Arete Technology Conference

Company Participants

- Ian Buck, Vice President of Accelerated Computing Business Unit
- Simona Jankowski, Investor Relations

Other Participants

- Brett Simpson, Analyst, Arete Research
- Jim Fontanelli, Analyst, Arete Research

Presentation

Brett Simpson {BIO 3279126 <GO>}

Okay. Thanks very much, and hi, everyone. Again, it's Brett Simpson here. It's my pleasure to welcome Ian Buck, who you all know, runs the Data Center division at NVIDIA, one of the key growth engines of the business. I think this is a particularly interesting time to connect with Ian and Simona as we are heading into a new product cycle with Hopper. We were looking back at when NVIDIA launched Ampere, I think it was back in early 2020, and the business was doing about $1 billion of quarterly revenue back then. And looking at it today, it's almost $4 billion in quarterly revenue.

So it's been a really strong period for the business and well done. Now many of you know Ian's the inventor of CUDA, and we're going to try to touch on some of the sort of software opportunities that NVIDIA sees ahead during this presentation.

So Ian, thanks very much for joining us today.

Ian Buck {BIO 18454865 <GO>}

Thanks for having me. Thank you.

Brett Simpson {BIO 3279126 <GO>}

We also have Simona Jankowski on the line, who you all know is VP of IR at NVIDIA. So I'm going to pass you over to Simona to read out the safe harbor. So Simona, over to you.

Simona Jankowski {BIO 7131672 <GO>}

Thank you, Brett, for hosting us. As a quick reminder, our comments today may contain forward-looking statements, and investors are advised to read our reports filed with the SEC for information related to the risks and uncertainties facing our business. Back to you.

Questions And Answers

Q - Brett Simpson {BIO 3279126 <GO>}

Thanks, Simona. And maybe just to start, Ian, can we maybe just recap the 2022 period? And what's to-date for you the most looking at the division and the market opportunity you see ahead, and maybe touch on some of the sort of customer reactions to the new products and what you think is happening at the bleeding edge of AI and how you see this all evolving into 2023?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. I mean it's been a whirlwind. It probably hit us back in, obviously, 2020, and that's when we launched Ampere, which feels like a lifetime ago, but it was only back in 2020. We -- throughout COVID, we launched Ampere. The following year, we announced our Grace CPU work. And then this year, of course, we've announced the Hopper. NVIDIA is on a new clip with the investment and the importance of computing across the industry, so are the innovations that allow us to continue to invent new GPUs, new architectures, new algorithms and serve the growing market of AI and computing in the data center in general with a biannual clip at this point. We're making new GPU architectures every two years.

We've committed to a new CPU architecture also every two years. And it was very exciting to launch Hopper this year. Hopper is a GPU that was specifically designed to advance AI as well as HPC. It was -- has specialization for the transformer-based models, which is the foundational model for -- used today in large language models, in generative AI, and being applied to pretty much every domain where computers want to see, listen, learn and generate back.

So we're very excited about Hopper's introduction. It's in production now. The PC AI [ph] systems are coming available from the OEMs and the cloud hyperscalers, all have Hopper and are bringing it to market.

Q - Brett Simpson {BIO 3279126 <GO>}

Great. And maybe just given the environment run at the moment, I think a lot of investors are trying to understand whether the environment we're in right now, has it made you think any different about the opportunity you see ahead? When you look at the next sort of six to 12 months and you've got a lot of product rollouts to do, et cetera, but is the enterprise engagement still the same? Is the demand signal still as strong as you were thinking six or 12 months ago?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. The foundations of who and how we engage have only grown over the last two years and they've remained consistent, but increasing in activity and motions. It starts certainly with some of the big hyperscalers doing their own work, some of the best work, Microsoft partnering with OpenAI, Meta and Google even and others, taking advantage of our GPUs and the software solutions that we're layering on top in order to develop that next-generation AI technology capability, moonshot or foundational model. And a lot of the work is in software, not just hardware. In fact, NVIDIA has more software developers than hardware developers, for example. A lot of work on frameworks like PyTorch and TensorFlow. We also develop our own large language models ourselves with our own supercomputers and share that with the rest of the world to help move the ball forward and to give us the experience to continuously improve the -- our platform, along with taking the input from every other major hyperscalers as well as our own experiences. We build our own supercomputers to try that to test that and to advance the state of the art.

That's certainly on the big hyperscale side. Activations in the cloud have continued to grow. Certainly, every enterprise has a cloud-first strategy of some kind and they're bringing their technology, their workflows, or their way of doing business to the cloud. And as a result, our -- the public cloud activations of our platforms has certainly grown. And we saw that with A100, and we'll continue to see it grow with H100.

In the enterprise side, it has been -- it is certainly on an upward trend. The challenge on the enterprise side was meeting the enterprise developer where they are. They don't have the talent pool necessary that Google or Amazon or Meta may have. But they have -- they see the opportunity for AI and how it could be changing and advancing their businesses. So our role in that has been to help move that forward to provide an enterprise-supported platform, that's a big part of our NVIDIA AI Enterprise platform is to help provide a stable supported platform that enterprises can rely on for getting their best -- not only the best performance, but the support they need directly from NVIDIA, working together (inaudible) and cloud partners.

And then also helping activate the right kind of software that they can take advantage of, not just natively the AI frameworks but targeted services and support that comes from working with, there's lots of great activity happening, of course, in the start-up community, providing cloud-based services, where enterprises can share their data and get the results back without having to build the software from scratch themselves. For -- we're seeing also, of course, the cloud providers themselves providing those services, a lot of them developed or deployed on NVIDIA GPUs along with the -- growing out of a start-up into the broader ISV category as well.

Even traditional ISVs may be doing in the areas of numerical simulations are trying to deploy AI as a surrogate model for some of the first principles physics simulations are doing for product development and other sorts of things. And finally, the -- NVIDIA, ourselves are trying to move the ball forward as well. So in areas where we see an opportunity to move the market forward, we'll invest ourselves, and we've done that, particularly in the healthcare space with our Clara platform, applying AI to things like medical imaging or proteomics. We're also doing it in speech AI. We've

sort of provided some foundational support in our Riva platform to help enterprises take advantage of end-to-end speech AI, which offers really cool abilities to do customization and train new models, new voices and additional things like that.

Recently, in our last GDC, we announced our NeMo large language model service to build a service where enterprises can customize a very large model and we provide it open. So we do GPT-like models, a variety of different GPT sizes. We've done our own 530 billion parameter model called Megatron, and we make it available as a service. So enterprises can take our existing train model, a model super smart and then tune it -- fine-tune it basically with the technical prompting. We have only a couple of examples, a few hundred examples up to 500 or so and if your model learns how to answer the questions the way that customer is looking to [ph] be answered. You're not asking a question of the broader Internet to do your customer support, you've already given a few hundred examples of what customer support calls and answers look like, (inaudible) answer appropriately in the context for that business.

So for the enterprise side, we're at the beginning -- we're still at the beginning of AI adoption, we're seeing a lot of interest. And from an NVIDIA standpoint, companies can engage at all different levels, certainly consuming our sort of first-party offerings that we decided to invest in, but also through the ISVs and start-ups as well as, of course, the platform holistically.

And that's a big part of our strategy, just to be open to move the ball forward in helping companies and customers and developers and users get the benefit of what AI can do.

Q - Brett Simpson {BIO 3279126 <GO>}

I want to come back to the services model around Megatron and GPT more broadly a little bit later, but I wanted to -- I mean, I guess if you go back to some of the GPT-3 models, you were coming out with A100 around that time. And it was very much an NLP-centric upgrade cycle that we could see playing out for NVIDIA the last couple of years. Can we talk a little bit about where we are today? What's bleeding edge, what's being worked on in the labs that you see?

And I mean, when you were on our conference a year ago, Ian, you talked about model sizes are growing like 10x a year. Is that still happening? I mean we must be well into the trillions of parameters -- the model sizes are in the trillions, I guess, of parameters. Where does this go over the next couple of years during the Hopper cycle? And what type of workload is Hopper ideally positioned to deliver?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. Certainly, the natural -- the NLP or large language model, LLM community hasn't slowed down one bit, and you can see that. So many -- both major players, as well as now start-ups, and -- are taking advantage of what these models can do. They are large. They tend to be -- well, previous models may focus on things like computer vision, understanding what is this a picture of or where in this picture is the stop sign. That's a capability that -- if you think about it from a first principle sort of

understanding perspective, it's fairly straightforward. In fact, in nature, we see animals and bugs and other things, have basic vision capability like that, identify the objects, where is it and call it out.

Language is different. Language is unique. It encompasses -- to understand language, you not only need to know the words that I'm saying right now, but what they mean in order to make context and doing anything useful with them. And that means that we have to sort of -- these models have to encompass or understand the corpus of human understanding. As a result, they're trained on the entire Internet, literally, these data sets are basically big scrapes of the Internet that are then cleaned up and tuned and trained. So they tend to be obviously much bigger. And we -- there's the 530 billion parameter model, which I've talked about already. And certainly, new work -- new models are coming out that will be or are already in the trillion of parameters. We're still short of like brain scale, and there's a question about what -- which is the 150 trillion parameters, we're only (inaudible) trillion right now.

Q - Brett Simpson {BIO 3279126 <GO>}

So 2025, 2026.

A - Ian Buck {BIO 18454865 <GO>}

Yeah. We'll get there, Moore's Law. So as the models are -- as the capabilities are growing, so will the models. So basically, the large language models today, obviously, can provide very convincing chat dialogue back and forth. They can be used for sentiment analysis, they can do all sorts of things because they understand human knowledge.

The next step, obviously, is what we're seeing right now. That remains true. If the models are large, certainly, you don't train them on a single GPU or even a single server, you tend to train them across a pod or collection. We trained our Megatron model in about 4,000 GPUs and the final training runs took about a little over a month. Of course, there's a lot of R&D to get to that point to develop the model. Once you have it, by the way, it is certainly -- once it's been trained, it can be tuned and to different use cases. That tuning step is actually much more approachable and only requires less infrastructure than going to train the entire model. But at some point, people are developing new foundational models that can be then used for those different use cases. And that's where a lot of the very interesting research and development by the biggest players that have that infrastructure.

One of the goals with Hopper was to bring down the cost of large language model training to make it more applicable, and we've done that. The Hopper runs trick [ph] and train 6 times to 9 times faster than Ampere, requiring 6 times to 9 times less infrastructure depending on the model. And that's really because it was designed to do that transformer layer that we talked about, and take advantage of reduced precision and mixed precision up and down the stack to still maintain the accuracy.

It's easy to offer a bit [ph] floating point, but it's hard to make it work and work well. In fact, we use our Selene supercomputer to train all the different heuristics to then

[ph] baked it into the software/hardware of Hopper to make it successful. The other part of Hopper that made NLPs more applicable is that we dramatically improved the inference performance. This is the performance that takes to run the model in production, not just train it but deploy it. Hopper is 30 times more faster than Ampere was. And that allows people to take these largest possible models and still run them with a reasonable amount of infrastructure. And with a single DGX-like system to deliver reasonable real-time performance in running these models. That is going to dramatically broaden the applicability of NLP models. A lot of people run it for more and more use cases than what we probably saw previously, which is bespoke offerings for some of these large models. That's a goal, and we're definitely seeing a lot of interest in deploying Hopper for that.

And then, of course, the next step is what else can it generate. What else can these large language models produce other than question-and-answer kind of responses, chat box responses or customer service responses and sentiments responses? And you're starting to see that hit in the mixed modality space in generative AI. The work being done in stable diffusion by folks like Stability AI or Midjourney, Runway and others are showing how we can take large language models and connect them to image generation to the output, a picture rather than just -- than text.

And that's super exciting as an example of another place where these large language models and the underlying generative portion of AI, where you're understanding now the corpus of all images and what they mean, so we can (Multiple Speakers) is another great example. If I project farther out in the future, you can imagine AIs generating all sorts of things, generating potential chemical compounds for the next-generation therapeutics, material properties for next-generation manufacturing and material science. I just came back from the Supercomputing Conference in Dallas, Texas, where the talk of the town is we're building the foundational -- using supercomputers to build the foundational models for science and HPC for next-generation for all the use cases that we have across science community.

So all sorts of opportunities. The -- what's going to get us there is access to the platforms to the technologies to those researchers, those users, those companies can take advantage of it and deploy it for all those different use cases. And that's what makes part of my job fun, is to help get there.

Q - Brett Simpson {BIO 3279126 <GO>}

And maybe if we can focus a little bit on training because I guess what we've seen in the last 12 months, I mean, Meta has been quite public, they've built this massive cluster 16,000 A100s, I think in their earnings report, you talked about Microsoft rolling out tens of thousands of GPUs. I think Oracle is also pretty active in that in building these large training clusters, large exaflop supercomputers. Are we heading into a phase where the amount of people that can keep up and building these massive training clusters is going to consolidate down to a handful of players? Or do you see hundreds of potential training clusters like that getting developed? I'm just trying to get a sense of how that -- you see training evolving.

And I guess part of that also, I want to come back to the services opportunity where you can license a pre-trained model. Is that going to be where a lot of enterprises get involved in AI rather than train everything from the ground up, they will license something that's been pre-trained and they'll pay a services fee for that rather than necessarily buying a big hardware cluster?

A - Ian Buck {BIO 18454865 <GO>}

Well, one of the things you can sort of bet on is that AI still hasn't -- we haven't tapped out on the different applications of AI. Because fundamentally, AI is just a statistical trick, if you will, an algorithm to take data and write code. In the case of Microsoft, that's literally what it's doing. It's helping you write code with our co-pilot program. But those applications of AI seem to be balanced because we have -- everything we're doing in computing in the enterprise revolves around the data that we collect from customers, the data that our business generates, operates and teaches us and the communications that we have with customers and the flow of business is also generates massive amounts of data that can be interpreted and understood and taking advantage of AI to improve and make better.

So that has been -- that has continued to grow. And as a result, access to infrastructure for developing those AI, either as a -- from a first principles foundational level, or taking existing foundational models and applying them are both growing. So that is -- so there's definitely both interest from start-ups and others, major ISVs and big companies to explore both, explore building new foundational capabilities. For that, you do need infrastructure.

And what's interesting now is that the clouds are starting to provide that infrastructure before building an AI supercomputer was just that a bespoke supercomputer. Now certainly, Microsoft led the way there with their announcements providing their independent and interconnected GPU clusters in the cloud so people can rent that kind of infrastructure. We're seeing it now with Oracle and many others and companies like Meta, building out that infrastructure. They're one of many that are either going to build it themselves on-prem or be able to consume it rented from the clouds. So that's why we're seeing the growth of AI infrastructure and particularly scale-out infrastructures available in the cloud.

It's important to note that you don't have to rent all of it. They're designed to be fractionalized all the way down to just one or a few notes. But before everything was single, was focused around the single node capability of how many GPUs can we fit in a single server, which usually is maxing out about eight or 16. Now the infrastructure in the cloud is being scalable thinking about data -- renting entire data centers or rows [ph] or pods or just a collection or a half rack to do the various work for developing those foundational models that are going to serve those different industries.

And it's broad. It's not just the largest of largest models, those are the ones that tend to get all the excitement and press certain capability. But designing models, new kinds of models that can do different kinds of workload foundation level requires some level of infrastructure at some scale.

Then the second part, of course, is the -- applying these -- taking these foundation models and applying to different use cases and businesses, and we certainly see that in things like speech AI where you have a foundational model that's trained on English, but you even [ph] want to do it with a certain accident or for a certain voice. Those things don't need to be retrained from scratch. They can take those foundational models and deploy it. And just there's just more and more businesses that are figuring out how to take advantage of that or their services from the -- from all sorts of different providers that can then go and consume it. So that's causing the growth of GPUs across the cloud and still on-prem. I think there's an ebb and flow in terms of the on-cloud versus enterprise, we don't -- we're a platform of choice. We try to activate all channels. So that's where it's coming from.

So I don't think it's -- AI is not squirreling away into a corner where only a few people can afford to do it. The -- there's certainly more interest in exploring the outer limits of what AI can do with the large infrastructure, and that will continue to be so. But I believe what we're seeing is the diversification of different use cases for developing foundational models, all sorts of different scales by all different kinds of players. The ability now to consume it and get it from the cloud as well as just on-prem. And then finally, the breadth of different services like you mentioned, to take advantage of these -- of what's been trained, what's been learned for [ph] services.

Q - Brett Simpson {BIO 3279126 <GO>}

Yeah. Interesting. Interesting. And just thinking about the -- I mean, you mentioned we're still at the early phases of building out the workloads and the compute requirements that go with that. But I guess when we look at some of the hyperscalers that are at the bleeding edge today and making big investments for public cloud and serving instances to thousands of enterprises, I guess when we break it all down today, we can see that some of these guys are maybe spending multiple billions a year with NVIDIA. If you just break down your overall revenues, you can see that we're at that sort of stage. And these are large investments that are now being made by the hyperscalers.

How do you see this scaling? Are we looking at -- is it inevitable that hyperscalers will be spending $10 billion plus at some point soon? How do we think about Fortune 500s, sizing that investment opportunity? Do you see Fortune 500 companies spending $1 billion on infrastructure to really differentiate in AI? How do you think about this from years out, the development opportunity for NVIDIA specifically?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. I mean I think you're still at the early stages. If you think about the -- it's difficult for me to project the dollars and when. Obviously, it's been exciting and will continue to be exciting, and we definitely see the growth moving forward. One way to look at it is there's a logical case to me that every server inside of a hyperscale data center should be accelerated at some level because they're all operating on the data either flowing in or out of the data center or east-west within the data center.

And each -- every bit of that data logically should be interpreted, understood by an AI to make an insight and to improve the function and operation of that data center

and what it's trying to do and impact the outcome results. Today, probably less than -- around less than 10% of the hyperscale data center is accelerated. And what's preventing that -- what causes that growth is just more people identifying the capability of how AI could impact or improve that part of the workflow, that part of the user story, that part of the capability of that part of the data center.

And particularly at a time now when data center space is precious, we've certainly seen a growth in data centers and they don't pop up overnight. They take years to plan and years to build. So that is -- accelerated computing offers a great way of them optimizing the data center space they have by moving stuff from CPU-based services, which can consume a lot to using AI to reduce the amount of infrastructure they have or preserve the space they have and to do more to the norm [ph] to grow, a data center space is an important metric, a foundation for them to do more and to grow and accelerating computing allows them to do more with less data center space.

Likewise, with power and energy efficiency. We can do these operations at scale at a much lower total energy cost than a CPU infrastructure, may be able to do. So there's a lot of interest in identifying those workloads and shifting them to an accelerated portion of the data center in order to do all those things, optimize data center usage, improve the throughput of a workflow with consuming less data center space and be more efficient with power in order for them to naturally grow.

So I think the -- that will be the application. Now, of course, some of it may -- will happen inside the data center, and it's difficult for folks outside to see it. The first people to do it are the hyperscalers themselves and you're seeing that with some of the services they're building themselves, some of them also make those services publicly visible and you can do your own projections on how they get translated to inside and operational lines inside their business.

On the enterprise side, it's the activations by the different enterprise ISVs and major enterprise service companies, and we're seeing that. We're seeing it in -- and you can go and look at some of the success stories that we talked about or that Jensen's talked about in our GDC keynotes, you can look at the -- or the Oracle world where he joined on stage with Safra and talked about how the enterprises are adopting. So it's a question of -- it's still a fairly small percentage of the total enterprise software stack that can be (inaudible) AI. And there's lots of valid reasons for that, but it's something that's going to be changing fairly quickly.

Now it's becoming a point where in order to be competitive in the enterprise software world, you need to be deploying these, offering these capabilities and giving the CIOs or CFOs of the consumers of ISV software, the Fortune 500s, they need [ph] our strategy. So as you can imagine, they're asking all their ISVs, how are you leveraging AI to make my business run better as a service or incorporated into the services they already have? So we're already at the stage where a lot of the baseline functionality capability, obviously, the operations of enterprises are served by the major enterprise ISVs. They're all trying hard right now, and some have or others are doing it now, activating AI in their foundational enterprise software stacks.

And with that, we'll grow the -- they'll get faster, more efficient and you'll see more adoption of GPUs across the cloud data centers as well as the on-prem offerings so that those companies can run those workloads efficiently, low as possible latency in order to deliver what they need for those services. So I think that's the interesting thing to track. I think it's looking at the ISVs and where they're deploying AI for the major Fortune 500s, where they're deploying it. The other place, I think, as you look at by sector as well, the interesting things happening in retail. We're now talking -- I think McDonald's has publicly talked about doing pilots on -- for speech AI, sort of talking -- looking at different ways of providing different kinds of services. Those are obviously all AI-driven and very exciting to see the retail community take advantage of AI.

So you can certainly look at it from a segment by -- vertical by vertical segment. And once one or two players go, they all see the opportunity, they'll turn around and very excited to take advantage of it. So there's lots of investment and activity happening across the enterprise. Our role in that is obviously, working with all those ISVs and provide that foundational layer so that companies -- these Fortune 500s owners that need to get the direct support from NVIDIA now that we have their back and that's kind of what our NVIDIA AI enterprise offerings is doing for the market.

Q - Brett Simpson {BIO 3279126 <GO>}

Yeah. Yeah. That makes sense. Maybe I wanted to just talk a little bit about -- I mean, I guess, when we look at 2023, big architecture changes in compute generally, the DPU is going to start to become more of a thing. I think, obviously, you're removed with Grace on the CPU side and Hopper is a big architecture change and all the interconnect that sits around it. How would you describe this transition that we're seeing now from a -- very much a CPU-centric approach? Are we -- is the compute complex now becoming a platform sale for NVIDIA rather than just selling sort of discrete cards? And then what does that do for content? I guess looking at your content today within the server, do you see dramatic increases because of your -- the ramp of BlueField and Grace and the switching infrastructure that you're developing?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. I think a couple of things. One, obviously, our -- the existing business, and people are very interested in getting access to great infrastructure for computing, and we'll continue -- and that is today is x86 with our GPUs, either sort of a standard server accelerator, which looks like a PCIe card. It's quite a large one, but similar to what you may have seen in the GeForce products, they go into standard servers up to eight or 16 of them. They just don't have graphics connector, they're optimized for computing. And then they also come in smaller sizes for inference use cases, they tend to be much smaller edge like accelerators. So we go -- and then we go all the way down to the embedded space, which is measured in single-digit watts, and that's being deployed for edge kind of use cases even down to conference room or telco kind of applications of robotics. And then you go all the way up to servers that are explicitly designed for doing computing at scale, and you can see that in our HGX and DGX solutions.

Moving forward, I think the -- what's interesting is this just trend toward the data center as the unit of compute. And people are not just looking at servers or components or chips in order for the advancements in computing an AI and I think they are looking at the entire data center and how they can optimize the entire data center for compute. Many years ago before this, it was data centers at hyperscalers, where basically map-reduced [ph], kind of SQL-like data centers driven by I/O more than compute. So they would deliberately choose a smaller CPU than focus on interconnect at scale along with storage.

Now with compute being king, you're looking at the entire data centers, how can they improve their compute utility of this entire data center to optimize it as a unit -- total unit of computing. If you look at it that way, now we want to look at all capabilities of the data center. The accelerators, obviously, the CPUs, how they're connected, how the network is integrated and together and how far it can scale across the data center. And you can see NVIDIA investing in all three of those.

We've done -- obviously, we continued down the path of accelerators. We're now building DPUs. We just -- with both BlueField 2, we've announced BlueField 3. We've got a strong road map in there as well with going all the way to BlueField 4 for connecting the interconnect of the data center together and doing that intelligently, both to allow to do things like in-network computing, so some of the operations should be done at the line speed on the network to provide the security and isolation and also the -- do more of the software-defined networking that other people have had demonstrated to be very efficient at providing those kinds of services, particularly for cloud-based use cases, very important, and as well as needing to be performant. We're seeing cloud providers provide InfiniBand infrastructure now alongside with high-speed Ethernet, both are good in market depending on the different use cases, and we're there to serve both. But certainly, demand for high-speed and high-performance Ethernet and InfiniBand is growing quite a bit, and they want to (inaudible) it intelligently by applying the DPU.

So that's rolling out across seeing more and more traction there and then the CPU side. So one of the interesting things about CPUs is that traditionally, they're connected to the GPU via PCIe Express standard bus. We continue to support that. We'll support that for as long as we shall live, basically. But there is an opportunity there. So can we -- is there another capability we can provide, perhaps a bespoke capability by putting Grace next to the Hopper GPU and tightly integrate the two. And in fact, we've done that with our Grace Hopper product, which was announced this year, where we put the two chips next to each other and we built that high-speed interconnect, the NVLink chip-to-chip interconnect, which offers 900 gigabytes a second of bandwidth between the wo, and it's fully coherent.

So that's a big uplift from what you get with PCIe, which is at most maybe 100 gigabytes a second, between 50 and 100, we're getting 900 with Grace Hopper and it's fully coherent. So this GPU can now operate on the entire data that's contained in the entire server, all of both the CPU and (inaudible) GPU memory. What that really makes it allow us to do is work on those largest of models like we talked about, you basically have it for operating at a different scale, and we're seeing some interest in

Grace Hopper for doing those large-scale recommenders and deploying large-scale NLPs. The rest of the market is going to -- obviously, has been well optimized to run everything on the GPU, and so that will continue to exist for quite some time.

So we're seeing an interesting opportunity with Grace Hopper to sort of push the limits of where we can go in both training and inference, being able to deploy these very large models with minimal infrastructure that provides sort of that real-time latency performance or before you may have to spread the model across multiple GPUs, now you'll be able to execute on a single GPU combined with the Grace and its memory to deliver a large language model in real time.

So very excited, and there's another interest in the HPC and supercomputing communities as well. They really like the coherence in CPU and GPU, they really love Arm architecture and the ecosystem has come a long way in terms of Arm. So we're very -- and you can see everyone investing in Arm as well across all the different verticals. So excited to see that take shape and we'll be bringing Grace to market next year.

Q - Brett Simpson {BIO 3279126 <GO>}

Great. Very interesting. Well, I think because we're -- I'm conscious of time. I think this is a good segue perhaps to open up for Q&A. And I've got my colleague, Jim on the line. So Jim, can you relay any questions you've picked up from investors so far?

Q - Jim Fontanelli {BIO 20458321 <GO>}

Yeah, sure. So the first question we had in was, I think quite straightforward, how quickly will it take for Hopper to be the majority of your data center shipments?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. Every transition is a little -- we're not like a little different than what you see in the gaming side, which tends to be a bit more of a switchover. In data center, we're operating enterprise use cases. So people have qualified and deployed at scale, the Ampere A100 GPU is obviously quite successful and still quite difficult to get in the cloud if you try to ask for an instance, it still shows up and sold out in a lot of places.

So I expect that to continue to trend. And then usually, our transitions happen over a period of a few quarters or more as different -- as Hopper becomes available through different markets. First, we'll [ph] come to market -- is coming to market now is the PCIe products, you can get them from all the major OEMs. And then the HGX NVLink connected baseboard products will be coming to market Q1 of next year and then from the clouds as well. And in H1, we're going to transition differently and also by different regions. That's just the nature of how they do their work and the work -- the hard work it does to take to qualify not just a server but for hyperscale and entire hyperscale data center because once they deploy at scale, it's fire and forget, these data centers are massive. And the way they execute at scales to make sure they have the best possible quality and test it out to work at scale. So that will happen all throughout 2023 and expecting A100 to continue on to 2024 as well, but it will slowly blend over. And we saw that before with the A100, we'll see it again. Of

course, everything is influenced by economics and market conditions and trends in AI. Certainly, a lot of the largest customers doing those large (inaudible) are very interested in getting their access to Hopper quickly. So I'm looking forward to seeing those announcements.

Q - Brett Simpson {BIO 3279126 <GO>}

And just by reference there, Ian, how long did it take for A100 to cross over the 100? Any sense there just as a benchmark?

A - Ian Buck {BIO 18454865 <GO>}

Yeah. We saw -- I believe it was between six and eight quarters worth of -- and we continued to -- Voltaire [ph] at this point has largely tailed off, but it was at least, I think, a six to eight quarter transition. A lot goes into that, obviously, but that's kind of what we saw last time.

Q - Brett Simpson {BIO 3279126 <GO>}

And Jim, maybe we just got time for one more. So...

Q - Jim Fontanelli {BIO 20458321 <GO>}

Yeah. So we had maybe a longer-term question, which might be a good place to wrap up. So the question is this industry often looks at transitions in decade cycles. I think there are estimates out there of around 3% to 4% of servers accelerated globally in '22. How do you think about market TAM and adoption curves with regards to servers that will get accelerated as we go through the decade?

A - Ian Buck {BIO 18454865 <GO>}

It's certainly picking up. I mean -- and that's one of the reasons why we're investing in an accelerated roadmap ourselves by doing new GPUs every two years, new CPUs every two years and DPUs every two years. It is simply because of the demand and interest for -- this isn't like the CPUs before where you wait for a process technology transition or a memory technology transition. We've dialed and tuned in our manufacturing processes so that we can ship often our A01 Silicon is our production product like it was with Ampere and that allows people to get access to our technology even sooner.

So with that, the opportunities for acceleration just become richer sooner and that is unlike perhaps before an era of Moore's Law, where you just waited for the next technology transition to get faster. Now it's people want to get faster, faster. So that is -- the main limiter, I think one of the main limiters will be the number of companies, ISVs and Fortune 500s that are -- get building up the talent pool to figure out how to adopt the technology for their use case. NVIDIA, we can only -- we can provide those platforms, in some cases, the vertical platforms like we do with Clara or in Merlin for Recommender Systems or Maxine for teleconference and communications. Those again are not in product solutions all the way. They provide a foundational layer that allows the enterprise to meet that (inaudible) gap.

That -- as more and more players enter into that space and they can get access to our technology affordably, especially with Hopper reducing the cost of access to perform an AI infrastructure or even more perform an AI infrastructure, I should say, that is the ramp driver for the broader adoption of our platform across the enterprises. So I'm expecting that, like you said, we're still at the beginning of that little curve with 3% to 4%. I expect as more ISVs adopt, more companies invest, more services get launched by everyone, that number will start -- continue to curve up even further.

And certainly, in this era where compute is paramount and access to infrastructure is key and yet, we're in a bit of a data center crunch perhaps and the importance of running things more efficiently with greener capabilities, better usage of our -- the power that we do have access to, it's a compounding factor in driving up that growth. So I think it's going to be exciting. It is exciting. And this Hopper transition is just yet another turn of the crank, the 2023 will bring even more and again, '24. So I'm happy to keep coming back to you guys and telling you what the next thing is as we roll it [ph] off.

Q - Brett Simpson {BIO 3279126 <GO>}

Yeah. We'll definitely take you up on that, Ian. And if Hopper is anything as good as the results we saw from Ampere, it's going to be a very interesting couple of years. So good luck with the new product introductions into next year.

A - Ian Buck {BIO 18454865 <GO>}

Thank you.

Q - Brett Simpson {BIO 3279126 <GO>}

And Simona, thanks very much for your time. Great discussion and we could have gone on a couple of more pages of questions, but I guess we'll read that for another time. All right. Well, thanks very much for joining us, guys. We really appreciate it and here ends our session. Thanks very much.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you.

A - Ian Buck {BIO 18454865 <GO>}

Thank you.