#GTC2024
March 17th - 21st

WHITESIGHT

Level Up Your AI Game:
#9 Insights from Nvidia GTC 2024

NVIDIA

GTC GPU TECHNOLOGY CONFERENCE

www.whitesight.net

Swipe to know >>>

#GTC2024
March 17th - 21st

WHITESIGHT

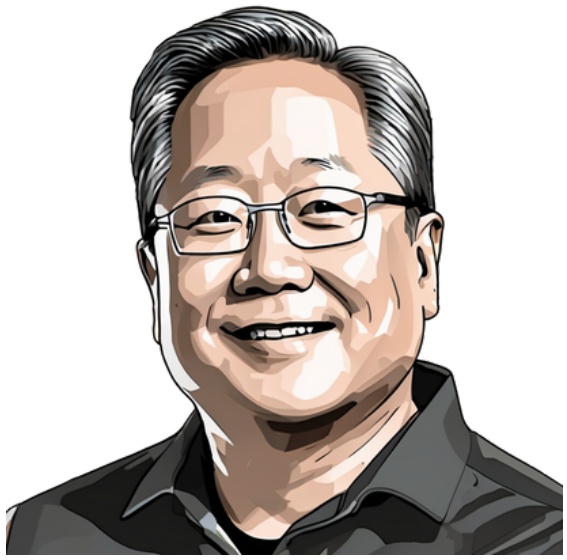
What: The Nvidia GTC, or GPU Technology Conference, is a biannual gathering for experts and enthusiasts in AI, computer graphics, data science, and related fields to explore the latest technological advancements.

Where: San Jose Convention Center, California, USA

When: March 17th - 21st, 2024

Topics:

- Accelerated computing tools & techniques
- AI models and deployment
- AR/VR
- Computer vision/video analytics
- Content creation/rendering/ray tracing
- Conversational AI/NLP
- Cybersecurity
- Data center/cloud
- Data science
- Edge computing
- Generative AI
- Networking
- Recommenders/personalisation
- Robotics
- Simulation/modeling/design
- Video streaming/conference



Jensen Huang

"The future is generative, which is the reason they call it generative AI, which is the reason why this is a brand new industry. The way we compute is fundamentally different."

#1 Blackwell : Nvidia's New GPU Platform

Blackwell GPU: A next-generation AI accelerator offering significant performance improvements:

- Up to 30x faster inference throughput and 4x faster training compared to its predecessor, the H100 GPU.
- Significantly less power consumption.

Key advancements: A second-generation transformer engine and a high-speed decompression engine.

GB200 Superchip: A powerful AI processing unit combining:

- 2 Blackwell GPUs for exceptional AI processing capabilities.
- A single Grace CPU for system management and control.

NVL72 Rack: A high-performance AI workstation featuring:

- 72x Blackwell GPUs enabling massive parallel processing.
- 36x Grace CPUs for efficient system coordination.
- Capable of training AI models with up to 27T parameters.

#2

To NIMfinity and Beyond: Nvidia's Turbocharged AI Deployment!

Streamlined Generative AI Deployment: Nvidia's Inference Microservices (NIM) concept simplifies the deployment process for GenAI applications, facilitating secure, stable, and scalable implementations within the NVIDIA AI Enterprise suite.

Pre-Packaged Solution: NIM offers pre-integrating essential components for deployment, including pre-built AI models, integration code, and a pre-configured Kubernetes Helm chart.

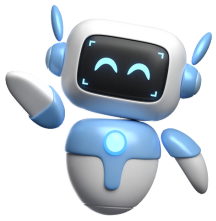
Enhanced Developer Efficiency: By streamlining deployment, NIM allows developers to focus on core tasks like AI model customisation and innovative application development.

Enterprise-Grade Security and Scalability: NIM ensures secure and scalable deployments, enabling businesses to confidently launch generative AI applications in production environments.

#3

I am GR00T: The Genesis of Nvidia's Robot Revolution

Project GROOT (Generalist Robot 00 Technology): A foundation model designed specifically to enhance the capabilities of humanoid robots.



The project aims to enable robots with two 2 functionalities:

- **Natural Language Processing:** GROOT equips robots with the ability to comprehend spoken language instructions, fostering a more intuitive human-robot interaction.
- **Observational Learning:** By observing human actions and movements, robots powered by GROOT can acquire new skills and progressively improve their physical dexterity.

#4 RAN Rampage: Nvidia's 6G Gold Rush

6G Research Platform: The platform leverages AI to accelerate advancements in radio access network (RAN) technology.

Applications: This platform aims to significantly shorten the development timeline for 6G technologies. The widespread adoption of 6G, facilitated by Nvidia's research platform, will enable the development of:

- Autonomous vehicles
- Smart spaces in home and workplaces
- AR,VR, and XR
- Collaborative robots

#5

From Asgard to Asphalt: Nvidia's Thor Powers Electric Valkyries

Drive Thor SoC Gains Traction in Electric Vehicles:

A Drive Thor centralised computer is a purpose-built AI platform designed specifically for autonomous vehicles.

Applications:

- Advanced performance.
- Cost Optimisation
- Combination of autonomous driving, in-cabin AI, and infotainment systems on a single platform

Potential Partners:

- BYD
- Hyper (under GAC Aion)
- Xpeng

#6 Cloud Connections: NVIDIA's AI Odyssey with Tech Titans

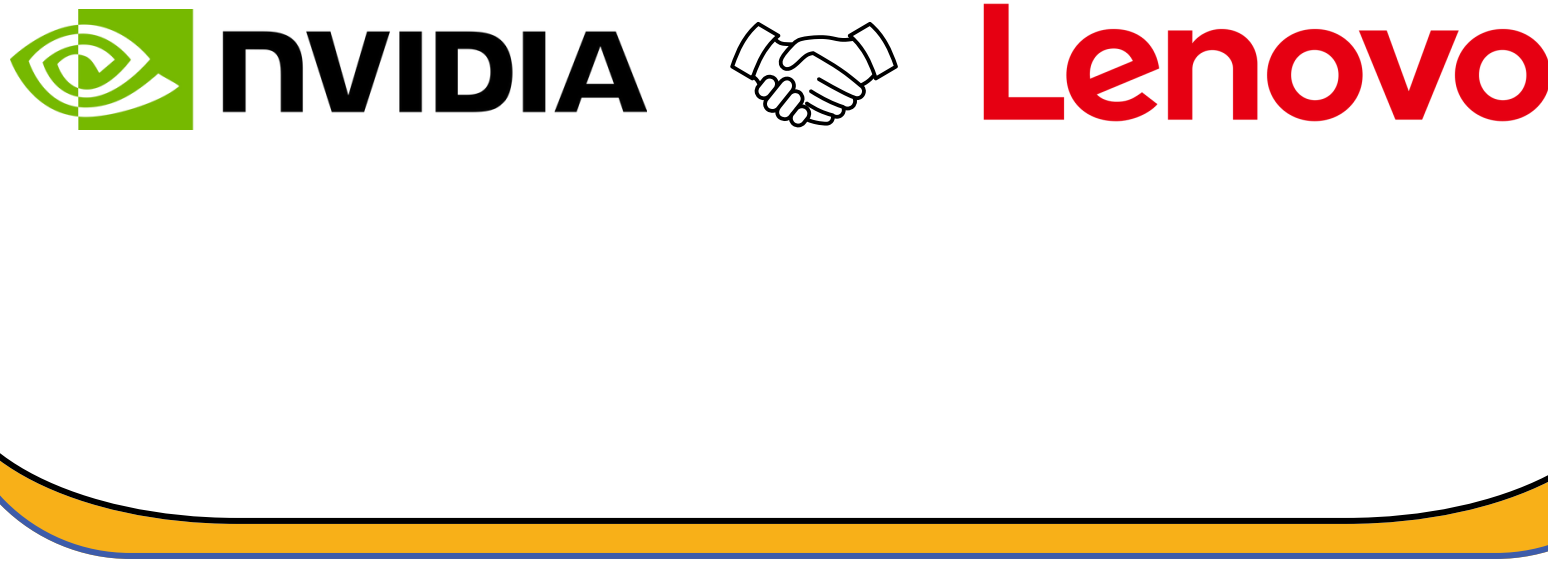
NVIDIA

Azure Google Cloud aws ORACLE

NVIDIA NIM is set to debut on **Azure AI, Google Cloud,** and **Oracle Cloud**, while more specific initiatives include:

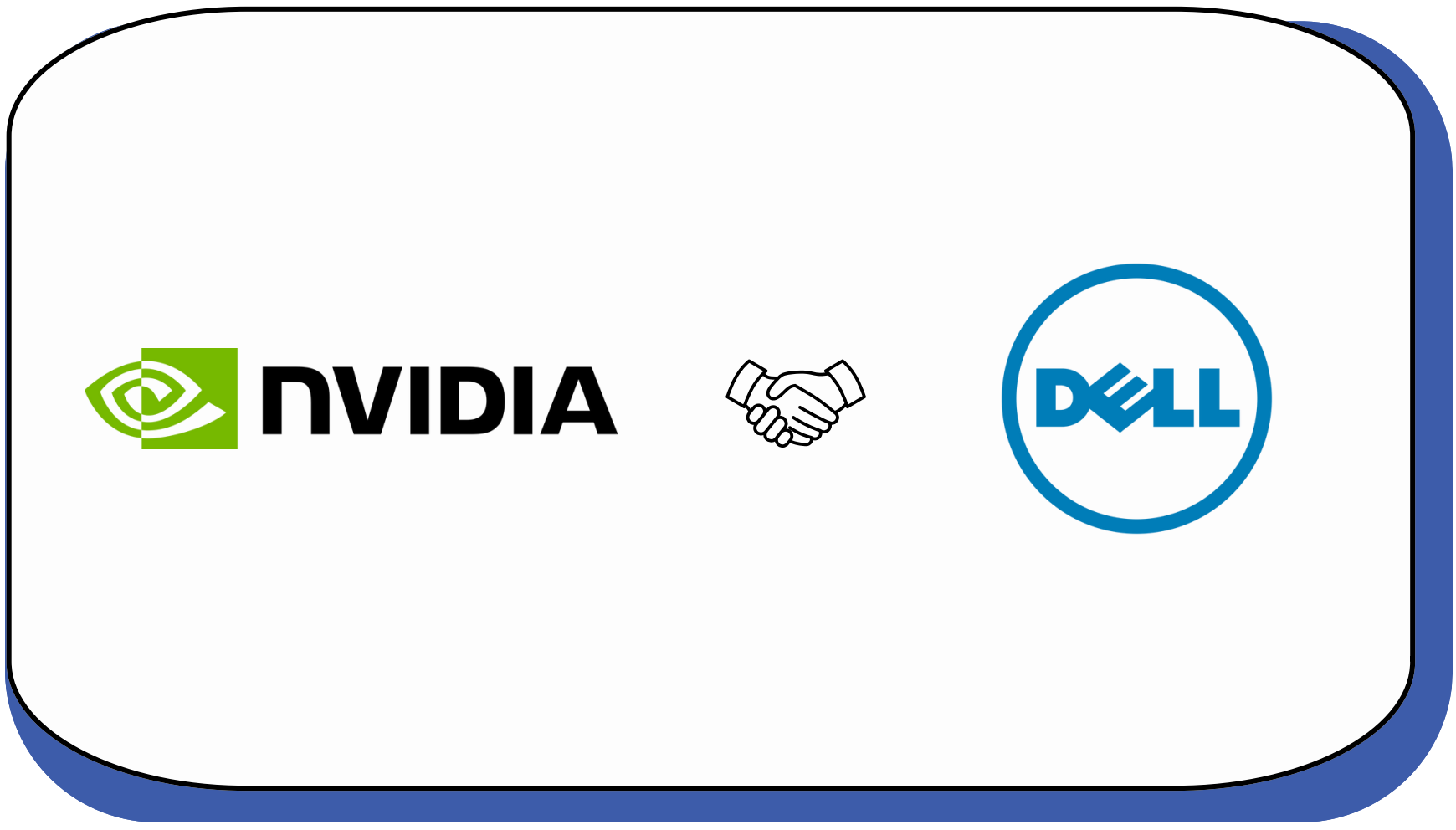
- **Amazon SageMaker** will integrate with NIM to optimise the price performance of foundation models running on GPUs, with more collaboration on healthcare.
- **Google** will adopt the Blackwell platform and the NVIDIA DGX Cloud service, boosting its generative AI capabilities.
- **Google** and **AWS** will work together on optimising open models like Gemma, supporting JAX on Nvidia GPUs, and using NIM inference microservices to give developers a flexible, open platform for training and deploying AI models.

#7 GenAIus Solutions: Lenovo's Nvidia-Powered Workstations & Servers



- Lenovo introduced AI servers and workstations powered by Nvidia technology to enhance GenAI capabilities for businesses and developers.
- The high-performance servers feature Nvidia GPUs, equipped for AI tasks and graphics-intensive workloads.

#8 AI Factory: Shaping Tomorrow with Dell



- Dell introduced the **Dell AI Factory**, developed jointly with NVIDIA to ensure data security and governance.
- The solution incorporates PowerEdge XE9680 servers that support the latest Nvidia GPUs.

#9 Potion Partners: Cognizant and Nvidia's GenAI Elixir

- Cognizant partnered with Nvidia to **advance drug discovery** for pharmaceutical clients using GenAI.
- The partnership combines **Nvidia's BioNeMo platform** with Cognizant's life sciences expertise.

WHITESIGHT

Discovering value in our content? [like & comment]

Found it helpful? Spread the knowledge! [repost]

Want a regular dose of this goodness? [subscribe]

www.whitesight.net