# NVDA Earnings Call – FY2025 Q1



Date: May 22, 2024

**Operator**

Good afternoon. My name is Regina and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's First Quarter Earnings Call. All lines have been placed on mute to prevent any background noise. After the speaker's remarks, there will be a question-and-answer session. [Operator Instructions] Thank you. Simona Jankowski, you may begin your conference.

**Simona Jankowski**

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the first quarter of fiscal 2025. With me today from NVIDIA are Jen-Hsun 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 2025. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent. During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent Forms 10-K and 10-Q and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, May 22, 2024, 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. Let me highlight some upcoming events. On Sunday, June 2nd, ahead of the Computex Technology Trade Show in Taiwan, Jensen will deliver a keynote which will be held in-person in Taipei as well as streamed live. And on June 5th, we will present at the Bank of America Technology Conference in San Francisco. With that let me turn the call over to Colette.

**Colette Kress**

Thanks, Simona. Q1 was another record quarter. Revenue of $26 billion was up 18% sequentially and up 262% year-on-year and well above our outlook of $24 billion. Starting with Data Center. Data Center revenue of $22.6 billion was a record, up 23% sequentially and up 427% year-on-year, driven by continued strong demand for the NVIDIA Hopper GPU computing platform. Compute revenue grew more than 5x and networking revenue more than 3x from last year. Strong sequential data center growth was driven by all customer types, led by enterprise and consumer internet companies. Large cloud providers continue to drive strong growth as they deploy and ramp NVIDIA AI infrastructure at scale and represented the mid-40s as a percentage of our Data Center revenue. Training and inferencing AI on NVIDIA CUDA is driving meaningful acceleration in cloud rental revenue growth, delivering an immediate and strong return on cloud provider's investment. For every $1 spent on NVIDIA AI infrastructure, cloud providers have an opportunity to earn $5 in GPU instant hosting revenue over four years. NVIDIA's rich software stack and ecosystem and tight integration with cloud providers makes it easy for end customers up and running on NVIDIA GPU instances in the public cloud. For cloud rental customers, NVIDIA GPUs offer the best time to train models, the lowest cost to train models and the lowest cost to inference large language models. For public cloud providers, NVIDIA brings customers to their cloud, driving revenue growth and returns on their infrastructure investments. Leading LLM companies such as OpenAI, Adept, Anthropic, Character.AI, Cohere, Databricks, DeepMind, Meta, Mistral, xAI, and many others are building on NVIDIA AI in the cloud. Enterprises drove strong sequential growth in Data Center this quarter. We supported Tesla's expansion of their training AI cluster to 35,000 H100 GPUs. Their use of NVIDIA AI infrastructure paved the way for the breakthrough performance of FSD Version 12, their latest autonomous driving software based on Vision. Video Transformers, while consuming significantly more computing, are enabling dramatically better autonomous driving capabilities and propelling significant growth for NVIDIA AI infrastructure across the automotive industry. We expect automotive to be our largest enterprise vertical within Data Center this year, driving a multibillion revenue opportunity across on-prem and cloud consumption. Consumer Internet companies are also a strong growth vertical. A big highlight this quarter was Meta's announcement of Llama 3, their latest large language model, which was trained on a cluster of 24,000 H100 GPUs. Llama 3 powers Meta AI, a new AI assistant available on Facebook, Instagram, WhatsApp and Messenger. Llama 3 is openly available and has kickstarted a wave of AI development across industries. As generative AI makes its way into more consumer Internet applications, we expect to see continued growth opportunities as inference scales both with model complexity as well as with the number of users and number of queries per user, driving much more demand for AI compute. In our trailing four quarters, we estimate that inference drove about 40% of our Data Center revenue. Both training and inference are growing significantly. Large clusters like the ones built by Meta and Tesla are examples of the essential infrastructure for AI production, what we refer to as AI factories. These next-generation data centers host advanced full-stack accelerated computing platforms where the data comes in and intelligence comes out. In Q1, we worked with over 100 customers building AI factories ranging in size from hundreds to tens of thousands of GPUs, with some reaching 100,000 GPUs. From a geographic perspective, Data Center revenue continues to diversify as countries around the world invest in Sovereign AI. Sovereign AI refers to a nation's capabilities to produce artificial intelligence using its own infrastructure, data, workforce and business networks. Nations are building up domestic computing capacity through various models. Some are procuring and operating Sovereign AI clouds in collaboration with state-owned telecommunication providers or utilities. Others are sponsoring local cloud partners to provide a shared AI computing platform for public and private sector use. For example, Japan plans to invest more than $740 million in key digital infrastructure providers, including KDDI, Sakura Internet, and SoftBank to build out the nation's Sovereign AI infrastructure. France-based, Scaleway, a subsidiary of the Iliad Group, is building Europe's most powerful cloud native AI supercomputer. In Italy, Swisscom Group will build the nation's first and most powerful NVIDIA DGX-powered supercomputer to develop the first LLM natively trained in the Italian language. And in Singapore, the National Supercomputer Center is getting upgraded with NVIDIA Hopper GPUs, while Singtel is building NVIDIA's accelerated AI factories across Southeast Asia. NVIDIA's ability to offer end-to-end compute to networking technologies, full-stack software, AI expertise, and rich ecosystem of partners and customers allows Sovereign AI and regional cloud providers to jumpstart their country's AI ambitions. From nothing the previous year, we believe Sovereign AI revenue can approach the high single-digit billions this year. The importance of AI has caught the attention of every nation. We ramped new products designed specifically for China that don't require an export control license. Our Data Center revenue in China is down significantly from the level prior to the imposition of the new export control restrictions in October. We expect the market in China to remain very competitive going forward. From a product perspective, the vast majority of compute revenue was driven by our Hopper GPU architecture. Demand for Hopper during the quarter continues to increase. Thanks to CUDA algorithm innovations, we've been able to accelerate LLM inference on H100 by up to 3x, which can translate to a 3x cost reduction for serving popular models like Llama 3. We started sampling the H200 in Q1 and are currently in production with shipments on track for Q2. The first H200 system was delivered by Jensen to Sam Altman and the team at OpenAI and powered their amazing GPT-4o demos last week. H200 nearly doubles the inference performance of H100, delivering significant value for production deployments. For example, using Llama 3 with 700 billion parameters, a single NVIDIA HGX H200 server can deliver

24,000 tokens per second, supporting more than 2,400 users at the same time. That means for every $1 spent on NVIDIA HGX H200 servers at current prices per token, an API provider serving Llama 3 tokens can generate $7 in revenue over four years. With ongoing software optimizations, we continue to improve the performance of NVIDIA AI infrastructure for serving AI models. While supply for H100 prove, we are still constrained on H200. At the same time, Blackwell is in full production. We are working to bring up our system and cloud partners for global availability later this year. Demand for H200 and Blackwell is well ahead of supply and we expect demand may exceed supply well into next year. Grace Hopper Superchip is shipping in volume. Last week at the International Supercomputing Conference, we announced that nine new supercomputers worldwide are using Grace Hopper for a combined 200 exaflops of energy-efficient AI processing power delivered this year. These include the Alps Supercomputer at the Swiss National Supercomputing Center, the fastest AI supercomputer in Europe. Isambard-AI at the University of Bristol in the UK and JUPITER in the Julich Supercomputing Center in Germany. We are seeing an 80% attach rate of Grace Hopper in supercomputing due to its high energy efficiency and performance. We are also proud to see supercomputers powered with Grace Hopper take the number one, the number two, and the number three spots of the most energy-efficient supercomputers in the world. Strong networking year-on-year growth was driven by InfiniBand. We experienced a modest sequential decline, which was largely due to the timing of supply, with demand well ahead of what we were able to ship. We expect networking to return to sequential growth in Q2. In the first quarter, we started shipping our new Spectrum-X Ethernet networking solution optimized for AI from the ground up. It includes our Spectrum-4 switch, BlueField-3 DPU, and new software technologies to overcome the challenges of AI on Ethernet to deliver 1.6x higher networking performance for AI processing compared with traditional Ethernet. Spectrum-X is ramping in volume with multiple customers, including a massive 100,000 GPU cluster. Spectrum-X opens a brand-new market to NVIDIA networking and enables Ethernet only data centers to accommodate large-scale AI. We expect Spectrum-X to jump to a multibillion-dollar product line within a year. At GTC in March, we launched our next-generation AI factory platform, Blackwell. The Blackwell GPU architecture delivers up to 4x faster training and 30x faster inference than the H100 and enables real-time generative AI on trillion-parameter large language models. Blackwell is a giant leap with up to 25x lower TCO and energy consumption than Hopper. The Blackwell platform includes the fifth-generation NVLink with a multi-GPU spine and new InfiniBand and Ethernet switches, the X800 series designed for a trillion parameter scale AI. Blackwell is designed to support data centers universally, from hyperscale to enterprise, training to inference, x86 to Grace CPUs, Ethernet to InfiniBand networking, and air cooling to liquid cooling. Blackwell will be available in over 100 OEM and ODM systems at launch, more than double the number of Hopper's launch and representing every major computer maker in the world. This will support fast and broad adoption across the customer types, workloads and data center environments in the first year shipments. Blackwell time-to-market customers include Amazon, Google, Meta, Microsoft, OpenAI, Oracle, Tesla, and xAI. We announced a new software product with the introduction of NVIDIA Inference Microservices or NIM. NIM provides secure and performance-optimized containers powered by NVIDIA CUDA acceleration in network computing and inference software, including Triton Inference Server and TensorRT LLM with industry-standard APIs for a broad range of use cases, including large language models for text, speech, imaging, vision, robotics, genomics and digital biology. They enable developers to quickly build and deploy generative AI applications using leading models from NVIDIA, AI21, Adept, Cohere, Getty Images, and Shutterstock and open models from Google, Hugging Face, Meta, Microsoft, Mistral AI, Snowflake and Stability AI. NIMs will be offered as part of our NVIDIA AI enterprise software platform for production deployment in the cloud or on-prem. Moving to gaming and AI PCs. Gaming revenue of $2.65 billion was down 8% sequentially and up 18% year-on-year, consistent with our outlook for a seasonal decline. The GeForce RTX Super GPUs market reception is strong and end demand and channel inventory remained healthy across the product range. From the very start of our AI journey, we equipped GeForce RTX GPUs with CUDA Tensor Cores. Now with over 100 million of an installed base, GeForce RTX GPUs are perfect for gamers, creators, AI enthusiasts and offer unmatched performance for running generative AI applications on PCs. NVIDIA has full technology stack for deploying and running fast and efficient generative AI inference on GeForce RTX PCs. TensorRT LLM now accelerates Microsoft's Phi-3-Mini model and Google's Gemma 2B and 7B models as well as popular AI frameworks, including LangChain and LlamaIndex. Yesterday, NVIDIA and Microsoft announced AI performance optimizations for Windows to help run LLMs up to 3x faster on NVIDIA GeForce RTX AI PCs. And top game developers, including NetEase Games, Tencent and Ubisoft are embracing NVIDIA Avatar Character Engine to create lifelike avatars to transform interactions between gamers and nonplayable characters. Moving to ProVis. Revenue of $427 million was down 8% sequentially and up 45% year-on-year. We believe generative AI and Omniverse industrial digitalization will drive the next wave of professional visualization growth. At GTC, we announced new Omniverse Cloud APIs to enable developers to integrate Omniverse industrial digital twin and simulation technologies into their applications. Some of the world's largest industrial software makers are adopting these APIs, including ANSYS, Cadence, 3DEXCITE at Dassault Systemes, Brand and Siemens. And developers can use them to stream industrial digital twins with spatial computing devices such as Apple Vision Pro. Omniverse Cloud APIs will be available on Microsoft Azure later this year. Companies are using Omniverse to digitalize their workflows. Omniverse power digital twins enable Wistron, one of our manufacturing partners to reduce end-to-end production cycle times by 50% and defect rates by 40%. And BYD, the world's largest electric vehicle maker, is adopting Omniverse for virtual factory planning and retail configurations. Moving to automotive. Revenue was $329 million, up 17% sequentially and up 11% year-on-year. Sequential growth was driven by the ramp of AI cockpit solutions with global OEM customers and strength in our self-driving platforms. Year-on-year growth was driven primarily by self-driving. We supported Xiaomi in the successful launch of its first electric vehicle, the SU7 sedan built on the NVIDIA DRIVE Orin, our AI car computer for software-defined AV fleets. We also announced a number of new design wins on NVIDIA DRIVE Thor, the successor to Orin, powered by the new NVIDIA Blackwell architecture with several leading EV makers, including BYD, XPeng, GAC's Aion Hyper and Neuro. DRIVE Thor is slated for production vehicles starting next year. Okay. Moving to the rest of the P&L. GAAP gross margin expanded sequentially to 78.4% and non-GAAP gross margins to 78.9% on lower inventory targets. As noted last quarter, both Q4 and Q1 benefited from favorable component costs. Sequentially, GAAP operating expenses were up 10% and non-GAAP operating expenses were up 13%, primarily reflecting higher compensation-related costs and increased compute and infrastructure investments. In Q1, we returned $7.8 billion to shareholders in the form of share repurchases and cash dividends. Today, we announced a 10-for-1 split of our shares with June 10th as the first day of trading on a split-adjusted basis. We are also increasing our dividend by 150%. Let me turn to the outlook for the second quarter. Total revenue is expected to be $28 billion, plus or minus 2%. We expect sequential growth in all market platforms. GAAP and non-GAAP gross margins are expected to be 74.8% and 75.5%, respectively, plus or minus 50 basis points, consistent with our discussion last quarter. For the full year, we expect gross margins to be in the mid-70s percent range. GAAP and non-GAAP operating expenses are expected to be approximately $4 billion and $2.8 billion, respectively. Full year OpEx is expected to grow in the low 40% range. GAAP and non-GAAP other income and expenses are expected to be an income of approximately, excuse me, approximately $300 million, excluding gains and losses from nonaffiliated investments. 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. I would like to now turn it over to Jensen as he would like to make a few comments.

**Jensen Huang**

Thanks, Colette. The industry is going through a major change. Before we start Q&A, let me give you some perspective on the importance of the transformation. The next industrial revolution has begun. Companies and countries are partnering with NVIDIA to shift the trillion-dollar installed base of traditional data centers to accelerated computing and build a new type of data center, AI factories, to produce a new commodity, artificial intelligence. AI will bring significant productivity gains to nearly every industry and help companies be more cost and energy efficient while expanding revenue opportunities. CSPs were the first generative AI movers. With NVIDIA, CSPs accelerated workloads to save money and power. The tokens generated by NVIDIA Hopper drive revenues for their AI services. And NVIDIA cloud instances attract rental customers from our rich ecosystem of developers. Strong and accelerated demand -- accelerating demand for generative AI

training and inference on Hopper platform propels our Data Center growth. Training continues to scale as models learn to be multimodal, understanding text, speech, images, video and 3D and learn to reason and plan. Our inference workloads are growing incredibly. With generative AI, inference, which is now about fast token generation at massive scale, has become incredibly complex. Generative AI is driving a from-foundation-up full stack computing platform shift that will transform every computer interaction. From today's information retrieval model, we are shifting to an answers and skills generation model of computing. AI will understand context and our intentions, be knowledgeable, reason, plan and perform tasks. We are fundamentally changing how computing works and what computers can do, from general purpose CPU to GPU accelerated computing, from instruction-driven software to intention-understanding models, from retrieving information to performing skills, and at the industrial level, from producing software to generating tokens, manufacturing digital intelligence. Token generation will drive a multiyear build-out of AI factories. Beyond cloud service providers, generative AI has expanded to consumer Internet companies and enterprise, Sovereign AI, automotive, and health care customers, creating multiple multibillion-dollar vertical markets. The Blackwell platform is in full production and forms the foundation for trillion-parameter scale generative AI. The combination of Grace CPU, Blackwell GPUs, NVLink, Quantum, Spectrum, mix and switches, high-speed interconnects and a rich ecosystem of software and partners let us expand and offer a richer and more complete solution for AI factories than previous generations. Spectrum-X opens a brand-new market for us to bring large-scale AI to Ethernet-only data centers. And NVIDIA NIMs is our new software offering that delivers enterprise-grade optimized generative AI to run on CUDA everywhere, from the cloud to on-prem data centers to RTX AI PCs through our expansive network of ecosystem partners. From Blackwell to Spectrum-X to NIMs, we are poised for the next wave of growth. Thank you.

**Simona Jankowski**

Thank you, Jensen. We will now open the call for questions. Operator, could you please poll for questions?

**Operator**

[Operator Instructions] Your first question comes from the line of Stacy Rasgon with Bernstein. Please go ahead.

**Stacy Rasgon**

Hi, guys. Thanks for taking my questions. My first one, I wanted to drill a little bit into the Blackwell comment that it's in full production now. What does that suggest with regard to shipments and delivery timing if that product is -- doesn't sound like it's sampling anymore. What does that mean when that's actually in customers' hands if it's in production now?

**Jensen Huang**

We will be shipping. Well, we've been in production for a little bit of time. But our production shipments will start in Q2 and ramp in Q3, and customers should have data centers stood up in Q4.

**Stacy Rasgon**

Got it. So this year, we will see Blackwell revenue, it sounds like?

**Jensen Huang**

We will see a lot of Blackwell revenue this year.

**Operator**

Our next question will come from the line of Timothy Arcuri with UBS. Please go ahead.

**Timothy Arcuri**

Thanks a lot. I wanted to ask, Jensen, about the deployment of Blackwell versus Hopper just between the systems nature and all the demand for GB that you have. How does the deployment of this stuff differ from Hopper? I guess I ask because liquid cooling at scale hasn't been done before, and there's some engineering challenges both at the node level and within the data center. So do these complexities sort of elongate the transition? And how do you sort of think about how that's all going? Thanks.

**Jensen Huang**

Yes. Blackwell comes in many configurations. Blackwell is a platform, not a GPU. And the platform includes support for air cooled, liquid cooled, x86 and Grace, InfiniBand, now Spectrum-X and very large NVLink domain that I demonstrated at GTC, that I showed at GTC. And so for some customers, they will ramp into their existing installed base of data centers that are already shipping Hoppers. They will easily transition from H100 to H200 to B100. And so Blackwell systems have been designed to be backwards compatible, if you will, electrically, mechanically. And of course, the software stack that runs on Hopper will run fantastically on Blackwell. We also have been priming the pump, if you will, with the entire ecosystem, getting them ready for liquid cooling. We've been talking to the ecosystem about Blackwell for quite some time. And the CSPs, the data centers, the ODMs, the system makers, our supply chain beyond them, the cooling supply chain base, liquid cooling supply chain base, data center supply chain base, no one is going to be surprised with Blackwell coming and the capabilities that we would like to deliver with Grace Blackwell 200. GB200 is going to be exceptional.

**Operator**

Our next question will come from the line of Vivek Arya with Bank of America Securities. Please go ahead.

**Vivek Arya**

Thanks for taking my question. Jensen, how are you ensuring that there is enough utilization of your products and that there isn't a pull-ahead or holding behavior because of tight supply, competition or other factors? Basically, what checks have you built in the system to give us confidence that monetization is keeping pace with your really very strong shipment growth?

**Jensen Huang**

Well, I guess, there's the big picture view that I'll come to, and then, but I'll answer your question directly. The demand for GPUs in all the data centers is incredible. We're racing every single day. And the reason for that is because applications like ChatGPT and GPT-4o, and now it's going to be multi-modality and Gemini and its ramp and Anthropic and all of the work that's being done at all the CSPs are consuming every GPU that's out there. There's also a long line of generative AI startups, some 15,000, 20,000 startups that in all different fields from multimedia to digital characters, of course, all kinds of design tool application -- productivity applications, digital biology, the moving of the AV industry to video, so that they can train end-to-end models, to expand the operating domain of self-driving cars. The list is just quite extraordinary. We're racing actually. Customers are putting a lot of pressure on us to deliver the systems and stand it up as quickly as possible. And of course, I haven't even mentioned all of the Sovereign AIs who would like to train all of their regional natural resource of their country, which is their data to train their regional models. And there's a lot of pressure to stand those systems up. So anyhow, the demand, I think, is really, really high and it outstrips our supply. Longer term, that's what -- that's the reason why I jumped in to make a few comments. Longer term, we're completely redesigning how computers work. And this is a platform shift. Of course, it's been compared to other platform shifts in the past. But time will clearly tell that this is much, much more profound than previous platform shifts. And the reason for that is because the computer is no longer an instruction-driven only computer. It's an intention-understanding computer. And it understands, of course, the way we interact with it, but it also understands our meaning, what we intend that we asked it to do and it has the ability to reason, inference iteratively to process a plan and come back with a solution. And so every aspect of the computer is changing in such a way that instead of retrieving prerecorded files, it is now generating contextually relevant intelligent answers. And so that's going to change computing stacks all over the world. And you saw a build that, in fact, even the PC computing stack is going to get revolutionized. And this is just the beginning of all the things that -- what people see today are the beginning of the things that we're working in our labs and the things that we're doing with all the startups and large companies and developers all over the world. It's going to be quite extraordinary.

**Operator**

Our next question will come from the line of Joe Moore with Morgan Stanley. Please go ahead.

**Joseph Moore**

Great. Thank you. I understand what you just said about how strong demand is. You have a lot of demand for H200 and for Blackwell products. Do you anticipate any kind of pause with Hopper and H100 as you sort of migrate to those products? Will people wait for those new products, which would be a good product to have? Or do you think there's enough demand for H100 to sustain growth?

**Jensen Huang**

We see increasing demand of Hopper through this quarter. And we expect to be -- we expect demand to outstrip supply for some time as we now transition to H200, as we transition to Blackwell. Everybody is anxious to get their infrastructure online. And the reason for that is because they're saving money and making money, and they would like to do that as soon as possible.

**Operator**

Our next question will come from the line of Toshiya Hari with Goldman Sachs. Please go ahead.

**Toshiya Hari**

Hi. Thank you so much for taking the question. Jensen, I wanted to ask about competition. I think many of your cloud customers have announced new or updates to their existing internal programs, right, in parallel to what they're working on with you guys. To what extent did you consider them as competitors, medium to long term? And in your view, do you think they're limited to addressing most internal workloads or could they be broader in what they address going forward? Thank you.

**Jensen Huang**

We're different in several ways. First, NVIDIA's accelerated computing architecture allows customers to process every aspect of their pipeline from unstructured data processing to prepare it for training, to structured data processing, data frame processing like SQL to prepare for training, to training to inference. And as I was mentioning in my remarks, that inference has really fundamentally changed, it's now generation. It's not trying to just detect the cat, which was plenty hard in itself, but it has to generate every pixel of a cat. And so the generation process is a fundamentally different processing architecture. And it's one of the reasons why TensorRT LLM was so well received. We improved the performance in using the same chips on our architecture by a factor of three. That kind of tells you something about the richness of our architecture and the richness of our software. So one, you could use NVIDIA for everything, from computer vision to image processing, the computer graphics to all modalities of computing. And as the world is now suffering from computing cost and computing energy inflation because general-purpose computing has run its course, accelerated computing is really the sustainable way of going forward. So accelerated computing is how you're going to save money in computing, is how you're going to save energy in computing. And so the versatility of our platform results in the lowest TCO for their data center. Second, we're in every cloud. And so for developers that are looking for a platform to develop on, starting with NVIDIA is always a great choice. And we're on-prem, we're in the cloud. We're in computers of any size and shape. We're practically everywhere. And so that's the second reason. The third reason has to do with the fact that we build AI factories. And this is becoming more an apparent to people that AI is not a chip problem only. It starts, of course, with very good chips and we build a whole bunch of chips for our AI factories, but it's a systems problem. In fact, even AI is now a systems problem. It's not just one large language model. It's a complex system of a whole bunch of large language models that are working together. And so the fact that NVIDIA builds this system causes us to optimize all of our chips to work together as a system, to be able to have software that operates as a system, and to be able to optimize across the system. And just to put it in perspective in simple numbers, if you had a $5 billion infrastructure and you improved the performance by a factor of two, which we routinely do, when you improve the infrastructure by a factor of two, the value too is $5 billion. All the chips in that data center doesn't pay for it. And so the value of it is really quite extraordinary. And this is the reason why today, performance matters everything. This is at a time when the highest performance is also the lowest cost because the infrastructure cost of carrying all of these chips cost a lot of money. And it takes a lot of money to fund the data center, to operate the data center, the people that goes along with it, the power that goes along with it, the real estate that goes along with it, and all of it adds up. And so the highest performance is also the lowest TCO.

**Operator**

Our next question will come from the line of Matt Ramsay with TD Cowen. Please go ahead.

**Matthew Ramsay**

Thank you very much. Good afternoon, everyone. Jensen, I've been in the data center industry my whole career. I've never seen the velocity that you guys are introducing new platforms at the same combination of the performance jumps that you're getting, I mean, 5x in training. Some of the stuff you talked about at GTC up to 30x in inference. And it's an amazing thing to watch but, it also creates an interesting juxtaposition where the current generation of product that your customers are spending billions of dollars on, it's going to be not as competitive with your new stuff, very, very much more quickly than the depreciation cycle of that product. So I'd like you to -- if you wouldn't mind speak a little bit about how you're seeing that situation evolve itself with customers. As you move to Blackwell, you're going to have very large installed bases, obviously software compatible, but large installed bases of product that's not nearly as performant as your new generation stuff. And it'd be interesting to hear what you see happening with customers along that path. Thank you.

**Jensen Huang**

Yes. I really appreciate it. Three points that I'd like to make. If you're 5% into the build-out versus if you're 95% into the build out, you're going to feel very differently. And because you're only 5% into the build-out anyhow, you build as fast as you can. And when Blackwell comes, it's going to be terrific. And then after Blackwell, as you mentioned, we have other Blackwells coming. And then there's a short -- we're in a one-year rhythm as we've explained to the world. And we want our customers to see our road map for as far as they like, but they're early in their build-out anyways and so they had to just keep on building, okay. And so there's going to be a whole bunch of chips coming at them, and they just got to keep on building and just, if you will, performance average your way into it. So that's the smart thing to do. They need to make money today. They want to save money today. And time is really, really valuable to them. Let me give you an example of time being really valuable, why this idea of standing up a data center instantaneously is so valuable and getting this thing called time to train is so valuable. The reason for that is because the next company who reaches the next major plateau gets to announce a groundbreaking AI. And the second one after that gets to announce something that's 0.3% better. And so the question is, do you want to be repeatedly the company delivering groundbreaking AI or the company delivering 0.3% better? And that's the reason why this race, as in all technology races, the race is so important. And you're seeing this race across multiple companies because this is so vital to have technology leadership, for companies to trust the leadership and want to build on your platform and know that the platform that they're building on is going to get better and better. And so leadership matters a great deal. Time to train matters a great deal. The difference between time to train that is three months earlier just to get it done, in order to get time to train on three-months project, getting started three months earlier is everything. And so it's the reason why we're standing up Hopper systems like mad right now because the next plateau is just around the corner. And so that's the second reason. The first comment that you made is really a great comment, which is how is it that we're doing -- we're moving so fast and advancing them quickly? Because we have all the stacks here. We literally build the entire data center and we can monitor everything, measure everything, optimize across everything. We know where all the bottlenecks are. We're not guessing about it. We're not putting up PowerPoint slides that look good. We're actually -- we also like our PowerPoint slides look good, but we're delivering systems that perform at scale. And the reason why we know they perform at scale is because we built it all here. Now one of the things that we do that's a bit of a miracle is that we build entire AI infrastructure here, but then we disaggregated and integrated into our customers' data centers however they liked. But we know how it's going to perform and we know where the bottlenecks are. We know where we need to optimize with them and we know where we have to help them improve their infrastructure to achieve the most performance. This deep intimate knowledge at the entire data center scale is fundamentally what sets us apart today. We build every single chip from the ground up. We know exactly how processing is done across the entire system. And so we understand exactly how it's going to perform and how to get the most out of it with every single generation. So I appreciate. Those are the three points.

**Operator**

Your next question will come from the line of Mark Lipacis with Evercore ISI. Please go ahead.

**Mark Lipacis**

Hi. Thanks for taking my question. Jensen, in the past, you've made the observation that general-purpose computing ecosystems typically dominated each computing era. And I believe the argument was that they could adapt to different workloads, get higher utilization, drive cost of compute cycle down. And this is a motivation for why you were driving to a general-purpose GPU CUDA ecosystem for accelerated computing. And if I mischaracterized that observation, please do let me know. So the question is, given that the workloads that are driving demand for your solutions are being driven by neural network training and inferencing, which on the surface seem like a limited number of workloads, then it might also seem to lend themselves to custom solutions. And so then the question is about does the general purpose computing framework become more at risk or is there enough variability or a rapid enough evolution on these workloads that support that historical general purpose framework? Thank you.

**Jensen Huang**

Yes. NVIDIA's accelerated computing is versatile, but I wouldn't call it general-purpose. Like for example, we wouldn't be very good at running the spreadsheet. That was really designed for general-purpose computing. And so there is a -- the control loop of an operating system code probably isn't fantastic for general-purpose compute, not for accelerated computing. And so I would say that we're versatile, and that's usually the way I describe it. There's a rich domain of applications that we're able to accelerate over the years, but they all have a lot of commonalities. Maybe some deep differences, but commonalities. They're all things that I can run in parallel, they're all heavily threaded. 5% of the code represents 99% of the run-time, for example. Those are all properties of accelerated computing. The versatility of our platform and the fact that we design entire systems is the reason why over the course of the last 10 years or so, the number of start-ups that you guys have asked me about in these conference calls is fairly large. And every single one of them, because of the brittleness of their architecture, the moment generative AI came along or the moment the fusion models came along, the moment the next models are coming along now. And now all of a sudden, look at this, large language models with memory because the large language model needs to have memory so they can carry on a conversation with you, understand the context. All of a sudden, the versatility of the Grace memory became super important. And so each one of these advances in generative AI and the advancement of AI really begs for not having a widget that's designed for one model. But to have something that is really good for this entire domain, properties of this entire domain, but obeys the first principles of software, that software is going to continue to evolve, that software is going to keep getting better and bigger. We believe in the scaling of these models. There's a lot of reasons why we're going to scale by easily a million times in the coming few years for good reasons, and we're looking forward to it and we're ready for it. And so the versatility of our platform is really quite key. And it's not -- if you're too brittle and too

specific, you might as well just build an FPGA or you build an ASIC or something like that, but that's hardly a computer.

**Operator**

Our next question will come from the line of Blayne Curtis with Jefferies. Please go ahead.

**Blayne Curtis**

Thanks for taking my question. Actually kind of curious, I mean, being supply constrained, how do you think about , I mean, you came out with a product for China, H20. I'm assuming there'd be a ton of demand for it, but obviously, you're trying to serve your customers with the other Hopper products. Just kind of curious how you're thinking about that in the second half. You could elaborate any impact, what you're thinking for sales as well as gross margin.

**Jensen Huang**

I didn't hear your questions. Something bleeped out.

**Simona Jankowski**

H20 and how you're thinking about allocating supply between the different Hopper products.

**Jensen Huang**

Well, we have customers that we honor and we do our best for every customer. It is the case that our business in China is substantially lower than the levels of the past. And it's a lot more competitive in China now because of the limitations on our technology. And so those matters are true. However, we continue to do our best to serve the customers in the markets there and to the best of our ability, we'll do our best. But I think overall, the comments that we made about demand outstripping supply is for the entire market and particularly so for H200 and Blackwell towards the end of the year.

**Operator**

Our next question will come from the line of Srini Pajjuri with Raymond James. Please go ahead.

**Srini Pajjuri**

Thank you. Jensen, actually more of a clarification on what you said. GB 200 systems, it looks like there is a significant demand for systems. Historically, I think you've sold a lot of HGX boards and some GPUs and the systems business was relatively small. So I'm just curious, why is it that now you are seeing such a strong demand for systems going forward? Is it just the TCO or is it something else or is it just the architecture? Thank you.

**Jensen Huang**

Yes. I appreciate that. In fact, the way we sell GB200 is the same. We disaggregate all of the components that make sense and we integrate it into computer makers. We have 100 different computer system configurations that are coming this year for Blackwell. And that is off the charts. Hopper, frankly, had only half, but that's at its peak. It started out with way less than that even. And so you're going to see liquid cooled version, air cooled version, x86 visions, Grace versions, so on and so forth. There's a whole bunch of systems that are being designed. And they're offered from all of our ecosystem of great partners. Nothing has really changed. Now of course, the Blackwell platform has expanded our offering tremendously. The integration of CPUs and the much more compressed density of computing, liquid cooling is going to save data centers a lot of money in provisioning power and not to mention to be more energy efficient. And so it's a much better solution. It's more expansive, meaning that we offer a lot more components of a data center and everybody wins. The data center gets much higher performance, networking from networking switches, networking. Of course, NICs, we have Ethernet now so that we can bring NVIDIA AI to a large-scale NVIDIA AI to customers who only operate only know how to operate Ethernet because of the ecosystem that they have. And so Blackwell is much more expansive. We have a lot more to offer our customers this generation around.

**Operator**

Our next question will come from the line William Stein with Truist Securities. Please go ahead.

**William Stein**

Great. Thanks for taking my question. Jensen, at some point, NVIDIA decided that while there are reasonably good CPUs available for data center operations, your ARM-based Grace CPU provides some real advantage that made that technology worth delivering to customers, perhaps related to cost or power consumption or technical synergies between Grace and Hopper, Grace and Blackwell. Can you address whether there could be a similar dynamic that might emerge on the client side, whereby while there are very good solutions, you've highlighted that Intel and AMD are very good partners and deliver great products in x86, but there might be some, especially in emerging AI workloads, some advantage that NVIDIA can deliver that others have more of a challenge?

**Jensen Huang**

Well, you mentioned some really good reasons. It is true that for many of the applications, our partnership with x86 partners are really terrific and we build excellent systems together. But Grace allows us to do something that isn't possible with the configuration, the system configuration today. The memory system between Grace and Hopper are coherent and connected. The interconnect between the two chips, calling it two chips is almost weird because it's like a superchip. The two of them are connected with this interface that's like a terabytes per second. It's off the charts. And the memory that's used by Grace is LPDDR. It's the first data center-grade low-power memory. And so we save a lot of power on every single node. And then finally, because of the architecture, because we can create our own architecture with the entire system now, we could create something that has a really large NVLink domain, which is vitally important to the next-generation large language models for inferencing. And so you saw that GB200 has a 72-node NVLink domain. That's like 72 Blackwells connected together into one giant GPU. And so we needed Grace Blackwells to be able to do that. And so there are architectural reasons, there are software programming reasons and then there are system reasons that are essential for us to build them that way. And so if we see opportunities like that, we'll explore it.

And today, as you saw at the build yesterday, which I thought was really excellent, Satya announced the next-generation PCs, Copilot+ PC, which runs fantastically on NVIDIA's RTX GPUs that are shipping in laptops. But it also supports ARM beautifully. And so it opens up opportunities for system innovation even for PCs.

**Operator**

Our last question comes from the line of C.J. Muse with Cantor Fitzgerald. Please go ahead.

**C.J. Muse**

Good afternoon. Thank you for taking the question. I guess, Jensen, a bit of a longer-term question. I know Blackwell hasn't even launched yet, but obviously, investors are forward-looking and amidst rising potential competition from GPUs and custom ASICs, how are you thinking about NVIDIA's pace of innovation and your million-fold scaling over the last decade, truly impressive. CUDA, Varsity, Precision, Grace, Cohere and Connectivity. When you look forward, what frictions need to be solved in the coming decade? And I guess, maybe more importantly, what are you willing to share with us today?

**Jensen Huang**

Well, I can announce that after Blackwell, there's another chip. And we are on a one-year rhythm. And so and you can also count that -- count on us having new networking technology on a very fast rhythm. We're announcing Spectrum-X for Ethernet. But we're all in on Ethernet, and we have a really exciting road map coming for Ethernet. We have a rich ecosystem of partners. Dell announced that they're taking Spectrum-X to market. We have a rich ecosystem of customers and partners who are going to announce taking our entire AI factory architecture to market. And so for companies that want the ultimate performance, we have InfiniBand computing fabric. InfiniBand is a computing fabric, Ethernet is a network. And InfiniBand, over the years, started out as a computing fabric, became a better and better network. Ethernet is a network and with Spectrum-X, we're going to make it a much better computing fabric. And we're committed -- fully committed to all three links, NVLink computing fabric for single computing domain to InfiniBand computing fabric, to Ethernet networking computing fabric. And so we're going to take all three of them forward at a very fast clip. And so you're going to see new switches coming, new NICs coming, new capability, new software stacks that run on all three of them. New CPUs, new GPUs, new networking NICs, new switches, a mound of chips that are coming. And all of it, the beautiful thing is all of it runs CUDA. And all of it runs our entire software stack. So you invest today on our software stack, without doing anything at all, it's just going to get faster and faster and faster and faster. And if you invest in our architecture today, without doing anything, it will go to more and more clouds and more and more data centers and everything just runs. And so I think the pace of innovation that we're bringing will drive up the capability, on the one hand, and drive down the TCO on the other hand. And so we should be able to scale out with the NVIDIA architecture for this new era of computing and start this new industrial revolution where we manufacture not just software anymore, but we manufacture artificial intelligence tokens and we're going to do that at scale. Thank you.

**Operator**

That will conclude our question-and-answer session and our call for today. We thank you all for joining and you may now disconnect.