ARETE VIRTUAL SEMIS CONFERENCE

Company Participants

- Brett Simpson, Arete Research Services LLP
- Ian Buck, NVIDIA Corporation
- Simona Jankowski, NVIDIA Corporation

Presentation

Brett Simpson {BIO 3279126 <GO>}

(Technical Difficulty) We also have Simona Jankowski. I'm going to pass over to Simona to read out forward-looking statements. So Simona, over to you.

Simona Jankowski {BIO 7131672 <GO>}

Thanks, Brett. As a reminder, this presentation contains forward-looking statements, and investors are advised to read our reports filed with the SEC for information related to risks and uncertainties facing our business. Back to you, Brett.

Questions And Answers

A - Brett Simpson {BIO 3279126 <GO>}

Thanks, Simona. So Ian, maybe wanted to set out, maybe just set the scene here in the market. I guess 2020 was a breakout year strategically and commercially for NVIDIA, for your division specifically.

I mean you launched A100. You finalize the Mellanox transaction and launched DPU. There's obviously the proposed acquisition with Arm. There's a Chinese entity list restrictions that deal with, I mean, surging sales. It was obviously just a crazy year.

It's clear -- at least it's clear to investors -- that we're in this giant kind of innovation phase for compute right now.

Can you maybe just break down what you see in front of you running this division? What do you see over the course of 2021, 2022? I think we can all recognize it's still really early in the adoption curve of accelerated compute. But what's going to define this year for you after a crazy 2020?

A - Ian Buck {BIO 18454865 <GO>}

It was a crazy 2020 and certainly from where I'm sitting. NVIDIA launched a whole new architecture, their Ampere GPU or A100. We had a launch from our homes. In fact, Jensen from his kitchen.

It was -- and I think while other companies slowed down or struggle to execute, I think we thrive in that environment. We're already a virtual global company, and we knew how to do what we do, and we've launched it from a kitchen virtually. And A100 has been a great success for us.

And it continues to be so. In fact, the market today is only just now really gives them enjoy the benefits of what Ampere can do for the businesses. It's a 20x improvement of our previous architecture. This is the kind of innovation that -- the kind of improvements that we make generation over generation.

We do that because -- we were able to achieve that because we look at it from a whole stack perspective. It's not just the chip and what the performance of the chip can do. Obviously, the chip is critically important, and Ampere added new technologies like Tensor Float 32 and the whole new Tensor Core architecture.

And on the other -- on the gaming and visualization side, it also added ray tracing -- the new ray tracing capabilities, which is amazing. But we optimize all the software stacks as well.

So what we're seeing right now, of course, is that adoption of the A100. It's targeted, first and foremost, of course, to the AI and cloud markets and broader data centers. So the ability to train some of the world's largest models are being done now with Ampere with A100.

It also, of course, is seeing success in HPC and scientific computing. Our strategy is to build 1 GPU, 1 platform that's highly leveraged, and at some point, different vertical capabilities and markets and libraries and SDKs to take advantage of that 1 GPU architecture underneath.

So what I see moving forward is, first, AI continues to obviously grow. It's software, writing software that is defining the next-generation applications that you and I are going to interact with in the cloud or use for business insights and other things. It's a totally new technology. It's something that the world is still learning how to use even from first principal standpoint.

And you see that in our business. It grew a lot originally in the hyperscalers, the Googles and the Facebooks and the Amazons and the Alibabas of the world, who have the people and the technology and the knowledge to go invent it and build it from scratch. We're starting to see a broadening of that.

So the cloud is starting to consume more and more of our GPUs as the rest of the world is learning how to use this technology, either NVIDIA's direct software or software partners in the AI stacks that we're working with. And finally, we're -- so on the training side for sure and also now increasingly in the enterprise.

The bigger companies who maybe need -- who now understand how they can apply their business, their decision-making and use this AI technology for things like understanding their data they're ingesting, understanding their product adoption, understanding their -- the recommend what to give to their customers and also changing how to interact with their customers, particularly here at COVID, where we're all kind of -- it's a whole new world where video chat is what we do.

A - Brett Simpson {BIO 3279126 <GO>}

Good. I mean I guess, since we -- this is largely a nontechnical audience. I wanted to spend a bit of time on the market and the strategic opportunities you see in the next couple of years.

And maybe, first of all, we've seen a lot of research breakthroughs in AI in the last 12 months. I mean BERT large, but now GPT-3. These model sizes are going crazy and algorithms, too.

But I mean, we're still -- it still looks like we're in the applied research domain here with a lot of these breakthroughs. And I think one of your colleagues, Bryan Catanzaro, was saying that maybe in a 5-year view, it's possible that companies could be spending $1 billion in compute time just to train a single language model. So I guess what's the practical applications we're going to see from all these breakthroughs that we're seeing on the model size? Yes.

A - Ian Buck {BIO 18454865 <GO>}

It is -- people are still figuring out the limits of AI, for sure. I think one of the areas where it started was, was in computer vision. And that was the idea of what is this a picture of a basic question. And there I think we're largely mature. We have well -- this was the starting point of AI 6, seven years ago was asked -- asking the computer that question.

And I think we're pretty good about that. We're getting pretty good also identifying different parts of the pictures, putting boxes around them, even identifying individual pixels of what's consisted of my face. And it's the background, that's AI green screening. And you're starting to see consumerization of that technology.

When Bryan talks about these -- these huge models, the 1 billion parameters that you hear about, that's in new areas particularly in natural language processing. That's an area where we're going beyond just understanding computer vision, which, if you think about it, bugs and cats and dogs and every intelligent life force, or some intelligence, can do some level of computer vision. Their brains are pretty small.

Language and language understanding, however, is a much broader, a bigger -- challenge, right? You not only need to understand what I am saying right now, but also what I mean, my intent and understand it and take action on it and respond to it and communicate back to me in the same way that I'm communicating with you.

A - Brett Simpson {BIO 3279126 <GO>}

Yes.

A - Ian Buck {BIO 18454865 <GO>}

That's a very more -- much, much more intelligent problem. But the opportunity there is very large, right? This is how we interact as humans and we're now going to interact as computers.

And how all of our data and how we interact with our businesses, actually, they do through language of understanding and document. Things like chat bots or recommender systems or sentiment analysis or understanding a dialogue and making decision or a call to a doctor's office or a help request or your financial adviser. We're understanding the meaning and intent.

We have technologies like virtual assistants or chat bots that want to do this kind of understanding to improve the user experience, and in the end, make it better for customers.

So that's why conversational AI is a big thrust right now. We call that the general conversational AI. There's -- and the market is huge. There's 500 million support calls a year. There's 200 million meetings going on. There's 200 million smart speakers in the world.

That is where -- and obviously, in the new world of COVID, everything is digital. So it's all going through digital medium, where it's an area where there could be an AI agent on this call to help transcribe it, understand it, summarize it, capture actions. And you can see people talking about those things. Those are the big models.

The other area is recommenders. I think recommenders -- or think about it, this is how you interact with the Internet now. Search, it really is a recommendation. When you go to Amazon or you're browsing website to buy something, they're recommending things to you.

When you visit your news feed, they're deciding what you see and you don't see, obviously, the implications of that. Those models not only have to present something that's accurate to you, but something you're likely to buy and also moderate content at the same time.

Here, we have a huge data problem of the amount of data across -- think about every user of social media, every news post and every product. That's a massive matrix. So those models tend to be very large. And certainly, the data they're trying to capture is very large and are getting even bigger.

So there, we're seeing significant growth in the challenges of AI and also where people are buying some of their infrastructures. And it goes right -- obviously, they're highly motivated because it often goes right to the bottom line of their economics.

One of the examples there, I guess, in conversational AI is Microsoft Office. So I don't know if your audience knows this, but if you turn on grammar checking, Microsoft is actually using an AI model based on BERT, which is a famous AI model for doing language understanding.

So it's finding the grammatical errors, highlighting it in blue. That model got so complicated because it's trying to understand language that they had to move it to a GPU. So today, that runs on an NVIDIA GPU. And as you're doing live Office 365, it's in beta now that will actually do a highly accurate grammar correction for you by running all the sentences through a GPU and running that AI in real-time for inference.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Yes. Interesting. Maybe just switching gears a little bit, Ian. In terms of the adoption curve here, the technology. I mean you mentioned computer vision. It was a 6-, 7-year ago thing, we've kind of solved that. We're moving on to the next thing.

But I guess when we do go to surveys, talking to CIOs, AI is not coming up and they are spending budgets as the big item or a big item today. And I guess, the technology still feels quite nascent outside of maybe the top 50 hyperscalers who are developing next-gen models and doing all the recommenders that you laid out. How many organizations do you think are actually using accelerated compute today at real scale?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I think the adoption curve followed, like I mentioned, the people who have the capability to use it. So if you think about the adoption curve or how it's getting consumed, there's 2 parts of it.

There's the training part, which is developing the AI service or model for your use case. And then there's a deployment part, which is the inference of deploying AI as live in a service and the infrastructure to do so.

The adoption curve follow the people. So first, you needed the people that understood data science, understood it like a machine learning ops flow and have the capabilities to understand what was being published in the AI community and apply it to a particular business. That was pretty small. In fact, it all started with hyperscaler.

So I think out of the Fortune 500, 5000 and others, it's still relatively small. Where we see the traction actually is them going through an ISV partner, a start-up. The start-up community is actually quite rich in this area.

They find the right partner that understands or has the particular technology and working with them to apply it to their particular business use case. I see a huge majority of the use cases outside of the big hyperscalers starting that way.

And I think your community probably sees the start-up activity in AI right now, and it makes a lot of sense. A lot of smart people with some great ideas that could be applied to many different possible business use cases.

At that point, you develop that service with that partner, probably with your own data scientists and with the company. And then you need to turn on and deploy it. That growth will come -- that's a different growth vector.

One is training and developing the model, which, of course, scales with the number of people, number of problems and model set size. So you kind of multiply those 3 together and you kind of get the size of the training opportunity.

On the other side, you have the inference and deployment. That is all about the data. So as the service is running, the amount of data coming in as an opportunity for AI to optimize. And that will scale with the amount of data and size of business.

And there is an opportunity for -- you want a GPU in front of every one of those connections to run and operate and execute the AI at the latencies required for that service, whether it be interactively on a news feed.

You click on it, you got to do hundreds of thousand inferences in milliseconds or a voice conversation where we have to maintain a 20-millisecond inference latency just to keep up with the number of utterances, at least at the speed that I talk.

A - Brett Simpson {BIO 3279126 <GO>}

I mean I guess this is obviously going to take time, and there's a lack of data scientists out there that really understand this as well. But we see a lot of guys prototyping today and kicking the tires in the cloud, trying to figure out what they're going to do with the technology.

But how do you see a path or what sort of path do you see to corporates building industrial scale in AI? And is this going to be more sort of a cloud hyperscale play in providing services and libraries in supporting your view?

A - Ian Buck {BIO 18454865 <GO>}

So I think a lot of it starts in the cloud. Obviously, it's easy to take that first step in the cloud when you're -- from a budgetary standpoint, you're paying by the hour. And certainly, we -- from NVIDIA's standpoint, we make assure our architecture is available and our platform is real everywhere.

I work with all the major cloud providers to activate and make sure they have the latest A100 GPUs in their cloud, so all the customers can use them or they're inferencing on their T4 GPUs. At the same time, we make sure all the major OEMs have access to our technology. And we offer the same base for the same GPUs both at the same time, so the market can consume it, however they want -- they choose. And we remain open that way.

The adoption tends to start usually where your data is. So if your data all resides in the cloud, it tends to start in the cloud. Or if it needs to be on-prem or your own data centers for privacy or security reasons, people start there, too. So that usually is the kicking point for where things get started. They can do experiments in the cloud, but when they get to real stakes, they start with where their data is.

As this scales up, certainly, people look at their cloud bills, and they decide whether they want to be -- do an on-prem or do it in the cloud. In many cases, I see also stratification. So people don't want to manage risk.

So they want to make sure they can work with multiple different cloud providers and have an on-prem use case, on-prem capability as well to manage their risk or where they want to run different services. So it's not necessarily locked into one particular cloud or sort of one particular on-prem or hybrid. I think people are going to necessarily naturally want to do both, and it will vary from industry to industry, for sure.

A - Brett Simpson {BIO 3279126 <GO>}

And just talking about the cloud, what role do you think they really play here? Because from a silicon perspective, on one hand, they're buying your GPUs, guys like Google and AWS.

And then the other, some of them are developing their own chips and software. So how do you see the sort of internal chip efforts developing? They obviously don't have the resources that NVIDIA has and the reach. So just love to get your thoughts in terms of these internal silicon efforts.

A - Ian Buck {BIO 18454865 <GO>}

Well, sure. So I think the -- first off, it's great that everyone's coming to conclusion they need accelerated computing. I mean the AI use case is a great one where compute can turn into the software, which can turn into these business opportunities. And we work with all of them.

One of the things that is unique about NVIDIA is that because we are at the forefront of a lot of this technology and that we are a full stack company -- remember, I'm releasing not just the GPUs that everyone can get access to and start and using, but I'm also releasing all the software stacks on top of it as well.

We work tightly with our friends at Google and TensorFlow or Facebook and PyTorch and many other stacks actually, some of the ML ops stacks as well. We also released some of that, our own container, our own optimized versions and make them freely available on our NGC container registry for download, and get the latest certified, tested, validated AI software that works, whether it be in the cloud or on-prem.

We're the only AI company that's working with all the other AI companies, including all the hyperscalers and we take all that learning and all that knowledge, all those engagements, whether it be a recommender system, a conversational one, we're doing stuff on AI for video chats just like this conversation as well. That informs us. We help the customer.

We make our products better, our libraries and SDKs better and then we make it better for them, and we incorporate them back into our platform. Some of those in corporations is just improving and making our frameworks better and more optimized or the vertical stacks that we optimize.

And that same feedback gets right back into the hardware teams and the architecture, and we benchmark and tune and test ourselves on how we are doing. And that team also sees how they can improve their architecture and make new investments in the architecture itself, whether it be the compute to the caching of the memory and define the next generation.

And as someone who stays in the middle of that, I can see how quickly that happens at NVIDIA. As a result, we're releasing new containers every month for these different workflows and the different frameworks as well as new architectures every year now to continuously innovate because we're super not done.

This is an area of rapid innovation, which is great for companies. They can get on that train on that bandwagon, and then they can ride the wave if they're programming to those interfaces at this high enough level. So that when NVIDIA comes along with the next GPU and the next version of our software, they just see that 20x that they saw with Ampere. And that story will repeat itself over and over again.

So everyone is investing. I think it makes total sense. AI is the platform where compute could just write software. I think from where we stand, and they know, is we're making our platform available everywhere to all those customers, and they can benefit and we can benefit from it. And I think what mostly advances is the adoption of AI inside of the broader enterprise, which is pretty exciting.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Yes. And then just, I guess, on the cloud specifically, we've seen in cloud compute, a big concentration amongst a few big players. And I listen to AWS, they talk about -- they've got 10,000 customers for machine learning today.

Does this become a contemplated services market, it's dominated by a few public cloud players? And if not, do you see a large opportunity to sell PODs or systems to corporates and build your own enterprise channel?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I think a couple of things. One is there's a -- there's not -- definitely, we see -- there's not one way to consume AI, for sure. And there will be services that are going to stand up different. AI is a capability. So different services will stand up to and serve that capability, and then we fine tune. This is not just one software program that runs them all.

The models are widely different. And the use cases are different. It's -- some off-line, some streaming, some more ensemble-based. Some are more -- and their latency requirements are wildly different. So I think you'll see a wide variety of different services and capabilities, and everyone will compete naturally on those verticals and those capabilities in the market.

In terms of the hybrid and the on-prem, that, of course, will be a decision that people want to make in terms of how they -- how much they want to consume in their utilization, first and foremost, and where their data is. From a SuperPOD standpoint and the PODs for your audience, this is the ability to actually put together multiple GPUs in a system and create an AI infrastructure for a broad data science team.

I think as AI gains more adoption, people are going to want that infrastructure internally. First off, it allows them to optimize that infrastructure for their workload and their model size. Some models can train up to thousands of GPUs. Some are more optimized to running an embedded use cases and maybe more a different size and different scaling.

So the scaling parameter is something we need -- that we can get today on-prem that we're starting to see more of it in the cloud, but also offers a lot more diversity of different architectures in the cloud.

And the hyperscalers have to make their own decisions on when to make that investment because, obviously, they're hyperscalers. They can decided the scale they want to deploy something.

We are seeing more and more options in the cloud, which is also very exciting because it offers different places and points for customers to choose from. In the end, it will boil down to rent versus buy decision on the -- for themselves. And that just comes down to the economics, which is a different -- which is the same conversation, I think, we've had all the way between cloud versus on-prem.

AI will be no different there, except there are differences in how you want to necessarily configure the machines, certainly for training. And in terms of how we know we want to put together going all the way to InfiniBand stack, going -- providing a high-speed storage solution that's tightly coupled to your compute, which is usually very important for training at scale.

And one of the reasons why we built our own SuperPOD, Selene, and made available to our customers is because some of those capabilities are obviously -- they're basically supercomputers, which we can offer to the community. I also do supercomputers, too, but we can offer the community to get that head start. And then the clouds are starting to figure out how they can participate in that as well.

And we are seeing some InfiniBand happen in the CSPs and the cloud providers, which is exciting, and some of the different storage options. Obviously, it's a different challenge for them to deploy that stuff at scale across multiple regions and making it available for rent, and making the economics work for them as well.

But I think you're going to see a lot more diversification in the kinds of instance types and capabilities of different clouds for people to get access to the different ways that technology is all connected together.

A - Brett Simpson {BIO 3279126 <GO>}

Yes, yes. Yes. Interesting. And then maybe you mentioned Selene, Ian. And I wanted to just touch a little bit on this infrastructure investments that NVIDIA is making. I look at the CapEx budget at NVIDIA, and it's going up and annualizing over $1 billion or so today.

And I know you're investing in new campuses, et cetera. But does NVIDIA see an opportunity to offer AI services to enterprises? Is that -- I mean we see the GeForce NOW business model, very early stage. But is there an opportunity, a wider strategy to be in the services business yourselves?

A - Ian Buck {BIO 18454865 <GO>}

The reason you see us building Selene and building our own infrastructure is, first off, to be successful in AI, you have to be a practitioner. It's not like you -- you have to understand this technology intimately to understand how to advance it. Because like I said, it's not just one chip or one core.

The problems that people want to do today, that capabilities are super exciting. But you need to think about like the data center as a whole. And you can't just think about programming on one chip. These models are too big to fit on a single GPU or you're going to -- in some cases, a single server. You have to think about, they have these spread across the entire rack or row or data center.

And certainly, to train them in any time that's reasonable, which most 2 weeks, you don't really want to go more than that, keep a data sciences team, which is also very expensive, busy. You need to do multi-node training at scale, which is batch-based training. So now I'm thinking about the entire data center.

And you can't do that on paper. In order to be successful in AI, you have to be thinking about that problem at scale. And as an engineering organization, we build it, and we do it. We do it to make our products better for sure. We also create our own products into vertical markets, and we've chosen a few. Self-driving cars is one of them.

So a huge portion of the infrastructure we're seeing today is being used for our self-driving car initiatives for NVIDIA DRIVE and the work we're doing with our self-driving car customers to give them a turnkey solution and partner with NVIDIA to deliver a self-driving car capability.

And that -- so we had to build the whole pipeline from the data ingestion to the labeling to the training to the simulated an environment where we run the car in simulation and crash it 100 times without ever hurting anyone and seeing what works and doesn't work before we ever put it in the car and have actually test it in real time. So we have to build out that infrastructure.

As a result, we also learned about our products that make them better. We certainly also have our own research teams, which are specializing in different capabilities and learning how our -- advancing the field forward ourselves. They also teach us a lot about our product.

And in terms of services and what we can do, we work with -- this is not just one market. This is a one use case, right? So our strategy is to make available our AI platform to every different start-up hyperscaler, OEM and enterprise, so they can consume and develop their own capability.

Remember, we're at the phase of a new form of computing, and people are figuring out how to deploy it, use it for those different use cases. For some markets, we do choose to just go vertical because we can advance it forward.

We've done it in self-driving cars. We've also done it in conversational AI, like I mentioned. So we have a software stack called Jarvis, which comes with pretrained models that can help do speech recognition and some language understanding, also some text-to-speech capabilities.

We've done it for a recommender system. So we've published a software stack, which is a baseline capability for doing large recommender systems at scale. And this was informed through our engagements with hyperscalers, we call that stack Merlin.

We also have a video collaboration stack, a UCaaS stack called Maxine, which is used for improving the experience that you're having right now over Zoom and helping those apply AI technologies to things like noise cancellation, and green screening, super resolution, et cetera.

So we'll choose some, and they're really there to help advance the AI field forward. And it's a different -- it usually doesn't go all the way to a turnkey solution. It usually goes to active enablement, and the rest of the market can then take it from there and deploy the technology and see the benefits of it.

A - Brett Simpson {BIO 3279126 <GO>}

Can I maybe just ask, Ian, if -- I guess if we look back at the V100 cycle for the data center division, it was a bit lumpy. You had good times and you had slower times.

And I guess it takes time for customers to digest what they're buying before they need the next amount of compute. Do you think that's going to be the -- is that just the behavior that we're going to see in A100? Or do you think that it will be a different type of market this time around?

A - Ian Buck {BIO 18454865 <GO>}

It's getting faster every time, and I think it's because the market's maturing. So what we saw in V100 was that -- remember, this was three years ago. That wave of AI, it requires some time for people to absorb and understand the technology and see the next ramp.

I think coming -- the software has matured, the way the go -- the markets and the technologies, software has matured such that it's going -- it's much faster for us. We certainly see some digestion in the sense that we do a product transition. Obviously, people want to refresh a large fleet like a hyperscaler.

That does happen. But it's what we have experienced from V100 to A100 is much more -- much -- happening much faster. And in fact, we grew in our data center business where many other companies, I think, struggled or saw some downturn.

And I think it's because of the rapid adoption of AI and then the observation of the 20x and the need to go to that next platform, really pulling the product forward and making sure that it gets into market as quickly as possible, making it available.

So I think that those transitions are happening faster. There's always going to be some transition. We have to manage them, we certainly spent a lot of time focusing on it. We prime the pump. We make sure that everyone gets their platforms ready, and we make our technologies ready early, so people can understand it. So when it hits the market, everyone is ready to absorb it.

Backwards compatibility is a big part of it, making sure all the frameworks and all the AI software stacks are already ready to go on day 1. We don't wait for people to test out our hardware or try it, reporting it ourselves ahead of that in simulation with early versions of the hardware so that when it gets ready to launch, we have all those stacks ready to go. And I think we've improved significantly since from V100 to A100.

A - Brett Simpson {BIO 3279126 <GO>}

Yes. Excellent. And maybe just on the competitive dynamic that you see ahead. I mean you've had -- you have a dominant franchise and particularly in training, but also, it's getting -- as accelerated compute comes to inference, you're doing great.

If you look out 2 to three years from now, how sustainable do you think your position, your market share position and accelerated compute looks like to you? There is a lot of start-ups. They've taken a lot longer to get into the market. But if we look 2 to three years from now, what's your perspective on market share?

A - Ian Buck {BIO 18454865 <GO>}

A lot changes in 2, three years. I mean if I think about AI 2, three years ago today, it continues to evolve and grow. The models and the things we're talking about are widely different. We used to talk about ImageNet.

We used to talk about ResNet. Now that's kind of like table stakes. If you can't do conversational AI or natural language processing, that's where a lot of the opportunity and the growth is coming from. The -- and we're well-established already in those other places.

So the -- I think the other -- that's one part of it. I think the technology of AI is evolving really rapidly. And it's why it's so important that you be a practitioner of AI in order to keep up with the trends because to understand what a model can do and actually do it, that's an art and a craft, and you learn a lot about system engineering, computer architecture, interconnect, storage systems, the whole data center.

So I think if I look out in the 2- to 3-year time frame, it's -- and really, it's happening now, is you're thinking about the data center as a computer because that's really what these new data centers are being designed to do.

They're designed to train the network at scale to give a data science team the throughput and keep those in the tools to develop those natural language or conversational agents that are going right to those use cases. And the problem sizes they're trying to solve are huge. It's every product, every user and every data point, every news feed.

So those are -- now we're optimizing, and NVIDIA has shifted into a company that's thinking about the data center as a unit of compute.

And that includes not just our GPUs, but the systems, the networking, increasing the CPUs and how they all fit together in a broader software stack and a Kubernetes workflow and how to manage enclaves with data scientists that are doing different services and developing those things and then turning around and flipping that infrastructure over and applying to inference.

You mentioned inference were -- before, 3, four years ago, the vast majority of inference was done on CPUs. Today, there are more compute for GPU -- compute for inference than there is for compute for CPUs in the hyperscalers, if you just add up the flops and the capabilities.

And that came from A100. When we designed A100, we started seeing this trend where the models are getting bigger because the capabilities want to get smarter.

And the CPU could not keep up in terms of executing that model at the latency necessary for the NLP models or the conversational agents that we talked about, particularly in text-to-speech, recommenders and some of the newer models. As a result, the workflow -- people started deploying GPUs for that. We saw it with our T4 GPU, which is why they're used in all the hyperscalers today for inference.

With A100, we made the architecture excellent both. It's a great training GPU, but also that same Tensor Core can run all the operations necessary for inference. So the reduced precision stuff either about (inaudible) or FP16. We also have this capability called MIG or multi-instance GPU.

This GPU actually can, inside itself, split itself up into 7 separate GPUs that presented independently in the system so that people can take the same GPU they bought for training and turnaround and serve up inference in use cases. And it's 2 to 3 faster than our previous GPUs, just each one of those single slices.

So we're seeing that adoption of -- you need to be good at both training and inference. Certainly having a workflow that allows you to do both and seamlessly go from training to inference is really important. And then being able to obviously run all these new models in the areas of conversation and recommender, that's super key.

A - Brett Simpson {BIO 3279126 <GO>}

And maybe just switching gears a little bit, Ian. Can we maybe just get an update on the Arm situation? I mean if there's an update on the Arm acquisition, that would be great. But I'm also interested in what do you do with this asset from a data center perspective?

I mean we can see the role of -- does NVIDIA see a general purpose CPU in its arsenal long time? Do you develop server architecture based on Arm general purpose server architecture? How do you see the Arm opportunity in the data center?

A - Ian Buck {BIO 18454865 <GO>}

Well, certainly, data center is the new unit of computing. So we should look at all the -- how we can advance or how NVIDIA can advance the data center. And certainly, that's my focus, and what I can talk to.

I think from an Arm standpoint, this creates a premier computing company for the age of AI. It combines the latest NVIDIA's leading AI computing platform with Arm's CPU expertise and helps position Arm and NVIDIA and all of Arm's customers and that entire ecosystem of customers for the next wave of computing, that age of AI.

So -- and of course, AI is also powering the Internet of Things, which thousands of items, bigger than the Internet of people at this point. So that makes it very exciting.

We certainly expand Arm's IP licensing opportunities and allows us to offer NVIDIA's technology to large end markets, including mobile and PC and a turbocharger Arm server CPU road map by help investing in the road map and advancing Arm and move faster and accelerate its adoption at the data center, at the edge and AI and of course in IoT.

And it certainly expands NVIDIA's computing platform from reach to -- from where we are today, about 2 million to over 15 million developers. So that's very exciting. Customers are very excited. I think the partners are very excited. That opportunity is great. And we can't succeed unless Arm's customers succeed.

A - Brett Simpson {BIO 3279126 <GO>}

But I guess, you see it as an important step to have CPU capability for data center in-house. Is that how we should think about it, particularly now that the business is scaling so much, the CPU becomes more fundamental to your solutions going forward?

A - Ian Buck {BIO 18454865 <GO>}

I think -- we always needed a fast CPU. In fact, Amdahl's Law is still very much a law unlike other laws. The -- I can make an infinitely fast GPU. But if we don't accelerate the entire workflow, the entire problem, I may -- I can accelerate 80% of the solution infinitely fast, but I'm still only 5x faster and they're stuck at 5x.

So that's why you have to think about AI and accelerated computing in general at the data center scale. By investing in the data center at data center scale, both CPU, DPU and GPU, all components of parallel serial and I/O networking, can you really get -- achieve those 20x speed ups that we talked about.

And you think about that entire canvas, and that's the innovation canvas that we're talking about in order to achieve the next-generation performance to see the AI continue to advance itself and continue to allow these breakthroughs to happen and turn them into business opportunities with all the world's enterprises.

A - Brett Simpson {BIO 3279126 <GO>}

Excellent. Very interesting. Well, I think we're out of time, Ian. Can I just say, really appreciate your time. Great, great discussion. We could go on for a lot longer with many other questions. But we really appreciate you coming on the event today and chatting with us.

A - Ian Buck {BIO 18454865 <GO>}

No problem. Any time. Thank you.

A - Brett Simpson {BIO 3279126 <GO>}

Our next speaker will be starting in a few minutes. We have Samsung's Network EVP, Woojune Kim. So please, can you click through on Zoom for that session. Thanks very much. Again, Ian, thanks for your time. Bye for now. Thank you. Thanks, Simona.