Company Participants

- Ian Buck, Vice President and General Manager, Accelerated Computing
- Simona Jankowski, Vice President of Investor Relations

Other Participants

- Harsh Kumar, Analyst, Piper Sandler

Presentation

Harsh Kumar {BIO 3235392 <GO>}

Hey, good afternoon everyone, thank you for joining us again on PSS Friday Series here. We have a pretty exciting lineup here. We've got Ian who had some -- you guys see him there. And I can tell you, there is no better person to talk about AI and Data Center compute than Ian right now in this world. And I'm going to turn it -- with that, I'm going to turn it over to Simona, she's got a couple of statements she wants to make and we'll go get started.

Simona Jankowski {BIO 7131672 <GO>}

Thanks very much, Harsh. Before we kick off the Q&A with Ian, I just wanted to remind everyone that we may make forward-looking statements as part of today's conversations. So investors are advised to look at the reports that we filed with the Securities and Exchange Commission for any of the risks and uncertainties that relate to our business. Back over to you guys.

Harsh Kumar {BIO 3235392 <GO>}

Thank you, Simona. So, Ian, I meant that very seriously, there's not a better person to talk about the trends in the data center business. You've been doing this for a while, you guys are at the top of the game when it comes to compute. Could you maybe talk about what you see as the primary role when you talk to your customers about data center today? And then perhaps what your customers are indicating or where that role is going to the next five, 10 even 15 years out?

Ian Buck {BIO 18454865 <GO>}

Yes, I mean one of the main drivers right now in the conversation is, now how to digitize or cloudify or accelerate different industries, whether it be in automotive and retail and social media and all things that we do engage with the Internet or engaging in retail. The biggest obvious driver for that was the prioritization and commercialization and enablement of AI for those use cases. And it's very challenging to AI as a bit -- is a scary word, it was pioneered by the hyperscalers

originally because of the fundamental research that had happened in order to activate it and obviously they took management first. What the thing that's trending now is the shift toward the commercialization of AI into the broader enterprises you on board, whether it be deciding how products are bought or recommended to you, we've all experienced improved ways of doing conversational AI, how we're talking to the cloud and our devices. And that's also extending more broadly into the broader area of data science. Now, with the era of -- now going from big data to big AI and big data science, we now have tools to actually understand our data much more intimately. And that's driving a whole new era of computing in our data centers, either in the cloud or on-prem and that's the opportunity we're seeing and driving.

NVIDIA while we do make the accelerators for AI and data science, we're now expanding in the conversations about how to accelerate the entire data center, what is the data center's future look like and needs to be built today to meet those credibly intense computing demand to achieve those business insights. So that's why you see NVIDIA as a company moving to becoming more of a data center company for both AI and data science and innovating not just the processor level or just the core software level, but of the data center as a whole.

Harsh Kumar {BIO 3235392 <GO>}

Okay. So that's pretty interesting, that gears towards some of the recent moves that you guys have made in terms of bringing in some critical pieces. So functionality is rising each year now more than ever with COVID. COVID just sort of put troubles [ph] on every aspect of the data center, I suppose. With functionality rising, what kind of stresses does that put if a company had a data center that's a year old, and all of a sudden you have a COVID economy, what kind of stresses does that put on a data center that's even just slightly old? And then how are you guys helping your customers overcome those stresses?

Ian Buck {BIO 18454865 <GO>}

Well, first off the -- it's driving an accelerated path toward more and more AI applications. And one of the challenges about AI is amazing technology is that people are discovering new ways of doing things and more intelligent networks, more intelligent AIs to give a better experience or to recommend better products when you click on things and of course increasing Ad revenue and user experiences which drives the economy of the cloud. The challenge is that AI is moving really quickly. If you look at state-of-the-art two years ago, three years ago, you have neural networks like ResNet 50 which allowed you to understand what's inside of an image.

The newer neural networks much -- have gone beyond just understanding images, understanding speech and language. And those things are much more higher level of understanding that they need to do and the neural networks results are much more intelligent and much larger. It's the true human intelligence to be able to understand the human language. Bugs, dogs, animals can recognize pictures. That's a very actually simple brain function. Understanding language is a much more intense one. As a result, those neural networks from like just a few years ago, are

literally 3000 times more computationally intensive. Layer like a ResNet 50 neural network, what it takes to train a network like that, to training some of the modern natural NLP natural language processing models like Megatron, which was recently published is a 3000x.

So that's what's driving a lot of the rapid innovation and rapid deployment of a latest possible technology to enable the data scientists to apply those networks to their particular problem to their particular domain. Whether it be understanding what's published on a web page and doing summarization, or if it's an automated call center for refilling a prescription or asking or understanding a financial call, and understanding the text and dialogue that's happening and be able to flag it across all the different information. That's natural language processing, and it's a much more intensive activity computationally and that's what is driving a lot of innovation today.

Both in the training side we have to develop these networks, there are now tools where people can take existing networks like Megatron, like Bert, and do transfer learning. So apply that pre-train network to their domain, it's getting much easier to apply AI technology, which is exciting. And so as a result, proliferating the need for AI data centers everywhere. There's the flip side of that, which is obviously the deployment side, you train and then you turn around and you deploy AI. The neural networks are getting big enough that they are requiring to be accelerated. In the past the early networks could still run on the legacy CPU based data centers, or CPU based edge devices or points of presence. Today, with the new neural natural language models, they just can't deliver the real time experience and run the neural network in real time. So we're seeing a big growth in accelerating at the edge, moving more of the -- with our edge accelerators, both at the edge of the data center, the points of presence and even in the embedded and telco use cases, where you want to put the acceleration as close as possible and provide the lowest possible latency and the best possible experience for the end user.

Harsh Kumar {BIO 3235392 <GO>}

So that's fascinating, it sounds like a lot more is to still come. Sounds like we've only gotten into the tip of what can happen with accelerated computing and AI. I want to talk about something you mentioned earlier, bringing an accelerating sort of the entire data center. So you guys recently acquired Mellanox, which was the bigger one, I think there was a software company acquired as well. So we know that Mellanox is computing at the surface -- sorry, connectivity at the surface. But Jensen was giving I think, a keynote for one of the events out of his kitchen and he talked about --- so that was fascinating to you. He talked about maybe offloading some things to perhaps parts and pieces of the Mellanox product set. Could you maybe expand on this thought process? Is this what you're referring to as sort of hyperdriving or sort of putting the turbo on the entire data center?

Ian Buck {BIO 18454865 <GO>}

Yes. The basic unit of compute is moving from a server to the entire data center. Now, people today when they think about building an AI data center, they don't

think about just building servers, well make sure they have GPUs in them to get great acceleration. The AI today is moving from a single server with multiple GPUs in it to training across the entire data center where you train a neural network a lot for language processing, for example, or for video understanding to understand video content. You typically train across many, many nodes, upwards of some largest neural networks we've trained have been up to over 2000 nodes across the entire data center, for a modern neural network.

The reason for that is it's time to train, we can't -- these data science's work is very important and these new neural networks need to complete in a certain amount of time, otherwise, it's just not productive for them to do their work. So today's data centers have to be multi-node and multi-node capable of doing AI training. And that's one of the important abilities that Mellanox with InfiniBand brings to the market. We can now optimize the entire computing stack from the server design -- for the GPU and its software stack to the server design to the interconnect to build a complete AI data center. And by having Mellanox as part of NVIDIA, we can now accelerate that roadmap even faster.

There's also a flip side of that which is of course the edge and deployment side. If I can take that neural network and deploy at the edge, my requirements are different, I need to make sure that I have the best possible network latency. I'm often -- and I care a lot about security. I want to make sure that I have my models are encrypted, my data that I'm transferring is encrypted and yet operating at full line speed. And the DPU, the data processing unit or SmartNIC that Mellanox provides can do all that line encryption, they can do a lot of the computing in the network where necessary both in the training steps and the deployment steps, plan it needs and provide the best possible encryption security for delivering -- for operating on those data which is the end user customer data which obviously is critically important.

And for the people who develop the technology, they can trust that their models, their IP are safe and secure because they never leave the accelerator itself. These are different ways we can accelerate and move the ball forward, both for building an AI data center that isn't just a composition of a bunch of servers, but truly operates as one data center GPU holistically. We are training the network, and then all the way down to the edge where we want to provide the best possible service and security for streaming the AI in production.

Harsh Kumar {BIO 3235392 <GO>}

Wow, fascinating, really fascinating. So it seems like pretty critical, seems like this is -- it's not one plus one equals to two plus some financial equation. It sounds like there's some pretty good technological advantage that you guys will have. In the minds of common folks, this is a question that I get a lot, believe it or not, like what's the difference between a hyperscale today versus a data center for hyperscale versus an enterprise data center? And the perception is, well the data centers at the guys like Google and Amazon are really savvy, really sophisticated, the ones by bigger companies like Walmart, et cetera are pretty good, but not quite the same. Is this actually the case? And are you seeing companies like larger, big retail

organizations and others, trying to push their data centers to be more like more cutting edge and therefore investing a lot of money into their data centers?

Ian Buck {BIO 18454865 <GO>}

I mean, we are honestly seeing data center growth in both vectors, both in hyperscalers for their internal use cases, hyperscalers for their cloud services, infrastructure as a service, as well as the enterprise on-prem. And that's always an equation that people have to make for themselves, whether or not they need to rent or buy and that's a TCL calculation. At NVIDIA, we serve both markets and both -- we see growth in both areas because of the strong demand for accelerated computing in general.

There is a difference between hyperscalers and enterprises. The hyperscalers obviously have a deep bench of technical talent. They invented many of the core technologies around AI that we all benefit from today. So their engagement -- I think our engagement model is a little bit different. We do provide basic capabilities and foundations for accelerating their workloads and they can design the next generation now themselves. We're -- bring it to market in a variety of ways, how we talk to our phones, or how we see recommendations off their websites for products we want to buy or use or links we should click on.

On the enterprise side, obviously they don't have that as necessarily deep of events to develop a foundational technologies. That's where we can -- NVIDIA can help too. So we have obviously and is experiencing ourselves of building AI technologies for our own use cases, which we use internally and developing our own products. The work we're doing for our self driving car, we've built up a very strong ability to develop those -- our own AIs for our own self driving vehicles as well as the work we're doing in robotics, healthcare, smart cities. So what we've done is open those software stacks to the rest of the world and in fact made them freely available.

So they can -- and we have our metropolis stack for smart city, Clara for health care and Jarvis for conversational AI, Merlin for a recommender systems. These are high-level stacks. STKs in some cases with solutions where enterprises can take them freely to their particular use case, their particular modality, apply some transfer learning maybe retrain that neural network to recognize the words that we all withstand like prescriptions, you're calling in or financial data. But they are starting from, a known good and highly intelligent neural network and software stack that's ready to deploy, making it easy for enterprises take advantage of AI and deploy it is our goal and we've done that through our -- by providing pre-built containers, pre-trained models that are starting point for it.

So while we are enabling AI across all those channels and vectors, the engagement model for it is a little bit different, but we're a good one AI company that's working with every AI company. And by doing that, we can learn across all those industries, decide how we can help move the ball forward and then make those technologies freely available to help accelerate the whole AI and data science acceleration.

Harsh Kumar {BIO 3235392 <GO>}

Fascinating. It sounds like you guys are able to cut down the time than a lot of these organizations would have, just starting from scratch and doing the training they can take a lot of what you provided, that moment is kind of not perfectly ready-made but close and then get to it. I wanted to ask about edge computing, you brought it up earlier. So we hear that edge computing is going to end up being one of the big things down the line, it's going to have some level of intelligence. But then, every time you get into a complex application, the edge device wants to talk back to the data center. So how -- how do you guys see that evolution at NVIDIA? Do you think the compute will reside at the edge? Or do you think for any kind of semi-complex application, you're pulled in into the data center again for analysis and infos [ph] whatever else you want to do?

Ian Buck {BIO 18454865 <GO>}

Well, devices become more intelligent, they're going to do something at the edge. In some cases, there is really three things that matter. One is latency. How long does it take to ingest the data, make an intelligent decision, and taking action as a result. And I can -- in some cases doing the -- connecting backups to the cloud or backup in the data center which may be hundreds of miles or states away. It just introduces too much latency that they can get a real-time experience.

The example might be online gaming. We need to have low-ping times, in order to have a good gaming experience. The same is true for AI. If I'm talking to my device, and it takes hundreds of milliseconds or even a half a second to a second of just getting through the different layers of the Internet to connect back to the cloud. Just to ask a question to my phone that's a bad user experience, and people aren't going to enjoy it. So latency is critically important.

The second is cost. Certainly, if we can push more of the competition into the device, and provide the right level of acceleration at the device level, we can save money by doing where it's needed, and not necessarily having to move into the servers and dealing with all the backhaul, including the networking cost, for that much bandwidth. It would just might be prohibited. The third, obviously is data sovereignty. There are certain use cases where it's too difficult to ensure data sovereignty to go all way back in the cloud. It's much more logical and simpler just do the AI in the device. We see this lot obviously in healthcare. Doing AI in the medical, at the CT scanner or the MRI machine, and there is work being done actually in COVID right now to understand, to process images. We never have to send that patient data into the cloud, where of course there is much more hit up considerations and it's just simply do it in the device.

So as a result, those three things are driving more and of course the overall interest in applying AI technologies and AI technology doing bigger and to deliver all those things we have to do them cost effectively and do them with acceleration, to meet the latency requirements. So that's really one of the big drivers around our EGX platform. We're building accelerators, which have both Mellanox networking, and

NVIDIA GPUs in a single accelerator and working with entire ecosystem edge system partners, OEMs and elsewhere to help bring edge acceleration right to where the data is being ingested and where intelligence needs to happen. Those devices will always be connected to the cloud as the AI makes decisions and sometimes makes mistakes or gets corrected. We always want to make sure that we provide the information back to the data center, so that we can build the next AI and build a virtuous cycle and do that of course safely and appropriately.

Harsh Kumar {BIO 3235392 <GO>}

Fascinating stuff. And -- so I'm going to go to a topic that I know is very near and dear to you guys, conversational AI. Every time, Jensen talks he gets really excited when it comes to that topic. So what are some of the new and exciting things that are going on in NVIDIA to the extent that you can talk about it publicly?

Ian Buck {BIO 18454865 <GO>}

Yes. Conversational AI is incredibly rapidly growing field. In fact there isn't an AI company out there that isn't trying to beat the other guy and getting the better BERT or the better NLP model, NVIDIA as well and we welcome all of it obviously. The conversational AI is a great demonstration of where we are in artificial intelligence. Before like I said understanding was in the picture. The simplest -- that is a basic intelligence tests, but one that is fairly common actually in biology, every insect to dogs and cats they all can see images and recognize what's in the picture.

Understanding language, being able to take a web page and summarize it, being able to ask a question of AI and get an answer back in a human and understandable and productive way. It's a real time test and it requires level of intelligence that obviously is far above and beyond. So -- and it's just revolutionizing the way we interact with computers and our kitchen counters and our cars. And so as a result, this is a very active area of investment, both in the research community and as well as driving in the software and engineering side as well.

For NVIDIA, we're really trying to help bring this technology to all the world's enterprises. We invested in our platform called Jarvis which is an end-to-end conversational AI stack that provide some of the basic foundational components of speech and conversational AI from Automated Speech Recognition ASR to Natural Language Processing NLP as well as text to speech taking stream text and making a human sound and voice out of it. One of the reasons why voices we're hearing from computers, from our phones, our kitchen counters sounds so really is because they're entirely generated from AI based models. And some of those models are most complex models and most intensive models to compute and require some of our largest GPUs to run even in real time. But the results are uncanny and quite convincing to get the right breath notes and tones.

Beyond just the foundation of ASR, NLP and TTS, we've also need to do a higher order technologies, like having proper Q&A engines, having proper dialog managers. So you can understand a thread of conversation which is going to be unique to each domain, whether it be ordered in prescriptions or asking for

directions when looking for a restaurant, dialog management is another area where we can help the community move forward, and we've actually provided early dialog measures. We are working with partners who provide those commercially.

The third is things like speaker diarization, understanding who's talking at one time. We've all been on Zoom chat where everyone is talking at the same time and the text to speech engine gets pretty confused. Having a good AI can help us there too. We can understand that I'm talking versus you talking versus someone else talking potentially in a room where most people are attending and be able to isolate it, this is also important in the car scenario. When are you actually talking to the car versus talking to the person next to you. And doing sensor fusion where you're combining camera based technology looking at where is this person looking at, doing gaze detection and combining it with a conversational AI engine to recognize when I should be paying attention and when is this is a separate conversation. As you can tell, this is an area where computers truly become more human. And as a result, radically accelerate how quickly AI gets deployed and really in use cases and makes it very exciting technology to work on.

Harsh Kumar {BIO 3235392 <GO>}

That's amazing, that is fascinating. So I wanted to ask about a business question, NVIDIA wins a lot, we can see that in the numbers. When you guys win, what do you think is the reason or the combination of reasons that allow NVIDIA to win so very often and particularly in compute situation with a data center?

Ian Buck {BIO 18454865 <GO>}

Yes, I mean it is important to recognize that AI and more broadly accelerated computing, it's not just a GPU question, it's not just who has the better chip. You can build a chip that's fast. There is a lot of transistors in 7-nanometer. What makes accelerated computing successful is the entire stack and the ability for the ecosystem to get their applications, whether it be AI, data science like Spark or HPC applications like for your home or Amber to be accelerated. And when we first started this journey, we produced an incredibly fast processor that could accelerate all those workloads, but actually had no users, had zero applications. Because in order for them to be successful, they had to be accelerated at the software level, at the application level. And there are many ways to do that, you can do that at the lowest level by programming the GPU directly, you can do it at the SDK level. We have many core libraries that people can use or you can opt the stack further to remain domain specific SDKs or tools kits like Jarvis I just talked about. Or all the ways to the application level or the ISV or the application liability itself has -- have already been accelerated.

So in order to be successful in accelerated computing you need to invest in an entire stack, you need to work with entire ecosystem. As one of the reasons why NVIDIA has more software engineers than hardware engineers is because what makes us successful is our 20 years of investment in the software platform, software ecosystem for accelerated computing wherever we can make a difference and that's the most important part of the equation.

Harsh Kumar {BIO 3235392 <GO>}

I think this comes back to earlier point, you guys are expanding your footprint turn to optimize data center with Mellanox and some of the other deals. I wanted to shift focus to go COVID-19. I am sure you guys are getting a lot of pressure from customers to help them out with some of the stresses on the data center and the enterprise level. What are you guys hearing from your customers these days that you probably weren't hearing six months ago?

Ian Buck {BIO 18454865 <GO>}

Well, there's certainly a lot of I can just say, the COVID itself is creating demand for understanding how the virus works, what can we do and how quickly can we come up with a cure, all these mitigate the symptoms to make it less deadly. Certainly we see the world supercomputers mandated and tasked with this problem. And so there's a lot of interest. In fact, one of our first -- we just released a new GPU called NIVIDA A100 and the DGX A100 system along with it. And one of the first deployments are on national labs for health, advancing and understanding COVID and it's different, how the virus works fundamentally. And that's actually doing a interesting combination of both traditional simulation of the atoms and molecules inside the virus as well as applying AI technologies to understand configuration and shape of it.

So we're seeing applications for just understanding the structure and the biology of it, of the virus as well as understanding what different drugs like remdesivir could be -- could intercept and muck up, gum up the works of the COVID virus to make it a little less potent and more survivable. In fact, we just ran working with our friends at Oak Ridge national labs, we tested over 1 billion compounds in a molecular dynamic simulation, we simulate -- there we're able to test different compounds against the particular ligand which is used, which is active in the infection process of COVID-19.

Harsh Kumar {BIO 3235392 <GO>}

Sorry, I've got to ask this. How long would this have taken a year ago? And how long did it take now or five years ago how long, would it even be possible five years ago?

Ian Buck {BIO 18454865 <GO>}

Arguably it would have taken about a year, just to run through all those 1 billion compounds on a traditional equivalently sized CBO datacenter or supercomputer with the some of supercomputer which has over 20,000 GPUs in it, they actually will complete in a single day. So that, now of course, that generates a massive storm of data. And in fact what they're doing right now is try to figure out how they can use the GPUs to digest all that data because they've just done all the document calculations and simulations in just a day. Now they're trying to figure out which ones yielded, interesting results versus I know exciting --- they're actually should be published in their work on archive fairly shortly and exciting to see the results.

Harsh Kumar {BIO 3235392 <GO>}

That is amazing, those are the applications you just really can't even buy with money, I mean that's just good work for humankind.

Ian Buck {BIO 18454865 <GO>}

Well it's incredibly important. I mean, I think, while there is obviously hope and interesting work being done with -- in the development of vaccine if we can just come up with a treatment to make the disease less deadly, we can intend it, not get back to and mitigate its risks. We never came up with the vaccine for HIV, but we've come up with the appropriate drug cocktail that made that virus at least manageable from a society standpoint. My hope here is that some of this work can accelerate the discovery of a treatment at least until we get to a proper vaccine.

Harsh Kumar {BIO 3235392 <GO>}

Fascinating. And do you think the COVID, COVID has put a lot of pressure on different things, people are working from home, that's putting a lot of pressure on data center. Do you think this is going to change, this is the way of things going forward and the rate of emphasis and acceleration and importance of the data center has gone up and we'll probably stay elevated from here. In other words, the curve has shifted up?

Ian Buck {BIO 18454865 <GO>}

Well, there isn't a company in the world that isn't rethinking their digital strategy, and how their company operates fundamentally, where the employees are, how they work, how do they interact. And the services that they are providing to the rest of the world who we're all now working and living from home and that is just -- that trend was always happening. Obviously, it's just, has been accelerated 10 times over because of this pandemic which has caused a lot of conversations and obviously a lot of stresses and investment in how we interact with cloud data centers. I do think there is a fundamental change that's been established as a result of this. I think companies are rethinking their own product strategies, their priorities and how their employees work.

It may be a case where this won't be the only pandemic unfortunately and so they all need to be prepared for should the next one strike, the companies be resilient to the need to potentially work from home or manage pandemic situations. Those conversations are certainly happening, the importance of the cloud is critical. The importance of their own data centers and how their own infrastructure can be resilient to those things is also important. And NVIDIA ourselves, have figured out methods and abilities to keep our data centers operating in these situations and growing in order to make sure that our engineers, our own teams have the best capabilities and resources to do their work.

In fact, when we first saw COVID happen and everyone started to work from home, it's utilization or reducing shot through the roof, not because they are suddenly needing them to work from home, they were just, I think people were spending less time in meetings and spending more times actually issuing jobs and trying to get their work done on their supercomputers. So that is an interesting dynamic perhaps in one way, and improvement in productivity given the situation.

Harsh Kumar {BIO 3235392 <GO>}

Definitely. I can tell you that with the people that I have talked to, everybody is implying that they are working harder while they're working from home, so less meetings, less travel, so on and so forth. Let's talk about -- you mentioned something interesting, you said a lot of things you guys do is sort of interoperable, it's open, you distribute a lot of things for free, which means you probably have amazing customer interfaces. But there are lot of customers, take you things or sort of like your ready made models if you will, Jarvis et cetera, and then tweeting. So that puts you in a pretty unique position to be able to have those customer interactions. Has that had an influence on your -- would you think that, that has had a major influence on the innovation rate at NVIDIA?

Ian Buck {BIO 18454865 <GO>}

Certainly. AI and data science is extremely fascinating fields, in order to build the next generation AI data center and the components and products that need to go into it, both from hardware and software standpoint. You need to be on the cutting edge and the conversations with people that define the technology. Fortunately, NVIDIA we have a great research organization, which is helping define the technology, but that great work is also happening at Facebook and Google and Alibaba and all over the industry, at Amazon, and Microsoft especially.

So by engaging with all those customers to see -- to helping them to get to the next level of their services technology or capabilities, whether it be Cloud ML or Sagemaker or Azure ML, whether it be PyTorch or TensorFlow, by engaging with them, we can help them move faster, we can accelerate our own platform, because again the data center features is built from many different pieces and is a complete whole stack solution. So -- we can take the right things and put them into our products and optimize our software stacks to be aligned to what's needed today and gives us perspective on what's needed in the future in the trends. So we can understand and figure out and develop new technologies. One example of that is in our latest GPU of A100, we introduced a new numerical format called TF32 or TensorFlow 32. It's actually a new number format for doing AI based calculations that delivers the same accuracy and the competitions for training neural networks, as the traditional IEEE FP13 result. But it runs up to 10x faster.

So that speed up in technology can take a lot of the legacy, all these neural networks that were traditionally done with 32-bit floating point calculation. And just without any software modification, run them 10 times faster and they experience that out of the box, which is without any code changes. That was really only possible because we work with every other AI company in the world and helping them optimize their

models, to then test and evaluate what is the right number format for doing AI calculations and come up with the next generation one that's going to move the whole industry forward.

And by having that perspective and engagement, we can build better products and innovate -- we need to innovate and also inform and help our end users and point them in the right directions, give them the heads up where things are coming so they can spend their time and focus on their problems and not let -- and just inform us and help us to help them, so they don't have to do some of that work themselves. So I think that creates a really healthy synergy in how we develop our products. It makes our business incredibly valuable to them. And of course helps NVIDIA move the needle, and generation over generation in AI.

Harsh Kumar {BIO 3235392 <GO>}

And Ian, as you look out, what are you most -- what are handful of things that you are excited about over the next 5 to 10 years that may come in the data center or might come out of the data center because of you guys?

Ian Buck {BIO 18454865 <GO>}

Well, I think especially with what COVID has imposed upon the cloud and data center, and everyone accelerating their perspective on it and investments. With the advancement of AI, and integration of AI in all the products we use in our daily lives. And there isn't a day that doesn't go by where I think any -- you don't -- your life isn't touched by AI at this point. It's only the beginning of a much broader application of this technology to improve the way we buy things, the way we engage with retailers, the way we communicate with the cloud or computers and services all around us. Super exciting.

Some of the work being done in recommender systems right now is fascinating. These are some of the -- we thought conversational AI was hard, the term test that's having a human like conversation with the computer. Think about recommender systems. Now I have every single product on Alibaba, every single product on Amazon. I have every single person in the world purchasing products and making buying decisions. Now we have to ask the AI, what's the right product for me. That is a not even a humanly comprehensible problem, that is a -- beyond human comprehension of problem, and solving is critically important.

Ad recommended -- we don't search for products on the Internet anymore, there's just too many. They're just recommended to us. So that is the next front and some of this technology of course, is just developed using the hyperscalers is now becoming available to the rest of the enterprise. And as we worked with many people to help, develop those recommended technologies, we started to package them up and make them activated for the rest of their process. We announced some of the stuff at our GTC conference and we call this the Merlin platform. It's only -- only the beginning and very exciting.

The other area that's very exciting is data science. Now that accelerated computing and data has become so critically important, if not just developing and deploying the AI technologies which is to understand the basics of the data itself. Some of the work we've been doing for accelerating Spark and I don't know if your folks noticed, but there's new version of Spark called Spark 3.0 that just came out. Spark is sort of follow-up from the original Hadoop, and that produce technologies for understanding data science across the cluster. Spark can do all that in memory and much faster. And with Spark 3.0 then we're getting it accelerated with GPUs. That is an area where we're activating all of data science now for understanding data and understanding in a much larger scale and much more intelligent modeling and of course now with GPUs in a reasonable timescale so they can get their work done.

So I'm really excited about the work being done in recommender systems, the work now being done in the broader data science community. And the third area I'll highlight is edge. A few years ago edge was manageable with legacy infrastructure, it's not today with the deploying AI at the edge, intelligent services at the edge, the ideas around sensor fusion, with the advent of 5G and having effectively infinite bandwidth to our devices in our homes for those services is naturally pushing a lot of the competition while the capabilities from the big isolated data centers down to the telco, down to the path, down to the edge where they can be best served and deliver that latency.

So those are three areas that I am super excited seeing in the next or at the least next few years over the broader decade, who knows, we've articulated $50 billion data center TAM by 2023 and that's not including the non-ox is going to be super exciting to see where this ends up in a decade. I could have imagined it 20 years ago when we started CUDA where this all would be going. I'm certainly not going to make a prediction on future again but I know it would surprise me.

Harsh Kumar {BIO 3235392 <GO>}

So at this point, I'd like to mention something that you and Simona and I were talking about. Ian, for those that are listening in, Ian is actually the creator of CUDA. So here is the man that sort of led it all. We've got a handful of minutes. I wanted to ask about, as you guys looked at Mellanox, and as you've owned it, now what are some of the amazing things or interesting things that perhaps you guys may not have known before something with their capability or technology or anything that you thought was interesting?

Ian Buck {BIO 18454865 <GO>}

They have -- so one thing is super interesting is the BlueField technology in the Mellanox SmartNic. That opportunity to do computing on the data on the wire is an area that is super exciting both for the data center as a whole. There are certainly optimizations that can made while we're doing the computing in the network to make things run much more efficiently, then pushing the data all the way out to the nodes and trying to do the competition there and resolving it there. When you're doing AI training or data science in general, if you can -- some of the competition

naturally wants to happen inside the network because that's where all the bits and data is transacted, it's moving anyway.

And as a result can be bandwidth multiplier if you will, on the order of 10x by doing some of the competition in network. Doing some of the intelligence on the wire for the edge as well from the SmartNic technology is super interesting and then supporting things like virtualization during virtualization offload into the SmartNic. I think the BlueField technology is very promising. We already announced and deploying it in certain areas. And as a computing company is obviously a new area, a new canvas where we can do some acceleration. I'm very excited about the BlueField technology and look forward to it also advancing the data center.

Harsh Kumar {BIO 3235392 <GO>}

Ian, I had a question in there, what innings is the data center in, but I'm not going to ask it. Listening to your talks, it's pretty obvious we're in very early innings. With that, we're almost out of time. I just wanted to thank you, Ian, for your time today, Simona thank you for yours and thanks to everybody that's signed in and appreciate everything you guys do particularly with respect to finding cures in healthcare.

Ian Buck {BIO 18454865 <GO>}

Thanks.