GPU Technology Conference 2018

Company Participants

- Fausto Milletari, Unknown
- Jensen Hsun Huang, Co
- Justin Ebert, Unknown
- Mark Daly, Unknown
- Ryan Olson, Unknown
- Steven Parker, Unknown

Presentation

Operator

(presentation)

Ladies and gentlemen, please welcome NVIDIA Founder and CEO, Jensen Huang.

Jensen Hsun Huang {BIO 1782546 <GO>}

Hi. Welcome to GTC. Welcome to Santa Clara. GTC is the GPU Computing Developers Conference. We do it for you, for all of you, whose work are simply impossible without a supercharged computer. Thousands of NVIDIA employees and our partners worked incredibly hard to pull GTC together for you. Let's give them a round of applause. Thank you, guys. I love you, guys. Thank you.

We have so much to talk about today. We have a lot of new products to show you. We're going to talk about amazing graphics, we're going to talk about amazing science, amazing AI and amazing robots. We have so much to cover. So let's get going.

Computer graphics is the driving force of the GPU. It is computationally insatiable. Recreating virtual reality is one of the most daunting computing tasks we know. And yet, on the other hand, it is an enormous industry. We want to visualize information, visualize experiences in all kinds of markets. As a result, the technology that we bring to bear and the size of the market combines into a gigantic R&D budget. It is the driving force of GPUs. Since 1979, when Turner Whitted wrote the paper on recursive ray tracing; in 1986, when Jim Kajiya described the rendering equation, the computer graphics industry and computer sciences all over the world have been pursuing this holy grail, this dream of recreating photorealistic images. This is the way film is done.

They trace rays through a scene as it strikes a surface, calculates the absorption, the reflection, the refraction, the luminescence, accumulates light from all the different sources and reflections and calculates the reflectance functions based on the materials that it strikes, accumulates all of that information and generates a photorealistic image. Well this is the way the film industry does it. And it takes thousands and thousands of CPUs and thousands of servers in order to calculate and compute each one of these frames. One CPU would take hours to compute one frame. And a movie has, as you know, hundreds of thousands of frames before they could create the final film. Well when you look at this image, it is utterly beautiful. It's just super, super hard to produce.

Well the computer graphics engineers and architects at NVIDIA isn't sitting still either. This is what modern computer graphics looks like. This is what a video game would look like. And in fact, it looks pretty amazing. The difference here is what was -- what I showed you earlier was generated one frame every many hours versus what's going on here with real-time rasterized computer graphics with programmable shading that we invented some 15 years ago, produces these images at 4K resolution at 60 frames per second. From one frame in many hours to 60 frames per second, that fundamental difference has been a gap we've been trying to close for literally 4 decades, literally 4 decades.

And in fact, it is incredibly hard. We look at this image today and I can't tell you how proud I am. I mean, this is modern computer graphics. Look how incredible it looks. And yet, there are so many places we like to improve. For example, we use all kinds of special techniques to trick your eyes to see reflections, to see shadows. Ambient occlusion is where light accumulates on a surface and because different surfaces are further away or occluded from the ambient light, it's darker. It's not a shadow. It's just simply wasn't illuminated, ambient occlusion. We use all kinds of tricks with our z-buffer, depth tricks, to calculate in our programmable shaders to figure out which one of those pixels are to be darker than the others. We use all kinds of tricks.

Game developers will even render the scene completely in advance. And for all the places where there's static lights, we will pre-bake the light. It's called light baking. But as a result, as you can see. So long as things don't move very much, the lighting conditions are relatively well reproduced. And it's quite beautiful. This is what ray traced image would look like with global illumination. Notice those little details, those little details that bring a scene to life. So ray traced global illumination where light is simply emitted from everything that has a light source or is reflected, defused into reflection where even objects that are matte in color could actually emit light as light reflects off of it.

Global illumination, we use Screen Space Reflection and environment maps. We render the environment and we create a cubemap, which is then used as a texture in combination with our shader programs to recreate reflections or we could do it using ray tracing. Incredibly computationally intensive. But much, much more beautiful. We could do refractions as light travels through that glass. But as you guys know, refractions are hard to do, especially with a glass because the glass has curved surfaces. And every time when you guys look at sun shining through a swimming

pool, the curved surfaces accumulate a light volume that causes the refraction to look with these spines of shapes of light, we call caustics.

Notice, this is very, very hard to do that here. In fact, it's not even visible here. We use all kinds of techniques to create the translucency of glass. And this is what it would look like if it was ray traced. It looks like crystal. And notice the beautiful caustics that are here, the beautiful caustics. Imagine if the light was moving. There are objects that absorb light. They absorb light. But they bounce inside and they reemit the light, subsurface scattering, JADE does that, gummy bears do that, car paint does that, your skin does that. And that's why we don't look like a brown blob. That's why we look alive. Light goes through, picks up a little bit of shades of red, bounces around in your subdermis and then reflects out, makes you look alive, subsurface scattering. We use all kinds of techniques to fake it today. Using ray tracing, the gummy bear looks like you can almost pick it up and eat it. Incredible.

Then here -- we didn't show it here. But this one, this gummy bear is actually being lit with the light from the caustics, all of those computationally intensive problems, the effects of light as it travels through your room, the environment, it's so hard to compute, it is so hard to compute. And that's the reason why ray tracing has become -- has been so popular in film. And it is the holy grail, the dream, for computer scientists for the last 40 years.

It's incredible to see this when you look at it in motion. And so here's an example of it in motion. Let's take a look at this.

(presentation)

Wow. Well first of all, what you were looking at earlier was not a scene from the last Star Wars movie. In fact, it wasn't a scene in a movie at all. What you just saw was completely rendered in real-time. Now let's show it to you. Okay. Who's driving this morning. Steve, is that you?

Steven Parker {BIO 4353775 <GO>}

Yes. We're here.

Jensen Hsun Huang {BIO 1782546 <GO>}

Okay. All right. This is Dr. Steven Parker. He's been working on ray tracing his entire career since joining NVIDIA 10 years ago. This has been the undertaking. This has been the endeavor of literally so many people in research, in software, in architecture, in hardware, everybody's working together to make this possible. Now here -- what he's going to show us now are some of the special effects that you could imagine that you can do if you have ray tracing.

Steven Parker {BIO 4353775 <GO>}

We've been working with Epic Games and ILMxLABs on a special project and are showing a prototype of RTX integrated into Unreal Engine 4 through Microsoft's DXR. In this case, we're showing area light shadows, which are quite important for cinematic visual effects. The tricky part is to get the contact hardening or the sharpening of the shadow near the base of the stormtroopers where they touch the ground. It helps -- it's an important visual cue that helps us understand that those people are standing on the ground. Then we can also see them in a more complicated environment.

Jensen Hsun Huang {BIO 1782546 <GO>}

Now Steve, the thing to say is, when you say aerial lights, aerial lights is where light is being emitted from an area instead of a spot. Spotlights are relatively easy to do. Aerial lights are really hard to do. And the reason for that is it's essentially a whole bunch of spotlights. And so the way that the shadow, the way that reflection works and the way that shadow works, shadows are so soft. And notice that there are so many different essentially the accumulation of a whole lot of shadows. And that's why there are so soft shadows.

Steven Parker {BIO 4353775 <GO>}

Exactly.

Jensen Hsun Huang {BIO 1782546 <GO>}

Keep going.

Steven Parker {BIO 4353775 <GO>}

And in cinematics, it's important to control the shape and scale of the light. So that you can control the shadows. So that they fall where they -- where the producer would like them to fall. So...

Jensen Hsun Huang {BIO 1782546 <GO>}

That's incredible. Well let's take a look at another scene.

Steven Parker {BIO 4353775 <GO>}

Let's see something shining...

Jensen Hsun Huang {BIO 1782546 <GO>}

This is shadows and lighting. Now this is reflection.

Steven Parker {BIO 4353775 <GO>}

So another thing that ray tracing is really good at is reflections. We can see them here in an environment, such as the elevator that we saw in the clip. You can see the reflections on the stormtroopers as well as them reflecting each other. But to really show them off, we need to bring in something even more shiny.

And so...

Jensen Hsun Huang {BIO 1782546 <GO>}

Look at this. This is reflections on reflections. Checkout Captain Phasma's gun reflecting on Phasma's chest right there. You see that? Reflections of reflections.

Steven Parker {BIO 4353775 <GO>}

So all those things have to be shaded multiple times, shading the primary surfaces as well as the reflections. And that's one of the things that makes it really difficult.

Jensen Hsun Huang {BIO 1782546 <GO>}

And these rays are bouncing all over this environment. These rays. Every single one of those rays are bouncing off the environment. And every time it strikes a surface, it has to figure out do I reflect, do I refract, do -- am I absorbed. And how much is it absorbed, where does other rays come from that I need to accumulate. And then it bounces off. And it goes to another place and it strikes another surface. Then it strikes all these different surfaces. This light is bouncing all over the place, as it strikes these surfaces, figures out what the rendering equation is at that moment as it strikes that surface, accumulates it for the entire scene with all of these billions and billions of rays. This is how this is recreated.

Now of course, doing so, we just take enormous number of computers. And that's why film studios have supercomputers. That's why Pixar has supercomputers. That's why Industrial Light and Magic has supercomputers. And they use these supercomputers to calculate these rays one at a time as it bounces and strikes these surfaces and creates this image. And the more reflections, the more refractions, the more ways that the light -- the ray can bounce before it's absorbed completely, the more reflections and refractions, the harder it is. There are so many little visual cues. Steven Parker, NVIDIA Ray Tracing team, what an amazing achievement. Ladies and gentlemen, everything that you're seeing here is completely in real-time. Now this complete demo -- yep, there you go. It's completely in real-time. And it's running on one DGX station, instead of a supercomputer rendering these scenes one frame every 10 hours, this is now running on one DGX computer with 4 Voltas in real-time.

This is what we can do now, $68,000 computer versus a supercomputer. That's great. Thank you, guys. So ladies and gentlemen, we're announcing the NVIDIA RTX technology. This has been 10 years in the making. You've seen us demonstrate pieces of it over time. But today, this is a very, very big deal. This is a very big deal because for the very first time, for the very first time -- and by the way, I want to thank the team at ILMxLABs and Epic and all the engineers at NVIDIA that worked on this,

this piece of work, the Star Wars, rendering in real-time is really one of the first times that ray tracing has ever been done at -- it is the first time that ray tracing has been done at this level in real-time. So thank you very much, guys. Super proud of you guys.

After 10 years, what makes this special is for the very first time we can bring real-time ray tracing to the market. People can actually use it. The technology has been encapsulated into multiple layers from our GPU architecture to the algorithms that makes it possible for us to do this. You are also seeing deep learning in action. Without deep learning, it would be impossible to have traced all of those rays. We use deep learning to essentially -- whereas, deep learning has been used in the past for super-resolution, we're using them for super-rays, predicting rays. So that we could fill in the spots that we know what the right answer is going to be using artificial intelligence. And so NVIDIA's Volta GPU, the RTX technology, the solvers, the architectures, the libraries has now been integrated into 3 of the most important rendering APIs: one, the NVIDIA Optics; two, Microsoft's DX12 extension, called DX ray tracing, DXR. The work that we did with them is fantastic. Then now it's also going to be available in OpenGL, Vulkan.

We also are announcing today the world's largest GPU. It's called the Quadro GV100. It is so large, we decided to paint it gold.

Ladies and gentlemen, this is the Quadro GV100. The world's first workstation GPU based on the Volta architecture. It also is the first one that has a brand-new interconnect between GPUs called NVLink 2. Super high-speed interconnect between 2 GPUs, that basically extends the programming model, the memory model, out of our GPU into the other GPU, which means all of the memory reads and writes. And all the atomics work exactly the same. Software doesn't have to change. The 2 GPUs connected through this new interconnect called NVLink is essentially one giant GPU. So these 2 GPUs working together, 2 GV100s will become a revolutionary new workstation. It's going to be available from HP, Dell and Lenovo. I think the announcement goes out today. They will be available relatively soon. These 2 GPUs combined have 10,000 cores, 10,000 CUDA cores, 236 teraflops of Tensor Cores, all used to revolutionize modern computer graphics. The largest frame buffer in the world, 32 gigabytes of HBM2 per GPU, 2 of them together working completely unified inside an application of 64-gigabyte frame buffer, HBM2. Ladies and gentlemen, the Quadro GV100 with the NVIDIA RTX technology.

Thank you, Paul. Creators and designers are going to love this. It is amazing how many frames are rendered each year. There are 400 games being made every single year. This is one of the largest industries in the world. And all of those beautiful lightings that you see is made possible by the technology -- that the technique that I described earlier called light baking, basically using ray tracing to render the entire game in advance. And wherever there are static lights, bake it, put it right into a light map. As a result, you get these wonderful crevices and shadows and all of these details come out. And the whole world looks -- comes to life. The film industry uses this technology. Mowgli and Bagheera wouldn't be possible if it wasn't because of rendering technology; 500 movies made each year. Every single frame has been

rendered multiple times during design and then finally has to be done in final film; 10 hours on the highest performance GPU per frame, just imagine how many CPUs -- excuse me, CPUs per frame, how many CPUs are in those render farms. Product design. When you see advertisements and you see these commercials and these cars are driving around and they're being launched all over the world at the same time, the cars don't exist. They're all completely created using photorealistic rendering. Architecture design. There are no square buildings anymore. All the buildings in the world are a beautiful, gigantic products. And the only way to imagine what it looks like inside, feels like inside, is to render it photorealistically, architecture.

A billion -- we estimate 1 billion frames are rendered each year. With the NVIDIA RTX technology and the Quadro GV100, I believe the number of frames that will be rendered will jump by a factor of 10, just so that you can iterate and try things before you get it done. You also get it done faster. But most importantly, you'll save money. As you know, everybody knows now this is like common sense, the more GPUs you buy, the more money you save. That's right, the more GPUs you buy, the more money you save. This is now common sense. Now let me illustrate it to you. This is what a render farm would look like. Look at this thing. It's incredible; 280 dual CPU servers. It consumes 168,000 watts. That's like a 168 families and their homes being used to render images. This is before, this is after. This is with the NVIDIA RTX technology, Quadro GV100. Look at this. 14 quad GPU servers, 24,000 watts. And you save millions of dollars. This is what it looks like before. Probably a few million dollars, call it $3 million, $4 million, $5 million. And this is what it looks like now, a few hundred thousand dollars. Save millions of dollars. Come to GTC, learn how to save millions of dollars.

Well the industry is just so excited. We're so grateful for the adoption and the engagement and all the work that we're doing with the industry, from tools makers, Adobe, Autodesk, Dassault; engine makers, Epic, Unity; film studios, Pixar, Industrial Light and Magic, they've all come out to adopt this technology. The NVIDIA RTX technology will completely revolutionize the way they do work. They could finally do ray tracing in real-time, try more frames, create more beautiful shots, deliver to the customer faster and on time. And more importantly, most importantly, save millions of dollars in the process. Incredible support from the industry. Gaming, design, film, architecture, this technology is the single most important advance in computer graphics in the last 15 years. And I say 15 years, because programmable shaders that we invented has basically defined modern computer graphics. I believe NVIDIA RTX is going to define the future of computer graphics. This is a very, very big deal. It's a very big market. Super proud of all the engineers that worked on it. Thanks a lot guys. Good job.

One of the best decisions we ever made was 15 years ago, when we decided that a GPU, which was a graphics accelerator as a beginning, would become more and more general purpose. We wanted to become more and more general purpose because we felt that in order to create virtual reality, we had to simulate reality, we had to simulate light, we had to simulate physics. And that simulation has so many different algorithms, from particle physics and fluid dynamics. And of course, what

you just saw earlier, ray tracing. The simulation of the physical world requires a general purpose supercomputing architecture.

My kids told me this morning, Spencer and Madison told me this morning that this is our 10th-year anniversary. The first year was called NVision. Since then, it was called GTC. This is our 10th-year anniversary. We started GTC because we wanted to promote and to enable the industry to create applications on our GPUs because our GPUs are going to become more and more flexible. Over the years, we've advanced the architecture, because accelerators tend to be very, very good at one thing, you accelerate one thing. Over the years, we've expanded the flexibility of our GPUs without sacrificing that one thing, which is accelerating computer graphics to its extreme or accelerating whatever application. And of course, one of the applications you guys know very about -- very well about is our work on deep learning.

Accelerating those applications to the speed of light, creating tools, creating systems, creating interconnect, evangelizing the programming model, all of you creating software and solvers on top of it, step by step by step, over the course of 10 years. And all of a sudden, one day, boom, the tipping point happened. It became a common sense computer architecture. It became a computer architecture that is available literally everywhere. And it couldn't have come at a better time and a more important time. At a time when all of these new applications are starting to show up, where AI software is writing software. But it needs a supercomputer to do it, where supercomputing is now the fundamental pillar of science, where people are starting to think about how do we create artificial intelligence, autonomous machines that collaborate with us, that augment us. The timing couldn't be more perfect.

Well we're at the tipping point. And the number of people who are jumping on top of GPU computing is really growing and growing at an exponential rate. We're almost up to 1 million GPU developers, up 10x in the last five years. The number of attendees of GTC, we are packed all the way to the end and we've got overflow. And Santa Clara is at its limit, Silicon Valley is at its limits. It has grown 4x in the five years. GTC is now a global event. The number of CUDA downloads has increased 5x in five years. But here's the amazing thing. Eight million downloads of CUDA, almost half of it was last year. Eight million downloads of CUDA, almost half of it was down -- was done last year. And remember, we started downloading it 10 years ago. The number of GPU flops in the top 50 supercomputers in the world has grown to 370 petaflops.

We just announced our nation's 2 fastest supercomputers, Summit and Sierra, both GPU accelerated, each one of them about 100 petaflops. Previous, the fastest supercomputer in United States was 20 petaflops, the Oak Ridge Natural Labs Titan. 370 petaflops is equivalent to 3 of the world's fastest supercomputers. 370 petaflops, an incredible amount. 15x in five years. So clearly the adoption of GPU computing is growing. And it's growing at quite a fast rate. And we need larger computers. Even then we need larger computers. The world needs larger computers because there are serious work to be done. There are serious groundbreaking science to be done and there are so many examples of them. Whether it's reinventing the way we store energy, trying to understand the earth core to predict future disasters, understanding and simulating weather or just understanding how HIV viruses work.

Each one of these simulations take days on the world's largest supercomputers. In the case of the work that is done at Caltech seven days on a Titan -- 2 weeks on a Titan, 840 days, almost three years on the Piz Daint supercomputer. And of course, Blue Water, almost 2 weeks to simulate the HIV virus.

Well what's going to happen here is this. We're going to go build an exascale computer. And all of these simulation times will be compressed down to one day. But what's going to happen at the same time is that we're going to increase the simulation model by a factor of 100 and we're back to three months. Then we'll go figure out a way to build a 10 exaflops computer. And we'll be able to reduce that simulation time. And then, of course, the size of science will be able to do, we'll grow again. Science needs supercharged computers. And that's the reason why we're building supercharged computers. If you take a look at the last five years, there is no question now GPU-accelerated computing is the right answer.

On the left -- excuse me, on this side, which is that, on the right is Fermi. This is our first GPU that you probably got to know about us. Where the GPU computing universe learned about my NVIDIA. This is the GPU server, circa 2013. This is the Volta server, five years later, circa 2018. The amazing thing is running the simulations of these molecular dynamics codes and there are so many of them trying to simulate the human biology. The speed up that we created in the last five years is 25x. While Moore's Law is 10x in five years. Moore's Law, the miracle of laws, the law that has enabled just about every single industry in progress of science and the progress of society was compounded over time 10x every five years. Our GPUs have accelerated these molecular dynamic simulations 25x in the last five years. There is a new law going on. It's a supercharged law. There's a new law going on. And I think this is the future of computing.

Well the question is how did that happen. Well the way that it happened, of course, is it starts out with a highly specialized parallel processor. A highly specialized parallel processes, we call the CUDA GPU. And of course, it is the largest and most complex semiconductor the world makes. But it's much, much more than that. The interconnect of these processors, the memory system of these processors, the way that it's architected into systems, all of the software on top of it. And I'm just showing you some of the stacks, look at this. From cuBLAS linear algebra, FFTs, random numbers, sparse linear algebra, the software stack in 2013 was 5.0. Amber 12. -- version 12. This is Amber 16, 9.0 across the board. Basically, we're continuously updating and continuously refining these algorithms and these solvers. We could design new architectures every single year. That's how we got from Kepler, to Maxwell, to Pascal, to Volta. On top of it, we create new interconnects, NVLink.

On top of that, we embrace new memory technologies and drive it to the speed of light. On top of that, we have new solvers and algorithms. On top of that, we work with system makers to create new systems. And on top of that, we work with all of the application developers to refactor, to optimize our software so that the entire algorithm could be highly accelerated. Now the benefit to science is incredible. This is what a supercomputing rack looks like. It's much, much more densely populated because most supercomputers deliver a lot more power per rack, 20,000, 30,000

watts per rack. This is 600 dual CPU servers, 360,000 watts. And this is what it looks like if it was GPU accelerated. That's right. Work with me here. The more you buy, the more you save.

This is not a complicated talk. Guys, this is not a complicated talk. I want to -- first of all, I want to congratulate -- speaking of talks, I want to congratulate John Hennessy and Dave Patterson for the recognition of the work that they've done, the Turing awards. Hey guys, isn't it amazing. John, you're my hero. Speaking of talks, the guy gives just the most amazing computer architecture talks. This is not like one of them. This is much easier. There are only 2 messages I want to deliver today. One of them is, the more you buy, the more you save. Okay. All right, there'll be a quiz at the end. There's a second quiz coming up.

NVIDIA Tesla B100, big savings for HPC. One of the most important and one of the -- the applications that I love the most of all of the work we do in HPC is the work that we do in revolutionizing modern medical imaging. Well it turns out that early detection is the most powerful weapon against disease, the sooner you detect the better. This is true for autonomous machines of all kinds. The sooner you detect, the better you detect, the better. Well whether it's CT, computer tomography; magnetic resonance; ultrasound; mammography, which is low-dose x-rays; the positron emission, PET, each one of these modalities of medical imaging has been revolutionized recently using computational approaches. In fact, CUDA is embraced and our GPUs are embraced throughout the medical imaging world for this very reason because it could take these sensors, these weak sensor information, do image processing on it, reconstruct a 3-dimensional image, visualize it, segment it, detect things using deep learning. All of these techniques have now come together in medical imaging.

On the left is a 15-year-old ultrasound image. In fact, it hasn't changed very much over the years. Most ultrasounds that you see in hospital still look like that. On the right is a brand-new medical imaging equipment from Philips Epic. It's unbelievable. It could do 3D reconstruction. It could visualize. It's as if there's a virtual light inside the womb. It's -- look how beautiful it is. It does 3D segmentation. So it detects the baby and lifts it out of all that noise. And then it does beautiful rendering. And here you could tell they're using some amount of subsurface scattering. So that the baby looks like there is the tone of flesh. Okay, modern ultrasound.

Well this technology is used everywhere as I mentioned. It's used in CT, MRIs, in PETs, in mammo and ultrasound. And in addition to that, modern advances in deep learning and artificial intelligence is about to revolutionize the entire industry. We can now reconstruct images better than we could ever before. We could identify and segment brain tumor better than ever before. And we can visualize the images in a way that reveals a lot more insight using cinematic rendering than we ever could before. The entire stack of NVIDIA from the GPU, the GPU containers, the virtualized GPUs, the CUDA, cuDNN, RTX, OpenGL, all of the solvers and all the libraries integrated into these imaging applications. That entire stack is identical to everything that I've talked to you about. It's identical to everything I've already talked about.

Well the unfortunate thing is this. There are some 3 million, 4 million, 5 million medical instruments that are installed all over hospitals, all over the world. So let's call it 3 million instruments. 100,000 new ones are being sold each year, which basically says it's going to take about 30 years before we can replace the installed base. And so the question is, how do we solve this problem. We can't wait for 30 years for doctors to be able to have early detection and have this powerful technology in their hands. And so the question is, can we take advantage of the same basic techniques that we've now seen in modern computing, where you have connected devices connected to data centers. And those data centers essentially have supercomputers that are virtualized. We've an initiative, a project in our company, we called project CLARA. CLARA, in short, is a medical imaging supercomputer. But what it is, is a data center virtualized, remoted, multimodality, multiuser medical computational medical instrument. Well that is such a long description of what it is. We decided to call it, it's a medical imaging supercomputer. You can put it in the data center, you can put it in the cloud. And because of the technology we've already talked to you about, it's possible for us to actually virtually upgrade every single medical instrument. Let me show it to you. Fausto. Hey Fausto, hey man...

Fausto Milletari

So what we are you seeing here is...

Jensen Hsun Huang {BIO 1782546 <GO>}

Wait.

Fausto Milletari

I'm sorry.

Jensen Hsun Huang {BIO 1782546 <GO>}

Yes, LUA, buddy. Forget it. Inside joke. I want to prepare his language so that you understand it. No, no, no, he -- this is one of our scientists in our company. And this is his field of expertise. This is where he dedicates his life. Dr. Fausto?

Fausto Milletari

Dr. Dr. Jensen.

Jensen Hsun Huang {BIO 1782546 <GO>}

That's right. Every time I call him doctor, he calls me doctor, doctor. And another inside joke. With your friends in the company, you've got a lot of inside jokes. And so Dr. Fausto has been working on medical imaging for some time. And so what he's about to show you is something that's really amazing. This -- go ahead, tell us.

Fausto Milletari

Sure. So what you are seeing here is a beating human heart. And this is an echocardiogram. Actually, we're seeing the left ventricle of the heart, which is pumping blood. And basically, this is the type of image that a doctor would look at when he diagnosis your heart functionality. And yes. So this is 2D. But our body, of course, is 3D. So there is more information from this scan that we can acquire. So when we switch to 3D, we see a more complete picture like this...

Jensen Hsun Huang {BIO 1782546 <GO>}

And so this basically is a scan. It's an ultra -- traditional ultrasound scan coming off of a 15-year-old instrument that we then take into this computational medical imaging supercomputer, this MIS. And this is what comes out of it. Does that makes sense? Okay, you take this ultrasound machine that's sitting in a hospitals, it's already on the network anyways, you stream the ultrasound information into your data center and your data center running this stack on top of a GPU server creates this miracle on the right. And now it's volumetric, keep going.

Fausto Milletari

Yes...

Jensen Hsun Huang {BIO 1782546 <GO>}

That's nothing, wait. He gets so excited.

Fausto Milletari

Okay. So at this point, we can -- we see the -- we barely see basically the chambers here and we can analyze more the motion of the left ventricle by using deep learning. And we resort to use a state-of-the-art deep learning method. It's a fully convolutional neural network in 3D called V-Net. We revisited it. We used the unprecedented resolution of 256 cubic voxel (inaudible) million values per scan per frame and we have asked deep learning to interpret it for us. And now we can see the chambers moving across the cardiac cycle, the left ventricle moving across the cardiac cycle and segmented very well.

Jensen Hsun Huang {BIO 1782546 <GO>}

Now just think, you got to realize what's happening. There's an artificial intelligence network here that is running on the original black and white gray scan that you saw. And from that scan -- from that ultrasound scan, it inferred what this left ventricle -- where the left ventricle is. It segmented it out. It segmented it out in motion and it segmented it out in 3D. Guys, if that's not amazing. That's pretty incredible. That's pretty incredible. Fausto, that's amazing work man. Go ahead.

Fausto Milletari

Of course, there is more even to this story. So now we can get values. We can get very precise explanation what we are seeing and markers that tell us how well the heart is functioning. We can get the ejection fraction, that's volume of the chamber and systolic and diastolic movements. And the amount of blood that is ejected with each heartbeat. And we can just augment the view of the doctor in this way, in an almost transparent way. The 2D view is augmented with the 3D information.

Jensen Hsun Huang {BIO 1782546 <GO>}

That's cool. Fausto, thank you. Your ejection volume was running about 80%.

Fausto Milletari

At this moment, yes, most probably.

Jensen Hsun Huang {BIO 1782546 <GO>}

Do you think so? Maybe 99%, everything coming in was going out. That's fantastic. And so here's what we're going to do. We're going to create this server with this stack -- this virtualized stack on top. And it's going to be remoted into all of these medical instruments all over the hospital. Then you'll be able to scan using your current installed base of scanners. And then we'll post-process it using AI to infer more information from it. We could also use all kinds of new computational techniques to improve the image and to visualize that image. We call this the medical imaging supercomputer. It's so great. We have partners all over the world working with us. And there's a -- this is an area that we've been working with and working on for quite some time. And so NVIDIA's new project CLARA, this brand-new platform of ours, we're getting a lot of advice and developing and working with startups and research hospitals and health care providers in all the different modalities. Okay. Ladies and gentlemen, CLARA.

Deep learning.

Deep learning has completely turbocharged modern AI. This incredible algorithm that can automatically detect the important features out of data. And from this algorithm, it could construct hierarchically knowledge representations. And these knowledge representations, if you give it enough information, enough data, it would become more and more robust and recognize a larger diversity of that space, become more intelligent. Deep learning has revolutionized modern AI. The thing that deep learning needs though is a ton of data. And more importantly, a ton of compute. And the reason for that is, it just iteratively goes through and tries to figure out what are the important features using all of these incredibly revolutionary algorithms.

Our strategy at NVIDIA is to advance GPU computing for deep learning for AI at the speed of light, from the processer development to the systems, the interconnects, the way we construct systems, to all of the software layers that are on top of it, making it available and partnering with cloud service providers and OEMs

everywhere. Irrespective of what AI framework you use or what AI network you are trying to create, wherever you are, we will support it. And we will support it from end-to-end. So that the frameworks that you use to develop the networks and models will be deployable in large scale. And we would even make it possible for you to access our platform in every possible way and in any sort of way. Whether you would like to build a personal supercomputer for a few thousand dollars with our TITAN V or to be able to rent time in the cloud for a few dollars an hour or to be able to build your own supercomputing cluster or just to buy one from us. We're going to make it possible for you to do your work at the speed of light and to be able to do it anywhere. Today, we're announcing several big things. First, we are doubling our GPU. The Volta V100 is now 32 gigabytes of HBM2 memory. It's twice the size of what it used to be.

Now this is so important. And the reason for that. And I'll show you later, that networks are getting larger and larger and larger. And so it doesn't fit in 16 gigabytes. Researchers would like to not worry about memory size while they're developing their neural networks. And 32 gigabytes gives them just a lot more space. HBM2, the fastest memory in the world, connected to the Volta V100. It'll be available as of now in the next DGXs you buy, it will be available in the cloud very, very shortly. It is in complete volume production everywhere.

The new Volta with 32 gig. Neural networks are growing and evolving at an extraordinary rate, at a lightning rate. What started out just five years ago with the AlexNet, Alex Krizhevsky's now world famous AlexNet, was 8 layers deep, had millions of parameters, 8 layers deep, millions of parameters, 8 layers deep and millions of parameters. It's a CNN. The CNN was first developed and produced or created by Yann LeCun. And the training technique, the Stochastic Gradient Descent by Geoff Hinton. And it was implemented in AlexNet for the very first time. And it won the ImageNet competition.

Well 5 short years later, thousands of species have emerged, thousands of species of artificial intelligence networks, models of all different types, from the CNNs of Yann LeCun to the RNNs that was first done at Schmidhuber's lab in Switzerland. The Generative Adversarial Networks of Ian Goodfellow, the deep reinforcement learning that was most recently done by DeepMind -- Demis' DeepMind and (Peter Beale) over at Berkeley, all of these different architectures have hundreds of different implementations. The number of species is growing so incredibly. And there are so many new ones coming out. Hinton just recently disclosed the capsule net. So that you could not only recognize images in its image form about to learn the geometric form of the images.

So many new techniques, neural collaborative filtering, NCF, which is used for recommenders. What started out just five years ago as 8 layers and a few million parameters is now hundreds of layers and billions of parameters, 8 layers and hundreds of -- and millions of parameters to hundreds of layers, billions of parameters. Neural network models have increased in complexity, not to mention the number of species, the thousands of species that are now out there. Neural network models have now grown in complexity 500x in just five years, 500x in five

years. Remember, Moore's Law would have only kept up with 10x of that. 500x in five years. So the world wants, the researchers all over the world wants just a gigantic GPU. Not a big one, a gigantic one. Not a huge one, a gigantic one. And so ladies and gentlemen, today, I would like to announce the world's largest GPUs.

This is the world's largest GPU. This is 16-Volta equivalence, connected by 12 brand-new high throughput switches that the world has never seen, it's called NVSwitch. These 16 Tesla V100 equivalence, each with 32 gigabytes creates virtually a 512 gigabyte memory. Not only that, these 512 gigabyte memories, the way you address the memory and the way that every single GPU could talk to the memory of another GPU, completely using the same memory model. The memory fabrics Symantec on our chip that connects all of the processors have been extended out of this chip, out of the GPU, onto the NVSwitches connecting each and every one of them. In total, 14 terabytes per second of aggregate bandwidth, 14 terabytes. So that's -- what is that? Let's think about that. 14,000 gigabytes. And let's say, if a high-resolution movie is 10 gigabytes, okay. So 14,000 gigabytes, 1,440 movies, 1,440 movies. More movies than any human has ever seen could be transferred across this switch in 1 second, like that. Ladies and gentlemen, 14,000 movies.

14,000 movies downloaded in 1 second. Dream come true. Altogether, 81,900 CUDA Cores, 2 petaflops. This GPU is 2 petaflops. I told you earlier, the fastest supercomputer on the planet is 125. The fastest supercomputer in America is 100 petaflops, this is 2 petaflops of Tensor Cores for AI on this GPU, the world's largest GPU. Let me show it to you.

Hang on a second. You know guys, here is the thing. We're still busy pulling the stuff together. We don't rehearse. Okay? Our rehearsal is basically grip it and rip it. Ladies and gentlemen, the world's largest GPU.

You didn't fall for that, did you? You didn't fall for that. Hang on a second. This isn't the world's largest GPU. This is the world's largest GPU. Guys, it's sitting in plain sight, hiding in plain sight, right here.

Sandy, if I could give that to you. Now I understood earlier that, that -- was that the game plan for me to do this myself? Oh, okay, it wasn't. All right. Ladies and gentlemen, grip it and rip it.

This -- this, ladies and gentlemen, is the world's largest GPU. Now what you're looking at is something that's really truly amazing. So let me show it to you this way. And I'll come back to that.

Okay. This switch has 2 billion transistors, made of TSMC's 12-nanometer FinFET. This -- great process. This -- every one of these switches -- and there are 12 of them. Every one of these switches has 18 links, which are 8 bits wide, 8 -- 18 NVLinks, 8 bits wide, with a SerDes that's moving at 25 gigabits per second, 25 gigabits per second now in just one signal. There is 8 of them per link, there is 18 links on this one chip. And it's bidirectional. Altogether, it creates 7.2 terabits per second or 900 gigabytes per

second. So 900 gigabytes per second, by the way, of bandwidth, well, that's a lot of movies too. So 7.2 gigabits -- 7.2 terabits per second, not 100 gigabytes per second goes through one switch. There are 12 switches, 12 switches. What that basically says is that every single GPU, every single GPU can communicate to every other single GPU at 20x the bandwidth of PCI Express.

Isn't that amazing? Every single GPU can talk to every other GPU. And every single GPU can talk to every single GPU in a nonblocking way, it's a fabric. It's not a network, it's a switch. It's a nonblocking fabric switch with a memory Symantec's, memory programming model, that's exactly the same as what's inside our chip. And therefore, all the reason, all the rights and all the atomics just works across this. The bandwidth and the performance and the latency of this switch is incredibly low. It's incredibly low. Unlike a network, this is a switch. Which means, every single GPU can talk to the other GPUs with extremely low latency. You want to read, you want to synchronize a parameter, it takes no time. 300 gigabytes per second. And all together, NVIDIA's largest graphics card. This is what it looks like. The largest graphics card the world has ever made. The largest graphics card the world has ever -- 2 petaflops, 512 gigabytes. Most graphics cards in the world today, 2 or 3 gigabytes, 2 or 3 gigabytes. This is 512 gigabytes. 10,000 watts, 10,000 watts only. Only 10,000 watts.

Only 10,000 watts. The amount of airflow is really quite amazing to cool it. And it -- and we designed this porous, this porous fabric air intake in the front. It looks beautiful. But it's incredibly functional. Air comes traveling through it, flows through all of our chips, 10,000 watts of energy power is being thermal managed. 350 pounds, no human can lift it. No -- 350 pounds. The first time I saw that, I was thinking I could lift that. It didn't move. And I'm fairly strong, upper body strength. It's 350 pounds. What's inside it is 16 Tesla V100 32 gigs with 12 NVSwitches. And it's got this incredible deck, it's called playing card. It basically has 200x, it has 200x the bandwidth of the highest speed NIC on the planet. So imagine, it's like 200 NICs of connecting the top to the bottom, connecting the top GPU tray to the bottom GPU tray. Each NIC, call it $1,000, $2,000, call it $1,000, $200,000 of NICs would be required to connect the top to the bottom. It has that much bandwidth. It has that much bandwidth. If we were to use networking cards, well, we still have to connect multiple of these.

And so we have 8 of the world's highest speed NICs from Mellanox to connect multiple of these systems together. It has the 2 fastest CPUs we can possibly buy the Xeon Platinums. 30 terabytes of storage, 30 terabytes of storage on the system because, as you know, we're going to crunch through a lot of data really, really fast. So all together, ladies and gentlemen, it looks like this. This is the NVIDIA DGX-2, the world's largest GPU, the world's largest adding card, 2 petaflops, 512 gigabytes, 350 pounds and just beautiful. This is what an engineer finds beautiful. You guys, this is sexy. This is beautiful. This is unbelievable. Incredibly beautiful.

And not to -- guys, did you guys notice the animation? NVIDIA's artists are so spectacular. This entire animation was shot with 112 shots. It's photographed, a stop motion animation. This is not 3D. You know that we could do it in 3D. But we chose

not to. Why? Because it's more fun. Do you guys want to see it again? Oh, stop motion animation guys. Check it out. Guys, I love you guys. It's incredible. Tim Burton, you've got nothing on us. Look at that. The world's largest GPU, 350 pounds. Well let's see what it can do? 10x faster than DGX-1. Does that even make any sense? 10x faster than DGX-1. And so this is six months ago. This is literally six months ago, we showed these numbers. And the entire stack, look at the entire stack. There is so much software that's necessary to build this. This is a high-performance computer. It's a supercomputer designed for deep learning.

And so look at this entire stack of software that was optimized. As a result, running (inaudible) to framework and training FAIR seek, the sequence-to-sequence model used for national language -- used for machine translation literally took 15 days on a DGX-1, which was a world record at the time and now it takes 1.5 days. Well the amazing thing is, you don't have -- the researchers don't have 13 days left to sit around. The researchers are now using all kinds of techniques to use AI to create AIs. And so the number it experiments, the number of permutations, the exploration space of AI, sweeping through all the different configurations of architecture and the way to compose it, the number of layers, the training rates, sweeping through all of those hyper-parameters and sweeping through all the different architectures at the same time is going to create just hundreds of experiments at one time. And so as a result, with bigger networks, more data, more experimentation, DGX-2 couldn't have come at a better time. And so it is 10x in six months.

Now the thing that's really amazing is this: how much should we charge? I mean, look, it took hundreds of millions of dollars of engineering. It took hundreds of millions of dollars of engineering. This is the first unit. So this is, call it, $250 million. And for a just an incredible friend-to-friend price of $1.5 million, it seems like just an -- it doesn't even seem fair. It doesn't seem fair at all, no. Ladies and gentlemen, $399,000.

You guys understand, there are only 2 themes in this conference. That's right. The more you buy, the more you save. Okay? This front row has got it. We're going to keep working. All right. The more you buy, the more you save. And so $399,000 for the world's most advanced, most powerful computer. This is what it replaces: 300 dual CPU servers easily for $3 million, easily for $3 million, easily for $3 million. But the important thing is 180,000 watts, 180 watts, in order to train FAIR seek in 1.5 days. Well it only takes one of these beautiful DGX-2, just one DGX-2, 10,000 watts, only 10,000 watts, 1/18th the power, 1/8 the cost. The more you buy, the more you save. It's incredible.

You know it's truly amazing how far we've come in just the last five years. We started talking about deep learning about 6 or seven years ago. And -- at the time, the effectiveness wasn't anywhere near what it is today. five years later since AlexNet, the progress that we've made is literally incredible. And I think it's just be kind of fun, just to do a report card. See what happened in the last five years. This is what it looks like. This is what Alex Krizhevsky trained AlexNet on. Two DGX -- 2 GTX 580s. And it took him six days. But those were 6 worthwhile days. Because he became world famous, he won ImageNet and kicked off the deep learning revolution. five years later, we

could train it in 18 minutes. five years later, we could train it in 18 minutes, 500x faster. Isn't that amazing, guys?

To be able to do a task 500x faster in five years. To be able to do any task 500x faster, any task of great importance, 500x faster, or to be able to do a task of great importance 500x larger in the same time. This is some kind of a supercharged law that we're experiencing. Well one of the things that made it possible for us is what I said earlier, this is an innovation that is not just about a chip, this is an innovation about the entire stack. Everything was touched. Thousands of engineers all over our company, from VLSI to process engineering, to package engineering, to circuit design, the architecture, the chip design, system software, solvers and libraries and algorithms and systems, all got involved. This is the effort. This is the work. This is the body of work of literally every single employee at NVIDIA. We are all in on deep learning. And this is the result. And so over the next several years, it's going to be just utterly incredible, because we're just picking up steam. This is some new type of computing.

The amount of data is growing exponentially. There is evidence that with GPU computing, the computation is growing exponentially. And as a result, deep learning networks and deep learning models, AI models, are growing in capability and effectiveness at a double exponential, at a double exponential. More data, more computing, both on an exponential growth, compounding together into some kind of a double exponential for AI. That's one of the reason why it's moving so fast. Really, really exciting times.

The system is complicated. And the software is complicated. And so what we've done is, we've created containers, to containerize all of these complicated software that are optimized into one container at a time. We put them into Tupperwares, if you will. And by putting them into these Tupperwares, we put them up in the cloud. We call it the NGC, the NVIDIA GPU Cloud. The NVIDIA GPU Cloud is not cloud computing. NVIDIA GPU Cloud is a cloud registry of all of these containers. 20,000 organizations have downloaded our containers. It's running in cloud. It's running in data centers. It's running all over the place. These NGC containers are just completely fabulous. And the reason for that is because irrespective of what cloud you run on, it's exactly the same stack. Because all the clouds have NVIDIA GPUs, all of the clouds now can run these stacks. All you have to do is come to that registry, you log in, you download, you run. You log in, you download, you run. It's about like that. 1, 2, 3, easy-peasy.

20,000 registered users or organizations, a lot more users, 30 containers. Last year, when I stood here, there were a handful. We've updated them, improved them, enhanced them and we added a whole bunch more to them. And so I'll -- NGC has now been certified, we've certified. We have a test each and every one of these architectures on the data centers that we run on. We've certified AWS. And just today, we're announcing that Google Cloud Platform GCP, AliCloud, Oracle cloud will also be certified -- has also been certified. Okay. So now they're available on multiple clouds. This is the only architecture that is any cloud. And in fact, it also runs,

obviously, on DGXs. So it runs on any cloud and in-prem. So that's hybrid cloud. So ladies and gentlemen, our latest release of NGC.

PLASTER. Programmability, because the number of network species is growing exponentially. And they're evolving. They're getting more and more complicated. Latency, because without low latency, you can't have a good quality of service. Accuracy, if your model isn't accurate, there is no way that it could predict the right answer, or more importantly, revenues get hurt. Quality of service is compromised. Prediction is compromised. Size, the smaller these networks are, the more of these services we could fit in a data center. Throughput, ultimately, these data centers are enormous, enormous capital investments. And the higher the throughput for the billions of people who are using it, the lower the cost, the more money you save. Energy efficiency, we are consuming a lot of energy all around the world. We need a computing model that dramatically reduces energy. Energy efficiency.

And rate of learning. Rate of learning, a machine learning. It's about the amount of data you have, the amount of computing you have. But ultimately, it's about how fast you're learning. The more new data you have, the faster computers you have, the faster rate of learning. You're going to deploy the model. You're going to learn from the new experiences. You're going to collect new data. You're going to train the current model. You're going to deploy it. That rate of learning is what defines success in the new world of machine learning. Ladies and gentlemen, I invented this word.

PLASTER. The second thing I want you to remember today. This is not a complicated conference. The more you buy, the more you save. And number two, inference is complicated. Anybody who thinks that inference is easy, you know what, I'm just going to go buy a ASIC, stick it to my data center. Here is an ASIC, here is an FPGA. I'm going to stick that in my data center. Unfortunately, we know the hyperscale data centers are the most complicated computers the world has ever made. It's the size of these rooms. Billions of dollars are dedicated to it. It serves billions of people. How could it be simple? Building one computer for one user is plenty hard enough as it is. Building a supercomputer, hyperscale, that serves billions of people, delivering great quality of service while minimizing operational cost, it can't possibly be easy. Inference is hard because everything matters. Programmability matters, latency matters, accuracy matters, size matters, throughput matters, energy efficiency matters and rate of learning matters. God, how did I do that? Come on, you guys, that's fairly good.

I was trying to figure out how I was going to explain the complexity of inference last night. This is it, brand-new slide. Just one slide. We're done. PLASTER. So we've dedicated just an enormous amount of resources to solving this problem for inference, for hyperscale data centers. Their workload is really complicated. From image and video workload, natural language understanding workload, recommendation systems, recommenders, speech synthesis, speech recognition. They've all kinds of different type of networks, sequence-to-sequence networks, deep generator networks -- something -- where's MLP? Where's Mike Houston? Multilayer Perceptrons, for God sake, neural collaboration filters. Okay. Recurring neural networks, CNN, of course.

And so we started working on TensorRT, which takes the output of these frameworks, which are these massively complex computational graphs. And we have to target for all of the parameters that I just told you about. We're -- we've got to make that network as small as possible, as high performance as possible and yet retain all of its accuracy. And each one of the target devices, of course, are a little different because our GPUs for deployment are different than the GPUs for training. And so initially, one year ago, we announced CNN. Then six months later, we announced 8-bit integer. So that we could have multiple types of precision. So that we could adapt for which one of the networks could lose a little precision at the benefit of a lot more throughput. We then support a Tensor Core six months ago.

Today, we're announcing the largest battery of new tools, the largest battery of new algorithms and new libraries for inference that we've ever announced. First of all, TensorRT 4.0. Brand-new TensorRT. It now has the ability to handle recurrent neural networks, sequence to sequence. It has deep integration into TensorFlow. I want to thank the engineers at Google who took TensorRT and our compilers and integrated it natively into TensorFlow. And so in the future, when you're done training the network, you could run it, run that network fully optimized right on the target device. We work with a team of Kaldi framework, the most popular voice recognition, speech recognition framework, called Kaldi has now been fully optimized for TensorRT for NVIDIA's architecture.

Then lastly, ONNX, a brand-new backend that supports PyTorch, supports MxNet, supports Windows and has the same backend for Windows ML. We now have full optimization across the stack. And as a result, the trained networks that we train and all of these different types of networks and all of these different types of frameworks can now have the opportunity to be deployed into the world's Hyperscale data centers. Four brand-new capabilities addressing literally the entire stack of the workloads. Still, lots and lots of work that we have to do in collaborating with all of you in your hyperscale data centers. But if we are successful in doing so, the millions of hyperscale servers can now be accelerated. Up until now, we've only been able to accelerate just those ones that are really focused on imaging, image and video. But in the future, starting with this generation, starting with today, we can now accelerate voice, speech, natural language understanding, recommender systems as well as images and videos.

While we've shown, we pulled together some of the latest results, it's really truly amazing. Images are accelerated at 190x. Natural language -- this is the Google Natural Machine Translator (sic) (Neural Machine Translation), GNMT for national language understanding, 50x. The neural collaborative filter that is in the SAP recommendation system, that IBM recently announced as well, those architectures recommend NCF has been accelerated 45x speech synthesis; WaveNet, 36x and speech recognition 60x. If we just assume that video and images and speech and voice and all of that stuff is all on balance approximately equal load, in aggregate, we're going to speed up hyperscale data centers with our GPUs with this generation of optimizations 100x. We would increase the throughput by 100x. Or another way of saying it is, we are going to save a lot of money. You guys -- come on, we've done this before. You guys know the drill. The more you buy, the more you save. And with 100x speed up, just imagine how much you will save.

Well now that we have all these accelerated frameworks and all these accelerated code, how do we deploy it into the world's data centers? Well it turns out, there is a thing called Kubernetes. Ladies and gentlemen, today, yes, it's okay. I know. Kubernetes on NVIDIA GPUs is going to bring so much joy. So much joy.

There are people watching this out on the web going, "That guy has lost it." That brings joy? Well yes, this is going to bring joy. And this is going to bring joy because Kubernetes allows us to take these massive workloads that are servicing billions of people. And it's an orchestration layer that orchestrates the workload that's coming in from the cloud and orchestrated across the resources of the hyperscale data center. Lots of CPUs and lots of memories, lots of networking, lots of storage. And these are orchestrated across the sea of servers. And today, it is GPU-aware and it is GPU-accelerated.

So Kubernetes is now GPU-aware. The docker container is now GPU-accelerated. You've got all of the frameworks that I talked about, which is GPU-accelerated. And now you've got all these inferencing workloads that are now GPU accelerated; and you've got NVIDIA GPUs in all these clouds; and NVIDIA GPUs in all these data centers and servers; and you've got this incredible wonderful, orchestration layers, system software called Kubernetes; life is complete. Okay. Let me show it to you. Guys, Justin, this is Justin. Justin is, of course, you guys have seen these flowers before. Now we're going to do the flowers a little bit differently today. So this is flowers, of course, running on CPUs, 4 images per second and the fastest Skylake that we have. And so Justin, take it away, sir.

Ryan Olson {BIO 17673249 <GO>}

Yes. Absolutely.

Jensen Hsun Huang {BIO 1782546 <GO>}

Oh, Ryan -- excuse me. I'm sorry. It's Ryan, who's coming up.

Ryan Olson {BIO 17673249 <GO>}

We're going to switch it up. But Justin's...

Jensen Hsun Huang {BIO 1782546 <GO>}

Yes, yes. No, no. That's all right.

Ryan Olson {BIO 17673249 <GO>}

You got nervous.

Jensen Hsun Huang {BIO 1782546 <GO>}

No, no. It's not me.

Ryan Olson {BIO 17673249 <GO>}

Absolutely. So as you said like, TensorRT and inference performance -- we've laid out the value for it, right here. So this is the miracle of deep learning, showing flowers using (Resident) 152 on Intel's latest Skylake CPU. So if we click on this, we can see these flowers.

Jensen Hsun Huang {BIO 1782546 <GO>}

Pelargonium. That's a Snapdragon. Hi, guys, wow. Come on. That's industry. That's industry stuff. People in the tech industry, we should go -- you recognize -- I didn't know that it was a flower. It's awesome. Fritillaria. Okay. Blanket flower. Okay. So now show us what it looks like on a GPU. So Ryan, this is now running on NVIDIA GPU. Okay. So this is Volta, right?

Ryan Olson {BIO 17673249 <GO>}

Yes.

Jensen Hsun Huang {BIO 1782546 <GO>}

This is what? Guys, yes, that's right. The more you buy, the more you save. Okay. Check this out. Ryan. So this is one GPU.

Ryan Olson {BIO 17673249 <GO>}

One GPU.

Jensen Hsun Huang {BIO 1782546 <GO>}

This is one GPU.

Ryan Olson {BIO 17673249 <GO>}

Exact same network.

Jensen Hsun Huang {BIO 1782546 <GO>}

This is one GPU. That's nuts. This is one GPU. How are we going to make a living, if this is one GPU? So this is one GPU. Okay. And so suppose -- let's Kubernete this thing.

Ryan Olson {BIO 17673249 <GO>}

Yes. We got it running in Kubernetes. So imagine like our app gets so popular, everyone wants to know about flowers and we get a larger load. So we're going to scale this up. Let's look at what a larger load looks like. There we go. That's a larger

load. So we need to handle that load. So what we do is, we could ask Kubernetes and say, "Hi. let's make multiple replicas of that same container." And so I'm going to add 8 replicas in. You're going to see them come in. And I'm going to add them into the load balancer. And as I add them in, you're going to see, it just get faster.

Jensen Hsun Huang {BIO 1782546 <GO>}

Wow. Guys, this is like magic. Kubernetes is orchestrating this data center. And what's amazing is this. So Kubernetes assigns this pod, which is basically a service-run application that contains a whole lot of containers, okay? It could assign this pod on one GPU, on many GPUs and one server, on many GPUs and many servers. It can also assign it across data centers. So you got to have some of it in the cloud and some of it in your data center, you got some of them in this cloud, some of that in that cloud. Truly amazing. And all of this stuff is happening completely invisibly because we've made Kubernetes GPU aware and because the NVIDIA architecture and all those containers are literally everywhere. And so as a result, this magic could happen. Because the foundation we've been layering and we've been growing now for several years has now come together. And this moment, the Kubernetes moment holding it all together. Ryan, what else you've got for us?

Ryan Olson {BIO 17673249 <GO>}

So as you said, not only can we run it in our local on-prem data center but we can scale and we can burst into the cloud. So let's take a look at that right now. So I'm going to...

Jensen Hsun Huang {BIO 1782546 <GO>}

SATURNV is at our NVIDIA data center. It's our supercomputer. There are 660 DGXs in SATURNV today. So 660 petaflops. It's about to get upgraded as you can imagine.

Ryan Olson {BIO 17673249 <GO>}

It's getting a bit exciting. So I'm going add 4 and more from the cloud. So this is going to be an example of cloud bursting or cloud fall over. So here we've got -- we've turned on 4 in the cloud, they are active. And what I'm going to do now is show you self-healing. So I'm going to kill off 4 of these GPUs that are running right now on SATURNV. And what you're going to see. And you're going to have to watch really closely, is we're going to take a bit of a performance hit. And then those AWS GPUs are going to jump right in and the performance is going to come back up. Ready? There they go down. You didn't even see it this time. Sometimes, it's that fast.

Jensen Hsun Huang {BIO 1782546 <GO>}

Wow, that's an incredibly good demo. Ladies and gentlemen, this is a demo of resilience where before and after is exactly the same. I bet you really enjoyed that. We don't have to show it again, because guess what guys, it looks exactly the same. Okay? A demonstration of resilience. And this is one of the most amazing things about modern hyperscale computing. And we now have cloud-to-cloud, we have

any cloud, we have cloud-to-prem, we have any prem. And so the NVIDIA architecture, all of these frameworks, all GPU accelerated is now possible orchestrated under Kubernetes. So that it can basically run everywhere. Ladies and gentlemen, Ryan.

So the NVIDIA Kubernetes on NVIDIA GPUs. Super exciting about that.

The inference movement on our GPUs is really moving fast. We now have up to 40,000 downloads, I think, of TRT and it's just picking up steam. And the reason for that is because all these different companies are developing their networks and models. And it's time to deploy them. And they want to deploy them into servers. They want to deploy them into PCs, WinML. You want to deploy them at the edge, IoT devices, on NVIDIA Jetsons. You want to deploy them on supercomputers. You want to deploy them on hyperscale. You want to deploy them inside companies with SAP. The number of ways you could deploy software is, of course, quite large. And now with all of our new battery of accelerations for inference and Kubernetess on NVIDIA GPUs, we can now literally blanket the world with this new approach. So super excited to do that. Ladies and gentlemen, the NVIDIA AI inference.

So this is the NVIDIA AI platform. Today, we announce that the GPU has been -- Volta has been doubled in memory size to 32 gig. It's now available in DGXs and available in the cloud soon. NGC is now on AWS GCP AliCloud and Oracle and more are being certified. We're certifying them as fast as we can. The NVIDIA GPU cloud now has 30 optimized GPU containers. And this year, we announced 4 new battery of acceleration capabilities for inference, new TensorRT 4.0 that is able to handle RNNs and sequenced-to-sequence type of models; TensorFlow deep integration; Kaldi acceleration, which is the most popular speech recognition system; ONNX, an industry standard backend for inference; and WinML, optimization into that. And sorry to say, TITAN V is still out of stock. Apparently, the popularity of TITAN V is quite high.

And so we're -- we'll make them as fast as we can. And we'll put them back in the store and if you could just be patient. Lots and lots of TITAN V is being sold.

The NVIDIA AI platform. You know all of this -- one of the things as I was reflecting on this keynote, one of the things that just gives me so much pride and it's such a great creation is NVIDIA Research. For 10 years, Bill Dally has been building NVIDIA Research. Bill Dally is my Chief Scientist. And he is the Head of NVIDIA Research. And we have -- it's 200 people strong. But the productivity of this organization is just utterly incredible. They're doing fundamental and basic research across the entire stack of computing. The number of fundamental basic research computer science, industrial computer science organizations, research organizations that are working across the entire stack all the way from circuits, circuits to networks to parallel processor architectures, programming models, all the way to computer graphics and deep learning, those aren't that many. And this is 200 people strong across the world. And the thing that's really special about the way we do NVIDIA Research, it's really a hub-and-spoke system. They're collaborating and doing research with all of the product research teams. We have hundreds of researchers around the company.

All working together to advance the future of computing. Some of their contributions, you have already seen. NVIDIA Research initially started RTX. And the 10 years' worth of work that they've done has resulted in the NVIDIA RTX or NVIDIA OptiX and NVIDIA RTX technology; NVIDIA Switch that was described today, that made the world's largest GPU; cuDNN, the library that has now revolutionized deep learning all originated at NVIDIA Research. I'm so proud of these guys.

And so -- and you guys have seen some of the really amazing work -- the progressive GAN that was recently described, that caught a lot of attention because of the amazingly realistic faces that it is able to generate. And this is one of my favorite technologies, noise-to-noise denoising. That's one too many times you said noise. But it's just -- that's got to be incredibly good technology. It -- I saw it. I showed it to you earlier today, it's basically super rays. So wherever these black dots are, ray tracing hasn't completed yet. But somehow using artificial intelligence, we're able to predict what color and what light it should have been and put that up on the screen in advance, finishing the frame much, much faster.

Okay. So let me show you one thing here. This is -- what you're about to see here is this: imagine we taught a network how to draw. But we wanted to teach this network how to draw photorealistic images. And so -- what you're looking at here is, of course, a road. You can imagine that is a road, just like that, based on our intelligence and based on all of our experience we know that, that's a road. The pink, what is that color? Is that purple? Fuchsia? That's Fuchsia. Okay. Fuchsia is roads; blue is cars; posts are yellow; green is, obviously, trees; and gray, that's got to be buildings. And we take this image and we put it into the artificial intelligence network. And we say to it, produce a photorealistic image. And this is what it did. Okay, we trained this network to do this. And now it's going to create photorealistic images, all from this. What do you guys think?

Is that just amazing? So this artificial intelligence network is actually synthesizing information. And we're actually -- we're teaching it to synthesize roads, things that you would see on roads and synthesize it in a photorealistic way. Ladies and gentlemen, video research is now all over the world. I'm so proud of this team. Here is Seattle, Redmond, Santa Clara of course, Salt Lake City, Austin, St. Louis, Westford, Virginia, Charlottesville, Durham, North Carolina, Helsinki, Lund and Berlin. Some of the greatest cities around the world, 200 people strong. We're growing. Reach out to us. We love to work with you.

The transportation industry is the largest industry -- one of the largest industries in the world, $10 trillion large. And we believe that someday everything that moves will be autonomous or have autonomous capabilities.

As you guys know, there are several very important dynamics that are happening to the transportation industry. One, first of all, the cities are getting too crowded and people are having to move further and further away. The online purchasing phenomenon, the Amazon phenomenon, is causing more and more people to not go to the store. But to wait for the store to come to us. Somebody has to drive those atoms to us. And as a result of so much increase in population, another billion

people are going to come into society in the next 12 years. We already have over a billion parking spaces. We have 2 or 3 billion parking spaces waiting for us right now as we speak in anticipation of us needing it. And it's being built in the most precious places in the world, city centers. Autonomous vehicles has the opportunity to reduce the cost of transportation, make it easier for people to get around, live further away from the cities and work and, of course, revolutionize the way we design cities.

However, it's hard. Safety is the single most important thing. It is really, really hard technology. It's probably the hardest computing problem that the world has ever encountered. Safety is the most important thing. And as we're reminded just last week where the fatal accident that happened, we're reminded that this work is, ultimately, vitally important that we have to dedicate ourselves, commit ourselves to solving this problem. We have to solve it step by step by step. We've just incredible amounts of capability because we know so much is at stake and we have the opportunity to save so many lives in the future if we do it right.

And so we -- at NVIDIA, we're dedicating ourselves to this problem, the grandest of computer problems. Just imagine this. You've got these cars with all these different sensors. Some of them are radar that could sense motion, LiDAR that can measure depth. But don't see very well in light and snow and fog; cameras with super-high resolution. But can't see in the dark; infrared that can see in the dark. But doesn't have very high resolution. We take all of these sensors. And we fuse them together into a super sensor.

This sensor -- this super sensor, we call them sensor fusion, this super sensor could measure depth. It could recognize motion. It could detect things. It can understand the scene. It could see in the dark. But in order to do this and to compute it in such a high rate of speed as the car is moving and to take action so quickly is incredibly hard.

The algorithms, this entire system has never been created in totality not even once. This system has not even been once created in totality. We're dedicating ourselves to solving this problem. We're going to -- this is the ultimate high-performance computing problem, this is the ultimate deep learning problem, this is the ultimate AI problem and we think we can solve this. We think we can make it -- just an enormous contribution to this. And so it's a hard problem. The software is hard. The algorithms are hard. But it's also a safety problem. Meaning, these computers have to continue to operate well even when faults are detected. We have to be able to detect the faults when it happens and we have to manage the faults, meaning continue to manage the car properly and safely even after we've detected the fault.

The level of functional safety required to make this thing happen is just really, really high. A level that, the computer industry has really not dedicated itself to understand because this is something in the transportation industry, something that is, obviously, very well understood in the airplane industry. We've dedicated our last 5 to seven years to understand this entire system. And what NVIDIA is dedicating ourselves to do is to solve the problem from end to end.

There are 4 basic pillars to what we do. There is collecting of data. These cars are collecting petabytes and petabytes of data. We take these data and we have, well, collecting of data, I'm going to back to that. Training models. These models, of course, cars, roads, signs, people, bikes, trains, motorcycles, trees, hydrants, you name it, roads, pot holes, lanes, curbs, things are all over the world. The world is very diverse. We have to train our neural network with extreme precision to do this. And then we have to simulate it.

Simulation is going to be the foundation of success. And I'm going to talk about that in a second. Then lastly, drive it. Creating the driving computer that's the computer that has the horsepower to compute everything in real-time, energy efficient like all the energy efficiency comments that I've and examples that I've shown you because, ultimately, energy efficiency translates to battery life and battery life translates to mileage, miles of operation, which for many car companies translates directly to their earnings. And so battery life matters. Battery life also allows us to put multiple redundant computers inside a system.

The more energy efficient we are, the more computing systems we can put inside the car. So that it could be redundant, diverse and provide functional safety. We have to drive the car, the computer, the system software and all the algorithms on top. We're building the end-to-end systems. First, let me talk to you about perception. In fact, you can almost piece together all of the work that I've talked about so far into this system. It's called the NVIDIA Perception infrastructure.

We are building a supercomputing infrastructure in our company. There are 660 petaflops of supercomputing horsepower already in our company. This is a giant data problem. I've already just said, each one of these cars are collecting petabytes and petabytes of data every single week. And from that data, we label it in our data factory. We use all kinds of tools, AI tools and trained professionals to label each and every one of these images so carefully and the reason for that is this. We label it once. And cars in the millions get the benefit of it forever. And so we take labeling incredibly seriously. This is not a crowdsourcing problem, this is a professional labeling problem. We have 1,500 people. We're labeling about 1 million images per month. And we're just going to keep on labeling.

And from these labels, we're able to create the perception networks that you'll see of incredible levels. We train them on our DGXs that I just mentioned. It's not just a supercomputing problem, it's a software infrastructure problem.

In the future -- in the past, your source code is your code. In the future, your source code is your data. And all the code and all the training methodologies and recipes that you applied in order to create that model, because you want to be able to trace it, you want to be able to repeat it. This infrastructure inside NVIDIA is called the Perception infrastructure. It's a massive investment and it's just something that I'm so proud of.

At the end of it, it produces networks. We already have 10 networks inside our car. The self-driving car is not one network, it's a whole bunch of networks and there'll be a whole bunch more. We already have 10. By the time that we ship in a couple of years, call it 20 or 30, each one of the 10 networks, there are 10 DGXs assigned to one network and they could produce experiments every single day. Then, we have to validate it and verify it. Many of those DGXs will then have to be used for testing and verification against the entire body of work and the test data that we've stored up over time.

Let me show you some of the results. And so guys let's go and play it. And so here is a whole bunch of different networks that we've created. This is Perception network. You're seeing car detectors and the bounding boxes. You're going to see lanes. Not only do we detect things that we shouldn't collide with, we also have to detect spaces. So that we could -- that it's safe to drive to. Meaning, on the one hand, we detect obstacles. On the other hand, we detect free space. So between those 2 algorithms, we have redundancy and diversity. We have so much redundancy and diversity all over through our system. We do everything, 2 or 3 different ways and 2 or 3 different times. So that we can make sure that what we, ultimately do, is right.

Distance perception. Here's using an artificial intelligent network that was trained with LiDAR labeling. So that it could predict the distance of objects, whether it's -- we've trained it with so much data that it is now robust across all kinds of poor weather conditions.

We use cameras for object detection. We use radars for object detection. We also use LiDARs for object detection, not just how far away they are. But based on the pattern of dots that are coming back our way, the resolution of LiDAR is not as high as camera. But the depth information is really good. Based on the point cloud that comes back, we do point cloud processing on CUDA, which is incredibly good. We do ICP processing, Iterative Closest Point processing, through a registration and then we use deep learning to detect what those objects are.

And so LiDAR perception. We do scene perception. One of the hardest things isn't just to detect the light intersections where you're supposed to stop or be cautious aren't always properly labeled. And so we have the ability to detect scenes of all kinds of intersection. So that we could be much, much more alert.

Our path perception is not just about calculating the path. But calculating to -- perceiving the trajectory of the path that is most comfortable and safe.

LiDAR localization. You got to figure out exactly where you are in space compared to an HD map. You want to figure out where you are relatively using -- relatively in space using your camera as well. Camera localization even to HD map. So if you had an HD map, we could figure out where we are -- it's an additional sensor in fact.

Then, lastly, every single car is also a mapping car. If you have all of this technology inside your car that I just mentioned, your car becomes a mapping car, which means

that when you drive it in the beginning, every time you drive it, you're mapping your space. And so as we -- as you know, 99% of the time, we're driving the same route. And most of those routes won't have an HD map in the beginning. And so we'll use our own cars -- your car to map and create a really high fidelity 3-dimensional map. Then after that, you could use it as an extra sensor to give you additional information to localize within it for safety. Okay? So camera-based mapping.

Why don't we show them, what we've done is every time we're together, I show you a homemade video. And this homemade video are some of the latest clips, if you will, the latest clips of some of our recent achievements. We're making progress all the time. We are far from finished.

There are several thousand engineers right now working on autonomous vehicles. We're going to work on this for another 2 or three years before we ship in volume cars. In the meantime, of course, we're going to develop and offer development systems and training systems, simulation systems of all kinds. But this is going to be a long journey that is step by step by step.

As I mentioned, this is really one of the most complex computing problems we've ever encountered because of the diversity of the problem, the complexity of problems, all the different types of algorithms involved and multiplied exponentially as a result of safety concerns.

And so let me just show you some of the latest clips. Guys?

(presentation)

Wow, God, I love you guys. It's amazing how much -- as I'm watching those, just as -- all the software and all the technology that has been created to make all of this stuff happen is just -- it's just -- is washing over me. And it's truly an amazing thing. And the reason why it's so hard, the question is why don't we just put up a car and demo a car? Well it turns out, our challenge is we're trying to create a AV computing system and an AV computing -- AV driving flow and infrastructure. So that the entire transportation industry can take advantage of this massive investment that we're making and create the future of autonomous vehicles.

And as you know, there are all kinds of different sizes and shapes and price points of cars. But every one of them should have the advantages of autonomous capability to either help you drive or keep you safe while you drive. Notice, we showed you earlier Janine -- a video of Janine and she is driving. She is pretending like she is sleepy, she is never sleepy. I've never seen her sleepy once. And so she is pretending like she is sleepy. And our AI inside the car is monitoring her, making sure that she is looking in the right direction that her gaze is still active that somehow she doesn't appear sleepy making sure that, that everything in the cabin is safe and nobody is doing crazy stuff.

And so using driver monitoring and cabin monitoring and AI, we're able to make sure that not only is the car doing the right thing. But the passengers inside are doing the right things. This entire infrastructure, this entire system is just such a great undertaking. And we're doing this -- we're doing this in a way that is mindful of the fact that we're not trying to build a chip, we're not trying to build a AV system. That in the end, this is an AV computer. And as you know, computer are software defined. One of the most powerful things about today's cloud computing is that the entire cloud infrastructure is software defined. One of the most amazing things about today's cellphones versus yesterday's feature phone is that today's cellphones are software defined. One of the reasons why the PC has become such a revolution instead of a typewriter is because the PC is software defined. Architecture matters.

If you're a software engineer, you care about the hardware architecture. And because NVIDIA is not a chip company, we are a computer architecture company, we're a software company, we care about computer architecture. We care about the fact that the software will run on all of these different computers for as long as we shall live. So that we can make the software better and better and better over time. And the entire install base of fleets and cars that we deployed gets better and better and better and safer and more comfortable over time.

And so we created one architecture road map. And it, basically, starts like this. Two years ago, we introduced DRIVE PX Parker and we created -- that's a 1-chip solution for a very basic car and then we created a 4-chip solution for DRIVE PX 2, which is the development system for most of the autonomous vehicles that you see around the world.

DRIVE PX 2 at the time, I -- when I brought it out, we were laughed out. Our -- people laughed at me for bringing out such a large computer. Well it turns out I did make a mistake. I didn't make the computer large enough. And the reason for that is because the amount of software that the world needed to develop is far more than what anybody realized. And so our basic methodology is this. Then, two years later, this year, we created the DRIVE Xavier. Basically, we took 4 chips. The computational capability of those 4 chips in those 300 watts. And we shrunk it into one miracle chip.

Xavier is the most complex SOC the world has ever created. It's great for high-performance computing, it's great for deep learning, it's great for computer vision, sensor fusion. It is the largest single chip we have ever made short of Volta, 9 billion transistors, unbelievable amount of engineers worked on it. Xavier is now sampling.

Then, from Xavier, we added -- we created a 4-chip system. Two Xaviers and 2 Voltas together, another running at a very low temperature and voltage, another 300-watt computer emerged. And that 300-watt computer is being used in robot taxi and self-driving car system development all over the world. We're sampling both of them today. Both of them will be in production by the end of this year. Completely architectural compatible, auto grade to the extreme temperatures and operating conditions of a car. It is super energy efficient because battery life matters, redundancy matters and it is also ASIL D. This is the first lineup of computers that

NVIDIA has ever made. That stands up to the highest standard of functional safety, ISO 26262, ASIL D is a rigor like you have never seen before. From processed -- because ASIL D has only been done on small computers, not computers of this type and not on a software stack of this complexity. This is just a gigantic investment on our part. But is so important and we've got to do this right.

All the engineers that are working on this, I want to thank you so much for doing this, everything from the architecture to the design, the resilience, the redundancy, the diversity, the -- even the documentation so that it can, ultimately, be certified and traced all the way back if anything were -- if anything were to happen because 100% of all the IP and all of those chips were created here in NVIDIA. We have the ability to produce traceability for as long as we shall live. This is just a gigantic investment. And we take it so seriously. But we're not stopping here.

DRIVE Pegasus as powerful as it is. Multiple of them are being used in self-driving cars, driverless cars. And the reason for that is because there are so many LiDARs. So many cameras. So many radars that are being deployed. So we're not stopping here.

Our next generation is called Orin. And we're going to take, basically, those 8 chips, 2 DRIVE Pegasuses and we're going to put it into a couple of Orins, incredible amounts of computation in one little chip. And of course, with Orin, there will be discrete GPUs that we will continue to build. There are some code names, I won't share with you today. I'll save some surprises for next time. It will create also an amazing processor for autonomous driving cars. Auto grade, super energy efficient, fully ASIL D, huge investment, NVIDIA's drive road map.

Let me tell you why this problem is just so incredibly hard? Well it turns out, the world civilization drives about 10 trillion miles per years, okay? 10 trillion miles, just -- you've to get the number in your head. It's 10 trillion miles. It's a 1,000 billion -- it was 10,000 billions, 10 trillion miles. And in the United States, 770 accidents happen from 1 billion miles, 770 accidents happen from 1 billion miles. The amount -- the safety work that has been done throughout the years in the United States has really reduced the amount of fatalities and accidents. And yet, it would take you driving 1 billion miles, society driving 1 billion miles in the United States to produce 770 accidents. And so the question you have to ask yourself is how confident are you when your car -- your fleet of test cars of 20 over one year has driven about 1 million miles. It takes 20 cars driving test drive all year long to test drive 1 million miles. And yet, it takes 1 billion miles to have experience 770 accidents. And we're trying to build a system that is better than humans at driving.

And so clearly, the amount of coverage -- scenario coverage and miles coverage is just not possible in real life. And this is where NVIDIA's skill can really shine. We know how to build virtual reality worlds. And so we imagine this that in the future, there will be thousands of these virtual reality worlds with thousands of different scenarios running at the same time. And our AI car is navigating itself and driving and testing itself in all of these virtual reality worlds. And if any of those tests were to fail, it would send us an e-mail, we would jump into it in virtual reality and figure out what's going

on. Do you guys see that? So we have all these virtual reality worlds. Well let me show it to you today.

Mark? So what we're going to show you is this. This is a virtual reality simulator. And before you start, this is the front view, your side mirrors and your rearview mirror. Okay. Go ahead, Mark. Hi. Mark, how's it going?

Mark Daly {BIO 20677145 <GO>}

Hi. good, Jensen.

Jensen Hsun Huang {BIO 1782546 <GO>}

One of NVIDIA's veterans.

Mark Daly {BIO 20677145 <GO>}

Let's start this thing. We're -- so we're going to -- we're jittering our world. We're running 4 simultaneous cameras, as you mentioned. We've got a pretty aggressive driver here in front of us. We defined the scenario. It's causing all the traffic to jam up. But as you mentioned, we got all the cameras. So we're running on multiple GPUs with a scalable system as cars get more and more sensors, more cameras, more LiDARs. We're going to need the horsepower of multiple GPUs. And we're showing that to you right here.

Jensen Hsun Huang {BIO 1782546 <GO>}

Okay? So we could do, of course -- we could simulate daylight scenarios. Let's simulate other scenarios.

Mark Daly {BIO 20677145 <GO>}

Okay. Let's -- why don't we turn the lights on and go to night. So here is the same setup. We now are doing all the dynamic lighting. So the headlights of the cars, the light posts that are lighting the road, the lights under the bridge as we drive under here, these are all dynamic real-time lights.

Jensen Hsun Huang {BIO 1782546 <GO>}

And the key here is that the fidelity of the simulation has to be sufficiently high that the sensors stack all the software that we create would just operate as they would in real life.

Mark Daly {BIO 20677145 <GO>}

That's exactly right.

Jensen Hsun Huang {BIO 1782546 <GO>}

Right. So -- and every one of the sensors needs to have GPU associated with it to generate its view of the universe, its view of the world that it's in. And So today, we're simulating with 4 GPUs, with 4 cameras. But obviously, we have the ability to spawn off a whole large number of GPUs. So if you have 8 LiDAR systems around your car, you have 8 cameras around your car, you have 6 radars, we have the ability to generate all of that and feed that information into our sensors. And if we were successful in doing that, our artificial intelligence network, all of the network, all the software that we ran, that we developed should just work. So Mark, let's turn that on.

Mark Daly {BIO 20677145 <GO>}

That's right. Okay. So let's, here we go. Now this is what our drive stack. It's sitting on DRIVE PX. This is all the detection of the cars in the world. The lane detection you're seeing going on. And we've been self-driving this whole time by the way. So as we ran into that little traffic jam, the DRIVE PX detected it and slowed down and kept us safe.

Jensen Hsun Huang {BIO 1782546 <GO>}

And notice, we're detecting all the cars, we're detecting all the lanes. But here's the amazing thing. Earlier the video that you saw were real cars in the real world, we didn't change one line of code iota. The code is exactly the same code that was developed by the engineers. We, of course, copy it over this, send it over to this data center and it's running on exactly the same DRIVE PX, DRIVE Xavier and DRIVE Pegasus computer that we would drive in the car. And as a result, it's exactly the same sensor response. And this -- the perception, the localization, the planning stack, all work exactly the same way. And now you're at the end of your...

Mark Daly {BIO 20677145 <GO>}

So we reached the end of that scenario. Can I show you another one?

Jensen Hsun Huang {BIO 1782546 <GO>}

Yes. Show me another scenario.

Mark Daly {BIO 20677145 <GO>}

Okay. So let's switch to the other system.

Jensen Hsun Huang {BIO 1782546 <GO>}

And so this is one of the beautiful things. In virtual reality, we could create whatever scenario we want. And as you know, in test car driving, you could go days, months, weeks and never run into a weird scenario.

Mark Daly {BIO 20677145 <GO>}

So there is...

Jensen Hsun Huang {BIO 1782546 <GO>}

And so we could create all kinds of strange and cornered conditions.

Mark Daly {BIO 20677145 <GO>}

Right.

Jensen Hsun Huang {BIO 1782546 <GO>}

In virtual reality. Go ahead, Mark.

Mark Daly {BIO 20677145 <GO>}

I'm sorry. So the same avid serial driver just came through. And now what we found one of San Jose's finest is in pursuit of him.

Jensen Hsun Huang {BIO 1782546 <GO>}

You see that the cop coming up on the back? The policeman coming up on the back and is coming up on the side, there you go.

Mark Daly {BIO 20677145 <GO>}

So -- and, of course, all the dynamic lighting. So you're seeing the red and the blue lights light up the world. Some of that strobing effect could affect the sensor. So we need to make sure we're modeling that correctly.

Jensen Hsun Huang {BIO 1782546 <GO>}

That's really terrific. And so one, the first is to recreate virtual reality in all kinds of weather conditions and day and night conditions. It has to be -- it has to have the fidelity and the performance that basically simulates reality. Two, that simulation of reality has to be so high and the computer itself has to be able to run the original unchanged software. So that we could test everything as is inside a virtual reality world. And when it passes that, then we can take it and put it into the car completely unchanged. We just OTA it directly into the car and it should just work. Then, third, we could use it to create extreme corner rare scenarios and it's completely repeatable. Every single time, that police car shows up exactly the same way.

Mark Daly {BIO 20677145 <GO>}

That's right.

Jensen Hsun Huang {BIO 1782546 <GO>}

And so we could have the ability, if -- with repeatability, engineers could debug systems.

Mark Daly {BIO 20677145 <GO>}

That's right.

Jensen Hsun Huang {BIO 1782546 <GO>}

These 3 fundamental capabilities are made possible, we call this DRIVE Sim.

Mark Daly {BIO 20677145 <GO>}

That's right.

Jensen Hsun Huang {BIO 1782546 <GO>}

Ladies and gentlemen, Mark Daly. And so DRIVE Sim, DRIVE Sim, DRIVE Sim, just imagine the amount of capable technology that has been brought to bear. DRIVE Sim, the simulator. But most importantly, the computer inside it and all the driving stacks. And so you got the image generator. This image generator is generating the world. Now this image generator in the context of a video game, this image generator is a gaming PC and the person that would be driving the car would be a person. A person driving the racing simulator. But no. We'd replace that person with a artificial intelligence, AV computer, a self-driving car computer. Inside the self-driving car computer is the DRIVE Xavier and the DRIVE Pegasus and they're connected. Literally, the output of the image generators go into and there's a sensor adapter and it goes directly into the drive computers. The exact same software stack runs and it performs the right actions, perceives the world, performs the right actions, sends back the driving control to the image generator, which controls this virtual reality world.

We will have thousands of -- we call this DRIVE Constellation. Pegasus in the sky, a constellation of them. Pegasus in the sky, a constellation of them. We're going to have thousands of constellations. These virtual-reality worlds, will all be running simultaneously. And hopefully, we could cover a large, large coverage of scenarios. With just 10,000 constellations, we can cover 3 billion miles a year. Incredible.

So ladies and gentlemen, this is the NVIDIA DRIVE Sim and our Constellation. And it is how we're going to bridge the gap between test -- actual test driving and the trillions of miles and the billions of miles that we need to experience over time.

Well as I mentioned, NVIDIA has created a platform, it's an open platform. And it takes a community to build the future of autonomous driving. It's about the car companies, it's about the trucking companies, about mobility service companies. It's

about the Tier 1 suppliers that are expert in building these systems and components for car companies. But it's about mapping companies, it's about sensor companies. We work so closely with all the LiDAR companies and all the camera companies to make sure that these future sensors are as high fidelity, as high performance and as robust as possible and that the algorithms they use are as efficient as possible. We work with startups all over the world.

NVIDIA's platform is an open platform. Last year, we had 320 partners. The year before that, we had 250 partners. And today, we have 370 partners. We're working with people all over the world. This is a one of the world's largest industries, we're making the fundamental investment for the future of AV computing, the AV computer, all the software driving stack and the end-to-end development system that will support this fleet and this industry for as long as we shall live and continue to make it better over time taking it step-by-step as we are so mindful of safety.

Well the self-driving car is the first. In the future, robots will augment and will assist us in all kinds of different places. One of the areas that we already talked about is, of course, the Amazon effect. We used to drive to pick up groceries. In the future, groceries will come to us. And this is going to happen in all kinds of different industries. Artificial intelligence and autonomous machines will help revolutionize just about every single industry, making us more productive, making us more efficient, do things that are otherwise impossible to do or too laborious to do or just simply too hard to do.

And so we've created a platform like we've done with DRIVE. We call this Isaac. Today, we're going to release the first little version of that. This is such a great achievement. And so basically in the future, you're going to be doing simulation as a development system because these robots have to work in a 3-dimensional world. And so we created a 3-dimensional world simulator, we call it the Isaac Lab. In the Isaac Lab, we will develop the perception capability, the localization capability, the mapping capability and the planning capability that is necessary for robots and autonomous machines to navigate these complex worlds. And so we already have our deep learning systems that I've mentioned. Then we take -- because it's hardware in the loop, that hardware is driving the simulator. When you are done with the simulator, that hardware should be able to run the entire software stack and we call that the Isaac SDK. The Isaac SDK running on top of our little computer, we call Jetson. We can deploy that. We made a quick little video for you. All you have to do is come to this website. And we'll take -- and to let us know if you're interested and we'll get you going on an Isaac SDK.

So ladies and gentlemen, this is the Isaac robotics platform ready to come to the world.

(presentation)

Ladies and gentlemen, Isaac, our new platform, the future of robotics. This is a development system. Lots of software developed for it. And of course, a simulator

that goes along with it. And this little tiny computer we call Jetson. It's been put together for use to start developing the future of autonomous machines.

Look, I have one thing that I want to share with you. This is something we've been working on and it's just so exciting. Now it's never been seen in the world before and it's far from finished. It's far from finished. And I thought that as a close, I would invite you into NVIDIA's Lab.

NVIDIA does a lot of things in the Holodeck today. We'd create the future in virtual reality because that's the only place you could create it. And we created this thing called Holodeck, as you guys know. And this project, this project is really about thinking through the entire path of how to deploy the self-driving cars into the world. And as you know, when you have a car that is driving by itself in autopilot mode, the driver is the backup system. The driver is the backup system.

If you have a driverless car, who is the backup system? And so the backup system likely will be another person. We would like humans to always be the backup system of AI. And so the question is how would -- how do we create -- how do we solve this problem where an autonomous machine has no operator? It could be a little tiny robot like Carter that you saw earlier. It could be a delivery vehicle, it could be a little agriculture, a little farming equipment, a little caterpillar tractor that's roaming around doing things and there are no operators. And so the question is, how do we create the backup system if something were to happen that we have to go and if you will, get it out of harm's way? Okay. And so we imagine this. We know that virtual reality has the ability to teleport us. Virtual reality has the ability to teleport us into a new world, into a different world. And so before I -- I'm going to show you a little something. This is Justin, Justin Ebert, please. Justin Ebert. Justin is going to help me with this demo. We are currently in the Holodeck. I'm going to explain a little bit more as we go through, okay? But imagine there is a car in the world somewhere and we need to help it. And so the question is, how do we do that?

Justin Ebert

Thanks, Jensen. Let's check it out.

Jensen Hsun Huang {BIO 1782546 <GO>}

So we're in Holodeck. We just created a virtual reality car.

Justin Ebert

We call this drive lab. And this is where we have a virtual car where we can take over. Our operator, Tim here, is sitting on the side stage and we can see sensors coming in from the vehicle, the remote vehicle that needs help.

Jensen Hsun Huang {BIO 1782546 <GO>}

Hi. don't go anywhere, Justin -- hey...

Justin Ebert

Oh, we're not going anywhere.

Jensen Hsun Huang {BIO 1782546 <GO>}

Yes. Okay. All right.

Justin Ebert

Yes, yes. We're not going anywhere. We're going to wait and talk this through. So you can see there is a...

Jensen Hsun Huang {BIO 1782546 <GO>}

This is so incredible, you guys. This is just utterly incredible. First of all, I'm going to hand it back to you in just a second.

Justin Ebert

Yes, no problem.

Jensen Hsun Huang {BIO 1782546 <GO>}

What you're looking at is live.

Justin Ebert

Yes.

Jensen Hsun Huang {BIO 1782546 <GO>}

This is Holodeck. This is Holodeck. There's a virtual reality car, okay? The virtual reality driver is sitting right there. His name is (Tim Wong), Tim, could you just raise your hand? There you go. Tim is the virtual reality driver. He is inside this virtual reality car. He is inside the Holodeck. He is not with us. That's just a sack of humanity, we call Tim Wong, okay? And so Tim is currently in Holodeck. He has got -- he is looking at this virtual world. And you see the video that's being piped in. That is literally live. That is live right now. Okay? All right. Justin, take it away.

Justin Ebert

Yes, yes. no problem. So -- yes, as Jensen said, this is actually out back here and -- in a private area. And let's bring in Tim. So we can show him taking over inside of this vehicle here. And let's show that he has control over the car as well. So there's the real car out there. (inaudible) You can turn the steering wheel a little bit. We can see that we've got...

Jensen Hsun Huang {BIO 1782546 <GO>}

Oh, come on, cut it out.

Justin Ebert

Come on, that's fun. Yes. So he engaged...

Jensen Hsun Huang {BIO 1782546 <GO>}

Are you guys putting this in your head right now? There is Tim in reality. But he is inside virtual reality controlling a car that's in reality. But it's not here. Are you guys following me? Okay. It's like 3 layers of inception right now. The mind clearly exploded.

Justin Ebert

We've engaged remote drive. You can see. So he knows he has...

Jensen Hsun Huang {BIO 1782546 <GO>}

I still remember when we were talking about this, creating this, the people that we're explaining to, we're going to build this thing, it was like, okay, explain to me one more time.

Justin Ebert

Yes. You can't understand until you see it. Then it all makes sense. Okay. So I think we've seen this. Let's take a look from Tim's perspective for a second here. And we can have him take off and actually drive the car. As you can see, he is blocked. There is a vehicle here that's doing some unloading. And Tim look down just a little bit. So we can see your perspective. And he can now operate this vehicle around this obstacle and maybe take it to a safe spot in a parking lot for us. And hopefully -- so we're doing this very slowly and safely in a cordoned off area here. And he's going to try to turn off into the parking lot. Now you can see a little bit on the left here. This view is...

Jensen Hsun Huang {BIO 1782546 <GO>}

Tim isn't there. Guys, there is Tim right there. The invisible man.

Justin Ebert

And Tim's view is very broad. He can see all 3 screens and get a full perception of everything that's going on in the car. In the future, we can represent all kinds of LiDAR systems and everything slowly, slowly. So yes, maybe we could take a look there.

Jensen Hsun Huang {BIO 1782546 <GO>}

Look at him. There he goes. Go, Tim.

Justin Ebert

There's the team racing after the car there. And see if we...

Jensen Hsun Huang {BIO 1782546 <GO>}

And he -- and look, what Tim is experiencing right now, he is sitting inside the car and he feels like he is inside the car. And he feels the car around him. And that's why he is able to -- look at that, he parked in a parking space.

Justin Ebert

Nice job, Tim.

Jensen Hsun Huang {BIO 1782546 <GO>}

Hi, guys, thank you, thank you. You guys are amazing. Thanks, guys. Ladies and gentlemen, what do we call this? I have no idea. Yes. That's right, we call this amazing. So guys, I'm so proud of you. This is such incredible work and this is such important work. And as we know, one of the amazing capabilities of the future of autonomous machines is we can go somewhere that we otherwise can't. We could be fixing something. The robot could be fixing something, rescuing someone. And now we could literally, through VR, teleport ourselves into the mind of the autonomous machine and be there. Does that make sense? Teleportation, teleportation and VR is our way of doing that. VR augmented with autonomous machines is our way of doing that, teleportation. The future has arrived. Pretty exciting. Thanks a lot, guys. I'm so proud of you.

Okay. We covered a lot of stuff today. We talked about a brand new, a revolutionary computer graphics technology, we call RTX. Our dream come true after all these years from architecture to algorithm, to AI, to integration, into all of the APIs in the world and the tools around the world. We have introduced real-time ray tracing to the world finally, we call it NVIDIA RTX. And it will run on a revolutionary workstation, called the Quadro GV100 with 32 gigabytes. We also announced our brand-new NVIDIA AI platform. All of the new updates from the largest GPU in the world, $1.5 million of value. All for my friends for $399,000. And NVIDIA's DGX-2, 2 petaflops of computing, incredible amounts of memory, 512 gigabytes of HBM2. TensorRT and integration -- deep integration into TensorFlow and Kaldi has expanded our ability to reach the hyperscale workflows around the world. In aggregate, speeding it up by 100x, which allows us to enable people to buy more GPUs and save more money. And not only that, we made it easy to deploy.

Kubernetes is now on NVIDIA GPUs, Kubernetes on NVIDIA GPUs. Just really, really exciting. And don't forget what makes TensorRT such an incredible achievement and

what inference is so hard is because of one word, you guys know that word, PLASTER, that's right. And so what a great audience, what a great audience.

Then, lastly, on our driving stack, we introduced our one architecture, we are in -- we'll be in production by the end of this year with our first full ASIL D, full driving computer called Xavier and Pegasus and the next generation called Orin. But one of the most important things we can imagine doing is to create a simulation environment. So that the automotive industry can cover the gap between 1 million cars -- 1 million miles and 1 trillion miles. That level of -- that many -- that orders -- that many orders of magnitudes of difference and coverage has to be covered in some clever way. And we call that DRIVE Sim, running on our server we call DRIVE Constellation. And we introduced 2 new platforms. And I'm going to talk about more and more and more every single year so that we can engage these new industries. One of them is Isaac, our robotics platform and CLARA, our medical imaging supercomputing platform.

Ladies and gentlemen, it's great having all of you here. Thanks for your support. And have a great GTC.