NVDA Earnings Call – FY2024 Q3

Date: November 21, 2023

Operator

Good afternoon. My name is JL, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Third 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] Simona Jankowski, you may now begin your conference.

Simona Jankowski

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2024. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the fourth quarter and fiscal 2024. 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 statements are made as of today, November 21, 2023, 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. Q3 was another record quarter. Revenue of $18.1 billion was up 34% sequentially and up more than 200% year-on-year and well above our outlook for $16 billion. Starting with Data Center. The continued ramp of the NVIDIA HGX platform based on our Hopper Tensor Core GPU architecture, along with InfiniBand end-to-end networking, drove record revenue of $14.5 billion, up 41% sequentially and up 279% year-on-year. NVIDIA HGX with InfiniBand together are essentially the reference architecture for AI supercomputers and data center infrastructures. Some of the most exciting generative AI applications are built and run on NVIDIA, including Adobe Firefly, ChatGPT, Microsoft 365 Copilot, CoAssist, now assist with ServiceNow and Zoom AI Companion. Our Data Center compute revenue quadrupled from last year and networking revenue nearly tripled. Investments in infrastructure for training and inferencing large language models, deep learning, recommender systems and generative AI applications is fueling strong broad-based demand for NVIDIA accelerated computing. Inferencing is now a major workload for NVIDIA AI computing. Consumer Internet companies and enterprises drove exceptional sequential growth in Q3, comprising approximately half of our Data Center revenue and outpacing total growth. Companies like Meta are in full production with deep learning, recommender systems and also investing in generative AI to help advertisers optimize images and text. Most major consumer Internet companies are racing to ramp up generative AI deployment. The enterprise wave of AI adoption is now beginning. Enterprise software companies such as Adobe, Databricks, Snowflake and ServiceNow are adding AI copilots and the systems to their platforms. And broader enterprises are developing custom AI for vertical industry applications such as Tesla in autonomous driving. Cloud service providers drove roughly the other half of our Data Center revenue in the quarter. Demand was strong from all hyperscale CSPs, as well as from a broadening set of GPU-specialized CSPs globally that are rapidly growing to address the new market opportunities in AI. NVIDIA H100 Tensor Core GPU instances are now generally available in virtually every cloud with instances in high demand. We have significantly increased supply every quarter this year to meet strong demand and expect to continue to do so next year. We will also have a broader and faster product launch cadence to meet the growing and diverse set of AI opportunities. Towards the end of the quarter, the U.S. government announced a new set of export control regulations for China and other markets, including Vietnam and certain countries in the Middle East. These regulations require licenses for the export of a number of our products, including our Hopper and Ampere 100 and 800 series and several others. Our sales to China and other affected destinations derived from products that are now subject to licensing requirements have consistently contributed approximately 20% to 25% of Data Center revenue over the past few quarters. We expect that our sales to these destinations will decline significantly in the fourth quarter. So we believe will be more than offset by strong growth in other regions. The U.S. government designed the regulation to allow the U.S. industry to provide data center compute products to markets worldwide, including China. Continuing to compete worldwide as the regulations encourage, promotes U.S. technology leadership, spurs economic growth and supports U.S. jobs. For the highest performance levels, the government requires licenses. For lower performance levels, the government requires a streamlined prior notification process. And for products even lower performance levels, the government does not require any notice at all. Following the government's clear guidelines, we are working to expand our Data Center product portfolio to offer compliance solutions for each regulatory category, including products for which the U.S. government does not wish to have advance notice before each shipment. We are working with some customers in China and the Middle East to pursue licenses from the U.S. government. It is too early to know whether these will be granted for any significant amount of revenue. Many countries are awakening to the need to invest in sovereign AI infrastructure to support economic growth and industrial innovation. With investments in domestic compute capacity, nations can use their own data to train LLMs and support their local generative AI ecosystems. For example, we are working with India's government and largest tech companies including Infosys, Reliance and Tata to boost their sovereign AI infrastructure. And French private cloud provider, Scaleway, is building a regional AI cloud based on NVIDIA H100 InfiniBand and NVIDIA's AI Enterprise software to fuel advancement across France and Europe. National investment in compute capacity is a new economic imperative and serving the sovereign AI infrastructure market represents a multi-billion dollar opportunity over the next few years. From a product perspective, the vast majority of revenue in Q3 was driven by the NVIDIA HGX platform based on our Hopper GPU architecture with lower contribution from the prior generation Ampere GPU architecture. The new L40S GPU built for industry standard servers began to ship, supporting training and inference workloads across a variety of consumers. This was also the first revenue quarter of our GH200 Grace Hopper Superchip, which combines our ARM-based Grace CPU with a Hopper GPU. Grace and Grace Hopper are ramping into a new multi-billion dollar product line. Grace Hopper instances are now available at GPU specialized cloud providers, and coming soon to Oracle Cloud. Grace Hopper is also getting significant traction with supercomputing customers. Initial shipments to Los Alamos National Lab and the Swiss National Supercomputing Center took place in the third quarter. The UK government announced it will build one of the world's fastest AI supercomputers called Isambard-AI with almost 5,500 Grace Hopper Superchips. German supercomputing center, Julich, also announced that it will build its next-generation AI supercomputer with close to 24,000 Grace Hopper Superchips and Quantum-2 InfiniBand, making it the world's most powerful AI supercomputer with over 90 exaflops of AI performance. All-in, we estimate that the combined AI compute capacity of all the supercomputers built on Grace Hopper across the U.S., Europe and Japan next year will exceed 200 exaflops with more wins to come. Inference is contributing significantly to our data center demand, as AI is now in full production for deep learning, recommenders, chatbots, copilots and text to image generation and this is just the beginning. NVIDIA AI

offers the best inference performance and versatility, and thus the lower power and cost of ownership. We are also driving a fast cost reduction curve. With the release of TensorRT-LLM, we now achieved more than 2x the inference performance for half the cost of inferencing LLMs on NVIDIA GPUs. We also announced the latest member of the Hopper family, the H200, which will be the first GPU to offer HBM3e, faster, larger memory to further accelerate generative AI and LLMs. It moves inference speed up to another 2x compared to H100 GPUs for running LLMs like Norma2 (ph). Combined, TensorRT-LLM and H200, increased performance or reduced cost by 4x in just one year. With our customers changing their stack, this is a benefit of CUDA and our architecture compatibility. Compared to the A100, H200 delivers an 18x performance increase for inferencing models like GPT-3, allowing customers to move to larger models and with no increase in latency. Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud will be among the first CSPs to offer H200-based instances starting next year. At last week's Microsoft Ignite, we deepened and expanded our collaboration with Microsoft across the entire stock. We introduced an AI foundry service for the development and tuning of custom generative AI enterprise applications running on Azure. Customers can bring their domain knowledge and proprietary data and we help them build their AI models using our AI expertise and software stock in our DGX cloud, all with enterprise grade security and support. SAP and Amdocs are the first customers of the NVIDIA AI foundry service on Microsoft Azure. In addition, Microsoft will launch new confidential computing instances based on the H100. The H100 remains the top performing and most versatile platform for AI training and by a wide margin, as shown in the latest MLPerf industry benchmark results. Our training cluster included more than 10,000 H100 GPUs or 3x more than in June, reflecting very efficient scaling. Efficient scaling is a key requirement in generative AI, because LLMs are growing by an order of magnitude every year. Microsoft Azure achieved similar results on a nearly identical cluster, demonstrating the efficiency of NVIDIA AI in public cloud deployments. Networking now exceeds a $10 billion annualized revenue run rate. Strong growth was driven by exceptional demand for InfiniBand, which grew fivefold year-on-year. InfiniBand is critical to gaining the scale and performance needed for training LLMs. Microsoft made this very point last week, highlighting that Azure uses over 29,000 miles of InfiniBand cabling, enough to circle the globe. We are expanding NVIDIA networking into the Ethernet space. Our new Spectrum-X end-to-end Ethernet offering with technologies, purpose built for AI, will be available in Q1 next year. With support from leading OEMs, including Dell, HPE and Lenovo. Spectrum-X can achieve 1.6x higher networking performance for AI communication compared to traditional Ethernet offerings. Let me also provide an update on our software and services offerings, where we are starting to see excellent adoption. We are on track to exit the year at an annualized revenue run rate of $1 billion for our recurring software, support and services offerings. We see two primary opportunities for growth over the intermediate term with our DGX cloud service and with our NVIDIA AI Enterprise software, each reflects the growth of enterprise AI training and enterprise AI inference, respectively. Our latest DGX cloud customer announcement was this morning as part of an AI research collaboration with Gentech, the biotechnology pioneer also plans to use our BioNeMo LLM framework to help accelerate and optimize their AI drug discovery platform. We now have enterprise AI partnership with Adobe, Dropbox, Getty, SAP, ServiceNow, Snowflake and others to come. Okay. Moving to Gaming. Gaming revenue of $2.86 billion was up 15% sequentially and up more than 80% year-on-year with strong demand in the important back-to-school shopping season with NVIDIA RTX ray tracing and AI technology now available at price points as low as $299. We entered the holidays with the best-ever line-up for gamers and creators. Gaming has doubled relative to pre-COVID levels even against the backdrop of lackluster PC market performance. This reflects the significant value we've brought to the gaming ecosystem with innovations like RTX and DLSS. The number of games and applications supporting these technologies has exploded in that period, driving upgrades and attracting new buyers. The RTX ecosystem continues to grow. There are now over 475 RTX-enabled games and applications. Generative AI is quickly emerging as the new pillar app for high performance PCs. NVIDIA RTX GPUs to find the most performance AI PCs and workstations. We just released TensorRT-LLM for Windows, which speeds on-device LLM inference up by 4x. With an installed base of over 100 million, NVIDIA RTX is the natural platform for AI application developers. Finally, our GeForce NOW cloud gaming service continues to build momentum. Its library of PC games surpassed 1,700 titles, including the launches of Alan Wake 2, Baldur's Gate 3, Cyberpunk 2077: Phantom Liberty and Starfield. Moving to the Pro Vis. Revenue of $416 million was up 10% sequentially and up 108% year-on year. NVIDIA RTX is the workstation platform of choice for professional design, engineering and simulation use cases and AI is emerging as a powerful demand driver. Early applications include inference for AI imaging in healthcare and edge AI in smart spaces and the public sector. We launched a new line of desktop workstations based on NVIDIA RTX Ada Lovelace generation GPUs and ConnectX, SmartNICs offering up to 2x the AI processing ray tracing and graphics performance of the previous generations. These powerful new workstations are optimized for AI workloads such as fine tune AI models, training smaller models and running inference locally. We continue to make progress on Omniverse, our software platform for designing, building and operating 3D virtual worlds. Mercedes-Benz is using Omniverse powered digital twins to plan, design, build and operate its manufacturing and assembly facilities, helping it increase efficiency and reduce defects. Oxxon (ph) is also incorporating Omniverse into its manufacturing process, including end-to-end simulation for the entire robotics and automation pipeline, saving time and cost. We announced two new Omniverse Cloud services for automotive digitalization available on Microsoft Azure, a virtual factory simulation engine and autonomous vehicle simulation engine. Moving to Automotive. Revenue was $261 million, up 3% sequentially and up 4% year-on year, primarily driven by continued growth in self-driving platforms based on NVIDIA DRIVE Orin SOC and the ramp of AI cockpit solutions with global OEM customers. We extended our automotive partnership of Foxconn to include NVIDIA DRIVE for our next-generation automotive SOC. Foxconn has become the ODM for EVs. Our partnership provides Foxconn with a standard AV sensor and computing platform for their customers to easily build a state-of-an-art safe and secure software defined car. Now we're going to move to the rest of the P&L. GAAP gross margin expanded to 74% and non-GAAP gross margin to 75%, driven by higher Data Center sales and lower net inventory reserve, including a 1 percentage point benefit from the release of previously reserved inventory related to the Ampere GPU architecture products. Sequentially, GAAP operating expenses were up 12% and non-GAAP operating expenses were up 10%, primarily reflecting increased compensation and benefits. Let me turn to the fourth quarter of fiscal 2024. Total revenue is expected to be $20 billion, plus or minus 2%. We expect strong sequential growth to be driven by Data Center, with continued strong demand for both compute and networking. Gaming will likely decline sequentially as it is now more aligned with notebook seasonality. GAAP and non-GAAP gross margins are expected to be 74.5% and 75.5%, respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $3.17 billion and $2.2 billion, respectively. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $200 million, excluding gains and losses from non-affiliated investments. GAAP and non-GAAP tax rates are expected to be 15%, plus or minus 1% excluding any discrete items. Further financial information 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 UBS Global Technology Conference in Scottsdale, Arizona, on November 28; the Wells Fargo TMT Summit in Rancho Palos Verdes, California on November 29; the Arete Virtual Tech Conference on December 7; and the J.P. Morgan Health Care Conference in San Francisco on January 8. Our earnings call to discuss the results of our fourth quarter and fiscal 2024 is scheduled for Wednesday, February 21. We will now open the call for questions. Operator, will you please poll for questions.

Operator

[Operator Instructions] Your first question comes from the line of Vivek Arya of Bank of America. Your line is open.

Vivek Arya

Thanks for taking my question. Just, Colette, wanted to clarify what China contributions are you expecting in Q4. And then, Jensen, the main question is for you, where do you think we are in the adoption curve in terms of your shipments into the generative AI market? Because when I just look at the trajectory of your data center, is growth -- it will be close to nearly 30% of all the spending in data center next year. So what metrics are you keeping an eye on to inform you that you

can continue to grow? Just where are we in the adoption curve of your products into the generative AI market? Thank you.

Colette Kress

So, first let me start with your question, Vivek, on export controls and the impacts that we are seeing in our Q4 outlook and guidance that we provided. We had seen historically over the last several quarters that China and some of the other impacted destinations to be about 20% to 25% of our Data Center revenue. We are expecting in our guidance for that to decrease substantially as we move into Q4. The export controls will have a negative effect on our China business. And we do not have good visibility into the magnitude of that impact even over the long-term. We are though working to expand our Data Center product portfolio to possibly offer new regulation compliance solutions that do not require a license, these products, they may become available in the next coming months. However, we don't expect their contribution to be material or meaningful as a percentage of the revenue in Q4.

Jensen Huang

Generative AI is the largest TAM expansion of software and hardware that we've seen in several decades. At the core of it, what's really exciting is that, what was largely a retrieval based computing approach, almost everything that you do is retrieved off of storage somewhere, has been augmented now, added with a generative method. And it's changed almost everything. You could see that text-to-text, text-to-image, text-to-video, text-to-3D, text-to-protein, text-to-chemicals, these were things that were processed and typed in by humans in the past. And these are now generative approaches. The way that we access data is changed. It used to be based on explicit queries. It is now based on natural language queries, intention queries, semantic queries. And so, we're excited about the work that we're doing with SAP and Dropbox and many others that you're going to hear about. And one of the areas that is really impactful is the software industry, which is about $1 trillion or so, has been building tools that are manually used over the last couple of decades. And now there's a whole new segment of software called copilots and assistants. Instead of manually used, these tools will have copilots to help you use it. And so, instead of licensing software, we will continue to do that, of course, but we will also hire copilots and assistants to help us use these -- use the software. We'll connect all of these copilots and assistants into teams of AIs, which is going to be the modern version of software, modern version of enterprise business software. And so the transformation of software and the way that software has done is driving the hardware underneath. And you can see that it's transforming hardware in two ways. One is something that's largely independent of generative AI. There's two trends: one is related to accelerated computing, general purpose computing is too wasteful of energy and cost. And now that we have much, much better approaches, call it, accelerated computing, you could save an order of magnitude of energy, you can save an order of magnitude of time or you can save an order of magnitudes of cost by using acceleration. And so, accelerated computing is transitioning, if you will, general purpose computing into this new approach. And that's been augmented by a new class of data centers. This is the traditional data centers that you were just talking about where we represent about a third of that. But there is a new class of data centers and this new class of data centers, unlike the data centers of the past, where you have a lot of applications running used by a great many people that are different tenants that are using the same infrastructure and that data center stores a lot of files. These new data centers are very few applications, if not one application, used by basically one tenant and it processes data, it trains models and then generates tokens and generates AI. And we call these new data centers AI factories. We're seeing AI factories being built out everywhere, and just by every country. And so if you look at the way where we are in the expansion, the transition into this new computing approach, the first wave you saw with large language model start-ups, generative AI start-ups and consumer Internet companies, and weren't in the process of ramping that. Meanwhile, while that's being ramped, you see that we're starting to partner with enterprise software companies who would like to build chatbots and copilots and assistants to augment the tools that they have on their platforms. You're seeing GPU specialized CSPs cropping up all over the world and they are dedicated to do really one thing, which is processing AI. You're seeing sovereign AI infrastructures, people -- countries that now recognize that they have to utilize their own data, keep their own data, keep their own culture, process that data and develop their own AI. You see that in India. Several -- about a year ago in Sweden, you are seeing in Japan. Last week, a big announcement in France. But the number of sovereign AI clouds that are being built is really quite significant. And my guess is that almost every major region will have and surely every major country will have their own AI clouds. And so I think you're seeing just new developments as the generative AI wave propagates through every industry, every company, every region. And so we're at the beginning of this inflection, this computing transition.

Operator

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

Aaron Rakers

Yeah. Thanks for taking the question. I wanted to ask about kind of the networking side of the business. Given the growth rates that you've now cited, I think, it's 155% year-over-year and strong growth sequentially, it looks like that business is like almost approaching $2.5 billion to $3 billion quarterly level. I'm curious of how you see Ethernet involved evolving and maybe how you would characterize your differentiation of Spectrum-X relative to the traditional Ethernet stack as we start to think about that becoming part of the networking narrative above and maybe beyond just InfiniBand as we look into next year? Thank you.

Jensen Huang

Yeah. Thanks for the question. Our networking business is already on a $10 billion plus run rate and it's going to get much larger. And as you mentioned, we added a new networking platform to our networking business recently. The vast majority of the dedicated large scale AI factories standardize on InfiniBand. And the reason for that is not only because of its data rate and not only just the latency, but the way that it moves traffic around the network is really important. The way that you process AI and a multi-tenant hyperscale Ethernet environment, the traffic pattern is just radically different. And with InfiniBand and with software defined networks, we could do congestion control, adaptive routing, performance isolation and noise isolation, not to mention, of course, the day rate and the low latency that -- and a very low overhead of InfiniBand that's natural part of InfiniBand. And so, InfiniBand is not so much just the network, it's also a computing fabric. We've put a lot of software-defined capabilities into the fabric including computation. We will do 40-point calculations and computation right on the switch, and right in the fabric itself. And so that's the reason why that difference in Ethernet versus InfiniBand or InfiniBand versus Ethernet for AI factories is so dramatic. And the difference is profound. And the reason for that is because you've just invested in a $2 billion infrastructure for AI factories. A 20%, 25%, 30% difference in overall effectiveness, especially as you scale up is measured in hundreds of millions of dollars of value. And if you will, renting that infrastructure over the course of four to five years, it really, really adds up. And so InfiniBand's value proposition is undeniable for AI factories. However, as we move AI into enterprise. This is enterprise computing what we'd like to enable every company to be able to build their own custom AIs. We're building customer AIs in our company based on our proprietary data, our proprietary type of skills. For example, recently we spoke about one of the models that we're creating, it's called ChipNeMo; we're building many others. There'll be tens, hundreds of custom AI models that we create inside our company. And our company is -- for all of our employee use, doesn't have to be as high performance as the AI factories we used to train the models. And so we would like the AI to be able to run in Ethernet environment. And so what we've done is we invented this new platform that extends Ethernet; doesn't replace Ethernet, it's 100% compliant with Ethernet. And it's optimized for East-West

traffic, which is where the computing fabric is. It adds to Ethernet with an end-to-end solution with Bluefield, as well as our Spectrum switch that allows us to perform some of the capabilities that we have in InfiniBand, not all but some. And we achieved excellent results. And the way we go to market is we go to market with our large enterprise partners who already offer our computing solution. And so, HP, Dell and Lenovo has the NVIDIA AI stack, the NVIDIA AI Enterprise software stack and now they integrate with Bluefield, as well as bundle -- take a market there, Spectrum switch, and they'll be able to offer enterprise customers all over the world with their vast sales force and vast network of resellers a fully integrated, if you will, fully optimized, at least end-to-end AI solution. And so that's basically it, bringing AI to Ethernet for the world's enterprise.

Operator

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

Joseph Moore

Great. Thank you. I'm wondering if you could talk a little bit more about Grace Hopper and how you see the ability to leverage kind of the microprocessor, how you see that as a TAM expander. And what applications do you see using Grace Hopper versus more traditional H100 applications?

Jensen Huang

Yeah. Thanks for the question. Grace Hopper is in production -- in high volume production now. We're expecting next year just with all of the design wins that we have in high performance computing and AI infrastructures, we are on a very, very fast ramp with our first data center CPU to a multi-billion dollar product line. This is going to be a very large product line for us. The capability of Grace Hopper is really quite spectacular. It has the ability to create computing nodes that simultaneously has very fast memory, as well as very large memory. In the areas of vector databases or semantic surge, what is called RAG, retrieval augmented generation. So that you could have a generative AI model be able to refer to proprietary data or a factual data before it generates a response, that data is quite large. And you can also have applications or generative models where the context length is very high. You basically store it in entire book into end-to-end system memory before you ask your questions. And so the context length can be quite large this way. The generative models has the ability to still be able to naturally interact with you on one hand. On the other hand, be able to refer to factual data, proprietary data or domain-specific data, you data and be contextually relevant and reduce hallucination. And so that particular use case for example is really quite fantastic for Grace Hopper. It also serves the customers that really care to have a different CPU than x86. Maybe it's a European supercomputing centers or European companies who would like to build up their own ARM ecosystem and like to build up a full stack or CSPs that have decided that they would like to pivot to ARM, because their own custom CPUs are based on ARM. There are variety of different reasons that drives the success of Grace Hopper, but we're off to a just an extraordinary start. This is a home run product.

Operator

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

Tim Arcuri

Hi. Thanks. I wanted to ask a little bit about the visibility that you have on revenue. I know there's a few moving parts. I guess, on one hand, the purchase commitments went up a lot again. But on the other hand, China bans would arguably pull in when you can fill the demand beyond China. So I know we're not even into 2024 yet and it doesn't sound like, Jensen, you think that next year would be a peak in your Data Center revenue, but I just wanted to sort of explicitly ask you that. Do you think that Data Center can grow even in 2025? Thanks.

Jensen Huang

Absolutely believe the Data Center can grow through 2025. And there are, of course, several reasons for that. We are expanding our supply quite significantly. We have already one of the broadest and largest and most capable supply chain in the world. Now, remember, people think that the GPU is a chip. But the HGX H100, the Hopper HGX has 35,000 parts, it weighs 70 pounds. Eight of the chips are Hopper. The other 35,000 are not. It is -- even its passive components are incredible. High voltage parts. High frequency parts. High current parts. It is a supercomputer, and therefore, the only way to test a supercomputer is with another supercomputer. Even the manufacturing of it is complicated, the testing of it is complicated, the shipping of it complicated and installation is complicated. And so, every aspect of our HGX supply chain is complicated. And the remarkable team that we have here has really scaled out the supply chain incredibly. Not to mention, all of our HGXs are connected with NVIDIA networking. And the networking, the transceivers, the mix, the cables, the switches, the amount of complexity there is just incredible. And so, I'm just -- first of all, I'm just super proud of the team for scaling up this incredible supply chain. We are absolutely world class. But meanwhile, we're adding new customers and new products. So we have new supply. We have new customers, as I was mentioning earlier. Different regions are standing up GPU specialist clouds, sovereign AI clouds coming out from all over the world, as people realize that they can't afford to export their country's knowledge, their country's culture for somebody else to then resell AI back to them, they have to -- they should, they have the skills and surely with us in combination, we can help them to do that build up their national AI. And so, the first thing that they have to do is, create their AI cloud, national AI cloud. You're also seeing us now growing into enterprise. The enterprise market has two paths. One path -- or if I could say three paths. The first path, of course, just off-the-shelf AI. And there are of course Chat GPT, a fabulous off-the-shelf AI, there'll be others. There's also a proprietary AI, because software companies like ServiceNow and SAP, there are many, many others that can't afford to have their company's intelligence be outsourced to somebody else. And they are about building tools and on top of their tools they should build custom and proprietary and domain-specific copilots and assistants that they can then rent to their customer base. This is -- they're sitting on a goldmine, almost every major tools company in the world is sitting on a goldmine, and they recognize that they have to go build their own custom AIs. We have a new service called an AI foundry, where we leverage NVS (ph) capabilities to be able to serve them in that. And then the next one is enterprises building their own custom AIs, their own custom chatbots, their own custom RAGs. And this capability is spreading all over the world. And the way that we're going to serve that marketplace is with the entire stacks of systems, which includes our compute, our networking and our switches, running our software stack called NVIDIA AI Enterprise, taking it through our market partners, HP, Dell, Lenovo, so on and so forth. And so we're just -- we're seeing the waves of generative AI starting from the start-ups and CSPs, moving to consumer Internet companies, moving to enterprise software platforms, moving to enterprise companies. And then ultimately, one of the areas that you guys have seen us spend a lot of energy on has to do with industrial generative AI. This is where NVIDIA AI and NVIDIA Omniverse comes together and that is a really, really exciting work. And so I think the -- we're at the beginning of a basically across-the-board industrial transition to generative AI to accelerated computing. This is going to affect every company, every industry, every country.

Operator

Your next question comes from the line of Toshiya Hari of Goldman Sachs. Your line is open.

Toshiya Hari

Hi. Thank you. I wanted to clarify something with Colette real quick, and then I had a question for Jensen as well. Colette, you mentioned that you'll be introducing regulation-compliant products over the next couple of months. Yet, the contribution to Q4 revenue should be relatively limited. Is that a timing issue and could it be a source of reacceleration in growth for Data Center in April and beyond or are the price points such that the contribution to revenue going forward should be relatively limited? And then the question for Jensen, the AI foundry service announcement from last week. I just wanted to ask about that, and hopefully, have you expand on it. How is the monetization model going to work? Is it primarily services and software revenue? How should we think about the long term opportunity set? And is this going to be exclusive to Microsoft or do you have plans to expand to other partners as well? Thank you.

Colette Kress

Thanks, Toshiya. On the question regarding potentially new products that we could provide to our China customers. It's a significant process to both design and develop these new products. As we discussed, we're going to make sure that we are in full discussions with the U.S. government of our intent to move products as well. Given our state about where we are in the quarter, we're already several weeks into the quarter. So it's just going to take some time for us to go through and discussing with our customers the needs and desires of these new products that we have. And moving forward, whether that's medium-term or long-term, it's just hard to say both the [Technical Difficulty] of what we can produce with the U.S. government and what the interest of our China customers in this. So we stay still focused on finding that right balance for our China customers, but it's hard to say at this time.

Jensen Huang

Toshiya, thanks for the question. There is a glaring opportunity in the world for AI foundry, and it makes so much sense. First, every company has its core intelligence. It makes up our company. Our data, our domain expertise, in the case of many companies, we create tools, and most of the software companies in the world are tool platforms, and those tools are used by people today. And in the future, it's going to be used by people augmented with a whole bunch of AIs that we hire. And these platforms just got to go across the world and you'll see and we've only announced a few; SAP, ServiceNow, Dropbox, Getty, many others are coming. And the reason for that is because they have their own proprietary AI. They want their own proprietary AI. They can't afford to outsource their intelligence and handout their data, and handout their flywheel for other companies to build the AI for them. And so, they come to us. We have several things that are really essential in a foundry. Just as TSMC as a foundry, you have to have AI technology. And as you know, we have just an incredible depth of AI capability -- AI technology capability. And then second, you have to have the best practice known practice, the skills of processing data through the invention of AI models to create AIs that are guardrails, fine-tuned, so on and so forth, that are safe, so on and so forth. And the third thing is you need factories. And that's what DGX Cloud is. Our AI models are called AI Foundations. Our process, if you will, our CAD system for creating AIs are called NeMo and they run on NVIDIA's factories we call DGX Cloud. Our monetization model is that with each one of our partners they rent a sandbox on DGX Cloud, where we work together, they bring their data, they bring their domain expertise, we bring our researchers and engineers, we help them build their custom AI. We help them make that custom AI incredible. Then that custom AI becomes theirs. And they deploy it on the runtime that is enterprise grade, enterprise optimized or outperformance optimized, runs across everything NVIDIA. We have a giant installed base in the cloud, on-prem, anywhere. And it's secure, securely patched, constantly patched and optimized and supported. And we call that NVIDIA AI Enterprise. NVIDIA AI Enterprise is $4,500 per GP per year, that's our business model. Our business model is basically a license. Our customers then with that basic license can build their monetization model on top of. In a lot of ways we're wholesale, they become retail. They could have a per -- they could have subscription license base, they could per instance or they could do per usage, there is a lot of different ways that they could take a -- create their own business model, but ours is basically like a software license, like an operating system. And so our business model is help you create your custom models, you run those custom models on NVIDIA AI Enterprise. And it's off to a great start. NVIDIA AI Enterprise is going to be a very large business for us.

Operator

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

Stacy Rasgon

Hi, guys. Thanks for taking my questions. Colette, I wanted to know if it weren't for the China restrictions would the Q4 guide has been higher or are you supply-constrained in just reshipping stuff that would have gone to China elsewhere? And I guess along those lines you give us a feeling for where your lead times are right now in data center and just the China redirection such as-is, is it lowering those lead times, because you've got parts that are sort of immediately available to ship?

Colette Kress

Yeah. Stacy, let me see if I can help you understand. Yes, there are still situations where we are working on both improving our supply each and every quarter. We've done a really solid job of ramping every quarter, which has defined our revenue. But with the absence of China for our outlook for Q4, sure, there could have been some things that we are not supply-constrained that we could have sold, but kind of we would no longer can. So could our guidance had been a little higher in our Q4? Yes. We are still working on improving our supply on plan, on continuing growing all throughout next year as well towards that.

Operator

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

Matt Ramsay

Thank you very much. Congrats, everybody, on the results. Jensen, I had a two-part question for you, and it comes off of sort of one premise. And the premise is, I still get a lot of questions from investors thinking about AI training as being NVIDIA's dominant domain and somehow inference, even large model inference takes more and more of the TAM that the market will become more competitive. You'll be less differentiated et cetera., et cetera. So I guess the two parts of the

question are: number one, maybe you could spend a little bit of time talking about the evolution of the inference workload as we move to LLMs and how your company is positioned for that rather than smaller model inference. And second, up until a month or two ago, I never really got any questions at all about the data processing piece of the AI workloads. So the pieces of manipulating the data before training, between training and inference, after inference and I think that's a large part of the workload now. Maybe you could talk about how CUDA is enabling acceleration of those pieces of the workload. Thanks.

Jensen Huang

Sure. Inference is complicated. It's actually incredibly complicated. If you -- we this quarter announced one of the most exciting new engines, optimizing compilers called TensorRT-LLM. The reception has been incredible. You got to GitHub, it's been downloaded a ton, a whole lot of stars, integrated into stacks and frameworks all over the world, almost instantaneously. And there are several reasons for that, obviously. We could create TensorRT-LLM, because CUDA is programmable. If CUDA and our GPUs were not so programmable, it would really be hard for us to improve software stacks at the pace that we do. TensorRT-LLM, on the same GPU, without anybody touching anything, improves the performance by a factor of two. And then on top of that, of course, the pace of our innovation is so high. H200 increases it by another factor of two. And so, our inference performance, another way of saying inference cost, just reduced by a factor of four within about a year's time. And so, that's really, really hard to keep up with. The reason why everybody likes our inference engine is because our installed base. We've been dedicated to our installed base for 20 years, 20-plus years. We have an installed base that is not only largest in every single cloud, it's in every available from every enterprise system maker, it's used by companies of just about every industry. And every -- anytime you see a NVIDIA GPU, it runs our stack. It's architecturally compatible, something we've been dedicated to for a very long time. We're very disciplined about it. We make it our, if you will, architecture compatibility is job one. And that has conveyed to the world, the certainty of our platform stability. NVIDIA's platform stability certainty is the reason why everybody builds on us first and the reason why everybody optimizes on us first. All the engineering and all the work that you do, all the invention of technologies that you build on top of NVIDIA accrues to the -- and benefits everybody that uses our GPUs. And we have such a large installed base, large -- millions and millions of GPUs in cloud, 100 million GPUs from people's PCs just about every workstation in the world, and they all architecturally compatible. And so, if you are an inference platform and you're deploying an inference application, you are basically an application provider. And as a software application provider, you're looking for large installed base. Data processing, before you could train a model, you have to curate the data, you have to dedupe the data, maybe you have to augment the data with synthetic data. So, process the data, clean the data, align the data, normalize the data, all of that data is measured not in bytes or megabytes, it's measured in terabytes and petabytes. And the amount of data processing that you do before data engineering, before that you do training is quite significant. It could represent 30%, 40%, 50% of the amount of work that you ultimately do. And what you -- and ultimately creating a data driven machine learning service. And so data processing is just a massive part. We accelerate Spark, we accelerate Python. One of the coolest things that we just did is called cuDF Pandas. Without one line of code, Pandas, which is the single most successful data science framework in the world. Pandas now is accelerated by NVIDIA CUDA. And just out-of-the box, without the line of code and so the acceleration is really quite terrific and people are just incredibly excited about it. And Pandas was designed for one purpose and one purpose only, really data processing, it's for data science. And so NVIDIA CUDA gives you all of that.

Operator

Your final question comes from the line of Harlan Sur of J.P. Morgan. Your line is open.

Harlan Sur

Good afternoon. Thanks for taking my question. If you look at the history of the tech industry like those companies that have been successful have always been focused on ecosystem; silicon, hardware, software, strong partnerships and just as importantly, right, an aggressive cadence of new products, more segmentation over time. The team recently announced a more aggressive new product cadence in data center from two years to now every year with higher levels of segmentation, training, optimization in printing CPU, GPU, DPU networking. How do we think about your R&D OpEx growth outlook to support a more aggressive and expanding forward roadmap, but more importantly, what is the team doing to manage and drive execution through all of this complexity?

Jensen Huang

Gosh. Boy, that's just really excellent. You just wrote NVIDIA's business plan, and so you described our strategy. First of all, there is a fundamental reason why we accelerate our execution. And the reason for that is because it fundamentally drives down cost. When the combination of TensorRT-LLM and H200 reduce the cost for our customers for large model inference by a factor of four, and so that includes, of course, our speeds and feeds, but mostly it's because of our software, mostly the software benefits because of the architecture. And so we want to accelerate our roadmap for that reason. The second reason is to expand the reach of generative AI, the world's number of data center configurations -- this is kind of the amazing thing. NVIDIA is in every cloud, but not one cloud is the same. NVIDIA is working with every single cloud service provider and not one of the networking control plane, security posture is the same. Everybody's platform is different and yet we're integrated into all of their stacks, all of their data centers and we work incredibly well with all of them. And not to mention, we then take the whole thing and we create AI factories that are standalone. We take our platform, we can put them into supercomputers, we can put them into enterprise. Bringing AI to enterprise is something generative AI Enterprise something nobody's ever done before. And we're right now in the process of going to market with all of that. And so the complexity includes, of course, all the technologies and segments and the pace. It includes the fact that we are architecturally compatible across every single one of those. It includes all of the domain specific libraries that we create. The reason why every computer company, without thinking, can integrate NVIDIA into their roadmap and take it to market. And the reason for that is, because there is market demand for it. There is market demand in healthcare, there is market demand in manufacturing, there is market demand, of course, in AI, including financial services, in supercomputing and quantum computing. The list of markets and segments that we have domain specific libraries is incredibly broaden. And then finally, we have an end-to-end solution for data centers; InfiniBand networking, Ethernet networking, x86, ARM, just about every permutation combination of solutions -- technology solutions and software stacks provided. And that translates to having the largest number of ecosystem software developers; the largest ecosystem of system makers; the largest and broadest distribution partnership network; and ultimately, the greatest reach. And that takes -- surely that takes a lot of energy. But the thing that really holds it together, and this is a great decision that we made decades ago, which is everything is architecturally compatible. When we develop a domain specific language that runs on one GPU, it runs on every GPU. When we optimize TensorRT for the cloud, we optimized it for enterprise. When we do something that brings in a new feature, a new library, a new feature or a new developer, they instantly get the benefit of all of our reach. And so that discipline, that architecture compatible discipline that has lasted more than a couple of decades now, is one of the reasons why NVIDIA is still really, really efficient. I mean, we're 28,000 people large and serving just about every single company, every single industry, every single market around the world.

Operator

Thank you. I will now turn the call back over to Jensen Huang for closing remarks.

Jensen Huang

Our strong growth reflects the broad industry platform transition from general purpose to accelerated computing and generative AI. Large language models start-ups consumer Internet companies and global cloud service providers are the first movers. The next waves are starting to build. Nations and regional CSPs are building AI clouds to serve local demand. Enterprise software companies like Adobe and Dropbox, SAP and ServiceNow are adding AI copilots and assistants to their platforms. Enterprises in the world's largest industries are creating custom AIs to automate and boost productivity. The generative AI era is in full steam and has created the need for a new type of data center, an AI factory; optimized for refining data and training, and inference, and generating AI. AI factory workloads are different and incremental to legacy data center workloads supporting IT tasks. AI factories run copilots and AI assistants, which are significant software TAM expansion and are driving significant new investment. Expanding the $1 trillion traditional data center infrastructure installed base, empowering the AI Industrial Revolution. NVIDIA H100 HGX with InfiniBand and the NVIDIA AI software stack define an AI factory today. As we expand our supply chain to meet the world's demand, we are also building new growth drivers for the next wave of AI. We highlighted three elements to our new growth strategy that are hitting their stride: CPU, networking, and software and services. Grace is NVIDIA's first data center CPU. Grace and Grace Hopper are in full production and ramping into a new multi-billion dollar product line next year. Irrespective of the CPU choice, we can help customers build an AI factory. NVIDIA networking now exceeds $10 billion annualized revenue run rate. InfiniBand grew five-fold year-over-year, and is positioned for excellent growth ahead as the networking of AI factories. Enterprises are also racing to adopt AI and Ethernet is the standard networking. This week we announced an Ethernet for AI platform for enterprises. NVIDIA Spectrum-X is an end-to-end solution of Bluefield SuperNIC, Spectrum-4 Ethernet switch and software that boosts Ethernet performance by up to 1.6x for AI workloads. Dell, HPE and Lenovo have joined us to bring a full generative AI solution of NVIDIA AI computing, networking and software to the world's enterprises. NVIDIA software and services is on track to exit the year at an annualized run rate of $1 billion. Enterprise software platforms like ServiceNow and SAP need to build and operate proprietary AI. Enterprises need to build and deploy custom AI copilots. We have the AI technology, expertise and scale to help customers build custom models with their proprietary data on NVIDIA DGX Cloud and deploy the AI applications on enterprise grade NVIDIA AI Enterprise. NVIDIA is essentially an AI foundry. NVIDIA's GPUs, CPUs, networking, AI foundry services and NVIDIA AI Enterprise software are all growth engines in full throttle. Thanks for joining us today. We look forward to updating you on our progress next quarter.

Operator

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