Wells Fargo TMT Conference

Company Participants

- Ian Buck, Vice President, General Manager of Datacenter GPU
- Simona Jankowski, Vice President, Investor Relations

Other Participants

- Aaron Rakers, Wells Fargo

Presentation

Aaron Rakers {BIO 6649630 <GO>}

Perfect. Thank you, and good afternoon or good morning, everybody, from wherever you are. I'm Aaron Rakers. I'm the IT hardware and semiconductor analyst here at Wells Fargo. I think this is the first time we get to host a meeting at Wells Fargo with NVIDIA. So extremely excited. Obviously, company has been extremely successful, and looking forward to the conversation. Before we jump to Ian Buck, the Vice President, General Manager of the Tesla Data Center business, I want to hand it over real quickly to Simona Jankowski to kind of go through a one-sentence disclaimer.

Simona Jankowski {BIO 7131672 <GO>}

Thanks, Aaron. 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.

Aaron Rakers {BIO 6649630 <GO>}

Perfect. Thank you, Simona. So Ian, my pleasure to host this discussion with you. Maybe just to kick it off, for those of you -- those in the audience that aren't familiar with yourself and kind of your role, maybe you can just give a real quick background of your responsibilities in NVIDIA, and then we can jump right into some questions.

Ian Buck {BIO 18454865 <GO>}

Yes. Sure. Hi. So I am the Vice President and General Manager of our Data Center GPU business, which focuses on all the GPUs that go into servers, either in the cloud or on-prem and on-prem data centers. Specifically, I focus on the hyperscale markets and AI as well as HPC and scientific computing and server [ph] computing as a result. My team focuses on the product lineup, bringing those new products to market for GPUs, also a lot of the software that goes into making that platform successful, both in AI as well as more generally accelerated computing.

Questions And Answers

Q - Aaron Rakers {BIO 6649630 <GO>}

(Question And Answer)

Yes. That's perfect. And I mentioned before we started the presentation, I watched you at the Supercomputing Conference a week or two ago, go through a ton of details on AI and kind of the strategy around it. But where I wanted to start the discussion was kind of the latter point you brought up, the software ecosystem, the importance of that. And I think you were clearly a visionary on CUDA and the importance of that CUDA ecosystem.

How do you see the landscape evolving around the importance of the software ecosystem, the stickiness of the CUDA platform? And I asked that a little bit in the context that we've seen others really validated. We've got Intel moving with oneAPI. We've got AMD with ROCm and Xilinx with Vitis and so on and so forth. So let's start there. Let's just talk about the stickiness of what you see at CUDA and the importance of that.

A - Ian Buck {BIO 18454865 <GO>}

Yes. I mean this all started probably over 15 years ago, when we recognized that there's a new kind of computing that was happening in the area. And it started in computer graphics, which really was a simulation of light in the environment around us. But it was a different kind of computing. We were focused -- back then, NVIDIA was focused on accelerating computer graphics, literally writing, encouraging developers, giving them a platform to write a program that determine the color of every pixel.

If you think about processing at the graphics or at the pixel level, that's 10,000 to 1 million programs all being run in parallel to generate an image. Folks like myself and others recognize that that computing horsepower could be used for other things, not just video games. And back in 2004, we started CUDA to build that platform for accelerated computing, taking the portions of applications, whether they are in HPC or anywhere, and accelerating those inner loops, those important computational portions and to get them on this new kind of platforms, new kind of architecture, this massive parallel processor.

It's been a long journey, 15 years of developing that sort of full-stack, accelerated computing platform that we are today. We're probably the most complete advanced and robust platform for it. It's consists of tens of thousands of engineering person years to create it. We're now on our 11th generation of CUDA. And it's not just the way of programming GPUs at this point, it's that entire platform.

We have an entire platform of libraries, which we call CUDA-X ranging from signal processing to numerics to linear algebra, of course, to AI, to machine learning, to computer vision and image processing as well as audio and other things. And in fact,

it's expanding now to even vertical workloads in conversational AI, with technologies like Jarvis, recommender systems with systems like Merlin or video collaboration platforms like we're doing now with Maxine.

Only through building up that platform have we made it successful and productive to so many people. At this point, we have over, I think, 18,000 -- 1,800 applications in our application catalog that have been accelerated and over two million developers developing on the platform. So I think the -- and that talks to the productivity of what accelerated computing, what NVIDIA GPUs and what CUDA has brought to that community. It just takes that -- those applications and make them do something that they've never even thought possible before because of the computing horsepower that GPUs offer.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. And one of the common questions I've gotten from time to time, and it seems to ebb and flow the stickiness of that. You hear around TensorFlow and PyTorch, and hey, we write directly on that. I mean, I guess -- I think, you've answered the question a little bit in the prior answer, which is it's much more than just an application layer that you write into. It's become verticalized. There's a library depth of it. Do you see anything that's competitively changing in the context of CUDA and that stickiness?

A - Ian Buck {BIO 18454865 <GO>}

Well, one of the things that made it really successful and broadly adopted was its availability. Very early on, we decided as a critical decision, a pivotal decision to make CUDA enabled and available in every GPU and video produced from then on. As a result, I think we've shipped over 1 billion CUDA-capable GPUs.

And we recognized early, even though we didn't make it bespoke to one platform or one product line, I mean, every product line. And that's because that innovation is happening all over the world in -- from research labs to dorm rooms. In fact, AI, Alex Krizhevsky, who did sort of the -- wrote the -- literally the first AI framework, cuda-convnet, did it as a student up in Canada, as he saw the stuff that we were doing at HPC, and he had a gaming GeForce GPU, and he ported it on to it.

So I think the access and availability of the platform, literally being able to consume and be available in every cloud, every OEM, every data center, every gaming laptop creates that innovation, makes it possible for us to accelerate so many different workloads. Because people see this opportunity and figure out how they apply it to themselves, either programming GPUs directly or using one of the libraries or going to vertical stacks or using the AI frameworks.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Yes.

A - Ian Buck {BIO 18454865 <GO>}

I think that is unmatched. And I think that is something that is truly unique in the marketplace. And because of its wide availability and the breadth of innovation has now happened, not just that NVIDIA provides, but now all the other people and players innovating on top of that platform that makes it so successful, as you say.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. That's great. And now we're at this point in time where we're tiptoeing towards another layer of accelerated compute, right? There's this thing that NVIDIA has driven and a few others, right, around data processing units/SmartNICs. And I guess, maybe it's a little bit lost on some people is that you're building a similar kind of software ecosystem stack on top of that DPU strategy, the Mellanox, BlueField. So can you talk a little bit about DOCA, the importance of DOCA and how you envision the role of DPUs evolving in a data center?

A - Ian Buck {BIO 18454865 <GO>}

Yes. So before I explain what it is, there's clearly a trend happening in data center in general. And it's pivoting around the use of the programmability that we see in the data center, where it used to be you programmed a single-instance VM or one -- the CPU, and that was it. And today, the data center itself is being redefined. It's becoming programmable at many different layers. We now program entire helm charts of concatenated many microservices together to build data center cloud application, running across many different nodes, many different instance types.

In addition, we program with GPUs and CPUs. So we have two kinds of computing platforms inside of the system. And CPU obviously continues to provide a really important role in that environment. And additionally, we're now connecting the GPU -- these nodes together in unique and fast and interesting ways with the NIC. And the NIC is becoming a critical part of that data center and also a more programmable part. Look at what people are doing by offloading some of the security operations, some of the virtualization stacks, moving it closer to the network, where the node connects the rest of the data center to offload a lot of that computation or workload or latency operations.

So one thing that we recognize is now, instead of programming just a CPU, the entire data center is becoming one programmable unit. It's the thing that people program, it's all aspects of that, whether it's connecting it up with Kubernetes containers by moving the different parts of the applications to the accelerators or accumulate it on the CPU as well as offloading many of the operational systems, security, virtualization tasks to the NIC and the networking platform. And even going a step further, we're starting to see computation happening in network where it -- just because of the way the traffic is flowing, it's the optimal place to do it may be in switch, for example. This is particularly popular in HPC.

So one of the things that we're doing with the recent acquisition of Mellanox is to accelerate that -- the networking stack, being able to program and that being a programmable part of the entire data center. Our solution for that is DOCA, data center on a chip. And it provides a programming platform for our BlueField-accelerated networking adapters. That allows people to do that kind of -- to program

and develop for that level of offload, running security on the chip as well as offering a virtualization offload onto that processor as well as take advantage of the many different accelerators that we have inside of our BlueField interconnect NICs [ph] to accelerate data center operations.

So we see that same opportunity that we saw in accelerated computing back when we started with CUDA, with DOCA but now programming at the data center scale for optimizing data center communications and data center applications.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Yes. That's awesome. Let me take a little bit of a step back, when you think about the data center business in total and you think about the TAM, the addressability of NVIDIA's opportunity, right? I think you laid out $50 billion going to $100 billion. What underpins that? How do you think about the attach rate of GPUs? How do you think about the evolution of DPUs as far as attach rate evolving when you're talking about customers? What are you seeing happening on that front? Because clearly, NVIDIA's data center business is far outpunching the server growth profile that we've seen just most recently in this last quarter.

A - Ian Buck {BIO 18454865 <GO>}

Yes. I mean the server and data center market, obviously, is a gigantic market in total. And actually, we're a fairly small part of that right now. I think what is causing the nonlinearity is the adoption of applications and use cases that benefit from GPUs and accelerated computing. Most notably, it started in HPC in simulation, which is a smaller market, but an important one. And they recognized it first, because they saw the end of Moore's Law before many of you guys thought and were talking about it, realized they needed an alternative path.

What's caused the growth, obviously, recently is the adoption of AI and more broadly, machine learning. The benefits that GPUs break from machine learning are dramatic. And often 1,000 CPU servers can be replaced with a single DGX-based like system. And that's from not just performance, but also cost, huge TCO savings, 20x in some cases. And it's because that processor or the GPU is designed to do those kinds of applications, not be a serial processor of how fast you can execute one thread of execution, but how fast you can execute 10,000 different programs operating on the entire data set.

AI works by data, ML works on data, and it's designed to do that kind of processing. And we've been continually optimizing it along the way since we first started down this path. With AI it was the Kepler generation. We're now with Pascal, Volta, Turing and now Ampere, continue to optimize that from an architecture standpoint. So what's driving a lot of that growth is the adoption of those techniques for different markets. It obviously started in the hyperscalers, where they had the talent to invent the technology from the ground up, folks like Google and Amazon and Microsoft and Alibaba and Baidu, Facebook and then the others, obviously, have the AI talent themselves.

What we're seeing now is them obviously take all that to market and continue to see the opportunities for AI and grow the use case internally, but also the rest of the cloud. They all offer public cloud offerings of GPUs. And now we're seeing all of the consumer Internet companies, the Snapchats and Pinterests and Zooms and Zillows of the world consume and benefit from AI to improve their recommenders, their conversational AI agents, their data processing.

And the next wave, I think, is the growth of enterprise in general, bringing more services and solutions to the enterprise. And they can take advantage of AI, where they don't necessarily have the AI brain trust that a Google might, but AI is maturing as a technology and machine learning in general, to see the benefits of GPUs and being served, either programming directly using -- understanding and using TensorFlow or Triton for inference serving or using some of the more managed services for delivering those capabilities.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. That's -- and how -- I mean do you have any thoughts around attach rate or the progression of if it's GPU attach rates? I think also on the DPU side, I mean, Jensen has been pretty clear, like, could we be evolving down a path where almost every or majority -- the majority of servers actually incorporate some form of DPU or SmartNICs?

A - Ian Buck {BIO 18454865 <GO>}

Yes. I think it's early on the DPU. It's early on DPU, but it -- in terms of today's attach rate. However, we believe every server should have a DPU in it, simply to do the offload of the virtualization and security and management tasks that logically should be happening, not unnecessarily on the CPU, but could be offloaded to the networking -- to the network and the rest of the system, leaving the rest of the CPU available to the end application and the user for their performance, what it was designed to do.

You saw similar experiences with AWS Nitro. They moved the infrastructure support to a separate processor that lets them free up all the CPU cores and CPU capability for their end customers. And they can manage the infrastructure themselves. The same is true for the vast majority of the market, and we believe DOCA and BlueField can do that.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Yes. Talk a little bit more about the system strategy. I mean we've seen NVIDIA with Selene show up on the top 500 list of supercomputers. The company has clearly pushed deeper with DGX, I think, recently announcing the DGX Station A100. So I guess, in your responsibility, how deep does NVIDIA go down the path of a full system strategy business?

A - Ian Buck {BIO 18454865 <GO>}

I'm glad you asked that question. A lot of people get confused about why NVIDIA is also -- is being a system provider. I think it's important to recognize, like I said before,

the data center is being reinvented now. And we've -- the business and the opportunity for accelerated computing has gotten large enough where we can break traditional molds or boxes, if you will, for where we can design in and really think of the entire data center as our design, our envisioned canvass [ph].

We did that earlier on by breaking out of the PCIe form factor, building custom baseboards and solutions that were better optimized to allow our GPUs to talk more efficiently with each other by exploring and leaning into NVLink-connected GPUs and potentially NVLink-connected nodes and systems as well as the work we're doing on InfiniBand to do data center scale interconnects.

So that innovation space is huge now and requires a lot of investment, a lot of invention. And it's difficult for traditional OEMs to invest the kind of money and capability in that leading-edge capability that starts with a $0 billion market. It costs -- it takes a significant R&D to think at that scale and invent at that scale. So we take on that burden ourselves now for two reasons. One, we are our own users. So we have our own self-driving car initiatives, own products, our DRIVE platform. We have our own research teams that are finding different ways to apply or innovate on AI in general for our own products as well as for general-purpose research.

And so our Saturn V infrastructure is used for that. So we optimize it for ourselves first. Then we take those designs and those innovations, and we package them up into products and bring them to market so that we can meet the tip of the spear for the rest of the market. Our DGX platforms, our DGX SuperPODs, which Selene is based on, provides that blueprint, that template for the rest of the market to follow and clears the way.

We bring it to market completely. So if customers do want to buy it directly and have that direct relationship with NVIDIA, they can. They're not configurable. This is the lighthouse juggernaut offering that helps pave the way for the market. And then we take different components, and we make them available to all our partners. So we take that same baseboard that's inside DGX or inside of Selene. We give it to HP, we give it to Dell, we give it to Inspur and the rest of the market. We give it to their cloud providers and it's the same baseboard that is available in Microsoft and Amazon and Dell and HP. And our software stack has well been optimized on that. So everyone gets the benefit of it.

So people could consume our technology however they see fit. My -- I'm -- our goal is to make it available everywhere, whether it be in the cloud, to rent and the server to buy or any which way. We do that system business both for ourselves to help lead the market, but more importantly, to accelerate the innovation in the data center space.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. And one question that got e-mailed to me, and I just kind of put it out there. I mean oftentimes, we're asked a question around the cloud's -- the hyperscale cloud internal development efforts as a competitor in the realm of AI. And we've seen what

AWS has done with the Annapurna assets and so on and so forth. So how do you see the competitive landscape vis-a-vis the cloud customers evolve?

A - Ian Buck {BIO 18454865 <GO>}

Sure. I think AI, in general, obviously, is a once-in-a-lifetime revolution in computing in general. I think the world has been able to achieve with AI has far outpaced even science fiction just 20 years ago. And so it makes sense that everyone obviously is investing in their own AI solutions and capabilities. And the clouds themselves have to compete. They're going to not only provide a channel for all the different options out there and capabilities, like you have seen, but also try to attempt to differentiate from each other. And we've seen that as well.

I think NVIDIA is the only AI company that works with every AI company. And by doing so, we accelerate all the different workloads and different capabilities and helps and improves and advance our platform and is able to innovate at a rate that is unmatched by any other. And I think you see that specifically in the results of MLPerf, the benchmark that was created by Google, to help separate the wheat from the chaff. There's a lot of claims out there, a lot of start-ups, people saying that they have more FLOPs and more capabilities. The proof of the pudding is in the eating, and that comes in the form of MLPerf, where they set standard guidelines for performance and accuracy.

As you see from our results, we've been able to outperform competitive hardware or just some of the start-ups to still show up. So I think that is an important aspect. Everyone is going to try to differentiate or specialize from each other in some ways. And that's fine. From a NVIDIA standpoint, our standpoint, our strategy doesn't change. It's to make sure we continue to work with every AI company, every AI customer. And that gives us the opportunity, the visibility and the capability to continue to move our platform forward.

Just going from Volta to Ampere, delivered over 20x more performance in AI, and that's what just over three years. So that comes from that innovation, that engagement with all those different customers, serving all those different needs as well as looking inward to what we can do in our architecture, and more importantly, in our software, to continue to advance AI for everyone to democratize it everywhere and keeps us on our toes. It makes my job super-interesting. I'm constantly learning about new applications of AIs and how to make them faster and have to optimize, both in the hyperscalers as well as some of the new emerging workloads in the enterprise.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Next question on -- to the extent, you can kind of go in any detail you can share on the Arm acquisition, the data center piece of Arm. We're still early in it, right, as Arm truly a competitor in the CPU side of the market. How do you think about the strategic merits of owning Arm? And whether or not it's all about kind of the better together, hey, I'm building a system stack strategy, we're optimizing data movement across CPU, GPU, memory interconnects, et cetera, et cetera. What -- how important is Arm to what you're doing on the data center side?

A - Ian Buck {BIO 18454865 <GO>}

Well, I think Arm is super-exciting. I mean, I think is there's announcements today. You're seeing people innovate on Arm across the board. Nations are investing in Arm supercomputers for their computing platform for doing COVID research. Clouds are investing in delivering Arm capabilities to think of new ways to improve cost performance or application performance. It provides a rich innovation platform, people to innovate themselves, whether they leverage Arm IP themselves or design their own. Super-exciting.

Arm is probably the most ubiquitous CPU platform in the world. I'm surrounded by Arm devices right now, for sure. I think combining it with NVIDIA as a leading AI technology platform, yes, there is -- better together. We can advance AI and that CPU ecosystem dramatically, both on the platforms that already serves today very well as well as the new opportunities in data center and cloud and computing. In a broader sense, the data center has three processors: the CPU, a DPU and a GPU. And by bringing all those three under one roof, we can innovate in new exciting ways.

So that acquisition is still ongoing. We're going to have to wait until that completes before we can talk more about what we're doing in that space, but that is the opportunity that we see in front of us. And it goes to the broader innovation that's happening in the data center as a whole.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes, I appreciate you answering that question. That was very helpful. I know it's probably a better question for maybe if I had Colette, but I'm just going to ask you and answer which way you can. The supply-demand balance or getting the supply relative to demand on what appears to be a very out-of-the-gate-successful A100 Ampere product cycle, how do you think about that? Or how are -- how is NVIDIA currently thinking about getting supply-demand in balance on the data center side at this point?

A - Ian Buck {BIO 18454865 <GO>}

I mean it's going to take several months to catch up some of the demand. I think the -- what's exciting is the sort of the interest and growth in both training and inference. We delivered a record number of quarter on T4, for example. And with the -- every time we introduce a new architecture, it's a game changer, right? So A100 is 20x better perf than V100, and with that comes a new wave of demand and interest in our products.

So we're meeting that demand. It's always an exciting time certainly for me to help bring all those and all those platforms to market, whether they're hyperscalers, who are just now bringing online all the OEMs as they're launching their products, as well as the rest of the market. So it will take several months to catch up with the demand, but it's always an exciting time to have on the platform refreshes that we experienced when you get the injection of new hardware. In the meantime, we continue to improve the software, too. So our value proposition doesn't stop when we launch a new GPU.

In fact, we don't even release the ISAs to our GPUs. It's all coming through the software that we're constantly updating. My team is responsible for delivering the new containers in TensorFlow and PyTorch from the base frameworks as well as the inferencing stack with Triton and Jarvis and Merlin and Maxine, and we're pumping out new versions of that every month. So it's a constantly moving platform, which makes it fun and exciting for all of our customers and, of course, operationally, to meet all the demand.

Q - Aaron Rakers {BIO 6649630 <GO>}

In the four minutes I have left because you hinted on it, T4 record quarter on the inferencing side this last quarter. But one thing with the A100 that you clearly emphasized is this convergence of training and inferencing. So outside of just T4, what has been the reaction? And what's been the -- is it harder for us to see what is inferencing versus training? And how successful has that kind of convergence been in your mind?

A - Ian Buck {BIO 18454865 <GO>}

Yes. It's going to get a little trickier just because, for those of you who may not know, this is our third-generation Tensor Core, Tensor Core being the core that is optimized for the AI operations. We started it in Volta architecture, which is focused on training. With the Turing architecture, we add -- we did a version for inference. And in Ampere, we combine the two, actually added in HPC as well. So we have FP16. That's intense cores [ph] capable of floating point calculations for HPC. We have FP16, Bfloat16 and TensorFloat-32 for training as well as FP16 and INT8 for inference all in its [ph] Tensor Core.

So what we're seeing is a lot of our customers can now take advantage of the performance and capabilities of an A100, both for training and for inference in the same instance in the same cloud. So it's hard for us to track how people are using our A100s in the cloud live. Obviously, it can vary. It can be used both for batch inference. So I reframe a model. We want to rerun all the inferences again that are going to use the same A100 as the most cost-effective and pro forma platform to do so. They can also use it for live inference and inference services. We've developed technologies like Triton, which allow multiple users, multiple models to be run in ensembles on that same one A100 GPU to optimize the infrastructure.

We're just rolling that out now. So -- but it's very exciting. If you compare the amount of inference capability that's now been shipped to the cloud in the form of GPUs, spanning the Volta, T4 and A100, was a huge boost in inference capability. We now have more total FLOPs available in the cloud for inference on GPUs than is available on CPUs for the first time. So I think that's an exciting tipping point. And A100 is a massively powerful inference GPU as well as training. So it's going to be a little bit tricker.

For sure, people are very excited about those training capabilities, and I expect that to be its first primary use case. And as more people adopt A100 for inference, we'll see some of those trends continue. I still expect T4 to run on. There's certainly a market for edge and that kind of server, which wants to accept one or two 70-watt

GPUs for doing more than just the inference capabilities. They obviously have work to do on the CPU as well. But that A100, particularly the A100 MIG capabilities, may change the game a lot. And the more new customers can learn to consolidate and optimize inference to using less infrastructure, the more they can save money and get the benefit of A100.

Q - Aaron Rakers {BIO 6649630 <GO>}

The final question I was just going to ask you, and this is just kind of a secular thematical kind of question around -- what really kind of got me extremely excited about the data center opportunity was this proliferation of complexity in these AI models, right, Google BERT [ph], NLG for Microsoft, et cetera. Does that slow down? Does this rate of expansion, parameter expansion, whatever you want to -- however we want to think about it in these AI models, do you see a part where it slows ever?

A - Ian Buck {BIO 18454865 <GO>}

Right now, it's a chicken and egg. It's -- they developed the model that the -- to the size of their infrastructure allows them to build, their customers are willing to buy. If it's a single GPU instances, a single A100 is all they have, that's what they'll design to. If they have a DGX-like system or SuperPOD to get bigger, and then you look at things like GPT-3 [ph] from OpenAI, which requires a massive number of GPUs to train. So that is the exciting part of this time.

The conversation -- where we're at with AI is the advancements of natural language processing, computer vision and recommender systems, particularly those last two, recommender systems and conversational AI. This is true human intelligence. This is not something as simple as what is this a picture of. Recognizing pictures is something conceptually humans do, but so can dogs and cats and even insects can recognize -- have some capability of recognizing what their vision system sees.

Language processing, being able to answer a question, understand not just the words I'm saying but what I meant and coming up with a sensible answer that because that helps me find the product that I want in the Internet or have that conversation with a bot that's constructive and helpful, that's really intelligence, human-level intelligence. And so far, no one has seen an endpoint to that number of parameters to solve that problem. There's no -- they're not even close to the number of neurons in our brains. And that will continue to grow for the foreseeable future until somebody figures out or gets to that point or builds the infrastructure that can get there.

A lot of the innovation right now is how to combine all those GPUs together, not just in a data center and electrically, mechanically and through computer technology, but also the algorithms and numerics to make it work as one. So it's a really hard problem. It's super-exciting. We get to work on it here at NVIDIA, but we also get to work with our partners in Microsoft, OpenAI and others to figure that out as well.

Q - Aaron Rakers {BIO 6649630 <GO>}

Awesome. Ian, we ran out of time. I could continue to go on, but I have to be -- I have to kind of jump. But I really appreciate it. Thank you so much for letting us host this conversation. Appreciate it.

A - Ian Buck {BIO 18454865 <GO>}

Sure. All right [ph].

Q - Aaron Rakers {BIO 6649630 <GO>}

Thanks. Thanks so much.