NVIDIA Corp. (NASDAQ:NVDA) Q1 2024 Earnings Conference Call May 24, 2023 5:00 PM ET

Company Participants

Simona Jankowski - VP, IR

Colette Kress - EVP & CFO

Jensen Huang - Co-Founder, CEO & President

Conference Call Participants

Toshiya Hari - Goldman Sachs

C.J. Muse - Evercore ISI

Vivek Arya - Bank of America Merrill Lynch

Aaron Rakers - Wells Fargo Securities

Timothy Arcuri - UBS

Stacy Rasgon - Sanford C. Bernstein & Co.

Joseph Moore - Morgan Stanley

Harlan Sur - JP Morgan

Matt Ramsay - Cowen

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 First 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'll 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 first 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 second 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, May 24, 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. Q1 revenue was $7.19 billion, up 19% sequentially and down 13% year-on-year. Strong sequential growth was driven by record data center revenue, with our gaming and professional visualization platforms emerging from channel inventory corrections.

Starting with data center, record revenue of $4.28 billion was up 18% sequentially and up 14% year-on-year, on strong growth by accelerated computing platform worldwide. Generative AI is driving exponential growth in compute requirements and a fast transition to NVIDIA accelerated computing, which is the most versatile, most energy-efficient, and the lowest TCO approach to train and deploy AI. Generative AI drove significant upside in demand for our products, creating opportunities and broad-based global growth across our markets.

Let me give you some color across our three major customer categories, cloud service providers or CSPs, consumer Internet companies, and enterprises. First, CSPs around

the world are racing to deploy our flagship Hopper and Ampere architecture GPUs to meet the surge in interest from both enterprise and consumer AI applications for training and inference. Multiple CSPs announced the availability of H100 on their platforms, including private previews at Microsoft Azure, Google Cloud, and Oracle Cloud Infrastructure, upcoming offerings at AWS, and general availability at emerging GPU specialized cloud providers like CoreWeave and Lambda. In addition to enterprise AI adoption, these CSPs are serving strong demand for H100 from Generative AI pioneers.

Second, consumer Internet companies are also at the forefront of adopting Generative AI and deep learning-based recommendation systems, driving strong growth. For example, Meta has now deployed it's H100 powered Grand Teton AI supercomputer for its AI production and research teams.

Third, enterprise demand for AI and accelerated computing is strong. We are seeing momentum in verticals such as automotive, financial services, healthcare, and telecom, where AI and accelerated computing are quickly becoming integral to customers' innovation roadmaps and competitive positioning. For example, Bloomberg announced it has a 50 billion parameter model, BloombergGPT, to help with financial natural language processing tasks such as sentiment analysis, named entity recognition, news classification, and question-answering.

Auto Insurance company, CCC Intelligent Solutions is using AI for estimating repairs. And AT&T is working with us on AI to improve fleet dispatches so their field technicians can better serve customers. Among other enterprise customers using NVIDIA AI are Deloitte for logistics and customer service and Amgen for drug discovery and protein engineering. This quarter, we started shipping DGX H100, our Hopper generation AI system, which customers can deploy on-prem. And with the launch of DGX Cloud through our partnership with Microsoft Azure, Google Cloud, and Oracle Cloud Infrastructure, we deliver the promise of NVIDIA DGX to customers from the cloud.

Whether the customers deploy DGX on-prem or via DGX Cloud, they get access to NVIDIA AI software, including NVIDIA Base Command, and AI frameworks, and pre-trained models. We provide them with the blueprint for building and operating AI, spanning our expertise across systems, algorithms, data processing, and training methods.

We also announced NVIDIA AI Foundations, which are model foundry services available on DGX Cloud, that enable businesses to build, refine, and operate custom large language models and generative AI models, trained with our own proprietary data,

created for unique domain-specific tasks. They include NVIDIA NeMo for large language models, NVIDIA Picasso for images, video, and 3D, and NVIDIA BioNeMo for life sciences.

Each service has six elements, pre-trained models, frameworks for data processing and curation, proprietary knowledge-based sector databases, systems for fine-tuning, aligning, and guardrailing, optimized inference engines, and support from NVIDIA experts to help enterprises fine-tune models for their custom use cases.

ServiceNow, a leading enterprise services platform is an early adopter of DGX Cloud and NeMo. They are developing custom large language models trained on data specifically for the ServiceNow platform. Our collaboration will let ServiceNow create new enterprise-grade generative AI offerings, with the 1,000s of enterprises worldwide running on the ServiceNow platform, including for IT departments, customer service teams, employees, and developers.

Generative AI is also driving a step-function increase in inference workloads. Because of their size and complexities, these workloads require acceleration. The latest MLPerf industry benchmark released in April showed NVIDIA's inference platform deliver performance that is orders of magnitude ahead of the industry, with unmatched versatility across diverse workloads.

To help customers deploy generative AI applications at scale, at GTC, we announced four major new inference platforms that leverage the NVIDIA AI software stack. These include L4 Tensor Core GPU for AI video, L40 for Omniverse, and graphics rendering, H100 NVL for large language models, and the Grace Hopper Superchip for LLMs and also, recommendation systems and vector databases.

Google Cloud is the first CSP to adopt our L4 inference platform with the launch of its G2 virtual machines for generative AI inference and other workloads such as Google Cloud Dataproc, Google AlphaFold, and Google Cloud's Immersive Stream, which render 3D and AR experiences. In addition, Google is integrating our Triton inference server with Google Kubernetes engine and its cloud-based Vertex AI platform.

In networking, we saw strong demand at both CSPs and enterprise customers for generative AI and accelerated computing, which require high-performance networking like NVIDIA's Mellanox networking platforms. Demand relating to general purpose CPU infrastructure remain soft. As generative AI applications grow in size and complexity, high performance networks become essential for delivering accelerated computing at data center scale to meet the enormous demand of all training and inferencing.

Our 400 gig Quantum-2 InfiniBand platform is the gold standard for AI dedicated infrastructure, with broad adoption across major cloud and consumer Internet platforms such as Microsoft Azure. With the combination of in-network computing technology and the industry's only end-to-end data center scale, optimized software stack, customers routinely enjoy a 20% increase in throughput for their sizable infrastructure investment.

For multi-tenant cloud transitioning to support generative AI our high-speed Ethernet platform with BlueField-3 DPUs and Spectrum-4 Ethernet switching, offers the highest available Ethernet network performance. BlueField-3 is in production and has been adopted by multiple hyperscale and CSP customers, including Microsoft Azure, Oracle Cloud, CoreWeave, Baidu, and others. We look forward to sharing more about our 400 gig Spectrum-4 accelerated AI networking platform next week at the COMPUTEX Conference in Taiwan.

Lastly, our Grace data center CPU is sampling with customers. At this week's International Supercomputing Conference in Germany, the University of Bristol announced a new supercomputer based on the NVIDIA Grace CPU Superchip, which is 6x more energy-efficient than the previous supercomputer. This adds to the growing momentum for Grace with both CPU only and CPU/GPU opportunities across AI and cloud and supercomputing applications. The coming wave of BlueField-3, Grace and Grace Hopper Superchips will enable a new generation of super energy efficient accelerated data centers.

Now, let's move to gaming. Gaming revenue of $2.24 billion was up 22% sequentially, and down 38% year-on-year. Strong sequential growth was driven by sales of the 40 Series GeForce RTX GPUs for both notebooks and desktops. Overall, end demand was solid, and consistent with seasonality, demonstrating resilience against a challenging consumer spending backdrop. The GeForce RTX 40 Series GPU laptops are off to a great start, featuring four NVIDIA inventions, RTX Path Tracing, DLSS 3 AI rendering, Reflex Ultra-Low Latency rendering, and Max-Q, energy efficient technologies. They deliver tremendous gains in industrial design, performance and battery life for gamers and creators.

And like our desktop offerings, 40 Series laptops support the NVIDIA Studio platform or software technologies, including acceleration for creative data science and AI workflows, and Omniverse, giving content creators unmatched tools and capabilities. In desktop, we ramped the RTX 4070, which joined the previously launched RTX 4090, 4080, and 4070 Ti GPUs. The RTX 4070 is nearly 3x faster than the RTX 2070 and offers our large installed-base a spectacular upgrade.

Last week, we launched the 60 family, RTX 4060, and 4060 Ti, bringing our newest architecture to the world's core gamers starting at just $299. These GPUs for the first time provide 2x the performance of the latest gaming console at mainstream price points. The 4060 Ti is available starting today, while the 4060 will be available in July.

Generative AI will be transformative to gaming and content creation from development to run time. At the Microsoft Build Developer Conference earlier this week, we showcased how Windows PCs and workstations with NVIDIA RTX GPUs will be AI-powered at their core. NVIDIA and Microsoft have collaborated on end-to-end software engineering, spanning from the Windows operating system to the NVIDIA graphics drivers, and NeMo's LLM framework to help make Windows on NVIDIA RTX Tensor Core GPUs, a supercharged platform for generative AI.

Last quarter, we announced a partnership with Microsoft to bring Xbox PC games to GeForce NOW. The first game from this partnership, Gears 5 is now available with more set to be released in the coming months. There are now over 1,600 games on GeForce NOW, the richest content available on any cloud gaming service.

Moving to Pro Visualization. Revenue of $295 million was up 31% sequentially, and down 53% year-on-year. Sequential growth was driven by stronger workstation demand across both mobile and desktop form factors, with strength in key verticals such as Public Sector, Healthcare, and Automotive. We believe the channel inventory correction is behind us. The ramp of our Ada Lovelace GPU architecture in workstations kicked-off a major product cycle.

At GTC, we announced six new RTX GPUs for laptops and desktop workstations, with further rollout planned in the coming quarters. Generative AI is a major new workload for NVIDIA-powered workstation. Our collaboration with Microsoft transformed windows into the ideal platform for creators and designers, harnessing generative AI to elevate their creativity and productivity.

At GTC, we announced NVIDIA Omniverse Cloud, an NVIDIA fully managed service running in Microsoft Azure that includes the full suite of Omniverse applications and NVIDIA OVX infrastructure. Using this full stack cloud environment, customers can design, develop, deploy, and manage industrial metaverse applications. NVIDIA Omniverse Cloud will be available starting in the second half of this year.

Microsoft NVIDIA will also connect Office 365 applications with Omniverse. Omniverse Cloud is being used by companies to digitalize their workflows from design and engineering to smart factories and 3D content generation for marketing. The

automotive industry has been a leading early adopter of Omniverse, including companies such as BMW Group, Geely Lotus, General Motors, and Jaguar Land Rover.

Moving to Automotive. Revenue was $296 million, up 1% sequentially, and up 114% from a year ago. Our strong year-on-year growth was driven by the ramp of the NVIDIA DRIVE Orin across a number of new energy vehicles. As we announced in March, our automotive design win pipeline over the next six years now stands at $14 billion, up from $11 billion a year ago, giving us visibility into continued growth over the coming years.

Sequentially, growth moderated as some NEV customers in China are adjusting their production schedules to reflect slower than expected demand growth. We expect this dynamic to linger for the rest of the calendar year. During the quarter, we expanded our partnership with BYD, the world's leading manufacturer of NEVs. Our new design win will extend BYD's use of the DRIVE Orin to its next-generation high-volume Dynasty, and Ocean series of vehicles, set to start production in calendar 2024.

Moving to the rest of the P&L. GAAP gross margins was 64.6%, and non-GAAP gross margins were 66.8%. Gross margins have now largely recovered to prior peak level, and we have absorbed higher costs, and offset them by innovating and delivering higher valued products as well as products incorporating more and more software.

Sequentially, GAAP operating expenses were down 3%, and non-GAAP operating expenses were down 1%, We have held OpEx at roughly the same level over the last past four quarters. We're working through the inventory corrections in gaming and professional visualization. We now expect to increase investments in the business while also delivering operating leverage.

We returned $99 million to shareholders in the form of cash dividends. At the end of the Q1, we have approximately $7 billion remaining under our share repurchase authorization through December 2023.

Let me turn to the outlook for the second quarter fiscal '24. Total revenue is expected to be $11 billion, plus or minus 2%. We expect this sequential growth to largely be driven by data center, reflecting a steep increase in demand related to generative AI and large language models. This demand has extended our data center visibility out a few quarters and we have procured substantially higher supply for the second half of the year.

GAAP and non-GAAP gross margins are expected to be 68.6% and 70% respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to

be approximately $2.71 billion and $1.9 billion, respectively. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $90 million, excluding gains and losses from non-affiliated investments.

GAAP and non-GAAP tax rates are expected to be 14%, plus or minus 1%, excluding any discrete items. Capital expenditures are expected to be approximately $300 million to $350 million. Further financial details are included in the CFO commentary and other information available on our IR website.

In closing, let me highlight some of the upcoming events, Jensen will give the COMPUTEX keynote address in person in Taipei this coming Monday, May 29 local time, which will be Sunday evening in the U.S. In addition, we will be attending the BofA Global Technology Conference in San Francisco on June 6. And Rosenblatt Virtual Technology Summit on The Age of AI on June 7, and the New Street Future of Transportation Virtual Conference on June 12. Our earnings call to discuss the results of our second quarter fiscal '24 is scheduled for Wednesday, August 23.

Well, that covers our opening remarks. We're now going to open the call for questions. Operator, would you please poll for questions?

Question-and-Answer Session

Operator

Thank you. [Operator Instructions] We'll take our first question from Toshiya Hari with Goldman Sachs. Your line is open.

Toshiya Hari

Hi. Good afternoon. Thank you so much for taking the question and congrats on the strong results, and incredible outlook. Just one question on data center. Colette, you mentioned the vast majority of the sequential increase in revenue this quarter will come from data center. I was curious what the construct is there, if you can speak to, what the key drivers are from April to July and perhaps more importantly, you talked about visibility into the second half of the year. I'm guessing it's more of a supply problem at this point, what kind of sequential growth beyond the July quarter can your supply chain support at this point? Thank you.

Colette Kress

Okay. So, a lot of different questions there. So, let me see if I can start and I am sure Jensen will have some following up comments. So when we talk about our sequential

growth that we're expecting between Q1 and Q2, our generative AI large language models are driving the surge in demand, and it's broad-based across both our consumer Internet companies, our CSPs, our enterprises, and our AI start-ups.

It is also interest in both of our architectures, both of our Hopper latest architecture as well as our Ampere architecture. This is not surprising as we generally often sell both of our architectures at the same time. This is also a key area where deep recommendators are driving growth. And we also expect to see growth both in our computing as well as in our networking business. So, those are some of the key things that we have baked in when we think about the guidance we provided to Q2.

We also surfaced in our opening remarks that we are working on both supply today for this quarter, but we have also procured a substantial amount of supply for the second half. We have significant supply chain flow to serve our significant customer demand that we see, and this is demand that we see across a wide range of different customers.

They are building platforms for some of the largest enterprises, but also setting things up at the CSPs and the large consumer Internet companies. So, we have visibility right now for our data center demand that has probably extended out a few quarters and that's led us to working on quickly procuring that substantial supply for the second half. I'm going to pause there and see if Jensen wants to add a little bit more.

Jensen Huang

I thought that was great color. Thank you.

Operator

Next we'll go to C.J. Muse with Evercore ISI. Your line is open.

C.J. Muse

Yeah. Good afternoon. Thank you for taking the question. I guess with data center, you are essentially doubling quarter-on-quarter, two natural kind of questions that relate to one another come to mind. Number one, where are we in terms of driving acceleration into servers to support AI? And as part of that, as you deal with longer cycle times with TSMC and your other partners, how are you thinking about managing their commitments there with where you want to manage your lead times in the coming years to best match that supply and demand? Thanks so much.

Jensen Huang

Yeah, C.J. Thanks for the question. I'll start backwards. The -- remember, we were in full production of both Ampere and Hopper when I -- when the ChatGPT moment came. And it helped everybody crystallize how to transition from the technology of large language models to a product and service based on a chatbot.

The integration of guardrails and alignment systems were through reinforcement learning human feedback, knowledge vector data bases for proprietary knowledge, connection to search, all of that came together in a really wonderful way and it's the reason why I call it the iPhone moment, all the technology came together and helped everybody realize what an amazing product that can be and what capabilities it can have. And so we were already in full production. NVIDIA's supply chain flow and our supply chain is very significant as you know.

And we build supercomputers in volume, and these are giant systems and we build them in volume. It includes, of course, the GPUs, but on our GPUs, the system boards have 35,000 other components. And the networking, and fiberoptics, and the incredible transceivers and the NICs, the Smart NICs, the switches, all of that has to come together in order for us to stand-up a data center. And so we were already in full production when the moment came. We had to really significantly increase our procurement substantially for the second half as Colette said.

Now, let me talk about the bigger picture and why the entire world's data centers are moving towards accelerated computing. It's been known for some time and you've heard me talk about it, that accelerated computing is a full stack problem, but it is full stack challenge, but if we could successfully do it in a large number of application domain has taken us 15 years.

If - sufficiently that almost the entire data centers' major applications could be accelerated you could reduce the amount of energy consumed and the amount of cost for our data center substantially by an order of magnitude. It takes -- it costs a lot of money to do it because you have to do all the software and everything and you have to build all the systems and so on and so forth, but we've been at it for 15 years.

And what happened is, when generative AI came along, it triggered a killer app for this computing platform that's been in preparation for some time. And so, now we see ourselves in two simultaneous transitions. The world's $1 trillion data center is nearly populated entirely by CPUs today, and $1 trillion, $250 billion a year, it's growing of course. But over the last four years, call it a $1 trillion worth of infrastructure installed. And it's all completely based on CPUs and dumb NICs. It's basically unaccelerated.

In the future, it's fairly clear now with this -- with generative AI becoming the primary workload of most of the world's data centers generating information, it is very clear now that -- and the fact that accelerated computing is so energy efficient, that the budget of the data center will shift very dramatically towards accelerated computing and you're seeing that now. We're going through that moment right now as we speak. While the world's data center CapEx budget is limited but at the same time we're seeing incredible orders to retool the world's data centers.

And so I think you're starting -- you're seeing the beginning of call it a 10-year transition to basically recycle or reclaim the world's data centers and build it out as accelerated computing. You'll have a pretty dramatic shift in the spend of the data center from traditional computing, and to accelerated computing with smart NICs, smart switches, of course, GPUs, and the workload is going to be predominantly generative AI.

Operator

And we'll move to our next question, Vivek Arya with BofA Securities. Your line is open.

Vivek Arya

Thanks for the question. Could I just wanted to clarify does visibility mean data center sales can continue to grow sequentially in Q3 and Q4 or do they sustain at Q2 levels? I just wanted to clarify that. And then Jensen, my question is that, given this very strong demand environment, what does that do to the competitive landscape? Does it invite more competition in terms of custom ASICs? Does it invite more competition in terms of other GPU solutions or other kinds of solutions? How do you see the competitive landscape change over the next two to three years?

Colette Kress

Yeah, Vivek. Thanks for the question. Let me see if I can add a little bit more color. We believe that the supply that we will have for the second half of the year will be substantially larger than H1. So, we are expecting not only the demand that we just saw in this last quarter, the demand that we have in Q2 for our forecast, but also planning on seeing something in the second half of the year. We just have to be careful here. But we are not here to guide on the second half of that. Yes, we do plan a substantial increase in the second half compared to the first half.

Jensen Huang

Regarding competition, we have competition from every direction. Start-ups really-really well-funded and innovative startups, countless of them all over the world. We have competitions from existing semiconductor companies. We have competition from CSPs with internal projects. And many of you know about most of these. And so, we're mindful of competition all the time, and we get competition all the time. But NVIDIA's value proposition at the core is, we are the lowest cost solution. We're the lowest TCO solution.

And the reason for that is, because accelerated computing is two things that I talk about often, which is it's a full stack problem, it's a full stack challenge, you have to engineer all of the software and all the libraries and all the algorithms, integrated them into and optimize the frameworks and optimize it for the architecture of not just one ship but the architecture of an entire data center, all the way into the frameworks, all the way into the models. And the amount of engineering and distributed computing, fundamental computer science work is really quite extraordinary. It is the hardest computing as we know.

And so, number one, it's a full stack challenge and you have to optimize it across the whole thing and across just the mind blowing number of stacks. We have 400 acceleration libraries. As you know, the amount of libraries and frameworks that we accelerate is pretty mind blowing.

The second part is that generative AI is a large scale problem, and it's a data center scale problem, it's another way of thinking that the computer is the data center or to data center is the computer, it's not the chip, it's the data center and it's never happened like this before. And in this particular environment, your networking operating system, your distributed computing engines, your understanding of the architecture of the networking gear, the switches and the computing systems, the computing fabric, that entire system is your computer and that's what you're trying to operate. And so in order to get the best performance, you have to understand full stack and you have to understand data center scale, and that's what accelerated computing is.

The second thing is that -- utilization, which talks about the amount of the types of applications that you can accelerate and diversity of our architecture keeps that utilization high. If you can do one thing and do one thing only and incredibly fast, then your data center is largely underutilized and it's hard to scale that up. And the thing is, universal GPU in fact that we accelerate so many stacks, makes our utilization incredibly high, and so number one is throughput, and that's software -- that's a software-intensive problems and data center architecture problems. The second is utilization versatility problem and the third is just data center expertise. We've built five

data centers of our own and we've helped companies all over the world build data centers and we integrate our architecture into all the world's clouds.

From the moment of delivery of the product to do standing up in the deployment, the time to operations of the data center is measured not -- if you're not good at it and not – not proficient at it, it could take months. Standing up a supercomputer, let's see, some of the largest supercomputers in the world were installed about a year and a half ago and now they're coming online, and so it's not – it unheard of to see a delivery to operations of about a year.

Our delivery to operation is measured in weeks. And we've taken data centers and supercomputers and we've turned it into products, and the expertise of the team in doing that is incredible, and so. So, our value proposition is in the final analysis, all of this technology translates in the infrastructure, the highest throughput in the lowest possible cost. And so I think -- our market is of course very, very competitive, very large. But the challenge is really-really great.

Operator

Next we go to Aaron Rakers with Wells Fargo. Your line is open.

Aaron Rakers

Yeah. Thank you for taking the question and congrats on the quarter. As we kind of think about unpacking the various different growth drivers of the data center business going forward, I'm curious, Colette, of just how we should think about the monetization effect of software, considering that the expansion of your cloud service agreements continues to grow? I'm curious of what -- where do you think we're at in terms of that approach in terms of the AI enterprise software suite and other drivers of software only revenue going forward?

Colette Kress

Thanks for the question. Software is really important to our accelerated platforms. Not only do we have a substantial amount of software that we are including in our nearest architecture and essentially, all products that we have. We are now with many different models to help customers start their work in generative AI and accelerated computing. So, anything that we have here from a DGX Cloud and providing those services, helping them build models or as we've discussed the importance of NVIDIA AI enterprise, essentially that operating system for AI. So, all things should continue to grow as we go forward, both the architecture and the infrastructure, as well as both availability of this

offering, our ability to monetize [indiscernible] as well. I'll turn it over to Jensen, if he needs to add.

Jensen Huang

Yeah. We can see in real-time the growth of generative AI and CSPs, both for training the models, refining the models, as well as deploying the models. As Colette said earlier, inference is now a major driver of accelerated computing because generative AI is used so capably in so many applications already.

There are two segments that requires a new stack of software and the two segments are enterprise and industrials. Enterprise requires a new stack of software, because many enterprises need to have all the capabilities that we've talked about, whether it's large language models, the ability to adapt, and for your proprietary use-case and your proprietary data, align it to your own principles, and your own operating domains.

You want to have the ability to be able to do that in a high performance computing sandbox, and we that DGX Cloud, and create a model. Then you want to deploy your chatbot or your AI in any Cloud, because you have services and you have agreements with multiple Cloud vendors and depending on the applications, you might deploy it on various clouds.

And for the enterprise, we have NVIDIA AI Foundation for helping you create custom models and we have NVIDIA AI Enterprise. NVIDIA AI Enterprise is the only accelerated stack, GPU accelerated stack in the world that is Enterprise safe, and Enterprise supported. There are a constant patching that you have to do, there are 4,000 different packages that buildup NVIDIA AI Enterprise and represents the operating engine, end-to-end operating engine of the entire AI workflow.

It's the only one of its kind from data ingestion, data processing, obviously, in order to train an AI model, you have a lot of data, you have to process and package up and curate, and align and there's just a whole bunch of stuff that you have to do to the data to prepare it for training. That amount of data, that could consume some 40%, 50%, 60% of your computing time and so, data processing is very big deal.

And then the second aspect of it is training the model, refining the model and the third is deploying model for inferencing. NVIDIA AI Enterprise supports and patches and security patches continuously all of those 4,000 packages of software. And for an Enterprise that wants to deploy their engines, just like they want to deploy Red Hat Linux, this is incredibly complicated software in order to deploy that in every cloud and

as well as on-prem, it has to be secure, it has to be supported. And so, NVIDIA AI Enterprise is the second point.

The third is Omniverse. Just as people are starting to realize that you need to align an AI to ethics, the same for robotics, you need to align the AI for physics. And aligning an AI for ethics includes a technology called reinforcement learning human feedback. In the case of industrial applications and robotics, it's reinforcement learning Omniverse feedback. And Omniverse is a vital engine for software defined in robotic applications and industries. And so, Omniverse also needs to be a cloud service platform.

And so our software stack, the three software stacks, AI Foundation, AI Enterprise and Omniverse runs in all of the world's clouds that we have partnerships, DGX Cloud partnerships with. Azure, we have partnerships on both AI as well as Omniverse. With GTP and Oracle, we have great partnerships in DGX Cloud for AI and AI Enterprise is integrated into all three of them and so I think the -- in order to for us to extend the reach of AI beyond the cloud, and into the world's Enterprise and into the world's industries, you need two new types of -- you need new software stacks in order to make that happen and by putting it in the cloud, integrate it into the world's CSP clouds, it's a great way for us to partner with the sales and the marketing team and the leadership team of all the cloud vendors.

Operator

Next we'll go to Timothy Arcuri with UBS. Your line is Open.

Tim Arcuri

Thanks a lot. I had a question and then I had a clarification as well. So, the question first is, Jensen, on the InfiniBand versus Ethernet argument, can you sort of speak to that debate and maybe how you see it playing out? I know you need the low late -- the low latency of InfiniBand for AI, but can you sort of talk about the attach rate of your InfiniBand solutions to what you're shipping on the core compute side and maybe whether that's similarly crowding out Ethernet like you are with on the compute side? And then the clarification, Colette, is that there wasn't a share buyback despite you still having about $7 billion on the share repo authorization. Was that just timing? Thanks.

Jensen Huang

Colette, how about you go first? You should take the question.

Colette Kress

That is correct. We have $7 billion available in recurrent authorization for repurchases. We did not repurchase anything in this last quarter, but we do repurchase opportunistically and we'll consider that as we go forward as well. Thankyou

Jensen Huang

InfiniBand and Ethernet are Target different applications in a data center. All right. They both have their place. InfiniBand had a record quarter. We're going to have a giant record year. And InfiniBand has a really -- NVIDIA's Quantum InfiniBand has an exceptional roadmap. It's going to be really incredible. But the two networks are very different. InfiniBand is designed for an AI factory, if you will. If that data center is running a few applications for a few people for a specific use case and it's doing it continuously and that infrastructure costs you, pick a number, $500 million.

The difference between InfiniBand and Ethernet could be 15%, 20% in overall throughput. And if you spent $500 million in an infrastructure and the difference is 10% to 20% and it's a $100 million, InfiniBand is basically free. That's the reason why people use it. InfiniBand is effectively free. The difference in data center throughput is just -- it's too great to ignore, and you're using it for that one application and so, however, if your data center is a cloud datacenter and its multi-tenant. It's a bunch of little jobs, a bunch of little jobs and is shared by millions of people.

Then Ethernet is really do I answer? There's a new segment in the middle where the Cloud is becoming a generative AI cloud. It's not only AI factory per se. But it's still a multi-tenant Cloud but it wants to run generative AI workloads. This new segment is a wonderful opportunity and at COMPUTEX, I referred to it at the last GTC. At COMPUTEX, we're going to announce a major product line for this segment, which is Ethernet focused generative AI application type of clouds. But InfiniBand is doing fantastically and we're doing record numbers quarter-on-quarter year-on-year.

Operator

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

Stacy Rasgon

Hi, guys. Thanks for taking my question. I had a question on inference versus training for generative AI. So, you're talking about inference as being a very large opportunity. I guess, two sub parts to that. Is that, besides inference basically scales with like the usage versus like training is more of a one-and-done. And can you give us some sort of even if it's just like qualitatively, like if do you think are influence is bigger than training or

vice-versa, like if it's bigger, how much bigger? Is it like the opportunity, is it 5x, is it 10x, is there anything you can give us on those two workloads within generative AI, it would be helpful.

Jensen Huang

Yeah. I'll work backwards. You're never done with training. You're always -- every time you deploy, you're collecting new data. When you collect new data, you train with the new data. And so, you're never done training. You're never done producing and processing a Vector database that augments the large language model. You're never done with vectorizing all of the collected structured -- unstructured data that you have.

And so, whether you're building a recommender system, a large language model, a vector database, these are probably the three major applications of the three core engines, if you will, of the future of computing. It's all a bunch of other stuff, but obviously these are very three very important ones. They are always running.

You're going to see that more-and-more companies realize they have a factory for intelligence, an intelligence factory and in that particular case, it's largely dedicated to training and processing data and vectorizing data and learning representation of the data, so on and so forth.

The inference part of it, are APIs that are either open APIs that can be connected to all kinds of applications, APIs that is integrated into workflows. But APIs of all kinds, there'll be 100s of APIs in the company, some of them they built themselves, some of them part that could -- many of them could come from companies like ServiceNow and Adobe that we're partnering with in AI Foundations. And they'll create a whole bunch of generative AI APIs that companies can then connect into their workflows or use as an application. And of course, there will be a whole bunch of Internet Service Companies.

So, I think you're seeing for the very first time simultaneously a very significant growth in the segment of AI Factories, as well as a market that -- a segment that really didn't exist before, but now it's growing exponentially, practically by the weak for AI inference with APIs. The simple way to think about it in the end, is that, the world has a $1 trillion of data center installed and they used to be 100% CPUs. In the future, we know we've heard it in enough places and I think this year there is a ISC keynote was actually about the end of Moore's Law.

We've seen it in a lot of places now that you can't reasonably scale-out data centers with general-purpose computing and that accelerated computing is the path forward and now it's got a killer app and it's got generative AI, and so the easiest way to think about

that is your $1 trillion infrastructure. Every quarters capital CapEx budget would lean very heavily into generative AI into accelerated computing infrastructure everywhere from the number of GPUs that would be used in the CapEx budget to the accelerated switches and accelerated net -- networking chips that connect them all. That the easiest way to think about that is over the next four or five, 10 years, most of that $1 trillion and then compensating adjusting for all the growth in data center still, it will be largely generative AI and so that's probably the easiest way to think about that and that's training as well as inference.

Operator

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

Joseph Moore

Great. Thank you. I want to follow-up on that, in terms of the focus on inference. It's pretty clear that this is a really big opportunity around large language models, but the cloud customers are also talking about trying to reduce cost per query by very significant amounts. You can talk about the ramifications of that for you guys, is that where some of the specialty insurance products that you launched at GTC come in and just how are you going to help your customers get the cost per query down?

Jensen Huang

Yeah. That's a great question. Whether your -- you start by building a large language model and you use that large language model very large version and you could distill them into medium, small and tiny size. And the tiny sized ones, you can put in your phone and your PC and so on and so forth and they all have good -- they all have -- it seems surprising, but they all can do the same thing. But obviously, the zero shot or the generalizeability of the large language model, the biggest one is much more versatile and it can do a lot more amazing things. And the large one would teach the smaller ones, how to be good AIs and so, you use the large one to generate prompts to align the smaller ones and so on and so forth. And so you start by building very large ones. And then you also have to train a whole bunch of smaller ones.

Now, that's exactly the reason why we have so many different sizes of our inference. You saw that I announced L4, L40, H100 NBL -- which also have H100 HGX and then we have H100 multi-node with NVLink and so there is -- you could have model sizes of any kind that you like. The other thing that's important is, these are models, but they are connected ultimately to applications. And the applications could have image in, video

out, video in, text out, image in, proteins out, text in, 3D out, video in, in the future, 3D graphics out.

So, the input and the output requires a lot of pre and post-processing. The pre and post-processing can't be ignored. And this is one of the things that most of the specialized chip arguments fall apart and it's because the length -- the model itself is only call it 25% of the data -- of the overall processing of inference. The rest of it is about preprocessing and post-processing, security, decoding, all kinds of things like that.

And so, I think the multimodality aspect of inference, the multi diversity of inference, that it's going to be done in the Cloud on-prem. It's going to be done in multi-cloud, that's the reason why we have the AI Enterprise in all the clouds. It's going to be done on-prem, it's the reason why we have a great partnership with Dell we just announced the other day, called project Helix, that's going to be integrated into third-party services. That's the reason why we have a great partnership with ServiceNow, and Adobe, because they're going to be creating a whole bunch of generative AI capabilities. And so, there's all the diversity, and the reach of generative AI is so broad, you need to have some very fundamental capabilities like what I just described, in order to really address the whole space of it.

Operator

Next we'll go to Harlan Sur with JP Morgan. Your line is open.

Harlan Sur

Hi. Good afternoon, and congratulations on the strong results and execution. I really appreciate more of the focus or some of the focus today in your networking products. I mean, it's really an integral part to sort of maximize the full performance of your compute platforms. I think so data center networking business is driving a part of $1 billion of revenues per quarter plus or minus, that's 2.5x growth from three years ago, right, when you guys acquired Mellanox.

So very strong growth, but given the very high attach of your InfiniBand, Ethernet solutions, your accelerated compute platforms, is the networking run-rate stepping up in line with your compute shipment? And then, what is the team doing to further unlock more networking bandwidth going forward just to keep pace with the significant increase in compute complexity, datasets, requirements for lower latency, better traffic predictability, and so on?

Jensen Huang

Yeah, Harlan. I really appreciate that. So, nearly everybody who thinks about AI, they think about that chip, the accelerator chip and in fact, it misses the whole point nearly completely. And I've mentioned before that accelerated computing is about the stack, about the software and networking, remember, we announced a very early-on this networking stack called DOCA and we have the acceleration library call Magnum IO. These two pieces of software are some of the crown jewels of our company. Nobody ever talks about it, because it's hard to understand, but it makes it possible for us to connect 10s of 1,000s of GPUs.

How do you connect 10s of 1000s of GPUs, if the operating system of the data center, which is the infrastructure, is not insanely great, and so that's the reason why we're so obsessed about networking in the company. And one of the great things that we have -- we have Mellanox as you know quite well, was the world's highest performance and the unambiguous leader in high performance networking, that's the reason why our two companies are together.

You also see that our network expands starting from NVLink, which is a computing fabric with a really super low latency and it communicates using memory references, not network package. And then we take NVLink, we connect it inside multiple GPUs, and I described, going beyond the GPU. And I'll talk a lot more about that at COMPUTEX in a few days. And then, that gets connected to InfiniBand, which includes the NIC, and the SmartNIC BlueField-3 that we're in full production with and the switches, all of the fiber optics that are optimized end-to-end. These things are running at an incredible line rates.

And then beyond that, if you want to connect the smart AI factory -- the smart fact -- this AI factory into your computing fabric, we have a brand new type of Ethernet that we'll be announcing at COMPUTEX, and so -- this whole area of the computing fabric extending connecting all of these GPUs and computing units together, all the way through the networking, through the switches, the software stack is insanely complicated. And so, we're -- I'm delighted you understand it, and but this -- we don't break it out, particularly, because we think of the whole thing as a computing platform as it should be.

We sell it to all of the world's data centers as components, so that they can integrate it into whatever style or architecture that they would like and we can still run our software stack. That's the reason why we break it up, it's way more complicated the way that we do it, but it makes it possible for NVIDIA's computing architecture to be integrated into anybody's data center in the world from Cloud of all different kinds to on-prem of all different kinds, all the way out to the edge to 5G and so this way of doing it is really complicated, but it gives us incredible reach.

Operator

And our last question will come from Matt Ramsay with TD Cowen. Your line is open.

Matt Ramsay

Thank you very much. Congratulations, Jensen, and to the whole team. One of the things I wanted to dig into a little bit is the DGX Cloud offering. You guys have been working on this for some time behind the scenes, where you sell in the hardware to your hyperscale partners and then lease it back for your own business, and the rest of us kind of found out about it publicly a few months ago. And as we look forward over the next number of quarters that Colette discussed to high visibility in the data center business.

Maybe you could talk a little bit about the mix you're seeing of hyperscale customers buying for their own first-party internal workloads versus their own sort of third-party, their own customers versus what of that big upside in data center going forward is systems that you're selling in, with potential to support your DGX Cloud offerings and what you've learned since you've launched it about the potential of that business. Thanks.

Jensen Huang

Yeah. Thanks Matt. It's -- without being too specific about numbers, but the ideal scenario, the ideal mix is something like 10% NVIDIA DGX Cloud and 90% the CSPs clouds, and the reason -- and our DGX Cloud is -- the NVIDIA stack is the pure NVIDIA stack. It is architected the way we like and achieves the best possible performance. It gives us the ability to partner very deeply with the CSPs to create the highest-performing infrastructure, number one.

Number two, it allows us to partner with the CSPs to create markets like, for example, we're partnering with Azure to bring Omniverse cloud to the world's industries. And the world's never had a system like that, the computing stack. Now with all the generative AI stuff and all the 3D stuff and the physics stuff, incredibly large database and really high-speed networks and low-latency networks, that kind of a virtual industrial virtual world has never existed before.

And so, we partnered with Microsoft to create Omniverse cloud inside Azure cloud. So, it allows us number two, to create new applications together and develop new markets together. And we go-to-market as one team and we benefit by getting our customers on our computing platform and they benefit by having us in their cloud, number one; but

number two, the amount of data and services and security services and all of the amazing things that Azure and GCP and OCI have, they can instantly have access to that through Omniverse cloud. And so it's a huge win-win.

And for the customers, the way that NVIDIA's cloud works for these early applications, they can do it anywhere. So one standard stack runs in all the clouds and if they would like to take their software and run it on the CSPs cloud themselves and manage it themselves, we're delighted by that, because NVIDIA AI Enterprise, NVIDIA AI Foundations. And long-term, this is going to take a little longer, but NVIDIA Omniverse will run in the CSPs clouds. Okay.

So, our goal really is to drive architecture to partner deeply in creating new markets and the new applications that we're doing and provide our customers with the flexibilities to run in their -- in their everywhere, including on-prem and so, that -- those were the primary reasons for it and it's worked out incredibly. Our partnership with the three CSPs and that we currently have DGX Cloud in and their sales force and marketing teams, their leadership team is really quite spectacular. It works great.

Operator

Thank you. I'll now turn it back over to Jensen Huang for closing remarks.

Jensen Huang

The computer industry is going through two simultaneous transitions, accelerated computing and generative AI. CPU scaling has slowed, yet computing demand is strong and now with generative AI supercharged. Accelerated computing, a full stack and data center scale approach that NVIDIA pioneered is the best path forward. There is $1 trillion installed in the global data center infrastructure, based on the general-purpose computing method of the last era. Companies are now racing to deploy accelerated computing for the generative AI era. Over the next decade, most of the world's data centers will be accelerated.

We are significantly increasing our supply to meet the surging demand. Large language models can learn information encoded in many forms. Guided by large language models, generative AI models can generate amazing content and with models to fine-tune, guardrail, align to guiding principles and ground the facts, generative AI is emerging from labs and is on its way to industrial applications.

As we scale with cloud and Internet service providers, we are also building platforms for the world's largest enterprises. Whether within one of our CSP partners or on-prem with

Dell Helix, whether on a leading enterprise platform like ServiceNow and Adobe or a bespoke with NVIDIA AI Foundations, we can help enterprises leverage their domain expertise and data to harness generative AI securely and safely.

We are ramping a wave of products in the coming quarters, including H100, our Grace and Grace Hopper super chips and our BlueField-3 and Spectrum 4 networking platform. They are all in production. They will help deliver data center scale computing that is also energy-efficient and sustainable computing. Join us next week at COMPUTEX and we'll show you what's next. Thank you.

Operator

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