What a keynote! At GTC 2025, NVIDIA CEO Jensen announced several advancements in accelerated computing, inference leadership, enterprise AI and infrastructure, and robotics and physical AI. Below is a summary of each announcement.

Key Messages for Customer Discussions:

- Shift to accelerated computing is accelerating; $1T worth of data centers becoming accelerated and driving demand for AI factories.
- AI investment is accelerating across every industry - beyond CSPs, to GPU clouds, enterprise, and robotics.
- AI reasoning inference requires significantly more computation than traditional one-shot inference. NVIDIA is the leader in inference with full-stack invention and optimization.
- AI is now mainstream - and the entire Enterprise stack has been reinvented with NVIDIA acceleration.
- Robots have arrived - a new wave of physical AI drives major data center workload consumption with simulation and synthetic data generation for post-training.

Accelerated Computing Announcements:

- NVIDIA Blackwell Ultra
- NVIDIA GB300 NVL72
- NVIDIA HGX B300 NVL16
- NVIDIA Vera Rubin and NVIDIA Vera CPU
- NVIDIA Photonics / Co-Packaged Optics
- NVIDIA Mission Control (Previously Advanced Cluster Management)
- NVIDIA DGX GB300
- NVIDIA DGX B300
- NVIDIA Omniverse Blueprint for AI Factory Design and Operations
- NVIDIA Dynamo
- Computer-Aided Engineering (CAE/EDA)
- Earth-2 - NVIDIA Omniverse Blueprint for Earth-2 Weather Analytics
- CUDA-X Libraries
- NVIDIA Accelerated Quantum Research Center (NVAQC)

Agentic/Enterprise Announcements:

- NVIDIA DGX Spark
- NVIDIA DGX Station
- NVIDIA RTX PRO
- NVIDIA DGX Cloud
- Llama Nemotron Models with Reasoning
- NVIDIA NeMo Microservices
- NVIDIA NeMo Retriever and NVIDIA AI Blueprint for RAG
- NVIDIA AI-Q Blueprint
- NVIDIA-Certified Storage
- NVIDIA AI Data Platform for Enterprise
- Leading Partners Are Ramping the Adoption of NVIDIA AI Enterprise
- NVIDIA Riva
- NVIDIA vGPU

Physical AI, Omniverse, Robotics, Metropolis Announcements:

- NVIDIA Omniverse
- NVIDIA Cosmos
- Open Source Dataset for Physical AI
- NVIDIA AI Blueprint for Video Search and Summarization (VSS)
- NVIDIA Isaac GR00T
- Newton Open Source Physics Engine

Industry Announcements:

- Healthcare and Life Sciences
- Telecommunications
- Energy
- Automotive

NVIDIA Blackwell Ultra

Blackwell Ultra is our newest accelerated computing platform built for the 3 scaling laws and the age of AI inference reasoning. Blackwell Ultra offers **50x** more AI factory output over Hopper and is available in two system configurations: GB300 NVL72, HGX B300 NVL16. Blackwell Ultra GPUs introduce several key technological breakthroughs:

- **AI Reasoning Inference:** Features **1.5x** more AI FLOPS compared to Blackwell GPUs and **2x** attention acceleration for long-context thinking.
- **More Memory**: Features **1.5x** more memory compared to Blackwell GPUs for up to 288 GB of HBM3e per GPU enabling more inference performance.
- **Faster Networking**: Features **800 Gb/s** of network connectivity for each GPU with NVIDIA ConnectX-8 SuperNIC with either NVIDIA Quantum-X800 InfiniBand or Spectrum-X Ethernet for multi-node workloads.

Pricing: Pricing will be disclosed from partners closer to Blackwell Ultra general availability. Contact OEM partners for Blackwell Ultra pricing.

Availability: Blackwell Ultra products will be available in the second half of 2025 from all major cloud service providers and server makers.

External Resources: Scaling Laws Blog, GB300 NVL72 Webpage, HGX Webpage, Blackwell Ultra Press Release

---

NVIDIA GB300 NVL72

The NVIDIA GB300 NVL72 features a fully liquid-cooled, rack-scale design that unifies 72 NVIDIA Blackwell Ultra GPUs and 36 Arm-based NVIDIA Grace CPUs in a single platform optimized for AI training and test-time scaling inference. Blackwell Ultra delivers breakthrough performance on the most complex workloads, from agentic systems and AI reasoning to real-time video generation for AI factories around the world.

- **NVIDIA GB300 Grace Blackwell Ultra Superchip**: Key building block for the NVIDIA GB300 NVL72 rack-scale solution. It features four NVIDIA Blackwell Ultra GPUs, two Grace CPUs, and four ConnectX-8 SuperNICs. Through NVIDIA NVLink Switch technology, 18 superchips combine into one giant GPU, GB300 NVL72, purpose-built for the age of AI reasoning.
- **Increased AI factory output:** GB300 NVL72 integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs, offering a 50X increase in AI factory output with optimized inference capabilities.
- **Specifications:** Relative to Hopper, GB300 NVL72 has 70X FP4 dense inference FLOPS, 30X GPU memory, 65X fast GPU + CPU memory, 20X networking bandwidth.

Pricing: Pricing will be disclosed from partners closer to Blackwell Ultra general availability. Contact OEM partners for Blackwell Ultra pricing.

Availability: GB300 NVL72 will be available in the second half of 2025 from all major cloud service providers and server makers.

External Resources: Scaling Laws Blog, GB300 NVL72 Webpage, Blackwell Ultra Press Release

---

NVIDIA HGX B300 NVL16

A Blackwell Ultra x86 platform connecting 16 Blackwell Ultra GPUs in a single NVLink domain. Relative to Hopper, HGX B300 NVL16 will deliver:

- **Performance:** 11X real-time inference and 4X training performance for Llama 3.1 405B
- **Specifications:** 7X FP4 dense inference FLOPS and 4X GPU memory

Pricing: Pricing will be disclosed from partners closer to Blackwell Ultra general availability. Contact OEM partners for Blackwell pricing.

Availability: HGX B300 NVL16 will be available in the second half of 2025 from all major cloud service providers and server makers.

External Resources: Scaling Laws Blog, HGX Webpage, Blackwell Ultra Press Release

---

NVIDIA Vera Rubin and NVIDIA Vera CPU

Next-generation accelerated computing platform shaping the future of AI

- **NVIDIA Rubin GPU:** The NVIDIA Rubin GPU is our next-generation data center GPU featuring: 2 reticle-sized dies, 50 petaFLOPS of FP4, 288 GB of HBM4 memory
- **NVIDIA Rubin Ultra GPU:** Next-generation of the Rubin platform. Rubin Ultra will feature: 4 reticle-sized dies, 100 petaFLOPS of FP4, 1TB of HBM4e memory
- **NVIDIA Vera CPU:** The NVIDIA Vera CPU is our next-generation data center CPU featuring 88 NVIDIA-designed, high-performance Olympus CPU cores, it delivers up to 1.2 terabytes per second (TB/s) of memory bandwidth while using only 50 watts of memory power. NVIDIA Scalable Coherency Fabric (SCF) maximizes performance and keeps data flowing. And with over 2x the performance of the prior generation, the NVIDIA Vera CPU is ideal for data processing, compute and memory-intensive workloads or pairing with the NVIDIA Rubin GPU to shape the future of high-performance computing (HPC) and AI.
- **NVIDIA Vera Rubin NVL144:** Next generation NVLink liquid cooled rack-scale architecture connecting Arm-based NVIDIA Vera CPUs and 144 Rubin GPUs in a single NVLink domain. Compared to GB300 NVL72, Vera Rubin NVL144 features: 2.5X FP4 inference FLOPS, 3.3X FP8 training FLOPS, 1.6X HBM memory bandwidth, 2X fast

memory (GPU + CPU), 2X all-to-all NVLink bandwidth, and 2X networking bandwidth (ConnectX-9)

- **NVIDIA Vera Rubin Ultra NVL576:** Next generation NVLink liquid cooled rack-scale architecture connecting Arm-based NVIDIA Vera CPUs and 576 Rubin Ultra GPUs in a single NVLink domain. Compared to GB300 NVL72, Vera Rubin NVL576 features: 10X FP4 inference FLOPS, 14X FP8 training FLOPS, 8X HBM memory bandwidth, 9X fast memory (GPU + CPU), 12X all-to-all NVLink bandwidth, and 8X networking bandwidth (ConnectX-9)

Pricing: Will be available closer to release

Availability: Vera & Rubin will be available in 2026. Rubin Ultra will be available in 2027.

NVIDIA Photonics / Co-Packaged Optics

NVIDIA Spectrum™-X and NVIDIA Quantum-X silicon photonics networking switches are the world's most advanced networking solution for the era of agentic AI, enabling AI factories to connect millions of GPUs across multi-sites. NVIDIA co-packaged optics (CPO) based networks simplify manageability and design while enabling more power for compute infrastructure. These benefits are critical to delivering the scale needed to enter the future of million-GPU AI factories. By replacing pluggable transceivers with silicon photonics on the same package as the ASIC, NVIDIA CPO innovations provide:

- 3.5x better power efficiency
- 10x higher network resiliency
- 1.3x faster time to deploy compared to traditional networks

Pricing: TBA

Availability: Quantum-X Photonics InfiniBand Switch (Q3450-LD) will be available later this year in 2025. Spectrum-X Photonics Ethernet switch models will follow and be available in 2026.

Enterprise Support: NVIDIA Enterprise Support will be available for NVIDIA networking solutions.

External Resources: Silicon Photonics Press Release, Silicon Photonics Webpage, Silicon Photonics/CPO Video, Quantum-X Photonic Switch Datasheet, Quantum-X800 Platform Solution Sheet

GTC Sessions:

- Data Expressways: Unlocking the Full Potential of AI With Next-Gen Networking [S71190] - Gilad Shainer & Michael Kagan
- Wired for AI: Lessons from Networking 100K+ GPU AI Data Centers and Clouds [S71145] - Gilad Shainer hosting CoreWeave, Azure, Oracle, Meta, Google

---

NVIDIA Mission Control (Previously Advanced Cluster Management)

Enterprises now have an unprecedented range of AI options and developers want a way to harness the power of AI infrastructure with a hyperscaler's agility, efficiency, and scale but without the burdens of cost, complexity, and expertise placed on IT. **NVIDIA Mission Control** powers every aspect of AI factory operations — from developer workloads to infrastructure to facilities — with the skills of a world-class operations team delivered as software. It powers NVIDIA Blackwell and Blackwell Ultra data centers for the newest frontiers of AI, bringing instant agility to inference and training workloads and full-stack intelligence to deliver world-class infrastructure resiliency and accelerate AI experimentation.

- NVIDIA Mission Control is the only unified operations and orchestration software platform that automates the complex management of AI factories and workloads
- Get unprecedented agility, boosting utilization by 5x, across both training and inference workloads with NVIDIA Run:ai technology for workload management
- Autonomous recovery identifies, isolates, and recovers from job and hardware failures 10x faster – leading to accelerated AI experimentation

Pricing: $4,000 per GPU per year ($12,000 per GPU for initial 3 year subscription)

Availability: NVIDIA Mission Control for NVIDIA DGX GB200 and DGX B200 systems is available now. NVIDIA GB200 NVL72 systems with Mission Control are expected to soon be available from Dell, HPE, Lenovo and Supermicro.

Enterprise Support: Enterprise Support is included with NVIDIA Mission Control

External Resources: Webpage, Corporate Blog, Solution Overview

GTC Sessions:

- Build a Strategic Foundation for Enterprise Generative AI [S72357]
- Next-Gen Data Centers: Intelligent Automation and Integrated Observability for Peak Developer Productivity [S72361]

NVIDIA DGX GB300

DGX GB300 is a liquid-cooled, rack-scale AI infrastructure solution that scales to tens of thousands of NVIDIA GB300 Grace Blackwell Ultra Superchips for training and inference of state-of-the-art AI reasoning models.

- High compute density, liquid-cooled, rack-scale design

- Powered by GB300 Grace Blackwell Ultra Superchips with 36 Grace CPUs and 72 Blackwell Ultra GPUs per rack, connected via fifth-generation NVLink
- Scalable to tens of thousands of GB300 Superchips
- Powered by NVIDIA Mission Control software, simplifying AI operations with agility, resilience, and hyperscale efficiency for enterprises.

Pricing: TBA

Availability: We expect DGX GB300 systems to be available later this year.

Enterprise Services: Enterprise Services are available for NVIDIA DGX GB300 solutions.

External Resources: Webpage, Press Release, Datasheet

GTC Sessions:

- Build a Strategic Foundation for Enterprise Generative AI [S72357]
- Next-Generation at Scale Compute in the Data Center [S73623]

NVIDIA DGX B300

The latest iteration of NVIDIA's air-cooled DGX systems, NVIDIA DGX B300 is powered by the groundbreaking NVIDIA Blackwell Ultra GPU. DGX B300 is purpose-built for the era of AI reasoning, capable of accelerating any workload across the entire AI development pipeline.

- The world's first system with the NVIDIA Blackwell Ultra GPUs, featuring 2.3TB of GPU memory.
- 4X more training performance, 11X more inference performance compared to AI factories built with NVIDIA Hopper.
- New chassis design to seamlessly integrate into NVIDIA MGX or traditional enterprise racks
- For the first time, customers can pick between an integrated PSU or external busbar for power supply
- 8x OSFP ports serving 8x single-port NVIDIA ConnectX-8 800Gbp/s InfiniBand and Ethernet, and 2x dual-port NVIDIA BlueField-3 DPUs
- DGX B300 systems are the building blocks of the next-generation NVIDIA DGX BasePOD and NVIDIA DGX SuperPOD

Pricing: TBA

Availability: We expect DGX B300 systems to be available later this year.

Enterprise Services: Enterprise Services are available for NVIDIA DGX B300 solutions.

External Resources: Webpage, Press Release, Datasheet

GTC Sessions:

- Build a Strategic Foundation for Enterprise Generative AI [S72357]
- Next-Generation at Scale Compute in the Data Center [S73623]

NVIDIA Omniverse Blueprint for AI Factory Design and Operations

To help design and optimize gigawatt-scale AI factories, NVIDIA unveiled the Omniverse Blueprint for AI Factory Design and Operations, enabling engineers to use digital twins to test and optimize power, cooling, and networking before construction. This reference workflow enables real-time simulation and collaboration, helping optimize energy and prevent downtime by integrating with partners like Cadence, ETAP, Schneider Electric, and Vertiv.

Availability: Blueprint coming soon (COMPUTEX timing)

External Resources: Announcement Blog, Physical AI Press Release, GTC Keynote Demo, AI Factories Blog

---

NVIDIA Dynamo

Models are growing in size and are increasingly being integrated into agentic AI workflows that require interaction with multiple other models. Deploying these models and workflows in production environments involves distributing them across multiple nodes of GPUs, which demands careful orchestration and coordination.

NVIDIA unveiled NVIDIA Dynamo, a new open source AI inference-serving software designed to maximize token revenue generation for AI factories deploying reasoning AI models. NVIDIA Dynamo orchestrates and accelerates inference communication across thousands of GPUs. It uses disaggregated serving to separate the processing and generation phases of LLMs on different GPUs, allowing each phase to be optimized independently for its specific needs and ensuring it can maximize GPU resource utilization.

NVIDIA Dynamo incorporates features that enable it to reduce costs. It can dynamically add, remove and reallocate GPUs in response to fluctuating request volumes and types. It can pinpoint specific GPUs in large clusters that can minimize response computations and route queries to them. And it offloads inference data to more affordable memory and storage devices, quickly retrieving them when needed, minimizing inference costs.

NVIDIA Dynamo provides:

- 30x more throughput running DeepSeek R1 models on NVIDIA GB200NVL72
- 2x more throughput running Llama 70B models on NVIDIA Hopper

Pricing: NVIDIA Dynamo is available as open source software on GitHub and will be included with NVIDIA NIM inference microservices, part of NVIDIA AI Enterprise.

Availability: Available today on GitHub. For simplified production deployment, Dynamo will be supported in NVIDIA NIM, part of NVIDIA AI Enterprise (target GA May 2025).

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments. Support for deploying models via Dynamo is also coming to NVIDIA NIM.

External Resources: Customer Deck, Press Release, Enterprise Webpage, Developer Webpage, Tech Blog

Partners: Cohere, Together AI, and Perplexity AI have provided quotes endorsing NVIDIA Dynamo and shared their interest in leveraging its capabilities. NVIDIA Dynamo will enable users to accelerate the adoption of AI inference, including at AWS, Cohere, CoreWeave, Dell, Fireworks, Google Cloud, Lambda, Meta, Microsoft Azure,Nebius, NetApp, OCI, Perplexity, Together AI and VAST, among others.

GTC Sessions:

- A Distributed Inference Serving Framework for Reasoning LLM Models [S73042]
- Generative AI Model Serving With Triton Inference Server [CWE73367]
- Generative AI Inference At Scale [CWE73249]

Computer-Aided Engineering (CAE/EDA)

NVIDIA Blackwell and CUDA-X Accelerates Computer-Aided Engineering by up to 50X making Real Time Digital Twins possible. Any end customer of a CAE ISV tool can significantly accelerate their workloads by leveraging Blackwell GPU.

The growing ecosystem integrating Blackwell into its software includes Altair, Ansys, BeyondMath, Cadence, COMSOL, ENGYS, Flexcompute, Hexagon, Luminary Cloud, M-Star, NAVASTO, an Autodesk company, Neural Concept, nTop, Rescale, Siemens, Simscale, Synopsys, and Volcano Platforms.

At Supercomputing we first announced our OV-RTDT, or Omniverse Blueprint for Real Time Digital Twin, which combines CUDA-X accelerated solvers with AI surrogate models all connected up for interactive investigation in Omniverse. Now at GTC we're making this blueprint generally available. This is a reference architecture that helps our ISV partners adopt the technologies needed to provide interactive design to their customers. At the show this week you'll see real-time digital twins of cars, supersonic jets, the human heart, and many other incredible applications from our partners and their customers.

Availability: Blackwell available today, Rescale CAE Hub Blackwell availability on the cloud. The blueprint is generally available at GTC25. Early access sign-up is available from the API catalog.

External Resources: Press Release, CAE Solutions Page, API Catalog

---

Earth-2 - NVIDIA Omniverse Blueprint for Earth-2 Weather Analytics

At GTC25, we announced the first ever blueprint for Earth-2 which is a reference architecture for climate tech ISVs'. Leveraging Earth-2 blueprint, climate tech companies can create their solutions for higher-resolution, energy-efficient, more accurate weather predictions and disaster preparedness. Blueprint accelerates ISVs and LHAs in building solutions while leveraging ecosystem partners that have data platforms to deliver products to end consumers. Leading adopters include, G42, JBA Risk Management, Spire, Esri, Orora Tech, Tomorrow.io, and more.

Availability: The blueprint will be generally available at GTC25 - access it from the API catalog

External Resources: Press Release, Earth-2 Webpage, API Catalog

---

CUDA-X Libraries

At GTC25, we announced new libraries/updates to existing libraries.

CUDA-X libraries accelerate every industry and every scientific research. NVIDIA has over 900 CUDA-X libraries and AI models that developers can leverage to accelerate their applications. We released a blog at GTC, highlighting how CUDA-X libraries accelerate a vast variety of applications.

- **cuDSS**: Math library for accelerating sparse direct solvers in matrix multiplications.
    - Adopters: Ansys, Gurobi, Siemens, Synopsis
    - Public beta

- **cuEquivariance:** New library for accelerating equivariant neural networks, directly integrated into the MACE foundation model, observing up to a 5x speed-up!
    - Public beta
- **cuOpt:** Delivers up to 70X faster decision optimization, finding feasible solutions in seconds
    - Adopters: Gurobi, IBM, FICO, HIGHS, SimpleRose, Lyric, COI-OR, COPT, AMPL, ZIB
    - Will be available open-source
- **cuPyNumeric:** Accelerated Python library extends cluster scale speed-up to SciPy for scientists and researchers.
    - Adopters: SLAC, NPCI (fraud detection), UMASS Boston, Australia National University, LANL, (skip others we don't have approval or not current)
- **PhysicsNemo (previously known as Modulus):** New NIM + retraining pipeline for external aerodynamics; helps to emulate high fidelity flow.
    - Adopters: Luminary Cloud, SimScale for CFD, Ansys for RedHawk SC (electro thermal application)
- **Warp:** A Python framework for high-performance simulation and graphics.
    - Adopters: Autodesk XLB and Google DeepMind

NVIDIA cuOpt

NVIDIA cuOpt is a GPU-accelerated optimization engine designed to tackle large-scale decision-making challenges with unprecedented speed and efficiency. Key capabilities include:

- Mixed-Integer Linear Programming (MILP) & Linear Programming (LP) solvers for production planning, resource allocation, workforce scheduling and factory placement.
- Vehicle Routing Problem (VRP) solver for fleet management, dispatch optimization and deliveries.
- Record breaking performance, including a MIPLIB and 23 world records in routing benchmarks.
- Delivers up to 60X speedup for MIP heuristics, 3000X for LP, and 240X for VRP, solving real-world problems with millions of variables and constraints.
- Hybrid optimization support, seamlessly integrating with CPU solvers for multi-cloud environments.

Pricing: cuOpt will be available as an open-source solution starting May 2025. For production-ready deployment, it is included in the NVIDIA AI Enterprise offering, providing security, reliability, and enterprise-class support.

Availability: cuOpt VRP (Route Optimization) - Available via API catalog or self-hosting through NVIDIA AI Enterprise. cuOpt LP/MIP available in May.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: NVIDIA cuOpt Webpage - Featuring customer and partner logos, NVIDIA cuOpt Get Started Webpage - Setup guides & resources, GTC Corporate Blog - Featuring 9 partner blogs supporting cuOpt open-source launch: COIN-OR, COPT, FICO, Gurobi, HiGHS, IBM, Lyric, SimpleRose, and ZIB, AMPL blog supporting cuOpt open-source launch.

GTC Sessions:

- Revolutionize Supply Chain Analytics with AI & Accelerated Computing [S74122] - EY, Databricks, & NVIDIA
- Turbocharging Prescriptive Analytics & Optimization with GPU Acceleration [S72603] - SimpleRose
- Advances in Optimization [S7220]
- Connect with Experts [CWE7329]
- Accelerating Portfolio Optimization [DLIT71690] - Cornell University, Cohen & Steers, NVIDIA

PhysicsNeMo

PhysicsNeMo is a framework for training and deploying physics driven AI models. It is a toolkit that provides utilities to train engineering scale, physics driven model architectures and domain-specific pre-trained models and benchmarks. It is a toolkit for developers to build AI models for developing their solutions. PhysicsNeMo is not an out of the box AI solution for CAE. “Modulus” is rebranded to PhysicsNeMo.

Availability: PhysicsNeMo Webpage

External Resources: PhysicsNeMo Webpage

NVIDIA cuML

NVIDIA cuML 25.02 introduces zero code change acceleration for scikit-learn algorithms in open beta. The latest release enables data scientists and machine learning engineers to keep their scikit-learn applications unchanged and achieve 50x faster performance on NVIDIA GPUs vs CPUs.

- NVIDIA cuML now brings zero code acceleration for scikit-learn, UMAP and HDBSCAN
- scikit-learn runs fastest on NVIDIA GPUs – 50x vs CPUs
- UMAP & HDBSCAN run in seconds vs hours, making it possible to process higher-dimensional data
- Integrates w/ Python workflows using pandas, matplotlib etc. and provides easy way to run models trained w/ GPUs on CPUs & GPUs

- Accelerating ML on NVIDIA GPUs delivers better models, faster results & higher productivity

Pricing: Available as open source as part of RAPIDS 25.02 from github

Availability: Download today in RAPIDS 25.02 as open beta!

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Product Page, Tech Blog, Demo Video, Google Colab Notebook, DLI - Bring Accelerated Computing to Data Science in Python

GTC Sessions and Workshops:

- RAPIDS in 2025: Accelerated Data Science Everywhere [S73290]
- Unlock the Speed of Light for Data Science Workflows With Gemini Coding Assistant [S73027]
- Build Lightning-Fast Data Science Pipelines in Industry With Accelerated Computing [S74334]
- Bring Accelerated Computing to Data Science in Python [DLIT73877]
- Accelerate Clustering Algorithms to Achieve the Highest Performance [DLIT74335]

RAPIDS Accelerator for Apache Spark/Project Aether

- RAPIDS Accelerator for Apache Spark reduces your TCO up to 80%. It accelerates analytic and machine learning workloads with no code change, and is available on Amazon Web Services, Cloudera, Databricks, Dataiku, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure.
- Project Aether automates the configuration, testing and optimization of hundreds of jobs onto GPUs and greatly shortens the time to value (lowered TCO)

Pricing: Free/Open Source + upsell for NVIDIA AI Enterprise. Project Aether for LHAs only - it is not open source

Availability: Spark RAPIDS is Available now. Project Aether is available for LHAs only

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Product Page, User Guide, Project Aether Blog

---

NVIDIA Accelerated Quantum Research Center (NVAQC)

The NVAQC is a NVIDIA research center where AI supercomputing will be used to solve some of quantum computing's most challenging problems, ranging from tackling qubit noise to the transformation of experimental quantum processors into practical devices. The center will run simulations and also integrate leading quantum hardware from partners.

- Starting operations later this year in Boston, the NVAQC is a GB200 NVL72 system with Quantum Infiniband 2 Networking - containing 576 Blackwell GPUs.
- The NVAQC is a research center where breakthroughs will be made leading to *accelerated quantum supercomputers* - devices combining classical and quantum hardware to solve meaningful problems. Though the NVAQC itself is not intended to be or become a useful quantum computer.
- Announced with founding quantum hardware collaborators QuEra, Quantinuum, Quantum Machines and academic researchers from the Engineering Quantum Systems group (EQuS) at MIT and the Harvard Quantum Initiative (HQI).
- The center will enable GPU-accelerated simulations of quantum algorithms and hardware, as well as research into how to integrate AI supercomputing with quantum processors (QPUs).
- Partners will bring QPUs into the center to experiment with using GPUs for control and error correction tasks needed to run and scale their hardware.
- The NVAQC will draw on NVIDIA's DGX Quantum reference architecture for integrating quantum and classical hardware, and NVIDIA's CUDA-Q platform for developing for and running hybrid quantum-classical applications.

Pricing: N/A

Availability: The center will primarily be used for NVIDIA research, with selected partners involved on specific projects.

External Resources: Press Release, Corporate Blog.

, FAQs.

GTC Sessions: Full targeted agenda for Quantum Computing at GTC. Quantum Day runs on March 20th, including a Fireside chat between JHH and 14 quantum industry leaders.

NVIDIA DGX Spark

Previously known as Project DIGITS, DGX Spark is part of a new class of computers created for the AI era. Designed from the ground up to build and run AI, DGX Spark is an ideal platform for AI developers who often need to work on public or private clouds due to insufficient memory or lack of access to the optimal software stack their workloads require.

- Formal product name: DGX Spark
- Customers will be able to reserve systems at the NVIDIA store after the keynote

- OEM partners will be announcing their own branded versions of DGX Spark

Pricing: Starting at $2,999 from select partners

Availability: Customers will be able to reserve DGX Spark systems at the NVIDIA.com store starting on 3/18 after the keynote. OEM partners ASUS, Dell, and HPI will be building their own branded systems and will provide details on availability as part of their GTC press efforts.

External Resources: DGX Spark Webpage, Press Release

GTC Session: Project DIGITS: From Project to Your Personal AI Supercomputer [S74680]

NVIDIA DGX Station

DGX Station is part of a new class of computers created for the AI era. Designed from the ground up to build and run AI, DGX Station brings data center-level performance to the desktop and is an ideal platform for AI developers who often need to work on public or private clouds due to insufficient memory or lack of access to the optimal software stack their workloads require.

- Powered by the NVIDIA GB3000 Grace Blackwell Ultra Desktop Superchip
- DGX Station provides 784GB of coherent memory for working with very large AI models and workflows locally
- The systems will include a NVIDIA ConnectX-8 SuperNIC for high speed network connectivity to other DGX Stations or other high-speed network infrastructure.
- There will be no NVIDIA branded DGX Station, OEM partners will be announcing their own branded versions of DGX Spark as part of their GTC PR.

Pricing: System pricing will be set by the OEM partners

Availability: DGX Station systems will be available through OEM partners including ASUS, Dell, HPI, and Lambda Labs starting later this year. OEM partners will provide their own timelines for availability.

External Resources: Webpage, Press Release

NVIDIA RTX PRO

The NVIDIA RTX PRO™ Blackwell series is a revolutionary generation of workstation and server GPUs redefining AI and visual computing for professionals and enterprises across industries.

Data Center - RTX PRO 6000 Blackwell Server Edition

The first NVIDIA Blackwell-powered data center GPU built for both visual computing and enterprise AI — the NVIDIA RTX PRO 6000 Blackwell Server Edition delivers breakthrough acceleration for a broad range of enterprise workloads, from multimodal and LLM inference to physical AI, scientific computing, 3D graphics, and video applications.

RTX PRO 6000 Blackwell Server Edition is the next-generation universal data center GPU, with supercharged performance compared to the previous generation L40S — up to 5x LLM inference, nearly 7x faster genomics sequencing, 3.3x speedups for text-to-video generation, nearly 2x faster inference for recommender systems, and 2x for rendering.

- Breakthrough Multimodal AI Inference:
    - 5th-Gen Tensor, 2nd-Gen Transformer Engine, FP4,
    - Full Media Pipeline: 4 NVENC/ NVDEC/ NVJPEG
- Powerful Graphics and Visual Computing:
    - 4th-Gen RTX, Neural Shaders, DLSS 4
- Data Center Ready:
    - 96GB GDDR7,1.6 TB/s Memory BW, 128MB L2 Cache
    - Multi-Instance GPU (MIG), TEE Confidential Compute

Pricing: Pricing for systems and cloud instances featuring RTX PRO 6000 Blackwell Server Edition will be determined by each partner.

Availability: RTX PRO 6000 Blackwell Server Edition will be available from our ecosystem of leading data center systems partners and cloud partners beginning in May 2025

External Resources: Webpage, Blog, RTX Press Release

NPN Webinars: Unlock Powerful Data Center Solutions with RTX GPUs & AI Virtual Workstations NALA/EMEA on Apr 15 | APAC on Apr 16

Workstation - RTX PRO Blackwell Desktop GPUs

NVIDIA is introducing a new family of desktop GPUs based on the NVIDIA Blackwell architecture, delivering a massive leap in workstation performance designed to empower designers, developers, engineers and creatives with unprecedented AI and graphics capabilities. The new NVIDIA RTX PRO Blackwell Desktop GPUs include the RTX PRO 6000 Blackwell Workstation Edition, RTX PRO 6000 Blackwell Max-Q Workstation Edition, RTX PRO 5000 Blackwell, RTX PRO 4500 Blackwell, and RTX PRO 4000 Blackwell.

The RTX PRO Blackwell desktop GPUs deliver groundbreaking advancements for AI-augmented workflows, 3D design, simulation, and content creation. Key features include:

- 4th-Gen RT Cores: Double the ray tracing performance for photorealistic scenes and intricate 3D models.
- 5th-Gen Tensor Cores: 4,000 AI TOPS and FP4 precision to accelerate generative AI, physics-based simulations, and real-time neural rendering.
- 9th-Gen NVENC & 6th-Gen NVDEC: up to 17x faster video encoding with hardware accelerated 4:2:2 support, 2x H.264 decode throughput, and additional engines enabling up to 3x simultaneous video streams over the previous generation.
- GDDR7 Memory: Up to 96GB, enabling seamless handling of massive datasets for AI training, CAD, and immersive XR environments.
- PCIe Gen5 & DisplayPort 2.1: Double bandwidth for data-heavy tasks and support for 8K displays, ideal for multi-monitor creative setups.
- MIG Technology (on 6000- and 5000-class): Secure GPU partitioning (up to four instances) for multitasking AI, rendering, and simulation workloads simultaneously.

Availability: RTX PRO 6000 Workstation Edition GPUs will be available starting in April from channel partners (PNY, TD Synnex) and from leading OEM partners (Dell, HP, Lenovo) beginning in May.

External Resources: Webpages: RTX PRO 6000, RTX PRO 6000 Max-Q, RTX PRO 5000, RTX PRO 4500, RTX PRO 4000, Press Release

Laptop - NVIDIA RTX PRO Blackwell Generation Laptop GPUs

NVIDIA is introducing a new family of laptop GPUs based on the NVIDIA Blackwell architecture to supercharge professional creative, design, engineering, and AI workflows from anywhere. The new NVIDIA RTX PRO™ Blackwell Generation Laptop GPUs include the RTX PRO 500, 1000, 2000, 3000, 4000, 5000 Blackwell Generation.

- RTX PRO Blackwell Generation Laptop GPUs deliver up to 2X faster performance compared to Ada - based on RTX 5000-class performance comparisons.
- Next-gen graphics with 4th Gen RT Cores for accelerating ray tracing, optimized for mega geometry, and new Streaming Multiprocessors optimized to bring AI to graphics with neural shaders.
- 5th Gen Tensor Cores for accelerating AI, including FP4 and DLSS 4 support.
- New NVENC/NVDEC video encoding and decoding engines introduce 4:2:2 support.
- Faster, larger GDDR7 memory - up to 24GB.
- Blackwell Max-Q technologies for optimizing performance and power efficiency.

Availability: Begins later this year, and varies by OEM (Dell, HP, Lenovo, Razer).

External Resources: Webpage, Press Release, Line Card

NVIDIA DGX Cloud

NVIDIA DGX Cloud is NVIDIA's exemplar cloud, delivering optimal performance for any application on any cloud infrastructure. New to DGX Cloud:

- **DGX Cloud Serverless Inference:** Powered by NVIDIA Cloud Functions (NVCF), DGX Cloud Serverless Inference provides scalable inferencing capabilities for all NVIDIA libraries and NIM as well ISVs across NVIDIA capacity on CSPs (public clouds) or NVIDIA's GeForce Network.
- **DGX Cloud Benchmarking:** A new benchmarking service for agentic AI to optimize performance for workloads across all stages of the AI Factory.
- **NeMo Curator and Post-Training on DGX Cloud**: New services to accelerate video curation and fine-tuning.
- **AI Development Accelerator program on DGX Cloud:** NVIDIA and leading VC partners will offer $100k toward DGX Cloud through the Inception Program to spur innovation through speed of light access to NVIDIA AI from one generation to the next.
- **DGX Cloud Create:** NVIDIA Run:ai is fully integrated into DGX Cloud Create serving as the AI workload optimization and orchestration interface. DGX Cloud Create is part of the DGX Cloud unified platform available in the AWS, Google Cloud, Microsoft Azure*, and Oracle Cloud* marketplaces (*Run:ai coming soon to these CSPs).

Pricing: DGX Cloud is a unified platform with a single SKU that includes all of the above features.

- DGX Cloud with H100 80GB, 8 GPU Node 1-2 months: $40,880/node/month
- DGX Cloud with H100 80GB, 8 GPU Node 3-5 months: $39,128/node/month
- DGX Cloud with H100 80GB, 8 GPU Node 6-8 months: $37,376/node/month
- DGX Cloud with H100 80GB, 8 GPU Node 9-11 months: $35,624/node/month
- DGX Cloud with H100 80GB, 8 GPU Node 12 months: $34,748/node/month

Higher Education & Research (HER) and Inception customers are allowed a 30% discount.

Availability: DGX Cloud Create, DGX Cloud Serverless Inference, and DGX Cloud Benchmarking are available now.
NeMo Curator and Post-Training on DGX Cloud are available as DA (directed availability) via a sign-up form.

Enterprise Support: Designated Technical Account Manager and Business Critical 24/7 support are included in DGX Cloud.

External Resources: DGX Cloud (developer.nvidia.com), DGX Cloud (nvidia.com), DGX Cloud Starting Options, DGX Cloud Benchmarking, DGX Cloud Serverless Inference, DGX Cloud Create, DGX Cloud Create Form Fill, DGX Cloud with NVIDIA GB200 Notify Me Form

GTC Sessions:

- Build a Resilient Cloud Strategy in the Era of Generative AI [S72772]
- Scalable, Configurable Multi-Modal Data Processing Pipelines to Enhance Gen AI Accuracy [S73287]
- DGX Cloud Deep Dive [CWE72870]
- Develop and Deploy Optimized Generative AI Apps in the Cloud [CWE73361]

Llama Nemotron Models with Reasoning

NVIDIA Llama Nemotron is a family of state-of-the-art reasoning models that achieves highest accuracy for a diverse set of agentic AI tasks including graduate level scientific reasoning, advanced math, and tool calling. The models are optimized for different computing platforms, achieving up to 5x higher throughput compared to leading open models and have a unique Reason ON/OFF capability to save compute on queries that don't require "thinking" to answer.

- **Nano** gives the highest reasoning accuracy for its class
- **Super** provides the best accuracy with highest throughput on a single data center GPU
- **Ultra** delivers maximum agentic accuracy on multi-GPU servers, at data-center scale

NVIDIA has also opened up the training datasets, post-training techniques, and tools for developers to build their custom reasoning models. Available as NVIDIA NIM inference microservices, developers can easily deploy agents using the reasoning models on their own infrastructure. Leading AI platform partners - Accenture, Amdocs, Atlassian, Box, Cadence, CrowdStrike, IQVIA, Microsoft, SAP, and Servicenow - are working with NVIDIA Llama Nemotron reasoning models.

Pricing: NIM microservices are available with a NVIDIA AI Enterprise license

Availability: Nano and Super available at GTC. Available as NIM microservice - try and/or download from build.nvidia.com. Download from Hugging Face.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: build.nvidia.com, Tech Blog

GTC Session: Building Custom Reasoning Models to Achieve Advanced Agentic AI Autonomy [S74781]

NVIDIA NeMo Microservices

NVIDIA NeMo™ microservices is an end-to-end platform for building agentic and generative AI applications that utilize LLMs and data flywheels, enabling enterprises to continuously optimize their AI agents with the latest information.

NeMo helps enterprise AI developers easily curate data at scale, customize large language models (LLMs) with popular fine-tuning techniques, consistently evaluate models on industry and custom benchmarks, and guardrail them for appropriate and grounded outputs.

- NeMo Curator: GPU-accelerated modules for curating high-quality, multi-modal training data
- NeMo Customizer: High-performance, scalable microservice that simplifies the fine-tuning of LLMs
- NeMo Evaluator: Automated evaluation of AI pipeline and models using academic and custom benchmarks
- NeMo Retriever: Fine-tuned microservices to build AI query engines with scalable document extraction and advanced retrieval-augmented generation (RAG) for multimodal datasets
- NeMo Guardrails: Seamless orchestrator for building robust safety layers to ensure accurate, appropriate, and secure agentic interactions
- NIM Operator: Kubernetes Operator that is designed to facilitate the deployment, management, and scaling of NeMo and NIM microservices on Kubernetes clusters

A data flywheel provides a feedback loop where data collected from interactions or processes is used to train and refine AI models, which in turn generate better outcomes and more valuable data. This iterative cycle enables continuous improvement, adaptability, and compounding value in AI-driven systems.

Pricing: NVIDIA NeMo microservices is part of a NVIDIA AI Enterprise license.

Availability: NeMo microservices will be generally available in April 2025

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Tech Blog, Data Flywheel Glossary

GTC Sessions:

- Building Scalable Data Flywheels for Continuously Improving AI Agents [S73280]
- Insights and Lessons Learned From Building LLM-Powered Applications [S73649]
- Build Generative AI With NVIDIA NeMo [CWE73358]
- Please review this doc for all the sessions

NVIDIA NeMo Retriever and NVIDIA AI Blueprint for RAG

NVIDIA NeMo Retriever is a collection of microservices for extraction, reranking, and embedding used to build state-of-the-art multimodal retrieval pipelines with high accuracy and maximum data privacy. It delivers quick, context-aware responses for AI applications like advanced RAG and agentic AI workflows. New NeMo Retriever microservices and features enable 15x faster multimodal PDF extraction and 50% fewer incorrect answers when compared to open-source alternatives. Part of the NVIDIA NeMo platform, NeMo Retriever commercial microservices, built with NVIDIA NIM, let developers connect AI applications to large enterprise datasets wherever they reside and fine-tune them to align with specific use cases.

The NVIDIA AI Blueprint for RAG is a reference architecture implemented in code enabling developers to build GPU-accelerated, scalable, context-aware retrieval pipelines tailored to enterprise data. Linking LLMs with an organization's existing knowledge base improves both accuracy and throughput, which are critical for modern generative AI applications. It can ingest and extract a variety of data types such as charts, tables, and infographics by leveraging the NeMo Retriever extraction, embedding, and reranking microservices. Together, the NVIDIA AI Blueprint for RAG and NeMo Retriever give customers a world-class solution for connecting data to AI applications

- NeMo Retriever extraction microservices increase multimodal retrieval accuracy with 50% fewer incorrect answers and 15x higher multimodal data extraction throughput.
- NeMo Retriever's embedding and reranking microservices deliver 3x higher embedding throughput, 1.6x better reranking throughput, and 35% fewer incorrect answers reflecting higher question-answering retrieval accuracy.
- NeMo Retriever embedding microservices enhance storage efficiency with its dynamic length and long context support, improving data storage efficiency by 35x.
- NeMo Retriever English text embedding microservice combined with GPU accelerated vector database result in 7x higher index build throughput on the GPU.
- By leveraging the NeMo Retriever reranking microservice, in addition to the embedding microservice, improve RAG accuracy with 10% fewer incorrect answers and reduce RAG costs by 20%.

Pricing: NeMo Retriever is part of NVIDIA AI Enterprise

Availability: New NeMo Retriever extraction microservices will be generally available at GTC, along with new features and updates to the embedding and reranking microservices that went GA in December 2024.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Enterprise Generative AI Sales Kit, Tech Blog, Explore Retrieval Models and Blueprints Using NeMo Retriever on Build

GTC Sessions:

- Transform an Enterprise Data Platform With Generative AI and RAG [S72205]
- Building Future-Ready AI with Agents and Data Flywheels [S72338]
- AI Agents for Real-Time Video Understanding and Summarization [S72784]
- How to Build Multimodal Retrieval-Augmented Generation and Agentic AI Pipelines [S72208]
- Scalable, Configurable Multi-Modal Data Processing Pipelines to Enhance Generative Accuracy [S73287]
- Best Practices and Techniques for Building Agentic and Retrieval Pipelines [CWE72181]
- Secrets for Scaling Gen AI from Proof of Concept to Production [CWE72612]
- Building RAG Agents With LLMs [DLIW73644]
- Make Retrieval Better: Fine-Tuning an Embedding Model for Domain-Specific RAG [DLIT71238]
- Blueprints for Success: Navigating Agent Workflows for Real-World Multimodal Retrieval [DLIT71592]

NVIDIA AI-Q NVIDIA Blueprint

AI-Q (pronounced IQ) is a NVIDIA Blueprint for multimodal queries across many data sources, leveraging state-of-the-art RAG and reasoning models to orchestrate AI agents. With AI-Q developers can build customized agentic systems that integrate with enterprise data sources and tools, and maintain data privacy.

The blueprint integrates AI agents, reasoning models, and NVIDIA NeMo Retriever for multimodal extraction and world-class information retrieval. It uses the new open-source NVIDIA AgentIQ toolkit to provide observability into the reasoning process—for how plans are structured, data sources are selected, and decisions unfold, step-by-step. This ensures transparency and trust in AI-driven decision-making, empowering the new digital workforce to tackle complex queries with confidence.

Pricing: AI-Q NVIDIA Blueprint and the AgentIQ toolkit are free. The AI-Q NVIDIA Blueprint is built with NVIDIA NIM microservices, NVIDIA NeMo Retriever, and Llama Nemotron with reasoning, part of NVIDIA AI Enterprise.

Availability: Sign up to be notified when the AI-Q NVIDIA Blueprint is available on build.nvidia.com and try the NVIDIA AgentIQ toolkit for free in GitHub after the announcement at GTC on March 18, 2025. Developers can download the AI-Q NVIDIA Blueprint in Q2 CY2025.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: AI-Q Corp Blog, AgentIQ Developer Page, AgentIQ Hackathon Webpage, AgentIQ Tech Blog

GTC Sessions:

- How to Build an Agentic AI System Using the Best Tools and Frameworks [S73739]
- How to Onboard Your Team of AI Agents to Transform Your Enterprise [S71784]
- See AI-Q in action at the NVIDIA booth

NVIDIA-Certified Storage

The new NVIDIA-Certified™ Storage program addresses the massive data demands of AI factories by offering high-performance storage certification for OEM enterprise deployments. It tests storage with selected NVIDIA-Certified Systems (including NVIDIA Spectrum-X networking and NVIDIA AI Enterprise software), to ensure they meet the highest standards of performance, quality, and reliability.

This new storage certification program complements NVIDIA Enterprise Reference Architectures (RAs) and NVIDIA-Certified Systems from our OEM partners, enabling partners and customers to efficiently deploy and scale enterprise AI infrastructure. By integrating certified storage with NVIDIA's accelerated computing, networking, and software, enterprise customers can confidently build AI factories that leverage data for faster, more accurate, and reliable AI models, unlocking new possibilities and driving innovation across industries.

- **Benchmarking:** Storage benchmark testing with NVIDIA-Certified servers, Spectrum-X networking, NVIDIA AI Enterprise software, and the partner's high-performance storage.
- **Enterprise Storage:** Certifies high-performance storage for accelerated computing use cases in enterprise deployments: Gen AI, Agentic AI, ML/DL, Data Analytics, HPC
- **AI Factory Building Block:** Complements NVIDIA-Certified Systems and NVIDIA Enterprise RAs for AI factory deployments. Empowers enterprises to deploy AI factories with the data required for faster, more accurate, and reliable AI models.
- **AI Data Platform Prerequisite:** Certification is a prerequisite for partners developing agentic AI infrastructure solutions built on the NVIDIA AI Data Platform.

Availability: Available now with NVIDIA-Certified Storage solutions from DDN, Dell, HPE, Hitachi, IBM, NetApp, Nutanix, PureStorage, VAST Data, and WEKA.

External Resources: Blog, Solution Brief, NVIDIA-Certified Systems Webpage

GTC Sessions:

- Unlock the Power of NVIDIA Certified Systems: Live Q&A With Experts [CWE74649]
- Everything You Want to Ask About NVIDIA Enterprise Reference Architectures [CWE74549]

NVIDIA AI Data Platform for Enterprise

The NVIDIA AI Data Platform is a customizable, reference design for a new class of AI infrastructure that integrates enterprise storage with NVIDIA-accelerated computing to power AI agents with near real-time business insights. Built on NVIDIA's expertise in AI workflow optimization, the platform unlocks the vast amount of enterprise data, making it readily available for AI agents to reason and solve complex queries.

- The NVIDIA AI Data Platform platform integrates GPUs, networking, and AI software directly into storage systems, enabling continuous data processing ('always-on embedding') to enable accurate, context-aware semantic search.
- NVIDIA NeMo Retriever delivers 16x faster data ingestion, and 97% lower vector storage needs. Achieves 50% faster data transfers with NVIDIA Spectrum-X networking for real-time insights. In addition, NVIDIA BlueField-accelerated storage systems provide 1.6x higher performance while reducing power consumption by 50%, resulting in 3x higher performance per watt.
- There are 10 NVIDIA-Certified Storage Partners announcing AI Data Platform-enabled solutions at GTC including DDN, Dell, Hitachi Ventara, HPE, IBM, NetApp, Nutanix, Pure Storage, VAST Data, and Weka. Additional partners are expected to adopt the reference design in the future.

Availability: The AI Data Platform reference design will become available in Q2.

External Resources: AI Data Platform Press Release, AI Data Platform Webpage, AI-Q Blueprint Corporate Blog, NVIDIA-Certified Storage Corporate Blog

GTC Sessions:

- Pioneering the Future of Data Platforms for the Era of Generative AI [S71650]
- Storage Innovations for AI Workloads [S72329]
- Enable Intelligent Storage to Process Data for AI Applications [S71937]

NVIDIA Riva

Introducing Riva Magpie (magpie-tts-multilingual) NIM, converts text into audio (speech). It is designed to address persistent issues in speech synthesis, such as audio hallucination and undesired vocalization. Riva Magpie improves intelligibility, speaker similarity, and naturalness of generated speech through a novel preference alignment framework and classifier-free guidance (CFG).

- Expanded Language Support: Magpie is available in English/Spanish/French and will include support for additional languages, significantly broadening its reach and usability for global applications this year. This expansion allows for more diverse and inclusive text-to-speech capabilities across various regions and languages.
- Enhanced Voice Customization: The new Magpie NIM introduces advanced voice customization features, enabling users to create highly personalized and expressive voices.
- Optimized Performance and Deployment: Magpie has been optimized for performance, providing low latency and high throughput. It is designed for seamless deployment across different environments, including cloud, on-premises, and edge devices, ensuring flexibility and scalability for various use cases.

Pricing: Riva is available as part of NVIDIA AI Enterprise.

Availability: Find magpie-tts-multilingual model at build.nvidia.com starting March 18, 2025

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Speech Category Page on Build, NVIDIA API Catalog, NVIDIA AI Enterprise 90-day License, NVIDIA Riva Webpage

GTC Sessions:

- Voice-to-Voice LLM for End-to-End Voice Generation [S71829]
- How Yum! Brands Serves Up Digital Innovation [S72790]
- Maximizing Contact Center Efficiency with NVIDIA AI: Boosting First Call Resolution and Cutting Handle Time [S71402]
- Speech AI Demystified [S73113]
- AI Agents and Digital Humans Shaping the Future of Interaction in Telecoms [S72988]
- Drive Integrated Customer Experience With Multi-Agents, From Speech to Action, Response, and Resolution (Presented by Cognizant) [S74559]

NVIDIA vGPU

vGPU 18 offers additional supported ecosystem platforms, new AI toolkits, and enhancements for GPU utilization.

- New supported hypervisors: Microsoft Windows Server 2025, Proxmox Virtual Environment (VE)
- Windows Subsystem for Linux (WSL) and live migration support with Windows Server 2025
- Heterogeneous vGPU support extended to Turing and Volta GPUs
- Omnissa Horizon version 2412 support

Other Announcements:

- New AI vWS Toolkit: Fine-tuning and Customizing LLMs
- Citrix and NVIDIA to announce AI Virtual Workstation running on Citrix DaaS and CVAD

Availability: Available now

Enterprise Services: Enterprise Support is available for vGPU

External Resources: vGPU Announcement Blog (live on March 19), vGPU 18 Release Notes, Installing vGPU on Proxmox VE video, AI vWS Toolkit: Fine-Tuning and Customizing LLMs and Video, vGPU Product Page, vGPU Resources Page, vGPU Product Guides, vGPU Linecard, vWS Solution Brief

Partner Resources: NPN Webinars: Unlock Powerful Data Center Solutions with RTX GPUs & AI Virtual Workstations NALA/EMEA on Apr 15 | APAC on Apr 16

GTC Sessions:

- Heritage Meets Technology: Leveraging Virtualized Platforms for Architectural Conservation and Business Innovation [S72407]
- Virtualization Unleashed From VDI to AI: Tactics for Successful Deployments [CWE72704]

NVIDIA Omniverse

Mega NVIDIA Omniverse Blueprint - Experience on Build.nvidia.com

The Mega NVIDIA Omniverse Blueprint, announced at CES, is now available to experience on build.nvidia.com. The reference workflow enables enterprises to bring physical AI to their

factories, warehouses, and industrial facilities through advanced simulation and testing of multi-robot fleets.

Availability: The blueprint is currently in early access. Developers can experience it on build.nvidia.com and sign up for notifications on general availability.

External Resources: CES Announcement Blog, CES Demo Video, Build.nvidia.com, Notify Me Form, Physical AI Press Release, Industrial Facility Digital Twins Use Case, GTC Robots Become Robots Demo

GTC Sessions:

- Revolutionizing Warehouse and Factory Management With NVIDIA "Mega" Blueprint for Robot Facilities [S73076]
- Reinvent Warehouses and Distribution Centers With AI-Powered Digital Twins [S74342]

Omniverse Development & Deployment in Every Cloud

Cloud-based NVIDIA Omniverse and Isaac Sim virtual workstations and Omniverse Kit App Streaming offerings provide developers the flexibility to develop, deploy, and scale out their applications via free and enterprise supported cloud instances, powered by NVIDIA GPUs.

Pricing:

- NVIDIA Omniverse Development Workstation - free with no enterprise support
- NVIDIA Omniverse Enterprise Workstation - $1/hr and includes enterprise support
- NVIDIA Isaac Sim Development Workstation - free with no enterprise support
- Note that instance pricing is in addition and is metered by the CSP

Availability: The development and enterprise workstation offerings launch at GTC on AWS and Azure marketplaces, with Omniverse Kit App Streaming coming to Azure as a private offer. They will launch on the OCI, and Google Cloud Platform marketplaces later this year. View WWFO FAQ for full breakdown and links to offerings.

- **AWS:** Check out the AWS marketplace for the listings**.** Omniverse and Isaac Sim development workstations run on EC2 G6e instances with NVIDIA L40S GPUs.

Omniverse Kit App Streaming will come to AWS later this year.

- **Azure:** Preconfigured Omniverse and Isaac Sim development workstations and Omniverse Kit App Streaming (Private Offer) launch at GTC on NVIDIA A10 GPUs.
- **OCI:** Omniverse and Isaac Sim Development and Enterprise workstations and Omniverse Kit App Streaming launch on the Oracle Cloud Infrastructure marketplace later this year, via compute bare-metal instances with NVIDIA L40S GPUs.
- **Google Cloud Platform:** Omniverse and Isaac Sim Development and Enterprise workstations and Omniverse Kit App Streaming launch later this year on NVIDIA RTX PRO 6000 Blackwell Server Edition.

Enterprise Support: NVIDIA Omniverse Enterprise Workstation includes up to three Enterprise Support cases.

GTC Session: Connecting Industrial Data to Digital Twins with Microsoft Power BI, NVIDIA Omniverse, and Open USD [DLIT71322]

NVIDIA Omniverse Kit 107 Release

The latest NVIDIA Omniverse Kit SDK 107 release is another major milestone for robotics applications. In addition to upgrades to OpenUSD version 24.05, updates to Python, C++, and Linux ABI, NVIDIA Isaac Sim 5.0 will also be developed on top of Kit 107. This release provides significant advances in sensor simulation, language support, and binary compatibility, which can greatly enhance the development and deployment of robotics applications.

Availability: Available Now

External Resources: Omniverse Press Release, OpenUSD for Robotics Blog

GTC Sessions:

- An Introduction to OpenUSD [S72488]
- OpenUSD Best Practices for Developing with NVIDIA Omniverse [CWE73248]
- Learn OpenUSD: Foundations to Applied Concepts [DLIT74499]
- Learn OpenUSD: Robotics Best Practices [DLIT71288]

NVIDIA Omniverse Blueprint for AV Simulation + Cosmos

Cosmos Transfer World Foundation Model (WFM) is now incorporated into the NVIDIA Omniverse Blueprint for Autonomous Vehicle (AV) Simulation. The blueprint is a reference workflow to create rich 3D worlds for training, testing, and validation. It contains APIs and services to build and enhance digital twins from real world sensor data, model physics and behavior of dynamic objects in a scene, and generate physically accurate and diverse sensor data.

Within the blueprint, Cosmos Transfer uses the physics-based scenes rendered by Sensor RTX APIs to generate new variations, including weather, lighting, and geolocations—turning one scenario into hundreds of drives.

The blueprint can be seamlessly integrated into existing workflows, enabling developers to replay driving data, generate new ground truth data, and perform closed-loop testing.

Availability: Later this year

External Resources: AV Sim Use Case Webpage, Cosmos Press Release

GTC Sessions:

- Advancing AV Development With Sensor Simulation and Cosmos [DD40002]
- An Introduction to NVIDIA Cosmos World Foundation Models [S72431]

NVIDIA Cosmos

NVIDIA Cosmos™ expands its suite of world foundation models (WFMs) with Cosmos Predict, Cosmos Transfer, and Cosmos Reason. These world foundation models are openly available and when coupled with NVIDIA Omniverse become the second computer for physical AI development in NVIDIA three computer strategy to drive adoption of NVIDIA software and hardware.

- **Cosmos Transfer** to generate scalable synthetic data for robot and autonomous vehicle development. It transforms 'ground truth' simulations from Omniverse and structural video inputs such as segmentation, LiDAR, and trajectory maps to photorealistic videos across varying environments, lighting, and weather conditions.
- **Cosmos Predict**, first revealed at CES, are purpose-built for post-training to develop custom specialized models like policy models for robots and autonomous vehicles. These are generalist models that generate world states as videos from text, video, and now from start and end images.
- **Cosmos Reason** is a suite of fully customizable reasoning models for developing intelligent downstream foundation models for vision AI, embodied AI and agentic AI. The model has spatiotemporal awareness and uses chain-of-thought reasoning to understand video data, plan and deliver best response to physical interactions - like what happens next when a box falls off the shelf.
- NVIDIA Cosmos go-to-market through NVIDIA Omniverse Blueprint for autonomous vehicle simulation and the NVIDIA Isaac GR00T Blueprint for synthetic manipulation motion generation. Post-training support for Cosmos WFMs available through NVIDIA AI Enterprise.

Pricing: Openly available under NVIDIA Open Model License.

Availability:

- Cosmos WFMs: Cosmos Predict and Cosmos Transfer on NGC, Hugging Face, GitHub. Cosmos Reason in private early access.
- Post-training tools for Cosmos WFMs: Implementation and post-training scripts on GitHub. NeMo Curator and managed service on DGX Cloud for video data curation in early access. Cosmos tokenizer and NeMo framework for training publicly available.

External Resources: Cosmos Platform Page, Cosmos Getting Started Page, NVIDIA Cosmos Press Release, Cosmos Tech Blog

GTC Sessions:

- An Introduction to NVIDIA Cosmos World Foundation Models [S72431]
- NVIDIA Cosmos World Foundation Models for Autonomous Driving Development [S73198]
- An Introduction to NVIDIA Cosmos for Physical AI [DLIT74500]
- Developing World Foundation Models with NVIDIA Cosmos [CWE74542]

Open Source Dataset for Physical AI

NVIDIA is releasing an open source Physical AI Dataset is a commercial-grade, pre-validated dataset to help researchers and developers kickstart physical AI projects that can be prohibitively difficult to start from scratch. Developers can either directly use the dataset for model pretraining, testing and validation — or use it during post-training to fine-tune world foundation models, accelerating the path to deployment.

The initial dataset is now available on Hugging Face, offering developers 15 terabytes of data representing more than 600 hours of synthetic footage, as well as 400 Universal Scene Description (USD) SimReady assets. Dedicated data to support end-to-end autonomous vehicle (AV) development — which will include 20-second clips of diverse traffic scenarios spanning over 1,000 cities across the U.S. and two dozen European countries — is coming soon.

Availability: Initial dataset to be released on 3/18 after keynote on Hugging Face

External Resources: Hugging Face Collections, Corporate Blog

NVIDIA AI Blueprint for Video Search and Summarization (VSS)

The blueprint combines VLMs, LLMs, RAG, and Metropolis media management microservices to develop visually perceptive and interactive AI agents deployed in warehouses, factories, retail stores, and smart cities to optimize processes.

- Accelerates development time - Reduces months/years of development effort by removing the heavy lifting of plumbing various generative AI models and NIM microservices together.
- Customizable & flexible - Common APIs allow you to easily swap VLM or LLM NIMs or 3rd party models and enable optimized deployments from enterprise edge to cloud.

- Reduce cost - Delivers summaries of long videos up to 200X faster than going through the videos manually

Features for upcoming GA include: single GPU deployment, RTSP short and live multi-streaming, burst mode video ingestion, audio transcription, one-click deployments through CSPs and Launchables, customizable video ingestion pipeline with CV and multi-view 3D tracker

Pricing: Free software for developers to build. Need NVIDIA AI Enterprise license for production workloads.

Availability: General Availability is coming in April. Blueprint is available now for early-access. Developers can preview on build.nvidia.com site and experiment with Launchables (only available after signing up for EA).

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Blueprint, VSS Launchable, Visual AI Agents Use-Case Page, Documentation, VLM Glossary, Webinar, Tech Blog 1, Tech Blog 2 , Tech Blog 3

GTC Sessions:

- AI Agents for Real-Time Video Understanding and Summarization [S72784]
- Build Next-Gen Agents With Large Vision Language Models [DLIT71406]

NVIDIA Isaac GR00T

NVIDIA Isaac GR00T is a platform including pre-trained models, simulation frameworks and synthetic data generation pipelines, and open datasets. NVIDIA Isaac GR00T is a platform for developing humanoid robots.

NVIDIA Isaac GR00T N1 Humanoid Robot Foundation Model

NVIDIA Isaac GR00T N1 is the world's first open foundation model for generalized humanoid robot reasoning and skills. This cross-embodiment model takes multimodal input, including language and images, to perform manipulation tasks in diverse environments. GR00T N1 was trained on an expansive humanoid dataset, consisting of real captured data, synthetic data generated using the components of the NVIDIA Isaac GR00T Blueprint, and internet-scale video data. It is adaptable through post-training for specific embodiments, tasks and environments.

Pricing: Free

Availability: Model available now on Hugging Face. Dataset and fine-tuning scripts available on GitHub.

External Resources: Robots Being Robots Keynote Demo, GR00T Demo, Press Release, Technical Blog, Isaac GR00T Webpage,

GTC Sessions:

- An Introduction to Building Humanoid Robots [S72590]
- Accelerating Robust Humanoid Robot Development for any Industry [CWE73360]

Isaac GR00T Blueprint for Synthetic Manipulation Motion Data Generation

A reference workflow for generating exponentially large synthetic motion data from a handful of human demonstrations for imitation learning, specifically for single arm manipulation.

- **Accelerates Data Collection:** Reduces data collection time from hours to minutes.
- **Streamlines Development:** Automates data generation, enabling faster iteration and quicker time to solution.
- **Enhances Model Generalization:** Adds diversity through domain randomization to improve model performance.
- **Bridges Sim-to-Real Gap:** NVIDIA Cosmos Transfer World Foundation Model (WFM) enhances images with photorealism, reducing the sim-to-real gap.

Pricing: Free

Availability: Available now on build.nvidia.com. Source code on GitHub.

External Resources: Press Release, GR00T Blueprint Technical Blog, Isaac GR00T Webpage, build.nvidia.com, Cosmos Technical Blog

---

Newton Open Source Physics Engine

Newton is an open-source, extensible physics engine being developed by NVIDIA, Google DeepMind, and Disney Research to advance robot learning and development. Built on NVIDIA Warp, which enables robots to learn how to handle complex tasks with greater precision, Newton is compatible with learning frameworks such as MuJoCo Playground or NVIDIA Isaac Lab—an open-source, unified framework for robot learning.

- **Open Source:** Newton is an open-source framework, allowing roboticists to freely use, distribute, and contribute to it.
- **NVIDIA-Accelerated:** Newton is built on NVIDIA Warp, offering a high-performance framework for physics-based simulations using NVIDIA GPUs
- **Powered by MuJoCo Warp:** Warp integration in MuJoCo boosts performance, achieving over 70X acceleration for humanoid simulations and 100X for in-hand manipulation tasks.
- **Differentiable Physics:** Generates forward-mode results and computes reverse-mode gradients for back-propagation and system optimization.

- **Extensible:** Newton supports rich multiphysics simulations with custom solvers for interactions with deformable objects
- **Built on OpenUSD:** Newton uses the OpenUSD framework. OpenUSD's flexible data model and composition engine aggregates data needed for describing robots and their surrounding environments

Pricing: N/A

Availability: Later this year

External Resources: Newton Tech Blog, Isaac Platform Updates Tech Blog

GTC Session: Physical AI for Humanoids: How Google Robotics Uses Simulation to Accelerate Humanoid Robotics Training [S72709]

Healthcare and Life Sciences

Digital Devices

Healthcare Robotics Announcement

At GTC, we are announcing NVIDIA and GE HealthCare are collaborating to advance development of autonomous diagnostic imaging with physical AI. Isaac for Healthcare is an AI robotic development platform of AI models, simulation frameworks and synthetic data generation pipelines, and accelerated runtime libraries purpose-built for healthcare robots. It enables developers to train, simulate, test, and deploy robots from digital environments into the physical world.

- GE HealthCare will leverage Isaac for Healthcare to fast-track its development of autonomous imaging systems for ultrasound and x-ray - the most common and widely used today.
- Isaac for healthcare is a domain-specific robotic development platform that leverages NVIDIA's three computers: MONAI for medical AI development, Ominiverse for simulating photo-realistic human anatomy, medical devices, and sensors; Holoscan for real-time sensor processing and deployment to real-world healthcare robots
- Isaac for Healthcare will offer at launch two reference workflows - surgical subtask automation and robotic autonomous ultrasound.

Availability: Isaac for Healthcare is available in early access

External Resources: Press Release, Isaac for Healthcare Demo Video, Isaac for Healthcare Dev Blog, Isaac for Healthcare EA Signup, Medical Devices Webpage

GTC Sessions:

- Advancing the Frontiers of Medical Devices with Physical AI [S72948]
- Accelerate the Future of Healthcare: AI-Powered Innovations for Imaging, Digital Health, and Beyond [S72493]
- Shaping the Future of Robotic Surgery: Innovations, Challenges, and Opportunities [S72930]
- Making Physical AI a Reality through Real-Time Edge Computing [S73298]

Holoscan 3.0 Product Release

NVIDIA Holoscan is the real-time AI sensor processing platform. This latest version brings dynamic flow control, empowering developers to build more robust, scalable, and efficient solutions for the edge.

- With physical AI rapidly evolving, Holoscan 3.0 is built to adapt alongside it, making it easier than ever to tackle the challenges of today's dynamic environments.
- Dynamic Flow Control reduces the need for complex code. With this feature, developers can now modify operator connections within a pipeline at runtime, enabling more flexible and adaptive workflows.

Availability: Tuesday, March 18

External Resources: GPU Genius Deck: Holoscan 101, New End to End Reference Application for Surgical AI Video Pipeline, Medical Devices Webpage, Holoscan 3.0 Dev Blog (Live 3/20)

GTC Sessions:

- Digital and Physical AI Helps Write a New Chapter in Medicine [S71353] - Kimberly Powell Special Address
- Advancing the Frontier of Medical Devices with Physical AI [S72948] - Prerna Dogra

MONAI Multimodal

MONAI Multimodal is an open-source toolkit of foundation models, reference workflows, and interoperable building blocks for enabling multimodal analysis of diverse healthcare data—from CT and MRI to EHRs and clinical documentation. It delivers advanced reasoning capabilities through specialized agentic architectures, and allows for the integration of custom models and Hugging Face components. MONAI Multimodal key features include:

- Agentic AI Framework: Utilizing autonomous agents for multi-step reasoning across images and text.
- Specialized LLMs and VLMs: Tailored models designed for medical applications that support cross-modal data integration.
- Data I/O components: Integrating diverse data readers like DICOM for medical imaging, EHR for clinical data, Video for surgical recordings, WSI for pathology images, Text for clinical notes, and Image for static images.

Pricing: Free

Availability: Available at GTC

External Resources: MONAI Multimodal Dev Blog (live 3/19), Updated MONAI 101 Deck, MONAI Multimodal README on GitHub, MONAI M3 on GitHub, Medical Imaging Webpage

GTC Sessions:

- Digital and Physical AI Helps Write a New Chapter in Medicine [S71352] - Kimberly Powell Special Address
- Advancing the Frontier of Medical Devices with Physical AI [S72948] - Prerna Dogra
- Accelerate the Future of Healthcare: AI-Powered Innovations for Imaging, Digital Health, and Beyond [S72493] – GE Healthcare
- Federated Learning in Medical Imaging: Enhancing Data Privacy and Advancing Healthcare [S73112] – University of Wisconsin-Madison
- Advanced Medical AI Development with MONAI: From Interactive Annotation to Foundation Models [DLIT71265] – Workshops & Trainings

Digital Biology

NVIDIA Parabricks v4.5 and NVIDIA AI Blueprints for Genomics and Single-Cell Analysis

NVIDIA Parabricks is a scalable genomics analysis software suite for secondary analysis that provides GPU-accelerated versions of trusted, open-source tools. The latest v4.5 release of Parabricks includes:

- Parabricks support for NVIDIA Blackwell, including the new NVIDIA RTX PRO 6000 Blackwell Server Edition GPU
- The ability to easily combine Giraffe and DeepVariant, combining the power of pangenome analysis with the gold standard for variant calling
- Improved features:
    - STAR acceleration (2x acceleration over existing speed)
    - Faster FQ2BAM (recovery mode improvements)
    - Faster force-calling mode in Haplotypecaller/Mutectcaller
    - Faster Giraffe (3.7x acceleration over existing speed)
    - Minimap2 acceleration (including splice-alignment)

Parabricks v4.5 is accompanied by new AI blueprints for genomics and single-cell analysis, enabling bioinformaticians and genomics platform providers to easily deploy and test NVIDIA Parabricks and NVIDIA RAPIDS without requiring local GPUs or self-managed cloud provisions

Availability: Available at GTC.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Parabricks v4.5 Blog (live 3/19), RTX PRO 6000 Corporate Blog, Single-Cell Analysis Blueprint (RAPIDS-singlecell), Genomics Analysis Blueprint (Parabricks), New Genomics Landing Page, Updated Genomics 101 Deck

GTC Sessions:

Digital and Physical AI Helps Write a New Chapter in Medicine [S71353] - Kimberly Powell Special Address

Enabling a Future of Personalized Healthcare: Tackling Data Rates [S72071] - Roche

Accelerating Genomic Analysis: Building the Foundation for Next-Generation Genomic Models in Healthcare [S72432] - Illumina

Interpret Health Data With AI to Change the Future of Medicine [S71671] - SOPHiA GENETICS

Performance-Optimized CUDA Kernels for Inference with Small Transformer Models - SOPHiA GENETICS

Performance-Optimized CUDA Kernels for Inference With Small Transformer Models [S73168] - Oxford Nanopore Technologies

New BioNeMo NIM microservices and updated BioNeMo Blueprint for generative virtual screening

NVIDIA BioNeMo is a platform for accelerating the development and deployment of AI models for the most time-consuming and costly stages of computational drug discovery.

The latest updates include:

- New NVIDIA BioNeMo NIM microservices: OpenFold2 NIM, MSA-Search NIM
- Updated NVIDIA BioNeMo Blueprint for generative virtual screening to include GenMol NIM, OpenFold2 NIM, MSA-Search NIM.
- BioNeMO support for NVIDIA Blackwell, including the new NVIDIA RTX PRO 6000 Blackwell Server Edition GPU.
    - For example, RTX PRO 6000 Blackwell Server Edition GPUs boost inference performance of OpenFold2 by up to 4.8x compared with L40S GPUs.
    - Also, B200 with FP8 shows a 2x improvement for Evo 2 pretraining on a single node (8 GPU) over H100 with FP8, and a 5x improvement over A100 with BF16.

Availability: Available at GTC.

Enterprise Support: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.

External Resources: Tech Blog: Guiding Generative Molecular Design with Experimental Feedback Using Oracles, RTX PRO 6000 Corporate Blog (Live 3/19), NVIDIA BioNeMo Blueprint for Generative Virtual Screening Page, Updated BioNeMo 101 Deck, Biopharma Webpage

GTC Sessions:

- Digital and Physical AI Helps Write a New Chapter in Medicine [S71353] - Kimberly Powell Special Address
- Toward Rational Drug Design With AlphaFold 3 and Beyond [S72684] - Isomorphic Labs
- Leveraging AI for Digital Biology: BioNeMo, Parabricks, and NVIDIA NIMs and Agent Blueprints [CWE71519] - NVIDIA

Digital Health

New Digital Health Agent Blueprint for Biomedical AI Research

We are introducing the Biomedical AI Research Agent (AIRA) Blueprint, an advanced AI-powered system designed to streamline literature reviews in biomedical research. Biomedical AIRA builds on the AI-Q Blueprint by integrating a dataset of biomedical literature PDFs, enabling researchers to efficiently analyze scientific findings.

- The ability to connect real-world data to reasoning models unlocks new opportunities to improve the efficiency and accuracy of various clinical development processes
- Reasoning LLMs enable state-of-the-art planning, validation, and insight generation capabilities during the research process, while agentic frameworks AgentIQ allow accelerated agent workflows.

Availability: Many blueprints and NIMs are available now, with more to come.

The Biomed AIRA Blueprint will be GA in May.

External Resources: Digital Health Web Page

GTC Sessions:

- Digital and Physical AI Helps Write a New Chapter in Medicine [S71353] - Kimberly Powell Special Address
- Clinical Research Agents Booth Demo [B021]
- Harnessing NVIDIA's Advanced Tools for Generative AI in Digital Health and Clinical Trials [DLIT71262]
- Build an AI Research Assistant With NVIDIA AI Blueprints [DLIT74486]
- Digital Health Session 1 [Brad Genereaux's Session]

Healthcare Reimagined: Coding the Future of Medicine Through Digital Acceleration [S72847]

Telecommunications

At GTC, NVIDIA Telecoms will be making announcements about developing AI-Native Wireless Networks for 6G with new partners, new Aerial research and commercialization tools to support that in addition to a demo about neural networks showcasing how future AI-Native wireless networks can provide extreme spectral efficiency. In the Gen-AI for Operations track, new Large Telco Models and Network AI Agents will be announced. In the AI Factories domain, a new customer will be announcing AI Factory based on NVIDIA platforms.

NVIDIA and Telecom Industry Leaders to Develop AI-Native Wireless Networks for 6G

NVIDIA is collaborating with T-Mobile, Cisco, MITRE, ODC and Booz Allen to develop an AI-native wireless network stack based on the NVIDIA AI Aerial platform which provides software-defined RAN on the NVIDIA accelerated computing platform.

- AI will be fully embedded into the network stack's software and hosted over a unified accelerated infrastructure, capable of running both network and AI workloads.
- End-to-end security and an open architecture to foster rapid innovation will be at the core of the solution.

External Resources: Press Release

GTC Sessions:

- The Telco AI Renaissance Is Here [S72984]
- AI-RAN in Action [S72987]
- Driving 6G Development With Advanced Simulation Tools [S72994]
- Defining AI-Native RAN for 6G [S72985]
- Pushing Spectral Efficiency Limits on CUDA-accelerated 5G/6G RAN [S72990]

NVIDIA Aerial Expands with New Tools for Building AI-Native Wireless Networks

New developer tools and ecosystem partnership enabling research breakthroughs in AI-RAN and 6G.

- NVIDIA announced significant advancements in its Aerial Research platform, unveiling new tools that equip researchers, developers and telecom leaders to pioneer AI-native wireless networks.
- With an expanded portfolio of solutions — the Aerial Omniverse Digital Twin on DGX Cloud, the Aerial Commercial Test Bed on the NVIDIA MGX, NVIDIA Sionna 1.0 open source library, the Sionna Research Kit on the NVIDIA Jetson — NVIDIA is accelerating AI-RAN and 6G research and enabling transformative solutions in radio signal processing.
- Industry leaders and 150+ higher education and research institutions from the U.S. and around the world are harnessing the NVIDIA Aerial Research platform to develop, train, simulate and deploy groundbreaking AI-native wireless innovations.

External Resources: Corp Blog, AODT Tech Blog (live on 3/19)

GTC Sessions:

- The Telco AI Renaissance Is Here [S72984]
- AI-RAN in Action [S72987]
- Driving 6G Development With Advanced Simulation Tools [S72994]
- Defining AI-Native RAN for 6G [S72985]
- Pushing Spectral Efficiency Limits on CUDA-accelerated 5G/6G RAN [S72990]

Telecom Leaders Call Up Agentic AI to Improve Network Operations

Amdocs, BubbleRAN, ServiceNow, SoftBank and Tech Mahindra harness NVIDIA AI Enterprise to develop large telco models and new network AI agents.

- SoftBank has developed a new LTM NIM. The 70B parameter LTM is trained on the NVIDIA DGX super pod using 4TB of its own network data covering 200K cells.
- Amdocs' Network Assurance Agent automates repetitive tasks such as fault prediction. Amdocs Network Deployment Agent simplifies open radio access network (RAN) adoption by automating integration, deployment tasks and interoperability testing, and providing insights to network engineers
- Tech Mahindra developed an LTM leveraging NVIDIA AI Enterprise to help address critical network operations, focused network observability and incident resolution
- BubbleRAN launched a Network Slicing agent using NVIDIA AI Enterprise, for Telenor, implementing a maritime use-case

External Resources: Corp Blog

GTC Session: AI Agents and Digital Humans Shaping the Future of Interaction in Telecoms [S72988]

Energy

NVIDIA and Electric Power Research Institute Launch Open Power AI Consortium to Transform the Future of Energy

Global consortium brings together utilities, technology companies, academia and more to build open AI models to transform the way we make, move and use electricity.

- Led by EPRI, the Open Power AI Consortium includes energy companies, technology companies and researchers developing AI applications to tackle domain-specific

challenges, such as adapting to an increased deployment of distributed energy resources and significant load growth on electric grids.

- As part of the consortium, EPRI, NVIDIA and Articul8, a member of the NVIDIA Inception program for cutting-edge startups, have developed a set of domain-specific, multi-modal large language models trained on massive libraries of proprietary energy and electrical engineering data from EPRI that can help utilities streamline operations, boost energy efficiency and improve grid resiliency.
- The domain-specific model—currently based on nearly 10,000 EPRI files, over 400,000 images, and around 230,000 tables—was developed using hundreds of NVIDIA H100 GPUs and is now available from EPRI.

Availability: Available from EPRI. Future availability as NIM microservice on build.nvidia.com

External Resources: Blog Post (Live 3/20)

GTC Sessions:

- Accelerate Energy Transformation With Industry Domain AI Models [S72520]
- Energy Transition: Impact of Generative AI in the Power Ecosystem of Generation, Transmission and Distribution [S73745]

Automotive

General Motors and NVIDIA Collaborate on AI for Next-Generation Vehicle Experience and Manufacturing

NVIDIA and General Motors are collaborating on next-generation vehicles, factories and robots using AI, simulation and accelerated computing. The companies will work together to build custom AI systems using NVIDIA accelerated compute platforms, including NVIDIA Omniverse™ with NVIDIA Cosmos™, to train AI manufacturing models for optimizing GM's factory planning and robotics. GM will also use NVIDIA DRIVE AGX™ for in-vehicle accelerated compute for advanced driver-assistance systems and in-cabin enhanced safety driving experiences.

External Resources: Press Release

GTC Session: How AI is Fueling Innovation in Automotive Software and Manufacturing [S74936]

NVIDIA Launches Halos, a Full-Stack Comprehensive Safety System for Autonomous Vehicles

NVIDIA Halos is a full-stack, comprehensive safety system that unifies vehicle architecture, AI models, chips, software, tools, and services to ensure the safe development of autonomous vehicles (AVs) from cloud to car. It applies design, deployment, and validation guardrails using

NVIDIA DGX™ for AI training, NVIDIA Omniverse™ with Cosmos™ for simulation, and NVIDIA AGX™ for deployment, ensuring explainability, regulatory compliance, and reliability in AV stacks.

Availability: Now

External Resources: Blog, Demo

GTC Sessions:

- Introduction to NVIDIA's Strategy for AV Safety [S74743]
- Guardrails for AV Safety Across the Product Life Cycle [S74744]
- Safety Regulation and Standardization in the Era of AI-Based AVs [S74745]
- Navigating the High-Stakes Safety Challenges of Autonomous Driving [S73122]

NVIDIA Expands Automotive Ecosystem to Bring Physical AI to the Streets

Leading global automakers, mobility innovators, suppliers and software providers harness NVIDIA accelerated computing to deliver AI from cloud to car.

External Resources: Blog