Arete Technology Conference

Company Participants

- Brett Simpson, Analyst
- Ian Buck, Vice President, General Manager
- Unidentified Participant, Analyst

Presentation

Brett Simpson {BIO 3279126 <GO>}

Okay. Thanks very much. And hi, everyone.

It's Brett Simpson at Arete. And it's my pleasure to welcome Ian Buck, who many of you know was the inventor of CUDA, NVIDIA's CUDA software and has headed up the Data Center Division for many years now -- I think longer than I've covered NVIDIA actually.

So Ian, excited to have you here today. And thanks for joining us.

Ian Buck {BIO 18454865 <GO>}

Thank you for having me. Happy to help.

Brett Simpson {BIO 3279126 <GO>}

I think it's a particularly interesting time to connect with Ian. And we're all keen to get Ian's perspective on how the AI market is going to trend over the next couple of years. It's obviously been a phenomenal year.

And I think if we go back last year when we spoke Ian, I think you were annualizing $16 billion of revenue in Data Center. And I think you're pretty much there in terms of quarterly sales. So phenomenal achievement. Great to see.

Before we start, I'm just going to hand over to Ian to maybe talk a little bit about the disclosures and forward-looking statements.

So Ian, over to you and then we can get started.

Ian Buck {BIO 18454865 <GO>}

Yes. Just really quick. I mean as a reminder, this presentation, what we'll talk about, that may contain forward statements.

Investors are advised to always read the reports filed with the SEC for information related to the risks and uncertainties facing our business.

Brett Simpson {BIO 3279126 <GO>}

Great. And just before we start, this is a 1-hour call. We're going to cover off our prepared questions for about 45 minutes. And then we're going to open up to investor Q&A. (Operator Instructions)

Questions And Answers

A - Brett Simpson {BIO 3279126 <GO>}

So maybe just jumping right in, Ian, to start with, can we maybe just have your thoughts on just reviewing the year that's gone? And obviously it's been a phenomenal year commercially, but what's the most for you looking through 2023? And what are you setting up as the big priorities for the next year or two in the Data Center division?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I mean obviously this was the -- while the official ChatGPT moment actually happened in '22, at the end of '22, really, its stride and its impact was felt in 2023. And that really was -- created the -- showed the world the opportunity that AI -- generative AI could do for how we interact with computers, how we interact with the cloud, how we, like, use software. I don't think there's ever been, at least in recent history, a technology introduction that -- unlike maybe PC to phone where you change the modality, you create a whole new ecosystem, a new opportunity, a new way of working. That's clearly happening with gen AI.

New services, new kinds of experiences, new consumer products, new business opportunities, but you're also making the old stuff way more interesting. Word just got sexy again. I always liked PowerPoint, but like we were actually like -- we're working in from both angles. So seeing how that commercialization of AI has happened to really make it a useful tool for literally everyone has given -- that killer app has -- it's been working on AI for, gosh, almost 10 years now since the first time NVIDIA GPUs were used. I mean it found us, we didn't find it.

Back in the AlexNet and ImageNet days, has really ramped. This year, on -- there's been a couple of great advances as a result. NVIDIA now really, we've talked about being a data center-scale company where data center is being the new unit of compute. And these language models are -- and generative AI is a data center-scale problem. It really has brought all of our different technologies together, whether it be GPUs, DPUs, CPUs, to provide that infrastructure and in addition, all of the software that goes on to that.

And NVIDIA as a company has more software engineers and hardware engineers because of the opportunity and capability and what we deliver. So certainly, LLM, generative AI ramps. From NVIDIA product perspective, I've been extremely thrilled

about the progress that we've made on Grace and Grace Hopper. You've seen the -- we've brought it to market now. You've seen the opportunity that it can provide as a single-node inferencing platform.

You've seen what it can do for our supercomputing and HPC business, which is often a leading indicator where we're going. We were obviously doing GPUs at scale for myself 20 years now. And we're now next year slated to deploy on the order of 200 exaflops of AI across supercomputing, all Grace Hopper. And then finally, of course what Grace Hopper and multi-node is going to do for AI this year, we saw first glimpse of that with the announcement we just made at re:Invent, taking Grace Hopper and putting it together in a 32-way NVLink (inaudible) rack and build out an AI supercomputer with it, AWS at 16,000 GPUs. It's a double benefit because they are also now DGX Cloud partners.

So now we have four DGX Cloud partners, Azure, OCI, GCP and now AWS, all working together. So as you can imagine, life is pretty busy on our front. And I think we'll just naturally carry over to accelerate to next year. We've got the enterprises adopting AI, building services or capabilities and partnering with NVIDIA to ask questions about their data, whether it be like ServiceNow or Dropbox or -- and that's happening through our DGX Cloud engagements. It's happening through the AI factories that we're standing up and all the software, either it be a framework or a service, that we're putting together in partnership with those enterprises.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Well, let me just take some of those points that you laid out, Ian. Maybe first on the AI factories strategy. Can you maybe just lay out -- I mean obviously we have public cloud who's -- mainly have conventional compute installed base, but building out significant GPUs now and you have 202 [ph] cloud. But explain to us what is an AI factory?

Whether it's in the scale of operations, the amount of power that it needs. What is -- what are we looking at here in terms of deployment? And how many AI factories do we need over the next couple of years? Just give us a sense as to how you're thinking about this. Yes.

A - Ian Buck {BIO 18454865 <GO>}

So it's a pivot and a change perhaps from the cloud of the past and even just GPU clouds in the past. You would rent $1 per GPU or a single GPU or maybe you can get a node, which is a eight GPU HGX platform, all fantastic, and it's how a lot of things are put together still today. But when we think about an AI factory, these model -- gen AI needs to work at scale. And the scales vary. I think it's within a rack or within a row or across the data center because AI is, at scale, training for sure, and in some cases, inference for large models are happening at scale.

They are a different kind of compute where you have the -- and as a result, the engagement with those enterprise is a little bit different. Instead of renting infrastructure, you're training a foundational model for your business. It starts with

the data. And one thing you may have noticed the leaders is the number of tokens to train has gone up quite a bit. I need to make that graph.

But it is -- we used to -- everyone feels satisfied maybe with a few hundred million to 1 billion, now one trillion token.

While the models may be growing, the number of tokens to get to the quality is increasing for foundation models, and that's also driving the sense of scale. So if you think about what the clouds and the unicorns and companies are building or using, they're using it not as an Infrastructure as a Service, but as an AI factory, a place where they can have access to their data, be able to train at various scales and be able to process and refine and transform the various amounts of data and turn them into a AI model and then quickly, of course monetize it, being able to take that model into a product to provide the insight, which creates the business opportunity for them, provide the additional service or capability. It turns that data into compute, into an AI asset that becomes a business for them in various ways they (inaudible) the product. It really wants one workload, right?

It's the AI model we're going to be training our inference. And as you know, the computing of training inference is very linked. In fact, to train a model, you first do inference on the model and teach it why it got it wrong. It's just the forward past version of it. So that is the -- they're building data centers, AI factories for that one use case, running and operating AI models at scale.

So our product and our engagement and how we work with the ecosystem there is, of course through the hardware platforms.

Today HGX H100. Now it's HGX H200, which hopefully you've have seen. We now have increased the memory sizes to 144 gigabytes per GPU and delivering five terabytes a second. The number again was just double the kind of the performance of AI on the existing Hopper.

And then going to Grace Hopper, which takes you up another click in terms of power optimization and, of course NVLink at scale, which is like the AWS work. That's on top of all the work we're doing in networking in Ethernet and InfiniBand, both of which we enable to build AI factories. I think it's early. You're seeing the first AI factories happen in the four leading customers. Our announcement we did with Azure around MLPerf is a great example.

It was -- we trained GPTs on the MLPerf benchmark. Training a GPT 175B across 10,000 GPUs in a full InfiniBand cluster gives you a sense of what -- the scale of an AI factory (inaudible). And we have a copy of that with our own Eos supercomputer that we use for ourselves and for our own research in AI factory. And our DGX Cloud business is an AI factory business, make that capability available to the Fortune 500s to be able to have access to an AI factory, while the rest of the world continues to build up this capability in region or in their clouds.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. And is it the thinking that fast forward 2, 3, four years, most of the Fortune 500 is going to have their own AI factory? And I mean we saw you did a partnership with Foxconn. It sounds like you're pretty close to deploying something on AI factory into early next year.

A - Ian Buck {BIO 18454865 <GO>}

Yes. I mean the -- to build an AI factory, we have to -- NVIDIA has to work with the entire ecosystem from taking -- from building out the servers and just the systems and baseboards. The technology will push the limits in things like NVLink in water cooling, in interconnect and then work with the data center ecosystem, which includes the Taiwan and all the system -- those in Taiwan, not just the systems, but the rack scale.

And then, of course the data centers themselves. Being able to have their -- helping guide them on what the -- what can be fit in today's data centers or retrofit in today's data centers, to provide that capability and value, and the forward-looking data center build-out plan to provide for things like water cooling, facility water and give a sense of the direction because data center planning, of course is a major capital expense and takes -- is a multiyear road map.

It's just big construction projects. So we have to -- what you see in the end is a product or a partnership or a customer using an AI factory, which you've seen probably less of, but get glimpse of is NVIDIA working with that end-to-end ecosystem to help accelerate the ability for everyone to stand up AI factories and having the ecosystem, right, or support the scale that we're talking about and the kind of data center, the kind of AI factory that people want to build.

A - Brett Simpson {BIO 3279126 <GO>}

Okay. And maybe just talking about DGX Cloud. I mean we talked briefly about -- I mean you mentioned the AWS agreement, which is great. But how many actual customers are using DGX Cloud today? I think the intention was you were going to bring your customers on to the platform.

Can you maybe just give us a snapshot in terms of the activity around DGX Cloud? And I think also, we talked about $1 billion of annualized software revenue in your division these days. How much of that is coming from DGX Cloud?

A - Ian Buck {BIO 18454865 <GO>}

Yes. So I'm actually not sure if we've provided up to date or what information we provided in terms of number of customers and others. I think Simona and the team can provide you more of what's been publicly disclosed in earnings and elsewhere. I'd guide there. I focus on many of -- and see many of the lighthouse customers.

And the customers are, in some case, obvious in the sense that they have data that they want to turn and capitalize on and train on to build the foundational model for their service and capability. So ServiceNow, IT data. What better minefield than everybody -- all the support and help tickets and back and forth engagements to provide a better experience for IT support and even potentially things like preventive maintenance and guide companies as a service to where they need help in IT. Dropbox is a data company who doesn't provide the ability to do train and learn on the data, to provide the services to be able to ask questions of your data. There's work being done with Genentech and the engine.

And these are now vertically applied use cases for drug discovery using our platform called BioNeMo to help them mine the vast majority of drug data and discovery data to accelerate the time to drug discovery, which anything -- the costs to develop the drug is measured in trillions. Anything that can be applied to an AI certainty to pull in that schedule not only reduces the cost of that development, but can dramatically increase time of revenue for a breakthrough drug.

So obviously these customers want to work with NVIDIA. They want to make sure they get the latest. They also inform us and help us drive the -- or accelerate our technology platforms. Obviously they inform our infrastructure, they inform our hardware side, but more importantly, it's the software engagements that we're doing. What we provide in DGX Cloud is an AI factory.

It's not just the infrastructure, but also the software. These are all the NeMo and services that have been talked about. NeMo framework itself is an open-source framework like PyTorch or (inaudible) for training large models and doing foundation monitoring at scale. There's many use cases of it across the industry. But NeMo, the rest of the services provide a way for enterprises to directly work with NVIDIA and work with our software to train those models in a service API format, so having -- be an AI ninja themselves.

Then of course they can take it out from DGX Cloud, they can containerize it and then the opportunity is then to go deploy it wherever they so choose. Often, it's in the same cloud that we're hosting DGX Cloud, which is why we have so many great partners or move it on-prem. It's entirely up to them. And then our monetization comes through the engagement through DGX Cloud, through those services and also through the NVIDIA AI Enterprise software support and run times that are attached to each of those containers as they deploy them. So they continuously maintain that engagement with the best of NVIDIA and get the support and rely on NVIDIA to basically be providing the support and software they need in order to continue to run their business and integrate and scale.

A - Brett Simpson {BIO 3279126 <GO>}

That's driving the software revenue for you guys -- services revenue, yes?

A - Ian Buck {BIO 18454865 <GO>}

Very much.

A - Brett Simpson {BIO 3279126 <GO>}

Okay. Can we talk a little bit about demand, Ian, because I guess we've seen obviously great trends in the last 12 months or so. But looking beyond, say 2024, how do you look at the demand for AI, whether it's at corporate level or government spending or even some of these co-pilots that a lot of the enterprise software companies are building out? How do you about demand and the sustainability of what we're seeing today the inflection point that we're seeing today?

A - Ian Buck {BIO 18454865 <GO>}

Yes, certainly. Obviously demand for gen AI is insatiable, and it creates a -- makes my job very entertaining to help -- and forecasting it makes a very exciting challenge. The aware -- how is it going to -- how is it growing? Certainly, the major cloud providers have the capability and the big muscles to scale. Where they would use to scale perhaps in standing up map-produced clusters and web server [ph] and back-end operations, they know how to operate at hyperscale, that hasn't changed.

The kinds of servers and infrastructure they're standing up is different. It used to be thin, very lightweight, very low-cost, I/O, a little bit of CPU, okay slow storage and simple Ethernet NIC to now standing up an AI supercomputer. But the skills and muscles they have to scale are obviously leveraged. And of course their ability to invest is also -- you can see that. They also have the ability to build data centers and build out data centers.

And so that is obviously also accelerating. So we're going to continue to see the majors obviously building out their AI factories, their AI services and pick up along with NVIDIA's road map. The enterprises then have their choice to consume it. So they can consume it directly from those clouds.

We still see on-prem as well and their ability to work with the major clouds, do on-prem with their own data center or work with these regional cloud providers, these GPU specialty cloud providers, which can operate not potentially at the hyperscales, but smaller, bare metal and sometimes can be a little quicker in the sense that they can in their -- how quickly they can bring to market a new technology because they can execute directly, one-on-one with a customer versus building out a cloud infrastructure.

That has definitely emerged as a go-to-market, which is great. That's just another way people can consume and get access to the latest NVIDIA technology. And we see customers doing all three of those. The -- you mentioned sovereign AI. That is a new thrust.

We've been talking about AI nations and sovereign AI a little bit, and you definitely have seen that tick up in 2023 and will continue to tick up moving forward. We recently announced a supercomputer in the U.K., Isambard-AI. It's being built with Grace Hopper. It is -- it was announced actually by the Prime Minister as an opportunity to build -- so that U.K. could have a resource for AI for the -- for its

nation, its companies, its industries and put itself right in the forefront of being an AI nation.

We obviously see it happening here in the U.S. And we're seeing it happen across Europe and other parts of the world as well. So that will continue to grow, and it dovetails very nicely into our supercomputing business, which has been building AI supercomputers for a while now because we actually build the same -- that same GPU that's being used to train these giant models, the same GPU, same capability that is going into supercomputing, which I have the pleasure of also help the camp [ph]. So it's a -- that is, I think a new growth angle, a new opportunity. And I think we'll be hearing more about sovereign AI projects around the world.

A - Brett Simpson {BIO 3279126 <GO>}

So I guess when you sit back and you look at the sort of early stage of gen AI in the G2000 market, you look at this opportunity for sovereign clouds to get built out. Looking beyond sort of like the next sort of six to 12 months, you must be pretty confident that we're in a pretty good growth trajectory over the sort of medium to long term just given how nascent a lot of these trends are for NVIDIA?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I think the challenge -- and I mentioned this, I think on these calls before. Historically, AI has been -- it's an entirely new computing capability. It required a different kind of software stack and a different kind of -- and had to be invented as a new kind of computer science. No surprise, it started at places like Google and Facebook back then first.

And for many years, we were sharing and educating that this is happening. And the next click was the -- when can it escape a search engine or a news feed and become a tool for the enterprise. And that's happened. I mean now you can see that. And so we're thrilled that -- and that's happened just not -- the technology has matured.

The world is -- the world's smartest people are helping mature it and also build out its capability and making it more adaptable to many different modalities. Then as a result, the opportunity for AI has expanded since the people know now how to tailor it, apply it to all of these different use cases in a way that enterprises can adopt it. A technology like RAG, using data along with the query to further tune and improve the experience. To be able to ask questions of your proprietary data, not just the data that the model was trained on, as a technology, was one of the -- that also has been talked about, but really happened this year in 2023. And that really cracks open the door for AI to allow enterprises to ask questions of their proprietary data.

The vast majority of data is proprietary. These models are trained on things -- the public Internet or what can be acquired in the market. But really, what the value is moving forward is the ability to connect with the proprietary data of the (inaudible) enterprises. And that's like 90% of the data. So we're at the very beginning because what you have seen has been what the -- internally, those hyperscalers use for their own services and the data they have, which obviously they benefited greatly.

Externally, what you've seen is what's been trained on the public Internet, which is -- is made of something that the rest of the entire world can now experiencing and become part of the zeitgeist of the vernacular of everyone, of every consumer [ph]. Moving forward, we're seeing the opportunity to the broadening of the AI market to be every business and every [ph] vertical to apply it with the maturity of AI in terms of how it can be applied, technology and software stack, including our own NeMo software, and also the ability to connect it with the proprietary data that your customer or yourself have, in the case of Dropbox.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Maybe switching gears a little bit. I wanted to get your perspective on inference, the opportunity you see ahead in inference. And I guess we haven't seen that much deployment yet. But I guess we're also looking -- we're hearing from a lot of folks in the industry that they want to see big efficiencies and savings.

And I guess a lot of companies are saying, maybe the GPU isn't the best architecture for inferencing LLMs and quite expensive. So I wanted to get your perspective on that. And do you think NVIDIA needs to look at a new class of accelerator where you strip down some of the double precision -- stuff that maybe isn't so relevant in inference to try to drive down the cost? And if not, how do you drive the cost per query down materially and sustain your leadership and cost of ownership advantages versus your peers?

A - Ian Buck {BIO 18454865 <GO>}

Yes. So first, let me address what we see in market. Certainly, I used to have to explain that you needed a GPU to -- that the CPU's inferencing was just limiting. And today you really can't deploy these models without some form of acceleration, without using GPUs. And we see that, too.

We used to have -- there was a clear -- probably a product segmentation between a lightweight T4 GPU and your A100 training GPU. The reality today is that there's a benefit, it's now blended entirely over. So it's going to be hard, unfortunately, to tease that apart. Just because the value of these models are so great, they need to be and they are large -- either large language model. They want a 100-class GPU, whether it be 1, 2, four or eight or 9, even multiple within a rack.

There's interest in doing that those Grace Hopper on a single Grace Hopper or a 32-way (inaudible) Grace Hopper based on the model. The -- let me talk a little bit, and so we see that. Let me talk about the second point, which is about how does NVIDIA reduce the cost in inferencing? We reduce the cost of inferencing by applying all of the engineering smarts of our software teams to study relentlessly what is the way we can improve the throughput of a GPU for inferencing. Whether it be a small GPU or big GPU, what can we do to improve the throughput, looking at the -- running the model?

And so we have a massive effort, and it's codified in the software called TensorRT. TensorRT is our run time for optimizing all things inference. And we recently this year actually announced TensorRT-LLM, which is an open source, so you don't get

encouraged. We push all of our innovations into TensorRT-LLM to make it the best possible software platform for running inference. When you do inference, you can do things even more extreme than you could do in training because you just need to make -- training needs to maintain a certain level of computational, numerical precision so that you can adjust the model and work with the derivatives.

AI is in turn is all about the derivatives, which is still the slopes and differences. And inference is just forecast. So for instance, we just announced on -- published a new performance numbers, new blogs on Monday this past Monday on H200. We ran the Falcon, the Llama2 models with Infor [ph] precision. So before you ran it with FP eight bits of precision, eight -- 0s and 1s in eight bits.

We did the hard work to figure out how to make only four bits of precision work with Infor. It's easy to say that Infor, I like to do it, but to do it with the accuracy of 95% to 99%, maintaining the accuracy of Llama2, that's -- and the other one was the Falcon-180B, it's 180 billion parameters on a single GPU. It's only possible if you can run the whole (inaudible) model into four bits of precision. In fact, I believe that is the world's largest model ever run on a single GPU period on a 180 billion parameters. So we just took the cost of inferencing and cut in half.

So it is difficult I know that to comprehend this is a fast-moving field. We're engaged with all of the entire ecosystem. We even have our own NVIDIA researchers, that technique of Infor used to -- was called -- used the metric called AWQ, which is actually invented by NVIDIA researchers. We put it all into (inaudible) open-sources NVIDIA (inaudible) orbit.

So I encourage everyone to be watching very closely our performance blogs, our TensorRT, our performance blogs on transformer engine and on the framework that's where we dumb all of this stuff constantly increasing the trip performance even after you buy your GPU. And we'll continue to make the GPU and throughputs and reduce the cost throughout. The last thing I'll say about the -- you asked about different kinds of GPUs. We do that today. So while everyone talks about A100 or H200, and they show the big iron infrastructure that we have.

We also sell a lot of the traditional PCIe form factor GPU very similar to what you may see in a gaming PC, but it has no fan, but it fits nicely in any server, and we have a 2-slot version and a single-slot version, even a half-height, half-length L4. In fact, both Amazon AWS and Google have announced bringing the L4 to market and the L40. These are two of these PCI GPUs. They are at a different price point. They're great universal GPUs and excellent value for deploying and running inference.

Effectively, with minimal changes in the economics, take any server and turn it into an AI server, just by adding an L4 to L4s and some people put an 8. But it's -- we have all of that available to us. And in fact, our gambit of GPUs and form factors continues to expand.

A - Brett Simpson {BIO 3279126 <GO>}

And how would Microsoft be doing copilot inference today?

A - Ian Buck {BIO 18454865 <GO>}

You should ask them that question about how they're doing the copilot. I worked very closely with that team. As you can see in our announcements together and the work that we're doing, you can see how closely NVIDIA and Microsoft and OpenAI, all work together.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Great. Maybe one question before we open up for Q&A. I just wanted to get your perspective on the competitive dynamic that you're seeing in the market? I mean I guess it's early innings for AMD, but they obviously came out last night with their MI300.

And hyperscalers have announced a few AI ASICs in the last month or so. How do you view the competition over the next year or 2, just factoring in? This is a much bigger TAM than it was 12 months ago and there is more competitive offerings coming to market?

A - Ian Buck {BIO 18454865 <GO>}

Yes. Yes. No. I mean - I've been in NVIDIA for 20 years, and we always have been -- you can see how important accelerated computing has become. And every computing company is an AI company now, and they can figure out their path, their way of contributing, and their way of maybe exploring alternatives or options for themselves, including building on silicon and certainly, the clouds have built here on silicon for many, many years, and it's very logical that they would also looking to explore. And you can see how close we are working in partnership with them together. When Google announced their latest CPU, it was a keynote here in San Francisco, and the next person to walk on was Jensen, talking about the collaboration that we're doing in DGX Cloud.

And the H100 instances and the work that we're collaborating on building a better cloud with them -- and with the Adam [ph] reinvent, with AWS. So the whole ecosystem is advancing AI and of course all these advancements were being done in plain sight with the optimizations like we've talked about for improving and providing that horizon tag with AI. They should talk to what their contribution wants to be and what they're looking to specialize and to offer, that was unique to themselves and how they want to -- they see the opportunity. But I can speak for NVIDIA. The -- and the other competitive aspects of it.

We are moving extremely fast. We -- and that often unfortunately requires people that have a sense of perspective or know what's going on and be thoughtful about the compares and what's being claimed and making sure that the information is -- that you have all the latest. We've announced H200. We've -- those blogs I talked about with TensorRT and NeMo and transformer engines and (inaudible) published all the numbers and they're real end-to-end. So we open source all that.

We provide the full accuracy and throughput with the latest NVIDIA software. And that's really important because the innovations are happening so quickly, you always have to have the latest NVIDIA software to understand where everything is at. With H200, we're the first provider with HPE and (inaudible) memory technology, we work very closely with the memory partners, as I mentioned, the whole ecosystem (inaudible) market. That's a 144-gigabyte GPU with five terabytes a second of memory balance. And then -- and which doubles performance in and of itself over a H100 and then the latest innovation on Monday just doubled in again.

So we're cooperating at like a 4x plus clip above A100, where A100 was just a few months ago. So that also has created this growth of AI, has also given us the benefit to accelerate our own investments. And I think as Jen-Hsun shared with the investor community, we're now pumping out new architectures even faster than before and have stuff on our road map for next year and continue. So this is a fast pacing -- fast-paced field. The technologies are advancing.

AI is definitely advancing. NVIDIA is continuously taking it -- after you purchase your data center full of GPUs, they know when you get increased performance and throughput and reduce costs for years after -- during those data centers. And those aren't just one-off purchases. Obviously there are capital expenditures that last for three to five years long. There are investments that need to be -- that are going to -- the AI factories that are producing the assets that are going to fuel the growth of those companies move forward.

Very difficult to predict with the next -- I mean I need to see some of it, but like where AI is going to be in three years. But you know that NVIDIA is continuously part of that system and optimizing the software stacks and making that -- making those data centers productive as AI evolves as partners in the journey. So -- and that obviously goes into all the economics and the decision-making that happens when you deploy at scale when we're deciding on your AI infrastructure that you're going to consume.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Yes. Good. I think this is probably a good junction to open up for Q&A. So my colleague Yanko [ph] is here to announce the questions.

So, Yanko [ph], over to you.

A - Unidentified Participant

I guess we can couple some of these questions together because everybody is interested in asking what is your strategy to compete with AMD on inference given their products -- the price point that the products are coming in compared to the H100 and the claims that they make against of superior performance to the H100?

A - Ian Buck {BIO 18454865 <GO>}

Like I said before, I encourage the community to look very closely at how the end-to-end workload is measured and the performance that is delivered. It's why we participate in industry standard benchmarks like MLPerf. MLPerf was created by

Google and Meta and others in the industry and has an end-to-end benchmarking in that, which includes accuracy and throughput both for training and for inference. So you can see the work that we've -- and contributing to the basket of models, we've been the only company that's actually smoothed every benchmark in that suite since its inception -- four or five years ago, we continue to do so. Make sure that we understand the latest NVIDIA performance, latest software.

And if you change the economics by 2x in a software organization that's available and TensorRT, it changes the math everywhere. The other part is that we're moving very quickly. our H200 has the new HBM3e memory, which is now the 144-gig and the 5-terawatt second and the -- which is -- which we shared all the numbers and have those and all of the blogs and the performance there. And then the last comment I'll make is that AI isn't -- and inference isn't just a chip benchmark. It is about delivering performance across the node and we're at scale.

And whether it be eight GB inside of system or 32 GB inside of system and providing that end-to-end throughput. So it's easy to talk about flops and tops and numbers. But in the end, the decisions we're making this happens is what's my throughput, that's my performance on end-to-end of my model. And then as that model evolves and changes and scales, you're making sure that they are getting the benefit of the ecosystem and the software improvement that is coming not just from NVIDIA engineering, but also all the partners (inaudible) platform. So please use that...

A - Unidentified Participant

The following -- the next question, does NVIDIA plan to pursue a chiplet architecture down the road? Or can you keep staying ahead of the competition with the monolithic architecture?

A - Ian Buck {BIO 18454865 <GO>}

So I can't talk about future products or Simona will come over here to you. The -- but I will talk about -- let me talk about compute density for a second. And if you kind of see what's happening in AI, where if you just go all the way back to the pre-AI and web scale, it was all about math-produce, where it was all about AI [ph], these datasets were filled with Ethernet at scale, if you will. And it wasn't about -- density was not a factor. Now if it over to AI and computing, computing was generating revenue.

More you can use your power, your budget, your space on computing on a number. And the less watts or dollar she spent on sending bites around which doesn't compute, it just moves data, you want to densify.

By densifying, by bringing the compute closer together, you spend less energy, less joules, less watts on moving data and you can optimize for costs as well going from optics to copper to PCB. That extends even further inside the chip. The cheapest way to move data within the piece of silicon.

Chiplets are great in that regard. It does provide for a better form of -- entire form of communication. But of course building it would -- it's still better to obviously be able to communicate on a chip than be able to drive signs across chip or across multichip models. We see the benefit of multichip models, that's what Grace Hopper is. Grace Hopper we call it super chip because there's basically two of the chips and put it right next to each other.

In this regard those two models circles together, we can drive it at 900 gigabytes a second of communication, basically taking a new NVIDIA technology and building one super chip. So you have to think about densification and trade-offs and what is also -- and you see that being built out, is driving things to technology that we've been cooling and building. But I always think you're going to see a large chips because of our ability to optimize the whole architecture for compute and apply the ideas of densification that isn't just at the silicon level or the package level, it's at the server and at the racking at the data center level as well.

A - Unidentified Participant

All right. The next question, how do you foresee the relative mix of customer types for data center GPUs evolving for NVIDIA in the future? Hyperscalers are the largest category today but it seems like governments, traditional enterprises, startups and other categories are emerging as a bigger percentage.

A - Ian Buck {BIO 18454865 <GO>}

So the -- yes. The hyperscalers are two kinds of customers. They obviously are providing the compute capability and infrastructure for their own services. Amazon talked about how they're using NVIDIA GPUs to improve the buyer and the seller experience. They use NVIDIA GPUs to help the seller write a effective, descriptive product description -- with the model that was deployed with NVIDIA GPUs.

They obviously use it for music search. So you search in Amazon Music right now, the query is not literally the query that goes into the infrastructure. It actually is processed by NVIDIA GPU with NVIDIA software for, I'll call, to apply an AI model to it to figure out what you really mean. (inaudible) song or a genre to get the right to provide a better search experience in music. They used NVIDIA GPUs and our NeMo framework actually to train the collection of those models that are in Amazon Titan that they have in Bedrock, for example.

So they're an internal customer, obviously. And then, of course they're -- they provide GPUs to the market through their public cloud instances. And you see that from Azure to GCP, the AWS. That will continue and continue to scale. The enterprises are becoming a much bigger portion, obviously of the consumption of AI factories.

And you're starting to see that now. So that will grow. The ratio of those 2, perhaps difficult, you're -- I'm trying to compare two exponentials and that's never a good idea. But they are both -- we will see the enterprise become a significant portion in terms of consuming the factory. It's early days now, but that is definitely a trend that's slash on.

The new one, I think is the sovereign and AI nations. The ability for every -- every country sees that has learned that AI can be a resource to their domestic industry and to help their nation advance, either solving important problems in healthcare or climate change or weather forecasting. This is probably more of the science side of things, but an important research and we now see how AI can make -- understand those things better to make better policy decisions or how they influence policymakers to provide -- so they can see what (inaudible) for their economies, landscapes and people. And then, of course be a resource for industry as well. And not every industry you can afford to have an AI supercomputer, but government can provide the capability.

So we see that as a new -- Japan has already talked about doing -- providing infrastructure for itself, the U.K. And you hear about in the U.S. as well, and that's definitely a new trend. It's hard for me to prescribe a mix of that, but that is the -- those are kind of the growth curves. And I think it's going to be probably more balanced moving as we continue to scale out AI for the next -- rest of the decade.

A - Unidentified Participant

The next question, there is a view in the industry of NVIDIA pursuing a kind of walled garden strategy similar to Apple, while it seems that your competitors have an open source approach to their software and also with partners with something like the ultra Ethernet consortium. To what extent is that true?

A - Ian Buck {BIO 18454865 <GO>}

Yes. Thank you for that question. And certainly, we do like to talk about our technology a lot. And -- because we're in engineering, we're a true technology company in that regard, which is probably pretty unique compared to our customers and others. The -- first off, we are an open company.

Look at our -- just look at the harbor, okay? We make available our GPUs as individual GPUs. We make available our GPUs as HGX baseboards. We make available our rack designs to the OEMs, to the cloud so they can take the reference architecture and then deploy it or modify it or make it different. If you look at our NGX initiative, it's a reference architecture of that.

Here's how to build a GPU server that -- we're going to continue to maintain these bounding boxes, these thermals and electricals. So as we move fast, you can invest once in a form factor and know that you're -- you don't have to redesign everything all that available to configure and build and allow every partner of the stack to innovate in a way.

Of course if they want to work directly with NVIDIA, we have our GX business to allow that, but it is a very small portion of the activation that we do across the entire sustainability system and whether it be the clouds and be sustainability, the on-prems or the supercomputers. From a software stack standpoint, the same is true. So we -- our NeMo framework, the framework that we use to train the megatron -- the megatron models at 530 billion parameters.

The same AI framework that Amazon used to do those models and to train some of those use cases, we open source. It's all on data. The TensorRT LLM software stack provide -- has all of those reference models of the latest Llamas and Falcons and you name it. We open sourced all of those techniques and how to train so that the world can have that access when need open source because likely they have variance or modifications where they need support themselves. So all of those technologies are open source and available to the ecosystem to modify or change.

Some go all the way down to program in the individual GPU, that's fine. But the software stack, those libraries are -- open sourced. There's a couple that are closed and since that -- they really truly are optimized for NVIDIA -- by NVIDIA engineers in terms of internal library. Frankly, it's stuff that would be very difficult for anyone other than NVIDIA engineers actually to understand. But the -- from the Hopper stack, everything we're doing in DGX Cloud, which is just sitting on top of their cloud, it's all sitting on top of the that (inaudible) sitting on top of (inaudible) an ODM or OEM platform, all the GPU is open.

And all the software innovations include ones we announced on Monday those technology pieces are all there really want to consume it direct as an end-to-end solution for NVIDIA or pick up those libraries and interface into integrate into their services. Unicorns, for example, just pick up those pieces and apply them to them or tell us, give us feedback so we can make even better. You guys may not see that, but we do that all the time to make that OpenStack more optimized. Another way of putting it is we're the one AI company that works with every other AI company. So everyone is on our platform wherever they stack to get the benefit of it.

It tends to gravitate toward the top just because they get the compound value of everything that we're doing, but they're not precluded from taking any bits and pieces and -- or tailoring and specializing for their workload.

A - Brett Simpson {BIO 3279126 <GO>}

Maybe -- maybe just one final question from my side, Ian. And on China. Can you maybe just share your thoughts on how might play out for NVIDIA in China? And I guess obviously we're going to see a compliant GPU shipping into that market fairly soon. But how do you think that's going to be received, maybe just firstly?

And then secondly, do you think that we're going to see training move offshore with -- in China? And then lastly, just indigenous platforms, and we are hearing a lot about Huawei scaling up or trying to scale up. Do you think we're going to see a shift towards indigenous platforms in the market there?

A - Ian Buck {BIO 18454865 <GO>}

Yes. Unfortunately, I can't go into the road map in that regard. So some of those questions are good questions, but it can't be answered in this forum. I will -- look, the opportunity for AI is one that is -- hits every nation, every country, every business and be able to meet that demand and obviously stay within the regulation, we'll continue to do so. I mean as we have.

So that is the -- we're adhering to serve the AI market. However we can and we get the guidance that we provide over -- have provided to us as a company. So that will continue. And for the forward stuff, I think we'll have to wait for those things to be talked about and announced.

A - Brett Simpson {BIO 3279126 <GO>}

Fair point. Good. Okay. Is there any last questions there, Yanko [ph] or should we...

A - Unidentified Participant

No further questions.

A - Brett Simpson {BIO 3279126 <GO>}

Okay. Great. Well, I think we're out of time. So Ian, thanks very much for coming on today and sharing your thoughts. I really appreciate it, as always.

And thanks everyone, for dialing in.

We're going to close out the call now. So thanks very much.

And we'll chat soon.

A - Ian Buck {BIO 18454865 <GO>}

Yes. Appreciate it. Thank you.

Bye.