# GTC 2024

**KEY ENTERPRISE ANNOUNCEMENTS**

- NVIDIA Blackwell: Platform, Architecture, Systems, and Networking
- NVIDIA AI including NIMs and Nemo Retriever
- CUDA-X Microservices, RAPIDS cuDF Accelerated pandas, and More
- NVIDIA Omniverse Cloud APIs, Earth-2, DRIVE, and Robotics
- Additional Announcements

---

**NVIDIA Blackwell**

NVIDIA Blackwell is our newest platform for building and deploying Trillion-Parameter Scale Generative AI. Relative to its predecessor, Blackwell achieves: **30X** more AI Inference Performance, **4X** faster AI Training, **25X** lower energy use, and **25X** lower TCO. It introduces groundbreaking advancements for accelerated computing through six technological innovations (below). Blackwell GPUs are available in three system configurations: GB200 NVL72, HGX B200, and HGX B100 (see sections below).

- **AI Superchip:** 208B Transistors built from two of the largest possible die
- **2nd Gen Transformer Engine:** Enabling FP4/FP6 Tensor Core for double the performance and model size
- **5th Generation NVLink:** Scales up to 576 GPUs in a single NVLink domain
- **RAS Engine:** 100% In-System Self-Test for reliability for GPU at scale AI
- **Secure AI**: Full Performance Encryption & TEE, Confidential compute without compromise
- **Decompression Engine:** 800GB/sec of performance to compute on compressed data without CPU decompression time

**Availability:** Later this year

**Resources:** Blackwell Architecture Webpage, HGX Webpage, Press Release

**GTC Session:** Jensen’s Keynote

---

**NVIDIA Blackwell Systems: GB200 NVL72**

GB200 NVL72 is our next generation NVLink rack scale architecture for Trillion Parameter Scale AI. It consists of 36 **GB200 Grace Blackwell Superchips** each with two Blackwell GPUs connected to one Grace CPU for a total of 72 Blackwell GPUs in a single NVLink domain. GB200 NVL72 is a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. The platform acts as a single GPU with 1.4 exaflops of AI performance and up to 30TB of fast memory, and is a building block for the newly announced DGX SuperPOD.

- **GB200 Superchip Compute Nodes:** Contains single or dual GB200 Superchips, PCIe gen 6 slots, and includes NVIDIA BlueField®-3 DPUs for cloud network acceleration, composable storage, zero-trust security, and GPU compute elasticity in hyperscale AI clouds. 18 compute nodes fit in a GB200 NVL72 rack.
- **NVLink Switch Tray:** Contains two NVLink Switch Chips per switch tray in GB200 NVL72. Nine switch trays in GB200 NVL72 provide 130TB/s total bandwidth for GPU communications. 72 switch trays create a 576 GPU NVLink domain with 1PB/s total bandwidth.

- **NVLink Switch Chip:** Creates the highest performance compute fabric for large AI and HPC models. Provides 72 NVLink ports at 100GB/s per chip, encryption for 128 GPUs in a Confidential Computing domain, and is managed by Unified Fabric Manager.

**Availability:** Later this year
**Resources**: GB200 NVL72 Webpage, NVLink/ NVLink Switch Webpage, NVIDIA GB200 NVL72 Delivers Trillion-Parameter LLM Training and Real-Time Inference
**GTC Session:** Jensen's Keynote

---

**NVIDIA Blackwell Systems: HGX B200 & HGX B100**

HGX B200 & HGX B100 are premier accelerated x86 scale-up platforms designed for the most demanding generative AI, data analytics, and high-performance computing (HPC) workloads

- **HGX B200:** HGX B200 delivers the best performance (15X more AI inference perf than HGX H100) and TCO (12X less than HGX H100) for x86 scale-up platforms and infrastructure.
- **HGX B100:** HGX B100 is designed for the fastest time to deployment with drop-in replacement compatibility for existing HGX H100 infrastructure.

**Availability:** Later this year
**Resources:** HGX Webpage, NVLink/NVSwitch Webpage
**GTC Session:** Jensen's Keynote

---

**NVIDIA DGX SuperPOD with DGX GB200 Systems**

DGX SuperPOD with DGX GB200 systems is liquid-cooled, rack-scale AI infrastructure with intelligent predictive management capabilities that scales to thousands of NVIDIA GB200 Grace Blackwell Superchips for training and inferencing trillion parameter generative AI models.

- High compute density, liquid-cooled, rack-scale design
- 36 GB200 Grace Hopper Superchips with 36 Grace CPUs and 72 Blackwell GPUs per rack, connected via fifth-generation NVLink
- Scalable to thousands of GB200 Superchips with InfiniBand
- Full-stack resiliency features to deliver constant uptime

**Availability:** Late 2024
**Enterprise Support**: Enterprise Support is available for NVIDIA DGX SuperPOD with DGX GB200 systems.
**Resources:** Press Release, Webpage, Datasheet
**GTC Sessions:**

- Solving the Generative AI Infrastructure Challenge in 2024 [S62227]
- The Next-Generation DGX Architecture for Generative AI [S62421]

---

**NVIDIA DGX B200**

The latest iteration of NVIDIA's legendary DGX systems, NVIDIA DGX B200 is powered by the groundbreaking NVIDIA Blackwell GPUs. DGX B200 is the proven choice for enterprises, a unified AI platform that accelerates any AI workload, from training, to fine-tuning, to inference.

- The world's first system with the NVIDIA Blackwell GPU, featuring 8 GPUs with a total of 1,440GB of GPU memory
- 3X more training performance, 15X more inference performance compared to DGX H100
- 4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 400Gbp/s InfiniBand and Ethernet, and 2x dual-port NVIDIA BlueField-3 DPUs
- DGX B200 systems are the building blocks of the next-generation NVIDIA DGX POD and NVIDIA DGX SuperPOD

**Availability:** DGX B200 will ship in late 2024
**Enterprise Support**: Enterprise Business Standard Support is included
**Resources:** Press Release, Webpage, Datasheet
**GTC Session:** Solving the Generative AI Infrastructure Challenge in 2024 [S62227]

---

**NVIDIA Quantum-X800 InfiniBand & Spectrum-X800 Ethernet Networking Platforms**

Quantum-X800 and Spectrum-X800 are NVIDIA's networking platforms for powering next-generation AI systems based on the Blackwell compute architecture.

- NVIDIA Quantum-X800 InfiniBand for Highest-Performance AI-Dedicated Infrastructure with inherent low latency and high effective bandwidth, as well as features such as in-network computing
- NVIDIA Spectrum-X800 Ethernet for AI-Optimized Networking in Every Data Center ideal for environments such as multi-tenant AI clouds and large enterprise hyperscale deployments

**Availability:** GA is expected end of 2024
**Resources:** Quantum-X800 Solution Sheet, Spectrum-X800 Solution Sheet, Press Release
**GTC Sessions:**

- Entering a New Frontier of AI Networking Innovation [S62293]
- Best Practices in Networking for AI: Perspectives from Cloud Service Providers [S62447]
- Enabling Enterprise Generative AI with Optimized Ethernet AI Networking [S62521]

---

**NVIDIA AI Enterprise 5.0**

NVIDIA AI Enterprise, an end-to-end, cloud-native software platform that brings generative AI within reach for every business, delivers the highest performance and most efficient runtime. NVIDIA AI Enterprise 5.0 includes easy-to-use microservices with enterprise-grade security, support, and stability to accelerate LLM inference and time to value.

- **NVIDIA NIM**, exclusive to NVIDIA AI Enterprise, is a set of easy-to-use microservices designed to accelerate deployment of generative AI models across cloud, data center, and workstations. Supporting a wide range of AI models, including NVIDIA AI Foundation and custom models, it ensures seamless, scalable AI inferencing, on-premises or in the cloud, leveraging industry standard APIs.
- **NVIDIA cuOpt**, a microservice to accelerate logistics and supply chain optimization to save time and reduce infrastructure costs.
- **NVIDIA API Catalog** provides quick access for enterprise developers to experiment, prototype and test NVIDIA-optimized foundation models powered by NIM at no initial cost. When ready to deploy, enterprise developers can export the enterprise-ready API with a valid NVIDIA AI Enterprise license, and run on a self-hosted system anywhere.
- **NVIDIA AI Workbench** is part of the enterprise platform and supported with NVIDIA AI Enterprise

- **Infra 5.0** brings the latest enhancement of infrastructure software and support for the new accelerated hardware, including support for Red Hat OpenStack Platform, Canonical Charmed Kubernetes, vGPU heterogeneous profiles, NVIDIA GH 200 96GB, GH200 144GB, HGX H100, HGX H200, and RTX 5880 Ada.

**Availability:** Now available
**Enterprise Support**: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.
**Resources:** Corp Blog, Web Page
**GTC Sessions:**

- Overcome the Complexities of Generative AI with a Secure, Scalable AI Foundry [S62953]
- What's Next in Generative AI [S62430]
- An AI Revolution in Insurance Claim Process [S62284]
- Unlock AI's Potential: Best Practices for Business-Led Digital Roadmaps and Implementation Challenges [S62432]
- Building an End-to-End Solution for Enterprise-Ready Generative [S62620]
- Unlock the Power of Generative AI in a Multi-Cloud Environment [S62013]
- A Guide to Building Safe Generative AI Copilots that Improve Productivity and Protect Company Data [S62954]

---

**NVIDIA NIM** *Available with NVIDIA AI Enterprise license*
NVIDIA NIM, part of NVIDIA AI Enterprise, is a set of easy-to-use microservices designed to accelerate deployment of generative AI models across cloud, data center, and workstations. Supporting a wide range of AI models, including NVIDIA AI Foundation and custom models, it ensures seamless, scalable AI inferencing, on-premises or in the cloud, leveraging industry standard APIs.

Many developers are running POCs doing inferencing on foundation models, but they don't have a way to deploy in production. NIM opens the path to production and scale of AI inference through a unique set of benefits:

- **Deploy anywhere** and maintain control of generative AI applications and data
- **Streamline AI application development** with industry standard APIs and tools tailored for enterprise environments
- **Prebuilt containers for the latest generative AI models**, offering a diverse range of options and flexibility right out of the gate
- **Industry-leading latency and throughput** for cost-effective scaling
- **Support for custom models** out of the box so models can be trained on domain specific data
- **Enterprise-grade software** with dedicated feature branches, rigorous validation processes, and robust support structures

Learn more and experiment with models through ai.nvidia.com and download NVIDIA NIM inference microservices to deploy models on your own infrastructure from the API catalog.

**Availability:** Microservices are available March 18, 2024 at ai.nvidia.com
**Enterprise Support**: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.
**Resources:** NVIDIA NIM Offers Optimized Inference Microservices for Deploying AI Models at Scale, Generative AI Microservices Press Release

**GTC Session:** Accelerating Enterprise: Tools and Techniques for Next-Generation AI Deployment [S63432]

---

**DGX Cloud**

DGX Cloud underpins NVIDIA's most important strategies, and significant work is happening to make it a truly cloud-first service that can power everything we do. Due to overwhelming demand from customers, we are currently capacity constrained, but are urgently building more capacity across all our CSP partners. Our announcements around Blackwell coming to DGX Cloud are the first proof point of this commitment by NVIDIA and our valued CSPs.

- GB200 NVL72 on AWS, Google Cloud, OCI, and YTL
- H100 DGX Cloud Instances on GCP
- Hugging Face Training-as-a-Service on DGX Cloud

**Resources:** Webpage

**GTC Sessions:**

- Customizing Generative AI with Your Own AI Foundry [S61968]
- Enable Hybrid Training and Inference With DGX Cloud and OCI GPU Infrastructure (Presented by Oracle) [S63030]
- Building Accelerated AI With Hugging Face and NVIDIA [S63149]
- Accelerating the Generative AI Transformation: Expert Insights for Rapid Innovation and Scale [S62494]

---

**NVIDIA cuOpt** *Available with NVIDIA AI Enterprise license*

NVIDIA announced the general availability of cuOpt. With 23 world-record benchmarks, NVIDIA cuOpt owns 100% of the world records on the largest routing benchmarks from the past three years. Operational research, logistics, supply chain, and industrial teams can now harness accelerated optimization to save time and reduce infrastructure costs. With billions of potential feasible moves to consider, NVIDIA cuOpt allows teams to unlock new use cases, reimagine solutions, find new routes, and discover possibilities previously unknown.

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

**Availability:** Available now in the API catalog

**Resources:** NVIDIA cuOpt Webpage, Route Optimization AI Workflow Webpage, New World Records Corporate Blog, World Record Technical Blog, NVIDIA Combines Digital Twins With Real-Time AI for Industrial Automation

**GTC Sessions:**

- Connect with Optimization AI Experts [CWE62383]
- DLI – Route Optimization [DLIT62051]
- Exploring Use of Accelerated Optimization AI for Route Planning [S62473]
- Advances in Optimization AI [S62495]
- Increase Safety, Save Time & Money [S61398]

---

**ai.nvidia.com and NVIDIA API Catalog** *Available with NVIDIA AI Enterprise license*

NVIDIA launched NVIDIA API catalog - a collection of performance optimized API endpoints packaged as enterprise-grade runtime that developers can experience, build POCs, and easily deploy in production anywhere.

- Leading community, proprietary, and NVIDIA-built models, optimized for highest performance on NVIDIA GPU-accelerated systems
- Enterprise developers can test the models with NVIDIA NIM at scale on NVIDIA-hosted API endpoints
- Integrated with popular frameworks like Langchain, LlamaIndex, and Haystack
- Models packaged as microservices to easily deploy in production on any Kubernetes platform w/ NVIDIA AI Enterprise

**Availability:** 30 NIM-supported models available in GA now
**Resources:** ai.nvidia.com
**GTC Sessions:**

- Power AI Anywhere With Google Gemma [GENAI63700]
- Mixtral of Experts Explained [GENAI63693]
- Scale AI in Production [GENAI63694]
- Fast-track AI Developments [GENAI63695]

---

**NVIDIA AI Workbench** *Available with NVIDIA AI Enterprise license*
NVIDIA AI Workbench is a toolkit that enables AI and machine learning developers with easy GPU environment setup and the freedom to work, manage, and collaborate across heterogeneous platforms regardless of skill-level.

- Offers streamlined setup and configuration for running AI and ML projects on GPU-accelerated hardware.
- Automates workflows across heterogeneous infrastructure, enabling developers to reproduce, collaborate, and migrate across any compute resources of their choice.
- Allows for greater productivity. Both novice and skilled developers can focus on execution without worrying about configuration and management overhead.

New features since the beta release include:

- **Visual Studio (VS) Code Support:** Directly integrated with VS Code to connect to Project containers for interactive development.
- **Choice of base images:** Users can choose their own container image as the project base image when creating projects. Base images need to follow a specification for image labels.
- **Improved package management:** Users can manage and add packages directly to containers through the AI Workbench user interface.
- **Installation improvements:** Users have an easier install path on Windows and MacOS. There is also improved support for the Docker container runtime.
- **New AI Workbench example projects:** Hybrid Retrieval Augmented Generation (RAG) example project makes it easy to chat with your documents. Customize a Llama-2 Project - fine-tune and run anywhere.

**Enterprise Support**: NVIDIA Enterprise Support is available with an NVIDIA AI Enterprise license
**Availability:** GA release available now at www.nvidia.com/workbench. Also, gain immediate, short-term access to try AI Workbench example projects on NVIDIA Launchpad.
**Resources:** NVIDIA AI Workbench Webpage, GA Announcement Blog, Visit NVIDIA on GitHub for example Workbench Projects to get started faster.
**GTC Sessions:**

- Breaking Barriers: How NVIDIA AI Workbench Makes AI Accessible to All [S62135]
- Getting the Most from NVIDIA AI Workbench for Generative AI Workflows [DLIT61929]
- Accelerate Your Developer Journey with Dell AI-Ready Workstations and NVIDIA AI Workbench [SE62279]

---

**NVIDIA NeMo Retriever for Retrieval-Augmented Generation** *Available with NVIDIA AI Enterprise license*
NVIDIA AI Enterprise provides the tools to enable enterprises to take their retrieval-augmented generation (RAG) applications from pilot to production. NVIDIA NeMo Retriever is a collection of generative AI microservices enabling organizations to seamlessly connect custom models to diverse business data and deliver highly accurate responses. NeMo Retriever enhances generative AI applications with enterprise-grade RAG capabilities which can be connected to business data wherever it resides.

With a NVIDIA AI Enterprise subscription, organizations have access to software and tools they need to build and deploy enterprise-ready RAG applications.

**Availability**: Customers can apply for early access
**Resources**: NeMo Retriever Tech Blog, 4 Steps for Taking your RAG Application from Pilot to Production Tech Blog, Generative AI Microservices Press Release, RAG Networking Tech Blog, Intro to Multimodal RAG TechBlog
**GTC Sessions:**

- Connect With the Experts: Building Generative AI Applications With Retrieval Augmented Generation [CWE62682]
- Streamlining Enterprise Data Operations with Multimodal RAG and LangChain [DLIT61342]
- Beyond RAG Basics: Building Agents, Co-Pilots, Assistants, and More! [S62533]
- Build a RAG-Powered Application With a Human Voice Interface [SE62869]
- Accelerating Enterprise: Tools and Techniques for Next-Generation AI Deployment [S63432]
- RAG Targeted Agenda

---

**NVIDIA NeMo Development Microservices** *Available with NVIDIA AI Enterprise license*
NeMo is an end-to-end platform for developing custom generative AI. To simplify building custom generative AI, we're announcing an early access program for NeMo Curator, NeMo Customizer, and NeMo Evaluator microservices - covering all development stages, from data curation and customization to evaluation.

- Part of CUDA-X microservices, the NeMo API endpoints are built on top of the NVIDIA libraries to provide the easiest path for enterprises to get started with building custom generative AI.
- NeMo Curator is a scalable and GPU-accelerated data-curation microservice that prepares high-quality datasets for pre-training and customizing generative AI models.
- NeMo Customizer is a high-performance, scalable microservice that simplifies fine-tuning and alignment of LLMs for domain-specific use cases.
- The NeMo Evaluator microservice provides automatic assessment of custom generative AI models across diverse academic and custom benchmarks on any cloud or data center.

**Availability:** Customers who want to gain early access can fill out the form
**Resources:** Tech Blog

**GTC Session:** Navigating the Large Language Models Frontier: Practical Strategies for Building Enterprise Applications Powered by LLMs [S62752]

---

**NVIDIA Maxine** *Available with NVIDIA AI Enterprise license*

The NVIDIA Maxine Developer Platform is set to transform the $10 Billion Video Conferencing Industry. Maxine enables developers to easily integrate AI features that turn commodity cameras and microphones into professional studios.

- Background Noise Removal 2.0 EA
- Studio Voice EA
- Speech Live Portrait EA
- Relighting EA
- Eye Contact UPDATE – Production
- Cloud Function APIs
- Maxine on AI Foundations
- Partner Ecosystem Expansion

**Availability:** Production features available now via NVIDIA AI Enterprise, EA features available via NVIDIA developer account on NGC

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

**Resources:** Maxine Webpage, Customer Deck, NVIDIA Maxine Blogs, NVIDIA AI Enterprise 4.1, FAQs

**GTC Sessions:**

- Using AI to Enhance Immersion and Communication in Video Conferencing [S62554]
- Secure AI-Driven Translation in Video Conferencing [S61718]
- Connect With the Experts: NVIDIA Maxine — Generative AI for Video, Audio, and AR [CWE62232]

---

**Picasso** *Available with NVIDIA AI Enterprise license*

NVIDIA Picasso is powering generative AI services from leading visual design service providers.

**Availability**: Available now

**Resources**: Picasso Webpage

**Key GTC Sessions**:

- Edify Model for Visual Content Generation [S62832]
- Unlocking Creative Potential: The Synergy 0f AI and Human Creativity [S62681]
- Artists Exploring Creativity and Language with Generative AI [S62958]
- Building a Generative Service to create 3D Assets from Text [S62683]
- Connect With Experts Picasso [CWE63132]

---

**NVIDIA Morpheus** *Available with NVIDIA AI Enterprise license*

At GTC, NVIDIA is demonstrating how event-driven RAG, powered by Morpheus, NVIDIA NIM, and NeMo Retriever, can be used to rapidly analyze and address common vulnerabilities and exposures (CVEs) in seconds versus days. NVIDIA is collaborating with leading cybersecurity and systems integrator partners to enhance cybersecurity with generative AI.

**Availability:** Available now on NGC and GitHub
**Resources:** NVIDIA Morpheus Web Page, Cybersecurity Web Page, Digital Fingerprinting Workflow Web Page, Spear Phishing Workflow Web Page, GTC 2024 Corporate Blog
**GTC Sessions:**

- Top Cybersecurity Conference Sessions
- Cybersecurity Developer Day [SE62821]

---

**NVIDIA Riva** *Available with NVIDIA AI Enterprise license*=
NVIDIA announced the following updates for Riva:

- Automatic speech recognition (ASR) now includes the fastest and most accurate models that top HuggingFace Open ASR Leaderboard—Canary, a multilingual (English, French, German & Spanish) and multitasking ASR and bidirectional translation model and English transcription model from Parakeet ASR family—and English and Spanish/Mandarin/Japanese language code-switch ASR models
- Text-to-speech (TTS) now offers a P-Flow model that enables creating a custom voice with 3 seconds of audio sample for enterprises only (model won the LIMMITS'24 Challenge) and fe/male voices in 5 languages—English, German, Italian, Mandarin, Spanish—with voice emotion (happy, calm, neutral, sad, fearful, angry) adjustment based on context
- Translation supports speech-to-text and speech-to-speech APIs and text-to-text translation for up to 32***** languages, as well as model customizations for domain-specific use cases
- Voice-powered retrieval-augmented generation (RAG) sample application for Q&A chatbot

**Availability:** Available in NVIDIA API catalog, on major CSP Marketplaces (Azure, AWS, GCP), and anywhere with NVIDIA NIM —cloud, data center, workstation, PC.
**Resources:** NVIDIA AI Enterprise Customer Deck, NVIDIA Riva Webpage, Audio Transcription Webpage, Intelligent Virtual Assistant Webpage
**GTC Sessions:**

- Speaking in Every Language: A Quick Start Guide to TTS Models for Accented, Multilingual Communication [S62517]
- Adapting Conformer-Based ASR Models for Conversations Over the Phone [S62441]
- Secure AI-Driven Translation in Video Conferencing [S61718]
- Multi-Speaker ASR with NVIDIA NeMo Toolkit —Training & Inference [CWE62255]
- Mastering Speech for Multilingual Multimedia Transformation [S62549] & [S62549a]
- Talk to Your Data in Your Native Language [DLI61469]
- Behind the Scenes of Running a Conversational Character in a 3D Scene [S62570]
- Build Speech AI for Multilingual Multimedia Transformation [SE62869]
- Speech AI Demystified [S61523]

---

**NVIDIA Metropolis**
NVIDIA is demonstrating how developers can leverage NVIDIA software and generative AI to merge digital twins and real-time simulations, facilitating the testing and refinement of robots and their interactions within industrial environments. Combining developer platforms like NVIDIA Omniverse, Metropolis, Isaac, and cuOpt, users can create an "AI gym"

environment to train AI agents built using NIMs, aiding both robots and humans in adapting to unpredictable scenarios and navigating complex surroundings.

NVIDIA also announceed Visual Insight Agent (VIA), a collection of workflows to build AI agents capable of processing large amounts of live or archived videos and images with Vision-Language Models (VLM) - whether deployed at the edge or cloud. This new generation of visual AI agents will help nearly every industry summarize, search, and extract actionable insights from video using natural language. Access to early access software will be available in Q2 this year for targeted accounts. Early access program is now open for applications.

**GTC Sessions:**

- Harnessing Generative AI and Large Language Model With Vision AI Agents [S62384]
- Augmenting Vision AI With Large Language Model Interfaces to Improve Productivity [S62394]
- A New Class of Cloud-Native Applications at the Far Edge With Generative AI [S62387]
- Harnessing Generative AI and Large Language Model With Vision AI Agents [S62384]
- Leveraging Microservices for Building Complex and Large-Scale Vision AI Apps [S62395]

**Resources:** Staying in Sync: NVIDIA Combines Digital Twins With Real-Time AI for Industrial Automation, Metropolis Corporate Page, Metropolis Developer Page, VIA Product Page, VIA Early Access Program Page

---

**Data Science - CUDA-X Microservices** *Available with NVIDIA AI Enterprise license*
CUDA-X microservices provide end-to-end building blocks for data preparation, customization and training to speed production AI development across industries. They include technologies, libraries, and tools for development that are packaged as APIs. Examples of CUDA-X microservices:

- NVIDIA Riva for customizable speech and translation AI
- NVIDIA cuOpt for routing optimization, as well as NVIDIA Earth-2 for high resolution climate and weather simulations.
- NeMo Retriever microservices let developers link their AI applications to their business data — including text, images and visualizations such as bar graphs, line plots and pie charts — to generate highly accurate, contextually relevant responses.

**Availability:** Available now
**Resource:** CUDA-X Webpage

---

**Data Science - CUDA-X Data Processing** *Available with NVIDIA AI Enterprise license*
Announcing CUDA-X Data Processing, a collection of GPU-accelerated libraries that speed up and scale out processing of image, text, tabular, and sensor data for AI workloads.

- Accelerates data processing operations by up to 150x, reducing operation processing from days to hours
- Easy integration into existing data pipelines with APIs that match popular CPU-only python, CV, and distributed computing libraries.
- Scalable from laptop to PC to workstation to data center and the cloud

**Availability**: Available today on NVIDIA Developer and Github; soon to be available on HP AI Studio summer 2024
**Resources**: CUDA-X Webpage, CUDA-X Developer Page, HP and NVIDIA Supercharge Data Processing Press Release
**GTC Sessions**:

- RAPIDS in 2024: Accelerated Data Science Everywhere [S62741]
- Accelerating Pandas with Zero Code Change using RAPIDS cuDF [S62168]
- Accelerate ETL and Machine Learning in Spark [S62257]
- Delivering Personalized Promotions at Retail POS Checkout With AI-Powered Recommendation Engine [S62947]
- Accelerating NetworkX: The Future of Easy Graph Analytics [S61674]
- Harnessing Grace Hopper's Capabilities to Accelerate Vector Database Search [S62339]
- Streamed Video Processing for Cloud-Scale Vision AI Services [S61437]
- NVIDIA GPU Video Technologies: New Features, Improvements, and Cloud APIs [S61767]
- Building GPU-Accelerated Streaming AI Pipelines With Holoscan SDK [S61294]

---

**Data Science - RAPIDS cuDF Accelerated pandas** *Available with NVIDIA AI Enterprise license*
RAPIDS cuDF accelerates pandas - the world's most popular Python data analytics library

- **Optimized Performance:** Up to 150x faster than CPU-only pandas, reducing processing time from hours to minutes
- **Zero Code Change Acceleration:** Just load the cudf.pandas extension at the top of your code or use a Python module flag on the command line.
- **Unified CPU/GPU Workflows:** Develop, test, and run in production with a single code path, regardless of hardware.

**Availability:** RAPIDS cuDF for pandas code now Generally Available as open source - download from rapids.ai; available on AI Workbench shortly after GTC; available on HP AI Studio in Summer 2024.
**Resources:** RAPIDS cuDF Accelerates pandas Nearly 150x with Zero Code Changes, cudf.pandas Open Source Landing Page
**GTC Session:** Accelerating Pandas with Zero Code Change using RAPIDS cuDF [S62168]

---

**NVIDIA HPC SDK - Update v24.3**
The NVIDIA HPC SDK is a comprehensive suite of compilers, libraries, and tools for developing accelerated HPC applications. The breadth of flexible support options enables users to create applications with the programming model that is the most relevant to their situation.

In addition to bug fixes and improvements in compile-time performance of the HPC Compilers, new HPC SDK 24.3 is released with new features supporting better development on NVIDIA's latest Grace Hopper systems. The NVIDIA HPC Compilers provide a unified memory compilation mode when using OpenMP Target Offload directives for GPU programming. This adds to the existing support for Grace Hopper and HMM systems unified memory in the OpenACC, CUDA Fortran, and Standard Parallelism (stdpar) programming models which is enabled in nvc++ and nvfortran via the -gpu=unified command line flag.

Additionally for CUDA Fortran programs, the 'unified' attribute has been added to provide additional type information that enables applications to be further optimized for unified memory systems like Grace Hopper. See the Docs for more information and details. Updates are made to the CUDA-X Math Libs throughout the year as well.

**Pricing:** Free download, HPC SDK
**Enterprise Support**: NVIDIA Enterprise Support is available for HPC Compiler
**Resources:** HPC SDK Webpage Articles: HPCwire Why Standards-Based Parallel Programming Should be in Your HPC Toolbox, HPCwire Leveraging Standards-Based Parallel Programming in HPC Applications, HPCwire New C++ Sender Library Enables Portable Asynchrony, Blog: Developing Accelerated Code with Standard Language Parallelism
**GTC Sessions:** For HPC-related sessions, please see the FAQs.

---

### CUDA Toolkit, Release 12.4

The latest release of CUDA Toolkit, version 12.4, continues to push accelerated computing performance using the latest NVIDIA GPUs. CTK 12.4 enhances support for NVIDIA Grace Hopper and Confidential Computing. New features and enhancements in this release include:

- CUDA driver update (R550)
- Access-counter-based memory migration for NVIDIA Grace Hopper systems
- Confidential computing support
- CUDA graphs conditionals
- CUB performance improvements
- Compiler updates
- Enhanced monitoring capabilities
- Enhanced NVIDIA Nsight Compute and NVIDIA Nsight Systems developer tools

**Availability:** Downloadable since March 6, 2024. License included with every NVIDIA GPU purchase.
**Resources:** CUDA Toolkit Website, DevForum for CUDA, NGC Catalog for CUDA Toolkit Downloads
**GTC Sessions:**

- CUDA: New Features and Beyond [S62400]
- How to Write a CUDA Program – The Ninja Edition [S62401]
- Introduction to CUDA Programming and Performance Optimization [S62191]
- Advanced Performance Optimization in CUDA [S62192]

---

### NVIDIA Warp

Warp is a Python framework for writing high-performance simulation and graphics code. Warp takes Python functions and JIT compiles them to efficient kernel code that can run on the CPU or GPU. Additionally, Warp uses automatic differentiation to integrate with ML frameworks.

The GTC Talk will cover the latest features in Warp for spatial computing, 3D data generation, computer-aided engineering, and robotics, and show how Warp seamlessly connects to machine learning frameworks such as PyTorch and JAX. Miles Macklin will illustrate these concepts with in-depth examples, including trajectory optimization for aerial vehicles, finite-element analysis, and large-scale computational fluid dynamics.

- New support for asynchronous CUDA allocators for fast array creation
- New support for native C++ snippets, allowing developers to access advanced CUDA features directly from Python
- Expanded Warp simulation library with Featherstone integrators for robotics
- New Warp FEM library for partial differential equations problems (PDEs) such as diffusion, fluid flow, and heat transfer
- Enhanced PyTorch, JAX, and CUTLASS compatibility and integration

**Availability:** Version 1.0.1 available today

**Resources:** Webpage, TechBlog Post, GitHub

**GTC Session:**

- Warp: Advancing Simulation AI with Differentiable GPU Computing in Python [S63345]

---

**Accelerated Python (Library)**

NVIDIA's collection of tools, libraries, & projects connecting Python to the GPU

- CUDA Python
- JAX, PyTorch & TensorFlow frameworks
- Legate runtime with Python libraries (cuNumeric, JAX, etc.)
- RAPIDS / Data Sciences
- NVIDIA Warp

**Availability:** Available for download today.

**Resource:** https://github.com/NVIDIA/cuda-python

**GTC Sessions:**

- Legate: A Productive Programming Framework for Creating Composable, Scalable, Accelerated Libraries [S62262]
- Legate Programming Framework for Creating Composable, Scalable, Accelerated Libraries [CWE63651]
- Fundamentals of Accelerated Computing with CUDA Python [DLIW61560]
- Warp: Advancing Simulation AI with Differentiable GPU Computing in Python [S63345]
- Profilers, Python, and Performance: Nsight Tools for Optimizing Modern CUDA Workloads [DLIT61667]
- More Data, Faster: GPU Memory Management Best Practices in Python and C++ [S62550]
- No More Porting: Accelerated Computing With Standard C++, Fortran, and Python [S61204]
- Accelerating NetworkX: The Future of Easy Graph Analytics [S61674]

---

**Math Libraries**

NVIDIA CUDA GPU-accelerated math libraries lay the foundation for compute-intensive applications in areas including AI and computational sciences. The latest CUDA math libraries releases and updates are as follows:

- cuDSS (Preview) is a GPU-accelerated direct sparse solver library for solving linear systems of very sparse matrices, common in autonomous driving and process simulations
- cuBLAS provides GPU-accelerated basic linear algebra subroutines. Available in CUDA Toolkit 12.4, cuBLAS adds Grouped Batched GEMM (General Matrix Multiplication) experimental

support allowing you to concurrently solve GEMMs of different dimensions, leading dimensions, and scaling factors

- cuBLASDx improves application performance by fusing numerical operations within a CUDA kernel. And cuFFTDx provides this same functionality for fast-fourier transforms, frequently used in deep learning and computer vision applications. Both libraries are available now for standalone download
- cuTENSOR 2.0 provides optimized routines for tensor computations that accelerate training and inference for neural networks. Version 2.0 upgrades the library in both performance and functionality, including support for just-in-time kernel compilation
- cuBLASMp (Preview) is a high performance, multi-process library for distributed basic dense linear algebra – available in the HPC SDK and for standalone download. NVIDIA also provides cuSOLVERMp for solving distributed dense linear systems and eigenvalue problems, as well as cuFFTMp to solve fast fourier transforms on multi-GPU multi-node platforms

NVIDIA Performance Libraries (NVPL) provide drop-in replacements for the industry standard math libraries many applications use today. NVPL is optimized for the Grace CPU and enables you to port applications to the Grace architecture with no source code changes required. NVPL is available now in HPC SDK 24.3. This release includes BLAS and LAPACK libraries implementing the Netlib API, an FFT library implementing the FFTW API, and random number generator and sparse matrix BLAS libraries. NVPL is also available for standalone download, which includes NVPL TENSOR for accelerating deep learning and inference on Grace CPUs.

**Availability:** Available now
**Resources:** Math Libraries Webpage, NVPL Webpage, cuTENSOR 2.0 Guide, Accelerating GPU Applications with NVIDIA Math Libraries

---

**Nsight Developer Tools**
As applications scale across data centers and clouds, NVIDIA Nsight Developer Tools are evolving to help. New features are being introduced to Nsight Systems 2024.2 to help you build and optimize microservices.

- Profiling support has been enhanced for container systems like Kubernetes and Docker, including CSP Kubernetes services from major providers including Azure, Amazon, Oracle, and Google.
- Python scripts called “recipes” allow you to do single and multi-node analysis as applications execute across the data center. Nsight Systems then visualizes key metrics using Jupyter Lab integration. Available now, recipes for networking analysis reveal how compute cold-spots relate to communication. You can generate multi-node heat maps that identify where to optimize InfiniBand and NVLink throughput for peak performance.
- Server development is enabled by a remote GUI streaming container. Nsight Systems also integrates seamlessly with Jupyter Lab, allowing you to profile code and view textual results directly in Jupyter, or launch the GUI streaming container for in-depth analysis.

**Availability:** Available now
**Resources:** Devtools Overview, Download, Tutorial Resources
**GTC Sessions:**

- Achieving Higher Performance From Your Multi-Node Application [S62388]
- Advances in Ray Tracing Developer Tools [S62398]
- Demystify CUDA Debugging and Performance with Powerful Developer Tools [S62256]

### NVIDIA cuLitho

Manufacturing computer chips requires a critical step called computational lithography – one of the largest compute workloads in semiconductor production, necessitating massive data centers. As silicon feature sizes become smaller, and the impacts of optical diffraction have to be offset, there arises a need to proactively manipulate mask patterns with optical proximity correction (OPC) or inverse lithography technology (ILT) to accurately image wafers.

NVIDIA cuLitho is a library with optimized tools and algorithms for GPU-accelerating computational lithography techniques (including OPC and ILT) by orders of magnitude over current CPU-based methods. This enables foundries to accelerate their fab development cycle time and deploy new solutions to continue semiconductor scaling.

**Benefits Include:**

- **Continued Future Silicon Scaling:** Faster OPC, and enabling new lithography innovations
- **Performance:** 40x-60x speed-up of producing semiconductor photomasks for different OPC flows
- **New Generative AI Support:** 2x speed-up on top of OPC flow acceleration
- **Productivity:** More masks per day, masks that took 2 weeks now are produced overnight
- **Savings:** 350 NVIDIA H100 GPU systems can now replace 40,000 CPU systems, accelerating time to solution, while simultaneously reducing costs, space, and power

**Resources:** cuLitho Webpage, Update Release GTC'24, Intro Press Release GTC23
**GTC Session:** Accelerating Computational Lithography: Enabling our Electronic Future [S52510]

### Quantum Computing - NVIDIA Quantum Cloud

NVIDIA Quantum Cloud is a cloud platform and microservices for quantum computing

- Seamless access to the most powerful computing platform - workloads run on GPUs and QPUs in the cloud, so customers don't need to have their own.
- Pre-built Quantum Cloud APIs - can be easily integrated from applications and experiences.
- Integrated developer experience - It's easy to build, verify and publish CUDA quantum applications with a developer experience in the cloud that supports the entire development cycle.

**Availability:** Early access applications open now.
**Resources:** Landing Page, Press Release, Corp Blog
**GTC Session**: Defining the Quantum-Accelerated Supercomputer [S62139]

### NVIDIA ACE *Available with NVIDIA AI Enterprise license*

NVIDIA ACE, NeMo, and RTX Neural Rendering technologies help developers bring digital humans to life for games, healthcare, and customer service applications

- The digital human technologies suite includes ACE for speech and animation, NeMo for language, and RTX rendering SDKs for graphics.
- ACE enables developers to create digital humans capable of AI-powered natural language interactions, making conversations more realistic and engaging, driving the future of healthcare assistants, gameplay experiences, and digital customer service.

**Availability:** Production NIMs (Audio2Face & ASR) available through ai.nvidia.com. In development NIMs available through ACE early access program at developer.nvidia.com/ace.

**Resources:** NVIDIA ACE, NVIDIA NeMo, NVIDIA RTX, FAQs, Press Release, Corp Blog, Tech Blog

**GTC Sessions:**

- Creating Lifelike Expressions in Digital Humans [S63409]
- Enhancing the Digital Human Experience with Cloud Microservices Accelerated by Generative AI [S63457]
- Behind the Scenes of Running a Conversational Character in a 3D Scene [S62570]
- How NVIDIA Accelerates Retailers on their Generative AI Journey [S62295]

---

## NVIDIA Omniverse Platform Updates

NVIDIA Omniverse is a platform of APIs, services, and software development kits (SDKs) that enable developers to build generative AI-enabled tools, applications, and services for industrial digitalization. And, to fully align with the platform's positioning as a development platform, we have released new licensing updates:

- **Pricing Update: NVIDIA Omniverse Enterprise**
  - List pricing: $4,500 per GPU per year
  - Includes NVIDIA Business Standard Support
  - Price independent of GPU type (A10G, L40 etc.)
  - Standard EDU and Inception program discounts apply
  - For Omniverse software that doesn't require a GPU (e.g. Nucleus), we require 1 GPU license per system/instance
- **Pricing Update: NVIDIA Omniverse Cloud Enterprise PaaS**
  - List pricing: $495,000 per OVX node per year
  - Includes NVIDIA Business Critical Support and Technical Account Manager
  - Includes OVX compute and networking
  - Available for purchase on Azure Cloud Marketplace
  - Managed by NVIDIA

**Availability:** Now available

**Resources:** Omniverse Enterprise Customer Deck

**GTC Sessions:**

- OpenUSD Day Conference Sessions
- Industrial Digitalization Conference Sessions
- Alliance for OpenUSD Panel [S62782]

---

## New NVIDIA Omniverse Cloud APIs

NVIDIA Omniverse Cloud APIs are a set of simple APIs that let developers integrate Omniverse technologies directly into their existing software applications for digital twins, or their simulation workflows for testing and validating autonomous machines like robots or self-driving vehicles. This is not to be confused with last year's announcement of Omniverse Cloud Platform-as-a-Service (PaaS) which is a fully managed service for developing and deploying Omniverse Kit-based applications.

There are 5 new Omniverse Cloud APIs which can be used individually or collectively, including:

- **USD Render:** Generates fully ray-traced RTX renders of OpenUSD data
- **USD Write:** Lets users modify and interact with OpenUSD data
- **USD Query:** Enables scene queries and interactive scenarios
- **USD Notify:** Tracks USD changes and provides updates
- **Omniverse Channel:** Connects users, tools, and worlds to enable collaboration across scenes

**Availability:** Omniverse Cloud APIs will be offered later this year on Microsoft Azure as self-hosted APIs on NVIDIA A10 GPUs, or as managed services deployed on NVIDIA OVX.
**Pricing:** For self-hosted APIs, developers will need an Omniverse Enterprise license per GPU per year; pricing for APIs as a managed service is not yet available
**Resources:** Press Release
**GTC Sessions:**

- OpenUSD Day Conference Sessions
- Industrial Digitalization Conference Sessions
- Alliance for OpenUSD Panel [S62782]

---

**NVIDIA Omniverse Cloud APIs for Sensor Simulation**
NVIDIA Omniverse Cloud APIs will enable high-fidelity sensor simulation to accelerate autonomous machine development. Sensor data is critical for AI systems' perception capabilities, enabling them to comprehend their environment and make informed decisions in real time. The APIs bring together a rich ecosystem of simulation tools, applications and sensors for developers to integrate physically based sensor simulation and behavior into their existing workflows. This ecosystem includes CARLA, Foretellix, IPG, MathWorks, MITRE, Parallel Domain, and Voxel51 in addition to a wide range of camera, radar, and lidar partners
**Pricing:** For self-hosted APIs, developers will need an Omniverse Enterprise license per GPU per year; pricing for APIs as a managed service is not yet available
**Availability:** Omniverse Cloud APIs will be offered later this year on Microsoft Azure as self-hosted APIs on NVIDIA A10 GPUs, or as managed services deployed on NVIDIA OVX.
**Resources:** Press Release, Blog
**GTC Sessions:**

- Advancing Autonomous Vehicle Simulation with Omniverse and Generative AI [SE63005]
- Open-Source Autonomous Vehicle Simulation With CARLA [S62468]
- Scaling Up: Achieving Mass Commercial AV Deployment With Real-World and Virtual Validation [S63237]
- Addressing AV Deployment Policy Issues and Introducing a New Digital Proving Ground [S62514]

---

**Earth-2**

NVIDIA Earth-2 is a full-tech stack offered as an open platform and as a suite of cloud services that accelerates high-resolution climate and weather simulations, climate and weather predictions, augments them with AI models, and enables interactive visualization of large scale data from multiple data sources. It includes physical CUDA accelerated simulation of numerical weather models like ICON and IFS; machine learning weather prediction models such as FourCastNet, GraphCast, and Deep Learning Weather Prediction (DLWP) through ; and data federation and visualization with NVIDIA Omniverse™. Running on NVIDIA DGX™ GH200, HGX™ H100, and OVX™ supercomputers, Earth-2 will provide a path to simulate and visualize the global atmosphere at unprecedented speed and scale.

Today NVIDIA announced the Earth-2 platform that enables the Earth Climate Digital Twin and Earth-2 Cloud APIs. Releasing 3 cloud API services:

- **AI as-a-service:** CorrDiff (first-of-its-kind km-scale genAI diffusion model): Regional Downscaling from 25 to 2km; 2km simulations require 25,000x more compute than 25 km simulations, computational growth that happens over 20+ years. Sidestepped by generative AI
    - 1000x faster and 2000x more energy efficient than NWP (numerical weather prediction)
    - Ability to generate massive ensembles essentially for free
    - Generative AI goes beyond just weather variables. It learns physics and synthesizes new data for insights and decision-making.
- **Simulation as-a-service:** GPU-accelerated acceleration provides 24X performance, 10x energy efficiency, 5X lower TCO
- **Visualization as-a-service:** Earth-2 viz service and Omniverse Digital Twins, to simulate the impact of weather on operations, facilities, assets, and people, creating enhanced industry resilience to weather events across a wide variety of short- and long-term scenarios to optimize decision-making, training, and planning.

**Availability:** AI service available in early-access now – request from Earth-2 Webpage
**Resources:** Earth-2 Webpage, Earth-2 Press Release
**GTC Sessions:**

- Sustainable Computing GTC Sessions
- Toward Kilometer-Scale Machine Learning Emulation of the Earth System [S62301]

---

**NVIDIA DRIVE**
NVIDIA DRIVE Powers Next Generation of Transportation — From Cars and Trucks to Robotaxis and Autonomous Delivery Vehicles
**Availability:** More details will be announced at a later date
**Resources:** Press Release

---

**NVIDIA IGX Orin**
NVIDIA AI Enterprise – IGX is now generally available. It offers an enterprise-grade software solution for NVIDIA IGX Orin platforms, providing unmatched performance, security, stability, and support for the entire software stack that consists of AI frameworks, SDKs, libraries, AI safety software, and enterprise drivers. Key benefits include:

- Enterprise-Grade Reliability, Security, and Support

- Optimized for Accelerated Mission-Critical AI
- Accelerated Time to Market

**Availability:** Available now in the NGC Public Catalog
**Resources:** Announcement Blog
**GTC Sessions:**

- Edge Computing 101: An Introduction to the Smart Edge [S62568]
- Functional Safety for Industry 4.0: Keeping Supply Chains Safe, Secure, and Efficient using AI at the Edge [S62299]

---

**NVIDIA Jetson Thor**
Built for deploying robotic foundational models on humanoid robots, Jetson Thor, based on NVIDIA Blackwell Architecture, is a new computing platform capable of performing complex tasks and interacting safely and naturally with people and machines. It has a modular architecture optimized for performance, power, and size. Jetson Thor has the following key features:

- GPU: Up to 8x Orin FP8 Transformer Performance
- CPU: Up to 2.6X Orin CPU Performance
- Memory: Up to 2x DRAM Capacity
- 10x Greater I/O: with up to 4x 25 Gbps Ethernet Ports

**Availability:** H1 2025
**Resources:** Corporate Blog, Press Release
**GTC Session:** AI Robotics: Driving Innovation for the Future of Automation [S63287]

---

**NVIDIA Isaac Perceptor**
NVIDIA Isaac™ Perceptor is a collection of hardware-accelerated packages for visual AI, tailored for Autonomous Mobile Robots (AMR) to perceive, localize, and operate robustly in unstructured environments. Robotics software developers can now easily access turnkey AI-based perception capabilities, ensuring reliable operations and obstacle detection in complex scenarios.
**Availability:** The developer preview is planned for Q2 2024
**Resources:** Product Page, Corp Blog: NVIDIA Isaac Taps Generative AI for Manufacturing and Logistics Applications, FAQs

---

**NVIDIA Isaac Manipulator**
Isaac Manipulator is a collection of foundation models and modular GPU-accelerated libraries that help build scalable and repeatable workflows for dynamic manipulation tasks by accelerating AI model training and task (re)programming. Isaac Manipulator brings new levels of dexterity and modular AI capabilities to robotic arms that face limitations in handling intricate tasks and dynamic environments due to their limited adaptability and manual re-tasking processes for every new scenario.
**Availability:** The developer preview is planned for Q2 2024
**Resources:** Product Page, Corp Blog: NVIDIA Isaac Taps Generative AI for Manufacturing and Logistics Applications

**NVIDIA Project GR00T**
GR00T is a general-purpose foundation model that promises to transform humanoid robot learning in simulation and the real world. Trained in NVIDIA GPU-accelerated simulation, GR00T enables humanoid embodiments to learn from a handful of human demonstrations with imitation learning and NVIDIA Isaac Lab for reinforcement learning, as well as generating robot movements from video data.
**Resources:** Product Page, Press Release, Corp Blog: NVIDIA Isaac Taps Generative AI for Manufacturing and Logistics Applications

**NVIDIA Isaac Lab**
Isaac Lab is a lightweight reference application built on the Isaac Sim platform specifically optimized for robot learning and is pivotal for robot foundation model training. Isaac Lab optimizes for reinforcement, imitation, and transfer learning, and is capable of training all types of robot embodiments, including the Project GR00T foundation model for humanoids.
**Resources:** Corp Blog: NVIDIA Isaac Taps Generative AI for Manufacturing and Logistics Applications

**NVIDIA OSMO**
OSMO is a cloud-native workflow orchestration platform that lets you easily scale your workloads across distributed environments—from on-premises to private and public clouds. It provides a single pane of glass for scheduling complex multi-stage and multi-container heterogeneous computing workflows.
**Resources:** Product Page, Technical Blog, Corp Blog: NVIDIA Isaac Taps Generative AI for Manufacturing and Logistics Applications
**Availability:** Customers can apply for early access on the product page.

**NVIDIA Clara - Healthcare and Life Sciences**

**Healthcare Microservices**
NVIDIA Clara now offers a suite of microservices that accelerates the building applications across computer-aided drug discovery, medtech, and digital health
BioNeMo. With over 25 microservices for healthcare at launch, including NVIDIA NIM for optimized inference models, we enable enterprise application developers at ISVs like Abridge, Hippocratic, Iambic, and Cadence Molecular sciences to augment their platforms with generative AI as they deploy to production.

**BioNeMo** *Available with NVIDIA AI Enterprise license*
BioNeMo introduces microservices, including NVIDIA NIM for optimized inference, and new foundation models to help accelerate drug discovery with generative AI.

- BioNeMo is now available as NVIDIA microservices, including NVIDIA NIM, offering domain-specific models accessible via API to build drug discovery applications.
- Users of BioNeMo microservices include Iambic and Cadence Molecular Sciences
- New BioNeMo foundation models are now available that can analyze DNA sequences and predict how proteins will interact with drug molecules, and a model that determines a single cell's function based on its RNA is coming soon.

**Availability:** Available at GTC
**Resources: Resources:** NVIDIA Healthcare Microservices Press Release, NVIDIA BioNeMo Corporate Blog, BioNeMo Webpage
**GTC Sessions:**

- Accelerated Computer-Aided Drug Discovery with BioNeMo [S62321]
- Lab in a Loop: AI to Transform Drug Discovery and Development [S62281]
- AI-Driven Drug Discovery Panel: Unraveling Biological Complexities [S62583]

**Parabricks 4.3** *Available with NVIDIA AI Enterprise license*
Parabricks 4.3 expands our genomics analysis software suite with the introduction of new tooling and workflows that bring acceleration and the latest AI techniques to multiple omics data types.

- NVIDIA Microservices now available for DeepVariant and FQ2BAM
- New workflow for single-cell and spatial omics data to enable rapid and high accuracy analysis, and new tool to accelerate sequencing alignment for DNA methylation data
- Further optimized germline analysis, bringing germline analysis down to under 10 minutes on DGX H100

**Enterprise Support**: NVIDIA Enterprise Support is available with NVIDIA AI Enterprise software for production AI deployments.
**Availability:** Available at GTC
**Resources:** NVIDIA Healthcare Microservices Press Release, Parabricks 4.3 Blog, Parabricks Web Page
**GTC Sessions:**

- Introduction to GPU-Accelerated Genomics with Parabricks [S62322]
- First-Ever Whole Transcriptome Imaging of Tissues using CosMx-SMI: Highest-Density Dataset Ever Collected [S61995]

**NVIDIA MONAI** *Available with NVIDIA AI Enterprise license*
NVIDIA MONAI microservices are creating unique state-of-the-art models and expanded modalities to meet the demands of the healthcare and biopharma industry.

- NVIDIA Microservices include NVIDIA NIM, optimized inference models for advanced imaging; VISTA-3D and 2-D for segmentation; MAISI for synthetic data generation
- CUDA-X Microservices for medical imaging enable AI-assisted annotation, inference and training, and Auto 3D segmentation workflows
- Flywheel and V7 are announcing integration with MONAI microservices

**Availability:** VISTA-3D NIM available at GTC; VISTA-2D and MAISI NIM in early access
**Resources:** NVIDIA Healthcare Microservices Press Release, MONAI Tech Blog (Live 3/19), NVIDIA MONAI Microservices Deck, MONAI Technical Deck, Healthcare Life Sciences Industry Deck, NVIDIA MONAI Page, FAQs
**GTC Sessions:**

- Learn how to Streamline Medical Imaging Annotation and Training workflows with MONAI Cloud APIs [S62320]
- AI Factory - Accelerating Research and Innovation in the Biopharmaceutical Industry [S62532]

- Weeks to Minutes: Improve Analysis Cycle Time in Drug Development with MONAI Label [S62502]

**Holoscan 1.0** *Available with NVIDIA AI Enterprise license*
NVIDIA Holoscan 1.0 helps developers more easily build production-ready applications for multimodal, real-time AI sensor processing. Holoscan 1.0 is now available with a NVIDIA AI Enterprise for IGX license, offering enterprise support with guaranteed API stability for the Holoscan software stack.
**Resources:** Holoscan 1.0 Dev Blog, Johnson and Johnson Corporate Blog, NVIDIA Holoscan Webpage
**GTC Sessions:**

- NVIDIA Holoscan, the AI Sensor Processing Platform, from Surgery to Satellites [S62323]
- Charting the Future of AI in MedTech [S62680]
- Realizing Augmented Reality's Benefits in Surgery through Real-Time Edge AI [S61354]
- Accelerating Development of Surgical Robotics with AI through NVIDIA's Ecosystem [S62318]

---

**6G Research Cloud Platform**
The NVIDIA 6G Research Cloud Platform is an open, flexible and interconnected platform, offering researchers a comprehensive suite of wireless and AI frameworks to advance AI for radio access network (RAN) technology. It gives developers the tools they need to successfully evolve current wireless stacks, deliver AI-native 6G and accelerate the development of 6G technologies. There are three key elements in the platform:

- **Aerial CUDA Accelerated RAN:** A framework for building commercial-grade, software-defined, and cloud-native 6G networks. It includes GPU-accelerated interoperable PHY and MAC layer libraries that can be easily modified and seamlessly extended with AI components.
- **Aerial Omniverse Digital Twin:** A system-level simulator for cutting edge 6G research and development of next-generation wireless systems, applying ray-traced channels to the PHY and MAC layers of NVIDIA Aerial RAN and simulated UEs.
- **Sionna Neural Radio Framework:** A Python programming framework to extend the GPU-accelerated PHY, MAC layer, and CUDA libraries for 5G/6G. It provides seamless integration with popular frameworks like PyTorch and TensorFlow, leveraging NVIDIA GPUs for generating and capturing data and training AI and machine learning models at scale. This also includes NVIDIA Sionna, the leading link-level research tool for AI/ML-based wireless simulations.

Free to academic and industry 6G researchers.
**Availability:** Mid April 2024
**Resources:** Press Release, Aerial Developer Page, Technical Blog
**GTC Sessions:**

- Telecom Special Address: Three Ways Artificial Intelligence is Transforming Telecommunications [S62965]
- Digital Twin Simulations for 6G Networks [S62343]
- Democratizing AI-RAN and 6G Research [S62784]

---

**Project Legate**
The Legate project team is giving a talk at GTC about a new programming framework for developers to achieve accelerated parallel computing across multi-GPU and multi-node

(MGMN) configurations using composable, scalable and accelerated libraries. Legate facilitates productive development of scalable, accelerated software packages targeting a variety of NVIDIA systems, from single-GPU cards to Grace Hopper Superchip and DGX Cloud. Libraries written with Legate interoperate seamlessly with each other, enabling true end-to-end acceleration of workflows.

- A programming framework for multi-GPU and multi-node (MGMN) configurations
- Transparently accelerates and scales existing JAX, NumPy and RAPIDS workflows
- Scales to up to thousands of GPUs optimally
- Requires minimal code changes to ensure developer productivity

**Availability:** Early access available today

**Resources:** Webpage, Documentation, TechBlog Post, GitHub

**GTC Sessions:**

- Legate: A Productive Programming Framework for Creating Composable, Scalable, Accelerated Libraries [S62262]
- Legate Programming Framework for Creating Composable, Scalable, Accelerated Libraries [CWE63651]

---

**NVIDIA Training - Gen AI Certification**

Training from NVIDIA is designed to help customers address their AI, HPC, and graphics skill development needs. We offer the combined courses, workshops, and certification exams from the Deep Learning Institute (DLI) and NVIDIA Academy. For the first time at GTC, we are offering certification. A limited number of on-site attendees can take one of our certification exams at a 50% discount.

NVIDIA is launching two new Gen AI certification exams at GTC:

- NVIDIA Certified Associate - Gen AI LLMs: Entry-level credential covering the developing, integrating, and maintaining AI-driven applications using Gen AI and LLMs with NVIDIA solutions
- NVIDIA Certified Associate - Gen AI Multimodal: Entry-level credential covering the foundational skills needed to design, implement, and manage AI systems that synthesize and interpret data across text, image, and audio modalities

**Availability:** Generally available on NVIDIA Training website

**Resources:** Certification Website, Corporate Blog, Training and Certification at GTC

**GTC Session:** Get NVIDIA Certified at GTC [DLIT62949]