NVDA Earnings Call – FY2024 Q4

Date: February 21, 2024

Operator

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

Simona Jankowski

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the fourth quarter and fiscal 2024. With me today from NVIDIA are Jen-Hsun Huang, President and Chief Executive Officer, and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the first quarter of fiscal 2025. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent. During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent Forms 10-K and 10-Q and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, February 21, 2024, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements. During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website. With that let me turn the call over to Colette.

Colette Kress

Thanks, Simona. Q4 was another record quarter. Revenue of $22.1 billion was up 22% sequentially and up to 265% year-on-year and well above our outlook of $20 billion. For fiscal 2024, revenue was $60.9 billion and up 126% from the prior year. Starting with data center. Data center revenue for the fiscal 2024 year was $47.5 billion, more than tripling from the prior year. The world has reached the tipping point of new computing era. The $1 trillion installed base of data center infrastructure is rapidly transitioning from general purpose to accelerated computing. As Moore's Law slows while computing demand continues to skyrocket, companies may accelerate every workload possible to drive future improvement in performance, TCO and energy efficiency. At the same time, companies have started to build the next generation of modern data centers, what we refer to as AI factories, purpose built to refine raw data and produce valuable intelligence in the era of generative AI. In the fourth quarter, data center revenue of $18.4 billion was a record, up 27% sequentially and up 409% year-over-year, driven by the NVIDIA Hopper GPU computing platform along with InfiniBand end-to-end networking. Compute revenue grew more than 5x and networking revenue tripled from last year. We are delighted that supply of Hopper architecture products is improving. Demand for Hopper remains very strong. We expect our next-generation products to be supply constrained as demand far exceeds supply. Fourth quarter data center growth was driven by both training and inference of generative AI and large language models across a broad set of industries, use cases and regions. The versatility and leading performance of our data center platform enables a high return on investment for many use cases, including AI training and inference, data processing and a broad range of CUDA accelerated workloads. We estimate in the past year approximately 40% of data center revenue was for AI inference. Building and deploying AI solutions has reached virtually every industry. Many companies across industries are training and operating their AI models and services at scale, enterprises across NVIDIA AI infrastructure through cloud providers, including hyperscales, GPU specialized and private clouds or on-premise. NVIDIA's computing stack extends seamlessly across cloud and on-premise environments, allowing customers to deploy with a multi-cloud or hybrid-cloud strategy. In the fourth quarter, large cloud providers represented more than half of our data center revenue, supporting both internal workloads and external public cloud customers. Microsoft recently noted that more than 50,000 organizations use GitHub Copilot business to supercharge the productivity of their developers, contributing to GitHub revenue growth accelerating to 40% year-over-year. And Copilot for Microsoft 365 adoption grew faster in its first two months than the two previous major Microsoft 365 enterprise suite releases did. Consumer internet companies have been early adopters of AI and represent one of our largest customer categories. Companies from search to e-commerce, social media, news and video services and entertainment are using AI for deep learning-based recommendation systems. These AI investments are generating a strong return by improving customer engagement, ad conversation and click-throughs rates. Meta in its latest quarter cited more accurate predictions and improved advertiser performance as contributing to the significant acceleration in its revenue. In addition, consumer internet companies are investing in generative AI to support content creators, advertisers and customers through automation tools for content and ad creation, online product descriptions and AI shopping assistance. Enterprise software companies are applying generative AI to help customers realize productivity gains. Early customers we've partnered with for both training and inference of generative AI are already seeing notable commercial success. ServiceNow's generative AI products in their latest quarter drove their largest ever net new annual contract value contribution of any new product family release. We are working with many other leading AI and enterprise software platforms as well, including Adobe, Databricks, Getty Images, SAP and Snowflake. The field of foundation of large-language models is thriving. Anthropic, Google, Inflection, Microsoft, OpenAI and xAI are leading with continued amazing breakthrough in generative AI. Exciting companies like Adept, AI21, Character.ai, Cohere, Mistral, Perplexity and Runway are building platforms to serve enterprises and creators. New startups are creating LLMs to serve the specific languages, cultures and customs of the world many regions. And others are creating foundation models to address entirely different industries like Recursion Pharmaceuticals and Generate:Biomedicines for biology. These companies are driving demand for NVIDIA AI infrastructure through hyperscale or GPU specialized cloud providers. Just this morning, we announced that we've collaborated with Google to optimize its state-of-the art new Gemma language models to accelerate their inference performance on NVIDIA GPUs in the cloud data center and PC. One of the most notable trends over the past year is the significant adoption of AI by enterprises across the industry verticals such as automotive, healthcare and financial services. NVIDIA offers multiple application frameworks to help companies adopt AI in vertical domains such as autonomous driving, drug discovery, low latency machine learning for fraud detection or robotics, leveraging our full stack accelerated computing platform. We estimate the data center revenue contribution of the automotive vertical through the cloud or on-prem exceeded $1 billion last year. NVIDIA DRIVE infrastructure solutions includes systems and software for the development of autonomous driving, including data ingestion, creation, labeling and AI training, plus validation through simulation. Almost 80 vehicle manufacturers across global OEMs, new energy vehicles, trucking, robotaxi and Tier 1 suppliers are using NVIDIA's AI infrastructure to train LLMs and other AI models for automated driving and AI cockpit applications. And in fact, nearly every automotive company working on AI is working with NVIDIA. As AV algorithms move to video transformers and more cars are equipped with cameras, we expect NVIDIA's automotive data center processing demand to grow significantly. In healthcare, digital biology and generative AI are helping to reinvent drug discovery, surgery, medical imaging and wearable devices. We have built deep domain expertise in healthcare over the past decade, creating the NVIDIA Clara healthcare platform and NVIDIA BioNeMo, a generative AI service to develop, customize and deploy AI foundation models for computer-aided drug discovery. BioNeMo features a

growing collection of pre-trained Biomolecular AI models that can be applied to the end-to-end drug discovery processes. We announced Recursion is making available for their proprietary AI model through BioNeMo for the drug discovery ecosystem. In financial services, customers are using AI for a growing set of use cases from trading and risk management to customer service and fraud detection. For example, American Express improved fraud detection accuracy by 6% using NVIDIA AI. Shifting to our data center revenue by geography. Growth was strong across all regions, except for China where our data center revenue declined significantly following the U.S. government export control regulations imposed in October. Although we have not received licenses from the U.S. government to ship restricted products to China, we have started shipping alternatives that don't require a license for the China market. China represented a mid-single digit percentage of our data center revenue in Q4. And we expect it to stay in a similar range in the first-quarter. In regions outside of the U.S. and China, sovereign AI has become an additional demand driver. Countries around the world are investing in AI infrastructure to support the building of large-language models in their own language, on domestic data and in support of their local research and enterprise ecosystems. From a product perspective, the vast majority of revenue was driven by our Hopper architecture along with InfiniBand networking. Together, they have emerged as the de-facto standard for accelerated computing and AI infrastructure. We are on track to ramp H200 with initial shipments in the second quarter. Demand is strong as H200 nearly doubles the inference performance of H100. Networking exceeded a $13 billion annualized revenue run rate. Our end-to-end networking solutions define modern AI data centers. Our Quantum InfiniBand solutions grew more than 5x year on year. NVIDIA Quantum InfiniBand is the standard for the highest performance AI-dedicated infrastructures. We are now entering the ethernet networking space with the launch of our new Spectrum-X end-to-end offering designed for an AI-optimized networking for the data center. Spectrum-X introduces new technologies over ethernet, that are purpose built for AI. Technologies incorporated in our Spectrum switch, BlueField DPU and software stack deliver 1.6x higher networking performance for AI processing compared with traditional ethernet. Leading OEMs, including Dell, HPE, Lenovo and Super Micro, with their global sales channels, are partnering with us to expand our AI solution to enterprises worldwide. We are on track to ship Spectrum-X this quarter. We also made great progress with our software and services offerings, which reached an annualized revenue run rate of $1 billion in Q4. We announced that NVIDIA DGX Cloud will expand its list of partners to include Amazon's AWS, joining Microsoft Azure, Google Cloud and Oracle Cloud. DGX Cloud is used for NVIDIA's own AI R&D and custom model development as well as NVIDIA developers. It brings the CUDA ecosystem to NVIDIA CSP partners. Okay, moving to gaming. Gaming revenue was $2.87 billion, was flat sequentially and up 56% year on year, better than our outlook on solid consumer demand for NVIDIA GeForce RTX GPUs during the holidays. Fiscal year revenue of $10.45 billion was up 15%. At CES, we announced our GeForce RTX 40 Super Series family of GPUs. Starting at $599, they deliver incredible gaming performance and generative AI capabilities. Sales are off to a great start. NVIDIA AI Tensor cores and the GPUs deliver up to 836 AI tops, perfect for powering AI for gaming, creating an everyday productivity. The rich software stack we offer with our RTX GPUs further accelerates AI. With our DLSS technologies, seven out of eight pixels can be AI generated, resulting up to 4x faster ray tracing and better image quality. And with the Tensor RT LLM for Windows, our open-source library that accelerates inference performance for the latest large-language models generative AI can run up to 5X faster on RTX AI PCs. At CES, we also announced a wave of new RTX 40 Series AI laptops from every major OEMs. These bring high-performance gaming and AI capabilities to a wide range of form factors, including 14 inch and thin and light laptops. With up to 686 tops of AI performance, these next-generation AI PCs increase generative AI performance by up to 60x, making them the best-performing AI PC platforms. At CES, we announced NVIDIA Avatar Cloud Engine microservices, which allowed developers to integrate state-of-the-art generative AI models into digital avatars. ACE won several Best of CES 2024 awards. NVIDIA has an end-to-end platform for building and deploying generative AI applications for RTX PCs and workstations. This includes libraries, SDKs, tools and services developers can incorporate into their generative AI workloads. NVIDIA is fueling the next wave of generative AI applications coming to the PC. With over 100 million RTX PCs in the installed-base and over 500 AI-enabled PC applications and games, we are on our way. Moving to Pro Visualization. Revenue of $463 million was up 11% sequentially and up 105% year on year. Fiscal year revenue of $1.55 billion was up 1%. Sequential growth in the quarter was driven by a rich mix of RTX Ada architecture GPUs continuing to ramp. Enterprises are refreshing their workstations to support generative AI-related workloads, such as data preparation, LLM fine-tuning and retrieval augmented generation. These key industrial verticals driving demand include manufacturing, automotive and robotics. The automotive industry has also been an early adopter of NVIDIA Omniverse as it seeks to digitize work flows from design to build, simulate, operate and experience their factories and cars. At CES, we announced that creative partners and developers including Brickland, WPP and ZeroLight are building Omniverse-powered car configurators. Leading automakers like LOTUS are adopting the technology to bring new levels of personalization, realism and interactivity to the car buying experience. Moving to Automotive. Revenue was $281 million, up 8% sequentially and down 4% year on year. Fiscal year revenue of $1.09 billion was up 21%, crossing the $1 billion mark for the first time on continued adoption of the NVIDIA DRIVE platform by automakers. NVIDIA DRIVE Orin is the AI car computer of choice for software-defined AV fleets. Its successor, NVIDIA DRIVE Thor, designed for vision transformers often -- offers more AI performance and integrates a wide range of intelligent capabilities into a single AI compute platform, including autonomous driving and parking, driver and passenger monitoring and AI cockpit functionality and will be available next year. There were several automotive customer announcements this quarter, Li Auto, Great Wall Motor, ZEEKR, the premium EV subsidiary of Geely and Jeremy Xiaomi EV all announced new vehicles built on NVIDIA. Moving to the rest of the P&L. GAAP gross margins expanded sequentially to 76% and non-GAAP gross margins to 76.7% on strong data center growth and mix. Our gross margins in Q4 benefited from favorable component costs. Sequentially, GAAP operating expenses were up 6% and non-GAAP operating expenses were up 9%, primarily reflecting higher compute and infrastructure investments and employee growth. In Q4, we returned $2.8 billion to shareholders in the form of share repurchases and cash dividends. During fiscal year '24, we utilized cash of $9.9 billion towards shareholder returns, including $9.5 billion in share repurchases. Let me turn to the outlook for the first quarter. Total revenue is expected to be $24 billion, plus or minus 2%. We expect sequential growth in data center and proviz, partially offset by seasonal decline in gaming. GAAP and non-GAAP gross margins are expected to be 76.3% and 77% respectively, plus or minus 50 basis-points. Similar to Q4, Q1 gross margins are benefiting from favorable component costs. Beyond Q1, for the remainder of the year, we expect gross margins to return to the mid-70s percent range. GAAP and non-GAAP operating expenses are expected to be approximately $3.5 billion and $2.5 billion respectively. Fiscal year 2025 GAAP and non-GAAP operating expenses are expected to grow in the mid-30% range as we continue to invest in the large opportunities ahead of us. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $250 million, excluding gains and losses from non-affiliated investments. GAAP and non-GAAP tax rates are expected to be 17%, plus or minus 1% excluding any discrete items. Further financial details are included in the CFO commentary and other information available on our IR website. In closing, let me highlight some upcoming events for the financial community. We will attend the Morgan Stanley Technology and Media and Telecom Conference in San Francisco on March 4 and the TD Cowen's 44th Annual Healthcare Conference in Boston on March 5. And of course, please join us for our Annual DTC conference starting Monday March 18 in San Jose, California, to be held in-person for the first time in five years. DTC will kick off with Jen-Hsun's keynote and we will host a Q&A session for financial analysts the next day, March 19. At this time, we will now open the call for questions. Operator, would you please poll for questions?

Operator

[Operator Instructions] Your first question comes from the line of Toshiya Hari from Goldman Sachs. Your line is open.

Toshiya Hari

Hi. Thank you so much for taking the question and congratulations on the really strong results. My question is for Jen-Hsun on the data center business. Clearly, you're doing extremely well in the business. I'm curious how your expectations for calendar '24 and '25 have evolved over the past 90 days. And as you answer

the question, I was hoping you can touch on some of the newer buckets within data center, things like software. Sovereign AI, I think you've been pretty vocal about how to think about that medium-to-long term. And recently, there was an article about NVIDIA potentially participating in the ASIC market. Is there any credence to that, and if so, how should we think about you guys playing in that market over the next several years? Thank you.

Jensen Huang

Thanks, Toshiya. Let's see. There were three questions, one more time. First question was -- can you -- well?

Toshiya Hari

I guess your expectations for data center, how they've evolved. Thank you.

Jensen Huang

Okay. Yeah. Well, we guide one quarter at a time. But fundamentally, the conditions are excellent for continued growth calendar '24, to calendar '25 and beyond. And let me tell you why? We're at the beginning of two industry-wide transitions and both of them are industry wide. The first one is a transition from general to accelerated computing. General-purpose computing, as you know, is starting to run out of steam. And you can tell by the CSPs extending and many data centers, including our own for general-purpose computing, extending the depreciation from four to six years. There's just no reason to update with more CPUs when you can't fundamentally and dramatically enhance its throughput like you used to. And so you have to accelerate everything. This is what NVIDIA has been pioneering for some time. And with accelerated computing, you can dramatically improve your energy efficiency. You can dramatically improve your cost in data processing by 20 to 1. Huge numbers. And of course, the speed. That speed is so incredible that we enabled a second industry-wide transition called generative AI. Generative AI, I'm sure we're going to talk plenty -- plenty about it during the call. But remember, generative AI is a new application. It is enabling a new way of doing software, new types of software are being created. It is a new way of computing. You can't do generative AI on traditional general-purpose computing. You have to accelerate it. And the third is it is enabling a whole new industry, and this is something worthwhile to take a step back and look at and it connects to your last question about sovereign AI. A whole new industry in the sense that for the very first time a data center is not just about computing data and storing data and serving the employees of a company. We now have a new type of data center that is about AI generation, an AI generation factory. And you've heard me describe it as AI factories. But basically, it takes raw material, which is data, it transforms it with these AI supercomputers that NVIDIA builds, and it turns them into incredibly valuable tokens. These tokens are what people experience on the amazing ChatGPT or Midjourney or, search these days are augmented by that. All of your recommender systems are now augmented by that, the hyper-personalization that goes along with it. All of these incredible startups in digital biology, generating proteins and generating chemicals and the list goes on. And so all of these tokens are generated in a very specialized type of data center. And this data center we call AI supercomputers and AI generation factories. But we're seeing diversity -- one of the other reasons -- so at the foundation is that. The way it manifests into new markets is in all of the diversity that you're seeing us in. One, the amount of inference that we do is just off the charts now. Almost every single time you interact with ChatGPT, that we're inferencing. Every time you use Midjourney, we're inferencing. Every time you see amazing -- these Sora videos that are being generated or Runway, the videos that they're editing, Firefly, NVIDIA is doing inferencing. The inference part of our business has grown tremendously. We estimate about 40%. The amount of training is continuing, because these models are getting larger and larger, the amount of inference is increasing. But we're also diversifying into new industries. The large CSPs are still continuing to build out. You can see from their CapEx and their discussions, but there's a whole new category called GPU specialized CSPs. They specialize in NVIDIA AI infrastructure, GPU specialized CSPs. You're seeing enterprise software platforms deploying AI. ServiceNow is just a really, really great example. You see Adobe. There's the others, SAP and others. You see consumer Internet services that are now augmenting all of their services of the past with generative AI. So they can have even more hyper-personalized content to be created. You see us talking about industrial generative AI. Now our industries represent multi-billion dollar businesses, auto, health, financial services. In total, our vertical industries are multi-billion dollar businesses now. And of course sovereign AI. The reason for sovereign AI has to do with the fact that the language, the knowledge, the history, the culture of each region are different and they own their own data. They would like to use their data, train it with to create their own digital intelligence and provision it to harness that raw material themselves. It belongs to them, each one of the regions around the world. The data belongs to them. The data is most useful to their society. And so they want to protect the data. They want to transform it themselves, value-added transformation, into AI and provision those services themselves. So we're seeing sovereign AI infrastructure is being built in Japan, in Canada, in France, so many other regions. And so my expectation is that what is being experienced here in the United States, in the West, will surely be replicated around the world, and these AI generation factories are going to be in every industry, every company, every region. And so I think the last -- this last year, we've seen a generative AI really becoming a whole new application space, a whole new way of doing computing, a whole new industry is being formed and that's driving our growth.

Operator

Your next question comes from the line of Joe Moore from Morgan Stanley. Your line is open.

Joe Moore

Great. Thank you. I wanted to follow up on the 40% of revenues coming from inference. That's a bigger number than I expected. Can you give us some sense of where that number was maybe a year before, how much you're seeing growth around LLMs from inference? And how are you measuring that? Is that -- I assume it's in some cases the same GPUs you use for training and inference. How solid is that measurement? Thank you.

Jensen Huang

I'll go backwards. The estimate is probably understated. And -- but we estimated it. And let me tell you why. Whenever -- a year ago, the recommender systems that people are -- when you run the internet, the news, the videos, the music, the products that are being recommended to you because as you know, the internet has trillions -- I don't know how many trillions, but trillions of things out there and your phone is 3-inches square. And so the ability for them to fit all of that information down to something, such a small real estate, is through a system, an amazing system called recommender systems. These recommender systems used to be all based on CPU approaches. But the recent migration to deep learning and now generative AI has really put these recommender systems now directly into the path of GPU acceleration. It needs GPU acceleration for the embeddings. It needs GPU acceleration for the nearest neighbor search. It needs GPU acceleration for the re-ranking and it needs GPU acceleration to generate the augmented information for you. So GPUs are in every single step of a recommender system now. And as you know, recommender system is the single largest software engine on the planet. Almost every major company in the world has to run these large recommender systems. Whenever you use ChatGPT, it's being inferenced. Whenever you hear about Midjourney and just the number of things that they're generating for consumers, when you when you see Getty, the work that we do with Getty and Firefly from Adobe. These are all generative

models. The list goes on. And none of these, as I mentioned, existed a year ago, 100% new.

Operator

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

Stacy Rasgon

Hi, guys. Thanks for taking my question. I wanted Colette -- I wanted to touch on your comment that you expected the next generation of products -- I assume that meant Blackwell, to be supply constrained. Could you dig into that a little bit, what is the driver of that? Why does that get constrained as Hopper is easing up? And how long do you expect that to be constrained, like do you expect the next generation to be constrained like all the way through calendar '25, like when do those start to ease?

Jensen Huang

Yeah. The first thing is overall, our supply is improving, overall. Our supply chain is just doing an incredible job for us, everything from of course the wafers, the packaging, the memories, all of the power regulators, to transceivers and networking and cables and you name it. The list of components that we ship -- as you know, people think that NVIDIA GPUs is like a chip. But the NVIDIA Hopper GPU has 35,000 parts. It weighs 70 pounds. These things are really complicated things we've built. People call it an AI supercomputer for good reason. If you ever look in the back of the data center, the systems, the cabling system is mind boggling. It is the most dense complex cabling system for networking the world's ever seen. Our InfiniBand business grew 5x year over year. The supply chain is really doing fantastic supporting us. And so overall, the supply is improving. We expect the demand will continue to be stronger than our supply provides and -- through the year and we'll do our best. The cycle times are improving and we're going to continue to do our best. However, whenever we have new products, as you know, it ramps from zero to a very large number. And you can't do that overnight. Everything is ramped up. It doesn't step up. And so whenever we have a new generation of products -- and right now, we are ramping H200's. There is no way we can reasonably keep up on demand in the short term as we ramp. We're ramping Spectrum-X. We're doing incredibly well with Spectrum-X. It's our brand-new product into the world of ethernet. InfiniBand is the standard for AI-dedicated systems. Ethernet with Spectrum-X --ethernet is just not a very good scale-out system. But with Spectrum-X, we've augmented, layered on top of ethernet, fundamental new capabilities like adaptive routing, congestion control, noise isolation or traffic isolation, so that we could optimize ethernet for AI. And so InfiniBand will be our AI-dedicated infrastructure. Spectrum-X will be our AI-optimized networking and that is ramping, and so we'll -- with all of the new products, demand is greater than supply. And that's just kind of the nature of new products and so we work as fast as we can to capture the demand. But overall, overall net-net, overall, our supply is increasing very nicely.

Operator

Your next question comes from the line of Matt Ramsay from TD Cowen. Your line is open.

Matt Ramsay

Good afternoon, Jensen, Colette. Congrats on the results. I wanted to ask I guess a two-part question, and it comes at what Stacy was just getting out on your demand being significantly more than your supply, even though supply is improving. And I guess the two sides of the question are, I guess, first for Colette, like how are you guys thinking about allocation of product in terms of customer readiness to deploy and sort of monitoring if there's any kind of build-up of product that might not yet be turned on? And then I guess Jen-Hsun, for you, I'd be really interested to hear you speak a bit about the thought that you and your company are putting into the allocation of your product across customers, many of which compete with each other, across industries to smaller startup companies, to things in the healthcare arena to government. It's a very, very unique technology that you're enabling and I'd be really interested to hear you speak a bit about how you think about quote/unquote fairly allocating sort of for the good of your company, but also for the good of the industry. Thanks.

Colette Kress

Let me first start with your question, thanks, about how we are working with our customers as they look into how they are building out their GPU instances and our allocation process. The folks that we work with, our customers that we work with, have been partners with us for many years as we have been assisting them both in what they set up in the cloud, as well as what they are setting up internally. Many of these providers have multiple products going at one time to serve so many different needs across their end customers but also what they need internally. So they are working in advance, of course, thinking about those new clusters that they will need. And our discussions with them continue not only on our Hopper architecture, but helping them understand the next wave and getting their interest and getting their outlook for the demand that they want. So it's always a moving process in terms of what they will purchase, what is still being built and what is in use for our end customers. But the relationships that we've built and their understanding of the sophistication of the build has really helped us with that allocation and both helped us with our communications with them.

Jensen Huang

First, our CSPs have a very clear view of our product road map and transitions. And that transparency with our CSPs gives them the confidence of which products to place and where and when. And so they know their -- they know the timing to the best of our ability. And they know quantities and of course allocation. We allocate fairly. We allocate fairly. We do the best of our -- do the best we can to allocate fairly and to avoid allocating unnecessarily. As you mentioned earlier, why allocate something when the data center's not ready. Nothing is more difficult then to have anything sit around. And so, allocate fairly, and to avoid allocating unnecessarily. And where we do -- the question that you asked about the end markets, that we have an excellent ecosystem with OEMs, ODMs, CSPs and, very importantly, end markets. What NVIDIA is really unique about is that we bring our customers, we bring our partners, CSPs and OEMs, we bring them customers. The biology companies, the healthcare companies, financial services companies, AI developers, large-language model developers, autonomous vehicle companies, robotics companies. There's just a giant suite of robotics companies that are emerging. There are warehouse robotics to surgical robotics to humanoid robotics, all kinds of really interesting robotics companies, agriculture robotics companies. All of these startups, large companies, healthcare, financial services and auto and such are working on NVIDIA's platform. We support them directly. And oftentimes, we can have a twofer by allocating to a CSP and bringing the customer to the CSP at the same time. And so this ecosystem, you're absolutely right that it's vibrant. But at the core of it, we want to allocate fairly with avoiding waste and looking for opportunities to connect partners and end users. We're looking for those opportunities all the time.

Operator

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

Timothy Arcuri

Thanks a lot. I wanted to ask about how you're converting backlog into revenue. Obviously, lead times for your products have come down quite a bit. Colette, you didn't talk about the inventory purchase commitments. But if I sort of add up your inventory plus the purchase commits and your prepaid supply, sort of the aggregate of your supply, it was actually down a touch. How should we read that? Is that just you saying that you don't need to make as much of a financial commitment to your suppliers because the lead times are lower or is that maybe you're reaching some sort of steady state where you're closer to filling your order book and your backlog? Thanks.

Colette Kress

Yeah. So let me, highlight on those three different areas of how we look at our suppliers. You're correct. Our inventory on hand given our allocation that we're on, we're trying to, as things come into inventory, immediately work to ship them to our customers. I think our customer appreciates our ability to meet the schedules that we've looked for. The second piece of it is our purchase commitments. Our purchase commitments have many different components into it, components that we need for manufacturing. But also, often we are procuring capacity that we may need. The length of that need for capacity or the length for the components are all different. Some of them may be for the next two quarters, but some of them may be for multiple years. I can say the same regarding our prepaids. Our prepaids are pre-designed to make sure that we have the reserve capacity that we need at several of our manufacturing suppliers as we look forward. So wouldn't read into anything regarding approximately about the same numbers as we are increasing our supply. All of them just have different lengths as we have sometimes had to buy things in long-lead times or things that needed capacity to be built for us.

Operator

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

Ben Reitzes

Yeah. Thanks. Congratulations on the results. Colette, I wanted to talk about your comment regarding gross margins and that they should go back to the mid-70s. If you don't mind unpacking that. And also, is that due to the HBM content in the new products and what do you think are the drivers of that comment? Thanks so much.

Colette Kress

Yeah. Thanks for the question. We highlighted in our opening remarks really about our Q4 results and our outlook for Q1. Both of those quarters are unique. Those two quarters are unique in their gross margin as they include some benefit from favorable component cost in the supply chain kind of across both our compute and networking and also in several different stages of our manufacturing process. So looking forward, we have visibility into a mid-70s gross margin for the rest of the fiscal year, taking us back to where we were before this Q4 and Q1 peak that we've had here. So we're really looking at just a balance of our mix. Mix is always going to be our largest driver of what we will be shipping for the rest of the year. And those are really just the drivers.

Operator

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

C.J. Muse

Yeah. Good afternoon, and thank you for taking the question. Bigger picture question for you, Jen-Hsun. When you think about the million-x improvement in GPU compute over the last decade and expectations for similar improvements in the next, how do your customers think about the long-term usability of their NVIDIA investments that they're making today? Do today's training clusters become tomorrow's inference clusters? How do you see this playing out? Thank you.

Jensen Huang

Hey, CJ. Thanks for the question. Yeah, that's the really cool part. If you look at the reason why we're able to improve performance so much, it's because we have two characteristics about our platform. One, is that it's accelerated. And two, it's programmable. It's not brittle. NVIDIA is the only architecture that has gone from the very, very beginning, literally the very beginning when CNN's and Alex Krizhevsky and Ilya Sutskever and Geoff Hinton first revealed AlexNet, all the way through to RNNs to LSTMs to every -- RLs to deep learning RLs to transformers to every single version. Every single version and every species that have come along, vision transformers, multi-modality transformers, every single -- and now time sequence stuff, and every single variation, every single species of AI that has come along, we've been able to support it, optimize our stack for it and deploy it into our installed base. This is really the great amazing part. On the one hand, we can invent new architectures and new technologies like our Tensor cores, like our transformer engine for Tensor cores, improved new numerical formats and structures of processing like we've done with the different generations of Tensor cores, meanwhile, supporting the installed base at the same time. And so, as a result, we take all of our new software algorithm invest -- inventions, all of the inventions, new inventions of models of the industry, and it runs on our installed base on the one hand. On the other hand, whenever we see something revolutionary we can -- like transformers, we can create something brand new like the Hopper transformer engine and implement it into future. And so we simultaneously have this ability to bring software to the installed base and keep making it better and better and better, so our customers installed base is enriched over time with our new software. On the other hand, for new technologies, create revolutionary capabilities. Don't be surprised if in our future generation, all of a sudden amazing breakthroughs in large-language models were made possible And those breakthroughs, some of which will be in software because they run CUDA, will be made available to the installed base. And so we carry everybody with us on the one hand. We make giant breakthroughs on the other hand.

Operator

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

Aaron Rakers

Yeah. Thanks for taking the question. I wanted to ask about the China business. I know that in your prepared comments you said that you started shipping some alternative solutions into China. You also put it out that you expect that contribution to continue to be about a mid-single digit percent of your total data center business. So I guess the question is what is the extent of products that you're shipping today into the China market and why should we not expect that maybe other alternative solutions come to the market and expand your breadth to participate in that in that opportunity again? Thank you.

Jensen Huang

Think of, at the core, remember the US government wants to limit the latest capabilities of NVIDIA's accelerated computing and AI to the Chinese market. And the U.S. government would like to see us be as successful in China as possible. Within those two constraints, within those two pillars if you will, are the restrictions, and so we had to pause when the new restrictions came out. We immediately paused. So that we understood what the restrictions are, reconfigured our products in a way that is not software hackable in any way. And that took some time. And so we reset -- we reset our product offering to China and now we're sampling to customers in China. And we're going to do our best to compete in that marketplace and succeed in that marketplace within the -- within the specifications of the restriction. And so that's it. We -- this last quarter, we -- our business significantly declined as we -- as we paused in the marketplace. We stopped shipping in the marketplace. We expect this quarter to be about the same. But after, that hopefully we can go compete for our business and do our best, and we'll see how it turns out.

Operator

Your next question comes from the line of Harsh Kumar from Piper Sandler. Your line is open.

Harsh Kumar

Yeah. Hey, Jen-Hsun, Colette and NVIDIA team. First of all, congratulations on a stunning quarter and guide. I wanted to talk about, a little bit about your software business and it's pleasing to hear that it's over a $1 billion but I was hoping Jen-Hsun or Colette if you could just help us understand what the different parts and pieces are for the software business? In other words, just help us unpack it a little bit, so we can get a better understanding of where that growth is coming from.

Jensen Huang

Let me take a step back and explain the fundamental reason why NVIDIA will be very successful in software. So first, as you know, accelerated computing really grew in the cloud. In the cloud, the cloud service providers have really large engineering teams and we work with them in a way that allows them to operate and manage their own business. And whenever there are any issues, we have large teams assigned to them. And their engineering teams are working directly with our engineering teams and we enhance, we fix, we maintain, we patch the complicated stack of software that's involved in accelerated computing. As you know, accelerated computing is very different than general-purpose computing. You're not starting from a program like C++. You compile it and things run on all your CPUs. The stacks of software necessary for every domain from data processing SQL versus -- SQL structure data versus all the images and text and PDF, which is unstructured, to classical machine-learning to computer vision to speech to large-language models, all --recommender systems. All of these things require different software stacks. That's the reason why NVIDIA has hundreds of libraries. If you don't have software, you can't open new markets. If you don't have software, you can't open and enable new applications. Software is fundamentally necessary for accelerated computing. This is the fundamental difference between accelerated computing and general-purpose computing that most people took a long time to understand. And now, people understand that the software is really key. And the way that we work with CSPs, that's really easy. We have large teams that are working with their large teams. However, now that generative AI is enabling every enterprise and every enterprise software company to embrace accelerated computing -- and when -- it is now essential to embrace accelerated computing because it is no longer possible, no longer likely anyhow to sustain improved throughput through just general-purpose computing. All of these enterprise software companies and enterprise companies don't have large engineering teams to be able to maintain and optimize their software stack to run across all of the world's clouds and private clouds and on-prem. So we are going to do the management, the optimization, the patching, the tuning, the installed-base optimization for all of their software stacks. And we containerize them into our stack. We call it NVIDIA AI Enterprise. And the way we go to market with it is that think of that NVIDIA AI Enterprise now as a run time like an operating system, it's an operating system for artificial intelligence. And we charge $4,500 per GPU per year. And my guess is that every enterprise in the world, every software enterprise company that are deploying software in all the clouds and private clouds and on-prem, will run on NVIDIA AI Enterprise, especially obviously for our GPUs. And so this is going to likely be a very significant business over time. We're off to a great start. And Colette mentioned that it's already at $1 billion run rate and we're really just getting started.

Operator

Thank you. I will now turn the call back over to Jen-Hsun Huang, CEO, for closing remarks.

Jensen Huang

The computer industry is making two simultaneous platform shifts at the same time. The trillion-dollar installed base of data centers is transitioning from general purpose to accelerated computing. Every data center will be accelerated so the world can keep up with the computing demand, with increasing throughput, while managing costs and energy. The incredible speed up of NVIDIA enabled -- that NVIDIA enabled, a whole new computing paradigm, generative AI, where software can learn, understand and generate any information from human language to the structure of biology and the 3D world. We are now at the beginning of a new industry where AI-dedicated data centers process massive raw data to refine it into digital intelligence. Like AC power generation plants of the last industrial revolution, NVIDIA AI supercomputers are essentially AI generation factories of this Industrial Revolution. Every company in every industry is fundamentally built on their proprietary business intelligence, and in the future, their proprietary generative AI. Generative AI has kicked off a whole new investment cycle to build the next trillion dollars of infrastructure of AI generation factories. We believe these two trends will drive a doubling of the world's data center infrastructure installed base in the next five years and will represent an annual market opportunity in the hundreds of billions. This new AI infrastructure will open up a whole new world of applications not possible today. We started the AI journey with the hyperscale cloud providers and consumer internet companies. And now, every industry is on board, from automotive to healthcare to financial services, to industrial to telecom, media and entertainment. NVIDIA's full stack computing platform with industry-specific applications frameworks and a huge developer and partner ecosystem, gives us the speed, scale and reach to help every company -- to help companies in every industry become an AI company. We have so much to share with you at next month's GTC in San Jose. So be sure to join us. We look forward to updating you on our progress next quarter.

Operator

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