NVIDIA Corp. (NASDAQ:NVDA) Q2 2024 Earnings Conference Call August 23, 2023 5:00 PM ET

Company Participants

Simona Jankowski - VP, IR

Colette Kress - EVP & CFO

Jensen Huang - Co-Founder, CEO & President

Conference Call Participants

Matt Ramsay - Cowen

Vivek Arya - Bank of America

Stacy Rasgon - Bernstein Research

Mark Lipacis - Jefferies

Atif Malik - Citi

Joseph Moore - Morgan Stanley

Toshiya Hari - Goldman Sachs

Timothy Arcuri - UBS

Benjamin Reitzes - Melius

Operator

Good afternoon. My name is David, and I'll be your conference operator today. At this time, I'd like to welcome everyone to NVIDIA's Second Quarter Earnings Call. Today's conference is being recorded. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer session. [Operator Instructions]

Thank you. Simona Jankowski, you may begin your conference.

Simona Jankowski

Thank you. Good afternoon, everyone and welcome to NVIDIA's conference call for the second 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 third quarter of 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 our statements are made as of today, August 23, 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.

And with that, let me turn the call over to Colette.

Colette Kress

Thanks, Simona. We had an exceptional quarter. Record Q2 revenue of $13.51 billion was up 88% sequentially and up 101% year-on-year, and above our outlook of $11 billion.

Let me first start with Data Center. Record revenue of $10.32 billion was up 141% sequentially and up 171% year-on-year. Data Center compute revenue nearly tripled year-on-year, driven primarily by accelerating demand from cloud service providers and large consumer Internet companies for HGX platform, the engine of generative AI and large language models.

Major companies, including AWS, Google Cloud, Meta, Microsoft Azure and Oracle Cloud as well as a growing number of GPU cloud providers are deploying, in volume, HGX systems based on our Hopper and Ampere architecture Tensor Core GPUs. Networking revenue almost doubled year-on-year, driven by our end-to-end InfiniBand networking platform, the gold standard for AI.

There is tremendous demand for NVIDIA accelerated computing and AI platforms. Our supply partners have been exceptional in ramping capacity to support our needs. Our data center supply chain, including HGX with 35,000 parts and highly complex networking has been built up over the past decade. We have also developed and qualified additional capacity and suppliers for key steps in the manufacturing process such as [indiscernible] packaging.

We expect supply to increase each quarter through next year. By geography, data center growth was strongest in the U.S. as customers direct their capital investments to AI and accelerated computing. China demand was within the historical range of 20% to 25% of our Data Center revenue, including compute and networking solutions.

At this time, let me take a moment to address recent reports on the potential for increased regulations on our exports to China. We believe the current regulation is achieving the intended results. Given the strength of demand for our products worldwide, we do not anticipate that additional export restrictions on our Data Center GPUs, if adopted, would have an immediate material impact to our financial results.

However, over the long term, restrictions prohibiting the sale of our Data Center GPUs to China, if implemented, will result in a permanent loss and opportunity for the U.S. industry to compete and lead in one of the world's largest markets.

Our cloud service providers drove exceptional strong demand for HGX systems in the quarter, as they undertake a generational transition to upgrade their data center infrastructure for the new era of accelerated computing and AI. The NVIDIA HGX platform is culminating of nearly two decades of full stack innovation across silicon, systems, interconnects, networking, software and algorithms.

Instances powered by the NVIDIA H100 Tensor Core GPUs are now generally available at AWS, Microsoft Azure and several GPU cloud providers, with others on the way shortly. Consumer Internet companies also drove the very strong demand. Their investments in data center infrastructure purpose-built for AI are already generating significant returns. For example, Meta, recently highlighted that since launching Reels, AI recommendations have driven a more than 24% increase in time spent on Instagram.

Enterprises are also racing to deploy generative AI, driving strong consumption of NVIDIA powered instances in the cloud as well as demand for on-premise infrastructure. Whether we serve customers in the cloud or on-prem through partners or direct, their applications can run seamlessly on NVIDIA AI enterprise software with access to our acceleration libraries, pre-trained models and APIs.

We announced a partnership with Snowflake to provide enterprises with accelerated path to create customized generative AI applications using their own proprietary data, all securely within the Snowflake Data Cloud. With the NVIDIA NeMo platform for developing large language models, enterprises will be able to make custom LLMs for advanced AI services, including chatbot, search and summarization, right from the Snowflake Data Cloud.

Virtually, every industry can benefit from generative AI. For example, AI Copilot such as those just announced by Microsoft can boost the productivity of over 1 billion office workers and tens of millions of software engineers. Billions of professionals in legal services, sales, customer support and education will be available to leverage AI systems trained in their field. AI Copilot and assistants are set to create new multi-hundred billion dollar market opportunities for our customers.

We are seeing some of the earliest applications of generative AI in marketing, media and entertainment. WPP, the world's largest marketing and communication services organization, is developing a content engine using NVIDIA Omniverse to enable artists and designers to integrate generative AI into 3D content creation. WPP designers can create images from text prompts while responsibly trained generative AI tools and content from NVIDIA partners such as Adobe and Getty Images using NVIDIA Picasso, a foundry for custom generative AI models for visual design.

Visual content provider Shutterstock is also using NVIDIA Picasso to build tools and services that enables users to create 3D scene background with the help of generative AI. We've partnered with ServiceNow and Accenture to launch the AI Lighthouse program, fast tracking the development of enterprise AI capabilities. AI Lighthouse unites the ServiceNow enterprise automation platform and engine with NVIDIA accelerated computing and with Accenture consulting and deployment services.

We are collaborating also with Hugging Face to simplify the creation of new and custom AI models for enterprises. Hugging Face will offer a new service for enterprises to train and tune advanced AI models powered by NVIDIA HGX cloud. And just yesterday, VMware and NVIDIA announced a major new enterprise offering called VMware Private AI Foundation with NVIDIA, a fully integrated platform featuring AI software and accelerated computing from NVIDIA with multi-cloud software for enterprises running VMware.

VMware's hundreds of thousands of enterprise customers will have access to the infrastructure, AI and cloud management software needed to customize models and run generative AI applications such as intelligent chatbot, assistants, search and

summarization. We also announced new NVIDIA AI enterprise-ready servers featuring the new NVIDIA L40S GPU built for the industry standard data center server ecosystem and BlueField-3 DPU data center infrastructure processor.

L40S is not limited by [indiscernible] supply and is shipping to the world's leading server system makers (ph). L40S is a universal data center processor designed for high volume data center standing out to accelerate the most compute-intensive applications, including AI training and inventing through the designing, visualization, video processing and NVIDIA Omniverse industrial digitalization.

NVIDIA AI enterprise ready servers are fully optimized for VMware, Cloud Foundation and Private AI Foundation. Nearly 100 configurations of NVIDIA AI enterprise ready servers will soon be available from the world's leading enterprise IT computing companies, including Dell, HP and Lenovo. The GH200 Grace Hopper Superchip which combines our ARM-based Grace CPU with Hopper GPU entered full production and will be available this quarter in OEM servers. It is also shipping to multiple supercomputing customers, including Atmos (ph), National Labs and the Swiss National Computing Center.

And NVIDIA and SoftBank are collaborating on a platform based on GH200 for generative AI and 5G/6G applications. The second generation version of our Grace Hopper Superchip with the latest HBM3e memory will be available in Q2 of calendar 2024. We announced the DGX GH200, a new class of large memory AI supercomputer for giant AI language model, recommendator systems and data analytics. This is the first use of the new NVIDIA [indiscernible] switch system, enabling all of its 256 Grace Hopper Superchips to work together as one, a huge jump compared to our prior generation connecting just eight GPUs over [indiscernible]. DGX GH200 systems are expected to be available by the end of the year, Google Cloud, Meta and Microsoft among the first to gain access.

Strong networking growth was driven primarily by InfiniBand infrastructure to connect HGX GPU systems. Thanks to its end-to-end optimization and in-network computing capabilities, InfiniBand delivers more than double the performance of traditional Ethernet for AI. For billions of dollar AI infrastructures, the value from the increased throughput of InfiniBand is worth hundreds of [indiscernible] and pays for the network. In addition, only InfiniBand can scale to hundreds of thousands of GPUs. It is the network of choice for leading AI practitioners.

For Ethernet-based cloud data centers that seek to optimize their AI performance, we announced NVIDIA Spectrum-X, an accelerated networking platform designed to

optimize Ethernet for AI workloads. Spectrum-X couples the Spectrum or Ethernet switch with the BlueField-3 DPU, achieving 1.5x better overall AI performance and power efficiency versus traditional Ethernet. BlueField-3 DPU is a major success. It is in qualification with major OEMs and ramping across multiple CSPs and consumer Internet companies.

Now moving to gaming. Gaming revenue of $2.49 billion was up 11% sequentially and 22% year-on-year. Growth was fueled by GeForce RTX 40 Series GPUs for laptops and desktop. End customer demand was solid and consistent with seasonality. We believe global end demand has returned to growth after last year's slowdown. We have a large upgrade opportunity ahead of us. Just 47% of our installed base have upgraded to RTX and about 20% of the GPU with an RTX 3060 or higher performance.

Laptop GPUs posted strong growth in the key back-to-school season, led by RTX 4060 GPUs. NVIDIA's GPU-powered laptops have gained in popularity, and their shipments are now outpacing desktop GPUs from several regions around the world. This is likely to shift the reality of our overall gaming revenue a bit, with Q2 and Q3 as the stronger quarters of the year, reflecting the back-to-school and holiday build schedules for laptops.

In desktop, we launched the GeForce RTX 4060 and the GeForce RTX 4060 TI GPUs, bringing the Ada Lovelace architecture down to price points as low as $299. The ecosystem of RTX and DLSS games continue to expand. 35 new games added to DLSS support, including blockbusters such as Diablo IV and Baldur’s Gate 3.

There's now over 330 RTX accelerated games and apps. We are bringing generative AI to gaming. At COMPUTEX, we announced NVIDIA Avatar Cloud Engine or ACE for games, a custom AI model foundry service. Developers can use this service to bring intelligence to non-player characters. And it harnesses a number of NVIDIA Omniverse and AI technologies, including NeMo, Riva and Audio2Face.

Now moving to Professional Visualization. Revenue of $375 million was up 28% sequentially and down 24% year-on-year. The Ada architecture ramp drove strong growth in Q2, rolling out initially in laptop workstations with a refresh of desktop workstations coming in Q3. These will include powerful new RTX systems with up to 4 NVIDIA RTX 6000 GPUs, providing more than 5,800 teraflops of AI performance and 192 gigabytes of GPU memory. They can be configured with NVIDIA AI enterprise or NVIDIA Omniverse inside.

We also announced three new desktop workstation GPUs based on the Ada generation. The NVIDIA RTX 5000, 4500 and 4000, offering up to 2x the RT core throughput and up to 2x faster AI training performance compared to the previous generation. In addition to traditional workloads such as 3D design and content creation, new workloads in generative AI, large language model development and data science are expanding the opportunity in pro visualization for our RTX technology.

One of the key themes in Jensen's keynote [indiscernible] earlier this month was the conversion of graphics and AI. This is where NVIDIA Omniverse is positioned. Omniverse is OpenUSD's native platform. OpenUSD is a universal interchange that is quickly becoming the standard for the 3D world, much like HTML is the universal language for the 2D [indiscernible]. Together, Adobe, Apple, Autodesk, Pixar and NVIDIA form the Alliance for OpenUSD. Our mission is to accelerate OpenUSD's development and adoption. We announced new and upcoming Omniverse cloud APIs, including RunUSD and ChatUSD to bring generative AI to OpenUSD workload.

Moving to automotive. Revenue was $253 million, down 15% sequentially and up 15% year-on-year. Solid year-on-year growth was driven by the ramp of self-driving platforms based on [indiscernible] or associated with a number of new energy vehicle makers. The sequential decline reflects lower overall automotive demand, particularly in China. We announced a partnership with MediaTek to bring drivers and passengers new experiences inside the car. MediaTek will develop automotive SoCs and integrate a new product line of NVIDIA's GPU chiplet. The partnership covers a wide range of vehicle segments from luxury to entry level.

Moving to the rest of the P&L. GAAP gross margins expanded to 70.1% and non-GAAP gross margin to 71.2%, driven by higher data center sales. Our Data Center products include a significant amount of software and complexity, which is also helping drive our gross margin. Sequential GAAP operating expenses were up 6% and non-GAAP operating expenses were up 5%, primarily reflecting increased compensation and benefits. We returned approximately $3.4 billion to shareholders in the form of share repurchases and cash dividends. Our Board of Directors has just approved an additional $25 billion in stock repurchases to add to our remaining $4 billion of authorization as of the end of Q2.

Let me turn to the outlook for the third quarter of fiscal 2024. Demand for our Data Center platform where AI is tremendous and broad-based across industries on customers. Our demand visibility extends into next year. Our supply over the next several quarters will continue to ramp as we lower cycle times and work with our supply

partners to add capacity. Additionally, the new L40S GPU will help address the growing demand for many types of workloads from cloud to enterprise.

For Q3, total revenue is expected to be $16 billion, plus or minus 2%. We expect sequential growth to be driven largely by Data Center with gaming and ProViz also contributing. GAAP and non-GAAP gross margins are expected to be 71.5% and 72.5%, respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $2.95 billion and $2 billion, respectively.

GAAP and non-GAAP other income and expenses are expected to be an income of approximately $100 million, excluding gains and losses from non-affiliated investments. GAAP and non-GAAP tax rates are expected to be 14.5%, 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 Jefferies Tech Summit on August 30 in Chicago, the Goldman Sachs Conference on September 5 in San Francisco, the Evercore Semiconductor Conference on September 6 as well as the Citi Tech Conference on September 7, both in New York. And the BofA Virtual AI conference on September 11. Our earnings call to discuss the results of our third quarter of fiscal 2024 is scheduled for Tuesday, November 21.

Operator, we will now open the call for questions. Could you please poll for questions for us? Thank you.

Question-and-Answer Session

Operator

Thank you. [Operator Instructions] We'll take our first question from Matt Ramsay with TD Cowen. Your line is now open.

Matt Ramsay

Yes. Thank you very much. Good afternoon. Obviously, remarkable results. Jensen, I wanted to ask a question of you regarding the really quickly emerging application of large model inference. So I think it's pretty well understood by the majority of investors that you guys have very much a lockdown share of the training market. A lot of the smaller market -- smaller model inference workloads have been done on ASICs or CPUs in the past.

And with many of these GPT and other really large models, there's this new workload that's accelerating super-duper quickly on large model inference. And I think your Grace Hopper Superchip products and others are pretty well aligned for that. But could you maybe talk to us about how you're seeing the inference market segment between small model inference and large model inference and how your product portfolio is positioned for that? Thanks.

Jensen Huang

Yeah. Thanks a lot. So let's take a quick step back. These large language models are fairly -- are pretty phenomenal. It does several things, of course. It has the ability to understand unstructured language. But at its core, what it has learned is the structure of human language. And it has encoded or within it -- compressed within it a large amount of human knowledge that it has learned by the corpuses that it studied. What happens is, you create these large language models and you create as large as you can, and then you derive from it smaller versions of the model, essentially teacher-student models. It's a process called distillation.

And so when you see these smaller models, it's very likely the case that they were derived from or distilled from or learned from larger models, just as you have professors and teachers and students and so on and so forth. And you're going to see this going forward. And so you start from a very large model and it has a large amount of generality and generalization and what's called zero-shot capability. And so for a lot of applications and questions or skills that you haven't trained it specifically on, these large language models miraculously has the capability to perform them. That's what makes it so magical.

On the other hand, you would like to have these capabilities in all kinds of computing devices, and so what you do is you distill them down. These smaller models might have excellent capabilities on a particular skill, but they don't generalize as well. They don't have what is called as good zero-shot capabilities. And so they all have their own unique capabilities, but you start from very large models.

Operator

Okay. Next, we'll go to Vivek Arya with BofA Securities. Your line is now open.

Vivek Arya

Thank you. Just had a quick clarification and a question. Colette, if you could please clarify how much incremental supply do you expect to come online in the next year? You

think it's up 20%, 30%, 40%, 50%? So just any sense of how much supply because you said it's growing every quarter.

And then Jensen, the question for you is, when we look at the overall hyperscaler spending, that buy is not really growing that much. So what is giving you the confidence that they can continue to carve out more of that pie for generative AI? Just give us your sense of how sustainable is this demand as we look over the next one to two years? So if I take your implied Q3 outlook of Data Center, $12 billion, $13 billion, what does that say about how many servers are already AI accelerated? Where is that going? So just give some confidence that the growth that you are seeing is sustainable into the next one to two years.

Colette Kress

So thanks for that question regarding our supply. Yes, we do expect to continue increasing ramping our supply over the next quarters as well as into next fiscal year. In terms of percent, it's not something that we have here. It is a work across so many different suppliers, so many different parts of building an HGX and many of our other new products that are coming to market. But we are very pleased with both the support that we have with our suppliers and the long time that we have spent with them improving their supply.

Jensen Huang

The world has something along the lines of about $1 trillion worth of data centers installed, in the cloud, in enterprise and otherwise. And that $1 trillion of data centers is in the process of transitioning into accelerated computing and generative AI. We're seeing two simultaneous platform shifts at the same time. One is accelerated computing. And the reason for that is because it's the most cost-effective, most energy effective and the most performant way of doing computing now.

So what you're seeing, and then all of a sudden, enabled by generative AI, enabled by accelerated compute and generative AI came along. And this incredible application now gives everyone two reasons to transition to do a platform shift from general purpose computing, the classical way of doing computing, to this new way of doing computing, accelerated computing. It's about $1 trillion worth of data centers, call it, $0.25 trillion of capital spend each year.

You're seeing the data centers around the world are taking that capital spend and focusing it on the two most important trends of computing today, accelerated computing and generative AI. And so I think this is not a near-term thing. This is a

long-term industry transition and we're seeing these two platform shifts happening at the same time.

Operator

Next, we go to Stacy Rasgon with Bernstein Research. Your line is open.

Stacy Rasgon

Hi, guys. Thanks for taking my question. I was wondering, Colette, if you could tell me like how much of Data Center in the quarter, maybe even the guide is like systems versus GPU, like DGX versus just the H100? What I'm really trying to get at is, how much is like pricing or content or however you want to define that [indiscernible] versus units actually driving the growth going forward. Can you give us any color around that?

Colette Kress

Sure, Stacy. Let me help. Within the quarter, our HGX systems were a very significant part of our Data Center as well as our Data Center growth that we had seen. Those systems include our HGX of our Hopper architecture, but also our Ampere architecture. Yes, we are still selling both of these architectures in the market. Now when you think about that, what does that mean from both the systems as a unit, of course, is growing quite substantially, and that is driving in terms of the revenue increases. So both of these things are the drivers of the revenue inside Data Center.

Our DGXs are always a portion of additional systems that we will sell. Those are great opportunities for enterprise customers and many other different types of customers that we're seeing even in our consumer Internet companies. The importance there is also coming together with software that we sell with our DGXs, but that's a portion of our sales that we're doing. The rest of the GPUs, we have new GPUs coming to market that we talk about the L40S, and they will add continued growth going forward. But again, the largest driver of our revenue within this last quarter was definitely the HGX system.

Jensen Huang

And Stacy, if I could just add something. You say it’s H100 and I know you know what your mental image in your mind. But the H100 is 35,000 parts, 70 pounds, nearly 1 trillion transistors in combination. Takes a robot to build – well, many robots to build because it’s 70 pounds to lift. And it takes a supercomputer to test a supercomputer. And so these things are technology marvels, and the manufacturing of them is really

intensive. And so I think we call it H100 as if it's a chip that comes off of a fab, but H100s go out really as HGXs sent to the world's hyperscalers and they're really, really quite large system components, if you will.

Operator

Next, we go to Mark Lipacis with Jefferies. Your line is now open.

Mark Lipacis

Hi. Thanks for taking my question and congrats on the success. Jensen, it seems like a key part of the success -- your success in the market is delivering the software ecosystem along with the chip and the hardware platform. And I had a two-part question on this. I was wondering if you could just help us understand the evolution of your software ecosystem, the critical elements. And is there a way to quantify your lead on this dimension like how many person years you've invested in building it? And then part two, I was wondering if you would care to share with us your view on the -- what percentage of the value of the NVIDIA platform is hardware differentiation versus software differentiation? Thank you.

A – Jensen Huang

Yeah, Mark, I really appreciate the question. Let me see if I could use some metrics, so we have a run time called AI Enterprise. This is one part of our software stack. And this is, if you will, the run time that just about every company uses for the end-to-end of machine learning from data processing, the training of any model that you like to do on any framework you'd like to do, the inference and the deployment, the scaling it out into a data center. It could be a scale-out for a hyperscale data center. It could be a scale-out for enterprise data center, for example, on VMware.

You can do this on any of our GPUs. We have hundreds of millions of GPUs in the field and millions of GPUs in the cloud and just about every single cloud. And it runs in a single GPU configuration as well as multi-GPU per compute or multi-node. It also has multiple sessions or multiple computing instances per GPU. So from multiple instances per GPU to multiple GPUs, multiple nodes to entire data center scale. So this run time called NVIDIA AI enterprise has something like 4,500 software packages, software libraries and has something like 10,000 dependencies among each other.

And that run time is, as I mentioned, continuously updated and optimized for our installed base for our stack. And that's just one example of what it would take to get accelerated computing to work. The number of code combinations and type of

application combinations is really quite insane. And it's taken us two decades to get here. But what I would characterize as probably our -- the elements of our company, if you will, are several. I would say number 1 is architecture.

The flexibility, the versatility and the performance of our architecture makes it possible for us to do all the things that I just said, from data processing to training to inference, for preprocessing of the data before you do the inference to the post processing of the data, tokenizing of languages so that you could then train with it. The amount of -- the workflow is much more intense than just training or inference. But anyways, that's where we'll focus and it's fine. But when people actually use these computing systems, it's quite -- requires a lot of applications. And so the combination of our architecture makes it possible for us to deliver the lowest cost ownership. And the reason for that is because we accelerate so many different things.

The second characteristic of our company is the installed base. You have to ask yourself, why is it that all the software developers come to our platform? And the reason for that is because software developers seek a large installed base so that they can reach the largest number of end users, so that they could build a business or get a return on the investments that they make.

And then the third characteristic is reach. We're in the cloud today, both for public cloud, public-facing cloud because we have so many customers that use -- so many developers and customers that use our platform. CSPs are delighted to put it up in the cloud. They use it for internal consumption to develop and train and to operate recommender systems or search or data processing engines and whatnot all the way to training and inference. And so we're in the cloud, we're in enterprise.

Yesterday, we had a very big announcement. It's really worthwhile to take a look at that. VMware is the operating system of the world's enterprise. And we've been working together for several years now, and we're going to bring together -- together, we're going to bring generative AI to the world's enterprises all the way out to the edge. And so reach is another reason. And because of reach, all of the world's system makers are anxious to put NVIDIA's platform in their systems. And so we have a very broad distribution from all of the world's OEMs and ODMs and so on and so forth because of our reach.

And then lastly, because of our scale and velocity, we were able to sustain this really complex stack of software and hardware, networking and compute and across all of these different usage models and different computing environments. And we're able to do all this while accelerating the velocity of our engineering. It seems like we're introducing a new architecture every two years. Now we're introducing a new

architecture, a new product just about every six months. And so these properties make it possible for the ecosystem to build their company and their business on top of us. And so those in combination makes us special.

Operator

Next, we'll go to Atif Malik with Citi. Your line is open.

Atif Malik

Hi. Thank you for taking my question. Great job on results and outlook. Colette, I have a question on the core L40S that you guys talked about. Any idea how much of the supply tightness can L40S help with? And if you can talk about the incremental profitability or gross margin contribution from this product? Thank you.

Jensen Huang

Yeah, Atif. Let me take that for you. The L40S is really designed for a different type of application. H100 is designed for large-scale language models and processing just very large models and a great deal of data. And so that's not L40S' focus. L40S' focus is to be able to fine-tune models, fine-tune pretrained models, and it'll do that incredibly well. It has a transform engine. It's got a lot of performance. You can get multiple GPUs in a server. It's designed for hyperscale scale-out, meaning it's easy to install L40S servers into the world's hyperscale data centers. It comes in a standard rack, standard server, and everything about it is standard and so it's easy to install.

L40S also is with the software stack around it and along with BlueField-3 and all the work that we did with VMware and the work that we did with Snowflakes and ServiceNow and so many other enterprise partners. L40S is designed for the world's enterprise IT systems. And that's the reason why HPE, Dell, and Lenovo and some 20 other system makers building about 100 different configurations of enterprise servers are going to work with us to take generative AI to the world's enterprise. And so L40S is really designed for a different type of scale-out, if you will. It's, of course, large language models. It's, of course, generative AI, but it's a different use case. And so the L40S is going to -- is off to a great start and the world's enterprise and hyperscalers are really clamoring to get L40S deployed.

Operator

Next, we'll go to Joe Moore with Morgan Stanley. Your line is open.

Joseph Moore

Great. Thank you. I guess the thing about these numbers that's so remarkable to me is the amount of demand that remains unfulfilled, talking to some of your customers. As good as these numbers are, you sort of more than tripled your revenue in a couple of quarters. There's a demand, in some cases, for multiples of what people are getting. So can you talk about that? How much unfulfilled demand do you think there is? And you talked about visibility extending into next year. Do you have line of sight into when you get to see supply-demand equilibrium here?

Jensen Huang

Yeah. We have excellent visibility through the year and into next year. And we're already planning the next-generation infrastructure with the leading CSPs and data center builders. The demand – easiest way to think about the demand, the world is transitioning from general-purpose computing to accelerated computing. That's the easiest way to think about the demand. The best way for companies to increase their throughput, improve their energy efficiency, improve their cost efficiency is to divert their capital budget to accelerated computing and generative AI. Because by doing that, you're going to offload so much workload off of the CPUs, but the available CPUs is -- in your data center will get boosted.

And so what you're seeing companies do now is recognizing this -- the tipping point here, recognizing the beginning of this transition and diverting their capital investment to accelerated computing and generative AI. And so that's probably the easiest way to think about the opportunity ahead of us. This isn't a singular application that is driving the demand, but this is a new computing platform, if you will, a new computing transition that's happening. And data centers all over the world are responding to this and shifting in a broad-based way.

Operator

Next, we go to Toshiya Hari with Goldman Sachs. Your line is now open.

Toshiya Hari

Hi. Thank you for taking the question. I had one quick clarification question for Colette and then another one for Jensen. Colette, I think last quarter, you had said CSPs were about 40% of your Data Center revenue, consumer Internet at 30%, enterprise 30%. Based on your remarks, it sounded like CSPs and consumer Internet may have been a larger percentage of your business. If you can kind of clarify that or confirm that, that would be super helpful.

And then Jensen, a question for you. Given your position as the key enabler of AI, the breadth of engagements and the visibility you have into customer projects, I'm curious how confident you are that there will be enough applications or use cases for your customers to generate a reasonable return on their investments. I guess I ask the question because there is a concern out there that there could be a bit of a pause in your demand profile in the out years. Curious if there's enough breadth and depth there to support a sustained increase in your Data Center business going forward. Thank you.

Colette Kress

Okay. So thank you, Toshiya, on the question regarding our types of customers that we have in our Data Center business. And we look at it in terms of combining our compute as well as our networking together. Our CSPs, our large CSPs are contributing a little bit more than 50% of our revenue within Q2. And the next largest category will be our consumer Internet companies. And then the last piece of that will be our enterprise and high performance computing.

Jensen Huang

Toshi, I'm reluctant to guess about the future and so I'll answer the question from the first principle of computer science perspective. It is recognized for some time now that general purpose computing is just not and brute forcing general purpose computing. Using general purpose computing at scale is no longer the best way to go forward. It's too energy costly, it's too expensive, and the performance of the applications are too slow.

And finally, the world has a new way of doing it. It's called accelerated computing and what kicked it into turbocharge is generative AI. But accelerated computing could be used for all kinds of different applications that's already in the data center. And by using it, you offload the CPUs. You save a ton of money in order of magnitude, in cost and order of magnitude and energy and the throughput is higher and that's what the industry is really responding to.

Going forward, the best way to invest in the data center is to divert the capital investment from general purpose computing and focus it on generative AI and accelerated computing. Generative AI provides a new way of generating productivity, a new way of generating new services to offer to your customers, and accelerated computing helps you save money and save power. And the number of applications is, well, tons. Lots of developers, lots of applications, lots of libraries. It's ready to be deployed.

And so I think the data centers around the world recognize this, that this is the best way to deploy resources, deploy capital going forward for data centers. This is true for the world's clouds and you're seeing a whole crop of new GPU specialty -- GPU specialized cloud service providers. One of the famous ones is CoreWeave and they're doing incredibly well. But you're seeing the regional GPU specialist service providers all over the world now. And it's because they all recognize the same thing, that the best way to invest their capital going forward is to put it into accelerated computing and generative AI.

We're also seeing that enterprises want to do that. But in order for enterprises to do it, you have to support the management system, the operating system, the security and software-defined data center approach of enterprises, and that's all VMware. And we've been working several years with VMware to make it possible for VMware to support not just the virtualization of CPUs but a virtualization of GPUs as well as the distributed computing capabilities of GPUs, supporting NVIDIA's BlueField for high-performance networking.

And all of the generative AI libraries that we've been working on is now going to be offered as a special SKU by VMware's sales force, which is, as we all know, quite large because they reach some several hundred thousand VMware customers around the world. And this new SKU is going to be called VMware Private AI Foundation. And this will be a new SKU that makes it possible for enterprises. And in combination with HP, Dell, and Lenovo's new server offerings based on L40S, any enterprise could have a state-of-the-art AI data center and be able to engage generative AI.

And so I think the answer to that question is hard to predict exactly what's going to happen quarter-to-quarter. But I think the trend is very, very clear now that we're seeing a platform shift.

Operator

Next, we'll go to Timothy Arcuri with UBS. Your line is now open.

Timothy Arcuri

Thanks a lot. Can you talk about the attach rate of your networking solutions to your -- to the compute that you're shipping? In other words, is like half of your compute shipping with your networking solutions more than half, less than half? And is this something that maybe you can use to prioritize allocation of the GPUs? Thank you.

Jensen Huang

Well, working backwards, we don't use that to prioritize the allocation of our GPUs. We let customers decide what networking they would like to use. And for the customers that are building very large infrastructure, InfiniBand is, I hate to say it, kind of a no-brainer. And the reason for that because the efficiency of InfiniBand is so significant, some 10%, 15%, 20% higher throughput for $1 billion infrastructure translates to enormous savings. Basically, the networking is free.

And so, if you have a single application, if you will, infrastructure or it's largely dedicated to large language models or large AI systems, InfiniBand is really a terrific choice. However, if you're hosting for a lot of different users and Ethernet is really core to the way you manage your data center, we have an excellent solution there that we had just recently announced and it's called Spectrum-X. Well, we're going to bring the capabilities, if you will, not all of it, but some of it, of the capabilities of InfiniBand to Ethernet so that we can also, within the environment of Ethernet, allow you to – enable you to get excellent generative AI capabilities.

So Spectrum-X is just ramping now. It requires BlueField-3 and it supports both our Spectrum-2 and Spectrum-3 Ethernet switches. And the additional performance is really spectacular. BlueField-3 makes it possible and a whole bunch of software that goes along with it. BlueField, as all of you know, is a project really dear to my heart, and it's off to just a tremendous start. I think it's a home run. This is the concept of in-network computing and putting a lot of software in the computing fabric is being realized with BlueField-3, and it is going to be a home run.

Operator

Our final question comes from the line of Ben Reitzes with Melius. Your line is now open.

Benjamin Reitzes

Hi. Good afternoon. Good evening. Thank you for the question, putting me in here. My question is with regard to DGX Cloud. Can you talk about the reception that you're seeing and how the momentum is going? And then Colette, can you also talk about your software business? What is the run rate right now and the materiality of that business? And it does seem like it's already helping margins a bit. Thank you very much.

Jensen Huang

DGX Cloud's strategy, let me start there. DGX Cloud's strategy is to achieve several things: number one, to enable a really close partnership between us and the world's

CSPs. We recognize that many of our -- we work with some 30,000 companies around the world. 15,000 of them are startups. Thousands of them are generative AI companies and the fastest-growing segment, of course, is generative AI. We're working with all of the world's AI start-ups. And ultimately, they would like to be able to land in one of the world's leading clouds. And so we built DGX Cloud as a footprint inside the world's leading clouds so that we could simultaneously work with all of our AI partners and help blend them easily in one of our cloud partners.

The second benefit is that it allows our CSPs and ourselves to work really closely together to improve the performance of hyperscale clouds, which is historically designed for multi-tenancy and not designed for high-performance distributed computing like generative AI. And so to be able to work closely architecturally to have our engineers work hand in hand to improve the networking performance and the computing performance has been really powerful, really terrific.

And then thirdly, of course, NVIDIA uses very large infrastructures ourselves. And our self-driving car team, our NVIDIA research team, our generative AI team, our language model team, the amount of infrastructure that we need is quite significant. And none of our optimizing compilers are possible without our DGX systems. Even compilers these days require AI, and optimizing software and infrastructure software requires AI to even develop. It's been well publicized that our engineering uses AI to design our chips.

And so the internal -- our own consumption of AI, our robotics team, so on and so forth, Omniverse teams, so on and so forth, all needs AI. And so our internal consumption is quite large as well, and we land that in DGX Cloud. And so DGX Cloud has multiple use cases, multiple drivers, and it's been off to just an enormous success. And our CSPs love it, the developers love it and our own internal engineers are clamoring to have more of it. And it's a great way for us to engage and work closely with all of the AI ecosystem around the world.

Colette Kress

And let's see if I can answer your question regarding our software revenue. In part of our opening remarks that we made as well, remember, software is a part of almost all of our products, whether they're our Data Center products, GPU systems or any of our products within gaming and our future automotive products. You're correct, we're also selling it in a standalone business. And that stand-alone software continues to grow where we are providing both the software services, upgrades across there as well.

Now we're seeing, at this point, probably hundreds of millions of dollars annually for our software business, and we are looking at NVIDIA AI enterprise to be included with many of the products that we're selling, such as our DGX, such as our PCIe versions of our H100. And I think we're going to see more availability even with our CSP marketplaces. So we're off to a great start, and I do believe we'll see this continue to grow going forward.

Operator

And that does conclude today's question-and-answer session. I'll turn the call back over to Jensen Huang for any additional or closing remarks.

Jensen Huang

A new computing era has begun. The industry is simultaneously going through 2 platform transitions, accelerated computing and generative AI. Data centers are making a platform shift from general purpose to accelerated computing. The $1 trillion of global data centers will transition to accelerated computing to achieve an order of magnitude better performance, energy efficiency and cost. Accelerated computing enabled generative AI, which is now driving a platform shift in software and enabling new, never-before possible applications. Together, accelerated computing and generative AI are driving a broad-based computer industry platform shift.

Our demand is tremendous. We are significantly expanding our production capacity. Supply will substantially increase for the rest of this year and next year. NVIDIA has been preparing for this for over two decades and has created a new computing platform that the world's industry -- world's industries can build upon. What makes NVIDIA special are: one, architecture. NVIDIA accelerates everything from data processing, training, inference, every AI model, real-time speech to computer vision, and giant recommenders to vector databases. The performance and versatility of our architecture translates to the lowest data center TCO and best energy efficiency.

Two, installed base. NVIDIA has hundreds of millions of CUDA-compatible GPUs worldwide. Developers need a large installed base to reach end users and grow their business. NVIDIA is the developer's preferred platform. More developers create more applications that make NVIDIA more valuable for customers. Three, reach. NVIDIA is in clouds, enterprise data centers, industrial edge, PCs, workstations, instruments and robotics. Each has fundamentally unique computing models and ecosystems. System suppliers like OEMs, computer OEMs can confidently invest in NVIDIA because we offer significant market demand and reach. Scale and velocity. NVIDIA has achieved

significant scale and is 100% invested in accelerated computing and generative AI. Our ecosystem partners can trust that we have the expertise, focus and scale to deliver a strong road map and reach to help them grow.

We are accelerating because of the additive results of these capabilities. We're upgrading and adding new products about every six months versus every two years to address the expanding universe of generative AI. While we increased the output of H100 for training and inference of large language models, we're ramping up our new L40S universal GPU for scale, for cloud scale-out and enterprise servers. Spectrum-X, which consists of our Ethernet switch, BlueField-3 Super NIC and software helps customers who want the best possible AI performance on Ethernet infrastructures. Customers are already working on next-generation accelerated computing and generative AI with our Grace Hopper.

We're extending NVIDIA AI to the world's enterprises that demand generative AI but with the model privacy, security and sovereignty. Together with the world's leading enterprise IT companies, Accenture, Adobe, Getty, Hugging Face, Snowflake, ServiceNow, VMware and WPP and our enterprise system partners, Dell, HPE, and Lenovo, we are bringing generative AI to the world's enterprise. We're building NVIDIA Omniverse to digitalize and enable the world's multi-trillion dollar heavy industries to use generative AI to automate how they build and operate physical assets and achieve greater productivity. Generative AI starts in the cloud, but the most significant opportunities are in the world's largest industries, where companies can realize trillions of dollars of productivity gains. It is an exciting time for NVIDIA, our customers, partners and the entire ecosystem to drive this generational shift in computing. We look forward to updating you on our progress next quarter.

Operator

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