

**High Performance Parallel Machine Learning and ai**

**Under Prof. Handan Liu**



**Nvidia Conference Summary Report**

**GTC September 2022 Keynote with Nvidia CEO Jensen Huang**

**Report By:  
Nikita Gaurihar - 002980962**

**Nvidia Conference 2022**

**GTC September 2022 Keynote with Nvidia CEO Jensen Huang**

**1. Introduction**:

Jensen Huang, CEO of Nvidia, addressed the audience as the keynote speaker for this session which was held in September 2022. He talks about how computing is advancing at an incredible speed. He also gave the analogy of the rocket engine. He mentions that the engine propelling this rocket is accelerated computing and its fuel is AI.

Further, he mentioned that they have launched a few new advances in their cloud services, such as, Nvidia RTX , Nvidia AI, Nvidia Omniverse. He claimed that these technological advancements will propel new breakthroughs in AI applications and the next wave of AI for science and industry. They plan to announce new chips that power these applications, thereby claiming that cloud services are expanding their horizons of the Nvidia platforms for the very first time. The detailed information is as follows -

**1.1 RacerX**

RacerX is a fully interactive simulation built with the help of Nvidia Omniverse services. It is physically simulated. With the help of this technology, the lighting , reflections, and Refractions are ray traced where nothing is manipulated. The future games will not have a pre-baked world, meaning it will be fully simulated. All the credit goes to RacerX Nvidia omniverse services which is running on a single GPU.

**1.2 RTX – Ada Lovelace**

He further explained that the 3D graphics were revolutionized by the GPUs. It created a medium of an infinite palette for artists. They launched a new GPU architecture called ‘Nvidia RTX’ back in 2018, at Siggraph. With help of RTX, Computer Scientists got flooded with a pool of amazing opportunities and ideas about new algorithms and gave birth to a new era of RTX neural rendering. With this information, he introduced Ada Lovelace , Nvidia’s 3rd generation of RTX and it was named after great mathematician Ada Lovelace who is world’s first computer programmer. One Ada GPU powers RacerX. This process allows us to integrate 76 billion transistors and over 18000 CUDA cores, 70% more than the Ampere generation. Ada evolved all the three RTX processors – **Shaders, Ray Tracing, Deep Learning.** GPUs are highly parallel and most efficient when processing similar work at the same time. These advancements will give up to 2-3times increase in ray tracing and 25% in overall game performance. RTXDI is used for emissive surfaces such as billboards, TV Screens, and neon tubes. Ada is up to 2 times faster for rasterized games and 4 times faster for ray-traced games.

**1.3 Nvidia Deep Learning: DLSS2**

DLSS uses a convolutional autoencoder AI Model and takes the low-resolution current frame and the high-resolution previous frame to predict on a pixel-by-pixel basis, a higher resolution current frame. AI Model is trained to predict an ultra-high resolution 16K reference image. DLSS 3 has four components, a new optical flow accelerator, game engine motion vectors, a convolutional autoencoder AI Frame generator and our reflex super low latency pipeline.

**1.4 Portal with RTX**

It is a mod, used majorly in gaming technologies. He claimed that the 9/10 competitive games owe their existence to mods. It is the most advanced game modeling tool ever created.Deep learning models are used through the AI assisted toolset to improvise the texture and assets with physically accurate properties.

1. **Metaverse - Omniverse JT Connecter**

Metaverse, the 3D internet, connects virtual 3D worlds described in USD and are viewed through a simulation engine. Omniverse is the platform used for building and operating metaverse platforms. Omniverse expands when physical and digital applications meet. It is a large-scale real-time 3D database which acts as a 3-D shared world. This is used as a portal to the virtual worlds. It is the world’s first sim ready asset library that is used for synthetic data generation using Replicator and digital Twin simulations to train self-driving cars, robots, and all kinds of computer vision models.

**2.1 Nvidia Omniverse Cloud Computer: Infrastructure-as-a-service**

Omniverse is a virtual world used to design , build, and operate things in the physical world, meaning, in future, everything made in the physical world will have a digital twin. He further explained this using many great examples of how companies use Omniverse to create digital twins of factories , logistics warehouses, industrial plants, etc. Omniverse is a new computing platform and requires a new computer system. There are 3 elements to omniverse –

* RTX computers for creators, designers, and engineers
* OVX servers to host connections to the database and run the virtual world simulations.
* Nvidia GDN – Portals to the Omniverse.

1. **Nvidia** **Drive Sim**

With the breakthrough in AI using Omniverse, Nvidia Drive simulations is the next big advancement which works on the concept of hardware in the loop. This will enable the cars with physical and digital cabin designs. Act as a virtual design studio. Drive Map with autonomous fleet mapping will also be prevalent in these models, with real time radar simulations.

1. **Nvidia Drive Orin:**

He explained about this product by giving examples of Robots. He says, there are robots that move and there are robotic systems that watch things that move. Orin is the robotic processor of industrial grade IGX Edge AI and computing platforms that Metropolis runs on.

OV Connectors like Drive Sim, Isaac ROS and Sim, Sim Ready Assets and Replicator, and Clara Holoscan bring up new vistas for accelerated computing. These AI models are implemented in several technological applications, including self-driving cars, chatbots for customer care, and e-commerce websites, along with workflows that adapt the AIs to use-cases and aid the models in learning from experience.

1. **Accelerating Computing across the stack:**

Scaling across multi-GPUs and multi-nodes is a datacenter scale challenge. They plan to enhance the performance and throughput of their fleet of computers with each new update by extending accelerating computing into new opportunities. These days, the two most significant AI models are recommender systems and large language models, and each has a different set of processing needs. Almost all the companies in various industries now use graph databases to store information and relationship between their data on nodes and edges.

1. **Nvidia RAPIDS**:

For accelerated Data science and Analytics, RAPIDS are used. Data scientists and Developers leverage RAPIDS through WSL to process data frames, SQL Array, ML models and Graph analytics. These computational high powers speed up processing techniques by 4 times and help to reduce costs by 70%.

1. **CV CUDA:**

This is an open source, GPU accelerated library for imaging and computer vision. The infrastructure-as-a-service known as NVIDIA Omniverse Cloud links Omniverse applications that are running on devices, on premises, or in the cloud. Avatars utilize cloud-scale, real-time computer vision, voice AI, language comprehension, and animation software. This is made possible through acceleration libraries like CV-CUDA, the Omniverse ACE Avatar Cloud Engine, the UCF Unified Computing Framework, and a prototype application for customer service avatars named Tokkio.

Nvidia claims that these systems will be available in the market by 2023. They plan to grow and expand their reach through these advancements. They are proud to believe that being the world’s largest professional services firm, they are bringing new services built on Nvidia AI AND Nvidia omniverse to the world’s enterprises.

**References**

<https://www.nvidia.com/gtc/keynote>