Goldman Sachs Technology and Internet Conference

Company Participants

- Ian Buck, NVIDIA Corporation
- Toshiya Hari, Goldman Sachs Group, Inc.

Presentation

Toshiya Hari {BIO 6770302 <GO>}

Good afternoon, everyone. Thank you, all for joining us at our Annual Technology and Internet Conference. I'm Toshiya Hari. I cover the Semiconductor and Semiconductor Capital Equipment space here at Goldman Sachs.

I'm very honored and very excited to have with us today Ian Buck, General Manager and Vice President of Accelerated Computing from NVIDIA.

We have about 40 minutes. I have a list of questions for Ian. But for those on the webcast, please feel free to type in your question. I'll try to get to them toward the latter part of the session.

With that, Ian, I'd like to get started. First of all, thank you so much for carving time out of your busy schedule. I'm sure you're being pulled into all sorts of directions. So I really appreciate it.

Questions And Answers

A - Toshiya Hari {BIO 6770302 <GO>}

This is a technology conference. So I suspect most investors in the audience know who you are and have heard you speak in the past.

But just to level set the audience, I was hoping you could talk a little bit about yourself, what your responsibilities are at NVIDIA -- and although I'm pretty sure there's no such thing as a typical day for you.

But I'm curious how you spend your time between internal responsibilities, customer-facing responsibilities and other things.

A - Ian Buck {BIO 18454865 <GO>}

Yes. Sure. Thank you for having me. No. It's no bother. This is a wonderful break from all the other things I do to get to talk a little bit about all the things we have done and where we're going.

My name is Ian Buck. So I'm the General Manager of Accelerated Computing here at NVIDIA. It's a long way of saying I'm responsible for the GPUs that go into the data centers themselves. And not less for graphics or gaming but more forward compute and AI.

My focus is on the CSP markets. So my day is filled with talking with the likes of Amazon, Microsoft and Facebook and Apple and Baidu and Tencent and Alibaba and you name it, the hyperscalers, which obviously, we, NVIDIA, of course, has a special relationship with all of them, given the work that we do.

My -- I'm also responsible for our work in HPC, traditionally my background. I joined NVIDIA in 2004 and started CUDA in 2006, came up and built it up through the engineering side, mostly focused on computing, HPC use cases, and still today, a huge part of our business.

But obviously, as it grew, it encompasses both HPC and AI today. So my time is spent with hyperscalers and some of the larger supercomputing projects that we do. And not surprisingly, those conversations tend to be very similar sometimes so...

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. That's a great intro. Thank you, so much. I wanted to kick off post that introduction, asking you to sort of look back and reflect on 2020 and potentially look forward to 2021. 2020 was clearly a very challenging year for many of us and for the global economy.

But at the same time, many of the secular trends in the technology space that we all expected to occur over the next 3, 5, 10 years seems to have been accelerated and pulled in. From where you sit, Ian, what were some of the key highlights of your business in 2020? And what are the top priorities for yourself and your team for 2021?

A - Ian Buck {BIO 18454865 <GO>}

Yes. So as I go and answer your questions, I want to remind your audience and my investor team reminds me. This presentation, this conversation may include forward-looking statements. So investors are advised to read our full reports filed with the SEC for information related to risks and uncertainties facing the business.

But with that out the way, I can get to some of your questions. And certainly, I don't think anyone could imagine this kind of year. And it was a real test to a lot of companies in the market.

Once COVID comes and we had to switch from an office environment, meeting customers' safe space, traveling to everyone at home and maintaining social distancing, and for our business, the inability to travel, the events, the conferences going virtual as well as engaging with our customers and our own data centers.

We are -- drive a lot of the AI innovation, and that comes through experimentation and development of AI. So we, of course, have our own, one of the world's largest AI computers that we had to upgrade and continue to invest in. In terms of achievements for the year, certainly, execution wise, I think everyone in NVIDIA is proud of what we're able to do.

We took COVID seriously early. We went home back in February. And -- but despite that and despite our culture, we're still able to execute. We launched the Ampere Generation GPUs, the A100 at our GTC Conference in the spring, and it was a home run.

Through all the work of pulling together that launch, finalizing the product, taking to market, getting all the hyperscalers and OEMs activated with the product, all happened during COVID. And as a result, the result -- the performance is great, 20x performance over the previous generation product, faster than anyone expected.

In the meantime, we upgraded our data centers to Ampere. We competed in competitions like MLPerf virtually through -- during COVID and delivering leadership performance and training and in inference. The challenge is AI is not slowing down.

The research community didn't slow down either. They continue to deliver faster, better, smarter neural networks. Basically, the neural networks are accelerating, doubling basically in capability and size in terms of number of parameters every 2 to three months.

And if you date back to ResNet-50, to GPT-3, that's like a 30,000x increase in computational complexity in those neural networks. And we've been racing to meet that demand through innovating the whole stack, of course, with the hardware, with A100 helped a lot this year.

The other big trend this year was that we finally saw the inflection on the inference side of the business. As we got our Ampere GPUs in, the growth of T4 in the cloud for inference workloads, if you add up all of the flops and GPUs available for inference and CPUs available for inference, we've actually tipped the scales now. And we have more compute on GPU for inference than we do on CPU, which is great.

Some of the things that are driving that are recommender systems, people figuring out how to apply AI for ad placement, content prioritization, research. The other one is NLP, natural language processing, speech workloads, being able to do transcription, virtual chat bots, all these kinds of workloads, which are an order of magnitude more complex than what the main use case they had in the past, which is more focused on computer vision.

The way I'd like to explain it is computer vision is a baseline capability that sort of when you think about it, we can see and understand what we're seeing with silicon dogs and cats and even bugs have basic computer vision understanding.

The other recognize it as an image, good or bad. Understanding language, that's a whole another level of intelligence. Not only do you have to understand what I said at the rate and speed at which I say it, but what was meant and come up with an answer.

And only humans can do that. And we're nowhere near to superhuman levels of NLP yet. We're trying to get there. That's one of the things that OpenAI demonstrated with GPT-3 and then turning it around to useful services and businesses.

So the company did really well during COVID. I mean it was a hard time for sure. We all have stories. We had to do data center upgrades, maintaining protocols, and we even had robots, supervisors walking around observing people, making sure the wiring is done right. But we're able to pull it off. And as a result, got the -- the market rewarded it and the customers appreciated it. And certainly, at a time when people care about cloud, we were there to deliver it with a next-generation platform.

A - Toshiya Hari {BIO 6770302 <GO>}

And Ian, maybe top priorities for the coming year?

A - Ian Buck {BIO 18454865 <GO>}

Certainly, we're seeing growth in continuing the rollout of Ampere and Ampere products. We'll see that, I mean you'll definitely see that. The emphasis on the vertical workloads, the conversational AI, some of that software is early still in beta. Now it will go more widely, also recommender systems.

We're also seeing the applying the technology to video and collaboration platforms. That's with our Maxim products. Those are obviously were -- some of these were invented last year. They're going to come to market now and activate.

So conversational has a huge market. We're having 500 million support calls, 200 million meetings a day. There's a huge area where AI has content and opportunity to add value. The same with recommender systems.

The challenge to recommenders is there's -- it's not a common input-output. We're not looking -- looking at one picture and then telling you what's in it. It's everything from customer sales data to web traffic to web content to stars and likes and reviews. So there's a much richer opportunity there. With that, people have to figure out different ways to apply the technology.

So I see inference definitely growing and conversational AI and recommenders being 2 big drivers in natural language processing in general. And the new neural networks are not going to stop.

People are going to continue to build bigger, more impressive, more capable AIs, and they'll need the infrastructure in order to deliver and get it done. There's a lot

going to happen this year, but certainly, from the AI front, that's the stuff I'm pretty excited about.

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. I definitely want to come back to some of the things you talked about there in terms of inference and the performance improvement with the A100. But before I go there, I wanted to ask about software.

As analysts that have a semiconductor background like myself, we tend to spend a little bit too much time on the hardware side and maybe too little on the software side. And it's pretty clear based on what you've accomplished and what NVIDIA has accomplished, software is a very critical component that supports growth in the business.

I was hoping you could take us back to when you were working on CUDA, I guess, in the early 2000s. What drove you and your company overall to create what is today a very critical platform that supports your business?

Explain to us how CUDA has evolved over the past 14 to 15 years. And how does it contribute to the competitive moat that you guys have in data center accelerated computing?

A - Ian Buck {BIO 18454865 <GO>}

So early on, certainly, we did a roadshow. We asked a lot of people in the industry who cared about computing or compute focused. They needed a way -- people have been doing GPGPU on graphics cards using graphics APIs. A couple of things came out very clear. First, they didn't want to learn a new language. They want easy access to take their programmers and developers and give them a way to program these for these systems.

That's why CUDA was based on C and extended to Fortran and other languages to make it really easy for developers to grok [ph] what a GPU can do, how to express the kind of parallelism it can do without having to learn a whole new language on a whole new platform.

I think we largely achieved that. It's very -- I can teach a developer CUDA in a day. As long as you understand the concept of a thread, you can -- and have decent C or a Fortran programming, you can get to it in the HPC space. That same model is extended to others.

One of the other -- that quickly evolved to building out the platform because it's not -- you can't just program the GPU. People expect a certain level of libraries and capabilities to get their work done. No software today is developed entirely on their own.

It's a combination of all the different libraries and technologies needed to make median application work certainly in the compute space. So our work has quickly grown, and it's quickly grown to building an entire platform of SDKs and libraries that enable it, starting with basic math libraries, single processing, linear algebra operations to now sparse algebra, DNN libraries, video codecs, DALI for input and output processing.

We have over 80 SDKs now which track the different industries and give them not just a way to program the GPU, that's one step, but all the libraries, all the capabilities that have already been optimized for our GPUs and are backward and forward compatible. So if you use them today, you go from Volta, if you're from 100, to Ampere, you just get the 10x, right, like that without having -- and the customers who are on our platform get to enjoy that.

The other thing that I think -- so our platform has grown substantially since that first day where we had to program with Kernel, had a program with one piece of code in a GPU, to providing those different -- it must be hundreds of libraries now.

And I mentioned over 80 SDKs to make that successful. The other thing we recognized, and this was a big secret to our success, is the fact that we're a full stack platform. We didn't try to like standardize or define the next-generation ISA. In fact, we don't even release the ISAs for our GPUs.

We decided to tackle the problem at a higher level. So you can meet the programmers up where they're programming and how they're expressing their problem in different ways they want to consume it. And then keep the innovation space deep so that we could innovate on the broader software stack underneath them as well as the hardware and change the hardware over time.

So as a result, we have a huge number of software developers. I think we actually have more software developers and hardware developers in NVIDIA, building and investing in those algorithms, those libraries and how to write code and optimize for our GPUs, to meet the developer up here, where they're doing the algorithmic work.

In the meantime, that also gives us the flexibility to completely redesign our architecture. We can break ISA compatibility all day long. In fact, we do in order to move the needle forward up here because we learn and engage with the customer.

So I think that's been a big part of our success is that we look at the whole stack. And we even have people that engage with the acquisitions themselves in each of the industries because we stay focused on certain ones that we know are going to add value and then optimize the whole stack from the chip to the library to the compiler to the runtime to all the vertical stacks that we have.

So that was a critical decision early on. Don't worry about ISA compatibility, meet them up here, provide this whole innovation space. And that's really how we get the 10 to 20x kind of performance that you see in some of our numbers is because we're

innovating all the way up here. This is where the programmers need us. They developed on Volta or Kepler or Fermi before that. They just -- they ride that wave, and that's why the platform is so enjoyable to so many people.

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. Thank you for that. And then Ian, just wanted to shift gears a little bit and talk about the growth drivers of the business. And you sort of touched on this earlier in your response. But hyperscale, obviously, has been a very important market for you. It's sort of the business that has turbocharged growth in your overall data center business.

You spoke to things like conversational AI, recommender systems as important drivers of your business this year and obviously going forward. Based on the conversation you're having with the Amazons and the Googles of the world, how are you thinking about growth? And what are sort of some of the killer applications for your hyperscale business?

A - Ian Buck {BIO 18454865 <GO>}

Yes, hyperscalers tend to be the tip of the spear because they both serve a market in terms of renting inference and -- sorry, infrastructure, as well as meeting customers themselves. And they often have the development teams and the engineering teams to go invest and build those first adopter kind of use cases like we saw at AI in the beginning.

So as a result, they're obviously a very important customer. We learn a lot from engaging with them. We partner with them to help them take those technologies to market, whether it be conversational AI, language processing or to the next model for doing recommender systems. We're -- we learn a lot from that engagement. We serve them well. They enjoy -- their developers enjoy our platform.

At the same time, they can turn around and turn into services themselves or provide the core infrastructure that the rest of the market can then rent from them and engage. Now it takes longer for the rest of the market to catch up to where Google is because they don't obviously don't have the brain trust of Google or an Amazon or an Alibaba, but that's happening.

In fact, our vertical industries, we're now running about 50-50, where it is half of our business is hyperscale, representing 50% of our data center revenue, where those vertical industries represent another 50%.

So they're catching up and learning those techniques. Part of it is because they're the SDKs, the libraries, the application frameworks are there now for them. They don't have to build it from scratch, those industries, which have the AI prowess, can consume it through a library service rather than developing it themselves.

Over time, I think both data center and edge use cases will be much larger than hyperscale as the world industries consume more of a footprint and learn to adopt

this technology. And you're obviously seeing it being applied everywhere. In the end, it will be a choice of how they want to consume it, whether it's in the cloud, in their own managed data center or push it to the edge based on the problem or use cases.

My job is basically to activate all 3 of those and let the customers figure and choose amongst themselves based on the use cases. But we're certainly seeing early adoption now in the vertical space, certainly, manufacturing, transportation, health care, retail, financial services, certainly.

And early adopter companies like BMW or GE, Walmart, our American Express, for example, are figuring out how to apply this technology. And we certainly get involved and engage with them. One of the fun parts about being NVIDIA is that we're the only company -- we're the only AI company that works with every other AI company.

So we are able to learn from and engage on Facebook at the same time on Amex and be able to bridge the technology and the capability from a recommender system that may be used for a social media site to looking for fraud in a credit card transaction.

While those are different use cases, the underlying system is a recommender system. It's trying to understand from a litany of unstructured data what the right choices are or what anomalies might be in the system.

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. So that was sort of my next question, Ian, the traction you're seeing on the enterprise side. As you mentioned, your business is about 50-50 roughly today, health care, financial services, manufacturing. You listed a couple of verticals where you're seeing traction.

But could you shed some light on how you're thinking about the enterprise market, medium to long term relative to hyperscale? It's probably hard to put an exact number on it, but how do you think about the relative growth profile of enterprise vis-à-vis hyperscale? And how do you think about the adoption cycle in enterprise relative to the adoption cycle [ph].

A - Ian Buck {BIO 18454865 <GO>}

The adoption of our technologies, there's a couple -- there's 2 camps. I think we've always been -- had a presence in the HPC side. Of course, oil and gas industry needs supercomputers with the seismic processing. We connect that market well and continue to do so. Likewise, for in the simulation space, as you can imagine.

I think AI adoption in enterprise is still fairly early, early days. So I expect that to grow significantly. The challenge is how can they consume it and adopt it and what's the right products for them to consume and adopt it and making it a little bit easier, then you can't just give them TensorFlow and hope that they can train their model.

In the end, we're focused on more vertical SDKs and solutions and stacks. That's one of the focuses around Jarvis, our conversational AI platform, providing them with a complete end-to-end ASR, NLU, TTS pipeline, which has been pretrained on all the data that we have in our own fleet of DGX systems. Enterprise customers get something which can do transcription out of the box.

They may have to do some last-mile training, but that's -- we even provide frameworks for doing that. So they can tune their -- they're a generically trained model, which can understand a phone conversation and then augment it with domain-specific information. If I'm ordering prescriptions, it teaches it to have the prescription names so they can recognize that.

But you're already teaching someone who is already well understood. It even knows how to speak English and knows how to have a conversation. Where in AI, if you start from the beginning, you're truly starting from a brand-new baby. They know nothing. And it's a huge engineering test to bring them up to that speed. You see this in models like BERT. They offer fine-tuning as a capability, so you train to a certain level of intelligence and you fine-tune for the specialization.

Those stacks don't really exist and are starting to exist. So that's the focus of our driver stack is to provide that for conversational AI.

Similarly, Merlin for recommender systems and Maxine for UCaaS, this kind of conversation collaboration platform, providing the -- what's needed for enterprise to adopt is more vertical base-focused SDKs and they need to be bought and sold like a product and managed and maintained, which is a little different go-to-market, a little different engagement than what you might see in a hyperscaler, which is obviously more technology-focused and developer to developer, which we've done pretty well at that, too. I have full confidence we know how to do this.

We certainly do this in different markets today. A lot of our Quadro rendering business was done this way. A lot of the early AI adoption that we've been doing, we have seen great success with Triton and some of the inference software that we've been building. So I'm excited to see that come to fruition this year and moving on.

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. Ian, I wanted to ask you about inference. A couple of years ago, one of the more common reactions from investors were: A, NVIDIA dominates the training market, but they don't have a plain inference and it's mostly CPU based and so on and so forth.

And here we are, obviously, you've had a ton of success with the T4 and now the A100. And I guess recently, you noted that the aggregate NVIDIA GPU compute capacity available for inference on the cloud has now exceeded that of CPUs. What's been sort of not the secret sauce, but what's been the big driver there in terms of how you've done so well in inference to date?

A - Ian Buck {BIO 18454865 <GO>}

It starts with the model complexity. Some of the -- obviously, training is orders of magnitude more complex than inference. When you train a model, you do an inference, but the back propagation is where a lot of the competition is and obviously requires where we started. But for many of those early AI models in the computer vision space and others, a CPU infrastructure was certainly suitable.

It can execute the forward pass of an AI in a reasonable amount of latency and it's also what people had, so that's the existing software. They simply call a user framework and just kick off the inference and that is capable. What's changed is the model complexity.

Model complexity, both in the models, the traditional use cases and the new use cases. At CV, when computer vision went to NASCAR CNM [ph], it got harder, where now you cannot only just put bounding boxes around people or things or objects but actually identify each and every pixel. Doing that, attempting to do that at real time with high throughput became prohibitively expensive.

NLP is the other driver. So BERT model is an example. Also DLRM, DLRM was the deep learning recommender model, which is -- basically, it's a reference recommended model from Facebook, which they submitted to MLPerf. It's much harder to execute in real-time with the latency requirements you might have for a recommender system or a language system.

As we're having a conversation, I can't take seconds to do inference. I have to respond in under 100 milliseconds. And once you take out all the network traffic load, then my time slot for actually doing inference is quite small in a real-time interaction use case. And a recommender system is even worse because you have to actually run that recommender on thousands of products and get the click and the answer immediately.

So that is driving some of the growth that we see in T4. We had record revenues of T4 and shipments in Q3. There is -- you can get T4, which is our inference-focused GPU. It's priced differently. It's a different product that's focused on inference from all the hyperscalers. And you can see that's driving a lot of the strength.

The other part is software. It doesn't -- the software problem hasn't gone away. Delivering the best performance on these problems isn't just having a chip with a lot of flops but having the algorithms and the optimizations and the software stack to execute them at the right precision and maintain the accuracy and continue to provide the high throughput.

We've been investing in software called TensorRT, which is sort of a deep learning inference compiler. It takes the model which you trained from TensorFlow or PyTorch or whatever and compiles it down to reduce precision, FP16 or an in date, and runs it in real-time.

On top of that, we have found that there have been challenges with deploying inference. And so we've made that also easier in a Kubernetes environment with our Triton software, which is now is like a turnkey inference engine.

You give it a -- throw it a -- you fire up Triton across a Kubernetes home chart, you just need to send it the data and it gives you the inference back. A lot of our customers are trying to optimize their inference infrastructure with Triton.

It's been super helpful to help people deploy it. It runs on CPUs, runs on GPUs, it supports all the different -- all the major AI frameworks and even includes some of them. So it's been a huge uplift and help in our business to make it easier for people to deploy inference.

Instead of trying to shoehorn it into an existing application, they can just call out to a micro-service. So again, software is a huge part of that. Training is one job. Inference is entirely different other where latency, throughput, accuracy, quantization, graph compilers and fusion matter a lot, and then diversity.

And you got to work with the Kubernetes infrastructure, a Prometheus monitoring system, all this other stuff has to work in order to get this stuff into production. It's a lot of software.

We've been investing in that for a couple of years now and just thrilled to see it now finally take off. Great use cases. Amex fraud detection is one. Walmart is using it for inventory management, even Microsoft Office. Your grammar correction right now is being done through a cloud -- in the cloud in Azure, checking your grammar on a GPU.

A - Toshiya Hari {BIO 6770302 <GO>}

Yes. Just a follow-up, Ian, in terms of how you think about growth in model complexity. I feel like if you and your customers manage to perfect things like conversational AI recommender systems, that's pretty darn complex. But is it fair to sort of extrapolate the current slope in terms of the rate at which models are becoming complex into the future? Or could there even be an acceleration? Or how do you think about that?

A - Ian Buck {BIO 18454865 <GO>}

So both male and [ph] recommender systems are still open and they haven't tapped out as in the curve is not slowing down. GPT-3 from OpenAI approved it. They can continue to go up on the right. And we've all seen the charts of the model complexity. No one has shown the end of language.

And no, these neural networks are not close to human level of neurons, if you just do the naive parameter neuron basis. So there's still clearly more -- at least one existence proof that we've got another one or 2 orders of magnitude to go.

And by the way, training at that scale gets really complicated to get net -- to have that whole thing learn and converge. So I expect as that to continue, as you do the application of NLU to the different use cases, there will be specialization of those networks. It's not as simple as what people want to extract from language varies compared to like images, where it's image in and recognition out, box -- bounding box out.

This is much more complicated. Language in, sentiment out. Or search, I want to find data or information or structured data from -- imagine if I can take the New York Times, and read The New York Times and out comes a fully structured table of the information that could be used in search for.

Cybersecurity, another example, like every network log and be able to build and extract structured data that I can then do and look for intrusion detection.

So the application space will go get really broad in the NLU. The exact same story plays out in recommender systems. That's even more complicated just because the data in and data out tends to be so unstructured as well. And the application use cases are all widely different.

So that's what makes AI super exciting is the application space keeps becoming broader, diverging, if you will. And as a result, the variety of different neural networks and a platform it needs to run gets more complicated. I think it's one of the reasons why we invest so much in the different vertical use cases to get that experience and bring it back into our core platforms and figure out where to optimize.

So I don't think -- the network complexity is not going to slow down. People continue to get to human levels of language understanding and intelligence. But more interestingly, I think the application use case and the applied use cases will continue to expand.

So the diversity of models, that Cambrian explosion, that big bang of AI is just going to get more and more exciting as people apply it to a variety of different use cases. Speech, too, by the way, creating humans -- reliable human speech. We've applied AI to that problem, but we are not done. Many of those use cases are still -- and turning a good AI speech use case is still being developed.

A - Toshiya Hari {BIO 6770302 <GO>}

It's fascinating stuff. Shifting gears a little bit. Wanted to ask you about the road map. As you mentioned earlier, the rate of innovation and in GPU accelerated computing has been staggering.

I think you spoke to a 20x improvement with the A100. Based on what you're working on today and the visibility you have internally into your road map, how would you characterize the sustainability of your technology cadence? And outside of 10 going to 7-nanometer and 7 going to 5, what are some of the levers that you

have that you can pull, both on the hardware side and the software side to maintain that cadence?

A - Ian Buck {BIO 18454865 <GO>}

Yes. A huge part of what we do is the full stack optimization. While going through new node technology changes some of the parameters, it's obviously good from a baseline. As you go to different new technologies, your efficiency or performance at a particular power level increases. Also, where you can run processors at different power levels can change and shift over time.

That's great. And it gives you new options in transistor performance. Where the huge product performance comes though is in the software and algorithms. We improved Volta's performance.

From the day we launched it out of GTC, the day we launched A100 three years later on a different GTC by 4x, that came through innovations in algorithms, compilers, working with the community to improve the overall holistic platform, running on the same V100 GPU or that same server.

So that's what -- when we are -- and as a result, we try to crank out our software and our optimization as fast as we can. I publish a new framework container every month, whether it be TensorFlow or PyTorch, to help improve and accelerate training. We look at bottlenecks of the training pipeline and see how we can make them faster.

As we make the core flops faster and the matrix multiples the layers, a lot of the I/O pipeline starts to show up. Amdahl's Law is a (expletive) and it shows up in a lot of places. And that front-end and back-end can burn you. So we've actually had to work and optimize those use cases.

We invented a library called DALI, which does all of the preprocessing and image processing for cropping and angling and preparing the data so it can be trained on a GPU because a CPU just couldn't keep up.

The same applies to audio and other data processing. NVTabular, a library for processing and structuring a lot of the data, used to be all done in a CPU. We now do it on the GPU, so that we can put it right there next to the framework, so the data can be optimized to move quickly.

So that's a huge part of what makes our platforms more successful is -- or generation over generation. And we don't wait for new hardware to release it. We're just constantly pumping out that technology.

It all comes together holistically when you look at the top to the bottom. So up here, you have the software, the algorithms, the domain-specific solutions that make things a lot faster. You then, of course, have to CUDA the compilers and the run times which, of course, work with the operating system to optimize all of the low level optimizations.

That, of course, goes to a GPU itself, which has been improved and redesigned throughout the old ISA, bring in a new one. We designed the core SM architecture as we see trends and shifts in both HPC and AI. But then it scales out again, it's -- as we look at this data center in itself.

We don't constrain ourselves to one GPU or a PCIe card. We broke the mold and turned the GPU on its side, turned it into a mezzanine product and built our HGX baseboard, which goes into our DGX platform, which goes -- the same HGX goes into all the hyperscalers and the OEMs. We then take a step further and we look at the interconnect.

And with the acquisition of Mellanox, we can now look at building data center scales, supercomputers, that work as one throughout the -- that solve some of these AI challenges -- help define the future of AI. So it is both a software stack and a data center scale optimization that's being done generation over generation.

And you kind of need to think that way. AI is getting too big for one GPU or one processor. It's not about a -- can you run this layer fast? It's can you apply all that technology, develop those next-generation AIs for those workloads because the science and the data sciences are not feeling constrained by that.

They want to push the limits. And if we can provide them ways of training at scale, like we've demonstrated, they will take advantage of it to develop the next-generation AI.

A - Toshiya Hari {BIO 6770302 <GO>}

Got it. I mean, you touched on Mellanox a little bit. Can you speak to the DPU opportunity that you see over the next couple of years?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I think it's super exciting. So with the acquisition of Mellanox, we can look at doing what we did for computing to the network and data center scale networking space. The NVIDIA BlueField, for example, and taking networking and building a DPU programmable data center on chip opens up another $10 billion TAM for NVIDIA.

DPUs with DOCA, which is our counterpart to CUDA for that platform, can help re-architect that modern data center to do the -- to optimize the data center scale networking.

This is an area of security or in network compute or being able to make our data centers malleable, turn a supercomputer into a cloud resource, all capable through by once you insert a programmable networking platform that can ensure security and isolation and the right levels of offload do not impact performance.

So that's some of the problems it's trying to solve, and we're early on in that journey. Certainly, we worked -- I worked with Mellanox for a little over a decade in the HPC space.

The cultures are very compatible. Very engineering driven, very technical. It's great to see them at NVIDIA. And certainly getting -- we can bring some of our platform strength and our capabilities that we've talked about from the full stack to that market and really help revolutionize it.

A - Toshiya Hari {BIO 6770302 <GO>}

Okay. And then with Arm, it's been about four months since you announced the acquisition. What's been the early feedback from your customers from the broader ecosystem? And remind us how Arm fits into your overall data center strategy.

A - Ian Buck {BIO 18454865 <GO>}

Sure, sure. Arm is also very exciting and certainly where I'm spending a lot of my time right now. A lot of interest in Arm, the talent. Our strategy from an AI -- from a company standpoint, I think we've made clear, is to create the premier computing company for the age of AI.

We can combine a lot of our NVIDIA's AI leading computing platform with Arm's vast ecosystem and help really move them all forward as we do it together and helps position for the next wave of computing in the age of AI.

And it's happening at the data center, it's happening at the desktop and, of course, it's happening at the edge and in the world of IoT. We can help expand Arm's IP licensing portfolio with NVIDIA's technology to meet some of those large markets, including mobiles and PCs.

We also -- given our background in data center and server, we can help advance -- turbocharge Arm's CPU adoption [ph] pace into the data center. Of course, there's a lot of interest in that with the work at Amazon and Graviton. Of course, in the consumer side, you see what Apple's doing with M1, it's a great time for Arm. We're certainly seeing it everywhere.

Customer's feedback, overall positive. The deal would not affect any way of customers getting access to Arm's technology, so they like that. We're fully committed to Arm's existing licensing model, preserve Arm's customer neutrality in that regard. In general, NVIDIA has been an open company.

We work with every major CPU provider, both x86, did a lot of work with IBM POWER and, of course, Arm even before the acquisition that was announced. Like I said, we're the only AI company that works with every other AI company. So my job is to activate our technology in all those places and help get it as we help ride and accelerate that tide of AI.

We have high confidence the deal will close. But in the long term, we can -- we treat Arm like a first-class citizen as we did before the acquisition, and we certainly already released our software stacks on Arm, and we're excited to see it come to market.

A - Toshiya Hari {BIO 6770302 <GO>}

Ian, I can't believe we're out of time. But before we let you go, I just wanted to ask you, I realize you probably don't spend too much time with the finance community. Maybe, you do. But to the extent you have a view, what is sort of the finance world missing about your prospects in data center or the world more broadly? What are we missing or underestimating about the story?

A - Ian Buck {BIO 18454865 <GO>}

Well, first off, it's hard to comprehend the growth of AI. I don't think none of us, unless you were around the original PC revolution, kind of experienced something like this before and what does it mean for models to be doubling every 2 to three months and what does that mean, not just for NVIDIA, but just in general, in all the different markets.

The people that figure this out first are going to really make waves into what they can do and how they impact their enterprise, their problems, their customers. It is hard.

That is also -- I think it's hard to put a number on that or put it on a spec sheet or say I've got a great product. I've lived this -- I've been doing it since AI -- since Oscar Dusky [ph] first did that work up in Montreal, and I worked with you on the core overriding [ph] work. But as soon as we did that Torch 7, where we started turning on GPUs, it's really hard.

And the complexity of the software stack is intense and the value of full stack innovation, so I think it's really important as you see different things happening in the industry, understand their software and how they expect to take it to market and where -- what they're doing.

It's -- that is a huge part of what we do, and it's hard to quantify other than the amount of constant optimizations and work that we do to move the ball forward in the software stack from the underlying hardware capability.

That said, and that's what the full stack innovation sort of business model, engineering approach has enabled us to do, and it keeps it super fun and that's moving so quickly. But don't underestimate the cost and energy it takes to bring this to -- to get this stuff done. So -- but it keeps me from -- it's very fun, and I get to learn about every different use case of AI. I wouldn't wish for any other job.

A - Toshiya Hari {BIO 6770302 <GO>}

Ian, that's a great place to end. Ian, thank you again for the precious time. Thank you for all the insights. Also a big thank you to all the investors that joined us this afternoon.

A - Ian Buck {BIO 18454865 <GO>}

Thank you. Good luck.

A - Toshiya Hari {BIO 6770302 <GO>}

Meanwhile everyone, thank you.

A - Ian Buck {BIO 18454865 <GO>}

Bye.