**NVIDIA Earnings Call**

Conference Operator: Good afternoon. My name is Krista, and

Stuart Stecker, Investor Relations, NVIDIA: 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.

Conference Operator: After the speakers' remarks, there will be a question and answer session. Thank you. Stuart Stecker, you may begin your conference.

Stuart Stecker, Investor Relations, NVIDIA: Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the fourth quarter of fiscal twenty twenty five. With me today from NVIDIA are Jensen Wong, President and Chief Executive Officer and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the first quarter of fiscal twenty twenty six.

The content of today's call is NVIDIA'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. 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 ks and 10 q, and the reports that we may file on Form eight ks with the Securities and Exchange Commission.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

All our statements are made as of today, 02/26/2025, 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 and 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 Kress, Executive Vice President and Chief Financial Officer, NVIDIA: Thanks, Stuart. Q4 was another record quarter. Revenue of $39,300,000,000 was up 12% sequentially and up 78% year on year, and above our outlook of thirty seven point five billion dollars For fiscal twenty twenty five, revenue was $130,500,000,000 up 114% in the prior year. Let's start with data center. Data center revenue for fiscal twenty twenty five was $115,200,000,000 more than doubling from the prior year.

In the fourth quarter, data center revenue of $35,600,000,000 was a record of 16% sequentially and 93% year on year. As the Blackwell ramp commenced and Hopper 200 continued sequential growth. In Q4, Blackwell sales exceeded our expectations. We delivered $11,000,000,000 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Post training and model customization are fueling demand for NVIDIA 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 created from the LAMA Foundation model. The scale of post training and model customization is massive and can collectively demand orders of magnitude more compute than pre training. Our inference demand is accelerating, driven by test time scaling and new reasoning models like OpenAI's O3, DeepSeq R1 and GROC3. Long thinking reasoning AI can require 100 X more compute per task compared to one shot inferences.

Blackwell was architected for reasoning AI inference. Blackwell supercharges reasoning AI models with up to 25 X higher token throughput and 20 X lower cost versus hover 100. It is revolutionary. Transformer engine is built for LLM and mixture of experts infants. And its NVLink domain delivers 14 X the throughput of PCIe Gen five, ensuring the response time, throughput and cost efficiency needed to tackle the growing complexity of inference at scale.

Companies across industries are tapping into NVIDIA's full stack inference platform to boost performance and slash costs. Now tripled inference throughput and cut costs by 66% using NVIDIA TensorRT for its screenshot feature. Perplexity sees four thirty five million monthly queries and reduced its inference costs 3 X with NVIDIA Triton Inference Server and TensorRT LLM. Microsoft (NASDAQ:MSFT) Bing achieved a five x speed up and major TCO savings for visual search across billions of images with NVIDIA, TensorRT and acceleration libraries. Blackwell has great demand for inference.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Many of the early GB200 deployments are earmarked for inference, a first for a new architecture. Blackwell addresses the entire AI market from pre training, post training to inference across clouds to on premise to enterprise. CUDA's programmable architecture accelerates every AI model and over 4,400 applications, ensuring large infrastructure investments against obsolescence in rapidly evaluating markets. Our performance and pace of innovation is unmatched. We're driven to a 200X 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 and our large ecosystem, including 5,900,000 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 GB200 systems to cloud regions around the world to meet surging customer demand for AI.

Regional cloud hosting NVIDIA 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. Or we've launched a 100,000 GV200 cluster based instance with NVLink switch and Quantum (NASDAQ:QMCO) two 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.

Jensen Huang, President and Chief Executive Officer, NVIDIA: For

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Colette Kress, Executive Vice President and Chief Financial Officer, NVIDIA: example, XAI is adopting the GB200 to train and inference its next generation of GROG AI models. Meta's cutting edge Andromeda advertising engine runs on NVIDIA's Grace Hopper Superchip, serving vast quantities of ads across Instagram, Facebook (NASDAQ:META) applications. Andromeda harnesses Grace Hopper's fast interconnect and large memory to boost inference throughput by three X, enhance ad personalization and deliver meaningful jumps in monetization and ROI. Enterprise revenue increased nearly 2X year on accelerating demand for model fine tuning, RAG and agentic AI workflows and GPU accelerated data processing. We introduced NVIDIA llama Numeron model family NIMS 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 (NYSE:NOW), are among the first to use new models. Healthcare leaders IQVIA, Illumina (NASDAQ:ILMN) and Mayo Clinic are well, as ARC Institute, are using NVIDIA AI to speed drug discovery, enhance genomic research and pioneer advanced healthcare services with generative and agentic AI. As AI expands beyond the digital world, NVIDIA 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 where virtually every AV company is developing on NVIDIA in the data center, the car or Bulls. NVIDIA's automotive vertical revenue is expected to grow to approximately $5,000,000,000 this fiscal year.

At CES, at CES, Hyundai Motor (OTC:HYMTF) Group announced it is adopting NVIDIA 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 AB development and will require 10X more compute. At TEDx, we announced the NVIDIA Cosmo 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 (NYSE:UBER), are among the first to adopt the platform.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

From a geographic perspective, sequential growth in our data center revenue was strongest in The US, driven by the initial ramp of BlackRock (NYSE:BLK). Countries across the globe are building their AI ecosystems and demand for compute infrastructure is surging. France's Two Hundred Billion Euro AI investment and the EU's two hundred billion euros Invest AI initiatives offer a glimpse into the build out to set redefined 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 on the onset of export controls. Absent any change in regulations, we believe that China shipments will remain roughly at the current percentage.

The market in China for data center solutions remains 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 eight with InfiniBand to large NVLink 72 with SpectrumX.

SpectrumX and NVLink switch revenue increased and represents a major new growth sector. We expect networking to return to growth in Q1. AI requires a new class of networking. NVIDIA offers NVLink switch systems for scale up compute. For scale out, we offer Quantum InfiniBand for HPC supercomputers and SpectrumX for Ethernet environments.

SpectrumX enhances the Ethernet for AI computing and has been a huge success. Microsoft Azure, OCI, CoreWeave and others are building large AI factories with SpectrumX. The first Stargate data centers will use SpectrumX. Yesterday, Cisco (NASDAQ:CSCO) announced integrating SpectrumX into their networking portfolio to help enterprises build AI infrastructure. With its large enterprise footprint and global reach, Cisco will bring NVIDIA Ethernet to every industry.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Now, moving to gaming and AI PCs. Gaming revenue of $2,500,000,000 decreased 22% sequentially and 11% year on year. Full year revenue of $11,400,000,000 increased 9% year on year. And demand remained 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 top. These GPUs deliver a 2x performance leap and new AI driven rendering, including neural shaders, digital human technologies, geometry and lighting.

The new DLSS four 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 the compute for unprecedented visual fidelity. We also announced a wave of GeForce Blackwell laptop GPUs with new NVIDIA 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 $511,000,000 was up 5% sequentially and 10% year on year. Full year revenue of $1,900,000,000 increased 21% year on year. Key industry verticals driving demand include automotive and healthcare. NVIDIA technologies and generative AI are reshaping design, engineering and simulation workloads. Increasingly, these technologies are being leveraged in leading software platforms from ANSYS (NASDAQ:ANSS), Cadence and Siemens (ETR:SIEGn), fueling demand for NVIDIA RTX workstation.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Now moving to automotive. Revenue was a record $570,000,000 up 27% sequentially and up 103% year on year. Full year revenue of 1,700,000,000 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 (NYSE:TM), the world's largest automaker, will build its next generation vehicles on NVIDIA Orin, running the safety certified NVIDIA DRIVE OS.

We announced Aurora and Continental will deploy driverless trucks at scale powered by NVIDIA DRIVE four. Finally, our end to end autonomous vehicle platform NVIDIA DRIVE Hyperion has passed industry safety assessments by Toussud and Toussud Ryland, two of the industry's foremost authorities for automotive grade safety and cybersecurity. NVIDIA is the first AV platform to receive a comprehensive set of third party assessments. Okay. Moving to the rest of the P and 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 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 70s. We initially, we are focused on expediting the manufacturing of Blackwell systems to meet strong customer demand as they race to build out Blackwell infrastructure.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

When fully ramped, we have many opportunities to improve the cost and gross margin will improve and return to the mid-70s 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 8,100,000,000 to shareholders in the form of share repurchases and cash dividends. Let me turn to the outlook in the first quarter. Total revenue is expected to be $43,000,000,000 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 70.671% respectively, plus or minus 50 basis points. GAAP and non GAAP operating expenses are expected to be approximately $5,200,000,000 and $3,600,000,000 respectively.

We expect full year fiscal year 2026 operating expenses to grow to be in the mid-30s. GAAP and non GAAP other income and expenses are expected to be an income of approximately $400,000,000 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

We will be at the TD Cowen Healthcare Conference in Boston on March 3, and at the Morgan Stanley Technology, Media and Telecom (BCBA:TECO2m) Conference in San Francisco on March 5. Please join us for our annual GTC conference starting Monday, March 17 in San Jose, California. Jensen will deliver a news packed keynote on March 18, and we will host a Q and A session for our financial analysts the next day, March 19. We look forward to seeing you at these events. Our earnings call to discuss the results for our first quarter of fiscal twenty twenty six is scheduled for 05/28/2025.

We are going to open up the call, operator, to questions. If you could start that, that would be great.

Conference Operator: And your first question comes from C. J. Muse with Cantor Fitzgerald. Please go ahead.

C.J. Muse, Analyst, Cantor Fitzgerald: Yes. Good afternoon. Thank you for taking the question. I guess, for me, Judson, as tough comp computing 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 and your customers? Thank you.

Jensen Huang, President and Chief Executive Officer, NVIDIA: Yes, I appreciate that, CJ. There are now multiple scaling laws. There's the pre training scaling law. And that's going to continue to scale because we have multi modality, we have, data that came from, reasoning that are now used to, do pretraining. And then the second is post training scaling law using reinforcement learning human feedback, reinforcement learning AI feedback, reinforcement learning verifiable rewards.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

The amount of computation you use for post training is actually higher than pre training. And it's kind of 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. And that's post training. And the third the third part, this is the part that you mentioned, is test time compute or reasoning, long thinking, inference scaling.

They're all basically the same ideas. And there is 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 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, is, in our future. And so, so the question is how do you design such an architecture? Some of it some of the models are autoregressive, some of the models are diffusion based, Some of it some of the times you want your, data center to have disaggregated inference, sometimes it's compacted. And so, it's hard to it's hard to figure out, what is the best configuration of a data center, which is the reason why NVIDIA's architecture is so popular. We run every model.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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 when 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 were tens of times faster, 25 times higher throughput. And so, Blackwell is going to 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 pre training 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.

Conference Operator: Your next question comes from the line of Joe Moore with JPMorgan. Please go ahead.

Joe Moore, Analyst, JPMorgan: Good morning, Stanley, actually. Thank you. I wonder if you could talk about GB200 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.

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 have you maintained your enthusiasm for the NVL 72 platforms?

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Jensen Huang, President and Chief Executive Officer, NVIDIA: Well, I'm more enthusiastic today than I was at CES. And the reason for that is because we've shipped a lot more to CES. We have some three fifty plants, manufacturing the 1,500,000 components that go into each one of the Blackwell racks, Grace Blackwell racks. Yes, it's extremely complicated and, we successfully, and incredibly ramped up Grace Blackwell, delivering some $11,000,000,000 in revenues last quarter. We're going to have to continue to scale as demand is quite high and customers are anxious and impatient to get their Blackwell systems.

You've 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's for our own engineering and our own design teams and software teams. CoreWeave has now, been quite public about the successful bring up of theirs. Microsoft has. Of course, OpenAI has.

And you're starting to see many, many come online. And, 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.

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

Vivek Arya, Analyst, Bank of America Securities: Thank you for taking my question. Could I just you wouldn't mind confirming if Q1 is the bottom for gross margin? 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 DeepSeek and whatever innovations they came up with, has that changed that view in any way?

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Thank you.

Colette Kress, Executive Vice President and Chief Financial Officer, NVIDIA: Let me first take the first part of the question, there regarding the gross margin. During our Blackwell ramp, our gross margins, will be in the low 70s. At this point, we are focusing on expediting our manufacturing, expediting our manufacturing to make sure that we can provide the customers as soon as possible. Our Blackwell has fully ramped, and once it does I'm sorry, once our Blackwell fully ramps, we can improve our cost and our gross margin. So we expect to probably be in the mid-70s later this year.

You know, walking through what you heard, Johnson speak about the systems and their complexity, they are customizable in some cases. They've got multiple networking options. They have liquid cooled and water cooled. So we know there is an opportunity for us to improve these gross margins going forward. But right now, we are going to focus on getting the manufacturing complete and to our customers as soon as possible.

Jensen Huang, President and Chief Executive Officer, NVIDIA: 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 going to 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, forecasts and plans from our top partners.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

And, we also know that there are many innovative, really exciting startups 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 startups are still quite vibrant and each one of them need a fair amount of computing infrastructure. And so, I think the whether it's the near term signals or the midterm signals near term signals, of course, are POs and forecasts and things like that. Midterm signals would be, the level of infrastructure and CapEx scale out compared to previous years. And then the long term signals has to do with the fact that we know fundamentally software has changed from hand coding that runs on CPUs to machine learning and AI based software that runs on GPUs and accelerated computing systems.

And 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 tapped 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 as different regions build out their AI for their own ecosystems.

And so, each one of these are fairly off the ground and we can see them. We can see them because, 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Conference Operator: Your next question comes from the line of Harlan Sur with JPMorgan. Please go ahead.

Harlan Sur, Analyst, JPMorgan: Yes, good afternoon. Thanks for taking my question. Your next generation Blackwall Ultra 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, President and Chief Executive Officer, NVIDIA: Yes. Blackwell Ultra is second half. As you know, the first Blackwell was, we had a hiccup that probably cost us a couple of months. We're fully recovered, of course. The team did an amazing job recovering.

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 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 and 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 eight system to a NVLink 72 based system. So the chassis, the architecture of the system, the hardware, the power delivery, all of that had to change.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

This was quite a challenging transition. But the next transition will slot right in, right? Blackwall 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 Verarubin.

And, all of our partners are, getting up to speed on, on the transition of that. And so preparing for that transition. And again, we're going to provide a big, big, huge step up. And so come to GCC and I'll talk to you about Blackwell Ultra, Verarubin and then show you what's the one click after that. Really exciting new products.

So come to GTC, please.

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

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

Thanks.

Jensen Huang, President and Chief Executive Officer, NVIDIA: Well, we build very different things than ASICs, in some ways completely different in some areas we intersect. We're different in several ways. One, NVIDIA's architecture is general. Whether you're you've optimized for autoregressive models or diffusion based models or vision based models or multimodal models or text models, we're great at all of it. We're great at all of it because our software stack is so our architecture is flexible, our software stack is ecosystem is so rich that we're the initial target of most exciting innovations and algorithms.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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, 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 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 in 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. And so, I think the third reason is performance. And then, the last thing that I would say is the software stack is incredibly hard.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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 10 times more complex today than it was two years ago. And that's fairly obvious because the amount of software that the world is 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 gets 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, in power and in time. And our technology, is 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 not 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.

Conference Operator: Your next question comes from the line of Ben Reitz with Melius Research. Please go ahead.

Ben Reitz, Analyst, Melius Research: Yes. Hi, Ben Reitz is here. Hey, thanks a lot for the question. Hey, Jensen, it's a geography related question. You did a great job explaining some of the demand underlying factors here on the strength.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

But U. S. Was up about $5,000,000,000 or so sequentially. And I think there is a concern about whether U. S.

Can pick up the slack if there's regulations towards other geographies. And I was just wondering as we go throughout the year, if this kind of surge in The U. S. Continues and it's going to 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 U. S?

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, President and Chief Executive Officer, NVIDIA: China is approximately the same percentage as Q4 and as, in 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, if you were to buy, you know, a quarter 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 every student will use AI as a tutor.

Healthcare services use AI. Financial services use AI. No fintech company will not use AI. Every fintech company will. Climate tech company use AI.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Mineral discovery now uses AI. The number of the number of at every higher education, every university, uses AI. And so, I think it is fairly safe to say that AI has gone mainstream, 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 society. And with that, 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 going to be 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 two 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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.

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

Stuart Stecker, Investor Relations, NVIDIA0: Hi, that's Mark Matapasas. Thanks for taking the question.

Stuart Stecker, Investor Relations, NVIDIA1: I had a clarification and a question. Colette, up for the clarification. Did you say that enterprise within the data center grew 2x year on year for the January? And if so, does that would that make it faster going 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 workloads and applications come up, would you expect enterprises to become a larger part of that consumption mix and decide to impact how you develop your, service your ecosystem? Thank you.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Colette Kress, Executive Vice President and Chief Financial Officer, NVIDIA: Sure. Thanks for the question regarding, our enterprise business. Yes, it grew 2x and 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 in 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, quite well.

Jensen Huang, President and Chief Executive Officer, NVIDIA: The CSPs are about half of our business. And, and, the CSPs have internal consumption and external consumption, as you say. And we're using of course, used for internal consumption. We we work very closely with all of them to optimize workloads that are internal to them, because they have a large infrastructure of NVIDIA 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, were fungible.

And so, the useful life of 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

So, So 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 share with you guys at GTC, those agentic systems are for employees to make employees more productive, to design, to market, to 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 today, there's a billion cars on the road. Someday, there'd be a billion cars on the road and every single one of those cars will be robotic cars. And they'll all be collecting data and we'll be improving them, using an AI factory.

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, the machineries, it could be, of course, it could be a tractor, it could be a lawnmower, it could be a humanoid 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.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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 type of things that are common sense to you and I, but, you know, AIs have 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 we've got some really great things going on.

And then there's the physical AI after that and then there are 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 them all, which kind of makes sense. The world of the world's GDP is representing represented by either heavy industries or industrials and companies that are providing for those.

Conference Operator: Your next question comes from the line of Aaron Rickers with Wells Fargo. Please go ahead.

Stuart Stecker, Investor Relations, NVIDIA2: Yes. 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 and Gen AI in general and 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 if it's GB300 or if it's the Ruben cycle where we start to see maybe some refresh opportunity. I'm just curious to how you look at that.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

Jensen Huang, President and Chief Executive Officer, NVIDIA: I appreciate it. First of all, people are still using Voltas (NSE:VOLT) and Pascals and Amperes. And the reason for that is because there are 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 and says, this is what happened and, I wasn't very good at it. You then take that response, the prompt, and you go and prompt an AI model to go find in your whole link of data, 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 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 they're all CUDA compatible and so everything runs on everything. But if you have infrastructure in place, then you can put the less intensive workloads onto the installed base of the past.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

All of our CPUs are very well employed.

Conference Operator: We have time for one more question. And that question comes from Atif Malik with Citi. Please go ahead.

Stuart Stecker, Investor Relations, NVIDIA0: Hi. Thank you for taking my question. I have a follow-up question on gross margins for Colette. Colette, I understand there are many moving parts, the BlackBull yields and ELink 72 and Ethernet mix. And you kind of tiptoed the earlier question if April is the bottom.

But second half would have to ramp like 200 basis points per quarter to get to the mid-70s 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

Stuart Stecker, Investor Relations, NVIDIA1: in

Stuart Stecker, Investor Relations, NVIDIA0: the back half of this year?

Colette Kress, Executive Vice President and Chief Financial Officer, NVIDIA: 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 going to probably start as soon as possible if we can. And 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 U.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

S. Government's plan is, both its timing, its where and how much. So, at this time, we are awaiting. But again, we would, of course, always follow, export controls and or tariffs in that manner.

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

Colette Kress, Executive Vice President and Chief Financial Officer, NVIDIA: you. No, no, we're going to open up to, Jensen.

Vivek Arya, Analyst, Bank of America Securities: I just want to thank you.

Jensen Huang, President and Chief Executive Officer, NVIDIA: I just want to thank you. Thank you, Colette. Demand for Blackwall 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, Broad3, DeepSeq R1 are reasoning models that apply inference time scaling. Reasoning models can consume 100 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 remains intact. Foundation models are being enhanced with multi modality and pre training is still growing.

3rd party Ad. Not an offer or recommendation by Investing.com. See disclosure here or remove ads.

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 pre training alone. Inference time scaling and reasoning, where a single query can demand 100 times more compute. We designed Blackwell for this moment, a single platform that can easily transition from pre training, post training and test time scaling.

Blackwell's FP4 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 its configurations is in full production. Each Grace Blackwell NVLink72 rack is an engineering marvel. 1,500,000 components produced across three fifty manufacturing sites by nearly 100,000 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 AI, 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 them either rented or self operated. I want to thank all of you for joining us today. Come join us at GTC in a couple of weeks. We're going to be talking about Blackwell Ultra, Rubin and other new computing, networking, reasoning AI, physical AI products and a whole bunch more. Thank you.