NVDA 2026 Q1 Earnings Call Transcript

28 May 2025

Participants

Toshiya Hari	executive
Colette Kress	executive
Jensen Huang	executive
Joseph Moore	analyst
Vivek Arya	analyst
Christopher Muse	analyst
Benjamin Reitzes	analyst
Timothy Arcuri	analyst
Jacob Wilhelm	analyst

Call transcript

Operator

Good afternoon. My name is Sarah, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's First Quarter Fiscal 2026 Financial Results Conference Call. [Operator Instructions] Toshiya Hari, you may begin your conference.

Toshiya Hari

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the first quarter of fiscal 2026. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer.

I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the second quarter of fiscal 2026. The content of today's call is NVIDIA's property. It can not be reproduced or transcribed without our prior written consent.

During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties, and our actual results may differ materially.

For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent Forms 10-K and 10-Q and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, May 28, 2025, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures.

You can find a reconciliation of these non-GAAP 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

Thank you, Toshiya. We delivered another strong quarter with revenue of $44 billion, up 69% year-over-year, exceeding our outlook in what proved to be a challenging operating environment. Data Center revenue of $39 billion grew 73% year-on-year. AR workloads have transitioned strongly to inference and AI factory build-outs are driving significant revenue.

Our customers' commitments are firm.

On April 9, the U.S. government issued new export controls on H20, our data center GPU designed specifically for the China market. We sold H20 with the approval of the previous administration.

Although our H20 has been in the market for over a year and does not have a market outside of China, the new export controls on H20 did not provide a grace period to allow us to sell through our inventory. In Q1, we recognized $4.6 billion in H20 revenue, which occurred prior to April 9, but also recognized a $4.5 billion charge as we wrote down inventory and purchase obligations tied to orders we had received prior to April 9.

We were unable to ship $2.5 billion in H20 revenue in the first quarter due to the new export controls. The $4.5 billion charge was less than what we initially anticipated as we were able to reuse certain materials.

We are still evaluating our limited options to supply data center compute products compliant with the U.S. government's revised export control rules. Losing access to the China AI accelerator market, which we believe will grow to nearly $50 billion, would have a material adverse impact on our business going forward and benefit our foreign competitors in China and worldwide.

Our Blackwell ramp, the fastest in our company's history, drove a 73% year-on-year increase in Data Center revenue. Blackwell contributed nearly 70% of Data Center compute revenue in the quarter with the transition from Hopper nearly complete. The introduction of GB200 NVL was a fundamental architectural change to enable data center-scale workloads and to achieve the lowest cost per inference token.

While these systems are complex to build, we have seen a significant improvement in manufacturing yields, and rack shipments are moving to strong rates to end customers. GB200 NVL racks are now generally available for motor builders, enterprises and sovereign customers to develop and deploy AI.

On average, major hyperscalers are each deploying nearly 1,000 NVL72 racks or 72,000 Blackwell GPUs per week and are on track to further ramp output this quarter. Microsoft, for example, has already deployed tens of thousands of Blackwell GPUs and is expected to ramp to hundreds of thousands of GB200s with OpenAI as one of its key customers. Key learnings from the GB200 ramp will allow for a smooth transition to the next phase of our product road map, Blackwell Ultra.

Sampling of GB300 systems began earlier this month at the major CSPs, and we expect production shipments to commence later this quarter. GB300 will leverage the same architecture, same physical footprint and the same electrical and mechanical specifications as GB200. The GB300 drop-in design will allow CSPs to seamlessly transition their systems and manufacturing used for GB200 while maintaining high yields. GB300 GPUs with 50% more HBM will deliver another 50% increase in dense FP4 inference compute performance compared to the B200.

We remain committed to our annual product cadence with our road map extending through 2028, tightly aligned with the multiple year planning cycles of our customers.

We are witnessing a sharp jump in inference demand. OpenAI, Microsoft and Google are seeing a step function leap in token generation. Microsoft processed over 100 trillion tokens in Q1, a fivefold increase on a year-over-year basis. This exponential growth in Azure OpenAI is representative of strong demand for Azure AI Foundry as well as other AI services across Microsoft's platform.

Inference serving startups are now serving models using B200, tripling their token generation rate and corresponding revenues for high-value reasoning models such as DeepSeek-R1 as reported by artificial analysis. NVIDIA Dynamo on Blackwell NVL72 turbocharges AI inference throughput by 30x for the new reasoning models sweeping the industry. Developer engagements increased with adoption ranging from LLM providers such as Perplexity to financial services institutions such as Capital One, who reduced agentic chatbox latency by 5x with Dynamo.

In the latest ELMO Perf inference results, we submitted our first results using GB200 NVL72, delivering up to 30x higher inference throughput compared to our [8-GPU] H200 submission on the challenging Llama 3.1 benchmark. This feat was achieved through a combination of tripling the performance for GPU as well as 9x more GPUs all connected on a single NVLink domain. And while Blackwell is still early in its life cycle, software optimizations have already improved its performance by 1.5x in the last month alone.

We expect to continue improving the performance of Blackwell through its operational life as we have done with Hopper and AMP Pro.

For example, we increased the inference performance of Hopper by 4x over 2 years. This is the benefit of NVIDIA's programmable CUDA architecture and rich ecosystem.

The pace and scale of AI factory deployments are accelerating with nearly 100 NVIDIA-powered AI factories in flight this quarter, a twofold increase year-over-year, with the average number of GPUs powering each factory also doubling in the same period. And more AI factory projects are starting across industries and geographies. NVIDIA's full stack architecture is underpinning AI factory deployments as industry leaders like AT&T, BYD, Capital One, Foxconn, MediaTek, and Telenor, are strategically vital sovereign clouds like those recently announced in Saudi Arabia, Taiwan and the UAE.

We have a line of sight to projects requiring tens of gigawatts of NVIDIA AI infrastructure in the not-too-distant future. The transition from generative to agentic AI, AI capable of receiving, reasoning, planning and acting will transform every industry, every company and country. We envision AI agents as a new digital workforce capable of handling tasks ranging from customer service to complex decision-making processes.

We introduced the Llama Nemotron family of open reasoning models designed to supercharge agentic AI platforms for enterprises. Built on the Llama architecture, these models are available as NIMs or NVIDIA inference micro services with multiple sizes to meet diverse deployment needs.

Our post training enhancements have yielded a 20% accuracy boost and a 5x increase in inference speed, leading platform companies, including Accenture, Cadence, Deloitte, and Microsoft or transforming work with our reasoning models.

NVIDIA NeMo micro services are generally available across industries are being leveraged by leading enterprises to build, optimize and scale AI applications. With NeMo, Cisco increased model accuracy by 40% and improved response time by 10x in its code assistant. NASDAQ realized a 30% improvement in accuracy and response time in its AI platform's search capabilities. And Shell's Custom LLM achieved a 30% increase in accuracy when trained with NVIDIA NeMo. NeMo's parallelism, techniques accelerated model training time by 20% when compared to other frameworks.

We also announced a partnership with Yum! Brands, the world's largest restaurant company to bring NVIDIA AI to 500 of its restaurants this year and expanding to 61,000 restaurants over time to streamline order-taking, optimize operations and enhance service across its restaurants.

For AI-powered cybersecurity leading companies like Check Point, CrowdStrike and Paladin Networks are using NVIDIA's AI security and software stack to build, optimize and secure agentic workflows, with CrowdStrike realizing 2x faster detection triage with 50% less compute cost.

Moving to networking. Sequential growth in networking resumed in Q1 with revenue up 64% quarter-over-quarter to $5 billion.

Our customers continue to leverage our platform to efficiently scale up and scale out AI factory workloads. We created the world's fastest switch, NVLink for scale up, our NVLink compute fabric in its fifth generation, offers 14x the bandwidth of PCIe Gen 5. NVLink 72 carries 130 terabytes per second of bandwidth in a single rack, equivalent to the entirety of the world's peak Internet traffic. NVLink is a new growth vector and is off to a great start with Q1 shipments exceeding $1 billion.

At Computex, we announced NVLink Fusion. Hyperscale customers can now build semi-custom CCUs and accelerators that connect directly to the NVIDIA platform with NVLink.

We are now enabling key partners, including ASIC providers such as MediaTek, Marvell, Alchip Technologies and Astera Labs as well as CPU suppliers, such as Fujitsu and Qualcomm to leverage and relink Fusion to connect our respective ecosystems.

For scale out, our enhanced Ethernet offerings delivered the highest throughput, low in its latency networking for AI.

Spectrum-X posted strong sequential and year-on-year growth and is now annualizing over $8 billion in revenue. Adoption is widespread across major CSPs and consumer Internet companies, including CoreWeave, Microsoft Azure and Oracle Cloud and xAI.

This quarter, we added Google Cloud and Meta to the growing list of Spectrum-X customers.

We introduced Spectrum-X and Quantum-X silicon photonics switches featuring the world's most advanced co-packaged optics. These platforms will enable next-level AI factory scaling to millions of DPUs through the increasingly power efficiency by 3.5x and network resiliency by 10x, while accelerating customer time to market by 1.3x.

Transitioning to a quick summary of our revenue by geography. China as a percentage of our Data Center revenue was slightly below our expectations and down sequentially due to H20 export licensing controls.

For Q2, we expect a meaningful decrease in China data center revenue.

As a reminder, while Singapore represented nearly 20% of our Q1 build revenue as many of our large customers use Singapore for centralized invoicing, our products are almost always shipped elsewhere.

Note that over 99% of H100, H200, and Blackwell data center compute revenue billed to Singapore was for orders from U.S.-based customers.

Moving to gaming and AI PCs. Gaming revenue was a record $3.8 billion, increasing 48% sequentially and 42% year-on-year. Strong adoption by gamers, creatives and AI enthusiasts have made Blackwell our fastest ramp ever. Against a backdrop of robust demand, we greatly improved our supply and availability in Q1 and expect to continue these efforts in Q2.

AI is transforming PC and creator and gamers. With a 100 million user installed base, represents the largest footprint for PC developers.

This quarter, we added to our AI PC laptop offerings, including models capable of running Microsoft's Copilot+. This past quarter, we brought Blackwell architecture to mainstream gaming with its launch of GeForce RTX 5060 and 5060 Ti starting at just $299. The RTX 5060 also debuted in laptop starting at $1,099. These systems that doubled the frame rate/latency. These GeForce RTX 50, 60 and 50-60TI desktop GPUs and laptops are now available.

In console gaming, the recently unveiled Nintendo Switch 2 leverages NVIDIA's neuro rendering and AI technologies, including next-generation custom RTX GPUs with DLSS technology to deliver a giant leap in gaming performance to millions of players worldwide. Nintendo has shipped over 150 million switch consoles to date, making it one of the most successful gaming systems in history.

Moving to Pro Visualization. Revenue of $509 million was flat sequentially and up 19% year-on-year. Tariff-related uncertainty temporarily impacted Q1 systems and demand for our AI workstations is strong, and we expect sequential revenue growth to resume in Q2. NVIDIA DGX Spark and station revolutionized personal computing. By putting the power of an AI supercomputer in a desktop form factor. DGX Spark delivers up to 1 petaflop of AI compute while DGX Station offers an incredible 20 petaflops and is powered by the GB300 Super Chip. DGX Spark will be available in calendar Q3 and DGX Station later this year.

We have deepened Omni versus integration and adoption into some of the world's leading software platforms, including Databricks, SAP and Schneider Electric, new Omniverse blueprints such as Mega for at-scale robotic fleet management are being leveraged in Kion Group, Pegatron, Accenture and other leading companies to enhance industrial operations. At Computex, we showcased Omni versus great traction with technology manufacturing leaders, including TSMC, Quanta, Foxconn, Pegatron.

Using Omniverse, TSMC saves months in work by designing fabs virtually. Foxconn accelerates thermal simulations by 150x, and Pegatron reduced assembly line defect rates by 67%.

Lastly with our automotive group. Revenue was $567 million, down 1% sequentially but up 72% year-on-year. Year-on-year growth was driven by the ramp of self-driving across a number of customers and robust end demand for NAVs.

We are partnering with GM to build the next-gen vehicles, factories and robots using NVIDIA AI, simulation and accelerated computing. And we are now in production with our full stack solution for Mercedes-Benz starting with the new CLA hitting roads in the next few months. We announced Isaac Group and one, the world's first open fully customizable foundation model for humanoid robots enabling generalized reasoning and skill development.

We also launched new open NVIDIA Cosmo World Foundation models. Leading companies include [OneX], Agility Robots, Robotics, Figure AI, Uber and Wobi. We've begun integrating Kosmos into their operations for synthetic data generation, while Agility Robotics, Boston Dynamics, and Robotics are harnessing Isaac's simulation to advance their humanoid efforts.

GE Healthcare is using the new NVIDIA Isaac platform for health care simulation built on NVIDIA Omniverse and using NVIDIA Cosmos. The platform speed, development of robotic imaging and surgery systems. The era of robotics is here, billions of robots, hundreds of millions of autonomous vehicles and hundreds of thousands of robotic factories and warehouses will be developed.

All right.

Moving to the rest of the P&L. GAAP gross margins and non-GAAP gross margins were 60.5% and 61%, respectively.

Excluding the $4.5 billion charge, Q1 non-GAAP gross margins would have been 71.3%, slightly above our outlook at the beginning of the quarter. Sequentially, GAAP operating expenses were up 7% and non-GAAP operating expenses were up 6%, reflecting higher compensation and employee growth.

Our investments include expanding our infrastructure capabilities and AI solutions, and we plan to grow these investments throughout the fiscal year. In Q1, we returned a record $14.3 billion to shareholders in the form of share repurchases and cash dividends.

Our capital return program continues to be a key element of our capital allocation strategy.

Let me turn to the outlook for the second quarter. Total revenue is expected to be $45 billion, plus or minus 2%.

We expect modest sequential growth across all of our platforms. In Data Center, we anticipate the continued ramp of Blackwell to be partially offset by a decline in China revenue. Note, our outlook reflects a loss in H20 revenue of approximately $8 billion for the second quarter. GAAP and non-GAAP gross margins are expected to be 71.8% and 72%, respectively, plus or minus 50 basis points.

We expect or Blackwell profitability to drive modest sequential improvement in gross margins.

We are continuing to work towards achieving gross margins in the mid-70s range late this year.

GAAP and non-GAAP operating expenses are expected to be approximately $5.7 billion and $4 billion, respectively, and we continue to expect full year fiscal year '26 operating expense growth to be in the mid-30% range. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $450 million, excluding gains and losses from nonmarketable and publicly held equity securities. GAAP and non-GAAP tax rates are expected to be 16.5%, 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 financially information AI agent.

Let me highlight upcoming events for the financial community.

We will be at the BofA Global Technology Conference in San Francisco on June 4. The Rosenblatt Virtual AI Summit and NASDAQ Investor Conference in London on June 10, and GTC Paris at VivaTech on June 11 in Paris. We look forward to seeing you at these events.

Our earnings call to discuss the results of our second quarter of fiscal 2026 is scheduled for August 27. Well, now let me turn it over to Jensen to make some remarks.

Jensen Huang
Thanks, Colette. We've had a busy and productive year.

Let me share my perspective on some topics we're frequently asked. On export control. China is one of the world's largest AI markets and a springboard to global success. With half of the world's AI researchers based there, the platform that wins China is positioned to lead globally. Today, however, the $50 billion China market is effectively closed to U.S. industry. The H20 export ban ended our hopper data center business in China. We cannot reduce hopper further to comply.

As a result, we are taking a multibillion-dollar write-off on inventory that cannot be sold or repurposed.

We are exploring limited ways to compete, but Hopper is no longer an option. China's AI moves on with or without U.S. chips. It has to compute to train and deploy advanced models. The question is not whether China will have AI, it already does. The question is whether one of the world's largest AR markets will run on American platforms.

Shielding Chinese chipmakers from U.S. competition only strengthens them abroad and weakens America's position. Export restrictions have spurred China's innovation and scale.

The AI race is not just about chips. It's about which stack the world runs on.

As that stack grows to include 6G and quantum, U.S. global infrastructure leadership is at stake. The U.S. has based its policy on the assumption that China cannot make AI chips. That assumption was always questionable and now it's clearly wrong. China has enormous manufacturing capability. In the end, the platform that wins the AI developers win AI -- wins AI. Export controls should strengthen U.S. platforms, not drive half of the world's AI talent to rivals.

On DeepSeek, DeepSeek and Q1 from China are among the most -- among the best open source models. Released freely, they've gained traction across the U.S., Europe and beyond. DeepSeek R1, like ChatGPT, introduced reasoning AI that produces better answers, the longer it thinks. Reasoning AI enables step-by-step problem solving, planning and tool use, turning models into intelligent agents.

Reasoning is compute-intensive, requires hundreds to thousands more thousands of times more tokens per task than previous one-shot inference. Reasoning models are driving a step-function surge in inference demand. AI scaling laws remain firmly intact, not only for training, but now Inference 2 requires massive scale compute. DeepSeek also underscores the strategic value of open source AI. When popular models are trained and optimized on U.S. platforms, it drives usage, feedback and continuous improvement, reinforcing American leadership across the stack.

U.S. platforms must remain the preferred platform for open source AI. That means supporting collaboration with top developers globally, including in China. America wins when models like DeepSeek and Q1 runs best on American infrastructure.

Regarding onshore manufacturing, President Trump has outlined a bold vision to reshore advanced manufacturing, create jobs and strengthen national security. Future plants will be highly computerized in robotics. We share this vision. TSMC is building 6 fabs and 2 advanced packaging plants in Arizona to make chips for NVIDIA. Process qualification is underway with volume production expected by year-end. SPIL and Amcor are also investing in Arizona, constructing packaging, assembly and test facilities.

In Houston, we're partnering with Foxconn to construct a 1 million square foot factory to build AI supercomputers. Wistron is building a similar plant in Fort Worth, Texas. To encourage and support these investments, we've made substantial long-term purchase commitments a deep investment in America's AI manufacturing future.

Our goal from chip to supercomputer built in America within a year. Each GB200 NVLink72 racks contains 1.2 million components and weighs nearly 2 tons. No 1 has produced supercomputers on this scale.

Our partners are doing an extraordinary job.

On AI diffusion rule, President Trump rescinded the AI diffusion rule, calling it counterproductive, and proposed a new policy to promote U.S. AI tech with trusted partners. On his Middle East tour, he announced historic investments. I was honored to join him in announcing a 500-megawatt AI infrastructure project in Saudi Arabia and a 5-gigawatt AI campus in the UAE. President Trump wants U.S. tech to lead. The deals he announced are wins for America, creating jobs, advancing infrastructure, generating tax revenue and reducing the U.S. trade deficit.

The U.S. will always be NVIDIA's largest market and home to the largest installed base of our infrastructure. Every nation now sees AI as core to the next industrial revolution, a new industry that produces intelligence and essential infrastructure for every economy. Countries are racing to build national AI platforms to elevate their digital capabilities. At Computex, we announced Taiwan's first AI factory in partnership with Foxconn and the Taiwan government.

Last week, I was in Sweden to launch its first national AI infrastructure. Japan, Korea, India, Canada, France, the U.K., Germany, Italy, Spain, and more are now building national AI factories to empower startups, industries and societies.

Sovereign AI is a new growth engine for NVIDIA. Toshiya, back to you.

Toshiya Hari
Operator, we will now open the call for questions. Would you please poll for questions?

Operator
[Operator Instructions] Your first question comes from the line of Joe Moore with Morgan Stanley.

Joseph Moore
You guys have talked about this scaling up of inference around reasoning models for at least a year now. And we've really seen that come to fruition as you talked about. We've heard it from your customers. Can you give us a sense for how much of that demand you're able to serve and give us a sense for maybe how big the inference business is for you guys? And do we need full on NDL72 rack scale solutions for reasoning inference going forward?

Jensen Huang
Well, we would like to serve all of it, and I think we're on track to serve most of it. Grace Blackwell NVLink72 is the ideal engine today, the ideal computer thinking machine, if you will, for reasoning AI. There's a couple of reasons for that.

The first reason is that the token generation amount, the number of tokens reasoning goes through, is 100x, 1,000x more than a one-shot chatbot.

It's essentially thinking to itself, breaking down a problem step-by-step. It might be planning multiple paths to an answer. It could be using tools, reading PDFs, reading web pages, watching videos and then producing a result, an answer. The longer it thinks, the better the answer, the smarter the answer is.

And so what we would like to do, and the reason why Grace Blackwell was designed to give such a giant step-up in inference performance, is so that you could do all this and still get a response as quickly as possible.

Compared to Hopper, Grace Blackwell is some 40x higher speed and throughput compared.

And so this is going to be a huge, huge benefit in driving down the cost while improving the quality of response with excellent quality of service at the same time.

So that's the fundamental reason. That was the core driving reason for Grace Blackwell NVLink 72.

Of course, in order to do that, we had to reinvent, literally redesign, the entire -- a way that these supercomputers are built. But now we're in full production. It's going to be exciting. It's going to be incredibly exciting.

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

Vivek Arya
Just a clarification for Colette first.

So on the China impact, I think previously, it was mentioned at about $15 billion, so you had the $8 billion in Q2.

So is there still some left as a headwind for the remaining quarters just Colette, how to model that?

And then a question, Jensen, for you. Back at GTC, you had outlined a path towards almost $1 trillion of AI spending over the next few years. Where are we in that build-out? And do you think it's going to be uniform that you will see every spender, whether it's ESP, sovereigns, enterprises or build-out, should we expect some periods of digestion in between? Just what are your customer discussions telling you about how to model growth for next year?

Colette Kress
Yes, Vivek. Thanks so much for the question regarding H20. Yes, we recognized $4.6 billion H20 in Q1. We were unable to ship $2.5 billion so the total for Q1 should have been $7 billion. When we look at our Q2, our Q2 is going to be meaningfully down in terms of China data center revenue. And we had highlighted in terms of the amount of orders that we had planned for H20 in Q2, and that was $8 billion.

Now going forward, we did have other orders going forward that we will not be able to fulfill. That is what was incorporated, therefore, in the amount that we wrote down of the $4.5 billion. That write-down was about inventory and purchase commitments, and our purchase commitments were about what we expected regarding the orders that we had received.

Going forward, though, it's a bigger issue regarding the amount of the market that we will not be able to serve. We assess that TAM to be close to about $50 billion in the future as we don't have a product to enable for China.

Jensen Huang
In fact, the -- probably the best way to think through it is that AI is several things.

Of course, we know that AI is this incredible technology that's going to transform every industry from, of course, the way we do software to health care and financial services to retail to, I guess, every industry, transportation, manufacturing. And we're at the beginning of that.

But maybe another way to think about that is where do we need intelligence, where do we need digital intelligence? And it's in every country, it's in every industry. And we know because of that, we recognize that AI is also an infrastructure. It's a way of developing a technology -- delivering a technology that requires factories and these factories produce tokens. And they, as I mentioned, are important to every single industry and every single country.

And so on that basis, we're really at the very beginning of it because the adoption of this technology is really kind of in its early, early stages.

Now we've reached an extraordinary milestone with AIs that are reasoning or thinking, what people call inference time scaling.

Of course, it created a whole new -- we've entered an era where inference is going to be a significant part of the compute workload. But anyhow, it's going to be a new infrastructure, and we're building it out in the cloud. The United States is really the early starter and available in U.S. clouds. And this is our largest market, our largest installed base and we continue to see that happening.

But beyond that, we're going to have to -- we're going to see AI go into enterprise, which is on-prem because so much of the data is still on-prem. Access control is really important. It's really hard to move all of every company's data into the cloud.

And so we're going to move AI into the enterprise. And you saw that we announced a couple of really exciting new products, our RTX Pro Enterprise AI server that runs everything enterprise and AI, our DGX Spark and DGX Station, which is designed for developers who want to work on-prem.

And so enterprise AI is just taking off.

Telcos. Today, a lot of the telco infrastructure will be, in the future, software defined and built on AI, and so 6G is going to be built on AI and that infrastructure needs to be built out. And I said, it's very, very early stages. And then, of course, every factory today that makes things will have an AI factory that sits with it. And the AI factory is going to be -- drive creating AI and operating AI for the factory itself but also to power the products and the things that are made by the factory.

So it's very clear that every company will have AI factories.

And very soon, there'll be robotics companies, robot companies and those companies will be also building AIs to drive the robots.

And so we're at the beginning of all of this build-out.

Operator
The next question comes from C.J. Muse with Cantor Fitzgerald.

Christopher Muse
There have been many large GPU cluster investment announcements in the last month, and you alluded to a few of them with Saudi Arabia, the UAE. And then also we heard from Oracle and xAI, just to name a few.

So my question, are there other that have yet to be announced of the same kind of scale and magnitude? And perhaps more importantly, how are these orders impacting your lead times for Blackwell and your current visibility sitting here today almost halfway through 2025?

Jensen Huang
Well, we have more orders today than we did at the last time I spoke about orders at GTC.

However, we're also increasing our supply chain and building out our supply chain. They're doing a fantastic job. We're building it here onshore in the United States. But we're going to keep our supply chain quite busy for several -- many more years coming.

And with respect to further announcements, I'm going to be on the road next week through Europe. And it's -- just about every country needs to build out AI infrastructure and their [umpteenth] AI factories being planned. We're -- I think in the remarks, Colette mentioned there's some 100 AI factories being built. There's a whole bunch that haven't been announced.

And I think the important concept here which makes it easier to understand is that like other technologies that impact literally every single industry, of course, electricity was one and it became infrastructure.

Of course, the information infrastructure, which we now know as the Internet affects every single industry, every country, every society. Intelligence is surely one of those things. I don't know any company, industry, country who thinks that intelligence is optional. It's essential infrastructure.

And so we've now digitalized intelligence.

And so I think we're clearly in the beginning of the build-out of this infrastructure. And every country will have it, I'm certain of that. Every industry will use it, that I'm certain of. And what's unique about this infrastructure is that it needs factories. It's a little bit like the energy infrastructure, electricity. It needs factories. We need factories to produce this intelligence, and the intelligence is getting more sophisticated.

We were talking about earlier that we had a huge breakthrough in the last couple of years with reasoning AI. And now there are agents that reason and there are super-agents that use a whole bunch of tools and then there's clusters of super agents where agents are working with agents, solving problems.

And so you could just imagine, compared to one-shot chatbots and the agents that are now using AI built on these large language models, how much more compute-intensive they really need to be and are.

So I think we're in the beginning of the build-out, and there should be many, many more announcements in the future.

Operator
Your next question comes from Ben Reitzes with Melius.

Benjamin Reitzes
I wanted to ask, first to Colette, just a little clarification around the guidance and maybe putting it in a different way. The $8 billion for H20 just seems like it's roughly $3 billion more than most people thought with regard to what you'd be foregoing in the second quarter.

So that would mean that with regard to your guidance, the rest of the business in order to hit [45] is doing $2 billion to $3 billion or so better.

So I was wondering if that math made sense to you.

And then in terms of the guidance, that would imply the non-China business is doing a bit better than the Street expected.

So wondering what the primary driver was there in your view. And then this second part of my question, Jensen, I know you guide 1 quarter at a time, but with regard to the AI diffusion rule being lifted and this momentum with sovereign, there's been times in your history where you guys have said on calls like this, where you have more conviction and sequential growth throughout the year, et cetera. And given the unleashing of demand with AI diffusion being revoked and the supply chain increasing, does the environment give you more conviction and sequential growth as we go throughout the year? So first 1 for Colette and then next 1 for Jensen.

Colette Kress
Thanks, Ben, for the question. When we look at our Q2 guidance and our commentary that we provided, that had the export controls not occurred, we would have had orders of about $8 billion for H20, that's correct. That was a possibility for what we would have had in our outlook for this quarter in Q2.

So what we also have talked about here is the growth that we've seen in Blackwell, Blackwell across many of our customers as well as the growth that we continue to have in terms of supply that we need for our customers.

So putting those together, that's where we came through with the guidance that we provided. I'm going to turn the rest over to Jensen to see how he wants to...

Jensen Huang

Yes. Thanks. Thanks, Ben. I would say compared to the beginning of the year, compared to GTC time frame, there are 4 positive surprises.

The first positive surprise is the step function demand increase of reasoning AI, I think it is fairly clear now that AI is going through an exponential growth, and reasoning AI really busted through. Concerns about hallucination or its ability to really solve problems, and I think a lot of people are crossing that barrier and realizing how incredibly effective agentic AI is and reasoning AI is.

So number 1 is inference reasoning and the exponential growth there, demand growth.

The second one, you mentioned AI diffusion. It's really terrific to see that the AI diffusion rule was rescinded. President Trump wants America to win, and he also realizes that we're not the only country in the race. And he wants the United States to win and recognizes that we have to get the American stack out to the world and have the world build on top of American stacks instead of alternatives.

And so AI diffusion happened, the rescinding of it happened at almost precisely the time that countries around the world are awakening to the importance of AI as an infrastructure, not just as a technology of great curiosity and great importance, but infrastructure for their industries and start-ups and society.

Just as they had to build out infrastructure for electricity and Internet, you got to build out an infrastructure for AI.

I think that, that's an awakening, and that creates a lot of opportunity.

The third is enterprise AI. Agents work and agents are doing -- these agents are really quite successful, much more than generative AI. Agentic AI is game-changing. Agents can understand ambiguous and rather implicit instructions and able to problem solve and use tools and have memory and so on.

And so I think this is -- enterprise AI is ready to take off.

And it's taken us a few years to build a computing system that is able to integrate and run enterprise AI stacks, run enterprise IT stacks but add AI to it. And this is the RTX Pro Enterprise server that we announced at Computex just last week. And just about every major IT company has joined us, super excited about that.

And so computing is 1 stack, 1 part of it. But remember, enterprise IT is really 3 pillars: it's compute, storage, and networking. And we've now put all 3 of them together for finally, and we're going to market with that.

And then lastly, industrial AI. Remember, one of the implications of the world reordering, if you will, is a region's onshoring manufacturing and building plants everywhere.

In addition to AI factories, of course, there are new electronics manufacturing, chip manufacturing being built around the world. And all of these new plants in these new factories are creating exactly the right time when Omniverse and AI and all the work that we're doing with robotics is emerging.

And so this fourth pillar is quite important. Every factory will have an AI factory associated with it. And in order to create these physical AI systems, you really have to train a vast amount of data.

So back to more data, more training, more AIs to be created, more computers.

And so these 4 drivers are really kicking into turbocharge.

Operator
Your next question comes from Timothy Arcuri with UBS.

Timothy Arcuri
Jensen, I wanted to ask about China. It sounds like the July guidance assumes there's no SKU replacement for the age 20. But if the President wants the U.S. to win, it seems like you're going to have to be allowed to ship something into China.

So I guess I had 2 points on that.

First of all, have you been approved to ship a new modified version into China? And you're currently building it but you just can't ship it in fiscal Q2?

And then you were sort of run rating $7 billion to $8 billion a quarter into China. Can we get back to those sorts of quarterly run rates once you get something that you're allowed to ship back into China? I think we're all trying to figure out how much to add back to our models and when.

So whatever you can say there would be great.

Jensen Huang
The President has a plan. He has a vision and I trust him. With respect to our export controls, it's a set of limits. And the new set of limits pretty much make it impossible for us to reduce hopper any further for any productive use.

And so the new limits, it's kind of the end of the road for Hopper.

We have some -- we have limited options.

And so we just -- the key is to understand the limits. The key is to understand the limits and see if we can come up with interesting products that could continue to serve the Chinese market.

We don't have anything at the moment, but we're considering it. We're thinking about it. Obviously, the limits are quite stringent at the moment. And we have nothing to announce today. And when the time comes, we'll engage the administration and discuss that.

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

Jacob Wilhelm
This is Jake on for Aaron. I was wondering if you could give some additional color around the strength you saw within the Networking business, particularly around the adoption of your Ethernet solutions at CSPs as well as any change you're seeing in network attach rates.

Jensen Huang
Yes, thank you for that. We now have 3 networking platforms, maybe 4.

The first 1 is the scale-up platform to turn a computer into a much larger computer. Scaling up is incredibly hard to do. Scaling out is easier to do but scaling up is hard to do. And that platform is called NVLink. NVLink is -- comes with it chips and switches and NVLink spines and it's really complicated. But anyway, that's our new platform, scale-up platform.

In addition to InfiniBand, we also have Spectrum-X. We've been fairly consistent that Ethernet was designed for a lot of traffic that are independent. But in the case of AI, you have a lot of computers working together. And the traffic of AI is insanely bursty. Latency matters a lot because the AI is thinking and it wants to get work on as quickly as possible, and you got a whole bunch of nodes working together.

And so we enhanced Ethernet, added capabilities like extremely low latency, congestion control, adaptive routing, the type of technologies that were available only in InfiniBand to Ethernet.

And as a result, we improved the utilization of Ethernet in these clusters. These clusters are gigantic, from as low as 50% to as high as 85%, 90%.

And so the difference is if you had a cluster that's $10 billion and you improve its effectiveness by 40%, that's worth $4 billion. It's incredible.

And so Spectrum-X has been really, quite frankly, a home run. And this last quarter, as we said in the prepared remarks, we added 2 very significant CSPs to the Spectrum-X adoption.

And then the last 1 is BlueField, which is our control plane.

And so in those 4 -- those -- the control plane network, which is used for storage. It's used for security and for many of these clusters that want to achieve isolation among its users, multi-tenant clusters and still be able to use and have extremely high-performance bare metal performance, BlueField is ideal for that and is used in a lot of these cases.

And so we have these 4 networking platforms that are all growing and we're doing really well. I'm very proud of the team.

Operator
That is all the time we have for questions. Jensen, I will turn the call back to you.

Jensen Huang
Thank you. This is the start of a powerful new wave of growth. Grace Blackwell is in full production. We're off to the races. We now have multiple significant growth engines. Inference, once the light workload is surging with revenue-generating AI services. AI is growing faster and will be larger than any platform shifts before, including the Internet, mobile and cloud.

Blackwell is built to power the full AI life cycle from training frontier models to running complex inference and reasoning agents at scale. Training demand continues to rise with breakthroughs in post training and like reinforcement learning and synthetic data generation. But inference is exploding. Reasoning AI agents require orders of magnitude more compute.

But foundations of our next growth platforms are in place and ready to scale.

Sovereign AI, nations are investing in AI infrastructure. They for electricity and Internet. Enterprise AI, AI must be deployable on prem and integrated with existing IT.

Our RTX Pro, DGX Park and DGX Station enterprise AI systems are ready to modernize the $500 billion IT infrastructure on-prem or in the cloud. Every major IT provider is partnering with us.

Industrial AI from training to digital twin simulation to deployment, NVIDIA Omniverse and Isaac are powering next-generation factories and humanoid robotic systems worldwide. The age of AI is here from AI infrastructures, inference at scale, sovereign AI, enterprise AI, and industrial AI, NVIDIA is ready.

Join us at GTC Paris, our keynote at VivaTech on June 11, talking about quantum GPU computing, robotic factories and robots and celebrate our partnerships building AI factories across the region. The NVIDIA band will tour France, the U.K., Germany, and Belgium. Thank you for joining us at the earnings call today. See you in Paris.

Operator
This concludes today's conference call.

You may now disconnect.