating expenses grow to grow to be in the mid-thirties. GAAP and non-GAAP other incoming expenses are expected to be an income of approximately $400 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. Further financial details are included in the CFO commentary and other information available on our IR website. Including a new financial information AI agent. In closing, let me highlight upcoming events for the financial community. We will be at the TD Cowen Healthcare Conference in Boston on March 3rd. And at the Morgan Stanley Technology, Media, and Telecom Conference in San Francisco. On March 5th. Please join us for our annual GTC conference starting Monday, March 17th, in San Jose, California. Jensen will deliver a news-packed keynote on March 18th, and we will host a Q&A session for our financial analysts. Next day, March 19th. We look forward to seeing you at these events. Our earnings call to discuss the results for our first quarter of fiscal 2026 is scheduled for May 28th, 2025. We are going to open up the call, operator. To questions. If you could start that, that would be great.

Christa

Thank you. At this time, I would like I also ask that you please limit yourself to one question. For any additional questions, please requeue. And your first question comes from C.J. Muse with Cantor Fitzgerald. Please go ahead.

C.J. Muse

Yeah. Good afternoon. Thank you for taking the question. I guess, for me, Jensen, as test time compute and reinforcement learning shows such promise, we're clearly seeing increasing blurring in the lines between training and inference. What does this mean for the potential future of potentially inference-dedicated clusters? And how do you think about the overall impact to NVIDIA Corporation and your customers? Thank you.

Jensen Huang

Yeah. I appreciate that, C.J. There are now multiple scaling laws. There's the pretrained scaling laws. And that's gonna continue to scale because we have multimodality. We have data that came from reasoning that are now used to pretraining. And then the second is post-training scaling law. Using reinforcement learning human feedback, reinforcement learning AI feedback, reinforcement learning verifiable rewards, the amount of computation you use for post-training is actually higher than pretraining. And it's kinda sensible in the sense that you could while you're using reinforcement learning, generate an enormous amount of synthetic data or synthetically generated tokens. AI models are basically generating tokens to train AI models. That's post-train. And the third part, this is the part that you mentioned, is test time compute or reasoning. Long thinking, inference scaling, basically the same ideas. And there's you have chain of thought, you have search. The amount of tokens generated, the amount of inference compute needed, is already a hundred times more than the one-shot examples and the one-shot capabilities of large language models in the beginning and that's just the beginning. This is just the beginning. The idea that the next generation could have thousands of times and even hopefully extremely thoughtful and simulation-based and search-based models that could be hundreds of thousands, millions of times more compute than today, is in our future. And so the question is how do you design such an architecture? Some of the models are autoregressive. Some of the models are diffusion-based. Some of the times you want your data center to have disaggregated inference. Sometimes it's compacted. And so it's hard to figure out what is the best configuration of a data center, which is the reason why NVIDIA Corporation's architecture is so popular. We run every model. We are great at training. The vast majority of our compute today is actually inference, and Blackwell takes all of that to a new level. We designed Blackwell with the idea of reasoning models in mind. And you look at training, it's many times more performant. But what's really amazing is for long-thinking, test time scaling reasoning AI models, we're tens of times faster, 25 times higher throughput. And so Blackwell is gonna be incredible across the board. And when you have a data center, that allows you to configure and use your data center based on are you doing more pretraining now, post-training now? Or scaling out your inference our architecture is fungible, and easy to use. In all of those different ways. And so we're seeing, in fact, much, much more concentration of a unified architecture than ever before.

Christa

Your next question comes from the line of Joseph Moore with JPMorgan. Please go ahead.

Joseph Moore

I wonder if you could talk about GV200 at CES. You sort of talked about the complexity of the rack-level systems and the challenges you have. And then as you said in the prepared remarks, we've seen a lot of general availability. You know, where are you in terms of that ramp? Are there still bottlenecks to consider at a systems level above and beyond the chip level? And just you know, have you maintained your enthusiasm for the NVLink 72 platforms?

Jensen Huang

Well, I'm more enthusiastic today than I was at CES. And the reason for that is because we shipped a lot more to CES. We have some 350 plants manufacturing the one and a half million components that go into each one of the Blackwell racks. Base Blackwell racks. Yes. It's extremely complicated. And we successfully and incredibly ramped up Grace Blackwell. Delivering some $11 billion of revenues last quarter. We're gonna have to continue to scale as demand is quite high and customers are anxious and impatient to get their Blackwell systems. You'd probably seen on the web a fair number of celebrations about Grace Blackwell Systems coming online and we have them, of course. We have a fairly large installation of Grace Blackwell for our own engineering and our own design teams and software teams. Coreweave has now gone public about the successful bring-up of theirs. Microsoft has. Of course, OpenAI has. And you're starting to see many come online. So I think the answer to your question is nothing is easy about what we're doing. But we're doing great, and all of our partners are doing great.

Christa

Your next question comes from the line of Vivek Arya with Bank of America Securities. Please go ahead.

Vivek Arya

Thank you for taking my question. Could I just you wouldn't mind confirming if Q1 is the bottom for gross margins? And then, Jensen, my question is for you. What is on your dashboard to give you the confidence that the strong demand can sustain into next year and has DeepSeq and whatever innovations they came up with, has that changed that view in any way? Thank you.

Colette Kress

Let me first take the first part of the question. Regarding the gross margin. During our Blackwell ramp, our gross margins will be in the low seventies. At this point, we are focusing on expediting our manufacturing. Expediting our manufacturing is to make sure that we can provide customers as soon as possible. Our Blackwell is fully ramped. And once it does, I'm sorry. Blackwell fully ramps, we can improve our cost and our gross margin. So we expect to probably be in the mid-seventies later this year. You know, walking through what you heard, Jensen speak about the systems and their complexity. They are customizable in some cases. They've got multiple networking options. Have liquid cool and water-cooled. So we know there is an opportunity for us to improve these gross margins going forward. But right now, we are gonna focus on getting the manufacturing plate into our customers as soon as possible.

Jensen Huang

We know several things, Vivek. We have a fairly good line of sight of the amount of capital investment that data centers are building out towards. We know that going forward, the vast majority of software is gonna be based on machine learning. And so accelerated computing and generative AI, reasoning AI, are going to be the type of architecture you want in your data center. We have, of course, forecast and plans from our top partners. And we also know that there are many innovative really exciting start-ups that are still coming online. As new opportunities for developing the next breakthroughs in AI, whether it's agentic AIs, reasoning AIs, or physical AIs. The number of start-ups are still quite vibrant and each one of them needs a fair amount of computing infrastructure. So I think the whether it's the near-term signals or the mid-term signals. Near-term signals, of course, are, you know, POs and forecasts and things like that. Mid-term signals, would be the level of infrastructure and CapEx scale out compared to previous years. And then the long-term signals it has to do with the fact that we know fundamentally software has changed. From hand coding that runs on CPUs through machine learning and AI-based software that runs on GPUs and accelerated computing systems. So we have a fairly good sense that this is the future of software. And then maybe as you roll it out, another way to think about that is we've really only touched consumer AI and search and some amount of consumer generative AI. Advertising, recommenders, kind of the early days of software. The next wave's coming. Agentic AI for enterprise, physical AI for robotics. And Sovereign AI has different regions build out their AI for their own ecosystems. And so each one of these are barely off the ground, and we can see them. We can see them because, you know, obviously, we're in the center of much of this development. And we can see great activity happening in all these different places. And these will happen. So near-term, mid-term, long-term.

Christa

Your next question comes from the line of Matt Ramsay with Cowen. Please go ahead.

Matt Ramsay

Yeah. Good afternoon. Thanks for taking my question. Your next generation Blackwell Ultra is set to launch in the second half of this year. In line with the team's annual product cadence. Jensen, can you help us understand the demand dynamics for Ultra given that you'll still be ramping the current generation Blackwell solutions? How do your customers and the supply chain also manage the simultaneous ramps of these two products and is the team still on track to execute Blackwell Ultra in the second half of this year?

Jensen Huang

Yes. Blackwell Ultra is second half. As you know, the first Blackwell was have we had a hiccup? That probably cost us a couple of months. We're fully recovered, of course. The team did an amazing job recovery. And all of our supply chain partners and just so many people helped us recover at the speed of light. And so now we've successfully ramped production of Blackwell. But that doesn't stop the next train. The next train is you know, it's on an annual rhythm. And, Blackwell Ultra with, new networking, new memories, and, of course, new processors and all of that is coming online. We've been working with all of our partners and customers laying this out. They have all of the necessary information. And we'll work with everybody to do the proper transition. This time between Blackwell, Blackwell Ultra, the system architecture is exactly the same. It's a lot harder going from Hopper to Blackwell because we went from an NVLink 8 system to a NVLink 72 base system. So the chassis, the architecture of the system, the hardware, the power delivery, all of that had to change. This was quite a challenging transition. But the next transition will slot right in. Grace Blackwell Ultra will slot right in. We've also already revealed and been working very closely with all of our partners on the click after that. And the click after that is called Vera Rubin. And, all of our partners are getting up to speed on the transition of that. And so preparing for that transition and, again, we're gonna provide a big, big, huge step up. And so come to GTC, and I'll hold on to you about Blackwell Ultra, Vera Rubin, and then show you what's the one click after that. Really, really exciting new product, so come to GTC, please.

Christa

Your next question comes from the line of Timothy Arcuri with UBS. Please go ahead.

Timothy Arcuri

Thanks a lot. Jensen, we hear a lot about custom ASICs. Can you kinda speak to the balance between custom ASIC and merchant GPU? We hear about some of these heterogeneous super clusters to use both GPU and ASIC. Is that something customers are planning on building or will these infrastructures remain fairly distinct? Thanks.

Jensen Huang

Well, we build very different things than ASICs. In some ways, completely different in some areas we intercept. We're different in several ways. One, NVIDIA Corporation's architecture is general. You know, whether you've optimized for autoregressive models or diffusion-based models or vision-based models or multimodal models or text models. We're great in all of it. We're great in all of it because our software stack is so our architecture is responsible. Our software stack is ecosystem is so rich that we're the initial target of, you know, most exciting innovations and algorithms. And so by definition, we're much, much more general than narrow. We're also really good from the end to end. From data processing, the curation of the training data, to the training of the data, of course, to reinforcement learning used in post-training. All the way to inference with test time scaling. So, you know, we're general. We're end to end. And we're everywhere. And because we're not in just one cloud, we're in every cloud, we could be on-prem. We could be in, you know, in a robot. Our architecture is much more accessible. And a great target initial target for anybody who's starting up a new company. And so we're everywhere. And then the third thing I would say is that our performance and our rhythm is so incredibly fast. Remember that these data centers are always fixed in size. They're fixed in size or they're fixed in power. And if our performance per watt is anywhere from 2x to 4x to 8x, which is not unusual. It translates directly to revenues. And so if you have a 100-megawatt data center, if the performance or the throughput that 100-megawatt or that gigawatt data center is four times or eight times higher your revenues for that gigawatt data center is eight times higher. And the reason that is so different than data centers of the past is because AI factories are directly monetizable through its tokens generated. And so the token throughput of our architecture being so incredibly fast is just incredibly valuable to all of the companies that are building these things for revenue generation reasons. And capturing the fast ROIs. So I think the third reason is performance. And then the last thing that I would say is the software stack is incredibly hard. Building an ASIC is no different than what we do. We have to build a new architecture. And the ecosystem that sits on top of our architecture is ten times more complex today than it was two years ago. And that's fairly obvious because the amount of software this world building on top of architecture is growing exponentially and AI is advancing very quickly. So bringing that whole ecosystem on top of multiple chips is hard. And so I would say that those four reasons and then finally, I will say this. Just because the chip is designed doesn't mean it gets deployed. And you've seen this over and over again. There are a lot of chips that get built. But when the time comes a business decision has to be made. And that business decision is about deploying a new engine, a new processor into a limited AI factory in size and power and find. And our technology is, you know, not only more advanced, more performant, it has much, much better software capability, and very importantly, our ability to deploy is lightning fast. And so these things are enough for the faint of heart as everybody knows now. And so there's a lot of different reasons why we do well. Why we win.

Christa

Your next question comes from the line of Ben Reitzes with Melius Research. Please go ahead.

Ben Reitzes

Yeah. Hi. Ben Reitzes here. Hey. Thanks a lot for the question. Hey, Jensen. It's a geography-related question. You know, you did a great job explaining some of the demand underlying, you know, factors here on the strength. But the US was up about $5 billion or so sequentially. And I think, you know, there is a concern about whether the US can pick up the slack if there's regulations towards other geographies. And I was just wondering as we go throughout the year, you know, if this kind of surge in the US continues and it's gonna be whether that's okay. And if that underlies your growth rate, how can you keep growing so fast with this mix shift towards the US? Your guidance looks like China is probably up sequentially. So just wondering if you could go through that dynamic and maybe Colette can weigh in. Thanks a lot.

Jensen Huang

China is approximately the same percentage as Q4. And as in as previous quarters. It's about half of what it was before the export control. But it's approximately the same in percentage. With respect to geographies, the takeaway is that AI is software. It's modern software. It's incredible modern software. But it's modern software. And AI has gone mainstream. AI is used in delivery services everywhere, shopping services everywhere. You know? You were to buy a quart of milk is delivered to you. AI was involved. And so almost everything that a consumer service provides AI's at the core of it. Every student will use AI as a tutor. Health care services use AI. Financial services use AI. No fintech company will not use AI. Every fintech company will. Climate tech company uses AI. Mineral Discovery now uses AI. The number of every higher education, every university, uses AI. So I think it is fairly safe to say that AI has gone mainstream. And that it's being integrated into every application. And our hope is that, of course, the technology continues to advance safely and advance in a helpful way to our society. And with that, you know, we're I do believe that we're at the beginning of this new transition. And what I mean by that in the beginning, is remember behind us has been decades of data centers and decades of computers that have been built. And they've been built for a world of hand coding and general-purpose computing. And CPUs and so on and so forth. And going forward, I think it's fairly safe to say that that world is going to be almost all software will be infused with AI. All software and all services will be based on ultimately based on machine learning, and the data flywheel is gonna part of improving software and services. And that the future computers will be accelerated. The future computers will be based on AI. And we're really three years into that journey. And in modernizing computers that have taken decades to build out. And so I'm fairly sure that we're in the beginning of this new era. And then lastly, no technology has ever had the opportunity to address a larger part of the world's GDP than AI. No software tool ever has. And so this is now a software tool that can address a much larger part of the world's GDP, more than any time in history. And so the way we think about growth and the way we think about whether something is big or small. Has to be in the context of that. And when you take a step back and look at it from that perspective, we're really just in the beginnings.

Christa

Your next question comes from the line of Aaron Rakers with Wells Fargo. Please go ahead. Erin, your line is open. Your next question comes from Mark Lipacis with Evercore ISI. Please go ahead.

Marshall Pappas

Hi. This is Marshall Pappas. Thanks for taking the question. Question. I had a clarification and a question. Colette, for the clarification. Did you say that enterprise within the data center grew 2x year on year for the January quarter? And if so, does that would that make it the faster growing than the hyperscalers? And then, Jensen, for you, the question, hyperscalers are the biggest purchasers of your solutions, but they buy equipment for both internal and external workloads, external workloads being cloud services that enterprises use. So the question is, can you give us a sense of how that hyperscale expense splits between that external workload and internal and as these new AI workflows and applications come up, would you expect enterprises to become a larger part of that consumption mix? And does that impact how you develop your service your ecosystem? Thank you.

Colette Kress

Sure. Thanks for the question regarding our enterprise business. Yes. It grew 2x. Very similar to what we were seeing with our large CSPs. Keep in mind, these are both important areas to understand. Working with the CSPs can be working on large language models. Can be working on inference on their own work? But keep in mind, that is also where the enterprises are surfacing. Your enterprises are both with your CSPs, as well as in terms of building on their own. They're both growing quite well.

Jensen Huang

The CSPs are about half of our business. And the CSPs have internal consumption, and external consumption, as you say. And we're using of course, used for internal consumption. We work very closely with all of them to optimize workloads that are internal to them because they have a large infrastructure of NVIDIA Corporation gear that they could take advantage of. And the fact that we could be used for AI on the one hand, video processing on the other hand, data processing like Spark. We're fungible. And so the useful life. Our infrastructure is much better. If the useful life is much longer, then the TCO is also lower. And so the second part is how do we see the growth of enterprise or not CSPs, if you will, going forward? And the answer is I believe, long term. It is by far larger. And the reason for that is because if you look at the computer industry today, and what is not served by the computer industry is largely industrial. Let me give you an example. When we say enterprise, and let's say let's use a car company as an example because they make both soft things and hard things. And so in the case of a car company, the employees would be what we call enterprise. And agentic AI and software planning systems and tools, and we have some really exciting things to with you guys at GTC. Those agentic systems are for employees to make employees more productive. To design, to market, plan, to operate their company. That's agentic AIs. On the other hand, the cars that they manufacture also need AI. They need an AI system that trains the cars treats this entire giant fleet of cars, and you know, today, there's some billion cars on the road. Someday, there'd be a billion cars on the road, and every single one of those cars will be, you know, robotic cars. And they'll all be collecting data, and we'll be improving them using an AI factory where they whereas they have a car factory today, in the future, they'll have a car factory and an AI factory. And then inside the car itself is a robotic system. And so as you can see, there are three computers involved. And there's the computer that helps the people. There's the computer that builds the AI for it. The machineries. It could be, of course. Could be a tractor. It could be a lawnmower. It could be a human or a robot that's being developed today. It could be a building. It could be a warehouse. These physical systems require a new type of AI we call physical AI. They can't just understand the meaning of words and languages but they have to understand the meaning of the world. Friction and inertia, object permanence, and cause and effect, and all of those types of things that are common sense to you and I. But you know, AI has to go learn those physical effects. So we call that physical AI. That whole part of using agentic AI to revolutionize the way we work inside companies. That's just starting. This is now the beginning of the agentic AI era. And you hear a lot of people talking about it and got some really great things going on. And then there's the physical AI after that, and then there's robotic systems after that. And so these three computers are all brand new. And my sense is that long term, this will be by far a larger of a mold which kinda makes sense. You know, the world the world's GDP is represented by either heavy industries industrials. And companies that are providing for those.

Christa

Your next question comes from the line of Aaron Rakers with Wells Fargo. Please go ahead.

Aaron Rakers

Yeah. Thanks for letting me back in. Jensen, I'm curious as we now approach the two-year anniversary of really the Hopper inflection that you saw in 2023 in Gen AI in general. We think about the roadmap you have in front of us, how do you think about the infrastructure that's been deployed from a replacement cycle perspective and whether, you know, if it's GV300 or if it's the Rubin cycle where we start to see maybe some refresh opportunity. I'm just curious to how you look at that.

Jensen Huang

Yeah. I appreciate it. First of all, people are still using Voltas. And Pascals, and Amperes. And the reason for that is because they're always things that because CUDA is so programmable, you could use it right well, one of the major use cases right now is data processing and data curation. You find a circumstance that an AI model is not very good at? You present that circumstance to a vision language model, let's say. Let's say it's a car? You present that circumstance to a vision language model, the vision language model actually looks at the circumstances. It's a this isn't this is what happened, and I wasn't very good at it. You then take that response, this the prompt, and you go and prompt an AI model to go find in your whole link of data of other circumstances like that. Whatever that circumstance was. And then you use an AI to do domain randomization and generate a whole bunch of other examples. And then from that, you can go train the model. And so you could use the Amperes to go and do data processing and data curation and machine learning-based search. And then you create the training dataset, which you then present to your Hopper systems for training. And so each one of these architectures are completely are you know, they're all CUDA compatible, and so everything runs on everything. But if you have infrastructure in place, and you can put the less intensive workloads onto the installed base of the past. All of our CPUs are very well employed.

Christa

We have time for one more question, and that question comes from Atif Malik with Citi. Please go ahead.

Atif Malik

Hi. Thank you for taking my question. I have a follow-up question on gross margins, Colette. I understand there are many moving parts that will yield and NVLink 72 and Ethernet mix. And you kind of tiptoed the earlier question if April quarter is the bottom. But second half would have to ramp, like, 200 basis point per quarter to get to the mid-seventies range that you're giving, for the end of the fiscal year. And we still don't know much about tariffs impact to broader semiconductor. So what kind of gives you the confidence in that trajectory in the back half of this year?

Colette Kress

Yeah. Thanks for the question. Our gross margins, they're quite complex. In terms of the material. And everything that we put together in a Blackwell system. Tremendous amount of opportunity to look at a lot of different pieces of that. On how we can better improve our gross margins over time. Remember, we have many different configurations as well. On Blackwell. That will be able to help us do that. So, together, working after we get some of these really strong ramping completed for our customers we can begin a lot of that work. If not, we're gonna probably start as soon as possible. If we can improve it in the short term, we will also do that. Tariffs, at this point, it's a little bit of an unknown. It's an unknown until we understand further what the US government's plan is, its timing, it's where, and how much. So at this time, we are awaiting but again, we would, of course, always follow export control and or tariffs in that manner.

Christa

Ladies and gentlemen, that does conclude our question and answer session. I'm sorry. Thank you.

Jensen Huang

No. No. I'm gonna just wanna thank you. Up to, Jensen? And, like, the medium, a couple things. I just wanna thank you. Thank you, Colette. Demand for Blackwell is extraordinary. AI is evolving beyond perception. And generative AI into reasoning. With reasoning AI, we're observing another scaling law. Inference time or test time scaling. The more computation the more the model thinks the smarter the answer. Models like OpenAI's Grok 3, DeepSeq R1, are reasoning models that apply inference time scale. Reasoning models can consume a hundred times more compute. Future reasoning models can consume much more compute. DeepSeq R1 has ignited global enthusiasm. It's an excellent innovation. But even more importantly, it has open-sourced a world-class reasoning AI model. Nearly every AI developer is applying R1. Or chain of thought and reinforcement learning techniques like R1. To scale their model's performance. We now have three scaling laws, as I mentioned earlier. Driving the demand for AI computing. The traditional scaling laws of AI remain intact. Foundation models are being enhanced with multimodality. And pretraining is still growing. But it's no longer enough. We have two additional scaling dimensions. Post-training scaling, where reinforcement learning fine-tuning, model distillation, require orders of magnitude more compute than pretraining alone. Inference time scaling and reasoning where a single query can demand a hundred times more compute. We designed Blackwell for this moment a single platform that can easily transition from pretraining, post-training, and test time scaling. Blackwell's MP4 transformer engine, and NVLink 72 scale-up fabric. And new software technologies let Blackwell process reasoning AI models 25 times faster than Hopper. Blackwell, in all of these configurations, is in full production. Each Grace Blackwell NVLink 72 rack is an engineering marvel. One and a half million components produced across 350 manufacturing sites by nearly a hundred thousand factory operators. AI is advancing at light speed. We're at the beginning of reasoning AI and inference time scaling. But we're just at the start of the age of AI. Multimodal AIs. Enterprise AI, Sovereign AI. And physical AI are right around the corner. We will grow strongly in 2025. Going forward, data centers will dedicate most of CapEx to accelerated computing and AI. Data centers will increasingly become AI factories. And every company will have a either rented or self-operated. I wanna thank all of you for joining us today. Come join us at GTC in a couple of weeks gonna be talking about Blackwell Ultra, Rubin, and other new computing networking, reasoning AI, physical AI products. And a whole bunch more. Thank you.

Christa

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