20-May-2026

NVIDIA Corp. (NVDA)

Q1 2027 Earnings Call

CORPORATE PARTICIPANTS

Toshiya Hari
Vice President-Investor Relations & Strategic Finance, NVIDIA Corp.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

OTHER PARTICIPANTS

Joseph Moore
Analyst, Morgan Stanley & Co. LLC

Vivek Arya
Analyst, BofA Securities, Inc.

Ben Reitzes
Analyst, Melius Research LLC

Stacy A. Rasgon
Analyst, Bernstein Research

C.J. Muse
Analyst, Cantor Fitzgerald & Co.

James Edward Schneider
Analyst, Goldman Sachs & Co. LLC

Timothy Arcuri
Analyst, UBS Securities LLC

Joshua Buchalter
Analyst, TD Cowen

MANAGEMENT DISCUSSION SECTION

Operator: Good afternoon. My name is Sarah, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's First Quarter Earnings 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. [Operator Instructions] Thank you.

Toshiya Hari, you may begin your conference.

Toshiya Hari

Vice President-Investor Relations & Strategic Finance, NVIDIA Corp.

Thank you, and good afternoon, everyone. Welcome to NVIDIA's conference call for the first 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 second quarter of fiscal 2027. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed, without our prior written consent.

During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties, and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent Forms 10-K and 10-Q, and the reports that we may file on Form 8-K with the Securities

and Exchange Commission. All our statements are made as of today, May 20, 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 M. Kress

Chief Financial Officer & Executive Vice President, NVIDIA Corp.

Thanks you, Toshiya. We delivered an exceptional quarter with revenue, operating income, and free cash flow exceeding our prior records. Total revenue of $82 billion was up 85% year-over-year and 20% sequentially. This marked our 3rd consecutive quarter of year-over-year acceleration and the 14th straight quarter of sequential growth, a significant feat given the sheer size and complexity of our manufacturing operations.

The $13.5 billion sequential revenue increase was also a record. We capitalized on the inflection in inference demand by ramping Blackwell systems across our diverse end customer base from hyperscalers to model makers to AI cloud providers and sovereign customers.

In Q1, we also allocated capital effectively across R&D, investments in our ecosystem, and share repurchases. We returned a record $20 billion to our shareholders, while executing strategic investments, both upstream supply chain and downstream go-to-market ecosystem. This is critical to the market's development and our long-term position.

Data Center revenue of $75 billion was up 92% year-over-year and 21% sequentially, driven by sustained strength in our Blackwell architecture. And demand for GB300 NVL72 was particularly strong with frontier model builders and hyperscalers, each having cumulatively deployed hundreds and thousands of Blackwell GPUs, marking the fastest product ramp in our company's history.

Grace Blackwell is the fastest training system as well as the lowest token generation cost at inference. Spectrum-X, our end-to-end Ethernet platform purpose built for AI, is now larger than all Ethernet network peers combined. InfiniBand has also had a very strong quarter, growing more than 4x year-over-year, driven by deployments of our next-generation XDR technology.

For your models, Data Center computing revenue of $60 billion was up 77% year-over-year, while Data Center networking revenue of $15 billion nearly tripled year-over-year. Before we deep dive into Data Center, we'd like to brief you on our transition to a new reporting framework that better reflects our current and future growth drivers.

We have two market platforms: Data Center and Edge Computing. Within Data Center, we will report two sub-markets; Hyperscale and ACIE, which incorporates AI clouds, industrial and enterprise. Hyperscale will include revenue from the public cloud and the world's largest consumer Internet companies, while ACIE addresses our growth opportunities in diverse AI purpose-built data centers and AI factories across industries and countries. Edge Computing highlights devices for agentic and physical AI, including PCs, gaming consoles, workstations, AI-RAN base stations, robotics, and automotive. For your reference, we have posted on our website a revenue breakdown based on our new platforms for the past nine quarters.

Moving back to our Data Center results. Hyperscale revenue of $38 billion was approximately 50% of Data Center revenue and increased 12% quarter-over-quarter. ACIE revenue was $37 billion and grew 31% quarter-over-quarter, including AI cloud revenue that more than tripled year-over-year.

Our customers have enabled rapid stand-up of AI compute capacity. The number of partner data centers exceeding 10 megawatts has nearly doubled in just one year, now surpassing 80 sites. Sovereign revenue increased more than 80% year-over-year. NVIDIA AI infrastructure is now deployed across nearly 40 countries, representing $50 trillion in GDP.

As evident to our Q1 results, our customer base is diverse and growing. Supported by our vast ecosystem and installed base, breadth of CUDA accelerated application, and the lowest token cost provider, we are well positioned to address a market opportunity that far exceeds that of any other AI computing platform.

Demand for AI infrastructure continues to expand at an unprecedented pace. The build-out of AI factories is accelerating. The value of NVIDIA AI infrastructure is rising. The price of renting an H100 has risen 20% year-to-date, while A100 cloud pricing is up nearly 15%. Benefiting from the versatility of our platform and continuous performance enhancements enhanced by our software stack, customers are generating profitable revenue beyond the depreciable life of their GPUs. The vast and trusted marketplace for NVIDIA Compute is a critical foundation on which billions in AI infrastructure spending is being financed by the ecosystem.

There are two primary drivers behind the accelerating build-out of AI infrastructure. First, from search and advertising to recommender systems and content understanding, the largest hyperscale workloads continue to transition from CPU to GPU-based accelerating computing.

Second, the adoption of products and services native to AI is inflecting. Since the advent of ChatGPT, we have witnessed mainstream AI transition from one-shot inference to reasoning and to now agentic. AI is no longer a nice to have. AI is now a necessity for enhancing productivity across all industries and roles. This is propelling revenue acceleration across all layers of the AI cake, including energy, chips, infrastructure, models, and applications. Growth in the model layer, particularly at Anthropic and OpenAI, has been incredible with momentum continuing to accelerate, including breakout growth in OpenAI's Codex since the launch of GPT-5.5.

With analysts now forecasting hyperscale CapEx to exceed $1 trillion in 2027 and agentic AI beginning to proliferate all industries, AI infrastructure spending is on track to reach $3 trillion to $4 trillion annually by the end of this decade.

Our Blackwell architecture is everywhere, adopted and deployed by every major hyperscaler, every cloud provider, and every major model maker. Last month, we celebrated OpenAI's launch of GPT-5.5, co-designed for, trained with, and served on Blackwell, currently positioned at the top of Artificial Analysis leaderboards.

Microsoft's Fairwater, the world's most powerful AI data center, is now live, ahead of schedule, powered by hundreds of thousands of Blackwell GPUs. Starting this year, AWS will add more than 1 million Blackwell and Rubin GPUs, and are collaborating on Spectrum networking. At Google, Blackwell will be offered to customers in the cloud, including confidential computing capability, a new foundation for secure, high-performance AI.

Our share of frontier AI compute is increasing. We have deepened our collaboration with Anthropic and are delighted to be a strategic partner to expand their compute capacity. We will support the company's growth trajectory through AWS, Azure, CoreWeave, SpaceXAI, and more. Now, with the addition of Anthropic to OpenAI,

Gemini, SpaceX xAI, Meta, MSL, Microsoft AI, TML, Reflection, Perplexity, Cursor, and other major frontier labs already building on NVIDIA, our share of frontier AI models will grow significantly.

Today's data centers are revenue-generating AI factories. Constrained by power and capital, AI factory operators must choose the right architecture. With our extreme co-design approach, we deliver the industry's lowest token cost, the highest token throughput, and the highest ROI.

MLPerf inference results are in and once again, we swept every benchmark as Blackwell Ultra delivered the highest throughput across the broad set of models and deployment scenarios. Full stack innovations drove the 2.7x increase in throughput and a 60% reduction in the cost per token on GB300 compared to just six months ago.

NVIDIA Compute is not just the highest performance AI infrastructure, it is the most economic and financeable. Customers do not buy GPUs. They build AI factories. And the right economic metric is not the purchase price of the GPU, it is the lifetime cost of an AI factory producing intelligence, token per watt, tokens per dollar, uptime, utilization, time to production, software durability, and asset life. NVIDIA excels at all of them.

Agentic AI and reinforcement learning represents new growth opportunities for CPUs. Building on the success of our Grace CPU, Vera is arriving just in time to meet this inflection. Built on custom Arm cores and co-designed end-to-end with Rubin GPUs and NVLink, Vera will deliver up to 1.5x faster performance per core, 2x performance per watt, and 4x density per rack, compared to x86-based alternatives.

Vera CPU opens a brand new $200 billion TAM for NVIDIA, a market we have never addressed before. And every major hyperscale and system maker is partnering with us to get it deployed. We have visibility to nearly $20 billion in total CPU revenue this year, setting us up to become the world-leading CPU supplier.

Our annual product cadence, a pace that is unmatched, remains a key pillar supporting our market position. We are on track to commence production shipments of Vera Rubin in the second half of this year, starting in Q3. By integrating seven purpose-built chips across five accelerated racks, Vera Rubin will deliver up to 35x higher inference throughput and up to 10x greater AI factory revenue compared with Blackwell.

As an early adopter, Google's A5X bare-metal instances, which can support up to 960,000 Rubin GPUs across multiple sites, can enable customers to run their largest AI workloads on NVIDIA's optimized infrastructure.

While the US government has approved licenses for H200 to be shipped to China-based customers, we have yet to generate any revenue, and we are uncertain whether any imports will be allowed into the country. As a result, consistent with last quarter, we are not including any China Data Center compute revenue in our outlook.

Let me move to Edge Computing. Our Edge Computing market platform generated $6.4 billion, up 10% quarter-over-quarter and 29% year-over-year. Robust Blackwell workstation demand was a strong contributor to the growth, while consumer demand fell modestly due to higher memory and system prices.

Our physical AI continues to gain momentum, exceeding $9 billion in revenue over the last 12 months. Our partnership with Uber will power the robotaxi fleet across nearly 30 cities and 4 continents by 2028. And in robotics, leading companies across a range of industrial, surgical, and humanoid applications are building on NVIDIA's technology to develop and deploy at scale.

We remain front-footed in securing sufficient supply to support our customers' growth. In Q1, we increased total supply, inclusive of inventory, purchase commitments, and prepaids to $145 billion. While we are not immune to supply challenges, we remain confident in our ability to support the growth opportunity ahead, with our intense focus, scale and longstanding partnerships with critical suppliers continuing to serve us well.

Let me move to the rest of the P&L. GAAP gross margin was 74.9% and non-GAAP gross margin was 75%, largely flat sequentially by Blackwell systems continued to account for most of our shipments. GAAP and non-GAAP operating expenses were up 12% sequentially, primarily due to higher compensation and an increase in compute and infrastructure costs.

Our non-GAAP effective tax rate of 16% came just below our prior outlook due to favorable geographic mix. And on our balance sheet, days sales outstanding was 45 days. Due to favorable timing of collections, we expect to return to the mid-50s in Q2. We generated record free cash flow $49 billion, up from $35 billion in Q4.

I'd now like to update you on our capital allocation plan. First, to reiterate, our intention is to prioritize R&D and strategic investment, both will enable us to cultivate our ecosystem, drive market growth, and strengthen our market position. As a key enabler of AI, we will make investments necessary to deliver the industry's lowest cost per token and the highest token throughput, which will help our customers and partners scale and expand the AI frontier.

Return program is another key component of our capital allocation strategy. Given confidence in our long-term free cash flow outlook and our commitment to sharing our success with shareholders, we are increasing our quarterly dividend from $0.01 to $0.20 (sic) [$0.25] per share. We plan to review our dividend on a regular basis, as we continue to scale our business. We are also announcing an $80 billion share repurchase authorization, which is in addition to the $39 billion remaining on our current plan. As we indicated at GTC, we plan to return roughly 50% of free cash flow to shareholders this year.

Let me turn to the outlook for the second quarter. Total revenue is expected to be $91 billion, plus or minus 2%. We expect sequential growth to be driven primarily by Data Center. We are continuing to work vigorously on our supply chain ecosystem to address the incredible demand we see ahead of us, giving us full confidence in the $1 trillion in Blackwell and Rubin revenue we foresee from 2025 through calendar 2027.

GAAP and non-GAAP gross margins are expected to be 74.9% and 75%, respectively, plus or minus 50 basis points. For the full year, we are still expecting to be in the mid-70s.

GAAP and non-GAAP operating expenses are expected to be approximately $8.5 billion and $8.3 billion, respectively. For the full year, we now expect OpEx to grow somewhere in the upper-40s on a year-over-year basis, driven by higher R&D and acceleration in the usage of AI tools to enhance productivity.

For the full year 2027, we expect GAAP and non-GAAP tax rates to be between 16% and 18%, excluding any discrete items from material changes to our tax environment. This is lower than our prior expectation of 17% to 19% due to changes in geographic mix.

That puts me at the end of this part, and I'm going to now turn this over to the Q&A with Toshiya.

Toshiya Hari

Vice President-Investor Relations & Strategic Finance, NVIDIA Corp.

Thanks, Colette. We will now transition to Q&A. Operator, please poll for questions.

QUESTION AND ANSWER SECTION

Operator: Thank you. [Operator Instructions] Your first question comes from Joseph Moore with Morgan Stanley. Your lie is open.

Joseph Moore
Analyst, Morgan Stanley & Co. LLC

Q

Great. Thank you for letting me ask a question. I guess, I'd like to ask, what drove the change in segmentation? What's the philosophy behind giving us the numbers that way? And then, can you talk about any competitive differences between the two segments and this kind of surprising CPU number that you talked about, how do you see that across the two segments as well? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah. Thanks, Joe. First of all, Colette meant to say we're increasing our quarterly dividend from $0.01 to $0.25. I think that extra $0.05 would mean a lot to the large shareholders.

So, anyhow, let's see, Joe, on the segmentation and the description of the business, we wanted you to understand our business better. AI is very diverse and computing is diverse. They're diverse in several ways. The first thing, of course, is AI includes languages, and depending on the different industries, it could be 3D graphics for manufacturing and industrial robotics. It could be proteins for life sciences. It could be small chemicals for life sciences or material sciences. It could be physics for the physical sciences, whether it's in the energy sector or, of course, the science labs, higher education, so on and so forth. So, AI is diverse.

The second thing is the applications are diverse. It could be in enterprise. It could be in the energy sector, manufacturing sector, and such. Where it runs is diverse. It could be in the hyperscale cloud. It could be AI natives. There's a whole network of AI natives that are cropping up around the world. Enterprises on-prem, industrial in the factories, in the plants, all the way to supercomputing centers and the edge. Edge, including, of course, what most people see, self-driving cars, robotics, but a large growing network of computers inside manufacturing plants, whether it's a chip plant or packaging or computer plants, all kinds of different types of manufacturing plants. And then, of course, in the future, every single base station, every single radio network would become an AI-powered radio network. And so where it runs.

And then lastly, how it's governed. It could be operated by public cloud, but it could also have industrial regulatory reasons that prevents it from being run in a regulatory cloud. It could be because of confidential computing. It could be because of national security reasons. Different data centers have to be built differently.

NVIDIA is quite unique in the sense that we are the only company that builds all of the technology components. We build it in an extreme co-design way, in a complete end-to-end way, in a full stack way. But then, we, of course, open the platform so that it could be integrated into all the different environments. But some environments just require – an enterprise, for example, require a company who has all of the technologies working together so that they don't have to build it. They would like to buy it and operate it. And so there's many different segments of the data center market where NVIDIA's total solution, fully integrated solution with full stack, but still open, that way of doing – of producing or delivering products is really, really important.

And so if you look at our different segments, the way we broke it out into three large segments. You take all of the words that I just said and you try to find the simplest factoring of it, it would be the hyperscale clouds, that would be one large segment. And within that segment, there's three different ways that we operate. First way is that we help the hyperscale clouds accelerate their data processing and machine learning workloads. We accelerate and support their AI processing inside. We also, of course, bring a lot of business – NVIDIA ecosystem business to their public clouds. And so that's one segment.

The second segment is AI natives, enterprise on-prems, industrial on-prems and sovereign AI. That segment is growing incredibly fast because everybody needs AI, and we're going to see AI being adopted by every industry, every country, every company. And so everybody wants to build it in a different way. And the fact that we provide the entire solution, it makes it much easier, it makes it possible at all for people to be able to build these things.

And then, of course, the robotic edge. Today, yesterday's computing was largely about personal computing. In the future, it's going to be about personal AI. And that personal AI, one example of it is the self-driving car. It's a car – it's a robotic system that's essentially your personal AI. And of course, there'll be all kinds of different types of robotic systems, including even the base station radio network, as I mentioned, is going to be essentially a robotic system.

And so that's the reason why we broke it all apart this way. It's the simplest way of understanding our business. Each one of them have different stacks in a lot of ways. They have different operating systems. They operate in a different way. We go to market very differently in each one of them. The easiest go-to-market, of course, is the hyperscaler, because there are only five or six of them. But the rest of them – the rest of the industry represents a couple of 250,000 companies around the world. That go-to-market is very complex, very diverse. Your understanding of AI has to be extremely diverse.

And as you know, NVIDIA has the largest suite of acceleration libraries in the world from computational lithography to fluid dynamics to particle physics to molecular dynamics to the list goes on. And all of those libraries are essential for us to engage the vertical industries that represents the second and the third category. Okay?

So, anyways, it's really about the fact that our business has now evolved and grown to such a large scale, it's helpful to segment it so that you have a better understanding of how our business works.

Operator: Your next question comes from Ben Reitzes with Melius Research. Your line is open.

Ben Reitzes

Analyst, Melius Research LLC

Q

Hey, guys. Thanks so much. I wanted to ask, Jensen, I want to ask you about your philosophy on growth. Your Data Center business ex-China grew about 120% in the quarter. And then, you're guiding about 100%. CapEx at the hyperscalers is forecast by many, including myself, to like grow 90% to 100% this year.

And you talked about Data Center still on track to be $3 trillion to $4 trillion by the end of the decade. I was just wondering, the goal for the company to grow faster than hyperscaler CapEx, are you comfortable in kind of endorsing that view? And do you still see hyperscaler CapEx kind of still growing after this year at a very rapid clip? Thanks a lot.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Yeah. Thanks, Ben. So, first of all, we should be growing faster than hyperscale CapEx. And the reason for that is illustrated by the segmentation that I just described. Our Data Center business has two large parts. It has more parts than that, but we combined it into two large parts for simplicity's sake. It's much more complex than the two large parts, but I combined it into two, so that it's at least easier to understand. Okay?

And so if you look at the first part is hyperscalers. That's the hyperscale CapEx that you were just talking about. And they're at $1 trillion this year. I have every expectation it's going to grow from here for fundamentally good reasons. This is the way computing is going to work in the future. And if they don't have the compute, they won't have the revenues. It is very clear compute is revenues, compute is profit. And so the world is changing. SaaS didn't used to use as much compute, but AI requires a tremendous amount of compute.

But you could do, of course, incredibly more, which is the reason why we heard about the frontier AI companies, both Anthropic and OpenAI, growing at an incredible pace. The fact that they can grow within one month, what some of the SaaS companies would have taken a decade to grow, tells you something. And so the first category is hyperscale and the CapEx is at $1 trillion, and it's growing towards the $3 trillion to $4 trillion.

The second category is all of the AI native clouds. They're regional. They're all over the place. There are startups all over the world supporting those companies. They're enterprise. 250,000 enterprise companies around the world, many of them will have to build or want to build AI factories for themselves to operate. Many industrial companies, there's no choice but to put the computer where the context is, where the action is. You can't put that in the cloud. It has to respond reliably, quickly every single time. Can't imagine a chip plant – a chip fab being connected to a cloud service provider, doesn't make any sense.

And so, the second category, and then sovereign AI clouds, and so there's a whole category of data centers that semi-custom chips just don't apply, because these data centers want to buy systems. They want to operate systems. They don't want to design – they don't want to build it themselves. And so, the second category is extremely diverse. Instead of five or six, seven companies representing the revenues associated with our first category, the second category is hundreds, thousands of companies. And in the future, it will be hundreds of thousands of companies with a large number of companies with smaller installations. And that category is going to continue to grow at incredible pace.

This second category, when I talk about physical AI and I talk about how the rest of the $100 trillion industry that has not been impacted by IT in the last 30 years, it's about to be impacted by AI, that is the segment that I'm talking about. The second cluster is growing incredibly fast. Our share of that, of course, is very, very large. We're fairly unique in our ability to be able to serve this industry. Our platform is built like it's vertically integrated, so that everything works. But when – then we disassemble it, so that people can build and buy it in the configuration they want and assemble it the way they like.

And so, this second category is fairly poorly understood because they're just so many small companies, or so many companies, and each one of the installations are relatively small compared to, of course, one of the hyperscalers. And so, if you look at the segmentation and the size of each, you could see that, in fact, we're growing share in the hyperscalers, because we now have much bigger support from Anthropic, a new partner of ours, and we're helping them expand their capacity greatly in the coming years. And then, the second, very few companies have exposure into the second category, because of the platform solution that we have.

Operator: Your next question comes from C.J. Muse with Cantor Fitzgerald. Your line is open.

C.J. Muse
Analyst, Cantor Fitzgerald & Co.

Q

Yeah. Good afternoon. Thank you for taking the question. You have Vera Rubin coming soon. And you, obviously, have great insight into coming updates to frontier models, new techniques to optimize around diverse AI workloads. With investors keenly focused on your market share in inference, how do you see Vera Rubin and your extreme co-engineering impacting your share of the inference market as we look into late 2026, 2027?

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Well, we are growing share in inference and we're growing share in inference very, very quickly. And the reason for that is, this year, the number of frontier model companies grew. And so there's Cursor and Perplexity, and there's some new model companies TML and Reflection and the list goes on. And so the number of frontier model companies has grown.

And we added Anthropic to our partnership this year. They're expanding incredibly fast. We've partnered with them to secure computing capacity across Azure, AWS, CoreWeave. I forget who else we've already announced, but there's a whole list of others that we are bringing online for them. And so the amount of capacity that we're going to bring online for Anthropic this year and next year is going to be quite significant, very significant. And our coverage of Anthropic has been largely zero until just recently.

And so we're gaining share tremendously fast in inference. Vera Rubin is going to be even more successful than Grace Blackwell at this point. Every single – I can't think of one, every single frontier model company will jump on Vera Rubin from the get go and that wasn't true before on Blackwell. And so Vera Rubin is off to a tremendous start, and it'll surely be more successful than even Grace Blackwell. So, I think the end of your answer, C.J., is that we're gaining share in inference.

Let me go back again to the question that Ben was asking. Remember, so far, everything that I've just explained in the inference question is really focused on hyperscale. Remember, there's a whole second category of AI data centers that we serve almost uniquely. This segment is very fragmented, requires a really well-integrated platform solution and a very large go-to-market, and that segment, all of the inference, 100% of that, the vast majority of that is NVIDIA.

And then, of course, physical AI. NVIDIA is practically the only company serving physical AI today, and we've been working on physical AI for a long time. And so, that is also growing. So, our share of inference is growing very quickly.

Operator: Your next question comes from Timothy Arcuri with UBS. Your line is open.

Timothy Arcuri
Analyst, UBS Securities LLC

Q

Thanks a lot. Jensen, I wanted to ask about the traction you're getting with some of these custom merchant things. You're doing stuff like CPX and LPX. And I just wanted to ask and see sort of you've talked before about [ph] SaaS LPX (00:37:50) being, I think, 20% of the market. So, I would imagine you're getting pretty good traction with LPX. So, can you just talk about that and maybe also how that fits into your broader platform strategy? Thanks.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

The LPX is designed for low latency and high token rate, but its throughput is low. Its model size capacity is low. And its context processing, its ability to absorb a lot of context, for example, for software coding, for agentic workloads, its ability to absorb a great deal of context is lower. And so the challenge is simply – and I've explained before that the use case for LPX is not broad. It's intended for somebody who has a fairly large portfolio of different types of token services. And for the high token rate, maybe these services are quite premium and the number of customers is not significant, but the token rate is very high. And so, that remains exactly consistent with what I've said before. And I still expect that. And so I expect that LPX and other SRAM-based, decode-focused, high token rate generated focused accelerators will be a niche product for some time to come.

As you know, Grace Blackwell and Vera Rubin, we support the entire life cycle of AI from the data processing, preparing for training, okay, data processing to pre-training, to post-training, reinforcement learning, all the way to inference, Grace Blackwell is the best platform in the world to do all of that. And if we insert certain circumstances, so long as the customer – the provider already has a high token rate service that they can offer, then we can tack on an LPX and they could deliver that service even better. And so that's how I see the market.

And I think whether it's 20% or 10%, just depends on where we are in the development of AI. I think today, it's a lot less than 20%. Someday, these premium tokens could be 20%. And we're ready to work with the service providers to enable this capability. I'm excited about it.

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

Vivek Arya
Analyst, BofA Securities, Inc.

Q

Thanks for taking my question. Jensen, there's a lot of excitement around CPU for agentic applications and just a lot of noise around the number of CPUs actually exceeding the number of GPUs. And I was just hoping that you could kind of give your perspective that, first of all, is this an incremental workload? Is this kind of cannibalizing what the GPU would have done otherwise?

And then, secondly, the $20 billion number that you gave, is that for standalone Vera CPUs, or is that kind of already included in that Vera as part of Vera Rubin? So just if you could educate us on the role of CPU versus GPU. Is it cannibalistic? Is it incremental? And then, the $20 billion number, how to kind of put that in context with what you sell, right, which is usually the CPU as part of the GPU. Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

The $20 billion is for standalone CPU. And remember, we have Vera is used in three ways, as a standalone – four ways, as a – let me just start with the one that you already know. The first way is Vera Rubin. And we'll sell millions of Rubins, and every two of them is connected to a Vera. And of course, we price those too. And they're properly priced. And so that's number one use case.

The second use case is Vera standalone CPU. The third is Vera with CX9 and the software stack for storage. And then Vera in a – with CX9 with a software stack for security and compute isolation and confidential computing. Okay? So each one of those use cases is built on Vera. And my sense is that we'll be supply constrained throughout the entire life of Vera Rubin. There are four different use cases of it. And – but anyhow, the answer to your question is of the $20 billion is a standalone.

With respect to CPUs, an agent is essentially what people call a harness. And the agent has a harness that does the – and the harness could be OpenClaw, it could be Hermes. Claude Code is essentially a harness around Claude, around the Opus model. OpenAI's Codex is a harness around the GPT-5.5 model. And so these are harnesses, and these harnesses provide for things like IO, orchestration, memory management, tool use, connected to tools, for example, browsers and things like that, C compilers, Python compilers. And so, the harness runs on CPU and the tool use runs on CPUs. For example, if the AI were to do a search or do a browser, use a browser, that would run on the CPU.

The world has 1 billion users, human users. My sense is that the world is going to have billions of agents. Not today. I mean, we're going to grow into it. But it will have billions of agents and those billions of agents will all use tools. And those tools are going to be like PCs. Just like us humans using PCs today, in the future, you'll have an agent using PC.

And so, if you kind of think along the lines of in the future, you pick your favorite number of agents at the moment, at the moment, call it, a few hundred thousand, but in the future, call it, eventually a few billion, I could imagine them all using – effectively having PCs that they can all use.

Every one of those agents are going to spin off sub-agents. And every time they spin these off, you're going to need to do inference. That's where the thinking happens. All of the thinking happens on GPUs. All of the orchestration essentially runs on CPUs and the sub-agents, when they're spun off, when they're thinking, they use GPUs.

Whenever the agents use simulators, those can run on CPUs or GPUs, which is the reason why we're working so closely with Cadence and Synopsys to accelerate all of the world's tools. We're accelerating all of the world's tools and data processing engines and database engines because agents use these tools and they have lower patience – tolerance than humans, and they want things to happen quickly.

And so we're accelerating all of the world's tools so that it runs on CUDA. And you could see us doing that, when I work with Cadence and Synopsys and Siemens and companies – and Adobe, that's because we're trying to get all of the world's tools to run on GPUs, because they already have GPUs and it's a lot faster.

So, we're going to need a lot more CPUs, and Vera was designed to be an agentic CPU. The CPUs of the past were designed to have many cores, so that it could be easily rentable. People rented cores. Well, agents don't rent cores. They just want the work to be done fast. The economics of the past was dollars per core. That's the economics of cloud computing of the past. The economics of AI of the future is tokens per dollar or dollars per token. And so, what we need to do in the future is to generate tokens, process tokens as fast as possible. And that's what Vera does incredibly well.

So we're expecting to be very successful with Vera. But ultimately, what we're doing is we're building infrastructure for AI and it needs incredibly great storage. That's the reason why we built STX. It needs incredibly good networking. That's why we have Spectrum-X. It needs incredibly great GPUs, of course, and inferencing ability. That's the reason why NVLink 72. It needs incredibly great security and confidential computing, which is the reason why Vera Rubin is the world's first platform with end-to-end confidential computing. And it needs great CPUs. We've got it all covered.

Operator: Your next question comes from Stacy Rasgon with Bernstein Research. Your line is open.

Stacy A. Rasgon
Analyst, Bernstein Research

Q

Hi, guys. Thanks for taking my question. I wanted to go back to the segmentation. So, first of all, I'm just curious, where do you put the neo clouds across those two segments? Are they in hyperscale or are they in the AI cloud? Part of me assumes the latter, but I'm not so sure. And then, by just looking...

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

Correct.

Stacy A. Rasgon
Analyst, Bernstein Research

Q

...at the magnitude of them, I mean, they're both about the same magnitude now. It almost sounded to me like you were suggesting that you thought the latter, the AI cloud would grow faster, maybe going forward than hyperscale. Is that what you were trying to say? Or do you see like the same kind of growth coming from both segments?

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

First of all, you're correct that AI native clouds. AI native clouds don't build chips – don't design their own chips and they can't really assemble unrelated parts together into an AI factory. And their patience, their tolerance for time to first token is extremely low. And their need for an architecture that has a great deal of offtake, so that it runs every model has customers from everywhere, is incredibly high. And so that's the reason why NVIDIA's architecture is so perfect for them.

We offer every component and whatever we don't offer, our ecosystems of partners offer it. And it's all fully integrated. It all works together. The number of customers that could rent it from an AI native is incredibly high, basically, every single AI builder, every AI native startup around the world, SaaS companies, enterprise companies, industrial companies.

And so, our computing – our architecture is the most rentable of any computing platform in the world. So it's the most performant. It's the easiest to put together. It's the most rentable. Has the best TCO. And it's the easiest to finance. And so all of those properties are quite unique to the needs of AI natives. It's in the second category. They're very similar to even OEMs and so on and so forth, large enterprises and so forth, surprisingly. Okay? So we put that in the second category.

If you look at that segment, it started growing after the AI ecosystem developed in the hyperscale. Hyperscale developed AI first for a lot of reasons. They have great computer science. They have excellent data center capability. And they also focus largely on consumer applications, which, if not perfect, is not the end of the world. It enhances the service, so long as it enhances the service.

And so for many of the other applications, industrial applications, enterprise applications, until the AI is very capable and does really productive work and does it safely, and it could do it in a way that can actually generate impact and income, it doesn't really get used. And so you expect the second category to develop slower than hyperscale, and you could see that in the numbers.

However, long term, if you look at industrial and enterprise, clearly, that's where future economics is going to be, because it represents some $50 trillion, $80 trillion of the world's economy. And it's going to be larger than that

because of AI. And so I expect the second category to be larger over time, both in the near term, over the next several years. I think it's a foregone conclusion, both are going to grow incredibly fast. I expect the second category to still grow faster, but both are going to grow incredibly fast. And then, I'm hoping that within the next five years, physical AI and robotics segment is going to grow incredibly fast.

Operator: Your next question comes from Jim Schneider with Goldman Sachs. Your line is open.

James Edward Schneider
Analyst, Goldman Sachs & Co. LLC

Q

Good afternoon. Thanks for taking my question. Back at GTC, I believe you discussed $1 trillion visibility into both your Rubin and Blackwell platform revenue. But I believe that excluded things like LPX, Rubin CPX and the Vera CPU racks. Can you maybe give us a sense about whether the Vera CPUs are going to be the biggest source of upside above and beyond that $1 trillion? Are you contemplating other sort of combinations of products, including CPUs that would allow you to gain an even greater share of that total TAM? Thank you.

Jen Hsun Huang
Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

A

In terms of incremental above the $1 trillion, I would say, one, the continued growing of share of the frontier AI models. I'm expecting to grow more share. And so I'm expecting that to grow. Number two, we didn't include any Vera CPU, standalone CPU in that number. And so I expect that to be the second largest. The TAM is, of course, quite large in agentic systems, and all of our customers are quite excited about Vera, and we're going to sell a whole bunch of Veras.

And then third would be LPX, because as I explained earlier, LPX is designed as a – because of its SRAM architecture, it has the benefit of very low latency and very high interactivity, but it's also its throughput and its context processing ability is also quite limited. And that's just kind of the nature of SRAM type-based systems. But the combination, we'll be able to address the entire spectrum of AI from pre-training to post-training to inference agentic systems through the combination of Vera Rubin and LPX.

Operator: Your next question comes from Joshua Buchalter with TD Cowen. Your line is open.

Joshua Buchalter
Analyst, TD Cowen

Q

Hey, guys. Thanks for taking my question, and congrats on the great results. Colette, I believe in your prepared remarks, you mentioned GB300 is sort of the fastest ramp in the company's history. How should we think about Vera Rubin against this benchmark? It's, obviously, a new architecture at the silicon level, but similar rack. Does that mean we should expect a similar slope to the Vera Rubin ramp as the GB300? Or should it be a bit more gradual given the new silicon? Thank you.

Colette M. Kress
Chief Financial Officer & Executive Vice President, NVIDIA Corp.

A

Yeah. Well, we have indicated for a while that we will be launching Vera Rubin in the second half. We will start in Q3. That will be our initial pieces together. And then once we get to Q4, we're probably going to start to see our ramping continue. It's hard to say at this point what will be a faster ramp. But again, we have demand already planned. We've got POs. We've got almost all of our major customers ready to go. And these are very complex

systems that we need to put together. So I think it's just about the timing that it's going to take for us to get that into market.

Nothing else other than getting from production of all of the different systems that we have ready for order. So, a little early to say, but yes, we're going to start in Q3 and continue to ramp into Q4 and Q1 of next year certainly is going to be very big as well.

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

Toshiya Hari

Vice President-Investor Relations & Strategic Finance, NVIDIA Corp.

Thank you. Before I hand it over to Jensen, please note Jensen will be giving a keynote at GTC Taipei at COMPUTEX on June 1. We will also be participating at the TD Cowen TMT Conference on May 28, and the Bank of America Global Technology Conference on June 4. Our earnings call to discuss the results of our second quarter of fiscal 2027 is scheduled for August 26.

With that, here's Jensen to close us out.

Jen Hsun Huang

Co-Founder, President, Chief Executive Officer & Director, NVIDIA Corp.

This was an extraordinary quarter. Demand has gone parabolic. The reason is simple. Agentic AI has arrived. AI can now do productive and valuable work. Tokens are now profitable. So model makers are in a race to produce more. In the AI era, compute capacity is revenue and profits. NVIDIA is the platform of this era. Of all the platforms in the world, NVIDIA Compute supports the richest diversity of demand. Let me highlight my top five things.

First, NVIDIA is the only platform that runs every frontier AI model. With the addition of Anthropic to our existing partners, OpenAI, xAI, Meta MSL, Gemini, and many others, our share of frontier AI is growing.

Second, we are in every hyperscale cloud, supporting their core data processing and machine learning workloads, internal AI services, as well as supporting their demand for NVIDIA users in their public cloud services.

Third, our full stack complete AI factory solution and vast global ecosystem let us uniquely address new AI data center segments, mew AI native clouds and sovereign AI clouds and on-premises enterprise and industrial infrastructure. This is that second category I was talking about earlier.

Fourth, NVIDIA CUDA extends all the way to the edge. Robotics, autonomous vehicles, embedded medical instruments. AI-RAN, telco base stations. The next wave is physical AI with billions of autonomous and robotic systems operating in the physical world. This is the third segment we were talking about earlier.

And rounding out the top five things, we have a major new growth driver, Vera, the world's first CPU purpose built for agentic AI. Vera opens a brand new $200 billion TAM for NVIDIA, a market we have never addressed before. And every major hyperscaler and system maker is partnering with us to deploy it.

The world is rebuilding computing for agentic AI and robotic physical AI. NVIDIA sits at the center of these transitions. We built NVIDIA Compute platform over three decades, one architecture, vast ecosystem, extreme

co-design across chips, systems, networking, and software. We built it ahead of this moment, so that when agentic AI arrived, NVIDIA would be ready. It has arrived.

Look forward to catching up next time.

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