Impact and Potential: Vizrt's Viz Engine and NVIDIA Ada Lovelace GPU Architecture

How real-time ray tracing, AI inference, and simulation are bringing realism and interactivity to 3D virtual sets, augmented reality graphics, and extended reality presentations.

A Joint Vizrt and NVIDIA White Paper

Gerhard Lang
Chief Technology Officer
Vizrt

Thomas True
NVIDIA Senior Applied Engineer
Professional Video and Image Processing Applications

Learn more at **vizrt.com**

GLOSSARY OF TERMS

AR - Augmented Reality Graphics

Augmented reality (AR) is the real-time use of information in the form of text, graphics, audio and other virtual enhancements integrated with real-world objects. It is this "real world" element that differentiates AR from virtual reality.

Global Illumination (GI)

Global Illumination (GI) is a group of algorithms used in 3D computer graphics that are meant to add more realistic lighting to 3D scenes.

Physically Based Rendering (PBR)

Physically based rendering (PBR) is a computer graphics approach that seeks to render images in a way that models the lights and surfaces with optics in the real world, with the aim to achieve photorealism.

Raster Graphics

Raster Graphics, which has its origin in television technology, is a collection of tiny, uniformly sized pixels, which are arranged in a two-dimensional grid made up of columns and rows to represent an image. Each pixel contains one or more bits of information, depending on the degree of detail in the image.

Ray Tracing

Ray tracing is a method of graphics rendering that simulates the physical behavior of light. NVIDIA has made real-time ray tracing possible with NVIDIA RTX.

Shaders

In computer graphics, a shader is a computer program that calculates the appropriate levels of light, darkness, and color during the rendering of a 3D scene—a process known as shading. Shaders have evolved to perform a variety of specialized functions in computer graphics special effects and video post-processing, as well as general-purpose computing on graphics processing units (GPU).

SM

SMs, or Streaming Multiprocessors, are the cores that execute the CUDA compute threads. The exact number of SMs available in a device depends on its NVIDIA processor family (Volta, Turing, Ampere, Ada), as well as the specific model reference of the processor.

Tensor Cores

Tensor Cores are hardware units introduced with NVIDIA RTX™ GPUs enabling mixed-precision computing, dynamically adapting calculations to accelerate throughput while preserving accuracy. The latest generation of NVIDIA Tensor Cores are faster than ever on a broader array of AI and high-performance computing (HPC) tasks.

XR - Extended Reality Graphics

Extended reality, or XR, is an umbrella category that covers a spectrum of newer, immersive technologies, including virtual reality, augmented reality and mixed reality.

NVIDIA RTX 6000 Ada Generation GPU

INTRODUCTION

The new NVIDIA Ada Lovelace GPU architecture raises the bar far above previous GPU generations to accelerate real-time ray tracing, AI inference, and simulation to bring incredible realism and interactivity to computer graphics applications. Vizrt's Viz Engine leverages the new features and performance of the NVIDIA RTX™ 6000 Ada Generation GPU to bring unparalleled levels of rendering performance and quality to 3D virtual sets, augmented reality graphics and extended reality presentations.

Vizrt is inextricably linked to the realm of 3D graphics. Three specific segments prioritize visual excellence and optimal performance – virtual sets, AR graphics, and XR sets.

A decade ago, Vizrt demonstrated some of the earliest XR sets utilizing NVIDIA GPUs and back projection, envisioning a future where this technology could potentially supplant virtual studios. With the advent of the Ada Lovelace architecture and the integration of high-quality video walls, the level of realism achievable is so remarkable that it becomes exceedingly challenging for viewers at home to discern whether the environment is genuine or virtual.

Vizrt relies upon NVIDIA's GPU hardware for rendering graphics and executing visual effects. However, an increasingly vital aspect for Vizrt is the capability to use convolutional neural networks (CNNs) for various tasks. Vizrt software now has the ability to leverage NVIDIA RTX at its full potential, using a mix of Vizrt custom AI modules and NVIDIA's pre-built AI SDKs for tasks such as pose estimations and super sampling. This allows Vizrt to enhance its capabilities in these areas by utilizing advanced AI technologies.

The NVIDIA Ada Lovelace GPU architecture takes its roots from the NVIDIA Turing™ GPU Architecture launched in 2018. The Turing GPU combined rasterization, real-time ray tracing, AI and simulation to enable incredible cinematic quality experiences in professional applications.

Two years later, the NVIDIA Ampere architecture incorporated more powerful ray tracing and tensor cores, along with a novel SM architecture that provided 2x the FP32 performance compared with Turing GPUs, for up to 1.7x increased performance in traditional raster graphics and up to 2x in ray tracing. Building upon the Ampere architecture, the NVIDIA Ada Lovelace GPU architecture is up to 2x faster in rasterized graphics applications and up to 4x faster in ray-traced applications.

GLOSSARY OF TERMS

AR - Augmented Reality Graphics

Augmented reality (AR) is the real-time use of information in the form of text, graphics, audio and other virtual enhancements integrated with real-world objects. It is this "real world" element that differentiates AR from virtual reality.

XR - Extended Reality Graphics

Extended reality, or XR, is an umbrella category that covers a spectrum of newer, immersive technologies, including virtual reality, augmented reality and mixed reality.

AI-based real-time pose estimation enables realistic talent reflections and shadow casting, produced by Reality Connect™, natively in Viz Engine 5.

GLOSSARY OF TERMS

Tensor Cores

Tensor Cores are hardware units introduced with NVIDIA RTX GPUs enabling mixed-precision computing, dynamically adapting calculations to accelerate throughput while preserving accuracy. The latest generation of NVIDIA Tensor Cores are faster than ever on a broader array of AI and high-performance computing (HPC) tasks.

Ray Tracing

Ray tracing is a method of graphics rendering that simulates the physical behavior of light. NVIDIA has made real-time ray tracing possible with NVIDIA RTX.

THE LINK BETWEEN TENSOR CORES AND AI

Tensor Cores are specialized high performance compute cores that are tailored for the matrix multiply and accumulate operations utilized for deep learning neural network training and inference. Compared to the previous Ampere GPUs, the Ada GPU Tensor Cores deliver more than double the Tensor TFLOPS performance for all types of calculations (FP16, BF16, TF32, INT8 and INT4). Ada GPUs also include the Hopper FP8 Transformer Engine that delivers over 1.3 PetaFLOPS of tensor processing in the RTX 6000.

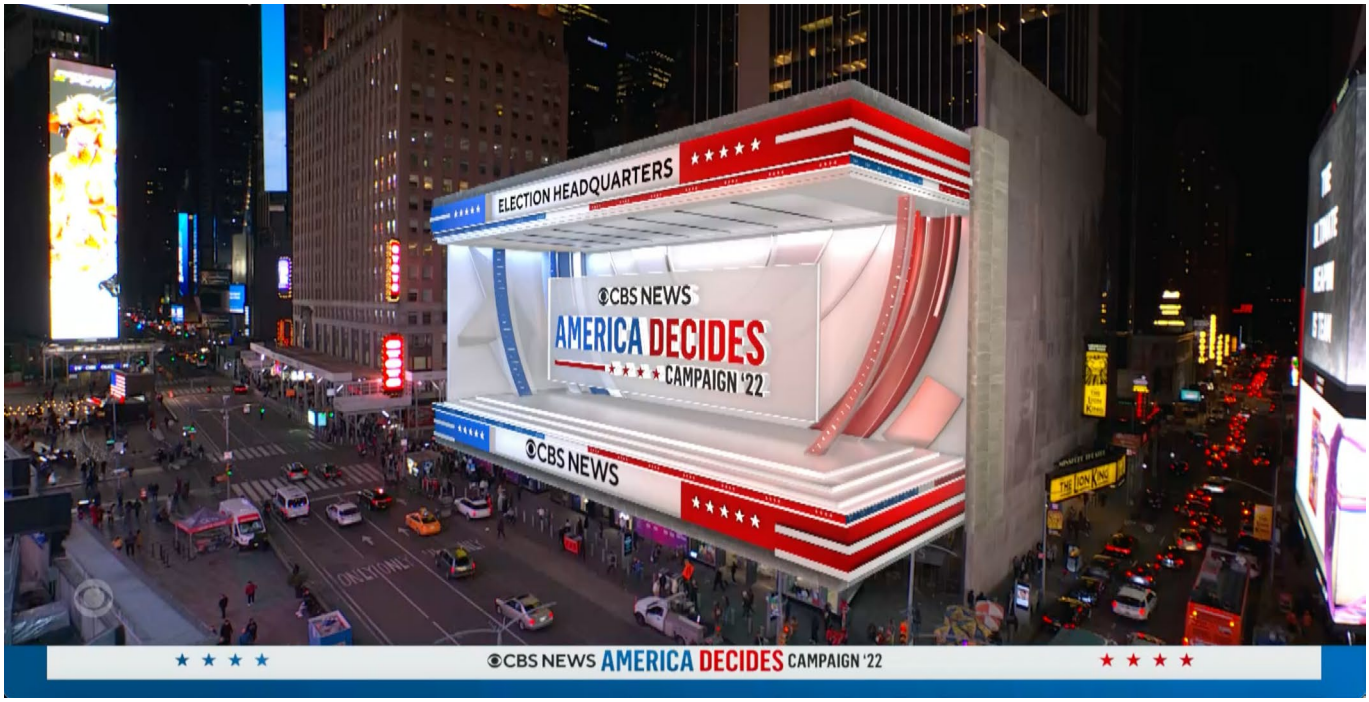
HOW DOES VIZ ENGINE USE TENSOR CORES?

To establish a convincing connection between real people and virtual worlds, it is crucial to incorporate realistic shadow casting and reflections of individuals within virtual studios and XR sets. Viz Engine 5.1's Reality Connect™ feature manifests these traits. By employing AI-based, real-time pose estimation and animated textured bone and skin models, these technologies enable realistic representations on the same hardware. Additionally, AI plays an inevitable role in keying, particularly in sports production. Furthermore, NVIDIA's DLSS (Deep Learning Super Sampling) enhances the overall image quality, providing a significant boost in visual fidelity.

However, AI is not solely employed for enhancing visual quality; its applications extend further. Vizrt places a significant emphasis on automation, and the integration of AI plays a vital role in achieving comprehensive automation. Unlike traditional algorithms, AI enables a more complete level of automation, allowing for advanced decision-making and intelligent processes that go beyond simple rule-based systems.

FASTER RAY TRACING

On previous NVIDIA GPU generations, although real-time ray traced effects could be achieved since the Turing architecture, fully physically correct lighting could not reach the real-time 60fps performance needed for virtual sets and broadcast graphics. As a result, real-time ray trace rendered scenes did not find their way into virtual set and broadcast graphics applications. To solve this problem, the Ada-generation RT Core has been enhanced to deliver 2x faster ray-triangle intersection testing and includes two new and important hardware units: 1) An Opacity Micromap Engine speeds up ray tracing of alpha-tested geometry by a factor of two, and 2) a Displaced Micro-Mesh Engine generates Displaced Micro-Triangles on-the-fly to create additional geometry. The Micro-Mesh Engine provides the benefit of increased geometric complexity without the traditional performance and storage costs of complex geometries.



An example of Viz Engine 5's physically-based rendering and global illumination techniques. *Image Courtesy CBS News.*

HOW DOES VIZ ENGINE TAKE ADVANTAGE OF GPU-BASED RAY TRACING?

Vizrt is developing Viz Engine to use ray tracing alongside physically-based rendering (PBR) and global illumination (GI) techniques for several reasons. Ray tracing provides a high level of realism and accuracy in rendering. It accurately simulates the behavior of light, including reflections, refractions, and shadows. This leads to more visually convincing and lifelike scenes, especially when rendering complex materials, intricate lighting scenarios, or highly reflective surfaces.

Ray tracing also allows for the accurate representation of various lighting scenarios. It can handle complex lighting setups, such as area lights, image-based lighting, and dynamic lighting, while accurately calculating light bounces and interactions. This flexibility enables the creation of visually rich and diverse scenes with realistic lighting effects, and it offers artists and designers greater control over the final look of their rendered scenes.

Ray tracing permits the precise manipulation of lighting parameters, material properties, and camera effects, thus empowering artists to achieve their desired aesthetics and bring their creative visions to life. Ray tracing enables the rendering of advanced effects that enhance visual quality and realism, and it can manage effects like caustics, subsurface scattering, volumetric lighting, and accurate ambient occlusion. All of these effects add depth, richness, and complexity to scenes, making them visually appealing and engaging.

Ray tracing is rapidly becoming an industry standard, with many software platforms adopting it as a core feature. By embracing ray tracing in Viz Engine, Vizrt future-proofs their software and ensures compatibility with evolving industry standards. It allows for seamless integration with other ray tracing-enabled workflows, such as content creation tools.

While PBR and GI techniques remain valuable in rendering pipelines, incorporating ray tracing into Viz Engine provides a powerful toolset for achieving unparalleled visual quality, accurate lighting, and realistic scene representation. Ultimately, ray tracing enhances the overall rendering capabilities of Viz Engine, and empowers artists and designers to create compelling and visually stunning content.

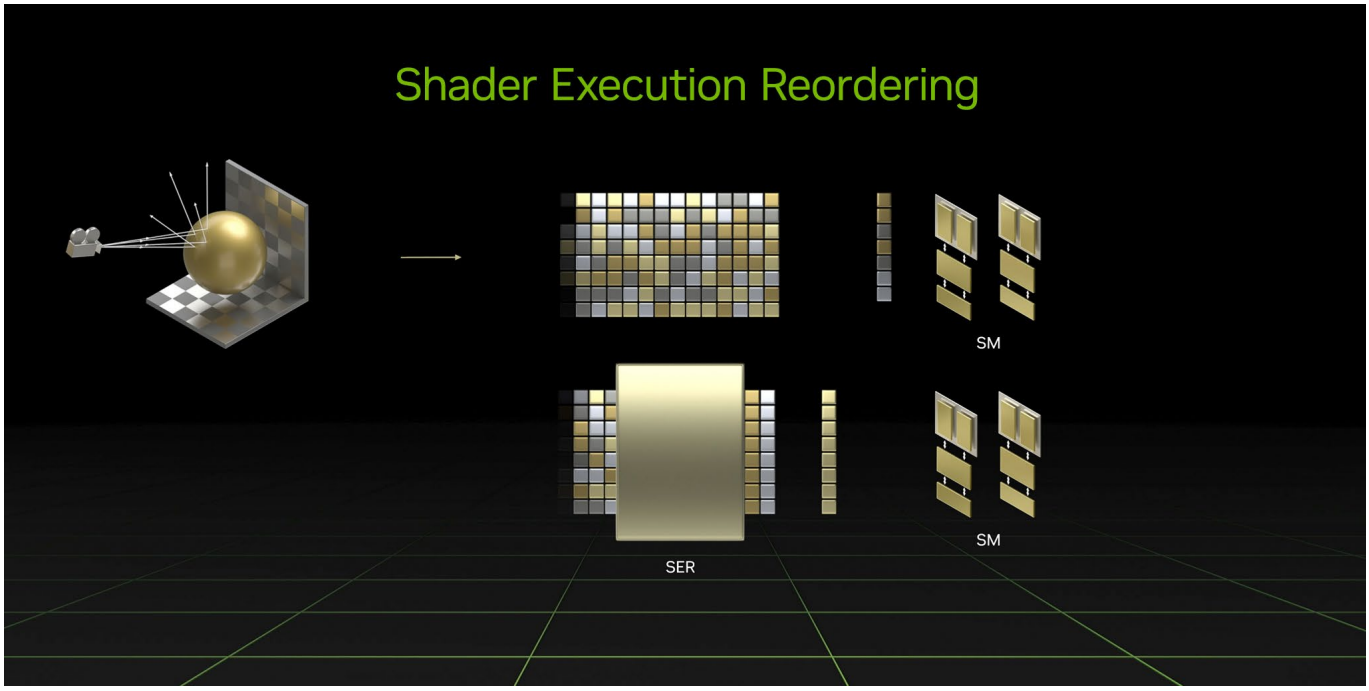
GLOSSARY OF TERMS

Physically Based Rendering (PBR)

Physically based rendering (PBR) is a computer graphics approach that seeks to render images in a way that models the lights and surfaces with optics in the real world, with the aim to achieve photorealism.

Global Illumination (GI)

Global Illumination (GI) is a group of algorithms used in 3D computer graphics that are meant to add more realistic lighting to 3D scenes.



SHADER EXECUTION REORDERING

Geometry, fragment and compute shaders contain the programmable instructions executed on the GPU for realtime rendering and visual effects. NVIDIA Ada GPUs support Shader Execution Reordering to dynamically and intelligently organize and reorder shading workloads to improve rendering performance. In a broadcast application, this translates to being able to render more complex geometry and effects within a broadcast frame.

Viz Engine often renders complex scenes with numerous shaders, each responsible for different visual aspects. By reordering shaders, Viz Engine can minimize the number of expensive state changes and texture accesses, optimizing the overall rendering performance. This improves frame rates and ensures smooth real-time rendering, especially when handling large-scale scenes or computationally demanding visual effects.

Reordering shaders can also help optimize GPU resource utilization. By grouping memory thrashing and efficiently utilizes available GPU memory and bandwidth. This ensures that the GPU is utilized to its fullest potential, allowing for faster and more efficient rendering. Additionally, shader reordering contributes to enhancing the visual quality of rendered scenes.

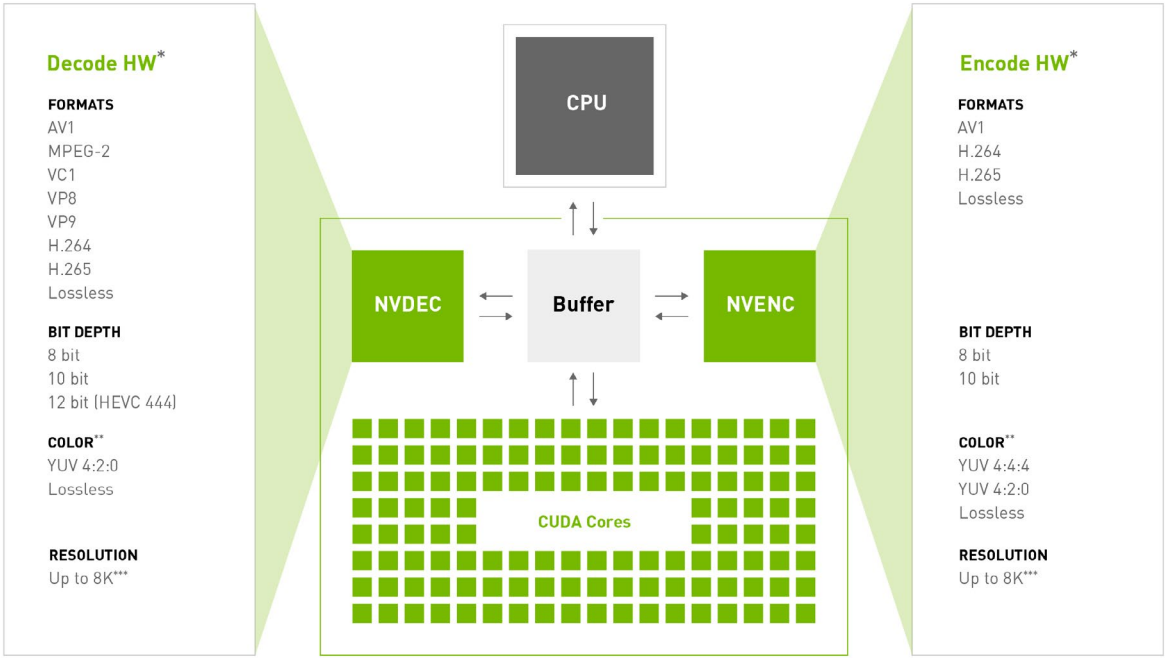
By strategically ordering shaders based on visual dependencies, Viz Engine can minimize artifacts and ensure that the correct shading calculations are performed. This leads to more accurate and visually appealing results, especially when dealing with complex materials or advanced rendering techniques.

In summary, shader reordering is crucial for Viz Engine to optimize performance, efficiently utilize GPU resources, enhance visual quality, and improve overall rendering efficiency.

GLOSSARY OF TERMS

Shaders

In computer graphics, a shader is a computer program that calculates the appropriate levels of light, darkness, and color during the rendering of a 3D scene—a process known as shading. Shaders have evolved to perform a variety of specialized functions in computer graphics special effects and video post-processing, as well as general-purpose computing on graphics processing units (GPU).



OPTICAL FLOW ACCELERATION

Frame generation techniques that combine optical flow estimation with Deep Learning Super Sampling (DLSS) can be utilized to insert accurately synthesized frames between existing video frames to improve frame rate and provide smoother motion. The Optical Flow Acceleration (OFA) unit in Ada GPUs, combined with new motion vector analysis algorithms, enable more accurate and performant frame generation capabilities. Furthermore, these motion vector computations happen without taxing the GPU SMs so they may be simultaneously utilized to render more pixels and more simulated effects.

Vizrt's sports-oriented AI Keyer utilizes optical flow analysis, and with the utilization of NVIDIA's Optical Flow Accelerator (OFA), Vizrt enhances the performance of its AI keyer. Optical flow data can be utilized to generate a precise matte for the keying operation. By understanding the motion between frames, Viz Engine can estimate the motion blur and subtle pixel shifts caused by subject movement.

This information assists in generating a more accurate and detailed matte, resulting in better foreground extraction and cleaner keying results. Optical flow analysis also aids in refining the edges of the keyed subject. By detecting the motion discontinuities at the edges, Viz Engine can apply edge-aware algorithms to improve the sharpness and accuracy of the subject's boundaries.

DEDICATED HARDWARE VIDEO TRANSCODING

NVIDIA's eighth generation dedicated hardware encoder (NVENC) in the Ada GPUs adds support for AV1 encoding providing 40% bitrate gains over x264 encoding. To further aid in encoding performance, the RTX 6000 is equipped with three NVENC encoders. This enables video encoding a single 8K/60 stream or four 4K/60 streams for professional applications.

In addition to NVENC, the RTX 6000 Ada also includes three fifth-generation hardware decoders (NVDEC). NVDEC supports hardware-accelerated video decoding of MPEG-2, VC-1, H.264 (AVCHD), H.265 (HEVC), VP8, VP9, and the AV1 video formats. 8K60 decoding is also fully supported.

GLOSSARY OF TERMS

SM

SMs, or Streaming Multi-processors, are the cores that execute the CUDA compute threads. The exact number of SMs available in a device depends on its NVIDIA processor family (Volta, Turing, Ampere, Ada), as well as the specific model reference of the processor.

HOW DOES VIZRT LEVERAGE DEDICATED HARDWARE VIDEO TRANSCODING?

Viz Engine leverages the dedicated hardware encoding capabilities of the RTX 6000 GPU to perform real-time video encoding, ensuring efficient processing of video streams while reducing CPU load. This results in smooth and high-quality video output, especially for formats like NDI HX and WebRTC.

Additionally, Viz Engine utilizes the GPU's capabilities for multi-channel encoding, enabling simultaneous encoding of multiple video streams. This is particularly useful for interactive previews and driving video walls. Furthermore, hardware acceleration for decoding is essential for seamlessly playing back large-format clips in real time.

As long term partners, Vizrt and NVIDIA are collaborating on new techniques in ray tracing, AI, and simulation bringing incredible realism and interactivity to computer graphics and driving innovation across the industry.

CONCLUSION

As long term partners, Vizrt and NVIDIA are collaborating on new techniques in ray tracing, AI, and simulation bringing incredible realism and interactivity to computer graphics and driving innovation across the industry. In their combined quest to help content creators achieve photorealism in a virtual environment, there are certain product development 'wins' which appear to be on the horizon.

For example, new approaches to ray tracing, which is a method of graphics rendering that simulates the physical behavior of light, promise to greatly improve the visual acuity and sophistication of talent shadow casting and reflection. And, as already mentioned in this paper, improved DLSS techniques will permit the future exploration of how to best utilize upscaled frames to reveal new XR use cases.

Beyond this, content creators who make creative use of Vizrt's new Reality Connect feature, which is currently exclusive to the Viz Engine Render Pipeline, will soon see this new creative wizardry extended to the Unreal Engine Pipeline.

And of course, users can expect the rush of AI development to continue unabated, accelerating both the pace of new product and feature introductions, leading to new forms of real-time 3D virtual sets, augmented reality graphics, and extended reality presentations.

For more information on the NVIDIA Ada Lovelace GPU Architecture, please read **Nvidia-ada-gpu-architecture.pdf**

For more information about Vizrt's Virtual Set, AR graphics, and XR Set solutions, please visit **vizrt.com**

About Vizrt

Vizrt® is the world's leading provider of innovative visual storytelling tools for media content creators in broadcast, enterprise, or new media – unlocking the power of a story for all.

Vizrt offers market-defining software-based solutions for real-time 3D graphics, video playout, studio automation, sports analysis, media asset management, and journalist story tools.

Vizrt offers Flexible Access to our workflows, and our platforms integrate with third-party products because we believe in enabling our customers' success, giving them the right tool for the job, and accelerating their creative excellence.

More than three billion people watch stories told by Vizrt customers every day including from media companies such as CNN, CBS, NBC, Fox, BBC, Sky Group, Al Jazeera, NDR, ZDF, Network 18, Tencent, and many more.

About NVIDIA

Since its founding in 1993, NVIDIA has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at **https://nvidianews.nvidia.com/.**