REFINITIV STREETEVENTS

EDITED TRANSCRIPT

NVDA.OQ - NVIDIA Corporation Presents at COMPUTEX 2023, May-29-2023 11:00 AM

EVENT DATE/TIME: MAY 29, 2023 / 3:00AM GMT

REFINITIV

CORPORATE PARTICIPANTS

Jensen Huang

PRESENTATION

Operator

(presentation)

Unidentified Participant

Ladies and gentlemen, please welcome NVIDIA Founder and CEO, Jensen Huang. (foreign language)

Jensen Huang

(foreign language) We're back. Our first live event in almost 4 years. I haven't given a public speech in 4 years. Wish me luck. I have a lot to tell you, very little time, so let's get going.

Ray tracing. Simulating the characteristics of light and materials is the ultimate accelerated computing challenge. Six years ago, we demonstrated, for the very first time, rendering this scene in less than a few hours. After a decade of research, we were able to render this scene in seconds, 15 seconds on our highest-end Huston GPU six years ago. And then we invented NVIDIA RTX and combined 3 fundamental technologies: hardware-accelerated ray tracing, artificial intelligence processing on NVIDIA Tensor Core GPUs and brand-new algorithms.

Let's take a look at the difference in just 5 years. Roll it. This is running on CUDA GPUs six years ago, rendering this beautiful image that would have otherwise taken a couple of hours on a CPU. So this was a giant breakthrough already. Enormous speed up running on accelerated computing.

And then we invented the RTX GPU. Run it, please.

(presentation)

Jensen Huang

The holy grail of computer graphics ray tracing is now possible in real time. This is the technology we have put into RTX. And this, after five years, is a very important time for us because for the very first time, we took our third-generation ADA architecture, RTX GPUs and brought it to the mainstream with 2 new products that are now completely in production.

Thank you. Hey, I got it backwards. Can I have it this way? I got that backwards. Everything looks different, inside out and upside down. Okay. This is our brand new -- right here, you're looking at an Ada GPU running ray tracing and artificial intelligence at 60 frames a second. It's 14-inch, it weighs almost nothing. It's more powerful than the highest end PlayStation.

And this is the RTX 4060 Ti for our core gamers. Both of these are now in production. Our partners here in Taiwan are producing both of these products in very, very large productions, and I'm really excited about them. Thank you very much.

I can almost put this is in my pocket.

AI made it possible for us to do that. Everything that you saw would have been utterly impossible without AI. For every single pixel we render, we use AI to predict 7 others. For every pixel we compute, AI predicted 7 others. The amount of energy we save, the amount of performance we get

is incredible. Now, of course, I showed you the performance on those 2 GPUs, but it wouldn't have been possible if not for the supercomputer back at NVIDIA, running all the time, training the model so that we can enhance applications. So the future is what I demonstrated to you just now.

You can extrapolate almost everything that I'm going to talk about for the rest of the talk into that simple idea, that there will be a large computer writing software, developing and deploying software that is incredible that can be deployed in devices all over the world.

We used AI to render this scene. We're going to also use AI to bring it alive. Today we're announcing NVIDIA ACE, Avatar Cloud Engine, that is designed for animating to bringing a digital avatar to life. It has several characteristics, several capabilities: speech recognition, text-to-speech, natural language understanding, basically a large language model, and using the sound that you will be generating with your voice, animate the face and using the sound and the expression that you're saying animate your gestures. All of this is completely trained by AI.

We have a service that includes pretrained models that you can come, developers can come and modify and enhance for your own application, for your own story because every game has a different story. And then you can deploy it in the cloud or deploy it on your device. Has a great backend, has a TensorRT. TensorRT is NVIDIA's deep learning, optimizing compiler, and you could deploy it on NVIDIA GPUs as well as output OnyX and industry standard backend so that you can run it on any device.

Let's take a look at this scene in just a second, but let me first tell you about it. It is completely rendered with ray tracing. Notice the beautiful lights. So many different lights and all of the different lights are projecting light from that source. So you have all kinds of direct lights. You have global illumination. You're going to see incredibly beautiful shadows and physics simulation. And notice the character, the beautiful rendering of the character. Everything is done in Unreal Engine 5. We partnered with a avatar framework and avatar toolmaker called XXComv AIXX. And together, we developed this demo you're about to see. Okay. Run it, please.

(presentation)

Jensen Huang

None of that conversation was scripted. We gave that AI, this Jin guy character, a back story. His story about his ramen shop and the story of this game. And all you have to do is go up and talk to this character. And because this character has been infused with artificial intelligence and large language models, it can interact with you. He can understand your meaning and interact with you in a really reasonable way. All of the facial animation completely done by the AI.

We have made it possible for all kinds of characters to be generated. They're all domain. They have their own domain knowledge. You can customize it so everybody's game is different. And look how wonderfully beautiful they are and natural they are. This is the future of video games. Not only will AI contribute to the rendering and the synthesis of the environment, AI will also animate the characters. AI will be a very big part of the future of video games.

The most important computer of our generation is unquestionably the IBM System/360. This computer revolutionized several things. The first computer in history to introduce the concept of a central processing unit, the CPU; virtual memory; expandable I/O, multitasking; the ability to scale this computer for different applications across different computing ranges. And one of the most important contributions and one of its greatest insights is the importance of preserving software investment.

The software ran across the entire range of computers, and it ran across multiple generations so that the software you develop, IBM recognized the importance of software, recognized the importance of preserving your investment and very importantly, recognized the importance of installed base.

This computer revolutionized not only computing -- and many of us grew up reading the manuals of this computer to understand how computer architecture worked, to even learn about DMA for the very first time. This computer not only revolutionized computing, it revolutionized the thinking of the computer industry. System/360 and the programming model of the System/360 has largely retained until today, 60 years. In 60 years, $1 trillion worth of the world's data center all basically use the computing model that was innovated all the way 60 years ago until now.

There are 2 fundamental transitions happening in the computer industry today. All of you are deep within it and you feel it. There are 2 fundamental trends. The first trend is because CPU scaling has ended. The ability to get 10x more performance every 5 years has ended. The ability to get 10x more performance every 5 years at the same cost is the reason why computers are so fast today. The ability to sustain 10x more computing every 5 years without increase in power is the reason why the world's data center hasn't consumed so much more power on earth. That trend has ended, and we need a new computing approach, and accelerated computing is the path forward.

It happened at exactly the time when a new way of doing software was discovered, deep learning. These 2 events came together and is driving computing today, accelerated computing and generative AI. This way of doing software, this way of doing computation is a reinvention from the ground up, and it's not easy. Accelerated computing is a full stack problem. It's not as easy as general-purpose computing. The CPU is a miracle, high-level programming languages, great compilers, almost anybody could write reasonably good programs because the CPU is so flexible.

However, its ability to continue to scale and performance has ended, and we need a new approach. Accelerate computing, is full stack. You have to reengineer everything from the top down and from the bottom up, from the chip to the systems, to the system software, new algorithms and of course, optimizing the applications.

The second is that it's a data center scale problem. And the reason why it's a data center scale problem is today the data center is the computer. Unlike the past when your PC was a computer or the phone was a computer, today, your data center is the computer. The application runs across the entire data center. And therefore, it's vital that you have to understand how to optimize the chips, the compute, the software across the NIC, the switch, all the way to the other end in a distributed computing way.

And the third, accelerated computing is multi-domain. It's domain specific. The algorithms and the software stacks that you create for computational biology and the software stack you create for computational fluid dynamics are fundamentally different. Each one of these domains of science need their own stack, which is the reason why accelerated computing has taken us nearly 3 decades to accomplish.

This entire stack has taken us nearly 3 decades. However, the performance is incredible, and I'll show you. After 3 decades, we realize now that we're at the tipping point. A new computing model is extremely hard to come by. And the reason for that is this: in order for there to be a new computing model, you need developers. But a developer would only come if they're -- and developers have to create applications that end users would buy. And without end users, there would be no customers, no computer companies to build computers. Without computer companies like yourself building computers there would be no installed base. Without installed base, there would be no developers. Without developers, there'll be no applications.

This loop has been suffered by so many computing companies in the 40 years that I've been in this industry. This is really one of the first major times in history a new computing model has been developed and created. We now have 4 million developers, 3,000-plus applications, 40 million CUDA downloads in history, 25 million just last year. 40 million downloaded in history, 25 million just last year. 15,000 start-up companies in the world built on NVIDIA today, building on NVIDIA today and 40,000 large companies, enterprises around the world are using accelerated computing.

We have now reached the tipping point of a new computing era. This new computing model is now enjoyed and embraced by just about every computer company and every cloud company in the world.

There's a reason for that. It turns out that every single computing approach, its benefit in the final analysis is lower cost. The PC revolution that started and that Taiwan enjoyed in 1984, starting in 1984, the year I graduated, that decade in the '80s was the PC revolution. PC brought computing to a price point nobody's ever seen before. And then, of course, mobile devices was convenient, and it also saved enormous amounts of money.

We aggregated and combined the camera, the music player, your PC, a phone. So many different devices were all integrated into one. And as a result, not only are you able to enjoy your life better, it also saves a lot of money and great convenience. Every single generation provided something new and saved money.

Well, this is how accelerated computing works. This is accelerated computing used for large language models. For large language models, basically the core of generative AI. This example is a $10 million server, and we costed everything. We costed the process, we costed all the chips, we costed

all the network, we costed literally everything. And so $10 million gets you nearly 1,000 CPU servers and to train to process this large language model takes 11 gigawatt hours, 11 gigawatt hours, okay? And this is what happens when you accelerate this workload with accelerated computing.

And so with $10 million, for a $10 million server, you buy 48 GPU servers. It's the reason why people say that GPU servers are so expensive. Remember, people say GPU servers are so expensive. However, the GPU server is no longer the computer. The computer is the data center. Your goal is to build the most cost-effective data center, not build the most cost-effective server.

Back in the old days, when the computer was the server, that would be a reasonable thing to do. But today, the computer is the data center. And so what you want to do is you want to create the most effective data center with the best TCO. So for $10 million, you buy 48 GPU servers. It only consumes 3.2 gigawatt hours and 44x the performance.

Let me just show it to you one more time. This is before, and this is after. And this is -- we want dense computers, not big ones. We want dense computers, fast computers, not big ones. And so that's ISO budget.

Let me show you something else. Okay. So this is $10 million again, 960 CPU servers. Now this time, this time, we're going to be ISO power. We're going to keep this number the same. We're going to keep this number the same, okay? So this number is the same. The same amount of power. This means your data center is power limited. In fact, most data centers today are power limited. And so with being power limited, using accelerated computing, you can get 150x more performance with 3x more cost.

But why is that such a great deal? - the reason for that is because it's very expensive and time-consuming to find another data center. Almost everybody is power limited today. Almost everybody is scrambling to break new ground to get more data centers. And so if you are power limited or if your customers are power limited, then what they can do is invest more into that data center, which already -- which has 11 gigawatts, and you can get a lot more throughput, continue to drive your growth.

Here's another example. This is my favorite. If your goal is to get the work done, if your goal is to get the work done, you don't care how. Your goal is to get the work done. You don't care how. And this is the work you want to get done. ISO work, okay? This is ISO work. Look at this, right?

(foreign language) Taiwanese people love that, right? Nice to see you, Carol. Nice to see you, Spencer. Okay. So let's do that one more time. It's so delightful. Look at this, look at this. Before, after. The more you buy, the more you save. That's right. The more you buy, the more you save. The more you buy, the more you save. That's NVIDIA. You don't have to understand the strategy. You don't have to understate technology. The more you buy, the more you save. That's the only thing you have to understand.

Data center. Now why is it you've heard me talk about this for so many years? In fact, every single time you saw me, I've been talking to you about accelerated computing. I've been talking about accelerated computing, well, for a long time, well over 2 decades And now why is it that finally is the tipping point?

Because the data center equation is very complicated. This equation is very complicated. This is the cost of building a data center. The data center TCO is a function of -- and this is the part where everybody mess up. It's a function of the chips, of course, no question. It's a function of the systems, of course, no question. But it's also, because there are so many different use cases, it's a function of the diversity of systems that can be created.

It is the reason why Taiwan is at the bedrock, at the foundation of the computer industry. Without Taiwan, why would there be so many different configurations of computers, big, small, powerful, cheap, enterprise, hyperscale supercomputing? So many different types of configurations, 1U, 2U, 4U, right? And all completely compatible.

The ability for the hardware ecosystem of Taiwan to have created so many different versions that are software compatible, incredible. The throughput of the computer, of course, is very important. It depends on the chip, but it also depends, as you know, the algorithm because without the algorithm, libraries, accelerated computing does nothing. It just sits there.

And so you need to algorithm software libraries. It's a data center scale problem. So networking matters. And networking matters, distributed computing is all about software. Again, system software matters. And before long, in order for you to present your system to your customers, you have to ultimately have a lot of applications that run on top of it. The software ecosystem matter.

Well, the utilization of a data center is one of the most important criteria of its TCO. Just like a hotel, if the hotel is wonderful, but it's mostly empty, the cost is incredible. And so you need the utilization to be high. In order for the utilization to be high, you have to have many different applications. So the richness of the applications matter, again, the algorithm and libraries and now the software ecosystem.

You purchase a computer, but these computers are incredibly hard to deploy. From the moment that you buy the computer to the time that you put that computer to work to start making money, that difference can be weeks, if you're very good at it, incredibly good at it. We can stand up a supercomputer in a matter of a couple of weeks because we built so many all around the world, hundreds around the world. But if you're not very good at it, it could take a year.

That difference, depriving yourself the year of making money and the year of depreciation, incredible cost, life cycle optimization. Because the data center is software defined, there are so many engineers that will continue to refine and continue to optimize the software stack. Because NVIDIA software stack is architecturally compatible across all of our generations, across all of our GPUs. Every time we optimize something, it benefits everybody. Every time we optimize something, it benefits everybody. So life cycle optimization. And of course, finally, the energy that you use, power. But this equation is incredibly complicated.

Well, because we have now addressed so many different domains of science, so many industries and in data processing, in deep learning, classical machine learning, so many different ways for us to deploy software from the cloud to enterprise to supercomputing to the edge, so many different configurations of GPUs, from our HGX versions to our Omniverse versions to our cloud GPU and graphics version, so many different versions, now the utilization is incredibly high. The utilization of NVIDIA GPU is so high almost every single cloud is overextended. Almost every single data center is overextended. There are so many different applications using it.

So we have now reached the tipping point of accelerated computing. We have now reached the tipping point of generative AI. And I want to thank all of you for your support and all of your assistance and partnership in making this dream happen. Thank you.

Every single time we announce a new product, the demand for every single generation increased, increased and increased. And then one generation, it hockey steps: we stick with it, we stick with it, we stick with it, Kepler and then Volta and then Pascal and then Volta and then Ampere. And now this generation of accelerated computing. The demand is literally from every corner of the world. And we are so, so, so excited to be in full volume production of the H100. I want to thank all of you for your support. This is incredible.

H100 is in full production, manufactured by companies all over Taiwan, used in clouds everywhere, enterprises everywhere. And let's take a look at a short video of how H100 is produced.

(presentation)

Jensen Huang

It's incredible. This computer, 35,000 components on that system board, 8 hopper GPUs. I'm going to show it to you. All right. This -- I would lift this, but I still have the rest of the keynote I would like to give. This is 60 pounds, 65 pounds. It takes robots to lift it, of course, and it takes robots to insert it because the insertion pressure is so high and has to be so perfect. This computer is $200,000. And as you know, it replaces an entire room of other computers.

So this, I know it's a very, very expensive computer. It's the world's single most expensive computer that you can say, the more you buy, the more you save. This is what a compute tray looks like. Even this is incredibly heavy. See that? So this is the brand-new H100 with the world's first computer that has a transformer engine in it. The performance is utterly incredible. Hoppers in full production.

We've been driving computing, this new form of computing, for 12 years. When we first met the deep learning researchers, we were fortunate to realize that not only was deep learning going to be a fantastic algorithm for many applications initially, computer vision and speech, but it would also be a whole new way of doing software. This fundamental new way of doing software that can use data to develop, to train a universal function approximator of incredible dimensionality.

It can basically predict almost anything that you have data for so long as the data has structure that it can learn from. And so we realized the importance of this new method of developing software and that it has the potential of completely reinventing computing, And we were right. 12 years later, we have reinvented literally everything. We have reinvented -- of course, we started by creating a new type of library, it's essentially like a SQL except for deep learning, for neural network processing. It's like a rendering engine, a solver for neuro network processing called cuDNN.

We reinvented the GPU. People thought that GPUs would just be GPUs. They were completely wrong. We dedicated ourselves to reinventing the GPU so that it's incredibly good at Tensor processing. We created a new type of packaging called SXM and worked with TSMC on colos so that we could stack multiple chips on the same die. NVLink so that we can connect these SXM modules together with high-speed chip-to-chip interconnect. Almost a decade ago, we built the world's first chip to chip SerDes so that we can expand the memory size of GPUs using SXMs and NVLink. And we create a new type of motherboard, we call it HGX, that I just showed you.

No computers has ever been this heavy before or consumed this much current. Every aspect of a data center had to be reinvented.

We also invented a new type of computer appliance so that we could develop software on it so that third-party developers could develop software on it with a simple appliance we call DGX. Basically a giant GPU computer, DGX. We also purchased Mellanox, which is one of the great strategic decisions of our company because we realized that in the future, if the data center is the computer, then the networking is the nervous system. If the data center is the computer, then the networking defines the data center. That was an incredibly good acquisition. And since then, we've done so many things together, and I'm going to show you some really, really amazing work today.

And then, of course, an operating system, if you have a nervous system, a distributed computer, it needs to have an operating system. And the operating system of this distributed computing, we call Magnum IO, some of our most important work. And then all of the algorithms and engines that sit on top of these computers, we call NVIDIA AI, the only AI operating system in the world that takes data processing from data processing to training, to optimization, to deployment and inference, end-to-end deep learning processing. It is the engine of AI today.

Well, every single generation since Kepler, which is K80, to Pascal, Volta, Ampere, Hopper, every 2 years, every 2 years, we took a giant leap forward. But we realized we needed more than even that, and which is the reason why we connected GPUs to other GPUs called NVLink, build one giant GPU, and we connected those GPUs together using InfiniBand into larger-scale computers. The ability for us to drive the processor and extend the scale of computing, made it possible for the AI research organization, the community, to advance AI at an incredible rate. We just kept pushing and pushing and pushing.

Hopper went into production August of last year, August 2022. 2024, which is next year, we'll have Hopper Next. Last year, we had Quantum. Two years from now or next year, we'll have Quantum Next. So every 2 years, we take giant leaps forward, and I'm expecting the next leap to be giant as well.

This is the new computer industry. Software is no longer programmed just by computer engineers. Software is programmed by computer engineers working with AI supercomputers. These AI supercomputers are a new type of factory. It is very logical that a car industry has factories. They build things that you can see, cars. It is very logical that computer industry has computer factories. You build things that you can see, computers. In the future, every single major company will also have AI factories, and you will build and produce your company's intelligence.

And it's a very sensible thing. We cultivate and develop and nourish our employees and continue to create the conditions by which they can do their best work. We are intelligence producers already. It's just that the intelligence producers, the intelligence are people. In the future, we will be intelligence producers, artificial intelligence producers. And every single company will have factories, and the factories will be built this way. This translates to your throughput. This translates to your scale. And you will build it in a way that is very, very good TCO.

Well, our dedication to pursuing this path and relentlessly increasing the performance, just think, in 10 years' time, we increased the throughput, we increased the scale, the overall throughput, across all of that stack by 1 millionx. 1 millionx in 10 years.

Well, just now, in the beginning, I showed you computer graphics. In 5 years, we improved the computer graphics by 1,000x in 5 years using artificial intelligence and accelerated computing. Using accelerated computing and artificial intelligence, we accelerated computer graphics by 1,000x in 5 years. Moore's Law is probably currently running at about 2x. 1,000x in 5 years. 1,000x in 5 years is 1 millionx in 10. We're doing the same thing in artificial intelligence.

Now question is, what can you do when your computer is 1 million times faster? What would you do if your computer was 1 million times faster? Well, it turns out that the friends we met at University of Toronto, Alex Krizhevsky and Sutskever -- Ilya Sutskever and Geoff Hinton, they -- and Ilya Sutskever, of course, was the founder of OpenAI. And he discovered the continued scaling of artificial intelligence and deep learning works and came up with the ChatGPT breakthrough. Well, in this general form, this is what has happened.

The transformer engine, transformer engine and the ability to use unsupervised learning, unsupervised learning, be able to learn from a giant amount of data and recognize patterns and relationships across a large sequence and using transformers to predict the next word, large language models were created. And the breakthrough, of course, is very, very clear, and I'm sure that everybody here has already tried ChatGPT.

But the important thing is this, we now have a software capability to learn the structure of almost any information. We can learn the structure of text, sound, images, their structure in all of this, physics, proteins, DNA, chemicals. Anything that has structure, we can learn that language, learn its language. Of course, you can learn English and Chinese and Japanese and so on and so forth, but you can also learn the language of many other things.

And then the next breakthrough came, generative AI, once you can learn the language, once you can learn the language of certain information, then with control and guidance from another source of information that we call prompts, we can now guide the AI to generate information of all kinds. We can generate text to text, text to image. But the important thing is this, information transformed to other information is now possible. Text to proteins, text to chemicals, images to 3D, images to 2D -- images to text captioning, video to video. So many different types of information can now be transformed.

For the very first time in history, we have a software technology that is able to understand the representation of information of many modalities. We can now apply computer science. We can now apply the instrument of our industry. We can now apply the instrument of our industry to so many different fields that were impossible before. This is the reason why everybody is so excited.

Now let's take a look at some of these. Let's take a look at what it can do. This here is a prompt and this prompt says hi, Computex. So this is a -- we typed in the word, hi Computex. I'm here to tell you how wonderful stinky tofu is. You can enjoy it right here in Taiwan. It's best from the night market. I was just there the other night.

Okay, play it.

(presentation)

Jensen Huang

The only input was words. The output was that video. And Paul, Tom, was the audio not very loud? It was hard for me to tell from here. Are you guys with me? Hi, am I alone? (foreign language)

Okay. Here's another prompt. Taiwanese, we tell this AI, okay? We tell this AI, yes, this is a Google text to music. Traditional Taiwanese music, peaceful. It gets warm and raining in a lush forest at daybreak. Please.

(presentation)

Jensen Huang

(foreign language) We send text in. AI says, okay, this music. Okay.

Here, this one. I am here at Computex. I will make you like me best. Sing with me. I really like NVIDIA. Okay? So this is the word. These are the words. These are the words. And I say, hey, voice mod. Could you write me a song? These are the words. Okay. Play it.

(presentation)

Jensen Huang

Can you do that? I couldn't do any of that. That is so good. Okay, you know karaoke, right? Taiwanese love karaoke. All right. This one. Okay, are you guys ready? You're going to join me on this one, okay?

Okay. Okay. (foreign language) 1, 2, 3, you've got to watch the dot. You guys know how karaoke works. (foreign language) Get ready. Okay. Here we go. Play it.

(presentation)

Jensen Huang

All right. Very good. Very good. Good job. With that kind of performance, I'm going to hire AI next time. Okay. So obviously, this is a very, very important new capability, and that's the reason why there's so many generative AI start-ups. We're working with some 1,600 generative AI start-ups. They're in all kinds of domain, in language domains, in media and biology. This is one of the most important areas that we care about.

Digital biology is going to go through its revolution. This is going to be an incredible thing. Just as we had Synopsys and Cadence help us create tools so that we can build wonderful chips and systems, for the very first time, we're going to have computer-aided drug discovery tools, and they'll be able to manipulate and work with proteins and chemicals and understands disease targets and try all kinds of chemicals that previously have never been thought of before, okay?

So really, really important area. Lots of start-ups, tools and platform companies. And let me show you a video of some of the work that they're doing. Play it, please.

(presentation)

Jensen Huang

Incredible, right? Just utterly incredible. There's no question that we're in a new computing era. There's just absolutely no question about it. Every single computing era, you could do different things that weren't possible before. And artificial intelligence certainly qualifies.

This particular computing era is special in several ways. One, it is able to understand information of more than just text to numbers. It can now understand multi-modality, which is the reason why this computing revolution can impact every industry, every industry.

Two, because this computer doesn't care how you program it, it will try to understand what you mean because it has this incredible large language model capability. And so the programming barrier is incredibly low. We have closed the digital divide. Everyone is a programmer now. You just have to say something to the computer.

Third, this computer, not only is it able to do amazing things for the future, it can do amazing things for every single application of the previous era, which is the reason why all of these APIs are being connected into Windows applications here and there and browsers and PowerPoint and Word. Every application that exists will be better because of AI.

You don't have to just AI this generation. This computing era does not need new applications, it can succeed with old applications. And it's going to have new applications. The rate of progress, the rate of progress, because it's so easy to use, is the reason why it's growing so fast. This is going to touch literally every single industry. And at the core, with -- just as with every single computing era, it needs a new computing approach. In this particular era, the computing approach is accelerated computing, and it has been completely reinvented from the ground up.

The last several years, I've been talking to you about the new type of processor we've been creating, and this is the reason we've been creating it. Ladies and gentlemen, Grace Hopper is now in full production. This is Grace Hopper.

Nearly 200 billion transistors in this computer (foreign language). Look at this. This is Grace Hopper. This processor is really quite amazing. There are several characteristics about it. This is the world's first accelerated processor, accelerated computing processor that also has a giant memory. It has almost 600 gigabytes of memory that's coherent between the CPU and the GPU. And so the GPU can reference the memory, the CPU can reference the memory and unnecessary -- any unnecessary copying back and forth could be avoided. The amazing amount of high-speed memory lets the GPU work on very, very large data sets. This is a computer. This is not a chip. Practically the entire computer is on here. This uses low-power DDR memory just like your cell phone, except this has been optimized and designed for high resilience data center applications. Incredible levels of performance. This took us several year to develop, and I'm so excited about it. And I'll show you some of the things that we're going to do with it. Janine, thank you.

Okay. So 4 petaflops transformer engine, 72 CPU cores, they're connected together by high-speed chip-to-chip link, 900 gigabytes per second. And so the local memory, 96 gigabytes of HBM3 memory, is augmented by LP DDR memory across this link, across a very, very large and high-speed cache. So this computer is like none other the world's ever seen.

Now let me show you some of its performance. So I'm comparing here on 3 different applications. And this is a very important application. If you've never heard of it, be sure to look into it, it's called Vector database. Vector database is a database that has tokenized, that has vectorized the data that you're trying to store. And so it understands the relationships of all of the data inside its storage. This is incredibly important for knowledge augmentation of the large language models to avoid hallucination.

The second is deep learning recommender systems. This is how we get news and music and all the texts that you see on your devices, recommend, of course, music and goods and all kinds of things. Recommender system is the engine of the digital economy. This is probably the single most valuable piece of software that any of the companies in the world runs. This is the world's first AI factory. There will be other AI factories in the future, but this is really the first one.

And then the last one is large language model inference, 65 gigabytes. 65 gigabytes is a fair -- 65 billion parameters is a fairly large language model. And you can see that on a CPU, it's just not possible. The CPU is simply not possible.

With Grace Hopper -- excuse me. With Hopper on an x86, it's faster. But notice it's memory limited. You could, of course, take this 400 gigabytes and cut it up into a whole bunch of small pieces, shard it and distribute it across more GPUs and distribute it across more GPUs. But in the case of Grace Hopper, Grace Hopper has more memory on this 1 module than all of these. Does that make sense?

And so as a result, you don't have to break the data into so many pieces. Of course, the amount of computation of this is higher, but this is so much easier to use. And if you want to scale out large language models, if you want to scale out vector databases, if you want to scale out deep learning recommender systems, this is the way to do it. This is so easy to use, plug this into your data center, and you can scale out AI, okay? So this is the reason why we built Grace Hopper.

The other application that I'm super excited about is the foundation of our own company. NVIDIA is a big customer of Cadence. We use all of their tools, and all of their tools run on CPUs. And the reason why they run on CPUs is because NVIDIA's data sets are very large. And the algorithms are refined over very long periods of time. And so most of the algorithms are very CPU-centric.

We've been accelerating some of these algorithms with Cadence for some time. But now with Grace Hopper and we've only been working on it for a couple of days and weeks. The performance speed up, I can't wait to show it to you, is insane. This is going to revolutionize an entire industry. One of the highest compute-intensive industries in the world, of course, designing chips, designing electronic systems, CAE, CAD, EDA and of course, digital biology. All of these markets, all of these industries require very large amounts of computation, but the data set is also very large. Ideal for Grace Hopper.

Well, 600 gigabytes is a lot, 600 gigabytes is a lot. This is basically a supercomputer I'm holding in my hands. This 600 gigabytes is a lot. But when you think about it, when we went from AlexNet of 62 million parameters 12 years ago and trained on 1.2 million images, it is now 5,000x bigger with Google's PaLM. 5,000x bigger with 340 billion parameters. And of course, we're going to make even bigger ones than that. And that's been trained on 3 million times more data. So literally, in the course of 10 years, the computing problem of deep learning increased by 5,000x for the software and 3 million times for the data set. No other area of computing has ever increased this fast. And so we've been chasing the deep learning advance for quite some time. This is going to make a big, big contribution. However, 600 gigabytes is still not enough. We need a lot more. So let me show you what we're going to do.

So the first thing is, of course, we have the Grace Hopper Superchip, put that into a computer. The second thing that we're going to do is want to connect 8 of these together using NVLINK. This is an NVLINK switch. So 8 of this connect into 3 switch trays into 8 Grace Hopper pod. These 8 Grace Hopper pods, each 1 of the Grace Hoppers are connected to the other Grace Hopper at 900 gigabytes per second. 600 gigabytes, 900 megabytes per second, 8 of have them connected together as a pod. And then we connect 32 of them together with another layer of switches. And in order to build in order to this, 256 Grace Hopper superchips connected into 1 ExaFLOPS, 1 ExaFLOPS. You know that countries and nations have been working on ExaFLOPS computing, and just recently achieved it. 256 Grace Hoppers for deep learning is 1 ExaFLOPS transformer engine, and it gives us 144 terabytes of memory that every GPU can see. This is not 144 terabytes distributed. This is 144 terabytes connected.

Why don't we take a look at what it really looks like? Play it, please.

(presentation)

Jensen Huang

This is 150 miles of cables, fiberoptic cables, 2,000 fans, 70,000 cubic feet per minute. It probably recycles the air in this entire room in a couple of minutes. 40,000 pounds, 4 elephants, 1 GPU.

If I can get up on here, this is actual size. I wonder if this can play Crysis. Only gamers know that joke. So this is this is our brand-new Grace Hopper AI supercomputer. It is 1 giant GPU. Utterly incredible. We're building in now all of the -- every component is in production. And we're so excited that Google Cloud, Meta and Microsoft will be the first companies in the world to have access, and they will be doing exploratory research on the pioneering front, the boundaries of artificial intelligence, with us. We will, of course, build these systems as products. And so if you would like to have an AI supercomputer, we would, of course, come and install it in your company. We also share the blueprints of the supercomputer with all of our cloud suppliers, so that our cloud partners so that they can integrate it into their networks and into their infrastructure. And we will also build it inside our company for us to do research ourselves and do development. So this is the DGX GH200. It is 1 giant GPU. Okay.

1964, the year after I was born, was a very good year for technology. IBM, of course, launched the System/360, and AT&T demonstrated to the world the first picture phone and coated, compressed, streamed over copper telephone wires and twisted pair and on the other end, decoded. Picture phone, little tiny screen, black and white. To this day, this very experience is largely the same, of course, at much, much higher volumes. For all of the reasons we all know well, video calls is now one of the most important things we do. Everybody does it. About 65% of the Internet's traffic is now video. And yet the way it's done is fundamentally still the same. Compress it on the device, stream it and decompress it on the other end. Nothing changed in 60 years.

We treat communications like it goes down to a dumb pipe. The question is, what would happen if we applied generative AI to that? We have now created a computer. I showed you, Grace Hopper. It can be deployed broadly all over the world easily. And as a result, every data center, every server will have generative AI capability. What would happen if instead of decompression, streaming and recovering compression, decompression, what if the cloud performed generative AI capability to it? Let's take a look.

(presentation)

Jensen Huang

(foreign language). Okay. So all of the words coming out of my mouth, of course, was generated by AI. So instead of compression, stream and decompression, in the future, communications will be perceive, stream and reconstruction, regeneration. And it can be generated in all kinds of different ways. It can be generated in 3D of course. It can regenerate your language in another language. So we now have a universal translator. This computing technology could be, of course, placed into every single cloud. But the thing that's really amazing, Grace Hopper is so fast, it can even run the 5G stack. A state-of-the-art 5G stack could just run in software in Grace Hopper, completely free, completely free. All of a sudden, a 5G radio runs in software, just like a video codec used to run in software. Now you can run a 5G stack in software. Of course, the layer 1 PHY layer, the layer 2 MAC layer, and the 5G core, all of that computation is quite intensive, and it has to be timing-precise, which is the reason why we have a BlueField-3 in the computer. So that kind of precision timing networking. But the entire stack can now run in a Grace Hopper.

Basically, what's happening here, this computer that you're seeing here allows us to bring generative AI into every single data center in the world today. Because we have software-defined 5G, then the telecommunication network can also become a computing platform like the cloud data centers. Every single data center in the future could be intelligent, every data center could be software-defined whether it's Internet-based, networking-based or 5G communications-based. Everything will be software defined. This is really a great opportunity. And we're announcing a partnership with SoftBank to partner to re-architect and implement generative AI and software-defined 5G stack into the network of SoftBank data centers around the world. Really excited about this collaboration.

I just talked about how we are going to extend the frontier of AI. I talked about how we're going to scale out generative AI, to scale out generative AI to advance generative AI. But the number of computers in the world is really quite magnificent, data centers all over the world. And all of them over the next decade will be recycled and reengineered into accelerated data centers and generative AI-capable data centers. But there are so many different applications in so many different areas. Scientific computing, data processing, large language model training that you've been -- we've been talking about generative AI inference that we just talked about. Cloud and video and graphics. EDAs, SDA, which we just mentioned, generative AI for enterprise and of course, the edge. Each one of these applications have different configurations of servers, different focus of applications, different deployment methods. And so security is different. Operating system is different, how it's managed is different, where the computers are will be different. And so each one of these diverse application spaces will have to be reengineered with a new type of computer.

Well, this is just an enormous number of configurations. And so today, we're announcing in partnership with so many companies here in Taiwan, the NVIDIA MGX. It's an open modular server design specification and is designed for accelerated computing. Most of the servers today are designed for general purpose computing. The mechanical, thermal and electrical is insufficient for a very highly dense computing system. Accelerated computers take, as you know, many servers and compress it into 1. You save a lot of money, you save a lot of floor space, but the architecture is different. And we designed it so that it's multi-generation standardized so that once you make an investment, our next-generation GPUs and next generation CPUs and next-generation DPUs will continue to easily configure into it so that we kind of best time to market and best preservation of our investment. We could -- configurable into hundreds of configurations for different diversities and different diverse applications and integrate into cloud or enterprise data center. So you could have either busbar or power regulators, you could have cabling in the hot aisle or cabling in the cold aisle. Different data centers have different requirements, and we've made this modular and flexible so that it could address all of these different domains.

Now this is the basic chassis. Let's take a look at some of the other things you could do with it. This is the Omniverse OVX server, that has x86, 4 L40s, BlueField-3, 2 CX7s, 6 PCI Express slots. This is the Grace Omniverse server. Grace, same, 4 L40s, BF3, BlueField-3 and 2 CX7s, okay? This is the Grace cloud graphic server. This is the Hopper NVLINK generative AI inference server, and we need sound effects like, like that. And then Grace Hopper 5G Aerial server, okay, for telecommunications, software-defined telco. Grace Hopper 5G Aerial Server Short. And of course, Grace Hopper

Liquid Cooled, okay, for very dense servers. And then this 1 is our dense general-purpose Grace superchip server. This is just CPU and has the ability to accommodate 4 CPU -- 4 Grace CPUs or 2 Grace superchips, enormous amounts of performance.

And if you were -- if you're a data center is powered limited, this has incredible capabilities. In a power limited environment, running PageRank, and there's all kinds of benchmarks you can run, but we ran PageRank in iSO performance, in iSO performance, Grace only consumes 580 watts for the whole server versus the latest generation CPU servers, x86 servers, 1,090 watts. It's basically half the power at the same performance or another way of saying, at the same power, if your data center is power constrained, you get twice the performance. Most data centers today are Power Limited. And this is really a terrific capability. There are all kinds of different servers that are being made here in Taiwan. Let me show you one of them. This is from Supermicro, get my exercise in today. I am the sound effect, okay. All right so this is a Supermicro server, you've got BlueField-3, got the CX7, you got the Grace Hopper. Okay. But there are so many systems. There are so many systems. So this is Grace Hopper, that's Supermicro. There are so many systems. Let me show you some of them.

And all of our partners, I'm so grateful. You're working on Grace, Grace Hopper, Hopper, L40s, L4s, BlueField-3s, just about every single processor that we're building are configured into the servers of all different types. And so this is Supermicro. This is Gigabyte. The 10s, I think it's like 70 different server configurations. This is Ingrasys, this is ASRock, Tyan, Wistron, Inventec. It's beautiful servers, PEGATRON. We love servers. I love servers. They're beautiful. They're beautiful to me. QCT, ASUS, Wiwynn, ZT Systems. And this ZT system, what you're looking at here is one of the pods of our Grace Hopper AI supercomputer. So I want to thank all of you. I want to thank all of you for your great support. Thank you.

We're going to expand AI into a new territory. If you look at the world's data centers, the data center is now the computer, and the network defines what that data center does. Largely, there are 2 types of data centers today. There's the data center that's used for hyperscale where you have application workloads of all different kinds. The number of CPUs, the number of GPUs you connect to it is relatively low. The number of tenants is very high. The workload is heterogeneous. The workloads are loosely coupled and you have another type of data center. They're like supercomputing data centers, AI supercomputers, where the workloads are tightly coupled. The number of tenants far fewer and sometimes just 1. Its purpose is high throughput on very large computing problems, okay? And it's basically a stand-alone. It's basically a stand-alone. And so supercomputing centers and AI supercomputers and the world's cloud, hyperscale cloud are very different in nature.

Ethernet is based on TCP. It's a lossy algorithm, and it's very resilient. And whenever there's a loss, packet loss, it retransmits. There's error correction that's done. It knows which one of the packets are lost and request the center to retransmit it. The ability for Ethernet to interconnect components of almost from anywhere is the reason why the world's Internet was created. If it required too much coordination, how could we have built today's Internet? So Ethernet's profound contribution, it's this lossy capability, its resilient capability. And because so it basically can connect almost anything together. However, a supercomputing data center can't afford that, you can't interconnect random things together because that $1 billion supercomputer, the difference between 95% networking throughput achieved versus 50% is effectively $500 million. So the cost of that 1 workload running across that entire supercomputer is so expensive that you can't afford to lose anything in the network.

InfiniBand is our -- relies on RDMA very heavily. It is a flow control so it's a lossless approach. It requires flow control, which basically means you have to understand the data center from end to end, to switch to the NIC to the software so that you can orchestrate the traffic with adaptive routing so that you could deal with congestion control and avoid the oversaturation of traffic in an isolated area, which results in packet loss. You simply can't afford that because in the case of InfiniBand, it's lossless. And so one is lossy, the other one is lossless, very resilient, very performant. These 2 data centers have lived separate lives. These 2 data centers have lived separate lives. But now we would like to bring generative AI to every data center. The question is how. The question is, how do we introduce a new type of Ethernet that's, of course, backwards compatible with everything but it's engineered in a way that achieves the type of capabilities that we can bring AI workloads to the world's any data center.

This is a really exciting journey. And at the core of this strategy is a brand-new switch that we've made. This is the Spectrum-4 switch. And this switch -- everything I'm showing today are very heavy. This is the Spectrum-4 switch, 128 ports, 400 gigabits per second. 128 ports of 400 gigabits per second, 51.2 terabits per second. This is the chip. It's gigantic, 100 billion transistors, 90 millimeters by 90 millimeters, 800 balls on the bottom. This is a 500-watt chip. This switch is 2,800 watts. It's air cooled. There are 48 PCBs that connect the switch together, 48 PCBs that build up the switch. And the switch is designed to enable a new type of Ethernet. Remember what I said, InfiniBand is fundamentally different in the sense that we build InfiniBand from end to end so that we could do adaptive routing so that we could do congestion control so that we can isolate performance, and we could keep noisy neighbors apart so that we could do in-fabric computing. All of these capabilities are simply not possible in a lossless

approach of the Internet, and of course, of Ethernet. And so the way that we do InfiniBand is design it from end to end, just the way supercomputers are built, this is the way AI supercomputers are built. And we are going to do the same thing now for the very first time for Ethernet. We've been waiting for the critical part and the critical part is the Spectrum-4 switch.

The entire system consists of several things. So our new Ethernet system for AI is this. The Spectrum-4 switch and the BlueField-3 Smart NIC or DPU. This BlueField-3 is 400 gigabits per second NIC. It connects directly to the Spectrum-4 switch in combination of 4 things: the switch, the BlueField-3, the cablings that connect them together, which are super important and the software that runs it all together represents the Spectrum-4. This is what it takes to build a high-performance network, and we're going to take this capability to the world's CSPs. The reception has been incredible. And the reason for that is, of course, every CSP, every data center would like to turn every single data center into a generative AI data center. There are some people that need -- they deployed Ethernet throughout their company. And they have a lot of users for that data center. The ability to have the capabilities of InfiniBand and isolating it within their data center is very difficult to do. And so for the very first time, we're bringing the capabilities of high-performance computing into the Ethernet market.

And we're going to bring to the Ethernet market several things. First, adaptive routing. Adaptive routing basically says based on the traffic that is going through your data center, depending on which one of the parts of that switch is overcongested, it will tell BlueField-3 to -- and will send it to another port. BlueField-3 on the other end, would reassemble it and present the data to the CPU -- present the data to the computer to the GPU without any CPU intervention. All completely in RDMA. Number one, adaptive routing. Second, congestion control. Congestion control, it is possible for certain different ports to become heavily congested, in which case the telemetry of the switch. Each switch will see how the network is performing and communicate to the senders, please don't send any more data right away because you're congesting the network. That congestion control requires basically an overriding system, which includes software, the switch, working with all of the end points to overall manage the congestion or the traffic and the throughput of the data center.

Now it's really important to realize that in a high-performance computing application, every single GPU must finish their job so that the application can move on. In many cases where you do all reductions, you have to wait until the results of every single 1. So if 1 node takes too long, everybody gets held back. This capability is going to increase Ethernet's overall performance dramatically. So Spectrum-X, really excited to roll this out. The world's applications, the world's enterprise has yet to enjoy generative AI. So far, we've been working with CSPs and the CSPs, of course, is going to be able to bring generative AI to many different regions and many different applications and industries. The big journey is still ahead of us. There are so many enterprises in the world, and everybody because of the multi-modality capability that I was mentioning before, every industry can now benefit from generative AI.

There are several things that we have to do. Number one, we have to help the industries build custom language models. Not everybody can use the language models that are available in a public service. Some customers need language models that are highly specialized for their particular modality, for example, proteins or chemicals. Each one of these industries have proprietary information, and so how can we help them do that? We have created a service called NVIDIA AI Foundation. It is a cloud service that captures NVIDIA's AI expertise and makes it possible for you to train your own AI models. We will help you develop your own AI models with supervised fine tuning, with guard railing, with proprietary knowledge bases and reinforcement learning human feedback so that this AI model is perfect for your application. We then deploy this model to run on NVIDIA AI enterprise. This is the operating system that I was talking to you about earlier. This operating system runs in every single cloud. This allows -- this very simple system with NVIDIA AI foundation for training large language models and deploying the language model into NVIDIA AI enterprise, which is available in every single model, every single cloud and on-prem. Allows every single enterprise to be able to engage.

Now one of the things that very few people realize is that today, there's only 1 software stack that is enterprise secure and enterprise grade. That software stack is CPU. And the reason for that is because in order to be enterprise grade, it has to be enterprise secure and has to be enterprise managed and enterprise supportive across its entire life, across its life cycle. There are so much software in accelerated computing. 4 -- over 4,000 software packages is what it takes for people to use accelerated computing today in data processing and training and optimization all the way to inference. So for the very first time, we are taking all of that software, and we're going to maintain it and manage it like Red Hat does for Linux. NVIDIA AI enterprise will do it for all of NVIDIA's libraries. Now enterprise can finally have an enterprise-grade and enterprise-secure software stack. This is such a big deal. Otherwise, even though the promise of accelerated computing is possible for many researchers and scientists is not available for enterprise companies.

And so let's take a look at the benefit for them. This is a simple image processing application. If you were to do it on a CPU versus on a GPU, running on enterprise NVIDIA AI enterprise, you're getting 31.8 images per minute or basically 24x the throughput, or you only pay 5% of the cost. This is really quite amazing. This is the benefit of accelerated computing in the cloud. But for many companies, enterprises is simply not possible unless you have the stack. NVIDIA AI Enterprise is now fully integrated into AWS, Google Cloud and Microsoft Azure and Oracle Cloud. And so when you go and deploy your workloads in those clouds and you want software that is enterprise grade or you have customers that need enterprise-grade software, NVIDIA AI Enterprise is ready for you. It is also integrated into the world's machine learning operations pipeline. As I mentioned before, AI is a different type of workload, and this new type of software, this new type of software has a whole new software industry. And this software industry, 100% of them, we have now connected with NVIDIA AI Enterprise.

Now let me talk to you about the next phase of where AI meets the digital twin. Now why does AI need a digital twin? I'm going to explain that in a second. But first, let me show you what you can do with it. In order for AI to have a digital twin, in order for AI to understand heavy industry, remember, so far, AI has only been used for light industry, information, words, images, music, so on and so forth. If we want to use AI for heavy industry, the $50 trillion of manufacturing, many of that you're part of. The trillions of dollars of health care. All of the different manufacturing sites, whether you're building chip fabs or battery plants or electric vehicle manufacturing factories, all of these would have to be digitized in order for artificial intelligence to be used to automate, to design, to build and to automate the future of your business. And so the first thing that we have to do is we have to create the ability for their world to be represented in digital, okay? So number 1 is digitalization.

Well, why does it -- how would you use that? So let me give you just a simple example. In the future, you would say to your robot, I would like you to do something, and the robot will understand your words, and it would generate animation. Remember, I said earlier, you can go from text to text. You can go from text to image. You can go from text to music. Why can't you go from text to animation? And so of course, in the future, robotics will be highly revolutionized by the technology we already have in front of us. However, how does this robot know that the motion that it is generating is grounded in reality? It is grounded in physics. You need a software system that understands the laws of physics. Now you've actually seen this already with ChatGPT. Whereas AI, NVIDIA AI would use NVIDIA Omniverse in a reinforcement learning loop to ground itself, you have seen ChatGPT do this using reinforcement learning, human feedback. Using human's feedback, ChatGPT was able to be developed by grounding it to human's, well, sensibility and align it with our principles. So reinforcement learning with human feedback is really important. Reinforcement learning for physics feedback is very important.

Let me show you. Everything that you're about to see as a simulation. Let's roll it, please.

(presentation)

Jensen Huang

Everything was a simulation, nothing was art. Everything was simulation. Isn't that amazing? In the last 25 years, I come to Taiwan, you sell me things. Omniverse will be the first thing I'm going to sell you. And this is -- because this will help you revolutionize your business and turn it into a digital business and automate it with AI. You will build products in digital first before you make it physical. You will build factories and plan it in digital first before you make it in physical. And so in the future, Omniverse is a big deal.

Now I'm going to show you very quickly Omniverse in the cloud. Omniverse, the entire stack, is so complicated. And so we put the whole thing into a cloud managed service, and it's hosted in Azure. This particular experience you're going to have, the computer is in California. And [Sean]? I'm sorry, I took so much time, so you're going to have to...

Unidentified Company Representative

I'll go quick. So this is -- let's take a look at the Omniverse Cloud. So this is just a web browser, and we're looking now into Omnibus factory Explorer. It's running 10,000 kilometers away in our Santa Clara headquarters and we're leveraging the power of our data center now to visualize this factory floor. We're using real factory data from Siemens and Autodesk Revit to take a look. It's a cloud application, so we could have multiple users collaborating. Let's go ahead and bring up LOE screen. And we can see that we have these 2 users in this environment. And [Jeff], on the left there

is going to look at some markup. We have this task to perform. We need to move this object. So we can have [Louis], just go ahead and grab that conveyor belt, move it over. And as he does so, you'll see that it's reflected accurately and completely in real time on [Jeff's] screen.

So we're able to collaborate with multiple users. And even in bringing up this demo, we had users from around the globe working on the process. East and West Coast, United States, Germany, even Sydney and course here in Taipei, to put this together. Now as we -- if we're modifying our production line, of course, one of the things we want to do is add the necessary safety equipment. So we're able to simply drag and drop items into Omniverse and modify this production environment and begin tweaking this and optimizing for performance even before we break around with construction.

Jensen Huang

That was so cool. This is in California, 6,264 miles away or something like that. 34 milliseconds by speed of light 1 way. And it's completely interactive. Everything is ray traced. No artist necessary, you bring everything, the entire CAD into Omniverse. Open up a browser, bring your data in, bring your factory in, no artist necessary. The lighting just does what the lighting does, physics does what the physics does. If you want to turn off physics, you can. If you want to turn on physics, you can. And multiple users, as many as you like, can enter the Omniverse at the same time and work together. One unified source of data across your entire company. You could virtually build, you can virtually design and build, and operate your factory before you break ground and not make the mistake, which usually in the beginning of the integration creates a lot of change orders, which cost a lot of money. Thank you very much, [Sean]. Good job.

Not only -- noticed just now, it was humans interacting with Omniverse. Humans interacting with Omniverse. In the future, Sean will even have a generative AI, and AI interact with him in Omniverse. We could, of course, imagine in the very beginning, there was [Jin]. That could be a character. That could be one of the users of Omniverse, interacting with you, answering questions, helping you. We can also use generative AI to help us create virtual worlds. So for example, this is a bottle that's rendered in Omniverse. It could be placed in a whole bunch of different type of environments, it can render beautifully physically. You can place it just by giving it a prompt by saying, I would like to put this life -- these bottles in a lifestyle photograph style backdrop for a modern, warm farmhouse bathroom, change the background, everything is all integrated and rendered again, okay? So generative AI will come together with Omniverse to assist the virtual -- the creation of virtual worlds.

Today, we're announcing that WPP, the world's largest advertising agency and advertising services company is partnering with NVIDIA to build a content generation engine based on Omniverse and generative AI. It integrates tools from so many different other partners, Adobe Firefly, for example, Getty, Shutterstock. And it integrates into this entire environment. And it makes it possible for them to generate unique content for different users for ad applications, for example. So in the future, whenever you engage in particular ad, it could be generated just for you. But yet the product is precisely rendered because, of course, the product integrity is very important. And so every time that you engage a particular ad in the future, today, it was retrieved. And today, the computing model, when you engage information, it is retrieved in the future. When you engage information, much of it will be generated. Notice the computing model has changed. WPP generates 25% of the ads that the world sees. 60% of the world's largest companies are already clients. And so they made a video of how they would use this technology.

(presentation)

Jensen Huang

No problem. We continue. Okay. So that's WPP. You could see that, that was an example -- if you think about it a second, that's an example of a company using digital information that was created in design and using that digital information all the way in marketing. I'm going to show you now how we're going to use Omniverse and AI here in Taiwan, and we're going to use it for manufacturing. Manufacturing as you know, is one of the largest industries in the world. We're going to use Omniverse to teach an AI. And then we're going to use Metropolis, our AI deployment, edge deployment system, to deploy the AI. Okay, run it.

(presentation)

Jensen Huang

Did you see that the whole factory is an Omniverse? It's completely digital. Imagine if you have digital information in your hands. What can you do with it? Almost everything. And so this is one of the things that's really exciting. What you just saw is basically every factory in the future will be digital, of course, first. Every factory will be a robot. Inside the factories, there will be other robots. The factory is orchestrating. We are also going to build robots that move themselves. So far, the robots that you saw are stationary. Now we're going to also have robots that move. Everything that move in the future will have artificial intelligence and will have robotic capability.

And so today, we're announcing our robot platform, NVIDIA Isaac AMR is now available as a reference design for anybody who wants to build robots, just like we did with our high-performance computing. NVIDIA builds the whole stack. And then we disaggregate it so that if you would like to buy the chip, that's fine, if you like to buy the system, that's fine, if you like to use your software that's fine. If you like to use your software, that's fine. If you like to use your own algorithm, that's terrific. If you'd like to ours, that's terrific. However you would like to work with us, we're open for business so that we can help you integrate accelerated computing wherever you like.

In the future, we're going to do the same with robotics. We built the entire robotic stack top to bottom from the chip, to the algorithms, we have state-of-the-art perception for multi-modality sensor, state-of-the-art mapping, state-of-the-art localization and planning and a cloud mapping system, everything has been created. However you would like to use it, you can use pieces of it. It's open available for you, including the cloud mapping systems. So this is Isaac AMR. It includes starts with the chip called Orin. It goes into a computer, and it goes into the NVIDIA Orin -- Nova Orin which is a reference system, a blueprint for AMRs. This is the most advanced AMR in the world today. And that entire stack has been built and let's take a look at it.

(presentation)

Jensen Huang

Nova cannot tell that it is not in the real environment. Nova thinks it is in the real environment. It cannot tell. And the reason for that is because all the sensors work, physics work, it can navigate, it can localize itself. Everything is physically based. So therefore, we could design the robot, simulate the robot, train the robot, all in Isaac. And then we take the brain, Isaac Sim, then we take the brain, the software and we put it into the actual robot. And with some amount of adaptation, it should be able to perform the same job. This is the future robotics, Omniverse and AI working together.

The ecosystem that we have been in the IT ecosystem is $0.25 trillion per year. $250 billion a year. This is the IT industry. For the very first time in our history together, we finally have the ability to understand the language of the physical world. We can understand the language of heavy industry. And we have a software tool, we have a software system called Omniverse that allows us to simulate, to develop, to build and operate our physical plants, our physical robots, our physical assets, as if they were digitally. The excitement in the hard industries, the heavy industries has been incredible. We have been connecting Omniverse all over the world with tools companies, robotics company, sensor companies, all kinds of industries.

There are 3 industries right now as we speak, that is putting enormous investments into the world. Number one, of course, is chip industry; number two, electric battery industry; number three, electric vehicle industry. Trillions of dollars will be invested in the next several years, trillions of dollars will be invested in the next several years. And they would all like to do it better, and they would like to do it in a modern way. For the very first time, we now give them a system, a platform, tools that allows them to do that. I want to thank all of you for coming today. I talked about many things. It's been a long time since I've seen you, so I had so much to tell you. There was too much. It was too much. Last night, I said, this is too much. This morning I said, this is too much. And now I realize, it's too much. (foreign language).

I told you several things. I told you that we are going through 2 simultaneous computing industry transition, accelerated computing and generative AI. Two, this form of computing is not like the traditional general-purpose computing. It is full stack. It is data center scale because the data center is the computer, and it is domain specific. For every domain that you want to go into, every industry you go into, you need to have the software stack. And if you have the software stack, then the utility, the utilization of your machine, the utilization of your computer will be high.

So number two, it is full stack data center scale and domain specific. We are in full production of the engine of generative AI, and that is HGX H100. Meanwhile, this engine that's going to be used for AI factories will be scaled out using Grace Hopper, the engine that we created for the era of generative AI. We also took Grace Hopper and realized that we can extend on the one hand, the performance, but we also had to extend the fabric so that we can make larger models trainable. And we took Grace Hopper connected to 256 node NVLINK and created the largest GPU in the world DGX GH200.

We're trying to extend generative AI and accelerated computing in several different directions at the same time. Number one, we would like to, of course, extend it in the cloud so that every cloud data center can be an AI data center. Not just AI factories and hyperscale, but every hyperscale data center can now be a generative AI data center. And the way we do that is the Spectrum-X. It takes 4 components to make Spectrum-X possible. The switch, the BlueField-3 NIC, the interconnects themselves. The cables are so important in high-speed communications, and the software stack that goes on top of it. We would like to extend generative AI to the world's enterprise. And there are so many different configurations of servers and the way we're doing that with partnership with our Taiwanese ecosystem, the MGX modular accelerated computing systems.

We put NVIDIA in the cloud so that every enterprise in the world can engage us to create generative AI models, and deploy it in a secure way, in an enterprise-grade, enterprise secure way in every single cloud. And lastly, we would like to extend AI to the world's heavy industries, the largest industries in the world. So far, our industry, our industry, that all of us have been part of, has been a small part of the world's total industry. For the very first time, the work that we're doing can engage every single industry and do that by automating factories, automating robots, and today, we even announced our first robotics full reference stack, the Nova Orin.

I want to thank all of you for your partnership over the years. Thank you.