

# NVIDIA CORP. R (NVD.HA) Q2 FY2027 earnings call transcript

Powered by Quartr

Aug 26, 2026, 11:00 PM GMT+2



## Earnings calls summary

Powered by Yahoo Scout ⓘ

NVIDIA delivered a record Q2 with revenue of \$96 billion, more than doubling year-over-year, driven by surging AI demand and continued data center strength. Management guides for ~70% revenue growth in fiscal 2028, citing ongoing supply constraints despite demand exceeding this level, and expects gross margins to moderate due to memory cost pressures.

### Operator

0:00:00

Good afternoon. My name is Tiffany, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's second 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. 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.

### Toshiya Hari

VP of Investor Relations and Strategic Finance

0:00:38

Thank you. Good afternoon, and welcome to NVIDIA's conference call for the second quarter of fiscal 2027. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer, and Colette Kress, Executive Vice President and Chief Financial Officer. 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 2027. The content of today's call is NVIDIA's property. It cannot 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.

### Toshiya Hari

VP of Investor Relations and Strategic Finance

0:01:26

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 26th, 2026, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements. During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website. With that, let me turn the call over to Colette.

**Colette Kress**

EVP and CFO

0:02:10

Thanks, Toshiya. We delivered another outstanding quarter with record revenue, operating income, and EPS. Total revenue of \$96 billion more than doubled year-over-year as growth accelerated for the fourth consecutive quarter. The surge in AI demand is driving a global infrastructure build-out, supported by an expanding and diverse set of growth opportunities, spanning hyperscalers, AI labs, AI natives, enterprises, and sovereign customers. We expect to grow revenue by approximately 70% in fiscal 2028. This is a supply-constrained outlook. Q2 data center revenue increased 18% quarter-over-quarter to \$89 billion, with strong contributions from both sub-segments, hyperscale and ACIE, which includes our neocloud, industrial, and enterprise customers. Hyperscale revenue of \$49 billion grew 13% sequentially, driven by sustained strength in Blackwell.

**Colette Kress**

EVP and CFO

0:03:21

Reinforcing that more compute drives more revenue as new GPU capacity comes online, our hyperscale customers delivered strong financial results in the quarter, with accelerating revenue growth and expanding margins. With cloud industry backlog now greater than \$2 trillion, CapEx by the top five hyperscalers is expected to reach nearly \$800 billion in 2026 and \$1.3 trillion in 2027. Today, we are delighted to announce an expansion of our partnership with AWS. Building on its already vast installed base of NVIDIA Compute, AWS is deploying an additional 2 million GPUs starting this quarter through the second quarter of fiscal 2029, along with Vera CPUs, some integrated with Rubin, others standalone. AWS will serve NVIDIA Nemotron family of open models on Amazon Bedrock and SageMaker. Amazon will also adopt our full physical AI stack, Omniverse, Cosmos, Isaac, and Jetson to power its fleet of warehouse robots.

**Colette Kress**

EVP and CFO

0:04:35

ACIE revenue of \$40 billion increased 25% sequentially and 138% year-over-year. Growth was driven by neocloud capacity additions to meet the rising demand from enterprises, AI startups, and sovereigns, as well as hyperscalers purchasing capacity to supplement their own build-outs. Using NVIDIA DSX reference designs, our neocloud partners are bringing capacity online faster and at lower token cost. They are expected to exit the year with 8 GW in total installed capacity, up from approximately 3 GW at the end of 2025. Incredibly, we are seeing demand acceleration even at our scale. Customers' forecasts point to our growth doubling next year. However, as I mentioned earlier, we expect to grow approximately 70% as we are supply-constrained. NVIDIA Compute is fully utilized across every cloud we serve. The economic value it generates for our hyperscale, neocloud, and AI lab partners keeps rising.

### **Colette Kress**

EVP and CFO

0:05:52

Besides building the best AI computing technologies and the most capable supply chain, NVIDIA has three unique capabilities that are engines powering our growth. First, NVIDIA's architecture runs every model, and we're growing share as closed and open model adoption grow. Closed and open models alike, adoption is skyrocketing. NVIDIA runs the leading closed models, OpenAI, Anthropic, Groq, Meta, Gemini, and the leading open models, TML, Mistral, Qwen, Kimi, GLM, DeepSeek, MiniMax, and Nemotron. We're great at small models and giant ones, large or video, auto, regressive or diffusion, in the cloud or in the edge. NVIDIA is great at training, great at inference, great at agentic workloads. One platform, fungible for every model and workload. Durable for the entire life cycle of AI. That combination of performance, fungibility, and durability is what makes NVIDIA the productive and financiaible compute infrastructure.

### **Colette Kress**

EVP and CFO

0:07:13

Our second unique capability is our full stack AI factory platform that is expanding our share of the data center TAM. Since Hopper, our revenue opportunity has grown from roughly \$18 billion per gigawatt to \$25 billion with Blackwell, to \$40 billion with Vera Rubin, which now spans Vera CPU, Rubin GPU, NVLink, InfiniBand or Ethernet, and Groq LPU, announced earlier this week. Our ability to extreme co-design across GPU, CPU, NVLink scale-up networking, scale-out networking, systems, algorithms, and software enables us to deliver X factor performance gain every generation. Vera Rubin exemplifies this, delivering 30x higher throughput per megawatt and 35x lower token cost relative to Grace Blackwell Ultra. We commenced production shipments of Vera Rubin earlier this month. Having already received purchase orders from every major hyperscaler, AI cloud, and system OEM, we expect Vera Rubin to mark the fastest product ramp in NVIDIA's history.

### **Colette Kress**

EVP and CFO

0:08:39

Our networking business had another record quarter, with revenue growing 18% on a sequential basis. Spectrum-X Ethernet, which grew 2.6x on a year-over-year basis, is already helping us become the largest and fastest-growing network company in the world. Rising adoption of agentic AI is driving an acceleration in demand for data center CPUs. Our Grace CPU, introduced in 2021, has been a great success, with revenue on a trailing 12-month basis exceeding \$5 billion. Today, we are in full production of our next generation Vera CPU. As a standalone product, Vera expands our TAM even further. Vera completes agentic tasks 1.8x faster on the spec benchmark and provides five times the bandwidth per watt than any other data center CPU.

**Colette Kress**

EVP and CFO

0:09:43

We expect Vera to be deployed by every major hyperscaler, neocloud, AI lab, and system OEM, with shipments already underway to our lead partners, including OCI, SpaceXAI, and starting this quarter, AWS. We continue to see demand for approximately \$20 billion in total server CPUs. Based on our customer demand and improving supply outlook, our preliminary expectation is for CPU revenue to more than double in fiscal 2028, positioning us as one of the world's leading server CPU suppliers. Since the announcement of our Groq partnership last year, we've been working to unite NVIDIA's high throughput and Groq's high interactivity architectures. At Hot Chips earlier this week, we announced that Groq 3 LPX, our first rack-scale LPU system, is in full production and already setting records, demonstrating nearly 4x the number of tokens per second against the next best alternative on our Artificial Analysis benchmark.

**Colette Kress**

EVP and CFO

0:11:01

We expect to ship Groq 3 LPX in volume later this quarter to early adopters. Nebius will be the first. Today, we're not just selling the best chips, we're selling a full stack AI factory platform, offering superior economics for customers and capturing a bigger share of the data center TAM. Our third unique capability is the combination of our full stack AI factory and rich CUDA ecosystem, allowing us to extend AI into markets a single chip alone can never reach. Beyond the hyperscalers lies a massive market anxious to adopt AI, customers with no interest in designing their own custom silicon.

**Colette Kress**

EVP and CFO

0:11:49

NVIDIA's fully proven full stack platform is uniquely suited to help sovereigns, neoclouds, and enterprises build their AI infrastructure, bring it to full operation, continuously optimize it through CUDA software, and connect it to offtake demand from our vast developer ecosystem.

Hyperscalers will remain a major growth driver, but non-hyperscaler growth, our ACIE segment spanning sovereign regional neoclouds, enterprise edge, and air gap data centers will represent roughly half of our data center business. Our AI native startup ecosystem, developed and running primarily on the NVIDIA compute platform, is scaling at a rapid pace. Global VC funding in AI, roughly 70% of which is spent on compute, exceeded \$400 billion in the first half of 2026, surpassing the \$265 billion raised in all of 2025.

**Colette Kress**

EVP and CFO

0:12:57

Nearly 20 companies, including Cursor, owned by SpaceX, Figma, and Together AI, now exceed \$1 billion in annualized run rate revenue, up from 13 companies in Q4 of last year, with vertical enterprise software logging the fastest growth. In enterprise, on a trailing 12-month basis, on-prem revenue in the automotive vertical reached \$8 billion, while financial services, manufacturing, and healthcare combined contributed \$7 billion in revenue. Hudson River Trading and Jane Street are leveraging NVIDIA-powered AI factories to accelerate quantitative trading. Samsung Electronics is using NVIDIA cuLitho to achieve up to 20x greater performance in computational lithography, while Bristol Myers Squibb is investing in Vera Rubin AI factory, a fast follow to the Roche and Lilly build-outs as drug R&D timelines compress from years to months. In sovereign AI, our business, primarily through the regional neoclouds, grew 35% sequentially and more than tripled year-over-year in Q2.

**Colette Kress**

EVP and CFO

0:14:15

A country or region can allocate land and power directly to a regional cloud partner in ways it never would to a foreign hyperscaler. We don't own a cloud ourselves. We are a neutral partner to every sovereign and neocloud. Because NVIDIA Compute is productive, fungible, rentable, and durable, regional cloud interest is surging around the world. We helped CoreWeave, Nebius, and Nscale build entire infrastructure businesses, and neoclouds are emerging everywhere. Firebird in Armenia, Cassava Technologies across Africa, GMI Cloud in Taiwan, Yotta and Neysa in India, Firmus in Australia, YTL AI Cloud in Malaysia, pairing local land, power, and operating expertise with our platform. Last month, we announced a partnership with Noetra, Japan's national AI company, to build an NVIDIA DSX AI factory that will create open models to power AI agents, digital twins, robotics, and physical AI applications.

**Colette Kress**

EVP and CFO

0:15:34

South Korea's LG and Hyundai Motor Group are partnering with NVIDIA to build and scale AI. In Europe, a record 35 new NVIDIA-powered AI supercomputers were unveiled to advance industry and scientific breakthroughs. Neoclouds are seeing strong demand pipelines for many diverse

offtakers. Rather than allocating their entire capacity to a single long-term offtake guarantee that lenders typically require to finance a data center independently, we have introduced a revenue-sharing structure. NVIDIA provides a take or pay commitment on a portion of the facility's capacity, a minimum revenue guarantee that gives lenders the confidence to underwrite the project, and in exchange, we share in a portion of the neocloud's revenue earned above that floor. Independent capital still underwrites every deal on its own merits. We're not making loans.

**Colette Kress**

EVP and CFO

0:16:36

In this model, we get paid twice, once on the hardware sale and again through the share of rental revenue, a highly reoccurring stream layered on top of a one-time equipment purchase. Over time, this model can expand our addressable market and create reoccurring usage-linked revenue stream alongside our core platform revenue, with the potential to drive billions in revenue over the medium to long term. Together, NVIDIA's three unique capabilities, a platform that runs every model, a full stack AI factory platform capturing more of the data center TAM, and a CUDA ecosystem that extends AI into markets no single chip could reach alone, reinforce one another and are the engines of our growth. Let me update you on our progress with our frontier AI labs.

**Colette Kress**

EVP and CFO

0:17:35

The frontier AI labs have extraordinary demand for training and inference compute, but they are growing faster than what their balance sheets and credit profiles can support. They have rapidly growing customer demand, yet still lack the decades-long infrastructure contracts and investment-grade financing capacity needed to secure the AI factory infrastructure independently. In other words, their growth is not limited by their technology or customer demand. It is limited by compute. For these companies, more compute means more intelligence, more users, and more revenue. NVIDIA is needed to help power this flywheel. First, we have invested nearly \$50 billion in the frontier AI labs. This was a meaningful commitment, but it represented a small fraction of our expected free cash flow over the same period.

**Colette Kress**

EVP and CFO

0:18:32

Further, to support the frontier labs infrastructure build-outs, we recently announced partnerships with six of the world's leading infrastructure capital providers, Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs, and KKR, to establish financing platforms that will raise over \$500 billion of third-party capital. With these partnerships, building on our unique, fungible, and durable computing platform, the AI labs will be able to build and assess AI infrastructure funded by long-term institutional capital at relatively attractive rates. Last week, we announced

that we secured land power shell capacity through our partnership with SoftBank Energy to exclusively host NVIDIA Compute at their Portsmouth campus. The initial deployment, expected to support 4.25 GW of AI factory capacity, will be utilized by OpenAI. Each generation of NVIDIA AI factory systems deployed at PORTS-Pike could represent approximately 1.5 million NVIDIA GPUs, and over 20 years, the site could support multiple upgrade cycles.

**Colette Kress**

EVP and CFO

0:19:49

Here is the essential economic point. The LPS commitment secures a long-lived AI factory site, while the NVIDIA Compute within the data center can be upgraded repeatedly. This project deepens our longstanding partnership with OpenAI. OpenAI has committed to substantial deployments of NVIDIA AI infrastructure through 2030. OpenAI's existing and planned commitments represent approximately 12 GW of NVIDIA Compute. For another frontier AI lab, we will provide selective credit enhancement for nearly 2 GW of compute. This complements the substantial NVIDIA Compute capacity they have secured independently without NVIDIA's credit support. We recognize the scale of this support, and we know some will call this circular financing. We see it differently. We are going through a major computing platform shift, the creation of one of the most important technologies in human history, and these are once-in-a-generation companies. Their technology leadership is proven, and their customer traction and usage are skyrocketing.

**Colette Kress**

EVP and CFO

0:21:08

We expect them to become the largest technology companies in history. We believe these investments, measured against the strength of their demand, the business they create for us, the ecosystem they build on NVIDIA's platform, and the equity returns on our invested capital will be excellent, and our risk is limited. The NVIDIA Compute platform is fungible and durable and can be redeployed to support other customers. For context, we expect demand from the AI labs for which we expect to leverage our balance sheet to contribute toward roughly a quarter of our business next year. This remains compute we ship will be consumed by investment-grade customers or those that are backed by one. In Q2, we shipped less than 1% of our total data center revenue in Hopper 200 products to customers based in China in accordance with the U.S. government licenses.

**Colette Kress**

EVP and CFO

0:22:14

Current Hopper shipments are dilutive to corporate gross margins, and given ongoing geopolitical uncertainty, there is no China data center compute revenue in our forward outlook. Moving to the rest of the P&L. GAAP and non-GAAP gross margins were both 75%, largely unchanged from last

quarter due to a similar product mix. GAAP and non-GAAP operating expenses were up 10% and 11% sequentially, primarily due to high compute infrastructure costs and compensation and benefits costs. Our non-GAAP effective tax rate of 16% increased from a year ago, primarily due to higher revenue. On our balance sheet, inventory increased to \$32 billion as we prepared for the Vera Rubin launch. Days of sales outstanding increased to 60 days, reflecting extended payment terms for large purchases by certain investment-grade customers to be shipped over multiple quarters.

**Colette Kress**

EVP and CFO

0:23:21

In Q2, we returned a record \$26 billion to shareholders, \$20 billion through share repurchases, and \$6 billion through our quarterly dividend of \$0.25 per share. Relative to our plan to return 50% or more of free cash flow, we have returned 60% on a year-to-date basis. Going forward, we intend to increase and return excess free cash flow net of strategic uses. Let me turn to the outlook for the third quarter. Total revenue is expected to be \$108 billion  $\pm 2\%$ . We expect sequential growth to be driven primarily by ACIE with data center, while growth in hyperscale is expected to re-accelerate in Q4 and into fiscal year 2028 as supply of Vera Rubin grows over time. We see Vera Rubin accounting for about 20% of data center revenue in Q3.

**Colette Kress**

EVP and CFO

0:24:25

Looking ahead, our preliminary expectation is for fiscal year 2028 revenue to grow approximately 70% year-over-year. Although we will work to close the supply-demand gap, we expect supply to remain a bottleneck at least through the end of fiscal year 2028. Many of you have expressed concerns regarding our gross margins, as component costs have risen significantly. As you are already aware, we are experiencing extreme pricing conditions in memory. The magnitude of the price increase has exceeded our prior expectations and are headed even higher into next year. As a result, we are resetting expectations today. For Q3, we expect GAAP and non-GAAP gross margins to be 74%  $\pm 50$  basis points. We expect margins to bottom in Q4 in the 71%-72% range before settling at 72%-73% in fiscal year 2028 as executed price increases take effect in Q1.

**Colette Kress**

EVP and CFO

0:25:40

We want to be direct about this rather than let it linger as an open question. Memory scarcity today is being driven in large part by the AI build-out itself, and unlike a component that simply raises our cost with no offset benefit. Tighter memory supply is a symptom of the same demand surge that is driving our own growth. We have longstanding, deep relationships with all three major memory suppliers, and we are working closely with them to further increase the capacity

our roadmap requires. GAAP and non-GAAP operating expenses are expected to be approximately \$9.2 billion and \$9.0 billion respectively.

**Colette Kress**

EVP and CFO

0:26:25

For the full year, we now expect OpEx to grow in the low 50s, driven by a broadening of our product portfolio and further increase in the usage of AI tools, which has already and will continue to enhance engineering productivity. For full year fiscal year 2027, we continue to expect GAAP and non-GAAP tax rates to be between 16% and 18%, excluding any discrete items and material changes to our tax environment. With that, we will now transition to Q&A. Operator, please poll for questions.

**Operator**

0:27:06

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 will pause for just a moment to compile the Q&A roster. Your first question comes from the line of Joseph Moore with Morgan Stanley. Your line is open.

**Joseph Moore**

Analyst

0:27:28

Great. Thank you. I wonder if you could give us color on the 70%, and what gives you the confidence to guide a full year out if you haven't been doing that? What is the key constraint that separates those numbers, and could you close those gaps over time?

**Jensen Huang**

President and CEO

0:27:50

Yeah. Thanks, Joe. As you probably are aware, AI has become useful. The AI agents that are being adopted everywhere use an enormous amount of compute. First of all, the large language models are larger than ever because they are smarter than ever. These agents go through reasoning and planning multiple turns of tool use. The amount of compute necessary for an agent versus a human using it is probably 15- 100 times, depending on the type of problem you are trying to solve. The amount of compute necessary is just extraordinary. That is a factor that almost everybody sees. The part that people don't see about our growth, because we are practically singular, because the nature of how we deliver products. We are the only company in the world that creates and builds, offers an entire AI factory platform, a full stack system.

**Jensen Huang**  
President and CEO  
0:28:52

Customers can still mix and match. However, most companies just don't have the skills to do that or a desire to do that. There is an entire part of the market that we experience growth. There is sovereign AI, there are regional AIs, there are neoclouds, there are AI startups at enterprises where we are seeing, which represents about half of our business, and that is growing 100% a year. That part of the world's computing is likely to be larger over time than even what we are currently experiencing in the cloud. I think the demand that we see is driven by all of those factors. It is also the case that you can no longer procure technology per se and stand up this infrastructure. You have got to go secure the Land, Power, and Shell, which oftentimes is a couple, two, three years out.

**Jensen Huang**  
President and CEO  
0:29:47

All of the rest of the supply chain necessary to align the construction, the power, the cooling, all of the labor that is necessary. AI infrastructure is creating so many jobs all over the United States and all around the world. It just takes a lot more planning. We are involved in securing infrastructure now, further down the pipeline. Just as a long time ago, people asked me why is it that we are working with memory suppliers when we are a chip company? Today, people understand it is really quite genius that we were working on our supply chain so far upstream. We work with power generator companies downstream. We work with Land, Power, and Shell companies all around the world.

**Jensen Huang**  
President and CEO  
0:30:35

That helps prepare all of this computing that is going to be built, that will ultimately deploy for our ecosystem and our customers. We just have a lot greater visibility now, upstream and downstream. It is the case that we have never forecasted or never guided to a year in advance. Even though our demand is much greater than 70%, our supply allows us to confidently deliver 70%. We are going to continue to work with our supply chain to increase on that.

**Jensen Huang**  
President and CEO  
0:31:10

What we wanted to do is to be consistent with everybody from our customers, our shareholders, our supply chain. Everybody sees the same view. The reason why that is important is because everybody is putting a lot of resources at play. We wanted to make sure that everybody has the

same set of information. We have got a huge year coming up next year, and it is going to be pretty extraordinary.

### **Operator**

0:31:41

Your next question comes from the line of C.J. Muse with Cantor Fitzgerald. Your line is open.

### **C.J. Muse**

Analyst

0:31:49

Yeah, good afternoon. Thank you for taking the question. There is tremendous investor focus on your inference market share. Can you speak to the evolving workloads you are seeing with agentic AI and how you see your share evolving here over time, particularly when you reflect on the growing value of the TAM you are seeing with each new full stack generation, your expectation for greater growth from ACIE, and then also including Groq 3 LPX? We would love to hear your thoughts.

### **Jensen Huang**

President and CEO

0:32:15

Yeah, thanks, C.J. The AI life cycle is getting way more complex than it used to be. It is playing into NVIDIA's architecture much more greatly than it used to be. You could kind of see it as four phases. There is the first phase, which is preparing all of the data that you need. Some of it is synthetic, some of it is real, some of it is human labeled and generated, pre-train the models. Then there is post-training, the third phase, and then there is the agentic inference. Agentic inference is extremely complicated. Every one of those phases are complicated. The thing that is really great about the NVIDIA architecture, and we created this with NVL72, it was a big surprise on the world when we first created the world's first rack scale architecture.

### **Jensen Huang**

President and CEO

0:33:11

It was hardly easy, and it was very challenging building the first generation. We are now in our third generation of NVL72 rack scale systems. We had to reinvent the entire supply chain, reinvent systems, reinvent the technology, redistribute our software, refactor our software, everything. Every aspect of it was hard. What it allowed us to do was to create one fungible system that allows us to transition from data creation, data preparation, to pre-training, to post-training, to agentic inference. The benefits to customers is incredible. The reason for that is because you have just spent, and we just mentioned, each gigawatt of technology and NVIDIA's revenue exposure in the Hopper timeframe with Hopper plus InfiniBand, and now Vera Rubin and

CPU and three types of different networking, because it takes that many types of networking to address the entire world's data center.

**Jensen Huang**

President and CEO

0:34:13

Not to mention the scale-in security networking and the scale across multi-campus networking. You could argue five different types of networking systems. Then, of course, Groq. All of that increased our revenue contribution or revenue opportunity per gigawatt to \$40 billion. Each gigawatt of data center increased from, say, \$30 billion about five years ago to now \$60 billion today. Of course, the productivity's tremendous. The performance is incredible in comparison. But you are talking about a \$60 billion investment. To the extent that you could use it across multiple phases of the AI life cycle, run every single type of model you can imagine running on it, whether it is diffusion or autoregressive or state space or some hybrid version of that, every version of attention mechanism you can think of, small or large models.

**Jensen Huang**

President and CEO

0:35:20

The investment that you make will be preserved and useful and productive for a lot longer time. I think our advantage in this new world is really quite extraordinary, and it could explain why it is that our growth is actually accelerating. It was already large, but now it is accelerating. Let us see. You asked about Groq. Super excited about Groq 3. We achieved a record token interactivity rate, extremely low latency, performance generation. The team is doing fantastically. We spent the last several months fusing the NVLink architecture, which will be the core engine.

**Jensen Huang**

President and CEO

0:36:13

Then for services that would like to have super high interactivity, super high speed token generation done, the throughput is going to be a lot lower. The cost per token will be higher. But you could associate it with high ASP services. For those companies, you could bolt on one of our Groq accelerators. I am super excited about that. But the vast majority of the world's data centers will just be Vera Rubin NVL72.

**Operator**

0:36:48

Your next question comes from the line of Stacy Rasgon with Bernstein Research. Your line is open.

**Stacy Rasgon**

Analyst

0:36:55

Hi, guys. Thanks for taking my question. The 70% growth in fiscal 2028, which I guess is sort of calendar 2027, that is something like, I don't know, a \$200 billion uptick versus the prior outlook if I back it out. The prior outlook was \$1 trillion over the three years, so this is probably \$200 billion more. I was just wondering if you could talk us through the contributors to that increase across the different products, Vera and CPUs and Groq and everything else. You talked about your price increase that takes effect in Q1, so I assume some of this is pricing. I guess I am also curious, maybe I am squeezing too many questions in here, but I am also curious because it is a constrained number. What would it be if it was not constrained?

**Jensen Huang**

President and CEO

0:37:42

The unconstrained would be a lot higher. We grew 100% year-over-year this year. The unconstrained is significant. We are just going to have to go work hard to get more capacity. We have a large supply chain. We have a really gigantic supply chain. We have incredible partners. We have secured a lot of supply, but we just need a lot more. To break it down, the way to think about that is most people see just hyperscalers. That is half of the picture. The other half of the picture is what we call ACIE, and that is all the enterprise, the neoclouds, the sovereign AIs. That part of the world is invisible to everybody. The reason for that is because they do not buy custom chips. They do not buy chips one at a time.

**Jensen Huang**

President and CEO

0:38:44

They really need an entire factory platform built for them. That is a space that we add just a tremendous amount of value. Now, of course, back in this hyperscale space, that is growing incredibly too, right? You know that they now have backlogs of \$2 trillion. You know that when they stand up NVIDIA Compute, when that happens, their revenues go up, their earnings contribution go up. Compute is profitable, very profitable today, and Compute directly translates into increased revenues. There is just a race to want to bring more NVIDIA Compute online, both at the hyperscalers. What you do not see is just really tremendous opportunities outside of the hyperscalers. The other part of it is, and this is the reason why we mapped it out for you. In the case of Hopper, we were at about \$18 billion per gigawatt.

**Jensen Huang**

President and CEO

0:39:43

For Grace Blackwell, we are about \$25 billion per gigawatt. For Vera Rubin, it is about \$40 billion per gigawatt. The productivity is X factors increase in each generation. Customers want to race to the next generation as fast as they can. Meanwhile, because NVIDIA's compute is so productive, the tokens they are generating, the GPU hours they are renting out is insanely profitable as you know. Their margins are fantastic. All of that is just simultaneously happening. I think the big picture is that we are going through this platform shift, and it affects every computer company.

**Jensen Huang**

President and CEO

0:40:31

Every industry in the world uses computers, so therefore every industry is affected, every company is affected. This new way of doing computing is intelligent. It is not based on retrieval of files, but it is now generative, generating intelligence, and that requires compute. The results you get is phenomenal. The results you get is tremendously better. We are just seeing that across the world. Everybody wants to be part of the AI revolution. Everybody will have to be part of this computing shift, and everybody has to build infrastructure.

**Operator**

0:41:12

Your next question comes from the line of Vivek Arya with Bank of America Securities. Your line is open.

**Vivek Arya**

Analyst

0:41:20

Thanks for taking my question. Thanks for providing all the transparency and all the commitments and guarantees that you have for a number of years. When I just add up everything that is in the CFO commentary, I get to a number of about \$500 billion or so, obviously over the next several years. So I had a few questions related to that. First is that the takeaway? That the sum of all your ecosystem investments over the next several years is in that ballpark, or are there other equity or other investments that could still be ahead? That is one.

**Vivek Arya**

Analyst

0:41:52

Secondly, if there is a specific cash part of that we should think about in fiscal 2028. Then, Jensen, a lot of these investments are designed to help the frontier labs, especially OpenAI and Anthropic, but both of them are designing their own custom chips. In fact, OpenAI just in the last few days spoke about Jalapeño and their claims about being better than Blackwell and so forth.

So how are you balancing this dynamic where you want to invest a lot in the ecosystem, but part of that ecosystem wants to develop competitive solutions? Thank you.

**Jensen Huang**

President and CEO

0:42:28

Well, we're building something very different. Whereas many of these XPU's are inference-specific chips for one cloud or one service, NVIDIA is a platform, an entire AI factory platform that spans the entire AI life cycle that you can use in any cloud. It's in every cloud. You can run anywhere. We'll help you set it up anywhere. We built something very different. All of the AI services, at some point, are going to want to go around the world, and those data centers won't necessarily be just built by them. They're going to run, and I think they're going to run on NVIDIA all around the world.

**Jensen Huang**

President and CEO

0:43:34

And of course, I think our technology, I have 100% confidence that our technology will continue to be extraordinary for them and that the economics of using our technology, whether it's from data processing to training, to post-training, to agentic processing, our technology's going to be extraordinary for them. They're going to use it. I'm very confident that they're going to be customers and partners of ours for a very long time. Now, having said that, taking a step backwards, investing in these two companies, or there are several AI labs that we've invested in, investing in these companies are a once-in-a-generation opportunity. I think the only regret that I have is that I didn't invest more and sooner. Two of the companies will likely go public soon, and others will follow. These will be some of the most consequential technology companies in history.

**Jensen Huang**

President and CEO

0:44:30

I'm delighted to be their friend. I'm delighted to partner with them. I'm delighted that they're building an ecosystem on top of the NVIDIA architecture. I'm delighted that they're counting on us to scale up, and I have 100% confidence that through quite a long period of time, they're going to be utilizing NVIDIA Compute for a lot of their computing. I feel great about it.

**Colette Kress**

EVP and CFO

0:44:59

So Vivek, let me add a little bit more regarding the commitments and the portion within those commitments, which is our supply commitments. This is essential. This is essential for the raising

of Vera Rubin today, as well as all next year. You can see that those commitments, the biggest parts of them are in the first three years, and we will use that to build the products that we need. That's what also gives us the confidence in terms of our growth and revenue, given how much we have already aligned in commitment in terms of our supply as well as capacity that we would need.

**Operator**

0:45:39

Your next question comes from the line of Timothy Arcuri with UBS. Your line is open.

**Timothy Arcuri**

Analyst

0:45:46

Hi. Thanks a lot. Jensen, I wanted to ask about open source. There is a lot of talk that these models could gain share for workload in the U.S. You are obviously well-positioned with Nemotron, but on the other hand, a lot of the end demand is being driven by these big frontier model companies. There is a lot of investors that equate open models as being negative for the growth of those companies. How do you put and take that? Do you see the rise of open models as being good for NVIDIA or ultimately negative? Thanks.

**Jensen Huang**

President and CEO

0:46:17

The world will need both closed models and open models. Both closed models and open models are skyrocketing in use. I would say nearly all open models run on NVIDIA, and the reason for that is because NVIDIA's footprint around the world is the highest, and our architecture is the most fungible. It is everywhere. It is in PCs and edge devices like DGX Spark, which is doing great, all the way to robots and workstations and your on-prem data centers. Open models are doing incredibly well. Closed models we know are doing incredibly well. The frontier labs, their sales are skyrocketing. Their margins are fantastic. They are generating profitable tokens. They are only limited by the amount of compute. That is equally true for open models. Our position in open models is very good because the CUDA ecosystem is literally everywhere.

**Jensen Huang**

President and CEO

0:47:31

The open models are also foundational to just about every AI startup and every enterprise company around the world. It is vital to them. The reason for that is because you should rent intelligence, strong intelligence, smart intelligence wherever you can, which is the reason why we rent it, and I encourage my employees to use the cloud service as much as they can. But every

major company and surely every country and every startup needs to build their domain-specific, their proprietary AI, their proprietary alpha. The open models reaching frontier levels has made it possible, has enabled them to all do that. One of the areas where frontier models is vital is cybersecurity. You see the number of cybersecurity companies that are enabled by frontier models so that they could have distributed, massively distributed, continuously running autonomous cybersecurity systems to defend. Those companies are emerging.

**Jensen Huang**

President and CEO

0:48:44

They are some amazing companies. They couldn't do it without open models. Open models is both incredibly successful and has finally reached the frontier, but it's also vital to the American economy, it's vital to the world economy, it's vital to companies to build their own proprietary AI. You can't do it without one or the other. Both are going to be extraordinarily successful. Lastly, as you know, our market footprint of all AI models, I think we're the only platform. I'm fairly certain we're the only platform that runs every frontier model, whether it's closed or open. Most of them were built on NVIDIA, so they run great on NVIDIA. So we're delighted by any model succeeding. So long as models succeed, I'm very happy. Both closed and open models are going to succeed, and they're both simultaneously driving our sales.

**Operator**

0:49:57

Your next question comes from the line of Ben Reitzes with Melius Research. Your line is open.

**Ben Reitzes**

Analyst

0:50:05

Yeah. Hey, thanks. I wanted to ask you a question, Jensen, about demand in a different way. You talked about demand growing 100% next year, and I wanted to get a sense for a couple of things driving that and even beyond that. There's two concepts here. There's recursive self-improvement, which apparently at Anthropic and OpenAI is going very well with AI that improves itself. Even OpenAI said they could hit AGI by the end of this year. With the developments in RSI as well as AGI, what happens to industry demand? Does it inflect further? What does it mean for NVIDIA when those things take place? How are you looking at that as a demand catalyst? Thanks.

**Jensen Huang**

President and CEO

0:50:54

I appreciate that. It's going to inflect further. Today, the vast majority of AI is prompted by people. I believe that this last month it has crossed. Most AI are now agentic. But in the future, every company will have a whole bunch of agents. We have 40,000 employees, roughly. In the future, we'll have 400,000 agents, 4 million agents. Those agents are running continuously. They're running in the background. If you know anybody who builds edge personal AI agents, and they run it on their DGX Spark. I know a lot of people who run it on DGX stations, this incredible workstation that we've built, and you can buy it from Dell, and they're incredible. These AI agents running on a DGX station runs 24/7, because you got stuff for it to do all the time.

**Jensen Huang**

President and CEO

0:51:58

When the world goes to agentic, fully agentic systems, you are going to have agents running all the time, working with other agents running all the time. Those will be working in the background, improving your company, improving your lives. In a lot of ways, we are kind of recursive at this point. You could argue it is coarse-grained, but every time you run through an agent, it reflects on how it could do a better job next time, and it updates the skill file. The skills document, the markdown, is updated at the end of every single one of them. Next time you run it is going to get better. It is a kind of a loosely coarse grain self-improvement. You see that always. In a lot of ways, and for many tasks, we could say that we have already achieved AGI.

**Jensen Huang**

President and CEO

0:52:52

I think all of those milestones and all those, they are kind of senseless at this point. I think the most important thing that matters for the industry is that, one, AI is now doing productive and useful work. Two, AI is generating profitable tokens. Three, if we had more compute, we could generate more profitable tokens, which results in more profit for all of the services. This is the exact phase where we are at, which is the reason why everybody is leaning in.

**Operator**

0:53:25

Your next question comes from the line of Jim Schneider with Goldman Sachs. Your line is open.

**Jim Schneider**

Analyst

0:53:33

Good afternoon. Thank you for taking my question. If you think about the 100% growth you talked about in terms of the plus unconstrained demand growth you are expecting, the 70% you expect to fulfill in terms of supply, can you maybe talk about some of the, or rank order some of the most

acute constraints, whether that be things like data center, power and shell availability, DRAM, wafer foundry availability, et cetera? If you could maybe help us understand which are the biggest among those, that would be very helpful. Thank you.

**Jensen Huang**

President and CEO

0:54:10

There's something funny I could say, but I'm going to just not. Last year, one of the funnest things to do is just to go figure out where I go for dinner and who I have dinner with, and their stock price doubles the next day. I think the answer is, our entire supply chain is challenged. Everybody is really running flat out. More capacity is coming online all the time, which is one of the advantages of what's going to happen this year. It's not going to come online in an instance in time, but it's going to come online every day. Yields are going to get improved. We're going to be doing yield improvement. We're going to work hard on working with every one of our suppliers.

**Jensen Huang**

President and CEO

0:55:05

We have a- It's not even next year yet, so we've got lots and lots of time to work hard every day. So at this moment, we have supply for 70%. We have more supply than 70%, but about 70%. Our demand is much higher than that, and we've got to go work hard, or we're going to be disappointing customers. We like not to disappoint our customers, and we like to work hard for them. So I'm going to need the help of the entire supply chain to help me out here. They all know that. What I'm telling you about our needs for next year is exactly consistent with what I've told them. Everybody's on the exact same song sheet, and I'm trying to be as transparent as we can because we're talking about big numbers.

**Operator**

0:56:04

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

**Aaron Rakers**

Analyst

0:56:12

Yeah. Thanks for taking the question. I want to go back to the gigawatts, the \$25 billion-\$40 billion, and maybe try and understand. I think, Jensen, you said at some recent conferences that's going to further scale. So, as we think about the path even beyond Vera Rubin, we think about Vera Rubin Ultra and so on and so forth. Should we really conceptualize \$40 billion goes to \$60 billion, \$80 billion? I guess underneath of that question is how do we think about your ability

to scale the capacity deployments? Is it a linear function, or is there something that unlocks your ability to supply more demand as we look through fiscal 2027?

**Jensen Huang**

President and CEO

0:56:54

Great question.

**Aaron Rakers**

Analyst

0:56:55

Or 2028, sorry.

**Jensen Huang**

President and CEO

0:56:56

Yeah, great question. Great question. It is very simple. Is our goal to put as much compute on a plot of land? Is our goal to put more compute into 1 GW or less? Obviously, we would like the speed of light answer. The perfect answer is actually infinity per gigawatt. If we could literally get \$1 trillion of compute into 1 GW and one piece of Land, Power, and Shell, it would be a fantastic outcome. The answer is directionally in that direction. We started in the world of general purpose computing during Moore's Law. We were probably, pick your favorite number, but I am going to go with something like \$5 billion, \$3 billion per gigawatt of compute with general purpose computing. Then eventually with Hopper, it was \$18 billion. Now Grace Blackwell is \$25 billion.

**Jensen Huang**

President and CEO

0:57:58

Next, Vera Rubin is \$40 billion, and after that it is going to be higher. That is excellent. That is fantastic for the industry. It is fantastic for customers. So long as the productivity of it continues to grow, the durability and the fungibility continues to grow, then people are happy to invest in assets that generates revenues, generates profits, and helps them recoup their returns so incredibly fast. I heard the other day that return on investment capital is now less than a year, and we are talking about \$50 billion data centers. That tells you something about the productivity of NVIDIA's technology and the rentability of it. I want to thank all of you for joining us today.

**Operator**

0:58:51

There are no further questions at this time. Toshiya Hari, I turn the call back over to you.

### **Toshiya Hari**

VP of Investor Relations and Strategic Finance

0:58:57

Thank you. Before we close, please note that Jensen will be participating in a keynote fireside chat at the Goldman Sachs Communacopia + Technology Conference in San Francisco on September 10th. He will also be giving a keynote at GTC Berlin on October 21st. Our earnings call to discuss the results of our third quarter of fiscal 2027 is scheduled for November 17th. Thank you for joining us today. Operator, please close the call.

### **Operator**

0:59:26

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

**yahoo!finance**

Copyright © 2026 Yahoo.

All rights reserved.



### **Trending Topics**

Dow Jones

S&P 500

DAX Index

Nvidia

Tesla

DJT

Tariffs

### **Explore Further**

Mortgages

Credit Cards

Sectors

Crypto Heatmap

Financial News

About

Data Disclaimer

Help

Feedback

Sitemap

Licensing

What's New

About Our Ads

Premium Plans

Terms and Privacy Policy

Your Privacy Choices 

 Quote Lookup

Trending tickers

|         |      |                   |
|---------|------|-------------------|
| NVDA    | N... | -3.39 (-20.96%)   |
| BTC-USD | B..  | +968.57 (+26.81%) |
| CRM     | S... | -0.07 (-20.56%)   |
| OKTA    | O... | +3.81 (+29.24%)   |
| CRWD    | C... | +3.80 (+28.51%)   |

Portfolio

Sign in to access your portfolio

Sign in

Recently viewed

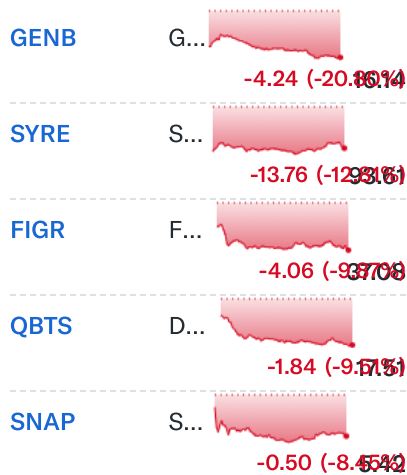
|        |      |                 |
|--------|------|-----------------|
| AMZ.HA | A... | -0.65 (-22.95%) |
| SPCX   | S... |                 |



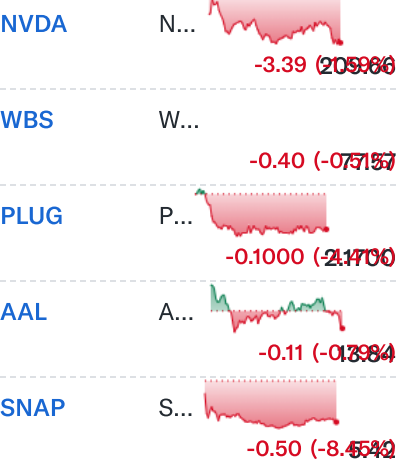
## Top gainers



## Top losers



## Most active



Top economic events

United States  ▾

Initial Jobless Clm \*  
Aug 27, 2026, 8:30 AM EDT  
Prior: P: 206 New: N: 203

U Mich Sentiment Final\*  
Aug 28, 2026, 10:00 AM EDT  
Prior: P: 51 New: N: -

ISM Manuf Employment Idx  
Sep 1, 2026, 10:00 AM EDT  
Prior: P: 52.8 New: N: -

Mortgage Market Index  
Sep 2, 2026, 7:00 AM EDT  
Prior: P: 245.3New: N: -  

< >

Earnings events  

Upcoming ▾

|           |                      |                                                                                     |
|-----------|----------------------|-------------------------------------------------------------------------------------|
| AUG<br>27 | MRVL                 |  |
|           | Marvell Technolog... |                                                                                     |
|           | IREN                 |  |
|           | IREN Limited         |                                                                                     |
| SEP<br>02 | AVGO                 |  |
|           | Broadcom Inc.        |                                                                                     |
|           | HPE                  |  |
|           | Hewlett Packard ...  |                                                                                     |
| SEP<br>03 | PATH                 |  |
|           | UiPath, Inc.         |                                                                                     |

Edit your Dock