NVIDIA Corporation (NVDA) Q2 2026 Earnings Conference Call August 27, 2025 5:00 PM ET

Company Participants

Colette M. Kress - Executive VP & CFO
Jen-Hsun Huang - Co-Founder, CEO, President & Director
Toshiya Hari - Vice President of Investor Relations & Strategic Finance

Conference Call Participants

Aaron Christopher Rakers - Wells Fargo Securities, LLC, Research Division
Benjamin Alexander Reitzes - Melius Research LLC
Christopher James Muse - Cantor Fitzgerald & Co., Research Division
James Edward Schneider - Goldman Sachs Group, Inc., Research Division
Joseph Lawrence Moore - Morgan Stanley, Research Division
Stacy Aaron Rasgon - Sanford C. Bernstein & Co., LLC., Research Division
Timothy Michael Arcuri - UBS Investment Bank, Research Division
Vivek Arya - BofA Securities, Research Division

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 Second Quarter Fiscal 2026 Financial Results Conference Call. [Operator Instructions] Thank you. Toshiya Hari, you may begin your conference.

Toshiya Hari

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the second 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 third quarter of fiscal 2026. 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, August 27, 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 M. Kress

Thank you, Toshiya. We delivered another record quarter while navigating what continues to be a dynamic external environment. Total revenue was $46.7 billion, exceeded our outlook as we grew sequentially across all market platforms. Data center revenue grew 56% year-over-year. Data center revenue also grew sequentially despite the $4 billion decline in H20 revenue. NVIDIA's Blackwell platform reached record levels, growing sequentially by 17%. We began production shipments of GB300 in Q2. Our full stack AI solutions for cloud service providers, neoclouds, enterprises and sovereigns are all contributing to our growth.

We are at the beginning of an industrial revolution that will transform every industry. We see $3 trillion to $4 trillion in AI infrastructure spend in the -- by the end of the decade. The scale and scope of these build-outs present significant long-term growth opportunities for NVIDIA.

The GB200 NVL system is seeing widespread adoption with deployments at CSPs and consumer Internet companies. Lighthouse model builders, including OpenAI, Meta and Mistral are using the GB200 NVL72 at data center scale for both training, next-generation models and serving inference models in production.

The new Blackwell Ultra platform has also had a strong quarter, generating tens of billions in revenue. The transition to the GB300 has been seamless for major cloud service providers due to its shared architecture, software and physical footprint with the GB200, enabling them to build and deploy GB300 racks with ease. The transition to the new GB300 rack-based architecture has been seamless. Factory builds in late July and early August were successfully converted to support the GB300 ramp, and today, full production is underway. The current run rate is back at full speed, producing approximately 1,000 racks per week. This output is expected to accelerate even further throughout the third quarter as additional capacity comes online.

We expect widespread market availability in the second half of the year as CoreWeave prepares to bring their GB300 instance to market as they are already seeing 10x more inference performance on reasoning models compared to H100. Compared to the previous Hopper generation, GB300 NVL72 AI factories promise a 10x improvement in token per watt energy efficiency, which translates to revenues as data centers are power limited.

The chips of the Rubin platform are in fab, the Vera CPU, Rubin GPU, CX9 SuperNIC, NVLink 144 scale up switch, Spectrum-X scale out and scale across switch, and the silicon photonics processor. Rubin remains on schedule for volume production next year. Rubin will be our third-generation NVLink rack scale AI supercomputer with a mature and full-scale supply chain. This keeps us on track with our pace of an annual product cadence and continuous innovation across compute, networking, systems and software.

In late July, the U.S. government began reviewing licenses for sales of H20 to China customers. While a select number of our China- based customers have received licenses over the past few weeks, we have not shipped any H20 based on those licenses. USG officials have expressed an expectation that the USG will receive 15% of the revenue generated from licensed H20 sales, but to date, the USG has not published a regulation codifying such requirement.

We have not included H20 in our Q3 outlook as we continue to work through geopolitical issues. If geopolitical issues reside, we should ship $2 billion to $5 billion in H20 revenue in Q3. And if we had more orders, we can bill more. We continue to advocate for the U.S. government to approve Blackwell for China. Our products are designed and sold for beneficial commercial use, and every license sale we make will benefit the U.S. economy, the U.S. leadership. In highly competitive markets, we want to win the support of every developer. America's AI technology stack can be the world's standard if we race and compete globally.

Notably, in the quarter was an increase in Hopper 100 and H200 shipments. We also sold approximately $650 million of H20 in Q2 to an unrestricted customer outside of China. The sequential increase in Hopper demand indicates the breadth of data center workloads that run on accelerated computing and the power of CUDA libraries and full stack optimizations, which continuously enhance the performance and economic value of our platform.

As we continue to deliver both Hopper and Blackwell GPUs, we are focusing on meeting the soaring global demand. This growth is fueled by capital expenditures from the cloud to enterprises, which are on track to invest $600 billion in data center infrastructure and compute this calendar year alone, nearly doubling in 2 years. We expect annual AI infrastructure investments to continue growing, driven by the several factors: reasoning agentic AI requiring orders of magnitude more training and inference compute, global build- outs for sovereign AI, enterprise AI adoption, and the arrival of physical AI and robotics.

Blackwell has set the benchmark as it is the new standard for AI inference performance. The market for AI inference is expanding rapidly with reasoning and agentic AI gaining traction across industries. Blackwell's rack scale NVLink and CUDA full stack architecture addresses this by redefining the economics of inference. New NVFP4 4-bit precision and NVLink 72 on the GB300 platform delivers a 50x increase in energy efficiency per token compared to Hopper, enabling companies to monetize their compute at unprecedented scale. For instance, a $3 million investment in GB200 infrastructure can generate $30 million in token revenue, a 10x return.

NVIDIA software innovation, combined with the strength of our developer ecosystem, has already improved Blackwell's performance by more than 2x since its launch. Advances in CUDA, TensorRT-LLM and Dynamo are unlocking maximum efficiency. CUDA library contributions from the open source community, along with NVIDIA's open libraries and frameworks are now integrated into millions of workflows. This powerful flywheel of collaborative innovation between NVIDIA and global community contribution strengthens NVIDIA's performance leadership. NVIDIA is a top contributor to OpenAI models, data and software.

Blackwell has introduced a groundbreaking numerical approach to large language model pretraining. Using NVFP4 computations on the GB300 can now achieve 7x faster training than the H100, which uses FP8. This innovation delivers the accuracy of 16-bit precision with the speed and efficiency of 4 bit, setting a new standard for AI factor efficiency and scalability.

The AI industry is quickly adopting this revolutionary technology with major players such as AWS, Google Cloud, Microsoft Azure and OpenAI as well as Cohere, Mistral, Kimi AI, Perplexity, Reflection and Runway already embracing it. NVIDIA's performance leadership was further validated in the latest MLPerf Training benchmarks, where the GB200 delivered a clean sweep. Be on the lookout for the upcoming MLPerf Inference results in September, which will include benchmarks based on the Blackwell Ultra.

NVIDIA RTX PRO servers are in full production for the world system makers. These are air-cooled PCIe-based systems integrated seamlessly into standard IT environments and run traditional enterprise IT applications as well as the most advanced agentic and physical AI applications. Nearly 90 companies including many global leaders are already adopting RTX PRO servers. Hitachi uses them for real-time simulation and digital twins, Lilly for drug discovery, Hyundai for factory design and AV validation, and Disney for immersive storytelling. As enterprises modernize data centers, RTX PRO servers are poised to become a multibillion-dollar product line.

Sovereign AI is one on the rise as the nation's ability to develop its own AI using domestic infrastructure, data and talent presents a significant opportunity for NVIDIA. NVIDIA is at the forefront of landmark initiatives across the U.K. and Europe. The European

Union plans to invest EUR 20 billion to establish 20 AI factories across France, Germany, Italy and Spain, including 5 gigafactories to increase its AI compute infrastructure by tenfold.

In the U.K., the Isambard-AI supercomputer powered by NVIDIA was unveiled at the country's most powerful AI system, delivering 21 exaflops of AI performance to accelerate breakthroughs in fields of drug discovery and climate modeling. We are on track to achieve over [20 billion] in Sovereign AI revenue this year, more than double than that last year.

Networking delivered record revenue of $7.3 billion, and escalating demands of AI compute clusters necessitate high efficiency and low latency networking. This represents a 46% sequential and 98% year-on-year increase with strong demand across Spectrum- X Ethernet, InfiniBand and NVLink. Our Spectrum-X enhanced Ethernet solutions provide the highest throughput and lowest latency network for Ethernet AI workloads. Spectrum-X Ethernet delivered double-digit sequential and year-over-year growth with annualized revenue exceeding $10 billion. At Hot Chips, we introduced Spectrum-XGS Ethernet, a technology design to unify disparate data centers into giga-scale AI super factories. [CoreWeave] is an initial adopter of the solution, which is projected to double GPU-to-GPU communication speed.

InfiniBand revenue nearly doubled sequentially, fueled by the adoption of XDR technology, which provides double the bandwidth improvement over its predecessor, especially valuable for the model builders. The world's fastest switch, NVLink, with 14x the bandwidth of PCIe Gen 5 delivered strong growth as customers deployed Grace Blackwell NVLink rack scale systems.

The positive reception to NVLink Fusion, which allows semi-custom AI infrastructure, has been widespread. Japan's upcoming FugakuNEXT will integrate Fujitsu's CPUs with our architecture via NVLink Fusion. It will run a range of workloads, including AI, supercomputing and quantum computing. FugakuNEXT joins a rapidly expanding list of leading quantum supercomputing and research centers running on NVIDIA's CUDA-Q quantum platform, including [ULIC], AIST, [NNF] and NERSC, supported by over 300 ecosystem partners, including AWS, Google Quantum AI, Quantinuum, QuEra and PsiQuantum.

Jetson Thor, our new robotics computing platform, is now available. Thor delivers an order of magnitude greater AI performance and energy efficiency than NVIDIA AGX Orin. It runs the latest generative and reasoning AI models at the edge in real time, enabling state-of-the-art robotics.

Adoption of NVIDIA's robotics full stack platform is growing at rapid rate, over 2 million developers and 1,000-plus hardware software applications and sensor partners taking our platform to market. Leading enterprises across industries have adopted Thor, including Agility Robotics, Amazon Robotics, Boston Dynamics, Caterpillar, Figure, Hexagon, Medtronic and Meta.

Robotic applications require exponentially more compute on the device and in infrastructure, representing a significant long-term demand driver for our data center platform. NVIDIA Omniverse with Cosmos is our data center physical AI digital twin platform built for development of robot and robotic systems. This quarter, we announced a major expansion of our partnership with Siemens to enable AI automatic factories. Leading European robotics companies, including Agile Robots, NEURA Robotics and Universal Robots are building their latest innovations with the Omniverse platform.

Transitioning to a quick summary of our revenue by geography. China declined on a sequential basis to low single-digit percentage of data center revenue. Note, our Q3 outlook does not include H20 shipments to China customers. Singapore revenue represented 22% of second quarter's billed revenue as customers have centralized their invoicing in Singapore. Over 99% of data center compute revenue billed to Singapore was for U.S.-based customers.

Our gaming revenue was a record $4.3 billion, a 14% sequential increase and a 49% jump year-on-year. This was driven by the ramp of Blackwell GeForce GPUs and strong sales continued as we increase supply availability. This quarter, we shipped GeForce RTX 5060 desktop GPU. It brings double the performance along with advanced ray tracing, neural rendering and AI-powered DLSS 4 gameplay to millions of gamers worldwide. Blackwell is coming to GeForce NOW in September. This is GeForce NOW's most significant upgrade, offering RTX 5080 cost performance, minimal latency and 5K resolution at 120 frames per second. We are also doubling the GeForce NOW catalog to over 4,500 titles, the largest library of any cloud gaming service.

For AI enthusiasts, on-device AI performs the best RTX GPUs. We partnered with OpenAI to optimize their open source GPT models for high-quality, fast and efficient inference on millions of RTX-enabled Window devices. With the RTX platform stack, Window developers can create AI applications designed to run on the world's largest AI PC user base.

Professional Visualization revenue reached $601 million, a 32% year-on-year increase. Growth was driven by an adoption of the high- end RTX Workstation GPUs and AI-powered workload like design, simulation and prototyping. Key customers are leveraging our solutions to accelerate their operations. Activision Blizzard uses RTX workstations to enhance creative workflows. While robotics innovator Figure AI powers its humanoid robots with RTX-embedded GPUs.

Automotive revenue, which includes only in-car compute revenue, was $586 million, up 69% year-on-year, primarily driven by self- driving solutions. We have begun shipments of NVIDIA Thor SoC, the successor to Orin. Thor's arrival coincides with the industry's accelerating shift to vision language model architecture, generative AI and higher levels of autonomy. Thor is the most successful robotics and AV computer we've ever created. Thor will power. Our full stack Drive AV software platform is now in production, opening up billions to new revenue opportunities for NVIDIA while improving vehicle safety and autonomy.

Now moving to the rest of our P&L. GAAP gross margin was 72.4% and non-GAAP gross margin was 72.7%. These figures include a $180 million or 40 basis point benefit from releasing previously reserved H20 inventory. Excluding this benefit, non-GAAP gross margins would have been 72.3%, still exceeding our outlook. GAAP operating expenses rose 8% and 6% on a non-GAAP basis sequentially. This increase was driven by higher compute and infrastructure costs as well as higher compensation and benefit costs. To support the ramp of Blackwell and Blackwell Ultra, inventory increased sequentially from $11 billion to $15 billion in Q2.

While we prioritize funding our growth and strategic initiatives, in Q2, we returned $10 billion to shareholders through share repurchases and cash dividends. Our Board of Directors recently approved a $60 billion share repurchase authorization to add to our remaining $14.7 billion of authorization at the end of Q2.

Okay. Let me turn it to the outlook for the third quarter. Total revenue is expected to be $54 billion, plus or minus 2%. This represents over $7 billion in sequential growth. Again, we do not assume any H20 shipments to China customers in our outlook. GAAP and non- GAAP gross margins are expected to be 73.3%, 73.5%, respectively, plus or minus 50 basis points. We continue to expect to exit the year with non-GAAP gross margins in the mid-70s. GAAP and non-GAAP operating expenses are expected to be approximately $5.9 billion and $4.2 billion, respectively. For the full year, we expect operating expenses to grow in the high 30s range year-over-year, up from our prior expectations of the mid-30s. We are accelerating investments in the business to address the magnitude of growth opportunities that lie ahead.

GAAP and non-GAAP other income and expenses are expected to be an income of approximately $500 million, excluding gains and losses from nonmarketable and public 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 data are included in the CFO commentary and other information available on our website.

In closing, let me highlight upcoming events for the financial community. We will be at the Goldman Sachs Technology Conference on September 8 in San Francisco. Our annual NDR will commence the first part of October. GTC data center begins on October 27, with Jensen's keynote scheduled for the 28. We look forward to seeing you at these events. Our earnings call to discuss the results of our third quarter of fiscal 2026 is scheduled for November 19. We will now open the call for questions. Operator, would you please poll for questions?

Question-and-Answer Session

Operator

[Operator Instructions] Your first question comes from CJ Muse with Cantor Fitzgerald.

Christopher James Muse

I guess with wafer in to rack out lead times of 12 months, you confirmed on the call today that Rubin is on track for ramp in the second half. And obviously, many of these investments are multiyear projects contingent upon power, cooling, et cetera. I was hoping perhaps could you take a high-level view and speak to your vision for growth into 2026. And as part of that, if you can kind of comment between network and data center would be very helpful.

Jen-Hsun Huang

Yes. Thanks, CJ. At the highest level of growth drivers would be the evolution, the introduction, if you will, of reasoning agentic AI. Where chatbots used to be one shot, you give it a prompt and it would generate the answer, now the AI does research. It thinks and does a plan, and it might use tools. And so it's called long thinking; and the longer it thinks, oftentimes, it produces better answers.

And the amount of computation necessary for 1 shot versus reasoning agentic AI models could be 100x, 1,000x and potentially even more as the amount of research and basically reading and comprehension that it goes off to do. And so the amount of computation that has resulted in agentic AI has grown tremendously. And of course, the effectiveness has also grown tremendously. Because of agentic AI, the amount of hallucination has dropped significantly. You can now use tools and perform tasks. Enterprises have been opened up. As a result of agentic AI and vision language models, we now are seeing a breakthrough in physical AI, in robotics, autonomous systems. So the last year, AI has made tremendous progress and agentic systems, reasoning systems is completely revolutionary.

Now we built the Blackwell NVLink 72 system, a rack scale computing system, for this moment. We've been working on it for several years. This last year, we transitioned from NVLink 8, which is a node scale computing, each node is a computer, to now NVLink 72, where each rack is a computer. That disaggregation of NVLink 72 into a rack scale system was extremely hard to do, but the results are extraordinary. We're seeing orders of magnitude speed up and therefore, energy efficiency and therefore, cost effectiveness of token generation because of NVLink 72.

And so over the next couple of years, you're going to -- well, you asked about longer term. Over the next 5 years, we're going to scale into with Blackwell, with Rubin and follow-ons to scale into effectively a $3 trillion to $4 trillion AI infrastructure opportunity. The last couple of years, you have seen that CapEx has grown in just the top 4 CSPs by -- has doubled and grown to about $600 billion. So we're in the beginning of this build-out, and the AI technology advances has really enabled AI to be able to adopt and solve problems to many different industries.

Operator

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

Vivek Arya

Colette, I just wanted to clarify the $2 billion to $5 billion in China. What needs to happen? And what is the sustainable pace of that China business as you get into Q4?

And then, Jensen, for you on the competitive landscape. Several of your large customers already have or are planning many ASIC projects. I think 1 of your ASIC competitors, Broadcom, signaled that they could grow their AI business almost 55%, 60% next year. Any scenario in which you see the market moving more towards ASICs and away from NVIDIA GPU? Just what are you hearing from your customers? How are they managing this split between the use of merchant silicon and ASICs?

Colette M. Kress

Thanks, Vivek. So let me first answer your question regarding what will it take for the H20s to be shipped. There is interest in our H20s. There is the initial set of license that we received. And then additionally, we do have supply that we are ready, and that's why we communicated that somewhere in the range of about $2 billion to $5 billion this quarter, we could potentially ship.

We're still waiting on several of the geopolitical issues going back and forth between the governments and the companies trying to determine their purchases and what they want to do. So it's still open at this time, and we're not exactly sure what that full amount will be this quarter. However, if more interest arrives, more licenses arrives, again, we can also still build additional H20 and ship more as well.

Jen-Hsun Huang

NVIDIA builds very different things in ASICs. So let's talk about ASICs first. A lot of projects are started. Many start-up companies are created. Very few products go into production. And the reason for that is it's really hard. Accelerated computing is unlike general- purpose computing. You don't write software and just compile it into a processor. Accelerated computing is a full-stack co-design problem. And AI factories in the last several years have become so much more complex because of the scale of the problems have grown so significantly. It is really the ultimate, the most extreme computer science problem the world's ever seen obviously.

And so the stack is complicated. The models are changing incredibly fast from generative based on auto regressive to degenerative based on diffusion to mixed models to multi-modality. The number of different models that are coming out that are either derivatives of transformers or evolutions of transformers is just daunting.

One of the advantages that we have is that NVIDIA is available in every cloud. We're available from every computer company. We're available from the cloud to on-prem to edge to robotics on the same programming model. And so it's sensible that every framework in the world supports NVIDIA.

When you're building a new model architecture, releasing it on NVIDIA is most sensible. And so the diversity of our platform, both in the ability to evolve into any architecture, the fact that we're everywhere, and also, we accelerate the entire pipeline, everything from data processing to pretraining to post training with reinforcement learning, all the way out to inference. And so when you build a data center with NVIDIA platform in it, the utility of it is best. The lifetime usefulness is much, much longer.

And then I would just say that in addition to all of that -- and it's just a really extremely complex systems problem anymore. People talk about the chip itself. There's one ASIC, the GPU that many people talk about. But in order to build Blackwell the platform and Rubin the platform, we had to build CPUs that connect fast memory, low -- extremely energy-efficient memory for large KB caching necessary for agentic AI to the GPU to a SuperNIC to a scale up switch, we call NVLink, completely revolutionary, we're in our fifth generation now, to a scale out switch, whether it's Quantum or Spectrum-X Ethernet, to now scale across switches so that we could prepare for these AI super factories with multiple gigawatts of computing all connected together. We call that Spectrum-XGS. We just announced that at Hot Chips this week. And so the complications, the complexity of everything that we do is really quite extraordinary. It's just done at a really, really extreme scale now.

And then lastly, if I could just say one more thing, we're in every cloud for a good reason. Not only are we the most energy efficient. Our perf per watt is the best of any computing platform. And in a world of power-limited data centers, perf per watt drives directly to revenues. And you've heard me say before that, in a lot of ways, the more you buy, the more you grow. And because our perf per dollar, the performance per dollar is so incredible, you also have extremely great margins.

So the growth opportunity with NVIDIA's architecture and the gross margins opportunity with NVIDIA's architecture is absolutely the best. And so there's a lot of reasons why NVIDIA is chosen by every cloud and every start-up and every computer company. We're really a holistic full-stack solution for AI factories.

Operator

Your next question comes from Ben Reitzes with Melius.

Benjamin Alexander Reitzes

Jensen, I wanted to ask you about your $3 trillion to $4 trillion in data center infrastructure spend by the end of the decade. Previously, you talked about something in the $1 billion range, which I believe was just for compute by 2028. If you take past comments, $3 trillion to $4 trillion would imply maybe $2 billion plus in compute spend. And just wanted to know if that was right and that's what you're seeing by the end of the decade. And wondering what you think your share will be of that. Your share right now of total

infrastructure compute-wise is very high, so I wanted to see. And also if there's any bottlenecks you're concerned about like power to get to the $3 trillion to $4 trillion.

Jen-Hsun Huang

Thanks. As you know, the CapEx of just the top 4 hyperscalers has doubled in 2 years. As the AI revolution went into full steam, as the AI race is now on, the CapEx spend has doubled to $600 billion per year. There's 5 years between now and the end of the decade, and $600 billion only represents the top 4 hyperscalers. We still have the rest of the enterprise companies building on-prem. You have cloud service providers building around the world. United States represents about 60% of the world's compute. And over time, you would think that artificial intelligence would reflect GDP scale and growth and so -- and would be, of course, accelerating GDP growth.

And so our contribution to that is a large part of the AI infrastructure. Out of a gigawatt AI factory, which can go anywhere from $50 billion to plus or minus 10%, let's say, $50 billion to $60 billion, we represent about $35 billion plus or minus of that and $35 billion out of $50 billion per gigawatt data center.

And of course, what you get for that is not a GPU. I think people -- we're famous for building the GPU and inventing the GPU, but as you know, over the last decade, we've really transitioned to become an AI infrastructure company. It takes 6 chips just to build -- 6 different types of chips just to build a Rubin AI supercomputer. And just to scale that out to a gigawatt, you have hundreds of thousands of GPU compute nodes and a whole bunch of racks. And so we're really an AI infrastructure company, and we're hoping to continue to contribute to growing this industry, making AI more useful and then very importantly, driving the performance per watt because the world, as you mentioned, limiters, it will always likely be power limitations or AI -- building limitations. And so we need to squeeze as much out of that factory as possible.

NVIDIA's performance per unit of energy used drives the revenue growth of that factory. It directly translates. If you have a 100- megawatt factory, perf per 100 megawatt drives your revenues. It's tokens per 100 megawatts of factory. In our case also, the performance per dollar spent is so high that your gross margins are also the best. But anyhow, these are the limiters going forward and $3 trillion to $4 trillion is fairly sensible for the next 5 years.

Operator

Next question comes from Joe Moore of Morgan Stanley.

Joseph Lawrence Moore

Great. Congratulations on reopening the China opportunity. Can you talk about the long-term prospects there? You've talked about, I think, half of AI software world being there. How much can NVIDIA grow in that business? And how important is it that you get the Blackwell architecture ultimately licensed there?

Jen-Hsun Huang

The China market, I've estimated to be about $50 billion of opportunity for us this year if we were able to address it with competitive products. And if it's $50 billion this year, you would expect it to grow, say, 50% per year. As the rest of the world's AI market is growing as well.

It is the second largest computing market in the world, and it is also the home of AI researchers. About 50% of the world's AI researchers are in China. The vast majority of the leading open source models are created in China. And so it's fairly important, I think, for the American technology companies to be able to address that market. And open source, as you know, is created in one country, but it's used all over the world.

The open source models that have come out of China are really excellent. DeepSeek, of course, gained global notoriety. Qwen is excellent. Kimi's excellent. There's a whole bunch of new models that are coming out. They're multimodal. They're great language models. And it's really fueled the adoption of AI in enterprises around the world because enterprises want to build their own custom proprietary software stacks. And so open source model's really important for enterprise. It's really important for SaaS who also would like to build proprietary systems. It has been really incredible for robotics around the world.

And so open source is really important, and it's important that the American companies are able to address it. This is -- it's going to be a very large market. We're talking to the administration about the importance of American companies to be able to address the Chinese market. And as you know, H20 has been approved for companies that are not on the entities list, and many licenses have been approved. And so I think the opportunity for us to bring Blackwell to the China market is a real possibility. And so we just have to keep advocating the sensibility of and the importance of American tech companies to be able to lead and win the AI race and help make the American tech stack the global standard.

Operator

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

Aaron Christopher Rakers

Yes. Thank you for the question. I want to go back to the Spectrum-XGS announcement this week and thinking about the Ethernet product now pushing over $10 billion of annualized revenue. Jensen, how -- what is the opportunity set that you see for Spectrum- XGS? Do we think about this as kind of the data center interconnect layer? Any thoughts on the sizing of this opportunity within that Ethernet portfolio?

Jen-Hsun Huang

We now offer 3 networking technologies. One is for scale up. One is for scale out and one for scale across. Scale up is so that we could build the largest possible virtual GPU, the virtual compute node. NVLink is revolutionary. NVLink 72 is what made it possible for Blackwell to deliver such an extraordinary generational jump over Hopper's NVLink 8. At a time when we have long thinking models, agentic AI reasoning systems, the NVLink basically amplifies the memory bandwidth, which is really critical for reasoning systems. And so NVLink 72 is fantastic.

We then scale out with networking, which we have 2. We have InfiniBand, which is unquestionably the lowest latency, the lowest jitter, the best scale-out network. It does require more expertise in managing those networks. And for supercomputing, for the leading model makers, InfiniBand, Quantum InfiniBand is the unambiguous choice. If you were to benchmark an AI factory, the ones with InfiniBand are the best performance.

For those who would like to use Ethernet because their whole data center is built with Ethernet, we have a new type of Ethernet called Spectrum Ethernet. Spectrum Ethernet is not off the shelf. It has a whole bunch of new technologies designed for low latency and low jitter and congestion control. And it has the ability to come closer, much, much closer to InfiniBand than anything that's out there. And that is -- we call that Spectrum-X Ethernet.

And then finally, we have Spectrum-XGS, a giga scale for connecting multiple data centers, multiple AI factories into a super factory, a gigantic system. And we're going to

-- you're going to see that networking obviously is very important in AI factories. In fact, choosing the right networking, the performance, the throughput improvement, going from 65% to 85% or 90%, that kind of step-up because of your networking capability effectively makes networking free. Choosing the right networking, you're basically paying -- you'll get a return on it like you can't believe because the AI factory, a gigawatt, as I mentioned before, could be $50 billion. And so the ability to improve the efficiency of that factory by tens of percent is -- results in $10 billion, $20 billion worth of effective benefit. And so this -- the networking is a very important part of it.

It's the reason why NVIDIA dedicates so much in networking. That's the reason why we purchased Mellanox 5.5 years ago. And Spectrum-X, as we mentioned earlier, is now quite a sizable business, and it's only about 1.5 years old. So Spectrum-X is a home run. All 3 of them are going to be fantastic. NVLink scale up, Spectrum-X and InfiniBand scale out, and then Spectrum-XGS for scale across.

Operator

Your next question comes from Stacy Rasgon with Bernstein Research.

Stacy Aaron Rasgon

I have a more tactical question for Colette. So on the guidance, we're up over $7 billion. The vast bulk of that is going to be from data center. How do I think about apportioning that $7 billion out across Blackwell versus Hopper versus networking? I mean it looks

like Blackwell was probably $27 billion in the quarter, up from maybe $23 billion last quarter. Hopper is still $6 billion or $7 billion post the H20. Like do you think the Hopper strength continues? Just how do I think about parsing that $7 billion out across those 3 different components?

Colette M. Kress

Thanks, Stacy, for the question. First part of it, looking at our growth between Q2 and Q3, Blackwell is still going to be the lion's share of what we have in terms of data center. But keep in mind, that helps both our compute side as well as it helps our networking side because we are selling those significant systems that are incorporating the NVLink that Jensen just spoke about.

Selling Hopper, we are still selling it. H100, H200s, we are. Again, they are HGX systems, and I still believe our Blackwell will be the lion's share of what we're doing on there. So we'll continue. We don't have any more specific details in terms of how we'll finish our quarter, but you should expect Blackwell again to be the driver of the growth.

Operator

Your next question comes from Jim Schneider of Goldman Sachs.

James Edward Schneider

You've been very clear about the reasoning model opportunity that you see, and you've also been relatively clear about technical specs for Rubin. But maybe you could provide a little bit of context about how you view the Rubin product transition going forward. What incremental capability does that offer to customers? And would you say that Rubin is a bigger, smaller or similar step-up in terms of performance from a capability perspective relative to what we saw with Blackwell?

Jen-Hsun Huang

Yes, thanks. Rubin. Rubin, we're on an annual cycle. And the reason why we're on an annual cycle is because we can do so to accelerate the cost reduction and maximize the revenue generation for our customers. When we increase the perf per watt, the token generation per amount of usage of energy, we are effectively driving the revenues of our customers. The perf per watt of Blackwell will be for reasoning systems in order of magnitude higher than Hopper. And so for the same amount of energy, and everybody's data center is energy limited by definition, for any data center, we -- using Blackwell, you'll be able to maximize your revenues compared to anything we've done in the past, compared to anything in the world today and because the perf per dollar, the performance is so good that the perf per dollar invested in the capital would also allow you to improve your gross margins.

To the extent that we have great ideas for every single generation, we could improve the revenue generation, improve the AI capability, improve the margins of our customers by releasing new architectures. And so we advise our partners, our customers to pace themselves and to build these data centers on an annual rhythm. And Rubin is going to have a whole bunch of new ideas.

I'll pause for a second because I've got plenty of time between now and a year from now to tell you about all the breakthroughs that Rubin is going to bring, but Rubin has a lot of great ideas. I'm anxious to tell you, but I can't right now. And I'll save it for GTC to tell you more and more about it. But nonetheless, for the next year, we're ramping really hard into now Grace Blackwell, GB200, and then now Blackwell Ultra, GB300, we're ramping really hard into data centers. This year is obviously a record-breaking year. I expect next year to be a record-breaking year. And while we continue to increase the performance of AI capabilities as we race towards artificial superintelligence on the one hand and continue to increase the revenue generation capabilities of our hyperscalers on the other hand.

Operator

Your final question comes from Timothy Arcuri with UBS.

Timothy Michael Arcuri

Jensen, I wanted to ask you, just answered the question. You threw out a number. You said 50% CAGR for the AI market. So I'm wondering how much visibility that you have into next year. Is that kind of a reasonable bogey in terms of how much your data center revenue should grow next year? I would think you'll grow at least in line with that CAGR? And maybe are there any puts and takes to that?

Jen-Hsun Huang

Well, I think the best way to look at it is we have reasonable forecasts from our large customers for next year, a very, very significant forecast. And we still have a lot of businesses that we're still winning and a lot of start-ups that are still being created. Don't forget that the number of start-ups for -- native-AI start-ups was $100 billion was funded last year. This year, the year is not even over yet, it's $180 billion funded. If you look at AI native, the top AI-native start-ups that are generating revenues last year was $2 billion. This year, it's $20 billion. Next year be 10x higher than this year is not inconceivable. And the open source models is now opening up large enterprises, SaaS companies, industrial companies, robotics companies to now join the AI revolution, another source of growth. And whether it's AI natives or enterprise SaaS or industrial AI or start-ups, we're just seeing just enormous amount of interest in AI and demand for AI.

Right now, the buzz is -- I'm sure all of you know about the buzz out there. The buzz is everything sold out. H100 sold out. H200s are sold out. Large CSPs are coming out renting capacity from other CSPs. And so the AI-native start-ups are really scrambling to get capacity so that they could train their reasoning models. And so the demand is really, really high.

But the long-term outlook between where we are today, CapEx has doubled in 2 years. It is now running about $600 billion a year just in the large hyperscalers. For us to grow into that $600 billion a year, representing a significant part of that CapEx isn't unreasonable. And so I think the next several years, surely through the decade, we see just a really fast growing, really significant growth opportunities ahead.

Let me conclude with this. Blackwell is the next-generation AI platform the world has been waiting for. It delivers an exceptional generational leap. NVIDIA's NVLink 72 rack scale computing is revolutionary, arriving just in time as reasoning AI models drive order of magnitude increases in training and inference performance requirement. Blackwell Ultra is ramping at full speed and the demand is extraordinary.

Our next platform Rubin, is already in fab. We have 6 new chips that represents the Rubin platform. They have all ticked up at TSMC. Rubin will be our third-generation NVLink rack scale AI supercomputer. And so we expect to have a much more mature and fully scaled up supply chain. Blackwell and Rubin AI factory platforms will be scaling into the $3 trillion to $4 trillion global AI factory build out through the end of the decade.

Customers are building ever greater scale AI factories from thousands of Hopper GPUs in tens of megawatt data centers to now hundreds of thousands of Blackwells in 100-megawatt facilities. And soon, we'll be building millions of Rubin GPU platforms, powering multi-gigawatt multisite AI super factories.

With each generation, demand only grows. One shot chatbots have evolved into reasoning agentic AI that research, plan and use tools, driving orders of magnitude jump in compute for both training and inference. Agentic AI is reaching maturity and has opened the enterprise market to build domain and company-specific AI agents for enterprise workflows, products and services.

The age of physical AI has arrived, unlocking entirely new industries in robotics, industrial automation. Every industrial company will need to build 2 factories: 1 to build the machines and another to build their robotic AI.

This quarter, NVIDIA reached record revenues and an extraordinary milestone in our journey. The opportunity ahead is immense. A new industrial revolution has started. The AI race is on. Thanks for joining us today, and I look forward to addressing you next week -- next earnings call. Thank you.