# GeForce RTX 30 Series Launch Presentation

## Company Participants

- Jensen Huang, Founder, President and Chief Executive Officer

## Presentation

### Jensen Huang {BIO 1782546 <GO>}

Welcome to my kitchen. I hope all of you are staying safe. We're going to talk about an amazing GPU today. Modern GPUs are technology marvels. It is the engine of large industries from design, cloud AI to scientific computing, but it is the gamers and their insatiable demand that is the driving force of the GPU. Pooling their GPUs to create the largest distributed computer ever, a million gamers united to counterstrike the COVID-19 coronavirus. The result was 2.8 exaflops, five times the processing power of the world's largest supercomputer to simulate the virus.

Folding@home was able to simulate 100 milliseconds, a tenth of a second in the life of the coronavirus and captured the moment it opens its mouth to infect a human cell. Scientists believe this is also its moment of weakness. Thank you all for joining this historic fight. We're going to talk about computer graphics and the work we're doing to push the boundaries.

We love computer graphics and have advanced it incredibly in the time of NVIDIA. As the technology advanced, the expressiveness of the medium has made graphics an invaluable tool to help us understand our world, create and explore new worlds, tell stories that inspire us. From science to industry, to the arts, computer graphics has made a profound impact on the world. And for that, we are privileged to have contributed.

We're going to talk about gaming and the infinite ways that gaming is expanding. GeForce PC gaming is large and thriving. It's open and rapidly advancing technology, combined with the amazing creativity of the community, makes magic. Anyone could be a broadcaster now. Add a GeForce and you have a personal broadcast station. Pros stream their practices. Experts stream tips and tricks. Friends stream to friends just to hang out. There are over 20 million streamers.

Games have become a new art medium. In Minecraft, gamers can build their work of art. Machinima artists create cinematics made from game assets. Tens of millions are using games to express their creativity. Inside a computer simulation, any sport can become eSport. Virtual NASCAR and F1 are already attracting top racers. Like sports, eSports captures the thrill of victory and the agony of defeat, and the human drama of athletic competition. eSports is on its way to be the biggest sport.

I have something special for all the GeForce gamers around the world, four gifts. I hope you like them and you will find new ways to game. First, big news, Fortnite is turning RTX on. Now Minecraft and Fortnite, the number one and number two most played games in world have RTX on. Fortnite will get ray trace shadows, reflections, ambient occlusion and DLSS 2. These effects look fantastic with the art style of Fortnite. I can't wait to see a Fortnite concert with RTX on. The last one with Travis Scott was watched by 28 million people. Epic made a trailer for you. Let's play it now.

(Video Presentation)

75% of GeForce gamers play eSports. eSports is a game of milliseconds. Reaction time is a combination of the gamer and the machine. Let me explain. This is Valorant. In this example, the opponent is traveling at 1,500 pixels per second, and is visible in this opening for only 180 milliseconds.

A typical gamer has a reaction time of 150 milliseconds from full time to action. You can only hit this opponent if your PC adds less than 30 milliseconds. Most gamers have latencies far greater than 30 milliseconds, many up to 100 milliseconds. Today, we're announcing a new eSports technology called NVIDIA Reflex.

NVIDIA Reflex optimizes the rendering pipeline across CPU and GPU to reduce latency by up to 50%. In September, we're releasing Reflex with our Game Ready Driver. Over 100 million GeForce gamers will instantly become more competitive. Valorant, Fortnite, Apex Legends, Call of Duty: Warzone and Destiny 2 will be the first to integrate Reflex technology.

ESports pros and enthusiasts strive for zero latency. For you, we're announcing an insanely fast and beautiful display, a 360-hertz G-SYNC display designed for eSports. This display has a built-in precision latency analyzer, just connect your mouse. The NVIDIA 360-hertz G-SYNC eSports displays are arriving this fall from Acer, Alienware, ASUS and MSI. We've made a video comparing gaming on a 60-hertz, 144-hertz and 360-hertz display. You can see immediately how 360-hertz display will help you target and track an opponent.

For the 20 million live streamers, we have something really cool for you. NVIDIA Broadcast turns any room into a broadcast studio. NVIDIA Broadcast runs AI algorithms trained by deep learning on NVIDIA's DGX supercomputer, one of the most powerful in the world. Effects like audio noise removal, virtual background effects, whether graphics or video, and webcam auto framing is a virtual camera person tracking you. These AI effects are amazing, available for download in September and runs on any RTX GPU.

Brandon and GeForce marketing will now show you NVIDIA Broadcast.

(Video Presentation)

A new form of art has emerged from gaming called Machinima. Artists are using game assets to create cinematics. There's been tens of billions of views on YouTube. Most are shorts. Some are even recreating entire classic movies. It's becoming a whole new art genre. Today, I'm going to show you an app that will make these cinematics amazing. It's called NVIDIA Omniverse Machinima. It's an app built on our Omniverse 3D workflow collaboration platform.

Omniverse is a universal design tool asset exchange with a viewer based on photorealistic path tracing. The engine is designed to be physically accurate, simulating light, physics, material and artificial intelligence. We have connectors for most third-party design tools like 3ds Max, Maya, Photoshop, Epic Unreal, Rhino and many more.

The Machinima app brings in elements and assets from games and third-party collections like TurboSquid, and lets you mix and compose them into a cinematic.

Creators can use their webcam to drive our AI-based pose estimator to animate characters, drive face animation AI with your voice, add high-fidelity physics like particles and fluids, make materials physically accurate, and then when done with your composition and mixing, render film quality cinematics with your RTX GPU. NVIDIA Omniverse Machinima, beta in October. Sign up at nvidia.com/machinima.

Let me show you a demo we created in a few days. We started with assets from Mount & Blade II: Bannerlord. You're going to love this.

(Video Presentation)

For 40 years since NVIDIA researcher, Turner Whitted, first published his paper on ray tracing, computer science researchers have chased this dream to create super-realistic virtual worlds with real-time ray tracing. NVIDIA, seeing the ultimate limits of rasterization approaching, focused intense efforts over the past 10 years to realize real-time ray tracing on a large scale.

At SIGGRAPH two years ago, we announced the NVIDIA RTX. Now two years later, it is clear we have reinvented computer graphics. NVIDIA RTX is a full stack invention. RTX starts with a brand-new GPU architecture, but it is so much more. It includes new engine tech and a bunch of new rendering algorithms. RTX is a home run. All major 3D APIs have been extended for RTX. RTX is supported by all major 3D tools.

RTX tech is incorporated into all major game engines. There are hundreds of games in development and thousands of research papers of new rendering and AI algorithms enabled by RTX. The RTX GPU has three fundamental processors. The programmable shader that we first introduced over 15 years ago, RT Core to accelerate the ray triangle and ray bounding box intersections and AI processing pipeline called Tensor Core.

Tensor Core accelerates linear algebra that is used for deep neural network processing, the foundation of modern AI. AI is the most powerful technology force of our time, computers that learn from data and write software that no humans can. The advances are nothing short of breathtaking. NVIDIA is doing groundbreaking work in this area. You might have seen our work in self-driving cars and robotics. Computer graphics and gaming will also be revolutionized by deep learning.

Let me show you some recent works in the art of the possible. The first video is a generative adversarial network that has learned to synthesize virtual characters of any artistic genre, including photo realistic. Second is a neural network that animates a 3D face directly from voice.

(Video Presentation)

The AI character can speak in any language, be any gender and even rap and sing. Third is a character locomotion of infinite number of positions. Imagine negotiating arbitrary paths and obstacles. The fourth is reconstructing 3D from video. Imagine the possibilities, record video, interact in 3D. This one is a deep learning model that learned the physics behavior of cloth animation.

Finally, this deep learning model of ray tracing can predict colors of missing pixels so that fewer rays need to be cast and fewer pixels need to be fully rendered. We can achieve orders of magnitude speedups. AI is starting to play a giant role in the future of computer graphics and gaming. The powerful Tensor Cores and RTX GPUs will let us do AI in real-time.

One of the first major AI computer graphics breakthroughs is DLSS. Here's the challenge. Real-time ray tracing is far more beautiful but requires a lot more computation per pixel than rasterization. So the solution is to ray trace fewer pixels and use AI on Tensor Cores to up res to super res, to a higher resolution and boost frame rate. DLSS took nearly two years of intensive research. We built a supercomputer to train the network.

The DLSS model is trained on extremely high-quality 16K offline rendered images of many kinds of content. Once trained, the model is downloaded into your driver. At runtime, DLSS 2.0 takes in low resolution alias image and motion vector of the current frame and the high resolution previous frame to generate a high resolution current frame. I think DLSS is one of our biggest breakthroughs in the last 10 years. Take a look at these images of Death Stranding, the latest game by Kojima San. DLSS is sharper than native 4K and create a detail from AI that native rendering didn't even show and the frame rate is higher.

Reviewers have loved DLSS 2.0. They say its quality beats out native rendering and runs even faster. You can play a 4K without a performance hit. Tensor Core effectively gives RTX a 2x performance boost. Let's look at one frame trace of a game to see the processors of RTX in action. Adding ray tracing to games dramatically increases the

computational workload. Using shaders to do rate reversal and object intersection reduces the frame rate.

We added the RT Core, which reduces shader workload by 60%. RT Core offloads to shaders by doing the ray triangle and ray bounding box intersection calculations. Using the same methodology as Microsoft Xbox, the RT Core is effectively a 34 teraflop shader. And Turing has an equivalent of 45 teraflops while ray tracing. Even with RT core, the amount of time consumed is significant. So, RT Core and shaders have to run concurrently. Even then, 20 milliseconds is only 50 frames per second and still a step back in performance relative to previous generations.

This is where the Tensor Core and DLSS come in. Rendering to a lower resolution, then using AI and super-fast Tensor Core to effectively double frame rate. Now you can get ray tracing, get high resolution and high frame rate at the same time. That's the magic of the three processors of RTX. Turing was our first-generation RTX GPU, combining ray tracing, programmable shading and AI. The flagship Turing had a ton of processing power; 11 shader teraflops, 34 RT teraflops and 89 Tensor teraflops.

Let me show you our new RTX GPU. Ampere is a giant leap in performance. Ampere does two shader calculations per clock versus one on Turing, 30 shader teraflops compared to 11. Ampere doubles ray triangle intersection throughput. Ampere's RT Core delivers 58 RT teraflops compared to Turing's 34. And Ampere's new Tensor Core automatically identifies and removes less important DNN weights. And the new Tensor Core hardware process the sparse network at twice the rate of Turing, 238 Tensor flops compared to 89.

Ladies and gentlemen, NVIDIA's new Ampere GPU, our second-generation RTX, 28 billion transistors built on Samsung 8N NVIDIA custom process. All three processors double rates over Turing, a triple-double. It connects to Micron's new G6X, the fastest memories ever made. The days of just relying on transistor performance scaling is over. Yet Ampere is an incredible two times the performance and energy efficiency of Turing.

At NVIDIA, we use every engineering lever to squeeze every drop of performance out of the system, from architecture, custom process design, circuit design, logic design, packaging, custom series I/O, memory, power and thermal design, PCB design, software and algorithms, thousands of engineers per generation, billions of dollars.

Full stack engineering and extreme craftsmanship is the hallmark of our GPUs. Our performance, energy efficiency and low power are all world-class. And real application performance highlights Ampere's new RT Core. The more ray tracing is done, the greater the Ampere speed up. Ampere RT Core doubles ray intersection processing. Its ray tracing is processed concurrently with shading, and Ampere can render cinematic images with motion blur eight times faster than Turing.

Let's take a look at Ampere in action. At our Kitchen GTC a few months ago, we showed Marbles, the world's first fully path-traced photorealistic real-time graphics. It was running on our highest end Turing Quadro RTX 8000. Turing was doing 720p, 25 frames per second. Today, we're going to run an enhanced version of Marbles with even more special effects and it is running at 1440p, 30 frames per second, over four times the performance.

Ladies and gentlemen, enjoy Marbles at night.

(Video Presentation)

Marbles is entirely path-traced. No rasterization, all real-time. There are hundreds of aerial lights, including spherical aerial lights. There is no pre-baking. Everything is dynamic. The depth of field is film quality and beautiful. Everything is dynamic. Diffuse GI, all dynamic. There are hundreds of rigid bodies. 80 million triangles. Materials are physically accurate, physics simulation and volumetric rendering in real time. DLSS 2.0 is doing the super-resolution and AI de-noising.

Let's compare Marbles Turing and Marbles Ampere. You could see dramatic visual quality jump of Ampere. Marbles on Turing runs at 720p, 25 frames per second. Marbles in Ampere runs at 1440p, 30 frames per second, more than four times the performance. And Ampere even did aerial lights and depth of field, a giant performance leap.

Today's games are giant worlds, indoor and out with photogrammetry, dense geometry and lots of characters. Games are over 200 gigabytes and getting bigger. This is like 50,000 songs or 400 hours of streaming video. Games have pushed PC I/O and file systems to the breaking point. CPUs copy files from disk and decompress the game image. This is fine when the storage system was slow, 50 megabytes to 100 megabytes per second. Now with Gen4 PCI Express and solid-state drives, PCs can transfer data at 7 gigabytes per second, 100 times faster. CPU copying data to memory and decompressing game images is now the bottleneck.

Decompressing data from 100 megabytes per second hard drives takes only a few CPU cores. However, decompressing from 7 gigabytes per second SSDs on PCIe Gen4 takes over 20 CPU cores. Today, we're announcing NVIDIA RTX IO with three new advances, new I/O APIs for fast loading and streaming directly from SSD to GPU memory, GPU losses decompression and collaboration with Microsoft on direct storage for Windows that streamlines the transfer of data from storage to GPU memory. With NVIDIA RTX IO, vast worlds will load instantly. Picking up where you left off will be instant. This is a very big deal for next-generation gaming.

Let me show you Ampere in action in one of the most anticipated games of 2020, CD PROJEKT Red's Cyberpunk. This trailer is called Scenes of Cyberpunk RTX. It shows ray trace reflections, diffused elimination, shadows, ambient occlusion and DLSS 2.0. Enjoy.

(Video Presentation)

Ladies and gentlemen, our new flagship GPU, the NVIDIA GeForce RTX 3080, powered by Ampere, second-generation RTX architecture.

(Video Presentation)

The NVIDIA RTX 3080. I have one right here. Let me show it to you. It is beautiful. Look at this, the RTX 3080. It is wonderfully crafted. It's going to look beautiful in your PC and it lights up.

Now, let me tell you about some of the other exciting technologies in sight. Turing use G6, the fastest memories at the time. The industry thought that was the limit. For Ampere, we had to push through that limit. Working with Micron, we designed the world's first memories with PAM4 signaling, pulse-amplitude modulation with four voltage levels that encode two bits of data each; 00, 01, 10, 11. Each voltage step is only 250 millivolts. So in the same period of time, G6X can transmit twice as much data as G6. PAM4 is extreme signaling technology, and it's just becoming used in high-speed networking.

The Ampere thermal architecture is the first-ever flow-through design, working harmoniously with PC chassis cooling system, pulling in cool air from the outside, flowing through the GPU and pushing hot air straight out the chassis. To allow room for a fan to flow air directly through the module, our engineers architected a super-dense PCB design that is 50% smaller than previous, while adding the bigger Ampere GPUs, HDMI 2.1, PCI Express 4.0 and G6X.

There are two independently controlled fans. The bracket front fan pulls cool air from the bottom and pushes the heated air out through the graphics card brackets. A backslide pull-through fan passes cool air over the fins of the heat pipe and directs the hot air to the top and back of the chassis to be exhausted by the system fan. The 3080 flow-through system is three times quieter and keeps the GPU 20 degrees cooler than the Turing design. It can cool 90 watts more than Turing.

The generational leap is ultimately the most important factor of new GPUs. A significant technology advance needed to inspire content developers to create the next level of content and for the installed base to upgrade. Let's see how the 3080 stacks up to previous generation architectures on the latest graphics-intensive games. 3080 is faster than 2080 Ti. 3080 is twice the performance of 2080 at the same price. Ampere is the biggest generational leap we've ever had.

Ladies and gentlemen, NVIDIA GeForce RTX 3080, our new flagship GPU, powered by Ampere, our second-generation RTX GPU architecture, incredible amounts of processing in the shader, RT ray tracing core and Tensor Core for processing AI, 10 gigabytes of G6X, twice the processing power of 2080 and at the same price, starting at $699, available September 17th.

One of our most popular GPUs is the 70 series. 970, 1070, 2070 were all hugely popular. You're going to love the new RTX 3070, faster than the 2080 Ti, the Turing enthusiast GPU priced at $1,200. Ladies and gentlemen, the new GeForce RTX 3070.

Let me show it to you. It's a work of art. 20 shader teraflops, 40 RT teraflops and 163 teraflops Tensor Core for AI processing. With 8 gigabytes of G6, RTX 3070 is faster than the $1,200 RTX 2080 Ti, starting at $499, available in October.

Every generation, we pack in our best ideas to increase performance, while introducing new features that enhance image quality. Every couple of generations, the stars align as it did with Pascal and we get a giant generational leap. Pascal was known as the perfect 10. Pascal was a huge success and set a very high bar.

It took the superfamily of Turing to meaningfully exceed Pascal on game performances without ray tracing. With ray tracing turned on, Pascal, using programmable shaders to compute ray triangle intersections, fell far behind Turing's RT Core. And Turing with ray tracing on reached the same performance as Pascal with ray tracing off. On a technical basis, this was a huge achievement. The images are far more beautiful and reflection in shadow artifacts are gone. But gamers wanted more.

They want every generation to be more realistic and higher frame rate at the same time. So, we doubled down on everything, twice the shader, twice the ray tracing and twice the Tensor Core, the triple-double. Ampere knocks the daylights out of Pascal on ray tracing and even with ray tracing on, crushes Pascal on frame rate. To all my Pascal gamer friends, it is safe to upgrade now.

Amazing ray tracing games are coming. Activision and developer, Treyarch, are launching a new Call of Duty on November 13. It's a masterpiece and it looks incredible. There are dynamic lights, ray tracing, shadows and ambient occlusion, DLSS 2.0 and NVIDIA Reflex super low latency technology. The last Call of Duty sold an amazing 30 million copies. Activision put together this trailer of never-before-seen footage.

Enjoy.

(Video Presentation)

Let me talk to you about one more thing. Several years ago, we started building the TITAN, pushing the GPU to the absolute limit to create the best graphics card of that generation. It was built in limited quantities and only through NVIDIA. The distribution was limited. The demand surprised us.

Creators were making 4K movies, rendering cinematics. Researchers built workstations for data science and AI. Bloggers built broadcast workstations. Flight

and racing simulation fans built sim rigs. There is clearly a need for a giant GPU that is available all over the world. So, we made a giant Ampere.

Ladies and gentlemen, the RTX 3090. 3090 is a beast, a ferocious GPU, a BF GPU, 36 shader teraflops, 69 RT teraflops, 285 Tensor teraflops and it comes with a massive 24 gigabytes of G6X. It comes with a silencer, a three slot dual-axle flow-through design, 10 times quieter and keeps the GPU 30 degrees cooler than the TITAN RTX design. But there is more.

The 3090 is so big that for the very first time, we can play games at 60 frames per second in 8K. This is insane. Because it's impossible for us to show you what it looks like on the stream, we invited some friends to check it out. Roll the clip.

(Video Presentation)

It's been 20 years since the NVIDIA GPU introduced programmable shading. The GPU revolutionized modern computer graphics. Developers jumped on and invented clever algorithms like shaders that simulate realistic materials or post-processing effects for soft shadows, ambient occlusion and reflections. Developers pushed the limits of rasterization beyond anyone's expectations.

Meanwhile, NVIDIA GPU processing increased a stunning 100,000-fold. Gaming became a powerful technology driver. Gamers grew to billions and gaming pushed into all aspects of entertainment and culture. If the last 20 years was amazing, the next 20 will seem nothing short of science fiction.

Today's Ampere launch is a giant step into the future. This is our greatest generational leap ever. The second-generation NVIDIA RTX, fusing programmable shading, ray tracing and artificial intelligence, gives us photorealistic graphics and the highest frame rates at the same time. Once the holy grail of computer graphics, ray tracing is now the standard. And Ampere is going to bring you joy beyond gaming.

NVIDIA Reflex to improve your response time. NVIDIA Broadcast turns any room into a studio. And Omniverse Machinima turns you into an animated filmmaker. We are super pleased with 3070, 3080 and 3090, the first three members of the Ampere generation. You're going to feel a boost like never before. I can't wait to go forward 20 years to see what RTX started. Homes will have holodecks. We will beam ourselves through time and space, traveling at the speed of light, sending photons, not atoms. In this future, GeForce is your holodeck, your lightspeed starship, your time machine. In this future, we will look back and realize that it started here.

Thank you for joining us today and to all of our fans for celebrating the arrival of Ampere.