Speaker 3
0s

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. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question and answer session. If you would like to ask a question during this time, simply press star, followed by the number one on your telephone keypad. If you would like to withdraw your question, press star one again. Thank you. Toshiya Hari, you may begin your conference.

Speaker 2
37s

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.

Speaker 2
1m 56s

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.

Speaker 5
2m 12s

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 to $4 trillion in AI infrastructure spend by the end of the decade. The scale and scope of these buildouts 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.

Speaker 5
4m 17s

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 third 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.

Speaker 5
6m 11s

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. U.S.G. officials have expressed an expectation that the U.S.G. will receive 15% of the revenue generated from licensed H20 sales, but to date, the U.S.G. 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. If we add 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 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.

Speaker 5
7m 55s

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 two 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 buildouts 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 address this by redefining the economics of inference. New NV FP4 4-bit precision and NVLink 72 on the GB300 platform deliver a 50x increase in energy efficiency per token compared to Hopper, enabling companies to monetize their compute at unprecedented scale.

Speaker 5
9m 48s

For instance, a $3 million investment in GB200 infrastructure can generate $30 million in token revenue, a 10x return. NVIDIA's 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 pre-training. Using NV FP4 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.

Speaker 5
11m 53s

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's 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 multi-billion 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 UK and Europe. The European Union plans to invest €20 billion to establish 20 AI factories across France, Germany, Italy, and Spain, including five gigafactories to increase its AI compute infrastructure by tenfold. In the UK, the Isambard AI supercomputer powered by NVIDIA was unveiled as the country's most powerful AI system, delivering 21 exaflops of AI performance to accelerate breakthroughs in fields of drug discovery and climate modeling.

Speaker 5
13m 53s

We are on track to achieve over $20 billion in sovereign AI revenue this year, more than double that of 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 Hotchips, we introduced Spectrum XGS Ethernet, a technology designed to unify disparate data centers into gigascale AI superfactories. 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 Fugaku Next will integrate Fujitsu's CPUs with our architecture via NVLink Fusion.

Speaker 5
16m 2s

It will run a range of workloads, including AI, supercomputing, and quantum computing. Fugaku Next joins a rapidly expanding list of leading quantum supercomputing and research centers running on NVIDIA's CUDA-Q quantum platform, including Ulric, AIST, NNF, and NERSC, supported by over 300 ecosystem partners, including AWS, Google Quantum AI, Quantinuum, QEra, and Seeq Quantum. Just in, 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 a rapid rate. Over 2 million developers and 1,000 plus hardware, software applications, and sensor partners are 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.

Speaker 5
18m 2s

Leading European robotics companies, including Agile Robots, NeuroRobotics, 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 % 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, as strong sales continued as we increased 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+ 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.

Speaker 5
19m 50s

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 Windows devices. With the RTX platform stack, Windows 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, which is now in production, opening up billions to new revenue opportunities for NVIDIA while improving vehicle safety and autonomy.

Speaker 5
21m 44s

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 prioritized 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. OK, let me turn 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% and 73.5% respectively, plus or minus 50 basis points.

Speaker 5
23m 30s

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 non-marketable 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 28th. 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.

Speaker 5
25m 4s

We will now open the call for questions. Operator, would you please poll for questions?

Speaker 3
25m 12s

Thank you. At this time, I would like to remind everyone, in order to ask a question, press star, then the number one on your telephone keypad. We'll pause for just a moment to compile the Q&A roster. As a reminder, please limit yourself to one question. Thank you. Your first question comes from C.J. Muse with Cantor Fitzgerald. Your line is open. Yes, good afternoon. Thank you for taking the question. 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 ramping the second half. Obviously, many of these investments are multi-year projects contingent upon power, cooling, et cetera. I was hoping perhaps you could take a high-level view and speak to your vision for growth into 2026. As part of that, if you can kind of comment between networking data center, it would be very helpful. Thank you. Yeah.

Speaker 1
26m 7s

Thanks, C.J. At the highest level, the growth drivers would be the evolution, the introduction, if you will, of reasoning agentic AI. You know, 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. It is called long thinking. The longer it thinks, oftentimes it produces better answers. The amount of computation necessary for one shot versus reasoning agentic AI models could be 100 times, 1,000 times, and potentially even more as the amount of research and basically reading and comprehension that it goes off to do increases. The amount of computation that has resulted in agentic AI has grown tremendously. Of course, the effectiveness has also grown tremendously. Because of agentic AI, the amount of hallucination has dropped significantly. You can now use it. 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. Over the last year, AI has made tremendous progress. Agentic systems, reasoning systems, are completely revolutionary.

Speaker 1
27m 47s

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. 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. Over the next couple of years, you're going to, you asked about longer term. Over the next five 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 four CSPs by double and grown to about $600 billion. We are in the beginning of this buildout. The AI technology advances have really enabled AI to be able to adopt and solve problems to many different industries.

Speaker 3
29m 25s

Your next question comes from Vivek Arya with Bank of America Securities. Your line is open. Thanks for taking my questions. Colette, just wanted to clarify the $2 to $5 billion in China, what needs to happen, and what is the sustainable pace of that China business as you get into Q4? Jensen, for you on the competitive landscape, several of your large customers already have or are planning many ASIC projects. I think one 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 their use of merchant silicon and ASICs? Thank you.

Speaker 5
30m 18s

Thanks, Vivek. Let me first answer your question regarding what it will take for the H20s to be shipped. There is interest in our H20s. There is the initial set of licenses that we received. Additionally, we do have a supply that we are ready. That is why we communicated that somewhere in the range of about $2 billion to $5 billion this quarter, we could potentially ship. We are 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. It is still open at this time, and we are not exactly sure what that full amount will be this quarter. However, if more interest arrives, more licenses arrive, we can also still build additional H20 and ship more as well.

Speaker 1
31m 17s

NVIDIA builds very different things than ASICs. Let's talk about ASICs first. A lot of projects are started. Many startup companies are created. Very few products go into production. 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. AI factories, in the last several years, have become so much more complex because the scale of the problems has grown so significantly. It is really the ultimate, the most extreme computer science problem the world's ever seen, obviously. The stack is complicated. The models are changing incredibly fast from generative based on autoregressive to generative based on diffusion to mixed models to multimodality. 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. 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.

Speaker 1
32m 56s

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 pre-training to post-training with reinforcement learning, all the way out to inference. When you build a data center with NVIDIA platform in it, the utility of it is best. The lifetime usefulness is much, much longer. In addition to all of that, it is 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. 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 KV caching necessary for agentic AI to the GPU to a SuperNIC to a scale-up switch we call NVLink, completely revolutionary when 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 superfactories with multiple gigawatts of computing all connected together. We call that Spectrum XGS. We just announced that at Hotchips this week. The complications, the complexity of everything that we do is really quite extraordinary.

Speaker 1
34m 47s

It's just done at a really, really extreme scale now. Lastly, if I could just say one more thing, you know, 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. In a world of power-limited data centers, perf per watt drives directly to revenues. You've heard me say before that in a lot of ways, the more you buy, the more you grow. Because our perf per dollar, the performance per dollar is so incredible, you also have extremely great margins. The growth opportunity with NVIDIA's architecture and the gross margins opportunity with NVIDIA's architecture is absolutely the best. There are a lot of reasons why NVIDIA is chosen by every cloud and every startup and every computer company. We're really a holistic full stack solution for AI factories.

Speaker 3
35m 59s

Your next question comes from Ben Reitzes with Barclays. Your line is open. Hey, thanks a lot. Jensen, I wanted to ask you about your $3 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 to $4 trillion would imply maybe $2 billion plus in compute spend. I just wanted to know if that was right and that's what you're seeing by the end of the decade. I'm wondering what you think your share will be of that. Your share right now of total infrastructure compute-wise is very high. I wanted to see if there's any bottlenecks you're concerned about, like power, to get to the $3 to $4 trillion. Thanks a lot.

Speaker 1
36m 52s

Yeah, thanks. As you know, the CapEx of just the top four hyperscalers has doubled in two 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 are five years between now and the end of the decade. $600 billion only represents the top four hyperscalers. We still have the rest of the enterprise companies building on-prem. You have cloud service providers building around the world. The United States represents about 60% of the world's compute. Over time, you would think that artificial intelligence would reflect GDP scale and growth. It would be, of course, accelerating GDP growth. Our contribution to that is a large part of the AI infrastructure. Out of a gigawatt AI factory, which can go anywhere from 50% to plus or minus 10%, let's say 50% to 60% billion, we represent about 35% plus or minus of that. 35% out of 50% or so billion dollars for a gigawatt data center. What you get for that is not a GPU. I think people, you know, we're famous for building the GPU and inventing the GPU. As you know, over the last decade, we've really transitioned to become an AI infrastructure company.

Speaker 1
38m 46s

It takes six chips, six different types of chips, just to build an AI Rubin AI supercomputer. Just to scale that out to a gigawatt, you have hundreds of thousands of GPU compute nodes and a whole bunch of racks. We're really an AI infrastructure company. 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 infrastructure building limitations. 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. Anyhow, these are the limiters going forward. $3 to $4 trillion is fairly sensible for the next five years.

Speaker 3
40m 26s

This question comes from Joe Moore of Morgan Stanley. Your line is open. Great. Thank you. Congratulations on reopening the China opportunity. Can you talk about the long-term prospects there? You've talked about, I think, half of the AI software world being there. You know, how much can NVIDIA grow in that business? You how important is it that you get the Blackwell architecture ultimately licensed there?

Speaker 1
40m 56s

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. 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. It's fairly important, I think, for the American technology companies to be able to address that market. 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 is excellent. There's a whole bunch of new models that are coming out. They're multimodal. They're great language models. It's really fueled the adoption of AI in enterprises around the world because enterprises want to build their own custom proprietary software stacks. Open source model is really important for enterprise.

Speaker 1
42m 34s

It's really important for SaaS, who also would like to build proprietary systems. It has been really incredible for robotics around the world. Open source is really important. It's important that the American companies are able to address it. 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. As you know, H20 has been approved for companies that are not on the entities list. Many licenses have been approved. I think the opportunity for us to bring Blackwell to the China market is a real possibility. We just have to keep advocating the sensibility 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.

Speaker 3
43m 48s

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

Speaker 1
43m 56s

Yeah, thank you for the question. I want to go back to the Spectrum XGS announcement this week. Thinking about the Ethernet product now pushing over $10 billion of annualized revenue, Jensen, 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? Thank you. We now offer three 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. NVLink 72 is fantastic. We then scale out with networking, which we have two. 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. For supercomputing, for the leading model makers, InfiniBand, Quantum InfiniBand is the unambiguous choice.

Speaker 1
45m 46s

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 our 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. It has the ability to come closer, much, much closer to InfiniBand than anything that's out there. That's what we call Spectrum X Ethernet. Finally, we have Spectrum XGS, a gigascale for connecting multiple data centers, multiple AI factories into a super factory, a gigantic system. 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. You know, 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. The ability to improve the efficiency of that factory by tens of % results in $10 billion, $20 billion worth of effective benefit.

Speaker 1
47m 37s

The networking is a very important part of it. It's the reason why NVIDIA dedicates so much in networking. It's the reason why we purchased Mellanox five and a half years ago. Spectrum X, as we mentioned earlier, is now quite a sizable business, and it's only about a year and a half old. Spectrum X is a home run. All three of them are going to be fantastic: NVLink, scale up; Spectrum X and InfiniBand, scale out; and then Spectrum XGS for scale across.

Speaker 3
48m 15s

Your next question comes from Stacy Raskin with Bernstein Research. Your line is open. Hi, guys. Thanks for taking my question. I have a more tactical question for Colette. On the dive up, you know, 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? It looks like Blackwell was probably $27 billion in the quarter, up from maybe $23 billion last quarter. Hopper is still $6 or $7 billion post the H20. Do you think the Hopper strength continues? How do I think about parsing that $7 billion out across those three different components?

Speaker 5
49m 1s

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. 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. They are HGX systems, and I still believe our Blackwell will be the lion's share of what we're doing on there. We don't have any more specific details in terms of how we'll finish our quarter. You should expect Blackwell, again, to be the driver of the growth.

Speaker 3
50m 1s

Your next question comes from Jim Schneider of Goldman Sachs. Your line is open. Good afternoon. Thanks for taking my question. You've been very clear about the reasoning model opportunity that you see. You've also been relatively clear about the technical specs for Rubin. 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? 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 at Blackwell? Thank you.

Speaker 1
50m 42s

Yeah, thanks. Rubin, we're on an annual cycle. 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, an order of magnitude higher than Hopper. For the same amount of energy, and everybody's data center is energy limited by definition, for any data center that we're 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. 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. We advise our partners, our customers, to pace themselves and to build these data centers on an annual rhythm.

Speaker 1
52m 33s

Rubin is going to have a whole bunch of new ideas. I paused for a second because I've got plenty of time between now and a year from now to tell you about all the break-throughs that Rubin is going to bring. Rubin has a lot of great ideas. I'm anxious to tell you, but I can't right now. I'll save it for GTC to tell you more and more about it. 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. 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.

Speaker 3
53m 43s

Your final question comes from Timothy Arcuri with UBS. Your line is open. Thanks a lot. Jensen, I wanted to ask you just to answer the question you threw at a number. You said 50% CAGR for the AI market. I'm wondering how much visibility 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. Maybe are there any puts and takes to that? Thanks.

Speaker 1
54m 15s

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. We still have a lot of businesses that we're still winning and a lot of startups that are still being created. Don't forget that the number of AI-native startups was $100 billion 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 startups that are generating revenues, last year was $2 billion. This year is $20 billion. Next year being 10 times higher than this year is not inconceivable. The open source models are now opening up large enterprises, SaaS companies, industrial companies, robotics companies to now join the AI revolution, another source of growth. Whether it's AI-natives or enterprise SaaS or industrial AI or startups, we're just seeing just an 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's sold out. H100s are sold out. H200s are sold out. Large CSPs are coming out, renting capacity from other CSPs.

Speaker 1
55m 59s

The AI-native startups are really scrambling to get capacity so that they could train their reasoning models. The demand is really, really high. The long-term outlook between where we are today, CapEx has doubled in two 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. I think the next several years, surely through the decade, we see just really fast growing, really significant growth opportunities ahead. Let me conclude with this. Blackwell is the next generation AI platform the world's 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 six new chips that represent the Rubin platform. They have all taped out to TSMC. Rubin will be our third generation NVLink rack scale AI supercomputer. We expect to have a much more mature and fully scaled-up supply chain.

Speaker 1
57m 44s

Blackwell and Rubin AI factory platforms will be scaling into the $3 to $4 trillion global AI factory buildout 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 hundred megawatt facilities. Soon we'll be building millions of Rubin GPU platforms powering multi-gigawatt, multi-site AI superfactories. 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 and industrial automation. Every industrial company will need to build two factories, one 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. I look forward to addressing you next week, next earnings call.

Speaker 1
59m 41s

Thank you.

Speaker 3
59m 45s

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