NVIDIA Corp. (NASDAQ:NVDA) Q4 2023 Earnings Conference Call February 22, 2023 5:00 PM ET

Company Participants

Simona Jankowski - VP, IR

Colette Kress - EVP & CFO

Jensen Huang - Co-Founder, CEO & President

Conference Call Participants

Aaron Rakers - Wells Fargo Securities

Vivek Arya - Bank of America Merrill Lynch

Christopher Muse - Evercore ISI

Stacy Rasgon - Sanford C. Bernstein & Co.

Joseph Moore - Morgan Stanley

Atif Malik - Citigroup

Timothy Arcuri - UBS

Mark Lipacis - Jefferies

Matthew Ramsay - Cowen and Company

Operator

Good afternoon. My name is Emma, and I will be your conference operator today. At this time, I would like to welcome everyone to the NVIDIA's Fourth Quarter Earnings Call. [Operator Instructions]. Thank you. Simona Jankowski, you may begin your conference.

Simona Jankowski

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the fourth quarter of fiscal 2023. 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 the financial results for the first 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, February 22, 2023, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website.

With that, let me turn the call over to Colette.

Colette Kress

Thank you, Simona. Q4 revenue was $6.05 billion, up 2% sequentially, while down 21% year-on-year. Full year revenue was $27 billion, flat from the prior year. Starting with data center. Revenue was $3.62 billion was down 6% sequentially and up 11% year-on-year. Fiscal year revenue was $15 billion and up 41%. Hyperscale customer revenue posted strong sequential growth, though short of our expectations as some cloud service providers paused at the end of the year to recalibrate their build plans. Though we generally see tightening that reflects overall macroeconomic uncertainty, we believe this is a timing issue as the end market demand for GPUs and AI infrastructure is strong. Networking grew but a bit less than our expected on softer demand for general purpose CPU infrastructure. The total data center sequential revenue decline was driven by lower sales in China, which was largely in line with our expectations, reflecting COVID and other domestic issues.

With cloud adoption continuing to grow, we are serving an expanding list of fast-growing cloud service providers, including Oracle and GPU specialized CSPs. Revenue growth from CSP customers last year significantly outpaced that of Data Center as a whole as more enterprise customers moved to a cloud-first approach. On a trailing 4-quarter basis, CSP customers drove about 40% of our Data Center revenue.

Adoption of our new flagship H100 data center GPU is strong. In just the second quarter of its ramp, H100 revenue was already much higher than that of A100, which declined sequentially. This is a testament of the exceptional performance on the H100, which is as much as 9x faster than the A100 for training and up 30x faster than [inferencing of] (ph) transformer-based large language models. The transformer engine of H100 arrived just in time to serve the development and scale out of inference of large language models.

AI adoption is at an inflection point. Open AI's ChatGPT has captured interest worldwide, allowing people to experience AI firsthand and showing what's possible with generative AI. These new types of neural network models can improve productivity in a wide range of tasks, whether generating text like marketing copy, summarizing documents like [indiscernible], creating images for ads or video games or answering customer questions. Generative AI applications will help almost every industry do more faster.

Generative large language models with over 100 billion parameters are the most advanced neural networks in today's world. NVIDIA's expertise spans across the AI supercomputers, algorithms, data processing and training methods that can bring these capabilities to enterprise. We look forward to helping customers with generative AI opportunities.

In addition to working with every major hyperscale cloud provider, we are engaged with many consumer Internet companies, enterprises and start-ups. The opportunity is significant and driving strong growth in the data center that will accelerate through the year.

During the quarter, we made notable announcements in the financial services sector, one of our largest industry verticals. We announced a partnership with Deutsche Bank to accelerate the use of AI and machine learning in financial services. Together, we are developing a range of applications, including virtual customer service agents, speech AI, fraud detection and bank process automation, leveraging NVIDIA's full computing stack, both on-premise and in the cloud, including NVIDIA AI enterprise software. We also announced that NVIDIA captured leading results for AI inference in a key financial services industry benchmark for applications such as asset price discovery. In networking, we see growing demand for our latest generation InfiniBand and HPC optimized Ethernet platforms fueled by AI.

Generative AI foundation model sizes continue to grow at exponential rates, driving the need for high-performance networking to scale out multi-node accelerated workloads.

Delivering unmatched performance, latency and in-network computing capabilities, InfiniBand is the clear choice for power-efficient cloud scale, generative AI.

For smaller scale deployments, NVIDIA is bringing its full accelerated stack expertise and integrating it with the world's most advanced high-performance Ethernet fabrics. In the quarter, InfiniBand led our growth as our Quantum 2 40 gigabit per second platform is off to a great start, driven by demand across cloud, enterprise and supercomputing customers. In Ethernet, our 40 gigabit per second Spectrum 4 networking platform is gaining momentum as customers transition to higher speeds, next-generation adapters and switches.

We remain focused on expanding our software and services. We released version 3.0 of NVIDIA AI enterprise with support for more than 50 NVIDIA AI frameworks and pretrained model and new workflows for contact center intelligent virtual assistance, audio transcription and cybersecurity. Upcoming offerings include our NeMo and BioNeMo large language model services, which are currently in early access with customers.

Now to Jensen to talk a bit more about our software and cloud business.

Jensen Huang

Thanks, Colette. The cumulation of technology breakthroughs has brought AI to an inflection point. Generative AI's versatility and capability has triggered a sense of urgency at enterprises around the world to develop and deploy AI strategies. Yet, the AI supercomputer infrastructure, model algorithms, data processing and training techniques remain an insurmountable obstacle for most. Today, I want to share with you the next level of our business model to help put AI within reach of every enterprise customer.

We are partnering with major service -- cloud service providers to offer NVIDIA AI cloud services, offered directly by NVIDIA and through our network of go-to-market partners, and hosted within the world's largest clouds. NVIDIA AI as a service offers enterprises easy access to the world's most advanced AI platform, while remaining close to the storage, networking, security and cloud services offered by the world's most advanced clouds.

Customers can engage NVIDIA AI cloud services at the AI supercomputer, acceleration library software or pretrained AI model layers. NVIDIA DGX is an AI supercomputer, and the blueprint of AI factories being built around the world. AI supercomputers are hard and time-consuming to build. Today, we are announcing the NVIDIA DGX Cloud, the

fastest and easiest way to have your own DGX AI supercomputer, just open your browser. NVIDIA DGX Cloud is already available through Oracle Cloud Infrastructure and Microsoft Azure, Google GCP and others on the way.

At the AI platform software layer, customers can access NVIDIA AI enterprise for training and deploying large language models or other AI workloads. And at the pretrained generative AI model layer, we will be offering NeMo and BioNeMo, customizable AI models, to enterprise customers who want to build proprietary generative AI models and services for their businesses. With our new business model, customers can engage NVIDIA's full scale of AI computing across their private to any public cloud. We will share more details about NVIDIA AI cloud services at our upcoming GTC so be sure to tune in.

Now let me turn it back to Colette on gaming.

Colette Kress

Thanks, Jensen. Gaming revenue of $1.83 billion was up 16% sequentially and down 46% from a year ago. Fiscal year revenue of $9.07 billion is down 27%. Sequential growth was driven by the strong reception of our 40 Series GeForce RTX GPUs based on the Ada Lovelace architecture. The year-on-year decline reflects the impact of channel inventory correction, which is largely behind us. And demand in the seasonally strong fourth quarter was solid in most regions. While China was somewhat impacted by disruptions related to COVID, we are encouraged by the early signs of recovery in that market.

Gamers are responding enthusiastically to the new RTX4090, 4080, 4070 Ti desktop GPUs, with many retail and online outlets quickly selling out of stock. The flagship RTX 4090 has quickly shot up in popularity on Steam to claim the top spot for the AI architecture, reflecting gamers' desire for high-performance graphics.

Earlier this month, the first phase of gaming laptops based on the Ada architecture reached retail shelves, delivering NVIDIA's largest-ever generational leap in performance and power efficiency. For the first time, we are bringing enthusiast-class GPU performance to laptops as slim as 14 inches, a fast-growing segment, previously limited to basic tasks and apps.

In another first, we are bringing the 90 class GPUs, our most performing models, to laptops, thanks to the power efficiency of our fifth-generation Max-Q technology. All in, RTX 40 Series GPUs will power over [170] (ph) gaming and creator laptops, setting up for a great back-to-schools season.

There are now over 400 games and applications supporting NVIDIA's RTX technology for real-time ray tracing and AI-powered graphics. The AI architecture features DLSS 3, our third-generation AI-powered graphics, which massively boosts performance. With the most advanced games, Cyberpunk 2077, recently added DLSS 3 enabling a 3 to 4x boost in frame rate performance at 4K resolution.

Our GeForce NOW cloud gaming service continued to expand in multiple dimensions, users, titles and performance. It now has more than 25 million members in over 100 countries. Last month, it enabled RTX 4080 graphics horsepower in the new high-performance ultimate membership tier. Ultimate members can stream at up to 240 frames per second from a cloud with full ray tracing and DLSS 3.

And just yesterday, we made an important announcement with Microsoft. We agreed to a 10-year partnership to bring to GeForce NOW Microsoft's lineup of Xbox PC games, which includes blockbusters like Minecraft, Halo and Flight Simulator. And upon the close of Microsoft's Activision acquisition, it will add titles like Call of Duty and Overwatch.

Moving to Pro Visualization. Revenue of $226 million was up 13% sequentially and down 65% from a year ago. Fiscal year revenue of $1.54 billion was down 27%. Sequential growth was driven by desktop workstations with strengths in the automotive and manufacturing industrial verticals. Year-on-year decline reflects the impact of the channel inventory correction, which we expect to end in the first half of the year.

Interest in NVIDIA's Omniverse continues to build with almost 300,000 downloads so far, 185 connectors to third-party design applications. The latest released Omniverse has a number of features and enhancements, including support for 4K, real-time path tracing, Omniverse Search for AI-powered search through large untagged 3D databases, and Omniverse cloud containers for AWS.

Let's move to automotive. Revenue was a record $294 million, up 17% from [indiscernible] and up 135% from a year ago. Sequential growth was driven primarily by AI automotive solutions. New program ramps at both electric vehicle and traditional OEM customers helped drive this growth. Fiscal year revenue of $903 million was up 60%.

At CES, we announced a strategic partnership with Foxconn to develop automated and autonomous vehicle platforms. This partnership will provide scale for volume, manufacturing to meet growing demand for the NVIDIA Drive platform. Foxconn will use NVIDIA Drive, Hyperion compute and sensor architecture for its electric vehicles.

Foxconn will be a Tier 1 manufacturer producing electronic control units based on NVIDIA Drive Orin for the global .

We also reached an important milestone this quarter. The NVIDIA Drive operating system received safety certification from TÜV SÜD, one of the most experienced and rigorous assessment bodies in the automotive industry. With industry-leading performance and functional safety, our platform meets the higher standards required for autonomous transportation.

Moving to the rest of the P&L. GAAP gross margin was 63.3%, and non-GAAP gross margin was 66.1%. Fiscal year GAAP gross margin was 56.9%, and non-GAAP gross margin was 59.2%. Year-on-year, Q4 GAAP operating expenses were up 21%, and non-GAAP operating expenses were up 23%, primarily due to the higher compensation and data center infrastructure expenses.

Sequentially, GAAP operating expenses were flat, and non-GAAP operating expenses were down 1%. We plan to keep them relatively flat at this level over the coming quarters. Full year GAAP operating expenses were up 50%, and non-GAAP operating expenses were up 31%.

We returned $1.15 billion to shareholders in the form of share repurchases and cash dividends. At the end of Q4, we had approximately %7 billion remaining under our share repurchase authorization through December 2023.

Let me look to the outlook for the first quarter of fiscal '24. We expect sequential growth to be driven by each of our 4 major market platforms led by strong growth in data center and gaming. Revenue is expected to be $6.5 billion, plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 64.1% and 66.5%, respectively, plus or minus 50 basis points. GAAP operating expenses are expected to be approximately $2.53 billion. Non-GAAP operating expenses are expected to be approximately $1.78 billion. GAAP and non-GAAP other income and expenses are expected to be an income of approximately $50 million, excluding gains and losses of nonaffiliated divestments. GAAP and non-GAAP tax rates are expected to be 13%, plus or minus 1%, excluding any discrete items. Capital expenditures are expected to be approximately $350 million to $400 million for the first quarter and in the range of $1.1 billion to $1.3 billion for the full fiscal year 2024. Further financial details are included in the CFO commentary and other information available on our IR website.

In closing, let me highlight upcoming events for the financial community. We will be attending the Morgan Stanley Technology Conference on March 6 in San Francisco and

the Cowen Healthcare Conference on March 7 in Boston. We will also host GTC virtually with Jensen's keynote kicking off on March 21. Our earnings call to discuss the results of our first quarter of fiscal year '24 is scheduled for Wednesday, May 24.

Now we will open up the call for questions. Operator, would you please poll for questions?

Question-and-Answer Session

Operator

[Operator Instructions]. Your first question comes from the line of Aaron Rakers with Wells Fargo.

Aaron Rakers

Clearly, on this call, a key focal point is going to be the monetization effect of your software and cloud strategy. I think as we look at it, I think, straight up, the enterprise AI software suite, I think, is priced at around $6,000 per CPU socket. I think you've got pricing metrics a little bit higher for the cloud consumption model. I'm just curious, Colette, how do we start to think about that monetization contribution to the company's business model over the next couple of quarters relative to, I think, in the past, you've talked like a couple of hundred million or so? Just curious if you can unpack that a little bit.

Colette Kress

So I'll start and turn it over to Jensen to talk more because I believe this will be a great topic and discussion also at our GTC.

Our plans in terms of software, we continue to see growth even in our Q4 results, we're making quite good progress in both working with our partners, onboarding more partners and increasing our software. You are correct. We've talked about our software revenues being in the hundreds of millions. And we're getting even stronger each day as Q4 was probably a record level in terms of our software levels. But there's more to unpack in terms of there, and I'm going to turn it to Jensen.

Jensen Huang

Yes, first of all, taking a step back, NVIDIA AI is essentially the operating system of AI systems today. It starts from data processing to learning, training, to validations, to inference. And so this body of software is completely accelerated. It runs in every cloud.

It runs on-prem. And it supports every framework, every model that we know of, and it's accelerated everywhere.

By using NVIDIA AI, your entire machine learning operations is more efficient, and it is more cost effective. You save money by using accelerated software. Our announcement today of putting NVIDIA's infrastructure and have it be hosted from within the world's leading cloud service providers accelerates the enterprise's ability to utilize NVIDIA AI enterprise. It accelerates people's adoption of this machine learning pipeline, which is not for the faint of heart. It is a very extensive body of software. It is not deployed in enterprises broadly, but we believe that by hosting everything in the cloud, from the infrastructure through the operating system software, all the way through pretrained models, we can accelerate the adoption of generative AI in enterprises. And so we're excited about this new extended part of our business model. We really believe that it will accelerate the adoption of software.

Operator

Your next question comes from the line of Vivek Arya with Bank of America.

Vivek Arya

Just wanted to clarify, Colette, if you meant data center could grow on a year-on-year basis also in Q1?

And then Jensen, my main question kind of relate to 2 small related ones. The computing intensity for generative AI, if it is very high, does it limit the market size to just a handful of hyperscalers? And on the other extreme, if the market gets very large, then doesn't it attract more competition for NVIDIA from cloud ASICs or other accelerator options that are out there in the market?

Colette Kress

Thanks for the question. First, talking about our data center guidance that we provided for Q1. We do expect a sequential growth in terms of our data center, strong sequential growth. And we are also expecting a growth year-over-year for our data center. We actually expect a great year with our year-over-year growth in data center probably accelerating past Q1.

Jensen Huang

Large language models are called large because they are quite large. However, remember that we've accelerated and advanced AI processing by a million x over the

last decade. Moore's Law, in its best days, would have delivered 100x in a decade. By coming up with new processors, new systems, new interconnects, new frameworks and algorithms and working with data scientists, AI researchers on new models, across that entire span, we've made large language model processing a million times faster, a million times faster.

What would have taken a couple of months in the beginning, now it happens in about 10 days. And of course, you still need a large infrastructure. And even the large infrastructure, we're introducing Hopper, which, with its transformer engine, it's new NVLink switches and its new InfiniBand 400 gigabits per second data rates, we're able to take another leap in the processing of large language models.

And so I think the -- by putting NVIDIA's DGX supercomputers into the cloud with NVIDIA DGX cloud, we're going to democratize the access of this infrastructure, and with accelerated training capabilities, really make this technology and this capability quite accessible. So that's one thought.

The second is the number of large language models or foundation models that have to be developed is quite large. Different countries with different cultures and its body of knowledge are different. Different fields, different domains, whether it's imaging or its biology or its physics, each one of them need their own domain of foundation models. With large language models, of course, we now have a prior that could be used to accelerate the development of all these other fields, which is really quite exciting.

The other thing to remember is that the number of companies in the world have their own proprietary data. The most valuable data in the world are proprietary. And they belong to the company. It's inside their company. It will never leave the company. And that body of data will also be harnessed to train new AI models for the very first time. And so we -- our strategy and our goal is to put the DGX infrastructure in the cloud so that we can make this capability available to every enterprise, every company in the world who would like to create proprietary data and so -- proprietary models.

The second thing about competition. We've had competition for a long time. Our approach, our computing architecture, as you know, is quite different on several dimensions. Number one, it is universal, meaning you could use it for training, you can use it for inference, you can use it for models of all different types. It supports every framework. It supports every cloud. It's everywhere. It's cloud to private cloud, cloud to on-prem. It's all the way out to the edge. It could be an autonomous system. This one architecture allows developers to develop their AI models and deploy it everywhere.

The second very large idea is that no AI in itself is an application. There's a preprocessing part of it and a post-processing part of it to turn it into an application or service. Most people don't talk about the pre and post processing because it's maybe not as sexy and not as interesting. However, it turns out that preprocessing and post-processing oftentimes consumes half or 2/3 of the overall workload. And so by accelerating the entire end-to-end pipeline, from preprocessing, from data ingestion, data processing, all the way to the preprocessing all the way to post processing, we're able to accelerate the entire pipeline versus just accelerating half of the pipeline. The limit to speed up, even if you're instantly passed if you only accelerate half of the workload, is twice as fast. Whereas if you accelerate the entire workload, you could accelerate the workload maybe 10, 20, 50x faster, which is the reason why when you hear about NVIDIA accelerating applications, you routinely hear 10x, 20x, 50x speed up. And the reason for that is because we accelerate things end to end, not just the deep learning part of it, but using CUDA to accelerate everything from end to end.

And so I think the universality of our computing -- accelerated computing platform, the fact that we're in every cloud, the fact that we're from cloud to edge, makes our architecture really quite accessible and very differentiated in this way. And most importantly, to all the service providers, because of the utilization is so high, because you can use it to accelerate the end-to-end workload and get such a good throughput, our architecture is the lowest operating cost. It's not -- the comparison is not even close. So -- anyhow those are the 2 answers.

Operator

Your next question comes from the line of C.J. Muse with Evercore.

Christopher Muse

I guess, Jensen, you talked about ChatGPT as an inflection point kind of like the iPhone. And so curious, part A, how have your conversations evolved post ChatGPT with hyperscale and large-scale enterprises? And then secondly, as you think about Hopper with the transformative engine and Grace with high-bandwidth memory, how have you kind of your outlook for growth for those 2 product cycles evolved in the last few months?

Jensen Huang

ChatGPT is a wonderful piece of work, and the team did a great job, OpenAI did a great job with it. They stuck with it. And the accumulation of all of the breakthroughs led to a

service with a model inside that surprised everybody with its versatility and its capability.

What people were surprised by, and this is in our -- and close within the industry is well understood. But the surprising capability of a single AI model that can perform tasks and skills that it was never trained to do. And for this language model to not just speak English, or can translate, of course, but not just speak human language, it can be prompted in human language, but output Python, output Cobalt, a language that very few people even remember, output Python for Blender, a 3D program. So it's a program that writes a program for another program.

We now realize -- the world now realizes that maybe human language is a perfectly good computer programming language, and that we've democratized computer programming for everyone, almost anyone who could explain in human language a particular task to be performed. This new computer -- when I say new era of computing, this new computing platform, this new computer could take whatever your prompt is, whatever your human-explained request is, and translate it to a sequence of instructions that you process it directly, or it waits for you to decide whether you want to process it or not.

And so this type of computer is utterly revolutionary in its application because it's democratized programming to so many people really has excited enterprises all over the world. Every single CSP, every single Internet service provider, and they're, frankly, every single software company, because of what I just explained, that this is an AI model that can write a program for any program. Because of that reason, everybody who develops software is either alerted or shocked into alert or actively working on something that is like ChatGPT to be integrated into their application or integrated into their service. And so this is, as you can imagine, utterly worldwide.

The activity around the AI infrastructure that we build Hopper and the activity around inferencing using Hopper and Ampere to inference large language models, has just gone through the roof in the last 60 days. And so there's no question that whatever our views are of this year as we enter the year has been fairly, dramatically changed as a result of the last 60, 90 days.

Operator

Your next question comes from the line of Matt Ramsay with Cowen & Company.

Matthew Ramsay

Jensen, I wanted to ask a couple of questions on the DGX Cloud. And I guess, we're all talking about the drivers of the services and the compute that you're going to host on top of these services with the different hyperscalers. But I think we've been kind of watching and wondering when your data center business might transition to more of a systems level business, meaning pairing and [indiscernible] InfiniBand with your Hopper product, with your Grace product and selling things more on a systems level. I wonder if you could step back, over the next 2 or 3 years, how do you think the mix of business in your data center segment evolves from maybe selling cards to systems and software? And what can that mean for the margins of that business over time?

Jensen Huang

Yes, I appreciate the question. First of all, as you know, our Data Center business is a GPU business only in the context of a conceptual GPU because what we actually sell to the cloud service providers is a panel, a fairly large computing panel of 8 Hoppers or 8 Amperes that's connected with NVLink switches that are connected with NVLink. And so this board represents essentially 1 GPU. It's 8 chips connected together into 1 GPU with a very high-speed chip-to-chip interconnect. And so we've been working on, if you will, multi-die computers for quite some time. And that is 1 GPU.

So when we think about a GPU, we actually think about an HGX GPU, and that's 8 GPUs. We're going to continue to do that. And the thing that the cloud service providers are really excited about is by hosting our infrastructure for NVIDIA to offer because we have so many companies that we work directly with. We're working directly with 10,000 AI start-ups around the world, with enterprises in every industry. And all of those relationships today would really love to be able to deploy both into the cloud at least or into the cloud and on-prem and oftentimes multi-cloud.

And so by having NVIDIA DGX and NVIDIA's infrastructure are full stack in their cloud, we're effectively attracting customers to the CSPs. This is a very, very exciting model for them. And they welcomed us with open arms. And we're going to be the best AI salespeople for the world's clouds. And for the customers, they now have an instantaneous infrastructure that is the most advanced. They have a team of people who are extremely good from the infrastructure to the acceleration software, the NVIDIA AI open operating system, all the way up to AI models. Within 1 entity, they have access to expertise across that entire span. And so this is a great model for customers. It's a great model for CSPs. And it's a great model for us. It lets us really run like the wind. As much as we will continue and continue to advance DGX AI supercomputers, it does take time to build AI supercomputers on-prem. It's hard no matter how you look at it. It takes

time no matter how you look at it. And so now we have the ability to really prefetch a lot of that and get customers up and running as fast as possible.

Operator

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

Timothy Arcuri

Jensen, I had a question about what this all does to your TAM. Most of the focus right now is on text, but obviously, there are companies doing a lot of training on video and music. They're working on models there. And it seems like somebody who's training these big models has maybe, on the high end, at least 10,000 GPUs in the cloud that they've contracted and maybe tens of thousands of more to inference a widely deployed model. So it seems like the incremental TAM is easily in the several hundred thousands of GPUs and easily in the tens of billions of dollars. But I'm kind of wondering what this does to the TAM numbers you gave last year. I think you said $300 billion hardware TAM and $300 billion software TAM. So how do you kind of think about what the new TAM would be?

Jensen Huang

I think those numbers are really good anchor still. The difference is because of the, if you will, incredible capabilities and versatility of generative AI and all of the converging breakthroughs that happened towards the middle and the end of last year, we're probably going to arrive at that TAM sooner than later. There's no question that this is a very big moment for the computer industry. Every single platform change, every inflection point in the way that people develop computers happened because it was easier to use, easier to program and more accessible. This happened with the PC revolution. This happened with the Internet revolution. This happened with mobile cloud. Remember, mobile cloud, because of the iPhone and the App Store, 5 million applications and counting emerged. There weren't 5 million mainframe applications. There weren't 5 million workstation applications. There weren't 5 million PC applications. And because it was so easy to develop and deploy amazing applications part cloud, part on the mobile device and so easy to distribute because of app stores, the same exact thing is now happening to AI.

In no computing era did 1 computing platform, ChatGPT, reached 150 million people in 60, 90 days. I mean, this is quite an extraordinary thing. And people are using it to create all kinds of things. And so I think that what you're seeing now is just a torrent of new companies and new applications that are emerging. There's no question this is, in every

way, a new computing era. And so I think this -- the TAM that we explained and expressed, it really is even more realizable today and sooner than before.

Operator

Your next question comes from the line of Stacy Rasgon with Bernstein.

Stacy Rasgon

I have a clarification and then a question both for Colette. The clarification, you said H-100 revenue's higher than A100. Was that an overall statement? Or was that at the same point in time like after 2 quarters of shipments?

And then for my actual question. I wanted to ask about auto, specifically the Mercedes opportunity. The Mercedes had an event today, and they were talking about software revenues for their MB Drive that could be single digit or low billion euros by mid-decade and mid billion euros by the end of the decade. And I know you guys were supposedly splitting the software revenues 50-50. Is that kind of the order of magnitude of software revenues from the Mercedes deal that you guys are thinking of and over that similar time frame? Is that how we should be modeling that?

Colette Kress

Great. Thanks, Stacy, for the question. Let me first start with your question you had about H-100 and A100. We began initial shipments of H-100 back in Q3. It was a great start. Many of them began that process many quarters ago. And this was a time for us to get production level to them in Q3. So Q4 was an important time for us to see a great ramp of H-100 that we saw. What that means is our H-100 was the focus of many of our CSPs within Q4, and they were all wanting to get both get up and running in cloud instances. And so we actually saw less of A100 in Q4 of what we saw in H-100 at a larger amount. We tend to continue to sell both architectures going forward, but just in Q4, it was a strong quarter for

Your additional questions that you had on Mercedes Benz. I'm very pleased with the joint connection that we have with them and the work. We've been working very diligently about getting ready to come to market. And you're right. They did talk about the software opportunity. They talked about their software opportunity in 2 phases, about what they can do with Drive as well as what they can also do with Connect. They extended out to a position of probably about 10 years looking at the opportunity that they see in front of us. So it aligns with what our thoughts are with a long-term partner of that and sharing that revenue over time.

Jensen Huang

One of the things that, if I could add, Stacy, to say something about the wisdom of what Mercedes is doing. This is the only large luxury brand that has, across the board, from every -- from the entry all the way to the highest end of their luxury cars, to install every single one of them with a rich sensor set, every single one of them with an AI supercomputer, so that every future car in the Mercedes fleet will contribute to an installed base that could be upgradable and forever renewed for customers going forward. If you could just imagine what it looks like if the entire Mercedes fleet that is on the road today were completely programmable, that you can OTA, it would represent tens of millions of Mercedeses that would represent revenue-generating opportunity. And that's the vision that Ola has. And what they're building for, I think, it's going to be extraordinary. The large installed base of luxury cars that will continue to renew with -- for customers' benefits and also for revenue-generating benefits.

Operator

Your next question comes from the line of Mark Lipacis with Jefferies.

Mark Lipacis

I think for you, Jensen, it seems like every year a new workload comes out and drives demand for your process or your ecosystem cycles. And if I think back facial recognition and then recommendation engines, natural language processing, Omniverse and now generative AI engines, can you share with us your view? Is this what we should expect going forward, like a brand-new workload that drives demand to the next level for your products?

And the reason I ask is because I found it interesting your comments in your script where you mentioned that your kind of view about the demand that generative AI is going to drive for your products and now services is -- seems to be a lot, better than what you thought just over the last 90 days. So -- and to the extent that there's new workloads that you're working on or new applications that can drive next levels of demand, would you care to share with us a little bit of what you think could drive it past what you're seeing today?

Jensen Huang

Yes, Mark, I really appreciate the question. First of all, I have new applications that you don't know about and new workloads that we've never shared that I would like to share

with you at GTC. And so that's my hook to come to GTC, and I think you're going to be very surprised and quite delighted by the applications that we're going to talk about.

Now there's a reason why it is the case that you're constantly hearing about new applications. The reason for that is, number one, NVIDIA is a multi-domain accelerated computing platform. It is not completely general purpose like a CPU because a CPU is 95%, 98% control functions and only 2% mathematics, which makes it completely flexible. We're not that way. We're an accelerated computing platform that works with the CPU that offloads the really heavy computing units, things that could be highly, highly paralyzed to offload them. But we're multi-domain. We could do particle systems. We could do fluids. We could do neurons. And we can do computer graphics. We can do . There are all kinds of different applications that we can accelerate, number one.

Number two, our installed base is so large. This is the only accelerated computing platform, the only platform. Literally, the only one that is architecturally compatible across every single cloud from PCs to workstations, gamers to cars to on-prem. Every single computer is architecturally compatible, which means that a developer who developed something special would seek out our platform because they like the reach. They like the universal reach. They like the acceleration, number one. They like the ecosystem of programming tools and the ease of using it and the fact that they have so many people they can reach out to, to help them. There are millions of CUDA experts around the world, software all accelerated, tool all accelerated. And then very importantly, they like the reach. They like the fact that you can see -- they can reach so many users after they develop the software. And it is the reason why we just keep attracting new applications.

And then finally, this is a very important point. Remember, the rate of CPU computing advance has slowed tremendously. And whereas back in the first 30 years of my career, at 10x in performance at about the same power every 5 years and then 10x every 5 years. That rate of continued advance has slowed. At a time when people still have really, really urging applications that they would like to bring to the world, and they can't afford to do that with the power keep going up. Everybody needs to be sustainable. You can't continue to consume power. By accelerating it, we can decrease the amount of power you use for any workload. And so all of these multitude of reasons is really driving people to use accelerated computing, and we keep discovering new exciting applications.

Operator

Your next question comes from the line of Atif Malik with Citi.

Atif Malik

Colette, I have a question on data center. You saw some weakness on build plan in the January quarter, but you're guiding to year-over-year acceleration in April and through the year. So if you can just rank order for us the confidence in the acceleration. Is that based on your H-100 ramp or generative AI sales coming through or the new AI services model? And also, if you can talk about what you're seeing on the enterprise vertical.

Colette Kress

Sure. Thanks for the question. When we think about our growth, yes, we're going to grow sequentially in Q1 and do expect year-over-year growth in Q1 as well. It will likely accelerate there going forward. So what do we see as the drivers of that? Yes, we have multiple product cycles coming to market. We have H-100 in market now. We are continuing with our new launches as well that are sometimes fueled with our GPU computing with our networking. And then we have grades coming likely in the second half of the year. Additionally, generative AI, it's sparked interest definitely among our customers, whether those be CSPs, whether those be enterprises, one of those be start-ups. We expect that to be a part of our revenue growth this year. And then lastly, let's just not forget that given the end of Moore's Law, there's an error here of focusing on AI, focusing on accelerated continuing. So as the economy improves, this is probably very important to the enterprises and it can be fueled by the existence of cloud first for the enterprises as they [indiscernible]. I'm going to turn it to Jensen to see if has any additional things he'd like to add.

Jensen Huang

No, you did great. That was great.

Operator

Your last question today comes from the line of Joseph Moore with Morgan Stanley.

Joseph Moore

Jensen, you talked about the sort of 1 million times improvement in your ability to train these models over the last decade. Can you give us some insight into what that looks like in the next few years and to the extent that some of your customers with these large language models are talking about 100x the complexity over that kind of time frame. I know Hopper is 6x better transformer performance. But what can you do to scale that

up? And how much of that just reflects that it's going to be a much larger hardware expense down the road?

Jensen Huang

First, I'll start backwards. I believe the number of AI infrastructures are going to grow all over the world. And the reason for that is AI, the production of intelligence, is going to be manufacturing. There was a time when people manufacture just physical goods. In the future, there will be -- almost every company will manufacture soft goods. It just happens to be in the form of intelligence. Data comes in. That data center does exactly 1 thing and 1 thing only. It cranks on that data and it produces a new updated model. Where raw material comes in, a building or an infrastructure cranks on it, and something refined or improved comes out that is of great value, that's called the factory. And so I expect to see AI factories all over the world. Some of it will be hosted in cloud. Some of it will be on-prem. There will be some that are large, and there are some that will be mega large, and then there'll be some that are smaller. And so I fully expect that to happen, number one.

Number two. Over the course of the next 10 years, I hope through new chips, new interconnects, new systems, new operating systems, new distributed computing algorithms and new AI algorithms and working with developers coming up with new models, I believe we're going to accelerate AI by another million x. There's a lot of ways for us to do that. And that's one of the reasons why NVIDIA is not just a chip company because the problem we're trying to solve is just too complex. You have to think across the entire stack all the way from the chip, all the way into the data center across the network through the software. And in the mind of 1 single company, we can think across that entire stack. And it's really quite a great playground for computer scientists for that reason because we can innovate across that entire stack. So my expectation is that you're going to see really gigantic breakthroughs in AI models in the next company, the AI platforms in the coming decade. But simultaneously, because of the incredible growth and adoption of this, you're going to see these AI factories everywhere.

Operator

This concludes our Q&A session. I will now turn the call back over to Jensen Huang for closing remarks.

Jensen Huang

Thank you. The accumulation of breakthroughs from transformers, large language model and generative AI has elevated the capability and versatility of AI to a remarkable

level. A new computing platform has emerged. New companies, new applications and new solutions to long-standing challenges are being invented at an astounding rate. Enterprises in just about every industry are activating to apply generative AI to reimagine their products and businesses. The level of activity around AI, which was already high, has accelerated significantly. This is the moment we've been working towards for over a decade. And we are ready. Our Hopper AI supercomputer with the new transformer engine and Quantum InfiniBand fabric is in full production, and CSPs are racing to open their Hopper cloud services. As we work to meet the strong demand for our GPUs, we look forward to accelerating growth through the year.

Don't miss the upcoming GTC. We have much to tell you about new chips, systems and software, new CUDA applications and customers, new ecosystem partners and a lot more on NVIDIA AI and Omniverse. This will be our best GTC yet. See you there.

Operator

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