NVIDIA Keynote at SIGGRAPH 2023

Company Participants

- Jensen Huang, Founder and Chief Executive Officer
- Unidentified Speaker, Unknown

Presentation

Unidentified Speaker

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

Jensen Huang {BIO 1782546 <GO>}

Twenty years, 20 years after we introduced to the world, the first programmable shading GPU, we introduced RTX at SIGGRAPH 2018 and reinvented computer graphics. You didn't know it at the time, but we did that it was a bet the company moment. The vision of RTX was to bring forward real time ray tracing something that was of course used in film rendering offline. It required that we invent, reinvent the GPU added ray tracing accelerators, reinvented the software of rendering, reinvented all the algorithms that we made four rasterization programmable shading. And that wasn't even enough. We had to bring together computer graphics and artificial intelligence for the very first time to make it possible.

In 2018, five years ago, this was the showcase demo. The first RTX GPU was called touring. As you can imagine, we did it on purpose. It was appropriately named to unify computer graphics and artificial intelligence for the very first time. This demo was called Star Wars: Reflections. It was created by the researchers at ILMxLABs, Epic, and Nvidia. It had two and a half million polygons or so, two rays per pixel, a couple of bounces per ray. We did ambient occlusion, area lights, specular reflections, it was a hybrid rasterization and ray-trace demo. We rendered it at 720p, 30 frames a second. And we use DLSS Super Resolution to scale it to 4K. The demonstration was frankly at the time, incredibly beautiful. That was five years ago.

Now five years later, Racer RTX 250 million polygons 100 times more geometry 10 rays per pixel, about 10 bounces per ray. We're using a unified lighting system for every effect, for the very first time. This entire scene is completely pap traced, no rasterization we're rendering it at 10 ADP 30 hertz, and using DLSS, using artificial intelligence infer something like one out of -- infer like seven out of eight pixels. Computing only one out of eight and as a result, we're able to render this at 4K, scale it up to 4K 6 -- 4K 30 hertz. Hit it.

Not bad for real time. Think it's safe to say that it was worth it to bet the company. We realized that rationalization was reaching his limits. And unless we did took such a

giant risk again, an introduced a brand-new way of doing computer graphics, combining CG and AI for the very first time. What you just saw would not be possible. Modern computer graphics has been reinvented. The bet has paid off.

While we were reinventing computer graphics with artificial intelligence, we're really reinventing the GPU altogether for artificial intelligence. The GPU, when I came to see you last time, five years ago, most people would say that this is what a GPU looks like. And in fact, this is the GPU that we announced this is touring. And this you guys might remember this, this is the Turing GPU. But this is what a GPU is today. This GPU is, I guess, let's see eight hoppers. Each one of them all together, something like between the hoppers, the eight hoppers connected with NVLink, the InfiniBand networking, the NVSwitches that are connecting them together, the NVLink switches, all together 1 trillion transistors.

This GPU has 35,000 parts. It's manufactured by a robot like an electric car, it weighs 70 pounds consumes 6000 watts. And this GPU revolutionized computer science all together. This is the third generation. This is Hopper GPU. This is the GPU that everybody writes music. There's a Billy Joel song written about this GPU. And so this GPU has gone on to reinvent artificial intelligence. And 12 years later, after 12 years working on artificial intelligence, something gigantic happen. The generative AI era is upon us. The iPhone moment of AI, if you will, where all of the technologies of artificial intelligence came together in such a way that is now possible for us to enjoy AI in so many different applications.

The revolutionary transformer model allows us to learn from a large amount of data that's across large spans of space and time to find patterns and relationships, to learn the representation of almost anything with structure, we learned a representation, how to represent language in mathematics and vectors and vector space, audio, animation, 3D video, DNA, proteins, chemicals. And with a generative model, and the learn language model, you can guide the auto regressive diffusion models to generate almost anything you like. And so we could learn the representation of almost anything with structure. We can generate almost anything that we can learn from structure, and we can guide it with our human natural language.

The journey of Nvidia accelerated computing met the journey of the deep learning researchers. And the big bang of modern AI happened. This is now 12 years later, the 12-year journey of our work in artificial intelligence, and it is incredible what is happening around the world. The generative AI era has clearly started. The combination of large language models and generative models, these auto regressive generative models has kicked off the generative AI era. Thousands of papers in just the last several years have been written about this area of large language models and generative AI. Billions of dollars are being invested into companies, and just about every single domain. And every single industry is pursuing ideas on generative AI. And the reason for that is very simple. The single most valuable thing that we do as humanity is to generate intelligent information. And now for the very first time, computers can help us augment our ability to generate information.

And a number of startups are just doing amazing things. Of course, they're doing content creation. But they're also using generative AI to steer the steering wheel of a self-driving car or animate, articulate the robotic arm, generate proteins, chemicals, discover new drugs, even learning the structure of physics so that we can generate physics of mesoscale multi-physics, maybe accelerate the understanding of climate change.

Well, here's some examples of some amazing things. This is the Adobe Firefly. Adobe Firefly does our painting, imagine the space around the image that we never captured? MOVE Ai does MOCAP from just video. This is on the upper right. You could -- you decide which one's real. I'm going with the left. This con this sketch to image guided by language prompt. This one's really cool. There are a lot of people who know how to sketch. And from the sketch and some guidance from your language, you could generate something photorealistic and rendered. The future of computer graphics is clearly going to be revolutionized. And this is really cool wonder dynamics, not only is the name of the company, cool. But they do pose and lighting detection and replace the actor with a CG character. They're just -- they've just goes on and on and on the number of generative AI startups around the world, it's I think we're coming up on something like 2000. And they were in just about every single industry, the generative AI era has arrived.

Well, what's really profound, though, is that when you take a step back and ask yourself, what is the meaning of generative AI? Why is this such a big deal? Why is it changing everything? Well, the reason for that is, first, human is the new programming language. We've democratized computer science. Everybody can be a programmer now, because human language, natural language is the best programming language. And it's the reason why ChatGPT has been so popular. Everybody can program that computer.

Large language model is a new computing platform. Because now the programming language is human. And what's your program that computer understands large language models, and generative AI is the new killer app. These three insights has gotten everybody just insanely excited. And because for the very first time, after 15 years or so, a new computing platform has emerged, like the PC, like the internet, and like mobile cloud computing, a new computing platform has emerged. And this new computing platform is going to enable all kinds of new applications, but very differently than the past. This new computing platform benefits every single computing platform before it.

Notice, one of my favorite -- the thing I'm looking forward to most is generative AI for office. There's so many different things that I do in office today, it would be great to plug in generative AI to help me be more productive in that. Generative AI is going to be plugged into just about every digital content creation tool, every single CAE tool, every single CAT tool. For the very first time, this new computing platform not only enables new applications in this new era, but helps every application in the old era. This is the reason why the industry is moving so fast.

Well, one of the one of the important things that's going to happen is this application spaces new way of doing computing is so profoundly different, that the computer will be reinvented, the computer itself, the computer itself, will of course process information in a very different way. And we need a new processor, the computers and world, computing is done in so many different places. And sometimes they're used for training, sometimes they're used for inference. Sometimes in the cloud. Sometimes it's for scale up sometimes it's for scale out. Sometimes it's for enterprise, sometimes it's underneath your desk in your workstation.

There's so many different ways that computing needs to be refactored. And Nvidia's accelerated computing will support every single one of those. But one particular area is extremely important, which is the basic scale out of the cloud. The basic skill out of the cloud historically, was based on off the shelf CPUs, x86 CPUs, while general purpose computing is a horrible way of doing generative AI, and you can see that in just a second. And so we created a brand-new processor for the era of generative AI. And this is it. This is the Grace Hopper. We announced Grace Hopper, in fact, just only recently several months ago, and today we're announcing that we're going to give it a boost. We're going to give this processor a boost with the world's fastest memory called HBM3e. The world's best memory -- fastest memory hub connected to Grace Hopper. We're calling it GH200. The chips are in production will sample it at the end of the year or so and be in production by the end of second quarter.

This processor is designed for scale out of the world's data centers, has 72 cores, gray CPU cores connected through this incredibly high-speed link cache-coherent, memory coherent link between the CPU and the GPU. This is the CPU and that's the GPU. The Hopper GPU is now connected to HBM3e, has four petaflops, have transformed engine processing capability. And now it has five terabytes per second of HBM3e performance. So this is the new GH200 based on the architecture, Grace Hopper, and a processor for this new computing era.

There's a whole lot of ways that we can connect Grace Hopper into a computer. This is one of my favorites, by connecting two of them into one computing node, connecting it together with NVLink. And this NVLink between these two processor modules is six terabytes per second. And it basically turns these two processors, these two super chips into a super-sized super chip, one giant GPU, one giant CPU, the CPU now has 144 cores. The GPU has 10 terabytes per second of frame buffer bandwidth, 10 terabytes per second of frame buffer bandwidth, and 282 gigabytes of HBM3e.

Well, pretty much, you could take just about any large language model you like, and put it into this, and it will inference like crazy. The inference cost of large language models will drop significantly, because look how small this computer is. And you could scale this out in the world's data centers. Because the servers are really, really easy to scale out, you can connect this with Ethernet, you can connect it with InfiniBand. And of course, there's all kinds of different ways that you can scale it out.

Let's take a look at what it means. If you were to scale it, take this and now scale it up into a giant system. This is two GPUs. But what if we would like to scale this up into a much, much larger GPU. Run it, please.

All right, this is actual size by the way. This is actual size. And it probably even runs crisis. The world's largest single GPU one exaflop, 4 petaflops per Grace Hopper, 256 connected my NVLink into one giant system. And so this is a modern GPU. So next time when you order a GPU on Amazon, don't be surprised if this shows up.

Okay, so that's how you take Grace Hopper and scale it up into of course, a giant system. Future frontier models will be built this way. The frontier models of the past, like GPT3 and GPT4 and Llama are the mainstream models of today. Only after a couple of years, these frontier models which was just gigantic to train on systems like this, and the future becomes mainstream. And once they become mainstream, they can be scaled out into all kinds of different applications and how would we scale these out? So let me show you this. This is how you would do it. And so now you would have a single Grace Hopper and each one of these nodes.

This is the way computing was done in the past. For the last 60 years ever since the IBM system 360 the central processing unit or general-purpose computing was relatively mainstream. And for the last 60 years, that's the way we've been doing computing. Well, now, general purpose computing is going to give way to accelerated computing and AI computing. And let me illustrate to you why, the canonical use case of the future is a large language model and the front end of just about everything, every single application, every single database, whenever you interact with an app, whether its computer, it will likely be first, you'll likely be first engaging a large language model. That large language models will figure out, what is your intention? What is your desire? What are you trying to do, given the context and present the information to you in the best possible way? It will do the smart query, maybe a smart search, augment that query and search with your question, with your prompt and generate whatever information necessary.

And so the canonical example that I'm using here is a Llama to large language model that has been influenced, it then does a query into a semantic database, a vector database of some kind. And the output of that is augmented and becomes a guide for a generative model. And here, the generative model I'm using is stable diffusion XL. So these three models, Llama 2 vector database, and stable diffusion SDSL, are relatively well understood as state of the art and the type of models that you could imagine running just by everywhere.

Well, if you were to have an ISO budget way of processing that workload, it would take, let me just choose the number $100 million and 100 million dollars and be a reasonably small data center these days, 100 million dollars will buy you about 8800 x86 GPUs, it would take about five megawatts to operate that, and I normalized the performance into 1x.

Using the exact same budget with accelerated computing Grace Hopper, it would consume only three megawatts. But your throughput goes up by an order of magnitude. Basically, the energy efficiency, the cost efficiency of accelerated computing for generative AI applications is about 20x, 20x in Moore's law, and just the current way of scaling CPUs, that would be a very, very long time. And so this is a giant step up in efficiency and throughput. So this is ISO budget.

Let's take a look at this now again, and let's go through ISO workloads. Suppose your intention was to provide a service. And that service has so many number of users. And so your workload is fairly well understood plus or minus. And so with ISO workload, this 1x, $100 million, using general purpose computing, and use the accelerated computing Grace Hopper, it would only cost $8 million dollars, $8 million, and only 260 -- not megawatts, yeah 260 kilowatts, so 20 times less power, and 12 times less cost. This is the reason why accelerated computing is going to be the path forward. And this is the reason why the world's data centers are very quickly transitioning to accelerated computing. And some people say, and you guys might have heard, I don't know who said it. But the more you buy, the more you save. And that's wisdom.

If I could just ask you to remember one thing from my talk today, that that would really be it, that the future is accelerated computing. And the more you buy, the more you save. Well, today I want to talk about something really, really important. And so the backdrop accelerated computing, the backdrop generative AI, the backdrop the things that real time ray tracing, the future of computer graphics unified with AI.

Let's talk about a couple of new things. And so today I want to talk about Omniverse and generative AI and how they come together. The first thing that we already established is that graphics and artificial intelligence are inseparable. That graphics needs AI and AI needs graphics. Graphics needs AI and AI needs graphics. And so the first thing that you could imagine doing for the future of artificial intelligence is the teacher common sense. All of us understand the consequences of the physical actions we take.

All of us understand that, gravity has effect. And all of us understand that even though you don't see something that object might still be there probably is still there because object presents. And so that common sense is known to humans ever since your babies. And yet, for most artificial intelligence agents that learn on large language models, it's unlikely it has that common sense that object permanence, the effects of gravity, the consequence of your actions, you have to learn it in a physically grounded way.

And so the thing that we could do is we can create a virtual world that is physically simulated, physics simulator that allows an artificial intelligence to learn how to perceive the environment, using a vision transformer maybe. And to use reinforcement learning to understand the impacts the consequences of its physical actions, and learn how to animate and learn how to articulate to achieve a particular goal. And so one mission of a connected artificial intelligence system. And a virtual

world system that we call Omniverse is so that the future of AI could be physically grounded. The number of applications is really quite exciting, because, as we know, the largest industries in the world are heavy industry. And those heavy industries are physics based, physically based. And so, first application is so that AI can learn in a virtual world.

The second application, the second reason why AI in computer graphics are inseparable, is that AI will help also to create these virtual worlds. Let me give you a couple of examples. This is an AI that is a large language models I mentioned that will be connected to almost every single application. However you -- the future user interface of almost every application is a large language model. And so it's sensible to imagine that this large language model could also be a query front end to a 3d database. And so here, find a Denza N7 SUV.

Now once you find this SUV, you might ask an AI agent to help you to turn this car to embed this car, to integrate this car into a virtual environment. And instead of designing a virtual environment, you ask the AI to help you. Give me a road in the desert at sunset.

Now, inside Omniverse we can then unify aggregate composite these two, this information together, and now the car is integrated rendered into position into a virtual world. And so here's an AI that helps you maybe create and find some manager data assets. You also have an AI that helps you generate a virtual world around it. And Omniverse allows you to integrate all this information.

Well, let's take a look at why WPP, the world's largest ad agency and BYD the world's largest electric vehicle maker, are used -- how they're using Omniverse and generative AI in their work. Play it, please.

Unidentified Speaker

WPP is building the next generation of car configurators for automotive giant BYDs Denza luxury brand powered by Omniverse cloud and generative AI, open USD and Omniverse cloud allows Denza to connect high fidelity data from industry leading CAD tools to create a physically accurate, real time digital twin of its N7. WPP artists can work seamlessly on this model in the same Omniverse cloud environment with their preferred tools from Autodesk, Adobe and side effects to deliver the next era of automotive digitalization and immersive experiences.

Today's configurators require hundreds of thousands of images to be prerendered to represent all possible options and variants. Open USD makes it possible for WPP to create a super digital twin of the card that includes all possible variants in one single asset deployed as a fully interactive 3D configurator on Omniverse Cloud GDN, a network that can stream high-fidelity real-time 3D experiences to devices in over 100 regions were used to generate 1000s of individual pieces of content that comprise a global marketing campaign. The USD model is placed in a 3D environment that can either be scanned from the real-world using LiDAR and virtual production are created in seconds with generative AI tools from organizations such

as Adobe and Shutterstock. This innovative WPP solution for BYD brings generative AI and cloud rendered real time 3D together for the first time powering the next generation of e-commerce.

Jensen Huang {BIO 1782546 <GO>}

Got to love that. Everything -- everything was rendered in real time nothing was pre-rendered. Every single scene that you saw was rendered in real time, every car, all of the beautiful integration with the background, all the rendering, everything is 100% real time. The car is the original CAD data set of BYD nothing was changed. You literally take the CAD, drag it into Omniverse, you tell an AI, synthesize and generate an environment. And all of a sudden the Car appears wherever you'd like it to be. So this is a one example of how generative AI and human designs come together to create these incredible applications.

And so how do we do this? Applications, generative AI models are making tremendous breakthroughs. And what you want to do, we all want to do this, there are millions of developers and artists and designers around the world and companies every single company would like to take advantage of. And certainly everybody has been video working hard to utilize large language models and generative AI in our work. In fact, the Hopper GPU was impossible designed by humans, we needed AIs and generative models to help us find the way to design this thing in such a high-performance way.

And so it augments our design engineers, it makes it possible for us to create some of these amazing things at all. And of course, the productivity of the teams go up tremendously. Well, we would like to do this in just like every single industry. So the first thing that we have to do is we have to go find a model that works for us, so that we can fine tune it. You can't just use the model as is you want to fine tune it for your curated data.

The second thing you want to do is to augment your engineers, your artists, your designers, your developers with the capability of these generative models, so augmenting it composing the information together in this particular case, I'm using media and entertainment where virtual world is an example. And this is the reason why Omniverse is central to that. We want to be able to run this in the cloud, of course, and we'll continue to run this in the cloud. But as, you know, computing is done literally everywhere.

AI is not some widget that is has a particular capability, AI is the way software is going to be done in the future. AI is the way computing will be done in the future. It will be literally in every application, it will be run in every single data center. You'll run it every single computer at the edge in a cloud. And so we want to have the ability to not just do AI -- generative AI in the cloud, but to be able to do it literally everywhere in the cloud and data center workstations, your PCs.

And we want to do this by making it possible for these really complicated stacks to run. There's a reason why the world's AI is done largely in the cloud today. We

partner very closely with the CSPs, the amount of acceleration libraries and all the runtimes. And from data processing to training, to inference, to deployment, the software stack is really complicated. The libraries and runtimes is getting it to run on a particular device and system is incredibly hard. And that's the reason why it's stood up as a managed service that everybody can use.

Well, we believe that in order for us to democratize this capability, we have to make it run literally everywhere. So we have to have these unified optimized stacks, be able to run on almost any device and make it possible for you to engage AI. Well, the first question is, where are the world's models? Well, the world's models are largely on Hugging Face today. It is the largest community of AI, the larger the AI community and world. Lots and lots of people use it, 50,000 companies, 2 million users I think, 50,000 companies engage Hugging Face. There's some tumors 75,000 models 50,000 datasets. Just about everybody who creates an AI model and wants to share with the community, puts it up in Hugging Face.

So today we're announcing that Hugging Face is going to build a new service to enable their community to train directly on NVIDIA DGX Cloud. NVIDIA DGX Cloud is the best way to train models. And its footprint is being set up, our DGX cloud footprints are being set up in Azure, OCI, Oracle Cloud, and GCP. So the footprint is going to be largely everywhere. And you'll be able to find from the Hugging Face portal, choose your model that you would like to train or you'd like to train a brand-new model and connect yourself to DGX cloud for training. So this is going to be a brand-new service to connect the world's largest AI community, the with the world's best AI training infrastructure. So that's number one. Where do you find the models.

But you want to do this in the cloud, but you might also want to do this everywhere else, and how do you build that infrastructure for yourself. And so the second thing we're announcing today is the NVIDIA AI Workbench. This Workbench is a collection of tools that make it possible for you to assemble -- to automatically assemble the dependent runtimes and libraries, the libraries to help you fine tune and guardrail to optimize your large language model, as well as assembling all of the acceleration libraries which are so complicated so that you could run a very easily on your target device. You could target a PC, you could target a workstation, you could target your own data center, or with one click you can migrate the entire project into any one of these different areas. Let's take a look at NVIDIA AI Workbench in action.

Unidentified Speaker

Generative AI is incredibly powerful, but getting accurate results customized with your secured proprietary data is challenging. NVIDIA AI Workbench streamline selecting foundation models building your project environment and fine tuning these models with domain specific data. Here AI Workbench is installed on a GeForce RTX 4090 laptop where we've been experimenting with an SD Excel project. As our project gets more complex, we need much more memory and compute power. So we use AI Workbench to easily scale to a workstation powered by 4 Nvidia RTX 6008 a generation GPUs. AI Workbench automatically creates your projects environment, building your container with all dependencies including Jupyter.

Now, in the Jupyter Notebook, we prompt our model to generate a picture of Toy Jensen in space. But because our model has never seen Toy Jensen, it creates an irrelevant result. To fix this, we fine tune the model with eight images of Toy Jensen, then prompt again. The result is much more accurate. Then with AI Workbench we deploy the new model in our enterprise application. This same simple process can be applied when customizing LLM such as Llama-2-70B.

To accommodate this much larger model. We use AI workbench to scale to the datacenter accessing a server with 8 NVIDIA L40S GPUs. We tune with 10,000 USD code snippets in nearly 30,000 USD functions built by Nvidia, which teaches the model to understand 3D USD based scenes. We call our new model CHATUSD. CHATUSD is a USD developers copilot, helping answer questions and generate USD Python code. With NVIDIA AI workbench, you can easily scale your generative AI projects from laptop to workstation to data center or cloud with a few clicks.

Jensen Huang {BIO 1782546 <GO>}

Everybody could do this. Just have to come to our website download, NVIDIA AI Workbench. Anybody could do this. Now turned out my parents gave me a Swedish name. As you know, it's Jensen, and I'm pretty sure that when they looked up Toy Jensen, that's why turned out that way. And it took a few more examples to turn them into Toy Jensen.

Okay, so everybody can do this, come to the website, Early Access download AI Workbench, it's for the creator of the project, you -- it helps you set up the libraries and the runtimes that you need, you can fine tune the model, if you want to migrate this project so that all of your colleagues can use it and fine tune other models. You could just tell it where you want to migrate it to and one click on migrate the entire dependency of the project all the runtimes all the libraries, all the complexities, and it runs on workstations and runs in the data center and runs in the cloud, one single body of code, one single project allows you to run literally everywhere. And so everybody can be a generative AI practitioner.

Well, what makes it possible to do all this is this other piece of code called NVIDIA AI Enterprise. This is essentially the operating system of modern data science and modern AI. It starts with data processing, data curation and just data processing represents some 40%, 50%, 60% of the amount of computation that is really done before you do the training of the model. So data processing, then training, then inference and deployment. All of those libraries, there are 4500 different packages that are inside the NVIDIA AI enterprise with 10,000 dependencies. This represents literally the NVIDIA 30-year body of work. Starting for CUDA, all the CUDA acceleration libraries, and everything else is accelerated for the GPUs that are the couple of 100 million GPUs that are all over the world, all CUDA compatible, make it -- make all of them run, it has the ability to support multi-GPU in a multi node environment, and every single version of our GPU 100% compatible with everything.

And that -- this operating system of AI, if you will, has been integrated into the cloud, integrated with leading operating systems like Linux and Windows WSL 2, windows

subsystem for Linux. The second version WSL 2 has been optimized for CUDA and supports VMware, the body of work that we've done with VMware is incredible, the stuff several years to do couple, two and a half years for us to make VMware be CUDA compatible, CUDA aware, multi-GPU aware, and still have all the benefits of an enterprise one pane of glass, resilient, virtualized data center.

And so this entire stack of immediate AI enterprise, this is really the giant body of work that makes all of this possible. As a result, literally everything that you would like to run will be supported by the ecosystem we're talking about here. It's also integrated above the stack into ML ops applications to help you with the management and the and the coordination of doing data processing, data driven software in your company, as well as will be integrated into AI models that will be provided by ServiceNow, and snowflake. Okay, so Nvidia and AI enterprise is what makes NVIDIA AI Workbench even possible in the first place.

Now we have these incredible models that are in Hugging Face that are pre trained and open sourced. We can now train them and fine tune them on AI Workbench, you could run it anywhere because of AI enterprise. Now, we just need some powerful machines. We have powerful machines in the cloud, of course, DGX Cloud has many, many, many footprints around the world. But would it be great if you had a powerful machine under your desk? So today we're announcing our latest generation Ada GPU, Ada Lovelace GPU. The most powerful GPU we've ever put into workstation is now -- Oh, Gosh, darn it. I just put my fingerprints on there. Did you guys -- can you guys see that?

That's not me. Could you? Hey, can I have this cleaned in the future? My bad. Yuck. Let me show -- that's the worst product launch ever you guys. The CEO pulls it out goes, yuck. This is the data center version of Ada Lovelace. Sorry everybody. My bad. They worked so hard. They worked so hard. It was perfect. It's like beautifully lacquered. Like, I'm sad. I'm super sad. I'm super sad. Okay, anyways. Thank you. Thank you.

And that's why you should rehearse. All right, it goes into these amazing workstations and these amazing workstations, packs up to four of these GPUs. It packs up the four Nvidia RTX 6000s, the most powerful GPUs ever created and run real time ray tracing for Omniverse as well as train fine tune and inference, large language models for generative AI, and it's available from box and Dell and HPI and Lambda and Lenovo, and it's available now. Okay, so we're in production with these workstations and with, as I mentioned, Hugging Face to AI Workbench, NVIDIA AI Enterprise running on Windows 11 with WSL 2, you have an amazing AI machine.

You could fine tune GPT 3, 40 billion per GPT 3 in about 15 hours on nearly a billion tokens. And so you could take your proprietary data, your curated data, you could maybe bring all of your PDFs, and you could fine tune this model before you ask it, prompted and ask a question. SDSL could be trained And after it can be fine-tuned, as I mentioned, we showed you an example of us fine tuning with Toy Jensen. You can now generate 40 images per minute. Okay, 40 images per minute, you're going to this workstation will pay for itself, and who knows, depending on how long, how

much you use generative AI these days, it could pay for itself in months. Like I said, the more you buy, the more you save. Okay.

Incredibly fast, incredibly powerful. And it's all yours. It produces answers in seconds, not minutes for in some of the services that are out there. Okay, man. So another incredible machine are the servers and these servers as you know, getting GPUs in the cloud these days is no easy feat. And now you can buy it, okay, you can have your company buy it for you, and put it in the data center. And there's a whole bunch of these servers, a whole bunch of different configurations. I don't know if you guys could see this. This is a server that has up to eight of the L40S Ada Lovelace GPUs.

And of course, these are not going to be used for frontier models. These are not designed to train large frontier models like GPT-4 or GPT-5. These are really used for mainstream models today that you can download from Hugging Face, or Nvidia could work with your company to create based on our language model called Nemo. We could create models that are mainstream today that you could use in just about all kinds of applications around your company. And you could fine tune it with these GPUs.

The fine tuning of a GPT-3 model. Okay, so this is GPT-3 40 billion parameters. It takes about seven hours for about a billion tokens. So 15 hours on a workstation with four GPUs, of course, takes less with eight GPUs. And just in fine tuning this is one and a half times faster than our last generation, A100. So L40S is a really terrific GPU for enterprise scale, fine tuning of mainstream large language models. You can also use it for course, synthesizing and generating images.

Okay. So generative AI for everyone, everybody could do it now. Hugging Face, NVIDIA AI Workbench and NVIDIA AI Enterprise, these amazing new enterprise systems that are in production today.

All right, let's change gears and talk about what's going on at SIGGRAPH this year. I'm pretty sure all of you have already heard about OpenUSD. OpenUSD is a very big deal. SIGGRAPH 2023 is all about OpenUSD. OpenUSD is visionary. And it's going to be a game changer.

OpenUSD is a framework, a universal interchange, for creating 3D worlds for describing, for compositing, for simulating for collaborating on 3D projects. OpenUSD is going to bring together the world onto one standard 3D interchange, and has the opportunity to do for the world and for computing, what HTML did for the 2D Web.

Finally, an industry standard, powerful and extensible 3D interchange that brings the whole world together, a really big deal.

Now, let's take a look at why it is such a visionary thing that Pixar did, it was invented. Well, I forget exactly when they invented it, but they open sourced it in 2015. And they've been, of course, using this framework for over a decade building amazing

3D content. While the 3D pipeline is incredibly complicated, the 3D workflow is specialized and complicated. You got designers and artists and engineers they all specialize in some part of the 3D workflow. It can be modeling and texturing, materials, physics simulation, animation, set design, scene composition. There are so many parts and so many different tools.

And because the tools are created by different companies, and largely incompatible, import and exporting data conversion is just part of the part of the workflow. And because they're incompatible, and because there's all this import and exporting, fundamentally, the workflow has to be serialized it's impossible to paralyze that and the converting of course all of this data is cumbersome and is error prone. So this workflow is fundamentally complex, you could argue that was just designed to be complex. And this is one of the reasons why creating these incredible 3D animation movies are so expensive and takes so much time.

Well, one of the visions, the first vision, of course, of OpenUSD, is to put the data at the center. Could you imagine if every single tool was natively compatible with USD, then as result, data gravitates to the center, everybody, you can work in parallel. The interchange and conversion goes away. And instead of a serialized model, you have a paralyzed spoken hub model. And so this way of doing work, of course, is incredibly appealing. And it's one of the reasons why the vision of OpenUSD has taken off.

Well, there's some 50 tools available now. The industry loves the vision. OpenUSD already has a rich ecosystem, some 50 tools are now compatible with OpenUSD natively. 170 contributors in the USD forum from about 100 companies. So you got a lot of people really interested in this. And the momentum is growing. It's being adopted in film, architecture, engineering, and construction, manufacturing, and so many different fields of robotics. We're engaged with companies in so many different parts of the industry, all excited about USD because everybody's workflow, like creating movies is complicated. This is no different than a company who's in manufacturing, a company who's trying to build a new building. So many different specialists, so many different contractors, so many different parties coming together to build something complex for the very first time, we have an interchange, a standard interchange that can bring everybody together. So super, super exciting.

Well, five years ago, we started working with Pixar, and we adopted USD as the foundation of Omniverse. Our vision was to create this these virtual worlds that make it possible for us to bring world design into the applications that I mentioned, industrial digitalization at the core of many things that we want to do, not just for, of course, creating amazing movies and broadcast and video games. But also to take 3D worlds physically-based real time into the world's industries, we felt that we could make a real impact.

And so we selected USD, it was a brilliant move, the team made a just a visionary move to partner with Pixar, to select USD as the foundation of Omniverse. This is probably the world's first major platform, incredible database and engine system that was built completely from the ground up for USD. Every single line of code was

designed with USD in mind. The platform for USD for describing simulating all the things that we just mentioned.

And so Omniverse was designed to connect. It's not a tool itself. It's a connector of tools. It's not intended to be a final production tool. It's intended to make be a connector that make it possible for everybody to collaborate, interchange share, live, work. Okay, so Omniverse is a connector. Well, let's take a look at how the vision of OpenUSD come together. And this is just a fantastic illustration.

Let's start from the left here. I think this is Adobe Stager Houdini, this is a modeling system, Maya or animation system, modeling system. This is Omniverse, blender, render man, Pixar's Minuteman and Unreal Engine from Epic, a game engine. Literally all OpenUSD. One dataset ingested into everybody's tools, and it looks basically the same. Everybody's rendering system is a little different. And so the quality of the rendering is a little different from tool to tool, but one data set available and usable by every tool. This is the vision of OpenUSD. So incredibly powerful.

Well, we've been investing in USD now for over five years. This SIGGRAPH is the five if you will, we've been working on USD now for about five years, the SIGGRAPH. And we've been working on extending USD to real time and physics-based systems for industrial applications. We brought RTX to it. We bought -- we extended USD with schema for physics, real time physics and offline physics. We added CAT to USD, Connect USD to a whole new industry. We made it possible to understand geospatial data to recognize and understand comprehend, consider the curvature of the earth.

We integrated it with AI runtime, as well as a framework to build generative AI. For example, the deep search that we showed you, ChatGPT or CHATUSD that I'll show you in just a second. We created a -- we extended USD for assets that are physics, accurate physics aware, so we call it SimReady. It's particularly interesting, particularly important for robotics applications, so that the joints move accordingly and such. And we took USD and made it hyperscale so that we can expand it and grow it, make it support datasets of enormous scale and put in the cloud, connected it with open XR and reality kit so that we can stream from the cloud to spatial computing devices.

Well, for the last five years we've been working on Omniverse has been building and working collaborating with the industry on USD. Let's take a look at this everything you're about to see is a simulation. Everything is real time. And so take a look at this, this is the latest of Omniverse.

Just make you happy. No art, all physics, physics make you happy, doesn't it? Physics makes you happy? Okay, well, we wanted to put Omniverse anywhere, you could download Omniverse from our website and run it on your PC in your workstation for enthusiast and designers that that's perfect. You could also license Omniverse for

enterprise. And for enterprises that are using it across many different organizations. We even set up a managed service for you inside your company.

We're putting Omniverse now in the cloud so that we could host and serve up API's that can be connected to developers and applications and services so that you can have the benefit of some of these amazing capabilities. And so we're setting up Omniverse cloud.

Now, Omniverse, as I mentioned before, is not a tool. It's a platform for tools. It's not a tool, it's a platform for tools. And we created a whole bunch of interesting tools that to help you get started. There are reference applications, and many of them are open sourced. And one application for example, that I really love is Isaac Sim. It's a gem for teaching robots for robots to learn how to be a robot. And because it's so physically accurate, the sim to real gap is reduced. And so theoretically, you should be able to learn as a robot how to be a robot inside Omniverse. And that neural network, that software could then be put into a local embedded device like a Jetson or one of the Nvidia's Jetson computers, robotic computers, and the robot can perform its task.

Okay. And so we would like to put Omniverse in as many places as possible, there's a whole bunch of different applications. And this SIGGRAPH, we're announcing a few new API's really cool API's. Now, we demonstrated a really cool API just recently, it's called ACE, Avatar Cloud Engine. And so it understands speech. So you could do speech with it. It talks, recognize your voice, speaks to you, based on what it's saying, based on the sound that it's making. It animates the face accordingly, audio to face. And so we call that the ACE engine in the cloud.

We're going to show you a couple of new APIs. And this one API, this is the run USD API. Of course, this should be the first API. What do you guys think? Pretty cute, huh? And so you send to the cloud USD. And what comes out of the cloud, what streams from the cloud onto your device in open XR or reality kit to your spatial computing device will be this incredibly, beautifully rendered, and very importantly, interactive USD. So let's take a look at it.

So for USD programmers, USD developers, you will have hours of joy, just hours and hours of joy. And you can just create your USD content, USD asset, loaded up on invidious Omniverse cloud and enjoy the device, enjoy the asset on your device.

Now, this allows you, of course, to test the compatibility of your USD. And so we now have a universal compatibility tester up in the cloud. And so whenever you have USD content, big or small, you can load it up on Omniverse Cloud and deploy it independent of which version of USD that you're using well tested for compatibility. Okay. And so this is going to be free for developers.

There's another API that we're creating. And we showed you earlier how we used AI workbench to train this model, to fine tune this model. We started with Llama 2. And we taught it, we fine-tuned it for USD. And so let's take a look at the video.

Unidentified Speaker

For USD developers, building profiling and optimizing large 3D scenes can be a very complex process. CHATUSD is an LLM that's fine tuned with USD functions and Python USD code snippets using Nvidia AI Workbench and the NeMo framework. This generative AI copilot is easily accessed as an Omniverse cloud API, simplifying your USD development tasks directly in Omniverse. Use CHATUSD for general knowledge like to understand the geometry properties of your USD schema, or complete previously tedious repetitive tasks, like generating code to find and replace materials on specific objects or to instantly expose all variants of a USD print.

ChatUSD can also help you build complex scenes, such as scaling a scene and organizing it in a certain way in your USD stage. Built bigger, more complex virtual worlds faster than ever with ChatUSD generative AI for USD workflows.

Jensen Huang {BIO 1782546 <GO>}

ChatUSD. ChatUSD, now everybody can speak USD and ChatUSD can be a USD teacher, it can be a USD copilot and help you create your virtual world. Okay. Enhance your productivity incredibly, and this is going to be also available on the on Omniverse cloud.

Well, I showed you some examples, probably the largest opportunity for the world of it for software, and for artificial intelligence is to help revolutionize the world's heavy industries. They're just enormous amounts of waste, as we all know, $50 trillion worth of industry. Over the next several years before the end of the decade, there will be trillions of dollars of new EV factories, battery factories, and new chip fabs that are going to be built all over the world. Not to mention the enormous number of factories that are already in operation, some 10 million factories are in operations today.

This industry would love to be digital. They would love the benefits of all of our industries. But unfortunately, their industry has to be physically coherent. It's physically based. They build things. They build and operate physical things. But they would love to do it digitally, just like us. And so how can we help them do that? Well, this is where Omniverse and artificial and generative AI comes together for us to be able to help the heavy industries of the world digitalize their workflow, just as modeling and texturing and lighting and animation and set design and so on so forth, are all done by different groups in a very complicated pipeline. This is very much the case of the world's heavy industries. Every one of their organizations from design to styling, to engineering and simulation and testing, to factory building and design, factory planning to build a products and even operating these software defined robotics assisted products in the future. All of this is done completely mechanically today.

That entire flow could be digitalized, and it could be integrated for the very first time using OpenUSD. This is the incredible vision of OpenUSD. Why we're so excited about it. If we can only augment it with real time capability and physics simulation

capability, and make it so that every single tool is connected to Omniverse. We can digitalize the world's industries.

Well, their excitement is enormous, because they all would love to have the productivity. We'd love to reduce the energy consumed. We love to reduce the waste. We love to reduce the mistakes in digital long before they have to build it in physical. And so this is Mercedes, they're using Omniverse to digitalize their manufacturing lines. This is Mercedes using Omniverse to simulate autonomous vehicles. This BMW using Omniverse to digitalize their global network of factories, some 30 factories that now building an Omniverse without breaking ground and doing the entire integration. A year before the factory is actually even built. Using Omniverse to simulate new electric vehicle production lines. Remember, the placement of the factory, the planning of the factory, and the programming of all the robotic systems are just incredibly complicated in the future. Entire factories will be software defined, the factory will be robotic, orchestrating a whole bunch of robots that are building cars that themselves are robotic. So robots building robot, orchestrating robots, building robots. So that's the future and everything is all software-driven, everything is all AI enabled.

This is BMW using AI to drive their operations. This is Westron using Omniverse to digitalize their production line to build this machine. As I mentioned, 35,000 parts incredibly complicated, one of the most expensive ones, the most valuable instruments that's made anywhere. And so having that line be completely robotic and automated is really important.

This is Pegatron using Omniverse to digitalize PCB manufacturing. Again, this is the PCB of this incredibly complicated PCB motherboard. This is the most complex motherboard the world's ever made. Techman is using Omniverse to test and simulate Cobots. It's not surprising to you but most Cobots, most robots today are not very autonomous, not very AI-driven, programming the robots usually cost way more than robots themselves. I heard it's a statistic that the robot for manufacturing arm for the automotive industry is something along the lines of $25,000. Not very much, but programming it could cost a quarter million dollars, which is very sensible. We would like to have Ai be self-programming these autonomous, autonomous limbs. So Techman is using Omniverse to tests its simulate Cobots.

They're also using Omniverse to build applications to automate optical inspection. Of course, for PCB lines, the cameras can be stationary, the product, the manufacturing system is just rolling by you. But for many things like cars and other very complicated systems, the optical inspection has to follow the curvature and of course, the various contours and shapes of the product.

Hexagon is using Omniverse to connect just as we connect to tools all over the world. Hexagon is using Omniverse to connect their own tools. This is one of the most powerful things that they observed in Omniverse. Whereas they had different groups and different teams are in silos and because they're incompatible tools, getting them to connect together was hard. And so this was a company organization and a company management challenge. So very first time using Omniverse. They

broke down the silos, they connected all the different teams, and they became one unified company for the very first time, really, really cool.

This is ready robotics using Omniverse to build applications that simplify the robot programming process. In the future, robot programming is probably going to be about explaining and prompts what you would like the robot to do, showing a few examples and just as we taught our language model, our generative model Toy Jensen, just as we taught the generative model USD, we're going to teach a generative model of the future a robotic generative model, a few examples, and it will be able to generalize and do that task.

Amazon is using Omniverse to digitalize their warehouse, the warehouse is robotic, giant system, help cut help their workers inside not have to work as -- work as -- walk as far. Amazon is using Omniverse to simulate their fleet of AMRs. These are autonomous moving robots and using Omniverse to generate synthetic data. To train the precession models the computer vision models of these robots.

You can also use Omniverse to create a digital twin. Nvidia is creating a digital twin of the Earth, the climate system of the earth. Deutsche Bahn is using Omniverse to create a digital twin of their entire railway network. So they could operate it completely in digital. Now, in order for that to happen Omniverse has to be real time.

Let me show you one more example. And this one example is about human designers, architects working side by side with generative AI models from different applications and different companies. And together they will automate and try to and help do industrial digitalization a lot more rapidly. Okay, roll it.

Unidentified Speaker

Planning industrial spaces like factories or warehouses is a long complex process. Let's see how you can use NVIDIA Omniverse in generative AI to connect your OpenUSD to fast-track planning concepts like a storage extension to an existing factory. You think twins Omniverse extension to quickly convert a 2D CAD floorplan into a 3D OpenUSD model. And populate it with SimReady OpenUSD assets using Omniverse' AI-enabled DeepSearch. Then use prompts to generate physically accurate lighting options with Blender GPT. Realistic floor materials with Adobe Firefly and an HDRI Skydome with BlockadeLabs.

To see the new space and context compose it on a cesium geospatial plane next year existing factory digital twin, then to share with stakeholders use one click to publish the proposal to Omniverse Cloud GDN which serves a fully interactive review experience to any device. Fast track your factory planning process with NVIDIA Omniverse and generative AI.

Jensen Huang {BIO 1782546 <GO>}

How incredible is that? Remember, it started with a 2D PowerPoint slide. And it ended with a virtual factory in spatial computing. That is incredible PowerPoint to a virtual factory, 2D to spatial 3D. So this is the future, this is the future and this is how everything comes together. USD, of course, is foundational in that journey, and Omniverse foundational in that journey and generative AI.

Well, this is what Omniverse is, we're super excited about the work that we're doing here. We're so happy that we chose OpenUSD as the foundation of Omniverse. And all the work that we've done to extend it into real time into physics-based applications. The number of the number of partners that we have is just growing incredibly, and it covers so many different industries, as I mentioned, from manufacturing to robotics, and others. And we hope, and this is the beginning of a journey, that we will finally be able to digitalize to bring software driven artificial intelligence powered workflows into the world's heavy industries, the $50 trillion worth of industries that are wasting enormous amounts of energy and money and time, all the time, because it was simply based, built on technology that wasn't available at the time, and so Omniverse for industrial digitalization.

Well, all of this momentum that we've already seen with OpenUSD is about to get trouble charged. Alliance for OpenUSD was announced with Pixar, Apple, Adobe, Autodesk, Nvidia as the founding members. The alliance's mission is to foster development and standardization of open USD and accelerate its adoption. So whatever momentum we've already enjoyed, the vision that we've already enjoyed, it's about to get kicked into turbo charge.

Well, I want to thank all of you for coming today. We talked about SIGGRAPH. SIGGRAPH 2023 for us is four things. It's about the transition of a new computing model. The very first time in decades that the computing architecture is going to be fundamentally redesigned from the processor to the data center, to the middleware, the AI algorithms and the applications that enables. The processor we created for this era of accelerated computing and generative AI is Grace Hopper. And we call it GH200.

We have Nvidia AI Workbench to make it possible for all of you to be able to engage generative AI, NVIDIA Omniverse now has a major release with generative AI and of course, release in support for OpenUSD. And then finally, whether you want to compute in the cloud or do AI in your company underneath your desk or in your data center, we now have incredibly powerful systems to help you all do that.

I want to thank all of you for coming. But before you go, I have a treat for you. This is an anniversary understand of computer graphics. SIGGRAPH has been a really special event and a very special place in computer graphics, as you know is the driving force of our company and very dear to us, we have dedicated probably more time, more engineering more R&D quarter of a century over quarter -- well, 30 years dedicated to advancing computer graphics. I don't know any company who has invested so much. This industry is dear to us. The work that you do is dear to us. And if not for the work that you do all of you who come to SIGGRAPH each year, how was it possible that AI would have achieved what it is today? How would it been possible

that Omniverse would be possible or OpenUSD would be possible. So we made something special for you. Please enjoy.

Unidentified Speaker

It all started over 50 years ago, with a simple question. What if a computer could make pictures? How could we use them? And what would they look like generations from now?

Hello, folks. I'm Mr. Computer Image.

Welcome in, I hope you're hungry.

Jensen Huang {BIO 1782546 <GO>}

Thank you all for coming. Have a great SIGGRAPH 2023. And remember accelerated computing and generative AI, the more you buy, the more you save.