REFINITIV STREETEVENTS

EDITED TRANSCRIPT

NVDA.OQ - NVIDIA Corp GTC China 2020 Keynote

EVENT DATE/TIME: DECEMBER 14, 2020 / NTS GMT

CORPORATE PARTICIPANTS

Ajay K. Puri *NVIDIA Corporation - EVP of Worldwide Field Operations*

Ashok Pandey

Greg Estes *NVIDIA Corporation - VP of Developer Marketing*

Kimberly Powell *NVIDIA Corporation - VP of Healthcare*

Raymond Teh

William J. Dally *NVIDIA Corporation - Chief Scientist & Senior VP of Research*

PRESENTATION

William J. Dally - *NVIDIA Corporation - Chief Scientist & Senior VP of Research*

(presentation)

Hello, and welcome to GTC China. For those of you who are meeting me for the first time, I'm Bill Dally, Chief Scientist and Senior Vice President of Research at NVIDIA. I lead NVIDIA's research labs and work with our product groups to transfer the technology we develop and research to make our products even better than they already are.

Today, I'm going to tell you about some of the great technology we're developing. We build the world's most high-performance computing devices and focus on the world's most demanding computing problems.

It all starts with our hardware, which today is Ampere. Here's an Ampere A100 SXM module. This is a tremendous amount of computing performance. I'll go into some of the details in a few minutes, and we can scale the power that's in Ampere from this module for very demanding computing problems all the way down to our Jetson line of products for embedded products.

And if you want to scale up, we can take 8 of these Amperes and put them in a DGX box, the gold box you see on the screen. And we can put a number of those into a rack with our Mellanox switches to build computers that are among the most powerful in the world. But hardware by itself doesn't solve the world's competing problems. It takes software to focus this tremendous computing power on demanding problems.

And so we've put a tremendous amount of effort into developing a software suite to do this. It all starts with CUDA, which is an outgrowth of work we did on stream processing back at Stanford. And since 2006, people have been using CUDA to harness the tremendous amount of power you can get out of GPUs and it lets you get it every last bit of that power. And then built on top of CUDA to simplify building applications, we have a whole bunch of libraries. If you're doing things having to do with linear algebra, we have CuBLAS and CuSPARSE. If you're doing things with spectral methods, we have cuFFT. We have an adaptive multigrid package.

For deep learning, we have cuDNN and TensorRT that simplify the task of getting very high-performance implementations of deep learning. On top of these, we build applications in a number of verticals. First, graphics is one of our most important areas, and I'll tell you a little bit about what we're building in the graphics area. We have a tremendous amount of software to support artificial intelligence, including software for natural language processing and recommender systems.

In health care, we have a Clara package that runs a gamut from our Parabricks that does genome analysis to image analysis the things that mine databases of medical papers. We have packages for intelligent video analytics. So you can take video streams and draw conclusions from what you're seeing. We have an entire package for autonomous vehicles, from curating the data set, training in the data center and deploying in the vehicle itself. And our Isaac package makes it simple to deploy robotic systems on top of our Jetsons.

So let's talk a little bit about Ampere. It's an amazing device, not only is it the world's largest 7-nanometer chip with 54 billion transistors. It has a number of innovations to make it a lot more powerful than our previous generation. It's our third-generation of Tensor Core, special hardware we add to our GPUs to accelerate deep learning. And in this generation, we've added support for new data type, Tensor Float 32, which solves the problem that in the past, you had to choose between bfloat16, which has a lot of bits of exponent and has a very high dynamic range. But very few bits of mantissa, and therefore, not enough precision to train many networks or FP16, which flips it the other way around. It has lots of bits of mantissa, it has a lot of precision, but limited dynamic range.

Tensor Float 32 gives you the best of both worlds. It does a lot of dynamic range, a lot of precision. And so far, it's been able to train every network that we've applied it to. So we get tremendous amount of computing performance using the TF32 data type.

One of the things I'm most excited about is Ampere has finally cracked how to exploit sparsity in neural networks to get better performance. And I'll have a whole slide on that in a minute. If you need to scale down in Ampere because it's such a powerful computing device, our MIG or multi instance GPU technology, let you treat 1 Ampere at 7 separate GPUs, so you can run separate tasks on each of them.

And if you need to scale up Ampere to solve even larger problems, our third-generation of NVLink and NVSwitch have twice the bandwidth with 600 gigabytes per second of bandwidth off of the GPU. So here are some details about Ampere. I'm not going to go through all the numbers here. But to me, what's most important are its performance on 3 data types for how well it does in 3 different types of applications.

For high-performance computing, Ampere has double precision Tensor Cores. And so for FP64 arithmetic, we can sustain 19.5 teraflops on executing a matrix multiply. For deep learning training with the new Tensor Float 32 data type, we have 156 teraflops of performance. It's a tremendous amount of performance for training neural networks.

And then for deep learning inference, with INT8, we have 1.25 petaOPS. If you can exploit the sparsity that I'll talk about in a minute. And if you can solve your problem with INT4, which many networks can, you have 2.5 petaOPS of inference performance. This is just an astounding amount of performance.

So let me tell you a little bit about sparsity. So it turns out that most neural networks can be pruned. I published a paper on this at the NeurIPS Conference in 2015, and showed that you could prune these networks, cutting out anywhere from 70% to 90% of the connections between neurons. What this means is that 70% to 90% of the weights in the network could be set to 0 without affecting the accuracy.

Now immediately, this gives you gains in terms of compression. You don't have to store all those 0s, so you can get more capacity out of your memory, more bandwidth out of your links. But until now, people have been unable to exploit this sparsity to get better arithmetic performance. And it basically was something we weren't able to take advantage of. Something sitting on the table, waiting for us to exploit.

With Ampere, we now are able to take advantage of it. And Ampere solves this problem by exploiting structured sparsity, by allowing 2 out of 4 of weights to be 0. We have a regular pattern that makes it easy for the hardware to exploit the sparsity without the overhead of irregularity, slumping the games from the sparsity.

So for a matrix multiply instruction, once you sparsify the weight to this 2 out of 4 pattern, you get double the performance. And even over an entire application where matrix multiply is only a part of the application, such as for the large inference natural language processing benchmark, we get 1.5x of performance.

This is a tremendous jump forward in architecture for deep learning. And we can take these A100s, and we can combine 8 of them in a box, along with a lot of SSD for storage, a bunch of RAM, and 9 of our Mellanox ConnectX-6 NICs. And this makes a great computing platform with 8x the performance of one of these GPUs since each of them is close to 20 teraflops double precision. This is close to 160 teraflops.

We can then put a number of these into a rack, and this is the DGX POD. The rack also includes one of our Mellanox switches to connect up those NICs and allow interconnection between those DGX boxes.

What's great about a DGX POD is that we've already solved the integration problems. We've made sure every piece of hardware here works together and everything is configured properly. So if you buy a DGX POD, you plug it in and it works on day 1 with all of the containers we have of software for you to run on them. There's no debugging, no figuring out what the right configuration is. And you can scale these DGX PODs up to what we call DGX SuperPOD.

In fact, we built a very large DGX SuperPOD, with 280 of our DGX A100s, that's 2,240 individual GPUs. And this is our Selene supercomputer. It's #5 on the TOP500 list and #5 in the Green500 list. It's the fastest industrial system in the world. It's the #1 industrial submission to the TOP500. What's really great about our Ampere technology is not only is it great at deep learning, that same technology is great for high-performance computing, and this simplifies the integration of AI into scientific applications.

You don't need to do part of it on one machine and part of it on the other. The same machine does both parts. In fact, in the recent TOP500 back at the Supercomputing Conference in November, NVIDIA technology was in 8 of the top 10 machines. Selene, that I just showed you, was #5 in both the TOP500 and the Green500. #1 in the Green500 was a different DGX SuperPOD.

We were also #1 in terms of academic machines, #1 in terms of industrial machines against Selene, and #1 in both the U.S. and China where the TaihuLight has a Mellanox network and #1 in Europe with the Jülich machine. Of all these accolades, the one I'm actually most proud of is the #1 on the Green500 list. Because you can move yourself up on the TOP500 by writing a bigger check, configure more nodes, connect them up with a bigger network, and you will get a higher high-performance LINPACK score. But the only way to move up in the Green500 list is to be more efficient. It actually takes better technology, better architecture, better circuit design to move up in the Green500 list. And so I'm very proud that an NVIDIA machine is #1 in the Green500. Because that's strictly a measure of our technology, not how big a check somebody wrote to build a machine.

Now for the AI applications, a lot of people these days are building dedicated AI accelerators and they claim that these will be more efficient than general purpose GPUs. This is not really the case. And this chart explains why. It shows sort of the evolution in many ways of our deep learning architectures.

Back in our Kepler generation, the biggest instruction we had to do deep learning with was a half-precision floating multiply accumulate. Normalizing this to technology since these are all going to be compared the same, that's about 1.5 picojoules of energy. And fetching and decoding the instruction, all of the overhead associated with doing that instruction is about 30 picojoules.

So here, the overhead swamps the payload. We're spending 20x as much energy on overhead as we are on the payload.

In our Pascal generation, we moved forward and had a half-precision dot product instruction, dot product, a 4 vector. Now we're doing 8 arithmetic operations. 4 multiplies 4 ads, 6 picojoules of energy. And the overhead is only 5x as much, still not really acceptable, but better.

Starting with the Volta generation, we introduced our Tensor Cores. And what Tensor Cores really do is they provide specialized instructions for matrix multiply accumulate. So in Volta, where we had a half-precision matrix multiply accumulate, HMMA, we now have the energy going into the payload, actually doing the 128 floating point operations associated with that 1 instruction, now completely dominating, amortizing out that overhead. So the overhead is only 22%.

And with Turing, when we had added the IMMA instruction, which is now doing 1,024 in 8 operations, the energy of the payload is 160 picojoules, and the overhead is only 16%. What this means is that if you build a dedicated accelerator that didn't have any programmability, you'd be getting a 16% advantage. But you'd also be getting a huge disadvantage. The programmability of a GPU allows you to track advances in neural networks, which happened at an amazing rate.

New models are coming out all the time, better training methods and to exploit these, you need a machine which is very programmable. The GPU offers you a completely programmable platform and by building Tensor Core specialized instructions amortized overhead. We're able to offer you that programmability with a negligible penalty compared to a dedicated accelerator. This chart tracks a little bit of the progress I showed in that table, but on a different scale.

This shows our single-chip inference performance from Kepler in 2012 up through Ampere A100 in May of this year. And what you see is that in this 8-year period, we've increased single-chip inference performance by 317x. This curve has come to be known as Huang's law, which is that inference performance doubles every year. Actually, we're more than doubling it every year. And this is in part because of those advances in Tensor Core that I mentioned on the previous slide, better circuit design, better architecture, very little of it is due to process technology.

There are only 3 generations of process technology on this graph, 28 nanometers at the beginning with Kepler, 16 nanometers in the middle and then most recently with Ampere in 7 nanometers. And those jumps in process technology gave us very little of this 317x, probably less than 2x overall, most of this is from better architecture. And so with Moore's Law going away, it's a good thing that we have Huang's law here to keep pushing up computing performance because we're going to need it for a lot of things we want to do in the future.

Now if you want to compare how well people are doing on deep learning performance, the way to do this is with the benchmarks we called MLPerf.

Most of the people who are players in the machine learning hardware space have gotten together and agreed on a set of fair rules for comparing the performance on both inference and training. And periodically, there is a submission where you have to have your results validated and everybody submits the results, and the results are then posted to the MLPerf website.

So here, I'm showing you the results from the most recent MLPerf training benchmarks. And what you see is that NVIDIA sweeps all of the categories. We win consistently. In fact, our biggest competition is our last generation. This is all normalized to 1x being a Volta V100. You see that Ampere is up to 2.5x faster on deep learning.

And in particular, in the areas that really matter like some of these large natural language models and recommender systems that's where 2.5x are. Most of the competitors don't even show up. There's not a single start-up that showed up. The only 2 other competitors that even showed up for a couple of the benchmarks are Google with their TPU V3 and Ampere beats them soundly. And Huawei showed up only for the image classification benchmark and got trounced 2x by an A100. The MLPerf benchmarks are broken into the training part, which I just showed you, and inference. And again, NVIDIA sweeps all of the categories. The inference part of MLPerf is itself broken into a data center inference and an edge inference category. The data center inference, we see as our biggest competitor is ourselves, and the gap here is even larger.

So compared to Turing T4, which is our previous solution for AI inference, the A100 is between 6x and 8x faster across the board on all of these benchmarks. The only competitors that show up here are Intel and Xilinx. And in all of these cases, they're actually beaten by the T4, Ampere just trounces them. None of the start-ups, and we're tracking numerous startups, all claiming to have better solutions for the inference space, show up here.

If they had something better, they should show it up on an MLPerf. But they don't bother to show up. The other side of the MLPerf inference benchmark is the edge inference benchmark.

This is for edge servers and embedded devices. Here we're showing our performance numbers for the A100 for the T4 and also for the Jetson AGX Xavier with our Tegra chip. You can see that again, NVIDIA sweeps all of the categories. Centaur shows up and beats Xavier and a few of them, but even against Centaur, the T4 and the A100 are quite a bit more powerful.

So as I said previously, we have software packages that focus the power of NVIDIA's GPU architecture on demanding problems of interest. And one of the most important of those for NVIDIA is computer graphics. We've been a graphics company for a long time. Now many of you will watch a motion picture or feature film that has portions of it done with computer graphics, with CGI. And you'll be impressed with how photorealistic that looks.

And that off-line computer graphics is typically done in a way where every frame takes hours of time on a compute farm to generate by casting tens of thousands of rays per pixel using a technique known as physically-based rendering with path tracing.

We've recently been able to come up with a number of innovations that approach that photorealism real-time at 60 frames per second or faster. So watch this movie. You'll notice as lights come on, they cast spots of light on the floor and they cast realistic shadows from objects that are in the pathway. The shadows are soft where it's appropriate. Each of these marbles has reflections showing things that are reflecting off of the marble and specular highlights showing where the lights are, that move around as the camera and the lights move.

As these glowing marbles roll down, they're casting light on the floor in a realistic way. And here, we see that the shadows moved properly as the lights in the scene move around. And there are proper soft shadows with the edges, reflecting the distance from the object of the surface that it's casting on. This is photorealistic. It's very close to what you would see in a motion picture. But it's rendered at 60 frames per second on a single NVIDIA GPU.

So let's look at how we're able to accomplish near photorealistic rendering at 60 frames per second. It comes from a number of techniques that I'm very proud of because they've been developed in NVIDIA's research lab. So the first is something we call RTXDI for direct illumination. And this image actually shows what things used to look like, what conventional graphics does with direct lighting.

You see here a number of emitters, these little Christmas lights, but they're not casting their light on the adjacent surfaces. They're not casting shadows. It looks wrong. This is incorrect lighting. With RTXDI each light is casting its light onto adjacent surfaces. So the yellow light is making the adjacent surface appear yellow, the green light, green and so on. And where things are between the light and the surface that's casting realistic shadows. And we can support up to millions of lights with this technique. It's using a technology called the reservoir important sampling, in which we call ReSTIR. And it was published in SIGGRAPH 2020 and is already now in our NVIDIA graphics products. And it produces wonderful direct lighting. So this is half of the problem.

The other half of the problem is indirect lighting. So RTXDI makes the lights extremely realistic for 1 bounce, from the light to a surface and then back to your eye or the camera. But lights bounce numerous times, infinite times. And to do that, we have a technology called RTXGI. RTXGI cast light from one surface to another using light probes, little spheres that we put in the image at various points to compute the indirect lighting as seen at that point.

So you can have an infinite number of bounces, a surface will illuminate another surface, which will illuminate another surface. This is done in a way where there's no leaks. So if you have a very bright room next to a very dark room, some little difference in polygon shapes won't let the light leak through the wall here. It's done in a watertight way. It's also a great example of multirate rendering. Because the indirect lighting doesn't change at a very rapid rate, we can render the scene at 60 frames per second, getting the things like visibility and shading that need to be done 60 frames per second done at that rate, but recompute the indirect lighting at a lower rate, depending on the available computing resources. This is a really wonderful technology. You can see how the upper part of the image looks very realistic with the indirect lighting. In this case, almost all the lighting is indirect, since only a little bit of light is coming in from the windows. The bottom of the scene is almost dark because without the indirect lighting, you don't have much. You can even appreciate it more if we turn off the textures and get the lighting only here, you can see how much richer the scene looks with good indirect lighting.

Another technology that lets us accomplish a lot at real-time rates is NVIDIA's DLSS or Deep Learning Super Sampling. This is a technology that is evolved. It's now in DLSS 2.0, which offers even better performance that original DLSS 1.0. So as illustrated in the graphic here, you start with a image at a certain rate, say, 1440, and we feed it into a neural network that upscales it to 4K.

And then to make this work, we take this upscale image and we compare it to ground truth, actually rendered at a much higher resolution, in this case, 16K. And the errors there go into a loss function that's used to train in neural network via one of our DGX SuperPODs, and that over certain iterations over data sets, we train the weight to the network to produce upscaled images in a very accurate manner.

Now there are 2 tricks to making this work. The first is to make it temporarily stable. It's relatively easy to upscale a still image. But if you were to upscale just a series of still images without worrying about motion, you would get a numerous artifacts by doing things a little bit inconsistently image to image. This would cause very objectionable things that would pop out at a viewer, things little wiggly worms and the like that appear where things are inconsistent from frame to frame.

So we've worked very hard and have solved that problem of temporal stability. We get very temporarily stable videos from this technology. The other thing which is difficult to do well is to make it generalize. And we've been able to do that so we can train 1 neural network and have it work on every level of the game and across multiple games. We don't have to retrain for each game, each level or each scene, 1 neural network to rule them all. Here's how the results look. So on the left is native 4K, on the right is 1440 image that's been upscaled to 4K. The frame rate is shown in the upper right-hand corner. Where you can see is not only are we running at a higher frame rate, in this case, 141 frames per second. But if we zoom in on the suitcase on the character's back, you can see it's a better image, it's sharper. You can see more details in the DLSS image than you can in the native 4K image.

So it's a wonderful technology. We get faster frame rate and higher quality graphics. So where are we ultimately going? In NVIDIA research, we're pursuing an agenda to push our rendering to be fully motion picture quality. And to do that, we also want to like the motion pictures, do physically-based path tracing, but we want to do it in real time.

And this image sort of shows what our vision of that is. We want to be casting rays from the camera and be able to bounce through some number of specular reflections and refractions, such as through the beer glasses in the image in the upper left of this slide. And then as we come out of some number of those specular bounces, we'll do a couple of bounces where each bounce, we do many light sampling using the reSTIR algorithm I mentioned in talking about direct illumination, this will give us wonderful direct illumination. And then after perhaps 2 bounces, which maybe all we can afford, depending on what graphics hardware we're running on. We will terminate into one of these RTXGI light probes to get very accurate indirect lighting.

And there are a number of technologies making this possible. I talked about the RTXDI and GI that give you good direct lighting and indirect lighting, respectively. What allows you to do these bounces, both the specular ones and then the diffuse ones, you're bouncing off of these intermediate paths is the RT Cores that are in our most recent GPUs that really accelerate ray tracing, making it possible for the first time to do ray tracing and real-time graphics.

Another technology, which is very important, is doing good de-noising. Because we can't afford to send 10,000 rays per pixel, the way the motion picture people do. And we need to get by with somewhere between 1 and 10 rays per pixel, and that will produce a very speckled image with a lot of sort of shot noise in it, as if you shot it at a very high ISO. But then by applying deep noising, in particular deep learning denoising, we're able to clean that image up and make it look really great.

So this is sort of our near-term vision for where computer graphics is going. In the longer term, we expect computer graphics to be generated by AI. The images of people you see on the left are not in fact images of people, they were generated by a neural network, these people never existed. And because they can generate images of real people, they can also generate images of stylize people as in the middle frame, and they can generate images of animals, cars, rooms, arbitrary scenes.

And in the long run, we expect that if you want to produce good graphics, you use AI, in particular generative networks to generate these images directly without ever having geometry. We're just at the beginning of this today, where we can generate single things, a single person, a single car, a single room. But we can't compose them well yet and we can't get the lighting to work well and the interactions to work well, but we'll get there. And eventually, computer graphics will be generated by AI. So the future of graphics is AI. In fact, the future of almost everything is AI. It's affecting almost every aspect of our lives on how we work, our entertainment, how we play.

Let's talk a little bit about how AI came to be where we are today. The current AI revolution was really created by GPUs. If you look at deep neural networks, there's 3 key pieces that come together to make them work. The algorithms, the deep neural networks themselves, the training data, large data sets that may be labeled in some cases and the hardware that it runs on.

The algorithms have been around since the 1980s, deep neural networks, convolutional neural networks, backpropagation, stochastic gradient descent, have all been known since the 1980s. The large data sets have been around since at least the early 2000s with things like the ImageNet data set. But it wasn't until 2012 when Alex Krizhevsky developed AlexNet, a deep neural network running on an NVIDIA GPU that the current revolution really took off. In that 1 year, he got a performance improvement on AlexNet, which was more than the previous 5 years of work on image net combined. Now GPUs-enabled deep learning, they're also gaining its progress.

If you look at this chart over just a few years moving from AlexNet to ResNet, the demand for compute performance increased by more than an order of magnitude. More recently, as shown on the right side of the chart, progress in natural language processing networks, moving from BERT to GPT-3 has been even more rapid and causing even larger demands on petaflop days of training time and the networks people can build are very much limited by the power of GPUs that can use to train them.

They would like to build larger models and train them in larger data sets, but they're limited by what they can train in a reasonable amount of time on an available GPU. So I'm going to show you this graph again of Huang's law. We're trying to meet this demand and allow progress in deep neural networks by doubling performance every year to allow people to build more powerful networks and train them on more powerful models. Let's look at one type of network. This is a generative adversarial network that was developed by Ian Goodfellow in 2014, and it's a way of synthesizing images, or actually, you can synthesize just about anything with these. And it really works by taking 2 networks and training them together.

On the left side, we have a generator network. The generator network takes a random number. We call this a latent variable. It's basically sampling a latent space, a space of the distribution that it's being trained on, and it generates an image. If it's an image generating GAN. The right side of the figure is the discriminator network. It's a neural network that takes us an input and image if the generator generating images could be a sound wave form or something else if it's generating a different modality.

And it has a switch. And so as you train it, you flip the switch, but the discriminator doesn't know which way the switch is flipped. And it's trying to tell just from the image, whether the image is a real image or whether it's generated by the generator. And by training these 2 networks together, the discriminator gets really good at deciding what a real image is. And to try to fool it, the generator gets really good at generating real images. And it will generate real images according to whatever training set of real images you use to train the pair.

So what he's really doing is it's learning the distribution of images in this training set. Now a lot of progress has been made since seeing Goodfellow's original work. And much of it actually by NVIDIA, we've come up with a number of very important movements forward. On a number of years ago, we developed something called progressive GAN. Whereby, training these networks with a curriculum, where we started with low resolution networks, originally 4x4 and 8x8 and on up to 1024x1024 pixels.

By learning the low resolution first and then moving out to the high resolution, we're able to make progress. Where in the past, before we did progressive GAN, people had not been able to produce sharp, high-resolution images with GANs, It was the breakthrough that enabled that. More recently, we developed something called styleGAN, which is shown on the right side of this figure, where we basically take that latent variable and feed it into its own neural network, which then decomposes that variable and feeds different parts of it into different levels of the generator network.

This allows us to independently control features at different scales, different sizes. And by doing this, it makes it much easier to disentangle the weight and variable to separate parts of the latent variable that control different parts of the image. So that we can, for example, control whether somebody is smiling or not, whether they have glasses or not, their hair color, by pulling apart that latent variable space.

Now one application of GANs is to video conferencing. In a normal video conferencing call, you would simply take an image, a video stream and each image of the video stream, we would motion code and take all the coated pixels and push them over the wire. This takes a lot of bandwidth. By using GANs, we could be much more efficient.

One aspect of our Maxine technology, we send a single still image over to the receiver, where that still image is used to prime a generator network. Then for every frame of the video, we extract key points and send just the key points over to the receiver, which is a very low bandwidth stream.

Generator network then combines the key points with the still image and generates an animated image. So we get very high-quality video with very low data rate for video conferencing, which particularly during the pandemic is one of the most key computer technologies around. Now what's neat about this technology is not only can you animate your own image. But if you choose one day to be a cartoon character with blue hair, you can do that.

In this video, TV reporter, Charlene Shen, is speaking, and we're taking the key points from what she's doing, along with her voice, and translating them to a still image, which is of this cartoon character with blue hair.

(presentation)

William J. Dally - *NVIDIA Corporation - Chief Scientist & Senior VP of Research*

Now just as we can use a GAN to generate great images, a lot of AI today involves speech and text and language. So NVIDIA has a system for doing that. It's called Jarvis. Jarvis is a multimode conversational AI service. So it basically takes the full gamut. You can speak to it. There's neural networks that will basically take your audio and recognize from that audio, text. We can then feed that text into natural language models to do querying, translation, question answering.

We then can take the answer and feed it back through a natural language model, generate text and feed that and detect the speech and produce an audio wave form out. Now what's really neat is what happens when we combine Jarvis, which allows us to interact with our AIs via natural language with GANs, and in particular, GauGAN, which lets you paint by numbers. You can paint where you want grass, where you want mountain, where you want water. And then it fills in the details, it produces pretty nice images.

Let's take a look at what happens when you do that.

(presentation)

William J. Dally - *NVIDIA Corporation - Chief Scientist & Senior VP of Research*

Now another thing we can do with our language models is build systems where we train the language model on a bunch of text. And then given prompts, we can have it write sentences. Here's an example of our Megatron large language model on the prompt case. We're going to give it a prompt and see how it writes realistic sentences given that prompt.

So our Megatron natural language processing model has been trained on a large corpus of data. And from what it has learned, it can generate sentences given a prompt. Here, we start with the prompt, male was bored on a weekend. From this prompt, the model suggests keywords decided and go, we reject these in instead type walk. Given this prompt and the keyword walk, Megatron generates, he decided to go for a walk on the local beach. It also suggests new keywords, saw and beach. Here, we decided to go with those, just hitting return. Megatron comes back with, at the beach, he saw a group of surfers and predicts new keywords, decided and serve, which we also accept. Megatron comes back with, he asked one of the surfers if he could join him to surf. No keywords are predicted, and we add the keyword happy. Megatron generates, he was happy as he got to surf for the first time ever.

Now after GANs and language models, another really key piece of AI that's revolutionizing a lot of aspects of life are recommenders. They are, in many ways, perhaps one of the most important pieces of AI in the data center. Since they're the ones that particularly decide which ad to put up in front of somebody. And that determines how a lot of revenue is generated. There also, for example, if you're watching movies on Netflix, will select which movie you're likely to be given next. For e-commerce, they'll recommend what products you might like. They use it on social media to recommend things and most importantly, ads. And these are very hard AI problem because in addition to having the neural networks, they involve embeddings. You may have 1 million items, and you don't want to represent that as a million bit vector. So instead, you take the items and you run them through an embedding table to generate a shorter length bit vector to then run through your neural networks.

These embedding tables are very large and they're very sparse. And the neural networks themselves wind end up being very sparse. So it's a very challenging aspect of computing. NVIDIA has simplified this for people with our NVIDIA Merlin package that basically makes this recommender technology accessible to pretty much anybody. And by accelerating with GPUs, we're able to take both the ETL, the extraction, translate and loading part of reading a large data set to train one of these networks as well as the training itself can turn hours into minutes, it's accelerating the ETL part by a factor of 90 and the training part by a factor of 60.

There are 2 aspects to deep learning, training and inference. Training is where you take curated data set possibly labeled and iteratively use it to compute the weights for a model, a network. Inferences where you take that train model, apply an input to it and get an output. You might be putting an imaging, getting a classification, putting speech in, getting text out, putting a question in, getting an answer out, putting a random number in and getting an image out for a GAN. Whatever it is, inference is where most of the horsepower in deep learning takes place.

And to provide that horsepower, we're making sure that Huang's law is continuing. So we're doubling inference performance every year, actually more than doubling. But just providing that horsepower isn't sufficient. There's more than just computation to handling inference. It's a complex problem involving many people as illustrated here.

It's the data scientist who will curate the data set, select which model to be used, possibly through trial and error, trying many models and different data and train that model. Then a machine learning engineer will optimize that to get better performance in the data center, optimizing the return on investment.

The optimized model goes into a model store and an operations team decides how to deploy those models, pulling them from the operation store, running them on a particular piece of hardware to take a query in and get a result for the end user.

To simplify this process of deploying inference, NVIDIA has developed the Triton Inference Server. It's open-source software that supports multiple different back ends for inference and makes it very simple to deploy inference in the data center. It supports the standard query reply interface to the AI application with either HTTP or remote procedure call, it takes these queries and batches them.

Each application is different in terms of its latency and throughput requirements. If it's very latency sensitive, you need to run an individual query itself for a very small batch, so you don't lose a lot of time waiting for other queries to show up. In other cases, you can afford to be more efficient by getting a larger batch. That's all handled by Triton in the dynamic batching module. Once you have a batch, it gets dropped into a per model scheduler queue to wait its turn to run.

And then eventually, gets deployed on either GPUs or CPUs, we have multiple back ends. And we can handle models that are written for different back ends. In a given data center, you may have some models written in PyTorch, others ONNX, others in Tensor Flow, we can handle all of those and more.

So it's very easy to interoperate models and allocate your GPU resources appropriately. It all makes it very simple.

In fact, typically, you're not running just a single model, but you're running an ensemble of models. As in the Jarvis examples previously, where you may have 1 module that does feature extraction, feeding into a neural network that does speech-to-text, feeding into a natural language model that does take question and answering and then back out through speech synthesis and ultimately the wave form generation. Triton can take an ensemble like this and understand how to schedule all these different modules and feed them together. Simplifying the deployment of complex inference in the data center.

Another key application area for artificial intelligence is health care. NVIDIA's Clara package is a suite of applications designed to accelerate health care on the GPU. As shown here, we can start with genomics with our Parabricks package and take a genome sequence, do assembly, do alignment, do variant calling to help personalize health care for an individual based on their genome.

Many pathogens like COVID had structure and help therapeutics interact with that structure is important. Packages like CryoSPARC help discover that structure. With CryoSPARC, one takes a bunch of images with x-ray crystallography. And then on the GPU, CryoSPARC constructs the structure of the virus from those images. Once you have a structure, you can do docking experiments with AutoDock.

With our RAPIDS database system, we can take billions of therapeutic compounds, access them out of the database, run AutoDock on to find which one is the most likely to interact with the COVID virus and then select those for further screening. That further screening would be done in various molecular dynamic simulations to find out which ones actually bind to the particular sites on the virus that are most interesting and are, therefore, suitable for advancing to do physical test on. Clara also includes a lot of facilities for imaging. It can analyze x-rays, mammograms, ultrasound, helping radiologists discover things that are important for their diagnosis, whether it's from a genomic analysis or image or anywhere else along this pipeline, a doctor may want to see what the literature has to say about something.

And the medical literature is huge with too many papers for any individual to keep up on. So we have our BioMegatron, which is the natural language processing system, so the doctor could query a particular thing that they're looking for. BioMegatron will search that medical literature and suggest which articles are most relevant and be able to answer questions about a particular condition.

As one example about how GPUs are accelerating health care at many different timescales, consider the examples shown here. Folding@home is a program where people with GPUs can donate unused cycles to be used to take protein sequences and run folding codes to try to discover structure from sequence. This takes months of time on many different GPUs, but it's 30x faster than other methods.

Moving up to the day scale, taking the x-ray crystallography data, these x-ray images of frozen virus, one can discover the structure of the virus in 12 days instead of 5 months as was previously required. More recently, DeepMind has released AlphaFold, which is a way of applying reinforcement learning to learn structure from sequence, which will take the amino acid sequence of a protein, for example, of a gene sequence of the virus and be able to discover its structure using artificial intelligence much faster than previously possible in minutes.

On the other side, once we have some therapeutic compounds, we want to screen them against a known structure, AutoDock has been GPU-accelerated to run 33x faster than previously possible. And the Oak Ridge National Laboratory and their Summit supercomputer based on our GPUs, was able to screen 2 billion compounds in 1 day, which would have previously required 3 months.

Once those compounds are discovered, they can be simulated using a code like Torch to screen millions of drugs in 8 minutes instead of being hundreds of days. If further analysis is needed, there are now full computational chemistry packages that run on the GPU that can be used for additional analyses of these drug candidates.

At the beginning of the pipeline, we start with genomics. And NVIDIA's Parabricks package accelerates genomics on the GPU. It includes modules for doing assembly. When you sequence a genome, you get a number of reads. They could be anywhere from a few hundred bases in length to tens of thousands of bases in length. But it's like a jigsaw puzzle. We need to assemble all of these reads, a big 1 dimension jigsaw puzzle into the full 3 billion based human genome. Parabricks does that assembly. And then once it's done, it will compare the assemble genome to a reference genome and call out variants, places where a particular individual's genome differs. And those variants can be very important for therapeutics. As shown in the middle plot here, there are many different ways of calling variance.

But Parabricks accelerates all of them by 10x to 40x, allowing people to very quickly discover what's really interesting about a particular genome. Parabricks can operate on 3 types of genomes. The germline genome is the one you were born with. The somatic genome is one that's taken from a tumor, which may be mutating very rapidly as a cancer advances. And an RNA sequence is one taken from a cell that's expressing a particular protein. It's the dynamics of genome expression.

The plot at the right shows if we have this RNA sequence data, Parabricks has modules for analyzing, letting us visualize what's happening with gene expression. It takes a very high dimensional space where there's a dimension for each protein that might be expressed and projects it down on a 2-dimensional space in a way that gives the most insight. And then let's an analyst zoom in on one part of that space to really visualize what's happening that's relevant to what they're trying to discover.

So overall, NVIDIA Parabricks accelerates this process of taking sequence data and getting insight from it. The revolution in artificial intelligence has been enabled by GPUs, has also enabled a revolution of robotics. To date, most robots are very precise positioning machines. They are programmed to millimeter accuracy to move an actuator, be it a spray gun, a spot welder or some gripper to do a very repetitive task.

But with no interaction with the environment, it's completely open-loop. But with deep learning, we can build robots that perceive their environment, interact with their environment. And at NVIDIA research robotics laboratory, we're doing this work.

Let me show you some of the interesting things we're doing. To interact with this human, pick up a block, this robot needs to compute pass and control its motion in smooth ways that avoid obstacles. We've developed a new technology for this called Riemannian Motion Policies, that basically are able to express mathematically this complex motion problem in a way that's simple to solve in real time.

To make it more difficult than just stacking these blocks and simulation will create some artificial obstacles, these purple cylinders and the robot will very quickly be able to compute a path through those cylinders to grip the ball despite where it moves. Once we can move in unknown environments and avoid obstacles, what we want to do now is to understand how to manipulate unknown objects.

A lot of people have done work on manipulation, but it usually involves training the robot for a very particular object. And then if it sees an object it hasn't been trained on, it can't grip that. These videos show robots learning how to grip objects they've never seen before. We've trained the gripper using artificial intelligence in a way that generalizes. So it can produce good grip points for unknown objects. You can also learn to grip objects in the presence of obstacles.

Here, we want to pick up the cup, but we see that the sugar box is blocking it. So we'll change our target initially and command the robot to pick up the sugar box, move it out of the way. Again, computing the grip on that, using our Riemannian Motion Policies for the motion and come back, retarget the cup with the obstacle out of the way, it's able to compute a good grip for the cup, grab the cup and carry on with its task.

One thing which makes GPUs particularly applicable to robotics is the fact that we can train robots in the simulated world and then have them take what they've learned in the simulator world and apply it in the real world. This video of 4-legged robots, learning how to walk and deal with obstacles is a great example of that. We start with the 4-legged robot knowing nothing and using reinforcement learning and a curriculum, we train it to walk on progressively more difficult surfaces.

We start with a flat surface, then we have stairs of different steepness, surfaces with various obstacles, blocks and irregular obstructions. And after having mastered the simple surfaces, our robot moves on to the more difficult ones. It learns how to deal with each of these. And then in the real world, a robot trained in the simulated world was able to walk over a set of blocks in its path. It's able to go up and down stairs, all using skills that it learned in the simulated world, training its policy network to come up with a particular action in response to each state that it encounters. It learns different gaits for dealing with different types of stairs and other surfaces.

So being able to train in the simulated world and apply in the real-world makes GPUs all the more applicable because we're able to train many robots in parallel using the perils of the GPU in the simulated world.

Another application of AI is to autonomous vehicles. This is an extremely complex problem, involving many types of sensors, cameras, radars, lidars, real-time computation, where the vehicle is moving at high speed and has to predict the action of the other vehicles and pedestrians and other actors around it.

But the stakes here are high. 1.3 million people are killed each year on the highway. And by building autonomous vehicles controlled by GPUs running artificial intelligence that don't text while they drive, don't drive while impaired, and completely focus and have attention, 360 degrees around the car at all points in time, we can greatly improve the safety of driving on the highways. To do this, it's not simply a matter of deploying some AI in a car, but it's an end-to-end problem. That starts with data collection. You have to produce a huge data set of labeled data from all of the sensors, from cameras, radars, lidars, ultrasound devices, and then take all that data, curate it because not all data is equally important. You don't want to just put all data into your training sets. You want to select the most relevant data to train your models.

And then train models in the data center on big DGX SuperPODs to produce the trained neural network models that will be deployed in the car. But before deploying them in the car, we want to simulate them with the hardware in the loop simulation. It will take the actual AI hardware that will be in the car, running those models and synthesize what that model will see. Will generate synthetic video streams for the cameras, synthetic lidar data for the lidars, synthetic radar returns for the radars and validate that those models work properly in simulation.

Then the actual software that runs in the car has a number of components to it. There's drive AV, which is the autonomous vehicle part. It's a part that drives the car. And I'll talk about the many different models that go into making that. But it has components for perception they use all those sensors to sense the environment around the car, components for planning that decide where the car should go, whether it should accelerate or brake and prediction. To predict what the other vehicles, pedestrians, other objects in the scene are going to do, so it can plan ahead given likely courses of action for the other entities.

There's also a component for in the car, DRIVE IX. That has a camera that monitors the driver, sees where their gaze is, what they're doing. They can monitor their gestures. They can control the car through gesture rather than having to press buttons or knobs. And then there's DRIVE RC, which provides remote control. Should something happen where somebody has to take over remotely, we have a package that allows remote operation of the vehicle. To do this requires world-class neural networks for perception.

We have to detect obstacles, understand the distance to objects, the time to collision. We have a neural network that finds free space, spaces where the car should go and conversely spaces the car should not go. We do this not just for cameras, but also with lidar and radar. We have a neural network that projects paths to help the path planner find the best direction, to find signs, to do mapping. It will find high beams on the highway. It will help you park. Will handle different characteristics, weather, different intersections, traffic lights and the like.

It's a huge computational load. Multiple cameras, multiple other sensors, lidars and radars, each running many models to pull out many different pieces of information needed for the planning process to work forward. To carry out this huge workload, we have a variation of our Ampere architecture that specialized for the edge, and in particular, specialized for autonomous vehicles.

And we can size this depending on the workload of a particular autonomous vehicle. If all we need is driver assist, we have a 10 teraOPS, 5-watt version of our Orin and peer-based embedded chip that can handle that task.

For a level 2 autopilot, we have a 45 watt, 200 tops Orin AGX that can carry that workload out. For full level 5 robot taxi that gives full autonomy, we have a supercomputer that goes in the car.

It's a pair of Orins and a pair of A100s that provides 2 petaOPS of performance at 800 watts. And is dual so that there's redundancy. If part of the system fails, a subset of the sensors is pressed on the other part of the system to continue operating the vehicle, at least until it can be safely stopped.

So I'm going to get to the favorite part, my favorite part of this keynote, which is to talk about some things that are going on in my research laboratories. These are 3 projects from NVIDIA Research, and I'd like to emphasize at the beginning that these are research projects. They may or may not turn into products at some point in time in the future. They are not NVIDIA products.

The first thing I'm going to talk about is how we're going to continue Huang's law and continue doubling inference performance every year. We've been able to more than double it each year over the past 8 years going from Kepler to Ampere. We're working on a number of research projects that are looking at different alternatives, any one of which could continue this doubling for the next several generations.

To do this, we built a bunch of very efficient deep learning inference accelerators in the research labs. The photo on the left shows RC18. RC stands for research chip. This was done in 2018, and it achieves 9 TeraOPS per watt and is scalable from 0.3 TeraOPS to 128 TeraOPS. It's a ray of processing elements, 16 processing elements on each of these small chips, then assembled with 36 of the chips on a multichip module. And by building a very efficient inference engine, this is able to achieve higher inference efficiency than deployed products today.

We learned a lot from building RC18 and in particular, we learned that a lot of the energy that goes into inference doesn't go into the payload arithmetic operations, even with very efficient Tensor Cores, it goes to moving data around. So we built a system called MAGNet, which does design space exploration. It allows us to explore different organizations of deep learning accelerators and different data flows, different schedules for moving the data from different parts of memory to different processing elements to carry out the computation.

Using MAGNet, we're able to simulate deep learning accelerators that look basically like this. We have a global controller controlling an array of processing elements that feed data between their local memories, a global buffer that's on the chip and DRAM that's off chip. Within each of the processing elements, there are buffers to hold weights and input activations, vector multiply accumulate units, MACs. Each of these vector MACs carries out multiply accumulate over a vector that can vary in length from 8 to 32.

And then we have from 8 to 32 lanes of these MACs. We have anywhere between 64 and 1,024 multiply cumulates happening per cycle.

The output of these multiply accumulates drop into the output activation unit at the bottom, where they're accumulated in the appropriate ways and stored in the output activation buffer waiting to be written back to the global buffer. The figure on the right shows the details of a vector MAC, just a bunch of multipliers going into a big accumulator, doing essentially a dot product operation.

Most of the popular data flows are ways of organizing the choreography of data on these chips, involves either holding the weight stationary while the input activation is run by and the output activations are sequenced. And as shown in the little high chart on the bottom, this results in most of the energy being spent in the accumulation buffer, 55%. And less than 1/3, only 29% is actually in the arithmetic of carrying out the inference.

And alternative then, the problem is that the outputs are not stationary, you're spending all your time and accessing that accumulation buffer is to hold the output stationary. And all this does is push the problem to the input side. And all the energy is now spent in the weight buffer because you have to constantly access new weights. So seeing this problem, we innovated and decided we had to add a level of storage, we call collectors, weight collectors and accumulation collectors, which then adds another level to our data flow, another level to the nested loop, which is carrying out this convolution in a deep neural network.

And by doing this, we're able to move the computation, so that rather than only 1/3 of the energy being in the arithmetic and the payload as it were of the computation, we're able to get 60% or more of the energy and the payload of the computation. And this particular unit was able to carry out inference at 29 TeraOPS per watt. This was published at the end of 2019.

And we've since advanced this architecture to the point that while it hasn't been published yet, we now are achieving 100 TeraOPS per watt for inference. So we're continuing this evolution of Huang's law, continuing to more than double inference performance each year. This isn't an announced product, but we expect that these techniques will be incorporated into future versions of Tensor Cores and future deep learning accelerators in our embedded chips for automotive and robotics.

Now as we scale up, we need to connect multiple GPUs together, and we do this with NVLink and NVSwitch, technologies that were developed in NV Research. But as we look ahead, we think that it's going to be increasingly difficult to continue to double the bandwidth of the SerDes that we use in NVLink each generation.

We're currently operating at 50 gigabits per second per wire pair. We can see our way going to 100 perhaps to 200. Beyond that, the waters are very murky. So looking at an alternative technology to actually signal out of our GPUs and in and out of our NVSwitches using light, using photonics. This shows a concept, where on the left, we show our existing DGX with electrical signaling. And on the right what an optical DGX might look like, we packaged 2 GPUs on a vertical card that then get interconnected with fiber optic bundles between the cards.

The part on the right side of each card are the light sources that provide multiple wavelengths of light, sort of the optical power supply as it were for the optical engines that modulate this light to signal at very high data rates. We see our way with silicon photonics to have many wavelengths for fiber. Operating each of those wavelengths at anywhere between 25 and 50 gigabits per second for aggregate data rates between 400 gigabits per second and a few terabits per second for fiber and a very conservative number would be -- we would do the signaling at 4 picojoules per bit, where the electrical signaling is more like 8 picojoules per bit.

And where the electrical signaling is limited in reach to about 1/3 of a meter, about a foot, the optical signaling would have a reach of 20 to 100 meters. So we can connect much larger systems with a single hop of NVLink rather than having to constantly repeat -- is it necessary with the electrical signaling. The way we plan to do this is with a relatively new optical technology called dense wavelength division multiplexing.

We start with an optical power supply, a comb laser source shown on the left, that produces many different colors of light. These colors are spaced very close together, perhaps 100 gigahertz spacing between the different colors. We feed this comb of laser light onto our optical engine chip on the transmit side, and it comes along of optical bus, where it gets modulated by ring resonators. Each ring resonator is tuned to one of these wavelengths and can either pick off that wavelength or allow it to pass.

So we can modulate a bit stream of, say, 25 gigabits per second on each of these colors of light. And so if we have, say, 32 colors of light at 25 gigabits per second each, then that gives us 800 gigabits per second aggregate on the fiber going out.

At the receiving side, we again have a set of ring resonators that are being used to pick off 1 color and feed it into a photo detector that goes into a transimpedance amplifier to feed out that channel for us to receive. This will be packaged with a GPU sitting on an interposer that communicates over an organic package to a photonic integrated circuit that contains those ring resonators and the wave guides to run the light around, and an electrical interface chip that basically receives the short reach interconnect from the GPU and controls the ring resonators on the PIC to modulate the light.

At the receiving side, again, there's an EIC and a PIC, but for receiving the part of the EIC that's important on the transimpedance amplifiers and the PIC includes the photo detectors.

So here is an artist conception of how DGX systems using this might look in the future. We can connect a very large GPU tray with these. In this case, we have 18 GPUs per row and 9 rows for over 160 GPUs connected. And their inputs and outputs all come out in a big fiber optic bundle. The NVSwitch tray on the right shows that we have a number of NVSwitch cards, each with 1 NVSwitch and optical power supply, and they're connected to the blue fibers that are sort of inputs, the orange fibers that are outputs, and the gray fibers are interconnecting the individual NV switches within this tray to form a co-network that allows us to route data from the input to the output.

So the final research project I'm going to tell you about is Legate. One of the key issues that faces us with GPUs is that there are many more applications that could benefit from the accelerations GPUs give, then we can take the time to code in CUDA even using all the libraries we have.

So we're constantly trying to make this programming process simpler. Now a lot of people already do in numeric analysis using Python, and particularly the Nanpy library within Python. And so we've developed a package called Legate, which sits on top of a data aware task scheduling run time system we have called Legion. And this allows us to take a Python program and run it transparently from a Jetson Nano all the way up through a DGX SuperPOD without changing a line of code.

All we have to do is take our original Python code and change the line that says import Nanpy as NP to say, import Legate that Nanpy as NP. That loads the library and everything from there on is done automatically. So we can run on a Jetson Nano, on a single A100 on a DGX A100 as shown here or a DGX SuperPOD.

Let's look at the results of what happens when we run this. Here's a relatively simple example of Jacobi iteration, the code is shown at the right side. And to start off with, let's look at what happens when we run on a single GPU. The real relevant mark here is the orange square, which is our current best library, CuPy, for running this Jacobi iteration. And what you see is that Legate running on the GPU actually runs slightly faster, that's the green line. If we run on DAS, which is an alternative task schedule, you see as we scale up the number of GPUs, performance falls off pretty rapidly.

Whereas if we run Legate, and this is Legate running on a CPU since the comparison was to Dask running on a CPU, you see that the performance remains flat because Legate is able to do a better job of parallelizing the task and keeping all of the GPUs busy. When we run Legate on the GPU, there's a little bit of a falloff in performance. I should say here that these curves show what is called weak scaling. Every time we increase the number of GPUs, we're increasing the problem size by a measured amount. So perfect speed up would be a straight line. And here we're a straight-line out to a few doublings of GPUs. Then we fall off towards the end because Legate is unable to get perfect parallelism out to a very large number of GPUs, something we're still working on.

So let me wrap up. At NVIDIA, we're constantly working to build the world's fastest computing devices and then build the software that focuses those computing devices on the world's most demanding and important problems. It all starts with our Ampere GPU, and we can scale the performance of this Ampere GPU from this A100 down to our small embedded Jetson devices, and we can scale it up through the DGX, the DGX SuperPOD up to some of the world's fastest supercomputers. And then we can focus that power at any scale from the Jetson to the supercomputers under problems from graphics, artificial intelligence, health care, video analytics, autonomous driving and robotics by having a huge set of libraries and then vertical stacks built on top of those. I've shown you a little bit of a glimpse of the future via what we're doing on our research labs. And it's really just the tip of the iceberg. We're doing very many exciting things, and I think the future is going to be a very exciting time as we're able to bring some of these to fruition, build even more powerful computing devices and apply them to a wider array of problems to make people's lives better. Thank you very much, and have a great day.

(Break)

Raymond Teh

Hello. Welcome to GTC China Nvidia Executive Panel. My name is Raymond Teh . I'm responsible for NVIDIA's Asia Pacific sales and marketing business, and I'm your moderator for this panel session. Today, we will hear from 4 NVIDIA leaders as they describe the company's latest breakthroughs and the relevance for the China market.

(foreign language)

Ladies and gentlemen, please allow me to introduce our 4 speakers for this panel. First speaker is Mr. Jay Puri, Executive Vice President of Worldwide Field operations. Jay is responsible for global sales and regional marketing for all NVIDIA's products and services. Our second speaker is Mr. Greg Estes, Vice President of Corporate Marketing and Developer Programs. He is leading the company's efforts engaging with more than 2 million developers and several thousand AI start-ups. Our third speaker is Ms. Kimberly Powell, Vice President Health Care, responsible for NVIDIA's Health Care Business globally. And our fourth speaker is Mr. Ashok Pandey, commonly referred in China as "Pandey". Pandey is Asia Pacific's Vice President responsible for operations and partners. We have a very diverse background and certainly a very -- a wealth of experience. We have a lot to cover today, so let's get started.

QUESTIONS AND ANSWERS

Raymond Teh

Jay, if I may, let me start with you with the first question. How important is the China market to NVIDIA?

Ajay K. Puri - *NVIDIA Corporation - EVP of Worldwide Field Operations*

Well, (foreign language). So Raymond, China, besides just the size of the market, is extremely important to NVIDIA. It's so strategic, and it's one that we make a very significant investment in. NVIDIA has been in China for more than 20 years. We first got started with our PC gaming platform, GeForce, there are millions of GeForce fans in China and it's absolutely fantastic. We have a terrific ecosystem of partners, AICs, distribution partners and so on, game developers. Some of the most important trends in PC gaming have actually come out of China, things like free-to-play, iCafes, esports and so on. So one really needs to understand the China market to be successful worldwide. So China is extremely important to NVIDIA.

And I want to thank all the GeForce fans. We just introduced our new lineup of RTX 30 products, which are just absolutely fantastic. I hope you guys are going to try them out. And then, of course, 10 to 15 years ago, we realized that Moore's Law was reaching its limits and we pioneered this new model of computing called accelerated computing, started out in high-performance computing and then -- for mostly scientific applications. But now over the last few years, this platform has been adapted for Artificial Intelligence and Data Analytics. And I have to say, I think running the AIPAC you fully understand this. China has been at the forefront of adopting AI to provide a competitive advantage to their industries, the universities are doing some really leading edge research, some of the most important researchers on AI in the world are in China. The startup ecosystem is so vibrant. So what's really verifying is most of all the work in AI in China is being done on NVIDIA's AI computing platform.

So to say that China is important to NVIDIA is a little bit of an understatement. China is extremely important. We try to learn from the China market to improve our platform, so we can continue to provide what our partners and customers need over there. And I would like to take this opportunity to really thank them for entrusting -- putting their trust in NVIDIA, and we will continue to innovate and deliver what you need.

Raymond Teh

Thank you, Jay. That was a great response to my question. Jay talked about our business. Greg, but Develop GDC is our developer conference. Can you share with us what you're doing with the developers?

Greg Estes - *NVIDIA Corporation - VP of Developer Marketing*

Well, I'd love to. And let me first also welcome everybody to GTC China. This kicks off another amazing event for us for our developers. We're going to have tens of thousands of people who join us here at GTC China with more than 200 talks of varying kinds from all different parts of the marketplace and from researchers and from core developers and from students and others.

China is our most -- in many ways, our most important developer market. We have more than 400,000 registered developers in our program in China, which is larger than any other country in the world. And so China is at the forefront of what we do. And those registered developers in our program, of which there's about 2.25 million of them right now all around the world. They get access to our SDKs. We put a lot into not only CUDA, but 100 other SDKs that we have of everything from health care, which Kimberly is going to talk about, to robotics, to financial services to really every market that we're in. And GTC is where all of that comes together because you get not only the developer sharing their most important work and their most important research on GPUs, but you have the rest of the ecosystem in the industry, too, right, all of the platform providers. And in brand new areas, including what the start-ups are doing. And Jay touched on that a little bit. We have hundreds of startups in our startup program Inception in China, and they are here and well represented as well. And that's kind of great because many times, the very leading edge work is coming out of start ups, right? And is coming out of the universities and our connection to the universities in China is better than really almost any country in the world, really any country in the world. And so we're very proud of that connection. Super glad that everybody is here and joining us and sharing their work.

Raymond Teh

Let me switch to a subject that is affecting all of us. And I'm talking about the global pandemic, COVID-19. Kimberly, from your perspective, can you tell us how the AI and accelerated computing contribute to global response to the COVID-19 pandemic?

Kimberly Powell - *NVIDIA Corporation - VP of Healthcare*

Yes, sure, Raymond. I hope all of you are staying healthy and safe, and coming back to health and being stronger than ever. The pandemic is a defining moment in the global health care industry. It marks the biggest threat to the global health in the last century. And so our race to track, test, discover test, discover vaccines and therapies has catalyzed the world really to use all of the modern technology available to us. And we know, we all share here at GTC that AI is the biggest technology for us of our lifetime. And COVID is really a supercharging moment, and I see the AI healthcare really coming upon us. And it's only enabled by everything Greg and Jay and Raymond have already touched on. It's a complete ecosystem approach, and we're so delighted that so many of our academic partners, our startup partners, our healthcare industry partners in China have harnessed every piece of technology to really fight this fight.

And it all starts with genomics. China is an industry leader in genomics. Alibaba and their genomic services offered GPU-accelerated gene comparisons so that governments and health officials could really understand the transmission and evolution of the virus. And then China was very quick to realize that medical imaging, something NVIDIA is very passionate about with our Clara imaging computational platform, that imaging can be an extreme source of understanding not only the detection of coronavirus, but also the treatment of coronavirus. Leaders like Ping An, United Imaging; startups, Infervision and Shukun, they put their medical imaging COVID AI technology into thousands of hospitals across China so that these tired, overworked frontline workers had the AI technology at their fingertips to make the best choices for their patients and get their patients back to health as quickly as possible. So this is an absolute defining moment. The AI health care era is here. It was supercharged, and 2021 is the beginning of this new era. It's a tragedy, but also a place where we can understand how important technology is and that the global ecosystem has come together to really fight this fight.

Raymond Teh

Great. Wow, thank you, Kim. It's fantastic to hear of all the good things that we are doing with -- from genomics to medical imaging, et cetera. We all look forward to AI improving health care.

Jay, I want to bring you back to the next question I have for you. NVIDIA recently announced plans to acquire Arm. Can you add some color to that? What were the reasons to do so?

Ajay K. Puri - *NVIDIA Corporation - EVP of Worldwide Field Operations*

Yes, Raymond, I'd be glad to -- yes, an extremely important development for -- not just NVIDIA, but for the entire industry. First of all, Arm is an amazing company, right? I mean, they sell more CPUs than any other company in the world, 22 billion -- that's with a b -- 22 billion Arm CPUs are sold every year. And they pioneered the sort of IP licensing model, the product is just fantastic. It's the most energy-efficient core. It's also a very high performance. So when ARM became available, of course, we are more than delighted to have them be part of NVIDIA. But that is only the beginning of the story. Really, I think by putting ARM and NVIDIA together, we are going to be able to make a huge contribution to our industry. ARM, as I said, has been extremely successful, but most of their success has been in the mobile space and the embedded space. But there is a lot more opportunity here. One of the most obvious places where Arm should be playing a role is in the cloud data center PC environment. And they're starting to address that market. But frankly, it's a really something -- a tough thing to do at this point. When they got established in the mobile space, it was basically greenfield, right? Nobody else quite had the energy-efficient technology or the open licensing model. And so the whole ecosystem developed around them, and they are very successful.

The data center PC cloud space is a completely different story. The x86 architecture completely dominates that space at this point, right? And so, of course, because Arm's (inaudible) is great technology, we are seeing success in that space. Amazon has the Graviton2 for the hyperscalers that they're using internally. The fastest supercomputer in the world, from Fujitsu, the Fugaku too -- #1 on the top 500 -- uses Arm technology. And then most recently, Apple has said that they're going to be using Arm in the M1 family of products, so all their Macs and so on.

So the technology is fantastic. But when people try to provide it as a merchant CPU, it's been very difficult. In fact, Qualcomm tried taking out a [mere footsie] Arm CPU on the general purpose market centers. Broadcom tried it, Marvell tried it. But all of them in the end, have not really been successful and redirected their programs. And again, it is not because of technology, the technology is fantastic. It's because there is just so much ecosystem that has been built around x86, it's hard for another architecture to come and take a significant acquisition regardless of the merits of the technology.

Now, as you know, going forward, the most important workloads in the data center are around accelerated computing and artificial intelligence and so forth. And NVIDIA's platform is extremely well-established in that space. And we have a full [staff] that is available. We have all the partners that are necessary. The ecosystem is huge: more than 2 million developers and lots of startups, lots of industry research going on, university research going on. So we have the platform that is -- what is necessary as to what is required to be successful in the data center going forward. And once I think Arm is part of NVIDIA and we both put our focus on making Arm successful in the data center, we're going to get this done. And then there will be a viable alternative to x86, not just in the mobile space, but for the data center and the PC and the cloud and -- all of it. And that is frankly extremely good for the whole industry. It's good to have competition: it drives innovation. And I'm really looking forward to it. And in particular, I think it's good for China.

Raymond Teh

So NVIDIA continue on with Arm's IP licensing model?

Ajay K. Puri - *NVIDIA Corporation - EVP of Worldwide Field Operations*

Absolutely. I mean that is 1 of the fundamental reasons why they have been successful. And so we completely believe in that. We admire the model. I think it's a fantastic business model, and we will 100% continue with that. In fact, I'm quite hopeful that once Arm is a part of NVIDIA, and we understand exactly how they do it, we will be able to take more of our technologies, both in the GPU and networking space, and make them available to the world using this licensing model.

Raymond Teh

So Jay, given the current geopolitical climate, some Chinese companies have expressed concerns that Arm's acquisition by NVIDIA may restrict their ability to access Arm technology. My question to you is, should they be worried?

Ajay K. Puri - *NVIDIA Corporation - EVP of Worldwide Field Operations*

Yes. Raymond, I'm glad you're asking that because there is a little bit of a misunderstanding about how export control laws work. The export control laws work based on where the technology is invented, the origin of the technology. They are not so focused on who owns the technology. So whether Arm is owned by a Japanese company like SoftBank as it is today or owned by NVIDIA in the future, that does not actually impact the export control regulations. And so of course, all of the Arms key technology today is invented in Cambridge, England, and we are -- we've already committed that, that's going to be the center of development for future Arm technology when they become part of NVIDIA. So there really will be no change in terms of export control regulations once ARM is part of NVIDIA.

Raymond Teh

Thank you, Jay. That was a great answer. And I'm going to ask Pandey the next question in Mandarin.

(foreign language)

Ashok Pandey

(foreign language)

Raymond Teh

(foreign language)

I want to switch from our data -- we talked a lot about data center and Pandey gave a very comprehensive answer on our CSP. I want to switch to another topic, which is very important, graphics -- and it's our heart and soul.

Greg, what are the most important new advancements for developers in gaming and graphics?

Greg Estes - *NVIDIA Corporation - VP of Developer Marketing*

Well, it's pretty obvious at this point that real-time rate tracing has been already started and is changing the industry to the point where I'm not sure that you can say that you're doing serious work as a technology company in graphics if you not put a lot of energy into ray tracing, and we've certainly been leading that over the last few years. Beginning with our RTX line and the work that we put in to having specific cores on the GPU called RT cores to accelerate ray tracing. And it's changed gaming, right? It is -- you're seeing it be adopted in the biggest games, in Fortnite, Justice in China, Cyberpunk 2077 coming up, the biggest games taking advantage of our latest technology is just fantastic and more and more, almost literally every day. And it's changing more than gaming, too, by the way, because, of course, there are other markets, particularly here in China where architecture is -- continues to be such an important marketplace to be able to visualize spaces in a way that uses the real reflections and light in the way that they physically act in the world that is important for so many different marketplaces, not just gaming, but particularly gaming, right?

The second thing that we're seeing is the adoption of artificial intelligence for graphics. One of those implementations that we're very proud of and we think is some of our most important work is DLSS, deep learning super sampling. And for those of you that aren't familiar with that, it's essentially using AI to render at one resolution and then use AI to make that at a much higher resolution, which, of course, increases your frame rate and improves game play. And the way that, that works is that there are models that are trained to use artificial intelligence to fill in pieces of information that would have been there if you had higher resolution and it gives you an incredibly sharp view in your monitor and extremely good gameplay because, again, we're doing the core rendering at a smaller resolution and then doing the super sampling using deep learning to do that. And again, we're seeing more and more games all the time take advantage of that.

And together, you see the trend here is to put the highest level of realism possible that you can for this and that gives such great enjoyment to people who are seeing shadows and reflections. And these things affect your gameplay, too, right? You can do things in gameplay that you can't do if you don't have this level of physical reality. And of course, that works for training systems and as I said, architecture and in media and entertainment and everything. So these core technologies that are being driven certainly out of gaming, they're affecting other markets well. And China has been 1 of the leaders in adoption of this as well. The Chinese games are -- I think some of the early adopters here. I mentioned Justice is a great example, and we're super proud of the work that our developers are doing, many of whom are here presenting their latest technology at GTC.

Raymond Teh

Great. Wow. Certainly, this year is a strange year for us and many of us are staying at home. We are seeing many gamers enjoying the new features in ray tracing and DLSS.

I want to switch back to the pandemic. Kimberly, can I ask you what is the biggest challenge or change in the health care industry as a result of the pandemic?

Kimberly Powell - *NVIDIA Corporation - VP of Healthcare*

Raymond, we're at a time in history that has never happened before, not only because of the pandemic, but actually because of all of the other healthcare technologies that have been invented over the last decade. Let me tell you something that most people don't know. Today, we can create more biomedical data than ever in our history. In 1 quarter, we can create more biomedical data than in the 300 year history of most of the pharmaceutical industry. This is creating with data, more data than we've ever had, AI and data center computing, that we have today. It's absolutely what I call the perfect storm for a computational global defense system. Think about it. If we can create more biomedical data today, we have the technology of artificial intelligence, and we have the computing capacity of things like what Pandey was mentioning, our DGX SuperPOD. We literally have the perfect storm. And governments all over the world are realizing this. They're realizing that our pharmaceutical industry that is still suffering from too long of a time frame to create a new therapy or vaccine, sometimes on the order of 10 years, it costing $2 billion and still having only a 10% success rate, they realized with this pandemic that we need to take more investment into the drug discovery industry.

And NVIDIA has been working on this problem actually for over a decade. Drug discovery really calls upon every single computer science domain that there is and that NVIDIA is world-class at. It starts with graphics to be able to visualize biology and chemistry. It has simulations so you can understand how biology and chemistry are interacting. And it has to use artificial intelligence so that we can tackle much larger problems and simulate much longer timescales and work on much larger data sets. And so we're at this critical time where we can, in fact, create a computational defense system. You can see that in academia some fantastic work just came out of Tsinghua University and Zhejiang University, where they were able to use cryo electron microscopes who can create terabytes of data in a day. They were -- they used that technology to create a complete picture, molecular architecture of the coronavirus. And when you create that data, they put it into a public database so that every other researcher or industry researcher can take that and start to search for vaccines and antiviral drugs. They published this work very recently in Cell, which is 1 of the most prestigious life sciences journals. And that work and the fact that the ecosystem can generate that data, create such insightful data, digitizing biology. We've never been able to digitize it at this accuracy -- at this scale before and contributing back to the entire industry at large. That's the global defense system. And it's a very fortuitous circle. So governments realize now that investing in academia, super computing, in their own drug discovery market is vial.

And so NVIDIA recently announced our own Clara discovery platform. As I said, we've been working on it for a decade. Everything Clara is, is domain specific. We have Clara Parabricks for Genomics. We have all of our life sciences applications in the area of cryo-electron microscopy, molecular docking, molecular dynamic simulations, Clara for Imaging. And recently, we also published our own state of the art natural language models and biomedical language models. A complete new language is inherent in the healthcare data that we just talked about. And so we're right at the cusp of this brand-new drug discovery process that is completely fueled by artificial intelligence and computing. And so I believe that forevermore changed is the idea that we have the ability -- we have essentially a time machine. Data, AI and compute is essentially a time machine and that's exactly what you need in a pandemic, so that we can accelerate our discovery of therapies and vaccines. And every government around the world and all of the pharmaceutical industries are taking a new approach at the process of drug discovery and investing in the R&D. And with Clara Discovery and DGX SuperPOD, we literally have it all packaged up into a complete AI-driven drug discovery infrastructure. And the world is taking to it.

We're going to build our very first supercomputer in Cambridge U.K. that is dedicated to biomedical research. We call it Cambridge-1. It's going to be the U.K.'s fastest supercomputer, completely dedicated to this cause because we know we're at this perfect storm and that we know we have no more time to wait for a vaccine and antiviral drugs to help us in this fight against COVID-19.

So the pharmaceutical industry is forever changed. And I think that's a good thing. It made us realize that we can now digitize completely biology, and we can couple that with all the other computer science domains out there to really build this time machine and this global defense system for all the future pandemics.

Raymond Teh

Thank you Kimberly for sharing with us so many use cases and also giving us an insight into the Clara platform. We certainly are so excited that NVIDIA's AI technology can help the healthcare industry in so many ways.

Let me switch to one of the most promising and growing businesses in China. And this is in our consumer internet space. It is the broadcasting business. And let me switch to Mandarin to ask Pandey.

(foreign language)

Ashok Pandey

(foreign language)

Raymond Teh

(foreign language), Pandey.

So Greg, I want to ask you the next question. In China, there are many AI startups. An important part of GTC China is our work with these AI startup companies. What are we doing with startups in China?

Greg Estes - *NVIDIA Corporation - VP of Developer Marketing*

Well, you think about the companies that are here at GTC, they're some of the most innovative companies in the world -- Baidu, Alibaba, Tencent, Xiaomi, ByteDance, DiDi and on and on, on just great innovative companies, and they were all startups at one-time or another, right? And that's 1 of the reasons why we've put such a strong investment in focusing on helping startups to grow, giving them the best access that we can to our technology. And in fact, we have more than 800 startups as part of our startup program, which we call Inception, in China. Right?

And what do we do for them? Well, the first thing that we do is we give them access to our technical resources, right? They get a personal connection with NVIDIA, if you will. And that's important to startups that are trying to take advantage of that absolute best technology. And a lot of times, it's much better to talk to a human than it is to try to get it out of documentation and others, right? The second thing that startups want to do is they want to be able to scale, right? A lot of them start with GPUs on-premises and then want to scale up into the cloud, and we have programs to make those GPUs more available to them. We have grant programs and discount prices, too, so that they can start their company and get going on there. And then if you think about it as a startup, what is the other thing where we can be most helpful to them? A lot of times, it's giving them visibility. So we help them with marketing programs. We do success stories with them, and we put NVIDIA's marketing muscle behind raising their visibility out to the world. And a lot of times, part of that is we have these special connection sessions with VCs. So we'll take venture capitalists and sometimes big companies too, and we'll do the matchmaking to put them together so that we're, again, raising the visibility of these startups to other companies or venture capital firms that could be interested in investing to them.

So there's a technical component to what we do. We try to help their business get going. We raise their visibility to others through marketing, and we give them a deep connection to us through training through our DLI program and other technical services that we have. And we think all of that together, we have 1 of the best and most comprehensive startup programs in technology. We have almost 7,000 startups worldwide in our program. Think about that. 7,000 AI companies all working with NVIDIA today. It's a wonderful program. And so for those of you that may be watching this that are with startups that may be interested, I encourage you to check out the Inception program. And if you're not a startup, but would like to get tapped into this community that's doing so much innovation, then contact us as well, and we can help connect you to some of these really great startups and GTC is the perfect place to do that.

Raymond Teh

Thank you, Greg. We really touched on a number of topics today. We touched from healthcare, to cloud service providers to startups to gamers and also to the whole entire business.

Jay, next -- my final question goes to you. Our route to market is through our partner ecosystem. Our China partners are so important for NVIDIA. I would like to ask you something about how we're working to promote AI with them? One of the recent announcements made by NVIDIA is the DGX A100 and the DGX SuperPOD. They are indeed a great AI supercomputing platform that is becoming very popular in China. My question to you is how do you differentiate NVIDIA's own DGX A100 supercomputer with the many forms of OEM branded GPU computer service made in the market? And what is your strategy and positioning between these OEM service and NVIDIA's own branding of DGX.

Ajay K. Puri - *NVIDIA Corporation - EVP of Worldwide Field Operations*

Sure, Raymond. First of all, let me just say what you just said, which is our partners are extremely important and our primary go-to-market for NVIDIA is through partners. There's absolutely no question about that.

So let me explain to you why we do DGX. AI is evolving at a breathtaking rate, right? I mean, Kimberly just told you what's happening in healthcare, and we're just scratching the surface. We have just come so far -- I mean it's only 5 years ago we were recognizing cats and feeling really great about it. And today, we are using AI for natural language understanding, conversational systems, recommendation systems, I mean the impact that AI is having, and can have, is just absolutely amazing. And so the data sets are getting huge -- believe it or not, the amount of AI computing performance that is required with some of these new neural networks like (inaudible) is increasing at 2x every couple of months. I mean, it's just absolutely hard to believe what's going on. And of course, NVIDIA as the leader in AI has to keep up with it. In fact, we have to drive this.

And so we do a lot of innovation, all the way up and down the stack starting with the GPU and then the boards around it, the systems, all the software going all the way up to the application frameworks like Clara and Jarvis and Merlin and so forth, right? So we have to innovate everywhere. And now because of the scale at which it has been done, in fact, frankly, it's not even possible just to stay at the server reference architecture -- we have to look at data center scale computing, and that's one of the reasons why we bought Mellanox. And now we have data center scaled products with DGX SuperPod and so forth. NVIDIA with ourselves, we have Selene as the DGX SuperPOD that we use, which is an AI supercomputer, #5 on the top 500 list overall, but definitely the fastest AI supercomputer that exists in the world. So we are making a huge amount of investment and we have to stay -- be the leader in AI.

But there is 1 other thing that everyone really needs to understand -- NVIDIA is a very open company. And so when we do all of this R&D at all levels, we make it available to our partners at different levels of integration, okay? So yes we designed DGX. And obviously, we're going to do all this work -- but there are some customers that want to have a very direct and deep relationship with NVIDIA and they buy our DGX product or they buy DGX SuperPOD. But by and large, we take what we do and then we give it to our partners so that they can build their own systems around this architecture. Not only do we provide this reference architecture to them, but we then work with them to certify their software stack all the way up to NGC and the containers in NGC that have all the applications, such as those for [Pyra] and other [verticals].

We look at the -- help the OEMs design these systems based on our reference architecture, we make sure they are fully compliant, make sure the performance is excellent and then we brand them as NVIDIA certified. Okay? And the market is huge. There's all kinds of requirements. There is no way NVIDIA can or wants to fulfill all of these. As I said, we really expect our OEMS and other system builders to deliver the -- NVIDIA's AI platform to the market to help us all succeed and to drive AI forward. And I'm happy to say, I think we're doing a fantastic job. We have some really great partners in China, Inspur and Lenovo, the new H3C, Netrix. There's a long list of partners that are investing in our platform and are providing solutions to the customers that they really seem to appreciate because our business is growing at a fantastic rate both for us and for our partners, and I'm really very grateful to them for working with us.

Raymond Teh

Thank you, Jay. And ladies and gentlemen, that's all the questions that I had. We tried to cover everything -- a broad, a wide area of our business that covers our partners, our customers, our industries, our ecosystem, et cetera. I hope you find the session informative. I want to thank Jay, Greg, Kimberly and Pandey for joining me in providing this session for all of you, and thank you to all our speakers for providing their great insights.

And please do not forget to get in touch with any member of the NVIDIA team if you need more information or if you have any follow-up questions. Please also take the time to view the 200-plus sessions and talks at GTC China given by experts from the industry, from our customers, our partners and from NVIDIA ourselves.

Thank you again. And on behalf of all the speakers and on behalf of NVIDIA, I hope everybody has a great GTC. So that's it for now. Thank you, and enjoy GTC China. Thank you.

DISCLAIMER

Refinitiv reserves the right to make changes to documents, content, or other information on this web site without obligation to notify any person of such changes.

In the conference calls upon which Event Transcripts are based, companies may make projections or other forward-looking statements regarding a variety of items. Such forward-looking statements are based upon current expectations and involve risks and uncertainties. Actual results may differ materially from those stated in any forward-looking statement based on a number of important factors and risks, which are more specifically identified in the companies' most recent SEC filings. Although the companies may indicate and believe that the assumptions underlying the forward-looking statements are reasonable, any of the assumptions could prove inaccurate or incorrect and, therefore, there can be no assurance that the results contemplated in the forward-looking statements will be realized.

THE INFORMATION CONTAINED IN EVENT TRANSCRIPTS IS A TEXTUAL REPRESENTATION OF THE APPLICABLE COMPANY'S CONFERENCE CALL AND WHILE EFFORTS ARE MADE TO PROVIDE AN ACCURATE TRANSCRIPTION, THERE MAY BE MATERIAL ERRORS, OMISSIONS, OR INACCURACIES IN THE REPORTING OF THE SUBSTANCE OF THE CONFERENCE CALLS. IN NO WAY DOES REFINITIV OR THE APPLICABLE COMPANY ASSUME ANY RESPONSIBILITY FOR ANY INVESTMENT OR OTHER DECISIONS MADE BASED UPON THE INFORMATION PROVIDED ON THIS WEB SITE OR IN ANY EVENT TRANSCRIPT. USERS ARE ADVISED TO REVIEW THE APPLICABLE COMPANY'S CONFERENCE CALL ITSELF AND THE APPLICABLE COMPANY'S SEC FILINGS BEFORE MAKING ANY INVESTMENT OR OTHER DECISIONS.